Θέμα: Credit Scoring – Retail Banking (Credit Cards)

1. Εισαγωγή

1.1 Τι είναι το credit scoring?

Credit scoring είναι το σύνολο προτύπων απόφασης και των τεχνικών τους που βοηθούν τους δανειστές στη χορήγηση της καταναλωτικής πίστης. Αυτές οι τεχνικές αποφασίζουν ποιος θα πάρει την πίστωση, πόση πίστωση πρέπει να πάρουν και ποιες λειτουργικές στρατηγικές θα ενισχύσουν την αποδοτικότητα των οφειλετών απέναντι στους δανειστές. Οι Credit scoring τεχνικές αξιολογούν τον κίνδυνο του δανεισμού για έναν συγκεκριμένο καταναλωτή.

Ένας δανειστής καλείται να λάβει δύο τύπους αποφάσεων – πρώτον, εάν θα χορηγήσει πίστωση σε έναν νέο υποψήφιο και, δεύτερον, πώς να χειριστεί υπάρχοντες υποψηφίους, συγκεκριμένα αν θα αυξήσει ή όχι τα πιστωτικά τους όρια. Οι τεχνικές που βοηθούν την πρώτη απόφαση καλούνται **credit scoring**, ενώ οι τεχνικές που βοηθούν το δεύτερο τύπο απόφασης καλούνται **behavioral scoring**.

Η φιλοσοφία στην οποία στηρίζεται το credit scoring είναι ο πραγματισμός και η εμπειριοκρατία. Ο στόχος του credit scoring και του behavioral είναι να προβλέψει τον κίνδυνο, και όχι να τον εξηγήσει. Τα τελευταία 50 έτη, ο στόχος ήταν να προβλεφθεί ο κίνδυνος ένας καταναλωτής να μην ανταποκριθεί στις υποχρεώσεις του που πηγάζουν από ένα δάνειο. Πρόσφατα, εμφανίστηκε μία νέα προσέγγιση σύμφωνα με την οποία υπάρχει ενδιαφέρον να προβλεφθεί ο κίνδυνος ένας καταναλωτής να μην ανταποκριθεί σε μια ταχυδρομική ενημέρωση για ένα νέο προϊόν, ο κίνδυνος ένας καταναλωτής να μην χρησιμοποιήσει ένα πιστωτικό προϊόν, ή ακόμα ο κίνδυνος ένας καταναλωτής να μεταφέρει τον λογαριασμό του σε έναν άλλο δανειστή. Οποιαδήποτε η χρήση, το ζωτικής σημασίας σημείο είναι ότι το credit scoring είναι προάγγελος του κινδύνου, και δεν είναι απαραίτητο ότι το προβλεπτικό μοντέλο εξηγεί επίσης γιατί μερικοί καταναλωτές θα ανταποκριθούν στις υποχρεώσεις τους και άλλοι όχι. Η δύναμη του credit scoring είναι ότι η μεθοδολογία του είναι υγιής και ότι τα στοιχεία που χρησιμοποιεί προκύπτουν εμπειρικά.

Οποιαδήποτε τεχνική χρησιμοποιηθεί, behavioral ή credit scoring, το ζωτικής σημασίας σημείο είναι ότι πρέπει να υπάρχει διαθέσιμο ένα πολύ μεγάλο δείγμα των προηγούμενων πελατών με τις λεπτομέρειες των αιτήσεων καθώς επίσης και η πιστωτική ιστορία τους, λόγω της εξάρτησης των συστημάτων credit scoring είναι

στην προηγούμενη απόδοση των καταναλωτών οι οποίοι είναι παρόμοιοι με εκείνους που θα αξιολογηθούν στο πλαίσιο του συστήματος. Η υλοποίηση ενός συστήματος credit scoring αρχίζει συνήθως με τη λήψη ενός δείγματος των προηγούμενων πελατών που υπέβαλαν αίτηση για το προϊόν όσο το δυνατόν πιο κοντά χρονολογικά και με όσο το δυνατόν αξιόπιστα στοιχεία τα οποία να αφορούν την πιστωτική τους ιστορία. Εάν αυτό δεν είναι δυνατό επειδή έχουμε ένα νέο προϊόν ή μόνο μερικοί καταναλωτές το έχουν χρησιμοποιήσει στο παρελθόν, τα συστήματα μπορούν να στηριχθούν σε μικρά δείγματα ή δείγματα τα οποία έχουν προέλθει από παρόμοια προϊόντα, αλλά το σύστημα που θα προκύψει σε μία τέτοια περίπτωση δεν θα είναι τόσο καλό στην πρόβλεψη του κινδύνου όσο ένα σύστημα το οποίο θα στηριζόταν στην απόδοση των προηγούμενων πελατών για ένα όμοιο προϊόν.

Όλες οι τεχνικές χρησιμοποιούν το δείγμα για να προσδιορίσουν τις συνδέσεις μεταξύ των χαρακτηριστικών των καταναλωτών και κατά πόσο ο καταναλωτής είναι "ενήμερος" ή "σε καθυστέρηση" βάσει της προηγούμενης ιστορίας τους.

Πολλές από τις μεθόδους οδηγούν σε μία scorecard, όπου στα χαρακτηριστικά δίνεται ένα αποτέλεσμα (score) και από το σύνολο αυτών των αποτελεσμάτων προκύπτει εάν ο κίνδυνος «ένας καταναλωτή να αποδειχτεί κακός » είναι πάρα πολύ μεγάλος για να γίνει αποδεχτός. Άλλες τεχνικές δεν οδηγούν σε τέτοιες scorecards αλλά αντ' αυτού δείχνουν άμεσα την πιθανότητα «ο καταναλωτής να είναι καλός» και έτσι εάν η αποδοχή του αξίζει.

1.2 Η Ιστορία του credit scoring.

Ενώ η ιστορία της πίστωσης αρχίζει σχεδόν πριν 5000 έτη, η ιστορία του credit scoring είναι μόνο 50 ετών. Το Credit scoring είναι ουσιαστικά ένας τρόπος να προσδιοριστούν οι διαφορετικές ομάδες σε έναν πληθυσμό όταν δεν μπορεί να δει κάποιος το χαρακτηριστικό που καθορίζει τις ομάδες αυτές.

Η πρώτη προσέγγιση στην επίλυση αυτού του προβλήματος τις ταξινόμησης σε ομάδες ενός πληθυσμό εισήχθη στην στατιστική από τον Fisher (1936).Ο οποίος επιδίωξε να ταξινομήσει σε δύο ποικιλίες της ίριδας βάσει των μετρήσεις του φυσικού μεγέθους των φυτών και να διαφοροποιήσει την προέλευση των χαρακτηριστικών τους χρησιμοποιώντας τις φυσικές τους μετρήσεις. Το 1941, ο Durand ήταν ο πρώτος που διατύπωσε την άποψη ότι θα μπορούσαν να χρησιμοποιηθούν οι ίδιες τεχνικές για να γίνει η διάκριση μεταξύ των ‘ενήμερων’ και ‘σε καθυστέρηση’ δανείων. Όμως το ερευνητικό πρόγραμμά του για το USA National Bureau Economic Research δεν χρησιμοποιήθηκε για οποιοδήποτε προβλεπτικό σκοπό.

Κατά τη διάρκεια το 1930 μερικές επιχειρήσεις παραγγελιών μέσω ταχυδρομείου είχαν εισαγάγει τα αριθμητικά συστήματα υπολογισμού score για να προσπαθήσουν να υπερνικήσουν τις ασυνέπειες στις πιστωτικές αποφάσεις των πιστωτικών αναλυτών. Με την έναρξη του δευτέρου παγκοσμίου πολέμου όλοι οι οίκοι χρηματοδότησης και οι επιχειρήσεις παραγγελιών μέσω ταχυδρομείου άρχισαν να αντιμετωπίζουν δυσκολίες με την πιστωτική διαχείριση. Οι πιστωτικοί αναλυτές καλούνταν να εκπληρώσουν τις στρατιωτικές τους υποχρεώσεις, και υπήρξε μια μεγάλη έλλειψη ανθρώπων με την πείρα αυτήν. Ως εκ τούτου οι εταιρίες ζήτησαν από τους αναλυτές τους να καταγράψουν τις εμπειροτεχνικές μεθόδους που χρησιμοποίησαν για να αποφασίσουν σε ποιους να δώσουν τα δάνεια. Μερικές από αυτές ήταν αριθμητικά συστήματα scoring που είχαν ήδη χρησιμοποιηθεί ενώ άλλα ήταν σύνολα όρων που έπρεπε να ικανοποιηθούν. Αυτοί οι κανόνες χρησιμοποιήθηκαν έπειτα από μη τους εμπειρογνώμονες για να βοηθηθούν να λάβουν τις πιστωτικές αποφάσεις – ήταν ένα πρώτο παράδειγμα των έμπειρων συστημάτων.

Δεν χρειάστηκε πολύς χρόνος μετά το τέλος του πολέμου για κάποιους για να συνδέσουν την αυτοματοποίηση των πιστωτικών αποφάσεων με τις τεχνικές ταξινόμησης που αναπτύσσονταν στην στατιστική και για να δουν το όφελος από την χρήση μοντέλων που προέρχονταν από την στατιστική στην λείψει αποφάσεων για δανεισμό (Wonderlic 1952). Η πρώτη γνωμοδότηση διαμορφώθηκε στο Σαν Φρανσίσκο από τον Bill Fair και τον Earl Issac στις αρχές της δεκαετίας του '50, και οι πελάτες τους ήταν κυρίως οίκοι χρηματοδότησης, λιανοπωλητές, και επιχειρήσεις παραγγελιών μέσω ταχυδρομείου.

Η άφιξη των πιστωτικών καρτών προς το τέλος της δεκαετίας του '60 έκανε τις τράπεζες και άλλους εκδότες πιστωτικών καρτών να διαπιστώσουν τη χρησιμότητα του credit scoring. Ο αριθμός ανθρώπων που υπέβαλλαν αίτηση για πιστωτικές κάρτες κάθε ημέρα κατέστησε αδύνατη την επεξεργασία μίας μίας αίτησης τόσο από οικονομικής πλευράς όσο και ανθρωπίνου δυναμικού και ώθησε στην αυτοματοποίηση της απόφαση δανεισμού. Η αύξηση της υπολογιστικής δύναμης κατέστησε αυτήν την αλλαγή δυνατή. Οι ενδιαφερόμενες εταιρίες βρήκαν το credit scoring ως καλύτερο κριτήριο απόφασης από οποιοδήποτε υποκειμενική μέθοδο, με αποτέλεσμα τα ποσοστά των ‘κακών’ δανείων να ελαττωθούν 50% ή ακόμα και περισσότερο.

Στη δεκαετία του '80, η επιτυχία του credit scoring στις πιστωτικές κάρτες ‘έπεισε’ τις τράπεζες οι οποίες άρχισαν το scoring και σε άλλα προϊόντα, όπως τα προσωπικά δάνεια, ενώ τα τελευταία έτη, το scoring εφαρμόζεται και στα εγχώρια δάνεια και τα δάνεια μικρών επιχειρήσεων. Στη δεκαετία του '90, η αύξηση του άμεσου μάρκετινγκ οδήγησε στη χρήση των scorecards για να βελτιωθεί το ποσοστό απάντησης στις διαφημιστικές εκστρατείες. Η πρόοδος στην πληροφορική και η

αύξηση της υπολογιστικής δύναμης των υπολογιστών επέτρεψαν και σε άλλες τεχνικές να δοκιμαστούν για την κατασκευή scorecards. Στη δεκαετία του '80 εισήχθησαν τα δύο κύρια εργαλεία κατασκευής credit scoring συστημάτων, η λογιστική παλινδρόμηση και το γραμμικό μοντέλο. Πιο πρόσφατα εισήχθησαν οι τεχνικές τεχνητής νοημοσύνης, όπως τα έμπειρα συστήματα και τα νευρωνικά δίκτυα.

2. Ο Στόχος μας

2.1 Εισαγωγή

Τα σημαντικά συστατικά στην παραγωγή ακριβών και ρεαλιστικών μοντέλων κινδύνου είναι ότι πρέπει να είναι ακριβείς προάγγελοι του μεμονωμένου κινδύνου και να διαθέτουν μια συστηματική μεθοδολογία κατασκευής τους. Αυτά θα αποτελέσουν το κύριο αντικείμενο αυτού της μελέτης. Προφανώς τέτοια μοντέλα κινδύνου είναι επίσης ενδιαφέροντα για τους ίδιους τους οικονομικούς μεσάζοντες. Σε αυτό το πλαίσιο θα συγκρίνουμε τις διαφορετικές μεθόδους ταξινόμησης ενός συνόλου στοιχείων πιστωτικών καρτών από μια εμπορική τράπεζα (π.χ. υπηκοότητα του υποψηφίου, σκοπός του δανείου, οικογενειακή κατάσταση, κ.λ.π. ...) . Η μελέτη αυτή γίνεται με σκοπό να καταλάβουμε τους περιορισμούς και τις δυνατότητες των διαφορετικών μεθόδων και, ειδικότερα, εκείνων οι οποίες είναι βασισμένες στις τεχνικές εκμάθησης των μηχανών. Θα ολοκληρώσουμε αυτήν την συστηματική σύγκριση χρησιμοποιώντας τις παραδοσιακές στατιστικές τεχνικές ταξινόμησης. Μια προσέγγιση πολλαπλής στρατηγικής θα χρησιμοποιηθεί, όπου τα αποτελέσματα διάφορων αλγορίθμων θα εφαρμοστούν στα ίδια στοιχεία και θα συγκριθούν για να βρούμε το καλύτερο πρότυπο. Αυτό δικαιολογείται από τη δυσκολία επιλογής ενός βέλτιστου προτύπου a priori χωρίς γνώση της πραγματικής πολυπλοκότητας ενός ιδιαίτερου συνόλου προβλημάτων ή στοιχείων.

Στην μελέτη μας θα αναλύσουμε συστηματικά ποικίλες μεθόδους συμπεριλαμβανομένου: Λογιστική Παλινδρόμηση (Logit Statistical regression), Δέντρα Αποφάσεων CART (decision-trees CART trees), Νευρωνικά Δίκτυα (neural networks), και κ-κοντινότεροι συγγενείς (k-nearest-neighbours) στο ίδιο σύνολο στοιχείων. Η διαδικασία αυτή παράγει κάποια πλεονεκτήματα: η προεπεξεργασία των στοιχείων είναι περισσότερο ομοιογενής και τα αποτελέσματα συγκρίνονται αμεσότερα. Αρχικά, πολλοί από αυτούς τους αλγορίθμους - μέθοδοι χρησιμοποιήθηκαν από τους στατιστικούς, τους φυσικούς ή τους επιστήμονες των υπολογιστών. Αλλά η χρήση τους έχει διαδοθεί τώρα επιτυχώς σε πολλές επιχειρησιακές εφαρμογές.

2.2. Στρατηγική και Μεθοδολογία

Το γενικό πρόβλημα που συναντάμε είναι αυτό της εύρεσης των αποτελεσματικών μεθοδολογιών και των αλγορίθμων για να παραγάγουμε τις μαθηματικές ή στατιστικές περιγραφές (πρότυπα) που αντιπροσωπεύουν τα σχέδια, τις τακτικότητες ή τις τάσεις στα οικονομικά ή επιχειρησιακά δεδομένα. Αυτό δεν είναι ένα νέο θέμα και επεκτείνει βασικά τις μεθόδους που χρησιμοποιούνται για δεκαετίες από τους στατιστικούς. Για τα σύνθετα πραγματικά στοιχεία, όπου ο θόρυβος, η μη γραμμικότητα και η ιδιοσυγκρασία είναι ο κανόνας, μια καλή στρατηγική είναι να υιοθετηθεί μια διεπιστημονική μέθοδος που συνδυάζει τις στατιστικές και τους αλγορίθμους εκμάθησης μηχανών. Αυτή η διεπιστημονική, στοιχείο-οδηγημένη, υπολογιστική προσέγγιση, που αναφέρεται μερικές φορές ως "Knowledge Discovery in Databases" είναι ιδιαίτερα σημαντική σήμερα λόγω της σύγκλισης τριών παραγόντων:

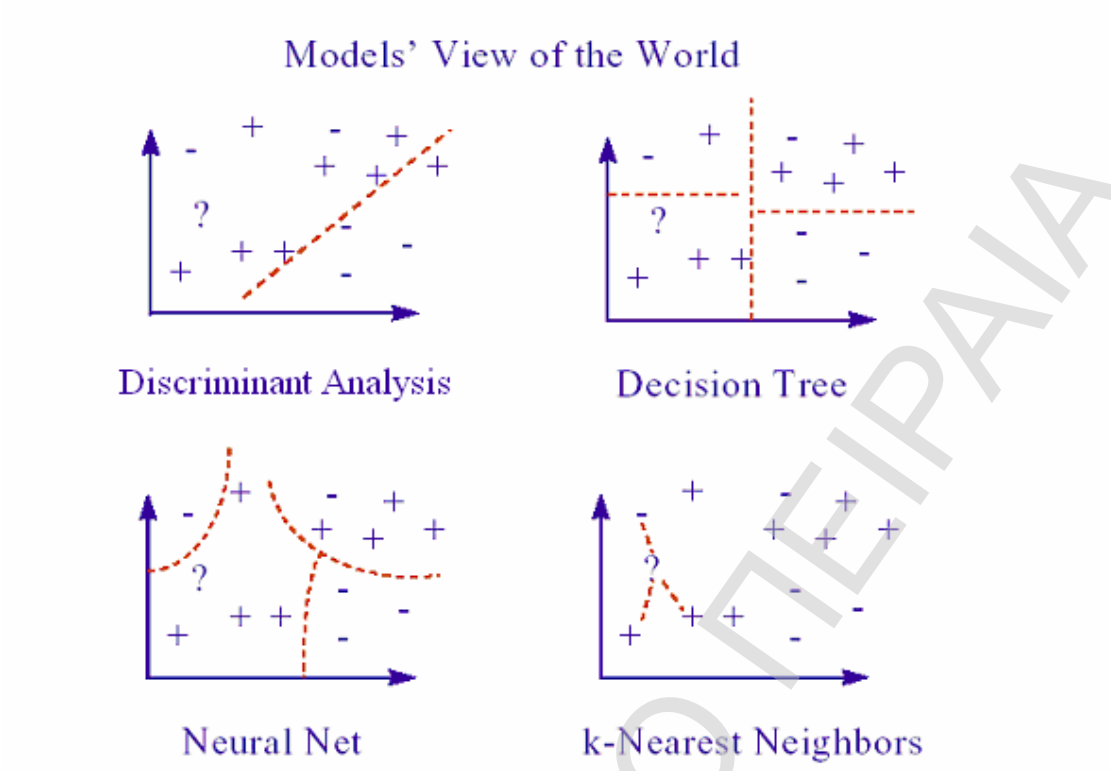
- Εταιρικές και κυβερνητικές οικονομικές βάσεις δεδομένων, όπου κάθε οικονομική συναλλαγή μπορεί να αποθηκευτεί και να παρασχεθεί για ανάλυση. Η ευρεία χρήση των βάσεων δεδομένων Datawarehouse και των εξειδικευμένων βάσεων δεδομένων έχει ανοίξει τη δυνατότητα παραγωγής οικονομικών μοντέλων σε μια πρωτοφανή κλίμακα.
- Ώριμες στατιστικές μέθοδοι και τεχνολογίες εκμάθησης μηχανών. Υπάρχει ένας μεγάλος αριθμός ώριμων και αποδεδειγμένων αλγορίθμων. Τα πρόσφατα αποτελέσματα για τις στατιστικές μεθόδους, τη θεωρία γενίκευσης, τις μηχανές που μαθαίνουν, και την πολυπλοκότητα έχουν παράσχει νέες οδηγίες και βαθιές ιδέες στα γενικά χαρακτηριστικά και τη φύση της πρότυπης διαδικασίας building / learning / fitting.
- Προσιτή υπολογιστική δύναμη συμπεριλαμβανομένων των κεντρικών υπολογιστών με πολλούς επεξεργαστές υψηλής απόδοσης, των ισχυρών υπολογιστών γραφείου, και των μεγάλων ικανοτήτων αποθήκευσης και δικτύωσης. Η τυποποίηση των λειτουργικών συστημάτων και των περιβαλλόντων έχει διευκολύνει την ολοκλήρωση και τη διασύνδεση των βάσεων δεδομένων, των repositories και των εφαρμογών.

Υπάρχουν πολλοί αλγόριθμοι διαθέσιμοι για την πρότυπη κατασκευή, έτσι ένα από τα κύρια προβλήματα είναι στην πράξη αυτό της επιλογής ή του συνδυασμού των αλγορίθμων. Δυστυχώς είναι δύσκολο να επιλεχτεί ένας αλγόριθμος a priori επειδή κάποιος είναι δύσκολο έως αδύνατο να γνωρίζει τη φύση και τα χαρακτηριστικά του συνόλου στοιχείων. Οι αλγόριθμοι ποικίλλουν πάρα πολύ στη

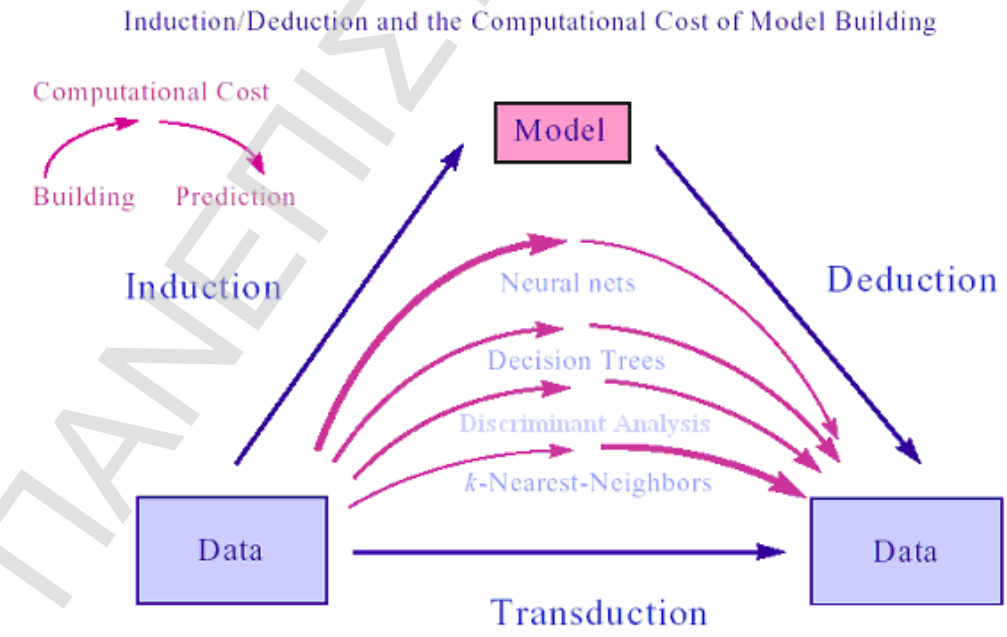
βασικά δομή, τις παραμέτρους και τους χώρους βελτιστοποίησής τους αλλά μπορούν κατά προσέγγιση να ταξινομηθούν σε μερικές βασικές οικογένειες:

- Παραδοσιακές στατιστικές μέθοδοι: linear, quadratic and logistic discriminants, regressions analysis, MANOVA etc. {Hand (1981), Lachenbruch and Mickey (1975), Eaton(1983)}, Bayesian Inference and Networks {Fayyad(1996), Berger(1985), Carlin and Louis(1996)}.
- Σύγχρονες στατιστικές: k-Nearest-Neighbors, projection pursuit, ACE, SMART, MARS etc. {Michie et al (1994), McLachan (1992), Weiss and Kulikowski(1991)}.
- Decision trees and rule based induction methods: CART, C5.0, decision trees, expert systems.{ Michie et al(1994), Mitchell(1997)}.
- Neural networks and related machines: feedforward ANN, self-organized maps, radial base functions, support vector machines, Genetic algorithms and intelligent-agents { Michie et al(1994), Mitchell(1997), Hassoun(1995), White(1992) , Goldberg (1989)}.
- Fuzzy logic, fractal sampling and hybrid approaches. {Zadeh (1994)}.
- Model combination methods: boosting and bagging. {Freund and Shapire(1995), Breiman (1996)}.

Κάθε αλγόριθμος υιοθετεί μια διαφορετική μέθοδο για να εγκαταστήσει τα στοιχεία και να προσεγγίσει τους κανόνες ή τους συσχετισμούς σύμφωνα με μια συγκεκριμένη δομή ή μια αντιπροσώπευση (representation). Σε αυτήν την μελέτη επιλέγουμε τέσσερις διαφορετικούς αλγορίθμους που αντιπροσωπεύουν τέσσερις σημαντικές κατηγορίες προβλέψεων: τα Δέντρα Αποφάσεων (Decision Trees), τα Νευρωνικά Δίκτυα (feedforward neural networks), την μέθοδο των κ-κοντινότερων γειτόνων (k-nearest-neighbors) και την λογιστική παλινδρόμηση (logit regression).



Η διαδικασία κατασκευής ενός μοντέλου και η εφαρμογή του σε νέα δεδομένα περιλαμβάνει υπολογιστικές δαπάνες. Αυτές οι δαπάνες μπορούν να περιορίσουν τον τύπο του μοντέλου που μπορεί να χρησιμοποιηθεί σε μία συγκεκριμένη κατάσταση. Το παρακάτω σχέδιο παρουσιάζει τις σχέσεις μεταξύ των στοιχείων, του μοντέλου, και των παραγωγικών, επαγωγικών και transductive διαδικασιών.



3. Μεθοδολογία

3.1. Εισαγωγή

Το πρόβλημα του Credit Scoring υπάγεται σε μια γενικότερη κατηγορία προβλημάτων λήψης αποφάσεων που αφορούν στην ταξινόμηση περιπτώσεων σε πληθυσμούς ή την πρόβλεψη της πιθανής ομάδας στην οποία οι περιπτώσεις ανήκουν.

Θεωρούμε ότι καθεμία από τις περιπτώσεις ανήκει σε μια και μόνο ομάδα η αλλιώς κλάση από ένα πεπερασμένο σύνολο αμοιβαία αποκλειόμενων κλάσεων. Η λύση τέτοιων προβλημάτων βασίζεται στην ανάπτυξη έξυπνων συστημάτων, που είτε στηρίζονται σε αλγόριθμους εκμάθησης (learning systems) είτε αποτελούν έμπειρα συστήματα (expert systems).

Ένα σύστημα εκμάθησης βασίζεται σε ένα δείγμα που αποτελείται από γνωστές περιπτώσεις, με την έννοια ότι είναι γνωστός ο πληθυσμός προέλευσής τους, για την κατασκευή κανόνων ταξινόμησης ή πρόβλεψης της κλάσης νέων περιπτώσεων. Τα δεδομένα που εισάγονται στο σύστημα αποτελούνται από παρατηρήσεις με μετρήσεις για συγκεκριμένα χαρακτηριστικά, τα οποία αντιπροσωπεύουν ιδιότητες που μπορεί να παίρνουν συνεχείς ή διακριτές τιμές, και από την αντίστοιχη σωστή κατάταξη τους σε ομάδες. Αντικειμενικός σκοπός του συστήματος εκμάθησης είναι η κατασκευή ενός κανόνα ταξινόμησης για το συγκεκριμένο πρόβλημα, με τον προσδιορισμό της σχέσης κάθε ξεχωριστού προτύπου που εμφανίζεται στις παρατηρήσεις με μία από τις προκαθορισμένες κλάσεις. Τελικός στόχος είναι, βέβαια, η αυτόματη λήψη απόφασης για την κατάταξη νέων περιπτώσεων. Επομένως, το σύστημα δέχεται ένα σύνολο δεδομένων με συγκεκριμένη δομή ως είσοδο (input), αναγνωρίζει τη δομή και παράγει έναν κανόνα αποφάσεων στην έξοδο (output).

Είναι προφανές ότι είναι ιδιαίτερα ουσιώδες τα δεδομένα να περιέχουν δομή. Από τη στιγμή που ο σκοπός του συστήματος είναι η πρόβλεψη και όχι η διάκριση των ήδη διαθέσιμων δειγματικών περιπτώσεων, τα κριτήρια της απόφασης κατάταξης σε ομάδες πρέπει να ισχύουν για νέες περιπτώσεις. Αυτό συμβαίνει όταν για τα χαρακτηριστικά που μετρώνται για κάθε παρατήρηση μπορούν να αναγνωριστούν ξεκάθαρα πρότυπα, δηλαδή όταν τα συγκεκριμένα χαρακτηριστικά έχουν καλή διακριτική και προβλεπτική ικανότητα, και όχι όταν τα διαθέσιμα δεδομένα είναι τυχαία και μπορούν να θεωρηθούν θόρυβος. Κάθε σύστημα επαγωγικής εκμάθησης προσπαθεί να εξάγει τη μέγιστη ποσότητα πληροφόρησης από ένα δείγμα και γι αυτό η προβλεπτική ικανότητα των χρησιμοποιούμενων χαρακτηριστικών έχει τεράστια σημασία για την επιτυχία του συστήματος.

Ένα σύστημα εκμάθησης συνδυάζει το γενικό μοντέλο του κανόνα ταξινόμησης που παράγεται από μια συγκεκριμένη μέθοδο με τα διαθέσιμα δειγματικά δεδομένα,

για την κατασκευή του κατάλληλου κανόνα ταξινόμησης που θα εφαρμοστεί στο συγκεκριμένο υπό μελέτη πρόβλημα.

Συνοπτικά, οι επιθυμητές ιδιότητες για έναν αλγόριθμο εκμάθησης είναι:

1. Ο παραγόμενος κανόνας ταξινόμησης να αναπαριστά γενικές αρχές ή πρότυπα που χαρακτηρίζουν το γνωστικό αντικείμενο του προβλήματος.
2. Ο κανόνας να μπορεί να προβλέπει αποδοτικά την κλάση ή ομάδα προέλευσης άγνωστων περιπτώσεων.
3. Η λύση στο πρόβλημα της ταξινόμησης που παρέχει το σύστημα να είναι εύκολα αντιληπτή και κατανοητή από τον άνθρωπο.
4. Ο αλγόριθμος εκμάθησης να μπορεί να χειριστεί μικτά, ελλιπή και ανακριβή δεδομένα.

Τα συστήματα εκμάθησης δίνουν τη δυνατότητα άντλησης πληροφορίας από εμπειρία συσσωρευμένη σε βάσεις δεδομένων, όμως αδυνατούν να εκμεταλλευτούν την πολύπλευρη γνώση που ένας ειδικός θα μπορούσε να εισάγει στην προσπάθεια αντιμετώπισης ενός δεδομένου προβλήματος. Από την άλλη πλευρά, μία άλλη κατηγορία έξυπνων συστημάτων, τα έμπειρα συστήματα, βασίζονται σε εμπειρικούς κανόνες που εμπεριέχουν τέτοιου είδους γνώση. Όμως, ακριβώς στο σημείο αυτό ελλοχεύει ο κίνδυνος ότι η αποτελεσματικότητα των συστημάτων θα περιορίζεται πάντα από το επίπεδο των δυνατοτήτων των εκάστοτε ειδικών. Επιπλέον, η διαδικασία κατασκευής ενός συστήματος κανόνων αποφάσεων με τη διεξαγωγή μιας σειράς συνεντεύξεων σε ειδικούς, όπως απαιτούν τα έξυπνα συστήματα, είναι μια διαδικασία ιδιαίτερα επίπονη και χρονοβόρα, ενώ τελικά και η απόφαση που βασίζεται σε μία τέτοια μέθοδο δεν μπορεί να είναι αυτόματη.

Στο σημείο αυτό πρέπει να αναφερθεί ότι ανάμεσα στους αλγορίθμους εκμάθησης δεν υπάρχει κάποιος που να θεωρείται συνολικά βέλτιστος, με την έννοια ότι αυτός δίνει τα καλύτερα αποτελέσματα για όλα τα προβλήματα εκμάθησης (Buchanan 1987). Για ένα συγκεκριμένο πρόβλημα ταξινόμησης οι διάφορες μέθοδοι θα δώσουν διαφορετικές λύσεις και τα αποτελέσματά τους θα πρέπει να συγκριθούν ώστε να χρησιμοποιηθεί η καλύτερη ως προς το συγκεκριμένο πρόβλημα. Στη συνέχεια θα περιγραφεί η διαδικασία αξιολόγησης των μεθόδων και θα ακολουθήσει η παρουσίαση των μοντέλων που χρησιμοποιήσαμε στην μελέτη μας.

3.2. Αξιολόγηση των συστημάτων

3.2.1. Μεροληψία και ακρίβεια

Κάθε αλγόριθμος εκμάθησης μπορεί να θεωρηθεί ως μία αναζήτηση γενικεύσεων οι οποίες καθιστούν δυνατή την πρόβλεψη της κλάσης νέων παρατηρήσεων. Όλοι οι αλγόριθμοι υιοθετούν κάποιες υποθέσεις για την επιλογή ενός κανόνα ταξινόμησης από το σύνολο των συνεπών με τα δεδομένα κανόνων. Το σύνολο αυτών των υποθέσεων εισάγει μεροληψία στο σύστημα εκμάθησης, η οποία περιορίζει τη δυνατότητα γενίκευσης της χρήσης του συστήματος στα διαφορετικά προβλήματα εκμάθησης. Η αξιολόγηση και σύγκριση των συστημάτων, σ' αυτό το πλαίσιο, βασίζεται στην πιθανότητα σωστής πρόβλεψης του χρησιμοποιούμενου αλγόριθμου εκμάθησης, όταν ο κανόνας ταξινόμησης ελέγχεται μόνο σε άγνωστες περιπτώσεις (Shaffer 1994).

Είναι προφανές ότι στην πράξη το ενδιαφέρον δεν εστιάζεται στην εύρεση ενός αλγορίθμου που να δίνει τα καλύτερα αποτελέσματα σε όλα τα προβλήματα ταξινόμησης. Το ζητούμενο είναι να διαπιστωθεί σε ποιες κατηγορίες προβλημάτων υπερέχει κάποιος αλγόριθμος ως προς την μεροληψία, κάτι που ελέγχεται με την αξιολόγηση των αποδόσεων διαφορετικών αλγορίθμων με πειράματα.

Η ικανότητα πρόβλεψης της κλάσης νέων περιπτώσεων των κανόνων ταξινόμησης που παράγονται από αλγόριθμους εκμάθησης για ένα δεδομένο πρόβλημα εκφράζεται με το γενικό όρο Ακρίβεια (Accuracy). Η πραγματική ακρίβεια ενός κανόνα ορίζεται ως η πιθανότητα σωστής ταξινόμησης ενός τυχαίου υποκειμένου που μπορεί να χρησιμοποιηθεί και ως μέτρο σύγκρισης μεταξύ διαφορετικών κανόνων ταξινόμησης (model selection).

Στα πλαίσια ενός πεπερασμένου δείγματος η ακρίβεια μπορεί μόνο να εκτιμηθεί και γι αυτό το λόγο πρέπει να αναζητηθούν αμερόληπτες μέθοδοι εκτίμησης. Για παράδειγμα, η φαινομενική ακρίβεια (apparent accuracy), που υπολογίζεται ως το ποσοστό των σωστών ταξινομήσεων στο σύνολο των διαθέσιμων δειγματικών περιπτώσεων, υπερεκτιμά την πραγματική ακρίβεια και δεν δίνει πληροφορία για τη συμπεριφορά του κανόνα σε νέες περιπτώσεις. Επομένως θα πρέπει να χρησιμοποιηθούν διαφορετικά δείγματα για την εκμάθηση και για τον έλεγχο του κανόνα. Για να είναι αξιόπιστη η πρόβλεψη, οι μέθοδοι εκτίμησης της ακρίβειας θα βασιστούν στη δειγματοληψία συνόλων εκμάθησης και συνόλων ελέγχου από το αρχικό δείγμα.

Ο στόχος είναι οι μέθοδοι εκτίμησης της ακρίβειας να έχουν μικρή στατιστική μεροληψία και μικρή διακύμανση. Η στατιστική μεροληψία εκφράζει πόσο κοντά είναι η εκτίμηση της ακρίβειας στην πραγματική ακρίβεια του κανόνα ταξινόμησης,

ενώ η διακύμανση εκφράζει πόσο αξιόπιστη είναι η εκτίμηση. Αναγωγή του χειρισμού των εννοιών στα προβλήματα ταξινόμησης έχει γίνει από τους Kong και Dietterich (1995) και Kohavi και Wolpret (1996).

3.2.2. Δείκτες Ταξινόμησης

Σκοπός ενός συστήματος εκμάθησης είναι η επιτυχής ταξινόμηση νέων περιπτώσεων και αυτό αποτελεί και το κριτήριο για την αξιολόγηση της λειτουργίας των χρησιμοποιούμενων από το σύστημα κανόνων. Η βελτιστοποίηση της προβλεπτικής ικανότητας ενός κανόνα ισοδυναμεί με την ελαχιστοποίηση του πραγματικού δείκτη λανθασμένης ταξινόμησης (true error rate) για τον κανόνα αυτό.

Ο πραγματικός δείκτης λανθασμένης ταξινόμησης ορίζεται ως ο δείκτης λάθους ενός κανόνα ταξινόμησης σε ασυμπτωτικά μεγάλο αριθμό νέων περιπτώσεων που προσεγγίζει στο όριό του την κατανομή του πληθυσμού (Weisss-Kulikowski, 1991). Εμπειρικά, ο δείκτης λάθους μπορεί να υπολογιστεί ως το πηλίκο του αριθμού των περιπτώσεων ενός δείγματος που ταξινομήθηκαν λανθασμένα από τον χρησιμοποιούμενο κανόνα προς το συνολικό αριθμό των δειγματικών περιπτώσεων, υποθέτοντας ότι όλες οι περιπτώσεις λανθασμένης ταξινόμησης έχουν την ίδια σημασία.

Έστω ότι ένας κανόνας ταξινόμησης χρησιμοποιείται για την τοποθέτηση υποκειμένων σε μία από δύο ομάδες (θετική και αρνητική) και έστω ότι οι συχνότητες σωστής και λανθασμένης ταξινόμησης των υποκειμένων ενός δείγματος μεγέθους N με βάση αυτό τον κανόνα δίνεται στον παρακάτω πίνακα:

		Πραγματική Ομάδα Υποκειμένου	
		Θετική	Αρνητική
Ομάδα Πρόβλεψης του κανόνα	Θετική	a	b
	Αρνητική	c	d

Οι αντίστοιχοι δείκτες δίνονται ως:

$\frac{a}{a+c}$: δείκτης σωστής θετικής ταξινόμησης (sensitivity)

$\frac{d}{b+d}$: δείκτης σωστής αρνητικής ταξινόμησης (specificity)

$\frac{a+d}{a+b+c+d}$: δείκτης σωστής ταξινόμησης (accuracy)

$1-\frac{a+d}{a+b+c+d}$: δείκτης λανθασμένης ταξινόμησης (misclassification error)

Συνήθως, για την εκτίμηση του πραγματικού δείκτη λάθους χρησιμοποιείται ο φαινομενικός δείκτης λανθασμένης ταξινόμησης (apparent error rate) ο οποίος υπολογίζεται ως δείκτης λάθους του κανόνα στο δείγμα των περιπτώσεων που χρησιμοποιήθηκαν για την κατασκευή του. Είναι φανερό ότι από το φαινομενικό δείκτη λάθους μπορεί να προκύψει η φαινομενική ακρίβεια του κανόνα ταξινόμησης. Ο φαινομενικός δείκτης τείνει να μεροληπτεί υποεκτιμώντας τον πραγματικό δείκτη λάθους και γι αυτό μια σωστότερη πρακτική είναι η κατασκευή του κανόνα ταξινόμησης και ο έλεγχος της απόδοσης του να βασιστούν σε διαφορετικά δείγματα (training and testing samples), για μια 'δικαιότερη' αξιολόγηση του κανόνα (Highleyman, 1962). Διαφορετικές τεχνικές εκτίμησης δεικτών λάθους μελέτησε ο Efron (1982).

3.2.3. Μέθοδοι Εκτίμησης της Ακρίβειας

Για την εκτίμηση της ακρίβειας ενός κανόνα ταξινόμησης έχουν προταθεί υπολογιστικές μέθοδοι (computational intensive methods) που στη βάση τους στηρίζονται στους δείκτες ταξινόμησης, έχουν όμως εξελιχθεί σε κατευθύνσεις που αντιμετωπίζουν τα προβλήματα τους. Στην συνέχεια παρουσιάζονται οι κυριότερες μέθοδοι.

3.2.3.1. Επαναληπτική Δειγματοληψία (Holdout with Random Sampling)

Σύμφωνα με την μέθοδο αυτή, ένα τυχαίο δείγμα περιπτώσεων που αποτελείται από προκαθορισμένο ποσοστό του συνολικού διαθέσιμου δείγματος χρησιμοποιείται για την κατασκευή του κανόνα ταξινόμησης και το σύνολο των περιπτώσεων που απομένουν χρησιμοποιείται για έλεγχο. Η διαδικασία επαναλαμβάνεται και κάθε φορά υπολογίζεται η ακρίβεια του κανόνα στο δείγμα ελέγχου. Η εκτίμηση της ακρίβειας είναι ο μέσος αριθμητικός των τιμών που λαμβάνονται με την παραπάνω ακολουθιακή διαδικασία.

Ένα μειονέκτημα της μεθόδου της επαναληπτικής δειγματοληψίας είναι ότι δεν χρησιμοποιεί αποδοτικά τα δεδομένα, με την έννοια ότι ένα υποσύνολο των δειγματικών περιπτώσεων χρησιμοποιείται κάθε φορά για την κατασκευή του κανόνα, με αποτέλεσμα η εκτίμηση της ακρίβειας να είναι απαισιόδοξη. Μια δεύτερη σημαντική παρατήρηση είναι ότι ενώ κάθε επανάληψη της δειγματοληψίας είναι ανεξάρτητη από την άλλη, τα υποσύνολα ελέγχου δεν είναι μεταξύ τους ανεξάρτητα,

καθώς περιέχουν κοινές περιπτώσεις. Έτσι όσο ο αριθμός των επαναλήψεων μεγαλώνει, το ποσοστό των κοινών περιπτώσεων στο συνολικό έλεγχο αυξάνει και κατά συνέπεια μειώνεται τεχνητά η διακύμανση (Kohavi, 1995).

3.2.3.2. Διεπικύρωση (Cross Validation)

Η μέθοδος της διεπικύρωσης προέκυψε ως γενίκευση της τεχνικής του μονοαποκλεισμού (leave-one-out) και οφείλεται στον Stone (1974). Στην ειδική περίπτωση του μονοαποκλεισμού (Lanchenberg-Michey, 1968), επιλέγεται μια μόνο παρατήρηση και αφήνεται στο δείγμα ελέγχου, ενώ οι υπόλοιπες χρησιμοποιούνται για την κατασκευή του κανόνα. Επειδή η μία παρατήρηση έχει αγνοηθεί κατά την κατασκευή του κανόνα, η σύγκριση της πραγματικής και της προβλεπόμενης από τον κανόνα κλάσης της παρέχει μια εικόνα για την καταλληλότητα του μοντέλου ταξινόμησης. Η διαδικασία επαναλαμβάνεται τόσες φορές όσες και οι δειγματικές περιπτώσεις και ο έλεγχος βασίζεται στο σύνολο των περιπτώσεων που αποκλείστηκαν σε κάθε επανάληψη.

Γενικεύοντας, ο αλγόριθμος μπορεί να αφήσει στο δείγμα ελέγχου περισσότερες από μια παρατηρήσεις σε κάθε επανάληψη και έτσι προκύπτει η μέθοδος της διεπικύρωσης. Σύμφωνα με αυτή λοιπόν, το αρχικό σύνολο δειγματικών περιπτώσεων χωρίζεται τυχαία σε k αμοιβαία αποκλειόμενα υποσύνολα, καθένα από τα οποία χωρίζεται και πάλι σε ένα σύνολο εκμάθησης και ένα σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται ακολουθιακά και έτσι προκύπτουν k σύνολα εκμάθησης. Η εκτίμηση της συνολικής ακρίβειας είναι ο μέσος αριθμητικός των k τιμών που υπολογίζονται από τα αντίστοιχα δείγματα ελέγχου.

Είναι φανερό ότι η μέθοδος της διεπικύρωσης επιχειρεί να διορθώσει τα μειονεκτήματα της επαναληπτικής δειγματοληψίας, αν και το γεγονός ότι η εκτίμηση της ακρίβειας είναι απαισιόδοξη παραμένει. Ωστόσο, η μεροληψία μειώνεται όσο αυξάνει το k, δυστυχώς όμως αντιστρόφως ανάλογη είναι η πορεία της διακύμανσης (Kohavi, 1995). Μ’ αυτή τη διαπίστωση, η βέλτιστη επιλογή του k μπορεί να τοποθετηθεί στο πλαίσιο της προσπάθειας εξισορρόπησης της επίδρασης της μεροληψίας και της διακύμανσης.

3.2.3.3. Bootstrap

Η μεθοδολογία Bootstrap βασίζεται στην ιδέα ότι τα δεδομένα ενός προβλήματος αποτελούν δείγμα από την άγνωστη κατανομή του πληθυσμού προέλευσής τους και άρα εμπειρική κατανομή τους είναι μια καλή εκτίμηση της πραγματικής κατανομής του πληθυσμού. Έτσι, οι μέθοδοι χρησιμοποιούν την προσομοίωση από την εμπειρική κατανομή των δεδομένων (επαναληπτικά δείγματα με επανάθεση από τα δεδομένα) για την εκτίμηση παραμέτρων και γενικά για στατιστική συμπερασματολογία για τον πληθυσμό. Η διαδικασία του υπολογιστικού αλγορίθμου είναι ανάλογη με αυτή της μεθόδου της διεπικύρωσης.

Στην πράξη, η μέθοδος διεπικύρωσης παρέχει σχεδόν αμερόληπτους εκτιμητές της ακρίβειας ενός κανόνα ταξινόμησης, ωστόσο, η μικρή μεροληψία πληρώνεται με μεγάλη μεταβλητότητα. Έχει δειχτεί ότι οι διαδικασίες Boostrap, που μπορούν να θεωρηθούν ως μια λεία εκδοχή της μεθόδου διεπικύρωσης, μπορούν να ελαττώσουν την μεταβλητότητα των προβλεπόμενων δεικτών ταξινόμησης, ενώ με κατάλληλη προσαρμογή διορθώνεται και η μεροληψία των εκτιμητών (Efron-Tibshirani, 1995).

Αντίθετα με τη μέθοδο διεπικύρωσης, η Boostrap εφαρμόζει δειγματοληψία με επανατοποθέτηση για την κατασκευή συνόλων εκμάθησης και χρησιμοποιεί τις περιπτώσεις που απομένουν στο αρχικό δείγμα για έλεγχο(Efron-Tibshirani, 1993)

Εμείς εφαρμόσαμε μια παραλλαγή της Boostrap στην οποία χωρίσαμε τα δεδομένα που είχαμε στην διάθεσή μας σε δύο σύνολα, το ένα το χρησιμοποιήσαμε για να δημιουργήσουμε δείγματα τα οποία χρησιμοποιήσαμε για την εκπαίδευση των κανόνων εκμάθησης και το άλλο για να ελέγξουμε την αποτελεσματικότητα των κανόνων που είχαν προκύψει από τα διάφορα δείγματα. Με τον τρόπο αυτό πετύχαμε να έχουμε στην διάθεσή μας ένα κοινό σύνολο στο οποίο να μπορούμε να εφαρμόσουμε τους κανόνες που είχαν προκύψει από τα διάφορα δείγματα και τα αποτελέσματα από την εφαρμογή των κανόνων να είναι συγκρίσιμα. Ως εκτιμήτρια της ακρίβειας χρησιμοποιήσαμε τόσο την φαινομενική ακρίβεια όσο και την εκτίμηση που προκύπτει από την παραπάνω διαδικασία με τον υπολογισμό του αριθμητικού μέσου.

3.3. Υποδείγματα

3.3.1. Λογιστική Παλινδρόμηση (Logit):

Το μοντέλο της λογιστικής παλινδρόμησης χρησιμοποιείται πια ευρύτερα σε εφαρμογές. Σύμφωνα με αυτό, όλοι οι υποψήφιοι δανειολήπτες ανήκουν στον ίδιο

πληθυσμό και για καθένα απ΄ αυτούς μπορεί να καθοριστεί η πιθανότητα αθέτησης της υποχρέωσης αποπληρωμής, η οποία εξαρτάται από το διάνυσμα χαρακτηριστικών του. Αυτό που παρατηρείται (y_i) στο διαθέσιμο δείγμα για κάθε πιστούχο είναι η κατάληξή του ως προς την αποπληρωμή, ως αποτέλεσμα μιας δοκιμής Bernoulli. Από το μοντέλο, όμως, παράγονται πιθανότητες (Greene, 1998).

Στα πλαίσια της θεωρίας των γενικευμένων γραμμικών μοντέλων (McCullagh-Nedler, 1989) , ως γραμμικός συνδυασμός των χαρακτηριστικών εκφράζεται μια λανθάνουσα μεταβλητή D_i (latent variable), η οποία μπορεί να θεωρηθεί ότι εκφράζει την 'τάση προς αθέτηση':

$$D_i = b_0 + b_1 X_{i1} + ... + b_p X_{ip} + e_i = bX_i + e_i, \quad e_i \sim l$$

όπου $l(.)$ η λογιστική συνάρτηση : $l(u) = \dfrac{1}{1+\exp(-u)}$ (μοντέλο Logit).

Η ψευδομεταβλητή y_i και η λανθάνουσα μεταβλητή D_i, που είναι ένα score για τον κίνδυνο, συνδέονται ως εξής:

$y_i = 1, \ an \ \ D_i > 0$

$y_i = 0, an \ \ D_i < 0$

Με βάση τα παραπάνω, η πιθανότητα το δάνειο να αποπληρωθεί σωστά δίνεται ως:

$$p_i = P(Y_i = 1 \mid X_i) = P(D_i > c \mid X_i) = P(bX_i + e_i > c) = P(e_i < bX_i) = l(bX_i)$$

είναι δηλαδή,

$$p_i = \frac{1}{1+\exp(-bX_i)}$$

κάτω από συγκεκριμένες υποθέσεις (Arminger et al., 1997)

Αφού εκτιμηθεί το λογιστικό μοντέλο, η πιθανότητα που προκύπτει για κάθε νέο υποψήφιο δανειολήπτη, με βάση το διάνυσμα χαρακτηριστικών του, θα καθορίσει την απόφαση παροχής δανείου. Στην πραγματικότητα, αυτό που εκτιμάται είναι το score που αντιστοιχεί στην πιθανότητα αθέτησης και σύμφωνα με αυτό ο υποψήφιος κατατάσσεται σε κατηγορία κινδύνου. Ο κανόνας που έχει προκύψει με την παραπάνω διαδικασία είναι ο κανόνας μεγίστης πιθανοφάνειας.

Μια θεωρητική προσέγγιση της διάκρισης, οδηγεί ίσως στην άποψη ότι η λογιστική παλινδρόμηση είναι μια πιο κατάλληλη στατιστική τεχνική για την αντιμετώπιση του προβλήματος του credit scoring σε σχέση με το απλό γραμμικό μοντέλο. Η φύση του προβλήματος, το γεγονός δηλαδή ότι η διάκριση αφορά σε δύο καθορισμένες διακριτές κατηγορίες κινδύνου και στην αντιμετώπιση του προβλήματος μπορούν να εμπλακούν οι πιθανότητες που αυτές εμφανίζουν, ενισχύει την άποψη αυτή.

Γενικά σε εφαρμογές credit scoring, η λογιστική παλινδρόμηση έχει δώσει πολύ καλά αποτελέσματα (Wiginton, 1980 και Gilbert et al.,1990).

3.3.2. Ταξινόμηση με μεθόδους Κοντινότερου Γείτονα

Το γεγονός ότι για τα δεδομένα του credit scoring δεν ικανοποιείται σχεδόν ποτέ η υπόθεση της κανονικότητας οδήγησε στην εισαγωγή των μη παραμετρικών μεθόδων smoothing, που δεν στηρίζονται σε υποθέσεις για την κατανομή των πληθυσμών των υποψηφίων, για την αντιμετώπιση του προβλήματος. Η σημαντικότερη μέθοδος μη παραμετρικής στατιστικής, η οποία και κυρίως χρησιμοποιήθηκε για την ταξινόμηση υποψηφίων δανειοληπτών σε κλάσεις πιστωτικού κινδύνου, είναι η οι μέθοδος του κοντινότερου γείτονα (ή κ-κοντινότερων γειτόνων).

Οι μέθοδοι κοντινότερου γείτονα δεν στηρίζονται σε καμία υπόθεση για την κατανομή των πληθυσμών, παρά μόνο στη διάθρωση ή τις αποστάσεις των δειγματικών σημείων (παρατηρήσεων). Η λογική είναι απλή: για κάθε νέα περίπτωση, ο κανόνας ταξινόμησης του κοντινότερου γείτονα την συγκρίνει με τα πρότυπα που περιέχονται στο δείγμα και τελικά την αποδίδει στην ομάδα που ανήκει το πιο παρεμφερές προς αυτή πρότυπο. Εναλλακτικά, αντί να προσδιορίζεται ένας μοναδικός κοντινότερος γείτονας εντοπίζονται οι k-κοντινότεροι γείτονες και η νέα παρατήρηση τοποθετείται στην ομάδα με τη μεγαλύτερη συχνότητα εμφάνισης για τις k αυτές δειγματικές περιπτώσεις. Συγκεκριμένα όταν εφαρμόζουμε την μέθοδο αυτή στο credit scoring συγκρίνουμε τα χαρακτηριστικά κάθε νέου υποψηφίου με τα αντίστοιχα χαρακτηριστικά των πιστούχων ενός ιστορικού δείγματος και τον τοποθετούμε στην κλάση που πλειοψηφεί ανάμεσα στους κ-κοντινότερους γείτονές τους. Προφανώς, πρέπει αρχικά να προσδιοριστούν κατάλληλα κριτήρια ομοιότητας, για να έχει ισχύ η αναγνώριση της πιστοληπτικής ικανότητας των νέων υποψηφίων ως προς την αντίστοιχη συμπεριφορά των χαρακτηριστικών ‘γειτόνων’ τους, με βάση τη σύγκριση με το ιστορικό δείγμα.

Τα μέτρα που συνήθως χρησιμοποιούνται για τον προσδιορισμό των αποστάσεων ανάμεσα σε μια νέα παρατήρηση και στα πρότυπα είναι:

- Η απόλυτη απόσταση
- Η Ευκλείδεια απόσταση
- Διάφορες κανονικοποιημένες αποστάσεις

Οι αποστάσεις αυτές λαμβάνουν υπόψη τις διαφορές των τιμών όλων των υπό μελέτη χαρακτηριστικών και άρα όταν τα χαρακτηριστικά είναι μετρημένα σε διαφορετικές κλίμακες η κανονικοποίηση είναι απαραίτητη.

Στην πραγματικότητα αυτό που ενδιαφέρει ως προς την σύγκριση μιας παρατήρησης με τα πρότυπα είναι ο προσδιορισμός των πιο ‘όμοιων’ προς αυτή

προτύπων. Γι αυτό το λόγο οι αλγόριθμοι κοντινότερου γείτονα χρησιμοποιούν τον ορισμό της έννοιας της ομοιότητας (similarity) μεταξύ δύο παρατηρήσεων.

Στην περίπτωση όπου τα χαρακτηριστικά εκφράζουν συνεχείς ιδιότητες η ομοιότητα μεταξύ των παρατηρήσεων x και y, η ομοιότητα μπορεί να οριστεί σύμφωνα με την Ευκλείδεια απόσταση ως:

$$S(x,y) = -\sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

όπου x_i, y_i οι τιμές του i χαρακτηριστικού για τις παρατηρήσεις x και y αντίστοιχα.

Ο ορισμός της ομοιότητας μπορεί να επεκταθεί ώστε να περιλαμβάνει και ιδιότητες που παίρνουν διακριτές τιμές ως εξής:

$$S(x,y) = -\sqrt{\sum_{i=1}^{d} f(x_i, y_i)}$$

όπου $f(x_i, y_i) = (x_i - y_i)^2$, για συνεχείς ιδιότητες και $f(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}$, για διακριτές ιδιότητες (Aha-kibler, 1991). Προκειμένου να συμμετέχουν με το ίδιο βάρος όλα τα χαρακτηριστικά στον υπολογισμό της ομοιότητας, οι τιμές των συνεχών ιδιοτήτων κανονικοποιούνται στο [0,1].

Στην πράξη, με ένα τυπικό δείγμα η μέθοδος έχει αρκετά καλή εφαρμογή αν χρησιμοποιεί χαρακτηριστικά με μεγάλη προβλεπτική ικανότητα.

Υπολογιστικά, οι μέθοδοι του κοντινότερου γείτονα δεν απαιτούν μεγάλη προσπάθεια για την εξαγωγή πληροφορίας από το δείγμα και αυτή είναι μία ουσιαστική διαφορά της από τα άλλα συστήματα εκμάθησης. Οι υπόλοιπες μέθοδοι απαιτούν κόπο και υψηλό οικονομικό κόστος για την ανάλυση του δείγματος, ενώ μετά η πρόβλεψη και ταξινόμηση νέων περιπτώσεων είναι μία αυτοματοποιημένη διαδικασία. Αντίθετα, οι μέθοδοι του κοντινότερου γείτονα εμπλέκουν μια χρονοβόρα διαδικασία στο στάδιο της ταξινόμησης, αφού κάθε νέα περίπτωση πρέπει να συγκριθεί με κάθε πρότυπο του δείγματος. Το μειονέκτημα αυτό τείνει, βέβαια, να εξαλειφθεί με την ανάπτυξη της τεχνολογίας των υπολογιστών και έτσι οι μέθοδοι μπορεί να θεωρηθεί ότι υπολογιστικά συμπεριφέρονται ανάλογα με τους αλγόριθμους των άλλων συστημάτων

Οι μέθοδοι κοντινότερου γείτονα έχουν κάποια πλεονεκτήματα για τις εφαρμογές του credit scoring. Ίσως η σημαντικότερη είναι αυτή που αναφέρθηκε παραπάνω, δηλαδή η δυνατότητα δυναμικής ανανέωσης των χρησιμοποιούμενων μοντέλων με τη συνεχή προσθήκη πελατών στο σχεδιασμό, όταν η πραγματική κατηγορία κινδύνου στην οποία ανήκουν γίνει γνωστή, και με την αφαίρεση των

παλαιότερων περιπτώσεων. Έτσι αντιμετωπίζεται και το πρόβλημα της πληθυσμιακής παρέκκλισης.

Οι μη παραμετρικές μέθοδοι smoothing, ειδικότερα οι μέθοδοι κοντινότερου γείτονα διερευνήθηκαν για εφαρμογές στο credit scoring από τους Chatterjee και Barcum (1970), οι οποίοι εξέτασαν αιτήσεις προσωπικών δανείων σε μία τράπεζα την Νέας Υόρκης και τις ταξινόμησαν ως προς τον πιστωτικό κίνδυνο που εμφάνιζαν με βάση το ποσοστό των περιπτώσεων με όμοια διανύσματα χαρακτηριστικών, οι οποίες ανήκαν στην ίδια κλάση. Τα αποτελέσματα για τις μη παραμετρικές μεθόδους ήταν ιδιαίτερα ενθαρρυντικά.

Τα πλεονεκτήματα και τα ικανοποιητικά αποτελέσματα από την εφαρμογή των μοντέλων κοντινότερου γείτονα είναι γενικά γνωστά, ωστόσο οι μη παραμετρικές , μέθοδοι δεν έχουν ευρέως διαδοθεί στη βιομηχανία του credit scoring. Ο λόγος είναι οι υπολογιστικές δυσκολίες που αντιμετωπίζουν, οι οποίες εμφανίζονται σημαντικές συγκριτικά με τις εναλλακτικές μεθόδους. Ένα πρόβλημα το οποίο έχει αρχίσει να βρίσκει την λύση του με την εξέλιξη της τεχνολογίας και του λογισμικού των ηλεκτρονικών υπολογιστικών.

Τα πλεονεκτήματα που συνοδεύουν την μέθοδο αυτή, όσον αφορά την εφαρμογή της, είναι τα εξής:

1. Η μη παραμετρική φύση της μεθόδου επιτρέπει τη μοντελοποίηση ανωμαλιών στο χώρο των χαρακτηριστικών.
2. Η μέθοδος έχει καλύτερη εφαρμογή από άλλες μη παραμετρικές τεχνικές, όπως οι μέθοδοι Kernel, όταν τα δεδομένα είναι πολυδιάστατα (Terrell-Scott, 1992).
3. Πρόκειται για μια διαισθητική διαδικασία, η οποία μπορεί εύκολα να εξηγηθεί στα διευθυντικά στελέχη που πρέπει να εγκρίνουν την υλοποίησή της και η οποία μπορεί να χρησιμοποιηθεί δυναμικά.

Μελέτες έχουν δείξει ότι η μέθοδος των κ-κοντινότερων γειτόνων μπορεί να δώσει μοντέλα με σταθερά καλή εφαρμογή για μεγάλο εύρος του δείκτη ‘κακών’ κινδύνων του πληθυσμού.

3.3.3. Μηχανική Εκμάθηση

Η βασική αιτία ανάπτυξης των συστημάτων Μηχανικής Εκμάθησης (Machine Learning) ήταν η ανάγκη για την εισαγωγή μεθόδων ταξινόμησης εύκολα κατανοητών και συμβατών με την ανθρώπινη κρίση. Οι αυτοματοποιημένες απαντήσεις των υπολογιστών, που αποτελούν το προϊόν των πολύπλοκων μαθηματικών τεχνικών που χρησιμοποιούν οι υπόλοιποι αλγόριθμοι εκμάθησης, δεν είναι εύκολο να αναλυθούν από τους χρήστες των συστημάτων. Το αποτέλεσμα είναι

οι λύσεις των συστημάτων σε προβλήματα ταξινόμησης να γίνονται δεκτές αδιάκριτα και χωρίς ανθρώπινη συμμετοχή στην πρόβλεψη, κάτι που μπορεί να οδηγήσει σε κακή αντιμετώπιση συγκεκριμένων προβλημάτων.

Η ιδέα των 'λογικών', εύκολα κατανοητών κανόνων αποφάσεων είναι ότι η ταξινόμηση ενός υποκειμένου θα πρέπει να βασιστεί σε μία καλή θεώρηση των χαρακτηριστικών του. Στην απλούστερη περίπτωση όλα τα χαρακτηριστικά του θα θεωρηθούν ως διχοτομικοί παράγοντες με τιμές 'σωστό' και 'λάθος'. Η πιο απλή μορφή κανόνα που μπορεί να προκύψει από μία τέτοια θεώρηση είναι η σύζευξη τέτοιων ενδείξεων.

Στην πραγματικότητα, για την κάλυψη μιας δεδομένης ομάδας απαιτούνται περισσότεροι από ένας κανόνες που μπορεί να βασίζονται σε συνθήκες είτε σύζευξης είτε διάζευξης ενδείξεων για διαφορετικά χαρακτηριστικά. Οι κανόνες αυτοί συνδυάζονται κατάλληλα για να καλύπτουν κάθε κατηγορία του προβλήματος και να διαχωρίζουν σωστά τα δειγματικά δεδομένα. Η πιο διαδεδομένη τεχνική για το σκοπό αυτό είναι η κατασκευή Δέντρων Αποφάσεων (Decision Trees).

3.3.3.1 Δέντρα Αποφάσεων (Decision Trees)

Η αρχική ανάπτυξη προγραμμάτων επαγωγής δέντρων αποφάσεων οφείλεται στους Morgan και Messemger (1973). Στην πραγματικότητα, τα δέντρα αποφάσεων αναπαριστούν κανόνες ταξινόμησης που προκύπτουν από μια οικογένεια αλγορίθμων. Οι κυριότεροι εκπρόσωποι αυτής της οικογένειας είναι ο αλγόριθμος CART (Breiman et al, 1984), ο αλγόριθμος ID3 (Quinlan, 1986) και ο C4.5 (Quinlan, 1993). Ακόμα, μέθοδοι επαγωγής δέντρων αποφάσεων μπορεί να προκύψουν από εμπειρικά δεδομένα και αυτές εμπεριέχονται στον γενικό όρο 'Αναδρομικός Διαχωρισμός' (Recursive Partitioning).

Ένα δέντρο αποφάσεων αποτελείται από κόμβους και κλαδιά και η διαδικασία κατασκευής του ξεκινά από έναν αρχικό κόμβο που ονομάζεται ρίζα του δέντρου. Κάθε κόμβος αντιπροσωπεύει μία απόφαση και στην απλούστερη περίπτωση ενός δυαδικού δέντρου οι δυνατές αποφάσεις είναι 'σωστό' και 'λάθος'. Κάθε κλαδί που φέρει την ένδειξη 'λάθος' καταλήγει σε τερματικό κόμβο, ενώ κάθε κλαδί που φέρει την ένδειξη 'σωστό' οδηγεί σε ένα κόμβο νέας απόφασης. Κάθε απόφαση συνδέεται με ένα συγκεκριμένο χαρακτηριστικό και η διαδικασία ανάπτυξης του δέντρου τελειώνει στον κόμβο όπου καταλήγει το κλαδί με την ένδειξη 'σωστό' για το τελευταίο χαρακτηριστικό. Οι τερματικοί κόμβοι αντιστοιχούν στις προβλεπόμενες από τον αλγόριθμο κλάσεις για την ταξινόμηση των νέων περιπτώσεων.

Εντελώς αναλογικά, ένα μη δυαδικό δέντρο αποφάσεων απλώς περιλαμβάνει κόμβους απ' όπου ξεκινάνε περισσότερα από δύο κλαδιά, τα οποία αντιπροσωπεύουν

τις τιμές του αντίστοιχου χαρακτηριστικού που σ' αυτή την περίπτωση είναι βέβαια, περισσότερες από δυο. Όταν τα κλαδιά αυτά καταλήγουν σε παραπάνω από ένα μη τερματικούς κόμβους, τότε η συγκεκριμένη απόφαση οδηγεί σε αντίστοιχο αριθμό εναλλακτικών σεναρίων ή αλλιώς μονοπατιών. Η εμπλοκή διαφορετικών σεναρίων στην απόφαση είναι ακριβώς και η ερμηνεία των διαζευκτικών κανόνων που μπορεί να περιέχονται στο σύστημα.

Η κατασκευή ενός δέντρου αποφάσεων καθορίζεται από την εισαγωγή κανόνων που βασίζονται σε ένα σύνολο δειγματικών δεδομένων S, σύμφωνα με τους οποίους όλοι οι κόμβοι διαμερίζονται σε ομάδες διακριτών επιλογών για την απόφαση που καθένας απ' αυτούς αντιπροσωπεύει. Ένας κόμβος καλείται τερματικός και δεν οδηγεί σε περαιτέρω διαχωρισμό των δεδομένων όταν όλες οι δειγματικές του μονάδες ανήκουν στην ίδια κλάση ή έστω αν οι συχνότητες εμφάνισης μιας συγκεκριμένης κλάσης ξεπερνά ένα κατώφλι.

Στο σημείο αυτό πρέπει να γίνει σαφές ότι τόσο οι ομάδες ταξινόμησης όσο και οι κανόνες είναι αμοιβαία αποκλειόμενοι. Δηλαδή, κάθε περίπτωση ανήκει σε μια και μόνο ομάδα και για κάθε περίπτωση ικανοποιείται ένας και μόνο ένας κανόνας.

Η σειρά με την οποία οι μεταβλητές (χαρακτηριστικά) θα εισαχθούν στην ανάπτυξη του δέντρου μπορεί να είναι τυχαία, όμως αυτό συνήθως οδηγεί στην κατασκευή μεγάλων δέντρων, αφού οι περισσότερες διαθέσιμες μεταβλητές είναι αρκετά 'θορυβώδεις'. Στην πράξη, επιδιώκεται η κατασκευή όσο το δυνατόν μικρότερων δέντρων, καθώς αυτά παρουσιάζουν τα εξής πλεονεκτήματα:

- Είναι ευκολότερα αντιληπτά από τον άνθρωπο
- Αποτυπώνουν καλύτερα τις γενικές αρχές που διέπουν το γνωστικό αντικείμενο.
- Ταξινομούν καλύτερα άγνωστες περιπτώσεις.

Γι' αυτούς τους λόγους είναι ιδιαίτερα σημαντική η επιλογή μίας μεταβλητής με καλή προβλεπτική ικανότητα κάθε φορά που ένας καινούργιος κόμβος πρόκειται να προστεθεί στο υπό κατασκευή δέντρο.

Οι μέθοδοι του αναδρομικού διαχωρισμού, ή απλά δέντρα αποφάσεων, αναπτύχθηκαν στο πεδίο της τεχνικής νοημοσύνης και χρησιμοποιούνται σε εφαρμογές και στη στατιστική. Στα πλαίσια του credit scoring, σε κάθε κόμβο ενός δέντρου αποφάσεων τοποθετείται ένα χαρακτηριστικό των υποψηφίων, στα κλαδιά του οι πιθανές τιμές των χαρακτηριστικών, ενώ οι τερματικοί κόμβοι αντιστοιχούν στις προβλεπόμενες από τον αλγόριθμο κατηγορίες κινδύνου για την ταξινόμηση των υποψηφίων.

Τα δέντρα αποφάσεων εμφανίζονται και σε άλλες μεθόδους κατασκευής συστημάτων διαχείρισης πιστωτικού κινδύνου. Ένα μειωτικό δέντρο είναι ένα μικρό δέντρο ταξινόμησης, το οποίο μπορεί να βοηθήσει στην αναγνώριση των αιτούντων με χαμηλό κίνδυνο και μ' αυτή την υπόθεση περιλαμβάνεται σε ένα μοντέλο credit

scoring ως ένα πρόσθετο χαρακτηριστικό. Έτσι, μη γραμμικές συσχετίσεις και αλληλεπιδράσεις ανάμεσα στις μεταβλητές μπορούν να συμπεριληφθούν σ' αυτό που επιφανειακά φαίνεται να είναι ένα γραμμικός συνδυασμός των χαρακτηριστικών (Hand-Henley,1997).

Σύμφωνα με τη φύση του προβλήματος του credit scoring στόχος του δέντρου αποφάσεων είναι ο διαχωρισμός του διαθέσιμου δείγματος εκμάθησης σε δύο υποσύνολα, καθένα από τα οποία να περιέχει πιστούχους από την ίδια κατηγορία πιστωτικού κινδύνου ('ενήμερος' ή 'σε καθυστέρηση').

Στο σημείο αυτό πρέπει να αναφερθεί ότι το σύνηθες επιχείρημα σε όφελος της κατασκευής δέντρων αποφάσεων για την επίλυση προβλημάτων ταξινόμησης είναι ότι αποτελούν συστήματα εύκολα αντιληπτά. Αυτή η ιδιότητα είναι ιδιαίτερα σημαντική στο ζήτημα της απόφασης παροχής πίστωσης, καθώς οι απαντήσεις που παρέχει το χρησιμοποιούμενο σύστημα για την αποδοχή ή την απόρριψη των αιτήσεων πρέπει να μπορούν να επεξηγηθούν στους υποψηφίους και μάλιστα με τρόπο συμβατό με την ανθρώπινη λογική. Επιπλέον, το σύστημα οφείλει να είναι κατανοητό στον ίδιο το χρήστη, ειδικά σε μία διαδικασία όπως η διαχείριση πιστωτικών κινδύνων που η λειτουργία της δεν μπορεί να είναι ανεξάρτητη από τον ανθρώπινο παράγοντα. Έτσι, λοιπόν, σαφές ότι εύκολα κατανοητοί κανόνες αποφάσεων, όπως είναι τα επαγόμενα δέντρα, είναι ευπρόσδεκτοι στο χώρο του credit scoring, με δεδομένη την καλή απόδοσή τους.

3.3.3.2. Νευρωνικά Δίκτυα

3.3.3.2.1. Συστήματα μίας εξόδου

Οι περισσότεροι κανόνες ταξινόμησης που βασίζονται στα Νευρωνικά Δίκτυα είναι μη παραμετρικοί, παρόλο που αρκετοί χρησιμοποιούν γραμμικές συναρτήσεις. Ο απλούστερος μηχανισμός νευρωνικού δικτύου είναι το σύστημα μίας εξόδου (single output perceptron), το οποίο οδηγεί σε απόφαση για την κατάταξη σε μία από δύο ομάδες ενός προτύπου που εισάγεται στο σύστημα.

Σε αναλογία με τη γραμμική διακριτική συνάρτηση, ο χρησιμοποιούμενος κανόνας ταξινόμησης βασίζεται σε ένα γραμμικό συνδυασμό των χαρακτηριστικών A_i που θεωρούνται οι είσοδοι (inputs) του συστήματος, με σταθμίσεις w_i:

$$\sum_i w_i A_i + q \text{ , όπου } q, w_i \text{ σταθερές}$$

Η σταθερά θ αντιστοιχεί στο σημείο που η γραμμική συνάρτηση τέμνει τον οριζόντιο άξονα.

Με βάση την παραπάνω γραμμική συνάρτηση και έναν καθορισμένο αριθμό χαρακτηριστικών A_i για ένα συγκεκριμένο πρόβλημα το σύστημα μιας εξόδου παράγει απόφαση (D) για την τοποθέτηση μιας παρατήρησης στην ομάδα 1 ή την ομάδα 0 σύμφωνα με τον κανόνα :

$$D = \begin{cases} 1, an \ \sum_i w_i A_i + q > 0 \\ 0, diaforetiká \end{cases}$$

Η σταθερά θ μπορεί να θεωρηθεί ως ένα κατώφλι για τον κανόνα, με την έννοια ότι η παρατήρηση τελικά θα ταξινομηθεί με βάση το αποτέλεσμα της σύγκρισης του αθροίσματος των σταθμισμένων τιμών των χαρακτηριστικών της με την τιμή του –θ.

3.3.3.2.2. Ο Μηχανισμός του Κανόνα Ταξινόμησης

Για τον προσδιορισμό των σταθμίσεων το σύστημα χρησιμοποιεί μια μη παραμετρική ακολουθιακή διαδικασία που στηρίζεται σε ένα δείγμα γνωστών περιπτώσεων. Αναλυτικότερα, οι δειγματικές παρατηρήσεις εισάγονται στο σύστημα σε σειρά και συμμετέχουν στην κατασκευή του κανόνα βελτιώνοντας τις τιμές για τις σταθερές q, w_i της συνάρτησης . Δηλαδή, αν η ταξινόμηση μιας παρατήρησης με βάση τον μέχρι στιγμής ισχύοντα κανόνα είναι λανθασμένη, τότε οι σταθμίσεις προσαρμόζονται κατάλληλα. Διαφορετικά, η μορφή της συνάρτησης παραμένει αμετάβλητη. Η παραπάνω διαδικασία ΄διόρθωσης των σφαλμάτων’ αποδίδεται στη συνέχεια με μαθηματικό συμβολισμό.

Οι νέες τιμές των σταθερών της συνάρτησης δίνονται από τις προηγούμενες με την προσθήκη ενός προσθετικού παράγοντα προσαρμογής, σύμφωνα με τους τύπους:

$$w_i(t+1) = w_i(t) + \Delta w_i(t)$$
$$q(t+1) = q(t) + \Delta q(t)$$

Συμβολίζοντας με Τ την πραγματική απάντηση στο πρόβλημα της ταξινόμησης και με D την απάντηση του μη προσαρμοσμένου κανόνα, οι παράγοντες προσαρμογής δίνονται αντίστοιχα ως:

$$\Delta w_i(t) = (T - D)A_i$$
$$\Delta q(t) = T - D$$

Δηλαδή, για κάθε σωστή πρόβλεψη του ισχύοντος συστήματος αυτό δεν μεταβάλλεται, ενώ για κάθε λανθασμένη τοποθέτηση στην ομάδα 0 ή 1, σε κάθε στάθμιση προστίθεται ή αφαιρείται αντίστοιχα η τιμή του συγκεκριμένου χαρακτηριστικού. Αναλογικά στο κατώφλι προστίθεται ή αφαιρείται αντίστοιχα μια μονάδα.

Είναι φανερό ότι η διαδικασία που περιγράφηκε παραπάνω είναι πολύ χρονοβόρα και ίσως ατέρμονη. Προκειμένου να επιταχυνθεί η προσέγγιση ενός

αποδεκτού κανόνα μπορούν να γίνουν κάποιες τροποποιήσεις των δεδομένων. Αυτό μπορεί να σημαίνει είτε κανονικοποίηση των δεδομένων, ώστε οι τιμές που εισάγονται στο σύστημα να μην εξαρτώνται από την κλιμάκωση των αντίστοιχων μεταβλητών, αλλά να βρίσκονται όλες στο διάστημα (0,1), είτε χρησιμοποίηση μιας διόρθωσης για παράγοντες προσαρμογής, ώστε οι αναθεωρήσεις των σταθμίσεων να είναι λιγότερο δραστικές.

3.3.3.2.3. Σύνθετα Συστήματα

Ένας συνθετότερος μηχανισμός νευρωνικών δικτύων θα είναι ένα σύστημα αποτελούμενο από πολλά απλά συστήματα μίας εξόδου. Πιο συγκεκριμένα, διαφορετικοί γραμμικοί συνδυασμοί των χαρακτηριστικών θα καταλήγουν σε διαφορετικές εξόδους ,οι οποίες με τη σειρά τους θα μπορούν να συνδυαστούν και να χρησιμοποιηθούν ως είσοδοι για νέα συστήματα, κ.ο.κ.

4. Διαδικασίας υλοποίησης.

4.1. 'Καθάρισμα' δεδομένων και σχεδιασμός δειγμάτων

4.1.1. Σχεδιασμός αρχικού δείγματος

Αρχίσαμε την υλοποίηση των συστήματων credit scoring με τη λήψη ενός δείγματος των προηγούμενων πελατών που υπέβαλαν αίτηση για ένα συγκεκριμένο τραπεζικό προϊόν, που στην συγκεκριμένη περίπτωση ήταν ένα συγκεκριμένο είδος πιστωτικής κάρτας με συγκεκριμένα χαρακτηριστικά, όσο το δυνατόν πιο κοντά χρονολογικά με όσο το δυνατόν αξιόπιστα στοιχεία τα οποία αφορούσαν την πιστωτική τους ιστορία καθώς και προσωπικά στοιχεία του υποψήφιου πελάτη. Αφαιρέσαμε από το δείγμα μας τις εγγραφές εκείνες οι οποίες είχαν αναξιόπιστα στοιχεία (data cleansing). Η επιλογή του συγκεκριμένου προϊόντός έγινε λόγο του μεγέθους και της ποιότητα των δεδομένων που μας δόθηκαν από την Τράπεζα.

Αφαιρέσαμε από τα δεδομένα μας παλιές εγγραφές καθώς επίσης και πολύ νέες έτσι ώστε να έχουμε όσο το δυνατόν περισσότερο αξιόπιστα δεδομένα. Συγκεκριμένα δεν συμπεριλάβαμε στο δείγμα μας δεδομένα πελατών που τους είχε χορηγηθεί πιστωτική κάρτα το τελευταίο εξάμηνο καθώς επίσης και στοιχεία πελατών που τους είχε χορηγηθεί κάρτα πριν τα τελευταία τρία χρόνια. Ο λόγος αυτού του φιλτραρίσματος στα δεδομένα μας είναι ότι δεν θέλαμε να συμπεριλάβουμε πιστωτικές κάρτες των οποίων οι δυνατότητες δεν έχουν χρησιμοποιηθεί σε ένα μεγάλο βαθμό καθώς επίσης δεν θέλαμε να χρησιμοποιήσουμε δεδομένα που έχουν προέλθει από διαφορετικό τρόπο χορήγησης, δεδομένου ότι ο τρόπος χορήγησης για την συγκεκριμένη πιστωτική κάρτα άλλαξε πριν 3 χρόνια και διατηρήθηκε ο ίδιος μέχρι την ημέρα που μας δόθηκαν τα δεδομένα. Μετά το πέρας της παραπάνω διαδικασίας είχαμε τελικά στην διάθεσή μας ένα δείγμα με τα παρακάτω χαρακτηριστικά:

	Population	Overdue	Normal
Absolute	**4700**	**967**	**3733**
Percentage	**100%**	**20.57%**	**79.43%**

Από τα χαρακτηριστικά που μας δόθηκαν για τον κάθε πελάτη επιλέξαμε αρχικά τα 20 σημαντικότερα και ποιο αξιόπιστα (ως προς την ποιότητα των δεδομένων) με την βοήθεια του υπευθύνου credit scoring της τράπεζας. Μετά από

διάφορες δοκιμές καταλήξαμε στα 9 πιο σημαντικά χαρακτηριστικά τα οποία παρουσιάζονται παρακάτω:

1. ***HasConnectedAccNo:*** Είναι μία διακριτή μεταβλητή η οποία παίρνει τιμές Yes/No. Και με την οποία παρουσιάζεται η πληροφορία αν για την συγκεκριμένη πιστωτική κάρτα υπάρχει συνδεδεμένος καταθετικός λογαριασμός.
2. ***HasAnotherAccNo:*** Είναι μία διακριτή μεταβλητή η οποία παίρνει τιμές Yes/No. Και με την οποία παρουσιάζεται η πληροφορία αν ο κάτοχος της πιστωτικής κάρτας έχει άλλον λογαριασμό με την τράπεζα.
3. ***HasAnotherProAccNo:*** Είναι μία διακριτή μεταβλητή η οποία παίρνει τιμές Yes/No. Και με την οποία παρουσιάζεται η πληροφορία αν ο κάτοχος της πιστωτικής κάρτας έχει άλλον λογαριασμό με την τράπεζα ο οποίος να έχει ποσά σε καθυστέρηση.
4. ***HasIDCard:*** Αν ο πελάτης μας έχει δώσει το αριθμό της αστυνομικής του ταυτότητας ή κάποιο άλλο αποδεικτικό στοιχείο. Είναι πάλι μια δια κριτή τυχαία μεταβλητή η οποία παίρνει τις τιμές Yes/No.
5. ***HasHomePhoneNumber:*** Με την μεταβλητή αυτή παρουσιάζεται η πληροφορία αν στην αίτηση ο πελάτης μας έχει δώσει το τηλέφωνο της οικίας του. Πρόκειται για μια διακριτή τυχαία μεταβλητή με τιμές Yes/No.
6. ***HasHomeWorkNumber:*** Με την μεταβλητή αυτή παρουσιάζεται η πληροφορία αν στην αίτηση ο πελάτης μας έχει δώσει το τηλέφωνο της εργασίας του. Πρόκειται για μια διακριτή τυχαία μεταβλητή με τιμές Yes/No.
7. ***StandingOrder:*** Στην μεταβλητή αυτή περιέχεται η πληροφορία αν η πιστωτική κάρτα έχει χρησιμοποιηθεί με κάποιο είδος πάγιας εντολής π.χ. για την πληρωμή ενός λογαριασμού τηλεφώνου. Είναι και αυτή μία διακριτή τυχαία μεταβλητή.
8. ***Gender:*** Πρόκειται για το φίλο του κατόχου της πιστωτικής κάρτας και παίρνει τρεις τιμές 1-Male , 2-Female , 3-Unknown.
9. ***ApprovedLimit:*** Πρόκειται ίσως ένα από τα σημαντικότερα χαρακτηριστικά στο οποίο απεικονίζεται το εγκριθέν πιστωτικό όριο που έχει εγκριθεί για τον πελάτη για την συγκεκριμένη πιστωτική κάρτα. Έχουμε χωρίσει τα όρια σε ranges και επομένως έχουμε μετατρέψει την συνεχή αυτή μεταβλητή σε διακριτή.
10. Η τελευταία μεταβλητή που έχουμε συμπεριλάβει στο δείγμα μας είναι και η ***μεταβλητή ενδιαφέροντος*** η οποία μπορεί να πάρει τις τιμές 0/1 (normal/overdue ή ‘ενήμερος’ / ‘σε καθυστέρηση’).

Αποφασίσαμε να ακολουθήσουμε την διαδικασία της Bootstrap με την εξής παραλλαγή, χρησιμοποιήσαμε ένα συγκεκριμένο δείγμα το οποίο προήλθε από το αρχικό πληθυσμό με τυχαία δειγματοληψία χωρίς επαναθέση ως σύνολο ελέγχου όλων των μοντέλων μας .Έτσι από τα validated δεδομένα δημιουργήσαμε ένα τυχαίο δείγμα μεγέθους 700 εγγραφών το οποίο χρησιμοποιήθηκε στην μελέτη μας ως σύνολο ελέγχου των μοντέλων μας. Η δομή του συγκεκριμένου συνόλου ήταν η εξής:

	Population	Overdue	Normal
Absolute	700	49	651
Percentage	100%	7%	93%

Το αμοιβαία αποκλειόμενο σύνολο που προέκυψε από την δημιουργία του συνόλου ελέγχου χρησιμοποιήθηκε για την κατασκευή των συνόλων κατάρτισης, των συνόλων δεδομένων επικύρωσης, των συνόλων δοκιμής. Το μέγεθος των συνόλων αυτών επιλέχθηκε σύμφωνα με το αποτέλεσμα της μελέτης των J.Galindo και P.Tamayo (1998), σύμφωνα με την οποία τα σύνολα αυτά θα πρέπει να έχουν τα εξής μεγέθη (ποσοστά ως προς το δείγμα) για να μπορεί το μοντέλο μας να έχει την βέλτιστη προσαρμογή:

Training Set (Σύν. Κατάρτισης)	Validation Set (Σύν. Επικύρωσης)	Testing Set (Συν. Ελέγχου)
50%	25%	25%

4.1.2. Δείγματα και Data Partitioning

Τα δείγματα, από το σύνολο που δημιουργήθηκε μετά από την αποκοπή του συνόλου ελέγχου, που χρησιμοποιήθηκαν για την εκμάθηση και την δημιουργία των τεσσάρων διαφορετικών μοντέλων σχεδιάστηκαν με δύο διαφορετικούς τρόπους. Η μία κατηγορία δειγμάτων σχεδιάστηκε με την υπόθεση ότι θέλαμε δείγματα με ίσα ποσοστά ‘ενήμερων’ και ‘σε καθυστέρηση’ περιπτώσεων ενώ τα δείγματα της άλλης κατηγορίας σχεδιάστηκαν με την υπόθεση ότι θέλαμε να διατηρήσουμε στα δείγματα τα ποσοστά των ‘ενήμερων’ και ‘σε καθυστέρηση’ όπως αυτά εμφανίζονταν στο σύνολο των δεδομένων. Σε ένα επόμενο βήμα δημιουργήσαμε για κάθε κατηγορία δύο υποκατηγορίες δειγμάτων η μία υποκατηγορία είχε μέγεθος 10% του συνόλου και η άλλη 40% του συνόλου. Σκοπός μας ήταν να εντοπίσουμε τον βέλτιστο τρόπο σχεδιασμού των δειγμάτων στα οποία θα στηρίζαμε τον σχεδιασμό των μοντέλων μας.

Μετά την δημιουργία των δειγμάτων ακολούθησε ο διαχωρισμός τους. Από τον διαχωρισμό αυτόν παρήχθησαν αμοιβαία αποκλειόμενα σύνολα δεδομένων: ένα

σύνολο δεδομένων κατάρτισης (Training Set), ένα σύνολο δεδομένων επικύρωσης (Validation Set), και ένα σύνολο δεδομένων δοκιμής (Testing Set) η χρήση των οποίων περιγράφεται παρακάτω.

Training Set: Το σύνολο δεδομένων κατάρτισης χρησιμοποιήθηκε για την εκπαίδευση, χτίσιμο του μοντέλου. Παραδείγματος χάριν, στην λογιστική παλινδρόμηση, το σύνολο δεδομένων κατάρτισης χρησιμοποιήθηκε για να εγκατασταθούν τα μοντέλα λογιστική παλινδρόμησης, π.χ. για τον υπολογισμό των coefficients. Στα νευρωνικά δίκτυα, το σύνολο δεδομένων κατάρτισης χρησιμοποιήθηκε για να υπολογιστούν τα βάρη των δικτύων.

Validation Set: Μετά την δημιουργία ενός μοντέλου από τα στοιχεία κατάρτισης, έπρεπε να ανακαλυφθεί η ακρίβεια του προτύπου σε στοιχεία τα οποία δεν είχαν χρησιμοποιηθεί για την κατάρτιση του. Για τον λόγο αυτό, το πρότυπο έπρεπε να χρησιμοποιηθεί σε ένα σύνολο δεδομένων που δεν είχε χρησιμοποιήθηκε στη διαδικασία κατάρτισης -- ένα σύνολο δεδομένων για το οποίο γνωρίζουμε την πραγματική αξία της μεταβλητής ενδιαφέροντος. Η απόκλιση μεταξύ της πραγματικής αξίας και της προβλεφθείσας αξίας της μεταβλητής ενδιαφέροντος είναι το λάθος της πρόβλεψη.

Εάν χρησιμοποιούσαμε τα στοιχεία κατάρτισης για τον υπολογισμό της ακρίβεια της πρότυπης ταξινόμησης, θα παίρναμε μια υπερβολικά αισιόδοξη εκτίμηση της ακρίβειας του μοντέλου. Αυτό οφείλεται στο ότι η κατάρτιση ή η πρότυπη διαδικασία συναρμολογήσεων εξασφαλίζει όσο το δυνατόν υψηλότερη ακρίβεια του προτύπου για τα στοιχεία κατάρτισης -- το πρότυπο σχεδιάζεται συγκεκριμένα για τα στοιχεία κατάρτισης. Για να έχουμε μια ρεαλιστικότερη εκτίμηση για το πώς το πρότυπο θα απέδιδε με άγνωστα στοιχεία, χρειάστηκε να αφήσουμε κατά μέρος ένα μέρος των αρχικών στοιχείων και να μην τα χρησιμοποιήσουμε στη διαδικασία κατάρτισης. Αυτό το σύνολο δεδομένων είναι γνωστό ως *σύνολο δεδομένων επικύρωσης*.

Testing Set: Το σύνολο δεδομένων επικύρωσης χρησιμοποιήθηκε για να καθορίσουμε με ακρίβεια τα μοντέλα. Παραδείγματος χάριν, να δοκιμάσαμε τα μοντέλα των νευρωνικών δικτύων με τις διάφορες αρχιτεκτονικές και εξετάσαμε την ακρίβεια για κάθε ένα στο σύνολο επικύρωσης και τελικώς επιλέξαμε κάποια μοντέλα μεταξύ των ανταγωνιστικών αρχιτεκτονικών. Σε αυτή την περίπτωση, όταν επιλέγεται δηλαδή ένα πρότυπο, η ακρίβειά της με το σύνολο επικύρωσης είναι ακόμα μια αισιόδοξη εκτίμηση όσον αφορά το πώς θα απέδιδε σε άγνωστα στοιχεία. Αυτό είναι επειδή το τελικό πρότυπο έχει επιλεχθεί ως το επικρατέστερο μεταξύ των ανταγωνιστικών προτύπων βασισμένων στο γεγονός ότι η ακρίβειά της στο σύνολο

επικύρωσης είναι η υψηλότερη. Κατά συνέπεια, χρειαζόμασταν ακόμα ένα σύνολο στοιχείων το οποίο δεν θα το χρησιμοποιούσαμε ούτε στην κατάρτιση αλλά ούτε και στην επικύρωση. Αυτό το σύνολο είναι γνωστό ως *σύνολο δεδομένων δοκιμής*. Η ακρίβεια του προτύπου στα στοιχεία δοκιμής δίνει μια περισσότερο ρεαλιστική εκτίμηση της απόδοσης του προτύπου στα απολύτως απαρατήρητα στοιχεία.

Όπως έχουμε αναφέρει στην μελέτη μας πέρα από το σύνολο δεδομένων δοκιμής χρησιμοποιήσαμε το σύνολο ελέγχου το οποίο μας έδωσε την δυνατότητα άμεσης σύγκρισης της ακρίβειας των μοντέλων, λόγω υπολογισμού της ακρίβειας ταξινόμησης με την χρήση των ίδιων δεδομένων.

4.2. Σχεδιασμός Υποδειγμάτων.

4.2.1. Λογιστική Παλινδρόμηση (Logistic Regression)

Στην Λογιστική παλινδρόμηση χρησιμοποιήσαμε το σύνολο κατάρτισης του κάθε δείγματος για τον σχεδιασμό των μοντέλων και τον υπολογισμό των coefficients της παλινδρόμησης. Κατά την διάρκεια του σχεδιασμού τον μοντέλων σε οχτώ διαφορετικά δείγματα (για κάθε μια από τις τέσσερις κατηγορίες δειγμάτων) λειτουργήσαμε ως εξής :

- Επιλέξαμε ως επίπεδο σημαντικότητας το 95%
- Δουλέψαμε για τον καταρτισμό του τελικού μοντέλου λογιστικής παλινδρόμησης βάση της μεθόδου του **backward selection**, δηλαδή μειώναμε τις μεταβλητές μία την φορά, αρχίζοντας από την ελάχιστα στατιστικά σημαντική και καταλήγαμε σε μοντέλα με μεταβλητές μόνο στατιστικά σημαντικές.
- Το μοντέλο που προέκυπτε από την προηγούμενη διαδικασία το εφαρμόζαμε στο σύνολο επικύρωσης και υπολογίζαμε την ακρίβειά του θέτοντας διαφορετικά σημεία διαχωρισμού (cut off points). Σημειώνουμε εδώ ότι από τα μοντέλα της λογιστικής παλινδρόμησης προκύπτουν πιθανότητες για το ενδεχόμενο κατά πόσο ο υποψήφιος πελάτης θα ανταποκριθεί ή όχι στις υποχρεώσεις του (για την συγκεκριμένη μελέτη), για την κάθε μία περίπτωση ξεχωριστά. Ορίζοντας σημείο διαχωρισμού επιλέγουμε από ποία τιμή και πάνω (ή κάτω) θα θεωρούμε ένα υποψήφιο πελάτη ‘ενήμερο’ (ή ‘σε καθυστέρηση’).
- Στο επόμενο στάδιο εφαρμόσαμε το μοντέλο στο σύνολο δοκιμής με τα σημεία διαχωρισμού που είχαν χρησιμοποιηθεί και το προηγούμενο στάδιο.
- Τέλος, εφαρμόσαμε το μοντέλο που είχε προκύψει στο σύνολο ελέγχου και υπολογίσαμε την ακρίβειά του.

Επαναλάβαμε την διαδικασία αυτή σε όλα τα τυχαία δείγματα που είχαμε δημιουργήσει και υπολογίσαμε την μέση τιμή και την διακύμανση της ακρίβειας (των μοντέλων) ως προς το σύνολο ελέγχου.

Παρακάτω παραθέτουμε τα αποτελέσματα που προέκυψαν από την εκτέλεση των προηγούμενων βημάτων για ένα δείγμα το οποίο έχει μέγεθος ίσο με το 40% του πληθυσμού(που έχει προκύψει αφού έχουμε αφαιρέσει το σύνολο ελέγχου) και τα ποσοστά των ‘ενήμερων’ / ‘σε καθυστέρηση’ πελατών είναι τα ίδια με του αρχικού πληθυσμού.

Στην πρώτη εκτέλεση της διαδικασίας σχηματίστηκε το εξής μοντέλο:

Data	
Source data worksheet	Data_Partition1
Training data used for building the model	Data_Partition1!$C$20:$M$819
Validation data	Data_Partition1!$C$820:$M$1219
# cases in the training data set	800
# cases in the validation data set	400

Variables									
Independent Variables	HasConnecte	HasAnotherA	HasAnotherPr	HasIDCard	HasHomePho	HasWorkPhon	StandingOrder	Gender	ApprovedL
Response Variable	Overdue								

The Regression Model

Predictor (indep. Var.)	Coefficient	Std. Error	p-value	Odds	CI	
HasConnectedAcct	0.52270788	0.29857917	0.0720426	1.68658863	0.95436214	2.98092175
HasAnotherAcc	-0.45226773	0.2540217	0.0718428	0.63301086	0.38475785	1.04144132
HasAnotherProAcc	2.78204083	0.55078864	0.00000044	16.1519508	5.4877305	47.5397759
HasIDCard	-0.26682481	0.19397241	0.16895129	0.76589721	0.52361232	1.1200295
HasHomePhoneNumber	0.12797414	0.33521897	0.70263731	1.13652364	0.58916855	2.19238782
HasWorkPhoneNumber	0.56534226	0.22318242	0.01130604	1.76005066	1.13646376	2.72682674
StandingOrder	-1.11258472	0.21617439	0.00000023	0.32870496	0.21560036	0.50114453
Gender	0.13009092	0.10777464	0.2274092	1.13993199	0.92206103	1.40881136
ApprovedL	0.00111731	0.00019191	0.00000001	1.00111794	1.00074148	1.00149453

Residual df	791
Std. Dev. Estimate	702.591064
% Success in training data	79.38
# Iterations	9
Residual SS	*

Variance-Covariance Matrix

	HasConnecte	HasAnother	HasAnoBorP	HasIDCard	HasHomePho	HasWorkPho	StandingOrd	Gender	ApprovedL
HasConnectedAcct	0.08443625	-0.02151302	0.08981541	-0.00361504	-0.00207241	-0.00083044	-0.04188534	-0.00051308	0.00000074
HasAnotherAcc	-0.02151302	0.06452702	-0.04910119	-0.00082705	0.00054114	-0.00162328	-0.00000181	0.00108917	0.00000571
HasAnotherProAcc	0.08981541	-0.04910119	0.30336815	-0.00323736	-0.0123762	-0.01356684	-0.09239198	-0.01971855	-0.00002731
HasIDCard	-0.00361504	-0.00082705	-0.00323736	0.03762529	-0.00558863	-0.00371987	-0.00223152	-0.00185102	-0.00000306
HasHomePhoneNumber	-0.00207241	0.00054114	-0.0123762	-0.00558863	0.11237177	0.00048364	0.00154897	0.00229079	-0.00000318
HasWorkPhoneNumber	-0.00083044	-0.00162328	-0.01356684	-0.00371987	0.00048364	0.04981039	0.00269359	-0.00190274	0.00000255
StandingOrder	-0.04188534	-0.00000181	-0.09239198	-0.00223152	0.00154897	0.00269359	0.04630002	-0.00072307	-0.00000322
Gender	-0.00051308	0.00108917	-0.01971855	-0.00185102	0.00229079	-0.00190274	-0.00072307	0.01161542	0.00000049
ApprovedL	0.00000074	0.00000571	-0.00002731	-0.00000306	-0.00000318	0.00000255	-0.00000322	0.00000049	0.00000004

Collinearity Diagnostics

Components	1	2
Eigenvalues	0.01632951	0.08306301
Condition numbers	18.8875408	8.37448597
HasConnectedAcct	0.45346216	0.01620089
HasAnotherAcc	0.04791016	0.35333708
HasAnotherProAcc	0.90565801	0.08716189
HasIDCard	0.00011844	0.0016046
HasHomePhoneNumber	0.00237026	0.00248196
HasWorkPhoneNumber	0.00760537	0.00036933
StandingOrder	0.87016926	0.12706347
Gender	0.00169662	0.36078862
ApprovedL	0.01153682	0.23787702

Χρησιμοποιώντας την μεθόδου του **backward selection** καταλήξαμε στο μοντέλο:

Data	
Source data worksheet	Data_Partition1
Training data used for building the model	Data_Partition1!$C$20:$M$819
Validation data	Data_Partition1!$C$820:$M$1219
New data	Final Test Sample!$B$2:$E$701
# cases in the training data set	800
# cases in the validation data set	400
# cases in the new data set	700

Variables				
Independent Variables	HasAnotherPr	HasWorkPhon	StandingOrder	ApproverL
Response Variable	Overdue			

The Regression Model

Predictor (Indep. Var.)	Coefficient	Std. Error	p-value	Odds	CI	
HasAnotherProAcc	2.33168364	0.4011803	0.00000001	10.2953504	4.68977647	22.6007385
HasWorkPhoneNumber	0.55921835	0.21982339	0.01096086	1.74930465	1.13692636	2.69140744
StandingOrder	-0.94617331	0.14881898	0	0.3882238	0.29012099	0.51949864
ApproverL	0.00114195	0.00019115	0	1.00114262	1.00076747	1.00151777

Residual df	796
Std. Dev. Estimate	710.696315
% Success in training data	79.38
# Iterations	9
Residual SS	*

Variance-Covariance Matrix

	HasAnotherP	HasWorkPho	StandingOrd	ApproverL
HasAnotherProAcc	0.16094592	-0.01633297	-0.0522709	-0.00002644
HasWorkPhoneNumber	-0.01633297	0.04832232	0.00124127	0.0000023
StandingOrder	-0.0522709	0.00124127	0.02208701	-0.00000177
ApproverL	-0.00002644	0.0000023	-0.00000177	0.00000004

Collinearity Diagnostics

Components	1	2
Eigenvalues	0.03117575	0.1658413
Condition numbers	9.91751992	4.06153631
HasAnotherProAcc	0.95686072	0.0357853
HasWorkPhoneNumber	0.01630136	0.0279602
StandingOrder	0.9166086	0.07428888
ApproverL	0.05454078	0.9164536

Το μοντέλο που προέκυψε από την προηγούμενη διαδικασία:

$$\log\left(\frac{p}{1-p}\right) = 2{,}33168X_1 + 0{,}55921X_2 - 0{,}94617X_3 + 0{,}00114X_4$$

$$X_1 = \text{HasAnotherProAcc}$$

$$X_2 = \text{HasWorkPhoneNumber}$$

$$X_3 = \text{StandingOrder}$$

$$X_4 = \text{ApproverL}$$

(όπου p η πιθανότητα το Y να πάρει την τιμή 1)

το εφαρμόσαμε στο σύνολο επικύρωσης και υπολογίσαμε την ακρίβειά του θέτοντας διαφορετικά σημεία διαχωρισμού. Στο επόμενο στάδιο εφαρμόσαμε το μοντέλο στο σύνολο δοκιμής με τα σημεία διαχωρισμού που είχαν χρησιμοποιηθεί και το

προηγούμενο στάδιο. Τέλος, εφαρμόσαμε το μοντέλο που προέκυψε στο σύνολο ελέγχου και υπολογίσαμε την ακρίβειά του.

Cutoff points	0,5	0,6	0,7	0,8	0,9
Misclassification Error	8.43%	13.71%	24.43%	42.14%	72.14%

Παρατηρούμε ότι όσο αυξάνεται το σημείο διαχωρισμού τόσο αυξάνει το σφάλμα ταξινόμησης.

4.2.2. Ταξινόμηση με μεθόδους Κοντινότερου Γείτονα

Η μέθοδος αυτή είναι μια τυποποιημένη μη παραμετρική προσέγγιση στο πρόβλημα ταξινόμησης (που προτάθηκε για πρώτη φορά από τους Fix και Hodges (1952)). Η ιδέα είναι να επιλεχτεί ένας μετρικός χώρος στο διάστημα των στοιχείων της αίτησης για να μετρήσει πόσο μακριά δύο οποιοιδήποτε υποψήφιοι είναι. Κατόπιν με την χρήση ενός δείγματος προηγούμενων αντιπροσωπευτικών υποψηφίων, ένας νέος υποψήφιος ταξινομείται ως ‘ενήμερος’ ή ‘σε καθυστέρηση’ ανάλογα με τα ποσοστά των ‘ενήμερων’ και ‘σε καθυστέρηση’ μεταξύ των K-κοντινότερων - υποψήφιων από το αντιπροσωπευτικό δείγμα .

Οι τρεις παράμετροι που απαιτούνται για την προσέγγιση αυτή είναι οι μετρικοί χώροι, πόσοι υποψήφιοι K αποτελούν το σύνολο κοντινότερων γειτόνων, και ποιο ποσοστό αυτών πρέπει να είναι ‘καλοί’ για να ταξινομηθεί ο υποψήφιος ως ‘ενήμερος’.

Στην μέθοδο των K–κοντινότερων-γειτόνων, το σύνολο των στοιχείων κατάρτισης χρησιμοποιείται για να προβλέψει την αξία μιας μεταβλητής ενδιαφέροντος για κάθε μέλος ενός συνόλου στοιχείων "στόχων". Στην δομή των στοιχείων πρέπει να υπάρχει μια μεταβλητή ενδιαφέροντος (π.χ. "αποδοχή πίστωσης"), και διάφορες πρόσθετες μεταβλητές που θα χρησιμοποιηθούν για την πρόβλεψη της μεταβλητής ενδιαφέροντος (π.χ. ηλικία, εισόδημα, θέση...). Γενικά, ο αλγόριθμος είναι ο ακόλουθος:

1. Για κάθε σειρά (περίπτωση) στο σύνολο των δεδομένων που πρέπει να προβλεφθεί η συμπεριφορά τους, εντοπίζει τα K ποιο κοντινά μέλη (οι κοντινότεροι γείτονες K) του συνόλου στοιχείων κατάρτισης. Ένα μέτρο ευκλείδειας απόστασης χρησιμοποιείται για να ορίσουμε το πόσο κοντά κάθε μέλος του συνόλου κατάρτισης είναι στη σειρά στόχων που εξετάζεται.
2. Υπολογίζεται το σταθμισμένο ποσό της μεταβλητής ενδιαφέρον για τους K κοντινότερους γείτονες (τα βάρη είναι το αντίστροφο των αποστάσεων).

3. Επαναλάβετε η ίδια διαδικασία για τις υπόλοιπες σειρές (περιπτώσεις) στο σύνολο στόχων.

Φυσικά ο χρόνος υπολογισμού αυξάνει καθώς το K αυξάνει, αλλά το πλεονέκτημα είναι ότι οι υψηλότερες τιμές του K παρέχουν τη λείανση που μειώνει την ευπάθεια στο θόρυβο στα στοιχεία κατάρτισης. Μετά από πολλές δοκιμές με διάφορα δείγματα (διαφορετικής σύνθεσης, διαφορετικού μεγέθους,…) παρατηρήσαμε ότι μοντέλα με K =15 είχαν αρκετή καλή προβλεπτική ικανότητα και για τον λόγο αυτό επιλέξαμε το K = 15 για την μελέτη μας. Φυσικά δεν υπάρχει κανόνας ο οποίος να μας ορίζει το K, έτσι είμαστε ελεύθεροι να επιλέξουμε μόνοι μας το μέγεθος του K, μία ελευθερία η οποία όμως εγκυμονεί κινδύνους π.χ. αν το K είναι πάρα πολύ μικρό, η ταξινόμηση μιας περίπτωσης (σειρά) θα μεταβάλετε αρκετά αφού εξαρτάται από την ταξινόμηση της περίπτωσης που είναι ποιο κοντά, ένα μεγαλύτερο K θα μείωνε αυτήν την μεταβλητότητα, αλλά ένα μεγάλο K θα είχε ως αποτέλεσμα την εισαγωγή μεροληψίας στην απόφαση ταξινόμησης. Φανταστείτε όπου K = «το ολόκληρο σύνολο στοιχείων», σε αυτήν την εκφυλισμένη περίπτωση, όλες οι περιπτώσεις θα ταξινομηθούν στην κατηγορία της πλειοψηφίας.

Κατά τον σχεδιασμό των μοντέλων ακολουθήσαμε τα εξής βήματα:

- Επεξεργαστήκαμε τα δεδομένα του συνόλου κατάρτισης, συγκεκριμένα τα κανονικοποιήσαμε. Η κανονικοποίηση των δεδομένων είναι απαραίτητη έτσι ώστε να εξασφαλίσουμε ότι το μέτρο απόστασης κατανέμει το ίδιο βάρος σε κάθε μεταβλητή -- χωρίς κανονικοποίηση, η μεταβλητή με τη μεγαλύτερη κλίμακα θα εξουσίαζε στο μέτρο.
- Επιλέξαμε ως K το 15 και καταρτίσαμε το μοντέλο των 15-κοντινότερων γειτόνων.
- Εφαρμόσαμε το μοντέλο στα σύνολα επικύρωσης και δοκιμής.
- Και στο τελευταίο στάδιο εφαρμόσαμε το μοντέλο στο σύνολο ελέγχου. Παρατηρήθηκε ότι κάποιες περιπτώσεις αποτύγχανε ο αλγόριθμος να της ταξινομήσει λόγω ότι δεν υπήρχαν στο σύνολο κατάρτισης περιπτώσεις με αντίστοιχα χαρακτηριστικά. Γι αυτές τις περιπτώσεις κάναμε την παραδοχή και τις ταξινομήσαμε ως ‘καλούς’ (Normal) πελάτες.

Επαναλάβαμε την διαδικασία αυτή σε όλα τα τυχαία δείγματα που είχαμε δημιουργήσει και υπολογίσαμε την μέση τιμή και την διακύμανση της ακρίβειας (των μοντέλων) ως προς το σύνολο ελέγχου.

Χρησιμοποιώντας τα δεδομένα του δείγματος που χρησιμοποιήθηκε και στην προηγούμενη μέθοδο, παραθέτουμε παρακάτω τα αποτελέσματα όπως αυτά προέκυψαν από την εφαρμογή της μεθόδου των κ κοντινότερων γειτόνων.

Data	
Source data worksheet	Data_Partition1
Training data used for building the model	Data_Partition1!$C$20:$N$819
Validation data	Data_Partition1!$C$820:$M$1219
# cases in the training data set	800
# cases in the validation data set	400
Normalization	TRUE
# nearest neighbors (k)	15

Variables									
Input variables	HasConnecte	HasAnotherA	HasAnotherPr	HasIDCard	HasHomePho	HasWorkPhon	StandingOrder	Gender	Approved_
Output variable	Overdue								

Output Classes

Classes
0
1

Validation Misclassification Summary

Classification confusion matrix		
Actual class	Computed class	
	0	1
0	6	80
1	14	298

Error Report			
Class	# Cases	# Errors	% Error
0	86	80	93.91
1	312	14	4.49
Overall	400	94	23.50

Από την εφαρμογή του μοντέλου στο σύνολο επικύρωσης έχουμε (ένα μικρό δείγμα της ταξινόμησης παραθέτετε στο επόμενο σχέδιο):

K-Nearest Neighbors - Classification of Validation Data

Predicted Class	% Predicted class in the nearest neighbors polled	Actual number of nearest neighbors	Actual Class	HasConnecte	HasAnotherA	HasAnotherP	HasIDCard	HasHomePho	HasWorkPhon	StandingOrd	Gender	Approved_
1	93.33	15	1	0	0	1	1	0	0	1	2	1000
1	70.59	17	1	1	1	1	0	0	0	3	1	900
1	93.33	15	1	0	1	1	1	1	1	2	1	2700
1	86.67	15	1	0	1	1	0	0	1	1	1	500
1	86.67	15	1	0	1	1	0	0	1	2	3	600
1	77.78	18	0	0	1	1	1	0	0	3	1	1500
1	86.67	15	1	0	0	1	1	0	1	1	1	1000
1	93.33	15	1	0	1	1	0	0	0	1	2	2000
1	93.33	15	1	0	0	1	0	0	0	2	2	3000
1	87.5	16	0	1	1	1	0	0	1	3	3	750
1	69.23	26	1	1	1	1	0	0	0	3	1	750
1	100	15	1	0	0	1	0	0	1	1	1	500
1	60	15	1	0	1	1	1	0	0	2	2	750
1	100	15	1	0	1	1	0	0	1	1	3	900
1	93.33	15	1	0	0	1	0	0	1	2	1	700
1	81.25	16	1	1	1	1	0	0	1	3	1	1000
1	62.5	16	1	0	1	1	0	0	0	2	2	500
1	60	15	1	0	1	1	1	0	1	2	2	900
1	81.25	16	1	0	1	1	0	0	0	3	1	750
1	100	15	1	0	0	1	1	0	0	2	2	2000
1	86.67	15	1	1	1	1	1	0	0	3	1	3500
1	93.33	15	1	1	1	1	1	0	0	3	3	2900

Εφαρμόζοντας το μοντέλο των κ κοντινότερων γειτόνων στο σύνολο ελέγχου είχαμε ως αποτέλεσμα ένα σφάλμα ταξινόμησης της τάξης του 16,56%.

4.2.3. Δέντρα Αποφάσεων CART (CART Decision Trees)

Τα δέντρα ταξινόμησης (Classification tree μία κατηγορία των decision trees) είναι μια καλή επιλογή μεθόδου όταν ο σκοπός του μοντέλου είναι η ταξινόμηση ή η πρόβλεψη των εκβάσεων και ο στόχος είναι να παραχθούν οι κανόνες που να μπορούν να γίνουν εύκολα κατανοητοί, να εξηγηθούν, ή να μεταφραστούν σε φυσική γλώσσα διατύπωσης ερωτήσεων.

Τα δέντρα ταξινόμησης καταγράφουν τις τιμές των χαρακτηριστικών των δεδομένων και τα κατατάσσουν σε διακριτές κατηγορίες. Τα δέντρα ταξινόμησης μπορούν επίσης να παρέχουν το μέτρο της εμπιστοσύνης ότι η ταξινόμηση είναι σωστή.

4.2.3.1. Μεθοδολογία

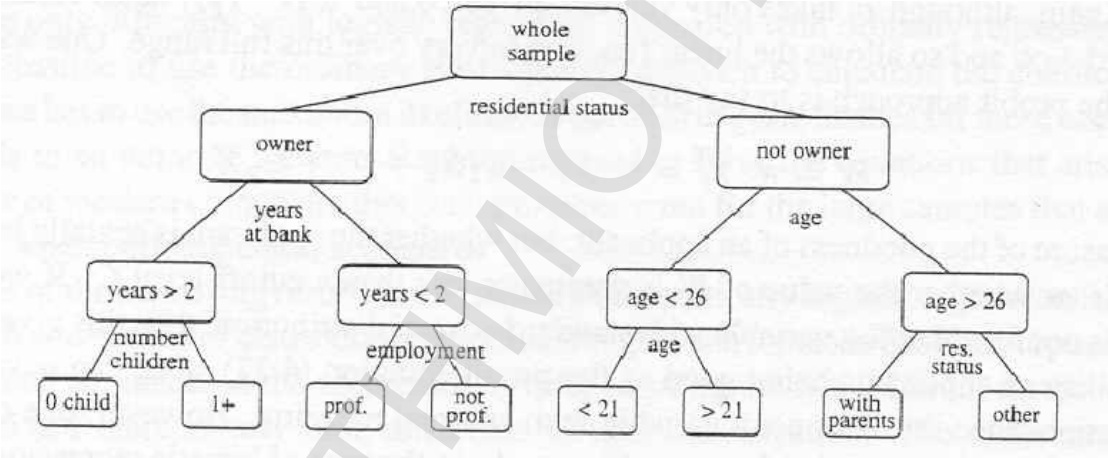
Το δέντρο ταξινόμησης (Classification tree) χτίζεται μέσω μιας διαδικασίας γνωστής ως «δυαδικός αναδρομικός διαχωρισμός». Αυτή είναι μια επαναληπτική διαδικασία κατά την οποία τα δεδομένα χωρίζονται σε υποσύνολα, και έπειτα χωρίζονται περαιτέρω σε κάθε ένα από τα κλαδιά.

Αρχικά, η διαδικασία αρχίζει με ένα σύνολο κατάρτισης στο οποίο η ταξινόμηση είναι γνωστή (προ-ταξινομημένος) για κάθε αρχείο. Όλα τα αρχεία στο σύνολο κατάρτισης είναι μαζί σε ένα ‘μεγάλο κουτί’. Ο αλγόριθμος έπειτα συστηματικά προσπαθεί να χωρίσει τα δεδομένα σε δύο μέρη, επεξεργάζεται μια μεταβλητή τη φορά χρησιμοποιώντας ως βάση μια διαχωριστική γραμμή για την συγκεκριμένη μεταβλητή. Ο σκοπός είναι να επιτευχθεί το όσο το δυνατόν περισσότερο ομοιογενές σύνολο ετικετών σε κάθε χώρισμα (partition). Αυτός ο διαχωρισμός εφαρμόζεται έπειτα σε κάθε ένα από τα νέα χωρίσματα. Η διαδικασία συνεχίζεται έως ότου εμφανιστούν μη χρήσιμες διασπάσεις. Η καρδιά του αλγορίθμου είναι ο κανόνας που καθορίζει τον αρχικό κανόνα διάσπασης.

4.2.3.2. Εύρεση του σημείου αρχικής διάσπασης

Η διαδικασία αρχίζει με ένα σύνολο κατάρτισης που αποτελείται από τα προ-ταξινομημένα αρχεία. Προ-ταξινομημένος σημαίνει ότι ο στόχος, ή η εξαρτώμενη

μεταβλητή, έχει μια γνωστή κατηγορία ή ετικέτα. Ο στόχος είναι να χτιστεί ένα δέντρο που να διακρίνει τις κατηγορίες . Στην περίπτωσή μας, η οποία είναι και η ποιο απλή, έχουμε μόνο δύο κατηγορίες στόχων και έτσι κάθε διάσπαση είναι ένας δυαδικός χωρισμός. Για να επιλέξει τον καλύτερο σημείο διαχωρισμού σε έναν κόμβο, ο αλγόριθμος ασχολείται με ένα εισαγόμενο πεδίο κάθε φορά. Στην ουσία, κάθε πεδίο ταξινομείται. Κατόπιν, κάθε πιθανή διάσπαση δοκιμάζεται και εξετάζεται, και το καλύτερο για διάσπαση είναι αυτό που παράγει τη μεγαλύτερη μείωση στην ποικιλομορφία της ετικέτας ταξινόμησης μέσα σε κάθε χώρισμα (αυτό είναι ακριβώς ένας άλλος τρόπος να περιγράψουμε "την αύξηση στην ομοιογένεια"). Αυτό επαναλαμβάνεται για όλα τα πεδία, και ο «νικητής» επιλέγεται ως το καλύτερο σημείο διαχωρισμού για εκείνο τον κόμβο. Η διαδικασία συνεχίζεται στον επόμενο κόμβο και, με αυτόν τον τρόπο, παράγεται ένα πλήρες δέντρο.



4.2.3.3. Περιορισμός του Classification tree

Ο περιορισμός του δέντρου ταξινόμησης είναι η διαδικασία κατά την οποία αφαιρούνται φύλλα και κλαδιά από το δέντρο αποφάσεων με σκοπό την βελτίωση της απόδοσης του δέντρου όταν χρησιμοποιηθούν σε πραγματικές εφαρμογές. Ο αλγόριθμος οικοδόμησης του δέντρου καθιστά το καλύτερο διαχωρισμό στον κόμβο ρίζας εκεί όπου υπάρχει ο μεγαλύτερος αριθμός στοιχείων και, ως εκ τούτου, πολλές πληροφορίες. Κάθε επόμενη διάσπαση έχει έναν μικρότερο και λιγότερο αντιπροσωπευτικό πληθυσμό με τον οποίο να συνεργαστεί. Προς το τέλος, η ιδιοσυγκρασία των στοιχείων κατάρτισης σε έναν ιδιαίτερο κόμβο επιδεικνύει τα χαρακτηριστικά που είναι ιδιαίτερα μόνο σε συγκεκριμένα στοιχεία. Αυτά τα χαρακτηριστικά μπορεί να μην έχουν νόημα και μερικές φορές να είναι επιβλαβή για

την πρόβλεψη εάν προσπαθήσουμε να επεκτείνουμε τους κανόνες αυτούς σε μεγαλύτερους πληθυσμούς.

Οι μέθοδοι περιορισμού λύνουν αυτό το πρόβλημα. Συγκεκριμένα αφήνουν το δέντρο να αυξηθεί στο μέγιστο μέγεθος, κατόπιν αφαιρούν τους μικρότερους κλάδους που αποτυγχάνουν να γενικεύσουν.

4.2.3.4. Εφαρμογή της μεθοδολογίας

Κατά τον σχεδιασμό των μοντέλων ακολουθήσαμε τα εξής βήματα:

- Επεξεργαστήκαμε τα δεδομένα του συνόλου κατάρτισης, συγκεκριμένα τα κανονικοποιήσαμε.
- Επιτρέψαμε στο δέντρο απόφασης αν αναπτυχθεί σε όλο το εύρος του.
- Για να διορθώσουμε το over fitting από την δημιουργία του πλήρες δέντρου εφαρμόσαμε την διαδικασία του Tree Pruning κατά την οποία με περικοπή του δέντρου τακτοποιούνται ξανά τα κλαδιά του δέντρου και παράγεται ένα λιγότερο σύνθετο δέντρο. Το "ελάχιστο δέντρο λάθους" είναι το δέντρο που παράγει το ελάχιστο ποσοστό λάθους ταξινόμησης όταν εξετάζεται στα στοιχεία επικύρωσης. Το "Best Pruned δέντρο" έχει τον μικρότερο αριθμό κόμβων, το οποίο έχει προκύψει υπό την υπόθεση του περιορισμού του λάθους κάτω από ένα συγκεκριμένο επίπεδο (το ελάχιστο ποσοστό λάθους συν το τυποποιημένο λάθος για το ποσοστό λάθους).
- Εφαρμόσαμε το μοντέλο στα σύνολα επικύρωσης και δοκιμής.
- Και στο τελευταίο στάδιο εφαρμόσαμε το μοντέλο στο σύνολο ελέγχου.

Επαναλάβαμε την διαδικασία αυτή σε όλα τα τυχαία δείγματα που είχαμε δημιουργήσει και υπολογίσαμε την μέση τιμή και την διακύμανση της ακρίβειας (των μοντέλων) ως προς το σύνολο ελέγχου.

Χρησιμοποιώντας τα δεδομένα του δείγματος που χρησιμοποιήθηκε και στις δύο προηγούμενες μεθόδους, παραθέτουμε παρακάτω τα αποτελέσματα όπως αυτά προέκυψαν από την εφαρμογή της μεθόδου των δέντρων αποφάσεων (με τον αλγόριθμο CART).

Στο πρώτο σχεδιάγραμμα παρουσιάζουμε την μείωση του σφάλματος κατά την εκπαίδευση του δέντρου:

Data	
Source data worksheet	Data_Partition1
Training data used for building the model	Data_Partition1!$C$20:$M$819
Validation data	Data_Partition1!$C$820:$M$1219
New data	Test Sample!$B$2:$J$701
# cases in the training data set	800
# cases in the validation data set	400
# cases in the new data set	700
Normalized	TRUE

Variables									
Input variables	HasConnected	HasAnotherAd	HasAnotherPh	HasIDCard	HasHomePhon	HasWorkPhon	StandingOrder	Gender	Approved
Output variable	Overdue								

Output Classes

Classes
0
1

Training Log

Growing the Tree	
#Nodes	Error
0	20.63
1	9.87
2	7.26
3	4.86
4	4.68
5	4.13
6	3.51
7	3.04
21	1.33
22	1.32
23	1.31
24	1.21
25	1
26	0.96
27	0.91
28	0.9
29	0.86
30	0.86
31	0.86
32	0.86
33	0.79
34	0.76
35	0.76
36	0.7
37	0.69
40	0.66
41	0.64
42	0.59
43	0.56
44	0.56
45	0.56
46	0.54
47	0.53
48	0.53
49	0.53
55	0.52
56	0.51
57	0.5
58	0.5
59	0.49
60	0.49
61	0.48
62	0.48

Growing the Tree	
#Nodes	Error
63	0.48
64	0.45
65	0.45
66	0.42
67	0.42
68	0.41
69	0.4
70	0.39
71	0.37
72	0.37
73	0.37
74	0.36
75	0.36
76	0.36
77	0.35
83	0.35
84	0.34
85	0.34
89	0.32
90	0.31
94	0.3
95	0.29
96	0.29
102	0.27
103	0.26
104	0.25
108	0.23
111	0.22
118	0.21
122	0.2
132	0.19
139	0.18
140	0.17
141	0.17
148	0.16
156	0.15
163	0.14
170	0.13
181	0.12
188	0.11
191	0.1
196	0.09

Validation Misclassification Summary

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	0	0
1	88	312

Error Report			
Class	# Cases	# Errors	% Error
0	0	0	Undefined
1	400	88	22.00
Overall	400	88	22.00

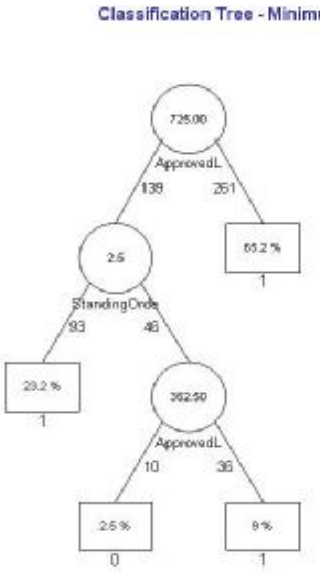
Στο επόμενο σχεδιάγραμμα εμφανίζουμε την εξέλιξη του σφάλματος κατά την διάρκεια του pruning.

Classification Tree - Prune Log

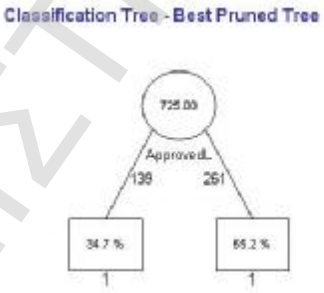
# Decision Nodes	Error
196	0.25999999
195	0.25999999
194	0.25
193	0.25
192	0.25
191	0.25
190	0.25
189	0.25
188	0.25
187	0.2475
186	0.2475
185	0.2475
184	0.2475
183	0.2475
182	0.2475
181	0.2475
180	0.245
179	0.245
178	0.245
177	0.245
176	0.245
175	0.2475
174	0.2475
173	0.2475
172	0.2475
171	0.2475
170	0.2475
169	0.2475
168	0.2475
167	0.2475
166	0.2475
165	0.2475
164	0.2475
163	0.2475
162	0.2475
161	0.2475
160	0.2475
159	0.2475
158	0.2475
157	0.2475
156	0.2475
155	0.24250001
154	0.24250001
153	0.24250001
152	0.24250001
151	0.24250001
150	0.23999999
149	0.23999999
148	0.23999999
147	0.23999999
146	0.23999999
145	0.24250001
144	0.24250001
143	0.245
142	0.245
141	0.245
140	0.245
139	0.245
138	0.245
137	0.245
136	0.245
135	0.245
134	0.245
133	0.245
132	0.245
131	0.245
130	0.245
129	0.245
128	0.245
127	0.245
126	0.23999999
125	0.23999999
124	0.23999999
123	0.23999999
122	0.23999999
121	0.2325
120	0.2325
119	0.2325
118	0.2325
117	0.22750001
116	0.22750001
115	0.22750001
114	0.22750001
113	0.22750001
112	0.22750001
111	0.22750001
110	0.22750001
109	0.22750001
108	0.22750001
107	0.22750001
106	0.22750001
105	0.22750001
104	0.22750001
103	0.22750001
102	0.22750001
101	0.22750001
100	0.22750001
99	0.22750001
98	0.22750001
97	0.22750001
96	0.23
95	0.23
94	0.23

# Decision Nodes	Error			
91	0.23			
90	0.22750001			
89	0.22750001			
88	0.22750001			
87	0.23			
86	0.23			
85	0.23			
84	0.23999999			
83	0.23999999			
82	0.23999999			
81	0.23999999			
80	0.23999999			
79	0.23999999			
78	0.2325			
77	0.2325			
76	0.2325			
75	0.2325			
74	0.2325			
73	0.23			
72	0.23			
71	0.23			
70	0.23			
69	0.23			
68	0.23			
67	0.23			
66	0.23			
65	0.23			
64	0.23			
63	0.22499999			
62	0.22499999			
61	0.22499999			
60	0.22499999			
59	0.22499999			
58	0.22499999			
57	0.22499999			
56	0.22499999			
55	0.22499999			
54	0.22499999			
53	0.22499999			
52	0.22499999			
51	0.22499999			
50	0.22499999			
49	0.2225			
48	0.2225			
47	0.2225			
46	0.2225			
45	0.2225			
44	0.2225			
43	0.22499999			
42	0.22499999			
41	0.22499999			
40	0.22499999			
39	0.22499999			
38	0.22499999			
37	0.22499999			
36	0.22499999			
35	0.22499999			
34	0.22499999			
33	0.22750001			
32	0.22750001			
31	0.22750001			
30	0.22750001			
29	0.2175			
28	0.2175			
27	0.2175			
26	0.2175			
25	0.2175			
24	0.2175			
23	0.2175			
22	0.22750001			
21	0.22750001			
20	0.22750001			
19	0.22			
18	0.22			
17	0.22			
16	0.22			
15	0.22			
14	0.22			
13	0.22			
12	0.22			
11	0.22			
10	0.22			
9	0.2225			
8	0.2225			
7	0.2225			
6	0.205			
5	0.205			
4	0.205			
3	0.205	<-- Minimum Error Prune	Std. Err	0.02018508
2	0.22			
1	0.22	<-- Best Prune		

Από την διαδικασία αυτή προκύπτουν δύο δέντρα αποφάσεων ένα το οποίο έχει το ελάχιστο σφάλμα,



και ένα το οποίο είναι best pruned.



Εφαρμόζοντας το δέντρο αποφάσεων που προέκυψε από την παραπάνω διαδικασία στο σύνολο ελέγχου είχαμε ως αποτέλεσμα ένα σφάλμα ταξινόμησης της τάξης του 19,71%.

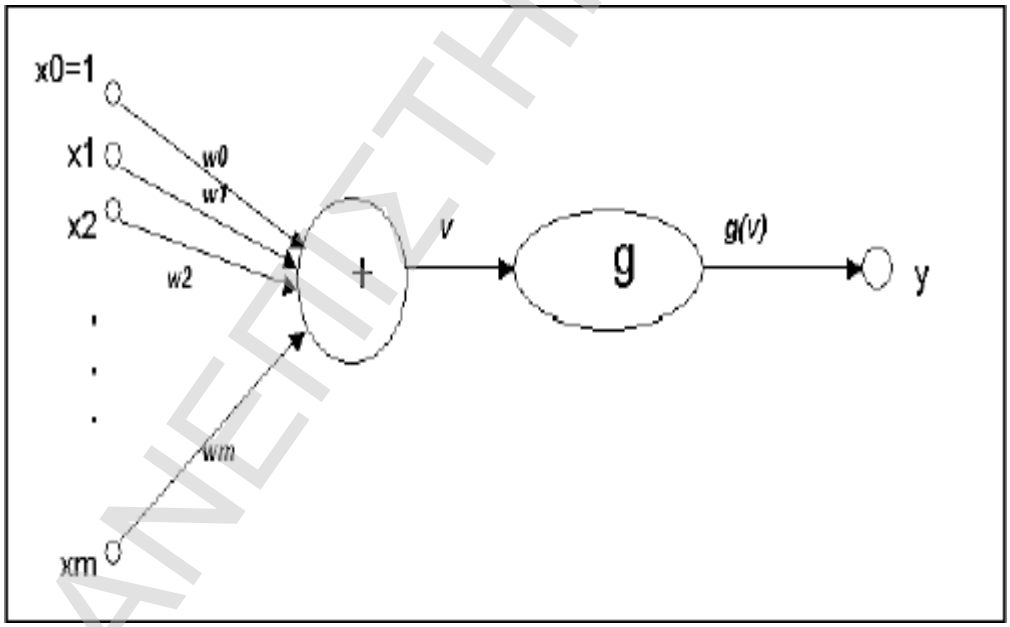
4.2.4. Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks) είναι σχετικά ακατέργαστα ηλεκτρονικά δίκτυα "νευρώνων" τα οποία στηρίζονται στη νευρική δομή του εγκεφάλου. Επεξεργάζονται τα στοιχεία ένα την φορά, και "μαθαίνουν" μέσω της σύγκρισης της ταξινόμησης η οποία έχει προέλθει από την διαδικασία εκμάθησης (που, στην αρχή, είναι κατά ένα μεγάλο μέρος αυθαίρετη) με τη γνωστή πραγματική ταξινόμηση του στοιχείου. Τα λάθη από την αρχική ταξινόμηση του πρώτου στοιχείου ανατροφοδοτούνται στο δίκτυο, και χρησιμοποιούνται για να τροποποιηθεί ο αλγόριθμος του δικτύου τη δεύτερη φορά, κάτι το οποίο επαναλαμβάνεται αρκετές φορές.

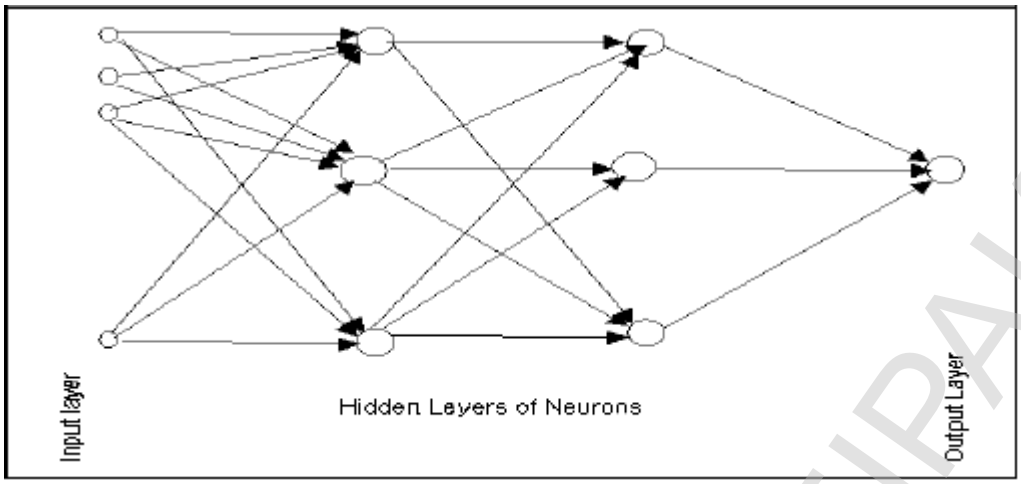
Σε γενικές γραμμές, ένας νευρώνας σε ένα τεχνητό νευρωνικό δίκτυο είναι

1. Ένα σύνολο τιμών εισαγωγής (x_i) και σχετικών βαρών (w_i)

2. Μια λειτουργία (ζ) που αθροίζει τα βάρη και χαρτογραφεί τα αποτελέσματα σε μια παραγωγή (ω)



Οι νευρώνες οργανώνονται σε στρώματα.



Το στρώμα εισαγωγής αποτελείται όχι από τους πλήρεις νευρώνες, αλλά μάλλον συνίσταται απλά από τις τιμές σε ένα αρχείο στοιχείων, το οποίο αποτελεί τις εισαγωγές στο επόμενο στρώμα των νευρώνων. Το επόμενο στρώμα καλείται κρυμμένο στρώμα, μπορούν να υπάρξουν διάφορα κρυμμένα στρώματα. Το τελικό στρώμα είναι το στρώμα παραγωγής, όπου υπάρχει ένας κόμβος για κάθε κατηγορία. Ένα ενιαίο σάρωμα προς τα εμπρός μέσω των δικτύων έχει ως αποτέλεσμα την ανάθεση μιας αξίας σε κάθε κόμβο παραγωγής, και η τιμή κατατάσσεται στον κόμβο της κατηγορίας η οποία έχει την υψηλότερη τιμή.

4.2.4.1. Κατάρτιση ενός Artificial Neural Network

Στη φάση της κατάρτισης, η σωστή κατηγορία για κάθε αρχείο είναι γνωστή (αυτό καλείται εποπτευμένη κατάρτιση), και στους κόμβους παραγωγής μπορούν επομένως να οριστούν οι "σωστές" τιμές -- "1" για τον κόμβο που αντιστοιχεί στη σωστή κατηγορία, και "0" για άλλες. (Στην πράξη έχει βρεθεί καλύτερο να χρησιμοποιηθούν οι τιμές 0,9 και 0,1, αντίστοιχα.) Είναι έτσι δυνατό να συγκριθούν οι υπολογισμένες τιμές του δικτύου για τους κόμβους παραγωγής με αυτές τις "σωστές" τιμές, και να υπολογιστεί ένας όρος λάθους για κάθε κόμβο (ο κανόνας "Delta"). Αυτοί οι όροι λάθους χρησιμοποιούνται έπειτα για να ρυθμίσουν τα βάρη στα κρυμμένα στρώματα έτσι ώστε, ενδεχομένως, η επόμενη φορά γύρω από τις τιμές παραγωγής θα είναι ποίο στενή στις "σωστές" τιμές.

4.2.4.2. Η επαναληπτική διαδικασία εκμάθησης

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων (neural networks) είναι ότι πρόκειται για μια επαναληπτική διαδικασία εκμάθησης στην οποία οι περιπτώσεις στοιχείων (σειρές) παρουσιάζονται στο δίκτυο μια την φορά, και τα βάρη που συνδέονται με τις τιμές εισαγωγής ρυθμίζονται κάθε φορά. Αφότου παρουσιάζονται όλες οι περιπτώσεις, η διαδικασία αρχίζει από την αρχή. Κατά τη διάρκεια αυτής της φάσης εκμάθησης, το δίκτυο μαθαίνει με τη ρύθμιση των βαρών ώστε να είναι σε θέση να προβλέψει τη σωστή κατηγορία των δειγμάτων εισαγωγής. Η εκμάθηση των Neural network αναφέρεται επίσης ως "διασυνδέτη εκμάθηση" (connectionist learning), λόγω των συνδέσεων μεταξύ των μονάδων. Στα πλεονεκτήματα των neural networks περιλαμβάνονται η υψηλή τους αντοχή στα θορυβώδη στοιχεία, καθώς επίσης και τη δυνατότητά τους να ταξινομήσουν τα στοιχεία στα οποία δεν έχουν εκπαιδευθεί. Ο δημοφιλέστερος neural network ο αλγόριθμος είναι back-propagation αλγόριθμος που προτάθηκε στη δεκαετία του '80.

Μόλις κτιστεί ένα δίκτυο για μια συγκεκριμένη εφαρμογή, το δίκτυο είναι έτοιμο να εκπαιδευθεί. Για να αρχίσουν αυτήν την διαδικασία, τα αρχικά βάρη (που περιγράφονται στο επόμενο τμήμα) επιλέγονται τυχαία. Κατόπιν η κατάρτιση, ή η εκμάθηση, αρχίζει.

Το δίκτυο επεξεργάζεται τα αρχεία δεδομένων των στοιχείων κατάρτισης ένα την φορά, χρησιμοποιώντας τα βάρη και τις λειτουργίες στα κρυμμένα στρώματα, κατόπιν συγκρίνει τα προκύπτοντα αποτελέσματα ενάντια στα επιθυμητά αποτελέσματα. Τα λάθη διαδίδονται έπειτα πίσω μέσω του συστήματος, αναγκάζοντας το σύστημα να ρυθμίσει τα βάρη και την εφαρμογή τους στο επόμενο αρχείο που υποβάλλεται σε επεξεργασία. Αυτή η διαδικασία επαναλαμβάνεται ξανά και ξανά και τα βάρη αναπροσαρμόζονται συνεχώς. Κατά τη διάρκεια της κατάρτισης ενός δικτύου το ίδιο σύνολο στοιχείων υποβάλλεται σε επεξεργασία πολλές φορές καθώς τα βάρη σύνδεσης καθαρίζονται συνεχώς.

Σημειώνουμε ότι μερικά δίκτυα δεν μαθαίνουν ποτέ. Αυτό θα μπορούσε να συμβεί επειδή τα δεδομένα εισόδου δεν περιέχουν τις συγκεκριμένες πληροφορίες από τις οποίες προέρχεται η επιθυμητή παραγωγή. Τα δίκτυα επίσης δεν συγκλίνουν εάν δεν υπάρχουν αρκετά στοιχεία τα οποία να επιτρέπουν την πλήρη εκμάθηση. Ιδανικά, πρέπει να υπάρξουν αρκετά στοιχεία έτσι ώστε μέρος των στοιχείων να μπορεί να συγκρατηθεί ως σύνολο επικύρωσης (validation set).

4.2.4.3. Feedforward, Back-Propagation

Η Feedforward, back-propagation αρχιτεκτονική αναπτύχθηκε στις αρχές της δεκαετίας του '70 από διάφορες ανεξάρτητες πηγές (Werbor Parker Rumelhart, Hinton και Ουίλιαμς). Αυτή η ανεξάρτητη παράλληλη ανάπτυξη είχε ως αποτέλεσμα τον πολλαπλασιασμό των άρθρων και διάφορες διασκέψεις που υποκίνησαν μια ολόκληρη βιομηχανία. Κατά την διάρκεια αυτής της περιόδου, αυτή η συνεργικά αναπτυγμένη back-propagation αρχιτεκτονική ήταν η δημοφιλέστερη, αποτελεσματικότερη, και ποιο εύκολη προς υλοποίηση ενός προτύπου εκμάθησης για τα σύνθετα, πολύ-στρωματικά δίκτυα. Η μέγιστη δύναμή της είναι οι μη γραμμικές λύσεις στα λάθος καθορισμένα προβλήματα. Το χαρακτηριστικό back-propagation δίκτυο έχει ένα στρώμα εισαγωγής, ένα στρώμα παραγωγής, και τουλάχιστον ένα κρυμμένο στρώμα. Δεν υπάρχει κανένα θεωρητικό όριο στον αριθμό κρυμμένων στρωμάτων αλλά συνήθως έχουμε ένα ή δύο. Μελέτες έχουν δείξει ότι ένας μέγιστος αριθμός πέντε στρωμάτων (ένα στρώμα εισαγωγής, τρία κρυμμένα στρώματα και ένα στρώμα παραγωγής) απαιτείται για να λύσει τα προβλήματα οποιασδήποτε πολυπλοκότητας. Κάθε στρώμα συνδέεται πλήρως με το επόμενο στρώμα.

Όπως αναφέρθηκε παραπάνω, η διαδικασία κατάρτισης χρησιμοποιεί κανονικά κάποια παραλλαγή του Delta Rule, η οποία αρχίζει με την υπολογισμένη διαφορά μεταξύ των πραγματικών αποτελεσμάτων και των επιθυμητών αποτελεσμάτων. Χρησιμοποιώντας αυτό το λάθος, τα βάρη σύνδεσης αυξάνονται αναλογικά ως προς το λάθος πολλαπλασιασμένα με ένα σταδιακό παράγοντα για τη σφαιρική ακρίβεια. Πραγματοποιώντας το προηγούμενο για έναν μεμονωμένο κόμβο σημαίνει ότι πρέπει η εισαγωγή, η παραγωγή, και η επιθυμητή παραγωγή να είναι όλες παρούσες στο ίδιο στοιχείο επεξεργασίας. Το σύνθετο μέρος αυτού του μηχανισμού εκμάθησης είναι για το σύστημα να καθοριστεί ποια εισαγωγή συνέβαλε περισσότερο σε μια ανακριβή παραγωγή και πώς το στοιχείο πρέπει να προσαρμοστεί για να διορθωθεί το λάθος. Ένας ανενεργός κόμβος δεν θα συνέβαλλε στο λάθος και δεν θα υπήρχε καμία ανάγκη να υποστεί αλλαγή στα βάρη του. Για να λυθεί το πρόβλημα αυτό, τα δεδομένα κατάρτισης (training inputs) εφαρμόζονται στο στρώμα εισαγωγής του δικτύου, και τα επιθυμητά αποτελέσματα συγκρίνονται με τα δεδομένα από το στρώμα παραγωγής. Κατά τη διάρκεια της διαδικασίας εκμάθησης, ένα σάρωμα μίας κατεύθυνσης πραγματοποιείται στο δίκτυο, και το αποτέλεσμα κάθε στοιχείου υπολογίζεται από στρώμα σε στρώμα. Η διαφορά μεταξύ της παραγωγής του τελικού στρώματος και της επιθυμητής παραγωγής διαδίδεται προς τα πίσω στα προηγούμενα στρώματα, και τροποποιείται συνήθως από την παράγωγο της συνάρτησης μεταφοράς, έτσι τα βάρη σύνδεσης ρυθμίζονται κανονικά χρησιμοποιώντας το Delta

Rule. Αυτή η διαδικασία προχωρά και στα προηγούμενα στρώματα έως ότου φτάσει στο στρώμα εισαγωγής.

4.2.4.4. Δόμηση του Network

Ο αριθμός στρωμάτων και ο αριθμός στοιχείων επεξεργασίας ανά στρώμα είναι πολύ σημαντικές αποφάσεις. Αυτές οι παράμετροι για μία feedforward, back-propagation τοπολογία είναι πολύ σημαντικές - είναι η "τέχνη" του σχεδιαστή δικτύων. Δεν υπάρχει καμία ποσοτικά προσδιορίσιμη, καλύτερη απάντηση για την δομή του δικτύου για οποιαδήποτε συγκεκριμένη εφαρμογή. Υπάρχουν μόνο γενικοί κανόνες που λαμβάνονται κατά τη διάρκεια του χρόνου και που ακολουθούνται από τους περισσότερους ερευνητές και μηχανικούς που εφαρμόζουν αυτήν την αρχιτεκτονική στα προβλήματά τους.

Rule One : *Όσο αυξάνει η πολυπλοκότητα στη σχέση μεταξύ των δεδομένων εισόδου και των επιθυμητών παραγόντων, τόσο πρέπει να αυξηθεί και ο αριθμός των στοιχείων επεξεργασίας στο κρυμμένο στρώμα.*

Rule Two: *Εάν η διαδικασία που μοντελοποιείτε διαχωρίζεται σε πολλαπλάσια στάδια, τότε ένα επιπλέον κρυμμένο στρώμα(τα) μπορεί να χρειάζεται. Εάν η διαδικασία δεν διαχωρίζεται σε στάδια, τότε τα επιπλέον στρώματα μπορούν απλά να επιτρέψουν την αποστήθιση του συνόλου κατάρτισης, με αποτέλεσμα μια μη αληθινή γενική λύση αποτελεσματική με άλλα στοιχεία.*

Rule Three: *Το διαθέσιμο ποσό στοιχείων κατάρτισης θέτει έναν ανώτερο όριο για τον αριθμό στοιχείων επεξεργασίας στο κρυμμένο στρώμα (τα). Για να υπολογίσετε αυτόν τον ανώτερο όριο, χρησιμοποιήστε τον αριθμό περιπτώσεων στο σύνολο στοιχείων κατάρτισης και διαιρέστε εκείνο τον αριθμό με το ποσό του αριθμού κόμβων στα στρώματα εισαγωγής και παραγωγής στο δίκτυο. Κατόπιν διαιρέστε ότι αποτέλεσμα πάλι με έναν παράγοντα μεταξύ πέντε και δέκα. Οι μεγαλύτεροι παράγοντες χρησιμοποιούνται για τα σχετικά λιγότερα θορυβώδη στοιχεία. Εάν χρησιμοποιήσετε πάρα πολλούς τεχνητούς νευρώνες το σύνολο κατάρτισης θα απομνημονευθεί. Εάν αυτό συμβεί, η γενίκευση των στοιχείων δεν θα εμφανιστεί, καθιστώντας το δίκτυο άχρηστο στα νέα σύνολα στοιχείων.*

4.2.4.5. Εφαρμογή της μεθοδολογίας

Κατά τον σχεδιασμό των μοντέλων ακολουθήσαμε τα εξής βήματα:

- Επεξεργαστήκαμε τα δεδομένα του συνόλου κατάρτισης, συγκεκριμένα τα κανονικοποιήσαμε. Η κανονικοποίηση των δεδομένων είναι απαραίτητη έτσι ώστε να εξασφαλίσουμε ότι το μέτρο απόστασης κατανέμει το ίδιο βάρος σε κάθε μεταβλητή -- χωρίς κανονικοποίηση, η μεταβλητή με τη μεγαλύτερη κλίμακα θα εξουσίαζε στο μέτρο.
- Σε όλες μας τις προσπάθειες δημιουργίας μοντέλων από τα διαφορετικά χρησιμοποιήσαμε 1και 2 hidden layers, στα αρχικά μοντέλα που δημιουργήσαμε είχαμε χρησιμοποιήσει και μεγαλύτερο αριθμό layers όμως είχαμε φαινόμενα over fitting.
- Για τον αριθμό των nodes επιλέξαμε μετά από δοκιμές τα 25.
- Επιλέξαμε για μέγεθος βημάτων ως προς την κάθοδο κλίσης το 0,1. Το μέγεθος βημάτων ως προς την κάθοδο κλίσης είναι ο πολλαπλασιάζοντας παράγοντας για τη διόρθωση λάθους κατά τη διάρκεια του backpropagation , είναι κατά προσέγγιση ισοδύναμο με το ποσοστό εκμάθησης για το νευρικό δίκτυο. Μια χαμηλή τιμή παράγει μια αργή αλλά σταθερή εκμάθηση, ενώ μια υψηλή τιμή παράγει γρήγορη αλλά ακανόνιστη εκμάθηση. Οι τιμές για το μέγεθος βημάτων κυμαίνονται χαρακτηριστικά από 0,1 έως 0,9,
- Επιλέξαμε για ταχύτητα αλλαγής βάρους το 0,6. Σε κάθε νέο κύκλο της διόρθωσης λάθους, κάποια μνήμη της προγενέστερης διόρθωσης διατηρείται έτσι ώστε ένα outlier δεν θα χαλάσει την συσσώρευμένη εκμάθηση. Οι τιμές ποικίλουν από 0 έως 2.
- Επιλέξαμε ως τιμή της Ανοχής λάθους (Error Tolerance) το 0,01. Το λάθος σε μια ιδιαίτερη επανάληψη είναι backpropagated μόνο εάν είναι μεγαλύτερο από την ανοχή λάθους. Τυπικά η ανοχή λάθους είναι μια μικρή αξία στη σειρά 0 έως 1.
- Επιλέξαμε ως τιμή Αποσύνθεσης βάρους (Weight decay) το 0. Για να αποτρέψουμε το over-fitting του δικτύου στο σύνολο στοιχείων κατάρτισης χρησιμοποιείται το Weight Decay για να τιμωρήσει το λάθος σε κάθε επανάληψη. Έτσι εάν e είναι το λάθος το οποίο θα γίνει back-propagated, εκείνο το οποίο θα γίνει back-propagated είναι το (e + w*e), όπου w είναι το weight decay, το οποίο μπορεί να πάρει τιμές από 0 έως 1.
- Τα μοντέλα μας τα σχεδιάσαμε με δύο cost functions την Squared Error και την Maximum Likelihood.

- Εφαρμόσαμε το μοντέλο στα σύνολα επικύρωσης και δοκιμής.
- Και στο τελευταίο στάδιο εφαρμόσαμε το μοντέλο στο σύνολο ελέγχου.

Επαναλάβαμε την διαδικασία αυτή σε όλα τα τυχαία δείγματα που είχαμε δημιουργήσει και υπολογίσαμε την μέση τιμή και την διακύμανση της ακρίβειας (των μοντέλων) ως προς το σύνολο ελέγχου.

Επαναλαμβάνουμε την διαδικασία που ακολουθήσαμε και στις προηγούμενες μεθόδους, δηλαδή εφαρμόζουμε στο δείγμα που είχαμε χρησιμοποιήσει και στις προηγούμενες περιπτώσεις την μέθοδο των Νευρωνικών δικτύων.

Ακολουθώντας την μεθοδολογία που αναπτύξαμε παραπάνω προκύπτουν τέσσερα μοντέλα:

1. Μοντέλα με ένα Hidden Layer και Cost function την Squared Error (NN1).
2. Μοντέλα με δύο Hidden Layers και Cost function την Squared Error (NN2).
3. Μοντέλα με ένα Hidden Layer και Cost function την Maximum Likelihood (NN3).
4. Μοντέλα με δύο Hidden Layers και Cost function την Maximum Likelihood (NN4).

Για το συγκεκριμένο δείγμα είχαμε τα εξής αποτελέσματα ανά μοντέλο:

Neural Net Model	NN 1	NN 2	NN 3	NN 4
Misclassification Error	15.71%	13.29%	20.57%	20.00%

5. Αποτελέσματα – Συμπεράσματα

Θα παρουσιάσουμε τα αποτελέσματα για κάθε κατηγορία δείγματος χωριστά και στο τέλος θα αναφερθούμε συνολικά για τα αποτελέσματα που προέκυψαν από τις διαφορετικές προσεγγίσεις που ακολουθήσαμε στην μελέτη μας.

5.1. Ισοκατανεμημένα Δείγματα.

Κατά τον σχεδιασμό των δειγμάτων αυτών ακολουθήσαμε, μια διαδικασία η οποία μας επέτρεψε να σχεδιάσουμε δείγματα τα οποία είχαν το ίδιο μέγεθος 'ενήμερων' και 'σε καθυστέρηση' πελατών. Συγκεκριμένα χωρίσαμε τα δεδομένα σε δύο υποσύνολα, ένα μόνο με 'ενήμερους' πελάτες και ένα με πελάτες 'σε καθυστέρηση'. Από αυτά τα υποσύνολα δημιουργήσαμε τυχαία δείγματα ίσου μεγέθους για κάθε υποσύνολο. Ενώνοντας τα δύο δείγματα προέκυπτε το δείγμα από το οποίο δημιουργούσαμε τα σύνολα κατάρτισης, επικύρωσης ,δοκιμής και ελέγχου.

5.1.1. Ισοκατανεμημένα δείγματα μεγέθους 10% του συνολικού πληθυσμού.

Τα δείγματα της κατηγορίας αυτής σχεδιάστηκαν με αυτό τον τρόπο ώστε να έχουν μέγεθος ίσο με το 10% του συνόλου από το οποίο προέκυψαν. Η σύνθεση αυτών των δειγμάτων και των συνόλων που δημιουργήθηκαν παρουσιάζεται στους παρακάτω πίνακες:

	'Καλοί' πελάτες	**'Κακοί' Πελάτες**	**Σύνολο**
# Παρατηρήσεων	**200**	**200**	**400**

Τα σύνολα που δημιουργήθηκαν από κάθε δείγμα είχαν την εξής δομή:

Σύνολο Κατάρτισης	**Σύνολο Επικύρωσης**	**Σύνολο Δοκιμής**
200	**100**	**100**

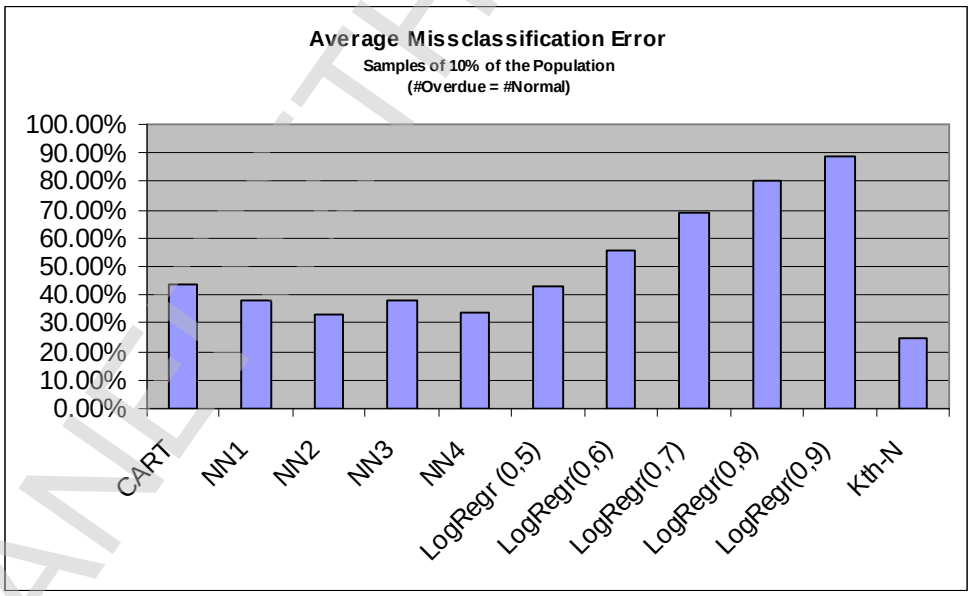
Επιπλέον χρησιμοποιήθηκε ένα ακόμα σύνολο το 'σύνολο ελέγχου' μεγέθους 700 παρατηρήσεων για το οποίο υπολογίσαμε το σφάλμα ταξινόμησης.

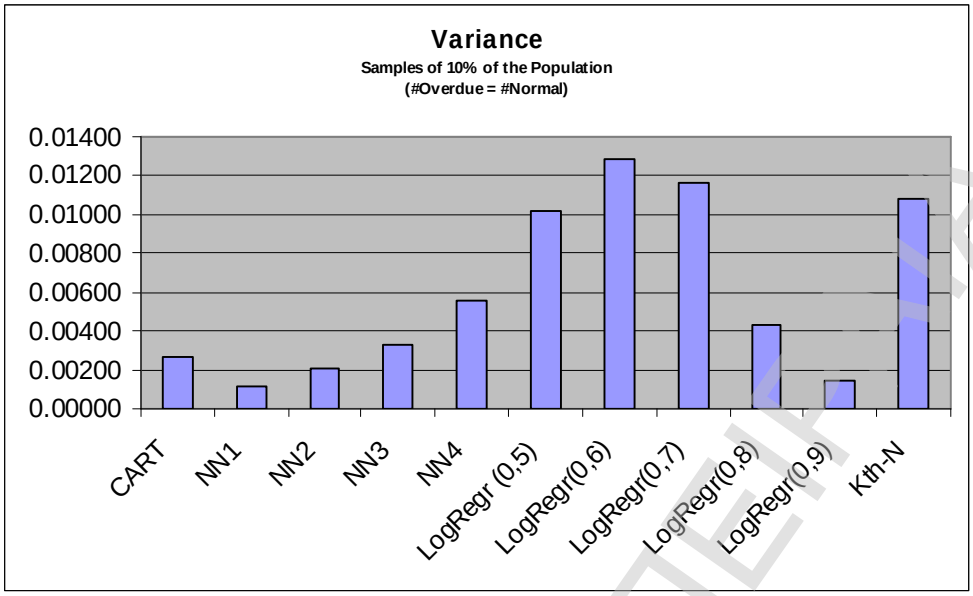
Εφαρμόζοντας τα μοντέλα, που προέκυψαν από κάθε δείγμα, στο σύνολο ελέγχου είχαμε τα εξής σφάλματα ταξινόμησης:

	CART	NN1	NN2	NN3	NN4	LogRegr (0,5)	LogRegr(0,6)	LogRegr(0,7)	LogRegr(0,8)	LogRegr(0,9)	Kth-N
Sample 1	32.57%	37.00%	30.86%	44.29%	34.43%	29.14%	49.43%	59.14%	72.14%	85.00%	10.86%
Sample 2	43.86%	34.86%	34.71%	37.29%	36.14%	36.57%	47.71%	69.00%	80.86%	92.43%	11.57%
Sample 3	41.00%	39.00%	28.71%	38.43%	39.86%	44.29%	58.29%	74.43%	81.71%	89.86%	31.57%
Sample 4	46.57%	41.29%	33.29%	32.71%	30.43%	39.14%	53.71%	69.00%	80.43%	92.71%	28.00%
Sample 5	48.29%	31.57%	33.43%	33.43%	30.43%	49.14%	65.57%	74.71%	87.43%	90.71%	14.29%
Sample 6	47.86%	41.43%	41.71%	50.00%	45.29%	43.71%	47.14%	53.14%	72.00%	81.57%	34.71%
Sample 7	40.29%	37.57%	34.00%	36.57%	34.43%	38.43%	42.57%	60.29%	77.43%	87.43%	25.71%
Sample 8	48.57%	41.86%	25.29%	32.29%	18.14%	65.43%	79.57%	90.14%	92.57%	92.71%	39.86%
Average Misclassification Error	**43.63%**	**38.07%**	**32.75%**	**38.13%**	**33.64%**	**43.23%**	**55.50%**	**68.73%**	**80.57%**	**89.05%**	**24.57%**
Variance	**0.00266**	**0.00113**	**0.00203**	**0.00335**	**0.00553**	**0.01015**	**0.01283**	**0.01166**	**0.00434**	**0.00146**	**0.01077**

Από τα αποτελέσματα προκύπτει ότι το μικρότερο μέσο σφάλμα ταξινόμησης για το συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα και χαρακτηριστικά που είχαμε στην διάθεσή μας και για τον συγκεκριμένο τρόπο σχεδιασμού των δειγμάτων προκύπτει από την μέθοδο των κ-κοντινότερων γειτόνων. Συγκεκριμένα η μέθοδος αυτή μας έδωσε μέσο σφάλμα 24,57% με δεύτερη καλύτερη την μέθοδο των νευρωνικών δικτύων όταν τα δίκτυα αυτά έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Squared Error η οποία έχει μέσο σφάλμα 32,75%. Αρκετά κοντά στα αποτελέσματα της δεύτερη αποτελεσματικότερης μεθόδου είναι και τα αποτελέσματα της μεθόδου των νευρωνικών δικτύων τα οποία έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Maximum Likelihood η οποία έχει μέσο σφάλμα 33,64%. Οι μέθοδοι των νευρωνικών δικτύων με ένα κρυφό επίπεδο ανεξαρτήτως της cost function που χρησιμοποιήσαμε είχαν σχεδόν τα ίδια αποτελέσματα (Squared Error : 38,07% , Maximum Likelihood: 38,13%). Τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των δέντρων αποφάσεων έδωσαν ‘φτωχά’ αποτελέσματα όσον αφορά της προβλεπτική τους ικανότητα στο σύνολο ελέγχου των 700 παρατηρήσεων, συγκεκριμένα το μέσο σφάλμα ταξινόμησης ήταν 43,63%. Ακόμα ποιο απογοητευτικά ήταν τα αποτελέσματα από τα μοντέλα που προέκυψαν από την μέθοδο της λογιστικής παλινδρόμησης. Υπολογίσαμε τα σφάλματα ταξινόμησης για τα μοντέλα της μεθόδου αυτής χρησιμοποιώντας διαφορετικά cut off points και από τα αποτελέσματα προέκυψε ότι την καλύτερη ταξινόμηση για την μέθοδο αυτή την έχουμε για cut off point 0,5 (μέσο σφάλμα ταξινόμησης 43,23%), ενώ την χειρότερη για cut off point 0,9 (μέσο σφάλμα ταξινόμησης 89,05%).

Στα παρακάτω διαγράμματα παρουσιάζουμε το μέσο σφάλμα ταξινόμησης για την κάθε μέθοδο, που προέκυψε από την εφαρμογή της στα διαφορετικά δείγματα, καθώς επίσης και την διακύμανση του σφάλματος.





5.1.2. Ισοκατανεμημένα δείγματα μεγέθους 40% του συνολικού πληθυσμού.

Τα δείγματα της κατηγορίας αυτής σχεδιάστηκαν με αυτό τον τρόπο ώστε να έχουν μέγεθος ίσο με το 40% του συνόλου από το οποίο προέκυψαν. Η σύνθεση αυτών των δειγμάτων και των συνόλων που δημιουργήθηκαν παρουσιάζεται στους παρακάτω πίνακες:

	Ενήμεροι πελάτες	Πελάτες σε Καθυστέρηση	Σύνολο
# Παρατηρήσεων	**800**	**800**	**1600**

Τα σύνολα που δημιουργήθηκαν από κάθε δείγμα είχαν την εξής δομή:

Σύνολο Κατάρτισης	Σύνολο Επικύρωσης	Σύνολο Δοκιμής
800	**400**	**400**

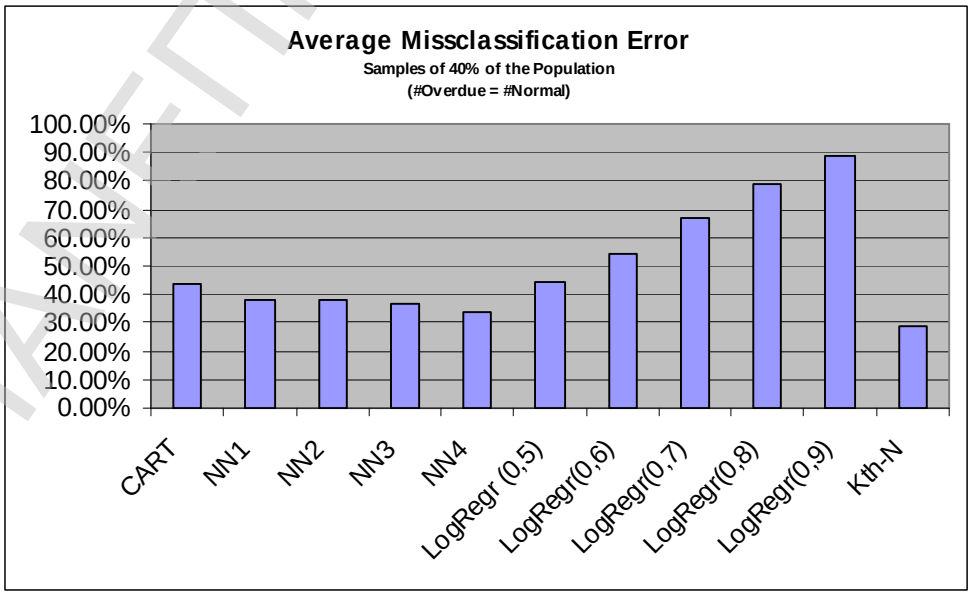
Επιπλέον χρησιμοποιήθηκε ένα ακόμα σύνολο το ‘σύνολο ελέγχου’ μεγέθους 700 παρατηρήσεων για το οποίο υπολογίσαμε το σφάλμα ταξινόμησης.

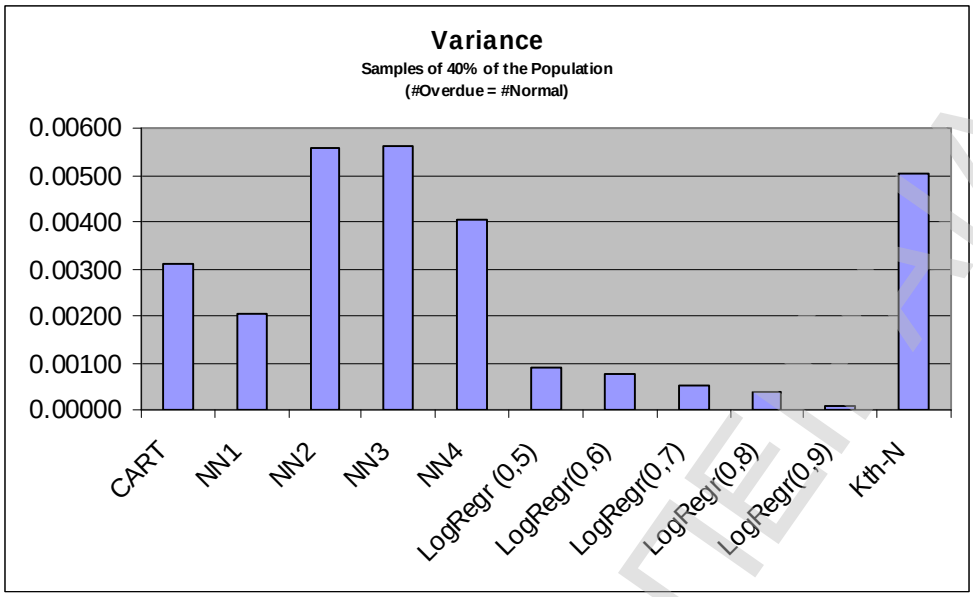
Εφαρμόζοντας τα μοντέλα, που προέκυψαν από κάθε δείγμα, στο σύνολο ελέγχου είχαμε τα εξής σφάλματα ταξινόμησης:

	CART	NN1	NN2	NN3	NN4	LogRegr (0,5)	LogRegr(0,6)	LogRegr(0,7)	LogRegr(0,8)	LogRegr(0,9)	Kth-N
Sample 1	41.57%	39.43%	32.86%	49.43%	32.86%	42.14%	51.00%	63.29%	76.71%	89.00%	28.17%
Sample 2	39.25%	32.86%	35.00%	39.14%	41.29%	41.71%	54.43%	67.57%	80.43%	89.57%	18.83%
Sample 3	47.29%	46.43%	51.00%	37.57%	36.29%	51.00%	53.00%	69.00%	81.71%	90.00%	18.16%
Sample 4	44.00%	40.71%	39.29%	42.86%	31.29%	43.29%	55.14%	67.14%	75.71%	87.57%	33.11%
Sample 5	49.57%	41.00%	40.71%	30.86%	27.86%	42.29%	50.57%	68.29%	80.57%	89.43%	35.00%
Sample 6	53.29%	32.14%	45.86%	39.29%	45.86%	44.43%	52.14%	64.00%	79.00%	88.29%	29.11%
Sample 7	34.71%	38.29%	29.14%	31.71%	27.29%	44.57%	56.00%	65.29%	77.71%	87.86%	40.19%
Sample 8	41.43%	34.43%	28.29%	23.43%	28.29%	47.57%	59.43%	70.29%	79.86%	89.86%	27.64%
Average Misclassification Error	**43.89%**	**38.16%**	**37.77%**	**36.79%**	**33.88%**	**44.63%**	**53.96%**	**66.86%**	**78.96%**	**88.95%**	**28.77%**
Variance	**0.00312**	**0.00204**	**0.00559**	**0.00561**	**0.00404**	**0.00089**	**0.00075**	**0.00053**	**0.00038**	**0.00008**	**0.00502**

Από τα αποτελέσματα προκύπτει ότι το μικρότερο μέσο σφάλμα ταξινόμησης για το συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα και χαρακτηριστικά που είχαμε στην διάθεσή μας και για τον συγκεκριμένο τρόπο σχεδιασμού των δειγμάτων προκύπτει από την μέθοδο των κ-κοντινότερων γειτόνων. Συγκεκριμένα η μέθοδος αυτή μας έδωσε μέσο σφάλμα 28,77% με δεύτερη καλύτερη την μέθοδο των νευρωνικών δικτύων όταν τα δίκτυα αυτά έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Maximum Likelihood η οποία έχει μέσο σφάλμα 33,88%. Τα αποτελέσματα των μοντέλων της μεθόδου των νευρωνικών δικτύων τα οποία έχουν σχεδιαστεί με ένα κρυφό επίπεδα και η cost function είναι η Maximum Likelihood έχουν την τρίτη καλύτερη προβλεπτική ικανότητα με μέσο σφάλμα 36,79%. Παρατηρούμε των μοντέλων τα οποία έχουν προκύψει από την εφαρμογή της μεθόδου των νευρωνικών δικτύων με cost function την Squared Error δεν έχουν καλή προβλεπτική ικανότητα, συγκεκριμένα δίνουν μέσα σφάλματα ταξινόμησης 38,16% με ένα κρυφό επίπεδο και 37,77% με δύο κρυφά επίπεδα. Τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των δέντρων αποφάσεων έδωσαν ‘φτωχά’ αποτελέσματα όσον αφορά της προβλεπτική τους ικανότητα στο σύνολο ελέγχου των 700 παρατηρήσεων, συγκεκριμένα το μέσο σφάλμα ταξινόμησης ήταν 43,89%. Ακόμα ποιο απογοητευτικά ήταν τα αποτελέσματα από τα μοντέλα που προέκυψαν από την μέθοδο της λογιστικής παλινδρόμησης. Υπολογίσαμε τα σφάλματα ταξινόμησης για τα μοντέλα της μεθόδου αυτής χρησιμοποιώντας διαφορετικά cut off points και από τα αποτελέσματα προέκυψε ότι την καλύτερη ταξινόμηση για την μέθοδο αυτή την έχουμε για cut off point 0,5 (μέσο σφάλμα ταξινόμησης 44,63%), ενώ την χειρότερη για cut off point 0,9 (μέσο σφάλμα ταξινόμησης 88,95%).

Στα παρακάτω διαγράμματα παρουσιάζουμε το μέσο σφάλμα ταξινόμησης για την κάθε μέθοδο, που προέκυψε από την εφαρμογή της στα διαφορετικά δείγματα, καθώς επίσης και την διακύμανση του σφάλματος.





5.2. Αντιπροσωπευτικά Δείγματα.

Κατά τον σχεδιασμό των δειγμάτων αυτών, ακολουθήσαμε μια διαδικασία η οποία μας επέτρεψε να σχεδιάσουμε δείγματα τα οποία είχαν την ίδια σύνθεση 'ενήμερων' και 'σε καθυστέρηση' πελατών όπως το αρχικό σύνολο σχεδιασμού. Συγκεκριμένα χωρίσαμε τα δεδομένα σε δύο υποσύνολα, ένα μόνο με 'ενήμερους' πελάτες και ένα με 'σε καθυστέρηση' πελάτες. Από αυτά τα υποσύνολα δημιουργήσαμε τυχαία δείγματα με μέγεθος ανάλογο προς το ποσοστό τους στο αρχικό υποσύνολο. Ενώνοντας τα δύο δείγματα προέκυπτε το δείγμα από το οποίο δημιουργούσαμε τα σύνολα κατάρτισης, επικύρωσης, δοκιμής και επικύρωσης.

5.2.1. Αντιπροσωπευτικά δείγματα μεγέθους 10% του συνολικού πληθυσμού.

Τα δείγματα της κατηγορίας αυτής σχεδιάστηκαν με αυτό τον τρόπο ώστε να έχουν μέγεθος ίσο με το 10% του συνόλου από το οποίο προέκυψαν. Η σύνθεση αυτών των δειγμάτων και των συνόλων που δημιουργήθηκαν παρουσιάζεται στους παρακάτω πίνακες:

	Ενήμεροι πελάτες	Πελάτες σε Καθυστέρηση	Σύνολο
# Παρατηρήσεων	**83**	**317**	**400**

Τα σύνολα που δημιουργήθηκαν από κάθε δείγμα είχαν την εξής δομή:

Σύνολο Κατάρτισης	Σύνολο Επικύρωσης	Σύνολο Δοκιμής
200	**100**	**100**

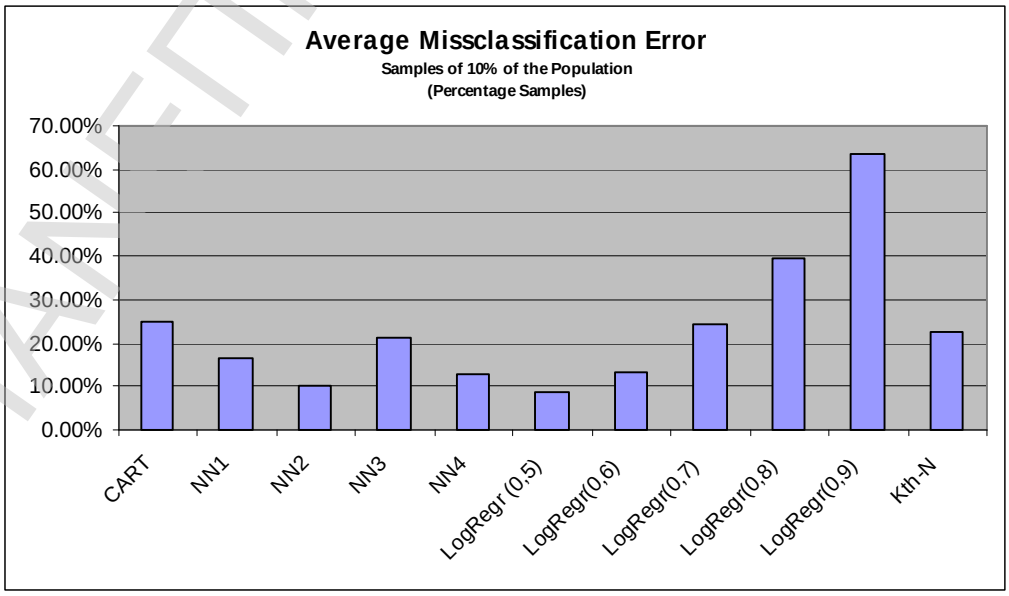
Επιπλέον χρησιμοποιήθηκε ένα ακόμα σύνολο το 'σύνολο ελέγχου' μεγέθους 700 παρατηρήσεων για το οποίο υπολογίσαμε το σφάλμα ταξινόμησης.

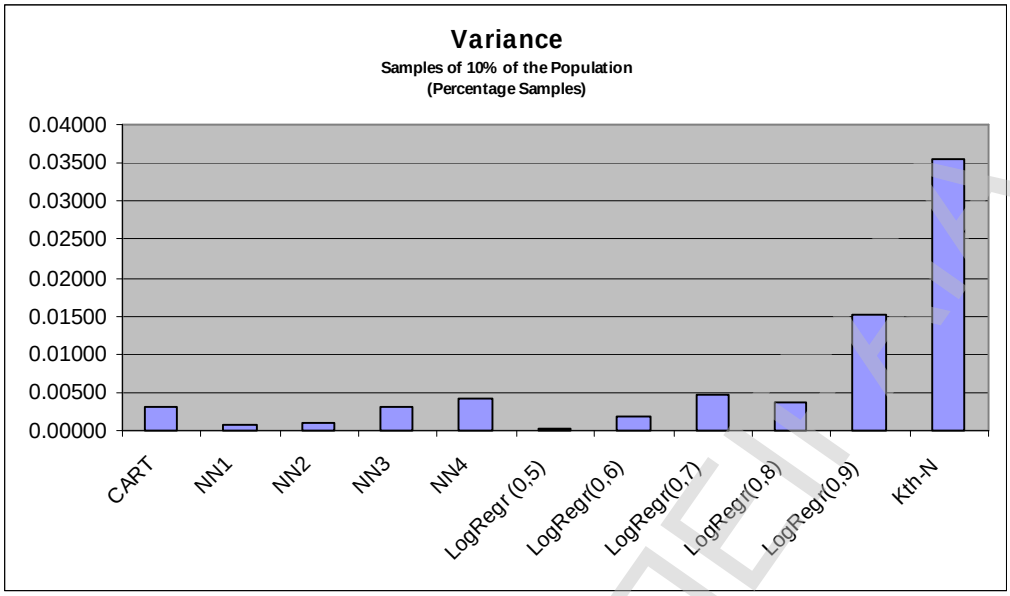
Εφαρμόζοντας τα μοντέλα, που προέκυψαν από κάθε δείγμα, στο σύνολο ελέγχου είχαμε τα εξής σφάλματα ταξινόμησης:

	CART	NN1	NN2	NN3	NN4	LogRegr (0,5)	LogRegr (0,6)	LogRegr (0,7)	LogRegr (0,8)	LogRegr (0,9)	Kth-N
Sample 1	27.86%	22.86%	10.29%	33.14%	7.00%	11.14%	11.14%	11.14%	50.00%	74.14%	72.43%
Sample 2	15.71%	16.29%	8.71%	19.43%	7.00%	7.14%	11.14%	25.86%	37.86%	62.57%	15.35%
Sample 3	21.86%	17.00%	11.71%	16.43%	16.29%	10.71%	15.43%	21.43%	37.43%	58.43%	15.35%
Sample 4	28.14%	16.86%	7.00%	25.86%	24.71%	9.43%	18.29%	26.71%	35.00%	47.71%	15.35%
Sample 5	30.29%	12.86%	7.00%	14.43%	20.00%	9.43%	18.29%	26.71%	35.00%	47.71%	15.35%
Sample 6	32.86%	15.86%	15.29%	23.14%	7.00%	7.57%	7.57%	36.43%	36.43%	74.14%	15.35%
Sample 7	20.57%	17.57%	15.14%	17.14%	13.00%	8.14%	18.57%	27.14%	34.57%	59.29%	15.35%
Sample 8	20.57%	14.29%	7.00%	21.71%	7.00%	7.57%	8.14%	18.57%	49.43%	84.57%	15.35%
Average Misclassification Error	**24.73%**	**16.70%**	**10.27%**	**21.41%**	**12.75%**	**8.89%**	**13.57%**	**24.25%**	**39.46%**	**63.57%**	**22.49%**
Variance	**0.00303**	**0.00075**	**0.00107**	**0.00320**	**0.00426**	**0.00020**	**0.00188**	**0.00479**	**0.00363**	**0.01517**	**0.03563**

Από τα αποτελέσματα προκύπτει ότι το μικρότερο μέσο σφάλμα ταξινόμησης για το συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα και χαρακτηριστικά που είχαμε στην διάθεσή μας και για τον συγκεκριμένο τρόπο σχεδιασμού των δειγμάτων προκύπτει από τα μοντέλα που σχεδιάστηκαν ακολουθώντας την μέθοδο της λογιστικής παλινδρόμησης με cut off point 0,5. Συγκεκριμένα η μέθοδος αυτή μας έδωσε μέσο σφάλμα 8,89% με δεύτερη καλύτερη την μέθοδο των νευρωνικών δικτύων όταν τα δίκτυα αυτά έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Squared Error, με μέσο σφάλμα 10,27%. Πολύ καλή προβλεπτική ικανότητα έδειξαν και τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των νευρωνικών δικτύων με δύο κρυφά επίπεδα και cost function την Maximum Likelihood, με μέσο σφάλμα ταξινόμησης 12,75%, μία επίδοση που ήταν πολύ κοντά σε αυτή της λογιστικής παλινδρόμησης με cut off point to 0,6 (13,57%). Οι υπόλοιπες μέθοδοι έδωσαν ποιο φτωχά αποτελέσματα. Συγκεκριμένα τα νευρωνικά δίκτυα με ένα κρυφό επίπεδο και cost function την Squared Error έδωσε μέσο σφάλμα ταξινόμησης ίσο με 16,70% έναντι των νευρωνικών δικτύων με ένα κρυφό επίπεδο και cost function την Maximum Likelihood τα οποία έδωσαν μέσο σφάλμα ταξινόμησης ίσο με 21,41%. Τα δέντρα αποφάσεων και τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των κ-κοντινότερων γειτόνων έδωσαν μέσα σφάλματα 24,73% και 22,49% αντίστοιχα. Παρατηρούμε επίσης ότι αυξάνοντας το cut off point στην λογιστική παλινδρόμηση αυξάνεται το μέσο σφάλμα ταξινόμησης συγκεκριμένα για cut off point 0.7 , 0.8 , 0.9 έχουμε μέσο σφάλμα ταξινόμησης 24,25% , 39,46% και 63,57% αντίστοιχα.

Στα παρακάτω διαγράμματα παρουσιάζουμε το μέσο σφάλμα ταξινόμησης για την κάθε μέθοδο, που προέκυψε από την εφαρμογή της στα διαφορετικά δείγματα, καθώς επίσης και την διακύμανση του σφάλματος.





5.2.2. Αντιπροσωπευτικά δείγματα μεγέθους 40% του συνολικού πληθυσμού.

Τα δείγματα της κατηγορίας αυτής σχεδιάστηκαν με αυτό τον τρόπο ώστε να έχουν μέγεθος ίσο με το 40% του συνόλου από το οποίο προέκυψαν. Η σύνθεση αυτών των δειγμάτων και των συνόλων που δημιουργήθηκαν παρουσιάζεται στους παρακάτω πίνακες:

	Ενήμεροι πελάτες	Πελάτες σε Καθυστέρηση	Σύνολο
# Παρατηρήσεων	**330**	**1270**	**1600**

Τα σύνολα που δημιουργήθηκαν από κάθε δείγμα είχαν την εξής δομή:

Σύνολο Κατάρτισης	Σύνολο Επικύρωσης	Σύνολο Δοκιμής
800	**400**	**400**

Επιπλέον χρησιμοποιήθηκε ένα ακόμα σύνολο το ‘σύνολο ελέγχου’ μεγέθους 700 παρατηρήσεων για το οποίο υπολογίσαμε το σφάλμα ταξινόμησης.

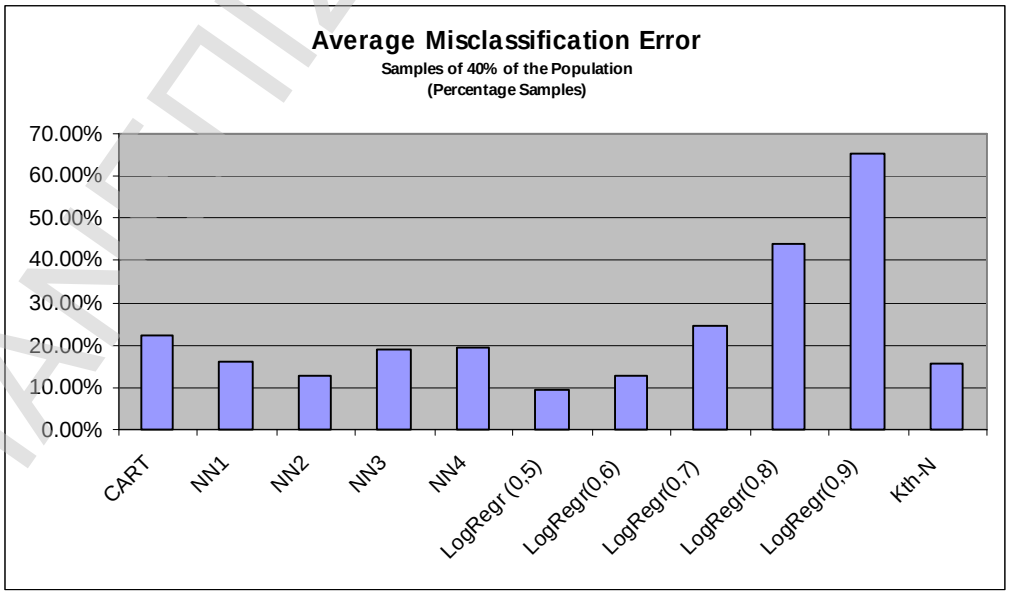
Εφαρμόζοντας τα μοντέλα, που προέκυψαν από κάθε δείγμα, στο σύνολο ελέγχου είχαμε τα εξής misclassification errors:

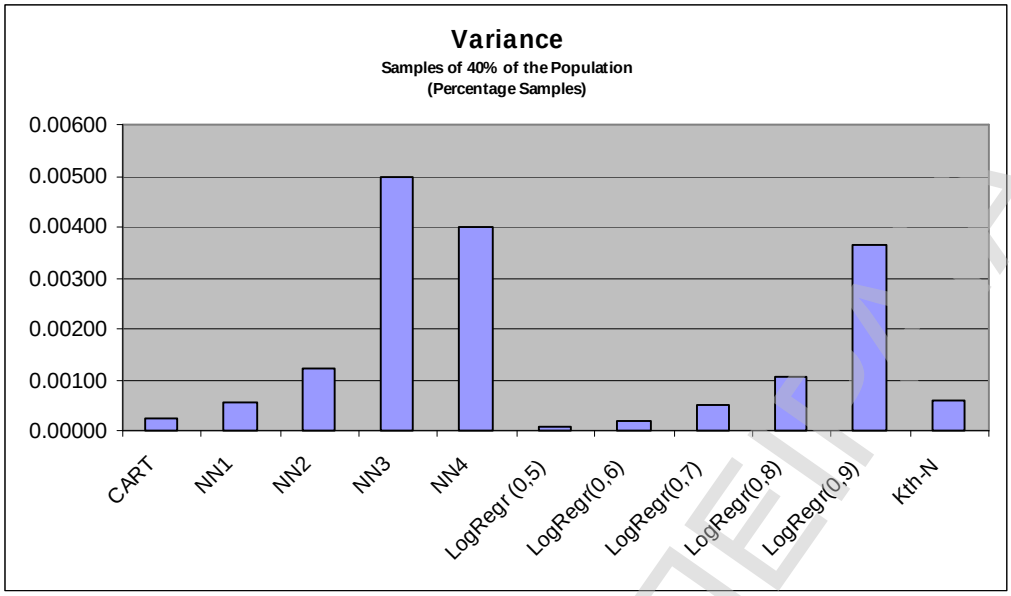
	CART	NN1	NN2	NN3	NN4	LogRegr (0,5)	LogRegr(0,6)	LogRegr(0,7)	LogRegr(0,8)	LogRegr(0,9)	Kth-N
Sample 1	21.86%	21.29%	12.29%	13.14%	13.57%	8.29%	12.43%	26.29%	40.00%	70.43%	15.35%
Sample 2	23.43%	12.86%	7.57%	23.29%	19.43%	9.29%	12.00%	21.29%	46.57%	64.57%	9.43%
Sample 3	20.29%	14.00%	13.57%	11.71%	14.57%	10.71%	11.14%	21.71%	42.86%	57.29%	15.35%
Sample 4	22.57%	15.71%	14.71%	21.00%	26.86%	9.71%	14.43%	23.86%	42.86%	72.43%	18.16%
Sample 5	21.86%	17.14%	13.29%	27.29%	19.71%	9.29%	12.00%	26.00%	42.57%	58.71%	15.62%
Sample 6	24.00%	14.57%	7.00%	7.00%	11.29%	9.29%	12.00%	26.00%	42.57%	58.71%	15.62%
Sample 7	19.71%	15.71%	13.29%	20.57%	20.00%	8.43%	13.71%	24.43%	42.14%	72.14%	16.56%
Sample 8	24.14%	16.14%	18.57%	27.57%	31.29%	10.86%	15.71%	28.57%	51.29%	69.43%	17.36%
Average Misclassification Error	**22.23%**	**15.93%**	**12.54%**	**18.95%**	**19.59%**	**9.48%**	**12.93%**	**24.77%**	**43.86%**	**65.46%**	**15.43%**
Variance	**0.00023**	**0.00057**	**0.00123**	**0.00498**	**0.00400**	**0.00008**	**0.00021**	**0.00053**	**0.00107**	**0.00365**	**0.00061**

Από τα αποτελέσματα προκύπτει ότι το μικρότερο μέσο σφάλμα ταξινόμησης για το συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα και χαρακτηριστικά που είχαμε στην διάθεσή μας και για τον συγκεκριμένο τρόπο σχεδιασμού των δειγμάτων προκύπτει από τα μοντέλα που σχεδιάστηκαν ακολουθώντας την μέθοδο της λογιστικής παλινδρόμησης με cut off point 0,5. Συγκεκριμένα η μέθοδος αυτή μας έδωσε μέσο σφάλμα 9,48% με δεύτερη καλύτερη την μέθοδο των νευρωνικών δικτύων όταν τα δίκτυα αυτά έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Squared Error, με μέσο σφάλμα 12,54%. Τα καλά αποτελέσματα της λογιστικής παλινδρόμησης διατηρούνται και όταν ορίσουμε cut off point το 0.6 , έχουμε μέσο σφάλμα 12,93%.Αρκετά καλή προβλεπτική ικανότητα έδειξαν τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των νευρωνικών δικτύων με ένα κρυφό επίπεδα και cost function την Squared Error, με μέσο σφάλμα ταξινόμησης 15,93% καθώς επίσης και τα μοντέλα που προέκυψαν από την εφαρμογή της μεθοδολογίας των κ-κοντινότερων γειτόνων με μέσο σφάλμα 15,43% .

Οι υπόλοιπες μέθοδοι έδωσαν ποιο φτωχά αποτελέσματα. Συγκεκριμένα τα νευρωνικά δίκτυα με ένα κρυφό επίπεδο και cost function την Maximum Likelihood έδωσε μέσο σφάλμα ταξινόμησης ίσο με 18,95% ενώ τα νευρωνικά δίκτυα με δύο κρυφά επίπεδα έδωσαν μέσο σφάλμα ταξινόμησης ίσο με 19,59%. Τα δέντρα αποφάσεων έδωσαν μέσα σφάλματα 22,23%. Παρατηρούμε επίσης ότι αυξάνοντας το cut off point στην λογιστική παλινδρόμηση αυξάνεται το μέσο σφάλμα ταξινόμησης συγκεκριμένα για cut off point 0.7 , 0.8 , 0.9 έχουμε μέσο σφάλμα ταξινόμησης 24.77% , 42.86% και 65.46% αντίστοιχα.

Στα παρακάτω διαγράμματα παρουσιάζουμε το μέσο σφάλμα ταξινόμησης για την κάθε μέθοδο, που προέκυψε από την εφαρμογή της στα διαφορετικά δείγματα, καθώς επίσης και την διακύμανση του σφάλματος.

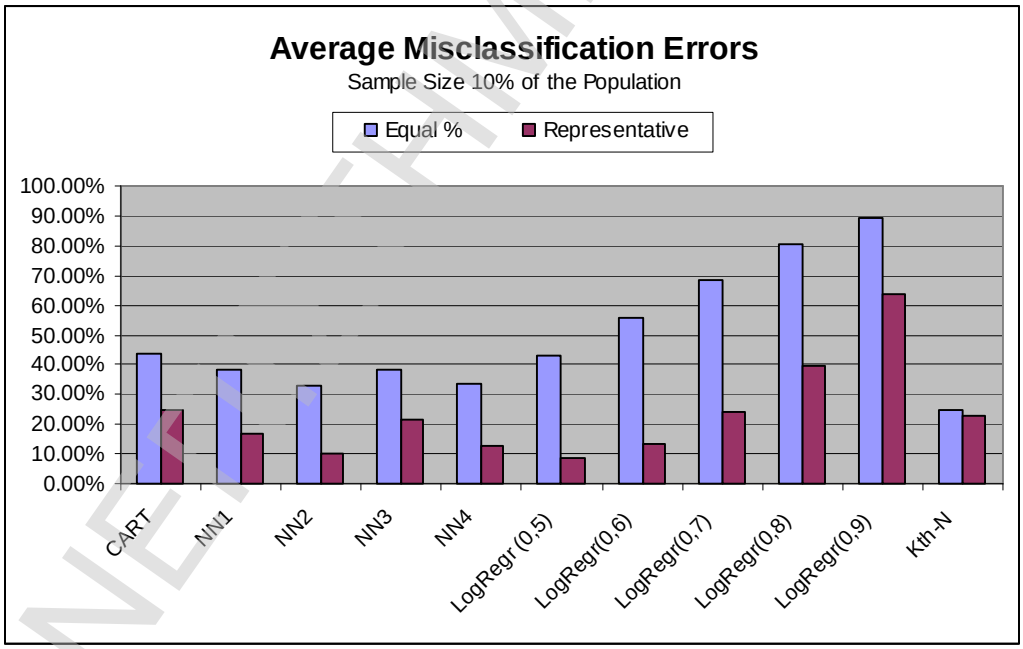


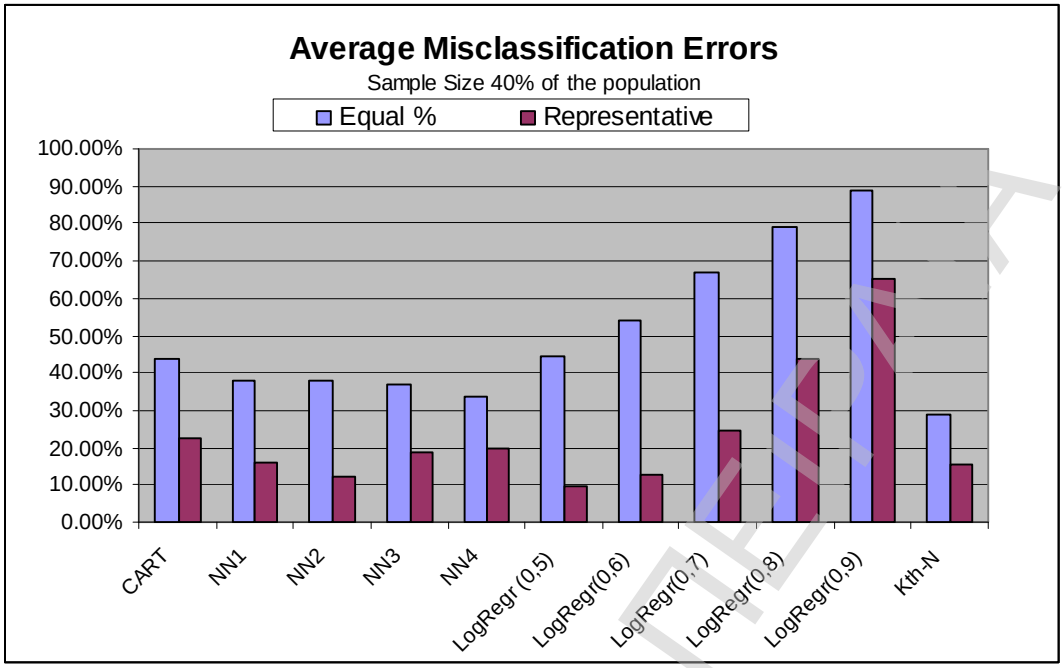


5.3. Συμπεράσματα.

Μετά την διενέργεια της μελέτης μας καταλήξαμε σε ορισμένα συμπεράσματα όσον αφορά τον τρόπο σχεδιασμού των δειγμάτων, το μέγεθος του δείγματος καθώς και της μεθόδου εκείνης της οποίας η ακρίβεια είναι μεγαλύτερη.

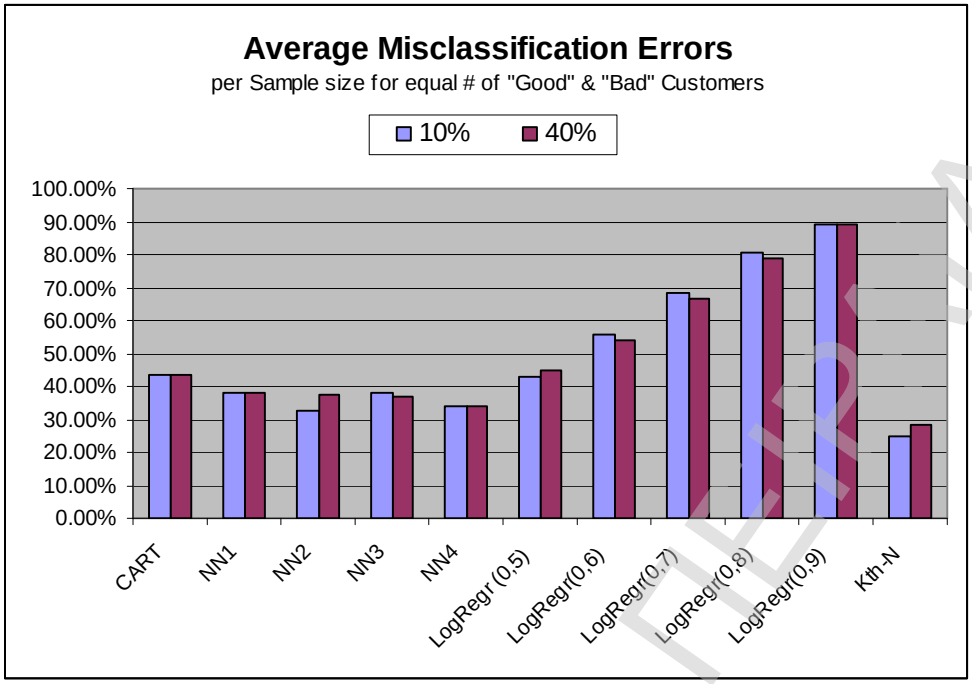
Το συμπέρασμα όσον αφορά τον τρόπο σχεδιασμού των δειγμάτων στο οποίο καταφέραμε να καταλήξουμε παρατηρώντας τα αποτελέσματα των υπολογισμών της μελέτης μας είναι ότι ακολουθώντας τον τρόπο σχεδιασμού δειγμάτων τα οποία είναι αντιπροσωπευτικά της σύνθεσης του αρχικού συνόλου προκύπτουν μοντέλα με καλύτερη προβλεπτική ικανότητα ανεξαρτήτως του μεγέθους τους. Συγκεκριμένα παρατηρούμε ότι για όλες τις μεθόδους σχεδιασμού μοντέλων τα μέσα σφάλματα ταξινόμησης είναι μικρότερα όταν τα δείγματα έχουν σχεδιαστεί με τέτοιων τρόπο έτσι ώστε το ποσοστό των 'ενήμερων' πελατών και των πελατών 'σε καθυστέρηση' να είναι ίσα με εκείνα του αρχικού συνόλου. Παρακάτω παραθέτουμε δύο διαγράμματα στα οποία συγκρίνουμε το μέσο σφάλμα για κάθε μια μέθοδο (με συγκεκριμένο μέγεθος δείγματος) όταν τα μοντέλα έχουν προκύψει από δείγματα με διαφορετικό τρόπο σχεδιασμού.

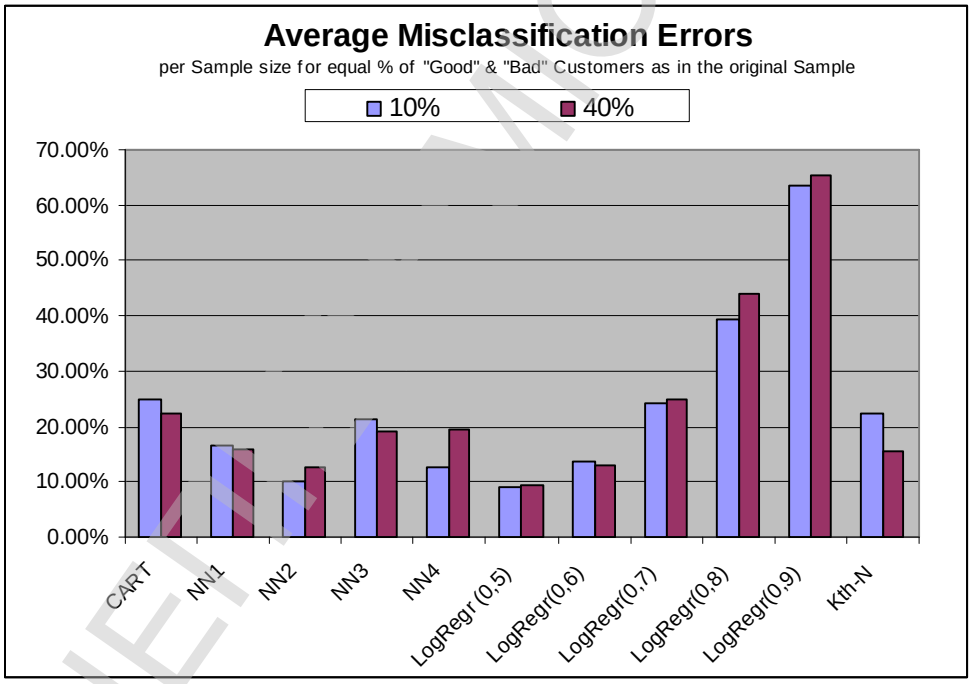




Όπως φαίνεται από τα παραπάνω διαγράμματα όλες οι μέθοδοι σχεδιασμού μοντέλων δίνουν πολύ καλύτερα μοντέλα, όσον αφορά την προβλεπτική τους ικανότητα με κριτήριο το μέσο σφάλμα ταξινόμησης , όταν χρησιμοποιούμε δείγματα τα οποία έχουν ίδια σύνθεση ως προς το αρχικό. Για παράδειγμα παρατηρούμε ότι για τα μοντέλα τα οποία έχουν προκύψει από την εφαρμογή της μεθόδου των δέντρων αποφάσεων σε δείγματα μεγέθους ίσο με το 40% του αρχικού συνόλου (1600 παρατηρήσεις) το μέσο σφάλμα ταξινόμησης είναι 22,23% όταν τα μοντέλα έχουν σχεδιαστεί με δεδομένα από αντιπροσωπευτικά δείγματα έναντι 43,89% όταν τα μοντέλα έχουν σχεδιαστεί με δεδομένα από δείγματα με ίσο αριθμό ‘ενήμερων’/ ‘σε καθυστέρηση’ πελατών.

Ένα άλλο συμπέρασμα το οποίο προέκυψε από την μελέτη μας είναι ότι το μέγεθος του δείγματος στο οποίο στηριζόμαστε για τον σχεδιασμό του μοντέλου μας δεν έχει σημαντικό ρόλο στην προβλεπτική ικανότητα των μοντέλων. Συγκεκριμένα παρατηρήσαμε ότι τα μοντέλα που προέκυψαν από τα δείγματα μεγέθους 10% του αρχικού συνόλου και τα δείγματα μεγέθους 40% του αρχικού συνόλου (400 και 1600 παρατηρήσεων αντίστοιχα) έδωσαν περίπου τα ίδια μέσα σφάλματα ταξινόμησης. Τα αποτελέσματα αυτά παρουσιάζονται στα παρακάτω διαγράμματα στα οποία απεικονίζονται τα μέσα σφάλματα ταξινόμησης για κάθε μέθοδο σχεδιασμού δειγμάτων που χρησιμοποιήσαμε για τα διαφορετικά μεγέθη δειγμάτων στα οποία στηριχθήκαμε για τον σχεδιασμό των μοντέλων μας.





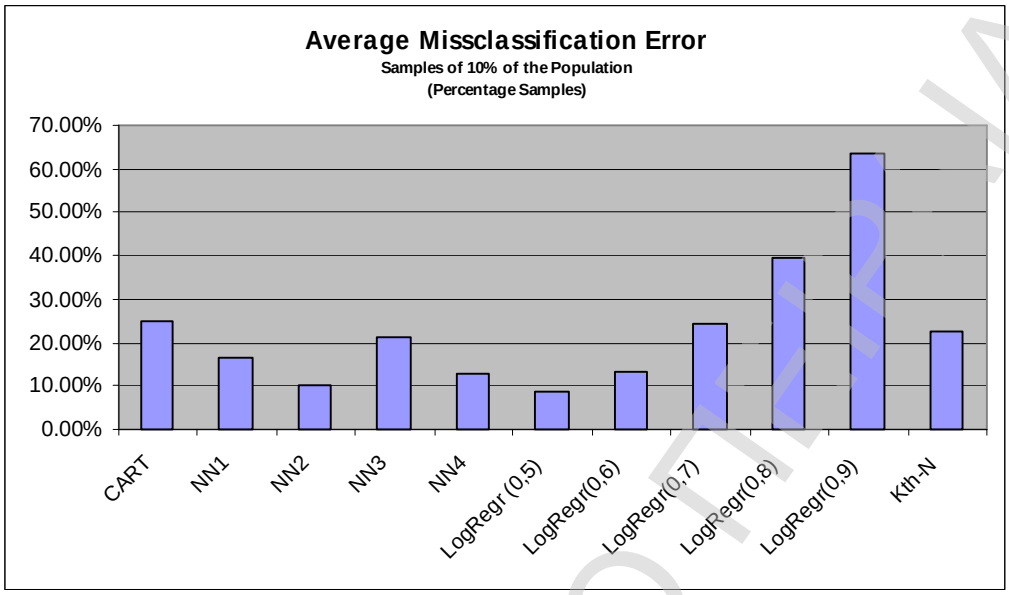
Σε μερικές μεθόδους, όπως τα νευρωνικά δίκτυα, δείγματα μικρού μεγέθους είναι προτιμότερα από δείγματα μεγάλου μεγέθους διότι οι μέθοδοι αυτοί προσαρμόζουν τα μοντέλα στα δεδομένα κατάρτισης με αποτέλεσμα να χάνεται η γενικότητα τους και η προβλεπτική τους ικανότητα να μειώνεται αισθητά όταν εφαρμόζονται σε άγνωστα προς αυτά δεδομένα όπως για παράδειγμα συμβαίνει για νευρωνικά δίκτυα με δύο κρυφά επίπεδα και cost function την Maximum Likelihood τα οποία έχουν εκπαιδευτεί σε 'αντιπροσωπευτικά' δείγματα. Παρατηρούμε ότι αυτά τα μοντέλα έχουν μέσο σφάλμα ταξινόμησης 12,75% όταν έχουν εκπαιδευτεί σε

δείγματα μεγέθους 10% του αρχικού συνόλου ενώ παρουσιάζουν μέσο σφάλμα ταξινόμησης 19,59% όταν έχουν εκπαιδευτεί σε δείγματα μεγέθους 10% του αρχικού συνόλου.

Σε πολλές περιπτώσεις τα δείγματα μεγάλου μεγέθους έχουν και άλλα αρνητικά αποτελέσματα λόγω της μεγάλης υπολογιστικής δύναμης που απαιτείται καθώς και του χρόνου που απαιτείται για την επεξεργασία τους.

Επομένως συνυπολογίζοντας τα συμπεράσματα που παραθέσαμε τα οποία προέκυψαν από την μελέτη μας όσον αφορά τον τρόπο σχεδιασμού του δείγματος καθώς και το μέγεθος του θα λέγαμε ότι ένας συνδυασμός δειγμάτων τα οποία είναι αντιπροσωπευτικά της σύνθεσης του αρχικού συνόλου και μεγέθους 10% του αρχικού συνόλου θα ήταν ο κατάλληλος για τον σχεδιασμό των μοντέλων μας ανεξαρτήτου μεθόδου. Συγκεκριμένα για τον συνδυασμό αυτό όπως είδαμε και προηγουμένως το μικρότερο μέσο σφάλμα ταξινόμησης για το συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα και χαρακτηριστικά που είχαμε στην διάθεσή μας και για τον συγκεκριμένο τρόπο σχεδιασμού των δειγμάτων προκύπτει από τα μοντέλα που σχεδιάστηκαν ακολουθώντας την μέθοδο της λογιστικής παλινδρόμησης με cut off point 0,5. Συγκεκριμένα η μέθοδος αυτή μας έδωσε μέσο σφάλμα 8,89% με δεύτερη καλύτερη την μέθοδο των νευρωνικών δικτύων όταν τα δίκτυα αυτά έχουν σχεδιαστεί με δύο κρυφά επίπεδα και η cost function είναι η Squared Error, με μέσο σφάλμα 10,27%. Πολύ καλή προβλεπτική ικανότητα έδειξαν και τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των νευρωνικών δικτύων με δύο κρυφά επίπεδα και cost function την Maximum Likelihood, με μέσο σφάλμα ταξινόμησης 12,75%, μία επίδοση που ήταν πολύ κοντά σε αυτή της λογιστικής παλινδρόμησης με cut off point to 0,6 (13,57%). Οι υπόλοιπες μέθοδοι έδωσαν ποιο φτωχά αποτελέσματα. Συγκεκριμένα τα νευρωνικά δίκτυα με ένα κρυφό επίπεδο και cost function την Squared Error έδωσε μέσο σφάλμα ταξινόμησης ίσο με 16,70% έναντι των νευρωνικών δικτύων με ένα κρυφό επίπεδο και cost function την Maximum Likelihood τα οποία έδωσαν μέσο σφάλμα ταξινόμησης ίσο με 21,41%. Τα δέντρα αποφάσεων και τα μοντέλα που προέκυψαν από την εφαρμογή της μεθόδου των κ-κοντινότερων γειτόνων έδωσαν μέσα σφάλματα 24,73% και 22,49% αντίστοιχα. Παρατηρούμε επίσης ότι αυξάνοντας το cut off point στην λογιστική παλινδρόμηση αυξάνεται το μέσο σφάλμα ταξινόμησης συγκεκριμένα για cut off point 0.7 , 0.8 , 0.9 έχουμε μέσο σφάλμα ταξινόμησης 24,25% , 39,46% και 63,57% αντίστοιχα.

Στο παρακάτω διάγραμμα παρουσιάζουμε το μέσο σφάλμα ταξινόμησης για την κάθε μέθοδο, που προέκυψε από την εφαρμογή της στα διαφορετικά δείγματα.



6. Επίλογος

Όπως έχει ήδη γίνει σαφές από τα αποτελέσματα της μελέτης μας, αλλά και από τα αποτελέσματα άλλων ερευνών μια συνολικά βέλτιστη μέθοδος δεν υπάρχει. Η καταλληλότητα κάθε μοντέλου κρίνεται ως προς το συγκεκριμένο πρόβλημα που αντιμετωπίζεται και με τα συγκεκριμένα δεδομένα που έχουμε στην διάθεσή μας, τόσο από πλευράς χαρακτηριστικών όσο και από πλευράς ποιότητας και ποσότητας.

Οι παράμετροι οι οποίοι επηρεάζουν μία μελέτη σχεδιασμού ενός συστήματος credit scoring είναι αρκετοί. Για παράδειγμα στο συγκεκριμένο πρόβλημα που είχαμε, δηλαδή της διερεύνησης της πιστοληπτικής ικανότητας υποψηφίων πελατών σε ένα συγκεκριμένο είδος πιστωτικής κάρτας, στον προσδιορισμό της καταλληλότερης μεθόδου ρόλο είχε η δομή και η ποιότητα των δεδομένων, τα υπό μελέτη χαρακτηριστικά ο τρόπος σχεδιασμού των δειγμάτων στα οποία θα στηριζόμασταν για την διεξαγωγή της ερευνάς μας, ο τρόπος αξιολόγησης της προβλεπτικής ικανότητας των μοντέλων καθώς επίσης και ο ορισμός της λάθος ταξινόμησης. Κατά την διάρκεια της μελέτης μας θεωρήσαμε ως λάθος ταξινόμηση τον χαρακτηρισμό ενός ‘ενήμερου’ πελάτη ως υποψήφιο ‘σε καθυστέρηση’ πελάτη καθώς και την ταξινόμηση ενός ‘σε καθυστέρηση’ πελάτη ως υποψήφιο ‘ενήμερο’. Βέβαια ο πιο σίγουρος τρόπος για να επιλεγεί η καλύτερη μέθοδος για μια συγκεκριμένη περίπτωση credit scoring είναι να μελετηθούν διαφορετικά μοντέλα και να συγκριθούν τα αποτελέσματά τους.

Πολύ σημαντικό στοιχείο σε μία μελέτη αυτού του είδους είναι η σύγκριση των μεθόδων στη βάση της διακριτής και προβλεπτικής τους ικανότητας η οποία αποτελεί μια καλή και συνήθης πρακτική. Πραγματικά, η δυνατότητα ενός μοντέλου να διαχωρίζει τις διαφορετικές κατηγορίες πιστωτικού κινδύνου σύμφωνα με τις διαθέσιμες μεταβλητές είναι ένα σημαντικό ζητούμενο του credit scoring και άρα οι δείκτες ταξινόμησης είναι ένα καλό κριτήριο για συγκρίσεις, το οποίο πάντα έχει σημαντική θέση στις συγκριτικές μελέτες εφαρμογών.

Ωστόσο, η ακρίβεια ταξινόμησης αποτελεί έναν μόνο συγκριτικό παράγοντα της καταλληλότητας των μεθόδων. Υπάρχουν ακόμα και πρέπει να ληφθούν υπόψη η ταχύτητα ταξινόμησης, η δυνατότητα αναπροσαρμογής των μοντέλων και η ευκολία στην κατανόηση των μεθόδων και των αποτελεσμάτων τους. Μία άμεση και εύκολη στην αιτιολόγησή της απόφαση είναι επιθυμητή τόσο στους υποψηφίους όσο και στους χρήστες των συστημάτων και αυτός είναι ένας λόγος που οι απλούστερες μέθοδοι, όπως η παλινδρόμηση ή τα δέντρα αποφάσεων, είναι συχνά προτιμότερες από τις δυσνόητες μεθόδους, όπως τα νευρωνικά δίκτυα.

Επίσης πολύ σημαντικό ρόλο στην αξιολόγηση των μοντέλων έχει και το σύνολο ελέγχου με τη χρήση του οποίου γίνεται ο έλεγχος των μοντέλων, δηλαδή

στην μελέτη μας ο υπολογισμός της ακρίβειας των μοντέλων. Πρέπει να ομολογήσουμε ότι θα αισθανόμασταν πολύ πιο σίγουροι για τα αποτελέσματα της μελέτης μας, την ακρίβεια και την καταλληλότητα των μοντέλων μας, αν τα μέσα σφάλματα ταξινόμησης είχαν προέλθει από τον έλεγχο των μοντέλων σε δεδομένα τα οποία δεν είχαν προέλθει από το αρχικό πληθυσμό.

Στα προβλήματα ταξινόμησης γενικά, η γνώση του αντικειμένου που πραγματεύεται το συγκεκριμένο πρόβλημα είναι ένας καθοριστικός παράγοντας στην επιλογή μεθόδου και πολλές φορές οι περιπτώσεις όπου δεν είναι πλήρως κατανοητή η δομή των δεδομένων παρουσιάζουν δυσκολία στο χειρισμό. Στον τομέα του credit scoring η κατανόηση του αντικειμένου βρίσκεται σε πολύ υψηλό επίπεδο, αφού για πολλά χρόνια κατασκευάζονται συστήματα ταξινόμησης με βάση παρόμοια δεδομένα. Σε αυτή την περίπτωση η βελτίωση των αποτελεσμάτων θα μπορούσε να επέλθει από έναν συνδυασμό μεθόδων από τον θα προέκυπταν μοντέλα με καλύτερη διακριτική και προβλεπτική ικανότητα.

Τέλος οφείλουμε να επισημάνουμε ότι στόχος αυτών των μοντέλων και οποιωνδήποτε μοντέλων τέτοιου τύπου, είναι να αποτελέσουν ένα κριτήριο για την λήψη δανειοδοτικών αποφάσεων αλλά δεν πρέπει να είναι το μοναδικό κριτήριο. Πρέπει να λαμβάνονται και άλλα στοιχεία υπόψη πριν την χορήγηση ή μη του δανείου.

7. Αναφορές

1. Breiman, L. Friedman, J.H., Olsen, R.A. and Stone, C.J., (1984): *Classification and Regression Trees.* Wadsworth International Group, Belmont.
2. Buchanan, B., (1989): *Can machine learning offer anything to expert systems?*
3. Chatterjee, S. and Barcum, S. (1970): *A nonparametric approach to credit screening.* J.Am. Statist. Ass., 65, 150-154.
4. Durand, D., (1941): *Risk Elements in Consumer Installment Financing. Studies in Consumer Installment Financing : Study 8.* National Bureau of Economic Research.
5. Efron, B. and Tibshirani, R., (1986): *Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy (with discussion).* Statist. Sience, 1, 54-77.
6. Efron, B. and Tibshirani, R., (1995): *Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule.* Technical Report 176, Dept of Statistics, Standford University.
7. Fisher, R.A., (1936). *The use of multiple measurement in taxonomic Problems.* Ann. Eugenics, 7, 179-188.
8. Gilbert, L.R., Menon, K. and Schwartz, K.B., (1990). *Predicting bankruptcy for firms in financial distress.* J.Bus. Finan. Account., 17(1), 161-171.
9. Greene, W. (1998). *Sample selection in credit scoring models.* Japan and World Economy, Vol. 10, No. 3, 299 – 316.
10. Highleyman, W., (1962). *The design and analysis of pattern recognition experiments.* Belt System Technical Journal, 41, 723-744.
11. Kohavi, R. (1995). *Wrappers for performance enhancement and oblivious decision graphs.*
12. Kong, E., B. and Dietterich T.G., (1995): *Error – correcting output coding corrects bias and variance.*
13. Lachenberg, P. and Mickey, M., (1982): *Estimation of Error Rates in Discriminant Analysis.*
14. Lyn C. Thomas, David D. Edelman, Jonathan N. Crook: *Credit Scoring and its applications.*
15. J. Galindo and P. Tamayo: *Credit Risk Assessment using Statistical and Machine Learning - Basic Methodology and Risk Modeling Applications*, February 12, 1998
16. McCullagh, P. and Nelder, J.A. (1989): *Generalized Linear Models.*
17. Shaffer, C., (1994): *A conservation law for generalization performance.* In: *Proceedings of the 11th International Conference on Machine Learning.* Morgan Kaufmann.
18. Stone, M., (1974): *Cross-Validatory Choice and Assesment of Statistical Predictions.* Journal of the Royal Statistical Society, 36, 111 -147.
19. Weiss, S. and Kulikowski, C., (1991). *Computer Systems that Learn.* Morgan Kaufmann.
20. Wiginton, J.C., (1980). *A note on the comparison of the logit and discriminant models of consumer credit behavior.* J. Finan. Quant. Anal. 15, 757-770.
21. Τσαντάς Νίκος, Μωυσιάδης Χρόνης, Μπαγιάτης Νίκος, Χατζηπαντελής Θόδωρος: *Ανάλυση Δεδομένων με τη βοήθεια στατιστικών πακέτων.*

Για την πραγματοποίηση της μελέτης μας χρησιμοποιήσαμε το εξής λογισμικό:

1. SQL Server
2. XL Miner
3. Analysis Services
4. SPSS
5. Excel