RESEARCH TOPICS ON CREDIT RISK MANAGEMENT

PANAGIOTA GIANNOULI

PhD THESIS

UNIVERSITY OF THE AEGEAN

DEPARTMENT OF STATISTICS &

ACTUARIAL-FINANCIAL MATHEMATICS

Supervisor: Professor Alex Karagrigoriou

July 13, 2021

RESEARCH TOPICS ON CREDIT RISK MANAGEMENT

PANAGIOTA GIANNOULI

PhD THESIS

UNIVERSITY OF THE AEGEAN
DEPARTMENT OF STATISTICS &
ACTUARIAL-FINANCIAL MATHEMATICS

Supervisor: Professor Alex Karagrigoriou

July 13, 2021

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Εγώ, η Παναγιώτα Γιαννούλη, δηλώνω υπεύθυνα ότι είμαι η αποκλειστική συγγραφέας της υποβληθείσας Διδακτορικής Διατριβής με τίτλο «Research Topics on Credit Risk Management». Η συγκεκριμένη Διδακτορική Διατριβή είναι πρωτότυπη και εκπονήθηκε αποκλειστικά για την απόκτηση του Διδακτορικού διπλώματος του Τμήματος. Κάθε βοήθεια, την οποία είχα για την προετοιμασία της, αναγνωρίζεται πλήρως και αναφέρεται επακριβώς στην εργασία. Επίσης, επακριβώς αναφέρω στην εργασία τις πηγές, τις οποίες χρησιμοποίησα, και μνημονεύω επώνυμα τα δεδομένα ή τις ιδέες που αποτελούν προϊόν πνευματικής ιδιοκτησίας άλλων, ακόμη κι εάν η συμπερίληψή τους στην παρούσα εργασία υπήρξε έμμεση ή παραφρασμένη. Γενικότερα, βεβαιώνω ότι κατά την εκπόνηση της Διδακτορικής Διατριβής έχω τηρήσει απαρέγκλιτα όσα ο νόμος ορίζει περί διανοητικής ιδιοκτησίας και έχω συμμορφωθεί πλήρως με τα προβλεπόμενα στο νόμο περί προστασίας προσωπικών δεδομένων και τις αρχές Ακαδημαϊκής Δεοντολογίας.

I, Panagiota Giannouli, declare that the research presented in this thesis is my own unless otherwise stated. This thesis has been prepared in Word. The programming language used for the development of the presented methods is R with some use of SPSS and Excel.

EVALUATION COMMITTEE

Karagrigoriou A. (Supervisor, member of the 3- and 7-member committee)

Professor, University of the Aegean,

Dept. of Statistics and Actuarial-Financial Mathematics.

Konstantinidis D. (Member of the 3- and 7-member committee)

Professor, University of the Aegean,

Dept. of Statistics and Actuarial-Financial Mathematics.

Xanthopoulos S. (Member of the 3- and 7-member committee)

Associate Professor, University of the Aegean,

Dept. of Statistics and Actuarial-Financial Mathematics.

Mavri M. (Member of the 7-member committee)

Associate Professor, University of the Aegean,

Dept. of Business Administration.

Stehlík M. (Member of the 7-member committee)

Professor, University of Valparaiso, Chile

Institute of Statistics &

Associate Professor, Johannes Kepler University Linz, Austria

Institute of Applied Statistics.

Xalidias N. (Member of 7-member committee)

Professor, University of the Aegean,

Dept. of Statistics and Actuarial-Financial Mathematics.

Xatzopoulos P. (Member of the 7-member committee)

Associate Professor, University of the Aegean,

Dept. of Statistics and Actuarial-Financial Mathematics.

Acknowledgements

First, I would like to thank my family and my husband for their encouragement and their support during my studies. I would like to express my gratitude to my supervisor Dr. Alex Karagrigoriou for the scientific, spiritual and moral support he provided me. Dr. Xanthopoulos also deserves special mention for his valuable advice, guidance and support. Also, I would like to thank Dr. Konstantinidis who helped me and trusted me with the laboratory exercises of his course and Dr. Kountzakis for his collaboration during the preparation of this thesis. I would also like to thank all seven members of the committee who read my thesis and provided helpful comments. Further, I wish to thank the PhD candidate Kimon Ntotsis for fruitful discussions and collaboration on regularization methods. Finally, I would like to thank TEIRESIAS S.A. for the support, assistance and provision of data because without them the present thesis could not be achieved.

Περίληψη

Ο πιστωτικός κίνδυνος είναι μια από τις μεγαλύτερες απειλές που αντιμετωπίζουν τα πιστωτικά ιδρύματα. Η σταδιακή ανάπτυξη του ελέγχου του πιστωτικού κινδύνου οδηγεί στη ανάγκη για συνεχή βελτίωση των μοντέλων πιστωτικού κινδύνου προκειμένου αυτός να αντιμετωπιστεί ή να προβλεφθεί. Για το λόγο αυτό, η παρούσα διατριβή εστιάζει στη συμβολή σε τομείς σχετικούς με μεθόδους πρόβλεψης και επιλογής επεξηγηματικών/ανεξάρτητων μεταβλητών με απώτερο σκοπό την ενίσχυση της αποδοτικότητας των μοντέλων πιστωτικού κινδύνου. Αρχικά, το ενδιαφέρον μας επικεντρώνεται στην κατηγορία μοντέλων πρόβλεψης, συμπεριλαμβανομένου του χαρακτήρα (τύπου) των επεξηγηματικών/ανεξάρτητων μεταβλητών που μπορούν να αξιοποιηθούν σε μοντέλα πιστωτικής βαθμολόγησης (credit scoring models) αλλά και νέοι αλγόριθμοι ταξινόμησης για έγκυρη και αξιόπιστη αξιολόγηση της απόδοσης των προτεινομένων μοντέλων.

Εκ πρώτης διερευνούμε την αποτελεσματικότητα των εναλλακτικών δεδομένων στα μοντέλα πιστωτικής βαθμολόγησης. Ως εναλλακτικά δεδομένα θεωρούμε τα δεδομένα που προέρχονται από μη παραδοσιακές πηγές και μπορούν να χρησιμοποιηθούν για τη συμπλήρωση παραδοσιακών δεδομένων προκειμένου να παρέχουν καλύτερες πληροφορίες που διαφορετικά δεν θα ήταν εφικτές και τα οποία θεωρούνταν μοναδικά, ασυνήθιστα ή ακριβά πριν από λίγα χρόνια.

Για το σκοπό αυτό, δημιουργήσαμε και εισαγάγαμε μεταβλητές οι οποίες προέρχονται από εναλλακτικές πηγές πληροφόρησης, σε ένα ήδη υπάρχον μοντέλο πρόβλεψης για ελληνικά ξενοδοχεία που χρησιμοποιεί μόνο δεδομένα πιστωτικής συμπεριφοράς. Για την ανάλυση αυτή χρησιμοποιήθηκε ένα πραγματικό σύνολο δεδομένων πιστωτικής βαθμολόγησης 678 ελληνικών ξενοδοχείων, το οποίο παραχωρήθηκε από την ιδιωτική βάση δεδομένων της ΤΕΙΡΕΣΙΑΣ Α.Ε. (μια εταιρεία που ιδρύθηκε από σχεδόν όλες τις τράπεζες της Ελλάδας). Συγκρίνοντας το «εναλλακτικό» μοντέλο με το ήδη υπάρχον χρησιμοποιώντας τους δείκτες απόδοσης K-S, Gini Index και ακρίβεια (accuracy), καταλήξαμε στο συμπέρασμα ότι τα εναλλακτικά δεδομένα συμβάλλουν στην απόδοση του μοντέλου. Πιο συγκεκριμένα, αυτή η συμβολή μπορεί να φανεί παρατηρώντας τις διαφορές μεταξύ των τιμών των δεικτών απόδοσης για αυτά τα δύο μοντέλα: K-S: 77,0%> 74,8%, ακρίβεια: 92,9> 91,4, Δείκτης Gini index: 0, 90> 0,88, όπου οι μεγαλύτερες τιμές αντιστοιχούν στο

εναλλακτικό μοντέλο. Έχοντας διαπιστώσει τη βελτίωση της απόδοσης του μοντέλου για τα ελληνικά ξενοδοχεία, μπορούμε εύκολα να συμπεράνουμε ότι θα ήταν συνετό να διερευνήσουμε τη χρησιμότητα εναλλακτικών δεδομένων και σε άλλους κλάδους.

Στη συνέχεια, πραγματοποιήθηκε μια συγκρητική μελέτη αξιολόγησης 12 αλγορίθμων ταξινόμησης στο ίδιο σύνολο δεδομένων για να συγκρίνουμε νέες με παραδοσιακές μεθόδους ταξινόμησης. Κατά την επιδίωξη αυτού του στόχου, συγκρίναμε αυτούς τους αλγόριθμους ταξινόμησης ως προς την AUC και την ακρίβεια. Τα αποτελέσματά μας έδειξαν ότι υπάρχουν μικρές διαφορές μεταξύ των τιμών των δεικτών απόδοσης σε κάθε ταξινομητή και πιθανόν αυτό να συμβαίνει επειδή εργαζόμαστε σε ένα ομοιογενές δείγμα. Συγκεκριμένα, παρατηρήσαμε ότι η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα είχαν καλύτερη απόδοση από άλλους (νέους ή μη) ταξινομητές και η λογιστική παλινδρόμηση είχε την υψηλότερη τιμή AUC. Το βασικό ερώτημα που δημιουργείται είναι αν τα νευρωνικά δίκτυα ή άλλοι «σύνθετοι» αλγόριθμοι ταξινόμησης μπορούν και πρέπει να αντικαταστήσουν την κλασική λογιστική παλινδρόμηση, λαμβάνοντας υπόψη τις μικρές διαφορές μεταξύ των τιμών των δεικτών απόδοσης. Με βάση την παραπάνω ανάλυση, η λογιστική παλινδρόμηση φαίνεται να αποδίδει σημαντικά ικανοποιητικά και δεν υπάρχει ζήτημα αντικατάστασής της, τουλάχιστον όσον αφορά (σχετικά) ομοιογενή δεδομένα. Σημειώνεται, ωστόσο, ότι η περαιτέρω έρευνα κρίνεται απαραίτητη για μη ομοιογενή δεδομένα.

Η συνεισφορά αυτής της ανάλυσης έγκειται αρχικά στην αξιοποίηση εναλλακτικών δεδομένων (alternative data) σε μοντέλα πρόβλεψης τα οποία παραδοσιακά χρησιμοποιούν μόνο κλασικά δεδομένα πιστωτικής συμπεριφοράς. Επιπλέον, συνεισφέρει στη σχετική βιβλιογραφία με τον καθορισμό και την αξιοποίηση μεταβλητών από εναλλακτικές πηγές πληροφόρησης με εφαρμογή στον ξενοδοχειακό τομέα. Επίσης, παρέχει πολύτιμες πληροφορίες για τους επαγγελματίες, καθώς μπορούν να εκμεταλευτούν νέους αλγόριθμους ταξινόμησης όσον αφορά τα μοντέλα πρόβλεψης. Επιπλέον, παρέχουμε μια αξιολόγηση των πρόσφατων μεθόδων βαθμολόγησης για να βοηθήσουμε τη μελλοντική έρευνα. Τέλος, αποδεικνύουμε την αποτελεσματικότητα και την ευστάθεια του μοντέλου λογιστικής παλινδρόμησης ‘τρέχοντάς το’ σε διαφορετική περίοδο και σε διαφορετικά δείγματα.

Σημαντική είναι επίσης η συνεισφορά της διατριβής όσον αφορά την εισαγωγή νέων μοντέλων πιστωτικού κινδύνου για την αξιολόγηση του

πιστωτικού κινδύνου Ελληνικών επιχειρήσεων. Συνεχίζοντας λοιπόν να στοχεύουμε στη μέγιστη απόδοση των μοντέλων πρόβλεψης και επιθυμώντας να συνεισφέρουμε στον τομέα των επιχειρήσεων καθώς και στον ευρύτερο βιομηχανικό κλάδο (όχι μόνο στον ξενοδοχειακό κλάδο), αποφασίσαμε να διερευνήσουμε ένα συνδυασμό δεδομένων σχετικά με τις ανεξάρτητες μεταβλητές που θα απαρτίζουν μοντέλα πρόβλεψης για εταιρείες. Δεδομένου ότι τα χρηματοοικονομικά δεδομένα είναι συνήθως τα μόνα δεδομένα που χρησιμοποιούνται στη μοντελοποίηση (τόσο στην Ελλάδα όσο και σε άλλες χώρες) για την αξιολόγηση του πιστωτικού κινδύνου μιας εταιρείας, αποφασίσαμε να χρησιμοποιήσουμε έναν συνδυασμό δεδομένων οικονομικής και πιστωτικής συμπεριφοράς. Σε αυτό το σημείο, η κύρια συμβολή της ανάλυσης είναι η κατασκευή νέων μοντέλων πιστωτικού κινδύνου που αξιολογούν τον πιστωτικό κίνδυνο για μικρές και μεγάλες ελληνικές επιχειρήσεις χρησιμοποιώντας έναν συνδυασμό δεδομένων χρηματοοικονομικής και πιστωτικής συμπεριφοράς.

Τα δεδομένα πιστωτικής συμπεριφοράς είναι ιδιωτικά και για τους σκοπούς αυτής της διατριβής που εστιάζει στην Ελλάδα, προέρχονται από τρία αξιόπιστα διατραπεζικά συστήματα, δηλαδή

- *the Credit Consolidation System (RCS),*
- *the Default Financial Obligation System (DFO) and*
- *the Mortgages and Prenotations to Mortgages System (MPS)*

τα οποία αναπτύχθηκαν από την ΤΕΙΡΕΣΙΑΣ Α.Ε. προκειμένου τα αποτελέσματα της ανάλυσης να βασίζονται σε πραγματικά δεδομένα και να είναι αντιπροσωπευτικά για την Ελλάδα.

Το *Credit Consolidation System (RCS)* περιέχει εταιρικά και προσωπικά δάνεια και πιστωτικές κάρτες. Περιέχει πληροφορίες σχετικά με την κατάσταση της πίστωσης (π.χ. τρέχον υπόλοιπο χωρίς καθυστέρηση, υπόλοιπο με καθυστέρηση). Η λειτουργία της τράπεζας δεδομένων είναι να διασφαλίζει τη συλλογή δεδομένων από πιστωτικά / χρηματοπιστωτικά ιδρύματα σχετικά με πιθανό χρέος από δάνεια, την επεξεργασία τους, τον έλεγχο πληρότητας και τη διάδοση των επεξεργασμένων πληροφοριών. Τα δεδομένα του RCS διοχετεύονται στην ΤΕΙΡΕΣΙΑΣ από πιστωτικά ιδρύματα, εταιρείες χρηματοδότησης, χρηματοδοτικές μισθώσεις, εταιρείες έκδοσης / διαχείρισης καρτών.

Το *Default Financial Obligation System (DFO)* περιέχει δεδομένα σχετικά με την πιστωτική συμπεριφορά ατόμων και εταιρειών (π.χ. ακάλυπτες επιταγές, ανακοινώσεις δημοπρασίας εκκαθάρισης, πτωχεύσεις).

Το *Mortgages and Prenotations to Mortgages System (MPS)* περιέχει δεδομένα σχετικά με υποθήκες, προσημάνσεις σε υποθήκες και μετατροπές προσημάνσεων σε υποθήκες. Τόσο το DFO όσο και το MPS επιτρέπουν στις τράπεζες να ολοκληρώνουν και να υποστηρίζουν μια πιο έγκυρη αξιολόγηση της οικονομικής αξιοπιστίας ενός πελάτη (τρέχουσα ή μελλοντική) από τις τράπεζες.

Στη συνέχεια, τα προτεινόμενα μοντέλα (με το συνδυασμό δεδομένων) συγκρίθηκαν με τα παραδοσιακά μοντέλα (που περιέχουν μόνο οικονομικά δεδομένα) χρησιμοποιώντας τρεις δείκτες απόδοσης, την ακρίβειας, το K-S και το Gini Index.

Μετά τη σύγκριση των μοντέλων, καταλήξαμε στο συμπέρασμα ότι τα νέα μοντέλα συμβάλλουν στην εκτίμηση του πιστωτικού κινδύνου όπως φαίνεται από την απόδοσή τους. Πράγματι, οι διαφορές φαίνονται εύκολα στους Πίνακες 1 και 2.

Τέλος, η αποτελεσματικότητα και η σταθερότητα των μοντέλων μελετήθηκαν και αποδείχτηκαν, προκειμένου να μπορούν να χρησιμοποιηθούν σε διαφορετικές χρονικές περιόδους καθώς μόνο σε μιά τέτοια περίπτωση είναι χρήσιμα.

- **Μικρές επιχειρήσεις:**

Δείκτες Απόδοσης	Μοντέλο με συνδυασμό δεδομένων	Μοντέλο μόνο με οικονομικά δεδομένα
Ακρίβεια	85,5%	71,0%
K-S	64,6%	33,8%
Gini Index	0,80	0,44

Πίνακας 1: Σύγκριση τιμών των δεικτών απόδοσης για τις μικρές επιχειρήσεις

- **Μεγάλες επιχειρήσεις:**

Δείκτες Απόδοσης	Μοντέλο με συνδυασμό δεδομένων	Μοντέλο μόνο με οικονομικά δεδομένα
Ακρίβεια	85,3%	78,8%
K-S	67,2%	41,2%
Gini Index	0,82	0,51

Πίνακας 2: Σύγκριση τιμών των δεικτών απόδοσης για τις μεγάλες επιχειρήσεις

Τέλος η διατριβή συμβάλει και στην ανάλυση μεγάλης κλίμακας δεδομένων αφού πραγματεύεται το πρόβλημα της επιλογής μεταβλητών που σε συνδυασμό με την αξιοποίηση της τεχνικής μείωσης της διάστασης επιτυγχάνει την κατασκευή ευέλικτων και αξιόπιστων μοντέλων ταξινόμησης που αφορούν στις Ελληνικές επιχειρήσεις (ένα μοντέλο για μικρές και ένα για μεγάλες επιχειρήσεις), βάσει της πιστωτικής συμπεριφοράς τους.

Πιο συγκεκριμένα, ο σκοπός αυτής της ανάλυσης είναι η ανάπτυξη μιας ευέλικτης και αξιόπιστης προσέγγισης μοντέλων ταξινόμησης για μια μεταβλητή απόκρισης που αντιπροσωπεύει την επιχειρηματική πιστωτική συμπεριφορά που χαρακτηρίζεται σύμφωνα με τη Βασιλεία ΙΙ, ως «καλή» (δηλαδή, χωρίς «παραβατικότητα») ή «κακή» (δηλαδή, με «σοβαρή παραβατικότητα») με μεταβλητές που σχετίζονται όχι μόνο με οικονομικά χαρακτηριστικά αλλά και με χαρακτηριστικά πιστωτικής συμπεριφοράς.

Πιο αναλυτικά, κακή πιστωτική συμπεριφορά αφορά επιχειρήσεις που:

Παρουσιάζουν αυστηρή καθυστέρηση δηλαδή:

- Συμβόλαια επιχειρήσεων, όχι overdrafts με μέγιστη καθυστέρηση το τελευταίο 12μηνο >=των 90 ημερών.
- Επιχειρήσεις με overdrafts με μέγιστη καθυστέρηση το τελευταίο 12μηνο >=90 ημερών ή με χρήση >102% για περισσότερο από 90 μέρες και ποσό >100 ευρώ.

Επίσης στην περίπτωση που υπάρχουν εγγυητές ισχύουν τα εξής:

- Συμβόλαια επιχειρήσεων, όχι overdrafts με μέγιστη καθυστέρηση το τελευταίο 12μηνο >=των 150 ημερών.
- Επιχειρήσεις με overdrafts με μέγιστη καθυστέρηση το τελευταίο 12μηνο >=150 ημερών ή με χρήση >102% για περισσότερο από 90 μέρες και ποσό >100 ευρώ.

➢ Επίσης, επιχειρήσεις με κακή πιστωτική συμπεριφορά θεωρούνται επιχειρήσεις με καταγγελία δανείου μέσα στο 12μηνο.

Σε αντιδιαστολή, καλή πιστωτική συμπεριφορά παρουσιάζουν οι επιχειρήσεις με μέγιστη καθυστέρηση το τελευταίο 12μηνο από 0 έως 29

ημέρες ή η χρήση να είναι >102% για 0 έως 29 μέρες για τις επιχειρήσεις με overdrafts.

Για τη μοντελοποίηση, προτείνουμε μια αλγοριθμική διαδικασία 3 (4) βημάτων για τη μείωση των διαστάσεων με ένα αρχικό στάδιο προκαταρκτικής επεξεργασίας δεδομένων (βήμα 0) το οποίο πραγματοποιήθηκε και στις προηγούμενες αναλύσεις και είναι το εξής:

Χρησιμοποιήσαμε την κωδικοποίηση Weight-of-Evidence (WOE) για να δημιουργήσουμε ψευδομεταβλητές προκειμένου να ομαδοποιήσουμε όλες τις ανεξάρτητες μεταβλητές. Οι ελλειπούσες τιμές (εγγραφές που δεν περιέχουν όλες τις πληροφορίες δεδομένων) ομαδοποιήθηκαν ξεχωριστά. Αυτή η διαδικασία έχει τα ακόλουθα οφέλη:

• Εξαλείφει το πρόβλημα των outliers και των σπάνιων κλάσεων (μικρής συχνότητας κατηγοριών-rare classes).

• Κατανοούμε καλύτερα τις σχέσεις με την ομαδοποίηση, καθώς ένα γράφημα που δείχνει τις σχέσεις μεταξύ των χαρακτηριστικών μιας μεταβλητής και της απόδοσης είναι ευκολότερα αντιληπτό από ένα στατιστικό στοιχείο μεταβλητής ισχύος ενώ ταυτόχρονα διευκολύνει την εξήγηση της φύσης αυτής της σχέσης, εκτός από την ισχύ της σχέσης.

• Οι μη γραμμικές εξαρτήσεις μπορούν να μοντελοποιηθούν με γραμμικά μοντέλα.

Η χρήση ψευδομεταβλητών για κατηγορικές μεταβλητές έχει ένα σοβαρό μειονέκτημα - υποθέτει ότι η διαφορά από τη μια ομάδα κατηγορικής μεταβλητής στην επόμενη είναι η ίδια. Ένας καλύτερος τρόπος αντιμετώπισης των ομαδοποιημένων μεταβλητών είναι να χρησιμοποιηθεί το WOE κάθε ομάδας ως μεταβλητή εισόδου. Αυτό όχι μόνο επιλύει τα προβλήματα διαφορετικών μονάδων εισόδου, αλλά επίσης λαμβάνει υπόψη την ακριβή τάση και κλίμακα της σχέσης από τη μία ομάδα στην άλλη. Επιπλέον, εάν η ομαδοποίηση γίνει σωστά, αυτό θα διασφαλίσει επίσης ότι η κατανομή των εγγραφών σε κάθε ομάδα κατά τη διάρκεια της κλιμάκωσης είναι λογική και αντιπροσωπεύει τη διαφορά στη σχέση μεταξύ των ομάδων.

Το κύριο τμήμα του αλγορίθμου βασίζεται σε τεχνικές μείωσης διάστασης λαμβάνοντας υπόψη το σταδιακό κριτήριο πληροφοριών Akaike και την ανάλυση κυρίων συνιστωσών (PCA). Η προτεινόμενη διαδικασία επιτρέπει ένα προαιρετικό 4ο βήμα που βασίζεται στην Elastic Net Regularization για περαιτέρω μείωση της διάστασης εάν ο ερευνητής πιστεύει ότι αυτό είναι χρήσιμο.

Τα ευρήματα αυτής της ανάλυσης δείχνουν σαφώς τη σημασία στη χρήση μεταβλητών πιστωτικής συμπεριφοράς, δεδομένου ότι ορισμένες από αυτές τις μεταβλητές βρέθηκαν να διαδραματίζουν βασικό ρόλο στη δημιουργία μοντέλων πιστωτικής βαθμολόγησης τόσο για τις μικρές όσο και για τις μεγάλες επιχειρήσεις. Πράγματι, στο τελικό μοντέλο για τις μικρές επιχειρήσεις, κάθε μεταβλητή PCA εξαρτάται από 6 μεταβλητές πιστωτικής συμπεριφοράς (από ένα σύνολο 15 μεταβλητών) ενώ για το τελικό μοντέλο μεγάλων επιχειρήσεων κάθε μεταβλητή PCA εξαρτάται από 10 μεταβλητές πιστωτικής συμπεριφοράς (από ένα σύνολο 18 μεταβλητών). Η χρήση τέτοιων συνδυασμών είναι μια από τις κύριες συνεισφορές της παρούσας διατριβής, δεδομένου ότι οι χώρες βασίζονται σχεδόν αποκλειστικά στις οικονομικές μεταβλητές. Αξίζει επίσης να σημειωθεί ότι η προτεινόμενη μεθοδολογία ανταποκρίνεται στην ανάγκη μείωσης των διαστάσεων για την κατασκευή ευέλικτων αλλά και αξιόπιστων μοντέλων πιστοληπτικής ικανότητας όχι μόνο για περιγραφικούς αλλά και κυρίως για προβλεπτικούς σκοπούς. Επιπλέον, η προτεινόμενη μεθοδολογία παρέχει, μεταξύ άλλων, στους ασφαλιστές, στους χρηματοοικονομικούς σχεδιαστές και στους δανειστές ένα αυτοματοποιημένο αξιόπιστο χρηματοοικονομικό εργαλείο αξιολόγησης της πιστοληπτικής ικανότητας σύμφωνα με μερικές στατιστικά σημαντικές χρηματοοικονομικές και πιστωτικές μεταβλητές και ταυτόχρονα τη λήψη πιστωτικών αποφάσεων γρηγορότερα και πιο δίκαια ενώ προσφέρει στους δανειολήπτες αυξημένες ευκαιρίες δανεισμού.

Επίσης, το μοντέλο μείωσης της διάστασης που προτείνεται μπορεί να εφαρμοστεί στη μοντελοποίηση της πιστοληπτικής ικανότητας φορολογικού χρέους. Η σημασία της πρόβλεψης που προκύπτει από τους οργανισμούς αξιολόγησης φορολογικού χρέους είναι ένα άλλο πεδίο πιθανών επεκτάσεων. Η σημασία των προβλέψεων που αφορούν τόσο τα μοντέλα αξιολόγησης πιστωτικού κινδύνου όσο και τα μοντέλα αξιολόγησης φορολογικού χρέους μπορεί να δοκιμαστεί χρησιμοποιώντας μη παραμετρικές μεθόδους.

Εν κατακλείδι, συνοψίζουμε παρακάτω τους κύριους στόχους, τα κύρια χαρακτηριστικά και τη συνεισφορά της παρούσας διατριβής:

• Ο κύριος στόχος αυτής της διατριβής είναι η πρόταση τόσο για περιγραφικούς όσο και για προβλεπτικούς σκοπούς, μιας καινοτόμου ευέλικτης και αξιόπιστης προσέγγισης για τη μοντελοποίηση της πιστοληπτικής ικανότητας, η οποία έχει σημαντική σημασία στη

χρηματοδότηση και την τραπεζική λόγω της άμεσης σύνδεσής της με την πιστοληπτική ικανότητα.

• Η πρωτοτυπία και μία από τις κύριες συνεισφορές της προτεινόμενης μεθοδολογίας μοντελοποίησης έγκειται στο γεγονός ότι συνδυάζουμε αποτελεσματικά οικονομικά χαρακτηριστικά μαζί με χαρακτηριστικά πιστωτικής συμπεριφοράς αλλά και εναλλακτικά δεδομένα που δεν έχουν εξεταστεί ποτέ πριν καθώς οι περισσότερες χώρες και ιδρύματα χρησιμοποιούν μόνο οικονομικά δεδομένα για τη μοντελοποίηση της πιστωτικής βαθμολόγησης.

• Πραγματοποιήθηκε μια συγκριτική μελέτη αξιολόγησης δώδεκα αλγορίθμων ταξινόμησης σε ένα πραγματικό σύνολο δεδομένων πιστοληπτικής ικανότητας για τη σύγκριση καινοτόμων και παραδοσιακών μεθόδων ταξινόμησης που προσφέρουν πολύτιμες γνώσεις τόσο στους επαγγελματίες όσο και στους μη επαγγελματίες.

• Μια αποτελεσματική και φιλική προς τον χρήστη αλγοριθμική διαδικασία που έχει προταθεί και εφαρμοστεί στη διατριβή αποτελεί μία ακόμα συμβολή δεδομένου ότι ανταποκρίνεται στην ανάγκη μείωσης της διάστασης, ένα ζήτημα που συναντάται συχνά στην πράξη, ειδικά σε προβλήματα που ταξινομούνται στην περιοχή της Ανάλυσης Μεγάλης Κλίμακας Δεδομένων (Big Data). Από όσο γνωρίζουμε, αυτή είναι η πρώτη φορά που ο συνδυασμός των παραπάνω τεχνικών πολλαπλών επιπέδων μείωσης διάστασης χρησιμοποιείται και εφαρμόζεται αποτελεσματικά, στη μοντελοποίηση της πιστοληπτικής ικανότητας.

• Τέλος, παρέχουμε μια αξιολόγηση των πρόσφατων μεθόδων πιστωτικής βαθμολόγησης για να συνδράμουμε τη μελλοντική έρευνα.

Table of Contents

Chapter 1
Introduction

1.1 Scope of the thesis

Credit risk is one of the major threats that financial institutions face. To that end, we are interested in contributing to areas such as predictive methods and the selection of independent variables for scorecard construction in order to boost the performance of credit risk models.

More specifically, the objective of this thesis is the proposal for descriptive (classification) as well as predictive purposes, of an innovative approach to flexible and accurate credit scoring modelling which is of significant importance in Finance and Banking due to its direct connection to one's creditworthiness. The originality and one of the main contributions of the proposed modelling methodology lies on the fact that we blend effectively financial features together with credit behavior characteristics and alternative data that have never been considered before and it is quiet original as most countries and institutions use only financial data for credit scoring modelling. Furthermore, we perform a benchmarking study of twelve classification algorithms on a real-world credit scoring data set in order to compare novel to traditional classification methods. This analysis provides valuable insights for professionals as they can see novel classification algorithms in predictive modelling. We also provide an evaluative survey of recent scoring methods to aid future research. Subsequently, an algorithmic procedure that has been proposed and implemented into the methodology constitutes yet, another contribution since it is responsive to the need for dimension reduction, an issue frequently encountered in practice, especially in problems classified as falling into the area of Big Data Analysis. For this, we rely on modern regularization and classification methods which ensure the construction of flexible yet, reliable credit scoring models. To the best of our knowledge this is the first time that the combination of the above multivariate techniques is being used and implemented effectively, into credit scoring modelling. Finally, the problem of dimension reduction in credit scoring modelling is addressed by combining Regularization methods and model identification techniques.

In what follows in this Chapter we will provide the basic definitions and characteristics of the topic under investigation for a better understanding of the subject together with the overview of the thesis (Section 1.4).

1.2 Definition of credit risk

Credit risk is the probability of loss due to inability of the borrower to fulfill contractual obligations of a contractor and it is linked to the following key risk parameters, namely:

- *Probability of Default*
- *Exposure at Default*
- *Loss given Default*
- *Recovery Rate*

Definition 1.1

a) *Exposure at default (EAD)* is the amount to which a financial or a credit institution is exposed to the borrower at the time of default, measured in currency.
b) *Loss given default (LGD)* is the magnitude of likely loss on the exposure, expressed as a percentage of the exposure.
c) *Probability of default (PD)* is the probability of default of a contractual obligation (i.e., debt repayment) within a certain period.
d) *Recovery rate* is the percentage of the defaulted amount that can be recovered and is equal to 100% - LGD.

Note that the first three risk parameters presented in the above definition are directly related to the *expected loss (EL)* which is defined as the product of EAD, LGD and PD.

A credit institution considers that the borrower is reasonably likely to default on all payment obligations when there is a delay of more than or equal to 90 days on a liability. (source: ISDA, 1999 Credit Derivative Definitions, http://credit-deriv.com/isdadefinitions.htm , Basel Committee on Banking Supervision, Basel 2 Accord, http://www.bis.org/publ/bcbsca.htm).

1.3 Risk assessment approaches

In order to assess the credit risk, the credit worthiness of the counterparty must be determined, namely the availability and the possibility of repayment. Risk assessment can be determined by:

- **Expert systems/rating** which consider quantitative and qualitative information and involve the analyst's experience and judgment (e.g., Holsapple and Whinston, 1987).

"Credit rating" means an opinion regarding the creditworthiness of an entity. (source: Regulation (EC) No 1060/2009 of the European Parliament and of the Council on Credit rating agencies).

- **DuPont model/analysis**

A DuPont analysis is used for the evaluation of the three components that constitute an institution's return on equity (ROE) and is given by:

*DuPont Analysis = (Profit Margin) * (Asset Turnover) * (Equity Multiplier)*

where

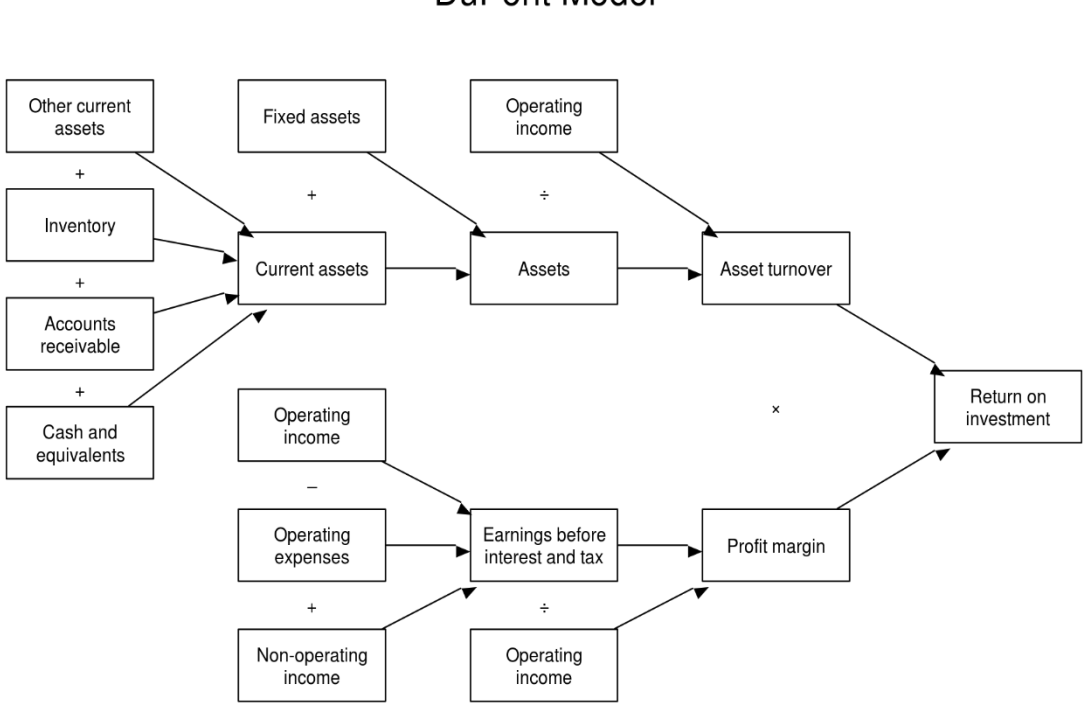
Profit Margin = Net Income / Revenue

Asset Turnover = Sales / Average Total Assets

Equity Multiplier = Average Total Assets / Average Equity

This type of modelling allows the researcher to understand the source of superior (or inferior) return by comparison with companies in similar industries (or between industries).

Diagram 1: Graphical representation of DuPont analysis.



(source: https://en.wikipedia.org/wiki/DuPont_analysis)

- **Market models** which utilize mathematical models like the Black-Scholes and Merton model (Black and Scholes, 1973; Merton, 1974; Shinde and Takale, 2012; Hull *et al*., 2005) which consider market information (e.g., stock indices) and require a developed and well-functioning market.
- **Credit Scoring**, which is the subject of this thesis and will be fully discussed in Chapter 2.

Credit scoring can be defined as "… *the use of statistical models to transform relevant data into numerical measures that guide credit decisions*." (Anderson, 2007) and concerns the methods and techniques of modeling the creditworthiness of the borrower.

Another definition of credit scoring was given by Thomas *et. al* (2002):

"Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders".

Credit scoring is based on credit history, i.e., on existing customer data it usually covers a specific period and is related to borrower's credit behavior like the repayment history, types of loans, the borrower's total debt etc. The purpose of these methods is to identify the most important factors and contribution of each in measuring the borrowers' creditworthiness. The proposed credit scoring models are interwoven with standard and advanced statistical methods including

- *Multivariate Statistical Analysis,*
- *Generalized Linear Models (GLM),*
- *Discriminant Analysis (DA) and*
- *Principal Component Analysis (PCA).*

1.4 Thesis overview

This thesis focuses on credit risk assessment in relation to credit rating models and consists of 6 Chapters. After the first Chapter which describes credit risk, **Chapter 2** presents in detail the concept of credit scoring.

In **Chapter 3**, considering that the interest in predictive modelling is endless and that there has been much advancement in this area, including

- the nature (character) of independent variables/factors that can be used in credit scoring models (e.g., Pendharkar, 2005; Fletcher and Goss, 1993; Jo *et al.*, 1997; Desai *et al.*, 1996; Tam and Kiang, 1992; Salchenberger *et al.*, 1992; Leshno and Spector, 1996) and
- benchmarking studies of classification algorithms for credit scoring (e.g., Xiao *et al*., 2006),

we wish to contribute to both these areas.

The first objective of this Chapter is to introduce alternative data (which we created) into predictive models that typically (traditionally) use only credit behavior data in order to explore their significance and contribution to models' performance.

The second objective is to perform a benchmarking study of twelve classification algorithms on a real-world credit scoring data set in order to compare novel to traditional classification methods. This analysis provides valuable insights for professionals as they can explore the capabilities of novel classification algorithms in predictive modelling. They can also estimate, based on the results, if it is advantageous to change 'traditional'

logistic regression analysis for novel classification algorithms in corporate practice. Furthermore, we provide an evaluative survey of recent scoring methods to aid future research.

In **Chapter 4** we continue to further improve the credit scoring system by combining different types of data. More specifically, we combine financial data which are more widely used with credit behavior data that are not usually used and we propose new credit risk models for small and large Greek enterprises. **The purpose of this Chapter** is to maximize the performance of these models and to investigate the impact and the contribution of credit behavior data to this performance. The use and evaluation of credit behavior variables is one of the main contributions of this Chapter since countries rely almost solely on the financial variables.

Chapter 5 describes and applies a dimension reduction algorithm in order to identify the statistically significant variables that prevail and then to build reliable models for predicting credit behavior. The proposed methodology is responsive to the need of dimension reduction for the construction of flexible yet reliable credit scoring models not only for descriptive but most importantly for predictive purposes. Furthermore, the proposed methodology provides among others, insurers, financial planners and lenders with an automated reliable financial tool of evaluating credit worthiness according to a few statistically significant financial as well as credit behavior covariates and at the same time making credit decisions faster and fairer while offering to borrowers increased lending opportunities.

Finally, in **Chapter 6** we have the summary and the concluding remarks of our study.

Chapter 2
Introduction to credit scoring

2.1 Credit scoring and scorecards

Credit scoring is an objective indicator of the probability of default that is attempting to distinguish 'good' from 'bad' borrowers (where 'good' are considered to be the good payers and 'bad', the bad payers) using available characteristics (e.g., demographics, economic behavior data), ranking on a numerical score the candidate borrowers. This score ranges from 0 to 600. The closer to 0 the score of the candidate borrower, the higher the risk of default while the closer to 600 the score, the lower the risk of default.

A scorecard is a tool which supports decision making in the credit industry. Let X be a n-dimensional vector of variables/components that characterizes an application for a credit product (e.g., loan). The performance of previously approved loans is known to the decision maker. Let δ be a binary response variable that shows whether a default event was observed for the loan taking the values of 0 and 1 corresponding respectively to performing and non-performing loans. When deciding on an application with characteristics x, it is important to have an estimate of the posterior probability $p(+1/x)$ that the loan will turn out to be non-performing if it is granted. A scorecard provides such an estimate. Then, the decision maker can compare the model-estimated $p(+1/x)$ to a threshold τ, approving the loan if $p(+1/x) \leq \tau$, and rejecting it otherwise. The problem of estimating $p(+1/x)$ belongs to the field of classification analysis (e.g., Hand, 1997). Specifically, a scorecard is the result of applying a classification algorithm to a data set of past loans.

For a better understanding of credit scoring an example of a scorecard with two borrowers and a 3-dimensional vector X, is presented below.

Consider the following scorecard:

SCORECARD		
CHARACTERISTIC	VALUE	SCORE
Total debt (€)	< 10.000 10.000-50.000 > 50.000	120 150 210
Maximum Current Delinquency (months)	0 ή 1 2 ≥ 3	250 80 -50
Age (date of previous approval)	< 2 έτη ≥ 2 έτη	80 120

Suppose that borrower 1 has the following characteristics:

SCORECARD		
CHARACTERISTIC	VALUE	SCORE
Total debt (€)	< 10.000 15.000 > 50.000	120 150 210
Maximum Current Delinquency (months)	0 2 ≥ 3	250 80 -50
Age (date of previous approval)	< 2 έτη ≥ 2 έτη	80 120

Then, borrower's 1 score is 150 + 250 + 80 = 480.

Suppose that borrower 2 has the following characteristics:

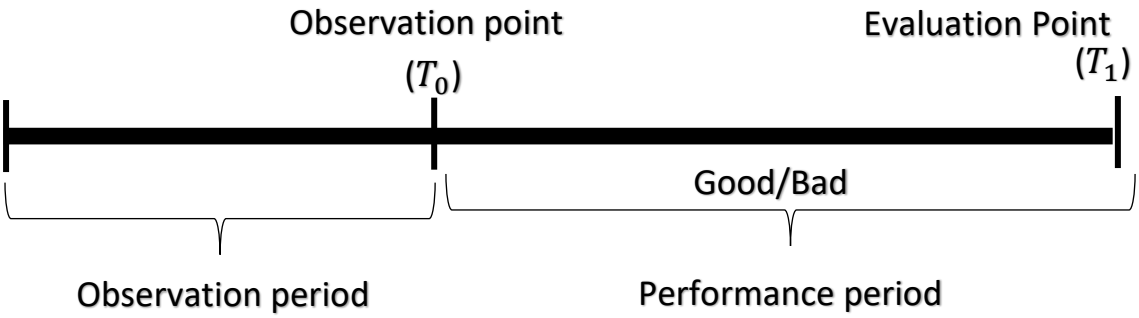
SCORECARD		
CHARACTERISTIC	VALUE	SCORE
Total debt (€)	< 10.000 15.000 >50.000	120 150 210
Maximum Current Delinquency (months)	0 2 ≥3	250 80 -50
Age (date of previous approval)	< 2 έτη ≥ 2 έτη	80 120

Then, borrower's 2 score is 210-50+120=280.

2.2 Behavioral score

There are different kinds of scores according to the stage of the analysis. Specifically, in the pre-application stage there is the response score. In the application stage there is the application score and the fraud score. In the performance stage there is the performance score, the behavioral score, the retention score and the early warning score. Finally, in the collection stage there is the collection score.

In this thesis we focus on the Behavioral score for the intention to assess the probability of default (PD) defined in Definition 1.1. As default we consider the period of 90 days or more of delinquency of an obligation. Also, the period of 24 to 60 months is the observation period, and the period of 12 months is the performance period where it is defined whether the candidate borrower is good or bad.



(source: Basel Committee on Banking Supervision, Basel 2 Accord, http://www.bis.org/publ/bcbsca.htm)

2.3 A brief review of the literature on credit scoring

Credit evaluation is one of the most critical procedures in banks' credit management decisions. This procedure contains collecting, analyzing and classifying different credit variables to assess credit decisions. Hand and Jacka (1998) stated that *'the process (by financial institutions) of modelling credit worthiness is referred as credit scoring'*. Several alternative definitions can be found in the literature (e.g., Anderson, 2007; Beynon, 2005; Lewis, 1992; Bailey, 2001; Mays, 2001; Siddiqi, 2006; Chuang and Lin, 2009; Sustersic *et al*., 2009), three of which are provided below:

Definition 2.1a (Beynon, 2005). *Credit scoring can be simply defined as the use of statistical models to transform relevant data into numerical measures that guide credit decisions. It is the industrialization of trust; a logical future development of the subjective credit ratings*

Definition 2.1b (Gup and Kolari, 2005). *Credit scoring is the use of statistical models to determine the likelihood that a prospective borrower will default on a loan. Credit scoring models are widely used to evaluate business, real estate, and consumer loans*.

Definition 2.1c (Thomas *et al*., 2002). *Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders*.

Due to the importance of credit scoring, the field remains in the center of attention with numerous research works over the last 30 years.

On the one hand, some studies have been focused on *the characteristics that should be used in credit scoring applications* (e.g., Pendharkar, 2005; Fletcher and Goss, 1993; Jo *et al*., 1997; Desai *et al*., 1996; Tam and Kiang, 1992; Salchenberger *et al*., 1992; Leshno and Spector, 1996). **On the other hand,** some studies have been focused on *the effectiveness of different classification algorithms for credit scoring* (e.g., Finlay, 2011; Paleologo *et al*., 2010; West *et al*., 2005). **Furthermore**, there are *benchmarking studies of classification algorithms for credit scoring* (e.g., Xiao *et al*., 2006). However, some of them are of limited scope and consider only a few classifiers (Dreiseitl and Ohno-Machado, 2002), while some others do not include most contemporary classification algorithms (e.g., King *et al.,* 1995).

Applications of credit scoring are scattered in various fields, including a comparison between different statistical techniques used in prediction purposes and classification problems. These applications can be classified into

- ***accounting and finance*** (Landajo *et al.,* 2007; Pendharkar, 2005; Altman *et al*., 1994),
- ***marketing*** (Chiang *et al.*, 2006),
- ***engineering and manufacturing*** (Dvir *et* al., 2006),
- ***health and medicine*** (Behrman *et al*., 2007) and
- ***general application*** (Nikolopoulos *et al*., 2007), as noted by (Paliwal and Kumar, 2009).

Particularly, ***in corporate credit scoring models*** (Altman, 2005; Paleologo *et al*., 2010) several steps must be included (Altman and Haldeman, 1995):

- The first step is to apply the primary client-data to credit scoring model.
- In the second step, the model requires test that cover the following issues: 'definition of risk, model development, test of time, stability, public versus private company data, probability of failure, credibility, model support and pilot testing' (Altman and Haldeman, 1995).
- The third step includes: 'smoothing out the wave, firm-capital market approach' (namely, using systematic beta risk) and 'firm econometric approach' (for further explanation of these steps, see Altman and Haldeman, 1995).

The classification of good and bad credit is the purpose of a credit scoring model. The question is what determines the classification of a new applicant. **For small businesses and corporate loans**, some of the characteristics that have been used in scoring applications are

- the main activity of the business,
- age of business,
- business location,
- credit amount and
- other financial ratios, such as profitability, liquidity, bank loans and leverage

(see Emel *et al*., 2003; Bensik *et al*., 2005; Zekic-Susac *et al*., 2004; Min and Lee, 2008; Min and Jeong, 2009; Lensberg *et al*., 2006; Cramer, 2004; Liang, 2003). Sometimes, the final selection of the variables is based on the statistical analysis used, namely

- *stepwise logistic regression,*
- *regression* or
- *neural networks*

(see Lee and Chen, 2005; Lenard *et al*., 1995; Steenackers and Goovarts, 1989).

Moreover, ***the classification techniques*** can be also discriminated into conventional and advanced methods. The first one includes

- *Weight of Evidence (WOE),*
- *Multiple Linear Regression,*
- *Discriminant Analysis,*
- *Probit Analysis and*
- *Logistic Regression.*

The other one comprises approaches and methods such as

- *Fuzzy Algorithms,*
- *Genetic Algorithms,*
- *Expert Systems and*
- *Neural Networks*

(see Hand and Henley, 1997).

The selection of the characteristics differs from study to study because of the nature of data and the cultural or economic variables that may affect the quality of the model and be appropriate to a particular market whose variables varies from country to country. In finance applications, a rank from only three variables (Pendharkar, 2005; Fletcher and Goss, 1993) to about twenty variables (Jo *et al*., 1997; Desai *et al*., 1996; Tam and Kiang, 1992) has been used in building scoring models. However, there are others who have used more variables, such as in (Salchenberger *et al*., 1992), who used twenty-nine variables and (Leshno and Spector, 1996), who used forty-one variables.

Finally, the determination of the sample size is another issue. The sample size depends on the data availability, the nature of the market and to what extend the sample is representative of the entire population. In some studies, a small number of observations have been used, around three-four dozen (Dutta *et al*., 1994; Fletcher and Goss, 1993), while others have used thousands of observations (Bellotti and Crook, 2009; Hsieh, 2004; Banasik *et al*., 2003).

2.4 Stages of credit scoring

A scorecard comprises different stages:

- collecting and preparing data,
- assessing a credit score using a formal induction algorithm,
- development,
- monitoring and
- recalibration of the scorecard.

These stages have been investigated in the literature. For example:

A. Data collecting and preparation:
1. how to handle missing values (e.g., Florez-Lopez, 2010),
2. the collection of a set of independent variables (e.g., Falangis and Glen, 2010; Liu and Schumann, 2005), and
3. how biases due to an underrepresentation of bad risk in scoring data sets (e.g., Brown and Mues, 2012; Marques *et al*., 2013; Paleologo *et al*., 2010) or the problem that
4. repayment behavior is only perceptible for previously accepted (seemingly good) customer, can be surmounted (e.g., Banasik and Crook, 2007; Banasik *et al*., 2003; Wu and Hand, 2007).

B. When the data set is available, many different prediction methods accommodate assessing different aspects of credit risk. Particularly, the Basel 2 Capital Accord demand financial institutions, who use an internal rating approximation, to develop the following types of prediction models, namely EAD, LGD, PD defined in Definition 1.1(a)-(c).

The development of EAD and LGD prediction models have been recently explored (e.g., Bellotti and Crook, 2012; Loterman *et al*., 2012; Somers and Whittaker, 2007).

Nevertheless, most credit scoring studies focus on PD modelling using either classification or survival analysis. Survival analysis models predict default probabilities for different time periods. This is important for estimating when a customer will default (e.g., Bellotti and Crook, 2009b; Stepanova and Thomas, 2002; Tong *et al*., 2012). Classification analysis benefits from many different modelling methods and represents the biggest part of the literature as a modelling approach.

C. Last but not least, is the consecutive monitoring of scorecard performance after development to explore its robustness towards changes in customer behavior and the recalibration of the scorecard when its performance relegates (e.g., Pavlidis *et al*., 2012; Sohn and Ju, 2014; Thomas *et al*., 2001).

2.5 Performance definition for companies

The evaluation of factors affecting sampled companies to shape their financial behavior in the final probability of repayment or not of their administration, will be via a statistical method used to predict the probability of a specific event to occur (the fact of repayment specifically). The models' specification is based on information that will prove to be characteristic of the future financial behavior. These models categorize businesses' rating based on the risk of default on their obligations. The characteristics that are contained in these models are data with information from the past and present. There are two periods that are studied during the models' creation, the observation period and the performance period. These models are intended to discriminate the 'bad' from 'good' behavior in the performance period. First, we must specify what we mean by 'bad' and 'good' credit behavior of a company:

'Bad' are these companies showing 'severe delinquency' in the service of their obligations, which means:

a) SME Contracts (business loans), not Overdrafts with maximum delinquency in the last 12 months greater or equal to 90 days past due.
b) SME Overdrafts (business overdrafts) with maximum delinquency in the last 12 months, greater or equal to 90 days past due or credit limit utilization over 102% for time period greater or equal to 90 days with over limit amount greater than 100 euros. Where

$$utilization = \frac{current\ balance}{credit\ limit} * 100$$

- For the case of Guarantor, the characterization "Bad" refers to case (b) above with 150 instead of 90 days or more.

It should be also mentioned that a company is characterized as having a "Bad" credit behavior if during the performance period, a new DFO (*loan denunciation*) has occurred.

Companies with a "Good" credit behavior are companies with no delinquency, namely with either maximum delinquency from 0 to 29 days past due, during the last 12 months or with credit limit utilization over 102% for 0 to 29 days, concerning Small and Medium-sized Enterprises (*SME*) Overdrafts.

There is also another category called 'Indeterminate' where there is a maximum delinquency in the last 12 months from 30 to 90 days. This type of companies does not take part in the analysis because they do not have discriminant ability.

2.6 Content and source of credit behavior data

In this Section we will discuss what type of information is contained in credit behavior data and from where they come from. We must first mention that credit behavior data are private and for the purpose of this Thesis which focuses on Greece, was taken from three reliable inter-bank systems, namely

- *the Credit Consolidation System (RCS),*
- *the Default Financial Obligation System (DFO) and*
- *the Mortgages and Prenotations to Mortgages System (MPS)*

developed by Tiresias S.A. (http://www.tiresias.gr/) (a company founded by all banks in Greece) in order for the results of our analysis to be based on real data and to be representative of Greece.

Credit Consolidation System (RCS) contains corporate and personal loans and credit cards. It contains information about the status of the credit (e.g., current balance with no delinquency, delinquent balance). The function of the Databank is to secure the collection of data from credit/financial institutions regarding possible debt from loans, their processing, the completeness control, and the dissemination of the processed information. The data of the RCS are channeled to Tiresias from credit institutions, funding companies, leasing, card issuing/ managing companies' provisions.

The Default Financial Obligation System (DFO) contains data concerning the credit behavior of individuals and companies (e.g., bounced checks, liquidation auction announcements, bankruptcies). More details are shown in Table 2.1.

The Mortgages and Prenotations to Mortgages System (MPS) contains data regarding mortgages, prenotations to mortgages, and conversions of prenotations to mortgages. Both DFO and MPS allows the banks to complete and support a more accurate assessment of a client's financial credibility (current or future) by the banks.

The data sources and the Default Financial Obligation System data categories are provided in Table 2.1.

Data sources	Categories
Banks and financial institutions	Bounced cheques
	Unpaid bills of exchange
	Filings of debt adjustment and discharge
	Debt adjustment judgment
	Termination of consumer/ housing/ business loan or credit card contracts
Courts of first instance	Filings for bankruptcy
	Judgments rejecting filings for bankruptcy due to insufficient wealth of the debtor
	Adjudicated bankruptcies
	Issued orders of payment
	Orders for the restitution of use of leased property
Magistrate's Court	Filings for bankruptcy
	Reconciliation/ rehabilitation procedures
	Liquidation auction announcements of real property
	Liquidation auction announcements of chattels
	Filings for debt adjustment and discharge
	Debt adjustment judgment
Registries of deeds/ Cadastral offices	Mortgages and prenotations of mortgages
	Conversions prenotations to mortgages
	Forfeitures of the Legislative Degree
Ministry of Finance	Administrative sanctions against tax law violators

Table 2.1: Source and category of data-DFO

(source: http://www.tiresias.gr/)

Chapter 3

Examining credit scoring methodologies with alternative data

In this Chapter, considering that the interest in predictive modelling is endless and that there has been much advancement in this area, including the character of independent variables that can be used in credit scoring models (e.g. Pendharkar, 2005; Fletcher and Goss, 1993; Jo *et al.*, 1997; Desai *et al.*, 1996; Tam and Kiang, 1992; Salchenberger *et al.*, 1992; Leshno and Spector, 1996) and benchmarking studies of classification algorithms for credit scoring (e.g. Xiao *et al.*, 2006), we wanted to contribute to both these areas.

The first objective of this Chapter is to introduce alternative data (which we created) to predictive models that use solely credit behavior data in order to explore their contribution to models' performance. More specifically, we are interested in creating variables by using information collected from alternative sources concerning Greek hotels. Hence, we introduce new variables that can be used in conjunction with already existing ones for increasing the predictive performance of the modelling process.

The second objective is to perform a benchmarking study of twelve classification algorithms on a real-world credit scoring data set in order to compare novel to traditional classification methods. In pursuing this objective, we compare these classification algorithms according to their performance indicators. The performance indicators that were used are accuracy and Area Under the ROC Curve (AUC). This analysis provides valuable insights for professionals as they can see novel classification algorithms in predictive modelling. They can also estimate, based on the results, whether it is advantageous to shift from 'traditional' logistic regression to novel classification algorithms in corporate practice. Furthermore, we provide an evaluative survey of recent scoring methods to aid future research. The code used for this analysis is presented in Appendix A. It is also important to mention that for this analysis a real set of credit scoring data was used, which was provided by the private database of Tiresias S.A.

Finally, we perform an out of time and out of sample validation for the 'best' classifier (with the highest performance) in order to see if its performance remains stable over time and for different population because otherwise the model would be useless.

3.1 Data description

In this Section, we present the details of alternative data, the credit scoring data set that was used for the analysis and the pre-processing operations of the data set for the subsequent analysis.

3.1.1 Alternative data in credit scoring

As data constantly changes and evolves, namely data that considered unique, unusual or expensive a few years ago, is now widely used, analysts should develop their thinking and data collection methods in order not to be left behind. Those who are exploited of these data sources, can gain competitive advantage before the others obviate them. This kind of data is often called alternative data and the endless availability increase of data gives the opportunity to gain competitive industry advantage. Put it simply, *alternative data is data that come from non-traditional sources and can be used to supplement traditional data in order to provide better analytical insights that would otherwise not have been achievable.*

3.1.2 Credit scoring data set

A real-world credit scoring data set which includes data from companies was granted for this Thesis, by the private database of Tiresias S.A.. Hotels with credit transactions with banks (800 hotels) were used for the analysis. Specifically, due to the relatively small number of hotels that had any transaction with a bank we did not use a sample, but instead we used

- the *678 hotels to build the model* and
- the remaining *122 hotels to verify the model* (out of time and out of sample validation).

The data set covers the period 1/1/2014 – 31/12/2016 and consists of independent variables related to information from the application form, the status of the credit and the credit behavior of the company. We expand this dataset by including the 'alternative' variables that we created by using

information from social media and customer reviews. This approach was chosen for the purpose of analyzing the alternative variables together with the already existing ones. The alternative variables included in the analysis are the following:

- hotel's registration in Facebook,
- hotel's registration in twitter,
- hotel's registration in Instagram,
- hotel's registration in LinkedIn,
- hotel's registration in YouTube,
- the number of hotel awards,
- hotel's rating in TripAdvisor,
- number of votes in TripAdvisor,
- hotel's rating in Booking and
- number of votes in Booking.

Using the above variables, we created various two-dimensional variables in order to increase the statistical significance (information value) of the variables. The two two-dimensional variables that stood out with the highest information values are:

- the combination of hotel's registration in twitter and Instagram and
- the average rating of TripAdvisor and Booking combined with the sum of votes in TripAdvisor and Booking.

In addition, this data set includes a binary response variable δ that indicates whether a default (of obligation) event was observed in a given period of time:

$$\delta = \begin{cases} 0 & \textit{no default} \\ \\ 1 & \textit{default} \end{cases}$$

Finally, it is important to note that two periods are studied during the model's creation (in this case a logistic regression model),

- ***the observation period*** *from 01/01/2014 to 31/12/2015 and*
- ***the performance period*** *01/01/2016 to 31/12/2016*

(see e.g., Siddiqi, 2006).

3.1.3 Data pre-processing

In this Section we use a standard pre-processing procedure to prepare the data for the upcoming analysis. Particularly, we used weight-of-evidence (WOE) coding to create dummy variables in order to group all the independent variables (Thomas *et al*., 2002). Missing values (records that do not contain all their data-information) were grouped separately. This procedure has the following benefits:

- Eliminates the problem of outliers and rare classes.
- We understand better the relationships with grouping since a chart displaying the relationships between attributes of a variable and performance is more understandable than a variable strength statistic and makes it easier to explain the nature of this relationship, in addition to the strength of the relationship.
- Non-linear dependencies can be modelled with linear models.

Using dummy variables for categorical variables has a serious drawback- it assumes that the difference from one categorical variable group to the next is the same. A better way to deal with grouped variables is to use the WOE of each grouping as the input. This not only solves the problems of differing input units, but also considers the exact trend and scale of the relationship from one group to the next. It also helps in the development of scorecards by keeping each characteristic intact. In addition, if the grouping is done right, this will also ensure that the allocation of points to each group during scorecard scaling is logical and represents the difference in the relationship between groups (Thomas *et al*., 2002).

Tables 3.1 and 3.2 present the way the two 2-dimensional alternative variables are grouped.

Bad Rate (twitter-Instagram)	Instagram No	Instagram Yes	Total
Twitter No	23,5%	15,0%	22,7%
Twitter Yes	15,7%	10,1%	13,9%
Total	21,1%	11,8%	19,5%

Table 3.1: Bad Rates of the variable registration in twitter and Instagram

In Table 3.1 we can see the bad rates of the variable registration in twitter and Instagram, where the bad rate is calculated as following:

$$\text{Bad rate } = \frac{Bad}{Bad + Good}$$

and it shows the percentage of bad that every group has. As it is noticed, the bad rate is not a column but a whole table as the variable is two-dimensional. We calculate the bad rate in order to group together cells with similar bad rate. For example, we group together the cells with bad rates 15,7% and 15,0%, which led to the creation of Table 3.2.

Registration in twitter and Instagram	Bad	Good	Bad Rate	WOE	IV
Neither registered in twitter nor Instagram	91	296	23,50%	-24,03	0,04
Registered either in twitter or Instagram	33	179	15,60%	27,11	0,02
Registered in both twitter and Instagram	8	71	10,10%	76,34	0,05
Total	132	546	19,50%	-	0,11

Table 3.2: Grouping of Registration in twitter and Instagram

In the first column of Table 3.2 we see how the variable of interest was grouped. Subsequently, we calculate the WOE needed for the coding (for the input values):

$$WOE = \ln\left(\frac{Distr.Good_i}{Distr.Bad_i}\right) * 100,$$

where

$$\text{Distr. Good} = \frac{Good}{Total\ Good} \text{ and Distr. Bad} = \frac{Bad}{Total\ Bad}.$$

The WOE measures the strength of each attribute, or grouped attributes, in separating good and bad accounts. It is a measure of the difference between the proportion of good and bad in each attribute (i.e., the odds of a person with that attribute being good or bad). Multiplication by 100 is done to make the numbers easier to work with. Negative numbers of WOE imply that the particular attribute is isolating a higher proportion of bad than good.

Finally, we calculate the Information Value (IV) which shows the statistical significance for each variable/attribute. More specifically, *Information*

Value or *Power Statistic* measures the distance between two distributions. Thus, it is a stage in which special attention should be given as it determines which variables will be kept for building the model and which ones will be abandoned:

$$IV = \sum_{i=1}^{n} (Distr.\ Good_i - Distr.\ Bad_i) * \ln(\frac{Distr.\ Good_i}{Distr.\ Bad_i})$$
$$= \sum_{i=1}^{n} (Distr.\ Good_i - Distr.\ Bad_i) * \text{WOE}$$

Loosely speaking, the *ln* term measures the deviation between the distributions involved while their difference describes the importance of this deviation. Observe that IV is the *J-divergence* (Jeffreys, 1946) which is the symmetric version of the well-known *Kullback-Leibler divergence measure* (Kullback-Leibler, 1951) given by

$$KL = \sum_{i=1}^{n} p_i * \ln(\frac{p_i}{q_i})$$

where $p_i = Distr.\ Good_i$ & $q_i = Distr.\ Bad_i$.

If

- Total IV < 0,02: the predictor is not useful for the model.
- Total IV = 0,02 - 0,1: weak relationship to good/bad odds ratio.
- Total IV = 0,1 – 0,3: medium strength relationship to good/bad odds ratio.
- Total IV = 0,3 – 0,5: strong relationship.
- Total IV > 0,5: suspicious relationship.

(source: https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html)

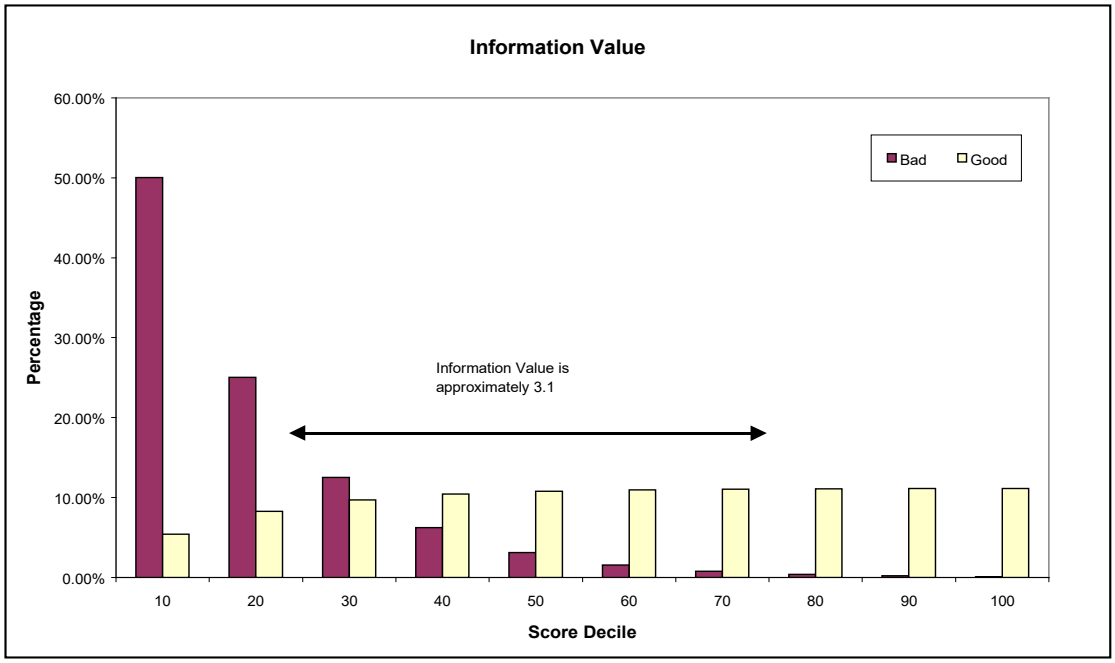


Figure 3.1: Information Value diagramme

In Figure 3.1 we see how the Information Value is shown diagrammatically.

By the same way, in Table 3.3, we have the grouping for the second two-dimensional variable which is the average rating of TripAdvisor and Booking combined with the sum of votes in TripAdvisor and Booking.

Average rating (a) and sum of votes (s)	Bad	Good	Bad Rate	WOE	IV
Registered either exclusively in TripAdvisor or Booking	23	80	22,3%	-17,33	0,00
Not registered in TripAdvisor nor Booking	10	0	100,0%	-893,27	-
[(a<7 & (s<77 or s>77)] or [(7,01<a<7,8 & s<479]	33	70	32,0%	-66,78	0,08
[(7,01<a<7,8)&s>480] or [(7,81<a<8,17)&(s<77ors>77)] or [(8,16<a<8,5)&s<888]	32	144	18,2%	8,43	0,00
[(8,51<a<8,9)&s<888] or [a>8,91&s<291]	27	132	17,0%	16,71	0,01
[(8,16<a<8,9)&s>889] or [a>8,91&s>292]	7	120	5,5%	142,18	0,24
Total	132	546	19,5%	-	0,33

Table 3.3: Grouping of average rating in TripAdvisor and Booking combined with the sum of votes in TripAdvisor and Booking

3.2 Models' comparison

In this Section, we report the explanatory credit behavior variables that are used in predictive models for Greek hotels by Tiresias S.A. We also mention the *K-S (Kolmogorov-Smirnov), Gini Index and accuracy values* of these models in order to compare them with the corresponding values of the 'alternative' model which we are going to build next. K-S and Gini Index are used to determine the degree of ability of a binary model to separate categories.

These two measures are defined as follows (Thomas *et al*., 2002):

$$K - S = Cumulative\ Good - Cumulative\ Bad$$

$$Total\ K - S = \max(Cumulative\ Good - Cumulative\ Bad),$$

where

$$Cumulative\ Bad = Cum\%\ Bad_i = Cum\ Bad_{i-1} + \frac{Bad_i}{Total\ Bad}$$

$$Cumulative\ Good = Cum\%\ Good_i = Cum\ Good_{i-1} + \frac{Good_i}{Total\ Good}$$

and

$$Gini\ Index_i = (Cum\ Bad_i - Cum\ Bad_{i-1}) * (Cum\ Good_i + Cum\ Good_{i-1})$$

$$Total\ Gini\ Index = 1 - \sum_{i=1}^{n} Gini\ Index_i$$

The model for Greek hotels includes the following explanatory variables:

1) x_1: Sum occurrence delinquency one plus (delinquencies) at last 24 months,
2) x_2: Utilization PJ (prime joint holders) update at last 12 months non-revolving, (where $utilization = \frac{current\ balance}{credit\ limit} * 100$ and an example of non-revolving is business loans),
3) x_3: Utilization PJ update at last 12 months revolving (by the term revolving we mean that someone has borrowed an amount and then he can borrow again),
4) x_4: Worst payment status PJ last month vs 24 months.

Tables 3.4 and 3.5 provide the K-S value for this model which is 74,8%, the Gini Index found to be equal to 0,88 and model's accuracy which is 91,4%.

Observed-Predicted	Bad	Good	Percentage Correct
Bad	95	37	72,0
Good	21	525	96,2
Overall Percentage	-	-	91,4

Table 3.4: Classification Table of the model without alternative variables

Score-Range	Bad	Good	Cum% Bad	Cum% Good	K-S	GINI Index
≤ ,20603	63	9	47,7%	1,6%	46,1%	
,20604-,61858	39	19	77,3%	5,1%	72,1%	0,02
,61859-,84963	17	56	90,2%	15,4%	74,8%	0,03
,84964-,93708	7	64	95,5%	27,1%	68,3%	0,02
,93709-,99281	6	258	100,0%	74,2%	25,8%	0,05
,99282+	0	141	100,0%	100,0%	0,0%	0,00
Total	132	546			74,8%	0,88

Table 3.5 K-S and Gini Index of the model without alternative variables

Subsequently, we introduce the two 2-dimensional alternative variables of Section 3.1.2, into this model, because they were statistically more significant than the one-dimensional alternative variables according to weight-of-evidence (WOE) and Information Value (IV). This resulted in the following 'alternative' model:

$$Ln(odds) = 1{,}55820 + 0{,}00610 * x_1 + 0{,}00587 * x_2 + 0{,}00750 * x_3 + 0{,}00494 * x_4 + 0{,}01191 * x_5 + 0{,}00932 * x_6$$

where in addition to the four variables defined previously, we have the variables

x_5= Hotel's registration in twitter and Instagram,

x_6= Hotel's average rating in TripAdvisor and Booking combined with the sum of votes in TripAdvisor and Booking.

Ln(odds) shows the possibility of a hotel to be good. It takes values between 0 and 1, and the closer to 0 the better is the hotel.

Table 3.6 is the Classification Table for the alternative model, and it shows that the inclusion of explanatory variables increases the proportion of cases (from the 50-50 case) of the dependent variable that are correctly predicted by the model. In this case, the model correctly predicts 92,9% (accuracy) of the observations. This percentage is slightly higher than the previous model's accuracy (which is 91,4%, see Table 3.4) which does not contain the alternative variables.

Observed-Predicted	Bad	Good	Percentage Correct
Bad	104	28	78,8
Good	20	526	96,3
Overall Percentage	-	-	92,9

Table 3.6: Classification Table of the alternative model

Table 3.7 contains K-S and Gini Index which are 77,0% and 0,90 respectively and they are used in order to verify whether the model is capable of distinguishing two populations (good-bad). We observe that both K-S and Gini Index are higher than they were in the previous model (which are 74,8% and 0,88 respectively, see Table 3.5) which contained only credit behavior variables.

Score-Range	Bad	Good	Cum% Bad	Cum% Good	K-S	GINI Index
≤ ,19102	64	3	48,5%	0,5%	47,9%	
,19103-,57739	42	26	80,3%	5,3%	75,0%	0,02
,57740-,85966	16	55	92,4%	15,4%	77,0%	0,03
,85967-,98958	9	193	99,2%	50,7%	48,5%	0,05
,98959 +	1	269	100,0%	100,0%	0,0%	0,01
Total	132	546			77,0%	0,90

Table 3.7 K-S and Gini Index of the alternative model

At this point, it is important to remember that we are working on a real-world and homogeneous dataset and for this reason the increase in accuracy, K-S and Gini Index that may seem small, is considered to be significant. Finally, based on the above results, we conclude that the alternative data contribute to the performance of the model for Greek hotels and it is our belief that it would be wise to investigate their usefulness in other industries as well.

3.3 Benchmarking experiment-experimental setup

In this part of the study, we compare twelve classification algorithms most of which offer some meta-parameters to emphasize specific tasks. Examples of such parameters are the number of hidden nodes in neural networks and the kernel function in Support Vector Machines (SVM). As our goal is to compare several algorithms, we create many classifiers for each algorithm by changing the parameters each time and keep the one with the best performance based on performance indicators. In fact, we compare the best classifiers of all the algorithms.

3.3.1 A short review of novel classification algorithms

In this Section we provide an overview of the novel classification algorithms considered here, in order to illustrate the philosophies underneath different classification algorithms.

1. Cubist is a powerful tool for generating rule-based models that balance the need for accurate prediction against the requirements of intelligibility. On the one hand, Cubist models give better results than those produced by simple techniques such as multivariate linear regression and on the other hand, are easier to understand than neural networks. Some important features of this algorithm are the following. Cubist has been designed to analyze substantial databases containing hundreds of thousands to millions of records and tens to thousands of numeric or nominal fields. Someone who has used neural networks or similar modelling tools, would be surprised by Cubist's speed. Also, to maximize interpretability, Cubist models are expressed as collections of rules, where each rule has an associated multivariate linear model. Whenever a situation matches a rule's condition, the associated model is used to calculate the predictive model (Quinlan, 1992)

2. Boosting is a general ensemble method that creates a strong classifier from several weak classifiers. This is achieved by building a model from the training data, then creating a second model that strives to correct the errors from the first model. Subsequently, models are added until the training set is predicted perfectly, or a maximum number of models are added. Adaptive Boosting (AdaBoost) was the first successful boosting algorithm developed for binary classification. AdaBoost is suitable for boosting the performance of any machine learning algorithm and it is best used with weak learners. These are models that achieve accuracy just above random probability on a classification problem, namely decision trees with one level, because these trees are so short and only contain one decision for classification. AdaBoost works by weighting the observations, putting more weight on difficult to classify instances and less on those already handled well. New weak learners are added sequentially that focus their training on the more difficult patterns. This process continues until a pre-set number of weak learners have been created or no further improvement can be made on the training data set. Once completed, we are left with a pool of weak learners each with a stage value. Finally, predictions are made by calculating the weighted average of the weak classifiers. Another point of view of boosting machine learning algorithms that we used in our study is Extreme Gradient Boosting (XGBoost), where the term 'Gradient Boosting' is proposed in the paper Greedy Function Approximation: A Gradient Boosting Machine, by (Friedman, 1999). XGBoost is based on this original model and is used for supervised learning problems. It is developed with both deep consideration in terms of systems optimization and principles in machine learning. The goal of XGBoost is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate algorithm to overcome classic Gradient Boosting.

3. Extreme learning machines (ELMs) are a recently introduced variant of neural networks. ELMs are based on a mathematical proof that a single-hidden layer feed-forward network with randomly generated hidden-layer-weights is a universal approximator if the weights connecting the hidden and the output layer, b, are appropriately chosen (Guang-Bin *et al*., 2006). This result accommodates building ELM classifiers without using resource-intensive training algorithm such as gradient descend. Instead, it can solve $y = H * b$, where $y = (y_1, \dots, y_n)$ are the training data class labels and H is the hidden layer output matrix (e.g., Huang *et al*., 2006).

4. The classification and regression trees (CART) classifier operate in a similarly with decision trees but uses the Gini-coefficient to guide tree

growing (e.g., Hastie *et al*., 2009). Decision trees tend to build a complex structure of many internal nodes and this often leads to overfitting. Therefore, the CART offers meta-parameters that allow you to influence when to stop growing trees or how to prune a fully developed tree. The success of an ensemble strategy depends on the accuracy of individual base models and the diversity among them (e.g., Kuncheva, 2004). Homogenous ensembles build several base models using the same classification algorithm and in order to manage diversity, they rely on sampling mechanisms. Specifically, given a training set of size n and some classification algorithm, bagging (Breiman, 1996) makes T bootstrap samples of size n from the training set and applies the classification algorithm to every sample. By this way, T base models are produced and their predictions are pooled using majority voting. Bagging works best with underlying classification algorithm sensitive to data perturbations (e.g., Marques *et al*., 2012). Therefore, we use bagging in conjunction with CART base classifiers.

3.3.2 Credit scoring data set

The credit scoring data set that was used for the benchmarking experiment, the independent variables that took part in the analysis and the data pre-processing have been described in Sections 3.1.2 and 3.1.3. The training period is the period of 24 months (01/01/2014 to 31/12/2015) and the testing period is the period of 12 months (01/01/2016 to 31/12/2016). Subsequently, we elaborate how we assess the predictive performance of competing classification algorithms.

3.3.3 Performance indicators

There are many indicators who measure predictive accuracy (Hand, 1997). In this study we consider the percentage correctly classified (PCC, also called classification accuracy) and the Area Under the ROC Curve (AUC).

PCC and other common indicators ground on a confusion matrix of actual versus predicted class labels. An example of these indicators is given in Table 3.8. Also, an example of PCC is Table 3.6 which presents the results that were found in section 3.2.

Actual-Predicted	Bad	Good
Bad	True Negative (TN)	False Positive (FP)
Good	False Negative (FN)	True Positive (TP)
Indicators	PCC=(TP+TN)/(TP+TN+FP+FN) Classif. error=(FP+FN)/(TP+TN+FP+FN)	TPR=TP/(TP+FN) Precision=TP/(TP+FP)

Table 3.8: Confusion matrix of actual and predicted class labels

AUC is an aggregated measure of classifier performance, namely it averages classifier performance in all possible thresholds (e.g., Flach *et al*., 2011). AUC takes values from 0.5 to 1, where 0.5 corresponds to the random classification while 1 corresponds to the perfect classification. In other words, the AUC equals the probability that a randomly chosen positive example will be ranked higher than a randomly chosen negative example. Also, AUC can be calculated in relation to Gini Index as follows:

$$AUC = \frac{Gini\ Index}{2}$$

3.4 Empirical results

3.4.1 Benchmarking results

Our experiment results consist of performance estimates of twelve classifiers (novel and traditional) in terms of accuracy (PCC) and Area Under the ROC Curve (AUC). Most classifiers offer some meta-parameters to emphasize to a particular task, like the number of hidden nodes in neural networks and the kernel function in Support Vector Machines (SVM). As our objective is to compare several classification algorithms to each other, we produce multiple models with a single classification algorithm by define several settings for such meta-parameters and we keep those with the highest values of performance indicators. In this way, we will end up by having a best performing classifier for each classification algorithm. Table 3.9 reports the benchmarking in terms of these performance indicators. The first seven classifiers were used as traditional classifiers and the other five as novel.

Classification algorithms (classifiers)	Accuracy (PCC)	AUC
Logistic Regression	92,89%	0,957
Decision tree	91,31%	0,931
Random forest	92,58%	0,954
SVM	92,26%	0,955
K-nearest neighbor	92,73%	0,954
Neural networks	93,84%	0,952
Naïve Bayes	92,42%	0,955
Extreme Gradient Boosting (XGBoost)	91,63%	0,898
Adaptive Boosting (Adaboost)	91,79%	0,878
Bagging CART	92,42%	0,949
Cubist	91,94%	0,932
Extreme learning machine (ELM)	85,62%	0,802

Table 3.9: Performance of classification algorithms in terms of accuracy and AUC

Table 3.9 indicates that the differences between the values of these performance indicators in each classifier are slight and possibly this happens because we worked on a homogenous sample. We say that the data set is homogeneous as it consists only of hotels and not of many different industries. We came to this conclusion after analyzes performed on homogeneous and not so homogeneous data sets and it was observed that the more homogeneous the sample, the smaller the differences in the values of the different methods. This does not mean, however, that the method that was best with a small difference in a homogeneous sample was no longer the best in a not so homogeneous sample. Specifically, we notice that logistic regression and neural networks perform better than other classifiers. It also seems that the Bagging CART (from the novel classifiers) has particularly good results in contrast to the Extreme learning machine (ELM) which has the lowest performance. In the following figure, Figure 3.2, we can graphically see the comparison of classification algorithms based on accuracy.

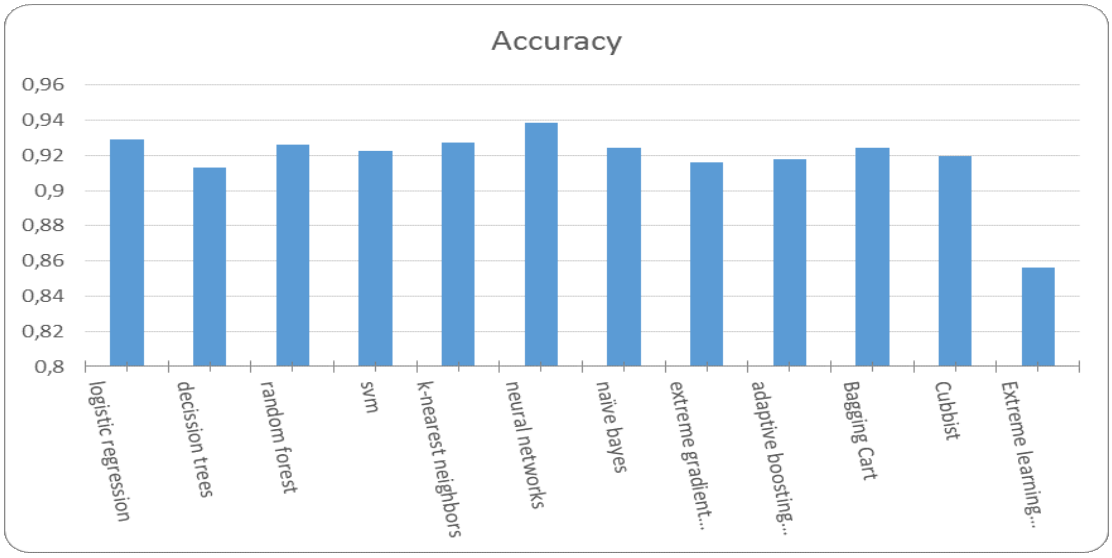


Figure 3.2: Accuracy comparison

As logistic regression has the highest AUC value, we take this model in order to test its stability in the following Section. In the following Figure 3.3 we can see the AUC of the logistic regression model.

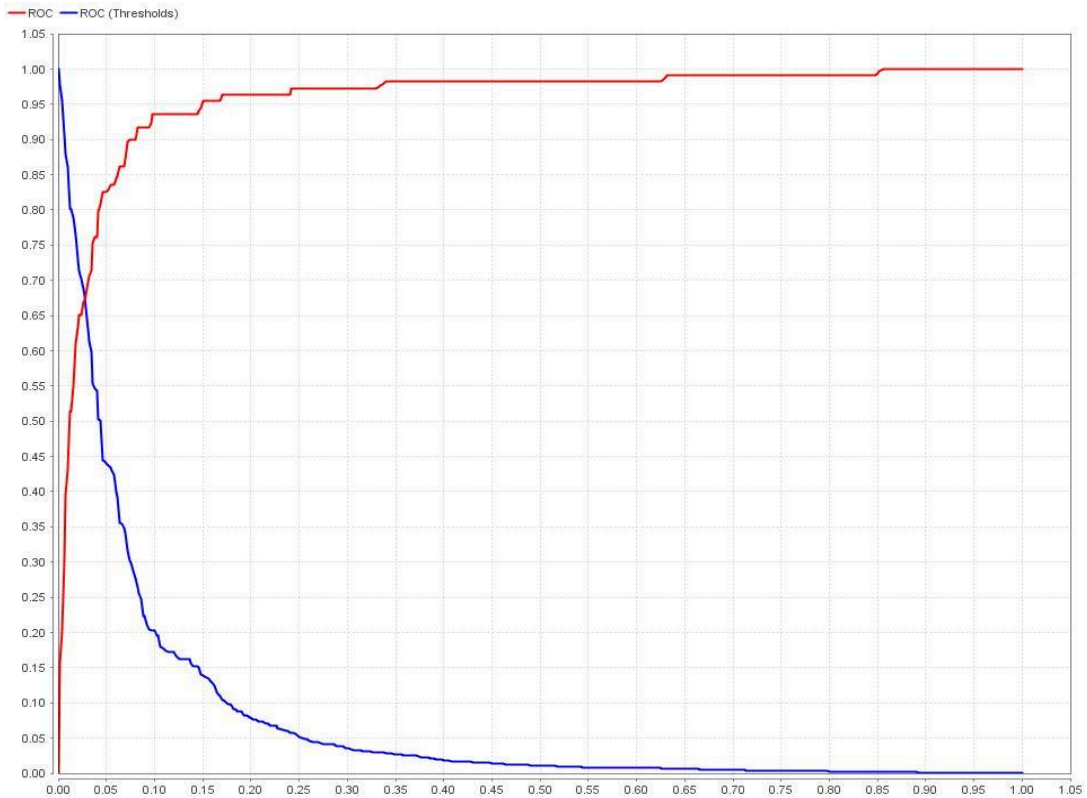


Figure 3.3: AUC of logistic regression

3.4.2 Out of time and out of sample validation

The following procedure verifies the logistic regression model by running it in another time (04/2016 to 04/2017) in order to see if it is still efficient and stable, as it will be useful only if it can be used over time. Observing the results in Table 3.10, it appears that K-S (79,0%) is better than before (K-S=77,0% section 3.2, Table 3.7) and Gini Index remains the same (0,90).

Predicted probability (Score-range)	Bad	Good	Bad Rate	K-S	Gini Index
≤ ,19102	54	6	90,0%	48,4%	-
,19103-,57739	37	24	60,7%	77,8%	0,02
,57740-,85966	11	46	19,3%	79,0%	0,02
,85967-,98958	5	193	2,5%	46,7%	0,03
,98959 +	2	254	0,8%	0,0%	0,03
Total	109	523	17,2%	79,0%	0,90

Table 3.10: Out of time validation (K-S and Gini Index)

Also, model's stability is confirmed once again in Table 3.11 as its stability value is 0,00.

Predicted probability (Score-range)	Development #	Validation #	Development %	Validation %	Stability Index
≤ ,19102	67	60	9,9%	9,5%	0,00
,19103-,57739	68	61	10,0%	9,7%	0,00
,57740-,85966	71	57	10,5%	9,0%	0,00
,85967-,98958	202	198	29,8%	31,3%	0,00
,98959 +	270	256	39,8%	40,5%	0,00
Total	678	632	100,0%	100,0%	0,00

Table 3.11: Stability

Where Development # is the sum of Bad and Good in every attribute at the period we created the model (see Table 3.7) and Validation # is the sum of Bad and Good in every attribute at the period of validation (see Table 3.10). Also,

$$Development\ \% = \frac{Development\ \#}{Total\ Development\ \#},$$

$$Validation\ \% = \frac{Validation\ \#}{Total\ Validation\ \#}$$

and

$$Stability\ Index = (Validation\ \% - Development\ \%) * ln(\frac{Validation\ \%}{Development\ \%})$$

Finally, we perform an out of sample validation utilizing the 122 hotels that were not used during the construction process of the model. In Table 3.12 we observe that KS (77.6%) and Gini Index (0.90) remain high compared to those of the sample with 678 hotels (K-S = 77.0%, Gini Index = 0.90, see Section 3.7, Table 3.7), proving that the model is also suitable for different samples (population).

Predicted probability (Score-range)	Bad	Good	Bad Rate	K-S	Gini Index
≤ ,19102	22	1	95,7%	61,7%	-
,19103-,57739	5	5	50,0%	70,2%	0,01
,57740-,85966	5	6	45,5%	77,6%	0,03
,85967-,98958	3	32	8,6%	49,4%	0,06
,98959 +	0	43	0,0%	0,0%	0,00
Total	35	87	28,7%	77,6%	0,90

Table 3.12: Out of sample validation (K-S and Gini Index)

3.5 Conclusion

In this Chapter we set out to explore the effectiveness of alternative data in credit scoring models. To that end, we created and introduced variables from alternative sources, to an already existing predictive model for Greek hotels which uses only credit behavior data. For this purpose, we used a real-world credit scoring data set of 678 Greek hotels. Comparing the 'alternative' model with the already existing one in terms of K-S, Gini Index and accuracy, we concluded that alternative data contribute to

model's performance. Indeed, the improvement can be easily seen by observing the differences between the values of performance indicators for these two models: K-S: 77,0% >74,8%, accuracy: 92,9 >91,4, Gini Index: 0,90 >0,88. After noticing this contribution in model's performance for Greek hotels, we can say that it would be prudent to explore alternative data's utility in other industries as well.

Subsequently, we conducted a benchmarking study of 12 classification algorithms on the same real-world data set in order to compare novel with traditional classification methods. In pursuing this objective, we compared these classification algorithms in terms of AUC and accuracy. Our results showed that there are slight differences between the values of these performance indicators in each classifier and possibly this happens because we worked on a homogeneous sample. Specifically, we noticed that logistic regression and neural networks performed better than other (novel or not) classifiers and logistic regression had the highest value of AUC. From a managerial perspective, the key question is whether neural networks or other 'complex' classification algorithms can and should take the place of the industry standard logistic regression in corporate practice, given the slight differences between the values of their performance indicators. Based on the above analysis, logistic regression seems to perform significantly satisfactorily and there is no question of replacing it, at least in terms of (relatively) homogeneous data. It is noted, however, that further investigation is deemed necessary for non-homogeneous data. Finally, we demonstrated the efficiency and stability of the logistic regression model by applying it at different period and in different samples.

Continuing to aim for the maximum performance of predictive models and wanting to contribute to a wider industry (not only hotels), we decided to explore a combination of data regarding the independent variables that will make up predictive models for enterprises. This analysis is presented in detail in the following Chapter.

Chapter 4

Combination of financial and credit behavior data for companies

As financial data are usually the only data that are used in Greek models (and other countries) in order to evaluate the credit risk of a company and our objective is to maximize their performance, we now proceed to use a combination of financial and credit behavior data. In this Chapter, the main contribution of our analysis is the introduction of new credit risk models which evaluate credit risk of small and large Greek enterprises (according to their revenues) by using a combination of financial and credit behavior data. Subsequently, the models with the combination of data are compared with models containing only financial data in terms of accuracy, K-S and Gini Index (Thomas *et al*., 2002). Finally, the stability of these new models is tested on samples after the period of the time-period of data-collection.

4.1 A review of financial data

In this Section we will describe the financial statements of a company, the meaning of the ratios, as well as their calculation. Although these concepts are widely known, one can find them on various financial websites like https://corporatefinanceinstitute.com/resources/knowledge/accounting/.

4.1.1 Financial statements of the company

Financial Statements of the Company include the following:

1. Balance sheet
2. Statement of income statements
3. Profit distribution status

1. Balance sheet: A company's balance sheet summarizes its financial situation at a given point in time. In particular, the assets of the company are grouped (on a time-share basis) into fixed assets, current assets, cash and its own funds (net worth capital, long-term and short-term liabilities). This is an equation described as follows:

$$\text{Assets} = \text{Liabilities (net worth capital + liabilities)}$$

The elements of the above equation are analyzed below:

Assets:

- Fixed assets: are the assets of the company which are not intended for sale but are used to serve the operation of the company, that is, they are held for use in the production of goods or services or for administrative purposes and are expected to be used for more than one accounting period (more than one year).

- Current assets: includes assets of a company that can be liquidated or disposed within a short-term period (less than one year). This category mainly includes the following:
  - Stocks of goods and raw materials,
  - Advances for purchases of goods,
  - Customer claims,
  - Checks and promissory notes receivable in the company's portfolio,
  - Various other debtors against which the company raises,
  - Debt securities (e.g., shares, third-party bonds),
  - Cash deposits and sight deposits which are distinguished for high and immediate liquidation.

- Cash: are assets in the form of cash or high liquidity positions (e.g., demand deposits, repos, deposits of one or few days) which are used in the context of current business relations of the company.

Liabilities:

- Net worth capital: include items categorized as follows:
  - Share capital (paid and payable),
  - Differences in the sale of shares at a price greater than their nominal value,
  - Revaluation differences (goodwill) from revaluation of assets,
  - Reserve funds formed from the profits of the period and are provided either by the Statute of the Amendment or by Development Laws or by Emergency Needs,
  - Retained earnings that offset the profit or loss for the period under review with any accumulated losses or gains from previous periods.

- Forecasts: are capital reserved for estimates of known liabilities (bad debts, employee indemnities)

- Long-term liabilities: mainly include loans that the company has entered with specific lenders and their repayment is expected to be over one year.

- Short-term liabilities: include the liabilities whose repayment is expected to occur within a period of less than one year following the balance sheet period. Such liabilities are as follows:

  - Suppliers (purchase of goods and services),
  - Accounts payable,
  - Bank loans,
  - Tax liabilities,
  - Insurance agencies,
  - Dividends payable to shareholders by profits,
  - Other debts to debtors.

2. Statement of income statements: Statement of income statements presents the total income and expenses incurred by the company during the year. Therefore, its figures are not static or cyclical but express, in terms of value, its activity over the entire use.

3. Profit distribution status: Profit distribution status shows how the net operating surplus is distributed between the company and its shareholders. Its figures reveal the dividend policy and the degree of self-financing of the company.

4.1.2 Financial ratios

Financial ratios express logical relationships between the balance sheet elements and the profit and loss account and are calculated as the mathematical effect of dividing one item by the other.

Categories of financial ratios

- Liquidity ratios: determine the short-term financial position of an entity as well as its ability to meet its short-term liabilities.
- Activity ratios: measure the extent to which an entity is effective in managing its assets.
- Profitability ratios: calculate the profitability of an entity profits.
- Financial structure & viability ratios: assess the ability of the entity to meet its long-term liabilities and the degree of protection received by its creditors.
- Investment ratios: associate the number of shares and its stock price with its profits, dividends and other assets.
- Operating expenses ratios: provide an indication of the management policy in managing operating expenses.

Rules for the development of financial ratios

Financial ratios shall be drawn up in accordance with the following rules:

1. The correlation of the sizes is made in such a way that the resulting indicators-numbers are directly proportional to the situation they represent, that is, the higher indicators correspond to more favorable situations and the lower to worse.
2. The correlation sizes are selected in a way that reduces to the minimum, for example, the errors or the effects of monetary fluctuations.
3. Indicators whose terms refer to a period of use of less than twelve months are not indicative of the overall situation of the entity and are always considered to be correlated with indicators of corresponding periods of previous years.
4. The indicators cover all the sections of activity of the entity. For this reason, they are grouped in a way that allows a much broader analysis for each activity.
5. An individual indicator has only relative utility. For this reason, it is necessary to compare several indicators to each other to ensure correct conclusions.

4.1.3 Calculation of some financial ratios

In this Section we show how some of the basic financial ratios, which also were used in our analysis, are calculated.

- Current ratio = Current assets / Current liabilities
- Quick ratio = Cash + Accounts Receivable / Current liabilities
- Cash ratio = Cash / Current liabilities
- Current liabilities to net worth ratio = Current liabilities / Net worth capital
- Total liabilities to net worth ratio = Total liabilities / Net worth capital
- Total liabilities to total assets = Total liabilities / Total assets
- Fixed assets to net worth = Fixed assets / Net worth capital
- Current assets to total liabilities = Current assets / Total liabilities
- Capital structure = Total assets / Net worth capital
- Working capital leverage = Short-term loans / (Stocks + accounts receivable – (current liabilities – short-term loans))
- Cash turnover ratio = Sales / Cash
- Collection period ratio (days) = (Accounts receivable / Sales) *365
- Inventory turnover ratio = Cost of sales / Inventory
- Total assets turnover ratio = Sales / Total assets
- Net working capital turnover ratio = Sales / Net working capital
- Accounts payable turnover ratio = (Cost of sales – depreciation embedded in cost of sales) / accounts payable
- Current liabilities turnover ratio = Sales / Current liabilities
- Current debt to sales = Short-term loans / Sales
- Gross profit margin = Gross profit / Sales
- Pretax profit margin = Pretax profits / Sales
- Net profit margin = After tax profits / Sales
- Return on assets (ROA) = After tax profits / Total assets
- Pretax return on assets = Pretax profits / Net worth capital
- Return on equity (ROE) = After tax profits / Net worth capital
- Earnings before interest and taxes (EBIT) to sales = Profits before interest and taxes / Sales
- EBITDA margin (EBTDA=Earnings before interest, taxes, depreciation and amortization) = Profits before interest, taxes, depreciation and amortization / Sales
- Profits before depreciation and after tax to sales = Profits before depreciation and after tax / Sales

4.2 Data description

For this analysis, the data were taken again by the private database of Tiresias S.A. A random sample of 7.315 companies was granted in order to be representative, credible and with no bias. The 3.256 of these enterprises are small (revenues ≤ 700.000) and the remaining 4.059 are large (revenues > 700.000). For the successful selection of the sample, enterprises without sufficient history (history of less than six months) were removed. Moreover, companies who did not wish to display their data in the system, as well as companies with negative behavior in the treatment and creation month of the model were erased. *The dependent variable is the investigation of the fact that the company's behavior is 'good' or 'bad'.* The definition of 'good' and 'bad' business behavior is given in section 2.5. The independent variables that were used in the analysis are:

- **The financial data for the years 2012-2013 that can be found in company's balance sheet and 35 financial ratios** that arise from them as well as their changes [e.g., (current ratio 2013 - current ratio 2012)/current ratio 2012]. Some of the financial ratios that were used are the following: current ratio, cash ratio, current liabilities to equity, debt equity ratio, total liabilities to total assets, fixed assets to equity, capital structure, current liabilities turnover ratio, net profit margin, ROA return on assets, ROE return on equity, equity net worth to total liabilities, net sales turnover to total assets.

- **Credit behavior data:** (e.g., delinquency index, credit limit, current balance, current balance delinquent, loan card type, approval date, number of 'instalments' (for loans), frequency of 'instalments' (for loans), amount of instalments, deletion flag and deletion flag date), from **which the following variables are given**:
- **Variable 1**: consecutive months with maximum utilization greater than 100 for the last 24 months.
- **Variable 2**: maximum current utilization.
- **Variable 3**: maximum delinquency updated last month (1 month) versus maximum delinquency of the last 24 months. Specifically, the grouping of this variable is showed in Table 4.1.

Variable	Delinquency last month	Delinquency last 24 months
5	Missing	≥ 0
59	0	0
38	0	1
16	0	≥ 2
-10	1	1
-29	1	≥ 2

Table 4.1: Grouping of variable 3.

- **Variable 4**: maximum number of consecutive months with credit utilization over 100% in last 6 months.
- **Variable 5**: number of occurrences with delinquency ≥ 1.
- **Variable 6**: $\frac{\text{current balance with delinquency}}{\text{current balance}}$.
- **Variable 7**: worst payment status for the last 3 months.

4.3 Initial characteristics analysis and performance indicators

Initial characteristic analysis involves two main tasks. The first step is to assess the strength of each characteristic/variable individually as a predictor of performance and is done to screen out weak or illogical characteristics.

Our models are produced using continuous characteristics. However, we grouped them by creating dummy variables using weight-of-evidence (WOE) coding as mentioned in Section 3.1.3. Once the strongest (statistically significant) characteristics are grouped and ranked, variable selection is done. At the end of the initial characteristic analysis, there will be a set of strong, grouped characteristics, preferably representing independent information types, for use in the regression step.

The strength of a characteristic is gauged using three main criteria:

- Predictive power of each attribute. WOE measure is used for this purpose.
- The range and trend of WOE across grouped attributes.
- Predictive power of characteristics. Information Value (IV) measure is used for this purpose as mentioned in Section 3.1.3.

An example of the variable *cash ratio* is shown in Table 4.2 where various statistical measures are given.

Cash ratio	Bad	Good	Indet.	Other	Total	Bad rate	WOE	IV
Missing	0	1	0	0	1	0,0%		
≤ 0,04	242	367	209	65	883	39,7%	-74,36	0,15
0,05-0,18	204	635	254	112	1205	24,3%	-2,45	0,00
0,19 +	196	1045	352	377	1970	15,8%	51,36	0,11
Total	642	2048	815	554	4059	23,9%	0,00	0,25

Table 4.2: Statistical measures of cash ratio

In reference to Table 4.2 the following notes should be made:

- Missing values are grouped separately.
- The bad rate and WOE are sufficiently different from one group to the next (namely, the grouping has been done in a way to maximize differentiation between good and bad). This is one of the objectives to identify and separate attributes that differentiate well. While, the absolute value of the WOE is important, the difference between the WOE of groups is the key for establishing differentiation. The largest the difference between subsequent groups, the higher the predictive ability of this characteristic.

The statistical strength is measured in terms of WOE and IV, however it is not the only factor in choosing a characteristic for further analysis or designating it as a strong predictor. The attribute strengths (bad rate and WOE) must also be in a logical order and make operational sense. As it can be clearly seen in Table 4.2, apart from 'missing', the other groups in this characteristic have a linear relationship with WOE; that is, they reveal a linear and logical relationship between the attributes in 'cash ratio values' and proportion of the 'bad'.

Subsequently, model tests are made with variable groups that are chosen (different each time), in order to find which is the best model (regression). Finally, when the choice of variables that will be used in the model in order to be optimum is made, the logistic regression model is constructed, as it is shown in the next Section.

4.4 Interpretation of results and models created by the combination of financial data and RCS, DFO, MPS data

4.4.1 Model for small enterprises

According to previous tests, it was observed that the model with the best results for the small enterprises (with revenues ≤ 700.000) using a combination of financial and credit behavior data is the following:

Ln(odds)= – 0,02769*total liabilities to total assets – 0,00807*ROE return on equity +0,02315*equity net worth to total liabilities +0,00773*net profit margin – 0,00780*net sales turnover to total assets – 0,00554*profit before tax depreciation amortization – 0,00177*Variable 3 – 0,00340*Variable 7 – 0,00404*Variable 4 – 0,00480*Variable 5 – 0,00247*Variable 1 – 0,00287*Variable 6.

Where Variables 3,7,4,5,1,6 are the variables that are described in Section 4.2.

Ln(odds) shows the possibility of a company to be good. It takes values between 0 and 1, and the closer to 0 the better is the company.

Table 4.3 is the Classification table, and it shows that the addition of independent variables increases the proportion of cases of the dependent variable that are correctly predicted by the model. In this case, the model correctly predicts 85,5% (accuracy) of the observations.

Observed-Predicted	Good	Bad	Percentage Correct
Good	1.169	119	90,8
Bad	153	438	74,1
Overall Percentage	-	-	85,5

Table 4.3: Classification table (small companies)

Table 4.4 indicates how well the good enterprises have been set apart from the bad. It is a way of verifying the chosen model as a K-S value of zero would indicate that the model in unable to make any distinction between two populations, while a K-S score of 100 would indicate that the model is capable of perfect distinction between two populations. The 64,6% is the maximum deviation that the bad companies have from the good in this model. Also, the Gini Index in this case is 0.80 and this is also a way of verifying the model as 1 is the highest value it can get.

Score-Range	Good	Bad	Bad rate	K-S	Gini index
≤ ,03251	191	3	1,5%	14,3%	
,03252 - ,05077	172	9	5,0%	26,2%	0,01
,05078 - ,07083	179	11	5,8%	38,2%	0,01
,07084 - ,09894	171	16	8,6%	48,8%	0,03
,09895 - ,15087	167	21	11,2%	58,2%	0,04
,15088 - ,25697	153	34	18,2%	64,3%	0,09
,25698 - ,48979	130	58	30,9%	64,6%	0,17
,48980 - ,76361	89	100	52,9%	54,6%	0,32
,76362 - ,93101	28	160	85,1%	29,7%	0,53
,93102 +	8	179	95,7%	0,0%	0,60
Total	1288	591	31,5%	64,6%	0,80

Table 4.4: K-S and Gini Index (small enterprises)

4.4.2 Model for large enterprises

Similarly, the model with the best results for the large enterprises using the combination of financial and credit behavior data is the following:

Ln(odds)= 0,00005*total liabilities to total assets – 0,00133*cash ratio – 0,00277*current liabilities turnover ratio – 0,00561*current liabilities to equity – 0,00460*long-term liabilities – 0,00422*net profit after tax – 0,00482*total fixed assets undeprec. – 0,00179*interest and related expenses to EBIT + 0,00348*Variable 2 – 0,00316*Variable 3 – 0,00452*Variable 7 – 0,00356*Variable 4 – 0,00410 *Variable 5 – 0,00272*Variable 1.

In this case, the model correctly predicts 85,3% (accuracy) of the observations and K-S and Gini Index are 67,2% and 0,82 respectively. We can see these results in the following Tables (Table 4.5 and Table 4.6).

Observed-Predicted	Good	Bad	Percentage Correct
Good	1778	270	86,8
Bad	125	517	80,5
Overall Percentage	-	-	85,3

Table 4.5: Classification table (large enterprises)

Score-Range	Good	Bad	Bad rate	K-S	Gini index
≤ ,03794	267	3	1,1%	12,6%	
,03795 - ,05633	263	4	1,5%	24,8%	0,00
,05634 - ,11170	520	19	3,5%	47,2%	0,02
,11171 - ,16676	252	17	6,3%	56,9%	0,03
,16677 - ,28776	237	32	11,9%	63,5%	0,07
,28777 - ,48557	223	46	17,1%	67,2%	0,12
,48558 - ,77295	185	84	31,2%	63,1%	0,24
,77296 - ,95615	88	181	67,3%	39,2%	0,55
,95616 +	13	256	95,2%	0,0%	0,79
Total	2048	642	23,9%	67,2%	0,82

Table 4.6: K-S and Gini Index (large enterprises)

4.5 Comparison with models created by financial data only

Making the same procedures as before and using only financial data this time, we conclude at the following results. The model for small enterprises contains the following variables: debt equity, capital structure, current liabilities turnover ratio, ROE return on equity, net profit margin, ROA return on assets, net sales turnover to total assets, total liabilities, interest and related expenses, profit before tax depreciation amortization and income tax. In this case, the model correctly predicts 71,0% (accuracy) of the observations, K-S is only 33,8% and Gini Index is 0,44. We can see these results in the following Tables (Table 4.7 and Table 4.8).

Observed-Predicted	Good	Bad	Percentage Correct
Good	1080	208	83,9
Bad	337	254	43,0
Overall Percentage	-	-	71,0

Table 4.7: Classification table (small enterprises-only financial data)

Score-Range	Good	Bad	Bad rate	K-S	Gini index
≤ ,11828	175	14	7,4%	11,2%	
,11829 - ,15958	160	28	14,9%	18,9%	0,02
,15959 - ,21124	158	31	16,4%	25,9%	0,03
,21125 - ,25842	152	39	20,4%	31,1%	0,06
,25843 - ,31555	139	48	25,7%	33,8%	0,09
,31556 - ,38302	124	65	34,4%	32,5%	0,14
,38303 - ,45383	120	72	37,5%	29,7%	0,18
,45384 - ,54100	102	85	45,5%	23,3%	0,24
,54101 - ,65828	92	96	51,1%	14,4%	0,29
,65829 +	71	118	62,4%	0,0%	0,39
Total	1293	596	31,6%	33,8%	0,44

Table 4.8: K-S and Gini Index (small enterprises-only financial data)

The model for large enterprises contains the following variables: cash ratio, debt equity, total liabilities to total assets, fixed assets to equity, capital structure, current liabilities turnover ratio, ROE return on equity, equity net worth to total liabilities, net sales turnover to total assets, interest and related expenses, net profit after tax, bank short-term payable, total fixed assets under-prec., short-term liabilities, current assets to total liabilities change, long-term liabilities change and interest and related expenses to EBIT. In this case, the model correctly predicts 78,8% (accuracy) of the observations, K-S is only 41,2% and Gini Index is 0,51. We can see these results in the following Tables (Table 4.9 and Table 4.10).

Observed-Predicted	Good	Bad	Percentage Correct
Good	1804	198	90,1
Bad	359	271	43,0
Overall Percentage	-	-	78,8

Table 4.9: Classification table (large enterprises-only financial data)

Score-Range	Good	Bad	Bad rate	K-S	Gini index
≤ ,11875	752	55	6,8%	28,2%	
,11876 - ,15950	238	31	11,5%	34,9%	0,04
,15951 - ,21346	225	44	16,4%	39,1%	0,07
,21347 - ,28395	215	54	20,1%	41,2%	0,11
,28396 - ,36731	197	72	26,8%	39,6%	0,17
,36732 - ,47056	184	85	31,6%	35,3%	0,22
,47057 - ,60989	145	125	46,3%	22,9%	0,36
,60990 +	92	176	65,7%	0,0%	0,54
Total	2048	642	23,9%	41,2%	0,51

Table 4.10: K-S and Gini Index (large enterprises -only financial data)

4.6 Out of time validation and stability

The following procedure verifies the original models (containing the combination of data) by running them in another time (2015) in order to see if they are still efficient and stable over time. Observing the results in the Tables below, it seems that the K-S is better than before for small enterprises, while in the case of large there is an exceedingly small drop. Specifically, K-S for small enterprises in 2015 is 69.0% while in the period when the model was built it was 64.6% (see Section 4.4.1) and for large enterprises, K-S is 66.2%, while before it was 67, 2% (see Section 4.4.2). We can see these results in Tables 4.11 and 4.12, respectively. Finally, the stability of the models for small and large enterprises is verified in Tables 4.13 and 4.14, respectively.

Score-Range	Good	Bad	Indet.	Other	K-S
≤ ,03251	37	0	0	0	3,8%
,03252 - ,05077	97	2	6	6	13,0%
,05078 - ,07083	175	4	11	7	29,5%
,07084 - ,09894	164	4	19	12	44,9%
,09895 - ,15087	172	13	21	20	58,1%
,15088 - ,25697	161	17	19	49	69,0%
,25698 - ,48979	109	49	78	71	64,1%
,48980 - ,76361	48	67	69	22	47,2%
,76362 - ,93101	16	86	43	6	20,8%
,93102 +	4	65	17	1	0,0%
Total	983	307	283	194	69,0%

Table 4.11: Out of time validation (small enterprises)

Score-Range	Good	Bad	Indet.	Other	K-S
≤ ,03794	75	1	3	5	5,1%
,03795 - ,05633	180	2	9	14	17,5%
,05634 - ,11170	429	16	43	17	44,1%
,11171 - ,16676	185	17	49	18	53,0%
,16677 - ,28776	220	28	61	32	61,7%
,28777 - ,48557	144	23	93	24	66,2%
,48558 - ,77295	105	86	159	4	52,0%
,77296 - ,95615	55	114	144	1	27,2%
,95616 +	7	110	41	0	0,0%
Total	1400	397	602	115	66,2%

Table 4.12: Out of time validation (large enterprises)

Score-Range	Develop. #	Validation #	Develop. %	Validation #	Stability
≤ ,03251	194	37	10,3%	2,9%	0,10
,03252 - ,05077	181	99	9,6%	7,7%	0,00
,05078 - ,07083	190	179	10,1%	13,9%	0,01
,07084 - ,09894	187	168	10,0%	13,0%	0,01
,09895 - ,15087	188	185	10,0%	14,3%	0,02
,15088 - ,25697	187	178	10,0%	13,8%	0,01
,25698 - ,48979	188	158	10,0%	12,2%	0,00
,48980 - ,76361	189	115	10,1%	8,9%	0,00
,76362 - ,93101	188	102	10,0%	7,9%	0,00
,93102 +	187	69	10,0%	5,3%	0,03
Total	1879	1290	100,0%	100,0%	0,19

Table 4.13: Stability (small enterprises)

Score-Range	Develop. #	Validation #	Develop. %	Validation #	Stability
≤ ,03794	270	76	10,0%	4,2%	0,05
,03795 - ,05633	267	182	9,9%	10,1%	0,00
,05634 - ,11170	539	445	20,0%	24,8%	0,01
,11171 - ,16676	269	202	10,0%	11,2%	0,00
,16677 - ,28776	269	248	10,0%	13,8%	0,01
,28777 - ,48557	269	167	10,0%	9,3%	0,00
,48558 - ,77295	269	191	10,0%	10,6%	0,00
,77296 - ,95615	269	169	10,0%	9,4%	0,00
,95616 +	269	117	10,0%	6,5%	0,01
Total	2690	1797	100,0%	100,0%	0,09

Table 4.14: Stability (large enterprises)

4.7 Conclusion

In this Chapter we set out to explore variables that are statistically significant enough to enhance the predictability of two credit risk assessment models (for small and large enterprises). To that end, we constructed two new credit risk models that evaluate Greek enterprises using a combination of financial and credit behavior data.

Subsequently, these models were compared with two other models based solely on financial data. The comparison was made in terms of accuracy, K-S and Gini Index. As it was shown in Sections 4.4 and 4.5, the financial variables that constitute the models based solely on financial data are not the same with those based on both financial and credit behavior data. Indeed, financial variables used in the traditional models (with financial data only) are no longer as statistically significant as the new financial variables used for the construction of the advanced models (based on the combined data).

After the models' comparison we concluded that the new models contribute to credit risk estimation as it is clearly shown by their respective performances. Tables 4.15 and 4.16 show the differences between the two competing models by comparing the above-mentioned performance indices.

- **Small Enterprises:**

Performance Indicators	Model with combination of data	Model with financial data only
Accuracy	85,5%	71,0%
K-S	64,6%	33,8%
Gini Index	0,80	0,44

Table 4.15: Comparison of performance indicators (small enterprises)

- **Large Enterprises:**

Performance Indicators	Model with combination of data	Model with financial data only
Accuracy	85,3%	78,8%
K-S	67,2%	41,2%
Gini Index	0,82	0,51

Table 4.16: Comparison of performance indicators (large enterprises)

Finally, the efficiency and stability of the models were studied with very satisfactory results which clearly confirm their applicability in different time periods.

Previously, we focused on the character of the independent variables and on classification methods that are used in predictive models. Subsequently, we will focus on a technique for dimension reduction, namely for reducing the number of variables contained in predictive models. More details on this issue are discussed in the following Chapter.

Chapter 5

Multiple dimension reduction for credit scoring modelling and prediction

Credit rating modelling is of great interest in Finance and Banking since as early as the 50's and the 60's (e.g., Durand, 1941; Myers and Forgy, 1963; Altman, 1968). Logistic Regression and Discriminant Analysis are considered classical parametric techniques for credit scoring modelling with the Discriminant Analysis introduced by Altman, 1968 and the Logistic Regression discussed among others, by Crook *et al*., 2007. For a general review of early methods in credit scoring modelling techniques, the interested reader may refer to Eisenbeis, 1978. A linear programming alternative method of analysis has been proposed by Hardy and Adrian, 1985. Artificial intelligence methods have also been considered for credit rating models some of which are based on Support Vector Machines (SVM) (e.g., Bellotti and Crook, 2009; Yu *et al*., 2010; Chen *et al*., 2011). Methods related to Neural Networks are discussed among others in Tam and Kiang, 1992; Boritz and Kennedy, 1995; Kumar, 2005. Furthermore, for highly unbalanced credit rating data, Paleologo *et al*., 2010 proposed the use of an ensemble classification technique called subagging.

In this Chapter, we contribute to the problem of dimension reduction in credit scoring modelling. More specifically, the purpose of this analysis is the development of a flexible and reliable forecasting modelling approach for a response variable representing the business credit behavior characterized according to Basel II [5], as "good" (i.e., with "no delinquency") or "bad" (i.e., with "severe delinquency") with covariates associated not only with financial characteristics but also with credit behavior characteristics. For the modelling, we propose a 3(4)-step algorithmic procedure for dimension reduction with an initial preliminary data pre-processing step (step0). The latter is used for creating dummy variables using Weight-of-Evidence (WOE) which measures the strength of each attribute in separating the bad from the good enterprises (see Section 3.1.3). The main part of the algorithm is based on dimension reduction techniques taking into consideration

- *a stepwise Akaike Information Criterion* (hereafter, stepAIC) and
- *a Principal Component Analysis* (PCA).

The proposed procedure allows for an optional 4th step based on **Elastic Net Regularization** (Zou and Hastie, 2005) for further dimension reduction if the researcher feels that it is of use. Note that in this analysis, we rely on logistic regression, as a general methodology for the final credit scoring model fitting, because it gives a prompt answer about the fitting of a credit scoring model, including the set of significant covariates. Moreover, logistic regression provides a direct estimation of the Probability of Default (PD) defined in Definition 1.1(c), both for an enterprise and for the entire Financial System. In some instances (Lee and Jung, 2000), the predictive power of logistic regression comparing to other techniques like e.g., the neural networks, relies on specific characteristics of subgroups existing in the same sample.

The proposed procedure is applied to the Greek system separately for small and large enterprises (according to their revenue). Through this analysis we expect to succeed in choosing the optimal predictive model for enterprises. The method is also appealing due to the use of the popular logistic regression analysis. Indeed, it should be noted that after the 2-level dimension reduction procedure we choose to use the standard logistic regression instead of other complex methods because it has proved its efficiency over the years, and it is easy to be explained to a general audience. Finally, the proposed methodology is responsive to the need of dimension reduction for the construction of a flexible yet reliable credit scoring model for descriptive as well as predictive purposes.

5.1 Introduction

5.1.1 Principal Component Regression (PCR)

In the field of statistics, principal component regression (PCR) is a regression analysis technique based on principal component analysis (PCA). Typically, it assumes that the deflection of the result (i.e., the response or otherwise the dependent variable) on a set of factors (or independent variables) is based on a standard linear regression model but uses PCA to estimate the unknown regression coefficients in the model.

In PCR, instead of regressing the dependent variable directly with the independent variables, main components are used as independent variables. Usually, only one subset of all the main components is used for regression. Often, the main components with higher variability (those based on eigenvectors corresponding to the higher eigenvalues of the variance-covariance matrix of the explanatory variables) are selected as independent variables. However, as the goal is to predict the outcome, the key components with lower fluctuations can also be important.

An important use of PCR is that it overcomes the problem of multicollinearity that arises when two or more of the explanatory variables are close to being collinear. PCR can treat such conditions by excluding some of the main components that have low variability in the regression step. In addition, as regression is performed with only one subset of all major components, PCR can reduce the dimension, substantially reducing the number of parameters that characterize the underlying model. This can be especially helpful in cases with oversized agents. Also, by selecting the main components to be used for the regression, PCR can lead to effective model-based prediction.

The PCR method can be divided into three steps:

1. Perform the PCR in the observed data matrix for the explanatory variables to obtain the main components and then select a subset of them based on appropriate criteria.

2. We now map the observed vector of the results to the selected main components as independent variables, using the usual least squares regression (linear regression) to get a vector of the estimated regression coefficients (with dimension equal to the number of selected squares).

3. We now transform this vector into the scale of real independent variables, using the selected PCA (the eigenvectors corresponding to the selected principal components) to get the final PCR estimator (dimension equal to the total number of covariates) for the estimation of the regression coefficients that characterize the original model.

5.1.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of correlated variables into a set of linearly unrelated variables called principal components. If there are *n*

observations and *p* variables, then the number of principal components is min (*n-1, p*). This transformation is defined in such a way that the first major component has the greatest possible variability (i.e., represents the greatest possible variability in the data), and each subsequent component in turn has the greatest possible variability if it is orthogonal with the previous components. The resulting vectors (each is a linear combination of the initial variables) are an unrelated orthogonal basis set.

Dimension reduction

The transformation $T = XW$ maps a data vector $x_{(i)}$ from an initial space of *p* variables to a new space of *p* variables that are not related to the data set. However, we do not need to keep all the key components. Retaining only the first *L* principal components, produced using only the first *L* eigenvectors, gives the transformation:

$$T_L = XW_L,$$

where the matrix T_L now has *n* rows and only *L* columns. In other words, PCA gives a linear transformation $t = W^T x, x \in R^p, t \in R^L$, where the columns of the $p \times L$ matrix *W* form an orthogonal basis for the *L* features. Based on all transformed matrixes with only *L* columns, this scoreboard maximizes the variability of the original data retained while minimizing the total squared reconstruction error:

$$|| TW^T - T_L W_L^T ||_2^2 \quad ή \quad || X - X_L ||_2^2.$$

Such a dimensionality reduction approach can be an especially useful step in displaying and processing a large data set, while maintaining the data set variability as much as possible. For example, selecting only L = 2 components the researcher identifies a 2-dimensional level through the large data set, in which the data is most spread out. So, if the data contains clusters, they can be more expansive, and therefore more visible to be drawn in a two-dimensional diagram, while if they randomly select two directions through the data (or two of the original variables), the clusters can be much less spread out between each other and may in fact be much more likely to overlay each other, which will make them difficult to separate.

Similarly, in regression analysis, the greater the number of explanatory variables allowed, the greater the likelihood of overfitting the model, drawing conclusions that cannot be generalized to other data sets. One

approach, especially when there are strong correlations between different possible explanatory variables, is to reduce them to a few key components and then run the regression. This method is the PCR described in Section 5.1.1.

5.2 Data description and pre-processing

The data for the analysis in this work constitute a representative random sample of 4579 Greek businesses chosen from the database of Tiresias S.A. The random sample used in this analysis consists of 1889 small enterprises (with revenue at most 700,000 euros) and 2690 large enterprises (with revenue at least 700,000 euros). In this work, as it is typical in such analyses (e.g., Siddiqi, 2006),

- a period of twelve (01/01/2014 to 31/12/2014) months is used as a performance period and

- a 2-year period (01/01/2012 to 31/12/2013) as an observation period.

The purpose of the analysis is the modelling of a response variable representing the business credit behavior characterized as either "good" or "bad". (see Section 2.5). The covariates used for the analysis are divided into two main categories that correspond to financial data and credit behavior data (see Section 4.2).

- **For small enterprises**, a total of 39 covariates (with 73,661 observations) have been used in this analysis, of which
  - ❖ **27 are financial** and
  - ❖ **12 are credit behavior** variables.

- **For large enterprises** we use a total of 49 covariates (with 131,752 observations) of which
  - ❖ **37 are financial** and
  - ❖ **12 are credit behavior** variables.

The above variables are presented in detail, in Appendix B.

In statistical terms the scope of this work is the modelling of a binary classification problem for credit scoring. For the analysis we will be using multivariate analysis techniques including logistic regression and model selection criteria for the identification of the most significant financial and credit behavior covariates for predictive purposes. The analysis is

performed separately for small and large enterprises with the description of the selected covariates given in Section 5.4.

In order to prepare the data for the main part of the analysis, we proceed into standard pre-processing operations by grouping the independent variables and creating dummy variables using weight-of evidence (WoE) coding (see Section 3.1.3).

5.3 Dimensionality reduction

In data analysis, the first and most crucial problem that a researcher should overcome is the correct data interpretation. Indeed, whenever we deal with big datasets like the ones in this work, we are entering into the field of Big Data Analysis where the existence of collinearity is, among others, one of the most serious problems encountered associated with unreliable results (e.g., Maheshwari 2020; Yoo *et al*., 2014). During a preliminary analysis, various models, techniques as well as combinations of techniques have been considered for both datasets for small and large enterprises, with the optimal combination resulting in a 3(4)-step algorithmic procedure consisting of the following:

- **Step 1** *Data Standardization*
- **Step 2** *step AIC*
- **Step 3** *PCA*
- **Step 4** *Elastic Net Regularization* (optional step)

The purpose of the above algorithm is the dimension reduction which is achieved in two levels (in steps 2 and 3): firstly, using the stepAIC procedure applied to the standardized variables of step 1 and later by performing PCA in the variables selected by stepAIC. Based on the data and the final results of this analysis, an additional step (step 4) is recommended to be included as optional, in the above algorithmic procedure. The use of this optional step is recommended if the data justify its use. More specifically, after the 3-step algorithm is completed, a logistic regression analysis is performed using the covariates selected in the latter step of the procedure. The optional (4th) step can be considered as a 3rd level dimension reduction technique which removes, via Elastic Net Regularization, those (PCA) variables that do not contribute significantly to the proposed logistic regression model. The optimal models for both

datasets (for small and large enterprises) were selected based on two frequent used criteria, namely AIC and Adjusted R^2. The proposed algorithmic procedure addresses and succeeds to resolve the problem of multicollinearity and any other consequence of dealing with Big Data and the limitation of the explanatory variables (covariates) and, on one hand, making it possible to identify a flexible and easy-to use model for predictive purposes and, on the other, a clear and precise interpretation of the results.

5.3.1 Step 1: Data Standardization

In standard data analysis, data standardization is often recommended before PCA. Indeed, if PCA is performed directly to the original explanatory variables, the new emerged PCA variables fail to be (fully) independent, although this is the main goal of the implementation of PCA. This phenomenon may be attributed to heavy multicollinearity between explanatory variables with different measurement scales (Ntotsis and Karagrigoriou, 2020). In our analysis we observed a high degree of multicollinearity as indicated by the Variance Inflation Factor (VIF; results not shown). In order to limit or eliminate it, a data standardization was done, which affects considerably the correlations involved. After the first step of the procedure, the multicollinearity in both datasets was observed to be significantly reduced although still existed.

5.3.2 Step 2: Stepwise Akaike Information Criterion

After the data standardization, the stepAIC procedure was applied as the first-dimension reduction / feature selection criterion. This technique is one of the most common techniques used which attempts to find the optimal subset of variables (features) by minimizing the AIC value among the competing candidate models. StepAIC can keep intact the larger possible part of the model's performance by simplifying it which results in the quantification of the amount of information loss. Note that, at each stage of the process, the technique checks whether variables that were removed in a previous phase become significant and are required to return to the model (for more, see Yamashita *et al*., 2007 and Zhang, 2016).

The overall results of step 2 of the algorithm for both small and large enterprises are presented in Table 5.1, while Table 5.2 contains the stepAIC selected variables to be used in step 3 of the analysis.

Small enterprises	Large enterprises
AIC=1165	AIC=1215
R^2=50%	R^2=50%
Adjusted R^2=50%	Adjusted R^2=49%
15 variables (from the original 39) remained for further analysis	18 variables (from the original 49) remained for further analysis

Table 5.1: Small and large enterprises model selection summary - stepAIC

After the implementation of stepAIC procedure, R^2 and Adjusted R^2 for both small and large enterprises remain unchanged as in the original full model. Nonetheless, a noteworthy decrease in the AIC value can be observed in both cases. The AIC of the full model drops from 1200 to 1165 and from 1235 to 1215 after the implementation of the first -stepAIC- dimension reduction technique, in small and large enterprises, respectively. Additionally, even if the AIC resulted in same values for the full and the stepAIC model, the second one would be preferred due to its simplicity. One substantial dexterity of stepAIC, is that the resulted models contain approximately only the 38% (small) and 36% (large) of the variables used in the full model. As a result, the proposed models are more flexible and thus, preferable for predictive purposes, than the original ones based on a well-established model selection criterion, AIC.

Small Enterprises	Large Enterprises
Financial Variables	
Debt equity ratio	Cash ratio
Return on equity	Current assets to total liabilities
Working capital leverage	Net profit margin
Total assets turnover ratio	Current liabilities turnover ratio
Return on assets	Fixed assets to equity
Total liabilities	Working capital turnover ratio
Short-term liabilities	Total liabilities
Return curried forward	Long-term liabilities
Profit before taxes depreciation and amortization expense	Total fixed assets
	Short-term liabilities

Behaviour Variables	
Worst Payment Status in Last 3 Months	Maximum Utilization Not Revolving
Maximum Number of Months Consecutive with Over 100% Utilization in Last 6 Months	Worst Payment Status Last Month vs Last 24 Months
Number of Occurrences with Delinquency 1+ in Last 12 Months	Worst Payment Status in Last 3 Months
Maximum Number of Months Consecutive with Over 100% Utilization in Last 24 Months	Maximum Number of Months Consecutive with Over 100% Utilization in Last 6 Months
Current Balance/Delinquency to Current Balance	Number of Occurrences with Delinquency 1+ in Last 12 Months
Worst Payment Status Last Month vs Last 24 Months	Maximum Number of Months Consecutive with over 100% Utilization in Last 24 Months
	Total Current Balance
	Current Balance/Delinquency to Current Balance

Table 5.2: Small and large enterprises emerged variables – stepAIC

For variables' interpretation please see Section 5.4

5.3.3 Step 3: Principal Component Analysis

The 2nd level dimension reduction procedure is applied to the 15 and 18 explanatory variables (see Table 5.2) by stepAIC for the small and large enterprises, respectively. For this purpose, the classical PCA technique based on the correlation matrix is used as the second-dimension reduction technique (for more about PCA see Jolliffee, 1972; Artemiou and Li, 2009). Note that all PCA assumptions are checked and found to be fulfilled for the variables of each dataset. Note further that the purpose of this procedure is to eliminate the remaining multicollinearity that still exists in the data and to further reduce the dimensionality, through PCA variables, by taking advantage of the fact that PCA classifies the variables from the

most important to the least important according to their contribution to the overall variability. In both datasets under consideration, we choose to retain the components that interpret approximately 90% to the overall variability of the original (standardized) variables. It is noted that various scenarios were studied with 80% and 75% variability as well as Kaiser's rule. The model with components (and by extension the PCA variables) interpreting 90% of the total variability was the one for which the AIC and Adjusted R^2 values coincide with the corresponding values of the model obtained by stepAIC at the end of step 2 of the process. Although there is no specific rule to determine which variables are significant within each component, a proportion is satisfactory when it can retain a sufficient amount of the original information (Ntotsis *et al*., 2019).

It is worth mentioning that this process simplifies the model (by reducing the number of PCA variables) without sacrificing the validity and effectiveness of the proposed model.

Remark 1: According to PCA methodology, each new PCA variable is a function of all variables selected through stepAIC (step 2). Also, the PCA components are ranked from the one with the largest to the one with the smallest percentage of variability. Finally, for the calculation of PCA components and the new PCA variables (V_i) see Ntotsis *et al*., 2020 (eq. (2) and (3)).

Based on the above Remark and in order to explain at least 90% of the total variability, for the small enterprises we retain only the first 9 out of 15, V_i variables while for the large enterprises we retain the first 11 out of 18, V_i variables. For forecasting purposes, the logistic regression will be applied to both datasets, using the model with the 90% variability. The results including the coefficient estimates for both cases under investigation are presented in Tables 5.3 and 5.4.

Coefficients				
	Estimate	**Std. Error**	**t value**	**p-value**
(Intercept)	0.3155109	0.0075925	41.556	< 2e-16
V_1	-0.0662298	0.0016387	-40.416	< 2e-16
V_2	-0.0406558	0.0031273	-13.000	< 2e-16
V_3	0.0307734	0.0045453	6.770	1.71e-11
V_4	0.0187028	0.0057085	3.276	0.00107
V_5	0.0049364	0.0066315	0.744	0.45673
V_6	0.0001217	0.0094362	0.013	0.98971
V_7	-0.0051696	0.0118533	-0.436	0.66279
V_8	0.0333873	0.0160883	2.075	0.03810
V_9	-0.0375664	0.0199572	-1.882	0.05994
Residual Standard Error:	0.33	**Degrees of Freedom:**		1879
Multiple R-square:	0.4984	**Adjusted R-squared:**		0.496
AIC:	1184.08			

Table 5.3: step 3 - Small enterprises regression and AIC results

Coefficients				
	Estimate	**Std. Error**	**t value**	**p-value**
(Intercept)	0.238662	0.005891	40.515	< 2e-16
V_1	-0.045877	0.001012	-45.323	< 2e-16
V_2	0.041277	0.002426	17.018	< 2e-16
V_3	0.032956	0.003291	10.015	< 2e-16
V_4	-0.027626	0.004825	-5.725	1.15e-08
V_5	0.033111	0.005120	-6.467	1.18e-10
V_6	-0.018376	0.006501	-2.827	0.00454
V_7	-0.018376	0.007557	-1.432	0.15239
V_8	-0.023566	0.009055	-2.603	0.00930
V_9	-0.031837	0.010407	-3.059	0.00224
V_{10}	0.026516	0.011492	2.307	0.02111
V_{11}	-0.049651	0.013419	-3.700	0.00022
Residual Standard Error:	0.30	**Degrees of Freedom:**		2678
Multiple R-square:	0.4886	**Adjusted R-squared:**		0.4865
AIC:	1268.686			

Table 5.4: step 3 - large enterprises regression and AIC results

One can easily observe that AIC and Adjusted R^2 values are remarkably close to the corresponding values of the model selected with stepAIC prior to PCA implementation for both categories of enterprises (see Table 5.1). In other words, both models selected at the end of step 3 of the algorithmic procedure are much simpler than the ones selected in step 2 and at the same time retain a considerable amount of information. Hence, the second-dimension reduction approach in step 3 chooses a set of variables for each class of enterprises (with 9 and 11, respectively for small and large

enterprises). Meanwhile no significant alterations of the Adjusted R^2 and AIC results were occurred compared to the full PCA models (with 15 and 18 variables respectively).

5.3.4 Step 4: Elastic Net Regularization – Optional dimension reduction procedure

After the dimension reduction is completed, the final model is obtained by using a logistic regression analysis separately for small and large enterprises using respectively the 9 and 11 covariates selected through the proposed algorithmic procedure.

Considering the results of the logistic regression in Tables 5.3 and 5.4, we can move on to an optional third level of dimensionality reduction if the results allow it. Specifically, the results extracted through logistic regression indicated, based on t-values, statistically non-significant PCA variables (e.g., at significance level α =5%). The reduction in the number of variables combined with the fact that the removed variables are statistically insignificant often results in models with a better AIC due to a lower penalty term. In order to ratify the above observation, an Elastic Net Regularization (ENR) that favors the LASSO regression was implemented. This technique aims to combine the penalties of LASSO and Ridge regression in order to get a hybrid regularization that highlights the benefits of both techniques (Zou and Hastie, 2005). Note that for both datasets examined, ENR with α parameter fluctuating from 0.5 to 0.9 led to the same results.

The implementation of ENR reveals that in the case of small enterprises the PCA variables V6 and V7 are statistically non-significant (a result also confirmed by the Student's t-test). The final proposed model, which can be used for predictive purposes, given in Table 5.5, has a better AIC than that of step 2 of the procedure and includes 7 PCA variables (with 15 initial - standardized variables each).

For large enterprises in Table 5.4 only V7 is statistically non-significant based on t-test and also confirmed by ENR. It can be also seen from Table 5.6 that the resulted model retains its credibility since it has the same AIC value as the model selected in step 2 although it has one less (PCA) variable.

Coefficients				
	Estimate	**Std. Error**	**t value**	**p-value**
(Intercept)	0.315511	0.007588	41.581	< 2e-16
V_1	-0.066228	0.0016387	-40.449	< 2e-16
V_2	-0.040645	0.003125	-13.005	< 2e-16
V_3	0.030745	0.004542	6.769	1.73e-10
V_4	0.0186688	0.005705	3.272	0.00109
V_5	0.033111	0.005120	-6.467	1.18e-10
V_8	0.033330	0.016078	2.073	0.03831
V_9	-0.037544	0.019945	-1.882	0.05994
Residual Standard Error:	0.32	**Degrees of Freedom:**		1882
Multiple R-square:	0.4983	**Adjusted R-squared:**		0.4967
AIC:	1180.042			

Table 5.5: step 4 - Small enterprises regression and AIC results

Coefficients				
	Estimate	**Std. Error**	**t value**	**p-value**
(Intercept)	0.238662	0.005892	40.515	< 2e-16
V_1	-0.045878	0.001012	-45.314	< 2e-16
V_2	0.041302	0.002426	17.025	< 2e-16
V_3	0.032987	0.003291	10.023	< 2e-16
V_4	-0.027586	0.004826	-5.716	1.21e-08
V_5	0.033181	0.005120	-6.480	1.09e-10
V_6	-0.018380	0.006502	-2.827	0.00474
V_8	-0.023551	0.009057	-2.600	0.00936
V_9	-0.031751	0.010409	-3.050	0.00231
V_{10}	0.026567	0.011494	2.311	0.02089
V_{11}	-0.049665	0.013422	-3.700	0.00022
Residual Standard Error:	0.30	**Degrees of Freedom:**		2679
Multiple R-square:	0.4882	**Adjusted R-squared:**		0.4863
AIC:	1268.743			

Table 5.6: step 4 - Large enterprises regression and AIC results

Remark 2: The procedure of the optional step 4 technique is applicable if there is at least one statistically significant variable in the final logistic regression model. In this particular case study, the contribution for both cases could be considered relatively limited since the comprehensive dimensionality is reduced by three (dimensions) which also contribute to the improvement of the overall performance of the model.

5.4 Definitions of the selected variables

In this Section, we give the definition of the selected variables for both models. The majority commentary was derived from www.investopedia.com and Tiresias private online library.

Variables appearing only in small enterprises:

1. Debt Equity Ratio = Total Liabilities / Shareholder Equity. This ratio is used to evaluate an enterprise financial leverage.

2. Return on Equity (ROE) = Net Income / Average Shareholders' Equity. Roe is considered a measure of how effectively management is using an enterprise's assets to create profits.

3. Working Capital Leverage = Current Liabilities / Working Capital. Working capital leverage refers to the impact of level working capital on business's profitability. The working capital management should improve the productivity of investments in current assets and ultimately it will increase the return on capital employed.

4. Total Assets Turnover Ratio = Net Sales / Total Assets. This ratio measures an enterprise's ability to generate sales from its assets by comparing net sales with average total assets. It calculates net sales as a percentage of assets to show how many sales are generated from each dollar of enterprise assets.

5. Return on Assets (ROA) = Net Income / Total Assets. ROA is an indicator of how profitable an enterprise is relative to its total assets. ROA gives an idea to how efficient a business management is at using its assets to generate earnings.

6. Result Curried Forward = profits / damages.

7. Profit Before Taxes Depreciation and Amortization Expense = a profitability measure that looks at an enterprise's profit before the enterprise must pay corporate income tax and depreciation and amortization expense.

Variables appearing only in large enterprises:

1. Cash Ratio = the ratio of an enterprise's total cash and cash equivalents to its current liabilities and signifies the enterprise's ability to pay short-term liabilities with its highest liquid assets.

2. Current Assets to Total Liabilities = measures the enterprise's ability to cover its total liabilities with its total current assets. This ratio is also used to estimate the liquidity of the enterprise by showing the enterprise can pay its creditors with its current assets if the business's assets ever had to be liquidated.

3. Net Profit Margin = net profit / revenue. This ratio is used to calculate the percentage of profit a business produces from its total revenue.

4. Current Liabilities Turnover Ratio = (short-term liabilities / net revenues from sales) * number of days in the period. This ratio indicates the number of days from the moment some liability arises to the moment it is paid.

5. Fixed Assets to Equity = fixed assets / equity. It measures the contribution of stockholders and the contribution of debt sources in the fixed assets of the enterprise.

6. Working Capital Turnover Ratio = net annual sales / average working capital. This ratio measures how efficiently an enterprise is using its working capital to support a given level of sales.

7. Long-term Liabilities = an obligation resulting from a previous event that is not due within one year of the date of the balance sheet.

8. Total fixed Assets (net book value) = Its formula is calculated by subtracting all accumulated depreciation and impairments from the total purchase price and improvement cost of all fixed assets reported on the balance sheet.

9. Maximum Utilization- Not Revolving = RCS Maximum percent credit utilization – Joint / Prime – Non-Revolving- SME – Updated in last 12 months.

10. Total Current Balance = RCS Total Current Balance – Joint / Prime – Open. When referring to a loan such as an auto loan or a mortgage, your current balance is the amount you currently still owe on the loan according to the date of your statement.

Variables appearing both in small and large enterprises:

1. Total Liabilities = the aggregate of all debts an individual or enterprise is liable for and can be calculated by summing all short-term and long-term liabilities.

2. Short-term Liabilities = a financial obligation that is to be paid within one year.

3. Worst Payment Status Last Month vs Last 24 Months = RCS Worst Payment Status – Joint / Prime – Last 1 Month vs. Last 24 Months.

4. Worst Payment Status in Last 3 Months = Worst Payment Status – SME – Joint/Prime – Last 3 Months.

5. Maximum Number of Months Consecutive with over 100% Utilization in Last 24 Months = RCS Maximum Number of Months Consecutive with over 100% of Percentage Credit Utilization in last 24 months - updated in last 12 months – Joint/Prime.

6. Number of Occurrences with Delinquency 1+ in Last 12 Months = RCS Number of Occurrences of Delinquency 1+ DPD – Joint/Prime – last 12 months.

7. Current Balance / Delinquency to Current Balance = RCS Ratio Current Balance / Delinquency to Current Balance – Joint / Prime – Open – updated in last 3 months.

8. Maximum Number of Months Consecutive with over 100% Utilization in Last 6 Months = RCS Maximum Number of Months Consecutive with over 100% of Percentage Credit Utilization in last 6 Months – Updated in last 3 months – Joint / Prime.

5.5 Conclusions

Summarizing, in this Chapter we proposed two credit scoring models, one for small and one for large enterprises using two datasets for Greek enterprises. Note that both models are constructed with PCA based variables, which means that each variable in the selected logistic regression contains all variables selected by the stepAIC procedure (see Table 5.2).

The findings of this work clearly show the importance in using credit behavior variables since a number of such variables have been found to play a key role in building credit scoring models both for small and large

enterprises. Indeed, in the final model for the small businesses each PCA variable depends on six (6) credit behavior covariates (out of a total of 15 covariates) while for the large enterprises final model each PCA variable depends on ten (10) credit behavior covariates (out of a total of 18 variables). The use of such covariates is one of the main contributions of the present work since countries rely almost solely on the financial covariates. It is also noteworthy that the proposed methodology is responsive to the need of dimension reduction for the construction of flexible yet reliable credit scoring models not only for descriptive but most importantly for predictive purposes. Furthermore, the proposed methodology provides among others, insurers, financial planners and lenders with an automated reliable financial tool of evaluating credit worthiness according to a few statistically significant financial as well as credit behavior covariates and at the same time making credit decisions faster and fairer while offering to borrowers increased lending opportunities.

The dimension reduction modelling proposed in this Chapter may be applied in the fiscal debt credit scoring modelling. The significance of the prediction arising from the fiscal debt rating agencies is another part of possible extensions. The significance of the predictions concerning both credit risk rating and fiscal debt rating models may be tested by using non -parametric methods (alike Wilcoxon test).

Chapter 6

Concluding remarks

In this Chapter we gather and present the goals and contributions of the present Thesis.

In Chapter 3, we set out to explore the effectiveness of alternative data in credit scoring models. To that end, we created and introduced variables from alternative sources, to an already existing predictive model for Greek hotels which uses only credit behavior data. For this purpose, we used a real-world credit scoring data set of 678 Greek hotels. Comparing the 'alternative' model with the already existing one in terms of K-S, Gini Index and accuracy, we concluded that alternative data contribute to model's performance. More specifically, this contribution can be seen by observing the differences between the values of performance indicators for these two models: K-S: 77,0% >74,8%, accuracy: 92,9 >91,4, Gini Index: 0,90 >0,88. Having established the improvement in model's performance for Greek hotels, we can easily conclude that it would be prudent to explore alternative data's utility in other industries as well.

Subsequently, we conducted a benchmarking study of 12 classification algorithms on the same real-world data set in order to compare novel with traditional classification methods. In pursuing this objective, we compared these classification algorithms in terms of AUC and accuracy (PCC). Our results showed that there are slight differences between the values of these performance indicators in each classifier and possibly this happens because we worked on a homogenous sample. Specifically, we noticed that logistic regression and neural networks performed better than other (novel or not) classifiers and logistic regression had the highest value of AUC. From a managerial perspective, the key question is whether neural networks or other 'complex' classification algorithms can and should take the place of the industry standard logistic regression in corporate practice, given the slight differences between the values of their performance indicators. Based on the above analysis, logistic regression seems to perform significantly satisfactorily and there is no question of replacing it, at least in terms of (relatively) homogeneous data. It is noted, however, that further investigation is deemed necessary for non-homogeneous data.

This analysis provides valuable insights for professionals as they can see novel classification algorithms in predictive modelling. Furthermore, we

provide an evaluative survey of recent scoring methods to aid future research. Finally, we demonstrated the efficiency and stability of the logistic regression model by applying it at different period and in different samples.

Continuing to aim for the maximum performance of predictive models and wishing to contribute to a wider industry (not only hotels), in Chapter 4, we decided to explore a combination of data regarding the independent variables that will make up predictive models for companies. As financial data are usually the only data used in Greek models (and other countries) for the evaluation of the credit risk of a company, we now use a combination of financial and credit behavior data. In that Chapter, the main contribution of our analysis is the construction of new credit risk models which evaluate credit risk for small and large Greek enterprises by using a combination of financial and credit behavior data. Subsequently, the proposed models (with the combined data) were compared with the typical ones (containing only financial data) in terms of accuracy, K-S and Gini Index.

After the models' comparison we concluded that the new models contribute to credit risk estimation as shown by their performance. Indeed, the differences can be easily seen in Tables 4.15 and 4.16 which are reproduced below.

Finally, the efficiency and stability of the models were studied with very satisfactory results so that they can be used in different time periods.

- **Small enterprises:**

Performance Indicators	Model with combination of data	Model with financial data only
accuracy	85,5%	71,0%
K-S	64,6%	33,8%
Gini Index	0,80	0,44

Table 4.15: Comparison of performance indicators (small enterprises)

- **Large enterprises:**

Performance Indicators	Model with combination of data	Model with financial data only
accuracy	85,3%	78,8%
K-S	67,2%	41,2%
Gini Index	0,82	0,51

Table 4.16: Comparison of performance indicators (large enterprises)

In Chapter 5 we do not focus on the character of the independent variables and on classification methods that are used in predictive models, but rather on a technique for reducing the variables contained in predictive models as we wanted to contribute to the problem of dimension reduction in credit scoring modelling. More specifically, the purpose of this analysis is the development of a flexible and reliable forecasting modelling approach for a response variable representing the business credit behavior characterized according to Basel II [5], as "good" (i.e., with "no delinquency") or "bad" (i.e., with "severe delinquency") with covariates associated not only with financial characteristics but also with credit behavior characteristics. For the modelling, we propose a 3(4)-step algorithmic procedure for dimension reduction with an initial preliminary data pre-processing step (step 0). The main part of the algorithm is based on dimension reduction techniques taking into consideration a stepwise Akaike Information Criterion and a Principal Component Analysis (PCA). The proposed procedure allows for an optional 4th step based on Elastic Net Regularization for further dimension reduction if the researcher feels that it is of use.

The findings of this work clearly show the importance in using credit behavior variables since a number of such variables have been found to play a key role in building credit scoring models both for small and large enterprises. Indeed, in the final model for the small enterprises each PCA variable depends on 6 credit behavior covariates (out of a total of 15 covariates) while for the large Enterprises final model each PCA variable depends on 10 credit behavior covariates (out of a total of 18 variables). The use of such covariates is one of the main contributions of the present work since countries rely almost solely on the financial covariates. It is also noteworthy that the proposed methodology is responsive to the need of dimension reduction for the construction of flexible yet reliable credit

scoring models not only for descriptive but most importantly for predictive purposes. Furthermore, the proposed methodology provides among others, insurers, financial planners and lenders with an automated reliable financial tool of evaluating credit worthiness according to a few statistically significant financial as well as credit behavior covariates and at the same time making credit decisions faster and fairer while offering to borrowers increased lending opportunities.

The dimension reduction modelling proposed in Chapter 5 may be applied in the fiscal debt credit scoring modelling. The significance of the prediction arising from the fiscal debt rating agencies is another part of possible extensions. The significance of the predictions concerning both credit risk rating and fiscal debt rating models may be tested by using non -parametric methods (alike Wilcoxon test).

In conclusion, we summarize below the main objectives, the main features and the contributions of the present Thesis:

- The main objective of this thesis is the proposal for both descriptive and predictive purposes, of an innovative flexible and reliable approach for credit scoring modelling which is of significant importance in Finance and Banking due to its direct connection to one's creditworthiness.
- The originality and one of the main contributions of the proposed modelling methodology lies on the fact that we blend effectively financial features together with credit behavior characteristics and alternative data that have never been considered before and it is quiet original as most countries and institutions use only financial data for credit scoring modelling.
- A benchmarking study has been performed of twelve classification algorithms on a real-world credit scoring data set for comparing novel to traditional classification methods which offer valuable insights to both professionals and practitioners.
- An effective and user-friendly algorithmic procedure that has been proposed and implemented into the methodology constitutes yet, another contribution since it is responsive to the need for dimension reduction, an issue frequently encountered in practice, especially in problems classified as falling into the area of Big Data Analysis. To the best of our knowledge this is the first time that the combination

of the above multivariate techniques is being used and implemented effectively, into credit scoring modelling.

- We finally provide an evaluative survey of recent scoring methods to aid future research.

References

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4),589-609.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparison using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance*, 18(3), 505-529.

Altman, E.I. (2005). An emerging market credit scoring system for corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609.

Altman, E.I., & Haldeman, R. (1995). Corporate credit scoring models: approaches and tests for successful implementation. *Journal of Commercial Lending*, 77(9), 10-22.

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation,* Oxford University, New York.

Artemiou, A., & Li, B. (2009). On principal components and regression: a statistical explanation of a natural phenomenon. *Statistica Sinica*, 19(4), 1557-1565.

Bailey, M. (2001). *Credit Scoring: The Principles and Practicalities,* White Box Publishing, Kingswood, Bristol.

Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8), 822-832.

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3), 1582-1594.

Behrman, M., Linder, R., Assadi, A. H., Stacey, B. R., & Backonja, M. M. (2007). Classification of patients with pain based on neuropathic pain symptoms: Comparison of an artificial neural network against an established scoring system. *European Journal of Pain*, 11(4), 370-376.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1), 171-182.

Bellotti, T., & Crook, J.N. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2), 3302-3308.

Bellotti, T., & Crook, J.N. (2009b). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60, 1699-1707.

Bensik, M., Sarlija, N., & Zekic-Susan, M. (2005). Modeling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent System in Accounting, Finance and Management*, 13(3), 133-150.

Beynon, M.J. (2005). Optimization object classification under ambiguity/ignorance: application to the credit rating problem. *Intelligent System in Accounting, Finance and Management*, 13(2), 113-130.

Black F. and Scholes M. (1973). The Pricing of Options and Corporate Liabilities, *The Journal of Political Economy*, 81 (3), 637-654.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453.

Boritz, J.E., & Kennedy, D.A. (1995). Effectiveness of Neural Networks Types for Prediction of Business Failure. *Expert Systems with Applications*, 9, 503-512.

Chen, N., Ribeiro, B., Vieira, A., Duarte, J.M.M., & Neves, J.C. (2011). A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Systems with Applications*, 38(10), 12939-12945.

Chiang, W. K., Zhang, D., & Zhou, L. (2006). Predicting and explaining patronage behavior toward web and traditional stores using neural networks: A comparative analysis with logistic regression. *Decision Support Systems*, 41(2), 514-531.

Chuang, C., & Lin, R. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, 36(2/1), 1685-1694.

Cramer, J.S. (2004). Scoring bank loans that may go wrong: a case study. *Statistica Neerlandica*, 58(3), 365-380.

Crook, J.N., Edelman, D.B., & Thomas, L.C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447-1465.

Desai, V.S., Crook, J.N., & Overstreet, G.A. (1996). A comparison of neural networks and linear scoring models in the credit union environment, *European Journal of Operational Research*, 95(1), 24-37.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352-359.

Durand, D. (1941). Risk elements in consumer instalment financing. National Bureau of Economic Research, New York.

Dutta, S., Shekhar, S., & Wong, W.Y. (1994). Decision support in non-conservative domains: generalization with neural networks. *Decision Support System*, 11(5), 527-544.

Dvir, D., Ben-Davidb, A., Sadehb, A., & Shenhar, A. J. (2006). Critical managerial factors affecting defense projects success: A comparison between neural networks and regression analysis. *Engineering Applications of Artificial Intelligence*, 19(5), 535-543.

Eisenbeis, R. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance*, 2, 205-219.

Emel, A., Oral, M., Reisman, A., & Yolalan, R. (2003). A credit scoring approach for the commercial banking sector. *Socio-Economic Planning Sciences*, 37(2), 103-123.

Falangis, K., & Glen, J.J. (2010). Heuristic for feature selection in mathematical programming discriminant analysis models. *Journal of the Operational Research Society*, 61(5), 804-812.

Finlay, S. (2011). Multiple classifier architectures and their applications to credit risk assessment. *European Journal of Operational Research*, 210(2), 368-378.

Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: An application using bankruptcy data. *Information and Management*, 24(3), 159-167.

Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of

sufficient data. *Journal of the Operational Research Society*, 61(3), 486-501.

Friedman, J.H. (1999). Greedy function approximation: A gradient boosting machine. *IMS 1999 Reitz Lecture*.

Guang-Bin, H., Lei, C., & Chee-Kheong, S. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions in Neural Networks*, 17(4), 879-892.

Gup, B. E., & Kolari, J. W. (2005). Commercial Banking: The management of risk. Alabama: John Wiley & Sons, Inc.

Hand, D. J., & Hanley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523-541.

Hand, D.J., & Jacka, S.D. (1998). *Statistics in Finance*. Arnold Applications of Statistics, London.

Hand, D.J. (1997). *Constructions and Assessment of Classification Rules*. Chichester: John Wiley.

Hardy, W.E.Jr., & Adrian, J.L. (1985). A linear programming alternative to discriminant analysis in credit scoring. *Agribusiness*, 1, 285-292.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (2nd ed.). Springer Series in Statistics, Springer, New York, USA.

Holsapple, C.W. and Whinston, A. B. (1987). *Business Expert Systems*, Homewood, IL, Irwin.

Hsieh, N-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers, *Expert Systems with Applications*, 27(4), 623-633.

Huang, G-B., Zhu, Q-Y., & Siew, C-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3), 489-501.

Hull, J.C., Nelken, I., & White, A.D. (2005). Merton's model, credit risk and volatility skews. *Journal of Credit Risk*, 1(1), 3-27.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007), 453-461.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks and discriminant analysis. *Expert Systems with Applications*, 13(2), 97-108.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2), 160-173).

King, R. D., Feng, C., & Sutherland, A. (1995): Stat Log: Comparison of Classification Algorithms on Large Real-World Problems. *Applied Artificial Intelligence*, 9(3), 289–333.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86.

Kumar, A.U. (2005). Comparison of neural networks and regression analysis: A new insight. *Expert Systems with Applications*, 29, 424-430.

Kuncheva, L.I. (2004). *Combining Pattern Classifiers Methods and Algorithms.* Hoboken, Wiley.

Landajo, M., Andres, J. D., & Lorca, P. (2007). Robust neural modelling for the cross-sectional analysis of accounting information. *European Journal of Operational Research*, 177(2), 1232-1252.

Lee, T.H., & Jung, S.C. (2000). Forecasting creditworthiness: Logistic vs. artificial neural net. *The Journal of Business Forecasting Methods and Systems*, 18, 28-30.

Lee, T., & Chen, I. (2005). A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines. *Expert Systems with Applications*, 28(4), 743-752.

Lenard, M. J., Alam, P., & Madey, G. R. (1995). The application of neural networks and a qualitative response model to the auditor's going concern uncertainly decision. *Decision Science*, 26(2), 209-227.

Lensberg, T., Eilifsen, A., & McKee, T. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research*, 169(2), 766-797.

Leshno, M., & Spector, Y. (1996). Neural network prediction analysis: the bankruptcy case. *Neurocomputing*, 10(2), 125-147.

Lewis, E.M. (1992). *An Introduction to Credit Scoring*. Fair, Isaac & Co., Inc., California.

Liang, Q. (2003). Corporate financial distress diagnosis in china: empirical analysis using credit scoring models. *Hitotsubashi Journal of Commerce and Management*, 38(1), 13-28.

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, 56(9), 1099-1108.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28(1), 161-170.

Maheshwari, A. (2020). Data analytics made accessible, ed. 2020. USA.

Marques, A.I., Garcia, V., & Sanchez, J.S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, 39(11), 10244-10250.

Mays, E. (2001). *Handbook of Credit Scoring*, Glenlake Publishing Company, Ltd., Chicago.

Merton, R. (1974). On the pricing of Corporate Debt: The Risk Structure of Interest Rates, *Journal of Finance*, 29(2), 449-470.

Min, J.H., & Jeong, C. (2009). A binary classification method for bankruptcy prediction. *Expert Systems with Applications*, 36(3), 5256-5263.

Min, J.H., & Lee, Y-C. (2008). A practical approach to credit scoring. *Expert Systems with Applications*, 35(4), 1762-1770.

Myers, J.H., & Forgy, E.W. (1963). The development of numerical credit evaluation systems. *Journal of the American Statistical Association*, 58(303), 799-806.

Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research*, 180(1), 354-368.

Ntotsis, K., Kalligeris, E.N., & Karagrigoriou, A. (2020). A comparative study of multivariate analysis techniques for highly correlated variable

identification and management. *International Journal of Mathematical, Engineering and Management Sciences*, 5(1), 45-55.

Ntotsis, K., & Karagrigoriou, A. (2020). The impact of multicollinearity on big data multivariate analysis modeling. In *Applied Modeling Techniques and Data Analysis,* ed. Dimotikalis et al., Wiley. *(to appear)*

Ntotsis, K., Papamichail, M., Hatzopoulos, P., & Karagrigoriou, A. (2019). On the modelling of pension expenditures in Europe. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 6(1), 50-68.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, 201(2), 490-499.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1), 2-17.

Pavlidis, N. G., Tasoulis, D. K., Adams, N. M., & Hand, D. J. (2012). Adaptive consumer credit classification. *European Journal of Operational Research Society*, 63(12), 1645-1654.

Pendharkar, P.C. (2005). A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers and Operations Research*, 32(10), 2561-2582.

Quinlan, J.R. (1992) Learning with Continuous Classes. *Proceedings of Australian Joint Conference on Artificial Intelligence*, Hobart 16-18 November 1992, 343-348.

Salchenberger, L.M., Cinar, E.M., & Lash, N.A. (1992). Neural networks: a new tool for predicting thrift failures. *Decision Sciences*, 23(4), 899-916.

Shinde, A.S., & Takale, K.C. (2012). Study of Black-Scholes Model and its Applications. *Procedia Engineering*, 38, 270-279.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, Inc., New Jersey.

Sohn, S.Y., & Ju, Y. H. (2014). Updating a credit-scoring model based on new attributes without realization of actual data. *European Journal of Operational Research*, 234(1):119–126.

Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183(3), 1477-1487.

Steenackers, A., & Goovaerts, M. J. (1989). A Credit Scoring Model for Personal Loans. *Insurance: Mathematics and Economics*, 8(8), 31-34

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operational Research*, 50(2), 277-289.

Sustersic, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications*, 36(3), 4736-4744.

Tam, K.Y., & Kiang, M.Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management Science*, 38(7), 926-947.

Thomas, L.C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and Its Applications*, SIAM, Monographs on Mathematical Modeling and Computation.

Thomas, L. C., Banasik, J., & Crook, J. N. (2001). Recalibration scorecards. *European Journal of Operational Research Society*, 52(9), 981-988.

Tong, E.N.C., Mues, C., & Thomas, L.C. (2012). Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research*, 218(1), 132-139.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, 32(10), 2543-2559.

Wu, I.D., & Hand, D.J. (2007). Handling selection bias when choosing actions in retail credit applications. *European Journal of Operational Research*, 183(3), 1560-1568.

Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of System Science and System Engineering*, 15(4), 419-435.

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Communications in Statistics – Theory and Methods*, 36(13), 2395-2403.

Yoo, W., Mayberry, R., Bae, S., Singh, K., Peter He, Q., & Lillard, J.W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, 4(5), 9-19.

Yu, L., Yue, W., Wang, S., & Lai, K.K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, 37(2), 1351-1360.

Zekic-Susan, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural networks, and decision tree models. *26th International Conference on Information Technology Interfaces*, Croatia.

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, 4(7), 136.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301-320.

Basel Committee on Banking Supervision, Basel 2 Accord, http://www.bis.org/publ/bcbsca.htm

ISDA, 1999 Credit Derivative Definitions, http://credit-deriv.com/isdadefinitions.htm , Basel Committee on Banking Supervision, Basel 2 Accord, http://www.bis.org/publ/bcbsca.htm.

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

https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

http://www.tiresias.gr/

Appendix A

Code for the benchmarking study

Using R and specifically R Studio Version 1.0.153 we implemented the same process with different machine learning algorithms in order to compare the results based on specific metrics (accuracy, AUC). We must point out that we worked on specific data sets so that it is meaningful to compare the above metrics. Subsequently, an indicative code follows in order to describe in detail the procedure followed.

First, we load the data set files. To achieve this, we use the read_ function from the readr library. So, we have the two data sets, training and control:

```
library(readr)

training_data <- read_delim("actual training data_numeric lbl.csv",";",
escape_double = FALSE)

testing_data <- read_delim("actual training data_numeric lbl.csv",";",
escape_double = FALSE)
```

The result is a table with data where in the first columns are the variables and then the category in which each observation belongs.

The goal is to use the set of training data derived from a machine learning algorithm in order to export the model. This model will make the predictions in the test data set. Finally, the export of metrics for comparing machine learning algorithms will be achieved by comparing the predictions with the actual values of the data set.

Subsequently, we separate the subset of data used by the algorithm function to calculate the model. In order a learning algorithm to train the model, it uses the variables and the group (label, class) in which each observation belongs. More specifically, in the resulting table we separate the variables from the group to which they belong.

```
trSet<-training_data[-1]

teSet<-testing_data[-1]

trcl <-as.matrix(actual_training_data_numeric_lbl[,1])

tecl <-as.matrix(actual_testing_data_numeric_lbl[,1])
```

Subsequently, we train the model by using the function that corresponds to the algorithm we want to study. For example, if we want Random Forest, we will use the randomForest(trSet, trcl) function. To call it we need to install the library

with the command install.packages ("randomForest") and then retrieve it using the command library(randomForest). So, having the following two commands:

rfFit <- randomForest(trSet, trcl)

ranForPred <- predict(rfFit, teSet)

we produce the rfFit model and we use it for predictions, resulting in ranForPred. It should also be noted that the result is not the group to which the observation belongs. The result is a score based on which we can set a threshold to distinguish the two groups. In other words, if the result of the score for an observation is 0.4 and the threshold is 0.5, then our observation belongs to category 0 and not to 1. Conversely, if threshold is 0.3, then the observation belongs to our category 1. This makes predictions more flexible and our models more efficient.

The above procedure is extended to most algorithms. Here we mentioned Random Forest as an example.

We can also use the AUC as a measure of comparison. To calculate it we use the pROC package:

roc_obj <- roc(tecl, ranForPred)

auc(roc_obj)

The result we get is a measure of comparison and the closer to 1 the value, the best prediction.

As an alternative measure of comparison, we have the accuracy (accuracy) which as mentioned is the percentage of correct predictions of the resulting model. There are several ways to calculate this and many functions have been written in different libraries. As mentioned before, we use (arbitrarily) a threshold with a value of 0.5 and see in which group the prediction belongs to.

To examine the group to which the prediction belongs:

Predcl <- ifelse(ranForPred < 0.5 , 0, 1)

and to calculate the differentiation of the two columns that emerged:

which(predcl!=tecl)

We calculate the number of different values in the two columns with:

FalsePrediction <- length (which(predcl!=tecl))

And the accuracy is equal to the percentage of correct predictions (1-percentage of incorrect predictions):

acc <- 1-falsePrediction/length(tecl)

Here, the initial thought is that the closer to 1 the predictive value, the better our predictive model.

Appendix B

List of the variables used

VARIABLES FOR SMALL ENTERPRISES

Financial Variables:

1. Current Liabilities to Equity = Current Liabilities / Equity.
2. Debt Equity Ratio = Total Liabilities / Shareholder Equity.
3. Current Assets to Total Liabilities = Current Assets / Total Liabilities.
4. Capital Structure = Total Liabilities / Total Equity.
5. Current Debt to Sales = Current Debt / Sales.
6. Current Liabilities Turnover Ratio = (short-term liabilities / net revenues from sales) * number of days in the period.
7. Return on Equity (ROE) = Net Income / Average Shareholders' Equity.
8. Equity Net Worth to Total Liabilities = Equity Net Worth / Total Liabilities.
9. Retained Earnings to Total Assets = Retained Earnings / Total Assets.
10. Working Capital Leverage = Current Liabilities / Working Capital.
11. Total Assets Turnover Ratio = Net Sales / Total Assets.
12. Pretax Profit Margin = Expenses - Taxes / Sales.
13. Net Profit Margin = Net Profit / Revenue.
14. Return on Assets (ROA) = Net Income / Total Assets.
15. Pretax Return on Equity = Net Income – Taxes / Average Shareholders' Equity.
16. Net Working Capital to Total Assets = Net Working Capital / Total Assets.
17. Earnings before Interest and Taxes to Total Assets = Earnings before Interest and Taxes / Total Assets.
18. Net Sales Turnover to Total Assets = Net Sales Turnover / Total Assets.
19. Total Liabilities
20. Short-term Liabilities
21. Interest and Related Expenses

22. Bank Short-term Payable
23. Long-term Liabilities
24. Pretax Profit
25. Profit before Tax Depreciation Amortization
26. Net Profit after Tax
27. Income Tax

Credit Behavior Variables:

28. RCS Maximum Percent Credit Utilization - Joint/Prime – Closed.
29. RCS Maximum percent credit utilization – Joint / Prime – Non-Revolving- SME – Updated in last 12 months.
30. RCS Maximum Percent Credit Utilization - Joint/Prime - Revolving-SME - Updated in Last 12 Months.
31. RCS Worst Payment Status - Joint/Prime - Last 1 Month vs. Last 24 Months.
32. RCS BAL Number of Months with Consecutive Increase in last 6 Months of Maximum Percent Credit Utilization - Joint/Prime - Open - Updated in Last 6 Months.
33. RCS Worst Payment Status - SME - Joint/Prime - Last 3 Months.
34. RCS Maximum Number of Months Consecutive with over 100% of Percent Credit Utilization in last 6 Months - Updated in Last 3 Months - Joint/Prime.
35. RCS Number of Occurrences of Delinquency 1+DPD - Joint/Prime - Last 12 Months.
36. RCS Maximum Number of Months Consecutive with over 100% of Percent Credit Utilization in last 24 Months - Updated in Last 12 Months - Joint/Prime.
37. RCS Total Current Balance - Joint/Prime – Open.
38. RCS Ratio Current Balance with Delinquency to Current Balance - Joint/Prime - Open - Updated in Last 3 Months.
39. RCS Maximum Percent Credit Utilization - Joint/Prime - Revolving-SME - Updated in Last 24 Months.

VARIABLES FOR LARGE ENTERPRISES

Financial Variables:

1. Quick ratio = Cash + Accounts Receivable / Current liabilities.
2. Cash ratio = Cash / Current liabilities.
3. Current Liabilities to Equity = Current Liabilities / Equity.
4. Debt Equity Ratio = Total Liabilities / Shareholder Equity.
5. Total Liabilities to Total Assets = Total Liabilities / Total Assets.
6. Current Assets to Total Liabilities = Current Assets / Total Liabilities.
7. Capital Structure = Total Liabilities / Total Equity.
8. Working Capital Leverage = Current Liabilities / Working Capital.
9. Total Assets Turnover Ratio = Net Sales / Total Assets.
10. Current Debt to Sales = Current Debt / Sales.
40. Pretax Profit Margin = Expenses - Taxes / Sales.
11. Net Profit Margin = Net Profit / Revenue.
12. Return on Assets (ROA) = Net Income / Total Assets.
13. Pretax Return on Equity = Net Income – Taxes / Average Shareholders' Equity.
14. Current Liabilities Turnover Ratio = (short-term liabilities / net revenues from sales) * number of days in the period.
15. Return on Equity (ROE) = Net Income / Average Shareholders' Equity.
16. Retained Earnings to Total Assets = Retained Earnings / Total Assets.
17. Equity Net Worth to Total Liabilities = Equity Net Worth / Total Liabilities.
18. Net Sales Turnover to Total Assets = Net Sales Turnover / Total Assets.
19. Current ratio = Current assets / Current liabilities.
20. Fixed Assets to Equity = Fixed Assets / Equity.
21. Working Capital Turnover Ratio = net annual sales / average working capital.
22. Net Working Capital to Total Assets = Net Working Capital / Total Assets.
23. Total Liabilities
24. Long-term Liabilities
25. Interest and Related Expenses
26. Pretax Profit

27. Net Profit After Tax
28. Bank Short-term Payable
29. Income Tax
30. Result Curried Forward
31. Total Assets
32. Tangible Intangible Acquisition
33. Total Fixed Assets Underappreciation
34. Short-term Liabilities
35. Interest and Related Expenses to Ebit = Interest and Related Expenses / Ebit.
36. Current Assets to Total Liabilities (=(2013-2012)/2012)
37. Long-term Liabilities (=(2013-2012)/2012)

Credit Behavior Variables:

38. RCS Maximum Percent Credit Utilization - Joint/Prime – Closed.
39. RCS Maximum Percent Credit Utilization - Joint/Prime - Non-Revolving-SME - Updated in Last 12 Months.
40. RCS Maximum Percent Credit Utilization - Joint/Prime - Revolving-SME - Updated in Last 12 Months.
41. RCS Worst Payment Status - Joint/Prime - Last 1 Month vs. Last 24 Months.
42. RCS BAL Number of Months with Consecutive Increase in last 6 Months of Maximum Percent Credit Utilization - Joint/Prime - Open - Updated in Last 6 Months.
43. RCS Worst Payment Status - SME - Joint/Prime - Last 3 Months.
44. RCS Maximum Number of Months Consecutive with over 100% of Percent Credit Utilization in last 6 Months - Updated in Last 3 Months - Joint/Prime.
45. RCS Number of Occurrences of Delinquency 1+DPD - Joint/Prime - Last 12 Months.
46. RCS Maximum Number of Months Consecutive with over 100% of Percent Credit Utilization in last 24 Months - Updated in Last 12 Months - Joint/Prime.
47. RCS Total Current Balance - Joint/Prime – Open.
48. RCS Ratio Current Balance w/ Delinquency to Current Balance - Joint/Prime - Open - Updated in Last 3 Months.
49. RCS Maximum Percent Credit Utilization - Joint/Prime - Revolving-SME - Updated in Last 24 Months.