ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΒΑΘΜΟΛΟΓΗΣΗΣ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ

Μαρία Γ. Βασιλάκη

Διπλωματική Εργασία
που υποβλήθηκε στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάιος 2010

Η παρούσα διπλωματική εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ……. συνεδρίαση σύμφωνα με τον Εσωτερικό Κανονισμό λειτουργίας του Προγράμματος Μεταπτυχιακών σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Αναπληρωτής Καθηγητής Αγιακλόγλου Χρήστος
- Επίκουρος Καθηγητής Πολίτης Κωνσταντίνος

Η έγκριση της Διπλωματικής Εργασίας στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

CREDIT SCORING MODELS

by

Maria G. Vasilaki

MSc Dissertation

Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

May 2010

Στην Οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που συνέβαλαν με τον τρόπο τους στην ολοκλήρωση της διπλωματικής μου εργασίας. Κατ' αρχάς, θέλω να ευχαριστήσω ιδιαίτερα τον Επιβλέποντα Καθηγητή κ. Μάρκο Κούτρα, αρχικά γιατί δέχτηκε να αναλάβει την επίβλεψη της εργασίας μου και μου έδωσε την ευκαιρία να ασχοληθώ με ένα ιδιαίτερα ενδιαφέρον θέμα. Επιπλέον, η εξαιρετική βοήθεια, οι συμβουλές και η υποστήριξη που μου προσέφερε από τη στιγμή της ανάθεσης του θέματος έως την ολοκλήρωση της παρούσας εργασίας ήταν πολύτιμες και αποτέλεσαν οδηγό στην πορεία της εργασίας αυτής. Δεν θα μπορούσα να παραλείψω να ευχαριστήσω τα μέλη της τριμελούς επιτροπής, τον Αναπληρωτή Καθηγητή κ. Αγιακλόγλου Χρήστο και τον Επίκουρο Καθηγητή Πολίτη Κωνσταντίνο για τον πολύτιμο χρόνο που αφιέρωσαν στη μελέτη και κριτική της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για την αγάπη τους και την αμέριστη υλική και ηθική υποστήριξη που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου.

.

Περίληψη

Τα τελευταία χρόνια και ιδιαίτερα μετά από το νέο κανονιστικό πλαίσιο που τέθηκε από τη Βασιλεία ΙΙ στα χρηματοπιστωτικά ιδρύματα για την κεφαλαιακή επάρκεια, τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας έχουν καταστεί απαραίτητα για την λήψη κρίσιμων αποφάσεων που αφορούν χρηματικές (και όχι μόνο) χορηγήσεις. Παρότι τα πρώτα μοντέλα στην περιοχή αυτή φαίνεται να ξεκίνησαν τη δεκαετία του 1960, υπάρχει διαρκής εξέλιξη με τη βοήθεια κυρίως της Στατιστικής αφού η εφαρμογή των μοντέλων γίνεται σε ένα σχετικά μεγάλο πλήθος παρελθοντικών δεδομένων με στόχο την πρόβλεψη της συμπεριφοράς των μελλοντικών πελατών του χρηματοπιστωτικού ιδρύματος. Τα στατιστικά μοντέλα βαθμολόγησης είναι κατάλληλα για δανειολήπτες που χρησιμοποιούν χρηματοδοτικά προϊόντα μικρού χρηματικού ποσού τα οποία όμως συνολικά είναι μεγάλου πλήθους (λιανική τραπεζική).

Στην παρούσα εργασία θα γίνει κριτική επισκόπηση της περιοχής των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας και θα παρουσιαστούν διάφορες στατιστικές τεχνικές που έχουν προταθεί για την ανάπτυξη πιθανοθεωρητικών/στατιστικών μοντέλων. Επιπλέον, θα γίνει μια σύντομη αναφορά στα πλεονεκτήματα και τα μειονεκτήματα κάθε τεχνικής. Τέλος θα αναπτυχθούν κάποια μοντέλα βαθμολόγησης με τις μεθόδους της Λογιστικής Παλινδρόμησης, της Διαχωριστικής Ανάλυσης και των Δέντρων Ταξινόμησης χρησιμοποιώντας πραγματικά δεδομένα.

Abstract

Nowadays and more particularly after the new regulatory framework set by Basel II for the banks capital adequacy, credit scoring models have become essential in obtaining critical decisions concerning financial lending. Despite of the fact that the initial models in this realm appear to have initiated in the 1960s, there is a constant development using mainly statistical methods, because the application of these models employs relatively large numbers of past data in an attempt to predict the behaviour of potential banks customers. Credit scoring is suitable for borrowers who are interested in low value transactions, but overall in large volumes (retail banking).

This dissertation will portray a critical overview of the credit scoring models and will present various statistical techniques which have been proposed for the development of probabilistic/statistical models. Additionally, there will be a brief reference to the pros and cons of each technique. Lastly, using real life data some credit scoring models exploiting the methods of Logistic Regression, Discriminant Analysis, and Classification Trees will be developed.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

AR	Accuracy Ratio – δείκτης ακρίβειας
ASM	Application Scoring Models – μοντέλα βαθμολόγησης αιτήσεων
AUC	Area Under Curve – εμβαδό κάτω από την καμπύλη ROC
BIS	Bank of International Settlements – τράπεζα διεθνών διακανονισμών
BSM	Behavioral Scoring Models – μοντέλα βαθμολόγησης συμπεριφοράς
CAP	Cumulative Accuracy Profile – συσσωρευτική καμπύλη ακρίβειας
CR	Credit Risk – πιστωτικός κίνδυνος
CRS	Credit Rating Systems – συστήματα πιστοληπτικής διαβάθμισης
CRT	Classification and Regression Trees – δέντρα ταξινόμησης και παλινδρόμησης
CSM	Credit Scoring Models – μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας
DR	Default Risk – Κίνδυνος αθέτησης υποχρεώσεων
EAD	Exposure At Default – χρηματοδοτικό άνοιγμα
EL	Expected Losses – αναμενόμενη ζημιά
IRB	Internal Ratimg Based models – μοντέλα εσωτερικής διαβάθμισης
LGD	Loss Given Default – ζημιά δεδομένης της αθέτησης
MAR	Missing At Random – δεδομένα που λείπουν τυχαία
MNAR	Missing Not At Random – δεδομένα που δεν λείπουν τυχαία
MSE	Mean Square Error – μέσο τετραγωνικό σφάλμα
PD	Probability of Default – πιθανότητα αθέτησης υποχρεώσεων
PIT	Point In Time – την στιγμή της ταξινόμησης
PSM	Profit Scoring Models – μοντέλα βαθμολόγησης κέρδους
RWA	Risk Weighted Assets – σταθμισμένο Ενεργητικό
ROC	Receiver Operating Characteristic Curve
RPA	Recursive Partitioning Algorithms – αλγόριθμοι επαναλαμβανόμενης διάσπασης
SA	Standardized Approach – τυποποιημένη μέθοδος
SMEs	Small Medium Enterprises – μικρομεσαίες επιχειρήσεις
TTC	Through The Cycle – κατά τη διάρκεια ενός πλήρους οικονομικού κύκλου
UL	Unexpected Loss – μη αναμενόμενη ζημιά
VaR	Value at Risk – Αξία σε κίνδυνο

ΓΠ	Γραμμική Παλινδρόμηση
ΔΑ	Διαχωριστική Ανάλυση
ΔΚΕ	Δείκτης κεφαλαιακής επάρκειας
ΔΤ	Δέντρα Ταξινόμησης
ΚΓ	Μέθοδος Κοντινότερου Γείτονα
ΛΠ	Λογιστική Παλινδρόμηση

ΚΕΦΑΛΑΙΟ 1

Μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας

1.1 Πιστωτικός κίνδυνος

Το σύγχρονο ευμετάβλητο χρηματοοικονομικό περιβάλλον και ο έντονος ανταγωνισμός μεταξύ των χρηματοπιστωτικών ιδρυμάτων έχουν καταστήσει αναγκαία την αποτελεσματική διαχείριση του πιστωτικού κινδύνου αφού η κύρια λειτουργία των ιδρυμάτων αυτών είναι η χορήγηση πιστώσεων σε ιδιώτες και επιχειρήσεις. Ο **πιστωτικός κίνδυνος (*credit risk* - CR)** αποτελεί τον κίνδυνο που μπορεί να αναλάβει ένας χρηματοοικονομικός οργανισμός και σχετίζεται άμεσα με την αξιολόγηση της δυνατότητας ενός δανειολήπτη να αποπληρώσει τις υποχρεώσεις του, βάσει των αμοιβαίως συμφωνηθέντων όρων που καλύπτουν την πίστωση αυτή.

Είναι πολύ σημαντικό για ένα χρηματοπιστωτικό ίδρυμα να γνωρίζει την πιστοληπτική ικανότητα του δανειολήπτη έτσι ώστε να μπορεί να καθορίζει την αμοιβή του από τον πιστωτικό κίνδυνο που αναλαμβάνει. Για παράδειγμα, μια τράπεζα μπορεί να χορηγήσει ένα καταναλωτικό δάνειο σε έναν ιδιώτη με ευνοϊκότερους όρους αν γνωρίζει ότι αυτός έχει μεγάλη πιστοληπτική ικανότητα. Επιπλέον, η γνώση της πιστοληπτικής ικανότητας των πελατών είναι ιδιαίτερα σημαντική διότι η απώλεια εσόδων λόγω της αύξησης των καθυστερήσεων αποπληρωμής των δόσεων αποτελεί απειλή για τη βιωσιμότητα του ιδρύματος αυτού. Ο πιστωτικός κίνδυνος όμως δεν είναι μόνο σημαντικός για τους δανειστές αλλά και για τους ίδιους τους δανειολήπτες καθώς και αυτοί θα πρέπει να γνωρίζουν την πιστοληπτική τους ικανότητα έτσι ώστε να μπορούν να υπολογίσουν το κόστος άντλησης κεφαλαίων και να δρουν ανάλογα. Επομένως, προκύπτει η αναγκαιότητα ανάπτυξης

μοντέλων βαθμολόγησης της πιστοληπτικής ικανότητας πελατών ενός χρηματοοικονομικού οργανισμού.

1.2 Εισαγωγικές έννοιες

Ως **μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας (*Credit Scoring Models - CSM*)** χαρακτηρίζουμε εκείνα τα στατιστικά μοντέλα, που έχουν ως στόχο να προβλέψουν μελλοντικές συμπεριφορές ήδη υπαρχόντων ή υποψηφίων πελατών και προκύπτουν από την ανάλυση μεγάλων στατιστικών δειγμάτων. Τα αποτελέσματα της ανάλυσης καθορίζουν έναν αλγόριθμο, ο οποίος βαθμολογεί τα στατιστικώς σημαντικά στοιχεία των αιτήσεων δίνοντας μια τελική **βαθμολογία (*score*)**, που σχετίζεται με την πιθανότητα κάποιος πελάτης να εξελιχθεί σε «καλό» ή «κακό». Αυτές οι πιθανότητες ή οι βαθμολογίες είναι το βασικότερο εργαλείο που χρησιμοποιούν οι δανειστές για να αποφασίσουν αν θα γίνει δεκτή ή όχι η αίτηση χορήγησης πίστωσης του υποψήφιου πελάτη.

Τα CSM χρησιμοποιούν δεδομένα από χαρακτηριστικά των δανειοληπτών, είτε για να μετρήσουν την **πιθανότητα αθέτησης των υποχρεώσεων (*probability of default - PD*)** ενός ιδιώτη ή μιας επιχείρησης είτε για να ταξινομήσουν τους δανειολήπτες σε διαφορετικές κατηγορίες πιστοληπτικής ικανότητας ως προς τον αναλαμβανόμενο βαθμό πιστωτικού κινδύνου. Σε γενικές γραμμές, η ανάπτυξη και η εφαρμογή ενός μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας αποσκοπεί στην αξιολόγηση της πιστοληπτικής ικανότητας κάθε υποψήφιου πελάτη με βάση τις πληροφορίες που μπορούν να αντληθούν για αυτόν από την υποβολή της αίτησής του για πίστωση. Το πρόβλημα έγκειται ουσιαστικά στην πρόβλεψη της μελλοντικής συμπεριφοράς μιας ομάδας πελατών που έχουν συγκεκριμένα κοινά χαρακτηριστικά, με βάση τη συμπεριφορά που επέδειξαν κατά το παρελθόν αντίστοιχες ομάδες πελατών με παρόμοια χαρακτηριστικά.

Για το σκοπό αυτό, λαμβάνεται ένα μεγάλο δείγμα δεδομένων σχετικά με την παρελθοντική συμπεριφορά διαφόρων ομάδων πελατών με γνωστά χαρακτηριστικά και στους οποίους η πίστωση έχει ήδη χορηγηθεί. Τα δεδομένα αυτά μπορεί να προέρχονται είτε από τα αρχεία της ενδιαφερόμενης επιχείρησης ή αν αυτά τα στοιχεία δεν επαρκούν υπάρχει η δυνατότητα να αποκτηθούν από τα αρχεία δημόσιων ή ιδιωτικών γραφείων που παρέχουν

υπηρεσίες στατιστικής παρακολούθησης διαφόρων χαρακτηριστικών της αγοράς που ονομάζονται **γραφεία πίστης (*credit bureaus*).** Στην Ελλάδα, χρησιμοποιούνται οι αρνητικές πληροφορίες που έχει ο Τειρεσίας και οι οποίες σχετίζονται με το αν ο πελάτης κατά το παρελθόν είχε προβλήματα στην εκπλήρωση των υποχρεώσεών του.

Όλα τα χαρακτηριστικά των διαφόρων ομάδων πελατών συγκρίνονται μεταξύ τους και συνδυάζονται στατιστικά για την εύρεση εκείνων των χαρακτηριστικών που ερμηνεύουν και προβλέπουν καλύτερα τον πιστωτικό κίνδυνο. Η στάθμιση της σπουδαιότητας των χαρακτηριστικών αυτών ολοκληρώνει την ανάπτυξη ενός απλού CSM. Στα περισσότερα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας η τελική βαθμολογία του υποψηφίου πελάτη διαμορφώνεται έτσι ώστε να προοιωνίζει την «ποιότητα» της αναμενόμενης συμπεριφοράς του. Συνήθως, όσο μεγαλύτερή είναι η βαθμολογία ενός πελάτη, τόσο μεγαλύτερη είναι η πιθανότητα αυτός να είναι «καλός».

Ως **βαθμολογία αποδοχής - απόρριψης** ή **σημείο αποκοπής (*cut-off score*)** θεωρούμε εκείνη την τιμή της βαθμολογίας κάτω από την οποία απορρίπτεται το αίτημα του πελάτη, ενώ για μεγαλύτερες βαθμολογίες γίνεται δεκτό το αίτημά του για τη λήψη της πίστωσης που ζητά. Όσο μεγαλύτερη είναι η βαθμολογία αποδοχής-απόρριψης, τόσο μεγαλώνει το ποσοστό των υποψηφίων πελατών που απορρίπτονται. Λαμβάνοντας υπόψη ότι οι στατιστικές μέθοδοι λήψης αποφάσεων ενέχουν τον κίνδυνο λάθους, ο σκοπός είναι να τεθεί η βαθμολογία απόρριψης σε τέτοιο επίπεδο ώστε να έχουμε μικρό ποσοστό απόρριψης χωρίς, ωστόσο, να αυξηθεί πολύ το ποσοστό των «κακών» πελατών που θα γίνουν αποδεκτοί.

Σε περίπτωση που προκύψουν νέα δεδομένα, θα πρέπει να επαναπροσδιοριστεί η βαθμολογία απόρριψης αναπτύσσοντας ξανά το μοντέλο (*redeveloping*), αξιοποιώντας την αυξανόμενη εμπειρία και μεριμνώντας ώστε να γίνονται δεκτοί οι πελάτες που μέχρι πρότινος θα απορρίπτονταν, αλλά που αποδεικνύεται εμπειρικά ότι έχουν υψηλή πιστοληπτική ικανότητα. Ταυτόχρονα πρέπει να καταβληθεί ιδιαίτερη προσοχή ώστε να απορρίπτονται οι πελάτες που, ενώ κατά το παρελθόν είχαν μεγάλη βαθμολογία τα νέα δεδομένα δείχνουν ότι έχουν χαρακτηριστικά που προοιωνίζουν μεγάλο πιστωτικό κίνδυνο.

Η βελτίωση του αναπτυσσόμενου μοντέλου πιστοληπτικής ικανότητας είναι αναγκαία και προϋποθέτει τη συστηματική παρακολούθηση της εξέλιξης των λογαριασμών των πελατών που έχουν γίνει αποδεκτοί, ώστε να ανακαλυφθούν τυχόν προβλεπτικές αδυναμίες του στατιστικού μοντέλου με στόχο την αναπροσαρμογή του, με την ενσωμάτωση όλης της συσσωρευμένης πληροφορίας.

Οι παραδοσιακές μέθοδοι που υποστηρίζουν την απόφαση αν θα χορηγηθεί πίστωση σε κάποιον πελάτη της επιχείρησης ή όχι βασίζονταν στην ανθρώπινη κρίση και την εμπειρία για την εκτίμηση του κίνδυνου αθέτησης υποχρεώσεων. Όμως, η οικονομική πίεση που δημιουργείται από τις αυξανόμενες αιτήσεις για χορήγηση πίστωσης, τον ανταγωνισμό μεταξύ των πιστωτικών ιδρυμάτων και από την ανάπτυξη των υπολογιστικών συστημάτων, έχουν οδηγήσει στην ανάπτυξη των περίπλοκων στατιστικών μοντέλων που έχουν ως στόχο να υποστηρίξουν τη λήψη αποφάσεων για χορήγηση πίστωσης. Τα αποτελέσματα που προκύπτουν από τα στατιστικά μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας ονομάζονται αλλιώς και **σκορόχαρτα (*scorecards*)**. Οι πιο συνηθισμένες στατιστικές μέθοδοι που χρησιμοποιούνται την ανάπτυξη αυτών των μοντέλων είναι η διαχωριστική ανάλυση, η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση και τα δέντρα αποφάσεων. Σε αυτά τα μοντέλα, οι επεξηγηματικές μεταβλητές ονομάζονται **χαρακτηριστικά** και οι τιμές που αυτές μπορούν να πάρουν ονομάζονται **ιδιότητες** (*attributes*).

1.3 Πεδία εφαρμογής και πλεονεκτήματα των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας

Απαραίτητη προϋπόθεση για την εκτίμηση του πιστωτικού κινδύνου κάθε υποψηφίου πελάτη είναι αυτή να καθίσταται αμερόληπτη και δίκαιη χωρίς να εμπλέκονται υποκειμενικές απόψεις, ενώ η διαδικασία λήψεως πιστοδοτικών αποφάσεων που προκύπτει να εξασφαλίζει ότι με τις ίδιες πληροφορίες θα παίρνουμε πάντοτε την ίδια απόφαση. Οι απαιτήσεις αυτές οδηγούν στην αναγκαιότητα αυτοματοποίησης των διαδικασιών αξιολόγησης των πελατών, δηλαδή τη χρήση μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας τα οποία βασίζονται σε μηχανογραφημένες εφαρμογές, ικανές να επεξεργαστούν μεγάλο όγκο δεδομένων.

Σημαντικό όφελος από τη χρήση των CSM είναι επίσης και η αύξηση του διαχειριστικού ελέγχου που οφείλεται στη στατιστική εκτίμηση όχι μόνο του πιστωτικού κινδύνου αλλά και του όγκου των παρεχομένων πιστώσεων, γεγονός που εξασφαλίζει ότι η εφαρμογή της διαχειριστικής πολιτικής δεν στηρίζεται σε υποκειμενικές ερμηνείες της. Επιπλέον, η εφαρμογή CSM έχει ως αποτέλεσμα τη μείωση των επισφαλειών με ταυτόχρονη διατήρηση ή ακόμα και αύξηση του όγκου των παρεχομένων πιστώσεων.

Γενικά, τα CSM είναι ένα συνεπές, ακριβές και ισχυρό εργαλείο αξιολόγησης της πιστοληπτικής ικανότητας. Ένα καλά σχεδιασμένο μοντέλο μπορεί να υπολογίσει το επίπεδο του κινδύνου, να μειώσει την υποκειμενικότητα των πιστοδοτικών αποφάσεων και γενικότερα να βοηθήσει έτσι τους πιστωτές να βελτιώσουν την αποδοτικότητα των δραστηριοτήτων τους. Λόγω των ειδικών απαιτήσεων των σύγχρονων, πολύπλοκων επιχειρηματικών λειτουργιών, τα CSM, εκτός από τον έλεγχο των συμπληρωμένων αιτήσεων για την αποδοχή ή απόρριψη των υποψηφίων πελατών, χρησιμοποιούνται και για την οργάνωση δραστηριοτήτων όπως:

- η επιλογή των «σωστών» (δηλαδή χαμηλού κινδύνου) πελατών για να προσφερθούν σε αυτούς πρόσθετα προϊόντα και βελτιωμένες υπηρεσίες.
- η επινόηση στρατηγικών για να προσδιοριστούν οι πελάτες που επιδεικνύουν αρνητική συμπεριφορά (μη πληρωμή, απάτη) αλλά και για να τους εξετάζουν αποτελεσματικά έτσι ώστε να ελαχιστοποιήσουν τις περαιτέρω απώλειες χρημάτων προσδιορίζοντας τις απαιτούμενες δραστηριότητες με σκοπό την είσπραξη των οφειλομένων ποσών όσο το δυνατόν γρηγορότερα. Τέτοιες ενέργειες μπορεί να είναι τηλεφωνικές ενημερώσεις πελατών, επικοινωνία με εγγυητές, προειδοποιητικές επιστολές διαφόρων βαθμών αυστηρότητας ή ακόμα και προσφυγή στη δικαιοσύνη.
- η παρακολούθηση της εξέλιξης της συμπεριφοράς των πελατών για να διαπιστωθεί η αξιοπιστία και η εγκυρότητα των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας και να αναπτυχθούν στρατηγικές διαφοροποιημένης αντιμετώπισης διαφορετικών ομάδων.

Αν και το μεγαλύτερο μέρος των CSM που αναπτύσσονται εστιάζουν στην εκτίμηση του κινδύνου αθέτησης υποχρεώσεων, ο κίνδυνος αυτός είναι μόνο ένα μέρος της απόφασης για την χορήγηση πίστωσης. Ο κύριος στόχος είναι να μεγιστοποιηθούν τα κέρδη και η αποδοτικότητα της ενδιαφερόμενης επιχείρησης στοιχεία τα οποία δεν είναι απαραίτητο να σχετίζονται μονότονα με τον κίνδυνο. Για παράδειγμα, οι πολύ χαμηλού κινδύνου πελάτες που αποπληρώνουν το λογαριασμό της πιστωτικής τους κάρτας κάθε μήνα και ο δανειστής δεν μπορεί να τους χρεώσει τόκο, δεν είναι κερδοφόροι. Αντιθέτως, οι πολύ υψηλού κινδύνου υποψήφιοι μπορούν να είναι κερδοφόροι, υπό τον όρο ότι χρεώνεται ένα αρκετά υψηλό επιτόκιο.

Η γνώση της πιθανότητας αθέτησης υποχρεώσεων διευκολύνει σημαντικά στην υιοθέτηση μιας ορθολογικής διαδικασίας χορήγησης πιστώσεων κάθε είδους. Επομένως, τα μοντέλα

αυτά μπορούν να χρησιμοποιηθούν σε όλες σχεδόν τις δραστηριότητες που σχετίζονται με την παροχή καταναλωτικής ή επιχειρηματικής πίστης. Ενδεικτικά αναφέρουμε τους επόμενους κλάδους στους οποίους τα CSM αποτελούν απαραίτητα εργαλεία:

- Τράπεζες και χρηματοπιστωτικοί οργανισμοί για την αξιολόγηση της πιστοληπτικής ικανότητας των πελατών τους για την καταναλωτική πίστη (πιστωτικές κάρτες, δάνεια αυτοκινήτων), για τα δάνεια προς μικρές επιχειρήσεις, για ανεξάρτητες εκτιμήσεις των εγγυήσεων στη στεγαστική πίστη και για πλήρη χρηματοοικονομική ανάλυση στα δάνεια προς μικρομεσαίες και μεγάλες επιχειρήσεις. Επίσης, μπορούν να χρησιμοποιηθούν για την αξιολόγηση του κινδύνου ένας καταναλωτής να μην χρησιμοποιήσει ένα πιστωτικό προϊόν, ή να μεταφέρει τον λογαριασμό του σε έναν άλλο δανειστή.
- Εταιρείες πωλήσεων μέσω ταχυδρομείου. Χρησιμοποιούνται με σκοπό να προβλεφθεί ο κίνδυνος ένας καταναλωτής να μην ανταποκριθεί θετικά σε μια ταχυδρομική ενημέρωση για ένα νέο προϊόν.
- Σε επιχειρήσεις λιανικών πωλήσεων για να εκτιμηθεί η πιστοληπτική ικανότητα των πελατών τους.
- Σε ασφαλιστικές εταιρείες που ενδιαφέρονται, χρησιμοποιώντας τα στοιχεία που υπάρχουν, να αποφασίσουν αν θα ασφαλίσουν έναν κίνδυνο ή όχι και με ποιους όρους.

Τα CSM χρησιμοποιούνται σχεδόν σε όλες τις μορφές καταναλωτικής πίστης, όπως είναι οι πιστωτικές κάρτες, τα προσωπικά δάνεια και η χρηματοδότηση αυτοκινήτων. Την τελευταία δεκαετία, αυξάνονται συνεχώς οι περιπτώσεις χρησιμοποίησης μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας για στεγαστικά δάνεια και για χορήγηση πίστωσης σε μικρομεσαίες επιχειρήσεις (*SMEs*). Επιπλέον, σε μερικές χώρες τα CSM χρησιμοποιούνται για να εκτιμήσουν τον κίνδυνο χαρτοφυλακίου του καταναλωτικού χρέους που ένας οικονομικός οργανισμός μπορεί να θελήσει να εκποιήσει σε άλλη.

Οι μέθοδοι που χρησιμοποιούνται για να προβλεφθεί η πιθανότητα αθέτησης υποχρεώσεων για ιδιώτες και καταναλωτική πίστη είναι διαφορετικές με αυτές που χρησιμοποιούνται για την χορήγηση πίστωσης σε μεγάλες επιχειρήσεις. Αυτό συμβαίνει γιατί τα CSM δεν είναι τόσο ακριβή στην περίπτωση της δανειοδότησης των μεγάλων επιχειρήσεων όσο στην καταναλωτική πίστη αφού στην πρώτη περίπτωση χρησιμοποιούνται εντελώς διαφορετικά χαρακτηριστικά (ερμηνευτικές μεταβλητές) στα στατιστικά μοντέλα και στη διαδικασία λήψης αποφάσεων. Τέτοια χαρακτηριστικά είναι κάποιοι χρηματοοικονομικοί δείκτες που προκύπτουν από τις λογιστικές καταστάσεις των επιχειρήσεων και είναι αρκετά

δύσκολο να αναλυθούν και να επικυρωθούν. Επιπλέον, τα χρηματικά ποσά που δίνονται για τη χρηματοδότηση των επιχειρήσεων είναι πολύ μεγαλύτερα σε σχέση με αυτά που δίνονται για καταναλωτικά δάνεια, επομένως αυτά τα ποσά ασκούν πολύ μεγαλύτερη επίδραση στην οικονομική κατάσταση του πιστωτικού ιδρύματος.

Οι μέθοδοι που χρησιμοποιούνται για τη χρηματοδότηση μεγάλων επιχειρήσεων ονομάζονται συστήματα πιστοληπτικής διαβάθμισης (*credit rating systems - CRS*) και σύμφωνα με αυτά, οι επιχειρήσεις κατατάσσονται ως προς το επίπεδο κινδύνου τους. Υπάρχουν διάφορα πρακτορεία πιστοληπτικής διαβάθμισης που προσφέρουν τις υπηρεσίες τους για την εκτίμηση της πιστοληπτικής ικανότητας επιχειρήσεων. Η διαβάθμιση υποδεικνύεται με κατηγορίες που συμβολίζονται από κάποια αλφαβητική κλίμακα και υποδηλώνουν διάφορα μεγέθη πιθανού κινδύνου: από ικανοποιητική πιστοληπτική συμπεριφορά ως πιθανόν προβληματικές χορηγήσεις, (οι οποίες τίθενται υπό παρακολούθηση), επισφαλείς απαιτήσεις και τέλος, χορηγήσεις με μηδενική πιθανότητα αποπληρωμής. Τα πρακτορεία πιστοληπτικής διαβάθμισης, όπως είναι το Standard & Poors (S&P) κατηγοριοποιούν τους δανειολήπτες σε τουλάχιστον επτά κατηγορίες με τις πρώτες τέσσερις κατηγορίες, AAA, AA, A και BBB να υποδεικνύουν τους πιο αξιόπιστους. Στις εθνικές τράπεζες απαγορεύεται η αγορά χρεογράφων, τα οποία δεν εκδίδονται από φορείς κάποιων από τις προηγούμενες κατηγορίες. Αντίθετα, οι κανονισμοί που ισχύουν στις ασφαλιστικές εταιρείες επιτρέπουν την αγορά χρεογράφων που ανήκουν στις κατηγορίες BB, B και CCC αλλά υπάρχει περιορισμός ως προς το ποσό που επιτρέπεται να εντάξουν στο χαρτοφυλάκιό τους.

Για τις μεγάλες επιχειρήσεις, τα CSM χρησιμοποιούνται συνήθως για να προβλέψουν τον κίνδυνο πτώχευσης της επιχείρησης. Η ανάπτυξη ενός καλά δομημένου συστήματος βαθμολόγησης πιστοληπτικής ικανότητας είναι μια δύσκολη και δαπανηρή διαδικασία, που απαιτεί από την ενδιαφερόμενη επιχείρηση σημαντική επένδυση χρόνου και εξειδικευμένης εργασίας. Η διαδικασία είναι εκ των προτέρων σύνθετη και αποτελείται από πολλά στάδια, με πιο σημαντικό αυτό που σχετίζεται με το σαφή καθορισμό των επιχειρηματικών στόχων και τη μελέτη της διάρθρωσης του περιβάλλοντος εντός του οποίου θα λειτουργήσει το σύστημα βαθμολόγησης. Τα υπόλοιπα στάδια αυτής της διαδικασίας σχετίζονται με την εξέταση της διαθεσιμότητας των απαιτούμενων δεδομένων, τη δραστηριοποίηση για την απόκτησή τους, την υλοποίηση του συστήματος βαθμολόγησης πιστοληπτικής ικανότητας καθώς και του ελέγχου της αξιοπιστίας του.

1.4 Ιστορική αναδρομή

Εδώ και 4000 χρόνια περίπου οι άνθρωποι δείχνουν εμπιστοσύνη ο ένας τον άλλον δανείζοντας χρήματα ή άλλα υλικά αγαθά. Ο Lewis (1992) αναφέρει ότι η πρώτη καταγραμμένη περίπτωση παροχής πίστωσης προέρχεται από μια πήλινη πινακίδα που βρέθηκε στην περιοχή της αρχαίας Βαβυλώνας και χρονολογείται περίπου στο 2000 π.Χ. Στην πινακίδα καταγράφεται η συναλλαγή δύο αγροτών οι οποίοι είχαν δανειστεί χρήματα και θα τα επέστρεφαν την εποχή της συγκομιδής, με τα εξής λόγια:
«ο Sin-Kalama-idi, γιος του Ulamasha και ο April-ilu-shu, γιος του Khayamdidou, δανείστηκαν από τον Arad-Sin έξι αργυρά νομίσματα για την αποθήκευση της σοδειάς τους. Στην γιορτή του Ab θα πληρώσουν το σιτάρι».
Συμπεραίνουμε λοιπόν ότι οι αγρότες εκείνης της εποχής αντιμετώπιζαν τα προβλήματα ρευστότητας λαμβάνοντας δάνειο κατά τη διάρκεια της σποράς και αποπληρώνανε τα χρέη τους μετά τη συγκομιδή.

Λίγο αργότερα, κατά τη διάρκεια της αυτοκρατορίας των Ασσυρίων, υπάρχουν κείμενα που καταγράφουν τις συναλλαγές μεταξύ πόλεων που βρίσκονταν στις όχθες του Τίγρη και κάποιων ανατολικών περιοχών της Τουρκίας. Όταν οι πωλητές έφθαναν στον προορισμό τους συνήθιζαν να χρεώνουν 33% τόκο στα αγαθά που πουλούσαν θέλοντας να καλύψουν τους κινδύνους που ενδεχομένως θα αντιμετώπιζαν κατά τη μεταφορά των αγαθών.

Μέχρι την εποχή των ελληνικών και ρωμαϊκών αυτοκρατοριών, τα όργανα τραπεζικών εργασιών και πίστωσης ήταν αρκετά προηγμένα. Την επόμενη χιλιετία, δηλαδή κατά τη διάρκεια του Μεσαίωνα της ευρωπαϊκής ιστορίας, υπήρξε μικρή ανάπτυξη στην παροχή πίστωσης, ωστόσο την εποχή των σταυροφοριών ($13^{ος}$ αιώνας), τα καταστήματα ενέχυρων είχαν αναπτυχθεί αρκετά. Αρχικά, αυτά τα καταστήματα είχαν ιδρυθεί για φιλανθρωπικούς σκοπούς και δεν χρέωναν κανέναν τόκο, αλλά όταν το 1350 μ.Χ. οι έμποροι αντιλήφθηκαν ότι τα κέρδη που θα μπορούσαν να αντλήσουν ήταν μεγάλα, τα εμπορικά καταστήματα ενέχυρων άρχισαν να χρεώνουν τόκο και το ενδιαφέρον αυτό επεκτάθηκε σε όλη την Ευρώπη. Κατά τη διάρκεια των μέσων χρόνων ($12^{ος}$ - $15^{ος}$ αιώνας π.Χ.) υπήρξε μια φιλοσοφική διαμάχη με το αν είναι ηθικά σωστή η καταβολή τόκου για ένα δάνειο. Αυτή η αντιπαράθεση συνεχίζεται μέχρι σήμερα στις ισλαμικές χώρες. Ο Αριστοτέλης στα Πολιτικά αναφέρει:

«…μισείται η τοκογλυφία, γιατί το κέρδος προέρχεται από αυτό το ίδιο το νόμισμα και όχι από την χρήση χάριν της οποίας εφευρέθηκε το νόμισμα, δηλαδή την ανταλλαγή».

Η Ίδρυση της Τράπεζας της Αγγλίας το 1694 ήταν ένα από τα πρώτα σημάδια της οικονομικής επανάστασης που επέτρεψε το μαζικό δανεισμό. Για τα επόμενα 150 χρόνια οι τράπεζες δάνειζαν στην αριστοκρατία και στους μεγαλοαστούς. Η αύξηση των μεσαίων τάξεων το 1800μ.Χ. οδήγησε στο σχηματισμό διάφορων ιδιωτικών τραπεζών, οι οποίες ήταν πρόθυμες να δώσουν μεγαλύτερα πιστωτικά όρια για τη χρηματοδότηση επιχειρήσεων και για έξοδα διαβίωσης. Εντούτοις, η έναρξη της καταναλωτικής πίστης περιορίστηκε την εποχή εκείνη σε ένα πολύ μικρό ποσοστό του πληθυσμού. Αργότερα, ο δανεισμός άρχισε σιγά - σιγά να προσφέρεται και από τους βιομήχανους και όχι μόνο από τις τράπεζες. Για παράδειγμα, το 1850 περίπου η εταιρία ραπτομηχανών Singer άρχισε να πουλάει στους πελάτες τις μηχανές με δόσεις.

Η πραγματική επανάσταση στη χορήγηση πίστωσης άρχισε το 1920, όταν οι καταναλωτές ξεκίνησαν να αγοράζουν αυτοκίνητα. Όταν ο Henry Ford και ο A. P. Sloan αντιλήφθηκαν ότι για να πουλήσουν αυτοκίνητα έπρεπε κάποιος να βρει τρόπο να χρηματοδοτήσει τις αγορές των καταναλωτών και έτσι προχώρησαν στη ίδρυση των πρώτων «πιστωτικών οίκων». Μετά το δεύτερο παγκόσμιο πόλεμο, άρχισαν να αυξάνονται οι επιχειρήσεις ταχυδρομικών παραγγελιών δεδομένου ότι οι καταναλωτές στις μικρότερες πόλεις ήθελαν να αποκτήσουν ενδύματα και οικιακά προϊόντα που ήταν διαθέσιμα μόνο στα μεγάλα αστικά κέντρα. Αυτά τα προϊόντα διαφημίζονταν σε καταλόγους, και οι επιχειρήσεις ήταν πρόθυμες να στείλουν τα αγαθά με πίστωση και να επιτρέψουν στους πελάτες να τα πληρώσουν μετά από μια συγκεκριμένη χρονική περίοδο.

Κατά τη διάρκεια των τελευταίων 50 ετών, η χορήγηση πίστωσης σε καταναλωτές και επιχειρήσεις αυξάνεται με ιλιγγιώδη ταχύτητα. Η εμφάνιση των πιστωτικών καρτών στη δεκαετία του '60 ήταν ένας παράγοντας που συνέβαλε αρκετά σε αυτήν την αύξηση. Η πιστωτική κάρτα είναι πλέον ένα αναπόσπαστο κομμάτι των αγορών μας. Σε πολλές περιπτώσεις μάλιστα οι αγορές μπορούν να γίνουν μόνο εάν κάποιος χρησιμοποιήσει πιστωτική κάρτα, όπως για παράδειγμα στις αγορές μέσω του Διαδικτύου.

Ενώ η ιστορία της χορήγησης πιστώσεως ξεκινάει περίπου πριν 6000 έτη, η ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας ξεκίνησε μόλις πριν από 60 έτη. Τα μοντέλα βαθμολόγησης είναι ουσιαστικά ένας τρόπος να προσδιοριστούν οι διαφορετικές ομάδες σε έναν πληθυσμό. Μια πρώτη προσέγγιση στην επίλυση αυτού του προβλήματος του

προσδιορισμού διαφορετικών ομάδων σε έναν πληθυσμό στη στατιστική έγινε από τον Fisher (1936). Ο Fisher επιδίωξε να ταξινομήσει δύο ποικιλίες φυτών χρησιμοποιώντας μετρήσεις κάποιων βασικών χαρακτηριστικών τους. Ο Durand (1941) ήταν ο πρώτος που αναγνώρισε ότι αυτές οι τεχνικές θα μπορούσαν να χρησιμοποιηθούν για να γίνει διάκριση μεταξύ των «καλών» και των «κακών» δανείων.

Κατά τη διάρκεια της δεκαετίας του '30, μερικές επιχειρήσεις ταχυδρομικών παραγγελιών είχαν εισάγει κάποια αριθμητικά συστήματα υπολογισμού βαθμολογιών με σκοπό να προσπαθήσουν να αντιμετωπίσουν τις ασυνέπειες στις πιστωτικές αποφάσεις των πιστωτικών αναλυτών. Με την έναρξη του δεύτερου παγκόσμιου πολέμου, όλοι οι οργανισμοί χρηματοδότησης και οι εταιρίες ταχυδρομικών παραγγελιών άρχισαν να αντιμετωπίζουν δυσκολίες με την πιστωτική διαχείριση. Οι πιστωτικοί αναλυτές καλούνταν να εκπληρώσουν τις στρατιωτικές τους υποχρεώσεις, και υπήρξε μια σοβαρή έλλειψη ανθρώπων που είχαν αυτήν την εμπειρία.

Αργότερα, οι εταιρίες ζήτησαν στους αναλυτές να καταγράψουν τις εμπειροτεχνικές μεθόδους που χρησιμοποίησαν για να αποφασίσουν σε ποιους θα δώσουν τα δάνεια. Κάποιες από αυτές τις μεθόδους ήταν αριθμητικά συστήματα υπολογισμού βαθμολογιών που είχαν ήδη εισαχτεί και κάποιες άλλες βασίζονταν σε όρους και προϋποθέσεις που έπρεπε να ικανοποιούνται από τον κάθε πελάτη για τη χορήγηση πίστωσης. Αυτοί οι κανόνες χρησιμοποιήθηκαν αργότερα για να λαμβάνονται αποφάσεις και αποτέλεσε ένα από τα πρώτα παραδείγματα των έμπειρων συστημάτων.

Λίγο καιρό μετά από τον δεύτερο παγκόσμιο πόλεμο, η αυτοματοποίηση των πιστωτικών αποφάσεων συνδέθηκε με τις τεχνικές ταξινόμησης που είχαν αναπτυχθεί στη στατιστική και έτσι έγινε ορατό το όφελος των στατιστικών μοντέλων για τη λήψη αποφάσεων για δανεισμό. Σύμφωνα με τον Wonderlic (1952) η πρώτη εταιρεία παροχής συμβουλών για χορήγηση πιστώσεων ιδρύθηκε στο Σαν Φρανσίσκο από τον Bill Fair και Earl Isaac στις αρχές του 1950, και οι πελάτες της ήταν κυρίως οργανισμοί χρηματοδότησης, εταιρείες λιανικών πωλήσεων και εταιρίες ταχυδρομικών παραγγελιών.

Η άφιξη των πιστωτικών καρτών προς το τέλος της δεκαετίας του '60, έκανε τις τράπεζες και άλλους εκδότες πιστωτικών καρτών να διαπιστώσουν τη χρησιμότητα των στατιστικών μοντέλων βαθμολόγησης της πιστοληπτικής ικανότητας. Η αύξηση του αριθμού ανθρώπων που υπέβαλλαν αίτηση για πιστωτικές κάρτες, κατέστησε αδύνατη την επεξεργασία μίας-μίας αίτησης τόσο από οικονομικής πλευράς όσο και ανθρωπίνου δυναμικού και έτσι

δημιουργήθηκε η ανάγκη αυτοματοποίησης των αποφάσεων δανεισμού. Η αύξηση της υπολογιστικής δύναμης κατέστησε δυνατή την αυτοματοποίηση των αποφάσεων δανεισμού. Οι ενδιαφερόμενες επιχειρήσεις βρήκαν τα στατιστικά μοντέλα πιστοληπτικής ικανότητας ως καλύτερο κριτήριο απόφασης από οποιοδήποτε άλλη υποκειμενική μέθοδο με αποτέλεσμα οι περιπτώσεις πελατών που αθετούσαν τις υποχρεώσεις τους να ελαττωθούν κατά 50% ή και περισσότερο.[1]

Υπήρχαν όμως κάποιες αντιθέσεις που αφορούσαν την εφαρμογή των CSM. Για παράδειγμα ο Capon (1982) υποστήριζε ότι «*η ωμή βία της εμπειριοκρατίας της βαθμολόγησης της πιστοληπτικής ικανότητας προσβάλλει τις παραδόσεις της κοινωνίας μας*». Ο Capon θεώρησε ότι πρέπει να υπάρξει περισσότερη εξάρτηση από την πιστωτική ιστορία δηλαδή από τα ιστορικά αρχεία του δανειολήπτη και θα πρέπει να είναι δυνατό να εξηγηθεί γιατί ορισμένα χαρακτηριστικά απαιτούνται σε ένα CSM και άλλα όχι. Επιπλέον, θεωρούσε ότι δεν είναι δίκαιο να παίρνονται αποφάσεις μόνο βάσει κάποιων χαρακτηριστικών για τους δανειολήπτες που δεν έχουν πιστωτικό παρελθόν. Το γεγονός που εξασφάλισε την πλήρη αποδοχή των στατιστικών μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας ήταν η εισαγωγή του νόμου των Ίσων Πιστωτικών Ευκαιριών (Equal Credit Opportunity Acts - ECOA) και των τροποποιήσεών του στις ΗΠΑ το 1975 και το 1976. Αυτή η νομοθεσία εισήχθηκε από το Κογκρέσο και απαγόρευε κάθε είδους διάκριση (λόγω φύλου, οικογενειακής κατάστασης, φυλής, θρησκείας, ηλικίας, καταγωγής, παραλαβή εισοδήματος από κάποιο δημόσιο πρόγραμμα βοήθειας) στη χορήγηση της πίστωσης εκτός αν η διάκριση αυτή έχει παραχθεί εμπειρικά και είναι στατιστικά έγκυρη.

Παράλληλα με την ανάπτυξη των CSM ο Altman (1968) ανέπτυξε και μια μέθοδο υπολογισμού βαθμολογιών με σκοπό να προβλέψουν τον κίνδυνο να χρεοκοπήσει μια επιχείρηση. Αυτή η μέθοδος παρουσιάζει ενδιαφέρον γιατί συνδέονται οι χρηματοοικονομικοί δείκτες της επιχείρησης με την επικείμενη πτώχευση. Παρ' όλα αυτά επειδή τα δείγματα αυτά είναι πολύ μικρότερα σε σχέση με τα δείγματα στην καταναλωτική πίστη, οι προβλέψεις είναι λιγότερο ακριβείς.

Στη δεκαετία του '80, η επιτυχία των στατιστικών μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας στις πιστωτικές κάρτες ώθησε τις τράπεζες να αρχίσουν να χρησιμοποιούν τα μοντέλα αυτά και για άλλα προϊόντα, όπως τα προσωπικά δάνεια, ενώ στα τελευταία έτη τα CSM έχουν χρησιμοποιηθεί για τα εγχώρια δάνεια και τα δάνεια μικρών επιχειρήσεων. Στη

[1] Για περισσότερες πληροφορίες βλέπε Myers and Forgy (1963).

δεκαετία του '90, η αύξηση του άμεσου μάρκετινγκ οδήγησε στη χρήση των σκορόχαρτων για να βελτιωθεί το ποσοστό απάντησης στις εκστρατείες διαφήμισης (στην πραγματικότητα αυτή ήταν μια από τις πιο πρόωρες χρήσεις των CSM στη δεκαετία του '50). Η πρόοδος στις δυνατότητες υπολογισμού επέτρεψε να δοκιμαστούν και άλλες τεχνικές για τη δόμηση των CSM. Στη δεκαετία του '80, η λογιστική παλινδρόμηση και ο γραμμικός προγραμματισμός ήταν τα δύο πιο σημαντικά εργαλεία για την ανάπτυξη των CSM. Πιο πρόσφατα, οι ειδικές τεχνικές της περιοχής της τεχνητής νοημοσύνης, όπως τα έμπειρα συστήματα (*expert systems*) και τα νευρωνικά δίκτυα (*neural networks*) έχουν χρησιμοποιηθεί για την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας.

Στη σύγχρονη εποχή υπάρχει μεγάλη ποικιλία πιστωτικών προϊόντων με διαφορετικά χαρακτηριστικά το καθένα. Αυτά τα προϊόντα μπορεί να είναι πιστωτικές κάρτες, καταναλωτικά ή στεγαστικά δάνεια, δάνεια αυτοκινήτου ή δάνεια επιχειρήσεων. Επομένως, ο ανταγωνισμός μεταξύ των πιστωτικών ιδρυμάτων ολοένα και αυξάνεται. Οι καταναλωτές πλέον ψάχνουν για προϊόντα που ταιριάζουν περισσότερο στις ανάγκες τους και αυτά που είναι πιο συμφέροντα γι’ αυτόν όπως για παράδειγμα, συνήθως επιλέγουν προϊόντα με το μικρότερο επιτόκιο. Οι καταναλωτές δεν επιδιώκουν μόνο περισσότερη πίστωση αλλά ελέγχουν εάν είναι καλύτερο να παραμείνουν με τον υπάρχοντα δανειστή τους ή να μετακινηθούν προς έναν άλλον που προσφέρει ένα ελκυστικότερο προϊόν. Κατά συνέπεια, σήμερα η τιμολόγηση και η προσαρμογή των προϊόντων δανεισμού γίνονται σημαντικότερες.

Στα σύγχρονα χρηματοπιστωτικά ιδρύματα δίνεται ιδιαίτερη προσοχή στον τρόπο με τον οποίο πρέπει να διαμορφωθούν οι επιχειρησιακοί στόχοι έτσι ώστε να ελαχιστοποιηθεί η πιθανότητα ένας πελάτης να μην καταφέρει να εκπληρώσει τις υποχρεώσεις του σε ένα συγκεκριμένο πιστωτικό προϊόν και στο πώς η ενδιαφερόμενη επιχείρηση μπορεί να μεγιστοποιήσει το κέρδος που θα της αποφέρει κάθε πελάτης. Επιπλέον, η ιδέα της εκτίμησης του κινδύνου μη αποπληρωμής των χρεών έχει επεκταθεί με τη χρήση μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας για να δοθούν απαντήσεις σε διάφορα ερωτήματα όπως:

- ποια είναι η πιθανότητα ο καταναλωτής να ανταποκριθεί σε μια άμεση αποστολή ενός νέου προϊόντος;
- ποια είναι η πιθανότητα ο καταναλωτής να χρησιμοποιήσει ένα προϊόν;
- ποια είναι η πιθανότητα ο καταναλωτής να συνεχίσει να χρησιμοποιεί το προϊόν όταν εισαγωγική περίοδο προσφοράς τελειώσει;

- ποια είναι η πιθανότητα ο πελάτης να αλλάξει δανειστή;
- εάν ο καταναλωτής αρχίζει να καθυστερεί δόσεις στο δάνειο, ποιες θα είναι οι κατάλληλες προσεγγίσεις για να αποτραπεί η αθέτηση των υποχρεώσεών του;
- πόσο πιθανό είναι η εφαρμογή του στατιστικού μοντέλου πιστοληπτικής ικανότητας να μην είναι έγκυρη;

Παραδοσιακά η μοντελοποίηση της πίστωσης κάθε είδους λάμβανε υπ' όψιν κάθε δάνειο ξεχωριστά. Σήμερα πλέον οι δανειστές ενδιαφέρονται για τα χαρακτηριστικά του χαρτοφυλακίου δανείων για καταναλωτές (λιανικό χαρτοφυλάκιο) ή για επιχειρήσεις. Η σημασία του πιστωτικού κινδύνου ενός χαρτοφυλακίου δανείων (το ποσοστό του κεφαλαίου που δανείζεται και εκτιμάται ότι θα χαθεί λόγω της αθέτησης υποχρεώσεων των πελατών) έχει αυξηθεί λόγω των αλλαγών που έχουν λάβει χώρα στους τραπεζικούς κανονισμούς που ενσωματώνονται στη νέο κανονιστικό πλαίσιο της Βασιλείας. Αυτή η συμφωνία εφαρμόστηκε το 2007 και καλείται **Βασιλεία ΙΙ (Basel II)** και είναι η δεύτερη τέτοια κύρια συμφωνία. Το νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ επιτρέπει στους χρηματοπιστωτικούς οργανισμούς να χρησιμοποιούν τα **μοντέλα εσωτερικής διαβάθμισης (*IRB – Internal Rating – Based models***) τα οποία βοηθούν τις τράπεζες να καθορίσουν πόσο κεφάλαιο πρέπει να διαθέτουν έτσι ώστε να μπορούν να αντιμετωπίσουν πιθανές απώλειες από το χαρτοφυλάκιο των δανείων τους. Για τα λιανικά χαρτοφυλάκια (*retail portfolio*), τα CSM ανήκουν στην κατηγορία των μοντέλων εσωτερικής διαβάθμισης. Εντούτοις, η χρησιμοποίηση των μοντέλων γι' αυτό το σκοπό απαιτεί κάποιες αλλαγές στις απαιτήσεις στα συστήματα βαθμολόγησης. Για τα CSM δεν είναι αρκετό πλέον να ταξινομήσουν σωστά τους πελάτες, αλλά τώρα, δεδομένου ότι αυτά χρησιμοποιούνται για την εκτίμηση της πιθανότητας αθέτησης των υποχρεώσεων, με σκοπό να προσδιοριστεί το κεφάλαιο που πρέπει να τεθεί στην άκρη, οι προβλέψεις πρέπει να ικανοποιούν κάποια ελάχιστα κριτήρια ορθής πρόβλεψης που τέθηκαν από το κανονιστικό πλαίσιο της Βασιλείας ΙΙ.

1.5 Περιεχόμενα διπλωματικής

Στην παρούσα εργασία γίνεται μια ανασκόπηση στη περιοχή των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας που χρησιμοποιούνται σε μεγάλο βαθμό από τα χρηματοπιστωτικά

ιδρύματα. Κύριος σκοπός είναι να περιγραφούν οι πιο βασικές στατιστικές μέθοδοι ανάπτυξης των μοντέλων αυτών και τα στάδια των διαδικασιών που απαιτούνται πριν και μετά την εφαρμογή τους.

Συγκεκριμένα σε αυτό το κεφάλαιο έγινε αναφορά στην έννοια του πιστωτικού κινδύνου, περιγράφηκε η γενική φιλοσοφία των CSM και διευκρινίστηκε σε ποιους τομείς αυτά χρησιμοποιούνται. Επιπλέον, έγινε μια ιστορική αναδρομή της χορήγησης πίστωσης και περιγράφηκε η εξέλιξη των CSM στην πάροδο του χρόνου.

Στο δεύτερο κεφάλαιο αναφέρονται τα βασικότερα είδη CSM τα οποία διαχωρίζονται ανάλογα με τον σκοπό για τον οποίο χρησιμοποιούνται και ανάλογα με το είδος των δεδομένων με βάση τα οποία αναπτύσσονται. Επίσης, περιγράφονται τα αρχικά στάδια της προετοιμασίας που απαιτούνται πριν την ανάπτυξη και εφαρμογή των CSM και επεξηγείται με ποιον τρόπο γίνεται ο καθορισμός του «καλού» και του «κακού» πελάτη. Τα στάδια αυτά περιλαμβάνουν την επιλογή δείγματος, την κατάτμηση των δεδομένων και τον καθορισμό των χαρακτηριστικών που χρησιμοποιούνται σε κάθε τύπο πίστωσης. Τέλος, προτείνονται τρόποι αντιμετώπισης όταν εμφανίζονται στα δεδομένα ελλείπουσες τιμές ή έκτροπες παρατηρήσεις.

Στο τρίτο κεφάλαιο περιγράφονται οι διαδικασίες που ακολουθούνται έτσι ώστε να ανακτηθεί η πληροφορία που χάνεται από τους απορριφθέντες πελάτες και αναφέρεται με ποιον τρόπο γίνεται η επιλογή των σημαντικότερων χαρακτηριστικών και πως αυτά αρχικώς ταξινομούνται. Επιπλέον αναλύονται οι σημαντικότερες στατιστικές μέθοδοι που χρησιμοποιούνται για την ανάπτυξη των CSM. Αυτές οι μέθοδοι είναι η διαχωριστική ανάλυση, η γραμμική διαχωριστική ανάλυση, η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, τα δέντρα ταξινόμησης και η μέθοδος του κοντινότερου γείτονα. Προς το τέλος του κεφαλαίου δίνεται μια συνοπτική περιγραφή των μεθόδων ανάπτυξης των μοντέλων βαθμολόγησης συμπεριφοράς και κέρδους.

Αφού αναπτυχθούν τα CSM η αμέσως επόμενη διαδικασία είναι η επικύρωση των μοντέλων. Στο τέταρτο κεφάλαιο περιγράφονται οι πιο συνηθισμένοι τρόποι αξιολόγησης της απόδοσης των CSM χρησιμοποιώντας κάποιο δείγμα ελέγχου. Για τη μέτρηση της απόδοσης κάθε CSM μπορεί να υπολογιστοούν τα ποσοστά σωστής ταξινόμησης των πελατών, να χρησιμοποιηθούν κάποια μέτρα διαχωριστικής ικανότητας ή να μετρηθεί η ακρίβεια προσαρμογής της εκτίμησης της πιθανότητας αθέτησης υποχρεώσεων. Επιπλέον,

προτείνονται κάποιες τεχνικές αξιολόγησης των μοντέλων σε περιπτώσεις όπου το δείγμα είναι αρκετά μικρό.

Στο πέμπτο κεφάλαιο περιγράφονται τα βασικότερα πλεονεκτήματα και μειονεκτήματα για κάθε μία από τις στατιστικές μεθόδους ανάπτυξης των CSM. Η περιγραφή αυτή γίνεται με σκοπό να γίνουν κατανοητοί οι περιορισμοί και οι δυνατότητες που έχει κάθε μέθοδος. Στη συνέχεια γίνεται σύγκριση των στατιστικών αυτών μεθόδων αναφέροντας μελέτες που έχουν γίνει στο παρελθόν γι' αυτόν τον σκοπό.

Στο έκτο κεφάλαιο της παρούσας εργασίας αναπτύσσονται στατιστικά μοντέλα βαθμολόγησης με χρήση πραγματικών δεδομένων, αφού προηγηθεί μια διερευνητική ανάλυση των δεδομένων. Για την ανάπτυξη των μοντέλων αυτών χρησιμοποιούνται οι μέθοδοι της λογιστικής παλινδρόμησης, της διαχωριστικής ανάλυσης και των δέντρων ταξινόμησης. Στο τέλος, γίνεται σύγκριση της απόδοσης των μοντέλων που αναπτύχθηκαν χρησιμοποιώντας τα ποσοστά σωστής ταξινόμησης και καμπύλες ROC.

Στο έβδομο και τελευταίο κεφάλαιο γίνεται μια σύντομη αναφορά στο κανονιστικό πλαίσιο της Βασιλείας ΙΙ. Σε αυτό το σημείο, περιγράφονται οι έννοιες της πιθανότητας αθέτησης υποχρεώσεων, της ζημιάς δεδομένης της αθέτησης (LGD) καθώς και του χρηματοδοτικού ανοίγματος (EAD) των πιστωτικών ιδρυμάτων όπως αναφέρονται σε αυτό το κανονιστικό πλαίσιο. Απώτερος σκοπός της εκτίμησης των παραπάνω παραμέτρων είναι η εκτίμηση της Αναμενόμενης Ζημίας (EL) του πιστωτικού ιδρύματος εξαιτίας των εν λόγω ανοιγμάτων και ο υπολογισμός των συνολικών κεφαλαιακών απαιτήσεων που πρέπει να διακρατούνται ως αντιστάθμιση έναντι των κινδύνων του συγκεκριμένου χαρτοφυλακίου. Εν κατακλείδι, αναφέρεται ο τρόπος με τον οποίο γίνεται η προσαρμογή των CSM σύμφωνα με τους κανονισμούς της Βασιλείας ΙΙ.

ΚΕΦΑΛΑΙΟ 2

Είδη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας και στάδια προετοιμασίας

2.1 Είδη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας

Τα στατιστικά μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας μπορούν να χωριστούν σε τρείς μεγάλες κατηγορίες ανάλογα με το σκοπό για τον οποίο χρησιμοποιούνται. Οι κατηγορίες αυτές είναι:

α_1. τα **μοντέλα βαθμολόγησης αιτήσεων (*application scoring models* - ASM*)***,**

α_2. τα **μοντέλα βαθμολόγησης συμπεριφοράς (*behavioral scoring models*- BSM*)*** και

α_3. τα **μοντέλα βαθμολόγησης κέρδους (*profit scoring models* - PSM).**

Η μεθοδολογία που ακολουθείται και τα δεδομένα που αντλούνται διαφοροποιούνται ανάλογα με το είδος του CSM που θέλει να εφαρμόσει η ενδιαφερόμενη επιχείρηση. Παρακάτω δίνονται κάποια χαρακτηριστικά για κάθε είδος μοντέλου.

α_1. Μοντέλα βαθμολόγησης αιτήσεων

Τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας αρχικά διαμορφώθηκαν με σκοπό να βοηθούν στην απόφαση αν σε κάποιο νέο πελάτη θα πρέπει να χορηγηθεί δάνειο ή όχι. Τα μοντέλα βαθμολόγησης αιτήσεων (*application scoring models* ή ASM) όπως ονομάζονται τα στατιστικά μοντέλα που χρησιμοποιούνται για αυτό το σκοπό βοηθούν ώστε να παίρνονται σωστές αποφάσεις γρήγορα και αυτοματοποιημένα για μεγάλο αριθμό υποψήφιων πελατών.

Τα ASM σκοπό έχουν μόνο να προβλέψουν την πιθανότητα αθέτησης υποχρεώσεων μελλοντικών δανειοληπτών και όχι να εξηγήσουν τους λόγους της απόφασης. Επομένως, μπορεί να χρησιμοποιηθεί οποιοδήποτε θεμιτό μέσο προκειμένου να αυξηθεί η προβλεπτική

ικανότητα του στατιστικού μοντέλου. Η γενική ιδέα των ASM είναι να συγκριθούν τα στοιχεία των υποψήφιων δανειοληπτών με των ήδη υπαρχόντων πελατών έτσι ώστε να ταξινομηθούν οι τρέχοντες υποψήφιοι ως προς τον κίνδυνο αθέτησης υποχρεώσεων. Τα δεδομένα εκείνων που υπέβαλλαν αίτηση για δάνειο 1 ή 2 έτη πριν, μαζί με την πιστωτική ιστορία τους, χρησιμοποιούνται για να δομηθούν τα ASM τα οποία χρησιμοποιούνται στη συνέχεια για να καθορίσουν τα επόμενα 2 έτη και συνεπώς υποστηρίζουν τη διαδικασία λήψης απόφασης σε ποιους τελικά θα χορηγηθεί πίστωση και σε ποιους όχι. Επομένως, απαραίτητη προϋπόθεση για την εφαρμογή του μοντέλου είναι οι παράγοντες που επηρεάζουν την πιστοληπτική ικανότητα να είναι σχετικά σταθεροί για τουλάχιστον ένα χρονικό διάστημα μερικών ετών.

Τα ASM παραδοσιακά αξιολογούσαν έναν πολύ συγκεκριμένο κίνδυνο αθέτησης υποχρεώσεων. Ο πιο κοινός κίνδυνος που εξετάζεται είναι ο πελάτης να έχει 90 μέρες ληξιπρόθεσμες οφειλές για τους επόμενους 12 μήνες. Το τι θα συνέβαινε κατά τη διάρκεια άλλων χρονικών περιόδων και εάν ο πελάτης είναι κερδοφόρος για το δανειστή είναι πτυχές που δεν εξετάζονται σε αυτήν την αξιολόγηση. Η απόφαση πόσοι ή ποιο ποσοστό των υποψηφίων θα γίνει δεκτό είναι μια διευθυντική απόφαση από την ενδιαφερόμενη επιχείρηση που βασίζεται σε διάφορα επιχειρησιακά μεγέθη, όπως στα αναμενόμενα κέρδη, στις αναμενόμενες απώλειες, και στο μερίδιο αγοράς. Κατά συνέπεια η βαθμολογία αποδοχής-απόρριψης, το σημείο «πάνω» από το οποίο οι υποψήφιοι γίνονται αποδεκτοί, αποφασίζεται συνήθως χρησιμοποιώντας εμπειρικά στοιχεία όπως η αναλογία των «καλών» προς τους «κακούς» πελάτες σε αυτή τη βαθμολογία.

Τα δεδομένα που ήταν διαθέσιμα για την εφαρμογή των ASM βελτιώθηκαν αισθητά με την ανάπτυξη των γραφείων πίστης. Αυτές οι επιχειρήσεις συγκεντρώνουν στοιχεία που αφορούν τη φερεγγυότητα ενός καταναλωτή τα οποία συλλέγονται από διαφορετικούς δανειστές που είχαν κατά καιρούς συναλλαγές με τον πελάτη και συνήθως αποτελούνται από επίσημες πληροφορίες όπως στοιχεία από εκλογικούς καταλόγους και έγγραφα δικαστηρίων σχετικά με τα πρακτικά πτώχευσης. Το είδος των δεδομένων που έχει στη διάθεσή του κάθε τέτοιο γραφείο διαφέρει από χώρα σε χώρα. Για παράδειγμα, στις ΗΠΑ τα γραφεία πίστης έχουν στη διάθεσή τους όλα τα δεδομένα που αφορούν την πιστωτική συμπεριφορά κάθε καταναλωτή, ενώ από την άλλη πλευρά στην Ευρώπη αυτά τα γραφεία έχουν στη διάθεσή τους μόνο επίσημα δεδομένα που είναι ευρέως διαθέσιμα, λόγω της νομοθεσίας που υπάρχει περί προστασίας προσωπικών δεδομένων.

α_2. Μοντέλα βαθμολόγησης συμπεριφοράς

Τα μοντέλα βαθμολόγησης συμπεριφοράς (*behavioural scoring models* ή BSM) έφεραν την επανάσταση στα CSM περίπου στα τέλη της δεκαετίας του ’80. Είναι μια προφανής επέκταση των ASM. Στα BSM χρησιμοποιούνται όλες οι πληροφορίες από την παρελθοντική πιστωτική συμπεριφορά που είναι διαθέσιμες στην ενδιαφερόμενη επιχείρηση και σκοπός τους είναι η αξιολόγηση συμπεριφοράς των ήδη υπαρχόντων πελατών.

Για τη δόμηση των μοντέλων βαθμολόγησης συμπεριφοράς χρησιμοποιούνται όλες οι διαθέσιμες πληροφορίες από το πρόσφατο παρελθόν για την πιστωτική συμπεριφορά των ήδη υπαρχόντων πελατών. Συνήθως, αυτές οι πληροφορίες λαμβάνονται από την εξέλιξη των λογαριασμών των πελατών για μια χρονική περίοδο 12 μηνών και η ανάπτυξή των BSM στηρίζεται συνήθως σε διαφορετικά χαρακτηριστικά από αυτά που χρησιμοποιούνται για την ανάπτυξη των ASM. Το είδος των χρηματοδοτούμενων δαπανών του εξεταζόμενου πελάτη και το ιστορικό εξόφλησης των δανείων του είναι τις περισσότερες φορές δύο από τα χαρακτηριστικά με τη μεγαλύτερη ικανότητα πρόβλεψης της μελλοντικής συμπεριφοράς του.

Για την ανάπτυξη των BSM μπορούν επίσης να χρησιμοποιηθούν διαθέσιμες πρόσφατες πληροφορίες από γραφεία πίστης καθώς επίσης και κάποια από τα στοιχεία που χρησιμοποιήθηκαν κατά την δόμηση των ASM. Όλα αυτά τα στοιχεία χρησιμοποιούνται για να προβλέψουν τον κίνδυνο αθέτησης υποχρεώσεων πελάτη κατά τη διάρκεια των επόμενων 12 μηνών ή κάποιου άλλου μελλοντικού χρονικού διαστήματος. Χρησιμοποιώντας τα BSM η ενδιαφερόμενη επιχείρηση μπορεί να αξιολογεί τους υπάρχοντες πελάτες της σε τακτά χρονικά διαστήματα (π.χ., σε ετήσια, μηνιαία ή και εβδομαδιαία βάση) ή και οποτεδήποτε άλλοτε η διοίκησή της κρίνει σκόπιμη μια τέτοια αξιολόγηση.

Τα BSM χρησιμοποιούν ακριβώς την ίδια προσέγγιση με τα ASM αντιπροσωπεύοντας την πιστοληπτική ικανότητα κατά τη διάρκεια της περιόδου παρατήρησης με ένα σύνολο στατιστικών στοιχείων (μέγιστη πιστωτική πληρωμή, αριθμός των φορών που έχει ξεπεραστεί το πιστωτικό όριο) και έπειτα αντιστοιχίζονται αυτά τα στατιστικά στοιχεία με τη μελλοντική κατάσταση του πελάτη. Μπορούν να υπάρξουν πολύ περισσότερες από 1000 μεταβλητές για να περιγράψουν την πρόσφατη απόδοση ενός πελάτη, αλλά η ανάλυση και η εκτίμηση του πιστωτικού κινδύνου (εκτίμηση της πιθανότητας αθέτησης υποχρεώσεων σε κάποιο προκαθορισμένο χρονικό διάστημα στο μέλλον) είναι η ίδια με αυτήν των ASM.

Οι βαθμολογίες των πελατών, δηλαδή η αξιολόγηση του κινδύνου που κάθε συγκεκριμένος πελάτης αντιπροσωπεύει, χρησιμοποιείται ως ένδειξη βάσει της οποίας μπορεί

να διαφοροποιηθεί η αντιμετώπιση διαφόρων κατηγοριών πελατών, ανάλογα με τον πιστωτικό κίνδυνο που αντιστοιχεί στην καθεμιά κατηγορία. Τα BSM χρησιμοποιούνται επίσης για να αξιολογηθούν οι διάφορες δραστηριότητες της επιχείρησης, έτσι ώστε να καταστεί δυνατός ο διαρκής έλεγχος της αποτελεσματικότητας των στρατηγικών προγραμμάτων της και η βελτίωσή τους.

Αυτά τα μοντέλα χρησιμοποιούνται από την ενδιαφερόμενη επιχείρηση για να αποφασίσει εάν θα προσφέρει στους οφειλέτες περαιτέρω πίστωση, ποιοι όροι πρέπει να τηρούνται για να την προσφέρει, εάν τη συμφέρει να πουλήσει άλλα προϊόντα σε αυτούς ή εάν χρειάζεται να βελτιώσει τους υπάρχοντες όρους τους επειδή είναι πιθανό οι πελάτες να μεταφέρουν το λογαριασμό τους αλλού. Παρ' όλα αυτά δεν είναι σαφές ότι ταξινομώντας τους πελάτες με βάση την πιθανότητα αθέτησής τους σε ένα δεδομένο χρονικό διάστημα, σύμφωνα με το ποσό του δάνειου που τους είχε χορηγηθεί, κάτω από τους περιορισμούς δανεισμού και κάτω από τους οικονομικούς όρους αγοράς που υπήρχαν στο παρελθόν, είναι ο καλύτερος τρόπος για να αποφασιστούν αυτές οι πιο σύνθετες αποφάσεις. Το BSM είναι ένα βασικό εργαλείο για τον τρέχοντα έλεγχο του πιστωτικού χαρτοφυλακίου και βοηθά στη μείωση του ποσοστού παροχής δανείων υψηλού κινδύνου.

α_3. Μοντέλα βαθμολόγησης κέρδους

Τα τελευταία χρόνια αναπτύσσεται μια νέα γενιά μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας που ονομάζονται μοντέλα βαθμολόγησης κέρδους (*profit scoring models* ή PSM). Η ανάπτυξη των μοντέλων αυτών βοηθούν στις αποφάσεις δανεισμού και αποβλέπουν στην εκπλήρωση των επιχειρησιακών στόχων. Η δόμηση των PSM βασίζεται σε αλλαγές στους στόχους των δανειστών, στις προσδοκίες και τις επιλογές των πελατών και σε ρυθμιστικές πιέσεις. Οι δανειστές πλέον δίνουν πολύ περισσότερη σημασία στις στρατηγικές που θα ακολουθήσουν για να πετύχουν τους επιχειρησιακούς τους στόχους όπως είναι η αύξηση του κέρδους τους και του μεριδίου αγοράς τους. Η εφαρμογή των μοντέλων βαθμολόγησης κέρδους από την ενδιαφερόμενη επιχείρηση στοχεύει περισσότερο στη βελτιστοποίηση όλων των αποφάσεων σχετικά με την πίστωση σε έναν καταναλωτή ή μια επιχείρηση σύμφωνα με τους επιχειρησιακούς στόχους παρά στην πρόβλεψη μόνο του κινδύνου αθέτησης υποχρεώσεων (*default risk - DR*) σε ένα προϊόν δανεισμού.

Με τα μοντέλα βαθμολόγησης κέρδους δίνεται η δυνατότητα να επιλεγεί το πιστωτικό όριο, το επιτόκιο (που ουσιαστικά αποτελεί την τιμή του προϊόντος) και άλλα

χαρακτηριστικά γνωρίσματα προϊόντων που προσφέρονται σε έναν πελάτη ώστε να μεγιστοποιηθεί το κέρδος που θα τους αποφέρει ο συγκεκριμένος πελάτης. Επιπλέον, η διαδικασία της χορήγησης πιστώσεως δεν τελειώνει μόλις γίνει αποδεκτό το προϊόν δανεισμού αφού όπως είναι φυσικό ο τρόπος που οι δανειστές λειτουργούν και η σχέση με τους οφειλέτες επηρεάζουν την αποδοτικότητα του καταναλωτή.

Συνοπτικά, τα PSM είναι ένα εργαλείο που όχι μόνο λαμβάνει υπόψη την πιθανότητα της αποπληρωμής δανείου, αλλά επιτρέπει παράλληλα τον υπολογισμό του κέρδους από τη συνεργασία με έναν ιδιαίτερο πελάτη. Αυτά είναι από τα πιο προηγμένα μοντέλα, δεδομένου ότι λαμβάνουν υπ' όψιν μια ευρεία σειρά πρόσθετων οικονομικών παραγόντων. Με τη βοήθεια των PSM, οι διάφορες κατηγορίες πελατών μπορεί να αντιμετωπιστούν με διαφορετικό τρόπο, όπως για παράδειγμα με:

- προσφορά των βελτιωμένων και πρόσθετων προϊόντων για τους «καλούς» πελάτες.
- αύξηση των πιστωτικών ορίων στις πιστωτικές κάρτες και στα καταναλωτικά δάνεια για τους συνεπείς πελάτες.
- παραχώρηση άδειας σε μερικούς πελάτες για ανακυκλούμενη πίστωση προσφέροντας τη δυνατότητα να υπερβούν τα πιστωτικά τους όρια.
- εντοπισμό των ενδεχομένως ψευδών συναλλαγών.
- προσφορά καλύτερων τιμών στις ανανεώσεις δανείου/ασφαλιστήριων συμβολαίων για τους «καλούς» πελάτες.
- λήψη απόφασης εάν πρέπει να επανεκδοθεί μια ληγμένη πιστωτική κάρτα.
- οδήγηση των «κακών» πελατών σε πιο αυστηρές μεθόδους όπως χρέωση υψηλότερων επιτοκίων.

Τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας δεν διαχωρίζονται μόνο ως προς τον σκοπό για τον οποίο χρησιμοποιούνται (ASM, BSM και PSM) αλλά μπορούν να διαχωριστούν και με βάση τον τύπο των δεδομένων που χρησιμοποιούνται για να αναπτυχθούν. Σύμφωνα με το δεύτερο διαχωρισμό προκύπτουν οι εξής κατηγορίες:

β_1. Πελατειακά σκορόχαρτα

Τα **πελατειακά σκορόχαρτα (*custom scorecards*)** είναι εκείνα που αναπτύσσονται χρησιμοποιώντας δεδομένα που αφορούν αποκλειστικά τους πελάτες μιας επιχείρησης. Για

παράδειγμα, μια τράπεζα χρησιμοποιεί τα δεδομένα απόδοσης των πελατών της για να δομήσει ένα μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας με στόχο την πρόβλεψη της πιθανότητα πτώχευσης (*probability of bankruptcy*). Η επιχείρηση αυτή μπορεί να χρησιμοποιήσει τα εσωτερικά της στοιχεία δηλαδή αυτά που είναι καταχωρημένα στις βάσεις δεδομένων της ή αυτά που λαμβάνονται από ένα γραφείο πίστης, αλλά τα δεδομένα αυτά θα αφορούν μόνο τους πελάτες της.

β_2. Γενικά σκορόχαρτα

Τα **γενικά σκορόχαρτα (*generic scorecards*)** είναι εκείνα τα μοντέλα τα οποία δομούνται χρησιμοποιώντας δεδομένα από πολλούς δανειστές. Σε περιπτώσεις όπου δεν υπάρχουν αρκετά δεδομένα ή δεν υπάρχει η δυνατότητα πρόσβασης σε αυτά ή η ποιότητα των δεδομένων είναι αμφισβητήσιμη, μπορεί να χρειαστεί η ανάπτυξη γενικών σκορόχαρτων από έναν εξωτερικό προμηθευτή ή ένα γραφείο πίστης. Τα γενικά σκορόχαρτα χρησιμοποιούνται συνήθως όταν μια επιχείρηση δραστηριοποιείται σε έναν νέο τομέα ή ένα προϊόν για τα οποία δεν έχει κανένα προηγούμενο στοιχείο. Επίσης, χρησιμοποιούνται όταν τα δεδομένα είναι μεν διαθέσιμα, αλλά το κόστος για την ανάπτυξη πελατειακών σκορόχαρτων είναι πολύ μεγάλο ή δεν υπάρχει επαρκής χρόνος για να αναπτυχθούν πελατειακά σκορόχαρτα.

Ως παράδειγμα αναφέρουμε την περίπτωση μιας ομάδας μικρών τραπεζών, καμία από τις οποίες δεν έχει αρκετά δεδομένα για να δομήσει τα πελατειακά της σκορόχαρτα για δάνεια αυτοκίνητων. Επομένως, αυτό που μπορούν να κάνουν είναι να δομήσουν ένα σκορόχαρτο με όλα τα δεδομένα που έχουν διαθέσιμα όλες οι τράπεζες μαζί και έπειτα να το «μοιραστούν» ή να προσαρμόσουν τα σκορόχαρτα με βάση τα ιδιαίτερα χαρακτηριστικά των χαρτοφυλακίων τους. Τα CSM που δομούνται χρησιμοποιώντας τα δεδομένα των κλάδων της βιομηχανίας και που πωλούνται από τα γραφεία πίστης, είναι ένας τύπος γενικών σκορόχαρτων (βλέπε Siddiqi (2000)).

Αξίζει να σημειώσουμε ότι σε μερικές περιπτώσεις, μπορεί να μην είναι δυνατό να χρησιμοποιηθούν στατιστικά μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας, πελατειακά ή γενικά. Αυτό οφείλεται συνήθως στον πολύ μικρό όγκο δεδομένων, στα μικρά οφέλη που δεν δικαιολογούν τις δαπάνες που συνδέονται με οποιαδήποτε ανάπτυξη CSM ή στο προϊόν για το οποίο κανένα γενικό σκορόχαρτο δεν είναι διαθέσιμο ή κατάλληλο. Σε αυτές τις περιπτώσεις, για να υποστηριχθεί η λήψη αποφάσεων, μπορεί να είναι απαραίτητο να αναπτυχθούν μοντέλα βαθμολόγησης βασισμένα στην κρίση και την εμπειρία. Τέτοια

μοντέλα είναι γνωστά ως «έμπειρα συστήματα» (*expert systems*). Η ανάπτυξη ενός τέτοιου μοντέλου περιλαμβάνει την επιλογή μιας ομάδας χαρακτηριστικών που θεωρούνται καλοί προάγγελοι του κινδύνου όπως με τα στατιστικώς αναπτυγμένα μοντέλα. Η ανάπτυξη βασίζεται στη συνολική εμπειρία και τη διαίσθηση, και το προκύπτον μοντέλο εφαρμόζεται σύμφωνα με τους όρους που πρέπει να τηρούνται.

2.2 Επιλογή δείγματος – περίοδος ωρίμανσης

Τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας αναπτύσσονται χρησιμοποιώντας την υπόθεση ότι «η προηγούμενη απόδοση απεικονίζει την μελλοντική απόδοση». Με βάση αυτήν την υπόθεση, η μελλοντική απόδοση των υποψήφιων πελατών εξαρτάται από την προηγούμενη απόδοση των ήδη υπαρχόντων πελατών οι οποίοι είναι παρόμοιοι, δηλαδή έχουν κοινά χαρακτηριστικά με εκείνους που πρόκειται να αξιολογηθούν με βάση το αναπτυχθέν μοντέλο. Η ανάπτυξη ενός CSM οποιουδήποτε είδους αρχίζει συνήθως με τη λήψη ενός αρκετά μεγάλου δείγματος στοιχείων προηγούμενων πελατών που είχαν υποβάλλει αίτηση για το συγκεκριμένο προϊόν στο παρελθόν και πρέπει να είναι όσο το δυνατόν πιο αξιόπιστο και πιο αντιπροσωπευτικό του πληθυσμού που θέλουμε να εξετάσουμε. Σε έρευνες τέτοιου είδους, τα δεδομένα συνήθως συμπληρώνονται και με στοιχεία σχετικά με την εξέλιξη των λογαριασμών (απόδοση) των εξεταζομένων πελατών για κάποιο προκαθορισμένο χρονικό διάστημα που ονομάζεται διάρκεια ωρίμανσης, έτσι ώστε να καθοριστεί εάν αυτοί ήταν «καλοί» ή «κακοί». Επίσης, χρειάζεται να προσδιοριστεί ο ελάχιστος αριθμός δείγματος «καλών» και «κακών» πελατών που θα χρησιμοποιηθούν για αυτόν το σκοπό.

Η ποσότητα των δεδομένων που απαιτούνται για την ανάπτυξη ενός CSM ποικίλλει, αλλά γενικά πρέπει να ικανοποιούνται οι προϋποθέσεις του στατιστικά σημαντικού και τυχαίου δείγματος. Το επαρκές μέγεθος του τυχαίου δείγματος εξαρτάται από τον καθορισμό του «καλού» και του «κακού» πελάτη. Γενικά, είναι δυσκολότερο να βρεθούν στον πληθυσμό αρκετοί «κακοί» λογαριασμοί παρά «καλοί». Στο συνολικό πληθυσμό η αναλογία των «καλών» σε σχέση με τους «κακούς» είναι στις περισσότερες περιπτώσεις περίπου 20:1 με αποτέλεσμα πολλές φορές, αν κρατηθεί η ίδια αναλογία στο δείγμα να μην υπάρχουν αρκετοί

«κακοί» ώστε να προσδιοριστούν τα χαρακτηριστικά τους. Για αυτόν το λόγο, το δείγμα των «καλών» και των «κακών» έχει διαφορετική σύνθεση απ' ότι στο συνολικό πληθυσμό (μια συνηθισμένη τακτική είναι να τηρείται μια αναλογία 50:50).

Σύμφωνα με τον Siddiqi (2000), για την ανάπτυξη ενός μοντέλου επιλέγονται τυχαία περίπου 2.000 «κακοί» και 2.000 «καλοί» ήδη υπάρχοντες πελάτες που παρακολουθούνται για ένα προκαθορισμένο χρονικό διάστημα. Για τα BSM, το δείγμα επιλέγεται από μια ομάδα ενεργών λογαριασμών σε ένα δεδομένο χρονικό σημείο, ή σε μια συγκεκριμένη κατάσταση ασυνέπειας. Για την ανάπτυξη των ASM μπορεί να χρειαστούν επιπλέον 2.000 απορριφθείσες αιτήσεις προκειμένου να γίνει εφικτή η **συμπερασματολογία απορριφθέντων (*reject inference*)**.

Τα στοιχεία που αφορούν την πιστωτική ιστορία ενός πελάτη που θα συμπεριληφθεί στο δείγμα συλλέγονται από λογαριασμούς που έχουν ανοιχτεί τα προηγούμενα δύο με πέντε χρόνια και αυτά τα στοιχεία συνήθως αφορούν:

- τον αριθμό λογαριασμού
- την ημερομηνία ανοίγματος λογαριασμού ή την ημερομηνία υποβολής της αίτησης για πίστωση
- τον αριθμό καθυστερήσεων πληρωμής δόσεων κατά τη διάρκεια του δανείου
- τον δείκτη αποδοχής/απόρριψης αριθμού δανείων του συγκεκριμένου πελάτη
- το προϊόν και τον καθορισμό άλλων παραγόντων που προσδιορίζουν την ομάδα που ανήκει ο πελάτης
- την κατάσταση τρεχούμενου λογαριασμού (π.χ. ενεργός, ανενεργός, κλειστός, απάτη).

Τα δεδομένα που συλλέγονται αφού οι πελάτες έχουν ταξινομηθεί σε «καλούς» και «κακούς» αποτελούν το **δείγμα ανάπτυξης (*development/training sample*)** μέσω του οποίου δημιουργείται το μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας. Η επιλογή των δειγμάτων ανάπτυξης από μια ώριμη ομάδα γίνεται για να ελαχιστοποιηθεί η πιθανότητα λανθασμένης κατάταξης των αποδόσεων κάθε πελάτη (δηλαδή σε όλους τους λογαριασμούς δίνεται αρκετός χρόνος μέχρι να θεωρηθούν «κακοί») και για να εξασφαλισθεί ότι ο καθορισμός του «κακού» δεν είναι αποτέλεσμα ενός ανώριμου δείγματος. Για παράδειγμα, εάν το δείγμα ανάπτυξης επιλέχτηκε από λογαριασμούς που άνοιξαν επτά μήνες πριν, περίπου 4.5% του δείγματος θα ταξινομούταν ως «κακοί». Ένα ώριμο δείγμα για αυτό το χαρτοφυλάκιο πρέπει να έχει ένα κακό ποσοστό περίπου 6%. Επομένως μερικοί λογαριασμοί από αυτήν την περίοδο όπου το προϊόν δεν έχει ωριμάσει και που στην πραγματικότητα είναι

«κακοί» θα χαρακτηρίζονταν λανθασμένα ως «καλοί» εάν το δείγμα ανάπτυξης επρόκειτο να ληφθεί από εκείνη την περίοδο. Έτσι καθίσταται εφικτή η εμπεριστατωμένη αξιολόγηση της συμπεριφοράς των εξεταζομένων πελατών και συνεπώς ο αξιόπιστος χαρακτηρισμός τους ως «καλοί» ή «κακοί».

Ο χρόνος ωρίμανσης ποικίλλει ανάλογα με το προϊόν και του τρόπου καθορισμού του «καλού» και του «κακού». Συνήθως, για τις πιστωτικές κάρτες ο λογαριασμός θεωρείται ώριμος μετά από 18 με 24 μήνες, ενώ για τα στεγαστικά δάνεια 3 με 5 έτη. Αυτό είναι λογικό δεδομένου ότι τα χαρτοφυλάκια πιστωτικών καρτών είναι υψηλότερου κινδύνου από αυτά των στεγαστικών δανείων, και επομένως αναμένεται να φτάσουν στο ίδιο επίπεδο μη εκπλήρωσης των υποχρεώσεων πολύ γρηγορότερα. Επίσης, για να καθοριστεί η περίοδος ωρίμανσης για τους λογαριασμούς που ανοίγουν, θα πρέπει να ληφθεί υπ' όψιν το φαινόμενο της εποχικότητας, δηλαδή το χρονικό διάστημα που θα χρειαστεί για ένα προϊόν να ωριμάσει συνήθως δε θα πρέπει να ξεπερνάει τα δύο χρόνια και θα πρέπει να είναι τέτοιο ώστε να εξαλείφονται οι πιθανές συνέπειες των εποχικών διακυμάνσεων της δραστηριότητας της αγοράς πάνω στη ζήτηση χρηματοπιστωτικών προϊόντων και υπηρεσιών. Αυτό εξασφαλίζει ότι το δείγμα ανάπτυξης δεν περιλαμβάνει οποιαδήποτε στοιχεία από τις «ανώμαλες» περιόδους.

Για την ανάπτυξη μοντέλων βαθμολόγησης συμπεριφοράς, επιλέγονται οι λογαριασμοί που βρίσκονται σε εξέλιξη σε κάποια χρονική στιγμή που ονομάζεται χρόνος παρατήρησης. Η περίοδος πριν από τον χρόνο παρατήρησης στην οποία αναλύεται η συμπεριφορά του πελάτη καλείται **περίοδος απόδοσης (*performance period*)** και διαρκεί συνήθως 6 έως 12 μήνες. Η περίοδος μετά από τον χρόνο παρατήρησης είναι η **περίοδος έκβασης (*outcome period*)**, η οποία είναι συνήθως 12 μήνες, και ο πελάτης, ταξινομείται ως «καλός» ή «κακός» ανάλογα με το αποτέλεσμά του στο τέλος αυτής της περιόδου έκβασης.

2.3 Καθορισμός «καλού – κακού» πελάτη

Τα άτομα του δείγματος ανάπτυξης ταξινομούνται αρχικά σε τρεις κατηγορίες: **«κακοί»**, **«καλοί»** και **«απροσδιόριστοι»** πελάτες. Ο σωστός προσδιορισμός του «καλού» και του «κακού» πελάτη είναι πολύ σημαντικός για την ανάπτυξη ενός μοντέλου βαθμολόγησης

πιστοληπτικής ικανότητας γιατί, χρησιμοποιώντας διαφορετικούς ορισμούς για τον «κακό» πελάτη, προκύπτει διαφορετικός αριθμός «κακών» λογαριασμών κάθε φορά. Σε περιπτώσεις λογαριασμών που έχει συμβεί πτώχευση, απάτη, μη καταβολή δόσεων για πολύ μεγάλο συνεχόμενο χρονικό διάστημα, οι πελάτες χαρακτηρίζονται αναμφισβήτητα «κακοί». Όμως, σε περιπτώσεις που υπάρχουν απλές καθυστερήσεις και ασυνέπειες στην αποπληρωμή του δανείου, ο χαρακτηρισμός του κάθε λογαριασμού εξαρτάται από το επίπεδο της ασυνέπειας του πελάτη.

Ο καθορισμός του «κακού» πελάτη εξαρτάται κυρίως από το είδος της πίστωσης, από την πιστωτική πολιτική της ενδιαφερόμενης επιχείρησης αλλά και από διάφορους άλλους παράγοντες κάποιοι από τους οποίους είναι οι εξής:

- Οι οργανωτικοί στόχοι της επιχείρησης. Εάν ο στόχος της επιχείρησης είναι να αυξηθεί το κέρδος, τότε για τον προσδιορισμό του «κακού» πελάτη πρέπει να χρησιμοποιηθεί ένα σημείο ασυνέπειας όπου ο λογαριασμός γίνεται ασύμφορος ή μη κερδοφόρος.
- Το προϊόν ή ο σκοπός για τον οποίο αναπτύσσεται το μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας. Για παράδειγμα, ο σκοπός μπορεί να είναι η πρόβλεψη πτώχευσης, απάτης ή ένα ποσοστό εισπράξεων μέχρι κάποια συγκεκριμένη χρονική στιγμή.
- Ο καθορισμός αυτός πρέπει να είναι εύκολα ερμηνεύσιμος και ανιχνεύσιμος, όπως για παράδειγμα η ασυνέπεια 90 ημερών, η πτώχευση και η επιβεβαιωμένη απάτη. Αντίθετα, προσδιορισμοί όπως «τρεις φορές ασυνέπεια 30 ημερών» ή «δύο φορές ασυνέπεια 60 ημερών» που μπορούν να είναι ακριβέστεροι, είναι πολύ πιο δύσκολο να ανιχνευτούν στο δείγμα και μπορεί να μην είναι κατάλληλοι για όλες τις επιχειρήσεις.
- Οι εξωτερικές απαιτήσεις ή άλλα ρυθμιστικά πλαίσια. Για παράδειγμα, το νέο πλαίσιο κεφαλαιακής επάρκειας της Βασιλείας ΙΙ γενικά θεωρεί ότι ένας πελάτης βρίσκεται σε κατάσταση αθέτησης υποχρεώσεων όταν έχει παρουσιάσει ασυνέπεια 90 ημερών (το κριτήριο αυτό θεωρείται λογικό γιατί, η μεγαλύτερη πλειοψηφία των λογαριασμών που παρουσιάζουν 90 μέρες ασυνέπεια, δεν «γιατρεύονται» αλλά γίνονται χειρότεροι και η περίοδος ασυνέπειας μεγαλώνει).

Μόλις καθοριστούν οι «κακοί» δανειολήπτες καθορίζονται και οι «καλοί». Ο προσδιορισμός του «καλού» πελάτη εξαρτάται επίσης από το είδος του πιστωτικού προϊόντος και από την πιστωτική πολιτική που ακολουθεί η ενδιαφερόμενη επιχείρηση. Κάποια στοιχεία που χαρακτηρίζουν «καλούς» λογαριασμούς είναι όταν ο πελάτης δεν έχει

παρουσιάσει ποτέ κάποιου είδους ασυνέπειας, δεν υπήρξε ποτέ πτώχευση, δεν υπήρξε ποτέ κάποιου είδους απάτη, όταν είναι κερδοφόρος ή παρουσιάζει θετική καθαρή παρούσα αξία.[2]

Οι λογαριασμοί που δεν ανήκουν σε καμία από τις δύο παραπάνω κατηγορίες, τελικά κατατάσσονται στην κατηγορία των «απροσδιόριστων» πελατών. Πελάτες που κατατάσσονται σε αυτήν την κατηγορία είναι συνήθως αυτοί που οι λογαριασμοί τους δεν έχουν αρκετή ιστορία ώστε να ταξινομηθούν ως «καλοί» ή «κακοί», ή που έχουν παρουσιάσει κάποια μικρής σημασίας περιστατικά ασυνέπειας. Για παράδειγμα, λογαριασμοί που κατατάσσονται σε αυτήν την κατηγορία μπορεί να είναι λογαριασμοί με 30 ή 60 ημέρες ασυνέπειας, ανενεργοί ή ακυρωμένοι λογαριασμοί, πιστώσεις που εγκρίθηκαν αλλά τελικά δεν χρησιμοποιήθηκαν ή λογαριασμοί που αποφέρουν μηδενική καθαρή παρούσα αξία.

Αυτό που ενδιαφέρει πραγματικά έναν δανειστή είναι αν ο πελάτης θα του αποφέρει κέρδος ή όχι. Εκείνοι που ορίζονται ως «καλοί» αφού ποτέ δεν βρίσκονται σε κατάσταση χρέους αποφέρουν σίγουρα κέρδος στην επιχείρηση, ενώ εκείνοι οι πελάτες που ορίζονται ως «κακοί» γιατί για παράδειγμα παρουσιάζουν τρεις μήνες καθυστέρηση στις οφειλές τους είναι σίγουρα μη κερδοφόροι. Επιπλέον, εκείνοι που ορίζονται ως «απροσδιόριστοι» μπορεί να είναι ή όχι κερδοφόροι λόγω των απρόβλεπτων οικονομικών αλλαγών πέρα από τους όρους του δανείου. Η γενική αρχή που χρησιμοποιείται είναι να κατασκευαστεί ένας κανόνας που θα χωρίζει σίγουρα τον κερδοφόρο από τον ασύμφορο πελάτη που είναι ένας λογικός στόχος.

Αξίζει να σημειωθεί ότι, στο δείγμα ανάπτυξης που θα χρησιμοποιηθεί για την ανάπτυξη ενός CSM, περιλαμβάνονται μόνο αυτοί οι πελάτες που έχουν χαρακτηριστεί ως «καλοί» ή «κακοί» και αυτό γιατί οι αιτήσεις για πίστωση των υποψήφιων πελατών ή θα εγκριθούν ή θα απορριφθούν (δεν είναι δυνατό να υπάρξει μεσαία κατηγορία). Η κατηγορία των «απροσδιόριστων» πελατών μπορεί να χρησιμοποιηθεί στην περίπτωση που ο σκοπός της ανάπτυξης των CSM είναι να κατατάξουν τους υποψήφιους πελάτες ως προς την πιθανότητα αθέτησης των υποχρεώσεών τους. Με αυτόν τον τρόπο μπορεί να απεικονιστεί επαρκώς ο πραγματικός πληθυσμός, δεδομένου ότι οι υποψήφιοι που θα βαθμολογηθούν και θα κριθούν μπορεί να ανήκουν και στις τρεις κατηγορίες.

---

[2] Η Καθαρή Παρούσα Αξία (ΚΠΑ) μιας επένδυσης είναι η διαφορά μεταξύ της παρούσας αξίας των Καθαρών Ταμειακών Ροών ΚΤΡ της επένδυσης, προεξοφλημένων στο παρόν με επιτόκιο i του αρχικού κεφαλαίου K_0 που απαιτείται για να πραγματοποιηθεί η επένδυση σήμερα.

$$ΚΠΑ = \sum_{t=1}^{n} \frac{ΚΤΡ_i}{(1+i)^n} - K_0$$

Σε περίπτωση που η παρούσα αξία των αναμενόμενων ταμειακών ροών από την επένδυση σήμερα είναι πιο υψηλή από το απαιτούμενο κόστος της επένδυσης, δηλαδή η $ΚΠΑ > 0$, η επένδυση γίνεται αποδοτική.

2.4 Κατάτμηση

Σε μερικές περιπτώσεις, η χρησιμοποίηση πολλών μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας για ένα χαρτοφυλάκιο παρέχει καλύτερη διαφοροποίηση κινδύνου σε σχέση με τη χρησιμοποίηση ενός μόνο μοντέλου. Αυτό συμβαίνει συνήθως όταν ένας πληθυσμός αποτελείται από ευδιάκριτα υποσύνολα πληθυσμού και επομένως η εφαρμογή ενός μόνο μοντέλου βαθμολόγησης δε θα λειτουργήσει αποτελεσματικά για όλους τους υποπληθυσμούς, αφού πιθανόν να απαιτούνται διαφορετικά χαρακτηριστικά για να προβλέψουν τον κίνδυνο για τα διάφορα υποσύνολα που υπάρχουν. Οι λόγοι για τους οποίους καταφεύγουμε στην παραπάνω διαδικασία μπορεί να είναι είτε στατιστικοί ή να αφορούν την πολιτική και τις στρατηγικές της ενδιαφερόμενης επιχείρησης. Για παράδειγμα, για την ανάπτυξη των BSM είναι συνηθισμένο να δομούνται άλλα μοντέλα για τους πρόσφατους πελάτες και άλλα για τους μακροχρόνιους πελάτες. Αυτό συμβαίνει επειδή μερικά χαρακτηριστικά, όπως το μέσο οφειλόμενο υπόλοιπο στους τελευταίους έξι μήνες, είναι διαθέσιμα για τους πιο παλιούς πελάτες αλλά όχι για τους πιο καινούργιους. Μια άλλη πολιτική της επιχείρησης είναι τα δεδομένα που αφορούν νεότερους σε ηλικία πελάτες να υποβάλλονται σε διαφορετική επεξεργασία από εκείνα που αφορούν τους μεγαλύτερους σε ηλικία πελάτες. Ένας τρόπος για να γίνει αυτό είναι να αναπτυχθούν διαφορετικά μοντέλα για τους πελάτες άνω των 25 ετών και κάτω από 25 ετών.

Η διαδικασία που βοηθάει στον ορθολογικό προσδιορισμό των διαφορετικών υποσυνόλων ενός πληθυσμού καλείται **κατάτμηση (*segmentation*)** και είναι ένα είδος ομαδοποίησης του πληθυσμού. Σύμφωνα με τον Siddiqi (2000) υπάρχουν δύο βασικοί τρόποι με τους οποίους μπορεί να γίνει κατάτμηση:

(α) Με βάση την εμπειρία και τους σκοπούς της ενδιαφερόμενης επιχείρησης (και εν συνεχεία επαλήθευση με χρήση αναλυτικών μεθόδων). Μια απλή μέθοδος για να επιβεβαιωθούν οι ομάδες κατάτμησης και να επαληθευτεί η ανάγκη για την κατάτμηση είναι να αναλυθεί η συμπεριφορά κάποιων χαρακτηριστικών στις διαφορετικά προκαθορισμένες ομάδες. Στην περίπτωση που το χαρακτηριστικό σύμφωνα με το οποίο έγινε η ομαδοποίηση (π.χ. ηλικία) εμφανίζει διαφορετική συμπεριφορά στις διαφορετικές ομάδες, τότε θεωρείται ότι αυτή η κατάτμηση έγινε σωστά. Αντιθέτως, εάν το χαρακτηριστικό αυτό συμπεριφέρεται με τον ίδιο τρόπο στις διαφορετικές

ομάδες, τότε τα πρόσθετα μοντέλα δεν είναι απαραίτητα, δεδομένου ότι δεν υπάρχει καμία διαφοροποίηση.

(β) Με χρήση διάφορων στατιστικών μεθοδολογιών, για παράδειγμα της ομαδοποίησης κατά συστάδες. Η **ομαδοποίηση κατά συστάδες (*clustering*)** είναι μια ευρέως χρησιμοποιημένη τεχνική για να προσδιοριστούν οι ομάδες που είναι παρόμοιες η μία με την άλλη όσον αφορά τις μεταβλητές που χρησιμοποιούνται. Η μέθοδος αυτή τοποθετεί τους πελάτες σε ομάδες ή «συστάδες» σύμφωνα με τα δεδομένα που υπάρχουν. Οι πελάτες κάποια έννοια, και οι πελάτες στις διαφορετικές συστάδες τείνουν να είναι διαφορετικοί. Μια μέθοδος που χρησιμοποιείται για να διαμορφωθούν οι συστάδες είναι η k-means. Η ομαδοποίηση κατά συστάδες μπορεί να εκτελεσθεί βάσει της Ευκλείδιας απόστασης (ή άλλων εναλλακτικών αποστάσεων). Οι παρατηρήσεις διαιρούνται σε συστάδες έτσι ώστε κάθε παρατήρηση να ανήκει σε μια και μόνο μία συστάδα. Πρέπει να σημειωθεί ότι η ομαδοποίηση κατά συστάδες προσδιορίζει τις ομάδες που έχουν παρόμοια χαρακτηριστικά και όχι παρόμοια απόδοση. Κατά συνέπεια, οι συστάδες μπορεί να είναι διαφορετικές, αλλά μπορεί να έχουν σχεδόν την ίδια απόδοση κινδύνου. Οι συστάδες επομένως πρέπει να αναλυθούν περαιτέρω, χρησιμοποιώντας, για παράδειγμα την ανάλυση του ποσοστού των «κακών» ώστε να εξασφαλιστεί ότι η κατάτμηση που έχει γίνει αφορά τις ομάδες με διαφορετικά προφίλ για την απόδοση κινδύνου.

Όποια μέθοδος και αν ακολουθηθεί και όποιες ομάδες και αν επιλεγούν, θα πρέπει οι ομάδες να είναι αρκετά μεγάλες έτσι ώστε να επιτρέπουν να γίνει ουσιαστική και σωστή επιλογή δείγματος για την ανάπτυξη του κάθε μοντέλου. Επιπλέον, οι ομάδες που έχουν ευδιάκριτη απόδοση κινδύνου αλλά δεν περιέχουν μεγάλο όγκο παρατηρήσεων για ξεχωριστή ανάπτυξη μοντέλων μπορούν να αντιμετωπιστούν διαφορετικά. Η κατάτμηση που γίνεται είτε χρησιμοποιώντας την εμπειρία είτε τις στατιστικές μεθόδους πρέπει επίσης να γίνεται λαμβάνοντας υπ' όψιν τα μελλοντικά σχέδια της ενδιαφερόμενης επιχείρησης. Οι περισσότερες αναλύσεις είναι βασισμένες στο παρελθόν, αλλά τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας πρέπει να εφαρμοστούν στο μέλλον, δηλαδή στις μελλοντικές υποψήφιες ομάδες. Ένας τρόπος για να επιτευχτεί αυτό είναι, για παράδειγμα, η κατάτμηση να γίνει με βάση τους επιδιωκόμενους πελάτες (πελάτες - στόχους) ώστε να προσδιοριστεί ένα βέλτιστο σύνολο ομάδων που θα μεγιστοποιήσει την απόδοση της επιχείρησης.

2.5 Είδη δεδομένων που χρησιμοποιούνται σε μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας

Αφού τελειώσει η διαδικασία της κατάτμησης και έχει καθοριστεί ο τρόπος προσδιορισμού του «καλού» και του «κακού» πελάτη, πρέπει να προσδιοριστεί το είδος των δεδομένων που θα χρησιμοποιηθούν για την ανάπτυξη των στατιστικών μοντέλων βαθμολόγησης. Θα πρέπει επίσης να προσδιοριστεί το μέγεθος δείγματος για κάθε κατηγορία τμήματος και απόδοσης (συμπεριλαμβανομένων και των απορριφθέντων πελατών), και ένας λεπτομερής κατάλογος χαρακτηριστικών από τις εσωτερικές και εξωτερικές πηγές που απαιτούνται στο δείγμα ανάπτυξης για το κάθε τμήμα.

Το είδος των δεδομένων που χρησιμοποιούνται για την ανάπτυξη των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας εξαρτάται από το είδος της πίστωσης και από τον σκοπό για τον οποίο αυτά εφαρμόζονται. Τα περισσότερα από τα δεδομένα που χρησιμοποιούνται είναι απαραίτητο να σχετίζονται με τον κίνδυνο αθέτησης υποχρεώσεων ώστε να ενισχύουν την προβλεψιμότητα του στατιστικού μοντέλου που αναπτύσσεται. Δεν είναι αναγκαίο να δικαιολογηθεί η ύπαρξη κάθε μεταβλητής που χρησιμοποιείται στα μοντέλα, αρκεί αυτή να βοηθάει στην πρόβλεψη.

Κάθε πιστωτικό ίδρυμα έχει το δικαίωμα να χρησιμοποιεί όποια νόμιμα χαρακτηριστικά θεωρεί ότι είναι καλύτερα για τη διαμόρφωση του μοντέλου. Βέβαια, κατά καιρούς υπήρξε μια διαμάχη για το αν θα πρέπει να χρησιμοποιούνται ή όχι για την ανάπτυξη των μοντέλων ορισμένες κατηγορίες δεδομένων με την έννοια της παραβίασης των προσωπικών δεδομένων των πελατών. Το είδος των δεδομένων που χρησιμοποιούνται είναι διαφορετικά σε κάθε χώρα και αυτό εξαρτάται από την ισχύουσα νομοθεσία σε αυτές. Για παράδειγμα, σε ορισμένες χώρες δεν επιτρέπεται η χρήση κάποιων χαρακτηριστικών όπως η φυλή, η θρησκεία ή το χρώμα δέρματος. Μια άποψη είναι ότι ο στόχος αυτής της νομοθεσίας είναι να εξασφαλιστεί ότι δεν επηρεάζεται η απόφαση από κριτήρια που έχουν να κάνουν με προκαταλήψεις, έτσι ώστε η ταξινόμηση να γίνει σύμφωνα με το στόχο και μόνο, δηλαδή την εξασφάλιση μειωμένου πιστωτικού κινδύνου.

Από την άλλη πλευρά, υπάρχουν κάποια χαρακτηριστικά που νομικά δεν απαγορεύεται η χρήση τους, αλλά παρ' όλα αυτά δε χρησιμοποιούνται για να προβλέψουν τον κίνδυνο αθέτησης υποχρεώσεων επειδή θεωρούνται κατακριτέα από το κοινωνικό σύνολο. Για παράδειγμα, ένα κακό ιστορικό υγείας ή επανειλημμένες παραβιάσεις του Κώδικα Οδικής

Κυκλοφορίας, ενώ είναι προάγγελοι αυξημένου κινδύνου, δε χρησιμοποιούνται από τους δανειστές επειδή φοβούνται την αποδοκιμασία των πολιτών. Γενικότερα, είναι υποκειμενικό ποια δεδομένα θα χρησιμοποιήσει η κάθε επιχείρηση για να κατατάξει τους πελάτες της ως προς τον κίνδυνο αθέτησης υποχρεώσεων ή για να τους χαρακτηρίσει «καλούς» ή «κακούς».

Στις περισσότερες περιπτώσεις, τα δεδομένα που χρησιμοποιούνται λαμβάνονται από τα στοιχεία που έχουν συμπληρώσει οι πελάτες στην αίτηση τους, επομένως είναι σημαντικό στην αίτηση να περιλαμβάνονται μόνο στοιχεία που η συμπλήρωσή τους είναι υποχρεωτική για όλους έτσι ώστε να αποφευχθούν περιπτώσεις ελλιπών δεδομένων. Όσο πιο πολλές μεταβλητές – χαρακτηριστικά περιλαμβάνονται στο μοντέλο τόσο καλύτερη θα είναι η προβλεψιμότητά του. Ωστόσο, υπάρχουν κάποια πρακτικά προβλήματα που πρέπει να αντιμετωπιστούν. Για παράδειγμα η ύπαρξη πολλών ερωτήσεων ή ερωτήσεων που απαιτούν πολύ χρόνο για να απαντηθούν λειτουργούν αποτρεπτικά για τους πελάτες οι οποίοι θα προτιμήσουν να απευθυνθούν σε κάποιον άλλο δανειστή. Επομένως, το είδος και ο αριθμός των ερωτήσεων που θα συμπεριλαμβάνονται στην αίτηση παίζουν πολύ σημαντικό ρόλο στην ανάπτυξη ενός CSM.

Ένα άλλο πρόβλημα κατά την κατάστρωση του ερωτηματολογίου είναι η καταγραφή χαρακτηριστικών, όπως ο τύπος επαγγέλματος και βιομηχανίας, τα οποία επιδέχονται υποκειμενική ερμηνεία. Διαφορετικοί άνθρωποι μπορεί να τοποθετήσουν το ίδιο πρόσωπο σε διαφορετικούς τύπους επαγγέλματος. Αυτός είναι ο λόγος που τις περισσότερες φορές η απάντηση «άλλο» στο ερώτημα που αφορά το επάγγελμα λαμβάνει αρκετά συχνά ποσοστό πάνω από 75%. Ακόμη και σε περιπτώσεις όπου τέτοια χαρακτηριστικά έχουν αποδειχθεί ότι είναι προφητικά, μια κακή ερμηνεία μπορεί να καταστήσει την ευρωστία τους αμφισβητήσιμη. Η υποκειμενικότητα αυτή μπορεί να μειωθεί αν η ερμηνεία των δεδομένων γίνεται από κάποιον που διαθέτει εμπειρία στον πιστωτικό κίνδυνο.

Στα δεδομένα που θα χρησιμοποιηθούν τελικά θα πρέπει να αποφεύγεται η δημιουργία δεικτών (αναλογιών) που μπορεί να έχουν μεγάλη προβλεψιμότητα αλλά είναι δύσκολο να ερμηνευτούν. Γενικά, οποιοσδήποτε δείκτης χρησιμοποιηθεί καλό θα είναι να επιδέχεται φυσική ερμηνεία. Τέλος, πρέπει να εξασφαλιστεί ότι οποιοδήποτε στοιχείο εξετάζεται για την ανάπτυξη του CSM είναι διαχρονικό δηλαδή μπορεί να συλλεχθεί και για τους μελλοντικούς πελάτες.

Πριν ξεκινήσει η διαδικασία για την ανάπτυξη του στατιστικού μοντέλου θα πρέπει να εξεταστεί εάν τα εσωτερικά δεδομένα που διαθέτει η ενδιαφερόμενη επιχείρηση είναι

αξιόπιστα ή αν έχουν υποστεί κάποια αλλαγή. Δημογραφικά στοιχεία και άλλα στοιχεία της αίτησης που δεν ελέγχονται, όπως για παράδειγμα το εισόδημα, είναι πιο ευαίσθητα στην παραποίηση. Σε περίπτωση που δεν είναι δυνατό να ελεγχθούν όλα τα δεδομένα, τότε η ενδιαφερόμενη επιχείρηση θα πρέπει να εφοδιαστεί με άλλα στοιχεία που διαθέτουν τα γραφεία πίστης, όπως στοιχεία ακίνητων περιουσιών και χρηματοοικονομικούς δείκτες που είναι πιο ανθεκτικά και μπορούν να χρησιμοποιηθούν.

Είναι γνωστό ότι ένα CSM θα αναπτυχθεί με βάση τα δεδομένα δύο έως τριών ετών, και αναμένεται να είναι σε λειτουργία για τα επόμενα δύο έτη περίπου. Άρα, οι προηγούμενες και οι αναμενόμενες μελλοντικές τάσεις πρέπει να εξεταστούν πριν την επιλογή του είδους των δεδομένων. Για παράδειγμα, αν ο σκοπός για την ανάπτυξη του CSM είναι η χορήγηση καταναλωτικών δανείων θα πρέπει η ενδιαφερόμενη επιχείρηση να ενημερωθεί από ένα γραφείο πίστης αν ο ανταγωνισμός έχει μειωθεί ή έχει αυξηθεί τα τελευταία δύο με τρία χρόνια. Ενώ αυτή η ενημέρωση δεν θα αλλάξει το είδος των δεδομένων ανάπτυξης, μπορεί να χρησιμοποιηθεί για να γίνει καλύτερη διαχείριση των προσδοκιών και να σχεδιαστούν κατάλληλες στρατηγικές για την επίτευξη μεγαλύτερου κέρδους. Μια αύξηση στον ανταγωνισμό θα αυξήσει το μέσο αριθμό αιτήσεων που θα κάνει ένας υποψήφιος στο γραφείο πίστης. Για παράδειγμα, όταν ένα μοντέλο αναπτύσσεται σε μια περίοδο χρηματοπιστωτικής κρίσης όπου οι χορήγηση δανείων είναι πολύ δύσκολη, το μοντέλο που αναπτύσσεται με βάση μόνο τα ιστορικά στοιχεία θα μεταχειριστεί εκείνους που έχουν κάνει πάνω από τέσσερις αιτήσεις σε 12 μήνες σαν υψηλού κινδύνου υποψήφιους ενώ τα νέα δεδομένα μπορούν να προτείνουν ότι οι τέσσερις αιτήσεις συνδέονται με μέσο κίνδυνο.

Για τα μοντέλα βαθμολόγησης συμπεριφοράς, μπορεί να προστεθεί ένα πλήθος δεδομένων που αφορούν τις συναλλαγές των πελατών με την ενδιαφερόμενη επιχείρηση. Το πιο κοινό χαρακτηριστικό που χρησιμοποιείται είναι ο μέσος όρος, το μέγιστο ή το ελάχιστο του οφειλόμενου υπολοίπου για τους περασμένους 6 ή 12 μήνες. Άλλα χαρακτηριστικά που θα μπορούσαν να συμπεριληφθούν είναι, η συνολική αξία των πιστωτικών συναλλαγών κατά τη διάρκεια τέτοιων περιόδων ή χαρακτηριστικά που καταδεικνύουν την ανεπαρκή συμπεριφορά, όπως ο αριθμός των φορών που έχει ξεπεραστεί το πιστωτικό όριο στις πιστωτικές κάρτες ή ο αριθμός των υπενθυμίσεων που έχουν γίνει στον πελάτη για αποπληρωμή των δόσεών του.

Για όλα τα είδη των CSM μπορεί να χρησιμοποιηθούν πολλά χαρακτηριστικά τα οποία είναι συσχετισμένα μεταξύ τους, ωστόσο από αυτά τα χαρακτηριστικά θα επιλεγούν τελικά

τα καλύτερα για να αναπτυχθεί το μοντέλο. Το είδος των δεδομένων που χρησιμοποιούνται για την ανάπτυξη ενός CSM εξαρτάται από το είδος του δανείου που ζητάει ο κάθε πελάτης. Παρακάτω περιγράφονται τα πιο σημαντικά στοιχεία που ζητούνται από τον κάθε πελάτη, ανάλογα με το είδος του δανείου που αιτείται.

α. Καταναλωτικά, προσωπικά, στεγαστικά δάνεια και πιστωτικές κάρτες

Για τη χορήγηση καταναλωτικών δανείων υπάρχουν διάφορα χαρακτηριστικά που χρησιμοποιούνται στην ανάπτυξη CSM. Κάποιες επιχειρήσεις θεωρούν σημαντική τη σταθερότητα του καταναλωτή σε μία κατάσταση και για το λόγο αυτό ενδιαφέρονται ιδιαίτερα για χαρακτηριστικά όπως ο χρόνος διαμονής στην τωρινή κατοικία ή η χρονική διάρκεια στην τωρινή απασχόληση. Κάποιες άλλες επιχειρήσεις λαμβάνουν υπόψη την οικονομική κατάσταση του καταναλωτή και θεωρούν ως σημαντικά στοιχεία την κατοχή (άλλων) πιστωτικών καρτών, το χρόνο συνεργασίας με την τράπεζα, τα τετραγωνικά μέτρα της μόνιμης κατοικίας, την απασχόληση, την απασχόληση του συζύγου, τον αριθμό παιδιών κ.α. Κάποια από τα πιο βασικά ήδη δεδομένων που χρησιμοποιούνται για την ανάπτυξη ενός CSM για καταναλωτικά, προσωπικά, στεγαστικά δάνεια και πιστωτικές κάρτες είναι:

- Ταχυδρομικός Κώδικας (περιοχή διαμονής).
- Συνολικά χρόνια στην τωρινή κατοικία.
- Κατάσταση κατοικίας (ιδιοκτήτης, με ενοίκιο).
- Απασχόληση.
- Χρόνος στην τωρινή εργασία.
- Οικογενειακή κατάσταση.
- Μηνιαίος μισθός.
- Άλλα εισοδήματα.
- Αριθμός προστατευόμενων μελών.
- Αριθμός τέκνων.
- Στοιχεία τρεχούμενων λογαριασμών.
- Στοιχεία λογαριασμών ταμιευτηρίου.
- Αριθμός πιστωτικών καρτών που έχει ο πελάτης στην κατοχή του και στοιχεία αυτών.
- Αιτηθέν ποσό.
- Διάρκεια δανείου.
- Ημερομηνία γέννησης.

- Γενικά μηνιαία έξοδα.
- Κινητή περιουσία.
- Ακίνητη περιουσία.

β. Δάνεια αυτοκινήτου

Για την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας για δάνεια αυτοκινήτου τα δεδομένα που χρησιμοποιούνται είναι συνήθως αυτά που αναφέρθηκαν στην περίπτωση (α) κάποια στοιχεία επιπλέον όπως:

- Το ποσοστό προκαταβολής.
- Ο αριθμός αυτοκινήτων που έχει στην κατοχή του.
- Τα χρόνια κατοχής διπλώματος οδήγησης.

γ. Δάνεια μικρών και μεσαίων επιχειρήσεων

Για την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας για δάνεια μικρών επιχειρήσεων[3] τα δεδομένα που χρησιμοποιούνται συνήθως είναι τα εξής:

- Κύρια δραστηριότητα της επιχείρησης (υφαντουργική βιομηχανία, πώληση αυτοκινήτων, παραγωγή τροφίμων, ιατρικές ή πνευματικές υπηρεσίες, κατασκευαστική εταιρεία κ.λ.π.).
- Χρόνια λειτουργίας της επιχείρησης.
- Αριθμός εργαζομένων.
- Ηλικία του επιχειρηματία.
- Περιοχή όπου στεγάζεται η επιχείρηση.
- Αριθμός προηγούμενων αιτήσεων για πίστωση.
- Επιτόκιο.
- Είδος επιτοκίου (μηνιαίο, τριμηνιαίο, εξαμηνιαίο).
- Περίοδος χάριτος (ναι ή όχι).
- Περίοδος αποπληρωμής.
- Αιτηθέν ποσό δανείου.
- Κέρδος από επανεπένδυση.
- Ανταγωνιστικότητα επιχείρησης (κανένας ανταγωνισμός, ευρύς ανταγωνισμός, τοπικός ανταγωνισμός, καμία απάντηση).

[3] Βλέπε Zekic-Susac et al (2004).

- Σε ποιον τομέα είναι πιο ανταγωνιστική η επιχείρηση (ποιότητα, παραγωγή, τιμή, εξυπηρέτηση, φήμη).
- Επίπεδα πωλήσεων (πωλήσεις σε μια συγκεκριμένη περιοχή, προκαθορισμένοι πελάτες, σε ολόκληρη τη χώρα, καμία απάντηση).
- Διαφήμιση επιχείρησης (τοπική διαφήμιση, σε όλα τα μέσα επικοινωνίας, στο διαδίκτυο, καθόλου διαφήμιση).

δ. Δάνεια μεγάλων επιχειρήσεων

Για μεγάλες επιχειρήσεις τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας συνήθως χρησιμοποιούνται για να προβλεφτεί η πιθανότητα πτώχευσης της επιχείρησης. Για την εκτίμηση της πιστοληπτικής ικανότητας γίνεται αξιολόγηση όλων των στοιχείων που έχουν σχέση με τη δραστηριότητα και την προοπτική εξέλιξης της εταιρείας. Κάποια από τα στοιχεία αυτά είναι τα εξής:

- Ιστορικό επιχείρησης (ίδρυση, μεταβολές, αντικείμενο).
- Διοίκηση επιχείρησης (στελέχωση, αρμοδιότητες, οργάνωση λειτουργίας επιχείρησης κ.λ.π.).
- Επιχειρηματική δράση διοικούντων (διοικητικές ικανότητες, εμπειρία, αποτελεσματικότητα, αξιοπιστία κ.λ.π.).
- Φήμη της επιχείρησης στην αγορά.
- Λειτουργία του συναλλακτικού κυκλώματος της επιχείρησης (πελάτες, προμηθευτές, παραγωγικά μέσα, όροι συναλλαγών κ.λ.π.).
- Εξέλιξη των πωλήσεων της επιχείρησης.
- Εξέλιξη αποσβέσεων (αυξημένες, μειωμένες).
- Εξέλιξη του τελικού αποτελέσματος.
- Εξέλιξη τραπεζικού δανεισμού.
- Εξέλιξη των βασικών στοιχείων του ενεργητικού και του παθητικού χρησιμοποιώντας διάφορους χρηματοοικονομικούς δείκτες όπως αυτοί που δίνονται στον Πίνακα 2.1 σύμφωνα με τους Ζουπουνίδης και Λεμονάκης (2009).

Πίνακας 2.1 *Βασικοί χρηματοοικονομικοί δείκτες*

Δείκτης	**Τύπος υπολογισμού**
Μικτού Περιθωρίου Κέρδους	$\frac{\text{Πωλήσεις} - \text{Κόστος Πωληθέντων}}{\text{Πωλήσεις}}$
Καθαρού Περιθωρίου Κέρδους	$\frac{\text{Καθαρά Κέρδη}}{\text{Πωλήσεις}}$
Βιομηχανικής Αποδοτικότητας	$\frac{\text{Κέρδη προ τόκων και φόρων}}{\text{Σύνολο Ενεργητικού}}$
Χρηματοοικονομικής Αποδοτικότητας	$\frac{\text{Καθαρά Κέρδη}}{\text{Ίδια Κεφάλαια}}$
Ταχύτητας Κυκλοφορίας Ενεργητικού	$\frac{\text{Πωλήσεις}}{\text{Σύνολο Ενεργητικού}}$
Ταχύτητας Κυκλοφορίας Ιδίων Κεφαλαίων	$\frac{\text{Πωλήσεις}}{\text{Ίδια Κεφάλαια}}$
Ταχύτητας Κυκλοφορίας Παγίων	$\frac{\text{Πωλήσεις}}{\text{Πάγιο Ενεργητικό}}$
Ταχύτητας Κυκλοφορίας Βραχ. Υποχρεώσεων	$\frac{\text{Πωλήσεις}}{\text{Βραχυπρόθεσμες Υποχρεώσεις}}$
Γενικής Ρευστότητας	$\frac{\text{Κυκλοφορούν Ενεργητικό}}{\text{Βραχυπρόθεσμες Υποχρεώσεις}}$
Άμεσης Ρευστότητας	$\frac{\text{Κυκλοφορούν Ενεργητικό} - \text{Αποθέματα}}{\text{Βραχυπρόθεσμες Υποχρεώσεις}}$
Ανακύκλωσης Αποθεμάτων	$\frac{\text{Πωλήσεις}}{\text{Αποθέματα}}$
Ταχύτητας Κυκλοφορίας Απαιτήσεων	$\frac{\text{Πωλήσεις}}{\text{Απαιτήσεις}}$
Ταχύτητας Κυκλοφορίας Κεφαλαίου Κίνησης	$\frac{\text{Πωλήσεις}}{\text{Κυκλοφορούν Ενεργητικό} - \text{Βραχυπρόθεσμες Υποχρεώσεις}}$
Ικανότητας Δανεισμού	$\frac{\text{Ίδια Κεφάλαια}}{\text{Συνολικές Υποχρεώσεις}}$
Φερεγγυότητας	$\frac{\text{Συνολικές Υποχρεώσεις}}{\text{Σύνολο Ενεργητικού}}$

2.6 Ελλείπουσες τιμές και έκτροπες παρατηρήσεις

Ένα από τα πλέον συνηθισμένα προβλήματα κατά την ανάλυση συνόλων δεδομένων είναι η ύπαρξη **ελλειπουσών τιμών (*missing values*)** ή τιμών που δεν έχουν νόημα για ένα ιδιαίτερο χαρακτηριστικό και πιθανότατα προήλθαν από λανθασμένη πληκτρολόγηση (αυτές ονομάζονται συνήθως **έκτροπες παρατηρήσεις (*outliers*)**). Στη δεύτερη περίπτωση, αν δεν υπάρχει δυνατότητα ανάκτησης της σωστής τιμής, έχουμε ελλείπουσα τιμή στα δεδομένα.

Κάποιες στατιστικές τεχνικές όπως τα δέντρα ταξινόμησης δεν επηρεάζονται από την ύπαρξη ελλειπουσών τιμών, κάποιες άλλες όμως όπως η λογιστική παλινδρόμηση απαιτούν τα πλήρη σύνολα δεδομένων. Σύμφωνα με τον Siddiqi (2000) υπάρχουν τέσσερις κύριες μεθοδολογίες για να αντιμετωπιστούν οι ελλείπουσες τιμές:

- Να αποκλειστούν όλα τα άτομα στα άτομα τα οποία υπάρχει μια τουλάχιστον ελλείπουσα τιμή, έτσι ώστε να προκύψει πλήρες σύνολο δεδομένων. Όμως με αυτόν τον τρόπο, στις περισσότερες περιπτώσεις, θα απομείνουν τελικά πολύ λίγα στοιχεία για να αναλυθούν.
- Να αποκλειστούν από το μοντέλο τα χαρακτηριστικά που έχουν τις περισσότερες ελλείπουσες τιμές (π.χ. περισσότερο από 50%), ειδικά εάν σε αυτό το χαρακτηριστικό αναμένεται να συνεχιστεί η ύπαρξη ελλειπουσών τιμών στο μέλλον.
- Να συμπεριληφθούν τα χαρακτηριστικά με τις ελλείπουσες τιμές στο μοντέλο και η περίπτωση όπου λείπει το χαρακτηριστικό να αντιμετωπιστεί ως ξεχωριστή ιδιότητα της μεταβλητής. Με την ανάπτυξη του μοντέλου βαθμολόγησης μπορεί να επιτραπεί να δοθούν βάρη σε αυτήν την ιδιότητα. Σε μερικές περιπτώσεις, το βάρος μπορεί να είναι κοντά στη μέση αξία του χαρακτηριστικού, αλλά σε περιπτώσεις όπου το βάρος είναι πιο κοντά σε μια άλλη ιδιότητα, μπορεί να εξηγήσει την ακριβή φύση των ελλειπουσών τιμών γιατί συνήθως αυτές δεν εμφανίζονται τυχαία.
- Αντικαθιστώντας τις ελλείπουσες τιμές ή τις έκτροπες παρατηρήσεις με το μέσο όρο του χαρακτηριστικού αυτού.

Εφαρμόζοντας την $3^{η}$ μεθοδολογία προκύπτουν αρκετά οφέλη σε σχέση με τις υπόλοιπες τρεις οι οποίες αλλοιώνουν την πληροφορία των δεδομένων. Αξίζει να σημειωθεί ότι, στις περισσότερες περιπτώσεις, οι ελλείπουσες τιμές είναι μέρος μιας τάσης, που μπορεί να συνδέεται με άλλα χαρακτηριστικά ή είναι ένδειξη της κακής απόδοσης των πελατών και δεν είναι συνήθως τυχαίες. Για παράδειγμα, εκείνοι οι πελάτες που είναι νέοι στην εργασία τους, είναι πιθανότερο να αφήσουν κενό το πεδίο «αριθμό ετών στην εργασία» και αυτοί

αποτελούν πράγματι ομάδα σχετικά υψηλού κινδύνου. Επομένως, συστήνεται οι ελλείπουσες τιμές να συμπεριλαμβάνονται στην ανάλυση και το χαρακτηριστικό στο οποίο εμφανίζονται να συμμετέχει στο μοντέλο. Αυτή η μέθοδος αναγνωρίζει ότι οι ελλείπουσες τιμές προσφέρουν κάποια αξία και ότι υπάρχει επιχειρησιακό όφελος στη συμπερίληψη τέτοιων στοιχείων στην ανάλυση.

Συνήθως γίνεται μια προκαταρκτική ανάλυση στις ελλείπουσες τιμές και εάν αποδειχτεί ότι αυτές είναι τυχαίες και επομένως δε δίνουν κάποια σημαντική πληροφορία για την ανάπτυξη του μοντέλου οι καταγραφές που τα περιέχουν (πελάτες) αποκλείονται από τη στατιστική επεξεργασία.

Οι έκτροπες παρατηρήσεις που βρίσκονται εκτός των συνηθισμένων ορίων για ένα συγκεκριμένο χαρακτηριστικό. Για παράδειγμα, η συνηθέστερη ηλικία αιτούντων βρίσκονται μεταξύ 18 έως 55 ετών. Αν στο χαρακτηριστικό αυτό βρεθούν συμπληρωμένοι οι αριθμοί 99, 112 ή 200 το πιθανότερο είναι να έχει γίνει κάποιο λάθος στην πληκτρολόγηση των δεδομένων ή ακόμα και από τον ίδιο τον πελάτη στην συμπλήρωση της αίτησής του. Αυτοί οι αριθμοί μπορούν να έχουν δυσμενείς επιπτώσεις στα τελικά αποτελέσματα και οι καταγραφές που τα περιέχουν, συνήθως αποκλείονται. Σε μερικές περιπτώσεις, στη θέση των έκτροπων παρατηρήσεων μπορούν να εισαχθούν οι μέσες τιμές του χαρακτηριστικού. Σε κάθε περίπτωση, πριν οδηγηθούμε στην τελική απόφαση ως προς τη διαχείριση των έκτροπων παρατηρήσεων θα πρέπει πρώτα να ερευνηθεί προσεκτικά ο λόγος που εμφανίστηκαν γιατί σε κάποιες περιπτώσεις μπορεί να αναδείξουν σημαντικά προβλήματα όπως η απάτη εκ μέρους του αιτούντος.

ΚΕΦΑΛΑΙΟ 3

Μέθοδοι δημιουργίας μοντέλων βαθμολόγησης

3.1 Εισαγωγή

Η διαδικασία της παραγωγής των βαθμολογιών πιστοληπτικής ικανότητας με τη χρήση στατιστικών μοντέλων είναι το βασικότερο κομμάτι για τη δημιουργία ενός σκορόχαρτου. Κατά καιρούς έχουν εφαρμοστεί και έχουν προταθεί διάφορες μεθοδολογίες δημιουργίας τέτοιων μοντέλων. Όταν ξεκίνησε η χρήση των CSM για την εκτίμηση της πιθανότητας αθέτησης υποχρεώσεων υποψήφιων πελατών, οι μόνες μέθοδοι που εφαρμόζονταν για τη δόμηση αυτών ήταν στατιστικές μέθοδοι διαχωρισμού και ταξινόμησης ενώ αργότερα η ραγδαία εξέλιξη των υπολογιστικών συστημάτων επέτρεψε και την εφαρμογή μη στατιστικών μεθόδων. Ακόμα και σήμερα, οι στατιστικές μέθοδοι χρησιμοποιούνται συχνότερα διότι έχουν ως πλεονέκτημα ότι επιτρέπουν την εκμετάλλευση των πολύ καλών ιδιοτήτων των εκτιμητών, των διαστημάτων εμπιστοσύνης και των ελέγχων υποθέσεων που προκύπτουν στο πλαίσιο της ανάπτυξης των μοντέλων αυτών. Επομένως, με τα στατιστικά μοντέλα δίνεται η δυνατότητα να αξιολογηθεί η διαχωριστική δύναμη του κάθε σκορόχαρτου και η σημαντικότητα του κάθε χαρακτηριστικού (μεταβλητής) σε αυτόν το διαχωρισμό που δημιουργείται. Επίσης, αυτές οι πληροφορίες μπορούν να φανούν χρήσιμες για να προσδιοριστούν ποιες αλλαγές πρέπει να γίνουν στις ερωτήσεις που υποβάλλονται στους πελάτες του κάθε χρηματοπιστωτικού ιδρύματος.

Αν και οι στατιστικές μέθοδοι ήταν οι πρώτες που χρησιμοποιήθηκαν για να δομηθούν τα CSM και παραμένουν οι σημαντικότερες μέθοδοι, έχουν υπάρξει αρκετές αλλαγές στις μεθόδους που χρησιμοποιούνται τα τελευταία πενήντα χρόνια. Αρχικά, οι μέθοδοι βασίστηκαν στις μεθόδους διαχωρισμού που προτάθηκαν από το Fisher (1936) για τα γενικά

προβλήματα ταξινόμησης. Αυτό οδήγησε σε ένα γραμμικό μοντέλο βασισμένο στη γραμμική διαχωριστική συνάρτηση του Fisher. Όμως, οι προϋποθέσεις που απαιτούνταν για να εξασφαλίσουν ότι αυτός ήταν ο καλύτερος τρόπος να γίνει ο διαχωρισμός μεταξύ των «καλών» και των «κακών» πιθανών πελατών ήταν εξαιρετικά περιοριστικές και δεν μπορούσαν να εφαρμοστούν στην πράξη, αν και τα σκορόχαρτα που παράγονταν ήταν πολύ ανθεκτικά. Η προσέγγιση του Fisher θα μπορούσε να αντιμετωπισθεί σαν μια μορφή γραμμικής παλινδρόμησης και αυτό οδήγησε σε μια έρευνα για βέλτιστες μορφές παλινδρόμησης που έχουν λιγότερο περιοριστικές υποθέσεις και όχι απλά διαχωρίζουν τους πελάτες σε «καλούς» και «κακούς» αλλά μπορούν και να τους ταξινομήσουν ως προς την πιστοληπτική τους ικανότητα με βάση τις βαθμολογίες που τους αντιστοιχούν. Η καλύτερη μορφή παλινδρόμησης που εφαρμόζεται πολύ συχνά ακόμα και σήμερα είναι η λογιστική παλινδρόμηση.

Μια άλλη προσέγγιση που χρησιμοποιείται ευρέως τα τελευταία 30 χρόνια είναι τα δέντρα ταξινόμησης. Με αυτή τη μέθοδο, το σύνολο των υποψηφίων χωρίζεται σε διάφορες υποομάδες ανάλογα με τις ιδιότητές τους και έπειτα κάθε υποομάδα ταξινομείται ως ικανοποιητική ή ανεπαρκής. Αν και αυτή η μέθοδος δε δίνει βάρος σε κάθε ένα από τα χαρακτηριστικά όπως γίνεται με τα γραμμικά μοντέλα, το αποτέλεσμα που προκύπτει είναι το ίδιο, δηλαδή ο χαρακτηρισμός ενός νέου υποψηφίου ως «καλός» ή «κακός». Μια ακόμα στατιστική μέθοδος που χρησιμοποιείται πολύ συχνά είναι μία μη παραμετρική μέθοδος και ονομάζεται μέθοδος του κοντινότερου γείτονα. Όλες αυτές οι μέθοδοι έχουν χρησιμοποιηθεί ευρέως στην πράξη για τη δόμηση μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας αλλά στη βιβλιογραφία υπάρχει πολύ μεγάλη ποικιλία σε στατιστικές μεθόδους για τη δημιουργία σκορόχαρτων.

Όμως, ο στόχος για οποιαδήποτε μέθοδο ανάπτυξης CSM, είναι να μπορεί αυτή να αντιμετωπίσει την πολύ μεγάλη ποσότητα των πληροφοριών που προέρχεται από τους προηγούμενους πελάτες και να ταξινομήσει τους νέους σε εκείνους που το χρηματοπιστωτικό ίδρυμα θα τους δεχτεί σαν πελάτες ή θα τους απορρίψει λόγω της εκτιμώμενης μελλοντικής συμπεριφοράς τους.

Στην παρούσα εργασία κυρίως περιγράφονται οι μεθοδολογίες ανάπτυξης ASM. Γι’ αυτού του είδους τα CSM αρχικά λαμβάνεται ένα δείγμα από προηγούμενους πελάτες και μελετάται η συμπεριφορά τους για ένα χρονικό διάστημα στο πρόσφατο παρελθόν. Συνήθως μελετώνται 50 έως 100 χαρακτηριστικά αυτών. Επιπλέον, ορίζεται και μια δυαδική μεταβλητή για κάθε

πελάτη στο δείγμα που δείχνει εάν η απόδοση του οφειλέτη κατά τη διάρκεια αυτού του χρονικού διαστήματος ήταν ικανοποιητική ή όχι (ήταν «καλός» ή «κακός»).

Συνοψίζοντας, οι πιο συνηθισμένες στατιστικές μέθοδοι που χρησιμοποιούνται για την ανάπτυξη ASM είναι **η διαχωριστική ανάλυση (*discriminant analysis*), η γραμμική παλινδρόμηση (*linear regression*)**, **η λογιστική παλινδρόμηση (*logistic regression*)**, τα **δέντρα ταξινόμησης (*classification trees*)** και η μέθοδος του **κ-κοντινότερου γείτονα (*k-nearest neighbour*)**. Υπάρχουν όμως και μη στατιστικές τεχνικές που χρησιμοποιούνται γι' αυτόν το σκοπό όπως τα **νευρωνικά δίκτυα (*neural networks*)**, ο **γραμμικός προγραμματισμός (*linear programming*)** και τα **έμπειρα συστήματα (*expert systems*)** αλλά δε θα ασχοληθούμε με αυτές στην παρούσα εργασία.

Σύμφωνα με τον Siddiqi (2000) η επιλογή της πιο κατάλληλης μεθόδου για την ανάπτυξη ενός μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας εξαρτάται από ζητήματα όπως:

- Η ποιότητα των διαθέσιμων δεδομένων. Για παράδειγμα, ένα δέντρο ταξινόμησης μπορεί να είναι πιο κατάλληλο για τις περιπτώσεις που υπάρχουν αρκετές ελλείπουσες τιμές ή σε περίπτωση πού η σχέση μεταξύ των χαρακτηριστικών και των μεταβλητών απόκρισης είναι μη γραμμική.
- Ο τύπος της μεταβλητής απόκρισης, δηλαδή αν έχουμε δυαδική μεταβλητή απόκρισης που παίρνει τιμές 0 ή 1 (καλός ή κακός) ή συνεχής (κέρδος ή απώλεια).
- Το μέγεθος δείγματος που είναι διαθέσιμο.
- Η ερμηνευσιμότητα των αποτελεσμάτων.
- Η νομική συμμόρφωση στη μεθοδολογία που χρησιμοποιείται από κάθε χρηματοπιστωτικό οργανισμό έτσι ώστε αυτή να είναι διαφανής και επεξηγήσιμη.
- Να υπάρχει η δυνατότητα να μετρηθεί η απόδοση των CSM που αναπτύσσονται.

Πριν δούμε αναλυτικά τις μεθοδολογίες ανάπτυξης των CSM θα περιγράψουμε κάποιες διαδικασίες που είναι απαραίτητες να γίνουν πριν ξεκινήσει η δόμηση αυτών των μοντέλων.

3.2 Συμβολισμοί που χρησιμοποιούνται για την περιγραφή των μεθόδων

Έστω $\mathbf{X}=(X_1,X_2,...,X_p)$ ένα τυχαίο διάνυσμα που αποτελείται από p τυχαίες μεταβλητές κάθε μια από τις οποίες περιγράφει τις πληροφορίες που έχουν αντληθεί από τον κάθε υποψήφιο πελάτη είτε από τα στοιχεία που έχουν συμπληρωθεί στην αίτηση χορήγησης δανείου είτε από κάποιο πιστωτικό γραφείο. Τα $X_1,X_2,...,X_p$ τα ονομάζουμε χαρακτηριστικά ή μεταβλητές. Οι τιμές που παίρνουν αυτά τα χαρακτηριστικά συμβολίζονται αντίστοιχα με $\mathbf{x}=(x_1,x_2,...,x_p)$ και καλούνται ιδιότητες των χαρακτηριστικών.

Αν n είναι το μέγεθος δείγματος θα μπορούσαμε να οργανώσουμε τα δεδομένα σε έναν πίνακα $n\times p$ ως εξής:

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}_{n\times p}$$

όπου x_{ij} είναι η τιμή της j μεταβλητής για την i παρατήρηση.

Υπάρχουν τρεις τρόποι για να περιγραφεί η πιθανότητα να είναι «καλός» ένας οφειλέτης. Ο πρώτος τρόπος είναι ο απευθείας καθορισμός της πιθανότητας να συμβεί το συγκεκριμένο γεγονός, ο δεύτερος είναι να υπολογιστεί η σχετική πιθανότητα (*odds*) αυτού του γεγονότος και ο τρίτος είναι να βρεθεί μια βαθμολογία (*score*) ή ένας δείκτης, στα οποία να περιέχονται όλες οι πληροφορίες που απαιτούνται για να υπολογιστεί η πιθανότητα που μας ενδιαφέρει. Στη συνέχεια θα δούμε πως συνδέονται μεταξύ τους οι τρεις αυτοί τρόποι.

Υποθέτουμε ότι υπάρχει ένας πεπερασμένος αριθμός διαφορετικών ιδιοτήτων $\mathbf{x}$, δηλαδή υπάρχει ένας πεπερασμένος αριθμός τρόπων για τη συμπλήρωση της αίτησης. Έχουμε επιπλέον και τους εξής συμβολισμούς:

p_G : το ποσοστό των «καλών» πελατών στον πληθυσμό (εκ των προτέρων πιθανότητα των «καλών» πελατών)

p_B: το ποσοστό των «κακών» πελατών στον πληθυσμό (εκ των προτέρων πιθανότητα των «κακών» πελατών)

$P(G|\mathbf{x})$: η πιθανότητα κάποιος πελάτης που έχει τις ιδιότητες $\mathbf{x}$ να είναι «καλός»

$P(B|\mathbf{x})$: η πιθανότητα κάποιος πελάτης που έχει τις ιδιότητες $\mathbf{x}$ να είναι «κακός»

Προφανώς,

$$p_G + p_B = 1 \quad \text{και} \quad P(G|\mathbf{x}) + P(B|\mathbf{x}) = 1.$$

Στην πράξη, τις περισσότερες φορές είναι ευκολότερο να εξεταστεί η **σχετική πιθανότητα (*odds*)** δηλαδή η πιθανότητα να συμβεί ένα γεγονός διαιρεμένο με την πιθανότητα να μη συμβεί. Με τη βοήθεια της σχετικής πιθανότητας των «καλών» (ή «κακών») μπορεί να υπολογιστεί η πιθανότητα ένας πελάτης να είναι «καλός» (ή «κακός»).

Η σχετική πιθανότητα των «καλών» ορίζεται ως η πιθανότητα ένας πελάτης με τις ιδιότητες $\mathbf{x}$ να είναι «καλός» προς την πιθανότητα να είναι «κακός», ενώ η σχετική πιθανότητα των «κακών» ορίζεται ως η πιθανότητα ένας πελάτης με τις ιδιότητες $\mathbf{x}$ να είναι «κακός» προς την πιθανότητα να είναι «καλός», δηλαδή

$$odds(G|\mathbf{x}) = \frac{P(G|\mathbf{x})}{P(B|\mathbf{x})}$$

$$odds(B|\mathbf{x}) = \frac{P(B|\mathbf{x})}{P(G|\mathbf{x})} = \frac{1}{odds(G|\mathbf{x})}.$$

Με τη χρήση του κανόνα του Bayes μπορούμε να πάρουμε πληροφορίες για την κατανομή των χαρακτηριστικών των υποψήφιων πελατών μεταξύ των «καλών» και των «κακών» εάν γνωρίζουμε τις πιθανότητες p_G και p_B. Πιο συγκεκριμένα συμβολίζοντας με $P(\mathbf{x}|G)$ την πιθανότητα να έχει τις ιδιότητες $\mathbf{x}$ ένας «καλός» πελάτης μπορούμε να γράψουμε:

$$P(\mathbf{x}|G) = f(\mathbf{x}|G) = \frac{P(\text{ο πελάτης έχει ιδιότητες } \mathbf{x} \text{ και είναι καλός})}{P(\text{ο πελάτης είναι καλός})} = \frac{P(G|\mathbf{x})P(\mathbf{x})}{p_G}. \quad (3.1)$$

Όμοια, για την πιθανότητα να έχει τις ιδιότητες $\mathbf{x}$ ένας «κακός» πελάτης παίρνουμε:

$$P(\mathbf{x}|B) = f(\mathbf{x}|B) = \frac{P(\text{ο πελάτης έχει ιδιότητες } \mathbf{x} \text{ και είναι κακός})}{P(\text{ο πελάτης είναι κακός})} = \frac{P(B|\mathbf{x})P(\mathbf{x})}{p_B}. \quad (3.2)$$

όπου $P(\mathbf{x})$ είναι η πιθανότητα ένας πελάτης να έχει τις ιδιότητες $\mathbf{x}$ και οι $f(\mathbf{x}|G)$ και $f(\mathbf{x}|B)$ είναι οι συναρτήσεις πυκνότητας.

Η **σχετική πιθανότητα του πληθυσμού (*population odds*)** ορίζεται ως εξής:

$$odds(pop) = \frac{p_G}{p_B}.$$

Η σχετική πιθανότητα του πληθυσμού αντικατοπτρίζει την αρχική πεποίθηση για την πιθανότητα ένας πελάτης να είναι «καλός» ή «κακός» πριν ακόμα υπάρξει κάποια πληροφορία για αυτόν.

Όταν η πληροφορία $\mathbf{x}$ είναι διαθέσιμη τότε η οι εκ των υστέρων πιθανότητες των «καλών» και των «κακών» μπορούν να δοθούν από τις ποσότητες $P(G|\mathbf{x})$ και $P(B|\mathbf{x})$. Διαιρώντας τις σχέσεις (3.1) και (3.2) θα πάρουμε την εκ των υστέρων σχετική πιθανότητα ένας πελάτης να είναι «καλός» ως προς το να είναι «κακός». Έχουμε ότι:

$$odds(G|\mathbf{x}) = \frac{P(G|\mathbf{x})}{P(B|\mathbf{x})} = \frac{f(\mathbf{x}|G)p_G}{f(\mathbf{x}|B)p_B} = \frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)} \cdot \frac{p_G}{p_B} \equiv I(\mathbf{x}) \cdot odds(pop) \qquad (3.3)$$

όπου $I(\mathbf{x}) = \frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)}$ και ονομάζεται **πληροφοριακή σχετική πιθανότητα (*information odds*)**.

Η σχέση (3.3) δείχνει ότι η σχετική πιθανότητα ενός πελάτη που έχει τις ιδιότητες $\mathbf{x}$ να είναι «καλός» είναι το γινόμενο της σχετικής πιθανότητας του πληθυσμού, η οποία είναι ανεξάρτητη από τις ιδιότητες που έχει ο κάθε πελάτης και της πληροφοριακής σχετικής πιθανότητας, που εξαρτάται από τις ιδιότητες του κάθε πελάτη. Όταν ο πληροφοριακή σχετική πιθανότητα $I(\mathbf{x})$ είναι μεγαλύτερη από τη μονάδα τότε οι υποψήφιοι δανειζόμενοι που έχουν αυτές τις ιδιότητες είναι πιο πιθανό να είναι καλοί σε σχέση με το γενικό πληθυσμό.

Μία βαθμολογία $s(\mathbf{x})$ είναι μια συνάρτηση των ιδιοτήτων των χαρακτηριστικών $\mathbf{x}$ ενός υποψήφιου πελάτη, η οποία κατά κάποιο τρόπο αντιπροσωπεύει την πιθανότητα ο πελάτης να είναι «καλός». Η βαθμολογία είναι μια επαρκής στατιστική συνάρτηση και θεωρείται ότι έχει μια αύξουσα μονότονη σχέση με την πιθανότητα ένας πελάτης να είναι «καλός» και γι' αυτόν το λόγο η βαθμολογία ονομάζεται και μονότονη βαθμολογία (*monotonic score*). Αυτές οι βαθμολογίες αποσκοπούν στη διάταξη των πελατών που σημαίνει ότι ένας πελάτης που έχει μεγαλύτερη βαθμολογία έχει ταυτόχρονα και μεγαλύτερη πιθανότητα να αποδειχτεί «καλός» σε σχέση με έναν πελάτη με μικρότερη βαθμολογία. Υποθέτουμε ότι $s(\mathbf{x})$ είναι η καλύτερη δυνατή βαθμολογία η οποία συμπεριλαμβάνει όλη την πληροφορία που απαιτείται για την πρόβλεψη της απόδοσης του πελάτη με ιδιότητες $\mathbf{x}$. Αυτό σημαίνει ότι αν $P(G|s(\mathbf{x}))$ είναι η πιθανότητα να είναι «καλός» ο πελάτης που έχει βαθμολογία $s(\mathbf{x})$ τότε ισχύει:

$$P(G \mid s(\mathbf{x})) = P(G \mid s(\mathbf{x}), \mathbf{x}) = P(G \mid \mathbf{x}).$$

Επομένως, με μια κατάλληλη βαθμολογία μπορούν να αντικατασταθούν τα πολυδιάστατα χαρακτηριστικά $\mathbf{x}$ περιγράφοντας έναν πελάτη με μία μόνο τιμή $s(\mathbf{x})$ χωρίς να επηρεάζεται η πιθανότητα ο πελάτης με τις ιδιότητες $\mathbf{x}$ να είναι «καλός». Αν χρησιμοποιήσουμε ένα διαφορετικό σύνολο μεταβλητών με διαφορετικό συνδυασμό ιδιοτήτων $\mathbf{y}$ μπορούμε να δημιουργήσουμε μια άλλη κατάλληλη βαθμολογία $s(\mathbf{y})$ [4]. Γενικά συμβολίζουμε:

$$P(s) = P(G \mid s(\mathbf{x}))$$

και

$$1 - P(s) = 1 - P(G \mid s(\mathbf{x})) = P(B \mid s(\mathbf{x})).$$

Οι συναρτήσεις πιθανοφάνειας συμβολίζονται με $f(s \mid G)$, $f(s \mid B)$ για τους «καλούς» και τους «κακούς» αντίστοιχα και

$$f(s) = f(\mathbf{x} \mid s(\mathbf{x}) = s).$$

Αν $s(\mathbf{x})$ είναι η καλύτερη δυνατή βαθμολογία τότε η σχέση (3.3) μπορεί να γραφεί

$$odds(G \mid s) = \frac{f(s \mid G)}{f(s \mid B)} \cdot \frac{p_G}{p_B} = I(s) \cdot odds(pop).$$

όπου $I(s)$ είναι ο πληροφοριακός δείκτης πιθανοτήτων που παράγεται από τη βαθμολογία s.

Μια άλλη πολύ σημαντική συνάρτηση της βαθμολογίας που χρησιμοποιείται στα CSM είναι η βαθμολογία του νεπέριου **λογαρίθμου της σχετικής πιθανότητας (*log odds score*)** η οποία ορίζεται ως εξής:

$$s(\mathbf{x}) = \ln(odds(G \mid \mathbf{x})) = \ln\left(\frac{P(G \mid \mathbf{x})}{P(B \mid \mathbf{x})}\right) \tag{3.4}$$

Προφανώς θα ισχύει

$$odds(G \mid \mathbf{x}) = e^{s(\mathbf{x})}. \tag{3.5}$$

Ο λογάριθμος της σχετικής πιθανότητας έχει την ιδιότητα να «μεγαλώνει» την κλίμακα μέτρησης για τις εξαιρετικά μικρές ή μεγάλες πιθανότητες. Η έμφαση στα σπάνια γεγονότα και τις πολύ μικρές πιθανότητες είναι σημαντική δεδομένου ότι η πιθανότητα της αθέτησης υποχρεώσεων για ένα μεμονωμένο οφειλέτη είναι μικρή αλλά οι συνέπειες για τα περισσότερα χαρτοφυλάκια τραπεζών είναι πολύ μεγάλη.

[4] Thomas (2009).

Ο προσδιορισμός της σχετικής πιθανότητας ή της βαθμολογίας ενός γεγονότος είναι ισοδύναμος με τον προσδιορισμό της απλής πιθανότητάς εμφάνισης του γεγονότος αφού η άλλη πιθανότητα μπορεί να γραφεί συναρτήσει της σχετικής πιθανότητας ως εξής:

$$P(G\,|\,\mathbf{x}) = \frac{odds(G\,|\,\mathbf{x})}{1+odds(G\,|\,\mathbf{x})}, \quad P\left(B|\mathbf{x}\right) = \frac{1}{1+odds(B\,|\,\mathbf{x})}.$$

Επιπλέον, χρησιμοποιώντας τη σχέση (3.5) η πιθανότητα μπορεί να γραφεί και συναρτήσει της βαθμολογίας $s(\mathbf{x})$ ως εξής:

$$P(G\,|\,\mathbf{x}) = \frac{odds(G\,|\,\mathbf{x})}{1+odds(G\,|\,\mathbf{x})} = \frac{e^{s(\mathbf{x})}}{1+e^{s(\mathbf{x})}} = \frac{1}{1+e^{-s(\mathbf{x})}}.$$

Το πιο σημαντικό θεωρητικό χαρακτηριστικό γνώρισμα της βαθμολογίας του λογαρίθμου της σχετικής πιθανότητας είναι ότι διαχωρίζει τελείως την πληροφορία που προκύπτει από τον πληθυσμό από την πληροφορία του μεμονωμένου πελάτη που πρόκειται να βαθμολογηθεί. Αυτό συμβαίνει γιατί αν πάρουμε τη σχέση (3.4) έχουμε:

$$s(\mathbf{x}) = \ln(odds(G\,|\,\mathbf{x})) = \ln\left(\frac{P(G\,|\,\mathbf{x})}{P(B\,|\,\mathbf{x})}\right) = \ln\left(\frac{p_G f(\mathbf{x}\,|\,G)}{p_B f(\mathbf{x}\,|\,B)}\right) = \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{f(\mathbf{x}\,|\,G)}{f(\mathbf{x}\,|\,B)}\right) = \tag{3.6}$$

$$= \ln(odds(pop)) + \ln(I(\mathbf{x})) = s(pop) + s_I(\mathbf{x})$$

όπου θέσαμε $s_I(\mathbf{x}) = \ln(I(\mathbf{x}))$.

Η σχέση (3.6) δείχνει ότι η βαθμολογία του λογαρίθμου της σχετικής πιθανότητας είναι το άθροισμα του όρου $s(pop)$ που εξαρτάται μόνο από τη σχετική πιθανότητα του πληθυσμού και του όρου $s_I(\mathbf{x})$ που εξαρτάται από τη πληροφορία που προσφέρει ο πελάτης με τις ιδιότητες $\mathbf{x}$. Ο όρος $s(pop)$ είναι μια βαθμολογία εκ των προτέρων που σημαίνει ότι αποτελεί τη βαθμολογία ενός τυχαία επιλεγμένου υποψηφίου από τον πληθυσμό. Ο όρος $s_I(\mathbf{x})$ είναι ο λογάριθμος της πληροφοριακής σχετικής πιθανότητας και ονομάζεται **βάρος ένδειξης (*weights of evidence*)** της πληροφορίας που προέρχεται από τις ιδιότητες $\mathbf{x}$ και μπορεί να συμβολιστεί επίσης και ως $w(\mathbf{x})$. Το βάρος ένδειξης είναι:

$$w(\mathbf{x}) = \ln(I(\mathbf{x})) = \ln\left(\frac{f(\mathbf{x}\,|\,G)}{f(\mathbf{x}\,|\,B)}\right) = \ln\left(\frac{\frac{P(G|\mathbf{x})P(\mathbf{x})}{p_G}}{\frac{P(B|\mathbf{x})P(\mathbf{x})}{p_B}}\right) = \ln\left(\frac{P(G|\mathbf{x})/P(B|\mathbf{x})}{p_G/p_B}\right). \tag{3.7}$$

Με τον ίδιο τρόπο μπορούμε να γράψουμε και το βάρος ένδειξης μιας βαθμολογίας s:

$$w(s) = \ln(I(s)) = \ln\left(\frac{f(s|G)}{f(s|B)}\right) = \ln\left(\frac{\frac{P(G|s)P(s)}{p_G}}{\frac{P(B|s)P(s)}{p_B}}\right) = \ln\left(\frac{P(G|s)/P(B|s)}{p_G/p_B}\right).$$

Τα βάρη ένδειξης είναι ο λογάριθμος του λόγου της σχετικής πιθανότητας των «καλών» που έχουν ιδιότητες **x** ή βαθμολογία s συγκρινόμενη με τη σχετική πιθανότητα του πληθυσμού. Εάν το βάρος ένδειξης είναι θετικό τότε θεωρείται ότι οι πελάτες που έχουν ιδιότητες **x** θα παρουσιάσουν καλύτερη απόδοση σε σχέση με το συνολικό πληθυσμό, ενώ αν είναι αρνητικό τότε πελάτες που έχουν ιδιότητες **x** θα παρουσιάσουν χειρότερη απόδοση.

3.3 Συμπερασματολογία απορριφθέντων

Ένα πρόβλημα που παρουσιάζεται κατά τη διαδικασία της ανάπτυξης ενός CSM είναι ότι στην πράξη, το δείγμα ανάπτυξης που χρησιμοποιείται για να κατασκευαστεί το μοντέλο σπάνια αποτελεί ένα τυχαίο δείγμα από ολόκληρο τον πληθυσμό. Στην πραγματικότητα είναι το σύνολο πελατών που χαρακτηρίστηκαν ως «καλοί» στο παρελθόν σύμφωνα με ένα προηγούμενο μοντέλο. Σε εκείνους τους πελάτες που απορρίφθηκαν δεν χορηγήθηκε η πίστωση και έτσι αυτοί δε μελετήθηκαν για να καθοριστεί ο πραγματικός τους κίνδυνος αθέτησης υποχρεώσεων και δεν είχαν ποτέ την ευκαιρία να αποδείξουν την πραγματική τους συμπεριφορά. Τα μόνα στοιχεία που είναι γνωστά για αυτούς τους απορριφθέντες πελάτες είναι αυτά που είχαν συμπληρώσει στις αιτήσεις τους ή ίσως και κάποιες συμπληρωματικές πληροφορίες που αφορούσαν προηγούμενη απόδοση αποπληρωμής κάποιου δανείου. Αυτή η διαστρέβλωση της κατανομής των υποψηφίων προφανώς έχει επιπτώσεις στην απόδοση οποιουδήποτε CSM που κατασκευάζεται. Η κατασκευή ενός μοντέλου χρησιμοποιώντας μόνο τις πληροφορίες των πελατών που έχουν γίνει αποδεκτοί οδηγεί συνήθως σε λανθασμένα αποτελέσματα γιατί πρώτον, υπάρχει περίπτωση οι εκτιμήσεις των παραμέτρων ενός τέτοιου μοντέλου να είναι μεροληπτικές για τον προοριζόμενο πληθυσμό και δεύτερον, χωρίς γνώση του πραγματικού ποσοστού των «καλών» και των «κακών» στον πληθυσμό των παλαιών πελατών δεν είναι δυνατόν να γίνει γνωστή η προβλεπτική απόδοση του νέου μοντέλου σε αυτόν τον πληθυσμό.

Οι εταιρείες πιστωτικού ελέγχου αντιμετωπίζουν αυτό το πρόβλημα χρησιμοποιώντας διάφορες τεχνικές οι οποίες καλούνται **συμπερασματολογία απορριφθέντων (*reject inference*)**. Αυτές οι τεχνικές χρησιμοποιούν τις πληροφορίες που υπάρχουν για τους απορριφθέντες πελάτες από τις αιτήσεις που είχαν συμπληρώσει στο παρελθόν, δηλαδή είναι γνωστές μόνο οι τιμές των χαρακτηριστικών τους αλλά όχι η κατηγορία στην οποία πραγματικά ανήκουν. Σε γενικές γραμμές, με τη συμπερασματολογία απορριφθέντων γίνεται προσπάθεια να προκύψει η πιθανή αληθινή κατηγορία των πελατών που είχαν απορριφθεί στο παρελθόν και έπειτα να κατασκευαστεί ένα νέο CSM χρησιμοποιώντας και τις πληροφορίες των απορριφθέντων πελατών.[5] Ωστόσο, οι τεχνικές της συμπερασματολογίας απορριφθέντων δεν εφαρμόζεται σε περιπτώσεις βαθμολόγησης συμπεριφοράς διότι τα δεδομένα όλων των πελατών είναι διαθέσιμα.

Μπορούμε να διακρίνουμε διάφορες περιπτώσεις των τεχνικών συμπερασματολογίας απορριφθέντων[6]. Η πιο απλή προσέγγιση είναι όλοι οι υποψήφιοι πελάτες που είχαν απορριφθεί λόγω αρνητικών πληροφοριών να χαρακτηριστούν ως «κακοί». Έπειτα δομείται ένα CSM βασισμένο στους απορριφθέντες πελάτες και σε αυτούς που είχαν γίνει αποδεκτοί. Χρησιμοποιώντας αυτή την τεχνική προκύπτουν κάποια προβλήματα όπως είναι η ενίσχυση της προκατάληψης για τους απορριφθέντες πελάτες και δεν διορθώνονται οι κακές αποφάσεις του παρελθόντος. Αυτή η τεχνική είναι λάθος και από στατιστικής αλλά και από ηθικής άποψης διότι όταν κάποιοι υποψήφιοι πελάτες απορρίπτονται και ταξινομούνται ως «κακοί» δεν έχουν την ευκαιρία να αποδείξουν σε ποια κατηγορία ανήκουν πραγματικά.

Οι Hand and Henley (1993), σε μια ανάλυση για τη συμπερασματολογία απορριφθέντων επισήμαναν ότι υπάρχουν δύο διαφορετικές περιπτώσεις που εξαρτώνται από τη σχέση μεταξύ των χαρακτηριστικών $\mathbf{X}_1$ που χρησιμοποιήθηκαν στο αρχικό μοντέλο (το μοντέλο σύμφωνα με το οποίο έγιναν αποδεχτοί ή απορρίφθηκαν οι παλιοί πελάτες) και των χαρακτηριστικών $\mathbf{X}_2$ που χρησιμοποιούνται για να δομηθεί το νέο σκορόχαρτο. Εάν οι επεξηγηματικές μεταβλητές $\mathbf{X}_1$ είναι υποσύνολο των $\mathbf{X}_2$, δηλαδή τα νέα χαρακτηριστικά περιλαμβάνουν όλα όσα χρησιμοποιήθηκαν στο αρχικό CSM, τότε δε γνωρίζουμε τίποτα για την απόδοση των πελατών που έχουν κάποιους συνδυασμούς ιδιοτήτων (για αυτούς που οι υποψήφιοι πελάτες απορρίφθηκαν), αφού όλοι οι πελάτες αυτοί απορρίφθηκαν. Όμως, για τους υπόλοιπους συνδυασμούς ιδιοτήτων των $\mathbf{X}_1$, για τους οποίους οι υποψήφιοι πελάτες

[5] Βλέπε Hand and Henley (1997).
[6] Βλέπε Thomas et al (2002).

έγιναν αποδεκτοί, παρέχεται η πληροφορία για την αναλογία «καλών»:«κακών». Κατόπιν, θα πρέπει να προχωρήσουμε στη μέθοδο της ***extrapolation***, δηλαδή θα πρέπει να δομηθεί ένα μοντέλο με βάση τους πελάτες που είχαν γίνει αποδεκτοί και έπειτα να εφαρμοστεί αυτό το μοντέλο χρησιμοποιώντας τις πληροφορίες αυτών που είχαν απορριφθεί. Με την τεχνική αυτή το μοντέλο προβλέπει την πιθανότητα ο πελάτης να είναι «καλός» σε κάθε έναν από τους πελάτες που είχαν απορριφθεί και έτσι το νέο σκορόχαρτο θα δομηθεί σε ολόκληρο τον πληθυσμό αφού η βαθμολογία των απορριφθέντων είναι πλέον γνωστή. Όπως είναι αναμενόμενο, το νέο CSM θα είναι καλύτερο σε σχέση με ένα μοντέλο που στηρίζεται μόνο σε εκείνους που είχαν γίνει αποδεκτοί.

Υπάρχει όμως και η άλλη περίπτωση όταν τα χαρακτηριστικά $\mathbf{X}_1$ δεν είναι υποσύνολο των $\mathbf{X}_2$. Σε αυτήν την περίπτωση, υπάρχουν άγνωστες μεταβλητές (ή λόγοι) με βάση τους οποίους απορρίφθηκαν οι αρχικοί πελάτες. Δηλαδή, η παρατηρηθείσα κατανομή των «καλών» και «κακών» πελατών για τους πελάτες που είχαν γίνει αποδεκτοί δεν είναι αντιπροσωπευτική της πραγματικής κατανομής και η περίπτωση αυτή είναι πιο περίπλοκη. Σε αυτήν την περίπτωση, μια ευρέως χρησιμοποιούμενη τεχνική είναι μια τεχνική ανάθεσης βαρών που είναι γνωστή ως **αύξηση (*augmentation*)**, η οποία περιγράφηκε από τον Hsia (1978). Έστω ότι A είναι το ενδεχόμενο να γίνει αποδεκτός ένας πελάτης και R αυτός να απορριφθεί. Πρώτα, κατασκευάζεται ένα CSM χρησιμοποιώντας μόνο τους πελάτες που είχαν γίνει αποδεκτοί για να εκτιμηθεί η ποσότητα $\mathrm{P}(\mathrm{G}\,|\,\mathbf{x}, A)$, η πιθανότητα ένας πελάτης να είναι «καλός» δεδομένου ότι έχει γίνει αποδεκτός και οι τιμές των χαρακτηριστικών του είναι $\mathbf{x}$. Έπειτα χρησιμοποιώντας τους απορριφθέντες πελάτες κατασκευάζεται ένα μοντέλο αποδοχής – απόρριψης με την ίδια μέθοδο για να βρεθεί η πιθανότητα

$$P(A\,|\,\mathbf{x}) = P(A\,|\,s(\mathbf{x})) = P(A\,|\,s),$$

όπου $s(\mathbf{x})$ είναι η βαθμολογία για τις τιμές $\mathbf{x}$ των χαρακτηριστικών που προκύπτουν από το αναπτυχθέν μοντέλο αποδοσής - απόρριψης. Η προσέγγιση του Hsia (1978) υποθέτει ότι η πιθανότητα ένας πελάτης να είναι «καλός» είναι ίδια ανάμεσα σε αυτούς που είναι αποδεχτοί και σε αυτούς που έχουν απορριφτεί, δηλαδή $P(G\,|\,s,A) = P(G\,|\,s,R)$, όπου

$$P(G\,|\,s,A) = \sum_{\mathbf{x};s(\mathbf{x})=0} P(G\,|\,\mathbf{x},A)P(\mathbf{x}\,|\,s(\mathbf{x}) = \mathbf{x}).$$

Αυτή η τεχνική είναι σαν να ξανά αναθέτονται βάρη στην κατανομή των δειγματικών πληθυσμών έτσι ώστε το ποσοστό των πελατών που είχαν βαθμολογία s και έγιναν

αποδεκτοί μετατράπηκε από $P(A,s)$ σε $P(s)$. Επομένως, μπορεί να κατασκευαστεί ένα νέο μοντέλο περιλαμβάνοντας και αυτούς που είχαν απορριφθεί. Δηλαδή, οι απορριφθέντες που είχαν βαθμολογία s έχουν πιθανότητα $P(G|s,A)$ να είναι «καλοί».

Κάποια βελτιωμένα μοντέλα βαθμολόγησης θα μπορούσαν να προκύψουν εάν ήταν διαθέσιμες οι πληροφορίες από κάποιους απορριφθέντες πελάτες, δηλαδή εάν μερικοί υποψήφιοι που θα απορρίπτονταν κανονικά, γινόντουσαν αποδεκτοί για μια μικρή χρονική περίοδο. Αυτό θα ήταν εμπορικά λογικό εάν η απώλεια λόγω του αυξανόμενου αριθμού «κακών» λογαριασμών αντισταθμίζονταν από την αυξανόμενη ακρίβεια στην ταξινόμηση των δανειοληπτών. Αλλά, αυτή η πρακτική εφαρμόζεται σε σπάνιες περιπτώσεις όπως για παράδειγμα από κάποιους οργανισμούς ή εταιρείες πωλήσεων μέσω ταχυδρομείου. Ωστόσο, μια πρακτική που χρησιμοποιείται όλο ένα και περισσότερο είναι να λαμβάνονται οι πληροφορίες για τους απορριφθέντες υποψηφίους από άλλους πιστωτικούς προμηθευτές που τους είχαν χορήγησαν πίστωση και διαθέτουν περισσότερες πληροφορίες για αυτούς.

3.4 Επιλογή χαρακτηριστικών – Αρχική ταξινόμηση

Τα ισχυρά σκορόχαρτα συνήθως περιέχουν από 10 έως 20 χαρακτηριστικά, παρ' όλα αυτά τις περισσότερες φορές υπάρχουν πολύ περισσότερα χαρακτηριστικά διαθέσιμα από αυτά που απαιτούνται για ένα «γερό» σκορόχαρτο. Επομένως, για την ανάπτυξη ενός στατιστικού μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας θα επιλεγούν τελικά τα πιο σημαντικά χαρακτηριστικά. Τα χαρακτηριστικά που αφαιρούνται, είτε προσφέρουν μικρή διάκριση μεταξύ των «καλών» και των «κακών» πελατών, είτε είναι έντονα συσχετισμένα με άλλα χαρακτηριστικά που υπάρχουν στο σκορόχαρτο, είτε οι τιμές τους δεν είναι ανθεκτικές με την πάροδο του χρόνου. Υπάρχουν διάφοροι τρόποι για να επιλεχτούν τα πιο σημαντικά χαρακτηριστικά κάποιοι από τους οποίους είναι οι εξής:

- Η χρησιμοποίηση των **βηματικών στατιστικών διαδικασιών (*stepwise procedures*)**. Για παράδειγμα, οι κατά μπροστινά βήματα μέθοδοι (*forward stepwise methods*) προσθέτουν σε κάθε βήμα τη μεταβλητή (ή μια ομάδα μεταβλητών) που οδηγούν στη μέγιστη βελτίωση της προβλεπτικής ακρίβειας του μοντέλου. Οι κατά πίσω βήματα μέθοδοι

(*backward stepwise methods*) αφαιρούν σε κάθε βήμα τη μεταβλητή με τη μικρότερη προβλεπτική αξία.

- Η χρησιμοποίηση ειδικής γνώσης, εμπειρίας ή απλά λειτουργώντας διαισθητικά για την επιλογή των πιο σημαντικών χαρακτηριστικών παρέχουν ένα καλό συμπλήρωμα στους επίσημους στατιστικούς χειρισμούς. Οι στατιστικοί χειρισμοί βοηθούν στη μη συμπερίληψη ή στην αφαίρεση χαρακτηριστικών που δεν συνεισφέρουν στην προβλεψιμότητα του μοντέλου ενώ η χρησιμοποίηση διαίσθησης ή εμπειρίας είναι σημαντική όταν υπάρχει η ανάγκη να δικαιολογηθεί η επιλογή των χαρακτηριστικών αυτών.
- Η επιλογή των μεμονωμένων χαρακτηριστικών με τη χρησιμοποίηση ενός μέτρου που εκφράζει τη διαφορά μεταξύ των κατανομών των «καλών» και «κακών» δανειοληπτών σε κάθε χαρακτηριστικό. Ένα τέτοιο μέτρο είναι η **τιμή πληροφορίας (*information value*)**, που ορίζεται ως:

$$IV_i = \sum_j (p_{ij} - q_{ij}) w_{ij}$$

όπου,

$w_{ij} = \ln(p_{ij} / q_{ij})$: τα βάρη ενδείξεων (*weights of evidence*) της *j*-οστής ιδιότητας του *i*-οστού χαρακτηριστικού

p_{ij}: ο αριθμός των «καλών» πελατών της ιδιότητας *j* του χαρακτηριστικού *i* διαιρεμένο με το συνολικό αριθμό των «καλών» που απαντούν στο χαρακτηριστικό *i* και

q_{ij}: ο αριθμός των «κακών» πελατών της ιδιότητας *j* του χαρακτηριστικού *i* διαιρεμένο με το συνολικό αριθμό των «κακών» που απαντούν στο χαρακτηριστικό *i*.

Σύμφωνα με τους Hand and Henley (1997), οποιοδήποτε χαρακτηριστικό με τιμή πληροφορίας πάνω από 0,1 θα εξεταστεί έτσι ώστε να συμπεριληφθεί στο μοντέλο.

Στην πράξη χρησιμοποιούνται όλες αυτές οι μέθοδοι, ίσως ξεκινώντας με μια αρχική επιλογή των χαρακτηριστικών με βάση την εμπειρία και έπειτα να αποβάλλονται κάποιες από αυτές χρησιμοποιώντας τις κατά βήμα μεθόδους.

Μόλις μειωθεί ο αριθμός χαρακτηριστικών τότε κάθε ένα από αυτά που έχουν απομείνει ταξινομείται «αρχικά» έτσι ώστε να αυξηθεί η ανθεκτικότητά του και να μπορεί να αντιμετωπιστεί οποιαδήποτε μη μονότονη σχέση μεταξύ του χαρακτηριστικού και του κινδύνου αθέτησης υποχρεώσεων. **Αρχική ταξινόμηση (*coarse classifying*)** μιας κατηγορικής μεταβλητής σημαίνει ότι ομαδοποιούνται οι ιδιότητες της μεταβλητής σε κάποιο

αριθμό «δοχείων» (κλάσεων) έτσι ώστε οι ιδιότητες που έχουν κατά προσέγγιση την ίδια αναλογία «καλοί»:«κακοί» πελατών να βρίσκονται στο ίδιο δοχείο. Ο αριθμός δοχείων είναι τέτοιος ώστε κάθε ένα να περιέχει τουλάχιστον το 5% του πληθυσμού. Οι διακριτές μεταβλητές χρειάζεται να ταξινομηθούν (αρχική ταξινόμηση) γιατί συνήθως κάθε μία από αυτές αποτελείται από πολλές ιδιότητες και ίσως να μην υπάρχει επαρκές δείγμα πελατών που να εμφανίζουν κάποιες ιδιότητες έτσι ώστε να έχουμε μια ανθεκτική ανάλυση. Για τις διατάξιμες μεταβλητές τα δοχεία ουσιαστικά είναι ομάδες όπου όλες οι συναφείς ιδιότητες βρίσκονται μαζί. Για τα συνεχή χαρακτηριστικά όπως η ηλικία, οι τιμές του χαρακτηριστικού χωρίζονται αρχικά σε 10 έως 20 κλάσεις και έτσι η μεταβλητή αυτή συμπεριφέρεται σαν μια διατάξιμη μεταβλητή με αυτές τις ιδιότητες. Έπειτα αποφασίζεται εάν οι συναφείς ιδιότητες πρέπει να παραμείνουν στην ίδια ομάδα ή όχι.

Για να περιγραφεί πόσο καλό είναι ένα χαρακτηριστικό με μια συγκεκριμένη αρχική ταξινόμηση όσον αφορά τη διαφοροποίηση των «καλών» και των «κακών» πελατών, συνήθως χρησιμοποιείται το στατιστικό X^2. Αυτό το στατιστικό είναι πολύ χρήσιμο γιατί βοηθάει στη δημιουργία της καταλληλότερης ομαδοποίησης του κάθε χαρακτηριστικού και στην κατάταξη των ομαδοποιημένων χαρακτηριστικών.

- **Στατιστικό X^2**

Το στατιστικό X^2 χρησιμοποιείται για να ελέγξουμε την υπόθεση ότι η αναλογία «καλοί»:«κακοί» πελάτες είναι ίδια σε κάθε ένα από τα δοχεία. Υποθέτουμε ότι υπάρχουν K δοχεία και έστω ότι στο k-οστό δοχείο, όπου $k=1,2,...,K$ υπάρχουν συνολικά n_k πελάτες από τους οποίους οι g_k είναι «καλοί» και οι $n_k - g_k$ είναι «κακοί». Στην περίπτωση που εξετάζουμε ταυτόχρονα I χαρακτηριστικά, αυτά τα συμβολίζουμε με $i=1,2,...,I$ και τις ιδιότητες του i χαρακτηριστικού τις συμβολίζουμε με $j=1,2,...,m_i$. Επομένως, έχουμε τους εξής συμβολισμούς:

g_{ij} : ο αριθμός των «καλών» πελατών που έχουν την ιδιότητα j του i χαρακτηριστικού,

b_{ij} : ο αριθμός των «κακών» πελατών που έχουν την ιδιότητα j του i χαρακτηριστικού.

n_{ij} : ο συνολικός των πελατών που έχουν την ιδιότητα j του i χαρακτηριστικού.

Η υπόθεση ότι το ποσοστό των «καλών» και των «κακών» πελατών σε κάθε δοχείο είναι ίδιο με το ποσοστό τους στο συνολικό πληθυσμό (p_G για τους «καλούς» και p_B για τους

«κακούς») σημαίνει ότι ο αριθμός των «καλών» στο δοχείο k είναι $n_k p_G$ και ο αριθμός των «κακών» είναι $n_k p_B = n_k(1-p_G)$.

Το στατιστικό X^2 καλής προσαρμογής είναι το άθροισμα των τετραγώνων των διαφορών των παρατηρούμενων από τις αναμενόμενες τιμές διαιρεμένο με τη διακύμανση. Ο αριθμός των «καλών» στο δοχείο k ακολουθεί διωνυμική κατανομή $B(n_k, p_G)$ με μέση τιμή $n_k p_G$ και διακύμανση $n_k p_G p_B = n_k p_G(1-p_G)$. Επομένως, το στατιστικό X^2 καλής προσαρμογής θα δίνεται από τον τύπο:

$$X^2 = \sum_{k=1}^{K} \frac{(g_k - n_k p_G)^2}{n_k p_G(1-p_G)} \sim \chi^2_{K-1}.$$

Για κάθε διαφορετική ομαδοποίηση των ιδιοτήτων των χαρακτηριστικών υπολογίζεται το στατιστικό X^2 και ελέγχεται εάν ισχύει ή όχι η υπόθεση ότι το ποσοστό των «καλών» και των «κακών» πελατών σε κάθε δοχείο είναι η ίδιο με το ποσοστό τους στο συνολικό πληθυσμό. Χρησιμοποιώντας αυτό το στατιστικό επιλέγουμε την ομαδοποίηση που δίνει τη μεγαλύτερη τιμή του στατιστικού X^2 γιατί αυτό σημαίνει ότι υπάρχουν σημαντικές διαφορές στην αναλογία «καλοί»:«κακοί» μεταξύ των διαφορετικών δοχείων. Πρακτικά όμως, όταν ένα χαρακτηριστικό ομαδοποιείται σε περισσότερα δοχεία ελέγχεται εάν υπήρξε μεγάλη αύξηση του στατιστικού X^2 με την εισαγωγή του νέου δοχείου και αν κριθεί ότι πράγματι υπήρξε διαφορά τότε η διαδικασία συνεχίζεται.

Αφού διασπαστούν οι ιδιότητες του χαρακτηριστικού X σε K δοχεία $C_1, C_2, ..., C_K$, συνήθως στη συνέχεια της ανάλυσης τα ταξινομημένα χαρακτηριστικά μετατρέπονται σε νέες μεταβλητές. Η μεταβλητή X περιγράφεται από $K-1$ δίτιμες μεταβλητές $X_1, X_2, ..., X_{K-1}$ οι οποίες ορίζονται ως εξής

$$X_i(x) = \begin{cases} 1, \text{ εάν } x \in C_i \\ 0, \text{ διαφορετικά} \end{cases}.$$

Η μεταβλητή X_K δεν περιλαμβάνεται γιατί διαφορετικά θα υπήρχε συγγραμμικότητα μεταξύ των μεταβλητών. Η C_K είναι η βασική κατηγορία ή το δοχείο με το οποίο συγκρίνονται τα υπόλοιπα. Μπορούμε να επιλέξουμε οποιοδήποτε από τα K δοχεία ως βασικό.

Μια άλλη προσέγγιση είναι να χρησιμοποιηθούν τα βάρη ενδείξεων (*weights of evidence*). Όταν έχουμε έναν λογάριθμο βαθμολογίας (*log score*) σε ένα μόνο χαρακτηριστικό X, τότε η βαθμολογία που αντιστοιχεί στην ιδιότητα x αυτού του χαρακτηριστικού δίνεται από τον τύπο (3.7).

Εάν g_k και b_k είναι ο αριθμός των «καλών» και των «κακών» αντίστοιχα στο δοχείο C_k τότε $n_G = \sum_{k=1}^{K} g_k$ και $n_B = \sum_{k=1}^{K} b_k$ είναι ο αριθμός των «καλών» και των «κακών» αντίστοιχα σε όλον τον πληθυσμό. Ορίζουμε λοιπόν μια νέα μεταβλητή X_W χρησιμοποιώντας τα βάρη ενδείξεων ως εξής:

$$X_W(x) = \ln\left(\frac{g_k}{b_k} \Big/ \frac{n_G}{n_B}\right) = \ln\left(\frac{g_k n_B}{b_k n_G}\right), \quad \text{εάν} \quad x \in C_k.$$

Η αναλογία $\frac{g_k n_B}{b_k n_G}$ είναι εκτίμηση της αναλογίας

$$\frac{P(G \mid x)}{P(B \mid x)} \Big/ \frac{p_G}{p_B}$$

διότι $\frac{n_G}{n_B} = \frac{p_G}{p_B}$ και $\frac{g_k}{b_k} = \frac{P(G \mid x \in C_k)}{P(B \mid x \in C_k)}$. Επομένως, οι τιμές της έκφρασης

$$w(x) = \ln\left(\frac{P(G \mid x)/P(B \mid x)}{p_G / p_B}\right)$$

είναι εκτιμήσεις των βαρών ενδείξεων για τις ιδιότητες του χαρακτηριστικού X. Αυτό εξασφαλίζει ότι οι τιμές που δίνονται στα δοχεία της μεταβλητής X έχουν την ίδια διάταξη με τις εμπειρικά σχετικές πιθανότητες σε κάθε κλάση.

Χρησιμοποιώντας τα βάρη των ενδείξεων έχουμε ως πλεονέκτημα ότι ο αριθμός χαρακτηριστικών δεν αυξάνεται και έτσι υπάρχει μικρότερη πιθανότητα να εμφανιστεί συσχέτιση μεταξύ των μεταβλητών και επιπλέον περισσότερη ανθεκτικότητα στη στατιστική εκτίμηση. Το βασικό μειονέκτημα είναι όμως ότι εάν υπάρχει μια ισχυρή αλληλεπίδραση με άλλα χαρακτηριστικά, αυτή η διάταξη μπορεί να μην απεικονίζει μόνο την επίδραση των ιδιοτήτων αυτού του χαρακτηριστικού. Με τα βάρη ενδείξεων θα πρέπει να κρατήσουμε όλες ή καμία από τις ιδιότητες του χαρακτηριστικού για τη δόμηση του CSM χωρίς να έχουμε κάποια δυνατότητα επιλογής. Χρησιμοποιώντας όμως τις δυαδικές μεταβλητές δίνεται η δυνατότητα για μερικές από τις ιδιότητες ενός χαρακτηριστικού να αφαιρεθούν ή να μείνουν στο σκορόχαρτο. Αυτό είναι πολύ χρήσιμο γιατί δίνεται έμφαση στις ιδιότητες που επιδρούν

περισσότερο στην πιθανότητα αθέτησης υποχρεώσεων ή στην πιθανότητα ο πελάτης να είναι «καλός».

3.5 Διαχωριστική Ανάλυση για κανονικούς πληθυσμούς

Η γενική ιδέα της διαχωριστικής ανάλυσης (ΔΑ) είναι να καταταχτούν πολυδιάστατες παρατηρήσεις σε πληθυσμούς με γνωστές κατανομές. Ο σκοπός της ανάπτυξης ενός μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας με τη μέθοδο αυτή είναι να εντοπιστούν οι «καλοί» και οι «κακοί» πελάτες πριν τη χορήγηση ενός δανείου. Συνεπώς, με τη χρήση ιστορικών στοιχείων σχετικά με άτομα που έλαβαν πίστωση από την ενδιαφερόμενη επιχείρηση μπορούν να σχηματιστούν κάποιοι κανόνες, ώστε να καταταχθεί ένας καινούργιος υποψήφιος πελάτης σε μία από τις δύο κατηγορίες και έπειτα να αποφασιστεί αν θα γίνει δεκτή ή όχι η αίτηση του πελάτη για χορήγηση πίστωσης με βάση το επίπεδο κινδύνου που έχει διαγνωστεί για αυτόν (Καρλής (2005)).

Υποθέτουμε ότι το A είναι το σύνολο όλων των πιθανών τιμών που οι τυχαίες μεταβλητές $\mathbf{X} = (X_1, X_2, ..., X_p)$ μπορούν να πάρουν, δηλαδή είναι το σύνολο όλων των διαφορετικών τρόπων με τους οποίους μπορεί να συμπληρωθεί η αίτηση για πίστωση. Ο στόχος εδώ είναι να βρεθεί ένας κανόνας που θα διαχωρίζει το σύνολο A σε δύο υποσύνολα A_G και A_B έτσι ώστε οι υποψήφιοι πελάτες των οποίων οι απαντήσεις ανήκουν στο σύνολο A_G να ταξινομούνται στους «καλούς» και να γίνονται δεκτές οι αιτήσεις τους. Επιπλέον, με βάση αυτόν τον κανόνα οι πελάτες των οποίων οι απαντήσεις ανήκουν στο σύνολο A_B θα ταξινομούνται στους «κακούς» και οι αιτήσεις τους θα απορρίπτονται. Με αυτόν τον τρόπο θα ελαχιστοποιείται το αναμενόμενο κόστος του δανειστή. Υπάρχουν δύο τύποι κόστους που αντιστοιχούν στους δύο τύπους λαθών που μπορούν να γίνουν με την κάθε απόφαση. Ο πρώτος τύπος λάθους είναι να ταξινομηθεί κάποιος πελάτης στους «κακούς» και να απορριφθεί ενώ στην πραγματικότητα είναι «καλός». Σε αυτήν την περίπτωση το πιθανό κέρδος από αυτόν τον πελάτη χάνεται. Υποθέτουμε ότι το αναμενόμενο κέρδος θα είναι το ίδιο για κάθε πελάτη και θα το συμβολίζουμε με L. Ο δεύτερος τύπος λάθους είναι να ταξινομηθεί κάποιος πελάτης στους «καλούς» και να γίνει αποδεκτός ενώ στην

πραγματικότητα αυτός είναι «κακός». Σε αυτήν την περίπτωση θα υπάρξει κόστος στην επιχείρηση όταν ο πελάτης αθετήσει τις υποχρεώσεις του. Υποθέτουμε λοιπόν ότι το αναμενόμενο κόστος θα είναι το ίδιο για κάθε υποψήφιο πελάτη και το συμβολίζουμε με D.

Υποθέτουμε ότι τα χαρακτηριστικά έχουν έναν πεπερασμένο αριθμό από διακριτές ιδιότητες έτσι ώστε το A να είναι πεπερασμένο και να υπάρχει μόνο ένας πεπερασμένος αριθμός διαφορετικών ιδιοτήτων $\mathbf{x}$, δηλαδή να υπάρχει ένας πεπερασμένος αριθμός τρόπων για τη συμπλήρωση της αίτησης. Το αναμενόμενο κόστος ανά υποψήφιο πελάτη, αν γίνονται δεκτοί αυτοί που έχουν ιδιότητες στο σύνολο A_G και απορρίπτονται αυτοί που έχουν ιδιότητες στο σύνολο A_B είναι:

$$L\sum_{\mathbf{x}\in A_B} P(\mathbf{x}\,|\,G)p_G + D\sum_{\mathbf{x}\in A_G} P(\mathbf{x}\,|\,B)p_B = L\sum_{\mathbf{x}\in A_B} P(G\,|\,\mathbf{x})P(\mathbf{x}) + D\sum_{\mathbf{x}\in A_G} P(B\,|\,\mathbf{x})P(\mathbf{x}).$$

Εάν ένα συγκεκριμένο διάνυσμα ιδιοτήτων $\mathbf{x}=(x_1,x_2,...,x_p)$ ταξινομηθεί στο σύνολο A_G θα υπάρξει κόστος μόνο αν στην πραγματικότητα αυτό το διάνυσμα ιδιοτήτων ανήκει στο σύνολο A_B. Σε αυτήν την περίπτωση το αναμενόμενο κόστος είναι $D\cdot P(\mathbf{x}\,|\,B)p_B$. Ενώ, αν το διάνυσμα των ιδιοτήτων $\mathbf{x}$ ταξινομηθεί στο υποσύνολο A_B, θα υπάρξει απώλεια μόνο αν στην πραγματικότητα αυτές οι τιμές χαρακτηρίζουν έναν «καλό» πελάτη και η αναμενόμενη απώλεια είναι $L\cdot P(\mathbf{x}\,|\,G)p_G$. Άρα, μια σκέψη είναι ο πελάτης που έχει τις ιδιότητες $\mathbf{x}$ να ταξινομείται στο σύνολο A_G εάν ισχύει η σχέση:

$$D\cdot P(\mathbf{x}\,|\,B)p_B \le L\cdot P(\mathbf{x}\,|\,G)p_G\,.$$

Επομένως, για να ελαχιστοποιηθεί το αναμενόμενο κόστος θα πρέπει να ικανοποιούνται οι αιτήσεις των πελατών των οποίων τα χαρακτηριστικά ανήκουν στο σύνολο

$$A_G=\left\{\mathbf{x}\middle|D\cdot P(\mathbf{x}\,|\,B)p_B \le L\cdot P(\mathbf{x}\,|\,G)p_G\right\}=\left\{\mathbf{x}\middle|\frac{D}{L}\le\frac{P(\mathbf{x}\,|\,G)p_G}{P(\mathbf{x}\,|\,B)p_B}\right\}.$$

Από τον τύπο (3.3) έχουμε ότι:

$$A_G=\left\{\mathbf{x}\middle|\frac{D}{L}\le\frac{P(G\,|\,\mathbf{x})}{P(B\,|\,\mathbf{x})}\right\}. \tag{3.8}$$

Στην περίπτωση που τα χαρακτηριστικά για την εφαρμογή της μεθόδου δεν προέρχονται από διακριτές τυχαίες μεταβλητές αλλά από συνεχείς τότε ακολουθείται η παραπάνω διαδικασία με τη διαφορά ότι οι δεσμευμένες συναρτήσεις πιθανότητας $P(\mathbf{x}\,|\,G)$ και $P(\mathbf{x}\,|\,B)$ θα αντικαθίστανται από τις δεσμευμένες συναρτήσεις πυκνότητας $f(\mathbf{x}\,|\,G)$ και $f(\mathbf{x}\,|\,B)$ και

τα ανωτέρω αθροίσματα θα αντικαθίστανται από ολοκληρώματα. Άρα, στην περίπτωση αυτή, το αναμενόμενο κόστος ανά υποψήφιο πελάτη αν γίνονται δεκτοί αυτοί που έχουν ιδιότητες στο σύνολο A_G και απορρίπτονται αυτοί που έχουν ιδιότητες στο σύνολο A_B είναι ίσο με

$$L\int_{\mathbf{x}\in A_B} f(\mathbf{x}\,|\,G)p_G d\mathbf{x} + D\int_{\mathbf{x}\in A_G} f(\mathbf{x}\,|\,B)p_B d\mathbf{x}.$$

Επομένως, σύμφωνα με τη σχέση (3.8) για να ελαχιστοποιηθεί το αναμενόμενο κόστος θα πρέπει να ικανοποιούνται οι αιτήσεις των πελατών των οποίων τα χαρακτηριστικά ανήκουν στο σύνολο

$$A_G = \left\{\mathbf{x}\middle| Df(\mathbf{x}\,|\,B)p_B \le Lf(\mathbf{x}\,|\,G)p_G\right\} = \left\{\mathbf{x}\middle|\frac{Dp_B}{Lp_G} \le \frac{f(\mathbf{x}\,|\,G)}{f(\mathbf{x}\,|\,B)}\right\}. \tag{3.9}$$

Ας εξετάσουμε στη συνέχεια την περίπτωση που υπάρχει μία μόνο συνεχής μεταβλητή (χαρακτηριστικό) X και η κατανομή $f(\mathbf{x}\,|\,G)$ των «καλών» οφειλετών είναι κανονική με μέση τιμή μ_G και διακύμανση σ^2 (δηλαδή όταν $X \in A_G$ τότε $X \sim N(\mu_G, \sigma^2)$), ενώ η κατανομή των «κακών» οφειλετών είναι επίσης κανονική με μέση τιμή μ_B και διακύμανση σ^2 (δηλαδή εάν $X \in A_B$ τότε $X \sim N(\mu_B, \sigma^2)$). Επομένως, θα έχουμε

$$f(x\,|\,G) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu_G)^2}{2\sigma^2}\right) \quad \text{και} \quad f(x\,|\,B) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu_B)^2}{2\sigma^2}\right).$$

απ' όπου προκύπτει ότι

$$\frac{f(x\,|\,G)}{f(x\,|\,B)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu_G)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu_B)^2}{2\sigma^2}\right)} = \frac{\exp\left(\frac{-(x-\mu_G)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(x-\mu_B)^2}{2\sigma^2}\right)} = \exp\left(\frac{(x-\mu_B)^2-(x-\mu_G)^2}{2\sigma^2}\right).$$

Επομένως, σύμφωνα με τη σχέση (3.9) για να ελαχιστοποιηθεί το αναμενόμενο κόστος θα πρέπει να ικανοποιούνται οι αιτήσεις των πελατών των οποίων τα χαρακτηριστικά ανήκουν στο σύνολο

$$A_G = \left\{x\middle|\frac{Dp_B}{Lp_G} \le \frac{f(x\,|\,G)}{f(x\,|\,B)}\right\} = \left\{x\middle|\frac{Dp_B}{Lp_G} \le \exp\left(\frac{(x-\mu_B)^2-(x-\mu_G)^2}{2\sigma^2}\right)\right\} =$$

$$= \left\{ x \left| 2\sigma^2 \ln\left(\frac{Dp_B}{Lp_G}\right) \le x^2 - 2x\mu_B + {\mu_B}^2 - x^2 + 2x\mu_G - {\mu_G}^2 \right. \right\} =$$

$$= \left\{ x \left| 2\sigma^2 \ln\left(\frac{Dp_B}{Lp_G}\right) \le 2x(\mu_G - \mu_B) + {\mu_B}^2 - {\mu_G}^2 \right. \right\} =$$

$$= \left\{ x \left| 2\sigma^2 \ln\left(\frac{Dp_B}{Lp_G}\right) - {\mu_B}^2 + {\mu_G}^2 \le 2x(\mu_G - \mu_B) \right. \right\} =$$

$$= \left\{ x \left| x(\mu_G - \mu_B) \ge 2\sigma^2 \ln\left(\frac{Dp_B}{Lp_G}\right) - {\mu_B}^2 + {\mu_G}^2 \right. \right\}.$$

Ως εκ τούτου, αν για την τιμή του χαρακτηριστικού x ισχύει

$$x(\mu_G - \mu_B) \ge 2\sigma^2 \ln\left(\frac{Dp_B}{Lp_G}\right) - {\mu_B}^2 + {\mu_G}^2$$

τότε ταξινομούμε τον πελάτη ως «καλό», αντίθετα τον ταξινομούμε ως «κακό».

Ας δούμε στη συνέχεια την περίπτωση της πολυμεταβλητής κανονικής κατανομής με κοινή διακύμανση. Έστω $\mathbf{X} = (X_1, X_2, ..., X_p)$ είναι ένα τυχαίο διάνυσμα που αποτελείται από p τυχαίες μεταβλητές (χαρακτηριστικά). Οι τιμές αυτών των τυχαίων μεταβλητών προέρχονται από την πολυμεταβλητή κανονική κατανομή σε οποιαδήποτε ομάδα και αν ανήκουν («καλοί» ή «κακοί»). Υποθέτουμε ότι οι πίνακες συνδιακύμανσης είναι ίσοι σε κάθε ομάδα και ο κοινός πίνακας συνδιακύμανσης θα συμβολίζεται με $\mathbf{\Sigma}$, δηλαδή

$$\mathbf{\Sigma} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \cdots & Var(X_p) \end{pmatrix}.$$

Η μέση τιμή των παρατηρήσεων που ανήκουν στην ομάδα των «καλών» είναι $\boldsymbol{\mu}_G = (\mu_{G1}, \mu_{G2}, ..., \mu_{Gp})$ και η μέση τιμή των τυχαίων διανυσμάτων που ανήκουν στην ομάδα των «κακών» είναι $\boldsymbol{\mu}_B = (\mu_{B1}, \mu_{B2}, ..., \mu_{Bp})$. Αυτό σημαίνει ότι:

$\mathbf{X}_G \sim N_p(\boldsymbol{\mu}_G, \mathbf{\Sigma})$, $\mathbf{X}_B \sim N_p(\boldsymbol{\mu}_B, \mathbf{\Sigma})$

$E(X_i \mid G) = \mu_{Gi}$, $E(X_i \mid B) = \mu_{Bi}$ και $Cov_B(X_i, X_j) = Cov_G(X_i, X_j) = \Sigma_{ij}$ ή

$$E[(X_i - E(X_i))(X_j - E(X_j)) \mid B] = E[(X_i - E(X_i))(X_j - E(X_j)) \mid G] = \Sigma_{ij}$$

για κάθε $i, j = 1,2,...,p$. Γνωρίζουμε ότι όταν ένα τυχαίο διάνυσμα $\mathbf{X} = (X_1, X_2, ..., X_p)$ ακολουθεί $p-$διάστατη κανονική κατανομή με μέσο $\boldsymbol{\mu}$ και πίνακα συνδιακύμανσης $\boldsymbol{\Sigma} > 0$, δηλαδή $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, τότε θα έχει συνάρτηση πυκνότητας της μορφής

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T}{2} \right).$$

Άρα, η συνάρτηση πυκνότητας για τους «καλούς» πελάτες και τους «κακούς» θα είναι αντίστοιχα

$$f(\mathbf{x} \mid G) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2} \right),$$

$$f(\mathbf{x} \mid B) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2} \right).$$

Έχουμε

$$\frac{f(x \mid G)}{f(x \mid B)} = \frac{\frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2} \right)}{\frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2} \right)} = \frac{exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2} \right)}{exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2} \right)} =$$

$$= exp\left(\frac{(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T - (\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2} \right)$$

και επομένως, σύμφωνα με τη σχέση (3.9) για να ελαχιστοποιηθεί το αναμενόμενο κόστος θα πρέπει να ικανοποιούνται οι αιτήσεις των πελατών των οποίων τα χαρακτηριστικά ανήκουν στο σύνολο

$$A_G = \left\{ x \left| \frac{Dp_B}{Lp_G} \le \frac{f(x \mid G)}{f(x \mid B)} \right. \right\} = \left\{ x \left| \frac{Dp_B}{Lp_G} \le exp\left(\frac{(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T - (\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2} \right) \right. \right\} =$$

$$= \left\{ x \left| 2\ln\left(\frac{Dp_B}{Lp_G} \right) \le (\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T - (\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T \right. \right\} =$$

$$= \left\{ \mathbf{x} \left| 2\ln\left(\frac{Dp_B}{Lp_G} \right) \le \mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}^T - \mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B{}^T - \boldsymbol{\mu}_B\boldsymbol{\Sigma}^{-1}\mathbf{x}^T + \boldsymbol{\mu}_B\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B{}^T - \mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}^T + \mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G{}^T + \right. \right. \right.$$
$$\left. \left. \left. + \boldsymbol{\mu}_G\boldsymbol{\Sigma}^{-1}\mathbf{x}^T - \boldsymbol{\mu}_G\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G{}^T \right. \right. \right\} =$$

$$= \left\{ \mathbf{x} \left| 2\ln\left(\frac{Dp_B}{Lp_G}\right) \le -2\mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B{}^T + \boldsymbol{\mu}_B{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B + 2\mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G{}^T - \boldsymbol{\mu}_G{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G \right.\right\} =$$

$$= \left\{ \mathbf{x} \left| 2\ln\left(\frac{Dp_B}{Lp_G}\right) \le 2\mathbf{x}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_G - \boldsymbol{\mu}_B\right)^T + \boldsymbol{\mu}_B{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B - \boldsymbol{\mu}_G{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G \right.\right\}.$$

Δηλαδή

$$A_G = \left\{ \mathbf{x} \left| \mathbf{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_G - \boldsymbol{\mu}_B)^T \ge \ln\left(\frac{Dp_B}{Lp_G}\right) + \frac{\boldsymbol{\mu}_G{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G - \boldsymbol{\mu}_B{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B}{2} \right.\right\}. \qquad (3.10)$$

Το αριστερό μέλος της ανισότητας στην (3.10) είναι ένα σταθμισμένο άθροισμα των τιμών των μεταβλητών και έχει μορφή $b_1x_1 + b_2x_2 + ... + b_px_p$ ενώ το δεξί μέλος της ανισότητας είναι μια σταθερά. Έτσι, το κριτήριο αυτό αποτελεί έναν γραμμικό κανόνα βαθμολόγησης, ο οποίος είναι γνωστός ως γραμμικός διαχωριστικός κανόνας. Όλα τα παραπάνω προϋποθέτουν ότι είναι γνωστές οι παράμετροι των πληθυσμών $\boldsymbol{\mu}_G$, $\boldsymbol{\mu}_B$ και ο πίνακας συνδιασποράς $\boldsymbol{\Sigma}$. Βέβαια στην πράξη αυτό δεν είναι εφικτό, επομένως θα πρέπει να εκτιμηθούν. Έστω,

$\mathbf{X}_{G1}, \mathbf{X}_{G2}, ..., \mathbf{X}_{Gn_G}$ τυχαίο δείγμα n_G «καλών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_G, \boldsymbol{\Sigma})$

και

$\mathbf{X}_{B1}, \mathbf{X}_{B2}, ..., \mathbf{X}_{Bn_B}$ τυχαίο δείγμα n_B «κακών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_B, \boldsymbol{\Sigma})$.

Τότε για την εκτίμηση της παραμέτρου $\boldsymbol{\mu}_G$ παίρνουμε το δειγματικό μέσο, δηλαδή έχουμε

$$\boldsymbol{\mu}_G = \mathbf{m}_G = (m_{G1}, m_{G2}, ..., m_{Gp}) \text{, όπου } m_{Gj} = \frac{1}{n_G}\sum_{i=1}^{n_G} x_{Gi},$$

για κάθε $j = 1,2,...,p$.

Ομοίως, για την εκτίμηση της παραμέτρου $\boldsymbol{\mu}_B$ παίρνουμε το δειγματικό μέσο, δηλαδή

$$\boldsymbol{\mu}_B = \mathbf{m}_B = (m_{B1}, m_{B2}, ..., m_{Bp}) \text{, όπου } m_{Bj} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{Bi},$$

για κάθε $j = 1,2,...,p$.

Για την εκτίμηση του πίνακα συνδιακύμανσης παίρνουμε το δειγματικό πίνακα συνδιακύμανσης, δηλαδή έχουμε

$$\boldsymbol{\Sigma} = \mathbf{S}_p = \frac{1}{n_G + n_B - 2}\left[\sum_{i=1}^{n_G}(\mathbf{x}_{Gi} - \mathbf{m}_G)^T(\mathbf{x}_{Gi} - \mathbf{m}_G) + \sum_{i=1}^{n_B}(\mathbf{x}_{Bi} - \mathbf{m}_B)^T(\mathbf{x}_{Bi} - \mathbf{m}_B)\right].$$

Επομένως, σύμφωνα με την παραπάνω διαδικασία ισχύει ο ακόλουθος κανόνας:

Αν $\mathbf{X}_{G1},\mathbf{X}_{G2},...,\mathbf{X}_{Gn_G}$ είναι ένα τυχαίο δείγμα n_G «καλών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_G,\boldsymbol{\Sigma})$ και $\mathbf{X}_{B1},\mathbf{X}_{B2},...,\mathbf{X}_{Bn_B}$ ένα τυχαίο δείγμα n_B «κακών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_B,\boldsymbol{\Sigma})$, τότε ο πελάτης με τις ιδιότητες $\mathbf{x}$ ταξινομείται στους «καλούς» δανειολήπτες αν ισχύει:

$$\mathbf{x}\mathbf{S}_p^{-1}(\mathbf{m}_G-\mathbf{m}_B)^T \geq \ln\left(\frac{Dp_B}{Lp_G}\right)+\frac{\mathbf{m}_G\mathbf{S}_p^{-1}\mathbf{m}_G^T-\mathbf{m}_B\mathbf{S}_p^{-1}\mathbf{m}_B^T}{2}$$

και στους «κακούς» δανειολήπτες αν ισχύει:

$$\mathbf{x}\mathbf{S}_p^{-1}(\mathbf{m}_G-\mathbf{m}_B)^T < \ln\left(\frac{Dp_B}{Lp_G}\right)+\frac{\mathbf{m}_G\mathbf{S}_p^{-1}\mathbf{m}_G^T-\mathbf{m}_B\mathbf{S}_p^{-1}\mathbf{m}_B^T}{2}.$$

Ας εξετάσουμε τέλος την περίπτωση πολυμεταβλητής κανονικής κατανομής με μη κοινή διακύμανση. Έστω $\mathbf{X}=(X_1,X_2,...,X_p)$ είναι ένα τυχαίο διάνυσμα που αποτελείται από p τυχαίες μεταβλητές (χαρακτηριστικά). Οι τιμές αυτών των τυχαίων μεταβλητών προέρχονται από την πολυμεταβλητή κανονική κατανομή σε οποιαδήποτε ομάδα και αν ανήκουν. Υποθέτουμε ότι οι πίνακες συνδιακύμανσης είναι ίσοι διαφορετικοί για κάθε ομάδα. Συμβολίζουμε με $\boldsymbol{\Sigma}_G$ τον πίνακα συνδιακύμανσης για την ομάδα των «καλών» πελατών και με $\boldsymbol{\Sigma}_B$ τον πίνακα συνδιακυμανσης για την ομάδα των «κακών» πελατών και θεωρούμε ότι $\boldsymbol{\Sigma}_G \neq \boldsymbol{\Sigma}_B$.

Επίσης, η μέση τιμή των παρατηρήσεων που ανήκουν στην ομάδα των «καλών» είναι $\boldsymbol{\mu}_G=(\mu_{G1},\mu_{G2},...,\mu_{Gp})$ και η μέση τιμή των τυχαίων διανυσμάτων που ανήκουν στην ομάδα των «κακών» είναι $\boldsymbol{\mu}_B=(\mu_{B1},\mu_{B2},...,\mu_{Bp})$. Αυτό σημαίνει ότι:

$$\mathbf{X}_G \sim N_p(\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_G),\ \mathbf{X}_B \sim N_p(\boldsymbol{\mu}_B,\boldsymbol{\Sigma}_B),$$

$$E(X_i\,|\,G)=\mu_{Gi},\ E(X_i\,|\,B)=\mu_{Bi},\ \text{και}\ \ Cov_G(X_i,X_j)=\Sigma_{Gij}\ \text{ή}\ Cov_B(X_i,X_j)=\Sigma_{Bij}$$

για κάθε $i,j=1,2,...,p$.

Άρα, στη συγκεκριμένη περίπτωση η συνάρτηση πυκνότητας είναι:

$$f(\mathbf{x}\,|\,G)=\frac{1}{(2\pi)^{p/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}_G|}}exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}_G^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right)$$

και

$$f(\mathbf{x}\,|\,B)=\frac{1}{(2\pi)^{p/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}_B|}}exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2}\right).$$

Έχουμε:

$$\frac{f(x\,|\,G)}{f(x\,|\,B)}=\frac{\frac{1}{(2\pi)^{p/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}_G|}}exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}_G^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right)}{\frac{1}{(2\pi)^{p/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}_B|}}exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2}\right)}=\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}\cdot\frac{exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}_G^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right)}{exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T}{2}\right)}=$$

$$=\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}exp\left(\frac{(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}_G^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right)$$

και σύμφωνα με τη σχέση (3.9) για να ελαχιστοποιηθεί το αναμενόμενο κόστος θα πρέπει να ικανοποιούνται οι αιτήσεις των πελατών των οποίων τα χαρακτηριστικά ανήκουν στο σύνολο

$$A_G=\left\{\mathbf{x}\left|\frac{Dp_B}{Lp_G}\le\frac{f(x\,|\,G)}{f(x\,|\,B)}\right.\right\}=$$

$$=\left\{\mathbf{x}\left|\frac{Dp_B}{Lp_G}\le\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}exp\left(\frac{(\mathbf{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B^{-1}(\mathbf{x}-\boldsymbol{\mu}_B)^T-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}_G^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right)\right.\right\}=$$

$$=\left\{\mathbf{x}\left|\ln\left(\frac{Dp_B}{Lp_G}\right)\le\ln\left(\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}\right)+\frac{1}{2}\left(\mathbf{x}\boldsymbol{\Sigma}_B^{-1}\mathbf{x}^T-2\mathbf{x}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B^T+\boldsymbol{\mu}_B\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B-\mathbf{x}\boldsymbol{\Sigma}_G^{-1}\mathbf{x}^T+2\mathbf{x}\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G^T-\boldsymbol{\mu}_G\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G\right)\right.\right\}=$$

$$=\left\{\mathbf{x}\left|\ln\left(\frac{Dp_B}{Lp_G}\right)\le\ln\left(\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}\right)+\frac{1}{2}\left(\mathbf{x}\left(\boldsymbol{\Sigma}_B^{-1}-\boldsymbol{\Sigma}_G^{-1}\right)\mathbf{x}^T+2\mathbf{x}\left(\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G-\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B\right)+\boldsymbol{\mu}_B\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B^T-\boldsymbol{\mu}_G\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G^T\right)\right.\right\}=$$

$$=\left\{\mathbf{x}\left|\mathbf{x}\left(\boldsymbol{\Sigma}_B^{-1}-\boldsymbol{\Sigma}_G^{-1}\right)\mathbf{x}^T+2\mathbf{x}\left(\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G-\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B\right)\ge 2\ln\left(\frac{Dp_B}{Lp_G}\right)-2\ln\left(\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_G|}}\right)+\boldsymbol{\mu}_G\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\mu}_G^T-\boldsymbol{\mu}_B\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B^T\right.\right\}.$$

Όλα τα παραπάνω προϋποθέτουν να είναι γνωστές οι παράμετροι των πληθυσμών $\boldsymbol{\mu}_G$, $\boldsymbol{\mu}_B$ και οι πίνακας συνδιασποράς $\boldsymbol{\Sigma}_G$ και $\boldsymbol{\Sigma}_B$. Οι παράμετροι αυτές θα πρέπει να εκτιμηθούν. Όπως και στην περίπτωση που έχουμε κοινό πίνακα συνδιασποράς έχουμε ότι για την εκτίμηση της παραμέτρου $\boldsymbol{\mu}_G$ παίρνουμε το δειγματικό μέσο, δηλαδή έχουμε

$$\boldsymbol{\mu}_G = \mathbf{m}_G = (m_{G1}, m_{G2}, ..., m_{Gp}) \text{, όπου } m_{Gj} = \frac{1}{n_G}\sum_{i=1}^{n_G} x_{Gi} \text{,}$$

για κάθε $j = 1,2,...,p$.

Ομοίως, για την εκτίμηση της παραμέτρου $\boldsymbol{\mu}_B$ δηλαδή έχουμε

$$\boldsymbol{\mu}_B = \mathbf{m}_B = (m_{B1}, m_{B2}, ..., m_{Bp}) \text{, όπου } m_{Bj} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{Bi} \text{,}$$

για κάθε $j = 1,2,...,p$.

Για την εκτίμηση των πινάκων συνδιακύμανσης $\boldsymbol{\Sigma}_G$ και $\boldsymbol{\Sigma}_G$ χρησιμοποιούμε το δειγματικό πίνακα συνδιακύμανσης και έχουμε:

$$\boldsymbol{\Sigma}_G = \mathbf{S}_G = \frac{1}{n_G - 1}\sum_{i=1}^{n_G} (\mathbf{x}_{Gi} - \mathbf{m}_G)^T (\mathbf{x}_{Gi} - \mathbf{m}_G) \text{, } \boldsymbol{\Sigma}_B = \mathbf{S}_B = \frac{1}{n_B - 1}\sum_{i=1}^{n_B} (\mathbf{x}_{Bi} - \mathbf{m}_B)^T (\mathbf{x}_{Bi} - \mathbf{m}_B).$$

Επομένως, σύμφωνα με την παραπάνω διαδικασία ισχύει ο ακόλουθος κανόνας:

Αν $\mathbf{X}_{G1}, \mathbf{X}_{G2}, ..., \mathbf{X}_{Gn_G}$ είναι ένα τυχαίο δείγμα n_G «καλών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)$ και $\mathbf{X}_{B1}, \mathbf{X}_{B2}, ..., \mathbf{X}_{Bn_B}$ είναι τυχαίο δείγμα n_B «κακών» πελατών από την κατανομή $N_p(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$ τότε ο πελάτης με τις ιδιότητες $\mathbf{x}$ ταξινομείται στους «καλούς» δανειολήπτες αν ισχύει:

$$\mathbf{x}\left(\mathbf{S}_B^{-1} - \mathbf{S}_G^{-1}\right)\mathbf{x}^T + 2\mathbf{x}\left(\mathbf{S}_G^{-1}\mathbf{m}_G - \mathbf{S}_B^{-1}\mathbf{m}_B\right) \geq 2\ln\left(\frac{Dp_B}{Lp_G}\right) - 2\ln\left(\frac{\sqrt{|\mathbf{S}_B|}}{\sqrt{|\mathbf{S}_G|}}\right) + \mathbf{m}_G\mathbf{S}_G^{-1}\mathbf{m}_G^T - \mathbf{m}_B\mathbf{S}_B^{-1}\mathbf{m}_B^T$$

και στους «κακούς» δανειολήπτες αν ισχύει:

$$\mathbf{x}\left(\mathbf{S}_B^{-1} - \mathbf{S}_G^{-1}\right)\mathbf{x}^T + 2\mathbf{x}\left(\mathbf{S}_G^{-1}\mathbf{m}_G - \mathbf{S}_B^{-1}\mathbf{m}_B\right) < 2\ln\left(\frac{Dp_B}{Lp_G}\right) - 2\ln\left(\frac{\sqrt{|\mathbf{S}_B|}}{\sqrt{|\mathbf{S}_G|}}\right) + \mathbf{m}_G\mathbf{S}_G^{-1}\mathbf{m}_G^T - \mathbf{m}_B\mathbf{S}_B^{-1}\mathbf{m}_B^T$$

Αυτή η περίπτωση της Διαχωριστικής Ανάλυσης ονομάζεται Τετραγωνική Διαχωριστική Ανάλυση (*Quadratic Discriminant Analysis*).

3.6 Η διαχωριστική συνάρτηση του Fisher – Μοντέλο του Altman

Ο στόχος της Διαχωριστικής Ανάλυσης είναι να ταξινομηθεί ένας ετερογενής πληθυσμός σε ομοιογενείς υποομάδες και ακολούθως να ληφθούν αποφάσεις πάνω σε αυτές τις υποομάδες. Μπορούμε να υποθέσουμε ότι για κάθε υποψήφιο υπάρχει ένας συγκεκριμένος αριθμός από επεξηγηματικές μεταβλητές. Σύμφωνα με τους Vojtek και Kočenda (2006) η βασική ιδέα είναι να βρεθεί ο κατάλληλος γραμμικός συνδυασμός των επεξηγηματικών μεταβλητών που διαχωρίζει περισσότερο τις υποομάδες. Στην απλή περίπτωση που θέλουμε να διαχωρίσουμε τον πληθυσμό σε δύο υποομάδες, ο σκοπός είναι να βρεθεί εκείνος ο γραμμικός συνδυασμός των επεξηγηματικών μεταβλητών σύμφωνα με τον οποίο η απόσταση των μέσων των δύο ομάδων θα είναι η μεγαλύτερη δυνατή.

Ο διαχωριστικός κανόνας του Fisher βασίζεται στη μετατροπή των χαρακτηριστικών $\mathbf{X}=(X_1,X_2,...,X_p)$ σε μονοδιάστατες βαθμολογίες μέσω μιας συνάρτησης, η οποία λέγεται διαχωριστική συνάρτηση (*discriminant function*). Οι βαθμολογίες των δύο ομάδων θα πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένες, έτσι ώστε με βάση αυτές να μπορεί να γίνει ο διαχωρισμός και η ταξινόμηση των δύο ομάδων. Έτσι λοιπόν ο Fisher (1936) πρότεινε τη χρήση γραμμικών συνδυασμών για τη δημιουργία αυτών των βαθμολογιών, χωρίς να γίνει κάποια υπόθεση για την κατανομή των ομάδων. Παρ' όλα αυτά υπέθεσε ισότητα των πινάκων συνδιακύμανσης, αφού χρησιμοποίησε τη συνδυασμένη κοινή εκτίμηση $\mathbf{S}_p$.

Ο Fisher ήταν ο πρώτος που εισήγαγε τη γραμμική διαχωριστική ανάλυση με σκοπό να βρεθεί ο συνδυασμός των μεταβλητών που διαχωρίζει καλύτερα τις δύο ομάδες των οποίων τα χαρακτηριστικών ήταν διαθέσιμα. Ο Eisenbeis (1977) όμως ανέφερε κάποια προβλήματα στον καθορισμό των «καλών» και των «κακών» πελατών στην περίπτωση που δεν υπάρχει σαφής διαχωρισμός μεταξύ τους και κάτω από την υπόθεση ότι οι πίνακες συνδιασποράς μεταξύ των δύο ομάδων είναι ίσοι. Επίσης, ο Eisenbeis (1978) άσκησε κριτική σε αυτή τη μέθοδο ισχυριζόμενος ότι ο κανόνας αυτός είναι βέλτιστος μόνο για μια μικρή κατηγορία κατανομών. Ωστόσο, οι Hand και Henley (1997) ισχυρίστηκαν ότι: «αν οι μεταβλητές ακολουθούν μια πολυμεταβλητή ελλειψοειδή κατανομή (για την οποία η κανονική κατανομή είναι ειδική περίπτωση) τότε αυτός ο γραμμικός διαχωριστικός κανόνας θα είναι βέλτιστος».

Έστω $Y = b_1X_1 + b_2X_2 + \ldots + b_pX_p = \mathbf{bX}^T$ είναι οποιοσδήποτε γραμμικός συνδυασμός των χαρακτηριστικών $\mathbf{X} = (X_1, X_2, ..., X_p)$ με αντίστοιχο διάνυσμα συντελεστών $\mathbf{b} = (b_1, b_2, \cdots, b_p)$. Ένα προφανές μέτρο διαχωρισμού των ομάδων θα μπορούσε να είναι το πόσο διαφορετικές είναι οι μέσες τιμές του Y για τις δύο διαφορετικές ομάδες των «καλών» και «κακών» πελατών στο δείγμα. Κατά συνέπεια, εξετάζεται η διαφορά μεταξύ των $E(Y|G)$ και $E(Y|B)$ και επιλέγονται οι συντελεστές b_i, $i = 1,2,...,p$ που μεγιστοποιούν τη διαφορά υπό τον περιορισμό $\sum_{i=1}^{p} b_i = 1$. Υποθέτοντας ότι οι δυο ομάδες έχουν κοινή δειγματική διασπορά ο Fisher πρότεινε ως μέτρο διαχωρισμού των ομάδων το εξής:

$$M = \frac{\text{Απόσταση μεταξύ δειγματικών μέσων των ομάδων}}{(\text{Δειγματική διασπορά της κάθε ομάδας})^{1/2}}.$$

Επισημαίνεται ότι διαιρούμε με την τετραγωνική ρίζα της δειγματικής διασποράς έτσι ώστε να μην υπάρχει εξάρτηση από την κλίμακα μέτρησης. Σκοπός της μεθόδου είναι να μεγιστοποιηθεί η ποσότητα M ή M^2 έτσι ώστε οι βαθμολογίες των δύο ομάδων να είναι όσο γίνεται πιο διαφορετικές μεταξύ τους. Εάν αλλαχτεί η μεταβλητή από Y σε cY, τότε το μέτρο M δεν αλλάζει.

Συμβολίζουμε με $\mathbf{m}_G$ και $\mathbf{m}_B$ τους δειγματικούς μέσους των ομάδων των «καλών» και των «κακών» αντίστοιχα και με $\mathbf{S}_p$ την κοινή δειγματική διασπορά.

Εάν $Y = b_1X_1 + b_2X_2 + \ldots + b_pX_p$ τότε η αντίστοιχη απόσταση διαχωρισμού των ομάδων είναι:

$$M = \mathbf{b}\frac{(\mathbf{m}_G - \mathbf{m}_B)^T}{(\mathbf{bS}_p\mathbf{b}^T)^{1/2}}.$$

Πράγματι λοιπόν έχουμε ότι

$$E(Y|G) = \mathbf{b}\cdot\mathbf{m_G}^T,\ E(Y|B) = \mathbf{b}\cdot\mathbf{m_B}^T \text{ και } Var(Y) = \mathbf{b}\cdot\mathbf{S}_p\cdot\mathbf{b}^T.$$

Επομένως, αν ο γραμμικός συνδυασμός είναι $\mathbf{bX}$ θα πρέπει να μεγιστοποιηθεί ως προς $\mathbf{b}$ η ποσότητα M. Παραγωγίζοντας την ποσότητα M ως προς $\mathbf{b}$ και εξισώνοντας την παράγωγο με το μηδέν βρίσκουμε ότι

$$\frac{(\mathbf{m}_G - \mathbf{m}_B)^T(\mathbf{bS}_p\mathbf{b}^T)^{\frac{1}{2}} - \mathbf{b}(\mathbf{m}_G - \mathbf{m}_B)^T\frac{1}{2}(\mathbf{bS}_p\mathbf{b}^T)^{-\frac{1}{2}}2\mathbf{S}_p\mathbf{b}^T}{\mathbf{bS}_p\mathbf{b}^T} = 0 \Leftrightarrow$$

$$\frac{(\mathbf{m}_G-\mathbf{m}_B)^T(\mathbf{bS}_p\mathbf{b}^T)^{\frac{1}{2}}}{\mathbf{bS}_p\mathbf{b}^T}-\frac{\mathbf{b}(\mathbf{m}_G-\mathbf{m}_B)^T(\mathbf{bS}_p\mathbf{b}^T)^{-\frac{1}{2}}\mathbf{S}_p\mathbf{b}^T}{\mathbf{bS}_p\mathbf{b}^T}=0\Leftrightarrow$$

$$\frac{(\mathbf{m}_G-\mathbf{m}_B)^T}{(\mathbf{bS}_p\mathbf{b}^T)^{\frac{1}{2}}}-\frac{\mathbf{b}(\mathbf{m}_G-\mathbf{m}_B)^T\mathbf{S}_p\mathbf{b}^T}{(\mathbf{bS}_p\mathbf{b}^T)^{\frac{3}{2}}}=0\Leftrightarrow$$

$$(\mathbf{m}_G-\mathbf{m}_B)^T(\mathbf{bS}_p\mathbf{b}^T)=(\mathbf{S}_p\mathbf{b}^T)\mathbf{b}(\mathbf{m}_G-\mathbf{m}_B)^T. \qquad (3.11)$$

Παρατηρούμε λοιπόν ότι η ποσότητα M μεγιστοποιείται για $\mathbf{b}^T=\mathbf{S}_p^{-1}(\mathbf{m}_G-\mathbf{m}_B)^T$ αφού ικανοποιεί τη σχέση (3.11). Με την παραπάνω μέθοδο αποδεικνύουμε ότι έχουμε ένα πιθανό σημείο μεγίστου. Γενικά, για να ικανοποιείται η ισότητα (3.11) θα πρέπει να ισχύει η σχέση:

$$\mathbf{b}^T\propto(\mathbf{S}_p^{-1}(\mathbf{m}_G-\mathbf{m}_B)^T). \qquad (3.12)$$

Υπολογίζουμε στη συνέχεια τη μέγιστη τιμή της ποσότητας

$$M^2=\frac{\left[\mathbf{b}(\mathbf{m}_G-\mathbf{m}_B)^T\right]^2}{\mathbf{bS}_p\mathbf{b}^T}.$$

Αν έχουμε δύο διανύσματα **a** και **u** με διάσταση $1\times p$ το καθένα, από την ανισότητα Cauchy – Schwarz θα ισχύει ότι $(\mathbf{au}^T)^2\le(\mathbf{aa}^T)(\mathbf{uu}^T)$. Αφού ο πίνακας συνδιακυμάνσεων $\mathbf{S}_p$ είναι θετικά ορισμένος, μπορούμε να θέσουμε $\mathbf{a}=\mathbf{bS_p^{1/2}}$ και $\mathbf{u}=(\mathbf{m_G}-\mathbf{m_B})\mathbf{S_p^{-1/2}}$ και η προηγούμενη ανισότητα δίνει

$$\left[\mathbf{b}(\mathbf{m_G}-\mathbf{m_B})^T\right]^2\le(\mathbf{bS_p^{1/2}S_p^{1/2}b}^T)(\mathbf{m_G}-\mathbf{m_B})\mathbf{S_p^{-1/2}S_p^{-1/2}}(\mathbf{m_G}-\mathbf{m_B})^T\Leftrightarrow$$

$$\left[\mathbf{b}(\mathbf{m_G}-\mathbf{m_B})^T\right]^2\le(\mathbf{bS_pb}^T)\left[(\mathbf{m_G}-\mathbf{m_B})\mathbf{S_p^{-1}}(\mathbf{m_G}-\mathbf{m_B})^T\right]\Leftrightarrow$$

$$M^2=\frac{\left[\mathbf{b}(\mathbf{m_G}-\mathbf{m_B})^T\right]^2}{\mathbf{bS_pb}^T}\le(\mathbf{m_G}-\mathbf{m_B})\mathbf{S_p^{-1}}(\mathbf{m_G}-\mathbf{m_B})^T.$$

Κατά συνέπεια οι συντελεστές είναι οι ίδιοι με αυτούς στην περίπτωση που έχουμε πολυμεταβλητή κανονική κατανομή με κοινή διακύμανση αλλά χωρίς να είναι απαραίτητο να ισχύει η υπόθεση της κανονικότητας. Επομένως, αυτός ο κανόνας διαχωρισμού των «καλών» και των «κακών» πελατών είναι σχετικά καλύτερος χωρίς να μας ενδιαφέρει τι κατανομή ακολουθεί το δείγμα μας. Αυτό το αποτέλεσμα ισχύει για όλες τις κατανομές επειδή το μέτρο M της απόστασης περιλαμβάνει μόνο το μέσο όρο και τη διακύμανση των κατανομών και έτσι δίνει τα ίδια αποτελέσματα για όλες τις κατανομές με τον ίδιο μέσο και την ίδια διακύμανση (Thomas (2002)).

Ο Altman (1968) ήταν ο πρώτος που εφάρμοσε τη Διαχωριστική Ανάλυση με σκοπό τη βαθμολόγηση πιστοληπτικής ικανότητας, δημιουργώντας το λεγόμενο μοντέλο Z-score, το οποίο περιλαμβάνει ένα γραμμικό συνδυασμό πολλών επεξηγηματικών μεταβλητών. Το μοντέλο αυτό χρησιμοποιείται στην περίπτωση του προβλήματος χορήγησης πίστωσης σε επιχειρήσεις. Χρησιμοποιώντας ως επεξηγηματικές μεταβλητές κάποιους χρηματοοικονομικούς δείκτες ο Altman βρήκε ότι το μοντέλο αυτό είναι εξαιρετικά ακριβές στο να προβλέπει σωστά την πτώχευση ή μη των υπό μελέτη επιχειρήσεων.

Το βασικό πλεονέκτημα της ΔΑ είναι ότι είναι μια σχετικά απλή μέθοδος και όταν εφαρμόζεται εμφανίζει πολύ καλά αποτελέσματα. Χρησιμοποιείται πολύ συχνά από χρηματοπιστωτικά ιδρύματα για σκοπούς βαθμολόγησης πιστοληπτικής ικανότητας. Το βασικότερο μειονέκτημα του είναι ότι για την εφαρμογή του μοντέλου του Altman απαιτούνται κανονικά κατανεμημένα δεδομένα, ενώ συνήθως τα δεδομένα που χρησιμοποιούνται για χορήγηση πίστωσης είναι μη κανονικά. Το μεγαλύτερο πρόβλημα προκύπτει όταν είναι απαραίτητο να ελεγχτεί η σημαντικότητα μεμονωμένων μεταβλητών και δεν ισχύει η υπόθεση της κανονικότητας, οπότε δεν μπορούμε να προχωρήσουμε σε στατιστική συμπερασματολογία. Οι γραμμικοί συνδυασμοί που παράγονται από τη μέθοδο της ΔΑ σύμφωνα με το υπόδειγμα του Altman Z-score έχουν την εξής μορφή:

$$Z = b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

όπου Z είναι η διαχωριστική βαθμολογία (*Z-score*), και b_i είναι οι αντίστοιχοι συντελεστές των χρηματοοικονομικών δεικτών X_i (μεταβλητές) για κάθε $i = 1,2,...,p$. Για παράδειγμα, ο Altman (1968) παίρνοντας ως δείγμα 66 επιχειρήσεις από τις οποίες οι 33 είχαν χρεοκοπήσει συγκέντρωσε στοιχεία για τις επιχειρήσεις αυτές και προσπάθησε να τις κατατάξει σε δύο ομάδες. Το προτεινόμενο υπόδειγμα περιελάμβανε 5 χρηματοοικονομικούς δείκτες που θεωρούνταν ως οι σημαντικότεροι και είχαν άμεση σχέση με τη βιωσιμότητα της κάθε επιχείρησης. Το υπόδειγμα Z-score του Altman ήταν το εξής:

$$Z = 0,012X_1 + 0,014X_2 + 0,033X_3 + 0,006X_4 + 0,999X_5 \qquad (3.13)$$

όπου,

$X_1 = \text{Κεφάλαιο Κίνησης}/\text{Σύνολο Ενεργητικού}$

$X_2 = \text{Παρακρατηθέντα Κέρδη}/\text{Σύνολο Ενεργητικού}$

$X_3 = \text{Κέρδη προ φόρων και τόκων}/\text{Σύνολο Ενεργητικού}$

$X_4 = \text{Ίδια Κεφάλαια}/\text{Σύνολο Υποχρεώσεων}$

$X_5 = \Pi\omega\lambda\acute{\eta}\sigma\epsilon\iota\varsigma/\Sigma\acute{\upsilon}\nu o\lambda o\,E\nu\epsilon\rho\gamma\eta\tau\iota\kappa o\acute{\upsilon}$.

Με βάση το μοντέλο (3.13) υπολογίζεται την τιμή της βαθμολογίας Z για κάθε επιχείρηση. Όσο υψηλότερη είναι η τιμή του Z τόσο χαμηλότερη είναι η κατάταξη της επιχείρησης με βάση την πιθανότητα χρεοκοπίας. Συνεπώς μικρές ή αρνητικές τιμές του Z να συνδυάζονται με υψηλή πιθανότητα αθέτησης των υποχρεώσεων. Σύμφωνα με τον Altman κάθε επιχείρηση που συγκεντρώνει βαθμολογία $Z < 1{,}81$ δεν πρέπει να της χορηγείται δάνειο γιατί έχει μεγάλη πιθανότητα να πτωχεύσει, ενώ αν $Z \geq 1{,}81$ τότε μπορεί να θεωρηθεί αξιόπιστη. Η τιμή $Z = 1{,}81$ προέκυψε από το μέσο όρο των υγειών και των μη υγειών επιχειρήσεων.

3.7 Γραμμική παλινδρόμηση

Μια εναλλακτική μέθοδος για τη βαθμολόγηση πιστοληπτικής ικανότητας είναι η γραμμική παλινδρόμηση (ΓΠ). Σκοπός αυτής της μεθόδου είναι να βρεθεί ο βέλτιστος γραμμικός συνδυασμός των χαρακτηριστικών

$$b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p = \mathbf{b}^* \mathbf{X}^{*T}$$

όπου $\mathbf{b}^* = (b_0, b_1, b_2, ..., b_p)$ και $\mathbf{X}^* = (1, X_1, X_2, ..., X_p)$. Ο παραπάνω γραμμικός συνδυασμός περιγράφει κατά κάποιον τρόπο την πιθανότητα αθέτησης των υποχρεώσεων. Επομένως, αν $p_i = P(G \mid \mathbf{x}_i)$ είναι η πιθανότητα ο i υποψήφιος του δείγματος να μπορέσει να ανταποκριθεί στις υποχρεώσεις του, δηλαδή να είναι «καλός», σκοπός είναι να βρεθεί το κατάλληλο $\mathbf{b}^*$ που προσεγγίζει περισσότερο τη σχέση:

$$p_i = b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} \qquad (3.14)$$

όπου $x_{i0} = 1$ για κάθε $i = 1,2,...,n$, με ένα σφάλμα ε_i. Υποθέτουμε ότι n_G είναι ο αριθμός των «καλών» πελατών του δείγματος και ότι αυτοί είναι οι πρώτοι n_G του δείγματος οπότε θα έχουμε ότι $p_i = 1$ για κάθε $i = 1,2,...,n_G$. Ο αριθμός των πελατών του δείγματος που απομένουν είναι n_B με $i = n_G + 1, n_G + 2, ..., n_G + n_B$ οι οποίοι αποτελούν τους «κακούς»

πελάτες και γι' αυτούς ισχύει ότι $p_i = 0$. Αφού n είναι ο συνολικός αριθμός του δείγματος θα ισχύει $n_G + n_B = n$.

Στην περίπτωση της γραμμικής παλινδρόμησης, επιλέγουμε το διάνυσμα των συντελεστών $\mathbf{b}^*$ που ελαχιστοποιεί το Μέσο Τετραγωνικό Σφάλμα (MSE) για τις διαφορές ανάμεσα στο δεξί και το αριστερό μέλος της σχέσης (3.14). Αυτό σημαίνει ότι πρέπει να ελαχιστοποιηθεί η ποσότητα:

$$\begin{aligned}\sum_{i=1}^{n} \varepsilon_i{}^2 &= \sum_{i=1}^{n_G}\left(p_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2 + \sum_{i=n_G+1}^{n}\left(p_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2 = \\ &= \sum_{i=1}^{n_G}\left(1 - \sum_{j=0}^{p} b_j x_{ij} \right)^2 + \sum_{i=n_G+1}^{n}\left(\sum_{j=0}^{p} b_j x_{ij} \right)^2 .\end{aligned} \tag{3.15}$$

Σε μορφή πινάκων η σχέση (3.14) μπορεί να γραφεί ως εξής:

$$\begin{pmatrix} \mathbf{I}_G & \mathbf{X}_G \\ \mathbf{I}_B & \mathbf{X}_B \end{pmatrix}\begin{pmatrix} b_0 \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_G \\ \mathbf{0} \end{pmatrix} \quad \text{ή} \quad \mathbf{Y}\mathbf{b}^T = \mathbf{d}^T$$

όπου, $\mathbf{I}_G$ και $\mathbf{I}_B$ είναι δύο διανύσματα $1 \times n_G$ και $1 \times n_B$ αντίστοιχα με όλα τα στοιχεία 1,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{I}_G & \mathbf{X}_G \\ \mathbf{I}_B & \mathbf{X}_B \end{pmatrix}$$

είναι ένας $n \times (p+1)$ πίνακας,

$$\mathbf{X}_G = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_G 1} & x_{n_G 2} & \cdots & x_{n_G p} \end{pmatrix}$$

είναι ένας $n_G \times p$ πίνακας,

$$\mathbf{X}_G = \begin{pmatrix} x_{(n_G+1)1} & x_{(n_G+1)2} & \cdots & x_{(n_G+1)p} \\ x_{(n_G+2)1} & x_{(n_G+2)2} & \cdots & x_{(n_G+2)p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

είναι ένας $n_B \times p$ πίνακας και

$$\mathbf{d}^T = \begin{pmatrix} \mathbf{I}_G \\ 0 \end{pmatrix}.$$

Σε μορφή πινάκων, για να βρεθούν οι συντελεστές της γραμμικής παλινδρόμησης θα πρέπει να βρεθεί το $\mathbf{b}$ που ελαχιστοποιεί την ποσότητα

$$(\mathbf{Y}\mathbf{b}^T - \mathbf{d}^T)^T(\mathbf{Y}\mathbf{b}^T - \mathbf{d}^T). \tag{3.16}$$

Για να ελαχιστοποιηθεί η ποσότητα (3.16) αρχικά την παραγωγίζουμε ως προς $\mathbf{d}$ και έπειτα εξισώνουμε την παράγωγο με το μηδέν. Δηλαδή, έχουμε:

$$\mathbf{Y}^T(\mathbf{Y}\mathbf{b}^T - \mathbf{d}^T) = \mathbf{0} \Leftrightarrow \mathbf{Y}^T\mathbf{Y}\mathbf{b}^T = \mathbf{Y}^T\mathbf{d}^T \tag{3.17}$$

όπου

$$\mathbf{Y}^T\mathbf{d}^T = \begin{pmatrix} \mathbf{I}_G & \mathbf{I}_B \\ \mathbf{X}_G & \mathbf{X}_B \end{pmatrix}\begin{pmatrix} \mathbf{I}_G \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} n_G \\ n_G\mathbf{m}_G \end{pmatrix}$$

και

$$\mathbf{Y}^T\mathbf{Y} = \begin{pmatrix} \mathbf{I}_G & \mathbf{I}_B \\ \mathbf{X}_G & \mathbf{X}_B \end{pmatrix}\begin{pmatrix} \mathbf{I}_G & \mathbf{X}_G \\ \mathbf{I}_B & \mathbf{X}_B \end{pmatrix} = \begin{pmatrix} n & n_G\mathbf{m}_G + n_B\mathbf{m}_B \\ n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T & \mathbf{X}_G^T\mathbf{X}_G + \mathbf{X}_B^T\mathbf{X}_B \end{pmatrix}. \tag{3.18}$$

Εάν θεωρήσουμε ότι οι αναμενόμενες τιμές είναι ίδιες με τις πραγματικές τότε παίρνουμε

$$\mathbf{X}_\mathbf{G}^T\mathbf{X}_\mathbf{G} + \mathbf{X}_\mathbf{B}^T\mathbf{X}_\mathbf{B} = nE(X_iX_j) = nCov(X_i, X_j) + n_G\mathbf{m}_\mathbf{G}\mathbf{m}_\mathbf{G}^T + n_B\mathbf{m}_\mathbf{B}\mathbf{m}_\mathbf{B}^T.$$

Εάν ο δειγματικός πίνακας συνδιακυμάνσεων είναι ο $\mathbf{S}_p$ τότε θα ισχύει

$$\mathbf{X}_\mathbf{G}^T\mathbf{X}_\mathbf{G} + \mathbf{X}_\mathbf{B}^T\mathbf{X}_\mathbf{B} = n\mathbf{S}_p + n_G\mathbf{m}_\mathbf{G}\mathbf{m}_\mathbf{G}^T + n_B\mathbf{m}_\mathbf{B}\mathbf{m}_\mathbf{B}^T. \tag{3.19}$$

Αντικαθιστώντας τη σχέση (3.19) στην (3.18) και χρησιμοποιώντας τη σχέση (3.17) προκύπτει το σύστημα εξισώσεων

$$\begin{cases} nw_0 + (n_G\mathbf{m}_G + n_B\mathbf{m}_B)\mathbf{b}^T = n_G \\ (n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T)b_0 + (n\mathbf{S}_p + n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T)\mathbf{b}^T = n_G\mathbf{m}_G^T \end{cases}$$

από όπου προκύπτει

$$\begin{cases} b_0 = \dfrac{n_G - (n_G\mathbf{m}_G + n_B\mathbf{m}_B)\mathbf{b}^T}{n} \\ (n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T)b_0 + (n\mathbf{S}_p + n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T)\mathbf{b}^T = n_G\mathbf{m}_G^T. \end{cases} \tag{3.20}$$

Αντικαθιστώντας την πρώτη σχέση της (3.20) στη δεύτερη έχουμε

$$\left(\frac{n_Gn_B}{n}\right)(\mathbf{m}_G - \mathbf{m}_B)\mathbf{b}^T + n\mathbf{S}_p\mathbf{b}^T = \left(\frac{n_Gn_B}{n}\right)(\mathbf{m}_G - \mathbf{m}_B)^T.$$

και επομένως,

$$\mathbf{S}_p\mathbf{b}^T = c(\mathbf{m}_G - \mathbf{m}_B)^T$$

απ' όπου βρίσκουμε

$$\mathbf{b}^T = c\mathbf{S}_p^{-1}(\mathbf{m}_G - \mathbf{m}_B)^T. \quad (3.21)$$

Η σχέση (3.21) δίνει την καλύτερη δυνατή επιλογή για τους συντελεστές της γραμμικής παλινδρόμησης, δηλαδή για το διάνυσμα $\mathbf{b} = (b_1, b_2, ..., b_p)$. Παρατηρούμε ότι η σχέση (3.21) είναι ίδια με τη σχέση (3.12). Αυτό σημαίνει ότι οι συντελεστές του μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας με τη μέθοδο των ελάχιστων τετραγώνων της γραμμικής παλινδρόμησης είναι ίδιοι με το διάνυσμα $\mathbf{b}$ που προέκυψε και από τη διαχωριστική συνάρτηση του Fisher.

Στην παρούσα ανάλυση έχουμε την προφανή περίπτωση που η αριστερή πλευρά της εξίσωσης παλινδρόμησης παίρνει την τιμή 1 όταν ο πελάτης είναι «καλός» και την τιμή 0 όταν ο πελάτης είναι «κακός». Για αυτούς τους πελάτες έχει βρεθεί κάποιο συγκεκριμένο σύνολο συντελεστών το οποίο το συμβολίζουμε με $\mathbf{b}^*(0,1)$. Στην περίπτωση όμως που πάρουμε διαφορετικές τιμές έτσι ώστε οι «καλοί» πελάτες να παίρνουν την τιμή g και οι «κακοί» να παίρνουν την τιμή b, τότε συμβολίζουμε τους συντελεστές της γραμμικής παλινδρόμησης με $\mathbf{b}^*(g,b)$ και διαφέρουν μόνο ως προς τον σταθερό όρο b_0 αφού ισχύει ότι:[7]

$$\mathbf{b}^*(g,b) = b + (g-b)\mathbf{b}^*(1,0).$$

3.8 Λογιστική παλινδρόμηση

Η χρήση της μεθόδου της γραμμικής παλινδρόμησης για την κατάταξη των υποψήφιων πελατών σε «καλούς» και «κακούς» παρουσιάζει ένα αδύναμο σημείο. Το δεξί μέλος της σχέσης (3.14) μπορεί να πάρει οποιαδήποτε τιμή από $-\infty$ μέχρι $+\infty$ ενώ το αριστερό μέλος της σχέσης παριστάνει πιθανότητα και μπορεί να πάρει τιμές μόνο μεταξύ 0 και 1. Θα ήταν καλύτερα λοιπόν εάν στο αριστερό μέλος της σχέσης (3.14) χρησιμοποιούσαμε μια συνάρτηση του p_i που θα μπορούσε να έχει ένα ευρύτερο πεδίο τιμών. Μια τέτοια συνάρτηση είναι για παράδειγμα ο λογάριθμος της σχετικής πιθανότητας (*odds*). Η μέθοδος

[7] Βλέπε Thomas et al (2002).

που προκύπτει με χρήση του λόγου πιθανότητας ως μεταβλητή απόκρισης σε ένα γραμμικό μοντέλο ονομάζεται λογιστική παλινδρόμηση (ΛΓ) και ο Wiginton (1980) ήταν ένας από τους πρώτους που χρησιμοποίησαν τη μέθοδο αυτή με σκοπό τη βαθμολόγηση της πιστοληπτικής ικανότητας.

Έστω Y_i είναι η δίτιμη μεταβλητή απόκρισης του i πελάτη του δείγματος, δηλαδή

$$Y_i = \begin{cases} 1, & \text{εάν ο πελάτης } i \text{ είναι καλός} \\ 0, & \text{διαφορετικά} \end{cases} \qquad \text{για κάθε } i = 1,2,...,n$$

και $p_i = P(Y_i = 1) = P(G \mid \mathbf{x}_i)$ είναι η πιθανότητα επιτυχίας, δηλαδή η πιθανότητα ο πελάτης i με τα χαρακτηριστικά $\mathbf{x}_i$ να είναι καλός. Η συνάρτηση σύνδεσης (*link function*) στο μοντέλο λογιστικής παλινδρόμησης είναι η

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \ln(odds(G \mid x)),$$

δηλαδή το μοντέλο της λογιστικής παλινδρόμησης συνδέει γραμμικά το λογάριθμο της σχετικής πιθανότητας της δίτιμης απόκρισης που ονομάζεται *logit* με τις επεξηγηματικές μεταβλητές. Επομένως, θα έχουμε

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p = s(\mathbf{x}), \quad i = 1,2,...,n \qquad (3.22)$$

όπου

$$s(\mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p = \mathbf{b} \cdot \mathbf{x}^{\mathbf{T}}.$$

Αφού η ποσότητα $\frac{p_i}{1-p_i}$ παίρνει τιμές από 0 έως ∞ τότε η ποσότητα $\ln\left(\frac{p_i}{1-p_i}\right)$ θα παίρνει τιμές από $-\infty$ έως $+\infty$. Η σχέση (3.22) μπορεί να γραφεί ισοδύναμα ως εξής

$$\frac{p_i}{1-p_i} = e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}} \Leftrightarrow p_i = e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}}(1-p_i) \Leftrightarrow p_i + e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}} \cdot p_i = e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}} \Leftrightarrow p_i = \frac{e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}}}{1+e^{\mathbf{b}\cdot\mathbf{x}^{\mathbf{T}}}}, \qquad (3.23)$$

ή ακόμη

$$p_i = \frac{e^{s(\mathbf{x})}}{1+e^{s(\mathbf{x})}}.$$

Αν υποθέσουμε ότι η κατανομή των χαρακτηριστικών των «καλών» και των «κακών» πελατών είναι πολυμεταβλητή κανονική τότε θα ικανοποιείται η έκφραση (3.22) της λογιστικής παλινδρόμησης. Πράγματι αν υποθέσουμε ότι η μέση τιμή των παρατηρήσεων

που ανήκουν στην ομάδα των «καλών» και των «κακών» είναι $\boldsymbol{\mu}_G$ και $\boldsymbol{\mu}_B$ αντίστοιχα και $\boldsymbol{\Sigma}$ ο κοινός πίνακας συνδιασποράς τότε η αντίστοιχη συνάρτηση πυκνότητας είναι

$$f(\mathbf{x}\,|\,G)=\frac{1}{(2\pi)^{p/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_G)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_G)^T}{2}\right).$$

Αν p_G είναι το ποσοστό των «καλών» πελατών και p_B το ποσοστό των «κακών» πελατών τότε ο λογάριθμος της σχετικής πιθανότητας για τον i πελάτη που έχει χαρακτηριστικά $\mathbf{x}$ είναι

$$\ln\left(\frac{p_i}{1-p_i}\right)=\ln\left(\frac{p_G f(\mathbf{x}\,|\,G)}{p_B f(\mathbf{x}\,|\,B)}\right)=\mathbf{x}\cdot\boldsymbol{\Sigma}^{-1}2(\boldsymbol{\mu}_{\mathrm{B}}-\boldsymbol{\mu}_G)^T+(\boldsymbol{\mu}_G\cdot\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_G^T+\boldsymbol{\mu}_B\cdot\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B^T)+\ln\left(\frac{p_G}{p_B}\right).$$

Αφού η παραπάνω σχέση είναι γραμμικός συνδυασμός του $\mathbf{x}$ ικανοποιεί την υπόθεση της λογιστικής παλινδρόμησης.

Υπάρχουν και άλλες κατανομές που ικανοποιούν τη λογιστική υπόθεση περιλαμβάνοντας ακόμα και αυτές τις κατανομές που δεν οδηγούν σε γραμμικές διαχωριστικές συναρτήσεις εάν μπορεί να εφαρμοστεί ο κανόνας του Bayes. Ας πάρουμε ως παράδειγμα την περίπτωση όπου όλες οι επεξηγηματικές μεταβλητές είναι δίτιμες και ανεξάρτητες μεταξύ τους. Αυτό σημαίνει ότι:

$$P(X_i=1\,|\,G)=p_G(i)\ \text{ και }\ P(X_i=0\,|\,G)=1-p_G(i),$$

$$P(X_i=1\,|\,B)=p_B(i)\ \text{ και }\ P(X_i=0\,|\,B)=1-p_B(i).$$

Επομένως, το X_i χαρακτηριστικό ακολουθεί κατανομή Bernoulli, δηλαδή $X_i\sim B(p_G(i))$ και p_G και p_B είναι οι εκ των προτέρων πιθανότητες των «καλών» και των «κακών» πελατών στον πληθυσμό. Τότε έχουμε:

$$P(G\,|\,\mathbf{x})=\frac{P(\mathbf{x}\,|\,G)p_G}{P(\mathbf{x})}=\frac{\prod_{i=1}^{p}p_G(i)^{x_i}(1-p_G(i))^{1-x_i}p_G}{P(\mathbf{x})}$$

και

$$P(B\,|\,\mathbf{x})=\frac{P(\mathbf{x}\,|\,B)p_B}{P(\mathbf{x})}=\frac{\prod_{i=1}^{p}p_B(i)^{x_i}(1-p_B(i))^{1-x_i}p_B}{P(\mathbf{x})}.$$

Οπότε,

$$\ln\left(\frac{P(G\,|\,\mathbf{x})}{P(B\,|\,\mathbf{x})}\right)=\sum_{i=1}^{p}x_i(\ln(1-p_G(i))-\ln(1-p_B(i)))+$$

$$+\sum_{i=1}^{p}(1-x_i)(\ln(1-p_G(i))-\ln(1-p_B(i)))+\ln\left(\frac{p_G}{p_B}\right),$$

δηλαδή,

$$\ln\left(\frac{P(G\,|\,\mathbf{x})}{P(B\,|\,\mathbf{x})}\right)=\sum_{i=1}^{p}x_i\left(\ln\left(\frac{p_G(i)(1-p_B(i))}{p_B(i)(1-p_G(i))}\right)\right)+\sum_{i=1}^{p}\left(\ln\left(\frac{1-p_G(i)}{1-p_B(i)}\right)\right)+\ln\left(\frac{p_G}{p_B}\right). \quad (3.24)$$

Παρατηρούμε ότι η σχέση (3.24) είναι ισοδύναμη με τη σχέση (3.22) και επομένως ικανοποιεί την υπόθεση της λογιστικής παλινδρόμησης..

Στην περίπτωση της λογιστικής παλινδρόμησης για τον υπολογισμό των συντελεστών **b** χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας. Για να μπορεί να κατασκευαστεί η σχέση (3.22) της λογιστικής παλινδρόμησης πρέπει πρώτα να επιλεγεί ένα δείγμα n παλαιότερων πελατών και να καταγραφούν οι τιμές των πιο σημαντικών χαρακτηριστικών σε αυτούς. Τελικά έχουμε ένα σύνολο τιμών $(\mathbf{x}_i, y_i)$ για κάθε έναν από τους n πελάτες, όπου $\mathbf{x}_i=(x_{1i},x_{2i},...,x_{pi})$ είναι οι τιμές των χαρακτηριστικών για τον i πελάτη και $y_i=1$ εάν ο πελάτης i ήταν «καλός» και $y_i=0$, εάν ο πελάτης i ήταν «κακός». Υποθέτουμε ότι κάθε πελάτης δεν έχει ακριβώς τις ίδιες τιμές για όλα τα χαρακτηριστικά με κάποιον άλλο και αυτό γιατί ο αριθμός των χαρακτηριστικών είναι συνήθως πολύ μεγάλος. Άρα, η μεταβλητή Y_i ακολουθεί διωνυμική κατανομή με πιθανότητα p_i, δηλαδή $Y_i \sim B(1,p_i)$. Η πιθανοφάνεια του δείγματος είναι ανάλογη του γινομένου των n ανεξάρτητων διωνυμικών, δηλαδή είναι

$$L(\mathbf{b})=\prod_{i=1}^{n}p_i^{\,y_i}(1-p_i)^{1-y_i}. \quad (3.25)$$

Αντικαθιστώντας στη σχέση (3.25) την (3.23) έχουμε

$$L(\mathbf{b})=\prod_{i=1}^{n}\left(\frac{e^{\mathbf{b}\mathbf{x}_i^T}}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right)^{y_i}\left(\frac{1}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right)^{1-y_i}. \quad (3.26)$$

Αρκεί λοιπόν να μεγιστοποιήσουμε το λογάριθμο της παραπάνω ποσότητας. Λογαριθμίζοντας και τα δύο μέλη της σχέσης (3.26) έχουμε

$$\ln(L(\mathbf{b}))=\sum_{i=1}^{n}y_i\ln\left(\frac{e^{\mathbf{b}\mathbf{x}_i^T}}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right)+\sum_{i=1}^{n}(1-y_i)\ln\left(\frac{1}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right). \quad (3.27)$$

Παραγωγίζοντας την έκφραση (3.27) ως προς b_j, $j=1,2,...,p$ και εξισώνοντας τις παραγώγους με το μηδέν, προκύπτει ότι η συνάρτηση πιθανοφάνειας γίνεται μέγιστη όταν ικανοποιούνται οι εξισώσεις

$$\sum_{i=1}^{n}\left(y_i-\left(\frac{e^{\mathbf{b}\mathbf{x}_i^T}}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right)\right)=0$$

και

$$\sum_{i=1}^{n}x_{ij}\left(y_i-\left(\frac{e^{\mathbf{b}\mathbf{x}_i^T}}{1+e^{\mathbf{b}\mathbf{x}_i^T}}\right)\right)=0 \quad \text{για κάθε} \quad j=1,2,...,p.$$

Για να λυθούν οι παραπάνω εξισώσεις χρησιμοποιείται η επαναληπτική μέθοδος Newton - Raphson. Η εξέλιξη των σύγχρονων υπολογιστικών συστημάτων έχει καταστήσει την εφαρμογή αυτής της μεθόδου πολύ απλή ακόμα και όταν διαθέτουμε πολύ μεγάλα δείγματα.

Οι εκτιμήσεις των συντελεστών b_i είναι πολύ σημαντικές. Αν μετά από τη διαδικασία της αρχικής ταξινόμησης όλες οι επεξηγηματικές μεταβλητές είναι δίτιμες, τότε οι συντελεστές b_i δίνουν τη βαθμολογία για την αντίστοιχη μεταβλητή. Όμως, εάν για τη δόμηση του μοντέλου της λογιστικής παλινδρόμησης οι μεταβλητές που χρησιμοποιούνται έχουν μετασχηματιστεί με βάση τη μέθοδο των βαρών ένδειξης, τότε η βαθμολογία για την ιδιότητα x_i του i χαρακτηριστικού είναι η εκτίμηση του συντελεστή b_i πολλαπλασιασμένη με το βάρος ένδειξης που αντιστοιχεί σε αυτή τη μεταβλητή.

Όταν κατασκευάζεται ένα μοντέλο λογιστικής παλινδρόμησης, ένα πολύ σημαντικό ζήτημα είναι να εντοπιστεί το βέλτιστο υποσύνολο επεξηγηματικών μεταβλητών χρησιμοποιώντας κάποια βηματική διαδικασία (*backward* ή *forward elimination*) ή θα πρέπει να εξεταστεί για κάθε μεταβλητή αν η ύπαρξή της συνεισφέρει πραγματικά στην προβλεπτική αξία του μοντέλου. Υποθέτοντας ότι οι εκτιμήσεις $\hat{b}_i$ είναι κανονικά κατανεμημένες (εάν υπάρχει μεγάλο μέγεθος δείγματος η υπόθεση αυτή είναι αρκετά ρεαλιστική), η τυπική απόκλιση $SE(\hat{b}_i)$ μπορεί να χρησιμοποιηθεί για την κατασκευή ενός διαστήματος εμπιστοσύνης του $\hat{b}_i$ της μορφής:

$$\hat{b}_i \pm Z_{a/2}\,SE(\hat{b}_i).$$

Για να ελεγχτεί αν ένα χαρακτηριστικό X_i συνεισφέρει στο μοντέλο λογιστικής παλινδρόμησης ή όχι χρησιμοποιούμε τη στατιστική συνάρτηση του Wald, $W = \frac{\hat{b}_i}{SE(\hat{b}_i)}$ και προχωράμε στον εξής έλεγχο:

$$H_0: b_i = 0 \text{ κατά της εναλλακτικής υπόθεσης } H_1: b_i \neq 0$$

Η ποσότητα W^2 ακολουθεί ασυμπτωτικά χ^2 κατανομή με 1 βαθμό ελευθερίας. Επομένως, αν $W^2 > \chi^2_{1,a}$ απορρίπτουμε τη μηδενική υπόθεση και η μεταβλητή X_i παραμένει στο μοντέλο. Παρ' όλα αυτά, ο έλεγχος του Wald μπορεί να αποδειχτεί αναξιόπιστος στην περίπτωση που υπάρχει μεγάλη επίδραση μεταξύ του χαρακτηριστικού και της βαθμολογίας (βλεπε Menard (2002)).

Ένας άλλος τρόπος για να ελεγχθεί αν μια μεταβλητή χρειάζεται να βρίσκεται στο μοντέλο ή όχι είναι να ελεγχθεί πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα με και χωρίς την υπό εξέταση μεταβλητή. Ο συνηθέστερος τρόπος για να ελεγχθεί αυτή η υπόθεση είναι χρησιμοποιώντας το στατιστικό του λογαρίθμου του λόγου πιθανοφανειών (*LR Statistic*) το οποίο δείχνει πόσο καλά προσαρμόζεται στα δεδομένα το μοντέλο που έχουμε επιλέξει σε σχέση με το κορεσμένο, το οποίο περιέχει όλες τις παραμέτρους και προσαρμόζεται σχεδόν τέλεια στα δεδομένα. Το στατιστικό που χρησιμοποιείται για τον έλεγχο του λόγου πιθανοφανειών καλείται και στατιστικό απόκλισης (*deviance*) X^2 και είναι ίσο με

$$LR = -2\ln\left(\frac{L(\text{εξεταζόμενο μοντέλου})}{L(\text{κορεσμένο μοντέλο})}\right) =$$

$$= 2\ln(L(\text{κορεσμένο μοντέλο})) - 2\ln(L(\text{εξεταζόμενο μοντέλου}))$$

όπου $L(\text{κορεσμένο μοντέλο}) = 1$.

Το στατιστικό LR ακολουθεί κατανομή χ^2 με βαθμούς ελευθερίας:

$$df(\text{κορεσμένο μοντέλο}) - df(\text{εξεταζόμενο μοντέλο}).$$

Γενικά, όσο πιο μικρή είναι η απόκλιση ενός μοντέλου, τόσο πιο κοντά είναι στο κορεσμένο μοντέλο, και αυτό παρέχει ένδειξη καλής προσαρμογής. Με το ίδιο στατιστικό μπορούμε να ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά στην απόκλιση ενός μοντέλου που περιέχει ένα συγκεκριμένο χαρακτηριστικό και ενός μοντέλου που δεν το περιέχει. Ο Agresti (1996) υποστήριξε ότι ο έλεγχος λόγου πιθανοφανειών είναι πιο αξιόπιστος για μικρό μέγεθος δείγματος. Επομένως, ένας εναλλακτικός τρόπος ελέγχου της σημαντικότητας μιας μεταβλητής. Έχουμε:

$$LR = -2\ln\left(\frac{L(\text{μοντέλο χωρίς τη μεταβλητή})}{L(\text{μοντέλο με τη μεταβλητή})}\right) =$$

$$= 2\ln(L(\text{μοντέλο με τη μεταβλητή})) - 2\ln(L(\text{μοντέλο χωρίς τη μεταβλητή})).$$

Εδώ το στατιστικό ακολουθεί χ^2 κατανομή με ένα βαθμό ελευθερίας. Αυτό το στατιστικό βοηθάει να αποφασιστεί αν το υπό μελέτη χαρακτηριστικό θα παραμείνει στο μοντέλο ή όχι. Επιπλέον, χρησιμοποιώντας αυτό το στατιστικό μπορούν να κατασκευαστούν κατά βήματα μέθοδοι όπως για παράδειγμα να ξεκινάμε με ένα μοντέλο και σε κάθε βήμα να εισάγεται στο μοντέλο εκείνο το χαρακτηριστικό που έχει τη μεγαλύτερη στατιστική σημαντικότητα.[8]

Αφού εκτιμηθεί το μοντέλο λογιστικής παλινδρόμησης, η πιθανότητα που προκύπτει για κάθε νέο υποψήφιο δανειολήπτη με βάση τις τιμές των χαρακτηριστικών του, καθορίζει την απόφαση παροχής πίστωσης. Στην πραγματικότητα, αυτό που εκτιμάται είναι η βαθμολογία του κάθε υποψήφιου που τον κατατάσσει σε μια κατηγορία κινδύνου. Εάν αυτή η βαθμολογία είναι μεγαλύτερη από το σημείο αποκοπής τότε ο πελάτης κατατάσσεται ως «καλός», ενώ διαφορετικά ως «κακός». Για να προσδιοριστεί επακριβώς το καταλληλότερο σημείο αποκοπής μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι επικύρωσης που περιγράφονται στο επόμενο κεφάλαιο.

Μια εναλλακτική προσέγγιση στο μοντέλο της λογιστικής παλινδρόμησης είναι το **κανονικό μοντέλο πιθανότητας (*probit model*)** το οποίο χρησιμοποιήθηκε για πρώτη φορά στη βαθμολόγηση πιστοληπτικής ικανότητας από τους Grablowsky και Talley (1981). Σε αυτό το μοντέλο η συνάρτηση σύνδεσης είναι η $\Phi^{-1}(p_i)$ για $i = 1,2,...,n$, όπου $\Phi(x)$ είναι η αθροιστική συνάρτηση της τυπικής κανονικής κατανομής

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy.$$

Επομένως το κανονικό μοντέλο πιθανότητας έχει τη μορφή:

$$\text{probit}(p_i) = \Phi^{-1}(p_i) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p, \qquad i = 1,2,...,n.$$

Όπως προαναφέραμε, το p_i είναι πιθανότητα και παίρνει τιμές μεταξύ του 0 και 1, οπότε η $\Phi^{-1}(p_i)$ παίρνει τιμές από $-\infty$ έως ∞ και επομένως έχουμε και πάλι ένα ευρύτερο πεδίο τιμών για να μπορέσουμε να βελτιώσουμε το μοντέλο της γραμμικής παλινδρόμησης.

[8] Περισσότερες πληροφορίες για τις προϋποθέσεις και τη στατιστική συμπερασματολογία ενός μοντέλου λογιστικής παλινδρόμησης μπορούν να βρεθούν στο βιβλίο των Hosmer and Lemeshow S. (2000).

3.9 Δέντρα ταξινόμησης

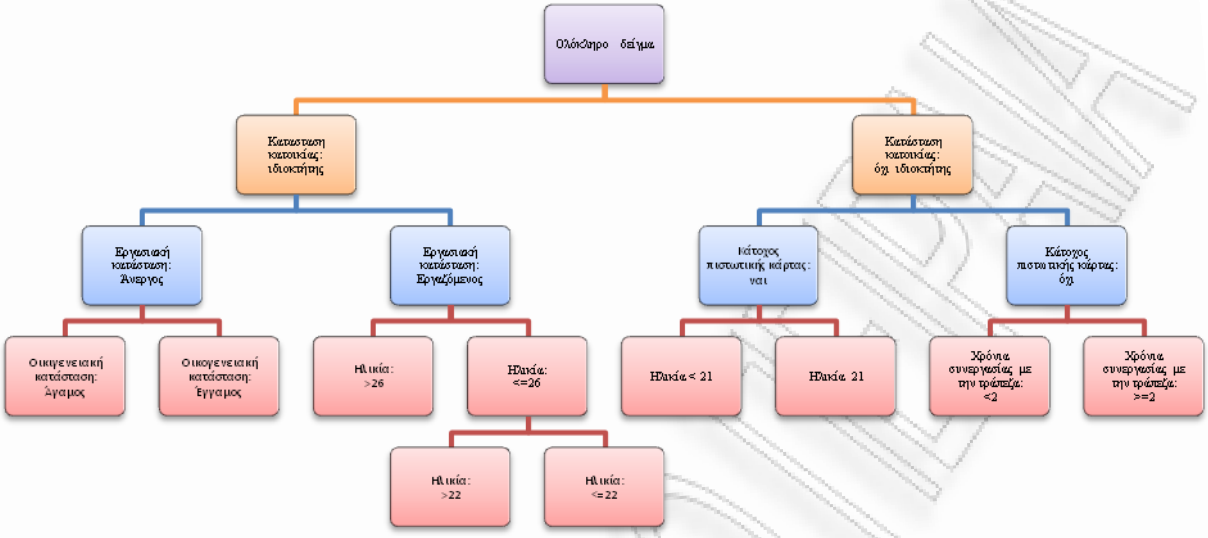
Μια διαφορετική προσέγγιση για το διαχωρισμό των «καλών» και των «κακών» πελατών είναι η ιδέα των δέντρων ταξινόμησης (ΔΤ). Αυτές τις μεθόδους τις ονομάζουμε αλλιώς αλγορίθμους επαναλαμβανόμενης διάσπασης (*Recursive Partitioning Algorithms* - RPA). Η βασική ιδέα των δέντρων ταξινόμησης είναι να διαχωριστεί το σύνολο των απαντήσεων της αίτησης χορήγησης πίστωσης σε διαφορετικά σύνολα και έπειτα να αξιολογηθεί κάθε ένα από τα σύνολα αυτά ως «καλό» ή «κακό» ανάλογα με το τι χαρακτηρίζει τη πλειοψηφία των πελατών που ανήκουν σε αυτό. Η ιδέα αναπτύχθηκε γενικά για τα προβλήματα ταξινόμησης από τους Breiman και Friedman (1984) οι οποίοι περιέγραψαν διάφορες στατιστικές εφαρμογές. Η χρήση των δέντρων ταξινόμησης για τη βαθμολόγηση της πιστοληπτικής ικανότητας προτάθηκε από τους Makowski (1985) και Coffman (1986).

Η μέθοδος ξεκινάει με όλες τις παρατηρήσεις του δείγματος (σύνολο δεδομένων A). Αρχικά χωρίζεται σε δύο υποσύνολα ανάλογα με τα χαρακτηριστικά τους (π.χ. κατάσταση κατοικίας: ιδιοκτήτης) έτσι ώστε εξετάζοντας το δείγμα των προηγούμενων υποψηφίων, αυτά τα δύο νέα υποσύνολα των ιδιοτήτων της αίτησης να είναι πολύ πιο ομοιογενή όσον αφορά κίνδυνο αθέτησης υποχρεώσεων υποψηφίων σε σχέση με το αρχικό σύνολο. Κάθε ένα από αυτά τα δύο υποσύνολα διασπάται πάλι σε δύο νέα πιο ομοιογενή υποσύνολα και η διαδικασία επαναλαμβάνεται. Γι' αυτό το λόγο η διαδικασία καλείται και επαναλαμβανόμενη διάσπαση και σταματάει όταν τα υποσύνολα που έχουν σχηματιστεί ικανοποιούν τα κριτήρια που χρειάζονται ώστε αυτά τα υποσύνολα να αποτελέσουν τους τελικούς κόμβους στο δέντρο. Κάθε τελικός κόμβος είναι ταξινομημένος ως μέλος του υποσυνόλου A_G ή του A_B και ολόκληρη η διαδικασία μπορεί να παρουσιαστεί γραφικά ως δέντρο, όπως φαίνεται στο Σχήμα 3.1.

Για την εφαρμογή της διαδικασίας των δέντρων ταξινόμησης είναι πρώτα απαραίτητο να καθοριστούν κάποιοι κανόνες αποφάσεων, μερικοί από τους οποίους είναι οι εξής:

- Ο κανόνας διάσπασης - ο κανόνας που πρέπει να χρησιμοποιηθεί για να διασπαστεί το σύνολο δεδομένων σε δύο υποσύνολα.
- Ο κανόνας τερματισμού – ο κανόνας με τον οποίο αποφασίζεται ότι ένα υποσύνολο είναι τελικός κόμβος.
- Ο κανόνας που καθορίζει αν κάθε τελικός κόμβος ανήκει στην κατηγορία των «καλών» ή των «κακών» πελατών.

Σχήμα 3.1 *Δέντρο ταξινόμησης*



Η απόφαση καθορισμού ποιοι από τους τελικούς κόμβους ανήκουν στην κατηγορία των «καλών» και ποιοι στους «κακούς» είναι πολύ εύκολο να ληφθεί. Μια απλή σκέψη είναι να χαρακτηρίζεται ο τελικός κόμβος ως «καλός» εάν η πλειοψηφία των περιπτώσεων του δείγματος σε αυτόν τον κόμβο είναι «καλοί» πελάτες. Μια εναλλακτική πρόταση είναι να ελαχιστοποιηθεί το κόστος λανθασμένης ταξινόμησης. Έστω D το αναμενόμενο κόστος για τη λάθος ταξινόμηση ενός «κακού» τελικού κόμβου ως «καλό» και L το αναμενόμενο κόστος που προκύπτει με λάθος ταξινόμηση ενός «καλού» κόμβου ως «κακό». Τότε το κόστος μπορεί να ελαχιστοποιηθεί εάν ο κόμβος ταξινομηθεί ως «καλός» όταν η αναλογία των καλών προς τους κακούς στο δείγμα υπερβαίνει την ποσότητα $\frac{D}{L}$ σε αυτόν τον κόμβο.

Στη διαδικασία σχηματισμού ενός δέντρου ταξινόμησης είναι απαραίτητο να χρησιμοποιηθεί κάποιος κανόνας τερματισμού έτσι ώστε να καθοριστεί ο τελικός κόμβος. Χρησιμοποιώντας κάποιον κανόνα τερματισμού, αναμένεται να δημιουργηθεί κάποιο πιο ανθεκτικό δέντρο όσον αφορά το διαχωρισμό των πελατών σε σχέση με το αρχικό που θα είναι πολύ μεγαλύτερο. Εάν είχαμε ένα δέντρο όπου κάθε τελικός κόμβος είχε μόνο μια περίπτωση από το δείγμα ανάπτυξης, αυτό το δέντρο θα περιέγραφε έναν τέλειο διαχωρισμό σε αυτό το δείγμα αλλά θα ήταν ένας πολύ «φτωχός» διαχωριστικός κανόνας για οποιοδήποτε άλλο σύνολο πελατών. Κατά συνέπεια, ένας κόμβος ορίζεται ως τελικός κόμβος σε δύο περιπτώσεις. Η πρώτη περίπτωση είναι όταν ο αριθμός περιπτώσεων στον κόμβο είναι

τόσο μικρός που δεν έχει κανένα νόημα να διαιρεθεί περαιτέρω. Πιο συγκεκριμένα, αυτό γίνεται συνήθως όταν υπάρχουν λιγότερες από 10 περιπτώσεις στον κόμβο. Η δεύτερη περίπτωση χαρακτηρισμού ενός κόμβου ως τελικού είναι όταν δεν υπάρχει κάποια αξιόλογη διαφορά αν διασπαστεί ο κόμβος (σε σχέση με το να παραμείνει ο κόμβος ως έχει). Επομένως, θα μπορούσαμε να πούμε ότι ο υποκείμενος κόμβος αξίζει να διασπαστεί όταν η διαφορά στην αξία μέτρησης είναι μεγαλύτερη από κάποιο ορισμένο επίπεδο β.

Εάν έχουμε δημιουργήσει ένα πολύ μεγάλο δέντρο μπορεί να ελαττωθεί το μέγεθός του αφαιρώντας κάποιες διασπάσεις. Ο καλύτερος τρόπος για να γίνει αυτό είναι να δημιουργηθεί ένα δείγμα πελατών το οποίο να μην έχει χρησιμοποιηθεί στην ανοικοδόμηση του δέντρου (*holdout sample*). Αυτό το δείγμα χρησιμοποιείται για να υπολογίσει εμπειρικά το αναμενόμενο κόστος για τις διάφορες πιθανές περικοπές του δέντρου. Χρησιμοποιώντας αυτό το δείγμα και ένα δέντρο ταξινόμησης T, ορίζουμε ως T_G το σύνολο των κόμβων που είναι ταξινομημένοι ως «καλοί» και ως T_B το σύνολο των κόμβων που είναι ταξινομημένοι ως «κακοί». Επιπλέον, συμβολίζουμε με $r(t,B)$ το ποσοστό του δείγματος (*holdout sample*) που ανήκει στον κόμβο t και είναι ταξινομημένο ως «κακό» και με $r(t,G)$ το ποσοστό του δείγματος που ανήκει στον κόμβο t και είναι ταξινομημένο ως «καλό». Η εκτίμηση του αναμενόμενου κόστους δίνεται από τον τύπο

$$r(T) = \sum_{i \in T_G} Dr(t,B) + \sum_{i \in T_B} Lr(t,G).$$

Εάν $n(T)$ είναι ο αριθμός των κόμβων στο δέντρο ταξινόμησης T, ορίζουμε την ποσότητα $c(T) = r(T) + dn(T)$ και δημιουργούμε ένα δέντρο T^* κοιτάζοντας από όλα τα δυνατά υποδέντρα του T ποιο ελαχιστοποιεί την ποσότητα $c(T)$. Εάν $d = 0$ τότε καταλήγουμε στο αρχικό μη «κομμένο» δέντρο, ενώ όσο το d μεγαλώνει το δέντρο θα περιλαμβάνει λιγότερους κόμβους. Κατά συνέπεια η επιλογή του d δίνει μια άποψη για το πόσο μεγάλο επιθυμούμε να είναι το δέντρο.

Οι πιο απλοί κανόνες διάσπασης είναι αυτοί που δίνουν έμφαση στο αποτέλεσμα που θα έχει η προτεινόμενη διάσπαση. Αυτό μπορεί να γίνει με την εύρεση της καλύτερης επικείμενης διάσπασης για κάθε χαρακτηριστικό, χρησιμοποιώντας κάποιο μέτρο που υπολογίζει πόσο καλή είναι η κάθε δυνατή διάσπαση. Κατόπιν, αποφασίζεται ποιου χαρακτηριστικού η διάσπαση είναι καλύτερη κάτω από αυτό το μέτρο. Εάν το X_i είναι μια κατηγορική μεταβλητή τότε εξετάζονται όλες οι πιθανές διασπάσεις των κατηγοριών σε δύο

και ελέγχεται το μέτρο για κάθε διάσπαση. Μπορούμε να πούμε ότι η διάσπαση μιας μεταβλητής είναι μια αρχική ταξινόμηση σε δύο κατηγορίες. Για οποιοδήποτε συνεχές χαρακτηριστικό X_i κοιτάμε τις διασπάσεις $\{x_i < s\},\{x_i \geq s\}$ για όλες τις τιμές του s και βρίσκουμε αυτήν την τιμή του s όπου το μέτρο που χρησιμοποιούμε είναι καλύτερο. Συνήθως μπορούμε να διατάξουμε σε αύξουσα σειρά τις αναλογίες του δείγματος «καλοί»:«κακοί» για κάθε κατηγορία γνωρίζοντας ότι η καλύτερη διάσπαση θα διαχωρίσει αυτή τη διάταξη σε δύο ομάδες. Το πιο κοινό μέτρο που χρησιμοποιείται είναι το κριτήριο **Kolmogorov-Smirnov**, αλλά υπάρχουν και άλλα μέτρα όπως είναι ο **βασικός δείκτης μη αγνότητας (*basic impurity index*)**, ο δείκτης **Gini**, ο **δείκτης εντροπίας (*entropy index*)** και το **ημιάθροισμα τετραγώνων (*half-sum of squares*)**. Δίνουμε στη συνέχεια σε συντομία την περιγραφή των δεικτών αυτών.

α. Κριτήριο Kolmogorov – Smirnov

Έστω ότι έχουμε ένα συνεχές χαρακτηριστικό X_i. Η $F(s|G)$ είναι η αθροιστική συνάρτηση κατανομής του X_i για τους «καλούς» πελάτες και η $F(s|B)$ είναι η αθροιστική συνάρτηση κατανομής για τους «κακούς» πελάτες. Υποθέτοντας ότι το χαρακτηριστικό X_i τείνει να παίρνει μικρότερες τιμές για τους «κακούς» πελάτες ο πιο λογικός κανόνας είναι να γίνει η διάσπαση στην τιμή s για την οποία ελαχιστοποιείται η ποσότητα

$$L \cdot F(s|G)p_G + D(1-F(s|B))p_B.$$

Εάν $Lp_G = Dp_B$ τότε ζητάμε να ελαχιστοποιηθεί η ποσότητα $F(s|G)-F(s|B)$ ή αντί αυτού να μεγιστοποιηθεί η ποσότητα $F(s|B)-F(s|G)$ και έτσι οδηγούμαστε στην κλασσική απόσταση Kolmogorov–Smirnov ανάμεσα στις δύο κατανομές όπως φαίνεται στο Σχήμα 3.2.

Εάν διαχωρίσουμε τα δύο υποσύνολα σε αριστερό (*left – l*) και δεξί (*right – r*) σύνολο τότε είναι σαν να μεγιστοποιούμε τη διαφορά μεταξύ των $P(l|B)$ και $P(l|G)$, όπου $P(l|B)$ είναι η πιθανότητα ένας «κακός» πελάτης να εμφανιστεί στην αριστερή ομάδα και $P(l|G)$ είναι η πιθανότητα ένας «καλός» πελάτης να εμφανιστεί στη δεξιά ομάδα (στην περίπτωση που έχουμε συνεχείς μεταβλητές θα έχουμε $F(s|B)$ και $F(s|G)$ αντίστοιχα).

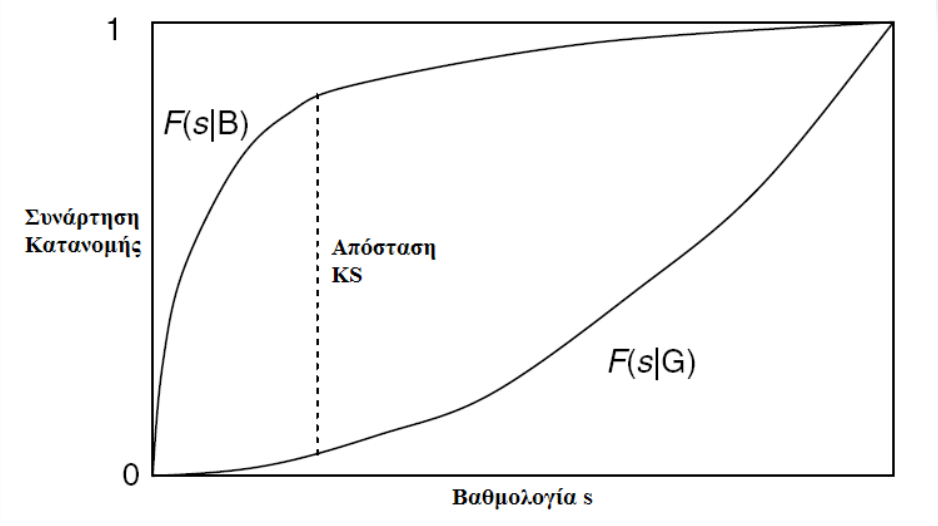
Χρησιμοποιώντας το θεώρημα του Bayes μπορούμε να πούμε ότι $P(l|B) = \dfrac{P(B|l)P(l)}{P(B)}$

και $P(l|G) = \dfrac{P(G|l)P(l)}{P(G)}$. Κατά συνέπεια, για κατηγορικές και συνεχείς μεταβλητές το

κριτήριο Kolmogorov-Smirnov (KS) είναι να βρεθεί ο κατάλληλος διαχωρισμός σε αριστερή και δεξιά ομάδα τέτοιος ώστε να μεγιστοποιείται η ποσότητα:

$$KS = \left|P(l \mid B) - P(l \mid G)\right| = \left|\frac{P(B \mid l)}{P(B)} - \frac{P(G \mid l)}{P(G)}\right| \cdot P(l). \qquad (3.28)$$

Σχήμα 3.2 *Στατιστικό Kolmogorov – Smirnov (πηγή: Thomas (2009))*



Για παράδειγμα, έστω ότι έχουμε το χαρακτηριστικό κατάσταση κατοικίας που έχει τρεις ιδιότητες (ιδιοκτήτης, ενοικιαστής και διαμονή με γονείς)[9]. Στον Πίνακα 3.1 δίνεται ο αριθμός των «καλών» και των «κακών» πελατών που είναι γνωστός για κάθε ιδιότητα και προέκυψε από ένα δείγμα προηγούμενων πελατών.

Πίνακας 3.1 *Πίνακας χαρακτηριστικού «κατάσταση κατοικίας»*

Κατάσταση Κατοικίας / Αποτέλεσμα	Ιδιοκτήτης	Ενοικιαστής	Με γονείς
Καλοί	1000	400	80
Κακοί	200	200	120
Αναλογία καλοί : κακοί	5:1	2:1	0,67:1

[9] Το παράδειγμα αυτό έχει ληφθεί από το βιβλίο των Crook et al. (2002)

Θα χρησιμοποιήσουμε το κριτήριο KS για να βρούμε ποιος είναι ο καλύτερος τρόπος για να διασπάσουμε το χαρακτηριστικό αυτό. Από την αναλογία «καλοί»:«κακοί» είναι φανερό ότι η καλύτερη διάσπαση είναι σε μία ομάδα να ανήκουν αυτοί που κατοικούν με τους γονείς τους και στην άλλη ομάδα να ανήκουν οι ιδιοκτήτες και οι ενοικιαστές μαζί. Μια άλλη πρόταση είναι σε μια ομάδα να ανήκουν αυτοί που μένουν με τους γονείς μαζί με τους ενοικιαστές και στην άλλη να ανήκουν οι ιδιοκτήτες. Χρησιμοποιώντας το κριτήριο KS και τον τύπο (3.28) θα βρούμε ποια διάσπαση είναι καλύτερη.

- Αν $l = \text{με γονείς}$, $r = \text{ιδιοκτήτης και ενοικιαστής}$, τότε

  $P(l \mid B) = \dfrac{120}{520} = 0{,}231$, $P(l \mid G) = \dfrac{80}{1480} = 0{,}054$.

  Άρα, $KS = \left|P(l \mid B) - P(l \mid G)\right| = \left|0{,}231 - 0{,}054\right| = 0{,}177.$

- Αν $l = \text{με γονείς και ενοικιαστής}$, $r = \text{ιδιοκτήτης}$, τότε

  $P(l \mid B) = \dfrac{320}{520} = 0{,}615$, $P(l \mid G) = \dfrac{480}{1480} = 0{,}324$.

  Άρα, $KS = \left|P(l \mid B) - P(l \mid G)\right| = \left|0{,}615 - 0{,}324\right| = 0{,}291.$

Παρατηρούμε ότι η ποσότητα KS είναι μεγαλύτερη στη δεύτερη περίπτωση οπότε εκεί θα έχουμε την καλύτερη δυνατή διάσπαση για το συγκεκριμένο χαρακτηριστικό.

β. Βασικός δείκτης μη αγνότητας *i*(*ν*)

Υπάρχει μια ολόκληρη κατηγορία δεικτών μη αγνότητας (*impurity index*) που έχουν ως σκοπό να αξιολογήσουν πόσο «καθαρός» είναι κάθε κόμβος v του δέντρου, όπου η «πλήρης αγνότητα» αντιστοιχεί σε εκείνο τον κόμβο που είναι ολόκληρος μια κατηγορία. Εάν ο κόμβος διασπαστεί σε έναν αριστερό κόμβο l με πιθανότητα $P(l)$ και σε ένα δεξί κόμβο r με πιθανότητα $P(r)$ τότε η αλλαγή στη μη αγνότητα που έχει προκύψει από αυτή τη διάσπαση μπορεί να μετρηθεί από την ποσότητα

$$DI = i(v) - P(l)i(l) - P(r)i(r). \tag{3.29}$$

Όσο μεγαλύτερη είναι αυτή η διαφορά, τόσο μεγαλύτερη είναι η αλλαγή στη μη αγνότητα, που σημαίνει ότι οι νέοι κόμβοι είναι καθαρότεροι το οποίο είναι αυτό που επιδιώκουμε. Επομένως, επιλέγουμε τη διάσπαση που μεγιστοποιεί την παραπάνω έκφραση. Αυτό είναι ισοδύναμο με την ελαχιστοποίηση της ποσότητας $P(l)i(l) - P(r)i(r)$. Είναι προφανές ότι εάν

δεν υπάρχει καμία διάσπαση με θετική διαφορά τότε δεν θα πρέπει να διασπαστεί καθόλου ο κόμβος.

Ένας βασικός δείκτης μη αγνότητας προκύπτει θεωρώντας ως $i(v)$ το ποσοστό της μικρότερης ομάδας σε αυτόν τον κόμβο έτσι ώστε

$$i(v)=\begin{cases} P(G\,|\,v), \text{ εάν } P(G\,|\,v)\le 0{,}5, \\ P(B\,|\,v), \text{ εάν } P(B\,|\,v)<0{,}5. \end{cases} \tag{3.30}$$

Χρησιμοποιώντας τον Πίνακα 3.1 από το προηγούμενο παράδειγμα θα βρούμε ποια είναι η καλύτερη δυνατή διάσπαση, όπου v είναι ολόκληρο το σύνολο πριν από κάθε διάσπαση. Σύμφωνα με το δείκτη μη αγνότητας και χρησιμοποιώντας τους τύπους (3.29) και (3.30) έχουμε:

- Αν $l=$ με γονείς, $r=$ ιδιοκτήτης και ενοικιαστής, τότε

  $i(v)=\frac{520}{2000}=0{,}26,\ P(l)=\frac{200}{2000}=0{,}1,\ i(l)=\frac{80}{200}=0{,}4,$

  $P(r)=\frac{1800}{2000}=0{,}9,\ i(r)=\frac{400}{1800}=0{,}22.$

  Άρα, $DI=0{,}26-0{,}1\cdot 0{,}4-0{,}9\cdot 0{,}22=0{,}02.$

- Αν $l=$ με γονείς και ενοικιαστής, $r=$ ιδιοκτήτης, τότε

  $i(v)=\frac{520}{2000}=0{,}26,\ P(l)=\frac{800}{2000}=0{,}4,\ i(l)=\frac{320}{800}=0{,}4,$

  $P(r)=\frac{1200}{2000}=0{,}6,\ i(r)=\frac{200}{1200}=0{,}167.$

  Άρα, $DI=0{,}26-0{,}4\cdot 0{,}4-0{,}6\cdot 0{,}167=0.$

Παρατηρούμε ότι η ποσότητα DI είναι μεγαλύτερη στην πρώτη περίπτωση όπου έχουμε και την καλύτερη δυνατή διάσπαση για το συγκεκριμένο χαρακτηριστικό.

Αν και ο δείκτης μη αγνότητας φαίνεται χρήσιμος, κάποιες φορές τα συμπεράσματα που προκύπτουν από τη χρήση του μπορεί να είναι παραπλανητικά. Παρατηρήσαμε ότι στη δεύτερη διάσπαση η ποσότητα DI είναι 0 επειδή οι «κακοί» πελάτες είναι μειοψηφία και στους τρεις κόμβους v, l και r. Αυτό θα συμβαίνει πάντα όταν η ίδια ομάδα είναι σε μειοψηφία και στους τρεις κόμβους και αυτό το φαινόμενο παρατηρείται σε πολλές περιπτώσεις ανάπτυξης μοντέλων πιστοληπτικής ικανότητας. Έτσι, αν όλες οι δυνατές διασπάσεις σε έναν κόμβο δίνουν στο DI την τιμή 0, το κριτήριο αυτό για την ανάδειξη της καλύτερης δυνατής διάσπασης δεν θα είναι ιδιαίτερα χρήσιμο.

γ. Δείκτης Gini

Ο δείκτης Gini (*Gini index*) δε σχετίζεται γραμμικά με τη πιθανότητα της αναλογίας της μη αγνότητας αλλά τετραγωνικά και έτσι βάζει σχετικά περισσότερο βάρος στους «καθαρότερους» κόμβους. Πιο συγκεκριμένα, ο δείκτης Gini ορίζεται μέσω του τύπου (3.29) θεωρώντας $i(v) = P(G|v)P(B|v)$, όποτε θα έχουμε

$$G = P(G|v)P(B|v) - P(l)i(l) - P(r)i(r) =$$
$$= P(G|v)P(B|v) - P(l)P(G|l)P(B|l) - P(r)P(G|r)P(B|r).$$

Όπως και στην περίπτωση του βασικού δείκτη μη αγνότητας επιλέγεται η διάσπαση που μεγιστοποιεί την παραπάνω έκφραση.

δ. Δείκτης εντροπίας

Ένας άλλος μη γραμμικός δείκτης είναι ο δείκτης εντροπίας (*entropy index*) ο οποίος ορίζεται ως εξής:

$$E = i(v) - P(l)i(l) - P(r)i(r)$$

όπου $i(v) = -P(G|v)\ln(P(G|v)) - P(B|v)\ln(P(B|v))$.

Η εντροπία σχετίζεται με την ποσότητα της πληροφορίας που υπάρχει για τη διάσπαση μεταξύ των καλών και των κακών πελατών στον κόμβο. Είναι ένα μέτρο που εκτιμά με πόσους διαφορετικούς τρόπους μπορούμε να καταλήξουμε στην πραγματική διάσπαση των καλών και των κακών.

ε. Ημιάθροισμα τετραγώνων

Το ημιάθροισμα τετραγώνων (*half – sum of square*) προέρχεται από τον έλεγχο X^2, ο οποίος εξετάζει τη μηδενική υπόθεση αν το ποσοστό των καλών πελατών είναι το ίδιο στους δύο νέους κόμβους που δημιουργούνται μετά τη διάσπαση. Εάν το στατιστικό X^2 είναι μεγάλο τότε απορρίπτουμε τη μηδενική μας υπόθεση και συμπεραίνουμε ότι τα δύο ποσοστά δεν είναι ίσα. Όσο πιο υψηλή τιμή έχει το στατιστικό τόσο πιο απίθανο είναι να προκύψουν ίδια ποσοστά ή τόσο μεγαλύτερη είναι η διαφορά μεταξύ των ποσοστών των «καλών» και «κακών» πελατών στους δύο νέους κόμβους. Επομένως, οδηγούμαστε στο ακόλουθο κριτήριο:

Εάν $n(l)$ και $n(r)$ είναι ο ολικός αριθμός του αριστερού και του δεξιού κόμβου αντίστοιχα τότε θα πρέπει να μεγιστοποιηθεί η ποσότητα

$$Chi = n(l)n(r)\frac{(P(G|l) - P(G|r))^2}{n(l) - n(r)}.$$

Στην βιβλιογραφία έχουν προταθεί και διάφορες άλλες μέθοδοι διάσπασης. Για παράδειγμα ο Breiman et al. (1984) πρότεινε ότι ένα καλύτερο κριτήριο από το να εξετάζεται ακριβώς η επόμενη διάσπαση θα ήταν να εξεταστεί ποια θα είναι η κατάσταση μετά από μερικές διασπάσεις. Με αυτόν τον τρόπο λαμβάνεται υπόψη όχι μόνο η άμεση βελτίωση που προκαλείται από την επόμενη διάσπαση αλλά και η μακροπρόθεσμη στρατηγική σημασία αυτής της διάσπασης. Μια τέτοια θεώρηση παρουσιάζει ενδιαφέρον όταν η διάσπαση γίνεται χρησιμοποιώντας τα διαφορετικά χαρακτηριστικά σε διαφορετικά επίπεδα του δέντρου.

3.10 Μέθοδος του κοντινότερου γείτονα

Η μέθοδος του κοντινότερου γείτονα (*nearest-neighbor method*) ανήκει στις μεθόδους ομαδοποίησης κατά συστάδες και είναι μια από τις πιο συνηθισμένες μη παραμετρικές προσεγγίσεις στο πρόβλημα ταξινόμησης πληθυσμών. Αυτή η μέθοδος αρχικά προτάθηκε από τους Fix και Hodges (1952), αλλά εφαρμόστηκε για πρώτη φορά στη βαθμολόγηση πιστοληπτικής ικανότητας πελατών από τους Chatterjee και Barcun (1970) και αργότερα από τους Henley και Hand (1996). Η βασική ιδέα της μεθόδου αυτής είναι να επιλεγεί ένα μέτρο απόστασης για τα χαρακτηριστικά, το οποίο θα προσδιορίζει πόσο «κοντά» βρίσκονται δύο οποιοιδήποτε υποψήφιοι πελάτες. Η λογική της μεθόδου είναι πολύ απλή: κάθε νέα παρατήρηση (υποψήφιος πελάτης) συγκρίνεται με τις υπόλοιπες παρατηρήσεις του δείγματος (που αποτελείται από παλαιότερους πελάτες), βρίσκονται οι *k* κοντινότεροι γείτονες σύμφωνα με το μέτρο απόστασης που έχει επιλεγεί και ο νέος υποψήφιος πελάτης τοποθετείται στην ομάδα με τη μεγαλύτερη συχνότητα εμφάνισης για τις *k* περιπτώσεις του δείγματος. Το πλεονέκτημα αυτής της τεχνικής είναι ότι δεν απαιτεί καμία υπόθεση για την κατανομή του πληθυσμού ή των χαρακτηριστικών του.

Οι παράμετροι που απαιτούνται για να εφαρμοστεί αυτή η μέθοδος είναι, ο αριθμός *k* των πελατών που αποτελούν το σύνολο των κοντινότερων γειτόνων, το μέτρο απόστασης και το ποσοστό αυτών που πρέπει να είναι «καλοί» έτσι ώστε ο νέος υποψήφιος να ταξινομηθεί ως «καλός». Η απάντηση στο τελευταίο θα μπορούσε να είναι ότι, εάν η πλειοψηφία των

γειτόνων είναι «καλοί» τότε ο νέος υποψήφιος μπορεί να ταξινομηθεί ως «καλός», διαφορετικά, ο υποψήφιος ταξινομείται ως «κακός». Παρ' όλα αυτά αν η αναμενόμενη απώλεια είναι D και το αναμενόμενο κέρδος είναι L ένας νέος υποψήφιος θα μπορούσε να ταξινομηθεί ως «καλός» αν τουλάχιστον $\frac{D}{D+L}$ από τους κοντινότερους γείτονες είναι «καλοί». Αυτό το κριτήριο ελαχιστοποιεί το αναμενόμενο κόστος εάν η πιθανότητα ενός νέου υποψηφίου να είναι «καλός» είναι ίση με το ποσοστό των κοντινότερων γειτόνων που είναι «καλοί».

Ένα από τα πιο κρίσιμα σημεία στην εφαρμογή της μεθόδου είναι η επιλογή της κατάλληλης μετρικής. Η απόσταση δύο παρατηρήσεων $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$ συμβολίζεται με $d(\mathbf{x}_i, \mathbf{x}_j)$. Στην περίπτωση που τα χαρακτηριστικά εκφράζουν συνεχείς ιδιότητες η απόσταση μεταξύ των παρατηρήσεων $\mathbf{x}_i$ και $\mathbf{x}_j$ μπορεί να είναι η ευκλείδεια απόσταση που δίνεται από τον τύπο

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}.$$

Στην περίπτωση που το σύνολο δεδομένων περιλαμβάνει και διακριτές μεταβλητές χρησιμοποιείται ο τύπος

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} f(x_i, x_j)}.$$

όπου $f(x_i, x_j) = (x_i - x_j)^2$ για τις συνεχείς μεταβλητές και

$$f(x_i, x_j) = \begin{cases} 0, & x_i = x_j \\ 1, & x_i \neq x_j \end{cases}$$

για τις διακριτές μεταβλητές.

Η ευκλείδεια απόσταση έχει το μειονέκτημα ότι εξαρτάται από την κλίμακα μέτρησης επομένως, αν χρησιμοποιηθεί σε μεταβλητές με διαφορετική κλίμακα μέτρησης, υπάρχει περίπτωση να πάρουμε τελείως διαφορετικές αποστάσεις. Ένας τρόπος για να φέρουμε όλες τις μεταβλητές σε συγκρίσιμη κλίμακα είναι να διαιρεθεί καθεμιά μεταβλητή με την τυπική της απόκλιση. Έτσι, αν συμβολίσουμε με s_r τη διακύμανση της r μεταβλητής μπορούμε να ορίσουμε την απόσταση με τον τύπο

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} \frac{(x_{ir} - x_{jr})^2}{s_r^2}}$$

όπου

$$s_r = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \bar{x}_r) \right) \quad \text{και} \quad \bar{x}_r = \frac{1}{n} \sum_{i=1}^{n} x_{ir} \, .$$

Μια τέτοια απόσταση λέγεται απόσταση του Pearson (βλέπε Κούτρας (2007)).

Υπάρχουν πάρα πολλά ακόμα μέτρα αποστάσεων που είναι κατάλληλα για να χρησιμοποιηθούν σε αυτή τη μέθοδο. Αν **b** είναι το p - διάστατο διάνυσμα που δίνεται από τον τύπο (3.12), οι Henley και Hand (1996) πρότειναν για την ταξινόμηση υποψήφιων πελατών ως προς την πιστοληπτική τους ικανότητα μια μετρική της μορφής

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left\{ (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} (\mathbf{I}_p + D\mathbf{w} \cdot \mathbf{w}^{\mathrm{T}}) (\mathbf{x}_i - \mathbf{x}_j) \right\}^{\frac{1}{2}}$$

όπου $\mathbf{I}_p$ είναι ο μοναδιαίος πίνακας. Για να προσδιοριστεί ποια είναι η πιο κατάλληλη τιμή για το D συνήθως εκτελείται ένας μεγάλος αριθμός πειραμάτων. Δοκιμάζοντας διάφορες τιμές του *k*, επιλέγεται και ο ιδανικότερος αριθμός *k* των κοντινότερων γειτόνων. Η επιλογή του *k* εξαρτάται από το μέγεθος του δείγματος ανάπτυξης. Αν και δεν υπάρχουν μεγάλες διαφορές στα αποτελέσματα, συνήθως η καλύτερη επιλογή του D βρίσκεται μεταξύ των τιμών 1,4 έως 1,8.

3.11 Μέθοδοι δημιουργίας στατιστικών μοντέλων βαθμολόγησης συμπεριφοράς και κέρδους

Στις προηγούμενες ενότητες μελετήθηκαν κάποιες μέθοδοι δημιουργίας μοντέλων βαθμολόγησης αιτήσεων (ASM). Σύμφωνα με αυτά τα μοντέλα ένας νέος υποψήφιος πελάτης μπορεί να χαρακτηριστεί ως «καλός» ή «κακός» με βάση τα στοιχεία που είχε συμπληρώσει στην αίτηση για πίστωση, χωρίς να είναι γνωστά κάποια στοιχεία που αφορούν τη πιστοληπτική συμπεριφορά του υποψηφίου ως προς τον χρηματοπιστωτικό οργανισμό. Εάν τελικά κάποιος υποψήφιος χαρακτηριστεί ως «καλός» και γίνει δεκτή η χορήγηση πίστωσης τότε ξεκινάει η συνεργασία μεταξύ του δανειολήπτη και της τράπεζας και πλέον είναι δυνατό

να υπάρξουν πληροφορίες για την πιστοληπτική του συμπεριφορά. Έπειτα από κάποιο χρονικό διάστημα συνεργασίας λοιπόν είναι πλέον εφικτό να κατασκευαστούν στατιστικά μοντέλα βαθμολόγησης συμπεριφοράς (*behavioral scoring*) και έτσι αποδίδεται σε κάθε πελάτη μια βαθμολογία συμπεριφοράς για κάποια συγκεκριμένη χρονική στιγμή. Τα μοντέλα βαθμολόγησης συμπεριφοράς χρησιμοποιούν τα χαρακτηριστικά της πρόσφατης συμπεριφοράς των πελατών για να προβλέψουν αν αυτοί πρόκειται να αθετήσουν τις υποχρεώσεις τους ή όχι.

Η μεθοδολογία ανάπτυξης των μοντέλων βαθμολόγησης συμπεριφοράς (BSM) είναι ίδια με αυτή των ASM Για την ανάπτυξη μοντέλων βαθμολόγησης συμπεριφοράς επιλέγονται οι λογαριασμοί που βρίσκονται σε εξέλιξη μέχρι το χρόνο παρατήρησης. Η συμπεριφορά κάθε πελάτη αναλύεται κατά την περίοδο απόδοσης (περίπου 12 μήνες) πριν το χρόνο παρατήρησης. Έπειτα από το χρόνο παρατήρησης υπάρχει η περίοδος έκβασης (12 μήνες περίπου), μετά το πέρας της οποίας οι πελάτης ταξινομείται ως «καλός» ή «κακός» ανάλογα με το αποτέλεσμά του στο χρόνο παρατήρησης.

Υπάρχει όμως μια χρονική καθυστέρηση περίπου 2 ετών μεταξύ της χρονικής περιόδου που συλλέχθηκαν οι πληροφορίες συναλλαγής που χρησιμοποιήθηκαν για να δομηθεί το σκορόχαρτο και της χρονικής στιγμής που αυτό χρησιμοποιείται. Σύμφωνα με τον Thomas et al (2001), αυτό σημαίνει ότι τα χαρακτηριστικά του πληθυσμού και το οικονομικό περιβάλλον μπορεί να έχει αλλάξει. Το πρόβλημα αυτό μεγαλώνει επειδή τα BSM τείνουν να μην έχουν κανένα εξωτερικό οικονομικό χαρακτηριστικό. Επομένως χρησιμοποιείται η υπόθεση ότι η σχέση μεταξύ των χαρακτηριστικών απόδοσης και της πιστοληπτικής ικανότητας του πελάτη παραμένει ίδια για τα επόμενα 2 χρόνια, εφ' όσον χρησιμοποιούνται τα ίδια χαρακτηριστικά και δεν επηρεάζεται από διάφορες αλλαγές στο οικονομικό περιβάλλον που έχουν συμβεί αυτή τη χρονική περίοδο.

Οι βαθμολογίες συμπεριφοράς περιγράφουν την πιθανότητα ένας ήδη υπάρχον πελάτης να αθετήσει τις υποχρεώσεις του μέσα σε ένα δεδομένο χρονικό διάστημα που είναι συνήθως 12 μήνες. Οπότε, όπως και στις βαθμολογίες αιτήσεων, οι βαθμολογίες συμπεριφοράς είναι ένα μέτρο κινδύνου αθέτησης υποχρεώσεων αλλά αυτές δεν χρησιμοποιούνται άμεσα για τη λήψη αποφάσεων που αφορούν τους οφειλέτες. Για παράδειγμα, μια τράπεζα δεν μπορεί να ανακαλέσει την απόφαση χορήγησης ενός καταναλωτικού δανείου σε έναν οφειλέτη αν η βαθμολογία συμπεριφοράς του έχει μειωθεί και εάν αυτός συνεχίζει να αποπληρώνει τις δόσεις σύμφωνα με τους όρους του δανείου. Εάν εντούτοις, ο οφειλέτης υπέβαλε αίτηση για

κάποιο άλλο δάνειο, τότε η τράπεζα μπορεί να χρησιμοποιήσει τη βαθμολογία συμπεριφοράς για να αποφασίσει αν θα εγκρίνει την αίτησή του ή όχι. Από την άλλη μεριά, μια υψηλή βαθμολογία συμπεριφοράς σημαίνει ότι η πιθανότητα ο οφειλέτης να αθετήσει τις υποχρεώσεις του είναι μικρή με βάση το τρέχον πιστωτικό όριο, αλλά αυτό δε σημαίνει ότι θα παραμείνει χαμηλή αν αυξηθεί αρκετά το πιστωτικό όριο.

Τα BSM εφαρμόζονται σε πελάτες για ένα προϊόν δανείου χρησιμοποιώντας τα χαρακτηριστικά συμπεριφοράς τους σε εκείνο το προϊόν. Όμως, η απόδοση πελατών σε ένα προϊόν μπορεί να δώσει καλές ενδείξεις για την πιθανότητα αθέτησης υποχρεώσεων σε άλλα προϊόντα. Για παράδειγμα, αν ένας πελάτης μια τράπεζας έχει έναν τρεχούμενο λογαριασμό αυτός θα μπορούσε να αποτελέσει μια πολύ καλή ένδειξη για τη γενική οικονομική κατάσταση του πελάτη. Επομένως, αλλαγές στη συμπεριφορά του πελάτη όσον αφορά το λογαριασμό του μπορεί να είναι μια ένδειξη κακής συμπεριφοράς του πελάτη και ως προς κάποιο λογαριασμό δανείου. Κατά συνέπεια τα σκορόχαρτα που έχουν κατασκευαστεί χρησιμοποιώντας τα χαρακτηριστικά από όλα τα προϊόντα που έχει ο πελάτης με την τράπεζα σκοπεύουν να εκτιμήσουν την πιθανότητα αθέτησης υποχρεώσεων σε μερικά ή όλα τα προϊόντα χορήγησης πίστωσης από την τράπεζα. Αυτό το είδος μοντέλου ονομάζεται μοντέλο βαθμολόγησης πελατών (*customer scoring*) και αποτελεί την πιο συνηθισμένη μέθοδο βαθμολόγησης συμπεριφοράς.

Μερικές αποφάσεις που πρέπει να λάβει ένας οργανισμός δεν απαιτούν μόνο τις βαθμολογίες συμπεριφοράς αλλά και άλλες πληροφορίες που αφορούν την πιστοληπτική κατάσταση των υπαρχόντων πελατών. Ο μεγάλος ανταγωνισμός που υπάρχει στη χρηματοπιστωτική αγορά αναγκάζει τους δανειστές να ενδιαφέρονται όχι μόνο για την πιθανότητα αθέτησης των υποχρεώσεών του αλλά και για το κέρδος που θα τους προσφέρει κάθε πελάτης. Το ιδανικότερο για έναν χρηματοπιστωτικό οργανισμό είναι να κατασκευαστεί το καταλληλότερο μοντέλο που να βαθμολογεί την κερδοφορία που προσφέρει κάθε πελάτης για ένα συγκεκριμένο προϊόν (*profit score*). Ακόμα πιο χρήσιμο είναι να αναπτυχθεί ένα σκορόχαρτο που θα βαθμολογεί την κερδοφορία του κάθε πελάτη για όλα τα πιστωτικά προϊόντα που του έχει χορηγήσει ο συγκεκριμένος οργανισμός (*customer profit score*). Ο χρηματοπιστωτικός οργανισμός μπορεί να θελήσει να ρυθμίσει το επιτόκιο ενός δανείου ή να το κάνει ελκυστικότερο και έτσι να αποφευχθεί η περίπτωση ο οφειλέτης να αποπληρώσει πρώρα το δάνειο, μπορεί να θέλει να αποφασίσει πως θα ενεργήσει σε περίπτωση πρόωρης αποπληρωμής του δανείου και ακόμα αν και με ποιο τρόπο μπορεί να προωθήσει πιο

επικερδή προϊόντα (π.χ. αν θα εκδώσει σε κάποιον πελάτη χρυσή πιστωτική κάρτα). Όλες οι προηγούμενες αποφάσεις λαμβάνονται με τη δόμηση μοντέλων βαθμολόγησης κέρδους. Για να δομηθούν τέτοιου είδους μοντέλα χρειάζεται πρώτα να μοντελοποιηθεί η αποδοτικότητα του πελάτη.

Υπάρχουν δύο ευρέως χρησιμοποιημένες προσεγγίσεις που επιτρέπουν τη μοντελοποίηση της δυναμικής φύσης της απόδοσης ενός οφειλέτη. Αυτές οι προσεγγίσεις βασίζονται στις **Μαρκοβιανές Αλυσίδες (*Markov Chains*)** και στην **Ανάλυση Επιβίωσης (*Survival Analysis*).**[10] Η πρώτη μέθοδος απαριθμεί τις διαφορετικές καταστάσεις που μπορεί να βρεθεί ο οφειλέτης και εκτιμάει την πιθανότητα να μετακινηθεί ο πελάτης από τη μία κατάσταση στην άλλη μεταξύ μίας και της αμέσως επόμενης χρονικής περιόδου. Ωστόσο οι καταστάσεις που μπορεί να βρεθεί ένας οφειλέτης ενδέχεται να είναι αρκετά πολύπλοκες αφού αποσκοπούν να περιγράψουν όλες τις πτυχές της θέσης του οφειλέτη, όπως είναι ο τρέχων κίνδυνος αθέτησης των υποχρεώσεων, η τρέχουσα συμπεριφορά αποπληρωμής και η άμεση χρήση του προϊόντος. Υποθέτουμε λοιπόν ότι η τωρινή κατάσταση του οφειλέτη περιέχει όλες τις πληροφορίες που χρειάζονται για να εκτιμηθεί η πιθανότητα σε ποια κατάσταση ο οφειλέτης θα μετακινηθεί αμέσως μετά. Αυτή η υπόθεση εκφράζει τη μαρκοβιανή ιδιότητα σύμφωνα με την οποία χρειάζεται μόνο η τρέχουσα συμπεριφορά του πελάτη για να προβλεφθεί η μελλοντική.

Η Ανάλυση Επιβίωσης συνήθως χρησιμοποιείται όταν εμφανίζονται ορισμένα γεγονότα και δεν ενδιαφέρει η κατάσταση του οφειλέτη ανάμεσα στις χρονικές στιγμές των γεγονότων. Στο πλαίσιο της καταναλωτικής πίστης αυτά τα γεγονότα θα μπορούσαν να ήταν η αθέτηση υποχρεώσεων ή η πρόωρη αποπληρωμή ενός δανείου. Αν και με την Ανάλυση Επιβίωσης υπάρχει η αδυναμία ότι δεν μπορεί να εκτιμηθεί τι συμβαίνει κατά τον ενδιάμεσο χρόνο μεταξύ των γεγονότων, έχει το πλεονέκτημα ότι δε χρειάζεται να ισχύει η μαρκοβιανή ιδιότητα. Επομένως, η Ανάλυση Επιβίωσης είναι ιδιαίτερα χρήσιμη σε περιπτώσεις προϊόντων καθορισμένης διάρκειας, για τα οποία οι δόσεις είναι προκαθορισμένες εκτός αν υπάρξει αθέτηση υποχρεώσεων ή πρόωρη αποπληρωμή.

Οι πιο συνηθισμένες προσεγγίσεις για την ανάπτυξη BSM που εφαρμόζονται στην πράξη μέχρι σήμερα είναι η ταξινόμηση των πελατών σε ομάδες σύμφωνα με ένα μέτρο κινδύνου και ένα μέτρο απόδοσης και η εφαρμογή διαφορετικών πολιτικών πίστωσης για κάθε ομάδα.

[10] Περισσότερες πληροφορίες για τις προσεγγίσεις των Μαρκοβιανών Αλυσίδων και της Αναλυσης Επιβίωσης για την ανάπτυξη μοντέλων βαθμολόγησης κέρδους δίνονται στο βιβλίο του Thomas (2009).

Για παράδειγμα, μια διαδικασία που ακολουθούν μερικοί οργανισμοί είναι να θέτουν τα πιστωτικά όρια με την κατασκευή ενός πίνακα που θα έχει ως γραμμές διάφορες κλάσεις από βαθμολογίες συμπεριφοράς (ως μέτρο κινδύνου) και ως στήλες κλάσεις του ύψους του οφειλόμενου υπολοίπου (ως μέτρο απόδοσης) όπως φαίνεται στον Πίνακα 3.2.

Πίνακας 3.2 *Πίνακας κινδύνου – απόδοσης για πιστωτικό όριο (Πηγή: Thomas et al (2001))*

Οφειλόμενο υπόλοιπο / **Βαθμολογία Συμπεριφοράς**	**<500€**	**500€ - 2.500€**	**1.000€**
>500	10.000€	12.500€	15.000€
300-500	2.000€	4.000€	10.000€
<300	Δε δίνεται όριο	500€	1.000€

Στον παραπάνω πίνακα, ο πιστωτικός αναλυτής χρησιμοποιεί την εμπειρία του για να καθορίσει ποιο θα είναι το πιστωτικό όριο που θα γραφεί σε κάθε κελί του πίνακα. Κατά συνέπεια δεν υπάρχει καμιά πραγματική μοντελοποίηση του συνολικού κέρδους ούτε του τρόπου λήψης αποφάσεων που αφορούν το κέρδος. Ο πίνακας κινδύνου - απόδοσης χρησιμοποιείται ως μια πρώτη προσέγγιση στις βέλτιστες πολιτικές μεγιστοποίησης κέρδους. Σε αυτήν την περίπτωση οι αποφάσεις λαμβάνονται υποκειμενικά. Η υπόθεση είναι ότι η κατάσταση του οφειλέτη δεν αλλάζει με την πάροδο του χρόνου και ούτε πρόκειται να επηρεαστεί από οποιαδήποτε μελλοντική απόφαση. Κατά συνέπεια είναι μια στατική προσέγγιση της βαθμολόγησης κέρδους. Η επιλογή των κλάσεων του κινδύνου και της απόδοσης είναι και αυτή υποκειμενική, δηλαδή τα σημεία διάσπασης επιλέγονται αυθαίρετα και συνήθως οι κλάσεις διασπώνται έτσι ώστε σε κάθε μία να προκύπτουν στρογγυλοποιημένοι αριθμοί και κατά συνέπεια συνήθως αγνοείται η ομοιογένεια των κλάσεων.

3.12 Ανακεφαλαίωση

Για την ανάπτυξη μοντέλων βαθμολόγησης αιτήσεων έχουν προταθεί πολλές στατιστικές μέθοδοι στο παρελθον. Οι κυριότερες μέθοδοι που χρησιμοποιούνται για την αξιολόγηση υποψηφίων πελατών ως προς την πιστοληπτική τους ικανότητα, είναι η Διαχωριστική Ανάλυση, η Γραμμική Παλινδρόμηση, η Λογιστική Παλινδρόμηση, τα Δέντρα Ταξινόμησης και η Μέθοδος του Κοντινότερου Γειτονα.

Ο κύριος σκοπός των Δέντρων Ταξινόμησης και της μεθόδου του Κοντινότερου Γείτονα είναι να προβλεφθεί σε ποια από τις δύο κατηγορίες («καλός» ή «κακός») ανήκει κάθε υποψήφιος πελάτης ενός χρηματοπιστωτικού οργανισμού. Από την άλλη, ο κύριος σκοπός της Γραμμικής Διαχωριστικής Ανάλυση, της Γραμμικής και της Λογιστικής Παλινδρόμησης είναι να προβλεφθεί η πιθανότητα αθέτησης υποχρεώσεων ή η πιθανότητα να είναι «καλός» ένας υποψήφιος πελάτης.

Η επιλογή της καταλληλότερης μεθόδου εξαρτάται από πολλές παραμέτρους κάποιες από της οποίες είναι το μέγεθος του δείγματος, το είδος της μεταβλητής απόκρισης, η ερμηνευσιμότητα των αποτελεσμάτων και η ποιότητα των διαθέσιμών δεδομένων.

ΚΕΦΑΛΑΙΟ 4

Αξιολόγηση της απόδοσης και επικύρωση των μοντέλων

4.1 Εισαγωγή

Αμέσως μετά το στάδιο της κατάτμησης, όπου έχει αποφασιστεί πόσα μοντέλα θα κατασκευαστούν και έχει διαμορφωθεί ένα σύνολο δεδομένων για κάθε τμήμα, τα δεδομένα χωρίζονται σε δύο κατηγορίες, το δείγμα ανάπτυξης με βάση το οποίο δομείται το μοντέλο και το **δείγμα ελέγχου (*test sample*)** ή **δείγμα επικύρωσης (*validation sample*)** ή **παρακρατημένο δείγμα (*hold-out sample*)** το οποίο χρησιμοποιείται για να ελεγχθεί πόσο καλή είναι η απόδοση του μοντέλου. Το ποσοστό σύμφωνα με το οποίο χωρίζεται το δείγμα εξαρτάται από μέγεθος του αρχικού δείγματος. Εάν αυτό είναι αρκετά μεγάλο τότε συνήθως το 70% από αυτό αποτελεί το δείγμα ανάπτυξης και το υπόλοιπο 30% το δείγμα ελέγχου.

Αφού κατασκευαστεί ένα στατιστικό μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας με βάση το δείγμα ανάπτυξης και αφού είναι γνωστή πλέον η κατάσταση των υποψήφιων πελατών, αν δηλαδή είναι «καλοί» ή «κακοί», το αμέσως επόμενο βήμα είναι να εξεταστεί πόσο καλό και αποδοτικό είναι αυτό το μοντέλο. Υπάρχουν αρκετοί διαφορετικοί τρόποι σύμφωνα με τους οποίους μπορεί να μετρηθεί η αποτελεσματικότητα ενός CSM και καθένας από αυτούς απεικονίζει τα διαφορετικά χαρακτηριστικά γνωρίσματα του CSM.

Ένας τρόπος για τη μέτρηση της απόδοσης ενός CSM, είναι η μέτρηση της ακρίβειας ταξινόμησης (εκτίμηση της σωστής ταξινόμησης) των οφειλετών σε «καλούς» και «κακούς». Η ακρίβεια ταξινόμησης δεν εξαρτάται μόνο από το μοντέλο αλλά και από τη βαθμολογία αποδοχής-απόρριψης που έχει προσδιοριστεί. Αυτό συμβαίνει γιατί η αποτελεσματικότητα κάθε μοντέλου είναι διαφορετική όταν χρησιμοποιούνται διαφορετικά σημεία αποκοπής.

Ένας από τους πιο σημαντικούς τρόπους μέτρησης της απόδοσης ενός CSM είναι να αξιολογηθεί η διαχωριστική ικανότητα του μοντέλου. Ένα καλό μοντέλο θα πρέπει να δημιουργεί έναν όσο γίνεται πιο σαφή διαχωρισμό μεταξύ των «καλών» και των «κακών» πελατών έτσι ώστε κάθε μία από τις δύο ομάδες να αντιμετωπιστεί με διαφορετικό τρόπο, όπως για παράδειγμα να απορριφθεί η αίτηση για χορήγηση πίστωσης των «κακών» και να και να γίνει δεκτή για τους «καλούς».

Ένας τρίτος τρόπος είναι να μετρηθεί η ακρίβεια προσαρμογής (*calibration*) της εκτίμησης της πιθανότητας ένας πελάτης να αθετήσει τις υποχρεώσεις του ή να είναι «καλός». Αυτή η μέθοδος απαιτεί να είναι γνωστή η συνάρτηση που μετασχηματίζει τη βαθμολογία σε πιθανότητα ενός γεγονότος. Για παράδειγμα, στην περίπτωση της λογιστικής παλινδρόμησης που η βαθμολογία εκφράζεται από το λογάριθμο του λόγου πιθανοτήτων, η πιθανότητα κάποιος να είναι καλός δίνεται από την έκφραση

$$p(s) = \frac{1}{1+e^{-s}}.$$

Εάν όμως έχουμε μια κλιμακούμενη βαθμολογία λογαρίθμου της σχετικής πιθανότητας (*scaled log odds score*)[11], τότε ο μετασχηματισμός είναι

$$p(s) = \frac{1}{1+e^{-(s-a)/b}}$$

αφού το $(s-a)/b$ είναι η βαθμολογία του λογαρίθμου της σχετικής πιθανότητας. Για άλλες βαθμολογίες, ο μετασχηματισμός της βαθμολογίας κάθε πελάτη σε πιθανότητα αυτός να είναι «καλός» γίνεται συγκρίνοντας τη βαθμολογία με τις εμπειρικές σχετικές πιθανότητες ή με το ποσοστό αυτών που αθετούν τις υποχρεώσεις τους στο δείγμα ανάπτυξης. Η μέτρηση της προσαρμογής της εκτίμησης της πιθανότητας αποτελεί ένα είδος σύγκρισης για το πόσο ανθεκτικός είναι ο μετασχηματισμός αυτός. Ο μετασχηματισμός λαμβάνεται εξετάζοντας τις σχετικές πιθανότητες σε διαφορετικές κλάσεις των βαθμολογιών του δείγματος ανάπτυξης και συγκρίνοντας τα αποτελέσματα των σχετικών πιθανοτήτων σε ένα δείγμα ελέγχου για τις ίδιες κλάσεις βαθμολογιών.

[11] Η κλιμακούμενη βαθμολογία χρησιμοποιείται συχνά στην πράξη όταν θέλουμε η βαθμολογία να έχει μια συγκεκριμένη κλίμακα, π.χ. από 1 έως 1000. Η δημιουργία μιας κλιμακούμενης βαθμολογίας επιτυγχάνεται με έναν απλό γραμμικό μετασχηματισμό της αρχικής βαθμολογίας, δηλαδή $s_{κλιμακούμενο} = \alpha + b \cdot s_{αρχικό}$, όπου α και b είναι οι σταθερές που εκφράζουν το σταθερό όρο και την κλίση της ευθείας γραμμής $\alpha + b \cdot s_{αρχικό}$ αντίστοιχα.

4.2 Ποσοστά λανθασμένης ταξινόμησης χρησιμοποιώντας δείγματα ελέγχου

Για να μετρηθεί η αποτελεσματικότητα των μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας αξιολογούνται οι προβλέψεις που χρησιμοποιήθηκαν για να ληφθεί η απόφαση εάν ένας πελάτης με μια συγκεκριμένη βαθμολογία θα απορριφθεί ή όχι. Δεδομένου ότι τα κέρδη κάθε χρηματοπιστωτικού οργανισμού μεγιστοποιούνται όταν όλοι οι πελάτες που γίνονται αποδεκτοί είναι στην πραγματικότητα «καλοί» και όλοι αυτοί που απορρίπτονται είναι «κακοί», η ομάδα αυτών που έγιναν αποδεκτοί είναι ομάδα των αναμενόμενα «καλών», ενώ αυτοί που απορρίφθηκαν είναι η ομάδα των αναμενόμενα «κακών». Ωστόσο, αυτή η διαδικασία της απόφασης δεν απαιτεί μόνο την ύπαρξη μιας βαθμολογίας για κάθε πελάτη αλλά και μια βαθμολογία αποδοχής – απόρριψης s, έτσι ώστε οι υποψήφιοι με βαθμολογία μεγαλύτερη από s να ταξινομούνται στην ομάδα των αναμενόμενα «καλών» και να γίνονται αποδεκτοί, ενώ αυτοί με βαθμολογία μικρότερη από s να ταξινομούνται στην ομάδα των «αναμενόμενα κακών» και να απορρίπτονται.

Με βάση το CSM που έχει αναπτυχθεί, ο υποψήφιος πληθυσμός χωρίζεται σε δύο ομάδες, στους πελάτες για τους οποίους έχει προβλεφτεί ότι θα είναι «καλοί» και γίνονται αποδεκτοί και σε αυτούς για τους οποίους έχει προβλεφτεί ότι θα αθετήσουν τις υποχρεώσεις τους και απορρίπτονται. Οι αποφάσεις που λαμβάνονται με βάση το κάθε CSM μπορούν να θεωρηθούν απλά ως προβλέψεις σε ποια ομάδα είναι πιθανό να ανήκει κάθε υποψήφιος. Επομένως, τα αποτελέσματα των αποφάσεων που προκύπτουν για το κάθε μοντέλο και εφαρμόζονται σε έναν πληθυσμό από n πελάτες μπορούν να απεικονιστούν σε έναν 2×2 πίνακα των προβλεφθέντων ομάδων έναντι των πραγματικών. Στις γραμμές αυτού του πίνακα δίνεται ο αριθμός (ή το ποσοστό) των πελατών που ανήκουν σε κάθε μία από τις δύο κατηγορίες με βάση τις προβλέψεις του CSM, ενώ στις στήλες δίνεται ο πραγματικός αριθμός (ή ποσοστό) των πελατών που ανήκουν σε κάθε κατηγορία. Αυτός ο πίνακας ονομάζεται **πίνακας συγχύσεως (*confusion matrix*)** και τα μη διαγώνια στοιχεία του εκφράζουν ποσοστά λανθασμένης ταξινόμησης.

Για να κατασκευαστεί ο πίνακας συγχύσεως χρησιμοποιείται κυρίως το δείγμα ελέγχου. Για κάθε πελάτη αυτού του δείγματος καθορίζεται μία βαθμολογία (με βάση το CSM που δομήθηκε από το δείγμα ανάπτυξης) σύμφωνα με την οποία αυτός απορρίπτεται ή γίνεται αποδεκτός. Όμως, για τους συγκεκριμένους πελάτες είναι γνωστό αν αυτοί είναι «καλοί» ή

«κακοί». Κατά συνέπεια τα αποτελέσματα που προκύπτουν για αυτό το CSM μπορούν να καταγραφούν σε έναν πίνακα συγχύσεως. Εάν g_G είναι ο αριθμός των πελατών που έγιναν αποδεκτοί ενώ στην πραγματικότητα είναι «καλοί», b_G είναι ο αριθμός των πελατών που απορρίφθηκαν ενώ στην πραγματικότητα είναι «καλοί» τότε ο συνολικός αριθμός που είναι πραγματικά «καλοί» στον πληθυσμό $n = n_G + n_B$ πελατών είναι $n_G = g_G + b_G$. Εάν g_B είναι ο αριθμός των πελατών που έγιναν αποδεκτοί ενώ στην πραγματικότητα είναι «κακοί» , b_B είναι ο αριθμός των πελατών που απορρίφθηκαν ενώ στην πραγματικότητα είναι «κακοί» τότε ο συνολικός αριθμός που είναι πραγματικά «κακοί» στον πληθυσμό $n = n_G + n_B$ πελατών είναι $n_B = g_B + b_B$. Ο Πίνακας 4.1 είναι ο πίνακας σύγχυσης συχνοτήτων και ο Πίνακας 4.2 είναι ο πίνακας σύγχυσης πιθανοτήτων που προκύπτουν για αυτόν τον πληθυσμό.

Πίνακας 4.1 *Πίνακας σύγχυσης συχνοτήτων*

Πραγματικές ομάδες / **Αναμενόμενες ομάδες**	**Καλοί**	**Κακοί**
Αποδεχτοί (αναμενόμενα καλοί)	g_G	g_B
Απορριφθέντες (αναμενόμενα κακοί)	b_G	b_B

Πίνακας 4.2 *Πίνακας σύγχυσης πιθανοτήτων*

Πραγματικές ομάδες / **Αναμενόμενες ομάδες**	**Καλοί**	**Κακοί**
Αποδεχτοί (αναμενόμενα καλοί)	g_G/n_G	g_B/n_B
Απορριφθέντες (αναμενόμενα κακοί)	b_G/n_G	b_B/n_B

Από τον Πίνακα 4.2 δίνεται ότι η πιθανότητα ένας πελάτης να ταξινομηθεί ως «καλός» και να γίνει αποδεκτός ενώ στην πραγματικότητα είναι «κακός» είναι g_B/n_B και η πιθανότητα ένας πελάτης να ταξινομηθεί ως «κακός» και να απορριφθεί ενώ στην πραγματικότητα είναι «καλός» είναι b_G/n_G. Αν δούμε το πρόβλημα αυτό ως ένα κλασσικό πρόβλημα ελέγχου

υποθέσεων με μηδενική υπόθεση ο πελάτης να είναι στην πραγματικότητα «καλός», τότε διαπιστώνουμε ότι η πιθανότητα να απορριφθεί ένας πελάτης ενώ είναι «καλός» (πιθανότητα λανθασμένης απόρριψης) είναι η πιθανότητα σφάλματος τύπου Ι που συμβολίζεται με α, δηλαδή $\alpha = b_G / n_G$. Η πιθανότητα να γίνει αποδεκτός ένας πελάτης ενώ αυτός είναι «κακός» (πιθανότητα λανθασμένης αποδοχής) είναι η πιθανότητα σφάλματος τύπου ΙΙ που συμβολίζεται με β, δηλαδή $\beta = g_B / n_B$.[12]

Επιπλέον, ορίζονται και οι παρακάτω πιθανότητες:

- Ειδικότητα $= P(\text{γίνεται αποδεκτός ένας καλός πελάτης}) = \frac{g_G}{n_G} = 1 - \frac{b_G}{n_G} = 1 - \alpha$
- Ευαισθησία $= P(\text{απορρίπτεται ένας κακός πελάτης}) = \frac{b_B}{n_B} = 1 - \frac{g_B}{n_B} = 1 - \beta$

Επομένως, ο Πίνακας 4.3 είναι ο πίνακας σύγχυσης με βάση τα μέτρα της ευαισθησίας, της ειδικότητας και των σφαλμάτων τύπου Ι και ΙΙ.

Πίνακας 4.3 *Πίνακας σύγχυσης πιθανοτήτων*

Πραγματικές ομάδες / **Αναμενόμενες ομάδες**	**Καλοί**	**Κακοί**
Αποδεχτοί (αναμενόμενα καλοί)	$1-\alpha$	B
Απορριφθέντες (αναμενόμενα κακοί)	α	$1-\beta$

Τα στοιχεία του πίνακα σύγχυσης εξαρτώνται πάντα από τη βαθμολογία αποδοχής-απόρριψης. Αυτό σημαίνει ότι διαφορετική βαθμολογία αποδοχής-απόρριψης δίνει διαφορετικό πίνακα σύγχυσης. Επομένως, για να είναι αποτελεσματικό το μοντέλο θα πρέπει να βρεθεί η καταλληλότερη βαθμολογία αποδοχής-απόρριψης, ώστε να ελαχιστοποιούνται και τα δύο ποσοστά λάθους ταξινόμησης ή ισοδύναμα να μεγιστοποιούνται και η **ευαισθησία (*sensitivity*)** και η **ειδικότητα (*specificity*).** Όμως τα σφάλματα τύπου Ι και ΙΙ δεν μπορούν να

[12] Η πιθανότητα σφάλματος τύπου Ι είναι η πιθανότητα να απορριφτεί η μηδενική υπόθεση δεδομένου ότι αυτή ισχύει. Η πιθανότητα σφάλματος τύπου ΙΙ είναι η πιθανότητα να γίνει αποδεκτή η μηδενική υπόθεση δεδομένου ότι αυτή δεν ισχύει.

ελαχιστοποιούνται ταυτόχρονα, άρα η ευαισθησία και η ειδικότητα δε μπορούν να μεγιστοποιούνται ταυτόχρονα.

Εάν υποτεθεί ότι το κόστος για έναν χρηματοπιστωτικό οργανισμό είναι μεγαλύτερο αν γίνουν αποδεκτοί πελάτες που τελικά δεν θα μπορέσουν να εκπληρώσουν τις υποχρεώσεις τους σε σχέση με το να απορριφθούν «καλοί» πελάτες, τότε το καταλληλότερο σημείο αποκοπής θα είναι αυτό που θα ελαχιστοποιεί το σφάλμα τύπου ΙΙ και θα διατηρεί σε ένα χαμηλό επίπεδο το σφάλμα τύπου Ι.

4.3 Μέτρα διαχωριστικής ικανότητας

Κατά καιρούς έχουν προταθεί αρκετά μέτρα που αξιολογούν τη διαχωριστική ικανότητα του κάθε μοντέλου. Εάν έχει αναπτυχθεί ένα CSM και έχει δοθεί κάποια βαθμολογία σε κάθε υποψήφιο πελάτη, τότε μπορούν να χρησιμοποιηθούν αυτά τα μέτρα για να περιγραφεί πόσο διαφέρουν οι βαθμολογίες των «καλών» και των «κακών» πελατών. Επομένως, κάποια από αυτά τα μέτρα (όπως είναι το στατιστικό Kolmogorov – Smirnov και η απόσταση Mahalanobis) μπορούν να χρησιμοποιηθούν μόνο για αυτά τα μοντέλα που δίνουν κάποια βαθμολογία, όπως είναι τα μοντέλα παλινδρόμησης. Όμως, δεν μπορούν να χρησιμοποιηθούν σε μοντέλα που έχουν δημιουργηθεί με μεθόδους διαχωρισμού του πληθυσμού όπως είναι τα δέντρα ταξινόμησης (βλέπε Thomas (2002)).

Τα περισσότερα μέτρα διαχωρισμού εξαρτώνται μόνο από το ίδιο το CSM και είναι ανεξάρτητα από τη σχετική πιθανότητα των «καλών» του πληθυσμού ($odds(pop)$). Αυτό σημαίνει ότι μπορεί να χρησιμοποιηθεί ένα δείγμα το οποίο να μην έχει την ίδια σχετική πιθανότητα «καλοί»:«κακοί» με τον πληθυσμό. Όμως υπάρχουν και κάποια διαχωριστικά μέτρα που εξαρτώνται από την αναλογία «καλοί»:«κακοί» στο δείγμα του πληθυσμού με βάση το οποίο αναπτύσσεται το μοντέλο.

Επιπλέον, τα μέτρα διαχωρισμού δεν εξαρτώνται από το σημείο αποκοπής. Κατά συνέπεια αυτά τα μέτρα μπορούν να αποτελέσουν μια ένδειξη για το πόσο ανθεκτικό είναι το κάθε μοντέλο εάν αλλάξει το σημείο αποκοπής. Με αυτόν τον τρόπο τα μέτρα διαχωρισμού μπορούν επίσης να χρησιμοποιηθούν και για τον προσδιορισμό του βέλτιστου σημείου αποκοπής. Για ακριβείς εκτιμήσεις, αυτά τα μέτρα υπολογίζονται από το δείγμα ελέγχου που

είναι ανεξάρτητο από το δείγμα ανάπτυξης σύμφωνα με το οποίο δομήθηκε το μοντέλο. Ωστόσο κάποιες φορές, αυτά τα μέτρα μπορούν να υπολογιστούν χρησιμοποιώντας μόνο το δείγμα ανάπτυξης.

Τα πιο συνηθισμένα μέτρα της διαχωριστικής ικανότητας ενός σκορόχαρτου είναι η **απόκλιση**, η **τιμή πληροφορίας,** η **απόσταση Mahalanobis**, το **στατιστικό Kolmogorov–Smirnov**, οι **καμπύλες ROC**, ο **δείκτης Gini,** οι **καμπύλες CAP** και ο **δείκτης ακρίβειας AR.** Στη συνέχεια δίνουμε μια περιγραφή των πιο σημαντικών μέτρων.

α. Απόκλιση

Ένα από τα πιο απλά μέτρα διαχωρισμού είναι η απόκλιση D (*divergence*). Εάν δίνεται μια βαθμολογία s και $f(s|G)$, $f(s|B)$ είναι οι δεσμευμένες συναρτήσεις πυκνότητας των βαθμολογιών των «καλών» και των «κακών» αντίστοιχα, τότε η απόκλιση D δίνεται από τον τύπο

$$D = \int_{-\infty}^{+\infty} (f(s|G) - f(s|B)) \ln\left(\frac{f(s|G)}{f(s|B)}\right) ds = \int_{-\infty}^{+\infty} (f(s|G) - f(s|B)) w(s) ds \qquad (4.1)$$

όπου $w(s) = \ln\left(\frac{f(s|G)}{f(s|B)}\right)$ είναι το βάρος ένδειξης μιας βαθμολογίας s. Η απόκλιση (απόκλιση $K-L$) εισήχθηκε για πρώτη φορά από τους Kullback και Leibler (1951) ως ένας τρόπος μέτρησης της σχετικής απόστασης μιας πραγματικής κατανομής πιθανότητας και μιας άλλης που εκτιμάται από κάποιο μοντέλο. Εάν για παράδειγμα $p(x)$ είναι η συνάρτηση πυκνότητας της πραγματικής κατανομής και $q(x)$ είναι η συνάρτηση πυκνότητας της εκτιμώμενης κατανομής τότε η απόκλιση $K-L$ δίνεται από τον τύπο

$$D_{K-L}(q|p) = \int_{-\infty}^{+\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx. \qquad (4.2)$$

Εάν οι δύο αυτές κατανομές με συναρτήσεις πυκνότητας $p(x)$ και $q(x)$ είναι ίδιες τότε η ποσότητα $\ln\left(\frac{p(x)}{q(x)}\right)$ θα είναι 0, επομένως και η απόκλιση θα είναι 0. Η απόκλιση θα είναι μεγάλη εάν για κάποιο x θα ισχύει $p(x) >> q(x)$ ή $p(x) << q(x)$. Από τις σχέσεις (4.1) και (4.2) η απόκλιση D μπορεί να γραφεί συναρτήσει της απόστασης $D_{K-L}(q|p)$ ως εξής:

$$D = \int_{-\infty}^{+\infty} f(s|G) \ln\left(\frac{f(s|G)}{f(s|B)}\right) ds + \int_{-\infty}^{+\infty} f(s|B) \ln\left(\frac{f(s|B)}{f(s|G)}\right) ds = \\ = D_{K-L}(f(s|B) \,|\, f(s|G)) + D_{K-L}(f(s|G) \,|\, f(s|B)). \tag{4.3}$$

Επομένως, για κάθε CSM που κατασκευάζεται υπολογίζεται η απόκλιση D με βάση τον τύπο (4.3). Όσο μεγαλύτερο είναι αυτό το άθροισμα των αποστάσεων, τόσο καλύτερα είναι διαχωρισμένος ο πληθυσμός στις δύο ομάδες των «καλών» και των «κακών» και για αυτόν το λόγο η απόκλιση D μπορεί να θεωρηθεί ως ένα μέτρο διαχωριστικής ικανότητας. Επίσης, παρατηρούμε ότι, για τον υπολογισμό αυτού του μέτρου, δεν χρειάζεται να είναι γνωστά τα ποσοστά των «καλών» και των «κακών» πελατών στον πληθυσμό, δηλαδή το μέτρο αυτό δεν εξαρτάται από τη σχετική πιθανότητα των «καλών» του πληθυσμού $odds(pop)$. Αυτό σημαίνει ότι η απόκλιση θα δώσει τα ίδια αποτελέσματα σε ένα οποιοδήποτε δείγμα που έχει διαφορετικά ποσοστά «καλών» και «κακών» πελατών σε σχέση με το αρχικό.

β. Τιμή πληροφορίας

Εάν οι βαθμολογίες ακολουθούν κάποια συνεχή κατανομή τότε η απόκλιση μπορεί να υπολογιστεί από το ολοκλήρωμα του τύπου (4.1). Όμως, συνήθως οι βαθμολογίες υπολογίζονται από έναν πεπερασμένο πληθυσμό όπως είναι το δείγμα προηγούμενων πελατών με βάση τους οποίους κατασκευάστηκε το CSM ή το δείγμα ελέγχου. Σε αυτήν την περίπτωση, οι βαθμολογίες χωρίζονται σε κλάσεις και το ολοκλήρωμα του τύπου (4.1) αντικαθίσταται με το άθροισμα για τις κλάσεις που έχουν δημιουργηθεί. Αυτή η νέα ποσότητα (άθροισμα) αποτελεί την τιμή της πληροφορίας για τις κλάσεις των βαθμολογιών. Εάν οι βαθμολογίες χωριστούν σε l κλάσεις και $f(s_i|G)$, $f(s_i|B)$ είναι οι δεσμευμένες συναρτήσεις πιθανοφάνειας που προέρχονται από τα αρχικά σκορόχαρτα για κάθε $i = 1, 2, ..., l$, τότε η τιμή της πληροφορίας δίνεται από τον τύπο

$$IV = \sum_{i=1}^{l} (f(s_i|G) - f(s_i|B)) \ln\left(\frac{f(s_i|G)}{f(s_i|B)}\right).$$

Εάν g_i είναι ο αριθμός των «καλών» που ανήκουν στην κλάση i, b_i είναι ο αριθμός των «κακών» που ανήκουν στην κλάση i, όπου $i = 1, 2, ..., l$ τότε η τιμή της πληροφορίας είναι:

$$IV = \sum_{i=1}^{l} (g_i/n_G - b_i/n_B) \ln\left(\frac{g_i/n_G}{b_i/n_B}\right) = \sum_{i=1}^{l} (g_i/n_G - b_i/n_B) \ln\left(\frac{g_i n_B}{b_i n_G}\right).$$

Όπως ισχύει και με την απόκλιση, όσο μεγαλύτερη είναι η τιμή της πληροφορίας που υπολογίζεται με βάση το δείγμα ελέγχου, τόσο καλύτερη είναι η διαχωριστική ικανότητα του μοντέλου.

γ. Απόσταση Mahalanobis

Στην περίπτωση που οι βαθμολογίες των «καλων» και των «κακών» με πιθανοφάνειες $f(s|B)$, $f(s|B)$ είναι κανονικά κατανεμημένες με μέση τιμή μ_G, μ_B και διακύμανση σ_G^2, σ_B^2 αντίστοιχα, ισχύουν οι παρακάτω σχέσεις:

$$\int_{-\infty}^{+\infty} f(s|G)ds = 1 \quad \text{και} \quad \int_{-\infty}^{+\infty} f(s|B) = 1$$

Ροπή $1^{ης}$ τάξης: $$\int_{-\infty}^{+\infty} sf(s|G)ds = \mu_G \quad \text{και} \quad \int_{-\infty}^{+\infty} sf(s|B)ds = \mu_B \tag{4.4}$$

Ροπή $2^{ης}$ τάξης: $$\int_{-\infty}^{+\infty} s^2 f(s|G)ds = \sigma_G^2 + \mu_G^2 \quad \text{και} \quad \int_{-\infty}^{+\infty} s^2 f(s|B)ds = \sigma_B^2 + \mu_B^2$$

Η ροπή πρώτης τάξης είναι η μέση τιμή και η ροπή $2^{ης}$ τάξης είναι η διακύμανση συν το τετράγωνο της μέσης τιμής.

Η συνάρτηση πυκνότητας για τους «καλούς» πελάτες είναι

$$f(s|G) = \frac{1}{\sigma_G\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{s-\mu_G}{\sigma_G}\right)^2\right)$$

ενώ η συνάρτηση πυκνότητας για την υποομάδα των «κακών» πελατών είναι

$$f(s|B) = \frac{1}{\sigma_B\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{s-\mu_B}{\sigma_B}\right)^2\right).$$

Επομένως, το βάρος ένδειξης μιας βαθμολογίας s είναι

$$w(s) = \ln\left(\frac{f(s|G)}{f(s|B)}\right) = \ln\left(\frac{\sigma_B}{\sigma_G}\right) - \frac{1}{2}\left(\frac{s-\mu_G}{\sigma_G}\right)^2 + \frac{1}{2}\left(\frac{s-\mu_B}{\sigma_B}\right)^2. \tag{4.5}$$

Αντικαθιστώντας στον τύπο (4.1) τη σχέση (4.5) και χρησιμοποιώντας τις σχέσεις (4.4) έχουμε:

$$D=\int_{-\infty}^{+\infty}(f(s\,|\,G)-f(s\,|\,B))\ln\left(\frac{f(s\,|\,G)}{f(s\,|\,B)}\right)ds=$$

$$=\int_{-\infty}^{+\infty}(f(s\,|\,G)-f(s\,|\,B))\left(\ln\left(\frac{\sigma_B}{\sigma_G}\right)-\frac{1}{2}\left(\frac{s-\mu_G}{\sigma_G}\right)^2+\frac{1}{2}\left(\frac{s-\mu_B}{\sigma_B}\right)^2\right)ds=$$

$$=\left(\frac{1}{2\sigma_B^2}-\frac{1}{2\sigma_G^2}\right)\int_{-\infty}^{+\infty}s^2(f(s\,|\,G)-f(s\,|\,B))ds+\left(\frac{\mu_G}{\sigma_G^2}-\frac{\mu_B}{\sigma_B^2}\right)\int_{-\infty}^{+\infty}s(f(s\,|\,G)-f(s\,|\,B))ds+$$

$$+\left(\ln\left(\frac{\sigma_B}{\sigma_G}\right)-\frac{\mu_G}{\sigma_G^2}+\frac{\mu_B}{\sigma_B^2}\right)\int_{-\infty}^{+\infty}(f(s\,|\,G)-f(s\,|\,B))ds=$$

$$=\left(\frac{1}{2\sigma_B^2}-\frac{1}{2\sigma_G^2}\right)((\mu_G^2+\sigma_G^2)-(\mu_B^2+\sigma_B^2))+\left(\frac{\mu_G}{\sigma_G^2}-\frac{\mu_B}{\sigma_B^2}\right)(\mu_G-\mu_B)+$$

$$+\left(\ln\left(\frac{\sigma_B}{\sigma_G}\right)-\frac{\mu_G}{\sigma_G^2}+\frac{\mu_B}{\sigma_B^2}\right)(1-1)=$$

$$=\frac{1}{2}\left(\frac{1}{\sigma_B^2}-\frac{1}{\sigma_G^2}\right)(\mu_G-\mu_B)^2+\frac{(\sigma_G^2-\sigma_B^2)^2}{2\sigma_G^2\sigma_B^2}.$$

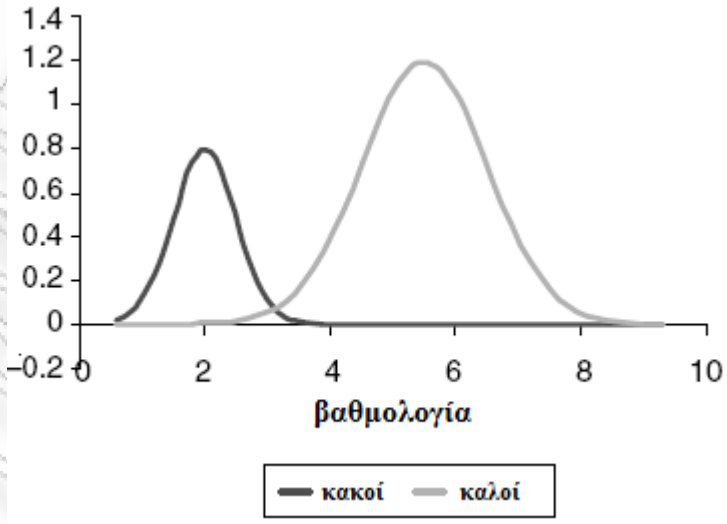
Στην ειδική περίπτωση που οι διακυμάνσεις των βαθμολογιών και των δύο ομάδων είναι ίσες, δηλαδή αν $\sigma_G^2=\sigma_B^2=\sigma^2$, τότε η απόκλιση δίνεται από τον τύπο

$$D=\frac{(\mu_G-\mu_B)^2}{\sigma^2}=\left(\frac{\mu_G-\mu_B}{\sigma}\right)^2=M^2$$

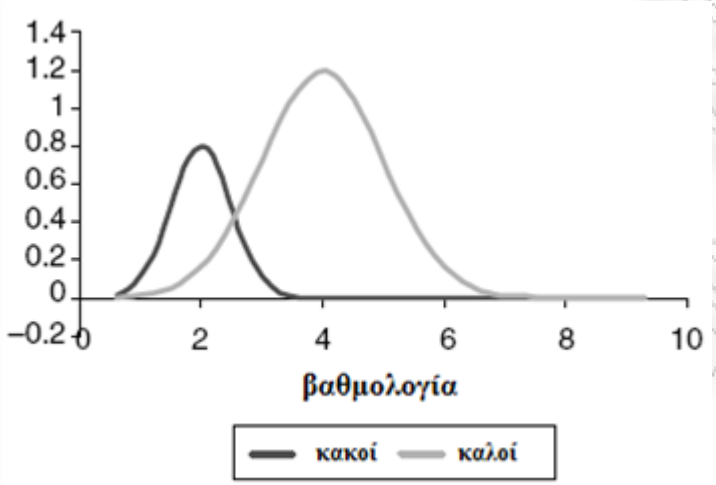
Παρατηρούμε ότι όταν οι βαθμολογίες των «καλών» και των «κακών πελατών είναι κανονικά κατανεμημένες με κοινή διακύμανση, η απόκλιση είναι το τετράγωνο της απόστασης Mahalonobis M. Την απόσταση αυτή τη συναντήσαμε στο προηγούμενο κεφάλαιο (Ενότητα 3.6), όπου αναλύθηκε η μέθοδος της διαχωριστικής συνάρτησης του Fisher και σκοπός ήταν να βρεθεί ο γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών $Y=\mathbf{bX}^T$ για τον οποίο μεγιστοποιείται η ποσότητα M που εκφράζει τη διαφορά μεταξύ του δειγματικού μέσου της Y για τους «καλούς» και του δειγματικού μέσου της Y για τους «κακούς» διαιρεμένη με την τυπική απόκλιση της Y. Αυτό το μέτρο εισήχθηκε για πρώτη φορά από τον Mahalanobis (1936) και στην περίπτωση της κατασκευής ενός CSM είναι ένα μέτρο που δείχνει πόσο διαφέρουν οι βαθμολογίες των «καλών» και των «κακών» πελατών. Όσο μεγαλύτερες είναι οι τιμές αυτού του μέτρου τόσο μεγαλύτερη είναι η διαχωριστική ικανότητα του μοντέλου.

Απαραίτητη προϋπόθεση για να θεωρηθεί η απόσταση Mahalanobis ως μέτρο διαχωρισμού είναι οι βαθμολογίες να είναι κανονικά κατανεμημένες και η διακύμανση των ομάδων να είναι κοινή. Η υπόθεση της κανονικότητας δεν ισχύει συνήθως για τις δεσμευμένες κατανομές βαθμολογιών, ωστόσο το μέτρο αυτό μπορεί να χρησιμοποιηθεί και σε περιπτώσεις που δεν ισχύουν οι προηγούμενες προϋποθέσεις γιατί η απόσταση Mahalanobis παρουσιάζει αρκετά πλεονεκτήματα. Σε ένα δείγμα, είναι πολύ εύκολο να υπολογιστούν η μέση τιμή και η διακύμανση των βαθμολογιών των «καλών» και των «κακών» και έτσι είναι δυνατό να αντιμετωπιστούν πολύ μεγάλα δείγματα. Επιπλέον, η απόσταση Mahalanobis είναι ένας εύκολος τρόπος να απεικονιστούν οι διαφορές μεταξύ των «καλών» και των «κακών» πελατών. Για παράδειγμα, στο Σχήμα 4.1 απεικονίζονται σε κοινό γράφημα οι ποσότητες $p_G f(s|G)$ και $p_B f(s|B)$ (για τους «καλούς» και τους «κακούς» πελάτες αντίστοιχα) για ένα CSM και στο Σχήμα 4.2 απεικονίζονται σε κοινό γράφημα οι ποσότητες $p_G f(s|G)$ και $p_B f(s|B)$ για ένα άλλο CSM. Παρατηρείται ότι το CSM του Σχήματος 4.1 έχει πολύ καλύτερη διαχωριστική ικανότητα σε σχέση με αυτό του Σχήματος 4.2. Αυτό συμβαίνει γιατί στο πρώτο σχήμα η διαφορά μεταξύ των μέσων είναι μεγαλύτερη σε σχέση με το δεύτερο.

Σχήμα 4.1 *Γράφημα των ποσοτήτων* $p_G f(s|G)$ και $p_B f(s|B)$ *Ισχυρή διαχωριστική ικανότητα (πηγή: Thomas (2009)).*



Σχήμα 4.2 *Γράφημα των ποσοτήτων* $p_G f(s|G)$ και $p_B f(s|B)$ *Ασθενής διαχωριστική ικανότητα (πηγή: Thomas (2009)).*



Στα δύο παραπάνω σχήματα η περιοχή κάτω από κάθε κατανομή αντιπροσωπεύει το ποσοστό του πληθυσμού που ανήκει σε εκείνη την ομάδα. Η απόσταση Mahalanobis είναι η απόσταση μεταξύ των μέσων των δύο καμπυλών.

Στην περίπτωση που δεν ισχύει η κανονικότητα των βαθμολογιών και η ισότητα των διακυμάνσεων, τότε ο αριθμός των «καλών» και των «κακών» που έχουν βαθμολογία s σε ένα δείγμα n πελατών όπου υπάρχουν n_G «καλοί» και n_B «κακοί» συμβολίζεται με $n_G(s)$ και $n_B(s)$ αντίστοιχα. Επομένως η πιθανότητα ένας «καλός» πελάτης να έχει βαθμολογία s είναι $P(s|G) = \frac{n_G(s)}{n_G}$, ενώ η πιθανότητα ένας «κακός» πελάτης να έχει βαθμολογία s είναι $P(s|B) = \frac{n_B(s)}{n_B}$. Οι μέσες βαθμολογίες των «καλών» και των «κακών» είναι αντίστοιχα

$$m_G = \sum_s sP(s|G) \text{ και } m_B = \sum_s sP(s|B).$$

Οι τυπικές αποκλίσεις σ_G και σ_B των βαθμολογιών των «καλών» και των «κακών» τότε υπολογίζονται αντίστοιχα από τους τύπους

$$\sigma_G^2 = \sum_s s^2 P(s|G) - m_G^2 \quad \text{και} \quad \sigma_B^2 = \sum_s s^2 P(s|B) - m_B^2.$$

Άρα, η διακύμανση του συνόλου των «καλών» και των «κακών» υπολογίζεται ως εξής:

$$s_p = \frac{n_G\sigma_G^2 + n_B\sigma_B^2}{n}.$$

Δηλαδή, η απόσταση Mahalanobis M είναι:

$$M = \left| \frac{m_G - m_B}{s_p} \right|.$$

Στις περιπτώσεις όπου δεν ισχύει η κανονικότητα και χρησιμοποιείται η απόσταση Mahalanobis, τα αποτελέσματα δεν είναι τόσο έγκυρα και είναι προτιμότερο να χρησιμοποιηθεί το μέτρο της απόκλισης ως μέτρο διαχωριστικής ικανότητας. Επίσης, παρατηρούμε ότι, εάν η τυπική απόκλιση είναι η ίδια για τους «καλούς» και τους «κακούς» τότε αυτή δεν εξαρτάται από τη σχετική πιθανότητα του πληθυσμού.

δ. Στατιστικό Kolmogorov - Smirnov

Ενώ η απόσταση Mahalanobis εκτιμά πόσο μακριά βρίσκονται οι μέσες τιμές των βαθμολογιών των «καλών» και των «κακών» πελατών, το στατιστικό Kolmogorov – Smirnov μετρά πόσο μακριά βρίσκονται οι αθροιστικές συναρτήσεις κατανομών των δύο αυτών υποπληθυσμών.

Η αθροιστική συνάρτηση κατανομής των «καλών» είναι :

$$F(s \mid G) = P(X \le s \mid G) = \sum_{x \le s} P(x \mid G)$$

και η συνάρτηση κατανομής των «κακών» είναι

$$F(s \mid B) = P(X \le s \mid B) = \sum_{x \le s} P(x \mid B).$$

(Στην περίπτωση που οι βαθμολογίες ακολουθούν συνεχή κατανομή τα αθροίσματα αντικαθίστανται από ολοκληρώματα).

Εάν οι ελάχιστες και οι μέγιστες βαθμολογίες συμβολιστούν με $s_{\min}$ και $s_{\max}$ αντίστοιχα τότε ισχύει ότι

$$F(s_{\min} \mid G) = F(s_{\min} \mid B) = 0 \quad \text{και} \quad F(s_{\max} \mid G) = F(s_{\max} \mid B) = 1.$$

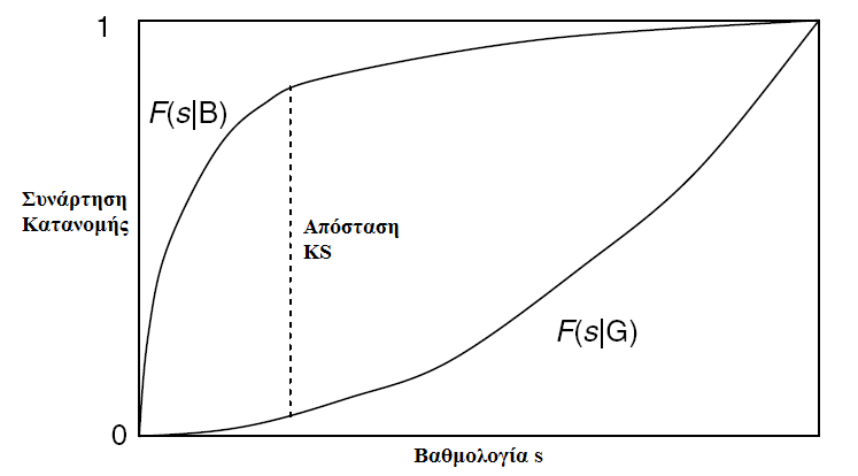
Αυτό σημαίνει ότι οι αθροιστικές συναρτήσεις κατανομών συμφωνούν και για τις δύο ομάδες μόνο για τις ακραίες βαθμολογίες, Όμως, αν οι συναρτήσεις κατανομών διαφέρουν αρκετά για άλλες τιμές των βαθμολογιών τότε το CSM έχει καλή διαχωριστική ικανότητα. Γι' αυτό το λόγο χρησιμοποιείται το στατιστικό Kolmogorov – Smirnov που ορίζεται ως εξής:

$$KS = \max_{s} \left| F(s \mid B) - F(s \mid G) \right|.$$

Στο Σχήμα 4.3 το στατιστικό Kolmogorov – Smirnov είναι το μήκος της διακεκομμένης γραμμής που μεγιστοποιεί την απόσταση μεταξύ των συναρτήσεων κατανομών $F(s \mid B)$ και $F(s \mid G)$. Αν συμβολιστεί με s^* η βαθμολογία που μεγιστοποιεί την ποσότητα KS τότε η παράγωγος της διαφοράς των συναρτήσεων κατανομών των «καλών» και των «κακών» $(F(s \mid B) - F(s \mid G))$ στο σημείο s^* θα είναι ίση με μηδέν. Αυτό σημαίνει ότι

$$f(s^* \mid B) - f(s^* \mid G) = 0.$$

Σχήμα 4.3 *Στατιστικό Kolmogorov – Smirnov (πηγή: Thomas (2009))*



Το μειονέκτημα του στατιστικού KS είναι ότι περιγράφει την κατάσταση για μια βέλτιστη βαθμολογία και αυτή σπάνια θα σχετίζεται με την καταλληλότερη βαθμολογία αποδοχής-απόρριψης με βάση την οποία λαμβάνεται η απόφαση. Γενικά, η απόσταση των δεσμευμένων συναρτήσεων κατανομών των δύο υποπληθυσμών για τη βαθμολογία αποδοχής – απόρριψης θα είναι μικρότερη από την ποσότητα KS. Αυτό σημαίνει ότι το στατιστικό KS αποτελεί το ανώτατο όριο ενός μέτρου απόστασης ή διαχωρισμού.

Αν υποθέσουμε ότι έχει καθοριστεί κάποιο σημείο αποκοπής s_c, οι πελάτες με υψηλότερες βαθμολογίες από σημείο αποκοπής ταξινομούνται ως «καλοί» και αυτοί με τις χαμηλότερες βαθμολογίες βαθμολογούνται ως «κακοί», τότε η ευαισθησία είναι η αναλογία των πραγματικών «κακών» που έχουν βαθμολογία χαμηλότερη από το σημείο αποκοπής κα η ειδικότητα θα είναι η αναλογία των πραγματικά «καλών» πελατών που έχουν βαθμολογία μεγαλύτερη από το σημείο αποκοπής.

Επομένως, το στατιστικό KS μπορεί να γραφεί ως εξής

$$KS = \max_s \left| F(s \mid B) - F(s \mid G) \right| = \max_s \left| \text{ευαισθησία} - (1 - \text{ειδικότητα}) \right| =$$

$$= \max_s \left| (\text{ευαισθησία+ειδικότητα}) \right| - 1.$$

Η βαθμολογία s_c που μεγιστοποιεί το άθροισμα της ευαισθησίας και της ειδικότητας αποτελεί τη βαθμολογία αποδοχής-απόρριψης.

ε. Καμπύλη ROC και συντελεστής Gini

Τα μέτρα διαχωριστικής ικανότητας που χρησιμοποιούνται συχνότερα είναι η καμπύλη ROC (*Receiver Operating Characteristic Curve*) και ο συντελεστής Gini (*Gini coefficient*) που σχετίζεται με την περιοχή (εμβαδό) κάτω από την καμπύλη ROC. Αυτή η μέθοδος εφαρμόστηκε για πρώτη φορά σε προβλήματα επεξεργασίας σήματος, ενώ τα τελευταία χρόνια χρησιμοποιείται για να εκτιμηθεί η ποιότητα ιατρικών διαγνώσεων.[13] Οι πρώτοι που εξήγησαν πως χρησιμοποιείται η καμπύλη ROC για την επικύρωση των CSM είναι οι Sobehart και Keeman (2001).

Η καμπύλη ROC είναι ένας τρόπος για να εκφραστεί η σχέση μεταξύ της αθροιστικής συνάρτησης κατανομής των βαθμολογιών των «καλών» και των «κακών». Πιο συγκεκριμένα, η καμπύλη ROC είναι η διδιάστατη γραφική απεικόνιση της αθροιστικής συνάρτησης κατανομής της βαθμολογίας των «κακών» $F(s \mid B) = P(X \le s \mid B)$ έναντι της αθροιστικής συνάρτησης κατανομής της βαθμολογίας των «καλών» $F(s \mid G) = P(X \le s \mid G)$. Εναλλακτικά, η καμπύλη ROC είναι η γραφική απεικόνιση της πιθανότητας της «σωστής απόρριψης ενός κακού πελάτη» (ευαισθησία) έναντι της πιθανότητας της λανθασμένης απόρριψης ενός καλού πελάτη» (1-ειδικότητα) για ένα εύρος διαφορετικών βαθμολογιών αποδοχής – απόρριψης. Η ευαισθησία ονομάζεται αλλιώς **πιθανότητα επιτυχίας (*hit rate*)**, ενώ η ποσότητα 1-ειδικότητα ονομάζεται αλλιώς **πιθανότητα εσφαλμένου συναγερμού (*false alarm rate*)**. Κατά τη διαδικασία αξιολόγησης ενός CSM, τέτοιου τύπου γραφικές παραστάσεις μπορούν να χρησιμοποιηθούν για να προσδιοριστεί η διαχωρισιμότητα του μοντέλου και ένα κατάλληλο σημείο αποκοπής.

[13] Για περισσότερες πληροφορίες και ενδιαφέρουσες εφαρμογές της καμπύλης ROC, βλέπε Swets (1988).

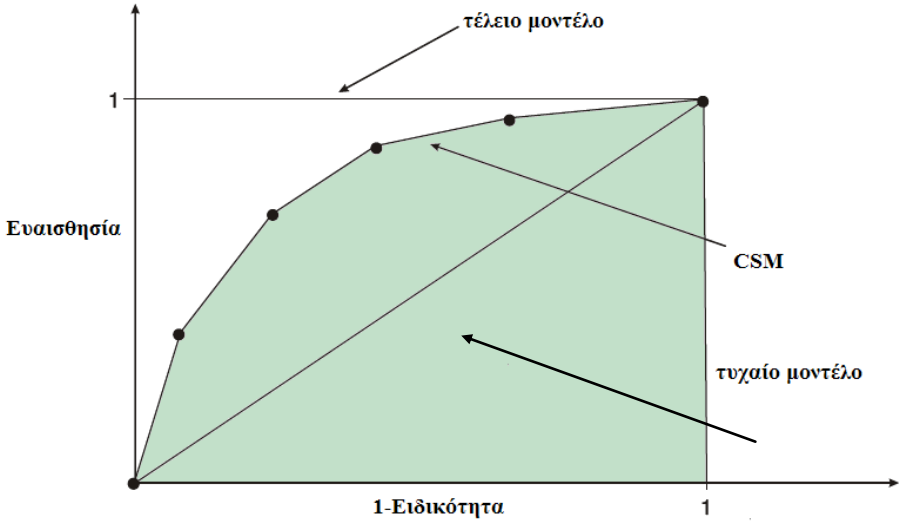
Για να κατασκευαστεί η καμπύλη ROC απαιτείται να υπολογιστούν οι ποσότητες ευαισθησία και 1 – ειδικότητα για κάθε δυνατό σημείο αποκοπής, όπου

$$a(s) = F(s \mid B) = P(X \le s \mid B) \text{ και}$$

$$1-\beta(s) = F(s \mid G) = P(X \le s \mid G).$$

Η καμπύλη των σημείων με συντεταγμένες $(a(s), 1-\beta(s))$ είναι η καμπύλη ROC, ένα παράδειγμα της οποίας δίνεται στο Σχήμα 4.4.

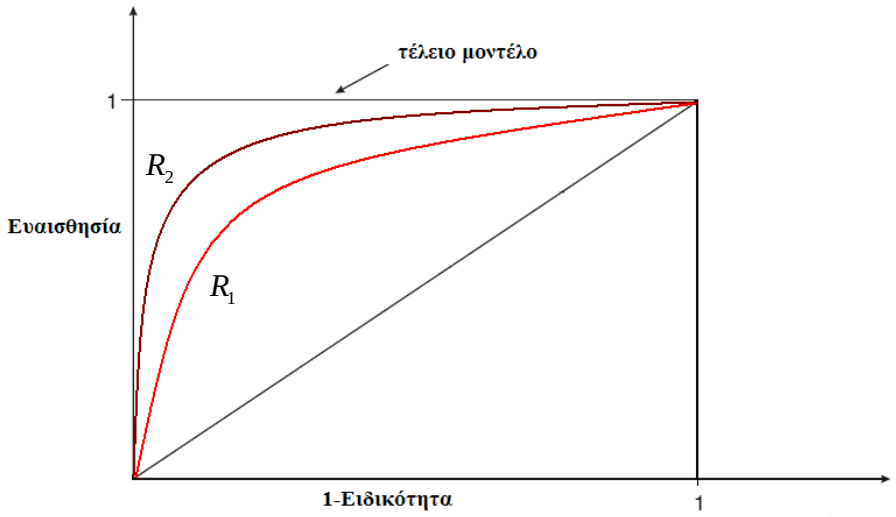
Σχήμα 4.4 *Καμπύλη ROC (πηγή: Tasche (2005))*



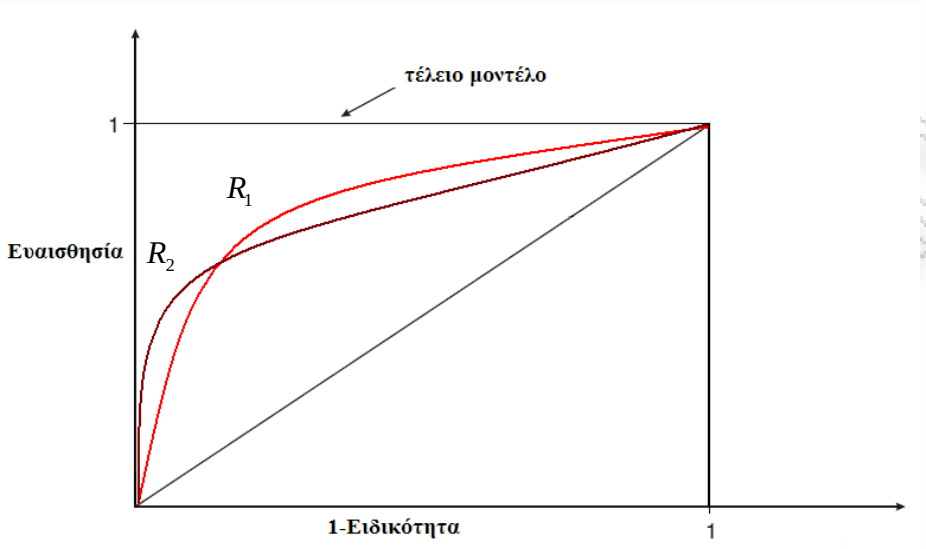
Εάν το CSM που αναπτύχθηκε είναι τέλειο, δηλαδή έχει την καλύτερη διαχωριστική ικανότητα, τότε θα υπάρχει μια βαθμολογία αποδοχής-απόρριψης s_c τέτοια ώστε όλοι οι πραγματικά «κακοί» να έχουν βαθμολογία μικρότερη από s_c και όλοι οι πραγματικά «καλοί» να έχουν βαθμολογία μεγαλύτερη από s_c. Για αυτό το σημείο ισχύει $F(s_c \mid G) = 0$ και $F(s_c \mid B) = 1$. Η καμπύλη ROC που αντιπροσωπεύει το τέλειο μοντέλο είναι η ευθεία γραμμή που φαίνεται στο Σχήμα 4.4. Επιπλέον, η διαγώνια γραμμή (ευαισθησία = 1 – ειδικότητα) αντιστοιχεί σε ένα μοντέλο που διαχωρίζει τυχαία τους «καλούς» και τους «κακούς» και δεν έχει καμιά διαχωριστική ικανότητα. Πράγματι, αν η κατανομή των βαθμολογιών στους «καλούς» και στους «κακούς» συμπίπτει, το μοντέλο δεν θα έχει κάποια διαχωριστική ικανότητα, δηλαδή αν $F(s \mid G) = F(s \mid B)$ τότε η καμπύλη ROC είναι η διαγώνιος.

Αν συγκρίνουμε δύο διαφορετικές καμπύλες ROC, R_1 και R_2 που η κάθε μία αντιστοιχεί σε διαφορετικό σκορόχαρτο, τότε το μοντέλο που έχει την καλύτερη διαχωριστική ικανότητα είναι αυτό του οποίου η αντίστοιχη καμπύλη βρίσκεται πιο κοντά στο σημείο $(0,1)$ ή θα είναι αυτό που πλησιάζει περισσότερο το τέλειο μοντέλο. Για παράδειγμα στο Σχήμα 4.5 φαίνεται ότι το μοντέλο που αντιστοιχεί στη καμπύλη R_1 διαχωρίζει καλύτερα τους «καλούς» και τους «κακούς» πελάτες γιατί η διαφορά $F(s|B)-F(s|G)$ θα είναι πάντα μεγαλύτερη σε σχέση με το άλλο μοντέλο. Εάν οι καμπύλες R_1 και R_2 τέμνονται σε κάποια σημεία όπως φαίνεται στο Σχήμα 4.6 τότε το ένα μοντέλο είναι καλύτερο σε μια περιοχή σημείων αποκοπής, ενώ το άλλο είναι καλύτερο σε μια άλλη περιοχή σημείων αποκοπής. Για παράδειγμα, στο Σχήμα 4.6 για χαμηλές βαθμολογίες αποδοχής – απόρριψης το μοντέλο που αντιστοιχεί στην καμπύλη R_2 είναι καλύτερο σε σχέση με το άλλο, ενώ για υψηλές βαθμολογίες αποδοχής – απόρριψης είναι καλύτερο το μοντέλο που αντιστοιχεί στην καμπύλη R_1. Συνήθως, οι χρηματοπιστωτικοί οργανισμοί προτιμούν να δεχτούν ένα μεγάλο ποσοστό από «καλούς» πελάτες (μεγάλη ευαισθησία), άρα προτιμάται το μοντέλο του οποίου τα σημεία αποκοπής τείνουν να βρίσκονται στην περιοχή πιο κοντά στα αριστερά του γραφήματος.

Σχήμα 4.5 *Καμπύλες ROC που δεν τέμνονται*



Σχήμα 4.6 *Καμπύλες ROC που τέμνονται*



Η καμπύλη ROC αποτελεί μια διδιάστατη περιγραφή της απόδοσης ενός CSM. Προκειμένου όμως να συγκριθούν διαφορετικά CSM θα ήταν προτιμότερο ένα μονοδιάστατο μέτρο απόδοσης. Ένα τέτοιο μέτρο που χρησιμοποιείται είναι το εμβαδό κάτω από την καμπύλη ROC που ονομάζεται AUROC ή AUC (*Area Under Curve*). Θεωρητικά, αφού το AUC είναι μέρος ενός τετραγώνου πλευράς 1, μπορεί να πάρει τιμές μεταξύ 0 και 1. Όμως, το τυχαίο μοντέλο που αντιστοιχεί στη διαγώνιο έχει εμβαδό $AUC = 1 \cdot 1 \cdot 0,5 = 0,5$. Αυτό σημαίνει ότι κανένα ρεαλιστικό μοντέλο δεν θα έχει AUC μικρότερο από 0,5, δηλαδή για κάθε μοντέλο αναμένεται το ποσοστό των «κακών» πελατών που απορρίπτονται να είναι υψηλότερο από το ποσοστό των «καλών» πελατών που απορρίπτονται. Κατά συνέπεια η καμπύλη ROC θα βρίσκεται πάνω από τη διαγώνιο. Είναι προφανές λοιπόν ότι, όσο η καμπύλη πλησιάζει την πάνω αριστερή γωνία του τετραγώνου της γραφικής παράστασης, δηλαδή όσο μεγαλύτερο είναι το εμβαδό κάτω από την καμπύλη (AUC), τόσο πιο καλή είναι η απόδοση του CSM. Συνήθως, το σημείο που είναι το πλησιέστερο σε αυτή την πάνω αριστερή γωνία του τετραγώνου επιλέγεται ως σημείο αποκοπής καθώς αυτό μεγιστοποιεί ταυτόχρονα και την ευαισθησία και την ειδικότητα (Κατέρη (2008)). Το εμβαδό αυτό υπολογίζεται από τον τύπο

$$AUC = \int_0^1 (1 - \beta(s)) d(a(s)) .$$

Η ποσότητα AUC σε σχέση με τις αθροιστικές συναρτήσεις κατανομών είναι

$$AUC = \int_0^1 F(s \mid B) dF(s \mid G) = \int_{-\infty}^{+\infty} F(s \mid B) f(s \mid G) ds .$$

Αντί της ποσότητας AUC θα ήταν προτιμότερο ως μέτρο διαχωρισμού να χρησιμοποιείται μια ποσότητα η οποία θα παίρνει την τιμή 0 όταν η διαχωριστική ικανότητα του μοντέλου δεν έχει νόημα (τυχαία) και την τιμή 1 όταν το μοντέλο θα έχει τέλεια διαχωριστική ικανότητα. Αυτό το νέο μέτρο που είναι ένας μετασχηματισμός του AUC είναι ο συντελεστής Gini. Ο τύπος με βάση τον οποίο υπολογίζεται ο συντελεστής Gini είναι

$$\begin{aligned} GINI &= 2AUC - 1 = 2\int_{-\infty}^{+\infty} F(s \mid B) f(s \mid G) ds - 1 = 2\int_{-\infty}^{+\infty} F(s \mid B) f(s \mid G) ds - 1 = \\ &= 2\int_{-\infty}^{+\infty} F(s \mid B) f(s \mid G) ds - 2\int_{-\infty}^{+\infty} F(s \mid G) f(s \mid G) ds = \\ &= 2\int_{-\infty}^{+\infty} \left(F(s \mid B) - F(s \mid G) \right) f(s \mid G) ds \end{aligned} \tag{4.6}$$

αφού με παραγοντική ολοκλήρωση έχουμε

$$\int_{-\infty}^{+\infty} F(s \mid G) f(s \mid G) ds = \left[F(s \mid G)^2 \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} f(s \mid G) F(s \mid G) ds \Rightarrow 2\int_{-\infty}^{+\infty} F(s \mid G) f(s \mid G) ds = 1.$$

Από την τελευταία ισότητα του τύπου (4.6) συμπεραίνουμε ότι ο συντελεστής Gini είναι δύο φορές το εμβαδόν της περιοχής μεταξύ της καμπύλης ROC και της διαγωνίου, αφού το ύψος της διαγωνίου στο σημείο όπου η καμπύλη ROC έχει συντεταγμένες $(F(s \mid G), F(s \mid B))$ ισούται με την απόσταση του σημείου από τον οριζόντιο άξονα, δηλαδή $F(s \mid G)$. Η περιοχή κάτω από τη διαγώνιο είναι 0,5. Άρα, το εμβαδό μεταξύ της καμπύλης ROC και της διαγωνίου είναι $AUC - 0,5$. Η ποσότητα $AUC - 0,5$ είναι ένας άλλος τρόπος υπολογισμού του συντελεστή Gini και μπορεί να περιγραφεί σαν μια βελτίωση της περιοχής του μοντέλου σε σχέση με το μοντέλο με την τυχαία διαχωριστική ικανότητα.

Για να σχεδιαστεί η καμπύλη ROC και για να υπολογιστεί η ποσότητα AUC ή ο συντελεστής Gini δεν χρειάζεται να είναι γνωστή η σχετική πιθανότητα του πληθυσμού *odds*(*pop*). Οπότε, όλα αυτά τα μέτρα είναι ανεξάρτητα από τον τρόπο με τον οποίο επιλέγεται το δείγμα (ανάπτυξης ή ελέγχου) με βάση το οποίο δομείται το CSM από τον αρχικό πληθυσμό, αλλά εξαρτώνται μόνο από τις ιδιότητες του μοντέλου.

Σε κάθε βαθμολογία s του σκορόχαρτου αντιστοιχεί ένα σημείο της καμπύλης ROC με συντεταγμένες $(F(s \mid G), F(s \mid B))$. Η απόσταση του κάθε σημείου από τη διαγώνιο ισούται με $\left| F(s \mid B) - F(s \mid G) \right|$, αφού η διαγώνιος έχει, σε αυτό το σημείο, ύψος $F(s \mid G)$. Παρατηρούμε λοιπόν ότι η καμπύλη ROC σχετίζεται με το στατιστικό KS, αφού όπως φαίνεται το στατιστικό KS είναι η μέγιστη δυνατή απόσταση μεταξύ της καμπύλης ROC και της διαγωνίου.

Ο συντελεστής Gini και η ποσότητα *AUC* είναι αριθμοί που δίνουν κάποια ένδειξη για την απόδοση του κάθε μοντέλου για όλες τις βαθμολογίες αποδοχής-απόρριψης. Όμως στην πράξη είναι πιο χρήσιμο να χρησιμοποιούνται μέτρα απόδοσης μοντέλων για ένα μικρό εύρος πιθανών βαθμολογιών αποδοχής-απόρριψης. Το ίδιο βέβαια συμβαίνει και για τα υπόλοιπα μέτρα όπως είναι το στατιστικό KS και η απόσταση Mahalnobis, διότι όλα αυτά περιγράφουν γενικές ιδιότητες του σκορόχαρτου, ενώ στην πράξη είναι σημαντικό να μετριέται η απόδοσή του για κάποιο συγκεκριμένο σημείο αποκοπής.

Η καμπύλη ROC εκτός από τη μέτρηση της απόδοσης ενός CSM μπορεί να χρησιμοποιηθεί και για να προσδιοριστεί το καταλληλότερο σημείο αποκοπής. Για παράδειγμα, το σημείο που μεγιστοποιεί το στατιστικό KS είναι αυτό που αντιστοιχεί στο σημείο της καμπύλης για το οποίο μεγιστοποιείται η απόσταση $|F(s|B)-F(s|G)|$. Ένα τέτοιο σημείο είναι το σημείο C στο Σχήμα 4.7. Εάν υποτεθεί ότι τα πραγματικά ποσοστά των «καλών» και των «κακών» πλατών στον πληθυσμό είναι p_G και p_B αντίστοιχα, και L, D είναι τα αντίστοιχα κόστη λανθασμένης ταξινόμησης, τότε το αναμενόμενο κόστος αν η βαθμολογία αποδοχής-απόρριψης είναι s_c υπολογίζεται από τον τύπο

$$L\cdot F(s_c|G)p_G + D\cdot(1-F(s_c|B))p_B. \qquad (4.7)$$

Αν συμβολιστεί με $f(x)$ η καμπύλη ROC τότε θα επιλεχτεί ως σημείο αποκοπής αυτό που ελαχιστοποιεί την ποσότητα $L\cdot p_G\cdot x + D\cdot p_B(1-f(x))$. Αυτή η ποσότητα ελαχιστοποιείται όταν

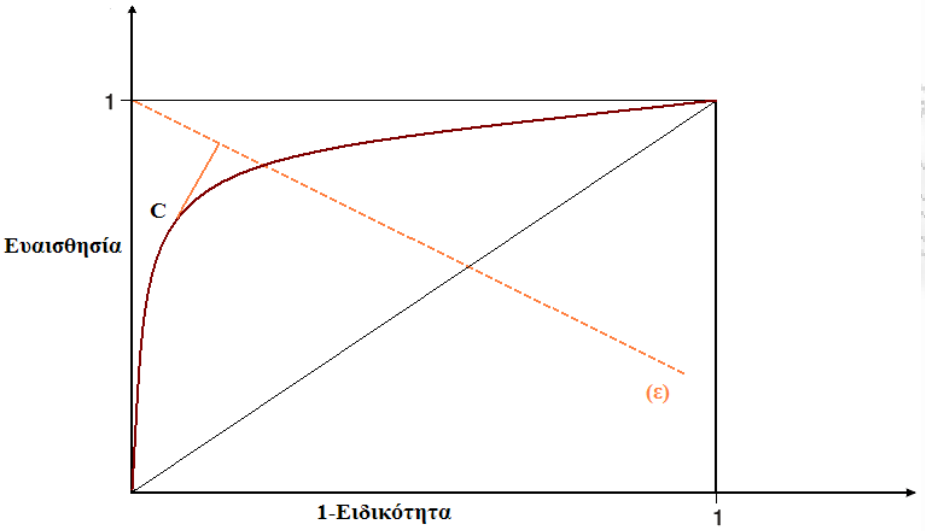
$$\frac{\partial(L\cdot p_G\cdot x + D\cdot p_B(1-f(x)))}{\partial x} = 0 \Rightarrow L\cdot p_G - D\cdot p_B\cdot f'(x) = 0 \Rightarrow f'(x) = \frac{L\cdot p_G}{D\cdot p_B}.$$

και

$$\frac{\partial(L\cdot p_G - D\cdot p_B\cdot f'(x))}{\partial x} = -D\cdot p_B\cdot f''(x) < 0, \text{ αφού } D>0 \text{ και } p_B>0.$$

Άρα, στο σημείο που ελαχιστοποιείται η ποσότητα (4.7), η κλίση της εφαπτομένης της καμπύλης ROC είναι $\frac{L\cdot p_G}{D\cdot p_B}$. Ένας τρόπος για να βρεθεί αυτό το σημείο είναι να σχεδιαστεί από το σημείο $(0,1)$ μια ευθεία (ε) με κλίση $-\frac{D\cdot p_B}{L\cdot p_G}$. Το σημείο της καμπύλης ROC που προβάλλεται στην ευθεία (ε) και βρίσκεται πιο κοντά στο σημείο $(0,1)$ θα είναι αυτό που αναζητάμε. Το βέλτιστό σημείο αποκοπής είναι το σημείο C στο Σχήμα 4.7.

Σχήμα 4.7 *Χρήση της καμπύλης ROC για να βρεθεί το σημείο αποκοπής*



Δοθέντων των βαθμολογιών που προκύπτουν από ένα CSM για κάθε πελάτη του δείγματος επικύρωσης, μια πολύ χρήσιμη ιδιότητα του AUC είναι ότι μπορεί να εκτιμηθεί με ένα μη παραμετρικό τρόπο χρησιμοποιώντας το στατιστικό Mann – Whitney. Θεωρώντας ότι το AUC είναι κανονικά κατανεμημένο είναι δυνατό να προκύψουν διαστήματα εμπιστοσύνης για αυτήν την περιοχή.[14]

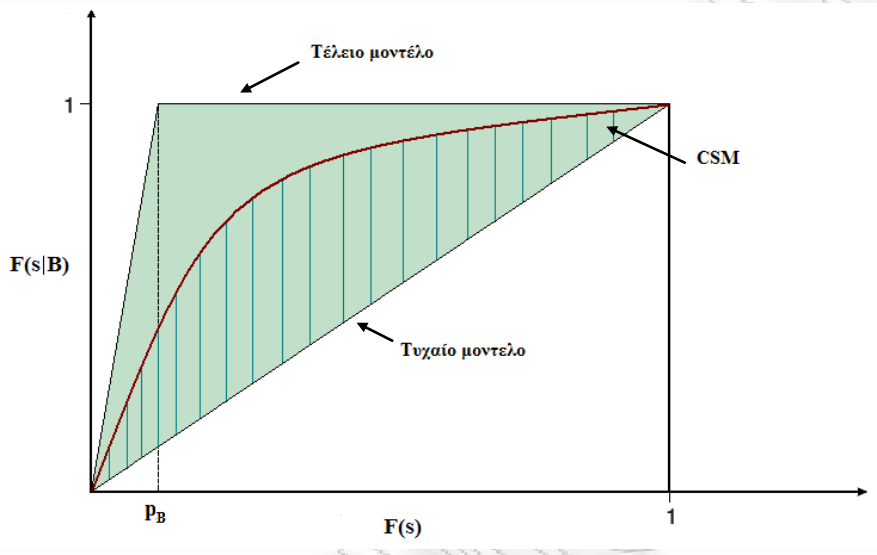
στ. Καμπύλη CAP και δείκτης ακρίβειας

Ένα άλλο μέτρο διαχωριστικής ικανότητας ενός CSM που είναι παρόμοιο με την καμπύλη ROC είναι η συσσωρευτική καμπύλη ακρίβειας (CAP - *Cumulative Accuracy Profile*). Η διαφορά της καμπύλης CAP με την καμπύλη ROC είναι ότι αντί να έχουμε τη διδιάστατη γραφική απεικόνιση της αθροιστικής συνάρτησης κατανομής της βαθμολογίας των «κακών» $F(s|B)$ έναντι της αθροιστικής συνάρτησης κατανομής της βαθμολογίας των «καλών» $F(s|G)$ για κάθε βαθμολογία s, έχουμε τη διδιάστατη γραφική απεικόνιση της αθροιστικής συνάρτησης κατανομής της βαθμολογίας των «κακών» $F(s|B)$, έναντι της αθροιστικής συνάρτησης κατανομής $F(s)$ για κάθε βαθμολογία s. Επομένως, ο οριζόντιος άξονας δίνει το ποσοστό των υποψηφίων που απορρίπτονται και ο κάθετος άξονας δίνει το ποσοστό των

[14] Περισσότερες πληροφορίες για το πώς προκύπτουν τα διαστήματα εμπιστοσύνης, Bamber (1975), Engleman, Hayden και Tasche (2003).

«κακών» που απορρίπτονται. Ένα παράδειγμα της μορφής που έχει η καμπύλη CAP δίνεται στο Σχήμα 4.8.

Σχήμα 4.8 *Καμπύλη CAP*



Η καμπύλη CAP για το CSM που έχει την καλύτερη διαχωριστική ικανότητα θα περνάει από το σημείο $(p_B,1)$, όπου p_B είναι το ποσοστό των «κακών» πελατών στον πληθυσμό. Η καμπύλη για το τέλειο μοντέλο αποτελείται από δύο ευθείες. Η μία είναι η ευθεία που καθορίζεται από τα σημεία $(0,0)$ και $(p_B,1)$ για την οποία ισχύει ότι όσο πηγαίνει ψηλότερα το σημείο αποκοπής θα είναι μεγαλύτερη η πιθανότητα απόρριψης των «κακών» υποψηφίων. Η άλλη ευθεία καθορίζεται από τα σημεία $(p_B,1)$ και $(1,1)$ για την οποία όσο πηγαίνει προς τα δεξιά το σημείο αποκοπής μεγαλώνει η πιθανότητα απόρριψης των «καλών» υποψηφίων. Η καμπύλη CAP για το CSM που ισοδυναμεί με το τυχαίο μοντέλο (του οποίου οι βαθμολογίες δε δίνουν καμιά πληροφορία) αντιστοιχεί στη διαγώνια γραμμή του σχήματος αφού για οποιοδήποτε σημείο αποκοπής η βαθμολογία των «καλών» που έχουν μικρότερη βαθμολογία από το σημείο αποκοπής είναι ίση με το ποσοστό του πληθυσμού που έχουν μικρότερη βαθμολογία από το σημείο αποκοπής, δηλαδή $F(s)=F(s\,|\,B)$.

Όπως και η καμπύλη ROC, η καμπύλη CAP είναι μια διδιάστατη περιγραφή της απόδοσης ενός CSM. Ένα μονοδιάστατο μέτρο της απόδοσης ενός CSM που προέρχεται από την καμπύλη CAP είναι το εμβαδό της περιοχής μεταξύ της καμπύλης και της διαγωνίου προς το

εμβαδό της περιοχής μεταξύ του μοντέλου με την τέλεια απόδοση και της διαγωνίου. Αυτό το πηλίκο ονομάζεται δείκτης ακρίβειας (*AR - Accuracy Ratio*), για το οποίο ισχύει ότι

$$AR = \frac{\text{Εμβαδό μεταξύ της CAP και διαγωνίου}}{\text{Εμβαδό μεταξύ καμπύλης τέλειου μοντέλου και διαγωνίου}} =$$

$$= \frac{\text{Εμβαδό μεταξύ της CAP και διαγωνίου}}{0,5(1-p_B)} = \frac{\int_{-\infty}^{+\infty} F(s \mid B) f(s) ds - 0,5}{0,5(1-p_B)} =$$

$$= \frac{2}{1-p_B}\left(\int_{-\infty}^{+\infty} F(s \mid B) f(s) ds - 0,5\right).$$

Αν και οι καμπύλες ROC και CAP είναι διαφορετικές, οι Engleman, Hayden και Tache (2003) έδειξαν ότι ο συντελεστής Gini και ο δείκτης ακρίβειας συνήθως συμπίπτουν, δηλαδή ο δείκτης ακρίβειας είναι ένας γραμμικός συνδυασμός της AUC διότι

$$AR = \frac{2}{1-p_B}\left(\int_{-\infty}^{+\infty} F(s \mid B) f(s) ds - 0,5\right) =$$

$$= \frac{2}{1-p_B}\left(\int_{-\infty}^{+\infty} F(s \mid B)(p_B f(s \mid B) + p_G f(s \mid G)) ds - 0,5\right) =$$

$$= \frac{2}{1-p_B}\left(p_B \int_{-\infty}^{+\infty} F(s \mid B) f(s \mid B) ds + p_G AUC - 0,5\right) =$$

$$= \frac{2}{1-p_B}(p_B \cdot 0,5 + (1-p_B) AUC - 0,5) =$$

$$= \frac{2}{1-p_B}((1-p_B) AUC - 0,5(1-p_B)) =$$

$$= \frac{2}{1-p_B}(1-p_B)(AUC - 0,5) = 2 \cdot AUC - 1 =$$

$$= GINI.$$

αφού με παραγοντική ολοκλήρωση έχουμε

$$\int_{-\infty}^{+\infty} F(s \mid B) f(s \mid B) ds = \left[F(s \mid B)^2\right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} f(s \mid B) F(s \mid B) ds \Rightarrow 2\int_{-\infty}^{+\infty} F(s \mid B) f(s \mid B) ds = 1.$$

Η καμπύλη ROC χρησιμοποιείται συχνότερα ως μέτρο διαχωριστικής ικανότητας ενός CSM σε σχέση με την καμπύλη CAP. Αυτό συμβαίνει διότι η καμπύλη ROC μπορεί να κατασκευαστεί σε οποιοδήποτε δείγμα ανεξάρτητα από τα ποσοστά των «καλών» και των «κακών» πελατών στον πληθυσμό. Από την άλλη πλευρά, για να κατασκευαστεί μια καμπύλη CAP πρέπει να είναι γνωστή η πληθυσμιακή σχετική πιθανότητα, διότι η καμπύλη θα είναι

διαφορετική για διαφορετική κατανομή του πληθυσμού. Αυτό προφανώς συμβαίνει γιατί για να κατασκευαστεί η καμπύλη CAP είναι απαραίτητο να είναι γνωστό το p_B που είναι το πραγματικό ποσοστό των «κακών» πελατών στον αρχικό πληθυσμό.

4.4 Ακρίβεια προσαρμογής της εκτίμησης της πιθανότητας

Αρχικά τα CSM αναπτύσσονταν με σκοπό την υποστήριξη των διαδικασιών λήψης αποφάσεων που αφορούν την πιστοληπτική συμπεριφορά υποψήφιων πελατών και διαχωρίζοντας τους ως προς την εκτίμηση αν αυτοί θα αθετήσουν τις υποχρεώσεις τους ή όχι. Πιο σημαντικό ήταν να δομηθεί ένα CSM που να διατάσσει σωστά τους υποψηφίους ως προς τον κίνδυνο αθέτησης των υποχρεώσεών τους παρά η σωστή εκτίμηση της πιθανότητας αθέτησης υποχρεώσεων. Ο προσδιορισμός της βαθμολογίας αποδοχής-απόρριψης γινόταν κυρίως εμπειρικά και εξαρτιόταν από διάφορους επιχειρηματικούς στόχους όπως να γίνει αποδεκτό κάποιο συγκεκριμένο ποσοστό των πελατών. Όμως, η βαθμολογία του κάθε υποψηφίου μπορεί να ληφθεί υπ' όψιν για να προβλεφθεί η πιθανότητα αθέτησης υποχρεώσεων. Για να προβλεφθεί η πιθανότητα αθέτησης υποχρεώσεων ή η πιθανότητα ένας υποψήφιος να είναι «καλός» χρειάζονται επιπλέον πληροφορίες εκτός από τις πιθανοφάνειες $f(s|G)$ και $f(s|B)$. Ενδεικτικά αναφέρουμε ότι, για να υπολογιστούν οι πιθανότητες $P(G|s)$ και $P(B|s)$ θα πρέπει να είναι γνωστή η σχετική πιθανότητα του πληθυσμού.

Όμως, σύμφωνα με τη νέα συμφωνία της Βασιλείας που εισήχθηκε το 2007, ένα χαρτοφυλάκιο δανείων πρέπει πρώτα να χωριστεί σε κλάσεις βαθμολογιών και έπειτα για κάθε κλάση να δοθεί μια πιθανότητα αθέτησης υποχρεώσεων (*PD – Probability of Default*) των υποψηφίων που ανήκουν σε αυτήν τη κλάση. Αν ένας υποψήφιος έχει πιθανότητα $P(s)=P(G|s)$ να είναι «καλός», τότε η πιθανότητα αθέτησης, δηλαδή η πιθανότητα ένας υποψήφιος να είναι «κακός» είναι $PD=1-P(s)=P(B|s)$. Η πιθανότητα αθέτησης υποχρεώσεων επικυρώνεται συγκρίνοντας τις προβλεπόμενες τιμές με τις πραγματικές, χρησιμοποιώντας το δείγμα ελέγχου. Επομένως, δημιουργείται η ανάγκη κατασκευής ελέγχων που θα επικυρώνουν την ακρίβεια πρόβλεψης της πιθανότητας αθέτησης υποχρεώσεων που προκύπτουν από τις βαθμολογίες. Η **ακρίβεια προσαρμογής (*accuracy***

***calibration*)** της εκτίμησης της πιθανότητας χρησιμοποιείται στο νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ ως μέθοδος μέτρησης της απόδοσης ενός CSM.

Όπως είδαμε, η λογιστική παλινδρόμηση δίνει τη σχέση μεταξύ της βαθμολογίας που εκφράζεται από το λογάριθμο της σχετικής πιθανότητας και πιθανότητας ένας πελάτης να είναι «καλός». Επομένως, αν η μέθοδος ανάπτυξης ενός CSM είναι η λογιστική παλινδρόμηση τότε προκύπτει άμεσα η πιθανότητα αθέτησης υποχρεώσεων. Όμως, υπάρχουν άλλες μέθοδοι κατασκευής CSM από τις οποίες δεν προκύπτει άμεσα η σχέση μεταξύ της βαθμολογίας και της πιθανότητας αθέτησης υποχρεώσεων, δηλαδή δεν έχουν την ιδιότητα που έχει ο λογάριθμος του λόγου πιθανοφανειών. Μια τέτοια συνάρτηση μετασχηματισμού της βαθμολογίας σε πιθανότητα αθέτησης υποχρεώσεων μπορεί να δημιουργηθεί με διάφορους τρόπους ένας από τους οποίους είναι η χρήση ιστορικών δεδομένων. Τα **μέτρα ακρίβειας προσαρμογής (*calibration measures*)** της εκτίμησης της πιθανότητας μετράνε πόσο καλές είναι οι προβλέψεις των πιθανοτήτων αθέτησης υποχρεώσεων με οποιοδήποτε τρόπο και αν έχουν εκτιμηθεί αυτές.

Για να μετρηθεί η προσαρμογή των βαθμολογιών έτσι ώστε οι εκτιμήσεις να αντικατοπτρίζουν την αναμενόμενη συμπεριφορά τους, οι βαθμολογίες χωρίζονται σε ζώνες (κλάσεις) και για κάθε ζώνη υπολογίζεται μια πιθανότητα αθέτησης υποχρεώσεων. Εάν οι βαθμολογίες χωρίζονται έτσι ώστε στην i ζώνη να περιλαμβάνεται το διάστημα των βαθμολογιών $[s_i, s_{i+1})$, τότε η πιθανότητα αθέτησης υποχρεώσεων PD_i για αυτή τη ζώνη ορίζεται ως εξής:

$$PD_i = \frac{\int_{s_i}^{s_{i+1}} P(B \mid s) f(s) ds}{\int_{s_i}^{s_{i+1}} f(s) ds} = \frac{\int_{s_i}^{s_{i+1}} s \cdot p_B f(s) ds}{\int_{s_i}^{s_{i+1}} f(s) ds}.$$

Εναλλακτικά, όπως προτείνεται με το νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ, η πιθανότητα αθέτησης υποχρεώσεων μπορεί να αντιμετωπιστεί ως συνάρτηση της βαθμολογίας $PD(s)$ για κάθε πελάτη και έπειτα να επιλεχτούν κλάσεις για την πιθανότητα PD. Οπότε, η κλάση i περιλαμβάνει όλους τους υποψηφίους με βαθμολογία s τέτοια ώστε να ισχύει

$$\left\{ s \middle| PD_i \le PD(s) < PD_{i+1} \right\} = \left\{ s \middle| PD_i \le PD(B \mid s) < PD_{i+1} \right\}$$

όπου η πιθανότητα PD σε ένα τέτοιο διάστημα είναι συνήθως ο μέσος του διαστήματος αυτού, δηλαδή $(PD_i + PD_{i+1})/2$.

Η προσαρμογή της ακρίβειας των προβλέψεων είναι πολύ σημαντική και χρησιμοποιείται εδώ και πολλά χρόνια σε αντικείμενα όπως είναι οι προβλέψεις των πωλήσεων μιας επιχείρησης, οι προβλέψεις για την οικονομία ακόμα και σε προβλέψεις καιρού. Όμως, αυτή η προσέγγιση για την εκτίμηση της απόδοσης των CSM έγινε σημαντική τα τελευταία χρόνια και έτσι έχουν προταθεί ελάχιστες μέθοδοι τέτοιου είδους ελέγχων. Ο σημαντικότερος από αυτούς τους ελέγχους είναι ο διωνυμικός έλεγχος και η κανονική προσέγγισή του όταν το δείγμα είναι μεγάλο. Υπάρχει ακόμα ο X^2 έλεγχος καλής προσαρμογής που έχει αποδειχτεί ένας πολύ καλός τρόπος υπολογισμού της ακρίβειας των προβλέψεων για τη λογιστική παλινδρόμηση. Σε όλους αυτούς τους ελέγχους οι υποψήφιοι που βαθμολογούνται χωρίζονται σε τμήματα ανάλογα με τη βαθμολογία ή την πιθανότητα αθέτησης υποχρεώσεων. Επιπλέον, σε όλους αυτούς τους ελέγχους υποθέτουμε ότι η πιθανότητα ένας πελάτης να αθετήσει τις υποχρεώσεις του δεν εξαρτάται από το αν ένας άλλος πελάτης που βρίσκεται στην ίδια κατάσταση θα αθετήσει τις υποχρεώσεις του ή όχι.

α. Διωνυμικός έλεγχος

Ο διωνυμικός έλεγχος (*binomial test*) είναι ένας τρόπος για να ελεγχθεί η εγκυρότητα της πιθανότητας αθέτησης υποχρεώσεων *PD* για κάθε κλάση βαθμολογιών. Υποθέτουμε ότι η πιθανότητα ένας πελάτης να αθετήσει τις υποχρεώσεις του δεν εξαρτάται από την πιθανότητα ένας άλλος πελάτης που ανήκει στην ίδια κλάση να αθετήσει τις υποχρεώσεις του. Σύμφωνα με αυτήν την υπόθεση, ο αριθμός των «κακών» πελατών στην κλάση i που περιλαμβάνει n_i οφειλέτες ακολουθεί διωνυμική κατανομή. Η πιθανότητα ένας πελάτης να είναι «κακός» είναι $PD_i = 1 - p_i$, όπου p_i είναι η πιθανότητα ένας πελάτης να είναι «καλός». Επομένως, σύμφωνα με τον Tasche (2005) ο διωνυμικός έλεγχος συγκρίνει τις δύο υποθέσεις

H_0: Η PD_i της κλάσης των βαθμολογιών i είναι σωστή $(PD_i = PD_i)$.

H_1: Η PD_i της κλάσης των βαθμολογιών i δεν είναι σωστή, είναι υποτιμημένη $(PD_i > PD_i)$.

Εάν ισχύει η μηδενική υπόθεση H_0, η πιθανότητα ότι υπάρχουν b_i «κακοί» στους n_i οφειλέτες είναι ίση με $\binom{n_i}{b_i}(PD)^{b_i}(1-PD)^{n_i-b_i}$. Επομένως, η μηδενική υπόθεση απορρίπτεται

σε επίπεδο σημαντικότητας α εάν για τον αριθμό των «κακών» k σε αυτήν την κλάση ισχύει $k \geq k^*$, όπου k^* είναι το κρίσιμο σημείο του ελέγχου και υπολογίζεται από τον ακόλουθο τύπο

$$k^* = \min\left\{k \middle| P(x \geq k) \leq 1-\alpha\right\} = \min\left\{k \middle| \sum_{x=k}^{n_i} \binom{n_i}{x} PD^x (1-PD)^{n_i-x} \leq 1-\alpha\right\} =$$

$$= \min\left\{k \middle| \sum_{x=k}^{n_i} \binom{n_i}{x} (1-p_i)^x (p_i)^{n_i-x} \leq 1-\alpha\right\}.$$

- ***Κανονική προσέγγιση του διωνυμικού ελέγχου***

Εάν ο αριθμός των πελατών που ανήκουν σε μια κλάση βαθμολογιών του δείγματος ελέγχου είναι μεγάλος, τότε οι υπολογισμοί της εκτίμησης των διωνυμικών πιθανοτήτων γίνονται πολύπλοκες. Όμως, λόγω του μεγάλου μεγέθους δείγματος η κατανομή του αριθμού των πελατών μπορεί να είναι ασυμπτωτικά κανονική λόγω του Κεντρικού Οριακού Θεωρήματος. Εάν υπάρχουν n_i πελάτες στην i κλάση και ισχύει η μηδενική υπόθεση H_0 ότι η PD_i της κλάσης των βαθμολογιών i είναι σωστή, τότε ο αναμενόμενος αριθμός των πελατών που θα αθετήσουν τις υποχρεώσεις τους είναι $n_i PD_i$ με διακύμανση $n_i PD_i(1-PD_i)$. Ο αριθμός των πελατών που πρόκειται να αθετήσουν τις υποχρεώσεις τους ακολουθεί την κανονική κατανομή $N(n_i PD_i, n_i PD_i(1-PD_i))$. Επομένως, το κρίσιμο σημείο k^*, πάνω από το οποίο απορρίπτεται η μηδενική υπόθεση σε ένα επίπεδο σημαντικότητας α είναι

$$k^* = \Phi^{-1}(\alpha)\sqrt{n_i PD_i(1-PD_i)} + n_i PD_i = \Phi^{-1}(\alpha)\sqrt{n_i p_i(1-p_i)} + n_i(1-p_i)$$

όπου $\Phi^{-1}(\alpha)$ είναι η αντίστροφη της αθροιστικής συνάρτησης κατανομής της τυπικής κανονικής κατανομής.

β. Έλεγχος $\mathbf{X^2}$

Ένας άλλος τρόπος επικύρωσης της προσαρμογής των μοντέλων είναι ό έλεγχος X^2 καλής προσαρμογής ή αλλιώς έλεγχος Hosmer-Lemeshow (Hosmer-Lemeshow (1980)). Ο έλεγχος X^2 είναι ένας γενικός έλεγχος που εκτιμά πόσο καλά προσαρμόζεται ένα μοντέλο στα δεδομένα, συγκρίνοντας τα πραγματικά αποτελέσματα με αυτά που προέβλεψε το

μοντέλο με βάση το δείγμα ελέγχου. Αυτό που απαιτείται για να εξεταστεί η προβλεπτική ικανότητα ενός CSM είναι να υπολογιστεί το άθροισμα των σφαλμάτων πρόβλεψης σταθμισμένο με την αντίστροφη διακύμανση. Η ιδέα είναι ακριβώς η ίδια όπως στην περίπτωση της αρχικής ταξινόμησης με μόνη διαφορά ότι σε εκείνη την περίπτωση η μηδενική υπόθεση ήταν ότι η σχετική πιθανότητα σε κάθε κλάση είναι ίδια με τη σχετική πιθανότητα του πληθυσμού. Σε αυτήν την περίπτωση η μηδενική υπόθεση είναι ότι οι πιθανότητες αθέτησης υποχρεώσεων σε κάθε κλάση i, $i=1,2,...,N$ έχουν συγκεκριμένες τιμές PD_i.

Οι βαθμολογίες που προκύπτουν για το κάθε σκορόχαρτο χωρίζονται σε N κλάσεις, όπου στην i κλάση, $i=1,2,...,N$ το ποσοστό των «καλών» εκτιμάται ότι είναι p_i και το ποσοστό των «κακών» είναι $1-p_i=PD_i$. Εάν στην κλάση i υπάρχουν συνολικά n_i πελάτες από τους οποίους οι g_i από αυτούς είναι «καλοί» και οι $b_i=n_i-g_i$ είναι «κακοί», τότε το στατιστικό X^2 είναι το άθροισμα των τετραγώνων των διαφορών μεταξύ των αναμενόμενων και των παρατηρούμενων αριθμών των «καλών» (ή «κακών») διαιρεμένο με τη διακύμανση. Στην κλάση i, ο αναμενόμενος αριθμός των καλών» είναι $n_i p_i$ και η διακύμανση του αριθμού των «καλών» ακολουθεί διωνυμική κατανομή είναι $n_i p_i(1-p_i)$. Επομένως, το στατιστικό X^2 είναι ίσο με

$$HL=\sum_{i=1}^{N}\frac{(n_i p_i-g_i)^2}{n_i p_i(1-p_i)}. \tag{4.8}$$

Συνήθως τα δεδομένα αναπαριστώνται σε ένα **πίνακα συνάφειας (*contingency table*)** με N γραμμές και 2 στήλες. Κάθε μία από τις γραμμές του πίνακα συνάφειας αντιστοιχεί στις κλάσεις των βαθμολογιών και κάθε μία από τις στήλες αντιστοιχούν στα 2 αποτελέσματα «καλοί» και «κακοί». Σε κάθε κελί του πίνακα αντιστοιχούν οι παρατηρούμενες συχνότητες του δείγματος που ανήκουν στην αντίστοιχη κατηγορία. Σε μερικούς πίνακες συνάφειας στα κελιά μπορεί να εμφανίζονται μέσα σε παρένθεση και οι αντίστοιχες αναμενόμενες συχνότητες.

Ο έλεγχος X^2 υπολογίζει για κάθε κελί του πίνακα το τετράγωνο της διαφοράς μεταξύ των παρατηρούμενων και των αναμενόμενων τιμών σε αυτό το κελί διαιρεμένο με τις αναμενόμενες τιμές. Με αυτόν τον τρόπο προκύπτει ξανά η σχέση (4.8) ως εξής:

$$HL = \sum_{i=1}^{N} \frac{(\text{Αναμενόμενος αριθ. καλών στην } i - \text{Παρατηρούμενος αριθ. καλών στην } i)^2}{\text{Παρατηρούμενος αριθ. καλών στην } i} +$$

$$+ \sum_{i=1}^{N} \frac{(\text{Αναμενόμενος αριθ. κακών στην } i - \text{Παρατηρούμενος αριθ. κακών στην } i)^2}{\text{Παρατηρούμενος αριθ. κακών στην } i} =$$

$$= \sum_{i=1}^{N} \left(\frac{(n_i p_i - g_i)^2}{n_i p_i} + \frac{(n_i(1-p_i) - b_i)^2}{n_i(1-p_i)} \right) = \sum_{i=1}^{N} \left(\frac{(n_i p_i - g_i)^2}{n_i p_i} + \frac{(g_i - n_i p_i)^2}{n_i(1-p_i)} \right) =$$

$$= \sum_{i=1}^{N} (n_i p_i - g_i)^2 \left(\frac{1}{n_i p_i} + \frac{1}{n_i(1-p_i)} \right) = \sum_{i=1}^{N} \frac{(n_i p_i - g_i)^2}{n_i p_i (1-p_i)}.$$

Το στατιστικό X^2 ακολουθεί κατανομή χ^2 με $N-2$ βαθμούς ελευθερίας. Απαραίτητη προϋπόθεση για να εφαρμοστεί ο έλεγχος X^2 είναι σε κάθε κελί του πίνακα συνάφειας να περιέχεται τουλάχιστον το 5% των παρατηρήσεων. Επομένως, η σωστή επιλογή του αριθμού N των κλάσεων και ο τρόπος με τον οποίο αυτές επιλέγονται παίζει πολύ σημαντικό ρόλο.

Ο τρόπος με τον οποίο αποδίδεται η πιθανότητα αθέτησης υποχρεώσεων σε μια κλάση βαθμολογιών εξαρτάται κυρίως από το είδος του CSM που κατασκευάστηκε. Για παράδειγμα, αν η σχέση μεταξύ της βαθμολογίας s και της πιθανότητας p (πιθανότητα ο πελάτης να είναι «καλός») είναι $s = \ln\left(\frac{p(s)}{1-p(s)}\right)$ ή $p(s) = \frac{1}{1+e^{-s}}$ τότε η πιθανότητα αθέτησης υποχρεώσεων είναι $PD(s) = \frac{1}{1+e^{s}}$. Επομένως η πιθανότητα αθέτησης υποχρεώσεων που θα αποδοθεί σε κάθε κλάση βαθμολογιών θα μπορούσε να είναι ο μέσος όρος των πιθανοτήτων $PD(s)$ για κάθε βαθμολογία που ανήκει σε αυτήν την κλάση. Αυτός ο τρόπος όμως απαιτεί πολλούς υπολογισμούς και πρέπει να επαναλαμβάνονται όταν επιλέγονται διαφορετικές κλάσεις.

Εναλλακτικά, για μια κλάση βαθμολογιών $[s_1, s_2]$ η πιθανότητα ο πελάτης να είναι «καλός» επιλέγεται να είναι αυτή για που αντιστοιχεί στο ενδιάμεσο σημείο των βαθμολογιών (*midpoint*) της κλάσης, δηλαδή $p = \frac{1}{1+e^{-(s_1+s_2)/2}}$ ή να είναι το ενδιάμεσο σημείο των πιθανοτήτων στην κλάση βαθμολογιών δηλαδή

$$p = \frac{p_1 + p_2}{2} = \frac{\left(\frac{1}{1+e^{-s_1}} + \frac{1}{1+e^{-s_2}}\right)}{2}.$$

Αν οι κλάσεις έχουν επιλεγεί αρχικά με βάση τις πιθανότητες, η δεύτερη προσέγγιση είναι καταλληλότερη.

Σύμφωνα με το νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ, σε κάθε κλάση λαμβάνεται ως αντιπροσωπευτική πιθανότητα το μικρότερο ποσοστό των «καλών» (ή το μεγαλύτερο ποσοστό των «κακών») για τις βαθμολογίες που ανήκουν σε αυτήν την κλάση. Αυτό σημαίνει ότι τα μέτρα προσαρμογής μπορεί να υπερεκτιμούν τις πιθανότητες αθέτησης υποχρεώσεων για κάθε κλάση, αλλά αυτό δεν αποτελεί πρόβλημα όταν χρησιμοποιείται ο διωνυμικός έλεγχος διότι η εναλλακτική υπόθεση είναι ότι η πιθανότητα αθέτησης υποχρεώσεων είναι υποτιμημένη.

4.5 Τεχνικές αξιολόγησης μοντέλων σε περιπτώσεις μικρού μεγέθους δείγματος

Σε μερικές περιπτώσεις κατά τη διαδικασία βαθμολόγησης πιστοληπτικής ικανότητας δεν είναι δυνατό να προκύψει ξεχωριστό δείγμα ελέγχου από αυτό του δείγματος ανάπτυξης λόγω του περιορισμένου μεγέθους δείγματος, όπως για παράδειγμα όταν έχουμε ένα νέο προϊόν δανεισμού. Για να αντιμετωπιστεί αυτό το πρόβλημα και για να προκύψουν αμερόληπτες εκτιμήσεις χωρίς απώλεια πληροφορίας μπορούν να χρησιμοποιηθούν διάφορες στατιστικές τεχνικές όπως η **διεπικύρωση (*cross validation*)** και η **μέθοδος *bootstrap*.**

α. Μέθοδος Διεπικύρωσης

Σύμφωνα με τη μέθοδο διεπικύρωσης κατασκευάζεται ένα CSM μοντέλο βασισμένο σε ένα υποσύνολο του συνολικού δείγματος D. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να καλυφθεί ολόκληρο το σύνολο D. Εάν για να καλυφτεί το σύνολο D χρειαστούν N υποσύνολα τότε δημιουργείται μια σειρά από N σκορόχαρτα που για καθένα από αυτά λαμβάνεται μια αμερόληπτη εκτίμηση ενός μέτρου απόδοσης του. Έπειτα, υπολογίζεται ο μέσος όρος του μέτρου απόδοσης για τα διαφορετικά υποσύνολα.

Υπάρχουν δύο διαφορετικοί τρόποι για να εφαρμοστεί η μέθοδος διεπικύρωσης. Ο πρώτος τρόπος είναι η κυκλική μέθοδος (*rotation method*). Σύμφωνα με αυτή τη μέθοδο το σύνολο

του δείγματος D χωρίζεται σε N διαφορετικά υποσύνολα $D_1, D_2, ..., D_N$ που δεν έχουν κανένα κοινό σημείο μεταξύ τους και δομούνται N διαφορετικά μοντέλα. Κάθε ένα από το μοντέλα κατασκευάζεται βασισμένο στο ποσοστό $(N-1)/N$ του συνολικού δείγματος και το δείγμα ελέγχου κάθε φορά είναι το υπόλοιπο $1/N$. Σε κάθε περίπτωση, εξετάζεται το μέτρο απόδοσης στα δεδομένα που δεν χρησιμοποιήθηκαν για να δομηθεί το μοντέλο. Μια αμερόληπτη εκτίμηση του μέτρου απόδοσης είναι ο μέσος όρος των μέτρων απόδοσης για κάθε ένα από τα N σκορόχαρτα. Βέβαια, η μεροληψία μειώνεται όσο αυξάνεται το N, αλλά η διακύμανση είναι αντιστρόφως ανάλογη του N.

Ένας δεύτερος τρόπος για να εφαρμοστεί η μέθοδος διεπικύρωσης είναι η προσέγγιση του μονοαποκλεισμού (*leave-one-out*), όπου για κάθε ένα στοιχείο του συνόλου $d \in D$ εξετάζεται η διάσπαση $\{d, D-\{d\}\}$ και σύμφωνα με αυτή κάθε φορά κατασκευάζεται ένα μοντέλο με δείγμα ανάπτυξης το σύνολο $D-\{d\}$ και μονοσύνολο ελέγχου το $\{d\}$. Εάν το σύνολο D αποτελείται από N στοιχεία, τότε θα χρειαστεί να κατασκευαστούν N διαφορετικά σκορόχαρτα για καθένα από τα οποία εκτιμάται και το μέτρο απόδοσής του. Επισημαίνεται ότι, και σε αυτήν την περίπτωση, μια αμερόληπτη εκτίμηση του μέτρου απόδοσης είναι ο μέσος όρος των μέτρων απόδοσης για κάθε ένα από τα N μοντέλα.

Αν και η μέθοδος μονοαποκλεισμού απαιτεί τη δημιουργία περισσότερων μοντέλων σε σχέση με την κυκλική μέθοδο, το δείγμα ανάπτυξης για το κάθε μοντέλο έχει πιο πολλά κοινά στοιχεία με το συνολικό δείγμα και επομένως η εκτίμηση του μέτρου απόδοσης κάθε μοντέλου είναι πιο ανθεκτική.

β. ***Μέθοδος bootstrap***

Σύμφωνα με τη μέθοδο στηρίγματος, εάν D είναι ένα σύνολο δεδομένων που αποτελείται από $|D|$ πελάτες, τότε μπορούν να κατασκευαστούν νέα σύνολα δεδομένων B λαμβάνοντας τυχαία δείγματα από το D με επανατοποθέτηση $|D|$ φορές έτσι ώστε ο ίδιος πελάτης να μπορεί να επιλεχτεί αρκετές φορές. Αυτή η δειγματοληψία με επανατοποθέτηση επαναλαμβάνεται N φορές μέχρι να δημιουργηθούν τα σύνολα $B_1, B_2, ..., B_N$. Έπειτα χρησιμοποιείται το δείγμα B_1 για να κατασκευαστεί ένα σκορόχαρτο και ελέγχεται με το B_1^c που αποτελείται από τα συμπληρωματικά στοιχεία του B_1. Αυτή η διαδικασία

επαναλαμβάνεται και για τα υπόλοιπα δείγματα $B_2, B_3, ..., B_N$. Συνολικά λοιπόν κατασκευάζονται N διαφορετικά σκορόχαρτα για το καθένα από αυτά εκτιμάται το μέτρο απόδοσης του κάθε μοντέλου. Αν $M_{B_i|B_i}$, $i = 1, 2, ..., N$ είναι το μέτρο απόδοσης για κάθε CSM με δείγμα ανάπτυξης και δείγμα επικύρωσης το B_i και $M_{B_i|B_i^c}$ είναι το μέτρο απόδοσης για κάθε CSM με δείγμα ανάπτυξης B_i και δείγμα επικύρωσης το B_i^c τότε ο μέσος όρος των σφαλμάτων $M_{B_i|B_i} - M_{B_i|B_i^c}$ είναι μια καλή εκτίμηση του σφάλματος $M_{D|D} - M_{D|V}$, όπου $M_{D|D}$ είναι το μέτρο απόδοσης του μοντέλου που έχει δομηθεί με δείγμα ανάπτυξης και δείγμα ελέγχου D και $M_{D|V}$ είναι το μέτρο απόδοσης του μοντέλου που έχει δομηθεί με δείγμα ανάπτυξης D και κάποιο δείγμα επικύρωσης V για το οποίο ισχύει $D \cap V = \{\varnothing\}$. Επομένως, ισχύει ότι

$$M_{D|V} = M_{D|D} - \left(M_{D|D} - M_{D|V}\right) \approx M_{D|D} - \left(\frac{\sum_{i=1}^{N} (M_{B_i|B_i} - M_{B_i|B_i^c})}{N} \right). \qquad (4.9)$$

Το σύνολο δεδομένων D χωρίζεται στους πελάτες που ανήκουν στο σύνολό B και σε αυτούς που δεν ανήκουν. Η πιθανότητα ένας πελάτης να μην ανήκει στο σύνολο B είναι $\left(1 - \frac{1}{|D|}\right)^{|D|}$. Όμως για $|D| \to \infty$ ισχύει

$$\lim_{|D|\to\infty} \left(1 - \frac{1}{|D|}\right)^{|D|} = e^{-1} = 0,368.$$

Επομένως, υπάρχει 0,368 πιθανότητα ο συγκεκριμένος πελάτης να μην ανήκει στο σύνολο B_i, όπου $i = 1, 2, ..., N$ και $1 - 0,368 = 0,632$ πιθανότητα να ανήκει στο σύνολο αυτό. Αυτό σημαίνει ότι το σύνολο B_i προσφέρει λιγότερη πληροφορία σε σχέση με το D διότι περιέχει μόνο 63% των στοιχείων του D. Υποθέτοντας ότι η ποσότητα $M_{B_i|B_i^c}$ είναι μια καλή εκτίμηση του $M_{D|V}$ και το μοντέλο που δομήθηκε με το δείγμα ελέγχου B_i^c περιέχει μόνο το 63% της πληροφορίας που χρησιμοποιείται για να δομηθεί το μοντέλο με το πραγματικό δείγμα επικύρωσης τότε μια καλύτερη προσέγγιση είναι

$$M_{D|V} = \sum_{i=1}^{N} (0,632 M_{B_i|B_i^c} + 0,368 M_{B_i|B_i})$$

Αυτή η προσέγγιση δίνει πιο σταθερά αποτελέσματα σε σχέση με την προσέγγιση bootsrap που δίνεται από τη σχέση (4.9). Όμως για την πλειοψηφία των CSM τα δείγματα είναι αρκετά μεγάλα και μπορεί να υπολογιστεί απευθείας η απόδοση M_V του μοντέλου.

4.6 Ανακεφαλαίωση

Λόγω των μεγάλων συνόλων δεδομένων που είναι διαθέσιμα, η επικύρωση των CSM και η μέτρηση της απόδοσής του μπορεί να βασίζεται συνήθως σε ένα δείγμα ελέγχου που αποτελεί μέρος του συνολικού δείγματος. Στην σπάνια περίπτωση που δεν είναι διαθέσιμο μεγάλο μέγεθος δείγματος και δεν μπορεί να υπάρξει ξεχωριστό δείγμα ελέγχου, χρησιμοποιούνται μέθοδοι όπως είναι η μέθοδος διεπικύρωσης και η μέθοδος bootstrap. Υπάρχουν τρεις βασικές κατηγορίες μεθόδων επικύρωσης, η μέτρηση της ακρίβειας ταξινόμησης των πελατών, η χρήση διάφορων μέτρων που αξιολογούν τη διαχωριστική ικανότητα του μοντέλου και η μέτρηση της ακρίβειας προσαρμογής της εκτίμησης της πιθανότητας αθέτησης υποχρεώσεων των υποψηφίων.

Η μέτρηση της ακρίβειας ταξινόμησης πελατών βασίζεται σε έναν 2×2 πίνακα, τον πίνακα συγχύσεως όπου τα μη διαγώνια στοιχεία του εκφράζουν τα ποσοστά λανθασμένης ταξινόμησης. Δεδομένου ότι σε κάθε υποψήφιο αντιστοιχεί μια βαθμολογία που έχει προκύψει από τις μεθόδους που περιγράφηκαν στο προηγούμενο κεφάλαιο, τα μέτρα διαχωριστικής ικανότητας που χρησιμοποιούνται είναι η απόκλιση, η τιμή πληροφορίας, η απόσταση Mahalanobis και το στατιστικό Kolmogorov – Smirnov. Όμως ο πιο συνηθισμένος τρόπος της μέτρησης της απόδοσης ενός CSM είναι η καμπύλη ROC στην οποία απεικονίζονται τα ποσοστά της «σωστής απόρριψης ενός κακού πελάτη» (ευαισθησία) έναντι των ποσοστών της λανθασμένης απόρριψης ενός καλού πελάτη» (1 – ειδικότητα) για ένα εύρος διαφορετικών βαθμολογιών. Ένα σημαντικό μέτρο διαχωρισμού που προκύπτει από αυτήν την καμπύλη είναι το εμβαδό της περιοχής μεταξύ της καμπύλης και των αξόνων ή ένας μετασχηματισμός αυτού που αποτελεί το συντελεστή Gini. Η καμπύλη ROC είναι ισοδύναμη με την καμπύλη CAP στην οποία απεικονίζεται η αθροιστική συνάρτηση κατανομής της βαθμολογίας των «κακών» $F(s|B)$, έναντι της αθροιστικής συνάρτησης κατανομής $F(s)$ για κάθε βαθμολογία s. Ένα σημαντικό μέτρο διαχωρισμού που προκύπτει

από την καμπύλη CAP είναι ο δείκτης ακρίβειας AR που ισοδυναμεί με το συντελεστή Gini. Αυτά τα μέτρα παρέχουν μια σφαιρική αντίληψη της απόδοσης ενός CSM, που ενσωματώνεται για όλες τις πιθανές επιλογές της βαθμολογίας αποδοχής-απόρριψης. Σύμφωνα λοιπόν με αυτό το γεγονός μπορούμε να ορίσουμε ως βαθμολογία αποδοχής-απόρριψης τη βαθμολογία αυτή που μεγιστοποιεί την απόδοση του μοντέλου.

Τα μέτρα ακρίβειας προσαρμογής της εκτίμησης της πιθανότητας υπολογίζουν πόσο καλές είναι οι προβλέψεις των πιθανοτήτων αθέτησης υποχρεώσεων αφού πρώτα χωριστούν σε κλάσεις οι βαθμολογίες. Τα πιο συνήθη μέτρα ακρίβειας προσαρμογής είναι ο διωνυμικός έλεγχος και ο X^2 έλεγχος καλής προσαρμογής.

ΚΕΦΑΛΑΙΟ 5

Σύγκριση των μεθόδων ανάπτυξης μοντέλων βαθμολόγησης

5.1 Πλεονεκτήματα και μειονεκτήματα των μεθόδων

Έχοντας περιγράψει τις πιο βασικές μεθόδους ανάπτυξης μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας προκύπτει το ερώτημα «ποια από αυτές τις μεθόδους είναι καλύτερη;». Η απάντηση σε αυτό το ερώτημα δεν είναι καθόλου εύκολη γιατί κάθε πιστωτικός οργανισμός ανάλογα με τον τελικό του στόχο και τα στοιχεία που διαθέτει χρησιμοποιεί διαφορετική μεθοδολογία ως καλύτερη.

Η μέθοδος που χρησιμοποιείται συχνότερα για την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας είναι η λογιστική παλινδρόμηση. Παρ' όλα αυτά, και οι υπόλοιπες μέθοδοι όπως η γραμμική παλινδρόμηση, η διαχωριστική ανάλυση και τα δέντρα ταξινόμησης χρησιμοποιούνται συχνά στην πράξη. Όσον αφορά τη μέθοδο του κοντινότερου γείτονα, παρότι έχουν γίνει πολλές μελέτες πάνω σε αυτή δε συνηθίζεται να χρησιμοποιείται στην πράξη.

Στη συνέχεια παρουσιάζονται κάποια πλεονεκτήματα και μειονεκτήματα για κάθε μέθοδο ξεχωριστά.

Διαχωριστική Ανάλυση

Η μέθοδος της διαχωριστικής ανάλυσης είναι αρκετά καλή όταν ικανοποιούνται κάποιες προϋποθέσεις. Οι βασικές προϋποθέσεις που πρέπει να ικανοποιούνται είναι τα δεδομένα να προέρχονται από κανονική κατανομή, οι πίνακες συνδιακυμάνσεων και για τις δύο ομάδες να

είναι ίσοι ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$) και επιπλέον να είναι γνωστές οι εκ των προτέρων πιθανότητες που είναι η αναλογία των «καλών» και των «κακών» αντίστοιχα μέσα στον πληθυσμό.

Επομένως, το κύριο μειονέκτημα της γραμμικής διαχωριστικής ανάλυσης είναι ότι για να είναι σωστή η εφαρμογή της θα πρέπει να ισχύουν όλες οι προηγούμενες προϋποθέσεις. Όμως, δυστυχώς τις περισσότερες φορές δεν είναι δυνατό να ισχύουν αυτές οι προϋποθέσεις. Αν, ελέγχοντας την υπόθεση της ισότητας των πινάκων συνδιακύμανσης προκύπτει ότι αυτή δεν ισχύει, τότε η συνάρτηση διαχωρισμού δεν είναι γραμμική αλλά τετραγωνική (*Quadratic Discriminant Analysis*).

Το βασικότερο πλεονέκτημα είναι ότι, σε περίπτωση που ισχύουν όλες οι προϋποθέσεις για την εφαρμογή της ΔΑ (πράγμα το οποίο είναι εξαιρετικά δύσκολο) τότε τα αποτελέσματα της είναι πολύ καλύτερα σε σχέση με όλες τις υπόλοιπες μεθόδους ανάπτυξης CSM. Από τα σημαντικότερα πλεονεκτήματα της γραμμικής διαχωριστικής ανάλυσης είναι ότι η μέθοδος αυτή είναι αρκετά απλή και κατανοητή στην εφαρμογή της. Επίσης, αυτή η μέθοδος λαμβάνει υπ' όψιν την αλληλεπίδραση των χαρακτηριστικών των πελατών και μετασχηματίζει τις τιμές των ανεξάρτητων μεταβλητών σε ένα μόνο Z-score. Για τη γραμμική διαχωριστική ανάλυση που εισήγαγε για πρώτη φορά ο Fisher δεν χρειάζεται να γίνει καμιά υπόθεση για την κατανομή των δεδομένων και η μόνη υπόθεση που λαμβάνεται υπ' όψιν είναι η ισότητα των πινάκων συνδιακυμάνσεων των δύο ομάδων.

Γραμμική Παλινδρόμηση

Όπως αναφέρθηκε, το μοντέλο της γραμμικής παλινδρόμησης έχει άμεση σχέση με τη γραμμική διαχωριστική συνάρτηση του Fisher. Επομένως, τα πλεονεκτήματα αυτής σχετίζονται και με αυτά της γραμμικής διαχωριστικής ανάλυσης.

Η μέθοδος της γραμμικής παλινδρόμησης παρουσιάζει και αυτή με τη σειρά της κάποια μειονεκτήματα. Ένα βασικό μειονέκτημα της μεθόδου αυτής είναι ότι, τα σφάλματα εξαρτώνται από την τιμή της μεταβλητής απόκρισης (φαινόμενο ετεροσκεδαστικότητας) και εάν η κατανομή των καταλοίπων δεν είναι κανονική τότε η μέθοδος των ελαχίστων τετραγώνων δεν είναι αποτελεσματική. Όμως, το σημαντικότερο μειονέκτημα είναι ότι η δεσμευμένη πιθανότητα να συμβεί το επιθυμητό γεγονός δεδομένης μιας τιμής του $\mathbf{x}$ μπορεί πρακτικά να βγει εκτός του διαστήματος (0,1) εντός του οποίου πρέπει να βρίσκεται η τιμή μιας πιθανότητας.

Λογιστική Παλινδρόμηση

Θεωρητικά η λογιστική παλινδρόμηση αποτελεί τη βέλτιστη μέθοδο ταξινόμησης για μια ευρύτερη κατηγορία κατανομών σε σχέση με τη γραμμική παλινδρόμηση. Παρ' όλα αυτά όταν γίνονται συγκρίσεις των δύο διαφορετικών μεθόδων πάνω στα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας που έχουν ήδη αναπτυχθεί στο ίδιο σύνολο στοιχείων, δεν υπάρχει μεγάλη διαφορά στην τελική τους ταξινόμηση. Η διαφορά είναι ότι η γραμμική παλινδρόμηση προσπαθεί να εκφράσει την πιθανότητα αθέτησης υποχρεώσεων p με ένα γραμμικό συνδυασμό των επεξηγηματικών μεταβλητών, ενώ η λογιστική παλινδρόμηση αποσκοπεί στο να εκφράσει το λογάριθμο $\log\left(\frac{p_i}{1-p_i}\right)$ με ένα γραμμικό συνδυασμό των επεξηγηματικών μεταβλητών.

Γενικά όσον αφορά τα CSM τα αποτελέσματα που παράγονται με τις δύο μεθόδους είναι παρόμοια εκτός από εκείνες τις περιπτώσεις όπου η πιθανότητα αθέτησης των υποχρεώσεων είναι πολύ μικρή ή πολύ μεγάλη. Αυτοί είναι οι υποψήφιοι πελάτες για τους οποίους είναι εύκολο να προβλεφθεί εάν θα αθετήσουν ή όχι τις υποχρεώσεις τους. Για τις πιο δύσκολες περιοχές πρόβλεψης της αθέτησης υποχρεώσεων, δηλαδή όταν $p = 0{,}5$ περίπου, τα αποτελέσματα και των δύο μεθόδων είναι σχεδόν ίδια.

Η λογιστική παλινδρόμηση μπορεί να θεωρηθεί ειδική περίπτωση της γραμμικής παλινδρόμησης. Ωστόσο, η δυαδική μεταβλητή απόκρισης παραβιάζει την υπόθεση της κανονικότητας των μοντέλων γραμμικής παλινδρόμησης. Ένα μοντέλο λογιστικής παλινδρόμησης ουσιαστικά εκφράζει ότι μια κατάλληλη συνάρτηση της πιθανότητας ένας πελάτης να είναι «καλός», είναι μια γραμμική συνάρτηση των παρατηρούμενων τιμών των διαθέσιμων ερμηνευτικών μεταβλητών. Τα σημαντικότερα πλεονεκτήματα αυτής της προσέγγισης είναι ότι μπορεί να παραγάγει έναν απλό πιθανολογικό τύπο της ταξινόμησης και ότι το μοντέλο της λογιστικής παλινδρόμησης σε σχέση με τη γραμμική είναι η πρόβλεψη για τη μεταβλητή απόκρισης δεν ξεφεύγει από το διάστημα (0,1).

Οι κύριες αδυναμίες του μοντέλου της λογιστικής παλινδρόμησης είναι ότι δεν μπορεί να αντιμετωπίσει επαρκώς τα προβλήματα των μη γραμμικών αλληλεπιδράσεων των ερμηνευτικών μεταβλητών.

Γενικά, τα μοντέλα logit και probit παράγουν όμοια αποτελέσματα με εξαίρεση την περίπτωση των πολύ μεγάλων δειγμάτων που παρουσιάζουν έκτροπες παρατηρήσεις.

Δέντρα Ταξινόμησης

Οι μέθοδοι που δημιουργούν ομάδες όπως είναι τα δέντρα ταξινόμησης έχουν το πλεονέκτημα ότι εξετάζουν αυτόματα τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, ενώ για τις γραμμικές μεθόδους, αυτές οι αλληλεπιδράσεις πρέπει να προσδιοριστούν εκ των προτέρων και να καθοριστούν τα κατάλληλα χαρακτηριστικά. Παρακάτω, ο Πίνακας 5.1 περιγράφει τα ποσοστά των «καλών» πελατών για διαφορετικές κατηγορίες κατάστασης κατοικίας και κατοχής ή όχι σταθερού τηλεφώνου.

Πίνακας 5.1 *Ποσοστό των «καλών» για τα χαρακτηριστικά «κατάσταση κατοικίας» και «κάτοχος σταθερού τηλεφώνου» (πηγή: Thomas et al. (2002)).*

Κάτοχος Τηλεφώνου / **Κατάσταση Κατοικίας**	**Ναι**	**Όχι**	
Ιδιοκτήτης	95%	50%	90%
Ενοικιαστής	75%	65%	70%
	91%	60%	

Από τον παραπάνω πίνακα παρατηρείται ότι το ποσοστό των «καλών» πελατών είναι μεγαλύτερο για τους ιδιοκτήτες (90%) σε σχέση με τους ενοικιαστές (70%) και ανάμεσα στους κατόχους σταθερού τηλεφώνου οι «καλοί» είναι περισσότεροι (91%) σε σχέση με αυτούς που δεν έχουν σταθερό τηλέφωνο (60%). Παρ' όλα αυτά, σε ένα γραμμικό σύστημα τη μεγαλύτερη βαθμολογία θα την έχουν οι ιδιοκτήτες που διαθέτουν σταθερό τηλέφωνο και τη μικρότερη βαθμολογία θα την έχουν οι ενοικιαστές που δε διαθέτουν σταθερό τηλέφωνο. Στην πραγματικότητα όμως και σύμφωνα με τον Πίνακα 5.1, οι ιδιοκτήτες που δεν έχουν σταθερό τηλέφωνο θα πρέπει να είναι χειρότεροι σε βαθμολογία αλλά ένα γραμμικό σύστημα δεν μπορεί να το δείξει αυτό. Τα δέντρα ταξινόμησης τείνουν να εντοπίζουν και να λαμβάνουν υπ' όψιν τέτοιες αλληλεπιδράσεις. Επίσης, τα δέντρα ταξινόμησης χρησιμοποιούνται για να προσδιοριστούν σημαντικές αλληλεπιδράσεις και όταν αυτές προσδιοριστούν γίνεται ομαδοποίηση με βάση το χαρακτηριστικό που εμφάνισε αυτές τις αλληλεπιδράσεις. Στο συγκεκριμένο παράδειγμα, μπορούν να χρησιμοποιηθούν διαφορετικές γραμμικές μέθοδοι για ιδιοκτήτες και ενοικιαστές. Επομένως, το βασικό πλεονέκτημα των δέντρων ταξινόμησης είναι ότι συνήθως δημιουργούνται από απλούς και κατανοητούς

κανόνες ταξινόμησης και μπορούν να χειριστούν τις μη γραμμικές αλληλεπιδράσεις των ερμηνευτικών μεταβλητών.

Τα δέντρα ταξινόμησης εφαρμόζονται στην περίπτωση όπου η μεταβλητή απόκρισης είναι διακριτή και εκτελούν ταξινόμηση των παρατηρήσεων με βάση όλες της επεξηγηματικές μεταβλητές που είναι διαθέσιμες και εποπτεύεται από την μεταβλητή απόκρισης. Η διαδικασία της ομαδοποίησης τυπικά διεξάγεται χρησιμοποιώντας μόνο μια επεξηγηματική μεταβλητή κάθε φορά. Τα δέντρα ταξινόμησης βασίζονται στην ελαχιστοποίηση της μη αγνότητας, η οποία αναφέρεται και ως μέτρο μεταβλητότητας των τιμών των παρατηρήσεων. Το βασικότερο μειονέκτημα των δέντρων ταξινόμησης είναι, σύμφωνα με τον Che-hui Lien (2009), ότι η διαδοχική φύση και η αλγοριθμική πολυπλοκότητα αυτής της προσέγγισης την κάνει να εξαρτάται από τις παρατηρηθείσες τιμές και ακόμα και μια πολύ μικρή αλλαγή μπορεί να αλλάξει ολοκληρωτικά τη δομή του δέντρου. Επομένως, καθίσταται δύσκολο να χρησιμοποιηθεί ένα δέντρο που έχει δομηθεί με βάση ένα συγκεκριμένο πλαίσιο και να χρησιμοποιηθεί για διαφορετικά πλαίσια.

Μέθοδος του κοντινότερου γείτονα

Η μέθοδος του κοντινότερου γείτονα, αν και δεν χρησιμοποιείται ευρέως από τους χρηματοπιστωτικούς οργανισμούς για την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας όπως συμβαίνει με τις μεθόδους της γραμμικής και της λογιστικής παλινδρόμησης, έχει μερικά ελκυστικά χαρακτηριστικά γνωρίσματα όσον αφορά την πρακτική της εφαρμογή. Με αυτή τη μέθοδο είναι πολύ εύκολο να ενημερωθεί το δείγμα ανάπτυξης με τη προσθήκη νέων υποθέσεων στο δείγμα ανάπτυξης όταν είναι γνωστό αν οι νέοι πελάτες είναι «καλοί» ή «κακοί» και κατά συνέπεια να αφαιρεθούν παλαιότεροι πελάτες. Έτσι, κατά κάποιον τρόπο ικανοποιείται η ανάγκη να ενημερώνεται τακτικά το σύστημα βαθμολόγησης λόγω των συχνών αλλαγών στον πληθυσμό και έτσι αντανακλάται η τάση του πληθυσμού. Ένα άλλο πλεονέκτημα αυτής της τεχνικής είναι ότι δε στηρίζεται σε καμία υπόθεση για την κατανομή του πληθυσμού ή των χαρακτηριστικών του.

Το βασικότερο μειονέκτημα αυτής της μεθόδου είναι ότι για μια νέα περίπτωση χρειάζονται να γίνουν πολλοί υπολογισμοί για να βρεθούν ποιοι είναι οι k κοντινότεροι γείτονες, αλλά στη σημερινή εποχή αυτό το πρόβλημα έχει ξεπεραστεί διότι οι σύγχρονοι υπολογιστές μπορούν να κάνουν τέτοιους υπολογισμούς σε μερικά μόνο δευτερόλεπτα.

Επιπλέον, αυτή η μέθοδος δεν παράγει μια απλή πιθανότητα ταξινόμησης αλλά η ακρίβεια πρόβλεψής της επηρεάζεται πάρα πολύ από τη μέτρο απόστασης που έχει επιλεγεί και από τον αριθμό k των κοντινότερων γειτόνων.

Εντούτοις, από πολλές απόψεις, το να βρεθεί μια καλή μετρική είναι κατά πρώτον σχεδόν ισοδύναμο με την προσέγγιση της γραμμικής παλινδρόμησης. Όπως συμβαίνει με την προσέγγιση των δέντρων ταξινόμησης η τεχνική του κοντινότερου γείτονα δεν είναι ικανή να δώσει ένα αποτέλεσμα για τα χαρακτηριστικά κάθε ιδιαίτερου υποψηφίου, και στερεί από τους χρήστες τη δυνατότητα κατανόησης τι κάνει πραγματικά το σύστημα.

Όπως λοιπόν γίνεται κατανοητό, η απάντηση στο ερώτημα ποια μέθοδος δημιουργίας μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας είναι η ιδανικότερη, δεν είναι καθόλου απλή και εξαρτάται από πολλούς παράγοντες και κυρίως από τις προτιμήσεις των πιστωτικών ιδρυμάτων, από τη διαθεσιμότητα των δεδομένων, από τα χαρακτηριστικά κ.λ.π. Δεν υπάρχει κάποια μέθοδος που να είναι πάντα καλύτερη από τις υπόλοιπες για το διαχωρισμό των πελατών σε «καλούς» και «κακούς», αλλά αυτή που είναι ευρέως χρησιμοποιούμενη είναι η μέθοδος της λογιστικής παλινδρόμησης λόγω του ότι δεν χρειάζεται να γίνουν υποθέσεις για τις μεταβλητές και το μόνο εμπόδιο συναντάται σε περιπτώσεις που υπάρχουν ελλείπουσες τιμές ή πολυσυγγραμικότητα ή συσχετίσεις μεταξύ των μεταβλητών. Αντιθέτως, οι μη παραμετρικές μέθοδοι μπορούν να αντιμετωπίσουν τέτοιου είδους προβλήματα αλλά συνήθως έχουν μεγάλες υπολογιστικές απαιτήσεις και οι κανόνες που κατασκευάζονται βασισμένοι σε αυτές τις μεθόδους είναι δύσκολό να γίνουν κατανοητές από τους περισσότερους πιστωτικούς αναλυτές και πόσο μάλλον από τους πελάτες.

Γενικά, όλες οι τεχνικές απαιτούν κάποια επιλογή παραμέτρων από τον πιστωτικό αναλυτή. Για παράδειγμα, στη λογιστική παλινδρόμηση η υπόθεση που πρέπει να ισχύει είναι ότι η συνάρτηση logit διαχωρίζει όσο το δυνατόν καλύτερα τις δύο κατηγορίες. Στα δέντρα ταξινόμησης πρέπει πρώτα να προσδιοριστούν τα καλύτερα δυνατά κριτήρια, ο κανόνας διάσπασης και τερματισμού.

5.2 Σύγκριση των στατιστικών μεθόδων

Λαμβάνοντας υπ' όψιν τα πλεονεκτήματα και τα μειονεκτήματα κάθε μίας από τις μεθόδους ανάπτυξης CSM, γίνεται κατανοητό ότι δεν υπάρχει η «ιδανική» μέθοδος για την κατασκευή ενός σκορόχαρτου. Η καλύτερη μέθοδος εξαρτάται από πολλούς παράγοντες όπως είναι κάποια λεπτομερή στοιχεία της κάθε περίπτωσης, η δομή των δεδομένων, τα χαρακτηριστικά που χρησιμοποιούνται και κυρίως ο αντικειμενικός στόχος του κάθε προβλήματος.

Επιπλέον, οι συγκρίσεις των μεθόδων που έχουν γίνει από ακαδημαϊκές μελέτες δεν μπορούν να απεικονίσουν ακριβώς την κατάσταση που επικρατεί στην πραγματικότητα αφού κάποια από τα πιο σημαντικά δεδομένα, όπως αυτά που παρέχονται από τα γραφεία πίστης είναι πολύ ακριβά ή επισημαίνονται ως ευαίσθητα και δεν μπορούν να αποκτηθούν εύκολα. Κατά συνέπεια τα αποτελέσματα αυτών των μελετών είναι απλά ενδεικτικά της πραγματικότητας και γενικά υπάρχει μόνο μικρή διαφορά μεταξύ των λαθών ταξινόμησης των διαφορετικών μεθόδων. Στον Πίνακα 5.2 φαίνονται τα αποτελέσματα οκτώ διαφορετικών μελετών που αφορούν τη σύγκριση στατιστικών μεθόδων βαθμολόγησης πιστοληπτικής ικανότητας.

Πίνακας 5.2 *Σύγκριση της ακρίβειας ταξινόμησης διαφορετικών μεθόδων (πηγή: Crook (2007))*

Μέθοδος / Συγγραφέας	Γραμμική Διαχωριστική Ανάλυση	Λογιστική Παλινδρόμηση	Δέντρα Ταξινόμησης	*K*-Κοντινότεροι Γείτονες
Srinivisan (1987)	87,3	89,3	93,2	-
Boyle (1992)	77,5	-	75	-
Henley (1995)	43,4	43,3	43,8	-
Desai (1997)	66,5	67,3	-	-
Yobas (1997)	68,4	-	62,3	-
West (2000)	79,3	81,8	77	76,7
Lee (2002)	71,4	73,5	-	-
Baesens (2003)	79,3	79,3	77	78,2

Οι αριθμοί που απεικονίζονται στον πίνακα εκφράζουν το ποσοστό των σωστά ταξινομημένων πελατών που έχουν γίνει αποδεκτοί για την κάθε μέθοδο. Η σύγκριση των ποσοτήτων αυτών μπορεί να γίνει μόνο ανά στήλη και όχι ανά γραμμή γιατί κάθε ερευνητής για να διεξάγει την έρευνά του χρησιμοποίησε διαφορετικό πληθυσμό και διαφορετικό ορισμό «καλού» και «κακού» πελάτη. Για παράδειγμα, τα αποτελέσματα στην έρευνα του Henley δείχνουν πολύ μικρότερα ποσοστά σωστά ταξινομημένων πελατών επειδή στο δείγμα που χρησιμοποίησε η αναλογία των πελατών που είχαν αθετήσει τις υποχρεώσεις τους ήταν μεγαλύτερη σε σχέση με τις υπόλοιπες μελέτες.

Παρατηρούμε λοιπόν ότι, σύμφωνα με τις μελέτες των Henley (1995) και Srinivasan (1987) η καλύτερη μέθοδος είναι τα δέντρα ταξινόμησης. Σύμφωνα με τους Boyle (1992) και Yobas (1997) η γραμμική διαχωριστική ανάλυση ταξινομεί καλύτερα τους πελάτες, ενώ σύμφωνα με τους Desai (1997), West (2000) και Lee (2002) η καλύτερη μέθοδος είναι η λογιστική παλινδρόμηση. Τέλος, ο Baesens (2003) βρήκε το ίδιο ικανοποιητικές τις μεθόδους της γραμμικής και της λογιστικής παλινδρόμησης. Παρ' όλο που τα συμπεράσματα κάθε έρευνας είναι διαφορετικά, σχεδόν σε όλες τις περιπτώσεις η διαφορά του ποσοστού σωστής ταξινόμησης για κάθε μέθοδο δεν είναι σημαντική. Στις συγκεκριμένες μελέτες η μέθοδος της γραμμικής παλινδρόμησης συγχέεται με τη γραμμική διαχωριστική ανάλυση.

Η ακρίβεια της ταξινόμησης όμως αντιπροσωπεύει μόνο ένα μέρος της απόδοσης ενός σκορόχαρτου. Άλλοι πολύ σημαντικοί παράγοντες που επηρεάζουν την απόδοση ενός CSM είναι η ταχύτητα ταξινόμησης, η ταχύτητα που ένα σκορόχαρτο μπορεί να αναδομηθεί και η απλότητα της μεθόδου δηλαδή κατά πόσο είναι κατανοητή η συγκεκριμένη μέθοδος ταξινόμησης και μπορούν να εξηγηθούν τα αποτελέσματά της. Όσον αφορά την ταχύτητα ταξινόμησης, μια άμεση απόφαση είναι πολύ πιο ελκυστική σε έναν πιθανό πελάτη παρά να χρειαστεί να αναμένει τα αποτελέσματα της βαθμολόγησής του αρκετές μέρες. Επιπλέον, ένα μοντέλο που είναι ανθεκτικό στην αλλαγή τάσης του πληθυσμού και μπορεί να αναδομηθεί σύμφωνα με τη νέα τάση πιο γρήγορα και με λιγότερο κόστος είναι πολύ πιο ελκυστικό και επιθυμητό.

Επειδή όμως διαφορετικές μέθοδοι δημιουργίας CSM δίνουν σχεδόν το ίδιο επίπεδο ακρίβειας ταξινόμησης, η μέθοδος που τελικά θα χρησιμοποιηθεί για τη δόμηση ενός σκορόχαρτου εξαρτάται κυρίως από τα χαρακτηριστικά που εμφανίζει κάθε μέθοδος. Για παράδειγμα, οι μέθοδοι παλινδρόμησης επιτρέπουν την εφαρμογή διάφορων στατιστικών ελέγχων για να προσδιοριστεί πόσο σημαντικό είναι κάθε χαρακτηριστικό που

χρησιμοποιείται στο μοντέλο. Κατά συνέπεια, σε περίπτωση που δεν είναι σημαντική κάποια μεταβλητή αυτή μπορεί να εξαλειφθεί από το μοντέλο, και το CSM να γίνει πιο απλό και πιο ανθεκτικό. Επιπλέον, μελετώντας τις συσχετίσεις των μεταβλητών βρίσκουμε πόσο συσχετίζονται οι επιδράσεις διαφορετικών χαρακτηριστικών και αν χρειάζεται να βρίσκονται όλα στο σκορόχαρτο. Δηλαδή, με αυτές τις μεθόδους μπορεί να προσδιοριστεί πόσο σημαντική είναι κάθε ερώτηση που περιλαμβάνεται στην φόρμα αίτησης δανεισμού ή αν δύο ή περισσότερες ερωτήσεις αφορούν το ίδιο πράγμα. Αυτό μπορεί να ληφθεί υπ' όψιν σε μια μελλοντική φόρμα αίτησης που θα δημιουργηθεί.

Οι Lovie και Lovie (1986) πίστευαν ότι πολλά σκορόχαρτα έχουν την ίδια απόδοση με κάποια άλλα όσον αφορά την ταξινόμηση των πελατών. Αυτό σημαίνει ότι μπορεί να υπάρχουν σημαντικές διαφορές μεταξύ των βαρών των διαφορετικών σκορόχαρτων σε σχέση με το βέλτιστο αλλά μικρή επίδραση στην απόδοση τους. Αυτό το γεγονός ίσως θα μπορούσε να εξηγήσει τα παρόμοια αποτελέσματα που προκύπτουν για την απόδοση των σκορόχαρτων με τις διάφορες μεθόδους βαθμολόγησης πιστοληπτικής ικανότητας. Αυτό το φαινόμενο ονομάζεται επίπεδη μέγιστη επίδραση (*flat maximum effect*) και προέτρεψε τους ειδικούς να αναρωτηθούν εάν τα CSM που προκύπτουν είναι ανθεκτικά στη διαφορετικότητα του πληθυσμού με βάση τον οποίο δομείται (βλέπε Thomas (2000)). Δηλαδή προέκυψε το ερώτημα αν μπορεί να δομηθεί ένα CSM βασισμένο σε έναν πληθυσμό (π.χ. κάποια χώρα της Ευρώπης) και να εφαρμοστεί σε έναν άλλον πληθυσμό (π.χ. κάποια άλλη χώρα της Ευρώπης). Το αποτέλεσμα κάποιων ερευνών όπως αυτή των Platts και Howe (1997) έδειξαν ότι τα CSM είναι πολύ ευαίσθητα στη διαφορετικότητα του πληθυσμού με βάση τον οποίο αυτά δομούνται και προτείνουν ότι η πιο κατάλληλη εναλλακτική λύση είναι να γίνει κατάτμηση του πληθυσμού σε πιο ομογενείς ομάδες και να εφαρμοστούν σε άτομα που ανήκουν στις αντίστοιχες ομάδες.

ΚΕΦΑΛΑΙΟ 6

Ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας με χρήση πραγματικών δεδομένων

6.1 Γενικά χαρακτηριστικά για το σύνολο των δεδομένων

Στην παρούσα ανάλυση θα υλοποιηθούν τρία στατιστικά μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας με τρεις διαφορετικές μεθόδους (λογιστική παλινδρόμηση, διαχωριστική ανάλυση και δέντρα ταξινόμησης) χρησιμοποιώντας ένα δείγμα[15] 1000 πελατών μιας γερμανικής τράπεζας που υπέβαλαν αίτηση για κάποιο συγκεκριμένο προϊόν δανεισμού. Από αυτούς τους 1000 πελάτες, οι 700 έχουν αξιολογηθεί ως «καλοί» και οι 300 ως «κακοί». Τα δεδομένα αποτελούνται από 19 μεταβλητές κάποιες από τις οποίες αντιπροσωπεύουν την πιστωτική ιστορία κάθε πελάτη και κάποιες άλλες διάφορα προσωπικά στοιχεία. Πιο αναλυτικά, η συλλογή δεδομένων που χρησιμοποιήθηκε περιλαμβάνει τα εξής χαρακτηριστικά:

1. **CH_ACCT:** Είναι μια μεταβλητή που αντιπροσωπεύει την κατάσταση του τρεχούμενου λογαριασμού του πελάτη στη συγκεκριμένη τράπεζα. Οι κατηγορίες είναι οι εξής:
   0: λιγότερα από 0 DM[16] (ο τρεχούμενος λογαριασμός έχει αρνητικό υπόλοιπο που προκύπτει από υπέρβαση του υπολοίπου του – υπερανάληψη).
   1: από 0 έως 200 DM.

[15] Το δείγμα έχει ληφθεί από τους Asuncion and Newman (2007).
[16] Το γερμανικό μάρκο (Deutsche Mark - DM) το 1999 αντικαταστάθηκε από το ευρώ με ισοτιμία 1€=1,95583DM.

2: περισσότερα από 200 DM.

3: δεν υπάρχει τρεχούμενος λογαριασμός.

2. **DURATION:** Είναι μια συνεχής αριθμητική μεταβλητή που αντιπροσωπεύει τη συνολική διάρκεια του δανείου σε μήνες.
3. **HISTORY:** Είναι μία κατηγορική μεταβλητή με 5 επίπεδα και αντιπροσωπεύει την πιστωτική ιστορία κάθε πελάτη. Οι κατηγορίες είναι οι εξής:

0: δεν υπάρχει πιστωτική ιστορία, ο πελάτης δεν έχει πάρει δάνειο στο παρελθόν.

1: όλα τα δάνεια σε αυτήν την τράπεζα έχουν αποπληρωθεί εγκαίρως.

2: όλα τα ενεργά δάνεια έχουν αποπληρωθεί κανονικά μέχρι τη στιγμή της αίτησης για νέο δάνειο.

3: υπήρξε καθυστέρηση στην αποπληρωμή των δανείων στο παρελθόν.

4: κρίσιμη πιστωτική ιστορία (υπάρχουν και άλλα χρέη σε άλλες τράπεζες).

4. **PURPOSE:** Η μεταβλητή αυτή είναι κατηγορική με 7 επίπεδα, κάθε ένα από τα οποία αντιπροσωπεύει τον κύριο λόγο για τον οποίο κάθε πελάτης υπέβαλε την αίτηση δανειοδότησης. Οι κατηγορίες είναι οι εξής:

0: αγορά νέου αυτοκινήτου.

1: αγορά μεταχειρισμένου αυτοκινήτου.

2: αγορά επίπλων / οικιακού εξοπλισμού.

3: αγορά ραδιοφώνου / τηλεόρασης.

4: για εκπαιδευτικούς λόγους ή πληρωμή διδάκτρων.

5: για λόγους επανεκπαίδεσης.

6: άλλοι λόγοι που δεν ανήκουν στις παραπάνω κατηγορίες

5. **AMOUNT:** Είναι μια συνεχής ποσοτική μεταβλητή και αντιπροσωπεύει το συνολικό ποσό του δανείου σε γερμανικά μάρκα.
6. **SAV_ACCT:** Κατηγορική μεταβλητή με 5 επίπεδα και αντιπροσωπεύει το μέσο ποσό που έχει κάθε πελάτης σε λογαριασμούς ταμιευτηρίου. Οι κατηγορίες της μεταβλητής αυτής είναι οι εξής:

0: λιγότερα από 100 DM

1: από 100 μέχρι 500 DM

2: από 500 μέχρι 1000 DM

4: άγνωστο/δεν υπάρχει λογαριασμός ταμιευτηρίου

7. **EMPLOYMENT:** Είναι κατηγορική μεταβλητή αυτή είναι κατηγορική με 5 επίπεδα που αντιπροσωπεύουν τα χρόνια που εργάζεται κάθε πελάτης στην τωρινή του εργασία. Οι κατηγορίες της μεταβλητής αυτής είναι οι εξής:
   0: άνεργος
   1: λιγότερο από 1 χρόνο
   2: από 1 έως 4 χρόνια
   3: από 4 έως 7 χρόνια
   4: από 7 χρόνια και πάνω
8. **INSTALL_RATE:** Η μεταβλητή αυτή είναι ποσοτική και αντιπροσωπεύει το ποσοστό % της δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος.
9. **STATUS_SEX:** Η μεταβλητή αυτή είναι κατηγορική με 4 επίπεδα και αφορά το φύλο και την οικογενειακή κατάσταση κάθε πελάτη. Οι κατηγορίες είναι οι εξής:
   0: άντρας και διαζευγμένος.
   1: άντρας και άγαμος.
   2: άνδρας και έγγαμος ή χήρος.
   3: γυναίκα.
10. **OTHER_DEPTORS:** Η μεταβλητή αυτή είναι κατηγορική με 3 επίπεδα και δηλώνει αν θα υπάρξει κάποιος άλλος δανειολήπτης ή εγγυητής. Οι κατηγορίες είναι:
    0: υπάρχει συνδανειολήπτης.
    1: υπάρχει εγγυητής.
    2: δεν υπάρχει ούτε συνδανειολήπτης ούτε εγγυητής.
11. **PRESENT_RESIDENCE:** Η μεταβλητή αυτή είναι ποσοτική και αντιπροσωπεύει το χρόνο διαμονής του πελάτη στην παρούσα κατοικία του σε έτη.
12. **PROPERTY:** Η μεταβλητή αυτή είναι κατηγορική με 3 επίπεδα η και περιγράφει την περιουσιακή κατάσταση κάθε πελάτη. Οι κατηγορίες αυτής της μεταβλητής είναι:
    0: ιδιοκτήτης ακίνητης περιουσίας εκτός της μόνιμης κατοικίας.
    1: κάτοχος άλλου είδους περιουσίας (π.χ. αυτοκίνητο).
    2: άγνωστη περιουσιακή κατάσταση / δεν υπάρχει κάποιο περιουσιακό στοιχείο.
13. **AGE:** Η μεταβλητή αυτή είναι ποσοτική και αντιπροσωπεύει την ηλικία σε έτη κάθε πελάτη.

14. **OTHER_INSTALL:** Η μεταβλητή αυτή είναι δίτιμη κατηγορική η οποία δίνει την πληροφορία αν για τον πελάτη εκκρεμούν και άλλες πιστώσεις στην εν λόγω τράπεζα ή όχι.
    0: όχι.
    1: ναι.
15. **RES_STATUS:** Είναι κατηγορική μεταβλητή με 3 επίπεδα και περιγράφει την κατάσταση κατοικίας κάθε πελάτη. Τα επίπεδά της είναι τα εξής:
    0: διαμονή στην παρούσα κατοικία με καταβολή ενοικίου.
    1: ιδιοκτήτης της παρούσας κατοικίας του.
    2: άλλο (π.χ. φιλοξενούμενος).
16. **JOB:** Η μεταβλητή αυτή είναι κατηγορική με 4 επίπεδα και αντιπροσωπεύει την εργασιακή κατάσταση κάθε πελάτη. Τα επίπεδα της μεταβλητής αυτής είναι τα εξής:
    0: άνεργος/ανειδίκευτος / μη μόνιμη εργασία.
    1: ανειδίκευτος / μόνιμη εργασία.
    2: ειδικευμένος υπάλληλος.
    3: επιχειρηματίας / ελεύθερος επαγγελματίας / υψηλόβαθμο στέλεχος.
17. **NUB_DEPENDENTS:** Η μεταβλητή αυτή είναι διακριτή ποσοτική και αντιπροσωπεύει τον αριθμό των προστατευόμενων μελών κάθε πελάτη.
18. **TELEPHONE:** Η μεταβλητή αυτή είναι δίτιμη κατηγορική και δίνει την πληροφορία αν ο πελάτης είναι κάτοχος σταθερού τηλεφώνου ή όχι.
    0: όχι.
    1: ναι.
19. **FOREIGN:** Η μεταβλητή αυτή είναι δίτιμη κατηγορική και δίνει την πληροφορία αν ο πελάτης είναι μετανάστης ή όχι.
    0: όχι.
    1: ναι.
20. **RESPONSE:** Η μεταβλητή αυτή είναι η δίτιμη κατηγορική μεταβλητή απόκρισης και οι τιμές της είναι γνωστές αφού όλοι οι πελάτες του δείγματος έχουν ήδη αξιολογηθεί ως προς την πιστοληπτική τους ικανότητα. Οι κατηγορίες αυτής της μεταβλητής είναι:
    0: «κακός» πελάτης.
    1: «καλός» πελάτης.

6.2 Διερευνητική ανάλυση των δεδομένων

Πριν προχωρήσουμε στην δημιουργία στατιστικών μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας θα γίνει μια στατιστική ανάλυση των χαρακτηριστικών με χρήση του στατιστικού πακέτου SPSS 16.0. Στον Πίνακα 6.1 δίνονται κάποια βασικά περιγραφικά στατιστικά μέτρα των συνεχών χαρακτηριστικών για τους «καλούς» και τους «κακούς» πελάτες ξεχωριστά.

Πίνακας 6.1 *Περιγραφικά στατιστικά μέτρα*

Αξιολόγηση πελάτη		N	Minimum	Maximum	Mean	Std. Deviation
κακός	Διάρκεια πίστωσης σε μήνες	300	6	72	24,86	13,283
	Χρηματικό ποσό δανείου	300	433	18424	3938,13	3535,819
	Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	300	1	4	3,10	1,088
	Ηλικία σε έτη	300	19	74	33,96	11,222
	Χρόνια διαμονής στην παρούσα κατοικία	300	1	4	2,85	1,095
	Valid N (listwise)	300				
καλός	Διάρκεια πίστωσης σε μήνες	700	4	60	19,21	11,080
	Χρηματικό ποσό δανείου	700	250	15857	2985,46	2401,472
	Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	700	1	4	2,92	1,128
	Ηλικία σε έτη	700	19	75	36,22	11,381
	Χρόνια διαμονής στην παρούσα κατοικία	700	1	4	2,84	1,108
	Valid N (listwise)	700				

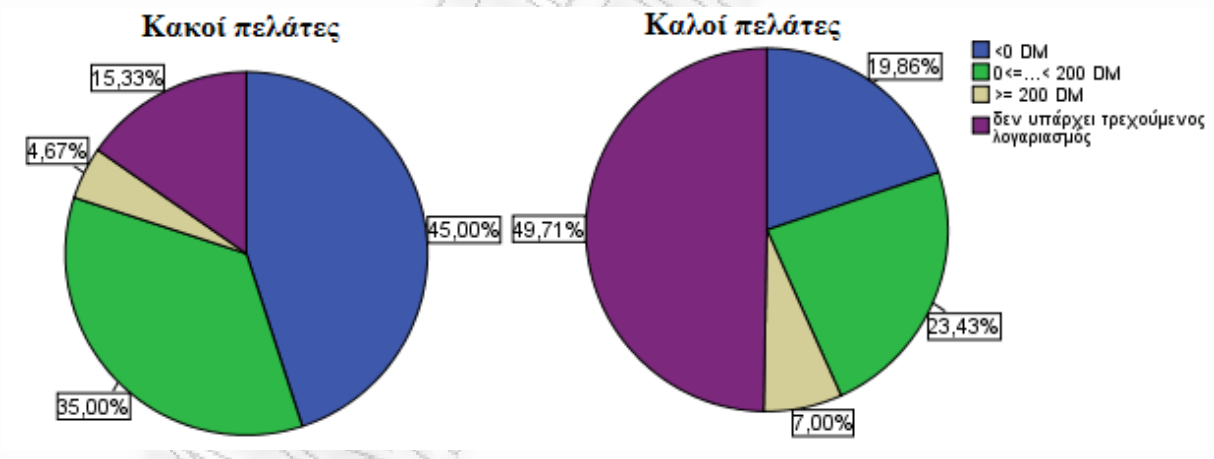
Από τον παραπάνω πίνακα παρατηρούμε ότι η μέση ηλικία των πελατών που έχουν καταφέρει να εκπληρώσουν τις υποχρεώσεις τους είναι τα 36 χρόνια ενώ γι' αυτούς που δεν τις εκπλήρωσαν είναι 34 χρόνια. Επιπλέον, οι πελάτες που αξιολογήθηκαν ως «κακοί» έλαβαν μεγαλύτερο ποσό πίστωσης και με μεγαλύτερη χρονική διάρκεια σε σχέση με τους «καλούς». Το ποσοστό δόσης αποπληρωμής επί του καθαρού εισοδήματος κυμαίνεται στα ίδια επίπεδα και για τις δύο κατηγορίες πελατών (για τους «καλούς» είναι 2,92% ενώ για τους «κακούς» λίγο μεγαλύτερο 3,1%). Τέλος, ο μέσος χρόνος διαμονής στην παρούσα κατοικία είναι σχεδόν ίδιος και για τις δύο κατηγορίες πελατών, (2,85 έτη για τους «κακούς» και 2,84

έτη για τους «καλούς»). Στον Πίνακα 6.2 δίνεται ο πίνακας συχνοτήτων της μεταβλητής CH_ACCT που αντιπροσωπεύει τον τρεχούμενο λογαριασμό, για κάθε κατηγορία πελατών και στο Σχήμα 6.1 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.2 *Τρεχούμενος λογαριασμός*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	<0 DM	135	45,0	45,0	45,0
		0<=...< 200 DM	105	35,0	35,0	80,0
		>= 200 DM	14	4,7	4,7	84,7
		δεν υπάρχει τρεχούμενος λογαριασμός	46	15,3	15,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	<0 DM	139	19,9	19,9	19,9
		0<=...< 200 DM	164	23,4	23,4	43,3
		>= 200 DM	49	7,0	7,0	50,3
		δεν υπάρχει τρεχούμενος λογαριασμός	348	49,7	49,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.1 *Τρεχούμενος λογαριασμός*



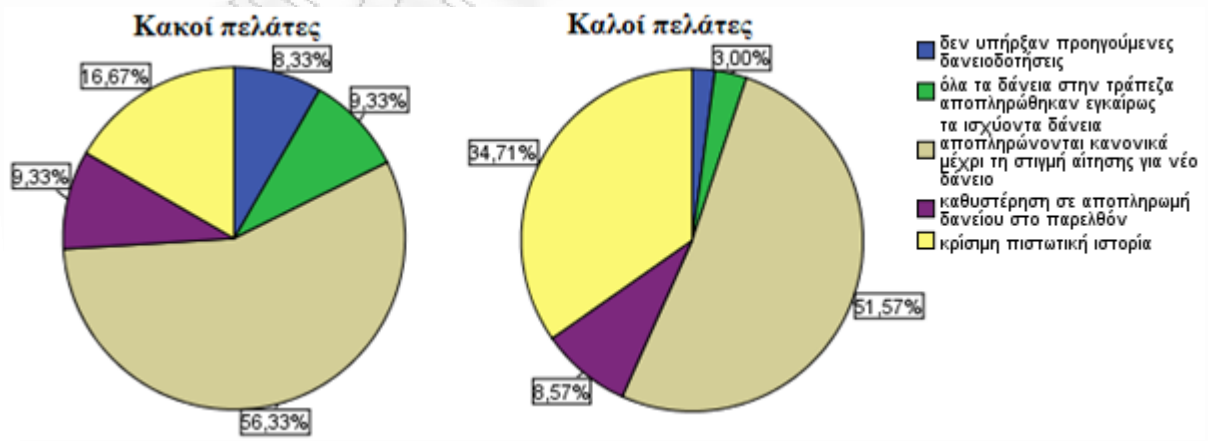
Παρατηρείται ότι το 45% των «κακών» πελατών έχουν αρνητικό υπόλοιπο στον τρεχούμενο λογαριασμό τους, ενώ το 49,7% των «καλών» πελατών δεν διαθέτουν τρεχούμενο λογαριασμό στην εν λόγω τράπεζα. Αυτό σημαίνει ότι υπάρχει σημαντική διαφοροποίηση στις δύο κατηγορίες πελατών όσον αφορά το χρηματικό ποσό που έχουν στον τρεχούμενο λογαριασμό τους και αυτό το γεγονός αποτελεί μια πρώτη ένδειξη ότι η μεταβλητή αυτή θα παίξει σημαντικό ρόλο ως προς την ταξινόμηση των πελατών σε δύο κατηγορίες. Στον Πίνακα 6.3 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής HISTORY

που αντιπροσωπεύει την πιστωτική ιστορία, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.2 δίνεται και το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.3 *Πιστωτική ιστορία*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	δεν υπήρξαν προηγούμενες δανειοδοτήσεις	25	8,3	8,3	8,3
		όλα τα δάνεια στην τράπεζα αποπληρώθηκαν εγκαίρως	28	9,3	9,3	17,7
		τα ισχύοντα δάνεια αποπληρώνονται κανονικά μέχρι τη στιγμή αίτησης για νέο δάνειο	169	56,3	56,3	74,0
		καθυστέρηση σε αποπληρωμή δανείου στο παρελθόν	28	9,3	9,3	83,3
		κρίσιμη πιστωτική ιστορία	50	16,7	16,7	100,0
		Total	300	100,0	100,0	
καλός	Valid	δεν υπήρξαν προηγούμενες δανειοδοτήσεις	15	2,1	2,1	2,1
		όλα τα δάνεια στην τράπεζα αποπληρώθηκαν εγκαίρως	21	3,0	3,0	5,1
		τα ισχύοντα δάνεια αποπληρώνονται κανονικά μέχρι τη στιγμή αίτησης για νέο δάνειο	361	51,6	51,6	56,7
		καθυστέρηση σε αποπληρωμή δανείου στο παρελθόν	60	8,6	8,6	65,3
		κρίσιμη πιστωτική ιστορία	243	34,7	34,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.2 *Πιστωτική ιστορία*



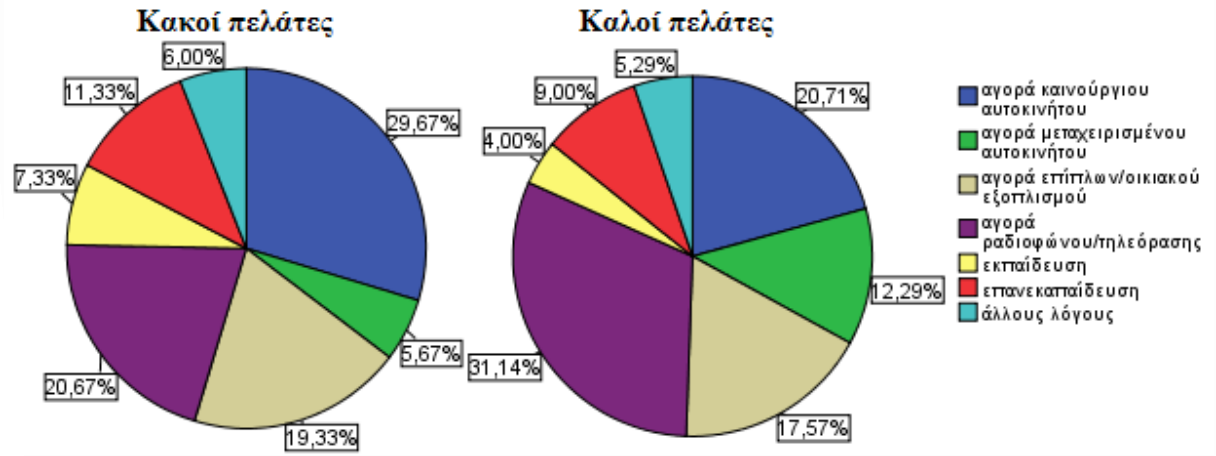
Παρατηρούμε ότι η πλειοψηφία και των «καλών» (51,57%) και των «κακών» πελατών (56,33%) αποπληρώνουν κανονικά τα δάνειά τους μέχρι τη στιγμή της αίτησης για νέο

δάνειο. Παρ' όλα αυτά αποτελεί έκπληξη το γεγονός ότι ένα μεγάλο ποσοστό των «καλών» πελατών είχαν κρίσιμη πιστωτική ιστορία. Στον Πίνακα 6.4 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής PURPOSE που αντιπροσωπεύει τον σκοπό δανειοδότησης, για κάθε κατηγορία πελατών και στο Σχήμα 6.3 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.4 *Σκοπός δανειοδότησης*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	αγορά καινούργιου αυτοκινήτου	89	29,7	29,7	29,7
		αγορά μεταχειρισμένου αυτοκινήτου	17	5,7	5,7	35,3
		αγορά επίπλων/οικιακού εξοπλισμού	58	19,3	19,3	54,7
		αγορά ραδιοφώνου/τηλεόρασης	62	20,7	20,7	75,3
		εκπαίδευση	22	7,3	7,3	82,7
		επανεκπαίδευση	34	11,3	11,3	94,0
		άλλους λόγους	18	6,0	6,0	100,0
		Total	300	100,0	100,0	
καλός	Valid	αγορά καινούργιου αυτοκινήτου	145	20,7	20,7	20,7
		αγορά μεταχειρισμένου αυτοκινήτου	86	12,3	12,3	33,0
		αγορά επίπλων/οικιακού εξοπλισμού	123	17,6	17,6	50,6
		αγορά ραδιοφώνου/τηλεόρασης	218	31,1	31,1	81,7
		εκπαίδευση	28	4,0	4,0	85,7
		επανεκπαίδευση	63	9,0	9,0	94,7
		άλλους λόγους	37	5,3	5,3	100,0
		Total	700	100,0	100,0	

Σχήμα 6.3 *Σκοπός δανειοδότησης*

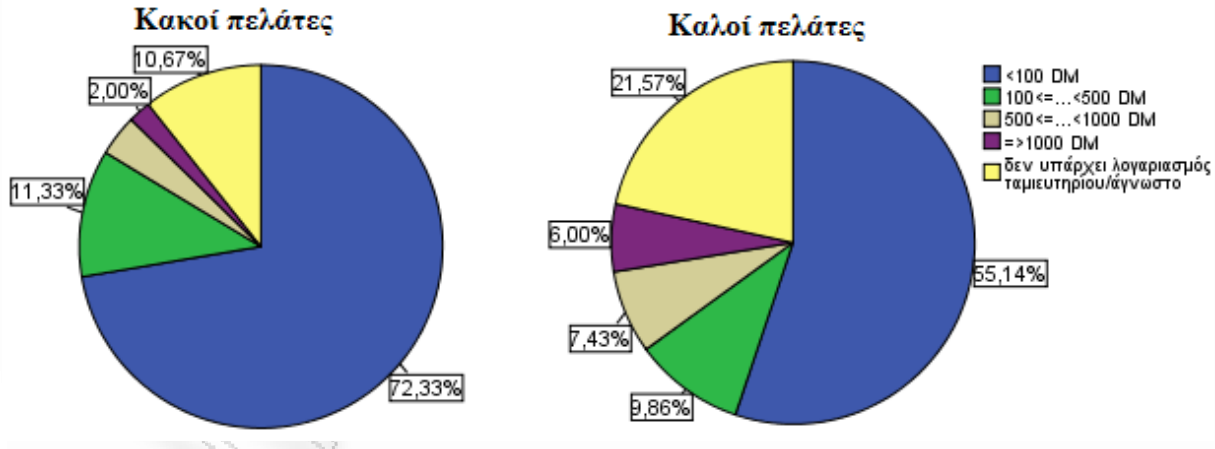


Παρατηρείται ότι το 29,7% των «κακών» πελατών ήθελαν να γίνει δεκτή η αίτησή τους για πίστωση με σκοπό να αγοράσουν καινούργιο αυτοκίνητο ενώ το 31,1% των «καλών» πελατών σκόπευαν να αγοράσουν καινούργια τηλεόραση ή ραδιόφωνο. Στον Πίνακα 6.5 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής SAV_ACCT (λογαριασμός ταμιευτηρίου) για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.4 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.5 *Λογαριασμός ταμιευτηρίου*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	<100 DM	217	72,3	72,3	72,3
		100<=...<500 DM	34	11,3	11,3	83,7
		500<=...<1000 DM	11	3,7	3,7	87,3
		=>1000 DM	6	2,0	2,0	89,3
		δεν υπάρχει λογαριασμός ταμιευτηρίου/άγνωστο	32	10,7	10,7	100,0
		Total	300	100,0	100,0	
καλός	Valid	<100 DM	386	55,1	55,1	55,1
		100<=...<500 DM	69	9,9	9,9	65,0
		500<=...<1000 DM	52	7,4	7,4	72,4
		=>1000 DM	42	6,0	6,0	78,4
		δεν υπάρχει λογαριασμός ταμιευτηρίου/άγνωστο	151	21,6	21,6	100,0
		Total	700	100,0	100,0	

Σχήμα 6.4 *Λογαριασμός ταμιευτηρίου*



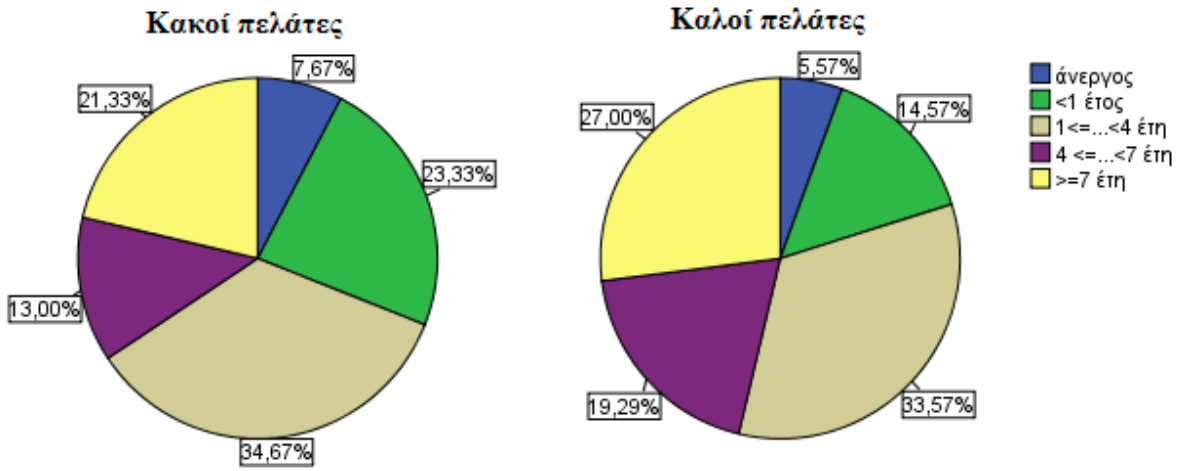
Παρατηρείται ότι η πλειοψηφία των πελατών και των δύο κατηγοριών (το 72,33% των «κακών» και το 55,14% των «καλών») έχουν στο λογαριασμό ταμιευτηρίου τους λιγότερα από 100 γερμανικά μάρκα. Στον Πίνακα 6.6 παρατίθεται ο πίνακας συχνοτήτων της

μεταβλητής EMPLOYMENT που αντιπροσωπεύει τη χρονική διάρκεια στην παρούσα εργασία, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.5 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.6 *Χρονική διάρκεια στην παρούσα εργασία*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	άνεργος	23	7,7	7,7	7,7
		<1 έτος	70	23,3	23,3	31,0
		1<=...<4 έτη	104	34,7	34,7	65,7
		4 <=...<7 έτη	39	13,0	13,0	78,7
		>=7 έτη	64	21,3	21,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	άνεργος	39	5,6	5,6	5,6
		<1 έτος	102	14,6	14,6	20,1
		1<=...<4 έτη	235	33,6	33,6	53,7
		4 <=...<7 έτη	135	19,3	19,3	73,0
		>=7 έτη	189	27,0	27,0	100,0
		Total	700	100,0	100,0	

Σχήμα 6.5 *Χρονική διάρκεια στην παρούσα εργασία*

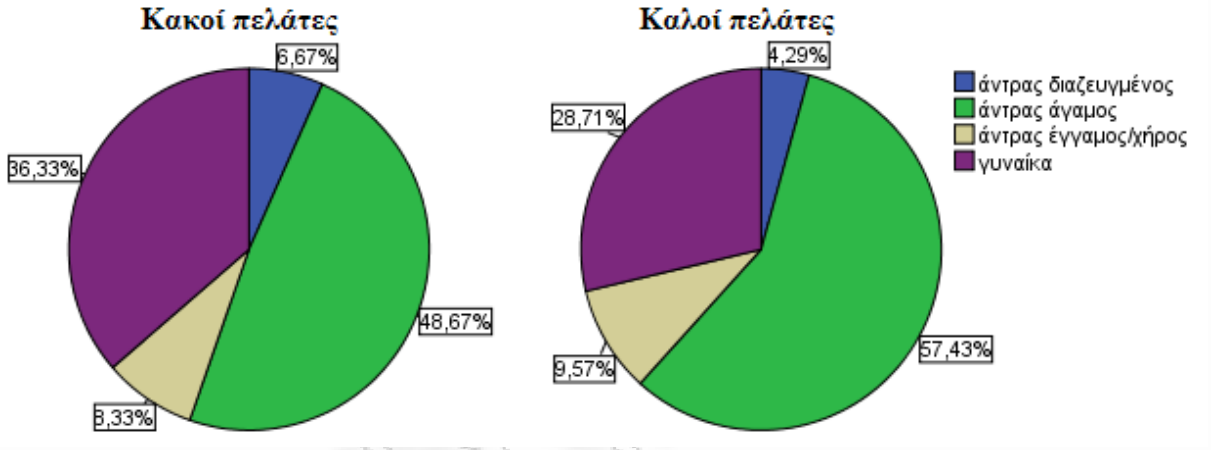


Από τα παραπάνω αποτελέσματα βλέπουμε ότι το 34,7% των «κακών» και το 33,6% των «καλών» πελατών εργάζονται από 1 έως 4 έτη στην παρούσα εργασία τους. Στον Πίνακα 6.7 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής STATUS_SEX που αντιπροσωπεύει την προσωπική κατάσταση και το φύλο, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.6 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.7 *Προσωπική κατάσταση και φύλο.*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	άντρας διαζευγμένος	20	6,7	6,7	6,7
		άντρας άγαμος	146	48,7	48,7	55,3
		άντρας έγγαμος/χήρος	25	8,3	8,3	63,7
		γυναίκα	109	36,3	36,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	άντρας διαζευγμένος	30	4,3	4,3	4,3
		άντρας άγαμος	402	57,4	57,4	61,7
		άντρας έγγαμος/χήρος	67	9,6	9,6	71,3
		γυναίκα	201	28,7	28,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.6 *Προσωπική κατάσταση και φύλο.*

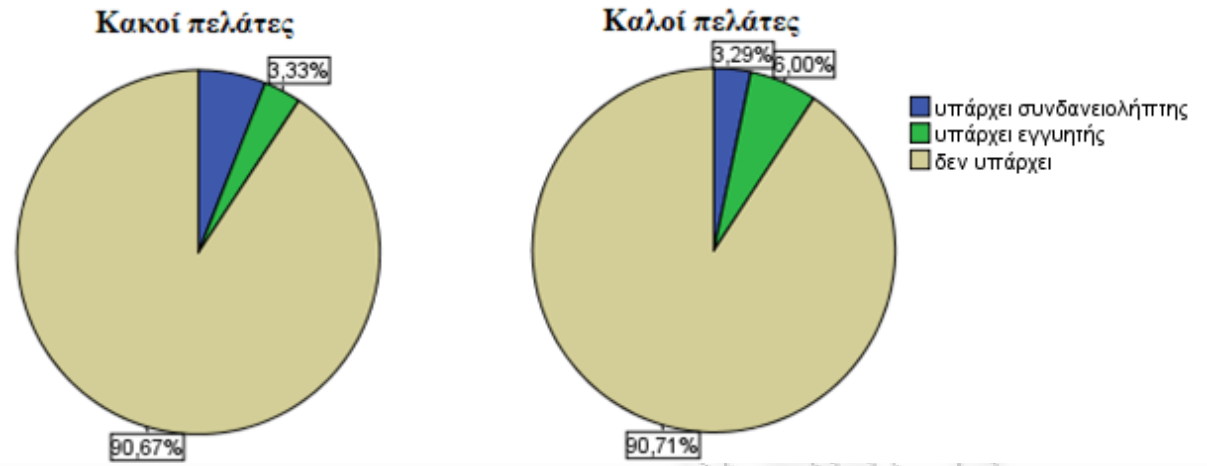


Επομένως, ότι το 48,6% των «κακών» και το 57,43% των «καλών» πελατών είναι ανύπανδροι άντρες, ενώ μόνο το 36,33% των «κακών» και το 28,71% των «καλών» πελατών είναι γυναίκες. Στον Πίνακα 6.8 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής OTHER_DEPTORS που δείχνει αν υπάρχουν συνδανειολήπτες ή εγγυητές, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.7 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.8 *Ύπαρξη άλλου δανειολήπτη ή εγγυητή*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	υπάρχει συνδανειολήπτης	18	6,0	6,0	6,0
		υπάρχει εγγυητής	10	3,3	3,3	9,3
		δεν υπάρχει	272	90,7	90,7	100,0
		Total	300	100,0	100,0	
καλός	Valid	υπάρχει συνδανειολήπτης	23	3,3	3,3	3,3
		υπάρχει εγγυητής	42	6,0	6,0	9,3
		δεν υπάρχει	635	90,7	90,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.7 *Ύπαρξη άλλου δανειολήπτη ή εγγυητή*

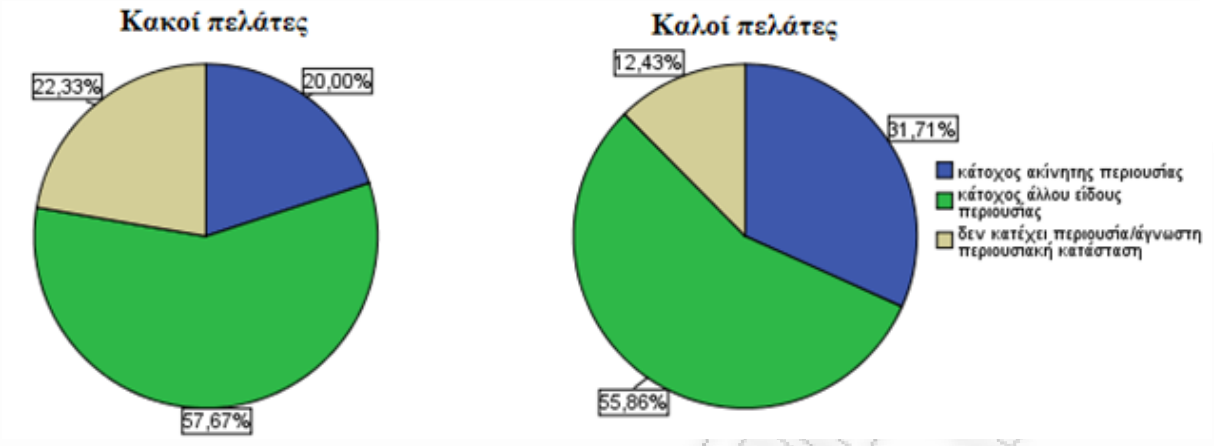


Παρατηρείται λοιπόν ότι η συντριπτική πλειοψηφία (90,7%) και των δύο κατηγοριών των πελατών δεν είχαν ούτε εγγυητή ούτε κάποιον άλλο συνυποψήφιο για την πίστωση που έλαβαν. Στον Πίνακα 6.9 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής PROPERTY που αντιπροσωπεύει την περιουσιακή κατάσταση, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.8 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.9 *Περιουσιακή κατάσταση*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	κάτοχος ακίνητης περιουσίας	60	20,0	20,0	20,0
		κάτοχος άλλου είδους περιουσίας	173	57,7	57,7	77,7
		δεν κατέχει περιουσία/άγνωστη περιουσιακή κατάσταση	67	22,3	22,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	κάτοχος ακίνητης περιουσίας	222	31,7	31,7	31,7
		κάτοχος άλλου είδους περιουσίας	391	55,9	55,9	87,6
		δεν κατέχει περιουσία/άγνωστη περιουσιακή κατάσταση	87	12,4	12,4	100,0
		Total	700	100,0	100,0	

Σχήμα 6.8 *Περιουσιακή κατάσταση*

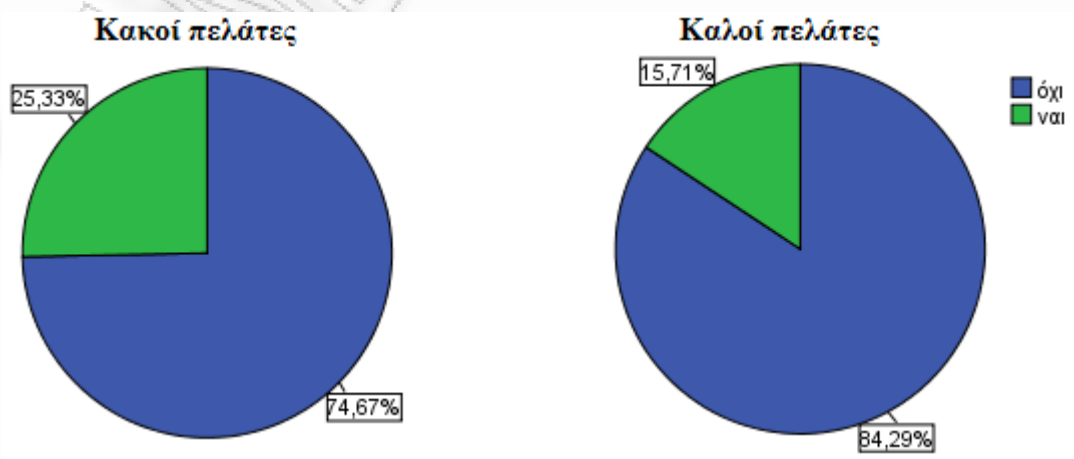


Οπότε, η πλειοψηφία και των δύο κατηγοριών πελατών είναι κάτοχοι κάποιου είδους περιουσίας εκτός από ακίνητη (57,67% για τους «κακούς» και 55,86% για τους «καλούς»). Στον Πίνακα 6.10 δίνεται ο πίνακας συχνοτήτων της μεταβλητής OTHER_INSTALL η οποία πληροφορεί για την ύπαρξη ή μη άλλων δανείων σε εξέλιξη στην εν λόγω τράπεζα, για κάθε κατηγορία πελατών. Στο Σχήμα 6.9 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.10 *Ύπαρξη άλλων δανείων σε εξέλιξη στην ίδια τράπεζα*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	όχι	224	74,7	74,7	74,7
		ναι	76	25,3	25,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	όχι	590	84,3	84,3	84,3
		ναι	110	15,7	15,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.9 *Ύπαρξη άλλων δανείων σε εξέλιξη στην ίδια τράπεζα*

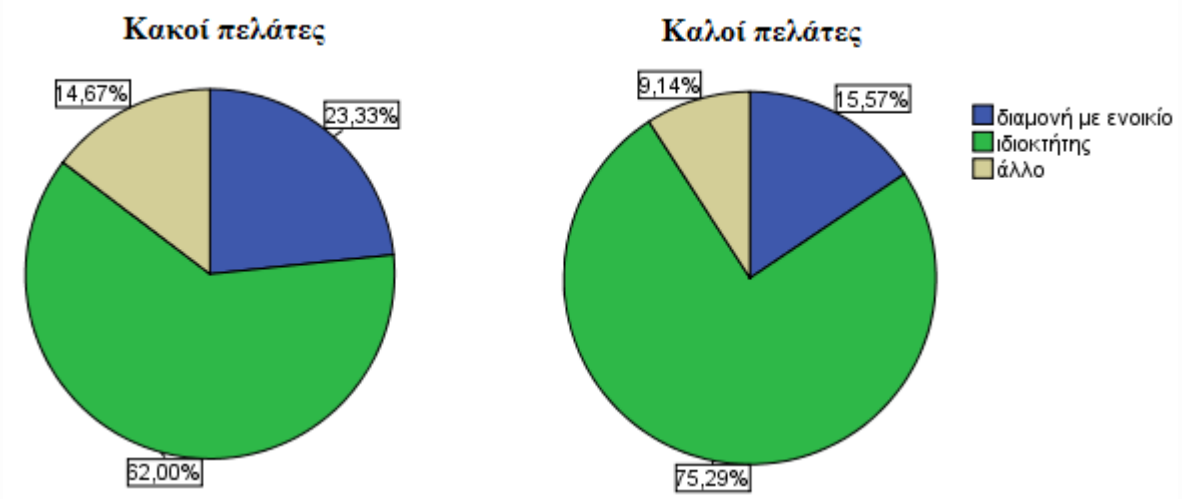


Συμπεραίνουμε λοιπόν ότι η συντριπτική πλειοψηφία και των δύο κατηγοριών των πελατών (74,67% για τους «κακούς» και το 84,29% για τους «καλούς») δεν είχαν άλλα δάνεια για αποπληρωμή στην εν λόγω τράπεζα. Στον Πίνακα 6.11 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής RES_STATUS που αντιπροσωπεύει την κατάσταση κατοικίας, για κάθε κατηγορία πελατών και στο Σχήμα 6.10 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.11 *Κατάσταση κατοικίας*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	διαμονή με ενοικίο	70	23,3	23,3	23,3
		ιδιοκτήτης	186	62,0	62,0	85,3
		άλλο	44	14,7	14,7	100,0
		Total	300	100,0	100,0	
καλός	Valid	διαμονή με ενοικίο	109	15,6	15,6	15,6
		ιδιοκτήτης	527	75,3	75,3	90,9
		άλλο	64	9,1	9,1	100,0
		Total	700	100,0	100,0	

Σχήμα 6.10 *Κατάσταση κατοικίας*

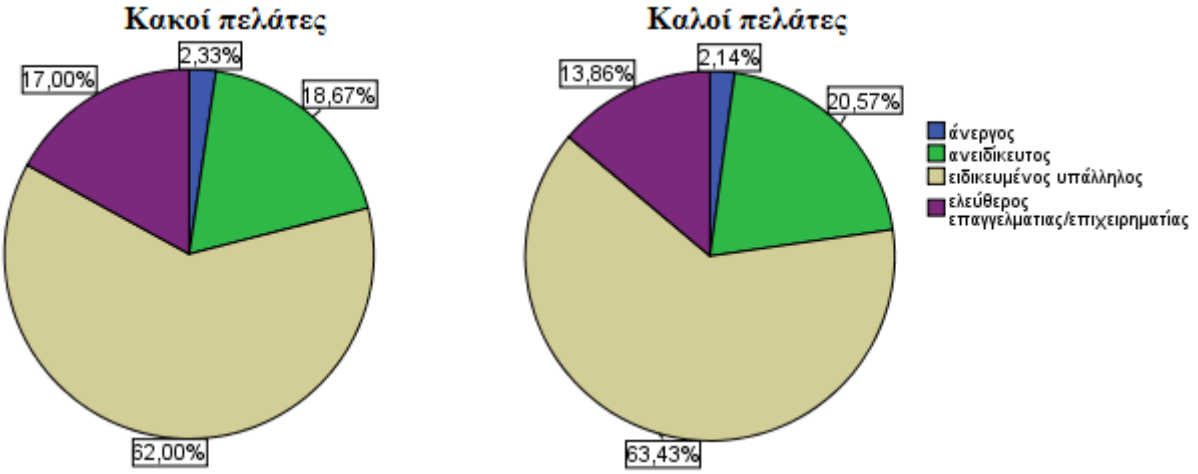


Παρατηρούμε ότι το 62% των «κακών» πελατών και το 75,29% των «καλών» είναι ιδιοκτήτες της κατοικίας στην οποία διαμένουν. Στον Πίνακα 6.12 δίνεται ο πίνακας συχνοτήτων της μεταβλητής JOB η οποία αντιπροσωπεύει την εργασιακή κατάσταση, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.11 παρατίθεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.12 *Εργασιακή κατάσταση*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	άνεργος	7	2,3	2,3	2,3
		ανειδίκευτος	56	18,7	18,7	21,0
		ειδικευμένος υπάλληλος	186	62,0	62,0	83,0
		ελεύθερος επαγγελματιας/επιχειρηματίας	51	17,0	17,0	100,0
		Total	300	100,0	100,0	
καλός	Valid	άνεργος	15	2,1	2,1	2,1
		ανειδίκευτος	144	20,6	20,6	22,7
		ειδικευμένος υπάλληλος	444	63,4	63,4	86,1
		ελεύθερος επαγγελματιας/επιχειρηματίας	97	13,9	13,9	100,0
		Total	700	100,0	100,0	

Σχήμα 6.11 *Εργασιακή κατάσταση*

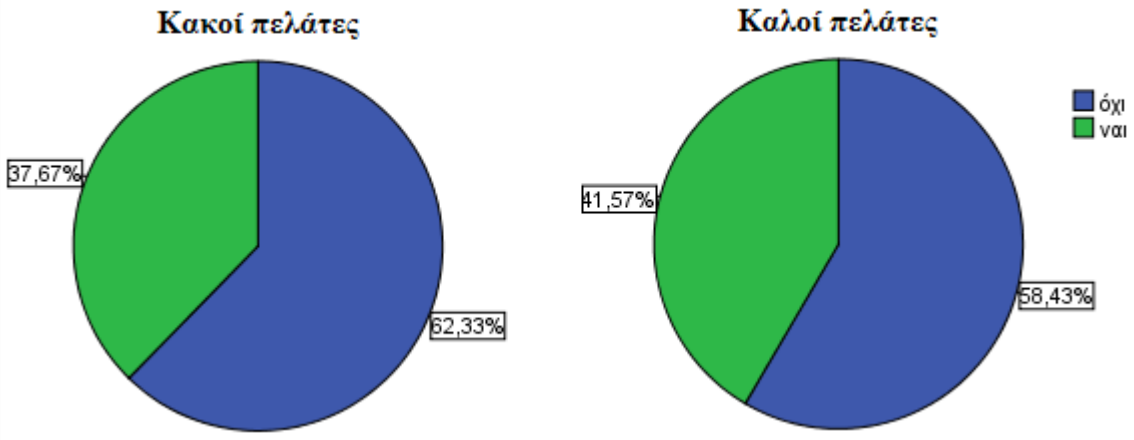


Από το παραπάνω διάγραμμα παρατηρούμε ότι το 62% των «κακών» πελατών και το 63,43% των «καλών» είναι ειδικευμένοι υπάλληλοι και μόνο ένα πολύ μικρό ποσοστό και για τις δύο κατηγορίες είναι άνεργοι. Επίσης, στον Πίνακα 6.13 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής TELEPHONE που παρέχει την πληροφορία αν ο πελάτης είναι κάτοχος ή όχι σταθερού τηλεφώνου, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.12 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.13 Κάτοχος σταθερού τηλεφώνου

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	όχι	187	62,3	62,3	62,3
		ναι	113	37,7	37,7	100,0
		Total	300	100,0	100,0	
καλός	Valid	όχι	409	58,4	58,4	58,4
		ναι	291	41,6	41,6	100,0
		Total	700	100,0	100,0	

Σχήμα 6.12 Κάτοχος σταθερού τηλεφώνου

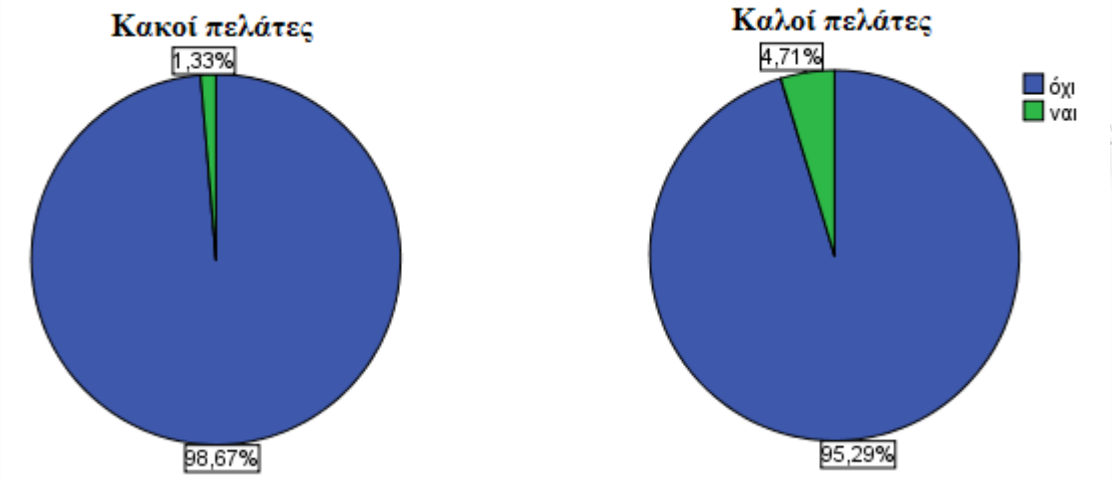


Σύμφωνα με τα παραπάνω η πλειοψηφία των πελατών σε κάθε κατηγορία (62,33% για τους «κακούς» και 58,43% για τους «καλούς») είναι κάτοχοι σταθερού τηλεφώνου. Στον Πίνακα 6.14 δίνεται ο πίνακας συχνοτήτων της μεταβλητής FOREIGN που δίνει την πληροφορία αν ο πελάτης είναι μετανάστης ή όχι, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.13 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.14 *Μετανάστης*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	όχι	296	98,7	98,7	98,7
		ναι	4	1,3	1,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	όχι	667	95,3	95,3	95,3
		ναι	33	4,7	4,7	100,0
		Total	700	100,0	100,0	

Σχήμα 6.13 *Μετανάστης*

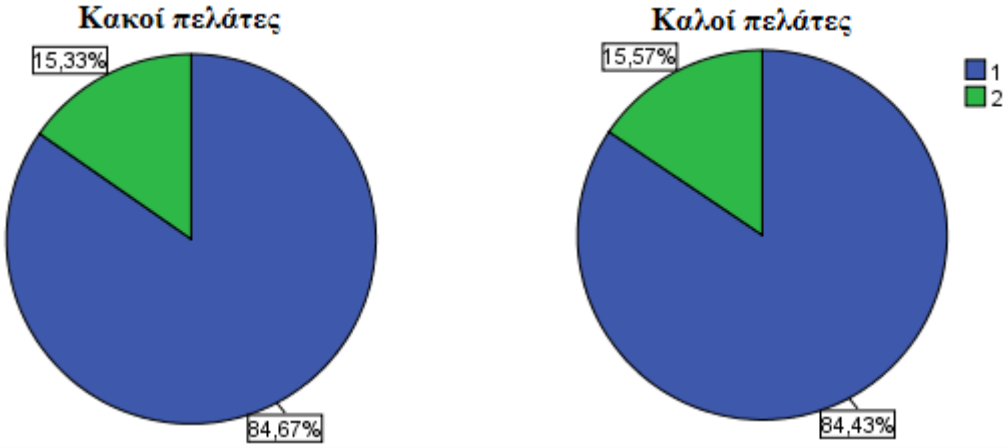


Παρατηρείται ότι η συντριπτική πλειοψηφία και των δύο κατηγοριών (98,67% για τους «κακούς» και τι 95,29% για τους «καλούς») δεν είναι μετανάστες. Τέλος, στον Πίνακα 6.15 παρατίθεται ο πίνακας συχνοτήτων της μεταβλητής NUB_DEPENDETS που αντιπροσωπεύει τον αριθμό των προστατευόμενων μελών, για κάθε κατηγορία πελατών, ενώ στο Σχήμα 6.14 δίνεται το αντίστοιχο κυκλικό διάγραμμα.

Πίνακας 6.15 *Αριθμός προστατευόμενων μελών*

Αξιολόγηση πελάτη			Frequency	Percent	Valid Percent	Cumulative Percent
κακός	Valid	1	254	84,7	84,7	84,7
		2	46	15,3	15,3	100,0
		Total	300	100,0	100,0	
καλός	Valid	1	591	84,4	84,4	84,4
		2	109	15,6	15,6	100,0
		Total	700	100,0	100,0	

Σχήμα 6.14 *Αριθμός προστατευόμενων μελών*



Άρα, συμπεραίνουμε ότι η πλειοψηφία των πελατών και των δύο κατηγοριών έχουν υπό την προστασία τους ένα μέλος (το 84,67% των «κακών» και το 84,43% των «καλών).

Ένα γενικό συμπέρασμα που προκύπτει από τα παραπάνω αποτελέσματα είναι ότι οι πελάτες και των δύο ομάδων (καλοί, κακοί) τείνουν να εμφανίζουν τις ίδιες ιδιότητες για τα περισσότερα χαρακτηριστικά. Αυτό συμβαίνει γιατί στο δείγμα αυτό περιλαμβάνονται μόνο πελάτες που έγιναν αποδεκτοί στο παρελθόν και έχουν παρόμοιο προφίλ. Παρακάτω, θα αναπτυχθούν τρία διαφορετικά στατιστικά μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας με τις μεθόδους της Λογιστικής Παλινδρόμησης, της Διαχωριστικής Ανάλυσης και των Δέντρων Ταξινόμησης. Η ανάπτυξη και των τριών μοντέλων θα γίνει με τη χρήση του στατιστικού πακέτου SPSS 16.0.

6.3 Ανάπτυξη μοντέλου Λογιστικής Παλινδρόμησης

Για τη δόμηση του μοντέλου της λογιστικής παλινδρόμησης χρησιμοποιούνται οι πρώτες 700 παρατηρήσεις του δείγματος ενώ οι υπόλοιπες 300 χρησιμοποιούνται για την επικύρωση του μοντέλου. Δηλαδή, το 70% των πελατών θα αποτελεί το δείγμα ανάπτυξης και το υπόλοιπο 30%, το δείγμα επικύρωσης. Από αυτούς τους 700 πελάτες που αποτελούν το δείγμα ανάπτυξης, οι 239 είναι «κακοί» και οι 461 είναι «καλοί». Επιπλέον, για την ανάπτυξη του μοντέλου λογιστικής παλινδρόμησης όλες οι κατηγορικές μεταβλητές μετατρέπονται σε δίτιμες και χρησιμοποιείται ως επίπεδο σύγκρισης το πρώτο. Η διαδικασία επιλογής των σημαντικότερων μεταβλητών γίνεται με χρήση βηματικής διαδικασίας (*forward LR*), στην οποία κάθε φορά προστίθεται μια μεταβλητή που είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%. Το μοντέλο που προκύπτει από την παραπάνω διαδικασία εφαρμόζεται στο δείγμα επικύρωσης και με αυτόν τον τρόπο μπορεί να υπολογιστεί η ακρίβεια του μοντέλου για διαφορετικά σημεία αποκοπής. Εδώ ως βαθμολογία αποδοχής-απόρριψης χρησιμοποιείται το 0,5 που είναι το προεπιλεγμένο σημείο αποκοπής του στατιστικού πακέτου που χρησιμοποιείται. Το τελικό μοντέλο έχει τη μορφή:

$$\log it\left(p_i\right) = \log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_1 + ... + b_n x_n ,$$

όπου x_j, $j=1,...,n$ είναι οι τιμές των αντίστοιχων μεταβλητών που συμπεριλαμβάνονται σε αυτό το μοντέλο. Στον Πίνακα 6.16 δίνονται όλες οι μεταβλητές μαζί με τους συντελεστές του τελικού υποδείγματος και τα p-values των στατιστικών ελέγχων για τη σημαντικότητα των αντίστοιχων μεταβλητών.

Πίνακας 6.16 *Μεταβλητές που περιλαμβάνονται στο τελικό μοντέλο*

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 12	CHK_ACCT			55,465	3	,000	
	CHK_ACCT(1)	,381	,239	2,547	1	,111	1,464
	CHK_ACCT(2)	1,025	,405	6,399	1	,011	2,788
	CHK_ACCT(3)	1,884	,265	50,546	1	,000	6,578
	DURATION	-,031	,010	9,026	1	,003	,969
	HISTORY			20,815	4	,000	
	HISTORY(1)	-,488	,640	,582	1	,445	,614
	HISTORY(2)	1,074	,465	5,343	1	,021	2,927
	HISTORY(3)	,987	,528	3,489	1	,062	2,683
	HISTORY(4)	1,487	,492	9,131	1	,003	4,426
	PURPOSE			25,500	6	,000	
	PURPOSE(1)	1,666	,422	15,577	1	,000	5,292
	PURPOSE(2)	,929	,295	9,932	1	,002	2,531
	PURPOSE(3)	,933	,282	10,963	1	,001	2,541
	PURPOSE(4)	-,013	,426	,001	1	,977	,988
	PURPOSE(5)	,788	,390	4,080	1	,043	2,198
	PURPOSE(6)	,852	,437	3,807	1	,051	2,345
	AMOUNT	,000	,000	3,808	1	,051	1,000
	EMPLOYMENT			11,461	4	,022	
	EMPLOYMENT(1)	,145	,450	,104	1	,747	1,156
	EMPLOYMENT(2)	,377	,417	,818	1	,366	1,458
	EMPLOYMENT(3)	1,170	,461	6,436	1	,011	3,224
	EMPLOYMENT(4)	,336	,424	,628	1	,428	1,400
	INSTALL_RATE	-,296	,098	9,108	1	,003	,744
	STATUS_SEX			10,272	3	,016	
	STATUS_SEX(1)	1,000	,417	5,753	1	,016	2,718
	STATUS_SEX(2)	,627	,508	1,525	1	,217	1,872
	STATUS_SEX(3)	,410	,427	,925	1	,336	1,507
	OTHER_DEPTORS			6,002	2	,050	
	OTHER_DEPTORS(1)	1,583	,647	5,992	1	,014	4,867
	OTHER_DEPTORS(2)	,747	,467	2,562	1	,109	2,111
	AGE	,021	,010	4,345	1	,037	1,021
	FOREIGN(1)	1,690	,801	4,453	1	,035	5,418
	Constant	-2,597	1,044	6,186	1	,013	,075

Το κριτήριο του Wald εξετάζει τη μηδενική υπόθεση $H_0: b_i = 0$ με εναλλακτική $H_1: b_i \neq 0$ για συγκεκριμένο $i = 1,2,...,n$. Ο έλεγχος αυτός γίνεται με τη βοήθεια της ποσότητας $[\hat{b}_i / se(\hat{b}_i)]^2$, η οποία, αν ισχύει η μηδενική υπόθεση, ασυμπτωτικά ακολουθεί την κατανομή χ^2 με 1 βαθμό ελευθερίας (Γναρδέλλης (2006)).

Από τον παραπάνω πίνακα και σύμφωνα με τον κριτήριο του Wald παρατηρούμε ότι οι μεταβλητές που περιλαμβάνονται στο τελικό μοντέλο είναι 27 αλλά κάποιες από αυτές δεν είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 5% αφού τα $p-value$ τους είναι μεγαλύτερα από 0,05 (π.χ. η μεταβλητή AMOUNT). Όμως, όλες οι μεταβλητές είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 10%. Αυτό σημαίνει ότι σημαντική επίδραση στη διαμόρφωση των τιμών της μεταβλητής απόκρισης έχουν οι μεταβλητές CHK_ACCT, DURATION, HISTORY, PURPOSE, AMOUNT, EMPLOYMENT, INSTALL_RATE, STATUS_SEX, OTHER_DEPTORS, AGE, FOREIGN. Οι μεταβλητές που δεν περιλαμβάνονται στο τελικό μοντέλο δίνονται στον Πίνακα 6.17, ενώ στον Πίνακα 6.18 φαίνεται ποια μεταβλητή εισέρχεται στο μοντέλο σε κάθε βήμα.

Πίνακας 6.17 *Μεταβλητές που δεν περιλαμβάνονται στο τελικό μοντέλο*

			Score	df	Sig.
Step 12	Variables	SAV_ACCT	8,393	4	,078
		SAV_ACCT(1)	,217	1	,642
		SAV_ACCT(2)	,481	1	,488
		SAV_ACCT(3)	2,289	1	,130
		SAV_ACCT(4)	4,204	1	,040
		PRESENT_RESIDENCE	,385	1	,535
		PROPERTY	2,944	2	,230
		PROPERTY(1)	,021	1	,886
		PROPERTY(2)	2,315	1	,128
		OTHER_INSTALL(1)	3,444	1	,063
		RES_STATUS	2,084	2	,353
		RES_STATUS(1)	2,038	1	,153
		RES_STATUS(2)	,912	1	,340
		JOB	,387	3	,943
		JOB(1)	,072	1	,788
		JOB(2)	,111	1	,739
		JOB(3)	,114	1	,736
		NUM_DEPENDENTS	3,429	1	,064
		TELEPHONE(1)	1,273	1	,259
		Overall Statistics	20,585	15	,151

Πίνακας 6.18 *Περίληψη της βηματικής διαδικασίας*

	Improvement			Model				
Step	Chi-square	df	Sig.	Chi-square	df	Sig.	Correct Class %	Variable
1	93,490	3	,000	93,490	3	,000	66,4%	IN: CHK_ACCT
2	28,051	1	,000	121,540	4	,000	71,0%	IN: DURATION
3	21,405	4	,000	142,945	8	,000	72,9%	IN: HISTORY
4	26,579	6	,000	169,524	14	,000	73,9%	IN: PURPOSE
5	15,208	4	,004	184,732	18	,000	73,7%	IN: EMPLOYMENT
6	8,378	1	,004	193,110	19	,000	73,7%	IN: FOREIGN
7	8,877	2	,012	201,986	21	,000	74,4%	IN: OTHER_DEPTORS
8	4,789	1	,029	206,776	22	,000	76,3%	IN: AGE
9	4,725	1	,030	211,500	23	,000	76,1%	IN: INSTALL_RATE
10	9,126	3	,028	220,627	26	,000	76,7%	IN: STATUS_SEX
11	3,859	1	,049	224,486	27	,000	77,4%	IN: AMOUNT

Η ερμηνεία που θα μπορούσε να δοθεί για να δικαιολογηθεί η απουσία των εν λόγω μεταβλητών από την ανάλυση είναι ότι στο διαχωρισμό των πελατών ως προς την πιστοληπτική τους ικανότητα δεν παίζει σημαντικό ρόλο το μέσο ποσό που έχει ο πελάτης σε λογαριασμούς ταμιευτηρίου, η περιουσιακή του κατάσταση, η κατάσταση κατοικίας ή τα χρόνια που διαμένει στην παρούσα κατοικία του. Επίσης, φαίνεται ότι δεν παίζει σημαντικό ρόλο αν για τον πελάτη βρίσκονται υπό εξέλιξη και άλλες πιστώσεις σε αυτήν την τράπεζα, το είδος της εργασίας του, αν αυτός διαθέτει ή όχι σταθερό τηλέφωνο ή ο αριθμός των ατόμων που βρίσκονται υπό την προστασία του.

Αντιθέτως, σημαντικό ρόλο στο διαχωρισμό των πελατών παίζει το ποσό που θέλει κάθε πελάτης να δανειστεί (σε γερμανικά Μάρκα), η διάρκεια του δανείου ή αν διαθέτει τρεχούμενο λογαριασμό στην εν λόγω τράπεζα και σε τι ύψος αυτός κυμαίνεται. Επιπλέον, κάποια άλλα χαρακτηριστικά που φαίνεται να είναι χρήσιμα στη διαδικασία του διαχωρισμού είναι το φύλο, η οικογενειακή του κατάσταση, ο τρόπος που θα διαθέσει τα χρήματα που θέλει να δανειστεί, το ποσοστό της δόσης του δανείου ως προς το εισόδημά του, καθώς και από το αν ήταν συνεπής στις υποχρεώσεις του απέναντι σε προηγούμενες πιστώσεις.

Ως διαγνωστικά της λογιστικής παλινδρόμησης προσφέρονται τα τυποποιημένα υπόλοιπα, τα υπόλοιπα του Pearson και τα deviance υπόλοιπα. Όμως στα πλαίσια τη λογιστικής παλινδρόμησης, αφού η μεταβλητή απόκρισης παίρνει μόνο τις τιμές 0 ή 1, τα υπόλοιπα δεν παρουσιάζουν ενδιαφέρον (Κατερή (2008)). Η αξιολόγηση της προσαρμογής του μοντέλου στα δειγματικά δεδομένα δίνεται στον Πίνακα 6.19 και γίνεται με το λόγο των μέγιστων τιμών της συνάρτησης πιθανοφάνειας (*likelihood ratio statistics*) για το εξεταζόμενο μοντέλο (L_F) και το μοντέλο που περιλαμβάνει μόνο τον σταθερό όρο (L_0).

Πίνακας 6.19 *Αξιολόγηση του μοντέλου*

		Chi-square	df	Sig.
Step 12	Step	3,859	1	,049
	Block	224,486	27	,000
	Model	224,486	27	,000

Από τον παραπάνω πίνακα παρατηρείται ότι στο 12ο βήμα της βηματικής διαδικασίας είναι $-2\ln\left(\frac{L_0}{L_F}\right)=224,486$. Η μηδενική υπόθεση είναι $H_0: b_1=b_2=...=b_{27}=0$, ενώ η πιθανότητα να προκύψει μια τιμή τόσο μεγάλη για την κατανομή χ^2 με 27 βαθμούς ελευθερίας είναι περίπου ίση με 0,00 και το μοντέλο μας είναι στατιστικά σημαντικό αφού το $p-value$ που αντιστοιχεί στο μοντέλο είναι 0,000 < 0,05 (οπότε απορρίπτεται η μηδενική υπόθεση ότι όλοι οι συντελεστές *b* είναι ίσοι με μηδέν).

Στον Πίνακα 6.20 δίνεται η τιμή της ποσότητας $-2\ln L_F$ για το τελικό μοντέλο μαζί με το συντελεστή προσδιορισμού των Cox & Snell και το συντελεστή προσδιορισμού του Nagelkerke στο 12ο βήμα της διαδικασίας. Αυτό σημαίνει ότι περίπου το 37,9% της μεταβλητότητας της μεταβλητής απόκρισης ερμηνεύεται από τις 27 ανεξάρτητες μεταβλητές του τελικού μοντέλου.

Πίνακας 6.20 *Ερμηνευσιμότητα τελικού μοντέλου*

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
12	674,284[a]	,274	,379

Το R^2 των Cox & Snell ισούται με $R^2=1-\left(\frac{L_0}{L_F}\right)^{2/n}$, όπου n είναι το μέγεθος του δείγματος. Το πρόβλημα με το συγκεκριμένο συντελεστή προσδιορισμού είναι ότι ποτέ δεν καταλήγει να πάρει μέγιστη τιμή το 1. Ο Nagelkerke (1991) πρότεινε μια τροποποίηση του συντελεστή Cox & Snell, προκειμένου να παρακαμφθεί το συγκεκριμένο πρόβλημα. Ο συντελεστής που πρότεινε ο Nagelkerke είναι

$$\tilde{R}^2=\frac{R^2}{R^2_{\max}}\in(0,1), \text{ όπου } R^2_{\max}=1-(L_0)^{2/n}.$$

Στον Πίνακα 6.21 δίνονται τα αποτελέσματα του ελέγχου προσαρμογής Hosmer and Lemeshow. Αυτός ο έλεγχος εξετάζει πόσο «κοντά» βρίσκονται οι παρατηρηθείσες τιμές προς τις αναμενόμενες πιθανότητες. Η μηδενική υπόθεση αυτού του ελέγχου είναι H_0: το μοντέλο προσαρμόζεται καλά στα δεδομένα μας. Το συμπέρασμα που προκύπτει είναι ότι στο $12^{ο}$ βήμα της διαδικασίας το τελικό μοντέλο προσαρμόζεται καλά στα δεδομένα μας αφού $p-value \approx 0,986 > 0,05$ και δεν μπορεί να απορριφθεί η μηδενική υπόθεση.

Πίνακας 6.21 *Έλεγχος Hosmer and Lemeshow*

Step	Chi-square	df	Sig.
12	1,972	8	,982

Contingency Table for Hosmer and Lemeshow Test

		Αξιολόγηση πελάτη = κακός		Αξιολόγηση πελάτη = καλός		
		Observed	Expected	Observed	Expected	Total
Step 12	1	57	57,772	13	12,228	70
	2	48	47,000	22	23,000	70
	3	37	37,260	33	32,740	70
	4	31	29,468	39	40,532	70
	5	21	22,508	49	47,492	70
	6	15	17,042	55	52,958	70
	7	15	12,075	55	57,925	70
	8	7	8,083	63	61,917	70
	9	6	5,279	64	64,721	70
	10	2	2,511	68	67,489	70

Όπως είδαμε σε προηγούμενο κεφάλαιο, ένας τρόπος για να μετρηθεί η απόδοση ενός στατιστικού μοντέλου ταξινόμησης είναι να υπολογιστούν τα ποσοστά σωστής και λανθασμένης ταξινόμησης με βάση το δείγμα ελέγχου ή το δείγμα επικύρωσης. Στον Πίνακα 6.22 δίνεται ο πίνακας ταξινόμησης (πίνακας σύγχυσης) ο οποίος δίνει για το αρχικό και τελικό βήμα της βηματικής διαδικασίας το ποσοστό ακρίβειας (ποσοστό σωστής πρόβλεψης) του μοντέλου με σημείο αποκοπής το 0,5.

Πίνακας 6.22 *Πίνακας ταξινόμησης*

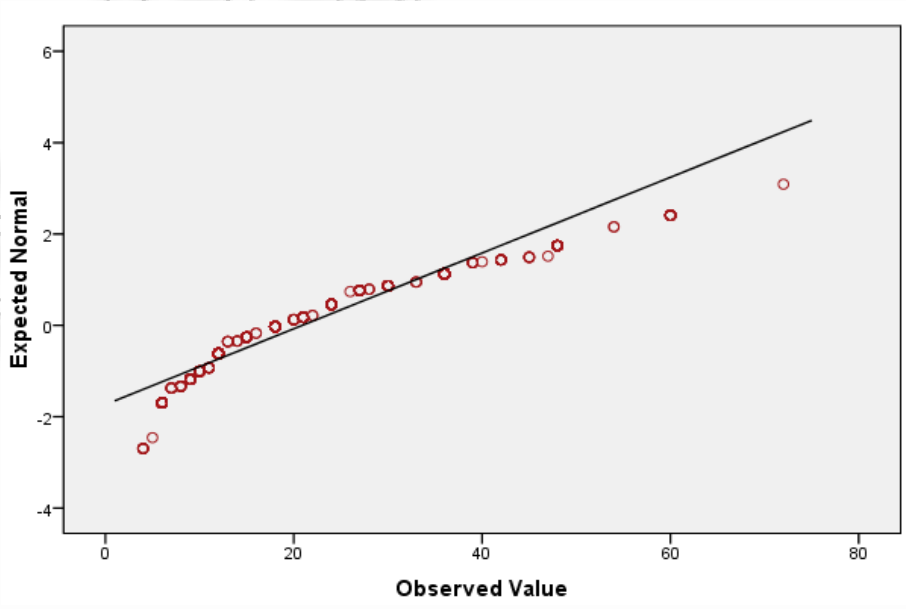
			Predicted					
			Selected Cases[a]			Unselected Cases[b]		
			Αξιολόγηση πελάτη			Αξιολόγηση πελάτη		
Observed			κακός	καλός	Percentage Correct	κακός	καλός	Percentage Correct
Step 12	Αξιολόγηση πελάτη	κακός	138	101	57,7	31	30	50,8
		καλός	57	404	87,6	40	199	83,3
		Overall Percentage			77,4			76,7

Στο τελικό μοντέλο που προέκυψε παρατηρείται ότι το ποσοστό των πελατών που έγιναν αποδεκτοί και αποδείχτηκαν «καλοί» για το δείγμα ελέγχου (*unselected cases*) είναι 83,3%, ενώ το συνολικό ποσοστό ακρίβειας του τελικού μοντέλου είναι 76,7%. Δηλαδή οι παρατηρούμενες και οι εκτιμώμενες από το μοντέλο τιμές της μεταβλητής απόκρισης συμφωνούν στο 76,7% περίπου του συνόλου των παρατηρήσεων για το δείγμα ελέγχου.

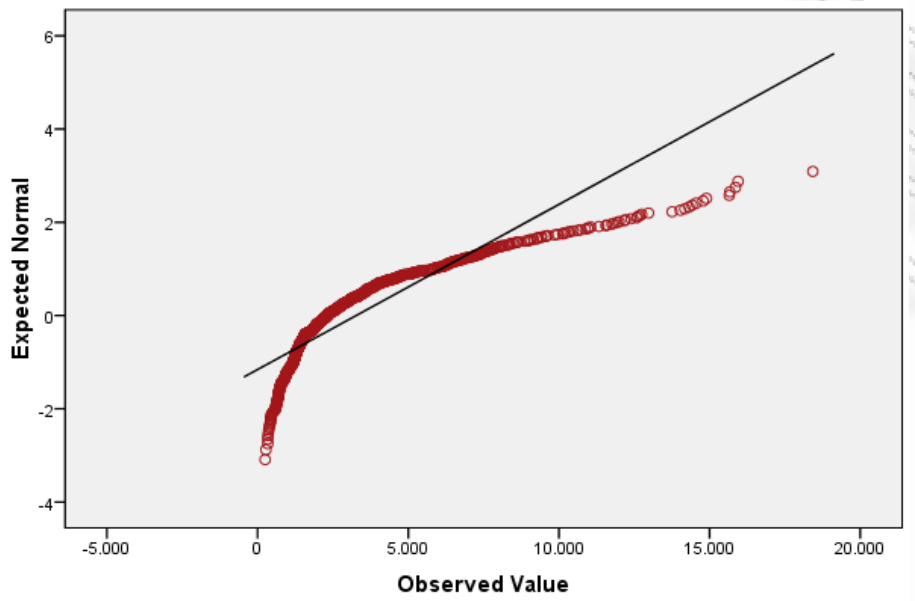
6.4 Ανάπτυξη μοντέλου Διαχωριστικής Ανάλυσης

Πριν ξεκινήσει η διαδικασία της διαχωριστικής ανάλυσης κρίνεται αναγκαίο να ελεγχτεί αν οι συνεχείς μεταβλητές είναι κανονικά κατανεμημένες. Μπορούμε να έχουμε μια πρώτη ένδειξη για την κανονικότητα των συνεχών μεταβλητών παρατηρώντας τα παρακάτω διαγράμματα στα οποία απεικονίζονται τα Normal QQ Plots για κάθε μια από τις συνεχείς μεταβλητές. Στο Σχήμα 6.15 δίνεται το Normal QQ Plot για τη μεταβλητή DURATION (διάρκεια πίστωσης σε μήνες), στο Σχήμα 6.16 για τη μεταβλητή AMOUNT (χρηματικό ποσό δανείου), στο Σχήμα 6.17 για τη μεταβλητή INSTALL_RATE (ποσοστό % της δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος) και στο Σχήμα 6.18 για τη μεταβλητή AGE (ηλικία σε έτη).

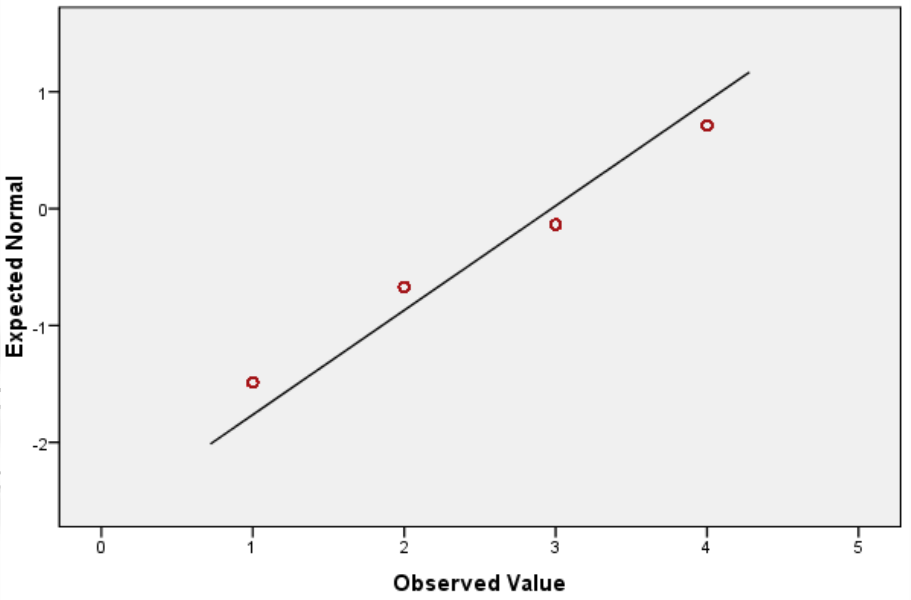
Σχήμα 6.15 *Normal QQ Plot για διάρκεια πίστωσης σε μήνες*



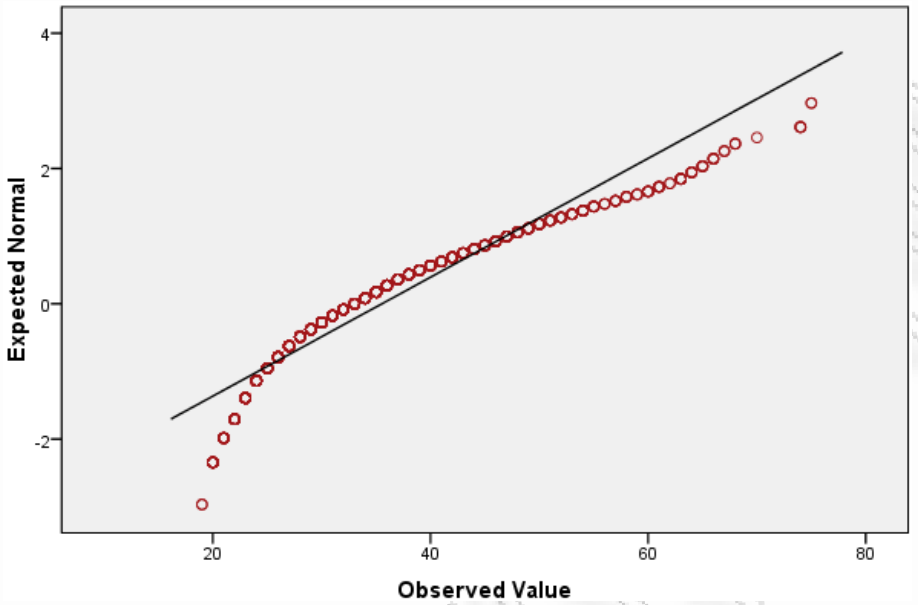
Σχήμα 6.16 *Normal QQ Plot για χρηματικό ποσό δανείου*



Σχήμα 6.17 *Normal QQ Plot για το ποσοστό % της δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος*



Σχήμα 6.18 *Normal QQ Plot για την ηλικία σε έτη*



Από τα παραπάνω σχήματα παρατηρείται ότι για όλες τις μεταβλητές τα σημεία απέχουν σημαντικά από την ευθεία, επομένως υπάρχει η ένδειξη ότι τα χαρακτηριστικά αυτά δεν είναι κανονικά κατανεμημένα. Επίσης, στον Πίνακα 6.23 δίνεται ο έλεγχος κανονικότητας των συνεχών μεταβλητών μέσω των κριτηρίων Kolmogorov – Smirnov και Shapiro Wilk.

Πίνακας 6.23 *Έλεγχος κανονικότητας*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Διάρκεια πίστωσης σε μήνες	,169	1000	,000	,900	1000	,000
Χρηματικό ποσό δανείου	,165	1000	,000	,793	1000	,000
Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	,297	1000	,000	,789	1000	,000
Ηλικία σε έτη	,111	1000	,000	,917	1000	,000

a. Lilliefors Significance Correction

Με βάση τον παραπάνω πίνακα απορρίπτεται η μηδενική υπόθεση της κανονικότητας για όλες τις συνεχείς μεταβλητές αφού $p-value = 0,00 < 0,05$ όπου 0,05 είναι το επίπεδο σημαντικότητας. Αφού απορρίφθηκε η κανονικότητα των συνεχών μεταβλητών τα αποτελέσματα της διαχωριστικής ανάλυσης αναμένονται να μην είναι τόσο ικανοποιητικά. Όμως, η ανάλυση συνεχίζεται κανονικά παρ' όλο που είναι γνωστό πως δεν ισχύει η υπόθεση

της κανονικότητας. Σύμφωνα με τους Reichert et al (1983) «το γεγονός ότι ένα σημαντικό ποσοστό της πιστωτικής πληροφορίας δεν είναι κανονικά κατανεμημένο, μπορεί να μην αποτελεί σημαντικό περιορισμό».

Για τη δόμηση του μοντέλου της διαχωριστικής ανάλυσης χρησιμοποιούνται όπως και στο μοντέλο της Λογιστικής Παλινδρόμησης οι 700 πρώτες παρατηρήσεις του δείγματος, ενώ οι υπόλοιπες 300 χρησιμοποιούνται για την επικύρωση του μοντέλου. Επιπλέον, η διαδικασία επιλογής των σημαντικότερων μεταβλητών γίνεται με τη χρήση βηματικής διαδικασίας. Στον Πίνακα 6.24 δίνεται ο έλεγχος της υπόθεσης ισότητας των μέσων της κάθε ομάδας για κάθε μεταβλητή.

Πίνακας 6.24 *Έλεγχος υπόθεσης των μέσων της κάθε ομάδας*

	Wilks' Lambda	F	df1	df2	Sig.
Τρεχούμενος λογαριασμός	,880	95,057	1	698	,000
Διάρκεια πίστωσης σε μήνες	,950	36,630	1	698	,000
Πιστωτική ιστορία	,956	32,273	1	698	,000
Σκοπός δανειοδότησης	1,000	,271	1	698	,603
Χρηματικό ποσό δανείου	,975	17,860	1	698	,000
Λογαριασμός ταμιευτηρίου	,981	13,445	1	698	,000
Χρονική διάρκεια στην παρούσα εργασία	,986	9,884	1	698	,002
Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	,993	5,157	1	698	,023
Προσωπική κατάσταση και φύλο	,995	3,841	1	698	,050
Άλλοι δανειολήπτες/εγγυητές	,999	,586	1	698	,444
Χρόνια διαμονής στην παρούσα κατοικία	1,000	,163	1	698	,687
Περιουσιακή κατάσταση	,977	16,350	1	698	,000
Ηλικία σε έτη	,991	6,439	1	698	,011
Υπάρξη άλλων πιστώσεων με δόσεις	,988	8,306	1	698	,004
Κατάσταση κατοικίας	1,000	,114	1	698	,736
Φύση εργασίας	,999	,935	1	698	,334
Προστατευόμενα μέλη	1,000	,111	1	698	,739
Κάτοχος σταθερού τηλεφώνου	,999	,548	1	698	,460
Μετανάστης	,989	7,947	1	698	,005

Από τον παραπάνω πίνακα παρατηρείται ότι για τις μεταβλητές OTHER_DEPTORS, PRESENT_RESIDENCE, RES_STATUS, JOB, NUB_DEPENDENTS, TELEPHONE οι

μέσοι για τα δύο επίπεδα της μεταβλητής RESPONSE είναι ίσοι αφού τα $p-value$ είναι μεγαλύτερα από το 0,05, οπότε δεν απορρίπτεται η μηδενική υπόθεση (H_0 : οι μέσοι των δύο επιπέδων της μεταβλητής RESPONSE είναι ίσοι.) Αναμένεται λοιπόν ότι οι μεταβλητές αυτές δεν θα είναι στατιστικά σημαντικές για το διαχωρισμό των πελατών ως προς την πιστοληπτική τους ικανότητα και ότι θα αφαιρεθούν με τη βηματική μέθοδο από τη διαχωριστική ανάλυση. Στον Πίνακα 6.25 φαίνονται οι μεταβλητές που χρησιμοποιούνται στην ανάλυσή μας στο βήμα 8 που είναι και το τελικό βήμα.

Πίνακας 6.25 *Μεταβλητές που χρησιμοποιούνται στην ανάλυση*

Step		Tolerance	F to Remove	Min. D Squared	Between Groups
8	Τρεχούμενος λογαριασμός	,957	62,747	,727	κακός and καλός
	Διάρκεια πίστωσης σε μήνες	,936	22,126	1,021	κακός and καλός
	Πιστωτική ιστορία	,968	13,469	1,088	κακός and καλός
	Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	,975	9,216	1,122	κακός and καλός
	Χρονική διάρκεια στην παρούσα εργασία	,947	7,635	1,134	κακός and καλός
	Περιουσιακή κατάσταση	,920	5,377	1,152	κακός and καλός
	Λογαριασμός ταμιευτηρίου	,953	6,052	1,147	κακός and καλός
	Υπάρξη άλλων πιστώσεων με δόσεις	,977	4,140	1,162	κακός and καλός

Παρατηρούμε ότι σε αυτές τις μεταβλητές δεν συμπεριλαμβάνονται οι μεταβλητές για τις οποίες δεν μπορεί να απορριφθεί η ισότητα των μέσων των δύο επιπέδων της μεταβλητής RESPONSE. Στον Πίνακα 6.26 δίνονται τα αποτελέσματα του ελέγχου ισότητας των δύο πινάκων συνδυακύμανσης του Box για τις δύο κατηγορίες πελατών.

Πίνακας 6.26 *Έλεγχος Box για την ισότητα των πινάκων συνδυακύμανσης*

Log Determinants

Αξιολόγηση πελάτη	Rank	Log Determinant
κακός	8	3,886
καλός	8	3,625
Pooled within-groups	8	3,851

Test Results

Box's M		95,811
F	Approx.	2,625
	df1	36,000
	df2	824128,775
	Sig.	,000

Ο έλεγχος Box ελέγχει αν ισχύει η μηδενική υπόθεση $H_0 : \mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$. Από τους παραπάνω πίνακες παρατηρείται ότι απορρίπτεται η μηδενική υπόθεση της ισότητας των πινάκων συνδυακύμανσης αφού $p-value = 0,00 < 0,05$. Ο έλεγχος αυτός είναι ευαίσθητος στην κανονικότητα και επειδή στα δεδομένα μας δεν ισχύει η κανονικότητα ήταν αναμενόμενο να απορριφθεί η μηδενική υπόθεση. Συνήθως σε πραγματικά δεδομένα είναι πολύ σπάνιο να βρεθεί ότι ικανοποιείται η υπόθεση της κανονικότητας και πόσο μάλλον η υπόθεση της ισότητας των πινάκων διακύμανσης-συνδιακύμανσης. Γι' αυτό το λόγο η διαχωριστική ανάλυση εφαρμόζεται ακόμα και όταν δεν ισχύουν οι προηγούμενες υποθέσεις. Ο Πίνακας 6.27 δίνει τις εκ των προτέρων πιθανότητες των ομάδων. Αφού δεν είναι διαθέσιμη κάποιου είδους πληροφόρηση για τα πραγματικά ποσοστά των ομάδων στον πραγματικό πληθυσμό τότε θεωρούμε ότι στο δείγμα διατηρείται και η ίδια αναλογία.

Πίνακας 6.27 *Εκ των προτέρων πιθανότητες των δύο κατηγοριών*

Αξιολόγηση πελ...	Prior	Cases Used in Analysis	
		Unweighted	Weighted
κακός	,500	239	239,000
καλός	,500	461	461,000
Total	1,000	700	700,000

Για κάθε ομάδα υπολογίζεται μια βαθμολογία με βάση κάποιο μοντέλο το οποίο είναι γραμμικό ως προς τα χαρακτηριστικά. Οι συντελεστές των γραμμικών συναρτήσεων των βαθμολογιών με την μέθοδο του Fisher δίνεται στον Πίνακα 6.28.

Πίνακας 6.28 *Συντελεστές διαχωριστικής ανάλυσης*

	Αξιολόγηση πελάτη	
	κακός	καλός
Τρεχούμενος λογαριασμός	,351	,951
Διάρκεια πίστωσης σε μήνες	,140	,104
Πιστωτική ιστορία	1,841	2,157
Λογαριασμός ταμιευτηρίου	-,038	,108
Χρονική διάρκεια στην παρούσα εργασία	,882	1,097
Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος	2,232	1,985
Περιουσιακή κατάσταση	1,589	1,255
Υπάρξη άλλων πιστώσεων με δόσεις	1,501	1,021
(Constant)	-10,023	-10,340

Επομένως, το μοντέλο της διαχωριστικής ανάλυσης για τους «κακούς» πελάτες θα είναι:

$w_0 = -10,023 + 0,351x_1 + 0,14x_2 + 1,841x_3 - 0,038x_4 + 0,882x_5 + 2,232x_6 + 1,589x_7 + 1,501x_8$,

όπου x_i, $i = 1,...,8$ είναι οι τιμές των χαρακτηριστικών που περιλαμβάνονται στο μοντέλο με τη σειρά που δίνονται στον παραπάνω πίνακα (από πάνω προς τα κάτω). Όμοια, μπορεί να εκφραστεί και το μοντέλο για τους «καλούς» πελάτες :

$w_1 = -10,340 + 0,951x_1 + 0,104x_2 + 2,157x_3 + 0,108x_4 + 1,097x_5 + 1,985x_6 + 1,255x_7 + 1,021x_8$

Άρα, κάθε νέος υποψήφιος πελάτης της τράπεζας μπορεί καταταχτεί σε μία από τις δύο κατηγορίες με βάση τα παραπάνω μοντέλα. Πιο αναλυτικά, χρησιμοποιώντας τις ιδιότητες των χαρακτηριστικών του πελάτη υπολογίζονται οι ποσότητες w_0 και w_1. Αν $w_0 > w_1$ τότε ο πελάτης κατατάσσεται στην πρώτη κατηγορία, δηλαδή δεν είναι άξιος για να λάβει από την τράπεζα αυτού του είδους πίστωσης, διαφορετικά κατατάσσεται στην άλλη κατηγορία.

Παρακάτω, στον Πίνακα 6.29 δίνεται ο πίνακας ταξινόμησης στον οποίο φαίνονται τα ποσοστά σωστής ταξινόμησης με βάση το δείγμα επικύρωσης.

Πίνακας 6.29 *Πίνακας ταξινόμησης*

			Αξιολόγηση πελ...	Predicted Group Membership		
				κακός	καλός	Total
Cases Selected	Original	Count	κακός	173	66	239
			καλός	131	330	461
		%	κακός	72,4	27,6	100,0
			καλός	28,4	71,6	100,0
Cases Not Selected	Original	Count	κακός	49	12	61
			καλός	65	174	239
		%	κακός	80,3	19,7	100,0
			καλός	27,2	72,8	100,0

a. 71,9% of selected original grouped cases correctly classified.

b. 74,3% of unselected original grouped cases correctly classified.

Στο τελικό μοντέλο που προέκυψε παρατηρείται ότι το ποσοστό των πελατών που έγιναν αποδεκτοί και αποδείχτηκαν «καλοί» για το δείγμα ελέγχου (*unselected cases*) είναι 72,8%, ενώ το συνολικό ποσοστό ακρίβειας του τελικού μοντέλου είναι 74,3%. Δηλαδή οι παρατηρούμενες και οι εκτιμώμενες από το μοντέλο τιμές της μεταβλητής απόκρισης συμφωνούν στο 74,3% περίπου του συνόλου των παρατηρήσεων για το δείγμα ελέγχου.

6.5 Ανάπτυξη ενός μοντέλου με Δέντρα Ταξινόμησης

Για τη δόμηση του δέντρου ταξινόμησης σε αυτό το σύνολο δεδομένων χρησιμοποιείται η μέθοδος CRT (*Classification and Regression Trees*). Η μέθοδος αυτή χρησιμοποιεί αναδρομική διαμέριση ώστε να διασπαστεί το σύνολο που χρησιμοποιείται για την δόμηση του δέντρου σε υποσύνολα με κύριο στόχο σε κάθε επανάληψη να μεγιστοποιείται η ακεραιότητα ή η καθαρότητα ή ομοιογένεια ενός κόμβου του δέντρου. Ένας κόμβος θεωρείται ακέραιος ή καθαρός (αγνός) όταν το 100% των περιπτώσεων που αντιστοιχούν στην μεταβλητή την οποία αναπαριστά ο κόμβος ανήκουν σε μια συγκεκριμένη κατηγορία της μεταβλητής απόκρισης, δηλαδή όταν όλοι οι πελάτες που ανήκουν σε αυτόν τον κόμβο είναι μόνο «καλοί» ή μόνο «κακοί».

Με αυτήν τη μέθοδο, οι μεταβλητές εισόδου (επεξηγηματικές μεταβλητές) μπορούν να παίρνουν είτε συνεχείς είτε διακριτές τιμές αφού κάθε διαμέριση είναι δυαδική, δηλαδή δημιουργούνται δύο υποομάδες σε κάθε επανάληψη, άρα κάθε κόμβος του δέντρου έχει δύο παιδιά.

Όπως και στις δύο προηγούμενες μεθόδους έτσι και σε αυτήν, για τη δόμηση του δέντρου ταξινόμησης χρησιμοποιούνται οι 700 πρώτες παρατηρήσεις, ενώ οι υπόλοιπες 300 χρησιμοποιούνται ως δείγμα επικύρωσης.

Η μέθοδος CRT αποσκοπεί στη μεγιστοποίηση της ομοιογένειας σε κάθε κόμβο. Για να μετρηθεί η αγνότητα κάθε κόμβου χρησιμοποιείται ο δείκτης Gini, ο οποίος παίρνει την τιμή 0 όταν όλες οι περιπτώσεις που ανήκουν στον κόμβο ανήκουν στην ίδια κατηγορία της μεταβλητής απόκρισης. Επίσης, για το δέντρο ταξινόμησης που αναπτύσσεται παίρνουμε ίσες εκ των προτέρων πιθανότητες για τις δύο κατηγορίες αφού δεν είναι γνωστές οι πιθανότητες στον πληθυσμό.

Στον Πίνακα 6.30 δίνονται κάποια γενικά χαρακτηριστικά που αφορούν το δέντρο ταξινόμησης. Ο πίνακας αυτός δείχνει ότι μόνο 10 από τις 19 επεξηγηματικές μεταβλητές συνεισφέρουν σημαντικά και περιλαμβάνονται στο τελικό μοντέλο. Αυτά τα χαρακτηριστικά είναι ο τρεχούμενος λογαριασμός, ο λογαριασμός ταμιευτηρίου, πιστωτική ιστορία, διάρκεια πίστωσης σε μήνες, ηλικία σε έτη, χρηματικό ποσό δανείου, μετανάστης, ποσοστό % της δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος, προσωπική κατάσταση και φύλο και ύπαρξη άλλων πιστώσεων με δόσεις. Επιπλέον, το δέντρο έχει 11 κόμβους από τους οποίους οι 6 είναι τελικοί.

Πίνακας 6.30 *Γενικά χαρακτηριστικά του δέντρου ταξινόμησης*

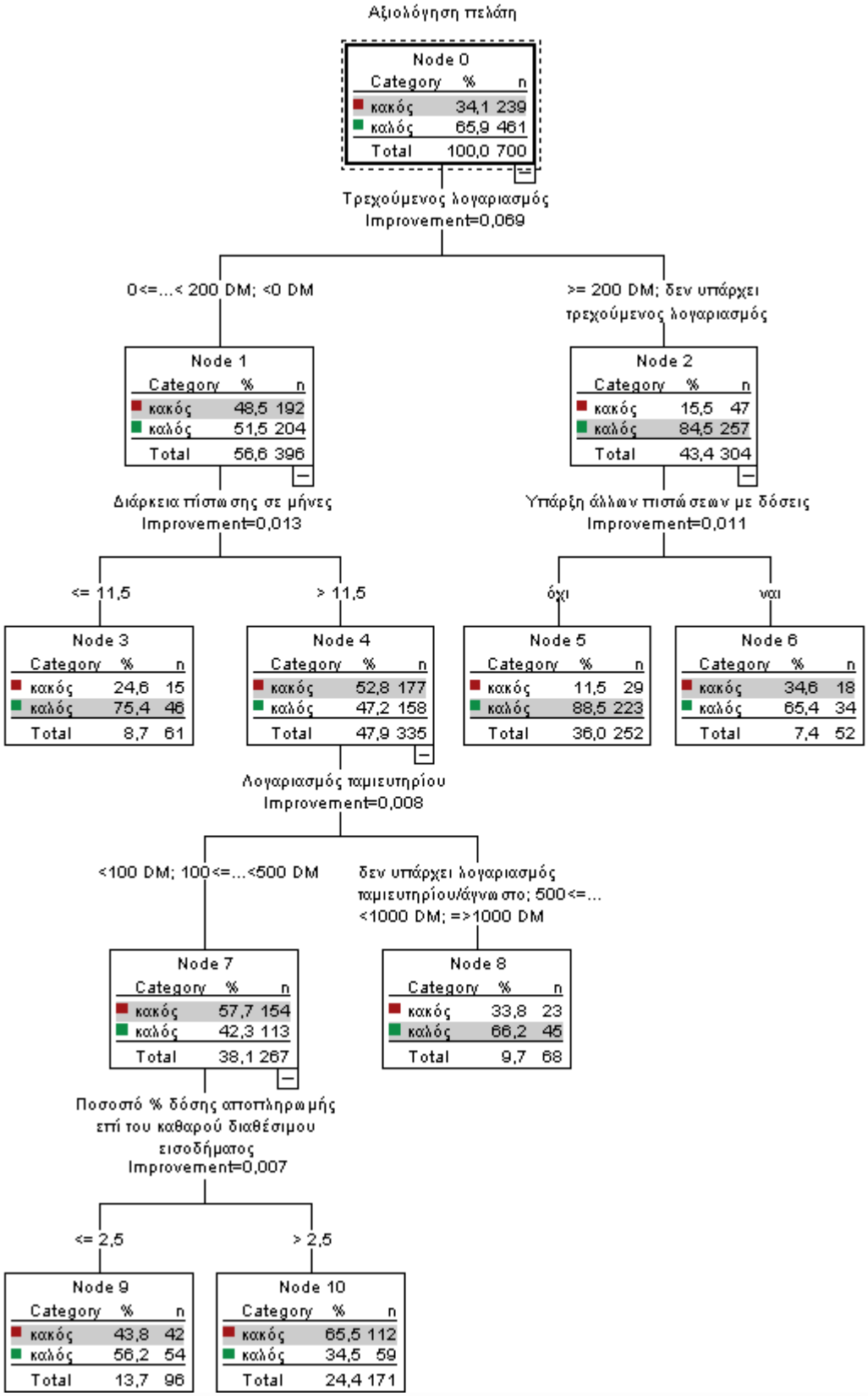
Specifications	Growing Method	CRT
	Dependent Variable	Αξιολόγηση πελάτη
	Independent Variables	Τρεχούμενος λογαριασμός, Διάρκεια πίστωσης σε μήνες, Πιστωτική ιστορία, Σκοπός δανειοδότησης, Χρηματικό ποσό δανείου, Λογαριασμός ταμιευτηρίου, Χρονική διάρκεια στην παρούσα εργασία, Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος, Προσωπική κατάσταση και φύλο, Άλλοι δανειολήπτες/εγγυητές, Χρόνια διαμονής στην παρούσα κατοικία, Περιουσιακή κατάσταση, Ηλικία σε έτη, Ύπαρξη άλλων πιστώσεων με δόσεις, Κατάσταση κατοικίας, Φύση εργασίας, Προστατευόμενα μέλη, Κάτοχος σταθερού τηλεφώνου, Μετανάστης
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Τρεχούμενος λογαριασμός, Λογαριασμός ταμιευτηρίου, Πιστωτική ιστορία, Διάρκεια πίστωσης σε μήνες, Ηλικία σε έτη, Χρηματικό ποσό δανείου, Μετανάστης, Ποσοστό % δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος, Προσωπική κατάσταση και φύλο, Ύπαρξη άλλων πιστώσεων με δόσεις
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	4

Το δέντρο ταξινόμησης για το δείγμα ανάπτυξης δίνεται στο Σχήμα 6.19, ενώ το αντίστοιχο δέντρο ταξινόμησης για το δείγμα επικύρωσης απεικονίζεται στο Σχήμα 6.20 και η γραμμοσκιασμένη κατηγορία σε κάθε τελικό κόμβο αποτελεί την κατηγορία στην οποία προβλέπεται ότι ανήκουν οι πελάτες σε αυτόν τον τελικό κόμβο.

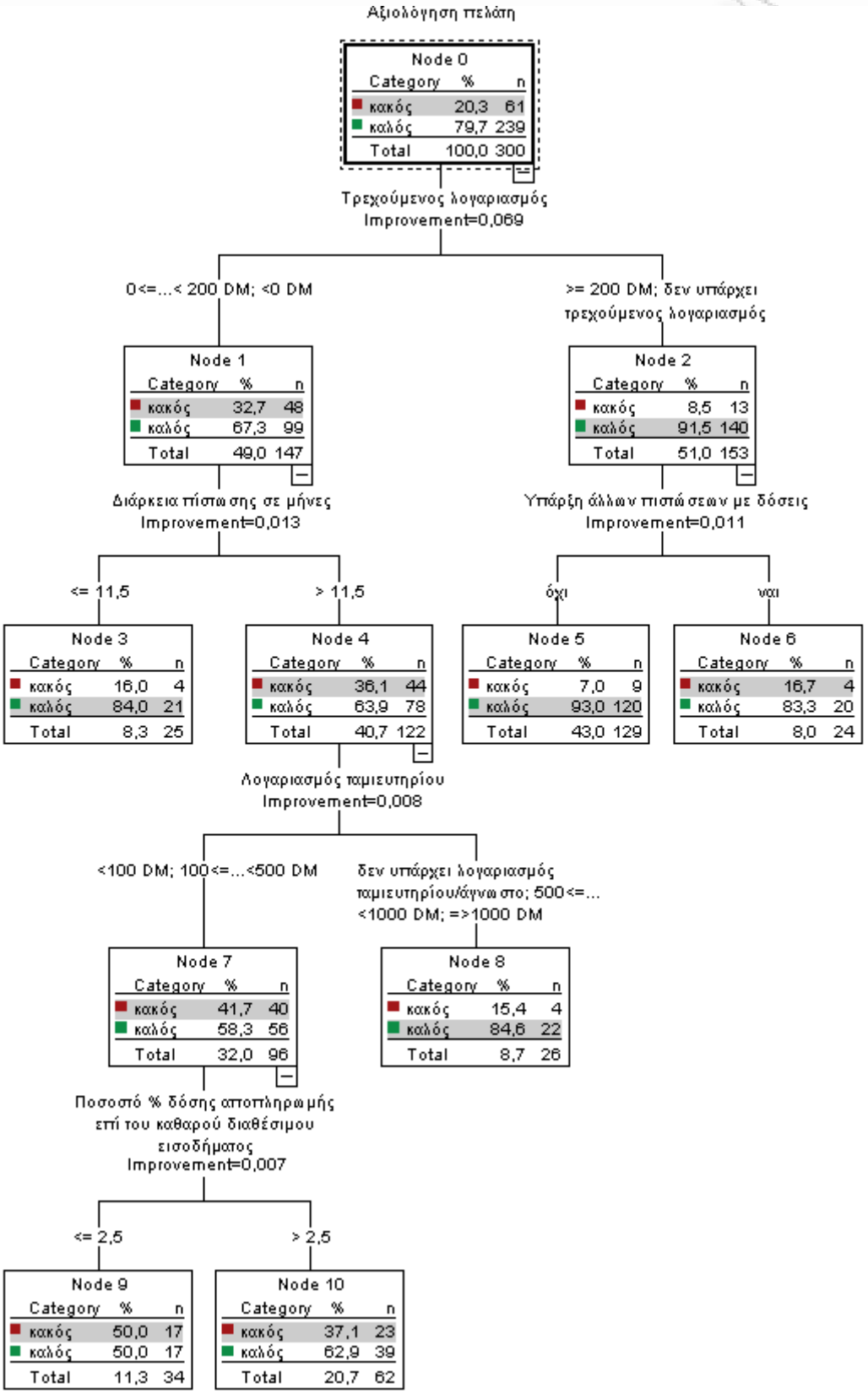
Ο τρόπος με τον οποίο προκύπτει η προβλεπόμενη κατηγορία σε κάθε τελικό κόμβο για κάθε ένα από τα δέντρα ταξινόμησης, εξηγείται παρακάτω στην ανάλυση των Πινάκων 6.32 και 6.33. Σε αυτούς τους πίνακες δίνεται το αναμενόμενο ποσοστό των περιπτώσεων που ανήκουν σε κάθε τελικό κόμβο για κάθε κατηγορία («καλοί» και «κακοί» πελάτες). Η κατηγορία στην οποία ανήκει το μεγαλύτερο αναμενόμενο ποσοστό αποτελεί και την προβλεπόμενη κατηγορία.

Στα δέντρα που απεικονίζονται στα παρακάτω σχηματα, σε κάθε κόμβο δίνεται το ποσοστό των «καλών» και των «κακών» του εξεταζόμενου δείγματος που ανήκουν σε αυτόν τον κόμβο. Αυτά τα ποσοστά, δεν έχουν σχέση με την προβλεπόμενη κατηγορία.

Σχήμα 6.19 *Δέντρο ταξινόμησης για το δείγμα ανάπτυξης*



Σχήμα 6.20 *Δέντρο ταξινόμησης για το δείγμα επικύρωσης*



Σύμφωνα με το διάγραμμα που απεικονίζεται στο Σχήμα 6.19 η καλύτερη επεξηγηματική μεταβλητή για την αξιολόγηση κάθε πελάτη αυτού του χρηματοπιστωτικού οργανισμού η CH_ACCT (τρεχούμενος λογαριασμός). Για τις κατηγορίες των πελατών που ο τρεχούμενος λογαριασμός τους κυμαίνεται από 0 έως 200 γερμανικά μάρκα ή εμφανίζει αρνητικό υπόλοιπο (υπερανάληψη), η αμέσως επόμενη καλύτερη επεξηγηματική μεταβλητή είναι η DURATION (διάρκεια πίστωσης σε μήνες). Για την κατηγορία όπου η διάρκεια πίστωσης σε μήνες είναι μικρότερη ή ίση με 11,5 μήνες, το 75,4% των πελατών της τράπεζας δεν έχουν καταφέρει να εκπληρώσουν τις υποχρεώσεις τους και αξιολογούνται ως «κακοί». Αφού δεν υπάρχει άλλος κόμβος παρακάτω από αυτήν την κατηγορία, αυτός θα αναπαριστά και έναν τελικό κόμβο.

Για την κατηγορία των πελατών που η διάρκεια πίστωσης σε μήνες είναι μεγαλύτερη από 11,5 μήνες η επόμενη σημαντικότερη επεξηγηματική μεταβλητή είναι η SAV_ACCT (λογαριασμός ταμιευτηρίου). Οι πελάτες που ανήκουν σε αυτήν την κατηγορία και έχουν λογαριασμό ταμιευτηρίου με περισσότερα από 500 γερμανικά μάρκα ή δεν έχουν καθόλου λογαριασμό ταμιευτηρίου ταξινομούνται στους «καλούς» και ο κόμβος αυτός αποτελεί και τερματικό κόμβο. Για τους πελάτες που έχουν λιγότερα από 500 γερμανικά μάρκα σε λογαριασμό ταμιευτηρίου η επόμενη πιο σημαντική επεξηγηματική μεταβλητή είναι η INSTALL_RATE (ποσοστό της δόσης αποπληρωμής επί του καθαρού διαθέσιμου εισοδήματος). Παρατηρείται ότι οποιαδήποτε και αν είναι η τιμή αυτού του ποσοστού οι πελάτες ταξινομούνται στους «κακούς» και οι κόμβοι 9 και 10 είναι τερματικοί.

Για τους πελάτες που δεν διαθέτουν τρεχούμενο λογαριασμό ή διαθέτουν πάνω από 200 γερμανικά μάρκα σε αυτόν, η αμέσως επόμενη καλύτερη επεξηγηματική μεταβλητή είναι η OTHER_INSTALL (ύπαρξη άλλων πιστώσεων με διακανονισμό δόσεων). Αυτοί που δεν έχουν σε εκκρεμότητα άλλες πιστώσεις ταξινομούνται ως «καλοί» και αυτοί που έχουν σε εκκρεμότητα άλλες πιστώσεις ταξινομούνται ως «κακοί» παρ' όλο που μόνο το 34,6% των πελατών που ανήκουν σε αυτήν την κατηγορία είναι «κακοί».

Στα παραπάνω σχήματα φαίνεται και η βελτίωση (*improvement*) ή η αύξηση στην καθαρότητα (ομοιογένεια) που επέρχεται από τη διάσπαση κάθε κόμβου. Στον Πίνακα 6.31 εμφανίζεται το δέντρο ταξινόμησης για την αξιολόγηση των πελατών της τράπεζας σε μορφή πίνακα. Αυτός ο πίνακας δίνει τον αριθμό και το ποσοστό των περιπτώσεων που ανήκουν στον κάθε κόμβο για κάθε μία κατηγορία της μεταβλητής απόκρισης. Επιπλέον, δίνεται και η

προβλεπόμενη κατηγορία για τους πελάτες που ανήκουν σε κάθε κόμβο με βάση το δέντρο ταξινόμησης.

Πίνακας 6.31 *Δέντρο ταξινόμησης σε μορφή πίνακα*

		κακός		καλός		Total			
Sample	Node	N	Percent	N	Percent	N	Percent	Predicted Category	Parent Node
Training	0	239	34,1%	461	65,9%	700	100,0%	κακός	
	1	192	48,5%	204	51,5%	396	62,3%	κακός	0
	2	47	15,5%	257	84,5%	304	37,7%	καλός	0
	3	15	24,6%	46	75,4%	61	8,1%	καλός	1
	4	177	52,8%	158	47,2%	335	54,2%	κακός	1
	5	29	11,5%	223	88,5%	252	30,3%	καλός	2
	6	18	34,6%	34	65,4%	52	7,5%	κακός	2
	7	154	57,7%	113	42,3%	267	44,5%	κακός	4
	8	23	33,8%	45	66,2%	68	9,7%	καλός	4
	9	42	43,8%	54	56,2%	96	14,6%	κακός	7
	10	112	65,5%	59	34,5%	171	29,8%	κακός	7
Test	0	61	20,3%	239	79,7%	300	100,0%	κακός	
	1	48	32,7%	99	67,3%	147	60,1%	κακός	0
	2	13	8,5%	140	91,5%	153	39,9%	καλός	0
	3	4	16,0%	21	84,0%	25	7,7%	καλός	1
	4	44	36,1%	78	63,9%	122	52,4%	κακός	1
	5	9	7,0%	120	93,0%	129	32,5%	καλός	2
	6	4	16,7%	20	83,3%	24	7,5%	κακός	2
	7	40	41,7%	56	58,3%	96	44,5%	κακός	4
	8	4	15,4%	22	84,6%	26	7,9%	καλός	4
	9	17	50,0%	17	50,0%	34	17,5%	κακός	7
	10	23	37,1%	39	62,9%	62	27,0%	κακός	7

Στους Πίνακες 6.32 και 6.33 δίνεται το κέρδος για κάθε τερματικό κόμβο για την κατηγορία των «κακών» και των «καλών» πελατών αντίστοιχα.

Πίνακας 6.32 *Πίνακας κέρδους για τους «κακούς» πελάτες*

		Node		Gain			
Sample	Node	N	Percent	N	Percent	Response	Index
Training	10	171	29,8%	112	46,9%	78,5%	157,1%
	9	96	14,6%	42	17,6%	60,0%	120,0%
	6	52	7,5%	18	7,5%	50,5%	101,0%
	8	68	9,7%	23	9,6%	49,6%	99,3%
	3	61	8,1%	15	6,3%	38,6%	77,2%
	5	252	30,3%	29	12,1%	20,1%	40,1%
Test	10	62	27,0%	23	37,7%	69,8%	139,6%
	9	34	17,5%	17	27,9%	79,7%	159,3%
	6	24	7,5%	4	6,6%	43,9%	87,9%
	8	26	7,9%	4	6,6%	41,6%	83,2%
	3	25	7,7%	4	6,6%	42,7%	85,5%
	5	129	32,5%	9	14,8%	22,7%	45,4%

Πίνακας 6.33 *Πίνακας κέρδους για τους «καλούς» πελάτες*

		Node		Gain			
Sample	Node	N	Percent	N	Percent	Response	Index
Training	5	252	30,3%	223	48,4%	79,9%	159,9%
	3	61	8,1%	46	10,0%	61,4%	122,8%
	8	68	9,7%	45	9,8%	50,4%	100,7%
	6	52	7,5%	34	7,4%	49,5%	99,0%
	9	96	14,6%	54	11,7%	40,0%	80,0%
	10	171	29,8%	59	12,8%	21,5%	42,9%
Test	5	129	32,5%	120	50,2%	77,3%	154,6%
	3	25	7,7%	21	8,8%	57,3%	114,5%
	8	26	7,9%	22	9,2%	58,4%	116,8%
	6	24	7,5%	20	8,4%	56,1%	112,1%
	9	34	17,5%	17	7,1%	20,3%	40,7%
	10	62	27,0%	39	16,3%	30,2%	60,4%

Στους παραπάνω πίνακες δίνονται κάποιες πληροφορίες για τους τελικούς κόμβους. Για παράδειγμα, στον Πίνακα 6.33, στην πρώτη στήλη Node δίνεται ο αριθμός και το ποσοστό των περιπτώσεων που εμφανίζονται σε κάθε τερματικό κόμβο για το σύνολο των πελατών. Στη δεύτερη στήλη Gain δίνεται ο αριθμός και το ποσοστό των «καλών» πελατών που ανήκουν στον κάθε τερματικό κόμβο στο σύνολο των «καλών» πελατών του δείγματος. Η στήλη Response δίνει το αναμενόμενο ποσοστό των περιπτώσεων που ανήκουν σε αυτόν τον κόμβο για την κατηγορία στόχο που είναι οι «καλοί» πελάτες (για τον κόμβο 6 του δείγματος ανάπτυξης το αναμενόμενο ποσοστό των «καλών» είναι $\frac{7,4\% \cdot 50\%}{7,5\%} = 49,5\%$, όπου 50% είναι η εκ των προτέρων πιθανότητα των «καλών» πελατών, το 7,4% δίνεται από τη στήλη Gain και το 7,5% δίνεται από τη στήλη Node).

Τέλος, στη στήλη Index δίνεται η αναλογία του ποσοστού των περιπτώσεων που ανήκουν σε αυτόν τον κόμβο για την κατηγορία στόχο προς το ποσοστό των περιπτώσεων που ανήκουν σε αυτόν τον κόμβο για ολόκληρο το δείγμα (για τον κόμβο 6 του δείγματος ανάπτυξης είναι $\frac{7,4\%}{7,5\%} = 99\%$). Μια τιμή αυτής της αναλογίας που είναι μεγαλύτερη από 100% σημαίνει ότι υπάρχουν περισσότερες περιπτώσεις που ανήκουν στην κατηγορία στόχο σε σχέση με το ολικό ποσοστό που ανήκει σε αυτήν την κατηγορία. Αντιθέτως, αν αυτή η αναλογία παίρνει τιμές μικρότερες από 100% σημαίνει ότι υπάρχουν λιγότερες περιπτώσεις που ανήκουν στην κατηγορία στόχο σε σχέση με το ολικό ποσοστό που ανήκει σε αυτήν την

κατηγορία. Στον Πίνακα 6.34 δίνεται ο κίνδυνος που μπορεί να έχει το συγκεκριμένο μοντέλο.

Πίνακας 6.34 *Κίνδυνος*

Sample	Estimate	Std. Error
Training	,300	,018
Test	,298	,032

Ο παραπάνω πίνακας δίνει την πληροφορία πόσο καλά δουλεύει αυτό το μοντέλο για το δείγμα ανάπτυξης και για το δείγμα επικύρωσης. Η εκτίμηση κινδύνου 0,3 για το δείγμα ανάπτυξης και 0,298 για το δείγμα επικύρωσης δείχνει ότι η κατηγορία που προβλέπεται από το μοντέλο είναι λάθος περίπου στο 30% των περιπτώσεων. Επομένως, ο κίνδυνος λάθους ταξινόμησης ενός πελάτη είναι 30%. Στον Πίνακα 6.35 δίνεται ο πίνακας ταξινόμησης στον οποίο φαίνονται τα ποσοστά σωστής ταξινόμησης με βάση το δείγμα ανάπτυξης και το δείγμα επικύρωσης.

Πίνακας 6.35 *Πίνακας ταξινόμησης*

		Predicted		
Sample	Observed	κακός	καλός	Percent Correct
Training	κακός	172	67	72,0%
	καλός	147	314	68,1%
	Overall Percentage	45,6%	54,4%	69,4%
Test	κακός	44	17	72,1%
	καλός	76	163	68,2%
	Overall Percentage	40,0%	60,0%	69,0%

Στο μοντέλο που προέκυψε με τη δόμηση του δέντρου ταξινόμησης παρατηρείται ότι το ποσοστό των πελατών που έγιναν αποδεκτοί και αποδείχτηκαν «καλοί» για το δείγμα ελέγχου είναι 68,2%, ενώ το συνολικό ποσοστό ακρίβειας του τελικού μοντέλου είναι 69%. Δηλαδή οι παρατηρούμενες και οι εκτιμώμενες από το μοντέλο τιμές της μεταβλητής απόκρισης συμφωνούν στο 69% περίπου του συνόλου των παρατηρήσεων για το δείγμα επικύρωσης.

6.6 Σύγκριση των μεθόδων

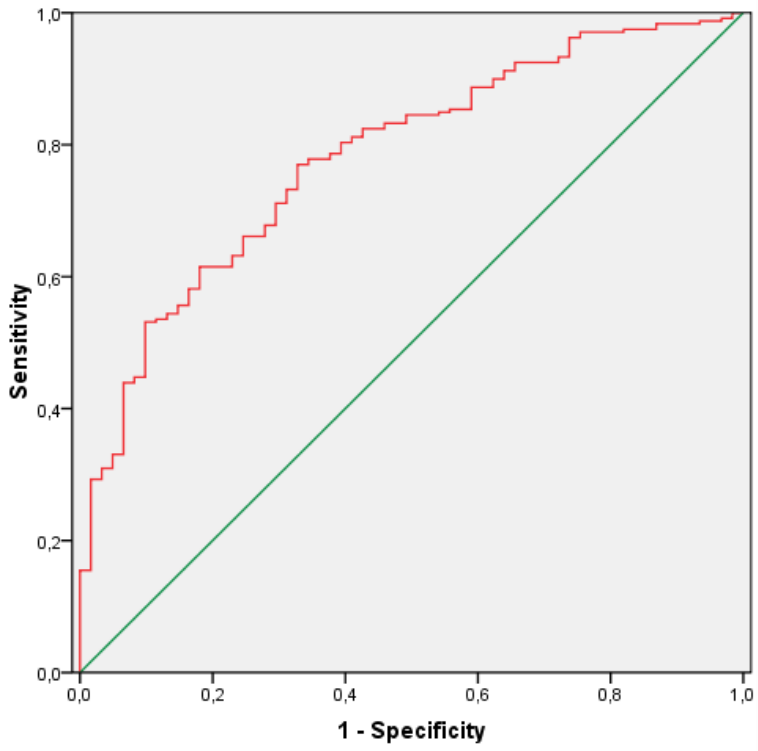
Στο παράδειγμα που επεξεργαστήκαμε στο παρόν κεφάλαιο, για την αξιολόγηση των πελατών της τράπεζας ως προς την πιστοληπτική τους ικανότητα χρησιμοποιήθηκαν μοντέλα

βαθμολόγησης πιστοληπτικής ικανότητας τα οποία αναπτύχθηκαν με τρεις διαφορετικές μεθόδους, τη λογιστική παλινδρόμηση, τη διαχωριστική ανάλυση και τα δέντρα ταξινόμησης. Για να βρεθεί ποια από αυτές τις μεθόδους είναι προτιμότερο να χρησιμοποιηθεί, αρκεί να μετρηθεί η απόδοση κάθε μοντέλου.

Για τη μέτρηση της απόδοσης κάθε μοντέλου ένας τρόπος είναι να συγκριθούν τα ποσοστά σωστής ταξινόμησης για κάθε μοντέλο. Όπως μπορούμε να παρατηρήσουμε από τους Πίνακες ταξινόμησης 6.22, 6.29 και 6.35 για κάθε μέθοδο αντίστοιχα, για τη λογιστική παλινδρόμηση το ποσοστό σωστής ταξινόμησης είναι 76,7%, για τη διαχωριστική ανάλυση είναι 74,3% και για τα δέντρα ταξινόμησης είναι 69%. Το υψηλότερο ποσοστό σωστής ταξινόμησης εμφανίζεται για τη μέθοδο της λογιστικής παλινδρόμησης. Παρ' όλα αυτά το ποσοστό για την διαχωριστική ανάλυση πλησιάζει αρκετά αυτό το ποσοστό.

Ένας άλλος τρόπος για την μέτρηση της απόδοσης ενός CSM είναι ο σχηματισμός της καμπύλης ROC και η μέτρηση του εμβαδού της περιοχής κάτω από αυτήν την καμπύλη. Όσο μεγαλύτερο είναι το εμβαδόν αυτό τόσο καλύτερη θα είναι και η απόδοση του μοντέλου. Στο Σχήμα 6.21 παρατίθεται η καμπύλη ROC που προέκυψε για το μοντέλο της λογιστικής παλινδρόμησης με βάση το δείγμα επικύρωσης.

Σχήμα 6.21 *Καμπύλη ROC για το μοντέλο της Λογιστικής Παλινδρόμησης*



Στον Πίνακα 6.36 δίνεται ότι το εμβαδό της περιοχής κάτω από την καμπύλη ROC (AUC) είναι 0,781 ενώ το 95% διάστημα εμπιστοσύνης για αυτήν την ποσότητα είναι $[0,736,\ 0,853]$.

Πίνακας 6.36 *Εμβαδό κάτω από τη καμπύλη ROC για το μοντέλο της Λογιστικής Παλινδρόμησης*

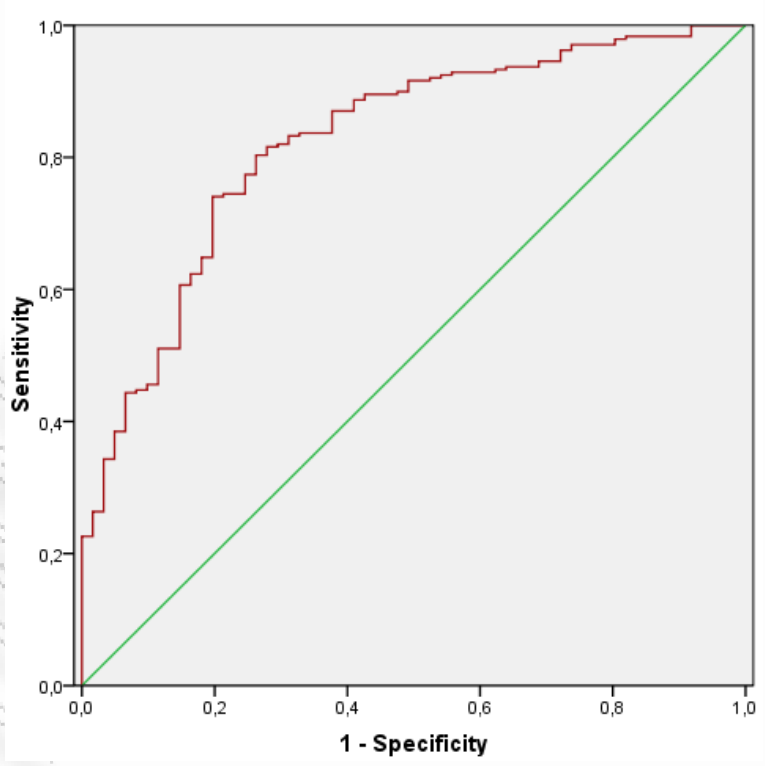
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,781	,031	,000	,721	,842

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Στο Σχήμα 6.22 παρατίθεται η καμπύλη ROC που προέκυψε για το τελικό μοντέλο της διαχωριστικής ανάλυσης με βάση το δείγμα επικύρωσης.

Σχήμα 6.22 *Καμπύλη ROC για το μοντέλο της Διαχωριστικής Ανάλυσης*



Στον Πίνακα 6.37 δίνεται ότι το εμβαδό της περιοχής κάτω από την καμπύλη ROC είναι 0,823 ενώ το 95% διάστημα εμπιστοσύνης για αυτήν την ποσότητα είναι $[0,766,\ 0,881]$.

Πίνακας 6.37 *Εμβαδό κάτω από τη καμπύλη ROC για το μοντέλο της Διαχωριστικής Ανάλυσης*

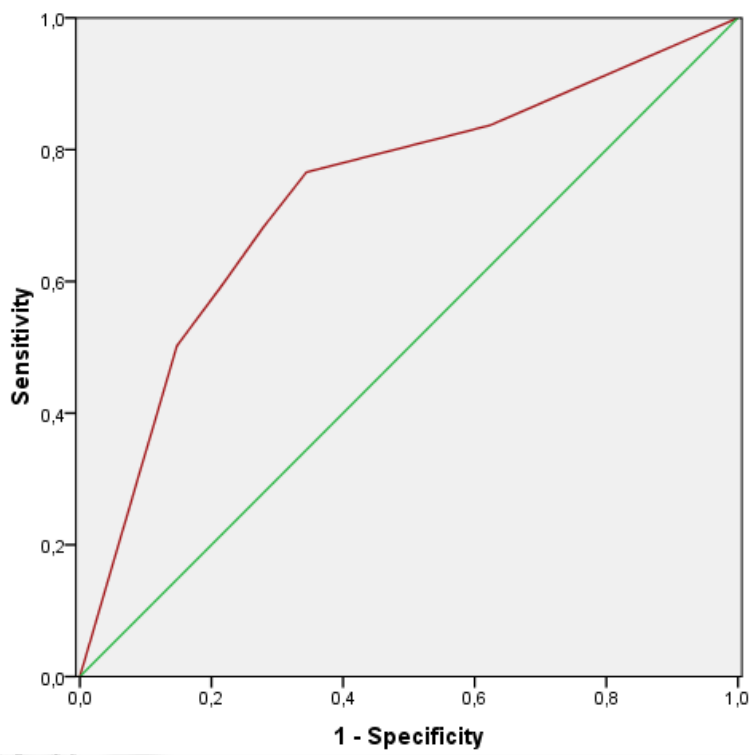
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,823	,029	,000	,766	,881

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Παρακάτω, στο Σχήμα 6.23 παρατίθεται η καμπύλη ROC που προέκυψε για το τελικό μοντέλο του δέντρου ταξινόμησης με βάση το δείγμα επικύρωσης.

Σχήμα 6.23 *Καμπύλη ROC για το Δέντρο Ταξινόμησης*



Αξίζει να σημειωθεί ότι στο παραπάνω σχήμα, κάθε σημείο στο οποίο «σπάει» η τεθλασμένη γραμμή αντιπροσωπεύει το σημείο (1-ειδικότητα, ευαισθησία) που αντιστοιχεί σε κάθε έναν από τους τελικούς κόμβους του δέντρου ταξινόμησης..

Στον Πίνακα 6.38 δίνεται ότι το εμβαδό της περιοχής κάτω από την καμπύλη ROC είναι 0,733 ενώ το 95% διάστημα εμπιστοσύνης για αυτήν την ποσότητα είναι $[0,663,\ 0,800]$.

Πίνακας 6.38 *Εμβαδό κάτω από τη καμπύλη ROC για το Δέντρο Ταξινόμησης*

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,732	,035	,000	,663	,800

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Συγκρίνοντας την ποσότητα AUC για τα τρία είδη μοντέλων βγαίνει το συμπέρασμα ότι καλύτερη απόδοση έχει το μοντέλο της διαχωριστικής ανάλυσης και αμέσως μετά ακολουθεί το μοντέλο της λογιστικής παλινδρόμησης. Όμως, όπως είδαμε στην προηγηθείσα ανάλυση δεν ικανοποιούνται οι προϋποθέσεις της διαχωριστικής ανάλυσης και τα αποτελέσματα δεν είναι και τόσο αξιόπιστα. Επομένως, αυτό το γεγονός σε συνδυασμό με το ότι το μεγαλύτερο ποσοστό σωστής ταξινόμησης δίνεται για τη μέθοδο της λογιστικής παλινδρόμησης μας επιτρέπει να συμπεράνουμε ότι η μέθοδος με την καλύτερη απόδοση και προβλεπτική ικανότητα είναι αυτή της λογιστικής παλινδρόμησης, ενώ αυτή με την χειρότερη απόδοση είναι η δόμηση ενός δέντρου ταξινόμησης.

Σε γενικές γραμμές, και οι τρεις μέθοδοι δεν έχουν απόλυτα ικανοποιητική απόδοση και συγκρίνοντάς τις μεταξύ τους δεν έχουν μεγάλη διαφορά όσον αφορά την προβλεπτική τους ικανότητα. Ένας τρόπος για να βελτιωθεί η απόδοση όλων των παραπάνω μοντέλων είναι να γίνει αλλαγή στο σημείο αποκοπής ή στις εκ των προτέρων πιθανότητες που χρησιμοποιούνται για τη δόμηση τους. Για να βρεθεί το βέλτιστο σημείο αποκοπής, μπορεί να χρησιμοποιηθεί η καμπύλη ROC. Δηλαδή, μπορεί να επιλεχτεί ως σημείο αποκοπής το σημείο της καμπύλης που είναι το πλησιέστερο στην πάνω αριστερή γωνία του σχήματος. Το σημείο αυτό αντιστοιχεί στο μέγιστο άθροισμα ευαισθησίας και ειδικότητας.

ΚΕΦΑΛΑΙΟ 7

Μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας και Βασιλεία ΙΙ

7.1 Προληπτική εποπτεία και κανονιστικό πλαίσιο της Βασιλείας Ι

Ο πιστωτικός κίνδυνος έχει άμεση σχέση με τη φερεγγυότητα των χρηματοπιστωτικών οργανισμών. Τα περισσότερα πιστωτικά ιδρύματα που χρεοκόπησαν αντιμετώπισαν προβλήματα με τους ισολογισμούς τους ή είχαν μεγάλες ζημιές από επισφαλείς απαιτήσεις. Όπως είναι λογικό οι επισφαλείς απαιτήσεις μειώνουν τη κερδοφορία και τα αποθεματικά της τράπεζας, γεγονός που έχει δυσμενείς επιπτώσεις στην πιστοληπτική της ικανότητα και το κόστος δανεισμού της στη διατραπεζική αγορά (βλέπε Ζοπουνίδης – Λεμονάκης (2009)).

Κάθε πιθανή έλλειψη φερεγγυότητας ή ρευστότητας ενός τραπεζικού οργανισμού έχει αρνητικές επιπτώσεις στην εμπιστοσύνη του κοινού και αυτή η καχυποψία επεκτείνεται σε όλο το τραπεζικό σύστημα. Η εξαιρετική σημασία του ελέγχου του πιστωτικού κινδύνου για τη σταθερότητα των τραπεζικών οργανισμών και τη διατήρηση εμπιστοσύνης του κοινού προς τα τραπεζικό σύστημα, οδήγησε στην άσκηση προληπτικής εποπτείας από αρμόδιους φορείς. Η προληπτική τραπεζική εποπτεία στοχεύει στη διασφάλιση της οικονομικής σταθερότητας των τραπεζών και ειδικότερα στην ικανότητά τους να εκπληρώσουν εμπρόθεσμα και στο ακέραιο τις υποχρεώσεις τους προς τους καταθέτες και τους λοιπούς πιστωτές τους. Ακόμη σημαντικότερο, η εποπτεία στοχεύει στην αποτροπή ή τουλάχιστον τον περιορισμό του συστηματικού κινδύνου που δημιουργείται όταν η αδυναμία μιας ή περισσοτέρων τραπεζών κλονίζει τη σταθερότητα του συνόλου του πιστωτικού συστήματος

και την ομαλή λειτουργία των συστημάτων πληρωμών και της οικονομίας γενικότερα (βλέπε Γκέκας (2005)).

Το βασικότερο αντικείμενο ενδιαφέροντος των εποπτικών αρχών είναι η κεφαλαιακή επάρκεια, δηλαδή οι εποπτικές αρχές εστιάζουν σε εκ των προτέρων ενέργειες, υποχρεώνοντας τις τράπεζες σε παρακράτηση κεφαλαίων αντίστοιχη με τους κινδύνους που αντιμετωπίζουν, προκειμένου να περιορίσουν τον κίνδυνο αθέτησης υποχρεώσεων. Με τον όρο «εποπτικά κεφάλαια»[17], εννοούμε τα κεφάλαια τα οποία η εποπτική αρχή θεωρεί ότι η τράπεζα πρέπει να «τοποθετήσει στην άκρη» έτσι ώστε να καλύπτεται απέναντι στους διάφορους κινδύνους που αναλαμβάνει. Μέσω της ύπαρξης ενός διεθνούς εποπτικού οργανισμού, μπορεί να επιτευχθεί η βέλτιστη λύση για την παρακράτηση οικονομικού κεφαλαίου που απαιτείται από μεριάς των πιστωτικών ιδρυμάτων για την αντιμετώπιση των διαφόρων κινδύνων του χρηματοπιστωτικού συστήματος. Το ρόλο αυτής της διεθνούς επιτροπής τον έλαβε η Επιτροπή της Βασιλείας (*Basel Committee*).

Η Τράπεζα Διεθνών Διακανονισμών (*Bank of International Settlements* – BIS) ιδρύθηκε το 1930 στην πόλη Βασιλεία της Ελβετίας με σκοπό την επιτήρηση των πληρωμών για την αποκατάσταση των ζημιών που προκλήθηκαν από τον πόλεμο, αλλά γρήγορα αποτέλεσε το μέσο για τη συνεργασία μεταξύ των Κεντρικών Τραπεζών με σκοπό τη νομισματική και οικονομική σταθερότητα. Τον Σεπτέμβρη του 1974, έπειτα από την κατάρρευση μεγάλων διεθνών τραπεζικών οργανισμών, όπως η γερμανική τράπεζα Herstatt, και η αμερικανική Franklin National, το ενδιαφέρον των οικονομικών αρχών στράφηκε στην αλληλεξάρτηση των επιμέρους εθνικών τραπεζικών συστημάτων. Αυτό οδήγησε στη δημιουργία της Επιτροπής της Βασιλείας Ι για την Τραπεζική Εποπτεία (*Basel I Committee on Banking Supervision*) υπό την αιγίδα της Τράπεζας Διεθνών Διακανονισμών για να καθιερώσει αποδεκτούς κανόνες ελέγχου των κινδύνων των τραπεζικών χαρτοφυλακίων χορηγήσεων. Η επιτροπή αυτή δεν επέβαλλε αυτούς τους κανόνες αλλά συνιστούσε στα μέλη που την αποτελούσαν[18] να εφαρμόσουν αυτές τις αρχές με τον πιο κατάλληλο τρόπο για αυτά. Σήμερα, η Επιτροπή της Βασιλείας δεν έχει νομική προσωπικότητα, ούτε μορφή υπερεθνικής εποπτικής αρχής και εδρεύει στην Τράπεζα Διεθνών Διακανονισμών της Βασιλείας της Ελβετίας η οποία και της παρέχει γραμματειακή υποστήριξη. Επομένως, οι προτάσεις της Βασιλείας δεν έχουν νομική δεσμευτική ισχύ και οι εποπτικές αρχές κάθε χώρας έχουν

[17] Τα εποπτικά κεφάλαια προκύπτουν από τα λογιστικά Ίδια Κεφάλαια με κάποιες προσαρμογές.

[18] Στη σύνθεση της Επιτροπής της Βασιλείας συμμετέχουν 13 χώρες: Βέλγιο, Καναδάς, Γαλλία, Γερμανία, Ιταλία, Ιαπωνία, Λουξεμβούργο, Ολλανδία, Ισπανία, Σουηδία, Ελβετία, Ηνωμένο Βασίλειο, ΗΠΑ.

ευθύνη για την ενσωμάτωση και την εφαρμογή των κανόνων και των πρακτικών σε εθνικό επίπεδο (βλέπε Bryan (2008)). Στην χώρα μας, η Τράπεζα της Ελλάδος είναι αρμόδια για την εποπτεία των πιστωτικών ιδρυμάτων σε ατομική και ενοποιημένη βάση.

Το ισχύον σύστημα κανόνων της Επιτροπής της Βασιλείας αναφορικά με την κεφαλαιακή επάρκεια των διεθνών τραπεζών διαμορφώθηκε σταδιακά από τον Ιούλιο του 1988, όταν δημοσιεύτηκε το Σύμφωνο της Βασιλείας για την Κεφαλαιακή Επάρκεια με τίτλο «Διεθνής Σύγκληση της Κεφαλαιακής Μέτρησης και των Κεφαλαιακών Προτύπων» (*International Convergence of Capital Measurement and Capital Standards*). Το κείμενο αυτό, το οποίο αφορούσε τον πιστωτικό κίνδυνο, τροποποιήθηκε και συμπληρώθηκε πολλές φορές στο παρελθόν με σημαντικότερη τροποποίηση εκείνη του 1996 προκειμένου να συμπεριληφθούν και οι κίνδυνοι αγοράς.

Η πρώτη αυτή συμφωνία για την κεφαλαιακή επάρκεια εστίασε στον πιστωτικό κίνδυνο. Ο δείκτης κεφαλαιακής επάρκειας (*ΔΚΕ*) συγκρίνει τα Ίδια Κεφάλαια της Εταιρίας με το σταθμισμένο ενεργητικό έναντι των κινδύνων αγοράς, του πιστωτικού κινδύνου και του λειτουργικού κινδύνου και υπολογίζεται από τον ακόλουθο τύπο:

$$\Delta KE = \frac{EK}{\Pi K + KA + \Lambda K},$$

όπου *ΕΚ* είναι τα Εποπτικά Κεφάλαια, *ΠΚ* είναι ο Πιστωτικός Κίνδυνος, *ΚΑ* είναι ο Κίνδυνος της Αγοράς και *ΛΚ* είναι ο Λειτουργικός Κίνδυνος. Στόχος της συμφωνίας της Βασιλείας Ι ήταν να πετύχει ένα ελάχιστο ύψος κεφαλαιακών απαιτήσεων της τάξεως του 8% (4% για τα στεγαστικά δάνεια), ώστε να επιτευχθεί η προστασία των πιστωτικών ιδρυμάτων από τον αναλαμβανόμενο πιστωτικό κίνδυνο. Η βασική ιδέα ήταν ότι τα πιο επικίνδυνα στοιχεία του ενεργητικού έπρεπε να «απασχολούν» μεγαλύτερο μέρος από τα κεφάλαια της τράπεζας. Ωστόσο, σύμφωνα με την Βασιλεία Ι, το σύνολο του ενεργητικού των τραπεζών κατηγοριοποιούνταν μόνο σε τέσσερις κατηγορίες κινδύνου (σταθμισμένο ενεργητικό *Risk-Weighted Assets* - RWA)[19], οι οποίες ήταν:

- Μηδενική στάθμιση σε περιουσιακά στοιχεία μηδενικού κινδύνου όπως τα κρατικά ομόλογα.
- στάθμιση 20% σε δάνεια με χαμηλό επίπεδο κινδύνου όπως ο διατραπεζικός δανεισμός.

[19] Συμφώνα με τους Ζουπουνίδης – Λεμονάκης (2009) σταθμισμένο ενεργητικό καλείται το ενεργητικό που είναι σταθμισμένο ανάλογα με το δείκτη επικινδυνότητας (πιθανότητα αθέτησης υποχρεώσεων) κάθε στοιχείου του ενεργητικού του χρηματοπιστωτικού οργανισμού.

- στάθμιση 50% για στεγαστικά δάνεια με υποθήκη πάνω σε περιουσιακά στοιχεία.
- στάθμιση 100% για καταναλωτικά δάνεια, συμπεριλαμβανομένων και των δανείων ανακυκλούμενης πίστης όπως είναι οι πιστωτικές κάρτες.

Έτσι, σύμφωνα με το πλαίσιο κεφαλαιακής επάρκειας της Βασιλείας Ι, οι σταθμίσεις δεν αποτελούσαν αντιπροσωπευτικά μεγέθη έναντι των κινδύνων που αντιστοιχούσαν στα διάφορα τραπεζικά προϊόντα. Με αυτό τον τρόπο, μπορεί το μεγάλο πλεονέκτημα της Βασιλείας Ι να ήταν η απλότητα ως προς την εφαρμογή της, όμως κατακρίθηκε για την απλότητα της στην κατηγοριοποίηση των διαφόρων στοιχείων του ενεργητικού των τραπεζών. Η αποτυχία διάκρισης καταναλωτικών δανείων διαφορετικού βαθμού κινδύνου δημιούργησε το κίνητρο να απομακρυνθούν τα δάνεια με το μικρότερο κίνδυνο και να διατηρηθούν τα δάνεια που ενέχουν μεγαλύτερο κίνδυνο. Αυτή η αυθαιρεσία υπονόμευσε την αποτελεσματικότητα της συμφωνίας της Βασιλείας Ι και έτσι αναπτύχθηκε μια νέα συμφωνία της Βασιλείας κατά τη διάρκεια των πέντε πρώτων ετών αυτής της χιλιετίας μέσω μιας σειράς συμβουλευτικών εγγράφων και ερευνών.

7.2 Νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ

Με τις εξελίξεις και τις νέες προκλήσεις στον τραπεζικό τομέα το σύμφωνο της Βασιλείας Ι σταμάτησε να ανταποκρίνεται αποτελεσματικά στους κινδύνους στους οποίους εκτίθενται οι τράπεζες. Έτσι, η Επιτροπή της Βασιλείας εξέδωσε στις 26 Ιουνίου 2004 το νέο Σύμφωνο για την Κεφαλαιακή επάρκεια. Οι οδηγίες του νέου συμφώνου ξεκίνησαν να εφαρμόζονται διεθνώς το 2007, ενώ η πρώτη εφαρμογή του νέου πλαισίου στην Ελλάδα έγινε τον Ιανουάριο του 2008.

Αυτή η δεύτερη συμφωνία διαφέρει από την πρώτη κυρίως στο γεγονός ότι διαχωρίζει τους κινδύνους που περιλαμβάνονται στη δανειοδότηση στον κίνδυνο αγοράς, στον πιστωτικό κίνδυνο, και στο λειτουργικό κίνδυνο[20]. Η δομή του νέου συμφώνου της Επιτροπής της Βασιλείας ΙΙ περιλαμβάνει αλληλένδετες και συμπληρωματικές θεματικές ενότητες που ονομάζονται πυλώνες, οι οποίες συμβάλλουν στην ασφάλεια και σταθερότητα του

[20] Ο κίνδυνος αγοράς είναι ο κίνδυνος μεταβολών της αξίας του χαρτοφυλακίου λόγω της μεταβλητότητας των τιμών της αγοράς ενώ ο λειτουργικός κίνδυνος αναφέρεται στην πιθανότητα ζημιών λόγω προβλημάτων στις εσωτερικές διαδικασίες, στα συστήματα, στο ανθρώπινο δυναμικό και σε εξωτερικά γεγονότα.

χρηματοπιστωτικού συστήματος αφού επιτρέπουν στις τράπεζες και στις εποπτικές αρχές να αξιολογούν σωστά τους κινδύνους. Στον Πίνακα 7.1 δίνεται η περιγραφή των τριών πυλώνων της Βασιλείας ΙΙ.

Πίνακας 7.1 *Δομή Βασιλείας ΙΙ*

Πυλώνας Ι	**Πυλώνας ΙΙ**	**Πυλώνας ΙΙΙ**
Ελάχιστες κεφαλαιακές απαιτήσεις (*Minimum Capital Requirements* – MCR). Ποσοτικοποίηση και υπολογισμός των κινδύνων και των ελάχιστων κεφαλαιακών απαιτήσεων έναντι του πιστωτικού κινδύνου με την προσθήκη απαιτήσεων για την κάλυψη και του λειτουργικού κινδύνου.	Εποπτικές διαδικασίες για τον έλεγχο της κεφαλαιακής επάρκειας των χρηματοπιστωτικών ιδρυμάτων, σε μόνιμη βάση.	Δημοσιοποίηση πληροφοριών σχετικά με τους αναλαμβανόμενους κινδύνους και την κεφαλαιακή επάρκεια με σκοπό την ενδυνάμωση της πειθαρχίας που επιβάλει η αγορά στις τράπεζες.

Για τον πιστωτικό κίνδυνο, οι κανονισμοί χωρίζουν το χαρτοφυλάκιο δανεισμού σε πέντε κατηγορίες – εταιρικό, λιανικό, δανεισμού σε ξένο νόμισμα, διατραπεζικών συναλλαγών και κοινών μετοχών μεταβλητού επιτοκίου. Η Επιτροπή της Βασιλείας ΙΙ προτείνει ουσιαστικά δύο εναλλακτικές μεθόδους υπολογισμού των εποπτικών κεφαλαίων για τον υπολογισμό κεφαλαιακών απαιτήσεων. Η πρώτη είναι η **τυποποιημένη μέθοδος (*Standardized Approach*)**[21], σύμφωνα με την οποία οι σταθμίσεις κινδύνου των στοιχείων του ενεργητικού μιας τράπεζας παραμένουν προκαθορισμένες αλλά λίγο πιο αναλυτικές σε σχέση με την Βασιλεία Ι. Η δεύτερη μέθοδος είναι η **μέθοδος εσωτερικών διαβαθμίσεων (*Internal Ratings Based* - IRB)**. Σύμφωνα με αυτήν την προσέγγιση, κάθε δανειακό χαρτοφυλάκιο ομαδοποιείται σε τμήματα και για κάθε ένα από αυτά κατασκευάζονται μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας από τους ίδιους τους οργανισμούς έτσι ώστε να δοθούν οι εκτιμήσεις των παραμέτρων αθέτησης υποχρεώσεων για κάθε τμήμα. Με αυτόν τον τρόπο οι τράπεζες αποφεύγουν την περίπτωση ακραίων τιμών σε κάθε κατηγορία των

[21] Για περισσότερες πληροφορίες σχετικά με την τυποποιημένη μέθοδο βλέπε έγγραφο Basel Committee on Banking Supervision (BCBS) (2006).

στοιχείων των χαρτοφυλακίων τους. Οι παράμετροι αυτές τροφοδοτούν τις συναρτήσεις στάθμισης κινδύνου, δηλαδή το μαθηματικό πλαίσιο που παρέχει η Βασιλεία ΙΙ και παράγουν τις σταθμίσεις κινδύνου για κάθε στοιχείο του ενεργητικού της τράπεζας, έτσι ώστε να υπολογιστεί το εποπτικό κεφάλαιο (ή ελάχιστη κεφαλαιακή απαίτηση) που απαιτείται για κάθε τμήμα του δανειακού χαρτοφυλακίου του οργανισμού. Το κεφάλαιο που ο χρηματοπιστωτικός οργανισμός θα πρέπει να «τοποθετήσει στην άκρη» είναι το άθροισμα του κεφαλαίου για κάθε τμήμα του δανειακού χαρτοφυλακίου.

Η μέθοδος εσωτερικών διαβαθμίσεων παρέχει στις τράπεζες δύο εναλλακτικούς τρόπους υπολογισμού των εποπτικών κεφαλαίων, ανάλογα με το βαθμό εξέλιξης των εσωτερικών τους συστημάτων διαβάθμισης κινδύνου, τη θεμελιώδη μέθοδο (*Foundation Approach*) και την προηγμένη μέθοδο (*Advanced Approach*). Και στις δύο μεθόδους, τα χρηματοπιστωτικά ιδρύματα είναι υποχρεωμένα να υπολογίζουν την πιθανότητα αθέτησης υποχρεώσεων *PD* των δανειακών χαρτοφυλακίων τους. Ωστόσο, τα χρηματοπιστωτικά ιδρύματα που χρησιμοποιούν την εξελιγμένη προσέγγιση, είναι υποχρεωμένα να προβαίνουν παράλληλα στον υπολογισμό κάποιων παραμέτρων κινδύνου όπως η **ζημιά δεδομένης της αθέτησης (*Loss Given Default - LGD*)** και το **χρηματοδοτικό άνοιγμα (*Exposure At Default - EAD*)**. Η ζημιά δεδομένης της αθέτησης *LGD* είναι η χρηματική ζημιά μετά την αθέτηση εκπλήρωσης υποχρεώσεων, δηλαδή η ζημιά που θα υποστεί η τράπεζα σε περίπτωση αθέτησης. Το μέγεθος εκφράζεται σε όρους ποσοστού επί του συνολικού ύψους του δανεισμού του χαρτοφυλακίου (χρηματικές μονάδες ζημιάς έναντι χρηματικών μονάδων συνολικού ανοίγματος κατά την αθέτηση). Το χρηματοδοτικό άνοιγμα *EAD* είναι το χρηματικό ποσό ανοίγματος που εκτίθεται σε πιστωτικό κίνδυνο την συγκεκριμένη χρονική στιγμή της αθέτησης.

Έχοντας εκτιμήσει τις τρεις παραπάνω παραμέτρους, τα χρηματοπιστωτικά ιδρύματα προβαίνουν στη διενέργεια προβλέψεων για την κάλυψη των αναμενόμενων ζημιών από αθέτηση υποχρεώσεων των οφειλετών και παρακρατούν το αντίστοιχο οικονομικό κεφάλαιο που αντικατοπτρίζει τις εν λόγω προβλέψεις και το μέγεθος του πιστωτικού κινδύνου που αναλαμβάνουν. Σύμφωνα με την Basel Committee on Banking Supervision (2006), η αναμενόμενη **ζημιά (*Expected Losses – EL*)** εκτιμάται, αγνοώντας την επίδραση της χρονικής διάρκειας των δανείων, από τον τύπο:

$$EL = PD \times LGD \times EAD \tag{7.1}$$

Ο παραπάνω τύπος υπολογισμού της αναμενόμενης ζημιάς εξαρτάται από την εκτίμηση των παραμέτρων *PD* και *LGD* σε τακτά χρονικά διαστήματα. Όμως, οι οικονομικές και κοινωνικές συνθήκες αλλάζουν διαρκώς και οι οφειλέτες που δραστηριοποιούνται στο οικονομικό και κοινωνικό περιβάλλον επηρεάζονται σε μεγάλο βαθμό από αυτές τις αλλαγές. Αυτό σημαίνει ότι οι τιμές των παραμέτρων *PD* και *LGD* για κάθε στοιχείο του ενεργητικού των πιστωτικών ιδρυμάτων αλλάζουν διαρκώς.

Η επιλογή των κατάλληλων τμημάτων του χαρτοφυλακίου είναι πολύ σημαντική. Η επιτροπή της Βασιλείας ΙΙ θεωρεί ότι η κατάλληλη ομαδοποίηση του λιανικού χαρτοφυλακίου πρέπει να γίνει αρχικά σε τρεις κατηγορίες (στεγαστικά δάνεια, ανακυκλούμενη πίστη και τα υπόλοιπα στην τρίτη κατηγορία). Σύμφωνα με τις οδηγίες, για κάθε μία από αυτές τις τρεις κατηγορίες, πρέπει να γίνουν επιπλέον διαχωρισμοί για τα διαφορετικά είδη δανείων και παράλληλα να δημιουργηθεί ένα ξεχωριστό τμήμα για τα μη αποπληρωμένα δάνεια. Τέλος, από κάποιους οργανισμούς συνηθίζεται τα ίδια είδη δανείων να διαχωρίζονται επιπλέον σε τμήματα ανάλογα με την κατηγορία της αθέτησης υποχρεώσεων στην οποία ανήκουν.

Στο κανονιστικό πλαίσιο της Βασιλείας ΙΙ, ο τύπος που προσδιορίζει τις ελάχιστες κεφαλαιακές απαιτήσεις ενός πιστωτικού ιδρύματος βασίζεται στην εκτίμηση της κατανομής των ζημιών. Μια τράπεζα, ενώ δεν είναι δυνατόν να γνωρίζει εκ των προτέρων τις ζημιές που ενδέχεται να υποστεί σε ένα συγκεκριμένο έτος, είναι εφικτό να προβλέψει το μέσο επίπεδο των ζημιών από τα δάνεια που έχει χορηγήσει. Έπειτα οι κεφαλαιακές απαιτήσεις ρυθμίζονται έτσι ώστε η πιθανότητα η ζημιά να είναι μεγαλύτερη από το εποπτικό κεφάλαιο να διατηρείται σε ένα σταθερό προκαθορισμένο επίπεδο. Οι ζημιές οι οποίες υπερβαίνουν αυτό το επίπεδο ονομάζονται μη αναμενόμενες ζημιές (*unexpected losses – UL*). Αν και τα πιστωτικά ιδρύματα γνωρίζουν ότι αυτές οι ζημιές μπορούν να συμβούν, τόσο στο παρόν όσο και στο μέλλον, εν τούτοις δεν είναι δυνατόν να προβλέψουν εκ των προτέρων το χρόνο που θα προκύψουν και τη σοβαρότητά τους.

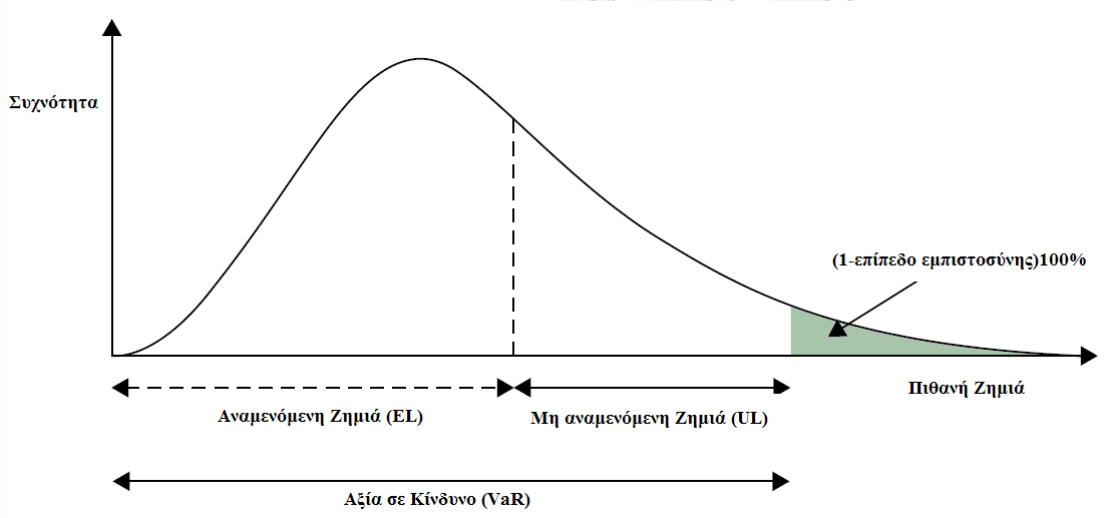
Στο Σχήμα 7.1 η γραμμοσκιασμένη περιοχή αντιπροσωπεύει την πιθανότητα η ζημιά να ξεπεράσει τις κεφαλαιακές απαιτήσεις, το ποσό των οποίων έχει επιλεχτεί κατάλληλα έτσι ώστε αυτή η πιθανότητα να έχει μια προκαθορισμένη τιμή. Αυτή η προσέγγιση είναι η ίδια με εκείνη που χρησιμοποιείται για τον υπολογισμό της Αξίας σε Κίνδυνο (*Value at Risk – VaR*) που προσδιορίζει το οικονομικό κεφάλαιο στον επιχειρησιακό δανεισμό. Σε αυτήν την περίπτωση, η πιθανότητα η ζημιά να μην ξεπερνάει το κατώτατο όριο του κεφαλαίου

καλείται επίπεδο εμπιστοσύνης και το κατώτατο όριο αποτελεί την ποσότητα *VaR* σε αυτό το επίπεδο εμπιστοσύνης. Κατά συνέπεια αν γίνει αποδεκτό ότι η ζημιά θα ξεπεράσει τις κεφαλαιακές απαιτήσεις με πιθανότητα p, τότε οι ελάχιστες κεφαλαιακές απαιτήσεις (*regulatory capital*) θα ισούνται με τη διαφορά μεταξύ της ποσότητας *VaR* σε ένα επίπεδο εμπιστοσύνης $1-p$ και της αναμενόμενης ζημιάς $E[L_{ex}]$.

$$RC_p = VaR_{1-p} - E[L_{ex}].$$

Γενικότερα, η καμπύλη του παρακάτω σχήματος δείχνει ότι μικρές σχετικά ζημιές με τιμές κοντά στην *EL*, προκύπτουν συχνότερα απ' ότι μεγάλες ζημιές.

Σχήμα 7.1 *Κατανομή Ζημιών, Αναμενόμενη Ζημιά, Μη αναμενόμενη Ζημιά, Αξία σε Κίνδυνο. (Πηγή: Thomas (2009))*



Είναι σαφές ότι η κατανομή ζημιών ενός χαρτοφυλακίου εξαρτάται από τις συσχετίσεις μεταξύ των πιθανοτήτων αθέτησης υποχρεώσεων των διαφορετικών δανείων που το αποτελούν. Η ακριβής μοντελοποίηση αυτών των συσχετίσεων είναι εξαιρετικά δύσκολη και ακόμα πιο δύσκολο είναι να επικυρωθεί ένα τέτοιο μοντέλο. Επιπλέον, η χρήση ενός τέτοιου μοντέλου, με τις άμεσες εξαρτήσεις μεταξύ των χαρακτηριστικών των μεμονωμένων δανείων, στην απόφαση χορήγησης ή όχι του δανείου θα ήταν ανεπιτυχής γιατί η απόφαση θα εξαρτιόταν από τα άλλα δάνεια του χαρτοφυλακίου, το οποίο δεν είναι λογικό. Επομένως, η επιτροπή της Βασιλείας αποφάσισε ότι οποιοδήποτε μοντέλο χρησιμοποιείται για τον υπολογισμό του RC_P θα πρέπει να εμφανίζει την ιδιότητα της ευστάθειας του χαρτοφυλακίου (*portfolio invariance*). Αυτό σημαίνει ότι το απαιτούμενο κεφάλαιο για κάθε δάνειο θα

εξαρτάται μόνο από τους κινδύνους που περιλαμβάνονται σε αυτό το δάνειο και όχι από το χαρτοφυλάκιο στο οποίο ανήκει.

Η ευστάθεια χαρτοφυλακίου είναι μια απαραίτητη προϋπόθεση ώστε να παρέχεται ένα εφαρμόσιμο ρυθμιστικό μοντέλο αλλά είναι ταυτόχρονα και μια περιοριστική συνθήκη για τα μοντέλα που ενδέχεται να χρησιμοποιηθούν. Η πρώτη προϋπόθεση που πρέπει να ισχύει είναι ότι το χαρτοφυλάκιο πρέπει να αποτελείται από έναν άπειρο αριθμό δανείων με μικρά ποσά το καθένα. Αυτό συμβαίνει γιατί με αυτόν τον τρόπο εξασφαλίζεται ακριβώς το επίπεδο εμπιστοσύνης *p*. Εάν μπορεί να γίνει αποδεκτό ένα επίπεδο εμπιστοσύνης περίπου *p*, το μόνο που απαιτείται είναι το χαρτοφυλάκιο να περιέχει μεγάλο αριθμό δανείων από τα οποία κανένα να μην καταλαμβάνει μεγάλο μέρος του χαρτοφυλακίου. Τα χαρτοφυλάκια των καταναλωτικών δανείων τείνουν να έχουν αυτήν την ιδιότητα.

Η δεύτερη προϋπόθεση για να ισχύει η ευστάθεια χαρτοφυλακίου είναι ότι μπορεί να υπάρξει μόνο ένας παράγοντας κινδύνου που επηρεάζει όλα τα δάνεια. Αυτό σημαίνει ότι δεν μπορούν να μοντελοποιηθούν όλες οι αλληλεπιδράσεις μεταξύ των διαφορετικών δανείων του χαρτοφυλακίου αλλά ότι οποιαδήποτε αλληλεπίδραση μεταξύ των πιθανοτήτων αθέτησης υποχρεώσεων που αφορούν τα δάνεια γίνεται μέσω αυτού του παράγοντα, ο οποίος περιγράφει κατά κάποιο τρόπο την κατάσταση της παγκόσμιας οικονομίας. Για τα χαρτοφυλάκια καταναλωτικής πίστης που περιλαμβάνουν τον ίδιο τύπο δανείων, στην ίδια χώρα η υπόθεση αυτή μπορεί να είναι αποδεκτή αλλά δεν ισχύει για τα περισσότερα χαρτοφυλάκια επιχειρηματικής πίστης γιατί σε αυτήν την περίπτωση συνήθως περιλαμβάνεται και ο δανεισμός εταιρειών διαφορετικών χωρών.

Όπως αναφέρθηκε παραπάνω, ευστάθεια χαρτοφυλακίου σημαίνει ότι στο μοντέλο που χρησιμοποιείται για να προσδιοριστούν οι κεφαλαιακές απαιτήσεις επιτρέπεται να υπάρχει μόνο ένας οικονομικός παράγοντας. Αυτός ο παράγοντας μπορεί να χρησιμοποιηθεί για να συνδέσει τα δύο είδη πιθανοτήτων αθέτησης υποχρεώσεων, τη μακροπρόθεσμη μέση *PD* και την *PD* που εμφανίζεται σε μια περίοδο οικονομικής ύφεσης. Η μακροπρόθεσμη μέση *PD* είναι η αναμενόμενη πιθανότητα αθέτησης υποχρεώσεων κάτω από συνηθισμένες επιχειρησιακές και οικονομικές συνθήκες και εκτιμάται με βάση τις ιστορικές αθετήσεις υποχρέωσης που έχουν συμβεί τα προηγούμενα έτη. Όμως οι ελάχιστες κεφαλαιακές απαιτήσεις πρέπει να υπολογιστούν με σκοπό να καλύπτουν τις ζημιές όταν αυτός ο οικονομικός παράγοντας παίρνει πολύ συντηρητικές τιμές σε συνθήκες ύφεσης της οικονομίας. Επομένως ο τύπος που προτείνει η Βασιλεία ΙΙ απεικονίζει τη μακροπρόθεσμη

μέση *PD* σε μία συντηρητική τιμή της *PD* η οποία συμβολίζεται με $K(PD)$ και αντιστοιχεί στην πιθανότητα αθέτησης υποχρεώσεων όταν οι οικονομικές συνθήκες βρίσκονται σε κακή κατάσταση.

7.3 Ο τύπος της Βασιλείας ΙΙ για την Καταναλωτική Πίστη

Το νέο ρυθμιστικό πλαίσιο της Βασιλείας ΙΙ επηρεάζει τον τρόπο με τον οποίο δομούνται τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας και τις επιχειρηματικές αποφάσεις που αυτά υποστηρίζουν για τα λιανικά δανειακά χαρτοφυλάκια. Είναι γνωστό ότι, χρησιμοποιώντας CSM, προκύπτει μια τελική βαθμολογία αθέτησης υποχρεώσεων η οποία αποτελεί ένα τέλειο εργαλείο αρχικά για τη μέτρηση και ακολούθως για τον έλεγχο του πιστωτικού κινδύνου μέσω των επιχειρηματικών αποφάσεων στα λιανικά χαρτοφυλάκια. Όμως, σύμφωνα με τη Βασιλεία ΙΙ, οι επιχειρηματικές αποφάσεις και η πιστωτική πολιτική επιδρούν στον τρόπο με τον οποίο δομούνται τα CSM. Χωρίς τα CSM δεν θα μπορούσε να χρησιμοποιηθεί η προσέγγιση εσωτερικών διαβαθμίσεων που προτείνεται από τους κανονισμούς της Βασιλείας ΙΙ για τα χαρτοφυλάκια της καταναλωτικής πίστης. Η μέθοδος των εσωτερικών διαβαθμίσεων στην καταναλωτική πίστη βασιζόμενη στα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας είναι αυτή που καθιστά ελκυστικούς τους κανόνες της Βασιλείας ΙΙ στους δανειστές.

Αφού εκτιμηθεί η πιθανότητα αθέτησης υποχρεώσεων υπό συνθήκες οικονομικής ύφεσης, μπορεί να προσδιοριστούν οι κεφαλαιακές απαιτήσεις για να καλυφθεί η αναμενόμενη ζημιά των δανείων σε ένα τμήμα του χαρτοφυλακίου. Το χρηματικό ποσό που βρίσκεται σε κίνδυνο, η ποσότητα *EAD* πολλαπλασιάζεται με την ποσότητα *PD* υπό συνθήκες οικονομικής ύφεσης και το γινόμενο αυτό ισούται με το ποσοστό του χρηματοδοτικού ανοίγματος που εκτίθεται σε κίνδυνο (λόγω του πολύ μεγάλου αριθμού των μικρών δανείων θεωρείται ότι το χρηματοδοτικό άνοιγμα για κάθε δάνειο είναι 1 μονάδα). Αυτό το γινόμενο δίνει το ύψος χρηματικού ποσού που αναμένεται να χαθεί. Όμως δεν υφίσταται απώλεια ολόκληρου αυτού του χρηματικού ποσού διότι το πιστωτικό ίδρυμα μπορεί να ανακτήσει το μεγαλύτερο ποσοστό αυτού με εξειδικευμένη τιμολόγηση των διαφόρων χρηματοδοτικών ανοιγμάτων, με τη διενέργεια προβλέψεων ή με τη διαγραφή απαιτήσεων. Για παράδειγμα, σε περίπτωση που

προκύψει αθέτηση υποχρεώσεων σε ένα στεγαστικό δάνειο, ο δανειστής μπορεί να κατάσχει μία κατοικία και να την πουλήσει για να ανακτήσει το χρηματικό ποσό που δάνεισε. Όμως η ποσότητα *LGD* αναπαριστά το ποσοστό που δεν είναι ανακτήσιμο. Επομένως, η αναμενόμενη ζημιά κάτω από συνθήκες οικονομικής ύφεσης είναι $K(PD)\times LGD\times EAD$ αλλά από αυτή τη ζημιά το ποσό $PD\times LGD\times EAD$ πρόκειται να καλυφθεί από τα κέρδη δανεισμού. Επομένως, για ένα τμήμα του χαρτοφυλακίου, οι κεφαλαιακές απαιτήσεις είναι

$$K(PD)\times LGD\times EAD - PD\times LGD\times EAD \tag{7.2}$$

ή διαφορετικά για κάθε μία μονάδα χρηματοδοτικού ανοίγματος η μη αναμενόμενη ζημιά είναι

$$K(PD)\times LGD - PD\times LGD.$$

Σύμφωνα με τις επίσημες οδηγίες της επιτροπής της Βασιλείας ΙΙ, τα καταναλωτικά δάνεια χωρίζονται σε τρεις κατηγορίες, τα στεγαστικά, τα ανακυκλούμενα (πιστωτικές κάρτες) και στην τρίτη κατηγορία ανήκουν όλα τα υπόλοιπα. Για όλες αυτές τις περιπτώσεις ο τύπος με βάση τον οποίο υπολογίζεται το κεφάλαιο που ο χρηματοπιστωτικός οργανισμός πρέπει να τοποθετήσει στην άκρη είναι ένα ποσοστό του εκτιμώμενου χρηματοδοτικού ανοίγματος *EAD*. Ο τύπος αυτός είναι ο εξής:

$$\tilde{K} = LGD\times K(PD) - LGD\times PD =$$

$$= LGD\times \Phi\left(\left(\frac{1}{1-R}\right)^{1/2}\Phi^{-1}(PD)+\left(\frac{R}{1-R}\right)^{1/2}\Phi^{-1}(0,999)\right) - LGD\times PD.$$

όπου Φ είναι η αθροιστική συνάρτηση της τυπικής κανονικής κατανομής, Φ^{-1} η αντίστροφη αυτής και R είναι ο συντελεστής συσχέτισης της απόδοσης διάφορων ανοιγμάτων. Για παράδειγμα, για τα στεγαστικά δάνεια ο συντελεστής συσχέτισης είναι $R=0,15$ για την ανακυκλούμενα δάνεια είναι $R=0,04$ και για τα υπόλοιπα καταναλωτικά δάνεια όπως είναι τα προσωπικά η συσχέτιση εξαρτάται από την πιθανότητα αθέτησης υποχρεώσεων και υπολογίζεται από τον τύπο:

$$R = 0,03\left(\frac{1-e^{-35PD}}{1-e^{-35}}\right)+0,16\left(1-\frac{1-e^{-35PD}}{1-e^{-35}}\right).$$

Οι παραπάνω επιλογές του συντελεστή συσχέτισης προκύπτουν από την ανάγκη να ληφθούν κεφαλαιακές απαιτήσεις οι οποίες να φαίνονται σωστές όταν συγκρίνονται με εμπειρικά δεδομένα. Δεν υπάρχει κάποιος οικονομικός παράγοντας που να παίζει ρόλο στην επιλογή τιμών των παραπάνω συντελεστών. Ο λόγος είναι ότι οι παραπάνω τύποι στηρίζονται

στο θεωρητικό πλαίσιο της αποτίμησης των δικαιωμάτων προαίρεσης (*options*) του Merton (1974) σύμφωνα με το οποίο ο πιστωτικός κίνδυνος σχετίζεται με την κεφαλαιακή δομή της επιχείρησης. Δηλαδή, το υπόδειγμα αυτό μπορεί να χρησιμοποιηθεί σε διάφορα είδη δανείων.

7.4 Καθορισμός αθέτησης υποχρεώσεων

Ένα μεγάλο μέρος των μεθόδων δημιουργίας μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας μπορούν να ικανοποιήσουν τις προϋποθέσεις που πρέπει να πληρούνται για να εγκριθεί η εφαρμογή της μεθόδου εσωτερικών διαβαθμίσεων για τον υπολογισμό των κεφαλαιακών απαιτήσεων προς κάλυψη του πιστωτικού κινδύνου. Οι απαιτήσεις αυτές αφορούν τη δομή των συστημάτων εσωτερικών διαβαθμίσεων, την ποσοτικοποίηση των παραμέτρων κινδύνου (*PD, LGD, EAD*), την επικύρωση των μοντέλων, το βαθμό χρησιμοποίησης των μοντέλων (*use test*) και την ύπαρξη ανεξάρτητης μονάδας ελέγχου (*control*) πιστωτικού κινδύνου.[22] Κάποιος οφειλέτης θεωρείται ότι αθετεί τις υποχρεώσεις του όταν είναι πιθανό να μην εκπληρώσει τις υποχρεώσεις πληρωμών προς το χρηματοπιστωτικό ίδρυμα ή όταν εμφανίσει καθυστέρηση 90 ημερών για την αποπληρωμή των υποχρεώσεών του. Η αθέτηση υποχρεώσεων δεν είναι απαραίτητο πάντα να οδηγεί σε ζημιά. Υπάρχουν περιπτώσεις στις οποίες, αν και επήλθε αθέτηση υποχρεώσεων ο οφειλέτης πλήρωσε τελικά όλες τις υποχρεώσεις του. Η ζημιά από οφειλέτες οι οποίοι εμφάνισαν αθέτηση υποχρεώσεων είναι το δεύτερο συστατικό υπολογισμού της αναμενόμενης ζημιάς *LGD*. Σε μερικές περιπτώσεις, όπως για παράδειγμα στο δανεισμό νομικών προσώπων δημοσίου τομέα, η αθέτηση υποχρεώσεων επέρχεται όταν ο οφειλέτης εμφανίσει καθυστέρηση 180 ημερών. Αυτό προκαλεί κάποιες δυσκολίες στον προσδιορισμό ποιος οφειλέτης είναι πιθανό να εμφανίσει 180 ημέρες καθυστέρησης διότι πολλοί λιγότεροι πελάτες είναι πιθανό να φτάσουν σε αυτό το σημείο καθυστέρησης.

Η διαβάθμιση των οφειλετών μπορεί να γίνεται με βάση την υπάρχουσα κατάσταση τη στιγμή της ταξινόμησης (*point in time - PIT*) ή με βάση την αναμενόμενη συμπεριφορά τους κατά τη διάρκεια ενός πλήρους οικονομικού κύκλου (*through-the-cycle - TTC*). Τα

[22] ΤτΕ έγγραφο διαβούλευσης ΙΙ (2004).

συστήματα PIT χρησιμοποιούνται όταν η διαβάθμιση των αντισυμβαλλομένων γίνεται με βάση τις διαθέσιμες πληροφορίες όπως διαμορφώνονται τη στιγμή της διαβάθμισης χωρίς να γίνεται καμία υπόθεση σχετικά με την εξέλιξη των παραγόντων κινδύνου. Από την άλλη πλευρά, τα συστήματα TTC χρησιμοποιούνται όταν η διαβάθμιση των αντισυμβαλλομένων λαμβάνει υπόψη τη συμπεριφορά τους σε όλες τις φάσεις του οικονομικού κύκλου καθώς και τυχόν δυσμενή γεγονότα που μπορεί μακροπρόθεσμα να επηρεάσουν αρνητικά τη συναλλακτική συμπεριφορά. Η φιλοσοφία των συστημάτων αυτών είναι ότι εκτιμούν τη σταθερότητα των αποτελεσμάτων στην περίπτωση αυξημένης οικονομικής διακύμανσης και καλύπτουν συνήθως μακροχρόνιο ορίζοντα.[23]

Οι βαθμολογίες αθέτησης υποχρεώσεων για την καταναλωτική πίστη υποδεικνύουν την πιθανότητα ένας οφειλέτης να αποτύχει να εκπληρώσει τις υποχρεώσεις του μέσα σε ένα δεδομένο μελλοντικό χρονικό διάστημα. Η επιτροπή της Βασιλείας προτείνει αυτό το χρονικό διάστημα να είναι 12 μηνών. Όμως, οι βαθμολογίες που προκύπτουν από ένα CSM είναι συνήθως PIT βαθμολογίες που δίνουν την πιθανότητα αθέτησης υποχρεώσεων μέσα στους επόμενους 12 μήνες. Όμως οι εξωτερικοί οίκοι αξιολόγησης όπως οι Standard & Poor και Moodys θεωρούν ότι η περίπτωση της επιχειρησιακής πίστης όπως η έκδοση ομολόγων ανήκει στην κατηγορία TTC. Επομένως, αυτές οι αξιολογήσεις δεν πρέπει να αλλάξουν εξαιτίας των οικονομικών καταστάσεων αλλά μόνο λόγω παραγόντων που έχουν επιπτώσεις στο μεμονωμένο χρηματοπιστωτικό οργανισμό. Δεδομένου ότι η διαχείριση κινδύνου στα περισσότερα διεθνή πιστωτικά ιδρύματα επηρεάζεται κυρίως από την επιχειρηματική πίστη, η φιλοσοφία της συμφωνίας της Βασιλείας επηρεάζεται και αυτή από την επιχειρηματική πίστη, όπου ως πιθανότητα αθέτησης υποχρεώσεων λαμβάνεται ο μακροπρόθεσμος μέσος όρος της πιθανότητας του οφειλέτη να αθετήσει τις υποχρεώσεις του στους επόμενους 12 μήνες.

Κατά συνέπεια, για να κατασκευαστεί ένα σκορόχαρτο καταναλωτικής πίστης με βάση τις οδηγίες της Βασιλείας ΙΙ, θα πρέπει η PIT πιθανότητα αθέτησης υποχρεώσεων να μεταφραστεί σε μια TTC εκτίμηση πιθανότητας αθέτησης υποχρεώσεων. Για να επιτευχθεί αυτό είναι αρκετά δύσκολο, ειδικότερα όταν δεν υπάρχουν επαρκή ιστορικά δεδομένα για να σχεδιαστεί η πιθανότητα αθέτησης υποχρεώσεων των οφειλετών σε έναν οικονομικό κύκλο. Αλλά ακόμα και αν αυτές οι πληροφορίες είναι διαθέσιμες, είναι αμφισβητήσιμο αν αυτές μπορούν να χρησιμοποιηθούν για την καταναλωτική πίστη αφού είναι σχεδόν απίθανο τα

[23] ΤτΕ έγγραφο διαβούλευσης V (2004).

χαρακτηριστικά γνωρίσματα και η τιμή του δανείου, οι επιχειρηματικές πολιτικές να μείνουν σταθερά κατά τη διάρκεια μιας τέτοιας χρονικής περιόδου. Ως εκ τούτου ο τύπος οφειλετών που υποβάλλουν αίτηση για ένα ιδιαίτερο προϊόν δανεισμού και η απόδοσή τους είναι πιθανό να αλλάξουν και είναι αρκετά δύσκολο να καταγραφούν αυτές οι αλλαγές λόγω του οικονομικού κύκλου. Παρ' όλες αυτές τις προειδοποιήσεις, οι δανειστές πρέπει να προσδιορίσουν ποια είναι η κατάλληλη PIT πιθανότητα αθέτησης υποχρεώσεων για το μετασχηματισμό της σε TTC πιθανότητας αθέτησης υποχρεώσεων με βάση το ρυθμιστικό πλαισίο της κεφαλαιακής επάρκειας.

Έστω ότι $PD_t(s,t_0)$ είναι η PIT πιθανότητα αθέτησης υποχρεώσεων τη χρονική στιγμή t για το τμήμα του δανειακού χαρτοφυλακίου που παρουσίασε βαθμολογία αθέτησης υποχρεώσεων s τη χρονική στιγμή t_0, όπου η βαθμολογία είναι ο λογάριθμος της σχετικής πιθανότητας. Επομένως, απαιτείται να υπολογιστεί το $\overline{PD}(s,t_0)$, που είναι ο μακροπρόθεσμος μέσος όρος της PIT πιθανότητας αθέτησης υποχρεώσεων για το ίδιο τμήμα του δανειακού χαρτοφυλακίου. Αν θεωρήσουμε ότι το μήκος του οικονομικού κύκλου είναι T τότε

$$\overline{PD}(s,t_0)=\frac{1}{T}\int_0^T PD_t(s,t_0)dt\,.$$

Μια πιθανή προσέγγιση είναι να χρησιμοποιηθεί η ανάλυση της βαθμολογίας στο λογάριθμο της σχετικής πιθανότητας του πληθυσμού και στο λογάριθμο της πληροφοριακής σχετικής πιθανότητας που περιγράφεται στη σχέση (3.18). Η πιθανότητα αυτών που αθετούν τις υποχρεώσεις τους (κακοί) και αυτών που καταφέρνουν να τις εκπληρώσουν (καλοί) τη χρονική στιγμή t είναι $p_B(t)$ και $p_G(t)=1-p_B(t)$ αντίστοιχα. Αν ο λογάριθμος της σχετικής πιθανότητας της βαθμολογίας s, δομείται από ένα σύνολο χαρακτηριστικών $\mathbf{x}$, τα $P_t(B|\mathbf{x})$ και $P_t(G|\mathbf{x})$ είναι τα ποσοστά αυτών που εμφανίζουν χαρακτηριστικά $\mathbf{x}$ και είναι «κακοί» και «καλοί» αντίστοιχα τη χρονική στιγμή t. Επομένως, σύμφωνα με το θεώρημα του Bayes και τον ορισμό του λογαρίθμου της σχετικής πιθανότητας της βαθμολογίας, η βαθμολογία $s_t(\mathbf{x})$ τη χρονική στιγμή t δίνεται από τον τύπο

$$s_t(\mathbf{x})=\ln\left(\frac{P_t(G|\mathbf{x})}{P_t(B|\mathbf{x})}\right)=\ln\left(\frac{p_G(t)}{p_B(t)}\right)+\ln\left(\frac{f_t(\mathbf{x}|G)}{f_t(\mathbf{x}|B)}\right)=s_t(pop)+s_{It}(\mathbf{x}) \qquad (7.3)$$

όπου $f_t(\mathbf{x}|G)$ είναι η συνάρτηση πιθανοφάνειας.

Όμως, ο όρος $s_{It}(\mathbf{x}) = \ln\left(\frac{f_t(\mathbf{x}|G)}{f_t(\mathbf{x}|B)}\right)$ που εξαρτάται από την πληροφορία που προσφέρει ο πελάτης με τις ιδιότητες **x** δεν εξαρτάται από τη χρονική στιγμή t, επομένως, ισχύει ότι:

$$s_{It}(\mathbf{x}) = \ln\left(\frac{f_t(\mathbf{x}|G)}{f_t(\mathbf{x}|B)}\right) = \ln\left(\frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)}\right) = s_I(\mathbf{x})$$

και από τη σχέση (7.3) προκύπτει ότι:

$$s_t(\mathbf{x}) = s_t(pop) + s_I(\mathbf{x}). \qquad (7.4)$$

Συμβολίζοντας με $X(s,t_0) = \{\mathbf{x}: s_{t_0}(\mathbf{x}) = s\} = \{\mathbf{x}: s_{It_0}(\mathbf{x}) = s - s_{t_0}(pop) = \bar{s}\}$ το τμήμα του πληθυσμού που μας ενδιαφέρει τη χρονική στιγμή t, μπορούμε να εισάγουμε τους συμβολισμούς

$$P_t(B \mid X(s,t_0)) = P_t(B \mid s,t_0) = PD_t(s,t_0) \text{ και}$$

$$P_t(G \mid X(s,t_0)) = P_t(G \mid s,t_0) = 1 - PD_t(s,t_0).$$

Χρησιμοποιώντας λοιπόν τη σχέση (7.4) προκύπτει ότι:

$$s_t(s,t_0) = \ln\left(\frac{P_t(G \mid s,t_0)}{P_t(B \mid s,t_0)}\right) = \ln\left(\frac{p_G(t)}{p_B(t)}\right) + \ln\left(\frac{f(s,t_0 \mid G)}{f(s,t_0 \mid B)}\right) = s_t(pop) + s_I(\bar{s}).$$

Δηλαδή,

$$\ln\left(\frac{1 - PD_t(s,t_0)}{PD_t(s,t_0)}\right) = \ln\left(\frac{p_G(t)}{p_B(t)}\right) + s_I(\bar{s}) \Rightarrow PD_t(s,t_0) = \frac{1}{1 + e^{s_t(s,t_0)}} = \frac{1}{1 + \frac{p_G(t)}{p_B(t)} e^{s_I(\bar{s})}}.$$

Αφού όμως τη χρονική στιγμή t_0 η βαθμολογία s για το τμήμα του πληθυσμού που μας ενδιαφέρει ικανοποιεί τη σχέση $s = s_{t_0}(pop) + s_I(\tilde{s})$ θα έχουμε

$$s = \ln\left(\frac{p_G(t_0)}{p_B(t_0)}\right) + s_I(\tilde{s}).$$

Επομένως, προκύπτει ότι η πιθανότητα αθέτησης υποχρεώσεων τη χρονική στιγμή t_0 για τους πελάτες με βαθμολογία s είναι

$$PD_t(s,t_0) = \frac{1}{1 + \frac{p_G(t)\,p_B(t_0)}{p_B(t)\,p_G(t_0)} e^{s}}.$$

Άρα,

$$\overline{PD}(s,t_0)=\frac{1}{T}\int_0^T PD_t(s,t_0)dt=\frac{1}{T}\int_0^T\frac{1}{1+\frac{p_G(t)p_B(t_0)}{p_B(t)p_G(t_0)}e^s}dt.$$

Κατά συνέπεια, για να γίνει αυτός ο μετασχηματισμός πρέπει να είναι γνωστή η σχετική πιθανότητα του πληθυσμού των «καλών» και των «κακών» για ολόκληρο τον οικονομικό κύκλο. Όμως, αυτός ο μετασχηματισμός βασίζεται στην υπόθεση ότι η πληροφοριακή σχετική πιθανότητα δεν αλλάζει με την πάροδο του χρόνου, γεγονός που στην πραγματικότητα δε συμβαίνει συχνά. Για παράδειγμα, για τις βαθμολογίες συμπεριφοράς οι οποίες χρησιμοποιούνται για να διαπιστωθεί αν η συμπεριφορά του πελάτη θα είναι «καλή» ή «κακή» σε μια δεδομένη χρονική στιγμή, η πληροφοριακή σχετική πιθανότητα δίνεται έμμεσα για την κάθε βαθμολογία και αλλάζει με την πάροδο του χρόνου. Αυτό είναι μια χρήσιμη προσέγγιση για να γίνει κατανοητός ο μετασχηματισμός της PIT πιθανότητας αθέτησης στον μακροπρόθεσμο μέσο όρο της πιθανότητας αθέτησης και δείχνει πως ο μετασχηματισμός ακόμα και στην πιο απλή περίπτωση δεν περιλαμβάνει τον πολλαπλασιασμό της τρέχουσας πιθανότητας αθέτησης υποχρεώσεων με έναν παράγοντα.

7.5 Χρηματοδοτικό άνοιγμα ή Έκθεση κατά τη στιγμή της αθέτησης

Στις περισσότερες περιπτώσεις η εκτίμηση του χρηματοδοτικού ανοίγματος *EAD* είναι αρκετά απλή και ενδεχομένως ντετερμινιστική. Σε αυτή τη φάση είναι γνωστό το οφειλόμενο υπόλοιπο και υπολογίζεται ο αριθμός των χαμένων, μη αποπληρωμένων δόσεων μαζί με το επιτόκιο αυτών που απαιτούνται για να σημειωθεί αθέτηση υποχρεώσεων. Ο κίνδυνος έκθεσης περιγράφεται από το ποσό που οφείλει ο δανειστής κατά τη στιγμή της αθέτησης υποχρεώσεων και ονομάζεται «Έκθεση κατά τη στιγμή της αθέτησης». Το άθροισμα αυτών των ποσοτήτων θα μπορούσε να αποτελεί μια καλή εκτίμηση του *EAD*, αλλά χρειάζεται πρώτα να ελεγχθούν κάποιες υποθέσεις, για παράδειγμα αν θα υπάρξει κάποια πληρωμή πριν εμφανιστεί ή αθέτηση υποχρεώσεων ή απλά για να καλυφθεί το επιτόκιο που έχει δημιουργηθεί.

Όμως, η ανακυκλούμενη πίστη όπως είναι οι πιστωτικές κάρτες και οι υπεραναλήψεις αποτελούν εξαίρεση για τις οποίες για να εκτιμηθεί η ποσότητα *EAD* χρησιμοποιείται το τρεχούμενο οφειλόμενο υπόλοιπο και το πιστωτικό όριο. Για την εκτίμηση του *EAD* μπορούν να χρησιμοποιηθούν κάποιες παράμετροι όπως είναι η διαφορά μεταξύ του πιστωτικού ορίου και του οφειλόμενου υπόλοιπου. Αυτή η διαφορά μοντελοποιείται χρησιμοποιώντας παλινδρόμηση, μοντέλα probit ή δομούνται μοντέλα Μαρκοβιανών Αλυσίδων ανάλογα με την περίπτωση. Στη συνέχεια, με βάση τις οδηγίες της Βασιλείας ΙΙ επιλέγεται το μοντέλο που είναι πιο αποτελεσματικό.

7.6 Ζημιά δεδομένης της αθέτησης

Ο κίνδυνος ανάκτησης περιγράφει το μέρος του ποσού που οφείλεται κατά τη στιγμή της αθέτησης το οποίο κατάφερε ο δανειστής να ανακτήσει από το δανειολήπτη. Το ποσό που δεν κατάφερε να ανακτήσει ο δανειστής ως προς την συνολική οφειλή ονομάζεται Ζημία δεδομένης της αθέτησης (*LGD*). Εάν το πιστωτικό ίδρυμα ανακτήσει πλήρως το οφειλόμενο ποσό (*EAD*), η επερχόμενη ζημία θα είναι μηδενική. Ωστόσο, η διαδικασία ανάκτησης μπορεί να διαρκέσει αρκετά χρόνια και μέρος των χρημάτων μπορεί να μην ανακτηθεί ποτέ.

Αν και η ποσότητα *LGD* ασκεί μεγάλη επίδραση στον υπολογισμό του ρυθμιστικού κεφαλαίου σύμφωνα με το κανονιστικό πλαίσιο της Βασιλείας ΙΙ, δεν έχει λάβει ιδιαίτερη προσοχή. Από το 2000 οι ερευνητές του επιχειρηματικού πιστωτικού κινδύνου ξεκίνησαν να μοντελοποιούν το *LGD* ως μια σταθερή τιμή που λαμβανόταν από ιστορικές μέσες τιμές και δεν επηρεαζόταν από τις οικονομικές συνθήκες, τον τύπο και τη διάρκεια του δανείου. Όμως σήμερα, σύμφωνα με τους Altman et al. (2005), είναι αναγνωρισμένο ότι τα *LGD* και *PD* συσχετίζονται και επομένως χρειάζεται πλέον να δημιουργηθούν νέα μοντέλα για να αντιμετωπιστεί αυτό το γεγονός.

Η δυσκολία στη μοντελοποίηση του *LGD* είναι κυρίως η έλλειψη δεδομένων, ειδικότερα στα προϊόντα δανεισμού όπου τα ποσοστά αθέτησης υποχρεώσεων είναι ιστορικώς χαμηλά. Για να αντιμετωπιστεί το πρόβλημα των δανειακών χαρτοφυλακίων με χαμηλά ποσοστά αθέτησης υποχρεώσεων έχουν προταθεί διάφοροι τρόποι. Για να εκτιμηθεί η πιθανότητα αθέτησης υποχρεώσεων σε τέτοιου είδους χαρτοφυλάκια, ένας τρόπος είναι να ληφθεί υπόψη

κάθε αθέτηση υποχρεώσεων χωρίς να έχει σημασία πριν πόσο καιρό σημειώθηκε αυτή μέσα στη διάρκεια ισχύος του δανείου. Έπειτα, χρησιμοποιώντας τις τεχνικές της Ανάλυσης Επιβίωσης επαναπροσδιορίζονται τα αποτελέσματα με βάση την τυποποιημένη αθέτηση υποχρεώσεων των δώδεκα μηνών. Παρ' όλα αυτά για την εκτίμηση του *LGD* δεν υπάρχει διέξοδος εάν δεν υπάρχει καμιά πληροφορία διαθέσιμη που να επιτρέπει τη δόμηση των πραγματικών απωλειών στις λίγες αθετήσεις υποχρέωσης που έχουν συμβεί. Η μοναδική εξαίρεση σε αυτήν την έλλειψη στοιχείων είναι συνήθως το χαρτοφυλάκιο στεγαστικών δανείων όπου η αθέτηση υποχρεώσεων καταλήγει συνήθως σε ανάκτηση περιουσίας και ακολούθως μεταπώληση και υπάρχει η ελπίδα δόμησης μοντέλων εκτιμώντας τους παράγοντες που επηρεάζουν αυτήν την αξία μεταπώλησης.

7.7 Επικύρωση και προσαρμογή μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας

Οι κανονισμοί της Βασιλείας έχουν επισημάνει διάφορες ανεπάρκειες όσον αφορά τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας. Όπως είδαμε στο προηγούμενο κεφάλαιο η επικύρωση και η μέτρηση απόδοσης των μοντέλων γίνεται κυρίως με τρεις διαφορετικές κατηγορίες μεθόδων: τη μέτρηση της ακρίβειας ταξινόμησης των πελατών, τη χρήση διάφορων μέτρων που αξιολογούν τη διαχωριστική ικανότητα του μοντέλου και τη μέτρηση της ακρίβειας προσαρμογής της εκτίμησης της πιθανότητας αθέτησης υποχρεώσεων των υποψηφίων. Οι δύο πρώτες κατηγορίες είναι αυτές που παραδοσιακά χρησιμοποιούνται στην ανάπτυξη των CSM, ενώ η τρίτη κατηγορία μεθόδων που εκτιμά αν οι βαθμολογίες αντικατοπτρίζουν ακριβώς την πιθανότητα αθέτησης υποχρεώσεων είναι λιγότερο σημαντική σε σχέση με την ακρίβεια ταξινόμησης των βαθμολογιών και αυτό είναι λογικό διότι ο κύριος στόχος είναι να γίνει αποδεκτή η αίτηση δανεισμού των «καλών» πελατών. Επίσης, η βαθμολογία αποδοχής – απόρριψης συνήθως επιλέγεται υποκειμενικά από τους υπεύθυνους λήψης αποφάσεων λαμβάνοντας υπ' όψιν διαφορετικά ζητήματα από την πιθανότητα αθέτησης υποχρεώσεων.

Σύμφωνα με την επιτροπή της Βασιλείας ΙΙ, η πιο σημαντική μέθοδος επικύρωσης και μέτρησης απόδοσης ενός CSM είναι η μέθοδος της προσαρμογής, δηλαδή η μέτρηση κατά

πόσο οι εκτιμώμενες πιθανότητες αθέτησης υποχρεώσεων αντικατοπτρίζουν τις πραγματικές τιμές, διότι στο κανονιστικό πλαίσιο της Βασιλείας ΙΙ η πιθανότητα αθέτησης υποχρεώσεων είναι αυτή που χρησιμοποιείται στις εξισώσεις κεφαλαιακής επάρκειας. Για να βρεθούν οι πραγματικές τιμές των πιθανοτήτων αθέτησης υποχρεώσεων χρειάζεται να είναι διαθέσιμη μια ομάδα με παρόμοιο είδος δανείων. Στη συνέχεια, η εκτιμώμενη πιθανότητα αθέτησης υποχρεώσεων συγκρίνεται με το ποσοστό αθέτησης υποχρεώσεων που παρατηρείται σε αυτό το τμήμα. Ο Tache (2005) ισχυρίζεται ότι δεν είναι εύκολο να γίνει αυτή η σύγκριση γιατί δεν μπορεί να είναι σίγουρο αν οι διαφορές προέρχονται από συστηματικά λάθη του CSM ή από τυχαίες διακυμάνσεις. Επιπλέον, είναι πολύ περιορισμένοι οι έλεγχοι που μπορούν να χρησιμοποιηθούν. Για να επιβεβαιωθεί ότι ένα CSM είναι καλά προσαρμοσμένο, συγκρίνεται η βαθμολογία με τον λογαρίθμου της σχετικής πιθανότητας των «καλών» προς τους «κακούς» που χρησιμοποιείται συνήθως στα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας. Αλλά σε αυτήν την περίπτωση συγκρίνεται η τελική βαθμολογία με τη σχετική πιθανότητα των «καλών» προς τους «κακούς» του δείγματος των οφειλετών με βάση τους οποίους δομήθηκε το μοντέλο. Σύμφωνα με τις οδηγίες της Βασιλείας ΙΙ, η μέθοδος της προσαρμογής απαιτεί να γίνει σύγκριση της βαθμολογίας με τη μελλοντική σχετική πιθανότητα αθέτησης υποχρεώσεων του τμήματος των δανειοληπτών για τους οποίους έχουν εκτιμηθεί οι τιμές της αθέτησης υποχρεώσεων για μια δεδομένη ζώνη πιθανοτήτων αθέτησης υποχρεώσεων.

Οι κανονισμοί της Βασιλείας ΙΙ προτείνουν ο δανειστής να έχει αρκετά ιστορικά στοιχεία (τουλάχιστον 5 ετών) για να δύναται ένας χρηματοπιστωτικός οργανισμός να προχωρήσει σε τέτοιες εκτιμήσεις προσαρμογής. Όμως, συνήθως δεν είναι δυνατό να συγκρατούνται πληροφορίες για τόσο μεγάλες χρονικές περιόδους. Ένας τρόπος λοιπόν για να αυξηθεί η ποσότητα αυτών των πληροφοριών που χρησιμοποιούνται στα CSM έτσι ώστε να ελεγχτεί ακριβέστερα η προσαρμογή τους είναι να εισαχθούν στα μοντέλα οικονομικές μεταβλητές. Παλαιότερα, δεν εισάγονταν στα μοντέλα οικονομικές μεταβλητές γιατί θεωρούνταν ότι οι χρονικές περίοδοι παρατήρησης και απόδοσης ήταν αρκετά μικρές για να υπάρχουν μεταβολές στην οικονομία.

7.8 Ανακεφαλαίωση

Το νέο κανονιστικό πλαίσιο της Βασιλείας ΙΙ αποτελείται από τρεις πυλώνες, ο πρώτος από αυτούς αφορά την ποσοτικοποίηση και τον υπολογισμό των κινδύνων και των ανάλογων κεφαλαιακών απαιτήσεων. Ο δεύτερος πυλώνας αναφέρεται στην εποπτεία των συστημάτων διαχείρισης κινδύνου της τράπεζας από την Κεντρική Τράπεζα. Οι εποπτικές αρχές απαιτούν από έναν χρηματοπιστωτικό οργανισμό να αναθεωρεί συχνά τα μοντέλα και τα συστήματα διαχείρισης κινδύνου που χρησιμοποιεί και αναλόγως να αυξάνει το ρυθμιστικό κεφάλαιο του ή να μειώνει το δανειακό του χαρτοφυλάκιο. Τέλος, ο τρίτος πυλώνας αφορά τη δημοσίευση και γνωστοποίηση των στοιχείων προς την αγορά που αφορούν την κεφαλαιακή επάρκεια και τη διαχείριση κινδύνων της τράπεζας.

Για τον πιστωτικό κίνδυνο, οι κανονισμοί χωρίζουν το χαρτοφυλάκιο δανεισμού σε πέντε κατηγορίες. Ένας χρηματοπιστωτικός οργανισμός έχει τη δυνατότητα να επιλέξει την τυποποιημένη μέθοδο ή ένα από τα δύο είδη μοντέλων εσωτερικών διαβαθμίσεων (θεμελιώδης και προηγμένη μέθοδος). Στη θεμελιώδη μέθοδο ο χρηματοπιστωτικός οργανισμός χρειάζεται να αναπτύξει μοντέλα που να εκτιμούν μόνο την πιθανότητα αθέτησης υποχρεώσεων, ενώ η προηγμένη μέθοδος απαιτεί και την εκτίμηση των ποσοτήτων *LGD* και *EAD*. Για το λιανικό χαρτοφυλάκιο που περιλαμβάνει την καταναλωτική πίστη επιτρέπεται να χρησιμοποιηθεί μόνο η προηγμένη μέθοδος ανάπτυξης μοντέλων εσωτερικών διαβαθμίσεων.

Όταν χρησιμοποιούνται μοντέλα εσωτερικών διαβαθμίσεων για ένα λιανικό χαρτοφυλάκιο, είναι παραίτητο το χαρτοφυλάκιο να διαχωριστεί αρχικά σε έναν αριθμό από τμήματα και για κάθε ένα από αυτά να παρέχονται οι εκτιμήσεις των ποσοτήτων *PD*, *LGD* και *EAD*. Όμως υπάρχει μια στοιχειώδης διαφορά με τον τρόπο που υπολογίζονται τα *PD* και *LGD*. Η ποσότητα *PD* είναι η μακροπρόθεσμη πιθανότητα να αθετήσει τις υποχρεώσεις του ένας δανειολήπτης μέσα στους επόμενους 12 μήνες. Παρ' όλα αυτά δεν υπάρχει πρόβλημα στην εκτίμηση της πιθανότητας αθέτησης σε περιόδους οικονομικής ύφεσης διότι η ποσότητα *K(PD)* δίνει αυτήν την πιθανότητα σε άσχημες οικονομικές συνθήκες. Το πραγματικό πρόβλημα δημιουργείται με τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας αφού η *PD* που εκτιμάται με τη βοήθεια αυτών στην πραγματικότητα αντιπροσωπεύει την PIT *PD*, δηλαδή την πιθανότητα αθέτησης υποχρεώσεων μέσα στους επόμενους 12 μήνες. Οι οδηγίες της συμφωνίας της Βασιλείας ΙΙ απαιτούν η εκτίμηση της *PD* να είναι η μέση

μακροπρόθεσμη *PD*, η οποία είναι η μέση τιμή των PIT πιθανοτήτων αθέτησης υποχρεώσεων μέσα σε έναν οικονομικό κύκλο. Όμως αυτός ο υπολογισμός είναι δύσκολο να γίνει στην πράξη διότι ένα CSM μπορεί να αναπροσαρμοστεί και να ξανακατασκευαστεί πολλές φορές μέσα σε έναν οικονομικό κύκλο. Ο μετασχηματισμός των PIT πιθανοτήτων αθέτησης υποχρεώσεων σε μακροπρόθεσμες αποτελεί ακόμα θέμα διαμάχης μεταξύ των ρυθμιστικών αρχών και των χρηματοπιστωτικών ιδρυμάτων.

Επιπλέον, σύμφωνα με τους κανονισμούς της Βασιλείας ΙΙ, για τα λιανικά χαρτοφυλάκια, η αναμενόμενη ζημιά ενός οργανισμού σε ένα τμήμα δίνεται από τον τύπο (7.1) και αυτή η προβλεπόμενη ζημιά αναμένεται να καλυφθεί από τα κέρδη του δανεισμού. Επομένως, δεν χρειάζεται να μπει στην άκρη επιπλέον κεφάλαιο. Όμως, η μη αναμενόμενη ζημιά δίνεται από τον τύπο (7.2) και η ποσότητα αυτή αποτελεί το ελάχιστο κεφάλαιο που απαιτείται να κρατηθεί σύμφωνα τα συστήματα εσωτερικής διαβάθμισης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Γκέκας Β. Δημήτριος (2005). Ομιλία με θέμα «Τράπεζα της Ελλάδος και εποπτεία του πιστωτικού συστήματος», Θεσσαλονίκη.
[URL: http://www.eap.gr/programmes/tra/tra50/docs/gr_bank.doc].

Γναρδέλλης Χαράλαμπος (2006). *Ανάλυση Δεδομένων με το SPSS* 14.0 *for Windows*, Εκδόσεις Παπαζήση, Αθήνα.

Ζοπουνίδης Κωνσταντίνος και Λεμονάκης Χρήστος (2009). *Διαχείριση Πιστωτικού Κινδύνου*, Εκδόσεις Κλειδάριθμος, Αθήνα.

Καρλής Δημήτρης (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*, Εκδόσεις Αθ. Σταμούλη, Αθήνα.

Κατέρη Μαρία (2008). *Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία*, Σημειώσεις παραδόσεων, Πανεπιστήμιο Πειραιώς, Πειραιάς.

Κούτρας Μάρκος. (2007). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά Συστάδες*, Πανεπιστημιακές παραδόσεις για το ΠΜΣ «Εφαρμοσμένη Στατιστική», Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς.

Τράπεζα της Ελλάδος, Έγγραφο Διαβούλευσης ΙΙ (2004). *Μέθοδος Εσωτερικών Διαβαθμίσεων*, έκδοση ΤτΕ, Αθήνα.

Τράπεζα της Ελλάδος, Έγγραφο Διαβούλευσης V (2004). *Βασικές Προϋποθέσεις για την Ανάπτυξη της Μεθόδου των Εσωτερικών Διαβαθμίσεων*, έκδοση ΤτΕ, Αθήνα.

Ξένη

Agresti A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

Altman I. Edward (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, **23**, 589-609.

Altman E.I., Brady B., Resti A. and Sironi A. (2005). The link between default and recovery rates; theory, empirical evidence and implications, *Journal of Business*, **78**, 2203–2222.

Asuncion A. and Newman D.J. (2007). UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science.
[URL: http://www.ics.uci.edu/~mlearn/MLRepository.html].

Baesens B. (2003). *Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques*, Doctoral Thesis no 180 Faculteit Economische en Toegepaste Economische Wetebnschappen, Katholieke Universiteit, Leuven.

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

Basel Committee on Banking Supervision (BCBS) (2006). *International Convergence of Capital Measurement and Capital Standards — A Revised Framework (Comprehensive Version*), Bank for International Settlements, Basel.

[URL: http://www.bis.org/publ/bcbs128.htm].

Boyle M., Crook J.N., Hamilton R. and Thomas L.C. (1992). Methods for credit scoring applied to slow payers, in *Credit Scoring and Credit Control,* Thomas L.C., Crook J.N. and Edelman D.B., Eds., Oxford University Press, Oxford, p75-90.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984). *Classification and Regression Trees,* Wadsworth, Belmont, CA.

Bryan J. Balin (2008). *Basel I, Basel II, and Emerging Markets: A Nontechnical Analysis,* the Johns Hopkins School of Advanced Studies in Washington DC.

Capon Noel (1982). Credit Scoring Systems: A Critical Analysis, *The Journal of Marketing*, **46**, 82-91.

Chatterjee S. and Barcun S. (1970). A nonparametric approach to credit screening, *Journal of the. American Statistical Association,* **65**, 150-154.

Che-hui Lien and I-Cheng Yeh (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*, **36**, 2473–2480.

Coffman J.Y. (1986). *The Proper Role of Tree Analysis in Forecasting the Risk Behavior of Borrowers,* MDS Reports, Management Decision Systems, Atlanta, GA, 3, 4, 7, 9.

Crook N. Jonathan, Edelman B. David and Thomas C. Lyn (2007). Recent developments in consumer credit risk assessment, *European Journal of Operational Research*, **183**, 1447-1465.

Desai V.S., Conway D.G., Crook J.N. and Overstreet G.A. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry,* **8**, 323–346.

Durand D. (1941). Risk Elements in Consumer Installment Financing, *National Bureau of Economic Research,* New York.

Eisenbeis R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics, *Journal Finance,* **32**, 875-900.

Eisenbeis R.A. (1978). Problems in applying discriminant analysis in credit scoring models, *Journal Banking Finance, **2***, 205-219.

Engelmann B. Hayden E. and Tasche D. (2003). Measuring the Discriminative Power of Rating Systems, *Discussion paper series 2: Banking and Financial Supervision*, Deutsche Bundesbank.

Fisher R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* **7**, 179-188.

Fix E. and Hodges J. (1952). *Discriminatory Analysis, Nonparametric Discrimination, Consistency Properties,* Report 4, Project 21-49-004, School of Aviation Medicine, Randolph Field, TX.

Grablowsky B.J. and Talley W.K. (1981). Probit and discriminant functions for classifying credit applicants: A comparison, *Journal of Economic Business,* **33**, 254-261.

Hand D.J. and Henley W.E. (1993). Can reject inference ever work?, *IMA Journal of Management Mathematics, **5***, 45-55.

Hand J.D. and Henley E.W. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review, *Journal of the Royal Statisical Society A*, **160**, 523-541.

Henley W.E. (1995). *Statistical Aspects of Credit Scoring*, Ph.D. thesis, Open University, Milton Keynes, U.K.

Henley W.E. and Hand D.J. (1996). A k-nearest-neighbour classiffer for assessing consumer credit risk, *Statistician*, **45**, 77-95.

Hosmer D.W., Lemeshow S. (1980). A goodness of fit test for the multiple logistic regression model, *Communications in Statistics*, **A10**, 1043–1069.

Hosmer D.W. and Lemeshow S. (2000). *Applied Logistic Regression*, 2nd Edition, Wiley, New York.

Hsia D. C. (1978). Credit scoring and the Equal Credit Opportunity Act, *Hastings Law J.*, 30, 371-448.

Kullback S. and Leibler R.A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79–86.

Lee T.S., Chiu C.C., Lu C.J. and Chen I.F. (2002). Credit scoring using the hybrid neural discriminant technique, *Expert Systems with Applications*, **23**, 245–254.

Lewis E. M. (1992). *An Introduction to Credit Scoring*, Athena Press, San Rafael, CA.

Lovie A.D. and Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction, *Journal of Forecasting*, **5**, 159–186.

Mahalanobis P.C. (1936). On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India*, **2**, 49–55.

Makowski P. (1985). Credit scoring branches out, *Credit World*, **75**, 30-37.

Menard S. (2005). *Applied Logistic Regression Analysis* (2nd edition), Sage University Paper series on Quantitative Applications in the Social Sciences, Beverly Hills, CA.

Merton R.C. (1974). On the Pricing of corporate debt: the risk structure of interest rates, *Journal of Finance*, **29**, 449-470.

Myers J.H. and Forgy E.W. (1963). The development of numerical credit evaluation systems, *Journal of American Statistical Association*, **58**,799—806.

Platts G. and Howe I. (1997). A single European scorecard. *In: Proceedings of Credit Scoring and Credit Control V*, Credit Research Centre, University of Edinburgh.

Reichert A. K., Cho C. C. and Wagner G.M. (1983). An examination of the conceptual issues involved in developing credit scoring models. *Journal of Business and Economic Statistics*, ***1***, 101-114.

Siddiqi Naeem (2000). *Credit Risk Scorecards – Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Sobehart Jorge R and Sean C Keenan (2001). Measuring Default Accurately, *Risk Magazine*, **11**, 31-33.

Srinivasan V. and Kim Y.H. (1987). Credit granting: A comparative analysis of classification procedures, *Journal of Finance*, **42**, 665-683.

Swets J.A. (1988). Measuring the accuracy of diagnostic systems, *Science*, **240**, 1285-1293.

Tasche D. (2005). Rating and Probability of Default Validation, *Studies on the Validation of Internal Rating Systems*, BIS Working Paper No 14., pp. 28-59.

Thomas C. Lyn (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, **16**, 149-172.

Thomas C. Lyn, Ho J. and Scherer T.W. (2001). Time will tell: Behavioural scoring and the dynamics of consumer credit assessment, *University of Southampton - Department of Accounting and Management Science in its series*, Papers with number 01-174.

Thomas C. Lyn, Edelman B. David and Crook N. Jonathan (2002). *Credit Scoring and Its Applications*, Siam, Philadelphia.

Thomas C. Lyn (2009). *Consumer Credit Models – Pricing, Profit, and Portfolios*, Oxford University Press, New York.

Vojtek M. and Kočenda E. (2006). Credit Scoring Methods, *Czech Journal of Economics and Finance*, **56**, 152-167.

West D. (2000). Neural network credit scoring models. *Computers and Operational Research*, **27**, 1131–1152.

Wiginton J.C. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour, *Journal of Financial and Quantitative Analysis*, **15**, 757-770.

Wonderlic E.F. (1952). An analysis of factors in granting credit, *Indiana University Bull.*, **50**, 163-176.

Yobas M.B., Crook J.N. and Ross P. (1997). *Credit Scoring Using Neural and Evolutionary Techniques*, Working Paper 97/2, Credit Research Centre, University of Edinburgh, Edinburgh, Scotland.

Zekic-Susac Marijana, Sarlija Natasa and Mirta Bensic (2004). Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models, *26th International Conference on Information Technology Interfaces*, **1**, 265-270.