

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ & ΔΙΟΙΚΗΣΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
Τομέας: Οργάνωσης και Διοίκησης

Συνθετική Μεταπτυχιακή Διατριβή
**“Διαδικασίες Εξόρυξης Δεδομένων για την Πρόβλεψη και
Ανάλυση της Απώλειας Πελατών”**

Διατριβή που υπεβλήθη για την μερική ικανοποίηση των
απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης
Υπό την

ΙΩΑΝΝΑ ΑΓΓΕΛΙΔΑΚΗ

Χανιά, Μάιος 2010

Ευχαριστώ το δάσκαλο και καθηγητή μου,
Δούμπο Μιχάλη, για την καθοδήγησή του.

Περιεχόμενα

ΕΙΣΑΓΩΓΗ.....	5
1. ΑΝΑΛΥΣΗ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ	7
1.1. ΔΙΑΧΕΙΡΙΣΗ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ	8
1.1.1. Παράμετροι διαχείρισης απώλειας πελατών.....	9
1.1.2. Παράγοντες που υποδηλώνουν την πρόθεση ενός πελάτη να αποχωρήσει από μία εταιρία	11
1.1.3. Αιτίες απώλειας πελατών και ανακάλυψή τους.....	11
1.1.4. Διατήρηση πελατών.....	14
1.1.5. Είναι όλοι οι πελάτες συμφέρον να διατηρηθούν;	16
1.2. ΔΙΑΧΕΙΡΙΣΗ ΣΧΕΣΕΩΝ ΠΕΛΑΤΩΝ	19
1.3. ΑΙΤΙΕΣ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ	20
1.3.1. Βιβλιογραφική ανασκόπηση	20
1.3.2. Διαχωρισμός πελατών με βάση τη συμπεριφορά	22
1.3.2.1. Αφοσιωμένοι πελάτες	22
1.3.2.2. Χαμένοι πελάτες	23
1.3.3. Παράγοντες απώλειας πελατών.....	25
1.3.3.1. Δυσaréσκεια πελατών.....	26
1.3.3.2. Κόστος αλλαγής εταιρίας.....	26
1.3.3.3. Χρήση υπηρεσιών	27
1.3.3.4. Μεταβλητές σχετικές με τον πελάτη	27
1.3.3.5. Κατάσταση πελατών	28
1.3.4. Σύνοψη	29
2. ΜΕΘΟΔΟΛΟΓΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ	30
2.1. ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....	30
2.1.1. Πλεονεκτήματα	33
2.2. ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ	33
2.2.1. Τύποι δέντρων απόφασης.....	34
2.2.2. Πλεονεκτήματα	35
2.2.3. Περιορισμοί	35
2.3. ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ.....	36
2.3.1. Πλεονεκτήματα & Μειονεκτήματα	39
2.4. ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ.....	41
2.4.1. Προσεγγιστικές μηχανές διανυσμάτων υποστήριξης.....	41
2.4.2. Προσθετικές μηχανές διανυσμάτων υποστήριξης	44
2.4.3. Πλεονεκτήματα	46
2.5. ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ ΜΟΝΤΕΛΩΝ	47

2.5.1.	<i>ROC Curve</i>	47
2.5.2.	<i>Καμπύλη Top Decile Lift</i>	48
2.6.	ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΑΠΩΛΕΙΑΣ ΠΕΛΑΤΩΝ	49
2.6.1.	<i>Ανάλυση παλινδρόμησης</i>	50
2.6.2.	<i>Δέντρα απόφασης</i>	50
2.6.3.	<i>Τεχνητά νευρωνικά δίκτυα</i>	51
2.6.4.	<i>Μηχανές υποστήριξης διανυσμάτων</i>	52
2.6.5.	<i>Γενετικοί αλγόριθμοι</i>	52
2.6.6.	<i>Άλλες μέθοδοι</i>	53
2.6.7.	<i>Σύνοψη</i>	55
3.	ΔΕΔΟΜΕΝΑ, ΜΕΘΟΔΟΛΟΓΙΑ ΑΝΑΛΥΣΗΣ, ΑΠΟΤΕΛΕΣΜΑΤΑ	57
3.1.	ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ.....	57
3.2.	ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΚΑΙ ΚΩΔΙΚΟΠΟΙΗΣΗ	60
3.2.1.	<i>Μονομεταβλητή ανάλυση</i>	65
3.2.2.	<i>Πολυμεταβλητή ανάλυση</i>	66
3.2.3.	<i>Μετασχηματισμός ποιοτικών μεταβλητών</i>	68
3.3.	ΕΦΑΡΜΟΓΕΣ ΜΕΘΟΔΩΝ	70
3.3.1.	<i>Αποτελέσματα μεθόδων</i>	71
3.3.2.	<i>Συνδυασμός μοντέλων</i>	75
4.	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΕΠΕΚΤΑΣΕΙΣ	79
5.	ΒΙΒΛΙΟΓΡΑΦΙΑ	82
6.	ΠΑΡΑΡΤΗΜΑ Α: ΑΡΧΙΚΟ ΣΥΝΟΛΟ ΜΕΤΑΒΛΗΤΩΝ ΚΑΙ ΠΕΡΙΓΡΑΦΗ	
ΤΟΥΣ		92

Εισαγωγή

Στην τελευταία δεκαετία η απώλεια πελατών (customer churn), η τάση τους δηλαδή να μετακινούνται από μία εταιρία σε μία άλλη, αποτελεί μία από τις κυριότερες προκλήσεις που οι εταιρίες παγκοσμίως πρέπει να αντιμετωπίσουν προκειμένου να επιβιώσουν σε μία όλο και περισσότερο ανταγωνιστική αγορά. Επιπλέον, μέσα από έρευνες στον τομέα του μάρκετινγκ έχει αποδειχτεί ότι είναι ακριβότερο για μια εταιρία να προσελκύσει ένα νέο πελάτη, από ότι είναι να διατηρήσει έναν ήδη υπάρχοντα. Προς το σκοπό αυτό, πολλές επιχειρήσεις στρέφονται στην ανάλυση απώλειας πελατών με σκοπό τη διατήρηση και αύξηση των κερδών τους μέσα από την ενίσχυση των σχέσεων με τους πελάτες τους και τη βελτίωση της ικανοποίησής τους.

Τα συστήματα Διαχείρισης Σχέσεων Πελατών (Customer Relationship Management-CRM) έχουν αντικαταστήσει τις παραδοσιακές στρατηγικές μαζικού μάρκετινγκ με επιλεκτικές πρακτικές. Οι επιλεκτικές αυτές πρακτικές marketing βασίζονται στη χρήση τεχνικών εξόρυξης δεδομένων (data mining) για την αναγνώριση των πελατών εκείνων οι οποίοι είναι πιο πιθανόν να σταματήσουν να χρησιμοποιούν τα προϊόντα ή υπηρεσίες μιας εταιρίας με σκοπό την πρόβλεψη μελλοντικών απωλειών. Ο έγκαιρος εντοπισμός τέτοιων πελατών, δίνει την ευκαιρία στις εταιρίες να τους στοχεύσουν πιο δραστικά, για να αποτρέπουν την απώλειά τους.

Η παρούσα διατριβή αφορά στην ανάλυση και πρόβλεψη της απώλειας πελατών με χρήση στατιστικών τεχνικών, τεχνικών εξόρυξης δεδομένων, επιχειρηματικής και τεχνητής νοημοσύνης. Η εργασία βασίζεται σε ένα μεγάλο σύνολο δεδομένων ενός τηλεπικοινωνιακού πάροχου (περίπου 200.000 πελάτες), που περιλαμβάνει δημογραφικές πληροφορίες, πληροφορίες για το χρόνο χρήσης, το εισόδημα, την τηλεφωνική συσκευή κ.α. Στόχος της ανάλυσης είναι η ανάπτυξη αξιόπιστων μοντέλων για την πρόβλεψη της απώλειας πελατών καθώς και την ανάλυση των παραγόντων που σχετίζονται με την απώλεια πελατών.

Αναλυτικότερα, το πρώτο κεφάλαιο της εργασίας είναι μία θεωρητική προσέγγιση του θέματος, όπου παρουσιάζονται βασικές έννοιες και ορισμοί που έχουν να κάνουν με την ανάλυση και διαχείριση της απώλειας πελατών. Στην ίδια ενότητα ερευνώνται οι αιτίες που προκαλούν την απώλεια πελατών, και γίνεται

βιβλιογραφική ανασκόπηση του θέματος. Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο των μεθόδων και των κριτηρίων αξιολόγησης των μοντέλων που εφαρμόστηκαν στην εργασία. Επιπλέον, γίνεται βιβλιογραφική ανασκόπηση και άλλων δημοφιλών μεθόδων και εφαρμογών τους στον τομέα της ανάλυσης και πρόβλεψης απώλειας πελατών. Στο τρίτο κεφάλαιο περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν στην ανάλυση, η διαδικασία ανάλυσης και επεξεργασίας των δεδομένων, η εφαρμογή των μεθόδων και αναλύονται τα αποτελέσματα. Τέλος, στο τέταρτο κεφάλαιο παρουσιάζονται τα συμπεράσματα και ορισμένες μελλοντικές επεκτάσεις.

1. Ανάλυση απώλειας πελατών

Η απώλεια πελατών (customer churn) ορίζεται ως η κλίση των πελατών να διακόπτουν τις σχέσεις και συναλλαγές τους με μία εταιρία σε κάποια χρονική περίοδο και αποτελεί μία από τις κύριες προκλήσεις που οι επιχειρήσεις σήμερα καλούνται να αντιμετωπίσουν (Chandar et al. 2006). Οι επιχειρήσεις στις οποίες γίνεται αναφορά ανήκουν σε διάφορους κλάδους, όπως εκδόσεις, επενδυτικές υπηρεσίες, ασφάλειες, τραπεζικές υπηρεσίες και άλλους. Για παράδειγμα, στην παροχή υπηρεσιών διαδικτύου, τα ετήσια ποσοστά απώλειας πελατών κυμαίνονται από 21% σε 63.2%. Στην ασύρματη τηλεφωνία, τα αντίστοιχα ποσοστά κυμαίνονται από 23.4% (Wireless Review, 2000) έως 46% (Telephony Online, 2002). Τα ποσοστά αυτά μπορεί να ποικίλλουν ανάλογα με την εταιρία και τον κλάδο, αλλά όλες όμως οι επιχειρήσεις αυτές χάνουν τουλάχιστον ένα τέταρτο από τους πελάτες τους από τον ένα χρόνο στον άλλο λόγω αυτού του φαινομένου.

Στα πλαίσια αυτά συνεπώς, και λόγω του έντονου ανταγωνισμού που παρουσιάζει, ο χώρος των τηλεπικοινωνιών αντιμετωπίζει σε μεγάλο εύρος το συγκεκριμένο φαινόμενο. Ο κλάδος παρουσιάζει δυναμικό χαρακτήρα, με καινούριες υπηρεσίες, αναδυόμενες τεχνολογίες, και νέες εταιρίες οι οποίες συνεχώς αλλάζουν το επιχειρηματικό τοπίο. Οι επιχειρήσεις, προκειμένου να υπερτερούν έναντι των ανταγωνιστών τους, κάνουν προσφορές στους πελάτες σε εβδομαδιαία βάση, ελπίζοντας να κρατήσουν τους ήδη υπάρχοντες και να δελεάσουν νέους συνδρομητές. Αν και υπάρχει σημαντικός χώρος για περεταίρω ανάπτυξη στις περισσότερες αγορές, ο δείκτης ανάπτυξης του κλάδου μειώνεται, και ο ανταγωνισμός αυξάνεται. Συνεπώς, ο έλεγχος της απώλειας πελατών αποτελεί σημαντικό παράγοντα στις τηλεπικοινωνιακές επιχειρήσεις.

Για καλύτερη κατανόηση του προβλήματος, ας ληφθεί υπόψη ότι το μηνιαίο ποσοστό απώλειας πελατών σε έναν τηλεπικοινωνιακό πάροχο στην Ευρώπη ή την Βόρεια Αμερική το 1998 ήταν στα επίπεδα του 2-3% της πελατειακής βάσης της εταιρίας. Λαμβάνοντας υπόψη ότι το μέσο κόστος απόκτησης ενός νέου πελάτη ήταν περίπου \$400, η απώλεια πελατών κόστιζε στη βιομηχανία σχεδόν \$6.3 δισεκατομμύρια. Επιπλέον, η συνολική ετήσια απώλεια έφτανε τα \$9.6 δισεκατομμύρια λαμβάνοντας υπόψη τα χαμένα μηνιαία έσοδα από τις ακυρώσεις

των συνδρομητών, και αυτό 11 χρόνια πριν, όταν τα επίπεδα του ανταγωνισμού στην αγορά ήταν πολύ χαμηλότερα (Mozer, 2000). Σήμερα, είναι γνωστό στον τομέα του μάρκετινγκ ότι κοστίζει περίπου 5 φορές περισσότερο σε μία εταιρία να κερδίσει ένα νέο πελάτη, από το να διατηρήσει έναν ήδη υπάρχοντα (Siber, 1997). Για μία εταιρία λοιπόν με 1.5 εκατομμύρια συνδρομητές, η μείωση του μηνιαίου ποσοστού απώλειας πελατών από 2% σε 1% θα έφερνε αύξηση στα ετήσια κέρδη του επιπέδου των \$54 εκατομμυρίων και αύξηση αξίας των μετόχων του επιπέδου των \$150 εκατομμυρίων. Επιπλέον, έχει αποδειχτεί ότι, αν μία εταιρία δεν αποκτήσει νέους πελάτες, ο μέσος χρόνος ζωής ενός ήδη υπάρχοντος πελάτη ισούται με $1/c$, όπου c το ετήσιο ποσοστό απώλειας πελατών. Για μία εταιρία με 25% ποσοστό, ο μέσος χρόνος ζωής ενός πελάτη είναι 4 χρόνια, ενώ για μία εταιρία με 50% ποσοστό απώλειας πελατών, ο μέσος χρόνος ζωής ενός πελάτη είναι 2 χρόνια (Neslin et al. 2004). Είναι φανερό λοιπόν η σημαντικότητα της διαχείρισης απώλειας πελατών (**churn management**) προκειμένου οι επιχειρήσεις να σταθεροποιήσουν τη θέση τους στην αγορά και να διατηρήσουν ή και να αυξήσουν την αποδοτικότητα και κερδοφορία τους.

1.1. Διαχείριση απώλειας πελατών

Από την οπτική του τμήματος μάρκετινγκ μιας επιχείρησης, υπάρχουν δύο βασικές προσεγγίσεις στη διαχείριση της απώλειας πελατών (**churn management**), οι στοχευμένες (targeted) και οι μη στοχευμένες (untargeted) (Burez et al. 2007). Οι **μη στοχευμένες προσεγγίσεις** βασίζονται στη μαζική διαφήμιση για να αυξήσουν την αφοσίωση στη μάρκα και να διατηρήσουν πελάτες. Οι **στοχευμένες προσεγγίσεις** βασίζονται στην αναγνώριση πελατών οι οποίοι είναι πιθανότερο να φύγουν από την επιχείρηση, και μετά είτε τους παρέχεται ένα άμεσο κίνητρο, είτε προσαρμόζεται στα μέτρα τους ένα πλάνο εξυπηρέτησης προκειμένου να αποφευχθεί η φυγή τους. Επιπλέον, υπάρχουν δύο είδη στοχευμένων προσεγγίσεων: reactive και proactive. Με την **reactive προσέγγιση**, η εταιρία περιμένει μέχρις ότου ο πελάτης επικοινωνήσει για να διακόψει τη συνδρομή του, και στη συνέχεια του προσφέρει ένα κίνητρο, π.χ. μία έκπτωση, για να τον πείσει να μείνει. Με την **proactive προσέγγιση**, η εταιρία προσπαθεί να εντοπίσει πελάτες οι οποίοι είναι πιθανό να αφήσουν την επιχείρηση σε μελλοντική χρονική περίοδο. Η εταιρία στη συνέχεια στοχεύει αυτούς τους πελάτες με ειδικά προγράμματα ή προσφορές για να προλάβει την απώλειά τους. Τα στοχευμένα proactive προγράμματα έχουν το πλεονέκτημα να

διαθέτουν λιγότερο δαπανηρά κίνητρα, αφού τα κίνητρα αυτά δεν είναι τόσο δραστικά, και άρα δαπανηρά, όσο θα ήταν στην περίπτωση που ο πελάτης έχει ήδη πάρει την απόφαση να φύγει και άρα θα έπρεπε να δελεαστεί την τελευταία στιγμή προκειμένου να μείνει. Παρόλα αυτά, τα συγκεκριμένα συστήματα μπορεί να είναι ανούσια εάν οι προβλέψεις απώλειας πελατών είναι ανακριβείς και λανθασμένες, επειδή με αυτό τον τρόπο οι εταιρίες θα σπαταλούσαν χρήματα για να κρατήσουν πελάτες οι οποίοι δεν είχαν σκοπό να εγκαταλείψουν. Συνεπώς, σκοπός είναι η υψηλή ακρίβεια στην πρόβλεψη της απώλειας πελατών (Lariviere και Van den Poel, 2004).

1.1.1. Παράμετροι διαχείρισης απώλειας πελατών

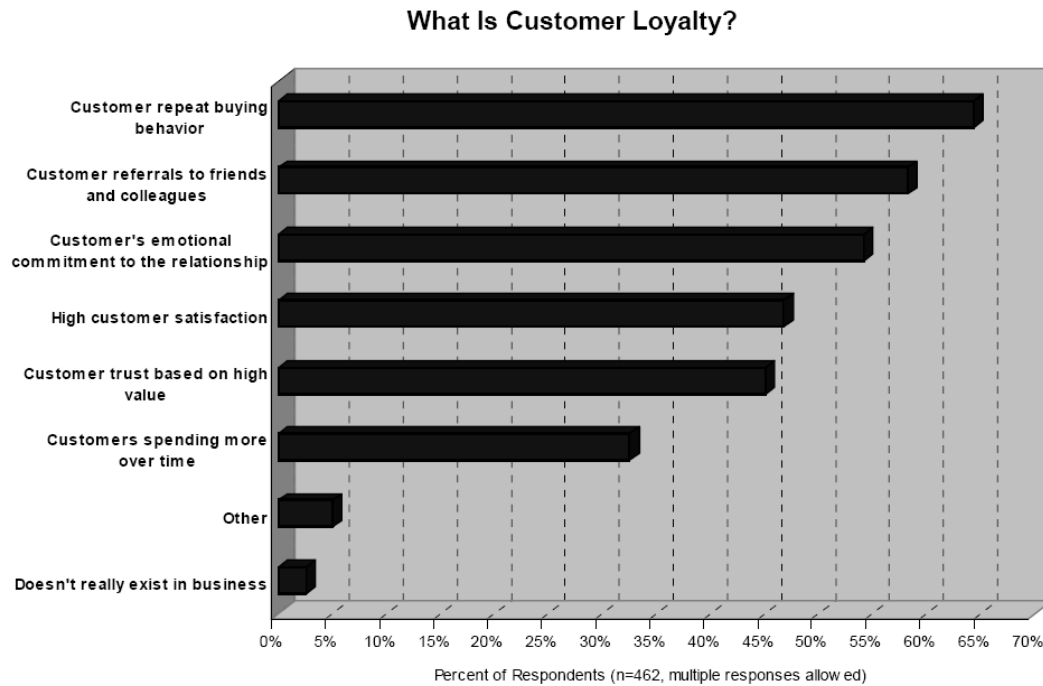
Μέσα από τις δύο προσεγγίσεις που παρουσιάστηκαν προκύπτουν εύλογοι προβληματισμοί, οι οποίοι και αποτελούν παραμέτρους που πρέπει να ληφθούν υπόψη κατά τη διαχείριση απώλειας πελατών: Γνωρίζει τελικά μία εταιρία ότι ένας πελάτης πρόκειται να την εγκαταλείψει; Αν όχι, μπορεί να μάθει τη στάση που έχουν οι πελάτες απέναντί της; Είναι όλοι οι πελάτες συμφέρον να διατηρηθούν; Και πώς μπορεί να διατηρήσει τους πελάτες της;

Πριν απαντηθούν οι προβληματισμοί αυτοί, καλό είναι να γίνει αναφορά σε μία έννοια που συνδέεται στενά με την απώλεια πελατών και στην οποία θα γίνεται εκτενώς αναφορά παρακάτω, την αφοσίωση (**customer loyalty**). Οι αφοσιωμένοι πελάτες παραμένουν σε μία εταιρία περισσότερο και εμφανίζουν πρόθεση επαναγοράς (Dick et al. 1994). Αυτή η συμπεριφορά οφείλεται στη θετική στάση και επιθυμία τους να συνεχίσουν να έχουν επιχειρηματικές σχέσεις με μια εταιρία, η οποία κατά συνέπεια αποκτά σημαντικό πλεονέκτημα στην αύξηση εσόδων (Thomson 2005, Kim et al. 2004). Για να γίνει πιο αντιληπτή η σχέση μεταξύ αφοσίωσης και απώλειας πελατών αρκεί να ληφθεί υπόψη ότι ο όρος απώλεια πελατών περιγράφει τον αριθμό ή ποσοστό των πελατών που σταματούν τις σχέσεις τους με έναν παροχέα υπηρεσιών, και άρα ισχύει μια σχέση της μορφής (Hwang et al. 2004 και Kim, 2006):

$$\text{Αφοσίωση πελατών} = 1 - \text{ποσοστό απώλειας πελατών}$$

Έρευνα που πραγματοποιήθηκε σε πρόσωπα του επιχειρηματικού τομέα το Δεκέμβριο του 2004, στα πλαίσια της διαδικτυακής κοινότητας επιχειρηματιών CRMGuru (www.customerthink.com) έδειξε τα παρακάτω αποτελέσματα όσον αφορά στην έννοια της αφοσίωσης (σχήμα 1): Το 64% των ερωτηθέντων όρισε την

αφοσίωση ως συμπεριφορά πρόθεσης επαναγοράς, το 58% ως συμπεριφορά καταναλωτών που προτείνουν μια εταιρία σε φίλους και γνωστούς και το 54% ως τη συναισθηματική δέσμευση του πελάτη σε μία εταιρία. Μόνο το 32% των ερωτηθέντων όρισε την αφοσίωση ως τη συμπεριφορά του πελάτη εκείνου που ξοδεύει όλο και περισσότερο μέσα στο χρόνο.



Σχήμα 1: Αποτελέσματα έρευνας για τον ορισμό της αφοσίωσης πελατών (Thompson, 2005)

Η αφοσίωση, όπως χαρακτηρίστηκε από τον Reichheld (2001), είναι μία γεννήτρια κερδών.

Όπως έχουν γίνει προσπάθειες ορισμού της αφοσίωσης, έτσι έχουν γίνει και προσπάθειες μέτρησής της. Η διαδικτυακή συμβουλευτική εταιρία Walker Information (www.walkerinformation.com) για παράδειγμα ανέπτυξε ένα δείκτη που συνδυάζει την εμπειρία του πελάτη με την πρόθεση του να παραμείνει σε μία εταιρία. Ερευνήθηκε ο βαθμός με τον οποίο οι πάροχοι ικανοποιούσαν τις προσδοκίες των πελατών, ο βαθμός με τον οποίο ο πελάτης σκόπευε να συνεχίσει τη σχέση του με μία εταιρία, και το μερίδιο που ένας πελάτης ήταν διατεθειμένος να δώσει στην εταιρία από το πορτοφόλι του. Ο δείκτης αυτός επέτρεψε στην εταιρία να ορίσει μια κλίμακα αφοσίωσης του πελάτη.

1.1.2. Παράγοντες που υποδηλώνουν την πρόθεση ενός πελάτη να αποχωρήσει από μία εταιρία

Στη σημερινή ανταγωνιστική αγορά δεν είναι εύκολο να γίνει αντιληπτή η πρόθεση ενός πελάτη να εγκαταλείψει μια εταιρία. Αυτό συμβαίνει αφενός επειδή μόνο ένα μικρό ποσοστό των πελατών εκφράζει τα παράπονά του. Μία τυπική για παράδειγμα επιχείρηση ακούει μόνο το 4% των δυσαρεστημένων πελατών της, ενώ το υπόλοιπο 96% απλά φεύγει (Γρηγορούδης, 2003). Επίσης ο τομέας στον οποίο ανήκει η επιχείρηση και οι σχέσεις που διατηρεί με τους πελάτες της επηρεάζει το πως εντοπίζονται οι πελάτες που επιθυμούν να φύγουν. Για παράδειγμα στο χώρο των πιστωτικών καρτών, οι πελάτες μπορούν εύκολα να αρχίσουν να χρησιμοποιούν μια άλλη πιστωτική κάρτα και ο μόνος δείκτης για την πρώτη πιστωτική εταιρία είναι οι μειούμενες συναλλαγές. Αντίθετα, στον κλάδο της ασύρματης τηλεφωνίας, ένας πελάτης μπορεί να αλλάξει εταιρία και να κρατήσει το ίδιο νούμερο τηλεφώνου, οπότε σε αυτή την περίπτωση η πρώτη εταιρία θα καταλάβει τι συμβαίνει μόνο τη στιγμή που ο πελάτης θα αποσυνδέσει το λογαριασμό του (Thomson, 2005).

Για να αντιμετωπιστεί αυτή η κατάσταση έχουν γίνει πολλές έρευνες εντοπισμού πράξεων που μπορεί να θεωρηθούν δείκτες. Ενδεικτικά αναφέρεται μία τέτοια έρευνα (Thompson, 2005) σε 200 εταιρίες για την αφοσίωση πελατών, η οποία κατέληξε στην ακόλουθη λίστα των 6 δεικτών που δείχνουν ότι ένας πελάτης είναι έτοιμος να εγκαταλείψει:

- Η αποδοχή των προσφορών από τον πελάτη καθυστερεί.
- Η ροή των δεδομένων του πελάτη επιβραδύνεται.
- Τα μελλοντικά σχέδια γίνονται σταδιακά βραχυπρόθεσμα.
- Ένα ή περισσότερα προϊόντα ή υπηρεσίες διακόπτονται.
- Ο όγκος των συναλλαγών του πελάτη με την εταιρία μειώνεται.

1.1.3. Αιτίες απώλειας πελατών και ανακάλυψή τους

Ανεξάρτητα με το μέγεθος μιας εταιρίας, πρέπει να γνωρίζει γιατί οι πελάτες της μένουν ή φεύγουν. Αυτή η οπτική προσανατολισμένη στον πελάτη θα διευκολύνει τη διαχείριση της συμπεριφοράς των πελατών κατά τρόπο ώστε τελικά να δημιουργηθεί μια στρατηγική αφοσίωσης που θα αποφέρει κέρδη. Εδώ παρατίθενται ορισμένοι οδηγοί αφοσίωσης, δηλαδή παράγοντες που έχουν μεγάλη

συσχέτιση με τη συμπεριφορά αφοσίωσης, οι οποίοι σκοπό έχουν να δείξουν που μπορεί να οφείλεται η τάση απώλειας πελατών (Thompson, 2005):

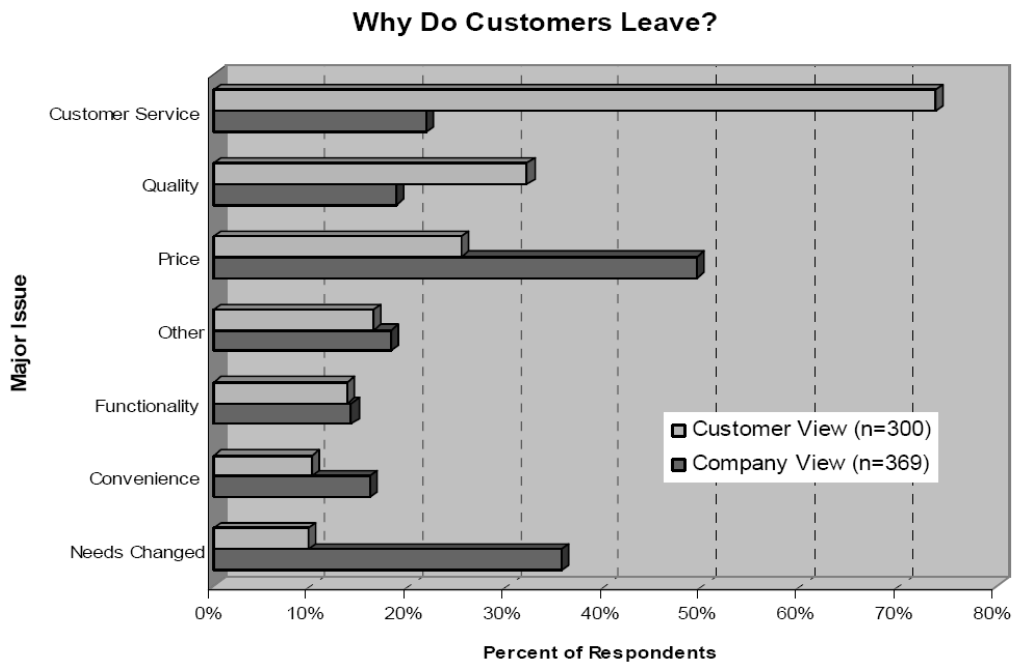
- Στάση-για την επωνυμία, την ποιότητα και την εστίαση στον πελάτη
- Εικόνα της επωνυμίας-ως ηγέτης βιομηχανίας, αξιόπιστη εταιρία και καινοτόμο πάροχο
- Εμπειρία πελάτη-με την ποιότητα του προϊόντος, τη διαδικασία αγοράς και την τεχνική υποστήριξη

Αξίζει να σημειωθεί ότι έρευνες πάνω στο θέμα έδειξαν ότι η τιμή συσχετίζεται με την αφοσίωση υπό τη μορφή του πως οι πελάτες αντιλαμβάνονται την αξία του προϊόντος που αγόρασαν/υπηρεσίας που χρησιμοποίησαν, δηλαδή υπό τη μορφή αντιλαμβανόμενης αξίας. Από μόνη της η τιμή ενός προϊόντος/υπηρεσίας δε συνεισφέρει στην αφοσίωση του πελάτη (Thompson, 2005).

Εξαιρουμένων των μονοπωλίων, οι σημερινοί πελάτες μπορούν να αποκτήσουν παρόμοια προϊόντα/υπηρεσίες σε παρόμοιες τιμές, από διάφορους παρόχους, και στο παρόν σενάριο αγοράς δεν υπάρχουν πολλές διαφοροποιήσεις στις προσφορές, τις τιμές και τις υπηρεσίες που προσφέρονται λόγω του ισχυρού ανταγωνισμού. Συνεπώς οι υπηρεσίες εξυπηρέτησης πελατών και οι υπηρεσίες προστιθέμενης αξίας (value-added) λαμβάνουν μεγάλη σημασία στην αντίληψη του πελάτη. Αν μια επιχείρηση δεν προβλέπει αγοραστικές ανάγκες ή δεν παρέχει υπηρεσίες προστιθέμενης αξίας, οι οποίες όμως προσφέρονται από ανταγωνιστές, τότε ο πελάτης είναι πολύ πιθανό να προτιμήσει μια ανταγωνιστική εταιρία. Με βάση την παρούσα κατάσταση της αγοράς λοιπόν, έχει αποδειχτεί ότι το συναισθημα είναι αυτό που διαδραματίζει ρόλο μεγαλύτερο ακόμα και από την τιμή ή την ποιότητα στην απόφαση διακοπής σχέσεων με μία εταιρία, και η φτωχή εξυπηρέτηση πελατών είναι ο κύριος ένοχος για τη δημιουργία αρνητικών συναισθημάτων. Πρακτικά, οι περισσότεροι άνθρωποι φεύγουν από μία εταιρία επειδή νιώθουν ότι δεν τους έχουν συμπεριφερθεί σωστά, σύμφωνα με τον Thompson (2003), η διοίκηση όμως συνήθως πιστεύει ότι ο λόγος που φεύγουν είναι η τιμή (και φυσικά αυτό αληθεύει σε κάποιες περιπτώσεις). Οι αιτίες απώλειας πελατών θα μελετηθούν εκτενέστερα παρακάτω (ενότητα 1.3), όπως αυτές συγκεντρώθηκαν από τη σχετική βιβλιογραφία.

Έρευνα που πραγματοποιήθηκε το Νοέμβριο του 2004, στα πλαίσια της διαδικτυακής κοινότητας επιχειρηματιών CRMGuru (www.customerthink.com) έδειξε (σχήμα 2): Από τους ερωτηθέντες πελάτες που δήλωσαν ότι σταμάτησαν να

χρησιμοποιούν ένα προϊόν/υπηρεσία, το 74% έδωσε την εξυπηρέτηση πελατών ως το σημαντικότερο παράγοντα για αυτή την απόφαση. Το δεύτερο πιο συχνό πρόβλημα-η ποιότητα-επιλέχτηκε από το 32% των ερωτηθέντων, και το 25% δήλωσε ως λόγο την τιμή. Η λειτουργικότητα του προϊόντος/υπηρεσίας επιλέχτηκε ως σημαντικός παράγοντας από το 14% των ερωτηθέντων. Η αντίστοιχη έρευνα που έγινε σε διευθυντικά στελέχη επιχειρήσεων έδειξε ένα σημαντικό χάσμα. Τα ανώτερα στελέχη έδωσαν την τιμή ως τον πιο κρίσιμο παράγοντα στην απώλεια πελατών με 49%, ακολουθεί η αλλαγή των αναγκών του πελάτη με 36% και η εξυπηρέτηση πελατών επιλέχτηκε μόλις από το 22% των ερωτηθέντων.



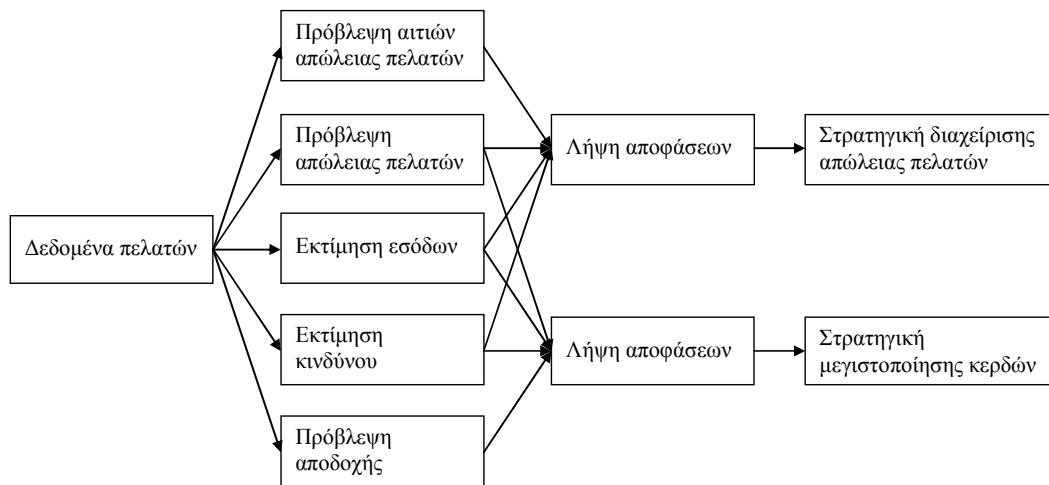
Σχήμα 2: Αποτελέσματα έρευνας για τους λόγους απώλειας πελατών (Thompson, 2005)

Και ενώ κάθε επιχείρηση είναι διαφορετική, αυτή η μελέτη υποδεικνύει ότι οι υπεύθυνοι πρέπει να εξετάζουν προσεκτικά τις υποθέσεις τους και να κατανοούν τους πραγματικούς λόγους για τους οποίους οι πελάτες μένουν ή φεύγουν.

Η ανακάλυψη του προβλήματος μπορεί να είναι δύσκολη, και εδώ βοηθούν οι έρευνες. Αλλά δεν είναι δυνατόν πάντα να γίνονται έρευνες στους πελάτες, και αν γίνεται αυτό, μερικές φορές είναι πολύ αργά. Εδώ εισάγεται η έννοια του **marketing analytics**, διαδικασία η οποία δίνει μία ακριβή εικόνα της πελατειακής βάσης, και επιτρέπει την εκτίμηση της παρελθοντικής, παρούσας, και μελλοντικής αξίας των πελατών, ενώ είναι μία cost-effective επένδυση για μεσαίες και μεγάλες επιχειρήσεις.

1.1.4. Διατήρηση πελατών

Την τελευταία δεκαετία, πολλές εταιρίες θεωρούν τη διατήρηση του πελάτη ως κεντρικό θέμα στις αποφάσεις μανάτζμεντ που παίρνουν. Η έμφαση στη διατήρηση βασίζεται στην υπόθεση ότι υπάρχει μεγάλη σχέση μεταξύ αυτής και της αποδοτικότητας: οι μακροχρόνιοι πελάτες αγοράζουν περισσότερο και είναι λιγότερο δαπανηρό να εξυπηρετηθούν (Hwang et al. 2006), ενώ η αντικατάσταση υπάρχοντων πελατών από νέους είναι πιο ακριβή και ριψοκίνδυνη στρατηγική, αφού είναι εύκολο να υποθεθεί ότι οι νέοι αυτοί πελάτες είναι πιο επιρρεπείς στο να συνεχίσουν την συμπεριφορά μετακίνησής τους στο κοντινό μέλλον, και άρα χαρακτηρίζονται από υψηλή πιθανότητα απώλειας (Colgate και Danaher, 2000).



Σχήμα 3: Πλαίσιο για τη διατήρηση πελατών και τη μεγιστοποίηση αποδοτικότητας (Mozier et al. 2000)

Όπως φαίνεται στο σχήμα 3, η πρόβλεψη της απώλειας πελατών (**churn prediction**), δηλαδή η εκτίμηση της πιθανότητας που έχει ένας πελάτης να φύγει από μια εταιρία, είναι τμήμα του πλαισίου απόφασης για τη διατήρηση των πελατών και τη μεγιστοποίηση του κέρδους. Τα δεδομένα μπορούν επίσης να βοηθήσουν στην ανάλυση των αιτιών για τους οποίους οι πελάτες αλλάζουν επιχειρήσεις, των κερδών που πραγματοποιούνται σε μία χρονική περίοδο, και της ανταπόκρισης σε μία προσφορά μάρκετινγκ. Οι επιχειρήσεις μπορούν να χρησιμοποιούν αυτούς τους πιθανολογικούς υπολογισμούς και τις εκτιμήσεις εσόδων για να καθορίζουν τη στρατηγική αντιμετώπισης της απώλειας πελατών και τη στρατηγική μεγιστοποίησης

των κερδών, στα πλαίσια ενός συστήματος διαχείρισης σχέσεων πελατών, στο οποίο θα γίνει αναφορά παρακάτω.

Η πρόβλεψη απώλειας πελατών χρησιμοποιεί τεχνικές εξόρυξης δεδομένων (**data mining**) προκειμένου να ερευνήσει τις βάσεις δεδομένων μιας εταιρίας με σκοπό την ανακάλυψη προηγούμενως άγνωστης και χρήσιμης γνώσης, όπως ο εντοπισμός τάσεων και επαναλαμβανόμενων προτύπων, συσχετίσεων μεταξύ στοιχείων κ.α., η οποία θα βοηθήσει στη λήψη αποφάσεων και θα οδηγήσει σε σημαντικά επιχειρηματικά οφέλη (Rygielski et al. 2002). Η εξόρυξη δεδομένων χρησιμοποιεί στατιστικές διαδικασίες και αλγορίθμους μηχανικής εκμάθησης για την κατάρτιση μοντέλων πρόβλεψης της απώλειας πελατών, όπως τα δέντρα απόφασης (decision trees), η λογιστική παλινδρόμηση (logistic regression), τα νευρωνικά δίκτυα (neural networks), οι μηχανές υποστήριξης διανυσμάτων (support vector machines), στα οποία και θα γίνει εκτενή αναφορά σε επόμενο κεφάλαιο. Όταν χρησιμοποιείται μια τεχνική ταξινόμησης οι μεταβλητές εισόδου είναι οι δείκτες που προέρχονται από τα δεδομένα των συνδρομητών, όπως δημογραφικές πληροφορίες, πληροφορίες υπηρεσιών συμβολαίου, ιστορικό συναλλαγών, λεπτομέρειες λογαριασμών κ.α.. Οι έξοδοι απόφασης ορίζουν την κατάσταση του πελάτη στην περίοδο πρόβλεψης, «churner» ή «non churner». Με την προϋπόθεση λοιπόν της ύπαρξης μιας αξιόπιστης βάσης δεδομένων πελατών με αναλυτική και συστηματική πληροφορία συναλλαγών, οι τεχνολογίες εξόρυξης δεδομένων μπορούν να δώσουν αποτελεσματικές λύσεις στο πρόβλημα της κατανόησης του πελάτη: μοντέλα που αξιολογούν (πιστωτικός κίνδυνος) (Ramachandran, 2004), προβλέπουν (πρόβλεψη απώλειας πελατών) (Hung, et al. 2006), ή ομαδοποιούν πελάτες με χρήση σύνθετων σχημάτων σε ομάδες με υψηλό βαθμό ομοιογένειας και τη μέγιστη απόσταση μεταξύ τους (clustering, τμηματοποίηση) (Mitra et al. 2002). Η συστηματική χρήση των τεχνολογιών αυτών μπορεί να οδηγήσει σε σημαντική γνώση για τους πελάτες, τις συμπεριφορές και τους διάφορους συσχετισμούς τους, που κατ' επέκταση εμπλουτίζουν την πελατειακή βάση παρέχοντας έτσι στη διοίκηση της εταιρίας τη δυνατότητα για άμεση και αποτελεσματική ενημέρωση.

Υπενθυμίζεται ότι απώτερος σκοπός της διαδικασίας είναι η ανάπτυξη της αποδοτικότητας της επιχείρησης, μέσα από τις προσπάθειες διατήρησης των πελατών εκείνων που κινδυνεύουν να εγκαταλείψουν, και μέσα από τον περιορισμό ή και αποφυγή περεταίρω απωλειών. Λαμβάνοντας όμως υπόψη ότι υπάρχουν περιορισμοί πόρων, όπως ο απλούστατος περιορισμός ότι τα κέντρα εξυπηρέτησης πελατών έχουν

μόνο ένα συγκεκριμένο αριθμό εργαζομένων που επικοινωνεί με ένα μικρό τμήμα συνδρομητών, γίνεται κατανοητή η σημαντικότητα της αναγνώρισης εκείνων των πελατών με τη μεγαλύτερη **πιθανότητα απώλειας**, ώστε αυτοί να είναι οι πρώτοι πελάτες με τους οποίους το κέντρο εξυπηρέτησης θα επικοινωνήσει. Άρα λοιπόν, στόχος δεν είναι μόνο να γίνει πρόβλεψη για το αν ένας πελάτης θα φύγει ή όχι από μια επιχείρηση, αλλά να υπολογιστεί η πιθανότητα απώλειας ενός πελάτη, διαφορετικά θα ήταν δύσκολο για μια εταιρία να εκμεταλλευτεί τα αποτελέσματα λόγω περιορισμένων πόρων.

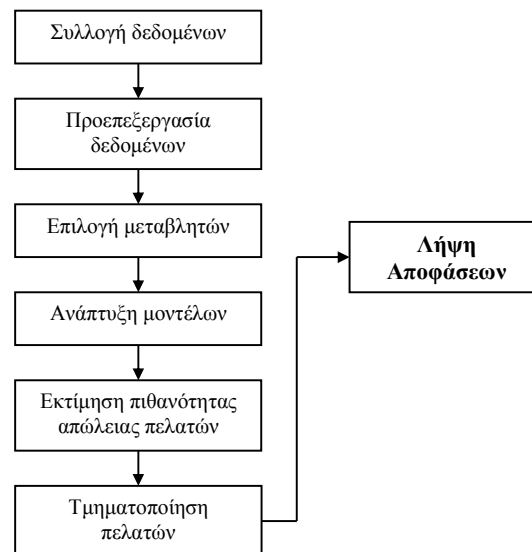
1.1.5. Είναι όλοι οι πελάτες συμφέρον να διατηρηθούν;

Όταν αποφασίζεται η εφαρμογή μιας στρατηγικής διατήρησης πελατών, πρέπει να καθοριστεί σε ποιους πελάτες θα πρέπει να δοθεί προσοχή: Ποιοι είναι οι πιο πολύτιμοι για την εταιρία και θα ήταν πλήγμα να χαθούν; Η ερώτηση αυτή ουσιαστικά απαντάται μέσα από τη λογική της πρόβλεψης απώλειας πελατών και τη σημαντικότητα της εκτίμησης πιθανοτήτων. Με την προσεκτική επιλογή των πιο αξιόλογων πελατών με υψηλό ρίσκο απώλειας, είναι δυνατόν να σχεδιαστούν και να εκτελεστούν πολύ αποτελεσματικές καμπάνιες διατήρησης (**retention campaigns**).

Εδώ όμως πρέπει να γίνει αναφορά σε μία ακόμα έννοια που συσχετίζεται με την απώλεια πελατών, την αξία διάρκειας ζωής του πελάτη (**customer lifetime value-CLV**). Οι Hoekstra et al. (1999) δίνουν έναν ορισμό: «Η αξία διάρκειας ζωής του πελάτη είναι η ολική αξία των άμεσων και έμμεσων συνεισφορών στα γενικά έξοδα και έσοδα ενός πελάτη καθόλη τη διάρκεια ζωής του με την επιχείρηση.» Η αξία διάρκειας ζωής του πελάτη αντιπροσωπεύει τη συνολική προσδοκώμενη συνεισφορά του πελάτη στα κέρδη της εταιρίας και βασίζεται σε κάποια μέτρα: 1) στον προσδοκώμενο χρόνο συνεργασίας του πελάτη με την εταιρία, 2) τα ανά περίοδο έξοδα (π.χ. μηνιαίως) που πληρώνει ο πελάτης στην εταιρία, 3) τα ανά περίοδο έξοδα που κάνει η εταιρία προκειμένου να παράσχει στον πελάτη μια υπηρεσία. Αυτά τα μέτρα συνδυάζονται μαθηματικά για κάθε πελάτη για να παράγουν έναν αριθμό, ή σκορ αξίας διάρκειας ζωής του πελάτη. Να σημειωθεί ότι η παράμετρος αυτή δεν είναι μια σταθερή τιμή αλλά μεταβάλλεται με το χρόνο και τη δυναμικότητα της αγοράς, και επηρεάζεται από τις προσπάθειες μάρκετινγκ που γίνονται από μια εταιρία. Είναι φανερό ότι το ποσοστό απώλειας πελατών, το οποίο ορίζεται ως ο

αριθμός των συνδρομητών που διακόπτουν τη χρήση μιας υπηρεσίας διαιρεμένος με το μέσο αριθμό του συνόλου των συνδρομητών εκείνη την περίοδο, έχει αντίκτυπο στη αξία διάρκειας ζωής του πελάτη επειδή επηρεάζει τη διάρκεια της υπηρεσίας και τα μελλοντικά έσοδα (Jain και Singh, 2002).

Η εταιρία, μέσα από την καλύτερη κατανόηση της αξίας διάρκειας ζωής του πελάτη μπορεί να μετρήσει τα τωρινά και τα μελλοντικά έσοδα από τους πελάτες, και να καλλιεργήσει τη διατήρηση των πελατών και την αφοσίωσή τους που θα οδηγήσουν σε υψηλότερη κερδοφορία (Jain, 2003). Γίνεται φανερό λοιπόν η σημαντικότητα της αξιολόγησης των πελατών μιας επιχείρησης υπό το πρίσμα της αξίας διάρκειας ζωής του πελάτη. Με το συνδυασμό πρόβλεψης απώλειας πελατών και ανάλυσης αξίας διάρκειας ζωής του πελάτη μπορούν να σχεδιαστούν οικονομικά αποδοτικές στρατηγικές διατήρησης πελατών. Αυτό επιτυγχάνεται με την επιλογή κατάλληλων κινήτρων (ατομικά για κάθε συνδρομητή), ώστε το κόστος αυτών να μην υπερβαίνει το προβλεπθέν κέρδος από τη διατήρηση του πελάτη.



Σχήμα 4: Πλαίσιο διαδικασίας λήψης απόφασης διατήρησης πελατών
(<http://www.expresscomputeronline.com/20030922/indiatrends01.shtml>)

Η επιλογή των κατάλληλων πελατών μπορεί να γίνει μέσα από μία διαδικασία που λέγεται τμηματοποίηση πελατών (**customer segmentation**), η οποία διαχωρίζει μια πελατειακή βάση δεδομένων σε διακριτά, ομογενή γκρουπ βάσει συγκεκριμένης μεθοδολογίας (kim et al. 2006). Σκοπός της είναι η κατανόηση της πελατειακής βάσης, και η απόκτηση γνώσης για τους πελάτες που θα βοηθήσει στην

υιοθέτηση της σωστής αντιμετώπισης στο σωστό σύνολο πελατών, τη σωστή στιγμή. Η χρήση της σε συνδυασμό με άλλες τεχνικές μπορεί να οδηγήσει σε:

- ανταγωνιστικά πλεονεκτήματα μέσα από ευέλικτες, στοχευμένες δράσεις και στρατηγικές μάρκετινγκ
- ικανοποίηση πελατών και αφοσίωση (Churn management)
- αποδοτική διαχείριση ρίσκου πελατών (Consumer Risk Management)
- αποτελεσματική παρακολούθηση απόδοσης και λήψη αποφάσεων

Η τμηματοποίηση πελατών μπορεί να είναι προσανατολισμένη στην αγορά (Market Driven), προκειμένου να ενσωματώνει συγκεκριμένα χαρακτηριστικά της αγοράς, ή προσανατολισμένη στα δεδομένα (Data Driven) προκειμένου να ενσωματώνει δομές ή σχέδια βάσει των χαρακτηριστικών των πελατών.

Ένας τρόπος αξιολόγησης πελατών στα πλαίσια της τμηματοποίησης μπορεί να γίνει με τον υπολογισμό του πίνακα ρίσκου/εσόδων (risk/revenue), ο οποίος συνδυάζει πληροφορίες πιθανοτήτων απώλειας πελατών και αξίας διάρκειας ζωής πελάτη, μέσα από δύο βήματα (σχήμα 5): 1. Υπολογισμός της αξίας διάρκειας ζωής των πελατών και διαχωρισμός της σε υψηλή, μέτρια και χαμηλή, 2. Υπολογισμός της πιθανότητας απώλειας πελάτη και διαχωρισμός των πελατών σε 3 τμήματα κατά όμοιο τρόπο:

Πίνακας Ρίσκου/Εσόδων				
		Πιθανότητα Απώλειας του Πελάτη		
		Υψηλή	Μεσαία	Χαμηλή
Αξία Διάρκειας Ζωής του Πελάτη	Υψηλή	προτεραιότητα Α	προτεραιότητα Β	προτεραιότητα Γ
	Μεσαία	προτεραιότητα Β	προτεραιότητα Β	προτεραιότητα Γ
	Χαμηλή	προτεραιότητα Γ	προτεραιότητα Γ	προτεραιότητα Γ

Σχήμα 5: Πίνακας ρίσκου/εσόδων (Thomson, 2005)

Οι πελάτες με τους οποίους πρέπει να ασχοληθεί η εταιρία είναι εκείνοι που ανήκουν στα τμήματα προτεραιότητας Α και Β, με την υψηλότερη αξία διάρκειας ζωής του πελάτη και την υψηλότερη πιθανότητα να φύγουν. Με χαμηλή αξία διάρκειας ζωής, οι πελάτες που είναι πιθανότερο να μείνουν στην εταιρία είναι στο

τμήμα προτεραιότητας Γ, και δε χρειάζεται να σπαταληθούν πόροι για τη διατήρησή τους.

1.2. Διαχείριση σχέσεων πελατών

Η προϋπόθεση για την επιβίωση των επιχειρήσεων σε ένα όλο και πιο ανταγωνιστικό περιβάλλον αγοράς και η διατήρηση της κερδοφορίας τους έγκειται στη δυνατότητα εφαρμογής του σλόγκαν «κατανοείτε τους πελάτες σας». Λαμβάνοντας υπόψη όλα τα παραπάνω είναι κατανοητό ότι για να γίνει αυτό απαραίτητη είναι η ύπαρξη αποτελεσματικών και αξιόπιστων μηχανισμών αξιολόγησης πελατών, ώστε να μπορούν να δοθούν απαντήσεις σε ερωτήματα σχετικά με την ποιότητά τους, την αποδοτικότητα, τον πιστωτικό κίνδυνο, το βαθμό ικανοποίησης, την αφοσίωση, την πιθανότητα απώλειας, και την τοποθέτησή τους σε κάποιο πελατειακό τμήμα. Έτσι οι σύγχρονες επιχειρήσεις την τελευταία δεκαετία καλύπτουν την ανάγκη αυτή χρησιμοποιώντας Συστήματα Διαχείριση Σχέσεων Πελατών (Customer Relationship Management-CRM), τα οποία έχουν αντικαταστήσει τις παραδοσιακές στρατηγικές μαζικού μάρκετινγκ με επιλεκτικές πρακτικές οι οποίες βασίζονται στη μοντελοποίηση και ανάλυση πληροφοριών στο χρόνο σε σχέση με τα προφίλ των πελατών, δίνοντας τη δυνατότητα για διαχείριση απώλειας πελατών, διαχείριση εκστρατείας (campaign management), και εκτίμηση πιστωτικού κινδύνου (Credit Risk assessment) (Coussement et al. 2008).

Το σύστημα διαχείρισης σχέσεων πελατών αποσκοπεί στη βελτίωση των οικονομικών αποτελεσμάτων της επιχείρησης μέσα από συστηματική αύξηση του επιπέδου ικανοποίησης του πελάτη (η οποία επιτυγχάνεται μέσω της διαρκούς, εύστοχης ανταπόκρισης της επιχείρησης στις ιδιαίτερες ανάγκες, επιθυμίες ή ακόμα και στις αρνητικές εμπειρίες των πελατών σχετικά με τις υπηρεσίες-complaint management) (Kumar et al. 2004). Η φιλοσοφία της χρήσης αυτού του συστήματος πλαισιώνεται από ένα σύνολο πρακτικών και μεθοδολογιών καθώς και από εξειδικευμένα πληροφοριακά συστήματα που εφαρμόζουν την εκάστοτε επιχειρηματική πολιτική: αξιοποιεί με βέλτιστο τρόπο τα δεδομένα που περιγράφουν τους πελάτες και τις συμπεριφορές τους, με σκοπό την ολοκληρωμένη και αντικειμενική αξιολόγησή τους και κατ' επέκταση το σχεδιασμό αποτελεσματικών προϊόντων, υπηρεσιών, πολιτικών επικοινωνίας, προγραμμάτων επιβράβευσης και πληθώρας άλλων εφαρμογών. Επιπλέον, η συνεργασία ενός συστήματος διαχείρισης

σχέσεων πελατών με ένα σύστημα διαχείρισης απώλειας πελατών (**churn management**) προσφέρει end-to-end λύσεις για τη διατήρησή τους, που υποστηρίζουν όλη τη διαδικασία της διαχείρισης απώλειας πελατών, από τη συλλογή και αποθήκευση δεδομένων, μέχρι την πρόβλεψη και την παρουσίαση των αποτελεσμάτων δράσης στους λήπτες αποφάσεων. Η λύση αυτή επιτρέπει στην εταιρία να έχει καλύτερη κατανόηση των μεταβλητών που επηρεάζουν την απώλεια πελατών, ενώ οι προβλεπτικές πληροφορίες που προσφέρει είναι κρίσιμες αφού δίνουν στην υπηρεσία ένα παράθυρο να διορθώσει εγκαίρως τα προβλήματα και να περιορίσει την απώλεια πελατών. Συνεπώς ένα τέτοιο σύστημα μπορεί να βοηθήσει στην επινόηση ελκυστικών κινήτρων, πλάνων αφοσίωσης και εξυπηρέτησης πελατών μαζί με στρατηγικές απόκτησης για την προσέλκυση του σωστού πελάτη-όλα κλειδιά για ένα προσοδοφόρο μέλλον (Kim et al. 2006).

Σε μακρο-επίπεδο, σκοπός είναι να κατανοηθεί η πελατειακή βάση δεδομένων, να παρασταθεί σε ουσιαστική γκρουπ πελατών, να παρακολουθούνται και να γίνονται κατανοητές αλλαγές μέσα στο χρόνο, να υποστηρίζονται κρίσιμες στρατηγικές και λειτουργίες, προγράμματα αφοσίωσης και ανάπτυξη προϊόντων. Σε μικρο-επίπεδο, σκοπός είναι η υποστήριξη εκστρατειών, εμπορικών πολιτικών, δραστηριοτήτων cross-selling και up-selling, ανάλυση και διαχείριση απώλειας πελατών και αφοσίωσης.

Στόχος είναι η σωστή μεταχείριση του σωστού πελάτη στο σωστό χρόνο μέσω του σωστού καναλιού.

1.3. Αιτίες απώλειας πελατών

Η παρουσίαση των αιτιών απώλειας πελατών που ακολουθεί αφορά κατά κύριο λόγο στον τομέα των τηλεπικοινωνιών, στον οποίο στηρίζεται η μεταπτυχιακή αυτή εργασία.

1.3.1. Βιβλιογραφική ανασκόπηση

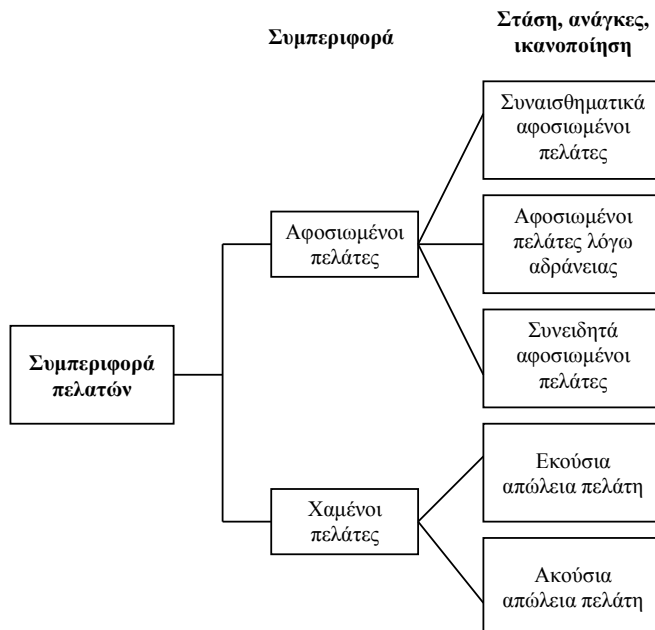
Οι αιτίες απώλειας πελατών έχουν μελετηθεί εκτενώς στη βιβλιογραφία, με μελέτες που ερευνούν πιθανούς προγόνους (Bolton et al. 2000, Dick και Basu 1994), καθοριστικούς παράγοντες και πιθανά ενδιάμεσα στοιχεία (Bolton 1998, Evanschitzky και Wunderlich 2006, Keaveney 1995). Γενικότερα, οι έρευνες στον τομέα του μάρκετινγκ λαμβάνουν υπόψη τη χρησιμότητα ή αξία μιας σχέσης

(οικονομική και ψυχολογική), ενώ οι έρευνες στον οικονομικό τομέα προσεγγίζουν τη συμπεριφορά απώλειας από την οπτική του κόστους αλλαγής εταιρίας (switching cost), το οποίο οι πελάτες αντιμετωπίζουν ως εμπόδιο κατά την απόφασή τους να εγκαταλείπουν μια εταιρία. Στη βιβλιογραφία που αφορά στην αφοσίωση των πελατών, οι παράγοντες που την επηρεάζουν κατατάσσονται σε 2 μεγάλες ομάδες: οικονομικοί (ή γνωστικοί) παράγοντες και κοινωνικοί (ή συναισθηματικοί) παράγοντες (Geyskens και Steenkamp 2000). Οι οικονομικοί παράγοντες εστιάζουν στην οικονομική αξία της σχέσης του πελάτη με μια εταιρία, συμπεριλαμβάνοντας την (οικονομική) ικανοποίηση, τα αντιλαμβανόμενα επίπεδα τιμής, και τους δείκτες τιμή-αξία (price-value). Οι κοινωνικοί παράγοντες αφορούν σε κοινωνικά και συναισθηματικά θέματα, όπως η εμπιστοσύνη και το δέσιμο (commitment) (Verhoef et al. 2002).

Ενδεικτικά θα αναφερθούν συγκεκριμένες μελέτες πάνω στο θέμα. Οι Gerpott et al. (2001) ανέλυσαν ένα μοντέλο όπου η συνολική ικανοποίηση πελατών έχει σημαντική επίδραση στην αφοσίωση, η οποία με τη σειρά της επηρεάζει τις προθέσεις του πελάτη να τερματίσει τις σχέσεις του με μία εταιρία. Οι Lee et al. (2003) βρήκαν ορισμένους παράγοντες της απώλειας πελατών στην αγορά τηλεπικοινωνιών της Κορέας. Οι Kim et al. (2004) ερεύνησαν τις επιδράσεις των κοστών αλλαγής εταιρίας στην ικανοποίηση και αφοσίωση πελατών. Άλλες έρευνες έχουν εξετάσει τις διαφορές στην ικανοποίηση και αφοσίωση μεταξύ των πελατών εκείνων που εγκαταλείπουν μια εταιρία και εκείνων που μένουν σε αυτή (Ganesh et al. 2000), ενώ πληθώρα μελετών υποστηρίζει τη σημαντικότητα της σχέσης ικανοποίησης και ποιότητας υπηρεσιών στη συμπεριφορά απώλειας πελατών (Anderson και Sullivan 1993, Boulding et al. 1993, Oliva et al. 1995, Reichheld και Sasser 1990). Σε μία σχετική έρευνα, ο Keaveney (1995) εντόπισε 8 βασικές αιτίες απώλειας πελατών. Κάποιες σχετίζονται με την αίσθηση της δυσαρέσκειας με μία υπηρεσία (π.χ. αποτυχία εξυπηρέτησης), και άλλες αποτελούν παράγοντες όπως η τιμή και ο ανταγωνισμός. Και ενώ η θεωρία για το κόστος αλλαγής εταιρίας προϋποθέτει την ύπαρξη συνειδητής απόφασης του πελάτη, η θεωρία για την αδράνεια (inertia) στηρίζεται στο γεγονός ότι ένας πελάτης εμφανίζει αγοραστική συμπεριφορά λόγω συνήθειας. Ο πελάτης δηλαδή συνεχίζει να έχει τον ίδιο πάροχο υπηρεσιών μόνο και μόνο επειδή αυτό έκανε και πριν, χωρίς να λαμβάνει υπόψη πιθανές εναλλακτικές. Πολλές μελέτες τονίζουν τη σημαντικότητα της αδράνειας (Rust et al. 2004) ως πιθανό (αρνητικό) παράγοντα της απώλειας πελατών.

1.3.2. Διαχωρισμός πελατών με βάση τη συμπεριφορά

Οι πελάτες, όπως αντιμετωπίζονται στη βιβλιογραφία απώλειας πελατών, πρωταρχικά χωρίζονται σε αφοσιωμένους και μη (πελάτες που έχουν εγκαταλείψει μία εταιρία), και στη συνέχεια χωρίζονται σε περεταίρω ομάδες με βάση τη στάση και συμπεριφορά τους (σχήμα 6).



Σχήμα 6: Απώλεια πελατών με βάση τη συμπεριφορά (Kim et al. 2004)

1.3.2.1. Αφοσιωμένοι πελάτες

Συναισθηματικά αφοσιωμένοι πελάτες (emotive customers): οι πελάτες αυτοί είναι και οι πιο αφοσιωμένοι. Νιώθουν έντονα ότι οι αγορές τους είναι οι σωστές για εκείνους και ότι το επιλεγμένο προϊόν ή υπηρεσία είναι η καλύτερη επιλογή, ενώ σπάνια επαναξιολογούν τις αγοραστικές τους αποφάσεις. Αυτά τα συναισθήματα μπορεί να είναι αντίκτυπος μιας μεγάλης περιόδου καλών αποδόσεων των προϊόντων/υπηρεσιών που έχει αγοράσει ο πελάτης, ενώ συχνά καλλιεργούνται από απροσδιόριστους παράγοντες. Συνήθως οι συναισθηματικοί πελάτες ξοδεύουν παραπάνω και έχουν χαμηλότερο επίπεδο απώλειας πελατών από άλλους.

Αφοσιωμένοι πελάτες λόγω αδράνειας (inertial customers): όπως και οι συναισθηματικά αφοσιωμένοι, αυτοί οι πελάτες σπάνια επαναξιολογούν τις αγορές τους, αλλά αυτή τους η απραξία είναι αποτέλεσμα του περιορισμού που νιώθουν

λόγω του υψηλού κόστους αλλαγής εταιρίας ή της έλλειψης δεσίματος με ορισμένα προϊόντα/υπηρεσίες. Αν και αυτοί οι πελάτες δεν τείνουν να ξοδεύουν περισσότερο ή λιγότερο από ότι συνήθως, ο επηρεασμός τους προσφέρει τόσες ευκαιρίες όσες και ο επηρεασμός των συναισθηματικά αφοσιωμένων πελατών.

Συνειδητά αφοσιωμένοι πελάτες (deliberators): οι πελάτες αυτοί είναι πληροφορημένοι, ενημερώνονται για τυχόν αγοραστικές ευκαιρίες, και επανεξετάζουν τις επιλογές τους χρησιμοποιώντας λογικά κριτήρια, όπως η απόδοση, η τιμή και η ποιότητα. Αυτό το είδος των πελατών αποτελεί τη μεγαλύτερη αγοραστική ομάδα, αντιπροσωπεύοντας κατά μέσο όρο το 40% του συνόλου των πελατών. Τα οφέλη από τον επηρεασμό αυτών των πελατών έχουν διπλάσια αξία σε σχέση με αυτά των δύο προηγούμενων ομάδων, καθώς αυτοί είναι οι πελάτες με τις μεγαλύτερες πιθανότητες απώλειας (Kim et al. 2004).

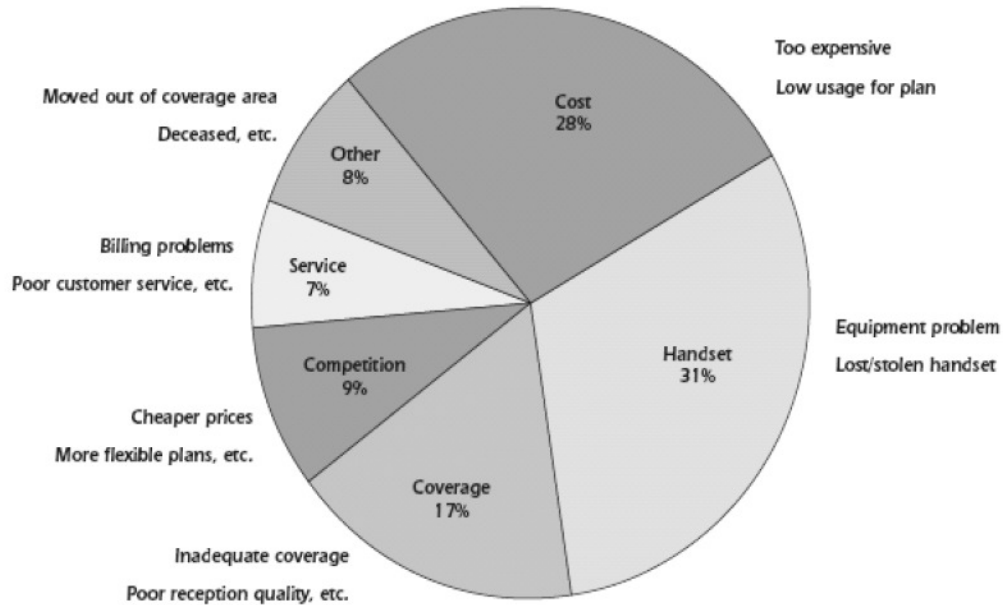
1.3.2.2. Χαμένοι πελάτες

Οι πελάτες που έχουν εγκαταλείψει μια εταιρία χωρίζονται σε εκείνους που έφυγαν εκούσια (voluntary churn-όταν επιλέγουν οι ίδιοι οι πελάτες να αφήσουν μία εταιρία ή να τερματίσουν ένα συμβόλαιο) και σε εκείνους που έφυγαν ακούσια (nonvoluntary churn-όταν η ίδια η εταιρία τερματίζει ένα συμβόλαιο ή απενεργοποιεί μία υπηρεσία) (Hadden et al. 2007).

Εκούσια απώλεια πελάτη

Η εκούσια απώλεια πελατών μπορεί να χωριστεί σε 2 κατηγορίες, τυχαία και εσκεμμένη. Η τυχαία απώλεια πελατών συμβαίνει όταν οι περιστάσεις καθιστούν τη χρήση μιας υπηρεσίας μη απαραίτητη. Π.χ. σε περίπτωση θανάτου, μετακόμισης σε άλλη γεωγραφική θέση, ή ραγδαία αλλαγή στην οικονομική κατάσταση του πελάτη ώστε να μην μπορεί να καλύψει τις οικονομικές του υποχρεώσεις. Η τυχαία απώλεια πελατών συνήθως είναι υπεύθυνη για ένα μικρό ποσοστό της εκούσιας απώλειας. Η εσκεμμένη απώλεια πελατών συμβαίνει όταν ένας πελάτης αποφασίζει συνειδητά να αλλάξει εταιρεία. Μερικοί λόγοι που μπορεί να οδηγήσουν στην απόφαση είναι αν ο πελάτης π.χ. ανακαλύψει ότι μια ανταγωνίστρια εταιρία προσφέρει καλύτερα προϊόντα, που ο τρέχον πάροχος δεν μπορεί να παρέχει, είναι η εύρεση του προϊόντος σε χαμηλότερη τιμή, και άλλοι παράγοντες όπως η ανεπαρκής κάλυψη, ή άσχημη εξυπηρέτηση κ.α. (Kim και Yoon, 2004). Έρευνα της συμβουλευτικής εταιρίας Booz· Allen (www.boozallen.com), υποστηρίζει ότι οι παράγοντες που σχετίζονται με

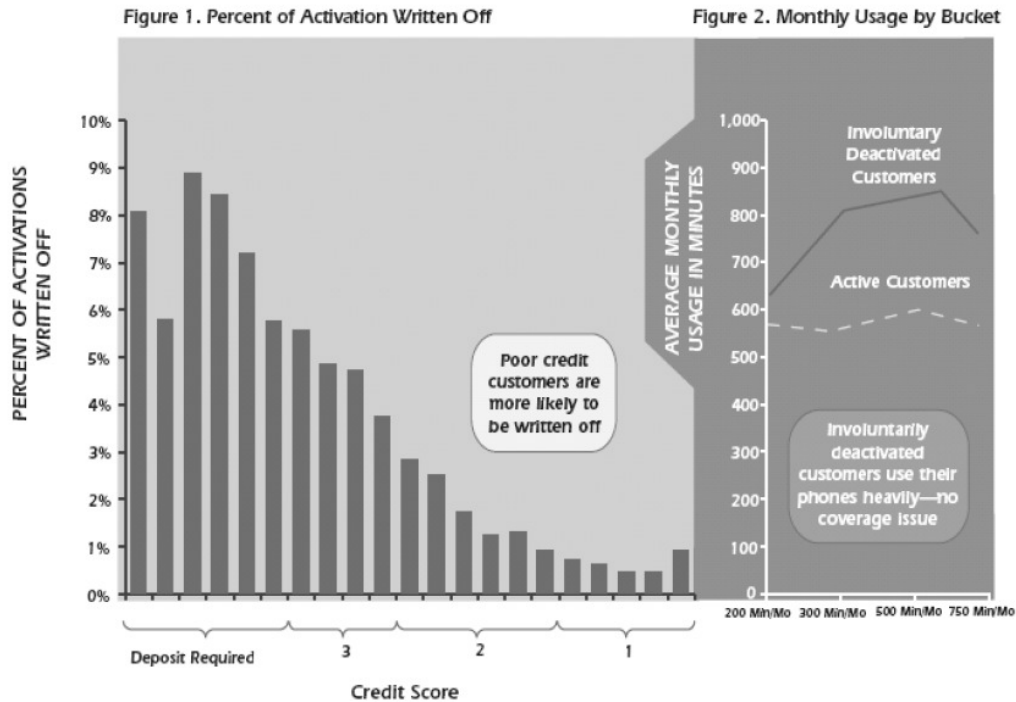
την τιμή αποτελούν τη σημαντικότερη αιτία για την εκούσια απώλεια πελατών (σχήμα 7). Επιπλέον, σημαντικοί είναι και παράγοντες όπως η ακαταλληλότητα των συμβολαίων σε σχέση με τα χαρακτηριστικά των πελατών και τις ανάγκες τους, αλλά και η έλλειψη ενημέρωσης των πελατών για νέα προϊόντα και υπηρεσίες, που μπορεί να οδηγήσει σε μη πληροφόρηση και αβεβαιότητα.



Σχήμα 7: Παράγοντες εκούσιας απώλειας πελατών (Desai, 2006)

Ακούσια απώλεια πελάτη

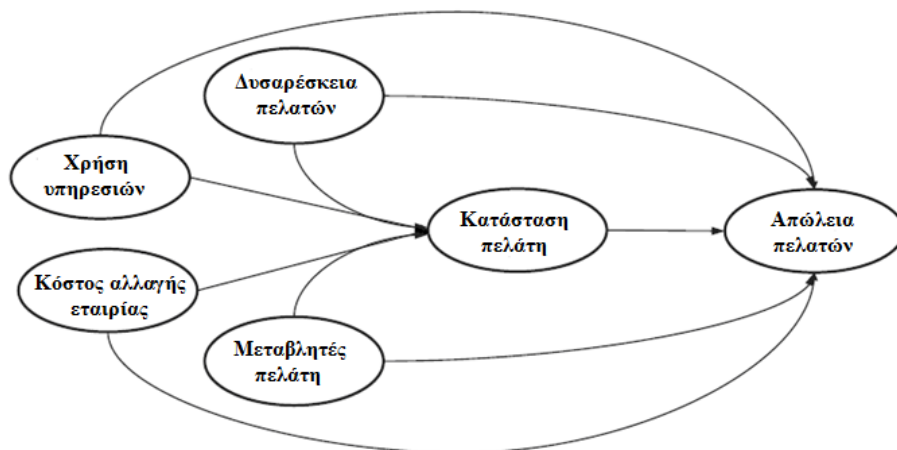
Οι ακούσια χαμένοι πελάτες είναι πιο εύκολο να αναγνωριστούν, αφού είναι εκείνοι οι πελάτες των οποίων οι υπηρεσίες διακόπηκαν από την ίδια την εταιρία. Υπάρχουν πολλοί λόγοι για τους οποίους μια εταιρία θα μπορούσε να διακόψει τη σχέση της με ένα πελάτη, όπως είναι η κατάχρηση μιας υπηρεσίας ή η μη κάλυψη των οικονομικών τους υποχρεώσεων (σχήμα 8). Από την πλευρά της εταιρίας αυτό το πρόβλημα μπορεί να οφείλεται στο ότι η ίδια δεν κατάφερε να παρακολουθεί τους πελάτες με τη μεγαλύτερη πιθανότητα απώλειας, και να παρέμβει πριν τα επίπεδα χρεώσεών τους γίνουν πολύ μεγάλα. Η ανάλυση που πραγματοποιήθηκε στα πλαίσια της έρευνας της εταιρίας Booz Allen έδειξε ότι οι πελάτες που εμφάνισαν ακούσια απώλεια είχαν μεγαλύτερα επίπεδα μηνιαίας χρήσης σε σχέση με τους ενεργούς πελάτες, και ότι οι πελάτες που έχασαν την υπηρεσία λόγω αθέτησης πληρωμών αντιμετώπισαν αυτό το πρόβλημα επειδή δεν γνώριζαν πώς να ελέγχουν το επίπεδο της χρήσης τους, ώστε να την περιορίσουν πριν να είναι πολύ αργά.



Σχήμα 8: Χαρακτηριστικά ακούσιας απώλειας πελατών (Desai, 2006)

1.3.3. Παράγοντες απώλειας πελατών

Παρακάτω περιγράφονται συγκεντρωτικά οι παράγοντες απώλειας πελατών και ο τρόπος επηρεασμού τους, όπως αυτοί προέκυψαν μέσα από σχετική βιβλιογραφική έρευνα στον τομέα των τηλεπικοινωνιών. Στο σχήμα 9 φαίνεται η γραφική απεικόνισή τους.



Σχήμα 9: Μοντέλο απεικόνισης παραγόντων απώλειας πελατών (Ahn et al. 2006)

1.3.3.1. Δυσαρέσκεια πελατών

Πληθώρα ερευνών έχει δείξει ότι η δυσαρέσκεια των πελατών σχετίζεται με την αύξηση της απώλειας πελατών (Keaveney 1995, Anderson και Sullivan 1993, Boulding et al. 1993, Oliva et al. 1995, Oliver 1997, Reichheld και Sasser 1990). Έρευνες απέδειξαν ότι η ποιότητα δικτύου και η ποιότητα κλήσεων είναι οδηγοί της ικανοποίησης (ή δυσαρέσκειας) των πελατών στον τομέα των τηλεπικοινωνιών (Gerpott et al. 2001, Kim et al. 2004), και ότι η αποτυχία εξυπηρέτησης επιταχύνει την απόφαση ενός πελάτη να διακόψει τη σχέση του με μία εταιρία (Bolton 1998, Bolton et al. 2000, Mozer et al. 2000). Παρόλα αυτά, πρέπει να σημειωθεί ότι οι αποτυχίες κλήσεων δεν οδηγούν απαραίτητα στην απώλεια πελατών επειδή αποδίδονται σε πολλά γεγονότα (π.χ., παρεμβολές, συσκευή, κ.α.), ενώ οι διακοπές κλήσεων οφείλονται στον τηλεπικοινωνιακό πάροχο. Επιπλέον η δυσαρέσκεια με μία υπηρεσία ή ένα προϊόν είναι θετικά συσχετισμένη με τη συμπεριφορά παραπόνων των πελατών και κατά συνέπεια με την απώλεια πελατών (Bearden και Teel 1983). Αυτό μπορεί να σημαίνει ότι είτε τα παράπονα που γίνονται στα κέντρα εξυπηρέτησης πελατών δεν λαμβάνουν την κατάλληλη διαχείριση ή ότι κάποια από αυτά δεν αντιμετωπίζονται άμεσα (χρειάζεται χρόνος για να διορθωθεί ένα πρόβλημα π.χ. κάλυψης δικτύου και οι πελάτες μπορεί να μην είναι διατεθειμένοι να περιμένουν).

1.3.3.2. Κόστος αλλαγής εταιρίας

Όπως αναφέρθηκε και παραπάνω, οι πελάτες διατηρούν τη σχέση τους με μία εταιρία είτε λόγω περιορισμών (πρέπει να διατηρήσουν τη σχέση) είτε λόγω αφοσίωσης (θέλουν να διατηρήσουν τη σχέση) (Bendapudi και Berry 1997). Το κόστος αλλαγής εταιρίας (switching cost) είναι ένας παράγοντας που δρα ως περιορισμός, εμποδίζοντας τους πελάτες από το να αλλάζουν ελεύθερα παρόχους υπηρεσιών. Στον τομέα των τηλεπικοινωνιών, οι πόντοι αφοσίωσης (loyalty points) και τα προγράμματα καρτών μέλους (membership card programs) είναι οι βασικοί παράγοντες που δρουν κατ' αυτό τον τρόπο, επειδή σκοπός των προγραμμάτων αυτών είναι να αυξήσουν το βαθμό της διατήρησης πελατών με την παροχή οφελών στα μέλη τους, τα οποία οφέλη χάνονται κατά την απώλεια πελατών (Kim et al. 2004). Επιπλέον, κατά τους Bolton et al. (2000), τα μέλη των προγραμμάτων επιβράβευσης μπορεί να αγνοήσουν ή να μην πάρουν σοβαρά αρνητικές αξιολογήσεις της εταιρίας με τους ανταγωνιστές της, και συνεπώς τέτοια

προγράμματα μπορούν να εμποδίσουν έμμεσα την απώλεια πελατών. Παρόλα αυτά η αποτελεσματικότητα των προγραμμάτων καρτών μέλους στην απώλεια πελατών δεν έχει καθοριστεί πλήρως, αντίθετα, έρευνα έχει αποδείξει ότι το ποσοστό απώλειας των ατόμων που συμμετέχουν σε τέτοια προγράμματα είναι υψηλότερο από εκείνο των ατόμων που δε συμμετέχουν (Ahn et al. 2006).

1.3.3.3. Χρήση υπηρεσιών

Το επίπεδο χρήσης μιας υπηρεσίας, σε μηνιαία χρέωση, είναι ένας από τους πιο σημαντικούς παράγοντες απώλειας πελατών (Buckinx και Poel 2005, Mozer et al. 2000). Παρόλα αυτά, δεν είναι ξεκάθαρο εάν η σχέση μεταξύ της χρήσης υπηρεσιών και της απώλειας πελατών είναι θετική ή αρνητική. Οι Oliva et al. (1995) και Oliver (1997) υποστηρίζουν ότι όταν ένας πελάτης χρησιμοποιεί μια υπηρεσία συχνά, αναπτύσσει θετική στάση απέναντί της, και συνεπώς ένας πελάτης με πολλές θετικές εμπειρίες σε μία υπηρεσία δε θα είναι ευαίσθητος σε μεμονωμένες περιπτώσεις αποτυχίας και άρα δε θα έχει μεγάλη πιθανότητα απώλειας. Αντίθετα, όπως υποστηρίζει ο Madden et al. (1999), οι συχνοί χρήστες είναι πιο πιθανό να εγκαταλείψουν μια εταιρία, και οι πελάτες με μεγάλο ποσοστό χρεώσεων είναι πιο ευαίσθητοι στην τιμή και πιο επιρρεπείς στην απώλεια, ενώ οι πελάτες που δεν τηρούν τις οικονομικές τους υποχρεώσεις δεν είναι επιρρεπείς στην απώλεια πελατών άμεσα, καθώς δεν μπορούν να φύγουν οικειοθελώς, παρά μόνο αν πληρώσουν ένα σημαντικό ποσοστό των οφελών τους.

1.3.3.4. Μεταβλητές σχετικές με τον πελάτη

Ενασχόληση

Οι πελάτες που ασχολούνται ενεργά με την επιλογή μιας υπηρεσίας ή μιας εταιρίας (involved consumers), έχει βρεθεί ότι εμφανίζουν μεγαλύτερα επίπεδα ικανοποίησης τα οποία και τείνουν να διατηρούνται (Richins και Bloch 1991). Αυτοί οι πελάτες κάνουν προαγοραστική έρευνα (Beatty και Smith 1987) και δείχνουν μεγαλύτερη προσοχή στην επιλογή (Celsi και Olson 1988). Μεγαλύτερα επίπεδα ενασχόλησης συνδέονται με μεγαλύτερα επίπεδα αφοσίωσης (commitment) με μία εταιρία και αντίσταση στην αλλαγή πεποιθήσεων (Pritchard et al. 1999). Έρευνες δείχνουν ότι σε περιπτώσεις υψηλής αφοσίωσης σε προϊόντα ή υπηρεσίες, οι ικανοποιημένοι πελάτες επιθυμούν να διατηρήσουν την κατάσταση ικανοποίησης, ακόμα και σε περιπτώσεις χαμηλών αποδόσεων (Keaveney και Parthasarathy 2001),

και ότι οι ικανοποιημένοι πελάτες αντέχουν σε μεγαλύτερο επίπεδο τη φτώχη απόδοση όσο αυξάνεται η αφοσίωσή τους με ένα προϊόν ή υπηρεσία (Oliva et al. 1995).

Δημογραφικά στοιχεία

Πελάτες με υψηλό εισόδημα και επίπεδο μόρφωσης μπορούν να αναπτύξουν σχετικά ακριβείς εκτιμήσεις του τι να περιμένουν από μια υπηρεσία, αφού για παράδειγμα οι πελάτες με μεγαλύτερο εισόδημα μπορεί να έχουν εμπειρία με πιο συχνή χρήση υπηρεσιών μεγαλύτερης ποικιλίας, γεγονός που τους προσδίδει εμπειρία. Οι πελάτες με υψηλότερο επίπεδο μόρφωσης μπορεί να διαθέτουν περισσότερες ικανότητες καθορισμού του επιπέδου μελλοντικής χρήσης μιας υπηρεσίας και των απαιτήσεών τους. Αντίθετα, οι πελάτες με χαμηλό εισόδημα και επίπεδο μόρφωσης μπορεί να έχουν ασαφείς προσδοκίες, αβέβαιες εκτιμήσεις, (Hoch και Deighton 1989), και να είναι πιο επιρρεπείς σε μεμονωμένα περιστατικά δυσαρέσκειας (Keaveney και Parthasarathy, 2001).

Άλλοι παράγοντες

Μεταξύ άλλων παραγόντων που σχετίζονται με τον πελάτη, ιδιαίτερου ενδιαφέροντος είναι το ότι η λειτουργική δυνατότητα της τηλεφωνικής συσκευής ενός πελάτη έχει σημαντικό αντίκτυπο στην πιθανότητα απώλειάς του. Όσο πιο πεπαλαιωμένη είναι η συσκευή, τόσο πιο πιθανό είναι ο πελάτης να αφήσει μια εταιρία. Καθώς οι εταιρίες τηλεπικοινωνιών επιταχύνουν το ποσοστό ανάπτυξης και προώθησης νέων υπηρεσιών, πιο εξελιγμένα χαρακτηριστικά γίνονται διαθέσιμα στην αγορά, και οι πελάτες επιθυμούν να έχουν ότι πιο τελευταίο κυκλοφορεί. Η αλλαγή συσκευής μπορεί να οδηγήσει στην απώλεια πελατών για πολλούς λόγους. Κατά την αλλαγή συσκευής ο πελάτης λαμβάνει υπόψη του πολλά μοντέλα, κάποια από τα οποία είναι συμβατά με συγκεκριμένους παρόχους υπηρεσιών τηλεπικοινωνίας. Επιπλέον διάφορα οφέλη που προσφέρονται από εταιρίες κατά την αγορά συσκευής μπορεί να ωθήσουν τους πελάτες να αλλάξουν πάροχο υπηρεσιών για να απολαύσουν τα οφέλη αυτά (Ahn et al. 2006).

1.3.3.5. Κατάσταση πελατών

Ο παράγοντας αυτός δεν αποτελεί άμεσα αιτία απώλειας πελατών, αλλά θεωρείται ως ενδιάμεσο στοιχείο μεταξύ των παραγόντων απώλειας πελάτη και της ίδιας της απώλειας, όπως φαίνεται και στο σχήμα 9, και επομένως η αλλαγή της

κατάστασης ενός πελάτη μπορεί να αποτελέσει στοιχείο προειδοποίησης για μία εταιρία. Τα διάφορα είδη κατάστασης πελατών ομαδοποιούνται σε 3 κατηγορίες: ενεργή χρήση, μη χρήση και καταστολή. Συμπερασματικά, πελάτες με μη χρήση ή σε καταστολή, είναι πιο πιθανό να εμφανίσουν συμπεριφορά απώλειας πελατών από τους ενεργητικούς πελάτες (Ahn et al. 2006).

1.3.4. Σύνοψη

Η απώλεια πελατών και οι παράγοντες που την επηρεάζουν έχουν μελετηθεί εκτενώς στη βιβλιογραφία που αφορά στην αφοσίωση, στο μάρκετινγκ, στην οικονομία, κ.α. Συμπερασματικά, σημαντικότερο παράγοντα της απώλειας πελατών αποτελεί η αφοσίωση, πρόγονος της οποίας είναι η ικανοποίηση, η οποία και εμποδίζει την απώλεια πελατών, και ενδυναμώνει τη διατήρησή τους (Fornell 1992, Reichheld 1996). Επιπλέον, επειδή η αφοσίωση επηρεάζεται από τη δομή της αγοράς, το είδος και τη συμπεριφορά του πελάτη, υπεισέρχονται επιπλέον παράγοντες στον επηρεασμό της. Αυτοί είναι το κόστος αλλαγής, το οποίο δρα ως αποτρεπτικός παράγοντας, η ποιότητα υπηρεσιών, η οποία επηρεάζει θετικά την ικανοποίηση, οι διαπροσωπικές σχέσεις, το χτίσιμο της εμπιστοσύνης και η κατάλληλη υποστήριξη πελατών που ενισχύουν σημαντικά την αφοσίωση, και άλλοι παράγοντες οι οποίοι εξαρτώνται από τον κλάδο στον οποίο ανήκει μία επιχείρηση.

2. Μεθοδολογικές προσεγγίσεις στην ανάλυση και πρόβλεψη απώλειας πελατών

Σε αυτή την ενότητα περιγράφονται περιληπτικά οι μέθοδοι που χρησιμοποιήθηκαν κατά την ανάλυση και πρόβλεψη της απώλειας πελατών, καθώς και τα κριτήρια αξιολόγησης που χρησιμοποιήθηκαν προκειμένου να εξεταστεί η απόδοση κάθε μεθόδου. Συγκεκριμένα θα παρουσιαστούν οι μέθοδοι:

- Λογιστική Παλινδρόμηση (Logistic Regression)
- Δέντρα Απόφασης (Decision Trees)
- Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

2.1. Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση (logistic regression) είναι ένα γενικευμένο γραμμικό μοντέλο που αποτελεί γνωστή τεχνική ταξινόμησης για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος με την προσαρμογή των δεδομένων σε μια λογιστική καμπύλη (logistic curve). Η μέθοδος αυτή χρησιμοποιείται σε περιπτώσεις που η εξαρτημένη μεταβλητή είναι διχοτομική, και όχι συνεχής, και όπου είναι επιθυμητή η πρόβλεψη της παρουσίας ή απουσίας ενός διακριτού χαρακτηριστικού ή αποτελέσματος βάσει κάποιων τιμών ενός συνόλου συνεχών ή/και διακριτών μεταβλητών πρόβλεψης (Mutanen, 2006). Αν και η λογιστική παλινδρόμηση έχει χρησιμοποιηθεί σε μία πληθώρα τομέων, όπως οι ιατρικές και κοινωνικές επιστήμες (Soldin et al. 2002), έχει χρησιμοποιηθεί και σε εφαρμογές μάρκετινγκ, όπως στην ανάλυση πελατών (Buckinx et al. 2005).

Το μοντέλο της γραμμικής παλινδρόμησης μπορεί να περιγραφεί ως εξής:

$$P(X) = \alpha + \beta X$$

όπου $X = (x_1, x_2, \dots, x_n)$ είναι ένα διάνυσμα n ανεξάρτητων τιμών.

Δεδομένου όμως ότι η πιθανότητα ενός γεγονότος $P(X)$ πρέπει να βρίσκεται μεταξύ του 0 και του 1, δεν είναι πρακτικό να μοντελοποιούνται οι πιθανότητες με τις

τεχνικές γραμμικής παλινδρόμησης, γιατί σε αυτή την περίπτωση η πιθανότητα θα μπορούσε να πάρει τιμές μεγαλύτερες του 1 ή μικρότερες του 0. Συνεπώς χρησιμοποιείται το μοντέλο της λογιστικής παλινδρόμησης, το οποίο επεκτείνει το γραμμικό μοντέλο με το να συνδέει το πεδίο των πραγματικών αριθμών στο $[0, 1]$.

Έχοντας ως δεδομένο ότι η εξαρτημένη μεταβλητή είναι διχοτομική και άρα ισχύει ότι μπορεί να πάρει την τιμή 1 με πιθανότητα $P(X)$ και την τιμή 0 με πιθανότητα $Q(X)$, παρουσιάζεται η λογιστική συνάρτηση:

$$P(X) = \frac{e^{-(\alpha + \beta X)}}{1 + e^{-(\alpha + \beta X)}}$$

$$Q(X) = 1 - P(X) = \frac{1}{1 + e^{-(\alpha + \beta X)}}$$

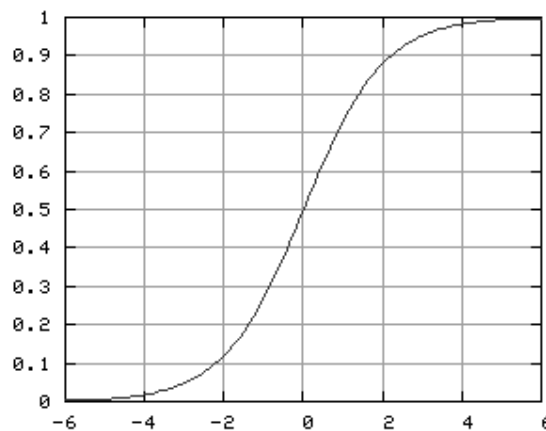
Άρα στο παραπάνω γραμμικό μοντέλο παλινδρόμησης με n εισόδους, η έξοδος προκύπτει από τον τύπο:

$$P(X) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Στο μοντέλο λογιστικής παλινδρόμησης ο αντίστοιχος τύπος της λογιστικής συνάρτησης είναι:

$$Q(X) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad \text{ή} \quad f(z) = \frac{1}{1 + e^{-z}},$$

Όπου είσοδος είναι το z και η έξοδος το $f(z)$ (σχήμα 10).



Σχήμα 10: Η λογιστική συνάρτηση, με το z στον οριζόντιο και το $f(z)$ στον κάθετο άξονα
(http://en.wikipedia.org/wiki/Logistic_regression)

Η λογιστική συνάρτηση είναι χρήσιμη επειδή μπορεί να πάρει ως είσοδο οποιαδήποτε τιμή από το αρνητικό άπειρο ως το θετικό, ενώ η έξοδος περιορίζεται σε

τιμές μεταξύ του 0 και του 1. Το $f(z)$ αντιπροσωπεύει την πιθανότητα ενός συγκεκριμένου αποτελέσματος, δεδομένου του συνόλου των παραγόντων ρίσκου (risk factors). Η μεταβλητή z , γνωστή ως λογιστική (logit) μεταβλητή, αντιπροσωπεύει ένα μέτρο της συνολικής συνεισφοράς όλων των παραγόντων ρίσκου που χρησιμοποιήθηκαν στο μοντέλο και μπορεί να θεωρηθεί ως η «κλίση» προς το γεγονός του ενδιαφέροντος. Η μεταβλητή z ορίζεται ως:

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Ο σταθερός όρος α είναι η τιμή του z όταν η τιμή όλων των παραγόντων ρίσκου είναι 0 (δίνει το P όταν το X είναι 0), ενώ κάθε ένας από τους συντελεστές παλινδρόμησης περιγράφει το μέγεθος της συνεισφοράς του κάθε παράγοντα. Ένας θετικός συντελεστής παλινδρόμησης σημαίνει ότι ο συγκεκριμένος παράγοντας ρίσκου αυξάνει την πιθανότητα εμφάνισης ενός αποτελέσματος, ενώ ένας αρνητικός συντελεστής παλινδρόμησης σημαίνει ότι ο παράγοντας αυτός μειώνει την πιθανότητα εμφάνισης αυτού του αποτελέσματος. Τα αποτελέσματα των συνεχών πιθανοτήτων που παράγονται από το μοντέλο της λογιστικής παλινδρόμησης διαχωρίζονται σε δύο ομάδες με τη χρήση ενός κατωφλιού.

Ένα σημαντικό στοιχείο στη λογιστική παλινδρόμηση είναι η συνάρτηση απώλειας (loss function). Η συνάρτηση αυτή είναι ένα μέτρο προσαρμογής μεταξύ του μαθηματικού μοντέλου και των δεδομένων. Οι παράμετροι επιλέγονται έτσι ώστε να μεγιστοποιείται η προσαρμογή του μοντέλου στα δεδομένα. Στα γραμμικά μοντέλα, υπάρχει αναλυτική λύση στο πρόβλημα που ελαχιστοποιεί το τετραγωνικό σφάλμα, αλλά στα μοντέλα της λογιστικής παλινδρόμησης δεν υπάρχει αναλυτική λύση που παράγει αντίστοιχες εκτιμήσεις των παραμέτρων, και γι αυτό το λόγο η συνάρτηση απώλειας που επιλέγεται είναι η συνάρτηση μέγιστης πιθανοφάνειας (maximum likelihood) προκειμένου να πραγματοποιηθεί η κατάλληλη προσαρμογή του μοντέλου στα δεδομένα χρησιμοποιώντας τεχνικές μη γραμμικής βελτιστοποίησης (Allison, 1999).

Γίνεται φανερό ότι η λογιστική παλινδρόμηση είναι ένας χρήσιμος τρόπος να περιγραφεί η σχέση μεταξύ ενός ή περισσότερων παραγόντων ρίσκου (π.χ. ηλικία, φύλλο κτλ) και ενός αποτελέσματος, εκφρασμένη ως πιθανότητα, που έχει μόνο δύο πιθανές τιμές.

2.1.1. Πλεονεκτήματα

Συμπερασματικά, η μέθοδος αυτή είναι πολύ δημοφιλής για τους παρακάτω λόγους:

1. Η λογιστική μοντελοποίηση είναι γνωστή, απλή, εύκολη στην ερμηνεία (σε σχέση με άλλα μοντέλα, π.χ. νευρωνικά δίκτυα) και συχνά χρησιμοποιούμενη στο μάρκετινγκ (Bucklin et al. 1992).
2. Έχει αποδειχτεί ότι παρέχει καλά και εύρωστα αποτελέσματα, σε σύγκριση με άλλες τεχνικές ταξινόμησης, σε μελέτες τόσο για την πρόβλεψη απώλειας πελατών όσο και για το credit scoring (Neslin et al. 2004).

2.2. Δέντρα απόφασης

Τα Δέντρα Απόφασης είναι δομές σε σχήμα δέντρου που αναπαριστούν ομάδες αποφάσεων και μπορούν να ταξινομούν δεδομένα μέσα από την εξαγωγή if-then κανόνων ταξινόμησης (Muata et al. 2004). Σκοπός τους είναι η δημιουργία ενός μοντέλου που προβλέπει την τιμή μιας εξαρτημένης μεταβλητής (target) με βάση τις τιμές ανεξάρτητων μεταβλητών εισόδου (predictor). Η μέθοδος των δέντρων απόφασης χρησιμοποιείται στην εξόρυξη δεδομένων, την ανάλυση αποφάσεων και τη μηχανική εκμάθηση, και αναπτύσσει ένα δέντρο απόφασης ως μοντέλο πρόβλεψης το οποίο χαρτογραφεί τις παρατηρήσεις για ένα αντικείμενο και τα συμπεράσματα για την τιμή του αντικειμένου αυτού, καθώς επίσης χρησιμοποιείται για να αναπαραστήσει οπτικά και ρητά τη διαδικασία λήψης αποφάσεων (http://en.wikipedia.org/wiki/Decision_tree).

Σε αυτές τις δομές, κάθε εσωτερικός κόμβος αντιστοιχεί σε μία από τις μεταβλητές εισόδου. Τα φύλλα αντιπροσωπεύουν τις κατηγορίες και τα κλαδιά τους συνδέσμους των χαρακτηριστικών που οδηγούν σε αυτές τις κατηγορίες. Η διαδικασία που ακολουθείται έγκειται στο διαχωρισμό του αρχικού συνόλου σε διμερή υποσύνολα, κάθε ένα από τα οποία περιγράφεται από έναν απλό κανόνα για ένα ή περισσότερα χαρακτηριστικά σχεδίων, πράγμα που επιτρέπει την αξιολόγηση των αποτελεσμάτων και τον προσδιορισμό των βασικών χαρακτηριστικών τους. Πιο συγκεκριμένα, η ανάπτυξη δέντρων απόφασης αποτελείται από δύο φάσεις, την ανάπτυξη του δέντρου (tree building) και τη περικοπή του (tree pruning). Η φάση της ανάπτυξης του δέντρου αποτελείται από την επαναλαμβανόμενη κατάτμηση

(recursive partitioning) των συνόλων εκμάθησης σύμφωνα με τις τιμές των χαρακτηριστικών τους. Η διαδικασία αυτή συνεχίζει μέχρι οι μεταβλητές του υποσυνόλου ενός κόμβου να έχουν την ίδια τιμή στην εξαρτημένη μεταβλητή, ή όταν η διχοτόμηση δεν προσθέτει επιπλέον αξία στις προβλέψεις. Η φάση περικοπής του δέντρου αφορά στην επιλογή και αφαίρεση των κλαδιών που περιέχουν το μεγαλύτερο ποσοστό λάθους. Η φάση αυτή βελτιώνει την ακρίβεια πρόβλεψης του δέντρου απόφασης, και μειώνει την πολυπλοκότητά του (Au et al. 2003).

Τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για:

- Κατάτμηση (Segmentation): αναγνώριση περιπτώσεων που είναι πιθανό να ανήκουν σε μία συγκεκριμένη ομάδα.
- Διαστρωμάτωση (Stratification): ανάθεση περιπτώσεων σε μία κατηγορία, όπως ομάδες υψηλού, μέτριου, και χαμηλού κινδύνου.
- Πρόβλεψη: δημιουργία κανόνων και χρήση τους για την πρόβλεψη μελλοντικών γεγονότων, όπως η πιθανότητα απώλειας πελατών.
- Μείωση μεταβλητών και εξέταση μεταβλητών προς ταξινόμηση (variable screening): επιλογή ενός χρήσιμου υποσυνόλου μεταβλητών από ένα μεγαλύτερο, για το χτίσιμο ενός μοντέλου.
- Αναγνώριση αλληλεπιδράσεων: αναγνώριση σχέσεων που αναφέρονται σε συγκεκριμένες υποομάδες δεδομένων και ορισμός τους σε ένα μοντέλο.

2.2.1. Τύποι δέντρων απόφασης

Οι αλγόριθμοι που χρησιμοποιούνται για την κατασκευή δέντρων απόφασης λειτουργούν επιλέγοντας μια μεταβλητή σε κάθε βήμα, η οποία είναι η επόμενη καλύτερη μεταβλητή προς χρήση για τη διαίρεση του συνόλου των δεδομένων. Ο όρος «η καλύτερη μεταβλητή» ορίζεται από το πόσο καλά η μεταβλητή διαχωρίζει το σύνολο σε υποσύνολα που έχουν την ίδια τιμή της εξαρτημένης μεταβλητής. Υπάρχουν διάφορες μέθοδοι για τη δημιουργία μοντέλων δέντρων, όπως:

CHAID: CHi-squared Automatic Interaction Detection. Πραγματοποιεί πολλαπλών επιπέδων διαιρέσεις, και σε κάθε βήμα η μέθοδος επιλέγει την ανεξάρτητη μεταβλητή που έχει την πιο ισχυρή αλληλεπίδραση με την εξαρτημένη μεταβλητή. Οι κατηγορίες κάθε ανεξάρτητης μεταβλητής συγχωνεύονται αν δεν είναι σημαντικά διαφορετικές αναφορικά με την εξαρτημένη μεταβλητή (Kass, 1980).

CRT: Classification and Regression Tree. Η μέθοδος CRT διαιρεί τα δεδομένα σε ομογενή τμήματα αναφορικά με την εξαρτημένη μεταβλητή, ώστε στην καλύτερη περίπτωση όλες οι περιπτώσεις σε ένα τελικό κόμβο να έχουν την ίδια τιμή για την εξαρτημένη μεταβλητή (ομογενής κόμβος) (Breiman et al. 1984).

QUEST: Quick, Unbiased, Efficient Statistical Tree. Μία μέθοδος γρήγορη που αποφεύγει την μεροληψία (bias) των άλλων μεθόδων υπέρ των ανεξάρτητων μεταβλητών με πολλές κατηγορίες. Η μέθοδος αυτή μπορεί να οριστεί μόνο αν η εξαρτημένη μεταβλητή είναι ονομαστική (Loh et al. 1997).

2.2.2. Πλεονεκτήματα

- Απλά στην κατανόηση και ερμηνεία, αφού χρησιμοποιούν ένα μοντέλο λευκού κουτιού (σε αντίθεση με τα τεχνητά νευρωνικά δίκτυα που λειτουργούν ως μαύρα κουτιά).
- Απαιτούν μικρή προετοιμασία δεδομένων. Άλλες τεχνικές συνήθως απαιτούν κανονικοποίηση δεδομένων, δημιουργία ψευδομεταβλητών κ.α.
- Μπορούν να διαχειριστούν τόσο αριθμητικά όσο και κατηγορικά δεδομένα. Άλλες τεχνικές συνήθως ειδικεύονται στην ανάλυση συνόλων δεδομένων που έχουν μόνο έναν τύπο μεταβλητών (οι κανόνες συσχέτισης μπορούν να χρησιμοποιηθούν μόνο με ονομαστικές μεταβλητές ενώ τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν μόνο με αριθμητικές μεταβλητές).
- Είναι πιθανόν να ελεγχθεί η αξιοπιστία ενός μοντέλου δέντρου απόφασης με χρήση στατιστικών ελέγχων.
- Ευρωστία: Αποδίδουν καλά ακόμα και αν τα δεδομένα δεν είναι αντιπροσωπευτικά του πραγματικού μοντέλου.
- Αποδίδουν καλά στην ανάλυση μεγάλου όγκου δεδομένων σε λίγο χρόνο.

2.2.3. Περιορισμοί

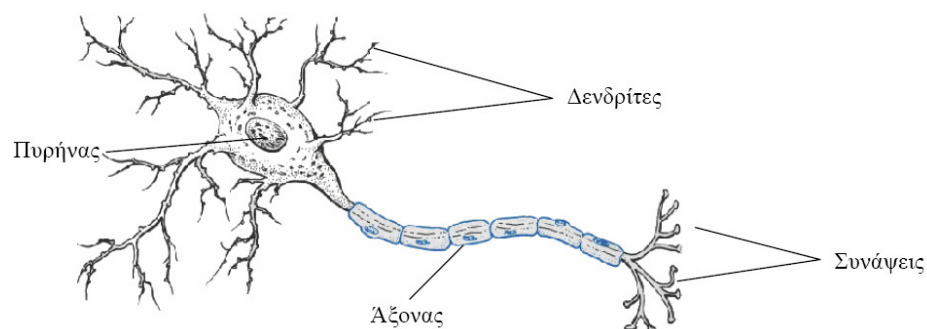
- Οι αλγόριθμοι εκπαίδευσης δέντρων απόφασης βασίζονται σε ευριστικούς κανόνες οι οποίοι οδηγούν σε τοπικά βέλτιστες αποφάσεις σε κάθε κόμβο. Τέτοιοι αλγόριθμοι δε μπορούν να εγγυηθούν την επιστροφή ενός ολικά βέλτιστου δέντρου απόφασης.

- Πολλές φορές μέσα από την εκτενή εκπαίδευση δέντρων απόφασης δημιουργούνται σύνθετα δέντρα χωρίς δυνατότητες γενίκευσης. Αυτό το πρόβλημα λέγεται υπερπροσαρμογή (overfitting)¹.

2.3. Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα αναπτύχθηκαν μέσα από έρευνα στην Τεχνητή Νοημοσύνη, και συγκεκριμένα, από προσπάθειες να γίνει μίμηση της ανοχής στο ελάττωμα και της ικανότητας μάθησης των βιολογικών νευρωνικών συστημάτων με τη μοντελοποίηση της χαμηλού επιπέδου δομής του εγκεφάλου.

Ο εγκέφαλος αποτελείται από έναν πολύ μεγάλο αριθμό νευρώνων που συνδέονται μεταξύ τους. Κάθε νευρώνας έχει μια διακλαδισμένη δομή εισόδου (dendrites), ένα σώμα κυττάρων, και μια δομή εξόδου (axon). Ο άξονας ενός κυττάρου συνδέεται με τους δενδρίτες ενός άλλου μέσω της σύναψης (αποτελεί το βάρος w_i , δηλαδή τη «σημαντικότητα» της αντίστοιχης εισόδου) η οποία μεταφέρει τα σήματα (ερεθισμούς) u_i μέσα στο νευρώνα. Όταν ένας νευρώνας ενεργοποιείται, στέλνει ένα ηλεκτροχημικό σήμα κατά μήκος του άξονα. Αυτό το σήμα διασχίζει τις συνάψεις προς άλλους νευρώνες, οι οποίοι μπορούν στη συνέχεια να ενεργοποιηθούν, παράγοντας ένα σήμα εξόδου z . Ένας νευρώνας ενεργοποιείται μόνο εάν το συνολικό σήμα εισόδου (σταθμισμένο άθροισμα $w_i \cdot u_i$) υπερβαίνει ένα ορισμένο επίπεδο (το κατώφλι ενεργοποίησης θ). Οι τιμές της εξόδου είναι συνεχείς ή δυαδικές (σχήμα 11).

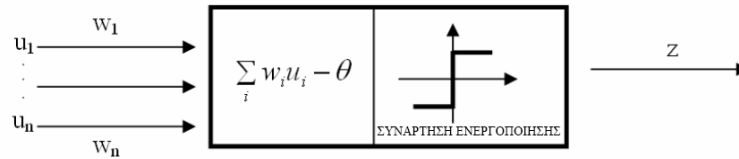


Σχήμα 11: Βιολογικός Νευρώνας

http://sourcetherapies.com/shift_moleculesofemotions.html

¹ (<http://en.wikipedia.org/wiki/Overfitting>)

Σε μαθηματικούς όρους ο νευρώνας μπορεί να περιγραφεί ως εξής (<http://www.statsoft.com/textbook/neural-networks/>):



$$y = \sum_i w_i u_i - \theta \text{ και } z = \psi(y)$$

όπου $\psi()$ είναι η συνάρτηση ενεργοποίησης (ή συνάρτηση μεταφοράς), η οποία καθορίζει τη μορφή των τιμών εξόδου του νευρώνα. Οι βασικές συναρτήσεις που χρησιμοποιούνται είναι:

Η σιγμοειδής συνάρτηση (sigmoid function):

$$\psi(x) = \frac{1}{1 + e^{-a(x)}}$$

Η βηματική συνάρτηση (threshold function):

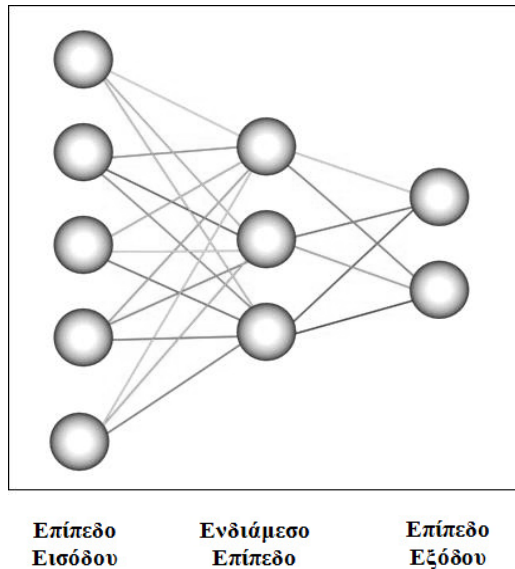
$$\psi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Η κατά τμήματα γραμμική συνάρτηση (piecewise linear function):

$$\psi(x) = \begin{cases} 1, & x \geq \frac{1}{2} \\ x, & -\frac{1}{2} < x < \frac{1}{2} \\ 0, & x < 0 \text{ ή } x \leq -\frac{1}{2} \end{cases}$$

Άλλες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται είναι η γραμμική και η γκαουσιανή συνάρτηση. Από όλες τις συναρτήσεις αυτές, η σιγμοειδής αποτελεί τη δημοφιλέστερη, καθώς έχει τη δυνατότητα γραμμικής και μη γραμμικής συμπεριφοράς ενώ είναι συνεχής και παραγωγίσιμη.

Ένα τεχνητό νευρωνικό δίκτυο αποτελείται από ένα σύνολο κόμβων συνδεδεμένων μεταξύ τους σε στρώματα ή επίπεδα (layers), το επίπεδο εισόδου (input layer), το/τα ενδιάμεσο/α επίπεδο/α (hidden layer) και το επίπεδο εξόδου (output layer) (το πλήθος των ενδιάμεσων επιπέδων μεταβάλλεται ανάλογα με την εφαρμογή) (σχήμα 12) (http://en.wikipedia.org/wiki/Neural_network).



Σχήμα 12: Τυπική δομή νευρωνικού δικτύου με ένα κρυμμένο επίπεδο
http://prlab.ceid.upatras.gr/courses/simeiwseis/NNET_HTML/Chapter2/NNET_chapter24.htm

Η σωστή σχεδίαση του δικτύου (διευθέτηση κόμβων και συνδέσεων) είναι ιδιαίτερα σημαντική, καθώς υπάρχει μια έντονη συσχέτιση μεταξύ του αλγορίθμου εκμάθησης και της τοπολογίας του δικτύου, γεγονός που κάνει τη σχεδίαση πρωταρχικής σημασίας (Haykin, 1994).

Τα τεχνητά νευρωνικά δίκτυα μπορούν να χωριστούν σε δύο βασικές κατηγορίες:

- **Feed-forward:** η πληροφορία ρέει από το επίπεδο εισόδου προς το επίπεδο εξόδου και κυρίως χρησιμοποιείται σε θέματα όπως η πρόβλεψη και η αναγνώριση προτύπων. Μια τέτοια δομή έχει σταθερή συμπεριφορά.
- **Recurrent:** η πληροφορία επιστρέφει σε προηγούμενο επίπεδο που με τη σειρά του ανατροφοδοτεί το ενδιάμεσο επίπεδο και κυρίως χρησιμοποιείται για τη συνειρμική μνήμη και για προβλήματα βελτιστοποίησης. Μια τέτοια δομή μπορεί να είναι ασταθής

Μόλις καθοριστεί η αρχιτεκτονική του νέου δικτύου, υπόκειται στη διαδικασία της εκπαίδευσης. Σε αυτή τη φάση οι νευρώνες εφαρμόζουν μια επαναληπτική διαδικασία στον αριθμό των εισόδων για να προσαρμόσουν τα βάρη w_i των συνάψεων ώστε το δίκτυο να είναι σε θέση να λύσει ένα πρόβλημα. Στόχος της εκπαίδευσης είναι η ελαχιστοποίηση του σφάλματος και η προσαρμογή των τιμών

των βαρών με τέτοιο τρόπο ώστε όταν εισάγεται στο δίκτυο μια είσοδος, να παράγεται η σωστή έξοδος.

Τα νευρωνικά δίκτυα εκπαιδεύονται κυρίως με τρεις διαφορετικούς τρόπους:

- Εκπαίδευση με εποπτεία (supervised): Το σύνολο προτύπων περιέχει ένα διανύσματα εισόδου και εξόδου και το δίκτυο μαθαίνει να συμπεραίνει τη σχέση μεταξύ των δύο. Το νευρωνικό δίκτυο εκπαιδεύεται χρησιμοποιώντας έναν από τους κατάλληλους αλγορίθμους μάθησης (με τον πιο διαδεδομένο να είναι ο αλγόριθμος «προς τα πίσω διάδοσης», back propagation, από τους Rumelhart et al. 1986) (Patterson 1996, Haykin 1994, Fausett 1994).
- Εκπαίδευση χωρίς εποπτεία (unsupervised): Το σύνολο προτύπων περιέχει μόνο διανύσματα εισόδου τα οποία ομαδοποιούνται με βάση τα κοινά τους χαρακτηριστικά, ενώ το νευρωνικό δίκτυο προσπαθεί να μάθει τη δομή των δεδομένων.
- Εκπαίδευση με ανταγωνισμό: Οι νευρώνες ανταγωνίζονται για το ποιος θα συμμετάσχει πιο ενεργά στην εκπαίδευση, με κάποιον από αυτούς τελικά να χαρακτηρίζεται «νικητής νευρώνας». Στόχος εδώ είναι η αναπαράσταση μιας κλάσης διανυσμάτων εισόδου από ένα και μοναδικό νευρώνα εξόδου. Με τον τρόπο αυτό η κλάση του κάθε διανύσματος ταυτοποιείται από τον μοναδικό νευρώνα που ενεργοποιείται σε κάθε περίπτωση.

Εάν το δίκτυο εκπαιδευτεί κατάλληλα, έχει μάθει να μοντελοποιεί την (άγνωστη) συνάρτηση που συσχετίζει τις μεταβλητές εισόδου με τις μεταβλητές εξόδου, και μπορεί στη συνέχεια να χρησιμοποιηθεί για να κάνει προβλέψεις εκεί όπου η έξοδος δεν είναι γνωστή.

2.3.1. Πλεονεκτήματα & Μειονεκτήματα

Τα νευρωνικά δίκτυα έχουν εφαρμοστεί με επιτυχία σε διάφορους τομείς όπως η οικονομία, η ιατρική, η εφαρμοσμένη μηχανική, η γεωλογία και η φυσική, και γενικά όπου υπάρχουν προβλήματα πρόβλεψης, ταξινόμησης ή ελέγχου. Η επιτυχία αυτή μπορεί να αποδοθεί σε πολλούς παράγοντες, όπως για παράδειγμα το γεγονός ότι μπορούν να παρέχουν μια πρόβλεψη με την πιθανότητά της, σε αντίθεση με άλλες τεχνικές ταξινόμησης όπως τα δέντρα απόφασης. Οι Fadalla et al. (2001) μετά από σχετική έρευνα καταλήγουν στο συμπέρασμα ότι στις περισσότερες συγκριτικές μελέτες τα νευρωνικά δίκτυα ξεπερνούν τα στατιστικά-οικονομετρικά μοντέλα.

Έρευνες δείχνουν ότι τα νευρωνικά δίκτυα αποδίδουν καλύτερα από τα δέντρα απόφασης και τη λογιστική παλινδρόμηση σε προβλήματα πρόβλεψης απώλειας πελατών (Au et al. 2003).

Πλεονεκτήματα Νευρωνικών Δικτύων	Μειονεκτήματα Νευρωνικών Δικτύων
Υψηλή ακρίβεια: τα νευρωνικά δίκτυα είναι σε θέση να επιλύουν ικανοποιητικά μη γραμμικά προβλήματα.	Έλλειψη διαφάνειας/κατανόησης μοντέλων: τα νευρωνικά δίκτυα λειτουργούν ως «μαύρα κουτιά».
Ανοχή στο θόρυβο: τα νευρωνικά δίκτυα είναι ευέλικτα σε θορυβώδη, ελλιπή δεδομένα.	Σχεδίαση δοκιμής και λάθους: η επιλογή των ενδιάμεσων κόμβων και η των παραμέτρων εκπαίδευσης είναι εμπειρική.
Ανεξαρτησία από υποθέσεις: τα νευρωνικά δίκτυα δεν κάνουν υποθέσεις για την κατανομή των δεδομένων, ή τη μορφή των αλληλεπιδράσεων μεταξύ των παραγόντων.	Ο υπολογισμός των βαρών του δικτύου απαιτεί μεγάλα ποσά δεδομένων, και αυτό συνεπάγεται μεγάλο υπολογιστικό κόστος.
Ευκολία συντήρησης: τα νευρωνικά δίκτυα μπορούν να ενημερωθούν με νέα δεδομένα, είναι συνεπώς ιδανικά για δυναμικά περιβάλλοντα.	Over-fitting: αν χρησιμοποιηθούν πάρα πολλά βάρη χωρίς κανονικοποίηση, αχρηστεύεται η δυνατότητα των νευρωνικών δικτύων να γενικεύουν σε νέα δεδομένα.
Τα νευρωνικά δίκτυα υπερνικούν κάποιους περιορισμούς άλλων στατιστικών μεθόδων ενώ τις γενικεύουν.	Δεν υπάρχει ρητό σύνολο κανόνων για να επιλεγεί ο πιο κατάλληλος αλγόριθμος νευρωνικού δικτύου.
Οι κρυμμένοι κόμβοι στα νευρωνικά δίκτυα με εποπτευμένη εκπαίδευση μπορούν να θεωρηθούν ως λανθάνουσες μεταβλητές.	Τα νευρωνικά δίκτυα εξαρτώνται σημαντικά από την ποσότητα και την ποιότητα των διαθέσιμων δεδομένων.
Τα νευρωνικά δίκτυα είναι μη γραμμικές τεχνικές μοντελοποίησης ικανές να μοντελοποιούν ιδιαίτερα σύνθετες συναρτήσεις.	Τα νευρωνικά δίκτυα μπορεί να συγκλίνουν σε τοπικό και όχι ολικό ελάχιστο.
Η απόδοση των νευρωνικών δικτύων μπορεί να αυτοματοποιηθεί, ελαχιστοποιώντας την ανθρώπινη παρουσία.	Τα νευρωνικά δίκτυα δεν έχουν κλασικές στατιστικές ιδιότητες, όπως τα διαστήματα εμπιστοσύνης και ο έλεγχος υπόθεσης.
Τα νευρωνικά δίκτυα είναι κατάλληλα στην αντιμετώπιση προβλημάτων σε μη-συμβατικά πεδία.	Οι τεχνικές νευρωνικών δικτύων συνεχίζουν να εξελίσσονται και δεν είναι ακόμα εύρωστες.

Πίνακας 1: Πλεονεκτήματα και μειονεκτήματα νευρωνικών δικτύων

2.4. Μηχανές διανυσμάτων υποστήριξης

Οι Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM) αποτελούν μια πρόσφατη προσέγγιση στο πρόβλημα της ταξινόμησης (Vapnik, 1998). Είναι ένας αλγόριθμος αναγνώρισης προτύπων, που διαφοροποιείται από τους προγενέστερους του, και αποτελεί ένα σύστημα εκμάθησης που χρησιμοποιεί γραμμικές διακρίνουσες συναρτήσεις σε ένα μετασχηματισμένο υψηλών διαστάσεων χώρο. Οι Μηχανές διανυσμάτων υποστήριξης χρησιμοποιούνται κυρίως για δυαδική κατηγοριοποίηση, χωρίς βέβαια να αποκλείεται η χρήση τους για ένταξη δεδομένων σε περισσότερες κατηγορίες, μετά την τροποποίησή τους, αφού έχουν τη δυνατότητα να προσαρμόζονται ανάλογα με την περίπτωση και τη φύση των δεδομένων προς ταξινόμηση (Burges 1998, Mangasarian 2000, Τσιώρης 2008). Τα μοντέλα αυτά μπορεί να είναι γραμμικά, αλλά και μη γραμμικά, και σκοπό έχουν τη μεγιστοποίηση του διαχωριστικού περιθωρίου μεταξύ αντικειμένων που ανήκουν σε διαφορετικές κατηγορίες. Τα γραμμικά μοντέλα μηχανών διανυσμάτων υποστήριξης είναι εύκολα στην εφαρμογή, ακόμα και σε μεγάλα σύνολα δεδομένων, και επιπλέον είναι εύκολα να ερμηνευτούν και να κατανοηθούν. Αντίθετα, τα μη γραμμικά μοντέλα μηχανών διανυσμάτων υποστήριξης χρησιμοποιούνται σε περιπτώσεις που ένας κανόνας γραμμικού διαχωρισμού δεν είναι αρκετός για τη δεδομένη περίπτωση ταξινόμησης. Η κατασκευή όμως μη γραμμικών μοντέλων απαιτεί αυξημένους υπολογιστικούς πόρους για μεγάλα σύνολα δεδομένων, και η δυνατότητα ερμηνείας τους είναι σχετικά περιορισμένη.

2.4.1. Προσεγγιστικές μηχανές διανυσμάτων υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης έχουν γίνει μια δημοφιλής, μη παραμετρική μεθοδολογία για την ανάπτυξη προβλημάτων ταξινόμησης. Σε ένα διχοτομικό περιβάλλον ταξινόμησης, η ανάπτυξη ενός μοντέλου αρχίζει με ένα δείγμα εκπαίδευσης $T = \{x_i, d_i\}$, $i=1, 2, \dots, n$, όπου το $x_i \in R^m$ είναι το διάνυσμα διάστασης m με τις πληροφορίες εισόδου για το εκπαιδευόμενο αντικείμενο i , ενός συνόλου ανεξάρτητων μεταβλητών m , και το $d_i \in \{-1, +1\}$ είναι το αντίστοιχο αποτέλεσμα εξόδου ανάλογα με την κατηγορία στην οποία ανήκει το x_i (εξαρτημένη μεταβλητή). Το αντικείμενο της ανάλυσης είναι η κατασκευή μιας συνάρτησης

απόφασης $f(x) \rightarrow d$ που να διαχωρίζει τις δύο κατηγορίες ταξινόμησης. Στην απλούστερη περίπτωση, η συνάρτηση $f(x)$ ορίζεται από το υπερεπίπεδο $xw = \gamma$ ως εξής:

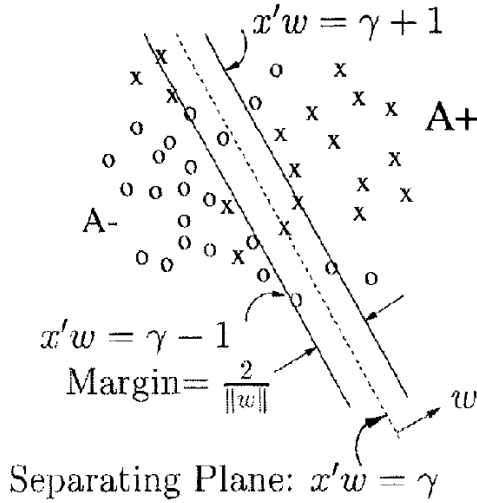
$$f(x) = \text{sgn}(xw - \gamma)$$

όπου w είναι το κάθετο στο υπερεπίπεδο διάνυσμα και γ είναι ένας σταθερός όρος. Από τη στιγμή που η f είναι σταθερή για κάθε θετική αύξηση (rescaling) του όρου μέσα στη συνάρτηση sign , ορίζεται το κανονικό υπερεπίπεδο (canonical hyperplane) $d(xw - \gamma) \geq 1$ ως το υπερεπίπεδο εκείνο που θέτει σε «συναρτησιακή απόσταση» μεγαλύτερη ή ίση της μονάδας τα διανύσματα των δύο κατηγοριών από το υπερεπίπεδο. Η βέλτιστη συνάρτηση απόφασης f είναι αυτή που μεγιστοποιεί το περιθώριο (margin) που προκαλείται από τον διαχωρισμό των κατηγοριών (μεγιστοποιεί την απόσταση του υπερεπιπέδου ταξινόμησης από τα κοντύτερα σημεία των δύο κλάσεων), το οποίο είναι $\frac{2}{\|w\|_2}$ (Vapnik, 1998). Συνεπώς, δεδομένου ενός

δείγματος εκπαίδευσης με n παρατηρήσεις, η μεγιστοποίηση του περιθωρίου μπορεί να πραγματοποιηθεί με τη λύση του ακόλουθου προβλήματος τετραγωνικού προγραμματισμού (1):

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C e^T y \\ \text{υ.π.} \quad & D(Xw - e\gamma) + y \geq e \\ & y \geq 0, \quad w, \gamma \in R \end{aligned}$$

όπου D είναι ένας $n \times n$ πίνακας τέτοιος ώστε $D_{ii} = d_i$ και $D_{ij} = 0, \forall i \neq j$, το X είναι ένας $n \times m$ πίνακας με τα δεδομένα εκπαίδευσης, το e είναι ένα διάνυσμα με μονάδες, το y είναι ένα $n \times 1$ διάνυσμα με θετικές μεταβλητές απόκλισης (slack variables) που σχετίζονται με τη λανθασμένη ταξινόμηση των αντικειμένων εκπαίδευσης όταν οι κατηγορίες δεν είναι γραμμικά διαχωρίσιμες, και το $C > 0$ είναι μια παράμετρος που χρησιμοποιείται για να τιμωρήσει τα σφάλματα ταξινόμησης.



Σχήμα 13: Μηχανές διανυσμάτων υποστήριξης, η γραμμικά μη διαχωρίσιμη περίπτωση (Mangasarian, 2001).

Από υπολογιστική οπτική γωνία, αντί της επίλυσης του προβλήματος (1), είναι πιο συμφέρον να ληφθεί υπόψη η δυική του λαγκρασιανή διατύπωση (2):

$$\max e^T u - \frac{1}{2} u^T D X X^T D u$$

$$\text{υ.π. } e^T D u = 0$$

$$0 \leq u \leq C e$$

Η συνάρτηση απόφασης κατά συνέπεια εκφράζεται σε όρους των δυικών μεταβλητών u ως εξής:

$$f(x) = \text{sgn}(x X^T D u - \gamma)$$

Μέχρι το σημείο αυτό, έχει παρουσιαστεί ο τρόπος με τον οποίο οι μηχανές διανυσμάτων υποστήριξης μπορούν να διαχωρίσουν δυο κλάσεις, τα όρια των οποίων είναι ή τείνουν να είναι γραμμικά. Στις περιπτώσεις που τα όρια είναι έντονα μη γραμμικά, οι μη γραμμικές μηχανές διανυσμάτων υποστήριξης δίνουν τη λύση. Προκειμένου να γενικευτεί το μοντέλων των γραμμικών μηχανών διανυσμάτων υποστήριξης σε ένα μη γραμμικό, τα δεδομένα εισόδου του προβλήματος αναπαριστώνται σε ένα υψηλότερων διαστάσεων χώρο H (χώρος χαρακτηριστικών) μέσα από μια μετατροπής της μορφής: $x_i x_j^T \rightarrow \phi(x_i) \phi^T(x_j)$. Η συνάρτηση ϕ που μεταφέρει τα δεδομένα στο χώρο υψηλότερων διαστάσεων ορίζεται εν δυνάμει από μια συμμετρική, θετικά ορισμένη συνάρτηση πυρήνα (kernel function)

$K(x_i, x_j) = \phi(x_i)\phi^T(x_j)$. Δημοφιλείς επιλογές για τη συνάρτηση πυρήνα συμπεριλαμβάνουν τον πυρήνα polynomial, τον πυρήνα Radial basis function (RBF), το σιγμοειδή πυρήνα κτλ (Scholkopf, et al. 2002). Η αναπαράσταση των δεδομένων που χρησιμοποιεί τη συνάρτηση πυρήνα επιτρέπει την ανάπτυξη ενός γραμμικού μοντέλου στο χώρο των χαρακτηριστικών H , στον οποίο δηλαδή τα δεδομένα είναι γραμμικά διαχωρίσιμα. Από τη στιγμή όμως που ο χώρος H είναι μια μη γραμμική χαρτογράφηση των αρχικών δεδομένων, το μοντέλο που αναπτύχθηκε συνεπώς είναι μη γραμμικό στον αρχικό χώρο εισόδων. Το μοντέλο τελικά αναπτύσσεται με την εφαρμογή της παραπάνω γραμμικής ανάλυσης στον χώρο χαρακτηριστικών H .

Για μεγάλα σύνολα εκπαίδευσης, μερικές υπολογιστικές διαδικασίες έχουν προταθεί ώστε να επιτρέψουν τη γρήγορη εκπαίδευση των μοντέλων μηχανών διανυσμάτων υποστήριξης. Εδώ χρησιμοποιήθηκε η μεθοδολογία των προσεγγιστικών μηχανών διανυσμάτων υποστήριξης (Fung et al. 2001). Σε αυτή τη μεθοδολογία, το βασικό πρόβλημα (1) μετατρέπεται στο ακόλουθο πρόβλημα βελτιστοποίησης (3):

$$\min \frac{1}{2} (w^T w + \gamma^2) + \frac{1}{2} Cy^T y$$

$$\text{υ.π. } D(Xw - e\gamma) + y = e$$

Αυτή η νέα διατύπωση οδηγεί σε έναν γρήγορο και απλό αλγόριθμο για την παραγωγή ενός γραμμικού ή/και μη γραμμικού ταξινομητή. Το βασικό πλεονέκτημά της είναι ότι έχει μόνο περιορισμούς ισότητας, συνεπώς δίνει τη δυνατότητα της διατύπωσης μιας βέλτιστης λύσης κλειστής μορφής άμεσα από τα δεδομένα εκπαίδευσης, χωρίς να απαιτεί τη χρήση κανενός αλγορίθμου τετραγωνικού προγραμματισμού, αλλά μόνο τη λύση ενός συνόλου γραμμικών εξισώσεων, απαιτώντας έτσι λιγότερο υπολογιστικό χρόνο.

2.4.2. Προσθετικές μηχανές διανυσμάτων υποστήριξης

Οι προσθετικές μηχανές διανυσμάτων υποστήριξης (Additive Support Vector Machines, ASVM) συνδυάζουν την απλότητα και διαφάνεια/δυνατότητα ερμηνείας των γραμμικών ταξινομητών με τη δυνατότητα γενίκευσης των μη γραμμικών

μοντέλων (Douplos et al. 2007). Αυτό είναι πολύ σημαντικό σε πολλά θέματα όπου τα συστήματα ταξινόμησης πρέπει όχι μόνο να είναι ακριβή αλλά και κατανοητά.

Τα προσθετικά μοντέλα είναι γενικεύσεις των γραμμικών. Ένα γενικευμένο τέτοιο μοντέλο έχει τη μορφή:

$$f(x) = \sum_{j=1}^n f_j(x_j) - \gamma$$

όπου f_j είναι μία ομαλή συνάρτηση του χαρακτηριστικού x_j .

Η μεθοδολογία των ASVM βασίζεται στο γεγονός ότι κάθε πραγματική συνάρτηση μπορεί να προσεγγιστεί από μία κατά τμήματα γραμμική συνάρτηση με έναν αυθαίρετα μεγάλο αριθμό γραμμικών τμημάτων.

Συνεπώς, ένα ομαλό προσθετικό μοντέλο μπορεί να αναπτυχθεί από το συνδυασμό κατά τμήματα γραμμικών συναρτήσεων, όπου κάθε μία συνάρτηση αναπτύσσεται χρησιμοποιώντας έναν αλγόριθμο απλών μηχανών διανυσμάτων υποστήριξης για την εκπαίδευση γραμμικών ταξινομητών.

Για την ανάπτυξη ενός κατά τμήματα γραμμικού μοντέλου, κάθε χαρακτηριστικό x_j χωρίζεται σε υποδιαστήματα, τα οποία ορίζονται από r_j διαχωριστικά σημεία $k_j^1 < k_j^2 < \dots < k_j^{r_j}$, όπου k_j^1 και $k_j^{r_j}$ είναι αντίστοιχα η ελάχιστη και η μέγιστη τιμή του χαρακτηριστικού από τα δεδομένα εκπαίδευσης. Έτσι η συνολική «βαθμολογία» του πελάτη i είναι:

$$f(x_i) = \sum_{j=1}^n \sum_{t=1}^{r_j} f_j(k_j^t) \bar{x}_{ij}^t - \gamma$$

όπου:

$$\bar{x}_{ij}^t = \begin{cases} \frac{x_{ij} - k_j^{t-1}}{k_j^t - k_j^{t-1}} & \text{εάν } k_j^{t-1} \leq x_{ij} \leq k_j^t \\ \frac{k_j^{t+1} - x_{ij}}{k_j^{t+1} - k_j^t} & \text{εάν } k_j^t \leq x_{ij} \leq k_j^{t+1} \\ 0 & \text{διαφορετικά} \end{cases}$$

Επομένως, το κατά τμήματα γραμμικό μοντέλο μπορεί να διατυπωθεί σε γραμμική μορφή ως $f(x) = b\bar{x} - \gamma$, όπου $\bar{x}$ είναι το διάνυσμα των κωδικοποιημένων δεδομένων και b είναι το διάνυσμα των συντελεστών των μεταβλητών στο μοντέλο, τα στοιχεία του οποίου καθορίζουν τις κατά τμήματα γραμμικές συναρτήσεις f_j : $b_j^t = f_j(k_j^t)$. Στα ASVM η εκτίμηση των παραμέτρων b και γ γίνεται μέσω της διαδικασίας των προσεγγιστικών μηχανών διανυσμάτων υποστήριξης.

Η ανάπτυξη του τελικού ομαλού προσθετικού μοντέλου επιτυγχάνεται διαμορφώνοντας ℓ κατά τμήματα γραμμικά μοντέλα (η ανάλυση των Doumpos et al. 2007, έδειξε ότι αρκεί η ανάπτυξη περίπου 20 κατά τμήματα γραμμικών μοντέλων). Κάθε κατά τμήματα γραμμικό μοντέλο διαμορφώνεται χρησιμοποιώντας διαφορετικά διαχωριστικά σημεία για τα χαρακτηριστικά. Έτσι, για κάθε χαρακτηριστικό x_j διαμορφώνονται ℓ συναρτήσεις $f_j^1(x_j), f_j^2(x_j), \dots, f_j^\ell(x_j)$, ο συνδυασμός των οποίων οδηγεί σε μια ομαλή συνάρτηση:

$$f_j(x_j) = \frac{1}{\ell} \sum_{h=1}^{\ell} f_j^h(x_j)$$

Οι Doumpos et al. (2007) εφάρμοσαν τη μεθοδολογία αυτή σε διάφορα σύνολα δεδομένα και διαπίστωσαν ότι παρέχει καλύτερα αποτελέσματα από άλλες μεθοδολογίες βασισμένες στη θεωρία των μηχανών διανυσμάτων υποστήριξης.

2.4.3. Πλεονεκτήματα

Ακολουθεί η παράθεση των κυριότερων πλεονεκτημάτων των μηχανών διανυσμάτων υποστήριξης (Kim et al. 2005):

- Βασίζονται σε πολύ απλές και ξεκάθαρες ιδέες από τη θεωρία στατιστικής μάθησης (Vapnik, 1995).
- Έχουν πολύ καλή ικανότητα γενίκευσης, ενώ ξεπερνούν σε σημαντικό βαθμό το πρόβλημα υπερπροσαρμογής στα δεδομένα (overfitting).
- Η λύση τους είναι μοναδική, βέλτιστη και ολική αφού η εκπαίδευσή τους γίνεται με την επίλυση ενός προβλήματος τετραγωνικού κυρτού προγραμματισμού.
- Η μέθοδος δεν απαιτεί τη γνώση της στατιστικής κατανομής των δεδομένων.
- Μπορούν να παράγουν περίπλοκα μη γραμμικά μοντέλα, που έχουν συγκεκριμένη συναρτησιακή διατύπωση.
- Έχουν ισχυρό θεωρητικό υπόβαθρο και έτσι μπορεί να ακολουθηθεί η διαδικασία βήμα προς βήμα.
- Σε πληθώρα πραγματικών εφαρμογών, έχουν επιδείξει ισάξια η καλύτερη επίδοση συγκριτικά με άλλες ανταγωνιστικές μεθόδους.
- Το πλήθος των παραμέτρων που απαιτούν ρύθμιση στις μηχανές διανυσμάτων υποστήριξης είναι σημαντικά μικρότερο από το αντίστοιχο άλλων μεθοδολογιών.

2.5. Κριτήρια αξιολόγησης μοντέλων

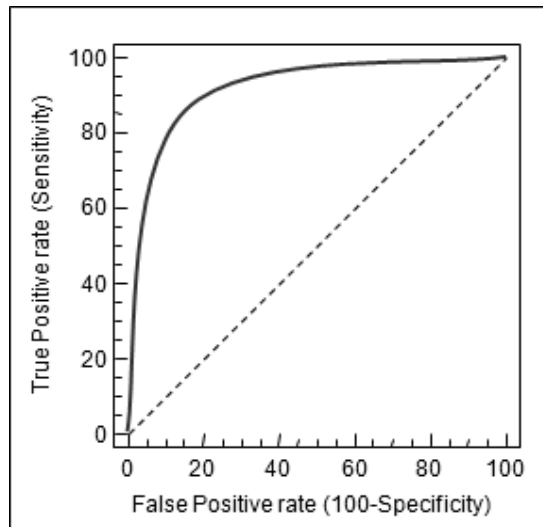
Τα μοντέλα κατάταξης που παρουσιάστηκαν παραπάνω χρησιμοποιούνται για την πρόβλεψη της μελλοντικής συμπεριφοράς απώλειας πελατών, και συνεπώς είναι σημαντικό να γίνει αξιολόγηση της απόδοσης και ακρίβειας της κατάταξής τους. Προκειμένου να ποσοτικοποιηθεί η ακρίβεια των μοντέλων αυτών χρησιμοποιήθηκε η καμπύλη Top decile lift και ο συντελεστής Gini.

2.5.1. ROC Curve

Για να εκτιμηθεί η ικανότητα πρόβλεψης των μοντέλων ταξινόμησης που χρησιμοποιήθηκαν, επιλέχθηκε το κριτήριο της καμπύλης ROC (Receiver Operating Characteristic) (Fawcett, 2004). Το μέτρο αυτό βασίζεται στη σύγκριση της προβλεφθείσας κατάστασης του γεγονότος και της πραγματικής κατάστασης του πελάτη σε σχέση με αυτό το γεγονός. Συγκεκριμένα, η καμπύλη αυτή αποτελείται από ένα διάγραμμα δύο αξόνων, όπου στην περίπτωση της απώλειας πελατών, ο οριζόντιος άξονας δείχνει το ποσοστό των πελατών που μένουν με μια εταιρία και το μοντέλο έχει θεωρήσει ότι την εγκαταλείπουν (false positives, FP), και ο κάθετος άξονας δείχνει το ποσοστό των πελατών που εγκαταλείπουν μια εταιρία και το μοντέλο τους έχει ταξινομήσει στην ίδια κατηγορία (true positives, TP). Η αποτελεσματικότητα του μοντέλου βάσει του διαγράμματος αυτού εξετάζεται από το εμβαδόν της περιοχής κάτω από την καμπύλη (AUC, area under curve):

$$AUC = \int_0^1 \frac{TP}{P} d \frac{FP}{N} = \frac{1}{P * N} \int_0^N TP dFP$$

όπου $TP+FN=P$, και $TN+FP=N$



Σχήμα 14: Καμπύλη ROC (Neslin et al. 2002)

Όσο πιο κυρτωμένη είναι η καμπύλη προς το πάνω αριστερό μέρος του γραφήματος, τόσο καλύτερο είναι το μοντέλο στο να διαχωρίζει με ακρίβεια τους αφοσιωμένους πελάτες από εκείνους που επιθυμούν να εγκαταλείψουν μια εταιρία. Συγκεκριμένα το εμβαδόν για ένα άριστο μοντέλο είναι ίσο με τη μονάδα (συνεχής καμπύλη) και για ένα ανεπαρκές μοντέλο είναι μικρότερο του 0,5 (διακεκομμένη γραμμή) (Douprios 2007).

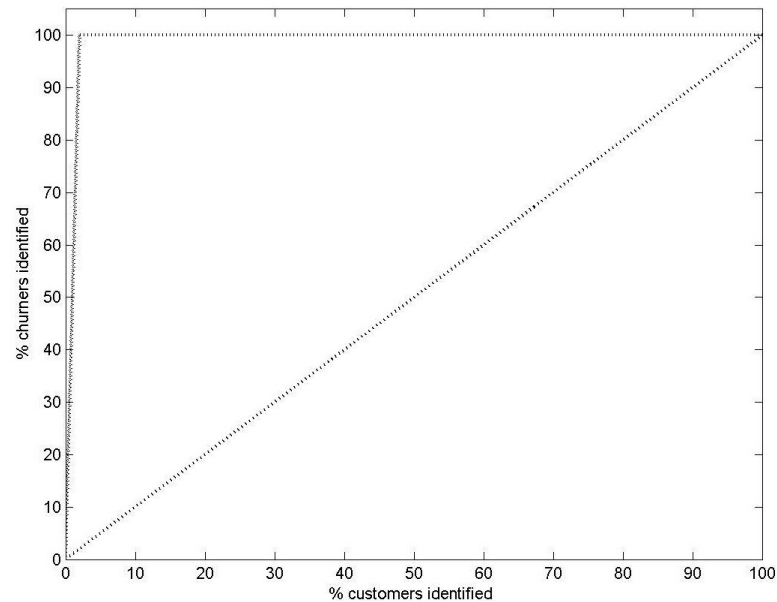
Στη συγκεκριμένη εργασία, χρησιμοποιήθηκε ο συντελεστής Gini, ο οποίος σχετίζεται στενά με την καμπύλη ROC ως εξής (Alker, 1965):

$$Gini = 2 * AUC - 1$$

2.5.2. Καμπύλη Top Decile Lift

Η καμπύλη top-decile lift μετρά τον αριθμό των πελατών που πραγματικά εγκαταλείπουν μια εταιρία, και οι οποίοι εντάσσονται στο 10% των πελατών που το μοντέλο έχει υπολογίσει ότι εμφανίζουν τη μεγαλύτερη πιθανότητα απώλειας. Όσο υψηλότερη η κλίση της καμπύλης, τόσο πιο ακριβές είναι και το μοντέλο. Στο σχήμα 8 φαίνεται μια καμπύλη lift με τέλειο διαχωρισμό των κατηγοριών των πελατών που εγκαταλείπουν μία εταιρία από εκείνους που παραμένουν αφοσιωμένοι σε αυτή (καμπύλη πάνω αριστερά): όλοι οι πελάτες που πρόκειται να φύγουν εντοπίζονται από το μοντέλο πρόβλεψης. Επίσης αναπαρίσταται και μία κατάσταση όπου κανένας διαχωρισμός μεταξύ των πελατών δεν έχει γίνει (μεσαία καμπύλη), πράγμα που

συμβαίνει όταν οι πιθανότητες απώλειας πελατών είναι τυχαίες. Ο στόχος συνεπώς είναι να παραχθεί μία καμπύλη όσο πιο μακριά γίνεται από την τυχαία καμπύλη lift (Neslin et al. 2004).



Σχήμα 15: Καμπύλη Lift για την απεικόνιση τέλει διαχωρισμού (πάνω αριστερά καμπύλη) και μη διαχωρισμού (μεσαία καμπύλη) πελατών που επιθυμούν να εγκαταλείψουν μια εταιρία και μη (Mutanen, 2006).

Η καμπύλη Top Decile Lift προτιμάται για την αξιολόγηση και τη σύγκριση της απόδοσης μοντέλων στην ανάλυση απώλειας πελατών. Ο λόγος για τον οποίο η καμπύλη αυτή χρησιμοποιείται ευρέως στον τομέα των τηλεπικοινωνιών είναι επειδή τα κέντρα εξυπηρέτησης πελατών έχουν ορισμένο προσωπικό διαθέσιμο για να επικοινωνεί με έναν ορισμένο αριθμό συνδρομητών. Οι τηλεπικοινωνιακές εταιρίες συνεπώς ενδιαφέρονται να υπολογίσουν ποιο τμήμα των πελατών που επιθυμούν να εγκαταλείψουν την επιχείρηση θα εντοπίσουν αν επικοινωνήσουν με το 10% της πελατειακής βάσης τους (Mutanen, 2006).

2.6. Εφαρμογές στην ανάλυση και πρόβλεψη απώλειας πελατών

Σκοπός αυτού του υποκεφαλαίου είναι η ανασκόπηση δημοφιλών μεθόδων και εφαρμογών που εντοπίστηκαν στη βιβλιογραφία της ανάλυσης και πρόβλεψης απώλειας πελατών. Μεταξύ των γνωστών μοντέλων για την πρόβλεψη της απώλειας

πελατών είναι τα νευρωνικά δίκτυα (Mozar et al. 2000), οι μηχανές υποστήριξης διανυσμάτων (Zhao et al. 2005), τα δέντρα απόφασης (Lemmens et al. 2003), η λογιστική παλινδρόμηση (Baesens et al. 2002, Coussement et al. 2008, Hung et al. 2006) και η γενετική μοντελοποίηση (Eiben et al. 1998).

Ακολουθεί μια πιο εκτενής παρουσίαση των μεθόδων και των εφαρμογών τους, όπως αυτές εντοπίστηκαν μετά από βιβλιογραφική ανασκόπηση.

2.6.1. Ανάλυση παλινδρόμησης

Μια προσέγγιση της γραμμικής παλινδρόμησης χρησιμοποίησαν οι Mihelis et al. (2001) προκειμένου να εκτιμήσουν την ικανοποίηση των πελατών, ενώ οι Rust et al. (1993) στην προσπάθειά τους να κάνουν το ίδιο, χρησιμοποίησαν λογιστική παλινδρόμηση για να συνδέσουν την ικανοποίηση πελατών με τα χαρακτηριστικά διατήρησής τους, υποστηρίζοντας ότι η λογιστική συνάρτηση μπορεί να αποδώσει την πιθανότητα διατήρησης των πελατών. Οι Hwang et al. (2004) στη μελέτη τους έδειξαν ότι η λογιστική παλινδρόμηση απέδωσε ικανοποιητικά στην πρόβλεψη της απώλειας πελατών, συγκρινόμενη με τα νευρωνικά δίκτυα και τα δέντρα απόφασης, και τόνισαν ότι η λογιστική παλινδρόμηση είναι το καλύτερο μοντέλο για αυτό το σκοπό, σημειώνοντας παράλληλα ότι πολλοί παράγοντες θα μπορούσαν να επηρεάσουν αυτά τα αποτελέσματα (όπως οι παράμετροι των νευρωνικών δικτύων και τα δεδομένα πάνω στα οποία λαμβάνει χώρα το πείραμα). Οι Kim et al. (2004) χρησιμοποίησαν ένα λογιστικό μοντέλο για να καθορίσουν την απώλεια πελατών στον τομέα των τηλεπικοινωνιών, και οι Neslin et al. (2004) δήλωσαν ότι η λογιστική μοντελοποίηση μπορεί ακόμα και να υπερτερεί έναντι πιο εκλεπτυσμένων τεχνικών, όπως τα νευρωνικά δίκτυα (Buckinx et al. 2005, Lariviere et al. 2005).

2.6.2. Δέντρα απόφασης

Οι Kitayama et al. (2002) χρησιμοποίησαν μια τεχνική βασισμένη στα δέντρα απόφασης για να αναπτύξουν ένα μοντέλο ανάλυσης του προφίλ του πελάτη. Η πελατειακή βάση χωρίστηκε σε τμήματα ομάδων πελατών με βάση το πόσο πολύτιμοι είναι για την επιχείρηση και το δέντρο απόφασης εφαρμόστηκε στα τμήματα αυτά με σκοπό τον καθορισμό των απαραίτητων μέτρων που πρέπει να ληφθούν για να αποφευχθεί η απώλειά πελατών και να αυξηθεί η διατήρησή τους. Πειράματα που έγιναν από τους Hwang et al. (2004) συμπεριελάμβαναν ένα δέντρο

απόφασης, ένα νευρωνικό δίκτυο και ένα μοντέλο λογιστικής παλινδρόμησης. Το δέντρο απόφασης έδειξε σχετικά καλύτερη ακρίβεια σε σχέση με τις άλλες τεχνικές, παρόλα αυτά οι Hwang et al. (2004) τόνισαν ότι τα αποτελέσματα αυτά δεν αποδεικνύουν ότι τα δέντρα απόφασης είναι η καλύτερη μέθοδος για όλες τις περιπτώσεις, γεγονός που υποστηρίζεται και από τους Mozer et al. (2000). Οι Datta et al. (2001) ανέπτυξαν το μοντέλο CHAMP (Churn Analysis Modeling and Prediction), το οποίο χρησιμοποιεί δέντρα απόφασης για να προβλέψει την απώλεια πελατών στον τομέα των τηλεπικοινωνιών.

2.6.3. Τεχνητά νευρωνικά δίκτυα

Οι Mozer et al. (2000) χρησιμοποίησαν ένα μοντέλο λογιστικής παλινδρόμησης, ένα μοντέλο δέντρου απόφασης, και ένα μοντέλο με νευρωνικά δίκτυα, για να προβλέψουν την απώλεια πελατών και να καθορίσουν τα θέλγητρα που πρέπει να δοθούν στους πελάτες μιας αμερικάνικης εταιρίας ασύρματης τηλεφωνίας για να βελτιωθεί η διατήρησή τους και να αυξηθεί η αποδοτικότητα της εταιρίας. Μετά τη σύγκριση των μεθόδων με βάση την καμπύλη lift, προέκυψε ότι τα νευρωνικά δίκτυα απέδωσαν καλύτερα από τις δυο άλλες μεθόδους. Ο Pendharkar (2009) χρησιμοποίησε νευρωνικά δίκτυα που βασίζονται σε γενετικούς αλγορίθμους για να προβλέψει την απώλεια πελατών σε μία υπηρεσία κινητής τηλεφωνίας. Μετά από σύγκριση του μοντέλου αυτού με ένα στατιστικό μοντέλο z-score (χρησιμοποιώντας ως κριτήριο αξιολόγησης την ακρίβεια πρόβλεψης, την καμπύλη top decile lift και την καμπύλη ROC), αποδείχτηκε ότι τα νευρωνικά δίκτυα υπερτερούν έναντι του στατιστικού μοντέλου σε όλα τα κριτήρια, αν και το δεύτερο έχει το πλεονέκτημα του μικρότερου υπολογιστικού φόρτου. Ο Boone (2002) εφάρμοσε τα νευρωνικά δίκτυα στην τμηματοποίηση πελατειακών βάσεων δεδομένων στον τομέα του λιανικού εμπορίου, ενώ ο Rygielski (2002) τα χρησιμοποίησε στη διαχείριση πελατειακών σχέσεων, και μέσα από τη μελέτη του υποστηρίζει ότι τα νευρωνικά δίκτυα παρέχουν ένα πιο ακριβές και εύρωστο μέσο πρόβλεψης, ενώ ενέχουν πιθανό ρίσκο εύρεσης υπο-βέλτιστων λύσεων, και πρέπει να ληφθεί υπόψη το γεγονός ότι δεν παρουσιάζουν σχέδια εύκολα κατανοητά από το χρήστη. Οι Au et al. (2003) έδειξαν ότι τα νευρωνικά δίκτυα υπερτερούν έναντι των δέντρων απόφασης και της λογιστικής παλινδρόμησης στην πρόβλεψη απώλειες πελατών, ενώ επιπλέον, μελέτες έδειξαν ότι τα νευρωνικά δίκτυα που βασίζονται σε

γενετικούς αλγόριθμους υπερτερούν έναντι των παραδοσιακών νευρωνικών δικτύων (Pendharkar 2007, Pendharkar et al. 2006).

2.6.4. Μηχανές υποστήριξης διανυσμάτων

Οι μηχανές υποστήριξης διανυσμάτων είναι μια ακόμα τεχνολογία που μελετάται προς χρήση στη διαχείριση απώλειας πελατών. Οι Lee et al. (2004) δήλωσαν μέσα από τη μελέτη τους ότι, σε ένα τυπικό PC, χρειάστηκαν περίπου 20 ώρες για να αναλυθεί ένα σύνολο δεδομένων 8000 πελατών (χρησιμοποιώντας MATLAB), γεγονός που αποτελεί σημαντικό μειονέκτημα των μηχανών υποστήριξης διανυσμάτων, ενώ η μέθοδος έδωσε παρόμοια αποτελέσματα με τα νευρωνικά δίκτυα. Η μελέτη των Coussement et al. (2008) ερευνά την αποτελεσματικότητα των μηχανών υποστήριξης διανυσμάτων (SVM) στην αναγνώριση απώλειας πελατών σε συνδρομητικές υπηρεσίες. Αυτή η έρευνα εφαρμόζει ένα μοντέλο μηχανών υποστήριξης διανυσμάτων σε έναν βέλγικο εκδοτικό οίκο προκειμένου να προβλέψει την απώλεια πελατών με όσο το δυνατόν υψηλότερη ακρίβεια. Τα αποτελέσματα έδειξαν ότι οι μηχανές υποστήριξης διανυσμάτων δείχνουν καλή απόδοση γενίκευσης όταν εφαρμόζονται σε θορυβώδη δεδομένα μάρκετινγκ. Επίσης φαίνεται ότι μόνο όταν εφαρμόζεται η διαδικασία επιλογής των βέλτιστων παραμέτρων, οι μηχανές υποστήριξης διανυσμάτων υπερτερούν έναντι της λογιστικής παλινδρόμησης, ενώ τα random forests υπερτερούν καθαρά των μηχανών υποστήριξης διανυσμάτων με βάση τα κριτήρια αξιολόγησης που χρησιμοποιήθηκαν (καμπύλη ROC και top decile lift).

2.6.5. Γενετικοί αλγόριθμοι

Σχετική βιβλιογραφία υποστηρίζει ότι η χρήση γενετικών αλγορίθμων προσφέρει υψηλότερη ακρίβεια, έχει όμως μεγαλύτερο υπολογιστικό φόρτο (Hadden et al. 2007). Οι Sun et al. (2004) πρότειναν μια προσέγγιση που χρησιμοποιεί γενετικούς αλγορίθμους, τονίζοντας ότι μπορεί να παρέχει ένα απλό, γενικευμένο και εύρωστο πλαίσιο για την επιλογή καλών χαρακτηριστικών. Οι Datta et al. (2001) ανέπτυξαν μια τεχνική για να προβλέπουν την απώλεια πελατών στον τομέα της κινητής τηλεφωνίας, στην οποία πραγματοποιείται προς τα εμπρός επιλογή χαρακτηριστικών (forward feature selection-FFS), όπου ένα δέντρο απόφασης ταξινομεί τα στοιχεία σύμφωνα με το ποσοστό λάθους τους, και ένας γενετικός αλγόριθμος χρησιμοποιεί λογιστική παλινδρόμηση για να αποτιμήσει κάθε πιθανό

σύνολο χαρακτηριστικών. Επιπλέον η σύγκριση της μεθόδου των δέντρων απόφασης με τη μέθοδο του k-κοντινότερου γείτονα δεν έδειξε σημαντικές διαφορές στην ακρίβεια και απόδοση μεταξύ των μοντέλων. Οι Ng et al. (2001) επίσης πρότειναν την προσέγγιση των δέντρων απόφασης για την επιλογή χαρακτηριστικών, τονίζοντας ότι αυτή η μέθοδος είναι ακριβής και αποδοτική, γεγονός που υποστηρίζουν και οι Crespo et al. (2004).

2.6.6. Άλλες μέθοδοι

Οι αλγόριθμοι που έχουν αναφερθεί μέχρι στιγμής έχουν κάποιους περιορισμούς εξαιτίας της φύσης του προβλήματος της απώλειας πελατών: τα δεδομένα είναι μη ισορροπημένα, δηλαδή ο αριθμός των πελατών που εγκαταλείπουν μια επιχείρηση συνήθως αποτελεί ένα πολύ μικρό μέρος των δεδομένων (Zhao et al. 2005), υπάρχει θόρυβος στα δεδομένα και η πρόβλεψη απώλειας πελατών απαιτεί την κατάταξη των συνδρομητών σύμφωνα με την πιθανότητα που έχουν να αλλάξουν εταιρία (Au et al. 2003). Ορισμένες προσεγγίσεις έχουν προσπαθήσει να αντιμετωπίσουν αυτά τα προβλήματα, όπως οι αλγόριθμοι που βασίζονται στα δέντρα απόφασης, οι οποίοι μπορούν να επεκταθούν ώστε να καθορίζουν την κατάταξη αλλά είναι πιθανό κάποια φύλλα του δέντρου να έχουν όμοιες πιθανότητες και επίσης η προσέγγιση αυτή είναι ευπαθής στο θόρυβο (Xie et al. 2009). Έρευνες από τους Au et al. (2003) έδειξαν ότι αν και η ανακάλυψη κανόνων με βάση τους γενετικούς αλγορίθμους μπορεί να παράγει ακριβή μοντέλα ταξινόμησης, αυτά δεν μπορούν να ορίσουν την πιθανότητα που σχετίζεται με αυτές προβλέψεις. Συνεπώς, οι ίδιοι πρότειναν ένα νέο αλγόριθμο, τον Data Mining by Evolutionary Learning (DMEL) ικανό να εξορύσσει κανόνες σε βάσεις δεδομένων, να χειρίζεται προβλήματα ταξινόμησης όπου η πιθανότητα πρόβλεψης είναι σημαντική, και να αντιμετωπίζει την περίπτωση χαμένων τιμών στα δεδομένα. Ο αλγόριθμος αυτός χρησιμοποιήθηκε για την πρόβλεψη απώλειας πελατών σε μια βάση δεδομένων μιας τηλεπικοινωνιακής εταιρίας στη Μαλαισία, και έγινε σύγκριση της ακρίβειας πρόβλεψής του με ένα μοντέλο νευρωνικών δικτύων και δέντρου απόφασης. Τα αποτελέσματα έδειξαν ότι ο DMEL απέδωσε καλύτερα σε σχέση με τις άλλες μεθόδους.

Οι Xie et al. (2009) παρουσίασαν στη μελέτη τους τη μέθοδο Improved Balanced Random Forests (IBRF) στην πρόβλεψη απώλειας πελατών, όπου ερευνάται η απόδοση της μεθόδου σε σύγκριση με τα νευρωνικά δίκτυα, τα δέντρα απόφασης και τις class-weighted core μηχανές υποστήριξης διανυσμάτων (CWC-SVM). Η

μέθοδος εφαρμόστηκε σε μία κινεζική τράπεζα, και προέκυψε ότι το μοντέλο των IBRF έχει καλύτερα αποτελέσματα πρόβλεψης, τόσο συγκριτικά με τις προαναφερθείσες μεθόδους, όσο και σε σχέση με άλλους αλγορίθμους random forests (balanced random forests και weighted random forests), ενώ αντιμετώπισαν ικανοποιητικά και τους προαναφερθέντες περιορισμούς.

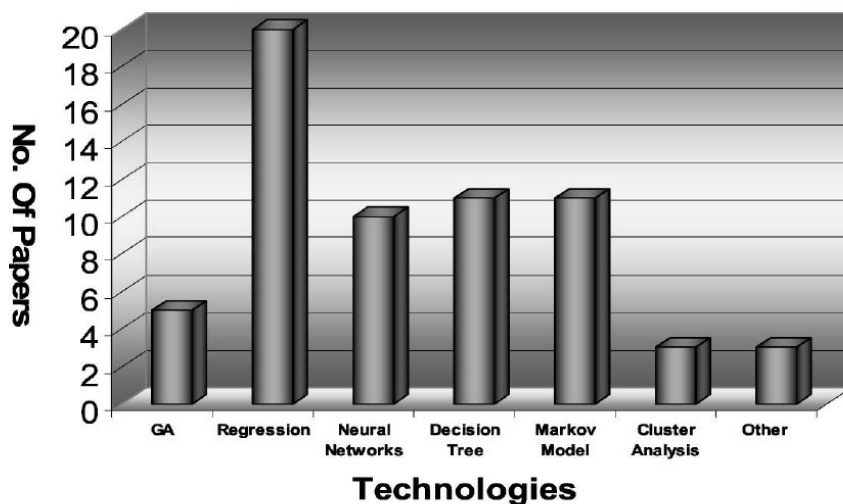
Οι Burez et al. (2009) στη μελέτη τους προσπάθησαν να αντιμετωπίσουν τη μη ισορροπία των δεδομένων στην απώλεια πελατών, χρησιμοποιώντας 4 μεθόδους και κατάλληλα μέτρα απόδοσης (AUC, καμπύλη lift). Χρησιμοποιήθηκαν μοντέλα weighted random forests, δειγματοληψία-sampling (random και advanced under-sampling-CUBE), cost-sensitive learning και boosting, τα οποία εφαρμόστηκαν σε 6 ευρωπαϊκές επιχειρήσεις για να ελεγχθεί η απόδοσή τους (δύο τράπεζες, ένα σούπερ μάρκετ, μια εταιρία κινητής τηλεφωνίας, έναν εκδοτικό οίκο, και μια εταιρία pay TV). Τα αποτελέσματα έδειξαν ότι το μοντέλο under-sampling μπορεί να οδηγήσει σε βελτιωμένη ακρίβεια πρόβλεψης, ενώ δε φαίνεται αύξηση στην απόδοση πρόβλεψης κατά τη χρήση της προχωρημένης τεχνικής δειγματοληψίας CUBE. Επιπλέον τα weighted random forests απέδωσαν καλύτερα από τα απλά random forests, ενώ το boosting αποδείχτηκε εύρωστος ταξινομητής, αλλά δεν απέδωσε καλύτερα από κάποια άλλη τεχνική.

Οι Neslin et al. (2004) ερεύνησαν το πώς μεθοδολογικοί παράγοντες συμβάλλουν στην ακρίβεια των μοντέλων πρόβλεψης απώλειας πελατών. Η μελέτη βασίστηκε σε ένα τουρνουά με 45 συμμετοχές ειδικών οι οποίοι κατασκεύασαν ένα μοντέλο απώλειας πελατών με βάση τα δεδομένα μιας εταιρίας ασύρματης τηλεφωνίας. Η απόδοση των μοντέλων που αναπτύχθηκαν αξιολογήθηκε με βάση την καμπύλη top-decile lift και το συντελεστή Gini. Τα αποτελέσματα έδειξαν ότι η ακρίβεια πρόβλεψης ποικίλει ανάλογα με τη μέθοδο που χρησιμοποιήθηκε, ότι τα μοντέλα μπορούν να χρησιμοποιηθούν επιτυχώς και σε μελλοντικό διάστημα και ότι τα μοντέλα λογιστικής παλινδρόμησης και δέντρων απόφασης έχουν σχετικά υψηλότερη απόδοση, η προσέγγιση Novice (χρησιμοποιήθηκε κοινή λογική για την επιλογή μεταβλητών, και δε χωρίστηκαν τα δεδομένα σε calibration και holdout) έχει μέτρια απόδοση πρόβλεψης, ενώ τα μοντέλα που χρησιμοποίησαν discriminant ανάλυση και η προσέγγιση Explain (δόθηκε βάση στη θεωρία και σε εργαλεία απλοποίησης δεδομένων όπως η παραγοντική ανάλυση για την επιλογή μεταβλητών) έχουν χαμηλότερη απόδοση.

Κλείνοντας, θα αναφερθεί μία ακόμα μέθοδος που ασχολείται με την πρόβλεψη της συμπεριφοράς πελατών, και δεν συμπεριελήφθη στην παρούσα εργασία, αυτή των αλυσίδων Markov, τις οποίες χρησιμοποίησαν οι Jenamani et al. (2003) για να αναπτύξουν ένα μοντέλο που λαμβάνει υπόψη τη συμπεριφορά διαδικτυακών πελατών. Και οι Burez et al. (2007) ανέπτυξαν ένα μοντέλο αλυσίδας Markov για την πρόβλεψη απώλειας πελατών και ένα μοντέλο random forest, τα οποία και σύγκριναν με ένα μοντέλο λογιστικής παλινδρόμησης. Χρησιμοποιήθηκαν δεδομένα από μια εταιρία pay-TV, και τα αποτελέσματα έδειξαν ότι τα random forests υπερτερούν έναντι των άλλων δύο μεθόδων, ενώ η απλή λογιστική παλινδρόμηση και η λογιστική παλινδρόμηση με αλυσίδες Markov έχουν ουσιαστικά το ίδιο αποτέλεσμα, με τη δεύτερη να υπερτερεί ελαφρώς (οι συγκρίσεις έγιναν με βάση την αθροιστική καμπύλη lift και την καμπύλη ROC).

2.6.7. Σύνοψη

Συμπερασματικά, όπως φαίνεται από τη βιβλιογραφική ανασκόπηση, μια πλειάδα μεθόδων χρησιμοποιείται στον τομέα ανάλυσης και πρόβλεψης απώλειας πελατών, η οποία επεκτείνεται όλο και περισσότερο, αρχίζοντας από τα παραδοσιακά στατιστικά μοντέλα μέχρι την εκμάθηση μηχανών και τις πιο σύνθετες μεθόδους. Το σχήμα 16 απεικονίζει τις περισσότερο προτιμητέες μεθόδους πρόβλεψης συμπεριφοράς πελατών των ετών 2002-2007. Την πρώτη θέση κατέχει η ανάλυση παλινδρόμησης, ενώ ακολουθούν τα δέντρα απόφασης, τα μοντέλα Markov και τα νευρωνικά δίκτυα.



Σχήμα 16: Ανάλυση των μελετών απώλειας πελατών σχετικά με τις τεχνολογιών που χρησιμοποίησαν (Hadden, 2007)

Η έρευνα στον τομέα της εξόρυξης δεδομένων δείχνει ότι οι τεχνικές εκμάθησης μηχανών, όπως τα νευρωνικά δίκτυα, συχνά υπερτερούν έναντι των στατιστικών τεχνικών (Baesens et al. 2002, Bhattacharyya et al. 1998). Παρόλα αυτά είναι σημαντικό να σημειωθεί ότι η επιτυχία μιας μεθόδου και η ικανοποιητική της απόδοση εξαρτάται από ένα σύνολο πραγμάτων, όπως είναι το είδος του προβλήματος, η ποιότητα των δεδομένων, η ποιότητα των παραμέτρων που ορίζουν την μέθοδο προς εφαρμογή και άλλα. Συνεπώς δεν μπορεί να προκύψει ένα πάγιο συμπέρασμα ως προς το ποια μέθοδος υπερτερεί των υπολοίπων στον τομέα της ανάλυσης και πρόβλεψης απώλειας πελατών.

3. Δεδομένα, μεθοδολογία ανάλυσης, αποτελέσματα

3.1. Περιγραφή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία παρείχε το τμήμα Teradata Center for Customer Relationship Management του πανεπιστημίου Duke University, και αφορούν σε πληροφορίες πελατών μιας μεγάλης εταιρίας ασύρματης τηλεφωνίας. Τα δεδομένα αυτά έγιναν διαθέσιμα στους ενδιαφερόμενους στα πλαίσια του Διαγωνισμού Μοντελοποίησης Απώλειας Πελατών (Churn Modeling Tournament), το 2001. Σκοπός του διαγωνισμού αυτού ήταν να ερευνήσει ποιές μέθοδοι λειτουργούν αποτελεσματικότερα στην πρόβλεψη της απώλειας πελατών, και με αυτό τον τρόπο να εμπλουτιστεί την γνώση πάνω στην μοντελοποίηση πρόβλεψης (Predictive modeling²), η οποία αποτελεί σημαντικό εργαλείο μάρκετινγκ αλλά και σημαντικό παράγοντα του τμήματος διαχείρισης των πελατειακών σχέσεων μιας εταιρίας.

Οι πελάτες που συμμετέχουν στην ανάλυση επιλέχτηκαν ως εξής: έγινε δειγματοληψία ώριμων πελατών, πελατών δηλαδή που έμειναν στην εταιρία για τουλάχιστον 6 μήνες, κατά τους μήνες Ιούλιο, Σεπτέμβριο, Νοέμβριο και Δεκέμβριο του 2001. Για κάθε πελάτη, οι μεταβλητές υπολογίστηκαν με βάση τους τέσσερις αυτούς μήνες, και στη συνέχεια υπολογίστηκε ο δείκτης απώλειας πελατών με βάση το αν ο πελάτης αυτός έφυγε από την εταιρία κατά τη διάρκεια 31-60 ημερών μετά από την αρχική δειγματοληψία του. Το κενό τουλάχιστον ενός μηνός μεταξύ της δειγματοληψίας και της καταγραφής της απώλειας πελατών υιοθετήθηκε λαμβάνοντας υπόψη το χρόνο που χρειάζεται μία τηλεπικοινωνιακή εταιρία να εφαρμόσει τα αποτελέσματα ενός μοντέλου πρόβλεψης απώλειας πελατών στα πλαίσια του τμήματος μάρκετινγκ και διαχείρισης πελατειακών σχέσεων. Σε μία τέτοια εφαρμογή θα χρειαζόταν τουλάχιστον ένας μήνας για να οργανωθεί μία καμπάνια, να επιλεγθούν οι κατάλληλοι πελάτες, να γίνουν προβλέψεις, να τεθούν προτεραιότητες μεταξύ των πελατών, και να γίνουν οι κατάλληλες ενέργειες διατήρησής τους. Συνεπώς, από την οπτική της αποτελεσματικότητας μιας

² Predictive modeling: η στατιστική διαδικασία κατά την οποία υπολογίζονται τα σκορ των πελατών και γίνεται η κατάλληλη στοχοθέτησή τους για την καμπάνια μάρκετινγκ.

καμπάνιας, η χρονική περίοδος ενός με δύο μήνες μετά την καταγραφή των τελευταίων μεταβλητών πρόβλεψης, είναι μια σημαντική περίοδος που πρέπει να ληφθεί υπόψη.

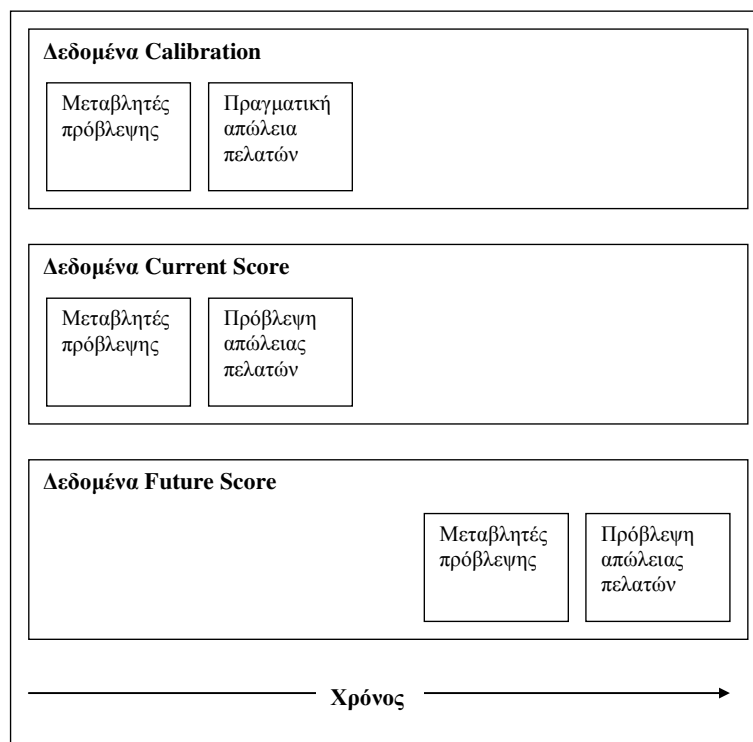
Τα δεδομένα οργανώνονται σε τρία αρχεία: ένα αρχείο δεδομένων εκπαίδευσης (Calibration data), και 2 αρχεία ελέγχου (validation), το αρχείο δεδομένων Current Score data, και το αρχείο δεδομένων Future Score data, ως εξής (πίνακας 2):

	Δεδομένα Calibration	Δεδομένα Current Score	Δεδομένα Future Score
Μέγεθος δείγματος	100.000 πελάτες	51.306 πελάτες	100.462 πελάτες
Αριθμός μεταβλητών πρόβλεψης	173	173	173
Δείκτης απώλειας πελατών	Ναι	Όχι	Όχι
% απώλεια πελατών	50%	1,80%	1,80%

Πίνακας 2: Τα δεδομένα των πελατών, όπως αυτά οργανώθηκαν στα τρία αρχεία.

Το αρχείο δεδομένων Calibration data περιλαμβάνει τις 173 μεταβλητές που περιγράφουν τις πελατειακές πληροφορίες της εταιρίας, και συγκεντρώθηκαν στο τετράμηνο Σεπτεμβρίου-Δεκεμβρίου του 2001, ενώ επιπλέον περιλαμβάνει και το δείκτη απώλειας πελατών που ορίζει ποιοι πελάτες είχαν φύγει ή μείνει στην εταιρία κατά την περίοδο 31-60 ημερών από τη δειγματοληψία τους. Ο δείκτης απώλειας πελατών ορίζεται από την εξαρτημένη μεταβλητή «churn», η οποία παίρνει δύο διακριτές τιμές, 0 εάν ο πελάτης παραμένει με την εταιρία και 1 εάν την έχει εγκαταλείψει μέσα στην προαναφερθείσα περίοδο. Τα αρχεία δεδομένων Current και Future Score data περιλαμβάνουν επίσης τις 173 μεταβλητές αλλά όχι το δείκτη απώλειας πελατών. Τα αρχεία αυτά δηλαδή περιλαμβάνουν ένα σύνολο πελατών, το 1,80% των οποίων εγκαταλείπουν την εταιρία, χωρίς όμως να υπάρχει γνώση για την ταυτότητα του ποσοστού αυτού των πελατών που τελικά έφυγαν. Το σύνολο δεδομένων Current Score data περιλαμβάνει διαφορετικό σύνολο πελατών από εκείνο του αρχείου Calibration data, η δειγματοληψία των οποίων όμως έγινε την ίδια χρονική στιγμή με εκείνους του συνόλου Calibration data. Το αρχείο Future Score data περιλαμβάνει επίσης διαφορετικό σύνολο πελατών, η δειγματοληψία του οποίου

έγινε σε μελλοντική χρονική στιγμή σε σχέση σύνολα δεδομένων των άλλων δύο αρχείων (περίπου τρεις μήνες αργότερα) (σχήμα 17). Αυτό συμβαίνει για να «αντιγραφεί» πιστά μια κατάσταση ενός μοντέλου που εκτιμάται σε μία δεδομένη χρονική στιγμή και εφαρμόζεται σε μελλοντική χρονική στιγμή.



Σχήμα 17: Δομή βάσης δεδομένων

Συνεπώς, όπως γίνεται κατανοητό από τη δομή των τριών αρχείων, τα μοντέλα πρόβλεψης που κατασκευάζονται εκτιμούνται με βάση το σύνολο δεδομένων calibration και χρησιμοποιούνται για πρόβλεψη πάνω στα σύνολα δεδομένων current και future score. Εάν η ακρίβεια πρόβλεψης μειώνεται αξιοσημείωτα για αυτά τα δεδομένα, σημαίνει ότι το μοντέλο απώλειας πελατών θα πρέπει να επανεκτιμάται συχνά μέσα στο χρόνο. Εάν η ακρίβεια πρόβλεψης δεν μειωθεί σημαντικά, αυτό σημαίνει ότι το μοντέλο πρόβλεψης πελατών διατηρεί την ισχύ του (staying power), και μπορεί να χρησιμοποιηθεί για τουλάχιστον τρεις μήνες προτού επανεκτιμηθεί.

Πρέπει να σημειωθεί ότι το ετήσιο ποσοστό πελατών που εγκαταλείπουν τη συγκεκριμένη επιχείρηση σε ένα δεδομένο μήνα είναι περίπου 1,80%. Το ποσοστό αυτό είναι σημαντικά χαμηλό, με αποτέλεσμα να μην υπάρχει αρκετή πληροφορία που να διαχωρίζει τους πελάτες που την εγκαταλείπουν από εκείνους που παραμένουν, και έτσι μπορεί να τεθεί σε κίνδυνο η αποτελεσματικότητα εκμάθησης

του μοντέλου που θα χρησιμοποιηθεί, λόγω μη ισορροπίας των κατηγοριών (class imbalance) (King and Zeng 2001). Συνεπώς το σύνολο δεδομένων Calibration δημιουργήθηκε έτσι ώστε να υπάρχει περίπου 50-50 διαχωρισμός μεταξύ των πελατών που εγκατέλειψαν την εταιρία και αυτών που της έμειναν πιστοί (ο ακριβής αριθμός είναι 49.562 πελάτες που εγκαταλείπουν και 50.438 που δεν εγκαταλείπουν την επιχείρηση), χρησιμοποιώντας την τεχνική «over sampling³». Η τεχνική αυτή δε χρησιμοποιήθηκε κατά τη δημιουργία των συνόλων δεδομένων Current Score και Future Score data, και αυτό για να υπάρξει μια πιο ρεαλιστική δοκιμή πρόβλεψης με βάση τα πραγματικά δεδομένα.

Τα πελατειακά δεδομένα που χρησιμοποιούνται στην ανάλυση περιλαμβάνουν τρία είδη μεταβλητών: *δεδομένα συμπεριφοράς πελατών* όπως λεπτά χρήσης, έσοδα, τηλεφωνική συσκευή κ.α., *δεδομένα αλληλεπίδρασης με την εταιρία* όπως κλήσεις πελατών στο κέντρο εξυπηρέτησης πελατών κ.α., και *δημογραφικά στοιχεία πελατών* όπως ηλικία, εισόδημα, γεωγραφική τοποθεσία κ.α. Τα δεδομένα συνολικά αποτελούνται από 173 μεταβλητές, 58 από τις οποίες είναι κατηγορικές και 115 συνεχείς μεταβλητές. Όλες οι μεταβλητές παρατίθενται αναλυτικά στο παράρτημα Α.

3.2. Επιλογή μεταβλητών και κωδικοποίηση

Εξαιτίας του μεγάλου αριθμού των διαθέσιμων μεταβλητών, κρίθηκε απαραίτητη η επιλογή των καταλλληλότερων. Αρχικά, μετά από μελέτη της σχετικής βιβλιογραφίας όσον αφορά στα σημαντικότερα χαρακτηριστικά των πελατών και με χρήση κοινής λογικής, διαγράφησαν μεταβλητές οι οποίες δεν ήταν σχετικές με το πεδίο μελέτης της εργασίας, και είχαν πληθώρα κενών στοιχείων στα δεδομένα τους (πίνακας 3):

³ Over sampling: τεχνική που εξαλείφει ή ελαχιστοποιεί τη σπανιότητα ορισμένων δεδομένων (π.χ. αριθμός πελατών που εγκαταλείπουν μια επιχείρηση σε σχέση με τον αριθμό εκείνων των πελατών που της μένουν πιστοί) με το να μεταβάλλει την κατανομή των δειγμάτων εκπαίδευσης, επιφέροντας έτσι ισορροπία θετικών και αρνητικών δειγμάτων.

	Συνεχείς μεταβλητές που διαγράφησαν
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
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
EQPDAYS	Number of days (age) of current equipment
	Κατηγορικές μεταβλητές που διαγράφησαν
CARTYPE	Dominant vehicle lifestyle
CHILDREN	Children present in household
CRTCOUNT	Adjustments made to credit rating of individual
CSA	Communications local service area
CUSTOMER_ID	Unique tournament specific customer ID for scoring purposes
DWLLSIZE	Dwelling size
DWLLTYPE	Dwelling unit type
EDUC1	Education of first household member
HHSTATIN	Premier household status indicator
LAST_SWAP	Date of last phone swap
LOR	Length of residence
NUMBCARS	Known number of vehicles
OCCU1	Occupation of first household member
OWNRENT	Home owner/renter status
PRE_HND_PRICE	Previous handset price
PRIZM_SOCIAL_ONE	Social group letter only
PROPTYPE	Property type detail
RV	RV indicator
TRUCK	Truck indicator

Πίνακας 3: Μεταβλητές που διαγράφησαν από την ανάλυση

Για να αντιμετωπιστούν τα εναπομείναντα κενά στην ανάλυση και οι χαμένες τιμές, διαγράφησαν εκείνοι οι πελάτες των οποίων τα ελλιπή στοιχεία έφταναν ή ξεπερνούσαν το 1,7%. Επιπλέον, στην περίπτωση των συνεχών μεταβλητών, ορισμένες από αυτές αφενός εμφάνιζαν ένα μεγάλο αριθμό ελλিপών τιμών, αφετέρου θεωρήθηκαν πολύ σημαντικές για την ανάλυση ώστε να διαγραφούν. Το πρόβλημα αυτό αντιμετωπίστηκε με την αντικατάσταση των κενών στοιχείων με μηδενικά, στις ακόλουθες μεταβλητές (πίνακας 4):

Μεταβλητές	Εξήγηση	Ελλιπή στοιχεία
RETDAYS	Number of days since last retention call	96,017%
RMCALLS	Total number of roaming calls	85,777%
RMMOU	Total minutes of use of roaming calls	85,777%
RMREV	Total revenue of roaming calls	85,777%

Πίνακας 4: Συνεχείς μεταβλητές στις οποίες έγινε αντικατάσταση των κενών στοιχείων με μηδενικά

Όμοια και στην περίπτωση των κατηγορικών μεταβλητών, ορισμένες από αυτές, σημαντικές στην ανάλυση, παρουσίαζαν κενά. Τα κενά αυτά αντικαταστάθηκαν με το μηδέν, σε περιπτώσεις που η μηδενική τιμή είχε λογικό νόημα (πίνακας 5):

Μεταβλητές	Εξήγηση	Ελλιπή στοιχεία
ADULTS	Number of adults in household	23,0%
PCOWNER	PC owner dummy variable	81,5%
REF_QTY	Total number of referrals	95,5%
TOT_RET	Total calls into retention team	96,0%
WRKWOMAN	Working woman in household	87,5%

Πίνακας 5: Κατηγορικές μεταβλητές στις οποίες έγινε αντικατάσταση των κενών στοιχείων με μηδενικά

Αντίθετα, σε περιπτώσεις που η μηδενική τιμή δεν μπορούσε να έχει λογικό νόημα, τα κενά αυτά αντικαταστάθηκαν με έναν αριθμό ο οποίος υποδήλωνε την έλλειψη τιμής (συνήθως το 9) (πίνακας 6):

Μεταβλητές	Εξήγηση	Ελλιπή στοιχεία
DIV_TYPE	Division type code	81,5%
HND_PRICE	Current handset price	0,8%
INFOBASE	InfoBase match	22,1%
MAILORD	Mail order buyer	64,4%
MAILRESP	Mail responder	62,9%
MAILFLAG	DMA: Do not mail flag	98,5%
SOLFLAG	Infobase no phone solicitation flag	98,0%
TOT_ACPT	Total offers accepted from retention team	96,0%

Πίνακας 6: Κατηγορικές μεταβλητές στις οποίες έγινε αντικατάσταση των κενών στοιχείων με το 9

Αφού αντιμετωπίστηκαν οι χαμένες τιμές και τα κενά στην ανάλυση, τα κατηγορικά δεδομένα χρειάστηκε να τροποποιηθούν προκειμένου να αποκτήσουν την κατάλληλη μορφή για την περαιτέρω επεξεργασία και χρήση τους στην κατασκευή των μοντέλων πρόβλεψης. Συνεπώς, οι ποιοτικές απαντήσεις των κατηγορικών μεταβλητών κωδικοποιήθηκαν σε αριθμούς, π.χ. τα δεδομένα της μορφής Yes, No μετατράπηκαν σε 1, 0 αντίστοιχα.

Ορισμένες ποιοτικές μεταβλητές με πληθώρα κατηγοριών, οι οποίες εμφάνιζαν λίγες τιμές και μπορούσαν να υποστούν συγχώνευση, ομαδοποιήθηκαν. Η ομαδοποίηση αυτή έγινε με χρήση ψευδομεταβλητών, με αριθμητικές τιμές οι οποίες αντικατέστησαν τις ποιοτικές τιμές κάθε μεταβλητής (αυτές οι μεταβλητές

μετονομάστηκαν σε “παλιό όνομα_NEW”). Οι μεταβλητές στις οποίες εφαρμόστηκε αυτή η κωδικοποίηση είναι οι εξής (πίνακας 7 έως 12):

κατηγορική μεταβλητή CRCLSCOD					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
A	17.00%	1	GY	0.03%	6
A2	0.86%	1	H	0.01%	6
A3	0.00%	1	I	0.24%	6
AA	36.51%	1	IF	0.01%	6
B	4.11%	2	J	0.19%	6
B2	0.11%	2	JF	0.19%	6
BA	12.67%	2	K	0.11%	6
C	1.46%	3	L	0.01%	6
C2	0.18%	3	M	0.17%	6
C5	0.08%	3	O	0.03%	6
CA	8.33%	3	P1	0.01%	6
CC	0.03%	3	S	0.00%	6
CY	0.19%	3	TP	0.01%	6
D	0.25%	4	U	0.48%	6
D2	0.01%	4	U1	0.07%	6
D4	0.30%	4	V	0.00%	6
D5	0.11%	4	V1	0.07%	6
DA	3.67%	4	W	0.10%	6
E	0.34%	5	Y	0.05%	6
E2	0.02%	5	Z	0.19%	6
E4	0.83%	5	Z1	0.01%	6
EA	6.42%	5	Z2	0.01%	6
EC	0.05%	5	Z4	0.25%	6
EF	0.00%	5	Z5	0.06%	6
EM	0.06%	5	ZA	3.49%	6
G	0.24%	6	ZF	0.00%	6
GA	0.28%	6	ZY	0.12%	6

Πίνακας 7: κωδικοποίηση ποιοτικής μεταβλητής CRCLSCOD

HND_PRICE					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
\$9.99	4.36%	1	\$149.99	22.04%	3
\$29.99	21.74%	1	\$159.99	0.00%	4
\$39.99	0.66%	1	\$179.99	0.08%	4
\$59.99	8.69%	2	\$199.99	10.25%	4
\$79.99	9.57%	2	\$239.99	0.04%	4
\$99.99	7.73%	2	\$249.99	0.24%	4
\$119.99	0.00%	3	\$299.99	0.09%	4
\$129.99	13.41%	3	\$399.99	0.19%	4
			\$499.99	0.06%	4

Πίνακας 8: κωδικοποίηση ποιοτικής μεταβλητής HND_PRICE

PHONES					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
1	63.27%	1	8	0.05%	2
2	25.04%	1	9	0.03%	2
3	7.82%	1	10	0.01%	2
4	2.47%	1	11	0.00%	2
5	0.85%	2	14	0.00%	2
6	0.33%	2	15	0.00%	2
7	0.12%	2	16	0.00%	2

Πίνακας 9: κωδικοποίηση ποιοτικής μεταβλητής PHONES

INCOME					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
1	4.03%	1	5	8.28%	3
2	2.26%	1	6	18.80%	3
3	5.83%	2	7	11.60%	4
4	7.79%	2	8	5.14%	4
			9	10.83%	4

Πίνακας 10: κωδικοποίηση ποιοτικής μεταβλητής INCOME

ETHNIC					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
B	1.37%	1	M	0.15%	3
C	0.24%	1	N	33.39%	4
D	0.87%	2	O	4.20%	1
F	2.13%	2	P	0.49%	1
G	5.91%	2	R	1.05%	1
H	13.62%	5	S	12.75%	6
I	3.81%	2	U	10.95%	3
J	2.84%	1	X	0.09%	3
			Z	4.43%	0

Πίνακας 11: κωδικοποίηση ποιοτικής μεταβλητής ETHNIC

AREA					
Τιμή	% του συνόλου	κωδικοποίηση	Τιμή	% του συνόλου	κωδικοποίηση
Atlantic South	6.13%	5	Midwest	6.35%	4
California North	5.91%	9	New England	5.45%	1
Central/South Texas	4.30%	7	New York City	11.10%	2
Chicago	5.14%	3	North Florida	4.35%	5
Dallas	5.43%	7	Northwest/Rocky Mountain	4.33%	8
DC/Maryland/Virginia	6.57%	5	Ohio	4.58%	3
Great Lakes	4.65%	3	Philadelphia	2.45%	2
Houston	4.34%	7	South Florida	3.33%	5
Los Angeles	6.64%	9	Southwest	6.19%	7
			Tennessee	2.74%	6

Πίνακας 12: κωδικοποίηση ποιοτικής μεταβλητής AREA

Μετά την αυτή την επεξεργασία των μεταβλητών, ο αριθμός τους μειώθηκε στις 149, από τις οποίες οι 40 είναι κατηγορικές και οι 109 συνεχείς μεταβλητές.

3.2.1. Μονομεταβλητή ανάλυση

Στο σημείο αυτό πραγματοποιείται μονομεταβλητή ανάλυση (Univariate analysis), στην οποία εξετάζεται η διακριτική ικανότητα μιας μεταβλητής κάθε φορά, ώστε να υπάρξει μια πρώτη εικόνα σχετικά με τις σχέσεις των μεταβλητών με τον δείκτη απώλειας πελατών. Η περιγραφή των σχέσεων κάθε μίας από τις μεταβλητές με το δείκτη απώλειας πελατών έγινε με χρήση των στατιστικών ελέγχων Mann-Whitney test, για τον έλεγχο των ποσοτικών μεταβλητών, και τον χ^2 , για τον έλεγχο των ποιοτικών, με χρήση του στατιστικού πακέτου SPSS.

Πιο συγκεκριμένα, ο έλεγχος Mann-Whitney είναι μια μη παραμετρική διαδικασία. Από τον έλεγχο απορρίφθηκαν οι μεταβλητές οι οποίες δεν εμφάνιζαν διαφοροποίηση στη σχέση τους με το δείκτη απώλειας πελατών, και κρατήθηκαν εκείνες για τις οποίες υπήρξε θετική ένδειξη ότι επηρεάζουν το συγκεκριμένο δείκτη (επίπεδο στατιστικής σημαντικότητας $\alpha < 0.05$). Ο έλεγχος αυτός βασίστηκε μόνο στα δεδομένα του αρχείου calibration, όπου υπάρχει ο δείκτης απώλειας πελατών. Οι μεταβλητές οι οποίες εξαιρέθηκαν από την ανάλυση ήταν οι εξής (πίνακας 13):

rmmou	mou_range
rmcalls	recv_sms_mean
rmrev	recv_sms_range
roam_mean	

Πίνακας 13: Ποσοτικές μεταβλητές που εξαιρέθηκαν από την ανάλυση μετά τη χρήση Mann-Whitney test

Στην περίπτωση των ποιοτικών μεταβλητών και της περιγραφής των σχέσεών τους με το δείκτη απώλειας πελατών, χρησιμοποιήθηκε ο έλεγχος Pearson's χ^2 . Ο έλεγχος αυτός αποτελεί μία τεχνική ανάλυσης της σχέσης δύο κατηγορικών μεταβλητών, όπου γίνεται έλεγχος ανεξαρτησίας και ορίζονται μέτρα συσχέτισης. Με όμοιο τρόπο όπως και στην περίπτωση των συνεχών μεταβλητών, δηλαδή ελέγχοντας το επίπεδο στατιστικής σημαντικότητας κάθε μεταβλητής συγκριτικά με το δείκτη «churn», έμειναν στην ανάλυση οι ακόλουθες μεταβλητές (πίνακας 14):

crclscod_NEW	mailresp
hnd_price_NEW	wrkwoman
phones_NEW	mtrcycle
income_NEW	marital
thnic_NEW	mailord
area_NEW	hnd_webcap
car_buy	refurb_NEW
creditcd	dualband
infobase	div_type
kid0_2	asl_flag

Πίνακας 14: Ποιοτικές μεταβλητές που έμειναν στην ανάλυση μετά τον έλεγχο χ^2

Τα αποτελέσματα αυτά δίνουν μια μικρή εικόνα για τα χαρακτηριστικά των πελατών που μένουν ή φεύγουν από μια εταιρία, με συνέπεια να μην υπάρχει δυνατότητα κατάληξης σε ασφαλή συμπεράσματα για το ποιες μεταβλητές περιγράφουν επαρκώς την κατάσταση των πελατών αυτών. Το πρόβλημα αυτό σε συνδυασμό με το ότι οι αλληλεπιδράσεις των μεταβλητών δεν μπορούν να εκτιμηθούν μέσω της μονομεταβλητής ανάλυσης, κάνει πιο έντονη την ανάγκη χρήσης πολυμεταβλητών μεθόδων.

3.2.2. Πολυμεταβλητή ανάλυση

Μέσω της πολυμεταβλητής ανάλυσης (Multivariate analysis) γίνεται εκτίμηση των αλληλεπιδράσεων των μεταβλητών στο δείκτη απώλειας πελατών. Εδώ χρησιμοποιήθηκαν οι μεταβλητές που προέκυψαν από τους ελέγχους Mann-Whitney test και χ^2 , στα δεδομένα του αρχείου calibration.

Πιο συγκεκριμένα, προκειμένου να ελεγχθεί το πώς επηρεάζεται η τιμή της εξαρτημένης μεταβλητής (απώλεια πελατών) από τις τιμές των ανεξάρτητων, και να γίνει πρόβλεψη της τιμής της εξαρτημένης μεταβλητής στην περίπτωση που η μεταβλητή αυτή είναι διχοτομική, χρησιμοποιήθηκε η λογιστική παλινδρόμηση (logistic regression). Η μέθοδος αυτή είναι κατάλληλη σε καταστάσεις όπου είναι επιθυμητή η πρόβλεψη της παρουσίας (ή απουσίας) ενός χαρακτηριστικού ή αποτελέσματος βάσει ορισμένων τιμών ενός συνόλου μεταβλητών πρόβλεψης, ενώ είναι μια στατιστική διαδικασία που μπορεί να επεξεργαστεί εύκολα πολλές μεταβλητές. Οι μεταβλητές που τελικά προέκυψαν από τη μέθοδο επιλογής forward stepwise⁴ είναι (πίνακας 15):

⁴ Η μέθοδος επιλογής μεταβλητών forward stepwise, στα πλαίσια της διαδικασίας Δυναμικής Λογιστικής Παλινδρόμησης (Binary Logistic Regression), φροντίζει ώστε κατά την κατασκευή ενός

crclscod_NEW	adults	totmou
hnd_price_NEW	age1	totrev
thnic_NEW	age2	adjrev
area_NEW	models	avg3rev
creditcd	tot_acpt	rev_mean
infobase	actvsbs	mou_mean
kid0_2	unqsubs	totmrc_mean
mailresp	drop_vce_mean	rev_range
mtrcycle	unan_vce_mean	totmrc_range
marital	threeway_mean	da_range
hnd_webcap	peak_vce_mean	ovrmou_range
refurb_NEW	unan_vce_range	roam_range
dualband	mou_rvce_range	drop_blk_mean
div_type	iwylis_vce_range	custcare_range
asl_flag	avgmou	peak_vce_range
retdays	avg3mou	attempt_range

Πίνακας 15: Μεταβλητές που έμειναν στην ανάλυση μετά τη χρήση της λογιστικής παλινδρόμησης

Παράλληλα, και για τον ίδιο σκοπό, χρησιμοποιήθηκε και η μέθοδος των δέντρων απόφασης (classification tree), στο ίδιο σύνολο δεδομένων (ο αλγόριθμος CRT χρησιμοποιήθηκε για την κατασκευή του δέντρου). Η μέθοδος, εκτός από το γεγονός ότι προβλέπει τιμές μιας εξαρτημένης μεταβλητής με βάση τις τιμές ανεξαρτήτων μεταβλητών πρόβλεψης, δίνει τη δυνατότητα επιλογής μεταβλητών χωρίς να γίνεται υπόθεση γραμμικού διαχωρισμού, αλλά πραγματοποιώντας μη γραμμικές συσχετίσεις. Οι μεταβλητές που προέκυψαν από αυτή τη διαδικασία είναι (πίνακας 16):

months	mou_mean
totmrc_mean	callwaitmean
totmrc_range	age1
rev_range	age2
asl_flag	mou_peav_range
hnd_webcap	totacpt
adjrev	avg3mou
unqsubs	avgqty
	hnd_price_NEW

Πίνακας 16: Μεταβλητές που έμειναν στην ανάλυση μετά τη χρήση του δέντρου απόφασης

Στα πλαίσια της επιλογής των καταλληλότερων μεταβλητών πρόβλεψης, τα δύο νέα σύνολα μεταβλητών που προέκυψαν από τις δύο προαναφερθείσες μεθόδους ελέχθησαν ως προς τη σημαντικότητα των μεταβλητών τους στην ανάλυση. Προς

μοντέλο να συμπεριληφθούν μεταβλητές πρόβλεψης οι οποίες συνεισφέρουν σημαντικά σε αυτό. Η μέθοδος ξεκινά με ένα μοντέλο που δεν περιέχει καμία μεταβλητή και σε κάθε βήμα οι μεταβλητές με το μεγαλύτερο στατιστικό σκορ των οποίων η τιμή σημαντικότητας είναι μικρότερη από το 0.05 προστίθενται στο μοντέλο.

αυτό το σκοπό έγινε η παρατήρηση ότι ορισμένες από τις ποιοτικές μεταβλητές μπορούν να τροποποιηθούν περαιτέρω ώστε να απλοποιηθεί η κλίμακα μέτρησής τους, με στόχο την ανάπτυξη απλών, εύχρηστων και αποτελεσματικών μοντέλων.

3.2.3. Μετασχηματισμός ποιοτικών μεταβλητών

Εδώ έγινε έλεγχος της σημαντικότητας στην ανάλυση των δύο νέων συνόλων μεταβλητών που προέκυψαν από τη λογιστική παλινδρόμηση και το δέντρο απόφασης. Παρατηρήθηκε ότι η στατιστική σημαντικότητα κάποιων βαθμίδων στην κλίμακα ορισμένων ποιοτικών μεταβλητών ήταν πάνω από το επιτρεπτό όριο (0,05) και συνεπώς αποφασίστηκε η συγχώνευση των κατηγοριών αυτών κατάλληλα μεταξύ τους. Οι συγχωνεύσεις αυτές έγιναν τόσο με βάση τη στατιστική σημαντικότητά τους, λαμβάνοντας όμως παράλληλα υπόψη και τη λογική συσχέτισή τους, π.χ. στην περίπτωση της μεταβλητής AREA, οι συγχωνεύσεις των τιμών της έγιναν λαμβάνοντας υπόψη και τη γεωγραφική θέση των περιοχών που αντιστοιχούν στις επιμέρους βαθμίδες της μεταβλητής αυτής.

Η πρώτη τροποποίηση έγινε μετά από έλεγχο των μεταβλητών (πίνακας 15) που προέκυψαν από τη λογιστική παλινδρόμηση που παρουσιάστηκε παραπάνω, και είναι η εξής (πίνακας 17):

asl_flag	1,0	ως έχει
refurb_NEW	1,0	ως έχει
mtrcycle	1,0	ως έχει
kid0_2	1,2	ως έχει
creditcd	1,0	ως έχει
hnd_price_NEW	1,2,3,4	ως έχει
marital	1,2,3,4,5	εκτός
hnd_webcap	1,2,3,4	εκτός
div_type	1,9,12	εκτός
dualband	0,1,2,3	{2,3}
infobase	0,4,9	{4,9}
area_NEW	1,2,3,4,5,6,7,8,9	{1,2},{5,6,7}
ethnic_NEW	0,1,2,3,4,5,6	{2,3,4}
crclscod_NEW	1,2,3,4,5,6	{1,2,3}

Πίνακας 17: Μετασχηματισμός μεταβλητών στα πλαίσια της 1^{ης} παλινδρόμησης

Στη συνέχεια έγινε έλεγχος της στατιστικής σημαντικότητας των μεταβλητών του νέου αυτού συνόλου μέσα από λογιστική παλινδρόμηση, τα αποτελέσματα της οποία υπόδειξαν την ανάγκη περεταίρω μετασχηματισμών. Η ίδια διαδικασία επαναλήφθηκε μέχρις ότου οι μεταβλητές που προέκυψαν ήταν όλες στατιστικά σημαντικές (πίνακας 18):

asl_flag	1,0
refurb_NEW	1,0
mtrcycle	1,0
kid0_2	1,2
creditcd	1,0
hnd_price_NEW	1,2,3,4
dualband2	0,1,2
infobase2	0,4
area_NEW3	1,3,4,8
ethnic_NEW4	0,1,6
crclscod_NEW4	1,6

Πίνακας 18: Τελικοί μετασχηματισμοί μεταβλητών

Πέρα από τους παραπάνω μετασχηματισμούς, έγινε δοκιμή και μιας δεύτερης, περισσότερο γενικευμένης επεξεργασίας, πάντα στο αρχείο δεδομένων calibration. Σε αυτή την ανάλυση εξετάστηκαν όλες οι ποιοτικές μεταβλητές που είχε το πρόβλημα εξ αρχής, εκτός των κατηγορικών μεταβλητών που εξαιρέθηκαν και παρουσιάζονται στον πίνακα 3. Στην ανάλυση αυτή επίσης συμπεριελήφθησαν οι μεταβλητές των πινάκων 7 έως 12, οι οποίες τροποποιήθηκαν κατά το πρώτο σενάριο μετασχηματισμού των δεδομένων (οι μεταβλητές με σκούρα γράμματα στον πίνακα που ακολουθεί). Συγκεντρωτικά, οι μεταβλητές που εξετάστηκαν στο 2^ο σενάριο μετασχηματισμών είναι οι εξής (πίνακας 19):

newCell	infobase	carBuy
aslFlag	mailflag	area_NEW
divType	solflag	ethnic_NEW
dualband	forgntvl	income_NEW
refurbNew	pcowner	phones_NEW
hndWebcap	kid02	hndprice_NEW
mtrcycle	kid35	crclscod_NEW
marital	kid610	area_NEW3
mailordr	kid1115	ethnic_NEW4
wrkwoman	kidy617	crclscod_NEW4
mailresp	creditcd	

Πίνακας 19: Σύνολο των ποιοτικών μεταβλητών που εξετάστηκαν στο 2^ο εναλλακτικό σενάριο μετασχηματισμού των δεδομένων

Οι μεταβλητές αυτές αναλύθηκαν εξετάζοντας για κάθε βαθμίδα της κλίμακας μέτρησής τους τον αριθμό των πελατών που έμεινε στην εταιρία, τον αριθμό εκείνων που έφυγε καθώς και το σύνολο. Με βάση τα αριθμητικά αυτά δεδομένα υπολογίστηκε το log of odds κάθε βαθμίδας i σε κάθε μεταβλητή j:

$$\ln\left(\frac{P_{ij}}{1 - P_{ij}}\right),$$

όπου το p_{ij} είναι το ποσοστό των πελατών στη βαθμίδα i της μεταβλητής j που έμειναν στην εταιρία σε σχέση με το σύνολο των πελατών στη βαθμίδα αυτή.

Το μέγεθος αυτό βοηθά στη μέτρηση της πληροφορίας που παρέχει κάθε βαθμίδα ενός κριτηρίου στο διαχωρισμό των πελατών. Με βάση αυτή τη λογική έγινε η δεύτερη ομαδοποίηση κάποιων βαθμίδων που παρέχουν παρόμοια πληροφορία στην ταξινόμηση των πελατών (παρόμοια log of odds). Το νέο σύνολο των ποιοτικών μεταβλητών που τελικά κρατήθηκε στην ανάλυση είναι (πίνακας 20):

asl_flag	1,0
dualband2_2	0,1,2
refurb_NEW	1,0
mailord	1,9
mailresp	1,9
infobase2	0,4,9
solflag	0,9
ethnic_NEW2_2	0,1,2
phones_NEW	1,2
hnd_price_NEW	1,2,3,4
crclscod_NEW2_2	1,3,4
area_NEW3_2	1,4,8

Πίνακας 20: Εναλλακτικός μετασχηματισμός των ποιοτικών μεταβλητών

Ο δεύτερος μετασχηματισμός των μεταβλητών χρησιμοποιήθηκε στην ανάλυση επειδή είναι περισσότερο εμπειρικός από τον πρώτο. Στον εναλλακτικό αυτό μετασχηματισμό ο αναλυτής έχει τη δυνατότητα να ελέγξει ο ίδιος ποιες βαθμίδες των ποιοτικών μεταβλητών παρέχουν παρόμοια πληροφορία στην ταξινόμηση των πελατών, και να τις ενσωματώσει με σκοπό τη μεγιστοποίηση της ισχύς του μοντέλου. Η διαδικασία αυτή στον πρώτο μετασχηματισμό γίνεται αυτοματοποιημένα, μέσα από τη χρήση της λογιστικής παλινδρόμησης και του δέντρου απόφασης, ελαχιστοποιώντας την συμμετοχή του ανθρώπινου παράγοντα στην ανάλυση.

3.3. Εφαρμογές μεθόδων

Μετά την κατάλληλη επιλογή και τροποποίηση των μεταβλητών, ακολούθησε η χρήση τους στην ανάλυση και πρόβλεψη της απώλειας πελατών. Όπως έχει αναφερθεί και παραπάνω, χρησιμοποιήθηκαν μοντέλα δέντρων απόφασης, λογιστικής παλινδρόμησης και νευρωνικών δικτύων, τα οποία υλοποιήθηκαν στο στατιστικό πακέτο SPSS 16.0, καθώς και μοντέλα προσεγγιστικών και προσθετικών μηχανών διανυσμάτων υποστήριξης (PSVM και ASVM αντίστοιχα), τα οποία υλοποιήθηκαν

στο περιβάλλον της MATLAB R2008a. Εδώ πρέπει να σημειωθεί ότι τα μοντέλα προσεγγιστικών μηχανών διανυσμάτων υποστήριξης που χρησιμοποιήθηκαν ήταν γραμμικά. Τα αποτελέσματα των μεθόδων συγκρίνονται μεταξύ τους με κριτήρια αξιολόγησης την καμπύλη top decile lift και το συντελεστή gini.

3.3.1. Αποτελέσματα μεθόδων

Τα αποτελέσματα παρουσιάζονται υπό μορφή πινάκων ανάλογα με το σύνολο των μεταβλητών που χρησιμοποιήθηκαν στις μεθόδους. Τα σύνολα αυτά κωδικοποιούνται για την ευκολότερη ανάγνωση των πινάκων:

Σ1: Μεταβλητές που προέκυψαν μετά τη χρήση λογιστικής παλινδρόμησης (πολυμεταβλητή ανάλυση, βλ. πίνακα 15)

Σ2: Ποσοτικές μεταβλητές που προέκυψαν μετά τη χρήση λογιστικής παλινδρόμησης (πολυμεταβλητή ανάλυση, βλ. πίνακα 15), και ποιοτικές που προέκυψαν μετά την πρώτη τροποποίηση των ποιοτικών μεταβλητών (μετασχηματισμός ποιοτικών μεταβλητών, βλ. πίνακα 18)

Σ3: Ποσοτικές μεταβλητές που προέκυψαν μετά τη χρήση λογιστικής παλινδρόμησης (πολυμεταβλητή ανάλυση, βλ. πίνακα 15), και ποιοτικές που προέκυψαν μετά τη δεύτερη τροποποίηση των ποιοτικών μεταβλητών (μετασχηματισμός ποιοτικών μεταβλητών, βλ. πίνακα 20)

Σ4: Μεταβλητές που προέκυψαν μετά τη χρήση του δέντρου απόφασης (πολυμεταβλητή ανάλυση, βλ. πίνακα 16)

Σ5: Μεταβλητές που προέκυψαν μετά τη χρήση μονομεταβλητής ανάλυσης

Όπως έχει αναφερθεί και στην ενότητα 3.2, χρησιμοποιήθηκαν δύο διαφορετικές προσεγγίσεις για να γίνει επιλογή των κατάλληλων μεταβλητών από το αρχικό σύνολο. Αυτές είναι η λογιστική παλινδρόμηση, κατά την οποία γίνεται γραμμικός διαχωρισμός ομάδων, και το δέντρο απόφασης, κατά το οποίο γίνεται μη γραμμικός διαχωρισμός ομάδων. Από τη χρήση της λογιστικής παλινδρόμησης για την επιλογή μεταβλητών προκύπτουν τα σύνολα μεταβλητών Σ1, Σ2, Σ3. Το σύνολο μεταβλητών Σ1 χρησιμοποιείται για να τρέξει ένα μοντέλο νευρωνικών δικτύων, ένα μοντέλο προσεγγιστικών μηχανών διανυσμάτων υποστήριξης (PSVM) και ένα μοντέλο προσθετικών μηχανών διανυσμάτων υποστήριξης (ASVM) (πίνακας 21):

	Νευρωνικά Δίκτυα	PSVM	ASVM
current score data			
Συντελεστής gini	0.306	0.260	0.270
top decile lift	2.1720	2.1279	2.1389
future score data			
Συντελεστής gini	0.296	0.257	0.262
top decile lift	2.1729	2.0705	2.1672

Πίνακας 21: Αποτελέσματα με τις μεταβλητές συνόλου Σ1

Λαμβάνοντας υπόψη ότι τα αποτελέσματα της πρώτης εφαρμογής της λογιστικής παλινδρόμησης (σύνολο Σ1) αποτελούν ενδιάμεσο στάδιο, καθώς το μοντέλο που έχει αναπτυχθεί μέχρι εδώ δεν είναι ικανοποιητικό (όπως αναφέρθηκε στην ενότητα 3.2.3), δεν υπάρχει στον πίνακα 21 καταγραφή των αποτελεσμάτων της λογιστικής παλινδρόμησης. Τα αποτελέσματα της λογιστικής παλινδρόμησης καταγράφονται και αξιολογούνται στους ακόλουθους δύο πίνακες (22 και 23), μετά τις απαραίτητες επεξεργασίες των δεδομένων που αναλύθηκαν στην ενότητα 3.2.3. Με βάση τη λογική αυτή, πραγματοποιείται η 1^η τροποποίηση των ποιοτικών μεταβλητών, για να προκύψει το σύνολο μεταβλητών Σ2. Για αυτό το σύνολο μεταβλητών έτρεξαν τα ακόλουθα μοντέλα, όπως φαίνεται στον πίνακα 22:

	Λογιστική Παλινδρόμηση	Νευρωνικά Δίκτυα	PSVM	ASVM
current score data				
Συντελεστής gini	0.252	0.252	0.255	0.276
top decile lift	2.0507	2.0617	2.0728	2.1610
future score data				
Συντελεστής gini	0.238	0.238	0.248	0.266
top decile lift	1.9226	1.9852	2.0478	2.1103

Πίνακας 22: Αποτελέσματα με τις μεταβλητές συνόλου Σ2

Η λογική παραμένει ίδια και για τον πίνακα 23, ο οποίος περιλαμβάνει το σύνολο μεταβλητών Σ3, δηλαδή τα αποτελέσματα της 2^{ης} τροποποίησης των ποιοτικών μεταβλητών στα πλαίσια βελτίωσης των αποτελεσμάτων της λογιστικής παλινδρόμησης.

	Λογιστική Παλινδρόμηση	Νευρωνικά Δίκτυα	PSVM	ASVM
current score data				
Συντελεστής gini	0.262	0.264	0.255	0.272
top decile lift	2.0397	2.0728	2.1058	2.1610
future score data				
Συντελεστής gini	0.242	0.256	0.252	0.258
top decile lift	1.9340	2.1445	1.9966	2.0421

Πίνακας 23: Αποτελέσματα με τις μεταβλητές συνόλου Σ3

Στην περίπτωση της επιλογής μεταβλητών με τη χρήση λογιστικής παλινδρόμησης, παρατηρείται ότι καθώς οι βαθμίδες των ποιοτικών μεταβλητών μειώνονται, και συνεπώς οι μεταβλητές αυτές απλοποιούνται, τα αποτελέσματα των νευρωνικών δικτύων εμφανίζουν μείωση. Συγκρίνοντας ειδικότερα τα αποτελέσματα των νευρωνικών δικτύων της 1^{ης} παραλλαγής του συνόλου μεταβλητών (πίνακας 22) με εκείνα της 2^{ης} (πίνακας 23), παρατηρείται ότι τα αποτελέσματα της δεύτερης παραλλαγής είναι σαφώς καλύτερα της πρώτης, σε όλα τα κριτήρια αξιολόγησης. Στην περίπτωση της λογιστικής παλινδρόμησης κατά την ίδια σύγκριση (των δύο παραλλαγών των συνόλων δεδομένων), τα αποτελέσματα παρουσιάζουν μικρή βελτίωση, με το τελικό μοντέλο να παρουσιάζεται ελαφρώς βελτιωμένο σε σχέση με το παλιό. Συγκεκριμένα, οι τιμές της καμπύλης top decile lift στο σύνολο current score data παρουσιάζουν μικρή επιδείνωση, ενώ στο σύνολο future score data εμφανίζουν μικρή βελτίωση. Παρόμοια κινούνται και οι τιμές του συντελεστή gini, οι οποίες δείχνουν να βελτιώνονται στο σύνολο future score data, αλλά παρουσιάζουν μικρή μείωση στο σύνολο current score data. Όσον αφορά στα αποτελέσματα των PSVM κατά την απλοποίηση των μεταβλητών από το σύνολο Σ1 στο Σ2 και το Σ3, παρατηρείται επιδείνωση. Συγκρίνοντας ειδικότερα τα αποτελέσματα των PSVM της 1^{ης} παραλλαγής του συνόλου μεταβλητών (πίνακας 22) με εκείνα της 2^{ης} (πίνακας 23), παρατηρείται ότι τα αποτελέσματα της δεύτερης παραλλαγής είναι ελαφρώς καλύτερα της πρώτης. Πιο συγκεκριμένα, οι τιμές του κριτηρίου top decile lift στο σύνολο current score data εμφανίζουν σαφή βελτίωση, όπως επίσης βελτίωση εμφανίζουν και οι τιμές του συντελεστή gini στο σύνολο future score data. Στην περίπτωση του μοντέλου ASVM, κατά την απλοποίηση των μεταβλητών από το σύνολο Σ1 στο Σ2 και το Σ3, παρατηρείται βελτίωση των αποτελεσμάτων. Συγκρίνοντας πιο ειδικά τα αποτελέσματα των ASVM της 1^{ης} παραλλαγής του

συνόλου μεταβλητών (πίνακας 22) με εκείνα της 2^{ης} (πίνακας 23), παρατηρείται μείωση των τιμών στο σύνολο των κριτηρίων αξιολόγησης.

Στη συνέχεια ακολουθεί η μέθοδος του δέντρου απόφασης για την επιλογή μεταβλητών, και το σύνολο μεταβλητών Σ4 χρησιμοποιείται για να τρέξει ένα μοντέλο νευρωνικών δικτύων, ένα μοντέλο προσεγγιστικών μηχανών διανυσμάτων υποστήριξης (PSVM) και ένα μοντέλο προσθετικών μηχανών διανυσμάτων υποστήριξης (ASVM) (πίνακας 24):

	Δέντρο Απόφασης	Νευρωνικά Δίκτυα	PSVM	ASVM
current score data				
Συντελεστής gini	0.238	0.266	0.213	0.256
top decile lift	1.7972	2.1940	1.9184	1.9184
future score data				
Συντελεστής gini	0.220	0.230	0.217	0.218
top decile lift	1.8146	2.0364	1.7577	1.7975

Πίνακας 24: Αποτελέσματα με τις μεταβλητές συνόλου Σ4

Συγκρίνοντας τα αποτελέσματα των δύο προσεγγίσεων επιλογής μεταβλητών, παρατηρείται ότι τα αποτελέσματα των μοντέλων στην περίπτωση του δέντρου απόφασης είναι σαφώς χειρότερα από εκείνα της λογιστικής παλινδρόμησης.

Παρατηρώντας την επίδοση των μοντέλων συνολικά, είναι φανερό ότι το δέντρο απόφασης έχει τη χαμηλότερη τιμή στα κριτήρια αξιολόγησης σε σχέση με τα άλλα μοντέλα, αν και ως μέθοδος χρησιμοποιήθηκε μόνο σε ένα σύνολο δεδομένων και συνεπώς μπορεί να θεωρηθεί ότι δεν υπάρχει επαρκής πληροφορία για την εξαγωγή ασφαλών συμπερασμάτων. Ακολουθεί η μέθοδος της λογιστικής παλινδρόμησης με αισθητά καλύτερα αποτελέσματα σε σχέση με το δέντρο απόφασης, και εκείνη των προσεγγιστικών μηχανών διανυσμάτων υποστήριξης. Μεγαλύτερη ακρίβεια πρόβλεψης εμφάνισαν οι προσθετικές μηχανές διανυσμάτων υποστήριξης (ASVM), τα αποτελέσματα των οποίων φαίνονται περισσότερο ευσταθή σε σχέση με εκείνα των άλλων μοντέλων, καθώς επίσης αξιόλογη επίδοση παρουσίασαν και τα νευρωνικά δίκτυα, τα οποία λειτουργούν επιτυχημένα με μεγάλο όγκο πληροφοριών. Στα πλαίσια αξιολόγησης των μεθόδων πρέπει επίσης να σημειωθεί ότι η φάση εκπαίδευσης των δέντρων απόφασης, αλλά και των

νευρωνικών δικτύων, διήρκησε σημαντικά περισσότερη ώρα από εκείνη της λογιστικής παλινδρόμησης και των μηχανών διανυσμάτων υποστήριξης.

3.3.2. Συνδυασμός μοντέλων

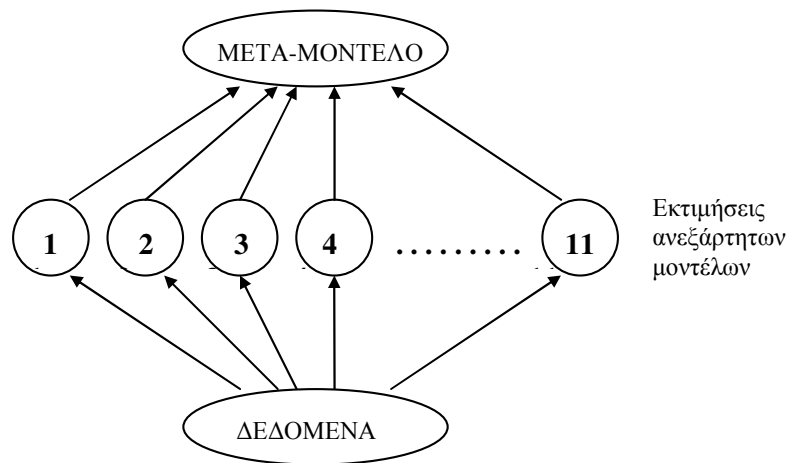
Σε αυτό το τμήμα έγινε προσπάθεια συνδυασμού των αποτελεσμάτων των μεθόδων που χρησιμοποιήθηκαν παραπάνω, ώστε με αυτό τον τρόπο να αξιολογηθεί η ακρίβεια πρόβλεψης επιλεγμένων μοντέλων σε νέα, καλύτερη βάση. Οι μέθοδοι οι οποίες επιλέχθηκαν να αναλύσουν την ικανότητα πρόβλεψης του νέου αυτού συνόλου δεδομένων, είναι εκείνες που θεωρήθηκαν περισσότερο επιτυχημένες από τις υπόλοιπες όσον αφορά την ακρίβεια πρόβλεψής τους στα κριτήρια αξιολόγησης που αναλύθηκαν προηγούμενα (νευρωνικά δίκτυα, προσθετικές μηχανές διανυσμάτων υποστήριξης).

Λαμβάνοντας υπόψη τα αποτελέσματα της προηγούμενης ενότητας, επιλέχθηκαν εκείνοι οι συνδυασμοί των μοντέλων-συνόλων μεταβλητών που εμφάνισαν την καλύτερη ακρίβεια πρόβλεψης. Πιο συγκεκριμένα, επιλέχθηκαν οι εξής συνδυασμοί (οι τιμές των μοντέλων αυτών στα κριτήρια αξιολόγησης υπάρχουν στους πίνακες 21 μέχρι 24):

1. Μοντέλο νευρωνικού δικτύου στο σύνολο μεταβλητών Σ1
2. Μοντέλο νευρωνικού δικτύου στο σύνολο μεταβλητών Σ2
3. Μοντέλο νευρωνικού δικτύου στο σύνολο μεταβλητών Σ3
4. Μοντέλο νευρωνικού δικτύου στο σύνολο μεταβλητών Σ4
5. Μοντέλο νευρωνικού δικτύου στο σύνολο μεταβλητών Σ5⁵
6. Μοντέλο PSVM στο σύνολο δεδομένων Σ1
7. Μοντέλο PSVM στο σύνολο δεδομένων Σ2
8. Μοντέλο PSVM στο σύνολο δεδομένων Σ3
9. Μοντέλο ASVM στο σύνολο δεδομένων Σ1
10. Μοντέλο ASVM στο σύνολο δεδομένων Σ2
11. Μοντέλο ASVM στο σύνολο δεδομένων Σ3

⁵ Το σύνολο Σ5 αφορά στις μεταβλητές που προέκυψαν μετά τη χρήση της μονομεταβλητής ανάλυσης, οι οποίες αποτελούν ένα σύνολο μεταβλητών αρχικού σταδίου. Θεωρήθηκε καλό να χρησιμοποιηθεί αυτό το σύνολο σε ένα μοντέλο νευρωνικών δικτύων (αφού τα νευρωνικά δίκτυα έχουν ικανοποιητική ακρίβεια πρόβλεψης) ώστε συμμετέχουν στην ανάλυση και αυτά τα αποτελέσματα.

Κάθε ένα από τα μοντέλα αυτά παρέχει εκτός των άλλων και την πιθανότητα που έχει κάθε πελάτης να φύγει από την επιχείρηση (churn probability). Το γεγονός αυτό δίνει τη δυνατότητα δημιουργίας ενός νέου συνόλου δεδομένων (πιθανοτήτων), το οποίο περιλαμβάνει έντεκα στήλες πιθανοτήτων απώλειας πελατών (μία για κάθε μοντέλο 1 μέχρι 11, ως ανεξάρτητες μεταβλητές), και το οποίο χρησιμοποιήθηκε για την ανάπτυξη ενός μοντέλου νευρωνικών δικτύων και ενός μοντέλου προσθετικών μηχανών διανυσμάτων υποστήριξης (σχήμα 18).



Σχήμα 18: Πλαίσιο συνδυασμού μοντέλων

Με βάση λοιπόν τα νέα αυτά δεδομένα προέκυψε ο ακόλουθος πίνακας αποτελεσμάτων (πίνακας 25):

	Νευρωνικά Δίκτυα	ASVM
current score data		
Συντελεστής gini	0.294	0.292
top decile lift	2.4366	2.3704
future score data		
Συντελεστής gini	0.286	0.288
top decile lift	2.2241	2.2298

Πίνακας 25: Αποτελέσματα από το συνδυασμό των αποτελεσμάτων επιλεγμένων μοντέλων

Συγκρίνοντας τον πίνακα 25 με τους πίνακες 21 έως 24, είναι εμφανές ότι οι ακρίβειες πρόβλεψης βελτιώθηκαν αισθητά. Συγκρίνοντας τώρα το μοντέλο των νευρωνικών δικτύων με εκείνο των ASVM παρατηρείται μικρή υπεροχή των πρώτων έναντι των δεύτερων.

Επιπλέον, στα πλαίσια αξιολόγησης της ακρίβειας πρόβλεψης των μοντέλων, πραγματοποιήθηκε ένας ακόμα, εναλλακτικός συνδυασμός τους. Επιλέχτηκε ένα σύνολο δεδομένων με τον ίδιο τρόπο όπως και παραπάνω (δεδομένα πιθανοτήτων απώλειας πελατών), το οποίο όμως προήλθε από ένα ευρύτερο σύνολο μοντέλων. Τα μοντέλα που χρησιμοποιήθηκαν στο σημείο αυτό αποτελούν όλα τα μοντέλα (συνολικά 16 στον αριθμό) που παρουσιάζονται στην ενότητα 3.3.1 (βλ. πίνακες 21 έως 24), και πιο συγκεκριμένα, χρησιμοποιήθηκαν τα μοντέλα 1 έως 11, που παρουσιάστηκαν προηγούμενα, με τα εξής επιπλέον μοντέλα:

12. Μοντέλο PSVM στο σύνολο δεδομένων Σ4
13. Μοντέλο ASVM στο σύνολο δεδομένων Σ4
14. Μοντέλο λογιστικής παλινδρόμησης στο σύνολο δεδομένων Σ2
15. Μοντέλο λογιστικής παλινδρόμησης στο σύνολο δεδομένων Σ3
16. Μοντέλο δέντρου απόφασης στο σύνολο δεδομένων Σ5

Τα αποτελέσματα (πιθανότητες απώλειας πελατών) των μοντέλων που μόλις παρουσιάστηκαν (1 έως 16) εισήχθησαν σε ένα μοντέλο λογιστικής παλινδρόμησης ώστε να εκτιμηθεί η σημαντικότητά τους σε σχέση με το δείκτη απώλειας πελατών. Τα αποτελέσματα της λογιστικής παλινδρόμησης μείωσαν το σύνολο των μεθόδων κατά πολύ, και πιο συγκεκριμένα, τα μοντέλα που έμειναν στην ανάλυση είναι πέντε μοντέλα νευρωνικών δικτύων (1, 2, 3, 4, 5), δύο μοντέλα PSVM (7, 12), δύο μοντέλα ASVM (9, 13), και ένα μοντέλο δέντρου απόφασης (16).

Οι πιθανότητες απώλειας πελατών των μοντέλων αυτών αποτέλεσαν συνεπώς ένα νέο σύνολο δεδομένων το οποίο χρησιμοποιήθηκε ως είσοδος ενός μοντέλου νευρωνικών δικτύων και ενός μοντέλου προσθετικών μηχανών διανυσμάτων υποστήριξης, για να προκύψουν τα αποτελέσματα Νευρωνικά δίκτυα και ASVM του πίνακα 26:

	Νευρωνικά Δίκτυα	ASVM
current score data		
Συντελεστής gini	0.342	0.336
top decile lift	2.4035	2.4035
future score data		
Συντελεστής gini	0.332	0.334
top decile lift	2.3891	2.2810

Πίνακας 26: Αποτελέσματα από τον εναλλακτικό συνδυασμό μοντέλων

Συγκρίνοντας τα αποτελέσματα του πίνακα 26 με εκείνα του 25, φαίνεται ότι τα μοντέλα των νευρωνικών δικτύων εμφανίζουν αρκετή βελτίωση στο συντελεστή gini, αλλά εμφανίζουν αυξομειώσεις στις τιμές της καμπύλης top decile lift. Συγκεκριμένα οι τιμές του συντελεστή gini παρουσιάζουν αύξηση και στα δύο σύνολα δεδομένων, αλλά οι τιμές της καμπύλης top decile lift εμφανίζονται μειωμένες στον πίνακα 26 στο σύνολο current score data, ενώ αντίθετα παρουσιάζουν βελτίωση στο σύνολο future score data.. Τα μοντέλα των ASVM εμφανίζουν σαφή βελτίωση τόσο στον συντελεστή gini, στα αποτελέσματα του πίνακα 26 σε σχέση με εκείνα του 25, όσο και στην καμπύλη top decile lift. Συγκρίνοντας τώρα τα αποτελέσματα του μοντέλου των νευρωνικών δικτύων με εκείνα του μοντέλου ASVM στον πίνακα 26, παρατηρείται υπεροχή των νευρωνικών δικτύων έναντι των προσθετικών μηχανών διανυσμάτων υποστήριξης.

4. Συμπεράσματα και επεκτάσεις

Η προβλεπτική μοντελοποίηση (η στατιστική διαδικασία της αξιολόγησης και της κατάλληλης στοχοθέτησης πελατών για τους σκοπούς μιας εκστρατεία μάρκετινγκ) είναι ένα σημαντικό εργαλείο του τμήματος διαχείρισης σχέσεων πελατών (CRM). Ο στόχος της προβλεπτικής μοντελοποίησης είναι να προβλέπει τις επόμενες κινήσεις των πελατών, και συνεπώς να επιτρέπει στις εταιρίες να οργανώνουν τις προσπάθειές τους στον τομέα του μάρκετινγκ πιο αποδοτικά. Η τάση που εμφανίζουν οι πελάτες σήμερα να εγκαταλείπουν την εταιρία με την οποία συνεργάζονται λόγω των δελεαστικών προσφορών άλλων εταιριών, έρχεται να τονίσει ακόμα περισσότερο τη σημαντικότητα της πρόβλεψης της συμπεριφοράς των πελατών. Πιο συγκεκριμένα, το αυξημένο επίπεδο του ανταγωνισμού στο χώρο της παροχής τηλεπικοινωνιακών υπηρεσιών και η αύξηση των ετησίων ποσοστών απώλειας πελατών, απαιτεί την ανάπτυξη μοντέλων που προβλέπουν με ακρίβεια ποιοι πελάτες είναι πιθανότερο να φύγουν από μια εταιρία και συνεπώς, δίνουν τη δυνατότητα λήψης κατάλληλων προληπτικών μέτρων.

Στην παρούσα μεταπτυχιακή εργασία έγινε ανάλυση του φαινομένου αυτού, της απώλειας πελατών, από θεωρητικής βάσης και πραγματοποιήθηκε βιβλιογραφική έρευνα των παραγόντων που την επηρεάζουν. Επιπλέον, χρησιμοποιήθηκε μία εκτενής βάση δεδομένων ενός τηλεπικοινωνιακού πάροχου για να αναπτυχθούν μοντέλα πρόβλεψης απώλειας πελατών και να γίνει αξιολόγηση της ακρίβειάς τους, των δυνατών και αδύναμων σημείων τους. Στα πλαίσια αυτά μπορούν να γίνουν ορισμένες συμπερασματικές παρατηρήσεις. Κατ' αρχάς, η ανάπτυξη και αξιολόγηση ενός μοντέλου πρόβλεψης απώλειας πελατών αφορά σε μία πληθώρα παραγόντων, όπως στην επιλογή της κατάλληλης μεθοδολογίας ανάλογα με τον όγκο και τη μορφή των δεδομένων, στην επιλογή της κατάλληλης τεχνικής επιλογής μεταβλητών, στις τεχνικές αντικατάστασης των ελλিপών τιμών, στις απαιτήσεις σε υπολογιστικό φόρτο και βέβαια στο χρόνο που χρειάζεται ένα μοντέλο για να περατώσει τη φάση εκπαίδευσης.

Όσον αφορά στην ακρίβεια πρόβλεψης των μοντέλων που χρησιμοποιήθηκαν στην εργασία, το δέντρο απόφασης παρουσίασε τη χαμηλότερη τιμή στα κριτήρια αξιολόγησης (top decile lift και gini coefficient), με τη μέθοδο της λογιστικής

παλινδρόμησης να ακολουθεί με αισθητά καλύτερα αποτελέσματα. Στα ίδια πλαίσια κινήθηκε επίσης και η μέθοδος των προσεγγιστικών μηχανών διανυσμάτων υποστήριξης (PSVM). Μεγαλύτερη ακρίβεια πρόβλεψης, και συνεπώς τα πιο ικανοποιητικά αποτελέσματα, εμφάνισαν οι προσθετικές μηχανές διανυσμάτων υποστήριξης (ASVM), καθώς επίσης αξιολογη επίδοση παρουσίασαν και τα νευρωνικά δίκτυα. Ειδικότερα μπορεί να σημειωθεί ότι οι προσθετικές μηχανές διανυσμάτων υποστήριξης έδειξαν ευσταθή χαρακτήρα, με τα αποτελέσματά τους να συμφωνούν στα κριτήρια αξιολόγησης και στα σύνολα δεδομένων (current score data και future score data), και να μην εμφανίζονται περιπτώσεις αυξομειώσεων των τιμών από τις οποίες δεν μπορούν να βγουν σαφή συμπεράσματα. Όσον αφορά στα νευρωνικά δίκτυα, φάνηκε να λειτουργούν επιτυχημένα με μεγάλο όγκο πληροφοριών, ενώ στα πλαίσια αξιολόγησης των μεθόδων πρέπει επίσης να σημειωθεί ότι η φάση εκπαίδευσης των δέντρων απόφασης, αλλά και των νευρωνικών δικτύων, διήρκεσε σημαντικά περισσότερη ώρα από εκείνη της λογιστικής παλινδρόμησης και των μηχανών διανυσμάτων υποστήριξης. Τέλος παρατηρήθηκε ότι συνδυάζοντας τα αποτελέσματα των πιο αποδοτικών μεθόδων μπορεί να επιτευχθεί σημαντική βελτίωση των αποτελεσμάτων.

Το πρόβλημα της διαχείρισης της απώλειας πελατών και τα στατιστικά προβλήματα πρόβλεψής της με την καλύτερη ακρίβεια θα συνεχίσουν να απασχολούν την έρευνα σε αυτό τον τομέα, καθώς η σημαντικότητά του θέματος είναι επιτακτική και οι επεκτάσεις του ποικίλες.

Η αντιμετώπιση των ελλিপών τιμών στη βάση δεδομένων στην παρούσα μεταπτυχιακή εργασία έγινε με τη διαγραφή των πελατών εκείνων που έφεραν τα περισσότερα κενά, και με τη χρήση ψευδομεταβλητών που υποδήλωναν την έλλειψη κάποιας τιμής. Αν και αυτή η αντιμετώπιση τελικά έφερε ικανοποιητικά αποτελέσματα στην ακρίβεια των μοντέλων, θα μπορούσε να εξεταστεί η χρήση πιο εξελιγμένων μεθόδων διαχείρισης ελλিপών στοιχείων. Επιπλέον, αξίζει να μελετηθεί η απόδοση και άλλων μοντέλων, με τεχνικές όπως οι γενετικοί αλγόριθμοι ή οι αλυσίδες markov, οι οποίες πρόσφατα μπήκαν στο προσκήνιο της πρόβλεψης απώλειας πελατών, με σκοπό να βελτιώσουν την ακρίβειά της.

Επιπλέον, πέρα από την ανάλυση και πρόβλεψη της απώλειας πελατών, εξίσου σημαντική είναι η διαχείρισή της και η ενίσχυση της αφοσίωσης των πελατών. Σε αυτά τα πλαίσια, τα δεδομένα που χρησιμοποιήθηκαν στην εργασία ήταν κατά κύριο λόγο δεδομένα συμπεριφοράς πελατών, δεδομένα αλληλεπίδρασης της εταιρίας

με τον πελάτη, δημογραφικά δεδομένα κ.α., αλλά η βάση δεδομένων δεν συμπεριελάμβανε δεδομένα όσον αφορά στις κινήσεις μάρκετινγκ της επιχείρησης και στην ανταπόκριση που αυτές δέχτηκαν. Δηλαδή, δεν υπήρχαν δεδομένα πάνω σε παλαιότερες προσφορές μάρκετινγκ για τη μείωση της απώλειας πελατών, τα οποία θα επέτρεπαν τη στοχοθέτηση πελατών με βάση τόσο τη συμπεριφορά τους σε παλαιότερες προσφορές μάρκετινγκ, όσο και την πιθανότητα που έχουν να φύγουν από την εταιρία.

Τέλος να σημειωθεί ότι σαφή επέκταση του θέματος αποτελεί η επέκταση της ανάλυσης και πρόβλεψης απώλειας πελατών σε άλλους τομείς, πέραν εκείνου των τηλεπικοινωνιών, καθώς ανάλογα με την εταιρία και τον τομέα στον οποίο ανήκει, τα δεδομένα και οι παράγοντες που επηρεάζουν την απώλεια πελατών διαφοροποιούνται σημαντικά.

5. Βιβλιογραφία

- Ahn, J. H., S. P. Han and Y. S. Lee (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy* 30, 552–568.
- Alker, H. R. (1965). *Mathematics and politics*. New York: The Macmillan Company.
- Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*, Cary, NC: SAS Institute.
- Anderson, E. and M. Sullivan (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing Science* (12), 125-143.
- Au, W. H., C. K. Chan and X. Yao (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on evolutionary computation* 7(6), 532-545.
- Bae, S. M., S. H. Ha and S. C. Park (2005). A web-based system for analyzing the voices of all center customers in the service industry. *Expert Systems with Applications* 28(1), 29–41.
- Baesens, B., S. Viaene, D. Van den Poel, J. Vanthienen and G. Dedene (2002). Bayesian neural network learning for repeat purchase modeling in direct marketing. *European Journal of Operational Research*, 138(1), 191–211.
- Bearden, W. O. and J. E. Teel (1983). Selected determinants of consumer satisfaction and complaint reports. *Journal of Marketing Research* 20(2), 21–28.
- Beatty, S. E. and S. M. Smith (1987). External search effort: An investigation across several product categories? *Journal of Consumer Research* (1), 83-95.
- Bhattacharyya, S. and P. C. Pendharkar (1998). Inductive, evolutionary and neural computing techniques for discrimination: A comparative study. *Decision Sciences* 29(4), 871–900.
- Bendapudi, N. and L. L. Berry (1997). Customers' motivations for maintaining relationships with service providers. *Journal of Retailing* 73(1), 15-37.
- Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science* 17(1), 45-65.

- Bolton, R. N., P. K. Kannan and M. D. Bramlett (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science* 28(1), 95–108.
- Boone, D. S. and M. Roehm (2002). Retail segmentation using artificial neural networks. *International journal of research in marketing* 19(3), 287-301.
- Boulding, W., A. Kalra, R. Staelin and V. A. Zeithaml (1993). A dynamic process model of service quality: From expectations to behavioral intentions. *Journal of Marketing Research* (30), 7-27.
- Breiman, L., K. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and regression trees*. Chapman and Hall.
- Bucklin, R. E. and J. M. Lattin (1992). A two-state model of purchase incidence and brand choice. *Marketing Science* 10, 24–39.
- Buckinx, W. and D. Van den Poel, (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research* 164(1), 252–268.
- Buckinx W., G. Verstraeten and D. Van den Poel (2005). Predicting customer loyalty using the internal transactional database. *Expert Systems with Applications* 32(1), 125-134.
- Burez, J. and D. Van den Poel, (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications* 32(2), 277–288.
- Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications* 36(3), 4626–4636
- Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2), 121-167.
- Celsi, R. L. and J. C. Olson (1988). The role of involvement in attention and comprehension processes. *Journal of Consumer Research* (15), 210-224.
- Cerny, P. A. (2001). Data mining and neural networks from a commercial perspective. *ORSNZ Conference Twenty Naught One*.
- Chandar, M., A. Laha and P. Krishna (2006). Modeling churn behavior of bank customers using predictive data mining techniques. *National conference on soft computing techniques for engineering applications*.
- Chen, M. C., A. L. Chiu, and H. H. Chang (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications* 28(4), 773-781.

- Chiang, D., Y. Wang, S. Lee and C. Lin (2003). Goal oriented sequential pattern for network banking churn analysis. *Expert systems with applications* 25(3), 293-302.
- Colgate, M. and P. J. Danaher (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science* 28(3), 375-387.
- Colgate M. and B. Lang (2001). Switching barriers in consumer markets: an investigation of the financial services industry. *Journal of consumer marketing* 18(4), 332-347.
- Coussement, K. and D. Van den Poel (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications* 34(1), 313–327.
- Crespo, F. and R. Weber (2004). A methodology for dynamic data mining based on fuzzy clustering, *Fuzzy Sets and Systems* 150(2), 267-284.
- Datta, P., B. Masand, D. R Mani and B. Li (2001). Automated cellular modeling and prediction on a large scale. *Issues on the application of data mining* 14(6), 485-502.
- Desai, M. (2006). Customer Churn-A reason to churn. Available at SSRN: <http://ssrn.com/abstract=957208>.
- Dick, A. S., and K. Basu (1994). Customer loyalty: Toward an integrated conceptual framework. *Journal of the Academy of Marketing Science* 22(2), 99–113.
- Doumpos, M., C. Gaganis and F. Pasiouras (2005). Explaining qualifications in audit reports using a support vector machine methodology. *Intelligent Systems in Accounting, Finance & Management* 13(4), 197-215.
- Doumpos, M., C. Zopounidis and V. Golfinopoulou (2007). Additive Support Vector Machines for pattern classification. *IEEE Transactions on systems, Man, and Cybernetics, Part B: Cybernetics* 37(3), 540-550.
- Drew, J. H., D. R. Mani, A. L. Betz and P. Datta (2001). Targeting customers with statistical and data-mining techniques. *Journal of Service Research* 3 (3), 205-219.
- Eiben, A. E., A. E. Koudijs and F. Slisser (1998). Genetic modelling of customer retention. *Lecture Notes in Computer Science* 1391, 178–186.

- Evanschitzky, H. and M. Wunderlich (2006). An examination of moderator effects in the four-stage loyalty model. *Journal of Service Research* (8), 330-345.
- Fadalla, A., C. H. Lin (2001). An analysis of the applications of neural networks in finance. *Interfaces* 31(4), 112-122.
- Fawcett, T. (2004). ROC Graphs: Notes and practical considerations for researchers. *Machine Learning* 31, 1-38.
- Fornell, C. (1992). A national customer satisfaction barometer: The Swedish experience. *Journal of Marketing* (56), 6-21.
- Fung, G. and O. L. Mangasarian, (2001). Proximal Support Vector Machine Classifiers. *International Conference on Knowledge Discovery and Data Mining*, 77-86.
- Ganesh, J., M. J. Arnold and K. E. Reynolds (2000). Understanding the customer base of service providers: An examination of the differences between Switchers and Stayers? *Journal of Marketing* (64), 65-87.
- Gerpott, T. J., W. Rams and A. Schindler (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. *Telecommunications Policy* 25(4), 249-269.
- Geyskens, I. and Jan-Benedict E. M. Steenkamp (2000). Economic and social satisfaction: Measurement and relevance to marketing channel relationships. *Journal of Retailing* (76), 11-32.
- Gupta, S. and V. Zeithaml (2005). Customer metrics and their impact on financial performance. *Marketing science* 25(6), 718-739.
- Hadden, J., A. Tiwari, R. Roy and D. Ruta (2006). Churn prediction: Does technology matter? *International Journal of Intelligent Systems and Technologies* 1(1), 104-110.
- Hadden, J., A. Tiwari, R. Roy, and D. Ruta (2007). Computer Assisted Customer Churn Management: State-Of-The-Art and Future Trends. *Computers & Operations Research* 34(10), 2902-2917.
- Haykin S., M. (1994). *Neural networks: a comprehensive foundation*. Cambridge University Press New York.
- Hoch, S. J. and J. Deighion. (1989). Managing what consumers learn from experience. *Journal of Marketing* (53), 1-20.
- Hoekstra, J. and E. Huizingh (1999). The lifetime value concept in customer based marketing. *Journal of Market Focused Management* 3, 257-274.

- Huang, C. S., Y. J. Lin and C. C. Lin (2008). Implementation of classifiers for choosing insurance policy using Decision Trees: A Case Study. *WSEAS Transactions on computers* 10(7), 1679-1689.
- Hung, S. Y., D. C. Yen, and H. Y. Wang (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515–524.
- Hwang, H., T. Jung, and E. Suh (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications* 26(2), 181-188.
- Jain, D. and S. S. Singh (2002). Customer lifetime value research in marketing: a review and future directions. *Journal of interactive marketing* 16(2), 34-46.
- Jenamani, M., P. K. J. Mohapatra and S. Ghose (2003). A Stochastic Model of E-Customer Behaviour. *Electronic commerce research and applications* 2(1), 81-94.
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics*, 29(2), 119–127.
- Keaveney, S. M. and M. Parthasarathy (2001). Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. *Journal of the Academy of Marketing Science* 29(4), 374-390.
- Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing*, 59(2), 71–82.
- Kim, M. K, M. C. Park, and D. H. Jeong, (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy* 28(2), 145–159.
- Kim, S. Y., T. S. Jung, E. H. Suh and H. S. Hwang (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications* 31(1), 101–107.
- Kim, S., K. S. Shin, and K. Park (2005). An application of support vector machines for customer churn analysis: Credit card case. *Lecture Notes in Computer Science*, 3611, 636–647.
- Kim, H. S. and C. H. Yoon (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, 28(9/10), 751–765.

- Kitayama, M., R. Matsubara and Y. Izui (2002). Application of data mining to customer profile analysis in the power electric industry. Power Engineering Society Winter Meeting 1, 632-634.
- Kumar, V. and G. Ramani (2004). Taking customer lifetime value analysis to the next level. *Journal of Integrated Communications*, 27–33.
- Lariviere, B. and D. Van den Poel (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2), 472–484.
- Lariviere, B., and D. Van den Poel (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications* 27(2), 277–285.
- Lee, T., C. Chiu, Y. Chou and C. Lu (2004). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis* 50(4), 1113-1130.
- Lee, M. and L. F. Cunningham (2001). A cost/benefit approach to understanding service loyalty. *Journal of Services Marketing* 15(2), 113–130.
- Lee, G.H., S.M. Shin and Y.W. Park (2003). Studies on the determinants of subscribers' churning in the high speed internet service market. KISDI report.
- Lemmens, A. and C. Croux (2003). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research* 43(2), 276-286.
- Loh, W. Y. and Y. S. Shih (1997). Split selection methods for Classification Trees, *Statistica Sinica* (7), 815-840.
- Lu, H., R. Setiono and H. Liu (1996). Effective Data Mining Using Neural Networks. *IEEE Transactions on knowledge and data engineering* 8(6), 957-961
- Madden, G. and G. Savage (1999). Subscriber Churn in the Australian Isp Market, *Information economics and policy* 11, 195-207.
- Malthouse, E. C. (2001). Assessing the performanse of direct marketing scoring models. *Journal of interactive marketing* 15(2), 49-62.
- Mangasarian, O. L. (2000). Generalized support vector machines. A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Edts.), *Advances in Large Margin Classifiers*, 135-146, Cambridge, MA, MIT Press.

- Mangasarian, O. L. and D. R. Musicant (2001). Active support vector machine classification. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, 577–583.
- Mihelis, G., E. Grigoroudis, Y. Siskos, Y. Politis and Y. Malandrakis (2001). Customer satisfaction measurement in the private bank sector, *European Journal of Operational Research* 130(2), 347-360.
- Mitra, S., S. K. Pal and P. Mitra (2002). Data mining in soft computing Framework: A Survey. *IEEE Transactions on neural networks* 13(1), 3-14.
- Mozer, M. C., R. Wolniewicz, D. B. Grimes, E. Johnson and H. Kaushansky (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunication industry. *IEEE Transactions on Neural Networks* 11(3), 690–696.
- Muata, K. and O. Bryson (2004). Evaluation of Decision Trees: A MultiCriteria Approach. *Computers and Operational Research* 31(11), 1933-1945.
- Mutanen T. (2006). Customer churn analysis-a case study. Helsinki University of technology, systems analysis laboratory, department of engineering physics and mathematics, Independent research project in applied mathematics.
- Neslin, S. A., S. Gupta, W. Kamakura, J. Lu and C. H. Mason (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* 43(2), 204–211.
- Neslin, S. A., S. Gupta, W. Kamakura, J. Lu and C. Mason (2004). Defection detection: Improving predictive accuracy of customer churn models. Working Paper, Teradata Center at Duke University.
- Neslin, S., S. Siang, S. Gupta, W. Kamakura, J. Lu, and C. Mason (2002) Churn Modeling Tournament. Teradata center for customer relationship management at Duke University.
- Ng, K. and H. Liu (2001). Customer retention via data mining. *Artificial Intelligence Review* 14(6), 569–590.
- Oliva, T. A., R. L. Oliver, and W. O. Bearden. 1995. The relationships among consumer satisfaction, involvement, and product performance: A catastrophe theory application? *Behavioral Science* (40), 104-132.
- Oliver, R. L. (1997). *Satisfaction: A behavioral perspective on the consumer*. Boston: Irwin McGraw-Hill.

- Pendharkar, P. C. (2007). A comparison of gradient ascent, gradient descent and genetic-algorithm based artificial neural networks for the binary classification problem. *Expert Systems*, 24(2), 65–86.
- Pendharkar, P. C. and S. Nanda (2006). A misclassification cost minimizing evolutionary-neural classification approach. *Naval Research Logistics*, 53(5), 432–447.
- Pendharkar, P. C. (2009). Genetic algorithm based neural network approaches for predicting churn, in cellular wireless network services. *Expert Systems with Applications* 36(3), 6714–6720.
- Pritchard, M. P., E. H. Mark and R. H. Dennis 1999. Analyzing the commitment-loyalty link in service contexts. *Journal of the Academy of Marketing Science* 27(3), 333-348.
- Qi, J., Zhang, Y., Shu, H., Li, Y. and L. Ge, (2006). Churn prediction with limited information in fixed-line telecommunication. *Proceedings of the 5th International Symposium on Communication Systems Networks and Digital Signal Processing*, New York: Springer, 423–426.
- Ramachandran, R. (2004). Applications of neural networks and data mining. A Paper Presented to Dr. Beach, The School of Technology, Indiana State University Terre Haute, Indiana, In Partial Fulfillment of the Requirements for ECT 680.
- Reichheld, F. F. (1996). *The loyalty effect*. Boston: Harvard Business School Press.
- Reichheld, F. F., W.E Jr. Sasser (1990). Zero defections: quality comes to services. *Harvard business review*, 105-111.
- Richins, M. L. and P. H. Bloch. (1991). Post-purchase product satisfaction: incorporating the effects of involvement and time? *Journal of Business Research* (23), 145-158.
- Rumelhart D.E., G.E. Hinton, R.J Williams (1986). Learning internal representations by error propagation. In Rumelhart D.E and McClelland J. A. The MIT press.
- Rust, R. T., K. N. Lemon and V. A. Zeithaml (2004). Return on Marketing: Using Customer Equity to Focus Marketing Strategy. *Journal of Marketing* (68), 109-127.
- Rust, R. T. and A. J. Zahorik (1993). Customer Satisfaction, Customer Retention, and Market Share, *Journal of retailing* 69(2), 193-215.
- Rygielski, C., Wang, J. and Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society* 24(4), 483-502.

- Schölkopf B. and A. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press.
- Siber, R. (1997). Combating the churn phenomenon. *Telecommunications*, 31(10), 77–81.
- Soldin O., A. Nandedkar, K. Japal, M. Stein, S. Mosee, P. Magrab, S. Lai and S. Lamm (2002). Newborn thyroxine levels and childhood ADHD. *Clinical Biochemistry* 35(2), 131-136.
- Suh, E., S. Lim, H. Hwang and S. Kim (2004). A prediction model for the purchase probability of anonymous customers to support real time web marketing: A case study. *Expert Systems with Applications* 27(2), 245–255.
- Sun, Z., G. Bebis and R. Miller (2004). Object detection using feature subset selection. *Pattern Recognition* 37, 2165-2176.
- Telephony Online (2002). Standing by Your Carrier. [available at http://telephonyonline.com/mag/telecom_standing_carrier/index.html].
- Thompson, B. (2005). *The Loyalty Connection: Secrets to Customer Retention and Increased Profits*. RightNow Technologies and CRMguru.
- Tsai, C. Y. and C. C. Chiu (2004). A purchased-based market segmentation methodology. *Expert Systems with Applications* 27(2), 265–276.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York.
- Vapnik V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons. New York USA.
- Verhoef, P. C. (2002). *The Joint Effect of Relationship Perceptions, Loyalty Program and Direct Mailings on Customer Share Development*. Working Paper, Erasmus Research Institute in Management.
- Wang, Y. F., D. A. Chiang, M. H. Hsu, C. J. Lin and I. L. Lin (2009). A recommender system to avoid customer churn: A case study. *Expert Systems with Applications* 36(4), 8071–8075.
- Weiss G. M. (2004). Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorations Newsletter* 6(1), 7-19.
- Wieringa, J. E. and P. C. Verhoef (2007). Understanding customer switching behavior in a liberalizing service market: An exploratory study. *Journal of Service Research* 10(2), 174-186.
- Wireless Review (2000). They love me, they love me not. 38–42, [available at http://connectedplanetonline.com/wireless/mag/wireless_love_love_not/]

- Xie, Y., X. Li, E.W. T. Ngai and W. Ying (2009). Customer churn prediction using improved balanced random forests. *Expert systems with applications* 36(3), 5445–5449.
- Yan, L., R. H. Wolniewicz, and R. Dodier (2004). Predicting customer behavior in telecommunications. *IEEE Intelligent System* 19(2), 50-58.
- Zhao, Y., B. Li, X. Li, W. Liu and S. Ren, (2005). Customer churn prediction using improved one-class support vector machine. *Lecture Notes in Computer Science* 3584, 300–306.
- Γρηγορούδης, Ε. (2003). Εισαγωγή στην ποιότητα των υπηρεσιών και την ικανοποίηση πελατών. Σημειώσεις μαθήματος. Πολυτεχνείο Κρήτης, Χανιά.
- Τσιώρης, Γ. (2007). Μεθοδολογίες Υπολογιστικής Νοημοσύνης στην Πρόβλεψη Χρηματιστηριακών Τάσεων: Εφαρμογή στο Χρηματιστήριο Αξιών Αθηνών. Διπλωματική διατριβή. Πολυτεχνείο Κρήτης, Χανιά.
- http://en.wikipedia.org/wiki/Decision_tree
- http://en.wikipedia.org/wiki/Logistic_regression
- http://en.wikipedia.org/wiki/Neural_network
- <http://en.wikipedia.org/wiki/Overfitting>
- http://prlab.ceid.upatras.gr/courses/simeiwseis/NNET_HTML/Chapter2/NNET_chapter24.htm
- http://sourcetherapies.com/shift_moleculesofemotions.html
- www.boozallen.com
- www.customerthink.com
- www.datamine.gr
- www.expresscomputeronline.com/20030922/indiatrends01.shtml
- www.statsoft.com/textbook/neural-networks/
- www.walkerinformation.com

6. Παράρτημα Α: Αρχικό σύνολο μεταβλητών και περιγραφή τους

Ποσοτικές Μεταβλητές	Επεξήγηση
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_PEA_V_MEAN	Mean unrounded minutes of use of peak voice calls
MOU_PEA_V_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

RET DAYS	Number of days since last retention call
REV_MEAN	Mean monthly revenue (charge amount)
REV_RANGE	Range of revenue (charge amount)
RM CALLS	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
THREWAY_MEAN	Mean number of three way calls
THREWAY_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

Ποιοτικές Μεταβλητές	Επεξήγηση
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