**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΕ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ ΕΛΛΗΝΙΚΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΕΛΙΣΑΒΕΤ Φ. ΚΑΝΤΑ

Διπλωματική Εργασία

Που υποβλήθηκε στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιά ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

ΠΕΙΡΑΙΑΣ

ΣΕΠΤΕΜΒΡΙΟΣ 2017

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Νικόλαος Πελάκης (Επιβλέπων)

- Χαράλαμπος Ευαγγελάρας

- Ελευθέριος Κοφίδης

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

**UNIVERSITY OF PIRAEYS**

**DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE**

**POSTGRADUATE PROGRAMM IN APPLIED STATISTICS**

# APPLICATION OF DATA MINING METHODS IN BANKING DATA OF GREEK ENTERPRISES

By

KANTA F. ELISAVET

MSc Dissertation

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

PIRAEUS, GREECE

SEPTEMBER 2017

Στην οικογένειά μου

# ***Ευχαριστίες***

Θα ήθελα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Νικόλαο Πελέκη, για την πολύτιμη βοήθειά του στην υλοποίηση της παρούσας εργασίας καθώς και για την υπομονή που έδειξε όλο αυτό το διάστημα. Επίσης ευχαριστώ θερμά τους καθηγητές κ. Χαράλαμπο Ευαγγελάρα και κ. Ελευθέριο Κοφίδη για την συμμετοχή τοτς στην τριμελή επιτροπή και υποστήριξή τους.

Επίσης θέλω να εκφράσω τις ευχαριστίες μου στους συναδέλφους μου για την συμπαράσταση και την υπομονή που έδειξαν κατά την διάρκεια της συγγραφής της εργασίας αυτής.

Τέλος, θέλω να πω ένα μεγάλο ευχαριστώ στην οικογένειά μου για την υποστήριξη και την συμπαράσταση κατά την διάρκεια των σπουδών μου και για όλα όσα μου προσέφερε αυτά τα χρόνια.

Ελισάβετ Κάντα

Πειραιάς, Σεπτέμβριος 2017

# ΠΕΡΙΛΗΨΗ

Η οικονομική κρίση που ξεκίνησε στην Ευρώπη το 2008 και διαρκεί ως και σήμερα στην Ελλάδα αναδεικνύει το πόσο επιτακτική είναι η ανάγκη παρακολούθησης του ρίσκου που αναλαμβάνει ένα χρηματοπιστωτικό ίδρυμα όταν δίνει ένα δάνειο. Ένα μοντέλο που μπορεί να προβλέψει έγκαιρα και έγκυρα δάνεια που θα καταστούν μη εξυπηρετούμενα στο άμεσο μέλλον αποτελεί εργαλείο ζωτικής σημασίας για κάθε τραπεζικό οργανισμό.

Στόχος μας είναι η περιγραφή και σύγκριση τέτοιων μοντέλων μέσω της εφαρμογής τους σε δεδομένα από μεγάλη ελληνική τράπεζα που αφορούν πελάτες-επιχειρήσεις.

Αναλυτικότερα, κατασκευάσαμε ένα μοντέλο ταξινόμησης το οποίο προβλέπει για μια επιχείρηση αν στο τέλος του επόμενου εξαμήνου θα καταστεί μη εξυπηρετούμενο δηλαδή, θα βρίσκεται σε καθυστέρηση μεγαλύτερη των 90 ημερών ή όχι.

Εφαρμόσαμε στα δεδομένα μας την μέθοδο της Λογιστικής Παλινδρόμησης (Logistic Regression), τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) και τα Δέντρα Ταξινόμησης (Classification Trees) κάνοντας χρήση είτε όλων των μεταβλητών είτε κάνοντας χρήση μόνο όσων χαρακτηρίστηκαν σημαντικές μετά από την κατάλληλη προεπεξεργασία. Συγκεκριμένα, εφαρμόστηκε η βηματική μέθοδος επιλογής μεταβλητών και η ανάλυση ευαισθησίας για τον εντοπισμό των μεταβλητών που παίζουν σημαντικό ρόλο στο εξαγόμενο αποτέλεσμα.

# ABSTRACT

The economic crisis that started in Europe in 2008 and lasts until today in Greece raises the urgent need to monitor the risk a financial institution takes when it gives a loan. A model that can predict timely and valid loans that will become inactive in the near future is a vital tool for any banking organization.

Our aim is to describe and compare such models by applying them to data from a large Greek bank that concern business customers.

More specifically, we have developed a classification model that predicts if an enterprise's loan becomes non-serving at the end of the next six months.

We have applied to our data the method of Logistic Regression, Artificial Neural Networks, and Classification Trees using either all variables or using only those that were identified as significant after proper pre-processing. Specifically, stepwise and sensitivity analysis was applied to identify the variables that play an important role in the result.

# CONTENTS

# ΚΕΦΑΛΑΙΟ 1

## 1.1 ΕΙΣΑΓΩΓΗ

Στη σημερινή κοινωνία ο όγκος καινούργιων πληροφοριών ολοένα και μεγαλώνει με αποτέλεσμα η συγκέντρωση και καταγραφή τους να δημιουργεί μια τεράστια βάση δεδομένων. Όσοι κάνουν χρήση των συγκεκριμένων δεδομένων αναζητούν μια πιο εξειδικευμένη γνώση. Την ανάγκη καταγραφής, διαχείρισης των μεγάλων βάσεων δεδομένων αλλά και ανακάλυψης καινούργιας πληροφορίας μέσω αυτών έρχεται να ικανοποιήσει η εξόρυξη γνώσης από δεδομένα. Η ιατρική, η δημογραφία, η στατιστική και η οικονομία αποτελούν κάποιους από τους κλάδους στους οποίους εφαρμόζεται η τεχνική της εξόρυξης γνώσης.

Ένα παράδειγμα στον κλάδο της οικονομίας είναι η εφαρμογή της εξόρυξης γνώσης στο τραπεζικό σύστημα. Συγκεκριμένα η τράπεζα μπορεί να πάρει απόφαση για το αν θα δανειοδοτήσει κάποιο φυσικό πρόσωπο ή μια επιχείρηση λαμβάνοντας υπόψιν την πληροφορία που θα λάβει μέσω της διαδικασίας της εξόρυξης γνώσης.

## 1.2 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ

Η εξόρυξη γνώσης από δεδομένα (data mining) αποτελεί μια προηγμένη τεχνική που στόχος της είναι η εξαγωγή χρήσιμων πληροφοριών και προτύπων που κρύβονται μέσα σε βάσεις δεδομένων. Πιο συγκεκριμένα εφαρμόζονται τεχνικές πρόβλεψης και περιγραφής σε μεγάλο όγκο δεδομένων. Η διαδικασία της ανεύρεσης γνώσης και ανάλυσης από βάσεις δεδομένων ονομάζεται Knowledge Discovery in Databases (KDD), ενώ ο όρος εξόρυξη δεδομένων αναφέρεται στις διάφορες μεθόδους που χρησιμοποιούνται για την ανάλυση αυτή. Η ανεύρεση γνώσης αποτελεί μια επαναληπτική διαδικασία μιας σειράς βημάτων τα οποία έχουν ως αποτέλεσμα την συλλογή δεδομένων και την ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά. (M.H.Dunham, 2004)

Η ανεύρεση γνώσης αποτελείται από τα ακόλουθα βήματα :

- **Καθαρισμός δεδομένων (Data cleaning)** : αφαιρείται ο θόρυβος ή τα outliers δηλαδή, όλα εκείνα τα στοιχεία που πιθανόν να επηρεάσουν το αποτέλεσμα.

- **Ενσωμάτωση δεδομένων (Data integration)** : όλα τα δεδομένα που ίσως έχουν συλλεχθεί από διαφορετικές πηγές ενσωματώνονται σε μία κοινή βάση δεδομένων.
- **Επιλογή δεδομένων (Data selection)** : επιλέγονται προσεκτικά τα δεδομένα που σχετίζονται με την ανάλυση που θα ακολουθήσει.
- **Τροποποίηση δεδομένων (Data transformation)** : τροποποιούνται τα δεδομένα που έχουν επιλεχθεί στο προηγούμενο βήμα έτσι ώστε να έχουν την κατάλληλη μορφή για την διαδικασία της εξόρυξης γνώσης.
- **Εξόρυξη δεδομένων (Data mining)** : ένα από τα πιο σημαντικά βήματα καθώς εδώ χρησιμοποιούνται διάφορες τεχνικές με σκοπό την εξαγωγή προτύπων που θα μπορούσαν να είναι πολύ χρήσιμα.
- **Αξιολόγηση προτύπων (Pattern evaluation)** : βασιζόμενοι σε συγκεκριμένα μέτρα αξιολόγησης (evaluation measures) αναγνωρίζονται πρότυπα που έχουν ενδιαφέρον και αντιπροσωπεύουν την γνώση.
- **Αναπαράσταση γνώσης (Knowledge representation)** : παρουσίαση της γνώσης που έχει ανακαλυφθεί με σκοπό να ερμηνευτούν τα αποτελέσματα της εξόρυξης δεδομένων με τον πιο κατανοητό τρόπο.

Συμπεραίνουμε λοιπόν ότι παρόλο που η εξόρυξη δεδομένων είναι μια διαδικασία-κλειδί για την ανεύρεση γνώσης αποτελεί μόνο ένα μικρό κομμάτι της όλης διαδικασίας λόγω της πολυπλοκότητάς της. Δεδομένου της επαναληπτικής φύσης της διαδικασία ανεύρεσης γνώσης υπάρχει η δυνατότητα μέσα από τροποποιήσεις των μέτρων αξιολόγησης και ενδεχόμενης εισαγωγής νέων δεδομένων, να εξαχθούν διαφορετικά και ίσως πιο κατάλληλα συμπεράσματα.

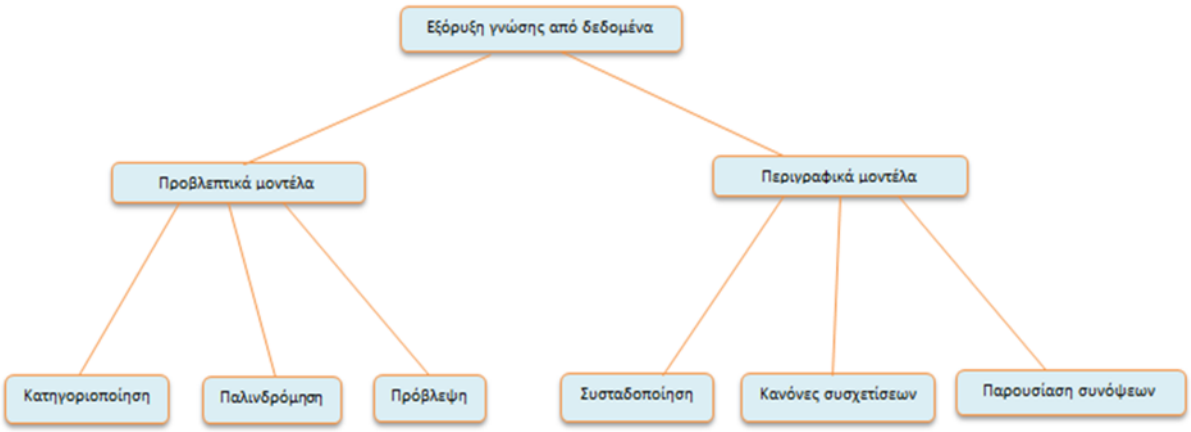
## 1.3 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η εξόρυξη γνώσης επιτυγχάνεται μέσα από ένα ευρύ φάσμα αλγόριθμων οι οποίοι χρησιμοποιούν τεχνικές από διαφορετικούς τομείς όπως είναι η στατιστική, η μηχανική μάθηση και η αναγνώριση προτύπων. Υπάρχει μια πληθώρα υπολογιστικών μεθόδων εξόρυξης γνώσης που μεταξύ άλλων περιλαμβάνουν την στατιστική ανάλυση (statistical analysis) την κατηγοριοποίηση (classification), την συσταδοποίηση (clustering), τους κανόνες συσχέτισης (association rules) και τα νευρωνικά δίκτυα (neural networks). Η χρησιμότητα τέτοιων μεθόδων οφείλεται στο γεγονός ότι συμβάλλουν στην εύρεση προτύπων, συσχετίσεων και δομών σε βάσεις δεδομένων που συνεχώς αυξάνονται.

Όπως έχουμε ήδη αναφέρει η εφαρμογή τεχνικών πρόβλεψης και περιγραφής σε μεγάλες βάσεις δεδομένων αποτελούν βασικούς στόχους της εξόρυξης γνώσης.

- Το προβλεπτικό μοντέλο χρησιμοποιεί μεταβλητές των οποίων οι τιμές είναι γνωστές με σκοπό να προβλέψει και να εκτιμήσει μελλοντικές ή άγνωστες τιμές των μεταβλητών που παρουσιάζουν ενδιαφέρον.
- Το περιγραφικό μοντέλο εστιάζει στην ανακάλυψη προτύπων αλλά και στην αναπαράσταση των δεδομένων μιας πολύπλοκης βάσης δεδομένων με σκοπό να γίνουν πιο κατανοητά και πιο εύκολα στην χρήση.

Στο επόμενο σχήμα παρουσιάζονται οι πιο συνήθεις τεχνικές εξόρυξης γνώσεων από δεδομένα χωρισμένες ανάλογα με το είδος του μοντέλου που χρησιμοποιείται κάθε φορά.




*Εικόνα 1.1 : Τεχνικές εξόρυξης γνώσης από δεδομένα*

### 1.3.1 ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

Μια από τις πλέον γνωστές και χρήσιμες μεθόδους της εξόρυξης γνώσης είναι η συσταδοποίηση (clustering). Σύμφωνα με την μέθοδο αυτή πραγματοποιείται τμηματοποίηση (partitioning) ενός συνόλου αντικειμένων από μια βάση δεδομένων σε συστάδες με τέτοιο τρόπο ώστε να υπάρχει ομοιότητα μεταξύ των στοιχείων που ανήκουν στην κάθε συστάδα. Με αυτό τον τρόπο επιτυγχάνεται η περιγραφή των δεδομένων αυτών. Στην συσταδοποίηση ο αριθμός αλλά και η σύσταση των ομάδων δεν είναι προκαθορισμένα αλλά προσδιορίζονται από τον αλγόριθμο συσταδοποίησης που θα χρησιμοποιηθεί. Αυτό είναι και το βασικό χαρακτηριστικό της συσταδοποίησης που την καθιστά διαδικασία μη εποπτευόμενης μάθησης.

Σκοπός της συσταδοποίησης είναι να δημιουργηθούν ομάδες των οποίων τα στοιχεία να παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους και ταυτόχρονα η κάθε συστάδα να διαφέρει από τις υπόλοιπες σε μεγάλο βαθμό έτσι ώστε να μην υπάρχει ο κίνδυνος

σύγχυσης. Η διαδικασία της συσταδοποίησης ανάλογα με το κριτήριο που θα χρησιμοποιηθεί οδηγεί σε διαφορετικές τμηματοποιήσεις ενός συνόλου δεδομένων. Παρακάτω περιγράφονται τα βήματα για την ανάπτυξη της διαδικασίας της συσταδοποίησης.

1. Επιλογή χαρακτηριστικών : Σκοπός είναι η επιλογή των κατάλληλων γνωρισμάτων στα οποία θα εφαρμοστεί η συσταδοποίηση. Με αυτό επιτυγχάνεται μεγαλύτερη ομοιογένεια μέσα σε κάθε συστάδα. Γίνεται αντιληπτό ότι η προ επεξεργασία των δεδομένων κρίνεται απαραίτητη πριν την εφαρμογή της διαδικασίας της συσταδοποίησης.
2. Επιλογή αλγορίθμων συσταδοποίησης : Με σκοπό τη δημιουργία ενός καλού σχήματος συσταδοποίησης (clustering scheme) για ένα σύνολο δεδομένων κρίνεται απαραίτητη η επιλογή του βέλτιστου αλγορίθμου. Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης χρησιμοποιούνται για να καθορίσουν τον αλγόριθμο που θα τμηματοποιήσει ένα συγκεκριμένο σύνολο δεδομένων με τον πιο κατάλληλο τρόπο.
3. Επικύρωση αποτελεσμάτων : Χρησιμοποιώντας τα κατάλληλα κριτήρια εξακριβώνεται η ακρίβεια των αποτελεσμάτων του αλγορίθμου συσταδοποίησης που εφαρμόστηκε. Η ποιότητα της συσταδοποίησης εξαρτάται από την ομοιότητα (μεγάλ ομοιότητα μεταξύ των γνωρισμάτων που βρίσκονται εντός μιας συστάδας και μικρή ομοιότητα μεταξύ των συστάδων) καθώς και την μέθοδο υλοποίησης της συσταδοποίησης. Ένα κριτήριο για την μέτρηση της ακρίβειας είναι η σύγκριση των αποτελεσμάτων με κάποια που υπάρχουν ήδη ή με κάποια που έχουν προκύψει από διαφορετικό τρόπο συσταδοποίησης.
4. Ερμηνεία των αποτελεσμάτων : Σαν τελευταίο βήμα είναι η εξαγωγή συμπερασμάτων μέσω των συστάδων που έχουν παραχθεί αλλά και σε συνδυασμό με άλλες αναλύσεις που πιθανόν έχουν γίνει με σκοπό εγκυρότερα αποτελέσματα.

Ανάλογα με τον ποιον αλγόριθμο συσταδοποίησης θα χρησιμοποιήσουμε κάθε φορά η μέθοδος της συσταδοποίησης διακρίνεται σε τρεις βασικές κατηγορίες :

**Μέθοδοι διαχωρισμού (Partitioning methods)**

Σύμφωνα με την μέθοδο διαχωρισμού τα δεδομένα ομαδοποιούνται έτσι ώστε κάθε ομάδα να θεωρείται ως μία κλάση και να ισχύουν οι επόμενες δύο συνθήκες : α)Σε κάθε κλάση θα πρέπει να περιέχεται τουλάχιστον ένα αντικείμενο και β)Κάθε αντικείμενο θα πρέπει να ανήκει σε μία μόνο κλάση.

**Ιεραρχικές μέθοδοι (Hierarchical methods)**

Εδώ έχουμε μια μορφή ιεραρχικών εμφωλιασμένων συσταδοποιήσεων όπου κάθε συστάδα μπορεί να περιέχει ένα αντικείμενο και άλλες συστάδες που με την σειρά τους μπορεί να περιέχουν μικρότερες συστάδες. Οι ιεραρχικοί αλγόριθμοι με την σειρά τους χωρίζονται σε δύο κατηγορίες : τους συσσωρευτικούς και τους διαιρετικούς ανάλογα με ποιον τρόπο θα γίνει η διάσπαση κάθε φορά. Οι ιεραρχικοί αλγόριθμοι μπορούν να αναπαρασταθούν με δενδρογράμματα στα οποία παρουσιάζεται ο τρόπος με τον οποίο έχει δημιουργηθεί η ιεραρχική συσταδοποίηση.

**Μέθοδοι βασισμένες σε μοντέλα (Model-based methods)**

Γίνεται η υπόθεση ότι κάθε κλάση περιγράφεται από κάποιο μαθηματικό μοντέλο και στόχος είναι να βρεθούν τα αντικείμενα που ανήκουν στην κλάση ώστε να ικανοποιούν αυτή τη μαθηματική σχέση.

Η συσταδοποίηση εφαρμόζεται σε διάφορους κλάδους. Μπορεί να εφαρμοσθεί σε περίπτωση διαχωρισμού μιας μεγάλης βάσης δεδομένων πελατών με κριτήριο παρόμοιων προτύπων αγοράς. Άλλη μια εφαρογή μπορεί να είναι όταν υπάρχει ανάγκη αναγνώρισης όμοιων προτύπων όσο αφορά την χρήση που γίνεται στο διαδίκτυο. Γενικά όπου χρειάζεται η περιγραφή δεδομένων με όμοια χαρακτηριστικά. Καθώς η συσταδοποίηση δεν είναι μέθοδος που θα χρησιμοποιηθεί στην παρούσα εργασία δεν θα την αναλύσουμε περισσότερο. Αναλυτικότερα για τη συσταδοποίηση και για τους αλγόριθμους συσταδοποίησης η βιβλιογραφία είναι πλούσια. (Μ.ΧΑΛΚΙΔΗ & ΒΑΖΙΡΓΙΑΝΝΗΣ, 2005).

### 1.3.2 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΣ (ASSOCIATION RULES)

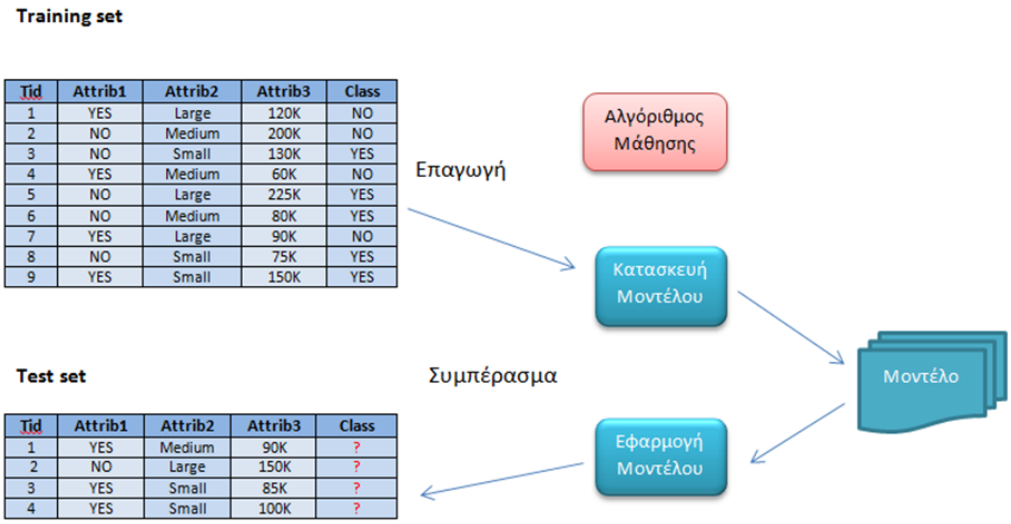
Μια από τις νεότερες και πλέον σημαντικές μεθόδους εξόρυξης γνώσης είναι η ανάλυση συσχέτισης. Η ανάλυση συσχέτισης αναζητά κρυμμένες συσχετίσεις ανάμεσα στα χαρακτηριστικά μιας μεγάλης βάσης δεδομένων. Δηλαδή, προσπαθεί να ανακαλύψει κάποιους κανόνες με σκοπό να ποσοτικοποιήσει σχέσεις μεταξύ δύο ή περισσοτέρων χαρακτηριστικών. Αυτοί οι κανόνες λέγονται κανόνες συσχέτισης. Δύο βασικά στοιχεία τους που είναι απαραίτητα για την εφαρμογή τους είναι το κατώφλι στήριξης (support threshold) και το κατώφλι εμπιστοσύνης (confidence threshold). Το κατώφλι στήριξης αναγνωρίζει όλα εκείνα τα χαρακτηριστικά μιας βάσης δεδομένων των οποίων η εμφάνιση είναι συχνή και το κατώφλι εμπιστοσύνης είναι η υπό συνθήκη πιθανότητα κάποιο χαρακτηριστικό να εμφανίζεται σε μια διαδικασία όταν εμφανίζεται και ένα άλλο. Η μορφή ενός κανόνα συσχέτισης είναι A, B, C,..□ Ω ,όπου τα A, B, C,.. είναι τα σύνολα των στοιχείων του αριστερού μέλους του κανόνα και Ω είναι το σύνολο των στοιχείων του δεξιού μέλους του κανόνα.

### 1.3.3 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ (CLASSIFICATION)

Η κατηγοριοποίηση θεωρείται από τις πιο δημοφιλείς τεχνικές που εφαρμόζεται στην ανακάλυψη προτύπων σε διάφορους κλάδους. Η κατηγοριοποίηση ή αλλιώς ταξινόμηση στηρίζεται στην εξέταση χαρακτηριστικών (attributes) κάποιου αντικειμένου (object) βάσει των οποίων αντιστοιχεί το αντικείμενο αυτό σε ένα προκαθορισμένο σύνολο κλάσεων. Η διαφορά της με την συσταδοποίηση είναι ότι το σύνολο των κλάσεων είναι προκαθορισμένο και αυτό την καθιστά ως μέθοδο μάθησης με επίβλεψη (supervised learning methods). Οι τεχνικές της ταξινόμησης χρησιμοποιούν αρχικά ένα σύνολο εκπαίδευσης (training set) τα αντικείμενα του οποίου ανήκουν σε γνωστές κλάσεις. Με αυτόν τον τρόπο ο αλγόριθμος ταξινόμησης μαθαίνει από το σύνολο εκπαίδευσης με σκοπό να κατασκευάσει ένα μοντέλο που θα του επιτρέπει να ταξινομεί τα καινούργια αντικείμενα σε κάποια από τις προκαθορισμένες κλάσεις.

Η εφαρμογή της ταξινόμησης μπορεί πρακτικά να χωριστεί σε δύο βήματα. Αρχικά, χρησιμοποιώντας το σύνολο εκπαίδευσης δημιουργείται το μοντέλο πάνω στο οποίο θα βασιστούμε για να ταξινομήσουμε τα νέα αντικείμενα που θα εισέλθουν στη βάση δεδομένων. Το δεύτερο βήμα αποτελεί την εφαρμογή του μοντέλου στο σύνολο των δεδομένων. Στην επόμενη εικόνα βλέπουμε εν συντομία την όλη διαδικασία της κατηγοριοποίησης. Καταρχήν, το σύνολο εκπαίδευσης περιέχει εγγραφές των οποίων οι κατηγορίες στις οποίες ανήκουν είναι γνωστές και σωστές. Το σύνολο εκπαίδευσης

χρησιμοποιείται για να δημιουργηθεί το μοντέλο κατηγοριοποίησης. Το μοντέλο αυτό εφαρμόζεται στο σύνολο ελέγχου (test set). Ως αποτέλεσμα έχουμε την πρόβλεψη των κατηγοριών στις οποίες ανήκει η κάθε εγγραφή του συνόλου ελέγχου. Σκοπός της παρούσας διπλωματικής είναι να προβλέψουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια το αν μια επιχείρηση θα μπορέσει να εξυπηρετήσει το δάνειο που θα πάρει από την τράπεζα.




*Εικόνα 1.2 : Διαδικασία κατηγοριοποίησης*

Το μοντέλο της κατηγοριοποίησης χρησιμοποιείται με δύο τρόπους :

- **Μοντέλο πρόβλεψης** : Με δεδομένο ότι έχουμε στην διάθεση μας χαρακτηριστικά κάποιου αντικειμένου έχουμε την δυνατότητα να προβλέψουμε σε ποια κλάση ανήκει.
- **Περιγραφικό μοντέλο** : Κυρίως χρησιμοποιείται με επεξηγηματικό τρόπο δηλαδή, για ποιο λόγο ένα αντικείμενο ανάλογα με τα χαρακτηριστικά που έχει ανήκει στην συγκεκριμένη κλάση.

Οι αλγόριθμοι κατηγοριοποίησης χωρίζονται στις ακόλουθες κατηγορίες :

- ✓ Δέντρα απόφασης
- ✓ Νευρωνικά δίκτυα
- ✓ Αλγόριθμοι κανόνων
- ✓ Αλγόριθμοι βασισμένοι στην απόσταση
- ✓ Στατιστικοί αλγόριθμοι

## 1.4 ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ ΣΤΟΝ ΤΡΑΠΕΖΙΚΟ ΤΟΜΕΑ

### 1.4.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ

Η τεχνητή νοημοσύνη και ιδιαίτερα η μηχανική μάθηση (machine learning) έχει αποκτήσει μεγάλο εύρος εφαρμογής τα τελευταία χρόνια. Κύριος στόχος είναι η αντιμετώπιση του προβλήματος της υπερ-πληροφόρησης (information overload), μέσω της ανάπτυξής συστημάτων τα οποία θα μπορούν αυτόματα τα φιλτράρουν τον ολοένα και αυξανόμενο όγκο δεδομένων, αναζητώντας σχετική πληροφορία για τον τελικό χρήστη. Πληθώρα εφαρμογών έχουν επιτυχώς αναπτυχθεί τα τελευταία χρόνια οι οποίες χρησιμοποιούν τη μηχανική μάθηση σε διάφορους τομείς όπως για παράδειγμα η ανακάλυψη γνώσης σε βάση δεδομένων (knowledge discovery in databases (KDD)) από μεγάλο όγκο βάσεων δεδομένων.

Η μηχανική μάθηση μπορεί να διακριθεί στην μάθηση με επίβλεψη (supervised learning) στη μάθηση χωρίς επίβλεψη (unsupervised learning) (Kotsiantis, 2007). Στη μάθηση χωρίς επίβλεψη, δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα μάθησης χωρίς επίβλεψη χωρίζονται σε, άγνωστες εκ των προτέρων, ομάδες με βάση τα χαρακτηριστικά τους. Χαρακτηριστικό παράδειγμα της μάθησης χωρίς επίβλεψη αποτελεί η συσταδοποίηση.

Το μεγαλύτερο τμήμα της ερευνητικής δραστηριότητας στο χώρο της μηχανικής μάθησης αφορά την μάθηση με επίβλεψη, τυπικό παράδειγμα της οποίας είναι τα προβλήματα ταξινόμησης. Κάθε παράδειγμα μάθησης αντιστοιχεί σε ένα διάνυσμα (χ1,χ2,..,χν,y), όπου χ1,χ2,..,χν είναι ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, και y είναι μια τιμή κλάσης η οποία περιγράφει ένα συγκεκριμένο γεγονός για μια θεματική περιοχή, ή αλλιώς, την έννοια στόχο. Στην παρούσα εργασία θα ασχοληθούμε με προβλήματα επιβλεπόμενης μηχανικής μάθησης.

Ένας λόγος που έχει οδηγήσει τους ερευνητές στην ανάπτυξη μη παραμετρικών τεχνικών από πεδία όπως είναι η τεχνητή νοημοσύνη είναι το γεγονός ότι οι στατιστικές και οι οικονομετρικές τεχνικές ταξινόμησης χρησιμοποιούν υποθέσεις με σκοπό να προσεγγίσουν ιδιότητες και σχέσεις που μπορεί να υπάρχουν στα δεδομένα. Ακολούθως θα περιγράψουμε κάποιες μη παραμετρικές τεχνικές που χρησιμοποιούνται στον τραπεζικό τομέα και που εφαρμόσαμε στην παρούσα εργασία.

**Νευρωνικά δίκτυα**

Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) αποτελούν μια σημαντική μέθοδο μοντελοποίησης σύνθετων προβλημάτων πρόβλεψης με μεγάλο αριθμό εξαρτημένων μεταβλητών. Μοντελοποιούνται με βάση τις λειτουργίες του ανθρώπινου εγκεφάλου. Στην πραγματικότητα, τα νευρωνικά δίκτυα είναι συστήματα επεξεργασίας πληροφορίας που αποτελούνται από ένα γράφο και διάφορους αλγόριθμους που προσπελαύνουν το γράφο.

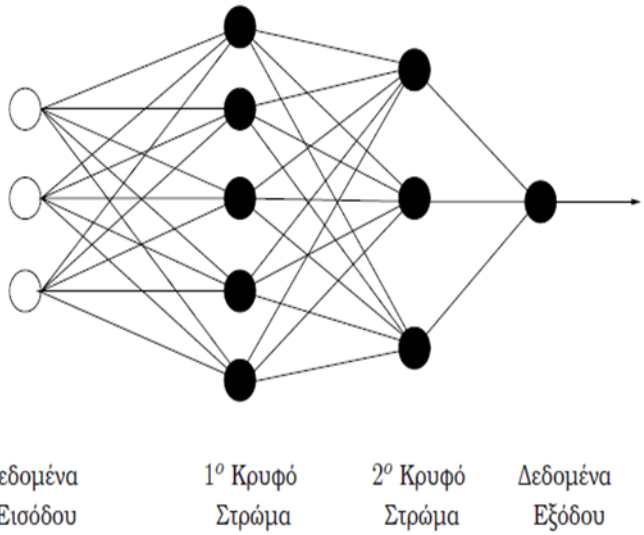
Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι.

Οι κόμβοι του γράφου είναι σαν ανεξάρτητοι νευρώνες, ενώ τα τόξα είναι σύνδεσμοι νευρώνων. Κάθε ένας από τους κόμβους είναι στοιχείο επεξεργασίας που λειτουργεί ανεξάρτητα από τους άλλους και χρησιμοποιεί μόνο τοπικά δεδομένα που καθοδηγούν την επεξεργασία. Ένα νευρωνικό δίκτυο μπορεί να θεωρηθεί σαν ένας κατευθυνόμενος γράφος οι νευρώνες του οποίου χωρίζονται σε τρεις βασικές κατηγορίες :

- Νευρώνες εισόδου (input neurons): δέχονται τις πληροφορίες που θα υποστούν επεξεργασία.
- Νευρώνες εξόδου (output neurons): καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας.
- Κρυφοί νευρώνες (hidden neurons): βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου.

Στην Εικόνα 1.3 παρουσιάζεται η διάταξη ενός πολυεπίπεδου προς τα εμπρός τροφοδοτούμενου νευρωνικού δικτύου για την περίπτωση δύο κρυφών επιπέδων. Για συντομία το δίκτυο στην Εικόνα 1.3 αναφέρεται ως ένα 3-5-3-1 δίκτυο, γιατί έχει τρεις (3) κόμβους πηγής, πέντε (5) και τρεις (3) νευρώνες στο $1^{ο}$ και στο $2^{ο}$ κρυφό επίπεδο, αντίστοιχα και ένα (1) νευρώνα εξόδου.

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους, και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου. (M.T. Hagan, 1996) (Livieris & Pintelas, 2008)




*Εικόνα 1.3: Εμπρόσθιας τροφοδότησης νευρωνικό δίκτυο πολλαπλών επιπέδων*

**Εκπαίδευση ενός Τεχνητού Νευρωνικού Δικτύου**

Η εκπαίδευση του νευρωνικού δικτύου πραγματοποιείται με συστηματική τροποποίηση των συνοπτικών βαρών μέχρι να ληφθεί η επιθυμητή έξοδος. Δηλαδή τα συνοπτικά βάρη είναι οι άγνωστοι παράμετροι που εκτιμώνται μέσω μιας διαδικασίας εκπαίδευσης. Μια από τις πιο διαδεδομένες τεχνικές εκπαίδευσης ενός νευρωνικού δικτύου ονομάζεται διάδοση (propagation).

Το νευρωνικό δίκτυο θεωρείται ότι έχει εκπαιδευτεί, όταν τα συνοπτικά βάρη, με τις τιμές που έχουν λάβει, δίδουν την επιθυμητή έξοδο. Ένας βασικός τύπος εκπαίδευσης είναι ο αλγόριθμος οπίσθιας διάδοσης του σφάλματος (backpropagation algorithm) (Rumelhart, et al., 1986) που υλοποιείται γενικά σε δύο φάσεις :

1. Φάση προώθησης : λαμβάνεται ένα σύνολο αρχικών συνοπτικών βαρών και με αυτά προσδιορίζεται η τιμή στον κόμβο εξόδου.
2. Φάση οπισθοδρόμησης : υπολογίζεται το σφάλμα στην έξοδο, δηλαδή η διαφορά μεταξύ της τιμής που υπολογίστηκε και της επιθυμητής τιμής (πραγματικής τιμής που βρίσκεται σε σύνολο δοκιμής). Στη συνέχεια το σφάλμα αυτό κατανέμεται στους κόμβους του κρυμμένου επιπέδου αναλογικά ως προς τα συνοπτικά βάρη. Με τη διαδικασία αυτή σε κάθε κόμβο εξόδου και κάθε κρυμμένο κόμβο αντιστοιχίζεται ένα σφάλμα. Στη συνέχεια το σφάλμα του κάθε κόμβου

χρησιμοποιείται για να προσαρμοστεί το εισερχόμενο στον κόμβο αυτό συνοπτικό βάρος.

Η παραπάνω διαδικασία εκπαίδευσης επαναλαμβάνεται για κάθε παράδειγμα του συνόλου εκπαίδευσης. Όταν εξαντληθεί το σύνολο εκπαίδευσης λέμε ότι συμπληρώνεται μια εποχή εκπαίδευσης (epoch). Το σύνολο εκπαίδευσης χρησιμοποιείται επαναληπτικά εώς ότου το σφάλμα στην έξοδο σταματήσει να μειώνεται. Στην βιβλιογραφία έχουν προταθεί αρκετοί αλγόριθμοι εκπαίδευσης οι οποίοι βασίζονται στον αλγόριθμο της οπίσθιας διάδοσης του σφάλματος. Στο σημείο αυτό το νευρωνικό δίκτυο θεωρείται ότι έχει εκπαιδευτεί και ότι μπορεί να αναπαράγει τις τιμές εξόδου στο σύνολο δοκιμής.

**Δέντρα Απόφασης**

Μια ευρέως χρησιμοποιούμενη μέθοδος μηχανικής μάθησης είναι εκείνη που βασίζεται σε δέντρα απόφασης (decision trees), κατά την οποία επιχειρείται η προσέγγιση μιας κατηγορικής συνάρτησης στόχου, ακολουθώντας την τεχνική του διαίρει και βασίλευε (divide and conquer). Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποιο χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δέντρο απόφασης. Τα δέντρα απόφασης έχυν χρησιμοποιηθεί ευρέως για τρεις λόγους :

1. Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δέντρο απόφασης σε ένα σύνολο κανόνων (if-then rules), προς διευκόλυνση της κατανόησής του. Έτσι, ο ταξινομητής στον οποίο καταλήγουν τα δέντρα αποφάσεων, όπως και στις μεθόδους μάθησης κανόνων, είναι η μορφή ταξινομητή πιο κοντά σην ανθρώπινη γλώσσα.
2. Η ευρωστία που επιδεικνύει ο ταξινομητής αναφορικά με το θόρυβο που ενδέχεται να παρουσιαστεί στα δεδομένα που απαρτίζουν το χώρο του προβλήματος.
3. Ένας άλλος λόγος για την χρήση των μεθόδων αυτών είναι όταν η συνάρτηση στόχος ξέρουμε ότι είναι διάζευξη συζεύξεων καθώς τα δέντρα αποφάσεων αποτελούν τέτοιες εκφράσεις.
4. Τέλος ένα από τα βασικά πλεονεκτήματα των δέντρων αποφάσεων είναι ότι δεν επηρεάζονται από τα λάθη στις τιμές των στιγμιότυπων αλλά ούτε και από την έλλειψη των τιμών τους (missing values). Αυτό οφείλεται στο γεγονός ότι για την κατασκευή των δέντρων αποφάσεων ασχολούμαστε με το σύνολο μάθησης και τα υποσύνολά του αντί να μας απασχολεί ξεχωριστά το κάθε στιγμιότυπο. Αυτό

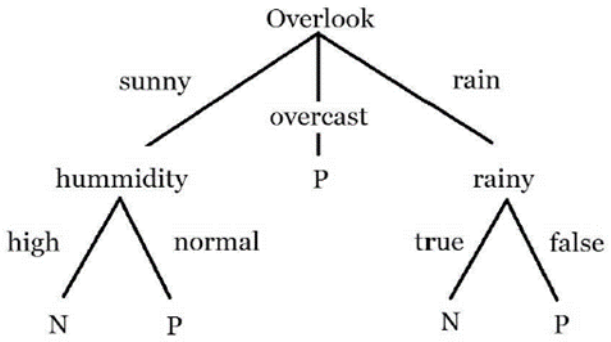
σημαίνει πως ένα λάθος σε μια τιμή ή η απουσία κάποιας τιμής δεν επηρεάζει ουσιαστικά την ανάπτυξη του δέντρου.

Η κατασκευή ενός δέντρου απόφασης από το σύνολο μάθησης είναι μια επαγωγική διαδικασία. Η περισσότερο χρησιμοποιούμενη κατηγορία επαγωγικής εκμάθησης για την κατασκευή δέντρων απόφασης είναι η επαγωγική κατασκευή δέντρων απόφασης από την κορυφή προς τα κάτω (top down induction of decision trees (TDIDT)).

Γενικά, ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από Αν-Τότε κανόνες (IF-THEN) που συνδυάζονται μεταξύ τους από την ρίζα του δέντρου προς τα φύλλα. Οι κόμβοι του δέντρου χαρακτηρίζονται με τα ονόματα των χαρακτηριστικών, οι ακμές ονομάζονται με τις δυνατές τιμές που μπορεί να πάρει ένα χαρακτηριστικό και τα φύλλα με τις διάφορες κλάσεις. Τα αντικείμενα ταξινομούνται ακολουθώντας ένα μονοπάτι που οδηγεί προς τα κάτω στο δέντρο, λαμβάνοντας τις ακμές που αντιστοιχούν στις τιμές των χαρακτηριστικών ενός αντικειμένου.

Αναλυτικότερα για την ταξινόμηση, μια εγγραφή εισέρχεται στο δέντρο από τον κόμβο της κορυφής. Στη ρίζα, εφαρμόζεται έλεγχος για να καθορισθεί ποιο παιδί θα ακολουθήσει στη συνέχεια η εγγραφή. Η επεξεργασία αυτή επαναλαμβάνεται μέχρι η εγγραφή να φτάσει στο κόμβο φύλλο. Όλες οι εγγραφές οι οποίες καταλήγουν σε ένα συγκεκριμένο φύλλο ταξινομούνται με τον ίδιο τρόπο. Υπάρχει ένα μοναδικό μονοπάτι που οδηγεί από την ρίζα σε κάθε φύλλο. Το μονοπάτι αυτό είναι μια έκφραση του κανόνα που χρησιμοποιείται για να ταξινομήσουμε τις εγγραφές.

Στην Εικόνα 1.4 παρουσιάζεται ένα παράδειγμα κάποιων αντικειμένων το οποίο περιγράφει το καιρό σε μία δεδομένη στιγμή. Κάποια ανικείμενα τα οποία είναι θετικά παραδείγματα δηλώνονται ως P και άλλα τα οποία είναι αρνητικά δηλώνονται ως N. Η ταξινόμηση στην περίπτωση αυτή είναι η κατασκευή ενός δέντρου το οποίο μπορεί να χρησιμοποιηθεί για να ταξινομήσει τα αντικείμενα με σωστό τρόπο.




*Εικόνα 1.4 : Δέντρο απόφασης*

Κατά την μάθηση, το δέντρο χτίζεται με την επαναλαμβανόμενη διάσπαση του δοσμένου συνόλου δεδομένων σύφωνα με τις διάφορες ανεξάρτητες μεταβλητές. Η «σειρά» με την οποία χρησιμοποιούνται οι ανεξάρτητες μεταβλητές στη δόμηση του δέντρου εξαρτάται από την δυνατότητα ταξινόμησης της κάθε ανεξάρτητης μεταβλητής. Υπάρχουν διάφοροι αλγόριθμοι για την επιλογή της σειράς (Murthy, 1998), αλλά ο στόχος είναι πάντα ο ίδιος, δηλαδή, να επιλέξουμε την μεταβλητή εκείνη που διαχωρίζει καλύτερα τις τελικές κλάσεις. Ο αλγόριθμος σταματά όταν φτάνει σε κόμβο από τον οποίο δεν είναι δυνατό να ξεκινήσει μία νέα διάσπαση. Τότε ο κόμβος αυτός δεν έχει παιδιά και αποτελεί φύλλο του δέντρου. Ως προς τον τρόπο ανάπτυξής τους, τα δέντρα αποφάσεων διακρίνονται σε :

- Δυαδικά : Από κάθε κόμβο διακλαδίζονται δύο νέοι κόμβοι.
- Σύνθετα : Από κάθε κόμβο διακλαδίζονται δύο ή περισσότεροι νέοι κόμβοι.

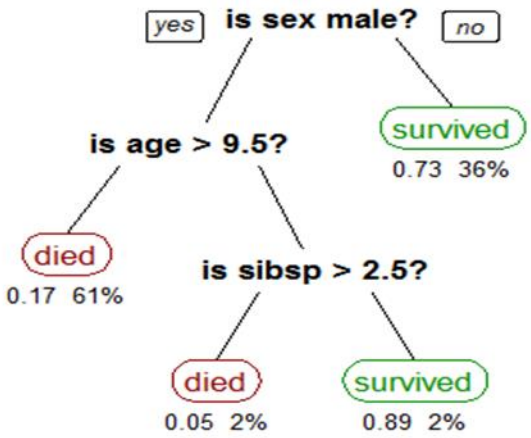
Κάθε κόμβος και προφανώς η ρίζα του δέντρου θέτουν ένα ερώτημα : «Το χαρακτηριτικό Α τι τιμή παίρνει ; » Μια εύλογη απορία είναι λοιπόν ποιο χαρακτηριστικό θα χρησιμοποιηθεί σε κάθε κόμβο. Ο τρόπος επιλογής του χαρακτηριστικού είναι μία από τις διαφορές μεταξύ των αλγορίθμων. Η επιλογή του καλύτερου διαχωριστή γίνεται βάσει μια συνάρτησης-κριτηρίου (impurity function), που εφαρμόζεται στους κόμβους που παράγονται. Κάποια μέτρα διαχωρισμού είναι τα εξής :

- Εντροπία (Entropy)
- Κριτήριο Gini
- Λάθος ταξινόμησης (Misclassification error)

**Μέθοδος CART**

Στην παρούσα εργασία μια από τις μεθόδους που θα εφαρμόσουμε είναι και η μέθοδος CART (Breiman, et al., 1984) (Classification and Regression Trees). Ο διαχωρισμός γίνεται με βάσει την εξαρτημένη μεταβλητή, αν είναι συνεχής τότε παράγεται ένα δέντρο παλινδρόμησης ενώ αν είναι διακριτή ένα δέντρο ταξινόμησης. Από τις πρώτες έρευνες οι οποίες χρησιμοποίησαν την συγκεκριμένη μεθοδολογία σε χρηματοοικονομικά προβήματα ταξινόμησης ήταν αυτές των (Marais, et al., 1985) (Frydman, et al., 1985).

Τα δέντρα απόφασης αναπαριστώνται από μια σειρά ερωτήσεων σε κάθε κόμβο με την βοήθεια των οποίων το δείγμα εκμάθησης χωρίζεται σε μικρότερα μέρη. Στην μέθοδο CART οι ερωτήσεις είναι της μορφής ναι/όχι. Ένα παράδειγμα θα μπορούσε να είναι η ερώτηση : "Είναι φύλο θηλυκό ;". Ο αλγόριθμος θα δοκιμάσει όλες τις πιθανές τιμές από όλες τις μεταβλητές που έχει διαθέσιμες για να βρει την καλύτερη «ερώτηση» που θα διαχωρίσει τα δεδομένα σε δύο μέρη πετυχαίνοντας παράλληλα την μέγιστη ομοιογένεια. Η διαδικασία αυτή επαναλαμβάνεται έως ότου σχηματιστεί το δέντρο. Ένα παράδειγμα δέντρου ταξινόμησης απεικονίζεται στην επόμενη εικόνα[1].




*Εικόνα 1.5 : Δέντρο ταξινόμησης που δείχνει την πιθανότητα επιβίωσης των επιβατών του πλοίου Τιτανικός*

Στην πράξη βέβαια μπορούν να υπάρξουν πολυπλοκότερα δέντρα που θα περιλαμβάνουν πολλά επίπεδα και ακόμα περισσότερες μεταβλητές. Ο αλγόριθμος CART έχει την δυνατότητα να διαχειριστεί και αριθμητικά και κατηγορικά δεδομένα. Γενικά η συγκεκριμένη μεθοδολογία αποτελείται από τα εξής τρία μέρη :

1. Κατασκευή του μέγιστου δέντρου
2. Επιλογή του κατάλληλου μεγέθους δέντρου

3. Ταξινόμηση νέων δεδομένων με την χρήση του δέντρου που έχει δημιουργηθεί

Δεδομένου ότι είναι γνωστή η κατηγορία στην οποία ανήκει η κάθε παρατήρηση από το δείγμα εκμάθησης μπορούμε να χρησιμοποιήσουμε τα λεγόμενα δέντρα ταξινόμησης (classification trees). Τα δέντρα ταξινόμησης χτίζονται βάσει ενός δείγματος λαμβάνοντας υπόψιν τα χαρακτηριστικά των μεταβλητών, την κατηγορία στην οποία ανήκουν στην πραγματικότητα, τις εκ των προτέρων πιθανότητες ταξινόμησής τους και τα κόστη εσφαλμένης ταξινόμησης. Το δέντρο αποτελείται από κόμβους σε καθέναν από τους οποίους αντιστοιχεί ένας κανόνας απόφασης. Όπως έχουμε ήδη περιγράψει ο σχηματισμός του δέντρου ξεκινά από έναν αρχικό κόμβο που αντιστοιχεί στην καλύτερη μεταβλητή. Η μεταβλητή αυτή επιλέγεται αφού έχουν δοκιμαστεί όλες οι μεταβλητές και όλοι οι πιθανοί κανόνες διαχωρισμού που αντιστοιχούν στην κάθε μεταβλητή. Η επιλογή του βέλτιστου διαχωριστή στην περίπτωση της μεθόδου CART για δέντρα ταξινόμησης γίνεται συνήθως με την συνάρτηση διαχωρισμού Gini. Έτσι, κάθε παραγόμενος κόμβος αντιστοιχεί σε μία κατηγορία. Η παραπάνω διαδικασία συνεχίζεται μέχρι τον σχηματισμό του τελικού δέντρου ταξινόμησης. Λόγω όμως της φύσης του συγκεκριμένου αλγορίθμου υπάρχει περίπτωση το δέντρο στο οποίο θα καταλήξουμε να είναι αρκετά μεγάλο και με τον τρόπο αυτό να καταλήξουμε στο πρόβλημα της υπερπροσαρμογής (overfitting) των δεδομένων. Αυτό σημαίνει ότι όταν θα εισέλθουν στην βάση καινούργια δεδομένα δεν θα παρέχουν ικανοποιητικά αποτελέσματα.

Για να μπορέσουμε να αποφύγουμε αυτό το πρόβλημα της υπεπροσαρμογής των δεδομένων αφαιρούμε σταδιακά κόμβους από το τελικό δέντρο (Tmax) σύμφωνα με την διαδικασία του «κλαδέματος» (pruning) και με τον τρόπο αυτό δημιουργούνται όλο και μικρότερα δέντρα. Η συγκεκριμένη τεχνική στηρίζεται στον προσδιορισμό της παραμέτρου μέτρησης της πολυπλοκότητας (cost-complexity parameter) των δέντρων αυτών. Η παράμετρος αντικατοπτρίζει την σχέση μεταξύ της πολυπλοκότητας του κάθε δέντρου και τους σφάλματός του. Ο κόμβος του δέντρου θεωρείται μέτρο μέτρησης της πολυπλοκότητάς του.

$$\alpha = \frac{E(T_i) - E(T_{i-1})}{\tilde{T}_{i-1} - 1}$$

Όπου, E(T_i) το κόστος εσφαλμένων ταξινομήσεων του νέου δέντρου, E(T_(i-1)) το κόστος εσφαλμένων ταξινομήσεων του προηγούμενου και T̃_(i-1)-1 ο αριθμός των κόμβων του καινούργιου δέντρου. Σύμφωνα με την διαδικασία pruning δημιουργείται ένα σύνολο από δέντρα τα οποία προκύπτουν από το αρχικό Tmax. Αφαιρώντας όσο

περισσότερους κόμβους γίνεται καταλήγουμε στην δημιουργία του πρώτου δέντρου T0 από το σύνολο το οποίο πρέπει να έχει την ίδια ακρίβεια (δηλαδή για το T0 , α=0). Έτσι λοιπόν για το δέντρο Ti αφαιρούνται οι κόμβοι από το δέντρο Ti-1 τόσοι ώστε η μεταβολή στην ακρίβεια να είναι η μικρότερη δυνατή. Το τελευταίο δέντρο που θα δημιουργηθεί θα έχει μόνο έναν τελικό κόμβο αλλά η παράμετρος α θα έχει την μέγιστη τιμή.

Για να μπορέσουμε να επιλέξουμε το βέλτιστο δέντρο ή από πόσους κόμβους θα πρέπει να αποτελείται το βέλτιστο δέντρο χρησιμοποιούμε την τεχνική επαναληπτικής δειγματοληψίας cross-validation (Stone, 1974). Σύμφωνα με αυτή την τεχνική παράγονται νέα δείγματα , συνήθως δέκα, από το δείγμα εκμάθησης. Σε κάθε ένα από αυτά τα νέα δείγματα εφαρμόζεται η διαδικασία του «κλαδέματος» (pruning) όπως έχουμε ήδη περιγράψει. Δηλαδή, θα δημιουργηθούν δέκα νέες αλληλουχίες δέντρων με τον ίδιο αριθμό κόμβων με τα δέντρα της αρχικής αλληλουχίας που προέκυψε μέσα από το δείγμα εκμάθησης. Τα δέντρα αυτά ομαδοποιούνται σύμφωνα με το αριθμό των κόμβων τους και υπολογίζεται η μέση ακρίβειά τους. Το βέλτιστο δέντρο επιλέγεται από την ομάδα που θα έχει το μικρότερο κόστος και κατά συνέπεια την μεγαλύτερη ακρίβεια. Περισσότερες πληροφορίες για την μεθοδολογία CART μπορούμε να ανατρέξουμε στους (Esposito, et al., 1997) και (Yohannes & Webb, 1999).

### 1.4.2 ΣΤΑΤΙΣΤΙΚΕΣ ΚΑΙ ΟΙΚΟΝΟΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ

Η στατιστική επιστήμη έχει ως αντικείμενο την ανάλυση δειγμάτων με σκοπό να εξαχθούν συμπεράσματα για ολόκληρο τον πληθυσμό. Στα πλαίσια της στατιστικής θεωρίας η ταξινόμηση (classification) αντιμετωπίζεται σαν ένα τέτοιο πρόβλημα. Ο Ficher το 1936 ήταν ο πρώτος που έθεσε τις βάσεις των πολυδιάστατων στατιστικών μεθόδων ταξινόμησης αναπτύσσοντας τη γραμμική διακριτική ανάλυση (linear discriminant analysis), η οποία μέθοδος αργότερα και στην τετραγωνική της μορφή (quadratic discriminant analysis) από τον Smith (1974) υπήρξε για δεκαετίας από τις πιο διαδεδομένες για την ανάπτυξη υποδειγμάτων ταξινόμησης.

#### Γραμμική διακριτική ανάλυση

Αν υποθέσουμε ότι έχουμε ένα σύνολο επιχειρήσεων των οποίων η κατηγορία στην οποία ανήκουν είναι ήδη γνωστή. Σύμφωνα με την γραμμική διακριτική ανάλυση αναπτύσσεται ένας γραμμικός συνδυασμός των χαρακτηριστικών τους x1, x2, x3,.., xn που έχει την ακόλουθη μορφή :

$$Z= w_0+x_1w_1+x_2w_2+x_3w_3 +\ldots+xnwn$$

, όπου $w_0$ σταθερός όρος και $w_1$, $w_2$, $w_3$,…, wn οι συντελεστές των χαρακτηριστικών των εταιρειών.

Για να εφαρμοσθεί η διακριτική ανάλυση θα πρέπει να ισχύουν δύο υποθέσεις. Η πρώτη είναι ότι οι επιδόσεις των επιχειρήσεων στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή και η δεύτερη υπόθεση είναι ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι σύμφωνα με τις οποίες υπολογίζονται ο σταθερός όρος $w_0$ και οι $w_1$, $w_2$, $w_3$,…, wn . Βάσει των παραπάνω υποθέσεων οι υπολογισμοί στη περίπτωση των δύο κατηγοριών γίνεται ως εξής :

- $w = \sum^{-1}[\mu_1 - \mu_2]$
- $w_0 = -[\mu_1 + \mu_2]' \cdot \frac{w}{2}$

, όπου μ_1, μ_2 τα διανύσματα των μέσων τιμών των χαρακτηριστικών των επιχειρήσεων που αντιστοιχούν στις κατηγορίες 1 και 2 και αντίστοιχα Σ ο πίνακας διακύμανσης-συνδιακύμανσης μεταξύ των δύο κατηγοριών.

Με τον τρόπο αυτό υπολογίζεται το σκορ διάκρισης Z (discriminant score) που αντιστοιχεί στην κάθε επιχείρηση το οποίο στη συνέχεια αφού συγκριθεί με την βέλτιστη τιμή διαχωρισμού εντάσσει την επιχείρηση σε μια από τις κατηγορίες. Για να υπολογιστεί η βέλτιστη τιμή διαχωρισμού θα πρέπει να καθοριστούν οι εκ των προτέρων πιθανότητες μια επιχείρηση να ανήκει σε κάποια κατηγορία και τα κόστη εσφαλμένων ταξινομήσεων. Συγκεκριμένα, η επιχείρηση i θα ταξινομηθεί στην κατηγορία $C_1$ αν :

$$Z1\geq \ln\frac{K_{12}P_1}{K_{21}P_2}$$

Όπου $K_{12}$ είναι το κόστος εσφαλμένης ταξινόμησης μιας επιχείρησης η οποία ενώ ανήκει στην κατηγορία $C_1$ εντάσσεται στην κατηγορία $C_2$ (όμοια για K21) και $P_1$, $P_2$ οι εκ των προτέρων πιθανότητες να ανήκει μια επιχείρηση στην κατηγορία $C_1$, $C_2$, αντίστοιχα. Το γεγονός όμως ότι ο υπολογισμός των δύο μεγεθών είναι δύσκολος πολλοί θεωρούν ίσα τα κόστοι εσφαλμένων ταξινομήσεων και τις εκ των προτέρων πιθανότητες ίδιες. Ένα ακόμα μειονέκτημα της γραμμικής διακριτικής ανάλυσης είναι η υπόθεση περί κανονικότητας κάτι που είναι δύσκολο να ισχύει σε χρηματοοικονομικά δεδομένα. Κάποιοι ερευνητές έχουν προτείνει κατάλληλους μετασχηματισμούς στις τιμές των χαρακτηριστικών με σκοπό την επίτευξη της κανονικότητας. (Balcaen & Ooghe, 2004)

**Λογιστική Παλινδρόμηση**

Πρώτος ο Olson το 1980 χρησιμοποίησε το λογιστικό υπόδειγμα πιθανότητας με σκοπό την πρόβλεψη της πτώχευσης επιχειρήσεων. Για να υπολογιστεί η πιθανότητα μια επιχείρηση να ανήκει σε κάποια από τις προκαθορισμένες κατηγορίες μπορεί να υπολογιστεί με την εφαρμογή της λογιστικής παλινδρόμησης μέσω της ανάπτυξης μιας μη γραμμικής συνάρτησης. Συγκεκριμένα γίνεται χρήση της λογιστικής συνάρτησης :

$$P_i = F(w_0 + xw_i) = \frac{1}{1 + e^{-w_0 - wx_i}}$$

Όπου, $w_0$ ένας σταθερός όρος, $w$ το διάνυσμα των συντελεστών των μεταβλητών και $x_i$ το διάνυσμα των χαρακτηριστικών της εκάστοτε επιχείρησης. Αν $P_i$ είναι η πιθανότητα η επιχείρηση i να ανήκει στην κατηγορία συνεπών επιχειρήσεων και $1$- $P_i$ η πιθανότητα να ανήκει στην κατηγορία μη συνεπών επιχειρήσεων τότε οι συντελεστές $w_0$ και $w$ υπολογίζονται με την μέθοδο μέγιστης πιθανοφάνειας μεγιστοποιώντας την συνάρτηση :

$$\ln L = \sum \ln(P_i) + \sum \ln(1 - P_i)$$

Σύμφωνα με την πιθανότητα $P_i$ η κάθε επιχείρηση ταξινομείται στην κατάλληλη κατηγορία αφού συγκριθεί με την τιμή διαχωρισμού των κατηγοριών (cut-off point). Για τον προσδιορισμό της βέλτιστης τιμής διαχωρισμού πολλοί ερευνητές ελαχιστοποιούν το συνολικό σφάλμα ταξινόμησης υποθέτοντας ότι τα κόστη εσφαλμένης ταξινόμησης I και II είναι ίσα. Αν ο αριθμός των επιχειρήσεων που ταξινομούνται ως συνεπείς στο δείγμα εκμάθησης είναι ίσος με αυτόν των μη συνεπών επιχειρήσεων τότε η τιμή διαχωρισμού είναι 0.5. Αν το πλήθος συνεπών και μη συνεπών επιχειρήσεων στο δείγμα εκμάθησης διαφέρει τότε μια σύνηθης τακτική είναι αυτή της στάθμισης των επιχειρήσεων. Υπάρχουν όμως και οι προσεγγίσεις που αποδίδουν διαφορετικά κόστη εσφαλμένης ταξινόμησης. Τέλος, η επιλογή του βέλτιστου μοντέλου γίνεται βάσει της ακρίβειας ταξινόμησης.

Το λογιστικό υπόδειγμα διαφέρει από αυτό της διακριτικής ανάλυσης στο γεγονός ότι δεν απαιτούνται στατιστικοί περιορισμοί γεγονός που επιτρέπει και την χρήση ποιοτικών μεταβλητών. Ακόμα μέσω του λογιστικού υποδείγματος είναι δυνατό να εκτιμηθεί η

σημαντικότητα του εκάστοτε χαρακτηριστικού σε σχέση με το αποτέλεσμα που θα προκύψει και πολλοί είναι οι ερευνητές που χρησιμοποιούν το λογιστικό υπόδειγμα για να επιλέξουν αρχικά ποια χαρακτηριστικά θα χρησιμοποιήσουν στην ανάλυσή τους. Ύστερα, προχωρούν στην δημιουργία ενός μοντέλου ταξινόμησης είτε χρησιμοποιώντας το λογιστικό υπόδειγμα είτε με κάποια άλλη τεχνική.

# ΚΕΦΑΛΑΙΟ 2

# ΠΙΣΤΩΤΙΚΟΣ ΚΙΝΔΥΝΟΣ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΕΚΤΙΜΗΣΗΣ ΤΟΥ

## 2.1 ΕΙΣΑΓΩΓΗ

Στο παρόν κεφάλαιο γίνεται μια εισαγωγή σε ότι αφορά τα συστήματα εκτίμησης πιστωτικού κινδύνου. Αναφέρονται βασικές έννοιες σχετικές με τα εν λόγω συστήματα και οι προδιαγραφές που πρέπει να πληρεί ένα σύστημα για να είναι αποτελεσματικό. Επίσης, παρουσιάζεται το πλαίσιο που ορίζεται από την Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας σύμφωνα με το οποίο πρέπει να εναρμονιστούν όλα τα τραπεζικά ιδρύματα για να εκτιμήσουν και να διαχειριστούν ορθά τον πιστωτικό κίνδυνο. Τέλος, γίνεται μια ανασκόπηση από διάφορες μελέτες που έχουν γίνει και βασικά μοντέλα που έχουν χρησιμοποιηθεί για την πρόβλεψη της πτώχευσης μιας επιχείρησης.

## 2.2 ΚΙΝΔΥΝΟΙ ΠΟΥ ΚΑΛΕΙΤΑΙ ΝΑ ΑΝΤΙΜΕΤΩΠΙΣΕΙ ΤΟ ΤΡΑΠΕΖΙΚΟ ΣΥΣΤΗΜΑ

Η παγκόσμια κρίση, οι ραγδαίες τεχνολογικές και πολιτικές εξελίξεις των τελευταίων χρόνων έχουν διαφοροποιήσει σε μεγάλο βαθμό τον τρόπο με τον οποίο δραστηριοποιούνται οι οικονομικοί οργανισμοί. Οι τράπεζες σαν οικονομικοί οργανισμοί δεν θα μπορούσαν να μην επηρεαστούν από το σημερινό πολύπλοκο περιβάλλον στο οποίο ζούμε. Τα τραπεζικά ιδρύματα είναι υποχρεωμένα να βρεθούν αντιμέτωπα με μία σειρά από κινδύνους οι οποίοι εξαρτώνται από διάφορους παράγοντες όπως το οικονομικό περιβάλλον της εκάστοτε χώρας στην οποία βρίσκονται, την οικονομική κατάσταση του πελατολογίου που διατηρούν, των διαφόρων προϊόντων που προσφέρουν και της παγκοσμιοποίησης.

Το πλαίσιο στο οποίο λειτουργούν τα τραπεζικά ιδρύματα έχει προσδιοριστεί με διάφορες ρυθμίσεις που έχουν εκδοθεί από την Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας σύμφωνα με τις οποίες πρέπει να λειτουργούν όλα τα πιστωτικά ιδρύματα διεθνώς. Οι τράπεζες θα πρέπει με κατάλληλες μεθόδους και διαδικασίες να εκτιμήσουν και να διαχειριστούν τους κινδύνους που μπορεί να υπάρξουν. Με αυτό τον τρόπο θα μπορούν να εξασφαλίσουν την βιωσιμότητά τους.

Οι κυριότεροι κίνδυνοι είναι οι εξής :

- **Κίνδυνος Αγοράς (Market Risk):** αφορά τον κίνδυνο μείωσης του επιπέδου τιμών της αγοράς στο σύνολο της ή σε κάποια από τα στοιχεία του ενεργητικού κάποιου επενδυτικού προϊόντος. Για παράδειγμα η μεταβολή αυτή μπορεί να οφείλεται στην αυξομείωση των επιτοκίων ή των τιμών των επενδυτικών τίτλων.
- **Πιστωτικός Κίνδυνος (Credit Risk):** είναι ο κίνδυνος που διατρέχει μια επιχείρηση ή ένας οικονομικός οργανισμός να μην εισπράξει τις απαιτήσεις του.
- **Επιτοκιακός Κίνδυνος (Interest Rate Risk)** : είναι ο κίνδυνος να μεταβληθεί η αξία μιας επένδυσης κάτι το οποίο οφείλεται σε μεταβολές στο επίπεδο των επιτοκίων.
- **Κίνδυνος Ρευστότητας (Liquidity Risk)** : οφείλεται στην αβεβαιότητα που δημιουργείται όταν κάποια επένδυση δεν μπορεί να ρευστοποιηθεί έγκαιρα. (Saunders & Cornett, 2003)

Δεδομένου ότι η αποδοτικότητα των τραπεζικών ιδρυμάτων επηρεάζεται πολύ από την απώλεια χορηγούμενων κεφαλαίων, ο πιστωτικός κίνδυνος καθιστάται ο πιο σημαντικός. Κρίσεις που είναι πιθανόν να εκδηλωθούν στο τραπεζικό σύστημα έχουν ένα μόνο κοινό στοιχείο, η χορήγηση δανείων που είναι δύσκολο να αποπληρωθούν. Αυτό έχει ως αποτέλεσμα την μείωση της αποδοτικότητας αλλά και των κεφαλαίων των πιστωτικών ιδρυμάτων. Η Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας εξέδωσε το εποπτικό πλαίσιο της Βασιλείας ΙΙ (Basel II) στο οποίο αναφέρεται η διαδικασία που πρέπει να ακολουθηθεί από τα τραπεζικά ιδρύματα για να εκτιμήσουν το συνολικό πιστωτικό κίνδυνο που αναλαμβάνουν. Παρακάτω θα αναλύσουμε το σύμφωνο αυτό και θα αναφέρουμε συστήματα εκτίμησης πιστωτικού κινδύνου. (BCBS, 2001)

## 2.3 ΕΠΙΤΡΟΠΗ ΤΡΑΠΕΖΙΚΗΣ ΕΠΙΘΕΩΡΗΣΗΣ ΤΗΣ ΒΑΣΙΛΕΙΑΣ (BASEL COMMITTEE ON BANKING SUPERVISION, BCBS)

Εκπρόσωποι των κεντρικών τραπεζών των χωρών του Καναδά, του Βελγίου, της Γαλλίας, της Γερμανίας, της Ιταλίας, της Ιαπωνίας, του Λουξεμβούργου, της Ολλανδίας, της Ελβετίας, της Σουηδίας, της Αγγλίας και των Ηνωμένων Πολιτειών της Αμερικής αποτελούν τη σύσταση της Επιτροπής Τραπεζικής Επιθεώρησης της Βασιλείας. (BCBS, 2004)

Με σκοπό να υπολογιστεί το κεφάλαιο που θα καταστεί επαρκές και ανάλογο του κινδύνου που πρόκειται να αναλάβουν οι τράπεζες εκδίδονται μια σειρά από ρυθμιστικές διατάξεις. Το περιεχόμενο αυτών των διατάξεων σχετίζεται με τις διαδικασίες που πρέπει

να εφαρμόζουν τα πιστωτικά ιδρύματα έτσι ώστε να είναι σε θέση να εκτιμήσουν και να διαχειριστούν τους κινδύνους με τους οποίους έρχονται αντιμέτωπα. Την εποπτεία για τη ορθή εφαρμογή αυτών των ρυθμιστικών διατάξεων την έχει η αρμόδια ελεγκτική αρχή της εκάστοτε χώρας. Στην Ελλάδα η αρμόδια ελεγκτική αρχή είναι η Τράπεζα της Ελλάδος.

Το εποπτικό πλαίσιο γνωστό και ως Βασιλεία ΙΙ τέθηκε σε ισχύ τον Ιούνιο του 2004 και αποτελείται από τρεις θεμελιώδεις άξονες εποπτείας γνωστούς και ως πυλώνες. Οι τρείς αυτοί πυλώνες είναι οι εξής :

**Πυλώνας 1**: σχετίζεται με τις μεθόδους που θα χρησιμοποιήσουν τα πιστωτικά ιδρύματα έτσι ώστε να προσδιοριστούν οι κεφαλαιακές απαιτήσεις ως αντιστάθμιση του κινδύνου που θα αναλάβουν από χορηγήσεις που είτε είναι επισφαλείς είτε όχι.

**Πυλώνας 2**: σχετίζεται με τον έλεγχο της διαδικασίας της αξιολόγησης της επάρκειας των κεφαλαίων και του συστήματος διαχείρισης κινδύνου που υιοθετούν τόσο τα πιστωτικά ιδρύματα όσο και η εποπτική αρχή.

**Πυλώνας 3**: σχετίζεται με την διαφάνεια και την πειθαρχία της αγοράς μέσω της δημοσιοποίησης στοιχείων έτσι ώστε να μπορεί να υπάρχει η σύγκριση του τρόπου διαχείρισης του κινδύνου από τα πιστωτικά ιδρύματα. Ακόμα, συνεπάγεται και η βελτίωση στον τρόπο που εφαρμόζονται διάφορες μέθοδοι και πρακτικές από την εκάστοτε εποπτική αρχή.

## 2.4 ΣΥΣΤΗΜΑΤΑ ΕΚΤΙΜΗΣΗΣ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ (CREDIT RATING SYSTEMS)

Το σύστημα εκτίμησης πιστωτικού κινδύνου συμβάλλει στην απόφαση που θα λάβει ένα πιστωτικό ίδρυμα για το αν θα δανειοδοτήσει ή όχι μια επιχείρηση υπολογίζοντας τον κίνδυνο που θα αναλάβει σε περίπτωση που προχωρήσει με την δανειοδότηση. Η πιστοληπτική ικανότητα των επιχειρήσεων δηλαδή, η ικανότητα να μπορούν να ανταπεξέλθουν στις δανειακές υποχρεώσεις τους προκύπτει από την έρευνα και αξιολόγηση ποιοτικών και ποσοτικών τους στοιχείων και συνδέεται άμεσα με τον πιστωτικό κίνδυνο. Η αξιολόγηση των παραπάνω στοιχείων γίνεται σήμερα στο πλαίσιο των μοντέλων που προκύπτουν μέσω των συστημάτων εκτίμησης πιστωτικού κινδύνου.

Δύο προσεγγίσεις προτείνονται από την Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας με σκοπό την εκτίμηση του πιστωτικού κινδύνου, η τυποποιημένη

(standardized approach) και η προσέγγιση των εσωτερικών αξιολογήσεων (Internal ratings-based approach, IRB). Στην παρούσα εργασία τα μοντέλα που θα αναπτυχθούν βασίζονται στην προσέγγιση των εσωτερικών αναλύσεων γι' αυτό το λόγο θα είναι και η μόνη που θα αναλύσουμε. Στηρίζεται κατά κύριο λόγο στα συστήματα που χρησιμοποιούν τα τραπεζικά ιδρύματα για να υπολογίσουν την πιθανότητα ασυνέπειας (Probability of default) της εκάστοτε επιχείρησης που έχει αιτηθεί δάνειο. Το σύστημα αυτό εξετάζει τα χαρακτηριστικά της επιχείρησης και αφού υπολογίσει την πιθανότητα αθέτησης των υποχρεώσεων της την κατατάσσει σε μια από τις προκαθοσμένες κατηγορίες του. Κάθε κατηγορία δείχνει τον βαθμό πιστωτικού κινδύνου των επιχειρήσεων που αντιστοιχούν σε αυτή. Όσο πιο υψηλός είναι ο κίνδυνος τόσο πιο δύσκολο είναι να γίνει δεκτή η αίτηση δανείου. Ακόμα, το σύστημα αυτό έχει την δυνατότητα να εκτιμήσει και τις απώλειες που μπορεί να υποστεί ο τραπεζικός οργανισμός σε περίπτωση που χορηγήσει το δάνειο και η επιχείρηση δεν μπορεί να ανταπεξέλθει στις υποχρεώσεις της.

Η εκτιμώμενη πιθανότητα αθέτησης των επιχειρήσεων-πελατών με σκοπό την κατάταξή τους στις κατηγορίες πιστωτικού κινδύνου υπολογίζεται, ανάλογα με το πόσο εξελιγμένα είναι τα συστήματα εσωτερικής διαβάθμισης του πιστωτικού ιδρύματος, με δύο διαφορετικούς τρόπους :

I. Θεμελιώδης Μέθοδος (Foundation Approach)
II. Προηγμένη Μέθοδος (Advanced Approach)

Για να υπολογιστούν οι σταθμίσεις των κινδύνων που θα δώσει το μέτρο του ρίσκου για τα τραπεζικά ιδρύματα πρέπει να εκτιμηθούν οι ακόλουθες παράμετροι :

**Πιθανότητα αθέτησης υποχρέωσης του αντισυμβαλλόμενου (Probability of Default-PD) :** Αφορά την πιστοληπτική ικανότητα της επιχείρησης που αιτείται το δάνειο και προσδιορίζει την πιθανότητα αθέτησης των υποχρεώσεων της προς το δάνειο αυτό.

**Εκτίμηση της αναμενόμενης ζημιάς (Loss Given Default-LGD) :** εκτιμά την μέση αναμενόμενη ζημιά δηλαδή, μας δίνει το ποσοστό του κεφαλαίου που δεν θα εισπραχθεί από το εκάστοτε πιστωτικό ίδρυμα λόγω της αδυναμίας της επιχείρησης-πελάτη να ανταπεξέλθει στις υποχρεώσεις της.

**Έκθεση αντισυμβαλλόμενου σε περίπτωση αθέτησης της υποχρέωσης του (Exposure at Default-EAD) :** μας δίνει το ποσό του κεφαλαίου που κινδυνεύει να χαθεί σε περίπτωση αθέτησης των υποχρεώσεων.

Εναπομένουσα διάρκεια μέχρι την λήξη των απαιτήσεων (Maturity-M) : χρησιμοποιείται για την ενσωμάτωση των απωλειών που πιθανόν να προκύψουν λόγω της διάρκειας του δανείου.

Η ποσοτικοποίηση των παραμέτρων LGD, EAD και M στην Θεμελιώδη Μέθοδο γίνεται από την Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας, ενώ της PD γίνεται από το εκάστοτε πιστωτικό ίδρυμα στηριζόμενο στο χαρτοφυλάκιό της. Στην περίπτωση της Προηγμένης Μεθόδου οι εκτιμήσεις των παραμέτρων LGD, EAD, PD και M γίνονται από την τράπεζα χρησιμοποιώντας ιστορικά δεδομένα που έχει στην διάθεσή της.

Η πιθανότητα ασυνέπειας (PD) υπολογίζεται με τρεις διαφορετικούς τρόπους (BCBS, 2001):

1. Στηριζόμενοι στην εμπειρία των πιστωτικών αναλυτών της τράπεζας
2. Κάνοντας χρήση δεδομένων από εξωτερικές πηγές όπως, οργανισμούς αξιολόγησης επιχειρήσεων (π.χ. Moody's)
3. Χρησιμοποιώντας μοντέλα αξιολόγησης (Credit Scoring Models)

Στην περίπτωση που η τράπεζα δεν έχει επαρκεί στοιχεία για άλλους πιστολήπτες των οποίων η πιστοληπτική ικανότητα είναι γνωστή και θέλει να εκτιμήσει την πιθανότητα ασυνέπειας μιας καινούργιας επιχείρησης χρησιμοποιεί έναν από τους δύο πρώτους τρόπους που αναφέραμε. Όμως, αυτές οι δύο μέθοδοι παρουσιάζουν κάποια μειονεκτήματα.

Όσον αφορά την μέθοδο στην οποία η τράπεζα βασίζεται στην γνώμη των πιστωτικών της αναλυτών, ο προσδιορισμός του προβλήματος καθώς και των παραμέτρων του γίνεται με τρόπο υποκειμενικό και αυτό έχει ως αποτέλεσμα να μην μπορεί να γίνει σωστά η αξιολόγηση των αποτελεσμάτων.

Στην δεύτερη μέθοδο υπάρχει περίπτωση τα κριτήρια με τα οποία εκτιμούν τις επιχειρήσεις και ορίζουν την ασυνέπεια οι διάφοροι οργανισμοί αξιολόγησης να διαφέρουν από αυτά που χρησιμοποιεί η τράπεζα με αποτέλεσμα η εκτίμηση της πιθανότητας της ασυνέπειας να μην έχει μεγάλη ακρίβεια. Επιπλέον, αυτοί οι οργανισμοί αξιολογούν επιχειρήσεις μεγάλων οικονομικά μεγεθών και είναι πολύ πιθανόν να μην υπάρχουν στο χαρτοφυλάκιο μιας εμπορικής τράπεζας.

Σύμφωνα με την τρίτη μέθοδο  δηλαδή, μέσω των μοντέλων αξιολόγησης δίνεται σε κάθε επιχείρηση μία βαθμολογία (score) η οποία αντιπροσωπεύει την πιθανότητα να μην ανταπεξέλθει στις υποχρεώσεις της. Σε επόμενη παράγραφο θα αναλύσουμε με περισσότερες λεπτομέρειες την εν λόγω μέθοδο.

Τέλος αφού εκτιμηθεί η πιθανότητα ασυνέπειας της επιχείρησης που έχει πάρει το δάνειο υπάρχει η δυνατότητα να υπολογιστούν και οι αναμενόμενες απώλειες (Expected Losses-EL) που θα έχει το πιστωτικό ίδρυμα στην περίπτωση που η επιχείρηση δεν μπορέσει να εκπληρώσει της υποχρεώσεις της. Οι αναμενόμενες απώλειες που ίσως προκύψουν μέσα σε ένα έτος υπολογίζονται βάσει του επόμενου τύπου (Aas, 2005) :

$$EL=PD\times EAD\times LGD$$

όπου ,

LGD (Loss Given Default): το ποσοστό του κεφαλαίου που δεν θα εισπραχθεί από το εκάστοτε πιστωτικό ίδρυμα λόγω της αδυναμίας της επιχείρησης-πελάτη να ανταπεξέλθει στις υποχρεώσεις της.

EAD (Exposure at Default): το ποσό του κεφαλαίου που κινδυνεύει να χαθεί σε περίπτωση αθέτησης των υποχρεώσεων.

## 2.5 ΕΞΩΤΕΡΙΚΑ ΣΥΣΤΗΜΑΤΑ ΑΞΙΟΛΟΓΗΣΗΣ

Οι οίκοι αξιολόγησης αποτελούν ιδιωτικές εταιρίες οι οποίες λειτουργούν συμβουλευτικά και αξιολογούν την πιστοληπτική ικανότητα μιας χώρας ή μιας επιχείρησης. Η δημιουργία των εξωτερικών συστημάτων ξεκίνησε αρχικά για τη αξιολόγηση κρατικών και επιχειρηματικών ομολόγων. Δεδομένου ότι η διαδικασία της αγοράς ενός ομολόγου είναι παρεμφερής με την χορήγηση ενός επιχειρηματικού δανείου, οι διεθνείς οργανισμοί αξιολόγησης χρησιμοποιούν αυτά τα εξωτερικά συστήματα αξιολόγησης για να αξιολογήσουν την κάθε επιχείρηση και να μπορούν να την κατατάξουν σε μία από τις προκαθορισμένες τους κατηγορίες. Από τους πιο γνωστούς οίκους είναι οι Moody’s, οι Standard & Poor’s και οι Fitch.

Στην επόμενη εικόνα παρουσιάζονται οι κατηγορίες αξιολόγησης των παραπάνω οίκων:

| | Moody's | S&P | Fitch | Meaning |
|---|---|---|---|---|
| Investment Grade | Aaa | AAA | AAA | Prime |
| | Aa1 | AA+ | AA+ | High Grade |
| | Aa2 | AA | AA | |
| | Aa3 | AA- | AA- | |
| | A1 | A+ | A+ | Upper Medium Grade |
| | A2 | A | A | |
| | A3 | A- | A- | |
| | Baa1 | BBB+ | BBB+ | Lower Medium Grade |
| | Baa2 | BBB | BBB | |
| | Baa3 | BBB- | BBB- | |
| Junk | Ba1 | BB+ | BB+ | Non Investment Grade Speculative |
| | Ba2 | BB | BB | |
| | Ba3 | BB- | BB- | |
| | B1 | B+ | B+ | Highly Speculative |
| | B2 | B | B | |
| | B3 | B- | B- | |
| | Caa1 | CCC+ | CCC+ | Substantial Risks |
| | Caa2 | CCC | CCC | Extremely Speculative |
| | Caa3 | CCC- | CCC- | In Default w/ Little Prospect for Recovery |
| | Ca | CC | CC+ | |
| | | C | CC | In Default |
| | | | CC- | |
| | D | D | DDD | |

*Εικόνα 2.1 : Κατηγορίες αξιολόγησης των Moody's, οι Standard & Poor's και οι Fitch*

Όπως έχουμε ήδη αναφέρει οι διεθνείς οργανισμοί αξιολόγησης κρίνουν επιχειρήσεις που είναι οικονομικά ισχυρές και ως επακόλουθο τα στοιχεία που χρειάζονται για την αξιολόγηση είναι εύκολο να βρεθούν.

Όσο αφορά μεσαίες επιχειρήσεις έχουν αναπτυχθεί από τους διεθνείς οργανισμούς κάποια ποσοτικά μοντέλα αξιολόγησης όπως, το RiskCalc από την Moody's και το Credit Model από την Standard & Poor's. Τα μοντέλα που ανήκουν στην κατηγορία της RiskCalc χρησιμοποιούν το κανονικό ή το λογιστικό υπόδειγμα πιθανότητας για να ταξινομήσουν τις επιχειρήσεις σε μια από τις προκαθορισμένες κατηγορίες ενώ το Credit Model χρησιμοποιεί τα Proximal Support Vector Machines.

Οι οίκοι αξιολόγησης πιστοληπτικής ικανότητας γνωστοποιούν τα κριτήρια με τα οποία αξιολογούν τις επιχειρήσεις καθώς και την κατηγορία στην οποία τις κατατάσσουν όμως δεν δίνουν πληροφορίες για την όλη διαδικασία που έχει ακολουθηθεί. Από την άλλη μεριά τα πιστωτικά ιδρύματα έχουν όλες τις πληροφορίες για τον τρόπο που λειτουργεί το σύστημα αξιολόγησης καθώς ή ανέπτυξαν ή συμμετείχαν στην ανάπτυξή του. Με αυτόν τον τρόπο έχουν μια ολοκληρωμένη άποψη και μπορούν να πάρουν μια ουσιαστική απόφαση. Αυτή είναι και μια σημαντική διαφορά μεταξύ των οίκων αξιολόγησης και των τραπεζικών οργανισμών.

## 2.6 ΜΕΘΟΔΟΙ CREDIT SCORING

Τα υποδείγματα credit scoring χρησιμοποιούνται για να αξιολογήσουν την πιστοληπτική ικανότητα μικρών επιχειρήσεων δηλαδή, με τζίρο μέχρι 2,5εκατ €. Η μέθοδος πιστωτικής βαθμολόγησης είναι κατάλληλη σε περιπτώσεις μικρών και ομοιογενών ομάδων πιστούχων που χρησιμοποιούν τραπεζικά προϊόντα μικρού ποσού όπως καταναλωτικά ή στεγαστικά δάνεια. Η συγκεκριμένη μέθοδος πραγματοποιείται μέσω ενός συστήματος αξιολόγησης του πιστωτικού ιδρύματος όπου αφού εισέλθουν στη βάση του τα απαραίτητα δεδομένα κατατάσσει τον πελάτη-δανειζόμενο σε μια από τις κατηγορίες κινδύνου.

Η βάση δεδομένων που χρησιμοποιείται θα πρέπει να αποτελείται από ποσοτικά και ποιοτικά δεδομένα για μεγάλο αριθμό επιχειρήσεων τα οποία θα αντιστοιχούν σε ένα χρονικό ορίζοντα τουλάχιστον πέντε ετών. Τα ποσοτικά δεδομένα αφορούν χρηματοοικονομικούς δείκτες (financial ratios) οι οποίοι πρέπει να αντιπροσωπεύουν την οικονομική κατάσταση της επιχείρησης και έχουν υπολογιστεί βάσει του ισολογισμού της. Τα ποιοτικά χαρακτηριστικά περιγράφουν την θέση που κατέχει η επιχείρηση στον χώρο μέσα στον οποίο δραστηριοποιείται, την οργάνωση και διοίκηση, την συνέπεια και φερεγγυότητα, τις προοπτικές ανάπτυξης που μπορεί να έχει κ.α.

Χρησιμοποιώντας αυτούς τους δείκτες και με την κατάλληλη στατιστική μέθοδο δημιουργείται ένα μοντέλο με το οποίο δίνεται στον πελάτη μια βαθμολογία (score) που τον χαρακτηρίζει «καλό» ή «κακό». Τα δεδομένα που υπάρχουν ήδη στη βάση και τα αποτελέσματα που απορρέουν από τους πελάτες που υπάρχουν μπορεί να γίνει η πρόβλεψη για την μελλοντική «συμπεριφορά» του υποψήφιου πελάτη και να παρθεί η απόφαση για το αν θα του δοθεί το δάνειο ή όχι. Για την ανάπτυξη συστημάτων πιστωτικής βαθμολόγησης υπάρχουν οι ακόλουθες βασικές προσεγγίσεις : Τα μοντέλα διακριτικής ανάλυσης, τα μοντέλα Logit και Probit και Νευρωνικά δίκτυα (Neural Network).

# ΚΕΦΑΛΑΙΟ 3

# ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

## 3.1 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Κατά το παρελθόν διάφοροι επιστήμονες μελέτησαν μοντέλα ή υποδείγματα που θα μπορούσαν να χρησιμοποιηθούν με τέτοιο τρόπο ώστε να συμβάλλουν στην πρόβλεψη της πτώχευσης μιας επιχείρησης. Για την αντιμετώπιση του προβλήματος της πτώχευσης αλλά και του πιστωτικού κινδύνου που είναι συνδεδεμένα, οι μελετητές βασίστηκαν στην προσέγγιση της ταξινόμησης των εταιρειών κατά κύριο λόγο, σε δύο ομάδες. Με τον τρόπο αυτό μια επιχείρηση είτε ανήκει στην κατηγορία των «υγειών» δηλαδή, των οικονομικά εύρωστων επιχειρήσεων, είτε στην ομάδα των «πτωχευμένων» επιχειρήσεων.

Στην βιβλιογραφία, υπάρχουν διάφορες μέθοδοι ταξινόμησης όσον αφορά τα υποδείγματα πτώχευσης επιχειρήσεων. Οι πρώτες μελέτες αφορούσαν την ανάλυση χρηματοοικονομικών δεικτών ή αναλογιών και η όποια ανάλυση γινόταν ήταν μονομεταβλητή. Αυτό σημαίνει ότι οι μελέτες επικεντρώνονταν σε ατομικές αναλογίες (ratios) και συνέκριναν τις αναλογίες των αποτυχημένων επιχειρήσεων με αυτές των υγειών.

### 3.1.1 ΜΟΝΟΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ

Το πρώτο υπόδειγμα που χρησιμοποιήθηκε για την πρόβλεψη μιας επιχείρησης κατασκευάστηκε από τον (Beaver, 1966) και στηρίχτηκε στην μονομεταβλητή ανάλυση. Στο υπόδειγμα αυτό εξετάζεται ένας συγκεκριμένος αριθμός χρηματοοικονομικών δεικτών σύμφωνα με τους οποίους αφού εξεταζόταν ένας ένας δείκτης ξεχωριστά, οι επιχειρήσεις ταξινομούνται στις δύο κατηγορίες. Αφού γίνει η σύγκριση της τιμής που λαμβάνει η επιχείρηση για έναν συγκεκριμένο χρηματοοικονομικό δείκτη με την τιμή αναφοράς του ίδιου δείκτη γίνεται η ταξινόμηση των επιχειρήσεων σε υγιείς και πτωχευμένες. Ο υπολογισμός της τιμής αναφοράς γίνεται για κάθε ένα δείκτη ξεχωριστά. Η τιμή αναφοράς ελαχιστοποιεί τα λάθη ταξινόμησης των επιχειρήσεων στις δύο ομάδες. Σύμφωνα με τον Beaver οι αριθμοδείκτες που υπολογίζονται μέσα από τα λογιστικά μεγέθη του ενεργητικού και παθητικού των επιχειρήσεων μπορούν να προβλέψουν την πορεία που θα ακολουθήσει μια επιχείρηση στο μέλλον και κατά συνέπεια να δώσουν μια εικόνα για το αν θα οδηγηθεί στην πτώχευση. Ο Beaver προσδιορίζοντας 30 δείκτες

(ratios) εφάρμοσε μονομεταβλητή διακριτική ανάλυση (univariate discriminant analysis) σε δείγμα από 79 ζεύγη επιχειρήσεων υγειών και μη. Το δείγμα επιλέχθηκε με την τεχνική κατά ζεύγη επιλογής (pair sample design) δηλαδή, σε κάθε υγιή επιχείρηση αντιστοιχεί μια πτωχευμένη. Από τους 30 δείκτες κατέληξε σε έξι οι οποίοι θεωρήθηκαν κατάλληλοι για την πρόβλεψης της πτώχευσης των επιχειρήσεων. Τέλος, στην συγκεκριμένη μελέτη αναφέρονται κάποια εμπειρικά στοιχεία σύμφωνα με τα οποία κυρίως ο δείκτης Ταμειακή ροή προς Συνολικό Χρέος (Cash Flow to Total Debt) έδινε στατιστικά σημαντικές ενδείξεις πολύ πριν την πτώχευση.

Η μονομεταβλητή ανάλυση γενικά θεωρείται απλή μέθοδος στη χρήση της αφού κάθε φορά εξετάζει έναν μόνο αριθμοδείκτη και επιπλέον δεν απαιτεί εξειδικευμένες στατιστικές γνώσεις. Όμως το γεγονός ότι το φαινόμενο της πτώχευσης μιας επιχείρησης είναι πολυδιάστατο σημαίνει ότι δεν θα μπορούσε ένας μόνο δείκτης να την προβλέψει σωστά χωρίς να ληφθούν υπόψιν και άλλοι παράγοντες δηλαδή, ένα σύνολο δεικτών οι οποίοι θα χρησιμοποιηθούν ταυτόχρονα. Έτσι η συγκεκριμένη μέθοδος είναι δυνατόν να οδηγήσει σε εσφαλμένα αποτελέσματα στην περίπτωση που γίνουν λάθη στους υπολογισμούς των σημαντικών χρηματοοικονομικών δεικτών, ή αν παραποιηθούν λογιστικά στοιχεία αλλά ακόμα και όταν υπάρξει έλλειψη δεδομένων.

Επόμενες μελέτες όπως του (Wilcox, 1970) τονίζουν ότι ο Beaver αφήνει τον αναγνώστη χωρίς μια θεωρητική εξήγηση σχετικά με το γιατί οι συγκεκριμένοι δείκτες μπορούν να προβλέψουν την πτώχευση. Έτσι, μέσα από την μελέτη του ο Wilcox δίνει τα θεμέλια ενός θεωρητικού μοντέλου το οποίο βασιζόταν σε απλά πιθανοτικά μοντέλα με σκοπό να δώσει μια εξήγηση στα εμπειρικά αποτελέσματα του Beaver. Το γεγονός ότι όλες οι μελέτες που χρησιμοποίησαν την μονομεταβλητή ανάλυση δεν κατέληξαν στους ίδιους ή όμοιους χρηματοοικονομικούς δείκτες είχε ως συνέπεια την εμφάνιση μιας δεύτερης κατηγορίας που στηρίζεται στην πολυδιάστατη στατιστική δηλαδή, την διακριτική ανάλυση (Discriminant Analysis).

### 3.1.2 ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ

Οι πολυμεταβλητές μέθοδοι εξετάζουν την προβλεπτική ικανότητα πολλών δεικτών ταυτόχρονα. Οι συγκεκριμένες τεχνικές χρησιμοποιούνται για την εφαρμογή μελετών σε πολλές διαστάσεις λαμβάνοντας υπόψη τα αποτελέσματα όλων των μεταβλητών παράλληλα. Ο Ficher, το 1936 ήταν εκείνος που παρουσίασε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης, την λεγόμενη Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis, LDA) όπως έχει περιγραφεί σε προηγούμενο κεφάλαιο. Αργότερα ο (Smith,

1946) στηριζόμενος στην μέθοδο LDA ανέπτυξε την Τετραγωνική Διακριτική Ανάλυση (Quadric Discriminant Analysis) που χρησιμοποιείται στην περίπτωση που οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών δεν είναι ίσοι. Ωστόσο, ο (Altman, 1968) ήταν εκείνος που εισχώρησε στον τομέα της πρόβλεψης της πτώχευσης των επιχειρήσεων χρησιμοποιώντας την μέθοδο της Πολυμεταβλητής Διακριτικής Ανάλυσης (Multiple Discriminant Analysis, MDA).

**Πολυμεταβλητή Διακριτική Ανάλυση-Υπόδειγμα Altman**

Το 1968 ο Altman δημοσίευσε ένα άρθρο με τίτλο «Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy». Στο συγκεκριμένο άρθρο προσπάθησε να χρησιμοποιήσει την ανάλυση δεικτών σαν μια τεχνική ανάλυσης που αφορά την πρόβλεψη πτώχευσης επιχειρήσεων. Χρησιμοποίησε δείγμα 66 βιομηχανικών εταιρειών από τις οποίες οι 33 είναι επιχειρήσεις που έχουν πτωχεύσει από το 1946 έως το 1965 και είχαν μέσο επίπεδο ενεργητικού $6,4 εκατομμύρια με ένα εύρος ενεργητικού από $0,7 εκατομμύρια μέχρι $25,9 εκατομμύρια. Οι υπόλοιπες 33 επιχειρήσεις είναι «υγιείς» και επιλέχθηκαν τυχαία σε αντιστοιχία με της πτωχευμένες την ίδια περίοδο με κριτήρια τον κλάδο της βιομηχανίας και το εύρος ενεργητικού να κυμαίνεται μεταξύ $1 και $25 εκατομμυρίων.

Ο Altman επέλεξε 22 δείκτες που πιθανόν θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη της πτώχευσης μιας επιχείρησης. Η επιλογή έγινε βάση της δημοτικότητάς τους στην αρθρογραφία την και την ενδεχόμενη συνεισφορά τους στην μελέτη του. Μέσα στους 22 αυτούς δείκτες συμπεριέλαβε κάποιους καινούργιους που εμπειρικά θεωρήθηκαν σημαντικοί. Οι δείκτες χωρίστηκαν σε πέντε ομάδες όπως, δείκτες ρευστότητας, χρηματοοικονομικής μόχλευσης, αποδοτικότητας, φερεγγυότητας και δραστηριότητας. Από τους 22 δείκτες οι 5 θεωρήθηκε ότι συνδυάζοντάς τους μπορούν να προβλέψουν καλύτερα την πτώχευση. Η τελική συνάρτηση διαφοροποίησης ήταν η εξής :

$$Z=0,012X_1+0,014X_2+0,033X_3+0,006X_4+0,999X_5$$

όπου :

X1=Κεφάλαιο Κίνησης/Σύνολο Ενεργητικού

X2=Μη διανεμόμενα Κέρδη/Σύνολο Ενεργητικού

X3=Κέρδη προ φόρων και τόκων/Σύνολο Ενεργητικού

X4=Τρέχουσα αξία μετοχών/Λογιστική Αξία Συνολικού Χρέους

X5=Πωλήσεις/Κόστος Πωληθέντων

- Όταν το Z >2,67 η επιχείρηση θεωρείται «υγιής» και δεν υπάρχει κίνδυνος αποτυχίας εντός του έτους.
- Όταν το Z< 1,81 η επιχείρηση οδηγείτε σε αποτυχία εντός του τρέχοντος έτους.
- Όταν το 1,81<Z<2,67 δεν είναι δυνατή η ταξινόμηση (grey area).

Με τον έλεγχο F-test συσχετίζεται η διαφορά μεταξύ των μέσων τιμών των δεικτών κάθε ομάδας με την μεταβλητότητα των τιμών των δεικτών κάθε ομάδας. Έτσι, ελέγχεται η διακριτική ικανότητα κάθε μιας μεταβλητής.

Το μοντέλο στο οποίο κατέληξε ο Altman κατατάσσει το 95% των επιχειρήσεων του δείγματος σωστά ένα χρόνο πριν την πτώχευση ενώ, όταν στο δείγμα εξετάζονται στοιχεία οικονομικών καταστάσεων δύο ετών πριν την πτώχευση παρουσιάζεται μείωση στην ακρίβεια της πρόβλεψης. Το υπόδειγμα του Altman ήταν η αρχή για την χρήση της τεχνικής της πολυμεταβλητής ανάλυσης διαχωρισμού στην μελέτη της πτώχευσης επιχειρήσεων.

**Πιθανοτικά Υποδείγματα Probit/Logit**

Όπως έχουμε ήδη αναφέρει και σε προηγούμενο κεφάλαιο οι μέθοδοι probit/logit για να εφαρμοστούν δεν έχουν ως προϋπόθεση αυστηρές στατιστικές υποθέσεις και για το λόγο αυτό υπερέχουν της Ανάλυσης Διαφοροποίησης και παρουσιάζουν σημαντική ακρίβεια στην πρόγνωση της πτώχευσης.

Ο (Martin, 1977) ήταν ο πρώτος που εφάρμοσε το μοντέλο logit με σκοπό την κατασκευή ενός μοντέλου έγκαιρης προειδοποίησης για την πτώχευση τραπεζικών ιδρυμάτων συναρτήσει δεικτών που προήλθαν μέσα από λογιστικές καταστάσεις τρέχουσας περιόδου. Συγκεκριμένα χρησιμοποίησε σαν δείγμα 58 τράπεζες που ήταν αποτυχημένες την περίοδο 1970- 1976 και 25 δείκτες που προήλθαν από λογιστικές καταστάσεις δύο χρόνια πριν την αποτυχία. Οι δείκτες αυτοί ανήκουν στις κατηγορίες της ρευστότητας, κεφαλαιακής επάρκειας, κερδοφορίας και asset risk.

Ο Daniel Martin κατέληξε σε ένα μοντέλο που αποτελούταν από τέσσερις δείκτες οι οποίοι αφορούσαν το έτος 1974 για πρόβλεψη για την χρονική περίοδο 1975-1976. Το μοντέλο logit κατέταξε τις αποτυχημένες τράπεζες με ακρίβεια 87% ενώ τις υγιείς με ακρίβεια 88,6%.

## 3.2 ΣΥΓΚΡΙΤΙΚΕΣ ΜΕΛΕΤΕΣ

Η τράπεζα διαθέτει μια μεγάλη βάση δεδομένων η οποία περιλαμβάνει μια πληθώρα μεταβλητών για το πιστωτικό όριο κάθε επιχείρησης-πελάτη. Μέσα από την αξιολόγηση αυτών των μεταβλητών το εκάστοτε τραπεζικό ίδρυμα έχει την δυνατότητα να ταξινομήσει τους πελάτες του και να προσδιορίσει το πως επιδρά η κάθε μεταβλητή στην εκτίμηση της πιθανότητας αθέτησης των πληρωμών τους. Η σημασία των επεξηγηματικών μεταβλητών είναι από τα πιο σημαντικά στοιχεία για την κατασκευή της σκοροκάρτας. Κάποιοι ερευνητές εξέτασαν μέσω της μελέτης τους παράγοντες που θα μπορούσαν να επηρεάσουν την αποτελεσματικότητα των μοντέλων και συγκεκριμένα το κατά πόσο η εξαίρεση των μη στατιστικά σημαντικών μεταβλητών μπορεί να επιφέρει βελτίωση στα ήδη ανεπτυγμένα μοντέλα. Οι (Fritz & Hosemann, 2000) θεωρούν ότι η ανάπτυξη διαδικασιών επιλογής των σημαντικότερων μεταβλητών οι οποίες στηρίζονται στην μέθοδο που χρησιμοποιείται φαίνεται να παρέχουν καλύτερα αποτελέσματα και χρήζουν περαιτέρω ερευνητικού ενδιαφέροντος.

Οι (Liu & Schumann, 2005) εφάρμοσαν τέσσερεις διαδικασίες επιλογής μεταβλητών και κατέληξαν στο αποτέλεσμα ότι μεγαλύτερη βελτίωση παρατηρήθηκε στο μοντέλο που δημιουργήθηκε με την μέθοδο των K-πλησιέστερων γειτόνων. Την παρατήρησή τους αυτή την στήριξαν στο γεγονός ότι οι υπόλοιπες μέθοδοι που χρησιμοποιήθηκαν έχουν ενσωματωμένη την διαδικασία της επιλογής των μεταβλητών. Συγκεκριμένα, η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα στις μεταβλητές που είναι λιγότερο σημαντικές δίνουν μικρότερα βάρη σε σχέση με τις πιο σημαντικές. Επίσης, τα δέντρα ταξινόμησης κάνουν επιλογή των πιο σημαντικών μεταβλητών για να καταλήξουν στο τελικό δέντρο.

Άλλο ένα σημαντικό θέμα στην κατασκευή μοντέλων βαθμολόγησης της πιστοληπτικής ικανότητας είναι η έμφαση που πρέπει να δοθεί στην διαδικασία έγκρισης ενός δανείου και όχι μόνο στον αν θα μπορέσει να εξυπηρετηθεί τελικά. Οι (Boyes, et al., 1989) θεωρούν ότι και οι δύο διαδικασίες είναι απαραίτητες για την κατασκευή ενός μοντέλου credit scoring. Στην ουσία υποστηρίζουν ότι παρουσιάζεται μεροληψία στην επιλογή του δείγματος που θα χρησιμοποιηθεί όταν το μοντέλο βαθμολόγησης της πιστοληπτικής ικανότητας που θα εφαρμόσει η τράπεζα στηρίζεται στην διαδικασία της έγκρισης του δανείου και μόνο.

Υπάρχουν διάφορες έρευνες που συγκρίνουν κάποιες εναλλακτικές μεθοδολογικές προσεγγίσεις που χρησιμοποιούνται με σκοπό την ανάπτυξη μοντέλων εκτίμησης του πιστωτικού κινδύνου. Το γεγονός όμως ότι σε κάποιες μελέτες συγκρίνεται ένας σχετικά μικρός αριθμός μεθοδολογιών και χρησιμοποιείται μια μικρή βάση δεδομένων οδηγεί

στην εξαγωγή αντικρουόμενων συμπερασμάτων. Ο (West, 2000) στην έρευνά του ανέπτυξε μοντέλα με σκοπό την εκτίμηση της πιστοληπτικής ικανότητας κατόχων πιστωτικών καρτών και κατέληξε στο συμπέρασμα ότι τα νευρωνικά δίκτυα και η λογιστική παλινδρόμηση ήταν οι μέθοδοι που παρείχαν καλύτερα αποτελέσματα σε σύγκριση με τα δέντρα ταξινόμησης, την διακριτική ανάλυση και τους K-πλησιέστερους γείτονες. Από την άλλη οι (Ong, et al., 2005) χρησιμοποιώντας την ίδια βάση δεδομένων κατέληξαν στο γεγονός ότι οι γενετικοί αλγόριθμοι σε σύγκριση με την λογιστική παλινδρόμηση, τα δέντρα ταξινόμησης, τα προσεγγιστικά σύνολα και τα νευρωνικά δίκτυα αποτελούν μια πιο αποτελεσματική προσέγγιση στο πρόβλημα της εκτίμησης του πιστωτικού κινδύνου.

# ΚΕΦΑΛΑΙΟ 4

## ΕΦΑΡΜΟΓΗ ΜΟΝΤΕΛΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

### 4.1 ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Η βάση δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία αποτελείται από ένα μεγάλο αριθμό εταιρειών που συναλλάσσονται με επιχειρηματικά κέντρα και έχουν τζίρο από 2 - 20 εκ. €. Τα δεδομένα αφορούν score (1-100) για ποσοτικές μεταβλητές που προέρχονται από επεξεργασία των πραγματικών τιμών συγκεκριμένων δεικτών (ratios) καθώς και απαντήσεις ποιοτικών ερωτημάτων. Το έτος στο οποίο αναφέρονται τα στοιχεία που έχουμε στην διάθεσή μας μπορεί να είναι έτος εντός τις δεκαετίας 2002-2011. Δηλαδή, για μια επιχείρηση μπορεί να υπάρχουν διαθέσιμα στοιχεία για κάποιο ή κάποια έτη της χρονικής περιόδου που αναφέραμε. Ακόμα είναι γνωστή η κατηγορία στην οποία ανήκει η κάθε εταιρεία δηλαδή, αν ανήκει στην κατηγορία των συνεπών ή μη συνεπών επιχειρήσεων.

Η βάση δεδομένων περιέχει συνολικά 15.937 «παρατηρήσεις» από τις οποίες 1.288 ανήκουν στην κατηγορία ασυνεπών και οι 14.649 στις συνεπείς. Χρησιμοποιείται ο όρος παρατηρήσεις επειδή όπως ήδη αναφέρθηκε για μια επιχείρηση μπορεί να υπάρχουν διαθέσιμα στοιχεία για δύο ή περισσότερα έτη. Σκοπός της παρούσας έρευνας είναι να προβλέψουμε στον επόμενο χρόνο την πιθανότητα κάποιος πελάτης να αθετήσει την δανειακή του υποχρέωση (δηλαδή να καθυστερήσει την πληρωμή περισσότερων των τριών δόσεων, 90 μέρες καθυστέρησης). Με αυτόν τον τρόπο μπορούμε να δημιουργήσουμε ένα μοντέλο ταξινόμησης βάσει του οποίου θα προβλέπουμε την συμπεριφορά κάθε πελάτη με όμοια δανειακά χαρακτηριστικά. Στον πίνακα 4.1 παρουσιάζεται ο αριθμός των συνεπών και ασυνεπών «παρατηρήσεων» για κάθε κλάδο βιομηχανίας.

| | Εμπόριο | Κατασκευαστικές | Βιομηχανία | Ναυτιλιακά | Κτηματομεσιτικές | Υπηρεσίες |
|---|---|---|---|---|---|---|
| Συνεπείς | 9.702 | 837 | 3.078 | 1.604 | 249 | 2.662 |
| Ασυνεπείς | 810 | 146 | 260 | 221 | 26 | 273 |
| Σύνολο | 10.512 | 983 | 3.338 | 1.825 | 275 | 2.935 |

*Πίνακας 4.1: Σύσταση βάσης δεδομένων*

Για την ανάλυση χρησιμοποιήθηκαν χρηματοοικονομικοί δείκτες που έχουν προκύψει από λογιστικά στοιχεία των επιχειρήσεων και αποτελούνται κυρίως από στοιχεία κερδοφορίας, ρευστότητας και αποδοτικότητας αλλά και δείκτες που έχουν προκύψει από ποιοτικά κριτήρια τα οποία σχετίζονται με την οργάνωση των επιχειρήσεων, την φήμη που διατηρούν, τις προοπτικές εξέλιξης και ανάπτυξης που έχουν κ.ά. Στον πίνακα που ακολουθεί παρουσιάζονται οι ανεξάρτητες μεταβλητές (δείκτες) που χρησιμοποιήθηκαν στην ανάλυση καθώς και μια συνοπτική περιγραφή τους.

| | Μεταβλητές | Περιγραφή |
|---|---|---|
| X1 | **Liquidity Ratio** | δείχνει συνοπτικά την ικανότητα της επιχείρησης να αποπληρώνει τις υποχρεώσεις της, χωρίς να διαταράσσεται η έκρυθμη λειτουργία της |
| X2 | **Current Ratio** | δείκτης που είναι ίσος με τον λόγο του κυκλοφορούντος ενεργητικού προς τις βραχυπρόθεσμες υποχρεώσεις. Μια υψηλή τιμή δείχνει ισχυρή ικανότητα της εταιρείας να αντιμετωπίζει τις βραχυπρόθεσμες υποχρεώσεις της από περιουσιακά στοιχεία τα οποία θα μπορούσαν να ρευστοποιηθούν την περίοδο που αναφέρονται οι συγκεκριμένες υποχρεώσεις. |
| X3 | **ARDAYS Ratio** | μέσος όρος ημερών που οι εισπρακτέοι λογαριασμοί της επιχείρησης είναι πολύ μεγάλοι. Όσο μεγαλύτερος είναι ο αριθμός τόσο μεγαλύτερη η πιθανότητα ασυνέπειας προς τους πιστωτές. Σκοπός είναι να προσδιοριστεί η αποτελεσματικότητα των |
| X4 | **Quick Ratio** | μετράει τον βαθμό στον οποίο οι βραχυπρόθεσμες υποχρεώσεις της εταιρείας καλύπτονται από τα περρισότερα και πιο πρόσφατα ρευστοποιημένα περιουσιακά στοιχεία. Είναι ένας δείκτης όμοιος με τον δείκτη κυκλοφοριακής ρευστότητας με την διαφορά ότι έχουν αφαιρεθεί τα αποθέματα της επιχείρησης τα οποία δεν μπορούν να ρευστοποιηθούν εύκολα και χωρίς να υπάρξει ζημία. |
| X5 | **Inventory Days Ratio** | υπολογίζει το χρονικό διάστημα (ημέρες) που τα εμπορεύματα παραμένουν ως απόθεμα στην επιχείρηση μέχρι την στιγμή της πώλησης ή διαφορετικά το χρονικό διάστημα που απαιτείται μέχρι να ανανεωθούν τα αποθέατα της επιχείρησης. Χρησιμοποιείται σαν μέτρο των δυντοτήτων που έχει η εταιρεία στις βραχυπρόθεσμες πωλήσεις και με τον τρόπο αυτό ελέγχεται η υπερθεματοποίηση η οποία είναι πιθανόν να προκαλέσει προβλήματα στην οικονομική εξέλιξη της εταιρείας. Ένας αριθμός πάνω από τον κανόνα βιομηχανίας δείχνει πρόβλημα με τις προβλέψεις πωλήσεων ενώ ένας αριθμός κάτω από τον κανόνα δείχνει απώλεια πωλήσεων που οφείλονται στην αδυναμία της εταιρείας να καλύψει την ζήτηση. |
| X6 | **APDAYS Ratio** | μετρά την μέση χρονική διάρκεια συναλλαγών χρέους (creditor days). Είναι το χρονικό διάστημα που χρειάζεται η επιχείρηση για να εκκαθαρίσει το βραχυπρόθεσμο χρέος. Καθορίζει αν η εταιρεία μπορεί να ανταπεξέλθει αποτελεσματικά στις βραχυπρόθεσμες υποχρεώσεις που έχει. |
| X7 | **Sales to Working Capital Ratio** | εκτίμηση τογ δείκτη των πωλήσεων σε κεφάλαιο κίνησης. Αποτελεί μέτρο τόσο για την αποδοτικότητα της επιχείρησης όσο και για την βραχυπρόθεσμη οικονομική της υγεία. Μια χαμηλή τιμή υποδηλώνει αποτελεσματική χρήση των κεφαλαίων κίνησης, ενώ μια υψηλή τιμή μπορεί να υποδεικνύει πωλήσεις που υπερβαίνουν τον όγκο πραγωγής ειδικά όταν συνδυάζεται με χαμηλές ή αρνητικές ταμειακές ροές από λειτουργικές δραστηριότητες. |
| X8 | **Operating Ratio** | συνοψίζει την χρηματοοικονομική επίδοση της εταιρείας δηλαδή, δείχνει την αποτελεσματικότητα της διαχείρισης μιας επιχείρησης , συγκρίνοντας λειτουργικά έξοδα σε καθαρές πωλήσεις . Όσο μικρότερος είναι ο δείκτης , τόσο μεγαλύτερη είναι η ικανότητα του οργανισμού να παράγει κέρδη δηλαδή, έχει μεγαλύτερη αποδοτικότητα |
| X9 | **Gross Profit Ratio** | δείχνει το μεικτό κέρδος που έχει η επιχείρηση αφού έχει αφαιρεθεί το κόστος πωληθέντων. Μέσω του συγκεκριμένου δείκτη μας δίνεται η δυνατότητα να γνωρίζουμε την λειτουργική αποτελεσματικότητα της επιχείρησης. Μια μεγάλη τιμή δείχνει ότι η εν όγω επιχείρηση έχει την δυνατότητα να καλύπτει τα λειτουγικά και άλλα έξοδά της. |
| X10 | **Operating Margin Ratio** | είναι ο λόγος των λειτουργικών εσόδων προς τις καθαρές πωλήσεις της εταιρείας και ουσιαστικά δίνει μια καλή εικόνα όσο αφορά την κερδοφορία της επιχείρησης. Χρησιμοποιείται για την μέτρηση της στρατιγικής της τιμολόγησης της εταιρείας και για την λειτουργική της αποδοτικότητα.Ένα υγιές λειτουργικό είναι απαραίτητο για να είναι σε θέση η εταιρεία να πληρώσει τα πάγια έξοδά της, τόκοι επί του χρέους κλπ. |
| X11 | **NPBTSALES Ratio** (net profit before tax/sales (%)) | δείκτης που δείχνει τα καθαρά κέρδη που πετυχαίνει η επιχείρηση από τις πωλήσεις της. |
| X12 | **NPBTTA Ratio** (net profit before tax/total assets (%)) | δείκτης καθαρών κερδών προ φόρου για το σύνολο του ενεργητικού της εταιρείας. Παρέχει ένα μέτρο της ικανότητας της διοίκησης να αξιοποιήσει τα περιουσιακά στοιχεία της εταιρείας. |
| X13 | **NPBTTNW Ratio** (net profit before tax/tangible net worth) | δείκτης καθαρών κερδών προ φόρου σε πάγια καθαρή θέση. Μια υψηλή απόδοση υποδεικνύει γενικά αποτελεσματική διαχείριση |

| | Μεταβλητές | Περιγραφή |
|---|---|---|
| X14 | **Sales to total assets Ratio** | αξιολογεί την ικανότητα της εταιρείας να παράγει επαρκείς πωλήσεις σε σύγκριση με το επίπεδο των περιουσιακών στοιχείων |
| X15 | **Sales Growth Ratio** | εκτίμηση του ιστορικού ρυθμού αύξησης των πωλήσεων. Λαμβάνεται υπ'όψιν και ο βαθμός μεταβλητότητας στην αλλαγή οτυ ρυθμού αυτού με την πάροδο του χρόνου. |
| X16 | **Capital Structure Ratio** | συνοψίζει την θέση ισχύος της επιχείρησης. Το πως μια επιχείρηση χρηματοδοτεί την συνολική λειτουργία και την ανάπτυξή της με χρήση διαφορετικών πηγών χρηματοδότησης. |
| X17 | **DebtTNW Ratio** (the debt to tangible net worth) | λόγος των υποχρεώσεων της επιχείρησης προς την καθαρή θέση της. Συγκρίνει το πόσο χρέος μεταφέρει η εταιρεία σε σχέση με το πόση καθαρά αξία έχει. Είναι ο βαθμός προστασίας που παρέχεται από τους ιδιωκτήτες στους πιστωτές. Όσο μεγαλύτερη είναι η αναλογία τόσο μεγαλύτερος είναι ο κίνδυνος που αναλαμβάνουν οι πιστωτές και πιο ευάλωτη είναι η εταιρεία σε περιόδους εντάσεων στις χρηματοπιστωτικές αγορές. Χαμηλή αναλογία δείχνει μεγαλύτερη μακροπρόθεσμη οικονομική ασφάλεια. |
| X18 | **Debt Coverage Ratio** | συνολική αξιολόγηση της ικανότητας της εταιρείας να ανταποκριθεί στις απαιτήσεις εξηπυρέτησης χρέους από τα κέρδη και τις ταμειακές ροές. (Κατά πόσο είναι ο δανειολήπτης σε μεγάλο χρέος) |
| X19 | **Cash Impact of Management Variables Ratio** | αποτελεί δείκτη αποτελεσματικότητας διοίκησης ροών από λειτουργικές δραστηριότητες. |
| X20 | **Cash Flow Coverage Ratio** | μετρά την ικανότητα της εταιρείας να χρηματοδοτήσει τις ανάγκες εξυπηρέησης χρέους. |
| X21 | **Earnings Coverage Ratio** | μετρά το επίπεδο στο οποίο οι απαιτήσεις εξυπηρέτησης χρέους καλύπονται από τα προσαρμοσμένα κέρδη της εταιρείας |
| X22 | **Industry Risk Ratio** | συνοψίζει τις διάφορες προκλήσεις και κινδύνους που αντιμετωπίζει η εταιρεία από το εξωτερικό περιβάλλον. Επηρεάζεται από το επίπεδο ευαισθησίας σε οικονομικούς και πολιτικούς παράγοντες. |
| X23 | **Economic Conditions Ratio** | βασίζεται στην υποκειμενική αξιολόγηση των κινδύνων που σχετίζονται με τις οικονομικές ευαισθησίες που υπάρχουν στο κλάδο |
| χ24 | **Structural Factors Ratio** | αποτελείται από κινδύνους που σχετίζονται με τη βιομηχανία σε σχέση με την ανταγωνιστική δομή και το περιβάλλον του κλάδου |
| X25 | **Industry Performance Ratio** | βασίζεται στην υποκειμενική αξιολόγηση των κινδύνων που σχετίζονται με τη συνολική απόδοση λειτουργίας της βιομηχανίας όσον αφορά την ανάπτυξη , την κερδοφορία και τη σταθερότητα |
| X26 | **Management Quality Ratio** | είναι η συνολική αξιολόγηση της ικανότητας της ομάδας διαχείρισης. Η εκτίμηση της ποιότητας της διαχείρισης προέρχεται από την αξιολόγηση των οργανωτικών ικανοτήτων, την εμπειρία και τον προγραμματισμό διαδοχής |
| X27 | **Management Organization Ratio** | είναι μια συνολική αξιολόγηση των γενικών παραγόντων διαχείρισης. |
| X28 | **Company Reputation Ratio** | Φήμη |
| X29 | **Company Status Ratio** | προφίλ εταιρείας |

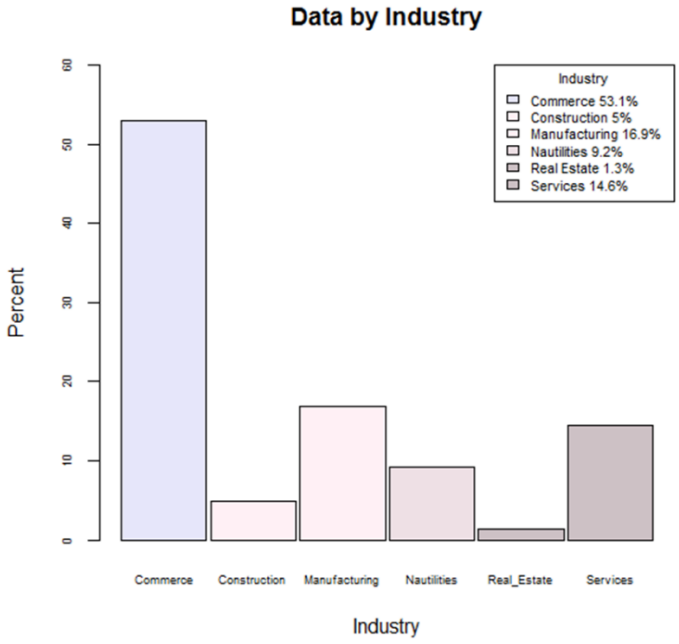
*Πίνακας 4.2: Πίνακας Δεδομένων*

Τέλος η μεταβλητή απόκρισης είναι η "Status" η οποία παίρνει τιμές 1 και 0 αν η επιχείρηση-παρατήρηση χαρακτηρίζεται "Ασυνεπής" ή "Συνεπής" αντίστοιχα.

Η ανάλυση πραγματοποιήθηκε λαμβάνοντας δείγμα από την βάση δεδομένων στο σύνολό της αλλά και χωρίζοντας την ανάλογα με τον επιχειρηματικό κλάδο που ανήκει η εκάστοτε εταιρεία. Αρχικά εφαρμόσαμε τα μοντέλα ταξινόμησης στο δείγμα εκμάθησης (train set) που καλύπτει το 75% του συνόλου των δεδομένων και η ύστερα η αξιολόγησή τους έγινε με την εφαρμογή τους στο δείγμα ελέγχου (test set) δηλαδή, στο υπόλοιπο 25%. Σε επόμενη φάση, η βάση δεδομένων χωρίστηκε σε έξι διαφορετικές ομάδες και σε κάθε μια από αυτές εφαρμόστηκαν οι αλγόριθμοι ταξινόμησης. Ο διαχωρισμός αυτός έγινε με σκοπό να γίνει αντιληπτό αν η ανάπτυξη διαφορετικών συστημάτων εκτίμησης του πιστωτικού κινδύνου για κάθε κλάδο βιομηχανίας ξεχωριστά, οδηγεί στην αποτελεσματικότερη αξιολόγηση των επιχειρήσεων.

Στο επόμενο γράφημα παρουσιάζεται το ποσοστό των επιχειρήσεων που αντιστοιχεί σε κάθε κλάδο βιομηχανίας. Παρατηρούμε ότι η βάση δεδομένων του εμπορικού κλάδου ξεπερνάει το 50% της συνολικής βάσης και αντίστοιχα η βάση δεδομένων των κτηματομεσιτικών επιχειρήσεων καταλαμβάνει μόνο το 1.3% . Κατά τον τρόπο αυτό

καθίσταται δυνατό η εξέταση της αποτελεσματικότητας των μεθόδων ταξινόμησης συναρτήσει του μεγέθους των δεδομένων.




*Γράφημα 4.1: Κατανομή δεδομένων ανά κλάδο βιομηχανίας*

### 4.1.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η εκάστοτε βάση δεδομένων που έχουμε στην διάθεσή μας για την επίλυση διάφορων προβλημάτων μπορεί να περιέχει ελλιπή ή δεδομένα με θόρυβο. Για το λόγο αυτό, πριν εφαρμόσουμε τον οποιοδήποτε αλγόριθμο θα πρέπει να προβούμε στην προ επεξεργασία των δεδομένων που υπάρχουν διαθέσιμα. Η προεπεξεργασία των δεδομένων αποτελείται από κάποια στάδια όπως ο καθαρισμός δεδομένων, ο μετασχηματισμός τους με στόχο την δημιουργία μεταβλητών που θα συμβάλλουν στην καλύτερη περιγραφή του προβλήματος και με τον τρόπο αυτό να οδηγήσουν στην καλύτερη προσαρμογή του μοντέλου, την μείωση της διάστασής τους αλλά και την συμπλήρωσή τους.

**Διαχείριση ελλειπουσών τιμών**

Σαν ελλείπουσες τιμές χαρακτηρίζονται αυτές οι οποίες για τον οποιαδήποτε λόγο δεν έχουν καταγραφεί. Οι ελλείπουσες τιμές μπορεί να αποδειχθούν ένα όχι και τόσο ασήμαντο πρόβλημα κατά την ανάλυση ενός συνόλου δεδομένων και η αντιμετώπισή τους δεν είναι συνήθως τόσο απλή.Τα ελλιπή δεδομένα χωρίζονται στις επόμενες δύο κατηγορίες :

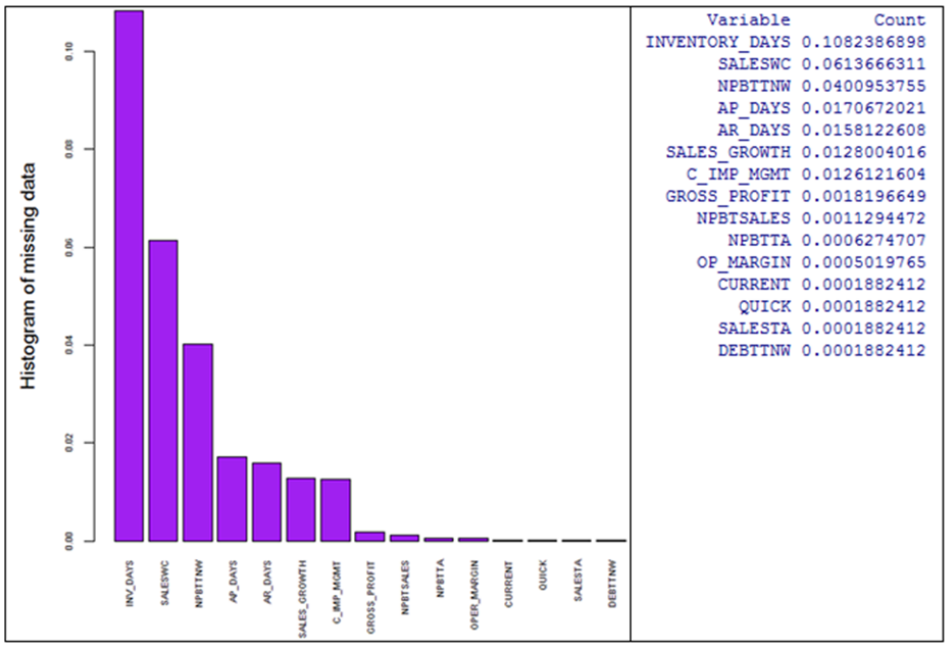
- MCAR : η έλλειψη των δεδομένων είναι εντελώς τυχαία και αυτό είναι το επιθυμητό σενάριο. Σε αυτή την κατηγορία τα ελλιπή δεδομένα ανάλογα με τον ποσοστό που καλύπτουν δύναται να διαγραφούν ή να συμπληρωθούν με την κατάλληλη μέθοδο.
- MNAR : η έλλειψη των δεδομένων δεν είναι τυχαία ίσως οφείλεται σε συγκεκριμένους παράγοντες και σε αυτή την περίπτωση θα πρέπει να ελεγχθεί περαιτέρω η διαδικασία συλλογής δεδομένων και να γίνει προσπάθεια κατανόησης της ελλιπής πληροφορίας.

Αν η ποσότητα των δεδομένων που λείπουν είναι πολύ μικρή σε σχέση με το μέγεθος του συνόλου δεδομένων, τότε διαγράφοντας κάποια χαρακτηριστικά με ελλείπουσες τιμές ίσως είναι η καλύτερη στρατηγική για να αποφευχθεί η μεροληψία στην ανάλυση. Παρόλα αυτά αφήνοντας κάποια δεδομένα εκτός της ανάλυσης είναι πιθανόν να στερηθούμε σημαντική πληροφορία η οποία δυνητικά θα μπορούσε να χρησιμοποιηθεί ανάλογα με το εκάστοτε πρόβλημα. Η απώλεια δεδομένων έχει ως αποτέλεσμα μεγαλύτερα τυπικά σφάλματα , ευρύτερα διαστήματα εμπιστοσύνης αλλά και την μείωση της ισχύος στους ελέγχους υποθέσεων και ίσως θα ήταν καλύτερο να καταφύγουμε σε διάφορες μεθόδους αντιμετώπισης των ελλειπουσών τιμών. Όσο αφορά τις ποιοτικές μεταβλητές η αντικατάσταση ελλειπουσών τιμών από παρατηρούμενες δεν συνίσταται. Μία λύση που θα μπορούσε να χαρακτηριστεί πρακτική είναι ο καθορισμός ακόμα μιας κατηγορίας για την μεταβλητή η οποία θα μπορούσε να συμπεριληφθεί στην ανάλυση. Όσο για τις ποσσοτικές μεταβλητές κάποιες αποτελεσματικές μέθοδοι αντιμετώπισης των ελλειπουσών τιμών είναι οι εξής (Han & Kamper, 2001) :

- Συμπλήρωση της ελλείπουσας τιμής με την μέση τιμή της μεταβλητής της συγκεκριμένης παρατήρησης.
- Σε ένα πρόβλημα ταξινόμησης κατηγοριοποιώντας την βάση δεδομένων σύμφωνα με τις κατηγορίες μιας μεταβλητής η ελλείπουσα τιμή θα μπορούσε να καλυφθεί με την μέση τιμή της κατηγορίας που ανήκει.
- Χρήση τις πιο πιθανής τιμής για την συμπλήρωση της ελλείπουσας τιμής. Για να εκτιμηθεί η πιθανή τιμή μπορεί να γίνει χρήση εργαλείων που βασίζονται στην στατιστική και χρησιμοποιούν τύπους του Bayes, ή μέσω της κατασκευής ενός δέντρου απόφασης.

Στο επόμενο γράφημα παρουσιάζονται οι μεταβλητές με το αντίστοιχο ποσοστό ελλειπουσών τιμών. Η μεταβλητή με το μεγαλύτερο ποσοστό ελλιπών δεδομένων είναι η

INVENTORY_DAYS, συγκεκριμένα 10.82% με την μεταβλητή SALESWC να ακολουθεί με ποσοστό 6.13%.

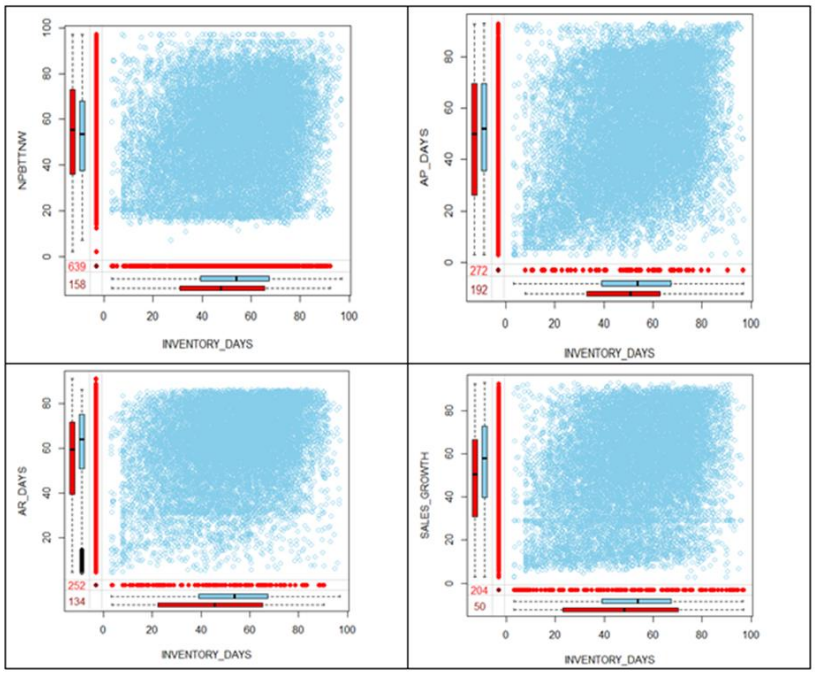



| Variable | Count |
|---|---|
| INVENTORY_DAYS | 0.1082386898 |
| SALESWC | 0.0613666311 |
| NPBTTNW | 0.0400953755 |
| AP_DAYS | 0.0170672021 |
| AR_DAYS | 0.0158122608 |
| SALES_GROWTH | 0.0128004016 |
| C_IMP_MGMT | 0.0126121604 |
| GROSS_PROFIT | 0.0018196649 |
| NPBTSALES | 0.0011294472 |
| NPBTTA | 0.0006274707 |
| OP_MARGIN | 0.0005019765 |
| CURRENT | 0.0001882412 |
| QUICK | 0.0001882412 |
| SALESTA | 0.0001882412 |
| DEBTTNW | 0.0001882412 |

*Γράφημα 4.2: Ποσοστό ελλειπουσών τιμών*

Εν συνεχεία θα προσπαθήσουμε να εντοπίσουμε γραφικά τυχόν συγκεκριμένες μορφές (patterns) ελλιπών δεδομένων με σκοπό την ταξινόμησή τους σε μία από τις κατηγορίες MCAR ή MNAR. Ενδεικτικά παρουσιάζουμε κάποια ειδικά θηκογράμματα που συγκρίνουν τις μεταβλητές με τα ελλιπή δεδομένα ανά δύο έτσι ώστε να μπορέσουμε να διακρίνουμε αυτές τις ειδικές μορφές.

Σε κάθε ένα από τα επόμενα γραφήματα το κόκκινο θηκόγραμμα απεικονίζει την κατανομή των ελλιπών δεδομένων της κάθε μεταβλητής ενώ το γαλάζιο θηκόγραμμα την κατανομή των υπολοίπων. Το γεγονός ότι τα κόκκινα και γαλάζια θηκογράμματα είναι πολύ παρόμοια οδηγούν στην παραδοχή ότι τα δεδομένα μας ανήκουν στην κατηγορία MCAR, δηλαδή η έλλειψή τους είναι τυχαία και μπορούν να αντικατασταθούν. Η μέθοδος που επιλέχθηκε για να αντικατασταθούν τα ελλιπή δεδομένα είναι η μέθοδος καταλογισμού (imputation method). Κάθε ελλιπής τιμή αντικαθίσταται από μια παρατηρούμενη λαμβάνοντας υπ'όψιν τιμές που υπάρχουν σε παρόμοια χαρακτηριστικά ή παρατηρήσεις. Με το τρόπο αυτό επιλέγεται τυχαία μια τιμή από αυτές για να συμπληρωθεί η ελλιπής τιμή.




*Γράφημα 4.3: Κατανομή ελλιπών δεδομένων*

### 4.1.2 ΜΗ ΙΣΟΡΡΟΠΗΜΕΝΑ ΔΕΔΟΜΕΝΑ

Πριν την εφαρμογή των μοντέλων ταξινόμησης θα πρέπει να γίνει αναφορά στο πρόβλημα της μη ισορροπίας των δεδομένων που έχουμε στην διάθεσή μας. Συγκεκριμένα, ο αριθμός των παρατηρήσεων της ομάδας των συνεπών επιχειρήσεων είναι πολύ μεγαλύτερος από τον αριθμό των ασυνεπών και αυτό έχει ως αποτέλεσμα ο αλγόριθμος ταξινόμησης να έχει την τάση να κατατάσσει όλες τις παρατηρήσεις στην μεγαλύτερη ομάδα.

Οι μέθοδοι αντιμετώπισης του παραπάνω προβλήματος διακρίνονται σε τρεις κατηγορίες, οι οποίες είναι οι εξής :

- Μείωση του πλήθους των στοιχείων της ομάδας με το μεγαλύτερο πλήθος παρατηρήσεων (under-sampling).
- Αύξηση του πλήθους των στοιχείων της ομάδας με το μικρότερο πλήθος παρατηρήσεων (over-sampling).
- Συνδυασμός των δύο παραπάνω μεθόδων.

Η εφαρμογή της πρώτης μεθόδου έχει ως αποτέλεσμα την απώλεια σημαντικής πληροφορίας η οποία θα ήταν απαραίτητη στην εκπαίδευση του ταξινομητή. Από την άλλη μεριά, η αύξηση του δείγματος είναι πιθανό να οδηγήσει σε υπερπροσαρμογή του

μοντέλου δεδομένου ότι θα δημιουργηθούν πολλά πανομοιότυπα στοιχεία της κατηγορίας με το μικρότερο πλήθος παρατηρήσεων. Κάθε μία από τις παραπάνω μεθόδους είναι κατάλληλη για τον εκάστοτε ταξινομήτη. Στην παρούσα εργασία θα εφαρμόσουμε την τυχαία δειγματοληψία με επανάθεση στο σύνολο των ασυνεπών επιχειρήσεων και θα αυξήσουμε το μέγεθος του και τα μοντέλα ταξινόμησης θα εφαρμοστούν στα ισορροπημένα δεδομένα.

## 4.2 ΑΝΑΛΥΣΗ

### 4.2.1 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Έστω ότι *p* είναι πιθανότητα το δάνειο της επιχείρησης να χαρακτηριστεί ως μη εξυπηρετούμενο στο τέλος του επόμενου χρόνου. Οι συντελεστές που προκύπτουν μετά την εφαρμογή της λογιστικής παλινδρόμησης φαίνονται στον επόμενο πίνακα.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 3.8606829 | 0.3098937 | 12.458 | < 2e-16 | *** |
| LIQUIDITY | -0.0794085 | 0.0098136 | -8.092 | 5.88e-16 | *** |
| CURRENT | 0.0159446 | 0.0024483 | 6.512 | 7.39e-11 | *** |
| QUICK | 0.0064280 | 0.0016424 | 3.914 | 9.09e-05 | *** |
| AR_DAYS | 0.0271148 | 0.0050078 | 5.414 | 6.15e-08 | *** |
| INVENTORY_DAYS | 0.0125605 | 0.0019469 | 6.452 | 1.11e-10 | *** |
| AP_DAYS | 0.0163327 | 0.0032160 | 5.079 | 3.80e-07 | *** |
| SALESWC | 0.0077661 | 0.0016792 | 4.625 | 3.75e-06 | *** |
| OPERATIONS | -0.0215512 | 0.0038473 | -5.602 | 2.12e-08 | *** |
| GROSS_PROFIT | -0.0007733 | 0.0009033 | -0.856 | 0.391975 | |
| OP_MARGIN | 0.0042778 | 0.0011155 | 3.835 | 0.000126 | *** |
| NPBTSALES | -0.0072544 | 0.0021649 | -3.351 | 0.000805 | *** |
| NPBTTA | 0.0107662 | 0.0029119 | 3.697 | 0.000218 | *** |
| NPBTTNW | 0.0102748 | 0.0020034 | 5.129 | 2.92e-07 | *** |
| SALESTA | 0.0028127 | 0.0021575 | 1.304 | 0.192347 | |
| SALES_GROWTH | 0.0014494 | 0.0013635 | 1.063 | 0.287796 | |
| CAP_STRACTURE | -0.0323695 | 0.0023270 | -13.910 | < 2e-16 | *** |
| DEBTTNW | 0.0046253 | 0.0014674 | 3.152 | 0.001621 | ** |
| DEBT_COVERAGE | -0.0951280 | 0.0081129 | -11.726 | < 2e-16 | *** |
| C_IMP_MGMT | 0.0403874 | 0.0038594 | 10.465 | < 2e-16 | *** |
| CASH_COVERAGE | 0.0218174 | 0.0022751 | 9.590 | < 2e-16 | *** |
| EARNINGS_COVERAGE | 0.0369710 | 0.0045534 | 8.119 | 4.68e-16 | *** |
| INDUSTRY_RISK | -0.0720192 | 0.0074676 | -9.644 | < 2e-16 | *** |
| ECONOMIC_CONDITIONS | 0.0259412 | 0.0061649 | 4.208 | 2.58e-05 | *** |
| STRAUCTURAL_FACTORS | 0.0045697 | 0.0019964 | 2.289 | 0.022080 | * |
| INDUSTRY_PERFOMANCE | 0.0051270 | 0.0033958 | 1.510 | 0.131095 | |
| MANAGEMENT_QUALITY | -0.0075641 | 0.0006580 | -11.496 | < 2e-16 | *** |
| MANAGEMENT_ORGANIZATION | -0.0117238 | 0.0007226 | -16.224 | < 2e-16 | *** |
| COMPANY_REPUTATION | -0.0047459 | 0.0010361 | -4.580 | 4.64e-06 | *** |
| COMPANY_STATUS | -0.0029735 | 0.0011153 | -2.666 | 0.007675 | ** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Πίνακας 4.3 : Συντελεστές του μοντέλου λογιστικής παλινδρόμησης

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = 3{,}86 - 0{,}079\cdot X_1 - 0{,}015\cdot X_2 - 0{,}006\cdot X_3 - 0.027\cdot X_4 + 0{,}012\cdot X_5 + 0{,}016\cdot X_6 + 0{,}007\cdot X_7 - 0{,}021\cdot X_8 - 0{,}001\cdot X_9 + 0{,}004\cdot X_{10} - 0{,}007\cdot X_{11} + 0{,}01\cdot X_{12} + 0{,}010\cdot X_{13} + 0{,}002\cdot X_{14} + 0{,}001\cdot X_{15} - 0{,}032\cdot X_{16} + 0{,}004\cdot X_{17} - 0{,}095\cdot X_{18} + 0{,}040\cdot X_{19} + 0{,}021\cdot X_{20} + 0{,}036\cdot X_{21} - 0{,}072\cdot X_{22} + 0{,}025\cdot X_{23} + 0{,}004\cdot X_{24} + 0{,}005\cdot X_{25} - 0{,}007\cdot X_{26} - 0{,}011\cdot X_{27} - 0{,}004\cdot X_{28} - 0{,}002\cdot X_{29}$$

Η πιθανότητα p ως συνάρτηση των μεταβλητών η εξίσωση διαμορφώνεται ως εξής :

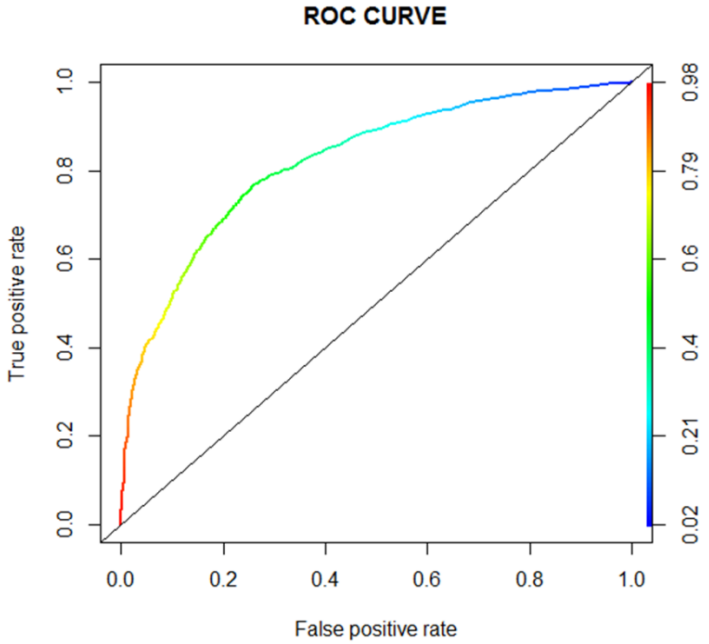
$$\hat{p} = \frac{e^{\hat{\beta}\cdot x}}{1+e^{\hat{\beta}\cdot x}}$$

Έτσι, έχουμε ότι από τους αξιόπιστους πελάτες του δείγματος ελέγχου σωστά ταξινομείται το 74%. Όσο αφορά τους αναξιόπιστους πελάτες, σωστά ταξινομείται το 76,3%. Συνεώς αυτό που παρατηρούμε είναι το μοντέλο προβλέπει εξίσου καλά και τις δύο κατηγορίες. Σύμφωνα με τον παρακάτω πίνακα το μοντέλο της λογιστικής παλινδρόμησης παρέχει ακρίβεια 75,18%.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2692  875
         1  943 2815

               Accuracy : 0.7518
                 95% CI : (0.7417, 0.7617)
    No Information Rate : 0.5038
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5035
 Mcnemar's Test P-Value : 0.1161

            Sensitivity : 0.7406
            Specificity : 0.7629
         Pos Pred Value : 0.7547
         Neg Pred Value : 0.7491
             Prevalence : 0.4962
         Detection Rate : 0.3675
   Detection Prevalence : 0.4870
      Balanced Accuracy : 0.7517

       'Positive' Class : 0
```

*Πίνακας 4.4 :Πίνακας ταξινόμησης μοντέλου λογιστικής παλινδρόμησης*

Η μορφή της καμπύλης ROC αποτελεί ένα κριτήριο για την αποτελεσματικότητα και ακρίβεια ενός μοντέλου και συγκεκριμένα η περιοχή που βρίσκεται κάτω από την καμπύλη. Για ένα «τέλειο» μοντέλο το εμβαδόν κάτω από την καμπύλη θα είναι ίσο με 1, ενώ για ένα μη αξιόπιστο η τιμή που θα λαμβάνει δεν θα ξεπερνάει το 0,5.Συγκερκιμένα, το εμβαδόν της περιοχής που βρίσκεται κάτω από την καμπύλη στο σχήμα (0,8215) εκφράζει την πιθανότητα το μοντέλο ταξινόμησης να δώσει μεγαλύτερη πιθανότητα αναξιοπιστίας σε έναν πελάτη-επιχείρηση που έχει επιλεχθεί τυχαία από τον πληθυσμό των αναξιόπιστων επιχειρήσεων απ'ότι να είχε επιλεχθεί από τον πληθυσμό των αξιόπιστων επιχειρήσεων.




*Γράφημα 4.4: Διάγραμμα ROC για το μοντέλο λογιστικής παλινδρόμησης*

Όπως διαπιστώνουμε από τον πίνακα (με τους συντελεστές) το μοντέλο δεν περιλαμβάνει όλες τις μεταβλητές που είναι στατιστικά σημαντικές. Δηλαδή, οι μεταβλητές GROSSPROFIT (p-value=0.3919>0,05), SALESTA (pvalue=0.1923>0,05), SALES_GROWTH (p-value=0.2877>0,05), INDUSTRY_PERFOMANCE (pvalue=0.131>0,05) δεν είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 5%.

Για να εντοπιστεί ο βέλτιστος συνδυασμός των ανεξάρτητων μεταβλητών που θα εισαχθούν στο λογιστικό υπόδειγμα έγινε χρήση της βηματικής μεθόδου επιλογής μεταβλητών (stepwise) και συγκεκριμένα της προς τα πίσω περιορσμός (backward elimination). Ο αλγόριθμος ξεκινά με το μοντέλο που περιλαμβάνει όλες τις μεταβλητές και διαδοχικά απομακρύνει κάποιες από αυτές αφού εξεταστεί η στατιστική τους

σημαντικότητα. Η στατιστική σημαντικότητα κάθε μεταβλητής αποτυπώνεται με βάση το $X^2$ τεστ (Altman, 1968)(chi-square test) με το οποίο ελέγχεται αν ο συντελεστής της ισούται με μηδέν. Το επίπεδο σημαντικότητας βάσει του οποίου κρίνεται αν μια μεταβλητή είναι στατιστικά σημαντική ή όχι είναι το 5%. Η μεταβλητή που εξαιρείται σε κάθε βήμα είναι εκείνη με το μεγαλύτερο p-value. Στο επόμενο βήμα πραγματοποιείται έλεγχος για το αν κάποια από τις μεταβλητές που έχουν απομακρυνθεί πλέον θεωρείται σημαντική και μπορεί να εισαχθεί στο μοντέλο. Ο αλγόριθμος σταματάει όταν όλοι οι συντελεστές των μεταβλητών που έχουν μείνει στο μοντέλο είναι στατιστικά σημαντικοί. Η τελική επιλογή των ανεξάρτητων μεταβλητών φαίνεται στον επόμενο πίνακα.

```
Call:  glm(formula = train_status1 ~ LIQUIDITY + CURRENT + QUICK + AR_DAYS +
    INVENTORY_DAYS + AP_DAYS + SALESWC + OPERATIONS + OP_MARGIN +
    NPBTSALES + NPBTTA + NPBTTNW + CAP_STRACTURE + DEBTTNW +
    DEBT_COVERAGE + C_IMP_MGMT + CASH_COVERAGE + EARNINGS_COVERAGE +
    INDUSTRY_RISK + ECONOMIC_CONDITIONS + STRAUCTURAL_FACTORS +
    INDUSTRY_PERFOMANCE + MANAGEMENT_QUALITY + MANAGEMENT_ORGANIZATION +
    COMPANY_REPUTATION + COMPANY_STATUS, family = binomial, data = train_data)

Coefficients:
             (Intercept)                LIQUIDITY                  CURRENT
               -3.954946                 0.080421                -0.015976
                   QUICK                  AR_DAYS            INVENTORY_DAYS
               -0.006538                -0.027683                -0.012713
                 AP_DAYS                  SALESWC               OPERATIONS
               -0.016923                -0.007859                 0.017787
               OP_MARGIN                NPBTSALES                   NPBTTA
               -0.003759                 0.008225                -0.009775
                 NPBTTNW             CAP_STRACTURE                  DEBTTNW
               -0.009526                 0.031198                -0.003955
          DEBT_COVERAGE               C_IMP_MGMT           CASH_COVERAGE
                0.095319                -0.040566                -0.021821
      EARNINGS_COVERAGE            INDUSTRY_RISK      ECONOMIC_CONDITIONS
               -0.037086                 0.072959                -0.026593
    STRAUCTURAL_FACTORS       INDUSTRY_PERFOMANCE       MANAGEMENT_QUALITY
               -0.004611                -0.005621                 0.007579
MANAGEMENT_ORGANIZATION        COMPANY_REPUTATION           COMPANY_STATUS
                0.011722                 0.004712                 0.002967

Degrees of Freedom: 21972 Total (i.e. Null);  21946 Residual
```

*Πίνακας 4.5: Συντελεστές του μοντέλου λογιστικής παλινδρόμησης-stepwise*

Επομένως το μοντέλο με τις ανεξάρτητες μεταβλητές που επιλέχθηκαν με την εφαρμογή της stepwise είναι το ακόλουθο :

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = -3{,}954 + 0{,}080\cdot X_1 - 0{,}015\cdot X_2 - 0{,}006\cdot X_3 - 0.027\cdot X_4 + 0{,}012\cdot X_5 - 0{,}016\cdot X_6 - 0{,}007\cdot X_7$$

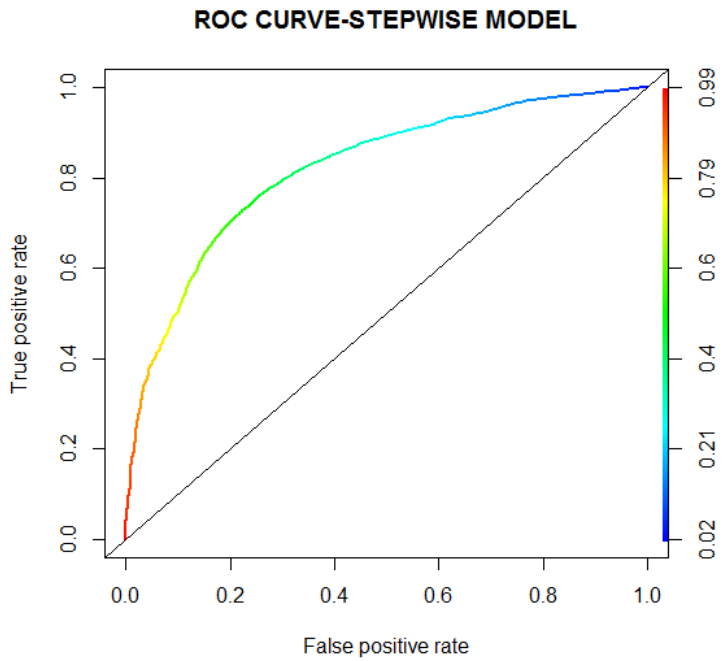
$+0{,}017\cdot X_8 - 0{,}003\cdot X_{10} + 0{,}008\cdot X_{11} - 0{,}009\cdot X_{12} - 0{,}009\cdot X_{13} + 0{,}031\cdot X_{16} - 0{,}003\cdot X_{17} - +0{,}095\cdot X_{18} - 0{,}040\cdot X_{19} - 0{,}021\cdot X_{20} - 0{,}037\cdot X_{21} + 0{,}072\cdot X_{22} - 0{,}026\cdot X_{23} - 0{,}004\cdot X_{24} - 0{,}005\cdot X_{25} + 0{,}007\cdot X_{26} + 0{,}011\cdot X_{27} + 0{,}004\cdot X_{28} + 0{,}002\cdot X_{29}$

Σύμφωνα με το παραπάνω μοντέλο έχουμε ότι από τους αξιόπιστους πελάτες σωστά ταξινομείται το 73,7 %. Ενώ, όσο αφορά τους αναξιόπιστους πελάτες, σωστά ταξινομείται το 76,3%. Παρατηρούμε ότι, το μοντέλο της *stepwise* ταξινομεί σωστά το ίδιο ποσοστό αναξιόπιστων πελατών όπως και το κορεσμένο μοντέλο όμως ανάμεσα στα δύο αυτά μοντέλα θα επιλέγαμε αυτό της stepwise αφού θεωρείται πιο οικονομικό καθώς περιλαμβάνει λιγότερες μεταβλητές. Τα ποσοστά ταξινόμησης σε κάθε κατηγορία αλλά και η ακρίβεια που παρέχει το συγκεκριμένο μοντέλο φαίνονται στον επόμενο πίνακα.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2682  872
         1  957 2814

               Accuracy : 0.7503
                 95% CI : (0.7402, 0.7602)
    No Information Rate : 0.5032
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5005
 Mcnemar's Test P-Value : 0.04951

            Sensitivity : 0.7370
            Specificity : 0.7634
         Pos Pred Value : 0.7546
         Neg Pred Value : 0.7462
             Prevalence : 0.4968
         Detection Rate : 0.3661
   Detection Prevalence : 0.4852
      Balanced Accuracy : 0.7502

       'Positive' Class : 0
```

*Πίνακας 4.6: Πίνακας ταξινόμησηςμοντέλου stepwise*

Στην περίπτωση της βηματικής μεθόδου επιλογής μεταβλητών το εμβαδόν της περιοχής που βρίσκεται κάτω από την καμπύλη όπως φαίνεται στο επόμενο σχήμα είναι 0,8205,μικρότερο σε σχέση με αυτό του κορεσμένου μοντέλου.

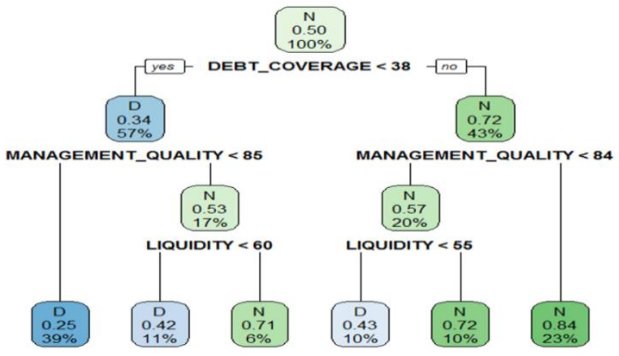



*Γράφημα 4.5: Διάγραμμα ROC για το μοντέλο λογιστικής παλινδρόμησης-stepwise*

### 4.2.2 ΔΕΝΤΡΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (CLASSIFICATION TREES – CART)

Η επόμενη μέθοδος που θα εφαρμοστεί στα δεδομένα μας είναι η CART (classification and regression trees). Η συγκεκριμένη μέθοδος είναι ιδανική για οικονομετρικές εφαρμογές καθώς ένα από τα στοιχεία που την χαρακτηρίζουν είναι η διαφάνεια. Μέσω της μεθόδου CART αναπτύσσεται ένα δέντρο απόφασης του οποίου οι κόμβοι αποτελούνται από τα κριτήρια αξιολόγησης (δείκτες ),τα κλαδιά του από τις συνθήκες που πρέπει να ικανοποιούνται για να γίνει ο διαχωρισμός και τέλος, τα φύλλα στα οποία βρίσκονται οι κατηγορίες ταξινόμησης.

Συνεπώς, βάσει των παραπάνω το δέντρο απόφασης που δημιουργείται είναι το εξής :



*Εικόνα 4.1: Δέντρο ταξινόμησης –CART*

Το δέντρο απόφασης που αναπτύχθηκε αποτελείται από 4 κόμβους, 6 φύλλα και αναπτύχθηκε σε 4 επίπεδα.Η βαρύτητα του κάθε κριτηρίου παρουσιάζεται στον επόμενο πίνακα.

| Κριτήριο | Βαρύτητα |
|---|---|
| $X_{18}$ | 20 |
| $X_{26}$ | 10 |
| $X_{16}$ | 8 |
| $X_{6}$ | 2 |
| $X_{19}$ | 2 |
| $X_{21}$ | 15 |
| $X_{8}$ | 8 |
| $X_{27}$ | 6 |
| $X_{28}$ | 2 |
| $X_{2}$ | 1 |
| $X_{20}$ | 10 |
| $X_{12}$ | 8 |
| $X_{1}$ | 5 |
| $X_{4}$ | 2 |
| $X_{7}$ | 1 |

*Πίνακας 4.7: Παρουσίαση βαρύτητας κάθε κριτηρίου CART μοντέλου*

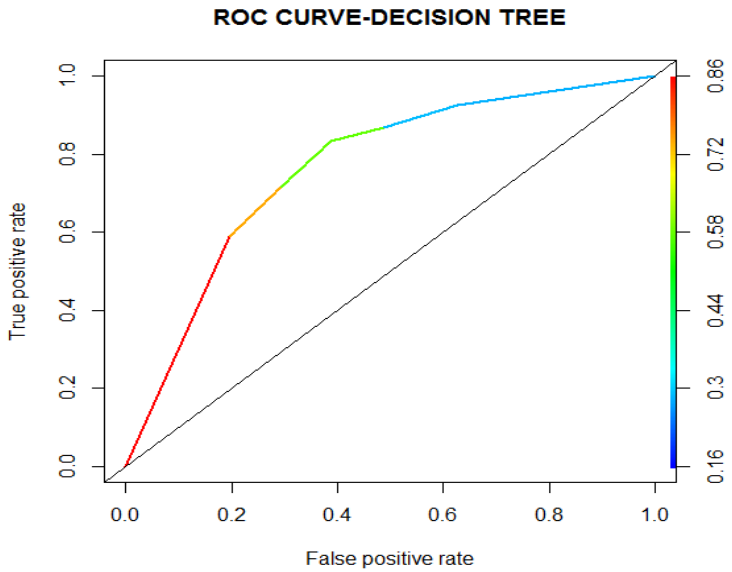
Την μεγαλύτερη βαρύτητα την έχει το κριτήριο X18 το οποίο αντιστοιχεί στον δείκτη Dept Coverage και το οποίο χρησιμοποιήθηκε πολλές φορές με σκοπό τον διαχωρισμό των επιχειρήσεων. Ακολουθούν τα κριτήρια X21, X26, X20, οι δείκτες Earnings Coverage, Management Quality και Cash coverage αντίστοιχα. Όσο αφορά τους δείκτες Current και SalesWC η συμβολή τους στην ανάπτυξη του δέντρου μπορεί να χαρακτηριστεί αμελητέα.

Το μοντέλο σύμφωνα με τον παρακάτω πίνακα παρέχει ακρίβεια 72,29% όμως, έχουμε ότι από τους αξιόπιστους πελάτες του δείγματος ελέγχου σωστά ταξινομείται το 83,2% ενώ όσο αφορά τους αναξιόπιστους πελάτες, σωστά ταξινομείται το 61,2% ποσοστό μη ικανοποιητικό αφού το βάρος της ανάλυσης πέφτει στην σωστή ταξινόμηση των αναξιόπιστων πελατών.

```
Confusion Matrix and Statistics

          Reference
Prediction    D    N
         D 3068 1412
         N  618 2227

               Accuracy : 0.7229
                 95% CI : (0.7125, 0.7331)
    No Information Rate : 0.5032
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4449
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8323
            Specificity : 0.6120
         Pos Pred Value : 0.6848
         Neg Pred Value : 0.7828
             Prevalence : 0.5032
         Detection Rate : 0.4188
   Detection Prevalence : 0.6116
      Balanced Accuracy : 0.7222

       'Positive' Class : D
```

*Πίνακας 4.8: Πίνακας ταξινόμησηςμοντέλου CART*

Στην περίπτωση της τεχνικής CART το εμβαδόν της περιοχής που βρίσκεται κάτω από την καμπύλη ROC όπως φαίνεται στο επόμενο σχήμα είναι 0,7628.




*Γράφημα 4.6: Διάγραμμα ROC για το μοντέλο CART*

### 4.2.3 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (ARTIFICIAL NEURAL NETWORKS)

Η τρίτη και τελευταία μέθοδος που θα εφαρμοστεί είναι αυτή των τεχνητών νευρωνικών δικτύων. Υπάρχουν διάφορες αρχιτεκτονικές νευρωνικών δικτύων. Στην παρούσα εργασία θα γίνει χρήση του multi-layer perceptron (MLP). Το συγκεκριμένο μοντέλο αποτελείται από δύο κόμβους στα επίπεδα εισόδου και εξόδου και από ενδιάμεσα επίπεδα. Στους κόμβους εισόδου εισέρχονται οι τιμές έστω $X$ και συγκεκριμένα οι δείκτες και υπολογίζεται ένα σταθμισμένο άθροισμα σε κάθε έναν από τους κόμβους. Το σταθμισμένο αυτό άθροισμα υπολογίζεται από την τιμή εισόδου του κόμβου και από το αντίστοιχο βάρος που υπάρχει στην σύνδεση δύο διαδοχικών κόμβων. Στο άθροισμα αυτό εφαρμόζεται η συνάρτηση $g(x)$ προκειμένου να προκύψει η τιμή εξόδου του κόμβου. Η $g(x)$ που χρησιμοποιείται είναι η hyperbolic tangent. :

$$y(v_i)=tanh(v_i)$$

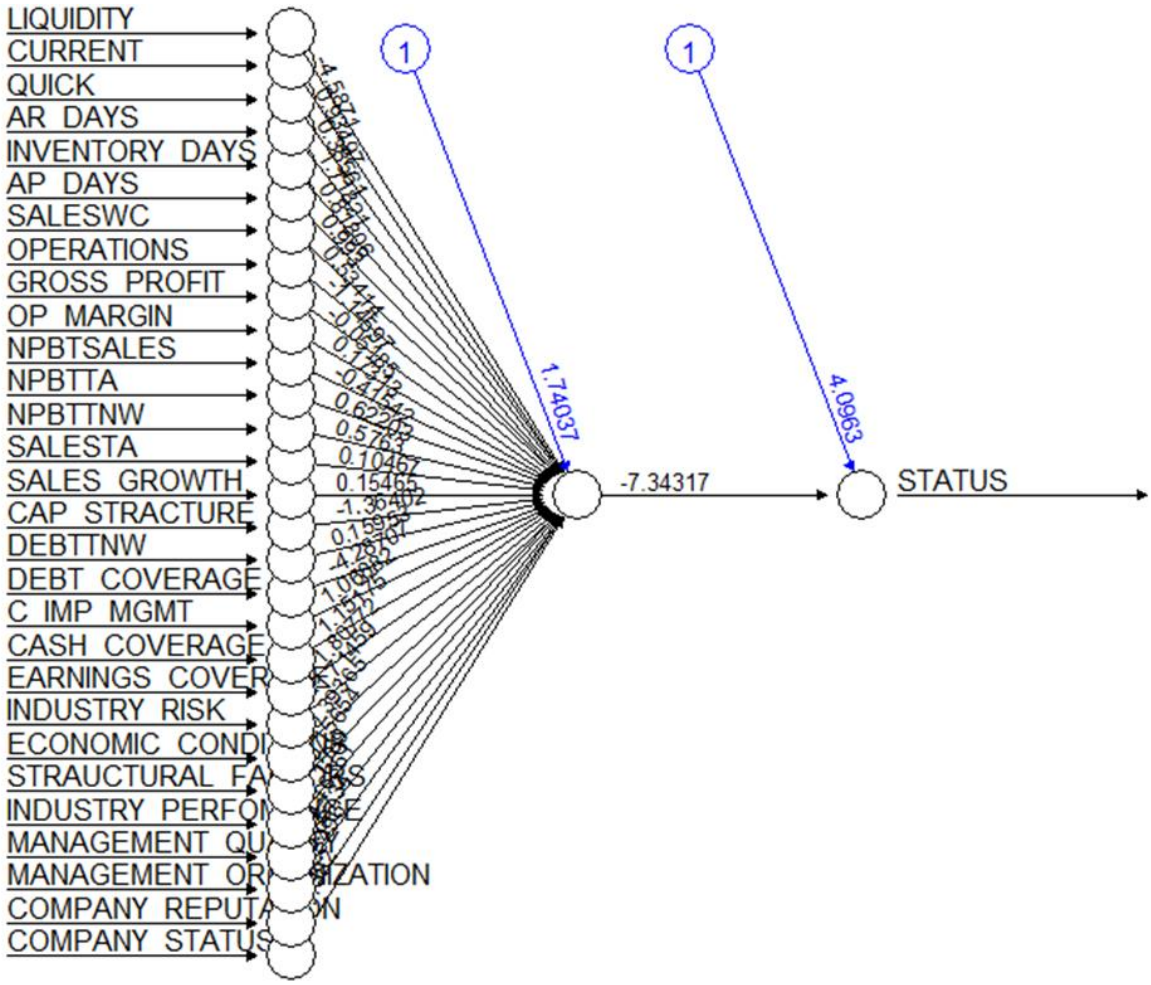
Όσο αφορά τα νευρωνικά δίκτυα μια επιπλέον προεπεξεργασία που έγινε στα δεδομένα είναι η κανονικοποίηση τους. Δεδομένου ότι ο αλγόριθμος δυσκολεύεται να συγκλίνει πριν τον μέγιστο αριθμό επαναλήψεων εάν οι τιμές εισόδου δεν κυμαίνονται στην περιοχή [0,1] κρίνεται σημαντικό να γίνει αυτός ο μετασχηματισμός. Η μέθοδος που χρησιμοποιήθηκε είναι αυτή της κανονικοποίησης ελάχιστου-μέγιστου. Στην συγκεκριμένη μέθοδο οι αριθμητικές τιμές αντιστοιχίζονται μέσω ενός γραμμικού συνδυασμού με άλλες που κυμαίνονται εντός μια προκαθορισμένης περιοχής τιμών. Έστω μια μεταβλητή A και maxA και minA η μεγαλύτερη και η μικρότερη τιμή της τιμή αντίστοιχα. Όλες οι τιμές της μεταβλητής αντιστοιχίζονται με άλλες που κυμαίνονται εντός μιας περιοχής με κατώτερο όριο το 0 ($new\_min_A$ ) και ανώτερο όριο το 1($new\_max_A$ ).

$$x' = \frac{x - \min_A}{\max_A - \min_A} = (new\_\max_A - new\_\min_A) + new\_\min_A$$

,όπου $\chi$ η τιμή της μεταβλητής και $\chi'$η νέα τιμή.

Για την ανάπτυξη του νευρωνικού δικτύου χρησιμοποιήθηκαν 29 κόμβοι εισόδου δηλαδή, όσα και τα κριτήρια αξιολόγησης, 2 κόμβοι εξόδου όσες και οι κατηγορίες ταξινόμησης και ένα ενδιάμεσο επίπεδο που περιείχε 1 κόμβο.

Τα βάρη των κριτηρίων φαίνονται στις επόμενες δύο εικόνες :




*Εικόνα 4.2: Αναπαράσταση νευρωνικού δικτύου με όλα τα κριτήρια*

| | |
|---|---|
| Intercept.to.1layhid1 | 1.738258708253 |
| LIQUIDITY.to.1layhid1 | -4.578285916861 |
| CURRENT.to.1layhid1 | 0.933345905684 |
| QUICK.to.1layhid1 | 0.384795888706 |
| AR_DAYS.to.1layhid1 | 1.714697444699 |
| INVENTORY_DAYS.to.1layhid1 | 0.816535855686 |
| AP_DAYS.to.1layhid1 | 0.990936381563 |
| SALESWC.to.1layhid1 | 0.533004373733 |
| OPERATIONS.to.1layhid1 | -1.144549229663 |
| GROSS_PROFIT.to.1layhid1 | -0.051731290509 |
| OP_MARGIN.to.1layhid1 | 0.172845708239 |
| NPBTSALES.to.1layhid1 | -0.414617534207 |
| NPBTTA.to.1layhid1 | 0.620948874836 |
| NPBTTNW.to.1layhid1 | 0.575664744058 |
| SALESTA.to.1layhid1 | 0.104709900183 |
| SALES_GROWTH.to.1layhid1 | 0.154522018850 |
| CAP_STRACTURE.to.1layhid1 | -1.361794448590 |
| DEBTTNW.to.1layhid1 | 0.159316592702 |
| DEBT_COVERAGE.to.1layhid1 | -4.280327853684 |
| C_IMP_MGMT.to.1layhid1 | 1.067106457705 |
| CASH_COVERAGE.to.1layhid1 | 1.149945082392 |
| EARNINGS_COVERAGE.to.1layhid1 | 1.804981796128 |
| INDUSTRY_RISK.to.1layhid1 | -1.712122392817 |
| ECONOMIC_CONDITIONS.to.1layhid1 | 0.393200811871 |
| STRAUCTURAL_FACTORS.to.1layhid1 | 0.156323974608 |
| INDUSTRY_PERFOMANCE.to.1layhid1 | 0.158699131771 |
| MANAGEMENT_QUALITY.to.1layhid1 | -0.467788693206 |
| MANAGEMENT_ORGANIZATION.to.1layhid1 | -0.656280591392 |
| COMPANY_REPUTATION.to.1layhid1 | -0.254466089332 |
| COMPANY_STATUS.to.1layhid1 | -0.137962813539 |
| Intercept.to.STATUS | 4.102653281021 |
| 1layhid.1.to.STATUS | -7.353762053234 |

*Εικόνα 4.3: Βάρη κριτηριών Τεχνητού Νευρωνικού Δικτύου*

Από τους αξιόπιστους πελάτες του δείγματος ελέγχου σωστά ταξινομείται το 78% ενώ όσο αφορά τους αναξιόπιστους πελάτες, σωστά ταξινομείται το 71%.

```
          Reference
Prediction    0    1
         0 2859 1042
         1  803 2620

               Accuracy : 0.7480885
                 95% CI : (0.7379807, 0.7579985)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.496177
 Mcnemar's Test P-Value : 0.00000003009548

            Sensitivity : 0.7807209
            Specificity : 0.7154560
         Pos Pred Value : 0.7328890
         Neg Pred Value : 0.7654105
             Prevalence : 0.5000000
         Detection Rate : 0.3903605
   Detection Prevalence : 0.5326324
      Balanced Accuracy : 0.7480885

       'Positive' Class : 0
```

*Πίνακας 4.9: Πίνακας ταξινόμησησμοντέλου Τεχνητών Νευρωνικών Δικτύων*

## ΑΝΑΛΥΣΗ ΕΥΑΙΣΘΗΣΙΑΣ

Όπως έχουμε αναφέρει τα νευρωνικά δίκτυα μαθαίνουν από τα δεδομένα. Είναι αρκετές οι φορές όμως που ο όγκος των δεδομένων είναι τόσο μεγάλος γεγονός που καθιστά δύσκολο από τους αναλυτές να ανακαλύψουν ποιες μεταβλητές εισόδου επηρεάζουν σημαντικά το τελικό αποτέλεσμα και ποιες μπορούν να παραλειφθούν.

Εφαρμόζοντας την ανάλυση ευαισθησίας σε ένα εκπαιδευμένο νευρωνικό δίκτυο δίνεται η δυνατότητα να εξεταστεί η σχέση της κάθε μεταβλητής με το εξαγόμενο αποτέλεσμα. Η διαδικασία περιγράφεται συνοπτικά ως εξής:

1. Προσδιορισμός του εύρους τιμών κάθε μεταβλητής στο δείγμα εκμάθησης. Έστω [α,β] το εύρος τιμών της σύμφωνα με το οποίο υπολογίζονται πέντε τιμές ($t_k$) για την i μεταβλητή μέσω της σχέσης $t_{ik} = a + k(b-a)/4, k = 0,1,2,3,4, i = 1,..,n$
2. Η επίδοση της παρατήρησης στην μεταβλητή i λαμβάνει την τιμή $t_{i0}$ και η τιμή αυτή εισέρχεται στο μοντέλο. Το ίδιο συμβαίνει για όλες τις τιμές $t_{ik}$ της μεταβλητής $i$. Έστω $dmax_i$ η μέγιστη διαφορά των επιδόσεων των «νέων» παρατηρήσεων από την επίδοση της αρχικής παρατήρησης, η οποία καταγράφεται για όλες τις παρατηρήσεις και για όλες τις μεταβλητές.
3. Κανονικοποίηση της μέγιστης διαφοράς κάθε μεταβλητής σύμφωνα με τον τύπο :

$$d\max_i = \frac{d\max_i}{\sum_{i=1}^{n} d\max_i}$$

4. Σύμφωνα με το προηγούμενο βήμα θα έχουν υπολογιστεί m μέγιστες διαφορές (έστω m το πλήθος των παρατηρήσεων). Υπολογίζοντας για κάθε μεταβλητή την μέση τιμή των διαφορών αυτών δύναται να εκτιμηθεί η σημαντικότητά της.

Μέσω της παραπάνω διαδικασίας προσδιορίζεται η «ευαισθησία» $w_1$, $w_2,\ldots,w_n$ των n μεταβλητών , στην ουσία ο βάρος τους, όπου $w_i \geq 0$ για κάθε i=1,2,…,n και $w_1+w_2+..+w_n$ =1. Η «ευαισθησία» των μεταβλητών είναι ένα μέγεθος το οποίο δείχνει κατά πόσο μια μεταβολή στην τιμή της μεταβλητής μπορεί να επηρεάσει το εξαγόμενο αποτέλεσμα. Όσες μεταβλητές λαμβάνουν υψηλή τιμή στο μέγεθος αυτό θεωρούνται σημαντικές.

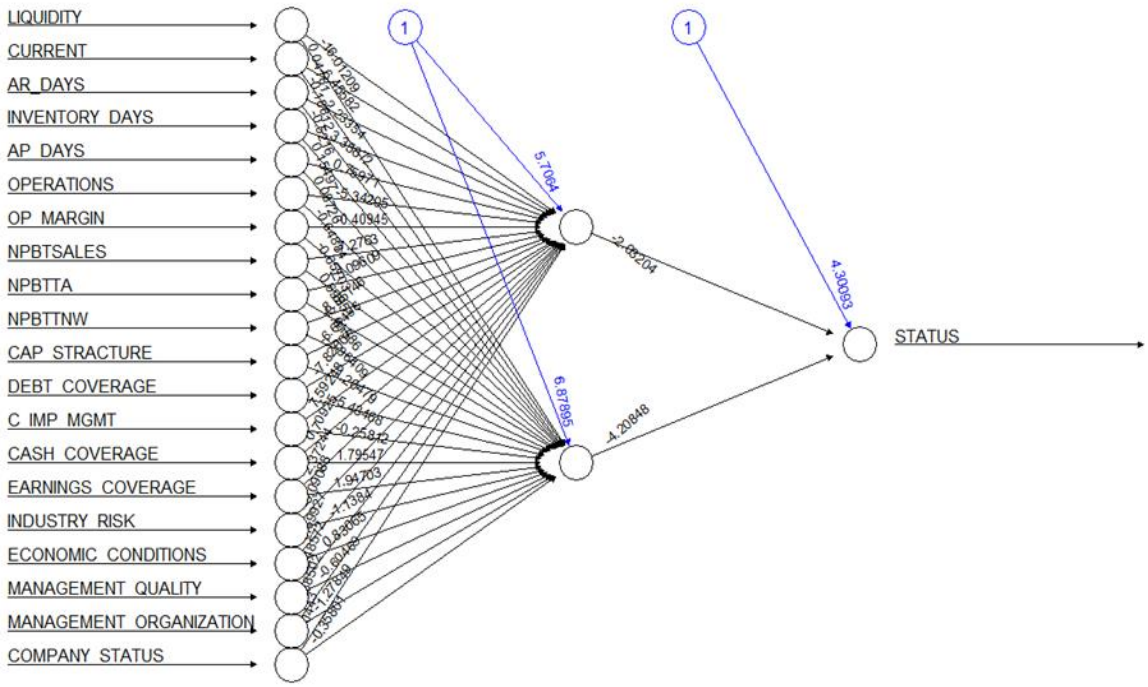
Έχοντας υπολογίσει την ευαισθησία των μεταβλητών επιλέγονται εκείνες που έχουν συνολικό βάρος τουλάχιστον 95%. Σημειώνεται ότι, σημαντικές θεωρήθηκαν εκείνες που προέκυψαν από την εφαρμογή της παραπάνω διαδικασίας σε νευρωνικά δίκτυα με ένα ενδιάμεσο επίπεδο.

Οι μεταβλητές που επιλέχθηκαν σύμφωνα με την ανάλυση ευαισθησίας παρουσιάζονται στον επόμενο πίνακα με ξεκινώντας με την μεταβλητή που έχει το μεγαλύτερο βάρος.

| **Κριτήριο** | **Βάρος** | **Κριτήριο** | **Βάρος** |
|---|---|---|---|
| X18 | 14.024 | X11 | 2.926 |
| X1 | 13.994 | X20 | 2.866 |
| X22 | 8.403 | X5 | 2.722 |
| X16 | 6.691 | X10 | 2.688 |
| X21 | 5.299 | X12 | 2.666 |
| X29 | 4.938 | X27 | 2.636 |
| X8 | 4.89 | X13 | 2.541 |
| X3 | 4.694 | X6 | 2.403 |
| X2 | 3.882 | X26 | 2.054 |
| X19 | 3.577 | X23 | 1.829 |

Σύμφωνα με τον παραπάνω πίνακα μεγαλύτερη βαρύτητα έχουν τα κριτήρια X18 και X1 με τα κριτήρια X22 και X16 να ακολουθούν.

Τα παραπάνω βάρη του τεχνητού νευρωνικού δικτύου φαίνονται στις επόμενες δύο εικόνες




*Εικόνα 4.4: Αναπαράσταση νευρωνικού δικτύου*

| | |
|---|---|
| Intercept.to.1layhid1 | 5.706402778714 |
| LIQUIDITY.to.1layhid1 | -16.012091452681 |
| CURRENT.to.1layhid1 | 6.455822350509 |
| AR_DAYS.to.1layhid1 | 2.233544787336 |
| INVENTORY_DAYS.to.1layhid1 | 3.358720808424 |
| AP_DAYS.to.1layhid1 | 0.759712260733 |
| OPERATIONS.to.1layhid1 | -5.342054159289 |
| OP_MARGIN.to.1layhid1 | -0.409453588209 |
| NPBTSALES.to.1layhid1 | -7.276300333863 |
| NPBTTA.to.1layhid1 | 15.096094987170 |
| NPBTTNW.to.1layhid1 | 8.951456618732 |
| CAP_STRACTURE.to.1layhid1 | -6.684959327718 |
| DEBT_COVERAGE.to.1layhid1 | -7.820071928052 |
| C_IMP_MGMT.to.1layhid1 | 7.592476711260 |
| CASH_COVERAGE.to.1layhid1 | 0.709251241997 |
| EARNINGS_COVERAGE.to.1layhid1 | 2.372440016321 |
| INDUSTRY_RISK.to.1layhid1 | -9.090876234479 |
| ECONOMIC_CONDITIONS.to.1layhid1 | 0.699268608076 |
| MANAGEMENT_QUALITY.to.1layhid1 | -2.485719131296 |
| MANAGEMENT_ORGANIZATION.to.1layhid1 | -2.385017720715 |
| COMPANY_STATUS.to.1layhid1 | 0.004430967780 |
| Intercept.to.1layhid2 | 6.878948625535 |
| LIQUIDITY.to.1layhid2 | 0.041812223755 |
| CURRENT.to.1layhid2 | -0.188121257442 |
| AR_DAYS.to.1layhid2 | -0.521602929599 |
| INVENTORY_DAYS.to.1layhid2 | 0.154967449393 |
| AP_DAYS.to.1layhid2 | 0.087262467304 |
| OPERATIONS.to.1layhid2 | -0.648839117595 |
| OP_MARGIN.to.1layhid2 | -0.657031615127 |
| NPBTSALES.to.1layhid2 | 0.699512252673 |
| NPBTTA.to.1layhid2 | -2.915860229237 |
| NPBTTNW.to.1layhid2 | -0.864093768012 |
| CAP_STRACTURE.to.1layhid2 | -1.204785507326 |
| DEBT_COVERAGE.to.1layhid2 | -5.434677949026 |
| C_IMP_MGMT.to.1layhid2 | -0.258122666516 |
| CASH_COVERAGE.to.1layhid2 | 1.795468786433 |
| EARNINGS_COVERAGE.to.1layhid2 | 1.947033914151 |
| INDUSTRY_RISK.to.1layhid2 | -1.138398292401 |
| ECONOMIC_CONDITIONS.to.1layhid2 | 0.830652002361 |
| MANAGEMENT_QUALITY.to.1layhid2 | -0.604691566098 |
| MANAGEMENT_ORGANIZATION.to.1layhid2 | -1.278486076377 |
| COMPANY_STATUS.to.1layhid2 | -0.358008167064 |
| Intercept.to.STATUS | 4.300934390162 |
| 1layhid.1.to.STATUS | -2.032038577402 |
| 1layhid.2.to.STATUS | -4.208481132608 |

*Εικόνα 4.5: Βάρη Τεχνητού Νευρωνικού Δικτύου*

Εφαρμόζοντας τον MLP(Multilayer Perceptron) με εισαγωγικούς κόμβους τα κριτήρια που βρέθηκαν να είναι πιο σημαντικά με βάσει την ανάλυση ευαισθησίας έχουμε τα εξής αποτελέσματα :

Από τους αξιόπιστους πελάτες του δείγματος ελέγχου σωστά ταξινομείται το 90,8% ενώ όσο αφορά τους αναξιόπιστους πελάτες, σωστά ταξινομείται το 97,2%.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3327  101
         1  335 3561

               Accuracy : 0.9404697
                 95% CI : (0.9348048, 0.9457807)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.8809394
 Mcnemar's Test P-Value : < 0.00000000000000022204

            Sensitivity : 0.9085199
            Specificity : 0.9724194
         Pos Pred Value : 0.9705368
         Neg Pred Value : 0.9140144
             Prevalence : 0.5000000
         Detection Rate : 0.4542600
   Detection Prevalence : 0.4680502
      Balanced Accuracy : 0.9404697

       'Positive' Class : 0
```

*Πίνακας 4.10: Πίνακας ταξινόμησησμοντέλου Τεχνητού Νευρωνικού Δικτύου μετά την ανάλυση ευαισθησίας*

Στον επόμενο πίνακα φαίνεται συνοπτικά η ακρίβεια των μοντέλων ταξινόμησης που αυτή την φορά εφαρμόστηκαν σε κάθε κλάδο βιομηχανίας ξεχωριστά.

| | | ΜΟΝΤΕΛΑ ΤΑΞΙΝΟΜΗΣΗΣ | | |
|---|---|---|---|---|
| **ΚΛΑΔΟΣ ΒΙΟΜΗΧΑΝΙΑΣ/ΑΚΡΙΒΕΙΑ ΤΑΞΙΝΟΜΗΣΗΣ** | **ΚΑΤΗΓΟΡΙΑ ΤΑΞΙΝΟΜΗΣΗΣ** | **ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ** | **ΔΕΝΤΡΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ** | **ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ** |
| **ΕΜΠΟΡΙΟ** | ΣΥΝΕΠΗΣ | 77% | 83% | 80% |
| | ΑΣΥΝΕΠΗΣ | 78% | 70% | 75% |
| **ΑΚΡΙΒΕΙΑ** | | 78% | 77,2% | 77,6% |
| **ΚΑΤΑΣΚΕΥΑΣΤΙΚΕΣ** | ΣΥΝΕΠΗΣ | 76,22% | 81,4% | 77% |
| | ΑΣΥΝΕΠΗΣ | 68% | 91,4% | 70,8% |
| **ΑΚΡΙΒΕΙΑ** | | 72,0% | 86% | 74,4% |
| **ΒΙΟΜΗΧΑΝΙΑ** | ΣΥΝΕΠΗΣ | 74,9% | 90% | 80,02% |
| | ΑΣΥΝΕΠΗΣ | 77,4% | 79% | 80,7% |
| **ΑΚΡΙΒΕΙΑ** | | 76,22% | 84,7% | 80,5% |
| **ΝΑΥΤΙΛΙΑΚΑ** | ΣΥΝΕΠΗΣ | 75% | 74% | 66% |
| | ΑΣΥΝΕΠΗΣ | 76% | 76,2% | 75% |
| **ΑΚΡΙΒΕΙΑ** | | 73% | 75,18% | 70% |
| **ΚΤΗΜΑΤΟΜΕΣΙΤΙΚΕΣ** | ΣΥΝΕΠΗΣ | 90% | 91% | 87% |
| | ΑΣΥΝΕΠΗΣ | 99,95% | 80% | 75% |
| **ΑΚΡΙΒΕΙΑ** | | 94,85 | 85% | 81% |
| **ΥΠΗΡΕΣΙΕΣ** | ΣΥΝΕΠΗΣ | 76% | 77% | 84,9% |
| | ΑΣΥΝΕΠΗΣ | 80% | 76,8% | 70,4% |
| **ΑΚΡΙΒΕΙΑ** | | 78% | 77% | 77,6% |

*Πίνακας 4.6: Η ακρίβεια των μοντέλων ταξινόμησης σε κάθε κλάδο βιομηχανίας ξεχωριστά*

# ΚΕΦΑΛΑΙΟ 5

# ΣΥΜΠΕΡΑΣΜΑΤΑ

Η δανειοδότηση επιχειρήσεων αλλά και ιδιωτών είναι μια από τις κύριες δραστηριότητες των πιστωτικών ιδρυμάτων και παράλληλα αποτελεί την μεγαλύτερη πηγή εξόδων τους. Όμως, η συγκεκριμένη διαδικασία οδηγεί στην έκθεση των τραπεζών σε διάφορους κινδύνους όπως ο πιστωτικός γεγονός που μπορεί να επηρεάζει την ομαλή λειτουργίας τους και να οδηγήσει ακόμα και στην πτώχευση. Συνεπώς, η διαχείριση των κινδύνων αυτών είναι απαραίτητη έτσι ώστε να μην εκτίθεται σε ρίσκο η βιωσιμότητα των τραπεζών.

Για να περιοριστεί ο πιστωτικός κίνδυνος είναι απαραίτητο να έχουν ληφθεί οι σωστές αποφάσεις που αφορούν την δανειοδότηση. Προκειμένου να γίνει αυτό θα πρέπει να έχει προηγηθεί η μέτρηση και διαχείρηση του πιστωτικού κινδύνου, σε συνδυασμό πάντα με τους κανόνες και την εποπτεία της κάθε εποπτικής αρχής.

Για την εκτίμηση του πιστωτικού κινδύνου απαιτείται ο υπολογισμός της πιθανότητας ασυνέπειας δηλαδή, της πιθανότητας ο δανειολήπτης να μην ανταπεξέλθει στις υποχρεώσεις του. Για την παραπάνω διαδικασία, η οποία θεωρείται αρκετά πολύπλοκη είναι απαραίτητη η χρήση αποτελεσματικών συστημάτων αξιολόγησης πιστοληπτικής ικανότητας των δανειοληπτών. Ακόμα, τα τραπεζικά ιδρύματα θα πρέπει να είναι σε θέση να υπολογίσουν και τις ελάχιστες κεφαλαιακές απαιτήσεις σύμφωνα με τις προτάσεις των συμφώνων της Επιτροπής της Βασιλείας.

Διάφοροι αναλυτές έχουν ασχοληθεί με την κατασκευή μοντέλων αξιολόγησης πιστωτικού κινδύνου και μέχρι και σήμερα έχουν αναπτυχθεί διάφορα συστήματα των οποίων η αποδοτικότητα και η αξιοπιστία είναι ευρέως αναγνωρισμένη. Μέσω αυτών των συστημάτων κάθε δανειολήπτης εντάσσεται στην βαθμίδα πιστωτικού κινδύνου που ανήκει και με τον τρόπο αυτό αποδίδεται η πιστοληπτική του ικανότητα. Έτσι δεν γίνεται γνωστή μόνο η κατηγορία στην οποία ανήκει δηλαδή, στους συνεπείς ή όχι αλλά γίνονται γνωστές περισσότερες πληροφορίες σχετικά με την ικανότητα που έχει ο κάθε πελάτης να ανταπεξέλθει στις δανειακές του υποχρεώσεις.

Στην παρούσα εργασία, εξετάστηκε η αποδοτικότητα τριών διαφορετικών μοντέλων αξιολόγησης πιστωτικού κινδύνου της Λογιστικής Παλινδρόμησης (Logistic Regression), των Δέντρων Κατηγοριοποίησης (Classification Tress) και των Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks). Οι παραπάνω μέθδοδοι εφαρμόστηκαν σε μεγάλο δείγμα επιχειρήσεων από έξι διαφορετικούς κλάδους βιομηχανίας χρησιμοποιώντας χρηματοοικονομικούς δείκτες ως κριτήρια αξιολόγησης με στόχο την ταξινόμησή τους σε δύο κατηγορίες πιστωτικού κινδύνου, τους συνεπείς και τους ασυνεπείς.

Οι μέθοδοι εφαρμόστηκαν  χρησιμοποιώντας και όλα τα κριτήρια αξιολόγησης αλλά και όσα κρίθηκαν σημαντικά ανάλογα με την μέθοδο που χρησιμοποιήθηκε. Έχουμε λοιπόν ότι, το μοντέλο της λογιστικής παλινδρόμησης έχει ακρίβεια 75,18% και ταξινομεί σωστά τους αναξιόπιστους πελάτες σε ποσοστό 76% ποσοστά αρκετά ικανοποιητικά αφού το βάρος της ανάλυσης πέφτει στην σωστή ταξινόμηση των αναξιόπιστων πελατών. Επίσης, το μοντέλο της λογιστικής παλινδρόμησης που προκύπτει μετά την βηματική μέθοδο επιλογής μεταβλητών έχει την ίδια περίπου ακρίβεια δηλαδή, 75,03%και ταξινομεί σωστά τους αναξιόπιστους πελάτες σε ποσοστό 76%. Παρόλου που η ακρίβεια των δύο μοντέλων είναι περίπου η ίδια ανάμεσα στα δύο θα επιλέγαμε αυτό που προέκυψε από την βηματική μέθοδο επιλογής μεταβλητών καθώς με λιγότερες μεταβλητές πετυχαίνουμε την ίδια ακρίβεια ταξινόμησης και εξοικονομούμε χρόνο και χρήμα.

Η επόμενη μέθοδος που εφαρμόστηκε είναι αυτή των δέντρων κατηγοριοποίησης. Συγκεκριμένα εφαρμόστηκε ο αλγόριθμος CART και η ακρίβεια που παρέχει το συγκεκριμένο μοντέλο είναι ελάχιστα χαμηλότερη από αυτήν της λογιστικής παλινδρόμησης δηλαδή, 72,2%. Όμως, αυτό που αξίζει να σημειωθεί είναι ότι το ποσοστό σωστής ταξινόμησης των αναξιόπιστων πελατών είναι μόνο 61,2% ενώ, το ποσοστό σωστής ταξινόμησης των αξιόπιστων φτάνει το 83% γεγονός που μας οδηγεί στο συμπέρασμα ότι η μέθοδος των δέντρων κατηγοριοποίησης δεν δουλεύει και τόσο καλά σε σχέση με αυτό της λογιστικής παλινδρόμησης.

Η Τρίτη και τελευταία μέθοδος που εφαρμόστηκε είναι αυτή τον νευρωνικών δικτύων και συγκεκριμένα εφαρμόστηκε η αρχιτεκτονική του multi-layer perceptron (MLP). Στην συγκεκριμένη περίπτωση οι αναξιόπιστοι πελάτες ταξινομήθηκαν σωστά σε ποσοστό 75% και το μοντέλο παρείχε ακρίβεια 74%. Καθώς τα νευρωνικά δίκτυα είναι μια μέθοδος που απαιτεί μεγάλους χρόνους εκπαίδευσης εφαρμόσαμε την ανάλυση ευαισθησίας με σκοπό να μειώσουμε το πλήθος των εισαγόμενων κριτηρίων. Αυτό είχε

σαν αποτέλεσμα στην δημιουργία ενός μοντέλου που όχι μόνο παρείχε μεγάλη ακρίβεια δηλαδή, 94% αλλά ταξινομούσε σωστά τους αναξίοπιστους πελάτες σε ποσοστό 97%.

Σύμφωνα με τα παραπάνω η μέθοδος που παρείχε την μεγαλύτερη ακρίβεια ταξινόμησης είναι αυτή των νευρωνικών δικτύων αφού είχε προηγηθεί η ανάλυση ευαισθησίας. Χρήσιμη θα ήταν μια περαιτέρω έρευνα του συγκεκριμένου μοντέλου εφαρμόζοντας και άλλες μεθόδους επιλογής βέλτιστων χαρακτηριστικών. Ακόμα, θα μπορούσε η συγκεκριμένη μέθοδος θα μπορούσε να εφαρμοστεί και σε άλλα προβλήματα ταξινόμησης με ίσως περισσότερες κατηγορίες ταξινόμησης με σκοπό να σχηματιστεί μια γενική εικόνα για την ακρίβεια των αποτελεσμάτων.

Ενδιαφέρον θα παρουσίαζε η εφαρμογή της συγκεκριμένης μεθόδου σε περισσότερα δεδομένα με περισσότερα κριτήρια αξιολόγησης αλλά και σε μη ισορροπημένα δεδομένα. Τέλος θα μπορούσε να πραγματοποιηθεί σύγκριση μοντέλων ταξινόμησης με τα ίδια κριτήρια αξιολόγησης αλλά κάνοντας χρήση δεδομένων και από ξένες τράπεζες και σε χώρες που δεν διανύουν την οικονομική κρίση.

# ΠΑΡΑΡΤΗΜΑ

# ΒΙΒΛΙΟΓΡΑΦΙΑ


Μ. Χαλκίδης & Μ. Βαζιργιάννης, (2005). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό* Β' έκδοση, Εκδόσες Τυπωθήτω

S. B. Kotsiantis, (2007). Supervised machine learning: A review of classification techniques, Informatica **31**, 249-268

I. E. Livieris, P. Pintelas (2008). A survey on algorithms for training artificial neural networks, Technical Report TR08-01, Department of Mathematics, University of Patras

K. Aas (2005). The Basel II IRB approach for credit portfolios: A survey, Norsk Regnesentral, Norwegian Computing Center

E. I. Altman., (1968). Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy, *The Journal of Finance*, **23**(4), 589-609

S. Balcaen, & H. Ooghe, (2004). 35 years of studies on business failure : an overview of the classical statistical methodologies and their related problems, Gent University, Belgium: Working Paper, Department of Accountancy and Corporate Finance

BCBS, (2001). The Internal Ratings-Bases Approach : Supporting Document to the New Basek Accord, Bank for International Settlements.

BCBS, (2001). The New Basel Capital Accord, Bank of International Settlements.

BCBS, (2004). International Convergence of Capital Measurement and Capital Standards, Bank for International Settlements

W. H. Beaver, (1966). Financial Ratios As Predictors of Failure, *Journal of Accounting Research*, **4**, 71-11

W. J. Boyes, D. L. Hoffman & S. A. Low (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, **40**, 3-14

L. Breiman, J. H. Friedman, R. A. Olsen & C. J. Stone, (1984). Classification and regression trees, Wadsworth International (California)

F. Esposito, D. Malerba & G. Semerano, (1997). A comparative analysis of methods for pruning decision tress, IEEE Transactions on Pattern Analysis and Machine Intelligence, **19**(5), 476-491

S. Fritz & D. Hosemann, (2000). Restructuring the credit process: Behavior scoring for Germans corporates, *International Journal of Intelligent Systems in Accounting,* Finance & Management, **9**, 9-21

H. Frydman, E. I. Altman, & D. L. Kao, (1985). Introducing recursive partitioning for financial classification: The case of financial distress, *Journal of Finance*, 269-291

J. Han & M, Kamper, (2001). Data Mining : Concepts and Techniques, Morgan Kaufman Publishers

Y. Liu & M. Schumann, (2005). Data mining feature selection for credit scoring models, *Journal of the Operational Research Society*, 1099-1108

M. H. Dunham, (2004). *Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα*, Εκδόσεις Νέων Τεχνολογιών

M. T. Hagan, H. Demuth & M. Beale, (1996). Neural Network Design, Boston: PWS Publishing Company

M. L. Marais, J. M. Patell & M. A. Wolfson, (1985). The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank, *Journal of Finance*, **22**, 87-114

D. Martin, (1977). Early warning of bank failure: A logit regression Approach, *Journal of Banking and Finance*, **1**(3), 249-276

K. Murthy, (1998). Automatic construction of decision trees from data: A multidisciplinary survey, Data Mining and Knowledge Discovery, **2**(4), 345-389

C. S. Ong, J. J. Huang & G. H. Tzeng, (2005). Building credit scoring models using genetic programming, *Expert Systems with Applications: An International Journal*, **29**, 41-47

D. Rumelhart, G. E. Hinton & R. J. Williams, (1986). Learning internal representations by error propagation. Cambridge, MA, 318-362

A. Saunders & M. M. Cornett, (2003). Financial Institutions Management: A Risk Management Approach, 6th edition, McGraw-Hill Education

C. A. Smith, (1946). Some Examples of Discrimination. Annals of Eugenics, **13**(1), 272-282.

M. Stone, (1974). Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society B*, **36**, 111-147

D. West, (2000). Neural network credit scoring models. Computers and Operations Research, **27**(11-12), 1131-1152

J. W. Wilcox, (1970). A Simple Theory of Financial Ratios as Predictors of Failure, *Journal of Accounting Research*, **2**, 389-395

Y. Yohannes & P. Webb, (1999). Classification and regression trees, Cart. A user manual for identifying indicators of vulnerability to famine and chronic food.., International Food Policy Research Institute, Microcomputer in Policy Research 3, Washington D.C.

https://en.wikipedia.org/wiki/Decision_tree_learning