ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνολογίας Λογισμικού

Μελέτη και Κατασκευή Μοντέλων για Πρόβλεψη του Ρυθμού Εγκατάλειψης σε Πρόγραμμα Αποταμιεύσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Γεωργακόπουλος

Επιβλέπων: Παπασπύρου Νικόλαος
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνολογίας Λογισμικού

Μελέτη και Κατασκευή Μοντέλων για Πρόβλεψη του Ρυθμού Εγκατάλειψης σε Πρόγραμμα Αποταμιεύσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Γεωργακόπουλος

Επιβλέπων: Παπασπύρου Νικόλαος
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Ιουλίου 2017.

…….	…….	…….
Ν. Παπασπύρου	Κ. Πελεχρίνης	Δ. Ασκούνης
Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής U.o.Pitt.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

…….

Παναγιώτης Γεωργακόπουλος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών

Περίληψη

Στη σύγχρονη τεχνολογική κοινωνία γεμάτη με σύνθετα συστήματα όπου αλληλοεπιδρούν παίκτες-άτομα με περίπλοκη συμπεριφορά, μία από τις πλέον επιτακτικές ανάγκες αποτελεί η πρόβλεψη των μελλοντικών κινήσεων των συμμετεχόντων. Στη παρούσα διπλωματική εργασία, μελετάμε τη συμπεριφορά αυτών των παικτών στα πλαίσια ενός τραπεζικού προγράμματος και προσπαθούμε να κατασκευάσουμε ένα σύστημα που με ικανοποιητική επιτυχία θα προβλέπει τις μελλοντικές τους κινήσεις. Συγκεκριμένα, αξιοποιώντας οικονομικά και δημογραφικά δεδομένα ενός μεγάλου πλήθους πελατών για ένα εύλογο χρονικό διάστημα, προσπαθούμε να κατασκευάσουμε μοντέλα που να προβλέπουν ικανοποιητικά το ρυθμό οικειοθελούς αποχώρησης των πελατών από το τραπεζικό πρόγραμμα (churn rate). Κάθε μοντέλο αξιολογείται βάσει κλασικών (στη σχετική βιβλιογραφία) μετρικών πάνω στην ακρίβεια και πληρότητα του συστήματος. Αντλώντας πληροφορίες από ορισμένες πρωταρχικές προσεγγίσεις μέσω μαρκοβιανών αλυσίδων και αναλύσεις χρονοσειρών, καταλήγουμε σε πιο εξειδικευμένα μοντέλα που βασίζονται σε ανάλυση παλινδρόμησης και δέντρα απόφασης, ενώ δοκιμάζουμε και τη δυνατότητα κατηγοριοποίησης της συμπεριφοράς των χρηστών μέσω μεθόδων Clustering. Εν τέλει, με χρήση των συμπερασμάτων που έχουμε εξάγει, καταλήγουμε και παρουσιάζουμε ένα υβριδικό μοντέλο που καταφέρνει να παρουσιάζει τόσο από πλευράς ποιότητας όσο και ευστάθειας υψηλές επιδόσεις, ενώ παράλληλα προτείνουμε μια πληθώρα καινοτόμων δομών που δύνανται να αυξήσουν ακόμη περισσότερο τη συνολική ποιότητα προβλέψεων του συστήματος.

Λέξεις-Κλειδιά: Ανάλυση Επιβίωσης, Ανάλυση Παλινδρόμησης, Churn Rate, Δέντρα Απόφασης, Γενική Λογιστική Παλινδρόμηση, Ομαδοποίηση, Συστήματα Προβλέψεων, Χρονοσειρές

Abstract

In today's technology society full of complex systems where players-people with complex behavior interact, one of the most pressing needs is to predict the future movements of the participants. In this diploma thesis, we study the behavior of these players in the context of a banking program and we are trying to build a system that will successfully anticipate their future moves. In particular, exploiting the economic and demographic data of a large number of customers over a reasonable period of time, we are trying to build models that adequately anticipate the willingness of customers to withdraw from the banking program (churn rate). Each model is evaluated on the basis of classical (amongst related bibliography) metrics on the accuracy and completeness of the system. Drawing information from some primitive approaches through Markov chains and time series, we come up with more specialized models based on regression analysis and decision trees, and we also test the possibility to categorize user behavior through Clustering methods. Ultimately, using the conclusions we have drawn, we end up and present a hybrid model that manages to deliver both high quality and stability in terms of performance, while also proposing a plethora of innovative structures that can further increase the overall quality of the system's predictions.

Keywords: Survival Analysis, Regression Analysis, Churn Rate, Decision Trees, General Logistic Regression, Clustering, Forecast Systems, Time series

Αφιερώνεται στη μητέρα μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον κ. Παπασπύρου, επιβλέποντα της παρούσας διπλωματικής, τόσο για την υποστήριξη του κατά την εκπόνηση της εργασίας αυτής, όσο και για τη συνεισφορά του στην ακαδημαϊκή και επαγγελματική μου σταδιοδρομία, για την μετάδοση της αγάπης για την εκπαίδευση γενικά αλλά και για την επιστήμη των υπολογιστών και των γλωσσών προγραμματισμού ειδικότερα. Τον κ. Πελεχρίνη, που μέσα και έξω από τα πλαίσια της παρούσας διπλωματικής με βοήθησε να εξερευνήσω το πεδίο του data science όπως αυτό διαμορφώνεται στην εποχή των μεγάλων δεδομένων, με επιμονή και υπομονή. Τον κ. Ασκούνη για την εισαγωγή της καινοτομίας στη σχολή μέσω δράσεων όπως το ennovation και της επιχειρηματικής διάστασης στο ιδιαίτερα τεχνικό και επιστημονικό πρόγραμμα σπουδών. Ιδιαίτερη αναφορά οφείλω να συμπεριλάβω, στον κ Ζάχο και τον κ Φωτάκη που μου γνώρισαν τον κόσμο των υπολογισμών και των αλγορίθμων.

Επί τη ευκαιρία, θα ήθελα να ευχαριστήσω τον κ. Προδρομίδη για την έμπνευση του θέματος και την πολύτιμη συνεργασία και καθοδήγησή του. Τους κ. Ματθαίο, κ. Χουλιαρά και κα. Μπουλουγούρα για την υποστήριξη τους ενάντια στην καθημερινότητα αλλά και την εμπιστοσύνη και την καθοδήγηση που απλόχερα μου παρέχουν. Τον συμφοιτητή και φίλο Αντώνη Μητρόπουλο για την φιλία, τη συμπαράσταση και την υποστήριξη του από τα φοιτητικά μας χρόνια ως τώρα, καθώς και τους Άρη, Κώστα, Αδάμ, Δημήτρη, Θέμη, Στράτο για τους ίδιους λόγους. Τον αδερφό μου Γιάννη, επίσης φοιτητή της ΣΗΜΜΥ, για τη συνεισφορά του στην διεκπεραίωση των τελευταίων μαθημάτων της σχολής, τον Πλάτωνα για την υποστήριξη σε άλυτα μαθηματικοφιλοσοφικά προβλήματα και τους υπόλοιπους ανώνυμους συνοδοιπόρους που χωρίς να το ξέρουν επηρέασαν την ακαδημαϊκή μου ζωή.

Ευχαριστώ την Αλεξάνδρα που είναι μαζί μου και με στηρίζει με υπομονή και αγάπη για το σύνολο των ακαδημαϊκών χρόνων, τον αγαπημένο μου φίλο Ρήγα που είναι τα τελευταία 21 χρόνια ο 5^{ος} μου αδελφός, τα 4 πρώτα, Ζωή, Γιάννη, Νίκο, Στέφανο, και φυσικά τους γονείς μου στους οποίους οφείλω όλα όσα είμαι.

Περιεχόμενα

Πίνακας Εικόνων

Κεφάλαιο 1
Εισαγωγή

1.1 Γενικό Πλαίσιο

Η ραγδαία εξέλιξη της τεχνολογίας μας έχει οδηγήσει σε ένα επίπεδο όπου είναι ευρέως εφικτή η μαζική συλλογή και αποθήκευση δεδομένων που αφορούν μια συγκεκριμένη ομάδα ατόμων σε σχέση με κάποιο συγκεκριμένο προϊόν, δραστηριότητα, κοινωνικό χαρακτηριστικό κ.ο.κ. Η ανάπτυξη κατάλληλων αλγοριθμικών και προγραμματιστικών εργαλείων έχει επίσης καταστήσει σε μεγάλο βαθμό εφικτή τη παράλληλη και αποδοτική επεξεργασία αυτών των δεδομένων για την εξαγωγή διαφόρων χρήσιμων συμπερασμάτων. Σε ένα διαρκώς όλο και πιο ανταγωνιστικό περιβάλλον, είτε εξαιτίας των τρεχόντων οικονομικών συνθηκών (π.χ. μεταξύ σελίδων κοινωνικής δικτύωσης που προσπαθούν να εγγυηθούν στους πελάτες τους στοχευμένη παροχή διαφημίσεων) είτε λόγω της ίδιας της τρέχουσας κοινωνικοπολιτικής κατάστασης (π.χ. άμεση εκτίμηση κινδύνου για μελλοντική οικονομική κρίση, τρομοκρατικό πλήγμα κ.ο.κ.), είναι προφανές ότι η ανάγκη εξεύρεσης αποδοτικών μεθόδων που θα εγγυούνται την όσο το δυνατόν ακριβέστερη πρόβλεψη των κινήσεων των εν λόγω ατόμων (βάσει, ασφαλώς, πάντοτε των συλλεχθέντων δεδομένων) είναι σήμερα πιο επιτακτική από ποτέ.

Στην παρούσα εργασία γίνεται μελέτη και αξιολόγηση των τρόπων με τους οποίους μπορεί να συμβεί η ανάπτυξη αλγορίθμων και λογισμικών ικανών να υπολογίζουν τα ανωτέρω και συγκεκριμένα ασχολούμαστε με την υποπερίπτωση της συμπεριφοράς των χρηστών σε σχέση με την διακοπή της συμμετοχής τους σε κάποια συγκεκριμένη υπηρεσία. Η γενική μετρική που σχετίζεται με το ρυθμό εγκατάλειψης μιας υπηρεσίας από τους πελάτες είναι γνωστή στη βιβλιογραφία ως **churn rate** (ή attrition rate) και αποτελεί έναν από τους βασικούς παράγοντες που καθορίζουν την ευστάθεια του συστήματος παροχής της αντίστοιχης υπηρεσίας (βλ. (Gorgoglione & Panniello, 2011) και (Hadden, Tiwari, Roy, & Ruta, 2006)).

Η μελέτη του churn rate αποτελεί ένα από τα κεντρικότερα ζητούμενα κατά την επιχειρησιακή ανάλυση και διοίκηση των σύγχρονων εταιριών. Βρίσκεται στο επίκεντρο των αποφάσεων σε πληθώρα τύπων επιχειρήσεων, όπως εταιρίες τηλεπικοινωνιών και παροχής υπηρεσιών διαδικτύου (π.χ. ηλεκτρονικό ταχυδρομείο, μέσα κοινωνικής δικτύωσης, cloud services κ.λπ.), συνδρομητικών προγραμμάτων ψυχαγωγίας (π.χ. μουσική, τηλεόραση κ.λπ.), τραπεζικών κι επενδυτικών προγραμμάτων και γενικότερα σε κάθε μορφή επιχείρησης που σχετίζεται με πελάτες οι οποίοι έχουν τη δυνατότητα της (συνήθως) οικειοθελούς αποχώρησης από προϊόντα της (ενώ σε πιο αφηρημένα πλαίσια μπορεί να εμπεριέχεται και σε διάφορα κοινωνιολογικά ή βιολογικά μοντέλα). Σημείο αναφοράς της παρούσας

εργασίας είναι η περίπτωση ενός τραπεζικού προγράμματος αποταμίευσης, το οποίο στηρίζεται σε υπαρκτά στοιχεία πραγματικής βάσης δεδομένων (λεπτομέρειες επί της οποίας ακολουθούν στο επόμενο κεφάλαιο).

Γενικά, υπάρχουν διάφορα πεδία στα οποία μπορεί να εντάσσεται η γενική μορφή του αλγορίθμου πρόβλεψης. Σε αρκετές περιπτώσεις, όπου δε δίνεται εξαιρετικά μεγάλη προσοχή στο churn rate, χρησιμοποιούνται απλές μέθοδοι, όπως π.χ. η κατηγοριοποίηση των πελατών ανάλογα με το μήκος της χρονικής περιόδου όπου έχουν μείνει ανενεργοί όσον αφορά τη χρήση του προϊόντος. Σε περιπτώσεις, ωστόσο, μεγαλύτερων εταιριών που η διατήρηση των υπαρχόντων πελατών εντός του προγράμματος έχει μεγάλη προτεραιότητα, χρησιμοποιούνται πιο ανεπτυγμένοι αλγόριθμοι, οι οποίοι μπορούν να χωριστούν σε γενικές γραμμές σε δύο γενικά μοτίβα. Το πρώτο εντάσσεται στη μηχανική μάθηση και περιλαμβάνει συχνά τη χρήση νευρωνικών δικτύων και γενετικών αλγορίθμων, οι οποίοι έχουν σκοπό να «μάθουν» τη συμπεριφορά του χρήστη κι έτσι να προβλέψουν τη μελλοντική του συμπεριφορά. Το δεύτερο, το οποίο είναι και αυτό στο οποίο εστιάζουμε περισσότερο στην παρούσα εργασία, είναι αυτό της λογιστικής παλινδρόμησης, όπου αξιοποιούνται δεδομένα των πελατών για την εξαγωγή μιας όσο το δυνατόν πιο ακριβούς συνάρτησης (η οποία μπορεί να είναι είτε σε αριθμητική μορφή, σε μορφή δέντρου ή γενικά κατηγοριοποίησης σε σύνολο), η τιμή της οποίας να ισούται με την απόφαση (ή γενικά με τη πιθανότητα αυτής) που πρόκειται να πάρει ο πελάτης.

1.2 Στάδια Συστημάτων Προβλέψεων

Στην παραγωγή κάθε τέτοιου συστήματος πρόβλεψης, όπως αυτό με το οποίο θα ασχοληθούμε, μπορούμε να ξεχωρίσουμε τρία βασικά στάδια: τη συλλογή των δεδομένων, την επεξεργασία-φιλτράρισμα τους και τη χρήση κάποιου κατάλληλου αλγορίθμου για την εξαγωγή της πρόβλεψης βάσει των επεξεργασμένων δεδομένων.

1.2.1 Συλλογή Στατιστικών Δεδομένων

Η συλλογή των δεδομένων μπορεί να γίνει με διάφορους τρόπους, είτε άμεσα και με ενεργή συμμετοχή του χρήστη (π.χ. με χρήση ερωτηματολογίων, φορμών συμπλήρωσης κ.λπ.) είτε έμμεσα με παθητική συμμετοχή του χρήστη (π.χ. με χρήση ειδικών προγραμμάτων συλλογής δεδομένων/προτιμήσεων των χρηστών (cookies), ειδικού εξοπλισμού καταγραφής αντιδράσεων σε διάφορα ερεθίσματα κ.λπ.). Κατά τη συλλογή αυτών των δεδομένων υπεισέρχονται ασφαλώς διάφορα ζητήματα που αφορούν την αξιοπιστία ή/και στατιστική καταλληλόλητα των συλλεχθέντων δεδομένων, τα οποία γενικά δε θα μας απασχολήσουν στην παρούσα εργασία και θα θεωρούμε κατά κανόνα ότι τα στατιστικά δεδομένα που έχουμε στη διάθεση μας είναι ακριβή και δεν προσθέτουν κάποιο άγνωστο επιπλέον τεχνικό σφάλμα.

1.2.2 Επεξεργασία-Φιλτράρισμα Δεδομένων

Στη συνέχεια ακολουθεί η επεξεργασία των δεδομένων, η οποία με τη σειρά της αποτελείται από ένα μάλλον τεχνικό κομμάτι μετατροπής τους σε μορφή κατάλληλη για είσοδο και ανάγνωση από τον αλγόριθμο πρόβλεψης, αλλά και από ένα πιο εξεζητημένο κομμάτι της επιλογής των κατάλληλων μόνο μεταβλητών για την εξαγωγή της πρόβλεψης και τον παράλληλο αποκλεισμό των μη χρήσιμων. Το στάδιο αυτό συχνά κρίνεται και αναπροσαρμόζεται βάσει της αξιολόγησης των αποτελεσμάτων του τρίτου σταδίου όπως στο μοντέλο καταρράκτη. Στην παρούσα εργασία, μελετάμε και παρουσιάζουμε, μεταξύ άλλων, τις δυσκολίες που προκύπτουν σε αμφότερα τα μέλη αυτού του σταδίου (με έμφαση στο δεύτερο και πιο ουσιώδες) όσο και τις μεθόδους που ακολουθήσαμε για την αντιμετώπιση τους.

1.2.3 Αλγόριθμος Πρόβλεψης

Τέλος στο τρίτο και πιο κρίσιμο στάδιο με χρήση κάποιου κατάλληλου αλγορίθμου αποφασίζεται για κάθε χρήστη (ή ομάδα χρηστών) ποια προβλέπεται να είναι η μελλοντική του κίνηση (συνήθως από ένα διακριτό φάσμα δυνατών επιλογών). Εδώ βρίσκεται η καρδιά της όλης διαδικασίας καθώς εξερευνάται ο τρόπος με τον οποίο τα εκάστοτε (όχι κατ' ανάγκη άμεσα συσχετιζόμενα) δεδομένα συνδέονται με τις μελλοντικές αποφάσεις κάθε χρήστη. Υπάρχουν διάφορα γενικά πλαίσια, στα οποία μπορεί να κινηθεί αυτός ο αλγόριθμος, όπως η μελέτη χρονοσειρών, στοχαστικών ανελίξεων, μοντέλων παλινδρόμησης κ.ο.κ. Η εύρεση της καλύτερης δυνατής μεθόδου γίνεται μέσω της αξιολόγησης των αποτελεσμάτων της, το οποίο αποτελεί έναν ξεχωριστό υποκλάδο ανάδρασης του τρίτου σταδίου και γίνεται βάσει διάφορων κριτηρίων, τα οποία επίσης μελετάμε διεξοδικά στα επόμενα κεφάλαια.

1.3 Δομή της Εργασίας

Ακολουθεί μια συνοπτική παρουσίαση της δομής της παρούσας εργασίας μαζί με σύντομη περιγραφή των περιεχομένων κάθε κεφαλαίου:

- **Κεφάλαιο 1:** *Εισαγωγή*
  Παρουσιάζεται το γενικό πλαίσιο του υπό μελέτη προβλήματος μαζί με τη γενική δομή της υπόλοιπης εργασίας.
- **Κεφάλαιο 2:** *Επισκόπηση του Προβλήματος*
  Παρατίθενται τα βασικά χαρακτηριστικά του συνόλου των δεδομένων που χρησιμοποιήθηκαν στο στιγμιότυπο που ασχοληθήκαμε. Μεταξύ άλλων γίνεται μια πλήρης παρουσίαση των δεδομένων που είχαμε για κάθε πελάτη του προγράμματος, μαζί με διάφορα στατιστικά αποτελέσματα για το σύνολο τους, τα οποία δίνουν μια εποπτική εικόνα της μορφολογίας της βάσης δεδομένων που είχαμε στη διάθεση μας. Παράλληλα γίνεται αναφορά στα κρίσιμα πεδία που θα αποτελέσουν το επίκεντρο της προσοχής μας στη δημιουργία του αλγορίθμου πρόβλεψης.

- **Κεφάλαιο 3:** *Πρωταρχικά Μοντέλα*
  Παρουσιάζονται οι πρώτες προσπάθειες πρόβλεψης του ρυθμού της αποχώρησης των πελατών, με μεθόδους δανεισμένες από τη μελέτη χρονοσειρών, όπως η χρήση κυλιόμενων παραθύρων υπό προϋποθέσεις μαρκοβιανής συμπεριφοράς. Παρότι οι μέθοδοι αυτές δεν απέδωσαν καρπούς, έδωσαν, μέσω της ανάλυσης των λόγων αποτυχίας τους, χρήσιμα στοιχεία και οδηγίες ως προς το πού πρέπει να κινηθούν οι επόμενες προσπάθειες για την αποφυγή επανάληψης τους.
- **Κεφάλαιο 4:** *Μοντέλα Παλινδρόμησης*
  Δίνεται μια γενική εισαγωγή στην Ανάλυση Παλινδρόμησης και ακολουθούν τα μοντέλα που στηρίχθηκαν σε αυτήν και κατάφεραν εν τέλει να δώσουν σημαντικά αποτελέσματα. Συγκεκριμένα παρατίθενται αρχικά οι προσεγγίσεις μέσω χρήσης δέντρων απόφασης με χρήση κριτηρίων βασισμένων σε συγκεκριμένα οικονομικά στοιχεία των πελατών (μαζί με πλήρη εξήγηση της αιτίας εστίασης σε αυτά) και στη συνέχεια παρουσιάζεται η προσέγγιση με χρήση μεθόδων clustering μαζί με σύγκριση των δύο μεθόδων. Στο τέλος παρουσιάζονται τα αποτελέσματα όταν υβριδοποιούνται οι δύο μέθοδοι, τα οποία και αποτελούν τις βέλτιστες προβλέψεις.
- **Κεφάλαιο 5:** *Συμπεράσματα και Επεκτάσεις*
  Γίνεται μια συνοπτική επισκόπηση των κύριων συμπερασμάτων της εργασίας και στη συνέχεια παρατίθεται μια πληθώρα δυνατών επεκτάσεων του συστήματος για την τελική κατασκευή ενός πλήρους και ευέλικτου εμπορικού προϊόντος.

Στο τέλος της εργασίας παρατίθενται η Βιβλιογραφία που χρησιμοποιήθηκε και το Παράρτημα με τους κώδικες που χρειάστηκαν για την επεξεργασία και τον υπολογισμό των δεδομένων και αποτελεσμάτων της εργασίας.

Κεφάλαιο 2
Επισκόπηση του Προβλήματος

2.1 Περιγραφή του Προβλήματος

2.1.1 Στιγμιότυπο Ανάλυσης

Στην παρούσα εργασία μελετάμε, όπως είπαμε, ένα συγκεκριμένο στιγμιότυπο του γενικότερου προβλήματος που μόλις περιγράψαμε και παρουσιάζουμε όλη τη πορεία από τους πρώτους σχετικά απλούς αλγορίθμους πρόβλεψης με χρήση χρονοσειρών, την αναπροσαρμογή του δεύτερου σταδίου φιλτραρίσματος των στατιστικών δεδομένων βασιζόμενοι στις αιτίες αποτυχίας των πρώτων προσεγγίσεων μέχρι και την κατασκευή μοντέλων βασιζόμενων σε μοντέλα παλινδρόμησης, τα οποία να παρουσιάζουν ικανοποιητικά ποσοστά επιτυχίας.

Συγκεκριμένα, σκοπός ήταν η ανάλυση ενός συνόλου στατιστικών δεδομένων που αφορούν τους πελάτες του προγράμματος "Affluent" μιας τράπεζας, με στόχο να αναπτυχθεί μια μέθοδος πρόβλεψης της οικειοθελούς αποχώρησης των πελατών από το πρόγραμμα. Στη δική μας περίπτωση, ένας πελάτης ορίζεται ως affluent (εύπορος) όταν η συνολική του θέση στην τράπεζα είναι άνω των 50.000€. Έτσι, το ερώτημα που έπρεπε να απαντηθεί είναι κατά πόσον τα δεδομένα που έχουμε για κάποιον πελάτη μπορούν να προβλέψουν (με ανεκτή επίδοση) το αν, σε κάποια στιγμή στο μέλλον, η συνολική του θέση θα είναι κάτω από τις 50.000€.

2.1.2 Σύνολο Δεδομένων

Τα δεδομένα που είχαμε διαθέσιμα για κάθε πελάτη απαρτίζονται από ορισμένα γενικά δημογραφικά στοιχεία, καθώς και από ένα πλήθος βασικών οικονομικών δεδομένων. Συνολικά, είχαμε στη διάθεση μας δεδομένα από 89698 πελάτες της τράπεζας για μια περίοδο 20 μηνών, συγκεκριμένα από τον Ιανουάριο του 2014 έως τον Αύγουστο του 2015. Ακολουθεί λεπτομερέστερη περιγραφή των πεδίων που διατηρούνταν για κάθε πελάτη μαζί με ορισμένα αντιπροσωπευτικά στατιστικά για κάθε κατηγορία, τα οποία δείχνουν τη γενική μορφολογία της βάσης μας.

Δημογραφικά Στοιχεία

Για κάθε πελάτη (έκαστος εκ των οποίων αναγνωρίζεται από ένα ξεχωριστό πεδίο-κλειδί *Id*) τηρούνται κάποια γενικά δημογραφικά στοιχεία, στιγμιότυπο των οποίων φαίνεται στην ακόλουθη εικόνα για 6 πελάτες:

Id	Sex	Age	Postal	MaritalStatus	Education	HomeStatus
1	F	49	15126	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
2	F	60	19014	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
3	F	58	10445	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΑΛΛΟ
4	M	49	17562	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΜΕΤΑΠΤΥΧΙΑΚΑ	UNKNOWN VALUE
5	M	54	14234	ΧΩΡΙΣΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
6	M	84	16232	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	UNKNOWN VALUE

Profession	Email	InternetConn	Phone	Months	Stay
ΕΛ.ΕΠ.ΔΗΜΟΣΙΟΓΡΑΦΟΣ	0	1	1	5	FALSE
ΔΗΜ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	2	FALSE
ΙΔ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	19	FALSE
ΙΔ.ΥΠΑΛ.ΛΟΙΠΕΣ ΕΙΔΙΚΟΤΗΤΕΣ	1	1	1	3	FALSE
ΙΔ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	3	FALSE
ΣΥΝΤΑΞ.ΟΑΕΕ	0	0	0	1	FALSE

Εικόνα 1: Δημογραφικά Στοιχεία Πελατών

Ακολουθεί σύντομη επεξήγηση των εμφανιζόμενων πεδίων μαζί με περιγραφή του πεδίου τιμών κάθε ενός:

Sex: Φύλο Πελάτη, *Διμελές Πεδίο Τιμών ('M' και 'F')*

Age: Ηλικία Πελάτη, *Θετική Ακέραια μεταβλητή*

Postal: Ταχυδρομικός Κώδικας Πελάτη, *Πενταψήφια Αριθμητική Συμβολοσειρά*

MaritalStatus: Οικογενειακή Κατάσταση Πελάτη, *Ολιγομελές Πεδίο Τιμών (χαρακτηριστικές τιμές φαίνονται στην εικόνα)*

Education: Επίπεδο Εκπαίδευσης Πελάτη, *Ολιγομελές Πεδίο Τιμών (χαρακτηριστικές τιμές φαίνονται στην εικόνα)*

HomeStatus: Οικιακή Κατάσταση Πελάτη, *Ολιγομελές Πεδίο Τιμών (χαρακτηριστικές τιμές φαίνονται στην εικόνα)*

Profession: Επάγγελμα Πελάτη, *Πολυμελές Πεδίο Τιμών (αρκετά μεγάλη εξειδίκευση στο είδος της κατηγορίας που ανήκει το επάγγελμα)*

Email: Τήρηση εκ μέρους του Πελάτη κάποιου Ηλεκτρονικού Ταχυδρομικού Λογαριασμού, *Boolean μεταβλητή*

InternetConn: Σύνδεση του Πελάτη με το Διαδίκτυο, *Boolean μεταβλητή*

Phone: Τήρηση εκ μέρους του Πελάτη κάποιου Λογαριασμού Κινητής Τηλεφωνίας, *Boolean μεταβλητή*

Months: Χρόνος Παραμονής του Πελάτη στο Πρόγραμμα, *Θετική Ακέραια μεταβλητή*

Stay: Διατήρηση του Πελάτη εντός του Προγράμματος, *Boolean μεταβλητή*

Μπορούμε να αποκτήσουμε μια συνοπτική εικόνα της γενικής μορφολογίας της βάσης δεδομένων, στηριζόμενοι στα ακόλουθα στατιστικά αποτελέσματα:

Sex: 68% Άντρες

Age: [Q1, Median, Mean, Q3] [1]= [49, 59, 59.57, 70]

Postal Codes: Κάλυψη σε όλη την Ελλάδα

Marital Status: 73% Παντρεμένοι

Education: (38% Τριτοβάθμια Εκπαίδευση, 31% Λύκειο)

Home Status: 50% Άγνωστο

Profession: Πολλές Κατηγορίες (κοκκώδη δεδομένα)

Email: 25%

Internet: 42%

Phone: 56%

Stay: 50.7%

Οικονομικά Στοιχεία

Παράλληλα με τα ανωτέρω στοιχεία, έχουμε στη διάθεση μας και τα ακόλουθα οικονομικά στοιχεία για τα οποία υπάρχει ξεχωριστή εγγραφή για κάθε πελάτη (πεδίο Id) και για κάθε μήνα (πεδίο Month), όπως φαίνεται στην εικόνα που ακολουθεί.

Id	Month	Immediate	Insurance	Investment	Business	Consumer	Closed	MB	Housing
1	2014.01	376.01	0	0	0	0	0	0	0.0
1	2014.02	497.30	0	0	0	0	0	0	0.0
1	2014.03	591.67	0	0	0	0	60000	3	0.0
1	2014.04	17035.37	0	0	0	0	0	0	-127749.1
1	2014.05	18035.37	0	0	0	0	0	0	-224712.6
1	2014.06	18028.70	0	0	0	0	0	0	-223748.9

Contributions	Sums	Status	Flag
0	60376.01	current	TRUE
0	60497.30	current	TRUE
0	60591.67	current	TRUE
0	77035.37	current	TRUE
0	78035.37	current	TRUE
0	78028.70	lost	FALSE

Εικόνα 2: Οικονομικά Στοιχεία Πελατών

Ακολουθεί σύντομη επεξήγηση των υπόλοιπων όρων και πεδίου τιμών τους:

Immediate: Άμεσα Διαθέσιμο Υπόλοιπο, *Πραγματική μεταβλητή*

Insurance: Ασφάλεια, *Πραγματική μεταβλητή*

[1] Υπενθυμίζουμε ότι Mean είναι η μέση τιμή και Q1, Median (Διάμεσος) και Q3 είναι αντίστοιχα η μέγιστη τιμή του χαμηλότερου 25%, 50% και 75% των τιμών (όταν αυτές είναι αύξοντα διατεταγμένες).

Investment: Επενδύσεις, *Πραγματική μεταβλητή*

Business: Επιχειρήσεις, *Πραγματική μεταβλητή*

Consumer: Καταναλωτικό Δάνειο, *Πραγματική μεταβλητή*

Closed: Προθεσμιακές Καταθέσεις, *Πραγματική μεταβλητή*

MB: Μονάδες Βάσης, *Πραγματική μεταβλητή*

Housing: Στεγαστικό Δάνειο, *Πραγματική μεταβλητή*

Contributions: Συνεισφορές, *Πραγματική μεταβλητή*

Sums: Αθροιστικές Καταθέσεις, *Πραγματική μεταβλητή*

Status: Κατάσταση Πελάτη, *ένα εκ των '', 'NEW', 'CURRENT', 'LOST' (ακολουθεί επεξήγηση)*

Flag: Σημαία Τρέχουσας Κατάστασης Πελάτη, *Boolean μεταβλητή*

Ακολουθούν και πάλι ορισμένα συνοπτικά στατιστικά για τις ανωτέρω μεταβλητές στη μορφή [Q1, Median, Mean, Q3]. Τα στατιστικά έχουν προκύψει από τις συνολικά 89698x20 εγγραφές (όπως προκύπτει από το γινόμενο πελατών και μηνών):

Immediate: [0, 2.8k, 15k, 13.6k]

Insurance: [0, 0, 4.9k, 0]

Investment: [0, 0, 15.6k, 0]

Business: [0, 0, -5.9k, 0]

Consumer: [0, 2.8k, 15k, 13.6k]

Closed: [0, 0, 4.9k, 0]

MB: [0, 0, 0.227, 0]

Housing: [0, 0, -4k, 0]

Contributions: [0, 0, 2.9k, 0]

Μήνες ως 'Affluent': [10, 20, 14.65, 20]

Γενικές Στατιστικές Πληροφορίες

Παρατηρούμε ότι σε πολλά από τα παραπάνω οικονομικά στατιστικά η μέση τιμή ρυθμίζεται από μόλις το άνω 25% του δείγματος καθώς $Q3 = Median = Q1 = 0$ (που σημαίνει ότι τουλάχιστον το 75% των πελατών (πελατομηνών για την ακρίβεια) δεν έχουν ενεργή συμμετοχή). Η έντονη επιρροή των μεγάλων τιμών φαίνεται και από το γεγονός ότι το άθροισμα των μέσων θέσεων των πελατών εντός του προγράμματος έχει υπολογιστεί σε 10Bn €, το οποίο με τη σειρά του αντιστοιχεί σε 146k € ανά πελάτη (111.8k € για όσους αργότερα έφυγαν από το πρόγραμμα), ενώ

το άθροισμα των μέσων θέσεων των πελατών εκτός του προγράμματος σε 0.3Bn € ή αλλιώς 7.2k € ανά πελάτη.

Γενικά θεωρούμε 4 δυνατές τιμές που μπορεί να έχει το Status ενός πελάτη για κάθε μήνα:

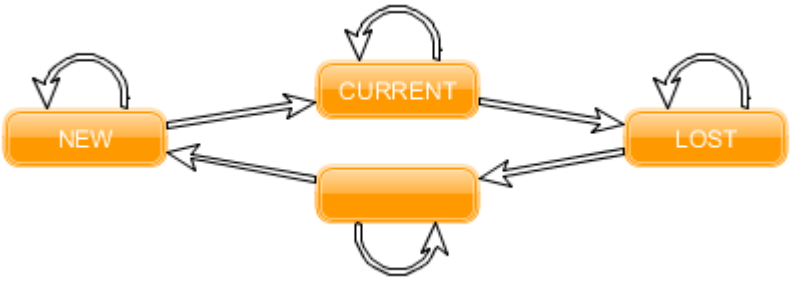
‘’: Εκτός Προγράμματος (για περισσότερο από έναν μήνα)

‘NEW’: Η κατάσταση για τον πρώτο μήνα εντός του προγράμματος

‘CURRENT’: Εντός προγράμματος (για περισσότερο από έναν μήνα)

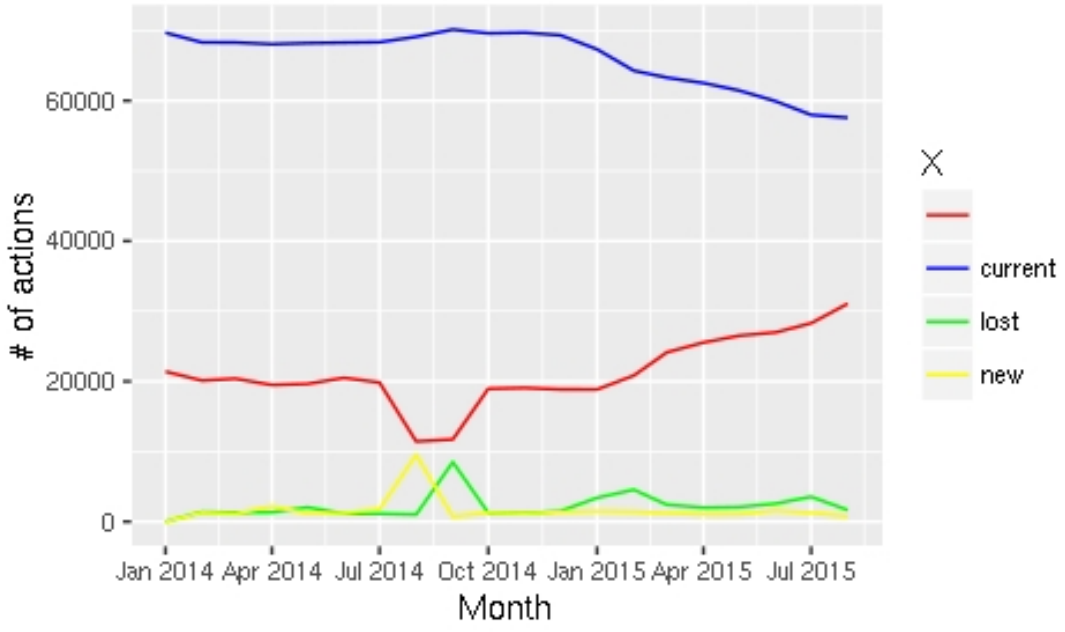
‘LOST’: Η κατάσταση για τον πρώτο μήνα εκτός του προγράμματος

Κάθε μήνα το status ενός πελάτη μπορεί είτε να παραμένει ίδιο είτε να μεταβαίνει κυκλικά στο επόμενο με τη σειρά που παρουσιάστηκαν ανωτέρω, όπως φαίνεται και στο σχετικό διάγραμμα που ακολουθεί:



Εικόνα 3: Διάγραμμα Μεταβάσεων Status

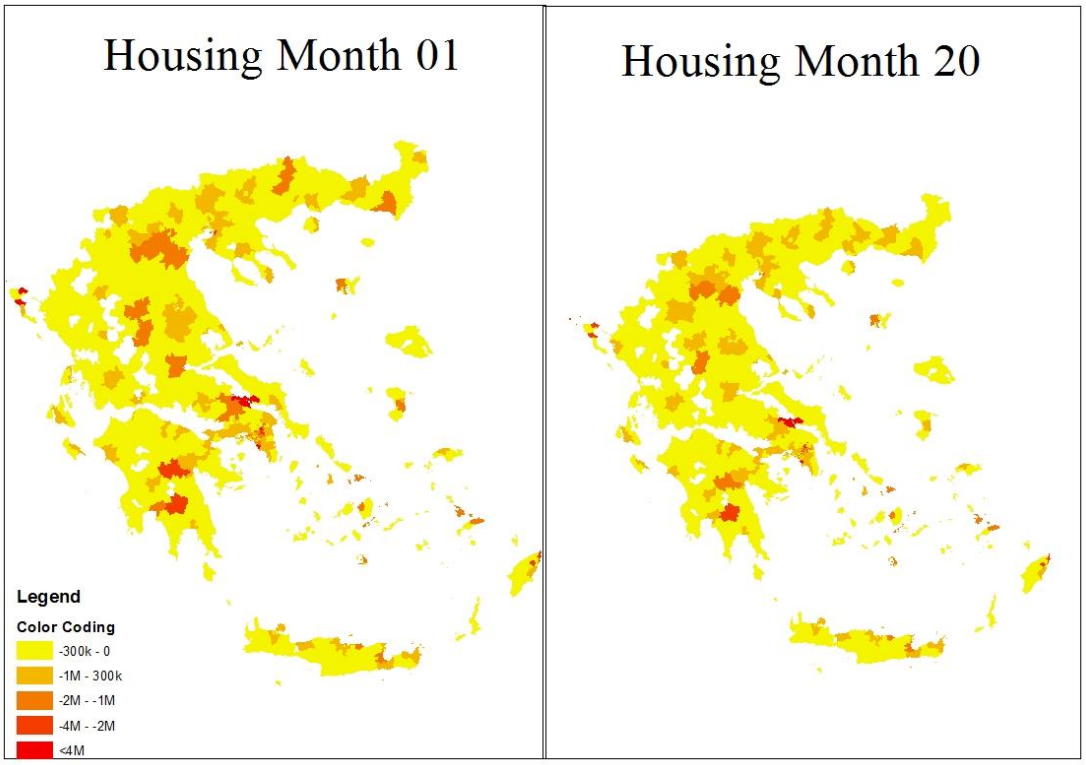
Για το διάστημα των 20 αυτών μηνών έχουμε το ακόλουθο γράφημα που περιγράφει τα ποσά των εντολών για κάθε status ανά μήνα:



Εικόνα 4: Εντολές ανά status και μήνα

Παρατηρούμε ότι αν και υπάρχει μια σχετική πτώση στα **'current'** (με παράλληλη αύξηση των εκτός προγράμματος **''**), όπως επίσης και μια απότομη καμπή στα μέσα του 2014, σε γενικές γραμμές βλέπουμε ότι έχουμε μια ομαλή καμπύλη, στο πλαίσιο αυτών των 20 μηνών (κάτι που θα δικαιολογήσει και την υιοθέτηση ορισμένων συγκεκριμένων μοντέλων όπως θα πούμε στη συνέχεια).

Τέλος, για λόγους πληρότητας, μπορούμε να πάρουμε, στην εικόνα που ακολουθεί, μια εποπτική εικόνα σύνδεσης ενός δημογραφικού κι ενός οικονομικού στοιχείου και συγκεκριμένα την εξέλιξη των στεγαστικών ανά ταχυδρομικό κώδικα από την αρχή έως το τέλος της περιόδου δειγματοληψίας.



Εικόνα 5: Μεταβολή των Στεγαστικών Δανείων ανά Περιοχή

2.2 Μοντέλα Επιβίωσης

2.2.1 Survival Analysis

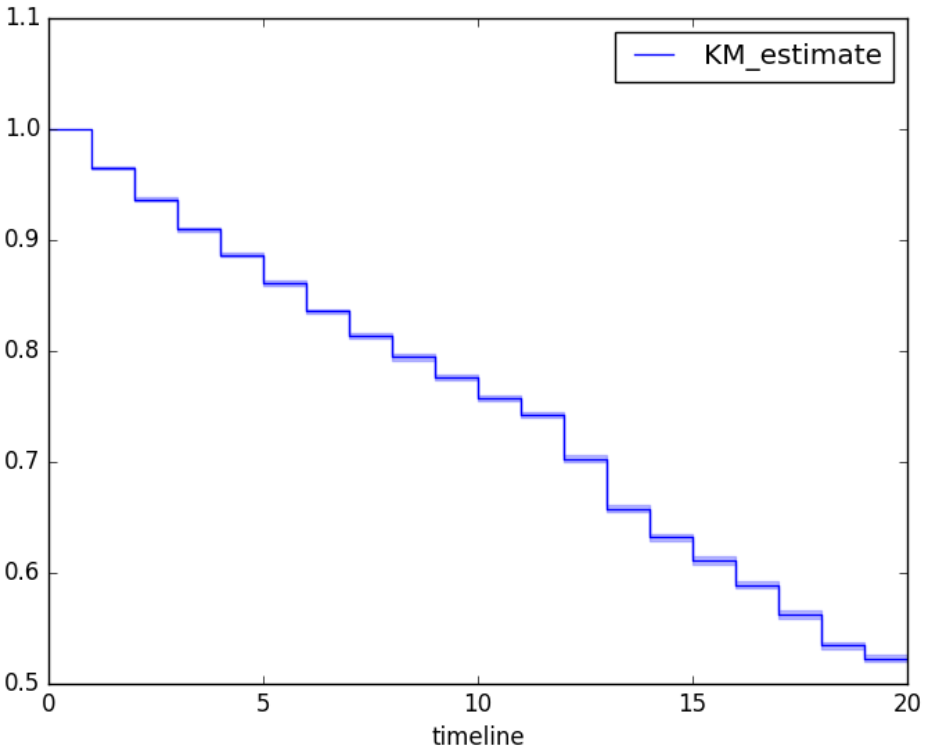
Το ανωτέρω πρόβλημα υπάγεται σε μια γενικότερη κατηγορία προβλημάτων, που απαρτίζουν το πεδίο μελέτης της Ανάλυσης Επιβίωσης (Survival Analysis) (βλ. και (Zhang, 2017)). Η Survival Analysis είναι υποκλάδος της Στατιστικής που μελετάει μετρικές σχετιζόμενες με μοντέλα στα οποία συμβαίνουν μονόδρομα γεγονότα σε ένα σύνολο χρηστών (π.χ. θάνατος σε ομάδα ζωντανών οργανισμών, κατάρρευση σε μονάδα υπολογιστικού συστήματος κ.τ.λ.). Βρίσκει εφαρμογές σε διάφορα βιολογικά, μηχανικά και κοινωνιολογικά συστήματα και σκοπός του είναι η ανάπτυξη εργαλείων ικανών να προβλέψουν με καλή πιθανότητα τους χρόνους και ρυθμούς αποβίωσης των αντικειμένων του συστήματος.

Οι βασικοί όροι που σχετίζονται με το πεδίο είναι τα «αντικείμενα» του συστήματος, τα «γεγονότα» (που συνήθως αντιστοιχούν στο θάνατο ή γενικά αποχώρηση ενός «αντικειμένου» από το σύστημα), ο «χρόνος» του «γεγονότος» και ο «χρόνος» παρατηρήσεων (που αντιστοιχεί στο χρονικό διάστημα ή βήματα μέχρις ότου να συμβεί το «γεγονός» (ή γενικά κάποιο «γεγονός») και το χρονικό διάστημα παρατήρησης αντίστοιχα) και η συνάρτηση επιβίωσης $S(t)$ (που αντιστοιχεί στην πιθανότητα ένα αντικείμενο να επιβιώσει περισσότερο από χρόνο t). Ένας ακόμη συνήθης όρος σε ειδικές περιπτώσεις δεδομένων είναι οι λογοκριμένες παρατηρήσεις που αντιστοιχούν σε αντικείμενα για τα οποία δεν έχει συμβεί το «γεγονός» για τον χρόνο παρατήρησης (ή και νωρίτερα σε περίπτωση που για οποιοδήποτε τεχνικό ή μη λόγο παύσει να συνεχίζεται η παρακολούθηση του). Στην περίπτωση που μελετάμε εμείς, το σύνολο των πελατών απαρτίζει τα «αντικείμενα», στην οποία συμβαίνουν τα γεγονότα «αποβίωσης» και συγκεκριμένα αποχώρησης από το τραπεζικό πρόγραμμα ‘Affluent’.

2.2.2 Μετρικές Kaplan-Meier και Nelson-Aalen

Εκτιμήτρια Kaplan-Meier

Ένας εύγλωττος τρόπος να ελέγξουμε τη συμπεριφορά του συστήματος μας ως προς την τάση αποχώρησης από το πρόγραμμα είναι τα γραφήματα εκτίμησης Kaplan-Meier (Meier, 1958). Στα γραφήματα αυτά παρουσιάζεται η πιθανότητα επιβίωσης (εν προκειμένω μη ερχομού μήνα αποχώρησης από το πρόγραμμα ‘Affluent’) κατά τη χρονική διάρκεια εξέλιξης του συστήματος.



Εικόνα 6: Εκτίμηση Kaplan-Meier για το τραπεζικό μοντέλο

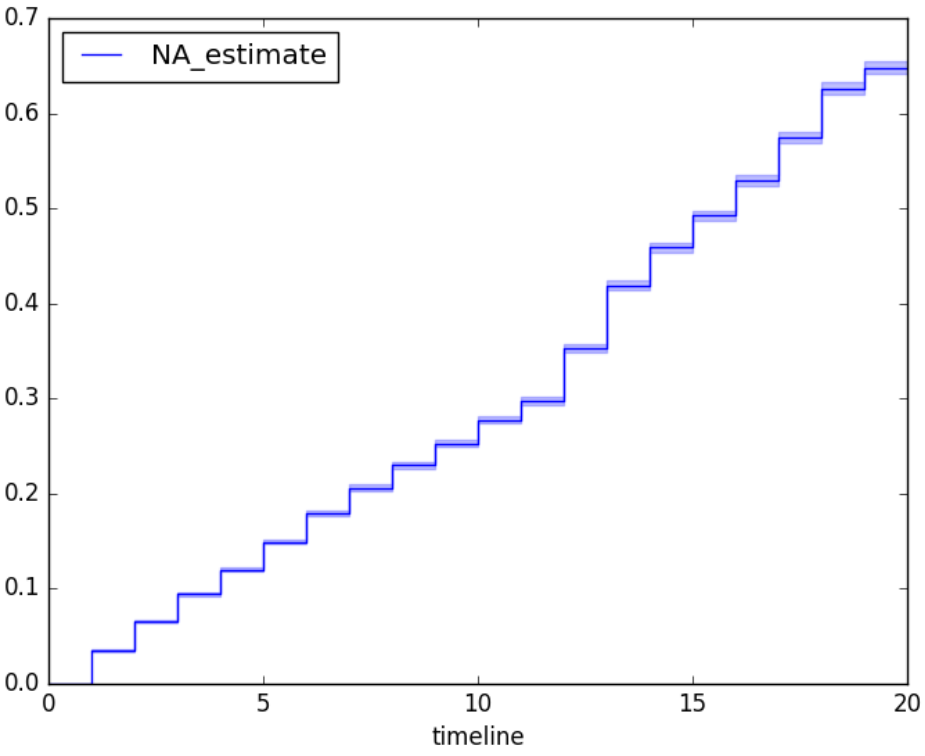
Συγκεκριμένα η μετρική αυτή ακολουθεί το πλήθος των επιβιωσάντων στο σύστημα παρουσιάζοντας έτσι μια συνοπτική εικόνα της πορείας και των εξάρσεων (ή σίγασης) «γεγονότων» κατά την πορεία της εξέλιξης. Γενικά έχει το χαρακτηριστικό ότι μπορεί να διαχειριστεί τυχόν λογοκριμένες παρατηρήσεις και ότι ασφαλώς τείνει στην τελική συνάρτηση επιβίωσης $S(t)$ του συστήματος.

Παραπάνω φαίνεται το σχετικό γράφημα για την περίπτωση μας, όπου παρατηρούμε ότι, όπως είναι αναμενόμενο, πρόκειται για μια φθίνουσα βηματική συνάρτηση (εφόσον σε κάθε μήνα φεύγει ένα θετικό ποσό πελατών) και ότι τείνει στην τελική τιμή που παρουσιάσαμε προηγουμένως.

Εκτιμήτρια Nelson-Aalen

Συμπληρωματική της εκτιμήτριας Kaplan-Meier και επίσης κατάλληλη για την περίπτωση ύπαρξης λογοκριμένων παρατηρήσεων είναι η μετρική Nelson-Aalen (βλ. (Colosimo, Ferreira, Oliveira, & Sousa, 2002)). Η μετρική αυτή υπολογίζει αθροιστικά το ποσοστό των «θανούντων» μέχρι κάθε χρονική στιγμή. Γενικά έχει πληθώρα εφαρμογών σε όλα τα αντίστοιχα επιστημονικά πεδία μελέτης, ενώ μπορεί να χρησιμοποιηθεί και για έλεγχο στατιστικής προσαρμοστικότητας των δεδομένων, όπως π.χ. για το αν τα «γεγονότα» ακολουθούν μια συγκεκριμένη κατανομή εμφάνισης κ.ο.κ.

Ακολουθεί το σχετικό διάγραμμα για την περίπτωση μας:



Εικόνα 7: Εκτίμηση Nelson-Aalen για το τραπεζικό μοντέλο

Γενικός σκοπός μας, λοιπόν, είναι η μελέτη της τάσης και των ιδιαίτερων χαρακτηριστικών αυτών των εγγραφών στο πλαίσιο αυτού του μοντέλου επιβίωσης με στόχο την εξαγωγή κάποιας αποτελεσματικής συσχέτισης των κατάλληλων μόνο εξ αυτών με την οικειοθελή αποχώρηση κάποιου πελάτη από το πρόγραμμα. Στα κεφάλαια που ακολουθούν παρουσιάζεται η πορεία που ακολουθήθηκε από παράλληλη ανάγνωση όλων των στοιχείων και διακριτοποίηση τους βάσει της φύσης τους (σε οικονομικά και δημογραφικά), μέχρι το φιλτράρισμα και εν τέλει τη διατήρηση μόνο ορισμένων από τα ανωτέρω πεδία, καθώς ασφαλώς και οι εναλλακτικοί τρόποι προσέγγισης και επεξεργασίας αυτών των δεδομένων προς αυτόν το σκοπό.

Κεφάλαιο 3
Πρωταρχικά Μοντέλα

3.1 Μελέτη Χρονοσειρών μέσω Μαρκοβιανών Μοντέλων

Μία πρώτη προσέγγιση στη πρόβλεψη της μελλοντικής συμπεριφοράς των πελατών ως προς την παραμονή τους εντός του προγράμματος 'Affluent' είναι η μελέτη των χρονοσειρών που προκύπτουν από τα εκάστοτε οικονομικά δεδομένα κάθε πελάτη. Ένα γενικό μοτίβο για τη μελέτη τέτοιου είδους χρονοσειρών, είναι η αρχική παραδοχή ότι η γενική τους μορφή πηγάζει από μια συγκεκριμένη οικογένεια συναρτήσεων (π.χ. εκθετικές-μαρκοβιανές στοχαστικές διαδικασίες, ταλάντωση με τυχαίο θόρυβο, γενική τάση μαζί με ημιτονοειδής εποχικότητα κ.ο.κ.) και στη συνέχεια η προσπάθεια εντοπισμού των ακριβών παραμέτρων της εν λόγω συνάρτησης, έτσι ώστε να μπορεί να γίνει η πρόβλεψη των μελλοντικών καταστάσεων.

Ο συνήθης τρόπος που γίνεται αυτή η μελέτη είναι αφότου έχει αποφασιστεί το γενικό μοντέλο στο οποίο υπάγεται η χρονοσειρά, να χρησιμοποιείται ένα κομμάτι της (π.χ. οι πρώτες τιμές μέχρι κάποιο συγκεκριμένο σημείο) για να «προπονήσουμε» τον αλγόριθμο πρόβλεψης. Στο σημείο αυτό της προπόνησης, γίνεται η προσπάθεια της εξαγωγής των τιμών των ειδικών παραμέτρων του συγκεκριμένου στιγμιότυπου, έτσι ώστε να χαρακτηριστεί πλήρως η υποβόσκουσα συνάρτηση και να είναι δυνατή η εφαρμογή του μοντέλου σε μελλοντικές τιμές. Στις περισσότερες περιπτώσεις ασφαλώς η αξιολόγηση του μοντέλου πρέπει να γίνει με τα ήδη υπάρχοντα δεδομένα και αυτός είναι ο λόγος που χρησιμοποιούμε μόνο ένα μέρος των δεδομένων που έχουμε στη διάθεση μας για την προπόνηση του αλγορίθμου: τα υπόλοιπα δεδομένα θα χρησιμοποιηθούν για την αξιολόγηση του προσομοιώνοντας μελλοντικές τιμές αφότου έχουν αποφασιστεί οι επί μέρους παράμετροι.

Στην περίπτωση που η αξιολόγηση προκύψει μη ικανοποιητική, τότε υπάρχουν διάφορα ενδεχόμενα, με βέλτιστο το να μην είναι αντιπροσωπευτικό το δείγμα πάνω στο οποίο έγινε η προπόνηση (π.χ. αφορούσε κάποια ειδική περίοδο κρίσης – μη αντιπροσωπευτική της μέσης κατάστασης πάνω στην οποία έγινε η αξιολόγηση) και χείριστο το να έχει χρησιμοποιηθεί εξαρχής ακατάλληλο μοντέλο για την περιγραφή των κινήσεων της χρονοσειράς. Η αρχική μας προσέγγιση, την οποία παρουσιάζουμε αναλυτικά στο υπόλοιπο αυτού του κεφαλαίου, αν και ιδιαιτέρως εύλογη, δε φάνηκε να αποτελεί ένα αρκετά ικανοποιητικό μοντέλο για την περίπτωση μας (οπότε και χρειάστηκε αντικατάσταση του, την οποία και μελετάμε εκτενέστερα στο επόμενο κεφάλαιο), παρόλα αυτά παρείχε πληθώρα χρήσιμων πληροφοριών για τα επόμενα.

3.1.1 Μαρκοβιανές Στοχαστικές Ανελίξεις

Η αρχική προσέγγιση, λοιπόν, ως προς το γενικό παράδειγμα που καθορίζει την εξέλιξη των χρονοσειρών μας ήταν η προσέγγιση με μαρκοβιανές στοχαστικές ανελίξεις (ή αλλιώς μαρκοβιανές αλυσίδες) (βλ. και (Migueis, Van den Poel, Camanho, & Falcao e Cunha, 2012) καθώς και μια εφαρμογή στο (Wu, 2012)). Χαρακτηριστικό των μαρκοβιανών αλυσίδων είναι η έλλειψη μνήμης. Συγκεκριμένα μια μαρκοβιανή αλυσίδα $X = \{X_0, X_1, X_2, \dots, X_n, \dots\}$ (διακριτού χρόνου) με καταστάσεις σε έναν χώρο $\mathbb{X}$ απαρτίζεται από αριθμήσιμου πλήθους τυχαίες μεταβλητές X_i (με κάθε μία να αντιστοιχεί στην κατάσταση που βρίσκεται η αλυσίδα την στιγμή i – ακριβώς όπως σε κάθε στοχαστική ανέλιξη διακριτού χρόνου) αλλά έχει την επιπρόσθετη ιδιότητα ότι:

$$\Pr[X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1, X_0 = x_0] = \Pr[X_{n+1} = x_{n+1} | X_n = x_n]$$

δηλαδή κάθε επόμενη κατάσταση της αλυσίδας (ισοδύναμα κάθε μετάβαση) εξαρτάται μόνο από την τωρινή κατάσταση και δεν επηρεάζονται καθόλου από το ιστορικό της (για αυτό και χαρακτηρίζονται ως memoryless).

3.1.2 Ομοιογενείς Μαρκοβιανές Στοχαστικές Ανελίξεις

Ας θεωρήσουμε την περίπτωση όπου το πλήθος των καταστάσεων είναι διακριτό και πεπερασμένο[2] $\mathbb{X} = \{x_0, x_1, \dots, x_k\}$. Αν π_n το διάνυσμα k θέσεων της κατανομής της X_n, τότε έχουμε για την ομοιογενή ή χρονικά ανεξάρτητη περίπτωση (όπου οι πιθανότητες μετάβασης μεταξύ δύο οποιονδήποτε καταστάσεων δεν εξαρτώνται από τον χρόνο n) ότι ισχύει:

$$\pi_{n+1} = \pi_n * A$$

όπου $A = [a_{ij}]$ ο πίνακας που στο (i, j) στοιχείο του έχει την πιθανότητα μετάβασης στο x_j όταν βρισκόμαστε στο x_i (εύκολα επαληθεύουμε ότι αυτές οι πιθανότητες είναι επαρκείς για να υπολογίσουμε κάθε στοιχείο του π_{n+1}, εφόσον λόγω της βασικής ιδιότητας των μαρκοβιανών αλυσίδων της έλλειψης μνήμης και εξάρτησης του X_{n+1} μόνο από το X_n έχουμε:

$$\pi_{n+1}^i = \sum_{j=0}^{k} \pi_n^j * \Pr\left[X_{n+1} = x_i \middle| X_n = x_j\right]$$

όπου όπως είπαμε $a_{ji} = \Pr[X_{n+1} = x_i | X_n = x_j]$). Ισοδύναμα, μπορούμε να αντιμετωπίσουμε αυτές τις αλυσίδες ως ένα τυχαιοκρατικό πεπερασμένο (ή μη) αυτόματο όπου τα a_{ji} αντιστοιχούν στις πιθανότητες μετάβασης μεταξύ των αντίστοιχων καταστάσεων του αυτομάτου.

Επομένως είναι εύκολο να καταλήξουμε, από τους παραπάνω τύπους, με απλή επαγωγή ότι:

[2] Αντίστοιχους ορισμούς έχουμε, ασφαλώς, και για τους συνεχείς ή/και άπειρους χώρους καταστάσεων.

$$\pi_n = \pi_0 * A^n$$

όπου π_0 η αρχική κατάσταση του συστήματος και η οποία συνήθως είναι ένα διάνυσμα που έχει έναν μοναδικό άσσο στην γνωστή αρχική κατάσταση (και στις άλλες θέσεις 0). Άρα για την χρονικά ανεξάρτητη περίπτωση, η γνώση του πίνακα μεταβάσεων A είναι αρκετή για τον χαρακτηρισμό όλης της στοχαστικής ανέλιξης.

Το παραπάνω αποτελεί ένα εύγλωττο παράδειγμα του γενικού μοτίβου που περιγράψαμε στην εισαγωγή του κεφαλαίου. Συγκεκριμένα, κάνουμε την παραδοχή ότι πίσω από τις κινήσεις των χρονοσειρών κρύβεται μια συνάρτηση που ανήκει στην οικογένεια των στοχαστικών ανελίξεων και αυτό που έπρεπε να κάνουμε για να καταφέρουμε να εξάγουμε την πληροφορία που χρειαζόμαστε για το μέλλον είναι ο υπολογισμός των παραμέτρων του συγκεκριμένου στιγμιότυπου, δηλαδή εν προκειμένω των πιθανοτήτων μετάβασης του πίνακα A.

3.1.3 Μαρκοβιανές Αλυσίδες Ανώτερης Τάξης

Στην παραπάνω περίπτωση είχαμε μηδενική μνήμη στο σύστημα, καθώς κάθε επόμενη κατάσταση εξαρτιόταν αποκλειστικά από την παρούσα. Εναλλακτικά μπορούμε να ορίσουμε μαρκοβιανές αλυσίδες όπου δίνουμε περισσότερη μνήμη στο σύστημα, συγκεκριμένα έστω ότι κάθε επόμενη κατάσταση εξαρτάται από τις m προηγούμενες:

$$\Pr[X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0] = \Pr[X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_{m'} = x_{m'}]$$

με $m' = n - m + 1$.

Μία τέτοια αλυσίδα ονομάζεται μαρκοβιανή αλυσίδα τάξης m (ή μνήμης m). Παρότι εκ πρώτης όψεως μπορεί να φαίνεται ότι προσφέρει περισσότερες δυνατότητες στην εκφραστικότητα της, εν τούτοις προκύπτει ότι κάθε μαρκοβιανή αλυσίδα τάξης m σε έναν χώρο $\mathbb{X}$ αντιστοιχεί σε μία μαρκοβιανή αλυσίδα στον χώρο $\mathbb{X}^m$. Πράγματι αν θεωρήσουμε τις τ.μ. $Y_i = [X_i, X_{i+1}, \dots, X_{i+m-1}]$, τότε είναι εύκολο να εξακριβώσουμε ότι το Y_{i+1} εξαρτάται μόνο από το Y_i (πράγματι τα $X_{i+1}, \dots, X_{i+m-1}$ καθορίζουν μονοσήμαντα τις αντίστοιχες συνιστώσες του Y_{i+1}, ενώ η τελευταία συνιστώσα του εξαρτάται μόνο από τις συνιστώσες του Y_i όπως ορίζει η παραπάνω εξίσωση). Οι πιθανότητες μετάβασης για τις Y_i προκύπτουν άμεσα από τις αντίστοιχες για τις X_i. Ακριβώς αντίστοιχα, μπορούμε να ορίσουμε και τις ομοιογενείς μαρκοβιανές αλυσίδες ανώτερης τάξης (για περισσότερες λεπτομέρειες επ' αυτών βλ. (Norris, 1997)).

3.1.4 Εφαρμογές Μαρκοβιανών Αλυσίδων

Το μοντέλο των μαρκοβιανών αλυσίδων είναι ένα από τα πλέον ευρέως χρησιμοποιούμενα μοντέλα σε πληθώρα επιστημών, όπως φυσική, χημεία, πληροφορική, οικονομετρία, στατιστική κ.λπ. Χαρακτηριστικά παραδείγματα της επιτυχίας των μαρκοβιανών αλυσίδων είναι η μοντελοποίηση των τυχαίων περίπατων (κίνηση Brown, εφαρμογές στην οικονομία), η Bayesian στατιστική, μια πληθώρα εφαρμογών στη μοντελοποίηση συστημάτων ελέγχου σε μηχανικούς και άλλους τομείς, αλλά ασφαλώς και ο διάσημος αλγόριθμος PageRank που έκανε τη

Google τη δημοφιλέστερη μηχανή αναζήτησης όπως και πρακτικά όλοι οι σύγχρονοι αλγόριθμοι πρόβλεψης διακλαδώσεων που χρησιμοποιούνται στους σημερινούς επεξεργαστές (με τους τελευταίους να βασίζονται κυρίως σε μαρκοβιανά μοντέλα ανώτερης τάξης). Τόσο η μεγάλη ζήτηση γύρω από αυτό το μοντέλο (λόγω των πολλών εφαρμογών του), όσο και η τεχνική ευκολία που προσφέρει το γεγονός της απώλειας μνήμης από πλευράς μαθηματικής ανάλυσης (σε πλήρη αντιστοιχία με ένα πεπερασμένο αυτόματο συγκρινόμενο με κάποια ανώτερη υπολογιστική μηχανή) έχουν οδηγήσει στην ανάπτυξη μιας μεγάλης συλλογής θεωρημάτων και εργαλείων γύρω από τις μαρκοβιανές αλυσίδες και τις έχουν καταστήσει ένα από τα πιο πρόσφορα πεδία μελέτης και μοντελοποίησης των περισσότερων παρατηρούμενων τυχαιοκρατικών διαδικασιών.

3.2 Ομοιογενή Μαρκοβιανά Μοντέλα για Πρόβλεψη του Churn Month

3.2.1 Χρήση Κυλιόμενων Παραθύρων

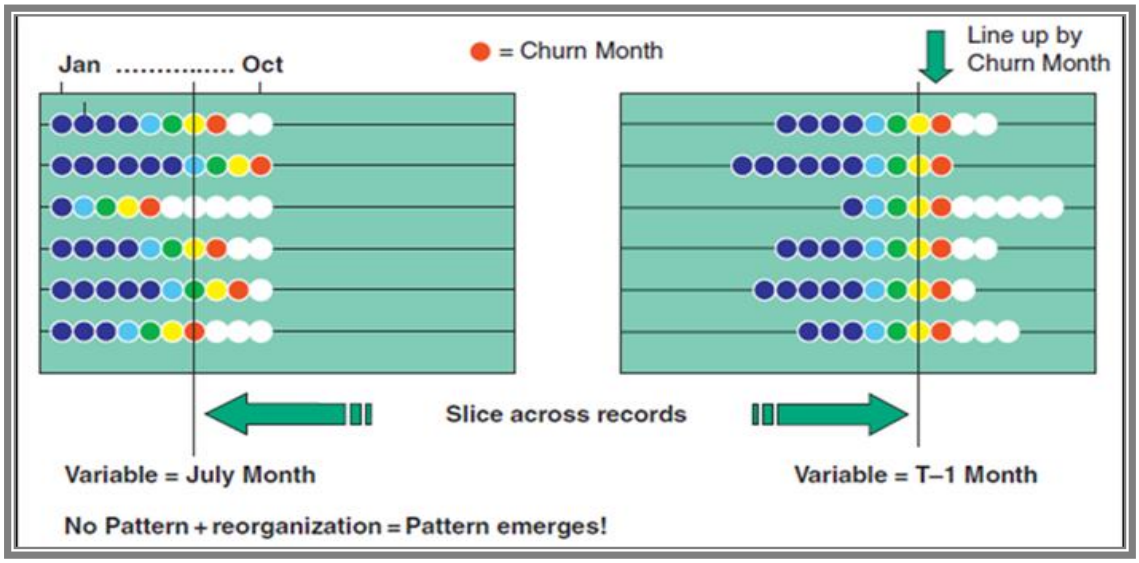
Θεωρήσαμε λοιπόν ότι η εξέλιξη της χρονοσειράς ρυθμίζεται βάσει ενός χρονικά ομοιογενούς μαρκοβιανού μοντέλου, όπου ο χώρος καταστάσεων, εδώ, δεν είναι άλλος από την σχετική θέση των οικονομικών παραμέτρων. Πρώτα από όλα, να σημειώσουμε ότι η επιλογή της χρονικής ομοιογένειας δικαιολογείται τόσο λόγω του σχετικά σύντομου χρονικού διαστήματος δειγματοληψίας, όσο και επειδή θεωρούμε ότι κάθε πελάτης δρα αυτόνομα, επομένως η χρονική στιγμή που θα συμβεί το churn month είναι ανεξάρτητη από την αντίστοιχη άλλων πελατών.

Ασφαλώς το να πάρουμε μαρκοβιανή αλυσίδα τάξης 1 δε πρόκειται να αποδώσει καρπούς, καθώς αφ' ενός από τα δεδομένα προκύπτει ότι οι μελλοντικές κινήσεις για συγκεκριμένα εύρη τιμών δεν φαίνεται να ακολουθούν συγκεκριμένα μοτίβα ώστε να δικαιολογούν μια τέτοια τάξη, όσο και αφ' ετέρου οι εμπειρικές παρατηρήσεις σπανίως δικαιολογούν μια τέτοια επιλογή. Πράγματι οι αποφάσεις που λαμβάνει ένας πελάτης δε μπορούμε να υποθέσουμε ότι ρυθμίζονται αποκλειστικά από την τωρινή του οικονομική κατάσταση, εφόσον είναι προφανές ότι εμπεριέχεται πληροφορία στην τάση που ακολουθεί η οικονομική χρονοσειρά (η οποία τάση απαιτεί περισσότερες από μία διαδοχικές τιμές). Θεωρούμε λοιπόν μαρκοβιανή αλυσίδα τάξης 2, έτσι ώστε η κατάσταση στον μήνα T να εξαρτάται από τις καταστάσεις των μηνών $T-1$ και $T-2$ και να μπορεί μεταξύ άλλων να υπολογιστεί και η κλίση της χρονοσειράς σε κάθε σημείο (για την οποία προφανώς αρκούν δύο σημεία).

Στο σημείο αυτό, πρέπει να κάνουμε ορισμένες παρατηρήσεις για τη συγκεκριμένη εφαρμογή του μαρκοβιανού μοντέλου. Σκοπός μας, εν προκειμένω, δεν είναι ένας καθολικός υπολογισμός των κατανομών των χρονοσειρών για κάθε χρονική στιγμή (ισοδύναμα ο υπολογισμός όλων των πιθανοτήτων του πίνακα μετάβασης A). Σκοπός μας, όπως έχουμε πει, είναι ο υπολογισμός του εάν ένας πελάτης θα αποχωρήσει από το πρόγραμμα κάποια στιγμή στο μέλλον. Δεδομένου ότι, εξ υποθέσεως, κάθε μελλοντική απόφαση εξαρτάται μόνο από τους προηγούμενους 2 μήνες, αρκεί να

θεωρούμε ένα κυλιόμενο παράθυρο $(T-2, T-1, T)$ από το οποίο μελετώντας την κατάσταση στους μήνες $T-2, T-1$ προσπαθούμε να προβλέψουμε αν ο T θα είναι ο churn month.

Γενικά η επιλογή του κατάλληλου μήκους παραθύρου εξαρτάται από διάφορους παράγοντες και στην πράξη είναι συνήθως απόρροια συνδυασμού εμπειρικών μετρήσεων και δοκιμών. Ένα μεγάλο παράθυρο έχει αυξημένες πιθανότητες επιτυχίας σε πολύπλοκα συστήματα με μεγάλο εύρος τιμών και έντονη εποχικότητα, ενώ εμπεριέχει τον κίνδυνο να κάνει υπερφόρτωση δεδομένων με αποτέλεσμα να προκύπτουν πολλά διαφορετικά μοτίβα τα οποία δε μπορούν να ομαδοποιηθούν και κατηγοριοποιηθούν βάσει της προβλεπόμενης τιμής. Επιπλέον σε περίπτωση βραχέος διαστήματος παρελθοντικών δεδομένων, ένα μεγάλο παράθυρο κάνει απαγορευτική την αξιολόγηση της μεθόδου, ενώ σε περίπτωση μακρού υπάρχει η πιθανότητα να «μολυνθούν» τα παράθυρα από τιμές που αντιστοιχούν σε διαφορετικής συμπεριφοράς περιόδους (καθώς σε μεγαλύτερα χρονικά διαστήματα είναι και πιθανότερη η έξαρση περισσότερων περιόδων-κρίσεων που δε συμβαδίζουν με τη μέση περίπτωση). Ένα μικρό παράθυρο, από την άλλη, εμπεριέχει τον κίνδυνο της υπεραπλούστευσης του μηχανισμού εξέλιξης των χρονοσειρών, αλλά αποτελεί τη καλύτερη λύση για σχετικά σύντομα (και άρα μάλλον σταθερά) χρονικά διαστήματα δειγματοληψίας. Η σχετική ευστάθεια που προκύπτει και από το γράφημα της Εικόνας 4 (καθώς και η γενικότερη χρήση του τριμήνου ως τυπικό μέτρο στις οικονομικές μελέτες) δικαιολογεί λοιπόν την επιλογή του κυλιόμενου παραθύρου σε ένα εύρος τριών μηνών.



Εικόνα 8: Ευθυγράμμιση ως προς Churn Month για τον εντοπισμό του μοτίβου αποχώρησης.

Ικανό και αναγκαίο βήμα για τον χαρακτηρισμό των καταστάσεων που οδηγούν σε αποχώρηση από το πρόγραμμα είναι ο εντοπισμός των επιπέδων που οδηγούν στον churn month. Στην περίπτωση της αλυσίδας δεύτερης τάξης, αυτό που ψάχνουμε είναι τα μεγέθη που υπάρχουν στους μήνες $T-2, T-1$ όταν T είναι ο churn month (ισοδύναμα, ψάχνουμε τις καταστάσεις που με μεγάλη πιθανότητα οδηγούν την ισοδύναμη μαρκοβιανή αλυσίδα σε κατάσταση αποχώρησης). Αν το καταφέρουμε

αυτό, τότε θα έχουμε κατασκευάσει ένα εργαλείο, τέτοιο ώστε όταν εμφανίζονται σε δύο διαδοχικούς μήνες κάποιο από τα μοτίβα που με μεγάλη πιθανότητα σημαίνουν αποχώρηση, θα μπορούμε να προβλέπουμε ότι στον επόμενο μήνα με ικανοποιητική ακρίβεια ότι όντως θα αποχωρήσει ο αντίστοιχος πελάτης (όπως ακριβώς ζητούσαμε).

Εν προκειμένω έχουμε την ευχέρεια να γνωρίζουμε ακριβώς τις ζητούμενες τελικές καταστάσεις καθώς και τη στιγμή που συνέβησαν, επομένως για να ανακύψει το ζητούμενο μοτίβο, εκμεταλλευόμαστε τη χρονική ομοιογένεια που υποθέσαμε για το μοντέλο μας, και μετακινούμε τις χρονοσειρές έτσι ώστε σημείο αναφοράς να είναι ο μήνας που προηγείται της αποχώρησης, όπως φαίνεται και στο ανωτέρω σχήμα, με στόχο τον εντοπισμό του ζητούμενου μοτίβου. Ο λόγος που έχουμε το δικαίωμα να κάνουμε αυτή την ολίσθηση, είναι επειδή τυχόν μαρκοβιανό μοτίβο που θα ανακύπτει πριν τον ερχομό churn month οφείλει, ακριβώς λόγω της χρονικής ομοιογένειας, να εμφανίζεται στην ίδια μορφή ανεξάρτητα της χρονικής στιγμής που αυτός συμβαίνει.

Στο Παράρτημα παρατίθεται ο ακριβής κώδικας που χρησιμοποιήθηκε για την παραγωγή του συγκεκριμένου μοντέλου σε γλώσσα R.

3.2.2 Λόγοι ανεπάρκειας του μαρκοβιανού μοντέλου

Παρότι για τους λόγους που εξηγήσαμε, η μοντελοποίηση μέσω ενός μαρκοβιανού μοντέλου ήταν μια εύλογη επιλογή, προέκυψε κατά την αξιολόγηση ότι δεν σχηματιζόταν κάποια επαρκής μορφολογία που να μπορεί να προβλέψει με ικανοποιητική επιτυχία τον ερχομό του churn month. Υπήρχαν διάφοροι λόγοι για τους οποίους συνέβη αυτό, με έναν από τους κυριότερους να αποτελεί η «μόλυνση» των δεδομένων με «παραπλανητικές» πληροφορίες.

Πρώτα από όλα, όλα τα στοιχεία καθορίζονταν από τα δεδομένα ενός πελάτη και το παραπάνω μοντέλο δεν είχε τη δυνατότητα να συσχετίσει τις διάφορες εγγραφές ενός πελάτη μεταξύ τους. Αυτό είχε ως συνέπεια να συγχρωτιστούν οι επί μέρους συνεισφορές κάθε εγγραφής σε μία κοινή με αποτέλεσμα να υπεισέρχονται αλλοιωτικές πληροφορίες από εγγραφές που (από ό,τι φάνηκε εκ των υστέρων) δε σχετίζονταν με το αποτέλεσμα που ζητούνταν να προβλεφθεί. Ως αποτέλεσμα, υπεισερχόταν στη φάση της «προπόνησης» του αλγορίθμου αυξημένος θόρυβος, ο οποίος εκ κατασκευής δεν μπορούσε να ακυρωθεί από τον αλγόριθμο πρόβλεψης.

Ο ίδιος θόρυβος ασφαλώς υπήρχε και στη φάση της αξιολόγησης (αντίστοιχα σε πιθανές μελλοντικές τιμές) και η μη ικανότητα απαλοιφής του, είχε ως αποτέλεσμα να προκύπτουν πολλά διαφορετικά μοτίβα τα οποία (αφότου κατηγοριοποιηθούν σε μία κοινή κατάσταση του αντίστοιχου μαρκοβιανού μοντέλου) να προκύπτουν ότι ανήκουν στην ίδια κατάσταση εξαιτίας διαφορετικών, όμως, αιτιών και έτσι σε κάθε κατάσταση να συγχρωτίζονται χρονοσειρές διαφορετικής συμπεριφοράς ως προς τον churn month, εξαιτίας της αλλοίωσης που εισάγει αυτός ο θόρυβος. Η εικόνα που υπήρχε εν τέλει στο μαρκοβιανό μοντέλο που δημιουργούνταν, ήταν η ενσωμάτωση στις ίδιες καταστάσεις τόσο χρονοσειρών που οδηγούσαν σε churn month όσο και

μη. Ήταν προφανές λοιπόν ότι το μαρκοβιανό μοντέλο δε μπορούσε να χαρακτηρίσει τις ζητούμενες χρονοσειρές κι άρα επρόκειτο για μια ακατάλληλη προσέγγιση ως προς το εν λόγω πρόβλημα.

Εν τούτοις, ακόμη και με αυτή την έκβαση, ο πειραματισμός με αυτό το μοντέλο και η εύρεση των αιτιών για τις οποίες απέτυχε, στάθηκε θεμελιώδους σημασίας για την εξαγωγή των χαρακτηριστικών που οφείλει να έχει το ορθό μοντέλο (και εν τέλει την εύρεση του). Πράγματι, τα συμπεράσματα που βγάζουμε, λοιπόν, από αυτή την προσέγγιση, είναι ότι κατ' αρχήν χρειαζόμαστε ένα μοντέλο, που κάνει χρήση μόνο όσων δεδομένων έχουν πραγματική συσχέτιση με τον πιθανό ερχομό churn month και όχι μαζική τους χρήση, καθώς όπως φάνηκε υπάρχουν ποσοστά που μπορούν να αλλοιώσουν σημαντικά τη χρήσιμη πληροφορία. Ακόμη και σε αυτή την περίπτωση, ωστόσο, θα υπάρχει θόρυβος, ο οποίος αναπόφευκτα θα δημιουργεί παραπλανητικά μοτίβα σε ένα κυλιόμενο παράθυρο σταθερού μήκους, κάνοντας συχνά αδύνατο τον σωστό χαρακτηρισμό των επιθυμητών χρονοσειρών. Πράγματι, αν δώσουμε μεγάλη εκλέπτυνση στο σύστημα, τότε προκύπτουν πολλά μοτίβα, για έκαστο εκ των οποίων πρέπει να δίνουμε ξεχωριστή πρόβλεψη (το οποίο προφανώς είναι άστοχο καθώς σκοπός είναι η δημιουργία clusters από χρονοσειρές όμοιας συμπεριφοράς), ενώ από την άλλη όταν δεν δίνουμε αρκετή εκλέπτυνση, ο εισερχόμενος θόρυβος στα πεπερασμένου μήκους παράθυρα, οδηγεί στον αναγκαστικό συγχρωτισμό διαφορετικής φύσεως χρονοσειρών στην ίδια ομάδα εξαιτίας του αναπόφευκτου θορύβου.

Συνοψίζοντας, χρειαζόμαστε ένα μοντέλο που να διατηρεί μόνο τις πληροφορίες που εμπεριέχουν ουσιώδη συσχέτιση με τη πληροφορία της ζητούμενης πρόβλεψης, αλλά παράλληλα ακόμη και αυτές οι κατάλληλες εγγραφές να εξαπλώνονται σε όσο το δυνατόν περισσότερο εύρος (χρονικά) έτσι ώστε να ελαχιστοποιείται η επίδραση του θορύβου. Στο επόμενο κεφάλαιο παρουσιάζουμε ένα εναλλακτικό μοντέλο που λαμβάνει υπόψιν του τα παραπάνω και φαίνεται να παρουσιάζει δραστικά καλύτερες επιδόσεις.

Κεφάλαιο 4
Μοντέλα Παλινδρόμησης

4.1 Ανάλυση Παλινδρόμησης

Όπως αναφέραμε στο τέλος του προηγούμενου κεφαλαίου, υπάρχει πρωταρχική ανάγκη εύρεσης των μεταβλητών του συστήματος μας, οι οποίες να έχουν έντονη συσχέτιση με το προβλεπτέο, δηλαδή την εμφάνιση churn month. Σε συνέχεια των προηγούμενων προσπαθειών, χρησιμοποιήθηκε για την εξεύρεση τέτοιων σχέσεων μία από τις κλασικότερες μεθόδους προς αυτό το σκοπό : η ανάλυση παλινδρόμησης.

Η ανάλυση παλινδρόμησης είναι ένα γενικό πλαίσιο υπολογιστικών μεθόδων, που προσπαθούν να εξάγουν την ακριβή σχέση αλληλεπίδρασης μεταξύ δύο (ή παραπάνω) παρατηρούμενων μεγεθών. Συνήθως ένα από τα δύο θεωρείται ως η ανεξάρτητη μεταβλητή και η άλλη ως η εξαρτημένη (αν και το μοντέλο μπορεί να εφαρμοστεί ακόμη και σε γενικότερες περιπτώσεις). Πρόκειται για μια ευρεία μέθοδο που χρησιμοποιείται στη στατιστική και η οποία από ένα πειραματικό δείγμα ζευγών τιμών, θεωρώντας γνωστή τη γενική μορφή της συνάρτησης που διέπει τη μεταξύ τους σχέση, προσπαθεί να εξάγει τις παραμέτρους εκείνες, οι οποίες να συμβαδίζουν όσο το δυνατό περισσότερο με αυτό το δείγμα (ισοδύναμα, τις παραμέτρους για τις οποίες το σφάλμα που αντιστοιχεί σε αυτά τα παρατηρούμενα ζεύγη τιμών να ελαχιστοποιείται). Παρατηρούμε για ακόμη μια φορά, ότι αναπόφευκτα και πάλι υποθέτουμε ένα συγκεκριμένο περιβάλλον δυνατών συναρτήσεων και στόχος της ανάλυσης είναι η εξαγωγή των καλύτερων δυνατών τιμών για αυτή τη συνάρτηση.

4.1.1 Ανάλυση Ελαχίστων τετραγώνων

Θεωρούμε λοιπόν δύο μεταβλητές $\boldsymbol{X}, \boldsymbol{Y}$ όπου $\boldsymbol{X}$ είναι η ανεξάρτητη μεταβλητή και $\boldsymbol{Y}$ η εξαρτημένη (σημειωτέον ότι τα $\boldsymbol{X}, \boldsymbol{Y}$ μπορεί να είναι διανύσματα, πολυδιάστατες μεταβλητές κ.ο.κ.) καθώς και ότι διέπονται από μια σχέση της μορφής $\boldsymbol{Y} = F(\boldsymbol{X}, \boldsymbol{a})$ όπου $\boldsymbol{a}$ το διάνυσμα των άγνωστων παραμέτρων προς προσδιορισμό. Θεωρούμε, συνήθως όπως είπαμε, ότι η F, δηλαδή αυτή που διέπει τη γενική μορφή της σχέσης μεταξύ των $\boldsymbol{X}, \boldsymbol{Y}$ είναι γνωστή συνάρτηση. Διαθέτουμε επίσης ένα σύνολο από μετρήσεις (x_i, y_i) και στόχος είναι να βρούμε την τιμή εκείνη του $\boldsymbol{a}$ για την οποία οι αποκλίσεις των παρατηρούμενων μετρήσεων να ελαχιστοποιείται. Δεδομένου ότι αυτές οι αποκλίσεις μπορεί να έχουν πρόσημο, προσπαθούμε να ελαχιστοποιήσουμε το τετράγωνο τους και συγκεκριμένα, ζητάμε το $\boldsymbol{a}$ για το οποίο το

$$\sum |y_i - F(x_i, \boldsymbol{a})|^2$$

να ελαχιστοποιείται.

Για παράδειγμα στη περίπτωση της γραμμικής παλινδρόμησης μεταξύ δύο μονοδιάστατων μεταβλητών, όπου η F είναι μια γραμμική συνάρτηση, έχουμε ότι τα X, Y συνδέονται με μια σχέση της μορφής $Y = \beta X + \gamma$ με $\boldsymbol{a} = (\beta, \gamma)$. Επομένως προσπαθώντας να ελαχιστοποιήσουμε τη ποσότητα που προκύπτει από το ανωτέρω άθροισμα, μπορούν να προκύψουν με χρήση στοιχειώδους ανάλυσης οι πλέον κατάλληλες τιμές για τα β, γ συναρτήσει των (x_i, y_i) [3]. Αντίστοιχες, αλλά ασφαλώς πιο περίπλοκες είναι οι αναλύσεις για πιο σύνθετες F (στη θέση των οποίων μπαίνουν συνήθως πολυώνυμα ανώτερης τάξης, καθώς η ακριβής επίλυση για πιο περίπλοκες συναρτήσεις παραμένει ανοιχτό πρόβλημα).

4.1.2 Πρόβλεψη με διαστήματα εμπιστοσύνης

Θεωρώντας, μάλιστα, δεδομένη κάποια στατιστική ομοιομορφία στην κατανομή που ακολουθούν οι τιμές του πειραματικού δείγματος, το μοντέλο της παλινδρόμησης έχει το μεγάλο πλεονέκτημα ότι πέραν της εκτίμησης μιας συνάρτησης που φαίνεται να ακολουθεί τη σχέση μεταξύ των δύο υπό ερεύνηση μεταβλητών, δίνει παράλληλα και μια εκτίμηση της αξιοπιστίας αυτής της συνάρτησης. Πράγματι, μελετώντας τη διασπορά των τιμών γύρω από τη κεντρική τους τιμή, είμαστε σε θέση να εξάγουμε μετρικές και διαστήματα εμπιστοσύνης, τα οποία να μας ενημερώνουν για το κατά πόσο είναι «συνεπή» τα δεδομένα με το επιλεχθέν μοντέλο, αλλά ακόμη περισσότερο να δίνουν μια εικόνα της αξιοπιστίας των προβλεπόμενων τιμών για τα διάφορα εύρη που τυχόν μπορεί να κυμαίνεται η ανεξάρτητη μεταβλητή. Ορισμένες τέτοιες μετρικές αξιοπιστίας θα εξάγουμε και στα διάφορα μοντέλα που θα εφαρμόσουμε στη συνέχεια και περισσότερες λεπτομέρειες θα παρουσιάσουμε στις σχετικές ενότητες.

4.2 Δέντρα Απόφασης

4.2.1 Εκμάθηση Δέντρων Απόφασης: Ανάλυση CART

Ασφαλώς υπάρχει η περίπτωση όπου δεν γνωρίζουμε αν υπάρχει κάποια απλή γραμμική ή πολυωνυμική εξάρτηση μεταξύ των υπό μελέτη μεταβλητών, ή ακόμη περισσότερο να έχουμε αρκετά στοιχεία που να μαρτυρούν ότι δε δικαιολογείται μια τέτοια μορφή συσχέτισης οπότε να αποκλείεται και η πιθανότητα χρήσης κάποιου τέτοιου προσεγγιστικού μοντέλου. Σε αυτή την περίπτωση οφείλουμε να πάρουμε την ευρύτερη δυνατή μορφή μιας συνάρτησης, που δεν είναι άλλη από τον εξαντλητικό ανεξάρτητο ορισμό κάθε δυνατής εξόδου για κάθε πιθανή είσοδο. Ασφαλώς για αυτή την περίπτωση δε μπορούμε να χρησιμοποιούμε συνεχές φάσμα και θα πρέπει να κάνουμε κάποια διακριτοποίηση του σε συστάδες. Το μήκος περιγραφής της αντίστοιχης συνάρτησης θα είναι προφανώς τότε όσο το γινόμενο του πλήθους των συστάδων για κάθε μεταβλητή του ανεξάρτητου διανύσματος $\boldsymbol{X}$.

Ένας εύγλωττος τρόπος να αναπαρασταθεί μια τέτοια συνάρτηση είναι με ένα δέντρο απόφασης (βλ. (Quinlan, 1986)). Στα δέντρα απόφασης, κάθε εσωτερικός κόμβος

[3] Θεωρούμε ασφαλώς πάντοτε ότι το δείγμα των πειραματικών τιμών είναι (σημαντικά) μεγαλύτερο από τη διάσταση του προς καθορισμό διανύσματος $\boldsymbol{a}$.

είναι μια ερώτηση (πάνω στις ανεξάρτητες μεταβλητές), η απάντηση της οποίας[4] οδηγεί στο επόμενο επίπεδο ερωτήσεων μέχρι να καταλήξουμε σε κάποιο φύλλο. Στα φύλλα δίνεται η απόφαση του δέντρου για την εξαρτημένη μεταβλητή. Αν η τελευταία μπορεί να πάρει διακριτό πλήθος τιμών (εν προκειμένω αν θα υπάρξει churn month ή όχι), τότε τα δέντρα αυτά λέγονται και δέντρα κατάταξης, αλλιώς όταν η έξοδος αφορά συνεχή τιμή λέγονται δέντρα παλινδρόμησης (εμείς θα ασχοληθούμε κυρίως με δέντρα κατάταξης, τα οποία γενικά θα αποκαλούμε δέντρα απόφασης).

Επομένως είναι προφανές ότι πρέπει να βρεθεί μια χρυσή τομή μεταξύ της ανάγκης παραγωγής πολλών συστάδων για την εξασφάλιση ενός καλού επιπέδου ακρίβειας αλλά και παράλληλα η αποφυγή εμφανίσεων πολλών διακλαδώσεων στο αντίστοιχο δέντρο καθώς το μέγεθος του αυξάνεται εκθετικά με το πλήθος των ανεξάρτητων μεταβλητών. Στην πραγματικότητα για να είναι εφικτή η χρήση ενός τέτοιου δέντρου θα πρέπει να προτιμηθεί κάποια ιεραρχία που να κλαδεύει μεγάλα τμήματα του δέντρου κάτι που ρυθμίζεται από την κατάλληλη τοποθέτηση των ερωτημάτων απόφασης.

Ακόμη κι αν είχαμε όμως αλγόριθμο που να βρίσκει σε αποδοτικό χρόνο ποια είναι αυτή η τοποθέτηση που ελαχιστοποιεί το μήκος του δέντρου (κάτι το οποίο προκύπτει ότι είναι NP-hard), αποδεικνύεται ότι στις περισσότερες συναρτήσεις, ήδη το μήκος του δέντρου είναι κοντά στο μέγιστο δυνατό (κι άρα εκθετικού μεγέθους). Επομένως, γνωρίζουμε ότι αναγκαστικά ένα δέντρο απόφασης για ένα μεγάλο πλήθος δεδομένων είναι προσεγγιστικό δέντρο και κάθε απόφαση θα έχει μια πιθανότητα επιτυχίας (γενικά μικρότερη του 1). Επομένως κάθε φύλλο (και αναδρομικά κάθε κόμβος) του δέντρου συνοδεύεται από δύο τιμές:

- Μία τιμή στο $[0,1]$ που δείχνει την ομοιογένεια της απόφασης. Συγκεκριμένα, αν έχουμε δύο πιθανές κλάσεις της εξαρτημένης μεταβλητής (ισοδύναμα μία κλάση και το συμπλήρωμα της), η τιμή αυτή αντιστοιχεί στο ποσοστό των ατόμων που καταλήγουν σε αυτό το φύλλο και ανήκουν στη κλάση. Τα φύλλα με μεγάλες τιμές αντιστοιχίζονται στην περίπτωση που η εξαρτημένη μεταβλητή ανήκει στη κλάση, ενώ οι μικρές στη συμπληρωματική (προφανώς ένα δέντρο είναι τόσο πιο ακριβές στις αποφάσεις του, όσο πιο κοντά στις ακραίες τιμές 0 και 1 είναι οι αντίστοιχες πιθανότητες).
- Μία τιμή (συνήθως σε μορφή ποσοστού %) που δείχνει το ποσοστό των παρατηρήσεων που καταλήγουν σε αυτό το φύλλο. Το ποσοστό αυτό είναι πολύ χρήσιμο για την αξιολόγηση της ακρίβειας του δέντρου (καθώς π.χ. ένα φύλλο με ακρίβεια κοντά στο 50% δε μειώνει τη συνολική ποιότητα του δέντρου αν έχει αμελητέο ποσοστό παρατηρήσεων να καταλήγουν σε αυτό).

Σκοπός, λοιπόν, είναι να βρούμε ορισμένα μονοπάτια ερωτήσεων που να κατασκευάζουν φύλλα με μεγάλη ομοιογένεια. Για να γίνει αυτό χρειαζόμαστε μια

[4] Συνήθως 'Ναι' ή 'Όχι' – διαφορετικά θεωρούμε πάντοτε πεπερασμένο πλήθος επιλογών, οπότε και πάλι μπορεί να γίνει μετατροπή σε δυαδικό δέντρο κατά τα γνωστά.

επιλογή κατάλληλης φύσης ερωτημάτων (αντίστοιχα παραμέτρους συστοιχιών) και κατάλληλη ακολουθία της τοποθέτησης τους, έτσι ώστε να ελαχιστοποιείται το μέγεθος του δέντρου ενώ παράλληλα να μεγιστοποιείται το ποσοστό επιτυχίας των φύλλων του. Η κατασκευή ενός τέτοιου δέντρου υπάγεται σε έναν τομέα αλγορίθμων, γνωστό ως εκμάθηση δέντρων απόφασης.

Το πλέον γνωστό μοτίβο εκμάθησης είναι το recursive partitioning, ένα γενικό πλαίσιο άπληστων αλγορίθμων όπου αντλώντας πληροφορία από τις δοθείσες εγγραφές-δυάδες $(\boldsymbol{X}, Y)$, θέτουμε σε κάθε κόμβο μια ερώτηση πάνω στις ανεξάρτητες μεταβλητές και ανοίγουμε νέους κόμβους-παιδιά μέχρις ότου να καταλήξουμε σε κάποιον κόμβο που δεν αναλύεται περαιτέρω.

Η μεγάλη ζήτηση για εξαγωγή τέτοιων δέντρων έχει οδηγήσει στην κατασκευή πληθώρας σχετικών αλγορίθμων. Η ανάλυση Classification and Regression Tree (CART) είναι μία τεχνική που αφορά αμφότερα τα δέντρα κατάταξης και παλινδρόμησης και που εισήχθη από τον Breiman κ.α. στο (Breiman, 1984). Η τεχνική αυτή είναι αναδρομική, και ανήκει στο παραπάνω πλαίσιο αλγορίθμων, δηλαδή εφαρμόζεται σε κάθε κόμβο-παιδί του δέντρου μέχρις ότου να πληροίται κάποια συνθήκη τερματισμού. Συγκεκριμένα σε κάθε κόμβο, η τεχνική αυτή βρίσκει τη καταλληλότερη τιμή αλλά και σημείο τομής της (έτσι ώστε να πετυχαίνει τη καλύτερη δυνατή ομοιομορφία στα παιδιά που δημιουργούνται με τρόπο που θα δούμε, όμως, πιο αναλυτικά στην ενότητα που ακολουθεί) και η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να φτάσουμε σε κάποιο κόμβο που δε μπορεί να αναλυθεί περαιτέρω (είτε είναι όλος ομοιογενής ή έχουν προσδιοριστεί όλες οι δυνατές ανεξάρτητες μεταβλητές). Αφότου συμβεί αυτό εξαντλητικά, μετά εφαρμόζεται κλάδεμα (pruning) στο παραγόμενο δέντρο έτσι ώστε να ομαδοποιηθούν φύλλα κοινών προγόνων που καταλήγουν στην ίδια απόφαση και έτσι να μειωθεί το μέγεθος του δέντρου.

Ασφαλώς για όλα τα παραπάνω χρειαζόμαστε ένα κατάλληλο πρόγραμμα υλοποίησης (δοθέντος και του μεγάλου όγκου δεδομένων) κι η γλώσσα που επιλέχθηκε για αυτό το σκοπό ήταν η R (βλ. και (R Team, 2011)). Η R είναι ένα ανοιχτού κώδικα προγραμματιστικό περιβάλλον, που υποστηρίζει πληθώρα paradigms προγραμματισμού όπως προστακτικό, αντικειμενοστραφή, συναρτησιακό και διαδικασιακό προγραμματισμό ενώ επίσης παρέχει μεταξύ των άλλων και δυνατότητα επεξεργασίας της ροής του προγράμματος κατά την ώρα της εκτέλεσης. Η R είναι μια κατάλληλη κι ευρέως χρησιμοποιούμενη γλώσσα για στατιστικούς υπολογισμούς καθώς και την εξαγωγή σχετικών γραφημάτων. Συγκεκριμένα, για την παραγωγή των αποτελεσμάτων της παρούσας εργασίας, χρησιμοποιήθηκε κατά κύριο λόγο το πακέτο rpart της R, το οποίο μεταξύ άλλων παρέχει σχεδόν όλα τα εργαλεία της τεχνικής CART για την παραγωγή των σχετικών δομών (είναι μάλιστα βασισμένη στο ομώνυμο βιβλίο των (Breiman, 1984) όπως φαίνεται κι από το σχετικό εγχειρίδιο προδιαγραφών της (Therneau, Atkinson, & Ripley, 2017), όπου μπορούν μάλιστα να βρεθούν και εκτενέστερες λεπτομέρειες για τη λειτουργικότητα των συναρτήσεων που χρησιμοποιήθηκαν).

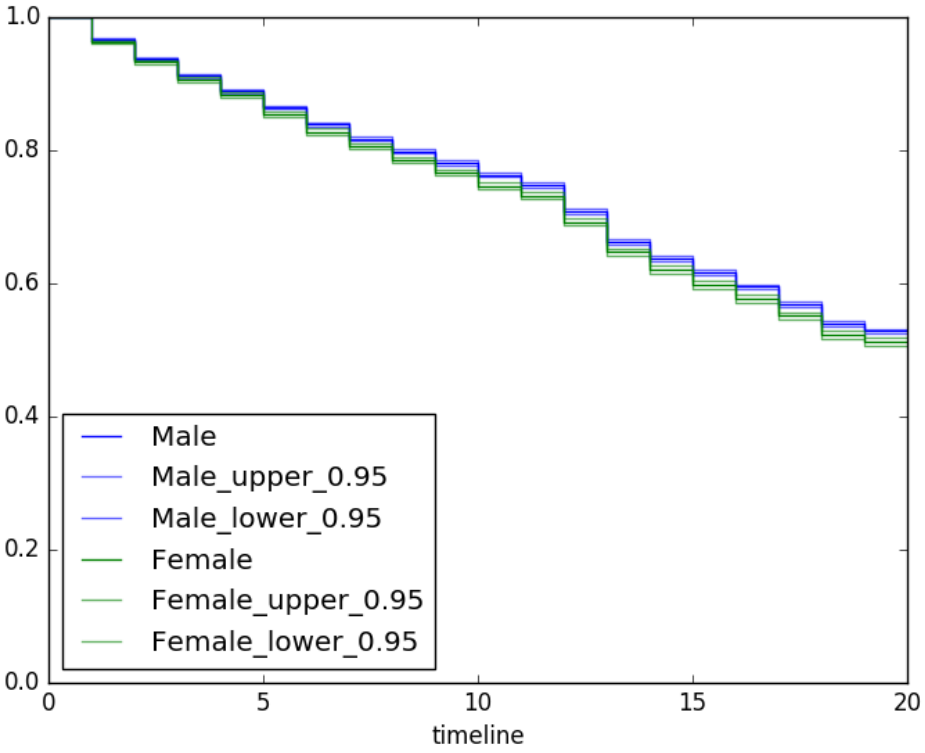
4.2.2 Αξιολόγηση συσχέτισης δημογραφικών στοιχείων με χρήση των πακέτων survival και rpart

Όπως καταλήξαμε στο προηγούμενο κεφάλαιο, θα πρέπει να χρησιμοποιήσουμε έναν αλγόριθμο που αξιοποιεί πληροφορία μεγαλύτερου εύρους, ενώ παράλληλα επικεντρώνεται στις πλέον χρήσιμες εγγραφές. Ασφαλώς δε γνωρίζουμε την ακριβή σχέση μεταξύ των εκάστοτε εγγραφών και του προβλεπτέου (δεδομένου ότι κατά βάσει αυτήν ακριβώς τη σχέση ψάχνουμε σε πρώτο πλάνο), αλλά ακόμη περισσότερο δεν υπάρχει κανένα εύλογο επιχείρημα που να υποστηρίζει μια απλή γραμμική σχέση (χωρίς αυτό να σημαίνει βέβαια ότι αναγκαστικά δεν υπάρχει), επομένως θα πρέπει να χρησιμοποιήσουμε ένα πιο γενικευμένο μοντέλο. Σε κάθε περίπτωση, μια πιο ασφαλής επιλογή, ώστε να αποφύγουμε ακόμη μια ακατάλληλη μοντελοποίηση, είναι να πάρουμε τη γενικότερη περίπτωση συνάρτησης, η οποία όπως είπαμε μπορεί να αναπαρασταθεί από ένα δέντρο αποφάσεων (αφότου ασφαλώς χωρίσουμε το διάστημα τιμών σε κατάλληλες συστάδες).

Το επόμενο βήμα είναι η επιλογή των κατάλληλων μόνο εγγραφών για τον αλγόριθμο πρόβλεψης. Να παρατηρήσουμε σε αυτό το σημείο, ότι ένα κομμάτι από τα δεδομένα που έχουμε διαθέσιμα και δε χρησιμοποιήσαμε μέχρι στιγμής είναι τα δημογραφικά στοιχεία. Γενικά, για λόγους που θα αναλύσουμε περαιτέρω στη συνέχεια, δεν αναμένουμε από μόνα τους να μπορούν να προσφέρουν κάποια κρίσιμη πληροφορία, εν τούτοις ασφαλώς αξίζει να μελετηθούν τόσο για τον πιθανό συμπληρωματικό τους ρόλο, όσο και επειδή αποτελούν μια εύπεπτη εισαγωγή στη χρήση των δέντρων απόφασης ως μοντέλο παλινδρόμησης.

Χρησιμοποιώντας λοιπόν τις κατάλληλες συναρτήσεις, ψάξαμε αν υπάρχει κάποια από τις μεταβλητές των δημογραφικών που να παρουσιάζει σχετική εξάρτηση με τον ερχομό churn month. Ένας εύγλωττος τρόπος να το ελέγξουμε αυτό (πέραν των παραμέτρων του δέντρου απόφασης, έτσι ώστε να φαίνεται και η χρονική πορεία) είναι τα γραφήματα εκτίμησης Kaplan-Meier. Στα γραφήματα αυτά παρουσιάζεται, όπως είπαμε η πιθανότητα επιβίωσης (εν προκειμένω μη ερχομού churn month για το πρόγραμμα ‘Affluent’) κατά τη χρονική διάρκεια εξέλιξης του συστήματος. Για την παραγωγή αυτών των γραφημάτων χρησιμοποιήθηκε το πακέτο survival της γλώσσας R (Therneau & Lumley, 2017).

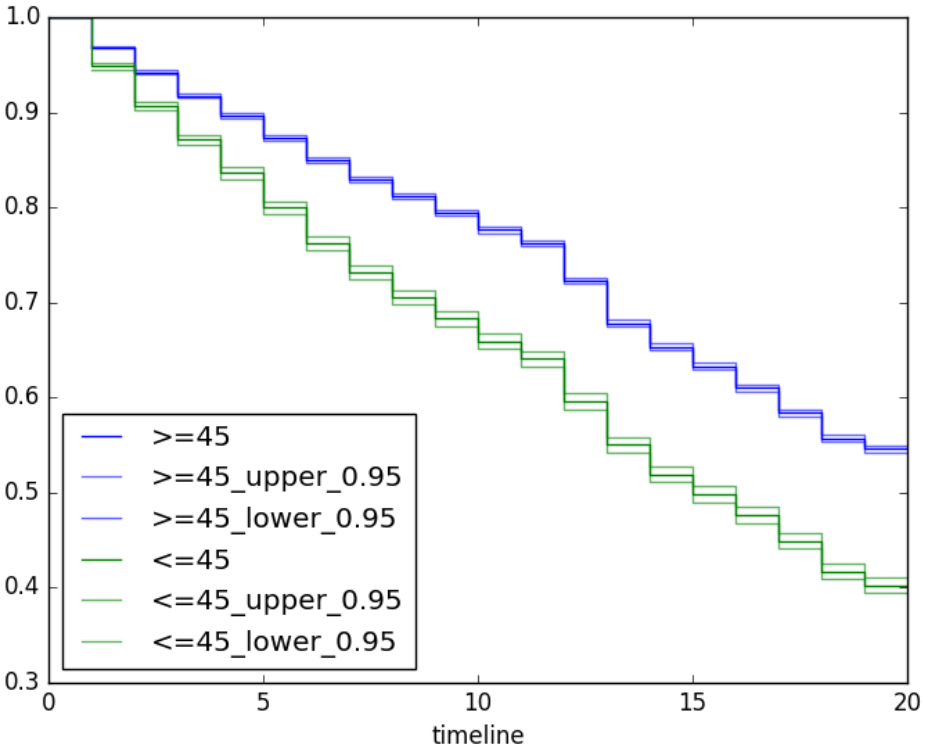
Ένα δημογραφικό στοιχείο, αναμένουμε να παρουσιάζει συσχέτιση με την επιβίωση των πελατών στο πρόγραμμα, όταν το αντίστοιχο γράφημα της ακολουθεί αισθητά διαφορετική πορεία από το καθολικό γράφημα που είδαμε στο κεφάλαιο της παρουσίασης του προβλήματος και αφορούσε την πορεία της αντίστοιχης μετρικής για το σύνολο των παρατηρήσεων μας. Δυστυχώς, όμως, τα περισσότερα δημογραφικά στοιχεία δε φαινόντουσαν ικανά να μπορούν να παρέχουν χρήσιμα στοιχεία, όπως για παράδειγμα φαίνεται χαρακτηριστικά στο γράφημα που ακολουθεί για το Sex των πελατών:



Εικόνα 9: Πιθανότητα επιβίωσης μεταξύ γυναικών και αντρών

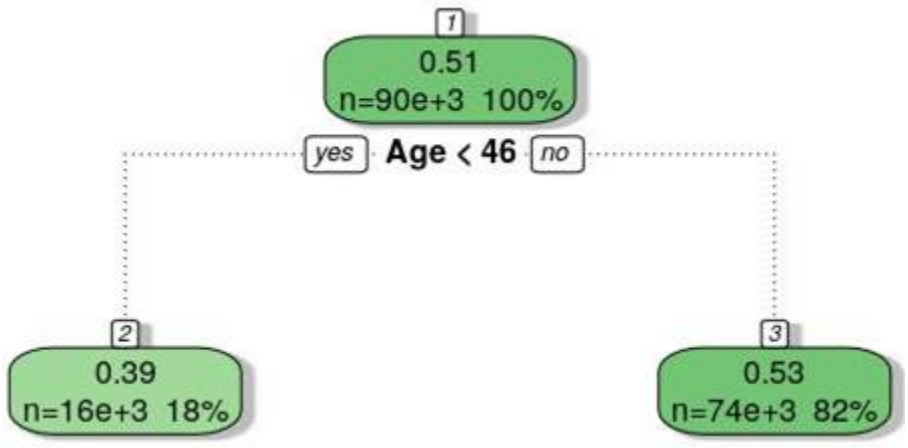
Όπως βλέπουμε, η πορεία που ακολουθούν είναι η ίδια με αυτή της γενικής περίπτωσης με αμελητέα απόκλιση προς το τέλος της περιόδου, που σημαίνει ότι δεν υπάρχει κάποια εξάρτηση από το φύλο. Ασφαλώς στο ίδιο συμπέρασμα μπορούμε να καταλήξουμε και μόνο από την παρατήρηση ότι οι δύο κεντρικές γραμμές του παραπάνω γραφήματος ακολουθούν επίσης ταυτόσημη πορεία, κάτι που κάνει προφανή την ανεξαρτησία τους από το ποσοστό επιβίωσης. Αντίστοιχα γραφήματα είχαμε και στα περισσότερα από τα υπόλοιπα δημογραφικά πεδία, όπου η ταύτιση των γραμμών κάνει ξεκάθαρη την παρόμοια συμπεριφορά τους ως προς τα ποσοστά αποχώρησης από το πρόγραμμα (και συνεπώς την ανεξαρτησία τους από αυτό).

Περισσότερο πετυχημένη, ωστόσο, φάνηκε να είναι η περίπτωση του Age, όπου όπως φαίνεται και στο σχετικό γράφημα υπάρχει ξεκάθαρη απόκλιση και αυξημένη πιθανότητα επιβίωσης των πελατών άνω των 45 ετών (αν και από ό,τι φαίνεται η απόκλιση αυτή δεν αποτελεί επαρκές στοιχείο καθώς για να μπορέσουμε να τη χρησιμοποιήσουμε ως κριτήριο, θα πρέπει να υπάρχει και κατάλληλη πληθυσμιακή κατανομή μεταξύ των δύο ομάδων):



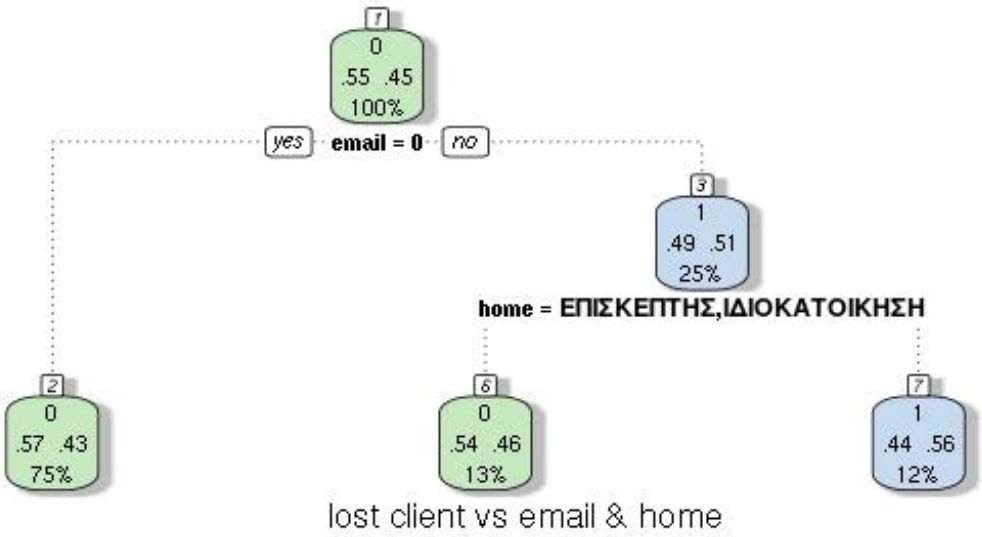
Εικόνα 10: Πιθανότητα επιβίωσης μεταξύ πελατών άνω και κάτω των 45 ετών

Ασφαλώς η διαφοροποίηση δεν είναι αρκετή ώστε να μπορεί να χαρακτηρίσει ικανοποιητικά τις δύο κλάσεις και αυτό φαίνεται τόσο από τις τιμές στον κάθετο άξονα του παραπάνω γραφήματος, όσο και στο συνοπτικό δέντρο παρακάτω, που εξάγεται από αυτό, και στο οποίο φαίνεται ότι οι πιθανότητες επιτυχίας των δύο τελικών φύλλων είναι κοντά στο 0.4 και 0.5 αντίστοιχα (με μεγαλύτερη πληθυσμιακή εμφάνιση μάλιστα στη χειρότερη τιμή >45), τιμές που επ' ουδενί δε μπορούν να χαρακτηριστούν επαρκείς για τον υπολογισμό του προβλεπτέου.



Εικόνα 11: Δέντρο Αποφάσεων με κριτήριο την Ηλικία των Πελατών

Αντιπροσωπευτικό παράδειγμα της αδυναμίας των υπόλοιπων δημογραφικών στοιχείων να δώσουν οποιαδήποτε ουσιώδη πληροφορία είναι το ακόλουθο δέντρο όπου φαίνεται ότι όλα τα σχετικά ποσοστά κυμαίνονται στο 50%.



Εικόνα 12:Ανεπαρκές δέντρο απόφασης για τα Δημογραφικά.

4.3 Περιληπτικές Μετρικές

Ας κάνουμε μία σύνοψη των συμπερασμάτων που έχουμε καταλήξει μέχρι στιγμής. Από τη μελέτη της μαρκοβιανής προσέγγισης, καταλήξαμε ότι χρειαζόμαστε ένα μοντέλο που περιλαμβάνει όλο το δοθέν μήκος πληροφοριών για όσο ένας πελάτης βρίσκεται εντός του προγράμματος (και όχι ένα σταθερό παρελθοντικό κομμάτι κάθε φορά), καθώς αλλιώς υπεισέρχονται μεγάλες αποκλίσεις λόγω τυχαίου θορύβου στο υποβόσκον μοτίβο και έτσι δεν είναι δυνατόν να το εξάγουμε. Επιπλέον έγινε σαφές ότι θα πρέπει να περιοριστούμε μόνο σε όσες μεταβλητές έχουν ουσιώδη πληροφορία καθώς η εισαγωγή πολλών που έχουν ανεξαρτησία από το προβλεπτέο, απλώς μειώνουν τη συνεισφορά των πράγματι συσχετιζόμενων. Επιπλέον από την προηγούμενη ενότητα έγινε σαφές, ότι τα δημογραφικά στοιχεία δεν έχουν σοβαρές πιθανότητες να βρίσκονται εντός αυτών των μεταβλητών που πρέπει να ξεχωρίσουμε και έτσι μπορούμε να τις αποκλείσουμε από το σύνολο των πιθανών.

Μία προσέγγιση που φαίνεται, επίσης, να λαμβάνει περισσότερο υπόψιν όλο το εύρος των δεδομένων είναι όπως είδαμε η ανάλυση παλινδρόμησης στην οποία αναζητείται συνήθως η συσχέτιση μεταξύ συνεχών μεταβλητών. Το πρόβλημα με τη περίπτωση μας είναι ότι οι πρωτογενείς ανεξάρτητες μεταβλητές, δηλαδή οι χρονοσειρές, είναι περισσότερων διαστάσεων από μια απλή συνεχή μεταβλητή, αφού πρόκειται για συνεχείς συναρτήσεις. Θα πρέπει λοιπόν με κάποιο τρόπο να συμπυκνώσουμε τη πληροφορία τους σε μία (ή περισσότερες) συνεχείς μεταβλητές (π.χ. η μέση τιμή τους, η διάμεσος μαζί με τα ολικά μέγιστα κ.ο.κ.). Ασφαλώς με κάθε τέτοιο μέτρο, η περισσότερη πληροφορία από μία χρονοσειρά χάνεται και θα πρέπει

να φροντίσουμε οι περιγραφικές μετρικές που θα επιλέξουμε να είναι τέτοιες ώστε να εμφωλεύουν σε ικανοποιητικό βαθμό τα κρίσιμα (για το σκοπό μας) χαρακτηριστικά των χρονοσειρών.

4.3.1 Αποσύνθεση χρονοσειρών στα δομικά τους μέρη

Οικονομική Προσέγγιση

Στις οικονομικές επιστήμες, η κατηγοριοποίηση των χρονοσειρών βάσει των χαρακτηριστικών τους είναι μία από τις πλέον κεντρικές μεθόδους για την πρόβλεψη της εξέλιξης τους. Για το σκοπό αυτό, έχουν ορισθεί τρεις βασικές μετρικές που είναι από τις πλέον χρησιμοποιούμενες σε τέτοιες εφαρμογές, καθώς όπως θα δούμε έχουν άμεση φυσική ερμηνεία και επειδή έχει προκύψει ότι όντως εμφανίζονται εντός των χαρακτηριστικών ενός πολύ μεγάλου πλήθους χρονοσειρών που πηγάζουν από πραγματικά στοιχεία (βλ. και (Koopman & Lee, 2009)). Τα χαρακτηριστικά αυτά είναι:

- **Τάση**: Πρόκειται για τη γενική πορεία που ακολουθεί η χρονοσειρά σε βάθος χρόνου. Μπορεί να μην είναι μονότονη και να ακολουθεί αυξομειώσεις, οι οποίες όμως πρέπει να γίνονται σε βάθος ετών και όπου κάθε τοπική μονοτονία να διαρκεί επίσης για ένα μεγάλο χρονικό διάστημα.

- **Εποχικότητα**: Πρόκειται για μοτίβα-αυξομειώσεις που επαναλαμβάνονται με σταθερή περίοδο (π.χ. ανά μήνα, τρίμηνο κ.ο.κ.) και που εξαρτώνται από γνωστές ειδικές περιόδους που εμφανίζονται ανά τακτά χρονικά διαστήματα.

- **Κυκλικότητα**: Πρόκειται για μοτίβα που αλλοιώνουν τη μονοτονία της συνάρτησης και εμφανίζονται με μη σταθερή περίοδο. Εξαρτώνται από την εμφάνιση ειδικών (συνήθως απρόβλεπτων) οικονομικών περιόδων και απέχουν μεγάλα χρονικά διαστήματα μεταξύ τους (π.χ. κάποια έτη). Γενικά έχουν μεγαλύτερη διάρκεια και μεγέθη παραμέτρων από ό,τι η εποχικότητα.

Προσπαθούμε, λοιπόν, να κατηγοριοποιήσουμε κάθε χρονοσειρά βάσει των τιμών που έχει για κάθε ένα από τα παραπάνω χαρακτηριστικά. Έτσι, σκοπός είναι, βάσει αυτών των παραμέτρων να δημιουργηθούν συστάδες, οι οποίες θα εμφανίζουν ομοιόμορφη συμπεριφορά και άρα να μπορεί να προβλεφθεί η πορεία τυχόντων νέων χρονοσειρών που προκύπτει ότι κατηγοριοποιούνται στις ίδιες συστάδες.

Για να συμβεί αυτό, όμως, πρέπει πρώτα να ορίσουμε αριθμητικές μεταβλητές που να χαρακτηρίζουν ορθά και αποδοτικά τα παραπάνω μεγέθη. Για αρχή, να σημειώσουμε ότι επειδή έχουμε μια σχετικά μικρή περίοδο μόλις 20 μηνών (λιγότερο από 2 χρόνια), μπορούμε να αγνοήσουμε την κυκλικότητα και να αφομοιώσουμε τυχόν επιρροές στην γενική τάση. Χρειαζόμαστε λοιπόν μετρικές που να χαρακτηρίζουν την *τάση* και την *εποχικότητα* της συνάρτησης. Διαισθητικά, για ένα τέτοιο χρονικό διάστημα, όπου υπάρχει σχετική σταθερότητα, όπως έχουμε δει στα

αντίστοιχα γραφήματα, η τάση μπορεί να περιγραφθεί από τη κεντρική τιμή μιας χρονοσειράς (μέση τιμή), ενώ για την εποχικότητα (δεδομένου ότι δε μπορεί να περιγραφθεί από τη μέση τιμή εξαιτίας της περιοδικής της φύσης (με κέντρο 0)) χρειάζεται ένα ακόμη μέγεθος που να περιγράφει τη διακύμανση της χρονοσειράς, το οποίο δεν είναι άλλο από τη συνώνυμη μετρική.

Όλα τα παραπάνω θεμελιώνονται και πιο αυστηρά με τη μαθηματική προσέγγιση που ακολουθεί.

Μαθηματική Προσέγγιση

Υπό κάποιες συγκεκριμένες συνθήκες, τις οποίες (τόσο για χάριν απλότητας όσο και επειδή μπορούμε να υποθέσουμε όποια επέκταση θέλουμε εκτός των 20 μηνών κι άρα δε προκύπτουν ζητήματα περιοδικότητας) θα τις θεωρήσουμε ότι ισχύουν για τις χρονοσειρές μας, μπορούμε να αναλύσουμε μια συνάρτηση σε ένα άθροισμα Fourier[5]:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \sin(n\omega t)$$

Είναι εύκολο να δούμε ότι σύμφωνα με όσα είπαμε προηγουμένως, η τάση αντιστοιχεί στο σταθερό μη περιοδικό όρο a_0 και η εποχικότητα στη κυρίαρχη περιοδικότητα της συνάρτησης, δηλαδή εν προκειμένω στον όρο $\alpha_1 \sin(\omega t)$. Οι επόμενοι όροι θεωρούμε ότι περιέχουν λιγότερη πληροφορία και επομένως προσεγγίζουμε τη συνάρτηση από την

$$\hat{f}(t) = a_0 + a_1 \sin(\omega t)$$

και σκοπός μας, για να καταφέρουμε να χαρακτηρίσουμε τη συνάρτηση, είναι να εντοπίσουμε τις άγνωστες παραμέτρους a_0 και a_1.

Ως μέση τιμή μιας συνάρτησης $\hat{f}$ σε ένα διάστημα $[-T, T]$ ορίζεται το

$$\mu_T = \frac{\int_{-T}^{T} \hat{f}(t)dt}{2T}$$

από το οποίο εύκολα υπολογίζουμε ότι στο εύρος μιας περιόδου (ή όταν $T \to +\infty$) έχουμε

$$\mu = \alpha_0$$

καθώς ο όρος $\sin(\omega t)$ (όπως άλλωστε και οι ψηλότεροι) δίνει προφανώς μηδενικό ολοκλήρωμα.

Όμοια έχουμε ότι η διακύμανση μιας μεταβλητής X ορίζεται ως

$$Var[X] = E[(X - \mu_X)^2]$$

[5] Στη γενική έκφραση υπάρχουν και όροι $\cos(n\omega t)$, όμως θεωρούμε ότι προσαρμόζουμε το εύρος της δειγματοληψίας, έτσι ώστε να απαλείφονται αυτοί οι όροι, βλ. και (Bloomfield, 2004).

και επομένως ορίζουμε αντίστοιχα για τη συνάρτηση μας

$$\sigma^2 = \frac{\int_{-T}^{T}\left(\hat{f}(t) - \alpha_0\right)^2 dt}{2T}$$

από το οποίο προκύπτει ότι είναι

$$\sigma^2 = \left(\frac{a_1}{\sqrt{2}}\right)^2$$

με χρήση στοιχειωδών πράξεων.

Εν τέλει συνδυάζοντας τα παραπάνω, καταλήγουμε ότι η προσέγγιση μας μπορεί να γραφτεί ισοδύναμα ως

$$\hat{f}(t) = \mu + \sqrt{2}\sigma\sin(\omega t)$$

Από την παραπάνω έκφραση γίνεται σαφές ότι η μέση τιμή και η διασπορά της συνάρτησης είναι επαρκείς για τον πλήρη χαρακτηρισμό της (βάσει αυτής της προσέγγισης). Πιο λεπτομερή ανάλυση μπορούμε να βρούμε και στα (Anderson & W., 1950) και (Koopman & Lee, 2009).

Έχουμε βρει, λοιπόν, ποιες είναι οι μετρικές (moments) που πρέπει να κρατάμε και χρησιμοποιούμε για κάθε συνάρτηση ως κριτήρια για την κατηγοριοποίηση τους. Μένει να διευθετήσουμε το ζήτημα της επιλογής των σχετικών μόνο μεταβλητών. Όπως είπαμε, ένας από τους λόγους που απέτυχαν προηγούμενες προσεγγίσεις ήταν η συμπερίληψη μεταβλητών που ήταν ανεξάρτητες από τον ερχομό churn month και επομένως απλώς προσέθεταν θόρυβο υποβαθμίζοντας τη συνεισφορά των όντως συσχετιζόμενων.

Μπορούμε, ωστόσο, με σχετική ασφάλεια να χαρακτηρίσουμε ορισμένες από τις οικονομικές εγγραφές ως σίγουρα συσχετιζόμενες με τη πιθανότητα επιβίωσης εντός του προγράμματος και αυτές δεν είναι άλλες, ασφαλώς, από τα βασικά πεδία Immediate, Closed και Sum. Η φύση αυτών των πεδίων είναι τέτοια έτσι ώστε αναπόφευκτα να συνδέονται άρρηκτα με την πορεία της χρονοσειράς ως προς την εμφάνιση churn month (εφόσον τα πεδία αυτά εκφράζουν άμεσα το χρηματικό επίπεδο στο οποίο εμπιστεύεται ο πελάτης την τράπεζα). Βεβαίως, αυτό δεν αποκλείει την ύπαρξη συσχέτισης και κάποιων άλλων φαινομενικά ανεξάρτητων μεταβλητών με το ζητούμενο (π.χ. του Housing ή του Investment). Εν τούτοις, αν εφαρμόσουμε το ξυράφι του Occam και εμείνουμε μόνο στις τρεις αυτές ασφαλείς μεταβλητές, έχουμε με μεγάλη σιγουριά αποφύγει τον κίνδυνο εισαγωγής ακατάλληλων δεδομένων – έστω με το αντάλλαγμα της πιθανότητας να μην έχουμε απορροφήσει όλη τη πιθανή πληροφορία (θα προτιμήσουμε, δηλαδή, μία έστω περικομμένη ορθή πληροφορία παρά την τυχούσα εισαγωγή παραπλανητικών ασυσχέτιστων δεδομένων).

4.3.2 Δέντρο Απόφασης με Περιληπτικές Μετρικές

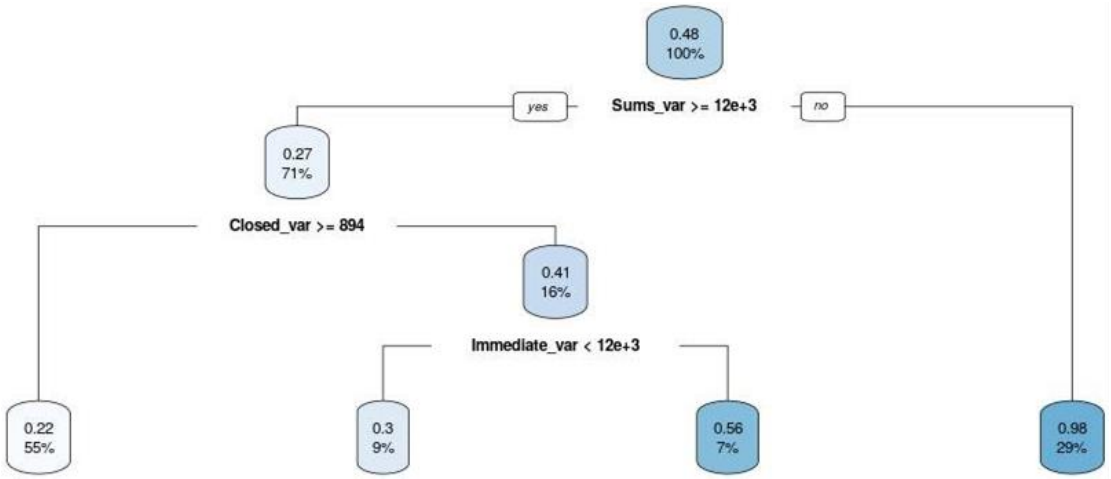
Καταλήξαμε, λοιπόν, ότι θα πρέπει να παίρνουμε πληροφορία μόνο από τα πεδία Immediate, Closed και Sum και ότι από αυτά αρκεί να χρησιμοποιούμε τις μετρικές mean και variance. Δεδομένου ότι ο μέσος όρος εκφράζει τη γενική τάση μιας χρονοσειράς (η οποία παραμένει σε σχετικά σταθερά επίπεδα), την ουσιώδη πληροφορία περιέχει το variance των μεταβλητών, το οποίο εκφράζει την αστάθεια και ευμεταβλητότητα των εγγραφών και από τις οποίες μπορούμε να εξάγουμε την αυξημένη ή μη πιθανότητα της χρονοσειράς να εξέλθει από το πρόγραμμα – πράγματι, άλλωστε, όπως θα δούμε και στη συνέχεια από το αντίστοιχο δέντρο, αυξημένη variance στα οικονομικά οδηγεί σε μεγαλύτερη πιθανότητα εμφάνισης churn month για κάποιον πελάτη.

Εύλογα, λοιπόν, επιλέχθηκε βάσει των παραπάνω η χρήση μόνο των Immediate_var, Sums_var, Closed_var και Sums_mean (με τη προφανή σημασιολογία) για τη κατασκευή του αντίστοιχου δέντρου. Σε αυτό το σημείο να σχολιάσουμε ότι για να γίνει αυτό προκύπτουν πρώτα ορισμένα ζητήματα συλλογής, επεξεργασίας και αρχειοθέτησης των δεδομένων. Γενικά, ένα πολύ κρίσιμο τεχνικό κομμάτι τόσο για την παρουσίαση όσο και για τη διεπαφή μεταξύ των σταδίων επεξεργασίας ενός μεγάλου συνόλου δεδομένων είναι η τακτοποίηση τους σε κατάλληλη μορφή. Στο (Wickham, 2014) παρουσιάζονται οι βασικές αρχές που πρέπει να ακολουθούνται προς αυτό το σκοπό, τις οποίες και ακολουθήσαμε κατά την παραγωγή του αντίστοιχου αλγορίθμου που παρατίθεται στο Παράρτημα σε γλώσσα R. Ακολουθεί στιγμιότυπο των σχετικών αποτελεσμάτων:

```
> summary_econ
# A tibble: 87,070 × 7
      Id Immediate_var    Sums_var Closed_var Sums_mean Months  Lost
   <int>         <dbl>       <dbl>      <dbl>     <dbl>  <int> <dbl>
1      1  9.344054e+03  9344.05425   26832.82  67307.14      5     1
2      2  4.868388e+03  8656.45778       0.00 153551.83      2     1
3      3  1.068055e+04 28696.22962   31241.37  94943.54     19     1
4      4  8.587174e+03 44113.23676   40104.03 107841.38      3     1
5      5  5.291195e+03  6971.11296   69282.03  57942.59      3     1
6      6            NA          NA         NA 105000.00      1     1
7      7  4.706220e+04 72556.82079  134703.84 151656.65      2     1
8      8  6.963639e+03  4099.46512   20412.41  63994.94      6     1
9      9  9.192388e-02    20.90915       0.00 434549.09      2     1
10    10  1.453005e+03  5406.26268   38890.87  54413.71      2     1
# ... with 87,060 more rows
```

Εικόνα 13: Περιληπτικές μετρικές ανά Πελάτη

Ακολουθεί το τελικό δέντρο που κατασκευάστηκε με τη παραπάνω μέθοδο παλινδρόμησης, κάνοντας χρήση του πακέτου rpart της R, όπου φαίνονται τα σημεία τομής που επιλέχθηκαν για κάθε μεταβλητή μαζί με τα ποσοστά πληθυσμού και ακριβείας κάθε φύλλου:



Εικόνα 14: Τελικό Δέντρο Απόφασης για τις Περιληπτικές Μετρικές.

Το ποσοστό επιτυχίας του παραπάνω δέντρου αναμένεται, όπως φαίνεται, να είναι της τάξης του 80% το οποίο αποτελεί το πρώτο ιδιαίτερα αξιόλογο αποτέλεσμα από τα μοντέλα πρόβλεψης που κατασκευάσαμε μέχρι στιγμής και αποτελεί μια πρώτη επιβεβαίωση της καταλληλόλητας των συγκεκριμένων περιληπτικών μετρικών που επιλέχθηκαν.

4.3.3 Κατασκευή και Αξιολόγηση του Γενικευμένου Γραμμικού Μοντέλου με Περιληπτικές Μετρικές

Εν συνεχεία, χρησιμοποιήσαμε την εντολή glm του σχετικού πακέτου (generalized linear model) με όρισμα τη συσχέτιση του Lost από τις παραπάνω μεταβλητές και επιλογή διωνυμικού (binomial) μοντέλου για την κατασκευή ενός γενικευμένου γραμμικού μοντέλου. Η συγκεκριμένη επιλογή είναι η πιο συνήθης όταν η ανεξάρτητη μεταβλητή μπορεί να πάρει μία από δύο δυνατές τιμές, έκαστη με κάποια άγνωστη πιθανότητα προς προσδιορισμό. Συγκεκριμένα θεωρούμε ότι το αποτέλεσμα ακολουθεί διωνυμική κατανομή η οποία προκύπτει από τις ανεξάρτητες μεταβλητές ως εξής: Θεωρούμε μια ενδιάμεση μεταβλητή που εξαρτάται κλασικά γραμμικά από αυτές και στη συνέχεια υπάρχει μια συνάρτηση σύνδεσης λογαριθμικής μορφής (link=”logit”) η οποία συνδέει την ενδιάμεση μεταβλητή με την έξοδο (η επιλογή της συγκεκριμένης μορφής γίνεται έτσι ώστε το αποτέλεσμα να κυμαίνεται στο διάστημα $[0,1]$ και να μπορεί να μοντελοποιεί ορθά μια διωνυμική κατανομή – σε αντίθεση π.χ. με το κλασικό γραμμικό μοντέλο).

```
glm(formula = Lost ~ Immediate_var + Sums_var + Closed_var +
    Sums_mean, family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4555  -1.0343  -0.3371   0.9203   8.4904

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.373e-01  5.949e-02   7.351 1.96e-13 ***
```

```
Immediate_var  1.315e-05  1.840e-06   7.147 8.90e-13 ***
Sums_var      -6.307e-05  3.230e-06 -19.525  < 2e-16 ***
Closed_var    -9.695e-06  1.538e-06  -6.305 2.89e-10 ***
Sums_mean      8.139e-06  9.041e-07   9.002  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6670.4  on 4837  degrees of freedom
Residual deviance: 5706.9  on 4833  degrees of freedom
  (162 observations deleted due to missingness)
AIC: 5716.9

Number of Fisher Scoring iterations: 6
```

Μαζί με το τελικό αποτέλεσμα υπολογίζονται, όπως φαίνεται και στην παραπάνω έξοδο, και διάφορα στατιστικά δεδομένα που δείχνουν κατά πόσο ταιριάζουν τα δεδομένα, πόσες επαναλήψεις χρειάστηκαν για να συγκλίνει ο αλγόριθμος κ.ο.κ.

Το μόνο που αξίζει να σχολιάσουμε είναι το γεγονός ότι το Intercept φαίνεται να είναι 5 με 6 τάξεις μεγέθους μεγαλύτερο από τις υπόλοιπες μεταβλητές, κάτι που από ό,τι φαίνεται δεν έχει ουσιώδη αντίκτυπο στο σύστημα, καθώς ακόμη και όταν το αφαιρούμε, τα αποτελέσματα συνεχίζουν να βρίσκονται στα ίδια πλαίσια:

```
glm(formula = Lost ~ 0 + Immediate_var + Sums_var + Closed_var
+
   Sums_mean, family = binomial(link = "logit"), data = df)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.2131  -0.9922  -0.4039   1.0269   8.4904

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Immediate_var  1.752e-05  1.462e-06  11.990  < 2e-16 ***
Sums_var      -5.839e-05  2.167e-06 -26.940  < 2e-16 ***
Closed_var    -7.918e-06  1.027e-06  -7.707 1.28e-14 ***
Sums_mean      9.297e-06  4.999e-07  18.599  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’1
```

Στο ίδιο συμπέρασμα μπορούμε να καταλήξουμε μελετώντας τους standardized coefficients του μοντέλου, με χρήση του πακέτου lm.beta της R:

```
Standardized Coefficients::
 (Intercept) Immediate_var      Sums_var    Closed_var     Sums_mean
   0.0000000     0.9385457    -6.5905918    -0.5869024     2.5251501
```

Όπως φαίνεται ο συντελεστής του Intercept είναι προφανώς μηδενικός, όπως αναμενόταν. Μεγάλο ενδιαφέρον παρουσιάζουν όμως και οι τιμές που αποδίδονται στις υπόλοιπες μεταβλητές του μοντέλου. Όπως βλέπουμε κυρίαρχες είναι οι Sums_mean και Sums_var και ακολουθούν πιο χαμηλά οι Immediate_var και Closed_var, κάτι που είναι αναμενόμενο, εφόσον οι πρώτες περιέχουν τη συνολική πληροφορία όσον αφορά το επίπεδο και την μεταβλητότητα των καταθέσεων αντίστοιχα.

Περισσότερο ενδιαφέρον παρουσιάζει η αξιολόγηση των αποτελεσμάτων του αλγορίθμου μέσω των εντολών της μεθόδου evalModel (βλ. Παράρτημα για τις ακριβείς εντολές), οι οποίες εξάγουν μια πληθώρα μετρικών χρήσιμων για την κρίση του συστήματος ως προς την ακρίβεια και ποιότητα της εξόδου του.

Σε τέτοιου είδους συστήματα, όπως φαίνεται και από τον κώδικα, επιλέγουμε ένα τυχαίο κομμάτι των δεδομένων για «προπόνηση» του αλγορίθμου και μετά ο έλεγχος του γίνεται σε ένα άλλο τυχαίο ανεξάρτητο κομμάτι (όπου θεωρούμε ομοιομορφία μεταξύ των δύο). Για τη δημιουργία του τελικού μοντέλου (όπως π.χ. του δέντρου που προηγήθηκε) στη φάση της προπόνησης, ο αλγόριθμος πρέπει να αποφασίσει ποια είναι τα καλύτερα κατώφλια (εν προκειμένω τιμή πιθανότητας της υποτιθέμενης μεταβλητής που ακολουθεί η διωνυμική κατανομή) για την μετάβαση της δυϊκής τιμής της εξόδου (ερχομός ή μη μήνα Lost).

Μετρικές Αξιολόγησης

Στη συνέχεια, θα χρειαστεί να ορίσουμε ορισμένες κλασικές μετρικές που σχετίζονται με αυτές τις επιλογές και την ακρίβεια τους. Όπως έχουμε δει, το μοντέλο θα εξάγει αποφάσεις θετικές ή αρνητικές (αντίστοιχα ερχομός ή μη Churn Month), ορισμένες εκ των οποίων θα είναι ορθές (true positive (TP) και true negative (TN)) και ορισμένες λανθασμένες (false positive (FP) και false negative (FN)). Ορίζονται, λοιπόν, τα εξής εύλογα μέτρα:

Accuracy: Το ποσοστό των ορθών προβλέψεων:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

(Εκφράζει την ακρίβεια του συστήματος ως προς την αληθοτιμή των προβλέψεων του.)

Precision: Το ποσοστό των ορθών θετικών προβλέψεων:

$$\frac{TP}{TP + FP}$$

(Εκφράζει την ακρίβεια του συστήματος ως προς τις αληθείς προβλέψεις.)

Recall: Το ποσοστό των εντοπισμένων θετικών εκβάσεων:

$$\frac{TP}{TP + FN}$$

(Εκφράζει την ακρίβεια του συστήματος ως προς το σύνολο των αληθών γεγονότων.)

Specificity: Το ποσοστό των εντοπισμένων αρνητικών εκβάσεων:

$$\frac{TN}{TN + FP}$$

(Εκφράζει την ακρίβεια του συστήματος ως προς το σύνολο των ψευδών γεγονότων.)

F-measure: Ο αρμονικός μέσος των Precision και Recall

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

(Δίνει μια συνοπτική εικόνα του ισοζυγίου μεταξύ Precision και Recall.)

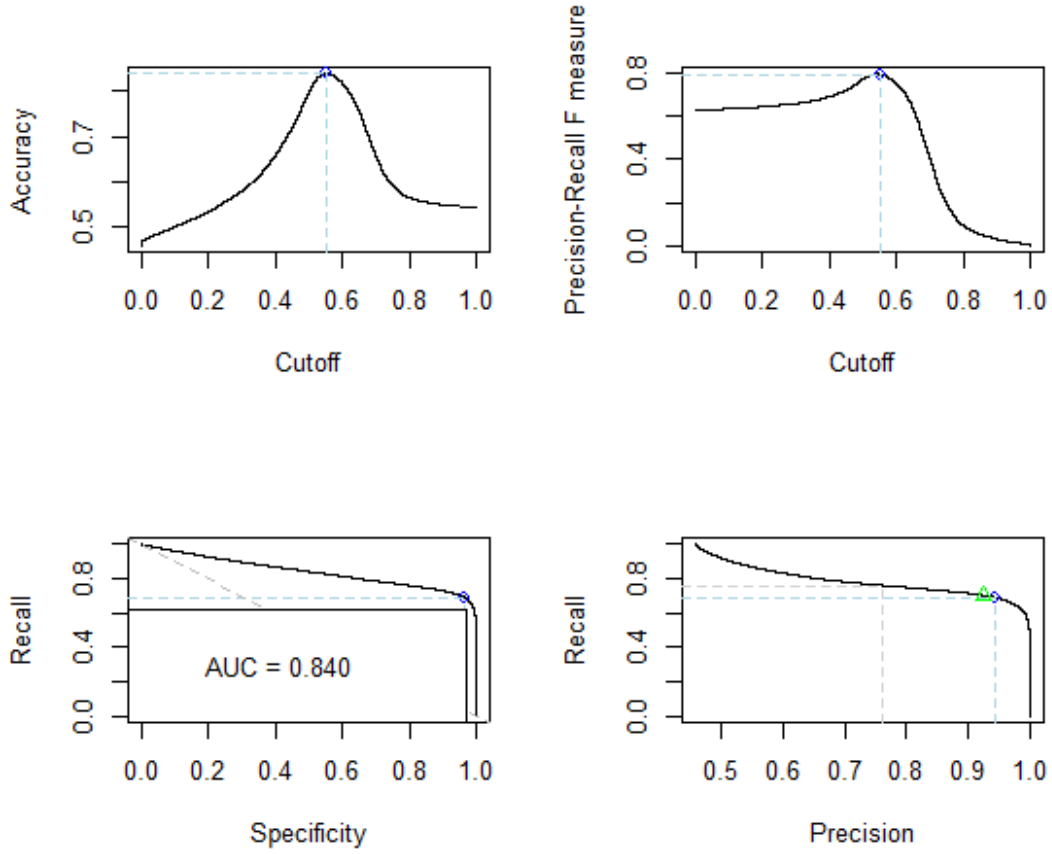
Brier Score: Η μέση τιμή των τετραγώνων της διαφοράς της προβλεπόμενης πιθανότητας και του τελικού αποτελέσματος

$$\frac{1}{N} * \sum_{i=1}^{N} (p_i - o_i)^2$$

(Εκφράζει το βαθμό συμφωνίας πιθανοτήτων πρόβλεψης και τελικών συμβάντων.)

Κάθε αλγόριθμος, λοιπόν, προσπαθεί συνήθως να βρει το σημείο τομής που μεγιστοποιεί κάποια από τις παραπάνω μετρικές (συνήθως το Accuracy). Συγκεκριμένα, στη περίπτωση μας, βλέπουμε δύο περιπτώσεις όπου επιλέγεται σημείο τομής βάσει του καλύτερου Accuracy και βάσει του καλύτερου F-measure. Στα υπόλοιπα γραφήματα, που ακολουθούν, αναπαρίσταται το ισοζύγιο μεταξύ Recall και Specificity (αλλά και Precision) μαζί με τις τιμές που τελικά επιλέχθηκαν.

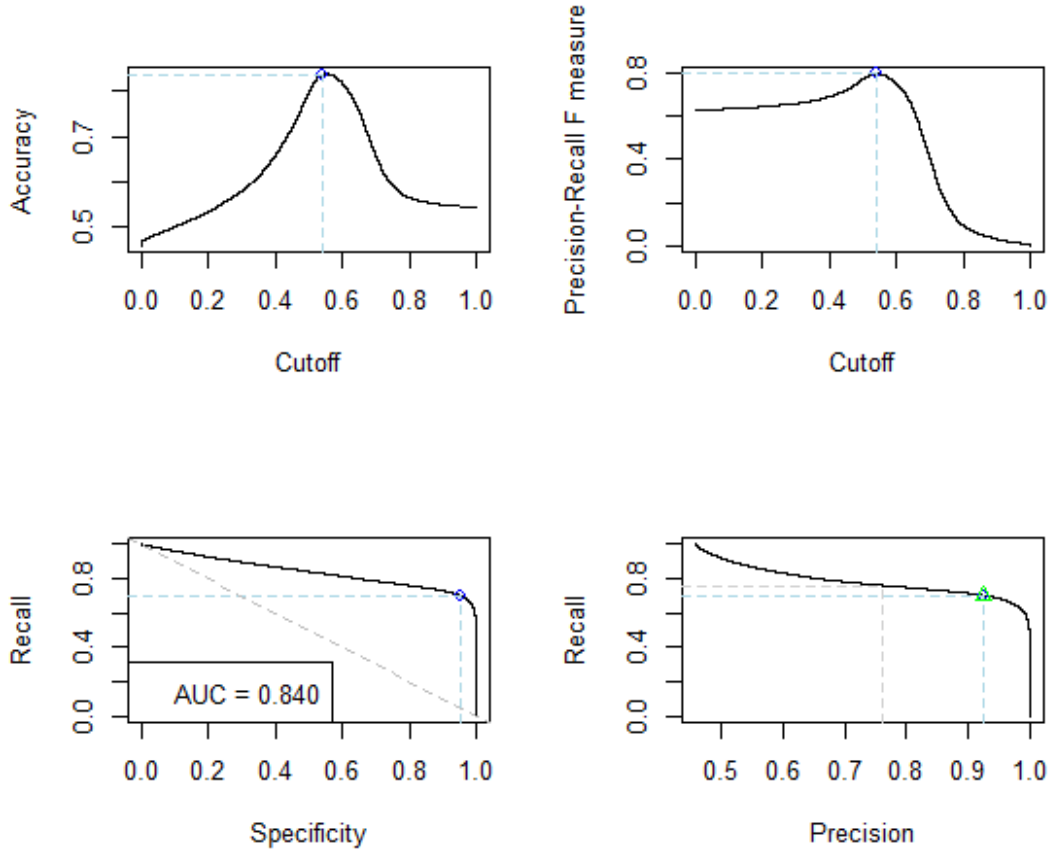
```
AUC = 0.840
Using acc-best cutoff 0.551534
Acc 0.839,   F 0.797, Prec 0.944, Rec 0.689, Spec 0.965, BalAcc 0.670
```



Εικόνα 15: Γραφήματα Αξιολόγησης χρησιμοποιώντας accuracy-best τομή

Το AUC (Area Under Curve) είναι μια εναλλακτική μετρική, η οποία απλώς μετράει το εμβαδό στο αντίστοιχο γράφημα κι έχει το πλεονέκτημα ότι είναι ανεξάρτητη από ακραίες μορφές της κατανομής θετικών-αρνητικών εκβάσεων, σε αντίθεση π.χ. με την Accuracy, η οποία μπορεί να πάρει πολύ μεγάλη τιμή ακόμη κι αν το Precision είναι μηδενικό, απλώς επειδή υπήρχε μεγάλη επικράτεια αρνητικών εκβάσεων. Εν αντιθέσει,, το AUC δίνει μια γενική εικόνα του γραφήματος Recall-Specificity το οποίο εύκολα εξακριβώνουμε ότι σε κάποια τέτοια περίπτωση θα έπαιρνε τιμή κοντά στο 0.5 εκδηλώνοντας ορθά τη χαμηλή ποιότητα του συστήματος (παρότι με υψηλό Accuracy).

```
AUC = 0.840
Using f-best cutoff 0.540576w
Acc 0.837,   F 0.799, Prec 0.923, Rec 0.704, Spec 0.951, BalAcc 0.670
```



Εικόνα 16: Γραφήματα Αξιολόγησης χρησιμοποιώντας F-best τομή

Όσον αφορά το Brier Score, σκοπός της συγκεκριμένης μετρικής είναι κατά βάση η αξιολόγηση του κατά πόσο συμβαδίζουν τα τελικά αποτελέσματα με τις πιθανότητες που προβλέπονται ανεξάρτητα από την ποιότητα των τελευταίων (για παράδειγμα ένα σύστημα που δίνει 50% πιθανότητα σε όλα τα ενδεχόμενα έχει Brier Score 0.25 (σχετικά ικανοποιητικό, εφόσον η μετρική βρίσκεται πάντα στο εύρος $[0,1]$), τη στιγμή που προφανώς δεν παρέχει κάποια ουσιώδη πληροφορία).

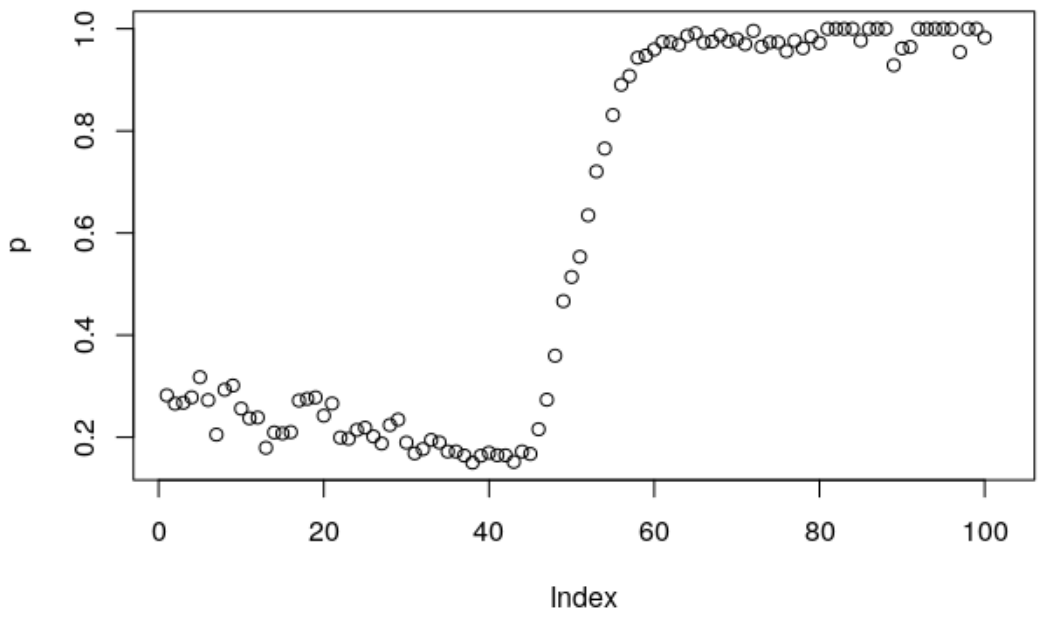
Χρησιμοποιώντας το πακέτο verification της R, παίρνουμε τα παρακάτω αποτελέσματα για την περίπτωση του μοντέλου μας όσον αφορά το Brier Score:

```
$baseline.tf
[1] FALSE
$bs
[1] 0.1820737
$bs.baseline
```

```
[1] 0.2483035
$ss
[1] 0.266729
$bs.reliability
[1] 0.04558172
$bs.resol
[1] 0.1118115
$bs.uncert
[1] 0.2483035
$y.i
[1] 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
$obar.i
[1] 0.2461344 0.2177314 0.2000661 0.1858546 0.1922898 0.4948365 0.9659598 0
.9970696
[9] 0.9941292 1.0000000
$prob.y
[1] 0.082066554 0.059588243 0.078311537 0.131814062 0.223022142 0.150459666
0.185627347
[8] 0.070697915 0.013233200 0.005179334
$obar
[1] 0.4588113
$thres
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
$check
[1] 0.1820737
$bins
[1] TRUE
```

Βλέπουμε ότι στο μοντέλο μας το Brier Score ισούται με 0.18, δηλαδή εντός των αποδεκτών πλαισίων, αλλά και πολύ χαμηλότερα από το baseline που είναι 0.248.

Μια ακόμη σχετική καμπύλη που συνδέεται με την αξιοπιστία του συστήματος είναι η Probability Validation Curve, η οποία εξετάζει αν οι πιθανότητες που δίνει το σύστημα εμφανίζονται με τη συχνότητα που τους αναλογεί.



Εικόνα 17: Καμπύλη Εγκυρότητας Πιθανότητας (PVC)

Όπως φαίνεται από τις τιμές των παραπάνω μετρικών, πλέον έχουμε καταφέρει σε πολύ μεγάλο βαθμό να έχουμε μια ικανοποιητική πρόβλεψη του Churn Month. Συγκεκριμένα τα ποσοστά επιτυχίας μπορούν να εξαχθούν από τον ακόλουθο περιληπτικό πίνακα:

Actual / *Predict*	**Stay**	**Leave**
Stay	49.49%	15.8%
Leave	4.27%	30.16%

Το ποσοστό των ορθών προβλέψεων (Accuracy) είναι 79.65%.

Είναι προφανές ότι πλέον η παλινδρόμηση έχει αναμφίβολα αποδείξει την ικανότητα της να προβλέπει ορθά τον ερχομό churn month σε πολύ ικανοποιητικό βαθμό (8 στις 10 περιπτώσεις υπάρχει ορθή πρόβλεψη). Η επιτυχία αυτή της μεθόδου αποτελεί μια εγγύηση ότι τα μοντέλα παλινδρόμησης είναι εκείνα που μπορούν να δώσουν ακριβείς και επιτυχημένες προβλέψεις, ενώ ακόμη περισσότερο παρέχουν μια

επιπλέον επιβεβαίωση ότι οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν ήταν όντως αυτές (ή έστω ένα υποσύνολο αυτών) που είχαν συσχέτιση με την πιθανότητα επιβίωσης στο πρόγραμμα.

4.4 Μοντέλο Παλινδρόμησης με χρήση Clustering

4.4.1 Αλγόριθμοι Ομαδοποίησης

Σήμερα, μία από τις κυριότερες μεθόδους που ακολουθείται για την αντιμετώπιση της υπολογιστικής δυσκολίας που παρουσιάζει η πληθώρα των big data είναι το clustering. Οι κυρίαρχοι αλγόριθμοι που προσπαθούν να χαρακτηρίσουν τη συμπεριφορά χρηστών σε έναν τεράστιο πληθυσμό (όπως στα μοντέλα των δημοφιλών μέσων κοινωνικής δικτύωσης) είναι αυτοί που ομαδοποιούν πρώτα τους χρήστες βάσει κάποιων σημαντικών (για το ζητούμενο) χαρακτηριστικών και στη συνέχεια εφαρμόζουν τον αντίστοιχο αλγόριθμο επεξεργασίας πάνω στις ομάδες (clusters) αντί σε κάθε χρήστη ξεχωριστά, μειώνοντας έτσι κατά πολύ τον απαιτούμενο υπολογιστικό φόρτο. Ασφαλώς, το clustering πέρα από την υπολογιστική ελάφρυνση, αποτελεί προφανώς ένα αποδοτικό μέσο χαρακτηρισμού της συμπεριφοράς κάποιου χρήστη. Ακριβώς όπως στα δέντρα απόφασης, οι χρήστες χαρακτηρίζονται βάσει του φύλλου που καταλήγουν, έτσι και στο clustering, θεωρούμε ότι οι χρήστες που καταλήγουν στο ίδιο cluster (εξαιτίας κάποιων συγκεκριμένων τιμών τους) παρουσιάζουν στατιστική ομοιομορφία ως προς αυτό που ερευνούμε (ακριβώς επειδή υποθέσαμε μεγάλη συσχέτιση μεταξύ αυτών των χαρακτηριστικών και του προβλεπτέου).

Η προηγούμενη ανάλυση μας έχει επιβεβαιώσει ότι μεταβλητές που χρησιμοποιήθηκαν στο μοντέλο `glm` έχουν στατιστική σημασία στην πρόβλεψη της αποχώρησης ενός πελάτη. Τις ίδιες μεταβλητές θα χρησιμοποιήσουμε λοιπόν για να εφαρμόσουμε το clustering στους πελάτες. Ο αλγόριθμος που θα ακολουθήσουμε είναι αυτός του πακέτου Mclust της R, λεπτομέρειες της υλοποίησης του οποίου μπορούν να βρεθούν στο (Fraley, Raftery, Scrucca, Murphy, & Fop, 2017).

Σε γενικές γραμμές, αυτός ο αλγόριθμος υλοποιεί τον Expectation-Maximization (EM) αλγόριθμο για την βέλτιστη προσαρμογή ενός μείγματος Gaussian μοντέλων σε ένα δείγμα παρατηρήσεων. Πιο συγκεκριμένα, ο αλγόριθμος υποθέτει ότι τα δεδομένα προέρχονται από ένα σύμπλεγμα ανεξάρτητων Gaussian κατανομών, έκαστο εκ των οποίων αντιστοιχεί και σε κάποιο cluster. Σκοπός του είναι να εντοπίσει στα δεδομένα το πλήθος, τα χαρακτηριστικά και τις παραμέτρους αυτών των κατανομών. Για το πλήθος και τα χαρακτηριστικά των κατανομών, όταν αυτά είναι άγνωστα (όπως στην περίπτωση μας), εφαρμόζεται το Bayesian Information Criterion (BIC) (βλ. (Schwarz, 1978)) το οποίο προσπαθεί να εντοπίσει τη πιο εφαρμοστή μορφή και πλήθος clusters για τα δεδομένα. Από εκεί και πέρα, ο EM αλγόριθμος είναι ένας επαναληπτικός αλγόριθμος, ο οποίος ξεκινώντας από μία τυχαία αρχική κατάσταση (π.χ. τοποθετώντας τυχαία στο επίπεδο τα κέντρα των clusters) υπολογίζει για κάθε σημείο ποιο είναι το πιο πιθανό cluster στο οποίο να ανήκει. Στη συνέχεια,

μεταβάλλει τις παραμέτρους αυτών των cluster, έτσι ώστε να αυξάνει κάθε φορά τη συνολική πιθανοφάνεια, μέχρις ότου να συγκλίνει σε ένα τοπικό μέγιστο. Γενικά, ο EM αλγόριθμος (με εξαίρεση τη σχετικά μη ομαλή εκτέλεση του όταν δεν υπάρχουν αρκετοί αντιπρόσωποι από κάθε cluster) είναι ένας αποδοτικός και μορφολογικά ευέλικτος στατιστικός αλγόριθμος ομαδοποίησης που χρησιμοποιείται ευρέως σε τέτοιες περιπτώσεις.

4.4.2 Εφαρμογή Clustering για πρόβλεψη Churn Month

Όπως προείπαμε, τα κριτήρια ομαδοποίησης που χρησιμοποιήσαμε είναι οι μεταβλητές που χρησιμοποιήθηκαν και στα δέντρα απόφασης και GLM. Ένα τεχνικό ζήτημα που έπρεπε να διευθετηθεί πρώτα, όμως, είναι ότι εξαιτίας του μεγάλου υπολογιστικού φόρτου που έχουν οι αλγόριθμοι clustering, η εφαρμογή έπρεπε να γίνει πάνω σε ένα τυχαίο υποσύνολο 5000 δειγμάτων. Οι ακριβείς εντολές για να γίνει αυτή η τυχαία δειγματοληψία και η μετέπειτα κατασκευή του μοντέλου μπορούν να βρεθούν στο Παράρτημα.

```
Mclust(data = km_smr.df[sample(84728, 5000), c("immediate.var"
, "closed.var", "sums.var", "sums.mean")])
```

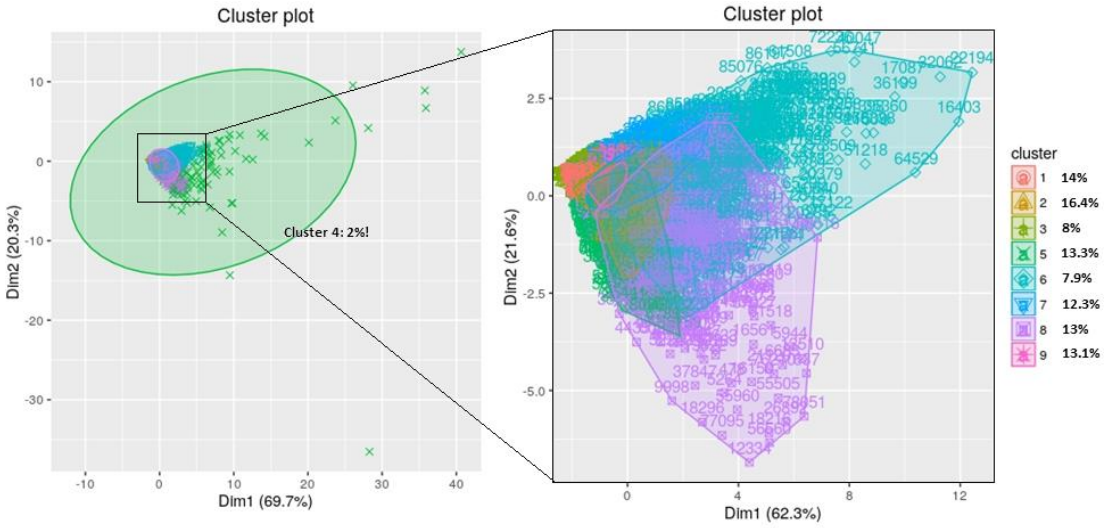
Παρατίθεται η απόκριση της εκτέλεσης, όπου παρουσιάζεται μεταξύ άλλων και η επιλογή του μοντέλου μαζί με το πλήθος των συνιστωσών (με τη βοήθεια του κριτηρίου BIC), η πιθανοφάνεια που επιτεύχθηκε από τον EM, καθώς και η κατανομή του πλήθους των 5000 δειγμάτων ανά συστάδα:

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------
Mclust VEV (ellipsoidal, equal shape) model with 9 compone
nts:

 log.likelihood    n  df       BIC       ICL
      -223022.2 5000 110 -446981.3 -448597.4
Clustering table:
  1   2   3   4   5   6   7   8   9
719 801 417  95 653 392 649 636 638
```

Όπως φαίνεται, λοιπόν, επιλέχθηκε μοντέλο VEV με 9 συνιστώσες. Το μοντέλο VEV σημαίνει ότι κατά το clustering, χρησιμοποιήθηκαν ελλειψοειδή χωρία μεταβλητού (Variable) όγκου, ίδιου (Equal) σχήματος και μεταβλητής (Variable) διεύθυνσης. Επίσης, παρατηρούμε ότι όντως υπάρχει σχετική ομοιομορφία στην κατανομή του

πληθυσμού στα 9 clusters. Στην εικόνα που ακολουθεί υπάρχει οπτικοποίηση των ανωτέρω clusters, όπως προέκυψε από το μοντέλο:



Εικόνα 18: Οπτικοποίηση του Clustering

Αρχικός σκοπός του clustering ήταν η δημιουργία ομοιογενών ομάδων (ως προς τις γνωστές περιληπτικές μετρικές) με τελικό σκοπό την κατασκευή clusters πελατών που να συμφωνούν ως προς την εμφάνιση ή μη churn month. Εξερευνήσαμε, λοιπόν, κατά πόσο υπάρχει ομοιομορφία σε κάθε ομάδα ως προς την τελική κατάσταση των πελατών:

Cluster	Stay	Leave	N	Verdict
1	27.2	72.8	11867	Leave
2	85.1	14.9	13870	Stay
3	1.9	98.1	6787	Leave
4	57.5	42.5	1684	Outliers
5	33.7	66.3	11278	Leave
6	60.3	39.7	6653	Stay?
7	42.5	57.5	10413	Leave?
8	65.2	34.8	11042	Stay?
9	91.1	8.9	11134	Stay

Σε ενδεχόμενη εφαρμογή του παραπάνω μοντέλου, λοιπόν, ανάλογα και με το cluster που φέρεται να ανήκει κάποιος πελάτης θα βγάζαμε και την αντίστοιχη ετυμηγορία (Verdict) με την ακρίβεια που προκύπτει από τα παραπάνω ποσοστά. Γενικά, παρατηρούμε ότι για το 64.8% των πελατών, μπορούμε να εξάγουμε βέβαιη απόφαση, για το 33.2% από λίγο έως περισσότερο αμφίρροπη και μόλις για το 2.0% αδυνατούμε να βγάλουμε βέβαιη απόφαση. Παρότι, το συνολικό Accuracy έχει πέσει σε σχέση με το προηγούμενο μοντέλο (για την ακρίβεια είναι πλέον 74.21%) υπάρχει

το μεγάλο πλεονέκτημα ότι μπορούμε χάρη σε αυτή την εκλέπτυνση να περιγράφουμε πληρέστερα τη βεβαιότητα μας για κάθε απόφαση ανάλογα με το ποιο cluster καταλήγει ένας πελάτης. Η συνεισφορά αυτού του μοντέλου δηλαδή είναι περισσότερο ποιοτικής φύσεως παρά ποσοτικής (σε συμφωνία με το (Hu, Comparison of Classification Methods for Customer Attrition Analysis, 2002)).

4.4.3 Σύγκριση με Μοντέλο Παλινδρόμησης

Προκύπτει ασφαλώς το ερώτημα, πόσο καλύτερο είναι το προηγούμενο μοντέλο παλινδρόμησης σε σχέση με το clustering και κατά πόσο μπορούμε να συνδυάσουμε τις δύο προσεγγίσεις έτσι ώστε να έχουμε ένα τόσο ποιοτικά όσο και ποσοτικά βελτιωμένο μοντέλο. Παρουσιάζει ενδιαφέρον, για αρχή, να δούμε πώς συμπεριφέρεται το προηγούμενο μοντέλο παλινδρόμησης ανά cluster, κάτι το οποίο φαίνεται συνοπτικά στον ακόλουθο πίνακα:

Predict / *Cluster*	**Stay**	**Correct**	**Leave**	**Correct**	**Accuracy**
1	34%	76%	65%	99%	90.2%
2	85%	85%	15%	20%	75.2%
3	7%	0%	95%	94%	89.3%
4	80%	50%	20%	75%	55.0%
5	41%	63%	59%	85%	75.8%
6	86%	70%	14%	100%	74.2%
7	85%	50%	15%	100%	57.5%
8	66%	85%	34%	75%	81.6%
9	99%	90%	1%	0%	89.1%

Ας ελέγξουμε τώρα τι γίνεται αν συνδυάσουμε τις δύο μεθόδους (με κριτήριο το accuracy). Συγκεκριμένα, για κάθε cluster επιλέγουμε είτε την απόφαση εξαιτίας του cluster είτε από το μοντέλο παλινδρόμησης, ανάλογα με το ποια από τις δύο μεθόδους φαίνεται να προσφέρει τη μεγαλύτερη ακρίβεια. Σε αυτή τη περίπτωση θα είχαμε το εξής αποτέλεσμα:

Cluster	Ποσοστό Πληθυσμού	Μοντέλο που Ακολουθείται	Accuracy
1	14.0%	Regression	90.2%
2	16.4%	Clustering	85.1%
3	8.0%	Clustering	98.1%
4	2.0%	Clustering	57.5%
5	13.3%	Regression	75.8%
6	7.9%	Regression	74.2%
7	12.3%	Clustering/Regression	57.5%
8	13.0%	Regression	81.6%
9	13.1%	Clustering	91.1%

Το τελικό Accuracy προκύπτει ότι αυξάνεται σε 81.14%.

Παρότι η αύξηση είναι κοντά στις 2 ποσοστιαίες μονάδες (οι οποίες σε επιχειρησιακό επίπεδο αποτελούν ασφαλώς μια διόλου ασήμαντη ποσότητα) από το απλό μοντέλο παλινδρόμησης (και κοντά στις 8 ποσοστιαίες μονάδες σε σχέση με το clustering), η μεγάλη συνεισφορά του παραπάνω μοντέλου είναι η επίτευξη του συνδυασμού αυξημένων ποσοστών επιτυχίας με επίσης αυξημένη ποιότητα εκτίμησης της βεβαιότητας της απόφασης (ανάλογα το cluster).

Σε κάθε περίπτωση, η εκλέπτυνση του παραπάνω μοντέλου μας πληροφορεί ότι η χρήση τέτοιων υβριδικών προσεγγίσεων δύναται να προσφέρει ένα τόσο ποιοτικά όσο και ποσοτικά ανεπτυγμένο μοντέλο προβλέψεων. Με εφαλτήριο αυτές τις παρατηρήσεις, θα προτείνουμε και μελετήσουμε στο επόμενο κεφάλαιο ορισμένες δυνατές προεκτάσεις του παραπάνω μοντέλου που ενδεχομένως να βελτιώνουν σε ακόμη υψηλότερα επίπεδα την ακρίβεια και τα χαρακτηριστικά του συστήματος μας.

Κεφάλαιο 5

Συμπεράσματα και Επεκτάσεις

5.1 Αποτελέσματα

5.1.1 Αρνητικά Αποτελέσματα

Στην πορεία μας μέχρι αυτό το σημείο, είδαμε ποια μοντέλα δύνανται να αποτελέσουν πλαίσια κατάλληλα για τη δημιουργία ενός συστήματος πρόβλεψης της συμπεριφοράς των πελατών ως προς την παραμονή τους στο τραπεζικό πρόγραμμα 'Affluent' και για να φτάσουμε εκεί χρειάστηκε να αντλήσουμε τις πληροφορίες που παίρναμε από τα αρνητικά αποτελέσματα των πρωταρχικών ακατάλληλων μοντέλων.

Συγκεκριμένα αποκλείσαμε τα απλοϊκά μαρκοβιανά μοντέλα κυλιόμενου παραθύρου (σταθερού μήκους) εξαιτίας της αδυναμίας τους να παρέχουν μια πλήρη και καθαρή εικόνα των επί μέρους μοτίβων που εμφανίζουν οι χρονοσειρές που καταλήγουν σε 'Lost'. Επιπλέον, είδαμε ότι τα περισσότερα από τα πεδία που είχαμε στη διάθεση μας (μεταξύ των οποίων η ολοκληρία των δημογραφικών) παρουσίαζε σχετική ανεξαρτησία από το προβλεπτέο και συγκεκριμένα η ανάμιξη τους μαζί με τις όντως συσχετιζόμενες μεταβλητές προκαλούσε αλλοίωση της συνεισφοράς των τελευταίων. Με κατάλληλα μοντέλα παλινδρόμησης καταφέραμε να εντοπίσουμε και εν τέλει να απομονώσουμε στο μοντέλο μας μόνο τις πραγματικά χρήσιμες.

5.1.2 Θετικά Αποτελέσματα

Εν συνεχεία, εκμεταλλευόμενοι όσα εξήγαμε από τη προηγούμενη φάση, καταλήξαμε σε μοντέλα παλινδρόμησης και συγκεκριμένα σε μοντέλα παλινδρόμησης που εξαρτιόνταν μόνο από το κατάλληλο υποσύνολο μεταβλητών και παρουσίαζαν για πρώτη φορά υψηλά ποσοστά επιτυχίας μαζί με καλά επίπεδα και στις υπόλοιπες μετρικές αξιολόγησης.

Σε συνέχεια αυτής της αναζήτησης, δοκιμάστηκε μία από τις συνηθέστερες μεθόδους αντιμετώπισης μεγάλων δεδομένων: η ομαδοποίηση. Με χρήση clustering λοιπόν μπορέσαμε να ομαδοποιήσουμε τους πελάτες βάσει των πεδίων που δείξαμε ότι παρουσιάζουν συσχέτιση με την πιθανότητα επιβίωσης στο πρόγραμμα και καταλήξαμε ότι προκύπτουν συστάδες όπου οι πελάτες παρουσιάζουν μεγάλη ομοιογένεια και στις οποίες μπορούμε με μεγάλη ποιότητα βεβαιότητας να εξάγουμε (για τη πλειοψηφία των clusters) ποια θα είναι η συμπεριφορά των πελατών που ανήκουν σε αυτό. Τέλος, κατασκευάζοντας ένα υβριδικό μοντέλο, στο οποίο για τα

clusters όπου η Ακρίβεια δεν ήταν τόσο υψηλή, χρησιμοποιούσαμε το καλύτερης επίδοσης μοντέλο παλινδρόμησης, καταλήξαμε σε ένα μοντέλο προβλέψεων που παρείχε τόσο συνολική ποσοτική ποιότητα προβλέψεων όσο και ποιοτική εκτίμηση της βεβαιότητας ανά απόφαση.

5.2 Δυνατές Επεκτάσεις του Συστήματος

5.2.1 Μορφολογικές και Σημασιολογικές Βελτιώσεις

Υπάρχουν διάφορες κατευθύνσεις ως προς τις οποίες μπορεί να επεκταθεί η μέχρι τώρα πορεία. Από λειτουργικής πλευράς, πολλές δυνατότητες παρουσιάζει η εξερεύνηση εναλλακτικών παραμέτρων και μορφολογιών για τα διάφορα μοντέλα που δοκιμάσαμε.

Σε πολύ πρώτη φάση, θα μπορούσε να γίνει πειραματισμός με εναλλακτικά μήκη παραθύρου και συνδυασμών πεδίων για την περίπτωση των μαρκοβιανών μοντέλων, αν και από τη σχετική ανάλυση που έγινε είναι μάλλον απίθανο να υπάρξει ουσιώδης βελτίωση.

Περισσότερο ενδιαφέρον παρουσιάζει η περίπτωση των μοντέλων παλινδρόμησης. Για αρχή υπάρχει η δυνατότητα εξερεύνησης εναλλακτικών επιλογών στις βασικές παραμέτρους, όπως για παράδειγμα η επιλογή άλλων κατανομών (Normal, Poisson κ.τ.λ. αντί για Binomial) στο μοντέλο παλινδρόμησης και έλεγχος για το αν κάποιο από αυτά δίνει πιο ταιριαστά αποτελέσματα. Παρόμοια στο clustering θα μπορούσαμε να εξερευνήσουμε διαφορετικές μεθόδους συσταδοποίησης (ή εναλλακτικά του BIC κριτηρίων) για την περίπτωση που κάποια διαφορετική μορφολογία προκύψει να προσφέρει καλύτερης ποιότητας αποτελέσματα.

Ασφαλώς, ένα από τα κυριότερα ζητήματα παραμένει η αύξηση της Ακρίβειας και η αντιμετώπιση των σκοτεινών περιοχών, τις οποίες τα βασικά μοντέλα που έχουμε μελετήσει δε φάνηκε να καλύπτουν με ικανοποιητική ακρίβεια, όπως τα clusters 4 και 5-7 στο προηγούμενο κεφάλαιο. Μία πρώτη κατεύθυνση θα μπορούσε να είναι η προσπάθεια ανακάλυψης (με κάποια πιο σύνθετη μέθοδο) κάποιας τυχούσας συσχέτισης επιπλέον μεταβλητών (ή κάποιων συναρτήσεων με όρισμα συνδυασμούς τους) με στόχο την άντληση όσο το δυνατόν περισσότερης (αλλά πάντοτε ορθής) πληροφορίας από τα δεδομένα μας.

Μία εναλλακτική και πιο ελπιδοφόρα προσέγγιση φάνηκε να είναι αυτή των υβριδικών μοντέλων. Φάνηκε ότι, εκ κατασκευής, το τελικό υβριδικό μοντέλο δε θα μπορούσε παρά να έχει περισσότερη Ακρίβεια από τα δύο προηγούμενα ως εκλέπτυνση αυτών. Παρατηρήσαμε, μάλιστα, ότι σε μερικά clusters, το μοντέλο παλινδρόμησης ήταν εκείνο που παρουσίαζε υπεροχή σε σχέση με το clustering καθώς και γενικά υψηλότερα ποσοστά επιτυχίας από τη μέση ακρίβεια και του ίδιου του GLM. Προς αυτή την κατεύθυνση, λοιπόν, παρουσιάζουν, πολύ μεγάλο ενδιαφέρον οι διάφορες δυνατότητες εκλέπτυνσης και υβριδοποίησης του μοντέλου μέσω clustering. Συγκεκριμένα δεδομένου ότι ανά ορισμένα cluster φαίνεται η υφιστάμενη ομοιογένεια να οδηγεί σε μεγαλύτερα ποσοστά επιτυχίας, είναι ιδιαίτερα εύλογο να δοκιμαστεί η κατασκευή ξεχωριστού δέντρου απόφασης ή GLM

για κάθε cluster ή ακόμη περισσότερο να εφαρμοστεί και από ένα διαφορετικό μοντέλο ανά cluster (χρονοσειρές, δέντρο απόφασης, επαναληπτικό clustering κ.ο.κ.) και να επιλέγεται για κάθε ένα το βέλτιστο με σκοπό τη συνολική αύξηση της επιτυχίας του συστήματος (βλ. και (Kotthoff, 2012)). Τέλος έχει νόημα να διερευνηθούν και συμμετέχουν πιο ενεργά στην αξιολόγηση και λοιπές μετρικές χαρακτηρισμού της επιτυχίας του συστήματος πέρα από το Accuracy, όπως το Precision, το F-measure ή το AUC ανά cluster έτσι ώστε να εξασφαλίσουμε ένα τελικό ομοιόμορφο σύστημα με τις ποιοτικά και ποσοτικά βέλτιστες δυνατές και ευσταθείς προβλέψεις.

5.2.2 Υλοποίηση Τελικού Προϊόντος

Έχοντας βελτιστοποιήσει το συνολικό αλγόριθμο, απομένει η μετατροπή του σε ένα ολοκληρωμένο εμπορικό προϊόν για την εκμετάλλευση του και εφαρμογή του σε πραγματικά συστήματα. Τα κυρίως τεχνικά κομμάτια που απομένουν σε πρώτο επίπεδο, προς αυτό τον σκοπό, είναι η υλοποίηση ενός κατάλληλου UI καθώς και η προτυποποίηση του με χρήση κατάλληλων διεπαφών που να κάνουν πιο εύκολη τη συντήρηση του κώδικα (σε εμπορικά επίπεδα) για αρχή.

Από εκεί και πέρα εγείρονται διάφορα πιθανά ζητήματα κλιμάκωσης του κώδικα, όπως η εφαρμογή του σε ακόμη μεγαλύτερου μεγέθους δεδομένα καθώς και η ευελιξία του σε σχέση με το πλήθος και τη στατιστική μορφολογία των παρελθοντικών δειγμάτων. Παρότι η έρευνα μας επικεντρώθηκε στην περίπτωση του τραπεζικού προγράμματος, μεγάλο ενδιαφέρον παρουσιάζει και η μελέτη της δυνατότητας προσαρμογής του αλγορίθμου σε διαφορετικής φύσεως υπηρεσίες και δεδομένα.

Τέλος σε έσχατο στάδιο, σκοπός είναι να κατασκευαστεί ένα πλήρες κι ολοκληρωμένο προγραμματιστικό προϊόν που ανάλογα με τα διαθέσιμα δεδομένα θα έχει προσαρμοστικό χαρακτήρα καθ’ όλη την πορεία της εκτέλεσης του. Συγκεκριμένα στη φάση της προπόνησης θα κρίνει από το πλήθος και τη μορφολογία των παρελθοντικών δεδομένων, την αξιοπιστία τους ανά περίοδο και θα επιλέγει τον καταλληλότερο από τους δυνατούς αλγορίθμους. Παράλληλα θα υλοποιεί και τεστ που θα απομονώνουν και χρησιμοποιούν μόνο τις εγγραφές (ή συναρτήσεις αυτών) που παρουσιάζουν συσχέτιση με το ζητούμενο. Στη συνέχεια, αναλόγως (και σε ανάδραση) με τα αποτελέσματα της προπόνησης, θα διερευνά ποια μέθοδος είναι η καλύτερη ανά cluster (ή γενικότερα ανά δομικό ιστό) και θα φροντίζει για την αποδοτική εφαρμογή της. Ταυτόχρονα έχει νόημα να γίνει αναβάθμιση σε online αλγόριθμο, όπου κάθε νέος πελάτης ελέγχεται ως προς την προσαρμογή του στο υπάρχον σύστημα, ενώ ανά τακτά χρονικά διαστήματα γίνεται επαναξιολόγηση της τρέχουσας δομής του συστήματος και μεταβολή των επί μέρους παραμέτρων προς την κατεύθυνση που θα αυξάνει τη συνολική επιτυχία του, εξασφαλίζοντας έτσι ένα χρονικά ευσταθές και καθολικά συνεπές σύστημα προβλέψεων.

Βιβλιογραφία

Anderson, R. L., & W., A. T. (1950). Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. *Ann. Math. Stat., 21*, 59-81.

Bloomfield, P. (2004). *Fourier Analysis of Time Series: An Introduction.* Wiley Series in Probability and Statistics.

Breiman, L. a. (1984). *Classification and Regression Trees.* Monterey, California, U.S.A.: Chapman & Hall.

Colosimo, E., Ferreira, F., Oliveira, M., & Sousa, C. (2002, April). Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation, 72*(4), 299-308. doi:10.1080/00949650212847

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., & Fop, M. (2017, May 21). *Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation.* Ανάκτηση από CRAN: https://cran.r-project.org/web/packages/mclust/mclust.pdf

Gorgoglione, M., & Panniello, U. (2011, April 8). Beyond Customer Churn: Generating Personalized Actions to Retain Customers in a Retail Bank by a Recommender System Approach. *Journal of Intelligent Learning Systems and Applications, 3*(2), 90-102. doi:10.4236/jilsa.2011.32011

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006, April). Churn Prediction: Does Technology Matter? *International Journal of Intelligent Technology, 1*(2), 104-110.

Hu, X. (2002). Comparison of Classification Methods for Customer Attrition Analysis. *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing* (σσ. 487-492). London, UK: Springer-Verlag. Ανάκτηση από http://dl.acm.org/citation.cfm?id=646473.692840

Hu, X. (2005). A Data Mining Approach for Retailing Bank Customer Attrition Analysis. *Applied Intelligence, 22*(1), 47-60. doi:10.1023/B:APIN.0000047383.53680.b6

Koopman, S. J., & Lee, K. M. (2009). Seasonality with Trend and Cycle Interactions in Unobserved Components Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 58*(4), 427-448.

Kotthoff, L. (2012). Hybrid Regression - Classification Models for Algorithm Selection. *Frontiers in Artificial Intelligence and Applications, 242*, 480-485.

Li, L., Noorian, F., Moss, D. J., & Leong, P. H. (2006). *Rolling Window Time Series Prediction using MapReduce.* NSW, Australia.

Meier, E. L. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Migueis, V. L., Van den Poel, D., Camanho, A. S., & Falcao e Cunha, J. (2012). Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification, 6*(4), 337-353. doi:10.1007/s11634-012-0121-3

Norris, J. R. (1997). *Markov Chains.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511810633

Quinlan, J. R. (1986). Induction of Decision Trees. *Mach. Learn., 1*(1), 81-106. doi:10.1023/A:1022643204877

R Team, D. C. (2011). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Schwarz, G. (1978, March). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Therneau, T. M., & Lumley, T. (2017, April 4). *Survival Analysis.* Ανάκτηση από CRAN: https://cran.r-project.org/web/packages/survival/survival.pdf

Therneau, T., Atkinson, B., & Ripley, B. (2017, April 21). *Recursive Partitioning and Regression Trees.* Ανάκτηση από CRAN: https://cran.r-project.org/web/packages/rpart/rpart.pdf

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software, 59*(10).

Wu, C.-J. C.-B.-Y. (2012). Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan. *Scientia Iranica, 19*(3), 849 - 855. doi:https://doi.org/10.1016/j.scient.2011.11.045

Zhang, B. (2017, April 2). Application of Survival Analysis for Predicting Customer Churn with Recency, Frequency, and Monetary. Orlando, Florida, U.S.A.

Παράρτημα:
Κώδικες των Προγραμμάτων που χρησιμοποιήθηκαν σε R

Πρόβλεψη με χρήση Κυλιόμενου Παραθύρου

```
#
# Commands used for timeseries prediction.
# Read Data, shift Window Length, find Pattern
#

#month shifting
data.df -> data_1.df
data.df -> data_2.df
data_1.df$time - 1 -> data_1.df$time
data_2.df$time - 2 -> data_2.df$time

#data merging
merge(data.df[,c("AA", "time", "labels")],
      data_1.df,
      by=c("AA", "time")) -> dataX.df
dataX.df$labels <- dataX.df$labels.x
dataX.df$labels.x <- NULL
dataX.df$labels.y <- NULL
merge(dataX.df, data_2.df, by=c("AA", "time")) -> dataXX.df

#find pattern using all economic variables
glm(labels ~ immediate.y + immediate.x +
insurance.y + insurance.x + investment.y + investment.x +
business.y + business.x + consumer.y + consumer.x +
closed.y + closed.x + housing.y + housing.x +
contrib.y + contrib.x + sums.x + sums.y,
family="binomial"(link='logit'),
data=dataXX.df
) -> glm_model
```

Τακτοποίηση Δεδομένων

```
#
# pankgeorg@programize.com
#
#
# Script used to make data tidy.
# Read data, split to categories, reshape to be useful
#
# Input: csv (mirroring given Excel) file
#
# Output: variables
# 'demographics': key: AA,       /  91089 rows
# 'economics'   : key: AA, month /1821780 rows
# 'status'      : key: AA, month /1821780 rows
#
# Months are from January 2014 to August 2015 (20 Months)

library(reshape2);
library(tidyr);
library(dplyr);

# Invoke garbage collection.
gc();
# raw_data
read.csv(
         "raw/ajnfn.csv",
         sep=";",
         stringsAsFactors = F
         ) -> raw_data;
print("CSV Loaded");
# print(head(raw_data))  # Too long

# Column Explanation
# 1      is the key
# 2:21   are the statuses for each month
# 22:31  are the ids/demographics
# 23:231 are the economics

keep   = names(raw_data)[c(1, 2:21, 22:31, 32:231)];
ids    = names(raw_data)[c(1, 22:31)];
melts1 = names(raw_data)[c(1, 32:231)];
melts2 = names(raw_data)[c(1, 2:21)]
# demographics
raw_data %>% subset(
       select=ids
       ) -> demographics;

print("Subsetted Demographics");
# print(head(demographics));
```

```
# economic data
raw_data %>% subset(
                     select=melts1
                     ) %>%
    reshape2::melt(
                    id='AA'
                    ) %>%
    tidyr::extract(
                    variable,
                    regex="([:alpha:]+)\\.?(.*)",
                    c("col", "month")
                    ) %>%
    reshape2::dcast(
                     formula = AA + month ~ col
                     ) -> economics;
print("Subsetted & Casted economics")

# Status
raw_data %>% subset(
                     select=melts2
                     ) %>%
    reshape2::melt(
                    id='AA'
                    ) %>%
    tidyr::extract(
                    variable,
                    regex="([:alpha:]+)\\.?(.*)",
                    c("col", "month")
                    ) %>%
    reshape2::dcast(
                     formula = AA + month ~ col
                     ) -> status;
print("Subsetted and Casted status")
print("Melting, Extracting and Casting complete!");

gc();
```

Προετοιμασία Δεδομένων σε κατάλληλη μορφή για Επεξεργασία

```
#
# pankgeorg@programize.com
#
#
# Script used to prepare data for the analysis
# Change names, remove #N/A lines

# Input: datasets
# 'demographics'
# 'economics'
# 'status'
# [raw_to_tidy.R]
#
# Output:
# Clean, consistent datasets for analysis

library(reshape2);
library(tidyr);
library(dplyr);

gc();

# Step 1: Fix names

# Names for demographics
c(
  "Id", "Sex", "Age", "Postal", "MaritalStatus",
  "Education", "HomeStatus", "Profession", "Email",
  "InternetConn", "Phone"
  ) -> nd
names(demographics) <- nd
print("Clean demographics names")

# Names for economics
c(
  "Id", "Month", "Immediate", "Insurance", "Investment",
  "Business", "Consumer", "Closed", "MB", "Housing",
  "Contributions", "Sums"
  ) -> ne
names(economics) <- ne
print("Clean economics names")
```

```
# Names for statuses
c(
  "Id", "Month", "Status"
  ) -> ns
names(status) <- ns
print("Clean status names")

### Vectorized Coersion (that's cool R)
as.boolYN_NA  <- function(x){
      return ((!(x == '#N/A') | NaN)*(x == 'Y'))
      }

# Step 2: Sanitize data (fix types, make consistent)
# Demographics
demographics %>%
    transform(
              Sex           = as.factor (Sex),
              Age           = as.numeric(Age),
              Postal        = as.factor (Postal),
              MaritalStatus = as.factor (MaritalStatus),
              Education     = as.factor (Education),
              HomeStatus    = as.factor (HomeStatus),
              Profession    = as.factor (Profession),
              Email         = as.numeric(Email),      # That's 0 or
1
              InternetConn  = as.boolYN_NA(InternetConn),
              Phone         = as.boolYN_NA(Phone)
              ) -> demographics
print("Transformed Demographics")

# Status
status %>%
    transform(
              StatusF = as.factor(Status),
              Flag    = Status == 'current'
              ) %>%
    subset(
           columns=c("Id", "Month", "StatusF", "Flag")
           ) -> status
print("Transformed Status")

# #### # ###################
# Step 3: Remove #N/As
# #### # ###################

rm_ids <- demographics[demographics$Sex == "#N/A", "Id"]
cdemg <- demographics[demographics$Sex != "#N/A",]
cecon <- economics[!(economics$Id %in% rm_ids),]
cstat <- status[   !(status   $Id %in% rm_ids),]
print("Removed #N/A - demographics")

# #### # ##################
```

```
# Step 4: Calculate Active
# #### # ##################
cstat %>%
    group_by(Id) %>%
    summarise(
              Months=sum(Flag)
              ) -> months_active

months_active %>%
    group_by(Months) %>%
    summarise(
              Count=length(Months)
              ) -> no_per_month
print("Calculate Active Months")

cdemg %>%
    left_join(months_active, by="Id") %>%
    transform(
              Stay = Months == 20
              ) -> cdemg

cecon %>% left_join(cstat) -> cecon_ext
print("Cleaning complete")
```

Επεξεργασία Δεδομένων και Εξαγωγή Μοντέλων Παλινδρόμησης και Clustering

```
#
#   pankgeorg@programize.com
#
#
#   This script demonstrates how we work
#   with the cleaned data.
#

library(mclust)
library(dplyr)
library(stats)
library(Hotelling)
library(verification)

# #### # ##############################
# Step 0: Extract useful variables    #
# #### # ##############################

# Useful Economics
cecon_ext %>%
    filter(Flag == TRUE) %>%
    group_by(Id) %>%
    summarise(
              Immediate_var = sqrt(var(Immediate)),
              Sums_var      = sqrt(var(Sums)),
              Closed_var    = sqrt(var(Closed)),
              Sums_mean     = mean(Sums),
              Months        = length(Month),
              Lost          = 0 + ((sum(Flag) < 20) |
                                   !(max(Month) == '2015.08'))
              ) -> summary_econ

# #### # ##############################
# Step 1: Functions to evaluate model #
# #### # ##############################

makeModel <- function(df){
    glm(
        Lost ~ Immediate_var + Sums_var + Closed_var + Sums_mean,
        family = binomial(link="logit"),
        data=df
        ) -> modelA
    return (modelA)
}
```

```
evalModel <- function(amodel, test){
    predict.glm(
                amodel,
                newdata=test,
                type="response"
                ) > 0.5 -> pv
    return (table(pv, test$Lost)/nrow(test))
}

brierModel <- function(amodel, test){
    predict.glm(
                amodel,
                newdata=test,
                type="response"
                ) -> prediction
    print(verification::brier(test$Lost, prediction))
    return (verification::brier(test$Lost, prediction))

}

pvc <- function(amodel, test){
    predict.glm(
                amodel,
                newdata=test,
                type="response"
                ) -> prediction
    round(100*prediction) -> bins
    rep(0, 101) -> p
    for(i in 1:102){
        q = test$Lost[bins == i]
        sum(q == 1, na.rm = T)/sum(q != 2, na.rm = T) -> p[i]
    }
    plot(p)
    print(p)
}

# #### # ##############################
# Step 2: Use logistic regression     #
# #### # ##############################

train  <- sample_n(summary_econ, 10000)
test   <- sample_n(summary_econ[!(summary_econ$Id %in% train$Id),],
40000)

modelA <- makeModel(train)
result <- evalModel(modelA, test)
brierModel(modelA, test)
pvc(modelA, test)
print(evalModel(modelA, test))
print(evalModel(modelA, train))


# #### # ##############################
# Step 3: Use mclust models           #
```

```r
# #### # ##############################

train_mclust = drop_na(train[,c(1:6)])
test_mclust  = drop_na(test [,c(1:6)])


mc <- mclust::Mclust(

      train_mclust[sample(nrow(train_mclust), 5000),]

)

res <- Predict.Mclust(mc, test_mclust)
```