ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ & ΔΙΟΙΚΗΣΗΣ

ΤΟΜΕΑΣ ΕΠΙΣΤΗΜΗΣ ΑΠΟΦΑΣΕΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

'ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΕΚΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΩΝ ΠΡΟΒΛΕΨΗΣ ΑΠΟΧΩΡΗΣΗΣ ΠΕΛΑΤΩΝ'

ΞΑΝΘΙΠΠΗ ΛΕΜΟΝΤΖΟΓΛΟΥ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ : ΜΙΧΑΛΗΣ ΔΟΥΜΠΟΣ

ΧΑΝΙΑ 2015

Πίνακας Περιεχομένων

1 ΕΙΣΑΓΩΓΗ

Τα αναλυτικά εργαλεία επιχειρηματικών αποφάσεων (Business Analytics – BA) και η επιχειρηματική ευφυΐα (Business Inteligence-BI) αποτελούν σημείο των καιρών, ιδίως σε μια εποχή που ο όγκος δεδομένων έχει οδηγήσει στην ανάπτυξη του κλάδου big data analytics. Βασική πρόκληση να εξαχθεί αξιοποιήσιμη πληροφορία και γνώση εγκαίρως για την υποστήριξη στρατηγικών αποφάσεων της εκάστοτε επιχείρησης. Σύμφωνα με μελέτη της εταιρείας Gartner, αποτελεί πρωταρχική λειτουργία της επιχείρησης για υποστήριξη στη λήψη αποφάσεων για τους Γενικούς Διευθυντές Οικονομικών και Γενικούς Διευθυντές Υπηρεσιών Πληροφορικής [1].

Η παρούσα εργασία εξετάζει την πρόβλεψη αποχώρησης πελατών από την οπτική μιας εκ των διαστάσεων του BA, customer analytics. Χρησιμοποιείται ως εργαλείο για την υποστήριξη στρατηγικών αποφάσεων στο τμήμα διαχείρισης πελατειακών σχέσεων (CRM). Κύριο πρόβλημα στο συγκεκριμένο τμήμα, με σοβαρό κόστος για την επιχείρηση, αποτελεί η απώλεια πελατών [2].

Στα πλαίσια ανάπτυξης στρατηγικών διαχείρισης απώλειας πελατών (churn management), έχουν αναπτυχθεί μεθοδολογίες πρόβλεψης αποχώρησης πελατών (churn prediction). Βασικό ζητούμενο είναι η σωστή ταξινόμηση του πελάτη σε εν δυνάμει ή μη αποχωρήσαντα μέσα από την ανάπτυξη κατάλληλων μοντέλων, με τη χρήση κατάλληλων αλγορίθμων. Τα μοντέλα αναπτύσσονται με τη χρήση predictive analytics με στόχο την ταξινόμηση (classification), εργασία (task) της εξόρυξης δεδομένων (data mining). Οι αλγόριθμοι που υλοποιούνται ανήκουν στην κατηγορία αλγορίθμων μηχανικής εκμάθησης (machine learning algorithms).

Οπότε, με τη χρήση αλγορίθμων μηχανικής εκμάθησης και ζητούμενο την ταξινόμηση του πελάτη αναπτύσσονται μοντέλα πρόβλεψης αποχώρησης πελατών. Με τη χρήση κατάλληλων δεικτών αξιολόγησης των αποτελεσμάτων, πραγματοποιείται συγκριτική αξιολόγηση των αλγορίθμων. Επίσης, εξετάζονται τα χαρακτηριστικά ως προς το πλήθος τους και το είδος τους. Στόχος είναι, να εξαχθεί συμπέρασμα για το αν και γιατί οι αλγόριθμοι που έχουν επιλέγει στην παρούσα εργασία δουλεύουν αποτελεσματικά σε τέτοιου είδους σύνολα δεδομένων.

2 BUSINESS ANALYTICS (BA) ΚΑΙ BIG DATA ANALYTICS

Το BA αναφέρεται στο κομμάτι του analytics, όπου έπειτα από κατάλληλη επεξεργασία και ανάλυση ενός όγκου δεδομένων εξάγει χρήσιμη πληροφορία που οδηγεί σε χρήσιμα συμπεράσματα. Προσανατολισμός των συμπερασμάτων, η υποστήριξη λήψης στρατηγικών και μη αποφάσεων της επιχείρησης.[3] Για να το επιτύχει αυτό συνδυάζονται επιστήμες από διαφορετικούς χώρους, ενοποιώντας την χρήση της τεχνολογίας με τη λήψη αποφάσεων και ποσοτικές/στατιστικές μεθοδολογίες.[4] Από την παραλαβή των δεδομένων μέχρι και την εξαγωγή και ανάλυση της πληροφορίας, υπόκειται στο κομμάτι της επιχειρηματικής ευφυίας.

Η χρηστικότητα προκύπτει τόσο από την πληροφόρηση και γνώση που παρέχει το BA, όσο και από το χρονικό ορίζοντα που παρέχεται. Εισαγάγετε η παράμετρος ευαισθησίας χρόνου (time – sensitivity), καθώς η γνώση έχει αξία όταν είναι αξιοποίησημη στο παρόν, ώστε να μπορέσει η επιχείρηση να προχωρήσει στις ανάλογες δράσεις, χωρίς να είναι πλέον ξεπερασμένες. Σε διαφορετική περίπτωση οδηγείται σε λανθασμένες αποφάσεις και ζημία. Στα πλαίσια αυτού αναπτύσσεται και το Operational BA, για την λήψη καθημερινών επιχειρησιακών αποφάσεων. Για να το επιτύχει αυτό αναπτύσσει μεθοδολογίες που μειώνουν την αναποτελεσματικότητα των αντίστοιχων διαδικασιών στη λήψη αυτών των αποφάσεων.[5]

Το BA προϋπήρχε εργασιακά και ακαδημαϊκά, αλλά δεν μελετιούνταν ως ξεχωριστός κλάδος μέχρι προσφάτως. Οι ρίζες του προέρχονται από την επιχειρησιακή έρευνα και την επιστήμη διοίκησης. Ξεκίνησε στην δεκαετία το 1920 από το κλάδο της επιστήμης διοίκησης, εξελίχθηκε ως τμήμα των συστημάτων υποστήριξης αποφάσεων το 1970, για να φτάσει να είναι κομμάτι της επιχειρηματικής ευφυίας από το 1980. Από το 2000 και έπειτα άρχισε να αναπτύσσεται το κομμάτι των analytics και να γίνεται αναφορά στο BA ως διακριτός κλάδος. [4]

Με την ραγδαία ανάπτυξη των εργαλείων της επιχειρηματικής ευφυίας και της δυνατότητας συλλογής δεδομένων, παρατηρήθηκε συλλογή μεγάλου όγκου δεδομένου προς επεξεργασία. Αυτό δημιούργησε επιπρόσθετες ανάγκες και προβλήματα προς επίλυση για την εξαγωγή ασφαλών συμπερασμάτων μέσω του BA. Ως αποτέλεσμα αυτού, αναπτύχθηκε ένας διακριτός κλάδος, το Big Data Analytics. Πάραυτα συνεχίζει να αποτελεί τμήμα του BA και αρκετές φορές μελετάται στα πλαίσια αυτού[6].

Το BA, πέραν του ότι περιλαμβάνει τα πεδία του Big Data Analytics και της επιχειρηματικής ευφυίας, μπορεί να διακριτοποιηθεί σε κατηγορίες ως προς τα πεδία δεδομένων προς ανάλυση (Πίνακας 1). Επίσης, χωρίζεται σε διαστάσεις, ως προς τον προσανατολισμό των συμπερασμάτων[3]. Ανάλογα τις σχέσεις που θέλουμε να εξαγάγουμε, χρησιμοποιούνται μεθοδολογίες ανάλογα την διαστάσεις, που βασίζονται σε μαθηματικά εργαλεία και αλγόριθμους μηχανικής εκπαίδευσης. Ως αποτέλεσμα οι διαστάσεις αυτές παρέχουν είτε περιγραφή/σύνοψη των δεδομένων, ώστε να εξαχθεί πληροφορία, (descriptive analytics) είτε προβλέψεις (predictive analytics). Μια εναλλακτική άποψη σχετικά με τα predictive analytics, σύμφωνα με το άρθρο “Smarter Data”[7] του Dr Michael Wu από την Lithium Technologies, είναι ότι με τα δεδομένα που ήδη έχουν συλλεχθεί εξάγεται συμπέρασμα για δεδομένα που υπολείπονται. Επίσης, μια τρίτη

διάσταση, όπως περιγράφεται από τον Bill Vorhies, Πρόεδρο και Διευθύνων Επιστήμονα Δεδομένων της Data-Magnum, είναι η βελτίωση της πρόβλεψης με εργαλεία βελτιστοποίησης ώστε αντί του τι θα συμβεί (predictive analytics) να περιγράφει τι θα έπρεπε να συμβεί (prescriptive analytics)[8]. Στη παρούσα εργασία, η ανάλυση δεδομένων προσανατολίζεται στην ανάπτυξη μοντέλου πρόβλεψης (διάσταση predictive analytics) ως προς τη συμπεριφορά των πελατών (πεδίο customer analytics).

Πίνακας 1: Περιγραφή πεδίων Business Analytics

ΠΕΔΙΟ ANALYTICS	**ΠΕΡΙΓΡΑΦΗ**
Web analytics	Παρέχει δυνατότητα να καταγραφεί η κινητικότητα μια ιστοσελίδας και ενός συνδέσμου. Επίσης, καταγράφεται από ποιο “κανάλι” προώθησης μια ενέργειας (διαδρομή) κατέληξε στην εκάστοτε ιστοσελίδα/σύνδεσμο ο χρήστης. Εξυπηρετεί στη μέτρηση -ποσοτικοποίηση των αποτελεσμάτων στρατηγικών προώθησης.[9]
Google analytics	Καταγραφή διαδικτυακής χρήσης ιστοσελίδας, στο τέλος του κύκλου χρήσης της εκάστοτε ιστοσελίδας. Λόγω καταγραφής από την Google, θεωρείται ότι παρέχει πιο αξιόπιστη καταγραφή της συμπεριφοράς του χρήστη.[10]
Software analytics	Συνδέει πολλαπλά παραδοτέα (artifacts) λογισμικού που έχουν υποστεί εξόρυξη δεδομένων, με στόχο τη λήψη αποφάσεων σε διάφορες φάσεις του κύκλου ζωής του λογισμικού.[11]
Learning &Knowledge analytics	Η αξία του προκύπτει από το γεγονός ότι μπορεί να παρέχει στον εκπαιδευόμενο εξατομικευμένες προτάσεις για κάλυψη γνωστικών κενών στο πεδίο αναζήτησης -ενδιαφέροντος του, μέσω εργαλείων που παρέχονται από το κομμάτι των analytics. [12]
Marketing analytics	Γίνεται χρήση του customer analytics με προσανατολισμό στις στρατηγικές marketing της επιχείρησης.[13]
Customer (CRM) analytics	Αναφέρεται στη συμπεριφορική ανάλυση των πελατών για την κατηγοριοποίηση τους ή την πρόβλεψη δράσης τους.[14]
Service analytics	Χρησιμοποιείται ως εργαλείο για τη ποσοτικοποίηση-μέτρηση χαρακτηριστικών, όπως συνδυασμό γνώσης ανθρωπίνων πόρων, λήψη αποφάσεων σχετικά με συστήματα υπηρεσιών και τον υπολογισμό του αντίστοιχου λογιστικού κόστους.[15]
Human resource analytics	Σύνδεση της απόδοσης του προσωπικού και των δεξιοτήτων τους, μέσα από δεδομένα, με στόχο τη μέτρηση της επίδρασης που έχει στην απόδοση της εταιρείας. Επίσης, χρησιμοποιείται και ως πηγή πληροφόρησης για τη λήψη αντίστοιχων στρατηγικών αποφάσεων για το τμήμα ανθρώπινου δυναμικού. [16]
Talent analytics	Θεωρούμενο από κάποιους ως υποκατηγορία του HR Analytics, εστιάζει στη διαχείριση ταλέντων της επιχείρησης, με στόχο την αποτελεσματική προσέλκυση και διαχείριση τους και τη λήψη αντίστοιχων αποφάσεων.[17]
Process analytics	Παράσχει στους ιδιοκτήτες, τους εμπλεκόμενους με τον καθορισμό των επιχειρησιακών διαδικασιών και σε όσους βρίσκονται σε αντίστοιχα κέντρα λήψης αποφάσεων , πληροφορίες για την επάρκεια και αποτελεσματικότητα των διαδικασιών του εκάστοτε οργανισμού.[18]
Sypply chain analytics	Συνδέει αποτελεσματικά , μέσω analytics, την προσφορά με τη ζήτηση, συναρτήσει της εφοδιαστικής αλυσίδας.[19]
Risk analytics	Αναφέρεται τόσο στη διαχείριση του χρηματοοικονομικού ρίσκου , όσο και του ρίσκου συσχετισμένου με τις επιχειρησιακές λειτουργίες, τις αντίστοιχες στρατηγικές αποφάσεις και τις τάσεις της αγοράς. Χρησιμοποιείται επίσης ως εργαλείο για την πρόβλεψη, διαχείριση και αποφυγή μελλοντικών κρίσεων (crisis management)[20]
Financial analytics	Χρησιμοποιείται για την λήψη χρηματοοικονομικών αποφάσεων και σχετικών με επενδύσεις καθώς και την εκτίμηση και έλεγχο του αντίστοιχου ρίσκου.[21]

2.1 ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΙΑ (BUSINESS INTELIGENCE- BI)

Η επιχειρηματική ευφυΐα ως εργαλείο του BA, χρησιμοποιεί ιστορικά δεδομένα για την εξαγωγή πληροφορίας και γνώσης, που μπορεί να αξιοποιηθεί από την επιχείρηση στην υποστήριξη λήψης αποφάσεων, με χρήση της επιστήμης δεδομένων (data science)[22]. Για να επιτευχθεί αυτό, η διαδικασία εφαρμογής της ξεκινάει από τη συλλογή και επεξεργασία καλών ποιοτικά δεδομένων. Προκειμένου να χαρακτηριστούν 'καλά', η επίβλεψη και επεξεργασία τους ξεκινάει από την παραλαβή μέχρι και το σημείο που θα χρησιμοποιηθούν ως είσοδο για την εξαγωγή πληροφορίας. [23]

Για να εξαχθεί πληροφορία πραγματοποιείται εξόρυξη δεδομένων (data mining). Στην ουσία χρησιμοποιούνται αλγόριθμοι που έχουν ως είσοδο σύνολα 'καλών' δεδομένων και δίνουν, ανάλογα το ζητούμενο, ως έξοδο την αντίστοιχη πληροφορία. Η πληροφορία μετέπειτα με κατάλληλη επεξεργασία και ερμηνεία μπορεί να οδηγήσει σε γνώση και συμπεράσματα.[24] Στη συγκεκριμένη εργασία ζητούμενο είναι η ταξινόμηση με χρήση predictive analytics μέσω αλγορίθμων μηχανικής εκμάθησης για την εξόρυξη των δεδομένων.

2.1.1 ΑΠΟ ΤΗΝ ΠΑΡΑΛΑΒΗ ΜΕΧΡΙ ΤΗΝ ΠΑΡΑΔΟΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Στην ενότητα αυτή θα γίνει ανάλυση των σταδίων από τα οποία περνάνε τα δεδομένα ώστε να καταλήξουν να είναι επεξεργάσιμα στη φάση της εξόρυξης δεδομένων. Τα κύρια στάδια είναι δύο, το πρώτο συλλογή και ανάκληση και το δεύτερο επεξεργασία. Όπως θα διαπιστωθεί, είναι ένα απαιτητικό κομμάτι που παίζει καθοριστικό ρόλο στο αν και τι πληροφορία θα εξαχθεί τελικά με τη χρήση της εξόρυξης δεδομένων.[25]

2.1.1.1 ΔΙΑΔΡΟΜΗ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα αρχικά κατατάσσονται, σε εσωτερικά ή εξωτερικά , ως προς τη πηγή προέλευσης τους. Τα εσωτερικά δεδομένα (internal data) προέρχονται από συστήματα διαχείρισης της εταιρείας και ενδοεπιχειρησιακές πηγές δεδομένων. Ενδεικτικά συστήματα είναι οι εφαρμογές διαχείρισης πελατειακών σχέσεων (CRM software) καθώς και τα συστήματα ενδοεπιχειρησιακού σχεδιασμού (ERP). Τα εξωτερικά προέρχονται από πηγές πέραν της εταιρείας και χωρίζονται σε δύο κατηγορίες, πρωτεύοντα και δευτερεύοντα. Τα πρωτεύοντα συλλέγονται από την εταιρεία άμεσα πηγές συνδεόμενες με την επιχείρηση, ενδεικτικά από έρευνα ικανοποίησης πελατών της επιχείρησής. Τα δευτερεύοντα είναι ήδη συγκεντρωμένα ή δημοσιευμένα δεδομένα από τρίτους, όπως ήδη δημοσιευμένες έρευνες αγοράς. Πρόκληση αποτελεί η αποτελεσματική ενοποίηση των δεδομένων από όλες τις πηγές.[26]

Πέραν της προέλευσης τους, μια άλλη κατηγοριοποίηση έγκειται στη μορφή τους. Αρχικά, η απεικόνιση των τιμών των χαρακτηριστικών/μεταβλητών, για τα οποία έχουν συλλεχθεί τα δεδομένα, μπορεί να είναι αλφαριθμητική ή αριθμητική. Με κατάλληλη τυποποίηση τα αλφαριθμητικά πεδία αντικαθίστανται από αριθμητικές τιμές. Ανάλογα την μοντελοποίηση αυτών των τιμών, μπορεί να απεικονίζουν ποιοτικές ή ποσοτικές μεταβλητές. Οι ποσοτικές είναι άμεσα μετρήσιμες και εκφράζουν ποσοτικά μεγέθη, είναι είτε συνεχείς είτε διακριτές. Οι ποιοτικές

εκφράζουν ονομαστικά δεδομένα, που τους έχουν αποδοθεί αριθμητικές τιμές. Μπορεί να είναι είτε κατηγορικές ή μη, αναλόγως αν εκφράζουν ομαδοποίηση ή αν ιεραρχούνται οι τιμές τους, κατά αντιστοιχία.[27]

Η φύλαξη των εσωτερικών δεδομένων γίνεται σε βάσεις δεδομένων. Παλαιότερα η φόρτωση των δεδομένων και τα αντίστοιχα ερωτήματα (queries), γίνονταν απευθείας από τις βάσεις δεδομένων. Με την απαίτηση όμως για δεδομένα πραγματικού χρόνου, όπου ο υπολογιστικός φόρτος δεν επέτρεπε, την αύξηση της πολυπλοκότητας τους και των αδόμητων δεδομένων, οι βάσεις δεδομένων δεν ήταν πια αποτελεσματικές. Παρουσιάστηκε η ανάγκη για επίλυση των παραπάνω και αποτελεσματική ενοποίηση εσωτερικών και εξωτερικών δεδομένων [28]

Στα πλαίσια αυτού αναπτύχθηκαν οι αποθήκες δεδομένων (data warehouses), όπου τραβάνε δεδομένα από τις βάσεις δεδομένων και ενοποιούν εσωτερικά με εξωτερικά δεδομένα. Οργανώνονται με τέτοιο τρόπο ώστε να μπορούν να τα διαχειρισθούν μετέπειτα οι αναλυτές. Η πρόσβαση στα δεδομένα και φόρτωση των απαραίτητων για ανάλυση, μέσω πολυδιάστατων ερωτημάτων, γίνεται μέσω της απευθείας αναλυτικής διαδικασίας (OLAP).[29] Τα δεδομένα της αποθήκης δεδομένων μπορούν να ανακατανεμηθούν και να διασυνδεθούν με μικρότερες αποθήκες (data marts) με ανάθεση καθεμιάς σε ένα τμήμα της επιχείρησης, προσαρμοσμένης με περιεχόμενο σχετικό με τις ανάγκες του αντίστοιχου τμήματος.[30]

Τα παραπάνω πραγματοποιούνται με τη χρήση εργαλείων λογισμικού που αυτοματοποιούν τις βασικές λειτουργίες εξαγωγής, μετασχηματισμού και φόρτωσης δεδομένων (ETL). Ξεκινώντας από την εξαγωγή των δεδομένων από τις αντίστοιχες πηγές, μετατρέπονται σε κατάλληλη μορφή μέσω των διαδικασιών που περιγράφηκαν, ώστε να φορτωθούν στις αποθήκες δεδομένων.[31] Από το σημείο αυτό, και έπειτα μπορεί να γίνει κατάλληλη προετοιμασία των δεδομένων και επιλογή μεταβλητών για να πραγματοποιηθεί εξόρυξη δεδομένων και να εξαχθεί πληροφορία.

2.1.1.2 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

Από την εξαγωγή των δεδομένων από την αποθήκη μέχρι και την είσοδο τους στον αντίστοιχο αλγόριθμο ή εφαρμογή, απαιτείται περαιτέρω επεξεργασία τους. Παραλαμβάνοντας τα δεδομένα από την αποθήκη δεδομένων, γίνεται προεπεξεργασία με διόρθωση των ανεπαρκών δεδομένων και επιλογή κατάλληλων μεταβλητών, με χρήση αντίστοιχων μεθοδολογιών. Έπειτα γίνεται μετασχηματισμός των δεδομένων όπου απαιτείται, για να πάρουν την κατάλληλη μορφή για τη χρήση τους από τον αντίστοιχο αλγόριθμο. Ωστόσο, έχουν δημιουργηθεί δύο σύνολα δεδομένων, που περιέχουν τις ίδιες μεταβλητές και έχουν υποστεί την ίδια επεξεργασία, το σύνολο εκπαίδευσης και το σύνολο ελέγχου.[32]

Η διαδικασία ξεκινάει με τον εντοπισμό των δεδομένων που υπολείπονται. Μπορεί να υπολείπονται είτε γιατί παραμελήθηκε να συμπληρωθεί είτε γιατί δεν υπήρχε επιλογή που να περιγράφει την τιμή που αντιστοιχούσε σε αυτό πεδίο. Σε αυτό το σημείο της επεξεργασίας είναι δύσκολο να εντοπιστεί σε ποια κατηγορία ανήκει, οπότε αντιμετωπίζονται όλα τα κενά πεδία ως δεδομένα που παραλήφθηκε να συμπληρωθούν. Κατά περίπτωση, επιλέγεται είτε να συμπληρωθεί αυτό το πεδίο μεσώ κάποιας μεθοδολογίας είτε να υπάρξει διαγραφή του αποφασίζοντα ή της μεταβλητής, υπό κάποιες προϋποθέσεις.[33]

Η διαγραφή μεταβλητών, γενικά αποφεύγεται, καθώς μπορεί να αλλοιώσει τα αποτελέσματα που θα προέκυπταν από το εκάστοτε σύνολο δεδομένων. Πάραυτα, όταν με χρήση κατάλληλων μεθοδολογιών ή βιβλιογραφικά, δεν προκύπτει σημαντική συσχέτιση της εκάστοτε μεταβλητής με το αποτέλεσμα του σεναρίου υπό μελέτη και υπάρχουν κενά πεδία κάτω κάποιου ορίου που έχει προκαθοριστεί, διαγράφεται. Επίσης, σε περιπτώσεις που οι κενές τιμές στα πεδία που αναλογούν σε ένα αποφασίζοντα ξεπερνάνε κάποιο προκαθορισμένο όριο, τότε αφαιρείται από το σύνολο δεδομένων.[34]

Στη συνέχεια τα εναπομείναντα κενά πεδία γεμίζονται με τιμές που προκύπτουν από συγκεκριμένες μεθοδολογίες. Έχει αξία να υπάρχει αποθηκευμένη η πληροφορία για την έλλειψη συμπλήρωσης πεδίων, καθώς μπορεί στα αποτελέσματα να προσφέρει χρήσιμα συμπεράσματα. Αρχικά μετράται η συσχέτιση των μεταβλητών, ως προς το τελικό αποτέλεσμα και τον αποφασίζοντα, για να εκτιμηθεί η επίδραση της τιμής που θα αποδοθεί στο κενό πεδίο,. Έπειτα, με βάση τα αποτελέσματα της συσχέτισης επιλέγεται ποιο μέτρο εκτίμησης της μεταβλητότητας των τιμών (διανύσματος) της μεταβλητής, θα χρησιμοποιηθεί για να μην την επηρεάσει (ως προς την αντίστοιχο μέτρο μεταβλητότητας).[32]

Εν συνέχεια, στα πλαίσια του δυνατού, μειώνεται ο όγκος δεδομένων, με στόχο την μείωση της πολυπλοκότητας, του υπολογιστικού φόρτου και του θορύβου. Είναι τακτική που επιλέγεται στον κλάδο του big data analytics, με στόχο την βελτιστοποίηση της γνώσης συναρτήσει του υπολογιστικού φόρτου. Ο όγκος δεδομένων αρχικά μειώνεται με την ομαδοποίηση (clustering) μεταβλητών που παρουσιάζουν ομοιότητες (similarity) μέσω των αντίστοιχων αλγορίθμων. Όπου εκτιμάται ότι μπορούν να συμψηφιστούν μεταβλητές χωρίς επίδραση στο τελικό αποτέλεσμα, δημιουργείται μια νέα μεταβλητή που τις εμπεριέχει. Επίσης, μπορεί να χρειαστεί κανονικοποίηση ή επεξεργασία της απεικόνισης των τιμών για να μειωθεί ο όγκος δεδομένων. .[35]

Μετέπειτα, πραγματοποιούνται περαιτέρω στατιστικοί έλεγχοι, για την εκτίμηση της συσχέτισης των μεταβλητών με το ζητούμενο, στην περίπτωση της παρούσας εργασίας, της ταξινόμησης στη κλάση. Ενδεικτικά κάποιοι στατιστικοί έλεγχοι που χρησιμοποιούνται παρουσιάζονται παρακάτω, όπου αναφέρονται ως μεθοδολογίες επιλογής μεταβλητών. Οι παρόντες είναι αυτοί που επιλέχθηκαν να χρησιμοποιηθούν στα πλαίσια αυτής της εργασίας.

Οι μεθοδολογίες διαλογής μεταβλητών χωρίζονται σε δύο κατηγορίες, στις μεθόδους που αναφέρονται ως wrapper και τις μεθόδους φιλτραρίσματος (filter).[36] Οι μεθοδολογίες που εντάσσονται στις wrapper, στα πλαίσια υλοποίησης τους ταξινομούν τις μεταβλητές σε σημαντικές ή μη. Η ταξινόμηση (classification) γίνεται με τη χρήση του αλγόριθμου που θα χρησιμοποιηθεί μετέπειτα για την εξόρυξη πληροφορίας από τα δεδομένα.[37] Παρότι αποτελεσματικές οι μεθοδολογίες wrapper, καθώς το σύνολο δεδομένων που εξάγεται είναι προσαρμοσμένο στις ανάγκες του εκάστοτε αλγόριθμου, έχει σημαντικό υπολογιστικό κόστος. Στους γενετικούς αλγόριθμους αυτή η διαδικασία πραγματοποιείται εσωτερικά.

Από την άλλη, οι μεθοδολογίες φιλτραρίσματος χρησιμοποιούν στατιστικές μεθόδους για την αξιολόγηση της συσχέτισης των μεταβλητών με την κλάση που επιλέγει να ενταχθεί ο αποφασίζοντας. Οι περισσότερες παρέχουν ως έξοδο απλά ταξινόμηση των μεταβλητών, οπότε συνδυάζονται με κάποιες μεθόδους αναζήτησης (search methods) για την επιλογή των μεταβλητών. Αντίθετα με τις wrapper , το σύνολο δεδομένων που εξάγεται είναι ανεξάρτητο του της μεθόδου ταξινόμησης, οπότε υλοποιείται μια φορά, πριν ξεκινήσει η χρήση των αντίστοιχων αλγορίθμων.[36]

Οι στατιστικοί έλεγχοι που πραγματοποιήθηκαν είναι ο X^2 , ο gain ratio και ο information gain. Στα πλαίσια του ελέγχου, εξετάζουν την πιθανότητα να συμβεί ένα γεγονός Α υπό την προϋπόθεση ότι έχει συμβεί το γεγονός Β, δηλαδή την εξάρτηση τους. Στην περίπτωση μας υπό την προϋπόθεση ότι συμβαίνει ή έχει λάβει συγκεκριμένη τιμή η μεταβλητή υπό εξέταση, εκτιμάται η πιθανότητα να λάβει συγκεκριμένη τιμή η κλάση. Επίσης, χρησιμοποιείται η εκτίμηση της εντροπίας σε κάποιες από τις μεθόδους. Η εντροπία εκφράζει στην περίπτωση μας την πληροφορία που περιέχει η πραγματοποίησης ενός γεγονότος υπό την προϋπόθεση ΄ότι συνέβη ένα άλλο. Όσο λιγότερη γνώση ή προφανής συσχέτιση υπάρχει γύρω από ένα γεγονός, τόσο μεγαλύτερη η τιμή της εντροπίας, καθώς μας παρέχει περισσότερη σχετική πληροφορία και εκφράζεται από τους παρακάτω τύπους. [38]

$$H(B)=E(-\ln(P(B))) \quad (1) \text{ και} \quad H(A/B)=\sum_{i,j} p(a_i,b_i)\log\left(\frac{p(b_i)}{p(a_i,b_i)}\right) \quad (2)$$

Ο έλεγχος X^2 test, χρησιμοποιείται για να εκφράσει τον βαθμό ανεξαρτησίας δύο γεγονότων, μέσω εργαλείων της X^2 κατανομής. Υπολογίζεται η αντίστοιχη στατιστική για τις δύο μεταβλητές και έπειτα υπολογίζονται οι αντίστοιχοι βαθμοί ελευθερίας. Επιλέγεται τιμή για τη εκτίμηση του διαστήματος εμπιστοσύνης (confidence), με επιλογή το αντίστοιχου εργαλείου μέτρησης, και υπολογίζεται για τις τιμές που αναφέρθηκαν παραπάνω, με χρήση διπλής κατευθύνσεως στην περίπτωση μας. Με βάση την τιμή που έχει προκαθοριστεί, αφαιρούνται από το σύνολο δεδομένων οι μεταβλητές που φαίνεται να μην συμμετέχουν στην τελική απόφαση του πελάτη.[39]

Η μεθοδολογία information gain χρησιμοποιεί ως εργαλείο τη εκτίμηση της ωφέλιμης πληροφορίας που δίνει η αντίστοιχη μεταβλητή για τη λήψη απόφασης του πελάτη. Μαθηματικά αυτό εκφράζεται με τη χρήση της εντροπίας, από τον τύπο:

$$IG(A,B)=H(A)-H(A/B) \quad (3)$$

Με χρήση της ταξινόμησης που προκύπτει και χρήση είτε κάποιας μεθόδου εύρεσης, είτε ορισμού ορίου ωφέλιμης πληροφορίας ή συγκεκριμένη επιλογή πλήθος μεταβλητών, που υποστηρίζεται βιβλιογραφικά, αποφασίζεται ποιες μεταβλητές θα χρησιμοποιηθούν παρακάτω. [40]

Βασικό μειονέκτημα αυτής της μεθόδου είναι ότι μπορεί να οδηγήσει σε επιλογή μεταβλητών που μαθηματικά παρέχουν πληροφορία αλλά λογικά δεν παρέχουν χρήσιμη πληροφορία.[36] Για να επιλυθεί αυτό, εισήχθη η μεθοδολογία information gain ratio, όπου υπολογίζεται από τον τύπο:

$$IGR(A,B)=\frac{H(A)-H(A/B)}{H(B)} \quad (4)$$

Η τελική επιλογή των μεταβλητών προς χρήση γίνεται όπως και στη μεθοδολογία information gain.

Τα δεδομένα έπειτα χωρίζονται σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Κάποιες φορές απαιτείται και η δημιουργία ενός τρίτου συνόλου (validation set), για τον έλεγχο στη διαδικασία εκπαίδευσης κατά την εκτέλεση του αλγορίθμου. Το μέγεθος τους καθορίζεται τόσο εμπειρικά όσο και βιβλιογραφικά. Σε περιπτώσεις big data δημιουργείται ένα υποσύνολο δεδομένων (sample size) , για τη δημιουργία του συνόλου εκπαίδευσης και ελέγχου, για τη μείωση του υπολογιστικού φόρτου και του θορύβου. Η επιλογή του δείγματος γίνεται μέσω συγκεκριμένων μεθοδολογιών για να μην προκύψουν προβλήματα κατά την εκπαίδευση από τον αντίστοιχο αλγόριθμο μηχανικής εκμάθησης, που οδηγούν σε λανθασμένα αποτελέσματα.[41]

Προβλήματα που μπορεί να προκύψουν κατά την επιλογή των δεδομένων για τη δημιουργία του συνόλου εκπαίδευσης, είναι το overfitting και το underfitting. Τα προβλήματα αυτά προκύπτουν όταν το δείγμα/σύνολο εκπαίδευσης δεν περιέχει ισορροπημένο ποσοστό των αντίστοιχων τιμών στις κλάσεις. Παρατηρείται είτε υπερπροσαρμογή είτε το αντίθετο, ως προς την αντίστοιχη κλάση όπου παρατηρείται ανισορροπία. Για το λόγο αυτό επιλέγεται συχνά να δημιουργηθεί το λεγόμενο σύνολο εκπαίδευσης 50-50 ως προς τις κλάσεις, όταν υπάρχουν δύο. Ουσιαστικά επιλέγεται 50% του δείγματος να αντιπροσωπεύει, στη περίπτωση πρόβλεψης αποχώρησης πελατών, αποχωρήσαντες και το υπόλοιπο 50% μη αποχωρήσαντες. Αυτή είναι η πιο απλή προσέγγιση η οποία με τη σειρά της μπορεί να δημιουργήσει προβλήματα, ιδίως σε δείγματα που η πραγματικότητα δεν αντιπροσωπεύεται από ισορροπημένα δείγματα.[42]

Τέλος, μπορεί να απαιτηθεί μετασχηματισμός δεδομένων για την κατάλληλη είσοδο τους στο αντίστοιχο αλγόριθμό ή εφαρμογή. Ως παράδειγμα μπορεί να χρησιμοποιηθεί η εισαγωγή των συνόλων στο πρόγραμμα υλοποίησης αλγορίθμων Weka. Τα σύνλα ήταν διαθέσιμα σε μορφή συμβατή με το πρόγραμμα επεξεργασίας excel. Μετατράπηκαν σε αρχείο csv για να εισαχθούν στο πρόγραμμα Weka. Έπειτα μετατράπηκαν, σε αρχείο arrf, μέσω του Weka, για να μπορούν να φορτωθούν από τους αντίστοιχους αλγόριθμους. Παρότι είχε γίνει κατάλληλη τροποποίηση των δεδομένων, χρειάστηκε να δηλωθεί ποιες μεταβλητές ήταν ποιοτικές και να υπάρξει κατάλληλος μετασχηματισμός των τιμών τους. Αυτή η διαδικασία ακολουθήθηκε τόσο στο σύνολο εκπαίδευσης

όσο και στο σύνολο ελέγχου για να ξεκινήσει η υλοποίηση των αλγορίθμων μηχανικής εκμάθησης στο αντίστοιχο περιβάλλον υλοποίησης.

2.1.2 ΑΠΟ ΤΑ ΔΕΔΟΜΕΝΑ ΣΤΗ ΠΛΗΡΟΦΟΡΙΑ

Έχοντας, πλέον μοντελοποιήσει κατάλληλα τα δεδομένα και δημιουργήσει δύο σύνολα δεδομένων, στόχος είναι να χρησιμοποιηθούν τα κατάλληλα εργαλεία ώστε να εξαχθεί πληροφορία. Η πληροφορία αυτή μπορεί να εξαχθεί με τη χρήση εξόρυξης δεδομένων (data mining). Η εξόρυξη δεδομένων αποτελεί ενοποίηση δύο κλάδων, της πολυκριτήρια και υπολογιστικής στατιστικής και της μηχανικής εκμάθησης (machine learning). Στόχος, είναι να πραγματοποιηθεί ανάλυση των δεδομένων, να ανακαλυφθούν σχέσεις και να εξαχθεί γνώση.[43]

Ουσιαστικά η εξόρυξη δεδομένων, όπως περιγράφεται από τον Paolo Giudici στο αντίστοιχο βιβλίο του, είναι η διαδικασία της επιλογής, αναζήτησης και μοντελοποίησης μεγάλου όγκου δεδομένων, ώστε να την ανακαλυφθούν συστηματοποιήσεις (regularities) ή συσχετίσεις που αρχικά ήταν άγνωστες με στόχο την απόκτηση ξεκάθαρων και χρήσιμων αποτελεσμάτων για τον κάτοχο της βάσης δεδομένων. Υπάρχουν δύο μεθοδολογίες ως προς τον προσανατολισμό του αναλυτή, η προβλεπτική και η περιγραφική. Η περιγραφική, προσφέρει περιγραφή των δεδομένων με στόχο την γενίκευση, με λειτουργίες όπως την ομαδοποίηση τους (clustering). Η προβλεπτική χρησιμοποιεί το σύνολο δεδομένων ώστε να προβλέψει τιμές που υπολείπονται οι μελλοντικές τιμές για μεταβλητές.[44] Η δυνατότητα αυτή παρέχεται μέσω των αλγορίθμων μηχανικής εκμάθησης με προσανατολισμό στην παρούσα εργασία την πρόβλεψη.

Ο τομέας την μηχανικής εκμάθησης είναι κλάδος της επιστήμης υπολογιστών, με στόχο την εκμάθηση από τα δεδομένα και την εξαγωγή προβλέψεων ως προς αυτά. Ο εκάστοτε αλγόριθμος εκπαιδεύετε ως προς την προβλεπτική του ικανότητα με βάση το σύνολο εκπαίδευσης που του παρέχεται και αποκτώντας “εμπειρία” εξάγει προβλέψεις. Δεδομένου ότι η παραπάνω διαδικασία είναι στοχαστική δεν μπορούν να υπάρξει ντετερμινισμός ως προς το αποτέλεσμα. Είθισται να εκτιμάται η αποτελεσματικότητα του μέσα από δείκτες αξιολόγησης των αποτελεσμάτων, με τη χρήση του συνόλου ελέγχου καθώς και με όρους υπολογιστικού φόρτου.[45]

Οι αλγόριθμοι μηχανικής χαρακτηρίζονται από το μοτίβο υλοποίησης και στυλ εκμάθησης που χρησιμοποιούν. Στο παρόν κεφάλαιο έχει επιλέγει να αναλυθούν αυτοί που χρησιμοποιήθηκαν για τη συγκριτική αξιολόγηση, με προσανατολισμό την πρόβλεψη και στόχο την ταξινόμηση (classification) για χρήση στο customer analytics. Στο κεφάλαιο της πρόβλεψης αποχώρησης πελατών υποστηρίζεται βιβλιογραφικά η επιλογή τους. Προτού περιγραφούν θα γίνει αναφορά στο στόχο-εργασία (task) της εξόρυξης δεδομένων και του αντίστοιχου αλγορίθμου, ως προς το ζητούμενο από τον αναλυτή.

2.1.2.1 ΣΤΟΧΟΣ ΜΟΝΤΕΛΟΥ-ΕΡΓΑΣΙΕΣ (TASKS)

Όπως προαναφέρθηκε ο προσανατολισμός του αναλυτή μπορεί να είναι η πρόβλεψη μιας κατάστασης ή η περιγραφή των δεδομένων. Αυτές οι δύο λειτουργίες μπορούν να αναλυθούν περαιτέρω ως προς το στόχο του μοντέλου που θα αναπτυχθεί με τη χρήση του αλγορίθμου μηχανικής εκμάθησης. Οι υποκατηγορίες χαρακτηρίζονται ως εργασίες (task) και συμμετέχουν ως παράμετρο στην επιλογή του κατάλληλου αλγορίθμου. Παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 2) με τις αγγλικές ονοματοθεσίες. [46]

Πίνακας 2: Περιγραφή εργασιών (Tasks) αλγρίθμων μηχανικής εκμάθησης

Εργασία	**Λειτουργία**	**Περιγραφή**
Classification	Προβλεπτική	Ταξινόμηση των αποφασιζόντων σε κλάσεις
Clustering	Περιγραφική	Ομαδοποίηση των αποφασιζόντων με βάση κοινά χαρακτηριστικά.
Association Rule Discovery	Περιγραφική	Ανακάλυψη συσχετίσεων των μεταβλητών.
Sequential Pattern Discovery	Περιγραφική	Ανακάλυψη επαναλαμβανόμενων μοτίβων, συμπεριφορών.
Forecasting	Προβλεπτική	Πρόβλεψη τιμών σε κενά πεδία- απόφασης πελάτη για κάποια μεταβλητή.
Regression	Προβλεπτική	Εναλλακτική του forecasting. Διαφορετική μεθοδολογία.

Στην παρούσα διπλωματική στόχος είναι το αποτελεσματικό classification μέσω αλγορίθμων μηχανικής εκμάθησης. Δεδομένου ότι ο προσανατολισμός είναι η πρόβλεψη, οι αλγόριθμοι που θα επιλεγούν θα χρησιμοποιούν μάθηση με επίβλεψη (supervised learning). Στους αλγόριθμους μηχανικής εκμάθησης αντιστοιχούν τρία στυλ ανάλογα το τρόπο με τον οποίο αλληλεπιδρούν κατά την εκπαίδευση. Αυτό καθορίζει την διαδικασία επεξεργασίας των δεδομένων και χαρακτηριστικά δημιουργίας των αντίστοιχων συνόλων δεδομένων που θα εισαχθούν στον αλγόριθμο. Υπάρχει η μάθηση υπό επίβλεψη, χωρίς επίβλεψη (unsupervised learning)και με ημι-επίβλεψη (semi – supervised). Στην πρώτη περίπτωση, υπάρχουν οι τιμές των κλάσεων κατά την εκπαίδευση ενώ στην δεύτερη απουσιάζουν και ζητούμενο είναι δημιουργία γενικών κανόνων. Η τρίτη αποτελεί ένα συνδυασμό των παραπάνω δύο.[47]

2.1.2.2 ΑΛΓΟΡΙΘΜΟΙ ΑΝΑΠΤΥΞΗΣ ΜΟΝΤΕΛΩΝ

Έχοντας αναφερθεί τόσο στο στυλ εκπαίδευσης, στις λειτουργίες και στις εργασίες, απομένει να γίνει αναφορά στην κατηγοριοποίηση των αλγορίθμων με βάση τις ομοιότητες που παρουσιάζουν στον τρόπο υλοποίησης τους. Οι ομάδες αλγορίθμων που περιγράφονται, χρησιμοποιούν μάθηση υπό επίβλεψη με προσανατολισμό την πρόβλεψη και στόχο την ταξινόμηση. Αξίζει να σημειωθεί ότι συνήθως οι απλοί αλγόριθμοι δουλεύουν ικανοποιητικά και η αποτελεσματικότητα μιας μεθόδου έχει άμεση συσχέτιση με το πεδίο στο οποίο εφαρμόζεται.

BAYESIAN ΑΛΓΟΡΙΘΜΟΙ

Χρησιμοποιεί το θεώρημα του Bayes, όπου εκφράζει την πιθανότητα να συμβεί το γεγονός A υπό την προϋπόθεση ότι συνέβη το γεγονός B, για προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression).

$$\text{Θεώρημα Bayes} \quad P(A/B)=\frac{P(A)*P(B/A)}{P(B)} \quad (5)$$

Οι πιο διαδεδομένοι αλγόριθμοι είναι:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

Θα αναλυθούν ο Naive Bayes αρχικά, όπου αποτελεί την απλούστερο αλγόριθμο της κατηγορίας αυτής και η υλοποίηση του Bayesian Network, καθώς χρησιμοποιείται στην συγκριτική μελέτη. Ο Naive Bayes , ονομάζεται Naive καθώς κάνει την παραδοχή ότι όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους. Για να συμπεριληφθούν όλα τα ενδεχόμενα που ανήκουν στο σύνολο της εκάστοτε μεταβλητής ο τύπος εκφράζεται ως εξής:

$$P(A/B)=\frac{P(A)*\prod_{i=1}^{n} P(B_i/A)}{P(B)} \quad (6)$$

όπου A η κλάση και B αναφέρεται στη μεταβλητή. Εισάγοντας και ένα κανόνα απόφασης (decision rule), για την τελική απόφαση καταλήγουμε στην ταξινόμηση. Ο κανόνας στην προκειμένη περίπτωση είναι η επιλογή της πιθανότητας με την μεγαλύτερη τιμή (argmax) και η κλάση που τις αντιστοιχεί.

$$\hat{y}=argmax\, P(A)*\prod_{i=1}^{n} P(B_i/A) \quad (7)$$

Ο αλγόριθμος Bayesian Net αντιμετωπίζει την αδυναμία του Naive Bayes να χρησιμοποιηθεί σε δεδομένα που δεν εκφράζονται από απλές κατανομές και την αδυναμία των δέντρων απόφασης,

όπου αναλύεται παρακάτω. Εκφράζει γραφικά τις κατανομές πιθανοτήτων και τις αντίστοιχες συσχετίσεις με την αποφυγή δημιουργίας κύκλων, ως προς τις συσχετίσεις των μεταβλητών. Σε κάθε μεταβλητή αντιστοιχεί ένας κόμβος. Ως είσοδο ο κάθε κόμβος τραβάει ένα σύνολο από τους προγόνους-γονείς και σαν έξοδο δίνει την αντίστοιχη πιθανότητα ή την κατανομή της μεταβλητής. [48]

Πίνακας 3: Περιγραφή πλεονεκτημάτων και μειονεκτημάτων Bayesian αλγορίθμων

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Άμεσος υπολογισμός των πιθανοτήτων των υποθέσεων	Απαιτεί ταυτόχρονο υπολογισμό/γνώση αρκετών πιθανοτήτων
Μπορεί να χρησιμοποιηθεί για εκμάθηση βάσει στατιστικής	Μπορεί να είναι ασύμφορο από άποψη υπολογιστικού κόστους.
Σε συγκεκριμένες περιπτώσεις δουλεύει πολύ καλύτερα από άλλους αλγόριθμους	
Μπορεί να συνδυάσει αποτελεσματικά νέα με παλαιά γνώση για να τη βελτιστοποιήσει	
Μπορεί να κάνει στοχαστικές προβλέψεις	

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Από τους πλέον διαδεδομένους αλγορίθμους που ανήκουν στην κατηγορία των αλγορίθμων παλινδρόμησης. Μοντελοποιεί και επαναπροσδιορίζει επαναληπτικά τη συσχέτιση μεταξύ των μεταβλητών ή μεταβλητών και κλάσης με τη μέτρηση του λάθους στις εκτιμήσεις του μοντέλου. Χρησιμοποιείται σε περιπτώσεις κατηγορικών κλάσεων.

Η είσοδος στην συνάρτηση πιθανότητας, που προκύπτει ως σχέση της λογιστικής παλινδρόμησης, είναι μια γραμμική συνάρτηση της μορφής

$$Y(\vec{x}) = \vec{a} * \vec{x}, \text{ όπου } \vec{a} = [a_0 a_1 .... a_n] x = [x_0 x_1 .... x_n], x_0 = 0 \quad (8)$$

όπου αποδίδονται βάρη (α) σε κάθε χαρακτηριστικό/ μεταβλητή απόφασης (x) ανάλογα την συμμετοχή στο τελικό αποτέλεσμα του. Η έξοδος Y είναι ένας δείκτης έκφρασης του τελικού αποτελέσματος συναρτήσει της κλάσης. Για την προσαρμογή της συνάρτησης αυτής στα δεδομένα, με στόχο την βελτίωση του αποτελέσματος, δηλαδή καλύτερη επιλογή τιμών στα βάρη ώστε ο αποφασίζοντας να ταξινομηθεί στην σωστή κλάση, χρησιμοποιείται η συνάρτηση μέγιστης πιθανοφάνειας ως συνάρτηση απώλειας (loss function)[49]

Η παρακάτω σχέση της λογιστική παλινδρόμησης εκφράζει την πιθανότητα να ανήκει πραγματικά στην κλάση που του αποδίδεται μέσω της τιμής Y κάποιος με τα χαρακτηριστικά x. Η σχέση που περιγράφει την λογιστική παλινδρόμηση:

$$P(Y=1) = \frac{e^{-Y}}{1-e^{-Y}} \text{ και } Q(Y=0) = 1 - P(Y=1) = \frac{1}{1+e^{-Y}}, \text{ όπου δουλέυουμε με αυτό τον τύπο.} \quad (9)$$

Ζητούμενο είναι η βελτιστοποίηση των τιμών του διανύσματος α.

Πίνακας 4: Περιγραφή πλεονεκτημάτων και μειονεκτημάτων λογιστικής παλινδρόμησης

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Γρήγορος και απλός αλγόριθμος	Χρειάζονται μεγαλύτερο όγκο δεδομένων για να προσφέρει καλά αποτελέσματα
Χειρίζεται μη γραμμικές εκφράσεις	
Δεν υποθέτει κανονική κατανομή για τις ανεξάρτητες μεταβλητές	
Δουλεύει καλά σε περιπτώσεις δύο κλάσεων	

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

Αναφέρεται στις μεθοδολογίες που χρησιμοποιούν δέντρα απόφασης στη διαδικασία της εκπαίδευσης, όπου χρησιμοποιούν κανόνες {if.. then} για τη δημιουργία συσχετίσεων. Χρησιμοποιούνται τόσο για ταξινόμηση όσο και για απόδοση συνεχών τιμών στην εξαρτημένη μεταβλητή. Τα δέντρα απόφασης βασίζονται στην αρχή “διαίρε-βασίλευε” και μπορούν να αναπαρασταθούν γραφικά, όπου αποτελούνται από κόμβους κλαδιά και φύλλα. Τα φύλλα ως ο τερματικός κόμβος αναπαριστούν τη πιθανότητα με την οποία ένας αποφασίζον με συγκεκριμένα χαρακτηριστικά ανήκει σε μια συγκεκριμένη κλάση. Οι κόμβοι αναπαριστούν τον έλεγχο μιας υπόθεσης για την μεταβλητή και τα κλαδιά το αποτέλεσμα με την αντίστοιχη πιθανότητα που τα συνοδεύει. Τα κλαδιά συνδέονται με λογικές σχέσεις με το επόμενο χαρακτηριστικό/μεταβλητή προς διερεύνηση προκειμένου να καταλήξει, με την ανάπτυξη του δέντρου, στα αντίστοιχα φύλλα. Ένα δέντρο μπορεί να αναπτυχθεί εξαντλητικά αλλά αυξάνει σημαντικά το υπολογιστικό κόστος. Με την ανάπτυξη του δέντρου, μπορούν να εξαχθούν στατιστικές σχέσεις που αποδίδουν την πιθανότητα να εντάσσεται κάποιος σε μια κλάση συναρτήσει των αντίστοιχων χαρακτηριστικών.

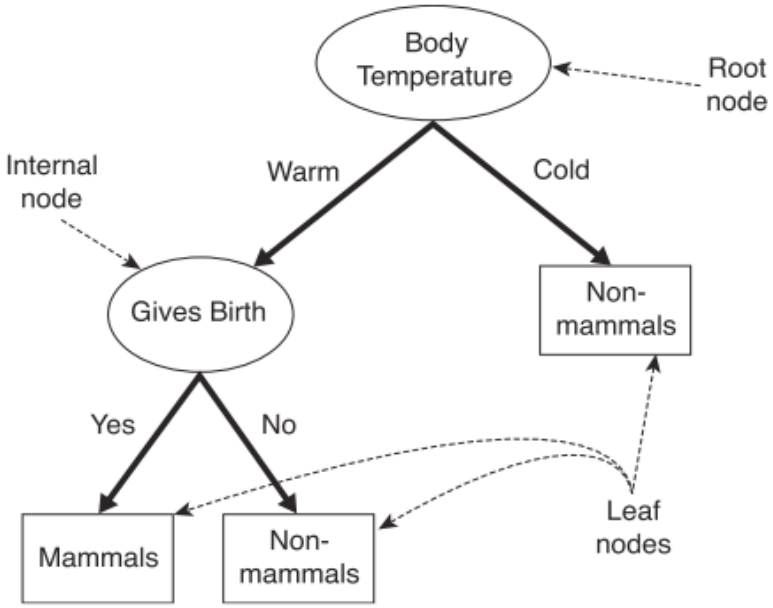


Figure 4.4. A decision tree for the mammal classification problem.

Εικόνα 1: Περιγραφή γραφικής αναπαράστασης δέντρων απόφασης (Tan,Steinbach, Kumar, Introduction to Data Mining)

Αποτελείται από δύο στάδια η υλοποίηση του δέντρου απόφασης την ανάπτυξη, και το κλάδεμα (pruning) . Κλάδεμα πραγματοποιείται όταν υπάρχουν πλεονάζουσες συγκρίσεις ή τίθεται ζήτημα βελτίωσης της απόδοσης και αφαίρεση υποδέντρα. Η υλοποίηση ενός αλγόριθμου που χρησιμοποιεί δέντρα απόφασης εξαρτάται από την επιλογή των κριτηρίων διάσπασης και τερματισμού. Το κριτήριο διάσπασης αφορά στη βέλτιστη επιλογή της μεταβλητής για τη διάσπαση ενός συνόλου σε υποσύνολα. Το κριτήριο τερματισμού τερματίζει την ανάπτυξη του δέντρου ή τμήματος όταν παρατηρηθούν προβλήματα όπως υπερπροσαρμογή, όταν έχουν αναλυθεί όλα τα χαρακτηριστικά ή όταν τεθεί κάποιο όριο συναλλαγής μεταξύ ακρίβειας και υπολογιστικής

πολυπλοκότητας.[50]

Με βάση την επιλογή αυτών των κριτηρίων δημιουργούνται διαφορετικού είδους αλγόριθμοι. Οι πλέον κλασικοί αλγόριθμοι είναι οι ID3 και C4.5. Πάραυτα για περιγραφή επιλέχθηκε ο CART (Classification and Regression Tree), που χρησιμοποιεί και τις δύο προσεγγίσεις των δέντρων απόφασης, ταξινόμηση και παλινδρόμηση. Αναπτύσσεται εξαντλητικά (gridy) και δυαδικά εξετάζοντας μια μεταβλητή τη φορά. Ως κριτήριο διάσπασης χρησιμοποιεί την τιμή του δείκτη gini, που αναλύεται στην ενότητα των δεικτών. Κατά περίπτωση, μπορεί να χρησιμοποιήσει εναλλακτικούς δείκτες όπως information gain , όπου περιγράφεται στις μεθοδολογίες διαλογής μεταβλητών. Επίσης, χρησιμοποιεί κριτήριο κλαδέματος, αφού ολοκληρώσει την ανάπτυξη του δέντρου. Εξετάζει αν δημιουργείται υπερπροσαρμογή στα δεδομένα και αφαιρεί τα αντίστοιχα τμήματα. Ξεκινάει από τους μικρούς κόμβους και έπειτα προχωράει στους μεγαλύτερους, εξετάζοντας όλο το δέντρο. Αρχικά, διαγράφει τα τμήματα που παρέχουν τη λιγότερη πληροφορία με όρους κέρδους ως προς την ακρίβεια. Διατηρεί όλα τα δέντρα και τα συγκρίνει για να επιλεγεί το βέλτιστο δέντρο ως προς την ακρίβεια και τον αντίστοιχο υπολογιστικό φόρτο, με προσανατολισμό τη μείωση του κόστους. [51]

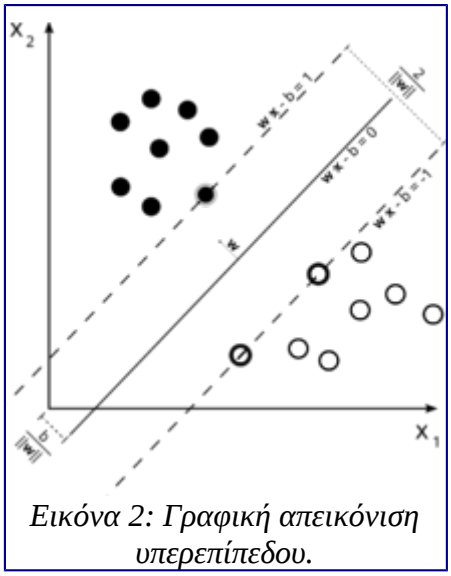
Ο αλγόριθμος που χρησιμοποιείται από την κατηγορία αυτή για τη συγκριτική μελέτη είναι ο SimpleCart. Μια υλοποίηση του CART που ουσιαστικά το μόνο κοινό που έχουν είναι η χρήση στρατηγικής κλαδέματος (μετά την ανάπτυξη του δέντρου) με το ελάχιστο κόστος που προκύπτει από την πολυπλοκότητα. Παρέχει δυνατότητες παραμετροποίησης, εν αντιθέσει με τον CART. Μπορεί να προκαθοριστεί το ελάχιστο πλήθος περιπτώσεων κάθε φύλλου, το ποσοστό που θα χρησιμοποιηθεί από το σύνολο δεδομένων ως σύνολο εκπαίδευσης για την ανάπτυξη του δέντρου και το πλήθος (folds) των υποσυνόλων δεδομένων που θα χρησιμοποιηθούν στη διαδικασία του κλαδέματος για διασταύρωση (cross-validation) των αποτελεσμάτων.[52]

Πίνακας 5: Περιγραφή πλεονεκτημάτων και μειονεκτημάτων δέντρων απόφασης[53]

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Εύκολη κατανόηση και ερμηνεία των αποτελεσμάτων	Μπορεί να οδηγήσει σε υπερπροσαρμογή των δεδομένων.
Απαιτεί ελάχιστη προετοιμασία των δεδομένων	Μικρές αλλαγές στα δεδομένα μπορεί να οδηγήσουν σε μεγάλες διαφοροποιήσεις στο μοντέλο
Εύκολο στην ερμηνεία το μοντέλο και οι αντίστοιχες καταστάσεις (white box).	Επειδή μέσω ευρετκών αλγορίθμων εντοπίζει τοπικά ακρότατα, στην πρακτική του υλοποίηση, δεν υπάρχει εγγύηση ότι θα επανέλθει στο ολικό ακρότατο.
Δουλεύει με ποσοτικές και ποιοτικές μεταβλητές.	Αντιμετωπίζει δυσκολία εκμάθησης κάποιων εκφράσεων.
Μπορεί να μετρηθεί η αξιοπιστία του μοντέλου μέσα από στατιστικούς ελέγχους.	Απαιτείται εξισορρόπηση του συνόλου εκπαίδευσης ως προς τις κλάσεις, για μην οδηγήσει μεροληπτικά αποτελέσματα ως προς μια κλάση.
Δουλεύει καλά ακόμα και αν παραβιάσει κάποιες υποθέσεις από το πραγματικό μοντέλο από όπου προέκυψαν τα δεδομένα.	Αδυναμία να κάνει χρήση νέων δεδομένων εκπαίδευσης για βελτίωση του ήδη υπάρχοντος μοντέλου (on-line learning)

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ

Αποτελούν αλγόριθμους που επιτυγχάνουν εκμάθηση γραμμικών ταξινομητών. Χρησιμοποιούν το υπερεπίπεδο μέγιστου περιθωρίου (maximum margin hyperplane), εργαλείο το οποίο επιτυγχάνει το μέγιστο διαχωρισμό των τάξεων χωρίς λάθος στο υπερεπίπεδο. Για τη δημιουργία τπερεπιπέδων, χρησιμοποιείται συναρτήσεις kernel. Τα υποδείγματα με τη μικρότερη απόσταση από αυτό το επίπεδο αποκαλούνται διανύσματα υποστήριξης. Στόχος η μεγιστοποίηση της απόσταση των διανυσμάτων υποστήριξης. Μαθηματικά αυτό επιτυγχάνεται με την επίλυση ενός προβλήματος τετραγωνικού προγραμματισμού και την βελτιστοποίηση της αντίστοιχης αντικειμενικής συνάρτησης. Μέσα από την επαναληπτική διαδικασία εκπαίδευσης παράγονται διανύσματα υποστήριξης που εκφράζουν το μέγιστο διαχωρισμό των κλάσεων.



Εικόνα 2: Γραφική απεικόνιση υπερεπίπεδου.
(en.wikipedia.org/wiki/Support_vector_machine)

Το υπερέπιπεδο αναπαρίσταται από τη σχέση **w.x-b,** όπου το όριο διαχωρισμού δίνει τιμη 0 για αυτή τη σχέση και για τις κλάσεις που εκφράζονται από τις τιμές 1 και -1, δίνει τις αντίστοιχες τιμές. Στόχος η μεγιστοποίηση της απόστασης, που εκφράζεται ως **2/||w||** μεταξύ των δύο κλάσεων ως προς το υπερεπίπεδο. Οπότε η αντικειμενική που προκύπτει έχει ως στόχο ελαχιστοποίηση της τιμής **w** και προϋπόθεση να μην υπερκαλυφθούν τα πεδία που εκφράζουν τις δύο κλάσεις.[54]

Ένας αλγόριθμος που επιλύει τα προβλήματα που μπορούν να ανακύψουν λόγω τετραγωνικού προγραμματισμού, είναι ο αλγόριθμος προσεγγιστικων μηχανών διανυσμάτων υποστήριξης (PSVM). Αντί ενός διαχωρίσιμου υπερεπιπέδου, χρησιμοποιεί δύο παράλληλα επίπεδα. Κάθε επίπεδο αντιστοιχεί σε μια κλάση και τα επίπεδα αρχικά τοποθετούνται κοντά. Στόχος , η κατά το δυνατόν μέγιστη απομάκρυνση τους. Από αυτή την υλοποίηση προκύπτει μια αντικειμενική συνάρτηση που επιλύεται με γραμμική προσέγγιση. Ως αποτέλεσμα μειώνεται ο υπολογιστικός φόρτος και η πολυπλοκότητα που προέκυπτε από την απλή υλοποίηση του, διατηρώντας την αποτελεσματικότητα του.[55]

Πίνακας 6: Περιγραφή πλεονεκτημάτων και μειονεκτημάτων μηχανών διανυσμάτων υποστήριξης[56]

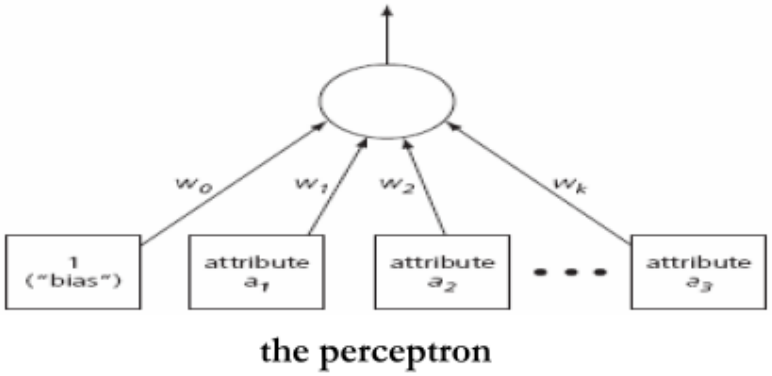
ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Ανθεκτικοί στην υπερπροσαρμογή	Δυσκολία στην διερεύνηση σε βάθος των αποτελεσμάτων
Χαμηλό υπολογιστικό κόστος ακόμα και σε περιπτώσεις μη γραμμικότητας	Η κατάλληλη επιλογή της συνάρτησης kernel
Δουλεύει γρήγορα στα αραιά δεδομένα	
Μπορούν να εφαρμοστούν αποτελεσματικά σε δεδομένα με θόρυβο	

ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Τα τεχνητά νευρωνικά δίκτυα αναπτύχθηκαν στα πλαίσια προσπάθειας της μίμησης λειτουργίας των νευρώνων του ανθρώπινου εγκεφάλου. Αναλογικά, αν οι νευρώνες θεωρηθούν απλά υπολογιστικά εργαλεία, ο στόχος είναι μέσω σύνθεσης απλών υπολογιστικών εργαλείων σε ένα περιβάλλων με πολύπλοκες και σύνθετες διασυνδέσεις να αναπτυχθεί ένας ευφυής τρόπος μηχανικής εκμάθησης. Χρησιμοποιείται τόσο στην αναλυτική όσο και στην περιγραφική προσέγγιση της εξόρυξης δεδομένων. Προτού, γίνει περαιτέρω ανάλυση είναι χρήσιμο να παρουσιαστεί η έννοια του perceptron, όπου αρχικά αναπτύχθηκε για την επίλυση προβλημάτων στους αλγορίθμους γραμμικής παλινδρόμησης και αποτέλεσε πρόγονο των νευρωνικών δικτύων.

Στις μεθόδους γραμμικής παλινδρόμησης παρουσιάζονταν αστοχίες, όπως αδυναμία κατάταξης στην ορθή κλάση, όταν υπήρχαν δεδομένα που παρουσίαζαν μη γραμμικές εξαρτήσεις. Στα πλαίσια αυτού είσηχθη η μεθοδολογία εκμάθησης του υπερεπίπεδου στον διαχωρισμό υποδείγματων διαφορετικών κλάσεων. Όπως αναλύθηκε και στη λογιστική παλινδρόμηση το υπερεπίπεδο καθώς και ο αντίστοιχος διαχωρισμός του ως προς τις κλάσεις περιγράφεται από μια εξίσωση, που ανάλογα την έξοδο ταξινομεί. Η τροποποίηση που εισήχθη, σε περίπτωση λανθασμένης κατάταξης μετέβαλε τις αντίστοιχες βαρύτητες ώστε να μετακινηθεί το υπερεπίπεδο στη σωστή κατεύθυνση, όταν οι κλάσεις είναι γραμμικά διαχωρίσιμες. Το υπερεπίπεδο που προκύπτει ονομάζεται perceptron και πλέον στα τεχνητά νευρωνικά δίκτυα αναφέρεται ως νευρώνας. Παρακάτω περιγράφεται η υλοποίηση του.[57]

```
Set all weights to zero
Until all instances in the training data are classified
   correctly
       For each instance I in the training data
              If I is classified incorrectly by the perceptron
                     If I belongs to the first class add it to
                            the weight vector
                     else subtract it from the weight vector
```



the perceptron

Εικόνα 3: Υλοποίηση Perceptron

(Διαλέξεις Επιχειρηματικής Ευφυΐας & διαχείρισης γνώσης (Κ.Λακιωτάκη), Π.Κ.)

Ένας νευρώνας έχει παρόμοια μορφή λειτουργία με τον perceptron, καθώς δέχεται ως είσοδο αριθμητικά επιτελεί κάποιους υπολογισμούς και ανάλογα τους αποτελέσματος δίνει μια έξοδο. Υπάρχουν τρία είδη νευρώνων, εισόδου όπου παραλαμβάνουν τα δεδομένα από για να μεταβιβάσουν στους υπόλοιπους νευρώνες (υπολογιστικούς), εξόδου, που εξάγουν το αποτέλεσμα από το δίκτυο και τέλος οι υπολογιστικοί νευρώνες. Οι υπολογιστικοί νευρώνες δέχονται ως είσοδο κάποια δεδομένα από τους νευρώνες/κόμβους που προηγούνται και τα πολλαπλασιάζουν με τα αντίστοιχα βάρη που αντιστοιχούν στην σύναψη απο την οποία προήλθαν (συναπτικά βάρη). Το σύνολο των εισόδων αφου πολλαπλασιαστεί με τα αντίστοιχα βάρη αθροίζεται. Η τιμή που προκύπτει χρησιμοποιείται ως είσοδο για την συνάρτηση που αντιστοιχεί στον εκάστοτε νευρώνα . Πραγματοποιείται εσωτερικά ο υπολογισμός και προκύπτουσα τιμή χρησιμοποιείται ως είσοδος για τη συνάρτηση ενεργοποίησης που έχει επιλέγει , όπου ανάλογα του αποτελέσματος , ενεργοποιείται ή όχι η μεταβολή της τιμής της εξόδου του νευρώνα. Τα επίπεδα που περιέχουν υπολογιστικούς νευρώνες ανήκουν στα κρυμμένα επίπεδα (hidden layer).[58]

Ουσιαστικά με πραγματοποιείται εκπαίδευση εσωτερικά με επαναληπτική διαδικασία ώστε να βελτιστοποιηθούν οι τιμές των βαρών που αντιστοιχούν στις συνδέσεις με τους νευρώνες, με βάση την ισχύ της συνεισφοράς κάθε νευρώνα στην τελική πρόβλεψή, και το κατώφλι για την ενεργοποίηση μεταβολής. Ο αλγόριθμος που θα χρησιμοποιηθεί από την αντίστοιχη κατηγορία στην υλοποίηση των δοκιμών είναι ο Multilayerperceptron. όπου πραγματοποιείται πολυεπίπεδη απλή ιεραρχική σύνδεση perceptron, με μεθοδολογία υλοποίησης όπως περιγράφεται για τα νευρωνικά δίκτυα και με επαναπροσδιορισμό των βαρών με ανάστροφη μετάδοση (back-propagation).

Πίνακας 7: Περιγραφή πλεονεκτημάτων και μειονεκτημάτων νευρωνικών δικτύων

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Δουλεύει αποτελεσματικά σε περιπτώσεις μη γραμμικότητας	Το κομμάτι της υλοποίησης αποτελεί ένα μαύρο κουτί
Δεν απαιτεί γνώση στατιστικής για την χρήση του	Είναι δύσκολο να ερμηνευτεί το μοντέλο που αναπτύσσεται ώστε να διερευνηθεί το αποτέλεσμα.
Αποτελεσματικό με όρους ακρίβειας	Ως μη στοχαστικό μοντέλο δεν είναι αποτελεσματικό στη εκτίμηση λάθους ταξινόμησης.
Εφαρμόζεται αποτελεσματικά σε άγνωστα πεδία	Υπολογιστικό κόστος
Ανταποκρίνεται καλά σε περιπτώσεις με θόρυβο	Δυσκολία στην επιλογή κατάλληλης σχεδίασης, π.χ. πλήθους επιπέδων

ΜΕΤΑΜΑΘΗΣΙΑΚΑ ΜΟΝΤΕΛΑ (Ensemble methods)

Ο συνδυασμός μαθησιακών μοτίβων, μέσω του συνδυασμού αλγορίθμων για να καλυφθούν αδυναμίες που εμφάνιζε ο καθένας ξεχωριστά και για τη περαιτέρω βελτιστοποίηση των αποτελεσμάτων, οδήγησε στις συνδυαστικές μεθόδους ή εναλλακτικά τα μεταμαθησιακά μοντέλα. Ο εκάστοτε αλγόριθμος που χρησιμοποιεί μεταμαθησιακά μοντέλα, χωρίζει με τυχαιότητα σε διάφορα σύνολα εκπαίδευσης το σύνολο δεδομένων για να αντιστοιχεί ένα σύνολο εκπαίδευσης σε κάθε μαθησιακό μοντέλο που περιέχει. Έπειτα μέσω ψηφοφορίας (voting) είτε αποδίδοντας σε όλα τα αποτελέσματα/μεθόδους ίδια βαρύτητα ως προς την ψήφο, είτε αποδίδοντας μεγαλύτερη βαρύτητα στις καλύτερες μεθόδους, με βάση δείκτες αξιολόγησης, επιλέγει την κλάση με τις περισσότερες ψήφους, στην περίπτωση της ταξινόμησης. Παρότι μπορεί να παράσχουν βελτιστοποίηση ως προς την ακρίβεια των αποτελεσμάτων, υπάρχει δυσκολία ερμηνείας τους συναρτήσει της μεθοδολογίας εκπαίδευσης. Επίσης, κατά περίπτωση εμφανίζουν αυξημένο υπολογιστικό φόρτο.[59] Παρακάτω αναλύονται οι τρεις αλγόριθμοι που χρησιμοποιούνται στην εργασία και οι αντίστοιχες κατηγορίες στις οποίες υπόκεινται.

Μια απλή και κλασσική μεθοδολογία συνδυαστικού αλγόριθμου περιγράφεται από τον αλγόριθμο bagging, που χρησιμοποιεί συνδυασμό ίδιων μεθόδων. Στην προκειμένη περιγραφή της μεθοδολογίας χρησιμοποιούνται πολλαπλά δέντρα απόφασης στα οποία επιλέγεται με τυχαιότητα το εκάστοτε σύνολο εκπαίδευσης που θα χρησιμοποιηθεί ως είσοδος. Τα σύνολα εκπαίδευσης,με τυχαία επιλεγμένα παραδείγματα, αποτελούν υποσύνολα του αρχικού συνόλου εκπαίδευσης. Σε κάθε επανάληψη της εκμάθησης, δημιουργούνται νέοι συνδυασμοί συνόλων εκπαίδευσης με αφαίρεση των ακατάλληλων, όπως προκύπτει από τη διαδικασία εκπαίδευσης του αλγορίθμου, παραδειγμάτων και με αναπαραγωγή ήδη υπαρχόντων για αντικατάσταση τους Οι βαρύτητες των αποτελεσμάτων είναι όμοιες για κάθε δέντρο απόφασης. Ο λόγος που προσφέρει καλύτερα αποτελέσματα από ένα απλό δέντρο απόφασης, κατά βάση, είναι ότι τα δέντρα απόφασης είναι ασταθή, με μικρές αλλαγές στα δεδομένα εμφανίζουν μεγάλες αλλαγές στα αποτελέσματα.[60]

Στην παραπάνω περιγραφή εισήχθη η έννοια της τυχαιότητας. Πάραυτα υπάρχουν μοντέλα που από κατασκευής περιλαμβάνουν τη παράμετρο της τυχαιότητας. Μια μέθοδος που περιέχει τυχαιότητα παράσχει συνήθως παρόμοια αποτελέσματα με τον bagging. Όμως δύναται να παράσχει ακόμα καλύτερα αποτελέσματα μια μέθοδος που περιέχει από κατασκευής την τυχαιότητα χρησιμοποιούμενη από τη μεθοδολογία του bagging. Μια τέτοια περίπτωση είναι ο αλγόριθμος Random Forest, όπου αντί κλασικών δέντρων απόφασης, σε κάθε επανάληψη παράγει τυχαίως παρηγμένα δέντρα απόφασης. Γενικά η τεχνική του Randomization δίνει πρακτική αξία στην μεθοδολογία bagging, καθώς μπορεί να δώσει διαφορετικά αποτελέσματα ακόμα και σε ευσταθής αλγορίθμους.[61]

Μια μεθοδολογία ακόμα είναι η Boosting, η οποία παρουσιάζει τόσο ομοιότητες όσο και διαφορές με τη bagging. Χρησιμοποιεί ψηφοφορία, όπου αποδίδει βάρη βάσει των επίπεδων βεβαιότητας (confidence) κάθε μοντέλου, και συνδυάζει μοντέλα από τον ίδιο αλγόριθμο. Πάραυτα, χρησιμοποιεί αλγορίθμους που αναπτύσσουν μοντέλα με συμπληρωματική αξία μεταξύ τους. Εκτελεί επαναληπτική διαδικασία κατά την οποία χρησιμοποιεί τη γνώση που προέκυψε από το προηγούμενο μοντέλο ώστε να βελτιώσει το επόμενο που θα αναπτυχθεί. Επίσης, εκπαιδεύει τα

μοντέλα, ώστε να βελτιωθούν στην ορθή ταξινόμησης, αποδίδοντας μεγαλύτερη βαρύτητα στα μη ορθώς ταξινομημένα παραδείγματα. [60]

Ο AdaBoostM1 αποτελεί μια υλοποίηση αυτής της μεθοδολογίας. Ο συγκεκριμένος αλγόριθμος αποδίδει βάρη στα παραδείγματα που εκφράζουν τη σωστή ή όχι ταξινόμηση τους. Αρχικά, όλα τα παραδείγματα έχουν την ίδια βαρύτητα κατά την εκπαίδευση. Έπειτα, αυξάνεται η βαρύτητα σε αυτά που ταξινομήθηκαν λανθασμένα και μειώνεται σε αυτά που τοποθετήθηκαν στη σωστή κλάση. Αυτό συμβαίνει επαναληπτικά εκπαιδεύοντας και δημιουργώντας ταξινομητές/μοντέλα που το καθένα λειτουργεί καλά σε διαφορετικά σύνολα δεδομένων. Με αυτό τον τρόπο βελτιστοποιείται η έξοδος των αλγορίθμου και δρουν συνδυαστικά οι ταξινομητές (classifiers).[62]

Model Generation

```
Assign equal weight to each training instance.
For each of t iterations:
  Apply learning algorithm to weighted dataset and store resulting
    model.
  Compute error e of model on weighted dataset and store error.
  If e equal to zero, or e greater or equal to 0.5:
    Terminate model generation.
  For each instance in dataset:
    If instance classified correctly by model:
       Multiply weight of instance by e / (1 - e).
  Normalize weight of all instances.
```

Classification

```
Assign weight of zero to all classes.
For each of the t (or less) models:
  Add -log(e / (1 - e)) to weight of class predicted by model.
Return class with highest weight.
```

Εικόνα 4: Αλγόριθμος AdaBoostM1 (Ian H. Witten , Eibe Frank ,Mark A. Hall, Data Mining)

ΕΞΕΛΕΓΚΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ (EVOLUTIONARY ALGORITHMS)

Οι εξελεγκτικοί αλγόριθμοι αν και δεν υπόκεινται κατά αποκλειστικότητα στα εργαλεία της εξόρυξης δεδομένων, έχουν χρήση και σε αυτή την επιστήμη.[63] Οι εξελεγκτικοί αλγόριθμοι προσπαθούν να μιμηθούν λειτουργίες της βιολογίας, όπως άλλωστε και τα τεχνητά νευρωνικά δίκτυα. Ακολουθούν τον εξελικτικό νόμο του Δαρβίνου “μια γενική αρχή, που οδηγεί στην εξέλιξη οργανικών όντων, ονομαστικά , πολλαπλασιάζονται , ποικίλλουν, και επιβιώνει ο πιο προσαρμοστικός” (Δαρβίνος, 1859).[64] Επιλέγεται να χρησιμοποιηθεί γενικά στην εξόρυξη δεδομένων γιατί αποτελεί μια εύρωστη και προσαρμοστική μέθοδο αναζήτησης που αναζητά το ολικό ακρότατο.[65]

Αρχικά είναι χρήσιμο να δοθούν οι ανάλογες υπολογιστικές ορολογίες της βιολογικής έκφρασης της εξέλιξης. Αναλύονται με προσανατολισμό την κατανόηση του διαφορικού εξελεγκτικού αλγόριθμου που χρησιμοποιήθηκε. Ξεκινώντας υπάρχει ένα πλήθος από δομές (π.χ. διανύσματα), ο λεγόμενος πληθυσμός (population), του οποίου το πλήθος διατηρείται σε κάθε γενεά (generation) σταθερό, εκτός αν μεταβάλλεται δυναμικά. Μέσω μια διαδικασίας επιλογής διατηρούνται οι βέλτιστες δομές από τον πληθυσμό (επιβίωση). Η επιβίωση καθορίζεται από την τιμή που θα προκύψει από μια συνάρτηση προσαρμογής (fitness function) ή κάποια άλλη μέθοδο εξαγωγής συμπεράσματος. Οι δομές που διατηρούνται μεταλλάσσονται. Από τον εκάστοτε αλγόριθμο καθορίζεται η μεθοδολογία επιλογής των δομών που θα μεταλλαχθούν. Η νέα λύση που θα προκύψει καθορίζεται από μια μαθηματική έκφραση που χρησιμοποιεί μια προκαθορισμένη σταθερά μετάλλαξης (mutation constant). Από την μετάλλαξη προκύπτουν οι γονείς, που μέσω κάποιου κανόνα που περιλαμβάνει τη τιμή της πιθανότητας διασταύρωσης (crossover probability), αναπαράγονται τα παιδιά -ένα νέο σύνολο λύσεων. Επαναξιολογούνται οι λύσεις που είχαν διατηρηθεί καθώς και τα παιδιά, και προκύπτει μέσω λογικών σχέσεων ποιοι θα προχωρήσουν στη νέα γενεά. [66]

```
procedure EA; {
t = 0;
initialize population P(t);
evaluate P(t);
until (done) {
        t = t + 1;
        parent_selection P(t);
        recombine P(t);
        mutate P(t);
        evaluate P(t);
        survive P(t);
}       }
```

Fig. 1. A typical evolutionary algorithm

Εικόνα 5: Απεικόνιση απλής υλοποίησης εξελεγκτικού αλγορίθμου (William M. Spears et al., An Overview of Evolutionary Computation)

Ανάλογα την υλοποίηση τους, οι απλοί αλγόριθμοι μπορούν να ενταχθούν σε μια από τις παρακάτω κατηγορίες[67]:

- Απλός Γενετικός αλγόριθμος (Genetic Algorithm)\
- Εξελεγκτικής στρατηγικής (Evolutionary Stategies)
- Εξελεγκτικού προγραμματισμού (Evolutionary Programming)
- Κατευθυντικής αναζήτησης (Direction based search)

Όπου ο Διαφορικός εξελεγκτικός αλγόριθμος (Differential Evolution Algorithm-DEA), ο αλγόριθμος που υλοποιείται στο υπολογιστικό κομμάτι της εργασίας, ανήκει στην τελευταία κατηγορία και συγκεκριμένα στην τυχαία (random) κατευθυντική αναζήτηση. Είναι ένας στοχαστικός αλγόριθμος για δύσκολα υπολογιστικά προβλήματα βελτιστοποίησης. Η υλοποίηση του είναι εύκολη και βιβλιογραφικά υποστηρίζεται ότι έχει ικανοποιητικά αποτελέσματα με συνεχείς μεταβλητές. Ως εξελεγκτικός αλγόριθμος παρουσιάζει προβλήματα υπολογιστικού φόρτου. [68] Η υλοποίηση του περιγράφεται από τη παρακάτω διαδικασία.

Ένας πληθυσμός λύσεων σε μια γενεά G εξελίσσεται πραγματοποιώντας τα ακόλουθα βήματα για κάθε λύση $\mathbf{x}_{iG}$ του πληθυσμού:

1. Μετάλλαξη (F=σταθερά μετάλλαξης)

   Επιλέγονται δύο τυχαίες λύσεις **y** και **z** (**y**≠**z**≠$\mathbf{x}_{iG}$) από τον πληθυσμό και συνδυάζονται με την καλύτερη λύση του πληθυσμού $\mathbf{b}_G$: $\mathbf{v}_{iG} = \mathbf{b}_G + F(\mathbf{y} - \mathbf{z})$ (10)

2. Διασταύρωση (CR = πιθανότητα διασταύρωσης)

   Συνδυασμός της λύσης $\mathbf{v}_{iG}$ με τη λύση $\mathbf{x}_{iG}$ για τη δημιουργία μιας νέας λύσης $\mathbf{u}_{iG}$, η οποία προέρχεται κατά CR% από τη $\mathbf{v}_{iG}$ και κατά (1–CR)% από την $\mathbf{x}_{iG}$.

3. Επιλογή

   Η λύση $\mathbf{x}_{iG}$ διατηρείται στην επόμενη γενεά μόνο εάν είναι καλύτερη της $\mathbf{u}_{iG}$, διαφορετικά αντικαθίσταται από την $\mathbf{u}_{iG}$.

Με μοντέλο απόφασης: $\boldsymbol{f(g) = b_1 g_1 + b_2 g_2 + \ldots + b_n g_n}$ **με** $\boldsymbol{\Pi b \Pi_2 = 1}$ (11)

όπου κάθε λύση στον διαφορικό αλγόριθμο περιλαμβάνει ένα πραγματικό μέρος με τους συντελεστές b_1, b_2, …, b_n των χαρακτηριστικών στο μοντέλο απόφασης και ένα δυαδικό μέρος με στοιχεία 0/1 που υποδεικνύει ποιες μεταβλητές συμμετέχουν στο μοντέλο. Τέλος, η συνάρτηση προσαρμογής περιγράφει τον συνδυασμό της ακρίβειας του μοντέλου και του πλήθος των χαρακτηριστικών που εξετάζει:

$$Z(\mathbf{x}) = \lambda A(\mathbf{x}) + N(\mathbf{x}) \quad (12)$$

$A(\mathbf{x})$: Ακρίβεια του μοντέλου που ορίζεται από τη λύση **x**,όπου εκφράζεται μέσω είτε του (α) δείκτη lift, (β) είτε του δείκτη Gini, (γ) είτε και των δύο (lift + Gini)/2 ανάλογα την επιλογή του αναλυτή.

$N(\mathbf{x})$: Ποσοστό των χαρακτηριστικών που εξαιρούνται από την ανάλυση και (λ) συντελεστή παραχώρησης μεταξύ της ακρίβειας και της πολυπλοκότητας του μοντέλου. Ο αλγόριθμός τερματίζεται όταν δημιουργήσει το πλήθος των γενεών που έχει προκαθοριστεί από τον αναλυτή.

2.2 ΣΥΝΟΨΗ

Το αποτέλεσμα μιας καλής ανάλυσης μέσω της εξόρυξης δεδομένων ξεκινάει από τη συλλογή των δεδομένων και εκχώρηση στην εκάστοτε βάση δεδομένων. Ένα σημαντικό κομμάτι προς αντιμετώπιση είναι των αδόμητων δεδομένων και των κενών τιμών. Επίσης, η ορθή επεξεργασία και τοποθέτηση κατά την εκχώρηση στην αποθήκη δεδομένων αποτελεί ζητούμενο για την περαιτέρω ανάλυση. Η επεξεργασία τους μετέπειτα και κατάλληλη επιλογή μεταβλητών είναι μια εργασία που είναι καλό να πραγματοποιείται με σεβασμό στο γεγονός ότι το σύνολο δεδομένων που θα προκύψει παίζει καθοριστικό ρόλο στην αποτελεσματικότητα του εκάστοτε αλγόριθμου.

Οι αλγόριθμοι που παρουσιάστηκαν, προέρχονται από διαφορετικές οικογένειες με διαφορετικές αδυναμίες ο καθένας. Δεν υπάρχει ένας αλγόριθμος που λειτουργεί σε όλες τις περιπτώσεις καλύτερα από τους υπόλοιπους. Κριτήριο για την αποτελεσματικότητα είναι τα χαρακτηριστικά του συνόλου εκπαίδευσης και το παραχώρηση μεταξύ ακρίβειας και υπολογιστικού φόρτου. Συνεπώς, η μελέτη των αντίστοιχων μοντέλων και η γνώση των ιδιομορφιών του συνόλου δεδομένων είναι ζωτικής σημασίας, ώστε να γίνει κατάλληλη αξιοποίηση των πόρων και να παραχθεί πληροφορία και γνώση αξιοποιήσιμη από την επιχείρηση.

Ένα εργαλείο για την εξαγωγή συμπερασμάτων ως προς τους αλγορίθμους και τα αντίστοιχα σύνολα δεδομένων , είναι οι συγκριτικές μελέτες. Στην παρούσα εργασία, μελετώνται αλγόριθμοι μηχανικής εκμάθησης με προσανατολισμό την πρόβλεψη και στόχο την ταξινόμηση των πελατών σε μη αποχωρήσαντες. Τα αποτελέσματα που θα προκύψουν θα έχουν άμεση εξάρτηση από τα δεδομένα. Οι ιδιαιτερότητες του συνόλου δεδομένων περιγράφονται σε αντίστοιχο κεφάλαιο. Τα παραπάνω αποτελούν εργαλείο πρόβλεψης αποχώρησης πελατών(churn prediction) για το τμήμα διαχείρισης πελατειακών σχέσεων (crm). Η πληροφορία που προκύπτει μετατρέπεται σε γνώση μέσω του custtomer analytics για να αξιοποιηθεί στις αντίστοιχες στρατηγικές που έχουν αναπτυχθεί σε συνεργασία με το τμήμα διαχείρισης αποχώρησης πελατών (churn prediction management). Τα ίδια αποτελέσματα ανάλογα τον προσανατολισμό της εκάστοτε στρατηγικής μπορεί να έχουν διαφορετική αξία για την επιχείρηση. Οπότε ακόμα και η κατάλληλη επιλογή δεικτών αξιολόγησης των αποτελεσμάτων, ως προς τη στρατηγική της επιχείρησης, αποτελεί παράμετρο αξιολόγησης της αποτελεσματικότητας του αλγόριθμου.

3 CUSTOMER ANALYTICS & ΔΙΑΧΕΙΡΙΣΗ ΠΕΛΑΤΕΙΑΚΩΝ ΣΧΕΣΕΩΝ (CRM)

Με τη πάροδο των χρόνων από το profiling (χαρακτηρισμός των πελατών με βάση μια ομάδα χαρακτηριστικών) των πελατών περάσαμε στη συμπεριφορική ανάλυση των πελατών της εκάστοτε επιχείρησης με χρήση εργαλείων analytics, τον κλάδο customer analytics. Η ανάπτυξη του συγκεκριμένου κλάδου προέκυψε από την ανάγκη κατανόησης των πελατών με πιο αποτελεσματικά εργαλεία από αυτά που προσέφερε η μέχρι τότε κατηγοριοποίηση (segmentation) [69]. Στόχος του customer analytics. η χρήση της εξαχθείσας γνώσης στις στρατηγικές αποφάσεις της επιχείρισης της διαχείρισης πελατειακών σχέσεων (customet relationship management).[14]

Η συλλογή των δεδομένων για τη συμπεριφορική ανάλυση των πελατών προέρχεται από ποικίλες πηγές, όπως τα κέντρα εξυπηρέτησης, τα δεδομένα χρήσης και δημογραφικά χαρακτηριστικά από τις καρτέλες των πελατών. Για να επιτευχθεί απαιτείται να υπάρχει η κατάλληλη δομή αποθήκευσης και ανάκλησης δεδομένων. Επίσης, καθοριστικό ρόλο παίζει η ποιότητα των δεδομένων. Τα παραπάνω στοιχειοθετούνται και αναλύονται στην προηγούμενη ενότητα. Σε αυτό το σημείο, αξίζει να αναφερθούν και τα ζητήματα που προκύπτουν σχετικά με τα προσωπικά δεδομένα των πελατών. Σύμφωνα με τους Thomas H. Davenport et al. [70], από την πλευρά της επιχείρησης πρέπει να γίνεται τέτοια διαχείριση τους, ώστε να συνάδει με τις αξίες της επιχείρησης. Λεπτή γραμμή αποτελεί, ως προς αυτό το ζήτημα αποτελεί η μεταπώληση αυτών των δεδομένων[71]. Παρότι έχουν νομοθετηθεί πλαίσια προστασίας των πελατών, αποτελεί φλέγον ζήτημα καθώς υπάρχουν αρκετές περιπτώσεις παραβίασης[72]. Επίσης, υπάρχουν κυβερνητικές προσπάθειας για την αντίστοιχη άρση απορρήτου[73].

Η επεξεργασία και διαχείριση της αντίστοιχης γνώσης γίνεται από τα τμήματα πληροφορικής (IT), Analytics, CRM, Marketing και των υπευθύνων για λήψη στρατηγικών αποφάσεων της επιχείρησης. Η εξαχθείσα γνώση μπορεί να χρησιμοποιηθεί ενδεικτικά για την ανάπτυξη στρατηγικών marketing και πρόληψης αποχώρησης πελατών . Ουσιαστικά υπάρχουν δύο χρήσεις της συμπεριφορικής ανάλυσης , η παραχώρηση του πελάτη με βάση κάποια χαρακτηριστικά σε μια κατηγορία (segmetation) και η πρόβλεψη αντίστοιχων αποφάσεων του (prediction), όπως αναφέρεται στην ενότητα του Business analytics Στην παρούσα εργασία εξετάζεται από την πλευρά της πρόβλεψης με στόχο το διαχείριση απώλειας πελατών (churn managment).

Πριν προχωρήσουμε παρακάτω, αξίζει να γίνει μια σύντομη αναφορά συνολικά στο τμήμα διαχείρισης πελατειακών σχέσεων. Αρμοδιότητες του τμήματος είναι συλλογή και επεξεργασία δεδομένων των πελατών, σε συνεργασία με άλλα τμήματα της επιχείρησης. Με με χρήση εφαρμογών CRM και ERP (Enterprise Resource Planning), προσανατολίζεται στη μοντελοποίηση της συμπεριφοράς και των προτιμήσεων των πελατών. Ταυτόχρονα, είναι υπεύθυνο για τη διαχείριση τους και ανάπτυξη στρατηγικών marketing/επικοινωνίας, ενώ διαχειρίζεται την απώλεια πελατών. Στόχος, η κατανόηση της πελατειακής βάσης της επιχείρησης και διατήρησης της. Επιπροσθέτως, αποσκοπεί στην αποτελεσματική εξυπηρέτηση των υπαρχόντων πελατών με ταυτόχρονη προσέλκυση νέων . Οι νέοι πελάτες προσελκύονται με το κριτήριο να είναι συμφέροντες για την επιχείρηση. Συνολικοί στόχοι η ικανοποίηση των πελατών, η κερδοφορία και

η ανάπτυξη της εταιρείας.[74]

3.1 ΔΙΑΧΕΙΡΙΣΗ ΑΠΟΧΩΡΗΣΗΣ ΠΕΛΑΤΩΝ

Αποτελεί βασικό κομμάτι της διαχείρισης πελατειακών σχέσεων. Το υψηλό ποσοστό αποχώρησης πελατών, ιδίως σε εταιρείες τηλεπικοινωνιών, το καθιστά ζωτικό για τη λειτουργιά της επιχείρησης. Το αντίστοιχο κόστος απώλειας και προσέλκυσης νέου πελάτη είναι υψηλό σε σχέση με τη διατήρηση υπάρχοντος, οπότε αναπτύσσονται στρατηγικές διαχείρισης αποχώρησης πελατών[75].

Προκειμένου να γίνουν στοχευμένες προσπάθειες πρόληψης αποχώρησης πελατών, απαιτείται με κάποιο τρόπο να εξαχθούν ασφαλή συμπεράσματα για το ποιοι ακριβώς πελάτες πρόκειται να αποχωρήσουν. Σε διαφορετική περίπτωση, εσφαλμένη πρόβλεψη, μπορεί να οδηγήσει την επιχείριση σε ζημία, επένδυση σε πελάτες που σκόπευαν να παραμείνουν και μη διατήρηση των τελικά αποχωρησάντων πελατών. Ακόμα και αν γίνει σωστή στόχευση ως προς τους εν δυνάμει αποχωρήσαντες μπορεί να οδηγηθεί σε σπατάλη πόρων,αν δεν έχουν ταξινομηθεί ορθά οι πελάτες με πρόθεση να παραμείνουν στην επιχείρηση. Οπότε, ακόμα και προληπτικές κινήσεις χωρίς αξιολόγηση των προβλέψεων δύναται να μην αποφέρουν τα επιθυμητά αποτελέσματα.[76]

Παραπάνω πραγματοποιείται η παραδοχή ότι η στρατηγική διαχείρισης πελατών θα είναι ορθή. Η ανάπτυξη αντίστοιχων στρατηγικών αποτελούν πολύπλευρο πρόβλημα, καθώς πέραν της σωστής πρόβλεψης πρέπει να πραγματοποιηθεί σωστή συμπεριφορική ανάλυση του πελάτη και να εφαρμοστεί με βάση αυτή η σωστή στρατηγική[77]. Ανάλογα το είδος πελάτη, ως προς τη κατηγορία που κατατάσσεται, διαφοροποιείται η αξία διατήρησης του[78].

Παρακάτω αναλύονται τα είδη πελατών, το κόστος απώλειας γενικευμένα και συγκεκριμένα για εταιρείες τηλεπικοινωνιών, όπου εμφανίζουν υψηλά ποσοστά αποχώρησης. Στα επόμενα βήματα, της ανάπτυξη στρατηγικών και της πρόληψης αποχώρησης, γίνεται σύντομη αναφορά στο τέλος της ενότητας. Στη παρούσα εργασία, μελετάται η ταξινόμηση των πελατών σε αποχωρήσαντες και μη, αξιολογώντας την προβλεπτική ικανότητα των μοντέλων πρόβλεψης που αναπτύσσονται, στα πλαίσια πρόβλεψης αποχώρησης πελατών (churn prediction).

3.1.1 ΚΑΤΗΓΟΡΙΑ ΑΠΟΧΩΡΗΣΗΣ

Υπάρχουν δύο είδη αποχώρησης, η οικειοθελής (voluntary) και η μη οικειοθελής (unvoluntary). Στην πρώτη εμπίπτουν οι πελάτες που οι ίδιοι επιλέγουν να αποχωρήσουν. Στη δεύτερη κατηγορία ανήκουν οι πελάτες όπου αναγκάζονται να αποχωρήσουν παρά τη θέληση τους. Από επιχειρησιακής πλευράς, αξίζει να αναπτυχθούν στρατηγικές διατήρησης συγκεκριμένων ομάδων πελατών που αποχωρούν οικειοθελώς. [79]

Οι μη οικειοθελώς αποχωρούντες αποτελούν, σε κάποιες περιπτώσεις, σημαντικό ποσοστό των αποχωρήσεων. Για το λόγο αυτό αναπτύσσονται τόσο στρατηγικές αντιμετώπισης τέτοιων περιπτώσεων, ώστε να αποφευχθεί η ζημιογόνος διατήρηση τέτοιων πελατών, όσο και να αποφευχθεί η προσέλκυση και ένταξη τους στο πελατολόγιο. Χωρίζονται σε τρεις βασικές κατηγορίες, διακοπή παροχής υπηρεσιών από την εταιρεία λόγω απάτης, λόγω προβλημάτων πίστωσης/εξόφλησης οφειλών και εξαιτίας μη χρήσης των αντίστοιχων υπηρεσιών.[80]

Από τους οικειοθελώς αποχωρήσαντες πελάτες, συγκεκριμένες κατηγορίες τους δύναται και αξίζει να διατηρηθούν. Μια κατηγορία για την οποία δεν μπορεί να προβλέψει η εταιρεία την αποχώρηση

και δε μπορεί να την εμποδίσει, είναι η οικειοθελής αλλά χωρίς πρόθεση/απρόσμενη αποχώρηση (incidental voluntary churn), η οποία αντιστοιχεί σε μικρό ποσοστό της συνολικής αποχώρησης. Αναφέρεται σε περιπτώσεις αλλαγής κατάστασης, π.χ. κατοικίας, όπου δεν παρέχεται εκεί η αντίστοιχη υπηρεσία/ προϊόν από την εταιρεία, αλλαγής οικονομικής κατάστασης και άλλες αντίστοιχες αλλαγές στη ζωή του πελάτη, όπως θάνατος, όπου ούτε ο πελάτης ούτε η εταιρεία μπορεί να κάνει κάτι για να αποφευχθεί η αποχώρηση.[79]

Οι υπόλοιπες κατηγορίες εθελουσίας αποχώρησης εμπεριέχουν τα είδη πελατών που αξίζει να αναπτυχθούν στρατηγικές διατήρησης. Η απόφαση αποχώρησης μπορεί να προκύπτει από κάποιους από τους παράγοντες που εμφανίζονται στον πίνακα. Δεδομένου ότι η εργασία υλοποιείται με δεδομένα από εταιρεία τηλεπικοινωνιών, η περιγραφή θα είναι από το αντίστοιχο πεδίο. [81]

Πίνακας 8: Αιτίες ηθελημένης με πρόθεση αποχώρησης

ΑΙΤΙΕΣ	ΠΕΡΙΓΡΑΦΗ
Επίπεδο τεχνολογικά παρεχόμενων υπηρεσιών & προϊόντων.	Δυνατότητες παρεχόμενης συσκευής, εύρος σήματος, λειτουργικές δυνατότητες υπηρεσιών
Ωρίμανσης του πελάτη	Τεχνολογικά προϊόντα/υπηρεσίες που πελάτες τα υιοθέτησαν πρώιμα , κατά την ανάπτυξη τους. Αποχωρούν όταν ωριμάσουν ως πελάτες.
Οικονομικοί παράγοντες	Τιμή συμβολαίου, χρεώσεις.
Παρεχόμενη ποιότητα	Εξυπηρέτηση πελατών, ποιότητα σήματος

3.1.2 ΕΙΔΗ ΠΕΛΑΤΩΝ ΚΑΙ ΑΞΙΑ ΓΙΑ ΤΗΝ ΕΠΙΧΕΙΡΗΣΗ

Οι πελάτες σε μια επιχείρηση είθισται να χωρίζονται σε αφοσιωμένους και μη, όπου οι μη αφοσιωμένοι αναφέρονται και ως χαμένοι πελάτες. Αυτό συμβαίνει καθώς οι εν δυνάμει αποχωρήσαντες πελάτες, συνήθως ανήκουν στους μη αφοσιωμένους. Οι αφοσιωμένοι πελάτες είναι πιθανοί προς αποχώρηση όταν παρατηρείται αλλαγή στη συμπεριφορά τους. Για το λόγο αυτό αναπτύσσονται στρατηγικές διατήρησης των αφοσιωμένων πελατών και μετατροπής των χαμένων σε αφοσιωμένων.[82] Η επιχείρηση προσανατολίζεται στη δημιουργία αφοσιωμένων πελατών γιατί φαίνεται να συμμετέχουν πιο δυναμικά στην κερδοφορία της.[83]

Υπάρχουν τριών ειδών σχέσεις πελάτη-εταρείας, που οδηγούν τον πελάτη να χαρακτηριστεί ως αφοσιωμένος. Η σχέση εξάρτησης του πελάτη λόγω περιορισμών παρατηρείται σε περιπτώσεις όπως, όταν ο πελάτης δεν είναι ικανοποιημένος αλλά η συγκεκριμένη εταιρεία αποτελεί την καλύτερη εναλλακτική του. Έπειτα, υπάρχουν οι σχέσεις εξάρτησης λόγω αφοσίωσης, όπου οι περιορισμοί τους είναι συναισθηματικοί, εναλλακτικά μπορεί να χαρακτηριστεί ως συναισθηματικά αφοσιωμένος πελάτης. Η τρίτη κατηγορία αναφέρεται στους πελάτες που είναι ικανοποιημένοι από την εταιρεία και επιλέγουν με αντικειμενικά κριτήρια να είναι αφοσιωμένοι. Στις δύο πρώτες περιπτώσεις ο πελάτης θεωρεί ότι η δέσμευση με την εταιρεία είναι η μόνη εναλλακτική ενώ στην τρίτη επιλέγει να δεσμευτεί με την εταιρεία. Γίνεται αναφορά στις σχέσεις αυτές καθώς τα προγράμματα αφοσίωσης οφείλουν να διαφοροποιούνται ως προς τη σχέση του αφοσιωμένου πελάτη με την εταιρεία.[84] Όσον αφορά τα προγράμματα αφοσίωσης, πρέπει να επιλέγει σοφά το

περιεχόμενο και η τακτική εφαρμογής τους για να μην καταστήσει τους αφοσιωμένους πελάτες ζημιογόνους για την επιχείρηση.[82]

Η αφοσίωση του εκάστοτε πελάτη δύναται να μετρηθεί μέσω κάποιων δεικτών αξιολόγησης των χαρακτηριστικών του, όπως αναφέρει το Loyalty Research Center. Η τιμή που εκφράζει συνολικά το ποσοστό των αφοσιωμένων πελατών μιας επιχείρησης, είναι πιο εύκολο να μετρηθεί μέσω της χρήσης ενός απλοϊκού τύπου[85]:

$$Αφοσίωση\ πελατών = 1 - Απώλεια\ πελατών \quad (13)$$

Ουσιαστικά εξαρτάται από το ποσοστό απώλεια πελατών, όπου είναι μετρήσιμο και το καθιστά εργαλείο για την εκτίμηση της κατάστασης της εταιρείας.

Οι υπόλοιποι πελάτες θεωρούνται ως χαμένοι και προβλέπεται να ενταχθούν σε μια από τις κατηγορίες οικειοθελούς αποχώρησης κατ' επιλογή. Όποτε στόχος, είναι να γίνουν προληπτικές κινήσεις ώστε να διατηρηθούν. Έπειτα μέσω προγραμμάτων αφοσίωσης, θα συνεχίσει να διατηρείται η πελατειακή βάση της εταιρείας, έχοντας μειώσει τη ζημία από τις απώλειες και αυξάνοντας ταυτόχρονα το τζίρο με προσέλκυση νέων πελατών. Η προσέλκυση μπορεί να είναι τέτοια ώστε αυτοί που θα ενταχθούν στο πελατολόγιο,να διακρίνονται από χαρακτηριστικά που επιθυμεί η εταιρεία και με υψηλή πιθανότητα να ενταχθούν στους αφοσιωμένους.

Γενικά, Η αξία ενός πελάτη εκφράζεται ως η αξία κύκλου ζωής του πελάτη (Lifecycle Value-LCV). Έχουν αναπτυχθεί ποίκιλες μεθοδολογίες εκτίμησης του, καθώς συμμετέχει στη λήψη στρατηγικών αποφάσεων της διαχείρισης των πελατειακών σχέσεων. Στη μέτρηση της συνυπολογίζεται το κόστος προσέλκυσης και διατήρησης του πελάτη. Όπου, το συνολικό κόστος μιας στρατηγικής προσέλκυσης απορροφάται από τους πελάτες που τελικά κατάφερε να προσελκύσει. Μια έκφραση του LTV:

$$LTV = \sum_{i=1}^{n} \frac{m_i * r^{(i-1)}}{(1+\delta)^{(i-1)}}, όπου\ r\ ποσοστό\ διατήρησης\ πελατών \quad (14)$$

Εκφράζεται συναρτήσει του χρόνου παραμονής στην επιχείρηση. Χρησιμοποιεί τη λογική της Καθαράς Παρούσας Αξίας.[86]

3.1.2.1 ΚΟΣΤΟΣ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ

Έστω ότι επιδιώκει η επιχείρηση και επιτυγχάνει να αντικαταστήσει κάθε αποχωρήσαντα με ένα νέο πελάτη. Σε αυτό το σημείο αξίζει να αναφερθεί ότι το κόστος προσέλκυσης είναι υψηλότερο από το κόστος διατήρησης. Επίσης, η αξία κύκλου ζωής του πελάτη, εφόσον έχουν παρθεί σωστές στρατηγικές αποφάσεις, είναι μια θετική τιμή- κέρδος για την επιχείρηση. Οπότε αυτή η αντικατάσταση σημαίνει αρνητικό ισοζύγιο για την επιχείρηση. Γενικά, σε επίπεδο επιχείρησης η απώλεια πελατών εκφράζεται ως:

$$Απώλεια\ πελατών = 1 - Διατηρούντες\ πελάτες(r) \quad (15)$$

Μια άλλη παράμετρο που καθορίζει η απώλεια πελατών είναι ο μέσος χρόνος παραμονής του πελάτη στην επιχείρηση. Όπως παρατηρείται από το LTV η αξία του πελάτη αυξάνεται σε βάθος χρόνου. Ο αναμενόμενος μέσος χρόνος παραμονής προκύπτει από:

$$Αναμενόμενος\ μέσος\ χρόνος\ παραμονής = \frac{1}{Απώλεια\ πελατών} \quad (16)$$

Οπότε μείωση της απώλεια αυξάνει το μέσο χρόνο παραμονής και αντιστρόφως Αν χρησιμοποιηθεί η παρούσα έκφραση στο LTV, φαίνεται ξεκάθαρα η επιρροή που έχει στα κέρδη της επιχείρησης. Υψηλά ποσοστά απώλειας πελατών εμφανίζουν οι κλάδοι παροχής υπηρεσιών, όπως ασφαλιστικές εταιρείες, τράπεζες, εταιρείες παροχής σύνδεσης στο διαδίκτυο και οι εταιρείες τηλεπικοινωνιών. [87]

3.1.3 ΣΤΡΑΤΗΓΙΚΕΣ ΔΙΑΤΗΡΗΣΗΣ ΠΕΛΑΤΩΝ

Υπάρχουν δύο είδη στρατηγικών που μπορούν να ακολουθηθούν, η στοχευμένη (targetted) και η γενική (untargetted). Ξεκινώντας με τη δεύτερη, προσπαθεί συνολικά να μεταβάλλει παραμέτρους, όπως αύξηση ικανοποίησης, μέσω ενεργειών που θα επηρεάσουν όλο το φάσμα των πελατών. Στη συγκεκριμένη κατηγορία ανήκουν και τα προγράμματα αφοσίωσης. Όπως, προαναφέρθηκε ένα ζητούμενο σε αυτά τα προγράμματα είναι η επιλογή της βέλτιστης επένδυσης ώστε να μην μειωθεί το καθαρό κέρδος που προκύπτει ανά πελάτη.

Οι στοχευμένες στρατηγικές, εντοπίζουν τους πιθανούς πελάτες προς αποχώρηση και προσανατολίζονται σε αυτούς. Η δράση μπορεί να είναι προληπτική (proactive) ή αντιδραστική (reactive), ανάλογα τον τρόπο που εντοπίστηκε ο πελάτης. Στην περίπτωση του reactive διαπιστώνεται ότι ένας πελάτης θα αποχωρήσει, τη στιγμή που ο ίδιος το κοινοποιεί μέσω κάποιας ενέργειας του, ενδεικτικά κλήση σε τηλεφωνικό κέντρο για ακύρωση της συνδρομής. Του γίνεται άμεσα κάποια προσφορά ή κίνηση, με βάση τις ανάγκες του, ώστε να αποτρέψει την αποχώρηση του.

Στις proactive στρατηγικές, εντοπίζεται μέσω κάποιων ενεργειών που παρέχουν πρόβλεψης, όπως predictive analytics, στοχαστικά ποιοι θα αποχωρήσουν. Έπειτα διευρενάται το γιατί και δομείται ένα πλάνο ενεργειών για την αποφυγή της αποχώρησης τους. Όμως, λόγω της στοχαστικότητας, δύναται ο αντίστοιχος πελάτης να μην είχε ποτέ την πρόθεση να αποχωρήσει. Οπότε η προσπάθεια διατήρησης του, θα είναι μια άστοχη κίνηση με κόστος για την επιχείρηση. Επιπροσθέτως, έχει διαπιστωθεί ότι πελάτες που είχαν την τάση για αποχώρηση, άλλα δεν το είχαν συνειδητοποιήσει ακόμα, αποχώρησαν μετά από την εκτέλεση των αντίστοιχων ενεργειών από την επιχείρηση. Πάραυτα, παρότι μπορεί να θεωρηθεί σπατάλη, η χρήση τέτοιων τεχνικών μπορεί να οδηγήσει σε χαμηλότερη δαπάνη διατήρησης ενός πελάτη που θα αποχωρούσε, λόγω της έγκαιρης διάγνωσης της πρόθεσης του.

Οπότε στις περιπτώσεις στοχευμένης στρατηγικής με προληπτικό χαρακτήρα έχουν αναπτυχθεί μοντέλα απόφασης για την εφαρμογή αντίστοιχων ενεργειών. Με παραμέτρους, όπως την πιθανότητα αποχώρησης, υλοποιούνται οι αντίστοιχες στρατηγικές και ταυτόχρονα αξιολογείται το μοντέλο για την αποτελεσματικότητα του. Με την εφαρμογή κατάλληλου μοντέλου απόφασης και κριτηρίων αξιολόγησης του, μπορεί αυτές οι στρατηγικές να είναι cost-effective. [87]

3.2 ΠΡΟΒΛΕΨΗ ΑΠΟΧΩΡΗΣΗΣ ΠΕΛΑΤΩΝ

Αποτελεί εργαλείο για τη διαχείριση αποχώρησης πελατών. Χρησιμοποιεί μεθοδολογίες από τον χώρο εξόρυξης δεδομένων, για την ανάπτυξη μοντέλων πρόβλεψης αποχώρησης. Οι αλγόριθμοι που χρησιμοποιεί προσανατολίζονται στην πρόβλεψη με στόχο-έξοδο την ταξινόμηση των πελατών. Οι πελάτες ταξινομούνται σε σε εν δυνάμει ή μη αποχωρήσαντες, με τη χρήση δύο κλάσεων. Κάνοντας αναδρομή στην ενότητα του Business Analytics, μέσω predictive analytics και με χρήση αλγόριθμων μηχανικής εκμάθησης και στόχο την ταξινόμηση (classification), αναπτύσσεται ένα μοντέλο πρόβλεψης αποχώρησς πελατών. Σκοπός, μέσω customer analytics να εξαχθεί γνώση από την πληροφορία που παράσχει το μοντέλο,ώστε να αναπτυχθούν στρατηγικές διαχείρισης απώλειας πελατών, που θα υλοποιηθούν από το τμήμα διαχείρισης πελατειακών σχέσεων.

3.2.1 ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ (CLASSIFICATION)

Στο κομμάτι αυτό θα πραγματοποιηθεί μια σύντομη βιβλιογραφική αναφορά σχετικών συγκριτικών μελετών που έχουν πραγματοποιηθεί. Θα αναφερθούν οι αλγόριθμοι που επιλέχθηκαν και τα κύρια συμπεράσματα. Επίσης, θα γίνει αναφορά στους δείκτες αξιολόγησης που χρησιμοποιήθηκαν.

Ξεκινώντας, οι B. Huang et al. επιλέγουν για την μελέτη τους σχετικά με “Customer churn prediction in telecommunications”[88] να συγκρίνουν Λογιστική Παλινδρόμηση, NaiveBayes, Multilayerpercetron, Δέντρα απόφασης, Μηχανές διανυσμάτων υποστήριξης, Εξελεγκτικό αλγόριθμο και γραμμικούς ταξινομητές, μέ ένα εκ των δεικτών αξιολόγησης AUC. Κατέληξαν στο συμπέρασμα ότι SVM και δέντρα απόφασης είναι καλύτερα για αξιολόγηση ποσοστών των πελατών που κατετάγησαν ορθά και λανθασμένα ως αποχωρήσαντες. Η λογιστική παλινδρόμηση είναι καλύτερη για τον υπολογισμό των πιθανοτήτων αποχώρησης. Όσον αφορά τον εξελεγκτικό αλγόριθμο, η εφαρμογή του δεν είναι πρακτική σε προβλήματα πρόβλεψης αποχώρησης πελατών.

Σε άλλη μελέτη στο κλάδο παροχής υπηρεσιών σχετιζόμενες με την τηλεόραση (δορυφορική σύνδεση), μεταξύ άλλων επιλέγονται λογιστική παλινδρόμηση και random forest, με δείκτες αξιολόγησης Top Decile Lift και AUC. Προκύπτει ότι στο Lift ο Random Forest αποδίδει καλύτερα και στο AUC οι δύο αλγόριθμοι δεν έχουν σημαντικές αποκλίσεις.[89] Σε αυτη την κατεύθυνση αναπτύσσονται εξελιγμένα μοντέλα Random Forest, όπου και αποδίδουν ακόμα καλύτερα. [90]Συνολικά παρατηρήθηκε, ότι επιλέγεται η χρήση αλγορίθμων από τις οικογένειες λογιστικής παλινδρόμησης, τεχνητών νευρωνικών δικτύων, SVM, δέντρων απόφασης και μεταμαθησιακών μοντέλων για την εκτίμηση της πρόβλεψης αποχώρησης πελατών. Οι δείκτες που επιλέχθηκαν κατά βάση ήταν Top decile Lift και Gini ή AUC. Μέσω της σύντομης βιβλιογραφικής επισκόπησης επίσης παρατηρήθηκε ότι λειτουργεί αποτελεσματικά ο Random Forest και αντίστοιχοι προσαρμοσμένοι αλγόριθμοι.

4 ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ

Στη παρούσα διπλωματική πραγματοποιείται συγκριτική αξιολόγηση αλγορίθμων μηχανικής εκμάθησης για την ανάπτυξη μοντέλων πρόβλεψης αποχώρησης πελατών. Σε περιβάλλον Matlab υλοποιήθηκαν ο διαφορικός εξελεγκτικός αλγόριθμος (Differential Evolution Algorithm) και ο αλγόριθμος μηχανών υποστήριξης διανυσμάτων (Proximal Support Vector Machines). Σε περιβάλλον Weka υλοποιήθηκαν οι αλγόριθμοι:

- ΑΠΛΟΙ
    - Λογιστικής Παλινδρόμησης -Logistic,
    - Δίκτυου Bayes -BayesNet,
    - Δέντρων Απόφασης -SimpleCart
    - Τεχνητών Νευρωνικών Δικτύων -MultiLayerPerceptron
- ΣΥΝΔΥΑΣΤΙΚΟΙ (Ensembles)
    - Bagging με χρήση Randomized-Random Forest,
    - Boosting-AdaBoostM1,
    - Bagging

Ως είσοδο χρησιμοποιούνται πραγματικά δεδομένα από εταιρεία τηλεπικοινωνιών. Υπάρχουν δύο σύνολα δεδομένων, το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Πραγματοποιήθηκε αρχικά κατάλληλη επεξεργασία για εκκαθάριση ανεπαρκών δεδομένων και διαλογή κατάλληλων μεταβλητών, με τη χρήση του προγράμματος SPSS. Στη πορεία για κάποιους από τους αλγορίθμους χρησιμοποιήθηκε επιπλέον μεθοδολογία επιλογής μεταβλητών με τη χρήση του προγράμματος εξόρυξης δεδομένων Weka. Λόγω υπολογιστικού φόρτου, για τον διαφορικό εξελεγκτικό αλγόριθμο χρησιμοποιήθηκε μικρότερο πλήθος δεδομένων. Επίσης, η τελευταία μεθοδολογία διαλογής μεταβλητών (wrapper) εφαρμόστηκε μόνο σε δύο αλγορίθμους μηχανικής εκμάθησης. Παρακάτω, παρουσιάζονται οι δοκιμές με τις αντίστοιχες εισόδους δεδομένων (με ή χωρίς πρόσθετη διαλογή δεδομένων). Για τους αλγόριθμους που χρησιμοποιήθηκαν όλοι οι αποφασίζαντες (94680 σύνολο εκπαίδευσης/ 92605 σύνολο ελέγχου) αντιστοιχεί η τιμή (1) , για αυτούς που χρησιμοποιήθηκε μικρότερο πλήθος (25000 σύνολο εκπαίδευσης/ 50000 σύνολο ελέγχου) αντιστοιχεί η τιμή (2).

Πίνακας 9: Περιγραφή τελικών συνόλων δεδομένων εισόδου

	Σύνολο εκπαίδευσης		Σύνολο ελέγχου	
	# πελατών	% αποχωρήσαντες	# πελατών	% αποχωρήσαντες
Data set (1)	94680	49.64	92605	1.80
Data set (2)	25000	49.46	50000	1.84

Πίνακας 10: Σύνολα δεδομένων εισόδου κάθε πειράματος

	Χωρίς	ChiSquared	Gainratio	Infogain	Wrapper
BayesNet	(1)	(1)	(1)	(1)	(1) & (2)
Logistic	(1)	(1)	(1)	(1)	-
SimpleCart	(1)	(1)	(1)	(1)	-
MultiLayerPerceptron	(1)	(1)	(1)	(1)	-
Bagging	(1)	(1)	(1)	(1)	-
AdaBoostM1	(1)	(1)	(1)	(1)	(1) & (2)
PSVM	(1)	-	-	-	-
DEA	(2)	-	-	-	-
RandomForest	(1)	(1)	(1)	(1)	-

Οι δείκτες αξιολόγησης των αποτελεσμάτων που επιλέχθηκαν ήταν Area Under Curve (AUC) και Top decile lift καθώς προτείνεται τόσο από την εταιρεία που παρείχε τα δεδομένα όσο και βιβλιογραφικά για τέτοιες συγκριτικές μελέτες. Επίσης, επιλέχθηκε να χρησιμοποιηθούν precision και recall καθώς χρησιμοποιούνται παραδοσιακά σε αλγόριθμους μηχανικής εκμάθησης.

4.1 ΔΕΔΟΜΕΝΑ

Στη συγκριτική αξιολόγηση χρησιμοποιούνται πραγματικά δεδομένα χρηστών του έτους 2001 που είχαν δοθεί στα πλαίσια του διαγωνισμού “Churn modeling tournament” του Teradata Center for Customer Relationship Management στο Duke University, από εταιρεία τηλεπικοινωνιών της Αμερικής το έτος 2002. (Παράρτημα/Περιγραφή δεδομένων) Παρακάτω περιγράφονται τα δεδομένα.

4.1.1 ΠΕΡΙΓΡΑΦΗ

Τα δεδομένα που χρησιμοποιούνται είχαν συλλεχθεί τον Ιούλιο, Σεπτέμβριο, Νοέμβριο και Δεκέμβριο του 2001, με διάστημα μερικών εβδομάδων για κάθε μέτρηση, για να μπορεί το δείγμα να παρέχει προβλεπτική ικανότητα. Οι κλάσεις είχαν υπολογιστεί από το γεγονός αν ο εκάστοτε πελάτης αποχώρησε σε μια περίοδο 1-2 μηνών από τις αρχικές μετρήσεις του. Οι πελάτες που επιλέχθηκαν να συμμετέχουν στα δείγματα, χαρακτηρίζονται ως ώριμοι καθώς είχαν παραμείνει στην εταιρεία τουλάχιστον για 6 μήνες.(Παράρτημα/ Περιγραφή δεδομένων από την εταιρεία)

Τα σύνολα δεδομένων που αναπτύχθηκαν από τα παραπάνω στοιχεία, ήταν το Calibration data set , το Current score data και το Future score data, όπου κάθε σύνολο περιέχει πληροφορία για διαφορετικούς πελάτες. Στη παρούσα μελέτη χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης το Calibration data set και ως σύνολο ελέγχου το Future score data. Στα αρχικά δεδομένα δεν είχε δοθεί η κλάση για το σύνολο ελέγχου, αλλά με τη λήξη του διαγωνισμού δόθηκαν. Για το σύνολο εκπαίδευσης χρησιμοποιήθηκε over sampling (τροποποίηση δείγματος ώστε να αλλάξει ο καταμερισμός των κλάσεων), ώστε να δημιουργηθεί ένα δείγμα κατά προσέγγιση 50-50 ως προς αποχωρήσαντες και μη, προκειμένου να μην υπάρξει υπερπροσαρμογή (overfitting) κατά την εκπαίδευση, που θα οδηγούσε σε αστοχία στην ανάπτυξη των μοντέλων, όπως έχει αναλυθεί και στο αντίστοιχο κεφάλαιο.

Πίνακας 11: Χαρακτηριστικά συνόλου εκπαίδευσης και ελέγχου προ επεξεργασίας

	Calibration data set	Future score data
Μέγεθος δείγματος	100000	100462
Πλήθος χαρακτηριστικών	171	171
Ποσοστό αποχωρήσαντων	49.56%	1.80%

Κριτήριο ταξινόμησης αποτελεί, εάν ένας πελάτης προβλέπεται να αποχωρήσει (τιμή κλάσης 1) ή να παραμείνει στην επιχείρηση(τιμή κλάσης 0). Οπότε αποτελείται από 2 κλάσεις την 1(churner) και την 0 (non churner).

Τα χαρακτηριστικά αποτελούνται από 114 ποσοτικές (συνεχείς) μεταβλητές και 57 ποιοτικές (κατηγορικές) (Παράρτημα/Περιγραφή μεταβλητών). Στην πορεία υπήρξε κατάλληλη επεξεργασία για τις τιμές που έλειπαν (missing values), για τη μετατροπή των ποιοτικών σε ποσοτικές, καθώς και μια αρχική αφαίρεση μεταβλητών που φαίνεται να μην έχουν σημαντική επίδραση στο τελικό αποτέλεσμα.

4.1.2 ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΔΙΑΛΟΓΗ

Χρησιμοποιήθηκε μια αρχική διαλογή για την εκκαθάριση των δεδομένων. Στην πορεία, στη μελέτη συγκεκριμένων αλγορίθμων χρησιμοποιήθηκαν μεθοδολογίες διαλογής μεταβλητών, για να μελετηθεί η επίδραση των μεταβλητών καθώς και η επίδοση του εκάστοτε αλγόριθμου συναρτήσει αυτών.

4.1.2.1 ΑΡΧΙΚΗ ΔΙΑΛΟΓΗ

Τα δεδομένα που παραλήφθηκαν για τη συγκεκριμένη διπλωματική είχαν ήδη υποστεί επεξεργασία στα πλαίσια της υλοποίησης της μεταπτυχιακής διατριβής της Ιωάννας Αγγελιδάκη[91]. Αρχικά κρίθηκε βιβλιογραφικά ποιες μεταβλητές, οι οποίες παρουσίαζαν πολλά κενά πεδία, δεν θα είχαν σημαντική αξία για την ανάπτυξη των μοντέλων και αφαιρέθηκαν. Επίσης. χρήστες που είχαν πληθώρα ελλιπών στοιχείων (άνω του 1.7%) διεγράφησαν.

Έπειτα, όπου υπήρχε δυνατότητα, τα κενά πεδία μεταβλητών (ποσοτικές και κάποιες ποιοτικές) που είχαν σημαντική αξία αντικαταστάθηκαν, κατά περίπτωση, από μηδενικά. Στις μεταβλητές που τα μηδενικά δεν αποτελούσαν ικανοποιητική λύση τοποθετήθηκαν τιμές από τις οποίες θα προέκυπτε νόημα (υπόλοιπες ποιοτικές- η τιμή που τοποθετήθηκε δήλωνε έλλειψη τιμής). Στην πορεία, οι ποιοτικές μετατράπηκαν σε ποσοτικές και όπου μπορούσε και υπήρχε ανάγκη έγινε συγχώνευση (ομαδοποίηση) τιμών, για μικρότερο εύρος τιμών κατά την τροποποίηση τους σε ποσοτικές, με αντικατάσταση τους από ψευδοτιμές (αντιστοιχεί σε αριθμητική κωδικοποίηση). Οπότε, για την παρούσα εργασία παραλήφθηκαν 'επειτα από αυτή την επεξεργασία 149 μεταβλητές, εκ των οποίων 40 ποιοτικές (με την αντίστοιχη κωδικοποίηση) και 109 ποσοτικές.

Παραλαμβάνοντας τα παραπάνω σύνολα δεδομένων, έγινε αξιολόγηση όλων των μεταβλητών συναρτήσει της κλάσης (αποχωρήσασας πελάτης ή μη) μέσω της στατιστικής εφαρμογής SPSS. Χρησιμοποιήθηκε η μεθοδολογία αξιολόγησης/διαλογής δεδομένων Pearson X^2-test. Με τις προεπιλεγμένες τιμές από την εφαρμογή, για αυτή την μεθοδολογία, προέκυπτε ένας πίνακας που περιείχε τις τιμές για assymptotic significance (2-tailed). Όσο μικρότερη η τιμή της, τόσο μεγαλύτερη η εξάρτηση της τελικής απόφασης από την μεταβλητή. Ως όριο τέθηκε το 0.01 και όσες ήταν κάτω από αυτό το όριο αφαιρέθηκαν από τα αντίστοιχα σύνολα. Οι μεταβλητές που τελικά διατηρήθηκαν είναι 15 ποιοτικές και 108 ποσοτικές. Στο παράρτημα υπάρχει αναλυτική περιγραφή των μεταβλητών που διατηρήθηκαν.

Οπότε, οι αρχικές δοκιμές για κάθε αλγόριθμο που αναφέρεται ότι χρησιμοποίησε ως είσοδο τα δεδομένα χωρίς μεθοδολογία διαλογής (plain), χρησιμοποίησε αυτά που προέκυψαν από την παραπάνω επεξεργασία. Έπειτα από την παραπάνω επεξεργασία το σύνολο δεδομένων εκπαίδευσης αποτελούνταν από 94680 αποφασίζαντες και 92605 το σύνολο ελέγχου με 123 μεταβλητές.

4.1.2.2 ΜΕΘΟΔΟΛΟΓΙΕΣ ΔΙΑΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Αφού υλοποιήθηκαν οι αλγόριθμοι μέ είσοδο τα δεδομένα χωρίς διαλογή μεταβλητών, χρησιμοποιήθηκαν περαιτέρω μεθοδολογίες επιλογής μεταβλητών. Προέκυψαν νέα σετ δεδομένων, ώστε να αξιολογηθεί μετέπειτα ο ρόλος των μεταβλητών. Τα συγκεκριμένα σετ δεδομένων χρησιμοποιήθηκαν για το μεγαλύτερο μέρος των αλγορίθμων που εξετάζονται, οι υπόλοιποι

παραλήφθηκαν κυρίως λόγω υπολογιστικού φόρτου. Παρακάτω περιγράφονται αναλυτικά οι μεθοδολογίες καθώς και οι μεταβλητές που επιλέχθηκαν. Για τις πρώτες τρεις επιλέχθηκαν οι επικρατέστερες 30. Σε όλες τις μεθοδολογίες χρησιμοποιήθηκε ως επιλογή για το σετ εκπαίδευσης 'full training set'. Σε όλες τις μεθοδολογίες, εκτός του WrapperSubsetEval, χρησιμοποιείται η μέθοδος αναζήτησης 'Ranker' . Στο WrapperSubsetEval χρησιμοποιείται 'Best First'. Στα υπόλοιπα πεδία επιλέγονται για κάθε μεθοδολογία οι προκαθορισμένες από την εφαρμογή τιμές. Αρχικά παρουσιάζονται οι μεθοδολογίες εκτός της WrapperSubsetEval.

Πίνακας 12: Περιγραφή απλών μεθοδολογιών διαλογής μεταβλητών

ChiSquaredAttributeEval	Με βάση το Pearson X^2-test εκτιμά τη αξία της κάθε μεταβλητής για την τελική απόφαση του πελάτη.
GainRatioAttributeEval	Εκτιμά το gain ratio της εκάστοτε μεταβλητής (συναρτήσει της εντροπίας (H)) ως προς την τελική απόφαση του πελάτη GainRatio(Κλάση, A) = (H(Κλάσης) - H(Κλάσης | Μεταβλητή)) / H(Μεταβλητής).
InfoGainAttributeEval	Εκτιμά information gain της εκάστοτε μεταβλητής (συναρτήσει της εντροπίας (H)) ως προς την τελική απόφαση του πελάτη InfoGain(Κλάση,Μεταβλητή) = H(Κλάση) - H(Κλάση| Μεταβλητή).

Οι μεταβλητές που επιλέγονται για τις παραπάνω μεθοδολογίες παρουσιάζονται στον παρακάτω πίνακα. Στο παράρτημα υπάρχουν οι ανάλογες αντιστοιχίες.

Πίνακας 13: Πρώτες 30 μεταβλητές κάθε μεθοδολογίας

A/A	ChiSquaredAttributeEval	GainRatioAttributeEval	InfoGainAttributeEval
1	months	retdays	months
2	eqpdays	tot_ret	eqpdays
3	hnd_price	tot_acpt	hnd_price
4	totmrc_Mean	eqpdays	totmrc_Mean
5	mou_Mean	months	mou_Mean
6	change_mou	asl_flag	change_mou
7	hnd_webcap	hnd_webcap	hnd_webcap
8	totmrc_Range	totmrc_Range	totmrc_Range
9	adjrev	hnd_price	avg3mou
10	avg3mou	change_mou	adjrev
11	avg3qty	mou_peav_Range	avg3qty
12	totrev	avg3qty	totrev

13	crclscod	mou_cvce_Range	retdays
14	retdays	mou_Mean	crclscod
15	mou_opkv_Mean	avg3mou	mou_opkv_Mean
16	mou_cvce_Mean	totrev	mou_cvce_Mean
17	asl_flag	opk_vce_Range	asl_flag
18	tot_acpt	adjrev	tot_acpt
19	tot_ret	totmrc_Mean	tot_ret
20	ethnic	rev_Mean	ethnic
21	opk_vce_Mean	mou_peav_Mean	opk_vce_Mean
22	mou_peav_Mean	drop_blk_Mean	mou_peav_Mean
23	comp_vce_Mean	unan_vce_Range	comp_vce_Mean
24	iwylis_vce_Mean	mou_rvce_Range	iwylis_vce_Mean
25	complete_Mean	avg6mou	complete_Mean
26	mouiwylisv_Mean	cc_mou_Mean	mouiwylisv_Mean
27	peak_vce_Mean	iwylis_vce_Range	peak_vce_Mean
28	attempt_Mean	peak_vce_Mean	attempt_Mean
29	plcd_vce_Mean	dualband	avg6mou
30	avg6mou	mou_cvce_Mean	plcd_vce_Mean

Στην πορεία χρησιμοποιήθηκε η μεθοδολογία WrapperSubsetEval που δημιουργεί τα νέα σύνολα δεδμένων με βάση το σύστημα εκμάθησης του εκάστοτε αλγόριθμου, αναλύεται στην πρώτη ενότητα.[37] Ως αλγόριθμοι επιλέχθηκαν να χρησιμοποιηθούν μόνο, λόγω υπολογιστικού φόρτου, AdaBoostM1 και BayesNet δημιουργώντας δύο ξεχωριστά νέα σύνολα δεδομένων με λιγότερες μεταβλητές. Η συγκεκριμένη μεθοδολογία υλοποιήθηκε τόσο για το μεγάλο σύνολο δεδομένων όσο και για το μικρότερο (δείκτης (2)) για να μπορέσει να υπάρξει μετέπειτα σύγκριση και με τον διαφορικό εξελεγκτικό αλγόριθμο. Οι υπόλοιπες παράμετροι της μεθοδολογίας παρέμειναν όπως είχαν προκαθοριστεί από το πρόγραμμα.

Πίνακας 14: Επιλογή μεταβλητών από μεθοδολογία wrapper ανά αλγόριθμο

A/A	AdaBoostM1	BayesNet
1	totmrc_Mean	ovrmou_Mean
2	totmrc_Range	totmrc_Range
3	crclscod	change_mou
4	tot_ret	months
5	dualband	crclscod
6	age1	totcalls
7	eqpdays	tot_ret

8		area
9		refurb_new
10		hnd_webcap
11		age1
12		ethnic
13		eqpdays

Πίνακας 15: Επιλογή μεταβλητών από μεθοδολογία Wrapper για κάθε σύνολο δεδομένων ανά αλγόριθμο

A/A	BayesNet	BayesNet (2)
1	ovrmou_Mean	ovrrev_Mean
2	totmrc_Range	vceovr_Mean
3	change_mou	totmrc_Range
4	months	change_mou
5	crclscod	mouiwylisv_Mean
6	totcalls	months
7	tot_ret	crclscod
8	area	totcalls
9	refurb_new	avg3qty
10	hnd_webcap	tot_ret
11	age1	tot_acpt
12	ethnic	hnd_price
13	eqpdays	hnd_webcap
14		marital
15		infobase
16		ethnic
17		kid0_2
18		retdays
19		eqpdays

A/A	AdaBoostM1	AdaBoostM1 (2)
1	totmrc_Mean	change_mou
2	totmrc_Range	plcd_vce_Mean
3	crclscod	custcare_Mean
4	tot_ret	months
5	dualband	hnd_price
6	age1	
7	eqpdays	

4.2 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΕΚΜΑΘΗΣΗΣ

Δεδομένου ότι περιγράφηκαν αναλυτικά οι οικογένειες των αλγορίθμων αυτών στη θεωρία στο κεφάλαιο 2.1.2.2 , θα γίνει συνοπτική αναφορά. Εξαιρουμένου του εξελεγκτικού αλγορίθμου, χρησιμοποιήθηκαν προκαθορισμένες τιμές σε όλους τους αλγορίθμους. Όπως προαναφέρθηκε οι 7 αλγόριθμοι υλοποιήθηκαν σε περιβάλλον Weka, χρησιμοποιώντας την εφαρμογή. Οι άλλοι δύο υλοποιήθηκαν σε περιβάλλον Matlab. Θα γίνει πιο εκτενής αναφορά στον εξελεγκτικό αλγόριθμο καθώς μελετήθηκε η βελτιστοποίηση της παραμετροποίησης του μέσα από μια σειρά πειραμάτων. Για τη σύγκριση με τους υπόλοιπους επιλέχθηκαν οι 10 καλύτερες δοκιμές, όπως ορίζεται μέσω των δεικτών αξιολόγησης. Παρακάτω περιγράφονται πρώτα οι αλγόριθμοι που υλοποιήθηκαν μέσω της εφαρμογής Weka. Σε όλες τις περιπτώσεις το σύνολο εκπαίδευσης και έπειτα το σύνολο ελέγχου, χωρίς καμιά μεταβολή στις υπόλοιπες παραμέτρους. Σαν έξοδο έδιναν το μοντέλο, την ταξινόμηση, τα αποτελέσματα συναρτήσει του συνόλου ελέγχου και τιμές δεικτών αξιολόγησης. Οι δείκτες AUC, Precision και Recall, δίδονταν από την εφαρμογή μαζί με τα αποτελέσματα του εκάστοτε αλγορίθμου. Ο δείκτης Top decile Lift υπολογήστηκε μετέπειτα με τη χρήση των εξόδων της εφαρμογής.

Logistic

Η υλοποίηση που χρησιμοποιείται από το πρόγραμμα Weka έχει κάποιες διαφοροποιήσεις σε σχέση με την κλασσική του υλοποίηση. Τροποποιείται ο αλγόριθμος ώστε να μπορεί να μεταχειρίζεται παραδείγματα (instances) που τους έχουν αποδοθεί βάρη. Ακολουθεί συνοπτική περιγραφή όπου i και j δείκτες.

k:κλάσεις, n:παραδείγματα m:μεταβλητές, B:πίνακας παραμέτρων μεγέθους m*(k-1)
Pj(Xi) = exp(XiBj)/((sum[j=1..(k-1)]exp(Xi*Bj))+1) Πιθανότητα για j-οστή κλάση, j=[1,..,k-1]
1-(sum[j=1..(k-1)]Pj(Xi))= 1/((sum[j=1..(k-1)]exp(Xi*Bj))+1) Για k-οστή κλάση
(αρνητική) πολυονιμική λογαριθμική πιθανοφάνεια
L = -sum[i=1..n]{sum[j=1..(k-1)](Yij * ln(Pj(Xi)))+(1 - (sum[j=1..(k-1)]Yij))* ln(1 - sum[j=1..(k-1)]Pj(Xi))} + ridge * (B^2)
Για εύρεση B τ.ω. minL, γίνεται χρήση Quasi-Newton μεθόδου για βελτιστοποίηση των μεταβλητών

Με προκαθορισμένες τιμές παραμέτρων:

- Τιμή Ridge για λογαριθμική συνάρτηση πιθανοφάνειας: $1 * 10^{-8}$
- Χωρίς προκαθορισμένο όριο μέγιστου πλήθους επαναλήψεων

Multilayeroerceptron

Χρησιμοποιεί την υλοποίηση που είχε περιγραφεί στη θεωρητική ενότητα των αλγορίθμων, στην υποενότητα των τεχνητών νευρωνικών δικτύων. Οι κόμβοι όταν είναι αριθμητικοί είναι σιγμοειδής.

Περιγραφή βασικών παραμέτρων:

- Ποσοστό εκμάθησης για τον backpropagation αλγόριθμο: 0.3
- Momentum Rate για τον backpropagation αλγόριθμο: 0.2
- Πλήθος εποχών για την εκπαίδευση: 500

- Ποσοστό μεγέθους validation σετ ως κριτήριο τερματισμού της εκπαίδευσης:0
- Τιμή που χρησιμοποιείται για τη γεννήτρια τυχαίων αριθμών (seed): 0
- Πλήθος αλληλουχίας λαθών που επιτρέπεται από έλεγχο επικύρωσης μέχρι να τερματιστεί το δίκτυο:20
- Τα κρυφά επίπεδα που θα δημιουργηθούν για το δίκτυο: (μεταβλητές + κλάσεις) / 2,

BayesNet

Το περιβάλλον υλοποίησης παράσχει πολλαπλές δυνατότητες επιλογής εκτιμητών και αλγορίθμων αναζήτησης, όπου ως προκαθορισμένες έχει:
- Εκτιμητής: SimpleEstimators
- Αλγόριθμος αναζήτησης:K2

AdaBoostM1

Χρησιμοποιεί τον κλασικό αλγόριθμο AdaBoost, όπου:

- Ταξινομητής: Decision Stump
- Πλήθος επαναλήψεων:10
- Seed:1
- Όριο βάρους για κλαδεμα:100

Bagging

Κλασική υλοποίηση με παραμέτρους:
- Ταξινομητής: RepTree
- Πλήθος επαναλήψεων:10
- Seed:1
- Ποσοστό κάθε bag, ως προς το σύνολο εκπαίδευσης: 100

Random Forest
Κλασσική υλοποίηση του, όπου προκαθορισμένοι παράμετροι:

- Μέγιστο βάθος των δέντρων: Χωρίς περιορισμό
- Πλήθος μεταβλητών που θα χρησιμοποιηθούν στην τυχαία επιλογή: Χωρίς περιορισμό
- Πλήθος δέντρων που θα παραχθούν: 100
- Seed:1

SimpleCart

- Seed: 1
- Ελάχιστο πλήθος παραδειγμάτων στους τερματικούς κόμβους:2
- Πλήθος folds για το κλάδεμα: 5
- Ποσοστό μεγέθους του συνόλου εκπαίδευσης:100

Οι αλγόριθμοι που υλοποιήθηκαν σε περιβάλλον Matlab ήταν ο PSVM και ο DEA. Κανένας από τους δύο δεν ήταν ενσωματωμένος στη βιβλιοθήκη της Matlab.

PSVM

Ο PSVM αποτελεί μια υλοποίηση του Proximal SVM από τους G. Fung and O. L. Mangasarian, όπως περιγράφεται από το άρθρο 'Proximal Support Vector Machine Classifiers". Χρησιμοποιείται η γραμμική του υλοποίηση. Σαν έξοδο έδινε το AUC. Μετέπειτα υπολογίστηκαν precision, recall & Top decile Lift.

DIFFERENTIAL EVOLUTION ALGORITHM (DEA)

Υλοποιείται όπως περιγράφεται από την αντίστοιχη ενότητα. Ως εισόδους έχει:

- Σύνολο εκπαίδευσης και ελέγχου
- Σταθερά μετάλλαξης
- Ποσοστό διασταύρωσης
- Πλήθος γενεών
- Πληθυσμός κάθε γενεάς (με είσοδο την παράμετρο κ, όπου Πληθυσμός=κ *#μεταβλητών)
- Σταθερά trade-off , όπου τιμές 1 βαρύτητα στον υπολογιστικό φόρτο, 10 βαρύτητα εξίσου σε υπολογιστικό φόρτο και ακρίβεια, 100 βαρύτητα σε ακρίβεια
- Επιλογή δείκτη προς βελτιστοποίηση, με τιμή J=1 Gini, J=2 Lift, J=3 (Gini + Lift)/2

Σαν έξοδο, ως προς τους δείκτες, δίνει Gini coefficient και Top decile Lift για σύνολο εκπαίδευσης και ελέγχου.

4.3 ΔΕΙΚΤΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Προκειμένου να αξιολογηθούν τα μοντέλα πρόβλεψης αποχώρησης πελατών, που αναπτύσσονται από τους αντίστοιχους αλγόριθμους μηχανικής εκμάθησης, επιλέχθηκαν τέσσερις δείκτες. Για όλους το κύριο αποτέλεσμα των μοντέλων που χρησιμοποιείται ως είσοδος, είναι κατά πόσο έγινε σωστά η ταξινόμηση (classification).των πελατών σε αποχωρήσαντες και μη. Οι ομάδες ταξινόμησης είναι 0 για όσους παρέμειναν ή εκτιμήθηκε προβλεπτικά ότι παρέμειναν στην επιχείρηση και αντίστοιχα 1 για τους αποχωρήσαντες. Με σύγκριση των αντίστοιχων προβλέψεων και πραγματικών δεδομένων έπειτα από συγκεκριμένη επεξεργασία προκύπτουν οι παρακάτω δείκτες.

Οι δύο πρώτοι δείκτες, AUC και Top decile lift είχαν χρησιμοποιηθεί στον αντίστοιχο διαγωνισμό από όπου προήλθαν τα δεδομένα και επιλέγονται κατά κόρον στην ανάπτυξη μοντέλων πρόβλεψης αποχώρησης σε εταιρείες τηλεπικοινωνιών και σε εφαρμογές big data analytics. Οι δείκτες recall και precision επιλέγονται, για την αξιολόγηση της ακρίβειας και της ανάκλησης των εκάστοτε μοντέλων σε γενικές εφαρμογές αλγορίθμων μηχανικής εκμάθησης.

Προτού γίνει αναλυτική αναφορά σε αυτούς, σκόπιμο είναι να παρουσιαστούν κάποιες ομάδες πληροφοριών που χρησιμοποιούνται για τον υπολογισμό τους. Δοθέντων των διανυσμάτων που περιέχουν την πραγματική καθώς και την προβλεπόμενη ταξινόμηση στην αντίστοιχη ομάδα (class), αποχωρήσαντας (τιμή 1) ή μη αποχωρήσαντας (τιμή 0) του πελάτη εξάγεται η παρακάτω πληροφορία. Με βάση τη προβλεπτική ικανότητα του μοντέλου ομαδοποιούνται τα αποτελέσματα (πραγματικά σε σχέση με προβλεπόμενα) σε τέσσερις κατηγορίες. Για αυτούς που κατετάγησαν στις σωστές ομάδες, true positives (TP), κατετάγησαν στο class 1 και όντως ανήκαν στο 1 και true negatives (TN), κατετάγησαν στο class 0 και όντως ανήκαν στο 0. Όσον αφορά αυτούς που ταξινομήθηκαν εσφαλμένα, οι false positives, κατετάγησαν στο class 1 και ανήκαν στο 0 και false negatives, κατετάγησαν στο class 0 και ανήκαν στο 1.

Πίνακας 16: Confusion matrix

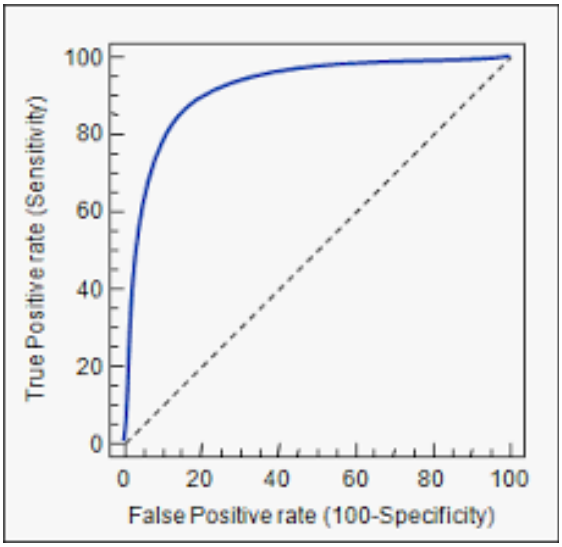
	Πρόβλεψη negatives (class 0)	Πρόβλεψη positives (class 1)
Ανήκουν σε class 0 (Negatives)	TN:# πελατών	FP:# πελατών
Ανήκουν σε class 1 (Positives)	FN:# πελατών	TP:# πελατών

Παρακάτω αναφέρεται αναλυτικά το μαθηματικό μοντέλο τους, η πληροφορία που χρησιμοποιούν ως είσοδο και η πρακτική αξία και σημασία τους για την επιχείρηση και τις αντίστοιχες λήψεις αποφάσεων. Επίσης, γί συνοπτική αναφορά στον υπολογιστικό φόρτο και το ρόλο του στη τελική επιλογή χρήσης αλγορίθμου σε συνδυασμό με τους παραπάνω δείκτες.

4.3.1 AREA UNDER CURVE (AUC)

Έχοντας τα διανύσματα από τις προβλεπόμενες και πραγματικές κλάσεις των πελατών καθώς και το διάνυσμα με τα αντίστοιχα αποτελέσματα (πιθανότητες) από το μοντέλο, υπολογίζεται αρχικά η καμπύλη λειτουργικών χαρακτηριστικών (ROC). Η καμπύλη προκύπτει υπολογίζοντας για διάφορα cutpoints τις αντίστοιχες τιμές των True Positives rates και False Positives rates. Τα παραπάνω

υπολογίζονται με τη χρήση των TP και NP, για το συγκεκριμένο διάστημα τιμών, προς τους πραγματικούς επί του συνόλου Positives και Negatives αντίστοιχα. Κατόπιν με τη χρήση αυτών των σημείων σχεδιάζεται η καμπύλη ROC[92]. Παρακάτω παρουσιάζεται γραφικά η καμπύλη ROC.



Εικόνα 6: Γράφημα απεικόνισης καμπύλης ROC (www.medcalc.org/manual/roc-curves.php)

Υπολογίζοντας το χωρίο κάτω από την αντίστοιχη καμπύλη προκύπτει ο δείκτης AUC (Area Under Curve- μτφ. περιοχή κάτω από καμπύλη). Ο αριθμός που προκύπτει δηλώνει με τι πιθανότητα ένα τυχαίο ζεύγος πελατών που ανήκουν σε διαφορετική κλάση, θα καταταγούν στη σωστή κλάση. Πρακτικά για να έχει ουσιαστική προβλεπτική αξία ένα μοντέλο και ο αντίστοιχος αλγόριθμος που χρησιμοποιήθηκε για την ανάπτυξη του, ο δείκτης πρέπει να είναι μεγαλύτερος του 0.5 [93]. Σε διαφορετική περίπτωση δεν έχει μεγαλύτερη αξία ο αλγόριθμος από ένα τυχαίο πείραμα. Για την επιχείρηση αποτελεί ένα δείκτη ποιότητας διαχωρισμού των πελατών στις αντίστοιχες κλάσεις. Αξιολογείται με αυτό τον τρόπο εάν η στρατηγική διατήρησης των πελατών τους θα απευθύνεται σε πραγματικά εν δυνάμει αποχωρήσαντες, που θα οδηγήσει σε μείωση των ποσοστών αποχώρησης πελατών. Ταυτόχρονα αξιολογεί εάν απευθύνεται εσφαλμένα σε πελάτες που δεν θα αποχωρούσαν, που θα μεταφραζόταν σε ζημία για την επιχείρηση. Συνήθως, αντί του AUC χρησιμοποιείται ο Gini coefficient σε τέτοιες εφαρμογές. Προκύπτει ως εξής:

$$Gini\,coefficient = 2 * AUC - 1 \quad (17)$$

Επιλέχθηκε η χρήση του AUC για ευκολότερη ερμηνεία των αποτελεσμάτων.

4.3.2 TOP DECILE LIFT

Ο συγκεκριμένος δείκτης εξετάζει τη προβλεπτική ικανότητα του μοντέλου στο 10% των πιο πιθανών χρηστών προς αποχώρηση, σύμφωνα με το εκάστοτε μοντέλο. Αξιολογείται αν με τη χρήση του μοντέλου υπάρχει καλύτερη προβλεπτική ικανότητα εν συγυρίσει με αυτή χωρίς μοντέλου. Ο αριθμός που προκύπτει εκφράζει πόσες φορές καλύτερα προβλέπεται το 10% των πιο

πιθανών πελατών προς αποχώρηση, βάσει των προβλέψεων από το μοντέλο σε σχέση με την απώλεια μοντέλου. Έχει χρηστική αξία καθώς, η επιχείρηση είθισται να επικεντρώνεται στους πιο επίφοβους προς αποχώρηση πελάτες. Τα δεδομένα που απαιτούνται για τον υπολογισμό του είναι τα διανύσματα πραγματικής και προβλεπόμενης κατάταξης στις κλάσεις και το διάνυσμα πιθανοτήτων αποχώρησης. Η διαδικασία που ακολουθείται είναι η εξής:[94]

1. Κατατάσσονται τα δεδομένα με βάση τη πιθανότητα αποχώρησης των πελατών, κατά φθίνουσα σειρά.
2. Χωρίζονται σε 10 (deciles) τμήματα τα δεδομένα. Το πρώτο τμήμα ονομάζεται top decile.
3. Υπολογίζεται για top decile lift το cumulative response rate.

$$cumulative\,response\,rate_{topdecile} = \frac{πλήθος\ πελατών\ top\,decile\ που\ όντως\ αποχώρησαν}{πλήθος\ πελατών\ που\ ανήκει\ σε\ top\,decile} \quad (18)$$

4. Υπολογίζεται total response rate

$$total\,response\,rate = \frac{πλήθος\ συνόλου\ πελατών\ που\ όντως\ αποχώρησαν}{σύνολο\ πελατών} \quad (19)$$

Όπου προκύπτει cumulative lift για top decile (θα αναφέρεται ως top decile lift)

$$Top\,decile\,lift = \frac{Cumulative\,response\,rate\,για\,top\,decile}{Total\,response\,rate} \quad (20)$$

Οι τιμές που προκύπτουν πρέπει να είναι μεγαλύτερες της μονάδας για να έχει αξία το μοντέλο.

4.3.3 PRECISION

Ο συγκεκριμένος δείκτης εκφράζει την ακρίβεια του μοντέλου, δηλαδή τι ποσοστό από αυτούς που προέβλεψε ότι θα αποχωρήσουν όντως αποχώρησαν. Προκύπτει από τη σχέση

$$Precision = \frac{TP}{FP + TP} \quad (21)$$

Για την επιχείρηση έχει αξία καθώς μικρές τιμές δηλώνουν ότι μπορεί να επικεντρωθεί εσφαλμένα και σε πελάτες που δεν είχαν ποτέ την πρόθεση να αποχωρήσουν. Οπότε θα έχει κόστος στην επιχείρηση χωρίς πιθανό όφελος. Σε μια τέτοια περίπτωση αξίζει να διερευνηθεί το κόστος διατήρησης πελάτη συναρτήσει κόστους απώλειας για κάθε πελάτη, προκειμένου να βρεθεί το όριο του recall που προκαλεί ζημία στην επιχείρηση συναρτήσει μιας αντίστοιχης στρατηγικής διατήρησης πελάτη.

Επίσης, γίνεται αναφορά στο precision (όπως και στο recall) και ως προς του προβλεπόμενους ως παραμείναντες πελάτες στην επιχείρηση. Για τον υπολογισμό τους αντιμετωπίζονται οι positives ως negatives και αντιστρόφως με τη χρήση των μαθηματικών τύπων που δόθηκαν.

4.3.4 RECALL

Ως recall γίνεται αναφορά στην ανάκληση, δηλαδή στο ποσοστό των αποχωρησάντων που προέβλεψε το μοντέλο ότι θα αποχωρήσουν. Πρακτικά δηλώνει σε τι ποσοστό θα εντοπίσει τους πελάτες που είχαν πρόθεση πραγματικά να αποχωρήσουν, χωρίς να αποκλείει ότι θα επικεντρωθεί λανθασμένα και σε πελάτες που δεν θα αποχωρούσαν (αυτό μετράται στο precision). Υπολογίζεται ως εξής: $Recall = \frac{TP}{TP + FN}$ (22)

4.3.5 ΥΠΟΛΟΓΙΣΤΙΚΟΣ ΦΟΡΤΟΣ

Αξίζει να γίνει αναφορά και σε αυτήν την παράμετρο καθώς η αξία ενός μοντέλου, μαζί με τις τιμές των παραπάνω δεικτών συνεκτιμάται με τον αντίστοιχο υπολογιστικό φόρτο. Ανατρέχοντας στο θεωρητικό τμήμα της εργασίας, παρατηρήθηκε ότι αποτελεί σημαντική παράμερο αξιολόγησης των αλγορίθμων. Πρακτικά αν κάποια μεθοδολογία για μικρές αποκλίσεις στις τιμές των δεικτών έχει μεγάλες αποκλίσεις στον υπολογιστικό φόρτο, θα προτιμηθεί αυτή με το μικρότερο υπολογιστικό φόρτο. Αυτό συμβαίνει καθώς ο υπολογιστικός φόρτος μεταφράζεται ως κόστος υλικοτεχνικά και χρονικά για μια επιχείρηση.

Για το τρέξιμο των αλγορίθμων χρησιμοποιήθηκαν δύο υπολογιστές. Για το 50% των πρώτων δοκιμών του γενετικού αλγορίθμου χρησιμοποιήθηκε φορητός υπολογιστής με επεξεργαστή 2GB Ram και λειτουργικό σύστημα Microsoft Windows. Για όλες τις υπόλοιπες δοκιμές και αλγορίθμους, χρησιμοποιήθηκε φορητός υπολογιστής με 8 GB Ram και λειτουργικό σύστημα Linux.

4.4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Αφού έγινε επεξεργασία των δεδομένων, όπως αναφέρεται παραπάνω, υλοποιήθηκαν δοκιμές για τους αλγορίθμους που περιγράφηκαν. Αρχικά υλοποιήθηκαν δοκιμές χωρίς τη χρήση κάποιας μεθοδολογίας επιλογής μεταβλητών. Στη πορεία επαναλήφθηκαν οι δοκιμές με είσοδο τα σύνολα δεδομένων που προέκυψαν από τις μεθοδολογίες επιλογής δεδομένων. Στη περίπτωση του wrapper πραγματοποιήθηκαν δοκιμές μόνο για τους αλγορίθμους που δεν αύξαναν σημαντικά τον υπολογιστικό φόρτο. Κριτήριο υλοποίησης του συναρτήσει των αλγορίθμων, ήταν ο υπολογιστικός φόρτος υλοποίησης των αλγορίθμων στο πρώτο σύνολο δοκιμών. Χρησιμοποιήθηκε για τους αλγορίθμους που έδιναν αποτελέσματα το πολύ εντός δύο λεπτών, στην αρχική τους υλοποίηση. Σε διαφορετική περίπτωση η συγκεκριμένη μεθοδολογία καθυστερούσε κάποιες ώρες να δώσει αποτελέσματα. Ο PSVM υλοποιήθηκε σε διαφορετικό περιβάλλον από τους υπόλοιπους (Matlab) , οι οποίοι υλοποιήθηκαν σε Weka, οπότε και δοκιμάστηκε μόνο στο αρχικό σύνολο δεδομένων.

Για την περίπτωση του διαφορικού εξελεγκτικού αλγόριθμου γίνεται ξεχωριστή αναφορά, καθώς πραγματοποιήθηκε πληθώρα δοκιμών για την επιλογή των βέλτιστων παραμέτρων εισόδου με κριτήριο τους δείκτες top decile lift και auc. Λόγω περιορισμένων υπολογιστικών δυνατοτήτων κατέστη δυνατό να εξεταστεί με μικρότερο σετ δεδομένων από ότι οι υπόλοιποι αλγόριθμοι. Ο συγκεκριμένος αλγόριθμος, ως διαφορικός εξελεγκτικός-γενετικός, μπορεί να συγκριθεί με τους αλγόριθμους που χρησιμοποίησαν σύνολα δεδομένων που προέκυπταν από τη μεθοδολογία wrapper. Για το λόγο αυτό οι αλγόριθμοι BayesNet και AdaBoostM1 έτρεξαν και με το αντίστοιχο μέγεθος συνόλου δεδομένων με τον DEA, ως προς το πλήθος των αποφασίζοντων, με επιλογή μεταβλητών από τη μεθοδολογία wrapper. Τα παραπάνω παρουσιάζονται σε ξεχωριστή ενότητα. Σε αυτό το σημείο αξίζει να αναφερθεί ότι μια παράμετρος που ελαχιστοποιεί την αποτελεσματικότητα του αλγόριθμου είναι ο υπολογιστικός φόρτος.

Η παρουσίαση αποτελεσμάτων αποτελείται από τρία σκέλη. Το πρώτο αφορά στη παρουσίαση των τιμών των δεικτών αξιολόγησης, με αντίστοιχη αναφορά και στο υπολογιστικό κόστος όπου κρίνεται χρήσιμο. Στη πορεία γίνεται αναφορά στις μεταβλητές που χρησιμοποιήθηκαν ως είσοδος, όπου πραγματοποιήθηκε επιλογή. Γίνεται παρουσίαση τόσο του πλήθους όσο και του είδους.

4.4.1 ΤΙΜΕΣ ΔΕΙΚΤΩΝ ΑΞΙΟΛΟΓΗΣΗΣ ΑΛΓΟΡΙΘΜΩΝ

Οι δείκτες που χρησιμοποιούνται για την αξιολόγηση είναι Top decile lift, AUC, precision και recall. Για τις περιπτώσεις recall και precision ο δείκτης 0 αναφέρεται στην κλάση μη αποχωρησάντων και ο δείκτης 1 στην κλάση αποχωρησάντων. Η συγκεκριμένη ενότητα απαρτίζεται από τρεις κύριες υποενότητες.

Στην πρώτη γίνεται παρουσίαση αποτελεσμάτων συγκεντρωτικά για όλους τους δείκτες και τους αλγορίθμους, χωρίς να έχει προηγηθεί η χρήση κάποιας μεθοδολογίας διαλογής μεταβλητών. Η δεύτερη καταγράφει τα αποτελέσματα για τις περιπτώσεις που έγινε χρήση μεθοδολογίας επιλογής μεταβλητών. Στην τελευταία γίνεται αναφορά στα πειράματα του διαφορικού εξελεγκτικού αλγορίθμου και εμφανίζονται οι δέκα καλύτερες δοκιμές. Στη πορεία χρησιμοποιούνται για σύγκριση με τα αποτελέσματα αλγορίθμων από wrapper, όπου χρησιμοποιήθηκε το αντίστοιχο σύνολο δεδομένων.

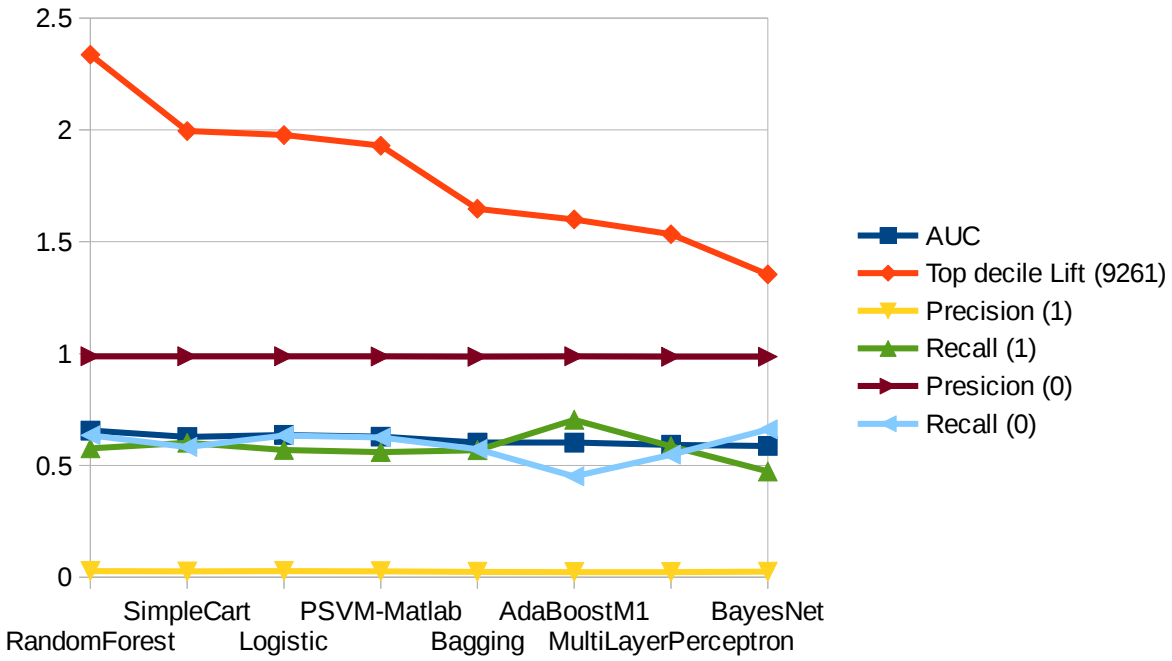
4.4.1.1 ΧΩΡΙΣ ΧΡΗΣΗ ΜΕΘΟΔΟΛΟΓΙΑΣ ΔΙΑΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Τα αποτελέσματα παρουσιάζονται συγκεντρωτικά για όλους τους αλγορίθμους (έκτος το DEA) και όλους τους δείκτες. Χρησιμοποιείται το σύνολο δεδομένων που χαρακτηρίζεται παραπάνω ως data set (1). Είναι το πλήρες σύνολο δεδομένων. Στη παρούσα ενότητα γίνεται χρήση όλων των μεταβλητών του αντίστοιχου σετ.

Πίνακας 17: Δείκτες αξιολόγησης αποτελεσμάτων για το πλήρες σύνολο δεδομένων (data set 1)

	AUC	Top decile Lift (9261)	Precision (1)	Recall (1)	Presicion (0)	Recall (0)
RandomForest	0.656	2.34	0.028	0.576	0.988	0.635
SimpleCart	0.627	2.00	0.026	0.601	0.988	0.582
Logistic	0.636	1.98	0.028	0.569	0.988	0.634
PSVM-Matlab	0.628	1.93	0.027	0.560	0.987	0.624
Bagging	0.602	1.65	0.024	0.568	0.986	0.573
AdaBoostM1	0.602	1.60	0.023	0.704	0.988	0.451
MultiLayerPerceptron	0.592	1.53	0.023	0.586	0.986	0.548
BayesNet	0.587	1.35	0.025	0.473	0.986	0.661

Γράφημα 1: Τιμές δεικτών αξιολόγησης αλγορίθμων για το πλήρες σετ δεδομένων



Παρατηρείται ότι τα καλύτερα αποτελέσματα καταφέρνει να τα δώσει ο RandomForest, χωρίς να είναι ικανοποιητικά, με προβλεπτική ικανότητα για τους 10% πιο πιθανούς προς αποχώρησης 2.34 φορές του μοντέλου πρόβλεψης έναντι της απουσίας μοντέλου. Επαρκής αλλά όχι καλές είναι οι τιμές που δίδονται για AUC, για τιμές από 0.6 και άνω, με πιθανότητα 60 % να κατατάσσεται ένα τυχαίο ζεύγος πελατών στη σωστή κλάση. Συνολικά από τη σκοπιά Top decile lift και AUC δίδονται με τη σειρά που καταγράφονται ,για Random Forest, Simple Cart, Logistic, PSVM επαρκή αποτελέσματα. Ανεπαρκή αποτελέσματα παρέχονται σε κάθε περίπτωση από τη σκοπιά του precision (1) με κάτω του 3% την καταχώρηση πραγματικών αποχωρησάντων, στους εν δυνάμει

αποχωρήσαντες. Ουσιαστικά κατατάσσει μεγάλο πλήθος πελατών που παραμένουν στην επιχείρηση ως εν δυνάμει αποχωρήσαντες. Αυτό είναι πιθανό να οφείλεται στη διαμόρφωση του συνόλου ελέγχου έναντι του συνόλου εκπαίδευσης, με αποχωρήσαντες κάτω του 2% στη μια περίπτωση και 50% στην άλλη για αποφυγή overfitting. Από τη σκοπιά ορθής πρόβλεψης αποχωρησάντων (recall1) αλλά και μη (recall0) , πέραν του AdaBoostM1 και BayesNet οι τιμές που προκύπτουν είναι επαρκής.

4.4.1.2 ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΟΛΟΓΙΩΝ ΔΙΑΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

ΩΣ ΠΡΟΣ ΤΙΣ ΜΕΘΟΔΟΛΟΓΙΕΣ

Για τη παρουσίαση των τιμών που προέκυψαν αφιερώνεται ένας πίνακας για κάθε δείκτη. Σε κάθε πίνακα καταγράφονται τα αποτελέσματα των αλγορίθμων συναρτήσει των αντίστοιχων μεθοδολογιών διαλογής μεταβλητών. Χρησιμοποιήθηκε το data set (1) με τις μεταβλητές που αναφέρονται στην ενότητα των δεδομένων/διαλογή μεταβλητών. Σε κάθε περίπτωση είναι μικρότερο το πλήθος μεταβλητών σε σύγκριση με το πλήρες του data set (1). Οι τιμές παρουσιάζονται ως αποκλίσεις σε σχέση με τη τιμή που έχει ο εκάστοτε αλγόριθμος χωρίς διαλογή δεδομένων, όπου

Τιμή πίνακα=Τιμή δείκτη (χρήση μεθοδολογίας διαλογής μεταβλητών)- Τιμή δείκτη(χωρίς χρήση μεθοδολογίας)

AUC

Πίνακας 18: Τιμές AUC αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

AUC	ChiSquared	Gainratio	Infogain	Wrapper
BayesNet	0.01	0.023	0.01	0.057
Logistic	-0.044	-0.037	-0.044	-
SimpleCart	-0.003	0.001	-0.003	-
MultiLayerPerceptron	0.001	-0.004	0.001	-
Bagging	-0.005	0.014	-0.005	-
AdaBoostM1	0	0	0	0.013
RandomForest	-0.031	-0.013	-0.031	-

Όπως φαίνεται στον πίνακα δεν παρατηρούνται αξιοσημείωτες αλλαγές ως προς AUC με τη εισαγωγή των μεθοδολογιών.

TOP DECILE LIFT

Table 19: Τιμές Top decile Lift αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

Lift (9261)	ChiSquared	Gainratio	Infogain	Wrapper
BayesNet	0.192	0.186	0.192	0.803
Logistic	-0.312	-0.174	-0.312	-
SimpleCart	-0.006	-0.156	-0.006	-
MultiLayerPerceptron	-0.120	-0.066	-0.120	-
Bagging	0.090	0.222	0.090	-
AdaBoostM1	0.000	0.000	0.000	0.000
RandomForest	-0.330	-0.144	-0.330	-

Παρατηρείται βελτίωση σημαντική στο BayesNet, ως προς τη προβλεπτική ικανότητα του μοντέλου για το 10% πιο επίφοβων προς αποχώρηση. Ακόμα και με παράμετρο τον υπολογιστικό φόρτο που προκύπτει από wrapper η βελτίωση είναι αξιοσημείωτη. Σε αυτό το σημείο αξίζει επίσης να παρατηρηθεί ότι chisquare και infogain παρουσιάζουν παρόμοια συμπεριφορά ως προς τα αποτελέσματα. Στο μεγαλύτερο πλήθος των αλγορίθμων δίνουν χειρότερα αποτελέσματα οι μεθοδολογίες. Η gainratio φαίνεται να δίνει ουσιαστική βελτίωση μόνο στη περίπτωση του Bagging.

PRECISION

Πίνακας 20: Τιμές Precision κλάσης 1 αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

Precision(1)	ChiSquared	GainRatio	InfoGain	Wrapper
BayesNet	0	0.001	0	0.002
Logistic	-0.004	-0.004	-0.004	-
SimpleCart	-0.001	0	-0.001	-
MultiLayerPerceptron	-0.002	-0.002	-0.002	-
Bagging	0	0.001	0	-
AdaBoostM1	0	0	0	0.001
RandomForest	-0.003	-0.001	-0.003	-

Για precision προσανατολισμένο στους αποχωρήσαντες, δεν παρουσιάζονται σημαντικές διαφοροποιήσεις με τη χρήση των μεθοδολογιών, για κανέναν από τους αλγορίθμους. Παραμένουν ανεπαρκείς οι τιμές. Το ίδιο παρατηρείται και για presicion ως προς τους παραμένοντες στην

επιχείρηση, όπου τα αποτελέσματα σε αυτή την περίπτωση χαρακτηρίζονταν ήδη ως “καλά”.

Πίνακας 21: Τιμές Precision κλάσης 0 αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

Precision (0)	ChiSquared	GainRatio	InfoGain	Wrapper
BayesNet	0	0	0	0.003
Logistic	-0.002	-0.002	-0.002	-
SimpleCart	-0.001	0	-0.001	-
MultiLayerPerceptron	0.003	0.002	0.003	-
Bagging	0	0.001	0	-
AdaBoostM1	0	0	0	0.001
RandomForest	-0.001	0	-0.001	-

RECALL

Για recall προσανατολισμένο στην κλάση 1 παρουσιάζονται σημαντικές βελτιώσεις για Bayes Net με wrapper και για MultiLayerperceptron συνολικά για όλες τις μεθοδολογίες (εκτός wrapper που δεν χρησιμοποιήθηκε).

Πίνακας 22: Τιμές Recall κλάσης 1 αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

Recall (1)	ChiSquared	GainRatio	InfoGain	Wrapper
BayesNet	0.035	0.024	0.035	0.175
Logistic	-0.048	-0.035	-0.048	-
SimpleCart	-0.003	0.003	-0.003	-
MultiLayerPerceptron	0.223	0.194	0.223	-
Bagging	0	0.013	0	-
AdaBoostM1	0	0	0	0
RandomForest	-0.007	0.025	-0.007	-

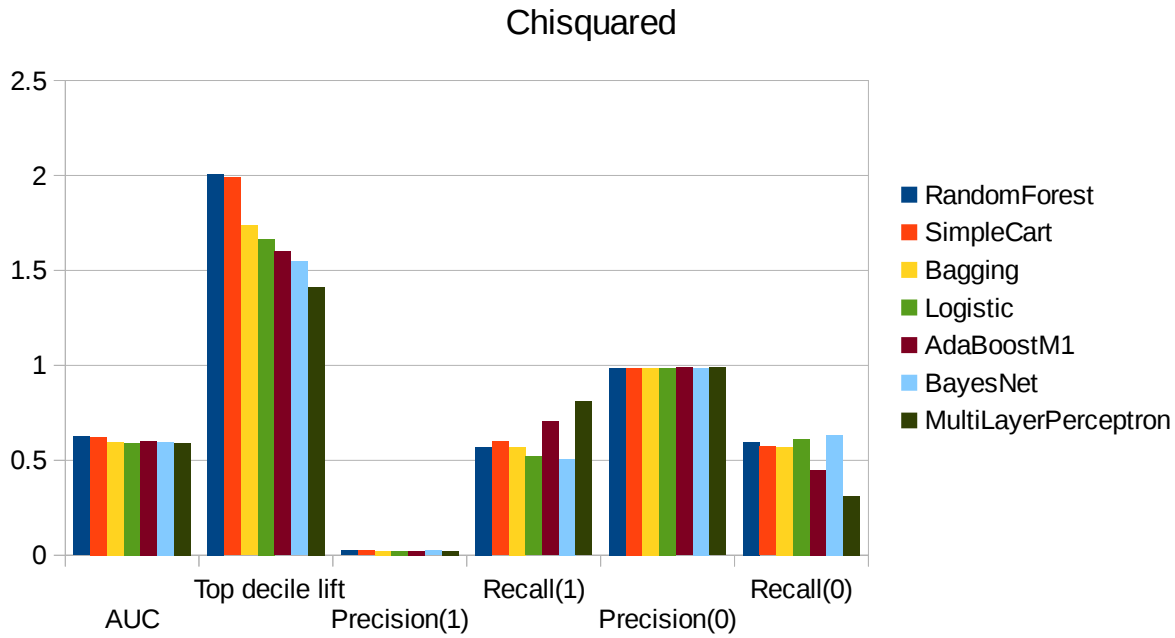
Ως προς την κλάση 0 φαίνεται να λειτουργεί ανταγωνιστικά ως προς την κλάση 1, καθώς για τους ίδιους αλγορίθμους και μεθοδολογίες παρουσιάζει επιδείνωση.

Πίνακας 23: Τιμές Recall κλάσης 0 αλγορίθμων για δοκιμές με πρόσθετη διαλογή μεταβλητών

Recall (0)	ChiSquared	GainRatio	InfoGain	Wrapper
BayesNet	-0.028	-0.004	-0.028	-0.091
Logistic	-0.023	-0.027	-0.023	-
SimpleCart	-0.01	-0.005	-0.01	-
MultiLayerPerceptron	-0.235	-0.209	-0.235	-
Bagging	-0.005	0.003	-0.005	-
AdaBoostM1	0	0	0	0.024
RandomForest	-0.038	-0.028	-0.038	-

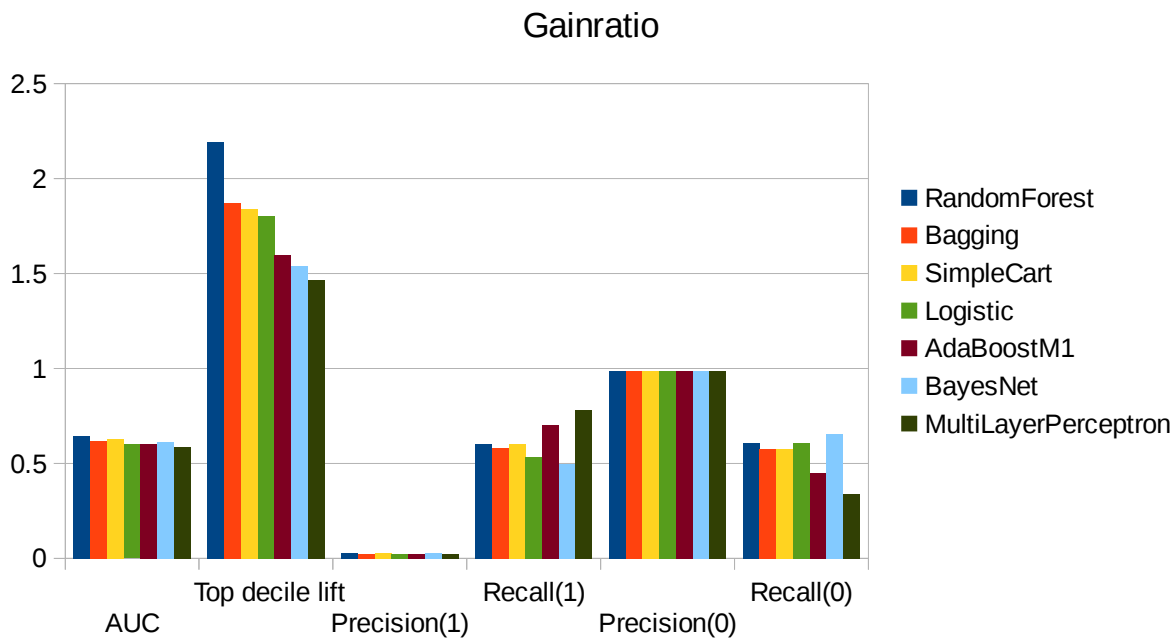
ΑΛΓΟΡΙΘΜΟΙ ΑΝΑ ΜΕΘΟΔΟΛΟΓΙΑ

Προκειμένου να υπάρξει συγκριτική αξιολόγηση των αλγορίθμων παρουσιάζονται γραφικά και συγκεντρωτικά τα αποτελέσματα των αλγορίθμων ανά μεθοδολογία/διαφορετικό σύνολο δεδομένων. Σε όλες τις περιπτώσεις παραμένει το πρόβλημα με το precision της κλάσης 1.



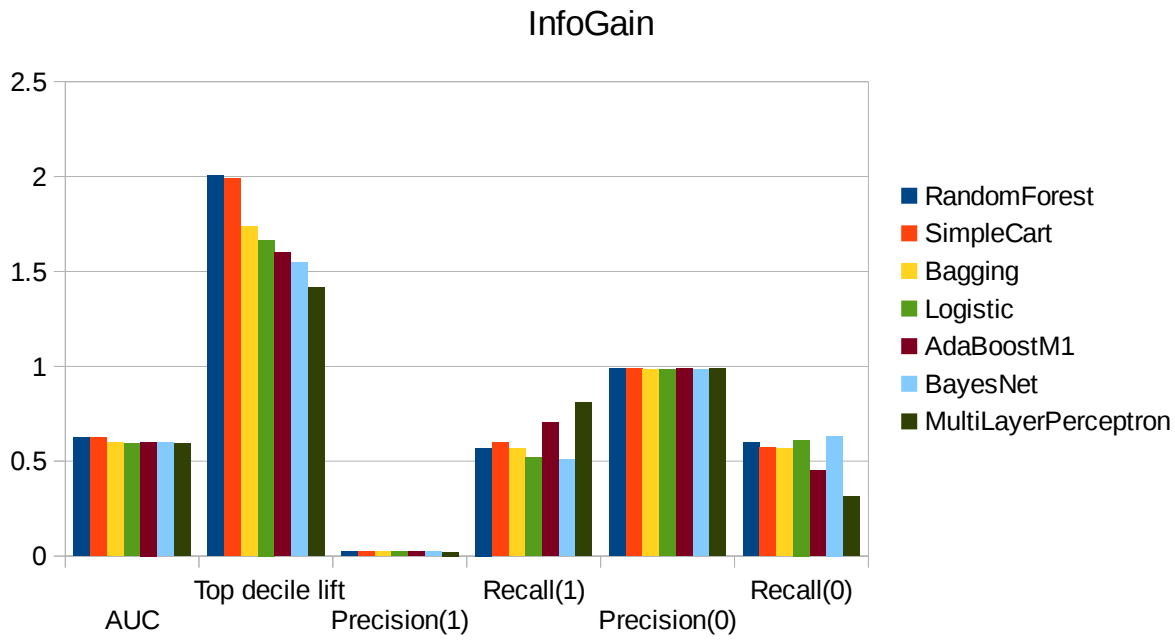
Γράφημα 2: Αποτελέσματα αλγορίθμων ανά δείκτη για μεθοδολογία Chisquared

Στην προκειμένη περίπτωση Random Forest και Simple Cart δίνουν τα καλύτερα αποτελέσματα από άποψη AUC και Top decile lift με ικανοποιητικό recall. Άρα κατατάσσουν καλύτερα ένα τυχαίο ζεύγος πελατών στις αντίστοιχες κλάσεις και περιγράφουν καλύτερα το 10% των εν δυνάμει αποχωρήσαντων. Από την άλλη AdaboostM1 και MultiLayerPerceptron δίνουν καλύτερα Recall(1), σωστή κατάταξη πραγματικών αποχωρησάντων, με προβληματικά όμως Recall(0) και όχι αρκετά καλή προβλεπτική ικανότητα στο 10%.



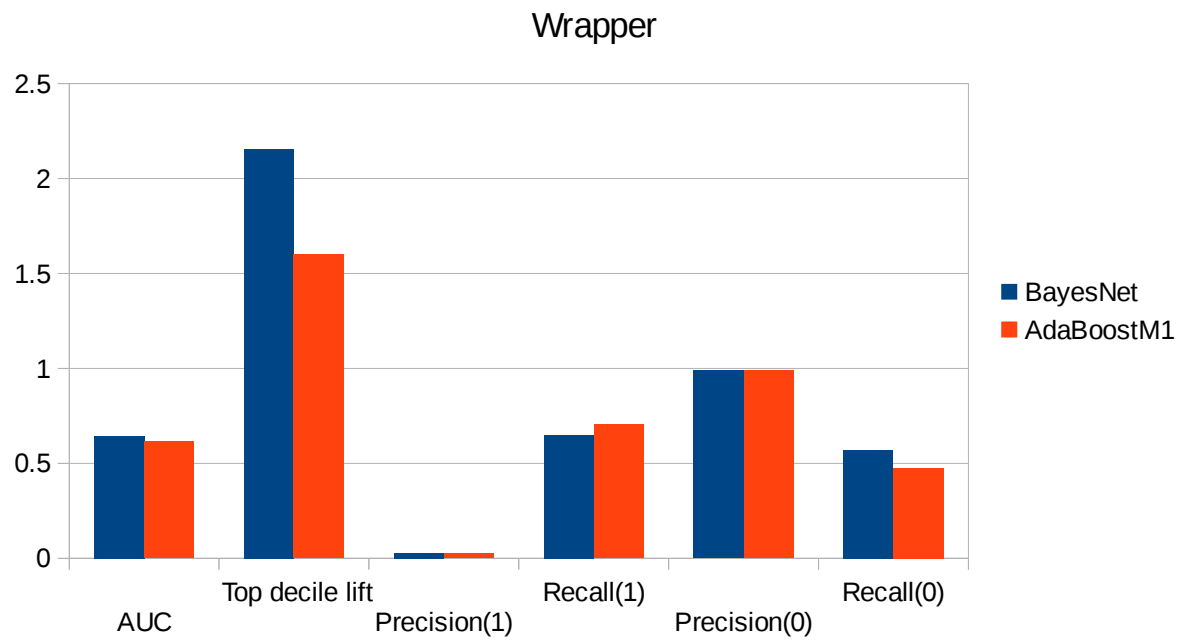
Γράφημα 3: Αποτελέσματα αλγορίθμων ανά δείκτη για μεθοδολογία Gainratio

Στην προκειμένη περίπτωση παρατηρείται το ίδιο μοτίβο με ξεκάθαρη υπεροχή ως προς Top decile lift του Random Forest. Έπειτα, το Bagging δουλεύει εξίσου καλά με το Simple CART. Επίσης, η λογιστική παλινδρόμηση δίνει επαρκή αποτελέσματα.



Γράφημα 4: Αποτελέσματα αλγορίθμων ανά δείκτη για μεθοδολογία Infogain

Ακριβώς ίδιο μοτίβο με chisquared. Όπως παρατηρείται αργότερα, chisquared και infogain παρέχουν ακριβώς το ίδιο σύνολο δεδομένων, με επιλογή των πρώτων 30 μεταβλητών.



Γράφημα 5: Αποτελέσματα αλγορίθμων ανά δείκτη για μεθοδολογία Wrapper

Για το wrapper φαίνεται να έχει ξεκάθαρη υπεροχή, λόγω top decile lift, το BayesNet, παρέχοντας για 1η φορά συνολικά ικανοποιητικά αποτελέσματα.

4.4.1.3 ΔΙΑΦΟΡΙΚΟΣ ΕΞΕΛΕΓΚΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΚΑΙ WRAPPER

Ο διαφορικός εξελεγκτικός αλγόριθμος (DEA) αποτελεί έναν αλγόριθμο που εσωτερικά υλοποιεί επιλογή μεταβλητών με βάση τον ίδιο τον αλγόριθμο. Οπότε συγκρίσιμα πειράματα είναι αυτά στα οποία έγινε χρήση του Wrapper.. Στα πειράματα που αναλύθηκαν προηγουμένως, οι αλγόριθμοι BayesNet και AdaBoostM1, χρησιμοποιήθηκαν από Wrapper και χρησιμοποίησαν τα αποτελέσματα από Wrapper για την ανάπτυξη μοντέλων.

Κατά την υλοποίηση των πειραμάτων με DEA, χρησιμοποιήθηκε μικρότερο σύνολο δεδομένων σε σχέση με τους υπόλοιπους αλγόριθμους. Ο υπολογιστικός φόρτος ήταν τέτοιος, που οι δυνατότητες του υπολογιστή δεν επέτρεπαν τη χρήση του data set 1, οπότε δημιουργήθηκε το data set 2, όπως έχει ήδη περιγραφεί. Ενδεικτικά, με το data set 2, μια υλοποίηση του μπορούσε να διαρκέσει από 20 λεπτά έως μιάμιση ώρα, ανάλογα την παραμετροποίηση του. Οι χρόνοι αυτοί αφορούν στον υπολογιστή με μνήμη RAM 8 GB.

Ο DEA προσφέρει αρκέτους βαθμούς ελευθερίας ως προς την παραμετροποίηση του σε σχέση με άλλους αλγορίθμους. Οπότε για την εύρεση των κατάλληλων τιμών ανάλογα το σύνολο δεδομένων απαιτείται να πραγματοποιηθεί μαι σειρά από πειράματα. Στα πειράματα αυτά, δοκιμάζονται διάφοροι συνδυασμοί των παραμέτρων εισόδου και αξιολογείται η αποτελεσματικότητα τους μέσω τον δεικτών Top Decile Lift και Gini coefficient. Με κριτήριο τις τιμές αυτών των δεικτών επιλέγονται τα 10 επικρατέστερα πειράματα. Έπειτα υλοποιείται από την αρχή Wrapper για AdaBoostM1 και BayesNet με χρήση του data set 2. Στη συνέχεια πραγματοποιούνται οι δοκιμές

σε AdaBoostM1 και BayesNet με το νέο σύνολο δεδομένων και μεταβλητών.

Παρακάτω παρουσιάζεται πίνακας με το σύνολο δοκιμών και τους αντίστοιχους συνδυασμούς παραμέτρων.

Πίνακας 24: Επιλογή τιμών παραμέτρων για το σύνολο πειραμάτων διαφορικού εξελικτικού αλγόριθμου

A/A	Trade-off	J	Generations	Mutation Constant	Crossover probability	k(Population size=k*123)	Runs
1	1	1	200	0.6:0.1:0.8	0.4:0.2:0.6	1	1
2	1,10,100	1,2,3	200	0.7	0.4	1	10
3	100	1	200,500	0.6:0.2:1.4	0.2:0.2:0.8	1	1
4	100	1	200	0.6:0.2:1.4	0.2:0.2:0.8	5	1

Αξίζει να γίνει αναφορά στη μεταβλητότητα των αποτελεσμάτων. Για το λόγο αυτό είχαν γίνει στο 2ο σετ δοκιμών 10 τρεξίματα σε κάθε συνδυασμό. Επίσης, το σύνολο ελέγχου έδινε καλύτερα αποτελέσματα ως προς τους δείκτες σε σχέση με το σύνολο εκπαίδευσης. Αυτό πιθανά να οφείλεται στο γεγονός ότι τα πραγματικά δεδομένα είχαν μόλις 2 % του συνόλου στη κλάση αποχωρούντων, όπως και το σύνολο ελέγχου, εν αντιθέσει με το σύνολο εκπαίδευσης που είχε διαμορφωθεί στο 50-50.

ΕΠΙΛΟΓΗ 10 ΚΑΛΥΤΕΡΩΝ ΠΕΙΡΑΜΑΤΩΝ

Επιλέχθηκαν οι 5 καλύτερες δοκιμές για Top Decile Lift και οι 5 καλύτερες για AUC (με κατάλληλη μετατροπή του Gini). Παρακάτω παρουσιάζεται ο πίνακας με τα 10 πειράματα. Παρατηρείται ότι επελέγη ένα παράδειγμα και ως ένα από τα 5 καλύτερα για AUC και για Top decile Lift.

Πίνακας 25: Τιμές παραμέτρων διαφορικού εξελικτικού αλγορίθμου για τα 5 καλύτερα πειράματα ως προς AUC και τα 5 καλύτερα ως προς Top decile Lift

A/A	Mutation Constant	Crossover probability	ɔulation size=k	Generations	Trade-off	J	
1	0.7	0.4	1	200	100	2	top5auc
2	0.7	0.4	1	200	10	2	top5auc
3	0.7	0.4	1	200	100	3	top5auc
4	0.8	0.2	1	500	100	1	top5auc
5	0.7	0.4	1	200	10	3	top5auc
6	1.4	0.4	1	500	100	1	top5lift
7	0.7	0.4	1	200	100	1	top5lift
8	1	0.4	5	200	100	1	top5lift
9	0.7	0.4	1	200	100	3	top5lift
10	1.4	0.2	1	500	100	1	top5lift

ΣΥΓΚΡΙΣΗ ΜΕ ΑΛΓΟΡΙΘΜΟΥΣ ΠΟΥ ΕΓΙΝΕ ΧΡΗΣΗ ΤΟΥ WRAPPER

Παρουσιάζεται ο πίνακας με τους δείκτες παραπομπής στον αντίστοιχο συνδυασμό DEA. Δεδομένου ότι το πείραμα με αριθμό 3 είναι το ίδιο με το 9, διατηρήθηκε μόνο το πείραμα με αριθμό 3.

Πίνακας 26: Σύγκριση τιμών δεικτών διαφορικού εξελεγκτικού αλγορίθμου με αλγορίθμους AdabbostM1 και BayesNet, όπου προέκυψε το σύνολο δεδομένων από wrapper

A/A	Lift future	AUC future	Precision1	Recall1	Precision0	Recall0
1	1.90	0.617	0.026	0.610	0.987	0.572
2	1.88	0.613	0.026	0.616	0.987	0.564
3	2.03	0.608	0.028	0.511	0.986	0.661
4	1.93	0.603	0.026	0.585	0.987	0.591
5	1.95	0.600	0.027	0.514	0.986	0.650
6	2.09	0.578	0.027	0.370	0.985	0.753
7	2.04	0.578	0.026	0.424	0.985	0.700
8	2.03	0.576	0.034	0.278	0.984	0.853
10	2.02	0.574	0.035	0.241	0.984	0.876
AdaBoostM1	1.67	0.600	0.024	0.601	0.986	0.544
BayesNet	1.82	0.615	0.026	0.538	0.986	0.623

Παρατηρείται ότι με όρους Lift και AUC είναι μόνο ο BayesNet συγκρίσιμος. Επίσης, ο συνδυασμός 1 είναι ο μόνος δίνει και καλύτερο Lift & καλύτερο AUC.

4.4.2 ΕΙΔΟΣ ΚΑΙ ΠΛΗΘΟΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Παρουσιάζεται το είδος των μεταβλητών για τις μεθοδολογίες διαλογής μεταβλητών & DEA. Για τις μεθοδολογίες που δεν συμμετείχε ο αλγόριθμος στην επιλογή των μεταβλητών γίνεται συγκριτική παρουσίαση των 30 πρώτων μεταβλητών. Δεδομένου ότι έχει προκαθοριστεί το πλήθος, δεν γίνεται αναφορά σε αυτό. Πάραυτα, αξίζει να παρουσιαστούν σε αυτό το σημείο τα κριτήρια επιλογής του ορίου των 30 μεταβλητών. Σε παλαιότερη υλοποίηση του DEA και προγενέστερη της υλοποίησης αλγορίθμων σε Weka, οι επικρατέστεροι συνδυασμοί, με ίσο trade-off μεταξύ ακρίβειας και πολυπλοκότητας, επέλεγαν τη χρήση περίπου 30 μεταβλητών. Επίσης, πείραμα με PSVM που είχε γίνει στα πλαίσια σύγκρισης με αυτούς τους συνδυασμούς, υπερείχε αυτών των συνδυασμών και χρησιμοποιούσε περίπου 30 μεταβλητές.

Για τις μεθοδολογίες που συμμετείχε ο αλγόριθμος στην επιλογή, γίνεται παρουσίαση των χαρακτηριστικών του συνδυασμού 1 από DEA. Δεδομένου ότι για τον συγκεκριμένο συνδυασμό είχαν γίνει 10 δοκιμές. υπολογίζεται το μέσο πλήθος χαρακτηριστικών. Επίσης, παρουσιάζονται και συγκρίνονται με Wrapper(BayesNe)t & Wrapper(AdaBoostM1) οι μεταβλητές που επιλέχθηκαν και στις 10 δοκιμές. Όσον αφορά τις 2 υλοποιήσεις Wrapper (data set 1 & data set 2) για κάθε αλγόριθμο, γίνεται ανά αλγόριθμο σύγκριση.

4.4.2.1 ΜΕΘΟΔΟΛΟΓΙΕΣ ΔΙΑΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΧΩΡΙΣ ΤΗ ΣΥΜΜΕΤΟΧΗ ΑΛΓΟΡΙΘΜΟΥ

Παρουσιάζονται οι πρώτες 30 μεταβλητές από Chisquared και Infogain και η αντίστοιχη ταξινόμηση των μεταβλητών αυτών που έχουν τόσο αυτές οι μεθοδολογίες όσο και η Gainratio.

Πίνακας 27: Πρώτες 30 μεταβλητές από μεθοδολογίες Chisquared και Infogain και αντίστοιχη ταξινόμηση ως προς όλες τις απλές μεθοδολογίες

	Chisquared	Gainratio	Infogain
month	1	5	1
eqpdays	2	4	2
hnd_price	3	9	3
totmrc_Mean	4	19	4
mou_Mean	5	14	5
change_mou	6	10	6
hnd_webcap	7	7	7
totmrc_Range	8	8	8
adjrev	9	18	10
avg3mou	10	15	9
avg3qty	11	12	11
totrev	12	16	12
crclscod	13		14
retdays	14	1	13
mou_opkv_Mean	15		15
mou_cvce_Mean	16	30	16
asl_flag	17	6	17
tot_acpt	18	3	18
tot_ret	19	2	19
ethnic	20		20
opk_vce_Mean	21	17	21
mou_peav_Mean	22	21	22
comp_vce_Mean	23		23
iwylis_vce_Mean	24	27	24
complete_Mean	25		25
mouiwylisv_Mean	26		26
peak_vce_Mean	27	28	27
attempt_Mean	28		28
plcd_vce_Mean	29		30
avg6mou	30	25	29

Παρατηρείται ότι Chisuared και Infogain έχουν τις ίδιες μεταβλητές στις πρώτες 30 θέσεις με μικρές αποκλίσεις στην ταξινόμηση. Όσον αφορά τη GainRatio χρησιμοποιεί 22 από τις 30 που χρησιμοποιούν οι άλλες δύο, στις πρώτες 30 θέσεις με σημαντικές διαφορές όμως στην ταξινόμηση τους. Οι 8 μεταβλητές που δεν είναι κοινές υπογραμμίζονται στον Πίνακα 27. Επίσης, 6 κοινές μεταβλητές κατατάσσονται και στις τρεις μεθοδολογίες στην 1η δεκάδα.

Πίνακας 28: Κοινές μεταβλητές στην πρώτη δεκάδα των απλών μεθοδολογιών διαλογής μεταβλητών

Κωδική ονομασία	**Περιγραφή**
month	Μήνες στην υπηρεσία
eqdays	Ηλικία παρεχόμενου εξοπλισμού
hnd_price	Τιμή συσκευής
change_mou	Ποσοστιαία μηνιαία μεταβολή χρήσης λεπτών ως προς τους τελευταίους 6 μήνες
hnd_webcap	Δυνατότητες συσκευής πρόσβασης στο internet
totmrc_range	Εύρος τιμής μηναίου παγίου

4.4.2.2 ΔΙΑΦΟΡΙΚΟΣ ΕΞΕΛΙΓΚΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ &WRAPPER

Ξεκινώντας από DEA, το μέσο πλήθος μεταβλητών είναι 58 με τιμή απόκλισης 4. Μόλις 6 από αυτές επιλέγονται και στα 10 πειράματα. Αυτές χρησιμοποιούνται για σύγκριση με Wrapper. Πάραυτα, γίνεται αναφορά της συχνότητας εμφάνισης σε DEA των μεταβλητών που επιλέγονται μέσω Wrapper.

Πίνακας 29: Σύγκριση επικρατέστερων μεταβλητών διαφορικού εξελικτικού αλγορίθμου με αντίστοιχα συγκρίσιμα πειράματα των αλγορίθμων AdaBoostM1 και BayesNet. Αναφορά επιλογής χρήσης από διαφορικό εξελικτικό αλγόριθμο, στο σύνολο των 10 πειραμάτων, των επικρατέστερων μεταβλητών των άλλων 2 αλγορίθμων

DEA	adaboostM1/DEA	ΧΡΗΣΗ ΣΕ DEA	BayesNet/DEA	ΧΡΗΣΗ ΣΕ DE
change_mou	change_mou		ovrrev_Mean	70.00%
uniqsubs	plcd_vce_Mean	60.00%	vceovr_Mean	20.00%
asl_flag	custcare_Mean	60.00%	totmrc_Range	90.00%
refurb_new	months	80.00%	change_mou	
age1	hnd_price	90.00%	mouiwylisv_Mean	70.00%
eqpdays			months	80.00%
			crclscod	40.00%
			totcalls	40.00%
			avg3qty	60.00%
			tot_ret	80.00%
			tot_acpt	80.00%
			hnd_price	90.00%
			hnd_webcap	40.00%
			marital	10.00%
			infobase	20.00%
			ethnic	60.00%
			kid0_2	30.00%
			retdays	60.00%
			eqpdays	

Παρατηρείται ότι, παρότι μόλις 2 μεταβλητές που επιλέγονται πάντα στο DEA επιλέγονται από wrapper. Πάραυτα οι μεταβλητές που χρησιμοποιούνται σε wrapper εμφανίζουν υψηλά ποσοστά χρήσης σε DEA. Κοινή μεταβλητή για όλες η change_mou. Γενικά παρατηρείται ότι επιλέγονται από wrapper, με υψηλά ποσοστά χρήσης από DEA, τέσσερις από τις “καλύτερες” έξι των

ChiSquared, Infogain και Gainratio.

Πίνακας 30: Επικρατέστερες μεταβλητές για το σύνολο των δοκιμών των αλγορίθμων

Κωδική ονομασία	**Περιγραφή**
month	Μήνες στην υπηρεσία
eqdays	Ηλικία παρεχόμενου εξοπλισμού
hnd_price	Τιμή συσκευής
change_mou	Ποσοστιαία μηνιαία μεταβολή χρήσης λεπτών ως προς τους τελευταίους 6 μήνες

Εν συνεχεία, πραγματοποιείται σύγκριση μεταξύ των δύο υλοποιήσεων Wrapper ανά αλγόριθμο.

Πίνακας 31: Μεταβλητές από Wrapper για BayesNet για τα δύο σύνολα δεδομένων

Bayes_Net	BayesNet/DEA
ovrmou_Mean	ovrrev_Mean
totmrc_Range	vceovr_Mean
change_mou	totmrc_Range
months	change_mou
crclscod	mouiwylisv_Mear
totcalls	months
tot_ret	crclscod
area	totcalls
refurb_new	avg3qty
hnd_webcap	tot_ret
age1	tot_acpt
ethnic	hnd_price
eqpdays	hnd_webcap
	marital
	infobase
	ethnic
	kid0_2
	retdays
	eqpdays

Πίνακας 32: Μεταβλητές από Wrapper για AdaBoostM1 για τα δύο σύνολα δεδομένων

AdaboostM1	AdaboostM1/DEA
totmrc_Mean	change_mou
totmrc_Range	plcd_vce_Mean
crclscod	custcare_Mean
tot_ret	months
dualband	hnd_price
age1	
eqpdays	

Εδώ μπορεί να αξιολογηθεί η επιρροή του συνόλου δεδομένων στα αποτελέσματα. Σε AdaBoost δεν παρατηρούνται σημαντικές διαφορές στα αποτελέσματα παρότι χρησιμοποιεί τελείως διαφορετικό σετ μεταβλητών και μικρότερο όγκο δεδομένων. Εν αντιθέσει σε BayesNet με Wrapper, όπου με data set 1 παρατηρήθηκε ότι απέδιδε καλύτερα από τα υπόλοιπα δεδομένων, με μικρή διαφοροποίηση των μεταβλητών, του πλήθους τους και σημαντική διαφοροποίηση του όγκου δεδομένων παρατηρήθηκε μείωση της απόδοσης.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 ΣΥΜΠΕΡΑΣΜΑΤΑ ΓΙΑ ΑΛΓΟΡΙΘΜΟΥΣ

Αρχικά, δεδομένου ότι ο προσανατολισμός μας είναι η υποστήριξη στρατηγικών αποφάσεων μέσω των αλγορίθμων, τα συμπεράσματα θα εξαχθούν με βαρύτητα στους δείκτες Gini και Top decile lift. Οπότε, ο αλγόριθμος που εμφανίζει ξεκάθαρη υπεροχή, σε όλα τα σετ δεδομένων, είναι ο Random Forest, ο οποίος δουλεύει καλύτερα με το πλήρες σετ μεταβλητών. Αξίζει να σημειωθεί ότι είναι ο μόνος αλγόριθμος που κατάφερε να δώσει αποτελέσματα λίγο πιο πάνω από το μέσο όρο των αποτελεσμάτων του διαγωνισμού. Πάραυτα δεν καταφέρνει να προσεγγίσει τις καλύτερες μεθοδολογίες. Παρακάτω εμφανίζονται οι αλγόριθμοι που έδωσαν ικανοποιητικά αποτελέσματα ιεραρχικά.

Πίνακας 33: Καλύτεροι αλγόριθμοι με ικανοποιητικά αποτελέσματα ως προς AUC και Top Decile Lift ανά μεθοδολογία διαλογής μεταβλητών

ΑΛΓΟΡΙΘΜΟΣ	ΜΕΘΟΔΟΛΟΓΙΑ
Random Forest	INFOGAIN/CHISQUARED
Simple CART	
Bagging (Μέτρια αποτελέσματα)	
Random Forest	GAINRATIO
Bagging	
Simple CART	
Logistic	
Random Forest	ΑΠΛΗ
Simple CART	
Logistic	
PSVM	

Σε όλες τις περιπτώσεις ο Simple Cart δίνει ικανοποιητικά αποτελέσματα. Ο Bagging αποδίδει καλύτερα με σύνολα δεδομένων, όπου έχουν αφαιρεθεί οι μεταβλητές με μικρότερη επιρροή στην ταξινόμηση. Η λογιστική παλινδρόμηση στην περίπτωση του infogain δεν αποδίδει καλά, συνολικά αποδίδει καλύτερα με το πλήρες σύνολο. Ο PSVM αποδίδει ικανοποιητικά, αλλά δεν ήταν διαθέσιμα αποτελέσματα από τις άλλες μεθοδολογίες για περαιτέρω σύγκριση. Οι μεθοδολογίες chissquared και infogain παράγουν το ίδιο σύνολο εκπαίδευσης, με επιλογή των 30 επικρατέστερων μεταβλητών.

Σε μια προσπάθεια γενίκευσης, μπορεί να ειπωθεί ότι αλγόριθμοι όπως RandomForest και ο SimpleCART που χρησιμοποιούν δέντρα απόφασης αποδίδουν καλύτερα. Επίσης ο AdaBoost σε σχέση με τους υπόλοιπους συνδυαστικούς αλγόριθμους, RandomForest και Bagging, δεν απέδωσε καλά. Πάραυτα, παρατηρήθηκε ότι με μεταβολή πλήθους δεδομένων ή μεταβλητών δεν μεταβάλλονται τα αποτελέσματα των δεικτών. Άρα για τέτοιες εφαρμογές θα μπορούσε να

χαρακτηριστεί ευσταθής.

Πίνακας 34: Περιγραφή αξιοσημείωτων μεταβολών στις τιμές των δεικτών με τη χρήση μεθοδολογιών διαλογής μεταβλητών

ΑΛΓΟΡΙΘΜΟΣ	ΜΕΤΑΒΟΛΗ	ΔΕΙΚΤΗΣ	ΜΕΘΟΔΟΛΟΓΙΑ
BayesNet	Βελτίωση	Top decile lift	Wrapper
Logistic	Χειροτέρευση	Top decile lift	Infogain/Chisquared
MultiLayerpercetron	Βελτίωση	Recall_1	Infogain/Chisquared
RandomForest	Χειροτέρευση	Top decile lift	Infogain/Chisquared

Γενικά με τη μεταβολή του συνόλου μεταβλητών, εμφανίστηκαν αξιοσημείωτες επιδράσεις σε συγκεκριμένες περιπτώσεις. Μια από τις περιπτώσεις που αξίζει να αναφέρουμε είναι του BayesNet, όπου παρουσιάστηκε σημαντική βελτίωση με τη χρήση Wrapper καθιστώντας τον αλγόριθμο ανταγωνιστικό σε σχέση με τους πρώτους δύο ως προς την αποτελεσματικότητα. Περαιτέρω, παρατηρήθηκε η μείωση της αποτελεσματικότητας του με τη χρήση του συνόλου εκπαίδευσης του DEA, οδηγώντας στην υπόθεση ότι ο συγκεκριμένος συνδυασμός παραδειγμάτων μπορεί να έχει επίδραση και στην αποτελεσματικότητα του DEA. Επίσης στην περίπτωση του AdaBoostm1(wrapper) οδήγησε σε διαφορετική επιλογή μεταβλητών σε σχέση με την περίπτωση του Dataset1 Αυτή η συσχέτιση γίνεται καθώς ο DEA υπερτερούσε του BayesNet(wrapper) με το dataset2 και στηνπερίπτωση του dataset1 BayesNet(wrapper) προσφέρει παρόμοια αποτελέσματα με τον Random Forest (χωρίς επιλογή μεταβλητών).

Ως προς τον DEA, συναρτήσει των αποτελεσμάτων του και του υπολογιστικού κόστους του, κρίνεται το αντίθετο του cost-effective. Αξίζει όμως να γίνει μια παρατήρηση ως προς τα αποτελέσματα του, που μπορεί να οδηγήσει σε συμπεράσματα σε σχέση με αδυναμίες στην προετοιμασία των δεδομένων. Όπως έχει ήδη αναφερθεί, ένα πρόβλημα ήταν ότι τα αποτελέσματα των δεικτών για τα σύνολο εκπαίδευσης ήταν χειρότερα από ότι για τα σύνολα ελέγχου. Το σύνολο εκπαίδευσης είχε τροποποιηθεί ώστε να περιέχει 50-50 παραδείγματα ως προς τις κλάσεις για αποφυγή overfitting. Όμως, τα πραγματικά δεδομένα περιγράφονται από 2% αποχωρήσαντες. Αυτό λοιπόν, πιθανολογείται ότι μπορεί να οδήγησε σε underfitting ως προς την κλάση 1. Με αυτό τον τρόπο μπορεί να ερμηνευτεί το προαναφερθέν πρόβλημα αλλά και οι προβληματικές τιμές ως προς το Presicion της κλάσης 1.

5.2 ΕΠΙΚΡΑΤΕΣΤΕΡΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Με την διαδικασία που περιγράφηκε στην ενότητα των χαρακτηριστικών, καταλήξαμε ότι τα επικρατέστερα χαρακτηριστικά είναι τα παρακάτω. Δεν αναλύεται βέβαια αν η συσχέτιση είναι θετική ή αρνητική. Σε μια προσπάθεια να παράσχουμε λογική ερμηνεία θα πραγματοποιηθεί μια σύντομη ανάλυση.

Πίνακας 35: Επικρατέστερες μεταβλητές για το σύνολο των δοκιμών των αλγορίθμων

Κωδική ονομασία	**Περιγραφή**
month	Μήνες στην υπηρεσία
eqdays	Ηλικία παρεχόμενου εξοπλισμού
hnd_price	Τιμή συσκευής
change_mou	Ποσοστιαία μηνιαία μεταβολή χρήσης λεπτών ως προς τους τελευταίους 6 μήνες

Βιβλιογραφικά έχει υποστηριχθεί ότι υπάρχουν χαμένοι πελάτες λόγω της μη ικανοποιητικά παρεχόμενης τεχνολογίας. Οπότε όσο μεγαλύτερη η ηλικία του εκάστοτε εξοπλισμού τόσο ξεπερασμένος θεωρείται. Αυτό μπορεί να οδηγήσει σε αποχώρηση, στα πλαίσια της μη ικανοποιητικά παρεχόμενης τεχνολογίας. Επίσης, η ωρίμανση του πελάτη αποτελεί έναν παράγοντα αποχώρησης. Από την άλλη αυξημένη παραμονή μπορεί να ερμηνεύεται και ως αφοσίωση. Η ποσοστιαία μεταβολής χρήσης, μπορεί να αποτελέσει ένα δείκτη πρόβλεψης, καθώς μπορεί να περιγράψει πότε ένας πελάτης οδηγείται σε αδράνεια ως προς τη χρήση των παρεχόμενων υπηρεσιών. Τέλος, η τιμή της συσκευής μπορεί να δημιουργήσει εξαρτημένη δέσμευση με την εταιρεία αν είναι υψηλή. Συνολικά, την κατεύθυνση ερμηνείας τους φαίνεται να τη δίνει η μεταβλητή ποσοστιαίας μεταβολής χρήσης λεπτών αν και απαιτούνται περαιτέρω πληροφορίες και ίσως μεταβλητές για την ορθή ερμηνεία τους.

5.3 ΣΥΝΟΨΗ

Έπειτα από μελέτη 9 αλγορίθμων μηχανικής εκμάθησης που προέρχονται από τα πεδία:

- Τεχνητών Νευρωνικών Δικτύων
- Λογιστικής Παλινδρόμησης
- NaiveBayes & BayesNet
- Δέντρων απόφασης
- Μηχανές διανυσμάτων υποστήριξης
- Εξελεγκτικών αλγορίθμων
- Μεταμαθησιακών μοντέλων

και τη χρήση διάφορων συνόλων δεδομένων για τη μελέτη τους, προέκυψε η ξεκάθαρη υπεροχή ενός αλγόριθμου, του RandomForest. Προέρχεται από την οικογένεια των μεταμαθησιακών μοντέλων και κάνει χρήση δέντρων απόφασης. Κύρια κριτήρια αξιολόγησης της αποτελεσματικότητας του οι δείκτες AUC και Top Decile Lift. Προσανατολισμός η ανάπτυξη μοντέλου πρόβλεψης τέτοιου ώστε να εξαχθεί γνώση και να υποστηριχθούν αποτελεσματικά στρατηγικές αποφάσεις για την πρόληψη αποχώρησης πελατών.

Τα αποτελέσματα που παρείχε, παρότι τα καλύτερα, δεν ήταν επαρκώς ικανοποιητικά Πάραυτα, προέκυψε πληροφορία προς επεξεργασία από την συγκριτική μελέτη των αλγορίθμων. Επίσης ανέκυψε ένα ακόμα ζητούμενο, οι ιδιαιτερότητες των δεδομένων και η επιρροή τους στην πιστότητα των αποτελεσμάτων. Τα παραπάνω ζητούμενα δημιουργούν ανάγκη για περαιτέρω μελέτη, για τέτοιου είδους δεδομένα και τα όρια απόδοσης χρήσιμης πληροφορίας και γνώσης για την επιχείρηση.

6 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Από την παρούσα εργασία ανέκυψαν περισσότεροι προβληματισμοί από ότι αποτελέσματα. Προβληματισμοί προς επίλυση που δημιουργούν προτάσεις για περαιτέρω έρευνα. Ξεκινώντας από τα δεδομένα, αξίζει να διερευνηθεί αν η προσπάθεια μέσω oversampling για αποφυγή overfitting, μπορεί να οδηγήσει σε underfitting για την ανταγωνιστική κλάση. Ιδίως σε δεδομένα που περιγράφουν μια πραγματική κατάσταση όπου δεν ισχύει ο κανόνας 50-50 για τις κλάσεις. Εφόσον επιβεβαιωθεί κάτι τέτοιο αξίζει να μελετηθεί πως μπορεί να επιλυθεί με σεβασμό στις εκάστοτε ιδιαιτερότητες κάθε κλάδου από όπου προέρχονται τα δεδομένα.

Ως προς τη μελέτη των αλγορίθμων συναρτήσει αυτών των δεδομένων, μπορεί να γίνει περαιτέρω μελέτη της αποτελεσματικότητας αλγορίθμων που χρησιμοποιούν κατά την υλοποίηση τους δέντρα αποφάσεων. Εναλλακτικά, μπορεί να επαναληφθεί η συγκριτική μελέτη για μεγαλύτερο δείγμα, δεδομένου του πλήθους των μεταβλητών και το μικρό ποσοστό που αντιστοιχεί στη μια κλάση. Σε αυτή την περίπτωση θα μπορούσε να μελετηθεί ένας εναλλακτικός εξελεγκτικός αλγόριθμος, με μικρότερο υπολογιστικό κόστος. Εάν είναι αποτελεσματικός μπορεί να εξάγει παραπάνω πληροφορία από άλλα μοντέλα και να προσφέρει ευελιξία την παραμετροποίηση.

Ένα ακόμα κομμάτι που θα άξιζε να μελετηθεί, είναι η επίδραση των χαρακτηριστικών και αν σε βάθος χρόνου μεταβάλλεται αυτή η συσχέτιση ως προς την αποχώρηση. Επίσης, θα μπορούσε να μελετηθεί, τελικά σε τι ποσοστό χρησιμοποιείται από την επιχείρηση η πληροφορία που εξάγεται. Κλείνοντας, με προσανατολισμό το κομμάτι του business analytics θα μπορούσε να αξιολογηθεί σε τι βαθμό αξιοποιείται ορθά η γνώση που προκύπτει από την ανάπτυξη των μοντέλων πρόβλεψης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1: , , , http://practicalanalytics.co/2011/04/24/gartner-says-bi-and-analytics-a-10-5-bln-market/
2: J.-H. Ahn et al. , Customer churn analysis: Churn determinants and mediationeffects of partial defection in the Korean mobiletelecommunications service industry, 2006,Telecommunications Policy,Elsevier
3: Clyde Holsapple , Anita Lee-Post , Ram Pakath, A unified foundation for business analytics, 2014,Decision Support Systems,Elsevier B.V.
4: Michael J. Mortenson, Neil F. Doherty, Stewart Robinson , Operational research from Taylorism to Terabytes: A research agendafor the analytics age, 2015,European Journal of Operational Research,Elsevier B.V.
5: Nadj, Mario; Morana, Stefan; Maedche, Alexander, Towards a situation-awareness-driven design of operational businessintelligence & analytics systems,
6: Chen et al., BUSINESS INTELLIGENCE AND ANALYTICS : FROM BIG DATA TO BIG I MPACT, 2012,MIS Quarterly,
7: Michael Wu, Smarter Data, ,
http://www.lithium.com/pdfs/documents/Lithium_AdMap_Smarter_Data.pdf
8: Bill Vorhies, Prescriptive versus Predictive Analytics – A Distinction without a Diffrence?, ,
http://data-magnum.com/prescriptive-versus-predictive-analytics-a-distinction-without-a-diffrence/
9: Joel Järvinen, Aarne Töllmen, Heiki Karjaluoto, Web Analytics and Social Media Monitoring in Industrial Marketing: Tools for Improving Marketing Communication Measurement, 2015
10: D.J. Clark, David Nicholas and Hamid R. Jamali, Evaluating information seeking and use in the changing virtual world: the emerging role of Google Analytics , 2014, Learned Publishing,
11: Tamer Mohamed Abdellatif, Luiz Fernando Capretz, and Danny Ho, Software Analytics to Software Practice: A Systematic Literature Review , 2015
12: Morten Søby, Learning Analytics, 2014,Nordic Journal of Digital Literacy,Universitetsforlaget
13: William J. Hauser, Marketing analytics: the evolution of marketing research in the twenty-firstcentury, 2007,Direct Marketing: An International Journal, Emerald Group Publishing Limited
14: Roman Lenzen, Customer Analytics: It's All About Behavior , 2004,DM Review,
15: W. Seiringer, J. Cardoso, and J.K. von Bischhoffshausen , Service System Analytics: Cost Prediction , 2013
16: Weena Yancey M Momin, Kushendra Mishra , HR Analytics as a Strategic Workforce Planning, 2015,International Journal of Applied Research,
17: Orijit Ghosh , Sannah Manuja , Ankita Sehrawat , Beas Banerjee , Talent Analytics from a Strategic Perspective, 2014,International Journal of Innovative Research & Development,
18: Michael zur Muhlen and Robert Shapiro, Business Process Analytics, 2010
19: Gilvan C. Souza, Supply chain analytics, 2014,Business Horizons,Elsevier
20: Bisias, Dimitrios and Flood, Mark D. and Lo, Andrew W. and Valavanis, Stavros, A Survey of Systemic Risk Analytics, 2012
21: Gautam Mitra, Leela Mitra, The Handbook of News Analytics in Finance, 2011
22: Ritu Agarwal, Vasant Dhar, Editorial—Big Data, Data Science, and Analytics: TheOpportunity and Challenge for IS Research, 2014,Information Systems Research,
23: ¨ yku Isık, Mary C. Jones, Anna Sidorova, Business intelligence success: The roles of BI capabilities and decisionenvironments, 2013,Information & Managemenr,Elsevier
24: R. Hema, Neha Malik, Data Mining and Business Intelligence, 2010
25: Ohbyung Kwon, Namyeon Lee, Bongsik Shin, Data quality management, data usage experience and acquisitionintention of big data analytics, 2014,International Journal of Information

Management ,Elsevier
26: Frachot, Antoine and Roncalli, Thierry , Mixing Internal and External Data for Managing Operational Risk, 2002,Social Science Research Network,
27: Π. Κικιλίας, Δ. Παλαμούρδας, Α. Πετράκης, Δ. Τσουκαλάς, Στατιστική - Πιθανότητες, 2001
28: Velicanu, Manole and Matei, Gheorghe, Database versus Data Warehouse, 2007,Social Science Research Network,
29: Surajit Chaudhuri , Umeshwar Daya, An overview of data warehousing and OLAP technology, 1997,ACM SIGMOD Record ,
30: Paulraj Ponniah, Data Warehousing Fundamentals for IT Professionals, 2010
31: Panos Vassiliadis, survey of Extract–transform–Load technology, 2009,International Journal of Data Warehousing & Mining,
32: DorianPyle, Data Preparation for Data Mining, 1999
33: ALANC. ACOCK, Working With Missing Values, 2005,Journal of Marriage and Family ,
34: Daniel A. Newman, Missing Data:Five Practical Guidelines, 2014,Organizational Research Methods,SAGE
35: EMILYNAMEY,GREGGUEST,LUCYTHAIRU,ANDLAURAJOHNSON, Data Reduction Techniques for Large Qualitative Data Sets, 2007,stanford.edu,
36: ASHA GOWDA KAREGOWDA, A. S. MANJUNATH & M.A.JAYARAM, COMPARATIVE STUDY OFATTRIBUTE SELECTION USING GAIN RATIOAND CORRELATION BASED FEATURE SELECTION, 2010,International Journal of Information Technology and Knowledge Management,
37: Ron Kohavi George H. John , Wrappers for feature subset selection, 1997,Artificial Intelligence,Elsevier Science Publishers Ltd.
38: Borda, Monica, Fundamentals in Information Theory and Coding, 2011
39: Greenwood, P.E., Nikulin, M.S, A guide to chi-squared testing, 1996
40: JOHN T. KENT, Information gain and a general measure of correlation, 1982,BIOMETRICA,
41: Hastie, Tibshirani and Friedman, The Elements ofStatistical Learning:Data Mining, Inference, and Prediction., 2009
42: W. M. P. van der AalstV. Rubin , H. M. W. Verbeek, B. F. van Dongen , E. Kindler,C. W. Günther , Process mining: a two-step approach to balance between underfitting and overfitting, 2010,Software & Systems Modeling,
43: PAOLO GIUDICI, Applied Data MiningStatistical Methods for Business and Industry, 2006
44: Fayyad, et.al., Advances in Knowledge Discovery and Data Mining, 1996
45: Tom Mitchell, Machine Learning, 1997
46: Pang-Ning Tan,Michael Steinbach, Vipin Kumar, , Introduction to Data Mining , 2006
47: Vapnik, Vladimir, The Nature of Statistical Learning Theory, 2000
48: William M. Bolstad, Understanding Computational Bayesian Statistics, 2010
49: Frank E. Harrell, Regression Modeling Strategies, 2001
50: Rokach, Lior; Maimon, Data Mining with Decision Trees: Theory and Applications, 2008
51: Breiman et al , Classification and Regression Trees by, 1984,,
52: Ian H. Witten, Eibe FrankMark, A. Hall, Data Mining:Practical Machine LearningTools and Techniques, 2010
53: , Pros and Cons of Decision Trees, , http://scikit-learn.org/stable/modules/tree.html
54: John Shawe-Taylor & Nello Cristianini , Support Vector Machines, 2010
55: Glenn Fung, Olga L. Mangasarian, Proximal Support Vector Machine Classifiers, ,,
56: Deutsches Institut für Wirtschaftsforschungwww.diw.deLaura Auria • Rouslan A. Moro, Support Vector Machines (SVM) as a Technique for Solvency Analysis, 2008,Dscussion papers,Deutsches Institut für Wirtschaftsforschungwww.diw.deLaura Auria • Rouslan A. Moro
57: Haykin, S. , Neural Networks: A Comprehensive Foundation, 1999
58: Ματσατσίνης Ν., Συστήματα Υποστήριξης Αποφάσεων, 2010

59: Opitz, D.; Maclin, R., Popular ensemble methods: An empirical study, 1999,Journal of Artificial Intelligence Research,
60: Ian H. WittenEibe FrankMark A. Hall, Data Mining, 2011
61: LEO BREIMAN, Machine Learning, 2001,,Kluwer Academic Publishers
62: Robert E. Schapire, Explaining AdaBoost, ,,
63: PAOLO GIUDICI, Applied Data MiningStatistical Methods for Business and Industry,
64: RYSZARD S. MICHALSKI , LEARNABLE EVOLUTION MODEL:Evolutionary Processes Guidedby Machine Learning, 2000,Machine Learning,Kluwer Academic Publishers
65: Alex A. Freitas, Evalutionary algorithms for data mining,
66: Thomas Back David B. Fogel Zbigniew Michalewicz, Handbook of Evolutionary Computation, 1997
67: Xinjie Yu · Mitsuo Gen, Introductionto Evolutionary Algorithms,
68: Price, K., Storn, R., Lampinen, J. , Differential Evolution: A Practical Approach to Global Optimization, 2005,, Springer
69: Marsella, Stone and Banks, Making customers analytics for you, 2005,ournal of Targeting, Measurement and Analysis for Marketing,Henry Stewart Publications 1479-1862
70: Thomas H. Davenport and Jeanne G. Harris, The Dark Side of Customer Analytics, 2007,HBR CASE STUDYThe Dark Side of Customer Analyticsby Thomas H. Davenport and Jeanne G. Harrisharvard business review ,
71: , Is your personal information really yours?, , olivercheek.com/bankruptcy-and-personal-data/
72: , Obama To Propose Laws On Hacking Notification, Student Privacy, , http://www.npr.org/2015/01/12/376623854/obama-to-propose-laws-on-hacking-notification-student-privacy
73: , Cyber Intelligence Sharing Act (CISA), ,
74: Injazz J. Chen and Karen Popovich, Understanding customerrelationship management(CRM)People, process and technology, 2003,Business Process Management Journal,
75: Carl Geppert, CUSTOMER CHURN MANAGEMENT :RETAINING HIGH - MARGIN CUSTOMERSWITH CUSTOMER RELATIONSHIPMANAGEMENT TECHNIQUES, ,,
76: Thomas H. Davenport, Competing on Analytics , 2006,decision making,harvard business review
77: Marsella, Stone and Banks, Making customer analytics workfor you!, 2005,,
78: Zhilin Yang, Robin T. Peterson, Customer Perceived Value,Satisfaction, and Loyalty:The Role of Switching Costs, 2004,Psychology & Marketing, ,
79: Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr , A Proposed Churn Prediction Model, 2012,International Journal of Engineering,
80: ROB MATTISON, CHURN TAXONOMY –PART 1: INVOLUNTARY CHURN, 2005
81: Rob Mattison, C HURN T AXONOMY –P ART 2: V OLUNTARYC HURN, 2005
82: Grahame R. Dowling and Mark Uncles, Do Customer Loyalty Programs Really Work?, 1997
83: Wouter Buckinx, Dirk Van den Poel*, Customer base analysis: partial defectionof behaviourally loyal clients in a non-contractualFMCG retail setting, 2005,European Journal of Operational Research,
84: NEELI BENDAPUDI, LEONARD1. BERRY, Customers' Motivations for MaintainingRelationships With Service Providers, 1997,Journalof Retailing ,
85: Kim, S. Y., T. S. Jung, E. H. Suh and H. S. Hwang (2006)., Customer segmentationand strategy development based on customer lifetime value: A case study, ,.Expert Systems with Applications,
86: Dipak JainSiddhartha S. Singh, , 2002,JOURNAL OF INTERACTIVE MARKETING,
87: Robert C. Blattberg,Byung-Do Kim,Scott A. Neslin, Database Marketing: Analyzing and Managing Customers, 2008
88: B. Huang et al., Customer churn prediction in telecommunications, 2012,Expert Systems with Applications,
89: Jonathan Burez, Dirk Van den Poel, CRM at a pay-TV company: Using analytical modelsto

reduce customer attrition by targeted marketingfor subscription services, 2007,Expert Systems with Applications ,
90: Yaya Xie a , Xiu Li a, * , E.W.T. Ngai b , Weiyun Ying c, Customer churn prediction using improved balanced random forests, 2009,Expert Systems with Applications ,
91: Ιωάννα Αγγελιδάκη, “Διαδικασίες Εξόρυξης Δεδομένων για την Πρόβλεψη καιΑνάλυση της Απώλειας Πελατών”,
92: , ROC CURVE, , https://www.medcalc.org/manual/roc-curves.php
93: , The Area Under an ROC Curve, , http://gim.unmc.edu/dxtests/roc3.htm
94: Bruce Ratner, Ph.D., Decile Analysis Primer:Cum Lift for Response Model, , http://www.dmstat1.com/res/DecileAnalysisPrimer.html

ΠΑΡΑΡΤΗΜΑ

ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΗΝ ΕΤΑΙΡΕΙΑ

Churn Modeling Tournament

TERADATA CENTER FOR
CUSTOMER RELATIONSHIP MANAGEMENT
AT DUKE UNIVERSITY

Scott Neslin, Director
Sanyin Siang, Managing Director

Sunil Gupta, Advisory Board
Wagner Kamakura, Advisory Board
Junxiang Lu, Advisory Board
Charlotte Mason, Advisory Board

Overview

Predictive modeling – the statistical process of "scoring" and targeting customers for a marketing campaign – is a significant database marketing tool and an important component of a firm's customer relationship management (CRM) effort. The promise of predictive modeling is the ability to predict what actions customers will take, thereby allowing firms to target their marketing efforts more effectively. One area of particular importance is customer "churn," in this case, customer voluntary churn, when current customers decide to take their business elsewhere or voluntarily terminate their service. Annual churn rates have been reported to be in the 20% - 40% range for telecommunication and other technology industries. This puts a premium on developing models that accurately predict which customers are most likely to churn, so proactive steps (e.g. appropriate communication and treatment programs) can be taken to prevent customers from churning. The purpose of the Churn Modeling Tournament is to learn which methods work best for predicting churn, thereby enhancing our overall understanding of predictive modeling.

The Teradata Center for Customer Relationship Management at Duke University (the Center) will provide data to those interested in participating in the tournament. The data consist of calibration and validation samples of customers from a major wireless telecommunications company. The calibration sample includes observed churn and a set of potential predictor variables. The two validation samples include the same predictor variables, but no churn variable. Participants will submit their predictions of likelihood to churn. We will merge those predictions with the actual churn records to evaluate predictive accuracy. The entries with the best prediction records will be the "winners." Cash prizes will be awarded.

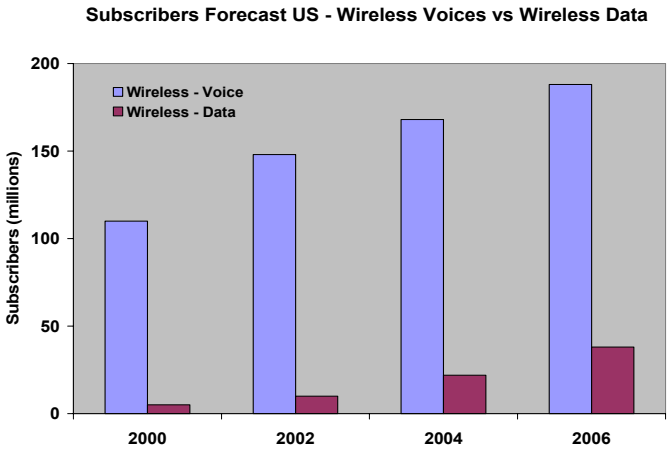
After the tournament is completed, we will conduct a "meta-analysis" of the results. We will determine which particular methodologies tend to work best and to what degree. We will make these results available to the general public and attempt to publish them in an academic, as well as a practitioner, journal. While the authors of any resulting paper will be the judges/organizers of the tournament, all entrants will be listed and acknowledged.

In summary, the tournament provides modelers with (1) the opportunity to gain recognition and win cash prizes, (2) the opportunity to participate in a collective effort that will enhance academic and practitioner knowledge of predictive modeling, and (3) the opportunity to learn first-hand about the challenges of predictive modeling in a particularly relevant context - churn.

Industry Background

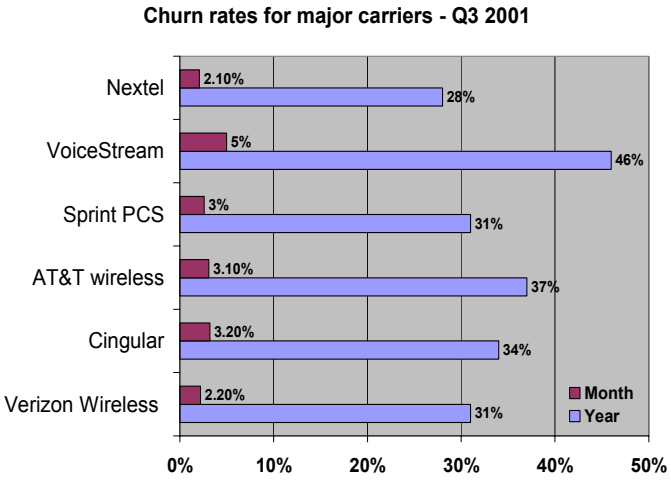
The Wireless Industry*:* During the last five years, the wireless sector has been one of the fastest-growing businesses in the economy. With a unique value proposition – freedom and connectivity – the number of subscribers doubled every two years during the 90's. Wireless stocks grew as fast as those of many dot-coms, start-ups emerged everywhere, and IPO's raised record amounts of money. These events shaped the new telecommunications landscape as we know it today. And there is promise for more developments to come (*Business Week* 2002; *Wireless News Factor* 2002):

- By 2003, 25% of all telephone minutes will be accounted for by wireless services.
- By 2006, the US penetration in the wireless-voice market is expected to hit 189 million subscribers, while that of the wireless-data market is expected to jump to 38 million subscribers.
- Of all wireless customers, 70% are using digital networks that allow carriers to efficiently offer more appealing services.
- Investment in network infrastructure has increased by 17% and the number of cell sites increased by 22.3%, indicating a clear upward trend in US coverage and quality.



Source: InfoTech (2002)

***Industry Turmoil**:* Despite the vertiginous levels of growth and promise, serious charges to industry profitability have recently emerged: (a) Consolidation: From the nearly 60 cellular companies, virtually all of them are now bankrupt, bought out, or struggling with heavy debts. Only six big players now account for 80% of the wireless pie. (b) Growth: Subscriber growth rates went from 50% yearly to 15% - 20% in 2002 and analysts predict a meager 10% growth rate in 2003 (*Business Week Online, 2002*). (c) Competition: As an obvious result (and to the consumer's delight), firms engaged in a devastating price war that not only eroded revenue growth but also endangered their ability to meet their titanic debts. (d) Customer Strategy: The industry paradigm has arguably changed from one of "make big networks, get customers" to "make new services, please customers." In short, the industry has moved from an acquisition orientation to a retention orientation.



Source: *Telephony Online*, 2002

***The Elusive Customer**:* Until now, firms have been able to acquire customers without much effort. Demand for wireless services has been such that if a customer decided to drop his service and switch to another carrier, another new customer was right behind him. The priority was to maintain the customer acquisition rate high, often at the expense of customer retention. But this situation has changed. As the well of wireless subscribers has begun to run dry, churn – the customer's decision to end the relationship and switch to another company – has become a major concern. Last year the industry average churn rate was 20% - 25% annually, which translates to approximately 2% churn per month. This means that companies lose 2% of their customers every month. Third quarter, 2001, statistics show annual churn rates in an even higher range, 28% - 46% annual churn.

The reasons for the high level of churn are: (a) variety of companies, (b) the similarity of their offerings, and (c) the cheap prices of handsets. In fact, the biggest current barrier to churn – the lack of phone number portability – is likely to change in the short term. Companies are now beginning to realize just how important customer retention is. In fact, one study finds that "the top six US wireless carriers would have saved $207 million if they had retained an additional 5% of customers open to incentives but who switched plans in the past year" (Reuters 2002). Over the next five years, the industry's biggest marketing challenge will be to control churn rates by identifying those customers who are most likely to leave and taking appropriate steps to retain them. The first step therefore is predicting churn likelihood at the customer level.

Data Description

The data provided have generously been provided to the Center by a major wireless carrier. The data are organized into three data files: Calibration, Current Score Data, and Future Score Data.

	Calibration	Current Score Data	Future Score Data
Sample Size	100,000	51,306	100,462
# of Predictor Variables	171	171	171
Churn Indicator	Yes	No	No
Customer ID	1,000,001 – 1,100,000	2,000,001 – 2,051,306	3,000,001 – 3,100,462

The Calibration Data contain the "dependent variable" – churn – as well as several potential predictors. The Current and Future Score Data contain the predictors but not churn. Participants in the tournament will therefore estimate models on the calibration data and use these models to predict for the Current and Future Score Data.

The "Data Documentation" spreadsheet provides detailed descriptions of all the variables. The predictors include three types of variables: *behavioral data* such as minutes of use, revenue, handset equipment; *company interaction data* such as customer calls into the customer service center, and *customer household demographics*.

Customers were selected as follows: mature customers, customers who were with the company for at least six months, were sampled during July, September, November, and December of 2001. For each customer, predictor variables were calculated based on the previous four months. Churn was then calculated based on whether the customer left the company during the period 31-60 days after the customer was originally sampled. The one-month treatment lag between sampling and observed churn was for the practical concern that in any application, a few weeks would be needed to score the customer and implement any proactive actions.

The actual percentage of customers who churn in a given month is approximately 1.8%. However, churners were over sampled when creating the Calibration sample to create a roughly 50-50 split between churners and non-churners (the exact number is 49,562 churners and 50,438 non-churners). Over sampling was not undertaken in creating the Current Score and Future Score validation samples. This is to provide a more realistic predictive test. The Current Score data contain a different set of customers from the Calibration data, but selected at the same point in time. The Future Score data contain a different set of customers selected at a future point in time. One interesting aspect of the tournament will be to investigate the accuracy for same-period versus future predictive accuracy.

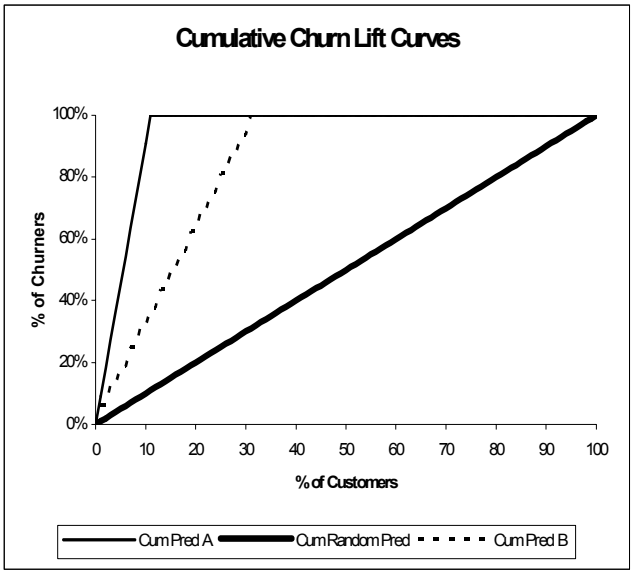
Prediction Criteria

We will calculate two measures of predictive accuracy for each submitted data file – Top Decile Lift and Gini Coefficient. Top Decile Lift measures whether the 10% of customers predicted most likely to churn actually churn. The Gini Coefficient measures predictive accuracy across the entire set of customers, not just the top 10%.

Top Decile Lift: We will sort the customers in the submitted file from predicted most likely to predicted least likely to churn. We then will take the top 10% and calculate the exact percentage that did in fact churn. To create an index, we will divide by the average churn rate across all customers. So for the Current Score Data, we will find the churn rate among the 10% x 51,306 = 5131 customers predicted most likely to churn[1]. Assume for example this turns out to be 3.9%, and that the churn rate among all customers in the Current Score Data turns out to be 1.80%. Therefore, the Top Decile Lift would be 3.9/1.80 = 2.17, or "2.17 to 1" lift.

Gini Coefficient: The Gini Coefficient is used in economics to measure phenomena such as income inequality (Wolff 2002; Sydsaeter and Hammond 1995). In database marketing, the Gini Coefficient works off the "Cumulative Lift Curves," shown in the figure to the right.

A Cumulative Lift Curve plots the top x% predicted customers versus the percentage of churners accounted for by these customers. For example, in the figure, the 10% of customers predicted most likely to churn by Method B account for 31.3% of all churners. The top 20% predicted customers account for 62.5% of all churners. That is better than random prediction (shown by the Cum Random line), where the top 10% would account for 10% of churners, and the top 20% would account for 20% of churners.



Generally, the higher the lift curve, the better. That is, a method predicts more accurately to the extent that the area between its cumulative lift curve and the lift curve for random prediction is large. In the figure, Method A is obviously better than Method B.

The Gini Coefficient is the area between a method's cumulative lift curve and the random lift curve. Technically, it should be calculated as an integral (Sydsaeter and Hammond 1995) but we will approximate it by a numerical measure (Alker 1965; Statistics.Com, 2002) since we have a finite number of customers and no closed form formula for the cumulative lift curve for a given method.

[1] Note the rounding. 10% of 51,306 is 5,130.6, which we round to 5131.

The formula we use is:

$$Gini = \left(\frac{2}{n}\right)\sum_{i=1}^{n}(v_i - \hat{v}_i)$$

where:

n = number of customers

v_i = % of churners who have predicted probability of churn equal to or higher than customer *i*.

$\hat{v}_i$ = % of customers who have predicted probability of churn equal to or higher than customer *i*.

v_i is the height of the method's cumulative lift curve at the i^{th} most likely predicted-to-churn customer, and $\hat{v}_i$ is the height of the random cumulative lift curve. The difference provides the "length" for calculating the area between the random and method prediction curves. The term *1/n* approximates the "width" on the x-axis. The Gini Coefficient sums these lengths-times-widths across customers, providing an approximation to the area between the method's lift curve and the random lift curve. The calculation is multiplied by "2" to ensure that the maximum possible Gini Coefficient is 1[2]. The Gini Coefficient for Method A in Figure 3 is .84; the Gini for Method B is .69. Random prediction will achieve a Gini of 0 (as seen in the formula above since for random prediction, $\hat{v}_i = v_i$) and higher Gini will correspond to more separation between the method's lift curve and random, which means better prediction.

Note that a method with a high first decile lift obviously has a head start toward achieving a high Gini Coefficient. But one method may do quite well on first decile lift but not well thereafter, and another method that doesn't do quite as well in first decile lift may do better overall.

[2] Note while this is the theoretical maximum, a Gini Coefficient exactly equal to 1 is possible with a finite data set only if there is only one churner and that churner is ranked first.

Procedures

Eligibility: Any interested party is welcome to participate. We anticipate receiving entries from four main groups: academic faculty, students, model builders working in industry, and software providers. Persons affiliated with the Center are not eligible to win.

Request for Data: Data are available for downloading from our website at http://faculty.fuqua.duke.edu/teradatacenter/ after the participant has registered. The data are in SAS dataset (v. 8.0) or CSV format. The data can also be placed on a CD and sent to the participant.

Requirements for Participation: The following are required for participating in the tournament:

1. The participant will submit predictions for both the Current Score Data and the Future Score Data, as well as for the Calibration Data.
2. When submitting the predictions, the participant will complete a brief questionnaire that asks questions regarding the methodology.
3. The participant agrees to be interviewed if follow-up clarifications on their method are needed[3].
4. The participant gives the Center permission to use the materials submitted to the tournament in resulting publications.

Submitting Predictions: The participant needs to create three files, one each for the Calibration, Current Score, and Future Score data. Each file will have two columns:

1. Customer ID: As stipulated in the original file.
2. Prediction: This can either be a rank order or a numeric score. If a rank order, we will assume that a lower rank order means more likely to churn (i.e., the customer ranked "1" is more likely to churn than the one ranked "2", etc.). If a numeric score, we will assume that a higher score means more likely to churn. In calculating top decile lift, if there are ties in the scores, i.e. two customers have the same score; the customer appearing first will be counted ahead of the subsequent customer.

In summary, the submitted Calibration prediction file will have 100,000 rows and 2 columns, the Current Score prediction file will have 51,306 rows and 2 columns, and the Future Score data will have 100,462 rows and 2 columns. Winners will be determined based on the predictive accuracy for the Current and Future Score Data, but we will use the Calibration predictions to measure out-of-sample prediction degradation.

[3] We understand that some entrants may be reluctant to discuss certain details of their methodology due to confidentiality. We will make every effort to ask reasonable questions in any follow-up.

Submissions can be uploaded onto the website under the subsection "Submit Results" in the modeling tournament section under the Current News & Events heading. The Center will also accept submissions on CDs sent to the Center address and addressed to "Churn Modeling Tournament". Please remember to include your username on all materials.

***Performance Feedback**:* Each entrant will receive his or her prediction evaluation scores for feedback purposes and so the entrant can compare with the statistics we will calculate across all entrants. This feedback will be made available after the tournament officially closes on January 1, 2003 and will be in the format of a scale from 1 – 10.

***Time Frame**:* The time frame for the tournament is August 1, 2002 through January 1, 2003. We will accept requests for data up to December 15, 2002, and will accept predictions that are received by 5:00 PM EDT, January 1, 2003.

***Governance**:* A governing board will monitor the process of the tournament to assure its integrity and to award prizes. Research assistants employed by the Teradata Center for Customer Relationship Management at Duke will calculate the predictive accuracy statistics as described above. Following are the members of the governing board:

Prof. Scott Neslin (Director), *Dartmouth College*
Sanyin Siang, *Teradata Center for Customer Relationship Management at Duke*
Prof. Sunil Gupta, *Columbia University*
Prof. Wagner Kamakura, *Duke University*
Dr. Junxiang Lu, *Industry Consultant*
Prof. Charlotte Mason, *University of North Carolina*

***Frequently Asked Questions (FAQs)**:* This description, the data, and the variable descriptions are intended to provide all information necessary for participating in the tournament. However, questions are allowed regarding the procedures, databases, and objectives. All questions and responses will be posted as FAQs on the Teradata Center website and any entrant can examine them.

***Prizes**:* For each of the four predictions (Current Score Data lift and Gini; Future Score Data lift and Gini), we will award $2000 to the entrant who receives the highest prediction score. It is possible for one entrant to win in each of the four categories. In case of a tie, the prize will be sub-divided. Prizes will be based on maximal scores. There will be no statistical testing for significant differences, etc.

Teradata Center for Customer Relationship Management at Duke University

The Teradata Center for Customer Relationship Management at Duke University advances the field of CRM through research and learning. Although various organizations exist internationally for studying CRM, the Center leverages the intellectual resources of a leading academic institution and business partnership to merge theory and practical business experience. The Center's overarching goal is to prioritize, facilitate and disseminate academic research and curriculum design aimed at advancing the field of CRM. It currently funds global CRM research, produces case studies and papers, develops curricula and provides data sets for use by other institutions and industry. The findings and offerings may influence the way corporations, students and academics view marketing.

The Center gratefully acknowledges the contributions of Emilio del Rio and Michael Kurima, who provided invaluable research assistance and data analysis in the development and organization of the tournament.

References

Alker, Hayward R., Jr. (1965) *Mathematics and Politics*, New York: The Macmillan Company.

Business Week (2002) "What Ails Wireless," April 1, 2002.

Business Week Online (2002) "Who'll Survive the Cellular Crisis?", February 15, 2002, http://www.businessweek.com/technology/content/feb2002/tc20020215_8884.htm.

InfoTech (2002) "Wireless Voice and Data Services Penetration," June 18, 2002.

Reuters (2002) http://news.com.com/2100-1033-276151.html?legacy=cnet.

Statistics.Com (2002) http://www.statistics.com/content/glossary/g/gini.html.

Sydsaeter, Knut, and Peter J. Hammond (1995) *Mathematics for Economic Analysis*, Englewood Cliffs, NJ: Prentice Hall.

Telephony Online (2002) "Standing by Your Carrier", March 19, 2002.

Wireless News Factor (2002) "Report: US Wireless Industry Flexes Muscles," May 21, 2002.

Wolff, Edward N. (2002) "The Impact of IT Investment on Income and Wealth Inequality in the Postwar US Economy," *Income Economics and Policy,* 14, 233-251.

ΠΕΡΙΓΡΑΦΗ ΜΕΤΑΒΛΗΤΩΝ

ΠΟΣΟΤΙΚΕΣ

Interval Variables	Explanation
ADJMOU	Billing adjusted total minutes of use over the life of the customer
ADJQTY	Billing adjusted total number of calls over the life of the customer
ADJREV	Billing adjusted total revenue over the life of the customer
ATTEMPT_MEAN	Mean number of attempted calls
ATTEMPT_RANGE	Range of number of attempted calls
AVG3MOU	Average monthly minutes of use over the previous three months
AVG3QTY	Average monthly number of calls over the previous three months
AVG3REV	Average monthly revenue over the previous three months
AVG6MOU	Average monthly minutes of use over the previous six months
AVG6QTY	Average monthly number of calls over the previous six months
AVG6REV	Average monthly revenue over the previous six months
AVGMOU	Average monthly minutes of use over the life of the customer
AVGQTY	Average monthly number of calls over the life of the customer
AVGREV	Average monthly revenue over the life of the customer
BLCK_DAT_MEAN	Mean number of blocked (failed) data calls
BLCK_DAT_RANGE	Range of number of blocked (failed) data calls
BLCK_VCE_MEAN	Mean number of blocked (failed) voice calls
BLCK_VCE_RANGE	Range of number of blocked (failed) voice calls
CALLFWDV_MEAN	Mean number of call forwarding calls
CALLFWDV_RANGE	Range of number of call forwarding calls
CALLWAIT_MEAN	Mean number of call waiting calls
CALLWAIT_RANGE	Range of number of call waiting calls
CC_MOU_MEAN	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls
CC_MOU_RANGE	Range of unrounded minutes of use of customer care calls
CCRNDMOU_MEAN	Mean rounded minutes of use of customer care calls
CCRNDMOU_RANGE	Range of rounded minutes of use of customer care calls
CHANGE_MOU	Percentage change in monthly minutes of use vs previous three month average
CHANGE_REV	Percentage change in monthly revenue vs previous three month average
COMP_DAT_MEAN	Mean number of completed data calls
COMP_DAT_RANGE	Range of number of completed data calls
COMP_VCE_MEAN	Mean number of completed voice calls
COMP_VCE_RANGE	Range of number of completed voice calls
COMPLETE_MEAN	Mean number of completed calls
COMPLETE_RANGE	Range of number of completed calls
CUSTCARE_MEAN	Mean number of customer care calls
CUSTCARE_RANGE	Range of number of customer care calls
DA_MEAN	Mean number of directory assisted calls
DA_RANGE	Range of number of directory assisted calls
DATOVR_MEAN	Mean revenue of data overage
DATOVR_RANGE	Range of revenue of data overage
DROP_BLK_MEAN	Mean number of dropped or blocked calls
DROP_BLK_RANGE	Range of number of dropped or blocked calls
DROP_DAT_MEAN	Mean number of dropped (failed) data calls
DROP_DAT_RANGE	Range of number of dropped (failed) data calls
DROP_VCE_MEAN	Mean number of dropped (failed) voice calls
DROP_VCE_RANGE	Range of number of dropped (failed) voice calls
EQPDAYS	Number of days (age) of current equipment
INONEMIN_MEAN	Mean number of inbound calls less than one minute
INONEMIN_RANGE	Range of number of inbound calls less than one minute
IWYLIS_VCE_MEAN	Mean number of inbound wireless to wireless voice calls
IWYLIS_VCE_RANGE	Range of number of inbound wireless to wireless voice calls
MONTHS	Total number of months in service
MOU_CDAT_MEAN	Mean unrounded minutes of use of completed data calls
MOU_CDAT_RANGE	Range of unrounded minutes of use of completed data calls
MOU_CVCE_MEAN	Mean unrounded minutes of use of completed voice calls
MOU_CVCE_RANGE	Range of unrounded minutes of use of completed voice calls
MOU_MEAN	Mean number of monthly minutes of use
MOU_OPKD_MEAN	Mean unrounded minutes of use of off-peak data calls
MOU_OPKD_RANGE	Range of unrounded minutes of use of off-peak data calls

MOU_OPKV_MEAN	Mean unrounded minutes of use of off-peak voice calls
MOU_OPKV_RANGE	Range of unrounded minutes of use of off-peak voice calls
MOU_PEAD_MEAN	Mean unrounded minutes of use of peak data calls
MOU_PEAD_RANGE	Range of unrounded minutes of use of peak data calls
MOU_PEAV_MEAN	Mean unrounded minutes of use of peak voice calls
MOU_PEAV_RANGE	Range of unrounded minutes of use of peak voice calls
MOU_RANGE	Range of number of minutes of use
MOU_RVCE_MEAN	Mean unrounded minutes of use of received voice calls
MOU_RVCE_RANGE	Range of unrounded minutes of use of received voice calls
MOUIWYLISV_MEAN	Mean unrounded minutes of use of inbound wireless to wireless voice calls
MOUIWYLISV_RANGE	Range of unrounded minutes of use of inbound wireless to wireless voice calls
MOUOWYLISV_MEAN	Mean unrounded minutes of use of outbound wireless to wireless voice calls
MOUOWYLISV_RANGE	Range of unrounded minutes of use of outbound wireless to wireless voice calls
OWYLIS_VCE_MEAN	Mean number of outbound wireless to wireless voice calls
OWYLIS_VCE_RANGE	Range of number of outbound wireless to wireless voice calls
OPK_DAT_MEAN	Mean number of off-peak data calls
OPK_DAT_RANGE	Range of number of off-peak data calls
OPK_VCE_MEAN	Mean number of off-peak voice calls
OPK_VCE_RANGE	Range of number of off-peak voice calls
OVRMOU_MEAN	Mean overage minutes of use
OVRMOU_RANGE	Range of overage minutes of use
OVRREV_MEAN	Mean overage revenue
OVRREV_RANGE	Range of overage revenue
PEAK_DAT_MEAN	Mean number of peak data calls
PEAK_DAT_RANGE	Range of number of peak data calls
PEAK_VCE_MEAN	Mean number of inbound and outbound peak voice calls
PEAK_VCE_RANGE	Range of number of inbound and outbound peak voice calls
PLCD_DAT_MEAN	Mean number of attempted data calls placed
PLCD_DAT_RANGE	Range of number of attempted data calls placed
PLCD_VCE_MEAN	Mean number of attempted voice calls placed
PLCD_VCE_RANGE	Range of number of attempted voice calls placed
RECV_SMS_MEAN	Mean number of received SMS calls
RECV_SMS_RANGE	Range of number of received SMS calls
RECV_VCE_MEAN	Mean number of received voice calls
RECV_VCE_RANGE	Range of number of received voice calls
RETDAYS	Number of days since last retention call
REV_MEAN	Mean monthly revenue (charge amount)
REV_RANGE	Range of revenue (charge amount)
RMCALLS	Total number of roaming calls
RMMOU	Total minutes of use of roaming calls
RMREV	Total revenue of roaming calls
ROAM_MEAN	Mean number of roaming calls
ROAM_RANGE	Range of number of roaming calls
THREEWAY_MEAN	Mean number of three way calls
THREEWAY_RANGE	Range of number of three way calls
TOTCALLS	Total number of calls over the life of the customer
TOTMOU	Total minutes of use over the life of the customer
TOTMRC_MEAN	Mean total monthly recurring charge
TOTMRC_RANGE	Range of total monthly recurring charge
TOTREV	Total revenue
UNAN_DAT_MEAN	Mean number of unanswered data calls
UNAN_DAT_RANGE	Range of number of unanswered data calls
UNAN_VCE_MEAN	Mean number of unanswered voice calls
UNAN_VCE_RANGE	Range of number of unanswered voice calls
VCEOVR_MEAN	Mean revenue of voice overage
VCEOVR_RANGE	Range of revenue of voice overage

Appendix

ADJMOU Return	Billings adjustments include any corrections to customer billing, including reimbursement for dropped calls, etc.
ATTEMPT_MEAN Return	PLCD_DAT_MEAN + PLCD_VCE_MEAN Lines 108 + 110 Attempted number of calls is equal to the sum of placed data calls and placed voice calls.
CCRNDMOU_MEAN Return	Rounded minutes refers to minutes rounded to the nearest whole minute, either rounding up (31 - 59 seconds) or rounding down (1 - 30 seconds). The minimum number of minutes of any single call is one minute. The value of 0 represents no calls were made.
COMPLETE_MEAN Return	COMP_DAT_MEAN + COMP_VCE_MEAN Lines 40 + 42 Completed number of calls is equal to the sum of completed data calls and completed voice calls.
CUSTCARE_MEAN Return	Customer care calls include any inbound calls to the company regarding complaints, disputes or questions (IVR Interactive Voice Response calls included).
DATOVR_MEAN Return	Overage represents calls or minutes of use over the number of minutes allowed by that customer's calling plan.
DROP_BLK_MEAN	BLCK_DAT_MEAN + BLCK_VCE_MEAN + DROP_DAT_MEAN + DROP_VCE_MEAN

ΠΟΙΟΤΙΚΕΣ

Class Variables	Explanation
ACTVSUBS	Number of active subscribers in household
ADULTS	Number of adults in household
AGE1	Age of first household member
AGE2	Age of second household member
AREA	Geographic area
ASL_FLAG	Account spending limit
CAR_BUY	New or used car buyer
CARTYPE	Dominant vehicle lifestyle
CHILDREN	Children present in household
CHURN	Instance of churn between 31-60 days after observation date
CRCLSCOD	Credit class code
CREDITCD	Credit card indicator
CRTCOUNT	Adjustments made to credit rating of individual
CSA	Communications local service area
CUSTOMER_ID	Unique tournament specific customer ID for scoring purposes
DIV_TYPE	Division type code
DUALBAND	Dualband
DWLLSIZE	Dwelling size
DWLLTYPE	Dwelling unit type
EDUC1	Education of first household member
ETHNIC	Ethnicity roll-up code
FORGNTVL	Foreign travel dummy variable
HND_PRICE	Current handset price
HHSTATIN	Premier household status indicator
HND_WEBCAP	Handset web capability
INCOME	Estimated income
INFOBASE	InfoBase match
KID0_2	Child 0 - 2 years of age in household
KID3_5	Child 3 - 5 years of age in household
KID6_10	Child 6 - 10 years of age in household
KID11_15	Child 11 - 15 years of age in household
KID16_17	Child 16 - 17 years of age in household
LAST_SWAP	Date of last phone swap
LOR	Length of residence
MAILFLAG	DMA: Do not mail flag
MAILORDR	Mail order buyer
MAILRESP	Mail responder
MARITAL	Marital status
MODELS	Number of models issued
MTRCYCLE	Motorcycle indicator
NEW_CELL	New cell phone user
NUMBCARS	Known number of vehicles
OCCU1	Occupation of first household member
OWNRENT	Home owner/renter status
PCOWNER	PC owner dummy variable
PHONES	Number of handsets issued
PRE_HND_PRICE	Previous handset price
PRIZM_SOCIAL_ONE	Social group letter only
PROPTYPE	Property type detail
REF_QTY	Total number of referrals
REFURB_NEW	Handset: refurbished or new
RV	RV indicator
SOLFLAG	Infobase no phone solicitation flag
TOT_ACPT	Total offers accepted from retention team
TOT_RET	Total calls into retention team
TRUCK	Truck indicator
UNIQSUBS	Number of unique subscribers in the household
WRKWOMAN	Working woman in household

ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ-ΜΕΤΑΒΛΗΤΕΣ ΧΩΡΙΣ ΔΙΑΛΟΓΗ

Α/Α	ΜΕΤΑΒΛΗΤΗ	Α/Α	ΜΕΤΑΒΛΗΤΗ	Α/Α	ΜΕΤΑΒΛΗΤΗ
1	rev_Mean	42	mou_peav_Mean	83	uniqsubs
2	mou_Mean	43	mou_pead_Mean	84	actvsubs
3	totmrc_Mean	44	opk_vce_Mean	85	crclscod
4	da_Mean	45	opk_dat_Mean	86	asl_flag
5	ovrmou_Mean	46	mou_opkv_Mean	87	totcalls
6	ovrrev_Mean	47	mou_opkd_Mean	88	totrev
7	vceovr_Mean	48	drop_blk_Mean	89	adjrev
8	datovr_Mean	49	attempt_Mean	90	adjqty
9	rev_Range	50	complete_Mean	91	avgmou
10	totmrc_Range	51	callwait_Mean	92	avgqty
11	da_Range	52	blck_vce_Range	93	avg3mou
12	ovrmou_Range	53	blck_dat_Range	94	avg3qty
13	ovrrev_Range	54	unan_vce_Range	95	avg3rev
14	vceovr_Range	55	unan_dat_Range	96	avg6mou
15	datovr_Range	56	plcd_dat_Range	97	avg6qty
16	change_mou	57	recv_vce_Range	98	avg6rev
17	drop_vce_Mean	58	comp_vce_Range	99	REF_QTY
18	drop_dat_Mean	59	comp_dat_Range	100	tot_ret
19	blck_vce_Mean	60	custcare_Range	101	tot_acpt
20	blck_dat_Mean	61	ccmdmou_Range	102	div_type
21	unan_vce_Mean	62	cc_mou_Range	103	area
22	unan_dat_Mean	63	inonemin_Range	104	dualband
23	plcd_vce_Mean	64	threeway_Range	105	refurb_new
24	plcd_dat_Mean	65	mou_cvce_Range	106	hnd_price
25	recv_vce_Mean	66	mou_cdat_Range	107	phones
26	comp_vce_Mean	67	mou_rvce_Range	108	last_swap
27	comp_dat_Mean	68	owylis_vce_Range	109	models
28	custcare_Mean	69	mouowylisv_Range	110	hnd_webcap
29	ccmdmou_Mean	70	iwylis_vce_Range	111	marital
30	cc_mou_Mean	71	mouiwylisv_Range	112	mailordr_mailresp
31	inonemin_Mean	72	peak_dat_Range	113	age1
32	threeway_Mean	73	mou_peav_Range	114	age2
33	mou_cvce_Mean	74	mou_pead_Range	115	infobase
34	mou_cdat_Mean	75	opk_vce_Range	116	income
35	mou_rvce_Mean	76	opk_dat_Range	117	numbcars
36	owylis_vce_Mean	77	mou_opkv_Range	118	ethnic
37	mouowylisv_Mean	78	mou_opkd_Range	119	kid0_2
38	iwylis_vce_Mean	79	drop_blk_Range	120	creditcd
39	mouiwylisv_Mean	80	complete_Range	121	car_buy
40	peak_vce_Mean	81	callwait_Range	122	retdays
41	peak_dat_Mean	82	months	123	eqpdays

Με 124η την κλάση churn.

ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΙΑΓΩΝΙΣΜΟΥ ΠΑΡΟΧΟΥ ΔΕΔΟΜΕΝΩΝ

Churn Modeling for Mobile Telecommunications:

Winning the Duke/NCR Teradata Center for CRM Competition

N. Scott Cardell, Mikhail Golovnya, Dan Steinberg

Salford Systems

http://www.salford-systems.com

June 2003

The Churn Business Problem

- Churn represents the loss of an existing customer to a competitor
- A prevalent problem in retail:
  - Mobile phone services
  - Home mortgage refinance
  - Credit card
- Churn is a problem for any provider of a subscription service or recurring purchasable.
  - Costs of customer acquisition and win-back can be high
  - Much cheaper to invest in customer retention
  - Difficult to recoup costs of customer acquisition unless customer is retained for a minimum length of time
- Churn is especially important to mobile phone service providers
  - easy for a subscriber to switch services.
  - Phone number portability will remove last important obstacle

Churn a Core CRM issue

- The core CRM issues include:
  - Customer acquisition
  - Customer retention
  - Cross-sell/Up Sell
  - Maximizing Lifetime Customer Value
- Churn can be combated by
  - Acquiring more loyal customers initially
  - Taking preventative measures with existing customers
  - Identifying customers most likely to defect
    - offering incentives to those customers you want to keep
- All CRM management needs to take churn into account

Predicting Churn: Key to a Protective Strategy

- Predictive modeling can assist churn management
  - by tagging customers most likely to churn
- High risk customers should first be sorted by profitability
- Campaign targeted to the most profitable at-risk customers
  - Typical retention campaigns include:
    - Incentives such as price breaks
    - Special services available only to select customers
- To be cost effective retention campaigns must be targeted to the right customers
  - Customers who would probably leave without the incentive
  - Costly to offer incentives to those who would stay regardless

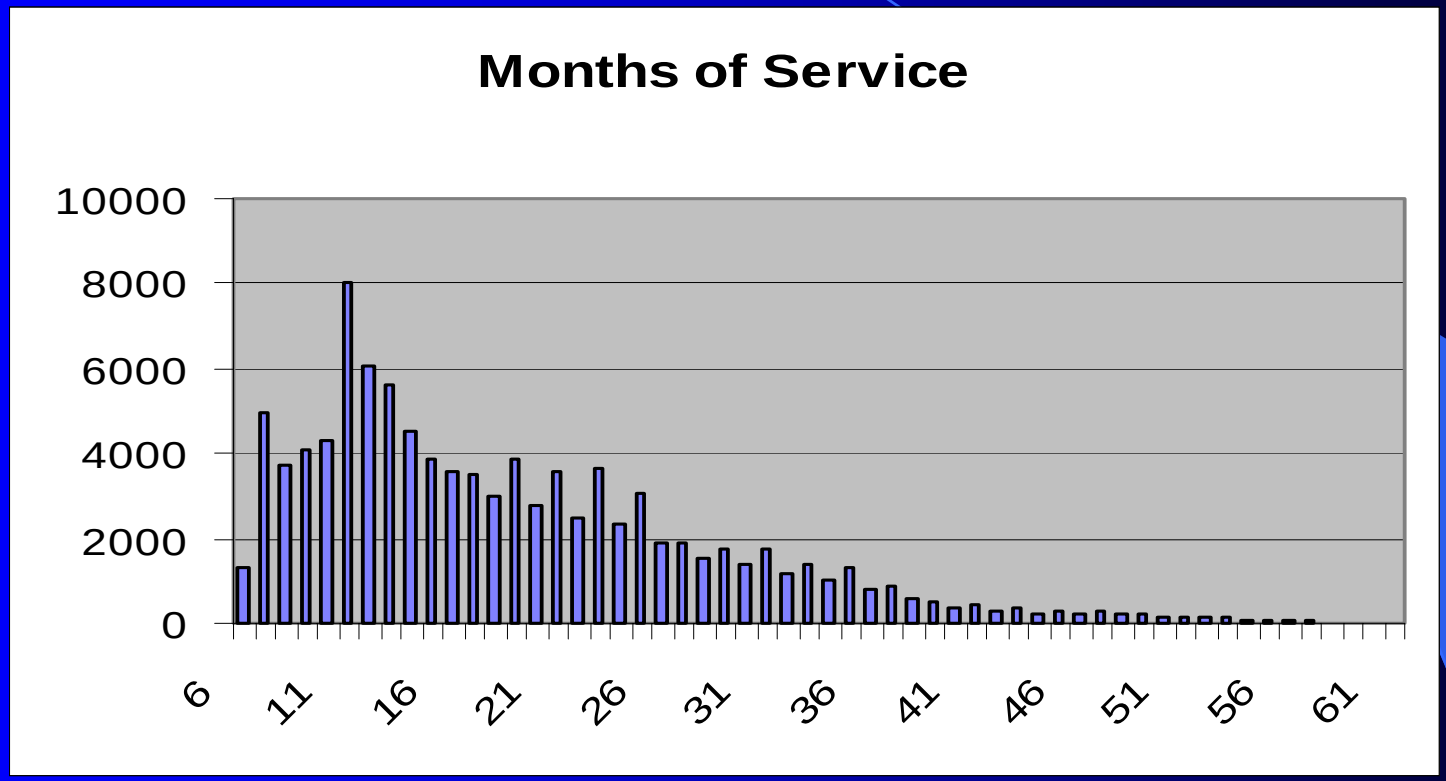
Duke/NCR Teradata 2003 Tournament

- The CRM Center sought to identify *best practice for churn modeling* in a real world context
- Solicited a major wireless telco to provide customer level data for an international modeling competition.
- Data suitable for churn modeling and prediction
- Competition was opened Aug 1, 2002 to all interested participants.
  - Publicized in a variety of data mining web sites, mailing lists, and SIGs (special interest groups)
  - Participants were given until January 10, 2003 to submit their predictions

Nature of the Data and Challenge

- Data were provided for 100,000 customers with at least 6 months of service history
  - One summary record per mobile phone account
  - Stratified into equal numbers of churners and non-churners
- Historical information provided in the form of
  - Type and price of current handset
  - Date of last handset change/upgrade
  - Total revenue (expenditure)
    - Broken down into recurring charges and non-recurring charges
  - Call behavior statistics (type, number, duration, totals, etc.)
  - Demographic and geographical information,
    - including familiar Acxiom style direct mail and marketing variables
    - Census-derived neighborhood summaries.

Accounts by Months of Service



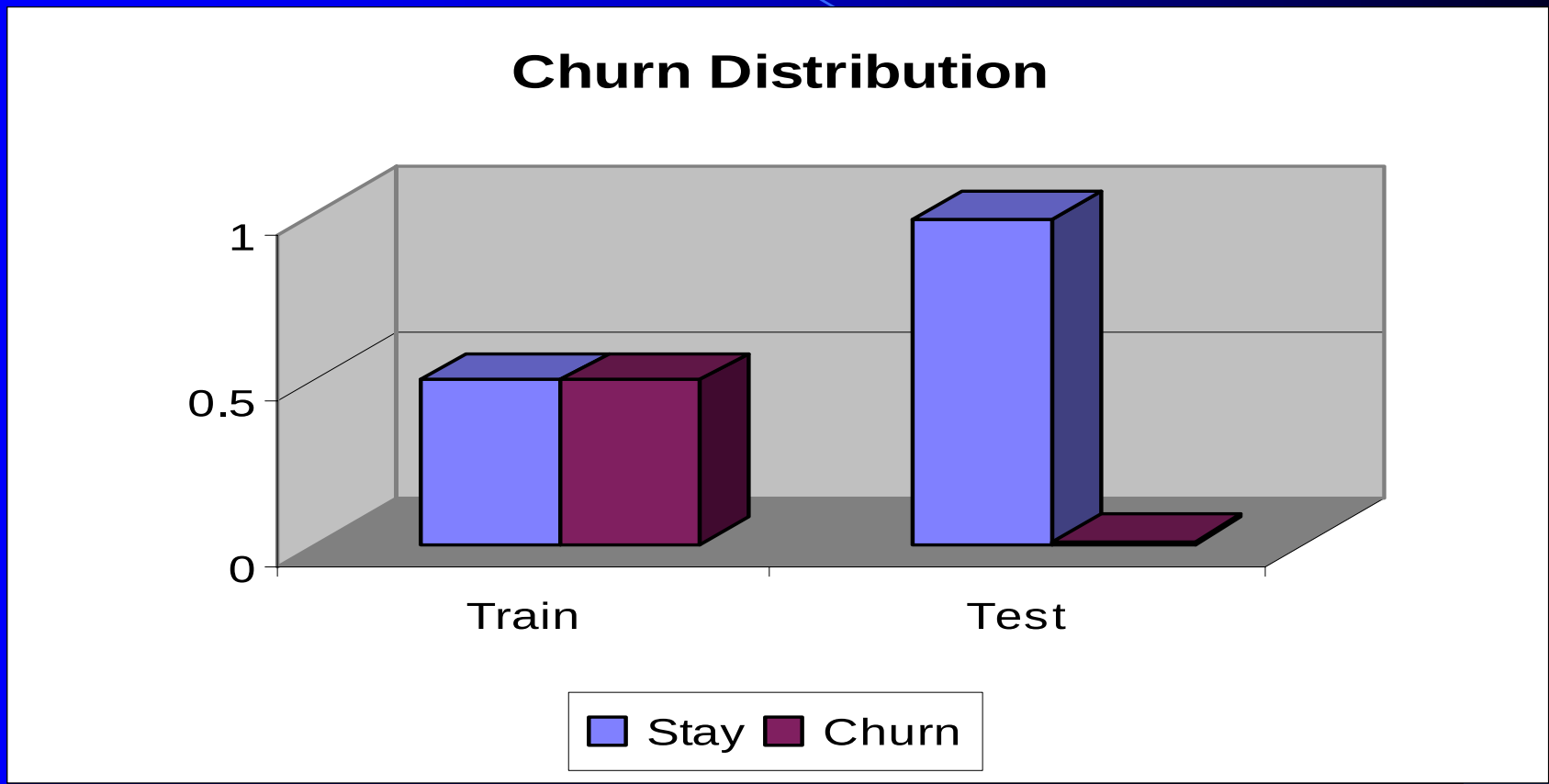
Time Structure of Data

- Data in the form of a “snapshot” of customer at a specific point in time
  - Month of data capture was one of:
  - July, September, November, December, 2001
- Historical data referred to
  - Current period (current plus prior 3 months)
  - Prior 3 months, prior 6 months
  - Lifetime (at least 6 months, as much as 5 years)
- Forecast Period
  - Churn or not in period 31-60 days after “snapshot”

Time Shape of the Data

- Data were captured for an account at a specific point in time
- From the perspective of that time retrospective summaries were computed
- Churn models were evaluated on forecast accuracy for the period 31 60 days
- To reflect data from different calendar months accounts were captured in July, September, November, December, 2001
- Which month a record was captured in was not available to the modelers

Train vs. "Test" Churn Distribution



Care required when forecasting from Train to Test data

Nature of Call Behavior Data

- Summary statistics describing number, duration, etc. of:
  - completed calls
  - failed calls
  - voice calls
  - data calls
  - call forwarding
  - customer care calls
  - directory info
- Statistics included mean and range for
    - Current period
    - Preceding 3 months,
    - Preceding 6 months
    - lifetime.

Evaluation Data

- Models were evaluated by performance on two different groups of accounts
  - "Current" data
    - Unseen accounts also drawn from July thru December 2001
  - "Future" data
    - Unseen accounts drawn from first half of 2002
- Evaluation on "current" data is the norm
  - Arranged by holding back some data for this purpose
- "Future" data is a more realistic and stringent test
  - Data used to test comes from a later time period
  - Markets, processes, behaviors change over time
  - Models tend to degrade over time due to changes in customer base and changes in offers by competitors, technology, etc.
- Model for best future performance may not be best for current performance, and vice verca

Modeling Observations

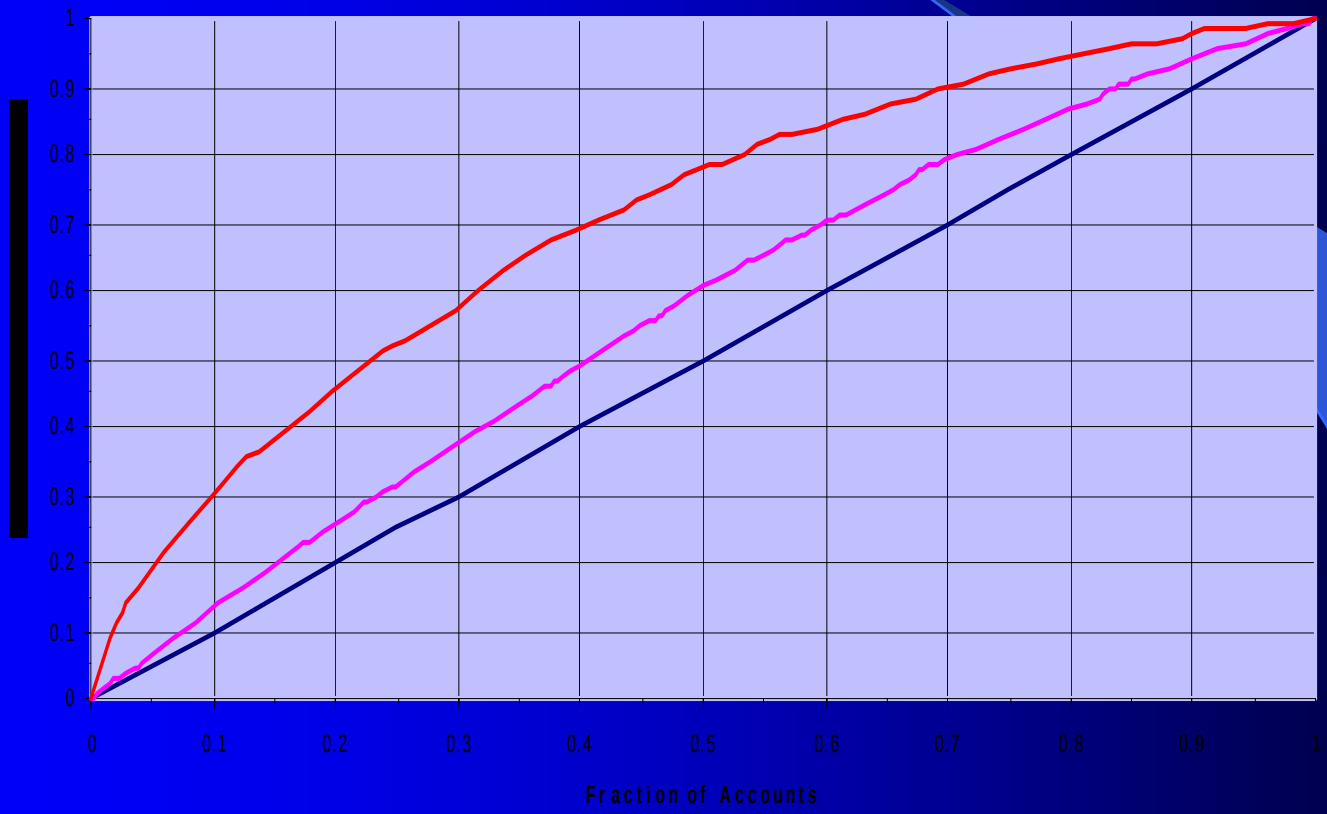
- Competition defined a sharply defined task:
  - churn within a specific window for existing customers of a minimum duration.
  - Objective was to predict probability of loss of a customer 30-60 days into the future.
- Challenge was defined in a way to avoid complications of censoring
  - Censored data could require survival analysis models.
- Each customer history was already summarized.
  - Only a modest amount of data prep required
  - No access to the raw data was provided so new summary construction was not possible
- Data quality was good
  - Perhaps because derived largely from operational database
- Majority of effort could be devoted to modeling

Model Evaluation

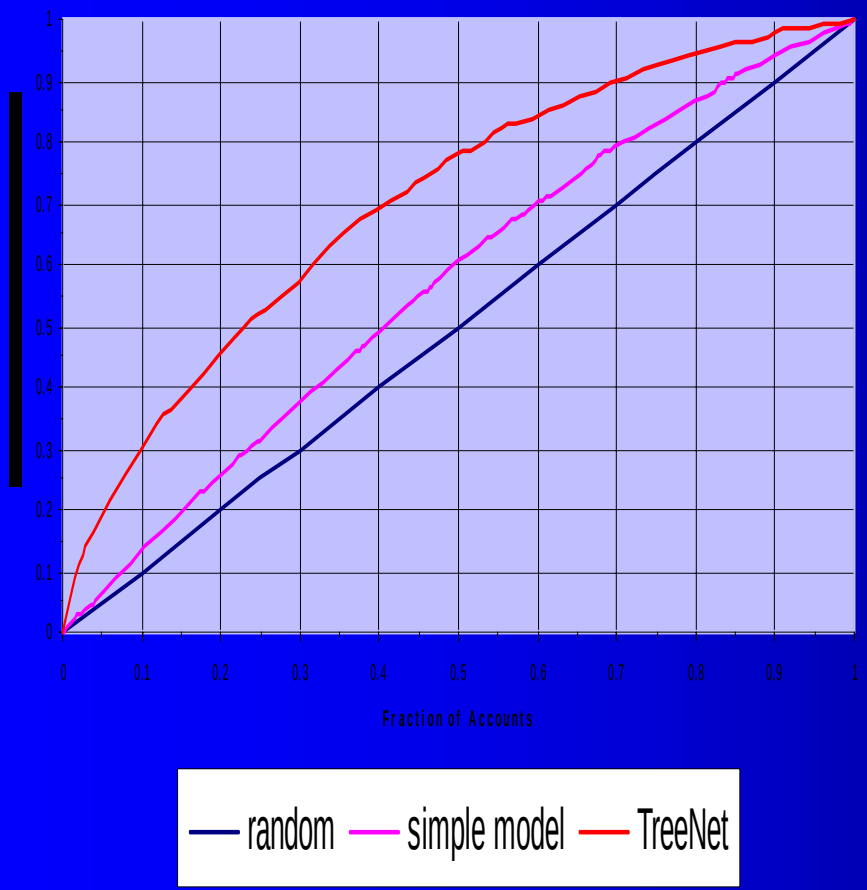
- Lift in top decile
  - Fraction of churners actually captured among the 10% “most likely to churn” as rated by the model
- Overall model performance as measured by the Gini coefficient
  - Area under the gains curve as illustrated on following slides
- Measures calculated for two different time periods
  - “Current” (June thru December 2001)
  - “Future” (First quarter 2002)

  Two measures calculated on each of two time periods yields 4 performance indicators in total
- Salford models were best in all four categories

Gains Chart: TreeNet vs Other



Top Decile Lift and Gini



- To produce the gains chart we order the data by the probability of event (churn)
- We plot the %target class captured as we move deeper into the ordered file
- In the example at left we capture 30% of churners in the top 10% of accounts and 70% in top 40% of accounts
- Models can also be evaluated by the “area under the curve” or Gini coefficient

Comparative Model Results: Top Decile Lift

	Future Data	Current data
Number of Accounts	100,000	51,036
Number Churning	1,808	924
Top decile capture		
Salford entry	525	278
2nd best model	506	253
Average of models	387	193
Salford Advantage		
Over runner up	3.8%	9.9%
Over average	35.7%	44.0%

Comparative Model Results: Overall Model Performance (Gini)

	Future Data	Current data
Number of Accounts	100,000	51,036
Number Churning	1,808	924
Gini Coefficient		
Salford entry	.409	.400
2nd best model	.370	.361
Average of models	.269	.261
Salford Advantage		
Over runner up	10.5%	10.8%
Over average	52.0%	53.2%

Comparative Model Results

Data Set	Measure	TreeNet Ensemble	Single TreeNet	2nd Best	Avg. (Std)
Current	Top Decile Lift	2.90	2.88	2.80	2.14 (.536)
Current	Gini	.409	.403	.370	.269 (.096)
Future	Top Decile Lift	3.01	2.99	2.74	2.09 (.585)
Future	Gini	.400	.403	.361	.261 (.098)

Model Observations

- Single TreeNet model always better than 2nd best entry in field.
- Ensemble of TreeNets slightly better than a single TreeNet 3 out of 4 times.
- TreeNet entries substantially better than the average.
- Minimal benefits from data preprocessing above and beyond that already embodied in the account summaries
- Virtually no manual, judgmental, or model guided variable selection
  - We let TreeNet do all the work of variable selection

Business Benefits of TreeNet Model

- In broad telecommunications markets the added accuracy and lift of TreeNet should yield substantially increased revenue
- For each 5 million customers over a one year period our models could capture as many as 20,000 more churn accounts in top decile
- Average revenue per month per account is $58.69 and $704 per year
- A customer retention rate of 15% should yield over $2 million per year in added revenue from the top decile alone
- The larger mobile telcos in the US and Europe have huge customer bases. Sprint PCS boasts 50 million accounts, so for Sprint the benefits could show in the vicinity of $20 million per year in added revenue.

Data Preparation

- A minimal amount of data preprocessing was undertaken to repair and extend original data.
- Some missing values could be recoded to "0."
- Select non-missing values were recoded to missing.
- Experiments with missing value handling were conducted, including the addition of missing value indicators to the data.
  - CART imputation
  - "All missings together" strategies in decision trees
    - Missings in a separate node
    - Missings go with non-missing high values
    - Missings go with non-missing low values

Modeling Tool of Choice:
TreeNet™ Stochastic Gradient Boosting

- TreeNet was key to winning the tournament.
  - Provided much greater accuracy and top decile lift than any other modeling method we tried.
- A new technology, different than standard boosting, developed by Stanford University Professor Jerome Friedman.
- Based on the CART® decision tree and thus inherits these characteristics:
  - Automatic feature selection
  - Invariant with respect to order-preserving transforms of predictors
  - Immune to outliers
  - Built-in methods for handling missing values

How TreeNet Works

- Goal is to model a target variable Y as a function of the data X
  - $Y = F(X)$
- We make this nonparametric by expressing the function as a series expansion

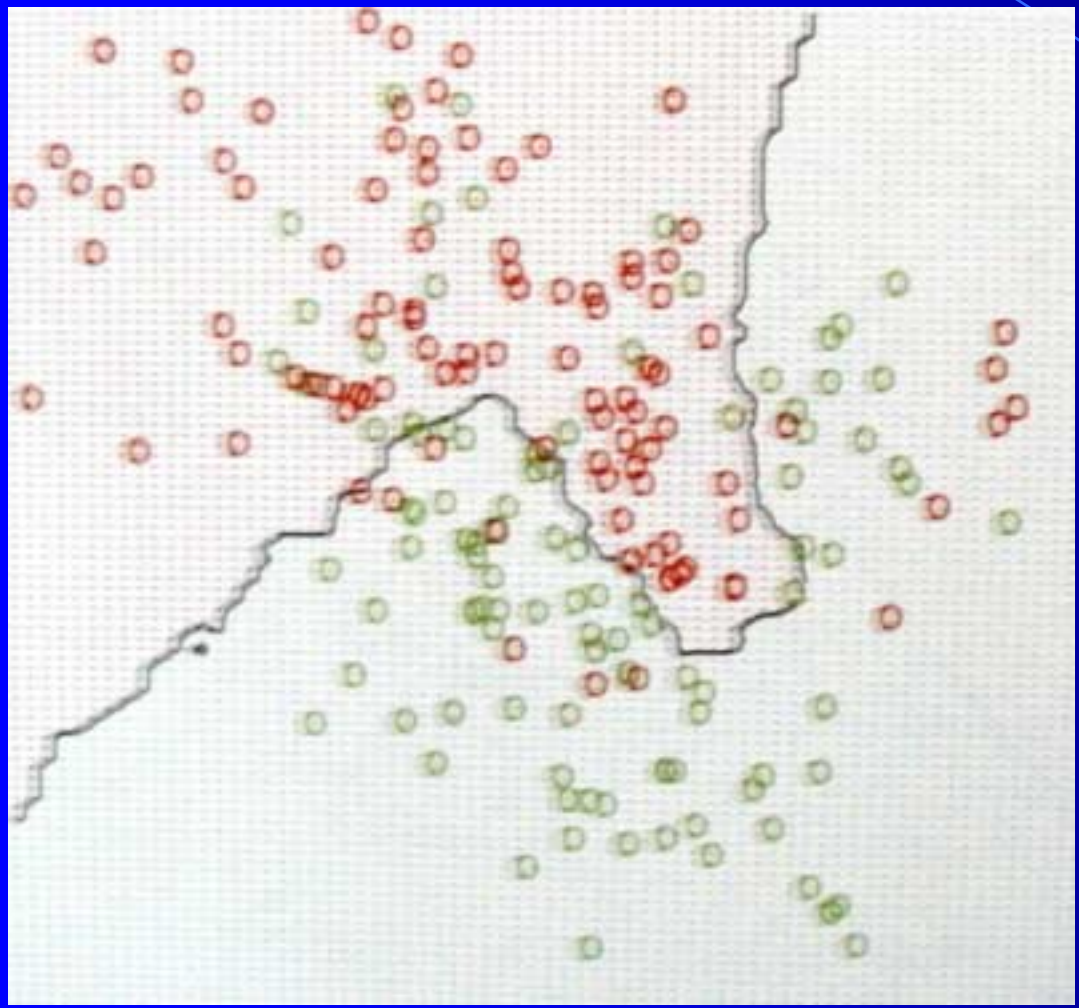
$$F(X) = F_0 + \beta_1 T_1(X) + \beta_2 T_2(X) + \ldots + \beta_M T_M(X)$$

- Expansion is developed one stage at a time
  - Each stage is a new learning cycle starting again using "all" the data
  - A term is learned once and never updated thereafter
- Each term of the series will typically be a small decision tree
  - As few as 2 nodes, typically 4 to 6 nodes, occasionally more
- Fit obtained by optimizing an objective function
  - e.g: likelihood function or sum of squared errors.

TreeNet Mechanics

- Stagewise function approximation in which each stage models transformed target (e.g. residuals) from last step model
  - Each stage uses a very small tree, as small as 2 nodes and typically in the range of 4-9 nodes
  - Each stage is intended to learn only a little
- Each stage learns from a fraction of the available training data, typically less than 50% to start
  - Slow learning intended to protect against overfitting
  - Data fraction used often falls to 20% or less by the last stag
- Each stage updates model only a little: severely downweighted contribution of each new tree (learning rate is typically 0.10, even 0.01 or less)
- In classification, focus is on points near decision boundary; ignores points far away from boundary even if the points are on the wrong side

Decision Boundary: 2 predictors



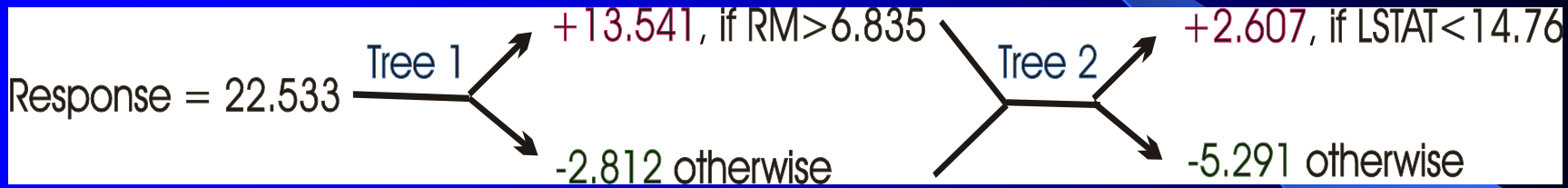
- Red dots represent YES (+1)
- Green dots represent NO (–1)
- Black curve is the current stage decision boundary
- TreeNet will not use data points too far from boundary to learn model update

TreeNet Objective Functions

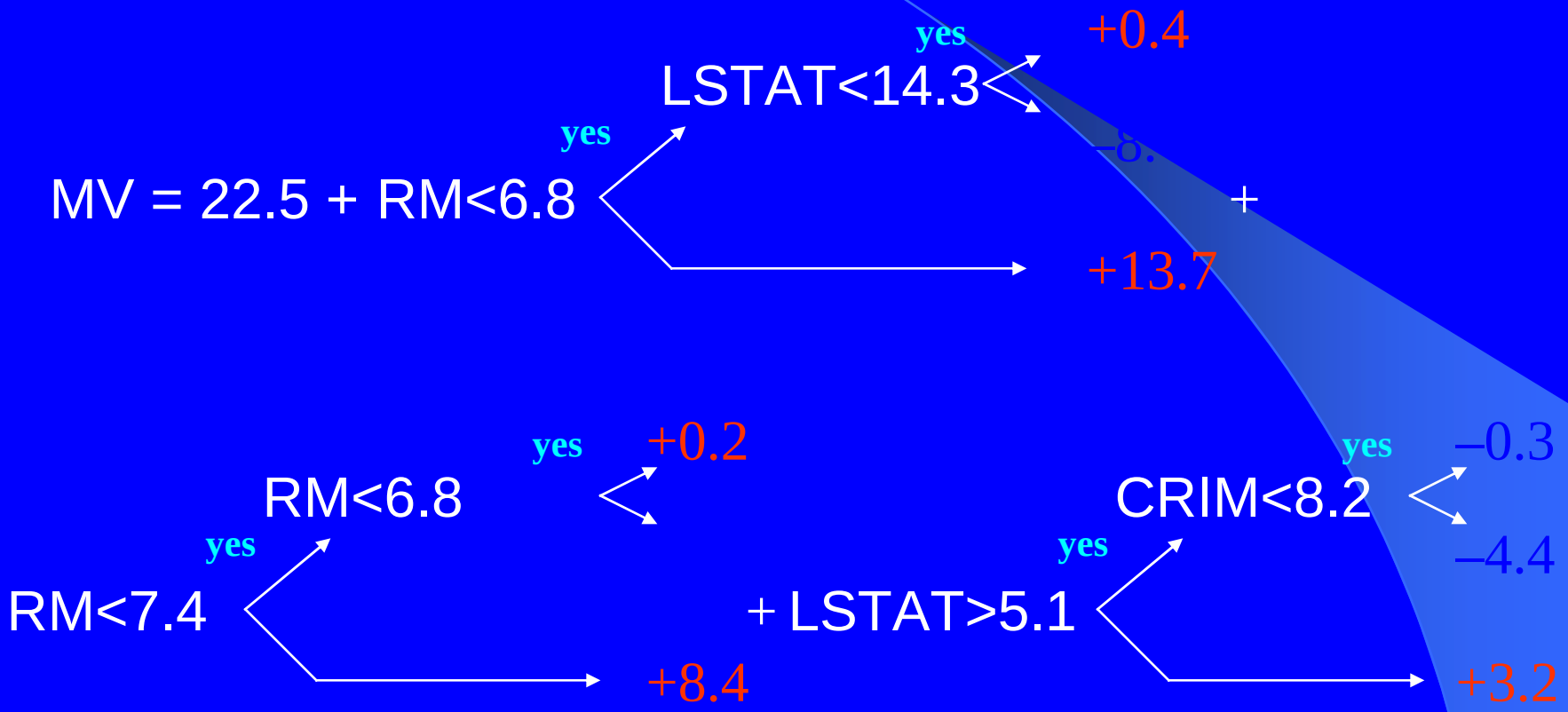
- For categorical targets (classification)
  - binary classification
  - multinomial classification
  - Logistic regression
- For continuous targets (regression)
  - least-squares regression
  - least-absolute-deviation regression
  - M-regression (Huber loss function)
- Other objective functions are possible and will be added in the future

Simple TreeNet Example

- **First two stages of a regression model**
  - **Each stage below is a 2 node tree**



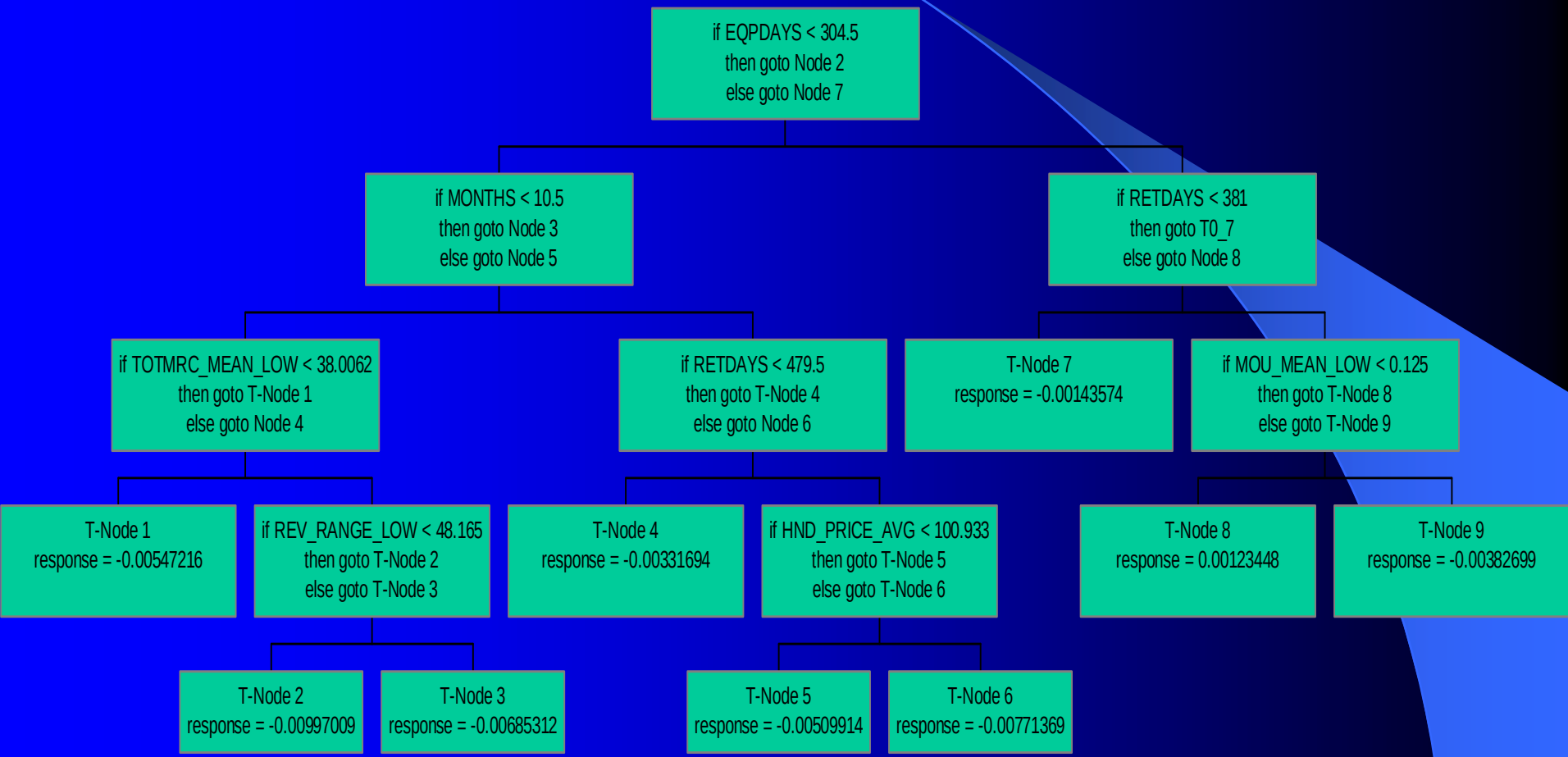
TreeNet Model with Three-Node Trees



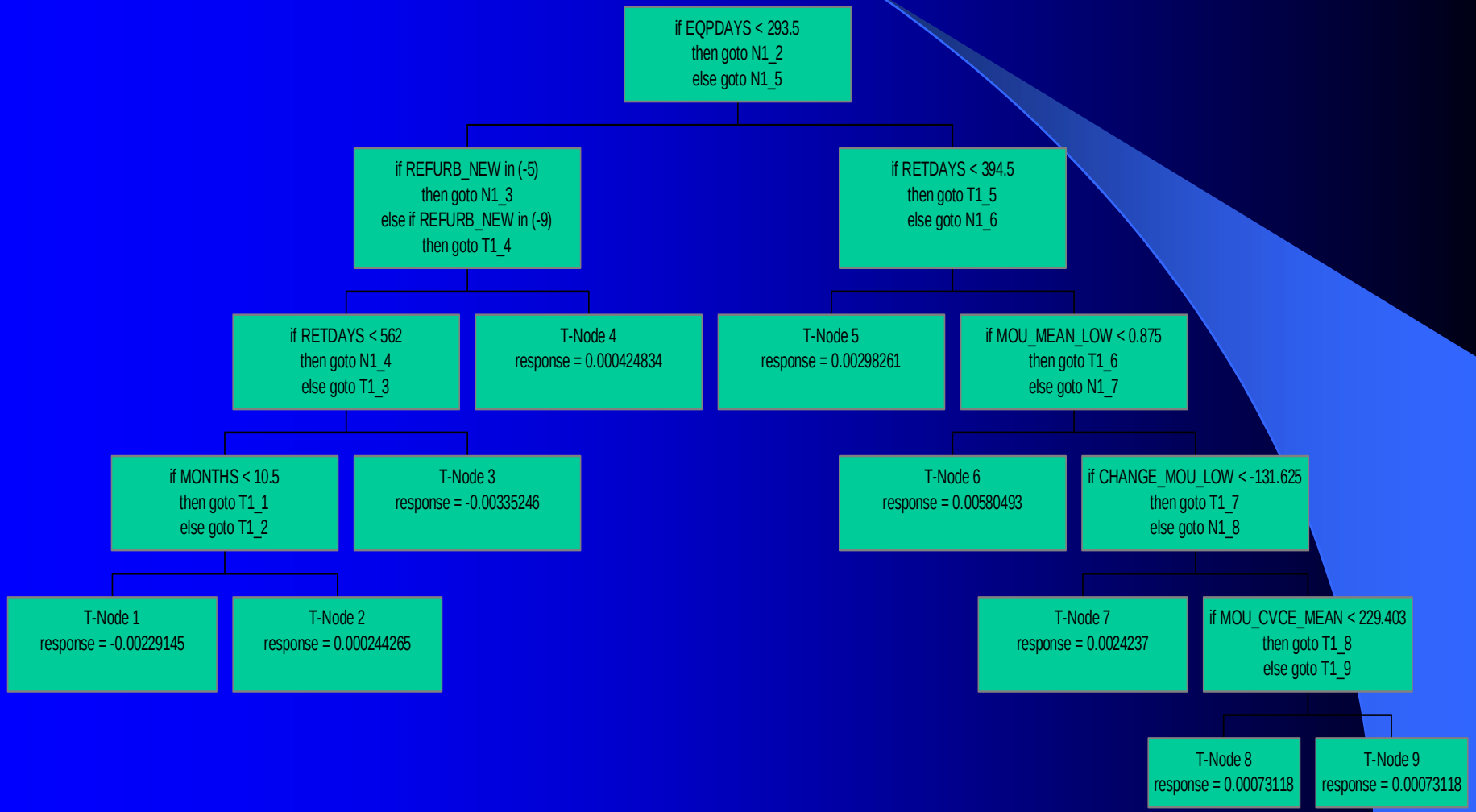
Each tree has three terminal nodes, thus partitioning data at each stage into three segments

Treenet churn model first tree

Tree 1 of 2912



Treenet churn model second tree



TreeNet Objective Function for Churn :
Logistic Log-Likelihood LL

- $$LL = \sum_i \log\left(1 + e^{-2\, y_i F(x_i)}\right) = \sum_i l\left(y_i, F(x_i)\right)$$

- The dependent variable, *y*, is coded (-1, +1)
- The target function, *F(x)*, is *1/2* the log-odds ratio.
- F_0 is initialized to the log odds on the full training data set.
  - equivalent to fitting data to a constant.

$$F_o(X) = \frac{1}{2}\log\left(\frac{1+\bar{y}}{1-\bar{y}}\right)$$

Patient Learning Strategy: Key to TreeNet Success

- **Do not use all training data in any one iteration.**
  - **Randomly sample from training data (we used a 50% sample).**
- **Compute log-likelihood gradient for each observation.**

$$G(y_i, x_i) = \frac{\partial l(y_i, F_m(x_i))}{\partial F_m(x_i)} = \frac{2y_i}{1+e^{2y_i F_m(x_i)}}$$

- **Build a K-node tree to predict $G(y_i, x_i)$.**
  - **$K=9$ gave the best cross-validated results.**
  - **Important that trees be much smaller than the size of an optimal single CART tree.**

Gradient Optimization

- Let

$$H(y_i, x_i) = -\frac{\partial^2 l(y_i, F_m(x_i))}{\partial F_m(x_i)^2}$$

- Update formula

$$H(y_i, x_i) = -\frac{\partial \frac{2y_i}{1+e^{2y_i F_m(x_i)}}}{\partial F_m(x_i)} = \frac{4e^{2y_i F_m(x_i)}}{\left(1+e^{2y_i F_m(x_i)}\right)^2} = |G(y_i, x_i)|\left(2-|G(y_i, x_i)|\right)$$

- Repeat until T trees grown.
- Select the value of $m \leq T$ that produces the best fit to the test data.

$$F_{m+1}(x_i) = F_m(x_i) + \sum_n \beta_{mn} 1(x_i \in \Phi_{mn})$$

Newton-Raphson Step

- Compute γ_{mn}, a single Newton-Raphson step for β_{mn}.

$$\gamma_{mn} = \frac{\sum_{i \in \Phi_{mn}} G(y_i, x_i)}{\sum_{i \in \Phi_{mn}} H(y_i, x_i)}$$

- Use only a small fraction, ρ of γ_{mn}. ($\beta_{mn} = \rho\gamma_{mn}$).
- Apply the update formula

$$F_{m+1}(x_i) = F_m(x_i) + \sum_n \beta_{mn} 1(x_i \in \Phi_{mn})$$

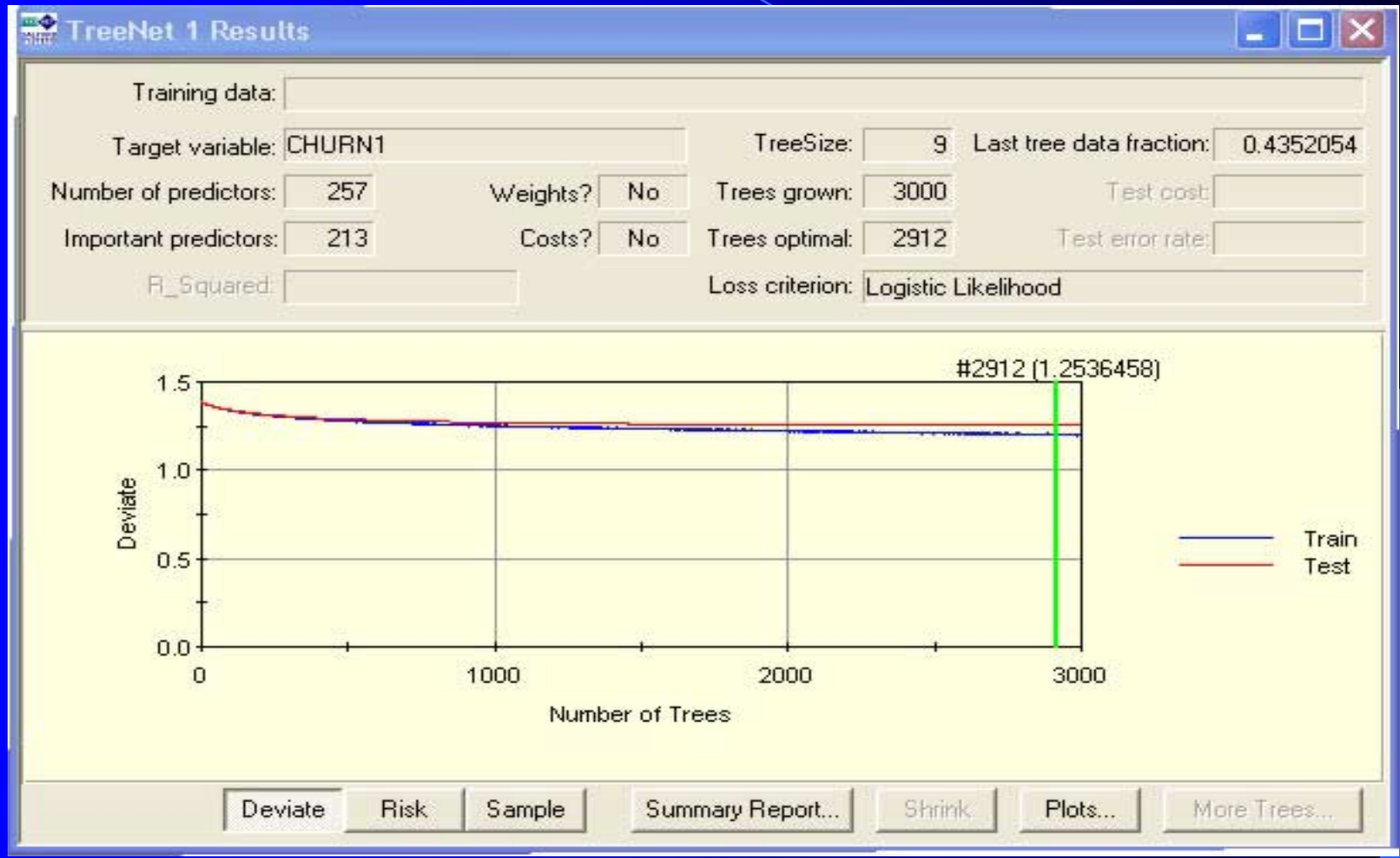
Learning Rates, Step Size, N Trees

- ρ is called the learning rate, T is the number of trees grown.
- The product ρT is the total learning.
  - Holding ρT constant, smaller ρ usually improves model fit to test data, but can require many trees.
- Reducing the learning rate tends to slowly increase the optimal amount of total learning.
- Very low learning rates can require many trees.
- Our CHURN models used values of ρ from .01 to .001.
- We used total learning of between 6 and 30.
- Our optimal models contained about 3,000 trees

The Salford CHURN Models

- All the models used to score the data for the entries used 9-node trees.
- Our final models used the following three combinations:
  - (ρ =.001; T = 6000; ρT = 6);
  - (ρ =. 005; T = 2500; ρT = 12.5);
  - (ρ =. 01; T = 3000; ρT = 30)
- One entry was a single TreeNet model (ρ =. 01; T = 3000; ρT = 30).
  - In this range all models had almost identical results on test data.
  - The scores were highly correlated ($r \geq .97$).
  - Within this range, a higher ρT was the most important factor.
  - For models with ρT = 6, the smaller the learning rate the better.

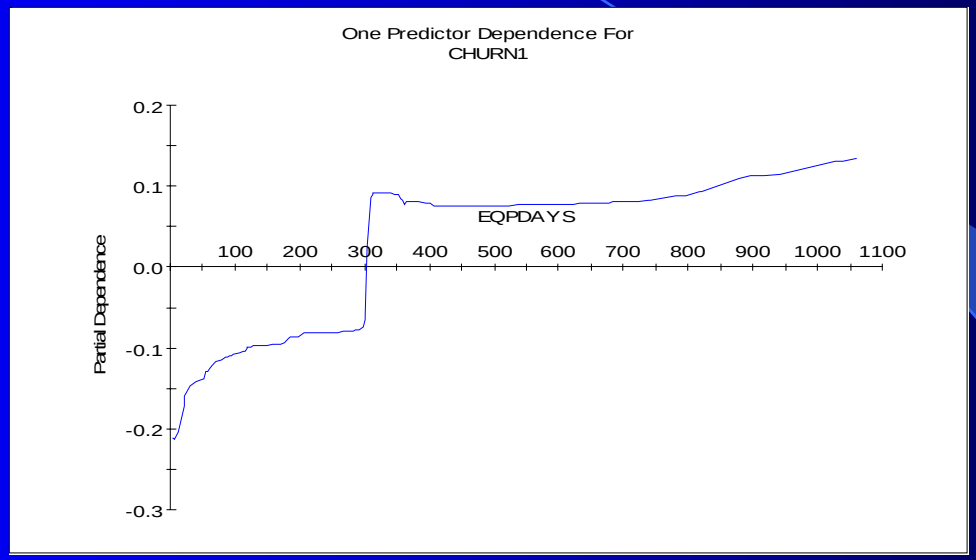
TreeNet Results Screen

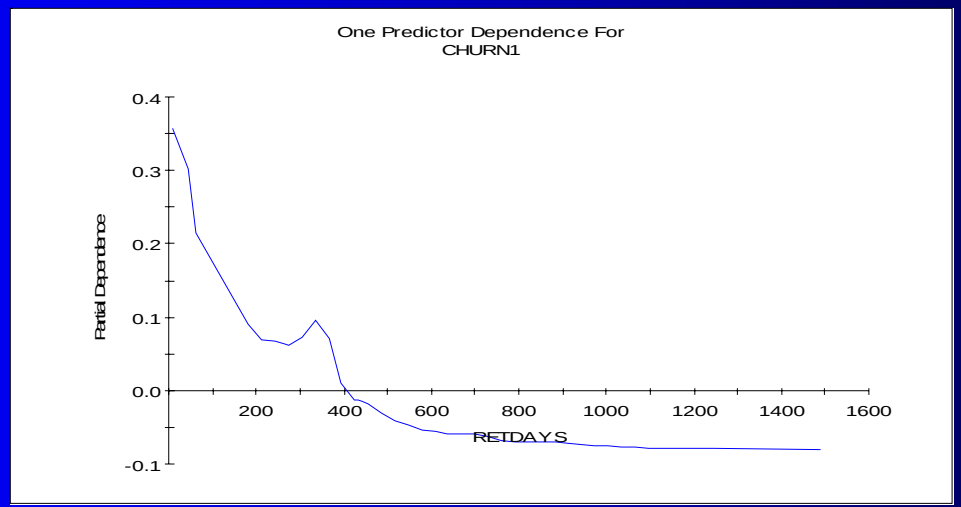


Model Results:
Variable Importance

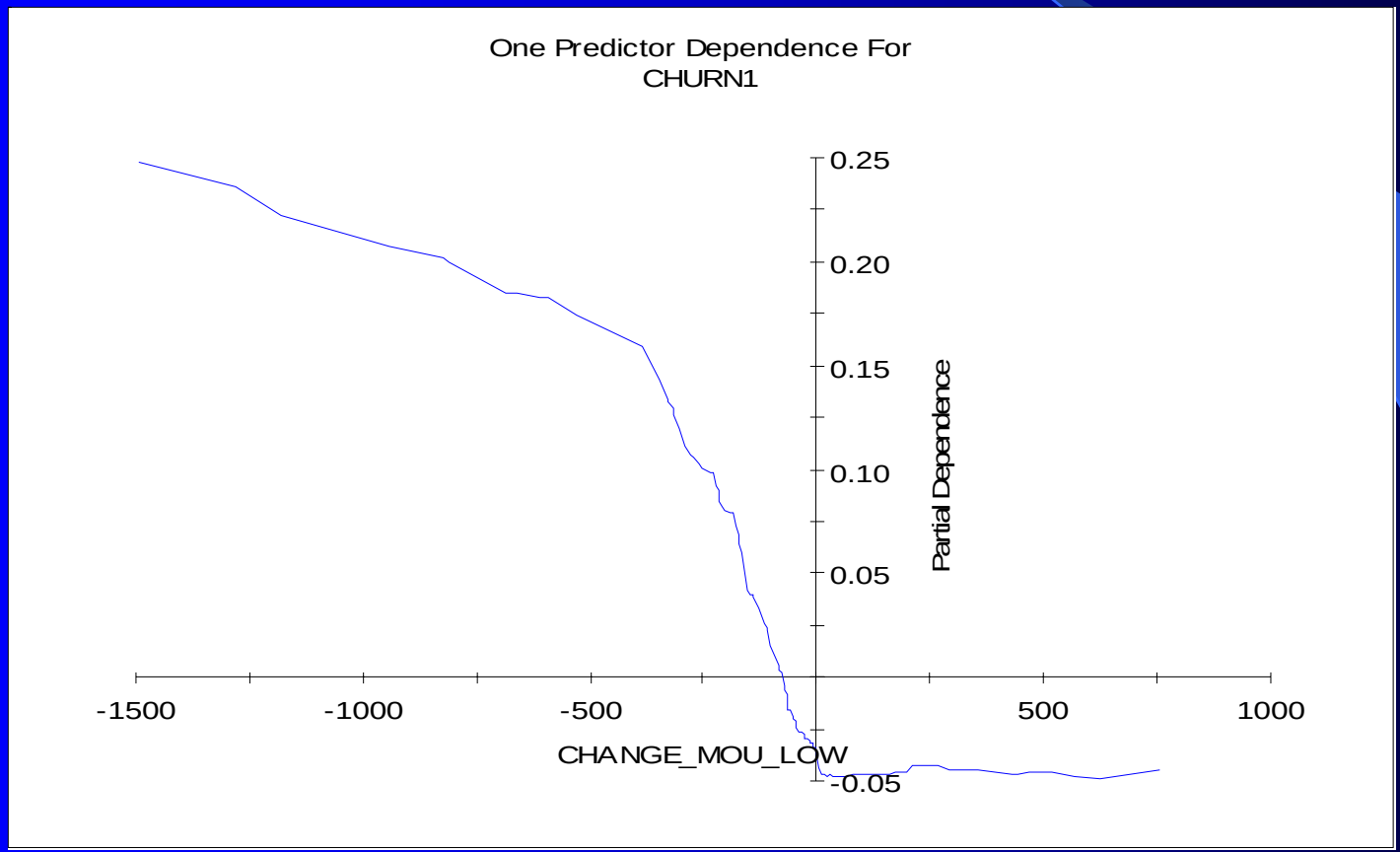
Variable	Description	Score	
CRCLSCOD$	Credit Rating Grade (A-Z)	100.00	|**********|
AREA$	Geographic Locale or Major City (19 levels)	85.67	|********** |
ETHNIC$	Race/Origin (17 Levels	51.91	|****** |
EQPDAYS	Age of current handset	46.39	|****** |
RETDAYS	Days since last retention call	45.95	|****** |
CHANGE_MOU	Recent Change in Monthly Minutes	37.51	|***** |
DWLLSIZE	Number of Households at address	36.10	|***** |
MOU_MEAN	Lifetime average minutes usage	35.40	|**** |
OCCU1$	Occupation (Blue/White, Self) (22 levels)	34.16	|**** |
MONTHS	Length of service to date	33.42	|**** |
TOTMRC_RANGE	Range of monthly recurring charges	31.38	|**** |
CSANODE	CSA condensed to 8 levels (CART nodes)	31.08	|**** |
AVGQTY	Avg monthly calls (lifetime)	26.35	|**** |
MOU_CVCE_MEAN	Avg Monthly Minutes (completed voice)	24.02	|*** |
AVGMOU	Avg Monthly Minutes (lifetime)	23.90	|*** |
HND_PRICE	Hand Set Price	23.43	|*** |

Impact of Hand Set Age and Time Since Retention Call

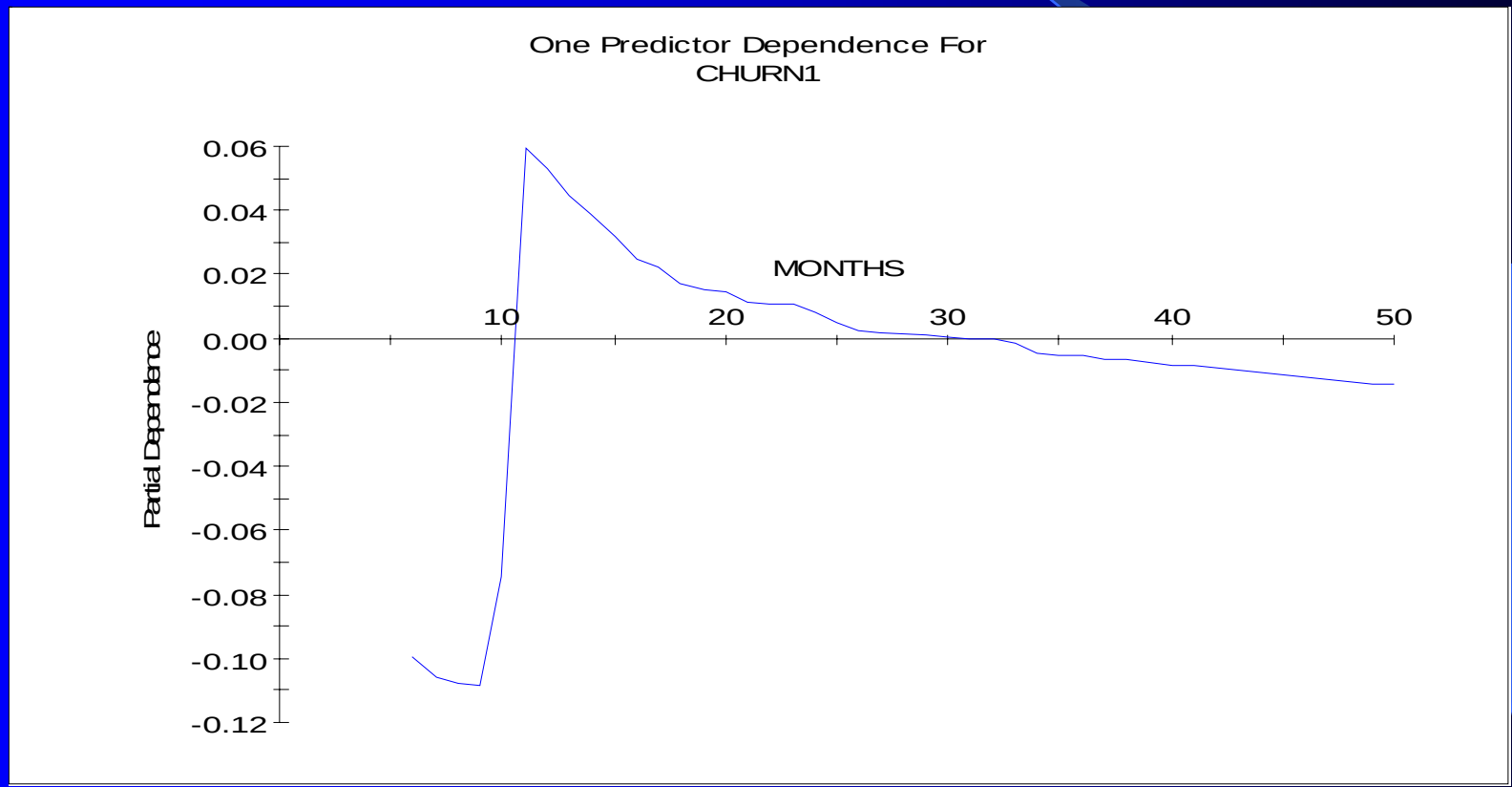




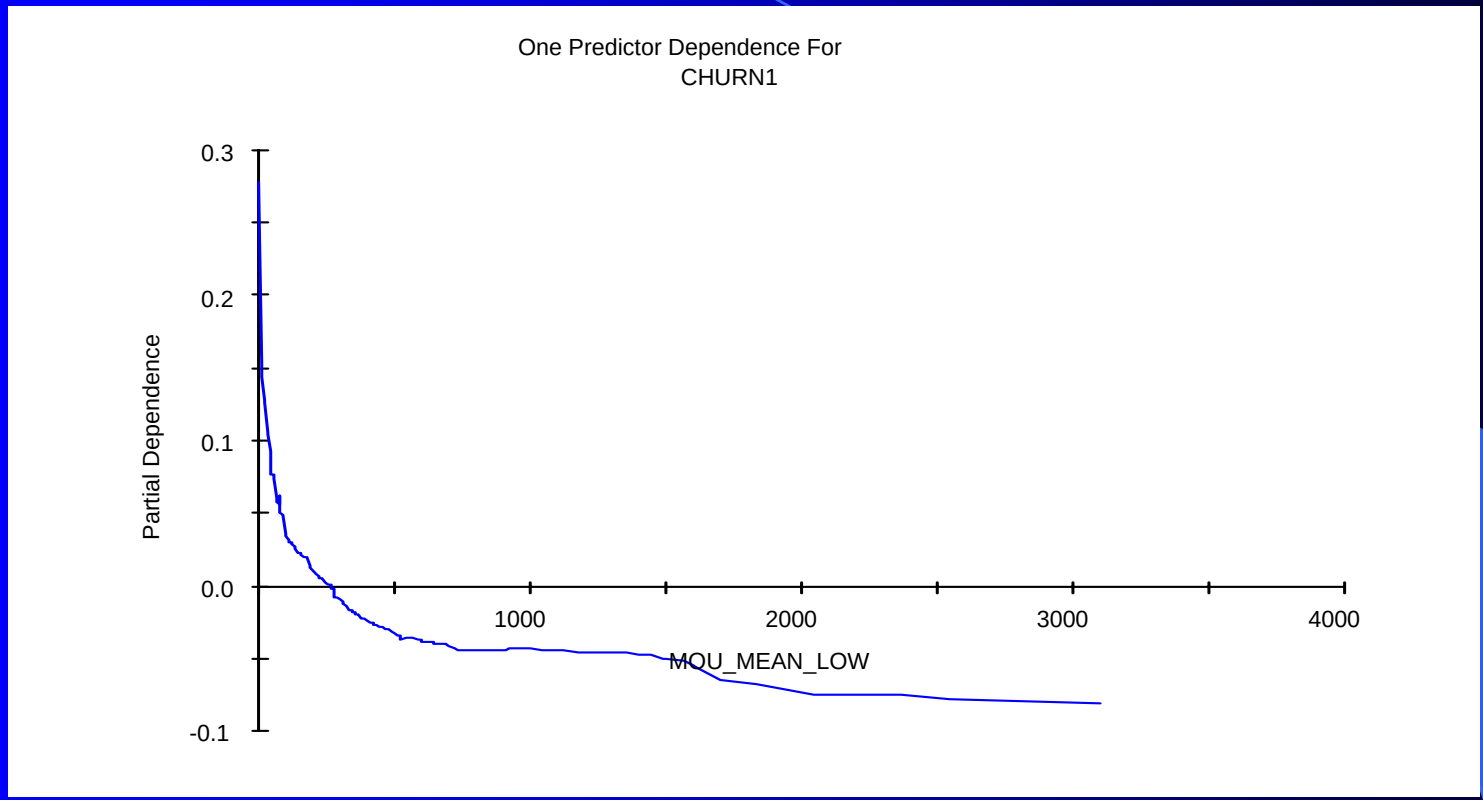
Effect of Recent Change in Minutes Usage



Effect of Tenure: The One-year Spike

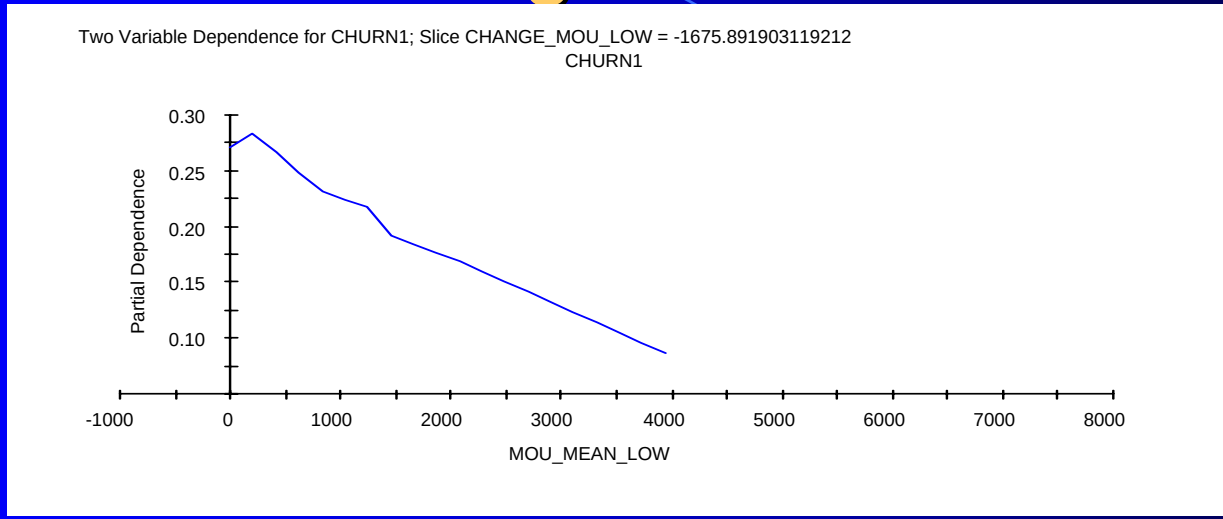


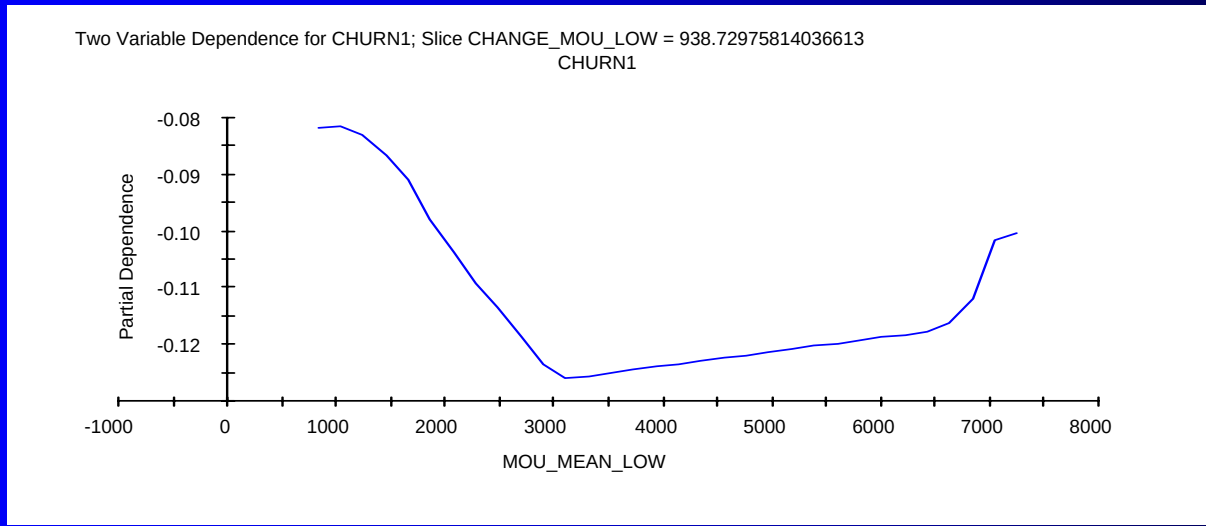
Prob of Churn vs Minutes



Unconditional; other variables varying in typical fashion

Interaction of Minutes and Change in Minutes





References

- Friedman, J.H. (1999). Stochastic gradient boosting. Stanford: Statistics Department, Stanford University.
- Friedman, J.H. (1999). Greedy function approximation: a gradient boosting machine. Stanford: Statistics Department, Stanford University.
- Salford Systems (2002) TreeNet™ 1.0 Stochastic Gradient Boosting. San Diego, CA.
- Steinberg, D., Cardell, N.S., and Golovnya, M. (2003) Stochastic Gradient Boosting and Restrained Learning. Salford Systems discussion paper.