Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων

τομεας μαθηματικων

Αξιολόγηση Πιστωτικής Βαθμολόγησης με Χρήση Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Εμμανουήλ Φωστέρης

Επιβλέπουσα: Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Αξιολόγηση Πιστωτικής Βαθμολόγησης με Χρήση Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Εμμανουήλ Φωστέρης

Επιβλέπουσα: Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Σεπτεμβρίου, 2022.

.....	.....	.....
Χρυσηίς Καρώνη	Βασίλης Παπανικολάου	Καλλιόπη Παυλοπούλου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	ΕΔΙΠ Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2022

.....

ΕΜΜΑΝΟΥΗΛ ΦΩΣΤΕΡΗΣ

Φοιτητής Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών Ε.Μ.Π.

Στη Μνήμη του Πατέρα μου

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη της πιστωτικής βαθμολόγησης με χρήση μηχανικής μάθησης. Ο σκοπός της εργασίας είναι η λεπτομερής ανάλυση των τεχνικών, των δεδομένων και η αξιολόγηση των αποτελεσμάτων τους. Στο πρώτο κεφάλαιο γίνεται μία πλήρης ανάλυση του ορισμού της πιστωτικής βαθμολόγησης καθώς και μία ιστορική αναδρομή περιγράφοντας την πορεία και την εξέλιξή της μέσα στα χρόνια. Στην συνέχεια παρουσιάζονται αναλυτικά ο ορισμός, οι στόχοι, αλλά και οι κίνδυνοι της μηχανικής μάθησης. Έπειτα περιγράφονται τα χαρακτηριστικά του κατάλληλου μοντέλου και των κατάλληλων δεδομένων. Στο επόμενο κεφάλαιο αναλύονται οι τεχνικές που θα χρησιμοποιηθούν για να αναλυθούν τα δεδομένα. Πιο συγκεκριμένα παρουσιάζεται η λογιστική παλινδρόμηση, η παλινδρόμηση κορυφογραμμής, η παλινδρόμηση λασσο, η ελαστική παλινδρόμηση και τα δένδρα απόφασης. Στο τελευταίο κεφάλαιο γίνεται μία εφαρμογή των τεχνικών που παρουσιάστηκαν στο δεύτερο κεφάλαιο σε ένα πρόβλημα πιστωτικής βαθμολόγησης με δείγμα 366 δανειζομένων, που έχει ως στόχο την πρόβλεψη της πιθανότητας αθέτησης ενός δανείου εντός μιας πενταετίας. Συμπερασματικά, τα αποτελέσματα της έρευνας κατέδειξαν πως οι συγκεκριμένες τεχνικές ήταν δυνατόν να επεξηγήσουν τα δεδομένα σε ικανοποιητικό βαθμό και δύναται να χρησιμοποιηθούν ακόμη και από χρηματοπιστωτικά ιδρύματα.

Λέξεις Κλειδιά — Πιστωτική Βαθμολόγηση, Μηχανική Μάθηση, Λογιστική Πάλινδρόμηση, Δένδρα Απόφασης

Abstract

The subject of this thesis is the study of credit scoring using machine learning. The aim of this study is the detailed analysis of the techniques, the data and the evaluation of their results. In the first chapter there is a complete analysis of the definition of credit scoring as well as a historical review describing its course and evolution over the years. The definition, goals, and risks of machine learning are then presented in detail. The characteristics of the appropriate model and appropriate data are then described. The next chapter analyzes the techniques that will be used to analyze the data. In particular, logistic regression, ridge regression, lasso regression, elastic regression and decision trees are presented in detail. In the last chapter, the techniques presented in the second chapter are applied on a credit scoring problem with a sample of 366 borrowers, which aims to predict the probability of a loan default within a five-year period. In conclusion, the results of the research showed that the techniques presented in this thesis were able to explain the data to a satisfactory degree and can be used by financial institutions.

Keywords — Credit Scoring, Machine Learning, Logistic Regression, Decision Trees

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους μου που μου συμπαραστάθηκαν καθ'όλη την διάρκεια των σπουδών μου καθώς και την κ. Καρώνη για τον επαγγελματισμό και την στήριξή της σε εμένα και για την δυνατότητα που μου έδωσε να μελετήσω στην παρούσα εργασία ένα τόσο ενδιαφέρον θέμα.

Εμμανουήλ Φωστέρης
Σεπτέμβριος 2022

Περιεχόμενα

Λίστα Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Πιστωτική Βαθμολόγηση και εφαρμογές της

1.1 Ο ορισμός της Πιστωτικής Βαθμολόγησης (Credit Scoring)

Οι τράπεζες και άλλα χρηματοπιστωτικά ιδρύματα λαμβάνουν χιλιάδες αιτήσεις πίστωσης καθημερινά (σε περίπτωση καταναλωτικών πιστώσεων μπορεί να είναι δεκάδες ή εκατοντάδες χιλιάδες κάθε μέρα). Δεδομένου ότι είναι αδύνατη η χειροκίνητη επεξεργασία τους, αυτοματοποιημένα συστήματα χρησιμοποιούνται ευρέως από αυτά τα ιδρύματα για την αξιολόγηση της πιστωτικής αξιοπιστίας των ατόμων που ζητούν πίστωση.

Ως πίστωση (credit) ορίζεται μία συμβατική συμφωνία ανάμεσα σε ένα δανειστή και ένα δανειολήπτη βάσει της οποίας ο δανειολήπτης λαμβάνει κατά τον τρέχοντα χρόνο κάτι που έχει αξία (ένα χρηματικό ποσό για παράδειγμα) και συμφωνεί να αποπληρώσει τον δανειστή σε μεταγενέστερο χρόνο, συνήθως με τόκο (De Servigny & Renault, 2004).

Η πιστωτική βαθμολόγηση είναι το σύνολο των προγνωστικών μοντέλων μηχανικής μάθησης και των υποκείμενων τεχνικών τους που βοηθούν τα χρηματοπιστωτικά ιδρύματα στη χορήγηση πιστώσεων. Αυτές οι τεχνικές αποφασίζουν ποιος θα λάβει πίστωση, πόση πίστωση μπορεί να λάβει και ποιες περαιτέρω στρατηγικές θα ενισχύσουν την κερδοφορία των δανειστών. Οι τεχνικές πιστωτικής βαθμολόγησης αξιολογούν τον πιστωτικό κίνδυνο (κίνδυνο δανεισμού) σε έναν συγκεκριμένο πελάτη. Ο πιστωτικός κίνδυνος (credit risk) είναι ο κίνδυνος που αναλαμβάνουν οι χρηματοπιστωτικοί οργανισμοί στην περίπτωση που ο δανειοδότης (επιχείρηση ή ιδιώτης καταναλωτής) αδυνατεί να εκπληρώσει τις χρηματοοικονομικές του υποχρεώσεις έναντι του οργανισμού (Anderson, 2007). Οι αιτήσεις για ένα πιστωτικό προϊόν δεν προσδιορίζονται ως

«καλές» ή «κακές» αιτήσεις σε μεμονωμένη βάση, αλλά προβλέπουν την πιθανότητα ότι ένας αιτών με οποιοδήποτε δεδομένη βαθμολογία θα είναι "καλός" ή "κακός". Αυτές οι πιθανότητες ή βαθμολογίες, μαζί με άλλες επιχειρηματικές εκτιμήσεις, όπως τα αναμενόμενα ποσοστά έγκρισης, το κέρδος, η ανατροπή και οι ζημίες, χρησιμοποιούνται στη συνέχεια ως βάση για τη λήψη αποφάσεων.

1.2 Ιστορική Αναδρομή

Οι άνθρωποι δανείζουν και να δανείζονται από την εποχή που άρχισαν να ζουν οργανωμένα. Είναι γεγονός λοιπόν πως η έννοια της αξιοπιστίας ενός ατόμου χρησιμοποιείται εδώ και χιλιάδες χρόνια. Τα παλαιότερα όμως χρόνια, την εποχή που δεν υπήρχε ο όγκος της πληροφορίας που υπάρχει σήμερα, οι άνθρωποι βασίζονταν περισσότερο στην φερεγγυότητα ενός ατόμου. Κατά κάποιο τρόπο μπορούμε να πούμε πως οι αποφάσεις για δανεισμό παίρνονταν στην τύχη. Γιαυτό και υπήρχε ανάγκη να δημιουργηθεί ένα αμερόληπτο σύστημα που θα διέκρινε τον πληθυσμό σε διαφορετικές ομάδες που θα είχαν κάποια υπό μελέτη χαρακτηριστικά.

Μία πρώτη προσέγγιση για την δημιουργία ενός συστήματος που θα διέκρινε τον πληθυσμό σε ομάδες έγινε το 1936 από τον Fisher. Ο Fisher προσπάθησε να διαχωρίσει δύο ποικιλίες ίριδας χρησιμοποιώντας κάποια φυσικά χαρακτηριστικά (Thomas et al., 2017). Ο πρώτος που προσπάθησε να χρησιμοποιήσει αυτήν την τεχνική για τον διαχωρισμό των δανείων σε καλά και κακά ήταν ο Durant το 1941. Παρόλα αυτά η έρευνά του για το Εθνικό Αμερικάνικο Γραφείο Οικονομικών Ερευνών δεν χρησιμοποιήθηκε στην πράξη. Η έννοια της πιστωτικής ικανότητας άρχισε να χρησιμοποιείται την δεκαετία του 50'. Πιο συγκεκριμένα, το 1956 ο μηχανικός Bill Fair μαζί με τον μαθηματικό Earl Isaac, δημιούργησαν την Isaac and Company, με στόχο να δημιουργήσουν ένα τυποποιημένο και αμερόληπτο σύστημα πιστωτικού κινδύνου, όπου και τα κατάφεραν. Οι πελάτες που εξυπηρετούσαν ήταν κυρίως λιανοπωλητές, εταιρίες ταχυδρομικών παραγγελιών και οίκοι χρηματοδότησης.

Η αξία της πιστωτικής βαθμολόγησης όμως έγινε ιδιαίτερα αντιληπτή από τις τράπεζες την δεκαετία το 60' όταν άρχισε να γίνεται ευρεία η χρήση πιστωτικών καρτών. Ο μεγάλος όγκος των πελατών που αιτούνταν για πιστωτικές κάρτες καθημερινά κατέστησε αναγκαίο να αυτοματοποιηθεί η διαδικασία έκδοσης τους για οικονομικούς λόγους αλλά και για εξοικονόμηση ανθρώπινου δυναμικού. Αυτό συνέπεσε σε μία περίοδο που υπήρχε αύξηση της υπολογιστικής ισχύς οπότε ήτανε και εφικτή η αυτοματοποίηση. Αποτέλεσμα ήτανε να μειωθούν οι αθετήσεις δανείων τουλάχιστον κατά 50%. Πριν από τη χρήση των επίσημων διαδικασιών στον τραπεζικό τομέα, οι αποφάσεις δεν ήταν αμερόληπτες, με τον διευθυντή της τράπεζας να αξιολογεί την πιστοληπτική ικανότητα ενός ατόμου με βάση τις προσωπικές γνώσεις του αιτούντος. Αυτό το σύστημα είχε πολλά ελαττώματα, συμπεριλαμβανομένου του ότι ήταν αναξιόπιστο (οι αποφάσεις

μπορεί να άλλαζαν ανάλογα με τη διάθεση του διευθυντή της τράπεζας), δεν μπορούσε να αναπαραχθεί (άλλος διευθυντής μπορεί να λάμβανε διαφορετική απόφαση και το σκεπτικό πίσω από αυτές τις αποφάσεις μπορεί να μην ήταν το ίδιο), ήταν δύσκολο να διδαχθεί, ανίκανο να χειριστεί μεγάλο αριθμός αιτήσεων και, γενικά, υποκειμενικό, με όλους τους κινδύνους παράλογης προσωπικής προκατάληψης που συνεπάγονται.

Φτάνοντας λοιπόν στην δεκαετία του 80' και βλέποντας πόσο επιτυχημένη είναι η χρήση της πιστωτικής βαθμολόγησης, οι τράπεζες άρχισαν να χρησιμοποιούν αυτό το σύστημα για άλλα πιστωτικά προϊόντα πέρα από πιστωτικές κάρτες. Πιο συγκεκριμένα, επεκτάθηκε η χρήση για προσωπικά καταναλωτικά δάνεια, για στεγαστικά δάνεια και για μικρά δάνεια για επιχειρήσεις. Αυτή η περίοδος ήταν πολύ κρίσιμη για τα χρηματοπιστωτικά ιδρύματα καθώς υπήρξε σημαντική ανάπτυξη τους και έγιναν μεγάλες αλλαγές (Thomas et al., 2017).

- Οι τράπεζες αλλάζουν σημαντικά τη θέση τους στην αγορά και αρχίζουν να διαφημίζουν τα προϊόντα τους για να προσελκύσουν περισσότερους πελάτες.

- Η χρήση πιστωτικών καρτών έχει αυξηθεί δραματικά. Λόγω των αδειών πώλησης του προϊόντος, απαιτούνταν ένα μέσο για τη λήψη μιας απόφασης δανεισμού γρήγορα και όλο το εικοσιτετράωρο. Επιπλέον, ο αριθμός των αιτήσεων ήταν τέτοιος που ο διευθυντής της τράπεζας ή άλλος εξειδικευμένος αναλυτής πιστώσεων δεν θα είχε το χρόνο ή τους πόρους για να πάρει συνέντευξη από όλους τους υποψηφίους.

- Ο στόχος των τραπεζών έχει μετατοπιστεί. Προηγουμένως, οι τράπεζες επικεντρώνονταν σχεδόν αποκλειστικά σε μεγάλα δάνεια και επιχειρηματικούς πελάτες. Όμως από το 1980 και μετά ο καταναλωτικός δανεισμός αποτελούσε πλέον ένα σημαντικό και αυξανόμενο μέρος των εργασιών της τράπεζας.Όσον αφορά τα επιχειρηματικά δάνεια, στόχος ήταν σχεδόν πάντα η αποφυγή τυχόν ζημιών. Βέβαια οι τράπεζες με τον καιρό άρχισαν να συνειδητοποιούν ότι ο στόχος με τον καταναλωτικό δανεισμό δεν πρέπει να είναι η αποφυγή τυχόν ζημιών, αλλά η μεγιστοποίηση των κερδών.

Γιαυτό και αξίζει να αναφερθεί ότι σήμερα οι τράπεζες δεν εστιάζουν τόσο στην μείωση των πιθανοτήτων αθέτησης υποχρεώσεων ενός πελάτη για ένα συγκεκριμένο προϊόν παρά στη μεγιστοποίηση του κέρδους που μπορεί να αποκομίσει η εταιρεία από αυτόν τον καταναλωτή.

Μία εφαρμογή της πιστωτικής βαθμολόγησης σήμερα είναι η επιλογή της γρήγορης δανειοδότητης (fast loan) που παρέχουν οι τράπεζες στους πελάτες τους. Ένας πελάτης μπορεί να κάνει αίτηση για ένα fast loan του οποίου το ύψος είναι σχετικά μικρό (συνήθως μέχρι 6000€) και να λάβει άμεσα απάντηση έγκρισης ή απόρριψης μέσω αυτοματοποιημένης διαδικασίας που κάνει χρήση πιστωτικής βαθμολόγησης. Το μόνο που χρειάζεται είναι ένας εκκαθαριστικός λογαριασμός. Τα υπόλοιπα στοιχεία που χρειάζονται για την απόφαση τα διαθέτει ήδη η τράπεζα.

1.3 Η Αίτηση και ο Πίνακας Βαθμολογίας

Φτάνοντας λοιπόν στο σήμερα, ας εξεταστεί η διαδικασία που θα ακολουθηθεί όταν κάποιος κάνει αίτηση για ένα πιστωτικό προϊόν, φερειπείν ένα καταναλωτικό δάνειο. Αρχικά ο πελάτης θα συμπληρώσει μία αίτηση στην οποία θα του ζητούνται κάποια βασικά του στοιχεία. Έπειτα, η τράπεζα με την χρήση μοντέλων θα βγάλει μια βαθμολογία από αυτά τα στοιχεία. Να σημειωθεί πως μπορεί να μην χρησιμοποιήσει όλα τα στοιχεία που δίνει ο αιτών για να εξάγει την βαθμολογία. Η τράπεζα μπορεί επίσης να αντλήσει περαιτέρω στοιχεία και από άλλες πηγές όπως θα δούμε στην παράγραφο 1.6.2. Όσο μεγαλύτερο σκορ έχει κάποιος τόσο πιο πιθανό είναι να εγκριθεί το αίτημά του (Capon, 1982). Για να γίνει πιο κατανοητό, ας εξετάσουμε πίνακα βαθμολογίας στο σχήμα 1.3.1:

Age	**Points**
Up to 25	10
26 to 40	25
41 to 65	38
66 and up	43
Income	
Up to 40k	16
40k to 70k	28
…	
Total score	(Sum of Points)

Σχήμα 1.3.1: Παράδειγμα Πίνακα Βαθμολόγησης

Ο συγκεκριμένος πίνακας λαμβάνει υπόψιν του δύο μεταβλητές, την ηλικία και το εισόδημα. Για παράδειγμα ένα εικοσιτετράχρονος με εισόδημα 35 χιλιάδες ευρώ έχει 26 βαθμούς ενώ ένας σαρανταπεντάχρονος με εισόδημα 60 χιλιάδες ευρώ έχει 66 βαθμούς άρα και είναι πολύ πιο πιθανό να εγκριθεί το αίτημά του και να λάβει πίστωση καθώς τα μοντέλα δείχνουν ότι έχει τα χαρακτηριστικά ενός αξιόπιστου πελάτη.

Σε αυτήν την εργασία θα γίνει μία μελέτη των τεχνικών μηχανικής μάθησης που χρησιμοποιούν οι τράπεζες για να καταλήξουν σε αυτά τα συμπεράσματα και στις συγκεκριμένες βαθμολογίες δηλαδή πως καταλήγουν στο ποιος θα λάβει πίστωση και ποιος όχι.

1.4 Μηχανική Μάθηση

Η μηχανική μάθηση έχει ως στόχο την δημιουργία υπολογιστών οι οποίοι βελτιώνονται αυτόματα μέσω εμπειρίας. Είναι ένας από τους πιο ταχέως αναπτυσσόμενους τεχνολογικούς τομείς του σήμερα, και αποτελεί ένα συνδυασμό της επιστήμης των υπολογιστών (computer science) και της στατιστικής (statistics) και πιο συγκεκριμένα της τεχνητής νοημοσύνης (artificial intelligence) και της επιστήμης δεδομένων (data science). Η πρόσφατη πρόοδος στη μηχανική

μάθηση οφείλεται τόσο στην ανάπτυξη νέων αλγορίθμων εκμάθησης και θεωρίας όσο και από τη συνεχιζόμενη έκρηξη στη διαθεσιμότητα διαδικτυακών δεδομένων και στο χαμηλό υπολογιστικό κόστος των υπερσύγχρονων υπολογιστών (Ramprasad et al., 2017). Η υιοθέτηση μεθόδων μηχανικής μάθησης με χρήση δεδομένων μπορεί να βρεθεί σε ποικίλους κλάδους επιστήμης, τεχνολογίας και εμπορίου οδηγώντας σε περισσότερες αποφάσεις βασισμένες σε τεκμήρια σε πολλούς τομείς της ζωής, συμπεριλαμβανομένης της υγειονομικής περίθαλψης, της εκπαίδευσης, της οικονομικής μοντελοποίησης και του μάρκετινγκ.

Η μηχανική μάθηση έχει αναδειχθεί ως η προτιμώμενη μέθοδος για τη δημιουργία πρακτικού λογισμικού για αναγνώριση ομιλίας, όραση υπολογιστή, επεξεργασία φυσικής γλώσσας, έλεγχο ρομπότ και άλλες εφαρμογές στην τεχνητή νοημοσύνη. Πολλοί προγραμματιστές συστημάτων τεχνητής νοημοσύνης αναγνωρίζουν τώρα ότι, για πολλούς σκοπούς, η εκπαίδευση ενός συστήματος παρέχοντάς του παραδείγματα επιθυμητής συμπεριφοράς εισόδου-εξόδου μπορεί να είναι σημαντικά ευκολότερη από τον μη αυτόματο προγραμματισμό του, που απαιτεί την πρόβλεψη απόκρισης για όλες τις πιθανές εισόδους.

Ένα πρόβλημα μηχανικής μάθησης περιγράφεται ως η δυσκολία στη βελτίωση κάποιου μέτρου απόδοσης κατά την εκτέλεση μιας εργασίας μέσω κάποιου τύπου εκπαίδευσης (Mitchell, 1997). Ο στόχος της εκμάθησης της ανίχνευσης της αθέτησης σε χρηματοπιστωτικά προϊόντα, για παράδειγμα, είναι να κατηγοριοποιήσουμε κάθε αίτηση για χρηματοπιστωτικό προϊόν ως "αθέτηση" ή "μη αθέτηση". Η ακρίβεια αυτού του ταξινομητή αθέτησης μπορεί να βελτιωθεί και η εκπαιδευτική εμπειρία θα μπορούσε να αποτελείται από μια συλλογή ιστορικών αιτήσεων, καθεμία από τις οποίες προσδιορίζεται ως αθέτηση ή μη-αθέτηση εκ των υστέρων. Εναλλακτικά, μπορεί να οριστεί μια ξεχωριστή μέτρηση απόδοσης που δίνει μεγαλύτερη ποινή όταν η λέξη "αθέτηση" επισημαίνεται λανθασμένα "μη-αθέτηση" αντί για "αθέτηση" λανθασμένα χαρακτηρίζεται "μη-αθέτηση".

Μια ποικιλία από αλγόριθμους μηχανικής μάθησης έχει αναπτυχθεί για να καλύψει τη μεγάλη γκάμα δεδομένων και τύπων προβλημάτων που παρουσιάζονται σε διαφορετικά προβλήματα μηχανικής μάθησης. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να θεωρηθούν ως αναζήτηση σε έναν τεράστιο χώρο υποψηφίων προγραμμάτων για την εύρεση ενός που βελτιστοποιεί τη μέτρηση απόδοσης, καθοδηγούμενη από την εμπειρία εκπαίδευσης.

Πολλοί αλγόριθμοι επικεντρώνονται σε προβλήματα προσέγγισης συναρτήσεων, στα οποία η εργασία ενσωματώνεται σε μια συνάρτηση (π.χ. έχοντας ως είσοδο μία αίτηση χρηματοπιστωτικού προϊόντος, έχει ως έξοδο την αθέτηση ή μη-αθέτηση) και το πρόβλημα μάθησης είναι να βελτιωθεί η ακρίβεια αυτής της συνάρτησης, με εμπειρία που αποτελείται από ένα δείγμα γνωστών ζευγών εισόδου-εξόδου της συνάρτησης (παλαιότερων δεδομένων). Σε ορισμένες περιπτώσεις, η συνάρτηση εκφράζεται σε παραμετρική μορφή ενώ σε άλλες, η συνάρτηση είναι σιωπηρή και

δημιουργείται μέσω μιας διαδικασίας αναζήτησης, παραγοντοποίησης, βελτιστοποίησης ή προσομοίωσης. Παρόλο που η συνάρτηση είναι σιωπηρή, συνήθως εξαρτάται από παραμέτρους ή άλλους βαθμούς ελευθερίας και η εκπαίδευση είναι η διαδικασία προσδιορισμού των καλύτερων τιμών για αυτές τις παραμέτρους για τη βελτίωση της μέτρησης απόδοσης.

Κύριο μέλημα είναι η ακρίβεια με την οποία ο αλγόριθμος μπορεί να να εκπαιδευτεί από ένα συγκεκριμένο τύπο και όγκο δεδομένων καθώς και πόσο καλά ανταποκρίνεται και εντοπίζει σφάλματα. Τέτοιοι θεωρητικοί χαρακτηρισμοί αλγορίθμων και προβλημάτων μηχανικής μάθησης συνήθως ακολουθούν τα πλαίσια της θεωρίας στατιστικών αποφάσεων και της θεωρίας υπολογιστικής πολυπλοκότητας.

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να διαχωριστούν σε δύο κατηγορίες, την εποπτευόμενη μάθηση (supervised learning) και μη εποπτευόμενη μάθηση (unsupervised learning). Η εποπτευόμενη μάθηση χαρακτηρίζεται από την χρήση δεδομένων με ετικέτα. Αυτά τα σύνολα δεδομένων έχουν σχεδιαστεί για να εκπαιδεύουν ή να «εποπτεύουν» αλγόριθμους για την ταξινόμηση δεδομένων ή την ακριβή πρόβλεψη των αποτελεσμάτων (Alpaydin, 2020). Χρησιμοποιώντας εισόδους και εξόδους με ετικέτα, το μοντέλο μπορεί να μετρήσει την ακρίβειά του και να βελτιωθεί με την πάροδο του χρόνου.

Η εποπτευόμενη μάθηση μπορεί να χωριστεί σε δύο τύπους προβλημάτων κατά την εξόρυξη δεδομένων: ταξινόμηση (classification) και παλινδρόμηση (regression):

- Τα προβλήματα ταξινόμησης χρησιμοποιούν έναν αλγόριθμο για την ακριβή αντιστοίχιση των δεδομένων δοκιμής (test sets) σε συγκεκριμένες κατηγορίες, όπως ο διαχωρισμός των μήλων από τα πορτοκάλια. Ή, στον πραγματικό κόσμο, οι εποπτευόμενοι αλγόριθμοι εκμάθησης μπορούν να χρησιμοποιηθούν για την ταξινόμηση των ανεπιθύμητων μηνυμάτων σε ξεχωριστό φάκελο από τα εισερχόμενά σας. Οι γραμμικοί ταξινομητές, οι μηχανές διανυσμάτων υποστήριξης (support vector machine), τα δέντρα αποφάσεων και το τυχαίο δάσος (random forest) είναι όλοι οι συνήθεις τύποι αλγορίθμων ταξινόμησης.

- Η παλινδρόμηση είναι ένας άλλος τύπος εποπτευόμενης μεθόδου μάθησης που χρησιμοποιεί έναν αλγόριθμο για την κατανόηση της σχέσης μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Τα μοντέλα παλινδρόμησης είναι χρήσιμα για την πρόβλεψη αριθμητικών τιμών με βάση διαφορετικά σημεία δεδομένων, όπως οι προβλέψεις εσόδων από δεδομένα πωλήσεων για μια δεδομένη επιχείρηση. Μερικοί δημοφιλείς αλγόριθμοι παλινδρόμησης είναι η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση και η πολυωνυμική παλινδρόμηση.

Η μη εποπτευόμενη μάθηση χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για την ανάλυση και τη ομαδοποίηση συνόλων δεδομένων χωρίς ετικέτα. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυφά μοτίβα σε δεδομένα χωρίς την ανάγκη ανθρώπινης παρέμβασης (επομένως, είναι «χωρίς

επίβλεψη»).

Τα μοντέλα μάθησης χωρίς επίβλεψη χρησιμοποιούνται για τρεις κύριες εργασίες: ομαδοποίηση, συσχέτιση και μείωση διαστάσεων (Ghahramani, 2003):

- Η ομαδοποίηση είναι μια τεχνική εξόρυξης δεδομένων για την ομαδοποίηση δεδομένων χωρίς ετικέτα με βάση τις ομοιότητες ή τις διαφορές τους. Για παράδειγμα, οι αλγόριθμοι ομαδοποίησης K-μέσων εκχωρούν παρόμοια σημεία δεδομένων σε ομάδες, όπου η τιμή K αντιπροσωπεύει το μέγεθος της ομαδοποίησης και την ευαισθησία. Αυτή η τεχνική είναι χρήσιμη για την κατάτμηση της αγοράς, τη συμπίεση εικόνας κ.λπ.

- Η συσχέτιση είναι ένας άλλος τύπος μεθόδου μάθησης χωρίς επίβλεψη που χρησιμοποιεί διαφορετικούς κανόνες για να βρει σχέσεις μεταξύ μεταβλητών σε ένα δεδομένο σύνολο δεδομένων. Αυτές οι μέθοδοι χρησιμοποιούνται συχνά για κινητήρες ανάλυσης καλαθιού αγοράς και συστάσεων.

- Η μείωση διαστάσεων είναι μια τεχνική εκμάθησης που χρησιμοποιείται όταν ο αριθμός των χαρακτηριστικών (ή διαστάσεων) σε ένα δεδομένο σύνολο δεδομένων είναι πολύ υψηλός. Μειώνει τον αριθμό των εισαγωγών δεδομένων σε ένα διαχειρίσιμο μέγεθος, ενώ παράλληλα διατηρεί την ακεραιότητα των δεδομένων. Συχνά, αυτή η τεχνική χρησιμοποιείται στο στάδιο της προεπεξεργασίας δεδομένων, όπως όταν οι αυτόματες κωδικοποιητές αφαιρούν το θόρυβο από τα οπτικά δεδομένα για να βελτιώσουν την ποιότητα της εικόνας.

Η κύρια διάκριση μεταξύ των δύο προσεγγίσεων είναι η χρήση συνόλων δεδομένων με ετικέτα. Με απλά λόγια, η εποπτευόμενη μάθηση χρησιμοποιεί δεδομένα εισόδου και εξόδου με ετικέτα, ενώ ένας αλγόριθμος μάθησης χωρίς επίβλεψη δεν το κάνει.

Στην εποπτευόμενη μάθηση, ο αλγόριθμος «μαθαίνει» από το σύνολο δεδομένων εκπαίδευσης κάνοντας επαναληπτικά προβλέψεις για τα δεδομένα και προσαρμόζοντας τη σωστή απάντηση. Ενώ τα μοντέλα εποπτευόμενης μάθησης τείνουν να είναι πιο ακριβή από τα μοντέλα μάθησης χωρίς επίβλεψη, απαιτούν άμεση ανθρώπινη παρέμβαση για την κατάλληλη επισήμανση των δεδομένων. Για παράδειγμα, ένα εποπτευόμενο μοντέλο εκμάθησης μπορεί να προβλέψει πόσο θα διαρκέσει η διαδρομή σας με βάση την ώρα της ημέρας, τις καιρικές συνθήκες και ούτω καθεξής. Αλλά πρώτα, θα πρέπει εκπαιδευτεί ώστε να γνωρίζει ότι ο βροχερός καιρός παρατείνει τον χρόνο οδήγησης.

Τα μοντέλα μάθησης χωρίς επίβλεψη, αντίθετα, λειτουργούν από μόνα τους για να ανακαλύψουν την εγγενή δομή των δεδομένων χωρίς ετικέτα. Να σημειωθεί ότι εξακολουθούν να απαιτούν κάποια ανθρώπινη παρέμβαση για την επικύρωση των μεταβλητών εξόδου. Για παράδειγμα, ένα μοντέλο μάθησης χωρίς επίβλεψη μπορεί να προσδιορίσει ότι οι διαδικτυακοί αγοραστές συχνά αγοράζουν ομάδες συγκεκριμένων προϊόντων ταυτόχρονα. Ωστόσο, για παράδειγμα,

ένας αναλυτής δεδομένων θα πρέπει να επιβεβαιώσει ότι είναι λογικό μια μηχανή συστάσεων να ομαδοποιεί τα βρεφικά ρούχα με μια παραγγελία από πάνες, σάλτσα μήλου και φλιτζάνια.

Σε αυτήν την εργασία θα γίνει χρήση μοντέλων εποπτευόμενης μάθησης.

1.4.1 Προβλήματα και Κίνδυνοι

Όπως βλέπουμε οι αλγόριθμοι μηχανικής μάθησης και γενικότερα η χρήση τεχνητής νοημοσύνης έχει αυξηθεί εκθετικά τα τελευταία χρόνια καθώς δίνει λύσεις σε πολλά προβλήματα. Παρόλα αυτά υπάρχουν και κίνδυνοι που ελλοχεύουν. Πιο συγκεκριμένα, με την χρήση των smartphones η συλλογή δεδομένων έχει γίνει πιο εύκολη από ποτέ και αρχίζουν να τίθενται ζητήματα παραβίασης δεδομένων. Χαρακτηριστικό παράδειγμα αποτελεί το σκάνδαλο που προέκυψε με την εταιρία της Facebook η οποία συνέλεξε εκατομμύρια δεδομένα χρηστών χωρίς την ενημέρωσή τους και τα πούλησε σε εταιρία πολιτικών διαφημίσεων (Cadwalladr & Graham-Harrison, 2018). Για να αντιμετωπιστούν τέτοιου είδους ζητήματα, η Ευρωπαϊκή Ένωση εισήγαγε το 2018 τον Γενικό Κανονισμό για την Προστασία Δεδομένων (GDPR) που είναι ο πιο σκληρός νόμος περί απορρήτου και ασφάλειας στον κόσμο. Επιβάλλει υποχρεώσεις σε οργανισμούς οπουδήποτε στον κόσμο, εφόσον στοχεύουν ή συλλέγουν δεδομένα που σχετίζονται με άτομα στην Ε.Ε. Ο GDPR επιβάλει σκληρά πρόστιμα σε όσους παραβιάζουν τα πρότυπα απορρήτου και ασφάλειας, με κυρώσεις που φτάνουν τα δεκάδες εκατομμύρια ευρώ (Albrecht, 2016).

Ένα άλλο ζήτημα που έχει προκύψει είναι το περιβαλλοντικό αποτύπωμα αυτών των διαδικασιών. Πιο συγκεκριμένα, η αποθήκευση δεδομένων στο cloud καθώς και η εκπαίδευση των μοντέλων είναι πολύ ενεργοβόρες διαδικασίες. Για παράδειγμα, το αποτύπωμα άνθρακα της εκπαίδευσης ενός μοντέλου μεγάλης κλίμακας είναι ίσο με περίπου 300.000 kg εκπομπών διοξειδίου του άνθρακα (Lu, 2019). Αυτή είναι της τάξης των 125 πτήσεων μετ' επιστροφής μεταξύ Νέας Υόρκης και Πεκίνου.

1.5 Το κατάλληλο μοντέλο

Όπως αναφέρθηκε πλέον χρησιμοποιούνται διάφορες τεχνικές μηχανικής μάθησης στην πιστωτική βαθμολόγηση με μία από αυτές να είναι η προσαρμογή μοντέλων παλινδρόμησης. Με πολύ απλά λόγια, δεδομένα από προηγούμενους πελάτες "τρέχουν" σε μοντέλα με σκοπό να εξαχθούν συμπεράσματα για μελλοντικούς πελάτες.

Ως «μοντέλο» («model») ορίζεται η μορφή της σχέσης ανάμεσα σε δύο ή και περισσότερες μεταβλητές. Ο βασικός σκοπός της μοντελοποίησης είναι η παραγωγή μίας μαθηματικής αναπαράστασης της σχέσης ανάμεσα σε μία μεταβλητή απόκρισης και ένα πλήθος επεξηγηματικών μεταβλητών, μαζί με ένα μέτρο της αντίστοιχης αβεβαιότητας για την σχέση αυτή (Mohri et al.,

2018).

Βασικό βήμα είναι να προσδιορισθεί η εξαρτημένη μεταβλητή του μοντέλου καθώς και η εύρεση δεδομένων για το χτίσιμο του μοντέλου. Στο πρόβλημα της πιστωτικής βαθμολόγησης ορίζεται η εξαρτημένη μεταβλητή να κάνει διάκριση μεταξύ πραγματοποίησης συμβάντων και μη-πραγματοποίησης συμβάντων (0-1 δίτιμη εξαρτημένη μεταβλητή). Γίνεται εστίαση στην πρόβλεψη αθέτησης.

Στην πιστωτική βαθμολόγηση, η αθέτηση χρησιμοποιείται για να περιγράψει την εξαρτημένη μεταβλητή. Η αβεβαιότητα σχετικά με την ικανότητα ενός δανειολήπτη να εξυπηρετήσει τα χρέη ή τις δεσμεύσεις του είναι γνωστή ως κίνδυνος αθέτησης. Ποσοτικοποιείται με τον υπολογισμό της πιθανότητας αθέτησης υποχρεώσεων. Η πιθανότητα αθέτησης αντικατοπτρίζει την πιθανολογική εκτίμηση της πιθανότητας ο οφειλέτης ή ο αντισυμβαλλόμενος να αθετήσει τη συμβατική του υποχρέωση εντός ορισμένης χρονικής περιόδου.

Συνεπώς, σε ένα πρόβλημα βαθμολόγησης πιστοληπτικής ικανότητας, ο ορισμός μιας εξαρτημένης μεταβλητής είναι διπλός: ένας όριο παραβατικότητας και μια χρονική περίοδος κατά την οποία δεν ξεπερνιέται αυτό το επίπεδο παραβατικότητας, το οποίο αναφέρεται ως περίοδος έκβασης. Ως εκ τούτου, η περίοδος έκβασης είναι η χρονική περίοδος κατά την οποία τα δάνεια του δείγματος παρατηρούνται για να ταξινομηθούν ως καλά ή κακά.

Η πιθανότητα που περιγράψαμε παραπάνω ορίζεται ως εξής:

$$p(y|\theta) \tag{1.5.1}$$

$$y = \begin{cases} 0, \text{ εάν αποπληρωθεί το δάνειο (αποπλήρωση)} \\ 1, \text{ εάν δεν αποπληρωθεί το δάνειο (αθέτηση)} \end{cases}$$

Όπου ϑ οι παράμετροι που οδηγούν στην πρόβλεψη της πιθανότητας.

Για τον υπολογισμό της πιθανότητας αθέτησης αρκετά συχνά γίνεται χρήση μοντέλων παλινδρόμησης. Τα μοντέλα παλινδρόμησης εξετάζουν πώς μια συγκεκριμένη μεταβλητή (η εξαρτημένη μεταβλητή) εξηγείται από μια άλλη μεταβλητή - ή, πιο τυπικά, από ένα ολόκληρο σύνολο άλλων μεταβλητών (επεξηγηματικές μεταβλητές). Το αποτέλεσμα που λαμβάνουμε από ένα μοντέλο παλινδρόμησης καθορίζεται από ένα σύνολο παραγόντων που ονομάζονται συντελεστές παλινδρόμησης (regression coefficients). Καθένα από αυτούς τους συντελεστές μπορεί να ερμηνευθεί ως η συσχέτιση μεταξύ της εξαρτημένης μεταβλητής που προσπαθεί κανείς να προβλέψει και της επεξηγηματικής μεταβλητής, διατηρώντας σταθερές όλες τις άλλες επιρροές στην επεξ-

ηγηματική μεταβλητή.

Ανεξάρτητα από την τεχνική που χρησιμοποιείται για την προσαρμογή του μοντέλου, θα πρέπει να τεθούν ορισμένα κριτήρια για τον προσδιορισμό της καλής προσαρμογής και της ικανότητας πρόβλεψης του μοντέλου.

Για να είναι χρήσιμο ένα μοντέλο πιστωτικής βαθμολόγησης, πρέπει να είναι εφαρμόσιμο στον γενικό πληθυσμό και όχι μόνο στο δείγμα ανάπτυξης. Πρέπει να ληφθεί μέριμνα ώστε να μην προσαρμόζεται υπερβολικά το μοντέλο στα δεδομένα που χρησιμοποιήθηκαν για την ανάπτυξή του. Ουσιαστικά χρησιμοποιείται ένα δείγμα ανάπτυξης που ονομάζεται training set και αφότου αναπτυχθεί το μοντέλο, εφαρμόζεται ένα δείγμα ελέγχου ή αλλιώς test set.

Αφότου έχει σχεδιαστεί και εφαρμοστεί το εκάστοτε μοντέλο είναι κρίσιμο να παρακολουθείτε το μοντέλο σε τακτά χρονικά διαστήματα. Είναι ιδιαίτερα ζωτικής σημασίας σε μια αναπτυσσόμενη οικονομία να διασφαλιστεί ότι το μοντέλο εξακολουθεί να προβλέπει με την ίδια ικανότητα και ότι ο πληθυσμός δεν έχει αλλάξει. Εάν ο πληθυσμός έχει αλλάξει, αυτό δεν σημαίνει πάντα ότι το μοντέλο δεν προβλέπει πλέον. Μπορεί απλώς να απαιτούνται μερικές τροποποιήσεις. Η προγνωστική ισχύς του μοντέλου παρακολουθείται επίσης, υποδεικνύοντας πότε πέφτει κάτω από τα αποδεκτά όρια και πότε πρέπει να ενημερωθεί.

1.6 Δεδομένα (Data)

Για να δημιουργηθεί και να δοκιμαστεί ένα μοντέλο σε ένα πρόβλημα, πρέπει να ευρεθούν και να ελεγχθούν οι κατάλληλες παράμετροι ϑ, όπως αναφέρθηκε και προηγουμένως. Αυτές οι παράμετροι μπορεί να είναι διαφορετικές κάθε φορά και για κάθε μοντέλο που εφαρμόζεται. Για παράδειγμα οι παράμετροι (ή χαρακτηριστικά) ενός πελάτη μπορεί να είναι το ιστορικό αποπλήρωσης κάποιου άλλου δανείου, η επαγγελματική κατάσταση του πελάτη, το ετήσιο εισόδημά του και πολλά άλλα. Είναι προφανές λοιπόν πως χρειάζεται ένας όγκος δεδομένων (dataset) από προηγούμενους πελάτες.

Η επιλογή δεδομένων είναι το πιο σημαντικό βήμα στη διαδικασία ανάπτυξης του μοντέλου. Αυτό είναι επίσης συνήθως το βήμα που απαιτεί τον περισσότερο χρόνο και προσπάθεια. Το να υπάρχουν καθαρά, ακριβή και κατάλληλα δεδομένα είναι εξαιρετικά σπουδαίο.

1.6.1 Ποιότητα των Δεδομένων (Data Quality)

Το πόσο καλή θα είναι η πρόβλεψη της διάκρισης μεταξύ καλών και κακών πελατών που κάνουν αίτηση για δάνειο εξαρτάται κατά πολύ από τα δεδομένα που θα χρησιμοποιηθούν για να εξαχθούν συμπεράσματα. Το πρώτο ζήτημα που πρέπει να αντιμετωπιστεί είναι η ποιότητα

των διαθέσιμων δεδομένων. Στη βιβλιογραφία η ποιότητα των δεδομένων ορίζεται από πολλά χαρακτηριστικά, όπως η ακρίβεια, η πληρότητα και η συνέπεια (Baesens et al., 2009).

Η ακρίβεια των δεδομένων σχετίζεται με τον βαθμό ακρίβειας των μετρήσεων ενός χαρακτηριστικού στην πραγματική του τιμή. Συνήθη παραδείγματα που αναφέρονται στις τυπικές αιτίες κακής ακρίβειας δεδομένων είναι τα σφάλματα εισαγωγής χρήστη και τα σφάλματα στο λογισμικό και τον κώδικα επεξεργασίας.

Η πληρότητα δεδομένων αναφέρεται στον βαθμό στον οποίο λείπουν οι τιμές από τα δεδομένα, τα λεγόμενα missing values. Δηλαδή ένα σύνολο δεδομένων στο οποίο λείπουν πολλές τιμές για μία σημαντική παράμετρο δεν είναι επιθυμητό. Ένας τρόπος αντιμετώπισης είναι να αφαιρεθούν αυτά τα δεδομένα από το σύνολο.

Η συνέπεια των δεδομένων σχετίζεται με καταστάσεις στις οποίες χρησιμοποιούνται πολλαπλές πηγές δεδομένων και λόγω έλλειψης τυποποίησης, δύο ή περισσότερα στοιχεία δεδομένων ενδέχεται να έρχονται σε σύγκρουση μεταξύ τους.

Πρώτο βήμα λοιπόν είναι να αποφασισθεί ποιες μεταβλητές έχουν ενδιαφέρον και μπορούν να αξιοποιηθούν στο μοντέλο προσφέροντας κάποια αξία στην πρόβλεψη.

1.6.2 Ποσότητα των Δεδομένων (Data Quantity)

Για να διασφαλιστεί η κατασκευή ενός μοντέλου υψηλής ποιότητας, απαιτείται επαρκής ποσότητα δεδομένων πελατών. Κατά την διάρκεια αυτής της φάσης αναζητούνται αξιόπιστες πηγές δεδομένων και δημιουργείται ένα πλάνο για την χρήση τους. Ο όγκος των δεδομένων που θα χρησιμοποιήσουμε εξαρτάται από τον στόχο της έρευνας και τις ιδιότητες των δεδομένων. Πλέον, με την ανάπτυξη των υπολογιστών, η ανεύρεση δεδομένων δεν είναι χρονοβόρα ούτε δαπανηρή διαδικασία. Στόχος είναι τα δεδομένα να προέρχονται από ποικίλες πηγές και όχι μόνο από μία. Για παράδειγμα, για την ανάπτυξη του μοντέλου της πιστωτικής βαθμολόγησης, μπορούν να χρησιμοποιηθούν δεδομένα για τους πελάτες από τις ακόλουθες πηγές (Lewis, 1992):

- Εσωτερικά - Αυτός ο τύπος δεδομένων περιγράφει λεπτομερώς προηγούμενες συναλλαγές πελατών και συμπεριφορών σχετικά με τον λογαριασμό τους. Για παράδειγμα, υπόλοιπο λογαριασμού, έτη ως πελάτης και υφιστάμενα δάνεια στην τράπεζα είναι πληροφορίες που αποθηκεύονται εσωτερικά από την κάθε τράπεζα.
- Εξωτερικά - Αυτός ο τύπος δεδομένων λαμβάνεται από έντυπα αιτήσεων και φορολογικές δηλώσεις. Παραδείγματα τέτοιων πληροφοριών περιλαμβάνουν: αριθμός εξαρτώμενων μελών, αριθμός ετών στην τρέχουσα διεύθυνση και εισόδημα.

Σε κάποιες χώρες υπάρχουν και δεδομένα που μπορούν να ανακτηθούν από δικαστικά αρχεία και συλλέγονται από πιστωτικά γραφεία. Τα πιστωτικά γραφεία είναι ιδρύματα που συλλέγουν

δεδομένα σχετικά με την απόδοση των χορηγούμενων δανείων από διαφορετικούς δανειστές. Όπως αναφέρθηκε προηγουμένως, ο όγκος του δείγματος πρέπει να είναι μεγάλος. Αυτό διότι το δείγμα πρέπει παρουσιάζει ίδιες ιδιότητες με τον πληθυσμό ενδιαφέροντος. Επιπλέον, τα δεδομένα πρέπει να περιλαμβάνουν πολλές περιπτώσεις ώστε να μην υπάρχουν μεταβλητές που συσχετίζονται και μπορεί να οδηγήσουν σε υπερβολική προσαρμογή του μοντέλου (overfitting).

1.6.3 Καθαρισμός Δεδομένων (Data Cleansing)

Αναφέρθηκε προηγουμένως ότι είναι σημαντική η ποσότητα και η ποιότητα των δεδομένων καθώς και το πόσο βασικό είναι να υπάρχει είναι αξιόπιστη η πηγή που εξάγονται τα δεδομένα. Παρόλα αυτά πολλές φορές όσο καλές και να είναι πηγές πληροφορίας, μπορεί να υπάρχουν σφάλματα (errors). Για προϋπάρχοντα σύνολα δεδομένων (datasets) είναι λογικό να γίνεται προσπάθεια για τον καθαρισμό τους. Στην πράξη αυτό είναι εξερεύνηση του σύνολου δεδομένων για πιθανά προβλήματα και προσπάθεια διόρθωσης των σφαλμάτων. Αυτό όμως δεν μπορεί να γίνει "με το χέρι" λόγω του χρόνου που θα χρειαζότανε κάτι τέτοιο. Τα τελευταία χρόνια γίνεται μια προσπάθεια για να αυτοματοποιηθεί αυτή η διαδικασία η οποία ονομάζεται Data Cleansing (Maletic & Marcus, 2000). Η αύξηση της υπολογιστικής δύναμης καθιστά κάτι τέτοιο εφικτό. Τα προβλήματα που χρήζουν αντιμετώπισης μπορεί να είναι δεδομένα που λείπουν (missing data), λανθασμένα δεδομένα κ.α.. Κάθε πρόβλημα έχει διαφορετική αντιμετώπιση αλλά κατά γενική ομολογία η διαδικασία έχει τρία στάδια:

- Ορισμός και προσδιορισμός τύπων σφαλμάτων
- Αναζήτηση και αναγνώριση περιπτώσεων σφαλμάτων
- Διόρθωση των σφαλμάτων

Στην συγκεκριμένη εργασία θα θεωρηθεί πως έχει γίνει ήδη αυτή η διαδικασία στα δεδομένα που έχουμε.

Κεφάλαιο 2

Τεχνικές και Μέθοδοι Αξιολόγησης της Πιστωτικής Βαθμολόγησης

Σε αυτό το κεφάλαιο θα γίνει ανάλυση των διαφόρων τεχνικών μηχανικής μάθησης που μπορούν να εφαρμοσθούν στο πρόβλημα της πιστωτικής βαθμολόγησης. Οι μέθοδοι που χρησιμοποιούνται γενικά στην πιστωτική βαθμολόγηση βασίζονται σε τεχνικές στατιστικής αναγνώρισης προτύπων. Ιστορικά, η διακριτική ανάλυση και η γραμμική παλινδρόμηση ήταν οι πιο ευρέως χρησιμοποιούμενες τεχνικές στην πιστωτική βαθμολόγηση. Και οι δύο έχουν το πλεονέκτημα ότι είναι εννοιολογικά απλές και ευρέως διαθέσιμες σε στατιστικά πακέτα λογισμικού. Η λογιστική παλινδρόμηση είναι τώρα πιθανώς η πιο χρησιμοποιούμενη τεχνική για τη βαθμολόγηση της πιστοληπτικής ικανότητας.

2.1 Λογιστική Παλινδρόμηση

Ένα μοντέλο που χρησιμοποιείται συχνά είναι αυτό της λογιστικής παλινδρόμησης και είναι και το πιο σύνηθες από την κατηγορία των γενικευμένων γραμμικών μοντέλων. Χρησιμοποιείται σε περιπτώσεις όπου η εξαρτημένη μεταβλητή του μοντέλου είναι κατηγορική και όχι ποσοτική και συνήθως παίρνει δύο μόνο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Για παράδειγμα, αν το εμβόλιο θα λειτουργήσει ή όχι, αν ο ασθενής θα επιζήσει ή όχι ή, όπως στην περίπτωση της πιστωτικής βαθμολόγησης, αν ο πελάτης θα αθετήσει το δάνειο εντός ενός χρονικού πλαισίου ή όχι. Συνήθως χρησιμοποιείται η κωδικοποίηση 0 και 1 για τα δύο ενδεχόμενα. Το μοντέλο της λογιστικής παλινδρόμησης μοντελοποιεί την πιθανότητα η εξαρτημένη μεταβλητή y να ανήκει σε μία συγκεκριμένη κατηγορία (Hosmer et al., 2013).

Θεωρούμε λοιπόν πως η δίτιμη τυχαία μεταβλητή (τ.μ) y πραγματοποιεί το ενδεχόμενο y=1 («επιτυχία» ή στην περίπτωσή μας αθέτηση) με πιθανότητα P=p και αντίστοιχα πραγματοποιεί

το ενδεχόμενο y=0 («αποτυχία» ή μη-αθέτηση) με πιθανότητα P=1-p. Προφανώς εφόσον μιλάμε για πιθανότητες, το $p \in [0, 1]$. Εστιάζουμε το ενδιαφέρον μας σε ένα από τα δύο ενδεχόμενα και πιο συγκεκριμένα στην επιτυχία. Για το πρόβλημα της πιστωτικής βαθμολόγησης η y περιγράφεται από την σχέση 1.5.1. Γνωρίζουμε ότι η y είναι τ.μ. της κατανομής Bernoulli, δηλαδή $y \sim B(p)$ και έχει σ.π.π.:

$$f(y) = p^y(1-p)^{1-y}$$

με μέση τιμή $E(y) = p$ και διασπορά $V(y) = p(1-p)$. Έχοντας πάντα στην διάθεση μας ένα δείγμα ποσότητας n (δηλαδή πραγματοποιήσεων ενδεχομένων), ορίζουμε την τ.μ. y=αριθμός επιτυχιών σε n δοκιμές. Υποθέτοντας ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε η y ακολουθεί την διωνυμική κατανομή δηλαδή

$$y \sim b(n, p)$$

με συνάρτηση πιθανότητας:

$$f(y) = \binom{n}{y} p^y(1-p)^{n-y}, \quad y = 0, 1, ..., n$$

όπου η πιθανότητα επιτυχίας p είναι παράμετρος της κατανομής. Η μέση τιμή της κατανομής είναι $E(y) = np$ και διασπορά της είναι $V(y) = np(1-p)$. Σε περίπτωση που n=1, μιλάμε για δυαδικά δεδομένα, αλλιώς για διωνυμικά.

Επειδή συνήθως η τ.μ. y εξαρτάται από κάποιες x επεξηγηματικές μεταβλητές, θέλουμε ένα μοντέλο το οποίο να μπορεί να λάβει υπόψιν αυτές τις συμμεταβλητές, και οι τιμές της τ.μ. y αυτού του μοντέλου να ανήκουν στο διάστημα [0, 1], αφού είναι πιθανότητες. Η σύνδεση της y από τις επεξηγηματικές μεταβλητές εισάγεται μέσω της εξάρτησης της πιθανότητας p από τις x. Για να επιτευχθεί αυτό, κατασκευάζεται το μοντέλο της λογιστικής παλινδρόμησης που ανήκει στην οικογένεια των γενικευμένων γραμμικών μοντέλων και εκφράζεται μέσω της σχέσης (Καρώνη & Οικονόμου, 2017):

$$\eta_x = g(E(y_x)) = g(\mu_x) = x'\beta$$

με την ακόλουθη δομή:

1. $y_x \sim b(n_x, \mu_x)$ ($n_x > 1$, δυωνυμικά δεδομένα)
   ή $y_x \sim B(\mu_x)$ ($n_x = 1$, δυαδικά δεδομένα)
2. $\eta_x = g(\mu_x) = \ln \frac{\mu_x}{\eta_x - \mu_x} = \ln \frac{p_x}{1-p_x} = logit(p_x) = x'\beta$ (συνάρτηση logit)

3. ανεξαρτησία μεταξύ των παρατηρήσεων y_x

όπου n_x είναι ο αριθμός των επαναλήψεων της τιμής του διανύσματος x των επεξηγηματικών μεταβλητών. Αν αντιστραφεί η συνάρτηση logit (σχέση 2) προκύπτει

$$p_x = \frac{e^{\eta_x}}{1 + e^{\eta_x}} \tag{2.1.1}$$

από όπου φαίνεται ότι ισχύει ο περιορισμός $0 < p_x < 1$ καθώς πρόκειται για πιθανότητα.

Για κάθε παρατήρηση i το μοντέλο γράφεται ως

$$\ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_\kappa x_{ik}, i = 1, ..., n \tag{2.1.2}$$

όπου

$$p_i = p_{xi} = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_\kappa x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_\kappa x_{ik})} \tag{2.1.3}$$

η πιθανότητα επιτυχίας και συνεπώς

$$E(y_i) = n_i p_i = n_i \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \tag{2.1.4}$$

Η συνάρτηση logit είναι η κανονική συνάρτηση σύνδεσης (link function) της διωνυμικής κατανομής και χρησιμοποιείται συχνά. Εναλλακτικές συναρτήσεις σύνδεσης που χρησιμοποιούνται για ειδικές περιπτώσεις είναι

- $g(\mu_x) = ln[-ln(1 - p_x)] = x'\beta$ (συνάρτηση complementary log-log)
- $g(\mu_x) = \Phi^{-1}(p_x) = x'\beta$ (συνάρτηση probit)

Στο απλό γραμμικό μοντέλο, για να προσαρμοσθεί το μοντέλο στα δεδομένα χρησιμοποιείται συνήθως η μέθοδος των ελαχίστων τετραγώνων. Επειδή κάτι τέτοιο δεν είναι εφικτό στην λογιστική παλινδρόμηση, γίνεται χρήση της μεθόδου μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας L ορίζεται από την σχέση

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \tag{2.1.5}$$

Επειδή οι πιθανότητες p_i εξαρτώνται από τα β, θεωρούμε την συνάρτηση πιθανοφάνειας ως συνάρτηση αυτών. Αν την λογαριθμίσουμε προκύπτει

$$l = \ln L(\beta) = \sum_{i=1}^{n} \ln \binom{n_i}{y_i} + y_i x_i' \beta - n_i \ln (1 + e^{x_i' \beta}) \tag{2.1.6}$$

Παραγωγίζοντας προκύπτει

$$\frac{\partial \ln Ln(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - n_i p_i)x_{ij} \tag{2.1.7}$$

Εξισώνοντας την σχέση (2.1.5) με το μηδέν θα προκύψει το σύστημα

$$\sum_{i=1}^{n}(y_i - n_i \hat{p_i})x_{ij} = 0,\ j = 0, 1, ..., k \tag{2.1.8}$$

το οποίο είναι ένα $k+1$ σύστημα μη-γραμμικών εξισώσεων που επιλύεται μόνο με επαναληπτικές μεθόδους και συνήθως γίνεται χρήση της μεθόδου Newton-Raphson, από όπου και προκύπτουν οι εκτιμήσεις των $\hat{\beta}$.

2.1.1 Εκτίμηση παραμέτρων μοντέλου Λογιστικής Παλινδρόμησης

Αφότου προσαρμοσθεί το μοντέλο στα δεδομένα, θα προκύψουν κάποιες εκτιμήσεις για τους συντελεστές όπως ειπώθηκε . Το πλεονέκτημα σε σχέση με άλλα μοντέλα είναι πως αυτές οι τιμές μπορούν να ερμηνευτούν. Αυτό επιτυγχάνεται μέσω του λόγου των συμπληρωματικών ή σχετικών πιθανοτήτων (odds), δηλαδή του

$$\frac{\hat{p}}{1-\hat{p}} = e^{x'\hat{\beta}} = e^{\hat{\beta_0}+\hat{\beta_1}x_1+...+\hat{\beta_k}x_k} \tag{2.1.9}$$

όπου $x_0 \equiv 1$. Από την παραπάνω σχέση προκύπτει ότι η ποσότητα $e^{\hat{\beta_j}}$ είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται ο λόγος των συμπληρωματικών πιθανοτήτων πραγματοποίησης του γεγονότος "επιτυχία" όταν αυξηθεί κατά μία μονάδα η αντίστοιχη μεταβλητή x_j χωρίς μεταβολή των τιμών των άλλων επεξηγηματικών μεταβλητών. Αν ο εκτιμώμενος συντελεστής β_j είναι θετικός, ο εκτιμώμενος παράγοντας $e^{\hat{\beta_j}}$ είναι μεγαλύτερος από την μονάδα, επομένως το odds αυξάνει με την αύξηση της x_j. Αντίστοιχα, αν ο συντελεστής είναι αρνητικός, ο παράγοντας, άρα και το odds, θα μειωθεί.

Οι παράμετροι της παλινδρόμησης μπορούν να εκφραστούν και μέσα από τον λόγο του λόγου των συμπληρωματικών πιθανοτήτων, δηλαδή μέσα από τον λόγο του odds (odds ratio). Έστω ότι έχουμε ένα άτομο με κάποια δεδομένα δηλαδή κάποιες συμμεταβλητές x_1 και ένα δεύτερο άτομο με κάποιες συμμεταβλητές x_2. Τότε ο λόγος των odds είναι

$$\frac{\hat{p_1}}{1-\hat{p_1}} / \frac{\hat{p_2}}{1-\hat{p_2}} = \frac{odds(y=1|x_1)}{odds(y=2|x_2)} = \frac{e^{x_1'\hat{\beta}}}{e^{x_2'\hat{\beta}}} = e^{(x_1'-x_2')\hat{\beta}} \tag{2.1.10}$$

Υπάρχουν δύο κύριοι λόγοι που χρησιμοποιείται το odds ratio. Πρώτον, παρέχουν μια εκτίμηση (με διάστημα εμπιστοσύνης) για τη σχέση μεταξύ δύο δυαδικών («ναι ή όχι») μεταβλητών. Δεύτερον, μας επιτρέπουν να εξετάσουμε τις επιδράσεις άλλων μεταβλητών σε αυτή τη σχέση (Bland & Altman, 2000).

2.1.2 Ελεγχοσυναρτήσεις καλής προσαρμογής

Οι ελεγχοσυναρτήσεις καλής προσαρμογής υποδεικνύουν πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα.

Ελεγχοσυνάρτηση Deviance

Η ελεγχοσυνάρτηση Deviance είναι ένας αριθμός που μετρά την καλή προσαρμογή ενός μοντέλου λογιστικής παλινδρόμησης. Μπορεί να θεωρηθεί ως η απόσταση από την τέλεια εφαρμογή — ένα μέτρο του πόσο αποκλίνει το μοντέλο λογιστικής παλινδρόμησης από ένα ιδανικό μοντέλο που ταιριάζει απόλυτα στα δεδομένα.

Η ελεγχοσυνάρτηση Deviance κυμαίνεται από 0 έως άπειρο. Όσο μικρότερος είναι ο αριθμός τόσο καλύτερα ταιριάζει το μοντέλο στα δεδομένα του δείγματος (deviance= 0 σημαίνει ότι το μοντέλο λογιστικής παλινδρόμησης περιγράφει τέλεια τα δεδομένα). Οι υψηλότερες τιμές του deviance αντιστοιχούν σε ένα λιγότερο ακριβές μοντέλο. Όμως αυτό δεν είναι γενικός κανόνας. Δεν πρέπει πάντα να επιλέγεται το μοντέλο μικρότερο deviance διότι μπορεί να έχουμε overfitting, δηλαδή να εξηγεί πολύ καλά τα συγκεκριμένα δεδομένα, αλλά με διαφορετικά δεδομένα να μην είναι τόσο καλό (Allison, 2014). Το πρόβλημα είναι ότι το μοντέλο με περισσότερους προγνωστικούς παράγοντες θα έχει πάντα χαμηλότερο deviance από το μικρότερο μοντέλο, δηλαδή δεν μπορεί να χαθεί η ακρίβεια προσθέτοντας περισσότερες συμμεταβλητές (στη χειρότερη περίπτωση, εάν οι πρόσθετες μεταβλητές δεν ήταν καθόλου σημαντικές, το μοντέλο μπορεί πάντα να ορίσει συντελεστές τους ίσους με 0). Οπότε αν επιλέγεται το μοντέλο με το μικρότερο deviance, θα συμπεριλαμβάνει πάντα όλες τις συμμεταβλητές

Το deviance δεν προορίζεται να ερμηνευτεί από μόνο του, αντίθετα, δύναται να χρησιμοποιηθεί για να συγκριθεί το μοντέλο λογιστικής παλινδρόμησης είτε με ένα μοντέλο αναφοράς είτε ένα άλλο μοντέλο που περιλαμβάνει είτε μεγαλύτερο είτε μικρότερο υποσύνολο συμμεταβλητών.

Ουσιαστικά είναι το αντίστοιχο των ελαχίστων τετραγώνων του γραμμικού μοντέλου παλινδρόμησης. Ο τύπος του deviance είναι

$$D(\hat{\beta}) = D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{\hat{\mu}_i} + (n_i - y_i) - \ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right] \tag{2.1.11}$$

Δηλαδή συγκρίνεται το μοντέλο μας, με ένα μοντέλο χωρίς καμία συμμεταβλητή δηλαδή των εκτιμώμενων $\hat{\mu}_i$. Υπό την υπόθεση ότι το μοντέλο είναι σωστό, $D(\hat{\beta}) \sim \chi^2_{n-p}$, ασυμπτωτικά, με n να είναι ο αριθμός των ομάδων παρατηρήσεων για τις οποίες οι συμμεταβλητές λαμβάνουν τις ίδιες τιμές και p=k+1 να είναι ο αριθμός των παραμέτρων στο μοντέλο (Καρώνη & Οικονόμου, 2017).

Στην ειδική περίπτωση που έχουμε δυαδικά δεδομένα, δηλαδή $n_i = 1, \forall i$, η ελεγχοσυνάρτηση deviance δεν παρέχει πληροφορίες για την καταλληλότητα του μοντέλου. Αυτό συμβαίνει επειδή εξαρτάται μόνο από τις εκτιμώμενες τιμές $\hat{\mu}_i$ και λαμβάνει την μορφή

$$D(\hat{\beta}) = -2\sum_{i=1}^{n} [\hat{\mu}_i logit(\hat{\mu}_i) + \ln(1 - \hat{\mu}_i)] \tag{2.1.12}$$

Ελεγχοσυνάρτηση Pearson

Η ελεγχοσυνάρτηση Pearson σπανίως οδηγεί σε διαφορετικά συμπεράσματα από την deviance. Και οι δύο είναι ασυμπτωτικά ισοδύναμες και ακολουθούν την χ^2 κατανομή όταν το μοντέλο είναι ορθό. Παρόλα αυτά οι τιμές που λαμβάνουν είναι διαφορετικές. Το μειονέκτημά της σε σχέση με την deviance είναι ότι δεν μπορεί να χρησιμοποιηθεί για την σύγκριση δύο εμφωλευμένων μοντέλων. Ο τύπος της ελεγχοσυνάρτησης Pearson είναι

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1 - \hat{p}_i)} \tag{2.1.13}$$

Έλεγχος Hosmer-Lemeshow

Όπως αναφέρθηκε προηγουμένως, οι ελεγχοσυναρτήσεις Pearson και Deviance δεν είναι χρήσιμες στην περίπτωση που έχουμε δυαδικά δεδομένα. Για αυτό χρησιμοποιείται ο έλεγχος των Hosmer-Lemeshow αρχικά για μη ομαδοποιημένα δυαδικά δεδομένα. Έπειτα οι παρατηρήσεις ομαδοποιούνται σύμφωνα με τις εκτιμώμενες πιθανότητες και διατάσσονται σε αύξουσα σειρά και χωρίζονται σε ομάδες του ίδιου περίπου αριθμού παρατηρήσεων. Έστω ότι στην i-οστή από τις g συνολικά ομάδες υπάρχουν m_i παρατηρήσεις όπου o_i ο συνολικός αριθμός επιτυχιών, e_i ο αναμενόμενος αριθμός επιτυχιών και $\hat{\pi}_i = \frac{e_i}{m_i}$. Τότε ο τύπος της ελεγχοσυνάρτησης Hosmer-Lemeshow είναι

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(o_i - m_i\hat{\pi}_i)^2}{i - m_i\hat{\pi}_i(1 - \hat{\pi}_i)} \tag{2.1.14}$$

Από προσομοιώσεις, έχει βρεθεί $Q^2_{HL} \sim \chi^2_{g-2}$ ότι ασυμπτωτικώς, όμως επειδή εξαρτάται από τον χωρισμό των παρατηρήσεων σε ομάδες και από το πλήθος τους σε κάθε μία από αυτές δεν προτιμάται.

2.1.3 Υπόλοιπα

Στην λογιστική παλινδρόμηση, όπως και στη γραμμική παλινδρόμηση, τα υπόλοιπα μπορούν να οριστούν ως η διαφορά από τις παρατηρούμενες μείον τις αναμενόμενες τιμές. Είναι απαραίτητη η εξέτασή τους καθώς προσφέρει λεπτομέρειες που δεν φαίνονται με τις ελεγχοσυναρτήσεις. Χρησιμοποιούνται κυρίως για τον έλεγχο της καταλληλότητας ενός μοντέλου μέσω γραφικών παραστάσεων. Η πιο απλή μέθοδος είναι η δημιουργία ενός γραφήματος δείκτη (index plot) των διαφόρων τύπων υπολοίπων ως προς την σειρά των παρατηρήσεων στο αρχείο δεδομένων. Η παρουσία ασυνήθιστα μεγάλων υπολοίπων υποδεικνύει ότι το μοντέλο δεν είναι καλό (Καρώνη & Οικονόμου, 2017). Επιπροσθέτως, από το ίδιο γράφημα μπορεί να εξεταστεί η συσχέτιση μεταξύ των υπολοίπων σε περίπτωση που οι παρατηρήσεις είναι σε χρονική σειρά. Επιπλέον, οι γραφικές παραστάσεις των υπολοίπων έναντι κάθε συμμεταβλητής ή έναντι γραμμικού συνδυασμού συμμεταβλητών μπορούν να δώσουν χρήσιμες πληροφορίες ώστε είτε να συμπεριληφθούν νέες μεταβλητές στο μοντέλο είτε να μετασχηματιστεί μια ήδη υπάρχουσα. Υποδεικνύουν επίσης έκτροπες τιμές (outliers) στα δεδομένα.

Υπόλοιπα Pearson

Τα υπόλοιπα Pearson δίνονται από την σχέση

$$r_i^P = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}},\ i = 1, ..., n \tag{2.1.15}$$

ενώ τα τυποποιημένα υπόλοιπα Pearson δίνονται από την σχέση

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1-\hat{h}_{ii}}} \tag{2.1.16}$$

όπου $\hat{h}_{ii}$ είναι το διαγώνιο στοιχείο του $n \times n$ πίνακα

$$\hat{H} = \hat{W}^{1/2}X(X'\hat{W}X)^{-1}X'\hat{W}^{1/2} \tag{2.1.17}$$

όπου X ο $n \times p$ πίνακας σχεδιασμού και $\hat{W}$ ο $n \times n$ διαγώνιος πίνακας, του οποίου κάθε στοιχείο είναι το $n_i\hat{p}_i(1-\hat{p}_i)$, που αποτελεί την εκτιμώμενη διασπορά της απόκρισης του y_i. Επίσης, ισχύει ότι $\sum_{i=1}^n r_i^P = X^2$, με X^2 να είναι το στατιστικό ελέγχου του Pearson.

Υπόλοιπο Deviance

Τα υπόλοιπα deviance υπολογίζονται από τον τύπο

$$r_i^D = sgn(y_i - \hat{\mu}_i)\left[2y_i \ln\frac{y_i}{\hat{\mu}_i} + 2(n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right]^{1/2} \tag{2.1.18}$$

και τα τυποποιημένα υπόλοιπα ορίζονται από την σχέση

$$r_i^{DS} = \frac{r_i^D}{\sqrt{(1 - \hat{h}_{ii})}} \tag{2.1.19}$$

Ισχύει ότι $\sum_{i=1}^{n} r_i^D = D(\hat{\beta})$. Τα γραφήματα των τυποποιημένων υπολοίπων deviance σε σχέση με τις εκτιμώμενες τιμές και με βάση τη σειρά των δεδομένων χρησιμεύουν για να ελεγχθεί η υπόθεση της ανεξαρτησίας των παρατηρήσεων.

Υπόλοιπο πιθανοφάνειας

Τα υπόλοιπα πιθανοφάνειας υπολογίζονται εύκολα μέσω των υπολοίπων deviance και Pearson και ο τύπος τους είναι

$$(r_i^L)^2 = \hat{h}_{ii}(r_i^{PS})^2 + (1 - \hat{h}_{ii})(r_i^{DS})^2 \tag{2.1.20}$$

Γενικότερα η χρήση των υπολοίπων deviance προτιμάται έναντι των άλλων υπολοίπων λόγω του ότι οι τιμές του $\hat{h}_{ii}$ είναι συνήθως χαμηλές με αποτέλεσμα τα τυποποιημένα υπόλοιπα πιθανοφάνειας να μην διαφέρουν πολύ με τα τυποποιημένα υπόλοιπα Deviance.

2.1.4 Σημεία Επιρροής

Πάντα στις τεχνικές και στα μοντέλα που χρησιμοποιούμε χρειαζόμαστε δεδομένα. Αυτά τα δεδομένα εξετάζονται ως σύνολο. Όμως πολλές φορές εντοπίζονται κάποιες παρατηρήσεις στο δείγμα οι οποίες έχουν μεγάλη επιρροή στην διαμόρφωση του μοντέλου. Δηλαδή χωρίς αυτές τις παρατηρήσεις θα λαμβάναμε ένα αρκετά διαφορετικό μοντέλο οπότε είναι και απαραίτητο να εντοπιστούν. Αυτές οι παρατηρήσεις ονομάζονται σημεία ή παρατηρήσεις επιρροής (influential observations). Μπορεί να οφείλονται σε κάποιο λάθος μέτρησης.

Με τα υπόλοιπα πιθανοφάνειας μπορούμε να εντοπίσουμε τέτοια σημεία, καθώς επίσης και με γραφικές παραστάσεις των υπολοίπων deviance σε σχέση με τα $\hat{h}_{ii}$

Απόσταση του Cook

Η απόσταση του Cook είναι ένα καλό μέτρο για τον εντοπισμό σημείων επιρροής. Ειδικότερα εξετάζει κατά πόσο η αφαίρεση μιας συγκεκριμένης παρατήρησης θα επηρεάσει τις εκτιμήσεις

των παραμέτρων ενός μοντέλου (Καρώνη & Οικονόμου, 2017). Η στατιστική συνάρτηση που χρησιμοποιείται είναι

$$CD_i = \frac{1}{p}(\hat{\beta}_{(i)} - \hat{\beta})' I(\hat{\beta})(\hat{\beta}_{(i)} - \hat{\beta}),\ i = 1, ..., n \tag{2.1.21}$$

όπου $\hat{\beta}$ οι εκτιμήσεις των συντελεστών του μοντέλου όταν συμπεριλαμβάνονται όλες οι παρατηρήσεις και $\hat{\beta}_{(i)}$ οι εκτιμήσεις όταν παραλείπετε η i-οστή παρατήρηση. Επίσης $I(\hat{\beta}) = X'\hat{W}X$ είναι η παρατηρούμενη πληροφορία κατά Fisher και $\hat{V}(\hat{\beta}) = I^{-1}(\hat{\beta})$. Η σχέση 2.1.21 μπορεί να προσεγγιστεί από την πιο απλή μορφή

$$CD_i = \frac{\hat{h}_{ii}(r_i^{PS})^2}{p(1 - \hat{h}_{ii})} \tag{2.1.22}$$

Τέλος, υπάρχει και η τροποποιημένη στατιστική συνάρτηση του Cook που ορίζεται ως

$$C_i = |r_i^L| \sqrt{\frac{(n-p)\hat{h}_{ii}}{p(1 - \hat{h}_{ii})}} \tag{2.1.23}$$

και είναι καλύτερη στον εντοπισμό σημείων αυξημένης επιρροής σε σχέση με την κλασική απόσταση Cook.

2.1.5 Μέτρα Προβλεπτικής Ισχύος

Τα μέτρα προβλεπτικής ισχύος δεν πρέπει να συγχέονται με τις ελεγχοσυναρτήσεις καλής προσαρμογής. Τα μέτρα προβλεπτικής ισχύος υποδεικνύουν πόσο καλά μπορούμε να εξηγήσουμε/προβλέψουμε την εξαρτημένη μεταβλητή με βάση τις ανεξάρτητες μεταβλητές. Μπορεί πολλές φορές σε ένα μοντέλο η ελεγχοσυνάρτηση να δείχνει ότι το μοντέλο είναι καλό και ένα μέτρο προβλεπτικής ισχύος το αντίθετο. Τα μέτρα προβλεπτικής ισχύος αποτελούν σημαντικά κριτήρια για την επιλογή του βέλτιστου μοντέλου λογιστικής παλινδρόμησης.

Κριτήρια AIC και BIC

Το κριτήριο AIC (Akaike's Information Criterion) χρησιμοποιείται ως κριτήριο για την επιλογή ενός μοντέλου με τον μικρότερο δυνατό αριθμό παραμέτρων. Γενικότερα όταν γίνεται σύγκριση μοντέλων, επιλέγεται αυτό με το μικρότερο AIC. Το AIC χρησιμοποιεί την εκτίμηση μέγιστης πιθανοφάνειας ενός μοντέλου (log-likelihood) ως μέτρο προσαρμογής. Το κριτήριο BIC αποτελεί παρόμοιο κριτήριο επιλογής με το AIC με την μόνη διαφορά πως είναι αυστηρότερο

με την εισαγωγή περισσότερων μεταβλητών.Το AIC δίνεται από την σχέση

$$AIC = -2\sum_{i=1}^{n}\left[ln\left(\binom{n_i}{y_i}\right) + y_i \ln(\hat{p}_i) + (n_i - y_i)\ln(1-\hat{p}_i)\right] + 2p \tag{2.1.24}$$

ενώ το κριτήριο BIC από την σχέση

$$BIC = -2\sum_{i=1}^{n}\left[n_i ln\left(\binom{n_i}{y_i}\right) + y_i \ln(\hat{p}_i) + (n_i - y_i)\ln(1-\hat{p}_i)\right] + p\ln(n) \tag{2.1.25}$$

όπου n το πλήθος των ομάδων παρατηρήσεων για τις οποίες οι συμμεταβλητές λαμβάνουν τις ίδιες τιμές και p=k+1 είναι το πλήθος των παραμέτρων του μοντέλου. Οι μικρότερες τιμές υποδεικνύουν το καλύτερο μοντέλο.

Συντελεστές συσχέτισης

Στα γενικευμένα γραμμικά μοντέλα οι συντελεστές R^2 δεν είναι ιδιαίτερα χρήσιμοι. Παρ'όλα αυτά χρησιμοποιούνται κάποια μέτρα τέτοιου είδους. Πιο συγκεκριμένα χρησιμοποιείται ο ψευδοσυντελεστής R^2_M και δίνεται από την σχέση

$$R^2_M = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{\frac{2}{m}} \tag{2.1.26}$$

όπου $m = \sum_{i=1}^{n} n_i$, $\hat{L}_0$ η πιθανοφάνεια του μοντέλου μόνο με τον σταθερό όρο, δηλαδή χωρίς καμία συμμεταβλητή, και $\hat{L}_1$ το μοντέλο που εξετάζουμε. Επειδή ο συντελεστής αυτός δεν μπορεί να λάβει την τιμή 1 προτάθηκε από τον Nagelkerke ο διορθωμένος συντελεστής του Nagelkerke ο οποίος δίνεται από την σχέση

$$R^2_N = \frac{R^2_M}{max(R^2_M)} \tag{2.1.27}$$

Στην λογιστική παλινδρόμηση οι δύο αυτοί συντελεστές δεν λαμβάνουν μεγάλες τιμές, ειδικά όταν έχουμε δυαδικά δεδομένα. Αυτό συμβαίνει, επειδή η τιμή τους αυξάνεται καθώς προστίθενται περισσότερες παράμετροι στο μοντέλο, καθώς και επειδή το μοντέλο εξηγεί ή προβλέπει μόνο την πιθανότητα «επιτυχίας» p=E(Υ) και όχι τις ατομικές τιμές y (0 ή 1). Επομένως, μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων δεν μπορεί να εξηγηθεί, άρα ένας δείκτης τύπου R^2 λαμβάνει χαμηλή τιμή αναγκαστικώς.

Καμπύλη ROC

Στο μοντέλο της λογιστικής παλινδρόμησης έχουμε πάντα δίτιμη μεταβλητή. Οπότε για κάθε παρατήρηση το μοντέλο υπολογίζει μία εκτιμώμενη πιθανότητα $\hat{p}$ πραγματοποίησης του γεγονότος έστω Y=1. Ορίζουμε λοιπόν εμείς ένα όριο p_0 για το οποίο θα ισχύει:

- εάν $\hat{p} > p_0$, τότε προβλέπουμε Y=1
- εάν $\hat{p} \leq p_0$, τότε προβλέπουμε Y=0

Έστω ότι το ενδεχόμενο Y=1 το αποκαλούμε θετικό ενώ το Y=0 αρνητικό. Τότε μπορούν να ορισθούν οι δύο ακόλουθες ποσότητες:

- Ευαισθησία (Sensitivity): Η πιθανότητα το μοντέλο να προβλέπει ένα θετικό αποτέλεσμα για μια παρατήρηση όταν το αποτέλεσμα είναι όντως θετικό.
- Ειδικότητα (Specificity): Η πιθανότητα το μοντέλο να προβλέπει ένα αρνητικό αποτέλεσμα για μια παρατήρηση όταν το αποτέλεσμα είναι όντως αρνητικό.

Ένας εύκολος τρόπος για να απεικονίσουμε αυτές τις δύο ποσότητες είναι η δημιουργία μιας καμπύλης ROC (Receiver Operating Characteristic), η οποία είναι μια γραφική παράσταση που εμφανίζει την ευαισθησία και την ειδικότητα ενός μοντέλου λογιστικής παλινδρόμησης και υποδεικνύει την προβλεπτική ικανότητα του μοντέλου ανάλογα με το όριο p_0 που ορίζουμε (Hosmer et al., 2013). Προκύπτει λοιπόν πως κάθε ενδεχόμενο ανήκει σε μία από τις παρακάτω κατηγορίες:

- Αληθώς Θετικό (True Positive): Το μοντέλο προβλέπει θετικό αποτέλεσμα και είναι θετικό στην πραγματική τιμή είναι θετικό.
- Ψευδώς Θετικό (False Positve): Το μοντέλο προβλέπει θετικό αποτέλεσμα ενώ η πραγματική τιμή είναι αρνητικό
- Ψευδώς Αρνητικό (False Negative): Το μοντέλο προβλέπει αρνητικό αποτέλεσμα ενώ στην πραγματικότητα είναι θετικό
- Αληθώς Αρνητικό (True Negative): Το μοντέλο προβλέπει αρνητικό αποτέλεσμα ενώ στην πραγματικότητα είναι αρνητικό.

Στον πίνακα 2.1 αποτυπώνονται οι παραπάνω ποσότητες.

Η ευαισθησία εκφράζει το ποσοστό αληθώς θετικών, αντιπροσωπεύει το ποσοστό των παρατηρήσεων που προβλέπεται να είναι θετικές όταν όντως είναι θετικές και ορίζεται ως

$$Sensitivity = \frac{A\Theta}{A\Theta + \Psi A} \tag{2.1.28}$$

		Πραγματική τιμή	
		Y=1	Y=0
Πρόβλεψη	Y=1	ΑΘ	ΨΘ
	Y=0	ΨΑ	ΑΑ

Πίνακας 2.1: 2x2 Πίνακας Συνάφειας

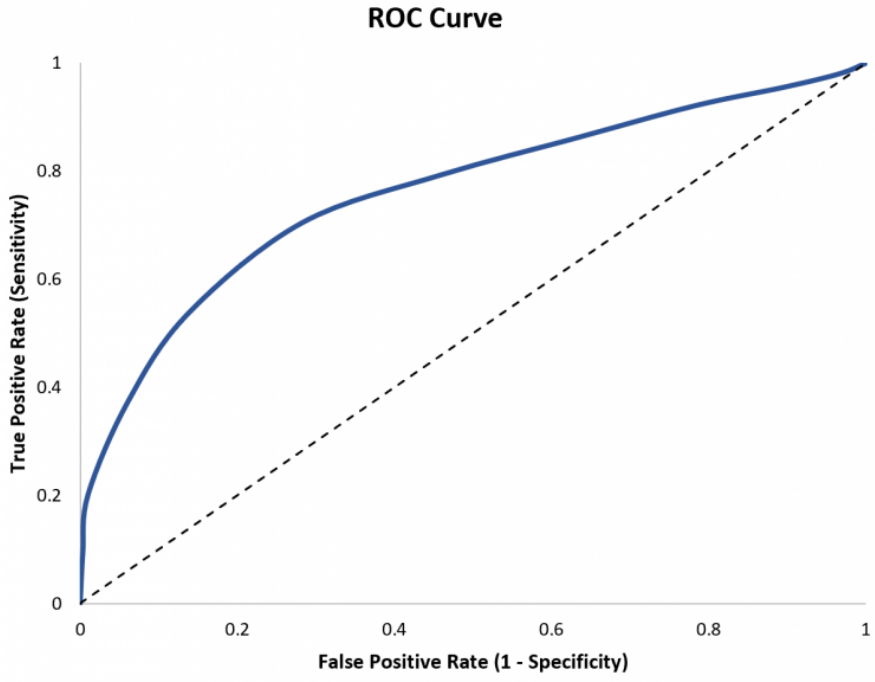
Η ειδικότητα εκφράζει το ποσοστό αληθώς αρνητικών, δηλαδή αντιπροσωπεύει το ποσοστό των παρατηρήσεων που προβλέπεται να είναι αρνητικές όταν είναι πραγματικά αρνητικές. και ορίζεται ως

$$Specificity = \frac{AA}{AA + \Psi\Theta} \tag{2.1.29}$$

Από την ειδικότητα μπορεί να ορισθεί το ποσοστό ψευδώς θετικών το οποίο αντιπροσωπεύει το ποσοστό των παρατηρήσεων που προβλέπεται να είναι θετικές όταν είναι πραγματικά αρνητικές. Δίνεται από την σχέση

$$1 - Specificity = \frac{\Psi\Theta}{\Psi\Theta + AA} \tag{2.1.30}$$

Όταν δημιουργούμε μια καμπύλη ROC, σχεδιάζουμε ζεύγη της ευαισθησίας έναντι του 1-ειδικότητα για κάθε πιθανό όριο απόφασης p_0 ενός μοντέλου λογιστικής παλινδρόμησης. Προκύπτει μια καμπύλη όπως φαίνεται στο σχήμα 2.1.1:



Σχήμα 2.1.1: Παράδειγμα Καμπύλης ROC

Η καμπύλη σχεδιάζεται στο μοναδιαίο τετράγωνο $[0,1] \times [0,1]$. Όσο περισσότερο αγκαλιάζει η καμπύλη ROC την επάνω αριστερή γωνία του γραφήματος, τόσο καλύτερα το μοντέλο ταξινομεί τα δεδομένα σε κατηγορίες. Για να ποσοτικοποιηθεί το πόσο καλά το κάνει ένα μοντέλο αυτό, υπολογίζεται το εμβαδόν κάτω από την καμπύλη και αποκαλείται AUC (Area Under Curve). Ένα μοντέλο με AUC=0,5 θα ήταν μια τέλεια διαγώνια γραμμή και θα αντιπροσώπευε ένα μοντέλο που δεν είναι καλύτερο από ένα μοντέλο που κάνει τυχαίες ταξινομήσεις (Καρώνη & Οικονόμου, 2017). Όσο πιο κοντινές τιμές λαμβάνει στο 1 το AUC, τόσο το καλύτερο. Η καμπύλης ROC είναι ιδιαίτερα χρήσιμη στην σύγκριση μοντέλων επειδή επιτρέπει να δούμε ποιο μοντέλο κάνει καλύτερες προβλέψεις.

2.2 Παλινδρόμηση Κορυφογραμμής (Ridge Regression)

Σε πολλές περιπτώσεις το μοντέλο που εξετάζεται μπορεί να υπάρχει το πρόβλημα της πολυσυγγραμικότητας. Η πολυσυγγραμμικότητα είναι η εμφάνιση υψηλής συσχέτισης μεταξύ δύο ή περισσότερων ανεξάρτητων μεταβλητών σε ένα μοντέλο πολλαπλής παλινδρόμησης. Η πολυσυγγραμμικότητα μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα όταν ένας ερευνητής ή αναλυτής προσπαθεί να προσδιορίσει πόσο καλά κάθε ανεξάρτητη μεταβλητή μπορεί να χρησιμοποιηθεί πιο αποτελεσματικά για την πρόβλεψη ή την κατανόηση της εξαρτημένης μεταβλητής σε ένα στατιστικό μοντέλο (James et al., 2013).

Ένας τρόπος για να παρακαμφθεί αυτό το ζήτημα χωρίς να αφαιρεθούν εντελώς ορισμένες μεταβλητές πρόβλεψης από το μοντέλο είναι να χρησιμοποιηθεί η μέθοδος της παλινδρόμησης κορυφογραμμής. Σε αυτήν την μέθοδο εισάγεται μία παράμετρος λ, η οποία έχει ως στόχο την μείωση της επίδρασης κάποιων συμμεταβλητών ώστε να προκύψει ένα μοντέλο με μικρότερη διασπορά.

Όταν εφαρμόζεται η μέθοδος της παλινδρόμησης κορυφογραμμής σε μοντέλο γραμμικής παλινδρόμησης, η ποσότητα που ελαχιστοποιείται είναι η

$$SSR + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.2.1}$$

όπου SSR είναι το άθροισμα τετραγώνων των αποκλίσεων $y - \hat{y}$ δηλαδή οι αποκλίσεις των τιμών του μοντέλου από τις τιμές των δεδομένων, $\lambda \geq 0$ και p το πλήθος των επεξηγηματικών μεταβλητών. Επειδή στην λογιστική παλινδρόμηση δεν εφαρμόζουμε την μέθοδο των ελαχίστων τετραγώνων για την προσαρμογή του μοντέλου, αλλά την μέθοδο της μέγιστης πιθανοφάνειας,

η σύγκριση των μοντέλων γίνεται μέσω της σχέσης

$$l^*_{ridge} = l(\beta) + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2 \tag{2.2.2}$$

όπου $l(\beta)$ το προσαρμοσμένο μοντέλο με όλες τις συμμεταβλητές και η παράμετρος λ ελέγχει το μέγεθος της συρρίκνωσης των β_j προς το μηδέν. Ο σταθερός όρος β_0 εξαιρείται από τον δεύτερο όρο με την παράμετρο λ, αποδίδοντας μια μέση προβλεπόμενη πιθανότητα ίση με το παρατηρούμενο ποσοστό συμβάντος.

2.3 Lasso Παλινδρόμηση

Η παλινδρόμηση lasso είναι άλλη μία μέθοδος συρρίκνωσης των συντελεστών β_j των συμμεταβλητών. Παρουσιάζει ομοιότητες με την παλινδρόμηση κορυφογραμμής καθώς υπάρχει και πάλι μία παράμετρος λ και μία ποσότητα που ελαχιστοποιείται η οποία είναι ελαφρώς διαφορετική και στο γραμμικό μοντέλο είναι:

$$SSR + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.3.1}$$

Το λ ελέγχει το μέγεθος της ποινής οπότε και το μέγεθος της συρρίκνωσης. Η βασική διαφορά με την παλινδρόμηση κορυφογραμμής είναι ότι στην lasso ορισμένοι συντελεστές μπορεί να μηδενιστούν και να αφαιρεθούν τελείως από το μοντέλο. Όταν $\lambda = 0$, καμία παράμετρος δεν εξαλείφεται και η εκτίμηση θα είναι ίση με αυτή που βρέθηκε στη γραμμική παλινδρόμηση. Καθώς το λ αυξάνεται, όλο και περισσότεροι συντελεστές μηδενίζονται και εξαλείφονται. Επίσης όσο μειώνεται το λ, αυξάνεται η διασπορά (James et al., 2013).

Όταν έχουμε το μοντέλο της λογιστικής παλινδρόμησης η ποσότητα που ελαχιστοποιείται είναι η

$$l + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.3.2}$$

όπου l η πιθανοφάνεια του μοντέλου με όλες τις συμμεταβλητές. Συνεπώς, λέμε ότι η lasso αποδίδει αραιά μοντέλα, δηλαδή μοντέλα που περιλαμβάνουν μόνο ένα υποσύνολο των μεταβλητών.

2.4 Ελαστική Παλινδρόμηση

Η ελαστική παλινδρόμηση εμφανίστηκε για πρώτη φορά ως αποτέλεσμα της κριτικής στην lasso, της οποίας η επιλογή μεταβλητών μπορεί να εξαρτάται πολύ από τα δεδομένα και επομένως να

είναι ασταθής. Η λύση που δίνει η ελαστική παλινδρόμηση είναι ο συνδυασμός των ποινών της παλινδρόμησης κορυφογραμμής και της lasso για να επιτευχθεί το καλύτερο αποτέλεσμα και από τις δύο μεθόδους. Η ελαστική παλινδρόμηση στοχεύει στην ελαχιστοποίηση της ακόλουθης ποσότητας:

$$SSR + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \tag{2.4.1}$$

όπου α είναι η παράμετρος ανάμειξης μεταξύ κορυφογραμμής (α=0) και lasso (α=1). Σε αυτήν την περίπτωση πρέπει να επιλεχθεί τιμή και για το λ αλλά και για το α.

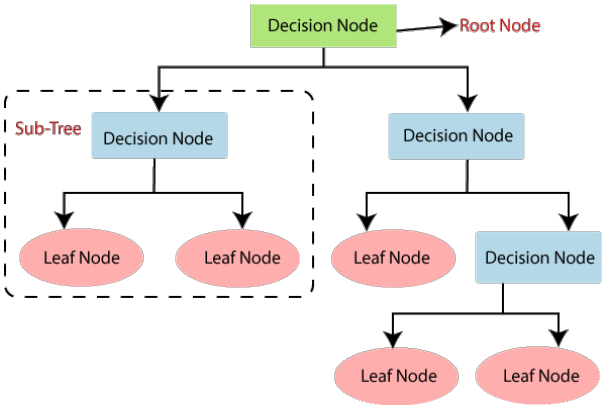
2.5 Δένδρα Απόφασης

Τα δένδρα αποφάσεων αποτελούν μία επιβλεπόμενη μέθοδο μηχανικής μάθησης η οποία χρησιμοποιείται για την επίλυση προβλημάτων παλινδρόμησης και ταξινόμησης. Ο στόχος της χρήσης ενός δέντρου αποφάσεων είναι να δημιουργηθεί ένα μοντέλο εκπαίδευσης που μπορεί να χρησιμοποιηθεί για να προβλέψει την κλάση ή την τιμή της μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που απορρέουν από προηγούμενα δεδομένα (training data).

Υπάρχουν δύο είδη δέντρων αποφάσεως. Τα δένδρα παλινδρόμησης, που έχουν ως εξαρτημένη μεταβλητή μία συνεχόμενη μεταβλητή, και τα δένδρα ταξινόμησης που έχουν ως εξαρτημένη μεταβλητή μία κατηγορική μεταβλητή. Στο πρόβλημα της πιστωτικής βαθμολόγησης είναι προφανές πως θα χρησιμοποιηθεί ένα δένδρο ταξινόμησης καθώς η εξαρτημένη μεταβλητή είναι δίτιμη κατηγορική.

Το πλεονέκτημα των δέντρων απόφασης είναι πως ερμηνεύονται εύκολα καθώς απεικονίζονται γραφικά εύκολα και κατανοητά (Kotsiantis, 2013). Αυτό οφείλεται στο ότι το δένδρο μιμείται την ανθρώπινη σκέψη κατά τη λήψη μίας απόφασης επομένως είναι εύκολα ερμηνεύσιμο. Επίσης, δύναται να διαχειρίζονται κατηγορικές μεταβλητές δίχως την ανάγκη δημιουργίας ψευδομεταβλητών. Ωστόσο, ένα μειονέκτημα είναι πως δεν παρουσιάζουν γενικώς το ίδιο επίπεδο προβλεπτικής ακρίβειας όπως άλλες αντίστοιχες μέθοδοι και, επίσης, μία μικρή μεταβολή στα δεδομένα δύναται να προκαλέσει μία μεγάλη μεταβολή στο τελικό εκτιμώμενο δένδρο. Ένα δένδρο αποφάσεων μπορεί να ενσωματώσει αριθμητικές και κατηγορικές κλάσεις. Στην περίπτωση των αριθμητικών κλάσεων, γίνεται κάποια κατηγοριοποίηση ώστε να γίνει κατηγορική. Στο σχήμα 2.5.2 βλέπουμε την μορφή ενός δένδρου απόφασης.

Στην κορυφή βρίσκεται ο κόμβος ρίζας (root node) από όπου ξεκινά το δέντρο απόφασης. Αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων, το οποίο περαιτέρω χωρίζεται σε δύο ή περισσότερα ομοιογενή σύνολα. Τα δεδομένα διαχωρίζονται όσο κατεβαίνουμε προς τα κάτω στο δένδρο. Κάθε κόμβος του δένδρου απόφασης έχει μόνο δύο παιδιά, κάτι που αποτελεί συνθήκη



Σχήμα 2.5.1: Παράδειγμα Δένδρου Απόφασης

δυαδικού δέντρου. Ο διαχωρισμός είναι η διαδικασία διαίρεσης του κόμβου απόφασης/κόμβου ρίζας σε υποκόμβους σύμφωνα με τις δεδομένες συνθήκες. Άμα τα δεδομένα ενός κόμβου κρίνεται πως δεν πρέπει να διαχωριστούν άλλο τότε λέμε πως έχουμε έναν κόμβο φύλλου (leaf node).

Κατά την εφαρμογή ενός δέντρου αποφάσεων, το κύριο ζήτημα που προκύπτει το πως θα γίνει η επιλογή για τον ριζικό κόμβο και τους υποκόμβους. Τυπικοί αλγόριθμοι μάθησης για δέντρα αποφάσεων δημιουργούν μια δομή δέντρου διαχωρίζοντας τα δεδομένα εκπαίδευσης σε όλο και μικρότερα υποσύνολα με αναδρομικό τρόπο από πάνω προς τα κάτω. Ξεκινώντας με όλα τα δεδομένα εκπαίδευσης στον ριζικό κόμβο, σε κάθε κόμβο έπειτα διαιρούνται τα δεδομένα εκπαίδευσης σε υποσύνολα. Η διαδικασία γίνεται αναδρομικά διαχωρίζοντας περαιτέρω καθένα από τα υποσύνολα (James et al., 2013). Ο διαχωρισμός συνεχίζεται έως ότου όλα τα υποσύνολα είναι "καθαρά", ή έως ότου η καθαρότητά τους δεν μπορεί να αυξηθεί περαιτέρω. Ένα υποσύνολο θεωρείται καθαρό αν περιέχει περιπτώσεις μιας μόνο κατηγορίας. Ο στόχος είναι να επιτευχθεί αυτό χρησιμοποιώντας όσο το δυνατόν λιγότερους διαχωρισμούς δηλαδή λιγότερα φύλλα έτσι ώστε το δέντρο απόφασης που προκύπτει να είναι μικρό και ο αριθμός των περιπτώσεων που υποστηρίζουν κάθε υποσύνολο να είναι μεγάλο. Για το σκοπό αυτό, έχουν σχεδιαστεί διάφορα κριτήρια διαχωρισμού επιλογής όπως το κέρδος πληροφοριών και ο δείκτης καθαρότητας Gini (Gini Impurity). Όλα παρέχουν τρόπους μέτρησης της καθαρότητας ενός διαχωρισμού.

2.5.1 Δείκτης Καθαρότητας Gini

Ο δείκτης καθαρότητας Gini χρησιμοποιείται ως οδηγός για την δημιουργία ενός δένδρου. Είναι μια συνάρτηση που καθορίζει πόσο καλά χωρίστηκε ένα δέντρο απόφασης. Ο καλύτερος διαχωρισμός αυξάνει την καθαρότητα των συνόλων που προκύπτουν από τη διάσπαση/διαχωρισμό. Αν L είναι ένα σύνολο δεδομένων με j διαφορετικές κλάσεις τότε ο τύπος για τον δείκτη Gini

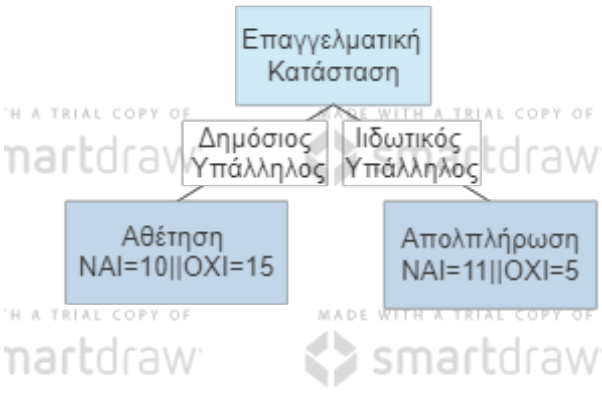
δίνεται από την σχέση

$$GINI(L) = 1 - \sum_{i=1}^{j} p_i^2 \tag{2.5.1}$$

Όπου p_i είναι σχετική συχνότητα της κλάσης i στο L. Εάν το σύνολο δεδομένων χωριστεί με βάση ένα χαρακτηριστικό i σε δύο υποσύνολα L1 και L2 με μεγέθη N1 και N2 αντίστοιχα, ο δείκτης Gini υπολογίζεται ως εξής

$$GINI_i(L) = \frac{N_1}{N} GINI_{(L_1)} + \frac{N_2}{N} GINI_{(L_2)} \tag{2.5.2}$$

Η επεξήγηση της συγκεκριμένης μεθόδου θα γίνει μέσω ενός παραδείγματος για να γίνει πιο κατανοητή. Έστω ότι θέλουμε να προβλέψουμε αν θα αθετήσει κάποιος το δάνειο του ή όχι με βάση την επαγγελματική του κατάσταση και το εισόδημά του. Το εισόδημα χωρίζεται σε δύο κατηγορίες αναφοράς (>10,000€ και <10,000€ τον χρόνο). Η επαγγελματική κατάσταση έχει και αυτή δύο κατηγορίες αναφοράς (δημόσιος υπάλληλος ή ιδιωτικός υπάλληλος). Το πρόβλημα που αντιμετωπίζουμε είναι ποία από τις δύο μεταβλητές να χρησιμοποιήσουμε αρχικά στην κορυφή του δένδρου. Δηλαδή να χωριστούν πρώτα τα δεδομένα με βάση την επαγγελματική τους κατάσταση ή με το εισόδημά τους. Για να αποφασισθεί ποια από τις δύο μεταβλητές θα χρησιμοποιηθεί, αρχικά θα υπολογισθούν και οι δύο δείκτες για τις δύο περιπτώσεις και θα επιλεχθεί αυτή με την μικρότερη τιμή. Ας εξετάσουμε τον πίνακα για να δούμε πως υπολογίζεται ο δείκτης καθαρότητας Gini.



Σχήμα 2.5.2: Παράδειγμα Δένδρου Απόφασης

Παρατηρούμε ότι σύμφωνα με αυτόν τον διαχωρισμό 10 δημόσιοι υπάλληλοι αθετήσουν και 15 όχι ενώ 11 ιδιωτικοί υπάλληλοι θα αθετήσουν ενώ 5 όχι. Υπολογίζονται πρώτα τα δείκτες

καθαρότητας Gini και για τα δύο φύλλα. Για του δημοσίου υπαλλήλου ισχύει

$$GINI_{\text{αριστερού φύλλου}} = 1 - \left(\frac{10}{10+15}\right)^2 - \left(\frac{15}{10+15}\right)^2 = 0.48 \tag{2.5.3}$$

Αντίστοιχα υπολογίζουμε και για το δεξιά φύλλο και στην συνέχεια υπολογίζεται ο σταθμισμένος δείκτης καθαρότητας GINI μέσω της σχέσης 2.5.2 . Η ίδια διαδικασία γίνεται και για το εισόδημα και όποια από τις δύο περιπτώσεις έχει μικρότερη τιμή αυτή επιλέγεται. Με αυτήν την διαδικασία χτίζεται όλο το δένδρο.

2.5.2 Κλάδεμα (Pruning)

Ένα από τα πιο συνηθισμένα προβλήματα κατά την εκμάθηση ενός δέντρου αποφάσεων είναι να βρεθεί το βέλτιστο μέγεθος του δέντρου που προκύπτει που οδηγεί σε καλύτερη ακρίβεια του μοντέλου (Mingers, 1989). Ένα δέντρο που έχει πάρα πολλά κλαδιά και στρώματα μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting) των δεδομένων εκπαίδευσης. Το κλάδεμα ενός δέντρου αποφάσεων βοηθά στην αποφυγή υπερβολικής προσαρμογής των δεδομένων εκπαίδευσης, έτσι ώστε το μοντέλο μας να γενικεύεται καλά και σε διαφορετικά δεδομένα. Το κλάδεμα ενός δέντρου αποφάσεων έχει ως στόχο την αφαίρεση ενός υποδέντρου που είναι περιττό και δεν χρησιμεύει στον διαχωρισμό των δεδομένων και στην αντικατάστασή του με ένα κόμβο φύλλου. Το κλάδεμα δέντρων απόφασης μπορεί να χωριστεί σε δύο τύπους: στο κλάδεμα στην αρχή της δημιουργίας του δένδρου και το μετα-κλάδεμα το οποίο γίνεται αφού δημιουργηθεί το δένδρο.

Το κλάδεμα στην αρχή της δημιουργίας του δένδρου είναι η μέθοδος όπου η κατασκευή του υποδένδρου διακόπτεται σε έναν συγκεκριμένο κόμβο μετά την αξιολόγηση κάποιου μέτρου. Αυτά τα μέτρα μπορεί να είναι το Gini Impurity ή το Information Gain. Στο προ-κλάδεμα, αξιολογούμε την κατάσταση κλαδέματος με βάση τα παραπάνω μέτρα σε κάθε κόμβο.

Όπως υποδηλώνει το όνομα, μετά-κλάδεμα σημαίνει το κλάδεμα μετά την κατασκευή του δέντρου. Αναπτύσσετε το δέντρο εξ ολοκλήρου χρησιμοποιώντας τον αλγόριθμο του δέντρου αποφάσεων και μετά κλαδεύονται τα υποδένδρα στο δέντρο με τρόπο από κάτω προς τα πάνω. Με βάση μέτρα όπως το Gini Impurity ή το Information Gain, αποφασίζεται εάν θα διατηρηθεί αυτός ο κόμβος απόφασης ή αν θα αντικατασταθεί με έναν κόμβο φύλλου.

2.5.3 Τυχαίο Δάσος (Random Forest)

Μία μέθοδος που χρησιμοποιεί δένδρα απόφασης είναι το τυχαίο δάσος. Το τυχαίο δάσος, όπως υποδηλώνει το όνομά του, αποτελείται από ένα μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο (Svetnik et al., 2003). Τα τυχαία δάση δημιουργούνται από υποσύνολα δεδομένων και το τελικό αποτέλεσμα βασίζεται στη μέση ή την πλειοψηφική κατάταξη

που δίνει ο αλγόριθμος και ως εκ τούτου αντιμετωπίζεται το πρόβλημα της υπερπροσαρμογής. Το training set χωρίζεται σε n υποσύνολα με σκοπό ο αλγόριθμος να εκπαιδευτεί σε κάθε υποσύνολο ξεχωριστά. Έπειτα αφού θα έχουμε n δένδρα αποφάσεων ανεξάρτητα μεταξύ τους, θα κρατάμε το αποτέλεσμα που δίνουν τα περισσότερα δένδρα, δηλαδή η πλειοψηφία. Αυτή η διαδικασία ονομάζεται bagging ή bootstrap aggregation. Το αρνητικό αυτής της μεθόδου είναι ότι χρειάζεται περισσότερη υπολογιστική δύναμη.

Κεφάλαιο 3

Ανάλυση των Δεδομένων

Σε αυτό το κεφάλαιο θα γίνει εφαρμογή των μεθόδων του κεφαλαίου 2 σε δεδομένα που αφορούν την πιστωτική βαθμολόγηση. Η ανάλυση των δεδομένων θα γίνει με την στατιστική γλώσσα προγραμματισμού R.

3.1 Παρουσίαση των Δεδομένων

Στο συγκεκριμένο κεφάλαιο θα γίνει μία προσπάθεια πρόβλεψης της πιθανότητας αθέτησης ενός πελάτη εντός μίας πενταετίας. Αυτό θα επιτευχθεί με έναν δείγμα δεδομένων μεγέθους n=366 που έχουμε στην διάθεσή μας από προηγούμενους πελάτες και με αυτό το δείγμα θα χτίσουμε τα μοντέλα. Η αθέτηση εντός πενταετίας αποτελεί την εξαρτημένη μεταβλητή y και ορίζεται ως εξής:

$$y = \begin{cases} 0, \text{Αθέτηση εντός πενταετίας=όχι} \\ 1, \text{Αθέτηση εντός πενταετίας=ναι} \end{cases} \tag{3.1.1}$$

Θεωρείται πως η προεργασία που περιγράφηκε στο κεφάλαιο 1.6 έχει γίνει καθώς επίσης δεν υπάρχουν και missing values που χρειάζονται επεξεργασία.

Οι συμμεταβλητές βάσει των οποίων θα γίνει η προσπάθεια επεξήγησης της εξαρτημένης μεταβλητής είναι η φερεγγυότητα, η σχέση περιουσίας σε σχέση με τον συνολικό δανεισμό, το ιστορικό του πελάτη και η επαγγελματική κατάσταση. Και οι τέσσερις είναι κατηγορικές μεταβλητές.

Η φερεγγυότητα (solvency) ορίζεται ως εξής

$$\text{Φερεγγυότητα} = \begin{cases} 0, \text{Τακτοποιημένα δυσμενή στοιχεία άνω των 1500€} \\ 1, \text{Τακτοποιημένα δυσμενή στοιχεία κάτω των 1500€} \end{cases} \tag{3.1.2}$$

Η σχέση περιουσίας με τον συνολικό δανεισμό (property) ορίζεται ως εξής

$$\text{Περιουσία σε σχέση με δανεισμό} = \begin{cases} 0, \text{έως και } 100\% \\ 1, \text{από } 100\% \text{ έως } 300\% \\ 2, \text{άνω του } 300\% \end{cases} \tag{3.1.3}$$

Το ιστορικό του πελάτη (history) ορίζεται ως εξής

$$\text{Ιστορικό} = \begin{cases} 0, \text{Κανένα δάνειο στο παρελθόν} \\ 1, \text{Υπάρχον δάνειο χωρίς πρόβλημα εξυπηρέτηση} \\ 2, \text{Προγούμενο δάνειο κανονικά αποπληρωμένο} \end{cases} \tag{3.1.4}$$

Για τις υποκατηγορίες 1 και 2 συμπεριλαμβάνεται το ενδεχόμενο μίας αποπληρωμένης καθυστέρησης εντός δύο μηνών μη επαναλαμβανόμενης.

Η επαγγελματική κατάσταση (employment) ορίζεται ως εξής

$$\text{Επαγγελματική Κατάσταση} = \begin{cases} 0, \text{Ιδιωτικός Υπάλληλος/ Συνταξιούχος Ιδιωτικού Τομέα} \\ 1, \text{Δημόσιος Υπάλληλος/ Υπάλληλος ΔΕΚΟ/Συνταξιούχος Δημοσίου} \\ 2, \text{Ελεύθερος Επαγγελματίας/ Επιστήμονας (π.χ. Γιατρός)} \\ 3, \text{Λοιποί Ελεύθεροι Επαγγελματίες (π.χ. Υδραυλικός)} \end{cases} \tag{3.1.5}$$

Οι 366 παρατηρήσεις χωρίστηκαν στο training set που αποτελείται από 302 παρατηρήσεις οι οποίες επιλέγονται τυχαία από το συνολικό δείγμα και το test set που αποτελείται από τις υπόλοιπες 64 παρατηρήσεις. Με το training set θα εκπαιδευτούν τα μοντέλα και με το test set θα εξεταστεί η προβλεπτική τους ισχύ.

Αρχικά θα εισαχθούν τα δεδομένα στην R και στην συνέχεια θα κατασκευαστούν οι πίνακες συχνοτήτων για τις επεξηγηματικές μεταβλητές και για τα δύο σύνολα δεδομένων ώστε να εξεταστούν με πιο τακτοποιημένο και εύκολο τρόπο. Για την Φερεγγυότητα προκύπτει:

	Training set		**Test set**	
Φερεγγυότητα (solvency)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**
0	17	0.056	2	0.031
1	285	0.944	62	0.969
Σύνολο	302	1	64	1

Πίνακας 3.1: Πίνακας συχνοτήτων για την Φερεγγυότητα

Παρατηρείται ότι στο training set το 94.4% των παρατηρήσεων ανήκει στην κατηγορία 1 δηλαδή έχει τακτοποιημένα δυσμενή στοιχεία κάτω των 1500€. Στο test set αυτό το ποσοστό ανέρχεται στο 96.9%.

Για την περιουσία προκύπτει:

	Training set		**Test set**	
Περιουσία (Property)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**
0	69	0.228	13	0.203
1	55	0.182	13	0.203
2	178	0.589	38	0.594
Σύνολο	302	1	64	1

Πίνακας 3.2: Πίνακας συχνοτήτων για την Περιουσία

Παρατηρείται ότι το πιο σύνηθες για έναν πελάτη είναι η περιουσία ενός πελάτη να είναι άνω του 300% σε σχέση με τον δανεισμό (περίπου 60% δηλαδή 6 στους 10).

Για το ιστορικό προκύπτει: Παρατηρούμε ότι στο training set 38.4% των πελατών δεν έχουν

	Training set		**Test set**	
Ιστορικό (History)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**
0	116	0.384	32	0.500
1	69	0.229	12	0.187
2	117	0.387	20	0..313
Σύνολο	302	1	64	1

Πίνακας 3.3: Πίνακας συχνοτήτων για το Ιστορικό

λάβει κανένα δάνειο στο παρελθόν, το 22.9% έχει τρέχων δάνειο στο οποίο όμως δεν υπάρχει κάποιο πρόβλημα αποπληρωμής και το 38.7% έχει προηγούμενο δάνειο κάτω των 1500€ το οποίο έχει αποπληρωθεί χωρίς κανένα πρόβλημα.

Για την επαγγελματική κατάσταση προκύπτει:

	Training set		**Test set**	
Επαγγελματική Κατάσταση (Employment)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**
0	118	0.391	21	0.328
1	138	0.457	34	0.531
2	13	0.043	2	0.031
3	33	0.109	7	0.109
Σύνολο	302	1	64	1

Πίνακας 3.4: Πίνακας συχνοτήτων για την Επαγγελματική Κατάσταση

Παρατηρούμε στο training set ότι 39.1% ανήκει στην κατηγορία 1, 45.7% στην κατηγορία 2, 4.3% στην κατηγορία 3 και 10.9% στην κατηγορία 4.

3.2 Πιστωτική Βαθμολόγηση με Λογιστική Παλινδρόμηση

Η πρώτη μέθοδος που θα εφαρμοσθεί για την ανάλυση των δεδομένων είναι αυτή της λογιστικής παλινδρόμησης. Επειδή κάποιες από τις συμμεταβλητές είναι κατηγορικές με περισσότερες από δύο κατηγορίες αναφοράς, για την προσαρμογή του μοντέλου θα χρησιμοποιηθούν ψευδομεταβλητές οι οποίες κατασκευάζονται αυτόματα από την R. Για παράδειγμα για το ιστορικό θα χρησιμοποιηθούν δύο ψευδομεταβλητές:

$$\text{Ιστορικό}[1] = \begin{cases} 1, \text{ αν ιστορικό=1} \\ 0, \text{ διαφορετικά} \end{cases} \tag{3.2.1}$$

$$\text{Ιστορικό}[2] = \begin{cases} 1, \text{ αν ιστορικό=2} \\ 0, \text{ διαφορετικά} \end{cases} \tag{3.2.2}$$

Η πρώτη κατηγορία επιλέγεται ως κατηγορία αναφοράς. Κρίνεται σκόπιμο προτού εφαρμοσθούν το μοντέλο της λογιστικής παλινδρόμησης, να κατασκευαστούν οι 2×2 πίνακες συνάφειας κάθε επεξηγηματικής μεταβλητής σε σχέση με την αθέτηση ώστε να υπάρχει μια εικόνα για το ποιες εμφανίζουν μεγαλύτερη συσχέτιση με την αθέτηση και θα είναι πιθανότερο να υπάρχουν στο τελικό μοντέλο.

Ο έλεγχος χ^2 που γίνεται για την κατασκευή του πίνακα συνάφειας έχει τον εξής έλεγχο υπόθεσης:

- H_0:γραμμές και στήλες είναι ανεξάρτητες
- H_1:γραμμές και στήλες είναι εξαρτημένες

Για την φερεγγυότητα προκύπτει:

			Αθέτηση		
Φερεγγυότητα (Solvency)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-τιμή χ^2 ελέγχου**
0	17	0.056	11	0.085	0.052
1	285	0.944	118	0.915	
Σύνολο	302	1	129	1	

Πίνακας 3.5: 2-way πίνακας συνάφειας για την εξάρτηση της αθέτησης από την Φερεγγυότητα

Παρατηρούμε ότι η p-value είναι ελάχιστα μεγαλύτερη του 0.05 επομένως οριακά θα μπορούσε να θεωρηθεί πως δεν έχει υψηλή συσχέτιση με την αθέτηση. Για την περιουσία, το ιστορικό και την επαγγελματική κατάσταση προκύπτει:

Παρατηρείται στον πίνακα 3.6 ότι p-value>0.05 επομένως θεωρείται πως δεν υπάρχει υψηλή συσχέτιση με την αθέτηση και επομένως είναι αρκετά πιθανό να μην συμπεριληφθεί στο τελικό μοντέλο. Για το ιστορικό παρατηρείται στον πίνακα 3.7 πως η p-value είναι μεγαλύτερη του 0.05 οπότε και πάλι θεωρείται πως δεν υπάρχει υψηλή συσχέτιση με την αθέτηση και επομένως είναι πιθανό να μην συμπεριληφθεί στο τελικό μοντέλο. Για την επαγγελματική κατάσταση παρατηρείται στον πίνακα 3.8 πως p-value<0.01 επομένως συσχετίζονται οι δύο μεταβλητές και θεωρείται πως η επαγγελματική κατάσταση θα συμπεριληφθεί στο τελικό μοντέλο.

			Αθέτηση		
Περιουσία (Property)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-τιμή χ^2 ελέγχου**
0	69	0.228	33	0.277	0.184
1	55	0.182	23	0.193	
2	178	0.589	63	0.529	
Σύνολο	302	1	119	1	

Πίνακας 3.6: 2-way πίνακας συνάφειας για την εξάρτηση της αθέτησης από την Περιουσία

			Αθέτηση		
Ιστορικό (History)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-τιμή χ^2 ελέγχου**
0	116	0.384	55	0.462	0.079
1	69	0.229	24	0.202	
2	117	0.387	40	0.336	
Σύνολο	302	1	119	1	

Πίνακας 3.7: 2-way πίνακας συνάφειας για την εξάρτηση της αθέτησης από το Ιστορικό

			Αθέτηση		
Επαγγελματική Κατάσταση (Employment)	**Απόλυτη συχνότητα**	**Σχετική συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-τιμή χ^2 ελέγχου**
0	118	0.391	62	0.521	$3.57\text{x}10^{-6}$
1	138	0.457	32	0.269	
2	13	0.043	7	0.059	
3	33	0.109	18	0.151	
Σύνολο	302	1	119	1	

Πίνακας 3.8: 2-way πίνακας συνάφειας για την εξάρτηση της αθέτησης από την Επαγγελματική Κατάσταση

3.2.1 Προσαρμογή του μοντέλου

Έχοντας μία εικόνα για τις επεξηγηματικές μεταβλητές μπορούμε να προχωρήσουμε στην εφαρμογή του μοντέλο της λογιστικής παλινδρόμησης. Με κατασκευή ψευδομεταβλητών που γίνεται αυτόματα στην R προκύπτει:

	Συντελεστής	**Τυπ. Απόκλιση**	**z-τιμή ελέγχου**	**p-τιμή**
Intercept	2.172	0.6429	3.378	<0.01
Φερεγγυότητα [2]	-1.319	0.5554	-2.375	0.01756
Περιουσία [2]	-0.490	0.3974	-1.233	0.2177
Περιουσία [3]	-0.850	0.3250	-2.615	<0.01
Ιστορικό [2]	-0.3450	0.3442	-1.003	0.3160
Ιστορικό [3]	-0.3568	0.2924	-1.220	0.2224
Επ. Κατάσταση [2]	-1.421	0.2943	-4.828	<0.01
Επ. Κατάσταση [3]	0.07267	0.6010	0.1209	0.9037
Επ. Κατάσταση [4]	0.1952	0.4085	0.4779	0.6327

Πίνακας 3.9: Λογιστική Παλινδρόμηση στο Training Set με χρήση των μεταβλητών Φερεγγυότητα, Περιουσία, Ιστορικό και Επαγγελματική Κατάσταση

Για να ελεγχθεί η καλή προσαρμογή του μοντέλου θα χρησιμοποιηθεί ο έλεγχος Hosmer-Lemeshow διότι οι έλεγχοι Deviance και Pearson δεν χρησιμοποιούνται σε μοντέλα με δυαδικά δεδομένα, δύναται όμως να χρησιμοποιηθεί ο έλεγχος Deviance για την σύγκριση μεταξύ δύο μοντέλων (Collett, 2003). Προκύπτει από τον έλεγχο ότι

$$X^2_{HL} = 1185.306 \text{ με p-value}<0.01 \text{ και } 8 \text{ βαθμούς ελευθερίας} \tag{3.2.3}$$

Επίσης ισχύει ότι AIC=378.51.

Στον πίνακα 3.9 παρατηρούμε ότι και οι δύο κατηγορίες του ιστορικού δεν είναι στατιστικά σημαντικές μεταβλητές επομένως το ιστορικό μπορεί να αφαιρεθεί από το μοντέλο. Κάτι τέτοιο έδειξε και ο χ^2 έλεγχος που έγινε νωρίτερα. Προσαρμόζοντας το νέο μοντέλο χωρίς το ιστορικό προκύπτει ο πίνακας 3.10 για τους συντελεστές.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.0012	0.6233	3.21	0.0013
solvency[1]	-1.2767	0.5485	-2.33	0.0199
property[1]	-0.5465	0.3945	-1.39	0.1659
property[2]	-0.9153	0.3204	-2.86	0.0043
employment[1]	-1.4843	0.2905	-5.11	0.0000
employment[2]	0.0571	0.5970	0.10	0.9238
employment[3]	0.1315	0.4030	0.33	0.7441

Πίνακας 3.10: Λογιστική Παλινδρόμηση στο Training Set με χρήση των μεταβλητών Φερεγγυότητα, Περιουσία και Επαγγελματική Κατάσταση

Το AIC του νέου μοντέλου είναι 376.27, μικρότερο από το αρχικό, άρα το μοντέλο βελτιώθηκε. Παρατηρούμε ότι οι περισσότεροι συντελεστές έχουν αρνητικό πρόσημο. Αυτό μεταφράζεται ως εξής: Αν συγκρίνουμε έναν πελάτη που έχει τακτοποιημένα δυσμενή στοιχεία άνω των 1500€ (Φερεγγυότητα [0]) με έναν δεύτερο που έχει τακτοποιημένα δυσμενή στοιχεία κάτω τον 1500€ (Φερεγγυότητα [1]) και οι υπόλοιπες μεταβλητές τιμές έχουν την ίδια τιμή και για τους δύο πελάτες, τότε από το αρνητικό πρόσημο θα συμπεράνουμε πως ο δεύτερος πελάτης έχει μικρότερη πιθανότητα να αθετήσει το δάνειο του. Όμοια ένας πελάτης που είναι Δημόσιος Υπάλληλος/ΔΕΚΟ/Συνταξιούχος Δημοσίου (Επαγγελματική Κατάσταση [1]) έχει μικρότερη πιθανότητα να αθετήσει το δάνειο του σε σχέση με έναν πελάτη που ανήκει στην κατηγορία Επαγγελματική Κατάσταση [0]. Παρ'όλα αυτά παρατηρούμε ότι ένας πελάτης που ανήκει ανήκει στην Επαγγελματική Κατάσταση [2] έχει μεγαλύτερη πιθανότητα να αθετήσει το δάνειο του σε σχέση με κάποιον που ανήκει στην Επαγγελματική Κατάσταση [1].

Παρατηρούμε από τον έλεγχο wald που έχει γίνει, πως η Φερεγγυότητα είναι στατιστικά σημαντική μεταβλητή. Όσο αναφορά την περιουσία και την επαγγελματική κατάσταση παρατηρούμε

κάποιες κατηγορίες τους δεν είναι στατιστικά σημαντικές ενώ άλλες είναι. Για αυτό το λόγο θα διατηρηθούν στο μοντέλο, καθώς αν αφαιρεθούν το μοντέλο θα υστερεί σε προβλεπτική ισχύ. Άλλωστε υπήρχαν ενδείξεις από τους χ^2 ελέγχους που έγιναν.

Γνωρίζοντας τους συντελεστές μπορούμε να υπολογίσουμε πλέον την πιθανότητα για κάθε ενδεχόμενο. Ας υπολογίσουμε την πιθανότητα ένας πελάτης με τα χαρακτηριστικά Φερεγγυότητα [1], Περιουσία [1] και Επαγγελματική Κατάσταση [3] να αθετήσει το δάνειο του. Η πιθανότητα αυτή υπολογίζεται από τον τύπο:

$$\hat{p}(x) = \frac{e^{2.0012-1.2767*1-0.5465*1+0.1315*1}}{1+e^{2.0012-1.2767*1-0.5465*1+0.1315*1}} = \frac{e^{0.3095}}{1+e^{0.3095}} = 0.5768 \tag{3.2.4}$$

Παρατηρούμε ότι ένας πελάτης με αυτά τα χαρακτηριστικά έχει πιθανότητα αθέτησης του δανείου του 57.67%. Συνεπώς δεν θα ήταν συνετό να του πιστωθεί κάποιο δάνειο. Όμως ακόμα και αν αυτή η πιθανότητα ήτανε κάτω του 50%, δεν σημαίνει ότι θα ήταν σωστό να του πιστωθεί το δάνειο. Εμείς θα ορίσουμε το ποσοστό-όριο που θέλουμε, ανάλογα με το ρίσκο που είναι διατεθειμένη η τράπεζα να πάρει.

Στον πίνακα 3.11 παρατηρούμε το 95% διάστημα εμπιστοσύνης για τους συντελεστές. Όπως αναφέρθηκε και παραπάνω, οι κατηγορίες που περιέχουν το 0 είναι στατιστικά μη σημαντικές, όμως θα παραμείνουν στο μοντέλο.

	2.5 %	97.5 %
(Intercept)	0.81	3.29
solvency[1]	-2.42	-0.23
property[1]	-1.33	0.22
property[2]	-1.55	-0.29
employment[1]	-2.07	-0.93
employment[2]	-1.12	1.26
employment[3]	-0.66	0.93

Πίνακας 3.11: 95% διάστημα εμπιστοσύνης για τους συντελεστές

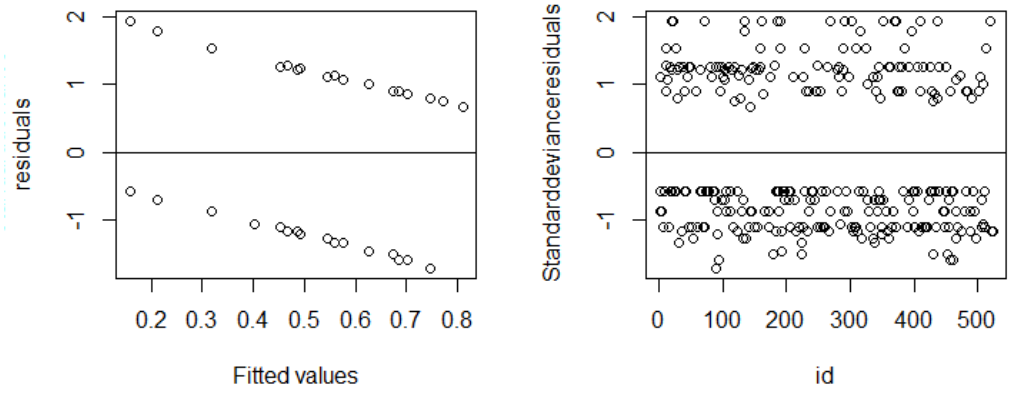
Όσο αναφορά την ελεγχοσυνάρτηση deviance μπορούμε να εξετάσουμε την μεταβολή της στο αρχικό μοντέλο σε σχέση με αυτό χωρίς το ιστορικό. Η μεταβολή των συναρτήσεων είναι 362.27-360.51=1.57, δηλαδή η μεταβολή είναι μικρή, κάτι που επιβεβαιώνει πως το ιστορικό δεν είναι στατιστικά σημαντικό.

3.2.2 Γραφήματα Υπολοίπων

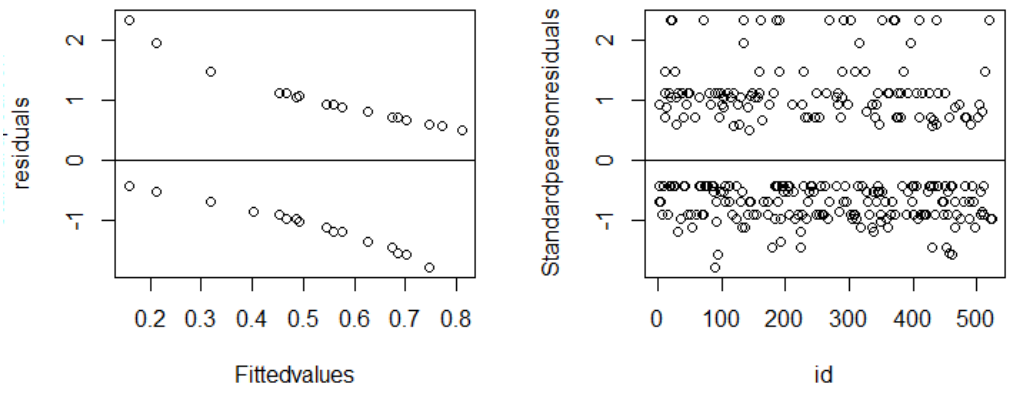
Για να ελεγχθεί η καταλληλότητα του μοντέλου είναι απαραίτητο να γίνουν τα γραφήματα υπολοίπων. Στους πίνακες 3.2.1, 3.2.2, 3.2.3 παρουσιάζονται τα γραφήματα των υπολοίπων deviance, pearson και πιθανοφάνειας . Επίσης στον πίνακα 3.2.4 παρουσιάζονται οι αποστάσεις Cook.

Από τα γραφήματα των τυποποιημένων υπολοίπων deviance σε σχέση με τις εκτιμώμενες τιμές και με βάση τη σειρά των δεδομένων παρατηρούμε ότι οι παρατηρήσεις είναι ανεξάρτητες και ότι δεν υπάρχουν παρατηρήσεις που να αποκλίνουν σημαντικά από τις υπόλοιπες. Ειδικότερα, το γράφημα των υπολοίπων deviance σε σχέση με τις εκτιμώμενες τιμές χωρίζεται σε δύο μέρη. Παρατηρούμε, λοιπόν, ότι στο "επάνω" τμήμα οι προσαρμοσμένες τιμές πλησιάζουν την τιμή 1 και έτσι τα τυποποιημένα υπόλοιπα deviance τείνουν στο 0. Αντιστοίχως, στο "κάτω" τμήμα οι προσαρμοσμένες τιμές πλησιάζουν την τιμή 0 και έτσι τα τυποποιημένα υπόλοιπα deviance τείνουν και πάλι στο 0. Αυτή είναι και η επιθυμητή συμπεριφορά για τα υπόλοιπα αυτά έτσι ώστε να μην υπάρχουν παρατηρήσεις που αποκλίνουν από τις υπόλοιπες.

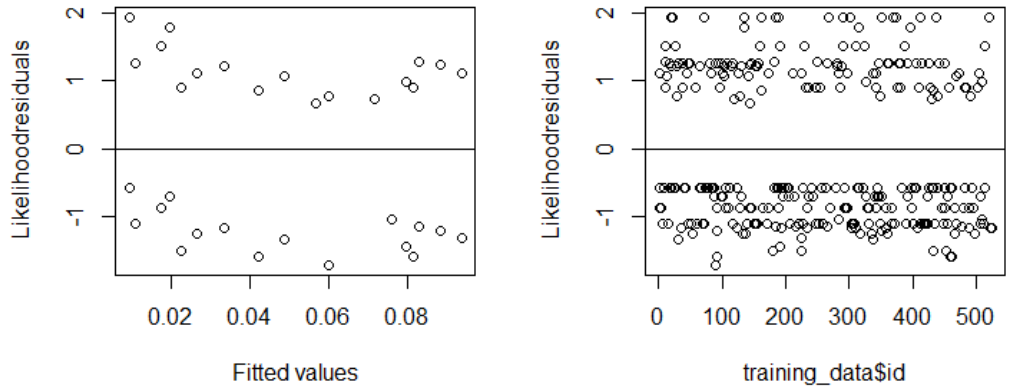
Αντίστοιχα, το γράφημα των τυποποιημένων υπολοίπων deviance σε σχέση με τη σειρά των δεδομένων χωρίζεται κι αυτό στα ίδια τμήματα. Παρατηρούμε λοιπόν ότι και στα δύο τμήματα η κατανομή των υπολοίπων είναι τυχαία και δεν ακολουθεί κάποιο μοτίβο, πράγμα που υποδεικνύει ότι οι παρατηρήσεις είναι ανεξάρτητες και δεν έχουν σχέση με την σειρά εμφάνισής τους στο δείγμα. Παρόμοια εικόνα δίνουν και τα υπόλοιπα pearson. Για τον εντοπισμό σημείων επιρροής κατασκευάστηκε το διάγραμμα των υπολοίπων πιθανοφάνειας ως προς τα $\hat{h}_{ii}$ αλλά και τα γραφήματα δείκτη των υπολοίπων πιθανοφάνειας, των $\hat{h}_{ii}$ και των αποστάσεων Cook. Από τα τέσσερα αυτά γραφήματα παρατηρούμε δεν υπάρχουν σημεία επιρροής στο μοντέλο μας.



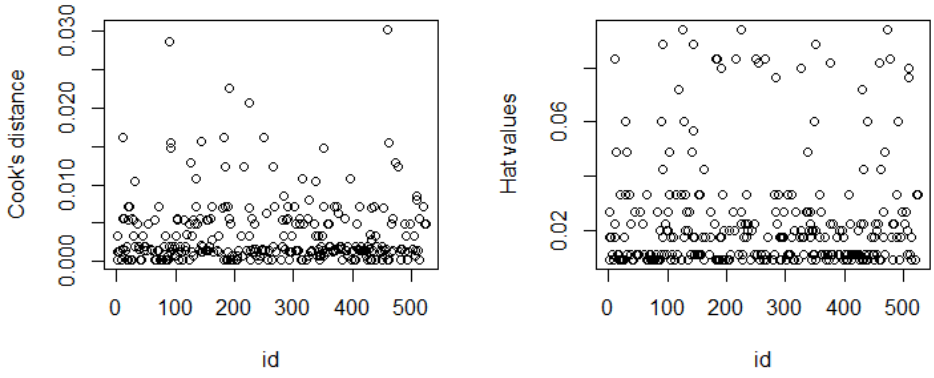
Σχήμα 3.2.1: Υπόλοιπα Deviance



Σχήμα 3.2.2: Υπόλοιπα Pearson



Σχήμα 3.2.3: Υπόλοιπα Πιθανφάνειας



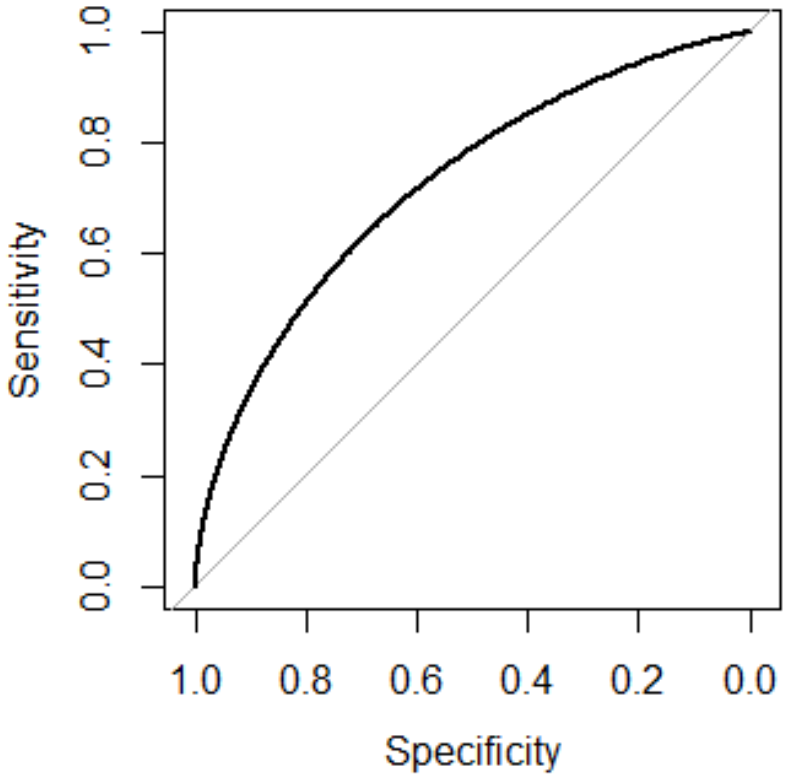
Σχήμα 3.2.4: Γραφήματα δείκτη των αποστάσεων Cook και των hat values

3.2.3 Προβλεπτική Ικανότητα

Ο κύριος στόχος ενός μοντέλου πιστωτικής βαθμολόγησης είναι να κατηγοριοποιήσει μία νέα αίτηση με την μικρότερη πιθανότητα λάθους. Δηλαδή είναι απαραίτητο το μοντέλο να έχει υψηλή προβλεπτική ισχύ ώστε να είναι χρήσιμο για τον δανειοδότη. Στο συγκεκριμένο μοντέλο χρησιμοποιήθηκε η καμπύλη ROC. Για να έχει υψηλή προβλεπτική ικανότητα, θα πρέπει το εμβαδόν κάτω από την καμπύλη (Area Under Curve) να είναι όσο πιο κοντά στην μονάδα γίνεται καθώς και για μικρές τιμές του (1-Specificity), το Sensitivity να λαμβάνει μεγάλες τιμές.

Training Set

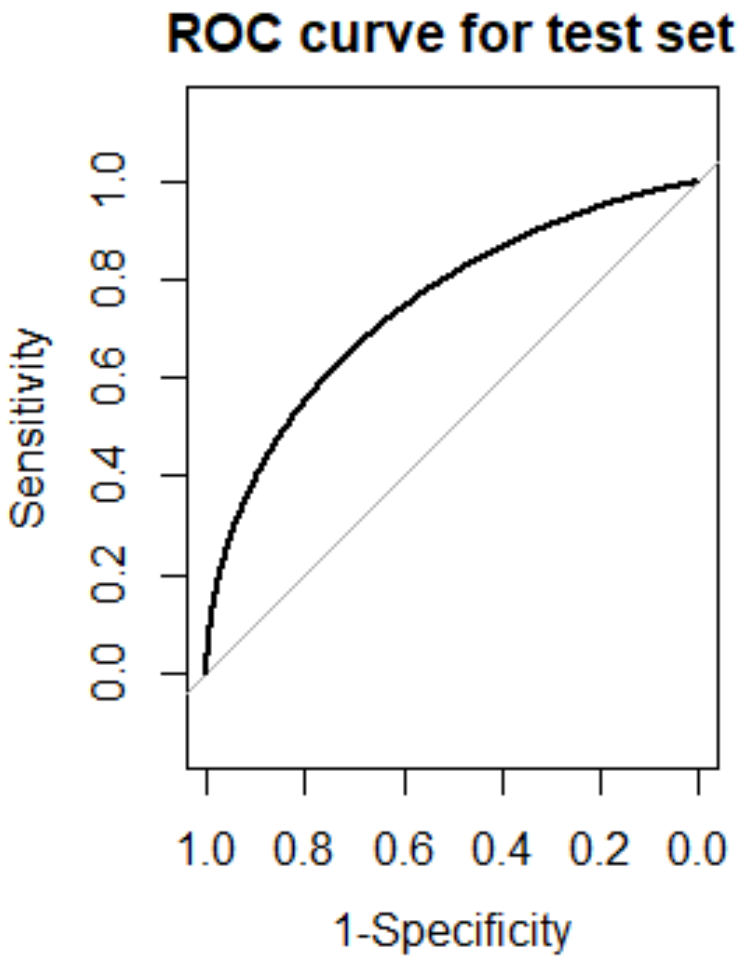
Στο σχήμα 3.2.5 βλέπουμε την καμπύλη ROC για το training set. Το AUC προκύπτει 0.7241 με 95% διάστημα εμπιστοσύνης το [0.6666, 0.7866]. Η τιμή αυτή είναι αρκετά ικανοποιητική και δείχνει ότι το μοντέλο έχει καλή προβλεπτική ισχύ.



Σχήμα 3.2.5: Καμπύλη ROC για το Training Set

Test Set

Δοκιμάζοντας το μοντέλο και στο test set μπορεί να ελεγχθεί αν μπορεί να αντεπεξέλθει το ίδιο καλά και σε διαφορετικά δεδομένα το μοντέλο. Στο σχήμα 3.2.6 βλέπουμε αυτήν την καμπύλη. Το AUC υπολογίστηκε 0.7458 με 95% διάστημα εμπιστοσύνης το [0.5889, 0.8791]. Η τιμή αυτή είναι αρκετά καλή και επιβεβαιώνει πως το μοντέλο ανταποκρίνεται καλά και σε διαφορετικά δεδομένα. Ο λόγος που το διάστημα εμπιστοσύνης λαμβάνει μεγάλες τιμές είναι λόγω του μικρού μεγέθους του test set.

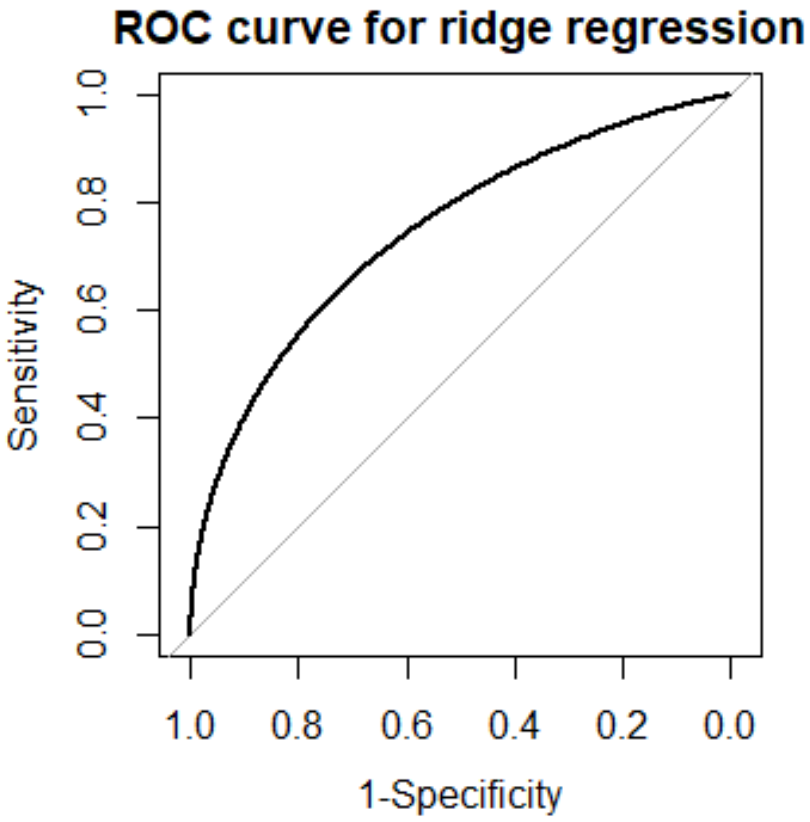


Σχήμα 3.2.6: Καμπύλη ROC για το Test Set

3.3 Εφαρμογή Παλινδρόμησης Κορυφογραμμής

Στην συγκεκριμένη παράγραφο θα γίνει η προσαρμογή του μοντέλου της παλινδρόμησης κορυφογραμμής που περιγράψαμε στο κεφάλαιο 2.2. Θα χρησιμοποιηθούν τα αρχικά δεδομένα του training set μαζί με το ιστορικό καθώς αυτό το μοντέλο χρησιμοποιεί όλες τις μεταβλητές και ελαχιστοποιεί κάποιες με την παράμετρο λ. Αφού προσαρμοσθεί το μοντέλο στο training set θα προκύψει το βέλτιστο λ, και με αυτό θα σχεδιαστεί η καμπύλη ROC στο test set για να εξετάσουμε εάν προκύπτει τελικά ένα καλύτερο μοντέλο από αυτό της λογιστικής παλινδρόμησης.

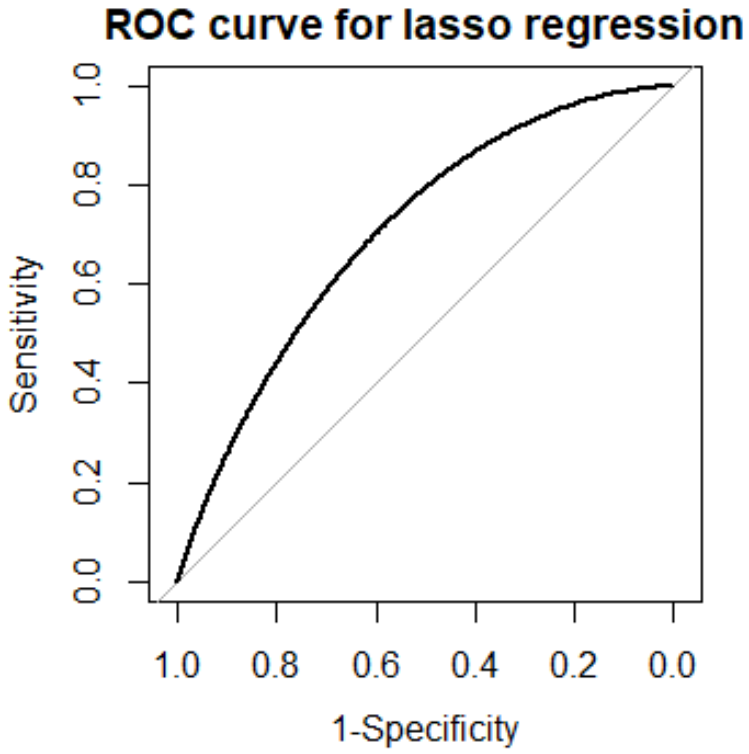
Εισάγοντας τα δεδομένα στην R, προκύπτει πως το βέλτιστο λ είναι λ=0.78125. Η καμπύλη ROC που προέκυψε παρουσιάζεται στο σχήμα 3.3.1. Το AUC για το συγκεκριμένο μοντέλο προκύπτει 0.7453. Άρα προκύπτει ένα αρκετά αξιόπιστο μοντέλο. Το AUC είναι ελαφρώς μεγαλύτερο σε σχέση με αυτό της λογιστικής παλινδρόμησης. Το 95% διάστημα εμπιστοσύνης για το AUC είναι [0.5825, 0.8799].



Σχήμα 3.3.1: Καμπύλη ROC για την Παλινδρόμηση Κορυφογραμμής

3.3.1 Εφαρμογή Παλινδρόμησης Λάσσο

Σε αυτήν την παράγραφο θα γίνει εφαρμογή της παλινδρόμησης Λάσσο, για να εξεταστεί εάν προκύπτει και πάλι ένα αξιόπιστο μοντέλο. Ο έλεγχος αυτός θα γίνει μέσω της καμπύλης ROC. Προσαρμόζοντας το μοντέλο στο εκ νέου στο training set, βρίσκουμε ότι το βέλτιστο λ είναι 0.02314. Κατασκευάζουμε την καμπύλη ROC που φαίνεται στο σχήμα 3.3.2 και υπολογίζουμε το AUC. Προκύπτει το AUC=0.7054 με 95% διάστημα εμπιστοσύνης το [0.5478, 0.8475]. Παρατηρούμε ότι το αυτό το μοντέλο έχει μικρότερη προβλεπτική ισχύ σε σχέση με την παλινδρόμηση κορυφογραμμής οπότε πιθανότατα να μην ήταν γινόταν χρήση του.



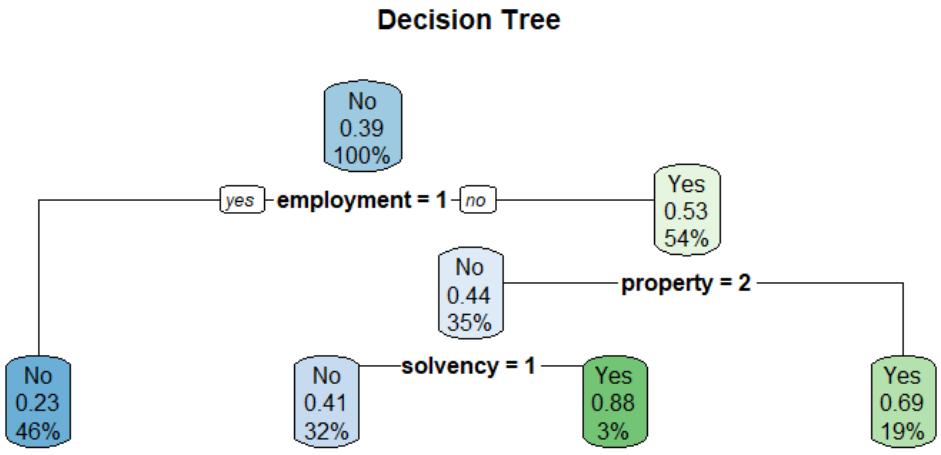
Σχήμα 3.3.2: Καμπύλη ROC για την Παλινδρόμηση Lasso

3.4 Πιστωτική Βαθμολόγηση με Δένδρα Αποφάσεως

Στο συγκεκριμένο κεφάλαιο θα γίνει μία προσπάθεια δημιουργίας ενός δένδρου αποφάσεως για το πρόβλημα της πιστωτικής βαθμολόγησης. Η εξαρτημένη αλλά και οι ανεξάρτητες μεταβλητές εξακολουθούν να είναι όπως περιγράφτηκαν στο κεφάλαιο 3.1.

3.4.1 Κατασκευή και ερμηνεία του μοντέλου

Εισάγοντας τα δεδομένα στην R, και με την χρήση κατάλληλων εντολών κατασκευάζεται το δένδρο ταξινόμησης (classification tree) του σχήματος 3.4.1. Η R χρησιμοποιεί το κριτήριο Gini για να κατασκευάσει το δένδρο.



Σχήμα 3.4.1: Δένδρο Απόφασης στην Πιστωτική Βαθμολόγησης

Παρατηρούμε πως η R επιλέγει ως ρίζα την μη αθέτηση του δανείου εντός πενταετίας και πως η πιο σημαντική συμμεταβλητή είναι το επάγγελμα καθώς βάσει αυτού γίνεται ο πρώτος διαχωρισμός στους επιμέρους κόμβους. Σύμφωνα με το δένδρο ταξινόμησης προκύπτει ότι:

1. Εάν ο αιτών είναι Δημόσιος Υπάλληλος/ Υπάλληλος ΔΕΚΟ/ Συνταξιούχος Δημοσίου (κατάσταση 1) τότε δεν θα αθετήσει το δάνειο.

2. Εάν ο αιτών δεν ανήκει στην παραπάνω επαγγελματική κατάσταση και η περιουσία του σε σχέση με τον δανεισμό δεν είναι άνω του 300% (κατάσταση 2) τότε θα αθετήσει το δάνειο.

3. Εάν ο αιτών δεν ανήκει στην επαγγελματική κατάσταση 1, δεν ανήκει στην περιουσιακή κατάσταση 2 αλλά έχει τακτοποιημένα δυσμενή στοιχεία άνω των 1500€ τότε δεν θα αθετήσει το δάνειο.

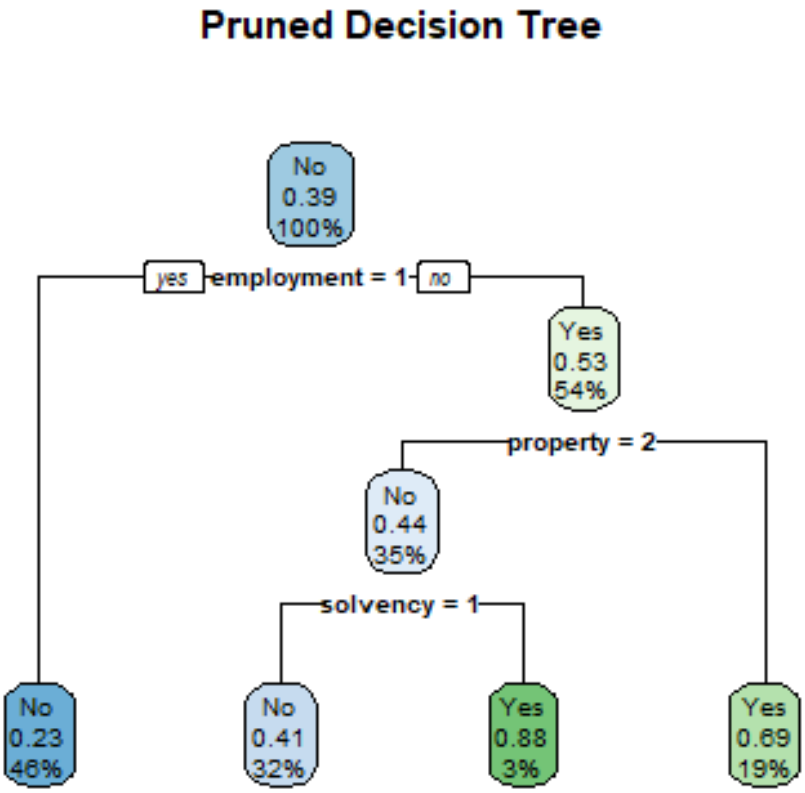
4. Εάν ο αιτών δεν ανήκει στην επαγγελματική κατάσταση 1, δεν ανήκει στην περιουσιακή κατάσταση 2 και έχει τακτοποιημένα δυσμενή στοιχεία κάτω των 1500€ τότε θα αθετήσει το δάνειο.

Βλέπουμε πως και πάλι δεν γίνεται χρήση της μεταβλητής Ιστορικό όπως έδειξε και το μοντέλο της λογιστικής παλινδρόμησης. Επίσης η σημαντικότητα κάθε μεταβλητής συμφωνεί με τις σημαντικές ψευδομεταβλητές που κατέδειξε η λογιστική παλινδρόμηση.

Σε κάθε κόμβο κόμβο του σχήματος 3.4.1 φαίνεται το ποσοστό του δείγματος το οποίο περιέχει. Δηλαδή η περίπτωση 1. που περιγράψαμε περιέχει το 46% του δείγματος. Τα δύο αριστερά φύλλα καταλήγουν στην περίπτωση μη αθέτησης, ενώ τα δύο δεξιά στην αθέτηση. Τελικώς προκύπτει πως το 32% του δείγματος ανήκει στην περίπτωση 2, το 3% στην περίπτωση 3 και το 19% στην περίπτωση 4.

3.4.2 Κλάδεμα του Δένδρου και Τυχαίο Δάσος

Με την χρήση της μεθόδου pruning θα προσπαθήσουμε να κατασκευάσουμε ένα ακριβέστερο δένδρο παλινδρόμησης το οποίο να ελαχιστοποιεί την εσφαλμένη ταξινόμηση. Χρησιμοποιώντας την R για το κλάδεμα του δέντρου προκύπτει το δένδρο του σχήματος 3.4.2 και βλέπουμε πως τελικά προκύπτει ακριβώς το ίδιο δένδρο με αυτό του δένδρου ταξινόμησης. Συνεπώς αυτή η μέθοδος δεν βελτίωσε κάπως το μοντέλο.

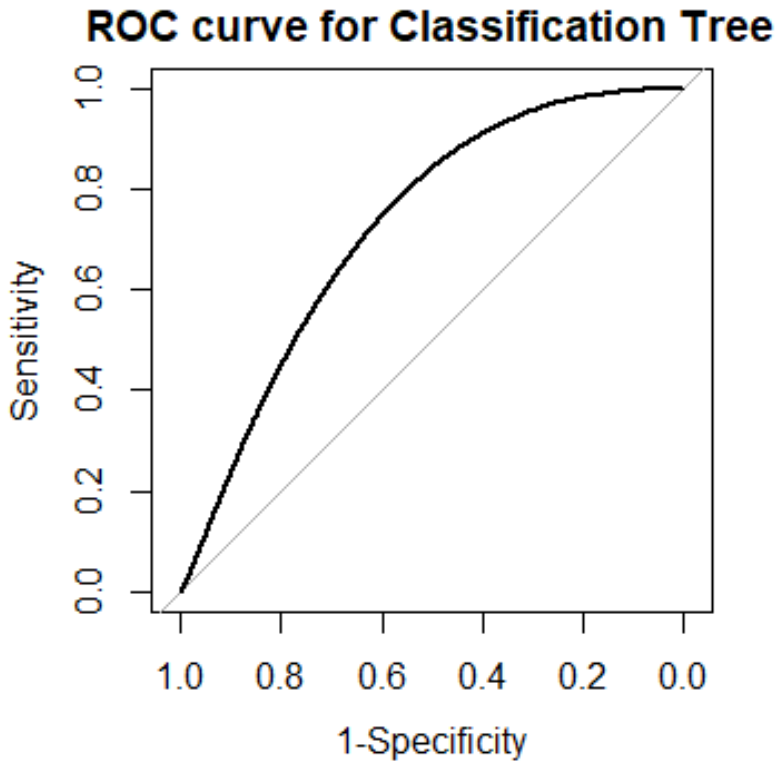


Σχήμα 3.4.2: Δένδρο Απόφασης στην Πιστωτική Βαθμολόγησης με χρήση κλαδέματος

Μία άλλη μέθοδος που δύναται να χρησιμοποιηθεί για να βελτιωθεί το εν λόγω μοντέλο είναι το τυχαίο δάσος (random forest). Το μοντέλο που προκύπτει θα συγκριθεί με αυτό του δέντρου ταξινόμησης μέσω της καμπύλης ROC για να ελεγχθεί η ακρίβεια και των δύο μοντέλων.

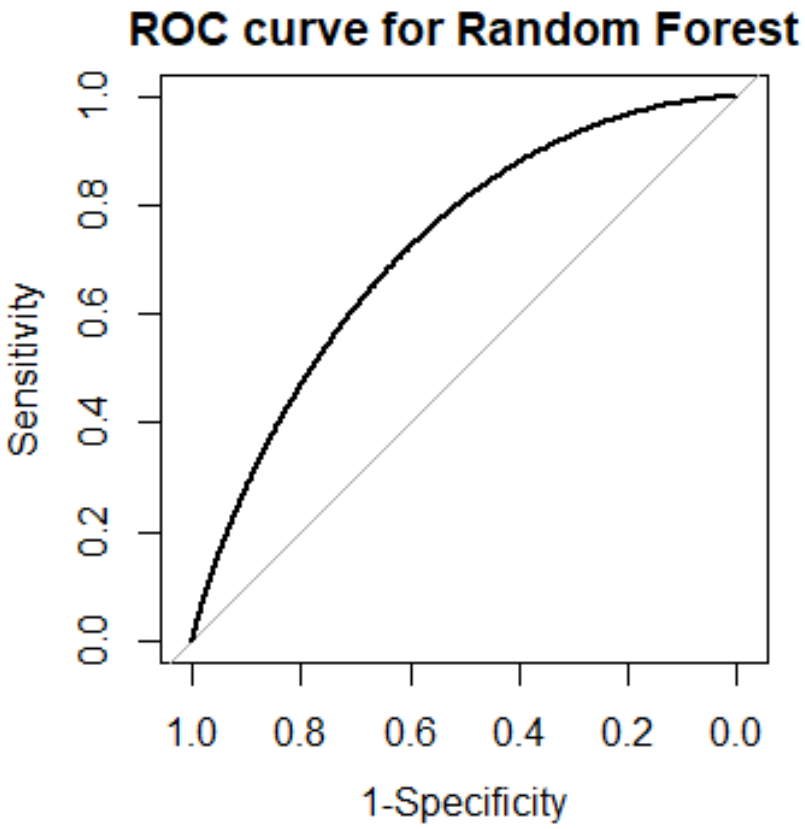
3.4.3 Προβλεπτική Ικανότητα

Θα γίνει χρήση του test set στο μοντέλο του δέντρου ταξινόμησης και του τυχαίου δάσους για να ελεγχθεί η ακρίβεια τους. Στο σχήμα 3.4.3 προκύπτει καμπύλη για το πρώτο μοντέλο ενώ στο σχήμα 3.4.4 για το δεύτερο.



Σχήμα 3.4.3: Καμπύλη ROC για το δένδρο ταξινόμησης

Το AUC για το δένδρο ταξινόμησης προκύπτει 0.7259 με 95% διάστημα εμπιστοσύνης το [0.5554, 0.8594]. Για το τυχαίο δάσος προκύπτει το AUC=0.7205 με 95% διάστημα εμπιστοσύνης το [0.5678, 0.8573]. Φαίνεται πως και τα δύο μοντέλα έχουν ένα καλό AUC και επειδή το τυχαίο δάσος αντιμετωπίζει και το πρόβλημα του overfitting, ίσως να αποτελεί και την προτιμότερη επιλογή.



Σχήμα 3.4.4: Καμπύλη ROC για το τυχαίο δάσος

3.5 Συμπεράσματα

Η πιστωτική βαθμολόγηση είναι ένα εργαλείο που χρησιμοποιείται σε μεγάλο βαθμό από τα χρηματοπιστωτικά ιδρύματα τις τελευταίες δεκαετίες. Με την χρήση μηχανικής μάθησης και ανάλυσης δεδομένων έχει καταστεί εφικτή η διαχείριση του μεγάλου όγκου δεδομένων και πελατών που διαχειρίζονται τα χρηματοπιστωτικά ιδρύματα. Αυτές οι τεχνικές μπορούν να βοηθήσουν στον υπολογισμό της πιθανότητας αθέτησης του εκάστοτε πελάτη με αποτέλεσμα τα χρηματοπιστωτικά ιδρύματα να τις αξιοποιούν και να αποφασίζουν ποιος θα δανειοδοτηθεί και ποιος όχι. Στην παρούσα εργασία έγινε εφαρμογή διαφόρων μοντέλων στο συγκεκριμένο πρόβλημα τα οποία κατέδειξαν ότι η λογιστική παλινδρόμηση, η lasso παλινδρόμηση, η παλινδρόμηση κορυφογραμμής αλλά και τα δένδρα αποφάσεως έχουν καλή προβλεπτική ικανότητα και προσαρμόζονται ικανοποιητικά στα δεδομένα. Το μοντέλο που φαίνεται να υπερέχει ελάχιστα είναι αυτό της λογιστικής παλινδρόμησης καθώς έχει AUC=0.7458 στο test set. Όμως αυτό δεν σημαίνει πως δεν μπορούν να χρησιμοποιηθούν και τα άλλα μοντέλα που παρουσιάστηκαν καθώς έδωσαν ικανοποιητικά αποτελέσματα.

Όλα τα μοντέλα κατέδειξαν την μεταβλητή ιστορικό ως περιττή οπότε και έτσι αφαιρέθηκε από όλα. Τα αποτελέσματα τα οποία προέκυψαν κρίνονται επαρκώς ικανοποιητικά, μιας και λήφθηκαν υπ' όψιν λίγες συμμεταβλητές. Επιπλέον το μέγεθος το δείγματος δεν είναι ιδιαίτερα μεγάλο. Επομένως, το ποιο μοντέλο θα μπορούσε να χρησιμοποιηθεί από ένα χρηματοπιστωτικό ίδρυμα εξαρτάται από τον εκάστοτε αναλυτή και τις γνώσεις του καθώς όλα τα μοντέλα που παρουσιάστηκαν είναι άξια χρήσης.

Κεφάλαιο 4

Βιβλιογραφία

Albrecht, J. P. (2016). How the GDPR will change the world. *Eur. Data Prot. L. Rev.*, *2*, 286–287.

Allison, P. D. (2014). Measures of fit for logistic regression. *Proceedings of the SAS Global Forum 2014 Conference*. https://doi.org/https://statisticalhorizons.com/wp-content/uploads/MeasuresOfFitForLogisticRegression-Slides.pdf

Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: Upcoming trends and challenges. *Journal of the Operational Research Society*, *60*(sup1), S16–S23.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ*, *320*(7247), 1468.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, *17*, 22.

Capon, N. (1982). Credit scoring systems: A critical analysis. *Journal of Marketing*, *46*(2), 82–91.

Collett, D. (2003). *Modelling binary data* (2nd ed.). Chapman & Hall/CRC.

De Servigny, A., & Renault, O. (2004). *Measuring and managing credit risk*. McGraw-Hill.

Ghahramani, Z. (2003). Unsupervised learning. *Summer school on machine learning*, 72–112. https://doi.org/https://link.springer.com/chapter/10.1007/978-3-540-28650-9_5

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283.

Lewis, E. M. (1992). *An introduction to credit scoring*. Fair, Isaac; Company.

Lu, D. (2019). Creating an AI can be five times worse for the planet than a car. *New Scientist*, *6*.

Maletic, J. I., & Marcus, A. (2000). Data cleansing: Beyond integrity analysis. *In Iq*, 200–209.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, *4*(2), 227–243.

Mitchell, T. M. (1997). Artificial neural networks. *Machine learning*, *45*, 81–127.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT Press.

Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects. *NPJ Computational Materials*, *3*(1), 1–13.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947–1958.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. SIAM.

Καρώνη, Χ., & Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης. 2η έκδοση, Εκδόσεις Συμεών Αθήνα*.

Παράρτημα Α

Κώδικας στην R

Βιβλιοθήκες που χρησιμποιήθηκαν:

```
library(dplyr)
library(pROC)
library(glmnet)
library(ggplot)
library(rpart)
library(cowplot)
library(randomForest)
library(tree)
library(xtable)
library(glmtoolbox)
library(lmtest)
library(rpart.plot)
```

Εισαγωγή Δεδομενων:

```
##eisagw dedomena
training_data<- read.table(choose.files(), header = TRUE)
##afairw id
training_data<-training_data[,-1]
for (i in 2:6){
  training_data[,i]<-factor(training_data[,i])
}
levels(training_data$athetisi)<-c('No','Yes')
levels(training_data$solvency)<-c('0','1')
```

```
levels(training_data$property)<-c('0','1','2')
levels(training_data$history)<-c('0','1','2')
levels(training_data$employment)<-c('0','1','2','3')
attach(training_data)
test.data<- read.table(choose.files(), header = TRUE)
test.data<-select(test.data,-1)
for (i in 1:5){
  test.data[,i]<-factor(test.data[,i])

}
levels(test_data$solvency)<-c('0','1')
levels(test_data$property)<-c('0','1','2')
levels(test_data$history)<-c('0','1','2')
levels(test_data$employment)<-c('0','1','2','3')
```

Πίνακες συνάφειας και χ^2 έλεγχοι

```
##
freq.table1<-table(athetisi,solvency)
freq.table2<-table(athetisi,property)
freq.table3<-table(athetisi,history)
freq.table4<-table(athetisi,employment)

## 2
chisq1 <- chisq.test(freq.table1)
chisq2 <- chisq.test(freq.table2)
chisq3 <- chisq.test(freq.table3)
chisq4 <- chisq.test(freq.table4)
```

Εφαρμογή μοντέλου Λογιστική Παλινδρόμησης με όλα τα δεδομένα και έπειτα χωρίς το ιστορικό:

```
#logistiki palindromisi
glm.fits<-glm(athetisi~.,data=training_data,family=binomial)
summary(glm.fits)
coef(glm.fits)
summary(glm.fits)$coef
##logistiki palindromisi xoris history
glm.fits2<-glm(athetisi~solvency+property+employment,
```

```
data=training_data,family=binomial)
```

Για τα γραφήματα υπολοίπων χρησιμοποιήθηκαν οι εντολές:

```
##grafimata ypoloipon

#pearson
pearson.res<-residuals(glm.fits2,type="pearson")
st.pearson.res<-residuals(glm.fits2,type="pearson")/
(sqrt(1-hatvalues(glm.fits2)))
par(mfrow=c(1,2))
plot(fitted.values(glm.fits2),st.pearson.res,xlab="Fitted values"
,ylab="Standard pearson residuals")
abline(h=0)
plot(id,st.pearson.res,ylab="Standard pearson residuals")
abline(h=0)



##deviance
deviance.res<-residuals(glm.fits2,type="deviance")
st.deviance.res<-residuals(glm.fits2,type="deviance")
/(sqrt(1-hatvalues(glm.fits2)))

par(mfrow=c(1,2))
plot(fitted.values(glm.fits2),st.deviance.res,xlab="Fitted values"
,ylab="Standard deviance residuals")
abline(h=0)
plot(id,st.deviance.res,ylab="Standard deviance residuals")
abline(h=0)

##pithanofaneias
y=as.numeric(training_data$athetisi)-1
res.lik<-sign(y-fitted.values(glm.fits2))*sqrt((hatvalues(glm.fits2)*
  (st.pearson.res)^2)+((1-hatvalues(glm.fits2))*(st.deviance.res)^2))
par(mfrow=c(1,2))
plot(hatvalues(glm.fits2),res.lik,        xlab="Fitted values",
ylab="Likelihood residuals")
```

```
abline(h=0)
plot(training_data$id,res.lik,  ylab="Likelihood residuals")
abline(h=0)

##cooks distance
par(mfrow=c(1,2))
plot(training_data$id,cooks.distance(glm.fits2),
ylab="Cook's distance",xlab='id')
plot(training_data$id,hatvalues(glm.fits2),      ylab="Hat values"
,xlab='id')
```

Παράδειγμα εντολής για την δημιουργία καμπύλης ROC (όμοια εντολή χρησιμοποιήθηκες για την δημιουργία όλων):

```
##roc gia training set
par(mfrow=c(1,2))
roc(athetisi, fitted.values(gml.step1),ci=TRUE, smooth=TRUE, plot=TRUE,
ci=TRUE,
    xlab='1-Specificity',
    main='ROC curve for training set')
```

Κατασκευή μοντέλου παλινδρόμηση κορυφογραμμής και λασσό:

```
##ridge regression
x.train<-model.matrix(athetisi~.,training_data )[,-1]
y.train<-training_data$athetisi
cv.ridge <- cv.glmnet(x.train, y.train, alpha = 0, family = "binomial")
ridge.model<-glmnet(x.train, y.train, alpha=0, family ="binomial",
                    lambda = cv.ridge$lambda.min)
                    x.test <- model.matrix(athetisi ~., test_data)[,-1]
y.test<-test_data$athetisi
ridge.probs<-predict(ridge.model,s=cv.ridge$lambda.min,type="response",
newx=x.test)
ridge.pred<-ifelse(ridge.probs > 0.5, "1", "0")
cv.ridge$lambda.min
mean(ridge.pred == y.test)
table(ridge.pred,test_data$athetisi )
###lasso regression
cv.lasso <- cv.glmnet(x.train, y.train, alpha = 1, family = "binomial")
```

```
lasso.model<-glmnet(x.train, y.train, alpha=1, family ="binomial",
                    lambda = cv.lasso$lambda.min)
lasso.probs<-predict(lasso.model,s=cv.lasso$lambda.min,type="response",
newx=x.test)
```

Κώδικας για την κατασκευή και ερμηνεία των δέντρων αποφάσεως:

```
###DECISION TREES



##classification tree
tree2<-rpart(athetisi~.,data=training_data,method='class')
rpart.plot(tree2, extra= 106,main='Decision Tree')

#pruned tree
pruned<-prune(tree2,cp=0.01)
rpart.plot(pruned, extra= 106,main='Pruned Decision Tree')

##tree predictions
tree.pred <- predict(pruned, newdata = test_data[,-5] ,
                     type = "prob")[, 2]
##random forest
random.forest.model<-randomForest(athetisi~.,data=training_data,
proximity=TRUE)
rpart.plot(random.forest.model, extra= 106)
##random forest predictions
random.forest.pred<-predict(random.forest.model,newdata= test_data[,-5]
                            , type = "prob")[, 2]
```