Αθήνα, 2018

Polyzogopoulos Pavlos (bapt1730)

Dimitris Karlis

Περιεχόμενα

[Εισαγωγή 3](#_Toc511594933)

[Μεθοδολογία 3](#_Toc511594934)

[Ανάλυση δεδομένων 4](#_Toc511594935)

[Αποτελέσματα ανάλυσης 5](#_Toc511594936)

[**1.** **Logistic Regression** 5](#_Toc511594937)

[**2.** **DECISION TREE** 7](#_Toc511594938)

[**3.** **Random forest** 8](#_Toc511594939)

[Συμπεράσματα 11](#_Toc511594940)

[Ομαδοποίηση πελατών 12](#_Toc511594941)

[Αποτελέσματα ομαδοποίησης και συμπεράσματα 14](#_Toc511594942)

# Εισαγωγή

Οι εταιρίες τηλεπικοινωνιών αντιμετωπίζουν παγκοσμίως μία μεγάλη πρόκληση, την απώλεια ενός όγκου πελατών τους. Το φαινόμενο αυτό είναι πιο έντονο στις εταιρίες της κινητής τηλεφωνίας όπου η αλλαγή παρόχου είναι πιο εύκολη καθώς πλέον ο καταναλωτής μπορεί να διατηρήσει τον αριθμό του αλλάζοντας εταιρία. Ως αποτέλεσμα αυτού και καθώς ο κλάδος των τηλεπικοινωνιών είναι ιδιαίτερα ανταγωνιστικός στις μέρες μας, οι καταναλωτές έχουν όλο και περισσότερες επιλογές - ευκαιρίες , τόσο από οικονομική όσο και από ποιοτική πλευρά. Προφανώς το φαινόμενο αυτό , είναι ιδιαίτερα ανησυχητικό για τις εταιρίες καθώς συνδέεται άμεσα με την πορεία των κερδών τους. Για το λόγο αυτό όλο και περισσότερες προσπαθούν να αναλύσουν την συμπεριφοράς των συνδρομητών τους και πιο συγκεκριμένα στην ανάλυση της αποχώρησης τους, προκειμένου να βελτιώσουν αλλά και να ενισχύσουν τη σχέση τους με αυτούς. Στην προηγούμενη ανάλυση ερευνήσαμε ακριβώς αυτό , δηλαδή ποιοι είναι οι πιο σημαντικοί παράγοντες που ωθούν έναν καταναλωτή να αλλάξει πάροχο (churn). Τώρα σκοπός της ανάλυσής θα είναι η δημιουργία ενός μοντέλου που θα μπορεί να προβλέψει με όσο το δυνατόν περισσότερη ακρίβεια την πιθανότητα αποχώρησης των καταναλωτών από την εταιρία. Επίσης θα προχωρήσουμε και στην ομαδοποίηση των πελατών σε σχέση με την καταναλωτική τους συμπεριφορά, προκειμένου να εξεταστούν πιθανές στρατηγικές-καμπάνιες από το τμήμα μάρκετινγκ.

Προκειμένου να μελετηθεί το φαινόμενο της αποχώρησης καταναλωτών της εταιρία , θα αναλυθούν τα ίδια ιστορικά δεδομένα ενός μέρους από το χαρτοφυλάκιο των πελατών μας. Τα δεδομένα αυτά αφορούν την προηγούμενη χρονική περίοδο. Πιο συγκεκριμένα θα γίνει ανάλυση για 3.333 διαφορετικούς πελάτες και τα στοιχεία που θα μπουν κάτω από το μικροσκόπιο αυτής της ανάλυσης αφορούν κυρίως την συμπεριφορά των συγκεκριμένων συνδρομητών όσο αφορά τον χρόνο, το πλήθος αλλά και τις αντίστοιχες χρεώσεις για το σύνολο των κλήσεων που πραγματοποίησαν. Είναι διαθέσιμα επίσης και μερικά δημογραφικά στοιχεία όπως το φύλλο και η πολιτεία από την οποία προέρχονται.

# Μεθοδολογία

Σκοπός αυτής της ανάλυσης είναι, όπως αναφέρθηκε και παραπάνω, είναι η πρόβλεψη για το εάν ένας πελάτης της εταιρίας θα αλλάξει πάροχο η όχι (churn). Η απάντηση σε ένα τέτοιο ερώτημα μπορεί να πάρει προφανώς μόνο δύο μόνο τιμές “NAI” ή “OXI”. Ένας καταναλωτής δηλαδή είτε θα παραμείνει είτε θα επιλέξει να αποχωρήσει. Η φύση αυτού του προβλήματος μας ωθεί στο να χρησιμοποιήσουμε μεθόδους-αλγόριθμους που μπορούν να προβλέψουν κατηγορηματικές μεταβλητές. Υπάρχουν αρκετοί και διαφορετικοί τρόποι προσέγγισης του προβλήματος αλλά χρησιμοποιηθούν τρείς εξ αυτών.

Πιο συγκεκριμένα οι μέθοδοι που θα χρησιμοποιηθούν παρακάτω είναι :

1. Logistic Regression
2. Decision Trees
3. Random Forest

Λεπτομέρειες για το πως ακριβώς λειτουργεί η κάθε μέθοδος αλλά και σχολιασμός των ευρημάτων θα ακολουθήσει παρακάτω. Για κάθε μία από αυτές και προκειμένου να μπορεί να ελεγχθεί η αποτελεσματικότητα της κάθε μεθόδου , θα χωριστούν τα δεδομένα σε δύο υποσύνολα. Στο πρώτο (training) , το οποίο και θα αποτελείται από το 80% των παρατηρήσεων , θα εφαρμοστεί κάθε μία μέθοδος ξεχωριστά και στο υπόλοιπο 20% (testing) θα ελεγχθεί η ακρίβεια-αποτελεσματικότητα της κάθε μεθόδου. Με τον τρόπο αυτό τα δύο υποσύνολο είναι ανεξάρτητα και έτσι μπορούμε να πούμε ότι αποφεύγεται το φαινόμενο της «υπερφόρτωσης» (overfitting). Η ανάλυση έχει γίνει με την βοήθεια της γλώσσας προγραμματισμού R και όλος ο κώδικάς υπάρχει σε ξεχωριστό αρχείο για περεταίρω μελέτη.

# Ανάλυση δεδομένων

Σε αυτό το σημείο καλό θα είναι να θυμηθούμε επιγραμματικά κάποια σημαντικά ευρήματα από την προηγούμενη ανάλυση καθώς θα χρησιμοποιηθούν τα ίδια δεδομένα.

Ο μέσος συνδρομητής λοιπόν είναι πελάτης της εταιρίας για περίπου 101 μήνες (8,4 χρόνια). Κατά την διάρκεια της τελευταίας χρονικής περιόδου, πραγματοποίησε περίπου ίδιο αριθμό κλήσεων (περίπου 100) στις διαφορετικές «ζώνες» της ημέρας (ημέρα , απόγευμα ,νύχτα). Την ημέρα μίλησε κατά μέσο όρο τον λιγότερο χρόνο αθροιστικά (180 λεπτά) σε σύγκριση με το απόγευμα και την νύχτα όπου χρεώθηκε για περίπου τον ίδιο χρόνο δηλαδή 201 λεπτά. Αντιθέτως , λόγω των διαφορετικών χρεώσεων που ισχύουν για τις διαφορετικές χρονικές περιόδους η μέση χρέωση για τις πραγματοποιθείσες κλήσεις ήταν 31€ για την ημέρα , 17€ για το απόγευμα και 9€ για το βράδυ. Συνολικά δηλαδή ένας μέσος πελάτης πραγματοποίησε 300 κλήσεις στις οποίες μίλησε για 983 λεπτά και χρεώθηκε συνολικά περίπου 57€. Όσο αφορά τις διεθνής κλήσεις κάθε καταναλωτής πραγματοποίησε περίπου 4 κλήσεις ενώ μίλησε για περίπου 10,2 λεπτά γεγονός που χρέωσε τον λογαριασμό του κατά 2,7€. Ένα σημείο που πρέπει να τονιστεί είναι ότι κάθε συνδρομητής κάλεσε κατά μέσο όρο 2 φορές το τμήμα εξυπηρέτησης πελατών. Επιπλέον μόνο το 10% (323 στους 3.333) των συνδρομητών διαθέτουν το πρόγραμμα για διεθνείς κλήσεις. Αντιθέτως το 28% (922 στους 3.333) έχουν ενεργοποιημένη την υπηρεσία για φωνητικά μηνύματα με τον κάθε συνδρομητή να πραγματοποιεί 8 φωνητικά μηνύματα κατά μέσο όρο. Τέλος το ποσοστό ανδρών και γυναικών είναι ίδιο και οι πελάτες προέρχονται από 51 διαφορετικές πολιτείες των Ηνωμένων Πολιτειών της Αμερικής.

Η πολιτική της εταιρίας είναι να έχει σταθερές χρεώσεις ανά χρονική ζώνη και διαμορφώνονται ως εξής :

• Ημέρα : 0,17€ / λεπτό

• Απόγευμα: 0,09€ / λεπτό

• Βράδυ : 0,04€ / λεπτό

• Διεθνής κλήσεις :0,27€ / λεπτό

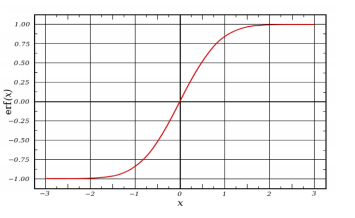
Επομένως υπάρχει απόλυτη συσχέτιση μεταξύ των μεταβλητών λεπτά και χρέωση για τις αντίστοιχες χρονικές ζώνες της ημέρας.

Τέλος από τους 3,333 πελάτες που εξετάζονται, 483 (14,5%) εξ αυτών αποφάσισαν να αλλάξουν πάροχο και οι κλήσεις προς το κέντρο εξυπηρέτησης πελατών , τα συνολικά λεπτά χρέωσης κατά την διάρκεια της ημέρας αλλά και προς διανθής κλήσεις να αποτελούν τους πιο σημαντικοί παράγοντές σε αυτήν τους την απόφαση.

# Αποτελέσματα ανάλυσης

## Logistic Regression

Η μέθοδος αυτή είναι συνδέει γραμμικά την εξαρτημένη μεταβλητή (churn) με τις υπόλοιπες. Η λογιστική παλινδρόμηση είναι στην ουσία ένα μοντέλο ταξινόμησης των τιμών μίας μεταβλητής απόκλισης Υ με βάση την θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή έχει δυαδικό χαρακτήρα (παίρνει δύο τιμές ΝΑΙ και ΌΧΙ) στοχεύετε η πρόβλεψη της έκβασης αυτής από το πλήθος των υπόλοιπων μεταβλητών. Η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος (στην περίπτωσή μας το αν ένας συνδρομητής θα αποφασίσει να αποχωρήσει ή όχι από την εταιρία) προσαρμόζοντας τα δεδομένα της μελέτης όπως αυτή παρουσιάζεται στο παρακάτω διάγραμμα.



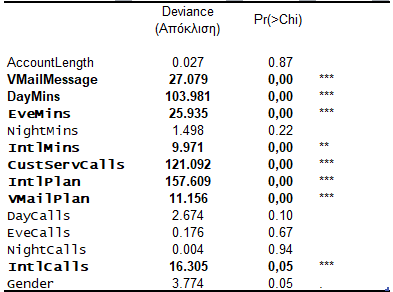
Διάγραμμα 1Τυπική ανάπτυξη σιγμοειδούς καμπύλης

Η έκφραση της Λογιστικής Παλινδρόμησης δίνεται από την εξίσωση



Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραµµικό συνδυασμό των ανεξάρτητων μεταβλητών που συµµμετέχουν στο µμοντέλο της παλινδρόμησης. Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής µε την μορφή του λογαρίθμου των odd δηλαδή, του λογαρίθμου της σχέσης: odds = prob/(1-prob) . Το odds εναλλακτικά ονομάζεται logit και ο όρος Prob εκφράζει την πιθανότητα να συμβεί το γεγονός που έχει ορισθεί σαν επιτυχία του πειράματος , το οποίο στην περίπτωσή μας είναι εάν ένας πελάτης θα αποχωρήσει από την εταιρία ή όχι.

Εφσρμόζοντας την μέθοδο αυτή και προχωρώντας στην ανάλυση της απόκλισης των μεταβλητών του μοντέλου παίρνουμε την παρακάτω εικόνα



Διάγραμμα 2: Ανάλυση της απόκλισης (Anova)

Αναλύοντας τον πίνακα βλέπουμε την μείωση της απόκλισης όταν προστίθεται κάθε μία μεταβλητή ξεχωριστά. Προσθέτοντας τις μεταβλητές «CustServCalls” , “DayMins” , “IntPlan” , μειώνεται σημαντικά η απόκλιση του μοντέλου. Η προσθήκη των υπόλοιπων μεταβλητών ( «VMailMessage” , “EveMins” , IntlMins” , “VMailPlan” , “IntCalls”) βελτιώνουν μεν το μοντέλο αλλά λιγότερο συγκριτικά με τις πρώτες.

H ικανότητα πρόβλεψης του παραπάνω μοντέλου υπολογίζεται στο 86% όπως φαίνεται και στον παρακάτω πίνακα. Το μοντέλο προέβλεψε σωστά στις 573 από τις 666 περιπτώσεις συνολικά. Με κόκκινο χρώμα φαίνονται τα λάθη που έκανε το μοντέλο. Επίσης φαίνεται ότι το μοντέλο μπορεί να προβλέψει καλύτερα τους καταναλωτές που δεν θα αποχωρήσουν (87%) σε σύγκριση με αυτούς που θα αποχωρήσουν (54,5%). Σε γενικές γραμμές το μοντέλο αποδίδει αρκετά καλά δεδομένου ότι το τυχαίο ποσοστό ακρίβειας υπολογίζεται στο 72,21% (Chance Accuracy Rate = (proportion of defaults)^2 + (proportion of non-defaults)^2). Η ζητούμενη πρόβλεψη όμως , δηλαδή αν ένας καταναλωτής θα αποχωρήσει όχι , έχει σχετικά μικρό ποσοστό επιτυχίας και για τον λόγο αυτό θα δοκιμάσουμε και τις επόμενες μεθόδους.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | NO | YES |
| Predicted | NO | **555** | *15* |
| YES | *78* | **18** |

Πίνακας 1 : Confusion Matrix for Logistic Regression

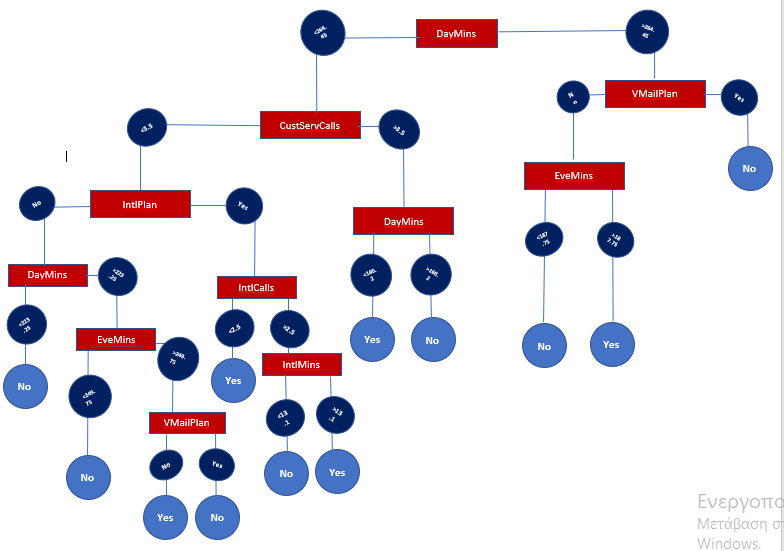
## DECISION TREE

Τα δέντρα απόφασης χρησιμοποιούνται ευρέως για την κατηγοριοποίηση και πρόβλεψη δεδομένων. Ένα δέντρο απόφασης κατασκευάζεται σύμφωνα με ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων. Κάθε εσωτερικός κόμβος προσδιορίζει τον έλεγχο των γνωρισμάτων και κάθε κλαδί που συνδέει τους εσωτερικούς με τους απόγονους αντιστοιχεί σε μία πιθανή τιμή για το γνώρισμα. Στο συγκεκριμένο παράδειγμα ο αλγόριθμος θα προσπαθήσει να προβλέψει εάν ένας πελάτης θα αποχωρήσει από την εταιρία ή όχι λαμβάνοντας υπόψιν τις τιμές που παίρνουν οι διάφορες μεταβλητές.

Χρησιμοποιώντας την συνάρτηση tree από το ομώνυμο πακέτο στην R παίρνουμε τα παρακάτω αποτελέσματα :

1. Ο αλγόριθμος χρησιμοποίησε 7 μεταβλητές για να φτιάξει το δέντρο απόφασης και αυτές είναι : “DayMins” , “CustServCalls” , “IntPlan” , “EveMins” , VmailPlan” , “IntCalls” και «IntMins”. Βλέπουμε δηλαδή ότι υπάρχει σχεδόν απόλυτη συμφωνία με τις μεταβλητές που η προηγούμενη μέθοδος θεώρησε σημαντικές για την πρόβλεψη.
2. Το δέντρο αποτελείται από 12 βρόγχους
3. Το ποσοστό εσφαλμένης ταξινόμησης είναι 5,7% , το οποίο είναι χαμηλό αλλά θα θέλαμε να το δοκιμάσουμε και στο testing dataset

To δέντρο που έχτισε ο αλγόριθμος φαίνεται στην παρακάτω εικόνα



Διάγραμμα 3: Δέντρο Απόφασης (decision tree)

Το δέντρο ερμηνεύεται ως εξής :

* Εάν ένας συνδρομητής μιλάει περισσότερα από 265 λεπτά κατά την διάρκεια της ημέρας , δεν έχει ενεργοποιημένη την υπηρεσία της φωνητικής κλήσης και μιλάει περισσότερα από 188 λεπτά κατά την διάρκεια του απογέματος είναι πολύ πιθανόν να αποχωρήσει από την εταιρία
* Το ίδιο πιθανόν είναι να αποχωρήσει ένας καταναλωτής ο οποίος μιλάει λιγότερο από 265 λεπτά, έχει καλέσει τουλάχιστον 4 φορές στο κέντρο εξυπηρέτησης πελατών αλλά έχοντας μιλήσει λιγότερο από 160 λεπτά κατά την διάρκεια της ημέρας.

Με αντίστοιχο τρόπο ερμηνεύονται και οι υπόλοιπες «διαδρομές» του δέντρου.

Το μοντέλο μπορεί να προβλέψει επιτυχώς το 93,8% των περιπτώσεων όπως φαίνεται στον παρακάτω πίνακα. Επίσης το ποσοστό σωστής ταξινόμησης για τους καταναλωτές που τελικά αποχώρησαν από την εταιρία είναι 64,5% και το αντίστοιχο για αυτούς που έμειναν στην εταιρία είναι 98%. Φαίνεται δηλαδή ότι αυτή η μέθοδος μπορεί να προβλέψει πιο αποτελεσματικά την αποχώρηση των καταναλωτών σε σύγκριση με την προηγούμενη

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | NO | YES |
| Predicted | NO | **563** | *34* |
| YES | *7* | **62** |

Πίνακας 2 : Confusion Matrix for Tree Model

Χτίζοντας ένα δεύτερο δέντρο με 11 βρόγχους αυτή τη φορά αντί για 12 αυξάνει το ποσοστό σωστής ταξινόμησης των καταναλωτών που είναι πιθανόν να αποχωρήσουν από την εταιρία κατά 2% περίπου (66,6%) που είναι άλλωστε και το ζητούμενο αυτής της ανάλυσης , με ταυτόχρονη μείωση όμως της συνολικής αποτελεσματικότητας πρόβλεψης του μοντέλου κατά 0,4% (93,4%).

## Random forest

Η μέθοδος αυτή χτίζει αντί για ένα πολλά δέντρα και στη συνέχεια συγχωνεύει τα αποτελέσματα. Για το χτίσιμο κάθε δέντρου χρησιμοποιείται η μέθοδος της δειγματοληψίας με αντικατάσταση (sample with replacement) . Κάθε φορά χρησιμοποιούνται περίπου τα 2/3 του δείγματος με τυχαία επιλογή. Το κύριο πλεονεκτήματα αυτής της μεθόδου σε σχέση με το δέντρο απόφασης είναι ότι μειώνεται ο κίνδυνος υπερφόρτωσης» (overfitting) καθώς με την συγχώνευση των αποτελεσμάτων κάθε δέντρου, μειώνεται η απόκλιση του μοντέλου. Επίσης η δυνατότητα πρόβλεψης γίνεται πιο σταθερή και ακριβής.

Χρησιμοποιώντας την συνάρτηση randomForest του ομώνυμου πακέτου στην R και δοκιμάζοντας την ικανότητά της στην επιτυχή ταξινόμηση των δεδομένων στο testing dataset παίρνουμε τον παρακάτω πίνακα.

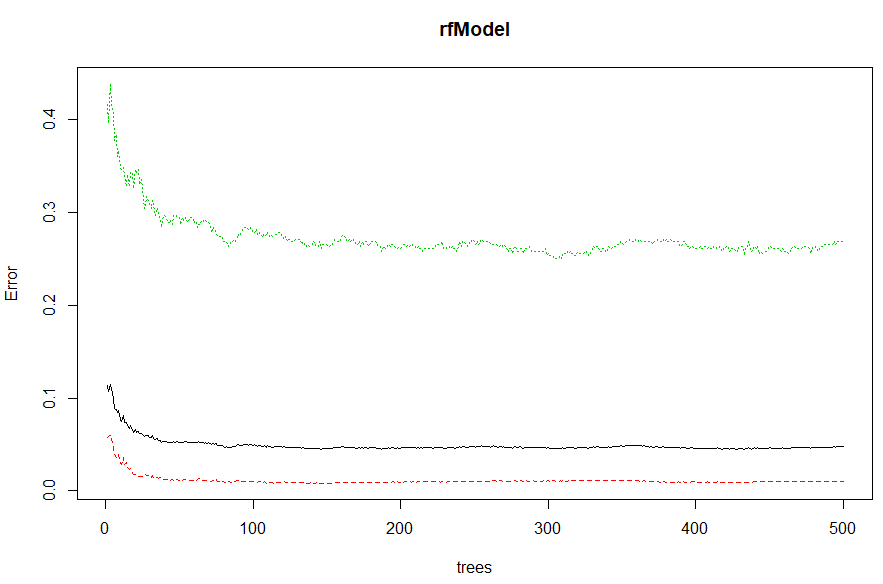
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | NO | YES |
| Predicted | NO | **565** | *28* |
| YES | *5* | **68** |

Πίνακας 3: Confusion Matrix for Random Forest

Η ακρίβεια του μοντέλου εκτιμάται στο 95% καθώς ταξινόμησε σωστά 633 περιπτώσεις από τις 666 συνολικά. Τα αντίστοιχα ποσοστά για τους πελάτες που δεν αποχώρησαν από την εταιρία είναι στο 99% ενώ τα αντίστοιχα για αυτούς που αποχώρησαν είναι στο 70,8%. Φαίνεται ότι αυτή η μέθοδος μπορεί να ταξινομεί με μεγαλύτερη επιτυχία τα δεδομένα από όλες τις προηγούμενες.

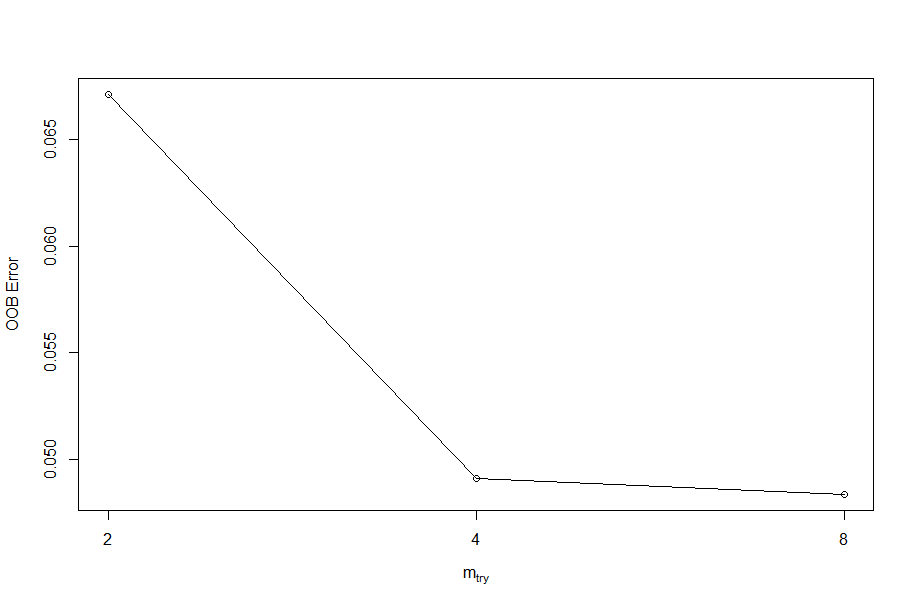
Μελετώντας περισσότερο το μοντέλο προκειμένου να βελτιωθεί η ακρίβεια του παίρνουμε τα εξής χρήσιμα συμπεράσματα.

Το παρακάτω διάγραμμα δείχνει πως μετά από περίπου 200-300 δέντρα το OOB Error σταματάει να μειώνεται και παραμένει σταθερό.



Διάγραμμα 4

Επίσης στο παρακάτω διάγραμμα φαίνεται ότι το OOB Error παίρνει την μικρότερη τιμή όταν το mtry=8. Αυτό σημαίνει το πόσες μεταβλητές χρησιμοποιούνται σε κάθε βρόγχο του δέντρου.



Διάγραμμα 5

Οπότε θα ξανατρέξουμε τον αλγόριθμο χρησιμοποιώντας αυτά τα συμπεράσματα και όπως φαίνεται και στον παρακάτω πίνακα βελτιώνεται η αποτελεσματικότητά του στην πρόβλεψη της φυγής των πελατών.

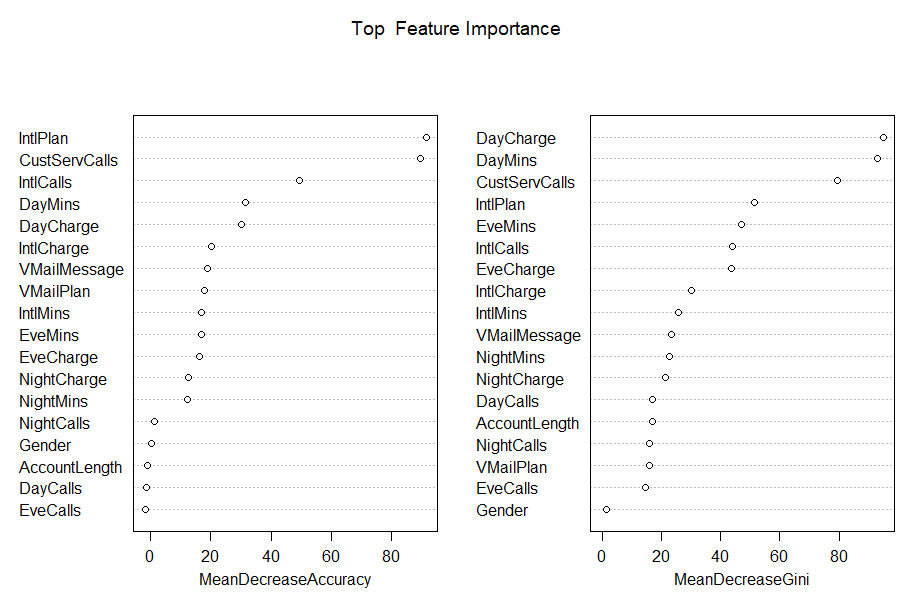
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | NO | YES |
| Predicted | NO | **565** | *23* |
| YES | *5* | **73** |

Πίνακας 4 : Confusion Matrix for Random Forest after Tuning

Τέλος θα θέλαμε να εξετάσουμε ποιες μεταβλητές είναι σημαντικές για το μοντέλο.

Στο αριστερό διάγραμμα φαίνεται ποιες μεταβλητές επηρεάζουν και σε τι βαθμό την ακρίβεια του μοντέλου. Όπως και στα προηγούμενα μοντέλα οι μεταβλητές «IntPlan” , “CustServCalls” , “IntCalls” , “DayMins” (και ως συνέπεια DayCharge) είναι οι πιο σημαντικές και η αφαίρεση αυτών θα επηρέαζε αρνητικά την ακρίβεια του μοντέλου.

Στο δεξιά διάγραμμα φαίνεται η κατά μέσο όρο μείωση του δείκτη gini εάν αφαιρεθεί κάποια μεταβλητή από το μοντέλο. Αυτός ο δείκτης επηρεάζει το πόσο «καθαρός» είναι ο κάθε διαχωρισμός στα δέντρα. Όσο μεγαλύτερη είναι η μείωση του δείκτη τόσο πιο «καθαρός» είναι ο διαχωρισμός στους βρόγχους το οποίο είναι και το ζητούμενο σε ένα μοντέλο που θέλουμε να προβλέπει με επιτυχία.



Διάγραμμα 6

# Συμπεράσματα

Από όλες τις μεθόδους που χρησιμοποιήθηκαν , η random forest είχε τα καλύτερα αποτελέσματα με ποσοστό επιτυχημένης ταξινόμησης 95,8% στο σύνολο των παρατηρήσεων. Επίσης είχε και το μεγαλύτερο ποσοστό όσο αφορά στην πρόβλεψη των πελατών που τελικά αποφάσισαν να αποχωρήσουν από την εταιρία (76%). Τέλος όλες οι μέθοδοι ανέδειξαν περίπου τις ίδιες μεταβλητές ως σημαντικές για την σωστή πρόβλεψη

# Ομαδοποίηση πελατών

Στην συνέχεια της ανάλυσης θα προσπαθήσουμε να ομαδοποιήσουμε τους πελάτες σύμφωνα με την καταναλωτική του συμπεριφορά προκειμένου να εξεταστούν διάφορες στρατηγικές-καμπάνιες από το τμήμα μάρκετινγκ με απώτερο σκοπό να αποτραπεί πιθανή φυγή πελατών.

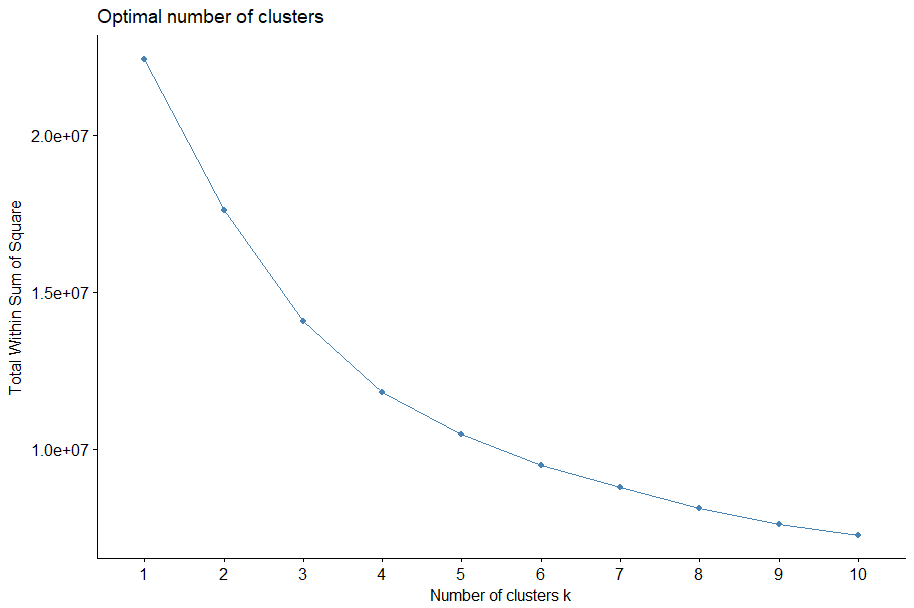
Αρχικά θα πρέπει να αφαιρεθούν από τα δεδομένα όσοι πελάτες τελικά αποχώρησαν από την εταιρία καθώς δεν αποτελούν πλέον συνδρομητές της εταιρίας και δεν θα ήταν σωστό η καταναλωτική τους συμπεριφορά να επηρεάσει τις στρατηγικές του τμήματος μάρκετινγκ. Επίσης, όπως έχει ήδη αναφερθεί οι μεταβλητές που αφορούν τα λεπτά χρέωσης για κάθε «χρονική ζώνη» και οι αντίστοιχες χρεώσεις είναι απόλυτα συσχετισμένες. Οπότε θα εξευρεθούν καθώς έχουν την ίδια συμπεριφορά.

Για την ομαδοποίησης των πελατών θα χρησιμοποιηθεί ο αλγόριθμος k-mean. Αυτός ο αλγόριθμος ομαδοποιεί τα δεδομένα σε k ομάδες ανάλογα με τις ομοιότητες που παρουσιάζουν τα δεδομένα της κάθε παρατήρησης (στην δική μας περίπτωση πελάτες). Ο αλγόριθμος χρησιμοποιεί την έννοια του κέντρου της ομάδας (centroid) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα µε την απόσταση τους από τα κέντρα όλων των ομάδων. Το κέντρο της ομάδας είναι η μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Αποτελείται από τέσσερα βήματα

* **Βήμα 1ο** Βρες τα αρχικά κέντρα
* **Βήµα 2ο** Κατάταξε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση
* **Βήµα 3ο** Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολόγισε τα νέα κέντρα.
* **Βήµα 4ο** Αν τα νέα κέντρα δεν διαφέρουν από τα παλιά σταμάτα αλλιώς πήγαινε στο βήμα 2.

Αρχικά θα θέλαμε να δούμε τον ιδανικό αριθμό των ομάδων (clusters) καθώς αυτός ο αλγόριθμος απαιτεί τον ορισμό τους από την αρχή.

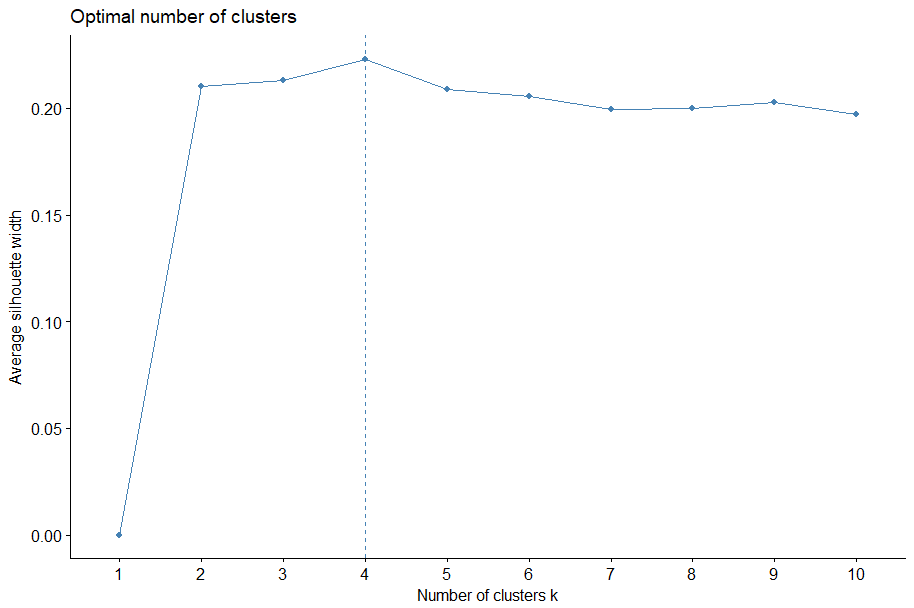
Τα παρακάτω διαγράμματα θα μας βοηθήσουν να αποφασίσουμε για τον αριθμό των ομάδων



Διάγραμμα 7

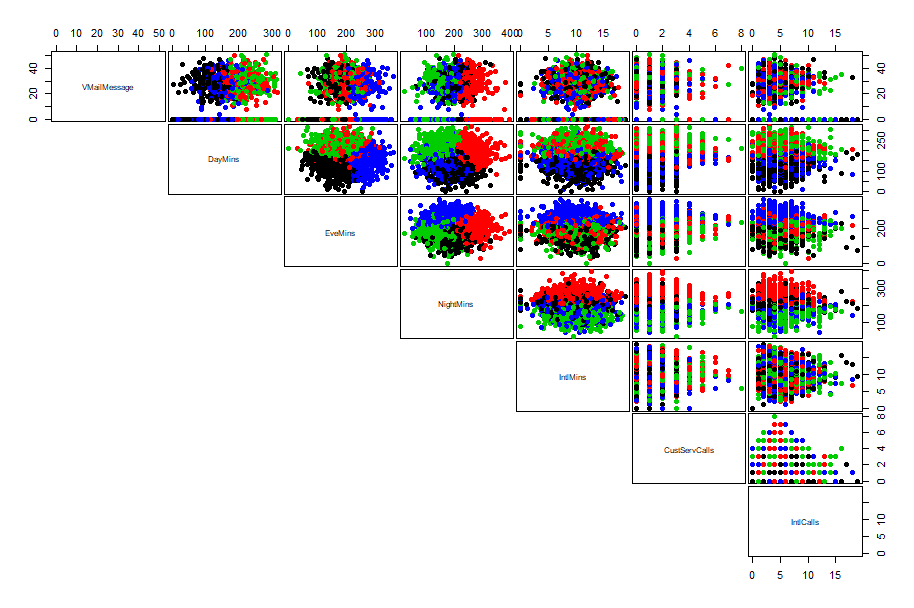
Από το παραπάνω διάγραμμα φαίνεται ότι 4 ομάδες θα ήταν ο ιδανικός αριθμός καθώς σε αυτό το σημείο φαίνεται να αρχίζει να αλλάζει απότομα το άθροισμα των τετραγώνων των αποστάσεων μεταξύ των παρατηρήσεων μέσα στις ομάδες (within-cluster sum of square)

Επίσης όπως φαίνεται και στο παρακάτω διάγραμμα οι τιμές silhouette παίρνουν την μέγιστη τιμή όταν οι ομάδες είναι 4. Οι τιμές silhouette το πόσο καλά έχει γίνει η ομαδοποίηση. Μεγάλες τιμές φανερώνουν μία καλή ομαδοποίηση.



Διάγραμμα 8

Η απεικόνιση των 4 ομάδων (clusters) φαίνεται στον παρακάτω πίνακα. Με ξεχωριστό χρώμα απεικονίζονται οι διαφορετικές ομάδες ενώ στον άξονα χ και y είναι κάθε πιθανός συνδυασμός των μεταβλητών.



Διάγραμμα 9 : Απεικόνιση των ομάδων σε συνάρτηση με τις μεταβλητές

# Αποτελέσματα ομαδοποίησης και συμπεράσματα

Από το παραπάνω διάγραμμα μπορούμε να συμπεράνουμε τα εξής για τις διάφορες ομάδες που δημιουργήθηκαν :

* **Ομάδα 1** ( πράσινο χρώμα) : αποτελείται από καταναλωτές που μιλούν πολύ κατά την διάρκεια της ημέρας αλλά λίγο το βράδυ. Όσο αφορά τα λεπτά ομιλίας το απόγευμα μπορούμε να πούμε ότι δεν παίρνουν μεγάλες τιμές σε σχέση με τους υπόλοιπους πελάτες.
* **Ομάδα 2** (κόκκινο χρώμα) : περιέχει τους συνδρομητές όπου σαν κύριο χαρακτηριστικό μιλάνε πολύ κατά την διάρκεια των νυχτερινών ωρών. Επίσης φαίνεται πως τα λεπτά ομιλίας τους στην διάρκεια της ημέρας είναι κατά κανόνα αρκετά.
* **Ομάδα 3** (μπλε χρώμα) : αποτελείται από καταναλωτές όπου μιλάνε πολύ κατά την διάρκεια του απογεύματος. Τα λεπτά ομιλίας τους στις πρωινές ώρες μπορούμε να πούμε ότι είναι στο μέσο όρο ή και λιγότερα.
* **Ομάδα 4** (μαύρο χρώμα) : απαρτίζεται από τους συνδρομητές όπου μιλούν λίγο κατά την διάρκεια της ημέρας και του απογεύματος , ενώ μπορούμε να πούμε ότι τα λεπτά ομιλίας τους το απόγευμα είναι στον μέσο όρο.

Ως γενικό συμπέρασμα θα πρέπει να σημειωθεί ότι όσο αφορά τις κλήσεις προς το κέντρο εξυπηρέτησης πελατών αλλά και τις διεθνής κλήσεις δεν παρατηρείται κάποια ομαδοποίηση και όλες οι ομάδες φαίνεται να παίρνουν παρόμοιες τιμές. Το παραπάνω συμπέρασμα θα μπορούσαμε να πούμε ότι είναι αναμενόμενο καθώς έχουν διαφορετική μονάδα μέτρησης με τις υπόλοιπες μεταβλητές (αριθμός κλήσεων και λεπτά ομιλίας). Επίσης αυτές οι μεταβλητές παίρνουν χαμηλές τιμές σε σχέση με τις υπόλοιπες. Αν θέλαμε να έχουν πιο πρωταγωνιστικό ρόλο στην ομαδοποίηση θα μπορούσαμε να σταθμίσουμε τις εν λόγω τιμές ώστε να έχουν μεγαλύτερη βαρύτητα. Στον αντίποδα καθώς και οι δύο μεταβλητές έχουν χαρακτηριστεί ως σημαντικές για την πρόβλεψη της φυγής των καταναλωτών ίσως θα έπρεπε να αντιμετωπιστούν με έναν γενικό κανόνα-στρατηγική για όλες τις ομάδες. Τέλος οι κλήσεις προς το κέντρο εξυπηρέτησης πελατών είναι κάτι το οποί μπορεί να απομονωθεί και να παρακολουθηθεί ξεχωριστά για κάθε πελάτη.

Τέλος ως συνάρτηση με την πρώτη ανάλυση ίσως θα έπρεπε να δοθεί ιδιαίτερη βαρύτητα στις ομάδες οι οποίες χαρακτηρίζονται από πολλά λεπτά ομιλίας κατά την διάρκεια της ημέρας (ομάδα 1 και 2). Όλα τα μοντέλα έχουν αναδείξει αυτόν τον παράγοντα ως σημαντικό για το ενδεχόμενο αποχώρησης από την εταιρία. Ως ιδανικό σενάριο θα μπορούσε για αυτές τις ομάδες να ακολουθηθεί μία διαφορετική πολιτική χρέωσης.