ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΦΑΡΜΟΓΕΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΕ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ

Κυριακή Β. Λυπάκη

Πειραιάς

Αύγουστος 2010

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΦΑΡΜΟΓΕΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΕ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ

Κυριακή Β. Λυπάκη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Αύγουστος 2010

Η παρούσα διπλωματική εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από την ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμό ..... συνεδρίαση σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Κατέρη Μαρία (Επιβλέπουσα)
- Κούτρας Μάρκος
- Πολίτης Κωνσταντίνος

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

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

APPLICATIONS OF CLASSIFICATION IN BANKING DATA

By

Kyriaki V. Lipaki

MSc Dissertation

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

Piraeus

August 2010

Στους γονείς μου
Βασίλη και Μαρία,
και στην αδελφούλα μου
Δέσποινα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την αναπληρώτρια καθηγήτρια και επιβλέπουσα, κα. Μαρία Κατέρη, για την πολύτιμη βοήθεια της στην υλοποίηση της παρούσας εργασίας, καθώς και για την συμπαράσταση και την υπομονή που έδειξε όλο αυτό το διάστημα. Επίσης θέλω να ευχαριστήσω τους καθηγητές κ. Μάρκο Κούτρα και κ. Κων/νο Πολίτη για την συμμετοχή τους στην τριμελή επιτροπή και την υποστήριξή τους.

Θέλω επίσης να εκφράσω τις ευχαριστίες μου στους συναδέλφους μου, και ιδιαίτερα στους προϊσταμένους μου, για την συμπαράσταση, την κατανόηση και την υπομονή που έδειξαν κατά την διάρκεια της συγγραφής της εργασίας αυτής.

Επίσης θέλω να ευχαριστήσω ένα πολύ καλό μου φίλο, τον Θοδωρή Σφακιανάκη, για την ηθική υποστήριξη και την συμπαράστασή του ώστε να ολοκληρώσω την διπλωματική μου εργασία.

Τέλος, θέλω να πω ένα μεγάλο ευχαριστώ στην οικογένεια μου, για την υποστήριξη, την συμπαράσταση, και όλα όσα μου προσέφερε καθ' όλη τη διάρκεια των σπουδών μου.

Κυριακή Λυπάκη

Πειραιάς, Αύγουστος 2010

Περίληψη

Η πρόσφατη οικονομική κρίση αναδεικνύει για άλλη μια φορά τον σημαντικό ρόλο της παρακολούθησης του ρίσκου, που έχει αναλάβει ένας τραπεζικός οργανισμός στο δανειακό του χαρτοφυλάκιο. Σημαντικό εργαλείο γι' αυτό αποτελεί ένα μοντέλο ικανό να εντοπίσει έγκαιρα πελάτες των οποίων τα δάνεια θα καταστούν μη-εξυπηρετούμενα στο άμεσο μέλλον.

Στόχος μας είναι να περιγράψουμε την διαδικασία κατασκευής ενός τέτοιου μοντέλου για την Καταναλωτική Πίστη, μέσω της εφαρμογής της σε δεδομένα μιας από τις μεγαλύτερες τράπεζες της Ελλάδας.

Αναλυτικότερα, κατασκευάσαμε ένα στατιστικό μοντέλο ταξινόμησης το οποίο μπορεί να προβλέψει, για ένα πελάτη που σήμερα οφείλει μια δόση σε ένα καταναλωτικό του δάνειο, αν στο τέλος του επόμενου εξαμήνου κάποιο δάνειο του πελάτη θα έχει χαρακτηριστεί ως μη-εξυπηρετούμενο (δηλαδή βρίσκεται σε καθυστέρηση μεγαλύτερη των 90 ημερών) ή όχι.

Για το σκοπό αυτό εφαρμόσαμε την μέθοδο της Λογιστικής Παλινδρόμησης (*Logistic Regression*) στα δεδομένα μας, τα οποία είχαμε προηγουμένως επεξεργαστεί ώστε να μειώσουμε την διάσταση του χώρου των μεταβλητών. Συγκεκριμένα, για τον περιορισμό των κατηγορικών μεταβλητών καταφύγαμε στη Ανάλυση Συστάδων Δύο Βημάτων (*Two-Step Cluster Analysis*), ενώ για την μείωση των συνεχών στην Ανάλυση Παραγόντων (*Factor Analysis*). Επίσης χρησιμοποιήσαμε την Ανάλυση Συστάδων Δύο Βημάτων για να σκιαγραφήσουμε το προφίλ των πελατών μας.

Abstract

The recent Economic crisis highlighted once again the significant role of Risk monitoring of loan portfolio for Banking institutions. A primary tool for this surely is a model able to identify future potentially non-payers that could become defaulted customers.

The goal for this dissertation is to describe how to build such a model for Consumer Lending, by actually constructing one, using data of a major bank of Greece.

Specifically, we constructed a statistical classification model, able to predict whether at the end of the next six months, a client that currently is at 30 days past overdue, will be over 90 days overdue, or in other words whether his loan will be classified as non-performing.

To serve this purpose we had, as a first step, to reduce the dimension of independent variables space. For categorical variables we applied the Two-Step Cluster Analysis, while for continuous variables Factor Analysis was chosen. In addition, Two-Step Cluster Analysis helped us in the clients' profiling process. The main part of the analysis utilized the Logistic Regression Method, which highlighted exposure, type of loan, client's marital status, and status of client's business loan (if any) as the key variables for increasing customer probability to become defaulted.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τι είναι η ταξινόμηση

«Ταξινόμηση είναι η διαδικασία ανάθεσης αντικειμένων σε μια εκ πολλών προκαθορισμένων κατηγοριών» (Tan, 2006).

Σε κάθε πρόβλημα ταξινόμησης τα αρχικά δεδομένα αποτελούνται από ένα σύνολο παρατηρήσεων, κάθεμια από τις οποίες χαρακτηρίζεται από ένα ζεύγος ($\vec{x}$, y). Το διάνυσμα $\vec{x}$ περιέχει τις τιμές των ανεξάρτητων μεταβλητών (*independent variables*) για κάθε παρατήρηση, ενώ η y είναι η εξαρτημένη μεταβλητή (*dependent variable*) που δηλώνει την κατηγορία στην οποία ανήκει η παρατήρηση.

Οι ανεξάρτητες μεταβλητές μπορεί να είναι συνεχείς, κατηγορικές (ονοματικές ή διατάξιμες) ή συνδιασμός τους. Όμως, η εξαρτημένη μεταβλητή πρέπει να είναι κατηγορική. Αυτή είναι και η ειδοποιός διαφορά της ταξινόμησης (*classification*) από την παλινδρόμηση (*regression*), όπου η εξαρτημένη μεταβλητή είναι συνεχής. Όταν η εξαρτημένη μεταβλητή παίρνει δύο μόνο τιμές (συνήθως {0,1}) τότε η ταξινόμηση λέγεται δυαδική (*binary classification*).

Έτσι ο μαθηματικός ορισμός του προβλήματος της ταξινόμησης διαμορφώνεται ακολούθως:

«Ταξινόμηση είναι η διαδικασία εκπαίδευσης μιας «συνάρτησης στόχου» (*target function*) f η οποία απεικονίζει κάθε σύνολο ανεξάρτητων μεταβλητών $\vec{x}$ σε μια από τις προκαθορισμένες ομάδες, την y.» (Tan, 2006)

Η συνάρτηση f ονομάζεται μοντέλο ταξινόμησης (*classification model*) ή ταξινομητής (*classifier*).

Τα μοντέλα ταξινόμησης εξυπηρετούν δύο σκοπούς, περιγραφή και πρόβλεψη. Με βάση ένα μοντέλο μπορεί να περιγραφούν τα χαρακτηριστικά εκείνα που παίζουν σημαντικό ρόλο στην διαφοροποίηση των κατηγοριών μεταξύ τους. Όμως, η βασικότερη χρήση τους είναι η πρόβλεψη της κλάσης στην οποία κατατάσσεται μια παρατήρηση της οποίας γνωρίζουμε τις τιμές των ανεξάρτητων μεταβλητών αλλά μας είναι άγνωστη η ομάδα στην οποία ανήκει.

Το μοντέλο ταξινόμησης θα πρέπει αφενός να προσαρμόζεται καλά στα αρχικά δεδομένα και αφετέρου να προβλέπει σωστά την κατηγορία που ταξινομείται μια παρατήρηση, την οποία δεν έχει προηγουμένως «διαβάσει». (Tan, 2006)

Για να μπορέσουμε να ελέγξουμε τα παραπάνω, πρέπει να χωρίσουμε το σύνολο των διαθέσιμων παρατηρήσεων σε δύο υποσύνολα: το εκπαιδευτικό δείγμα (*training set*) και το δείγμα ελέγχου (*test set*).

Το πρώτο χρησιμοποιείται από τον αλγόριθμο ταξινόμησης για την «εκμάθηση» του ταξινομητή, δηλαδή με βάση τα δεδομένα αυτά προκύπτει η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής.

Στη συνέχεια, το μοντέλο που προκύπτει εφαρμόζεται στο δείγμα ελέγχου, προκειμένου να εξεταστεί η ορθότητα με την οποία ταξινομεί παρατηρήσεις «άγνωστες» σε αυτό. Ο έλεγχος πραγματοποιείται με την βοήθεια του πίνακα συνάφειας (*crosstabs*) μεταξύ των παρατηρούμενων (*observed*) και των προβλεπόμενων (*predicted*) -βάσει μοντέλου- κλάσεων (Πίνακας 1.1). Όσο πιο ενισχυμένη είναι η διαγώνιος του πίνακα τόσο ισχυρότερη είναι και η ικανότητα πρόβλεψης του μοντέλου.

		Observed	
		0	**1**
Predicted	**0**	**a**	b
	1	c	**d**

Πίνακας 1.1: πίνακας συνάφειας για παρατηρούμενες και προβλεπόμενες τιμές

1.2 Εφαρμογές ταξινόμησης

Σύμφωνα και με τον ορισμό που δόθηκε παραπάνω, πληθώρα από προβλήματα ταξινόμησης προκύπτουν καθημερινά σε διάφορους τομείς. Παρακάτω αναφέρουμε ορισμένες από τις πιο σημαντικές εφορμογές:

Ιατρική: Για την πραγματοποίηση ιατρικών διαγνώσεων μέσω της ταξινόμησης ασθενών σε γκρουπ ασθενειών, βάση των συμπτωμάτων που παρουσιάζουν. Ενδεικτικά αναφέρουμε Stefanowski and Slowinski (1998), Tsumoto (1998), Michalowski et al. (2001), Dreiseitl et al. (2002).

Βιολογία – Γενετική: Για την ταξινόμηση των γονιδίων με βάση τα χαρακτηριστικά που εκφράζουν και τον διαχωρισμό των μίκρο-RNAs σε εκφραστικά και μη. (π.χ., Keller et al., 2000; Tran et al., 2008; Sinha et al., 2009; Xu et al., 2009)

Οικονομία: Πρόβλεψη αποτυχίας επιχειρήσεων, αξιολόγηση πιστωτικού κινδύνου για επιχειρήσεις και καταναλωτές, αξιολόγηση κινδύνου χαρτοφυλάκιου (π.χ., Martin, 1977; Altman et al., 1981; Sung et al., 1999)

Marketing-CRM: Μέτρηση ικανοποίησης πελατών, ανάλυση των χαρακτηριστικών διαφορετικών κατηγοριών πελατών, ανάπτυξη στρατηγικών διείσδυσης στις αγορές (*market penetration*) (π.χ., Siskos et al., 1998; Nalbantov et al., 2005; Ngai et al., 2009).

Διαχείριση περιβάλλοντος και ενέργειας – Οικολογία: Ανάλυση και μέτρηση των περιβαλλοντικών επιπτώσεων διαφορετικών ενεργειακών πολιτικών, έρευνα -αξιολόγηση της αποτελεσματικότητας ενεργειακών πολιτικών (π.χ., Dunn et al., 1984; Diakoulaki et al., 1999; Rossi et al. 1999; Flinkman et al., 2000).

Διαχείριση ανθρώπινου δυναμικού: Ανάθεση του ανθρώπινου δυναμικού σε κατάλληλες ομάδες επαγγελμάτων σύμφωνα με τα προσόντα τους (π.χ., Gochet et al., 1997; Jantan et al., 2009).

Διαχείριση συστημάτων παραγωγής και Τεχνικής διάγνωσης: Παρακολούθηση διαδικασίας σύνθετων συστημάτων παραγωγής για σκοπούς διάγνωσης βλαβών (π.χ., Nowicki et al., 1992; Catelani and Fort, 2000; Shen et al., 2000).

Οι πιο σύγχρονες εφαρμογές της ταξινόμησης αφορούν στην κατηγοριοποίηση ψηφιοποιημένου κειμένου και φωτογραφιών (π.χ., Yang, 1999; Antonie et al., 2001; Zhang and Oles, 2001; Baoli et al., 2003; Han and Zhao, 2008; Tang et al., 2008) καθώς και στον διαχωρισμό ηλεκτρονικων σελίδων: (π.χ., Dumais and Chen, 2000; Sun et al., 2002; Kwon and Lee, 2003; Schenker et al., 2004).

1.3 Κατηγοριοποίηση των μεθόδων ταξινόμησης

Η ταξινόμηση ξεκίνησε το 1936, όταν ο Fisher συνέλεξε ένα δείγμα από κυκλάμινα (*iris*) τα οποία με την βοήθεια ενός βοτανολόγου διαχώρισε σε τρία είδη (*species*). Στη συνέχεια πραγματοποίησε μετρήσεις για το μήκος και το πλάτος των πετάλων και των σεπάλων κάθε λουλουδιού. Χρησιμοποιώντας τις παραπάνω μετρήσεις, παρήγαγε μαθηματικούς τύπους οι οποίοι όριζαν επίπεδα (*hyperplanes*) τα οποία διαχώριζαν με τον καλύτερο δυνατό τρόπο τις τρεις κατηγορίες των κυκλαμίνων.

Από τότε έχει αναπτυχθεί μεγάλος αριθμός μεθόδων που χρησιμοποιούνται για την επίλυση των προβλημάτων ταξινόμησης.

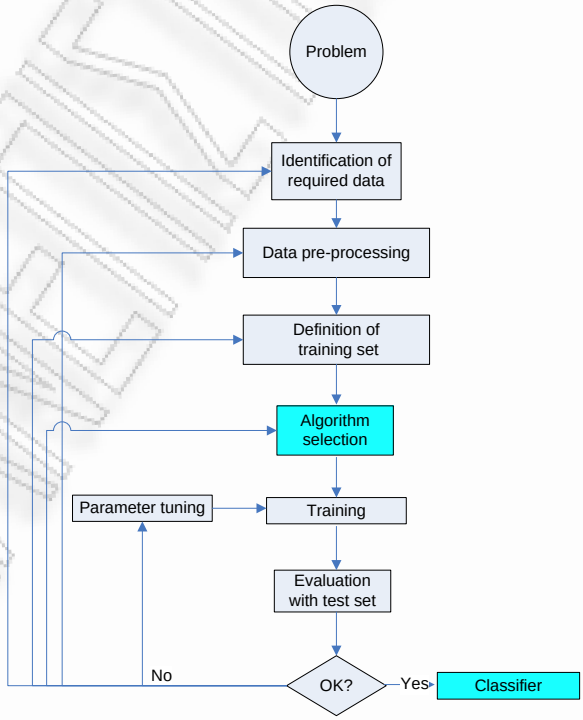
Σύμφωνα με τον Kotsianti (2007), μπορούν να διαχωριστούν σε 5 βασικές κατηγορίες:

- Λογικοί - Συμβολικοί αλγόριθμοι (*logic-based algorithms*), με κυριότερους τα δέντρα απόφασης (*decision trees*) και τους αλγορίθμους που βασιζονται σε κανόνες απόφασης (*decision rules-based algorithms*),
- Νευρωνικά Δίκτυα (*Neural Networks*), τεχνικές βασισμένες στην έννοια του perceptron,
- Στατιστικές μέθοδοι (*Statistical algorithms*), όπως η διαχωριστική ανάλυση (*discriminant analysis*) και η λογιστική παλινδρόμηση (*logistic regression*).
- Αλγόριθμοι με βάση τα χαρακτηριστικά (*Instance-based*), με γνωστότερο αυτόν του κοντινότερου γείτονα (*nearest neighbor algorithm*) και
- Μηχανές υποστηρικτικών διανυσμάτων (*Support Vector Machines-SVMs*)

Μια δεύτερη κατηγοριοποίηση των μεθόδων ταξινόμησης προτείνεται από τους Zopounidi and Doumpo (2004) οι οποίοι τις χωρίζουν σε:

- Στατιστικές και οικονομετρικές μεθόδους (π.χ. διαχωριστική ανάλυση (*discriminant analysis*), λογιστική ανάλυση (*logit analysis*)) και
- Μη-παραμετρικές τεχνικές (π.χ. νευρωνικά δίκτυα, δέντρα απόφασης)

Η διαδικασία που ακολουθείτε σε ένα τυπικό πρόβλημα ταξινόμησης απεικονίζεται στο Σχήμα 1.1.



Σχήμα 1.1: Διαδικασία επίλυσης των προβλημάτων ταξινόμησης (πηγή: Kotsiantis, 2007)

1.4 Περιεχόμενο και δομή της εργασίας

Αντικείμενο της παρούσας εργασίας είναι η αξιολόγηση του ρίσκου των πελατών μιας τράπεζας, ως προς την αποπληρωμή των καταναλωτικών τους δανείων. Πρόκειται για εφαρμογή της δυαδικής ταξινόμησης σε συγκεκριμένο χαρτοφυλάκιο της τράπεζας Χ.

Στο κεφάλαιο (2) περιγράφεται το πρόβλημα (§ 2.2) και ορίζονται οι μεταβλητές που το απαρτίζουν (§ 2.3). Στην παράγραφο 2.4 θα γίνει ο καθαρισμός των δεδομένων (*data cleaning*), η συμπλήρωση των ελλειπουσών τιμών (*missing values*) καθώς και η επιλογή του εκπαιδευτικού δείγματος (*training set*).

Μια ιδιαιτερότητα των δεδομένων μας είναι η ανισορροπία (*imbalanced data*), αφού το πλήθος των αξιόπιστων πελατών είναι πολύ μεγαλύτερο από των αναξιόπιστων. Για το λόγο αυτό θα κάνουμε μια σύντομη περιγραφή για τους τρόπους αντιμετώπισης των μη ισορροπημένων δεδομένων (§ 3.4)

Επίσης στο κεφάλαιο 3 γίνεται μια συνοπτική παρουσίαση της ανάλυσης συστάδων (*Cluster Analysis*) και της ανάλυσης παραγόντων (*Factor Analysis*), οι οποίες στη συνέχεια θα χρησιμοποιηθούν για την σκιαγράφηση του προφίλ των πελατών (*profiling*) (§ 4.1), τον εντοπισμό των κύριων κατηγορικών μεταβλητών (§ 4.2) και την μείωση της διάστασης των συνεχών μεταβλητών (§ 4.3).

Στο κεφάλαιο 5 ακολουθεί παρουσίαση της μεθόδου της λογιστικής παλινδρόμησης (§ 5.2 – § 5.3) και εφαρμογή της στα δεδομένα μας (§ 5.4). Έπεται η αξιολόγηση της προσαρμογής του μοντέλου και η ανάλυση των αποτελεσμάτων (§ 5.5)..

Η εφαρμογή όλων των παραπάνω μεθόδων θα γίνει με την χρήση του στατιστικού πακέτου SPSS.

ΚΕΦΑΛΑΙΟ 2

Προεπεξεργασία Δεδομένων

2.1 Εφαρμογές ταξινόμησης στην Τραπεζική

Οι τεχνικές της εξώρυξης δεδομένων (*Data Mining*) και ειδικότερα της ταξινόμησης (*Classification*) χρησιμοποιούνται ευρέως στον κλάδο της Τραπεζικής (*Banking*).

Κάποιες από τις εφαρμογές που αναφέρονται στην βιβλιογραφία είναι η πρόβλεψη χρεοκοπίας των τραπεζών (Sung et al., 1999) και ειδικότερα με την χρήση λογιστικής παλινδρόμησης (Martin, 1977), η πρόβλεψη αθέτησης πληρωμών δανείων (*loan-default prediction*) με νευρωνικά δίκτυα (Piramuthu et al., 1998; Yegorova et al., 2001; Hays et al., 2002) και η ανάλυση των οικονομικών χαρακτηριστικών των τραπεζών με την βοήθεια της λογιστικής παλινδρόμησης (Kosmidou et al., 2006).

Ειδικότερα, ένα μεγάλο μέρος των εφαρμογών έχει ως αντικείμενο την αξιολόγηση πιστωτικού κινδύνου με διάφορες μεθόδους όπως η λογιστική παλινδρόμηση (*Logistic Regression*) (Laitinen, 1999; Westgaard and Wijst, 2001), η διαχωριστική ανάλυση (*Discriminant Analysis*) (Desai et al., 1996; Bardos, 1998), οι μηχανές υποστηρικτικών διανυσμάτων (*Support Vector Machines (SVM)*) (Huang et al., 2007), τα νευρωνικά δίκτυα (Desai et al., 1996; West, 2000; Lee et al., 2002), τα δένδρα παλινδρόμησης & ταξινόμησης (*regression and classification tree*), τα πολυδιάστατα προσαρμοσμένα πολυπολυώνυμα παλινδρόμησης (*multivariate adaptive regression splines*) (Lee et al., 2006), καθώς και ο συνδυασμός μεθόδων (Twala, 2009). Εκτενής αναφορά για εφαρμογές στο συγκεκριμένο πεδίο υπάρχει στο βιβλίο του Altman (1981).

2.2 Περιγραφή προβλήματος

Η εφαρμογή τεχνικών ταξινόμησης για την πρόβλεψη και αξιολόγηση του πιστωτικού κινδύνου θεωρούνται υψίστης σημασίας στην Τραπεζική.

Ένας τραπεζικός οργανισμός πρέπει διαρκώς να γνωρίζει πως διαμορφώνεται το ρίσκο που έχει αναλάβει. Επίσης πρέπει να είναι σε θέση να προβλέψει και το πως θα εξελιχθεί. Έτσι μπορεί να διαμορφώσει την βέλτιστη στρατηγική. Στην παρούσα οικονομική συγκυρία, όπου το ποσοστό των πελατών με μη-εξυπηρετούμενα δάνεια αυξάνεται συνεχώς, η ανάγκη

των τραπεζικών οργανισμών για πρόβλεψη της συμπεριφοράς των πελατών γίνεται πιο επιτακτική από ποτέ.

Ειδικότερα, όταν εμφανιστούν οι πρώτες ενδείξεις «κακής» συμπεριφοράς ενός πελάτη (π.χ. καθυστέρηση στην πληρωμή των οφειλών ενός κατόχου καταναλωτικού δανείου), είναι σημαντικό για την τράπεζα να μπορεί να προβλέψει την μελλοντική του εξέλιξη ως προς τις απαιτήσεις της. Αυτή η πρόβλεψη αφενός θα οδηγήσει στην επανεκτίμηση του ρίσκου που αναλάμβανει η τράπεζα εξαιτίας του συγκεκριμένου πελάτη και αφετέρου θα της δώσει την δυνατότητα για έγκαιρη και στοχευμένη διαχείριση ώστε να κρατήσει τον πελάτη συνεπή στις υποχρεώσεις του. Όπως άλλωστε επισημαίνει ο Xavier, «στην διαχείριση των καθυστερημένων δανείων η πρόληψη είναι καλύτερη της θεραπείας».

Στην παρούσα εργασία θα θεωρήσουμε τους πελάτες που έχουν καθυστερήσει να πληρώσουν το πολύ μια δόση (30 ημέρες καθυστέρησης) σε κάποια από τα καταναλωτικά δάνεια τους. Στόχος μας είναι να τους χαρακτηρίσουμε σε «αναξιόπιστους» (*High risk*) και «αξιόπιστους» (*Low risk*) σύμφωνα με την συμπεριφορά που προβλέπεται να έχουν στους επόμενους έξι μήνες. Αν κάποιος θα έχει καθυστερήσει την πληρωμή περισσότερων των 3 δόσεων (90 ημέρες καθυστέρησης) τότε θα θεωρηθεί ως «High risk», διαφορετικά «Low risk».

2.3 Περιγραφή δεδομένων

Από την βάση δεδομένων της τράπεζας Χ επιλέγουμε τους πελάτες που στις 30 Μαϊου 2009 έχουν καθυστερήσει να πληρώσουν μόνο μια δόση σε κάποιο από τα καταναλωτικά προϊόντα τους (πιστωτικές κάρτες, προσωπικά δάνεια). Το δείγμα μας περιλαμβάνει 166.538 πελάτες.

Κύριος στόχος μας είναι, με βάση την υπάρχουσα πληροφόρηση, να κατασκευάσουμε ένα μοντέλο ταξινόμησης, το οποίο θα μπορεί, για πελάτες με αντίστοιχη δανειακή κατάσταση, να προβλέπει πως θα έχει διαμορφωθεί η συμπεριφορά τους ως προς την αποπληρωμή των δανείων τους, στο τέλος του επόμενου εξαμήνου.

Για το σκοπό αυτό έχουμε στη διάθεσή μας τις τιμές διάφορων μεταβλητών, οικονομικών και δημογραφικών, στο τέλος Μαϊου 2009 οι οποίες αποτελούν τις ανεξάρτητες μεταβλητές μας, όπως και τον χαρακτηρισμό των πελατών ως «High risk» και «Low risk», με βάση τον αριθμό των ληξιπρόθεσμων δόσεων τους (τουλάχιστον 4 και το πολύ 3 αντίστοιχα) 6 μήνες αργότερα (30 Νοεμβρίου 2009), που είναι η εξαρτημένη μεταβλητή.

Οι ανεξάρτητες μεταβλητές είναι οι ακόλουθες:

- **“Prod_type”**: ο τύπος του δανείου στο οποίο ο πελάτης έχει καθυστερήσει την πληρωμή μιας δόσης (π.χ. πιστωτική κάρτα, τοκοχρεωλυτικό δάνειο κτλ.)
- **“unpaid_amt”**: το ποσό της ληξιπρόθεσμης δόσης
- **“CL_prod”**: το πιστωτικό όριο του ληξιπρόθεσμου δανείου
- **“exposure_prod”**: η συνολική οφειλή του πελάτη για το ληξιπρόθεσμο δάνειο
- **“utilization”**: το ποσοστό του πιστωτικού ορίου το οποίο έχει χρησιμοποιήσει ο πελάτης για το ληξιπρόθεσμο δάνειο
- **“months_prod”**: οι μήνες που έχουν περάσει από την χορήγηση του ληξιπρόθεσμου δανείου
- **“acc”**: ο αρθμός των καταναλωτικών δανείων που έχει λάβει ο πελάτης από την συγκεκριμένη τράπεζα
- **“CL_total”**: το συνολικό πιστωτικό όριο όλων των δανείων του πελάτη στη συγκεκριμένη τράπεζα
- **“exposure_total”**: η συνολική οφειλή του πελάτη για όλα του τα καταναλωτικά δάνεια στην συγκεκριμένη τράπεζα
- **“income”**: το εισόδημα του πελάτη
- **“DtI”**: οφειλή του πελάτη/εισόδημα του πελάτη
- **“ML”**: παίρνει τιμές 1 και 0 αντίστοιχα αν ο πελάτης έχει ή όχι καθυστερημένο στεγαστικό δάνειο στην συγκεκριμένη τράπεζα
- **“BL”**: παίρνει τιμές 1 και 0 αντίστοιχα αν ο πελάτης έχει ή όχι καθυστερημένο επαγγελματικό δάνειο στην συγκεκριμένη τράπεζα
- **“Age”**: η ηλικία του πελάτη
- **“Occupation”**: το επάγγελμα του πελάτη (4 βασικές κατηγορίες επαγγελμάτων)
- **“Gender”**: το φύλο του πελάτη
- **“Family status”**: η οικογενειακή κατάσταση του πελάτη (4 κατηγορίες: έγγαμος, άγαμος, διαζευγμένος, χήρος)
- **“Chil_no”**: ο αριθμός των παιδιών του πελάτη
- **“Years_in_add”**: έτη διαμονής στην τρέχουσα διεύθυνση (την στιγμή αίτησης του δανείου)

- **“Year_in_comp”**: έτη παραμονής στη τρέχουσα εργασία (την στιγμή αίτησης του δανείου)

Τέλος η μεταβλητή απόκρισης είναι η **“Risk”** η οποία παίρνει τιμές 1 και 0 αν ο πελάτης χαρακτηριζεται «High risk» ή «Low risk» αντίστοιχα.

Στο δείγμα μας περιλαμβάνονται , σύμφωνα με τον Πίνακα 2.1, 158.457 «Low risk» (95,1%) και 8.081 «High risk» (4,9%) πελάτες.

RISK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	158457	95,1	95,1	95,1
	1	8081	4,9	4,9	100,0
	Total	166538	100,0	100,0	

Πίνακας 2.1: Κατανομή πελατών στην αρχική βάση δεδομένων

2.4 Προεπεξεργασία δεδομένων

Οι βάσεις δεδομένων που διαχειριζόμαστε για την επίλυση των διαφόρων προβλημάτων μπορεί να περιλαμβάνουν αναξιόπιστα δεδομένα ή δεδομένα που περιέχουν θόρυβο καθώς και κάποια να είναι ελλειπή. Επίσης ο όγκος των δεδομένων συχνά είναι τόσο μεγάλος που δυσκολεύει τον χειρισμό τους. Για την αντιμετώπιση λοιπόν των παραπάνω προβλημάτων θα πρέπει να γίνει προεπεξεργασία των διαθέσιμων δεδομένων πριν προχωρήσουμε στην εφαρμογή κάποιου μοντέλου ταξινόμησης.

Σύμφωνα με τους Olafsson et al. (2008), οι οποίοι περιγράφουν την ταξινόμηση στα πλαίσια της διαδικασίας της εξόρυξης δεδομένων (*Data Mining process*), η προεπεξεργασία περιλαμβάνει κάποια στάδια όπως τον «καθαρισμό» των δεδομένων (*data cleaning*), τη συμπλήρωση τους (*data integration*), το μετασχηματισμό (*data transformation*) τους με σκοπό την δημιουργία νέων μεταβλητών που θα περιγράφουν καλύτερα το πρόβλημα και έτσι θα βελτιώσουν την προσαρμογή του μοντέλου και την μείωση της διάστασης τους (*data reduction*) με στόχο την αύξηση της απόδοσης του αλγορίθμου.

2.4.1. Έλεγχος και «καθαρισμός» δεδομένων

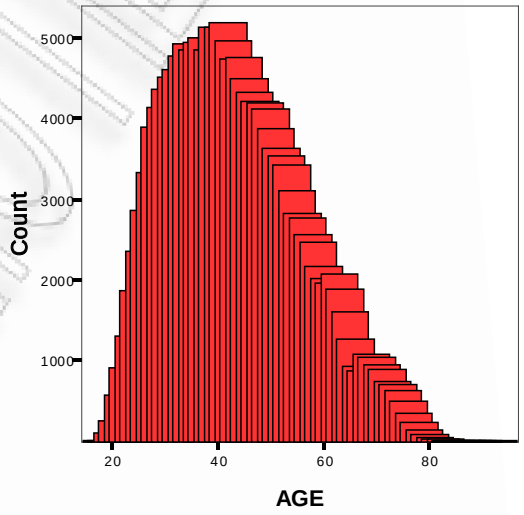
Στη βάση δεδομένων θα προσπαθήσουμε να εντοπίσουμε τυχόν αναξιόπιστες τιμές ή ασυνεπείς συνδιασμούς τιμών.

Από τους πίνακες συχνοτήτων των ανεξάρτητων μεταβλητών παρατηρήθηκε ότι για κάποιες από τις μεταβλητές οι τιμές που έχουν καταγραφεί δεν είναι αποδεκτές. Αυτό οφείλεται κυρίως στην λανθασμένη καταχώρηση των στοιχείων του πελάτη.

Στην μεταβλητή "Age" υπάρχουν τιμές μικρότερες του 18 (Πίνακας 2.2), που αντιστοιχούν στο 0,02% των παρατηρήσεων, όμως κάποιος δεν μπορεί να είναι κάτοχος δανείου εάν δεν είναι ενήλικας. Επίσης υπάρχουν ηλικίες μεγαλύτερες των 100 (0.09%) οι οποίες επίσης δεν θεωρούνται αξιόπιστες.

AGE	Frequency	Percent	Valid Percent	Cumulative Percent
1	1	0,00	0,00	0,00
2	1	0,00	0,00	0,00
3	1	0,00	0,00	0,00
4	2	0,00	0,00	0,00
5	2	0,00	0,00	0,00
6	3	0,00	0,00	0,01
7	1	0,00	0,00	0,01
9	3	0,00	0,00	0,01
11	2	0,00	0,00	0,01
12	2	0,00	0,00	0,01
13	1	0,00	0,00	0,01
14	2	0,00	0,00	0,01
16	3	0,00	0,00	0,01
17	2	0,00	0,00	0,02
•				
•				
•				
104	1	0,00	0,00	99,91
109	147	0,09	0,09	100,00
110	4	0,00	0,00	100,00

Πίνακας 2.2: Πίνακας συχνοτήτων της μεταβλητής "Age"



Σχήμα 2.1: Κατανομή της ηλικίας

Για να απαλείψουμε τα προβλήματα αυτά, θα θεωρήσουμε τις συγκεκριμένες παρατηρήσεις ως ελλείπουσες και θα τις διαχειριστούμε σύμφωνα με την διαδικασία που περιγράφεται στην επόμενη παράγραφο (§ 2.4.2) για την αντιμετώπιση των ελλειπουσών τιμών. Στο Σχήμα 2.1 φαίνεται η κατανομή της ηλικίας αφού εξαιρέθηκαν οι παρατηρήσεις αυτές, η οποία δείχνει κανονική.

Ομοίως, υπάρχουν τιμές για την μεταβλητή "Years_in_add" μεγαλύτερες του 70 (0,01%) και για την "Years_in_com" μεγαλύτερες του 55 (0,01%), οι οποίες δεν είναι αποδεκτές αφού η συγκεκριμένη τράπεζα δεν χορηγεί δάνεια σε πελάτες άνω των 70 ετών.

Στην μεταβλητή "Income" παρατηρούνται τιμές μικρότερες των 500€ (1,5%) και μεγαλύτερες των 200.000€ (0,1%), οι οποίες θεωρούνται μη ορθές, αφού όσον αφορά τις

πρώτες, η συγκεκριμένη τράπεζα δεν εγκρίνει δάνεια σε αιτούντες με εισόδημα μικρότερο των 500€, ενώ για την δεύτερη περίπτωση πελάτες με εισόδημα μεγαλύτερο των 200.000€ αποτελούν διακεκριμένο χαρτοφυλάκιο της τράπεζας (*Special Clients*) το οποίο δεν έχει συμπεριληφθεί στα δεδομένα μας. Οπότε οι τιμές αυτές θα ληφθούν υπόψην ως ελλείπουσες.

Επίσης, η μεταβλητή "Child_no" δείχνει μη αξιόπιστη καθώς το 82% των πελατών εμφανίζεται χωρίς παιδιά (Πίνακας 2.3). Επομένως είναι προτιμότερο να εξαιρεθεί από την ανάλυση του προβλήματός μας.

CHIL_NO

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	136985	82,4	82,4	82,4
	1	8600	5,2	5,2	87,6
	2	15064	9,1	9,1	96,7
	3	4110	2,5	2,5	99,1
	4	1114	,7	,7	99,8
	5	208	,1	,1	99,9
	6	53	,0	,0	100,0
	7	20	,0	,0	100,0
	8	6	,0	,0	100,0
	9	7	,0	,0	100,0
	10	2	,0	,0	100,0
	11	3	,0	,0	100,0
	12	1	,0	,0	100,0
	14	1	,0	,0	100,0
	17	1	,0	,0	100,0
	20	1	,0	,0	100,0
	30	1	,0	,0	100,0
	Total	166177	100,0	100,0	

Πίνακας 2.3: Πίνακας συχνοτήτων της μεταβλητής "Child_no"

Εκτός από τις μεμονωμένες τιμές των μεταβλητών που διορθώσαμε, θα πρέπει να ελέγξουμε τα δεδομένα μας και για ασυνεπείς συνδυασμούς τιμών.

Για παράδειγμα υπάρχουν εγγραφές όπου η ηλικία είναι μικρότερη των ετών διαμονής στην τρέχουσα διεύθυνση ή μικρότερη των ετών παραμονής στην εργασία. Ακόμα υπάρχουν συνδυασμοί όπου η διαφορά ηλικίας και χρόνου εργασίας είναι μικρότερη των 15 χρόνων. Αυτοί απορρίπτονται αφού κάποιος δεν μπορεί να δουλεύει σε ηλικία μικρότερη των 15 ετών.

Στις περιπτώσεις αυτές θεωρούμε ελλείπουσες τις τιμές και των δύο μεταβλητών που εμπλέκονται (age και years_in_add ή years_in_com).

Άλλος ένας έλεγχος που πρέπει να πραγματοποιηθεί είναι στους συνδυασμούς επαγγελμάτων-εισοδήματος, όπου όμως δεν εμφανίζεται κάποιος μη αξιόπιστος συνδυασμός.

Αφού λοιπόν καθαρίσαμε τα δεδομένα μας από αναξιόπιστες μεταβλητές και μη αποδεκτές τιμές, θα πρέπει εν συνεχεία να διαχειριστούμε τις ελλείπουσες τιμές που δημιουργήθηκαν καθώς και εκείνες που υπήρχαν εξ' αρχής στη βάση μας.

2.4.2 Διαχείριση ελλειπουσών τιμών

Οι μέθοδοι για την αντιμετώπιση των ελλειπουσών τιμών διακρίνονται σε δύο κατηγορίες, τις μεθόδους προεπεξεργασίας (*preprocessing methods or sequential methods*) και τις παράλληλες μεθόδους, όπου η συμπλήρωση των δεδομένων γίνεται κατά την κύρια διαδικασία «απόκτησης γνώσης» (Maimon and Rokach, 2005). Στην παρούσα ανάλυση θα εστιάσουμε στην πρώτη κατηγορία.

Οι βασικές μέθοδοι προεπεξεργασίας είναι οι ακόλουθες (η εξαρτημένη μεταβλητή είναι γνωστή για όλες τις παρατηρήσεις)

- Διαγραφή των παρατηρήσεων που έχουν ελλειπείς τιμές (*listwise deletion*)
- Αντικατάσταση της ελλείπουσας τιμής από την τιμή της αντίστοιχης μεταβλητής που εμφανίζεται με την μεγαλύτερη συχνότητα στο σύνολο της βάσης δεδομένων
- Αντικατάσταση από την τιμή που εμφανίζεται με την μεγαλύτερη συχνότητα ανά κατηγορία της εξαρτημένης μεταβλητής (για μη συνεχή εξαρτημένη μεταβλητή)
- Αντικατάσταση από την μέση τιμή της αντίστοιχης μεταβλητής (για συνεχείς ανεξάρτητες μεταβλητές)
- Αντικατάσταση από την μέση τιμή της αντίστοιχης μεταβλητής ανά κατηγορία της εξαρτημένης μεταβλητής (για συνεχείς ανεξάρτητες μεταβλητές και μη συνεχή εξαρτημένη)
- Αντικατάσταση της εγγραφής που περιέχει ελλειπή δεδομένα με ένα σύνολο εγγραφών όπου η ελλείπουσα τιμή έχει αντικατασταθει από κάθε δυνατή τιμή της αντίστοιχης μεταβλητής (μη συνεχείς ανεξάρτητες μεταβλητές)
- Αντικατάσταση της εγγραφής που περιέχει ελλειπή δεδομένα με ένα σύνολο εγγραφών όπου η ελλείπουσα τιμή έχει αντικατασταθει από κάθε δυνατή τιμή της αντίστοιχης μεταβλητής, ανά κατηγορία της εξαρτημένης μεταβλητής (μη συνεχείς ανεξάρτητες και εξαρτημένες μεταβλητές)
- Αντικατάσταση από την γνωστή τιμή μιας άλλης εγγραφής της οποίας το διάνυσμα έχει την «κοντινότερη απόσταση» από το διάνυσμα της εγγραφής με την ελλείπουσα τιμή (*Global closest fit*)

- Αντικατάσταση από την γνωστή τιμή μιας άλλης εγγραφής της οποίας το διάνυσμα έχει την «κοντινότερη απόσταση» από το διάνυσμα της εγγραφής με την ελλείπουσα τιμή, ανά κατηγορία της εξαρτημένης μεταβλητής (*Concept closest fit*)

Για να θεωρηθεί μια μέθοδος κατάλληλη για τον χειρισμό των ελλειπών δεδομένων θα πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

1. Ελαχιστοποίηση της μεροληψίας. Αν και είναι γνωστό ότι τα ελλειπή δεδομένα εισάγουν μεροληψία στις εκτιμήσεις παραμέτρων, μια καλή μέθοδος θα πρέπει να κάνει τη μεροληψία αυτή όσο το δυνατόν μικρότερη.
2. Μεγιστοποίηση της χρήσης της διαθέσιμης πληροφορίας. Θέλουμε να αποφύγουμε την απόρριψη δεδομένων και να χρησιμοποιήσουμε τις διαθέσιμες παρατηρήσεις για να παράγουμε αποτελεσματικές εκτιμήσεις των παραμέτρων.
3. Απόδοση καλών εκτιμήσεων της αβεβαιότητας. Πρέπει να πάρουμε ακριβείς εκτιμήσεις για τα τυπικά σφάλματα, τα διαστήματα εμπιστοσύνης και τα p-values.

Η μέθοδος της διαγραφής έχει το προφανές μειονέκτημα της απώλειας δεδομένων, τα οποία δυνητικά θα μπορούσαν να χρησιμοποιηθούν. Από τη μια, η απώλεια αυτή οδηγεί σε μεγαλύτερα τυπικά σφάλματα, ευρύτερα διαστήματα εμπιστοσύνης καθώς και μείωση της ισχύος στους ελέγχους υποθέσεων. Από την άλλη, οι εκτιμήσεις των τυπικών σφαλμάτων που προκύπτουν μετά την διαγραφή είναι συνήθως ακριβείς εκτιμήσεις των πραγματικών τυπικών σφαλμάτων. Οπότε μπορούμε να πούμε ότι η μέθοδος της διαγραφής είναι αξιόπιστη για την επεξεργασία των ελλειπουσών τιμών, αντίθετα με άλλες κλασσικές μεθόδους (Allison, 2002).

Οι μεταβλητές στις οποίες δημιουργήσαμε ελλείπουσες τιμές με τον καθαρισμό των δεδομένων είναι οι “Age”, “Years_in_comp”, “Years_in_add” και η “Income”.Ο συνολικός αριθμός των παραπάνω εγγραφών ανέρχεται σε 9.099 και αντιστοιχεί στο 5,5% του αρχικού δείγματος, το οποίο όμως είναι πολύ μεγάλο, γεγονός που μας επιτρέπει να τις διαγράψουμε με μεγαλύτερη ευκολία, χωρίς να χάσουμε μεγάλο μέρος της διαθέσιμης πληροφορίας. Έτσι η νέα βάση που προκύπτει περιλαμβάνει 157.439 πελάτες.

Εξ’ αρχής ελλειπή δεδομένα στην βάση είχαν οι μεταβλητές “Occupation”, “Family Status”, “Gender” και “Income”, τα οποία στη νέα βάση διαμορφώνονται σύμφωνα με τον Πίνακα 2.4.

Statistics

		OCCUPATION	FAMILY_STATUS	GENDER	INCOME
N	Valid	157.431	156.109	157.426	132.843
	Missing	8	1.330	13	24.596
%	Missing	0,01%	0,84%	0,01%	**15,62%**

Πίνακας 2.4: Στατιστικά ελλειπουσών τιμών

Όπως παρατηρούμε το ποσοστό των ελλιπών δεδομένων είναι αρκετά υψηλό για την μεταβλητή του εισοδήματος. Οπότε θα προσπαθήσουμε να εντοπίσουμε τυχόν ιδιαίτερα χαρακτηριστικά των πελατών αυτών. Πράγματι, κοιτώντας τις μέσες τιμές των υπόλοιπων μεταβλητών, για τους πελάτες με διαθέσιμη και μη την πληροφορία του εισοδήματος, προκύπτει ότι η μέση παλαιότητα του δανείου είναι 38 μήνες για την πρώτη κατηγορία και σχεδόν διπλάσια (61 μήνες) για την δεύτερη. Το γεγονός λοιπόν της μεγαλύτερης παλαιότητας δικαιολογεί την απώλεια της πληροφορίας, αφενός λόγω του τρόπου συμπλήρωσης των αιτήσεων για παροχή δανείου που στις περισσότερες περιπτώσεις ήταν χειρόγραφες και περιοριζόταν σε πολύ συγκεκριμένα πεδία βάσει πολιτικής εγκρίσεων και αφετέρου εξαιτίας της χρήσης όχι και τόσο εξελιγμένων συστημάτων αποθήκευσης δεδομένων (*data warehouse*).

Διαγράφοντας και αυτές τις παρατηρήσεις το τελικό μέγεθος του δείγματος ανέρχεται σε 132.102 παρατηρήσεις, παραμένοντας ικανοποιητικό. Η εικόνα για το ρίσκο των πελατών αυτών είναι όμοια με εκείνη του αρχικού δείγματος, 95% Low risk - 5% High risk (Πίνακες 2.1, 2.5).

RISK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	125142	94,7	94,7	94,7
	1	6960	5,3	5,3	100,0
	Total	132102	100,0	100,0	

Πίνακας 2.5: Κατανομή πελατών στο τελικό δείγμα

2.4.3 Επιλογή δείγματος εκμάθησης και ελέγχου

Όπως θα δούμε αναλυτικά στο Κεφάλαιο (5), για την αξιολόγηση της απόδοσης των μεθόδων ταξινόμησης που θα εφαρμόσουμε, χρειαζόμαστε και ένα δείγμα ελέγχου (*test set*). Το δείγμα αυτό θα πρέπει να είναι ανεξάρτητο από το εκπαιδευτικό και αντιπροσωπευτικό του αρχικού. Ποιος είναι όμως ο βέλτιστος τρόπος να μοιράσουμε την βάση μας στα δύο αυτά δείγματα;

Γενικά, όσο πιο μεγάλο είναι το δείγμα εκπαίδευσης τόσο καλύτερη είναι η ταξινόμηση, αν και η απόδοση αρχίζει να μειώνεται όταν το μέγεθος του εκπαιδευτικού δείγματος υπερβεί μια ορισμένη ποσότητα. Επίσης όσο πιο μεγάλο είναι το δοκιμαστικό δείγμα τόσο πιο ακριβής είναι η εκτίμηση του σφάλματος ταξινόμησης. Γεννιέται λοιπόν το δίλλημα: να βρούμε ένα καλό μοντέλο ταξινόμησης ή να πετύχουμε μια καλή εκτίμηση του σφάλματος; Αν ο διαθέσιμος όγκος των δεδομένων είναι μεγάλος, τότε το πρόβλημα εξαλοίφεται. Χρησιμοποιούμε ένα μεγάλο δείγμα για εκπαίδευση και ένα επίσης μεγάλο δείγμα για έλεγχο (Witten and Frank, 2005).

Στην περίπτωση μας το αρχικό δείγμα είναι αρκετά μεγάλο, οπότε μπορούμε να χρησιμοποιήσουμε το μισό για την εκμάθηση του αλγορίθμου και το υπόλοιπο για την εκτίμηση του λάθους στην ταξινόμηση. Θα επιλέξουμε λοιπόν με τυχαίο τρόπο δύο ισομεγέθη δείγματα από την αρχική μας βάση.

Τα δύο δείγματα διαμορφώνονται σύμφωνα με τον Πίνακα 2.6. Το δείγμα εκμάθησης αποτελεί από το 50,1% των υποθέσεων και το δοκιμαστικό το 49,9%.

Approximately 50 % of cases (SAMPLE)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Test	65941	49,9	49,9	49,9
	Training	66161	50,1	50,1	100,0
	Total	132102	100,0	100,0	

Πίνακας 2.6: Δείγματα εκπαίδευσης και ελέγχου

Η αρχική κατανομή των πελατών με βάση το ρίσκο τους (95%-5%) διατηρείται και στα δύο δείγματα (Πίνακας 2.7)

RISK

SET			Frequency	Percent	Valid Percent	Cumulative Percent
Test	Valid	0	62399	94,6	94,6	94,6
		1	3542	5,4	5,4	100,0
		Total	65941	100,0	100,0	
Training	Valid	0	62743	94,8	94,8	94,8
		1	3418	5,2	5,2	100,0
		Total	66161	100,0	100,0	

Πίνακας 2.7: Κατανομή πελατών στα δείγματα εκπαίδευσης και ελέγχου

Θα πρέπει επίσης να ελέγξουμε ότι το εκπαιδευτικό δείγμα ακολουθεί την ίδια κατανομή με το αρχικό συνολικό δείγμα σε όλα τα κύρια χαρακτηριστικά (*in all mayor aspects*). Πράγματι, για τις κατηγορικές ανεξάρτητες μεταβλητές καθώς και για την εξαρτημένη από

τον Πίνακα 2.8 προκύπτει ότι οι κατανομές τους δεν διαφοροποιούνται στο εκπαιδευτικό δείγμα σε σχέση με το αρχικό.

Variables		% Training	% Total
Risk	0	94,8	94,7
	1	5,2	5,3
Occupation	A	26,1	26,2
	B	13,9	13,8
	F	21,6	21,5
	G	38,4	38,5
ML	0	96,8	96,7
	1	3,2	3,3
BL	0	96,7	96,7
	1	3,3	3,3

Variables		% Training	% Total
Family Status	A	59,8	59,8
	B	31,6	31,7
	C	6,0	6,0
	E	2,6	2,6
Gender	H	37,6	37,6
	V	62,4	62,4
Prod_type	CARD	23,9	23,9
	OPEN	14,9	14,8
	TKX	61,1	61,2

Πίνακας 2.8: Κατανομές κατηγορικών μεταβλητων στα δείγματα εκπαίδευσης και ελέγχου

Για κάθεμια από τις συνεχείς μεταβλητές, με βάση το *t-test* σύγκρισης 2 κατηγοριών, προκύπτει ότι η διασπορά δεν διαφέρει μεταξύ του συνολικού δείγματος και του εκπαιδευτικού, αφού τα *p-values* στο τεστ ισότητας των διακυμάνσεων του Levene είναι όλα μεγαλύτερα του 0,05 (Πίνακας 2.9). Ομοίως και για τις μέσες τιμές όπου το *t-test* ισότητας των μέσων δίνει *p-values* επίσης μεγαλύτερα του 0,05 για όλες τις μεταβλητές.

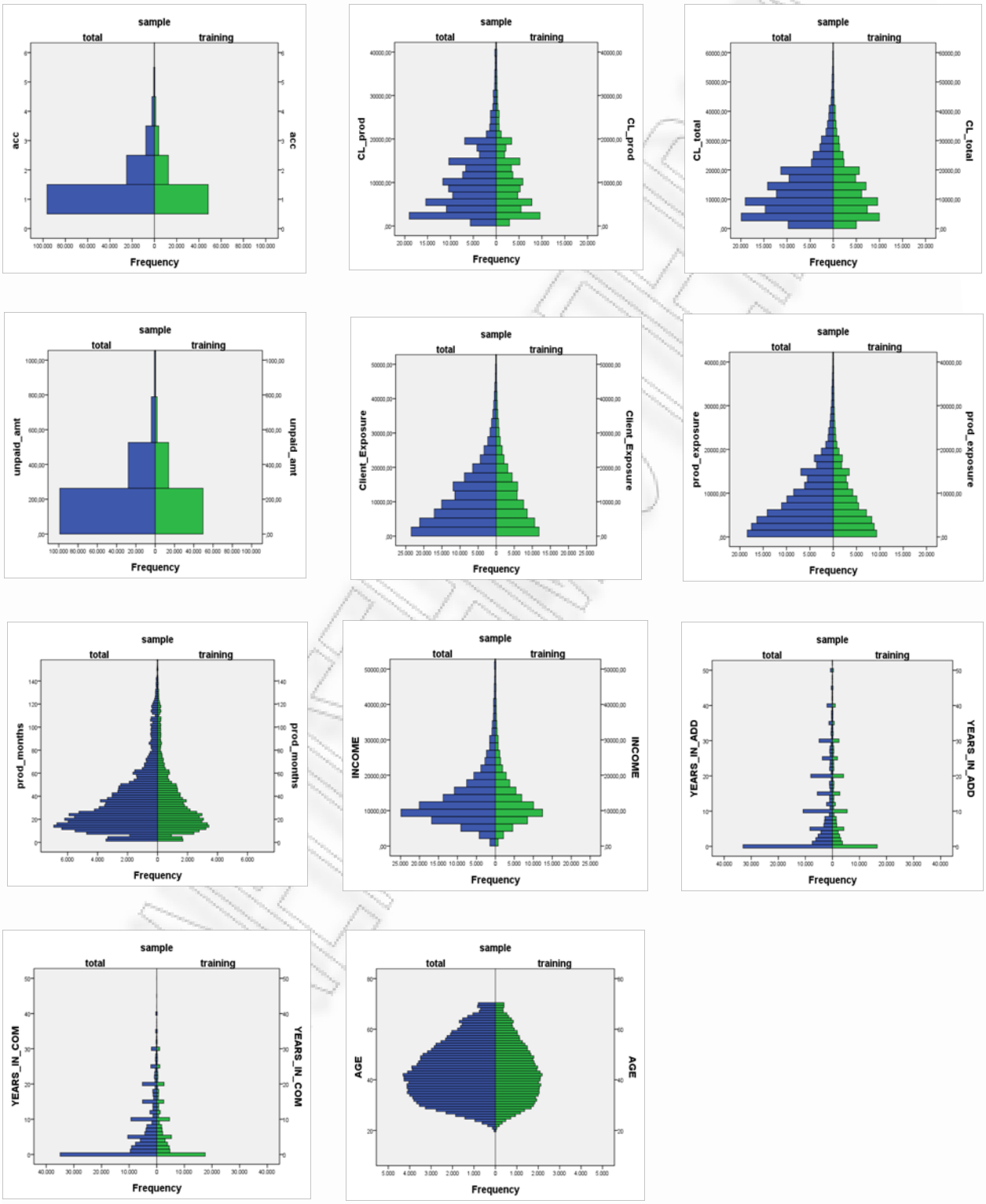
Independent Samples Test

		Levene's Test for Equality of Variances	t-test for Equality of Means
		Sig.	Sig. (2-tailed)
acc	Equal variances assumed	,624	,759
CL_prod	Equal variances assumed	,069	,065
CL_total	Equal variances assumed	,255	,122
unpaid_amt	Equal variances assumed	,934	,651
Client_Exposure	Equal variances assumed	,193	,190
prod_exposure	Equal variances assumed	,089	,122

		Levene's Test for Equality of Variances	t-test for Equality of Means
		Sig.	Sig. (2-tailed)
prod_months	Equal variances assumed	,927	,867
INCOME_ORIGIN	Equal variances assumed	,208	,304
YEARS_IN_ADD	Equal variances assumed	,452	,562
YEARS_IN_COM	Equal variances assumed	,367	,628
AGE_ORIGINAL	Equal variances assumed	,573	,608

Πίνακας 2.9: Έλεγχος ισότητας των διακυμάνσεων και των μέσων τιμών για το εκπαιδευτικό & το συνολικό δείγμα

Επίσης στο Σχήμα 2.2 φαίνονται οι κατανομές των μεταβλητών στα δύο δείγματα, οι οποίες δείχνουν όμοιες μεταξύ των δειγμάτων.



Σχήμα 2.2: Κατανομές των συνεχών μεταβλητών για το εκπαιδευτικό και το συνολικό δείγμα

Επομένως αφού έχουμε εξασφαλίσει ότι το εκπαιδευτικό μας δείγμα είναι αντιπροσωπευτικό του συνολικού μπορούμε να το χρησιμοποιήσουμε για την εκμάθηση του μοντέλου ταξινόμησης.

ΚΕΦΑΛΑΙΟ 3

Παρουσίαση Στατιστικών Μεθόδων

3.1 Εισαγωγή

Οι βάσεις δεδομένων που προκύπτουν σε σύγχρονες εφαρμογές συχνά περιέχουν μεγάλο αριθμό μεταβλητών, κάποιες από τις οποίες είτε είναι πλεονάζουσες είτε δεν έχουν σχέση με το υπό εξέταση πρόβλημα (Fayyad et al., 1996). Η παρουσία τέτοιων μεταβλητών μπορεί να έχει συγχυτική επίδραση στην κατανομή των πραγματικά χρήσιμων μεταβλητών και να μειώσει την απόδοση των μοντέλων ταξινόμησης (Zhao et al., 2009 και αναφορές). Επιπλέον, η αυξημένη διάσταση του χώρου όπου ορίζονται οι μεταβλητές αυξάνει την πολυπλοκότητα των αλληλεπιδράσεων μεταξύ των μεταβλητών καθώς και τον θόρυβο, περιορίζοντας έτσι την αποδοτικότητα και την αποτελεσματικότητα των μεθόδων ταξινόμησης (Piramuthu et al.,1998).

Για την αντιμετώπιση των παραπάνω προβλημάτων έχουν εφαρμοστεί διάφορες προσεγγίσεις (Olafson et al., 2008) οι οποίες μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες: την επιλογή (*selection*), την εξαγωγή (*extraction*) και την κατασκευή (*construction*) μεταβλητών (Liu and Motoda, 1998). Η επιλογή των μεταβλητών αναφέρεται στην επιλογή ενός «καλού» υποσυνόλου των αρχικών μεταβλητών το οποίο περιέχει το μεγαλύτερο μέρος της χρήσιμης πληροφορίας για ένα συγκεκριμένο πρόβλημα. Η εξαγωγή και η κατασκευή περιλαμβάνουν την εύρεση ενός συνόλου σύνθετων μεταβλητών οι οποίες είναι συναρτήσεις των αρχικών μεταβλητών. Η εξαγωγή προβάλλει ένα χώρο υψηλών διαστάσεων που ορίζονται οι μεταβλητές σε ένα νέο μικρότερων διαστάσεων, μέσω γραμμικών και μη μετασχηματισμών έτσι ώστε να διατηρείται ο βασικός όγκος της αρχικής πληροφορίας. Η κατασκευή διορθώνει το πρόβλημα των αλληλεπιδράσεων ανακαλύπτοντας «καλούς» συνδυασμούς των πρωτογενών μεταβλητών.

Στο πρόβλημα που εξετάζουμε περιλαμβάνονται 7 κατηγορικές και 11 συνεχείς μεταβλητές. Ο περιορισμός των κατηγορικών μεταβλητών θα γίνει με την βοήθεια της ανάλυσης κατά συστάδες (*Cluster Analysis*) ενώ η μείωση της διάστασης των συνεχών θα πραγματοποιηθεί με την μέθοδο της ανάλυσης παραγόντων (*Factor Analysis*).

Επίσης θα χρησιμοποιήσουμε την ανάλυση συστάδων για να περιγράψουμε το προφίλ των πελατών (*profiling*).

Άλλη μια ιδιαιτερότητα που εμφανίζουν τα δεδομένα μας, είναι η μη-ισορροπημένη συμμετοχή των δύο ομάδων (αξιόπιστοι και αναξιόπιστοι πελάτες) στο δείγμα μας (*imbalanced data*). Όπως είδαμε στην παράγραφο 2.3 οι αναξιόπιστοι πελάτες αποτελούν μόλις το 5% του συνόλου που έχουμε στη διάθεσή μας. Στην παράγραφο 3.4 θα δούμε αναλυτικά πως αντιμετωπίζονται τέτοιες περιπτώσεις.

3.2 Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες (*Cluster Analysis*) είναι μια πολυμεταβλητή μέθοδος, η οποία εξετάζει την ομοιότητα ορισμένων παρατηρήσεων ως προς ένα σύνολο μεταβλητών, με σκοπό να ομαδοποιήσει τις παρατηρήσεις που μοιάζουν μεταξύ τους. Μια επιτυχημένη εφαρμογή της μεθόδου θα πρέπει να καταλήξει σε ομάδες όπου οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, ενώ οι παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο το δυνατόν περισσότερο.(βλ. Κούτρας, 2007)

Υπάρχουν δύο είδη μεθόδων συσταδοποίησης με βάση το αν είναι γνωστός εξαρχής ή όχι ο αριθμός των συστάδων. Στην πρώτη περίπτωση οι μέθοδοι ονομάζονται μή ιεραρχικές (*non-hierarchical ή partitioning ή k-means clustering*) ενώ στη δεύτερη ιεραρχικές (*hierarchical*). Οι Han and Kamber (2001) προτείνουν τρεις επιπλέον κατηγορίες: μεθόδους βασισμένες στις συναρτήσεις πυκνότητας (*density-based methods*), σε μοντέλα που για κάθε συστάδα υποθέτουν ένα μοντέλο (*model-based clustering*) και σε μεθόδους που βασίζονται σε δομές πλέγματος πολλαπλών επιπέδων (*grid-based methods*).

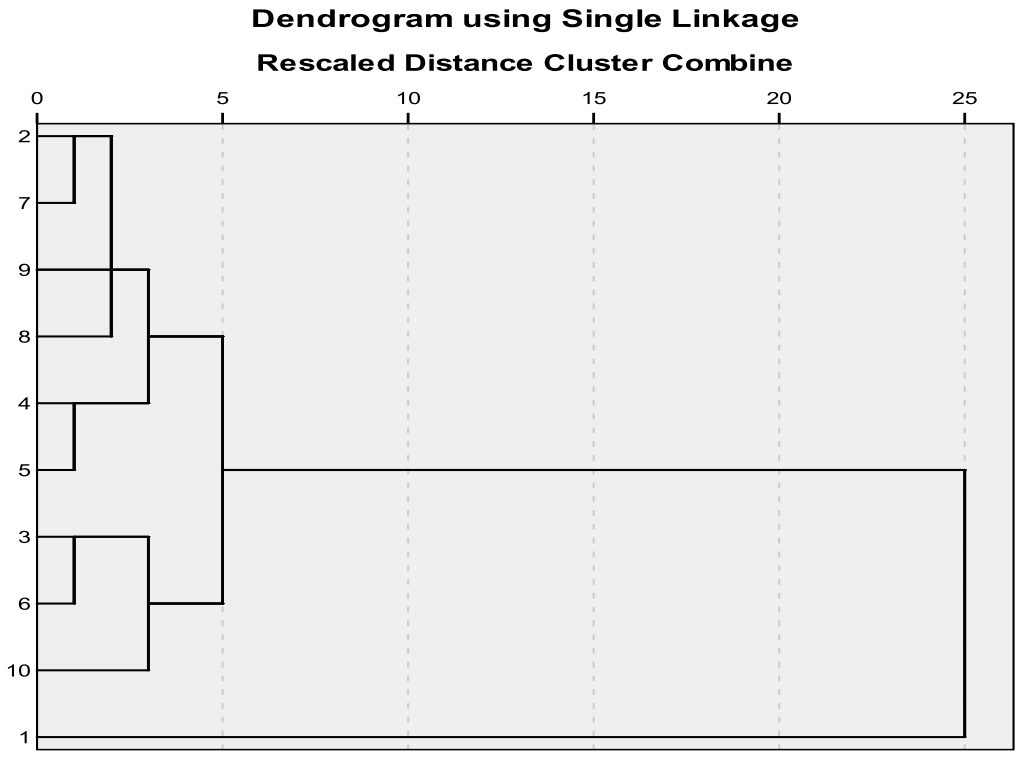
Οι ιεραρχικές μέθοδοι διακρίνονται επιπλέον σε:

- Συσσωρευτικές μεθόδους (*agglomerative methods*), όπου ξεκινώντας κάθε παρατήρηση αποτελεί μια συστάδα. Οι δύο πιο όμοιες συστάδες ενώνονται στο επόμενο βήμα και η διαδικασία επαναλαμβάνεται έτσι ώστε να καταλήξουμε σε μια συστάδα που περιέχει όλες τις παρατηρήσεις. Στο τέλος επιλέγουμε τον βέλτιστο αριθμό συστάδων από το σύνολο των λύσεων που έχουν παραχθεί με την βοήθεια του δενδρογράμματος όπως θα δούμε αναλυτικά παρακάτω.
- Διαιρετικές μεθόδους (*divisive methods*), όπου αρχικά όλες οι παρατηρήσεις ανήκουν σε μια συστάδα η οποία στη συνέχεια διαιρείται σε υπο-συστάδες και τελικά καταλήγουμε κάθε παρατήρηση να αποτελεί μια ομάδα. Κι εδώ ο αριθμός των συστάδων επιλέγεται αφού ολοκληρωθεί η παραπάνω διαδικασία όπως θα δούμε στη συνέχεια. (Maimon and Rokach, 2005).

Οι πιο διαδεδομένες και μάλιστα με διαφορά είναι οι συσσωρευτικές. (Tan et al., 2006; Κούτρας, 2007). Ο βασικός λόγος είναι ότι οι διαιρετικές μέθοδοι έχουν πολύ μεγαλύτερο υπολογιστικό κόστος (έχοντας n παρατηρήσεις, για το πρώτο βήμα το πλήθος των υπολογισμών που απαιτούνται είναι της τάξης του 2^{n-1}, ενώ για τις συσσωρευτικές είναι της τάξης του n^2). (βλ. Κούτρας, 2007)

Ένα βασικό μειονέκτημα των ιεραρχικών μεθόδων, πέραν του μεγάλου υπολογιστικού κόστους, είναι ότι παρατηρήσεις που θα ενωθούν κάποια στιγμή θα εξακολουθούν να μένουν μαζί χωρίς να υπάρχει δυνατότητα μετακίνησης τους σε άλλη ομάδα.

Το αποτέλεσμα των ιεραρχικών μεθόδων, είτε είναι συσσωρευτικές είτε διαιρετικές, είναι ένα δενδρόγραμμα (Σχήμα 3.1) στο οποίο αναπαριστάται η διαδοχική ομαδοποίηση καθώς και το επίπεδο ομοιότητας (απόσταση) που οδηγεί στην διαμόρφωση των ομάδων κάθε βήματος (Johnson and Wichern, 1998; Maimon and Rokach, 2005; Cios et al., 2007).



Σχήμα 3.1: Δενδρόγραμμα

Ο βέλτιστος αριθμός συστάδων μπορεί να προκύψει απεικονιστικά από το παραπάνω σχήμα . Εάν από το σημείο του δενδρογράμματος που υπάρχει η μεγαλύτερη μεταβολή της απόστασης ή της ομοιότητας (οριζόντιος άξονας) φέρουμε μια γραμμή παράλληλη προς τον κατακόρυφο άξονα, τότε ο βέλτιστος αριθμός των ομάδων είναι ίσος με τα σημεία που τέμνει η γραμμή (βλ. Κούτρας, 2007). Ένας δεύτερος τρόπος είναι να σταματάει η ένωση των

ομάδων μόλις ξεπεραστεί μια συγκεκριμένη κατώτατη τιμή της απόστασης, ώστε να αποφευχθεί η συνένωση διακριτών ομάδων (Cios et al., 2007).

Επειδή τα παραπάνω εμπεριέχουν το στοιχείο της υποκειμενικότητας, έχουν προταθεί πιο αντικειμενικές μέθοδοι, οι οποίες βασίζονται είτε στην ανάλυση διακύμανσης (ANOVA) είτε στις αποστάσεις από τα κέντρα των ομάδων. (Κούτρας, 2007).

3.2.1 Μέτρα απόστασης/ομοιότητας παρατηρήσεων

Μέχρι τώρα έχουμε αναφερθεί στην έννοια της απόστασης και της ομοιότητας μεταξύ υποκειμένων άλλά δεν τις έχουμε ορίσει.

Ας θεωρήσουμε ένα δείγμα *n* ατόμων για το οποίο έχουμε στη διάθεσή μας τις τιμές *p* μεταβλητών. Ο πίνακας δεδομένων που αντιστοιχεί στο παραπάνω πρόβλημα είναι ένας $(n \times p)$ πίνακας όπου το στοιχείο $(i\ ,\ j) = x_{ij}$ απεικονίζει την τιμή της μεταβλητής *j* που παρατηρήθηκε για το άτομο *i*. Το διάνυσμα $\vec{x}_i = (x_{i1},...,x_{ip})$ είναι το διάνυσμα που περιέχει τις τιμές των *p* μεταβλητών για το άτομο *i*.

Ως απόσταση μπορούμε να θεωρήσουμε μια συνάρτηση $d_{ij} = d(\vec{x}_i, \vec{x}_j)$ η οποία να ικανοποιεί τις εξής ιδιότητες:

1. $d_{ij} \geq 0$ για κάθε *i, j* και $d_{ij} = 0 \Leftrightarrow i = j$
2. $d_{ij} \leq d_{is} + d_{sj}$ (τριγωνική ανισότητα)
3. $d_{ij} = d_{ji}$ (συμμετρική ιδιότητα)

Όμως, πολλά από τα μέτρα που χρησιμοποιούνται στην πράξη δεν ικανοποιούν την τριγωνική ανισότητα, ενώ σε ορισμένες περιπτώσεις δεν απαιτείται η συμμετρία (Κούτρας, 2007).

Για συνεχή δεδομένα οι πιο συνηθισμένες αποστάσεις που χρησιμοποιούνται είναι οι ακόλουθες:

- Ευκλείδεια απόσταση: $d_{ij} = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$

  Η απόσταση αυτή έχει το μειονέκτημα ότι εξαρτάται από την κλίμακα μέτρησης και καθορίζεται σε πολύ μεγάλο βαθμό από τις μεταβλητές με μεγάλες απόλυτες τιμές.

- Απόσταση του Pearson: $d_{ij} = \sqrt{\sum_{r=1}^{p}\left(\frac{x_{ir} - x_{jr}}{s_r}\right)^2}$, όπου s_r είναι η διακύμανση της r μεταβλητής, $s_r = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ir} - \frac{1}{n}\sum_{i=1}^{n} x_{ir}\right)^2}$

  Με τον τρόπο αυτό κάνουμε συγκρίσιμη την κλίμακα των μεταβλητών, διαιρώντας τις με την τυπική τους απόκλιση.

- Απόσταση Mahalanobis: $d_{ij} = (\vec{x}_i - \vec{x}_j)\Sigma^{-1}(\vec{x}_i - \vec{x}_j)'$, όπου Σ είναι ο δειγματικός πίνακας διακύμανσης-συνδιακύμανσης των διανυσμάτων $\vec{x}_i$ και $\vec{x}_j$.

  Ο τύπος αυτός λαμβάνει υπόψη του τις συνδιακυμάνσεις μεταξύ των μεταβλητών, κάτι που δεν κάνουν οι προηγούμενες αποστάσεις που ορίσαμε.

- Απόσταση Manhattan ή City-block metric: $d_{ij} = \sum_{r=1}^{p}\left|x_{ir} - x_{jr}\right|$

  Είναι καταλληλότερη από την ευκλείδεια για περιπτώσεις που υπάρχουν έκτροπες παρατηρήσεις, γιατί τους δίνει μικρότερο βάρος, αφού δεν υψώνει τις διαφορές στο τετράγωνο.

- Απόσταση Minkowski: $d_{ij} = \left(\sum_{r=1}^{p}\left|x_{ir} - x_{jr}\right|^{\lambda}\right)^{1/\lambda}$

- Απόσταση max ή Chebyshev: $d_{ij} = \max_{r=1,\ldots,p}\left|x_{ir} - x_{jr}\right|$

- Απόσταση του Gower: $d_{ij} = \sum_{r=1}^{n}\frac{\left|x_{ir} - x_{jr}\right|}{R_r}$, όπου $R_r = \max_{i=1,\ldots,p} x_{ir} - \min_{i=1,\ldots,p} x_{ir}$

  Χρησιμοποιείται συνήθως όταν υπάρχουν και δίτιμες μεταβλητές στα δεδομένα μας.

- Απόσταση Bhattacharyya: $d_{ij} = \sqrt{\sum_{r=1}^{p}\left(\sqrt{x_{ir}} - \sqrt{x_{jr}}\right)^2}$

  Χρησιμοποιείται όταν τα δεδομένα αποτελούνται από ποσοστά.

- Canberra metric: $d_{ij} = \sum_{r=1}^{p}\frac{\left|x_{ir} - x_{jr}\right|}{\left|x_{ir}\right| + \left|x_{jr}\right|}$

Οι αποστάσεις στις οποίες αναφερθήκαμε έως τώρα αφορούν σε συνεχή δεδομένα. Στη συνέχεια θα δούμε πως ορίζονται αποστάσεις για δίτιμες μεταβλητές (*binary variables*) οι οποίες μπορούν να πάρουν μια από τις τιμές 0-1 (π.χ. απουσία-παρουσία κάποιου χαρακτηριστικού).

Συγκρίνοντας την τιμή της *r*-στης μεταβλητής στα άτομα *i* και *j* έχουμε ότι $(x_{ir} - x_{jr})^2 = \begin{cases} 0 \text{ αν } x_{ir} = x_{jr} \\ 1 \text{ αν } x_{ir} \neq x_{jr} \end{cases}$. Επομένως η ευκλείδεια απόσταση θα λαμβάνει υπόψη τον αριθμό των ασυμφωνιών μεταξύ των δύο ατόμων.

Τα μέτρα που συνήθως χρησιμοποιούνται στην πράξη αφενός λαμβάνουν υπόψη το ποσοστό των συμφωνιών/ασυμφωνιών αντί του πλήθους τους και αφετέρου μπορούν να δώσουν διαφορετική βαρύτητα σε καθένα συνδυασμό των τιμών 0 και 1.

Για την ευκολότερη κατανόηση των αποστάσεων d_{ij} που θα ορίσουμε παρακάτω, θεωρούμε τον εξής πίνακα συνάφειας

		άτομο *j*		
άτομο *i*		**0**	**1**	
	0	a	b	**a+b**
	1	c	d	**c+d**
		a+c	**b+d**	***p=a+b+c+d***

Πίνακας 3.1: Πίνακας συνάφειας δύο ατόμων

όπου φαίνεται το πλήθος των διάφορων συνδυασμών των τιμών 0 και 1.

Με βάση αυτό μπορούμε να κατασκευάσουμε τις ακόλουθες αποστάσεις

- Simple matching distance: $d_{ij} = \dfrac{b+c}{a+b+c+d}$
- Rogers and Tanimoto distance: $d_{ij} = \dfrac{2(b+c)}{(a+d)+2(b+c)}$
- Sokal and Sneath distance: $d_{ij} = \dfrac{b+c}{2(a+d)+(b+c)}$
- Jaccard distance: $d_{ij} = \dfrac{b+c}{a+b+c}$
- Dice and Sorensen distance: $d_{ij} = \dfrac{b+c}{2a+b+c}$

Αν τα δεδομένα μας αποτελούνται από μη διατάξιμες κατηγορικές μεταβλητές, τότε κατασκευάζουμε μια δίτιμη μεταβλητή για κάθε επίπεδο της κατηγορικής και υπολογίζουμε την απόσταση για τις νέες μεταβλητές.

Όταν οι κατηγορικές μεταβλητές είναι διατάξιμες, τις θεωρούμε ως συνεχείς και χρησιμοποιούμε κάποια από τις αποστάσεις που αναφέραμε προηγουμένως, προσέχοντας όμως να έχουμε την ίδια κλίμακα σε όλες τις μεταβλητές.

Όλα τα μέτρα απόστασης που ορίσαμε παραπάνω έχουν το χαρακτηριστικό γνώρισμα ότι παρατηρήσεις που μοιάζουν πολύ μεταξύ τους δίνουν πολύ μικρή τιμή στην απόσταση. Αντίστοιχα μπορούμε να ορίσουμε τα μέτρα ομοιότητας τα οποία θα πρέπει να έχουν πολύ μεγάλη τιμή για παρατηρήσεις που μοιάζουν πολύ μεταξύ τους.

Η συνάρτηση ομοιότητας $s_{ij} = s(\vec{x}_i, \vec{x}_j)$ μεταξύ των ατόμων i και j θα πρέπει να ικανοποιεί τις εξής ιδιότητες:

1. $s_{ij} \geq 0$ για κάθε i, j και $i = j \Rightarrow s_{ij} = 1$
2. $s_{ij} \leq 1$
3. $s_{ij} = s_{ji}$ (συμμετρική ιδιότητα)

Για συνεχείς μεταβλητές το πιο συνηθισμένο μέτρο ομοιότητας είναι ο συντελεστής συσχέτισης $s_{ij} = \dfrac{\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})(x_{jr} - \bar{x}_{j\cdot})}{\sqrt{\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})^2 \sum_{k=1}^{p}(x_{jk} - \bar{x}_{j\cdot})^2}}$, όπου $\bar{x}_{i\cdot} = \dfrac{1}{p}\sum_{r=1}^{p} x_{ir}$ και $\bar{x}_{j\cdot} = \dfrac{1}{p}\sum_{r=1}^{p} x_{jr}$.

Όταν πρόκειται για δίτιμες μεταβλητές και σύμφωνα με τον πίνακα συνάφειας 3.1 τα πιο γνωστά μέτρα ομοιότητας είναι τα ακόλουθα:

1. Simple matching: $s_{ij} = \dfrac{a+d}{a+b+c+d}$
2. Rogers and Tanimoto: $s_{ij} = \dfrac{a+d}{(a+d)+2(b+c)}$
3. Sokal and Sneath: $s_{ij} = \dfrac{2(a+d)}{2(a+d)+(b+c)}$
4. Jaccard coefficient: $s_{ij} = \dfrac{a}{a+b+c}$
5. Dice and Sorensen: $s_{ij} = \dfrac{2a}{2a+b+c}$
6. Russel and Rao: $s_{ij} = \dfrac{a}{a+b+c+d}$
7. Sokal and Sneath II: $s_{ij} = \dfrac{a}{a+2(b+c)}$
8. Sokal and Sneath III: $s_{ij} = \dfrac{a+d}{b+c}$

9. Kulczynski: $s_{ij} = \frac{a}{b+c}$

Όλα τα παραπάνω μέτρα ενσωματώνουν βαρύτητες για τις συμφωνίες και τις ασυμφωνίες των χαρακτηριστικών. Ειδικότερα τα μέτρα 1 έως 5 προκύπτουν από τις αντίστοιχες αποστάσεις με βάση τον τύπο $s_{ij} = 1 - d_{ji}$.

Στην περίπτωση που έχουμε μεικτά δεδομένα (συνεχείς, κατηγορικές και δίτιμες μεταβλητές) ο συντελεστής ομοιότητας του Gower είναι ο $s_{ij} = \frac{\sum_{r=1}^{p} w_{ij}(r) s_{ij}(r)}{\sum_{r=1}^{p} w_{ij}(r)}$ όπου

$\Rightarrow$ Αν η *r* μεταβλητή είναι συνεχής, τότε $s_{ij}(r) = 1 - \frac{\left| x_{ir} - x_{jr} \right|}{\max\limits_{i=1,\ldots,p} x_{ir} - \min\limits_{i=1,\ldots,p} x_{ir}}$

$\Rightarrow$ Αν η *r* μεταβλητή είναι διακριτή, τότε $s_{ij}(r) = \begin{cases} 1, & \text{αν } x_{ir} = x_{jr} \\ 0, & \text{αλλιώς} \end{cases}$

Επίσης τα βάρη $w_{ij}(r)$ παίρνουν τιμές 1 και 0 αντίστοιχα, αν έχει νόημα η σύγκριση στην *r* μεταβλητή ή όχι.

3.2.2 Απόσταση ομάδων

Για την συνένωση των ομάδων στις συσσωρευτικές μεθόδους απαιτείται και ο ορισμός της απόστασης μεταξύ των ομάδων. Ανάλογα με το είδος της απόστασης που χρησιμοποιούμε στις συσσωρευτικές μεθόδους διακρίνουμε τις ακόλουθες περιπτώσεις (βλ. Κούτρας, 2007)

- Μέθοδος της απλής συνένωσης (*Single Linkage*) ή του «κοντινότερου γείτονα» (*Nearest Neighbor*). Η απόσταση μεταξύ δύο ομάδων ορίζεται ως η μικρότερη απόσταση από μια παρατήρηση της μιας ομάδας σε μια παρατήρηση της άλλης ομάδας.
- Μέθοδος της πλήρους συνένωσης (*Complete Linkage*) ή του «μακρυνότερου γείτονα» (*Furthest Neighbor*). Η απόσταση μεταξύ δύο ομάδων ορίζεται ως η μεγαλύτερη απόσταση από μια παρατήρηση της μιας ομάδας σε μια παρατήρηση της άλλης ομάδας.
- Μέθοδος των σταθμισμένων μέσων (*Weighted Average Linkage*). Η απόσταση των δύο ομάδων είναι ο μέσος των αποστάσεων όλων των στοιχείων της μιας ομάδας με τα στοιχεία της άλλης.

- Μέθοδος των κέντρων βάρους (*Centroid Method*). Για τη μέθοδο αυτή η απόσταση που υπολογίζεται είναι η ευκλείδια απόσταση μεταξύ των κέντρων βάρους των ομάδων.
- Μέθοδος του Ward. Η λογική αυτή της μεθόδου διαφέρει από τις προηγούμενες αφού στόχος της είναι να ελαχιστοποιήσει την "Στατιστική" διακύμανση μέσα στις ομάδες.
- Μέθοδος του Gower. Είναι παραλλαγή της μεθόδου των κέντρων βάρους έτσι ώστε να μην εξαρτάται από το μέγεθος των ομάδων που ενώνονται.

Λόγω της χρήσης της ευκλείδιας απόστασης, οι τρεις τελευταίες μέθοδοι εφαρμόζονται μόνο σε ποσοτικά δεδομένα.

Μεταξύ των παραπάνω μεθόδων έχει διαπιστωθεί (βλ. Κούτρας, 2007) ότι η καλύτερη ομαδοποίηση προκύπτει από την μέθοδο του Ward και την μέθοδο των σταθμισμένων μέσων, ενώ την χειρότερη επίδοση έχει η μέθοδος του κοντινότερου γείτονα.

3.2.3 Ανάλυση συστάδων σε δύο βήματα

Όπως αναφέραμε και στην εισαγωγή του κεφαλαίου, θα χρησιμοποιήσουμε την ανάλυση συστάδων για δύο σκοπούς: την ανάλυση του προφίλ των πελατών και τον περιορισμό των κατηγορικών μεταβλητών.

Για την ανάλυση του προφίλ θα πρέπει να δουλέψουμε με μεικτά δεδομένα (συνεχείς και κατηγορικές μεταβλητές), οπότε θα καταφύγουμε στην ανάλυση συστάδων σε δύο βήματα (*Two-Step Cluster Analysis*) η οποία είναι η μοναδική που μας επιτρέπει να δουλέψουμε με τέτοιου είδους δεδομένα. Επίσης, και για τον περιορισμό των κατηγορικών θα εφαρμόσουμε την συγκεκριμένη μέθοδο γιατί, αν και έχουμε μόνο κατηγορικές μεταβλητές, το SPSS δεν έχει την δυνατότητα να εφαρμόσει κάποια άλλη μέθοδο σε κατηγορικά δεδομένα.

Η ανάλυση συστάδων σε δύο βήματα παρουσιάζει τα ακόλουθα πλεονεκτήματα έναντι των παραδοσιακών μεθόδων συσταδοποίησης

- Μπορεί να εφαρμοστεί σε μεικτά δεδομένα (συνεχείς και κατηγορικές μεταβλητές),
- Έχει την δυνατότητα να επιλέξει αυτόματα τον αριθμό των συστάδων που θα δημιουργηθούν και
- Μπορεί να αναλύσει αποτελεσματικά μεγάλα σετ δεδομένων.

Το βασικότερο πλεονέκτημά της, που είναι η εφαρμογή σε μεικτά δεδομένα, οφείλεται στην μέτρηση της απόστασης μεταξύ δύο συστάδων, που χρησιμοποιεί. Οι αποστάσεις που έχουμε δει έως τώρα εφαρμόζονται είτε σε συνεχείς μεταβλητές, είτε σε κατηγορικές, αλλά όχι σε συνδιασμό τους. Η καινούρια απόσταση είναι βασισμένη στις πιθανότητες, και σχετίζεται με την αύξηση στον λογάριθμο της πιθανοφάνειας (*log-likelihood*) που προκαλείται από την συνένωση των δύο συστάδων. Για τον υπολογισμό της υποθέτουμε κανονικότητα για τις συνεχείς μεταβλητές, πολυωνυμική κατανομή για τις κατηγορικές και ανεξαρτησία μεταξύ των μεταβλητών και μεταξύ των παρατηρήσεων. (βλ. SPSS, 2001)

Ο υπολογισμός της παραπάνω απόστασης γίνεται με βάση τον τύπο

$$d(j,s) = \xi_j + \xi_s + \xi_{<j,s>}$$

με

$$\xi_i = -n_i\left(\sum_{k=1}^{p_1}\frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{ik}^2) + \sum_{k=1}^{p_2}\hat{E}_{ik}\right) \qquad \textbf{(3.1)}$$

και

$$\hat{E}_{ik} = -\sum_{l=1}^{c_k}\left[\frac{n_{ikl}}{n_i}\log\left(\frac{n_{ikl}}{n_i}\right)\right]$$

όπου

$< j, s >$ η συστάδα που προκύπτει από την ένωση των συστάδων j και s

n_i το πλήθος των παρατηρήσεων που περιέχει η i-οστή συστάδα

n_{ikl} το πλήθος των παρατηρήσεων στην i-οστή συστάδα, που ανήκουν στην l-οστή κατηγορία της k-οστής κατηγορικής μεταβλητής

p_1 ο αριθμός των συνεχών μεταβλητών

p_2 ο αριθμός των κατηγορικών μεταβλητών

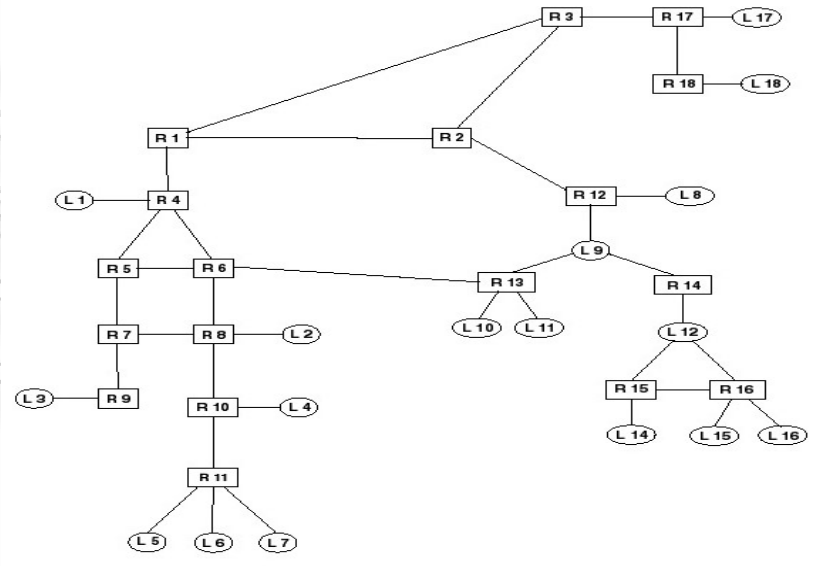
c_k ο αριθμός των κατηγοριών της k-οστής κατηγορικής μεταβλητής

$\hat{\sigma}_k^2$ η εκτίμηση της διασποράς της k-οστής συνεχούς μεταβλητής στο σύνολο των παρατηρήσεων

$\hat{\sigma}_{ik}^2$ η εκτίμηση της διασποράς της k-οστής συνεχούς μεταβλητής στην συστάδα i

Όπως δηλώνει και το όνομα της, η μέθοδος εκτελείται σε δύο στάδια. Στο πρώτο στάδιο (*pre-clustering*) οι παρατηρήσεις ομαδοποιούνται σε πολλές μικρές υποσυστάδες (*sub-clusters*). Στο δεύτερο βήμα, οι υποσυστάδες που δημιουργήθηκαν θεωρούνται ως αρχικές παρατηρήσεις για την συσταδοποίηση στον βέλτιστο αριθμό συστάδων. Αναλυτικότερα,

- **<u>Βήμα 1ο</u>:** Εφαρμόζεται μια ακολουθιακή προσέγγιση της συσταδοποίησης. Κάθε παρατήρηση ελέγχεται, βάση του κριτηρίου της απόστασης, αν μπορεί να ενταχθεί σε κάποια από τις προϋπάρχουσες συστάδες ή αν θα πρέπει να δημιουργήσει μια νέα. Ο αλγόριθμος μπορεί να περιγραφεί με ένα «δένδρο χαρακτηριστικών της συστάδας» (*Cluster Feature tree, CF-tree*) (Σχήμα 3.2) το οποίο αποτελείται από επίπεδα φύλλων-κόμβων (*levels of leaf-nodes*), όπου κάθε κόμβος περιέχει έναν αριθμό εισόδων (*entries*) και κάθε είσοδος περιέχει παρατηρήσεις του συνόλου δεδομένων. Κάθε είσοδος ενός κόμβου αποτελεί μια υποσυστάδα και χαρακτηρίζεται από την πληροφορία των μεταβλητών (αριθμός παρατηρήσεων, μέση τιμή και διασπορά για συνεχείς μεταβλητές, συχνότητα κάθε κατηγορίας για κατηγορικές μεταβλητές) που προκύπτει από τις παρατηρήσεις που την αποτελούν. Το SPSS χρησιμοποιεί δέντρα που επιτρέπουν το πολύ 3 επίπεδα και μέγιστο αριθμό είσόδων 8 ανά κόμβο, οπότε μπορεί να καταλήξει το πολύ σε 512 υποσυστάδες.
- **<u>Βήμα 2ο:</u>** Οι υποσυστάδες που προέκυψαν από το πρώτο βήμα αποτελούν τα αρχικά δεδομένα για την συσταδοποίση στο δεύτερο. Από την στιγμή που ο αριθμός των υποσυστάδων είναι πολύ μικρότερος των αρχικών παρατηρήσεων, μπορούμε να χρησιμοποιήσουμε αποτελεσματικά κάποια από τις παραδοσιακές μεθόδους. Το SPSS χρησιμοποιεί ιεραρχική συσσωρευτική μέθοδο (*hierarchical agglomerative method*) κυρίως εξαιτίας της καλής απόδοσής της στην αυτόματη επιλογή του βέλτιστου αριθμού συστάδων για την οποία θα μιλήσουμε αμέσως παρακάτω.



Σχήμα 3.2: Δέντρο χαρακτηριστικών συστάδας (CF-tree)

Για την εύρεση του αριθμού των συστάδων το SPSS εφαρμόζει μια διαδικασία δύο βημάτων. Στο πρώτο βήμα υπολογίζεται το BIC (*Schwarz's Bayesian Criterion*), σύμφωνα με την σχέση (3.2), για κάθε αριθμό συστάδων εντός συγκεκριμένου εύρους και χρησιμοποιείται για την αρχική εκτίμηση του αριθμού των συστάδων. Στο δεύτερο βήμα τελειοποιείται η αρχική εκτίμηση, με βάση την μεγαλύτερη αλλαγή στην απόσταση μεταξύ των δύο πιο κοντινών συστάδων σε κάθε βήμα της ιεραρχικής συσταδοποίησης.

Το κριτήριο BIC υπολογίζεται από τον τύπο

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(n) \qquad \textbf{(3.2)}$$

με $$m_J = J\{2p_1 + \sum_{k=1}^{p_2}(c_k - 1)\}$$

όπου

ξ_j όπως ορίστηκε από την σχέση (3.1)

J ο αριθμός των συστάδων

n ο αριθμός των συνολικών παρατηρήσεων

p_1, p_2, c_k όπως ορίστηκαν παραπάνω

Για περισσότερες λεπτομέρειες σχετικά με την μέθοδο ο αναγνώστης μπορεί να ανατρέξει στο white paper του SPSS (2001).

3.3 Ανάλυση Παραγόντων

Στη βιβλιογραφία έχουν καταγραφεί διάφορες μέθοδοι για την μείωση της διάστασης των συνεχών μεταβλητών. Χαρακτηριστικότερες είναι η ανάλυση παραγόντων (*Factor Analysis-FA*),(West,1985; Piramuthu et al., 1998), η ανάλυση κύριων συνιστωσών (*Principal Component Analysis-PCA*) (Kemsley,1996; Dai et al., 2006) και η μέθοδος των μερικών ελαχίστων τετραγώνων (*Partial Least Squares-PLS*), (Kemsley, 1996; Dai et al.,2006). Για περιγραφή άλλων μεθόδων παραπέμπουμε στους Maimon and Rokach (2005).

Η ανάλυση παραγόντων είναι ευρέως διαδεδομένη μέθοδος σε αντίστοιχες εφαρμογές με αυτήν που εξετάζουμε (π.χ., Ang and Willhour, 1976; Paxton, 1996; Chen et al., 2009; Gnanadhas and Geetha, 2009) και αυτήν θα χρησιμοποιήσουμε στην ανάλυση μας.

Βασική ιδέα της FA είναι να ελαττώσει την διάσταση ενός συνόλου δεδομένων που περιέχει ενα μεγάλο αριθμό συσχετισμένων μεταβλητών, δημιουργώντας ένα μικρότερο

σύνολο νέων ασυσχέτιστων μεταβλητών, που ονομάζονται παράγοντες. Οι παράγοντες αυτοί μπορούν να θεωρηθούν ως υφιστάμενα χαρακτηριστικά που δεν μπορούν να περιγραφούν με μια μόνο μεταβλητή. Οι μεταβλητές που χρησιμοποιούνται για την FA θα πρέπει να είναι υψηλά συσχετισμένες μεταξύ τους, διαφορετικά ο αριθμός των παραγόντων θα είναι σχεδόν ίδιος με τον αριθμό των αρχικών μεταβλητών, οπότε η εφαρμογή της μεθόδου δεν θα έχει καμία χρησιμότητα.

Το μοντέλο της FA μπορεί να περιγραφεί ως ακολούθως. Έστω ότι έχουμε *p* μεταβλητές, τις $X_1, X_2, ..., X_p$ των οποίων τις τιμές γνωρίζουμε για ένα δείγμα μεγέθους *n*. Κάθε μεταβλητή X_i μπορεί να γραφεί μέσω της FA ως γραμμικός συνδυασμός των *m* ασυσχέτιστων παραγόντων $F_1, F_2, ..., F_m$, όπου *m* < *p* σύμφωνα με τον τύπο:

$$X_i = \ell_{i1}F_1 + \ell_{i2}F_2 + ... + \ell_{im}F_m + u_i$$

Ο συντελεστής ℓ_{ij} ονομάζεται φορτίο (*loading*) ή σκορ (*score*) της μεταβλητής *i* στον παράγοντα *j*, ενώ το u_i είναι το σφάλμα (*error*) και απεικονίζει την μεταβλητότητα της X_i που δεν μπορεί να εξηγηθεί από τους παραπάνω παράγοντες.

Ο υπολογισμός των παραγόντων και των φορτίων μπορεί να γίνει με διάφορες μεθόδους όπως οι κύριες συνιστώσες (*principal components*), η μέγιστη πιθανοφάνεια (*maximum likelihood*) και οι κύριοι άξονες (*principal axis factoring*).

Η μέθοδος των κυρίων συνιστωσών ξεκινάει βρίσκοντας ένα γραμμικό συνδυασμό των αρχικών μεταβλητών (μια συνιστώσα), που αντιπροσωπεύει όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης τους. Ακολούθως, βρίσκει μια δεύτερη συνιστώσα η οποία να εξηγεί όσο το δυνατόν περισσότερη από την εναπομείνουσα μεταβλητότητα και να είναι ασυσχέτιστη με την προηγούμενη συνιστώσα. Συνεχίζοντας με τον ίδο τρόπο υπολογίζει τόσες ασυσχέτιστες συνιστώσες όσες και ο αριθμός των αρχικών μεταβλητών. Συνήθως ένας μικρός αριθμός συνιστωσών απεικονίζει το μεγαλύτερο μέρος της αρχικής συνολικής μεταβλητότητας, και αυτές οι συνιστώσες επαρκούν για την περιγραφή των αρχικών μεταβλητών, τις οποίες και αντικαθιστούν στην πράξη.

Στην πραγματικότητα το παραπάνω αποτελεί μοντελοποίηση του πίνακα συνδιακυμάνσεων Cov(**X**)=**Σ**, μέσω της σχέσης $Cov(X_i, F_j) = \ell_{ij}$. Αντίστοιχα μπορεί να εφαρμοστεί και στον πίνακα συσχετίσεων Cor(**X**)=**R**. (βλ. Ηλιόπουλος, 2007). Αν $(l_i, \vec{e}_i)_{i=1,...,p}$ είναι τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του αντίστοιχου πίνακα που

αναλύουμε και ισχύει $l_1 > ... > l_p$ τότε ο πίνακας των φορτίων $\underset{p\times m}{L} = (\ell_{ij})$ δίνεται από τον τύπο $(\ell_{ij}) = \left(\sqrt{l_1}\vec{e}_1 ... \sqrt{l_m}\vec{e}_m\right)$.

Ο αριθμός των παραγόντων που τελικά θα χρησιμοποιηθούν στην θέση των αρχικών μεταβλητών μπορεί να προσδιοριστεί με ένα από τους παρακάτω τρόπους

- Επιλέγουμε τόσους παράγοντες όσα τα ιδιοδιανύσματα με ιδιοτιμή μεγαλύτερη του 1 (αν χρησιμοποιούμε τον πίνακα των συσχετίσεων (*correlation matrix*))
- Διαλέγουμε τόσους παράγοντες ώστε να εξηγείται ένα συγκεκριμένο ποσοστό μεταβλητότητας (π.χ. 80%)
- Με βάση το scree plot των ιδιοτίμων, το οποίο δείχνει κατά πόσο υπάρχει προφανές σημείο στο οποίο να διαχωρίζονται οι μεγάλες από τις μικρές ιδιοτιμές, ο αριθμός των παραγόντων που κρατάμε είναι ίσος με τον αριθμό των μεγάλων ιδιοτιμών.

Αφού υπολογιστούν οι παράγοντες, περιστρέφονται για να είναι ευκολότερο να ερμηνευτούν. Αν υπάρχουν κάποιες «ομάδες» μεταβλητών (με ισχυρή συσχέτιση μέσα σε κάθε ομάδα), η περιστροφή προσπαθεί για τις μεταβλητές της κάθε υποομάδας να δημιουργήσει όσο το δυνατόν μεγαλύτερα φορτία σε ένα συγκεκριμένο παράγοντα και ταυτόχρονα να εξασφαλίσει ότι τα φορτία τους στους άλλους παράγοντες θα είναι όσο το δυνατό μικρότερα. Ουσιαστικά, στόχος της περιστροφής είναι όλες οι μεταβλητές να έχουν υψηλό φορτίο σε ένα μόνο παράγοντα.

Υπάρχουν δύο είδη περιστροφής, η ορθογώνια (*orthogonal*) και η πλάγια (*oblique*). Με την ορθογώνια περιστροφή οι παράγοντες παραμένουν ασυσχέτιστοι ενώ με την πλάγια οι νέοι παράγοντες θα είναι πλέον συσχετισμένοι (βλ. Ηλιόπουλος, 2007).

Για εκτενέστερη αναφορά στην μέθοδο ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο των Rummel (1979) και Johnson and Wichern (1998).

3.4 Μη ισορροπημένα δεδομένα

Το πρόβλημα της μη ισορροπίας των δεδομένων προκύπτει όταν σε ένα πρόβλημα ταξινόμησης ο αριθμός των παρατηρήσεων στη μια ομάδα είναι σημαντικά μεγαλύτερος από ότι στην άλλη (Maloof, 2003; Chawla et al., 2004). Στην περίπτωση αυτή ο αλγόριθμος ταξινόμησης τείνει να κατατάσσει όλες τις παρατηρήσεις στην μεγαλύτερη ομάδα (*majority*

class) (Kotsiantis et al., 2006). Συνήθως όμως, η μικρότερη ομάδα (*minority class*) είναι αυτή που παρουσιάζει το μεγαλύτερο ενδιαφέρον.

Το παραπάνω φαινόμενο είναι διαδεδομένο σε εφαρμογές όπως η ανίχνευση απάτης (*fraud detection*), η διοίκηση κινδύνου (*risk management*), η ταξινόμηση κειμένου, οι ιατρικές διαγνώσεις καθώς και πολλές άλλες (Chawla et al., 2002; Maimon and Rokach και αναφορές).

Στη βιβλιογραφία προτείνονται διάφοροι τρόποι επίλυσης, οι οποίοι λειτουργούν σε δύο διαφορετικά επίπεδα: είτε αφορούν την αναπροσαρμογή των δεδομένων (*data level*) είτε αναφέρονται στην τροποποίηση του αλγορίθμου (*algorithm level*) (Chawla et al., 2004; Hulse et al., 2007). Στην παρούσα εργασία θα αναφερθούμε μόνο στους πρώτους.

Οι μέθοδοι που χρησιμοποιούνται για την εξισορρόπηση των δεδομένων μπορούν να χωριστούν σε τρεις βασικές κατηγορίες (Liu et al., 2006):

- Μείωση του πλήθους των στοιχείων της πλειοψηφούσας ομάδας (*under-sampling*).
- Αύξηση του πλήθους των στοιχείων της μειοψηφούσας ομάδας (*over-sampling*).
- Συνδιασμός των δύο προηγούμενων.

Για κάθεμια από τις παραπάνω κατηγορίες μεθόδων έχουν αναπτυχθεί αρκετοί εναλλακτικοί τρόποι υλοποίησης, με τυχαία ή στοχευμένη/επιλεκτική δειγματοληψία. Ο απλούστερος τρόπος είναι η τυχαία δειγματοληψία.

Όμως, με την τυχαία μείωση του δείγματος (*random under-sampling*) χάνεται μέρος της διαθέσιμης πληροφορίας το οποίο είναι σημαντικό για την διαδικασία εκπαίδευσης του ταξινομητή. Από την άλλη, η αύξηση του δείγματος με τυχαίο τρόπο (*random over-sampling*) αφενός μπορεί να οδηγήσει σε υπερ-προσαρμογή (*overfitting*) του μοντέλου ταξινόμησης, αφού δημιουργούνται ακριβή αντίτυπα των στοιχείων της μειοψηφούσας ομάδας και αφετέρου αυξάνει το υπολογιστικό κόστος σε ήδη μεγάλα δείγματα.(Chawla et al., 2004; Batista et al., 2004). Το πρόβλημα της υπερ-προσαρμογής προκύπτει όταν το επιλεγμένο μοντέλο ερμηνεύει τον τυχαίο θόρυβο των δεδομένων, τόσο καλά όσο την πραγματική μεταβλητότητα (Banks and Feinberg, 2002).

Για περισσότερες λεπτομέρειες και εφαρμογές ο αναγνώστης μπορεί να ανατρέξει στα άρθρα των Joshi et al. (2001), Japkowicz and Stephen (2002), Chawla et al. (2002), Chawla et

al. (2004), Batista et al. (2004), Raskutti and Kowalczyk (2004), Liu et al. (2006), Kotsiantis et al. (2006), Hulse et al. (2007), καθώς και στο βιβλίο των Maimon and Rokach (2005).

Σύμφωνα με τους Hulse et al., (2007) κάθε τύπος δειγματοληψίας είναι καταλληλότερος για διαφορετικούς ταξινομητές. Η μεγέθυνση της μικρότερης ομάδας με τυχαία δειγματοληψία (*random over-sampling*) είναι αποτελεσματικότερη σε συνδυασμό με την εφαρμογή της λογιστικής παλινδρόμησης (*logistic regression*).

Επομένως, στην εφαρμογή μας θα χρησιμοποιήσουμε την τυχαία δειγματοληψία με επανάθεση (*bootstrap*) στο σύνολο των αναξιόπιστων πελατών, για να δεκαπλασιάσουμε το μέγεθος του.

ΚΕΦΑΛΑΙΟ 4

Προφίλ Πελάτων και Μείωση της Διάστασης των Δεδομένων

Στο κεφάλαιο 3 περιγράψαμε την ανάλυση συστάδων και την ανάλυση παραγόντων, τις οποίες θα εφαρμόσουμε για την περιγραφή των πελατών και την μείωση της διάστασης των μεταβλητών.

4.1 Περιγραφή πελατών

Με την μέθοδο της συσταδοποίησης θα σκιαγραφήσουμε την εικόνα των πελατών και θα τους ομαδοποιήσουμε με βάση τα χαρακτηριστικά τους ως προς τις διαθέσιμες μεταβλητές.

Η ανάλυση θα πραγματοποιηθεί μέσω της συσταδοποίησης σε δύο βήματα (*Two Step Cluster Analysis*), η οποία μπορεί να εφαρμοστεί σε μεικτά δεδομένα, όπως είδαμε στο προηγούμενο κεφάλαιο (§ 3.2.3)

Η παραπάνω ανάλυση οδηγεί σε δύο ομάδες (*clusters*) όπου η πρώτη αποτελείται από το 70% των πελατών και η δεύτερη από το 30%. Σύμφωνα με τον Πίνακα 4.1 η δεύτερη ομάδα περιλαμβάνει διπλάσιο ποσοστό αναξιόπιστων πελατών (8%) έναντι της πρώτης (4%). Οπότε πελάτες με χαρακτηριστικά όμοια με εκείνα των πελατών της δεύτερης ομάδας ενέχουν πολύ μεγαλύτερο κίνδυνο (*risk*) για τα συμφέροντα της τράπεζας και χρείζουν ιδιαίτερης ανάλυσης και διαχείρισης. Έτσι θα επικεντρωθούμε στην περιγραφή των πελατών αυτών.

		RISK		Total	% Risky clients
		0	1		
Cluster	1	44.500	1.819	**46.319**	**3,9%**
	2	18.243	1.599	**19.842**	**8,1%**
Total		**62.743**	**3.418**	**66.161**	5,2%

Πίνακας 4.1: Αποτελέσματα της two-step cluster analysis

Παρατηρώντας τις μέσες τιμές των συνεχών μεταβλητών για κάθε συστάδα (Πίνακας 4.2), βλέπουμε ότι οι πελάτες με το μεγαλύτερο ρίσκο φαίνεται να έχουν παλαιότερα και περισσότερα καταναλωτικά προϊόντα και συνεπώς μεγαλύτερη συνολική οφειλή και μεγαλύτερα πιστωτικά όρια. Ακόμα διαθέτουν μεγαλύτερο εισόδημα και περισσότερα χρόνια διαμονής στην ίδια διεύθυνση και στην εργασία τους.

	Cluster 1	2	Total	C2 vs Total	C2 vs C1
# δανείων	1,16	1,89	1,38	37%	62%*
Πιστ. Οριο δανείου	10.454	9.450	10.153	-7%	-10%*
Πιστ. Οριο πελάτη	11.501	17.357	13.257	31%	51%*
καθυστερημένη δόση	158	296	199	49%	88%*
Οφειλή πελάτη	8.861	15.518	10.857	43%	75%*
Ύψος δανείου	8.022	8.460	8.153	4%	5%*
Παλαιότητα δανείου (μήνες)	26	54	34	58%	109%*
Εισόδημα	12.305	17.446	13.847	26%	42%*
Έτη στην διεύθυνση	10	13	11	17%	26%*
Έτη στην εργασία	6	9	7	31%	51%*
Ηλικία	43	45	44	4%	5%*

*Στατιστικά σημαντική διαφορά σε επίπεδο σημαντικότητας 5% (βλ. Appendix 4.1: Πίνακας ANOVA)

Πίνακας 4.2: Μέσες τιμές των συστάδων

Όμως, για τις κατηγορικές μεταβλητές διαπιστώνουμε ότι οι συχνότητες εμφάνισης των διάφορων τιμών τους δεν διαφέρουν ιδιαίτερα μεταξύ των συστάδων και του γενικού πληθυσμού. Εξαίρεση στα παραπάνω αποτελεί η μεταβλητή "BL". Ενώ η συχνότητα των πελατών με επαγγελματικό δάνειο σε καθυστέρηση είναι της τάξης του 3,3% στο συνολικό δείγμα, στην ομάδα 2 το ποσοστό αυτό περίπου τριπλασιάζεται (9%). Έτσι η μεταβλητή αυτή είναι που παίζει τον κυρίαρχο ρόλο και καθορίζει την ομαδοποίηση, πρακτικά μη αφήνοντας τις άλλες μεταβλητές να επιδράσουν, όπως θα φανεί και στη συνέχεια.

Η εντονότερη παρουσία των πελατών με επαγγελματικό δάνειο σε καθυστέρηση στην πιο «επικίνδυνη ομάδα» δηλώνει ότι τα καταναλωτικά τους δάνεια έχουν μεγαλύτερη πιθανότητα να καταλήξουν μη-εξυπηρετούμενα. Αυτό επιβεβαιώνεται και από το γεγονός ότι ενώ στο συνολικό δείγμα οι αναξιόπιστοι πελάτες (Risk=1) αποτελούν το 5%, στο σύνολο των πελατών που δεν έχουν πλήρώσει τουλάχιστον μια δόση στο επαγγελματικό τους δάνειο το ποσοστό αυτό ανέρχεται στο 8% (Πίνακας 4.3), δηλαδή 38% μεγαλύτερο.

		RISK 0	1	Total	% Risky Clients
BL	**0**	60.746	3.240	**63.986**	**5,1%**
	1	1.997	178	**2.175**	**8,2%**
Total		**62.743**	**3.418**	**66.161**	5,2%

Πίνακας 4.3: Συνάφεια των μεταβλητών "BL" – "RISK"

Αφού η επίδραση της "BL" φαίνεται να είναι τόσο ισχυρή, θα εφαρμόσουμε την μέθοδο της συσταδοποίησης για καθένα από τα επίπεδα της μεταβλητής. Έτσι θα μπορέσουμε να εντοπίσουμε τυχόν διαφορετικά χαρακτηριστικά και ομαδοποιήσεις που προκύπτουν στις δύο κατηγορίες πελατών.

Πράγματι, η συσταδοποίηση σε δύο στάδια ομαδοποιεί τους πελάτες που δεν έχουν επαγγελματικό δάνειο σε καθυστέρηση σε έξι συστάδες αντί για δύο. Η πρώτη ομάδα της

αρχικής ανάλυσης έχει διασπαστεί στις νέες συστάδες 1-4 και η δεύτερη στις 5-6 (Πίνακας 4.4). Τα παρακάτω ποσοστά περαμένουν σχεδόν αμετάβλητα αν τα κοιτάξουμε στα επιμέρους σύνολα των αξιόπιστων και των αναξιόπιστων πελατών.

		Cluster (all BL)				Total	
		1	%	2	%		
Cluster (BL = 0)	1	12.266	27%	234	1%	12.500	20%
	2	8.944	19%	301	2%	9.245	14%
	3	12.061	26%	26	0%	12.087	19%
	4	9.819	21%	93	1%	9.912	15%
	5	2.846	6%	2.979	17%	5.825	9%
	6	0	0%	14.417	80%	14.417	23%
	Total	45.936	100%	18.050	100%	63.986	100%

Πίνακας 4.4: Συσταδοποίηση των πελατών με "BL"=0

Άρα έχοντας σταθεροποιήσει την τιμή της "BL", εμφανίζονται οι επιδράσεις και των υπόλοιπων μεταβλητών τις οποίες επισκίαζε η επίδραση της "BL", με αποτέλεσμα να έχουμε το περαιτέρω σπάσιμο των ομάδων.

Από τις έξι συστάδες που δημιουργήθηκαν, αυτές που περιέχουν πελάτες με μεγαλύτερη πιθανότητα να μην ανταποκριθούν στις απαιτήσεις του καταναλωτικού τους δανείου, είναι η 5 και η 6, τα χαρακτηριστικά των οποίων θα περιγράψουμε αμέσως παρακάτω, αφού σε αυτές το ποσοστό των αναξιόπιστων πελατών είναι 10% και 8% αντίστοιχα, έναντι 3,6% που είναι κατά μεσο όρο στις υπόλοιπες συστάδες (Πίνακας 4.5). Άλλωστε αυτό ήταν αναμενόμενο καθώς όπως είδαμε οι δύο αυτές συστάδες αντιστοιχούν στην αρχική ομάδα 2 που είναι η πιο «επικίνδυνη».

		RISK		Total	% Risky Clients
		0	1		
Cluster (BL = 0)	1	12.118	382	12.500	3%
	2	8.800	445	9.245	5%
	3	11.681	406	12.087	3%
	4	9.587	325	9.912	3%
	5	5.252	573	5.825	10%
	6	13.308	1.109	14.417	8%
	Total	60.746	3.240	63.986	5%

⇨ C2 (all BL)

Πίνακας 4.5: Κατανομή πελατών & ρίσκο για τις 6 συστάδες της ανάλυσης για "BL=0"

Ας δούμε τώρα ποια είναι η μέση εικόνα (*profile*) των πελατών με το μεγαλύτερο ρίσκο. Στις δύο «επικίνδυνες» ομάδες η μέση ηλικία των πελατών είναι 45 και τα έτη παραμονής στην διεύθυνση 12 και στην εργασία 9. Οι παραπάνω τιμές κυμαίνονται στο γενικό μέσο όρο των πελατών χωρίς επαγγελματικό δάνειο σε καθυστέρηση.

Στο σύνολο των πελατών με “BL=0” ο μέσος αριθμός καταναλωτικών δανείων που κατέχει κάθε πελάτης είναι 1,4, το συνολικό πιστωτικό όριο είναι €13.000, το συνολικό οφειλόμενο ποσό ανέρχεται σε €11.000 και το εισόδημα σε €14.000.

Στην συστάδα 5 οι πελάτες έχουν κατά μέσο όρο 1,5 καταναλωτικά προϊόντα, συνολικό πιστωτικό όριο €30.000, συνολικό οφειλόμενο ποσό €27.000 και εισόδημα €23.000. Επίσης οφείλουν 1 δόση με μέσο ποσό €360 στο τοκοχρεωλυτικό τους δάνειο το οποίο είναι ύψους €25.000 και παλαιότητας 1,5 έτους.

Επίσης, υπάρχει η παρουσία ορισμένων επαγγελμάτων σε μεγαλύτερο ποσοστό από ότι στο γενικό σύνολο, όπως δημόσιοι υπάλληλοι, ελεύθεροι επαγγελματίες, έμποροι, επιχειρηματίες, βιοτέχνες και ιατρικό προσωπικό. Διαφοροποιείται ακόμα η οικογενειακή κατάσταση των πελατών αυτών αφού είναι στο 73,5% έγγαμοι και στο 8,4% διαζευγμένοι, έναντι 59% και 6% που είναι ο μέσος όρος καθώς και το φύλο αφού οι άνδρες αποτελούν το 65% της ομάδας έναντι 62% του συνόλου.

Η ομάδα 6 αποτελείται από πελάτες που έχουν κατά μέσο όρο 2 καταναλωτικά προϊόντα, συνολικό πιστωτικό όριο €14.000, συνολικό οφειλόμενο ποσό €13.000 και εισόδημα €14.000. Ακόμα οφείλουν 1 δόση με μέσο ποσό €270 στην πιστωτική τους κάρτα, η οποία έχει πιστωτικό όριο €6.000 και παλαιότητα 5 έτη.

Όμοια με προηγουμένως, κι εδώ οι ελεύθεροι επαγγελματίες, οι έμποροι, οι επιχειρηματίες, οι βιοτέχνες και το ιατρικό προσωπικό αποτελούν μεγαλύτερο τμήμα της ομάδας σε σχέση με το σύνολο των πελατών χωρίς καθυστερημένο επαγγελματικό δάνειο. Επίσης η παρουσία των έγγαμων αλλά και των γυναικών είναι ενισχυμένη στην συστάδα αυτή.

Η ομάδα 5 λοιπόν περιέχει πελάτες με μεγάλες οφειλές, κυρίως σε τοκοχρεωλυτικά δάνεια, αλλά και υψηλά εισοδήματα, οι οποίοι στην πλειοψηφία τους είναι άνδρες έγγαμοι ή διαζευγμένοι. Αντίθετα, στην συστάδα 6 έχουμε πελάτες, κυρίως έγγαμες γυναίκες, με χαμηλό εισόδημα και μικρού μεγέθους δάνεια, κατά βάση πιστωτικές κάρτες, μεγαλύτερης παλαιότητας.

Επιπροσθέτως, από τον Πίνακα 4.5 διαπιστώνουμε ότι υπάρχει διαβάθμιση του ρίσκου μεταξύ των συστάδων που έχουν δημιουργηθεί. Οι συστάδες 5 και 6 ενέχουν το μεγαλύτερο ρίσκο (*High Risk*), όπως ήδη αναλύσαμε, η 2 μπορεί να χαρακτηριστεί μεσαίου κινδύνου (*Medium Risk*) αφού το ποσοστό των αναξιόπιστων πελατών είναι ίσο με το μέσο όρο του συνόλου για “BL”=0, ενώ για τις υπόλοιπες ομάδες (1, 3 και 4) το επίπεδο «επικινδυνότητας»

είναι 3%, δηλαδή 40% χαμηλότερο από το μέσο, οπότε μπορούμε να τις θεωρήσουμε χαμηλού ρίσκου (*Low Risk*).

Το κυριότερο χαρακτηριστικό της συστάδας 2 είναι ότι περιέχει στην συντριπτική πλειοψηφία δάνεια ανοιχτού τύπου. Όλα τα υπόλοιπα χαρακτηριστικά της κυμαίνονται στον μέσο όρο του συνόλου "BL"=0.

Έως τώρα αναλύσαμε τα χαρακτηριστικά των πελατών που δεν έχουν καθυστερημένο επαγγελματικό δάνειο. Στη συνέχεια θα επαναλάβουμε την παραπάνω ανάλυση και για τους πελάτες που έχουν καθυστερήσει τουλάχιστον μια δόση του επαγγελματικού τους δανείου στην συγκεκριμένη τράπεζα ("BL" = 1). Όπως προαναφέρθηκε αυτή η κατηγορία των πελατών έχει αυξημένο ρίσκο (8,2%).

Η συσταδοποίηση δύο βημάτων οδηγεί στην δημιουργία δύο ομάδων. Η πρώτη ομάδα αποτελείται από ολόκληρη την πρώτη ομάδα της αρχικής ανάλυσης (all BL) συν επιπλέον κάποιους πελάτες από την δεύτερη (Πίνακας 4.6). Η δεύτερη ομάδα περιέχει εξ' ολοκλήρου πελάτες που στην αρχική ανάλυση είχαν ενταχθεί επίσης στην δεύτερη ομάδα που είναι αυτή με το μεγαλύτερο ρίσκο. Αυτό έχει ως αποτέλεσμα η ομάδα με το μεγαλύτερο ποσοστό αναξιόπιστων πελατών (10%) να είναι και πάλι η δεύτερη (Πίνακας 4.7).

Οι παραπάνω αναλογίες για τις ομάδες των δύο αναλύσεων διατηρούνται και στα επιμέρους σύνολα τών αξιόπιστων και αναξιόπιστων πελατών.

		Cluster (all BL)				Total	
		1	%	2	%		
Cluster (BL = 1)	1	383	100%	526	29%	909	42%
	2	0	0%	1.266	71%	1.266	58%
Total		**383**	**100%**	**1.792**	**100%**	**2.175**	100 %

Πίνακας 4.6: Συσταδοποίηση για "BL=1"

		RISK 0	RISK 1	Total	Risky Clients
Cluster (BL = 1)	1	855	54	909	**6%**
	2	1.142	124	1.266	**10%**
Total		**1.997**	**178**	**2.175**	**8%**

Πίνακας 4.7: Κατανομή πελατών για "BL"=1

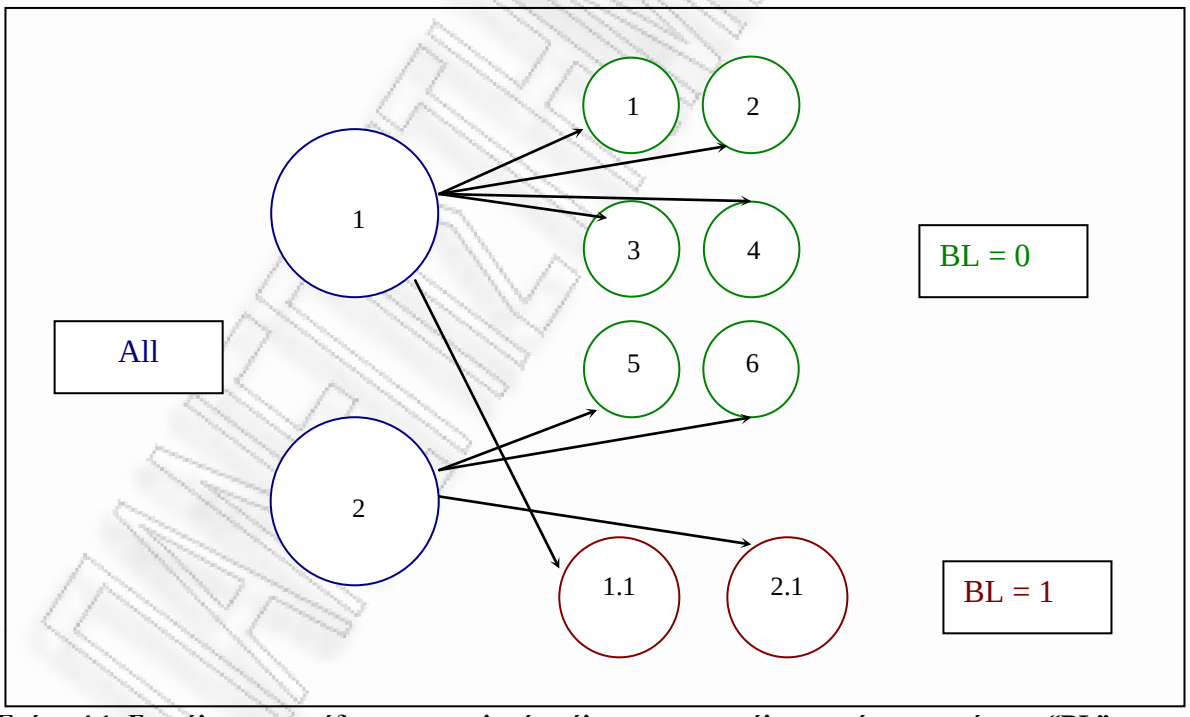
Οι πελάτες (με BL = 1) με την μεγαλύτερη πιθανότητα το δάνειο τους να καταλήξει μη-εξυπηρετούμενο μπορούν να περιγραφούν ως οι πελάτες που οφείλουν μια δόση €320 στην πιστωτική τους κάρτα, η οποία έχει πιστωτικό όριο €7.000 και παλαιότητα λίγο μεγαλύτερη των 5 ετών. Επίσης το εισόδημα τους είναι περίπου €20.000 και η συνολική τους οφειλή ανέρχεται σε €13.000 και αφορά σε 1,8 καταναλωτικά δάνεια, η μέση ηλικία τους είναι 46

και διαμένουν στην τρέχουσα διεύθυνση 14 χρόνια και στην εργασία 11, τιμές που δεν διαφοροποιούνται από το μέσο όρο για το σύνολο των πελατών με καθυστερημένο επαγγελματικό δάνειο (“BL” =1).

Συνοψίζοντας τα αποτελέσματα των δύο τρόπων ανάλυσης των πελατών, για το σύνολο τους αφενός και αφετέρου κοιτάζοντας τους χωριστά ανάλογα με το αν έχουν επαγγελματικό δάνειο σε καθυστέρηση ή όχι, διαπιστώνουμε ότι και στις δύο περιπτώσεις δημιουργούνται ομάδες πελατών με αυξημένο κίνδυνο να πάψουν να ανταποκρίνονται στις υποχρεώσεις του καταναλωτικού τους δανείου.

Στα χαρακτηριστικά των πελατών αυτών η τράπεζα δείχνει ιδιαίτερο ενδιαφέρον. Επιδιώκοντας να αποτρέψει τα αρνητικά για αυτήν αποτελέσματα, διαμορφώνει ανάλογα τις στρατηγικές της για τους υπάρχοντες πελάτες αλλά και τα κριτήρια της για υποψήφιους.

Στο Σχήμα 4.1 φαίνεται η σχέση των ομάδων που διαμορφώνονται στις δύο αναλύσεις. Η ομάδα 1 του γενικού συνόλου διασπάται στις 1.0, 2.0, 3.0 και 4.0 (πελάτες με “BL”=0) και στην 1.1 (πελάτες με “BL”=1). Αντίστοιχα, η $2^{η}$ συστάδα της γενικής ανάλυσης μοιράζεται στις 5.0, 6.0 (πελάτες με “BL”=0) και 2.1 (πελάτες με “BL”=1).



Σχήμα 4.1: Επικάλυψη συστάδων για συνολική ανάλυση κατην ανάλυση ανά κατηγορία της “BL”

4.2 Εντοπισμός κύριων κατηγορικών μεταβλητών μέσω ανάλυσης συστάδων

Εφαρμόζουμε την ανάλυση συστάδων σε δύο βήματα, μόνο για τις κατηγορικές μας μεταβλητές, για τα δύο διαφορετικά επίπεδα της μεταβλητής "BL" της οποίας την εξαιρετική σημασία διαπιστώσαμε στην προηγούμενη παράγραφο. Από τα αποτελέσματα προκύπτει ότι για κάποιες από τις μεταβλητές μας η συχνότητα των τιμών τους διαφέρει σημαντικά μεταξύ των συστάδων που δημιουργούνται. Ουσιαστικά, αυτές είναι και οι μεταβλητές που παίζουν ιδιαίτερο ρόλο στην ομαδοποίηση που εξάγεται.

Όμως, όπως αναφέρεται και στο προηγούμενο κεφάλαιο (§ 3.1), η ύπαρξη πολλών μεταβλητών δημιουργεί προβλήματα στην εφαρμογή των μεθόδων ταξινόμησης. Το γεγονός αυτό δημιουργεί την ανάγκη μείωσης της διάστασης του χώρου των μεταβλητών. Έτσι λοιπόν θα περιοριστούμε στις μεταβλητές εκείνες, των οποίων τον εξέχοντα ρόλο έχουμε ήδη εντοπίσει.

Σύμφωνα λοιπόν με τα αποτελέσματα που προέκυψαν, στην διαμόρφωση των ομάδων (και στα δύο επίπεδα ανάλυσης) επιδρά ο τύπος του καταναλωτικού δανείου στο οποίο ο πελάτης οφείλει μια δόση ("product type"), η οικογενειακή του κατάσταση ("family_status"), ο τύπος του επαγγέλματος του ("occupation") καθώς και το φύλο του ("gender"). Επίσης έχουμε ήδη επισημάνει την σημαντικότητα της ύπαρξης ή όχι επαγγελματικού δανείου σε καθυστέρηση (μεταβλητή "BL").

Άρα για το μοντέλο ταξινόμησης που θα εφαρμόσουμε στη συνέχεια (Κεφ. 5) θα χρησιμοποιήσουμε αυτές τις κατηγορικές μεταβλητές.

4.3 Μείωση διάστασης των συνεχών μεταβλητών μέσω της FA

Για την εφαρμογή της FA θα πρέπει πρώτα να τυποποιήσουμε τις μεταβλητές μας καθώς η κλίμακα των τιμών διαφέρει κατά πολύ μεταξύ των διάφορων μεταβλητών (π.χ. acc:1-5, income:500-200.000) (Rummel, 1979).

Πριν προχωρήσουμε όμως στην ανάλυση, εξετάζουμε κατά πόσο τα δεδομένα είναι κατάλληλα για την εφαρμογή της FA. Μέσω του τεστ KMO διαπιστώνουμε ότι περισσότερο από το 50% της συνολικής μεταβλητότητας είναι κοινή, δηλαδή θα μπορούσε να οφείλεται στους παράγοντες. Επίσης από το τεστ σφαιρικότητας του Barlett φαίνεται ότι υπάρχουν

σημαντικές συσχετίσεις μεταξύ των μεταβλητών (*p-value*=0,00). (Πίνακας 4.8) Επομένως έχει νόημα να χρησιμοποιηθεί η ανάλυση παραγόντων.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**,511**
Bartlett's Test of Sphericity	Approx. Chi-Square	559008,570
	df	55
	Sig.	**,000**

Πίνακας 4.8: Έλεγχος καταλληλότητας της FA

Εφαρμόζουμε την FA με την μέθοδο των κυρίων συνιστωσών, στον πίνακα των συσχετίσεων. Ο αριθμός των παραγόντων που θα επιλεγούν βασίζεται στο πρώτο από τα κριτήρια που προαναφέρθηκαν, δηλαδή είναι ίσος με τον αριθμό των ιδιοτιμών με τιμή >1. Για την περιστροφή των παραγόντων χρησιμοποιούμε την ορθογώνια μέθοδο Varimax Kaiser η οποία είναι η πλέον συνηθισμένη (Ηλιόπουλος, 2007).

Με βάση τα αποτελέσματα που εξάγονται (Πίνακας 4.9), προκύπτουν 4 ασυσχέτιστοι παράγοντες οι οποίοι ερμηνεύουν το 72,4% της αρχικής μεταβλητότητας.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,731	33,918	33,918	3,731	33,918	33,918	3,553	32,301	32,301
2	1,819	16,533	50,451	1,819	16,533	50,451	1,679	15,260	47,561
3	1,383	12,575	63,026	1,383	12,575	63,026	1,571	14,278	61,839
4	1,035	9,412	72,438	1,035	9,412	**72,438**	1,166	10,599	**72,438**
5	,867	7,879	80,317						
6	,735	6,686	87,002						
7	,604	5,491	92,494						
8	,552	5,020	97,514						
9	,150	1,366	98,880						
10	,117	1,067	99,947						
11	,006	,053	100,000						

Extraction Method: Principal Component Analysis.

Πίνακας 4.9: Αποτελέσματα της FA

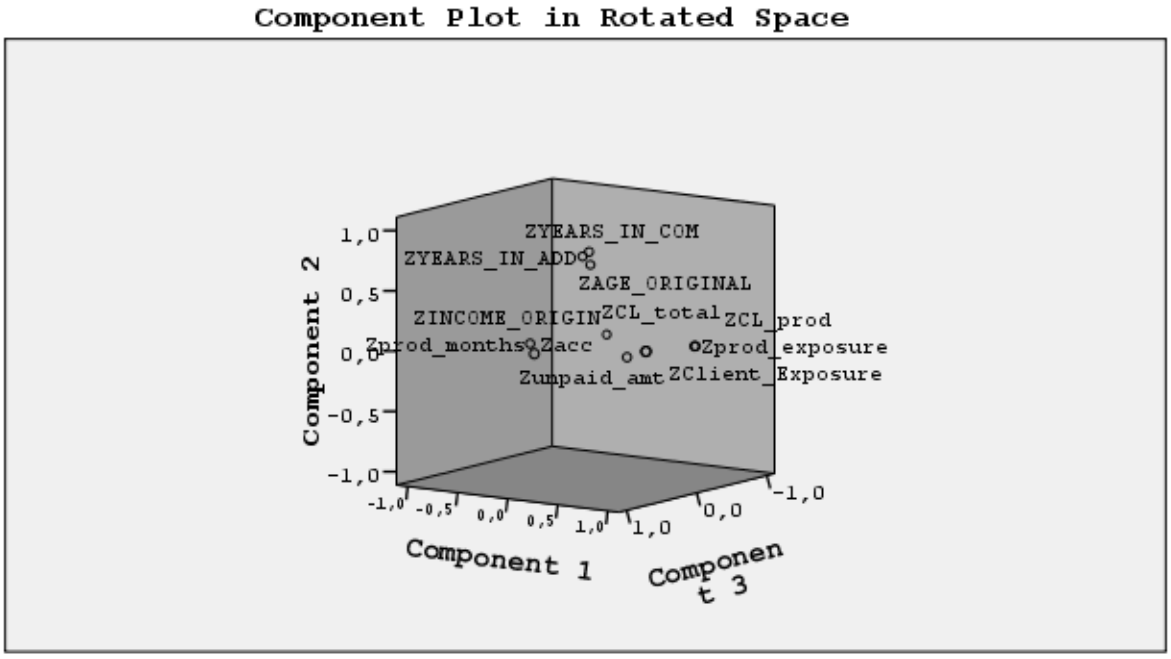
Μετά και την περιστροφή των παραγόντων, η απεικόνιση των τυποποιημένων μεταβλητών σε σχέση με τους παράγοντες φαίνεται στο Σχήμα 4.2 και τα αντίστοιχα φορτία στον Πίνακα 4.10. Από τα φορτία των μεταβλητών στους παράγοντες παρατηρούμε ότι σχηματίζονται 4 ομάδες μεταβλητών και κάθε ομάδα έχει υψηλά φορτία σε ένα από τους παράγοντες.

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Zscore(acc)	,133	,076	**,921**	-,041
Zscore(CL_prod)	**,905**	,044	-,278	,054
Zscore(CL_total)	**,872**	,091	,385	,044
Zscore(unpaid_amt)	**,496**	-,032	,116	,371
Zscore(Client_Exposure)	**,878**	,095	,388	,015
Zscore(prod_exposure)	**,930**	,054	-,226	,015
Zscore(prod_months)	-,194	,068	,513	**,532**
Zscore(INCOME_ORIGIN)	,164	,096	-,069	**,753**
Zscore(YEARS_IN_ADD)	,029	**,752**	,083	-,231
Zscore(YEARS_IN_COM)	,082	**,793**	,067	,120
Zscore(AGE_ORIGINAL)	,027	**,665**	-,032	,320

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 6 iterations.

Πίνακας 4.10: Φορτία των μεταβλητών στην FA



Σχήμα 4.2: Παράγοντες και μεταβλητές

Η ομάδα μεταβλητών που εκπροσωπείται από τον πρώτο παράγοντα περιέχει όλες τις μεταβλητές που απεικονίζουν τα στοιχεία οφειλής τόσο του συγκεκριμένου προϊόντος όσο και συνολικά του πελάτη. Στον δεύτερο παράγοντα συγκεντρώνονται τα δημογραφικά στοιχεία του πελάτη (ηλικία, έτη στην εργασία, έτη στην τρέχουσα διεύθυνση), ενώ ο τρίτος ουσιαστικά συμπίπτει με την μεταβλητή “acc” που δηλώνει τον αριθμό των καταναλωτικών προϊόντων που έχει κάθε πελάτης στην συγκεκριμένη τράπεζα. Τέλος στον τέταρτο

παράγοντα κυρίαρχο ρόλο έχουν η μεταβλητή του εισοδήματος και η μεταβλητή της παλαιότητας του λογαριασμού.

Είναι σύνηθες στις εφαρμογές της ανάλυσης παραγόντων να χρησιμοποιούνται δείκτες (*ratios*) αντί των απόλυτων μεγεθών των μεταβλητών (Ali and Charbaji, 1993; Öcal et al., 2007).

Σε προβλήματα αξιολόγησης πιστωτικού κινδύνου δύο διαδεδομένοι δείκτες που παρουσιάζουν ενδιαφέρον στους αναλυτές είναι ο $DtI = \frac{Client\ debt}{Income}$ (Yegorova et al., 2000; Zurada, 2007; Tsai et al., 2009;) και ο $utilization = \frac{Exposure}{Credit\ Limit}$ (Sarlija et al., 2006; Hederson and Jagtiani, 2010). Αν χρησιμοποιήσουμε τους δείκτες αυτούς αντί των μεταβλητών που τους δημιουργούν, η FA θα μας οδηγήσει σε 3 παράγοντες που εξηγούν μόλις το 54,3% της συνολικής μεταβλητότητας.

Επομένως παραμένουμε στο μοντέλο με τις αρχικές μεταβλητές το οποίο έχει καλύτερη προσαρμογή αφού μπορεί να ερμηνεύσει το 72% της μεταβλητότητας των δεδομένων μας.

4.4 Εξισορρόπηση δεδομένων

Αφού έχουμε επιλέξει τις μεταβλητές, κατηγορικές και συνεχείς, που είναι υποψήφιες να συμμετέχουν στο μοντέλο ταξινόμησης, και πριν προχωρήσουμε στην επιλογή του κατάλληλου μοντέλου, θα πρέπει να αντιμετωπίσουμε το πρόβλημα της ανισορροπίας των δεδομένων μας.

Όπως προαναφέραμε (§ 3.4), θα εφαρμόσουμε τυχαία δειγματοληψία με επανάθεση (*bootstrap*) για το σύνολο των αναξιόπιστων πελατών, δεκαπλασιάζοντας το μέγεθος του. Έτσι πλέον η αναλογία αξιόπιστων – αναξιόπιτων πελατών από 95%-5% μετατρέπεται σε 60%-40%.

Οπότε η εφαρμογή της λογιστικής παλινδρόμησης θα γίνει στο παραπάνω εξισορροπημένο δείγμα.

APPENDIX 4.1

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
acc	Between Groups	7.328	1	7.328	16.821,75	0,000
	Within Groups	28.821	66.159	0		
	Total	36.150	66.160			
CL_prod	Between Groups	13.999.734.580	1	13.999.734.580	236,94	0,000
	Within Groups	3.909.050.936.381	66.159	59.085.702		
	Total	3.923.050.670.961	66.160			
CL_total	Between Groups	476.346.400.837	1	476.346.400.837	5.118,74	0,000
	Within Groups	6.156.710.714.184	66.159	93.059.307		
	Total	6.633.057.115.021	66.160			
unpaid_amt	Between Groups	264.763.473	1	264.763.473	6.979,46	0,000
	Within Groups	2.509.719.182	66.159	37.935		
	Total	2.774.482.655	66.160			
Client_Exposure	Between Groups	615.571.205.000	1	615.571.205.000	7.480,06	0,000
	Within Groups	5.444.548.912.994	66.159	82.294.909		
	Total	6.060.120.117.993	66.160			
prod_exposure	Between Groups	2.666.874.408	1	2.666.874.408	49,68	0,000
	Within Groups	3.551.422.762.698	66.159	53.680.116		
	Total	3.554.089.637.106	66.160			
prod_months	Between Groups	10.889.723	1	10.889.723	18.128,35	0,000
	Within Groups	39.741.794	66.159	601		
	Total	50.631.517	66.160			
INCOME_ORIGIN	Between Groups	367.090.009.883	1	367.090.009.883	3.364,62	0,000
	Within Groups	7.218.147.646.347	66.159	109.103.034		
	Total	7.585.237.656.229	66.160			
YEARS_IN_ADD	Between Groups	93.179	1	93.179	634,68	0,000
	Within Groups	9.712.884	66.159	147		
	Total	9.806.063	66.160			
YEARS_IN_COM	Between Groups	135.914	1	135.914	2.091,28	0,000
	Within Groups	4.299.730	66.159	65		
	Total	4.435.644	66.160			
AGE_ORIGINAL	Between Groups	76.805	1	76.805	614,73	0,000
	Within Groups	8.266.063	66.159	125		
	Total	8.342.869	66.160			

Πίνακας ANOVA for για τις συστάδες 1 και 2

ΚΕΦΑΛΑΙΟ 5

Προτεινόμενο Μοντέλο

Προηγουμένως επιλέξαμε τις μεταβλητές, συνεχείς και κατηγορικές, που θα χρησιμοποιήσουμε για να εκπαιδεύσουμε τον αλγόριθμο ταξινόμησης (*classifier*). Στο κεφάλαιο αυτό θα περιγράψουμε το μοντέλο της δίτιμης λογιστικής παλινδρόμησης (*Binary Logistic Regression*) το οποίο στη συνέχεια θα εφαρμόσουμε στα δεδομένα μας.

5.1 Γιατί λογιστική παλινδρόμηση;

Οι δύο πιο ευρέως διαδεδομένες μέθοδοι για την μοντελοποίηση του πιστωτικού κινδύνου είναι η διαχωριστική ανάλυση (*discriminant analysis*) και οι διάφοροι τύποι μοντέλων παλινδρόμησης με την χρήση της μέγιστης πιθανοφάνειας (*maximum likelihood regression models*). Αν και η διαχωριστική ανάλυση φαίνεται να είναι η πιο συνηθισμένη, έχει αρχίσει να χάνει έδαφος σε σχέση με μοντέλα μέγιστης πιθανοφάνειας όπως η λογιστική παλινδρόμηση. (βλ. Yegorova, 2000)

Σύμφωνα με τον Anderson (βλ. Krishnaiah and Kanal, 1982), τα τρία ελκυστικότερα πλεονεκτήματα για την χρήση της λογιστικής συνάρτησης είναι τα εξής

- Μικρός αριθμός υποθέσεων για τα δεδομένα που χρησιμοποιούνται. (Έχει λιγότερο περιοριστικές παραδοχές για τις κατανομές που ακολουθούν οι μεταβλητές σε σχέση με άλλες μεθόδους όπως η διαχωριστική ανάλυση, δεν υποθέτει γραμμική σχέση μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών και επιτρέπει την ανισότητα των διασπορών ανάμεσα στις δύο ομάδες της εξαρτημένης μεταβλητής (Yegorova, 2000).)
- Εφαρμογή για ανεξάρτητες μεταβλητές οι οποίες είτε είναι συνεχείς, είτε κατηγορικές ή συνδυασμός τους και
- Ευχρηστία, αφού, από την στιγμή που θα έχουν υπολογιστεί οι παράμετροι του μοντέλου, για την ταξινόμηση ενός νέου υποκειμένου απαιτείται μόνο ο υπολογισμός μιας γραμμικής συνάρτησης.

5.2 Το μοντέλο της δίτιμης λογιστικής παλινδρόμησης

Θεωρούμε ότι έχουμε ένα δείγμα με *n* ανεξάρτητες παρατηρήσεις, και το ζεύγος $(\vec{x}_i, y_i)$ αποτελείται από το *k*-διάστατο διάνυσμα $\vec{x}$ των *k* ανεξάρτητων μεταβλητών και την τιμή *y* της εξαρτημένης μεταβλητής, για την *i*-οστή παρατήρηση. Οι ανεξάρτητες μεταβλητές μπορεί να είναι είτε συνεχείς, είτε κατηγορικές ή συνδυασμός τους. Η εξαρτημένη μεταβλητή είναι δίτιμη με δυνατές τιμές 1-0 για επιτυχία/παρουσία και αποτυχία/απουσία αντίστοιχα.

Με $p_i = P(y_i = 1 | \vec{x}_i)$ συμβολίζουμε την υπο συνθήκη (*conditional*) πιθανότητα επιτυχίας, δηλαδή την πιθανότητα επιτυχίας γνωρίζοντας τις τιμές των ανεξάρτητων μεταβλητών. Προφανώς, η αντίστοιχη πιθανότητα αποτυχίας είναι ίση με $P(y_i = 0 | \vec{x}_i) = 1 - p_i$.

Ο απλούστερος τρόπος εκτίμησης της παραπάνω πιθανότητας θα ήταν ένα μοντέλο γραμμικής παλινδρόμησης (*linear regression model*) της μορφής $p_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$. Όμως, η εκτιμώμενη από το παραπάνω μοντέλο πιθανότητα $\hat{p}_i$ ενδέχεται να μην ανήκει στο διάστημα [0,1].

Για να αποφύγουμε το παραπάνω πρόβλημα, χρησιμοποιούμε αντί της πιθανότητας *p* μια συνάρτηση g(*p*) τέτοια ώστε g: [0,1] → □ . Η συνάρτηση g ονομάζεται συνάρτηση σύνδεσης (*link function*).

Για δίτιμα δεδομένα οι συναρτήσεις σύνδεσης που χρησιμοποιούνται είναι οι ακόλουθες (βλ. Πολίτης, 2010)

1. **Logit**: $g(p) = \log it(p) = \log\left(\frac{p}{1-p}\right)$
2. **Probit**: $g(p) = \Pr obit(p) = \Phi^{-1}(p)$, όπου Φ είναι η αθροιστική συνάρτηση κατανομής της τυποποιημένης κανονικής.
3. **Complementary log-log**: $g(p) = \log[-\log(1-p)]$

Όταν για συνάρτηση σύνδεσης εφαρμόζεται η *logit*, τότε έχουμε το μοντέλο της λογιστικής παλινδρόμησης (*logistic regression*) που περιγράφεται από την συνάρτηση

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad \textbf{(5.1)}$$

Το βασικότερο πλεονέκτημα του παραπάνω μετασχηματισμού είναι ότι διατηρούνται όλες ο επιθυμητές ιδιότητες της γραμμικής παλινδρόμησης. Η συνάρτηση *logit* είναι γραμμική ως

προς τις εξαρτημένες μεταβλητές, συνεχής και μπορεί να πάρει τιμές από το -∞ έως το +∞. (Hosmer and Lemeshow, 2000)

Από τον τύπο (5.1) προκύπτει ότι η εκτιμώμενη πιθανότητα είναι ίση με

$$\hat{p} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)}}$$

και αυτήν θα χρησιμοποιήσουμε για να ταξινομήσουμε καθεμία από τις παρατηρήσεις μας σε κάποια από τις δύο ομάδες, $y = 1$ ή $y = 0$, σύμφωνα με τον κανόνα

$$\hat{y}_i = \begin{cases} 0, \text{ αν } \hat{p}_i \le cut - off \\ 1, \text{ αν } \hat{p}_i > cut - off \end{cases},$$

όπου το σημείο αποκοπής (*cut-off*) συνηθίζεται να είναι το 0,5, αλλά μπορεί να μετατοπιστεί ανάλογα με την φύση του προβλήματος που έχουμε να αντιμετωπίσουμε.

Η εκτίμηση των παραμέτρων β_i είναι βασισμένη στη μέθοδο της μέγιστης πιθανοφάνειας (*maximum likelihood*), όπως και στην απλή γραμμική παλινδρόμηση. Για το μοντέλο της λογιστικής παλινδρόμησης η συνάρτηση πιθανοφάνειας είναι η $l(\vec{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$, την οποία θα πρέπει να μεγιστοποιούν οι συντελεστές β_i που τελικά υπολογίζονται.

Στη συνέχεια θα εστιάσουμε στη φυσική ερμηνεία των συντελεστών β_i του μοντέλο της λογιστικής παλινδρόμησης.

Ο λόγος της πιθανότητας επιτυχίας προς την πιθανότητα αποτυχίας (σχετική πιθανότητα), δηλαδή ο $\frac{p}{1-p}$, ονομάζεται *odds* και θα τον συμβολίζουμε με *Ω*. Από την σχέση (5.1) προκύπτει ότι

$$\Omega = \frac{p}{1-p} = e^{(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)} \qquad \textbf{(5.2)}$$

Ο λόγος των *odds* (*odds ratio*) που προκύπτει αν αυξήσουμε την τιμή της *i*-οστης μεταβλητής κατά μια μονάδα, διατηρώντας τις υπόλοιπες σταθερές, ισούται με e^{β_i}, δηλαδή

$$odds\ ratio = \frac{\Omega_B}{\Omega_A} = \frac{e^{(\beta_0 + \beta_1 x_1 + ... + \beta_i (x_i + 1) + ... + \beta_k x_k)}}{e^{(\beta_0 + \beta_1 x_1 + ... + \beta_i x_i + ... + \beta_k x_k)}} = e^{\beta_i} \qquad \textbf{(5.3)}$$

Διαφορετικά, «η μοναδιαία αύξηση της τιμής της μεταβλητής X_i προκαλεί πολλαπλασιαστική αύξηση της σχετικής πιθανότητας κατά e^{β_i} (όταν οι τιμές των άλλων μεταβλητών παραμένουν σταθερές)» (Πολίτης, 2010).

Θα πρέπει να σημειώσουμε ότι οι κατηγορικές ανεξάρτητες μεταβλητές για να μπουν στο μοντέλο, πρέπει πρώτα να κωδικοποιηθούν σε δίτιμες (0-1) ψευδομεταβλητές (*dummy variables*) οι οποίες ονομάζονται και δείκτριες (*indicators*). Για κάθε κατηγορική μεταβλητή με *d* επίπεδα δημιουργούνται *d*-1 ψευδομεταβλητές, οι οποίες παίρνουν την τιμή 1 για το επίπεδο που εκφράζουν και 0 αλλιώς. Το επίπεδο για το οποίο δεν δημιουργείται ψευδομεταβλητή (ουσιαστικά 0 παντού) θεωρείται επίπεδο αναφοράς και με αυτό συγκρίνονται τα υπόλοιπα επίπεδα.

Έτσι η μοναδιαία αύξηση στην περίπτωση των κατηγορικών μεταβλητών ισοδυναμεί με την διαφοροποίηση του επιπέδου από το επίπεδο αναφοράς.

Για περισσότερες λεπτομέρειες σχετικά με την μέθοδο ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο των Hosmer and Lemeshow (2000) και στα βιβλία του Agresti (2002,2007).

5.3 Επιλογή του βέλτιστου μοντέλου

Βέλτιστο θεωρείται το μοντέλο το οποίο με όσο το δυνατόν λιγότερες μεταβλητές μπορεί να ερμηνεύσει με όσο το δυνατόν καλύτερο τρόπο την εξαρτημένη μεταβλητή.

Υπάρχουν τρία διαφορετικά κριτήρια με τα οποία επιλέγουμε ποια μεταβλητή θα εξαιρεθεί από ένα μοντέλο, βάσει της στατιστικής σημαντικότητας του αντίστοιχου συντελεστή. Τα κριτήρια αυτά είναι:

- *Wald statistic*: Αν η *i*-οστή μεταβλητή είναι συνεχής τότε υπολογίζεται από τον τύπο

$$Wald_i = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}.$$

  Αν είναι κατηγορική και $\hat{\vec{\beta}}_i$ είναι το διάνυσμα με τις εκτιμήσεις των συντελεστών των *d*-1 ψευδομεταβλητών και **C** είναι ο ασυμπτωτικός πίνακας των συνδιακυμάνσεων για το $\hat{\vec{\beta}}_i$, τότε

$$Wald_i = \hat{\vec{\beta}}_i' \mathbf{C}^{-1} \hat{\vec{\beta}}_i.$$

Η ασυμπτωτική κατανομή του στατιστικού είναι η χ^2 με βαθμούς ελευθερίας όσοι οι συντελεστές που εκτιμώνται.

- <u>*Likelihood ratio test*</u> (LRT): Αν *l* είναι η συνάρτηση πιθανοφάνειας ενός μοντέλου και M_1, M_2 είναι τα συγκρινόμενα μοντέλα, όπου το M_2 περιέχει υποσύνολο των μεταβλητών του M_1 τότε

$$LRT = -2\log\left(\frac{l(M_2)}{l(M_1)}\right)$$

Η ασυμπτωτική κατανομή του στατιστικού είναι η χ^2 με βαθμούς ελευθερίας ίσους με τη διαφορά του αριθμού των συντελεστών των δύο μοντέλων.

- <u>*Conditional statistic*</u>: Ο τύπος υπολογισμού είναι ίδιος με το LRT, με τη διαφορά ότι οι εκτιμήσεις των παραμέτρων του μοντέλου M_2, αντί να είναι βασισμένες στην μέγιστη πιθανοφάνεια (MLEs), είναι υπο συνθήκη εκτιμήσεις (*conditional estimates*). Για λεπτομέρειες βλ. Hosmer and Lemeshow (2000).

Σύμφωνα με τον Agresti (2007), αν και το τεστ του *Wald* είναι επαρκές για μεγάλα δείγματα, το *likelihood ratio test* είναι ισχυρότερο και πιο αξιόπιστο για τα μεγέθη δείγματος που συνήθως χρησιμοποιούμε στην πράξη.

Το κριτήριο για την εισαγωγή μιας μεταβλητής στο μοντέλο είναι το *Score statistic*. Για τον αναλυτικό υπολογισμό του βλ. Hosmer and Lemeshow (2000).

Υπάρχουν δύο μέθοδοι για την σταδιακή επιλογή του μοντέλου (*stepwise selection*), (Agresti, 2007)

- <u>Προς τα πίσω περιορισμός</u> (*backward elimination*): Ο αλγόριθμος ξεκινάει με το μοντέλο που περιέχει όλες τις μεταβλητές και διαδοχικά αποβάλει κάποιες από αυτές. Σε κάθε βήμα εξαιρείται η μεταβλητή με το μεγαλύτερο *p-value* στο τεστ που ελέγχει αν ο συντελεστής της ισούται με μηδέν. Αφού η μεταβλητή φύγει από το μοντέλο, ο αλγόριθμος επανελέγχει αν κάποια από τις μεταβλητές που είναι εκτός τώρα πλέον θεωρείται σημαντική και πρέπει να εισαχθεί στο μοντέλο. Η παραπάνω διαδικασία σταματάει όταν όλοι οι συντελεστές των εναπομείναντων μεταβλητών είναι στατιστικά σημαντικοί, οπότε οποιαδήποτε επιπλέον εξαίρεση θα υποβαθμίσει την προσαρμογή του μοντέλου.
- <u>Προς τα εμπρός επιλογή</u> (*forward selection*): Η διαδικασία αυτή είναι αντίστροφη της προηγούμενης. Ξεκινάει χωρίς καμία μεταβλητή στο μοντέλο και σε κάθε βήμα

προσθέτει εκείνη με το μεγαλύτερο *Score* και στη συνέχεια ελέγχει αν κάποια από τις ήδη υπάρχουσες έχει πάψει να είναι πλέον σημαντική, οπότε και την εξαιρεί. Ο αλγόριθμος τερματίζεται όταν δεν υπάρχει πλέον καμιά σημαντική μεταβλητή για να εισαχθεί στο μοντέλο.

Και στις δύο μεθόδους, για την απόρριψη των μεταβλητών από το μοντέλο, μπορεί να χρησιμοποιηθεί οποιοδήποτε από τα τρία κριτήρια που προαναφέραμε.

5.4 Εφαρμογή της λογιστικής παλινδρόμησης

Τα δεδομένα μας θα αναλυθούν με την μέθοδο της λογιστικής παλινδρόμης για να ταξινομήσουμε τους πελάτες μας σε αναξιόπιστους και αξιόπιστους. Η επιλογή του βέλτιστου μοντέλου θα γίνει με *Backward elimination,* χρησιμοποιώντας ως κριτήριο για την έξοδο μιας μεταβλητής από το μοντέλο το *likelihood ratio test*.

Η προεπιλεγμένη τιμή για το σημείο cut-off είναι το 0,5. Όμως, την συγκεκριμένη χρονική περίοδο η τράπεζα ακολουθώντας αμυντική πολιτική, δίνει πολύ μεγάλη έμφαση στην διαχείριση και αντιμετώπιση πελατών που είναι επιρρεπείς στην μη-αποπληρωμή των δανείων τους, θέλοντας να προλάβει αρνητικά αποτελέσματα. Οπότε, θέλουμε να δώσουμε μεγαλύτερη βαρύτητα στην ορθή πρόβλεψη των αναξιόπιστων πελατών. Στην περίπτωση αυτή η πιθανότητα 0,5 είναι αρκετά υψηλή και καταλληλότερη για το κρίσιμο σημείο θεωρείται η τιμή 0,3. (Yegorova, 2000) Δηλαδή κάθε πελάτης για τον οποίο, σύμφωνα με το μοντέλο, η πιθανότητα να έχει σταματήσει να εξυπηρετεί το δάνειό του στο τέλος του επόμενου εξαμήνου είναι μεγαλύτερη από 30% θα χαρακτηρίζεται ως αναξιόπιστος.

Στην αντίθετη περίπτωση, που η πολιτική της τράπεζας ήταν επιθετική με σκοπό την προώθηση νέων δανείων σε πελάτες που θεωρούνται αξιόπιστοι, η τιμή 0,5 κρίνεται ικανοποιητική για τον διαχωρισμό των πελατών.

Μεταβλητή	**Επίπεδο αναφοράς**
Τύπος δανείου	Πιστωτική Κάρτα
Επαγγ. Δάνειο σε καθυστέρηση	Όχι
Επάγγελμα	Ιδιωτικός υπάλληλος
Οικογενειακή κατάσταση	Έγγαμος
Φύλο	Γυναίκα

Πίνακας 5.1: Κατηγορία αναφοράς για τις κατηγορικές μεταβλητές

Τα επίπεδα αναφοράς για τις κατηγορικές μεταβλητές που είναι υποψήφιες για να εισαχθούν στο μοντέλο, διαμορφώνονται σύμφωνα με τον Πίνακα 5.1.

5.5 Αποτελέσματα - Ερμηνεία

Θεωρώντας λοιπόν ότι *p* είναι η πιθανότητα το δάνειο του πελάτη να καταλήξει μη εξυπηρετούμενο στο τέλος του επόμενου εξαμήνου, οι συντελεστές του μοντέλου που προκύπτει από την εφαρμογή της λογιστικής παλινδρόμησης φαίνονται στον Πίνακα 5.2.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for Lower	95% C.I.for Upper
Step 1[a]	prod_type			2806,156	2	,000			
	prod_type(1)	-,764	,023	1109,529	1	,000	,466	,445	,487
	prod_type(2)	-1,047	,020	2801,182	1	,000	**,351**	,338	,365
	BL(1)	,351	,035	103,278	1	,000	**1,420**	1,327	1,520
	OCC_CAT			751,574	3	,000			
	OCC_CAT(1)	,051	,018	8,139	1	,004	1,052	1,016	1,089
	OCC_CAT(2)	-,319	,023	188,694	1	,000	,727	,694	,761
	OCC_CAT(3)	,328	,018	323,050	1	,000	1,389	1,340	1,440
	FAMILY_STATUS			293,830	3	,000			
	FAMILY_STATUS(1)	,010	,016	,383	1	,536	1,010	,979	1,042
	FAMILY_STATUS(2)	,436	,027	262,862	1	,000	**1,546**	1,467	1,629
	FAMILY_STATUS(3)	,263	,042	39,547	1	,000	1,300	1,198	1,411
	GENDER(1)	,175	,014	149,205	1	,000	1,191	1,158	1,225
	FAC1	,400	,006	3960,055	1	,000	**1,492**	1,474	1,511
	FAC2	-,086	,007	138,169	1	,000	,917	,904	,930
	FAC3	,016	,007	4,959	1	,026	1,016	1,002	1,030
	FAC4	-,207	,007	818,881	1	,000	,813	,802	,825
	Constant	-,024	,020	1,395	1	**,238**	,977		

a. Variable(s) entered on step 1: prod_type, BL, OCC_CAT4, FAMILY_STATUS_1, GENDER_1, FAC1, FAC2, FAC3, FAC4.

Πίνακας 5.2: Συντελεστές του μοντέλου λογιστικής παλινδρόμησης

Με βάση τους παραπάνω συντελεστές το μοντέλο μας είναι το ακόλουθο:

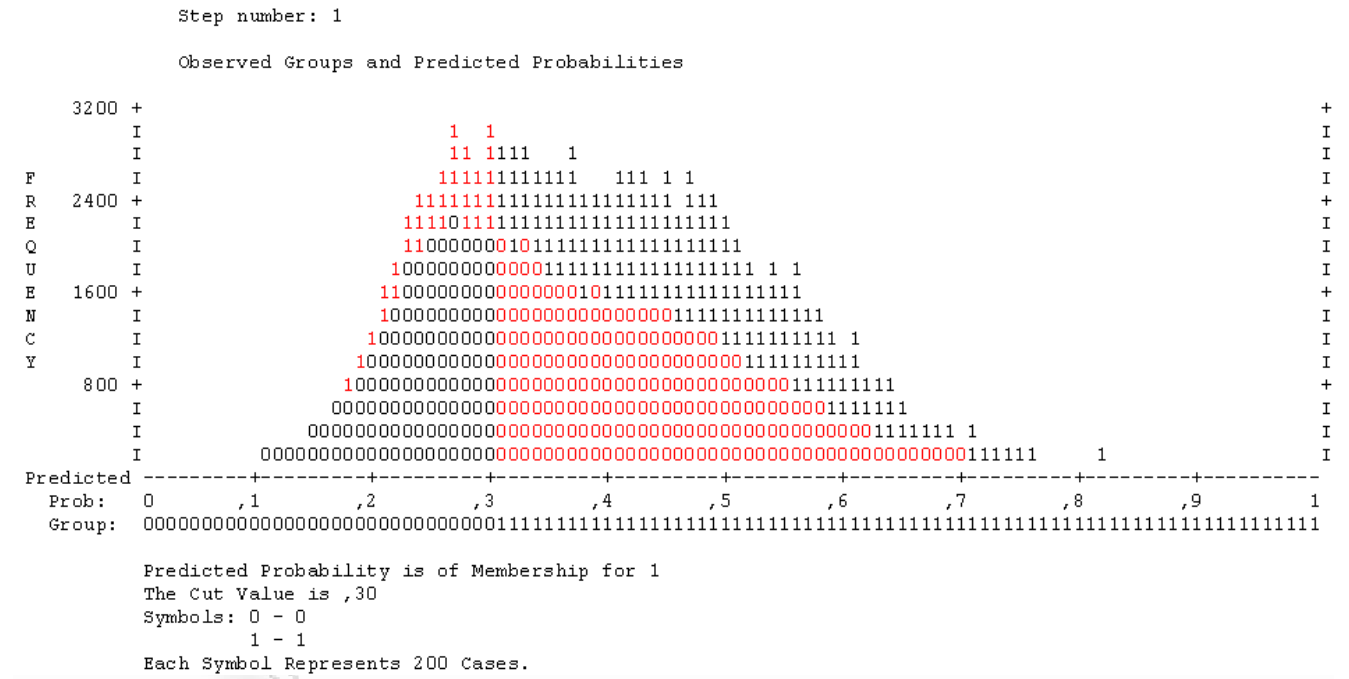
$$\log it(p) = \log\left(\frac{p}{1-p}\right) = -0,024 - 0,764 \cdot OL - 1,047 \cdot TKX + 0,351 \cdot BL + 0,051 \cdot Occ.A - 0,319 \cdot Occ.B +$$
$$+ 0,328 \cdot Occ.F + 0,010 \cdot FS.B + 0,436 \cdot FS.C + 0,263 \cdot FS.E + 0,175 \cdot male +$$
$$+ 0,4 \cdot FAC1 - 0,086 \cdot FAC2 + 0,016 \cdot FAC3 - 0,207 \cdot FAC4$$
$$= \hat{\vec{\beta}} \cdot \vec{x}$$

Αν θέλουμε να γράψουμε την πιθανότητα p ως συνάρτηση των μεταβλητών μας, τότε η εξίσωση διαμορφώνεται ως εξής:

$$\hat{p} = \frac{e^{\hat{\vec{\beta}} \cdot \vec{x}}}{1 + e^{\hat{\vec{\beta}} \cdot \vec{x}}}$$

Όπως διαπιστώνουμε (Πίνακας 5.2) το μοντέλο περιλαμβάνει όλες τις μεταβλητές ως στατιστικά σημαντικές, αλλά ο σταθερός όρος (*constant*) δεν είναι στατιστικά σημαντικός (αφού *p-value* = 0,238 > 0,05).

Με βάση το παραπάνω μοντέλο οι πελάτες του εκπαιδευτικού δείγματος και του δείγματος ελέγχου ταξινομούνται όπως δείχνει ο Πίνακας 5.3 καθώς και το Σχήμα 5.1.



Σχήμα 5.1: Διάγραμμα μη-ορθής ταξινόμησης

Για την αξιολόγηση του μοντέλου θα στηριχτούμε στα αποτελέσματα που προκύπτουν από την ταξινόμηση στο δείγμα ελέγχου (*Test set*) το οποίο περιέχει υπόθεσεις «άγνωστες» για το

μοντέλο. Διαφορετικά, αν χρησιμοποιούσαμε το εκπαιδευτικό δείγμα, που είναι αυτό στο οποίο βασίστηκε η εκμάθηση του μοντέλου, τα αποτελέσματα θα ήταν «αισιόδοξα».

Έτσι, έχουμε ότι από τους 62.399 αξιόπιστους πελάτες του δείγματος ελέγχου, σωστά ταξινομούνται οι 22.825, ποσοστό που ανέρχεται στο 36,6%. Αν και αυτό το ποσοστό δεν θεωρείται πολύ καλό, τα πράγματα διαφοροποιούνται κατά πολύ αν κοιτάξουμε τους αναξιόπιστους πελάτες. Από τους 3.542 ταξινομούνται σωστά ως αναξιόπιστοι οι 2.958, ποσοστό που φτάνει στο 83,5% και είναι αρκετά ικανοποιητικό.

Classification Table[c]

Observed			Predicted					
			Training Set [a]			Test Set [b]		
			RISK		Percentage	RISK		Percentage
			0	1	Correct	0	1	Correct
Step 1	RISK	0	22.928	39.815	36,5	**22.825**	39.574	36,6
		1	6.566	34.450	84,0	584	**2.958**	**83,5**
	Overall Percentage				55,3			39,1

a. Selected cases TRAINING EQ 1

b. Unselected cases TRAINING NE 1

c. The cut value is ,300

Πίνακας 5.3: Ταξινόμηση σύμφωνα με το μοντέλο στα δείγματα εκπαίδευσης και ελέγχου

Επειδή, όπως τονίσαμε και στην προηγούμενη παράγραφο, όλο το βάρος από την στρατηγική της τράπεζας πέφτει στην ορθή πρόβλεψη των πελατών που θα αποδειχθούν αναξιόπιστοι και θα σταματήσουν να εξυπηρετούν τις δόσεις των δανείων τους, το μοντέλο αυτό μας καλύπτει.

Εξάλλου, η καμπύλη ROC (Σχήμα 5.2) και τα αντίστοιχα στατιστικά για την περιοχή κάτω από την καμπύλη (Πίνακας 5.4) δείχνουν ότι είναι προτιμότερο να χρησιμοποιήσουμε το μοντέλο για την πρόβλεψη της ομάδας που ανήκει ο πελάτης από το να την μαντέψουμε (*ασυμπτωτικό p-value* < 0,05). Η περιοχή κάτω από την καμπύλη (0,668) εκφράζει την πιθανότητα το μοντέλο ταξινόμησης να δώσει μεγαλύτερη πιθανότητα αναξιοπιστίας σε ένα τυχαία επιλεγμένο άτομο από τον πληθυσμό των αναξιόπιστων πελατών σε σχέση με ένα τυχαία επιλεγμένο άτομο από τον πληθυσμό των αξιόπιστων πελατών. Το παραπάνω δείχνει την αξιοπιστία του μοντέλου μας.

Area Under the Curve

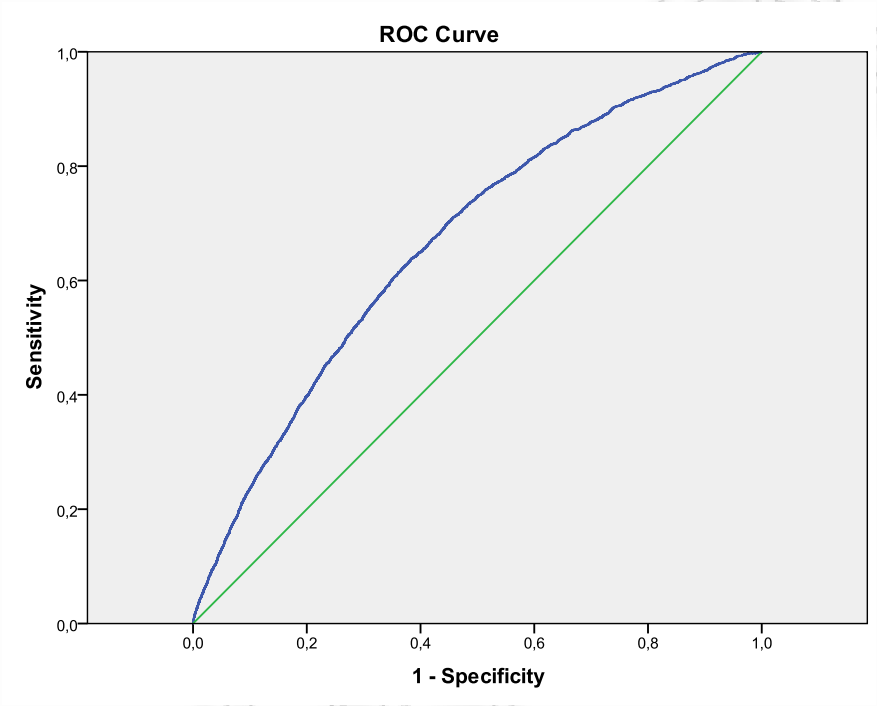
Test Result Variable(s):Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,668	,001	**,000**	,665	,671

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Πίνακας 5.4: Στατιστικά της καμπύλης ROC



Σχήμα 5.2: Καμπύλη ROC

Ακολουθούν τα βασικά συμπεράσματα που προκύπτουν από το μοντέλο μας.

Όπως δείχνουν οι μεταβλητές που συμμετέχουν στο μοντέλο, το αν ο πελάτης μετά από έξι μήνες συνεχίσει να εξυπηρετεί το δάνειό του ή όχι εξαρτάται από τον τύπο του δανείου, την οικογενειακή του κατάσταση, το φύλο του, το επαγγελμά του καθώς και το γεγονός ότι έχει επαγγελματικό δάνειο σε καθυστέρηση ή όχι. Ακόμα επηρεάζεται από τα οικονομικά χαρακτηριστικά του δανείου και την παλαιότητα του, το εισόδημα του πελάτη και δημογραφικά χαρακτηριστικά όπως η ηλικία και τα έτη διαμονής στην τρέχουσα διεύθυνση.

Επίσης, όλα τα επιμέρους επίπεδα των κατηγορικών μεταβλητών φαίνεται (Πίνακας 5.2) να διαφοροποιούνται από το επίπεδο αναφοράς στον τρόπο με τον οποίο επηρεάζουν την

συμπεριφορά των πελατών, αφού οι επιμέρους συντελεστές είναι όλοι σημαντικοί (*p-value* < 0,05). Εξαίρεση αποτελεί η οικογενειακή κατάσταση, όπου οι άγαμοι πελάτες (FS_1) συμπεριφέρονται όμοια με τους έγγαμους (FS_0) οι οποίοι είναι η κατηγορία αναφοράς.

Για παράδειγμα, με βάση την σχέση (5.3) και τον Πίνακα 5.2, ο λόγος της πιθανότητας ένας πελάτης να είναι αναξιόπιστος προς την πιθανότητα να είναι αξιόπιστος (*odds*) είναι 1,42 φορές μεγαλύτερος για αυτούς που έχουν επαγγελματικό δάνειο σε καθυστέρηση (BL_1), σε σχέση με αυτούς που δεν έχουν. Ακόμα, το *odds* αναξιοπιστίας προς αξιοπιστία είναι 1,546 φορές μεγαλύτερο για τους διαζευγμένους πελάτες (FS_3) σε σχέση με τους έγγαμους. Επιπλέον ο παραπάνω λόγος είναι 2,85 (1/0,351) φορές μεγαλύτερος για τους κατόχους πιστωτικής κάρτας (prod_type_0) έναντι των κατόχων τοκοχρεωλυτικού δανείου (prod_type_2). Τέλος, η αύξηση κατά μια μονάδα στον Παράγοντα 1 (FAC1), που εκπροσωπεί το οικονομικά στοιχεία του δανείου, αυξάνει την σχετική πιθανότητα αναξιοπιστίας 1,492 φορές.

Ομοίως, μπορούμε να ερμηνεύσουμε και τους υπόλοιπους λόγους και να καταλήξουμε σε ενδιαφέροντα συμπεράσματα για την συμπεριφορά των πελατών, που θα μας οδηγήσουν στην βελτιστοποίηση της στρατηγικής μας για την διαχείριση του πελατολογίου.

Για την σύγκριση των αποτελεσμάτων, που έχουν προκύψει σχετικά με τις μεταβλητές και τον τρόπο που επηρεάζουν την διαμόρφωση του ρίσκου, με άλλες παρόμοιες αναλύσεις παραπέμπουμε στο βιβλίο του Altman (1981, pp 188).

Βιβλιογραφία

Βασιλάκη, (2010). Στατιστικά Μοντέλα Βαθμολόγησης Πιστοληπτικής Ικανότητας. *Διπλωματική εργασία, ΠΜΣ στην Εφαρμοσμένη Στατιστική, Παν/μιο Πειραιώς*.

Ηλιόπουλος, (2007). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση. *Σημειώσεις, ΠΜΣ στην Εφαρμοσμένη Στατιστική, Παν/μιο Πειραιώς*.

Κούτρας, (2007). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά Συστάδες. *Σημειώσεις, ΠΜΣ στην Εφαρμοσμένη Στατιστική, Παν/μιο Πειραιώς*.

Πολίτης, (2010). Γενικευμένα Γραμμικά Μοντέλα. *Σημειώσεις, ΠΜΣ στην Εφαρμοσμένη Στατιστική, Παν/μιο Πειραιώς.*

Agresti, (2002). *Categorical Data Analysis*, Second Edition. Wiley.

Agresti, (2007). *An Introduction to Categorical Data Analysis*, Second Edition. Wiley.

Ali and Charbaji, (1993). Applying Factor Analysis to Financial Ratios of International Commercial Airlines. *International Journal of Commerce and Management*, **4**, 25 - 37.

Allison, (2002). *Missing Data*. Chapter 4. Thousand Oaks, Sage Publications.

Altman et al., (1981). *Application of Classification Techniques in Business, Banking and Finance*. JAI Press.

Altman et al., (1981). Application of Classification Techniques in Business, Banking and Finance. *Contemporary Studies in Economic and Financial Analysis*, **3**.

Ang and Willhour (1976). The Consumer Loan Supply Function of a Minority Bank: An Empirical Note: Comment. *Journal of Money, Credit and Banking*, **8**, 255-259.

Antonie et al., (2001). Application of Data Mining Techniques for Medical Image Classification. *Proceedings of the 2nd International Workshop on Multimedia Data Mining*, 94-101.

Banks and Fienberg, (2002). *Data Mining. Encyclopedia of Physical Science and Technology*, third edition. Academic Press.

Baoli et al., (2003). An Improved k-Nearest Neighbour Algorithm for Text Categorization. *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages.*

Bardos, (1998). Detecting the risk of company failure at the banquet de France. *Journal of Banking and Finance*, **22**, 1405-1419.

Batista et al., (2004). A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data. *Sigkdd Explorations*, **6**, 20-29.

Catelani and Fort, (2000). Fault diagnosis of electronic analog circuits using a radial basis function network classifier. *Measurement,* **28**, 147–158.

Chawla et al., (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* **16,** 321-357.

Chawla et al., (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. *Sigkdd Explorations,* **6**.

Chen et al., (2009). A model based on factor analysis and Support Vector Machine for Credit Risk Identification in small-and-medium enterprises. *International Conference on Machine Learning and Cybernetics*, 913-918.

Cios et al., (2007). *Data Mining, A Knowledge Discovery Approach*. Springer.

Dai et al., (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Modecular Biology*, **5**.

Desai et al., (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* **95**, 24-37.

Diakoulaki et al., (1999). The use of a preference disaggregation method in energy analysis and policy making. *Energy*, **24**, 157–166.

Dreiseitl et al., (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics,* **35**, 352-359.

Dumais and Chen, (2000). Hierarchical classification of Web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval,* 256-263.

Dunn et al., (1984). Pattern recognition for classification and determination of polychlorinated biphenyls in environmental samples. *Analytical Chemistry,* **56**, 1308–1313.

Fayyad et al., (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of ACM,* **39**, 27-34.

Fisher, (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7**, 179-188.

Flinkman et al., (2000). Use of rough sets analysis to classify Siberian forest ecosystem according to net primary production of phytomass. *INFOR,* **38**, 145–161.

Gnanadhas and Geetha, (2009). Repayment of Loan in Employees' Cooperative Thrift and Credit Societies. *Journal of Rural Development,* **28**, 485 - 490.

Gochet et al., (1997). Multigroup discriminant analysis using linear programming. *Operations Research*, **45**, 213–225.

Han and Kamber, (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Han and Zhao, (2008). A Scene Images Classification Method Based on Local Binary Patterns and Nearest-Neighbour Classifier. *Intelligent Systems Design and Applications,* **3**, 100-104.

Hays et al., (2002). Predicting micro-loan defaults using probabilistic neural networks. *The Credit and Financial Management Review,* **9**.

Hederson and Jagtiani, (2010). Can Banks Circumvent Minimum Capital Requirements? The Case of Mortgage Portfolios under Basel II. *Working paper, Research Department, Federal Reserve Bank of Philadelphia.*

Hosmer and Lemeshow, (2000). *Applied Logistic Regression,* Second Edition. Wiley.

Huang et al., (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications,* **33**, 847-856.

Hulse et al., (2007). Experimental Perspectives on Learning from Imbalanced Data. *Proceedings of 24th International Conference on Machine Learning, Corvallis.*

Jantan et al., (2009). Classification Techniques for Talent Forecasting in Human Resource Management. *Advanced Data Mining and Applications, Lecture Notes in Computer Science,* **5678**, 496-503.

Japkowicz and Stephen, (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis,* **6**, 203-231.

Johnson and Wichern, (1998). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Joshi et al., (2001). Evaluating boosting algorithms to classify rare cases: comparison and improvements. *In First IEEE International Conference on Data Mining,* 257-264.

Keller et al., (2000). Bayesian Classification of DNA Array Expression Data. *Technical Report UW-CSE-2000-08-01, Department of Computer Science & Engineering, University of Washington.*

Kemsley, (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems,* **33**, 47-61.

Kosmidou et al., (2006). A multivariate analysis of the financial characteristics of foreign and domestic banks in the UK. *Omega,* **36**, 189-195.

Kotsiantis et al., (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering,* **30**.

Kotsiantis, (2007). Supervised Machine Learning: A review of Classification Techniques. *Informatica,* **31**, 249-268.

Krishnaiah and Kanal, (1982). *Handbook of Statistics, Vol. 2: Classification, Pattern Recognition and Reduction of Dimensionality*. North-Holland Publishing Company.

Kwon and Lee, (2003). Text categorization based on k-nearest neighbour approach for Web site classification. *Information Processing & Management,* **39**, 25-44.

Laitinen, (1999). Predicting a corporate credit analyst's risk estimate by logistic and linear models. *International Review of Financial Analysis,* **8**, 97-121.

Lee et al., (2002). Credit Scoring using the hybrid neural discriminant technique. *Expert Systems with Applications,* **23**, 245-254.

Lee et al., (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, **50**, 113-1130.

Liu and Motoda, (1998). *Feature extraction, construction and selection, A data mining perspective,* Second Printing (2001). Kluwer Academic Publishers

Liu et al. (2006). Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. *Advances in Knowledge Discovery and Data Mining*, **3918**, 107-118.

Maimon and Rokach, (2005). *The Data Mining and Knowledge Discovery*. Springer.

Maloof, (2003). Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. *Workshop on Learning from Imbalanced Data Sets II, ICML, Washington.*

Martin, (1977). Early warning of bank failure: Logit regression approach. *Journal of Banking and Finance*, **1**, 249-276.

Michalowski et al., (2001). Triage of the child with abdominal pain: A clinical algorithm for emergency patient management. *Paediatrics and Child Health,* **6**, 23-28.

Michie et al., (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

Nalbantov et al., (2005). Solving and Interpreting Binary Classification Problems in Marketing with SVMs. *Econometric Institute Report E.I. 2005-46*.

Ngai et al., (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, **36**, 2592-2602.

Nowicki et al., (1992). Evaluation of vibroacoustic diagnostic symptoms by means of the rough sets theory. *Journal of Computers in Industry,* **20**, 141–152.

Öcal et al., (2007). Industry financial ratios-application of factor analysis in Turkish construction industry. *Building and Environment,* **42**, 385-392.

Olafson et al., (2008). Operations research and data mining. *European Journal of Operational Research,* **187**, 1429-1448.

Paxton, (1996). Determinants of successful group loan repayment: An application to Burkina Faso. *Dissertation, The Ohio State University*.

Piramuthu et al., (1998). Using Feature Construction to Improve the Performance of Neural Networks. *Management Science,* **44**, 416-430.

Raskutti and Kowalczyk, (2004). Extreme Re-balancing for SVMs: a case study. *Sigkdd Explorations*, **6**, 60-69.

Rossi et al., (1999). Rough set approach to evaluation of stormwater pollution. *International Journal of Environment and Pollution,* **12**, 232–250.

Rummel, (1979). *Applied Factor Analysis*, Fourth printing. Northwestern University Press.

Sarlija et al., (2006). Modelling customer revolving credit scoring using logistic regression, survival analysis and neural networks. *Proceedings of the 7th WSEAS International Conference on Neural Networks,* 164-169.

Schenker et al., (2004). Classification of Web documents using graph matching. *International Journal of Pattern Recognition and Artificial Intelligence,* **18**, 475-496.

Shen et al., (2000). Fault diagnosis using rough sets theory. *Computers in Industry,* **43**, 61–72.

Siddiqi, (2000). Credit Risk Scorecards – Developing and Implementing Intelligent Credit Scoring. Wiley.

Sinha et al., (2009). Performance and Evaluation of MicroRNA Gene Identification Tools. *Journal of Proteomics & Bioinformatics,* **2**, 336-343.

Siskos et al., (1998). Measuring customer satisfaction using a survey based preference disaggregation model. *Journal of Global Optimization,* **12**, 175-195.

SPSS, (2001). The SPSS TwoStep Cluster Component. *White paper - technical report.*

Stefanowski and Slowinski, (1998). Rough set theory and rule induction techniques for discovery of attribute dependencies in medical information systems. *Bulletin of the Polish Academy of Sciences, ser Technical Sciences,* **46**, 247-263.

Sun et al., (2002). Web classification using support vector machine. *Proceedings of the 4th international workshop on Web information and data management,* 96-99.

Sung et al., (1999). Dynamics of modelling in data mining: Interpretive approach to bankruptcy prediction. *Journal of Management Information Systems,* **16**, 63-85.

Tan et al., (2006). *Introduction to Data Mining*. Pearson Education.

Tang et al., (2008). A Vectorial Image Classification Method Based On Neighbourhood Weighted Gaussian Mixture Model. *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 1922-1925.

Tran et al., (2008). Prediction of human microRNA hairpins using only positive sample learning. *Biomedical Science and Engineering,* **1**, 141-146.

Tsai et al., (2009). The consumer loan default predicting model – An application of DEA–DA and neural network. *Expert Systems with Applications,* **36**, 11682-11690.

Tsumoto, (1998). Automated extraction of medical expert system rules from clinical databases based on rough set theory. *Information Sciences,* **112**, 67-84.

Twala, (2009). Combining classifiers for credit risk prediction. *Journal of Systems Science and Systems Engineering,* **18**, 292-311.

West, (1985). A factor analytic approach to bank condition. *Journal of Banking and Finance,* **9**, 253-266.

West, (2000). Neural Network Credit Scoring Models. *Computers and Operations Research,* **27**, 1131-1152.

Westgaard and Wijst, (2001). Default probabilities in a corporate bank portfolio: a logistic model approach. *European Journal of Operational Research,* **135**, 338-349.

Witten and Frank, (2005). *Data Mining: Practical Machine Learning Tools and Techniques,* Second edition. Morgan Kaufmann Publishers.

Xavier et al., Improving prediction accuracy of loan default - A case in rural credit. *Working paper, Institute for Financial Management and Research.*

Xu et al., (2009). Using default ARTMAP for cancer classification with microRNA expression signatures. *Proceedings of the 2009 international joint conference on Neural Networks,* 173-179.

Yang, (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval,* **1**, 69–90.

Yegorova et al., (2000). A successful loan default prediction model for small business. *Credit & Financial Management Review*, **6**.

Yegorova et al., (2001). A Successful Neural Network-Based Model for Predicting Small Business Loan Default. *The Credit and Financial Management Review*, **7**.

Zhang and Oles, (2001). Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval,* **4**, 5–31.

Zhao et al., (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications,* **36**, 2633-2644.

Zopounidis and Doumpos, (2004). *Multicriteria Decision Aid Classification Methods.* Kluwer Academic Publishers.

Zurada, (2007). Rule Induction Methods for Credit Scoring. *Review of Business Information Systems,* **11**, 11-22.