ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

«Μοντέλα Πιστωτικού Κινδύνου και Εφαρμογές με Χρήση της R»

Κωνσταντίνος Παπαγιάννης

Επιβλέπουσα:
Χρυσηίς Καρώνη, Καθηγήτρια Ε.Μ.Π.

Μέλη Επιτροπής:
Χρήστος Κουκουβίνος, Καθηγητής Ε.Μ.Π.
Βασίλειος Παπανικολάου, Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2019

Ευχαριστίες

Εκφράζω τις ευχαριστίες μου στην επιβλέπουσα καθηγήτρια της παρούσας εργασίας, κ. Καρώνη, για την ευκαιρία που μου έδωσε να αναλάβω μία εργασία με ένα τόσο ενδιαφέρον θέμα, καθώς και για τις πολύτιμες συμβουλές της καθ΄ όλη την διάρκεια εκπόνησης της εργασίας. Επίσης, εκφράζω τις ευχαριστίες μου σε όλους τους καθηγητές μου για τις σπουδαίες γνώσεις που μού μεταλαμπάδευσαν. Τέλος, ευχαριστώ την οικογένειά μου για την στήριξή της καθ΄ όλη την διάρκεια των σπουδών μου.

Περίληψη

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη μοντέλων πιστωτικού κινδύνου. Αυτό επιτυγχάνεται μέσω

- της ιστορικής, φιλοσοφικής και πρακτικής ανάλυσης των εννοιών της πίστωσης και της βαθμολόγησης πιστοληπτικής ικανότητας (Κεφάλαιο 1)
- της μελέτης μεθόδων αξιολόγησης πιστοληπτικής ικανότητας που περιλαμβάνουν την λογιστική παλινδρόμηση και τα δένδρα απόφασης, καθώς και την διακριτική ανάλυση, τα νευρωνικά δίκτυα και την ανάλυση επιβίωσης (Κεφάλαιο 2)
- της κατασκευής και αξιολόγησης μοντέλων λογιστικής παλινδρόμησης και δένδρων απόφασης, με χρήση δεδομένων δανειοληπτών και της γλώσσας προγραμματισμού R (Κεφάλαιο 3)

Abstract

The goal of this project is the study of credit risk models. This is achieved through

- the historical, philosophical and practical analysis of the meanings of credit and credit scoring (Chapter 1)
- the study of credit evaluation methods including logistic regression and decision trees, as well as discriminant analysis, neural networks and survival analysis (Chapter 2)
- the building and evaluation of logistic regression and decision tree models, using data on borrowers and employing the programming language R (Chapter 3)

Περιεχόμενα

(συνέχεια)

1. ΕΙΣΑΓΩΓΗ

Το Κεφάλαιο 1 περιέχει ιστορικές, φιλοσοφικές και πρακτικές πληροφορίες σχετικώς με την πίστωση και τον πιστωτικό κίνδυνο.

1.1 Ιστορικά και φιλοσοφικά στοιχεία

Η ενότητα 1.1 παρουσιάζει τα ιστορικά και φιλοσοφικά στοιχεία.

1.1.1 Η ιστορία της πίστωσης

Ως «πίστωση» («credit») ορίζεται μία συμβατική συμφωνία (ανάμεσα σε έναν δανειστή και έναν 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[1] (Investopedia, 2019).

Θα μπορούσε, ευλόγως, να υποθέσει κανείς ότι το φαινόμενο της πίστωσης ξεκίνησε από τότε που οι άνθρωποι άρχισαν να αναπτύσσουν τις συναλλαγές και το εμπόριο μεταξύ τους και ότι, με την πάροδο των ετών, απέκτησε διαφορετικά χαρακτηριστικά και ποικίλες μορφές. Αυτό, πράγματι, συνέβη και είναι χαρακτηριστική η σχετική αναδρομή στην ιστορία της πίστωσης που επιχειρείται παρακάτω (Thomas et al., 2002).

Το πρώτο καταγεγραμμένο παράδειγμα πίστωσης προέρχεται από την αρχαία Βαβυλώνα: πάνω σε μία πέτρινη πλάκα που χρονολογείται το 2000 π.Χ. περίπου, υπάρχει η επιγραφή: *«Δύο ασημένια νομίσματα έχει δανειστεί oMas - Schamach, υιός του Adadrimeni, από την ιέρεια του Ήλιου, Amat - Schamach, κόρη του Warad - Enlil. Θα πληρώσει τον τόκο του Θεού Ήλιου. Κατά την περίοδο της συγκομιδής, θα αποπληρώσει το ποσό και τον τόκο του».*

Έως την περίοδο της ελληνικής και της ρωμαϊκής αυτοκρατορίας, τα τραπεζικά και πιστωτικά ιδρύματα βρίσκονταν σε προχωρημένο στάδιο. Κατά τα επόμενα χίλια χρόνια, τα «Σκοτεινά Χρόνια» (Μεσαίωνας) της ευρωπαϊκής ιστορίας, υπήρξε μικρή ανάπτυξη στην πίστωση, αλλά ενεχυροδανειστήρια είχαν αναπτυχθεί μέχρι την εποχή των Σταυροφοριών στον 13^{ο} αιώνα. Κατ' αρχάς, επρόκειτο για φιλανθρωπικές οργανώσεις οι οποίες δεν χρέωναν τόκο, αλλά οι έμποροι διέβλεψαν γρήγορα τις δυνατότητες και, μέχρι το 1350, εμπορικά ενεχυροδανειστήρια που χρέωναν τόκο υπήρχαν σε όλη την Ευρώπη. Κατά την διάρκεια του Μεσαίωνα, υπήρχε μία διαρκής διαμάχη περί της ηθικής στην επιβολή τόκων σε δάνεια - μία διαμάχη η οποία συνεχίζεται σήμερα στις ισλαμικές χώρες. Το αποτέλεσμα αυτής της διαμάχης στην Ευρώπη

[1] Χρηματική αποζημίωση λόγω της δανειοδότησης. Σε περίπτωση δανεισμού χρημάτων, ο τόκος προκύπτει βάσει ενός ποσοστού, του επιτοκίου, που εφαρμόζεται επί της αξίας του χρηματικού ποσού που δανείστηκε.

ήταν ότι εάν ο δανειστής εισέπραττε μικρές χρεώσεις, τότε αυτό ήταν τόκος και ήταν αποδεκτός, αλλά μεγάλες χρεώσεις συνιστούσαν τοκογλυφία και αυτό ήταν κάτι κακό. Εκείνη την εποχή, επίσης, οι βασιλείς και οι ηγεμόνες ξεκίνησαν να δανείζονται προκειμένου να χρηματοδοτούν τους πολέμους τους. Ο δανεισμός σε αυτό το επίπεδο αποτελούσε περισσότερο ζήτημα πολιτικής παρά επιχειρηματικότητας.

Η άνοδος των μεσαίων τάξεων, κατά τον 19^{ο} αιώνα, οδήγησε στην δημιουργία ενός αριθμού ιδιωτικών τραπεζών, οι οποίες ήταν πρόθυμες να δώσουν τραπεζικές υπεραναλήψεις[2] προκειμένου να χρηματοδοτήσουν επιχειρήσεις και δαπάνες διαβίωσης. Ωστόσο, αυτή η καταναλωτική πίστωση περιορίστηκε σε ένα πολύ μικρό ποσοστό του πληθυσμού.

Η πραγματική επανάσταση ξεκίνησε την δεκαετία του 1920, όταν οι καταναλωτές άρχιζαν να αγοράζουν αυτοκίνητα. Εδώ υπήρξε για πρώτη φορά ένα αντικείμενο το οποίο ήθελαν οι περισσότεροι καταναλωτές και ήταν πολύ ευέλικτο για να κατατεθεί ως εγγύηση για ασφάλιση (όπως στον δανεισμό των ενεχυροδανειστηρίων) ή ακόμη και να χρησιμοποιηθεί ως ασφάλιση (όπως η γη και η ιδιοκτησία, των οποίων την τοποθεσία γνωρίζει συνεχώς ο δανειστής). Επιχειρήσεις χρηματοδότησης αναπτύχθηκαν προκειμένου να ανταποκριθούν σε αυτή την ανάγκη και γνώρισαν ραγδαία ανάπτυξη πριν τον Β’ Παγκόσμιο Πόλεμο.

Την ίδια στιγμή, οι εταιρείες ταχυδρομικών παραγγελιών άρχισαν να αναπτύσσονται, καθώς καταναλωτές σε μικρότερες πόλεις ζητούσαν ρουχισμό και οικιακά είδη τα οποία ήταν διαθέσιμα μόνο σε μεγάλα πληθυσμιακά κέντρα. Αυτά διαφημίζονταν σε καταλόγους και οι εταιρείες ήταν πρόθυμες να αποστείλουν τα αγαθά με πίστωση και να επιτρέψουν στους πελάτες να αποπληρώσουν σε μία εκτεταμένη περίοδο.

Κατά το τελευταίο μισό του 20^{ου} αιώνα, ο δανεισμός στους καταναλωτές είχε εκτοξευθεί στα ύψη. Η καταναλωτική πίστωση είχε έναν από τους μεγαλύτερους ρυθμούς ανάπτυξης σε κάθε επιχειρηματικό τομέα. Η έλευση πιστωτικών καρτών, την δεκαετία του 1960, ήταν μία από τις πιο εμφανείς ενδείξεις αυτής της ανάπτυξης και υπήρξε, πλέον, δύσκολο να λειτουργήσει κανείς, σε μία κοινωνία, χωρίς πιστωτική κάρτα.

Στην σύγχρονη εποχή, το φαινόμενο της πίστωσης έχει λάβει γιγαντιαίες διαστάσεις και παρουσιάζεται σε πολλές μορφές: πιστωτικές κάρτες, αγορές με δόσεις, υπεραναλήψεις, ενυπόθηκος δανεισμός. Μάλιστα, σε κάποιες περιπτώσεις, δεν είναι δυνατή η αγορά ενός αγαθού παρά μόνο με πίστωση: για παράδειγμα, ενδέχεται η αγορά ενός εμπορεύματος από

[2] Επεκτάσεις δανεισμού μέχρι ενός συγκεκριμένου ορίου, οι οποίες προσφέρονται από μία τράπεζα σε πελάτες της δίδοντάς τους την δυνατότητα να συνεχίσουν να πραγματοποιούν αναλήψεις χρημάτων αν και το υπόλοιπο του τραπεζικού λογαριασμού τους είναι μηδενικό.

κάποιο ηλεκτρονικό κατάστημα να πραγματοποιείται μόνο μέσω πιστωτικής κάρτας. Βεβαίως, πίστωση δεν λαμβάνουν μόνο οι άνθρωποι, αλλά και οι επιχειρήσεις και τα κράτη.

1.1.2. Η ιστορία της βαθμολόγησης πιστοληπτικής ικανότητας

Ως «πιστοληπτική ικανότητα / διαβάθμιση» («credit rating») ορίζεται η αξιοπιστία και η φερεγγυότητα ενός ατόμου ή μίας επιχείρησης ή ακόμη και ενός κράτους σχετικώς με την αποπληρωμή των χρεών που έχει (Ευρετήριο οικονομικών όρων, 2019).

Ως «βαθμολόγηση πιστοληπτικής ικανότητας» («credit scoring») ορίζεται η εφαρμογή στατιστικών τεχνικών στον καταναλωτικό δανεισμό, με σκοπό την έγκριση πίστωσης αφ' ενός και την παρακολούθηση πίστωσης αφ' ετέρου (Business Dictionary, 2019).

Ακολούθως, λοιπόν, ως σύστημα βαθμολόγησης πιστοληπτικής ικανότητας («credit scoring system») ορίζεται ένα αριθμητικό σύστημα που «μετρά» πόσο πιθανό είναι ένας δανειολήπτης να αποπληρώσει σχετικώς με τα χρήματα που δανείστηκε και, δημιουργείται αποδίδοντας πόντους σε διάφορα χαρακτηριστικά που σχετίζονται με την πιστοληπτική ικανότητα του αιτούντος (Bankrate, 2019).

Τέλος, ο αριθμός που προκύπτει μετά την διαδικασία της βαθμολόγησης πιστοληπτικής ικανότητας καλείται «βαθμός πιστοληπτικής ικανότητας» («credit score»).

Από τα παραπάνω, λοιπόν, προκύπτει ότι η βαθμολόγηση πιστοληπτικής ικανότητας αποτελεί ένα σύνολο στατιστικών μεθοδολογιών και τεχνικών που αναπτύσσονται από ειδικούς επιστήμονες και χρησιμοποιούνται από προμηθευτές πίστωσης (για παράδειγμα, τράπεζες) προκειμένου να εκτιμηθεί εάν ένας υποψήφιος δανειολήπτης δύναται να ανταποκριθεί στις οικονομικές υποχρεώσεις που προτίθεται (μέσω της δανειοδότησης) να αναλάβει. Βεβαίως, οι ακόλουθες τελικές αποφάσεις που θα λάβουν οι προμηθευτές πίστωσης είναι συγκεκριμένες - εάν ο υποψήφιος δανειολήπτης θα λάβει, όντως, πίστωση και πόση - και, επίσης, σκοπό τους αποτελεί η επίτευξη κερδοφορίας.

Οι προαναφερθείσες μεθοδολογίες και τεχνικές προαπαιτούν την ύπαρξη ενός μεγάλου δείγματος πελατών και την διαθεσιμότητα σχετικών στοιχείων τόσο της αιτήσεώς τους όσο και του πιστωτικού ιστορικού τους, προκειμένου να εντοπιστούν μοτίβα και σχέσεις τόσο ανάμεσα στους «καλούς» δανειολήπτες όσο και ανάμεσα στους «κακούς». Τελικώς, εάν ο προκύπτων βαθμός ενός υποψήφιου δανειολήπτη είναι μικρότερος από κάποιο αριθμητικό όριο, τότε πιθανολογείται πως ο κίνδυνος δανεισμού του είναι σχετικώς μικρός και, επομένως, ο προμηθευτής πίστωσης θα προβεί στον δανεισμό μάλλον. Στην αντίθετη περίπτωση, ο κίνδυνος αξιολογείται ως μεγάλος και πιθανολογείται ότι ο προμηθευτής πίστωσης δεν θα προβεί στον δανεισμό.

Όπως φάνηκε προηγουμένως, η ιστορία της πίστωσης είναι χιλιάδων ετών. Η ιστορία της βαθμολόγησης πιστοληπτικής ικανότητας, όμως, είναι μερικών δεκάδων ετών. Η έναρξή της εδράζεται στην παρατήρηση πως η ίδια αποτελεί, βασικώς, έναν τρόπο αναγνώρισης διακριτών ομάδων σε έναν πληθυσμό. Εντυπωσιακή είναι η ακόλουθη σχετική ιστορική αναδρομή (Thomas et al., 2002).

Η πρώτη προσέγγιση για επίλυση του προβλήματος της αναγνώρισης ομάδων σε έναν πληθυσμό εισήχθη στην Στατιστική όταν ο Fisher (1936) επεδίωξε την διαφοροποίηση ανάμεσα σε δύο ποικιλίες ίριδας μέσω των διαστάσεων του φυσικού μεγέθους των φυτών και την διαφοροποίηση της προέλευσης των κεφαλιών χρησιμοποιώντας τις φυσικές τους διαστάσεις. Παρατηρήθηκε από τον Durand (1941), όμως, ότι κάποιος μπορούσε να χρησιμοποιήσει τις ίδιες τεχνικές για να πραγματοποιήσει τον διαχωρισμό ανάμεσα στα «καλά» και τα «κακά» δάνεια.

Κατά την διάρκεια της δεκαετίας του 1930, κάποιες εταιρείες ταχυδρομικών παραγγελιών εισήγαγαν αριθμητικά συστήματα βαθμολόγησης προσπαθώντας να ξεπεράσουν τις ασυνέπειες στις πιστοδοτικές αποφάσεις μεταξύ των αναλυτών πιστωτικού κινδύνου.

Με την έναρξη του Β’ Παγκοσμίου Πολέμου, όλοι οι χρηματοπιστωτικοί οίκοι και οι εταιρείες ταχυδρομικών παραγγελιών άρχισαν να αντιμετωπίζουν δυσκολίες με την διαχείριση της πίστωσης. Οι αναλυτές πιστωτικού κινδύνου απορροφήθηκαν στις στρατιωτικές υπηρεσίες και υπήρξε μία σοβαρή έλλειψη ανθρώπων με αυτή την εξειδίκευση. Ως εκ τούτου, οι εταιρείες ζήτησαν από τους αναλυτές να καταγράψουν τους εμπειρικούς κανόνες που χρησιμοποιούσαν για να αποφασίσουν σε ποιούς θα δοθούν δάνεια. Κάποιοι από αυτούς τους κανόνες ήταν τα αριθμητικά συστήματα βαθμολόγησης που είχαν εισαχθεί ήδη. Άλλοι ήταν σειρές από συνθήκες που έπρεπε να ικανοποιηθούν. Αυτοί οι κανόνες, έπειτα, χρησιμοποιήθηκαν από μη ειδικούς προκειμένου να βοηθήσουν στην λήψη πιστοδοτικών αποφάσεων.

Σε όχι μεγάλο χρονικό διάστημα μετά το τέλος του πολέμου, κάποιοι συνέδεσαν την αυτοματοποίηση των πιστοδοτικών αποφάσεων με τις τεχνικές ταξινόμησης που είχαν αναπτυχθεί στην Στατιστική και είδαν το όφελος χρήσης μοντέλων προερχόμενων από την Στατιστική σε αποφάσεις δανεισμού.

Η έλευση των πιστωτικών καρτών, στα τέλη της δεκαετίας του 1960, οδήγησε τις τράπεζες και άλλες εταιρείες έκδοσης πιστωτικών καρτών στο να συνειδητοποιήσουν την χρησιμότητα της βαθμολόγησης πιστοληπτικής ικανότητας. Ο αριθμός των ανθρώπων που αιτούντο πιστωτικές κάρτες σε καθημερινή βάση κατέστησε αδύνατο - σε οικονομικούς όρους και σε όρους σε

ανθρώπινου δυναμικού - οτιδήποτε άλλο πλην της αυτοματοποίησης της απόφασης δανεισμού. Η ανάπτυξη της υπολογιστικής ισχύος κατέστησε αυτήν δυνατή.

Οι παραπάνω οργανισμοί θεώρησαν ότι η βαθμολόγηση πιστοληπτικής ικανότητας αποτελεί ένα πολύ καλύτερο μέσο πρόβλεψης απ' ό, τι οποιοδήποτε άλλο σχέδιο κρίσης, και τα ποσοστά αθέτησης υποχρεώσεων μειώθηκαν κατά 50% ή περισσότερο. Η μόνη εναντίωση προήλθε από κάποιους που υποστήριξαν ότι «*η ωμή βία του εμπειρισμού της βαθμολόγησης πιστοληπτικής ικανότητας καταπατά τις παραδόσεις της κοινωνίας μας».* Οι ίδιοι πίστευαν ότι έπρεπε να υπάρχει μεγαλύτερη εξάρτηση από το πιστωτικό ιστορικό και ότι θα έπρεπε να είναι δυνατό να εξηγηθεί γιατί ορισμένα χαρακτηριστικά χρειάζονται σε ένα σύστημα βαθμολόγησης ενώ κάποια άλλα όχι.

Κατά την δεκαετία του 1980, η επιτυχία της βαθμολόγησης πιστοληπτικής ικανότητας στις πιστωτικές κάρτες οδήγησε τις τράπεζες στο να ξεκινήσουν να χρησιμοποιούν την βαθμολόγηση και σε άλλα προϊόντα, όπως προσωπικά δάνεια. Η πρόοδος στους υπολογισμούς επέτρεψε την χρήση προηγμένων τεχνικών στην κατασκευή πινάκων βαθμολογίας, όπως είναι η λογιστική παλινδρόμηση[3].

Κατά την δεκαετία του 1990, η ανάπτυξη του άμεσου μάρκετινγκ οδήγησε στην χρήση πινάκων βαθμολογίας ώστε να βελτιωθεί το ποσοστό ανταπόκρισης στις διαφημιστικές καμπάνιες.

Μετέπειτα, εισήχθησαν πίνακες βαθμολογίας που εκτιμούν την ανταπόκριση (το πόσο πιθανό είναι ένας καταναλωτής να ανταποκριθεί σε άμεση ταχυδρόμηση ενός καινούργιου προϊόντος), την χρήση (το πόσο πιθανό είναι ένας καταναλωτής να χρησιμοποιήσει ένα προϊόν), την διατήρηση (το πόσο πιθανό είναι ένας καταναλωτής να συνεχίσει να χρησιμοποιεί το προϊόν αφού ολοκληρωθεί η περίοδος διαφημιστικής προσφοράς), την απώλεια (το εάν θα αλλάξει δανειστή ο καταναλωτής), την διαχείριση χρέους (το πόσο πιθανές είναι διάφορες προσεγγίσεις ώστε να αποτραπεί η αθέτηση των υποχρεώσεων εάν ο καταναλωτής αρχίσει να καθίσταται αμελής ως προς το δάνειο) και την βαθμολόγηση απάτης (το πόσο πιθανό είναι η αίτηση να είναι δόλια).

Στην σύγχρονη εποχή, φαίνεται πως - από πλευράς προμηθευτών πίστωσης - επιχειρείται μία διαφορετική αντιμετώπιση: έχει παραγκωνιστεί ο επιμέρους στόχος του πώς θα αποφευχθεί η αθέτηση των υποχρεώσεων ενός πελάτη ως προς ένα συγκεκριμένο προϊόν και έχει ενισχυθεί ο στόχος του πώς θα επιτευχθεί η μεγαλύτερη δυνατή κερδοφορία από αυτόν τον πελάτη εν συνόλω.

[3] Οι μέθοδοι της λογιστικής παλινδρόμησης έχουν καταστεί ένα σημαντικότατο συστατικό κάθε ανάλυσης δεδομένων που σχετίζεται με την περιγραφή της σχέσης ανάμεσα σε μία (δίτιμη) μεταβλητή απόκρισης (εξαρτημένη μεταβλητή) και μία ή περισσότερες ανεξάρτητες μεταβλητές (Abdou & Pointon, 2011).

1.1.3 Η φιλοσοφική προσέγγιση της βαθμολόγησης πιστοληπτικής ικανότητας

Η βαθμολόγηση πιστοληπτικής ικανότητας εδράζεται στον πραγματισμό και τον εμπειρισμό, δεδομένου ότι χρησιμοποιεί πραγματικά δεδομένα και επιστημονικές μεθόδους και τεχνικές προκειμένου να ποσοτικοποιήσει την εκτίμηση διαφόρων μορφών κινδύνου: να αθετηθούν οι υποχρεώσεις ενός δανείου ή να υπάρξει δόλια συμπεριφορά από πλευράς ενός δανειολήπτη, να μην υπάρξει ανταπόκριση στην διαφήμιση και στα προωθητικά εγχειρήματα, γενικώς, ενός προϊόντος, να αποτανθεί σε άλλον προμηθευτή πίστωσης ένας υποψήφιος δανειολήπτης.

Όπως αναφέρθηκε προηγουμένως, τα συστήματα βαθμολόγησης πιστοληπτικής ικανότητας χρησιμοποιούν ένα μεγάλο δείγμα πελατών με λεπτομέρειες της αιτήσεώς τους και του πιστωτικού ιστορικού τους προκειμένου να διακρίνουν μοτίβα και σχέσεις ανάμεσα στους «καλούς» και ανάμεσα στους «κακούς» δανειολήπτες. Πιο συγκεκριμένα, οι πελάτες που θα ληφθούν υπ' όψιν θα πρέπει, αφού εξετασθούν ορισμένα χαρακτηριστικά τους - όπως εισόδημα, ηλικία, αριθμός δανείων, οικογενειακή κατάσταση - να εμφανίζουν ικανοποιητική ομοιότητα με εκείνους που θα αξιολογηθούν στο πλαίσιο του συστήματος (τους υποψήφιους δανειολήπτες δηλαδή). Επίσης, οι πελάτες αυτοί θα πρέπει να έχουν αιτηθεί το ίδιο προϊόν και, μάλιστα, κατά το πρόσφατο παρελθόν ώστε τα δεδομένα που θα ληφθούν να οδηγήσουν σε πιο αντιπροσωπευτικά αποτελέσματα.

Εάν δεν είναι εφικτό να ικανοποιούνται όλα τα παραπάνω προαπαιτούμενα, τότε δύναται, και πάλι, να επιχειρηθεί η αξιολόγηση ενός υποψήφιου δανειολήπτη, αλλά η πρόβλεψη δεν θα είναι τόσο καλή. Για παράδειγμα, εάν το δείγμα των εξεταζομένων υπαρχόντων πελατών δεν είναι επαρκώς μεγάλο ή εάν λίγοι εξ' αυτών παρουσιάζουν ομοιότητα με τον υποψήφιο δανειολήπτη ή εάν το αιτούμενο πιστωτικό προϊόν είναι καινούργιο και, επομένως, η εξέταση θα πραγματοποιηθεί βάσει ενός παρεμφερούς προϊόντος για το οποίο υπάρχουν, πράγματι, δεδομένα, τότε ελλοχεύει ο κίνδυνος μίας κακής πρόβλεψης.

Είναι προφανές, λόγω του χαρακτήρα της, ότι η βαθμολόγηση πιστοληπτικής ικανότητας θα πρέπει να λαμβάνει υπ' όψιν οποιοδήποτε χαρακτηριστικό του υποψήφιου δανειολήπτη, αρκεί να πιθανολογείται ότι θα συνεισφέρει στην βελτίωση της επιχειρούμενης πρόβλεψης. Κάθε τέτοιο χαρακτηριστικό αντιπροσωπεύεται - στο σύστημα βαθμολόγησης πιστοληπτικής ικανότητας - από μία μεταβλητή. Συνήθεις μεταβλητές, λοιπόν, είναι: εισόδημα μετά φόρους, συνολικό εισόδημα νοικοκυριού, αριθμός δανείων και συνολικός δανεισμός, ηλικία, οικογενειακή κατάσταση, χρόνος στην παρούσα θέση εργασίας, θέση εργασίας συζύγου, αριθμός τέκνων, χρόνος στην παρούσα κατοικία, χρόνος τήρησης τραπεζικών λογαριασμών, χρόνος συνεργασίας με τον συγκεκριμένο προμηθευτή πίστωσης. Είναι προφανής η σύνδεση

των παραπάνω μεταβλητών με τον κίνδυνο αθέτησης των υποχρεώσεων από πλευράς του υποψήφιου δανειολήπτη, καθώς συνδέονται - είτε αμέσως είτε εμμέσως - με το χρηματοοικονομικό προφίλ του και τις προοπτικές αυτού.

Σε αυτό το σημείο, θα πρέπει να επισημανθεί πως τίθεται ένα ζήτημα σχετικώς με το κατά πόσο είναι ηθική η χρήση δεδομένων ατόμων, από πλευράς προμηθευτών πίστωσης, προκειμένου οι τελευταίοι να λάβουν τις αποφάσεις παροχής πίστωσης. Σε πολλές περιπτώσεις, μάλιστα, είναι αμφίβολο το κατά πόσο τα εν λόγω άτομα είναι ενήμερα ότι προσωπικά στοιχεία και πληροφορίες τους πρόκειται να χρησιμοποιηθούν από τρίτους. Σε πολλές χώρες, έχουν θεσπιστεί νόμοι προκειμένου να ρυθμίσουν το ζήτημα. Επιπλέον, κάποια δεδομένα, ακόμη κι εάν δεν απαγορεύονται νομικώς, είθισται να μην χρησιμοποιούνται, καθώς είναι πιθανό η χρήση τους να προκαλέσει κοινωνικές αντιδράσεις. Ενδεικτικό παράδειγμα αποτελούν τα ιατρικά δεδομένα ατόμων. Όπως παρατίθεται παρακάτω (Thomas et al., 2002), στο Ηνωμένο Βασίλειο έχουν υπάρξει έντονες προστριβές σχετικώς με την χρήση δεδομένων αφ' ενός και με την αποτελεσματικότητα και αξιοπιστία της βαθμολόγησης πιστοληπτικής ικανότητας αφ' ετέρου.

Πιο συγκεκριμένα, στο Ηνωμένο Βασίλειο έχει ξεσπάσει μία σφοδρή διαμάχη με την γραμματεία προστασίας δεδομένων σχετικώς με το εάν θα μπορούσαν να χρησιμοποιηθούν πληροφορίες σχετικές με ανθρώπους που έχουν ζήσει στην ίδια ταχυδρομική διεύθυνση με έναν καταναλωτή. Αυτή η διαμάχη υπογραμμίζει ότι η χρήση κάποιων χαρακτηριστικών - όπως φυλή, θρησκεία και φύλο - σε συστήματα βαθμολόγησης πιστοληπτικής ικανότητας είναι παράνομη. Παρεμπιπτόντως, ένας αριθμός μελετών έχει δείξει ότι εάν επιτρεπόταν η χρήση του φύλου, τότε περισσότερες γυναίκες θα ελάμβαναν πίστωση. Αυτό συμβαίνει επειδή άλλες μεταβλητές, όπως το χαμηλό εισόδημα και η θέση εργασίας μερικής απασχόλησης, είναι μέσα πρόβλεψης συμπεριφοράς καλής αποπληρωμής στις γυναίκες αλλά συμπεριφοράς πτωχής αποπληρωμής σε ολόκληρο τον πληθυσμό. Ωστόσο, οι νομοθέτες δεν επιτρέπουν την χρήση του φύλου διότι πιστεύουν πως αυτό θα αποτελέσει διάκριση σε βάρος των γυναικών.

Παρά τις θετικές επιδράσεις του πιστωτικού κινδύνου, λοιπόν, υπήρξαν ορισμένοι που εναντιώθηκαν στις νέες αυτές τεχνικές, όπως ο Capon (1982). Υπήρξαν διαμάχες, κατά τις αρχές της δεκαετίας του 1980, σχετικώς με την ηθική της βαθμολόγησης πιστοληπτικής ικανότητας, ανάμεσα σε εκείνους που την υποστήριξαν και εκείνους που άσκησαν κριτική ως προς την φιλοσοφία και υλοποίησή της σε σχέση με τα υποκειμενικά επικριτικά συστήματα που βασίζονταν σε αναλυτές πιστωτικού κινδύνου και απόψεις ασφαλιστικών εταιρειών.

Οι πρώτοι περιέγραψαν τα πλεονεκτήματα της βαθμολόγησης πιστοληπτικής ικανότητας, όπως η ικανότητά της να μεγιστοποιεί το αντιστάθμισμα κινδύνου - απόδοσης, το ότι προσδίδει

έλεγχο διαχείρισης αυτού του αντισταθμίσματος και το ότι είναι αποτελεσματική στην επεξεργασία των αιτήσεων. Ισχυρίστηκαν ότι η βαθμολόγηση πιστοληπτικής ικανότητας μείωσε την ανάγκη για πιστοληπτικές έρευνες, αλλά αυτό δεν ισχύει στην πραγματικότητα: η αναζήτηση τραπεζικής σύστασης μειωνόταν ούτως ή άλλως, καθώς αυτό επέτρεπε στην τράπεζα του αιτούντος την ευκαιρία να πραγματοποιήσει αντιπροσφορά (Thomas et al., 2002). Υποστήριξαν, επίσης, την συνέπειά της απέναντι στους καταναλωτές και επεσήμαναν ότι βελτίωσε τις πληροφορίες που διατίθενται στους λογαριασμούς, καθώς και την ποιότητα του συνολικού χαρτοφυλακίου των λογαριασμών.

Εκείνοι που αντιπαρατέθηκαν στην βαθμολόγηση της πιστοληπτικής ικανότητας επιτέθηκαν στην φιλοσοφία της και στην αξιοπιστία της μεθοδολογίας της. Άσκησαν κριτική ως προς το γεγονός ότι δεν έδιδε εξηγήσεις σχετικώς με τους συνδέσμους ανάμεσα στα χαρακτηριστικά που θεωρούσε σημαντικά και στην ακόλουθη πιστοληπτική συμπεριφορά. Υποστήριξαν ότι υπήρχε μία πολύπλοκη αλυσίδα αλληλεπιδρουσών μεταβλητών που συνέδεε τα αρχικά χαρακτηριστικά με την συμπεριφορά. Η αξιοπιστία της στατιστικής μεθοδολογίας επικρίθηκε λόγω μεροληψίας στο χρησιμοποιούμενο δείγμα και, επίσης, αμφισβητήθηκε το κατάλληλο μέγεθος δείγματος και επισημάνθηκε η συγγραμμικότητα ανάμεσα στις μεταβλητές (Thomas et al., 2002).

Σε αυτό το σημείο, θα πρέπει να επισημανθεί ότι παρά τις αναφερθείσες διαμάχες, συστήματα βαθμολόγησης έχουν εισαχθεί σε πολλά και διαφορετικά πιστωτικά προϊόντα από πληθώρα προμηθευτών πίστωσης ανά τον κόσμο. Στην πράξη, λοιπόν, αποδεικνύεται μάλλον πως η χρήση βαθμολόγησης επιφέρει θετικά αποτελέσματα ως προς την ελαχιστοποίηση των κινδύνων και την μεγιστοποίηση των κερδών από πλευράς πιστωτών. Εξ' άλλου, οι όποιες αδυναμίες της βαθμολόγησης - αν και εξακολουθούν να επισημαίνονται - φαίνεται πως λαμβάνονται υπ' όψιν από τους επιστήμονες οι οποίοι έχουν επινοήσει τρόπους παράκαμψής τους.

Τέλος, ας παρατηρηθεί ότι ο σκοπός της βαθμολόγησης πιστοληπτικής ικανότητας είναι μόνο η πρόβλεψη κινδύνου, όχι η τεκμηρίωσή του. Ως εκ τούτου, δεν είναι αναγκαίο ένα μοντέλο πρόβλεψης κινδύνου να εξηγεί πώς καταλήγει στα συμπεράσματά του και, επομένως, δεν γνωστοποιείται το γιατί πιθανολογείται πως κάποιος υποψήφιος δανειολήπτης θα αθετήσει τις υποχρεώσεις του ή όχι. Με αυτό το σκεπτικό, μία μεταβλητή συμπεριλαμβάνεται στο μοντέλο εάν έχει παρατηρηθεί ή πιθανολογείται πως θα βελτιώσει την πρόβλεψη δίχως να εξετάζεται το πώς ακριβώς.

1.2 Πρακτική προσέγγιση

Η ενότητα 1.2 προσεγγίζει πρακτικώς τις έννοιες της πίστωσης και του πιστωτικού κινδύνου.

1.2.1 Η εκτίμηση της πιστοληπτικής ικανότητας πριν και μετά την βαθμολόγηση

Κατ' αρχάς, η εκτίμηση της πιστοληπτικής ικανότητας ενός υποψήφιου δανειολήπτη πραγματοποιείτο βάσει του ενστίκτου εκείνου που καλείτο να λάβει την απόφαση της πίστωσης, αλλά ορισμένοι παράγοντες - συν τω χρόνω - συνέβαλαν στο να αντικατασταθεί αυτή η πρακτική από εκείνη της βαθμολόγησης. Οι Thomas et al. (2002) δίδουν σχετικές λεπτομέρειες, όπως καταγράφονται ακολούθως.

Όχι πολύ παλαιά - σίγουρα κατά την δεκαετία του 1970 στο Ηνωμένο Βασίλειο και στις Ηνωμένες Πολιτείες και ίσως, για λίγους δανειστές, ακόμη και στα τέλη της δεκαετίας του 1990 - η βαθμολόγηση της πιστοληπτικής ικανότητας δεν χρησιμοποιείτο. Η παραδοσιακή εκτίμηση της πιστοληπτικής ικανότητας βασιζόταν στην «έντονη αίσθηση» και σε μία εκτίμηση του χαρακτήρα του υποψήφιου δανειολήπτη, της ικανότητας αποπληρωμής, και της εγγύησης ή ασφάλισης. Αυτό σήμαινε πως ένας υποψήφιος δανειολήπτης δεν προσέγγιζε έναν διευθυντή τράπεζας ή χρηματοοικονομικού οργανισμού έως ότου να έχει πραγματοποιήσει αποταμιεύσεις ή χρησιμοποιήσει άλλες υπηρεσίες για αρκετά χρόνια.

Ο διευθυντής σκεπτόταν την πρόταση και, ανεξαρτήτως της διάρκειας της σχέσης με τον πελάτη, υπολόγιζε την πιθανότητα της αποπληρωμής και εκτιμούσε την σταθερότητα και τιμιότητα του ατόμου και του χαρακτήρα του. Αξιολογούσε, επίσης, την προτεινόμενη χρήση των χρημάτων και, έπειτα, ίσως ζητούσε μία ανεξάρτητη σύσταση από κάποιον επικεφαλής κοινότητας ή από τον εργοδότη του αιτούντος. Ίσως κανόνιζε ένα περαιτέρω ραντεβού με τον πελάτη και, έπειτα, ενδεχομένως ελάμβανε μία απόφαση και ενημέρωνε τον πελάτη. Αυτή η διαδικασία ήταν σχετικώς αργή και απρόβλεπτη.

Κατά την διάρκεια της δεκαετίας του 1980, στο Ηνωμένο Βασίλειο, πολλές αλλαγές προέκυψαν στο περιβάλλον δανεισμού. Μερικές από αυτές τις αλλαγές είναι οι ακόλουθες.

- Οι τράπεζες άλλαξαν αισθητά την θέση τους στην αγορά και άρχισαν να διαφημίζουν τα προϊόντα τους. Αυτό, ακολούθως, σήμαινε ότι έπρεπε να πουλήσουν προϊόντα σε πελάτες - όχι μόνο σε εκείνους που γνώριζαν αλλά και σε εκείνους που είχαν προσελκύσει.
- Υπήρξε πρωτοφανής άνοδος στις πιστωτικές κάρτες. Οι εξουσιοδοτήσεις πώλησης αυτού του προϊόντος συνεπάγονταν ότι έπρεπε να υπάρχει ένας μηχανισμός λήψης αποφάσεων δανεισμού πολύ γρήγορα και ανά πάσα στιγμή. Επίσης, οι όγκοι των

αιτήσεων ήταν τέτοιοι που ο διευθυντής τράπεζας ή ο οποιοσδήποτε άλλος εκπαιδευμένος αναλυτής πιστωτικού κινδύνου δεν είχε τον χρόνο ή την ευκαιρία να πραγματοποιήσει συνέντευξη με όλους τους αιτούντες.

- Η τραπεζική πρακτική άλλαξε έμφαση. Προηγουμένως, οι τράπεζες εστίαζαν σχεδόν αποκλειστικώς στους μεγάλους δανεισμούς και στους εταιρικούς πελάτες. Πλέον, ο καταναλωτικός δανεισμός αποτελούσε ένα σημαντικό και αναπτυσσόμενο τμήμα της τράπεζας. Με τον εταιρικό δανεισμό, ο στόχος ήταν συνήθως το να αποφευχθούν απώλειες. Ωστόσο, οι τράπεζες άρχισαν να συνειδητοποιούν ότι, με τον καταναλωτικό δανεισμό, ο στόχος δεν θα έπρεπε να είναι το να αποφευχθούν απώλειες αλλά το να μεγιστοποιηθούν τα κέρδη.

Στην σύγχρονη εποχή, λοιπόν, η βαθμολόγηση πιστοληπτικής ικανότητας χρησιμοποιείται ευρέως από πληθώρα πιστωτών. Οι πιστωτές που ενδέχεται να την αγνοούν είναι εκείνοι που προτιμούν να επενδύσουν στην πρόσληψη έμπειρων στελεχών και διευθυντών που θα λάβουν την απόφαση για το εάν θα δοθεί πίστωση ή όχι σε έναν αιτούντα.

Η βαθμολόγηση πιστοληπτικής ικανότητας, λοιπόν, χρησιμοποιείται σε διάφορες μορφές πίστωσης: στις πιστωτικές κάρτες, στις υπεραναλήψεις, στα δάνεια σταδιακής εξόφλησης, στις αγορές με δόσεις, στον ενυπόθηκο δανεισμό.

Πιο συγκεκριμένα, ο λόγος για τον οποίο η βαθμολόγηση χρησιμοποιείται ευρέως στις πιστωτικές κάρτες είναι ότι αυτές χρησιμοποιούνται κατά κόρον, από τεράστιο πλήθος καταναλωτών, οποιαδήποτε ώρα του 24ώρου και, ως εκ τούτου, απαιτούνται γρήγορες και αποδοτικές αποφάσεις που είναι σχεδόν απίθανο να ληφθούν από έναν άνθρωπο. Επιπλέον, αποφάσεις απαιτούνται και σε περαιτέρω φάσεις της πίστωσης: στο εάν θα αυξηθεί το πιστωτικό όριο ενός πελάτη - και για ποιό χρονικό διάστημα - ή όχι, καθώς και στο εάν θα επανεκδοθεί ή αλλαχθεί ή ακυρωθεί μία πιστωτική κάρτα.

Ομοίως, δεδομένου ότι πάρα πολλοί καταναλωτές χρησιμοποιούν ευρέως και ανά πάσα ώρα και στιγμή τις υπεραναλήψεις, απαιτείται ένας μηχανισμός για λήψη γρήγορων και αξιόπιστων αποφάσεων. Όπως καθίσταται εμφανές, οι εν λόγω αποφάσεις σχετίζονται με το ύψος του ορίου υπερανάληψης, καθώς και με το εάν θα επιτραπεί η πραγματοποίηση μίας ενδεχόμενης ηλεκτρονικής πληρωμής σε περίπτωση που μία τέτοια ενέργεια οδηγούσε σε υπέρβαση του ορίου υπερανάληψης. Τέλος, απόφαση απαιτείται και σε περίπτωση που κάποιος πελάτης αιτηθεί αύξηση του ορίου υπερανάληψης.

Στα δάνεια σταδιακής εξόφλησης, πρέπει να αποφασιστεί το εάν θα δοθεί το αιτούμενο δάνειο - και για ποιό χρονικό διάστημα - ή όχι, το ύψος του επιτοκίου που θα εφαρμοσθεί επί του ποσού

δανεισμού, το εάν θα πρέπει να υπάρξουν εγγυήσεις και ποιές θα είναι αυτές. Περαιτέρω αποφάσεις δύναται να απαιτηθούν σε περίπτωση που ο δανειολήπτης αμελεί την πραγματοποίηση των πληρωμών που οφείλει.

Στις αγορές με δόσεις, τα πράγματα καθίστανται λίγο πιο περίπλοκα. Τα εμπορεύματα που αγοράζονται με δόσεις είναι, συνήθως, κινητά αντικείμενα, όπως μία οικιακή συσκευή ή ένα αυτοκίνητο, τα οποία δεν ανήκουν στον καταναλωτή παρά μόνο αφού εξοφλήσει και την τελευταία δόση. Λόγω της φύσης της εν λόγω πίστωσης, ο πιστωτής αναλαμβάνει ιδιαίτερο κίνδυνο: ο καταναλωτής ενδέχεται να αποσκοπεί, απλώς, στην χρήση του εμπορεύματος για ένα χρονικό διάστημα και, έπειτα, να παύσει την πληρωμή των δόσεων. Ακόμη και στην περίπτωση που ο πιστωτής ανακτήσει το εν λόγω εμπόρευμα, τίθεται ένα ζήτημα ως προς την κατάσταση στην οποία θα βρίσκεται αυτό κατά την εν λόγω ανάκτηση και, ακολούθως, ένα ζήτημα ως προς την αξία του - ως μεταχειρισμένο - τότε. Τέλος, μιας που το εμπόρευμα είναι κινητό, ενδέχεται ο καταναλωτής να το αποκρύψει ή να παραχωρήσει την χρήση του σε τρίτο άτομο δίχως ο πιστωτής να είναι σε θέση να αποτρέψει αυτό το ενδεχόμενο ή έστω να εντοπίσει το αντικείμενο εκ των υστέρων.

Στον ενυπόθηκο δανεισμό, ο πιστωτής είναι πιο εξασφαλισμένος: ο δανειολήπτης διαθέτει ένα περιουσιακό στοιχείο που αποτελεί ιδιοκτησία του και, λόγω του δανείου που έλαβε, επωμίζεται μία (νόμιμη) χρέωση - σχετική με το εν λόγω περιουσιακό στοιχείο. Αυτή η χρέωση εξακολουθεί να υφίσταται έως ότου αποπληρώσει το δάνειο. Εάν δεν το αποπληρώσει, τότε το εν λόγω περιουσιακό στοιχείο - μετά από κάποιες νόμιμες διαδικασίες - θα περάσει στην ιδιοκτησία του πιστωτή. Περαιτέρω διασφάλιση του πιστωτή προκύπτει από το γεγονός ότι ο δανειολήπτης δεν δύναται να πουλήσει το περιουσιακό στοιχείο εάν δεν έχει αποπληρώσει το χρέος του. Ένας παράγοντας κινδύνου, ο οποίος θα πρέπει να ληφθεί υπ' όψιν από πλευράς δανειστή, είναι η ενδεχόμενη κακή κατάσταση στην οποία θα βρίσκεται το περιουσιακό στοιχείο σε περίπτωση που το αποκτήσει λόγω αθετήσεως των υποχρεώσεων του δανειολήπτη, καθώς και η ενδεχομένως χαμηλή αξία του τότε.

Ιδίως στις παραπάνω τρεις τελευταίες μορφές δανεισμού, ο πιστωτής θα πρέπει να λάβει υπ' όψιν τις ενδεχόμενες δικαστικές διαμάχες που θα προκύψουν σε περίπτωση που υπάρξουν ληξιπρόθεσμες οφειλές ή αθέτηση των υποχρεώσεων από πλευράς δανειολήπτη, καθώς αυτές ενέχουν ένα κόστος - τόσο σε οικονομικό επίπεδο όσο και σε επίπεδο χρόνου.

1.2.2 Οι σύμβουλοι

Σε περασμένες δεκαετίες, την ανάπτυξη μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας πραγματοποιούσε ένας ειδικός, ενώ μεταγενέστερα (και στην σύγχρονη εποχή βεβαίως) υπήρξαν συμβουλευτικές εταιρείες ή οργανωμένες ομάδες που ανέλαβαν αυτόν τον ρόλο.

Μέχρι τα μέσα της δεκαετίας του 1980, ένας δανειστής που ήθελε έναν πίνακα βαθμολογίας θα προσέγγιζε έναν δημιουργό πινάκων βαθμολογίας και θα προέβαινε σε σύναψη σύμβασης μαζί του. Αυτή η προσέγγιση περιελάμβανε την παροχή - από πλευράς δανειστή - ενός δείγματος των δεδομένων του και την παραγωγή ενός μοντέλου από πλευράς του δημιουργού της βαθμολόγησης. Υπήρχαν λίγοι δημιουργοί βαθμολόγησης σε κάθε αγορά, αλλά παρείχαν μία χρήσιμη υπηρεσία. Είχαν την τάση να έχουν πολύ μεγάλους ηλεκτρονικούς υπολογιστές που διέθεταν αρκετό χώρο ώστε να πραγματοποιούν τεράστιους υπολογισμούς. (Thomas et al., 2002)

Όπως προαναφέρθηκε, τα πράγματα έχουν αλλάξει. Πιο συγκεκριμένα, υπάρχουν, πλέον, τόσο (εξωτερικές) εταιρείες που αναπτύσσουν μοντέλα βαθμολόγησης όσο και επιμέρους (εσωτερικά) τμήματα πιστωτικών οργανισμών και επιχειρήσεων που έχουν αναλάβει αυτόν τον ρόλο.

Και στις δύο περιπτώσεις, οι πιστωτές δέχονται συμβουλές σχετικώς με το εάν μία ενδεχόμενη δανειοδότηση ενέχει υψηλό κίνδυνο ή όχι και με το πώς πρέπει να αντιμετωπιστεί η αίτηση ενός υποψήφιου δανειολήπτη, καθώς και με το ποιές στρατηγικές πρέπει να ακολουθήσουν προκειμένου να ελαχιστοποιήσουν τους κινδύνους πίστωσης και να μεγιστοποιήσουν τα οφέλη τους. Παρέχονται, επίσης, υπηρεσίες εκπαίδευσης, κυρίως από πλευράς των συμβουλευτικών εταιρειών.

Σε κάθε περίπτωση, δύναται να παρατηρηθεί πως είναι σημαντικό η ανάπτυξη ενός μοντέλου βαθμολόγησης να αποτελεί ένα συλλογικό εγχείρημα, με την έννοια του να εκφράζουν την άποψή τους και οι ίδιοι οι πιστωτές σχετικώς με την μεθοδολογία και τις τεχνικές που θα χρησιμοποιηθούν κατά την παραπάνω ανάπτυξη. Εξ' άλλου, αυτό θα συντελούσε στο να διαδραματίσουν οι ίδιοι έναν πιο ενεργό ρόλο και να αντιληφθούν, ακολούθως, περαιτέρω πτυχές του εγχειρήματος.

Οι (εξωτερικές) συμβουλευτικές εταιρείες υπερέχουν σε σύγκριση με τις (εσωτερικές) συμβουλευτικές ομάδες ως προς το γεγονός ότι, επειδή έχουν μεγάλο και ποικίλο πελατολόγιο, έχουν περισσότερη εμπειρία και μεγαλύτερη ευρύτητα σκέψης που τους βοηθά να αναγνωρίσουν μοτίβα με πιο αποδοτικό τρόπο και να εξάγουν τάσεις σε ολόκληρους τομείς χαρτοφυλακίου. Από την άλλη, οι (εσωτερικές) συμβουλευτικές ομάδες είναι πιο εξειδικευμένες

στα ζητήματα που σχετίζονται με τον συγκεκριμένο πιστωτή, έχουν καλύτερη αντίληψη των σχετικών δεδομένων και μπορούν, μάλλον, να πραγματοποιούν την παραγωγή μοντέλων με μικρότερο κόστος - οικονομικό και χρονικό.

1.2.3 Τα πιστωτικά γραφεία

Όπως θίχθηκε και στα προηγούμενα, πριν από μερικές δεκαετίες το φαινόμενο της πίστωσης δεν είχε τόσο μεγάλες διαστάσεις όσο έχει στην σύγχρονη εποχή, καθώς τα δάνεια δίδονταν με φειδώ. Όταν ένας υποψήφιος δανειολήπτης επιθυμούσε πίστωση, απευθυνόταν στον διευθυντή της αντίστοιχης τράπεζας ή του αντίστοιχου χρηματοοικονομικού οργανισμού και ο τελευταίος, προκειμένου να αποφασίσει εάν θα δοθεί η αιτηθείσα πίστωση ή όχι, ζητούσε μία σύσταση από τον εργοδότη του αιτούντος ή από μία τράπεζα με την οποία συναλλασσόταν εκείνος. Οι προκύπτουσες πληροφορίες, σε κάθε περίπτωση, αφορούσαν γενικά χαρακτηριστικά και ιδιότητες του αιτούντος και η τύχη της αίτησης πίστωσης εξαρτάτο από την υποκειμενική γνώμη του διευθυντή.

Καθώς, όμως, το φαινόμενο της πίστωσης άρχισε να προσλαμβάνει περισσότερες μορφές και ολοένα και μεγαλύτερες διαστάσεις, κατέστη αναγκαία η εύρεση, οργάνωση και αποθήκευση πληθώρας πληροφοριών και δεδομένων κατ' αρχάς. Έπειτα, προέκυψε σωρεία εργασιών σχετικών με τα εγχειρήματα και τις διαδικασίες πίστωσης και η ανάγκη διεκπεραίωσής τους οδήγησε στην ίδρυση των πιστωτικών γραφείων (ή υπηρεσιών πληροφοριών πιστοληπτικής ικανότητας). Αυτά καθιερώθηκαν στις Ηνωμένες Πολιτείες και στο Ηνωμένο Βασίλειο, καθώς και στα περισσότερα ανεπτυγμένα κράτη, αποτελώντας ιδιοκτησία του κράτους σε πολλές περιπτώσεις. Χρήσιμες πληροφορίες για τον ρόλο τους δίδονται, από τους Thomas et al. (2002), παρακάτω.

Στο περιβάλλον τόσο των Ηνωμένων Πολιτειών όσο και του Ηνωμένου Βασιλείου, οι υπηρεσίες πληροφοριών πιστοληπτικής ικανότητας ξεκίνησαν, κατ' αρχάς, να συγκεντρώνουν δημοσίως διαθέσιμες πληροφορίες και να τις τοποθετούν σε ένα κεντρικό σημείο.

Στην σύγχρονη εποχή, χρησιμοποιώντας την ισχύ των ηλεκτρονικών υπολογιστών, τα γραφεία δύνανται να συνδέουν ταχυδρομικές διευθύνσεις ούτως ώστε, όταν ένας καταναλωτής μετακομίζει, να μην αποκόπτεται από το χρέος του.

Τα γραφεία, επίσης, λειτουργούν ως πρακτορεία για τους δανειστές. Οι δανειστές συνεισφέρουν στα στοιχεία των γραφείων σχετικώς με την τρέχουσα κατάσταση των λογαριασμών των δανειοληπτών τους. Άλλοι δανειστές, όταν λαμβάνουν υπ' όψιν τους μία αίτηση πίστωσης, δύναται να δουν και χρησιμοποιήσουν αυτές τις καταστάσεις. Επίσης, άλλοι δανειστές, σε εγχειρήματα μάρκετινγκ, δύναται να δουν και χρησιμοποιήσουν τις καταστάσεις, αν και με

αυξημένους περιορισμούς σχετικώς με την χρήση των δεδομένων. (Στο Ηνωμένο Βασίλειο, αυτός ο διακανονισμός λειτουργεί για τους δανειστές σε βάση αμοιβαιότητας. Χονδρικώς, εάν ο δανειστής συνεισφέρει μόνο λεπτομέρειες σχετικώς με τους πελάτες του που έχουν αθετήσει τις υποχρεώσεις τους, τότε θα δει μόνο λεπτομέρειες σχετικώς με τους πελάτες άλλων δανειστών που έχουν αθετήσει τις υποχρεώσεις τους επίσης.)

Μία άλλη υπηρεσία που προσφέρουν τα γραφεία είναι η συγκέντρωση λεπτομερειών από όλα τα αιτήματα και η προσπάθεια εκτίμησης ασυμβατοτήτων και πιθανώς δολίων αιτήσεων. Σαφώς, όσο πιο υψηλό είναι το επίπεδο των λεπτομερειών με τις οποίες δουλεύουν τόσο το καλύτερο.

Μία περαιτέρω υπηρεσία που χρησιμοποιείται από πολλούς δανειστές είναι ένας γενικός βαθμός. Αυτός ο βαθμός υπολογίζεται από έναν πίνακα βαθμολογίας τον οποίο κατασκεύασε το γραφείο βάσει της εμπειρίας του με εκατομμύρια αιτήσεων και εκατομμύρια εγγραφών πιστωτικού ιστορικού. Είναι ιδιαιτέρως χρήσιμος είτε σε περιπτώσεις που ο δανειστής δεν είναι τόσο μεγάλος ώστε να αναπτύξει πίνακες βαθμολογίας για το δικό του χαρτοφυλάκιο είτε κατά το πρώτο έτος ή κατά τα δύο πρώτα έτη ενός καινούργιου προϊόντος. Χρησιμοποιείται, επίσης, για απόκτηση μίας επίκαιρης άποψης ως προς την πιστωτική θέση του δανειολήπτη, καθώς ενσωματώνει την πρόσφατη πιστωτική συμπεριφορά του. Πράγματι, κάποιοι δανειστές, ιδίως στα χαρτοφυλάκια πιστωτικών καρτών, αγοράζουν έναν βαθμό για κάθε κάτοχο κάρτας τους, κάθε μήνα, και χρησιμοποιούν αυτούς τους βαθμούς προκειμένου να εκτιμήσουν το πώς θα πρέπει να αντιμετωπίσουν περιπτώσεις απώλειας πληρωμών ή υπέρβασης ορίων ή το πότε και το κατά πόσο πρέπει να αυξήσουν το πιστωτικό όριο των πελατών τους.

Σε αυτό το σημείο, θα πρέπει να διευκρινιστεί πως τα πιστωτικά γραφεία παρέχουν πληροφορίες και εργαλεία ανάλυσης προκειμένου να βοηθήσουν τις επιχειρήσεις να λάβουν αποφάσεις σχετικώς με το εάν θα προσφέρουν πίστωση και με τί επιτόκιο, αλλά δεν λαμβάνουν τα ίδια αυτές τις αποφάσεις (The balance, 2019).

Επιπλέον, πρέπει να επισημανθεί πως κάθε γραφείο πίστωσης ενδέχεται να διαθέτει διαφορετικές πληροφορίες σχετικώς με το ίδιο άτομο και, επομένως, τα άτομα θα πρέπει να πραγματοποιούν συχνούς ελέγχους για ενδεχόμενα λάθη (Investing Answers, 2019).

1.3 Εφαρμογές

Η ενότητα 1.3 περιλαμβάνει εφαρμογές της πίστωσης και της αξιολόγησης πιστωτικού κινδύνου.

1.3.1 Η φόρμα αίτησης και τα απαιτούμενα δεδομένα

Όπως έχει αναφερθεί ήδη, ο αιτών πίστωση θα πρέπει να συμπληρώσει μία φόρμα αίτησης - είτε έντυπη είτε ηλεκτρονική. Προφανώς, όσο περισσότερα δεδομένα απαιτούνται στην φόρμα αίτησης τόσο πληρέστερο θα είναι το προφίλ του αιτούντος - πράγμα που θα συντελέσει σε μία καλύτερη πρόβλεψη του πιστωτικού κινδύνου.

Ωστόσο, δύναται να παρατηρηθεί πως η απαίτηση συμπλήρωσης και καταγραφής μεγάλου όγκου πληροφοριών προκαλεί μία δυσαρέσκεια στον αιτούντα. Αυτό θα μπορούσε να οδηγήσει στην συμπλήρωση μη ορθών δεδομένων, απλώς και μόνο προκειμένου να μην καταναλώσει πολύ χρόνο σε αυτήν την διαδικασία, ή ακόμη και στην αποθάρρυνσή του ως προς την ολοκλήρωση της αίτησης. Ένας τρόπος, λοιπόν, να ξεπεραστεί αυτό το πρόβλημα είναι η κατασκευή σύντομων φορμών αίτησης στις οποίες θα ζητείται να δοθούν κάποιες βασικές απαντήσεις. Έπειτα, οι εν λόγω απαντήσεις δύναται να πλαισιωθούν από περαιτέρω δεδομένα που θα παράσχει μία εναλλακτική πηγή - ένα πιστωτικό γραφείο για παράδειγμα.

Επιπλέον, ο πιστωτής - μία τράπεζα συνήθως - ενδέχεται να διατηρεί δικά του αρχεία με δεδομένα, απ' όπου μπορεί να αντλήσει χρήσιμες πληροφορίες.

Όσον αφορά στα άτομα, μία φόρμα αίτησης πίστωσης είθισται να συμπεριλαμβάνει προσωπικά στοιχεία - όπως ονοματεπώνυμο, ημερομηνία γέννησης, τόπο διαμονής - και στοιχεία εργασιακής απασχόλησης - όπως στοιχεία τρέχοντος και προηγούμενων εργοδοτών. Όσον αφορά στις επιχειρήσεις, μία φόρμα αίτησης πίστωσης είθισται να συμπεριλαμβάνει πληροφορίες επικοινωνίας και πληροφορίες σχετικές με την επιχείρηση και το χρηματοοικονομικό προφίλ της.

1.3.2 Ο πίνακας βαθμολογίας

Όπως έχει καταστεί εμφανές, ένας πίνακας βαθμολογίας κατασκευάζεται βάσει δεδομένων από προηγούμενες περιπτώσεις δανειοληπτών. Όπως έχει αναφερθεί και στα προηγούμενα, οι παρελθούσες περιπτώσεις που θα ληφθούν υπ' όψιν θα πρέπει να είναι παρόμοιες με την υπό εξέταση περίπτωση πίστωσης. Σημαντικό είναι, ωστόσο, να πραγματοποιηθεί πολύ προσεκτικά η σχετική σύγκριση, καθώς ελλοχεύει ο κίνδυνος ύπαρξης «τεχνητών» διαφοροποιήσεων, και παραχαραγμένων στοιχείων εν γένει, ως αποτέλεσμα μίας πράξης ανταγωνιστή πιστωτή.

Προκειμένου τα πράγματα να καταστούν πιο σαφή, ας επιχειρηθεί ένα απλό εγχείρημα βαθμολόγησης πιστοληπτικής ικανότητας στο οποίο θα χρησιμοποιηθεί ένα απλός πίνακας βαθμολογίας με τρεις μεταβλητές: ηλικία, ετήσιο εισόδημα μετά φόρους, χρόνος στην παρούσα θέση εργασίας.

Ηλικία (έτη)	Πόντοι	**Ετήσιο Εισόδημα Μετά Φόρους ($)**	Πόντοι	**Χρόνος στην παρούσα θέση εργασίας (έτη)**	Πόντοι
18 - 25	22	0 - 300	18	0 - 1	12
26 - 40	25	301 - 600	25	2 - 6	20
41 - 55	35	601 - 900	38	7 - 14	52
56+	40	901+	49	15+	63

Πίνακας 1.1 Ενδεικτικός πίνακας βαθμολογίας

Βάσει του Πίνακα 1.1, δύναται να γραφούν τα εξής:

- Ένα άτομο ηλικίας 18 ετών, το οποίο έχει ετήσιο εισόδημα μετά φόρους ίσο με $592 και βρίσκεται στην παρούσα θέση εργασίας επί 9 μήνες, θα λάβει βαθμό ίσο με 22+25+12=59 .
- Ένα άτομο ηλικίας 41 ετών, το οποίο έχει ετήσιο εισόδημα μετά φόρους ίσο με $994 και βρίσκεται στην παρούσα θέση εργασίας επί 14 έτη, θα λάβει βαθμό ίσο με 35+49+52=136 .
- Ένα άτομο ηλικίας 60 ετών, το οποίο έχει ετήσιο εισόδημα μετά φόρους ίσο με $305 και βρίσκεται στην παρούσα θέση εργασίας επί 2 έτη, θα λάβει βαθμό ίσο με 40+25+20=85 .

Ας παρατηρηθεί πως ενδέχεται να υπάρχουν συσχετίσεις ανάμεσα στις χρησιμοποιούμενες μεταβλητές: για παράδειγμα, ένα άτομο ηλικίας 18 ετών αποκλείεται να βρίσκεται σε κάποια θέση εργασίας επί 15 και πλέον έτη.

Σε αυτό το σημείο, θα πρέπει να επισημανθεί πως κατά την δημιουργία ενός συστήματος πιστοληπτικής ικανότητας, θα πρέπει να αποφασιστεί ποιός θα είναι ο βαθμός βάσης, δηλαδή εκείνο το όριο που θα διαχωρίζει τους υποψήφιους δανειολήπτες που πιθανολογείται πως θα είναι συνεπείς στις υποχρεώσεις τους από εκείνους που πιθανολογείται πως θα αθετήσουν τις υποχρεώσεις τους και, επομένως, δεν θα πρέπει να λάβουν την αιτούμενη πίστωση.

Ας υποτεθεί ότι στο παραπάνω σύστημα βαθμολόγησης πιστοληπτικής ικανότητας, ο βαθμός βάσης είναι το 90. Οι τρεις υποψήφιοι δανειολήπτες έλαβαν βαθμό ίσο με 59, 136 και 85 αντιστοίχως. Ας παρατηρηθεί, κατ’ αρχάς, ότι δεν ενδιαφέρει πώς ακριβώς προέκυψε ο βαθμός καθενός εξ’ αυτών. Θα μπορούσε, δηλαδή, κάποιος να είναι «αδύναμος» σε μία μεταβλητή αλλά «δυνατός» σε μία άλλη μεταβλητή. Ας παρατηρηθεί, ακόμη, αυτό που αναφέρθηκε στα προηγούμενα: το σύστημα δεν είναι αναγκαίο να εξηγήσει το πώς αποφαίνεται το εάν κάποιος

υποψήφιος δανειολήπτης θα λάβει την αιτούμενη πίστωση ή όχι και, επίσης, δεν είναι αναγκαίο να τεκμηριωθεί το γιατί μία μεταβλητή εισήχθη στο μοντέλο πρόβλεψης - αρκεί να συμβάλλει στην πραγματοποίηση πιο επιτυχημένων προβλέψεων.

Κατόπιν τούτων, προκύπτει ότι ο δεύτερος υποψήφιος δανειολήπτης θα λάβει την αιτούμενη πίστωση ενώ ο πρώτος όχι. Ο τρίτος υποψήφιος δανειολήπτης δεν θα λάβει την αιτούμενη πίστωση μάλλον. Ωστόσο, θα πρέπει να επισημανθεί πως η ακριβής αντιμετώπιση ενός υποψήφιου δανειολήπτη - ιδίως, εάν ο βαθμός του είναι αρκετά «κοντά» στον βαθμό βάσης - εξαρτάται από τον ακριβή τρόπο με τον οποίο θα χρησιμοποιήσει τον προκύπτοντα βαθμό ο δανειστής, όπως περιγράφεται στα επόμενα (Thomas et al., 2002).

Κάποιοι δανειστές επιχειρούν μία πολύ αυστηρή πολιτική σημείου διαχωρισμού. Εάν ο βαθμός είναι μεγαλύτερος από το σημείο διαχωρισμού ή ίσος με αυτό, τότε η αίτηση εγκρίνεται. Εάν είναι μικρότερος από το σημείο διαχωρισμού, τότε η αίτηση απορρίπτεται.

Κάποιοι δανειστές επιχειρούν μία απλή διαφοροποίηση σε αυτό. Δημιουργείται μία λωρίδα παραπομπής ή γκρίζα περιοχή. Αυτή δύναται να είναι 5 ή 10 πόντους στην μία πλευρά ή και στις δύο πλευρές του σημείου διαχωρισμού. Οι αιτήσεις που εμπίπτουν σε μία τέτοια γκρίζα περιοχή παραπέμπονται σε μία πιο προσεκτική εξέταση. Αυτή η εξέταση ενδέχεται να εμπεριέχει κάποια εγγύηση ή αναζήτηση περαιτέρω πληροφοριών.

Κάποιοι δανειστές χρησιμοποιούν κανόνες πολιτικής που θέτουν τις εν δυνάμει αποδεκτές περιπτώσεις σε μία λωρίδα παραπομπής. Για παράδειγμα, αυτό ενδέχεται να συμβαίνει όταν η αίτηση επιτυγχάνει το σημείο διαχωρισμού αλλά υπάρχει κάποιο δυσμενές περιστατικό στις πληροφορίες του πιστωτικού γραφείου - μία χρεοκοπία για παράδειγμα. Με άλλα λόγια, δεν θα ήταν επιτρεπτό η ισχύς των υπόλοιπης αίτησης να εξισορροπηθεί αυτομάτως με μία αδυναμία.

Κάποιοι δανειστές χρησιμοποιούν τιμολόγηση βασισμένη στον κίνδυνο ή διαφοροποιημένη τιμολόγηση. Εδώ ενδέχεται να μην υφίσταται η απλή προκαθορισμένη τιμή πλέον. Η τιμή προσαρμόζεται, μάλλον, σύμφωνα με τον κίνδυνο (ή την προοπτική κέρδους) που αντιπροσωπεύει η πρόταση. Αντί ενός σημείου διαχωρισμού, ο δανειστής ενδέχεται να έχει αρκετά. Ενδέχεται να υπάρχει ένα υψηλό σημείο διαχωρισμού που οριοθετεί τους «καλύτερους» αιτούντες που, ίσως, τους προσφερθεί ένα αναβαθμισμένο προϊόν, ένα άλλο σημείο διαχωρισμού για το τυπικό προϊόν με ένα χαμηλότερο επιτόκιο, ένα τρίτο σημείο διαχωρισμού για το τυπικό προϊόν στην τυπική τιμή και ένα τέταρτο σημείο διαχωρισμού για ένα υποβαθμισμένο προϊόν. Στον εμπορικό δανεισμό, μέχρι ένα σημείο, η εκτίμηση της τιμής λαμβάνει υπ’ όψιν τον κίνδυνο, αν και άλλα ζητήματα, όπως ο ανταγωνισμός και η πελατειακή σχέση, έχουν κάποια σχέση επίσης.

1.4 Στοιχεία μικροοικονομίας

Η ενότητα 1.4 προσεγγίζει το μικροοικονομικό περιβάλλον της πίστωσης.

1.4.1 Οικονομική ανάλυση της ζήτησης πίστωσης

Εξετάζοντας το πώς το εισόδημα ενός δανειολήπτη θα μπορούσε να επηρεάσει την ζήτηση πίστωσης από πλευράς του, προέκυψε μία δημοφιλής θεωρία που δύναται να καλύψει όσες περιόδους προσμένει να ζήσει ένα άτομο.

Βάσει αυτής της θεωρίας, προκύπτει το σκεπτικό ότι ένα άτομο θα δανειστεί, μάλλον, σε νεαρή ηλικία, καθώς είθισται τα έξοδά του να υπερβαίνουν τα έσοδά του τότε, θα αποταμιεύσει στην μέση ηλικία και θα καταναλώσει τις αποταμιεύσεις όταν συνταξιοδοτηθεί. Από αυτό το σκεπτικό, λοιπόν, προέκυψε η «υπόθεση του κύκλου ζωής».

Πρόκειται για μία οικονομική θεωρία σύμφωνα με την οποία τα άτομα προγραμματίζουν την καταναλωτική και αποταμιευτική συμπεριφορά τους σε μακροχρόνιο ορίζοντα, ώστε να επιτύχουν την καλύτερη δυνατή (διαχρονικώς) κατανομή της κατανάλωσής τους για όσα χρόνια ελπίζουν ότι θα ζήσουν. Βασίζεται στην υπόθεση ότι οι καταναλωτές - διαχρονικώς - δρουν ορθολογικώς και, για την τρέχουσα κατανάλωσή τους ,λαμβάνουν υπ’ όψιν, εκτός από το τρέχον εισόδημα, και τα προβλεπόμενα μελλοντικά εισοδήματά τους (Ευρετήριο οικονομικών όρων, 2019). Είναι φανερό ότι το κύριο στοιχείο αυτής της καταναλωτικής συμπεριφοράς είναι η εξασφάλιση ομαλής κατανομής της καταναλωτικής δαπάνης σε ολόκληρη την ζωή ενός ατόμου ή ενός νοικοκυριού.

Σύμφωνα με αυτήν την θεωρία, λοιπόν, αναμένεται ότι τα άτομα νεαρής ηλικίας έχουν αυξημένα χρέη λόγω δανεισμού, καθώς και ότι μία αύξηση στο ποσοστό του πληθυσμού ατόμων νεαρής ηλικίας θα συνοδευόταν από αυξήσεις χρεών ενώ μία μείωση στο ποσοστό αυτό θα συνοδευόταν από μειώσεις χρεών. Κάτι τέτοιο, ωστόσο, δεν δύναται να καταστεί οπωσδήποτε αποδεκτό: ακόμη κι εάν η αύξηση των ατόμων νεαρής ηλικίας διαδραμάτιζε ουσιαστικό ρόλο στην αύξηση των χρεών μακροπροθέσμως, τότε πως θα δικαιολογούντο αντίστοιχες αυξήσεις χρεών βραχυπροθέσμως; Βεβαίως, το αντίστοιχο ερώτημα διατυπώνεται και για μειώσεις χρεών: ακόμη κι εάν η μείωση των ατόμων νεαρής ηλικίας διαδραμάτιζε ουσιαστικό ρόλο στην μείωση των χρεών μακροπροθέσμως, τότε πως θα δικαιολογούντο αντίστοιχες μειώσεις χρεών βραχυπροθέσμως;

Σε αυτό το σημείο, θα πρέπει να επισημανθεί πως παρατηρείται το φαινόμενο κάποια άτομα ή και επιχειρήσεις να έχουν περιουσιακά στοιχεία και χρέη ταυτοχρόνως. Παρά το γεγονός ότι η λήψη πίστωσης συνοδεύεται από συγκεκριμένες υποχρεώσεις, συχνά αυξημένες λόγω τόκων,

αυτό δύναται να εξηγηθεί ως εξής: η λήψη πίστωσης προσφέρει άμεσους χρηματοοικονομικούς πόρους που, ακολούθως, προσφέρουν ευελιξία κινήσεων και πληρωμών για αγορές από πλευράς δανειοληπτών.

1.4.2 Περιορισμοί πίστωσης

Θα μπορούσε να σκεφθεί κανείς πως όσο πιο μεγάλο είναι το επιτόκιο τόσο περισσότερη πίστωση θα προθυμοποιούντο να παρέχουν οι τράπεζες και τόσο λιγότερη πίστωση θα ζητούσαν οι καταναλωτές. Η αγορά της πίστωσης δεν είναι απλή, όμως, όπως αιτιολογείται ακολούθως (Thomas et al., 2002).

Πρώτον, υπάρχουν στοιχεία ότι πολλοί καταναλωτές δεν λαμβάνουν όλη την πίστωση που επιθυμούν και το ποσό της χορηγούμενης πίστωσης που παρατηρείται, στην πραγματικότητα, είναι το ποσό που οι τράπεζες επιθυμούν να προσφέρουν με αυτό το επιτόκιο.

Δεύτερον, κάποιοι αιτούντες και τράπεζες διαφέρουν στην εκτίμησή τους ως προς το ποιό είναι το κατάλληλο επιτόκιο να εισπραχθεί από την τράπεζα προκειμένου να το αντισταθμίσει με την εικόνα του κινδύνου δανεισμού του αιτούντος που έχει.

Τρίτον, ενδέχεται να μην υπάρχει διαφορά ανάμεσα στα επιτόκια που περιγράφονται στην δεύτερη περίπτωση, αλλά ο συμφωνηθείς κίνδυνος είναι τόσο υψηλός που δεν υπάρχει κανένα επιτόκιο που θα μπορούσε να αποζημιώσει τον δανειστή και με το οποίο ο αιτών θα επιθυμούσε να δανειστεί οποιοδήποτε ποσό.

Τέταρτον, κάποιοι ενδέχεται να λάβουν πίστωση όταν κάποιοι άλλοι με ίδια χαρακτηριστικά όχι. Αυτό δύναται να προκύψει ως ακολούθως. Ας υποτεθεί ότι η τράπεζα επιλέγει ένα υψηλό επιτόκιο. Αυτό θα αυξήσει τα έσοδα από κάθε όγκο δανείων. Όμως, καθιστά, επίσης, κάθε δάνειο πιο επικίνδυνο διότι αυξάνει την πιθανότητα αθέτησης των υποχρεώσεων. Περαιτέρω αύξηση του επιτοκίου θα προκαλούσε στροφή των αιτούντων χαμηλού κινδύνου σε άλλους προμηθευτές ή στο να μην δανειστούν καθόλου (και, επομένως, μόνο δανειστές υψηλού κινδύνου θα παρέμεναν και θα απέφεραν χαμηλότερα κέρδη).

2. ΜΕΘΟΔΟΙ

Το Κεφάλαιο 2 περιλαμβάνει δημοφιλείς μεθόδους αξιολόγησης πιστοληπτικής ικανότητας με έμφαση σε εκείνες της λογιστικής παλινδρόμησης και των δένδρων απόφασης.

2.1 Λογιστική παλινδρόμηση

Η ενότητα 2.1 παρουσιάζει το θεωρητικό πλαίσιο της μεθόδου της λογιστικής παλινδρόμησης.

2.1.1 Οι βασικές έννοιες

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 (Collett, 2003).

Η **παλινδρόμηση** (**regression**) είναι μία στατιστική τεχνική μοντελοποίησης με σκοπό την διερεύνηση της συσχέτισης μίας μεταβλητής απόκρισης - εξαρτημένης μεταβλητής και μίας ή περισσότερων επεξηγηματικών μεταβλητών - ανεξάρτητων μεταβλητών. Υποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με κάποιο είδος συνάρτησης και, έπειτα, καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα εν λόγω δεδομένα. Αποτέλεσμα της παλινδρόμησης αποτελεί ένα μοντέλο που χρησιμοποιείται προκειμένου να προβλέψει τις τιμές της εξαρτημένης μεταβλητής για καινούργια δεδομένα.

Η **ανάλυση παλινδρόμησης** (**regression analysis**) αφορά στην μελέτη της μεταβολής της εξαρτημένης μεταβλητής όταν μεταβάλλεται μία από τις ανεξάρτητες μεταβλητές ενώ οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές και, αποσκοπεί στην εξακρίβωση της αιτιώδους επίδρασης μίας μεταβλητής σε μία άλλη (Βικιπαίδεια, 2019).

Πολύ συχνά η εξαρτημένη μεταβλητή είναι διακριτή και λαμβάνει μία ή περισσότερες δυνατές τιμές. Το μοντέλο λογιστικής παλινδρόμησης είναι το πιο συχνά χρησιμοποιούμενο μοντέλο παλινδρόμησης για την ανάλυση τέτοιου είδους δεδομένων (Hosmer et al., 2013).

2.1.2 Η παρουσίαση της μεθόδου

Το μοντέλο της λογιστικής παλινδρόμησης (logistic regression) αποτελεί ένα γενικευμένο γραμμικό μοντέλο[4], το οποίο είναι πολύ σημαντικό - έχει, επίσης, αναπτυχθεί εκτεταμένη

[4] Μία δημοφιλή κατηγορία στατιστικών μοντέλων παλινδρόμησης αποτελούν τα γενικευμένα γραμμικά μοντέλα (generalized linear models). Σε αυτά τα μοντέλα, η κατανομή της εξαρτημένης μεταβλητής, έστω Y, πρέπει να είναι μέλος της εκθετικής οικογένειας κατανομών. Εάν, δηλαδή, θεωρήσουμε ένα τυχαίο

σχετική βιβλιογραφία - και χρησιμοποιείται τόσο σε εφαρμογές της επιστήμης της ιατρικής όσο και σε εφαρμογές του χρηματοπιστωτικού τομέα. Η μέθοδος της λογιστικής παλινδρόμησης χρησιμοποιεί έναν μετασχηματισμό λογαριθμικής φύσεως προκειμένου να καταστήσει δυνατή την μοντελοποίηση μίας μη γραμμικής σχέσης με γραμμικό τρόπο (όπως θα διαπιστώσουμε, παρακάτω, κατά την εξήγηση του μοντέλου της λογιστικής παλινδρόμησης).

Σε πληθώρα περιπτώσεων, η εξαρτημένη μεταβλητή - η οποία είναι ποιοτική ή αλλιώς κατηγορική[5] και της οποίας η πρόβλεψη, βάσει ορισμένων ανεξάρτητων μεταβλητών, επιχειρείται - είναι δίτιμη. Οι δύο δυνατές τιμές της εξαρτημένης μεταβλητής συνδέονται με δύο αντίστοιχα ενδεχόμενα: ο ασθενής επέζησε / ο ασθενής απεβίωσε, ο δανειολήπτης αποπλήρωσε το δάνειό του / ο δανειολήπτης αθέτησε τις υποχρεώσεις του κ.ο.κ. Τα δύο αυτά ενδεχόμενα είναι συμπληρωματικά, αφού καθένα τους πραγματοποιείται όταν δεν πραγματοποιείται το άλλο και, επομένως, το άθροισμα των πιθανοτήτων τους ισούται με την μονάδα. Επειδή, όπως αναφέρθηκε παραπάνω, η εξαρτημένη μεταβλητή είναι κατηγορική, οι τιμές της, τελικώς, συνιστούν μία αυθαίρετη κωδικοποίηση των δύο αντίστοιχων ενδεχομένων, συνήθως 1 και 0. Το ένα ενδεχόμενο θεωρείται ως «επιτυχία» και το άλλο ως «αποτυχία» (Collett, 2003).

Ας θεωρήσουμε, λοιπόν, την δίτιμη διακριτή (λαμβάνει διακριτές τιμές, όχι συνεχείς) τ.μ. y, καθώς και το ενδεχόμενο «επιτυχία» (y=1) με πιθανότητα p και το ενδεχόμενο «αποτυχία» (y=0) με πιθανότητα q=1-p. Τότε, η τ.μ. y ακολουθεί την κατανομή Bernoulli - αυτό συμβολίζεται με y~B(p) - με συνάρτηση μάζας πιθανότητας (σ.μ.π.)

$$f(y) = p^{y}(1-p)^{1-y},\ y=0,1\ ,$$

E(y)=p (μέση τιμή) και V(y)=p(1-p) (διασπορά). Προφανώς, τα p και q, ως πιθανότητες και προκειμένου να μην μηδενίζεται η σ.μ.π., ανήκουν στο (ανοικτό) διάστημα (0,1). Σε αυτήν την περίπτωση, θεωρήσαμε τυχαίο πείραμα με μόνο μία λήψη ή αλλιώς δοκιμή (n=1). Τα δεδομένα αυτής της περιπτώσεως καλούνται «δυαδικά».

Εάν πραγματοποιήσουμε μία επέκταση, λαμβάνοντας n - το πλήθος - δοκιμές (δηλαδή

δείγμα (τ.δ.) από την εν λόγω κατανομή, έστω **Y**=(Y_1,…,Y_n) - οι τυχαίες μεταβλητές (τ.μ.) Y_1,…,Y_n είναι ανεξάρτητες και ισόνομες προκειμένου να αποτελούν τ.δ. - τότε θα πρέπει να διαπιστώσουμε ότι η πυκνότητα πιθανότητας του τ.δ. αυτού, δηλ. το γινόμενο των πυκνοτήτων πιθανότητας των τ.μ. Y_1,…,Y_n, δύναται να γραφεί σε εκθετική μορφή. Χαρακτηριστικά μέλη της εκθετικής οικογένειας κατανομών είναι η κατανομή Bernoulli και η Διωνυμική κατανομή που συναντούμε στο μοντέλο της λογιστικής παλινδρόμησης και, επομένως, θα αναλύσουμε παρακάτω.

[5] Σε αντίθεση με μία ποσοτική μεταβλητή, της οποίας η τιμή αποδίδεται με αριθμούς, μία ποιοτική ή αλλιώς κατηγορική μεταβλητή δείχνει μεταβολή ενός παράγοντα κατά είδος. Παραδείγματα τέτοιων μεταβλητών αποτελούν τα εξής: «φύλο», «επαγγελματική κατάσταση», «αθέτηση» με δυνατές τιμές, αντιστοίχως, «άνδρας», «γυναίκα» και, «ιδιωτικός υπάλληλος», «δημόσιος υπάλληλος», «ελεύθερος επαγγελματίας», «άνεργος» (ενδεικτικές τιμές) και, «ναι», «όχι».

πραγματοποιήσεις των ενδεχομένων) (n>1), τότε ορίζουμε την τ.μ. y=πλήθος επιτυχιών σε n - το πλήθος - δοκιμές. Υποθέτοντας ότι η πιθανότητα πραγματοποίησης του ενδεχομένου «επιτυχία», p, είναι σταθερή σε καθεμία από τις προαναφερθείσες δοκιμές και ότι οι δοκιμές είναι ανεξάρτητες μεταξύ τους, η τ.μ. y ακολουθεί την Διωνυμική κατανομή, - αυτό συμβολίζεται με y~b(n,p) - με συνάρτηση μάζας πιθανότητας (σ.μ.π.)

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \; y=0,1,2,\ldots,n \; ,$$

E(y)=np (μέση τιμή) και V(y)=np(1-p) (διασπορά). Τα δεδομένα αυτής της περιπτώσεως καλούνται «διωνυμικά».

Όπως αναφέρθηκε και παραπάνω, η τ.μ. y εξαρτάται, γενικώς, από κάποιες ανεξάρτητες μεταβλητές ή αλλιώς συμμεταβλητές **x**. Η εν λόγω εξάρτηση προκύπτει από την εξάρτηση της πιθανότητας επιτυχίας, p, από τις συμμεταβλητές. Για παράδειγμα, η πιθανότητα του ενδεχομένου επιβίωσης ενός ασθενούς εξαρτάται από την ηλικία του, το ιατρικό ιστορικό του κ.ά. και, η πιθανότητα του ενδεχομένου αποπληρωμής ενός δανείου εξαρτάται από την ηλικία του δανειολήπτη, τα εισοδήματά του κ.ά.

Το μοντέλο της λογιστικής παλινδρόμησης, λοιπόν, εκφράζεται μέσω της σχέσης $\eta_{\mathbf{x}}=g(E(y_{\mathbf{x}}))=g(\mu_{\mathbf{x}})=\mathbf{x}'\boldsymbol{\beta}$ με την ακόλουθη δομή (Καρώνη & Οικονόμου, 2017):

- $y_{\mathbf{x}}\sim b(n_{\mathbf{x}},\mu_{\mathbf{x}})$ ή αλλιώς $y_{\mathbf{x}}\sim b(n_{\mathbf{x}},n_{\mathbf{x}}\cdot p_{\mathbf{x}})$ ($n_{\mathbf{x}}>1$, διωνυμικά δεδομένα) ή

  $y_{\mathbf{x}}\sim B(\mu_{\mathbf{x}})$ ή αλλιώς $y_{\mathbf{x}}\sim B(p_{\mathbf{x}})$ ($n_{\mathbf{x}}=1$, δυαδικά δεδομένα)
- $\eta_{\mathbf{x}} = g(\mu_{\mathbf{x}}) = \ln\frac{\mu_{\mathbf{x}}}{n_{\mathbf{x}}-\mu_{\mathbf{x}}} = \ln\frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \mathrm{logit}(p_{\mathbf{x}}) = \mathbf{x}'\boldsymbol{\beta}$ (συνάρτηση logit)
- ανεξαρτησία μεταξύ των παρατηρήσεων $y_{\mathbf{x}}$

<u>Εξήγηση μοντέλου</u>

Κατ' αρχάς, οι συμμεταβλητές **x** και οι αντίστοιχες παράμετροι παλινδρόμησης (παράμετροι του μοντέλου) **β** ορίζονται ως εξής: $\mathbf{x}'=(x_0,x_1,\ldots,x_k)$ όπου το x_0 λαμβάνεται ίσο με 1, και $\boldsymbol{\beta}'=(\beta_0,\beta_1,\ldots,\beta_k)$ αντιστοίχως.

Επίσης, στα γενικευμένα γραμμικά μοντέλα γενικώς, θεωρούμε ότι οι συμμεταβλητές συνδέονται γραμμικώς, σχηματίζοντας την **γραμμική προβλέπουσα** (**linear predictor**) $\eta_{\mathbf{x}}=\mathbf{x}'\boldsymbol{\beta}$, η οποία συνδέεται με την μέση τιμή της εξαρτημένης μεταβλητής $y_{\mathbf{x}}$, $\mu_{\mathbf{x}}$, μέσω της **συνάρτησης σύνδεσης** (**link function**) g ως εξής: $\eta_{\mathbf{x}}=g(\mu_{\mathbf{x}})=\mathbf{x}'\boldsymbol{\beta}$.

Θα πρέπει να επισημανθεί ότι η εξάρτηση από τις συμμεταβλητές, για μία στατιστική μονάδα[6], εκφράζεται μέσω ενός κατάλληλου μετασχηματισμού g της μέσης τιμής της εξαρτημένης μεταβλητής, έτσι ώστε να ισχύει μία σχέση της μορφής $g(\mu_x)=\mathbf{x}'\boldsymbol{\beta}$. Στην ουσία, η σχέση μεταξύ των μ_x και $\mathbf{x}'\boldsymbol{\beta}$ είναι μη γραμμική αλλά «γραμμικοποιείται» μέσω της συνάρτησης g[7].

Πιο συγκεκριμένα τώρα λοιπόν, στο μοντέλο της λογιστικής παλινδρόμησης, ως συνάρτηση σύνδεσης χρησιμοποιείται η $\eta_x=g(\mu_x)=\ln\frac{\mu_x}{n_x-\mu_x}$. Επειδή, δε, η εξαρτημένη μεταβλητή ακολουθεί είτε την Διωνυμική κατανομή (με $\mu_x=n_xp_x$, $n_x>1$) είτε την κατανομή Bernoulli (με $\mu_x=n_xp_x$,$n_x=1$ δηλαδή $\mu_x=1\cdot p_x=p_x$), ο τελευταίος όρος δίδει (σε καθεμία από τις δύο περιπτώσεις, κατόπιν απλών πράξεων) την ισοδύναμη ποσότητα $\ln\frac{p_x}{1-p_x}$(δηλαδή $\ln\frac{\mu_x}{n_x-\mu_x}=\ln\frac{p_x}{1-p_x}$).

Επίσης, ολοκληρώνοντας την εξήγηση του μοντέλου, το n_x συμβολίζει, κατά τα προηγούμενα, το πλήθος των δοκιμών (κάθε δοκιμή θα καταλήξει ή σε «επιτυχία» ή σε «αποτυχία» ως αποτέλεσμα), δηλαδή το πλήθος των επαναλήψεων της τιμής του διανύσματος **x** των συμμεταβλητών.

Έπειτα, αντιστρέφοντας την συνάρτηση logit προκύπτει

$$p_{\mathbf{x}} = \frac{e^{\eta_{\mathbf{x}}}}{1+e^{\eta_{\mathbf{x}}}}$$

από την οποία είναι φανερό ότι ισχύει ο απαραίτητος περιορισμός $0<p_x<1$.

Για κάθε παρατήρηση - στατιστική μονάδα, i, λοιπόν, το μοντέλο γράφεται ως

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},\ i=1,\ldots,n$$

όπου

$$p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}$$

η πιθανότητα «επιτυχίας» και συνεπώς

$$E(y_i) = n_i p_i = n_i \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

($n_i = n_{x_i}$το πλήθος δοκιμών της στατιστικής μονάδας i και, $p_i = p_{x_i}$η αντίστοιχη πιθανότητα

[6] Ως «στατιστική μονάδα» («statistical unit») ορίζεται ένα μέλος κάποιου υπό διερεύνηση συνόλου, π.χ. ένας άνθρωπος, ένα ζώο, ένα φυτό, ένας δανειολήπτης.

[7] Όπως θα δούμε αμέσως παρακάτω, η μέθοδος της λογιστικής παλινδρόμησης χρησιμοποιεί συνάρτηση g λογαριθμικής φύσεως για την εν λόγω «γραμμικοποίηση» (πρόκειται, κατά τα προαναφερθέντα, για τον μετασχηματισμό λογαριθμικής φύσεως που χρησιμοποιεί προκειμένου να καταστήσει δυνατή την μοντελοποίηση μίας μη γραμμικής σχέσης με γραμμικό τρόπο).

«επιτυχίας»)

Παρατηρείται, τέλος, ότι η λογιστική παλινδρόμηση, αντί να μοντελοποιεί την απόκριση απ' ευθείας, μοντελοποιεί την πιθανότητα να ανήκει σε μία συγκεκριμένη κατηγορία (James et al., 2013).

2.1.3 Οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων β

Κατ' αρχάς, όσον αφορά στις παραμετρικές μεθόδους, η προσαρμογή ενός μοντέλου στα δεδομένα απαιτεί την εκτίμηση των αγνώστων παραμέτρων του μοντέλου. Η μέθοδος εκτίμησης των αγνώστων παραμέτρων στην λογιστική παλινδρόμηση είναι η **μέθοδος μέγιστης πιθανοφάνειας** (**maximum likehood method**). Πρόκειται για μία πιο γενική μέθοδο με καλύτερες στατιστικές ιδιότητες από άλλες μεθόδους (James et al., 2013).

Κατά την εφαρμογή της εν λόγω μεθόδου, κατασκευάζουμε την πιθανοφάνεια των αγνώστων παραμέτρων στο μοντέλο για το δείγμα - πρόκειται για την πιθανότητα τομής των παρατηρηθέντων δεδομένων και ερμηνεύεται περισσότερο ως συνάρτηση των αγνώστων παραμέτρων παρά ως συνάρτηση των δεδομένων (Collett, 2003). Ακολούθως, περιγράφεται η μέθοδος (Καρώνη & Οικονόμου, 2017).

Η συνάρτηση πιθανοφάνειας, L, ενός δείγματος τιμών $y_1,y_2,\ldots,y_n$ με μέσες τιμές $E(y_i)=\mu_i=n_i p_i$ και συμμεταβλητές $\mathbf{x}_i'=(x_{i0},x_{i1},\ldots,x_{ik})$,όπου n_i το πλήθος δοκιμών της στατιστικής μονάδας i, p_i η αντίστοιχη πιθανότητα επιτυχίας και x_{i0} λαμβάνεται ίσο με 1, γράφεται ως

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

Η πιθανοφάνεια εξαρτάται από τις άγνωστες πιθανότητες επιτυχίας, p_i , οι οποίες με την σειρά τους εξαρτώνται από τα β μέσω της προαναφερθείσης σχέσης,

$$p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}$$

Έτσι, η συνάρτηση πιθανοφάνειας μπορεί να θεωρηθεί ως συνάρτηση των **β** με

$$l=\ln L(\boldsymbol{\beta})$$

$$= \sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{\ln}\binom{\mathbf{n_i}}{\mathbf{y_i}} + \mathbf{y_i \ln p_i} + \mathbf{(n_i-y_i)\ln(1-p_i)}\}$$

$$= \sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{\ln}\binom{\mathbf{n_i}}{\mathbf{y_i}} + \mathbf{y_i \ln}\frac{\mathbf{p_i}}{\mathbf{1-p_i}} + \mathbf{n_i \ln(1-p_i)}\}$$

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

Παραγωγίζοντας έχουμε

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} =$$

$$= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n_i x_{ij} \, e^{x_i' \beta} \left(1 + e^{x_i' \beta}\right)^{-1} \quad j = 0, 1, \dots, k$$

$$= \sum_{i=1}^{n} \left[y_i - n_i e^{x_i' \beta} \left(1 + e^{x_i' \beta}\right)^{-1} \right] x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - n_i p_i) x_{ij}$$

Οι Εκτιμήτριες Μέγιστης Πιθανοφάνειας (Ε.Μ.Π.) των β_j , λοιπόν, προκύπτουν με την ικανοποίηση των εξισώσεων

$$\sum_{i=1}^{n} (y_i - n_i \hat{p}_i) x_{ij} = \sum_{i=1}^{n} (y_i - \hat{\mu}_i) x_{ij} = 0 \ , j = 0, 1, \dots, k \Rightarrow X'(y - \hat{\mu}) = 0$$

Εξισώνοντας, δηλαδή, τις μερικές παραγώγους με 0, παίρνουμε ένα σύστημα από (k+1) - το πλήθος - μη γραμμικές εξισώσεις. Η λύση του συστήματος μάς δίδει τις τιμές των Ε.Μ.Π., $\hat{\beta}$ (εκτιμήσεις). Η αντίστοιχη προσαρμοσμένη τιμή του αριθμού των επιτυχιών για την i-οστή παρατήρηση είναι $\hat{\mu}_i = n_i \hat{p}_i$.

2.1.4 Η ερμηνεία των εκτιμήσεων των παραμέτρων β

Αφού προσαρμοστεί ένα μοντέλο, η έμφαση μετακυλίεται από τον υπολογισμό και την εκτίμηση του πόσο στατιστικώς σημαντικές είναι οι εκτιμηθείσες παράμετροι στην ερμηνεία των τιμών τους (Hosmer et al., 2013). Ένα μεγάλο πλεονέκτημα της λογιστικής παλινδρόμησης, έναντι άλλων μοντέλων για διωνυμικά ή δυαδικά δεδομένα, είναι η δυνατότητα ερμηνείας των $\hat{\beta}$.

Αφού εκτιμηθούν οι παράμετροι, η σχέση μεταξύ της προσαρμοσμένης πιθανότητας απόκρισης, $\hat{p}$, και των τιμών των συμμεταβλητώνx$_0$,x$_1$,x$_2$,…,x$_k$ μπορεί να εκφραστεί ως

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

ή ισοδύναμα μέσω του λόγου των συμπληρωματικών ή σχετικών πιθανοτήτων (odds)

$$\frac{\hat{p}}{1 - \hat{p}} = e^{x' \hat{\beta}} = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \cdots + \hat{\beta}_k \cdot x_k}$$, όπου x$_0$=1.

Από το odds προκύπτει ότι η ποσότητα $e^{\hat{\beta}_j}$ είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος «επιτυχία», όταν η αντίστοιχη ανεξάρτητη μεταβλητή x_j αυξηθεί κατά μία μονάδα, με δεδομένο πάντα ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές (Καρώνη & Οικονόμου, 2017). Εάν η εκτιμημένη παράμετρος $\hat{\boldsymbol{\beta}}_j$ είναι θετική, ο παράγοντας $e^{\hat{\beta}_j}$ είναι μεγαλύτερος από τη μονάδα, γεγονός που σημαίνει πως το odds $\frac{\hat{p}}{1-\hat{p}}$ αυξάνεται με την αύξηση της x_j . Αντιθέτως, εάν η εκτιμημένη παράμετρος $\hat{\boldsymbol{\beta}}_j$ είναι αρνητική, τότε ο παράγοντας $e^{\hat{\beta}_j}$ είναι μικρότερος από τη μονάδα και η σχετική πιθανότητα μειώνεται με την αύξηση της x_j.

Οι παράμετροι της παλινδρόμησης μπορούν να εκφραστούν και μέσα από τον λόγο του λόγου των συμπληρωματικών πιθανοτήτων, δηλαδή μέσα από το λόγο των odds (odds ratio). Γενικώς, ο λόγος των odds ενός ατόμου με τιμές συμμεταβλητών $\mathbf{x}_1$ σε σχέση με ένα άτομο με τιμές $\mathbf{x}_2$ των ίδιων συμμεταβλητών προκύπτει ως

$$\frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = \frac{\mathbf{odds}(\mathbf{y}=\mathbf{1}|\mathbf{x_1})}{\mathbf{odds}(\mathbf{y}=\mathbf{1}|\mathbf{x_2})} = \frac{e^{\mathbf{x_1}'\hat{\boldsymbol{\beta}}}}{e^{\mathbf{x_2}'\hat{\boldsymbol{\beta}}}} = e^{(\mathbf{x_1}-\mathbf{x_2})'\hat{\boldsymbol{\beta}}}$$

2.1.5 Η ελεγχοσυνάρτηση deviance

Η υποενότητα 2.1.5 αναλύει πληροφορίες σχετικώς με την ελεγχοσυνάρτηση deviance.

2.1.5.1 Για διωνυμικά δεδομένα

Η συνάρτηση πιθανοφάνειας συνοψίζει την πληροφορία που μας παρέχουν τα δεδομένα για τις άγνωστες παραμέτρους του μοντέλου. Η τιμή της πιθανοφάνειας, όταν οι άγνωστες παράμετροι θεωρούνται ίσες με τις αντίστοιχες εκτιμήτριες πιθανοφάνειας, μπορεί να δείξει σε ποιόν βαθμό τα δεδομένα προσαρμόζονται στο υπό εξέταση μοντέλο. Αυτή είναι η μεγιστοποιημένη πιθανοφάνεια του υπό εξέταση μοντέλου και την συμβολίζουμε $\hat{\mathbf{L}}_0$. Αυτή δεν δύναται να εκτιμήσει από μόνη της την έλλειψη προσαρμογής του υπό εξέταση μοντέλου, αφού δεν είναι ανεξάρτητη του πλήθους των παρατηρήσεων του δείγματος.

Είναι, επομένως, αναγκαίο να συγκρίνουμε το υπό εξέταση μοντέλο με ένα εναλλακτικό βασικό μοντέλο για τα ίδια δεδομένα. Ως τέτοιο θεωρούμε αυτό για το οποίο οι προσαρμοσμένες τιμές συμπίπτουν με τις πραγματικές, δηλαδή ένα μοντέλο που προσαρμόζει τελείως στις τιμές. Αυτό το μοντέλο έχει ίσο αριθμό παρατηρήσεων και αγνώστων παραμέτρων. Καλείται «κορεσμένο» και η μεγιστοποιημένη πιθανοφάνειά του συμβολίζεται με $\tilde{L}_s$. Το κορεσμένο μοντέλο δεν είναι χρήσιμο από μόνο του, καθώς δεν παρέχει μία απλούστερη, σε σχέση με τις παρατηρήσεις τις ίδιες, περιγραφή των δεδομένων. Ωστόσο, συγκρίνοντας τις $\hat{L}_0$ και $\tilde{L}_s$, δύναται να εκτιμηθεί ο

βαθμός στον οποίο το υπό εξέταση μοντέλο αναπαριστά ικανοποιητικώς τα δεδομένα.

Πιο συγκεκριμένα, ας θεωρήσουμε την συμβολή της i-οστής παρατήρησης στην μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας υπό την υπόθεση, H_s, του κορεσμένου μοντέλου

$$\tilde{\mathbf{l}}_{\mathbf{is}} = \mathbf{ln}\binom{\mathbf{n_i}}{\mathbf{y_i}} + \mathbf{y_i ln}\tilde{\mathbf{p}}_{\mathbf{i}} + (\mathbf{n_i} - \mathbf{y_i})\mathbf{ln}(\mathbf{1} - \tilde{\mathbf{p}}_{\mathbf{i}}) .$$

Η αντίστοιχη συμβολή της i-οστής παρατήρησης στην μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας υπό την υπόθεση, H_0, του υπό εξέταση μοντέλου είναι

$$\hat{\mathbf{l}}_{\mathbf{i0}} = \mathbf{ln}\binom{\mathbf{n_i}}{\mathbf{y_i}} + \mathbf{y_i ln}\hat{\mathbf{p}}_{\mathbf{i}} + (\mathbf{n_i} - \mathbf{y_i})\mathbf{ln}(\mathbf{1} - \hat{\mathbf{p}}_{\mathbf{i}}) .$$

Η διαφορά των δύο τιμών είναι

$$\hat{\mathbf{l}}_{\mathbf{i0}} - \tilde{\mathbf{l}}_{\mathbf{is}} = \mathbf{y_i}(\mathbf{ln}\hat{\mathbf{p}}_{\mathbf{i}} - \mathbf{ln}\tilde{\mathbf{p}}_{\mathbf{i}}) + (\mathbf{n_i} - \mathbf{y_i})[\mathbf{ln}(\mathbf{1} - \hat{\mathbf{p}}_{\mathbf{i}}) - \mathbf{ln}(\mathbf{1} - \tilde{\mathbf{p}}_{\mathbf{i}})]$$

$$= \mathbf{y_i ln}\left(\frac{\hat{\mathbf{p}}_{\mathbf{i}}}{\tilde{\mathbf{p}}_{\mathbf{i}}}\right) + (\mathbf{n_i} - \mathbf{y_i})\mathbf{ln}\left(\frac{\mathbf{1} - \hat{\mathbf{p}}_{\mathbf{i}}}{\mathbf{1} - \tilde{\mathbf{p}}_{\mathbf{i}}}\right)$$

$$= \mathbf{y_i ln}\left(\frac{\hat{\boldsymbol{\mu}}_{\mathbf{i}}}{\mathbf{y_i}}\right) + (\mathbf{n_i} - \mathbf{y_i})\mathbf{ln}\left(\frac{\mathbf{n_i} - \hat{\boldsymbol{\mu}}_{\mathbf{i}}}{\mathbf{n_i} - \mathbf{y_i}}\right)$$

όπου $\tilde{\mathbf{p}}_i = \frac{\mathbf{y_i}}{\mathbf{n_i}}$ και $\hat{\mathbf{p}}_i = \frac{\hat{\boldsymbol{\mu}}_i}{\mathbf{n_i}}$.

Η ελεγχοσυνάρτηση deviance ορίζεται, λοιπόν, ως

$$\mathbf{D}(\hat{\boldsymbol{\beta}}) = \mathbf{D}(\mathbf{y}; \hat{\boldsymbol{\mu}}) = -\mathbf{2}(\hat{\mathbf{l}}_{\mathbf{0}} - \tilde{\mathbf{l}}_{\mathbf{s}}) = \mathbf{2}\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i ln}\left(\frac{\mathbf{y_i}}{\hat{\boldsymbol{\mu}}_{\mathbf{i}}}\right) + (\mathbf{n_i} - \mathbf{y_i})\mathbf{ln}\frac{\mathbf{n_i} - \mathbf{y_i}}{\mathbf{n_i} - \hat{\boldsymbol{\mu}}_{\mathbf{i}}} = \mathbf{2}\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{d_i}(\hat{\boldsymbol{\beta}})$$

όπου $\ln\hat{\mathbf{L}}_0 = \hat{\mathbf{l}}_0$ και $\ln\tilde{\mathbf{L}}_s = \tilde{\mathbf{l}}_s$

Η ποσότητα $d_i(\hat{\boldsymbol{\beta}})$, δε, εκφράζει την συμβολή της παρατήρησης i στην ελεγχοσυνάρτηση deviance.

Υπό την υπόθεση ότι το μοντέλο είναι σωστό, $\mathbf{D}(\hat{\boldsymbol{\beta}}) \sim \mathbf{X}^{\mathbf{2}}_{\mathbf{n-p}}$ ασυμπτωτικώς, όπου n είναι το πλήθος των ομάδων παρατηρήσεων για τις οποίες οι συμμεταβλητές λαμβάνουν τις ίδιες τιμές και p=k+1 είναι το πλήθος των παραμέτρων στο μοντέλο.

Η ελεγχοσυνάρτηση deviance αποτελεί ένα μέτρο σύγκρισης μεταξύ των παρατηρήσεων y_i και των εκτιμηθέντων $\hat{\boldsymbol{\mu}}_i$. Χρησιμοποιείται, κυρίως, για τη σύγκριση και ανάπτυξη μοντέλων: μεγάλες τιμές της D καταδεικνύουν ότι το υπό εξέταση μοντέλο δεν είναι ικανοποιητικό και αντιστρόφως - επομένως, το στατιστικό D «μετράει» τον βαθμό κατά τον οποίο το υπό εξέταση

μοντέλο αποκλίνει από το κορεσμένο.

2.1.5.2 Για δυαδικά δεδομένα

Στην ειδική περίπτωση των δυαδικών δεδομένων, **δηλαδή όταν n_i=1 για κάθε i, η** ελεγχοσυνάρτηση deviance δεν μας παρέχει πληροφορίες για την καταλληλότητα ενός μοντέλου, διότι εξαρτάται μόνο από τις προσαρμοσμένες ή εκτιμημένες τιμές $\widehat{\boldsymbol{\mu}}_i$ ως ακολούθως (Καρώνη & Οικονόμου, 2017).

Η συνάρτηση πιθανοφάνειας για n - το πλήθος - δυαδικές παρατηρήσεις είναι

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

με $E(y_i)=p_i=\mu_i$ και, επομένως, η λογαριθμοποιημένη πιθανοφάνεια είναι

$$\mathbf{l} = \mathbf{lnL} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{y_i ln\mu_i} + (\mathbf{1} - \mathbf{y_i})\mathbf{ln}(\mathbf{1} - \boldsymbol{\mu}_\mathbf{i})\}$$

Εκτιμώντας το κορεσμένο μοντέλο, ισχύει $\tilde{\boldsymbol{\mu}}_i= y_i$και, αφού οι όροι $y_i \ln y_i$ και $(1-y_i)\ln(1-y_i)$ είναι ίσοι με 0 για τις δύο δυνατές τιμές του y_i, 0 και 1, θα ισχύει $\tilde{\mathbf{l}}_s = 0$. Επομένως, η ελεγχοσυνάρτηση deviance για δυαδικές παρατηρήσεις (n_i=1) δίδεται από την σχέση

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = -\mathbf{2}\sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{y_i ln}\widehat{\boldsymbol{\mu}}_\mathbf{i} + (\mathbf{1} - \mathbf{y_i})\mathbf{ln}(\mathbf{1} - \widehat{\boldsymbol{\mu}}_\mathbf{i})\} = -\mathbf{2}\sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{y_i} \ln \frac{\widehat{\boldsymbol{\mu}}_\mathbf{i}}{\mathbf{1} - \widehat{\boldsymbol{\mu}}_\mathbf{i}} + \mathbf{ln}(\mathbf{1} - \widehat{\boldsymbol{\mu}}_\mathbf{i})\}$$

Για το υπό εξέταση μοντέλο, όπου n_i=1, η προαναφερθείσα λογαριθμοποιημένη συνάρτηση πιθανοφάνειας γράφεται ως

$$\mathbf{l} = \mathbf{lnL}(\boldsymbol{\beta}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \{\mathbf{y_i x_i'\boldsymbol{\beta}} - \mathbf{ln}(\mathbf{1} + \mathbf{e}^{\mathbf{x_i'\boldsymbol{\beta}}})\}$$

και παραγωγίζοντας ως προς τις παραμέτρους β_j, έχουμε ότι (βάσει, πάλι, προαναφερθείσης σχέσεως)

$$\frac{\partial \mathbf{lnL}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_\mathbf{j}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \mathbf{p_i})\mathbf{x_{ij}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{i})\, \mathbf{x_{ij}},$$

από την οποία συνεπάγεται ότι

$$\sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{\beta}_\mathbf{j} \frac{\partial \mathbf{lnL}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_\mathbf{j}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{i}) \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{\beta}_\mathbf{j} \mathbf{x_{ij}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{i}) \ln \frac{\boldsymbol{\mu}_\mathbf{i}}{\mathbf{1} - \boldsymbol{\mu}_\mathbf{i}}$$

όπου $\mu_i=p_i= \frac{e^{xi'\beta}}{1+e^{xi'\beta}}$ και p=k+1.

Επειδή η $\hat{\boldsymbol{\beta}}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της $\boldsymbol{\beta}$, η παράγωγος στην αριστερή πλευρά της εξίσωσης μηδενίζεται στο $\hat{\boldsymbol{\beta}}$. Επομένως, οι προσαρμοσμένες πιθανότητες $\hat{\boldsymbol{\mu}}_i = \hat{\mathbf{p}}_i$ πρέπει να ικανοποιούν την εξίσωση

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \hat{\boldsymbol{\mu}}_\mathbf{i})\, \mathbf{logit}(\hat{\boldsymbol{\mu}}_\mathbf{i}) \;=\; \mathbf{0}$$

και άρα

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}\, \mathbf{logit}(\hat{\boldsymbol{\mu}}_\mathbf{i}) \;=\; \sum_{\mathbf{i=1}}^{\mathbf{n}} \hat{\boldsymbol{\mu}}_\mathbf{i}\, \mathbf{logit}(\hat{\boldsymbol{\mu}}_\mathbf{i})$$

Τέλος, αντικαθιστώντας την

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}\, \mathbf{logit}(\hat{\boldsymbol{\mu}}_\mathbf{i})$$

στην σχέση για την $D(\hat{\boldsymbol{\beta}})$, λαμβάνουμε την τελική έκφραση για την ελεγχοσυνάρτηση deviance

$$\mathbf{D}(\hat{\boldsymbol{\beta}}) \;= -\,\mathbf{2} \sum_{\mathbf{i=1}}^{\mathbf{n}} \hat{\boldsymbol{\mu}}_\mathbf{i} \mathbf{logit}(\hat{\boldsymbol{\mu}}_\mathbf{i}) + \mathbf{ln}(\mathbf{1} - \hat{\boldsymbol{\mu}}_\mathbf{i})$$

η οποία εξαρτάται μόνο από τις προσαρμοσμένες τιμές $\hat{\boldsymbol{\mu}}_i$ και όχι άμεσα από τις παρατηρήσεις y_i. Συνεπώς, δεν μπορεί να κριθεί η καλή προσαρμογή του μοντέλου.

Τέλος, σε αυτήν την περίπτωση των δυαδικών παρατηρήσεων, όπου όλα τα n_i= 1, η deviance δεν ακολουθεί την X^2 - κατανομή ούτε προσεγγιστικώς.

2.1.6 Οι X^2 - έλεγχοι καλής προσαρμογής

Για την λογιστική παλινδρόμηση έχουν αναπτυχθεί και οι ακόλουθοι έλεγχοι καλής προσαρμογής.

2.1.6.1 Η ελεγχοσυνάρτηση Pearson

Για την μελέτη της καταλληλότητας του μοντέλου, εκτός από την ελεγχοσυνάρτηση deviance, χρησιμοποιείται και η ελεγχοσυνάρτηση Pearson, που δίδεται από τον τύπο

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

Υπό την υπόθεση ότι το μοντέλο είναι σωστό, $\mathbf{X^2} \sim \mathbf{X^2_{n-p}}$ ασυμπτωτικώς, όπου n είναι το πλήθος των ομάδων παρατηρήσεων για τις οποίες οι συμμεταβλητές λαμβάνουν τις ίδιες τιμές και p=k+1 είναι το πλήθος των παραμέτρων στο μοντέλο.

Η ελεγχοσυνάρτηση deviance και αυτή του Pearson είναι ασυμπτωτικώς ισοδύναμες και ακολουθούν την ίδια X^2 - κατανομή, όταν το προσαρμοσμένο μοντέλο είναι ορθό. Χρησιμοποιούνται για τον ίδιο έλεγχο. Οι τιμές τους διαφέρουν γενικώς, ωστόσο σπάνια στο βαθμό που να οδηγούν σε διαφορετικά συμπεράσματα. Μεγάλες διαφορές μεταξύ τους δύναται να θεωρηθούν ως ένδειξη ότι, για τη μία από τις δύο, η προσέγγιση της X^2 - κατανομής δεν είναι ικανοποιητική. Επίσης, έχει παρατηρηθεί ότι η ελεγχοσυνάρτηση του Pearson είναι, συχνά, καλύτερη από την deviance, διότι δεν επηρεάζεται ιδιαιτέρως από πολύ μικρές συχνότητες. Μάλιστα, ο έλεγχος καλής προσαρμογής με χρήση του στατιστικού Pearson έχει δημιουργήσει παραγωγή πολλής έρευνας κατά τα τελευταία έτη (Hosmer et al., 2013).

Τέλος, εάν κάθε παρατήρηση είναι τέτοια ώστε y=0 ή y=1 με n_i=1 (περίπτωση δυαδικών δεδομένων), τότε η ελεγχοσυνάρτηση deviance δεν είναι χρήσιμη (βλ. 2.1.5.2), όπως χρήσιμη δεν είναι και η ελεγχοσυνάρτηση του Pearson. Σε αυτήν την περίπτωση, προτιμότερη είναι η χρήση του ελέγχου των Hosmer - Lemeshow.

2.1.6.2 Η ελεγχοσυνάρτηση Hosmer - Lemeshow

Σε αντίθεση με την ελεγχοσυνάρτηση deviance και την ελεγχοσυνάρτηση του Pearson, λοιπόν, ο έλεγχος των Hosmer και Lemeshow είναι ένα μέτρο καταλληλότητας του μοντέλου που μπορεί να χρησιμοποιηθεί, αρχικώς, σε μη ομαδοποιημένα δυαδικά δεδομένα (n_i = 1). Στην συνέχεια, όμως, οι παρατηρήσεις ομαδοποιούνται σύμφωνα με τις εκτιμημένες πιθανότητες. Πιο συγκεκριμένα, για να υπολογιστεί ο έλεγχος αυτός, οι παρατηρήσεις διατάσσονται κατά αύξουσα σειρά σύμφωνα με την τιμή της εκτιμημένης πιθανότητας $\hat{p}_i$ και χωρίζονται σε ομάδες του ίδιου περίπου πλήθους παρατηρήσεων.

Ας υποθέσουμε ότι στην i-οστή από τις g, συνολικώς, ομάδες υπάρχουν m_i παρατηρήσεις, όπου το συνολικό πλήθος «επιτυχιών» είναι o_i, και το αντίστοιχο αναμενόμενο πλήθος «επιτυχιών» είναι e_i. (Οι συχνότητες o_i και e_i προκύπτουν από το άθροισμα των y_j και $\hat{\mu}_j$ των $j=1, \dots, m_i$ παρατηρήσεων της ομάδας i αντιστοίχως).

Τότε ο X^2 - έλεγχος καλής προσαρμογής των Hosmer - Lemeshow δίδεται από τον τύπο

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

όπου $\hat{\pi}_i = \frac{e_i}{m_i}$ η μέση πιθανότητα επιτυχίας της i-οστής ομάδας. Από προσομοιώσεις, έχει βρεθεί ότι $X^2_{HL} \sim X^2_{g-2}$ ασυμπτωτικώς - δηλαδή ότι ο έλεγχος αυτός ακολουθεί, προσεγγιστικώς, την X^2 - κατανομή με (g-2) βαθμούς ελευθερίας. Ωστόσο, αφού η τιμή της ελεγχοσυνάρτησης εξαρτάται από τον χωρισμό των παρατηρήσεων σε ομάδες και από το πλήθος τους σε καθεμία από αυτές,

θεωρείται ως ένα ανεπίσημο μέτρο αξιολόγησης προσαρμογής του μοντέλου (Καρώνη & Οικονόμου, 2017).

2.1.7 Τα υπόλοιπα

Τα υπόλοιπα χρησιμοποιούνται ως μέτρα συμφωνίας μεταξύ των παρατηρήσεων της μεταβλητής απόκρισης και τον αντίστοιχων προσαρμοσμένων τιμών και, κυρίως, προκειμένου να ελέγξουμε την καταλληλότητα του υπό εξέταση μοντέλου μέσω γραφικών παραστάσεων. Ο πιο απλός τρόπος αποτελεί ένα γράφημα 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 ικανοποιητικό. Το ίδιο γράφημα μπορεί, επίσης, να δείξει την παρουσία συσχέτισης των υπολοίπων, στην περίπτωση που οι παρατηρήσεις δίδονται σε χρονική σειρά (Καρώνη & Οικονόμου, 2017).

Επίσης, γραφικές παραστάσεις των υπολοίπων έναντι κάθε συμμεταβλητής χωριστά ή έναντι της εκτιμημένης γραμμικής προβλέπουσας, μπορούν να αποβούν πολύ χρήσιμες, ώστε είτε να συμπεριλάβουμε νέες συμμεταβλητές στο μοντέλο είτε να μετασχηματιστεί μία ήδη υπάρχουσα μεταβλητή. Κάποιο συστηματικό σχήμα ή εικόνα δείχνει την πιθανή ύπαρξη προβλήματος στο μοντέλο. Επίσης, όλες αυτές οι γραφικές παραστάσεις χρησιμεύουν στον εντοπισμό έκτροπων ή άτυπων τιμών (outliers) στα δεδομένα (Καρώνη & Οικονόμου, 2017).

2.1.7.1 Τα υπόλοιπα Pearson

Επειδή το μοντέλο της λογιστικής παλινδρόμησης είναι ένα γενικευμένο γραμμικό μοντέλο, τα υπόλοιπα Pearson ορίζονται, γενικώς, ως

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad , i=1,\ldots,n$$

Επειδή, εν προκειμένω, η συνάρτηση διασποράς είναι $V(\hat{\mu}_i) = n_i\hat{p}_i(1-\hat{p}_i)$ και, επίσης, $\hat{\mu}_i = n_i\hat{p}_i$, τα **υπόλοιπα Pearson** δίδονται από την σχέση

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

Τα **τυποποιημένα υπόλοιπα Pearson**, δε, ορίζονται μέσω της σχέσης

$$r_i^{PS} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-\hat{h}_{ii})}} = \frac{ri^P}{\sqrt{(1-\hat{h}_{ii})}}$$

όπου h_{ii} είναι το διαγώνιο στοιχείο του nxn πίνακα

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}^{\frac{1}{2}}$$

Στην τελευταία σχέση, **X** είναι ο nxp πίνακας σχεδιασμού (p=k+1), ο οποίος περιέχει τα x_{ij} , δηλαδή τις τιμές των συμμεταβλητών x_j για την i-οστή παρατήρηση. Πιο συγκεκριμένα, η $1^η$ στήλη του περιέχει n - το πλήθος - άσσους, η $2^η$ περιέχει τα $x_{11}, x_{21},\ldots,x_{n1}$, η $3^η$ περιέχει τα $x_{12}, x_{22},\ldots,x_{n2}$, … η (k+1)-οστή περιέχει τα $x_{1k}, x_{2k},\ldots,x_{nk}$. Επίσης, $\widehat{\mathbf{W}}$ είναι ο nxn διαγώνιος πίνακας, του οποίου το κάθε στοιχείο είναι το $n_i\hat{p}_i(1-\hat{p}_i)$, που αποτελεί την εκτιμημένη διασπορά της απόκρισης y_i .Ακόμη, $E(r_i^{PS})\approx 0$ και $V(r_i^{PS})\approx 1$, ωστόσο η κατανομή αυτών των υπολοίπων δεν προσεγγίζεται καλά από την Κανονική.

Τέλος, ισχύει ότι $\sum_{i=1}^{n} r_i^P = X^2$, με X^2 να είναι το στατιστικό ελέγχου του Pearson.

2.1.7.2 Τα υπόλοιπα deviance

Στα γενικευμένα γραμμικά μοντέλα, τα υπόλοιπα deviance ορίζονται, γενικώς, ως εξής

$$r_i^D = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i(y_i, \hat{\mu}_i)},\ i=1,\ldots,n$$

και αποτελούν την προσημασμένη τετραγωνική ρίζα της συνεισφοράς της i-οστής παρατήρησης σε ολόκληρη την ελεγχοσυνάρτηση deviance, D. Από τον ορισμό αυτό, προκύπτει ότι το άθροισμα των τετραγώνων των r_i^D ισούται με D.

Επίσης, εν προκειμένω (λογιστική παλινδρόμηση, διωνυμικά δεδομένα), τα υπόλοιπα deviance προκύπτουν από την τυποποιημένη συνάρτηση deviance

$$D(y, \hat{\mu}) = D(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} d_i(y_i, \hat{\mu}_i)$$

η οποία , όπως έχουμε δει, ορίζεται ως

$$D(\widehat{\boldsymbol{\beta}}) = D(y; \hat{\mu}) = -2(\hat{l}_0 - \tilde{l}s) = 2\sum_{i=1}^{n} y_i \ln\frac{y_i}{\hat{\mu}_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i} = 2\sum_{i=1}^{n} d_i(\widehat{\boldsymbol{\beta}}))$$

και ταυτίζεται με την συνάρτηση deviance.

Άρα,

$$\sqrt{d_i(y_i, \hat{\mu}_i)} = \sqrt{2y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + 2(n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}}$$

και, επομένως, τα **υπόλοιπα deviance** δίδονται από την σχέση

$$r_i^D = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{2y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + 2(n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}}$$

Το $\mathrm{sgn}(y_i - \hat{\mu}_i)$ μάς βεβαιώνει ότι το r_i^D θα έχει το ίδιο πρόσημο με το r_i^P, αφού $\hat{\mu}_i = n_i\hat{p}_i$.

Τα **τυποποιημένα υπόλοιπα deviance**, δε, κατ' αντιστοιχία με τα τυποποιημένα υπόλοιπα

Pearson, ορίζονται μέσω της σχέσης

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

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 (Hosmer et al., 2013).

2.1.7.3 Τα υπόλοιπα πιθανοφάνειας

Στα γενικευμένα γραμμικά μοντέλα, τα υπόλοιπα πιθανοφάνειας ορίζονται, γενικώς, ως εξής

$$r_i^L = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{\hat{h}_{ii}\left(r_i^{PS}\right)^2 + (1-\hat{h}_{ii})\left(r_i^{DS}\right)^2}\ ,\ i=1,2,\ldots,n$$

Τα r_i^L καλούνται «υπόλοιπα πιθανοφάνειας», επειδή δύναται να ερμηνευθούν ως οι τιμές της ελεγχοσυνάρτησης του λόγου των πιθανοφανειών για την εισαγωγή στο μοντέλο μίας επιπλέον παραμέτρου που αντιστοιχεί σε μία άτυπη παρατήρηση i, ώστε το μοντέλο να έχει τέλεια προσαρμογή σε αυτό το i-οστό σημείο, δηλαδή $\hat{\mu}_i=y_i$.

Ας επισημανθεί, σε αυτό το σημείο, ότι διαγνωστικά μέτρα μπορούν να βασιστούν στις μεταβολές που προκύπτουν, όταν ένα μοντέλο προσαρμόζεται ξανά στα δεδομένα αφαιρώντας από αυτά μία παρατήρηση κάθε φορά. Ένα μέτρο αυτού του τύπου είναι η μεταβολή της deviance (Καρώνη & Οικονόμου, 2017). Αποδεικνύεται ότι το $(r_i^L)^2$ προσεγγίζει την μεταβολή της deviance, αφού παραλειφθεί το i-οστό σημείο από τα δεδομένα. Ακριβής προσδιορισμός αυτής της μεταβολής πραγματοποιείται μόνο με την χρονοβόρα διαδικασία της διαδοχικής αφαίρεσης σημείων και την επαναπροσαρμογή του μοντέλου κάθε φορά. Αντιθέτως, όλα τα στοιχεία για τον υπολογισμό του $(r_i^L)^2$ προκύπτουν από μία μόνο προσαρμογή του μοντέλου στο σύνολο των δεδομένων.

Επομένως, τα **υπόλοιπα πιθανοφάνειας** προσδιορίζονται από τις μεταβολές στην deviance αφαιρώντας την κάθε παρατήρηση με την σειρά της αλλά, προς αποφυγήν της χρονοβόρας διαδικασίας της επανειλημμένης προσαρμογής του μοντέλου, δύναται η μεταβολή αυτή να προσεγγιστεί από τον σταθμισμένο συνδυασμό των παραπάνω υπολοίπων (Καρώνη & Οικονόμου, 2017)

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

Τα υπόλοιπα πιθανοφάνειας, r_i^L, ομοίως με το υπόλοιπα deviance r_i^D, λαμβάνουν το πρόσημο $\mathrm{sgn}(y_i-\hat{\mu}_i)$.

2.1.8 Η επιρροή: η απόσταση του Cook

Η εξέταση της **επιρροής** κάθε παρατήρησης στην προσαρμογή του μοντέλου είναι εξαιρετικά χρήσιμη. Πιο συγκεκριμένα, η εξέταση των σημείων επιρροής στην εκτίμηση ενός μοντέλου δύναται να πραγματοποιηθεί κατασκευάζοντας ορισμένα γραφήματα.

Τα υπόλοιπα πιθανοφάνειας αποτελούν μέτρα που χρησιμοποιούνται με σκοπό την εξέταση της επιρροής κάθε παρατήρησης στην προσαρμογή του μοντέλου. Πιο συγκεκριμένα, κατασκευάζονται το διάγραμμα των υπολοίπων πιθανοφάνειας ως προς τα $\hat{h}_{ii}$, καθώς και τα γραφήματα δείκτη (index plots) των υπολοίπων πιθανοφάνειας, των $\hat{h}_{ii}$ και των **αποστάσεων Cook**.

Ως «απόσταση του Cook» («Cook's distance») ορίζεται ένα μέτρο που «μετρά» το κατά πόσο η αφαίρεση μίας συγκεκριμένης παρατήρησης θα επηρεάσει τις εκτιμήσεις των παραμέτρων ενός μοντέλου. Η στατιστική συνάρτηση του Cook δίδεται από την σχέση

$$CD_i = \frac{1}{p} + \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' I\left(\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right),\ i = 1, \ldots, n\ ,$$

όπου $\hat{\boldsymbol{\beta}}_{(i)}$ και $\hat{\boldsymbol{\beta}}$ είναι οι εκτιμήσεις των παραμέτρων του μοντέλου όταν παραλείπεται από την ανάλυση η i-οστή παρατήρηση καθώς και, αντιστοίχως, όταν χρησιμοποιείται όλο το δείγμα και, επίσης, $I\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}$ είναι η παρατηρούμενη πληροφορία κατά Fisher, με $V\left(\hat{\boldsymbol{\beta}}\right) = I^{-1}\left(\hat{\boldsymbol{\beta}}\right)$,

ή από την πιο απλή σχέση

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

όπου p=k+1 είναι το πλήθος των παραμέτρων στο μοντέλο.

Τέλος, υπάρχει και η αποκαλούμενη ως «τροποποιημένη στατιστική συνάρτηση του Cook», η οποία παρουσιάζει πλεονεκτήματα έναντι της κλασσικής απόστασης Cook στον εντοπισμό σημείων αυξημένης επιρροής (Καρώνη & Οικονόμου, 2017),

$$C_i = \left|r_i^L\right| \sqrt{\frac{(n-p)\hat{h}_{ii}}{p\left(1-\hat{h}_{ii}\right)}}\ \ , i = 1, \ldots, n$$

2.1.9 Τα κριτήρια επιλογής μοντέλου

Σημαντικά κριτήρια για την επιλογή του βέλτιστου μοντέλου στην λογιστική παλινδρόμηση παρουσιάζονται στην συνέχεια.

2.1.9.1 Τα κριτήρια AIC, BIC

Για την σύγκριση δύο ή περισσότερων «ανταγωνιζομένων» μοντέλων, μπορούμε να χρησιμοποιήσουμε τα κριτήρια AIC (Akaike's Information Criterion) και BIC (Bayesian Information Criterion). Αυτά αποτελούν δύο μέτρα καταλληλότητας και εκφράζονται ως αριθμητικές ποσότητες που χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου και την σύγκριση διαφορετικών μοντέλων ως προς την σπουδαιότητά τους.

Το κριτήριο AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο πλήθος παραμέτρων. Το κριτήριο BIC αποτελεί, επίσης, ένα κριτήριο επιλογής του βέλτιστου μοντέλου και η χρήση του είναι όμοια με εκείνη του AIC - η διαφοροποίησή τους έγκειται στο ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC (Καρώνη & Οικονόμου, 2017).

Το κριτήριο AIC λαμβάνει τις μορφές

και

$$\mathrm{AIC} = -2\sum_{i=1}^{n}\{\ln\binom{n_i}{y_i} + y_i \ln\hat{p}_i + (n_i - y_i)\ln(1-\hat{p}_i)\} + 2p$$

$$= 2\sum_{i=1}^{n}\{n_i \ln(1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}) - y_i \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \ln\binom{n_i}{y_i}\} + 2p$$

ενώ το κριτήριο BIC λαμβάνει την μορφή

$$\mathrm{BIC} = 2\sum_{i=1}^{n}\{n_i \ln(1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}) - y_i \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \ln\binom{n_i}{y_i}\} + p \cdot \ln n$$

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.

Παρεμπιπτόντως, παρατηρείται συχνά ότι, στα στατιστικά πακέτα, παραλείπονται οι σταθεροί όροι. Στην λογιστική παλινδρόμηση, ο όρος αυτός είναι ο $\ln\binom{n_i}{y_i}$.

2.1.9.2 Οι συντελεστές συσχέτισης: τα κριτήρια R^2

Ένα μέτρο είναι ο **ψευδοσυντελεστής** R_M^2,

$$R_M^2 = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{\frac{2}{n}}$$

Όπου n είναι το συνολικό πλήθος των παρατηρήσεων που αποτελούν την βάση δεδομένων, $\hat{L}_0$ είναι η μεγιστοποιημένη πιθανοφάνεια για το μοντέλο που περιέχει έναν σταθερό όρο μόνο

(δεν έχει συμμεταβλητές) και, $\hat{L}_1$ είναι η μεγιστοποιημένη πιθανοφάνεια για το μοντέλο που μας ενδιαφέρει. Το πρόβλημα με τον συγκεκριμένο συντελεστή προσδιορισμού είναι ότι ποτέ δεν καταλήγει να πάρει μέγιστη τιμή το 1.

Ένα άλλο μέτρο, το οποίο προτάθηκε προκειμένου να παρακαμφθεί το παραπάνω πρόβλημα είναι ο **διορθωμένος συντελεστής του Nagelkerke**,

$$R_N^2 = \frac{R_M^2}{\max\{R_M^2\}}$$

που προτάθηκε από τον NagelKerke (1991).

Τα δύο αυτά ψευδο - R^2 μέτρα, συχνά παρουσιάζουν χαμηλές τιμές στην λογιστική παλινδρόμηση - ιδιαίτερα στην περίπτωση των δυαδικών δεδομένων - σε σύγκριση με αυτά που συνηθίζονται στα γραμμικά μοντέλα. Αυτό συμβαίνει, διότι η τιμή τους αυξάνεται καθώς προστίθενται περισσότερες παράμετροι στο μοντέλο, καθώς και επειδή το μοντέλο εξηγεί ή προβλέπει μόνο την πιθανότητα «επιτυχίας» p=E(Y) και όχι τις ατομικές τιμές y (0 ή 1). Δεδομένης της p, η «επιτυχία» ή «αποτυχία» είναι τυχαίο γεγονός που το μοντέλο δεν μπορεί να προβλέψει. Επομένως, μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων δεν μπορεί να εξηγηθεί, άρα ένας δείκτης τύπου R^2 λαμβάνει χαμηλή τιμή αναγκαστικώς.

2.1.10 Η καμπύλη ROC

Οι καμπύλες ROC (Receiver Operating Characteristic - Χαρακτηριστικό Λειτουργίας Δέκτη) συμβάλλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις. Χρονολογούνται στις αρχές της δεκαετίας του ’50. Η μέθοδος των καμπυλών ROC, δε, αποτελεί μία μη παραμετρική μέθοδο (Hosmer et al., 2013).

Κατ’ αρχάς, ας εξετάσουμε το πρόβλημα **πρόβλεψης** της αθέτησης ή μη των υποχρεώσεων ενός δανειολήπτη. Στο πρόβλημα αυτό, έστω $\hat{p}$ η εκτιμημένη πιθανότητα «επιτυχίας» (δηλαδή αθέτησης) για κάθε μονάδα. Ορίζουμε ένα κατώφλι, p_0, και διακρίνουμε δύο περιπτώσεις,

- εάν $\hat{p} > p_0$, τότε προβλέπουμε «Αθέτηση»
- εάν $\hat{p} \leq p_0$, τότε προβλέπουμε «Μη Αθέτηση»

Το αποτέλεσμα χαρακτηρίζεται ως «θετικό» (θ) (πραγματοποίηση του ενδεχομένου της εξαρτημένης μεταβλητής «Αθέτηση», δηλαδή τιμή 1) ή «αρνητικό» (α) (μη πραγματοποίηση του ενδεχομένου της εξαρτημένης μεταβλητής «Αθέτηση», δηλαδή τιμή 0).

Υπάρχουν οι ακόλουθες τέσσερις πιθανές εκβάσεις. Εάν το αποτέλεσμα της πρόβλεψης είναι θ και η πραγματική τιμή είναι θ επίσης, τότε αυτό ονομάζεται «Αληθώς Θετικό» (ΑΘ). Ωστόσο, εάν η πραγματική τιμή είναι α, τότε λέγεται «Ψευδώς Θετικό» (ΨΘ). Από την άλλη, εάν το

αποτέλεσμα της πρόβλεψης είναι α και η πραγματική τιμή είναι α επίσης, τότε αυτό ονομάζεται «Αληθώς Αρνητικό» (ΑΑ). Ωστόσο, εάν η πραγματική τιμή είναι θ, τότε λέγεται «Ψευδώς Αρνητικό» (ΨΑ).

Ορίζουμε ένα πείραμα με Θ - το πλήθος - θετικές και Α - το πλήθος - αρνητικές περιπτώσεις. Οι τέσσερις προαναφερθείσες δυνατές εκβάσεις μπορούν να αποτυπωθούν σε έναν πίνακα, τον πίνακα συνάφειας (contingency table):

	Πραγματική τιμή	
Προβλεφθείσα τιμή	**θ**	**α**
θ’	«Αληθώς Θετικό» (ΑΘ)	«Ψευδώς Θετικό» (ΨΘ)
α’	«Ψευδώς Αρνητικό» (ΨΑ)	«Αληθώς Αρνητικό» (ΑΑ)
Σύνολο	Θ	Α

Πίνακας 2.1 Πίνακας συνάφειας

Από έναν πίνακα συνάφειας, όπως είναι ο Πίνακας 2.1, ορίζονται τα ακόλουθα,

- Ποσοστό Αληθώς Θετικών ή Ευαισθησία :

$$\frac{\text{ΑΘ}}{\Theta} = \frac{\text{ΑΘ}}{\text{ΑΘ} + \Psi\text{Α}}$$

- Ποσοστό Αληθώς Αρνητικών ή Ειδικότητα:

$$\frac{\text{ΑΑ}}{\text{Α}} = \frac{\text{ΑΑ}}{\text{ΑΑ} + \Psi\Theta}$$

- Ποσοστό Ψευδώς Θετικών ή 1 - Ειδικότητα:

$$\frac{\Psi\Theta}{\text{Α}} = \frac{\Psi\Theta}{\Psi\Theta + \text{ΑΑ}}$$

- Ακρίβεια:

$$\frac{\text{ΑΘ} + \text{ΑΨ}}{\Theta + \text{Α}}$$

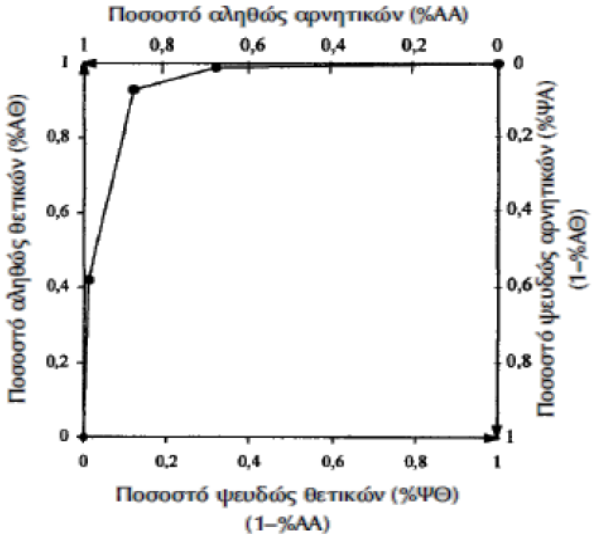
Θα πρέπει, σε αυτό το σημείο, να επισημανθούν τα εξής: οι δύο έννοιες που είναι βασικές στην θεωρία των καμπυλών ROC είναι οι έννοιες «Ευαισθησία» («Sensitivity») και «Ειδικότητα» («Specificity»). Οι έννοιες αυτές αποτελούν τις πιθανότητες η μέθοδος να κατατάσσει σωστά τις «θετικές» και τις «αρνητικές» περιπτώσεις αντιστοίχως (δηλαδή αυτές που συνδέονται με την πραγματοποίηση ή μη του ενδεχομένου της εξαρτημένης μεταβλητής «Αθέτηση», όπως

αναφέρθηκε προηγουμένως). Στην μελέτη μας, επίσης, μας ενδιαφέρει η έννοια «1 - Ειδικότητα» («1 - Specificity»).

Η μέθοδος των καμπυλών ROC κατασκευάζει ένα διάγραμμα ως εξής: στον οριζόντιο άξονα τοποθετούνται οι τιμές της 1 - Ειδικότητας και στον κατακόρυφο άξονα τοποθετούνται οι τιμές της Ευαισθησίας. Οπότε, κάθε σημείο της εμπειρικής καμπύλης ROC προσδιορίζεται από ένα ορισμένο ζεύγος (%ΨΘ, %ΑΘ) και, επομένως, η καμπύλη ROC εκφράζει την σχέση μεταξύ του ποσοστού των Αληθώς Θετικών (%ΑΘ) και του ποσοστού των Ψευδώς Θετικών (%ΨΘ=1-%ΑΑ). Ακολούθως, ένας χώρος ROC ορίζεται από το %ΨΘ στον οριζόντιο άξονα και από το %ΑΘ στον κατακόρυφο άξονα.

Επιπλέον, η καμπύλη ROC ορίζεται ως το μοναδιαίο τετράγωνο [0,1]x[0,1], το οποίο «ξεκινά» από το σημείο (0,0) για να «καταλήξει» στο σημείο (1,1). Η καλύτερη δυνατή μέθοδος πρόβλεψης θα απέφερε ένα σημείο στην επάνω αριστερή γωνία ή αλλιώς την συντεταγμένη (0,1) του χώρου ROC, που αντιπροσωπεύει το 100% Ευαισθησία (μηδέν ΨΑ) και 0% 1 - Ειδικότητα ή αλλιώς 100% Ειδικότητα (μηδέν ΨΘ).

Με άλλα λόγια, στο προκύπτον διάγραμμα, όσο πιο κοντά είναι η καμπύλη στο πάνω - αριστερά όριο του πλέγματος τόσο καλύτερη θεωρείται η απόδοση του μοντέλου. Μάλιστα, η οπτική ερμηνεία αυτή δύναται να περιγραφεί και αριθμητικώς: το εμβαδόν - περιοχή που βρίσκεται κάτω από την καμπύλη ROC, AUC (Area Under the Curve), χρησιμοποιείται ως μέτρο ολικής απόδοσης του μοντέλου και, λαμβάνει τιμές από 0.5 έως 1 - με το 1 να αφορά την περίπτωση του ιδανικού μοντέλου.



Εικόνα 2.1 Ενδεικτική καμπύλη ROC

2.2 Δένδρα απόφασης

Η ενότητα 2.2 παρουσιάζει το θεωρητικό πλαίσιο της μεθόδου των δένδρων απόφασης με έμφαση στα δένδρα ταξινόμησης και τον αλγόριθμο CART.

2.2.1 Οι βασικές έννοιες

Τα δένδρα απόφασης (decision trees) αποτελούν μία δημοφιλή τεχνική με σκοπό την ταξινόμηση και πρόβλεψη. Έχουν μεγάλη εφαρμογή στην διάγνωση ιατρικών περιπτώσεων, καθώς και στην εκτίμηση πιθανού ρίσκου από πιστοληπτικές τραπεζικές εργασίες.

Η δομή τους αποτελείται από «κόμβους» («nodes») και ακμές («edges»). Ο (μοναδικός) κόμβος του ανώτερου επιπέδου ονομάζεται «ρίζα» («root»). Οι κόμβοι του κατώτερου επιπέδου ονομάζονται «φύλλα» («leaves») (James et al., 2013).

Ένα δένδρο απόφασης, ξεκινώντας από την ρίζα και καταλήγοντας στα φύλλα, αντιπροσωπεύει μία ακολουθία από κανόνες της μορφής «εάν … αλλιώς …». Οι εσωτερικοί κόμβοι του περιλαμβάνουν τα γνωρίσματα του προβλήματος, οι ακμές του περιλαμβάνουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα του περιλαμβάνουν τις κλάσεις του προβλήματος.

Από άποψη μοντελοποίησης με δένδρο απόφασης, όπως και με λογιστική παλινδρόμηση άλλωστε, προαπαιτείται ένα σύνολο από στιγμιότυπα εκπαίδευσης (training set), τα οποία περιγράφονται από κάποια γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκουν.

Η διαδικασία που ακολουθούν οι αλγόριθμοι κατασκευής ενός δένδρου απόφασης συνοψίζεται στα ακόλουθα (Breiman et al., 1984): ξεκινώντας από την ρίζα του δένδρου, ο αλγόριθμος διασπά το training set σε υποσύνολα με βάση τη βέλτιστη ιδιότητα (best attribute) του κόμβου - η βέλτιστη ιδιότητα ενός κόμβου καθορίζεται από κάποιο κριτήριο. Επομένως, μπορούμε να πούμε ότι ως ρίζα επιλέγουμε εκείνο το χαρακτηριστικό που δίνει το μέγιστο κέρδος πληροφορίας και, για να το ποσοτικοποιήσουμε, χρησιμοποιούμε την έννοια της εντροπίας. Έτσι, προκύπτει ένα πλήθος υποσυνόλων που το καθένα περιέχει λιγότερα παραδείγματα από το αρχικό σύνολο. Για καθένα από αυτά τα επιμέρους υποσύνολα, εφαρμόζεται επαναληπτικώς η παραπάνω διαδικασία, χρησιμοποιώντας τα εναπομείναντα γνωρίσματα, οπότε η διάσπαση των στιγμιοτύπων προχωρά και σταματά όταν όλα τα στιγμιότυπα του υποσυνόλου ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα.

Όπως και στην περίπτωση της μοντελοποίησης με λογιστική παλινδρόμηση, εκτός από το training set, υπάρχει και το σύνολο ελέγχου (test set) ώστε να ελέγχεται η απόδοση του

μοντέλου. Το πλήθος των λανθασμένων απαντήσεων (δηλαδή τα στιγμιότυπα στα οποία το δένδρο απάντησε διαφορετική κλάση από την πραγματική) καθορίζει την ακρίβεια του δένδρου.

Οι αλγόριθμοι εκπαίδευσης δένδρων απόφασης πραγματοποιούν εξαντλητική αναζήτηση στον χώρο των πιθανών δένδρων απόφασης. Αρχίζουν με ένα κενό δένδρο και, προοδευτικώς, θέτουν πιο περίπλοκες προθέσεις με στόχο την εύρεση ενός δένδρου που ταξινομεί σωστά τα δεδομένα εκπαίδευσης.

Έτσι, η διαδικασία κατασκευής δένδρου απόφασης είναι η εξής (Breiman et al., 1984):

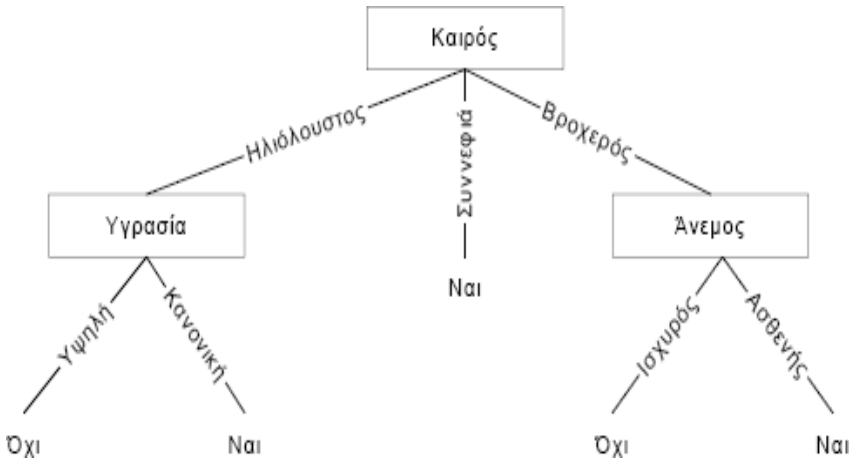
- επιλογή χαρακτηριστικού για την θέση της ρίζας και δημιουργία κλάδων για κάθε πιθανή τιμή του εν λόγω χαρακτηριστικού
- διάσπαση υποδειγμάτων σε υποσύνολα, ένα για κάθε κλάδο που εκτείνεται από την ρίζα
- επανάληψη των παραπάνω για κάθε κλάδο, με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου
- ολοκλήρωση της διαδικασίας όταν όλα τα υποδείγματα σε έναν κόμβο ανήκουν στην ίδια τάξη

Όταν ολοκληρωθεί η διαδικασία, τότε το δένδρο δύναται να αναπαρασταθεί ως ένα σύνολο κανόνων της μορφής

«Εάν <σύνολο συνθηκών> τότε <συμπέρασμα>».

Έστω, για παράδειγμα, το πρόβλημα που επιχειρεί να απαντήσει στο ερώτημα «Παίζεις τέννις;» και το οποίο έχει δυο κλάσεις: Ναι και Όχι. Η απάντηση στο πρόβλημα εξαρτάται από τους εξής παράγοντες: Καιρός (με δυνατές τιμές: Ηλιόλουστος, Βροχερός, Συννεφιά), Υγρασία (με δυνατές τιμές: Υψηλή, Κανονική) και Άνεμος (με δυνατές τιμές: Ισχυρός, Ασθενής).

Το δένδρο απόφασης του παραπάνω προβλήματος, το οποίο παρουσιάζεται στην Εικόνα 2.2, περιέχει τρεις εσωτερικούς κόμβους. Σε κάθε κόμβο πραγματοποιείται έλεγχος ως προς κάποιο από τα γνωρίσματα του προβλήματος, ενώ στα φύλλα περιέχονται οι κλάσεις του προβλήματος.



Εικόνα 2.2 Ενδεικτικό δένδρο απόφασης

Σύμφωνα με το δένδρο, λοιπόν, προκύπτουν τα εξής ως προς το ερώτημα «Παίζεις τέννις;»:

- **Όχι**, εάν ο καιρός είναι ηλιόλουστος και η υγρασία είναι υψηλή ή εάν ο καιρός είναι βροχερός και ο άνεμος ισχυρός .
- **Ναι**, εάν ο καιρός είναι ηλιόλουστος και η υγρασία είναι κανονική ή εάν ο καιρός είναι βροχερός και ο άνεμος είναι ασθενής ή εάν επικρατεί συννεφιά

Τα πλεονεκτήματα των δένδρων απόφασης είναι ότι αναπαριστώνται γραφικώς και, επίσης, δύνανται να διαχειρίζονται κατηγορικές μεταβλητές δίχως την ανάγκη να δημιουργούν ανούσιες μεταβλητές. Ωστόσο, 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 εκτιμούμενο δένδρο. (James et al., 2013)

2.2.2 Ο αλγόριθμος CART

Θα πρέπει να επισημανθεί, κατ' αρχάς, ότι τα δένδρα απόφασης για περιπτώσεις που η εξαρτημένη μεταβλητή είναι ποσοτική ονομάζονται δένδρα παλινδρόμησης (regression trees), ενώ για περιπτώσεις που είναι κατηγορική ονομάζονται δένδρα ταξινόμησης (classification trees). Κατά την διαδικασία της ταξινόμησης, χρησιμοποιούμε δυαδικό διαχωρισμό ώστε να αναπτύξουμε δένδρο ταξινόμησης (James et al., 2013).

Σε έναν αλγόριθμο CART (Classification And Regression Trees), τα δεδομένα χωρίζονται σε δύο υποσύνολα, ξεκινώντας με την ρίζα, η οποία περιέχει ολόκληρο το δείγμα εκπαίδευσης, με τρόπο ώστε κάθε υποσύνολο να εξασφαλίζει περισσότερη ομοιογένεια απ' ό, τι το προηγούμενο. Η διαδικασία αυτή επαναλαμβάνεται έως ότου επιτευχθεί το κριτήριο ομοιογένειας ή κάποιο άλλο κριτήριο διακοπής. Ο αλγόριθμος CART είναι ευέλικτος και δύναται να διαχειρίζεται δεδομένα με ελλείπουσες τιμές χρησιμοποιώντας υποκατάστατο διαχωρισμό.

Ένα θεμελιώδες πλεονέκτημα των αναδρομικών δυαδικών δένδρων, δε, είναι ότι ερμηνεύονται εύκολα (Hastie et al., 2009).

Οι αρχές του αλγορίθμου CART περιγράφονται ακολούθως (Breiman et al., 1984).

2.2.2.1 Τα πεδία συχνότητας και τα πεδία βάρους

Για την κατασκευή του μοντέλου, είναι απαραίτητο να πραγματοποιηθούν κάποιοι υπολογισμοί. Για παράδειγμα, για την μείωση του μεγέθους του συνόλου δεδομένων, χρειάζεται υπολογισμός των πεδίων συχνότητας και βάρους.

Είναι πολύ σημαντικό να γίνει ο σωστός διαχωρισμός μεταξύ των πεδίων βάρους και συχνότητας, διότι θα προκύψουν λανθασμένα αποτελέσματα σε διαφορετική περίπτωση. Στην περίπτωση που τα πεδία συχνότητας ή βάρους δεν ορίζονται, τότε η συχνότητα και τα βάρη για όλες τις καταχωρήσεις παίρνουν τις τιμές 1, 0.

Πεδία συχνότητας

Ένα πεδίο συχνότητας αναπαριστά τον συνολικό αριθμό των παρατηρήσεων που αντιπροσωπεύονται από κάθε καταχώρηση. Στην ανάλυση των συνολικών δεδομένων, είναι σημαντικό να γνωρίζουμε σε ποιό πεδίο μία καταχώρηση (συνδυασμός περιπτώσεων) αναπαριστά περισσότερες από μία παρατηρήσεις. Ο συνολικός αριθμός των παρατηρήσεων μέσα στο δείγμα πρέπει, πάντοτε, να είναι ίσος με το άθροισμα των τιμών στο πεδίο συχνότητας. Το αποτέλεσμα που προκύπτει εάν χρησιμοποιήσουμε πεδίο συχνότητας είναι ίδιο με αυτό που λαμβάνουμε χρησιμοποιώντας δεδομένα κατά περίπτωση.

Στον Πίνακα 2.2, παρατηρούμε ένα υποθετικό παράδειγμα, με τα πεδία πρόβλεψης Φύλο και Απασχόληση και το πεδίο - στόχο Απόκριση (Κάπνισμα). Το πεδίο συχνότητας λέει, για παράδειγμα, ότι 11 εργαζόμενοι άνδρες ανταποκρίθηκαν με «Ναι» στην ερώτηση για το εάν καπνίζουν και 18 άνεργες γυναίκες ανταποκρίθηκαν με «Όχι» στην ίδια ερώτηση.

Φύλο	Απασχόληση	Απόκριση	Συχνότητα
Άνδρας	Ναι	Ναι	11
Άνδρας	Ναι	Όχι	17
Άνδρας	Όχι	Ναι	12
Άνδρας	Όχι	Όχι	21
Γυναίκα	Ναι	Ναι	11
Γυναίκα	Ναι	Όχι	15
Γυναίκα	Όχι	Ναι	15
Γυναίκα	Όχι	Όχι	18

Πίνακας 2.2 Πίνακας - πεδίο συχνότητας

Στο συγκεκριμένο παράδειγμα, χρησιμοποιώντας το πεδίο συχνότητας επεξεργαζόμαστε έναν πίνακα 8 καταχωρήσεων ενώ, εάν χρησιμοποιούσαμε δεδομένα κατά περίπτωση θα ήταν απαραίτητες 120 καταχωρήσεις.

Πεδία βάρους

Κάνοντας χρήση ενός πεδίου βάρους, οδηγούμαστε σε μία άνιση μεταχείριση στις καταχωρήσεις, σε ολόκληρο το σύνολο δεδομένων. Έτσι, η συνεισφορά μίας καταχώρησης στην ανάλυση είναι σταθμισμένη (weighted) σε αναλογία με τον πληθυσμό των μονάδων που η καταχώρηση αναπαριστά μέσα στο δείγμα. Για παράδειγμα, στην ερώτηση μίας έρευνας, σε δείγμα 100000 καταναλωτών, για λογαριασμό μιας επώνυμης βιομηχανίας παραγωγής καπνού, για το εάν καπνίζουν ή όχι, 20000 ερωτηθέντες απάντησαν θετικώς και 80000 αρνητικώς. Σε μία προσπάθεια να μειώσουμε το μέγεθος των δεδομένων, πιθανότατα θα συμπεριλάβουμε όλους όσους είναι καπνιστές και μόνο 25% του δείγματος (20000) που δεν είναι καπνιστές. Κάτι τέτοιο μπορούμε να το πραγματοποιήσουμε εάν ορίσουμε μια περίπτωση βάρους ίση με 1 για αυτούς που καπνίζουν και 4 για αυτούς που δεν καπνίζουν.

2.2.2.2 Η κατασκευή ενός δένδρου CART

Η βασική ιδέα κατασκευής ενός δένδρου είναι να επιλέξουμε έναν διαχωρισμό (split) μεταξύ όλων των πιθανών διαχωρισμών σε κάθε κόμβο, έτσι ώστε οι θυγατρικοί κόμβοι που θα προκύψουν ως αποτέλεσμα να είναι οι καθαρότεροι. Με τον όρο «καθαρότητα»,

αναφερόμαστε στην ομοιότητα των τιμών του πεδίου - στόχου. Σε έναν εντελώς καθαρό κόμβο, όλες οι καταχωρήσεις έχουν την ίδια τιμή στο πεδίο - στόχο. Ο αλγόριθμος CART μετρά την καθαρότητα ενός διαχωρισμού σε έναν κόμβο ορίζοντας ένα μέτρο καθαρότητας.

Τα βήματα που χρησιμοποιούνται για την κατασκευή ενός δένδρου CART είναι τα ακόλουθα (Breiman et al., 1984), ξεκινώντας με τον αρχικό κόμβο - ρίζα, ο οποίος περιέχει όλες τις καταχωρήσεις.

1) Για κάθε πεδίο πρόβλεψης (predictor field), εντοπίζουμε τον καλύτερο δυνατό διαχωρισμό για αυτό ως ακολούθως.

    - **Αριθμητικά πεδία**. Ταξινομούμε τις τιμές των πεδίων στον κόμβο από την μικρότερη στην μεγαλύτερη. Επιλέγουμε κάθε σημείο, με την σειρά, ως σημείο διαχωρισμού και υπολογίζουμε το στατιστικό μη καθαρότητας για τους θυγατρικούς κόμβους που προκύπτουν ως αποτέλεσμα του διαχωρισμού. Έπειτα, διαλέγουμε ως σημείο διαχωρισμού, για το πεδίο, αυτό το οποίο αποδίδει την μεγαλύτερη μείωση στην μη καθαρότητα σε σύγκριση με την μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.

    - **Κατηγορικά πεδία (συμβολικά)**. Εξετάζουμε τον κάθε πιθανό συνδυασμό των τιμών ως 2 υποσύνολα. Για κάθε συνδυασμό, υπολογίζουμε την μη καθαρότητα των θυγατρικών κόμβων για τον διαχωρισμό που βασίζεται σε αυτόν τον συνδυασμό. Επιλέγουμε ως καλύτερο σημείο διαχωρισμού, για το πεδίο, αυτό το οποίο αποδίδει την μεγαλύτερη μείωση στην μη καθαρότητα σε σύγκριση με την μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.

2) Εντοπίζουμε τον καλύτερο διαχωρισμό για τον κόμβο και προσδιορίζουμε το πεδίο του οποίου ο καλύτερος διαχωρισμός δίδει την μεγαλύτερη μείωση στην μη καθαρότητα για τον κόμβο. Στην συνέχεια, επιλέγουμε αυτόν τον καλύτερο διαχωρισμό του πεδίου ως το βέλτιστο συνολικό διαχωρισμό για τον κόμβο.

3) Ελέγχουμε εάν ικανοποιούνται οι κανόνες διακοπής και επαναλαμβάνουμε. Εάν οι κανόνες διακοπής δεν ικανοποιούνται από τον διαχωρισμό ή από τον γεννήτορα κόμβο, τότε εφαρμόζουμε τον διαχωρισμό για να δημιουργήσουμε δύο θυγατρικούς κόμβους. Επαναλαμβάνουμε όλη τη διαδικασία σε κάθε θυγατρικό κόμβο.

2.2.2.3 Τα μέτρα μη καθαρότητας

Για την εύρεση διαχωρισμών στα μοντέλα CART, υπάρχουν δύο γνωστά διαφορετικά μέτρα μη καθαρότητας: οι δείκτες Gini και Towing (χρησιμοποιούνται για συμβολικά πεδία - στόχους).

Gini

Ο δείκτης μη καθαρότητας Gini σε έναν κόμβο t ενός δένδρου CART, g(t), ορίζεται ως:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t)$$

όπου i και j είναι κατηγορίες στο πεδίο - στόχο και

$$p(j|t) = \frac{p(j,t)}{p(t)}$$

$$p(j,t) = \frac{\pi(j)N_j(t)}{N_j}$$

$$p(t) = \sum_j p(j,t)$$

όπου π(j) είναι η τιμή της prior πιθανότητας για την κατηγορία j, $N_j(t)$ είναι το πλήθος των καταχωρήσεων στην κατηγορία j του κόμβου t, N_j είναι το πλήθος των καταχωρήσεων στην κατηγορία j στον αρχικό κόμβο - ρίζα. Ο δείκτης Gini αποτελεί ένα μέτρο της συνολικής διασποράς σε όλες τις κατηγορίες (James et al., 2013) και πρέπει να χρησιμοποιείται κατά την ανάπτυξη ενός δένδρου (Hastie et al., 2009).

Επίσης, όταν χρησιμοποιείται ο δείκτης Gini για την εύρεση της βελτίωσης για έναν διαχωρισμό κατά την διάρκεια της ανάπτυξης του δέντρου , για να υπολογιστούν το N_j και το $N_j(t)$ χρησιμοποιούνται οι καταχωρήσεις στον αρχικό κόμβο - ρίζα και στον κόμβο t αντιστοίχως που έχουν έγκυρες τιμές για το πεδίο διαχωρισμού (split predictor).

Μία άλλη μορφή του δείκτη μη καθαρότητας Gini είναι η

$$g(t) = 1 - \sum_j p^2(j|t)$$

Επομένως, όταν οι καταχωρήσεις σε έναν κόμβο διανέμονται ομαλά διά μέσου των κατηγοριών, ο δείκτης Gini λαμβάνει την μεγαλύτερη τιμή του, $1 - \left(\frac{1}{k}\right)$, όπου k είναι το πλήθος των κατηγοριών για το πεδίο - στόχο. Ο δείκτης Gini ισούται με 0 όταν όλες οι καταχωρήσεις σε έναν κόμβο ανήκουν στην ίδια κατηγορία.

Για τον διαχωρισμό, s, στον κόμβο t, η συνάρτηση του κριτηρίου Gini, Φ(s,t), ορίζεται ως

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

όπου p_L είναι η μερίδα των καταχωρήσεων στον κόμβο t οι οποίες αποστέλλονται στον αριστερό θυγατρικό κόμβο και p_R είναι η μερίδα των καταχωρήσεων στον κόμβο οι οποίες αποστέλλονται

στον δεξιό θυγατρικό κόμβο.

Οι λόγοι p_L και p_R ορίζονται ως εξής:

$$p_L = \frac{p(t_L)}{p(t)}$$

και

$$p_R = \frac{p(t_R)}{p(t)}$$

Επιλέγεται ο κατάλληλος διαχωρισμός, s, ούτως ώστε να μεγιστοποιηθεί η τιμή της συνάρτησης Φ(s,t).

Twoing

Ο δείκτης Twoing είναι βασισμένος στον διαχωρισμό των κατηγοριών στόχου σε δύο υπερκλάσεις και, ακολούθως, στην εύρεση του βέλτιστου διαχωρισμού στο πεδίο πρόβλεψης και, στηρίζεται στις δύο υπερκλάσεις. Οι υπερκλάσεις C_1 και C_2 ορίζονται ως εξής:

$$C_1 = \{j: p(j|t_L) \geq p(j|t_R)\}$$

και

$$C_2 = C - C_1$$

όπου C είναι το σύνολο των κατηγοριών του πεδίου - στόχου και $p(j|t_R)$, $p(j|t_L)$ είναι τα $p(j|t)$ όπως ορίζονται στο κριτήριο Gini για τους δεξιούς και αριστερούς θυγατρικούς κόμβους αντιστοίχως.

Η συνάρτηση του κριτηρίου του Twoing για τον διαχωρισμό s στον κόμβο t ορίζεται ως

$$\Phi(s,t) = p_L p_R [\sum_j |p(j|t_L) - p(j|t_R)|]^2$$

όπου t_L και t_R είναι οι κόμβοι που δημιουργούνται από τον διαχωρισμός . Ο διαχωρισμός που επιλέγεται είναι αυτός ο οποίος μεγιστοποιεί το κριτήριο Twoing.

2.2.2.4 Οι κανόνες διακοπής - ολοκλήρωσης της διαδικασίας

Οι κανόνες διακοπής - ολοκλήρωσης της διαδικασίας ελέγχουν εάν η διαδικασία κατασκευής δένδρου πρέπει να σταματήσει ή όχι.

Χρησιμοποιούνται οι εξής κανόνες διακοπής (Breiman et al., 1984).

- Εάν ο κόμβος καταστεί καθαρός: δηλαδή εάν όλες οι περιπτώσεις μέσα σε έναν κόμβο έχουν πανομοιότυπες τιμές της εξαρτημένης μεταβλητής, τότε ο κόμβος δεν θα διαχωριστεί.

- Εάν όλες οι περιπτώσεις μέσα σε έναν κόμβο έχουν πανομοιότυπες τιμές για κάθε μεταβλητή πρόβλεψης, τότε ο κόμβος δεν θα διαχωριστεί.
- Εάν το βάθος του πρόσφατου δένδρου πλησιάζει την τιμή του μεγίστου ορίου βάθους το οποίο καθορίζεται από τον χρήστη, τότε η διαδικασία κατασκευής δένδρου θα σταματήσει.
- Εάν το μέγεθος ενός κόμβου είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από τον χρήστη, τότε ο κόμβος δεν θα διαχωριστεί.
- Εάν ο διαχωρισμός ενός κόμβου έχει ως αποτέλεσμα έναν θυγατρικό κόμβο του οποίου το μέγεθος είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από τον χρήστη, τότε ο κόμβος δεν θα διαχωριστεί.
- Ο καλύτερος διαχωρισμός για έναν κόμβο αποδίδει μία μείωση στην μη καθαρότητα, η οποία είναι μικρότερη από την ελάχιστη αλλαγή στην μη καθαρότητα που ορίζεται από τον χρήστη.

2.2.2.5 Τα κέρδη - κόστη

Κέρδη

Τα κέρδη είναι αριθμητικές τιμές οι οποίες σχετίζονται με τις κατηγορίες ενός συμβολικού πεδίου - στόχου τα οποία μπορούν να χρησιμοποιηθούν για να εκτιμήσουν το κέρδος ή τη ζημιά που σχετίζεται με ένα τμήμα. Καθορίζουν την σχετική τιμή κάθε καταχώρησης του πεδίου - στόχου. Οι τιμές χρησιμοποιούνται στον υπολογισμό των κερδών, αλλά όχι κατά την διάρκεια ανάπτυξης του δέντρου. Το κέρδος για κάθε κόμβο στο δένδρο υπολογίζεται ως

$$\sum_{j} f_j(t) \cdot P_j$$

όπου j είναι η κατηγορία του πεδίου - στόχου, $f_j(t)$ είναι το άθροισμα των τιμών των πεδίων συχνότητας για όλες τις καταχωρήσεις στον κόμβο t με κατηγορία j για το πεδίο - στόχο, P_j είναι η τιμή κέρδους για την κατηγορία j (καθορίζεται από τον χρήστη).

Κόστη

- Gini: εάν τα κόστη καθορίζονται, τότε ο δείκτης Gini υπολογίζεται ως

$$g(t) = \sum_{j \neq i} C(i|j)p(j|t)p(i|t)$$

όπου $C(i|j)$ είναι το κόστος της λανθασμένης ταξινόμησης μίας κατηγορίας j ως

κατηγορία i.

- Twoing: τα κόστη ενσωματώνονται στην εκχώρηση κόμβου και στην εκτίμηση του ρίσκου.

2.2.2.6 Οι prior πιθανότητες

Οι prior (πιθανότητες ορισμένες εκ των προτέρων) είναι αριθμητικές τιμές οι οποίες επηρεάζουν τα ποσοστά λανθασμένης ταξινόμησης για τις κατηγορίες του πεδίου - στόχου. Καθορίζουν την αναλογία των καταχωρήσεων που αναμένεται να ανήκουν σε κάθε κατηγορία του πεδίου - στόχου πριν από την ανάλυση. Οι τιμές των prior εμπλέκονται στην ανάπτυξη του δένδρου, καθώς και στην εκτίμηση του ρίσκου.

Υπάρχουν οι εξής τρεις τύποι prior πιθανοτήτων (Breiman et al., 1984):

- εμπειρικές prior, που υπολογίζονται βάσει training set
- ίσες prior, των οποίων η επιλογή ορίζει την prior πιθανότητα για καθεμία από τις j κατηγορίες στην ίδια τιμή
- prior καθορισμένες από τον χρήστη, των οποίων οι καθορισμένες - εξειδικευμένες τιμές χρησιμοποιούνται στους υπολογισμούς οι οποίοι περιέχουν prior.

2.2.2.7 Η διαδικασία κλαδέματος

Το «κλάδεμα» («pruning») αναφέρεται στην διαδικασία του ελέγχου ενός πλήρους αναπτυσσόμενου δένδρου και της αφαίρεσης των διαχωρισμών των κάτω επιπέδων που δεν έχουν σημαντική συνεισφορά στην ακρίβεια του δένδρου. Το λογισμικό στο κλάδεμα του δένδρου προσπαθεί να δημιουργήσει το μικρότερο δένδρο του οποίου το ρίσκο λανθασμένης ταξινόμησης δεν είναι πολύ μεγαλύτερο από το ρίσκο λανθασμένης ταξινόμησης του μεγαλύτερου πιθανού δένδρου. Η διαδικασία αφαιρεί ένα κλαδί δένδρου εάν το κόστος το οποίο σχετίζεται με την μεγαλύτερη πολυπλοκότητα του δένδρου είναι μεγαλύτερο από το κέρδος το οποίο σχετίζεται με το εάν έχουμε ένα άλλο επίπεδο κόμβων (κλαδί). Χρησιμοποιεί έναν δείκτη, ο οποίος μετρά το ρίσκο λανθασμένης ταξινόμησης και την πολυπλοκότητα του δένδρου, αφού στόχος μας είναι να ελαχιστοποιήσουμε και τα δύο.

Το μέτρο κόστους πολυπλοκότητας (cost complexity) ορίζεται ως

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

όπου R(T) είναι το ρίσκο λανθασμένης ταξινόμησης του δένδρου, $|\tilde{T}|$ είναι το πλήθος των τερματικών κόμβων για το δένδρο και α είναι το κόστος πολυπλοκότητας ανά τερματικό κόμβο για το δένδρο. Η τιμή α υπολογίζεται από τον αλγόριθμο κατά την διάρκεια του κλαδέματος (James et al., 2013).

Κάθε δένδρο που μπορούμε να παράγουμε έχει ένα μέγιστο μέγεθος (T_{max}), όπου σε κάθε τερματικό κόμβο περιέχεται μόνο μία καταχώρηση. Στην περίπτωση που το κόστος πολυπλοκότητας είναι μηδενικό (α=0), το μέγιστο δένδρο έχει το χαμηλότερο ρίσκο, αφού κάθε εγγραφή προβλέπεται τέλεια. Επομένως, όσο μεγαλύτερη είναι η τιμή του α τόσο μικρότερος είναι ο αριθμός των τερματικών κόμβων στο T(α), δηλαδή το δένδρο με το μικρότερο κόστος πολυπλοκότητας για το δοσμένο α. Όταν το α αυξάνεται από το 0, τότε παράγει μία πεπερασμένη ακολουθία από υποδένδρα (T_1, T_2, T_3), - καθένα με λιγότερους τερματικούς κόμβους από το προηγούμενο. Το κλάδεμα κόστους πολυπλοκότητας λειτουργεί αφαιρώντας τον πιο αδύναμο διαχωρισμό. Οι εξισώσεις που ακολουθούν εκφράζουν το κόστος πολυπλοκότητας για τον κόμβο {t}, που είναι ένας οποιοσδήποτε ξεχωριστός - μόνος κόμβος, και για T_t, τον υπο - κλάδο του {t}:

$$R_\alpha(\{t\}) = R(T) + \alpha$$

και

$$R_\alpha(\{T_t\}) = R(T_t) + \alpha|\tilde{T}_t|$$

Στην περίπτωση που το $R_\alpha(T_t)$ είναι μικρότερο από το $R_\alpha(\{t\})$, το κλαδί T_t έχει μικρότερο κόστος πολυπλοκότητας από αυτό του ξεχωριστού κόμβου {t}.

Η διαδικασία ανάπτυξης του δένδρου εξασφαλίζει ότι, για α=0, ισχύει

$$R_\alpha(\{t\}) \geq R_\alpha(T_t)$$

Καθώς το α αυξάνεται από το 0, τα $R_\alpha(\{t\})$ και $R_\alpha(T_t)$ αυξάνονται γραμμικώς με το $R_\alpha(T_t)$ να αυξάνεται με ταχύτερο ρυθμό. Τελικώς, βρίσκουμε ένα άνω φράγμα α' τέτοιο ώστε $R_\alpha(\{t\})<R_\alpha(T_t)$ για όλα τα α>α' . Συμπεραίνουμε ότι όταν το α καθίσταται μεγαλύτερο από το α', το κόστος πολυπλοκότητας του δένδρου μειώνεται εάν κόψουμε το υπο - κλαδί T_t κάτω από το {t}.

Μπορούμε να υπολογίσουμε το όριο ούτως ώστε να βρούμε την μεγαλύτερη τιμή του α, για την οποία ισχύει η ανισότητα, η οποία συμβολίζεται και ως g(t). Προκύπτει

$$\alpha \leq g(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

Μπορούμε να ορίσουμε τον πιο αδύναμο σύνδεσμο στο δένδρο T τον κόμβο, $\bar{t}$, ο οποίος λαμβάνει την μικρότερη τιμή του g(t):

$$g(\bar{t}) = \min\{g(t): t \epsilon T\}$$

Κατά συνέπεια, καθώς το α αυξάνεται, ο $\bar{t}$ είναι ο πρώτος κόμβος για τον οποίο ισχύει ότι $R_\alpha(\{t\})=R_\alpha(T_t)$. Στο σημείο αυτό, το {t} προτιμάται από το $T_{\bar{t}}$και το υπο - κλαδί κλαδεύεται.

Συνοπτικώς, ο αλγόριθμος κλαδέματος βασίζεται στα εξής βήματα (Breiman et al., 1984)

- Ορίζουμε το $\alpha_1=0$ και ξεκινούμε με το δένδρο για το οποίο $T_1=T(0)$, δηλαδή το πλήρως αναπτυσσόμενο δέντρο.
- Αυξάνουμε το α μέχρι το κλάδεμα ενός κλαδιού. Έπειτα, κλαδεύουμε το κλαδί από το δένδρο και υπολογίζουμε την εκτίμηση του ρίσκου του δένδρου το οποίο έχουμε κλαδέψει.
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να απομείνει μόνο ο αρχικός κόμβος - ρίζα , αποδίδοντας μία σειρά από υπο - δένδρα $T_1, T_2, ...,T_k$.
- Στην περίπτωση που επιλέξουμε τον κανόνα του τυπικού σφάλματος, τότε διαλέγουμε το μικρότερο δέντρο, T_{opt}, για το οποίο

$$R(T_{opt}) \leq \min_k R(T_k) \ + \ m \ x \ SE(R(T))$$

- Στην περίπτωση που δεν επιλέγουμε τον κανόνα τυπικού σφάλματος, τότε διαλέγουμε το δένδρο με τη μικρότερη τιμή της συνάρτησης ρίσκου R(T).

2.3 Περαιτέρω μέθοδοι

Η ενότητα 2.3 παρουσιάζει τρεις επιπλέον μεθόδους αξιολόγησης πιστωτικού κινδύνου: την διακριτική ανάλυση, τα νευρωνικά δίκτυα και την ανάλυση επιβίωσης.

2.3.1 Η διακριτική ανάλυση

Η διακριτική ανάλυση (discriminant analysis) αποτελεί μία τεχνική της Στατιστικής για εκτίμηση της καταλληλότητας μίας ταξινόμησης και, επίσης, χρησιμοποιείται για τον διαχωρισμό δύο συνόλων (ομάδων). Σκοπός της είναι να αναγνωρίσει τις μεταβλητές που διαφοροποιούνται ανάμεσα στα δύο σύνολα. Η περιγραφή της μεθόδου είναι η ακόλουθη (Thomas et al., 2002).

Έστω $\mathbf{X}=(X_1,X_2,...,X_p)$ το σύνολο p - το πλήθος - τυχαίων μεταβλητών οι οποίες περιγράφουν τις πληροφορίες που διατίθενται για έναν αιτούντα πίστωση. Οι λέξεις «μεταβλητή» και «χαρακτηριστικό» χρησιμοποιούνται εναλλακτικώς ώστε να περιγραφεί ένα τυπικό X_i: η πρώτη όταν θέλουμε να δώσουμε έμφαση στην τυχαία φύση αυτής της πληροφορίας ανάμεσα σε αιτούντες και η δεύτερη όταν θέλουμε να θυμηθούμε τί είδος πληροφορίας είναι. Η τιμή των μεταβλητών για έναν συγκεκριμένο αιτούντα δηλώνεται ως $\mathbf{x}=(x_1,x_2,...,x_p)$.

Στην ορολογία της βαθμολόγησης πιστοληπτικής ικανότητας, οι διαφορετικές δυνατές τιμές ή απαντήσεις, x_i, για το χαρακτηριστικό X_i, καλούνται «ιδιότητες» εκείνου του χαρακτηριστικού. Επομένως, εάν ένα τυπικό χαρακτηριστικό είναι το καθεστώς διαμονής του αιτούντα, τότε οι

ιδιότητές του ενδέχεται να είναι «ιδιοκτήτης», «ενοικιαστής μη επιπλωμένης οικίας», «ενοικιαστής επιπλωμένης οικίας», «διαμονή με γονείς», ή άλλες. Διαφορετικοί δανειστές ενδέχεται να έχουν διαφορετικές ομάδες ιδιοτήτων για το ίδιο χαρακτηριστικό. Επομένως, ένας άλλος δανειστής ενδέχεται να αποφασίσει να ταξινομήσει το καθεστώς διαμονής σε «ιδιοκτήτης χωρίς υποθήκη», «ιδιοκτήτης με υποθήκη», «ενοικιαστής μη επιπλωμένης ιδιοκτησίας», «ενοικιαστής επιπλωμένης ιδιοκτησίας», «ιδιοκτησία σε επινοικίαση», «κινητή οικία», «προσφερόμενο κατάλυμα», «διαμονή με γονείς», «διαμονή με άλλους, όχι γονείς», ή αλλιώς. Δεν είναι ασύνηθες να συγχέει κανείς την έννοια της ιδιότητας με εκείνη του χαρακτηριστικού. Ένας εύκολος τρόπος να τα διαχωρίζει κανείς είναι ότι η ιδιότητα αποτελεί την απάντηση στην ερώτηση της φόρμας αίτησης ενώ το χαρακτηριστικό αποτελεί την ερώτηση.

Επιστρέφοντας στην απόφαση που πρέπει να ληφθεί από τον οργανισμό δανειοδότησης (για το εάν θα δοθεί η αιτούμενη πίστωση ή όχι) , ας υποθέσουμε ότι A είναι το σύνολο όλων των δυνατών τιμών που λαμβάνουν οι μεταβλητές $\mathbf{X}=(X_1,X_2,...,X_p)$ της αίτησης - δηλαδή όλοι οι διαφορετικοί τρόποι που δύναται απαντηθεί η φόρμα αίτησης. Ο σκοπός μας είναι να χωρίσουμε το σύνολο A σε δύο υποσύνολα, A_Gκαι A_B, ώστε η ταξινόμηση των αιτούντων των οποίων οι απαντήσεις ανήκουν στο A_Gως «καλοί» και η αποδοχή τους και η ταξινόμηση των αιτούντων των οποίων οι απαντήσεις ανήκουν στο A_Bως «κακοί» και η απόρριψή τους να ελαχιστοποιεί το αναμενόμενο κόστος για τον δανειστή.

Τα δύο είδη κόστους ανταποκρίνονται στα δύο είδη σφάλματος που δύναται να συμβούν σε αυτή την απόφαση. Κάποιος ενδέχεται να ταξινομήσει κάποιον που είναι «καλός» ως «κακό» και να τον απορρίψει. Σε αυτή την περίπτωση, το δυνητικό κέρδος από αυτόν τον αιτούνται χάνεται. Ας υποθέσουμε ότι το αναμενόμενο κέρδος είναι ίδιο, ίσο με L, για κάθε αιτούντα. Το δεύτερο σφάλμα είναι τα ταξινομηθεί ως «καλός» ένας «κακός» και, επομένως, να καταστεί αποδεκτός.

Σε αυτή την περίπτωση, θα προκύψει χρέος όταν ο πελάτης αθετήσει τις υποχρεώσεις του ως προς το δάνειο. Υποθέτουμε ότι το προκύπτον αναμενόμενο χρέος είναι ίδιο, ίσο με D, για όλους τους πελάτες.

Ας υποθέσουμε ότι p_G είναι το ποσοστό των αιτούντων που είναι «καλοί». Αντιστοίχως, έστω ότι p_B είναι το ποσοστό των αιτούντων που είναι «κακοί». Ας υποθέσουμε ότι τα χαρακτηριστικά της αίτησης έχουν ένα πεπερασμένο πλήθος διακριτών ιδιοτήτων, έτσι ώστε το A να είναι πεπερασμένο και να υπάρχει μόνο ένα πεπερασμένο πλήθος διακριτών ιδιοτήτων **x**. Αυτό είναι ισοδύναμο με τον ισχυρισμό ότι υπάρχει μόνο ένα πεπερασμένο πλήθος τρόπων να συμπληρωθεί η φόρμα αίτησης.

Έστω ότι p(**x**|G) είναι η πιθανότητα ένας «καλός» αιτών να έχει τις ιδιότητες **x**. Αυτή είναι μία δεσμευμένη πιθανότητα και αναπαριστά τον λόγο

$$p(\mathbf{x}|G) = \frac{\text{Πιθανότητα (ο αιτών είναι «καλός» και έχει τις ιδιότητες } \mathbf{x})}{\text{Πιθανότητα (ο αιτώνείναι «καλός»)}}$$

Αντιστοίχως, ορίζεται ως p(**x**|B) η πιθανότητα ένας «κακός» αιτών να έχει τις ιδιότητες **x**.

Εάν ορίζεται ως q(G|**x**) η πιθανότητα κάποιος με ιδιότητες αίτησης **x** να είναι «καλός», τότε

$$q(G|\mathbf{x}) = \frac{\text{Πιθανότητα (ο αιτών έχει τις ιδιότητες } \mathbf{x} \text{ και είναι «καλός»)}}{\text{Πιθανότητα (ο αιτών έχει τις ιδιότητες } \mathbf{x})}$$

και εάν p(**x**)= Πιθανότητα (ο αιτών έχει τις ιδιότητες **x**), τότε οι δύο τελευταίες σχέσεις δίδουνπιθανότητα (ο αιτών έχει τις ιδιότητες **x** και είναι «καλός») = $q(G|\mathbf{x})\cdot p(\mathbf{x}) = p(\mathbf{x}|G)\cdot p_G$

Επομένως, φθάνουμε στο θεώρημα Bayes, σύμφωνα με το οποίο

$$q(G|\mathbf{x}) = \frac{p(\mathbf{x}|G)\cdot p_G}{p(\mathbf{x})}$$

Ένα αντίστοιχο αποτέλεσμα προκύπτει για το q(B|**x**), την πιθανότητα κάποιος με χαρακτηριστικά αίτησης **x** να είναι «κακός»

$$q(B|\mathbf{x}) = \frac{p(\mathbf{x}|B)\cdot p_B}{p(\mathbf{x})}$$

Από τις δύο τελευταίες σχέσεις προκύπτει ότι

$$\frac{q(G|\mathbf{x})}{q(B|\mathbf{x})} = \frac{p(\mathbf{x}|G)\cdot p_G}{p(\mathbf{x}|B)\cdot p_B}$$

Το αναμενόμενο κόστος ανά αιτούντα, εάν αποδεχθούμε αιτούντες με ιδιότητες στο A_G και απορρίψουμε εκείνους με ιδιότητες στο A_B, είναι

$$L\cdot\sum_{\mathbf{x}\ \text{στο AB}} p(\mathbf{x}|G)\cdot p_G + D\cdot\sum_{\mathbf{x}\ \text{στο AG}} p(\mathbf{x}|B)\cdot p_B =$$

$$L\cdot\sum_{\mathbf{x}\ \text{στο AB}} q(G|\mathbf{x})\cdot p(\mathbf{x}) + D\cdot\sum_{\mathbf{x}\ \text{στο AG}} q(B|\mathbf{x})\cdot p(\mathbf{x})$$

Ο κανόνας που ελαχιστοποιεί το αναμενόμενο κόστος προκύπτει κατ' ευθείαν. Ας σκεφθούμε ποιά είναι τα δύο κόστη εάν κατηγοριοποιήσουμε ένα συγκεκριμένο $\mathbf{x}=(x_1,x_2,...,x_p)$ σε A_G ή A_B. Εάν τοποθετηθεί στο A_G, τότε υπάρχει μόνο ένα κόστος - εάν είναι «κακό» - στην οποία περίπτωση το αναμενόμενο κόστος είναι $D\cdot p(\mathbf{x}|B)\cdot p_B$. Εάν το x ταξινομηθεί στο A_B, τότε υπάρχει μόνο ένα κόστος - εάν είναι «καλό» - και, επομένως, το αναμενόμενο κόστος είναι $L\cdot p(\mathbf{x}|G)\cdot p_G$. Επομένως, ταξινομεί κανείς το x στο A_G εάν $D\cdot p(\mathbf{x}|B)\cdot p_B \leq L\cdot p(\mathbf{x}|G)\cdot p_G$. Άρα, ο κανόνας απόφασης προκειμένου να ελαχιστοποιηθούν τα αναμενόμενα κόστη δίδεται από

την σχέση (κριτήριο) $A_G = \{\mathbf{x}|D \cdot p(\mathbf{x}|B) \cdot p_B \leq L \cdot p(\mathbf{x}|G) \cdot p_G\} = \{\mathbf{x}|\frac{D}{L} \leq \frac{p(\mathbf{x}|G) \cdot p_G}{p(\mathbf{x}|B) \cdot p_B}\} = \{\mathbf{x}|\frac{D}{L} \leq \frac{q(G|\mathbf{x})}{q(B|\mathbf{x})}\}$.

Τέλος, κάποια παραδείγματα που χρησιμοποιείται η διακριτική ανάλυση είναι τα ακόλουθα

- ένας ερευνητής ερευνά παράγοντες που διαφοροποιούνται σημαντικά ανάμεσα σε ασθενείς που επεβίωσαν και σε ασθενείς που απεβίωσαν. Έπειτα, θέλει - με βάση τους παράγοντες αυτούς - να προβλέψει την πιθανότητα να επιβιώσεως ενός ασθενούς μελλοντικώς.
- ο υπεύθυνος πωλήσεων εταιρείας ενδιαφέρεται να εντοπίσει εμφανή γνωρίσματα ανάμεσα σε αγοραστές και μη των προϊόντων της και να χρησιμοποιήσει την πληροφορία που εξάγεται για πρόβλεψη των αγοραστικών σκοπών των καταναλωτών μελλοντικώς.

Βεβαίως, ένα ακόμη παράδειγμα αποτελεί το εξής: ένας πιστωτικός οργανισμός εξετάζει παράγοντες που διαφοροποιούνται ανάμεσα σε δανειολήπτες που αθέτησαν τις υποχρεώσεις τους και σε αυτούς που δεν τις αθέτησαν, επιδιώκοντας να προβλέψει την πιθανότητα αθέτησης από πλευράς ενός μελλοντικού δανειολήπτη.

2.3.2 Τα νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα (neural networks) αποτελούν μία μη στατιστική μέθοδο, η οποία αναπτύχθηκε από προσπάθειες μοντελοποίησης της επικοινωνίας στον ανθρώπινο εγκέφαλο.

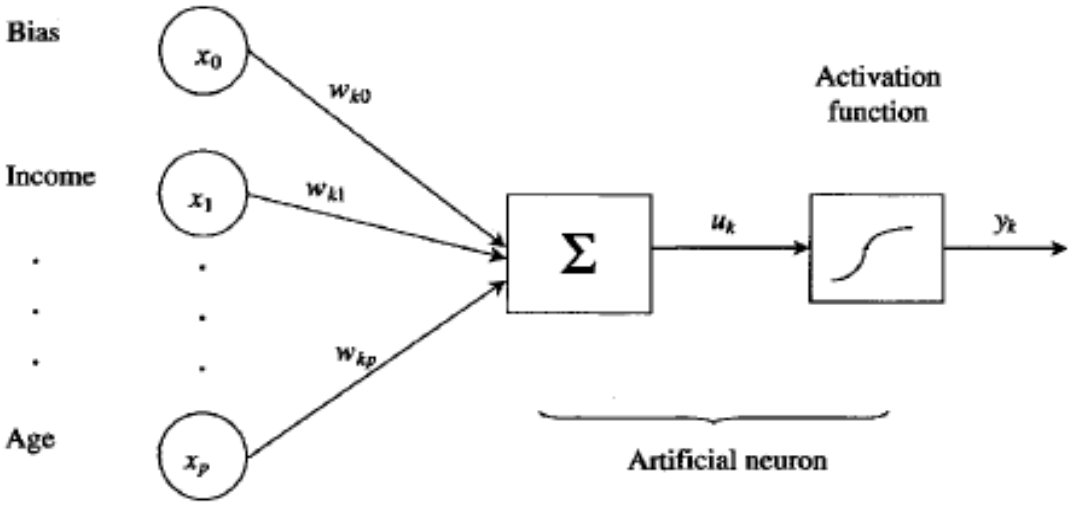
Νευρωνικό δίκτυο είναι ένας μαζικός παράλληλος επεξεργαστής, ο οποίος - εκ φύσεως - αποθηκεύει εμπειρική γνώση και την καθιστά διαθέσιμη για χρήση. Προσομοιάζει τον ανθρώπινο εγκέφαλο στους εξής δύο τομείς

- η γνώση αποκτάται από το δίκτυο μέσω μίας διαδικασίας μάθησης.
- οι ενδονευρωνικές συνδέσεις χρησιμοποιούνται για την φύλαξη γνώσης.

(Haykin, 1999)

Ένα νευρωνικό δίκτυο αποτελείται από ένα πλήθος εισόδων - μεταβλητών, καθεμία εκ των οποίων πολλαπλασιάζεται με ένα βάρος. Τα γινόμενα προστίθενται και μετασχηματίζονται σε έναν «νευρώνα» και, έπειτα, το αποτέλεσμα καθίσταται είσοδος - μεταβλητή για έναν άλλο νευρώνα. (Thomas et al., 2002)

Σε ένα **νευρωνικό δίκτυο μονού στρώματος** (**single - layer neural network**), η μετασχηματισμένη τιμή αποτελεί την αναζητούμενη τιμή - αντί να καταστεί είσοδος για έναν άλλο νευρώνα. Η Εικόνα 2.3 παρουσιάζει ένα νευρωνικό δίκτυο μονού στρώματος.



Εικόνα 2.3 Νευρωνικό δίκτυο μονού στρώματος

Μπορούμε να αναπαραστήσουμε ένα νευρωνικό δίκτυο μονού στρώματος αλγεβρικώς ως εξής (Thomas et al., 2002)

$$u_k = w_{k0}x_0 + w_{k1}x_1 + \ldots + w_{kp}x_p = \sum_{q=0}^{p} w_{kq}x_q$$

$$y_k = F(u_k)$$

Καθένα από τα x_1, ..., x_p είναι μία μεταβλητή, όπως ένα χαρακτηριστικό μίας πιστωτικής κάρτας αιτούντος. Καθεμία προσλαμβάνει μία τιμή, γνωστή ως «σήμα». Τα βάρη, γνωστά και ως «συναπτικά βάρη», εάν είναι θετικά τότε είναι γνωστά ως «διεγερτικά» καθώς αυξάνουν την τιμή της μεταβλητής απόκρισης και εάν είναι αρνητικά τότε καλούνται «κωλυτικά» καθώς μειώνουν την τιμή u_k για θετικές μεταβλητές.

Ας επισημανθεί ότι οι δείκτες σε κάθε βάρος γράφονται με την σειρά (k,p), όπου το k υποδεικνύει τον νευρώνα στον οποίο εφαρμόζεται το βάρος και το p υποδεικνύει την μεταβλητή. Σε ένα νευρωνικό δίκτυο μονού στρώματος, k=1 διότι υπάρχει μόνο ένας νευρώνας. Ας επισημανθεί, επίσης, ότι στην μεταβλητή x_0 αποδίδεται η τιμή +1 έτσι ώστε ο όρος $w_{k0}.x_0$ είναι, απλώς, w_{k0}, γνωστός και ως «μεροληψία» («bias»). Αυτός έχει την λειτουργία αύξησης ή μείωσης της τιμής u_k κατά μία σταθερή ποσότητα.

Η τιμή u_k μετασχηματίζεται λοιπόν, χρησιμοποιώντας μία συνάρτηση ενεργοποίησης (ή μεταφοράς ή σύνθλιψης). Στα πρώτα δίκτυα, αυτή η συνάρτηση ήταν γραμμική - κάτι που περιόριζε σε μεγάλο βαθμό την κλάση των προβλημάτων που μπορούσαν να αντιμετωπίσουν αυτά τα δίκτυα. Διάφορες εναλλακτικές συναρτήσεις μεταφοράς χρησιμοποιούνται και περιλαμβάνουν τις ακόλουθες.

- Συνάρτηση κατωφλίου:

F(u)=1 εάν u≥0

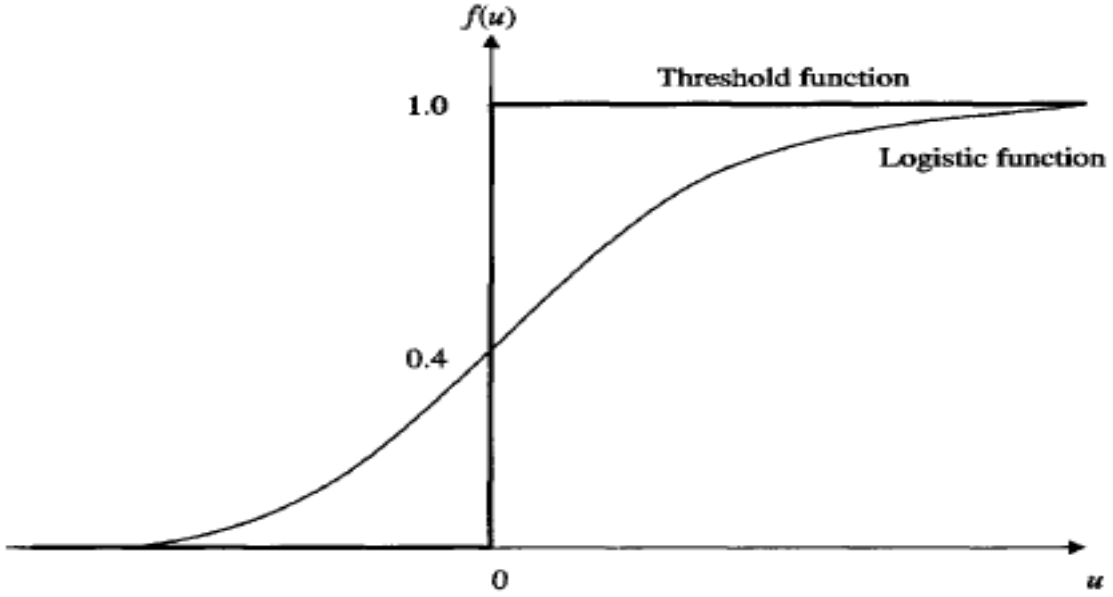
F(u)=0 εάν u<0

το οποίο σημαίνει ότι εάν το u ισούται με ή είναι μεγαλύτερο από το 0, τότε ο νευρώνας έχει την τιμή 1 ως έξοδο - αλλιώς, έχει την τιμή 0 ως έξοδο.

- Λογιστική συνάρτηση:

$$\mathrm{F(u)} = \frac{1}{1 + \mathrm{e}^{-\mathrm{au}}}$$

Και οι δύο συναρτήσεις αναπαριστώνται στην Εικόνα 2.4.



Εικόνα 2.4 Ενδεικτικές συναρτήσεις μεταφοράς

Η τιμή α στην λογιστική συνάρτηση καθορίζει την κλίση της καμπύλης. Και οι δύο συναρτήσεις περιορίζουν την έξοδο του δικτύου να βρίσκεται εντός του εύρους (0,1). Κάποιες φορές, επιθυμούμε η έξοδος να βρίσκεται εντός του εύρους (-1,+1) και, επομένως, χρησιμοποιούμε την συνάρτηση υπερβολικής εφαπτομένης, F(u)=tanh(h).

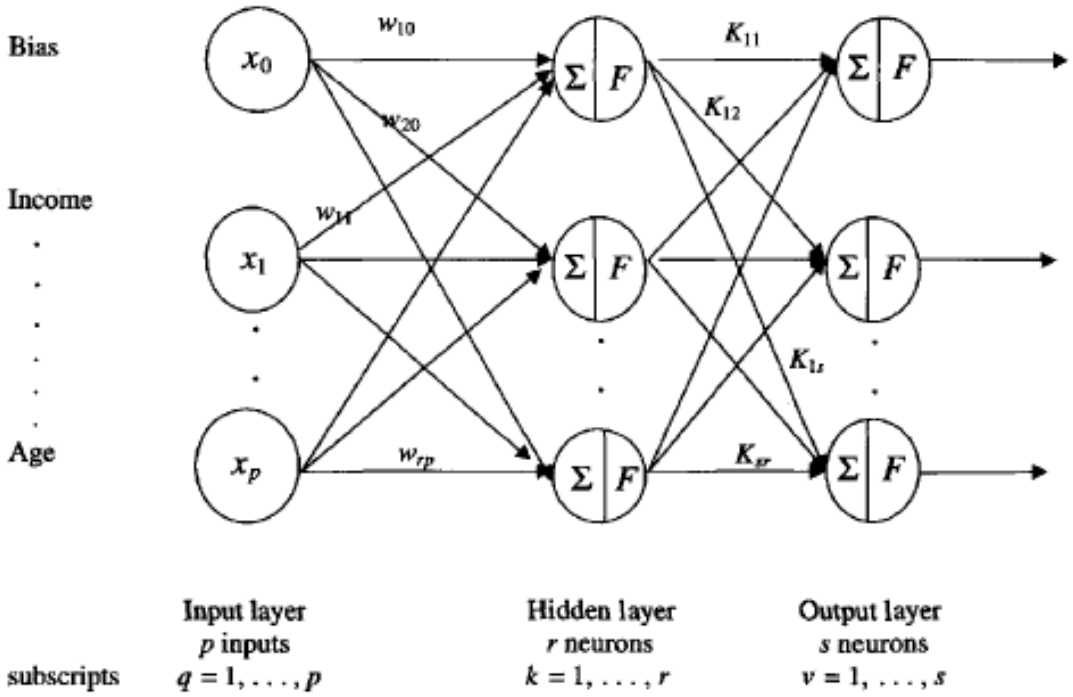
Δεδομένων των τιμών για τα βάρη και την συνάρτηση μεταφοράς, μπορούμε να προβλέψουμε εάν μία αίτηση για πιστωτική κάρτα πρόκειται να καταστεί αποδεκτή ή να απορριφθεί, υπολογίζοντας την τιμή y_k χρησιμοποιώντας τα χαρακτηριστικά του αιτούντος και συγκρίνοντας αυτή την τιμή με μία τιμή διαχωρισμού.

Ένα μοντέλο, το οποίο αποτελείται από έναν μοναδικό νευρώνα και μία συνάρτηση ενεργοποίησης κατωφλίου, καλείται «αισθητήρας» («perceptron»).

Από την άλλη, ένας **αισθητήρας πολλαπλών στρωμάτων** (**multilayer perceptron**) αποτελείται από ένα στρώμα εισόδου σημάτων, ένα στρώμα εξόδου σημάτων εξόδου (διαφορετικές y_v

τιμές) και ένα πλήθος από στρώματα νευρώνων ανάμεσα, που καλούνται «κρυμμένα στρώματα», και περιγράφεται ακολούθως (Thomas et al., 2002).

Κάθε νευρώνας σε ένα κρυμμένο στρώμα έχει ένα σύνολο από βάρη που εφαρμόζονται στις εισόδους του, τα οποία ενδέχεται να διαφέρουν από εκείνα που εφαρμόζονται στις ίδιες εισόδους που οδεύουν προς έναν διαφορετικό νευρώνα στο κρυμμένο στρώμα. Οι έξοδοι από κάθε νευρώνα σε ένα κρυμμένο στρώμα έχουν εφαρμοζόμενα βάρη και καθίστανται είσοδοι για νευρώνες στο επόμενο κρυμμένο στρώμα - εάν υπάρχει τέτοιο. Αλλιώς, καθίστανται είσοδοι στο στρώμα εξόδου. Το στρώμα εξόδου δίδει τις τιμές για καθέναν από τους νευρώνες - μέλη του, των οποίων οι τιμές συγκρίνονται με τιμές διαχωρισμού ώστε να ταξινομηθεί κάθε περίπτωση. Ένα δίκτυο τριών στρωμάτων παρουσιάζεται στην Εικόνα 2.5.



Εικόνα 2.5 Ενδεικτικό δίκτυο τριών στρωμάτων

Μπορούμε να αναπαραστήσουμε έναν αισθητήρα πολλαπλών στρωμάτων, αλγεβρικώς, όπως ακολούθως (θα το πραγματοποιήσουμε για τον αισθητήρα της Εικόνας 2.5). Έχουμε

$$y_k = F_1(\sum_{q=0}^{p} w_{kq}x_q)$$

όπου ο δείκτης 1 στο F υποδεικνύει ότι είναι το πρώτο στρώμα μετά το στρώμα εισόδου. Τα y_k, k=1,...,r είναι οι έξοδοι από το πρώτο κρυμμένο στρώμα. Καθώς η έξοδος από ένα στρώμα είναι η είσοδος στο επόμενο στρώμα, μπορούμε να γράψουμε

$$z_v = F_2(\sum_{k=1}^{r} K_{vk}y_k) = F_2(\sum_{k=1}^{r} K_{vk}(F_1(\sum_{q=0}^{p} w_{kq}x_q)))$$

όπου z_v είναι η έξοδος του νευρώνα v στο στρώμα εισόδου, v=1,...,s, F_2 είναι η συνάρτηση ενεργοποίησης στο στρώμα εξόδου, και K_{vk}είναι το βάρος που εφαρμόζεται στο στρώμα y_kτο οποίο συνδέεται με τον νευρώνα k στο κρυμμένο στρώμα και με τον νευρώνα v στο στρώμα εξόδου.

Ως προς τα πλεονεκτήματα της μεθόδου των νευρωνικών δικτύων, η μέθοδος διέπεται από την ικανότητα γενίκευσης των αποτελεσμάτων (έλλειψη ρητής περιγραφής του προβλήματος), του χειρισμού μεγάλου όγκου δεδομένων, της ύπαρξης λιγότερων στατιστικών υποθέσεων (Hooman et al., 2015).

Από την άλλη, τα κύρια μειονεκτήματά της είναι το κόστος εφαρμογής και η συντήρησή τους, η ταχύτητα εξαγωγής αποτελεσμάτων (εξ' αιτίας του πιθανόν μεγάλου πλήθους συσχετίσεων), αλλά και η ισχύς των αποτελεσμάτων δεδομένου ότι, πολλές φορές, τα βήματα που χρησιμοποιούνται δεν ερμηνεύονται από την οικονομική θεωρία (Θωμαδάκης & Ξανθάκης, 2006).

2.3.3 Η ανάλυση επιβίωσης

Τα συστήματα βαθμολόγησης πιστοληπτικής ικανότητας κατασκευάστηκαν προκειμένου να απαντήσουν στο ερώτημα «πόσο πιθανό είναι ένας αιτών πίστωση να αθετήσει τις υποχρεώσεις του σε ένα δεδομένο χρονικό διάστημα στο μέλλον;» Η μεθοδολογία συνίσταται στην λήψη ενός δείγματος προηγούμενων πελατών και στην ταξινόμησή τους σε «καλούς» ή «κακούς» αναλόγως της συμπεριφοράς αποπληρωμής τους κατά μία δεδομένη συγκεκριμένη χρονική περίοδο.

Ωστόσο, «κακή» συμπεριφορά αμέσως πριν την λήξη της χρονικής περιόδου αυτής συνεπάγεται ότι ο πελάτης ταξινομείται ως «κακός» - «κακή» συμπεριφορά αμέσως μετά την λήξη της χρονικής περιόδου αυτής δεν έχει σημασία και ο πελάτης ταξινομείται ως «καλός». (Thomas et al., 2002). Σε πολλές εφαρμογές του πιστωτικού κινδύνου, λοιπόν, επιθυμούμε να εκτιμήσουμε όχι μόνο εάν αλλά και πότε ένας δανειζόμενος θα αθετήσει την συμφωνία ενός δανείου[8]. Με το τελευταίο ασχολείται η «ανάλυση επιβίωσης» (survival analysis) και η χρήση μοντέλων επιβίωσης έχει αρκετά πλεονεκτήματα σε σχέση με την χρήση της παραδοσιακής τεχνικής της λογιστικής παλινδρόμησης (Shumway, 2001).

Στην ανάλυση επιβίωσης υπάρχουν δύο βασικές συναρτήσεις, η συνάρτηση επιβίωσης (survival function) και η συνάρτηση διακινδύνευσης (hazard function), οι οποίες περιγράφονται παρακάτω (Collett, 2003).

[8] Στις περιπτώσεις αυτές είναι πολύ χρήσιμο το ημιπαραμετρικό μοντέλο παλινδρόμησης που προτάθηκε από τον Cox (1972).

Ο πραγματικός χρόνος επιβίωσης μίας (υπό μελέτη) μονάδας, t, δύναται να θεωρηθεί ως η τιμή μίας μεταβλητής, T, που δύναται να λάβει οποιαδήποτε μη αρνητική τιμή. Οι διαφορετικές τιμές που δύναται να λάβει η T έχουν μία κατανομή πιθανότητας, και καλούμε την T ως την «τυχαία μεταβλητή» που σχετίζεται με τον χρόνο επιβίωσης.

Τώρα, ας υποθέσουμε ότι η T έχει μία κατανομή πιθανότητας με υποκείμενη συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) f(t). Τότε η συνάρτηση κατανομής (σ.κ.) της T δίδεται ως

$$\mathrm{F(t)} \;=\; \mathrm{P[T < t]} \;=\; \int_{0}^{\mathrm{t}} \mathrm{f(u)du}$$

και αναπαριστά την πιθανότητα ο χρόνος επιβίωσης να είναι μικρότερος από κάποια τιμή t.

Η συνάρτηση επιβίωσης, S(t), ορίζεται ως η πιθανότητα ο χρόνος επιβίωσης (T) να είναι μεγαλύτερος από ή ίσος με κάποια τιμή t, και επομένως

$$\mathrm{S(t)} = \;\mathrm{P[T \geq t]} = 1 - \mathrm{F(t)}$$

Η συνάρτηση επιβίωσης, επομένως, δύναται να χρησιμοποιηθεί προκειμένου να αναπαραστήσει την πιθανότητα μία μονάδα να επιβιώσει από την αρχική χρονική στιγμή μέχρι κάποια χρονική στιγμή μετά την t.

Η συνάρτηση διακινδύνευσης, h(t), χρησιμοποιείται ευρέως προκειμένου να εκφράσει τον κίνδυνο αποβίωσης σε κάποιον χρόνο t, και προκύπτει από την πιθανότητα h(t)δt μία μονάδα να αποβιώσει στο διάστημα [t, t+δt) δεδομένου ότι έχει επιβιώσει μέχρι εκείνη την χρονική στιγμή. Η h(t) ορίζεται ως

$$\mathrm{h(t)} = \lim_{\delta t \to 0} \{\frac{\mathrm{P[t \leq T < t + \delta t | T \geq t]}}{\delta t}\}$$

Από την δεσμευμένη πιθανότητα στον παραπάνω ορισμό και με F(t) την σ.κ. του T, λαμβάνεται

$$\mathrm{P[t \leq T < t + \delta t | T \geq t]} \;=\; \frac{\mathrm{P[t \leq T < t + \delta t]}}{\mathrm{P[T \geq t]}} = \frac{F(\mathrm{t} + \delta t) - \mathrm{F(t)}}{\mathrm{S(t)}}$$

Οπότε

$$\mathrm{h(t)} = \lim_{\delta \mathrm{t} \to 0} \{\frac{\mathrm{F(t + \delta t) - F(t)]}}{\delta \mathrm{t}}\} \cdot \frac{1}{\mathrm{S(t)}}$$

Και επειδή

$$\mathrm{f(t)} = \lim_{\delta \mathrm{t} \to 0} \{\frac{\mathrm{F(t + \delta t) - F(t)]}}{\delta \mathrm{t}}\}$$

έχουμε τελικώς

$$h(t) = \frac{f(t)}{S(t)}$$

Μία ακόμη χρήσιμη συνάρτηση στην ανάλυση επιβίωσης είναι η «σωρευτική συνάρτηση διακινδύνευσης» («cumulative hazard function»), H(t), η οποία ορίζεται ως

$$H(t) = \int_0^t h(u)du$$

Από τα ανωτέρω, υπολογίζεται η σχέση της συνάρτησης αυτής με τα υπόλοιπα μεγέθη,

$$H(t) = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t -\frac{S'(u)}{S(u)}du = -\ln S(t)$$

και, επομένως,

$$S(t) = e^{-H(t)}$$

Παρατηρείται ότι εάν γνωρίζουμε κάποια από τις συναρτήσεις h(t), f(t), S(t), F(t), H(t), τότε δύναται να υπολογίσουμε οποιαδήποτε από τις υπόλοιπες συναρτήσεις.

3. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Το Κεφάλαιο 3 περιλαμβάνει ανάλυση δεδομένων με χρήση της R. Η R είναι μία γλώσσα προγραμματισμού που χρησιμεύει κυρίως για ανάλυση δεδομένων και εφαρμογή διαφόρων «κλασσικών» και «σύγχρονων» στατιστικών τεχνικών και οι τελευταίες εκδόσεις της είναι αποτέλεσμα μίας συλλογικής προσπάθειας με συνεισφορές ερευνητών από όλον τον κόσμο (Φουσκάκης, 2013).

3.1 Παρουσίαση

Βάσει ενός δείγματος 10000 πελατών και με χρήση τριών χαρακτηριστικών τους, επιχειρήθηκε η εκτίμηση της πιθανότητας αθέτησης της υποχρέωσης αποπληρωμής της πιστωτικής κάρτας τους. Τα προαναφερθέντα τρία χαρακτηριστικά αποτυπώθηκαν στις εξής ισάριθμες συμμεταβλητές: ύπαρξη φοιτητικής ιδιότητας από πλευράς του πελάτη (student) (κατηγορική συμμεταβλητή), καθώς και μηνιαίο υπόλοιπο πιστωτικής κάρτας του πελάτη (balance) και ετήσιο εισόδημά του (income) (ποσοτικές συμμεταβλητές).

Κατ' αρχάς, οι 10000 παρατηρήσεις χωρίστηκαν σε δύο ομάδες - επιμέρους δείγματα: το training set και το test set. Το πρώτο, περιλαμβάνοντας το 75% των παρατηρήσεων, χρησιμοποιήθηκε για την προσαρμογή των μοντέλων και το δεύτερο, περιλαμβάνοντας το υπόλοιπο 25% των παρατηρήσεων, για την αξιολόγηση της προβλεπτικής ικανότητάς τους.

Έπειτα, η μεταβλητή student χωρίστηκε σε δύο κατηγορίες. Η κατηγορία που λαμβάνει την τιμή "1" περιλαμβάνει τους πελάτες που είναι φοιτητές και, αντιστοίχως, η κατηγορία που λαμβάνει την τιμή "2" περιλαμβάνει τους πελάτες που δεν είναι φοιτητές.

Επίσης, όσον αφορά στην **λογιστική παλινδρόμηση**, επιχειρήθηκαν δύο προσεγγίσεις. Κατά την πρώτη προσέγγιση, κατηγοριοποιήθηκαν οι συμμεταβλητές balance και income. Πιο συγκεκριμένα, η μεταβλητή balance χωρίστηκε σε δύο κατηγορίες. Η κατηγορία που λαμβάνει την τιμή "1" περιλαμβάνει τους πελάτες με μηνιαίο υπόλοιπο πιστωτικής κάρτας ≤ 1800 και η κατηγορία που λαμβάνει την τιμή "2" περιλαμβάνει τους πελάτες με μηνιαίο υπόλοιπο πιστωτικής κάρτας > 1800. Η μεταβλητή income χωρίστηκε σε τρεις κατηγορίες. Η κατηγορία που λαμβάνει την τιμή "1" περιλαμβάνει τους πελάτες με ετήσιο εισόδημα ≤ 20000, η κατηγορία που λαμβάνει την τιμή "2" περιλαμβάνει τους πελάτες με ετήσιο εισόδημα > 20000 και ≤ 40000 και, τέλος, η κατηγορία που λαμβάνει την τιμή "3" περιλαμβάνει τους πελάτες με ετήσιο εισόδημα > 40000. Κατά την δεύτερη προσέγγιση, η οποία χρησιμοποιήθηκε και στην περίπτωση των **δένδρων απόφασης**, οι συμμεταβλητές balance και income παρέμειναν ποσοτικές.

3.2 Ο πιστωτικός κίνδυνος με λογιστική παλινδρόμηση - πρώτη προσέγγιση

Για την ανάλυση των δεδομένων προσαρμόζοντας ένα μοντέλο λογιστικής παλινδρόμησης, η εξαρτημένη μεταβλητή θα είναι η Y, η οποία είναι δίτιμη και εκφράζει εάν υπήρξε αθέτηση της υποχρέωσης αποπληρωμής της πιστωτικής κάρτας - οπότε λαμβάνει την τιμή 1 - ή όχι - οπότε λαμβάνει την τιμή 0.

Κατ' αρχάς, για μία πιο εμπεριστατωμένη μελέτη αυτών των training set και test set, κρίθηκε σκόπιμο να κατασκευαστούν πίνακες συχνοτήτων για τις συμμεταβλητές. Οι εν λόγω πίνακες είναι οι 3.1, 3.2, 3.3.

	Training Set		**Test Set**	
student	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**
"1"	2176	29.01%	768	30.72%
"2"	5324	70.99%	1732	69.28%
Σύνολο	**7500**	**100.00%**	**2500**	**100.00%**

Πίνακας 3.1 Συχνότητες για την συμμεταβλητή student

	Training Set		**Test Set**	
balance	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**
"1"	7287	97.16%	2425	97.00%
"2"	213	2.84%	75	3.00%
Σύνολο	**7500**	**100.00%**	**2500**	**100.00%**

Πίνακας 3.2 Συχνότητες για την συμμεταβλητή balance

	Training Set		Test Set	
income	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**
“1”	1612	21.49%	550	22.00%
“2”	3245	43.27%	1096	43.84%
“3”	2643	35.24%	854	34.16%
Σύνολο	**7500**	**100.00%**	**2500**	**100.00%**

Πίνακας 3.3 Συχνότητες για την συμμεταβλητή income

Έπειτα, για τις 7500 παρατηρήσεις του training set, κατασκευάστηκαν οι Πίνακες 3.4, 3.5 και 3.6, οι οποίοι είναι two - way. πίνακες με βάση την αθέτηση της υποχρέωσης αποπληρωμής της πιστωτικής κάρτας (Y), και παρουσιάζουν τις συχνότητες εμφάνισης κάθε τιμής σε κάθε μεταβλητή, δίδοντας μία πρώτη εικόνα για την εξάρτηση της αθέτησης από κάθε ανεξάρτητη μεταβλητή.

			Αθέτηση		
student	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-value X^2 ελέγχου**
“1”	2176	29.01%	101	4.64%	0.0005827
“2”	5324	70.99%	160	3.01%	
Σύνολο	**7500**	**100.00%**	**261**	**3.48%**	

Πίνακας 3.4 Two - way πίνακας για την εξάρτηση της αθέτησης (Y) από την φοιτητική ιδιότητα (student)

Η p-value του παραπάνω ελέγχου υποδεικνύει ότι η φοιτητική ιδιότητα έχει υψηλή συσχέτιση με την αθέτηση της υποχρέωσης αποπληρωμής πιστωτικής κάρτας, οπότε είναι πιθανό να χρησιμοποιηθεί στο τελικό μοντέλο.

			Αθέτηση		
balance	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-value X^2 ελέγχου**
"1"	7287	97.16%	133	1.83%	< 0.001
"2"	213	2.84%	128	60.10%	
Σύνολο	**7500**	**100.00%**	**261**	**3.48%**	

Πίνακας 3.5 Two - way πίνακας για την εξάρτηση της αθέτησης (Y) από το μηνιαίο υπόλοιπο πιστωτικής κάρτας (balance)

Η p-value του παραπάνω ελέγχου υποδεικνύει ότι το μηνιαίο υπόλοιπο πιστωτικής κάρτας έχει πάρα πολύ υψηλή συσχέτιση με την αθέτηση της υποχρέωσης αποπληρωμής πιστωτικής κάρτας, οπότε είναι πάρα πολύ πιθανό να χρησιμοποιηθεί στο τελικό μοντέλο.

			Αθέτηση		
income	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**	**Πλήθος αθετήσεων**	**Ποσοστό αθετήσεων**	**p-value X^2 ελέγχου**
"1"	1612	21.49%	72	4.47%	0.05037
"2"	3245	43.27%	103	3.17%	
"3"	2643	35.24%	86	3.25%	
Σύνολο	**7500**	**100.00%**	**261**	**3.48%**	

Πίνακας 3.6 Two way πίνακας για την εξάρτηση της αθέτησης (Y) από το ετήσιο εισόδημα (income)

Η p-value του παραπάνω ελέγχου υποδεικνύει ότι το ετήσιο εισόδημα μπορεί, οριακώς, να θεωρηθεί πως δεν έχει υψηλή συσχέτιση με την αθέτηση της υποχρέωσης αποπληρωμής πιστωτικής κάρτας, οπότε είναι πιθανό να μην χρησιμοποιηθεί στο τελικό μοντέλο.

3.2.1 Η προσαρμογή του μοντέλου

Σε αυτό το σημείο, έχουμε (ήδη) μετατρέψει όλες τις μεταβλητές μας σε κατηγορικές. Προκειμένου να πραγματοποιηθεί η προσαρμογή του μοντέλου, και δεδομένου ότι όλες οι συμμεταβλητές κατέστησαν κατηγορικές όπως είπαμε, κατασκευάστηκαν αυτομάτως οι εξής ψευδομεταβλητές από την R:

- $student_2$ = 1εάν student=2 και, = 0 αλλιώς.
- $balance_2$ = 1 εάν balance=2 και, = 0 αλλιώς.
- $Income_2$ = 1 εάν income=2 και, = 0 αλλιώς.
- $Income_3$ = 1 εάν income=3 και, = 0 αλλιώς.

Αρχικώς, λοιπόν, κατασκευάστηκε ένα μοντέλο λογιστικής παλινδρόμησης που περιέχει όλες αυτές τις μεταβλητές. Τα αποτελέσματα της εν λόγω προσαρμογής παρουσιάζονται στον Πίνακα 3.7.

	Συντελεστής μοντέλου	**Τυπικό σφάλμα**	**p-value ελέγχου Wald**
Σταθερός όρος	-4.0730	0.17	<0.001
$student_2$	0.0795	0.25	0.75
$balance_2$	4.4422	0.17	<0.001
$income_2$	-0.0759	0.25	0.76
$income_3$	0.1595	0.30	0.60

Πίνακας 3.7 Αποτελέσματα προσαρμογής του μοντέλου που περιέχει όλες τις μεταβλητές

Σύμφωνα με τις p-values των ελέγχων Wald, συμπεραίνουμε ότι η μεταβλητή balance είναι στατιστικώς πολύ σημαντική (πράγμα που συμφωνεί με την p-value του αντίστοιχου X^2 ελέγχου που είδαμε παραπάνω), ενώ οι μεταβλητές student και income είναι στατιστικώς μη σημαντικές αφού καμμία κατηγορία τους δεν φαίνεται να χρειάζεται στο μοντέλο (όσον αφορά την μεταβλητή income, είχαμε λάβει μία πρώτη εικόνα από την p-value του αντίστοιχου X^2 ελέγχου που είδαμε παραπάνω, η οποία την εμφάνιζε ως οριακώς στατιστικώς μη σημαντική).

Έπειτα, όπως γνωρίζουμε, ούτε η συνάρτηση deviance ούτε η συνάρτηση Pearson μπορεί να χρησιμοποιηθεί για την σύγκριση του μοντέλου που προσαρμόστηκε με το κορεσμένο μοντέλο, επειδή τα δεδομένα μας είναι δυαδικά (n_i=1, για κάθε i).

Αντιθέτως, η ελεγχοσυνάρτηση Hosmer - Lemeshow είναι χρήσιμη στην περίπτωση των δυαδικών δεδομένων. Για το μοντέλο με όλες τις μεταβλητές, λαμβάνουμε X^2_{HL}= 1.2034 με p - value = 0.5479 και 2 βαθμούς ελευθερίας. Η p - value αυτή δείχνει ότι η προσαρμογή του συγκεκριμένου μοντέλου στα δεδομένα είναι ικανοποιητική.

Μάλιστα, υπολογίσαμε και τις προβλέψεις των παρατηρήσεων για τον έλεγχο Hosmer - Lemeshow για το μοντέλο που περιέχει όλες τις μεταβλητές. Τα δεδομένα χωρίστηκαν σε g = 4 ομάδες, αφού

προέκυψαν 2 βαθμοί ελευθερίας και το στατιστικό αυτό ακολουθεί την X^2_{g-2} κατανομή. Τα συγκεκριμένα αποτελέσματα παρουσιάζονται στον Πίνακα 3.8.

	Αθέτηση (Y=1)		**Αποπληρωμή (Y=0)**	
Ομάδα	**Observed**	**Expected**	**Observed**	**Expected**
1	39	33.71	2022	2027.29
2	41	41.95	2456	2455.05
3	53	57.34	2676	2671.66
4	128	128.00	85	85.00

Πίνακας 3.8 Προβλέψεις των παρατηρήσεων για τον έλεγχο Hosmer - Lemeshow για το μοντέλο που περιέχει όλες τις μεταβλητές

Έπειτα, από την διαφορά των συναρτήσεων deviance που αφορούν δύο μοντέλα, μπορούμε να ελέγξουμε ποιό από τα δύο είναι πιο κατάλληλο. Επομένως, θα χρησιμοποιήσουμε αυτή την μέθοδο, συγκρίνοντας το μοντέλο που περιέχει όλες τις μεταβλητές με ένα καινούργιο από το οποίο θα απουσιάζει μία διαφορετική συμμεταβλητή κάθε φορά.

Αφαιρεθείσα μεταβλητή	**Μεταβολή της deviance**	**Βαθμοί ελευθερίας**	**p - value**	**AIC**
student	0.10	1	0.75	1620.6
balance	640.49	1	0.00	2261.0
income	1.70	2	0.43	1620.2

Πίνακας 3.9 Μεταβολή της deviance για το μοντέλο που περιέχει όλες τις μεταβλητές

Από τα αποτελέσματα του Πίνακα 3.9, επαληθεύουμε ότι η μεταβλητή balance είναι στατιστικώς πάρα πολύ σημαντική, αφού η p - value για τον X^2 έλεγχο που της αναλογεί είναι 0 και, επίσης, η αφαίρεσή της συνεπάγεται μεγάλη αύξηση του AIC. Επίσης, επειδή το κριτήριο AIC λαμβάνει την τιμή 1622.5 για το μοντέλο που περιλαμβάνει όλες τις μεταβλητές, παρατηρούμε ότι μειώνεται μόνο όταν αφαιρούμε την μεταβλητή student ή την μεταβλητή income από το μοντέλο.

Σε αυτό το σημείο, έχουμε διαπιστώσει πως το καλύτερο δυνατό μοντέλο θα περιέχει i) την μεταβλητή balance και την μεταβλητή income (fit1) ή ii) την μεταβλητή balance και την μεταβλητή student (fit3) ή iii) μόνο την μεταβλητή balance (fit4).

Θα ελέγξουμε, λοιπόν, κατά πόσο χρειάζεται να αφαιρέσουμε την income από το μοντέλο fit1 και, αντιστοίχως, την student από το fit3.

	Συντελεστής μοντέλου	**Τυπικό σφάλμα**	**p-value ελέγχου Wald**
Σταθερός όρος	-3.9851	0.09	<0.001
$\mathbf{balance_2}$	4.3945	0.17	<0.001

Πίνακας 3.10 Αποτελέσματα προσαρμογής του μοντέλου που περιέχει μόνο την μεταβλητή balance

Αφαιρεθείσα μεταβλητή	**Μεταβολή της deviance**	**Βαθμοί ελευθερίας**	**p - value**	**AIC**
income	2.45	2	0.29	1619

Πίνακας 3.11 Μεταβολής της deviance για το μοντέλο fit1

Αφαιρεθείσα μεταβλητή	**Μεταβολή της deviance**	**Βαθμοί ελευθερίας**	**p - value**	**AIC**
student	0.85	1	0.36	1619

Πίνακας 3.12 Μεταβολή της deviance για το μοντέλο fit3

Από τα αποτελέσματα των Πινάκων 3.11 και 3.12, προκύπτει ότι και η income και η student είναι στατιστικώς μη σημαντικές, αφού η p - value για τον X^2 έλεγχο που τους αναλογεί είναι υψηλή και, επίσης, η αφαίρεση εκάστοτε εξ' αυτών συνεπάγεται μείωση του AIC. Το κριτήριο AIC λαμβάνει την τιμή μικρότερη δυνατή τιμή (1619) για το μοντέλο που περιλαμβάνει μόνο την μεταβλητή balance.

Από την ανωτέρω ανάλυση, καταλήγουμε ότι το καλύτερο δυνατό μοντέλο είναι αυτό που περιέχει μόνο την συμμεταβλητή balance. Άρα, το τελικό μοντέλο είναι το

$$\mathbf{ln}\left(\frac{\hat{\mathbf{p}}}{\mathbf{1}-\hat{\mathbf{p}}}\right) = \mathbf{4.3945balance_2} - \mathbf{3.9851}\,,$$

όπου $\hat{\mathbf{p}}$ είναι η εκτιμώμενη πιθανότητα αθέτησης των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας.

Για να καταλήξουμε στο καταλληλότερο μοντέλο για την περιγραφή των δεδομένων, μπορούμε να πραγματοποιήσουμε και την διαδικασία της διαδοχικής αφαίρεσης (backward elimination). Κατά την διαδικασία αυτή, προσαρμόζεται αρχικώς ένα μοντέλο που περιέχει όλες τις συμμεταβλητές και, σε κάθε βήμα, αφαιρείται η συμμεταβλητή που είναι περισσότερο μη σημαντική στατιστικώς. Το μοντέλο στο οποίο καταλήγει η διαδικασία αυτή είναι εκείνο που περιέχει μόνο τις στατιστικώς

σημαντικές συμμεταβλητές, τις οποίες δεν μπορούμε να αφαιρέσουμε από το μοντέλο. Εν προκειμένω, η διαδικασία καταλήγει, και πάλι, στο μοντέλο που καταλήξαμε παραπάνω.

Από την ανάλυσή μας προέκυψε ότι η αθέτηση των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας από έναν πελάτη εξαρτάται μόνο από το μηνιαίο υπόλοιπο της πιστωτικής κάρτας του.

3.2.2 Η ερμηνεία των συντελεστών

Όπως γνωρίζουμε, ένα από τα σημαντικότερα πλεονεκτήματα του μοντέλου της λογιστικής παλινδρόμησης είναι η εύκολη ερμηνεία των συντελεστών του. Για το μοντέλο στο οποίο καταλήξαμε, ο συντελεστής της μεταβλητής balance_2 ισούται με $4.3945 > 0$. Επομένως, η σχετική πιθανότητα αθέτησης των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας από έναν πελάτη αυξάνεται κατά $e^{4.3945} = 81.00412$, εάν το μηνιαίο υπόλοιπο της πιστωτικής κάρτας του αυξηθεί από την κατηγορία “1” στην κατηγορία “2”.

Το προκύπτον αποτέλεσμα είναι αναμενόμενο, καθώς όσο πιο μεγάλο είναι το μηνιαίο υπόλοιπο της πιστωτικής κάρτας ενός πελάτη τόσο πιο πιθανό είναι να αθετήσει την υποχρέωση αποπληρωμής της.

3.2.3 Τα γραφήματα των υπολοίπων

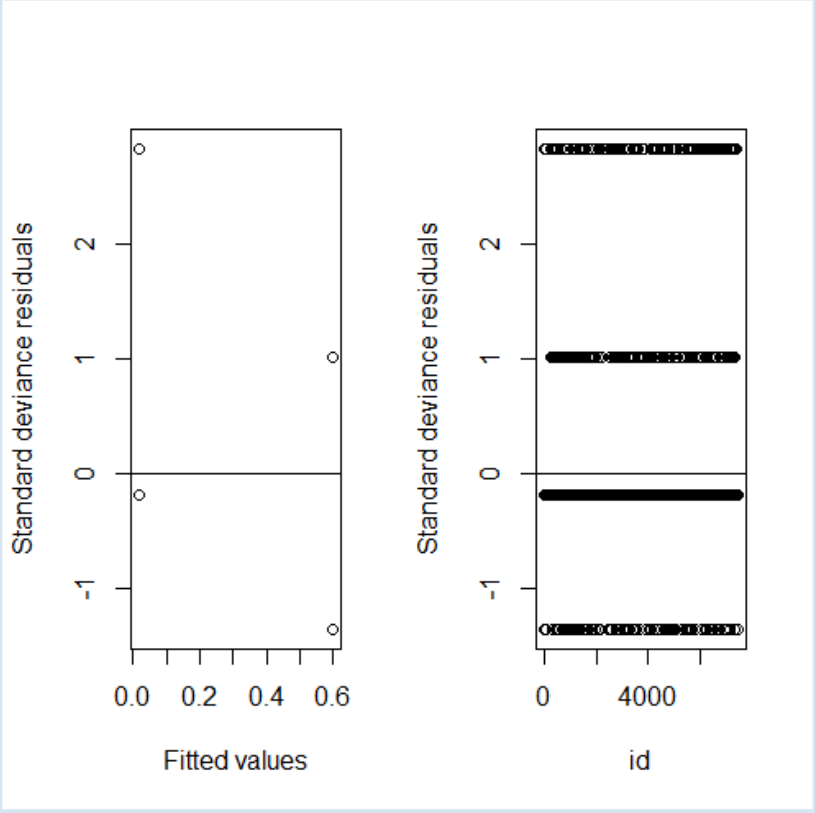
Προκειμένου να κατασκευαστούν τα διαγνωστικά γραφήματα, είναι απαραίτητος ο υπολογισμός των υπολοίπων του τελικού μοντέλου. Χρήσιμα είναι τόσο τα υπόλοιπα deviance, Pearson και πιθανοφάνειας όσο και οι αποστάσεις Cook.

Πιο συγκεκριμένα, όπως φαίνεται στην Εικόνα 3.1, το γράφημα των τυποποιημένων υπολοίπων deviance σε σχέση με τις εκτιμώμενες τιμές χωρίζεται σε δύο τμήματα. Κάτω από την ευθεία y=0 βρίσκεται το τμήμα για τα δεδομένα όπου η εξαρτημένη μεταβλητή λαμβάνει την τιμή “0” και πάνω από την ευθεία βρίσκεται το τμήμα για τα δεδομένα όπου η εξαρτημένη μεταβλητή λαμβάνει την τιμή “1”. 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.

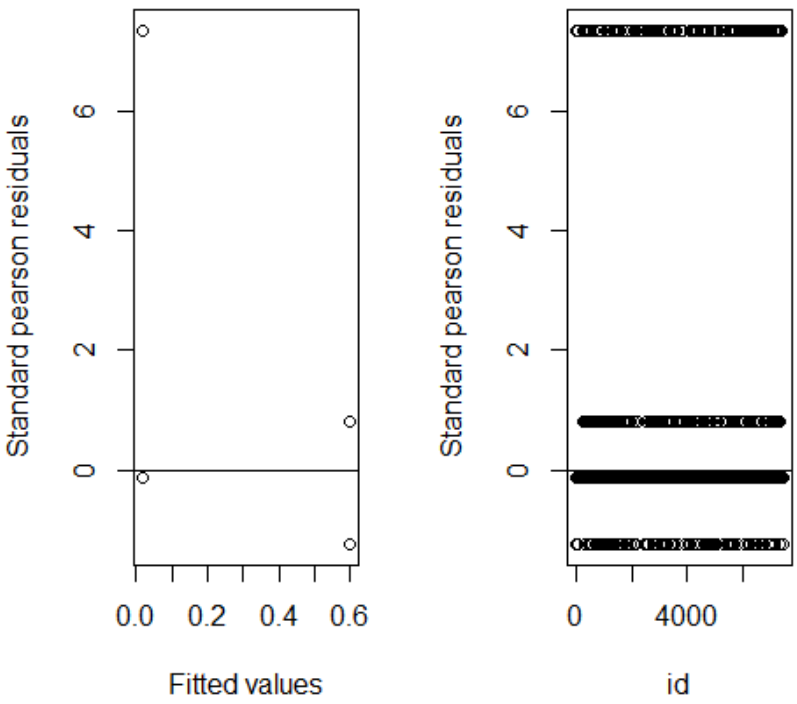
Αντιστοίχως, όπως φαίνεται στην Εικόνα 3.1, το γράφημα των τυποποιημένων υπολοίπων deviance σε σχέση με τη σειρά των δεδομένων χωρίζεται κι αυτό στα ίδια τμήματα. Παρατηρούμε ότι και στα δύο τμήματα η κατανομή των υπολοίπων δεν ακολουθεί κάποιο μοτίβο, πράγμα που υποδεικνύει ότι οι παρατηρήσεις είναι ανεξάρτητες και δεν έχουν σχέση με την σειρά εμφάνισής τους στο δείγμα.

Παρόμοια αποτελέσματα προκύπτουν και με τα γραφήματα των τυποποιημένων υπολοίπων Pearson (Εικόνα 3.2).

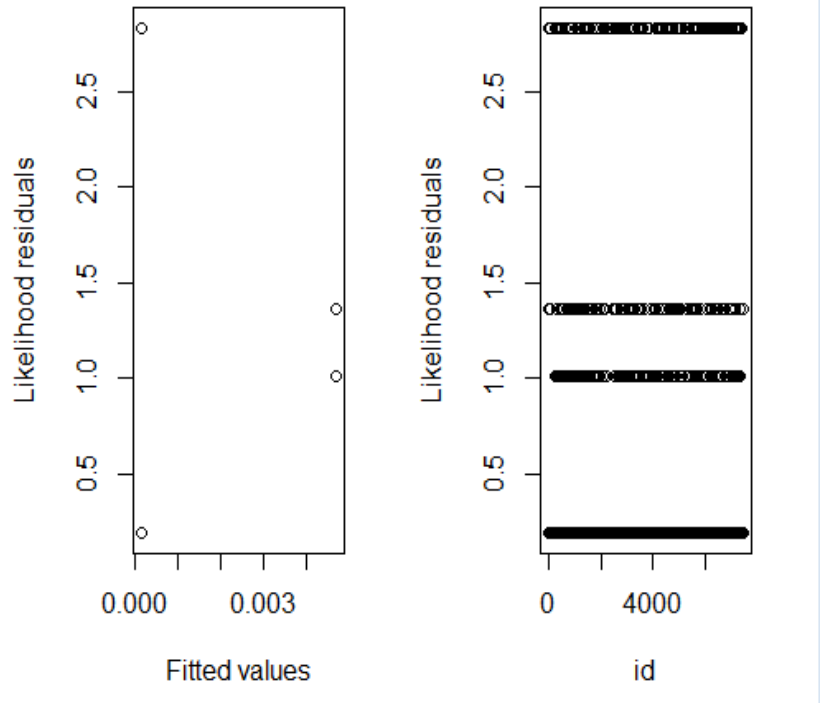
Τέλος, για τον εντοπισμό σημείων επιρροής κατασκευάστηκαν τα γραφήματα των υπολοίπων πιθανοφάνειας (Εικόνα 3.3), καθώς και τα γραφήματα δείκτη των αποστάσεων Cook και των h_{ii} (Εικόνα 3.4). Από τα αυτά τα γραφήματα, παρατηρούμε ότι δεν υπάρχουν σημεία επιρροής στο τελικό μοντέλο.



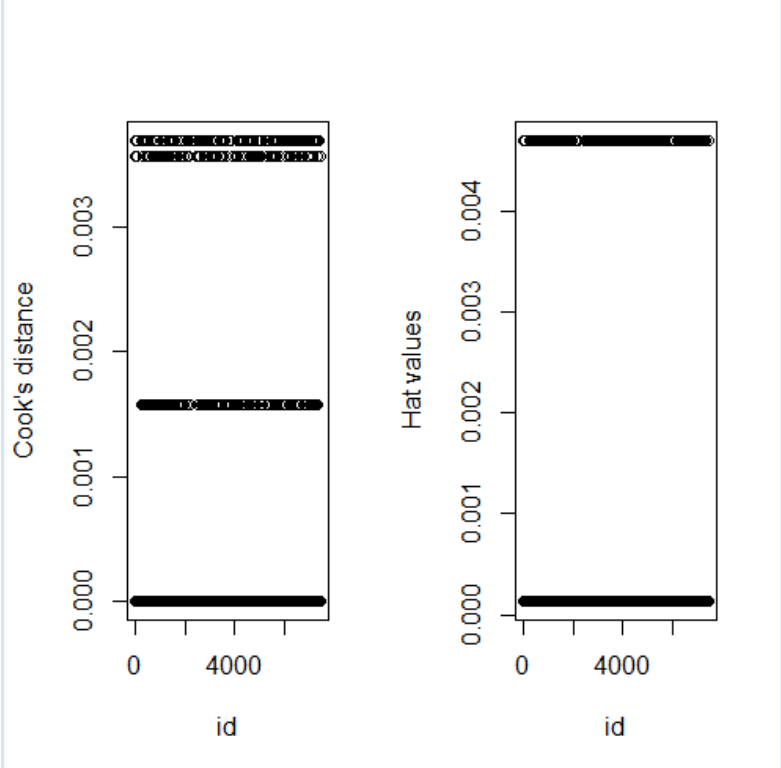
Εικόνα 3.1 Γραφήματα υπολοίπων deviance



Εικόνα 3.2 Γραφήματα υπολοίπων Pearson



Εικόνα 3.3 Γραφήματα υπολοίπων πιθανοφάνειας



Εικόνα 3.4 Γραφήματα δείκτη των αποστάσεων Cook και των hat values

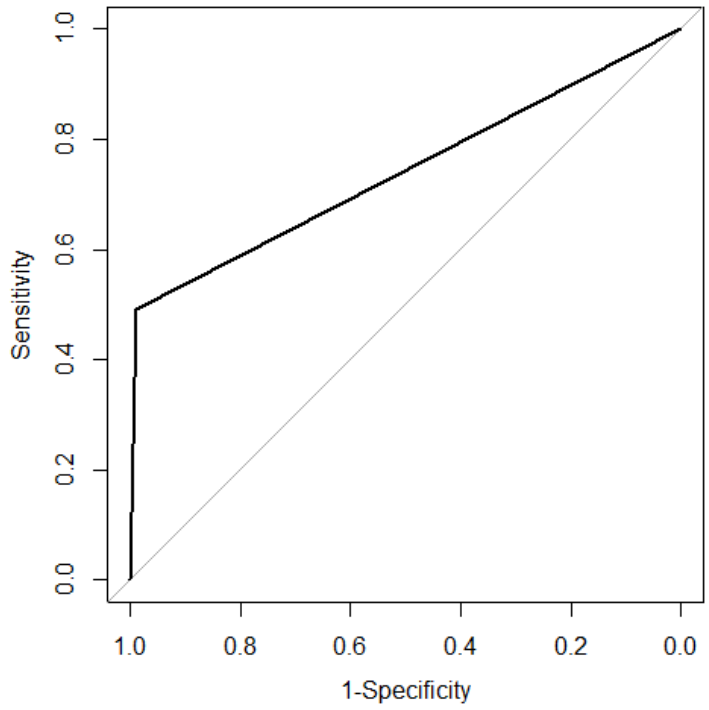
3.2.4 Η προβλεπτική ικανότητα

Το σημαντικότερο χαρακτηριστικό ενός μοντέλου πιστωτικού κινδύνου είναι η προβλεπτική ικανότητά του. Το ιδανικό είναι να είναι πολύ υψηλή έτσι ώστε να είναι χρήσιμο και αξιόπιστο για τον δανειστή που το χρησιμοποιεί, ελαχιστοποιώντας την πιθανότητα λάθους στην κατηγοριοποίηση μίας νέας αίτησης.

Η προβλεπτική ικανότητα του τελικού μοντέλου λογιστικής παλινδρόμησης ελέγχθηκε με την χρήση μίας καμπύλης ROC. Προκειμένου να είναι υψηλή, θα πρέπει το εμβαδόν κάτω από την καμπύλη να είναι όσο το δυνατόν πιο κοντά στην μονάδα και, για μικρές τιμές του 1-Specificity, το Sensitivity να λαμβάνει μεγάλες τιμές.

3.2.4.1 Training set

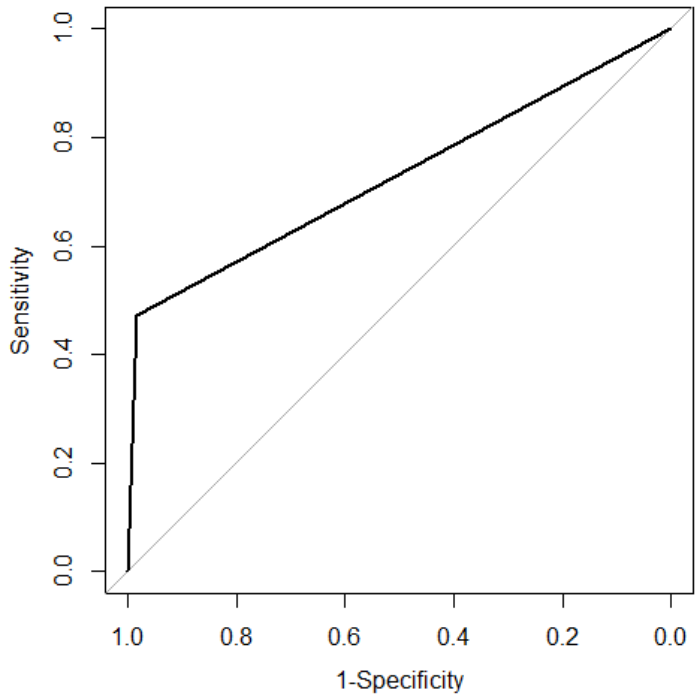
Η Εικόνα 3.5 παρουσιάζει την καμπύλη ROC για το μοντέλο λογιστικής παλινδρόμησης στο οποίο καταλήξαμε. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.7393 και, επομένως, υπερβαίνει αρκετά την ελάχιστη τιμή του που είναι ίση με 0.5. Ως εκ τούτου, η προβλεπτική ικανότητα του μοντέλου στο οποίο καταλήξαμε είναι υψηλή.



Εικόνα 3.5 Καμπύλη ROC τελικού μοντέλου λογιστικής παλινδρόμησης βάσει του training set

3.2.4.2 Test set

Η Εικόνα 3.6 παρουσιάζει την καμπύλη ROC για το μοντέλο στο οποίο καταλήξαμε. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.7277 και, επομένως, υπερβαίνει αρκετά την ελάχιστη τιμή του που είναι ίση με 0.5. Ως εκ τούτου, επαληθεύεται ότι η προβλεπτική ικανότητα του μοντέλου στο οποίο καταλήξαμε είναι υψηλή.



Εικόνα 3.6 Καμπύλη ROC τελικού μοντέλου λογιστικής παλινδρόμησης βάσει του test set

3.3 Ο πιστωτικός κίνδυνος με λογιστική παλινδρόμηση - δεύτερη προσέγγιση

Όπως αναφέραμε, σε αυτήν την προσέγγιση οι συμμεταβλητές balance και income θεωρήθηκαν ποσοτικές.

3.3.1 Η προσαρμογή του μοντέλου

Αρχικώς, κατασκευάστηκε ένα μοντέλο λογιστικής παλινδρόμησης που περιέχει όλες τις συμμεταβλητές. Κατασκευάστηκε από την R, επίσης, η ψευδομεταβλητή $student_2$ η οποία λαμβάνει την τιμή 1 εάν student = 2 και την τιμή 0 αλλιώς. Τα αποτελέσματα της εν λόγω προσαρμογής παρουσιάζονται στον Πίνακα 3.13.

	Συντελεστής μοντέλου	Τυπικό σφάλμα	p-value ελέγχου Wald
Σταθερός όρος	-1.199e+01	5.197e-01	<0.001
student$_2$	4.106e-01	2.758e-01	0.137
balance	5.973e-03	2.760e-04	<0.001
income	1.181e-05	9.397e-06	0.209

Πίνακας 3.13 Αποτελέσματα προσαρμογής του μοντέλου που περιέχει όλες τις μεταβλητές

Σύμφωνα με τις p-values των ελέγχων Wald, συμπεραίνουμε ότι η συμμεταβλητή balance είναι στατιστικώς πολύ σημαντική, ενώ οι συμμεταβλητές student και income φαίνονται στατιστικώς μη σημαντικές.

Έπειτα, θα χρησιμοποιήσουμε την μέθοδο της διαφοράς των συναρτήσεων deviance που αφορούν δύο μοντέλα, ώστε να συγκρίνουμε το μοντέλο που περιέχει όλες τις μεταβλητές με ένα καινούργιο από το οποίο θα απουσιάζει μία διαφορετική συμμεταβλητή κάθε φορά.

Αφαιρεθείσα μεταβλητή	Μεταβολή της deviance	Βαθμοί ελευθερίας	p - value	AIC
student	2.20	1	0.138	1179.4
balance	1077.66	1	0.000	2254.9
income	1.58	1	0.208	1178.8

Πίνακας 3.14 Μεταβολή της deviance για το μοντέλο που περιέχει όλες τις μεταβλητές

Από τα αποτελέσματα του Πίνακα 3.14, επαληθεύουμε ότι η μεταβλητή balance είναι στατιστικώς πάρα πολύ σημαντική, αφού η p - value για τον X^2 έλεγχο που της αναλογεί είναι 0 και, επίσης, η αφαίρεσή της συνεπάγεται μεγάλη αύξηση του AIC. Επίσης, επειδή το κριτήριο AIC λαμβάνει την τιμή 1179.2 για το μοντέλο που περιλαμβάνει όλες τις μεταβλητές, παρατηρούμε ότι μειώνεται μόνο όταν αφαιρούμε την μεταβλητή income από το μοντέλο.

Σε αυτό το σημείο, έχουμε διαπιστώσει πως το καλύτερο δυνατό μοντέλο θα περιέχει την μεταβλητή balance και την μεταβλητή student (fit3) ή μόνο την μεταβλητή balance (fit4).

Θα ελέγξουμε, λοιπόν, κατά πόσο χρειάζεται να αφαιρέσουμε την student από το fit3.

	Συντελεστής μοντέλου	Τυπικό σφάλμα	p-value ελέγχου Wald
Σταθερός όρος	-1.179e+01	4.928e-01	<0.001
student$_2$	6.814e-01	1.700e-01	<0.001
balance	5.982e-03	2.759e-04	<0.001

Πίνακας 3.15 Αποτελέσματα προσαρμογής του μοντέλου fit3

Αφαιρεθείσα μεταβλητή	Μεταβολή της deviance	Βαθμοί ελευθερίας	p - value	AIC
student	16.91	1	3.924302e-05	1193.7

Πίνακας 3.16 Μεταβολή της deviance για το μοντέλο fit3

Από τα αποτελέσματα των Πινάκων 3.15 και 3.16, προκύπτει ότι η student και η balance είναι στατιστικώς σημαντικές, αφού η p - value για τον X^2 έλεγχο που τους αναλογεί είναι πάρα πολύ υψηλή. Επίσης, το κριτήριο AIC λαμβάνει την τιμή μικρότερη δυνατή τιμή (1178.8) για το μοντέλο που περιλαμβάνει την student και την balance ενώ αφαίρεση της student από το εν λόγω μοντέλο συνεπάγεται αύξησή του (1193.7).

Από την ανωτέρω ανάλυση, καταλήγουμε ότι το καλύτερο δυνατό μοντέλο είναι αυτό που περιέχει μόνο τις συμμεταβλητές student και balance. Άρα, το τελικό μοντέλο είναι το

$$\ln\left(\frac{\hat{\mathbf{p}}}{\mathbf{1}-\hat{\mathbf{p}}}\right) = \mathbf{0.6814student_2} + \mathbf{0.005982balance} - \mathbf{11.79}$$

όπου $\hat{\mathbf{p}}$ είναι η εκτιμώμενη πιθανότητα αθέτησης των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας.

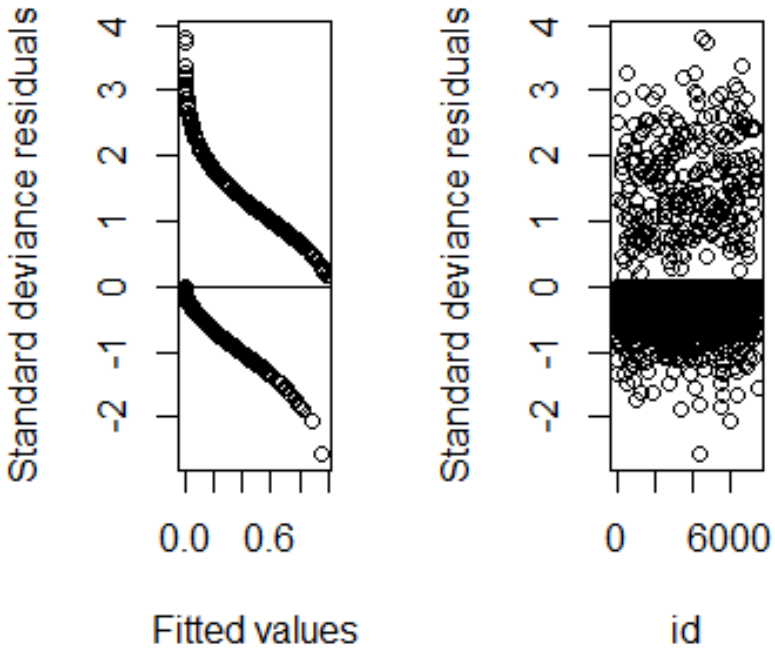
Και πάλι, η διαδικασία backward elimination καταλήγει στο ίδιο μοντέλο.

3.3.2 Η ερμηνεία των συντελεστών

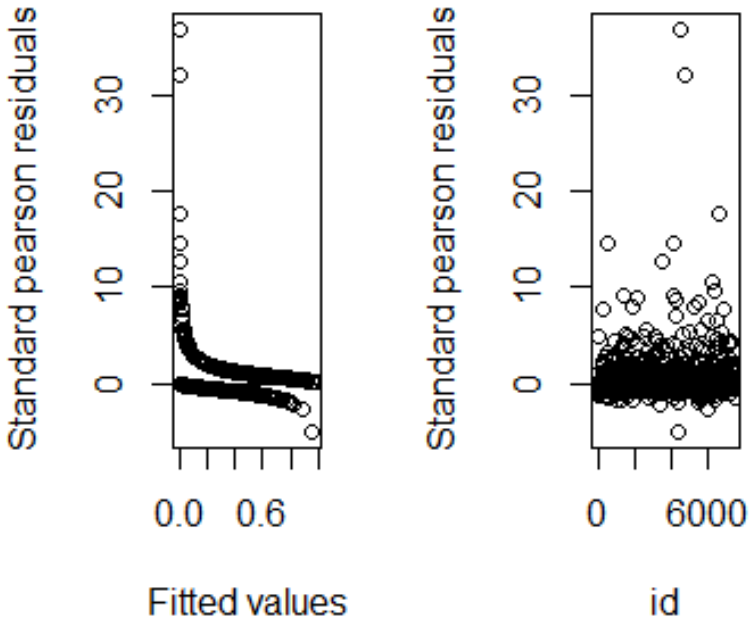
Για το μοντέλο στο οποίο καταλήξαμε, ο συντελεστής της μεταβλητής student$_2$ ισούται με 0.6814 > 0. Επομένως, η πιθανότητα αθέτησης των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας από έναν πελάτη αυξάνεται κατά $e^{0.6814}$ =1.976643, εάν η φοιτητική ιδιότητα μεταβληθεί από την κατηγορία “1” στην κατηγορία “2”. Επίσης, ο συντελεστής της balance ισούται με 0.005982 > 0 και, επομένως, αύξηση του μηνιαίου υπολοίπου της πιστωτικής κάρτας συνεπάγεται αύξηση της πιθανότητας αθέτησης των υποχρεώσεων αποπληρωμής της πιστωτικής κάρτας κατά $e^{0.005982}$ =1.005998.

3.3.3 Τα γραφήματα των υπολοίπων

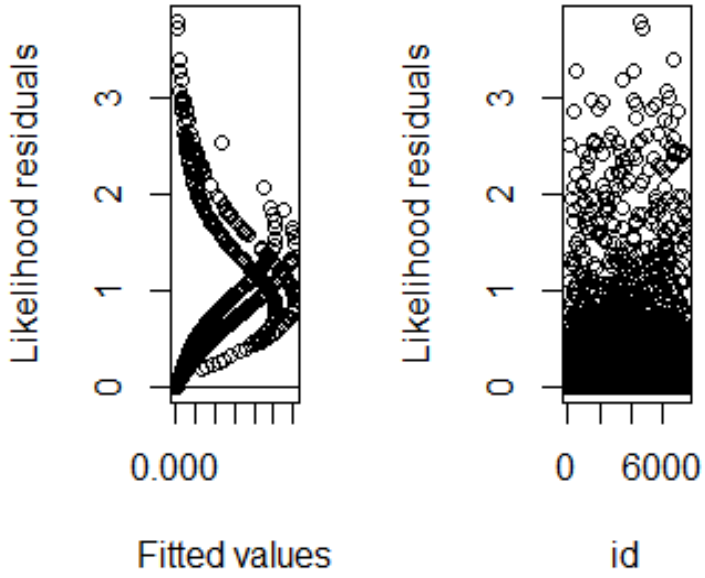
Κατασκευάστηκαν τα αντίστοιχα γραφήματα που είδαμε κατά την πρώτη προσέγγιση. Όπως φαίνεται στις Εικόνες 3.7, 3.8, 3.9 και 3.10, και πάλι, δεν υπάρχουν παρατηρήσεις που να αποκλίνουν από τις υπόλοιπες, οι παρατηρήσεις είναι ανεξάρτητες και δεν έχουν σχέση με την σειρά εμφάνισής τους στο δείγμα και δεν υπάρχουν σημεία επιρροής στο τελικό μοντέλο.



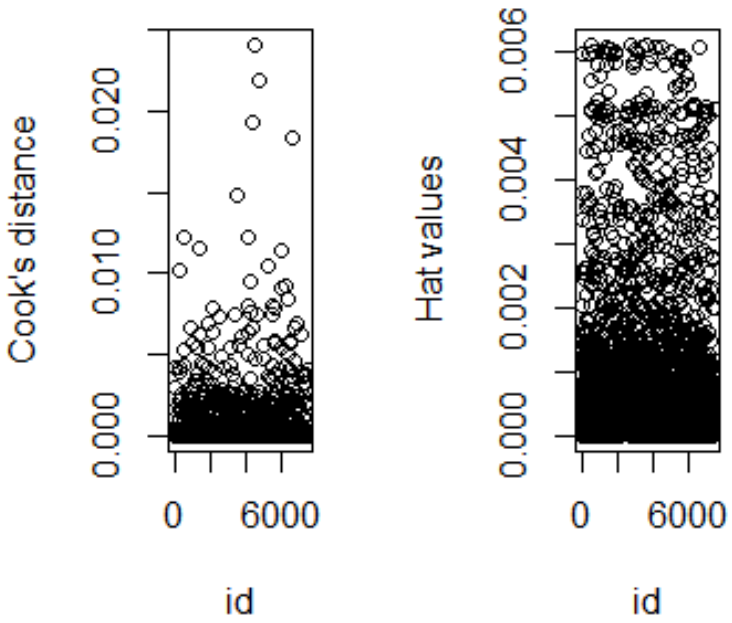
Εικόνα 3.7 Γραφήματα υπολοίπων deviance



Εικόνα 3.8 Γραφήματα υπολοίπων Pearson



Εικόνα 3.9 Γραφήματα υπολοίπων πιθανοφάνειας



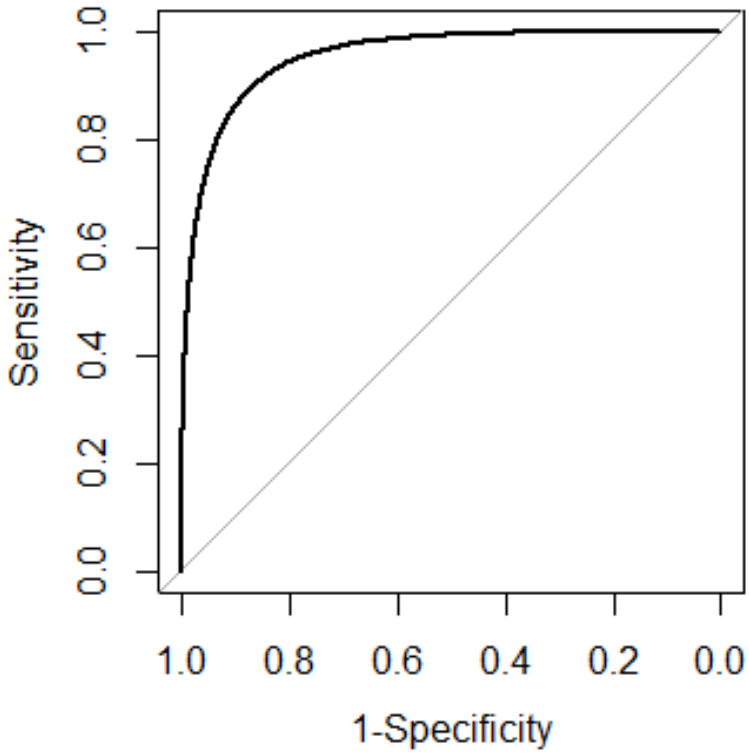
Εικόνα 3.10 Γραφήματα δείκτη των αποστάσεων Cook και των hat values

3.3.4 Η προβλεπτική ικανότητα

Και σε αυτήν την προσέγγιση, η προβλεπτική ικανότητα του τελικού μοντέλου λογιστικής παλινδρόμησης ελέγχθηκε με την χρήση μίας καμπύλης ROC.

3.3.4.1 Training set

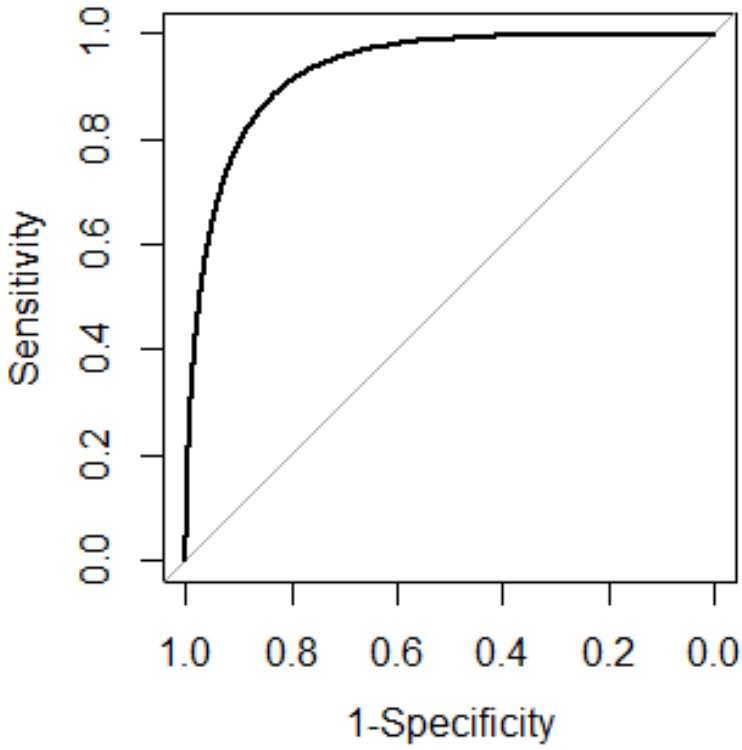
Η Εικόνα 3.11 παρουσιάζει την καμπύλη ROC, όσον αφορά στο training set, για το μοντέλο λογιστικής παλινδρόμησης στο οποίο καταλήξαμε. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.9538 και, επομένως, προσεγγίζει πάρα πολύ την μέγιστη τιμή που ισούται με την μονάδα. Ως εκ τούτου, η προβλεπτική ικανότητα του μοντέλου στο οποίο καταλήξαμε είναι πάρα πολύ υψηλή.



Εικόνα 3.11 Καμπύλη ROC τελικού μοντέλου λογιστικής παλινδρόμησης βάσει του training set

3.3.4.2 Test set

Η Εικόνα 3.12 παρουσιάζει την καμπύλη ROC, όσον αφορά στο test set, για το μοντέλο στο οποίο καταλήξαμε. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.934 και, επομένως προσεγγίζει πάρα πολύ την μέγιστη τιμή που ισούται με την μονάδα. Ως εκ τούτου, επαληθεύεται ότι η προβλεπτική ικανότητα του μοντέλου στο οποίο καταλήξαμε είναι πάρα πολύ υψηλή.



Εικόνα 3.12 Καμπύλη ROC τελικού μοντέλου λογιστικής παλινδρόμησης βάσει του test set

3.4 Ο πιστωτικός κίνδυνος με δένδρα απόφασης

Όπως και στην περίπτωση της κατασκευής μοντέλου λογιστικής παλινδρόμησης, έτσι και στην περίπτωση της κατασκευής μοντέλου δένδρου απόφασης, η εξαρτημένη μεταβλητή θα είναι η Y, η οποία είναι δίτιμη και εκφράζει εάν υπήρξε αθέτηση της υποχρέωσης αποπληρωμής της πιστωτικής κάρτας - οπότε λαμβάνει την τιμή 1 - ή όχι - οπότε λαμβάνει την τιμή 0.

3.4.1 Η κατασκευή και η ερμηνεία του μοντέλου δένδρου απόφασης

Κατ' αρχάς, ελέγξαμε πόσες αθετήσεις και πόσες μη αθετήσεις παρατηρούνται στο training set - διαπιστώνουμε πως παρατηρούνται 7239 μη αθετήσεις και 261 αθετήσεις.

Έπειτα, κατασκευάσαμε ένα μοντέλο δένδρου ταξινόμησης, αφού η εξαρτημένη μεταβλητή είναι κατηγορική. Χρησιμοποιήθηκε το κριτήριο Gini, το οποίο και υποστηρίζει η R.

1ος τρόπος: με προεπιλογή από την R

Με αυτόν τον τρόπο, κατασκευάσαμε ένα δένδρο ταξινόμησης, αφήνοντας την R να επιλέξει τις επιμέρους συνθήκες και παραμέτρους κατασκευής του.

Το αποτέλεσμα παρουσιάζεται στην Εικόνα 3.13. Προκύπτει ότι η πιο σημαντική συμμεταβλητή είναι η balance. Επίσης, όπως φαίνεται στην Εικόνα 3.13, επιλέχθηκε από την R η μη αθέτηση ως ρίζα και, επίσης, με αστερίσκο σηματοδοτήθηκαν τα φύλλα.

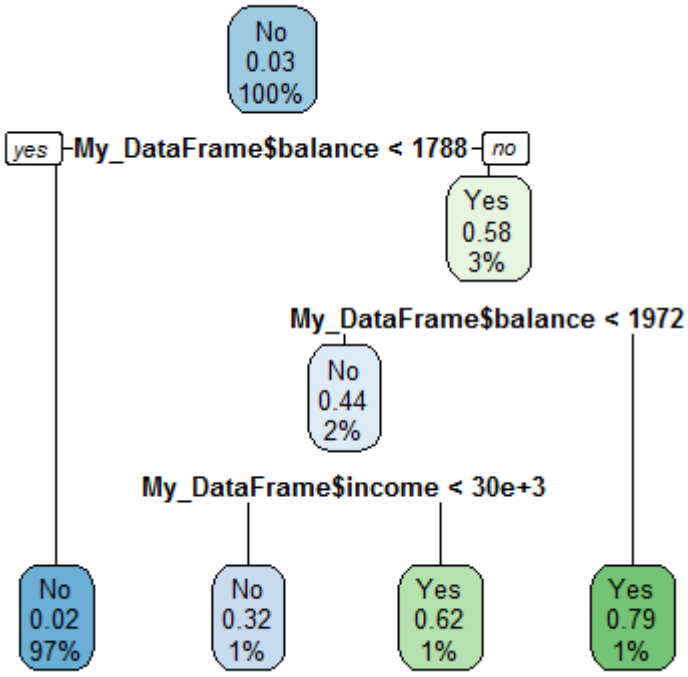
```
1) root 7500 261  No  (0.96520000 0.03480000)
  2) My_DataFrame$balance< 1788.349 7272 128  No  (0.98239824 0.01760176) *
  3) My_DataFrame$balance>=1788.349 228  95  Yes  (0.41666667 0.58333333)
    6) My_DataFrame$balance< 1971.915 134  59  No  (0.55970149 0.44029851)
     12) My_DataFrame$income< 30111.74 81  26  No  (0.67901235 0.32098765) *
     13) My_DataFrame$income>=30111.74 53  20  Yes  (0.37735849 0.62264151) *
    7) My_DataFrame$balance>=1971.915 94  20  Yes  (0.21276596 0.78723404) *
```

Εικόνα 3.13 Δένδρο ταξινόμησης με προεπιλογή από την R

Πιο συγκεκριμένα, σύμφωνα με το μοντέλο, προκύπτουν τα εξής:

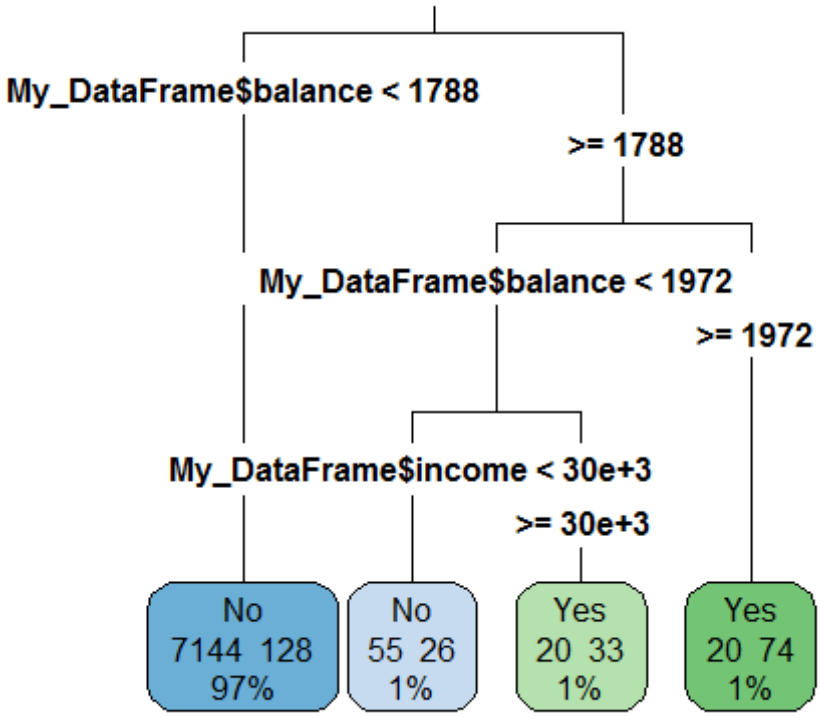
- εάν το μηνιαίο υπόλοιπο της πιστωτικής κάρτας είναι μικρότερο από 1788.349, τότε δεν προκύπτει αθέτηση **(1)**
- εάν το μηνιαίο υπόλοιπο της πιστωτικής κάρτας είναι μεγαλύτερο από / ίσο με 1788.349, τότε, εάν είναι και μεγαλύτερο από / ίσο με 1971.915, προκύπτει αθέτηση **(2)**. Αλλιώς (εάν, δηλαδή, είναι μεγαλύτερο από / ίσο με 1788.349 και μικρότερο από 1971.915), εάν το ετήσιο εισόδημα είναι μικρότερο από 30111.74 τότε δεν προκύπτει αθέτηση **(3)** ενώ εάν αυτό είναι μεγαλύτερο από / ίσο με 30111.74 τότε προκύπτει αθέτηση **(4)**

Η Εικόνα 3.14 παρουσιάζει την γραφική αναπαράσταση του προκύπτοντος δένδρου. Όπως φαίνεται και στην Εικόνα 3.14 , επιλέχθηκε ως ρίζα η μη αθέτηση (περιλαμβάνει το 100% του δείγματος) και, έπειτα, κάθε κόμβος αναγράφει το ποσοστό του δείγματος το οποίο περιέχει και, επίσης, οι αριστερές ακμές αντιπροσωπεύουν το «ναι» στον εκάστοτε αντίστοιχο έλεγχο και οι δεξιές το «όχι». Προκύπτει, τελικώς, ότι το 97% του δείγματος εμπίπτει στην περίπτωση **(1)**, το 1% στην περίπτωση **(2)**, το 1% στην περίπτωση **(3)** και το 1% στην περίπτωση **(4)**.



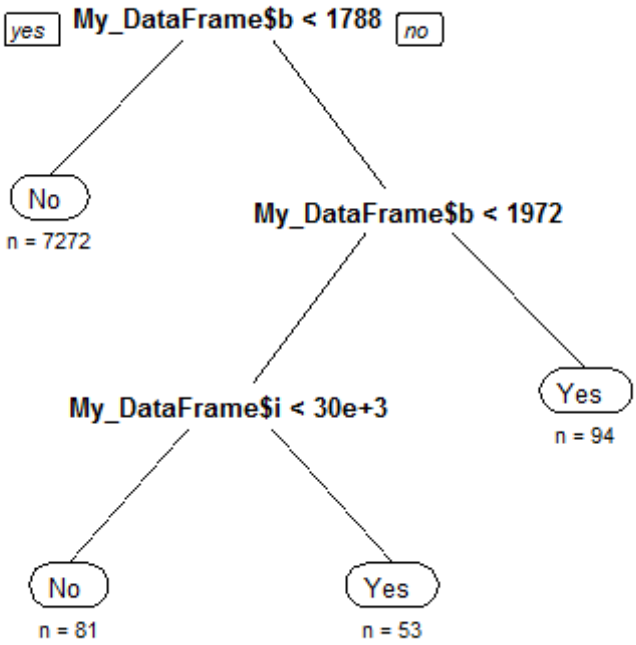
Εικόνα 3.14 Πρώτη σχηματική απεικόνιση δένδρου ταξινόμησης με προεπιλογή από την R

Το οπτικό αποτέλεσμα για το προκύπτον δένδρο βελτιώνεται στην Εικόνα 3.15, όπου όλα τα φύλλα είναι στο κατώτερο επίπεδο και εξαιρέθηκε τόσο η ρίζα όσο και οι εσωτερικοί κόμβοι. Τα φύλλα, μάλιστα, περιέχουν και το αντίστοιχο πλήθος των παρατηρήσεων: 7144+128 παρατηρήσεις εμπίπτουν στην περίπτωση **(1)**, 20+74 στην περίπτωση **(2),** 55+26 στην περίπτωση **(3)**, 20+33 στην περίπτωση **(4)**.



Εικόνα 3.15 Δεύτερη σχηματική απεικόνιση δένδρου ταξινόμησης με προεπιλογή από την R

Τέλος, το δένδρο καθίσταται πλέον αντιληπτό στην Εικόνα 3.16. Σηματοδοτείται, και πάλι, ότι οι αριστερές ακμές αφορούν το «ναι» ενώ οι δεξιές το «όχι» στον αντίστοιχο έλεγχο. Τα φύλλα αναγράφουν την κλάση (αθέτηση ή μη αθέτηση) και, κάτω από αυτά, αναγράφεται το αντίστοιχο συνολικό πλήθος των παρατηρήσεων: 7272 στην περίπτωση **(1)**, 94 στην περίπτωση **(2),** 81 στην περίπτωση **(3)**, 53 στην περίπτωση **(4)**.



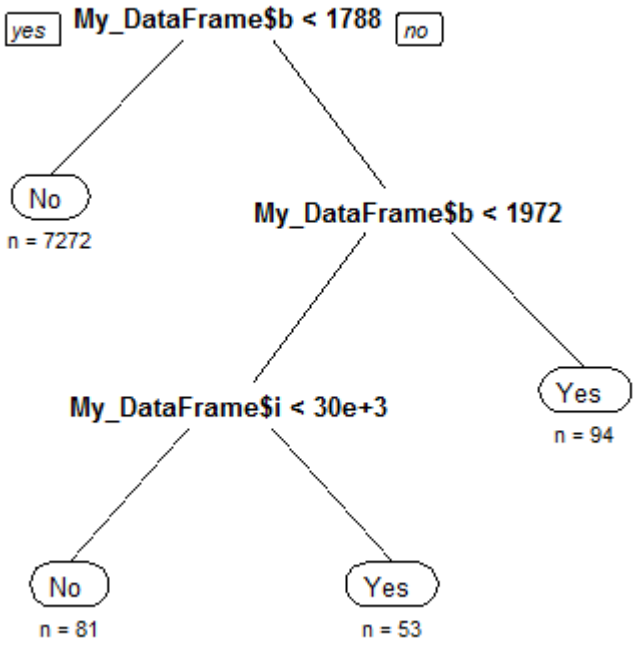
Εικόνα 3.16 Τρίτη σχηματική απεικόνιση δένδρου ταξινόμησης με προεπιλογή από την R

Τελικώς, προκύπτει ότι 7272+81=7353 περιπτώσεις προβλέφθηκαν ως μη αθετήσεις και 94+53=147 περιπτώσεις προβλέφθηκαν ως αθετήσεις.

2ος τρόπος: με ορισμό της καλύτερης δυνατής παραμέτρου πολυπλοκότητας

Με αυτόν τον τρόπο, κατασκευάσαμε ένα δένδρο ταξινόμησης, ορίζοντας την καλύτερη δυνατή παράμετρο πολυπλοκότητας προκειμένου να επιλεχθεί το μέγεθος δένδρου (κατόπιν κλαδέματος) που ελαχιστοποιεί το ποσοστό εσφαλμένης ταξινόμησης (δηλαδή εσφαλμένης πρόβλεψης)

Το αποτέλεσμα καταδεικνύει ότι, και πάλι, μόνο οι συμμεταβλητές balance και income χρησιμοποιήθηκαν στην κατασκευή του δέντρου, το οποίο και παρουσιάζεται στην Εικόνα 3.17.



Εικόνα 3.17 Σχηματική απεικόνιση δένδρου ταξινόμησης με την καλύτερη δυνατή παράμετρο πολυπλοκότητας

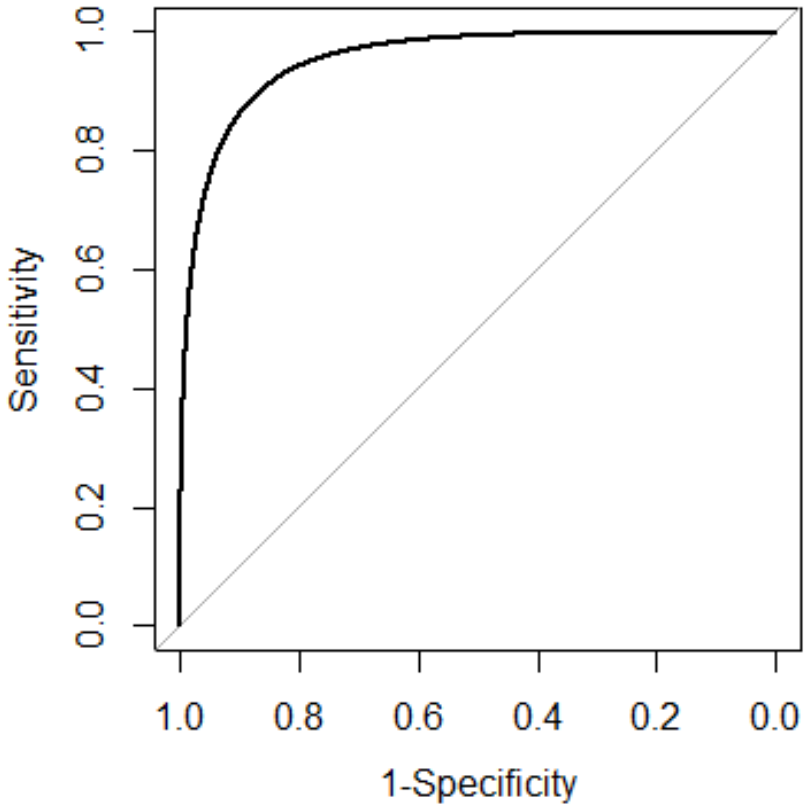
Όπως φαίνεται, προέκυψε το ίδιο επ' ακριβώς μοντέλο με εκείνο του 1^{ου} τρόπου.

3.4.2 Η προβλεπτική ικανότητα

Όπως έχουμε πει και παραπάνω, είναι πολύ σημαντικό το προκύπτον μοντέλο πιστωτικού κινδύνου να διέπεται από υψηλή προβλεπτική ικανότητα. Θα ελέγξουμε, λοιπόν, την προβλεπτική ικανότητα του προκύπτοντος μοντέλου δένδρου ταξινόμησης λαμβάνοντας υπ' όψιν τόσο το training set όσο και το test set. Ο εν λόγω έλεγχος θα προκύψει, όπως και στην περίπτωση της λογιστικής παλινδρόμησης, με κατασκευή καμπυλών ROC.

3.4.2.1 Training set

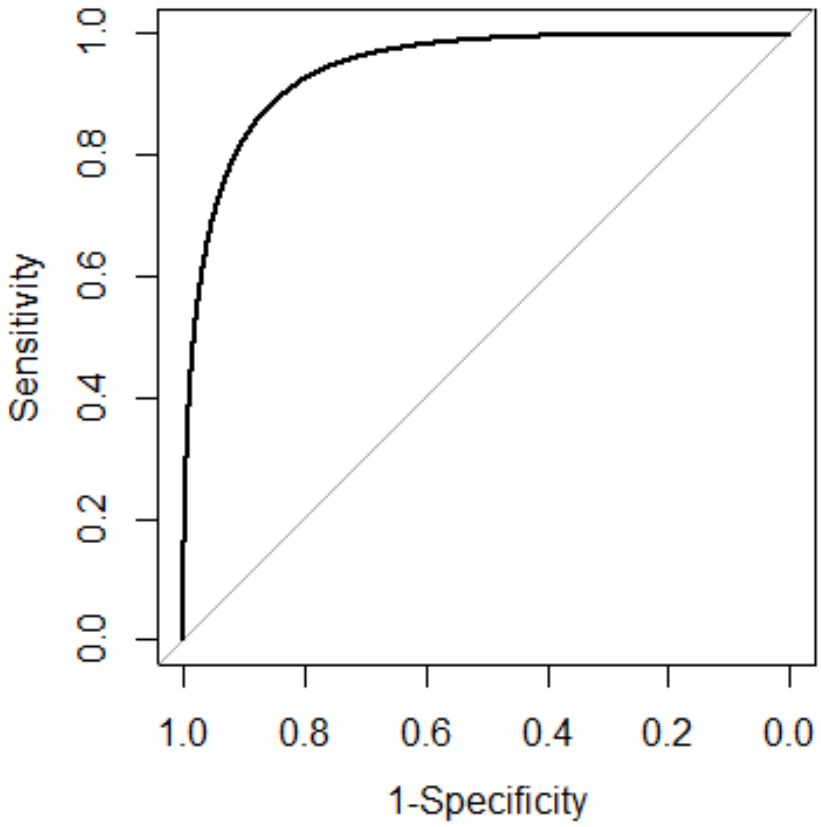
Η Εικόνα 3.18 παρουσιάζει την καμπύλη ROC για το μοντέλο δένδρου ταξινόμησης, χρησιμοποιώντας το training set. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.9538 και, επομένως, η προβλεπτική ικανότητα του εν λόγω μοντέλου είναι πάρα πολύ υψηλή.



Εικόνα 3.18 Καμπύλη ROC τελικού μοντέλου δένδρου ταξινόμησης βάσει του training set

3.4.2.2 Test set

Η Εικόνα 3.19 παρουσιάζει την καμπύλη ROC για το μοντέλο δένδρου ταξινόμησης, χρησιμοποιώντας το test set. Το εμβαδόν κάτω από την καμπύλη, AUC, ισούται με 0.9433 και, επομένως, επαληθεύεται ότι η προβλεπτική ικανότητα του εν λόγω μοντέλου είναι πάρα πολύ υψηλή.



Εικόνα 3.19 Καμπύλη ROC τελικού μοντέλου δένδρου ταξινόμησης βάσει του test set

3.5 Συμπεράσματα

Αποτελεί διαπίστωση ότι η αποπληρωμή ενός δανείου είναι μία δυναμική και συνεχώς εξελισσόμενη διαδικασία και δεν μπορεί να κατασκευαστεί ένα μοντέλο που να παρέχει τέλεια κατηγοριοποίηση των υποψήφιων πελατών (Hand, 2001). Γι' αυτό και πραγματοποιείται προσπάθεια τα μοντέλα που κατασκευάζονται να ελαχιστοποιούν την πιθανότητα λάθος κατηγοριοποίησης ενός αιτούντα χρησιμοποιώντας τα στοιχεία της αίτησής του.

Τα δύο είδη μοντέλων με τα οποία ασχοληθήκαμε θέτουν λίγους περιορισμούς για την προσαρμογή τους, αλλά στερούνται και κάποιες δυνατότητες ευελιξίας. Η μελέτη χρηματοπιστωτικών δεδομένων με μοντέλο λογιστικής παλινδρόμησης, για παράδειγμα, κατατάσσει έναν πελάτη ως «κακό ρίσκο» μόνο εάν αθετήσει τις υποχρεώσεις του πριν από μία προκαθορισμένη χρονική στιγμή (Banasik et al., 2001).

Ως προς τα δύο μοντέλα λογιστικής παλινδρόμησης που αναπτύξαμε, η πρώτη προσέγγιση που λαμβάνει ως κατηγορικές τις balance και income (το μοντέλο περιλαμβάνει μόνο την student) δίδει υψηλή προβλεπτική ικανότητα, αλλά η δεύτερη προσέγγιση που τις θεωρεί ως ποσοτικές (το

μοντέλο συμπεριλαμβάνει τις student και balance) βελτιώνει κατά πολύ την προβλεπτική ικανότητα. Θα πρέπει να επισημανθεί πως υπήρξε λογική η κατηγοριοποίηση των δεδομένων της πρώτης προσέγγισης καθώς, πολλές φορές, η λήψη μεταβλητών ως ποσοτικών επιβάλλει γραμμικότητα στο μοντέλο ενώ η κατηγοριοποίησή τους αποκαλύπτει την έκταση της μη γραμμικότητας. Ωστόσο, το κόστος της κατηγοριοποίησης είναι, πολλές φορές, η απώλεια πληροφοριών. Από την άλλη, το μοντέλο των δένδρων ταξινόμησης συμπεριέλαβε τις balance και income (τις οποίες θεώρησε ως ποσοτικές).

Η σύγκριση μοντέλων πραγματοποιείται βάσει της προβλεπτικής ικανότητάς τους. Για το δείγμα που χρησιμοποιήσαμε, το μοντέλο των δένδρων ταξινόμησης φαίνεται να έχει οριακώς υψηλότερη προβλεπτική ικανότητα από το μοντέλο λογιστικής παλινδρόμησης της δεύτερης προσέγγισης αφού $AUC_{LOGISTIC\ REGRESSION} = 0.9538 = AUC_{DECISION\ TREES}$ όσον αφορά στο training set και, $AUC_{LOGISTIC\ REGRESSION} = 0.934 < AUC_{DECISION\ TREES} = 0.9433$ όσον αφορά στο test set.

Συμπεραίνουμε ότι τα δύο μοντέλα μάς δίδουν τα ίδια αποτελέσματα περίπου. Τα αποτελέσματα τα οποία προέκυψαν είναι επαρκώς ασφαλή - καθώς λάβαμε υπ' όψιν λίγες συμμεταβλητές μεν αλλά το πλήθος των παρατηρήσεων (n=10000) ήταν πολύ μεγάλο - και προκύπτει ότι και τα δύο μοντέλα είναι πάρα πολύ αποδοτικά. Επομένως, το εάν κάποιος ερευνητής αποφασίσει να δουλέψει με το ένα ή το άλλο εξαρτάται από τις ειδικές γνώσεις του, τις προτιμήσεις του, καθώς και από το είδος της έρευνας που επιθυμεί να διεξάγει.

Βιβλιογραφία

Ξενόγλωσση

Abdou, H. A. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, **18**, 59-88.

Banasik, J., Crook J.N. & Thomas L.C. (2001). Scoring by usage, *Journal of the Operational Research Society*, **52**, 997 - 1006.

Breiman, L., Friedman, J.H, Olshen, R.A. & Stone C.J. (1984), *Classification and Regression Trees*. Wadsworth, Belmont.

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

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, Boca Raton, Florida.

Collett, D. (2003). *Modelling Binary Data*, (2nd ed.). Chapman and Hall/CRC, Boca Raton, Florida.

Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B*, **34**, 187 - 220.

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing*. National Bureau of Economic Research, New York.

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

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Hand, D.J. (2001). Modelling consumer credit risk, *IMA Journal of Management Mathematics*, **12**, 139 - 155.

Hastie, T., Tibshirani, R. & Friedman J. (2009). *The Elements of Statistical Learning*. (2nd ed.), Springer, New York.

Haykin, S. (1999). *Neural Networks, a comprehensive foundation*, (2nd ed.). McMaster University Hamilton, Ontario, Canada.

Hooman, A., Omidi, M., Marthandan G., Wan Yusoff, W.F. & Karamizadeh, S. (2015). *Statistical and data mining methods in credit scoring*. Proceedings of the Asia Pacific Conference on Business and Social Sciences, Kuala Lumpur (in partnership with The Journal of Developing Areas).

(συνέχεια)

Hosmer, D.W. Lemeshow, S. & Sturdivant, R.X. (2013). *Applied Logistic Regression*, (3rd ed.), Wiley, Hoboken, New Jersey.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination, *Biometrika*, **78**, 691 - 692.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model, *The Journal of Business*, **74**, 101 - 124.

Thomas, L. C, Edelman, D. B. & Crook, J. N. (2002). *Credit Scoring and its Applications*, Society for Industrial and Applied Mathematics, Philadelphia.

Ελληνική

Θωμαδάκης, Σ., Ξανθάκης Ε. (2006). *Αγορές Χρήματος και Κεφαλαίου,* (1η έκδοση). Εκδόσεις Σταμούλη, Αθήνα.

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

Φουσκάκης, Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R*. Εκδόσεις Τσότρας, Αθήνα.

Ηλεκτρονικές Πηγές

Bankrate. *Credit scoring system*.
Διαθέσιμο στο:https://www.bankrate.com/glossary/c/credit-scoring-system/
Ανακτήθηκε στις 17 Ιουνίου 2019.

Business Dictionary. *Creditscoring*.
Διαθέσιμο στο: http://www.businessdictionary.com/definition/credit-scoring.html
Ανακτήθηκε στις 11 Ιουνίου 2019.

Investing Answers. *Credit Bureau*.
Διαθέσιμο στο: https://investinganswers.com/dictionary/c/credit-bureau
Ανακτήθηκε στις 16 Ιουλίου 2019.

Investopedia. *Credit*.
Διαθέσιμο στο: https://www.investopedia.com/terms/c/credit.asp
Ανακτήθηκε στις 14 Αυγούστου 2019.

(συνέχεια)

The balance. *What are the 3 major credit reporting agencies?*

Διαθέσιμο στο: https://www.thebalance.com/who-are-the-three-major-credit-bureaus-960416

Ανακτήθηκε στις 11 Αυγούστου 2019.

Βικιπαίδεια. *Παλινδρόμηση (στατιστική)*.

Διαθέσιμο στο: https://el.wikipedia.org/wiki/Παλινδρόμηση_(στατιστική)

Ανακτήθηκε στις 18 Ιουνίου 2019.

Ευρετήριο οικονομικών όρων. *Πιστοληπτική Ικανότητα / Διαβάθμιση*.

Διαθέσιμο στο: https://www.euretirio.com/pistoliptiki-ikanotita/

Ανακτήθηκε στις 18 Ιουνίου 2019.

Ευρετήριο οικονομικών όρων. *Υπόθεση του κύκλου ζωής (Life cycle hypothesis)*.

Διαθέσιμο στο: https://www.euretirio.com/kyklos-zois/

Ανακτήθηκε στις 27 Ιουλίου 2019.

.

Παράρτημα Α: «Ενδεικτική φόρμα αίτησης πίστωσης ατόμου»

_____Company

Credit Application

APPLICANT INFORMATION		
Name:		
Date of birth:	SSN:	Phone:
Current address:		
City:	State:	ZIP Code:
Own Rent (Please circle)	Monthly payment or rent:	How long?
Previous address:		
City:	State:	ZIP Code:
Owned Rented (Please circle)	Monthly payment or rent:	How long?
EMPLOYMENT INFORMATION		
Current employer:		
Employer address:		How long?
Phone:	E-mail:	Fax:
City:	State:	ZIP Code:
Position:	Hourly Salary (Please circle)	Annual income:
Previous employer:		
Address:		How long?
Phone:	E-mail:	Fax:
City:	State:	ZIP Code:
Position:	Hourly Salary (Please circle)	Annual income:
Name of a relative not residing with you:		
Address:		Phone:
City:	State:	ZIP Code:
Relationship:		
CO-APPLICANT INFORMATION, IF FOR A JOINT ACCOUNT		
Name:		
Date of birth:	SSN:	Phone:
Current address:		
City:	State:	ZIP Code:
Own Rent (Please circle)	Monthly payment or rent:	How long?
Previous address:		
City:	State:	ZIP Code:
Owned Rented (Please circle)	Monthly payment or rent:	How long?
EMPLOYMENT INFORMATION		
Current employer:		
Employer address:		How long?
Phone:	E-mail:	Fax:
City:	State:	ZIP Code:
Position:	Hourly Salary (Please circle)	Annual income:
Previous employer:		
Address:		
Phone:	E-mail:	Fax:
City:	State:	ZIP Code:
Position:	Hourly Salary (Please circle)	Annual income:

_____Company

Credit Application

APPLICATION INFORMATION CONTINUED			
Name of a relative not residing with you:			
Address:			Phone:
City:	State:		ZIP Code:
Relationship:			
CREDIT CARDS			
Name	Account no.	Current balance	Monthly payment
MORTGAGE COMPANY			
Account no.:	Address:		
AUTO LOANS			
Auto loans	Account no.	Balance	Monthly payment
OTHER LOANS, DEBTS, OR OBLIGATIONS			
Description	Account no.	Amount	
OTHER ASSETS OR SOURCES OF INCOME			
Description		Amount per month or value	

I authorize Contoso, Ltd. to verify the information provided on this form as to my credit and employment history.

Signature of applicant	Date
Signature of co-applicant, if for joint account	Date

Παράρτημα Β: «Ενδεικτική φόρμα αίτησης πίστωσης επιχείρησης»

Credit Application Form

BUSINESS CONTACT INFORMATION

Title		Date business commenced	
Company name		☐ Sole proprietorship	
Phone | Fax		☐ Partnership	
E-mail		☐ Corporation	
Registered company address City, State ZIP Code		☐ Other	

BUSINESS AND CREDIT INFORMATION

City, State ZIP Code		Bank name:	
How long at current address?		Primary business address City, State ZIP Code	
Phone		Phone	
Fax		Account number	
E-mail		Type of account	☐Savings ☐ Checking ☐ Other

BUSINESS/TRADE REFERENCES

Company name		Phone	
Address		Fax	
City, State ZIP Code		E-mail	
Type of account		Other	
Company name		Phone	
Address		Fax	
City, State ZIP Code		E-mail	
Type of account		Other	
Company name		Phone	
Address		Fax	
City, State ZIP Code		E-mail	
Type of account	☐Savings ☐ Checking ☐ Other	Other	

agreement

1. All invoices are to be paid 30 days from the date of the invoice.
2. Claims arising from invoices must be made within seven working days.
3. By submitting this application, you authorize [Company Name] to make inquiries into the banking and business/trade references that you have supplied.

SIGNATURES

Signature		Signature	
Name and Title		Name and Title	
Date		Date	

Παράρτημα Γ: «οι εντολές για την ανάλυση των δεδομένων με R»

3.2 Ο πιστωτικός κίνδυνος με λογιστική παλινδρόμηση - πρώτη προσέγγιση

```
my_data = read.table(file = "clipboard",sep = "\t", header=TRUE)
table(table(my_data$student))
prop.table(table(my_data$student))
for (i in 1:7500){
 if (my_data[i,2]<=1800){
 my_data[i,2]=1}
 else{
my_data[i,2]=2}}
table(table(my_data$balance))
prop.table(table(my_data$balance))
for (i in 1:7500){
 if (my_data[i,3]<=20000){
my_data[i,3]=1}
else{
if (my_data[i,3]<=40000){
my_data[i,3]=2}
else{
my_data[i,3]=3}}}
table(table(my_data$income))
prop.table(table(my_data$income))
```

**Οι ανωτέρω εντολές αφορούν το training set.*
Αντιστοίχως προκύπτουν οι αντίστοιχες ποσότητες για το test set.

```
my_data$student=as.numeric(my_data$student)
for (i in 1:7500){
if (my_data[i,1]==2){
my_data[i,1]=1}
else{
my_data[i,1]=2}}
my_data$student=as.factor(my_data$student)
my_data$balance=as.factor(my_data$balance)
my_data$income=as.factor(my_data$income)
my_data$Y=as.factor(my_data$Y)
Way1=table(my_data$student,my_data$Y)
Chisq1 = chisq.test(Way1)
Way2=table(my_data$balance,my_data$Y)
Chisq2 = chisq.test(Way2)
Way3=table(my_data$income,my_data$Y)
Chisq3 = chisq.test(Way3)

fit=glm(my_data$Y~my_data$student+my_data$balance+my_data$income,family=binomial)
summary(fit)

MyDataFrame=as.data.frame(my_data)
fit=glm(MyDataFrame$Y~ MyDataFrame$student+ MyDataFrame $balance+ MyDataFrame
$income,family=binomial(logit),data= MyDataFrame)
h1= logitgof(MyDataFrame$Y,fitted(fit))
h1$observed
h1$expected
```

```
fit1=glm(my_data$Y~ my_data$balance+my_data$income,family=binomial)
summary(fit1)
fit2=glm(my_data$Y~my_data$student+ my_data$income,family=binomial)
summary(fit2)
fit3=glm(my_data$Y~my_data$student+my_data$balance,family=binomial)
summary(fit3)
ddev1=fit1$deviance-fit$deviance
ddf1=fit1$df.residual-fit$df.residual
pvalue1 =1-pchisq(ddev1, ddf1)
ddev2=fit2$deviance-fit$deviance
ddf2=fit2$df.residual-fit$df.residual
pvalue2 =1-pchisq(ddev2, ddf2)
ddev3=fit3$deviance-fit$deviance
ddf3=fit3$df.residual-fit$df.residual
pvalue3 =1-pchisq(ddev3, ddf3)

fit4=glm(my_data$Y~ my_data$balance,family=binomial)
summary(fit4)
ddev11=fit4$deviance-fit1$deviance
ddf11=fit4$df.residual-fit1$df.residual
pvalue11 =1-pchisq(ddev11, ddf11)
ddev13=fit4$deviance-fit3$deviance
ddf13=fit4$df.residual-fit3$df.residual
pvalue13 =1-pchisq(ddev13, ddf13)

step(fit, direction="backward", test="Chisq")
```

```
pearson.res=residuals(fit4,type="pearson")
deviance.res=residuals(fit4,type="deviance")
st.pearson.res=residuals(fit4,type="pearson")/(sqrt(1-hatvalues(fit4)))
st.deviance.res=residuals(fit4,type="deviance")/(sqrt(1-hatvalues(fit4)))
my_data$Y=as.numeric(my_data$Y)
res.lik=sign(my_data$Y-fitted.values(fit4))*sqrt((hatvalues(fit4)*(st.pearson.res)^2)+((1-
hatvalues(fit4))*(st.deviance.res)^2))
id=1:7500
par(mfrow=c(1,2))
plot(fitted.values(fit4),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)
par(mfrow=c(1,2))
plot(fitted.values(fit4),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)
par(mfrow=c(1,2))
plot(hatvalues(fit4),res.lik, xlab="Fitted values", ylab="Likelihood residuals")
abline(h=0)
plot(id,res.lik, ylab="Likelihood residuals")
abline(h=0)
par(mfrow=c(1,2))
plot(id,cooks.distance(fit4), ylab="Cook's distance")
plot(id,hatvalues(fit4), ylab="Hat values")


pred=fitted.values(fit4)
roc(my_data$Y,pred,plot=TRUE,xlab="1-Specificity")

MyTest=read.table(file="clipboard",sep="\t",header=TRUE)
MyTest$Y=as.factor(MyTest$Y)
test_set_bal=read.table(file="clipboard",sep="\t",header=TRUE)
for (i in 1:2500){
if (test_set_bal[i,1]<=1800){
test_set_bal[i,1]=1}
else{
test_set_bal[i,1]=2}}
prediction=function(bal2){
return(exp(-3.9851+4.3945*bal2)/(1+exp(-3.9851+4.3945*bal2)))}
my_predictions=vector(mode="integer",length=2500)
for(i in 1:length(my_predictions)){
my_predictions[i]=prediction(test_set_bal$balance[i])}
roc(MyTest$Y,my_predictions,plot=TRUE,xlab="1-Specificity")
```

3.3 Ο πιστωτικός κίνδυνος με λογιστική παλινδρόμηση - δεύτερη προσέγγιση

```
my_data = read.table(file = "clipboard",sep = "\t", header=TRUE)
my_data$student=as.numeric(my_data$student)
for (i in 1:7500){
if (my_data[i,1]==2){
my_data[i,1]=1}
else{
my_data[i,1]=2}}
my_data$student=as.factor(my_data$student)
my_data$Y=as.factor(my_data$Y)

fit=glm(my_data$Y~my_data$student+my_data$balance+my_data$income,family=binomial)
summary(fit)
fit1=glm(my_data$Y~ my_data$balance+my_data$income,family=binomial)
summary(fit1)
fit2=glm(my_data$Y~my_data$student+ my_data$income,family=binomial)
summary(fit2)
fit3=glm(my_data$Y~my_data$student+my_data$balance,family=binomial)
summary(fit3)
ddev1=fit1$deviance-fit$deviance
ddf1=fit1$df.residual-fit$df.residual
pvalue1 =1-pchisq(ddev1, ddf1)
ddev2=fit2$deviance-fit$deviance
ddf2=fit2$df.residual-fit$df.residual
pvalue2 =1-pchisq(ddev2, ddf2)
ddev3=fit3$deviance-fit$deviance
ddf3=fit3$df.residual-fit$df.residual
pvalue3 =1-pchisq(ddev3, ddf3)
fit4=glm(my_data$Y~ my_data$balance,family=binomial)
summary(fit4)
ddev13=fit4$deviance-fit3$deviance
ddf13=fit4$df.residual-fit3$df.residual
pvalue13 =1-pchisq(ddev13, ddf13)

step(fit, direction="backward", test="Chisq")
```

```
pearson.res=residuals(fit3,type="pearson")
deviance.res=residuals(fit3,type="deviance")
st.pearson.res=residuals(fit3,type="pearson")/(sqrt(1-hatvalues(fit3)))
st.deviance.res=residuals(fit3,type="deviance")/(sqrt(1-hatvalues(fit3)))
my_data$Y=as.numeric(my_data$Y)
res.lik=sign(my_data$Y-fitted.values(fit3))*sqrt((hatvalues(fit3)*(st.pearson.res)^2)+((1-
hatvalues(fit3))*(st.deviance.res)^2))
id=1:7500
par(mfrow=c(1,2))
plot(fitted.values(fit3),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)
par(mfrow=c(1,2))
plot(fitted.values(fit3),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)
par(mfrow=c(1,2))
plot(hatvalues(fit3),res.lik, xlab="Fitted values", ylab="Likelihood residuals")
abline(h=0)
plot(id,res.lik, ylab="Likelihood residuals")
abline(h=0)
par(mfrow=c(1,2))
plot(id,cooks.distance(fit3), ylab="Cook's distance")
plot(id,hatvalues(fit3), ylab="Hat values")

pred=fitted.values(fit3)
roc(my_data$Y,pred,plot=TRUE,xlab="1-Specificity")

my_data=read.table(file="clipboard",sep="\t",header=TRUE)
my_data$Y=as.factor(my_data$Y)
 my_data$student=as.numeric(my_data$student)
 for (i in 1:2500){
 if (my_data[i,1]==2){
 my_data[i,1]=1}
 else{
 my_data[i,1]=2}}
my_predictions=vector(mode="integer",length=2500)
prediction=function(stud2,bal){
+ return(exp(-11.79+0.6814*stud2+0.005982*bal)/(1+exp(-11.79+0.6814*stud2+0.005982*bal)))}
 for(i in 1:length(my_predictions)){
+ my_predictions[i]=prediction(my_data$student[i],my_data$balance[i])}
roc(MyTest$Y,my_predictions,smooth=TRUE,plot=TRUE,xlab="1-Specificity")
```

3.4. Κατασκευή και ερμηνεία του μοντέλου δένδρου απόφασης

```
my_data=read.table(file="clipboard",sep="\t",header=TRUE)
my_data$student=as.factor(my_data$student)
my_data$Y=as.factor(my_data$Y)
My_DataFrame=as.data.frame(my_data)
str(My_DataFrame)
table(My_DataFrame$Y)
TreeModel=rpart(My_DataFrame$Y~My_DataFrame$student+My_DataFrame$balance+
My_DataFrame$income, data=My_DataFrame, method="class")
summary(TreeModel)
rpart.plot(TreeModel)
 rpart.plot(TreeModel, type=3, extra=101, fallen.leaves=T)
tot_count=function(x, labs, digits, varlen)
 {
 paste(labs, "\n\nn =", x$frame$n)
 }
 prp(TreeModel, faclen = 0, cex = 0.8, node.fun=tot_count)
tree=rpart(My_DataFrame$Y~My_DataFrame$student+My_DataFrame$balance+
My_DataFrame$income, data = My_DataFrame, method="class", control = rpart.control(cp = 0.01))
printcp(tree)
bestcp = tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"]
tree.pruned = prune(tree, cp = bestcp)
tot_count <- function(x, labs, digits, varlen)
{
paste(labs, "\n\nn =", x$frame$n)
}
prp(tree.pruned, faclen = 0, cex = 0.8, node.fun=tot_count)

MyPrediction1_TreePruned=predict(tree.pruned, My_DataFrame, type="prob")
my_predictions=vector(mode="integer",length=7500)
for (i in 1:7500){
my_predictions[i]=MyPrediction1_TreePruned[i,2]}
roc(My_DataFrame$Y,my_predictions,plot=TRUE,xlab="1-Specificity")

my_data=read.table(file="clipboard",sep="\t",header=TRUE)
my_data$student=as.factor(my_data$student)
my_data$Y=as.factor(my_data$Y)
My_DataFrame=as.data.frame(my_data)
tree=rpart(My_DataFrame$Y~My_DataFrame$student+My_DataFrame$balance+
My_DataFrame$income, data = My_DataFrame, method="class", control = rpart.control(cp = 0.01))
bestcp = tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"]
tree.pruned = prune(tree, cp = bestcp)
MyPrediction1_TreePruned=predict(tree.pruned, My_DataFrame, type="prob")
my_predictions=vector(mode="integer",length=2500)
for (i in 1:2500){
my_predictions[i]=MyPrediction1_TreePruned[i,2]}
roc(My_DataFrame$Y,my_predictions,plot=TRUE,xlab="1-Specificity")
```