ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΙΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ

ΤΕΧΝΙΚΕΣ ΑΝΑΠΤΥΞΗΣ ΠΙΣΤΟΛΗΠΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΒΑΘΜΟΛΟΓΗΣΗΣ ΑΙΤΗΣΕΩΝ

ΑΝΝΑ ΔΗΜΑΚΗ

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

ΠΕΙΡΑΙΑΣ

ΣΕΠΤΕΜΒΡΙΟΣ 2022

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. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Τα μέλη της Επιτροπής ήταν:

1. Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
2. Αναπληρωτής Καθηγητής Αντζουλάκος Δημήτριος
3. Αναπληρωτής Καθηγητής Μπούτσικας Μιχαήλ

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

UNIVERSITY OF PIRAEUS

SCHOOL OF FINANCE AND STATISTICS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
ACTUARIAL SCIENCE AND RISK MANAGEMENT

Application credit scoring techniques

By

ANNA DIMAKI

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 Actuarial Science and Risk Management

PIRAEUS

SEPTEMBER 2022

Στους γονείς μου

Γιάννη & Μαρία

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να πω ένα πολύ με μεγάλο ευχαριστώ σε όλους τους ανθρώπους που συνέβαλλαν για την ολοκλήρωση της διπλωματικής μου εργασίας. Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κύριο Μάρκο Κούτρα, ο οποίος με καθοδήγησε καθ' όλη τη διάρκεια της συγγραφής της εργασίας και ήταν πάντα πρόθυμος να απαντήσει οποιαδήποτε ερώτηση και αν του έθετα. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου και ειδικότερα τους γονείς μου που πάντα είναι δίπλα μου και με πολλή αγάπη με στηρίζουν υλικά και ψυχολογικά.

ΠΕΡΙΛΗΨΗ

Σύμφωνα με το κανονιστικό πλαίσιο που τέθηκε από τη Βασιλεία στα χρηματοπιστωτικά ιδρύματα, για τη λήψη κρίσιμων αποφάσεων που αφορούν αιτήσεις χορηγήσεων δανείων είναι απαραίτητη η ανάπτυξη κατάλληλων μοντέλων πιστοληπτικής ικανότητας (application credit scoring models). Η ανάπτυξη τέτοιων μοντέλων βασίζεται κυρίως σε στατιστικές μεθόδους οι οποίες χρησιμοποιούν παρελθοντικά δεδομένα με στόχο την πρόβλεψη συμπεριφοράς των μελλοντικών πελατών του χρηματοπιστωτικού ιδρύματος.

Στην παρούσα εργασία αρχικά θα γίνει παρουσίαση των προτάσεων των συμφώνων της Βασιλείας για τον τρόπο ανάπτυξης πιστοληπτικών μοντέλων βαθμολόγησης αιτήσεων, μετέπειτα θα παρουσιαστούν οι διάφορες στατιστικές τεχνικές που έχουν προταθεί για την ανάπτυξη στατιστικών μοντέλων πιστοληπτικής ικανότητας και θα γίνει αναφορά στα πλεονεκτήματα και μειονεκτήματα κάθε μοντέλου. Τέλος θα γίνει εφαρμογή των μεθόδων σε μία μελέτη περίπτωσης.

ABSTRACT

According to the regulatory framework set by Basel for financial institutions, the development of appropriate application credit scoring models is necessary for making critical decisions regarding loan applications. The development of such models is based on statistical methods that exploit past data in order to predict the behaviour of future customers of the financial institution.

In this thesis we will first present the proposals of the Basel Accords on developing credit scoring models for applications, then we will present several statistical techniques that have been proposed for the development of statistical credit scoring models and we will discuss the advantages and disadvantages of each model. Finally, the methods will be applied to a case study.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΔΗ ΚΙΝΔΥΝΩΝ ΚΑΙ ΒΑΣΙΛΕΙΑ Ι & ΙΙ

Το χρηματοπιστωτικό σύστημα συνίστανται κυρίως από τις χρηματοπιστωτικές αγορές, τα χρηματοπιστωτικά προϊόντα και τα χρηματοπιστωτικά ιδρύματα. Πρόκειται, λοιπόν, για ένα σύνολο θεσμών και οικονομικών φορέων που έχουν ως βασικό σκοπό τη μεταφορά οικονομικών πόρων από τις πλεονασματικές οικονομικές μονάδες στις ελλειμματικές και αποτελείται από τη μετατροπή χρηματικών μέσων σε δανειακό κεφάλαιο. Είναι πολύ σημαντικό καθώς εκτελεί πληθώρα λειτουργιών με σκοπό την ανάπτυξη της οικονομίας κάθε χωράς με έναν ορθολογικό τρόπο. Βασικές λειτουργίες του είναι η διευκόλυνση και η μείωση του κόστους των συναλλαγών, η παραγωγή πληροφοριών, επιτυγχάνει την ορθολογικότερη κατανομή των διαθέσιμων κεφαλαίων και προωθεί νέα χρηματοπιστωτικά προϊόντα. Τέλος μία ακόμα πολύ σημαντική λειτουργία είναι αυτή της διαχείρισης και μείωσης του κινδύνου.

Το θεμέλιο του χρηματοπιστωτικού συστήματος είναι τα πιστωτικά ιδρύματα και κατά κύριο λόγο οι τράπεζες. Τα πιστωτικά ιδρύματα δέχονται καταθέσεις και διαθέτουν κεφάλαια μέσω του δανεισμού και των επενδυτικών δραστηριοτήτων στους ιδιώτες, τις επιχειρήσεις και τις κυβερνήσεις. Τα τελευταία χρόνια τα πιστωτικά ιδρύματα δεν έχουν μόνο τον ρόλο της μεταφοράς κεφαλαίων από τους χρηματοδότες στους χρηματοδοτούμενους, αλλά η λειτουργία τους περιλαμβάνει πλήθος λειτουργιών, παραγωγής και διαχείρισης χρηματοοικονομικών προϊόντων. Οι τράπεζες λοιπόν είναι οργανισμοί με μεγάλη και περίπλοκη λειτουργία και η έννοια του κινδύνου υπάρχει σχεδόν σε όλο το φάσμα των δραστηριοτήτων τους.

Οι τραπεζικοί κίνδυνοι είναι συνδεδεμένοι με οποιαδήποτε δραστηριότητα η οποία σχετίζεται με μελλοντικές εισροές και εκροές, των οποίων το καθαρό αποτέλεσμα είναι από μόνο του αβέβαιο. Επίσης τα οικονομικά μεγέθη των τραπεζών και άλλων χρηματοπιστωτικών ιδρυμάτων υπόκεινται σε πλήθος κινδύνων, ιδιαίτερα λόγω του διαμεσολαβητικού τους ρόλου. Αυτές οι πιθανολογούμενες οικονομικές ζημίες των χρηματοπιστωτικών ιδρυμάτων είναι κυρίως συνδεδεμένες με την αστάθεια των ταμειακών ροών, τις θέσεις και τις σχέσεις που αναπτύσσουν με τους πελάτες τους, καθώς και τις συναλλαγές που διενεργούν στις αγορές χρήματος και κεφαλαίου. Η

διάγνωση της μορφής κάθε κινδύνου, η εκτίμηση του μεγέθους της κάθε πιθανολογούμενης οικονομικής ζημίας που ενδέχεται να προκύψει και η ανάπτυξη μηχανισμών εξουδετέρωσης ή αντιστάθμισης, αποτελεί τον πυρήνα συγκεκριμένης ενασχόλησης εξειδικευμένου τομέα των χρηματοπιστωτικών ιδρυμάτων (risk management). Η αποτελεσματική διαχείριση του κινδύνου εμπεριέχει τον εντοπισμό και την κατανόηση των ειδών κινδύνων καθώς επίσης την εύρεση των πιο αποτελεσματικών μεθόδων ώστε να γίνει η κατάλληλη διαχείρισή τους. Παρόλο που ο κίνδυνος δεν είναι επιθυμητό στοιχείο, η ανάληψη του αποτελεί κίνητρο για την προσδοκία επίτευξης (μεγαλύτερου) κέρδους.

1.1 Τα κυριότερα είδη κινδύνων

Τα είδη των τραπεζικών κινδύνων είναι: Πιστωτικός κίνδυνος, Κίνδυνος επιτοκίου, Κίνδυνος αγοράς, Κίνδυνος ρευστότητας, Λειτουργικός κίνδυνος, Συναλλαγματικός κίνδυνος και άλλοι κίνδυνοι όπως ο επιχειρηματικός, κίνδυνος φήμης, κίνδυνος συμμόρφωσης κτλ.

α)Λειτουργικός Κίνδυνος

Ο κίνδυνος αυτός αφορά απώλειες ή ζημιές που προκαλούνται από ανεπαρκείς ή αποτυχημένες εσωτερικές διαδικασίες, από τους ανθρώπους και από εξωτερικά γεγονότα. Είναι κίνδυνος που δύσκολα μπορεί να προβλεφθεί και αυτό έχει και ως συνέπεια να η διαχείρισή του να είναι πολύ δύσκολη.

β)Κίνδυνος Αγοράς

Είναι ο κίνδυνος που αντιμετωπίζουν τα ΠΙ από τις διακυμάνσεις στις αξίες των στοιχείων του ενεργητικού τους. Σχετίζεται με τις ανεπιθύμητες μεταβολές των συναλλαγματικών ισοτιμιών, των επιτοκίων, των τιμών των μετοχών και γενικά των παραμέτρων της αγοράς. Όλοι όσοι συναλλάσσονται στην αγορά είναι εκτεθειμένοι στον κίνδυνο απώλειας τιμής δηλαδή μείωση της αξίας της επένδυσή της.

γ)Πιστωτικός κίνδυνος

Είναι ο πιο συχνός και ο πιο σοβαρός κίνδυνος ο οποίος σχετίζεται με τη φύση των δραστηριοτήτων μίας τράπεζας. Ουσιαστικά είναι ο κίνδυνος αθέτησης των υποχρεώσεων ενός δανειολήπτη. Πριν την χορήγηση οποιουδήποτε δανείου η τράπεζα μαζεύει τις απαραίτητες πληροφορίες ώστε να αξιολογήσει τον δανειστή, αυτός είναι ένας τρόπος υπολογισμού του πιστωτικού κινδύνου. Ο συγκεκριμένος κίνδυνος περιέχει 3 είδη.

Τον κίνδυνο αθέτησης(default risk), ο οποίος αφορά την πιθανότητα που αναλαμβάνει η τράπεζα, ο δανειολήπτης να μην είναι σε θέση να πραγματοποιήσει τις απαιτούμενες πληρωμές για την οφειλή του. Αυτό μπορεί να σημαίνει είτε καθυστέρηση στην αποπληρωμή μίας δόσης, μονοήμερη ακύρωση της συμφωνίας είτε και ολική αθέτηση

Τον κίνδυνο έκθεσης (exposure risk),ο οποίος είναι μια μέτρηση της μέγιστης πιθανής ζημίας για έναν δανειστή εάν ο δανειολήπτης αθετήσει την πληρωμή. Το μέγεθος, ο χρόνος και το εύρος της ζημίας είναι προφανώς δύσκολο να υπολογιστούν, οπότε γίνονται μόνο υποθέσεις ή προγνωστικά μοντέλα που βοηθάνε στη θέσπιση ορίων στην πιθανότητα απωλειών. Ένας τρόπος να περιοριστεί αυτός ο κίνδυνος είναι η λήψη εξασφαλίσεων για την κάλυψη του δανείου π.χ. υποθήκη.

Τέλος, τον κίνδυνο ανάκτησης (recovery risk), ο οποίος υποδηλώνει τον κίνδυνο ότι μετά από ένα γεγονός αθέτησης, οι υποχρεώσεις του δανειολήπτη δεν μπορεί να εκπληρωθούν πλήρως, οδηγώντας έτσι σε οικονομική ζημία για τον δανειστή.

Ο πιστωτικός κίνδυνος αποτελεί την μεγαλύτερη απειλή της φερεγγυότητας των τραπεζικών ιδρυμάτων, για αυτό και η διαχείριση του στοχεύει στην πρακτική του μετριασμού των ζημιών με την κατανόηση της επάρκειας των αποθεματικών κεφαλαίων και ζημιών από τα δάνεια μιας τράπεζας ανά πάσα στιγμή – μια διαδικασία που αποτελεί εδώ και καιρό πρόκληση για τα χρηματοπιστωτικά ιδρύματα.

1.2 Βασιλεία Ι

Όπως είδαμε και παραπάνω ο πιστωτικός κίνδυνος έχει άμεση σχέση με τη φερεγγυότητα των χρηματοπιστωτικών οργανισμών. Αυτό συμβαίνει κυρίως γιατί οι περισσότερες χρεοκοπίες τραπεζών έχουν προέλθει από τα προβλήματα που προκαλούν οι επισφαλείς απαιτήσεις στον ισολογισμό τους, οι οποίες προξενούν μείωση της κερδοφορίας τους και των αποθεματικών τους. Αυτόματα λοιπόν η πιστοληπτική τους ικανότητα μειώνεται και αυξάνεται το κόστος δανεισμού τους στη διατραπεζική αγορά και κάτι τέτοιο δημιουργεί σύγχυση σε όλο το τραπεζικό σύστημα με αντίκτυπο να μην υπάρχει εμπιστοσύνη από την μεριά του κοινού. Βλέποντας λοιπόν την σημασία και τα προβλήματα που φέρει ο πιστωτικό κίνδυνος ,θα πρέπει να υπάρχουν και οι κατάλληλοι κανόνες από τις εποπτικές αρχές των τραπεζικών συστημάτων, για να μπορεί να ελεγχθεί και να μετριαστεί με σκοπό τη σταθερότητα

και την εμπιστοσύνη του κοινού στο σύστημα. Οι κανόνες αυτοί είναι γνωστοί ως κανόνες της Βασιλείας.

Η Επιτροπή της Βασιλείας, που αρχικά ονομάστηκε Επιτροπή Τραπεζικών Κανονισμών και Εποπτικών Πρακτικών, ιδρύθηκε από τους Διοικητές των κεντρικών τραπεζών της Ομάδας των 10 χωρών στα τέλη του 1974 μετά από σοβαρές διαταραχές στις διεθνείς αγορές συναλλάγματος και τραπεζών (κυρίως η αποτυχία του Bankhaus Herstatt στη Δυτική Γερμανία). Η Επιτροπή, με έδρα την Τράπεζα Διεθνών Διακανονισμών στη Βασιλεία, ιδρύθηκε για να ενισχύσει τη χρηματοπιστωτική σταθερότητα μέσω της βελτίωσης της ποιότητας της τραπεζικής εποπτείας παγκοσμίως και να χρησιμεύσει ως φόρουμ τακτικής συνεργασίας μεταξύ των χωρών μελών της σε θέματα τραπεζικής εποπτείας. Η ισχύς της δεν είναι νομική ούτε δεσμευτική, κατά κύριο λόγο συνιστά στα μέλη να εφαρμόσουν αρχές και να ενσωματώσουν κανόνες οι οποίοι είναι κατάλληλοι για αυτά σε εθνικό επίπεδο.

Το επίκεντρο των δραστηριοτήτων της Βασιλείας Ι είναι η προώθηση μεθόδων υπολογισμού της κεφαλαιακή επάρκειας ,δηλαδή τα εποπτικά κεφάλαια τα οποία οι τράπεζες πρέπει να διατηρούν έτσι ώστε να περιορίσουν τον κίνδυνο αθέτησης υποχρεώσεων. Αρχικά καθιερώθηκε ελάχιστη τιμή 8% για το δείκτη κεφαλαιακής επάρκειας ή συντελεστή κεφαλαιακής επάρκειας ή συντελεστή φερεγγυότητας (Capital adequacy ratio - CAR) ο οποίος ορίστηκε ως ο λόγος των ιδίων κεφαλαίων (εποπτικά κεφάλαια)προς τα στοιχεία του ενεργητικού και τα εκτός ισολογισμού στοιχεία, σταθμισμένα σύμφωνα με τον (πιστωτικό) κίνδυνό τους.(Κούτρας, 2020)

$$\text{Minimum Capital Requirements} \geq 8\% \times \text{RWA}.$$

Όταν η Βασιλεία Ι έγινε το 1988 οι οικονομικές συναλλαγές δεν ήταν σύνθετες. Τα επόμενα χρόνια όμως, τα οικονομικά περιβάλλοντα σε όλο τον κόσμο εξελίχθηκαν. Δημιουργήθηκαν νεότερα χρηματοπιστωτικά ιδρύματα και παρουσιάστηκαν περισσότερα καινοτόμα προϊόντα και υπηρεσίες, οπότε η φύση των οικονομικών κινδύνων άρχισε να αλλάζει. Το βασικότερο μειονέκτημα της Βασιλείας Ι ήταν η έλλειψη ευαισθησίας στον κίνδυνο και επικεντρώθηκε σε βασικές μετρήσεις χρηματοοικονομικού κινδύνου αγνοώντας εντελώς την ανάγκη για μια ισχυρή διαδικασία διαχείρισης κινδύνου.

1.3 Βασιλεία ΙΙ

Οι αλλαγές , οι εξελίξεις και ο ανταγωνισμός στο διεθνές τραπεζικό σύστημα έφερε την επιτακτική ανάγκη, το πλαίσιο της Βασιλείας, να αναθεωρήσει τους κανόνες και τις αρχές που είχε θεσπίσει καθώς σταμάτησε να ανταποκρίνεται με αποτέλεσμα στους κινδύνους που εκτίθονταν οι τράπεζες. Έτσι η επιτροπή της Βασιλείας ξεκίνησε διαδικασία διαβούλευσης και αναθεώρησης του υπάρχοντος Κειμένου και παράθεσε νέες προτάσεις στο πλαίσιο της κεφαλαιακής επάρκειας, τη Βασιλεία ΙΙ. Η Βασιλεία ΙΙ βασίζεται σημαντικά στη Βασιλεία Ι αυξάνοντας την ευαισθησία του κεφαλαίου σε βασικούς τραπεζικούς κινδύνους. Επιπλέον, η Βασιλεία ΙΙ αναγνωρίζει ότι οι τράπεζες μπορούν να αντιμετωπίσουν πολλούς κινδύνους πέρα του πιστωτικού γι' αυτό και εισάγονται μέσα ο λειτουργικός και ο κίνδυνος αγοράς. Το νέο σύμφωνο της Βασιλείας ΙΙ ενσωματώνει τρεις πυλώνες οι οποίοι αλληλοϋποστηρίζονται επιτρέποντας τις τράπεζες και τις εποπτικές αρχές να σταθμίζουν καλύτερα τους κινδύνους που διαχειρίζονται.

- **Πρώτος πυλώνας:** Ελάχιστες κεφαλαιακές απαιτήσεις για την κάλυψη των αναλαμβανόμενων κινδύνων. Ορίζει πως οι τράπεζες πρέπει να υπολογίζουν τις κεφαλαιακές απαιτήσεις ανάλογα με τους κινδύνους που αντιμετωπίζουν στο πλαίσιο της άσκησης των δραστηριοτήτων τους. Ο δείκτης της κεφαλαιακής επάρκειας(ΔKE) διαμορφώνεται πλέον ως :

$$\Delta KE = \frac{Εποπτικά\ ίδια\ Κεφάλαια}{ΠK_{ΣE} + 12{,}5 * (ΛK + KA)} \geq 8\%$$

όπου:

$ΠK_{ΣE}$:Το σταθμισμένο , κατά το πιστωτικό κίνδυνο Ενεργητικό

$ΛK$: Κεφαλαιακές απαιτήσεις για τον Λειτουργικό Κίνδυνο

KA: Κεφαλαιακές απαιτήσεις για τον Κίνδυνο Αγοράς (Ζοπουνίδης Κ. και Λεμονάκης Χ, 2009)

- **Δεύτερος πυλώνας:** Διαδικασίες εποπτικής επιθεώρησης σχετικά με την κεφαλαιακή επάρκεια των τραπεζών. Πλέον με το νέο σύμφωνο η κάθε τράπεζα είναι υπεύθυνη για την διαχείριση των κινδύνων της, παρόλα αυτά βέβαια οι εποπτικές αρχές μπορούν να παίξουν ρόλο στην αξιολόγηση των πρακτικών διαχείρισης των κινδύνων και να διασφαλίζουν ότι οι αρνητικές εξωτερικές επιδράσεις που μπορεί να προκύψουν από την αποτυχία μιας τράπεζας μπορούν να ελαχιστοποιηθούν και να διαχειριστούν.

- **Τρίτος πυλώνας:** Πειθαρχία μέσω της αγοράς. Οι συμμετέχοντες στην αγορά έχουν ανάγκη να διασφαλίσουν ότι οι τράπεζες διαθέτουν τα κατάλληλα κεφάλαια και ανάλογα με το ενδιαφέρον και των ενεργειών τους, τις ενθαρρύνουν να συμπεριφέρονται συνετά προς τους καταναλωτές. Ένα βασικό συστατικό για την προώθηση της πειθαρχίας της αγοράς σε αυτό πλαίσιο είναι η διασφάλιση ότι οι πελάτες των τραπεζών, τα ιδρύματα και άλλοι συμμετέχοντες στην αγορά έχουν πρόσβαση στις κατάλληλες πληροφορίες που τους επιτρέπουν να παρακολουθούν τραπεζική απόδοση και ανάληψη κινδύνων. Ο Πυλώνας 3 το επιτυγχάνει αυτό απαιτώντας από τις τράπεζες να αποκαλύπτουν, σε τακτική βάση, τις ποσοτικές και ποιοτικές πληροφορίες σχετικά με την τη φύση των κινδύνων τους, τις διαδικασίες μέτρησης τους και την κεφαλαιακή επάρκεια.

Η Βασιλεία ΙΙ δίνει μεγαλύτερη ελευθερία στις τράπεζές να ακολουθούν αρχές οι οποίες ταιριάζουν με τη φύση των δραστηριοτήτων τους και εγκαταλείπει την αντίληψη ότι όλοι οι εποπτικοί κανόνες πρέπει να είναι ίδιοι για όλους. Τέλος προωθεί την ιδέα ότι όσο καλύτερα ένα χρηματοπιστωτικό ίδρυμα εκτιμά και διαχειρίζεται τους κινδύνους του τόσο λιγότερο κεφάλαιο θα χρειάζεται να διακρατά.

1.4 Βασιλεία ΙΙ και Πιστωτικός Κίνδυνος

Σύμφωνα με το κανονιστικό πλαίσιο της Βασιλείας ΙΙ, για να εκτιμηθεί ο πιστωτικός κίνδυνος προτείνονται δύο προσεγγίσεις. Η πρώτη είναι η τυποποιημένη μέθοδος (standardize approach) και η άλλη είναι αυτή των εσωτερικών διαβαθμίσεων (Internal Rating based approach-IRB), η οποία χωρίζεται στην Θεμελιώδη (Foundation IRB) και στην Εξελιγμένη (Advanced). Η τυποποιημένη προσέγγιση βασίζεται στην προσέγγιση της Βασιλείας Ι και είναι η βασική επιλογή για τον καθορισμό των ελάχιστων κεφαλαιακών απαιτήσεων. Η συγκεκριμένη μέθοδος διατηρεί τον απλό χαρακτήρα που έχει η Βασιλεία Ι, ενώ ταυτόχρονα αυξάνει την ευαισθησία των εποπτικών κεφαλαιακών απαιτήσεων ως προς τον κίνδυνο. Για να αυξηθεί λοιπόν αυτή η ευαισθησία χρησιμοποιούνται αξιολογήσεις πιστοληπτικής ικανότητας από επιλέξιμους οίκους αξιολόγησης, ώστε να αυξηθεί ο αριθμός των κατηγοριών στάθμισης κινδύνου που εφαρμόζονται στα υποκείμενα περιουσιακά στοιχεία.

Οι προσεγγίσεις των εσωτερικών διαβαθμίσεων(μοντέλα IRB) αποτελούν τη σημαντικότερη εξέλιξη του Συμφώνου για τον υπολογισμό των ελάχιστων κεφαλαιακών απαιτήσεων που χρειάζεται να έχει ένα πιστωτικό ίδρυμα για την ορθή λειτουργία του και την διασφάλιση της πελατείας του. Τα μοντέλα IRB βασίζονται ως επί το πλείστον στα συστήματα που διαθέτουν οι ίδιες οι τράπεζες, ώστε να κατατάξουν σε διακριτές κατηγορίες τους πελάτες, έναντι του πιστωτικού κινδύνου, με βάση την εκτιμώμενη πιθανότητα αθέτησης. Σύμφωνα με το ΠΔ.ΤΕ/2589 τα συστήματα, οι διαδικασίες και οι εσωτερικές διαβαθμίσεις, καθώς και οι εκτιμήσεις αθέτησης και ζημίας που χρησιμοποιούνται στον υπολογισμό των σταθμισμένων ανοιγμάτων παίζουν ουσιαστικό ρόλο, τόσο στη διαχείριση κινδύνου και τη λήψη αποφάσεων, όσο και στις λειτουργίες του πιστωτικού ιδρύματος, που αφορούν στην έγκριση πιστώσεων, την κατανομή του εσωτερικού κεφαλαίου και την εταιρική διακυβέρνηση. Για να μπορεί μία τράπεζα να αναπτύξει τα συγκεκριμένα μοντέλα θα πρέπει να έχει πάρει ειδική άδεια από την εποπτική αρχή της χώρας της, συγκεκριμένα στην Ελλάδα από την Τράπεζα της Ελλάδος, η οποία πιστοποίει ότι τα συστήματα που χρησιμοποιεί για την κατηγοριοποίηση των ανοιγμάτων εφαρμόζονται βάσει των κανόνων της συγκεκριμένης μεθόδου.

Για τον προσδιορισμό των ελάχιστων κεφαλαιακών απαιτήσεων για την κάλυψη των πιστωτικών κινδύνων, οι τράπεζες πρέπει να κατηγοριοποιήσουν το ενεργητικό του ισολογισμού τους σε πέντε μεγάλες ομάδες οι οποίες είναι :

1. Τραπεζικές απαιτήσεις έναντι επιχειρήσεων (corporate risk).
2. Τραπεζικές απαιτήσεις έναντι κρατών (sovereign risk).
3 Τραπεζικές απαιτήσεις έναντι εμπορικών τραπεζών (bank risk) αλλά και εταιρειών του χρηματοπιστωτικού τομέα.
4. Τραπεζικές απαιτήσεις έναντι λιανικής τραπεζικής (retail risk).
5. Τραπεζικές απαιτήσεις έναντι χρηματοοικονομικών αξιών (equity risk).

Οι σημαντικότερες παράμετροι οι οποίοι χρησιμοποιούνται για τον υπολογισμό των σταθμίσεων του κινδύνου είναι :

- PD (probability of default), είναι η πιο σημαντική παράμετρος του πιστωτικού κινδύνου και είναι ουσιαστικά η πιθανότητα ο αντισυμβαλλόμενος να μην μπορέσει να καλύψει τις υποχρεώσεις του μία χρονική στιγμή. Ουσιαστικά μέσα από αυτήν απεικονίζεται η πιστοληπτική του ικανότητα.

- LGD (Loss given default), είναι το μέγεθος της δυνητικής ζημιάς από το άνοιγμα, δηλαδή το ποσό που δεν θα λάβει η τράπεζα λόγω της αθέτησης του πελάτη
- EAD (exposure at default), πρόκειται για το προβλεπόμενο ποσό της ζημίας που μπορεί να αντιμετωπίσει μια τράπεζα σε περίπτωση της αθέτησης των υποχρεώσεων του δανειολήπτη. Η ζημία εξαρτάται από το ποσό στο οποίο ήταν εκτεθειμένη η τράπεζα έναντι του δανειολήπτη κατά τη στιγμή της αθέτησης, καθώς η αθέτηση επέρχεται σε άγνωστη μελλοντική ημερομηνία.
- M (maturity), είναι η εναπομείνασα διάρκεια μέχρι την λήξη των απαιτήσεων.

Η βασική διαφορά των δύο προσεγγίσεων(Θεμελιώδης και Εξελιγμένη) είναι ότι στην Θεμελιώδη Προσέγγιση οι τιμές της PD, καθορίζονται από την ίδια την τράπεζα ενώ οι υπόλοιπες παράμετροι παραχωρούνται από την επιτροπή, ενώ στην Εξελιγμένη τόσο η PD όσο και οι LGD , EAD , M εκτιμώνται από τα ίδια τα πιστωτικά ιδρύματα, αρκεί προφανώς να πληρούν τους κανόνες της Βασιλείας ΙΙ. Οι συνιστώσες αυτές (PD, LGD, EAD, M)αποτελούν τις βασικές εισροές στην προσέγγιση IRB και κατά συνέπεια, τις κεφαλαιακές απαιτήσεις που απορρέουν από αυτήν, ως εκ τούτου, οι περισσότερες πτυχές του IRB πλαισίου έχουν σχεδιαστεί για να παρέχουν εμπιστοσύνη ότι τα στοιχεία αυτά είναι αναγνωρίσιμα, μετρήσιμα και ικανά να επαληθεύονται τόσο από τις τράπεζες όσο και από τις εποπτικές αρχές. Παρόλες τις διαφορές που έχουν οι δύο προσεγγίσεις το σύμφωνο της Βασιλείας για τον υπολογισμό του σταθμισμένου στον κίνδυνο ενεργητικού(RWA) χρησιμοποιεί τον τύπο :

$$RWA = 1{,}25 \times K \times EAD$$

όπου το K είναι η κεφαλαιακή απαίτηση και το EAD εκφράζεται σε μία χρηματική μονάδα μέτρησης π.χ. ευρώ. Επιπλέον για τον υπολογισμό K στην λιανική τραπεζική ο τύπος που χρησιμοποιείται είναι :

$$K = LGD * \Phi[\frac{1}{(1-R)^{\frac{1}{2}}} * \Phi^{-1}(PD) + \sqrt{\left(\frac{1}{1-R}\right)} * \Phi^{-1}(0{,}999)] - PD * LGD$$

όπου Φ είναι η αθροιστική συνάρτηση πιθανότητας της κανονικής κατανομής και Φ^{-1} είναι η αντίστροφη αυτής και R είναι ο συντελεστής συσχέτισης της απόδοσης διαφόρων ανοιγμάτων.

Όπως λοιπόν καταλαβαίνουμε οι παράμετροι PD, EAD, LGD,είναι πολύ σημαντικοί για την εκτίμηση των κεφαλαιακών απαιτήσεων ενός πιστωτικού ιδρύματος και πώς αυτά επηρεάζουν την φερεγγυότητα του αλλά και την κερδοφορία του. Στη συνέχεια λοιπόν θα ασχοληθούμε με προτάσεις για τη μεθοδολογία υπολογισμού της πιθανότητας αθέτησης (PD) με τη χρήση μοντέλων πιστοληπτικής διαβάθμισης (Credit Scoring Models) όπως προτείνει και το κείμενο της Βασιλείας ΙΙ.

ΚΕΦΑΛΑΙΟ 2

ΜΟΝΤΕΛΑ ΒΑΘΜΟΛΟΓΗΣΗΣ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ

Όπως έχουμε δει και πιο πάνω βασικές λειτουργίες κάθε χρηματοπιστωτικού ιδρύματος είναι η παροχή πιστώσεων σε δανειολήπτες, η δημιουργία δανείων και πιστωτικών περιουσιακών στοιχείων. Αυτό λοιπόν μας δείχνει ότι ένα μεγάλο μέρος του κινδύνου μιας τράπεζας έγκειται στην ποιότητα των περιουσιακών στοιχείων η οποία θα πρέπει να είναι σύμφωνη με την διάθεση που έχει για ανάληψη και έκθεση στον κίνδυνο. Η τράπεζα λοιπόν για να μπορέσει να είναι επικερδής θα πρέπει να διαθέτει κατάλληλα και προηγμένα εργαλεία για την αποτελεσματικότερη διαχείριση και ποσοτικοποίηση του πιστωτικού κινδύνου. Λόγω λοιπόν του μεγάλου όγκου πελατών και την ανάγκη που είχαν οι τράπεζες να αποφέρουν κέρδη και να είναι φερέγγυες ξεκίνησαν να χρησιμοποιούνται τα μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας(credit scoring models,CSM). Πρόκειται για στατιστικά μοντέλα που στόχο έχουν να προβλεφθεί η μελλοντική συμπεριφορά υφιστάμενων ή μελλοντικών πελατών και την κατάταξη τους σε «καλούς» ή «κακούς» που προκύπτουν από την ανάλυση στατιστικών δειγμάτων.

Η μέτρηση του πιστωτικού κινδύνου μέσω των σχετικών μοντέλων είναι σημαντική όχι μόνο για την διαχείριση κινδύνου αλλά και για την ανάπτυξη της επιχείρησης μέσω του εντοπισμού των κατάλληλων πελατών, της βελτίωση της εμπειρίας του πελάτη, της ελαχιστοποίησης του χρόνου επεξεργασίας κ.α. Την σημερινή εποχή διάφορα χρηματοπιστωτικά ιδρύματα και κυρίως τα μεγαλύτερα από αυτά βασίζονται στα μοντέλα αυτά αφού πέρα των άλλων συμβάλλουν στην καλύτερη διαχείριση της σχέσης πελάτη και τράπεζας.

Σύμφωνα με τον επενδυτικό οίκο Standard and Poors η μέτρηση του πιστωτικού κινδύνου συνιστά την διαδικασία αξιολόγησης του πιστωτικού κινδύνου, την ποιότητας της πίστωσης, και ικανότητας και προθυμίας των δανειστών του να ανταποκριθούν στις δανειακές τους υποχρεώσεις. Οι τελευταίες συνίστανται στη πληρωμή των τόκων ή τόκο χρεολυσίων ανάλογα βέβαια και με τους όρους του δανείου.

2.1 Κατηγορίες και χρησιμότητα των CSM

Τα CSM, μπορούν να διαχωριστούν σε διάφορες κατηγορίες ανάλογα με το στάδιο πιστωτικού κύκλου στον οποίο χρησιμοποιούνται. Κάποιες από τα πιο βασικές κατηγορίες είναι τα μοντέλα βαθμολόγησης αιτήσεων, τα μοντέλα βαθμολόγησης συμπεριφοράς και τα μοντέλα βαθμολόγησης κέρδους.Η κύρια διαφορά μεταξύ τους συνίστανται στο σύνολο των μεταβλητών οι οποίες χρησιμοποιούνται για την μέτρηση της πιστοληπτικής ικανότητας του πελάτη, π.χ. όσο νωρίτερα βρισκόμαστε στον πιστωτικό κύκλο του πελάτη τόσο μικρότερος είναι ο αριθμός των πληροφοριών που έχει στη διάθεσή της η τράπεζα. Αυτό γενικά σημαίνει ότι τα μοντέλα βαθμολόγησης αιτήσεων έχουν χαμηλότερη ισχύς πρόβλεψης από τα μοντέλα συμπεριφοράς και κέρδους

Τα μοντέλα βαθμολόγησης αιτήσεων έχουν ως σκοπό την απόφαση για το αν σε ένα πελάτη θα χορηγηθεί ή όχι ένα δάνειο. Το κάθε χρηματοπιστωτικό ίδρυμα διαθέτει συστήματα τα οποία αξιολογούν χαρακτηριστικά πελατών τα οποία έχουν συμπληρωθεί από τους ίδιους τους πελάτες κατά την αίτηση τους για χορήγηση δανείων. Οι τράπεζες λοιπόν έχοντας σαν γνώμονα παλιότερους πελάτες που αιτήθηκαν δάνειο και συγκρίνοντας αυτά τα χαρακτηριστικά , κατατάσσουν τους νέους υποψήφιους πελάτες βάσει του score που έχουν λάβει από την αίτηση σε μία κλίμακα, η οποία καθορίζει αν θα τους δοθεί το δάνειο ή όχι. Ο καθορισμός αυτός ορίζεται με βάση το σημείο αποκοπής (cut off) και σε κάθε τράπεζα είναι διαφορετικός. Το σημείο αποκοπής το έχει επιλέξει το κάθε ίδρυμα βάσει της στρατηγικής του, την όρεξη που έχει για κίνδυνο καθώς και τα αναμενόμενα κέρδη.

Όπως καταλαβαίνουμε η γενική ιδέα των μοντέλων βαθμολόγησης αιτήσεων(ACSM) είναι να προβλέψει κατά πόσο ένας πελάτης θα αθετήσει ή όχι τις υποχρεώσεις του βάσει σύγκρισης στοιχείων παλαιότερων ατόμων που αιτήθηκαν δανειοδότηση. Σημαντικό ρόλο για την ανάπτυξη των κατάλληλων στατιστικών μοντέλων παίζουν τα παρελθοντικά δεδομένα. Τα ACSM, είναι ένα χρήσιμο εργαλείο για τους τραπεζικούς οργανισμούς, διότι τα αυτοματοποιημένα μοντέλα αξιολόγησης κατατάσσουν αυτόματα και άμεσα μεγάλο όγκο υποψηφίους εξοικονομώντας χρόνο σε μία εποχή που η αγορά έχει μεγάλο ανταγωνισμό και κινείται σε πολύ γρήγορους ρυθμούς.

Η χρησιμότητα των μοντέλων μέτρησης πιστωτικού κινδύνου έγκειται στα παρακάτω. Πρώτον βελτιώνουν την διαδικασία επιλογής των δανειοληπτών αφού το να κάνουν λάθος επιλογή με το να δανείσουν στον λάθος πελάτη συνεπάγεται την

δημιουργία σημαντικής μελλοντικής ζημίας. Δεύτερον, συμβάλλουν στην μείωση του κινδύνου συνεισφέροντας έτσι στην σταθεροποίηση του τραπεζικού συστήματος αφού μια λανθασμένη απόφαση λόγω αλληλεξάρτησης των τραπεζών, συνεπάγεται αρνητικές επιπτώσεις τις οποίες θα βιώσει όλο το τραπεζικό σύστημα. Τρίτον υποστηρίζουν την ταξινόμηση των δανείων ενώ προλαμβάνεται ο κίνδυνος. Τέταρτον δημιουργούνται πολιτικές διαχείρισης των πελατών όπου δημιουργούνται γκρουπ από αυτούς ανάλογα με το σκορ που έχουν λάβει. Πέμπτων ανάλογα με τον κίνδυνο του πελάτη θα υπάρξει και η ανάλογη χρέωση του επιτοκίου διαμορφώνοντας έτσι μια τιμολογιακή πολιτική (Huyen,2011).

Σύμφωνα με τους (Alinejad, 2013), τα πλεονεκτήματα της μέτρησης του πιστωτικού κινδύνου είναι ότι μειώνεται το κόστος της τράπεζας, γίνεται μια καλύτερη εκτίμηση του κινδύνου, διαχωρίζονται οι καλοί από τους κακούς πελάτες, μειώνεται η πιθανότητα αποσταθεροποίησης του τραπεζικού συστήματος και παρέχονται καλύτερα χρηματοπιστωτικά προϊόντα. Τα μειονεκτήματα είναι ότι αγνοούνται κάποιες πτυχές της διαδικασίας χορήγησης δανείων αφού είναι πιθανόν να αποκλειστούν φερέγγυοι πελάτες, τίθεται ζήτημα ασφάλειας προσωπικών δεδομένων και δεν υπάρχει ευελιξία στην όλη διαδικασία αφού γίνεται χρήση παρελθοντικών στοιχείων, κυρίως αν τα δεδομένα για τους εκάστοτε πελάτες μπορεί να αλλάξουν είτε προς το καλύτερο είτε προς το χειρότερο μελλοντικά.

Υπάρχουν διάφοροι παράγοντες που επηρεάζουν τον πιστωτικό κίνδυνο και οι οποίοι αναγράφονται στις αιτήσεις των δανείων όπως η οικονομική κατάσταση του υποψήφιου δανειολήπτη, το ποσό που ζητείται ως δάνειο, η οικογενειακή κατάσταση, η ανάλυση των ταμειακών ροών, το είδος του ενέχυρου κ.α. Ο κίνδυνος συνεπώς υπολογίζεται με την χρήση ποιοτικής και ποσοτικής ανάλυσης. Πολύ σημαντικοί παράγοντες είναι και οι παρελθούσες καταναλωτικές συμπεριφορές οι οποίες επηρεάζουν αρνητικά την βαθμολογία όσον αφορά τον πιστωτικό κίνδυνο όπως η καθυστέρηση πληρωμή δόσεων, η υπέρβαση του ορίου των πιστωτικών καρτών, η έλλειψη πιστωτικής ιστορίας κ.α. (Comarch, 2019)

Όπως αναφέραμε και προηγουμένως η βαθμολόγηση / ποσοτικοποίηση του πιστωτικού κινδύνου έχει ως στόχο να υπολογίσει τον κίνδυνο αθέτησης της υποχρέωσης από πλευράς δανειολήπτη και τη μείωση του κινδύνου έκθεσης της τράπεζας. Πάντως τα τελευταία χρόνια τα χρηματοπιστωτικά ιδρύματα εστιάζουν και σε άλλους παράγοντες όπως τη μεγιστοποίηση της κερδοφορίας, την αύξηση της

αποτελεσματικότητας των εκστρατειών marketing, την βαθμολόγηση της απάτης, το ποσοστό απώλειας των πελατών, την βελτίωση της διαχείρισης χρέους κ.α.

Όταν οι τράπεζες καθορίζουν το μέγεθος του πιστωτικού κινδύνου που θέλουν να αναλάβουν τότε υπολογίζουν την πιθανότητα αθέτησης του δανειολήπτη, την ζημία που θα υποστεί το χρηματοπιστωτικό σύστημα αν ο δανειολήπτης αθετήσει την υποχρέωση του και το ποσό ζημίας στο οποίο είναι εκτεθειμένη η τράπεζα. Οι τράπεζες μπορούν να ελέγχουν τον πιστωτικό κίνδυνο αλλά δεν μπορούν να προστατευτούν πλήρως από αυτόν άλλωστε τα μοντέλα που αναφέραμε πέρα από την χρησιμότητά τους παρουσιάζουν και κάποια μειονεκτήματα ενώ δεν πρέπει να λησμονούμε ότι η εκτίμησης του κινδύνου πριν το ξέσπασμα της κρίσης δεν ήταν επιτυχής.

Τελευταία, η βαθμολόγηση της πιστοληπτικής ικανότητας αποκτά νέα σημασία με τη νέα συμφωνία της Βασιλείας για το κεφάλαιο. Η λεγόμενη Βασιλεία ΙΙ αντικαθιστά την ισχύουσα Βασιλεία Ι και επικεντρώνεται σε τεχνικές που επιτρέπουν στις τράπεζες και τις εποπτικές αρχές να αξιολογούν σωστά τους διάφορους κινδύνους που αντιμετωπίζουν οι τράπεζες. Δεδομένου ότι η βαθμολόγηση της πιστοληπτικής ικανότητας συμβάλλει ευρέως στον εσωτερικό κίνδυνο διαδικασία αξιολόγησης ενός ιδρύματος, οι ρυθμιστικές αρχές έχουν επιβάλει αυστηρότερους κανόνες σχετικά με την ανάπτυξη, την εφαρμογή και την επικύρωση μοντέλων που πρέπει να ακολουθούνται από τις τράπεζες που επιθυμούν να χρησιμοποιήσουν τα εσωτερικά τους υποδείγματα για την εκτίμηση των κεφαλαιακών απαιτήσεων. Σε αυτό το σημείο βλέπουμε ότι κυρίως στη λιανική τραπεζική τα CSM, ανήκουν στην κατηγορία των IRB.

Ο κύκλος ζωής των μοντέλων πιστοληπτικής ικανότητας χωρίζεται σε διάφορες στάδια (αξιολόγηση, εφαρμογή, επικύρωση) και οι εποπτικές αρχές έχουν δημοσιεύσει συγκεκριμένες προϋποθέσεις που θα πρέπει να τηρούνται για κάθε ένα από αυτά.

2.2 Αξιολόγηση μοντέλου

Σύμφωνα με αυτά που είπαμε πιο πάνω μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας χρησιμοποιούν παρελθοντικά δεδομένα τα οποία έχουν συλλεχθεί από χαρακτηριστικά υφιστάμενων πελάτων. Η βασική ιδέα είναι ότι, αν κάποιος πελάτης με κάποια συγκεκριμένα χαρακτηριστικά ακολουθήσε ένα μοτίβο συμπεριφοράς στο παρελθόν, τότε και ο νέος πελάτης με παρόμοια χαρακτηριστικά θα ακολουθήσει την ίδια συμπεριφορά. Σημαντικό λοιπόν για την τράπεζα προκειμένου να αναπτύξει το

κατάλληλο μοντέλο, είναι να συλλέξει ένα μεγάλο δείγμα δεδομένων αιτούντων του παρελθόντος ή πελατών που αφορούν το προϊόν για το οποίο θέλει να χρησιμοποιήσει το μοντέλο βαθμολόγησης Ένα χρονικό διάστημα στο οποίο ένα δείγμα μπορεί να χρησθεί στατιστικά ωφέλιμο για τον έλεγχο της απόδοσης ενός πελάτη είναι τουλάχιστον ένας χρόνος. Απόδοση του πελάτη ορίζεται ουσιαστικά η πιθανότητα αθέτησης ή όχι που συνδέεται με αυτόν. Στα στατιστικά μοντέλα που χρησιμοποιούνται η εξαρτημένη μεταβλητή είναι ουσιαστικά η απόδοση του πελάτη και τα χαρακτηριστικά του πελάτη είναι οι προγνωστικοί παράγοντες. Γενικά το σύνολο των χαρακτηριστικών που πρόκειται να χρησιμοποιηθούν στα μοντέλα εξαρτάται από την κατηγορία μοντέλου που θέλουμε να χρησιμοποιήσουμε πχ στα μοντέλα βαθμολόγησης αιτήσεων που αποφασίζουν αν ο αιτών θα δανειοδοτηθεί ή όχι, βασίζονται στα προσωπικά στοιχεία του πελάτη καθώς αυτά είναι τα μόνα που έχει η τράπεζα στη διάθεσή της σε εκείνο το στάδιο. Έτσι λοιπόν αφού αναπτυχθεί το μοντέλο, δοκιμάζεται σε ένα δείγμα δοκιμής, για να επιβεβαιωθεί η εγκυρότητα των αποτελεσμάτων του.

2.3 Εφαρμογή του μοντέλου

Ένα από τα μεγαλύτερα πλεονέκτημα των μοντέλων βαθμολόγησης είναι ότι οι τράπεζες με αυτοματοποιημένες διαδικασίες μπορούν να παίρνουν αποφάσεις για τη διαχείριση των πελατών τους. Το σημείο αποκοπής (cut off) είναι αυτό που καθορίζει το αν ένας πελάτης θα πάρει το δάνειο ή όχι, οπότε αποτελεί μεγάλη πρόκληση να οριστεί ποιο θα είναι το κατώτατο όριο για κάθε μοντέλο βαθμολόγησης. Το βέλτιστο σημείο αποκοπής για να μπορέσει να βρεθεί χρειάζεται μελέτη, συνεκτίμηση των ιδιαιτεροτήτων κάθε τράπεζας (πχ στόχοι κέρδους, ανοχή στον κίνδυνο κλπ.). Η εξέλιξη των συστημάτων τεχνολογίας έχει βοηθήσει στην επέκταση των στρατηγικών που εξαρτιόνται από τα μοντέλα βαθμολόγησης. Οι τράπεζες είναι σε θέση να παρακολουθούν όλο τον κύκλο ζωής του πελάτη με μηνιαία επικαιροποιημένη βαθμολογία όπου υπολογίζεται από διαφορετικές κάρτες βαθμολογίας ο οποίες σχετίζονται με τη φάση του πιστωτικού κύκλου στην οποία βρίσκεται ο πελάτης. Τέλος πολλές καμπάνιες Marketing , εκστρατείες(π.χ. cross-selling, up-selling) έχουν σαν καθοδηγητή τα αποτελέσματα των μοντέλων βαθμολόγησης, τα οποία στοχεύουν στην αύξηση του κέρδους της κάθε τράπεζας.

2.4 Επικύρωση μοντέλου

Όπως είπαμε και πιο πάνω από τη στιγμή που τα credit score models εντάσσονται στα μοντέλα IRB σύμφωνα με τη Βασιλεία ΙΙ, η κάθε τράπεζα για να τα θέσει σε εφαρμογή και να υιοθετήσει την κουλτούρα τους, θα πρέπει να θεσπίσει ένα κύκλο επικύρωσης των μοντέλων στο οποίο θα υπάρχει συνεχής παρακολούθηση της απόδοσης και σταθερότητάς του, ελέγχει τις σχέσεις του υποδείγματος και δοκιμάζει τα αποτελέσματα του υποδείγματος έναντι άλλων αποτελεσμάτων(back testing). Γενικά λόγω του ότι ο κύκλος ζωής των CSM είναι αρκετά μικρός, εξαιτίας της μεταβλητότητας των αγορών, οι τράπεζες ούτως η άλλως είχαν κάποιες διαδικασίες επικύρωσης. Η Βασιλεία ΙΙ έδωσε μία πιο επίσημη μορφή και θέσπισε το ότι η επικύρωση των μοντέλων θα πρέπει να γίνεται από ομάδες ανεξάρτητες από αυτές που έχουν δημιουργήσει τα μοντέλα. Η σταθερότητα και η απόδοση είναι δύο στοιχεία πολύ σημαντικά για την ποιότητα των μοντέλων βαθμολόγησης, οπότε θα πρέπει να παρακολουθούνται και να αναλύονται πάρα πολύ συχνά, κυρίως όταν ακόμα και μικρές αλλαγές να γίνουν στα χαρακτηριστικά των πελάτων μπορούν να αλλοιώσουν τα αποτέλεσμα επιλογής.

Η δοκιμή και η συγκριτική αξιολόγηση είναι δύο βασικά στοιχεία για την επικύρωση των μοντέλων βαθμολόγησης. Με την δοκιμή(back testing) αξιολογείται η βαθμονόμηση, η οποία αντιστοιχεί τη βαθμολογία με ένα ποσοτικό μέτρο κινδύνου και η διάκριση, η οποία είναι πόσο καλά το μοντέλο παρέχει μία διαδοχική κατάταξη του προφίλ κινδύνου των παρατηρήσεων του δείγματος. Όσον αφορά τη συγκριτική αξιολόγηση, αποσκοπεί στην συνοχή των εκτιμώμενων μοντέλων βαθμολόγησης συγκριτικά με εκείνα που προέκυψαν με τη χρήση άλλων τεχνικών εκτίμησης, και ενδεχομένως χρησιμοποιήθηκαν και άλλες πηγές δεδομένων. Παρόλα αυτά, αυτή η ανάλυση είναι δύσκολο να εφαρμοστεί σε χαρτοφυλάκια λιανικής καθώς υπάρχει έλλειψη γενικών δεικτών αναφοράς στην αγορά.

Κλείνοντας η Βασιλεία απαιτεί όλες οι ουσιώδεις πτυχές των διαδικασιών αξιολόγησης να εγκρίνονται από το διοικητικό συμβούλιο της τράπεζας ή κάποια ορισμένη επιτροπή. Θα πρέπει να έχουν πλήρη κατανόηση των αναφορών διαχείρισης όπως και κατανόησης όλου του συστήματος. Η διοίκηση πρέπει επίσης να διασφαλίζει, σε συνεχή βάση ότι το σύστημα αξιολόγησης λειτουργεί σωστά. Η διοίκηση και το προσωπικό του πιστωτικού ελέγχου πρέπει να συναντώνται τακτικά για να συζητούν

την απόδοση της διαδικασίας αξιολόγησης, τους τομείς που χρήζουν βελτίωσης και την κατάσταση των προσπαθειών για τη βελτίωση των ελλείψεων που είχαν εντοπιστεί προηγουμένως. Τα μοντέλα CSM's είναι πολύ χρήσιμα εργαλεία για τη διαχείριση του πιστωτικού κινδύνου, για να μπορέσουν να λειτουργήσουν σωστά χρειάζεται η συνεργασία αρκετών τμημάτων μίας τράπεζας, ενημέρωση και κατανόηση των στατιστικών αυτών μοντέλων. Στη επόμενη ενότητα θα αναλύσουμε τις στατιστικές μεθόδους που χρησιμοποιούνται στη διαδικασία βαθμολόγησης αιτήσεων.

ΚΕΦΑΛΑΙΟ 3

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ

Όπως είδαμε και στην προηγούμενη ενότητα προκειμένου να μειωθεί ο κίνδυνος έκθεσης οι τράπεζες εφαρμόζουν ισχυρές πολιτικές για την έγκριση δανείων, αξιολογούν λεπτομερώς αιτήσεις δανείων έτσι ώστε να ικανοποιηθούν τα κριτήρια των τραπεζών να προσαρμόζονται σε ισχυρά πρότυπα για διαφοροποίηση του χαρτοφυλακίου βάσει γεωγραφικών κριτηρίων και μέσω της στενής παρακολούθησης αγορών έτσι ώστε οποιαδήποτε προβλήματα να εντοπίζονται αρκετά νωρίς έτσι ώστε να υπόκεινται σε επεξεργασία.

Ανάλογα με τον πιστωτικό κίνδυνο η τράπεζα θα αποφασίσει αν θα δώσει δάνεια βάσει κάποιων κριτηρίων που αναγράφονται στις αιτήσεις. Στην περίπτωση αυτή υπολογίζεται η πιθανότητα του να πραγματοποιηθεί η αθέτηση των υποχρεώσεων με κριτήριο κάποιες παραμέτρους που σχετίζονται όπως είδαμε με το εισόδημα, επάγγελμα, περιουσιακά στοιχεία κ.α. Για να υπολογιστεί αυτή η πιθανότητα γίνεται χρήση του μοντέλου λογιστικής παλινδρόμησης. Μέσω του μοντέλου και του υπολογισμού των σχετικών συντελεστών μπορούμε να δούμε πως θα μεταβληθεί η πιθανότητα να σημειωθεί αθέτηση του δανείου σε σχέση με το να μην γίνει, αν οι ανεξάρτητες μεταβλητές που συνιστούν τα κριτήρια των αιτήσεων μεταβληθούν οριακά. Ένα άλλο πολύ διαδεδομένο μοντέλο βάσει του οποίου η τράπεζα θα αποφασίσει αν χορηγήσει ή όχι ένα δάνειο είναι το μοντέλο διαχωριστικής ανάλυσης. Στην περίπτωση αυτή η τράπεζα έχει την δυνατότητα να δει μέσω των διαχωριστικών συναρτήσεων αν υπάρχουν διαφορές μεταξύ της ομάδας για την οποία έχει εγκριθεί δάνειο και της ομάδας για την οποία δεν έχει εγκριθεί δάνειο και παράλληλα να εντοπιστούν εκείνοι οι παράγοντες που συμβάλλουν περισσότερο στην διάκριση αυτών των ομάδων. Στα δέντρα αποφάσεων, ο στόχος είναι να δημιουργηθεί ένα δέντρο όπου ο αρχικός κόμβος θα αναπαριστά το βασικό κριτήριο λήψης δανείων και στην συνέχεια θα διακρίνονται οι υποπεριπτώσεις οι οποίες συνιστούν τις διακλαδώσεις των δέντρων. Τα μοντέλα αυτά θα αναλυθούν λεπτομερώς παρακάτω.

3.1 Λογιστική Παλινδρόμηση

Στο μοντέλο πολλαπλής γραμμικής παλινδρόμησης, ο στόχος είναι να εκτιμήσουμε την μαθηματική και στατιστική σχέση μεταξύ της εξαρτημένης και

ανεξάρτητων μεταβλητών. Οι συντελεστές των ανεξάρτητων μεταβλητών μας δείχνουν το πόσο θα μεταβληθεί η εξαρτημένη μεταβλητή αν η ανεξάρτητη μεταβλητή μεταβληθεί οριακά. Πολλές φορές ως ανεξάρτητες μεταβλητές χρησιμοποιούμε και ψευδομεταβλητές οι οποίες παίρνουν την τιμή 0 ή 1. Αυτό γίνεται λόγω του ότι οι μεταβλητές αυτές είναι ποιοτικές όπως το φύλο, το μορφωτικό επίπεδο κ.τ.λ.

Στο μοντέλο λογιστικής παλινδρόμησης η εξαρτημένη μεταβλητή είναι διχοτομική. Ορισμένα παραδείγματα όπου βρίσκει εφαρμογή αυτό το μοντέλο είναι στον έλεγχο ποιότητας, όπου μας ενδιαφέρει η πρόβλεψη της πιθανότητας αποτυχίας μιας διεργασίας παραγωγής προϊόντων σε ένα εργοστάσιο τροφίμων, στον χώρο της υγείας όπου ο στόχος είναι να προβλεφθεί η πιθανότητα εμφάνισης μίας νόσου σε ένα άτομο ανάλογα με τα χαρακτηριστικά του, στον χώρο του marketing όπου ο στόχος είναι να γίνε πρόβλεψη της απόφασης για τη αγορά ή μη ενός προϊόντος, στον χώρο της αγοράς δανειακών κεφαλαίων όπου στόχος είναι να προβλεφθεί η πιθανότητας ένα δανειολήπτης να αθετήσει ή όχι την αποπληρωμής του δανείου κ.α.

Υπάρχουν ουσιαστικά τρία είδη λογιστικής παλινδρόμησης όπως η δίτιμη λογιστική παλινδρόμηση, η πολυωνυμική λογιστική παλινδρόμηση και η διατακτική λογιστική παλινδρόμηση. Το μοντέλο που θα χρησιμοποιηθεί στη ανάλυσή μας θα είναι η δίτιμη λογιστική παλινδρόμηση. Σε αυτό το μοντέλο η εξαρτημένη μεταβλητή λαμβάνει δύο τιμές οι οποίες αντιστοιχούν σε δύο ενδεχόμενα: την αποτυχία και την επιτυχία. Στην περίπτωση της αποτυχίας η εξαρτημένη μεταβλητή κωδικοποιείται με την τιμή 1 και στην περίπτωση της επιτυχίας η εξαρτημένη μεταβλητή κωδικοποιείται με την τιμή 0.

Η βασική λοιπόν διαφορά μεταξύ του γραμμικού μοντέλου παλινδρόμησης και του μοντέλου λογιστικής παλινδρόμησης είναι ότι στο τελευταίο η εξαρτημένη μεταβλητή είναι διακριτή και δυαδική ενώ στο μοντέλο πολλαπλής παλινδρόμησης είναι συνεχής. Πρόκειται λοιπόν για ένα γραμμικό μοντέλο πιθανότητας όπου οι συντελεστές των ανεξάρτητων μεταβλητών αναφέρονται στη πιθανότητα του να λάβει χώρα ένα γεγονός. Για την εκτίμηση των πιθανοτήτων γίνεται χρήση των μοντέλων logit και probit. Τα δύο αυτά μοντέλα δεν έχουν σημαντικές διαφορές ενώ το μοντέλο logit χρησιμοποιείται πολύ συχνά στον χώρο της υγείας αφού οι συντελεστές μπορούν να ερμηνευτούν σε όρους αναλογίας (odd ratio). Γενικότερα πάντως το μοντέλο logit χρησιμοποιείται πιο συχνά αφού είναι ποιο εύκολη η ερμηνεία του. Να σημειώσουμε ότι ως ανεξάρτητες μεταβλητές μπορούμε να χρησιμοποιήσουμε και ψευδομεταβλητές.

Για παράδειγμα θα μπορούσαμε να εκτιμήσουμε πόσες φορές υψηλότερη είναι η πιθανότητα του να πάθει ένας άντρας καρδιακή προσβολή σε σχέση με μια γυναίκα όπου βέβαια το φύλο χρησιμοποιείται ως ψευδομεταβλητή.

Το λογιστικό μοντέλο παλινδρόμησης μπορεί να χρησιμοποιηθεί και στον τραπεζικό κλάδο. Όταν ένας δανειολήπτης επιθυμεί να λάβει ένα δάνειο τότε θα πρέπει να κάνει την σχετική αίτηση στην τράπεζα όπου αν π.χ.αν ο δανειολήπτης είναι μια επιχείρηση τότε στην αίτηση θα αναγράφεται ο κλάδος στον οποίο δραστηριοποιείται η επιχείρηση, το περιθώριο κέρδους η ταμειακές ροές οι πωλήσεις, τα χρόνια που λειτουργεί η επιχείρηση, η ικανότητά της στο να αποπληρώνει δάνεια βάσει παρελθόντων στοιχείων, οι δείκτες ταχύτητας κυκλοφορίας αποθεμάτων και πωλήσεων κ.α. Σε περίπτωση που ο δανειολήπτης είναι ιδιώτης τότε στην αίτηση του δανελιου θα αναγράφονται το επάγγελμα του, η οικογενειακή του κατάσταση, το εισόδημα του, το φύλο, η δανειακή του προϊστορία, τα αν είναι πελάτης της τράπεζας ή όχι, η ηλικία του κ.α. Ανάλογα με τα χαρακτηριστικά αυτά ο πελάτης είτε θα λάβει ένα δάνειο είτε όχι.

Στο σημείο αυτό ας δούμε την γενική μορφή ενός γραμμικού υποδείγματος πιθανότητας. Ας θεωρήσουμε το απλό υπόδειγμα

$$Y = \beta_0 + \beta_1 x_1 \ \dots + \beta_p x_p$$

$$Y_i = \begin{cases} 0, & \textit{αν ο πελατης i είναι καλός} \\ 1 & \textit{διαφορετικά} \end{cases} \quad \textit{για κάθε i=1,2,3 ,...n}$$

Η τιμή Y μπορεί να λάβει την τιμή 0 ή 1. Επανερχόμενοι στο παράδειγμα των αιτήσεων των δανείων το Y θα λάβει την τιμή 0 αν ο πελάτης λάβει το δάνειο και την τιμή 1 αν δεν λάβει το δάνειο. Το x_1ως ανεξάρτητη μεταβλητή θα μπορούσε να είναι το εισόδημα του δανειολήπτη, το x_2 η ηλικία του δανειολήπτη κτλ . Ουσιαστικά η παραπάνω εξίσωση μας δείχνει πως το εισόδημα του πελάτη, η ηλικία καθώς και άλλα χαρακτηριστικά μπορούν να επηρεάσουν τη πιθανότητα να λάβει ο πελάτης ένα δάνειο. Αν λάβουμε υπόψη το μοντέλο έχουμε

$$E(Y) = p = \beta_0 + \beta_1 X_1 \ \dots + \beta_p X_p$$

τότε μπορούμε να δούμε ποια είναι η πιθανότητα ο δανειολήπτης να λάβει δάνειο για δεδομένο εισόδημα, ηλικία και άλλα χαρακτηριστικά.

Στο μοντέλο logit λαμβάνουμε τον λογάριθμο και το μοντέλο είναι το εξής

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} \ \ldots + \beta_{p-1} x_{ip-1}$$

Η συνάρτηση του p_i , $\ln\left(\frac{p_i}{1-p_i}\right)$ ονομάζεται λογάριθμος της σχετικής πιθανότητας (odds) και παίρνει τιμές από -∞ έως +∞.

Για δεδομένες τιμές του X_1 , X_2 , …. X_p μπορούμε να βρούμε ποιο είναι το log του odds ratio. Στην συνέχεια αντιλογαριθμίζοντας βρίσκουμε το odd ratio. Για παράδειγμα και στην περίπτωση των αιτήσεων για δάνειο αν το odd ratio είναι 1,5 τότε η πιθανότητα του να λάβει ο δανειολήπτης δάνειο βάσει κάποιων συγκεκριμένων χαρακτηριστικών είναι 1,5 φορές μεγαλύτερη σε σχέση με τη πιθανότητα του να μην λάβει δάνειο. Στην περίπτωση αυτή το p_i είναι η πιθανότητα λήψης του δανείου και το $(1 - p_i)$ η πιθανότητα μη λήψης του δανείου.

Με την εφαρμογή του logit στην περίπτωση αυτή έχουμε μια εξαρτημένη μεταβλητή η οποία αναφέρεται στην πιθανότητα πραγματοποίησης δύο εκβάσεων την λήψη ή μη λήψη του δανείου.

Αν θέλουμε να υπολογίσουμε την πιθανότητα λήψης του δανείου για δεδομένες τιμές των ανεξάρτητων μεταβλητών και π.χ. έχουμε 2 ανεξάρτητες μεταβλητές, τότε θα ισχύει ότι:

$$p = \frac{e^{\beta_o} + e^{\beta_1 X_1} + e^{\beta_2 X_2}}{1 + e^{\beta_o} + e^{\beta_1 X_1} + e^{\beta_2 X_2}}$$

Τα β_0 , β_1 , β_2 μας δείχνουν πόσο αυξάνουν τα log odds για μια οριακή μεταβολή της ανεξάρτητης μεταβλητή. Για παράδειγμα αν X_1είναι η μεταβλητή του επαγγέλματος όπου το άτομο είτε θα είναι ιδιωτικός υπάλληλος είτε δημόσιος υπάλληλος, τότε αν $\beta_1 = 0{,}38$ θα έχουμε $e^{\beta_1} = 1{,}46$ και συνεπώς ένας ιδιωτικός υπάλληλος θα έχει 1,46 φορές παραπάνω πιθανότητα να λάβει ένα δάνειο σε σχέση με έναν ιδιωτικό υπάλληλο.

Για την εκτίμηση των παραμέτρων στο μοντέλο της Λογιστικής Παλινδρόμησης χρησιμοποιείται συνήθως η μέθοδος της μέγιστης πιθανοφάνειας. Μπορεί να χρησιμοποιηθεί επίσης και η μέθοδος των ελαχίστων τετραγώνων, ωστόσο η έλλειψη της υπόθεσης της κανονικότητας έχει ως αποτέλεσμα να προκύπτουν διαφορετικές εκτιμήτριες από ότι με τη μέθοδο της μέγιστης πιθανοφάνειας και οι εκτιμήτριες αυτές να μην είναι απαραίτητα αμερόληπτες .

Η συνάρτηση υπολογισμού της πιθανοφάνειας δίνεται από τον τύπο :

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} {p_i}^{y_i} (1 - p_i)^{n_i - y_i}$$

Η συνάρτηση πιθανοφάνειας όπως φαίνεται από την εξίσωση εξαρτάται από τα p_i τα οποία με την σειρά τους εξαρτώνται από τα β, άρα κατά συνέπεια και η συνάρτηση πιθανοφάνειας εξαρτάται από τα β.

3.2Έλεγχοι καλής προσαρμογής των δεδομένων λογιστικής παλινδρόμησης

- Έλεγχος Hosmer and Lemeshow

Ο έλεγχος Hosmer and Lemeshow μας δείχνει αν τα δεδομένα προσαρμόζονται ικανοποιητικά στο μοντέλο που έχουμε επιλέξει. Πιο αναλυτικά ελέγχεται η υπόθεση εάν οι παρατηρούμενες τιμές διαφέρουν από τις εκτιμώμενες , δηλαδή έχουμε :

H_0: Οι παρατηρούμενες τιμές της Y δεν διαφέρουν από τις εκτιμώμενες

H_1: Οι παρατηρούμενες τιμές της Y διαφέρουν από τις εκτιμώμενες.

Έχουμε απόρριψη της H_0, όταν το p-value(sig) είναι μικρότερο από το επίπεδο σημαντικότητας $\alpha = 0{,}05$. Από τη στιγμή λοιπόν που απορρίπτεται η μηδενική υπόθεση το συμπέρασμα που βγαίνει είναι ότι τα δεδομένα που έχουμε δεν προσαρμόζονται σωστά στο μοντέλο.

- Έλεγχος λόγου πιθανοφανειών(Likelihood ratio test)

Ένας ακόμα τρόπος για τον έλεγχο καλής προσαρμογής είναι ο λόγος των πιθανοφανειών, ο οποίος δίνεται από τον :

$$LRT = -2\,ln\left(\frac{L_0}{L_F}\right)$$

Η συνάρτηση LRT ακολουθεί κατανομή x^2 με βαθμούς ελευθερίας όσες και οι μεταβλητές όπου:

L_O: μέγιστη πιθανοφάνεια του μοντέλου που περιέχει μόνο τη σταθερά β_0.

L_F : μέγιστη πιθανοφάνεια του εξεταζόμενου μοντέλου.

Ελέγχουμε τις υποθέσεις ότι :

H_0: οι συντελεστές είναι μηδέν

H_1: οι συντελεστές διαφορετικοί του μηδέν.

- Συντελεστής προσδιορισμού R^2των Cox και Snell

Υπάρχει και ένα ακόμη μέτρο καλής προσαρμογής και αυτός είναι ο συντελεστής R^2 των Cox και Snell ο οποίος δίνεται από τον τύπο :

$$R^2 = 1 - \left(\frac{L_0}{L_F}\right)^{\frac{2}{n}}$$

Παρόλα αυτά ο συντελεστής αυτός ποτέ δεν παίρνει μέγιστη τιμή το 1, οπότε χρησιμοποιείται ο συντελεστής Nagelkerke που υπολογίζεται από τον τύπο:

$$\tilde{R}^2 = \frac{R^2}{R^2_{max}} \in (0,1)$$

όπου, $R^2_{\max} = 1 - [L_0]^{2/n}$.

Η λογιστική παλινδρόμηση είναι μία από τις πιο απλές μεθόδους πρόβλεψης είναι εύκολη στη χρήση και παρέχει μεγάλη αποτελεσματικότητα. Τα πλεονεκτήματα της λογιστικής παλινδρόμησης είναι τα παρακάτω. Πρώτον είναι εύκολο να εφαρμοστεί και να ερμηνευτεί. Δεύτερον δεν υπόκειται σε στατιστικούς περιορισμούς όπως η διακριτική ανάλυση και συνεπώς κατά την κατασκευή της μπορεί να γίνει χρήση και ποιοτικών μεταβλητών. Τρίτον μέσω του λογιστικού υποδείγματος είναι εφικτή η εκτίμηση της σημαντικότητας του κάθε χαρακτηριστικού, κάτι που δεν είναι εφικτό στην διακριτική ανάλυση. Τέταρτον οι μεταβλητές δεν είναι απαραίτητο να κατανέμονται κανονικά.. Τέλος έχει καλή απόδοση όταν το σύνολο δεδομένων είναι γραμμικά διαχωρισμένο ενώ είναι πολύ γρήγορο στη ταξινόμηση άγνωστων εγγραφών.

Το μοντέλο που εξετάζουμε έχει και κάποια μειονεκτήματα όπως το ότι προϋποθέτει την μη ύπαρξη πολυσυγγραμικότητας, οι διάφορες ανεξάρτητες μεταβλητές έχουν μια πολλαπλασιαστική σχέση μεταξύ τους όσον αφορά την συχνότητα εμφάνισης της μελετώμενης έκβασης, επιπροσθέτως απαιτείται μεγάλος αριθμός παρατηρήσεων ενώ υπάρχει πρόβλημα όταν η σχέση εξαρτημένης και ανεξάρτητων μεταβλητών είναι μη γραμμική.

3.3 Διαχωριστική Ανάλυση

Η διαχωριστική ανάλυση είναι μία τεχνική που στόχος της είναι να ταξινομήσει παρατηρήσεις σε πληθυσμούς με γνωστές κατανομές. Στον τραπεζικό κλάδο η διαχωριστική ανάλυση χρησιμοποιείται για να κατατάξει ένα πελάτης σε «καλό» ή «κακό». Έτσι, έχοντας στην κατοχή της ιστορικά στοιχεία από υφιστάμενους πελάτες, θέλει να διαμορφώσει κανόνες, ώστε να κατατάξει έναν πελάτη στις παραπάνω κατηγορίες και να κρίνει αν θα γίνει δεκτή η αίτηση πίστωσης που έχει ζητήσει ή όχι.

Έστω λοιπόν ότι έχουμε ένα σύνολο πιθανών τιμών Π που οι τυχαίες μεταβλητές $\boldsymbol{X} = (X_1, X_2, \ldots . X_P)$, μπορούν να πάρουν κατά τη συμπλήρωση της αίτησης δανειοδότησης ενός πελάτη. Στόχος της τράπεζας είναι να βρει ένα κανόνα ώστε να πετύχει την μεγιστοποίηση του κέρδους της, οπότε η αίτηση κάθε πελάτη θα πρέπει να ελεγχθεί ενδελεχώς και να ταξινομηθεί ο πελάτης σε «καλό» ή «κακό». Το σύνολο λοιπόν Π θα πρέπει να διαχωριστεί σε υποσύνολο Π_A για τους υποψήφιους πελάτες που οι απαντήσεις των αιτήσεων τους τους ταξινόμησαν σε «καλούς» και σε Π_B οι απαντήσεις εκείνων που ταξινομήθηκαν σε «κακούς». Παρόλα αυτά υπάρχουν περιπτώσεις που από την ταξινόμηση των πελατών προκύπτουν λάθη. Τα λάθη που γίνονται λοιπόν είναι πρώτον να ταξινομηθεί ένας «καλός» πελάτης στον υποσύνολο των «κακών» πελατών, άρα το αναμενόμενο κέρδος από τον πελάτη αυτόν χάνεται . Το αναμενόμενο κέρδος θα το συμβολίσουμε με G και θα υποθέσουμε ότι είναι το ίδιο για κάθε πελάτη. Ο δεύτερος τύπος λάθους είναι η ταξινόμηση ενός «κακού» πελάτη σε «καλό». Αυτό αυτόματα σημαίνει ότι θα αθετήσει τις υποχρεώσεις του προς την τράπεζα και θα επιφέρει κόστος σε αυτή. Το αναμενόμενο κόστος θα το ονομάσουμε L και θα είναι και αυτό ίδιο για τον κάθε πελάτη.

Επομένως η αναμενόμενη απώλεια της λάθος ταξινόμησης του «καλού» πελάτη είναι ίση με $G \times P(\boldsymbol{x}|A)p_A$ και το αναμενόμενο κόστος της λάθος ταξινόμησης του «κακού» πελάτη είναι ως $L \times P(\boldsymbol{x}|B)p_B$.

όπου:

p_A: η πιθανότητα ένας πελάτης να ταξινομηθεί σε «καλό»

p_B: η πιθανότητα ένας πελάτης να ταξινομηθεί σε «κακό»

Για την ελαχιστοποίηση του αναμενόμενου κόστους πρέπει ο πελάτης με τα χαρακτηριστικά $\boldsymbol{x}$ να ταξινομείται στο σύνολο Π_A αν ισχύει :

$$L \times P(\boldsymbol{x}|B)p_B \leq G \times P(\boldsymbol{x}|A)p_A$$

οπότε η δανειοδότηση θα γίνεται στους πελάτες των οποίων τα χαρακτηριστικά $\boldsymbol{x}$ ανήκουν στο σύνολο:

$$\Pi_A = \{\boldsymbol{x}|L \times P(\boldsymbol{x}|B)p_B \leq G \times P(\boldsymbol{x}|A)p_A\} = \left\{\boldsymbol{x}|\frac{L}{G} \leq \frac{P(\boldsymbol{x}|A)p_A}{P(\boldsymbol{x}|B)p_B}\right\}$$

Αν υποθέσουμε τώρα πως τα χαρακτηριστικά δεν προέρχονται από διακριτές μεταβλητές, αλλά από συνεχείς η συνάρτηση ελαχιστοποίησης του κόστους διαμορφώνεται ως:

$$\Pi_A = \{\boldsymbol{x}|\ Lf(\boldsymbol{x}|B)p_B \leq Gf(\boldsymbol{x}|A)p_A\} = \left\{\boldsymbol{x}|\ \frac{Lp_B}{Gp_A} \leq \frac{f(\boldsymbol{x}|A)}{f(\boldsymbol{x}|B)}\right\}$$

Μεγάλο ενδιαφέρον έχει η πολυμεταβλητή κανονική κατανομή ως προς την χρήση της στις περισσότερες διαδικασίες ταξινόμησης, αφού χαρακτηρίζεται από απλότητα και υψηλή αποδοτικότητα.

Έστω $\boldsymbol{X} = (X_1, X_2, \ldots . X_P)$ είναι ένα τυχαίο δείγμα χαρακτηριστικών ενός πελάτη που οι τιμές προέρχονται από την πολυμεταβλητή κανονική κατανομή. Η μέση τιμή των «καλών» πελατών είναι $\boldsymbol{\mu}_A = \left(\mu_{A_1}, \mu_{A_2}, \ldots \mu_{A_p}\right)$ και των «κακών» $\boldsymbol{\mu_B} = \left(\mu_{B_1}, \mu_{B_2}, \ldots \mu_{B_p}\right)$. Υποθέτουμε ότι η διακύμανση των δύο ομάδων είναι ίσες οπότε και οι συνδιακυμάνσεις τους είναι ίσες.

$$\boldsymbol{X}_A \sim N_P(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}) \text{ και } \boldsymbol{X}_B \sim N_P(\boldsymbol{\mu}_B, \boldsymbol{\Sigma})$$

Σ:ο κοινός πίνακας συνδιακύμανσης των 2 ομάδων.

Οπότε για την ελαχιστοποίηση του αναμενόμενου κόστους :

$$\Pi_A = \{\boldsymbol{x}|\ Lf(\boldsymbol{x}|B)p_B \leq Gf(\boldsymbol{x}|A)p_A\} = \left\{\boldsymbol{x}|\ \frac{Lp_B}{Gp_A} \leq \frac{f(\boldsymbol{x}|A)}{f(\boldsymbol{x}|B)}\right\} =$$

$$\left\{\boldsymbol{x}|\ \frac{Lp_B}{Gp_A} \leq exp\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu_B})\boldsymbol{\Sigma^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_B})^T - (\boldsymbol{x}-\boldsymbol{\mu_A})\boldsymbol{\Sigma^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_A})^T}{2}\right)\right\}$$

Έπειτα από πράξεις καταλήγουμε ότι:

$$\Pi_A = \left\{\boldsymbol{x}|\boldsymbol{x\Sigma^{-1}}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T \geq ln\left(\frac{LP_B}{GP_A}\right) + \frac{\boldsymbol{\mu}_A{}^T\boldsymbol{\Sigma^{-1}}\boldsymbol{\mu}_A - \boldsymbol{\mu}_B{}^T\boldsymbol{\Sigma^{-1}}\boldsymbol{\mu}_B}{2}\right\}$$

Θα πρέπει να εκτιμήσουμε τις παραμέτρους $\boldsymbol{\mu}_A$ και $\boldsymbol{\mu}_B$ χρησιμοποιώντας τους δειγματικούς μέσους όπου για $\boldsymbol{\mu}_A = \boldsymbol{m}_A = \left(m_{A_1}, m_{A_2}, \ldots, m_{A_p}\right)$ το $m_{Aj} = \frac{1}{n_A}\sum_{i=1}^{n_A} x_{A_i}$ και για $\boldsymbol{\mu}_B = \boldsymbol{m}_B = \left(m_{B_1}, m_{B_2}, \ldots, m_{B_p}\right)$ το $m_{B_j} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{B_i}$ για κάθε $j = 1,2,3 \ldots p$.

Για να εκτιμήσουμε το $\boldsymbol{\Sigma}$, κάνουμε χρήση του δειγματικού πίνακα συνδιακύμανσης

$$\boldsymbol{\Sigma} = \boldsymbol{S}_p = \frac{1}{n_A + n_B - 2}\left[\sum_{i=1}^{n_A}\left(\boldsymbol{x}_{A_i} - \boldsymbol{m}_A\right)^T\left(\boldsymbol{x}_{A_i} - \boldsymbol{m}_A\right) + \sum_{i=1}^{n_B}\left(\boldsymbol{x}_{B_i} - \boldsymbol{m}_B\right)^T\left(\boldsymbol{x}_{B_i} - \boldsymbol{m}_B\right)\right]$$

Έχοντας λοιπόν εκτιμήσει όλα τα παραπάνω ένας πελάτης με ιδιότητες χ θα ταξινομηθεί σε «καλό» αν ισχύει

$$\boldsymbol{x}\boldsymbol{S}^{-1}(\boldsymbol{m}_A - \boldsymbol{m}_B)^T \geq ln\left(\frac{LP_B}{GP_A}\right) + \frac{\boldsymbol{m}_A{}^T\boldsymbol{S}^{-1}\boldsymbol{m}_A - \boldsymbol{m}_B{}^T\boldsymbol{S}^{-1}\boldsymbol{m}_B}{2}$$

και σε «κακό» αν ισχύει

$$\boldsymbol{x}\boldsymbol{S}^{-1}(\boldsymbol{m}_A - \boldsymbol{m}_B)^T < ln\left(\frac{LP_B}{GP_A}\right) + \frac{\boldsymbol{m}_A{}^T\boldsymbol{S}^{-1}\boldsymbol{m}_A - \boldsymbol{m}_B{}^T\boldsymbol{S}^{-1}\boldsymbol{m}_B}{2}$$

Έστω τώρα ότι $\boldsymbol{X} = (X_1, X_2, \ldots . X_P)$ είναι ένα τυχαίο δείγμα χαρακτηριστικών ενός πελάτη που οι τιμές προέρχονται από την πολυμεταβλητή κανονική κατανομή. Η μέση τιμή των «καλών» πελατών είναι $\boldsymbol{\mu}_A = \left(\mu_{A_1}, \mu_{A_2}, \ldots \mu_{A_p}\right)$ και των «κακών» $\boldsymbol{\mu}_B = \left(\mu_{B_1}, \mu_{B_2}, \ldots \mu_{B_p}\right)$. Ας υποθέσουμε ότι οι πίνακες των διακυμάνσεων και συνδιακυμάνσεων αυτή τη φορά είναι διαφορετικοί για τους δύο πληθυσμούς. Αυτό σημαίνει ότι :

$$\boldsymbol{x}_A \sim N_P(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) \text{ και } \boldsymbol{x}_B \sim N_p(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$$

όπου,

$\boldsymbol{\Sigma}_A$: Ο πίνακας συνδιακύμανσης για την ομάδα των «καλών» πελατών

$\boldsymbol{\Sigma}_B$: Ο πίνακας συνδιακύμανσης για την ομάδα των «κακών» πελατών

Για την ελαχιστοποίηση του αναμενόμενου κόστους θα έχουμε:

$\Pi_A = \{\boldsymbol{x}|\, Lf(x|B)p_B \leq Gf(x|A)p_A\} =$

$$\left\{\boldsymbol{x}|\, \frac{Lp_B}{Gp_A} \leq \frac{f(\boldsymbol{x}|A)}{f(\boldsymbol{x}|B)}\right\} = \left\{\boldsymbol{x}|\, \frac{Lp_B}{Gp_A} \leq \frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_A|}} exp\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_B)\boldsymbol{\Sigma}_B{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_B)^T - (\boldsymbol{x}-\boldsymbol{\mu}_A)\boldsymbol{\Sigma}_A{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_A)^T}{2}\right)\right\}$$

Έπειτα από πράξεις λοιπόν καταλήγουμε:

$$\Pi_A = \begin{Bmatrix} \boldsymbol{x}|\boldsymbol{x}\left(\boldsymbol{\Sigma}_B{}^{-1} - \boldsymbol{\Sigma}_A{}^{-1}\right)\boldsymbol{x}^T + 2\boldsymbol{x}(\boldsymbol{\Sigma}_A{}^{-1}\boldsymbol{\mu}_A - \boldsymbol{\Sigma}_B{}^{-1}\boldsymbol{\mu}_B) \geq \\ 2ln\left(\frac{Lp_B}{Gp_A}\right) - 2ln\left(\frac{\sqrt{|\boldsymbol{\Sigma}_B|}}{\sqrt{|\boldsymbol{\Sigma}_A|}}\right) + \boldsymbol{\mu}_A\boldsymbol{\Sigma}_A{}^{-1}\boldsymbol{\mu}_A{}^T - \boldsymbol{\mu}_B\boldsymbol{\Sigma}_B{}^{-1}\boldsymbol{\mu}_B{}^T \end{Bmatrix}$$

Θα πρέπει να εκτιμήσουμε τις παραμέτρους $\boldsymbol{\mu}_A$ και $\boldsymbol{\mu}_B$ χρησιμοποιώντας τους δειγματικούς μέσους όπου για $\boldsymbol{\mu}_A = \boldsymbol{m}_A = \left(m_{A_1}, m_{A_2}, \ldots, m_{A_p}\right)$ το $m_{A_j} = \frac{1}{n_A}\sum_{i=1}^{n_A} x_{A_i}$ και για $\boldsymbol{\mu}_B = \boldsymbol{m}_B = \left(m_{B_1}, m_{B_2}, \ldots, m_{B_p}\right)$ το $m_{B_j} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{B_i}$ για κάθε $j = 1,2,3 \ldots p$.

Η εκτίμηση των πινάκων συνδιακύμανσης $\boldsymbol{\Sigma}_A$ και $\boldsymbol{\Sigma}_B$:

$$\boldsymbol{\Sigma}_A = \boldsymbol{S}_A = \frac{1}{n_A - 1}\sum_{i=1}^{n_A}(\boldsymbol{x}_{Ai} - \boldsymbol{m}_A)^T(\boldsymbol{x}_{Ai} - \boldsymbol{m}_A),$$

$$\boldsymbol{\Sigma}_B = \boldsymbol{S}_B = \frac{1}{n_B - 1}\sum_{i=1}^{n_B}\left(\boldsymbol{x}_{B_i} - \boldsymbol{m}_B\right)^T\left(\boldsymbol{x}_{Ai} - \boldsymbol{m}_A\right)$$

Έχοντας λοιπόν εκτιμήσει όλα τα παραπάνω ένας πελάτης με ιδιότητες $\boldsymbol{x}$ θα ταξινομηθεί σε «καλό» αν ισχύει

$$\boldsymbol{x}|\boldsymbol{x}\left({\boldsymbol{S}_B}^{-1} - {\boldsymbol{S}_A}^{-1}\right)\boldsymbol{x}^T + 2\boldsymbol{x}\left({\boldsymbol{S}_A}^{-1}\boldsymbol{m}_A - {\boldsymbol{S}_B}^{-1}\boldsymbol{m}_B\right) \geq$$

$$2ln\left(\frac{Lp_B}{Gp_A}\right) - 2ln\left(\frac{\sqrt{|\boldsymbol{S}_B|}}{\sqrt{|\boldsymbol{S}_A|}}\right) + \boldsymbol{m}_A{\boldsymbol{S}_A}^{-1}{\boldsymbol{m}_A}^T - \boldsymbol{m}_B{\boldsymbol{S}_B}^{-1}{\boldsymbol{m}_B}^T$$

και σε «κακό»

$$\boldsymbol{x}|\boldsymbol{x}\left({\boldsymbol{S}_B}^{-1} - {\boldsymbol{S}_A}^{-1}\right)\boldsymbol{x}^T + 2\boldsymbol{x}\left({\boldsymbol{S}_A}^{-1}\boldsymbol{m}_A - {\boldsymbol{S}_B}^{-1}\boldsymbol{m}_B\right) <$$

$$2ln\left(\frac{Lp_B}{Gp_A}\right) - 2ln\left(\frac{\sqrt{|\boldsymbol{S}_B|}}{\sqrt{|\boldsymbol{S}_A|}}\right) + \boldsymbol{m}_A{\boldsymbol{S}_A}^{-1}{\boldsymbol{m}_A}^T - \boldsymbol{m}_B{\boldsymbol{S}_B}^{-1}{\boldsymbol{m}_B}^T$$

Τα πλεονέκτημα την διαχωριστικής ανάλυσης είναι ότι η εφαρμογή και η χρήση της είναι απλή και κατανοητή. Επιπλέον η διαχωριστική ανάλυση έχει μειωμένα ποσοστά σφάλματος Επίσης όταν δεν παραβιάζονται οι υποθέσεις της είναι πιο αποτελεσματικό εργαλείο από τη λογιστική παλινδρόμηση. Κλείνοντας με τα πλεονεκτήματα, προσφέρει τη δυνατότητα ταξινόμησης των περιπτώσεων που είναι "μη ομαδοποιημένες" ως προς την εξαρτημένη μεταβλητή.

Στην προηγούμενη ενότητα εξετάσαμε το μοντέλο πιθανότητας της λογιστικής παλινδρόμησης. Το μοντέλο αυτά υπερέχει έναντι του μοντέλου διαχωριστικής ανάλυσης σε κάποια σημεία τα οποία συνιστούν και τα μειονεκτήματα του τελευταίου. Πιο συγκεκριμένα στο μοντέλο λογιστικής παλινδρόμησης δεν απαιτείται τα σφάλματα να κατανέμονται κανονικά, ενώ και η διαχείριση των γραμμικών όρων αλληλεπίδρασης είναι άμεση και πιο απλή. Ένα μειονέκτημα της διαχωριστικής ανάλυσης είναι ότι είναι ευαίσθητη στην ύπαρξη ακραίων τιμών ενώ για να είναι αποτελεσματική η εφαρμογή της θα πρέπει να ισχύουν οι προϋποθέσεις ότι τα δεδομένα προέρχονται από την κανονική κατανομή και οι πίνακες των συνδιακυμάνσεων να είναι ίσοι.

3.2 Δέντρα Ταξινόμησης

Ένα δέντρο ταξινόμησης κατασκευάζεται βάσει μιας επαναληπτικής διαδικασίας διαίρεσης του αρχικού συνόλου των δεδομένων για μια εξαρτημένη μεταβλητή σε 2 υποσύνολα. Όσον αφορά τον πιστωτικό κίνδυνο, η βασική ιδέα των δέντρων ταξινόμησης, είναι να διαχωριστεί το σύνολο των απαντήσεων της αίτησης για χορήγηση δανείου σε καλό ή κακό.

Ένα δέντρο απόφασης ή ταξινόμησης συνιστά μια σειρά του τύπου if then που συνδυάζονται μεταξύ τους από την ρίζα του δέντρου προς τα φύλλα. Οι κόμβοι του δέντρου χαρακτηρίζονται με τα ονόματα των χαρακτηριστικών, οι ακμές λαμβάνουν ονομασία βάση της τιμής που μπορεί να λάβει ένα χαρακτηριστικό και τα φύλλα τέλος χαρακτηρίζονται με τις διάφορες κλάσεις.

Το δέντρο ταξινόμησης χτίζεται βάσει μιας επαναλαμβανόμενης διαδικασίας όπου διασπάται ένα δοσμένο σύνολο δεδομένων βάσει διαφόρων ανεξάρτητων μεταβλητών. Η σειρά με τη οποία χρησιμοποιούνται οι ανεξάρτητες μεταβλητές στην δημιουργία του δέντρου εξαρτάται από την δυνατότητα ταξινόμησης της εκάστοτε μεταβλητής. Θα πρέπει να επιλεχθεί η μεταβλητή εκείνη η οποία διαχωρίζει με το καλύτερο δυνατό τρόπο τις τελικές κλάσεις. Με κριτήριο τον τρόπο ανάπτυξής τους, τα δέντρα αποφάσεων διακρίνονται σε δυαδικά όπου από κάθε κόμβο διακλαδίζονται δύο νέοι κόμβοι και σε σύνθετα όπου από κάθε κόμβο διακλαδίζονται δύο η περισσότεροι κόμβοι.

Το δέντρο ταξινόμησης βάσει εναλλακτικού ορισμού αποτελείται από εσωτερικού και εξωτερικούς κόμβους οι οποίοι συνδέονται με διακλαδώσεις. Ο εσωτερικός κόμβος συνιστά μια μονάδα λήψης απόφασης η οποία υπολογίζει την συνάρτηση απόφασης για τον προσδιορισμό του επόμενου κόμβου γνωστό και ως (child node) με τον οποίο θα συνδεθεί. Ο εξωτερικός κόμβος δεν έχει απογόνους αλλά συνδέεται με μια ονομασία ή τιμή που χαρακτηρίζει τα δεδομένα που οδηγούν σε αυτόν. Το αποτέλεσμα της συνάρτησης απόφασης που χρησιμοποιείται από τον εσωτερικό κόμβο θα καθορίσει τον τρόπο βάσει του οποίου το δέντρο θα διακλαδωθεί με τους απογόνους κόμβους.

Τα δέντρα ταξινόμησης χρησιμοποιούνται και για την εκτίμηση του πιστωτικού κινδύνου των πελατών. Βάση του αρχικού κόμβου που είναι η ρίζα του δέντρου και των σχετικών διακλαδώσεων μπορεί να αποφασιστεί ποιοι πελάτες θα λάβουν δάνειο και ποιοι όχι.

Στο σημείο αυτό ας παρουσιάσουμε ένα παράδειγμα όπου με κριτήριο εισοδήματος αποφασίζεται αν θα χορηγηθεί ένα δάνειο ή όχι. Έστω ότι τα άτομα διαχωρίζονται σε 2 κατηγορίες όπως το αν έχουν εισόδημα πάνω ή κάτω από 30.000 ευρώ. Η σχετικά απάντηση θα είναι ότι δεν έχουν ή έχουν. Αν δεν έχουν τότε μπορεί να εξεταστεί και ένα άλλο χαρακτηριστικό όπως το αν έχουν προϋπηρεσία μεγαλύτερη ή μικρότερη από 5 έτη. Αν έχουν τότε θα λάβουν δάνειο ενώ αν δεν έχουν δεν θα λάβουν. Στην άλλη διακλάδωση του δέντρου αν τα άτομα λαμβάνουν άνω των 30.000

ευρώ τότε μπορούν να εξεταστούν βάσει του χαρακτηριστικού του αν έχουν υψηλό ή χαμηλό χρέος. Αν έχουν υψηλό χρέος τότε θα λάβουν δάνειο ενώ αν δεν έχουν δεν θα λάβουν.

Η δημιουργία δέντρων ταξινόμησης μας βοηθάει να βρούμε τρόπους για να αναγνωρίσουμε κανόνες για την πρόβλεψη μελλοντικών ταξινομήσεων. Τα δέντρα ταξινόμησης χρησιμοποιούν συνήθως τη μέθοδο CHAID(Chi-squared automatic interaction detection) ή τη μέθοδο CRT (Classification and regression). Οι κύριες διαφορές ανάμεσα στις δύο αυτές μεθόδους είναι:

1. Χειρίζονται διαφορετικά τις τιμές που λείπουν. Η μέθοδος CRT ταξινομεί χρησιμοποιώντας τις υπόλοιπες ανεξάρτητες μεταβλητές που σχετίζονται πολύ με την ανεξάρτητη μεταβλητή που δεν έχει τιμή. Η μέθοδος CHAID αντιμετωπίζει όλες τις τιμές μιας ανεξάρτητης μεταβλητής που λείπουν σαν μία κατηγορία.
2. Η μέθοδος CHAID αποφασίζει για τη διάσπαση ενός δέντρου μέσω του Pearson's Chi-squared ενώ η μέθοδος CRT αποφασίζει χρησιμοποιώντας τον δείκτη Gini.
3. Η μέθοδος CRT διαιρεί μόνο δυαδικά ένα δέντρο. Αν όλες οι ανεξάρτητες μεταβλητές είναι δίτιμες τότε τόσο η μέθοδος CRT όσο και η μέθοδος CHAID θα δώσουν το ίδιο δέντρο ως αποτέλεσμα.
4. Η μέθοδος CRT τείνει να απλοποιεί το δέντρο που παράγει αφού οι κόμβοι του δέντρου που αυξάνουν την πιθανότητα της λανθασμένης ταξινόμησης αφαιρούνται αυτόματα.

Τα δέντρα ταξινόμησης έχουν σημαντικά πλεονεκτήματα όπως το ότι είναι αεπεξηγηματικά που σημαίνει ότι ένα δέντρο απόφασης με σχετικά μικρό αριθμό φύλλων μπορεί να χρησιμοποιηθεί εύκολα και να κατανοηθεί από έμπειρους χρήστες. Η αναπαράστασή τους είναι τόσο πλούσια που έχουν την δυνατότητα να αναπαραστήσουν οποιοδήποτε ταξινόμηση διακριτών τιμών και διαχειρίζονται σύνολα δεδομένων με ελλιπείς τιμές.

Συγχρόνως έχουν και κάποια μειονεκτήματα όπως ότι τα δέντρα ταξινόμησης μπορεί να είναι ασταθή, επειδή μικρές μεταβολές στα δεδομένα μπορεί να οδηγήσουν στη δημιουργία ενός εντελώς διαφορετικού δέντρου.

ΚΕΦΑΛΑΙΟ 4
ΕΦΑΡΜΟΓΗ ΜΟΝΤΕΛΩΝ ΒΑΘΜΟΛΟΓΗΣΗΣ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

4.1 Γενικά χαρακτηριστικά για το σύνολο των δεδομένων

Στην παρούσα ανάλυση θα υλοποιηθούν, με τη βοήθεια του πακέτου SPSS 27, τρία στατιστικά μοντέλα βαθμολόγησης της πιστοληπτικής ικανότητας με τρεις διαφορετικές μεθόδους (λογιστική παλινδρόμηση, διαχωριστική ανάλυση και δέντρα ταξινόμησης) πάνω σε ένα δείγμα 500 ατόμων που έχει παρθεί από μία τράπεζα της Αμερικής η οποία ασχολείται με τα στεγαστικά δάνεια. Τα δεδομένα είναι διαθέσιμα στον ιστότοπο «kaggle.com» . Από τα 500 αυτά άτομα, τα 345 έχουν χαρακτηριστεί ως "καλοί πελάτες" ενώ τα υπόλοιπα 155 ως "κακοί πελάτες". Το δείγμα περιγράφεται μέσα από 14 μεταβλητές που καταγράφουν είτε προσωπικά στοιχεία του κάθε πελάτη είτε στοιχεία του πιστωτικού του ιστορικού.

Συγκεκριμένα, οι μεταβλητές αυτές είναι οι εξής:

1. Loan_ID: Εκφράζει τον μοναδικό κωδικό κάθε δανείου και χρησιμοποιείται ως αναγνωριστικό της κάθε εγγραφής. Είναι κατηγορική μεταβλητή που δε θα χρειαστούμε κάπου.
2. Gender: Το φύλο του αιτούντος. Είναι δίτιμη κατηγορική μεταβλητή που έχει κωδικοποιηθεί ώστε η τιμή 1 να απεικονίζει άντρα και η τιμή 2 γυναίκα.
3. Married: Εκφράζει το αν ο αιτών είναι παντρεμένος ή όχι. Είναι δίτιμη κατηγορική μεταβλητή που έχει κωδικοποιηθεί ώστε η τιμή 0 να αντιστοιχεί στο 'όχι' και η τιμή 1 στο 'ναι'.
4. Dependents: Δείχνει το πλήθος των εξαρτώμενων μελών. Φαίνεται να είναι ποσοτική μεταβλητή αφού εκφράζει πλήθος, όμως αν τα εξαρτώμενα μέλη είναι από 3 και επάνω, τότε εμφανίζονται με την ίδια τιμή (3+), άρα η μεταβλητή αυτή είναι ιεραρχική.
5. Education: Εκφράζει το αν ο αιτών είναι κάτοχος πτυχίου ή όχι. Είναι δίτιμη κατηγορική μεταβλητή που έχει κωδικοποιηθεί ώστε η τιμή 1 να αντιστοιχεί στην ύπαρξη και η τιμή 2 στην μη ύπαρξη πτυχίου.

6. Self_employed: Δείχνει αν ο αιτών είναι αυτοαπασχολούμενος. Είναι δίτιμη κατηγορική μεταβλητή που έχει κωδικοποιηθεί ώστε η τιμή 1 να αντιστοιχεί σε κάποιον που είναι αυτοαπασχολούμενος και η τιμή 0 σε κάποιον που δεν είναι αυτοαπασχολούμενος.
7. ApplicantIncome: Το ετήσιο εισόδημα του αιτούντος. Η τιμή στα αρχικά δεδομένα ήταν σε δολάρια και μετατράπηκε σε ευρώ σύμφωνα με την ισοτιμία της 7ης Ιουλίου 2022, όπου 1$ = 0.99€. Η μεταβλητή είναι συνεχής αριθμητική.
8. CoapplicantIncome: Το ετήσιο εισόδημα του συναιτούντος. Η τιμή στα αρχικά δεδομένα ήταν σε δολάρια και μετατράπηκε σε ευρώ σύμφωνα με την ισοτιμία της 7ης Ιουλίου 2022, όπου 1$ = 0.99€. Η μεταβλητή είναι συνεχής αριθμητική.
9. LoanAmount: Συνεχής αριθμητική μεταβλητή που εκφράζει το ποσό δανείου σε χιλιάδες ευρώ. Αρχικά ήταν σε χιλιάδες δολάρια οπότε μετατράπηκε σε χιλιάδες ευρώ σύμφωνα με την ισοτιμία της 7ης Ιουλίου 2022, όπου 1$ = 0.99€.
10. Loan_Amount_Term: Συνεχής αριθμητική μεταβλητή που εκφράζει τη διάρκεια του δανείου σε μήνες.
11. Credit_History: Δίτιμη κατηγορική μεταβλητή η οποία παίρνει την τιμή 1 όταν το πιστωτικό ιστορικό του αιτούντος πληροί τις κατευθυντήριες γραμμές της τράπεζας και την τιμή 0 στην αντίθετη περίπτωση.
12. Property_Area: Κατηγορική μεταβλητή που περιγράφει την περιοχή της κατοικίας του αιτούντος. Η τιμή 1 εκφράζει αστική περιοχή, η τιμή 2 ημιαστική και η τιμή 3 εκφράζει αγροτική περιοχή.
13. Loan_Status: Δίτιμη κατηγορική μεταβλητή η οποία εκφράζει την απάντηση της τράπεζας και παίρνει την τιμή 1 όταν πρόκειται για καλό πελάτη και την τιμή 0 όταν πρόκειται για κακό πελάτη.
14. Total_Income: Συνεχής αριθμητική μεταβλητή που δείχνει το συνολικό ετήσιο εισόδημα, του αιτούντος και του συναιτούντος μαζί. Επειδή η αρχική τιμή ήταν σε δολάρια και για να μη προκύψουν λάθη από τις στρογγυλοποιήσεις στα 2 δεκαδικά ψηφία, αντί να μετατρέψουμε την τιμή από δολάρια σε ευρώ όπως πριν, προσθέσαμε τις τιμές των μεταβλητών ApplicantIncome + CoapplicantIncome.

Το δείγμα μας αποτελείται από 500 συνολικά άτομα. Όπως φαίνεται στον Πίνακα 1 και το Σχήμα 1, τα 500 αυτά άτομα μοιράζονται σε 402 άντρες (80.4%) και 98 γυναίκες (19.6%). Το άτομα του δείγματος είναι κατά κύριο λόγο (64.6%)

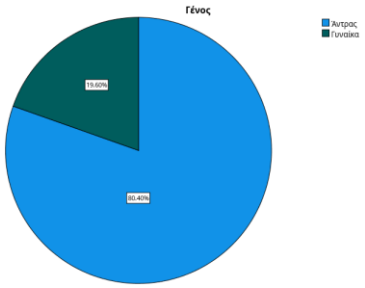
παντρεμένοι (Πίνακας 2 και Σχήμα 2), με το υπόλοιπο 35.4% ανύπαντροι. Ο Πίνακας 3 και το Σχήμα 3 δείχνει το πλήθος των εξαρτώμενων μελών του δείγματος και παρατηρούμε ότι η πλειοψηφία του δεν έχει εξαρτώμενα μέλη (58.2%), ενώ ακολουθεί το 17% με ένα εξαρτώμενο μέλος και το 16.6% με δύο. Ο Πίνακας 4 και το Σχήμα 4 δείχνει την εκπαίδευση από εκεί φαίνεται ότι το 78.6% του δείγματος είναι κάτοχοι πτυχίου ενώ το υπόλοιπο 21.4% όχι. Το μεγαλύτερο κομμάτι του δείγματος δηλώνει αυτοαπασχολούμενο, (83.6%) ενώ το υπόλοιπο 16.4% όχι (

Πίνακας 5 και Σχήμα 5). Τέλος, όπως δείχνει ο Πίνακας 6 και το Σχήμα 6, το δείγμα είναι σχετικά μοιρασμένο ως προς την περιοχή κατοικίας: το 37.6% μένει σε ημιαστική περιοχή, το 34.2% σε αστική ενώ το υπόλοιπο 28.2 σε αγροτική περιοχή.

Γένος

	N	%
Άντρας	402	80.4%
Γυναίκα	98	19.6%

Πίνακας 1: Γένος του δείγματος

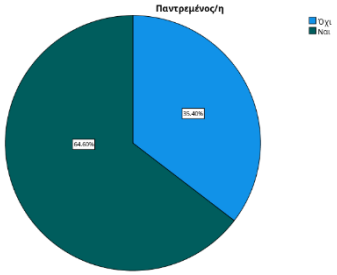


Σχήμα 1: Γένος του δείγματος

Παντρεμένος/η

	N	%
Όχι	177	35.4%
Ναι	323	64.6%

Πίνακας 2: Οικογενειακή κατάσταση του δείγματος

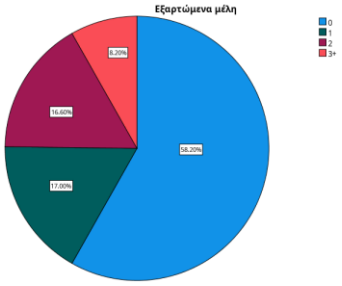


Σχήμα 2: Οικογενειακή κατάσταση του δείγματος

Εξαρτώμενα μέλη

	N	%
0	291	58.2%
1	85	17.0%
2	83	16.6%
3+	41	8.2%

Πίνακας 3: Εξαρτώμενα μέλη του δείγματος

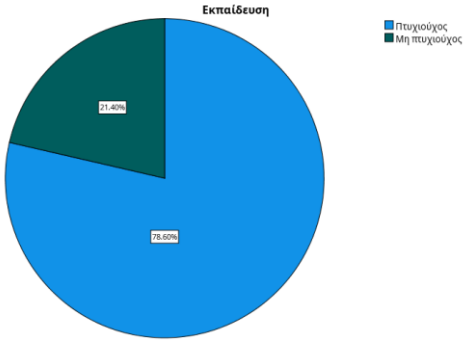


Σχήμα 3: Εξαρτώμενα μέλη του δείγματος

Εκπαίδευση

	N	%
Πτυχιούχος	393	78.6%
Μη πτυχιούχος	107	21.4%

Πίνακας 4: Εκπαίδευση του δείγματος

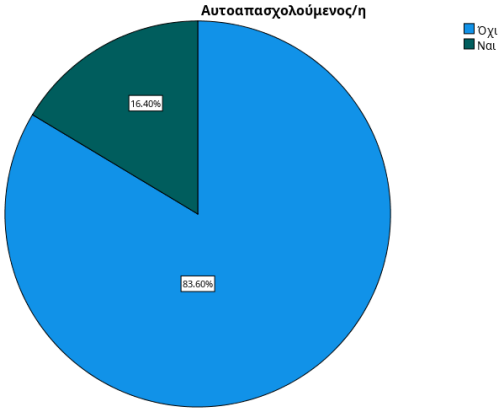


Σχήμα 4: Εκπαίδευση του δείγματος

Αυτοαπασχολούμενος/η

	N	%
Όχι	418	83.6%
Ναι	82	16.4%

Πίνακας 5: Αυτοαπασχόληση του δείγματος

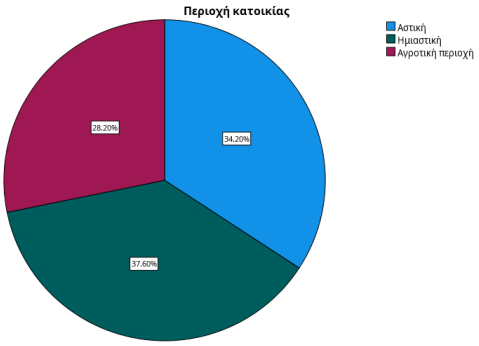


Σχήμα 5: Αυτοαπασχόληση του δείγματος

Περιοχή κατοικίας

	N	%
Αστική	171	34.2%
Ημιαστική	188	37.6%
Αγροτική περιοχή	141	28.2%

Πίνακας 6: Περιοχή κατοικίας του δείγματος



Σχήμα 6: Περιοχή κατοικίας του δείγματος

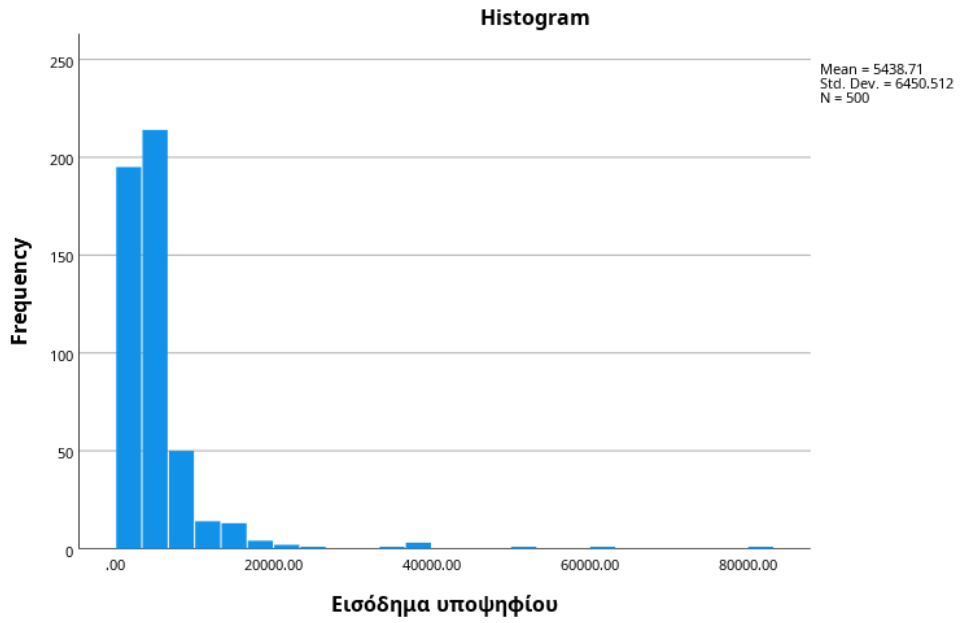
Εξετάζοντας τις ποσοτικές μεταβλητές του δείγματος παίρνουμε τον Πίνακα 7. Εκεί βλέπουμε πως το ετήσιο εισόδημα των υποψηφίων παίρνει τιμές από 148.50€ μέχρι 80190.00€, με μέση τιμή τα 5438.71€ και μεγάλη τυπική απόκλιση (s=6450.51). Το Σχήμα 7 δείχνει το ιστόγραμμα του ετήσιου εισοδήματος των υποψηφίων και μια πρώτη εικόνα για το είδος της κατανομής που ακολουθεί η μεταβλητή. Το ετήσιο εισόδημα των συνυποψηφίων παίρνει τιμές μέχρι 19800.00€, με μέση τιμή τα 14991.24€ και τυπική απόκλιση (s=2113.09€). Ομοίως, στο Σχήμα 8 βλέπουμε το ιστόγραμμα του ετήσιου εισοδήματος των συνυποψηφίων. Όπως και στο εισόδημα των υποψηφίων, το ιστόγραμμα δείχνει να μην ακολουθεί την κανονική κατανομή. Βάζοντας μαζί τα δύο εισοδήματα, το συνολικό εισόδημα παίρνει τιμές από 1427.58€ μέχρι 80190.00€, με μέση τιμή τα 6929.95€ και τυπική απόκλιση τα 6539.70€. Στο Σχήμα 11 βλέπουμε το ιστόγραμμα του ετήσιου συνολικού εισοδήματος που δε δείχνει να ακολουθεί την κανονική κατανομή, κάτι που περιμένουμε αφού καμία από τις δύο μεταβλητές που συνθέτουν το συνολικό εισόδημα δε φαίνεται να ακολουθεί την κανονική κατανομή.

Το ποσό δανείου παίρνει τιμές από 12870.00€ μέχρι και 693000.00€. Θυμίζουμε ότι η μεταβλητή αυτή εκφράζει ποσά σε χιλιάδες ευρώ. Η μέση τιμή του ποσού δανείου είναι 141461.10€ με τυπική απόκλιση 81993.02€. Το Σχήμα 9 δείχνει το ιστόγραμμα της μεταβλητής και σε αντίθεση με τις προηγούμενες μεταβλητές, η μεταβλητή που δείχνει το ποσό δανείου φαίνεται να ακολουθεί κανονική κατανομή. Τέλος, η διάρκεια του δανείου παίρνει τιμές από 12 έως 480 μήνες, με μέση τιμή τους 340.15 μήνες και τυπική απόκλιση 66.63 μήνες. Στο Σχήμα 10 φαίνεται το ιστόγραμμα της μεταβλητής και δε δίνει ενδείξεις ότι ακολουθεί κανονική κατανομή.

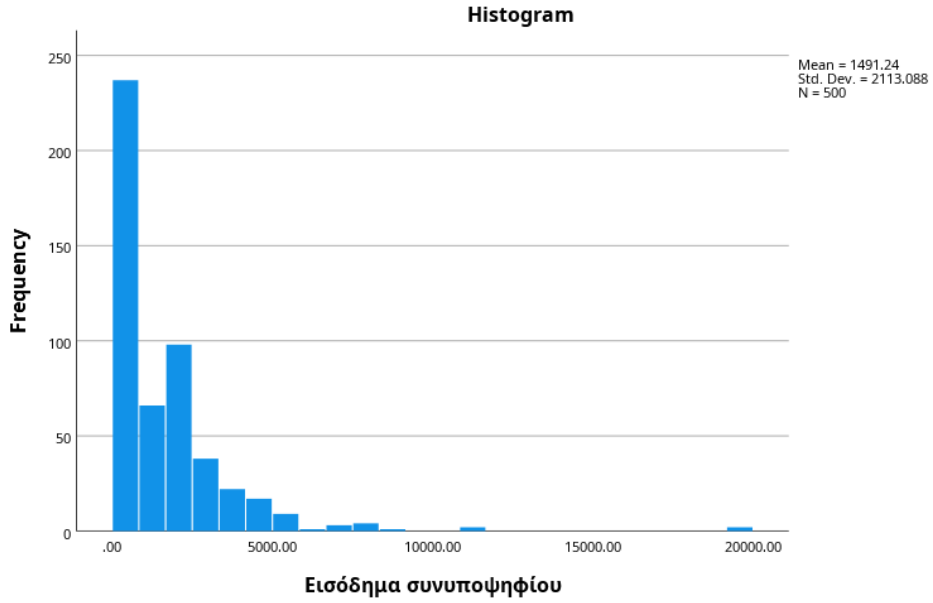
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Εισόδημα υποψηφίου	500	148.50	80190.00	5438.7076	6450.51228
Εισόδημα συνυποψηφίου	500	.00	19800.00	1491.2448	2113.08787
Ποσό δανείου	500	12.87	693.00	141.4611	81.99302
Διάρκεια του δανείου σε μήνες	500	12	480	340.15	66.632
Συνολικό εισόδημα	500	1427.58	80190.00	6929.9523	6539.70166
Valid N (listwise)	500				

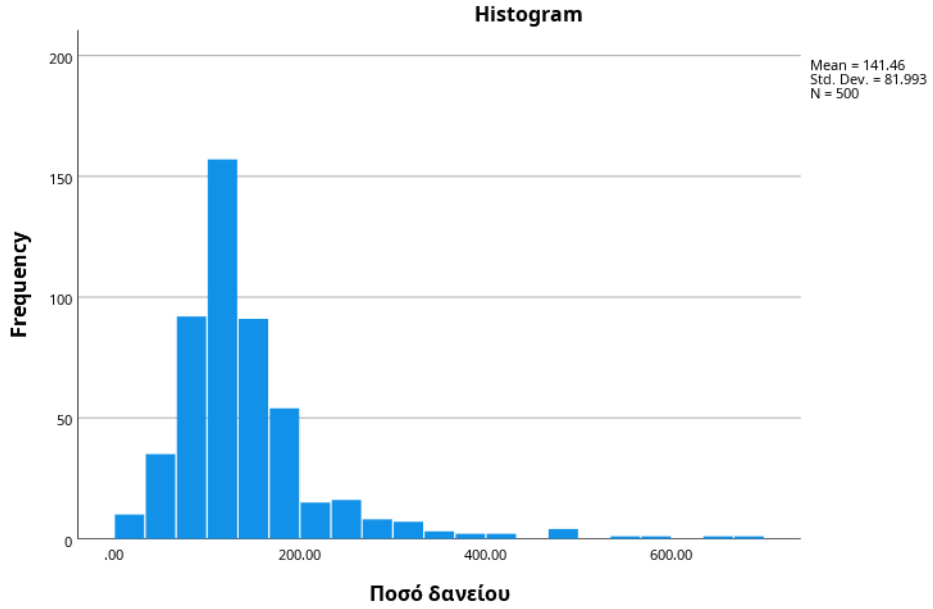
Πίνακας 7: Στατιστικά ποσοτικών μεταβλητών του δείγματος



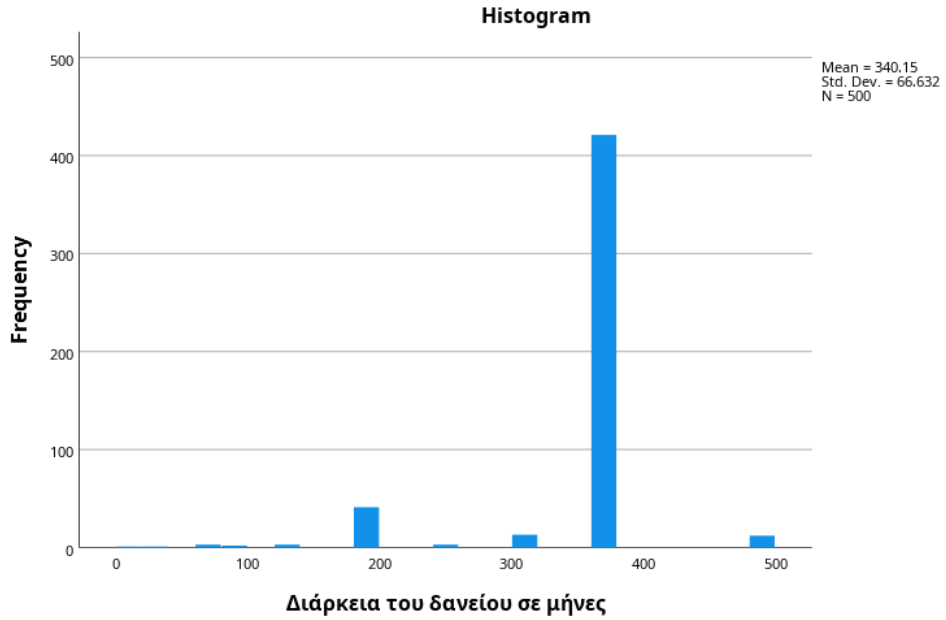
Σχήμα 7: Ιστόγραμμα εισοδήματος υποψηφίων



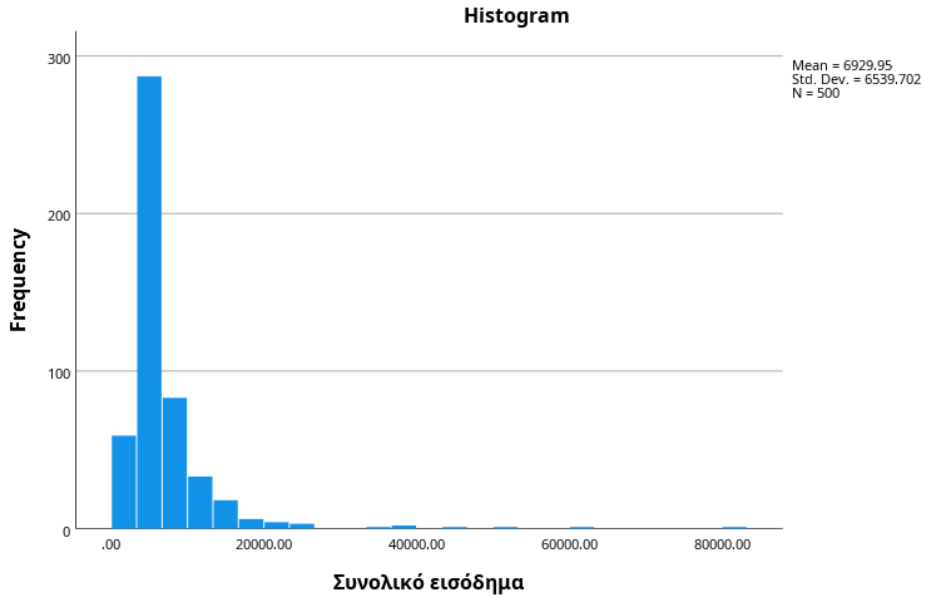
Σχήμα 8: Ιστόγραμμα εισοδήματος συνυποψηφίων



Σχήμα 9: Ιστόγραμμα ποσών δανείου



Σχήμα 10: Ιστόγραμμα διάρκειας δανείου



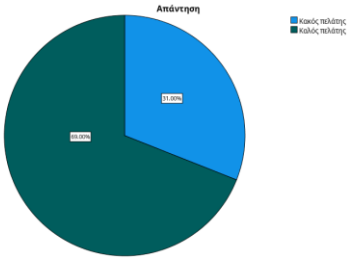
Σχήμα 11: Ιστόγραμμα συνολικού εισοδήματος

Όπως δείχνει ο Πίνακας 8 και το Σχήμα 12, τα 500 άτομα του δείγματος αποτελούνται από 345 "καλούς πελάτες" (69%) και 155 "κακούς πελάτες" (31%). Αν τώρα κοιτάξουμε τα δημογραφικά στοιχεία των δύο αυτών κατηγοριών ξεχωριστά, τότε θα

δούμε στον Πίνακα 9 το πώς κατανέμεται το φύλο των αιτούντων ανάμεσα στις δύο κατηγορίες. Δε φαίνονται ιδιαίτερες διαφορές αφού και στις 2 κατηγορίες, οι άντρες έχουν περίπου το ίδιο μεγαλύτερο ποσοστό (78.7% στους κακούς πελάτες και 81.2% στους καλούς πελάτες), κάτι που φαίνεται και στο Σχήμα 13. Στην οικογενειακή κατάσταση (Πίνακας 10 και Σχήμα 14) βλέπουμε ότι το 67% των "καλών πελατών" και το 59.4% των "κακών πελατών" είναι παντρεμένο. Ο Πίνακας 11 και το Σχήμα 15 δείχνουν τα εξαρτώμενα μέλη ανά κατηγορία: παρατηρούμε ότι, όπως και πριν χωρίσουμε το δείγμα σε δύο κατηγορίες, το μεγαλύτερο ποσοστό (58.3%) των "καλών πελατών" και των "κακών πελατών" (58.1%) δεν έχει εξαρτώμενα μέλη. Στη συνέχεια, ο Πίνακας 12 και το Σχήμα 16 δείχνουν την εκπαίδευση των καλών και των κακών πελατών και παρατηρούμε ότι το αν κάποιος είναι πτυχιούχος ή όχι αλλάζει λίγο ανάμεσα στις δύο κατηγορίες: οι πτυχιούχοι στους καλούς πελάτες αποτελούν το 81.2% ενώ στους κακούς πελάτες το 72.9%. Σχεδόν καμία διαφορά δε φαίνεται να υπάρχει στην αυτοαπασχόληση ανάμεσα στις δύο κατηγορίες πελατών: Το 16.2% των καλών πελατών και το 16.8% των κακών πελατών δηλώνουν αυτοαπασχολούμενοι. Τέλος, ο Πίνακας 15 και το Σχήμα 18 δείχνουν το που κατοικούν οι δύο κατηγορίες και παρατηρούμε διαφορές ανάμεσα στις κατηγορίες: Οι καλοί πελάτες κατοικούν (σε ποσοστό 43.2%) σε ημιαστικές περιοχές και μετά (σε ποσοστό) 32.2% σε αστικές, ενώ οι κακοί πελάτες κατοικούν, σε ποσοστό 38.8%, σε αστικές περιοχές και μετά, σε ποσοστό 36.1% σε αγροτικές περιοχές.

		Frequency	Percent
Valid	Κακός πελάτης	155	31.0
	Καλός πελάτης	345	69.0
	Total	500	100.0

Πίνακας 8: Καλοί/Κακοί πελάτες στο δείγμα



Σχήμα 12: Καλοί/κακοί πελάτες στο δείγμα

Γένος

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Άντρας	122	78.7
		Γυναίκα	33	21.3
		Total	155	100.0
Καλός πελάτης	Valid	Άντρας	280	81.2
		Γυναίκα	65	18.8
		Total	345	100.0

Πίνακας 9: Γένος του δείγματος ανά κατηγορία πελάτη

Παντρεμένος/η

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Όχι	63	40.6
		Ναι	92	59.4
		Total	155	100.0
Καλός πελάτης	Valid	Όχι	114	33.0
		Ναι	231	67.0
		Total	345	100.0

Πίνακας 10: Οικογενειακή κατάσταση του δείγματος ανά κατηγορία πελάτη

Εξαρτώμενα μέλη

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	0	90	58.1
		1	31	20.0
		2	19	12.3
		3+	15	9.7
		Total	155	100.0
Καλός πελάτης	Valid	0	201	58.3
		1	54	15.7
		2	64	18.6
		3+	26	7.5
		Total	345	100.0

Πίνακας 11: Εξαρτώμενα μέλη του δείγματος ανά κατηγορία πελάτη

Εκπαίδευση

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Πτυχιούχος	113	72.9
		Μη πτυχιούχος	42	27.1
		Total	155	100.0
Καλός πελάτης	Valid	Πτυχιούχος	280	81.2
		Μη πτυχιούχος	65	18.8
		Total	345	100.0

Πίνακας 12: Εκπαίδευση του δείγματος ανά κατηγορία πελάτη

Αυτοαπασχολούμενος/η

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Όχι	129	83.2
		Ναι	26	16.8
		Total	155	100.0
Καλός πελάτης	Valid	Όχι	289	83.8
		Ναι	56	16.2
		Total	345	100.0

Πίνακας 13: Αυτοαπασχόληση του δείγματος ανά κατηγορία πελάτη

Πιστωτικό ιστορικό

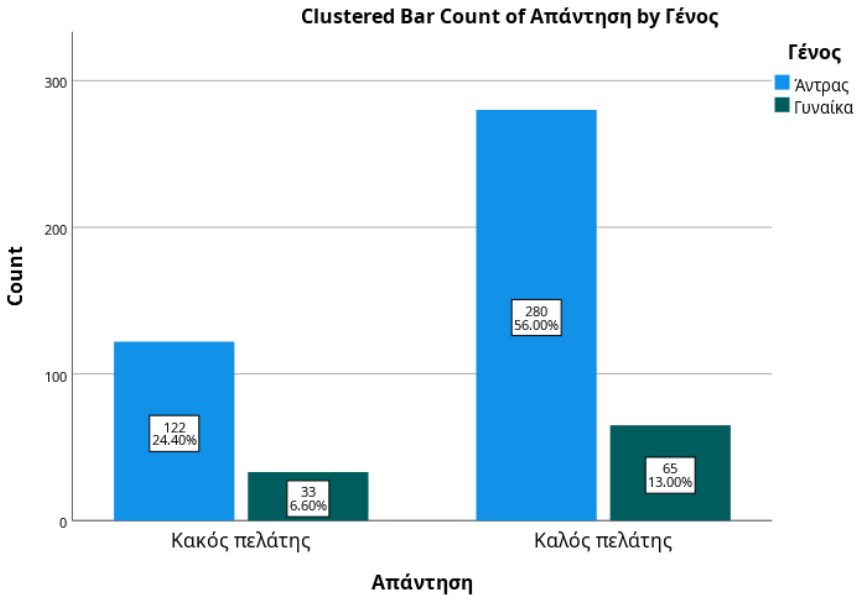
Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Όχι	68	43.9
		Το πιστωτικό ιστορικό πληροί τις κατευθυντήριες γραμμές της τράπεζας	87	56.1
		Total	155	100.0
Καλός πελάτης	Valid	Όχι	24	7.0
		Το πιστωτικό ιστορικό πληροί τις κατευθυντήριες γραμμές της τράπεζας	321	93.0
		Total	345	100.0

Πίνακας 14: Πιστωτικό ιστορικό του δείγματος ανά κατηγορία πελάτη

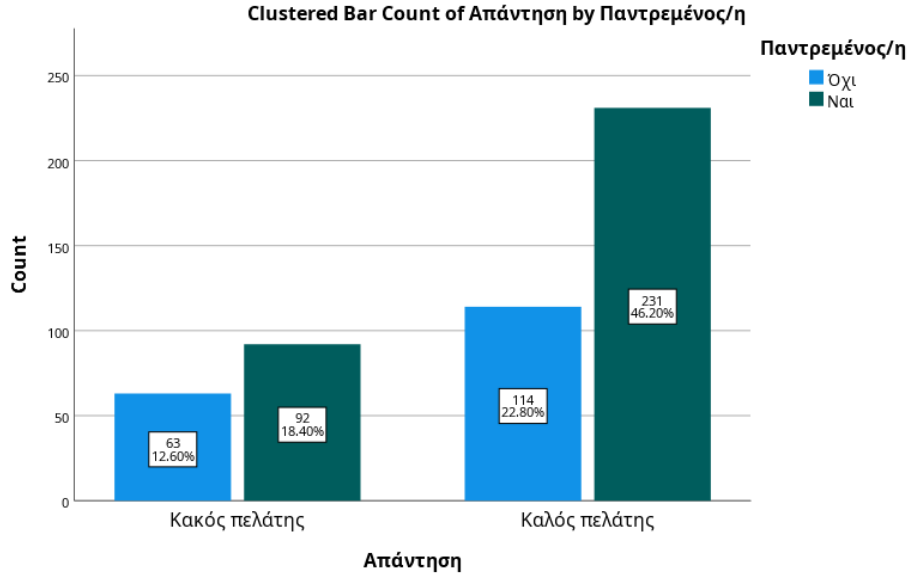
Περιοχή κατοικίας

Απάντηση			Frequency	Percent
Κακός πελάτης	Valid	Αστική	60	38.7
		Ημιαστική	39	25.2
		Αγροτική περιοχή	56	36.1
		Total	155	100.0
Καλός πελάτης	Valid	Αστική	111	32.2
		Ημιαστική	149	43.2
		Αγροτική περιοχή	85	24.6
		Total	345	100.0

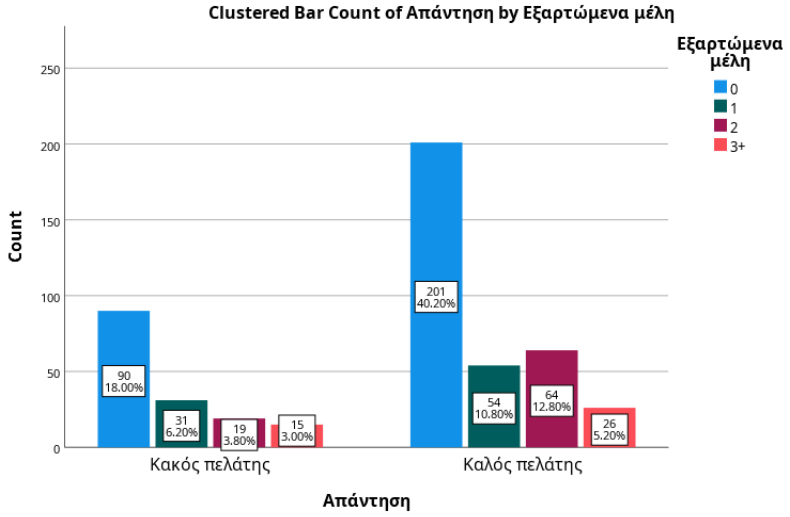
Πίνακας 15: Περιοχή κατοικίας του δείγματος ανά κατηγορία πελάτη



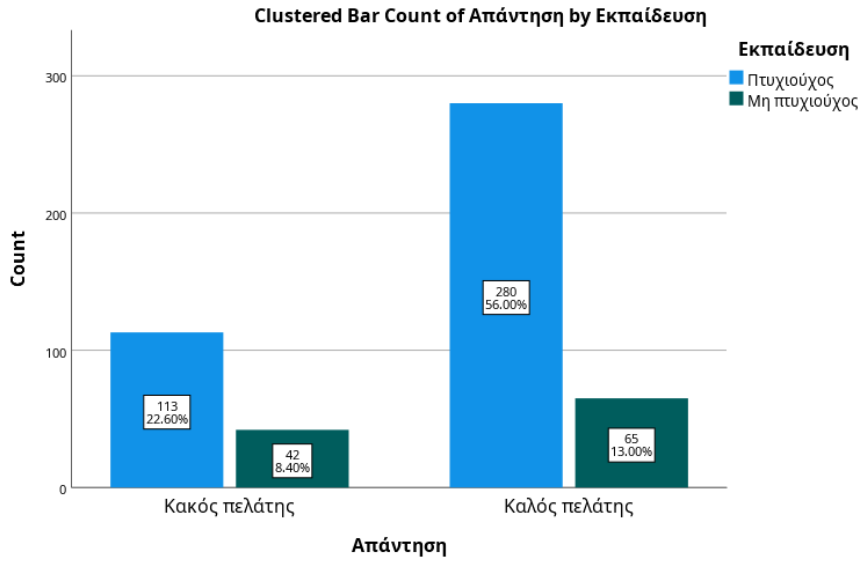
Σχήμα 13: Γένος του δείγματος ανά κατηγορία πελάτη



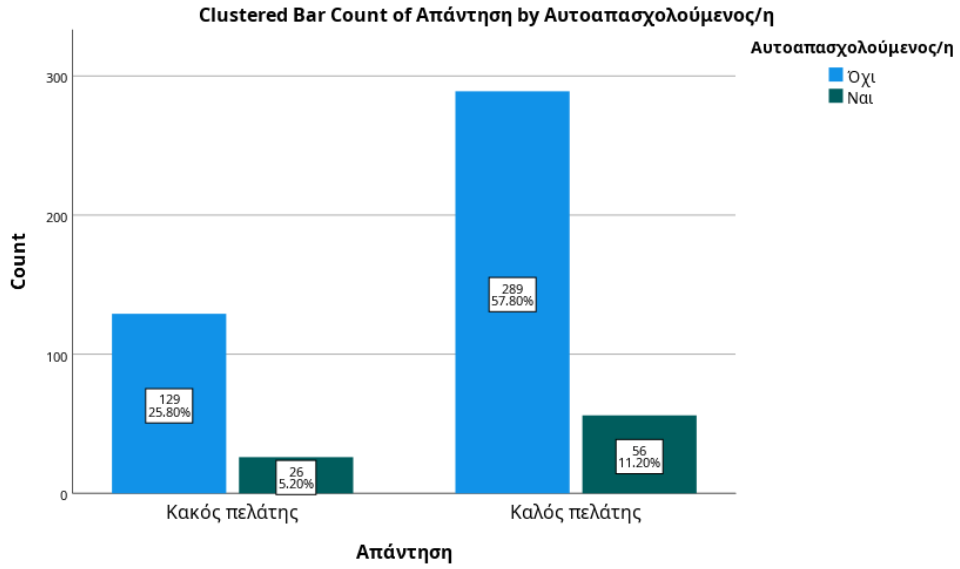
Σχήμα 14: Οικογενειακή κατάσταση του δείγματος ανά κατηγορία πελάτη



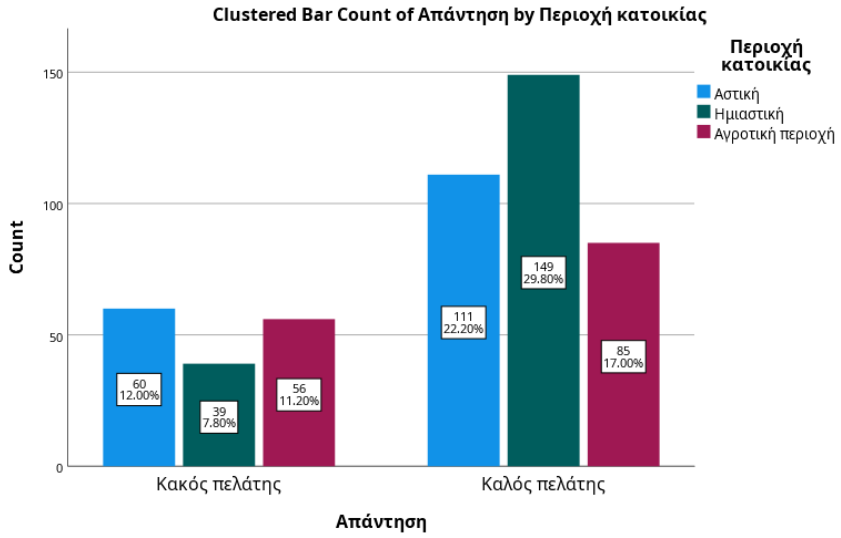
Σχήμα 15: Εξαρτώμενα μέλη του δείγματος ανά κατηγορία πελάτη



Σχήμα 16: Εκπαίδευση του δείγματος ανά κατηγορία πελάτη



Σχήμα 17: Αυτοαπασχόληση του δείγματος ανά κατηγορία πελάτη



Σχήμα 18: Περιοχή κατοικίας του δείγματος ανά κατηγορία πελάτη

Εξετάζοντας τις ποσοτικές μεταβλητές του δείγματος, όταν χωρίσουμε το δείγμα στις κατηγορίες "καλός" και "κακός" πελάτης, παίρνουμε τον Πίνακα 16. Εκεί βλέπουμε πως το ετήσιο εισόδημα των "καλών πελατών" παίρνει τιμές από 207.90€ μέχρι 62703.63€, με μέση τιμή τα 5314.08€ και μεγάλη τυπική απόκλιση (*s*=5973.92€). Το ετήσιο εισόδημα των "κακών πελατών" παίρνει τιμές από 148.50€ μέχρι 80190.00€, με μέση τιμή τα 5716.10€ και μεγάλη τυπική απόκλιση (*s*=7415.80€). Φαίνεται λοιπόν ότι οι κακοί πελάτες έχουν λίγο μεγαλύτερο εισόδημα κατά μέσο όρο και αρκετά μεγαλύτερη τυπική απόκλιση. Το ετήσιο εισόδημα των συνυποψηφίων των "καλών πελατών" παίρνει τιμές μέχρι 19800.00€, με μέση τιμή τα 1485.13€ και τυπική απόκλιση *s*=1950.69€. Αντίστοιχα, το ετήσιο εισόδημα των συνυποψηφίων των "κακών πελατών" παίρνει τιμές μέχρι 19800.00€, με μέση τιμή τα 1504.86€ και τυπική απόκλιση *s*=2442.94€.

Βάζοντας μαζί τα δύο εισοδήματα, το συνολικό εισόδημα των "καλών πελατών" παίρνει τιμές από 1943.37€ μέχρι 62703.63€, με μέση τιμή τα 6799.21€ και τυπική απόκλιση τα 5992.46€, ενώ των "κακών πελατών" παίρνει τιμές από 1427.58€ μέχρι 80190.00€, με μέση τιμή τα 7220.96€ και τυπική απόκλιση τα 7631.59€.

Το ποσό δανείου των "κακών πελατών" παίρνει τιμές από 29700.00€ μέχρι και 564300.00€, με μέση τιμή τις 150250.10€ και τυπική απόκλιση 87325.75€, ενώ των "καλών πελατών" παίρνει τιμές από 12870.00€ μέχρι και 693000.00€, με μέση τιμή τις 137512.40€ και τυπική απόκλιση 792952.50€

Τέλος, η διάρκεια του δανείου που ζήτησαν οι "καλοί" πελάτες παίρνει τιμές από 12 έως 480 μήνες, με μέση τιμή τους 338.96 μήνες και τυπική απόκλιση 64.82 μήνες, ενώ η αντίστοιχη διάρκεια για τους "κακούς" πελάτες κυμαίνεται από 36 μέχρι 480 μήνες, με μέση τιμή 342.81 και τυπική απόκλιση 70.63 μήνες.

Descriptive Statistics

Απάντηση		N	Minimum	Maximum	Mean	Std. Deviation
Κακός πελάτης	Εισόδημα υποψηφίου	155	148.50	80190.00	5716.1003	7415.80541
	Εισόδημα συνυποψηφίου	155	.00	19800.00	1504.8639	2442.94023
	Ποσό δανείου	155	29.70	564.30	150.2501	87.32575
	Διάρκεια του δανείου σε μήνες	155	36	480	342.81	70.637
	Συνολικό εισόδημα	155	1427.58	80190.00	7220.9642	7631.59073
	Valid N (listwise)	155				
Καλός πελάτης	Εισόδημα υποψηφίου	345	207.90	62703.63	5314.0818	5973.91567
	Εισόδημα συνυποψηφίου	345	.00	19800.00	1485.1260	1950.69903
	Ποσό δανείου	345	12.87	693.00	137.5124	79.29525
	Διάρκεια του δανείου σε μήνες	345	12	480	338.96	64.823
	Συνολικό εισόδημα	345	1943.37	62703.63	6799.2079	5992.46519
	Valid N (listwise)	345				

Πίνακας 16: Στατιστικά ποσοτικών μεταβλητών του δείγματος ανά κατηγορία πελάτη

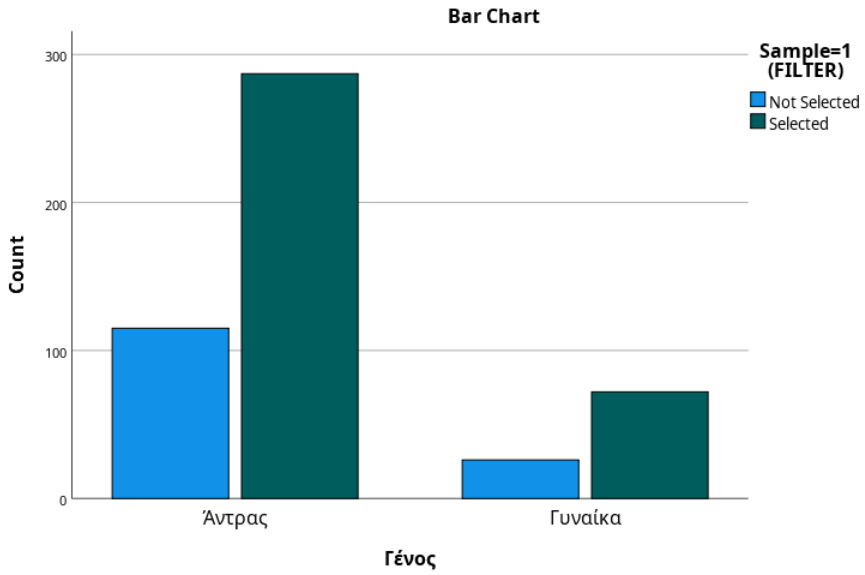
4.2 Δημιουργία δείγματος ανάπτυξης και δείγματος επικύρωσης

Θα χωρίσουμε το δείγμα των 500 πελατών σε δύο κατηγορίες:

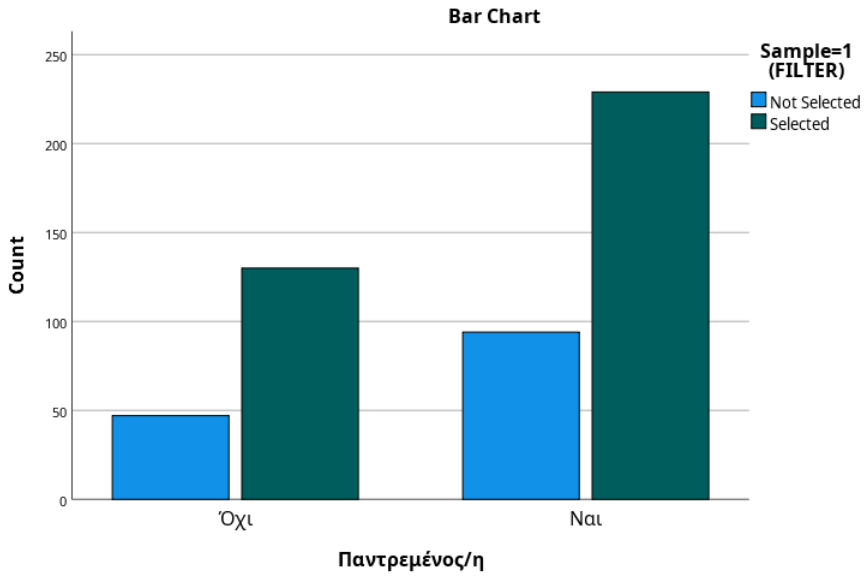
1. το δείγμα ανάπτυξης που χρησιμοποιείται για τη δημιουργία του μοντέλου.
2. το δείγμα επικύρωσης που χρησιμοποιείται για τη δοκιμή της απόδοσης του μοντέλου που δημιουργήσαμε με το δείγμα ανάπτυξη

Συνήθως το δείγμα χωρίζεται με αναλογία 70% στο δείγμα ανάπτυξης και 30% στο δείγμα επικύρωσης. Θα ζητήσουμε από το SPSS να μας διαλέξει στην τύχη το 70% του δείγματος και θα δημιουργήσουμε μία μεταβλητή με όνομα sample. Αυτή η μεταβλητή θα έχει την τιμή 1 αν ο πελάτης ήταν μέρος της 70% τυχαίας επιλογής του SPSS και την τιμή 0 αλλιώς ώστε να κρατήσουμε αυτή την τυχαία επιλογή του SPSS και να την ενεργοποιούμε όποτε χρειάζεται.. Στα Σχήματα 19 έως και 26 φαίνονται τα

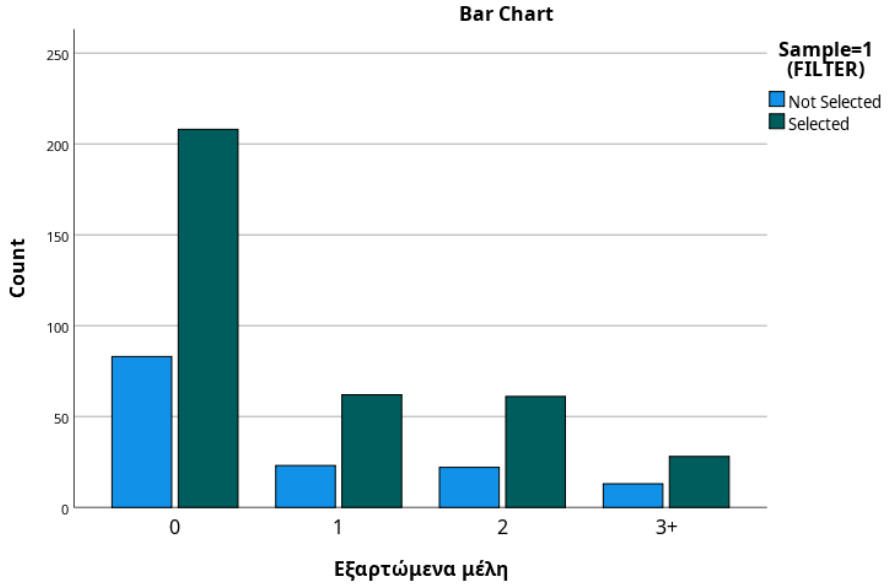
ραβδογράμματα των κατηγορικών μεταβλητών για το δείγμα ανάπτυξης και το δείγμα επικύρωσης.



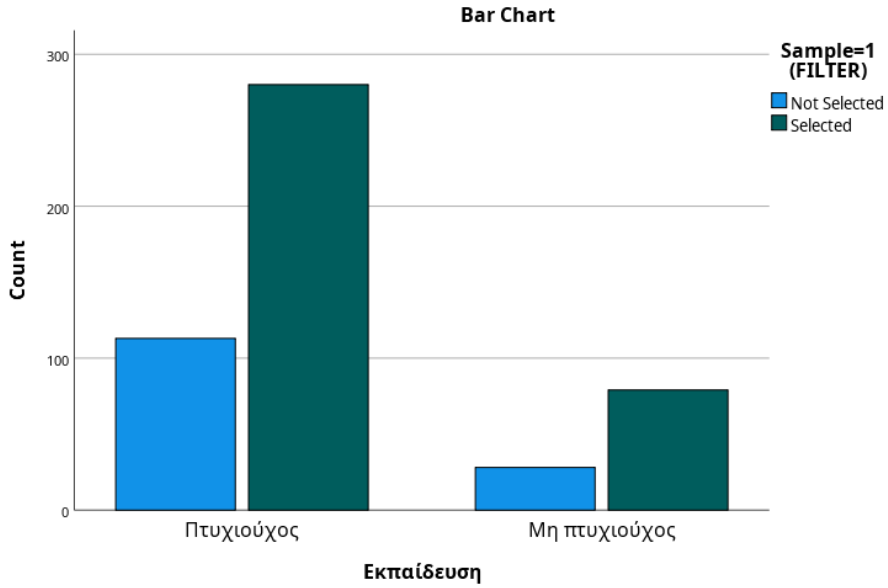
Σχήμα 19: Γένος των δειγμάτων ανάπτυξης και επικύρωσης



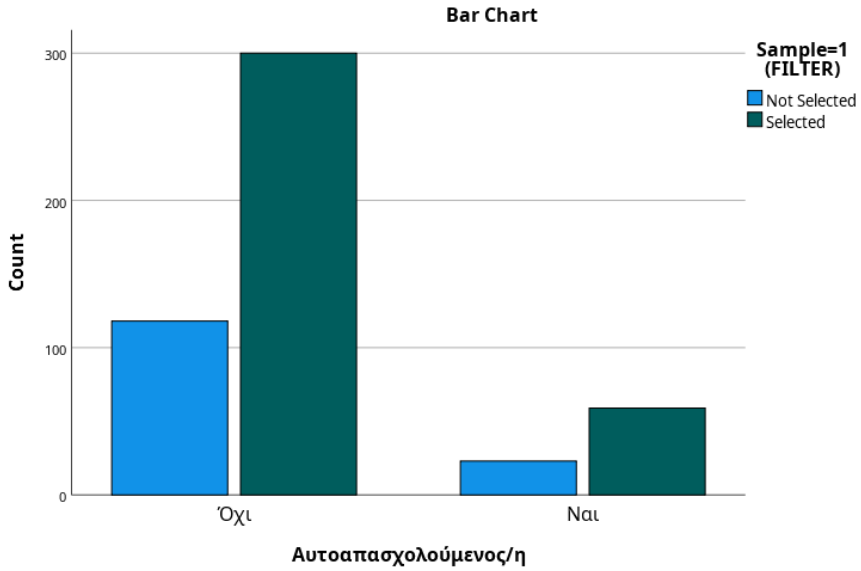
Σχήμα 20: Οικογενειακή κατάσταση των δειγμάτων ανάπτυξης και επικύρωσης



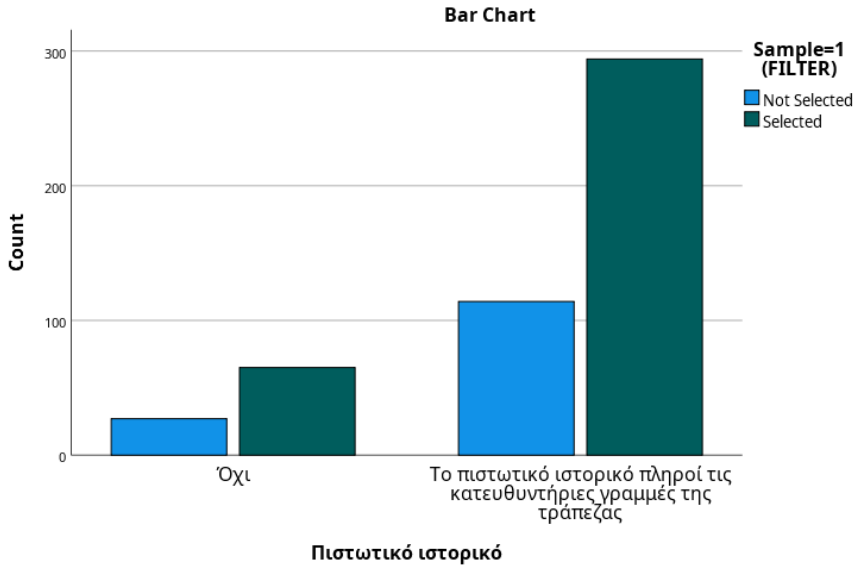
Σχήμα 21: Εξαρτώμενα μέλη των δειγμάτων ανάπτυξης και επικύρωσης



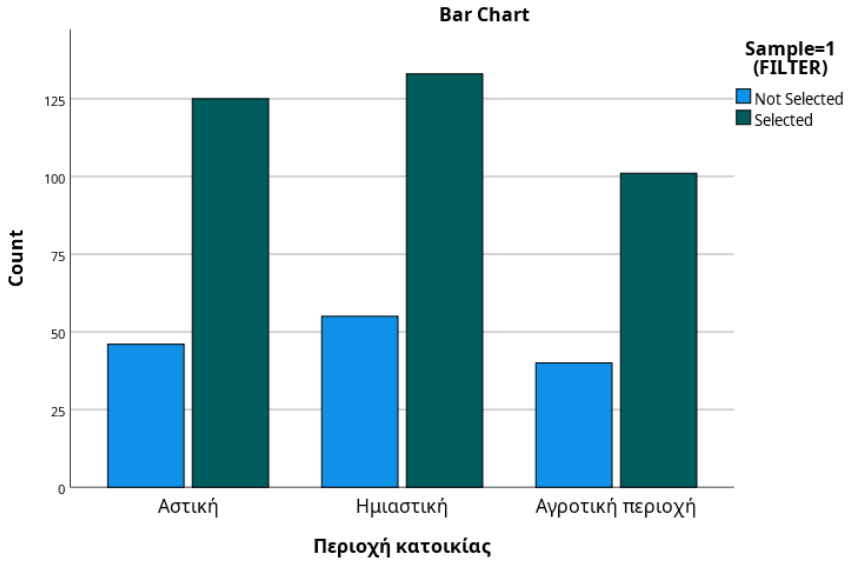
Σχήμα 22: Εκπαίδευση των δειγμάτων ανάπτυξης και επικύρωσης



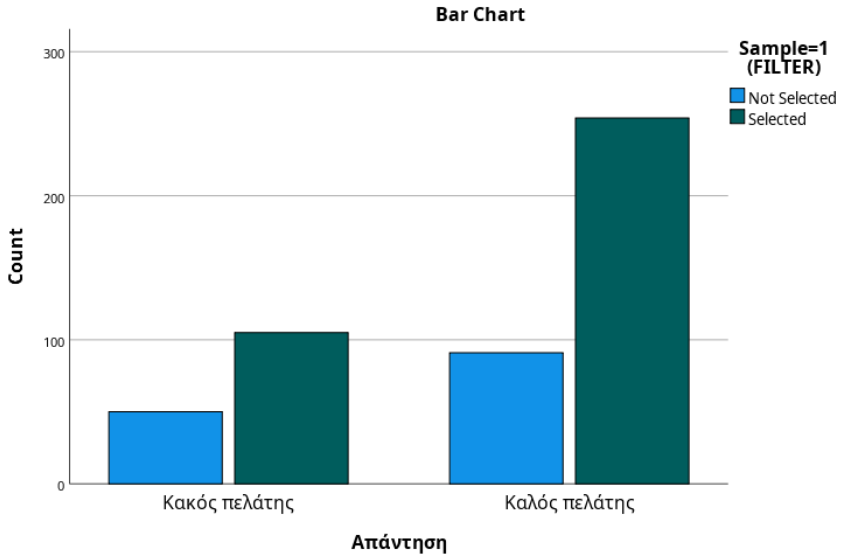
Σχήμα 23: Αυτοαπασχόληση των δειγμάτων ανάπτυξης και επικύρωσης



Σχήμα 24: Πιστωτικό ιστορικό των δειγμάτων ανάπτυξης και επικύρωσης



Σχήμα 25: Περιοχή κατοικίας των δειγμάτων ανάπτυξης και επικύρωσης



Σχήμα 26: Καλοί/κακοί πελάτεςστα δείγματα ανάπτυξης και επικύρωσης

4.3 Λογιστική Παλινδρόμηση

Στόχος μας είναι μπορέσουμε να προβλέψουμε και να κατατάξουμε έναν πελάτη σε μία από τις κατηγορίες “καλός” ή “κακός” χρησιμοποιώντας όλες τις υπόλοιπες μεταβλητές που εξηγήσαμε πριν. Θα χρησιμοποιήσουμε την τεχνική της λογιστικής παλινδρόμησης, για την οποία υπάρχουν οι εξής προϋποθέσεις:

- Να έχουμε ως εξαρτημένη μεταβλητή μία δίτιμη μεταβλητή: Εδώ για εξαρτημένη μεταβλητή έχουμε τη δίτιμη Loan_Status που εκφράζει την κατάταξη ενός πελάτη σε "καλό" ή "κακό".
- Οι ανεξάρτητες μεταβλητές να είναι ποσοτικές ή ποιοτικές: Στα δεδομένα μας όλες οι ανεξάρτητες μεταβλητές (οι μεταβλητές 2 έως και 12, και η 14) είναι ποσοτικές ή ποιοτικές.
- Θα πρέπει να έχουμε ανεξαρτησία ανάμεσα στις εγγραφές και η εξαρτημένη μεταβλητή θα πρέπει να έχει αμοιβαία αποκλειόμενες κατηγορίες: Δεν έχουμε λόγο να υποψιαζόμαστε ότι κάθε εγγραφή (κάθε πελάτης) επηρεάζει κάποιον άλλο πελάτη άρα μπορούμε να θεωρήσουμε ότι δεν υπάρχει θέμα μη-ανεξαρτησίας ανάμεσα στις εγγραφές. Επίσης, η εξαρτημένη μεταβλητή παίρνει αμοιβαία-αποκλειόμενες κατηγορίες που εξαντλούν όλο το δείγμα.

Αφού οι προϋποθέσεις της λογιστικής παλινδρόμησης επαληθεύονται, εφαρμόζουμε τη συγκεκριμένη τεχνική χρησιμοποιώντας το SPSS. Στο μοντέλο που θα δημιουργήσουμε θεωρούμε ως εξαρτημένη μεταβλητή την Loan_Status (απάντηση της τράπεζας) και στις ανεξάρτητες μεταβλητές τις Gender, Married, Dependents, Education, Self_employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area και Total_Income. Για τις κατηγορικές μεταβλητές (Gender, Married, Dependents, Education, Self_employed, Credit_History και Property_Area) χρησιμοποιήσαμε ως κατηγορία αναφοράς την πρώτη, δηλαδή την κατηγορία στην οποία αντιστοιχεί η μικρότερη αριθμητική τιμή κάθε μεταβλητής. Ο Πίνακας 17 μας δείχνει ότι το συνολικό μοντέλο είναι στατιστικά σημαντικό, $x^2(14)=94.225$.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	94.225	14	.000
	Block	94.225	14	.000
	Model	94.225	14	.000

Πίνακας 17: Συνολική αξιολόγηση του μοντέλου

Ο Πίνακας 18 δίνει τις τιμές του Cox & Snell R^2 και Nagelkerke R^2 που λένε πόση από την απόκλιση της εξαρτημένης μεταβλητής μπορεί να εξηγηθεί από το

μοντέλο που δημιούργησε η λογιστική παλινδρόμηση. Οι δύο αυτές τιμές συμπεριφέρονται σχεδόν σαν το R^2, οπότε μπορούμε να πούμε ότι η απόκλιση της εξαρτημένης μεταβλητής που μπορεί να εξηγηθεί από το μοντέλο μας κυμαίνεται ανάμεσα στο 23.1% και το 32.9%. Ο τελευταίος δείκτης είναι πιο αντιπροσωπευτικός και λέει ότι το 32,9% της διακύμανσης στο καλός/κακός πελάτης μπορεί να εξηγηθεί από το μοντέλο μας.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	339.703[a]	.231	.329

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Πίνακας 18: Σύνοψη μοντέλου

Ο Πίνακας 19 δείχνει τα αποτελέσματα του Hosmer and Lemeshow test που ελέγχει τη μηδενική υπόθεση, ότι δηλαδή οι προβλέψεις που κάνει το μοντέλο ταιριάζουν με τις παρατηρούμενες τιμές. Το στατιστικό x^2 υπολογίζεται συγκρίνοντας τις παρατηρούμενες συχνότητες με αυτές που θα περιμέναμε από το γραμμικό μοντέλο. Ένα μη στατιστικά σημαντικό x^2 μας δείχνει ότι τα δεδομένα ταιριάζουν καλά στο μοντέλο. Εδώ, έχουμε sig=.562>.05 άρα έχουμε καλή προσαρμογή του μοντέλου στα δεδομένα.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.768	8	.562

Πίνακας 19: Αποτελέσματα του test Hosmer and Lemeshow

Η λογιστική παλινδρόμηση υπολογίζει την πιθανότητα να χαρακτηριστεί ένας υποψήφιος ως “καλός” ή “κακός” πελάτης. Αν η πιθανότητα αυτή βρεθεί να είναι μεγαλύτερη ή ίση του 0.5 τότε ο πελάτης αυτός θα σημειωθεί ως “καλός” και αν η πιθανότητα είναι μικρότερη του 0.5 τότε ο πελάτης αυτός θα σημειωθεί ως “κακός”. Αυτό που προσπαθούμε να κάνουμε είναι να χρησιμοποιήσουμε τη λογιστική παλινδρόμηση για να ελέγξουμε το αν οι πελάτες μπορούν να χαρακτηριστούν σωστά από τις ανεξάρτητες μεταβλητές. Η λογιστική παλινδρόμηση δημιουργεί ένα τέτοιο

μοντέλο πρόβλεψης το οποίο πρέπει να αξιολογηθεί. Ο Πίνακας 20 είναι μία αξιολόγηση του μοντέλου προβλέψεων. Παρατηρούμε ότι οι κακοί πελάτες του δείγματος ανάπτυξης προβλέφθηκαν επιτυχώς ως κακοί πελάτες από το μοντέλο στο 44.8% (η ιδιαιτερότητα του μοντέλου) των περιπτώσεων (δηλαδή μόνο 47 από τους 105 κακούς πελάτες). Στην κατηγορία των καλών πελατών όμως, το μοντέλο είχε πολύ μεγαλύτερη επιτυχία, αφού έπιασε επιτυχώς το 93.8% (η ευαισθησία του μοντέλου) των καλών πελατών. Επίσης παρατηρούμε ότι 58 κακοί πελάτες κατατάσσονται από το μοντέλο ως καλοί πελάτες, δίνοντας μας τα ψευδώς θετικά (55.2%). Ομοίως, 16 καλοί πελάτες κατατάσσονται από το μοντέλο ως κακοί πελάτες, δίνοντας μας τα ψευδώς αρνητικά (6.3%). Αν βάλουμε μαζί αυτές τις πληροφορίες, το SPSS μας λέει ότι το συνολικό ποσοστό επιτυχών προβλέψεων του μοντέλου είναι 79.4% (η ακρίβεια κατάταξης του μοντέλου) που είναι αρκετά ικανοποιητικό.

Από όλους τους πελάτες που περιμέναμε να είναι καλοί, το 100 x (238 / (238+58)) = 80.40% προβλέφθηκε σωστά από το μοντέλο. Αντίστοιχα, από τους πελάτες που περιμέναμε να είναι κακοί, το 100 x (47 / (47+16)) = 74.60% προβλέφθηκε σωστά από το μοντέλο.

Classification Table[a]

Observed			Predicted		
			Απάντηση		Percentage
			Κακός πελάτης	Καλός πελάτης	Correct
Step 1	Απάντηση	Κακός πελάτης	47	58	44.8
		Καλός πελάτης	16	238	93.7
	Overall Percentage				79.4

a. The cut value is .500

Πίνακας 20: Πίνακας κατάταξης

Ο Πίνακας 21 δείχνει την κατηγορία αναφοράς για κάθε κατηγορική μεταβλητή καθώς και τις κατηγορίες που ονομάστηκαν 1, 2 και 3 για τις μεταβλητές αυτές. Για παράδειγμα, για την μεταβλητή "Περιοχή Κατοικίας" που παίρνει 3 συνολικά τιμές, η κατηγορία αναφοράς είναι αυτή που αντιστοιχεί στην Αστική περιοχή (το καταλαβαίνουμε γιατί έχει την τιμή 0 σε κάθε στήλη του Parameter coding), ενώ η κατηγορία με όνομα "1" είναι αυτή που αντιστοιχεί στην ημιαστική περιοχή και η κατηγορία με όνομα "2" είναι αυτή που αντιστοιχεί στην αγροτική περιοχή.

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
Εξαρτώμενα μέλη	0	208	.000	.000	.000
	1	62	1.000	.000	.000
	2	61	.000	1.000	.000
	3+	28	.000	.000	1.000
Περιοχή κατοικίας	Αστική	125	.000	.000	
	Ημιαστική	133	1.000	.000	
	Αγροτική περιοχή	101	.000	1.000	
Παντρεμένος/η	Όχι	130	.000		
	Ναι	229	1.000		
Εκπαίδευση	Πτυχιούχος	280	.000		
	Μη πτυχιούχος	79	1.000		
Αυτοαπασχολούμενος/η	Όχι	300	.000		
	Ναι	59	1.000		
Πιστωτικό ιστορικό	Όχι	65	.000		
	Το πιστωτικό ιστορικό πληροί τις κατευθυντήριες γραμμές της τράπεζας	294	1.000		
Γένος	Άντρας	287	.000		
	Γυναίκα	72	1.000		

Πίνακας 21: Κωδικοποιήσεις κατηγορικών μεταβλητών

Τέλος, ο Πίνακας 22 δείχνει τη συνεισφορά κάθε μίας από τις ανεξάρτητες μεταβλητές στο μοντέλο της λογιστικής παλινδρόμησης. Η στήλη B δείχνει τις τιμές των συντελεστών παλινδρόμησης, ενώ η στήλη S.E. δείχνει την τυπική απόκλιση (standard error). Η στατιστική Wald δείχνει τη στατιστική σημαντικότητα για κάθε ένα από τους συντελεστές αυτούς: υψηλότερες τιμές του Wald test δείχνουν μεγαλύτερη συνεισφορά στη πρόβλεψη ενός καλού πελάτη, ενώ η στήλη Sig δείχνει και αυτή τη στατιστική σημαντικότητα της συνεισφοράς αυτής. Η στήλη exp(B) (που λέγεται και odds ratio) διευκολύνει την ανάλυση των αποτελεσμάτων, δεδομένου ότι ο συντελεστής παλινδρόμησης B στη λογιστική παλινδρόμηση δείχνει τις μεταβολές του λογαρίθμου των odds της εξαρτημένης μεταβλητής. Έτσι, όταν το B=0, το exp(B) είναι 1, όταν το B είναι θετικό, το exp(B) παίρνει τιμές μεγαλύτερες της μονάδας, ενώ όταν είναι αρνητικό, παίρνει τιμές μικρότερες της μονάδας.

Παρατηρούμε λοιπόν ότι τη μεγαλύτερη συνεισφορά την έχει η μεταβλητή που

δείχνει το πιστωτικό ιστορικό με Wald = 54.517, exp(B) = 13.142 > 1 και Sig. < 0,01 (δηλαδή η ανεξάρτητη μεταβλητή είναι στατιστικώς σημαντική σε επίπεδο 1%). Αυτό μας λέει ότι όταν το πιστωτικό ιστορικό του αιτούντος πληροί τις κατευθυντήριες γραμμές της τράπεζας, τότε οι πιθανότητες να χαρακτηριστεί ως καλός πελάτες αυξάνονται κατά 13.142 φορές σε σχέση με τα άτομα που το πιστωτικό ιστορικό τους δεν πληροί τις κατευθυντήριες γραμμές της τράπεζας.

Ακολουθεί η μεταβλητή που δείχνει την περιοχή κατοικίας του αιτούντα, με Wald = 7.485 και Sig. = 0,024 <0.05. Επειδή η μεταβλητή "Περιοχή Κατοικίας" παίρνει τρεις τιμές, η πρώτη γραμμή για τη μεταβλητή στον Πίνακα 22 μας λέει ότι η μεταβλητή έχει στατιστικά σημαντικά συνεισφορά στην επιτυχημένη πρόβλεψη που κάνει το μοντέλο (αφού sig <0.05) και οι επόμενες δύο γραμμές συγκρίνουν την 2η και την 3η κατηγορία περιοχής κατοικίας με την κατηγορία αναφοράς. Θυμίζουμε ότι ο Πίνακας 21 μας δείχνει για κάθε μεταβλητή ποια είναι η κατηγορία αναφοράς και ποιες οι υπόλοιπες. Εδώ λοιπόν παρατηρούμε ότι το να μένει κάποιος σε ημιαστική περιοχή, με sig=0.023 < 0.05 και exp(B)=2.160, αυξάνει 2.160 φορές την πιθανότητα ενός πελάτη να καταταγεί ως καλός, σε σχέση με κάποιον που μένει στην κατηγορία αναφοράς, δηλαδή σε αστική περιοχή. Όμως, η 3η κατηγορία περιοχής κατοικίας (αγροτικές περιοχές), παρά το ότι όλη η μεταβλητή "Περιοχή Κατοικίας" συνεισφέρει σημαντικά στο μοντέλο, με sig=0.747>0.05 δε φαίνεται να επηρεάζει στατιστικώς σημαντικά το μοντέλο.

Η τελευταία μεταβλητή που έχει στατιστικά σημαντική επιρροή στο μοντέλο είναι αυτή που δείχνει την εκπαίδευση του πελάτη, με Wald = 4.754, Sig = 0.029 < 0.05 και exp(B)=0.492. Δηλαδή, για ύπαρξη πτυχίου αυξάνει την πιθανότητα να χαρακτηριστεί ο πελάτης ως καλός κατά 2.032 φορές, τιμή που προκύπτει αντιστρέφοντας το exp(B).

Για τη μεταβλητή που δείχνει τα εξαρτώμενα μέλη, η κατηγορία αναφοράς είναι η μη ύπαρξη εξαρτώμενων μελών (0), η οποία έχει sig=0.101>0.05 άρα η μη ύπαρξη εξαρτώμενων μελών δεν έχει στατιστικώς σημαντική επίδραση στο μοντέλο. Αν όμως ο αριθμός των εξαρτώμενων μελών αλλάξει από μηδέν και γίνει ένα, τότε με sig=0.037<0.05 η αλλαγή αυτή έχει στατιστικώς σημαντική επίδραση στο μοντέλο και περιγράφει μείωση της πιθανότητας να χαρακτηριστεί ένας πελάτης ως καλός κατά 2.13 φορές. Οι υπόλοιπες δύο κατηγορίες της μεταβλητής που αντιστοιχούν είτε σε δύο εξαρτώμενα μέλη είτε σε τρία ή παραπάνω δεν επηρεάζουν το μοντέλο αφού αν τα εξαρτώμενα μέλη από μηδέν γίνουν δύο, τότε με sig=0.557>0.05 η όποια επίδραση στο

μοντέλο δεν είναι σημαντική και το ίδιο συμβαίνει και για τρία ή παραπάνω μέλη αφού sig=0.724>0.05.

Η μεταβλητή Γένος έχει sig=0.945 > 0.05 άρα η όποια συνεισφορά της στο μοντέλο δεν είναι στατιστικά σημαντική. Ομοίως, η μεταβλητή Παντρεμένος/η, με sig=0.171 > 0.05 δεν έχει κάποια στατιστικά σημαντική συνεισφορά στο μοντέλο. Ομοίως, η μεταβλητή που εκφράζει την αυτοαπασχόληση έχει sig=0.771 > 0.05 άρα δεν έχει κάποια στατιστικά σημαντική συνεισφορά στο μοντέλο. Αυτό που φαίνεται πιο ενδιαφέρον είναι ότι ούτε το εισόδημα του υποψηφίου ούτε αυτό του συνυποψηφίου φαίνεται να έχουν στατιστικά σημαντική συνεισφορά στο μοντέλο, αφού το sig=0.762>0.05 για το εισόδημα υποψηφίου και το sig=0.799>0.05 για το εισόδημα συνυποψηφίου. Αν εκτελέσουμε νέα λογιστική παλινδρόμηση χωρίς τις δύο αυτές μεταβλητές, αφήνοντας όμως μέσα το συνολικό εισόδημα των υποψηφίων, τότε όπως δείχνει ο Πίνακας 22, το Cox & Snell R^2 όπως και το Nagelkerte R^2 παραμένουν ίδια. Βλέπουμε ότι το -2 Log likelihood αυξάνεται ελάχιστα κάτι που δείχνει ότι το μοντέλο αποδίδει λίγο χειρότερα από το προηγούμενο. Ο Πίνακας 24 δείχνει τα αποτελέσματα του Hosmer and Lemeshow test για το νέο μοντέλο και παρατηρούμε ότι με sig=.864>.05 δεν έχουμε λόγους να πιστεύουμε ότι το νέο μοντέλο έχει κακή προσαρμογή στα δεδομένα. Ο Πίνακας 25 δείχνει τη συνεισφορά των ανεξάρτητων μεταβλητών στο μοντέλο που δεν έχει τα επιμέρους εισοδήματα των υποψηφίων. Παρατηρούμε ότι η μεταβλητή που εκφράζει το συνολικό εισόδημα των υποψήφιων παραμένει μη-στατιστικά σημαντική, με sig=0.771>0.05 άρα μπορούμε να πούμε ότι το εισόδημα των υποψηφίων, είτε όταν το δούμε σαν άθροισμα των δύο εισοδημάτων (υποψηφίου, συνυποψηφίου) είτε σαν δύο διαφορετικές μεταβλητές δε φαίνεται να έχει σημαντική επίδραση στο μοντέλο, αν και το μοντέλο που περιέχει τα δύο εισοδήματα αποδίδει λίγο καλύτερα. Θα συνεχίσουμε λοιπόν με το αρχικό μοντέλο και εκεί (Πίνακας 22) βλέπουμε ότι η μεταβλητή που εκφράζει τη διάρκεια του δανείου σε μήνες έχει sig=0.073 > 0.05 άρα δεν έχει κάποια στατιστικά σημαντική συνεισφορά στο μοντέλο. Τέλος, η μεταβλητή "Ποσό Δανείου", με sig=0.207 > 0.05 δεν έχει κάποια στατιστικά σημαντική συνεισφορά στο μοντέλο.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Γένος(1)	-.025	.362	.005	1	.945	.975	.480	1.984
	Παντρεμένος/η(1)	.446	.325	1.878	1	.171	1.561	.826	2.953
	Εξαρτώμενα μέλη			6.233	3	.101			
	Εξαρτώμενα μέλη(1)	-.758	.364	4.333	1	.037	.468	.229	.957
	Εξαρτώμενα μέλη(2)	.252	.430	.345	1	.557	1.287	.554	2.988
	Εξαρτώμενα μέλη(3)	.196	.554	.125	1	.724	1.216	.410	3.603
	Εκπαίδευση(1)	-.708	.325	4.754	1	.029	.492	.260	.931
	Αυτοαπασχολούμενος/η(1)	-.106	.364	.085	1	.771	.899	.441	1.835
	Εισόδημα υποψηφίου	.000	.000	.091	1	.762	1.000	1.000	1.000
	Εισόδημα συνυποψηφίου	.000	.000	.065	1	.799	1.000	1.000	1.000
	Ποσό δανείου	-.003	.002	1.595	1	.207	.997	.993	1.001
	Διάρκεια του δανείου σε μήνες	-.004	.002	3.211	1	.073	.996	.991	1.000
	Πιστωτικό ιστορικό(1)	2.576	.349	54.517	1	.000	13.142	6.633	26.038
	Περιοχή κατοικίας			7.485	2	.024			
	Περιοχή κατοικίας(1)	.770	.339	5.160	1	.023	2.160	1.111	4.199
	Περιοχή κατοικίας(2)	-.106	.329	.104	1	.747	.899	.472	1.713
	Constant	.436	.891	.240	1	.624	1.547		

a. Variable(s) entered on step 1: Γένος, Παντρεμένος/η, Εξαρτώμενα μέλη, Εκπαίδευση, Αυτοαπασχολούμενος/η, Εισόδημα υποψηφίου, Εισόδημα συνυποψηφίου, Ποσό δανείου, Διάρκεια του δανείου σε μήνες, Πιστωτικό ιστορικό, Περιοχή κατοικίας.

Πίνακας 22: Συνεισφορά των ανεξάρτητων μεταβλητών στο μοντέλο

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	339.844[a]	.231	.329

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Πίνακας 23: Σύνοψη μοντέλου χωρίς εισοδήματα υποψηφίου/συνυποψηφίου

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.928	8	.864

Πίνακας 24: Αποτελέσματα του test Hosmer and Lemeshow για το μοντέλο χωρίς εισοδήματα υποψηφίου/συνυποψηφίου

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Γένος(1)	-.020	.362	.003	1	.956	.980	.482	1.992
	Παντρεμένος/η(1)	.424	.320	1.751	1	.186	1.527	.816	2.861
	Εξαρτώμενα μέλη			6.180	3	.103			
	Εξαρτώμενα μέλη(1)	-.754	.364	4.295	1	.038	.471	.231	.960
	Εξαρτώμενα μέλη(2)	.242	.429	.319	1	.572	1.274	.550	2.953
	Εξαρτώμενα μέλη(3)	.216	.552	.153	1	.696	1.241	.421	3.659
	Εκπαίδευση(1)	-.707	.325	4.745	1	.029	.493	.261	.932
	Αυτοαπασχολούμενος/η (1)	-.106	.365	.084	1	.772	.900	.440	1.839
	Ποσό δανείου	-.003	.002	1.684	1	.194	.997	.993	1.001
	Διάρκεια του δανείου σε μήνες	-.004	.002	3.267	1	.071	.996	.991	1.000
	Πιστωτικό ιστορικό(1)	2.584	.348	55.055	1	.000	13.244	6.693	26.205
	Περιοχή κατοικίας			7.404	2	.025			
	Περιοχή κατοικίας(1)	.762	.338	5.074	1	.024	2.142	1.104	4.157
	Περιοχή κατοικίας(2)	-.110	.329	.112	1	.738	.896	.470	1.706
	Συνολικό εισόδημα	.000	.000	.085	1	.771	1.000	1.000	1.000
	Constant	.427	.888	.231	1	.631	1.533		

a. Variable(s) entered on step 1: Γένος, Παντρεμένος/η, Εξαρτώμενα μέλη, Εκπαίδευση, Αυτοαπασχολούμενος/η, Ποσό δανείου, Διάρκεια του δανείου σε μήνες, Πιστωτικό ιστορικό, Περιοχή κατοικίας, Συνολικό εισόδημα.

Πίνακας 25: Συνεισφορά των ανεξάρτητων μεταβλητών στο μοντέλο (χωρίς εισοδήματα υποψηφίου/συνυποψηφίου)

4.4.Αξιολόγηση του μοντέλου με το δείγμα επικύρωσης

Ένας τρόπος για να αξιολογηθεί η απόδοση του μοντέλου που δημιουργήσαμε είναι να υπολογίσουμε τα ποσοστά σωστής και λανθασμένης ταξινόμησης με κριτήριο το δείγμα ελέγχου. Ο Πίνακας 27 δείχνει τον πίνακα ταξινόμησης ταυτόχρονα για το μοντέλο που σχηματίστηκε από το δείγμα ανάπτυξης και για το δείγμα ελέγχου που επιχειρεί να προβλέψει το μοντέλο. Όπως δείχνει ο Πίνακας 27 το μοντέλο που σχηματίστηκε από το δείγμα ανάπτυξης και εφαρμόζεται στο δείγμα ελέγχου, πιάνει σωστά το 93.4% των καλών πελατών, δηλαδή χαρακτηρίζει σωστά τους πελάτες που είναι "καλοί" σε ποσοστό 93.4%. Όταν όμως ένας πελάτης είναι "κακός", το μοντέλο τον προβλέπει ως "κακό" μόνο στο 32% των περιπτώσεων, δηλαδή δίνει ψευδώς θετική πρόβλεψη στο 68%. Το συνολικό ποσοστό ακρίβειας του τελικού μοντέλου που

συμψηφίζει τους πελάτες που ταξινομούνται σωστά ως "καλοί" και τους πελάτες που ταξινομούνται σωστά ως "κακοί" είναι 71.6%.

Classification Table[a]

Observed			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			Απάντηση		Percentage	Απάντηση		Percentage
			Κακός πελάτης	Καλός πελάτης	Correct	Κακός πελάτης	Καλός πελάτης	Correct
Step 1	Απάντηση	Κακός πελάτης	47	58	44.8	16	34	32.0
		Καλός πελάτης	16	238	93.7	6	85	93.4
	Overall Percentage				79.4			71.6

a. The cut value is .500

b. Selected cases Approximately 70% of the cases (SAMPLE) EQ 1

c. Unselected cases Approximately 70% of the cases (SAMPLE) NE 1

Πίνακας 26: Πίνακας ταξινόμησης για το δείγμα ελέγχου

Η λογιστική παλινδρόμηση δημιουργεί ένα μοντέλο με σκοπό να προβλέψει την πιθανότητα ένας πελάτης να είναι "καλός". Αν η πιθανότητα αυτή υπολογιστεί από το 0.5 και πάνω τότε ο πελάτης αυτός χαρακτηρίζεται ως "καλός", αλλιώς ως "κακός". Η τιμή του 0.5 λέγεται cut point (είναι η προεπιλεγμένη τιμή στη λογιστική παλινδρόμηση) και ορίζει το "όριο" πάνω από το οποίο ένας πελάτης πηγαίνει στην επόμενη βαθμίδα που στην περίπτωση μας είναι το "καλός". Το πρόβλημα είναι ότι η επιλογή του 0.5 ως τέτοιο όριο είναι αυθαίρετη και δεν είναι πάντα σωστή. Στόχος είναι να βρεθεί το κατάλληλο cut point ώστε να υπάρχει μεγαλύτερη ισορροπία ανάμεσα στις πιθανότητες εμφάνισης ψευδών θετικών και ψευδών αρνητικών συμβάντων. Για παράδειγμα, ο Πίνακας 27 δείχνει τον πίνακα ταξινόμησης για το μοντέλο λογιστικής παλινδρόμησης που δημιουργήθηκε με cut point στο 0.6. Παρατηρούμε ότι η πιθανότητα εμφάνισης ψευδώς θετικού έπεσε από το 68% στο 56% και αυξήθηκε ελαφρά η πιθανότητα εμφάνισης ψευδώς αρνητικού (από 6.6% στο 8.8%). Οι αλλαγές αυτές αύξησαν την συνολική πιθανότητα επιτυχίας του μοντέλου, από το 71.6% στο 74.5%. Δοκιμάστηκαν και άλλα cut points, όπως το 0.4 και το 0.7, όμως δεν έδωσαν καλύτερα αποτελέσματα από το μοντέλο με cut point στο 0.6.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			Απάντηση		Percentage	Απάντηση		Percentage
	Observed		Κακός πελάτης	Καλός πελάτης	Correct	Κακός πελάτης	Καλός πελάτης	Correct
Step 1	Απάντηση	Κακός πελάτης	54	51	51.4	22	28	44.0
		Καλός πελάτης	25	229	90.2	8	83	91.2
	Overall Percentage				78.8			74.5

a. The cut value is .600

b. Selected cases Approximately 70% of the cases (SAMPLE) EQ 1

c. Unselected cases Approximately 70% of the cases (SAMPLE) NE 1

Πίνακας 27: Πίνακας ταξινόμησης για το δείγμα ελέγχου με cut point στο 0.6

4.5 Δέντρα ταξινόμησης

Τα δέντρα ταξινόμησης βοηθούν στην αναγνώριση των χαρακτηριστικών ομάδων, εξετάζουν τις σχέσεις ανάμεσα στις ανεξάρτητες μεταβλητές και πώς αυτές σχετίζονται με την εξαρτημένη μεταβλητή. Η παρουσίαση των αποτελεσμάτων γίνεται με πιο φιλικό, οπτικό τρόπο.

Θα δημιουργήσουμε δέντρα ταξινόμησης για να δούμε ποιες από τις ανεξάρτητες μεταβλητές συνεισφέρουν στην κατάταξη ενός πελάτη ως καλού. Όπως και πριν, θα χρησιμοποιήσουμε το 70% τυχαίο δείγμα για τη δημιουργία του δέντρου ταξινόμησης και το υπόλοιπο 30% του δείγματος ως δείγμα επικύρωσης. Εκτελώντας ένα CRT δέντρο ταξινόμησης με μέγιστο βάθος τους 5 κόμβους παίρνουμε το δέντρο ταξινόμησης με γενικά χαρακτηριστικά όπως φαίνονται στον Πίνακα 28. Ζητήσαμε από το SPSS να κάνει επαλήθευση με το υπόλοιπο 30% του δείγματος (όπως το δείχνει η μεταβλητή στην αποθηκεύσαμε την 70% αρχική τυχαία επιλογή που έγινε). Επίσης, μειώσαμε τον απαιτούμενο αριθμό ελάχιστων περιπτώσεων σε γονικό κόμβο σε 40 και σε κόμβο-απόγονο σε 10, αφού το δείγμα μας έχει μόνο 350 περιπτώσεις που θα αξιοποιηθούν για τη δημιουργία του μοντέλου. Παρατηρούμε λοιπόν ότι μόνο 3 από τις 12 ανεξάρτητες μεταβλητές έμειναν στο δέντρο και άρα είναι αυτές που είχαν σημαντική συνεισφορά στο μοντέλο πρόβλεψης. Οι μεταβλητές αυτές είναι οι: “Πιστωτικό ιστορικό”, “Εισόδημα υποψηφίου” και “Συνολικό εισόδημα”. Το δέντρο έχει 5 κόμβους, εκ των οποίων οι 3 είναι τελικοί. Το βάθος του δέντρου είναι 2 κόμβοι.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	Απάντηση
	Independent Variables	Γένος, Παντρεμένος/η, Εξαρτώμενα μέλη, Εκπαίδευση, Αυτοαπασχολούμενος/η, Εισόδημα υποψηφίου, Εισόδημα συνυποψηφίου, Ποσό δανείου, Διάρκεια του δανείου σε μήνες, Πιστωτικό ιστορικό, Περιοχή κατοικίας, Συνολικό εισόδημα
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	40
	Minimum Cases in Child Node	10
Results	Independent Variables Included	Πιστωτικό ιστορικό, Εισόδημα υποψηφίου, Συνολικό εισόδημα
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Πίνακας 28: Γενικά χαρακτηριστικά του δέντρου ταξινόμησης

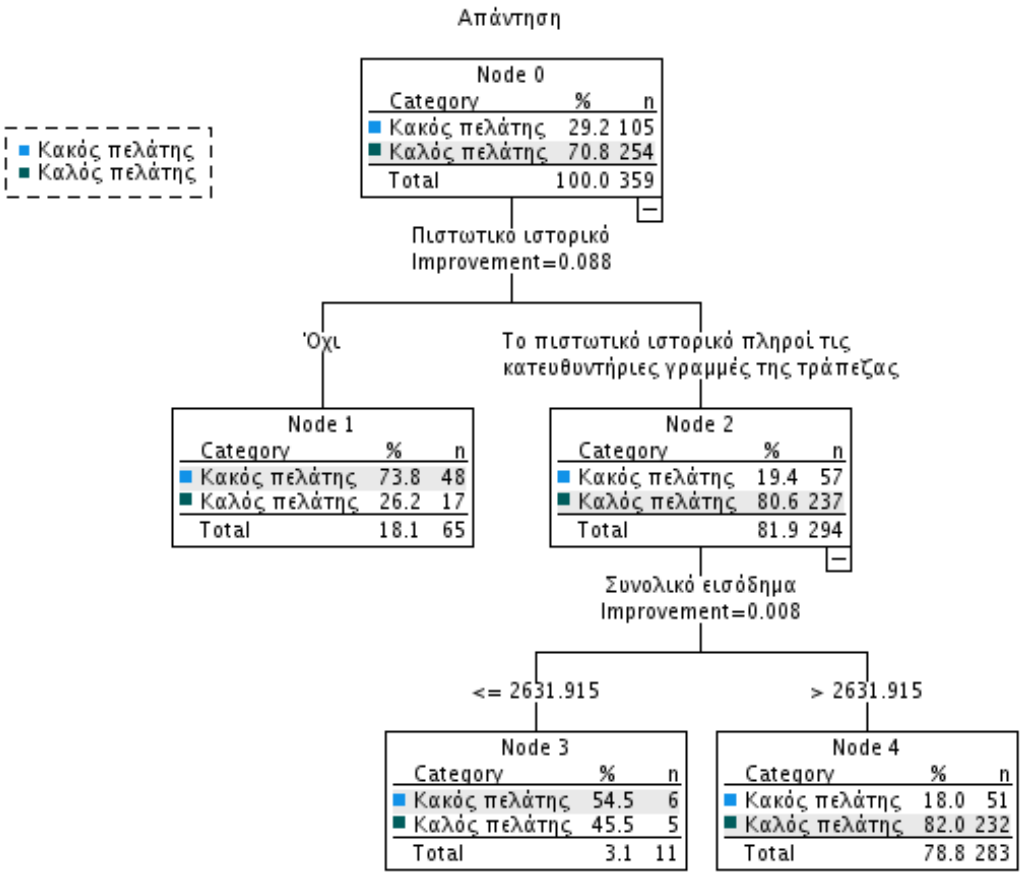
Το Σχήμα 27 δείχνει το δέντρο ταξινόμησης που δημιούργησε το SPSS. Οι διαιρέσεις των κόμβων γίνονται με σειρά σπουδαιότητας. Εδώ, η πιο σημαντική μεταβλητή είναι το "Πιστωτικό Ιστορικό", που χωρίζει τους 359 πελάτες του δείγματος σε δύο υποκατηγορίες, μία που περιλαμβάνει όσους έχουν πιστωτικό ιστορικό που πληροί τις κατευθυντήριες γραμμές της τράπεζας (294 πελάτες) και σε όσους δε τις πληρούν (65 πελάτες). Οι πελάτες που πληρούν τις κατευθυντήριες γραμμές της τράπεζας χωρίζονται σε άλλες δύο κατηγορίες ανάλογα με το συνολικό τους εισόδημα. Έτσι έχουμε συνολικά τρεις τελικούς κόμβους. Σε κάθε τελικό κόμβο φαίνεται το ποσοστό των περιπτώσεων που προβλέπεται να είναι "καλοί" ή "κακοί" πελάτες, και το SPSS υποδεικνύει ποια από τις δύο κατηγορίες υπερισχύει σε κάθε τέτοιο κόμβο. Επίσης, από το μονοπάτι που ακολουθήσαμε για να φτάσουμε στον κόμβο αυτό μπορούμε να καταλάβουμε τις προϋποθέσεις για να φτάσουμε εκεί. Αναλυτικά, οι τελικοί κόμβοι είναι οι εξής:

1. Ο τελικός κόμβος 1 κατατάσσει το 73.8% όσων δεν έχουν πιστωτικό ιστορικό που να πληροί τις κατευθυντήριες γραμμές της τράπεζας ως "κακούς" πελάτες.
2. Ο τελικός κόμβος 3 κατατάσσει το 54.5% όσων έχουν πιστωτικό ιστορικό που να πληροί τις κατευθυντήριες γραμμές της τράπεζας και ταυτόχρονα έχουν συνολικό ετήσιο εισόδημα μικρότερο ή ίσο των 2631.92€ ως "κακούς" πελάτες. Στον κόμβο αυτό βλέπουμε ότι συνολικά υπάρχουν μόνο 11 πελάτες και αυτό

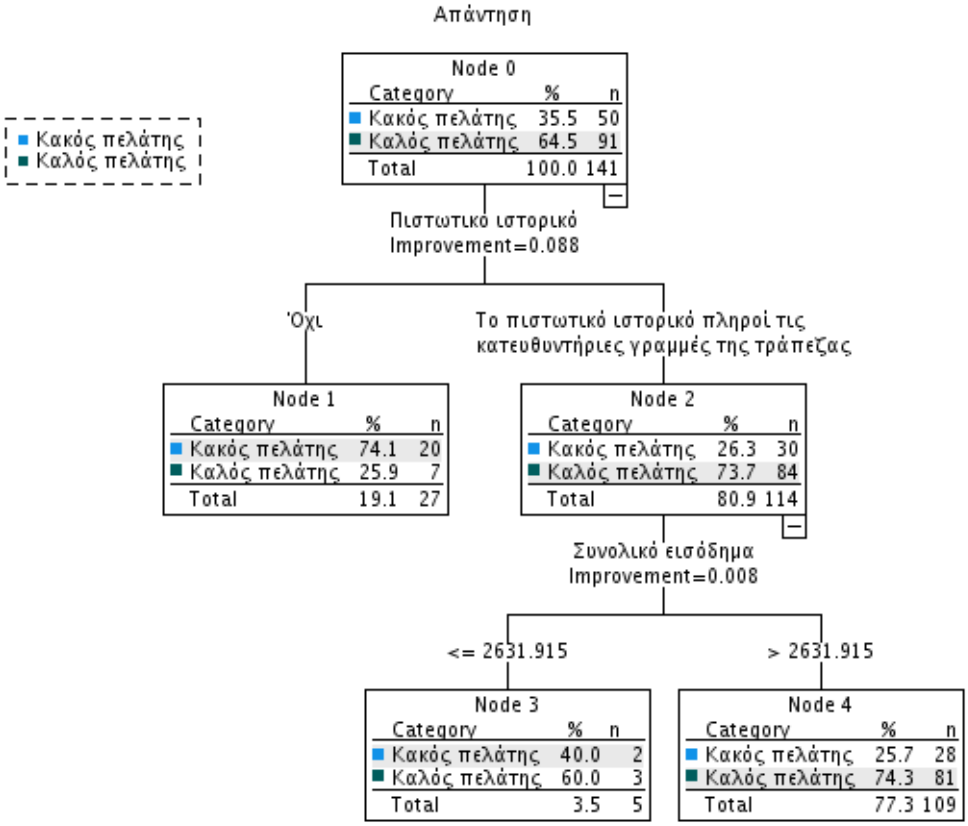
έχει να κάνει με το ότι μειώσαμε τον ελάχιστο αριθμό περιπτώσεων που απαιτεί ένας κόμβος στο SPSS, λόγω του σχετικά μικρού δείγματος.

3. Τέλος, ο τελικός κόμβος 4 κατατάσσει το 82% όσων έχουν πιστωτικό ιστορικό που να πληροί τις κατευθυντήριες γραμμές της τράπεζας και ταυτόχρονα έχουν συνολικό ετήσιο εισόδημα μεγαλύτερο των 2631.92€ ως "καλούς" πελάτες.

Το αντίστοιχο δέντρο ταξινόμησης για το δείγμα ελέγχου φαίνεται στο Σχήμα 28.



Σχήμα 27: Δέντρο ταξινόμησης για το δείγμα ανάπτυξης



Σχήμα 28: Δέντρο ταξινόμησης για το δείγμα ελέγχου

Ο Πίνακας 29 δείχνει στις στήλες Node τον αριθμό του κόμβου και δίπλα το πλήθος των πελατών που βρίσκονται στον κόμβο αυτό αλλά και το ποσοστό τους σε σχέση με το σύνολο των πελατών. Στις στήλες Gain φαίνεται το πλήθος των "καλών" πελατών που βρίσκονται στον κόμβο αυτό αλλά και το ποσοστό τους σε σχέση με το σύνολο των "καλών" πελατών. Η στήλη Response μας δίνει το αναμενόμενο ποσοστό "καλών" πελατών που αντιστοιχούν στον κόμβο αυτό. Τέλος, η στήλη Index δίνει το λόγο του ποσοστού των "καλών" πελατών που αντιστοιχούν στον κόμβο αυτό προς το ποσοστό των πελατών του κόμβου. Τιμή Index μεγαλύτερη του 100% σε κάποιο κόμβο σημαίνει ότι στον κόμβο αυτό υπάρχουν περισσότεροι "καλοί" πελάτες σε σχέση με το ολικό ποσοστό των "καλών" πελατών. Αντίστοιχα, τιμή index μικρότερη του 100% σε κάποιο κόμβο σημαίνει ότι στον κόμβο αυτό υπάρχουν λιγότεροι "καλοί" πελάτες σε σχέση με το ολικό ποσοστό των "καλών" πελατών.

Gains for Nodes

Sample	Node	Node N	Node Percent	Gain N	Gain Percent	Response	Index
Training	4	283	78.8%	232	91.3%	82.0%	115.9%
	3	11	3.1%	5	2.0%	45.5%	64.2%
	1	65	18.1%	17	6.7%	26.2%	37.0%
Test	4	109	77.3%	81	89.0%	74.3%	115.1%
	3	5	3.5%	3	3.3%	60.0%	93.0%
	1	27	19.1%	7	7.7%	25.9%	40.2%

Growing Method: CRT

Dependent Variable: Απάντηση

Πίνακας 29: Πίνακας κέρδους για τους "καλούς" πελάτες

Το "ρίσκο" εκφράζει την πιθανότητα ένας πελάτης να ταξινομηθεί λανθασμένα σύμφωνα με το μοντέλο που δημιουργήθηκε. Όπως δείχνει ο Πίνακας 30, το ρίσκο αυτό για το μοντέλο μας είναι 20.3%. Όταν δοκιμάσουμε το μοντέλο αυτό στο δείγμα ελέγχου, τότε η πιθανότητα λανθασμένης ταξινόμησης ανεβαίνει στο 27%.

Risk

Sample	Estimate	Std. Error
Training	.203	.021
Test	.270	.037

Growing Method: CRT

Dependent Variable: Απάντηση

Πίνακας 30: Πιθανότητα λανθασμένης ταξινόμησης

Ο Πίνακας 31 δίνει τα ποσοστά σωστής ταξινόμησης του μοντέλου, τόσο με το δείγμα ανάπτυξης όσο και με το δείγμα ελέγχου. Παρατηρούμε ότι το ποσοστό των πελατών που κατατάχθηκαν ως "καλοί" στο δείγμα ανάπτυξης είναι 91.3% ενώ το συνολικό ποσοστό ακρίβειας του μοντέλου είναι 79.7%. Όταν δοκιμάσουμε το μοντέλο στο δείγμα ελέγχου τότε το το ποσοστό των πελατών που κατατάχθηκαν ως "καλοί" πέφτει ελάχιστα στο 89% ενώ αντίστοιχα, το συνολικό ποσοστό ακρίβειας του μοντέλου πέφτει στο 73%. Αυτό μας δείχνει ότι για τη μεταβλητή που δείχνει την απάντηση της τράπεζας, οι παρατηρούμενες και οι εκτιμώμενες τιμές της συμφωνούν στο 73% του συνόλου των παρατηρήσεων για το δείγμα ελέγχου.

Classification

Sample	Observed	Predicted: Κακός πελάτης	Predicted: Καλός πελάτης	Percent Correct
Training	Κακός πελάτης	54	51	51.4%
	Καλός πελάτης	22	232	91.3%
	Overall Percentage	21.2%	78.8%	79.7%
Test	Κακός πελάτης	22	28	44.0%
	Καλός πελάτης	10	81	89.0%
	Overall Percentage	22.7%	77.3%	73.0%

Growing Method: CRT

Dependent Variable: Απάντηση

Πίνακας 31: Πίνακας ταξινόμησης

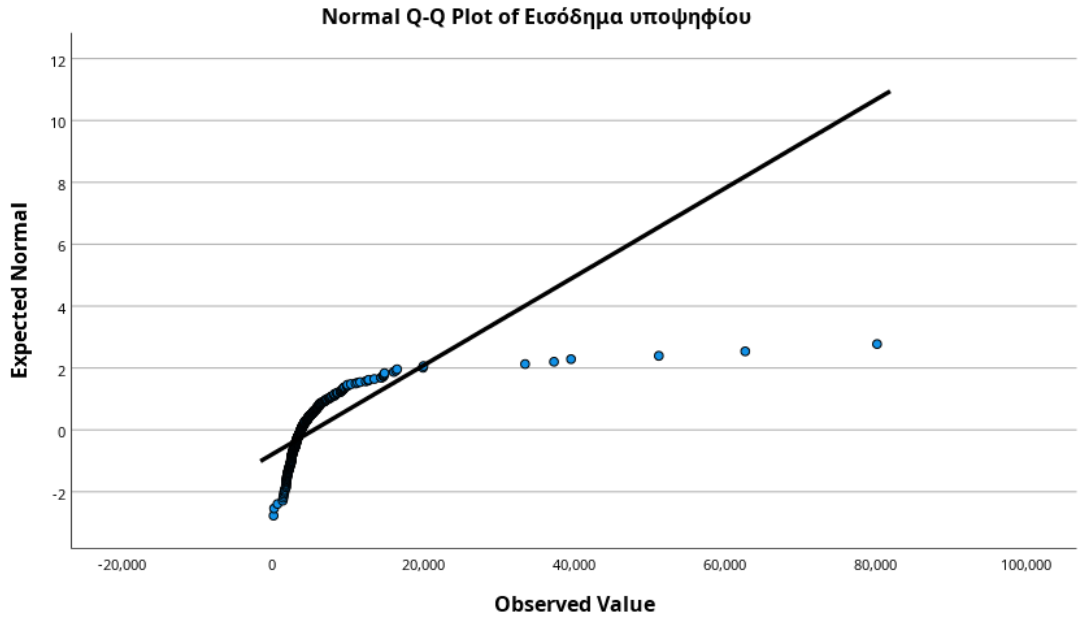
4.5 Διαχωριστική ανάλυση

Η διαχωριστική ανάλυση δημιουργεί ένα μοντέλο που προσπαθεί να προβλέψει την κατηγορία που θα ανήκει κάποιο μέλος. Το μοντέλο δημιουργείται από μια διαχωριστική συνάρτηση (ή ένα σύνολο διαχωριστικών συναρτήσεων αν έχουμε πάνω από δύο κατηγορίες) που βασίζεται σε γραμμικούς συνδυασμούς των ανεξάρτητων μεταβλητών που δίνουν την καλύτερη διάκριση ανάμεσα στις κατηγορίες. Οι συναρτήσεις αυτές θα δημιουργηθούν από το δείγμα ανάπτυξης και θα αξιολογηθούν από το δείγμα ελέγχου.

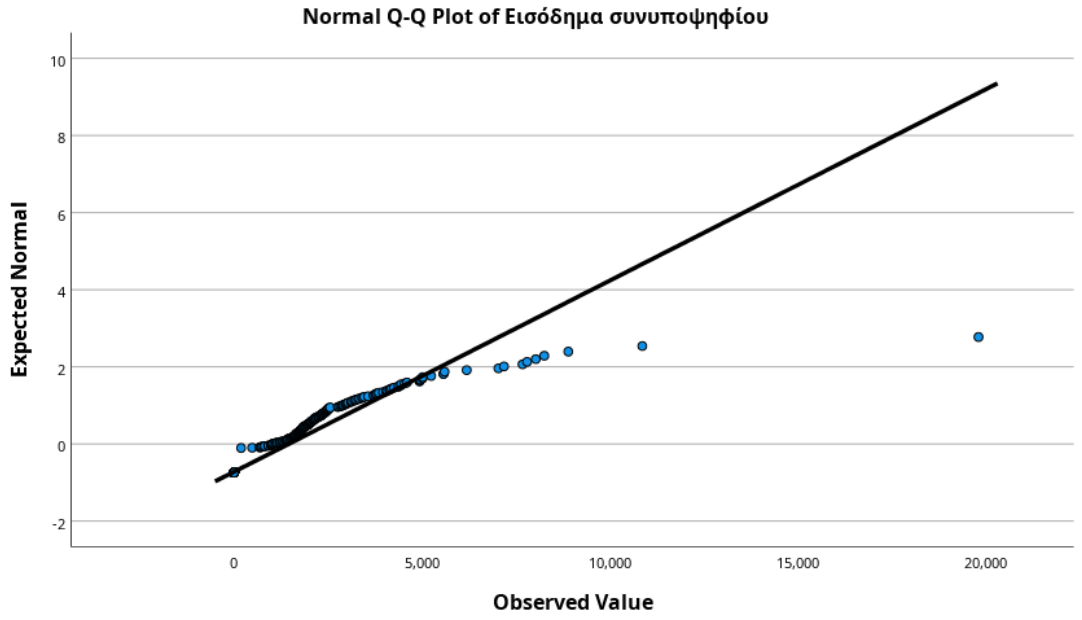
Αρχικά θα πρέπει να έχουμε ανεξαρτησία ανάμεσα στις περιπτώσεις. Δεν έχουμε λόγο να υποψιαζόμαστε ότι κάθε εγγραφή (κάθε πελάτης) επηρεάζει κάποιον άλλο πελάτη άρα μπορούμε να θεωρήσουμε ότι δεν υπάρχει θέμα μη-ανεξαρτησίας ανάμεσα στις εγγραφές.

Τέλος, οι συνεχείς ανεξάρτητες μεταβλητές θα πρέπει να κατανέμονται κανονικά. Ένας οπτικός τρόπος για να πάρουμε μια πρώτη εκτίμηση για το αν οι μεταβλητές αυτές ακολουθούν την κανονική κατανομή είναι τα Q-Q διαγράμματα. Σε ένα Q-Q διάγραμμα συγκρίνουμε τα ποσοστιαία σημεία των δεδομένων μας με την ιδανική κατανομή, δηλαδή την κανονική. Όσο πιο πολύ πέφτουν τα δεδομένα μας πάνω στην ευθεία γραμμή τόσο πιο κοντά πλησιάζουν στην κανονική κατανομή. Στα Σχήμα 29, 30, 31 και 33 βλέπουμε ότι τα δεδομένα δε φαίνεται να πέφτουν επάνω στην ιδανική ευθεία άρα υποψιαζόμαστε ότι οι μεταβλητές “Εισόδημα Υποψηφίου”, “Εισόδημα Συνυποψηφίου”, “Ποσό δανείου” και “Συνολικό εισόδημα” δεν ακολουθούν κανονική

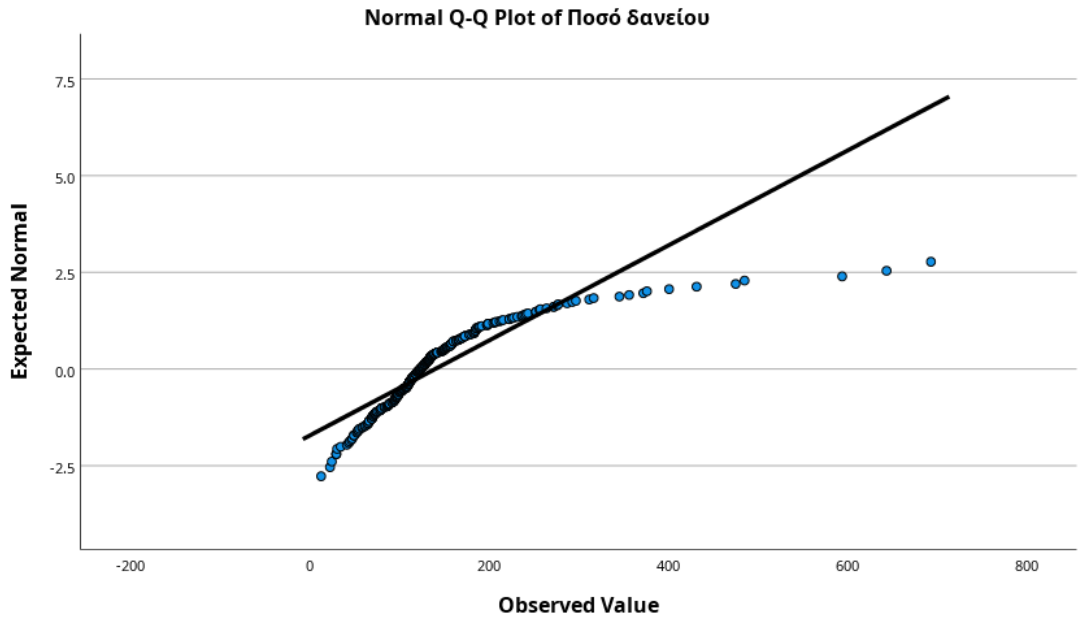
κατανομή. Το Σχήμα 32 που είναι το Q-Q διάγραμμα για τη διάρκεια του δανείου σε μήνες θα μπορούσε να θεωρηθεί αρκετά κανονικό.



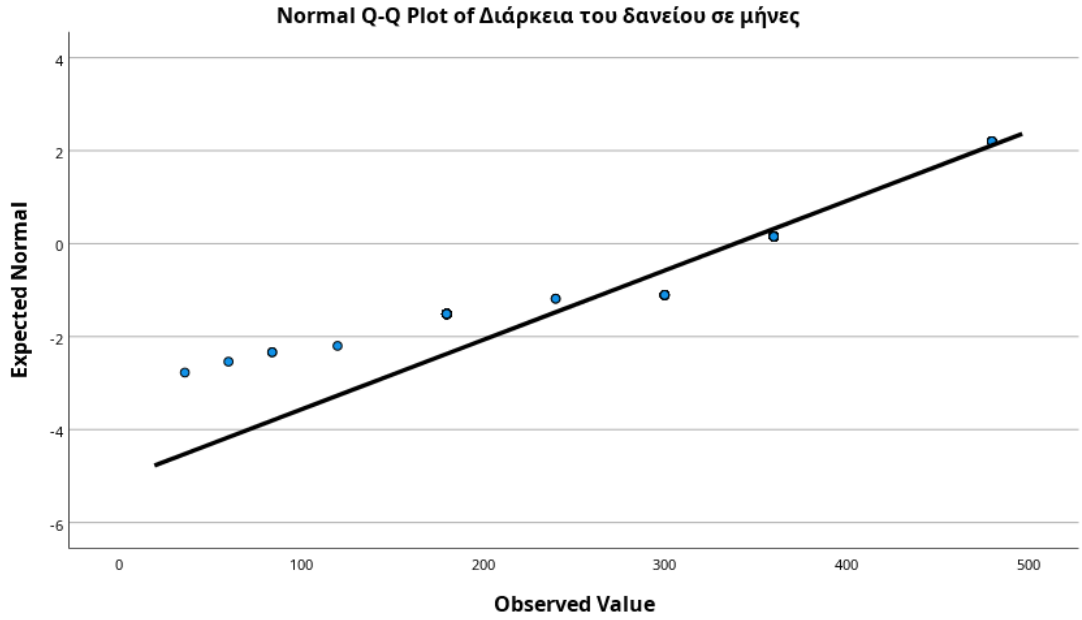
Σχήμα 29: Διάγραμμα Q-Q για το εισόδημα υποψηφίου



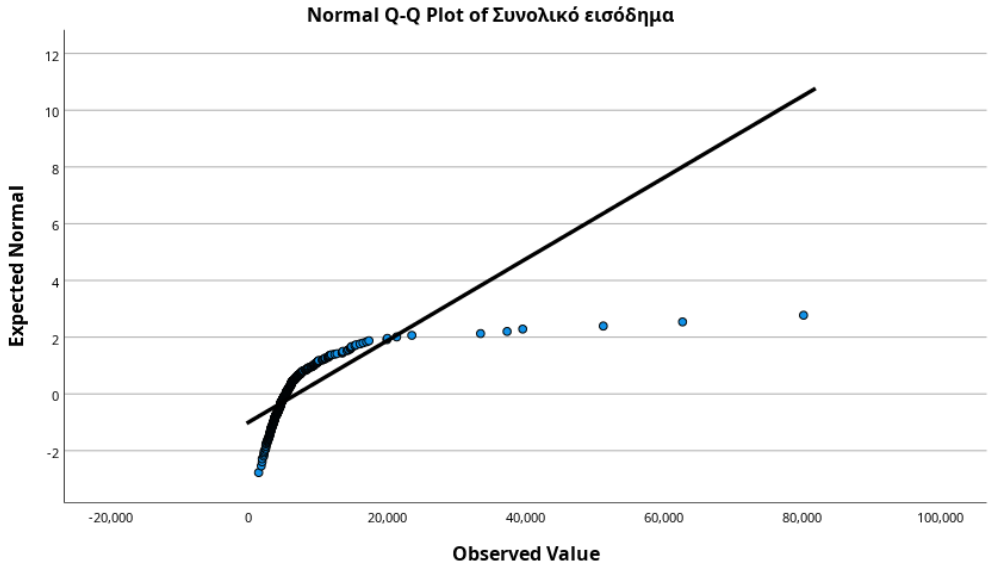
Σχήμα 30: Διάγραμμα Q-Q για το εισόδημα συνυποψηφίου



Σχήμα 31: Διάγραμμα Q-Q για το ποσό δανείου



Σχήμα 32: Διάγραμμα Q-Q για τη διάρκεια του δανείου



Σχήμα 33: Διάγραμμα Q-Q για το συνολικό εισόδημα

Ένας πιο αξιόπιστος τρόπος να καταλάβουμε αν οι μεταβλητές αυτές κατανέμονται κανονικά είναι ο έλεγχος κανονικότητας Kolmogorov–Smirnov και ο Shapiro–Wilk. Ο έλεγχος Shapiro–Wilk χρησιμοποιείται κυρίως για δείγματα <=50, άρα στο δείγμα ανάπτυξης θα χρησιμοποιήσουμε τον έλεγχο Kolmogorov–Smirnov. Ο Πίνακας 32 δείχνει την εκτέλεση του ελέγχου και για τις πέντε συνεχείς μεταβλητές και μας δείχνει ότι καμία από τις μεταβλητές δεν ακολουθεί κανονική κατανομή αφού το p-value για κάθε μία από αυτές είναι < 0.05.

Τόσο λοιπόν ο γραφικός έλεγχος κανονικότητας όσο και το τεστ Kolmogorov–Smirnov μας δείχνουν ότι τα αποτελέσματα που θα βγάλουμε από την διαχωριστική ανάλυση ίσως να μην είναι κατάλληλα. Πάντως σύμφωνα με τους Hastie(2009) και Reichert(1983) ακόμα και αν τα δεδομένα που χρησιμοποιήθηκαν στη διαχωριστική ανάλυση δεν είναι κανονικά κατανεμημένα, η διαχωριστική ανάλυση θα δώσει κάποια σημαντικά αποτελέσματα. Αυτό που επηρεάζει η μη-κανονικότητα κάποιων μεταβλητών είναι η εκτίμηση για το cut-off point, που μπορεί να είναι σημαντικά λανθασμένη.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Εισόδημα υποψηφίου	.270	359	.000	.417	359	.000
Εισόδημα συνυποψηφίου	.235	359	.000	.701	359	.000
Ποσό δανείου	.172	359	.000	.756	359	.000
Διάρκεια του δανείου σε μήνες	.477	359	.000	.521	359	.000
Συνολικό εισόδημα	.242	359	.000	.464	359	.000

a. Lilliefors Significance Correction

Πίνακας 32: Έλεγχοι κανονικότητας για τις συνεχείς μεταβλητές

Μετά από τον έλεγχο των προϋποθέσεων της διαχωριστικής ανάλυσης, θα προχωρήσουμε στην εκτέλεση της στο SPSS, στο δείγμα ανάπτυξης. Θα επιλέξουμε βηματική διαδικασία, στην οποία το SPSS προσπαθεί να εισάγει στο μοντέλο μία-μία τις ανεξάρτητες μεταβλητές όταν αυτές έχουν στατιστικώς σημαντική συνεισφορά σε αυτό. Ο Πίνακας 33 δείχνει τον έλεγχο υποθέσεων για τους μέσους κάθε ομάδας. Η μηδενική υπόθεση εδώ είναι ότι σε κάθε μεταβλητή οι δύο ομάδες (καλοί – κακοί πελάτες) έχουν ίδια μέση τιμή. Παρατηρούμε ότι μόνο οι μεταβλητές "Πιστωτικό Ιστορικό" και "Εκπαίδευση" με p-value <0.05 απορρίπτουν την H_0, άρα μόνο για αυτές τις μεταβλητές μπορούμε να πούμε ότι οι παρατηρούμενες διαφορές στις μέσες τιμές τους ανάμεσα στις δύο ομάδες δεν είναι τυχαίες. Όλες οι υπόλοιπες μεταβλητές με Sig.>0.05 δεν έχουν στατιστικώς σημαντικές διαφορές στις μέσες τιμές τους ανάμεσα στις δύο ομάδας και άρα δεν είναι σε θέση να προβλέψουν για έναν πελάτη την κατάταξη του σε καλό ή κακό, οπότε δε θα εμφανιστούν στο μοντέλο που θα δημιουργήσει η διαχωριστική ανάλυση.

Ο Πίνακας 34 δείχνει τη μεταβλητή που χρησιμοποιήθηκε στο τελικό μοντέλο πρόβλεψης. Η διαχωριστική ανάλυση χρειάστηκε μόλις ένα βήμα για να καταλήξει στο μοντέλο και ενδιαφέρον έχει ότι πέρα από τη μεταβλητή "Πιστωτικό Ιστορικό", που την περιμέναμε στο μοντέλο, η μεταβλητή "Εκπαίδευση" που φάνηκε πριν να έχει στατιστικώς σημαντικές διαφορές στους μέσους όρους ανάμεσα στις ομάδες δεν ήταν αρκετά καλή για το μοντέλο ώστε να μπει σε αυτό.

Ο Πίνακας 35 μας δίνει τα αποτελέσματα του ελέγχου Box ισότητας των πινάκων συνδιακύμανσης. Όπως βλέπουμε, με Sig<0.05, η μηδενική υπόθεση της ισότητας των πινάκων συνδιακύμανσης απορρίπτεται. Ο έλεγχος του Box είναι επιρρεπής σε λάθη όταν τα δεδομένα δεν πληρούν την υπόθεση της κανονικότητας,

όπως δηλαδή συμβαίνει στα δεδομένα του δείγματος, οπότε τέτοιο p-value ήταν αναμενόμενο. Όπως όμως ήδη αναφέραμε, η διαχωριστική ανάλυση μπορεί να εφαρμοστεί και όταν η υπόθεση της κανονικότητας αποτυγχάνει.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Γένος	.996	1.302	1	357	.255
Παντρεμένος/η	.990	3.727	1	357	.054
Εξαρτώμενα μέλη	.997	.946	1	357	.331
Εκπαίδευση	.986	4.927	1	357	.027
Αυτοαπασχολούμενος/η	.999	.297	1	357	.586
Εισόδημα υποψηφίου	1.000	.158	1	357	.691
Εισόδημα συνυποψηφίου	.999	.187	1	357	.666
Ποσό δανείου	.998	.606	1	357	.437
Διάρκεια του δανείου σε μήνες	.999	.262	1	357	.609
Πιστωτικό ιστορικό	.787	96.336	1	357	.000
Περιοχή κατοικίας	.999	.195	1	357	.659
Συνολικό εισόδημα	.999	.274	1	357	.601

Πίνακας 33: Έλεγχος υποθέσεων των μέσων κάθε ομάδας

Variables in the Analysis

Step		Tolerance	F to Remove
1	Πιστωτικό ιστορικό	1.000	96.336

Πίνακας 34: Μεταβλητές που χρησιμοποιήθηκαν στο μοντέλο πρόβλεψης

Test Results

Box's M		79.927
F	Approx.	79.668
	df1	1
	df2	232909.811
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Πίνακας 35: Αποτελέσματα ελέγχου Box για την ισότητα των πινάκων συνδιακύμανσης

Ο Πίνακας 36 μας δίνει τις εκ των προτέρων πιθανότητες των δύο ομάδων. Θα κάνουμε την υπόθεση ότι η αναλογία καλών/κακών πελατών στο δείγμα μας είναι παρόμοια με την αναλογία που βρίσκεται στον πληθυσμό. Έτσι, για κάθε μία από τις δύο ομάδες υπολογίζεται ένα σκορ που στηρίζεται σε κάποιο γραμμικό ως προς τα χαρακτηριστικά μοντέλο. Οι συντελεστές των γραμμικών συναρτήσεων των σκορ

υπολογίζονται με τη μέθοδο του Fisher και τα αποτελέσματα φαίνονται στον Πίνακα 37. Έτσι λοιπόν έχουμε τις δύο εξισώσεις:

$w_0 = -1.948 + 4.623 * ΠιστωτικόΙστορικό$, για τους κακούς πελάτες και

$w_1 = -4.4 + 7.946 * ΠιστωτικόΙστορικό$, για τους καλούς πελάτες.

Έτσι, για κάθε υποψήφιο πελάτη της τράπεζας υπολογίζονται οι ποσότητες w_0 και w_1 και αν ισχύει $w_0 > w_1$ τότε ο υποψήφιος χαρακτηρίζεται ως "κακός" πελάτης, αλλιώς χαρακτηρίζεται ως "καλός" και σε κάθε περίπτωση η τράπεζα δρα ανάλογα.

Prior Probabilities for Groups

Απάντηση	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Κακός πελάτης	.500	105	105.000
Καλός πελάτης	.500	254	254.000
Total	1.000	359	359.000

Πίνακας 36: Εκ των προτέρων πιθανότητες για τις δύο κατηγορίες

Classification Function Coefficients

	Απάντηση	
	Κακός πελάτης	Καλός πελάτης
Πιστωτικό ιστορικό	4.623	7.946
(Constant)	-1.948	-4.400

Fisher's linear discriminant functions

Πίνακας 37: Συντελεστές διαχωριστικής ανάλυσης

Ο Πίνακας 38 μας δίνει τα ποσοστά σωστής ταξινόμησης που δίνει το μοντέλο που δημιουργήθηκε από το δείγμα ανάπτυξης όταν αυτό εφαρμοστεί στα δεδομένα του δείγματος επικύρωσης. Παρατηρούμε ότι το ποσοστό των πελατών του δείγματος επικύρωσης που είναι "καλοί" και χαρακτηρίστηκαν "καλοί" από το μοντέλο είναι 92.3% ενώ το συνολικό ποσοστό ακρίβειας του τελικού μοντέλου είναι 73.8%. Μπορούμε δηλαδή να συμπεράνουμε ότι οι παρατηρούμενες και οι εκτιμώμενες από το μοντέλο τιμές της μεταβλητής "LoanStatus" συμφωνούν στο 73.8% του συνόλου των παρατηρήσεων του δείγματος ελέγχου.

Classification Results[a,b]

				Predicted Group Membership		
			Απάντηση	Κακός πελάτης	Καλός πελάτης	Total
Cases Selected	Original	Count	Κακός πελάτης	48	57	105
			Καλός πελάτης	17	237	254
		%	Κακός πελάτης	45.7	54.3	100.0
			Καλός πελάτης	6.7	93.3	100.0
Cases Not Selected	Original	Count	Κακός πελάτης	20	30	50
			Καλός πελάτης	7	84	91
		%	Κακός πελάτης	40.0	60.0	100.0
			Καλός πελάτης	7.7	92.3	100.0

a. 79.4% of selected original grouped cases correctly classified.

b. 73.8% of unselected original grouped cases correctly classified.

Πίνακας 38: Πίνακας ταξινόμησης της διαχωριστικής ανάλυσης

4.6 Σύγκριση των μεθόδων

Επιχειρήσαμε να δημιουργήσουμε μοντέλο πρόβλεψης για την κατάταξη ενός υποψήφιου δανειολήπτη σε καλό ή κακό, χρησιμοποιώντας τρεις διαφορετικές μεθόδους, τη λογιστική παλινδρόμηση, τα δέντρα ταξινόμησης και τέλος, τη διαχωριστική ανάλυση. Αυτό έχει σαν αποτέλεσμα να δημιουργηθούν τρία διαφορετικά μοντέλα οπότε μένει να βρούμε ποιο από τα τρία είναι το καλύτερο.

Ένας τρόπος για να αξιολογήσουμε την κάθε μέθοδο είναι να συγκρίνουμε τα ποσοστά σωστής ταξινόμησης που δίνει το κάθε μοντέλο της. Έτσι, δημιουργήσαμε τον Πίνακα 39 που συγκεντρώνει τους Πίνακες 26, 31, 38 που είναι οι Πίνακες Ταξινόμησης για τη λογιστική παλινδρόμηση, των δέντρων ταξινόμησης και της διαχωριστικής ανάλυσης αντίστοιχα, στο δείγμα ελέγχου. Βλέπουμε λοιπόν ότι στη συνολική σωστή ταξινόμηση, πρώτη έρχεται η διαχωριστική ανάλυση που προβλέπει σωστά έναν υποψήφιο πελάτη στο 73.76% των περιπτώσεων, ακολουθούν τα δέντρα ταξινόμησης με 73.05% σωστή πρόβλεψη και τέλος, η λογιστική παλινδρόμηση με ποσοστό σωστής πρόβλεψης στο 71.63% των περιπτώσεων. Κοιτώντας το συνολικό ποσοστό σωστής πρόβλεψης, συμπεραίνουμε ότι το μοντέλο που δημιουργεί η διαχωριστική ανάλυση έχει την καλύτερη πιθανότητα σωστής πρόβλεψης. Ενδιαφέρον έχει ότι αν κοιτάξουμε μόνο την πιθανότητα ενός καλού πελάτη να ταξινομηθεί σωστά ως καλός, η λογιστική παλινδρόμηση είναι αυτή που δίνει τα καλύτερα αποτελέσματα, στο 93.41% των περιπτώσεων. Ο λόγος που η λογιστική παλινδρόμηση δίνει συνολικά τα χειρότερα αποτελέσματα από τις τρεις μεθόδους παρά το ότι έχει το μεγαλύτερο ποσοστό σωστής

ταξινόμησης των καλών πελατών, είναι το μεγάλο ποσοστό των λανθασμένα ταξινομήσεων στους κακούς πελάτες: η πιθανότητα ένας κακός πελάτης να ταξινομηθεί λανθασμένα ως καλός στη λογιστική παλινδρόμηση είναι 68%.

Οι μεταβλητές που παρέμειναν συνολικά στα τρία μοντέλα είναι οι:

- Πιστωτικό ιστορικό (παρούσα και στα τρία μοντέλα)
- Εισόδημα υποψηφίου (μόνο στα δέντρα ταξινόμησης)
- Συνολικό εισόδημα (μόνο στα δέντρα ταξινόμησης)
- Περιοχή κατοικίας (μόνο στη λογιστική παλινδρόμηση)
- Εκπαίδευση (παρούσα στη λογιστική παλινδρόμηση και στη διαχωριστική ανάλυση)

Παρατηρούμε ότι τα τρία μοντέλα χρησιμοποιούν το πιστωτικό ιστορικό, κάτι που δείχνει τη σημαντικότητα αυτής της μεταβλητής για τη σωστή πρόβλεψη. Ενδιαφέρον αποτελεί το γεγονός ότι το εισόδημα (του υποψηφίου και το συνολικό) εμφανίζονται ως μεταβλητές πρόβλεψης μόνο στα δέντρα ταξινόμησης. Η διαχωριστική ανάλυση χρησιμοποιεί και το εκπαιδευτικό επίπεδο για τις προβλέψεις ενώ η λογιστική παλινδρόμηση χρησιμοποιεί και την περιοχή κατοικίας.

Μέθοδος	**Observerd**	**Predicted**		
		Κακός	**Καλός**	**Percent Correct**
Δέντρα Ταξινόμησης				
Μεταβλητές στο μοντέλο: Πιστωτικό ιστορικό, εισόδημα υποψηφίου, συνολικό εισόδημα	Κακός	22	28	44.00%
	Καλός	10	81	89.01%
	Overall Percentage	22.70%	77.30%	73.05%
Λογιστική Παλινδρόμηση				
Μεταβλητές στο μοντέλο: Πιστωτικό ιστορικό, περιοχή κατοικίας, εκπαίδευση	Κακός	16	34	32.00%
	Καλός	6	85	93.41%
	Overall Percentage	15.60%	84.40%	71.63%
Διαχωριστική ανάλυση				
Μεταβλητές στο μοντέλο: Πιστωτικό ιστορικό, εκπαίδευση	Κακός	20	30	40.00%
	Καλός	7	84	92.31%
	Overall Percentage	19.15%	80.85%	73.76%

Πίνακας 39: Σύγκριση των τριών μεθόδων

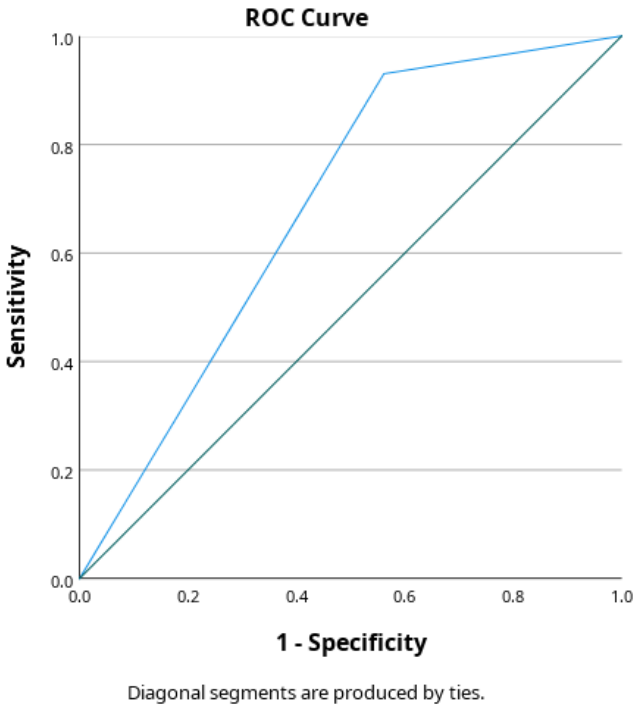
Ένας δεύτερος τρόπος για να συγκρίνουμε τις τρεις μεθόδους είναι να χρησιμοποιήσουμε την μέθοδο της καμπύλης ROC (Receiver Operating Characteristic), η οποία μπορεί να οργανώσει, να επιλέξει και να απεικονίσει ταξινομητές με βάση τη γραφική τους παράσταση. Με τη βοήθεια της καμπύλης ROC μπορούμε να εκτιμήσουμε τη διακριτική ικανότητα μίας δοκιμασίας ή να συγκρίνουμε τη διακριτική ικανότητα δύο ή περισσότερων δοκιμασιών.

Αυτό που μας ενδιαφέρει για την αξιολόγηση των μοντέλων μας, είναι το εμβαδό που βρίσκεται κάτω από την καμπύλη. Οι τιμές που παίρνει το εμβαδό είναι από 0.5 μέχρι 1. Αν το εμβαδό κάτω από την καμπύλη είναι 0.5 (δηλαδή η ελάχιστη τιμή που μπορεί να πάρει) τότε το μοντέλο που περιγράφει η καμπύλη δεν έχει καμία διακριτική ικανότητα και η καμπύλη ROC είναι μια διαγώνιος ευθεία γραμμή. Όσο μεγαλύτερη τιμή έχει το εμβαδό κάτω από την καμπύλη ROC τόσο μεγαλύτερη διακριτική ικανότητα έχει το μοντέλο, με μέγιστη τιμή το 1, δηλαδή μια ορθή γωνία.

Για να δημιουργήσουμε (ώστε μετά να συγκρίνουμε) τις καμπύλες ROC που αντιστοιχούν στα 3 μοντέλα πρέπει για κάθε μοντέλο να έχουμε την πιθανότητα κάθε υποψήφιου πελάτη να καταταχθεί ως "καλός", όπως αυτή εκτιμάται μέσω του μοντέλου. Για να γίνει αυτό θα ζητήσουμε από το SPSS σε κάθε μία από τις τρεις μεθόδους που εκτελέσαμε να αποθηκεύσει αυτές τις πιθανότητες, δημιουργώντας νέες μεταβλητές στο αρχείο μας (μέσω των επιλογών "Save" που έχει κάθε τέτοια μέθοδος). Μετά, παράγουμε την καμπύλη ROC για κάθε μέθοδο, συγκρίνοντας την πιθανότητα που δίνει το κάθε μοντέλο, να καταταγεί ένας υποψήφιος πελάτης ως "καλός" με την πραγματική κατάταξη της τράπεζας.

Έτσι, το Σχήμα 34 δίνει την καμπύλη ROC που αντιστοιχεί στο μοντέλο της διαχωριστικής ανάλυσης (με μπλε χρώμα). Ο Πίνακα 41 δίνει το εμβαδό που βρίσκεται κάτω από την καμπύλη και παρατηρούμε ότι είναι 0.685. Το Σχήμα 35 δίνει την καμπύλη ROC που αντιστοιχεί στο μοντέλο της λογιστικής παλινδρόμησης. Ο Πίνακας 42 δίνει το εμβαδό που βρίσκεται κάτω από την καμπύλη και παρατηρούμε ότι είναι 0.781. Τέλος, το Σχήμα 36 δίνει την καμπύλη ROC που αντιστοιχεί στο μοντέλο των δέντρων ταξινόμησης. Ο Πίνακας 43 δίνει το εμβαδό που βρίσκεται κάτω από την καμπύλη και παρατηρούμε ότι είναι 0.702.

Συγκρίνοντας τις τρεις μεθόδους χρησιμοποιώντας το εμβαδό της καμπύλης ROC, βλέπουμε ότι το μοντέλο της λογιστικής παλινδρόμησης είναι το καλύτερο. Άρα όταν συγκρίνουμε τις μεθόδους κοιτώντας τα ποσοστά σωστής ταξινόμησης τότε παίρνουμε ως καλύτερο μοντέλο αυτό που δημιουργεί η διαχωριστική ανάλυση, ενώ συγκρίνοντας τις μεθόδους κοιτώντας το εμβαδό των καμπύλων ROC παίρνουμε ως καλύτερο μοντέλο αυτό της λογιστικής παλινδρόμησης. Μπορούμε να συμπεράνουμε πως συγκρίνοντας τα τρία μοντέλα με την μέθοδο της καμπύλης ROC το πιο επιτυχημένο είναι αυτό που δημιούργησε η λογιστική παλινδρόμηση.



Σχήμα 34: Καμπύλη ROC για το μοντέλο της Διαχωριστικής Ανάλυσης

Area Under the Curve

Test Result Variable(s): Probabilities of Membership in Group 1 for Analysis 1

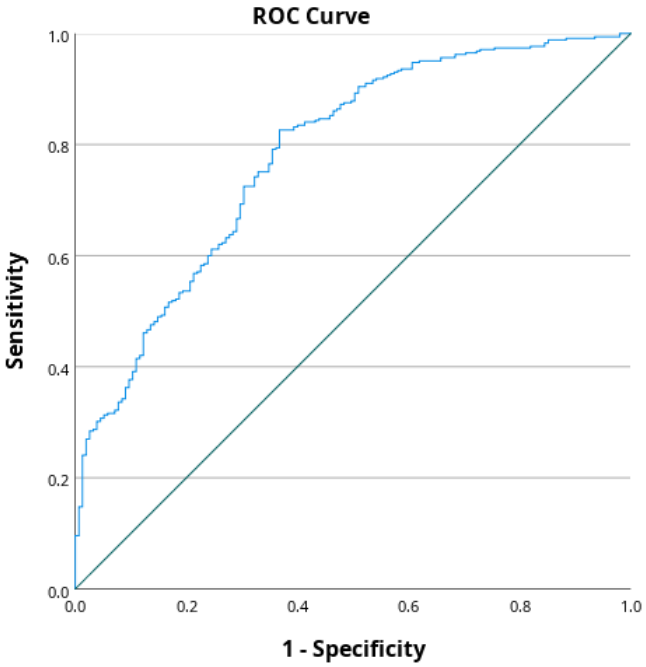
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.685	.028	.000	.630	.739

The test result variable(s): Probabilities of Membership in Group 1 for Analysis 1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Πίνακας 40: Εμβαδό κάτω από την καμπύλη ROC για το μοντέλο της Διαχωριστικής Ανάλυσης



Σχήμα 38: Καμπύλη ROC για το μοντέλο της Λογιστικής Παλινδρόμησης

Area Under the Curve

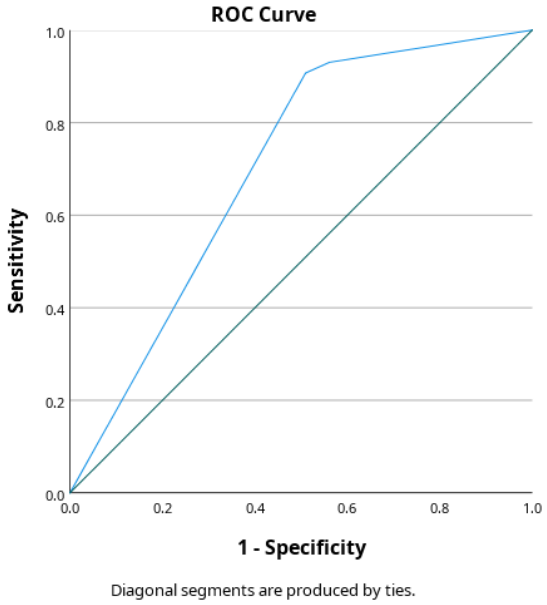
Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.781	.022	.000	.738	.825

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Πίνακας 41: Εμβαδό κάτω από την καμπύλη ROC για το μοντέλο της Λογιστικής Παλινδρόμησης



Σχήμα 39: Καμπύλη ROC για το μοντέλο του Δέντρου Ταξινόμησης

Area Under the Curve

Test Result Variable(s): Predicted Probability for Loan_Status=1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.702	.028	.000	.648	.756

The test result variable(s): Predicted Probability for Loan_Status=1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Πίνακας 42: Εμβαδό κάτω από την καμπύλη ROC για το μοντέλο του Δέντρου Ταξινόμησης

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ:

Αντωνόπουλος Ιωάννης. (2015). *Μοντέλα Αξιολόγησης Πιστοληπτικής Ικανότητας.* Διπλωματική Εργασία του Προγράμματος Μεταπτυχιακών Σπουδών στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου: Πανεπιστήμιο Πειραιώς.

Γάκη Ελένη. Σημείωσεις στην Εφαρμοσμένη Στατιστική. *Ενότητα 8 : Logistic Regression*. Πανεπιστήμιο Αιγαίου.

Ζοπουνίδης Κωνσταντίνος και Λεμονάκης Χρήστος . (2009). *Διαχείριση Πιστωτικού Κινδύνου.* Αθήνα: Εκδόσεις Κλειδάριθμος.

Κούτρας Βασίλειος. (2020). *Σημειώσεις για το μάθημα του Πιστωτικού Κινδύνου του τμήματος Αναλογιστικής Επιστήμης και Διοικητικής Κινδύνου.* Πανεπιστήμιο Πειραιά.

Κούτρας Μαρκος. (2020). *Σημειώσεις για το μάθημα του Πιστωτικού Κινδύνου του τμήματος Αναλογιστικής Επιστήμης και Διοικητικής Κινδύνου.* Πανεπιστήμιο Πειραιά.

Ξενή Μαρία. (2012). Λογιστική Παλινδρόμηση & Διαχωριστική Ανάλυση. Διπλωματική εργασία, Μεταπτυχιακό Πρόγραμμα σπουδών «Μαθηματικά των υπολογιστών και, των αποφάσεων» του τμήματος Μαθηματικών, Πανεπιστήμιο Πατρών.

ΞΕΝΟΓΛΩΣΗ:

Altman, E. I. (May 2002). *Revisiting Credit Scoring Models in a Basel 2 Environment.* London.

Chong, Mimi Mei Ling. (2016). Applications of Credit Scoring Models.Ph.D.thesis,The University of Western Ontario.

Comarch. (2019). *What is credit scoring ? About types, model and method.* Ανάκτηση από https://www.comarch.com/

Do, Hoai Linh, Luong, Thi Thu Hang, Nguyen, Xuan Thang, and Mai, Ngoc Linh. (2019). *Credit Scoring Application at Banks: Mapping to Basel II Journal of Social and Political Sciences, Vol.2, No.1, 83-89.* The Asian Institute of Research.

Gabriele Sabato. *Credit Risk Scoring Models.* Royal Bank of Scotland.

Hastie Trevor, T. R. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics.

Huyen, N. (2011). Corporate Credit Risk Management by Customer Risk Level. *International , Banking Magazine,*Vol.2, No.1, 83-89.

Majer Izabela. (n.d.). *Application scoring: logit model approach and the divergence method compared.* Warsaw School of Economics: Department of Applied Econometrics.

Naeem Siddiqi. *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring.* John Wiley & Sons, Inc.

Reichert A. K., Cho C.C. and Wagner G.M. (1983). *An examination of conceptual issues involved in developing credit scoring models.* Journal of Business and Economic Statistics,1,101-114

Supervision, Basel Committee on Banking. (n.d.). *The Basel Framework.* Bank for International Settlements.

Yeh Andrew, T. J. Basel II: A new capital framework. Financial Stability Department, Bulletin, Reserve Bank of New Zealand.

Alinejad, E. S. (2013). *Credit scoring models of customers in banks.* Singaporean Jouranl of Business Econcomics and Management Studies: Vol. 1,37-41