



**Πρόγραμμα Μεταπτυχιακών Σπουδών**

**στην Αναλυτική των Επιχειρήσεων και την Επιστήμη Δεδομένων**

**Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων**

**Διπλωματική Εργασία**

# Πρόβλεψη και Ερμηνεία της Απώλειας Πελάτη στον Κλάδο του Διαδικτυακού Στοιχηματισμού

**της**

**Αναστασίας Τσιρίκη του Ηλία**

**Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη Δεδομένων**

**Αύγουστος 2023**

Με μεγάλη χαρά και ευγνωμοσύνη, θέλω να εκφράσω τις θερμές μου ευχαριστίες για την υποστήριξη και την εμπιστοσύνη που έλαβα κατά την διάρκεια της εκπόνησης της διπλωματικής μου εργασίας από τον υπεύθυνο καθηγητή μου κύριο Ταραμπάνη Κωνσταντίνο, καθώς και τον επίκουρο καθηγητή κύριο Καλαμπόκη Ευάγγελο και την κυρία Καραμάνου Αρετή. Η διαδικασία αυτής της εργασίας ήταν μια πολύτιμη εμπειρία για μένα, και δεν θα μπορούσα να την είχα πραγματοποιήσει χωρίς τη βοήθεια και την καθοδήγησή σας. Οι συμβουλές και οι ενθαρρύνσεις σας με ώθησαν να ξεπεράσω τις προκλήσεις και να επιτύχω τα καλύτερα δυνατά αποτελέσματα.

Επίσης, θέλω να ευχαριστήσω τον άντρα μου, τους γονείς και την αδελφή μου για τη στήριξη και την κατανόησή τους κατά την διάρκεια αυτής της πορείας. Η συμπαράστασή τους ήταν απαραίτητη για την ολοκλήρωση αυτού του σημαντικού εγχειρήματος.

Με την ολοκλήρωση αυτής της διπλωματικής εργασίας, κλείνει ένα κεφάλαιο στην ακαδημαϊκή μου πορεία, αλλά οι εμπειρίες που απέκτησα θα με συνοδεύουν σε κάθε νέα επαγγελματική πρόκληση. Είμαι ευγνώμων προς όλους εσάς που κάνατε αυτό το ταξίδι τόσο σημαντικό.

# Περίληψη

Η πρόβλεψη της απώλειας πελάτη αποτελεί σημαντικό παράγοντα για την λειτουργικότητα και την ανάπτυξη μιας επιχειρήσεις. Επιτρέπει την αναγνώριση των πελατών που έχουν αυξημένη πιθανότητα να διακόψουν την χρήση των υπηρεσιών ή την αγορά των προϊόντων μιας επιχείρησης και επίσης, υποστηρίζει τον ορθότερο σχεδιασμό της στρατηγικής που θα ακολουθήσει ώστε να ενισχύσει την ανάπτυξη της επιχείρησης και να μειώσει το κόστος που προκαλείται από την απώλεια μέρους των πελατών της. Επίσης, η χρήση τεχνικών Εξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence XAI) δίνει την δυνατότητα να αναγνωριστούν τα χαρακτηριστικά εκείνα που επηρεάζουν σε μεγαλύτερο βαθμό την πρόβλεψη του μοντέλου, και κατ' επέκταση να ληφθούν αποφάσεις και μέτρα που εστιάζουν προς την κατεύθυνση που θα συμβάλουν στην επίτευξη των καλύτερων δυνατών αποτελεσμάτων, τα οποία βασίζονται στην χρήση μηχανικής μάθησης. Η μελέτη αυτή έχει ως σκοπό την χρήση μηχανικής μάθησης και Εξηγήσιμης Τεχνητής Νοημοσύνης για την πρόβλεψη απώλειας πελάτη στον τομέα του διαδικτυακού στοιχηματισμού ώστε να βοηθήσει τα στελέχη της εταιρίας να λάβουν ακριβείς και αξιόπιστες αποφάσεις στον τρόπο που θα διαχειριστούν το πρόβλημα της απώλειας πελατών. Τέλος, παρουσιάζεται μία μελέτη περίπτωσης στην οποία γίνεται χρήση του αλγορίθμου XGBoost για να δημιουργηθεί το μοντέλο που θα προβλέπει την απώλεια πελάτη, σε δεδομένα που έχουν ληφθεί από την εταιρία Kaizen Gaming, η οποία είναι προσφέρει υπηρεσίες διαδικτυακού στοιχηματισμού στην Ελλάδα και το εξωτερικό. Η αξιολόγηση του μοντέλου γίνεται με την χρήση των μετρικών Accuracy, Precision και Recall, καθώς και με την αυτή των AUROC και AUPRC και η επεξήγηση των αποτελεσμάτων με την μέθοδο ανάλυσης SHAP (SHapley Additive exPlanations). Τα αποτελέσματα που προέκυψαν από το μοντέλο πρόβλεψης δίνουν τιμή Accuracy 0.95, Precision 0.91, Recall 0.89, AUROC 0.921 και AUPRC 0.919. Σχετικά με την ανάλυση SHAP προκύπτει ότι μεγαλύτερη επίδραση στην ικανότητα πρόβλεψης του μοντέλου έχουν οι μεταβλητές που αφορούν στον αριθμό τον ημερών που μένει ενεργός ένας πελάτης τον τελευταίο μήνα, στο ποσό των καταθέσεων που κάνει τον τελευταίο μήνα, στον αριθμό τον ημερών που μένει ενεργός ένας πελάτης το τελευταίο εξάμηνο και στον αριθμό των δελτίων που τοποθετεί τον τελευταίο μήνα.

# Πίνακας Περιεχομένων

# Κατάλογος πινάκων



# Κατάλογος εικόνων

# 1 Εισαγωγή

## 1.1 Ορισμός προβλήματος και ορολογία

Ο όρος του διαδικτυακού στοιχηματισμού αναφέρεται στην δυνατότητα που έχουν οι χρήστες να στοιχηματίζουν μέσω διαδικτύου είτε από την ιστοσελίδα, είτε από την εφαρμογή μιας εταιρίας, χωρίς να υπάρχει φυσική παρουσία σε κάποιον συγκεκριμένο χώρο. Σε αντίθεση με άλλους κλάδους, στον διαδικτυακό στοιχηματισμό το γεγονός να προσδιοριστεί αν ένας πελάτης έχει διακόψει την δραστηριότητά του στην συγκεκριμένη εταιρία είναι δυσκολότερο, καθώς δεν υπάρχει κάποιο συμβόλαιο που να υποδεικνύει την διακοπή αυτή, όπως για παράδειγμα μπορεί να συμβαίνει σε άλλους κλάδους λόγου χάρη στις τηλεπικοινωνίες που υπάρχει συμβόλαιο ή τα χρηματοπιστωτικά ιδρύματα όπου λήγει μία κάρτα. Για τον λόγο αυτό, δεν υπάρχει συγκεκριμένος όρος που να έχει εφαρμογή συνολικά στον κλάδο, αλλά κάθε εταιρία ορίζει με τα δικά της κριτήρια πότε θεωρεί πως κάποιος πελάτης έχει πάψει την δραστηριότητά του σε αυτή. Αυτό αποτελεί και τον λόγο που είναι ακόμα χρησιμότερη η δημιουργία διαδικασιών που επιτρέπουν την έγκαιρη και έγκυρη πρόβλεψη πιθανής απώλειας πελατών, ώστε να υπάρχει η δυνατότητα στα αρμόδια στελέχη της εταιρίας να δράσουν για την αποφυγή μείωσης του μεριδίου τους στην αγορά.

Στην εργασία αυτή, ως απώλεια πελάτη θεωρούμε τον πελάτη ο οποίος παραμένει ανενεργός για διάστημα δύο μηνών και πάνω από την δημιουργία του λογαριασμού του. Ο πελάτης αυτός χαρακτηρίζεται ως απώλεια και έχει την ένδειξη 1 στην τελευταία στήλη των δεδομένων, ενώ ο πελάτης που χαρακτηρίζεται ως ενεργός έχει την ένδειξη 0 στην τελευταία στήλη των δεδομένων.

## 1.2 Σκοπός και στόχος

Σε έναν κλάδο με πολύ υψηλό ανταγωνισμό, όπως είναι ο διαδικτυακός στοιχηματισμός, είναι υψίστης σημασίας η κατανόηση της συμπεριφοράς και των απαιτήσεων των πελατών της εταιρίας. Η μηχανική μάθηση δίνει την δυνατότητα στις εταιρίες να γνωρίσουν καλύτερα την συμπεριφορά των πελατών τους και να λαμβάνουν

αποφάσεις οι οποίες βασίζονται στην ανάλυση δεδομένων, κάτι που συμβάλλει σημαντικά στην δημιουργία πιο προσωποποιημένων υπηρεσιών και εμπειριών κατά την δραστηριότητα των πελατών τους. Επιπλέον, η στρατηγική της λήψης αποφάσεων οι οποίες είναι βασισμένες στην ανάλυση δεδομένων επιτρέπει την προσέλκυση αλλά και τη διατήρηση των πελατών. Συνεπώς, ο σκοπός αυτής της έρευνας είναι να δημιουργηθεί ένα μοντέλο το οποίο θα προβλέπει την πιθανότητα απώλειας πελάτη, το οποίο μπορεί να ενισχύσει τις προσπάθειες του τμήματος Marketing για πιο εστιασμένες κινήσεις ώστε να μειωθεί το ποσοστό των πελατών που διακόπτουν την δραστηριότητά τους στην εταιρία. Πιο συγκεκριμένα, ο στόχος αυτής της έρευνας είναι να αναπτυχθεί και να εκπαιδευτεί ένα μοντέλο μηχανικής μάθησης το οποίο θα προβλέπει την πιθανότητα απώλειας πελάτη με την μεγαλύτερη δυνατή ακρίβεια χρησιμοποιώντας δεδομένα έξι μηνών. Αναλυτικότερα, η μελέτη αυτή χρησιμοποιεί έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης ώστε να λυθεί το πρόβλημα δυαδικής ταξινόμησης (binary classification problem), εάν δηλαδή ο πελάτης θα διακόψει την δραστηριότητά του ή όχι.

## 1.3 Δομή

Η δομή της εργασίας αυτής είναι αρχικά να παρουσιάσει την υπάρχουσα βιβλιογραφία στην πρόβλεψη απώλειας πελάτη στο δεύτερο κεφάλαιο, έπειτα στο τρίτο κεφάλαιο περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε για την δημιουργία του μοντέλου πρόβλεψης, στη συνέχεια στο τέταρτο κεφάλαιο αναλύονται τα αποτελέσματα της έρευνας και η έρευνα ολοκληρώνεται παραθέτοντας τα συμπεράσματα που προέκυψαν από την ανάλυση.

# 2 Ανασκόπηση βιβλιογραφίας

Στο σημερινό οικονομικό περιβάλλον, όπου ο ανταγωνισμός παίζει κύριο ρόλο στην ανάπτυξη μιας εταιρίας, το τμήμα διαχείρισης πελατειακών σχέσεων (Customer Relationship Management -CRM) καλείται να διαχειριστεί τις απώλειες πελατών,

δηλαδή να εντοπίσει τους πελάτες που είναι πιθανό να "μεταφερθούν" από τον ένα πάροχο μίας υπηρεσίας ή ενός προϊόντος σε έναν άλλο. Το πρόβλημα αυτό μπορεί να αναλυθεί σε δύο μέρη. Το πρώτο είναι να προβλεφθούν οι πελάτες που είναι πιθανό να στραφούν στον ανταγωνισμό και το δεύτερο είναι να αξιολογήσουμε τις δυνατότητες που έχουμε ώστε να διατηρήσουμε τους πελάτες (Ballings & Van den Poel, 2022), (Hung, et al., 2006). Τα δύο αυτά στάδια είναι ιδιαίτερα δύσκολα να εφαρμοστούν, ειδικά στις περιπτώσεις των εταιριών που δεν υπάρχει κάποιο συμβόλαιο μεταξύ της εταιρίας και των πελατών της (non-contractual). Σε αυτήν την κατηγορία ανήκει και ο κλάδος του διαδικτυακού στοιχηματισμού, που θα μελετήσουμε σε αυτή την εργασία. Στις εταιρίες που έχουν συμβόλαια με τους πελάτες τους, ο ορισμός ενός πελάτη ως churner είναι ευκολότερος καθώς, μπορεί να προκύψει από τις αλλαγές στο υφιστάμενο συμβόλαιο τους, δηλαδή είτε την λήξη του συμβολαίου χωρίς ανανέωση, είτε την ακύρωση ή διακοπή του ενεργού συμβολαίου (Ascarza, et al., 2018). Αντίθετα, για την εταιρίες που δεν συνάπτουν συμβόλαιο με τους πελάτες τους, ο καθορισμός ενός πελάτη πως έχει φύγει δεν συνοδεύεται από κάποια ενημέρωση του πελάτη προς την εταιρία, ούτε με την λήξη ή διακοπή κάποιου συμβολαίου. Οι εταιρίες αυτές μπορεί να παρατηρήσουν μια μείωση της δραστηριότητας των πελατών τους, αλλά για να συνδεθεί αυτή η μείωση με την πιθανότητα απώλειας του πελάτη απαιτείται μεγαλύτερη προσπάθεια, ώστε να εντοπιστούν και άλλα χαρακτηριστικά στην συμπεριφορά του πελάτη που επιτρέπουν να καταλήξουμε σε αυτό το συμπέρασμα. Αυτό οδηγεί στην υιοθέτηση υποκειμενικών κριτηρίων από την πλευρά της εταιρίας, βάσει των οποίων ορίζει αν κάποιος πελάτης έχει διακόψει την δραστηριότητα του ή όχι (Kaya, et al., 2018). Κατά κύριο λόγο, ο ορισμός της απώλειας πελάτη περιλαμβάνει δύο βασικές παραμέτρους, την δραστηριότητα του πελάτη και κάποιο όριο που έχει τεθεί βάσει των αρχών της εταιρίας (Clemente-Císcar, et al., 2014).

Όπως αναφέραμε, στις περιπτώσεις εταιριών που δεν υφίσταται κάποιο συμβόλαιο με τους πελάτες τους, και λόγω της δυσκολίας που υπάρχει στο να οριστεί εάν ένας πελάτης έχει διακόψει την δραστηριότητά του ή όχι, οδηγούμαστε στη χρήση υποκειμενικών κριτηρίων από την πλευρά της επιχείρησης και του αναλυτή που διεξάγει την πρόβλεψη. Ο τρόπος για να ελαχιστοποιηθεί όσο το δυνατό περισσότερο η υποκειμενικότητα που εισέρχεται στο πρόβλημα είναι να οριστεί με ακρίβεια και νόημα ο χαρακτηρισμός ενός πελάτη ως churner. Παρακάτω, αναφέρονται κάποιοι ορισμοί

που έχουν δοθεί από μελέτες του συγκεκριμένου θέματος, που έχουν διεξαχθεί μετά το 2010.

Ο ορισμός που προτάθηκε από τον Karnstedt και τους συνεργάτες του (Karnstedt, et al., 2010) είχε διάφορες παραλλαγές και εφαρμογή στον τομέα των κοινωνικών δικτύων, που όμως είναι εφαρμόσιμος σε κλάδους ευρείας εμβέλειας. Ο γενικός ορισμός που έδωσαν στηρίζεται στην αξιοσημείωτη και διατηρούμενη αλλαγή της δραστηριότητας του ατόμου ή /και της κοινωνίας. Επίσης, υποστηρίζουν ότι ο τρόπος που ορίζεται αν ένας πελάτης έχει διακόψει την δραστηριότητα του ή όχι, το χρονικό διάστημα που μελετάται η παρελθοντική του δραστηριότητα, το χρονικό διάστημα μέσα στο οποίο ερευνάται αν ο πελάτης έχει διακόψει την δραστηριότητά του ή όχι, ο τύπος της δραστηριότητας που έχει χρησιμοποιηθεί αλλά και κατάλληλο σύστημα παραγόντων, θα πρέπει να βασίζονται στον τομέα εφαρμογής και στα ρίσκα που προκύπτουν από την ανάλυση και την πρόβλεψη της απώλειας πελάτη. Αναλυτικότερα, στον ατομικό τύπο απώλειας πελάτη, όπως τον χαρακτήρισαν οι μελετητές, ένας πελάτης θεωρείται πως έχει διακόψει την δραστηριότητά του όταν η μέση δραστηριότητά του για ένα συγκεκριμένο χρονικό διάστημα έχει μειωθεί σε λιγότερο από ένα ποσοστό της μέσης δραστηριότητάς του το προηγούμενο διάστημα.

Ο Clemente και οι συνεργάτες του (Clemente-Císcar, et al., 2014) χρησιμοποίησαν μια παρόμοια προσέγγιση με τους Buckinx και Van den Poel (Buckinx & Van den Poel, 2005). Ορίζουν έναν πελάτη ως "churner" αν μεταβεί από την κατάσταση της πιστής συνεργασίας σε μια κατάσταση πιθανού αποχαιρετισμού ή βελτιώσιμης συνεργασίας. Από την άλλη πλευρά, κατηγοριοποιούν τους "πιστούς πελάτες" ως αυτούς που κάνουν συχνές αγορές (υπεράνω της μέσης συχνότητας) και εμφανίζουν μια συνεπή μορφή αγοραστικής συμπεριφοράς (με συντελεστή μεταβολής του διαστήματος μεταξύ των αγορών υποκάτω της μέσης τιμής). Η μελέτη αξιολόγησε τον αντίκτυπο κάθε ορισμού churn στο κέρδος και την απόδοση μιας εταιρείας, και εξέτασε περαιτέρω πως η επιλογή του ορισμού churn επηρέασε την οικονομική επίδοση της εταιρείας μακροπρόθεσμα.

Ορίζοντας την απώλεια πελατών (churn) ως αντίστοιχη της διατήρησης πελατών, ο Ascarza και οι συνεργάτες του (Ascarza, et al., 2018) εξέτασαν το θέμα, όπου η διατήρηση καθορίζεται για να περιλαμβάνει τις έννοιες της "συνέχειας" και της "συμπεριφοράς". Η διατήρηση πελατών σχετίζεται στενά με τη συνέχεια, καθώς οι

πελάτες συνεχίζουν να αλληλεπιδρούν με την εταιρεία. Αποτελεί μια μορφή πελατειακής συμπεριφοράς που οι εταιρείες προσπαθούν να διαχειριστούν. Αντίθετα, ο όρος "απώλειας πελάτη" (churner) αναφέρεται σε έναν πελάτη που έχει αποφασίσει να σταματήσει να συνεργάζεται με την εταιρία.

Η πρόβλεψη απώλειας πελάτη αποτελεί ένα σημαντικό θέμα για τις εταιρίες και έχει μελετηθεί ιδιαίτερα καθώς μπορεί να συμβάλει σε μεγάλο βαθμό στην διατήρηση και βελτίωση των σχέσεων μεταξύ μίας εταιρίας και των πελατών τους. Οι εταιρίες έχουν σημαντικά οικονομικά κίνητρα να δημιουργήσουν μεθόδους οι οποίες θα μπορούν να αναγνωρίσουν περιπτώσεις στις οποίες υπάρχει ένδειξη πως είναι πολύ πιθανό ο πελάτης να διακόψει τις συναλλαγές του με την εταιρία, ώστε να μπορέσει να λάβει δράση εγκαίρως και να το αποτρέψει. Τα κίνητρα, όμως, για την πρόβλεψη της απώλειας πελάτη δεν είναι μόνο άμεσα οικονομικά αλλά μπορούν πιο επηρεάσουν θετικά και στην φήμη της εταιρίας. Πιο αναλυτικά, υπάρχουν αρκετοί λόγοι που εξηγούν γιατί είναι προτιμότερο οι εταιρίες να εστιάσουν στην διατήρηση του πελατολογίου τους, συγκριτικά με την προσέλκυση νέων πελατών. Αρχικά, διατηρώντας μακροχρόνιες σχέσεις με τους πελάτες τους, είναι σε θέση να εστιάζουν στις ανάγκες και τις προτιμήσεις τους, και όχι αποκλειστικά στην απόκτηση νέων πελατών, οι οποίοι συνήθως εμφανίζουν μεγαλύτερο ποσοστό απώλειας (Dawes & Swailes, 1999) (Reinartz & Kumar, 2003). Επιπλέον, οι πελάτες που διακόπτουν τη χρήση προϊόντων ή υπηρεσιών μιας εταιρίας είναι πιθανό να επηρεάσουν και άλλους πελάτες της εταιρίας αυτής να πράξουν το ίδιο (Nitzan & Libai, 2011). Ένας άλλος θετικός παράγοντας είναι πως ένας ικανοποιημένος πελάτης είναι πολύ πιθανό να επηρεάσει νέους πελάτες στην χρήση προϊόντων ή υπηρεσιών της εταιρίας, συμβάλλοντας με αυτόν τον τρόπο στην αύξηση της κερδοφορίας της εταιρίας (Ganesh, et al., 2000). Επίσης, έχει φανεί πως οι ενέργειες μάρκετινγκ έχουν μικρότερη επίδραση στους μακροχρόνιους πελάτες κάτι που μπορεί να μειώσει το κόστος διατήρησης αυτών των πελατών (Colgate, et al., 1996). Τέλος, η απώλεια πελατών αυξάνει την ανάγκη απόκτησης νέων πελατών και μειώνει τα κέρδη από την απώλεια των συγκεκριμένων πελατών. Από αυτό, προκύπτει ότι το κόστος διατήρησης των υπαρχόντων πελατών είναι αρκετές φορές μικρότερο από την απόκτηση νέων (Torkzadeh, et al., 2006). Συνεπώς, η δυνατότητα πρόβλεψης της πιθανής απώλειας πελάτη είναι αναγκαία στην στρατηγική διατήρησης πελατών.

Οι παραπάνω λόγοι οδήγησαν στην μελέτη του συγκεκριμένου θέματος σε μεγάλη κλίμακα και με εφαρμογή σε σημαντικό εύρος επιχειρήσεων όπως είναι οι τηλεπικοινωνίες (Qureshi, et al., 2013) (Kim, et al., 2014) (Kirui, et al., 2013) (Jadhav & Pawar, 2011) (Richter, et al., 2010) (Kraljević & Gotovac, 2010) (Tsai & Lu, 2009) (Brandusoiu, et al., 2016), τα χρηματοπιστωτικά ιδρύματα (Xie & Li, 2008) (Prasad & Madhavi, 2012) (De Bock & Van den Poel, 2012) (Tang, et al., 2014),ο κλάδος των ασφαλίσεων (Günther, et al., 2011) και η ενέργεια (Moeyersoms & Martens, 2015). Για την πρόβλεψη απώλειας πελάτη έχουν προταθεί αρκετοί αλγόριθμοι μηχανικής μάθησης όπως είναι τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks-ANN), τα Δέντρα Αποφάσεων (Decision Trees), η ανάλυση Γραμμικής και Λογιστικής Παλινδρόμησης (Linear and Logistic Regression) και η Naïve Bayes. Ωστόσο, η μελέτη την πρόβλεψης απώλειας πελάτη είναι κάτι που έχει μελετηθεί σε πολύ μικρότερο βαθμό στον κλάδο του διαδικτυακού στοιχηματισμού. Στο παρελθόν οι μελέτες στον κλάδο αυτό εστίαζαν κυρίως στις ψυχολογικές πτυχές της συμπεριφοράς των παικτών (Bleichrodt & Schmidt, 2002) (Lam, 2006) (Lam, 2007) (McDaniel & Zuckerman, 2003) (Mowen, et al., 2009), στην δυνατότητα πρόβλεψης των αποτελεσμάτων των αθλητικών γεγονότων (Stekler, et al., 2010) και στην ατομική στοιχηματική συμπεριφορά των παικτών (Andrade & Iyer, 2009) (Grant & Xie, 2007) (Seybert & Bloomfield, 2009) (Smith, et al., 2009).

Οι μελέτες που έχουν γίνει για τον καθορισμό ενός μοντέλου πρόβλεψης απώλειας πελάτη στο διαδικτυακό στοιχηματισμό είναι περιορισμένες και αυτό οφείλεται στο γεγονός ότι ο διαδικτυακός στοιχηματισμός έχει γίνει ευρέως διαδεδομένος τα τελευταία χρόνια όπου η χρήση των "έξυπνων" κινητών (smartphones) έχει αναπτυχθεί σε σημαντικό βαθμό. Παρακάτω γίνεται αναλυτική αναφορά των μελετών που έχουν εστιάσει στον κλάδο του διαδικτυακού στοιχηματισμού.

## 2.1 Δεδομένα

Ως προς τις υπάρχουσες μελέτες στον κλάδο του διαδικτυακού στοιχηματισμού, τα δεδομένα που έχουν χρησιμοποιηθεί περιλαμβάνουν τόσο δημογραφικά στοιχεία, όσο και στοιχεία που προκύπτουν από την παικτική συμπεριφορά των χρηστών. Πιο συγκεκριμένα, οι Lovisa Gronros και Ida Janer (Gronros & Janer, 2018) στην μελέτη

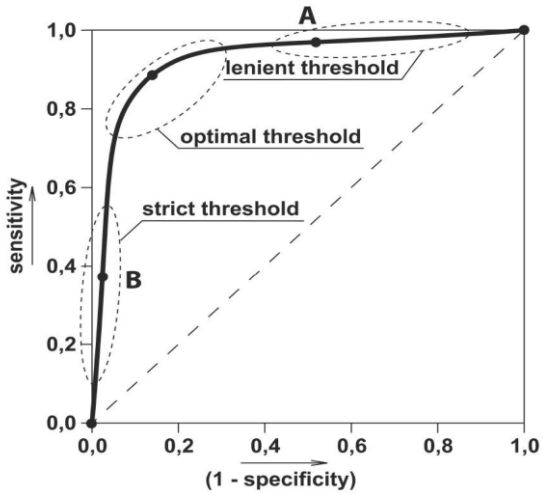
τους χρησιμοποίησαν δεδομένα 24 ωρών σε επίπεδο πελάτη που περιλάμβαναν πληροφορίες όπως είναι το φύλο, η ηλικία και η χώρα, αλλά και δεδομένα όπως ο συνολικός αριθμός και η μέση τιμή των καταθέσεων, ο αριθμός των δελτίων που έχουν παίξει και των αναλήψεων που έχουν κάνει οι πελάτες, οι πλατφόρμες από τις οποίες επιλέγουν οι πελάτες να χρησιμοποιήσουν τις υπηρεσίες, όπως και η ημερομηνίες και οι ωρες της πρώτης κατάθεσης τους. Όπως αναφέρουν οι συγγραφείς λόγω του κινδύνου εμφάνισης συσχέτισης στα δεδομένα, επέλεξαν να χρησιμοποιήσουν τους μέσους όρους κάποιων μεταβλητών όπως το σύνολο των χρημάτων που κερδήθηκαν ή χάθηκαν από τα δελτία στοιχηματισμού και τα δελτία στοιχηματισμού που έγιναν από διαφορετικά κανάλια (ιστοσελίδα, ιστοσελίδα από τον κινητό, Android, iOS). Αντίστοιχα, Eunju Suh και Matt Alhaery 2016 (Suh & Alhaery, 2016) που μελέτησαν την πρόβλεψη απώλειας πελάτη στο διαδικτυακό καζίνο, χρησιμοποίησαν δεδομένα από 26939 παίκτες για διάστημα 7 μηνών (Μάρτιος – Σεπτέμβριος 2015),που αφορούν τόσο το φύλο και την ηλικία των παικτών και την απόσταση από επίγεια καζίνο, όσο και την συχνότητα με την οποία παίζουν, το ποσό που δαπανούν, τα κέρδη ή τις ζημίες, τον αριθμό των ημερών από την πρώτη κατάθεση και τα ποσά που έχουν λάβει ως bonus. Οι Florian Merchie και Ernst Damien (Merchie & Ernst, 2022) μελέτησαν το θέμα αυτό χρησιμοποιώντας δεδομένα 24 ωρών που περιλάμβαναν τον αριθμό των παιχνιδιών στο καζίνο και των δελτίων που τοποθετήθηκαν στο στοίχημα, τα ακαθάριστα έσοδα από το καζίνο και το στοίχημα, τον αριθμό των καταθέσεων και αναλήψεων, τον αριθμό επισκέψεων στην ιστοσελίδα και τον αριθμό των ημερών από την τελευταία μέρα που υπήρξε δραστηριότητα του πελάτη. Τέλος, οι Kristof Coussement και Koen W. De Bock (Coussement & De Bock, 2013) μελέτησαν την πρόβλεψη απώλειας πελάτη με τη χρήση δεδομένων δύο ετών (Φεβρουάριος 2005 – Φεβρουάριος 2007) από 3729 πελάτες της bwin τα οποία αφορούν σε δημογραφικά χαρακτηριστικά όπως είναι η γλώσσα, η περιοχή και το φύλλο, όπως επίσης και συμπεριφορικές μεταβλητές που έχουν κατηγοριοποιηθεί σε μεταβλητές που εξηγούν πόσο πρόσφατα (Recency) και πόσο συχνά (Frequency) υπήρχε δραστηριότητα των παικτών και ποια είναι η χρηματική αξία (Monetary value) που προκύπτει από τον κάθε παίκτη. Πιο συγκεκριμένα, χρησιμοποιήθηκαν μεταβλητές όπως είναι ο αριθμός των ημερών από την τελευταία παικτική δραστηριότητα και το τελευταίο κέρδος ή ζημία, ο αριθμός των ημερών από την πρώτη κατάθεση και στοιχηματισμό, ο αριθμός των στοιχηματικών δελτίων την τελευταία εβδομάδα και τον τελευταίο μήνα, τα συνολικά κέρδη ή ζημίες και αν έχει λάβει bonus.

## 2.2 Μοντέλα

Η αντιμετώπιση του προβλήματος της απώλειας πελάτη μέσω της μηχανικής μάθησης γίνεται με τη χρήση μοντέλων που αφορούν στο πρόβλημα της δυαδικής ταξινόμησης (binary classification problem). Στόχος των μοντέλων είναι η υψηλότερη και ορθότερη πρόβλεψη ταξινόμησης της μεταβλητής στόχου, εάν δηλαδή ο πελάτης θα διακόψει τις συναλλαγές του με την εταιρεία ή όχι. Οι αλγόριθμοι πρόβλεψης της απώλειας πελάτη που έχουν χρησιμοποιηθεί στην βιβλιογραφία περιλαμβάνουν τόσο μεμονωμένους αλγόριθμους, όπως είναι τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Network RNN) ( (Merchie & Ernst, 2022), τα δέντρα αποφάσεων (E-CHAID decision tree) (Suh & Alhaery, 2016) αλλά και συνδυασμό βασικών μοντέλων όπως την δημιουργία stacking classifier με την χρήση των αποτελεσμάτων μοντέλων λογιστικής παλινδρόμησης (logistic regression), τυχαίων δέντρων (random forest) και linear discriminant analysis (LDA) (Gronros & Janer, 2018), όπως επίσης και η χρήση ensemble learners, που επιτρέπουν την καλύτερη και πιο εύρωστη πρόβλεψη συγκριτικά με τη χρήση μεμονωμένων αλγόριθμων (Coussement & De Bock, 2013).

## 2.3 Αξιολόγηση μετρήσεων/μοντέλων

Ολοκληρώνοντας το στάδιο της πρόβλεψης, ακολουθεί η αξιολόγηση των αποτελεσμάτων που προκύπτουν από τα μοντέλα πρόβλεψης, ώστε να κριθεί η αξιοπιστία τους. Οι μετρικές των μοντέλων που χρησιμοποιούνται κατά κύριο λόγο περιλαμβάνουν τον υπολογισμό της ακρίβειας της πρόβλεψης (accuracy) (Gronros & Janer, 2018) (Suh & Alhaery, 2016) (Merchie & Ernst, 2022), δηλαδή την αναλογία των αληθών αποτελεσμάτων (αληθώς θετικά και αληθώς αρνητικά) μεταξύ του συνολικού αριθμού των υποθέσεων που εξετάστηκαν, αλλά και της καμπύλης AUC (Area Under the ROC Curve) (Suh & Alhaery, 2016) όπως φαίνεται στην Εικόνα 1, η οποία δείχνει πόσο καλά οι θετικές κλάσεις είναι διαχωρισμένες από τις αρνητικές κλάσεις (Rahul, 2019 ).




*Εικόνα 1 AUC*

Παράλληλα με την ακρίβεια της πρόβλεψης όμως χρησιμοποιούνται και οι μετρικές της ευστοχίας (precision) και ανάκλησης (recall) (Merchie & Ernst, 2022). Πιο συγκεκριμένα, η ευστοχία (precision) αφορά στην αναλογία των αληθώς θετικών αποτελεσμάτων μεταξύ του συνολικού αριθμού θετικών αποτελεσμάτων (αληθώς θετικά και ψευδώς θετικά), ενώ η ανάκληση αφορά στην αναλογία των αληθώς θετικών αποτελεσμάτων μεταξύ του συνολικού αριθμού πραγματικά θετικών αποτελεσμάτων (αληθώς θετικών και ψευδώς αρνητικών) (Rahul, 2019 ).

*Πίνακας 1 Confusion matrix*

| | | **Actual Class** | |
|---|---|---|---|
| | | *Positive* | *Negative* |
| **Predicted Class** | *Positive* | True Positive (TP) | False Positive (FP) |
| | *Negative* | False Negative (FN) | True Negative (TN) |

Παρακάτω φαίνονται οι τύποι που υπολογίζονται οι μετρικές για την αξιολόγηση των μοντέλων:

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

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

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

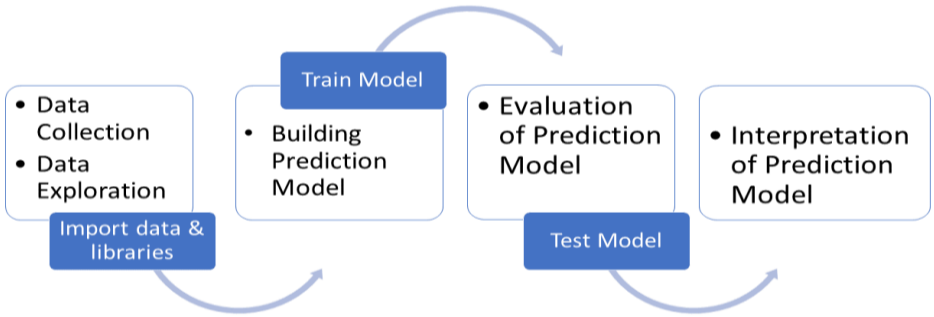
Κάποιες άλλες μετρικές για την αξιολόγηση των μοντέλων πρόβλεψης που έχουν χρησιμοποιηθεί είναι η Top-Decile List (TDL) και η Lift Index (LI), οι οποίες μετρούν την ικανότητα του μοντέλου πρόβλεψης να παράγει μία καλή κατάταξη των πελατών με βάση τις πιθανότητες απώλειας των πελατών που έχουν προβλεφθεί (Coussement & De Bock, 2013). Στον *Πίνακα 2* αναφέρονται συγκεντρωτικά οι μελέτες που έχουν γίνει στον κλάδο του διαδικτυακού στοιχηματισμού.

*Πίνακας 2 Σύνοψη βιβλιογραφικών αναφορών στον κλάδο του διαδικτυακού στοιχηματισμού*

| Authors | Title | Model | Metrics |
|---|---|---|---|
| **(Coussement & De Bock, 2013)** | *Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning* | CART decision tree, Generalized Additive Model, Ensemble learners | TDL, LI |
| **(Suh & Alhaery, 2016)** | *Customer retention:Reducing Online Casino Player Churn Through the Application of Predictive Modeling* | E-CHAID Decision Tree | Accuracy, AUC |
| **(Gronros & Janer, 2018)** | *Predicting Customer Churn in the iGmaing Industry using Supervised Machine Learning* | Logistic Regression, Random Forest, Linear Discriminant Analysis, Stacking Classifier, Voting Classifier | Accuracy |
| **(Merchie & Ernst, 2022)** | *Churn Prediction in online gambling* | Recurrent Neural Network RNN | Accuracy, Precision, Recall |

# 3 Μεθοδολογία

Στο παρακάτω διάγραμμα φαίνονται τα βασικά βήματα που ακολουθούμε σε αυτή την ενότητα για την δημιουργία του μοντέλου πρόβλεψης απώλειας πελάτη.



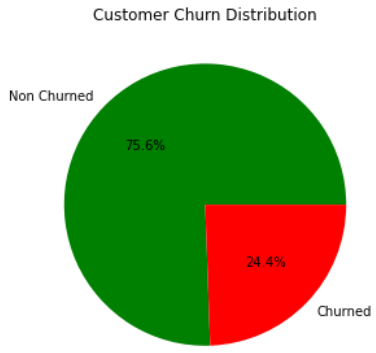

*Εικόνα 2 Workflow of Machine Learning Process*

## 3.1 Συλλογή και ανάλυση δεδομένων

Για την υλοποίηση της εργασίας έγινε χρήση της γλώσσας προγραμματισμού Python, στο περιβάλλον του Google Cloud Platform (GCP), καθώς το μέγεθος των δεδομένων δεν επέτρεπε να χρησιμοποιηθεί τοπικό περιβάλλον επεξεργασίας. Τα δεδομένα που χρησιμοποιούνται στη μελέτη αυτή προέρχονται από την εταιρεία Kaizen Gaming, η οποία δραστηριοποιείται σε 16 χώρες μεταξύ των οποίων είναι η Ελλάδα, η Κύπρος, ο Καναδάς, η Βραζιλία και άλλες. Τα δεδομένα έχουν ληφθεί από την βάση δεδομένων της εταιρείας ως ένα αρχείο με την μορφή CSV, διατηρώντας την ανωνυμία και την προστασία των προσωπικών δεδομένων των πελατών της και χωρίς να μπορεί με τον οποιονδήποτε τρόπο να γίνει συσχέτιση με κάποιον πελάτη. Το σύνολο των δεδομένων αποτελείται από 202.666 παρατηρήσεις και ένα σύνολο 35 μεταβλητών που περιλαμβάνουν τόσο δημογραφικά δεδομένα όπως η ηλικία, όσο και δεδομένα που περιγράφουν την παικτική δραστηριότητα των πελατών όπως είναι ο αριθμός των ημερών που ο χρήστης είναι ενεργός, ο αριθμός καταθέσεων και αναλήψεων κ.α.. Τα δεδομένα έχουν καταγραφεί ανά πελάτη και η περιγραφή όλων των μεταβλητών βρίσκεται στο *Παράρτημα 1*.

Συνεχίζοντας στη διαδικασία της διερευνητικής ανάλυσης δεδομένων (Exploratory Data Analysis - EDA), το πρώτο βήμα είναι εισάγουμε τις απαραίτητες βιβλιοθήκες της Python που θα μας επιτρέψουν να επεξεργαστούμε τα δεδομένα, να πάρουμε χρήσιμες πληροφορίες για αυτά και στην συνέχεια να δημιουργήσουμε το μοντέλο που θα μας επιτρέψει να κάνουμε προβλέψεις. Όλες οι βιβλιοθήκες της Python που χρησιμοποιήθηκαν παρουσιάζονται στο *Παράρτημα 2*. Έπειτα, εισάγουμε τα δεδομένα που θα αναλύσουμε και στα οποία θα βασιστεί το μοντέλο πρόβλεψης. Στην βάση δεδομένων της εταιρίας, τα ποσά των αναλήψεων καταγράφονται ως αρνητικές τιμές καθώς αποτελούν έξοδα, για τον λόγο αυτό θα κρατήσουμε τις απόλυτες τιμές αυτών των μεταβλητών.

Η διερευνητική ανάλυση μας επιτρέπει να έχουμε μια καλύτερη εικόνα για τα δεδομένα που έχουμε στη διάθεσή μας. Πιο αναλυτικά, στην *Εικόνα 3* φαίνεται η κατανομή της μεταβλητής στόχου και παρατηρούμε πως δεν υπάρχει ισορροπία μεταξύ των δύο τιμών που μπορεί να πάρει η μεταβλητή καθώς η τιμή 0 (Non Churned) συναντάται σε ποσοστό 75.6%, ενώ η τιμή 1 (Churned) συναντάται σε ποσοστό 24,4%, κάτι που υποδηλώνει ανισορροπία στην κατανομή της μεταβλητής στόχου.




*Εικόνα 3 Κατανομή μεταβλητής στόχου*

Ως προς τις μεταβλητές που αφορούν στο πλήθος και τα ποσά που καταθέτουν και λαμβάνουν ως αναλήψεις οι πελάτες, στους *Πίνακες 3* και *4*, αντίστοιχα, παρουσιάζονται τα περιγραφικά στατιστικά αυτών των μεταβλητών. Στα αρχικά δεδομένα, οι τιμές που έχουν στα ποσά των αναλήψεων έχουν αρνητικές τιμές καθώς αποτελούν έξοδο για την εταιρία, για τον λόγο αυτό χρησιμοποιήσαμε τις απόλυτες τιμές των μεταβλητών αυτών ώστε να

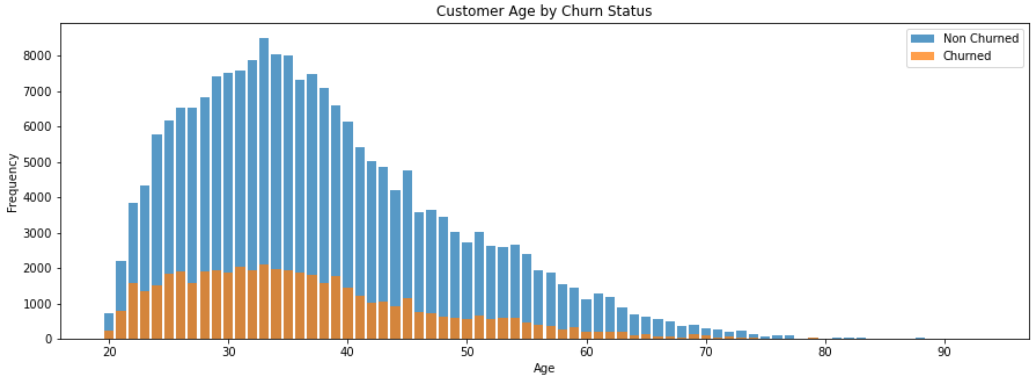
*Πίνακας 3 Περιγραφικά στατιστικά για τα ποσά και το πλήθος των καταθέσεων*

| | **Last Month** | | **Last 3 Months** | | **Last 6 Months** | |
|---|---|---|---|---|---|---|
| | Deposit Number | Deposit Amount | Deposit Number | Deposit Amount | Deposit Number | Deposit Amount |
| Mean | 3 | 189.50 | 8 | 496.00 | 13 | 815.00 |
| Median | 1 | 30.00 | 2 | 80.00 | 4 | 140.00 |
| Maximum | 361 | 115,984.00 | 960 | 241,255.50 | 1,570 | 542,974.00 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 |

*Πίνακας 4 Περιγραφικά στατιστικά για τα ποσά και το πλήθος των αναλήψεων*

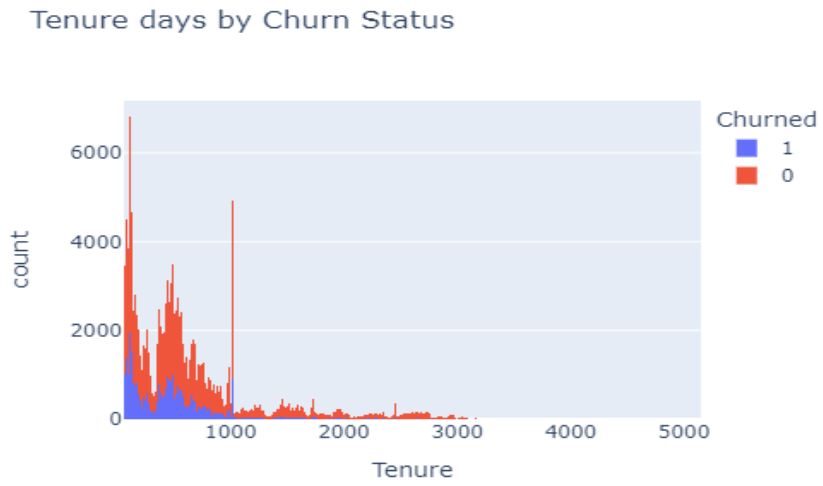
| | **Last Month** | | **Last 3 Months** | | **Last 6 Months** | |
|---|---|---|---|---|---|---|
| | Withdrawal Number | Withdrawal Amount | Withdrawal Number | Withdrawal Amount | Withdrawal Number | Withdrawal Amount |
| Mean | 1 | 102.00 | 2 | 266.00 | 2 | 444.00 |
| Median | 0 | 0 | 0 | 0 | 1 | 21.00 |
| Maximum | 66 | 89,347.00 | 169 | 182,490.00 | 388 | 434,264.00 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 |

Το επόμενο βήμα είναι να δούμε την κατανομή της ηλικίας των πελατών βάσει του διαχωρισμού τους σε αυτούς που έχουν αποχωρήσει (Churned) και σε αυτούς που παραμένουν πελάτες (Non Churned). Όπως φαίνεται στην Εικόνα 4 παρατηρούμε μεγαλύτερη συγκέντρωση στις ηλικίες μεταξύ 22 και 40 ετών και για τις δύο κατηγορίες πελατών.



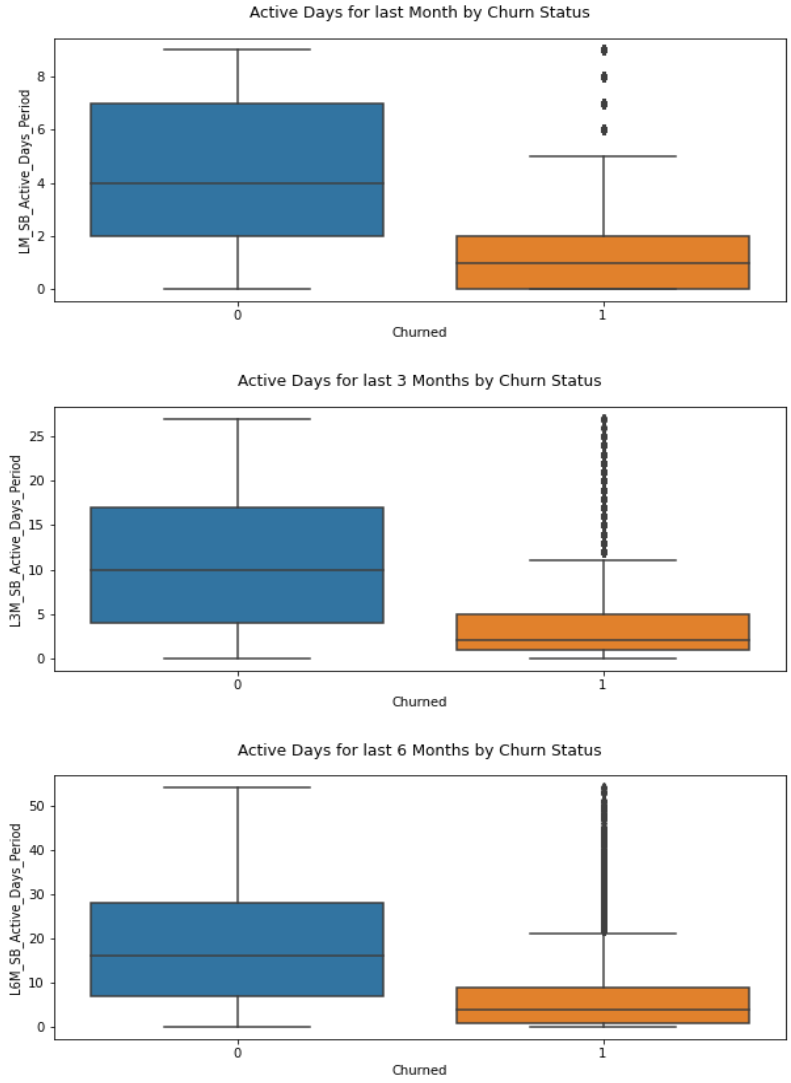

*Εικόνα 4 Κατανομή της ηλικίας ανά τιμές μεταβλητής στόχου*

Αντίστοιχα, εξετάζουμε την χρονική περίοδο σε ημέρες από όταν ο πελάτης έχει κάνει εγγραφή στην σελίδα ή την εφαρμογή βάσει του διαχωρισμού τους σε αυτούς που έχουν αποχωρήσει (Churned=1) και σε αυτούς που παραμένουν πελάτες (Non Churned=0). Όπως φαίνεται στην *Εικόνα 5* παρατηρούμε πως οι πελάτες που αποφασίζουν να αποχωρήσουν είναι κυρίως αυτοί που είναι λιγότερες από 1000 ημέρες χρήστες της πλατφόρμας και της ιστοσελίδας.



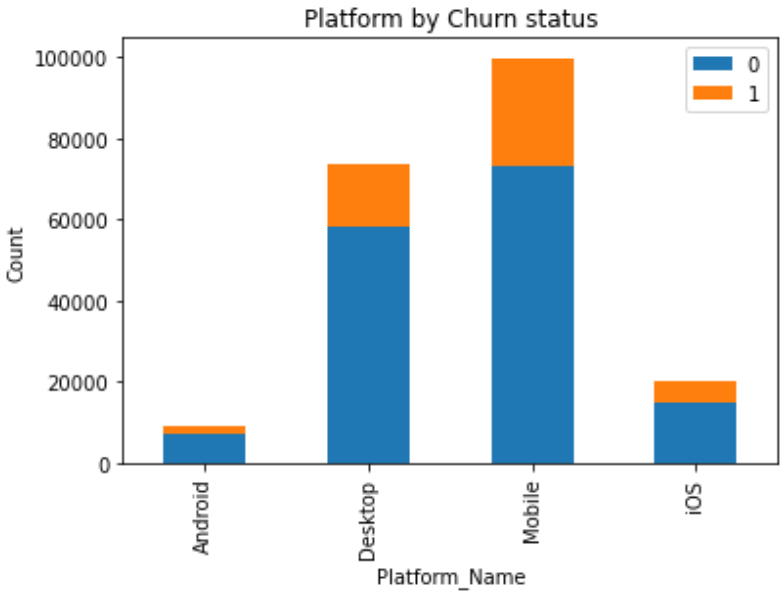

*Εικόνα 5 Κατανομή των ημερών από την εγγραφή ανά τιμές μεταβλητής στόχου*

Στην *Εικόνα 6* παρατηρούμε την διασπορά τριών μεταβλητών, ενεργές ημέρες τον τελευταίο μήνα, το τελευταίο τρίμηνο και το τελευταίο εξάμηνο ανά τιμή της μεταβλητής στόχου. Όπως μπορούμε να συμπεράνουμε και για τα τρία χρονικά διαστήματα, παρατηρούμε ότι ο αριθμός των ενεργών ημερών των πελατών που έχουν χαρακτηριστεί ως πελάτες που αποχωρούν από την εταιρία είναι σημαντικά μικρότερος συγκριτικά με τους πελάτες που παραμένουν. Αυτό αποτελεί μια αρχική ένδειξη πως όταν μειώνεται ο αριθμός των ενεργών ημερών αυξάνονται οι πιθανότητες ένας πελάτης να στραφεί στον ανταγωνισμό.




*Εικόνα 6 Ενεργές ημέρες για τον τελευταίο μήνα, τελευταίο τρίμηνο και τελευταίο εξάμηνα ανά τιμή μεταβλητής στόχου*

Τέλος, μελετώντας την μεταβλητή της πλατφόρμας μέσω της οποίας έχει εγγραφεί κάποιος πελάτης στην *Εικόνα 7*, παρατηρούμε μεγαλύτερη συγκέντρωση να έχει το κινητό τηλέφωνο και έπειτα ο ηλεκτρονικός υπολογιστής, ενώ δεν παρατηρείται να αλλάζει η αναλογία τόσο στους πελάτες που παραμένουν στην εταιρία (1), όσο και σε αυτούς που αποφασίζουν να στραφούν στον ανταγωνισμό (0).




*Εικόνα 7 Πλατφόρμα εγγραφής πελατών ανά τιμή μεταβλητής στόχου*

## 3.2 Προετοιμασία και εφαρμογή μοντέλου

Στο κεφάλαιο αυτό αναλύεται ο τρόπος που αναπτύχθηκε το μοντέλο πρόβλεψης απώλειας πελάτη. Πριν την παράθεση της ανάλυσης, αναφέρεται μια βασική περιγραφή του αλγορίθμου που χρησιμοποιήθηκε για την δημιουργία του μοντέλου, καθώς επίσης, και τους λόγους που συνέβαλαν στην επιλογή του συγκεκριμένου αλγορίθμου για την ανάλυση που πραγματοποιήθηκε.

Ο αλγόριθμος που επιλέχθηκε για την πρόβλεψη του μοντέλου είναι ο XGBoost. Οι αλγόριθμοι που ανήκουν στην κατηγορία gradient tree boosting, όπως είναι και ο XGBoost φαίνεται να παρουσιάζουν πολύ καλά αποτελέσματα σε διαφορετικά πεδία εφαρμογής, έναντι άλλων μεθόδων που χρησιμοποιούνται στην μηχανική μάθηση (Friedman, 2001).

Το σύστημα είναι προσβάσιμο μέσω πακέτου ανοιχτού κώδικα (open-source package). Η θετική του επιρροή έχει αναγνωριστεί ευρέως σε διάφορες προκλήσεις μηχανικής μάθησης και εξόρυξης δεδομένων. Αυτό μπορεί να απεικονιστεί μέσω περιπτώσεων όπως οι διαγωνισμοί μηχανικής εκμάθησης στο Kaggle. Από τις 29 λύσεις που κέρδισαν τους διαγωνισμούς που παρουσιάστηκαν στον ιστότοπο της Kaggle το 2015, οι 17 λύσεις χρησιμοποίησαν αποτελεσματικά τον XGBoost. Μεταξύ αυτών, οκτώ

λύσεις χρησιμοποιούσαν αποκλειστικά τον XGBoost για την εκπαίδευση των μοντέλων, ενώ η πλειοψηφία των άλλων συνδύαζαν τον XGBoost με νευρωνικά δίκτυα με τη χρήση μεθόδου ensemble. Για να παρέχει μια βάση για σύγκριση, η δεύτερη πιο διαδεδομένη προσέγγιση, τα deep neural networks, παρουσιάστηκε σε 11 λύσεις.

Ο πιο σημαντικός παράγοντας πίσω από την επιτυχία του XGBoost είναι η εξαιρετική του κλιμάκωση (scalability) σε όλα τα σενάρια. Το σύστημα λειτουργεί περισσότερο από δέκα φορές ταχύτερα από τις υπάρχουσες δημοφιλείς λύσεις σε ένα μόνο μηχάνημα και κλιμακώνεται σε δισεκατομμύρια παραδείγματα σε κατανεμημένα ή περιορισμένα από μνήμη περιβάλλοντα. Η κλιμάκωση του XGBoost οφείλεται σε αρκετές σημαντικές βελτιστοποιήσεις στα συστήματα και τους αλγορίθμους του. Αυτές οι καινοτομίες περιλαμβάνουν: ένα νέο αλγόριθμο μάθησης δέντρων για τον χειρισμό αραιών δεδομένων. Ο παράλληλος και κατανεμημένος υπολογισμός καθιστά τη μάθηση ταχύτερη, επιτρέποντας πιο γρήγορη εξερεύνηση μοντέλου. Επιπροσθέτως, ο XGBoost αξιοποιεί τον υπολογισμό εκτός πυρήνα, επιτρέποντας στους επιστήμονες δεδομένων να επεξεργαστούν εκατοντάδες εκατομμύρια παραδείγματα σε έναν απλό υπολογιστή. Επιπλέον, ο XGBoost εκμεταλλεύεται την εξωτερική επεξεργασία και επιτρέπει στους επιστήμονες δεδομένων να επεξεργαστούν εκατομμύρια παραδείγματα σε έναν απλό υπολογιστή. Τέλος, είναι ακόμη πιο ενθαρρυντικό ότι συνδυάζει αυτές τις τεχνικές για την δημιουργία ενός ολοκληρωμένου συστήματος που μπορεί να ανταπεξέλθει ακόμη και σε μεγαλύτερα δεδομένα με την ελάχιστη δυνατή χρήση πόρων συστήματος (Chen & Guestrin, 2016).

Επίσης, ο XGBoost συνδυάζει δέντρα παλινδρόμησης και gradient boosting. Σε κάθε δέντρο της διαδικασίας εκπαίδευσης, το υπόλοιπο ενός βασικού ταξινομητή χρησιμοποιείται στον επόμενο ταξινομητή για να βελτιώσει τη συνάρτηση στόχου. Ο αλγόριθμος XGBoost μειώνει την πολυπλοκότητα του μοντέλου και αποτρέπει τα προβλήματα που σχετίζονται με τη υπερπροσαρμογή (overfitting). Τέλος, ο συνδυασμός όλων των δέντρων παρέχει τον τελικό στόχο (Jabeur, et al., 2023).

Απαραίτητο κομμάτι για την δημιουργία του μοντέλου πρόβλεψης απώλειας πελάτη (Customer Churn Prediction - CCP) αποτελεί η προετοιμασία και επεξεργασία δεδομένων, ώστε να είναι κατάλληλα για την δημιουργία του μοντέλου πρόβλεψης. Στα δεδομένα που έχουμε λάβει, έπειτα από έλεγχο που πραγματοποιήθηκε με την χρήση της γλώσσας προγραμματισμού Python, δεν εντοπίστηκαν ελλείπουσες τιμές, συνεπώς δεν χρειάζεται να παρέμβουμε για να τις διαχειριστούμε, συνεπώς προχωράμε σε ενέργειες που θα διευκολύνουν την 'δόμηση' του μοντέλου μας. Ένας περιορισμός που έχει ο αλγόριθμος XGBoost είναι ότι δεν μπορεί να διαχειριστεί κατηγορικές μεταβλητές (Jabeur, et al., 2023), για τον λόγο αυτό επιβάλλεται να μετατρέψουμε τις δύο κατηγορικές μεταβλητές που περιγράφουν το είδος της πλατφόρμας από την οποία έκαναν εγγραφή οι χρήστες (Android, iOS, Mobile, Desktop) και το προϊόν που επιλέγουν (Στοίχημα, Καζίνο) σε αριθμητικές. Αυτό πραγματοποιήθηκε με την διαδικασία One-Hot Encoding και την χρήση ψευδομεταβλητών (dummy variables).

Έχοντας ολοκληρώσει την προετοιμασία των δεδομένων, προχωράμε με τον διαχωρισμό των δεδομένων σε δύο μέρη. Το πρώτο μέρος αποτελείται από το 70% του συνόλου των δεδομένων, δηλαδή 141.866 παρατηρήσεις, και αποτελεί τα δεδομένα βάσει των οποίων θα εκπαιδευτεί το μοντέλο πρόβλεψης. Το δεύτερο μέρος αποτελείται από το υπόλοιπο 30% των δεδομένων, δηλαδή 60.800 παρατηρήσεις και αποτελεί τα δεδομένα στα οποία θα ελέγξουμε την απόδοση του μοντέλου πρόβλεψης και τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευσή του.

Στο πρώτο στάδιο δοκιμάζουμε την χρήση ενός εικονικού αλγόριθμου (Dummy Classifier) ώστε να καθορίσουμε την βασική απόδοση του μοντέλου. Στην συνέχεια κατασκευάζουμε το μοντέλο με την χρήση του αλγόριθμου XGBoost. Ένα πολύ σημαντικό στάδιο για την δημιουργία του μοντέλου αποτελεί η σωστή επιλογή του επιθυμητού συνδυασμού παραμέτρων ώστε να ενισχυθεί και να αυξηθεί η προβλεπτική ικανότητα του μοντέλο. Με την χρήση της μεθόδου Grid Search, δηλαδή τον συνδυασμό όλων των πιθανών παραμέτρων ώστε να βρεθεί ο καλύτερος συνδυασμός, καταλήξαμε στην χρήση των ιδανικών παραμέτρων του μοντέλου, μεταξύ διάφορων τιμών που δόθηκαν προς έλεγχο. Στο *Παράρτημα 3* παρουσιάζεται ο κώδικας βάσει του οποίου προέκυψαν οι ιδανικές τιμές στις υπερπαραμέτρους του μοντέλου. Στον *Πίνακα 5* παρουσιάζονται οι βέλτιστες τιμές που λάβαμε με την χρήση cross-validation.

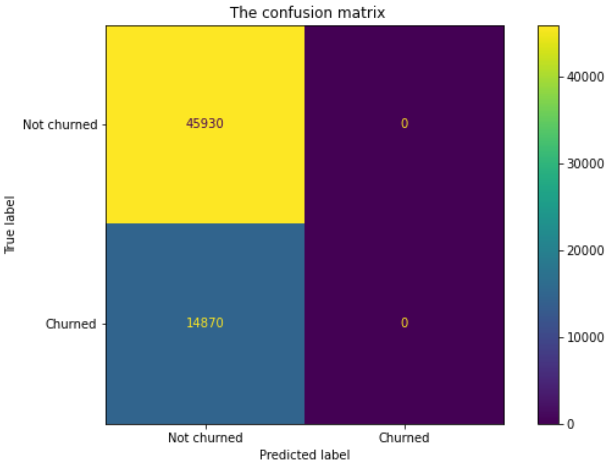
*Πίνακας 5 Βέλτιστες τιμές υπερπαραμέτρων*

| | Παράμετρος | Βέλτιστη Τιμή |
|---|---|---|
| 1 | n_estimators | 1000 |
| 2 | learning_rate | 0.5 |
| 3 | max_depth | 7 |
| 4 | min_child_weight | 3 |
| 5 | subsample | 1 |
| 6 | scale_pos_weight | 3 |
| 7 | colsample_bytree | 0.5 |

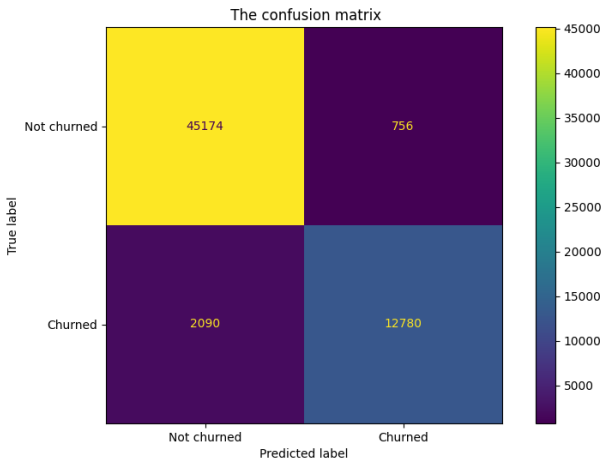
Τέλος, έχοντας καταλήξει στις ιδανικές τιμές που θα θέσουμε στις παραμέτρους του μοντέλου XGBoost, προχωράμε με την δημιουργία του. Ο κώδικας που παρουσιάζει αυτό το κομμάτι βρίσκεται στο *Παράρτημα 4*.

## 3.3 Αξιολόγηση μοντέλου πρόβλεψης

Στην Εικόνα 6 βλέπουμε πίνακα «σύγχυσης» (confusion matrix) του εικονικού αλγόριθμου, ενώ στην Εικόνα 7 βλέπουμε τον αντίστοιχο πίνακα με την χρήση του αλγόριθμου XGBoost και την χρήση των βέλτιστων παραμέτρων. Παρατηρούμε ότι βελτιώνεται σημαντικά η πρόβλεψη του μοντέλου καθώς η μετρική Accuracy στον εικονικό μοντέλο είναι 0.75, ενώ με την χρήση του αλγορίθμου XGBoost και τις βέλτιστες παραμέτρους η μετρική αυτή φτάνει στο 0.95.




*Εικόνα 8 Confusion Matrix of Dummy Classifier*




*Εικόνα 9 Confusion Matrix of XGBoost Classifier*

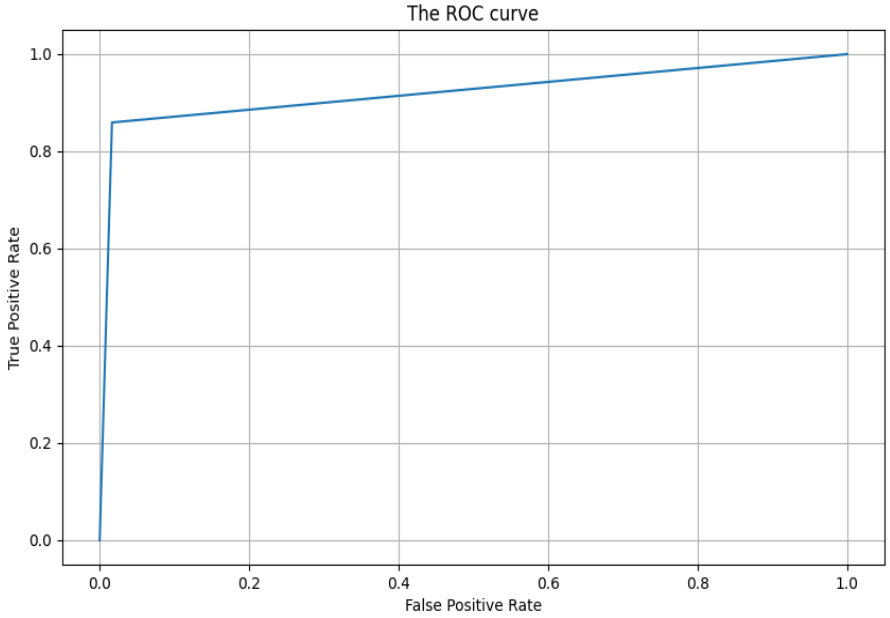
Το επόμενο στάδιο είναι η δημιουργία των καμπυλών ROC και AUC οι οποίες αποτελούν δείκτες απόδοσης στα προβλήματα κατηγοριοποίησης για διαφορετικές τιμές threshold. Έχοντας ελέγξει τις όλες τις τιμές από το 0 έως το 1 ανά 0.01, καταλήγουμε ότι η ιδανική τιμή threshold είναι 0.82, που στην τιμή αυτή έχουμε F1-score 0.90. Ο αντίστοιχος κώδικας βρίσκεται στο *Παράρτημα 5*. Η καμπύλη ROC (Εικόνα 10) παρουσιάζει την αντιστάθμιση μεταξύ του πραγματικού θετικού ρυθμού (True Positive Rate) και του ψευδώς θετικού ρυθμού (False Positive Rate) (Gattermann-Itschert & Thonemann, 2022), τα οποία ορίζονται ως:

$$True\ Positive\ Rate = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

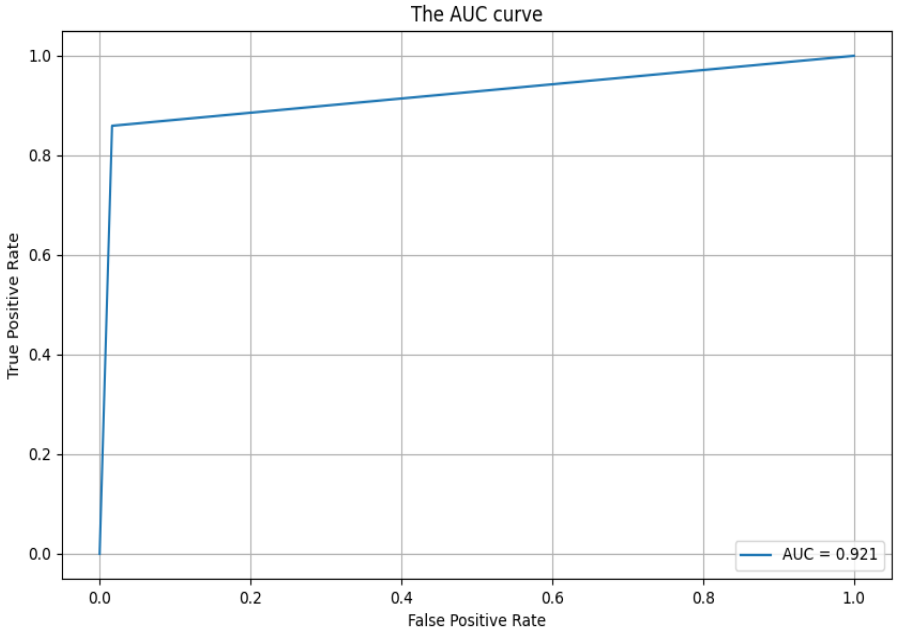
$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative} \quad (2)$$

Η καμπύλη AUC μετράει την περιοχή κάτω από την καμπύλη και επιτρέπει την ποσοτικοποίηση της απόδοσης του μοντέλου ως προς την σωστή πρόβλεψη της κάθε κλάσης (0,1) για την μεταβλητή στόχο στα δεδομένα που χρησιμοποιούμε στον έλεγχο του μοντέλου πρόβλεψης (test set). Μπορεί να πάρει τιμές από 0 έως 1 και όσο μεγαλύτερη τιμή έχει, τόσο καλύτερη πρόβλεψη έχει και στις 2 κλάσεις (0,1) (Lalwani,

et al., 2021). Το συγκεκριμένο μοντέλο έχει AUC=0.921, όπως φαίνεται στην Εικόνα 10.



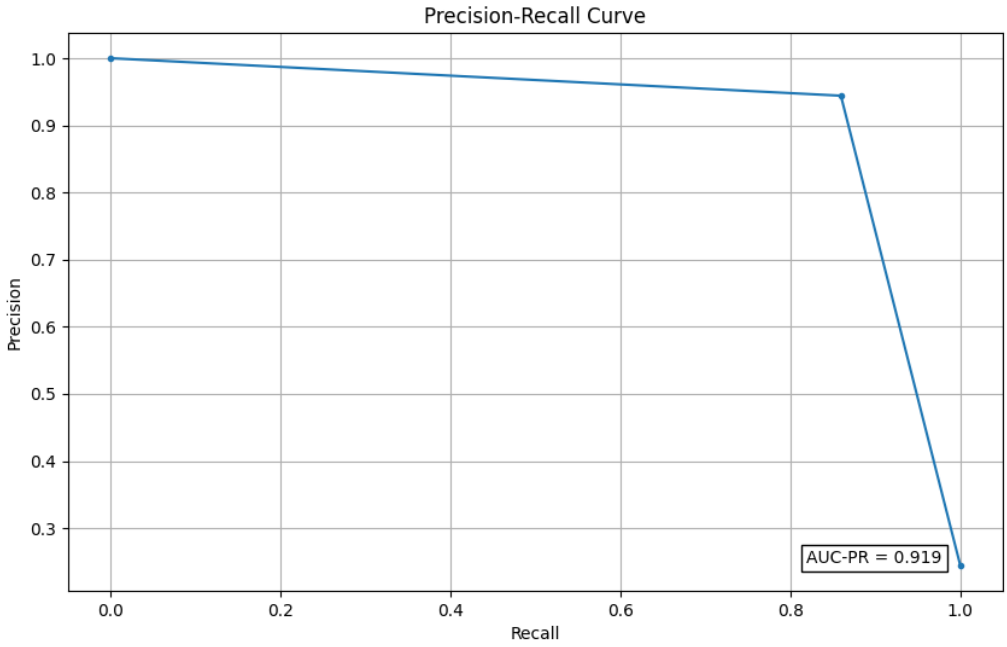

*Εικόνα 8 Καμπύλη ROC*




*Εικόνα 9 Καμπύλη AUC*

Μια μετρική ακόμα που λήφθηκε υπόψη για την αξιολόγηση της προβλεπτικότητας του μοντέλου είναι η καμπύλη Precision-Recall, η οποία δείχνει την αναλογία μεταξύ Precision (αριθμός αληθώς θετικών προβλέψεων προς το σύνολο των αληθών προβλέψεων) και Recall (αριθμός αληθώς θετικών προβλέψεων προς το σύνολο των

αληθώς θετικών προβλέψεων και των ψευδώς αρνητικών προβλέψεων) (Miao & Zhu, 2022). Το μέτρο για να αξιολογήσουμε την απόδοση είναι η καμπύλη Area under Precision Recall Curve (AUPRC) και φαίνεται στην Εικόνα 12, με AUC=0.919. Στον Πίνακα 6 συνοψίζονται όλες οι μετρικές του μοντέλου πρόβλεψης.



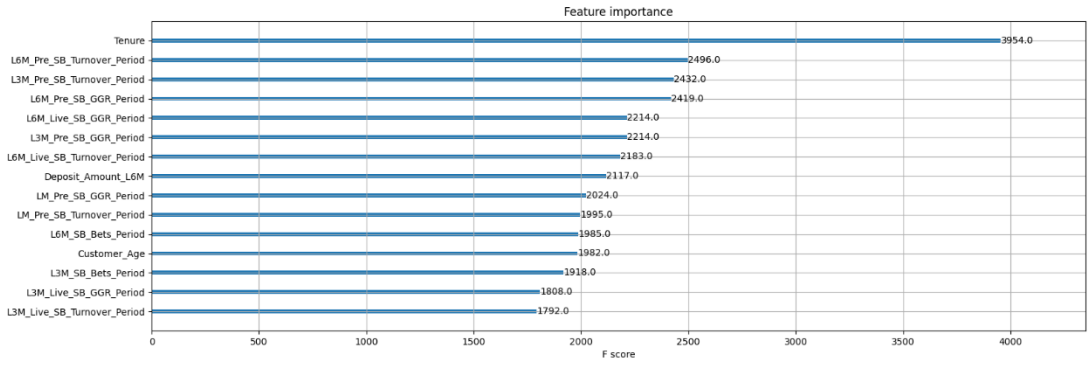

*Εικόνα 10 Καμπύλη Precision Recall*

*Πίνακας 6 Μετρικές μοντέλου πρόβλεψης*

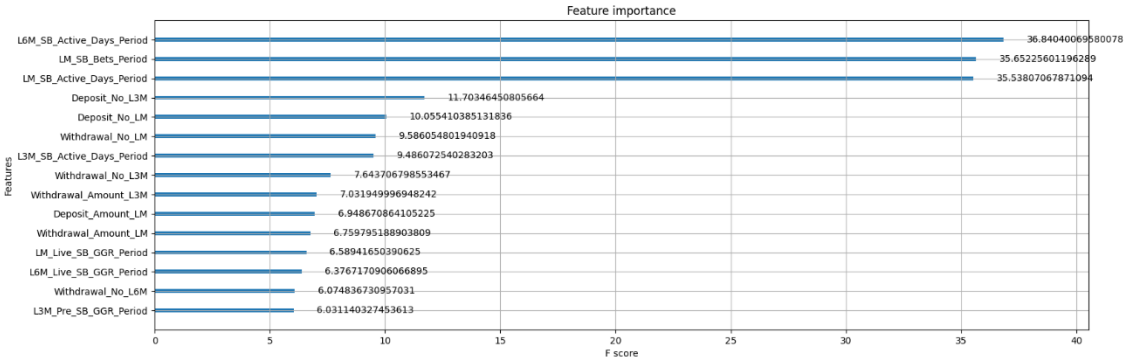
| | Accuracy | Precision | Recall | F1-Score | AUROC | AUPRC |
|---|---|---|---|---|---|---|
| XGBoost Classifier | 0.95 | 0.91 | 0.89 | 0.90 | 0.921 | 0.919 |

## 3.4 Ερμηνεία αποτελεσμάτων

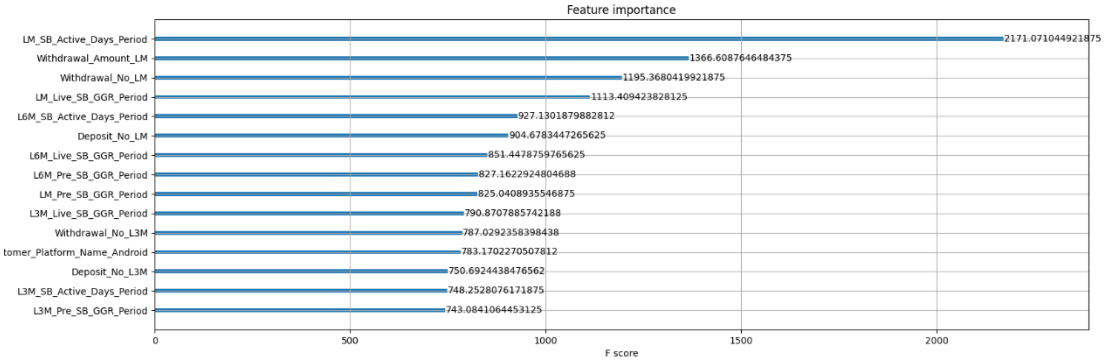
Για την ερμηνεία των αποτελεσμάτων των μοντέλου υπολογίσαμε την σημαντικότητα των μεταβλητών (feature importance) που χρησιμοποιήθηκαν για την πρόβλεψη του μοντέλου και το κατά πόσο συνέβαλαν σε αυτό. Η φύση του αλγορίθμου XGBoost έχει ως προκαθορισμένη μεταβλητή βάσει της οποίας υπολογίζεται η σημαντικότητα των μεταβλητών το "weight", το οποίο δείχνει τον αριθμό των φορών που μία μεταβλητή εμφανίζεται σε ένα δέντρο (tree). Εκτός όμως από αυτή την τιμή της παραμέτρου, μπορεί να υπολογιστεί και βάσει του "gain", το οποίο υπολογίζει την σημαντικότητα των μεταβλητών ανάλογα με το μέσο όφελος των διαχωρισμών που χρησιμοποιούν οι μεταβλητές, καθώς επίσης και με βάση το "cover", το οποίο υπολογίζει την σημαντικότητα των μεταβλητών ανάλογα με την μέση κάλυψη των διαχωρισμών που χρησιμοποιούν οι μεταβλητές, όπου η κάλυψη ορίζεται ως ο αριθμός των δειγμάτων που επηρεάζονται από τον διαχωρισμό (Anon., n.d.). Στις Εικόνες 13, 14 και 15 παρουσιάζεται η σημαντικότητα των πρώτων 15 μεταβλητών και για τις 3 τιμές του τύπου σημαντικότητας αντίστοιχα.




*Εικόνα 11 Σημαντικότητα μεταβλητών βάσει του Weight*




*Εικόνα 12 Σημαντικότητα μεταβλητών βάσει του Gain*



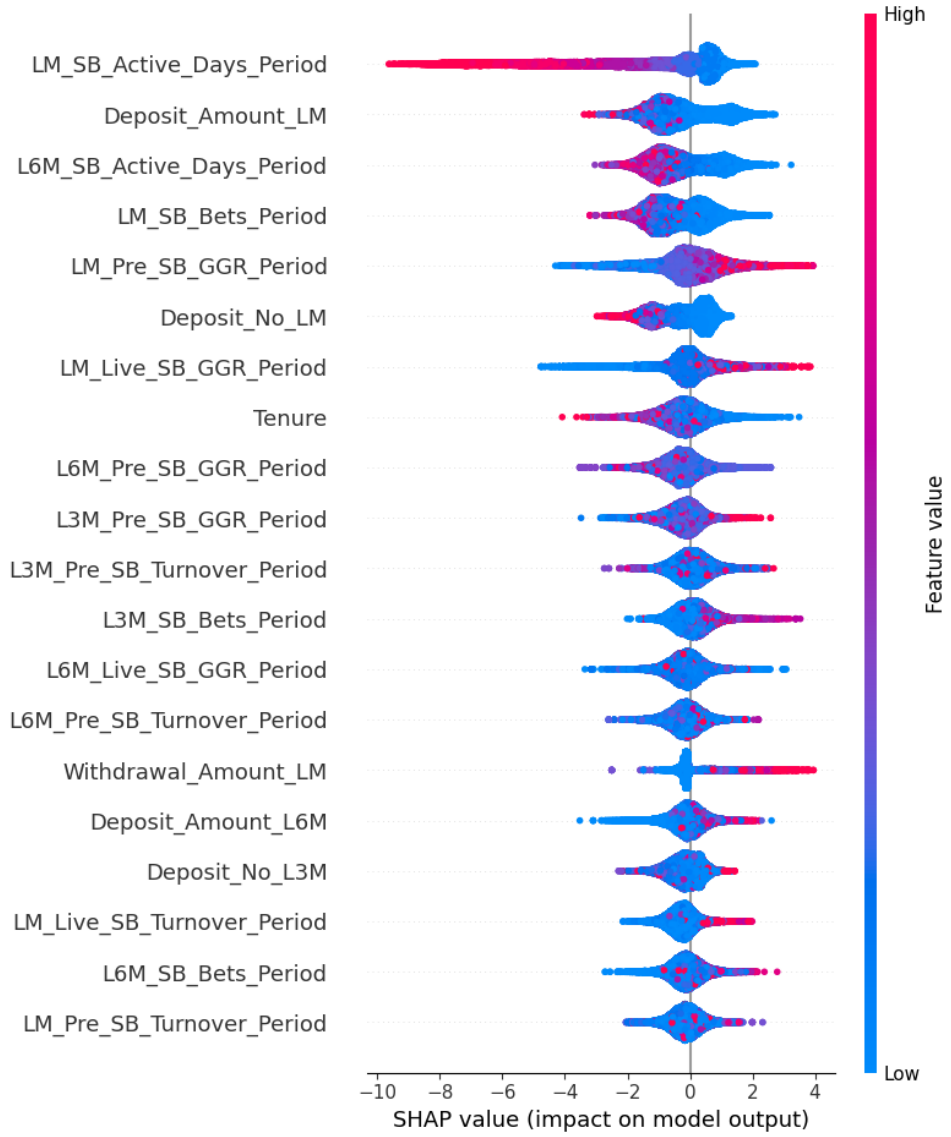

*Εικόνα 13 Σημαντικότητα μεταβλητών βάσει του Cover*

Παρατηρούμε πως υπάρχουν σημαντικές διαφοροποιήσεις, ανάλογα με το κριτήριο που υπολογίζεται η σημαντικότητα των μεταβλητών. Πιο συγκεκριμένα, στην ανάλυση που βασίζεται στο Weight, παρατηρούμε με σημαντική διαφορά να κυριαρχεί η μεταβλητή Tenure, ακολουθώντας οι μεταβλητές που δείχνουν το Turnover για το τελευταίο εξάμηνο και πολύ κοντά να βρίσκεται το Turnover του τελευταίου τριμήνου. Αντίστοιχα, μελετώντας την ανάλυση βάσει του Weight, φαίνεται να είναι τρεις οι μεταβλητές που παίζουν σημαντικό ρόλο στην πρόβλεψη του μοντέλου και αυτές είναι ο αριθμός των ημερών που είναι ενεργός ένας πελάτης το τελευταίο εξάμηνο, και ακολουθούν πολύ κοντά ο αριθμός των δελτίων που τοποθετήθηκαν τον τελευταίο μήνα και ο αριθμός των ημερών που ήταν ενεργό ένας πελάτης τον τελευταίο μήνα. Τέλος, η ανάλυση βάσει του Cover, φαίνεται να επηρεάζεται πολύ σημαντικά από τον αριθμό των ημερών που ήταν ενεργός ένας πελάτης τον τελευταίο μήνα, ακολουθώντας

έπειτα με μικρότερη επίδραση το ποσό των αναλήψεων που έκανε ένας πελάτης τον τελευταίο μήνα καθώς και το πλήθος αυτών.

Για την καλύτερη ερμηνεία των αποτελεσμάτων του μοντέλου πρόβλεψης αποφασίστηκε να χρησιμοποιηθεί η SHAP analysis. Η θεωρία στην οποία βασίζεται η SHAP analysis είναι βαθιά συνδεδεμένη με τη θεωρία των Shapley values από την θεωρία παιγνίων. Αναλυτικότερα, μέσω της SHAP analysis προσπαθούμε να ερμηνεύσουμε την πρόβλεψη που γίνεται από τα μοντέλα μηχανικής μάθησης για κάθε πρόβλεψη. Η μέθοδος Shapley values εκχωρεί αμοιβές σε παίκτες ανάλογα με τη συνεισφορά τους στη συνολική αμοιβή, όταν οι παίκτες συνεργάζονται σε συμμαχία (Shapley, 1953). Στην μηχανική μάθηση συγκεκριμένα, το «παίγνιο» είναι η πρόβλεψη για μία μόνο παρατήρηση του συνόλου των δεδομένων. Η «αμοιβή» είναι η πρόβλεψη μείον τη μέση πρόβλεψη όλων των περιπτώσεων και οι «παίκτες» είναι οι τιμές των μεταβλητών που συνεργάζονται για να λάβουν την αμοιβή. Η τιμή Shapley είναι η μέση οριακή συνεισφορά μίας τιμής μια μεταβλητής σε όλους τους πιθανούς συνδυασμούς (Kalampokis, et al., 2021).

Ο κώδικας βάσει του οποίου δημιουργήθηκε η SHAP analysis βρίσκεται στο *Παράρτημα 6*. Στην Εικόνα 16 παρουσιάζεται το διάγραμμα SHAP summary. Κάθε σημείο στο διάγραμμα αντιπροσωπεύει μία παρατήρηση, δηλαδή στην συγκεκριμένη μελέτη, έναν πελάτη. Ο άξονας Y παρουσιάζει τις μεταβλητές που χρησιμοποιήθηκαν για την δημιουργία του μοντέλου με σειρά σημαντικότητας, από την πιο σημαντική στην κορυφή του άξονα προς την λιγότερο σημαντική στη βάση του άξονα. Ο άξονας X παρουσιάζει την τιμή SHAP, η οποία δείχνει την επίδραση στην πρόβλεψη του μοντέλου. Το χρώμα με το οποίο απεικονίζεται κάθε παρατήρηση υποδηλώνει τον βαθμό που επηρεάζει την πρόβλεψη του μοντέλου, συγκεκριμένα, οι παρατηρήσεις που έχουν χρωματιστεί με κόκκινο υποδηλώνουν μεγάλη αξία στην πρόβλεψη, ενώ οι παρατηρήσεις που έχουν μπλε χρώμα υποδηλώνουν χαμηλότερη αξία στην πρόβλεψη. Από το διάγραμμα αυτό προκύπτει ότι η μεταβλητή με την μεγαλύτερη επίδραση στην πρόβλεψη του μοντέλου είναι ο αριθμός των ημερών που ήταν ενεργός ένας πελάτης τον τελευταίο μήνα. Ακολουθεί το ποσό των καταθέσεων που έχει πραγματοποιήσει ο πελάτης τον τελευταίο μήνα και στη συνέχεια ο αριθμός των ημερών που έμεινε ενεργός ένας πελάτης το τελευταίο εξάμηνο.




*Εικόνα 14 Ανάλυση SHAP*

Με την χρήση των SHAP dependence plot μπορούμε να δούμε πως αλλάζουν οι τιμές SHAP για τις διάφορες τιμές της κάθε μεταβλητής που εξετάζουμε. Πιο συγκεκριμένα, μπορούμε να διακρίνουμε τις τιμές που δείχνουν να έχουν μεγαλύτερη επίδραση στην πρόβλεψη του μοντέλου, όπως επίσης και ποιες παρατηρήσεις παρουσιάζουν μικρότερη επιρροή στην κατάταξη ενός πελάτη στην κατηγορία να διακόψει την δραστηριότητά του ή όχι. Στην Εικόνα 17 παρουσιάζονται τα SHAP dependence plots των έξι πρώτων μεταβλητών που εμφανίζονται στο SHAP summary. Ο κώδικας για τα SHAP dependence plot βρίσκονται στο *Παράρτημα 7-12*.

Στην Εικόνα 17α φαίνεται το dependence plot της μεταβλητής ‘Αριθμός ενεργών ημερών τον τελευταίο μήνα’. Ο άξονας Υ δείχνει τις τιμές Shapley και ο άξονας Χ δείχνει τις τιμές που παίρνει η μεταβλητή, οι οποίες είναι διακριτές και παίρνουν τιμές από 0 έως 9 και αυτό δικαιολογείται από το γεγονός ότι κυρίως τις ημέρες Σάββατο και Κυριακή υπάρχει μεγαλύτερη αθλητική δραστηριότητα οπότε και οι πελάτες εισέρχονται στην εφαρμογή ή την ιστοσελίδα. Από το διάγραμμα προκύπτει πως όσο αυξάνονται οι μέρες που ένας πελάτης είναι ενεργός, τόσο η τιμή μικραίνει σημαντικά παίρνοντας αρνητική τιμή κάτι που υποδηλώνει την μείωση της πιθανότητες ο πελάτης να φύγει από την εταιρία, και αντίστροφα, όσο λιγότερες μέρες ένας πελάτης μένει ενεργός τον τελευταίο μήνα τόσο αυξάνονται οι πιθανότητες να στραφεί στον ανταγωνισμό. Λόγου χάρη, ένας πελάτης που είναι ενεργός από μηδέν έως δύο ημέρες τον τελευταίο μήνα, αποτελεί ένδειξη πως μάλλον θα διακόψει την δραστηριότητά του στην εταιρία, αντίθετα οι πελάτες που είναι ενεργοί από έξι έως εννιά ημέρες έχουν ισχυρή ένδειξη πως θα συνεχίσουν την δραστηριότητά τους.

Επίσης, την Εικόνα 17β φαίνεται το dependence plot της μεταβλητής ‘Ποσό καταθέσεων τον τελευταίο μήνα’. Στον άξονα Υ βλέπουμε τις τιμές Shapley και στον άξονα Χ βλέπουμε τις τιμές που παίρνει η μεταβλητή του ποσού των καταθέσεων που έχει πραγματοποιήσει ένας πελάτης τελευταίο μήνα, που παίρνει τιμές από μηδέν έως εξήντα χιλιάδες. Παρατηρούμε πως για τις τιμές πολύ κοντά στο μηδέν είναι υπάρχει τόσο θετική όσο και αρνητική επίδραση στην πρόβλεψη του μοντέλου, ενώ όσο αυξάνονται τα ποσά των καταθέσεων τον τελευταίο μήνα, γίνεται πιο ξεκάθαρη επίδραση που υπάρχει στο μοντέλο και μάλιστα γίνεται σαφές πως επίδραση αυτή είναι αρνητική. Αυτό, εν προκειμένω, σημαίνει πως οι πελάτες που αποφασίζουν να καταθέσουν είναι δυσκολότερο να διακόψουν την δραστηριότητά τους, ενώ για τους πελάτες που δεν παρουσιάζουν σημαντικές καταθέσεις τον μήνα είναι δύσκολο να καταλήξουμε σε κάποιο συμπέρασμα, καθώς μπορεί είτε να σημαίνει πως έχουν διακόψει την δραστηριότητα τους, είτε πως εξακολουθούν να είναι ενεργοί χρησιμοποιώντας ίσως κεφάλαια που προέκυψαν νωρίτερα στον λογαριασμό τους, είτε κάποια bonus που μπορεί να έχουν λάβει.

Επιπλέον, στην Εικόνα 17γ φαίνεται το dependence plot της μεταβλητής ‘Αριθμός ενεργών ημερών τους τελευταίους 6 μήνες’. Ο άξονας Υ δείχνει τις τιμές Shapley και ο άξονας Χ δείχνει τις τιμές που παίρνει η μεταβλητή, οι οποίες είναι από μηδέν έως περίπου 55. Από το διάγραμμα αυτό προκύπτει ότι μεγαλύτερη θετική επίδραση έχουν

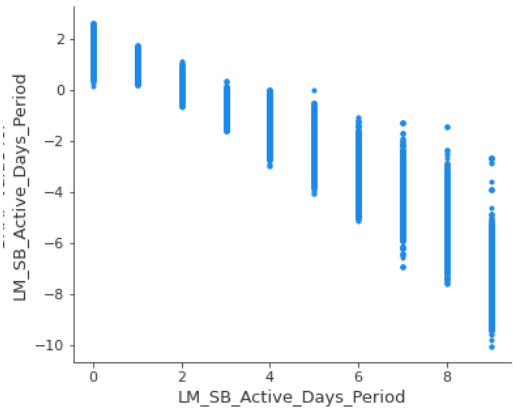
οι ημέρες μεταξύ μηδέν και πέντε, ενώ την μεγαλύτερη αρνητική επίδραση έχουν οι ημέρες μεταξύ δεκαοκτώ και είκοσι τρείς. Αναλυτικότερα, όπως και στην Εικόνα 17α παρατηρούμε πως όσο αυξάνονται οι μέρες που παραμένει ενεργός ένας πελάτης τόσο αυξάνονται οι πιθανότητες να παραμείνει ενεργός στην δραστηριότητά του στην εταιρία.

Στην Εικόνα 17δ φαίνεται το dependence plot της μεταβλητής 'Αριθμός δελτίων τον τελευταίο μήνα'. Ο άξονας Y δείχνει τις τιμές Shapley και ο άξονας X δείχνει τις τιμές που παίρνει η μεταβλητή, οι οποίες είναι από μηδέν έως χίλια εξακόσια. Στο διάγραμμα αυτό φαίνεται πως οι πελάτες που τοποθετούν πολύ μικρό αριθμό δελτίων τον τελευταίο μήνα έχουν μεγάλη θετική επίδραση στο μοντέλο, αντίθετα όσο αυξάνεται ο αριθμός των δελτίων που τοποθετούν οι πελάτες τον τελευταίο μήνα η επίδραση γίνεται αρνητική και σταδιακά μειώνεται ο βαθμός που επηρεάζει το μοντέλο. Πιο συγκεκριμένα, διαπιστώνουμε ότι ο μικρός αριθμός δελτίων τον τελευταίο μήνα μπορεί να αποτελέσει ένδειξη πως ένας πελάτης είναι πιθανό να διακόψει την δραστηριότητά του στην εταιρία και να αποφασίσει να συνεχίσει σε κάποια άλλη, απεναντίας, η αύξηση του αριθμού των δελτίων που τοποθετεί ένας πελάτης υποδεικνύει την πιθανότητα να διατηρήσει την δραστηριότητα του στην εταιρία.

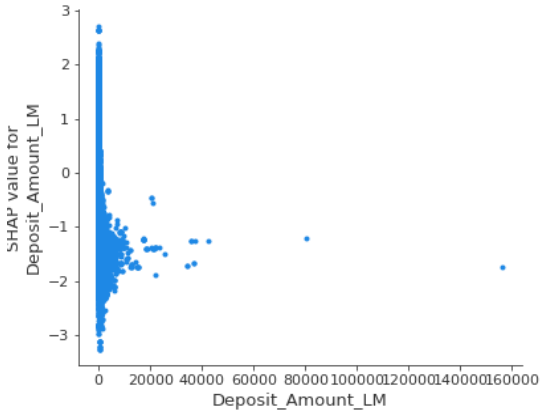
Στην Εικόνα 17ε φαίνεται το dependence plot της μεταβλητής 'Τα Gross Gaming Revenue-GGR τον τελευταίο μήνα'. Ο άξονας Y δείχνει τις τιμές Shapley και ο άξονας X δείχνει τις τιμές που παίρνει η μεταβλητή, οι οποίες είναι από -6000 έως περίπου 16000. Στο διάγραμμα αυτό φαίνεται πως οι πελάτες με αρνητικές τιμές GGR (δηλαδή οι πελάτες που είναι κερδισμένοι) τον τελευταίο μήνα έχουν αρνητική επίδραση στο μοντέλο, αντίθετα οι πελάτες με θετικές τιμές GGR (δηλαδή οι πελάτες που είναι χαμένοι) τον τελευταίο μήνα έχουν θετική επίδραση στο μοντέλο. Πιο συγκεκριμένα, διαπιστώνουμε ότι οι αρνητικές τιμές GGR τον τελευταίο μήνα μπορεί να αποτελέσει ένδειξη πως ένας πελάτης είναι πιθανό να συνεχίσει την δραστηριότητά του στην εταιρία ενώ, οι θετικές τιμές GGR εκφράζουν την πιθανότητα ο πελάτης να διακόψει τη δραστηριότητά του.

Τέλος, στην Εικόνα 17στ φαίνεται το dependence plot της μεταβλητής 'Αριθμός καταθέσεων τον τελευταίο μήνα'. Ο άξονας Y δείχνει τις τιμές Shapley και ο άξονας X δείχνει τις τιμές που παίρνει η μεταβλητή, οι οποίες είναι από μηδέν έως περίπου 260. Στο διάγραμμα αυτό δεν προκύπτει ξεκάθαρη σχέση για αριθμό καταθέσεων μικρότερο

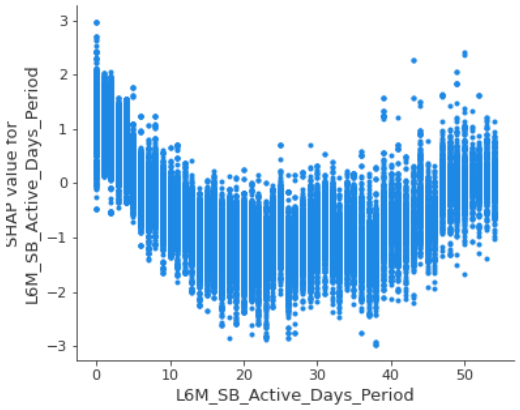
των 25, μπορούμε όμως να παρατηρήσουμε πως για πλήθος καταθέσεων άνω των 25 τον τελευταίο μήνα, η επίδραση στο μοντέλο είναι αρνητική. Αυτό σημαίνει πως ο μεγαλύτερος αριθμός καταθέσεων τον τελευταίο μήνα αποτελεί ένδειξη ότι ο πελάτης δεν θα στραφεί στον ανταγωνισμό και θα συνεχίσει την δραστηριότητά του στην εταιρία.



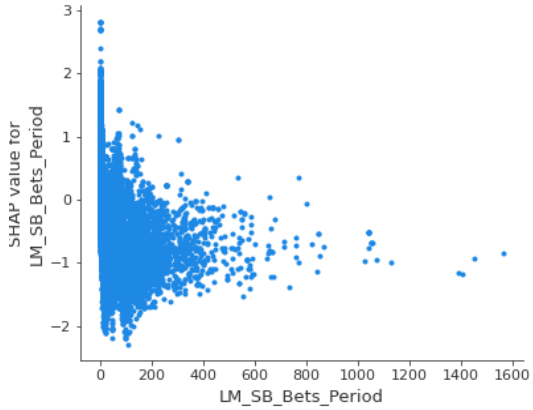

(α) Αριθμός ενεργών μερών τον τελευταίο μήνα



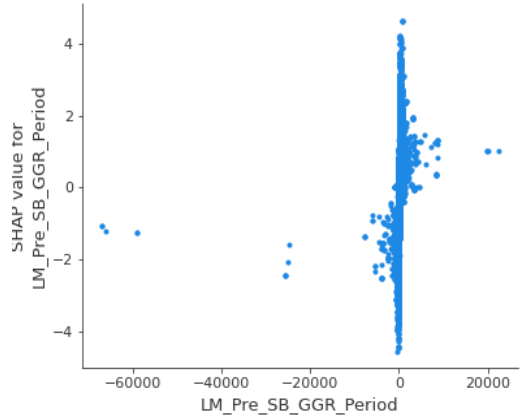

(β) Ποσό καταθέσεων τον τελευταίο μήνα



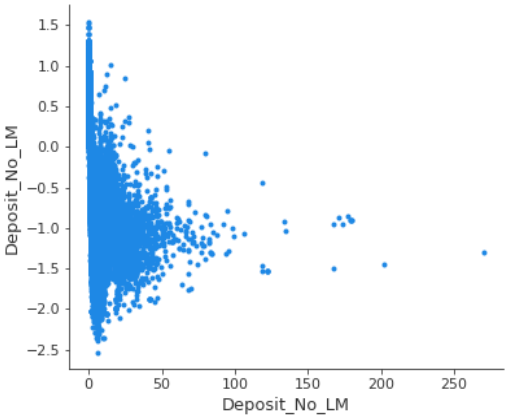

(γ) Αριθμός ενεργών ημερών τους τελευταίους 6 μήνες




(δ) Αριθμός δελτίων τον τελευταίο μήνα




(ε) Gross Gaming Revenue τον τελευταίο μήνα




(στ) Αριθμός καταθέσεων τον τελευταίο μήνα

*Εικόνα 15 SHAP dependence plot*

# 4 Συμπεράσματα

Η πρόβλεψη απώλειας πελάτη είναι ένα θέμα που έχει απασχολήσει ιδιαίτερα την βιβλιογραφία και αυτό γιατί μπορεί να έχει εφαρμογή σε πολλούς κλάδους όπως είναι οι τηλεπικοινωνίες, τα χρηματοπιστωτικά ιδρύματα και ο κλάδος των ασφαλίσεων. Επιπλέον, μπορεί να συμβάλει σημαντικά στην έγκαιρη αναγνώριση των πελατών με αυξημένη πιθανότητα να διακόψουν τις συναλλαγές τους σε την εταιρία, κάτι που αναπόφευκτα οδηγεί σε απώλεια κερδών. Αντίστοιχα, ο κλάδος του διαδικτυακού στοιχηματισμού που έχει αναπτυχθεί κυρίως τα τελευταία χρόνια με την ευρεία διάδοση του διαδικτύου και των έξυπνων κινητών, μπορεί να χρησιμοποιήσει την γνώση αυτή προς όφελος του, καθώς υπάρχει μεγάλος όγκος δεδομένων που προκύπτει από την δραστηριότητα των πελάτων τους που μπορεί να αξιοποιηθεί, αποφεύγοντας τον κίνδυνο μείωσης του πελατολογίου των εταιριών αυτών.

Σκοπός αυτής της εργασίας είναι με την χρήση της επιβλεπόμενης μηχανικής μάθησης (Supervised Machine Learning) να δημιουργηθεί ένα μοντέλο πρόβλεψης απώλειας πελάτη και με την χρήση εξηγήσιμης τεχνητής νοημοσύνης (Explainable Artificial Intelligence – XAI) να ερμηνεύσουμε ποιες με μεταβλητές επηρεάζουν και με ποιο τρόπο την πρόβλεψη του μοντέλου. Για τον λόγο αυτό παρουσιάστηκε μια μελέτη περίπτωσης, στην οποία χρησιμοποιήθηκαν δεδομένα από την εταιρία Kaizen Gaming ώστε να δημιουργήσουμε το μοντέλο πρόβλεψης και να εξηγήσουμε τις παραμέτρους που καθορίζουν την απώλεια πελατών στον κλάδο του διαδικτυακού στοιχηματισμού. Το μοντέλο δημιουργήθηκε με την χρήση 35 μεταβλητών εκ των οποίων η τελευταία είναι η μεταβλητή στόχος δηλαδή μια δυαδική μεταβλητή στην οποία δηλώνεται αν ο πελάτης έχει διακόψει την δραστηριότητα του στην εταιρία (τιμή 1) ή αν εξακολουθεί να είναι πελάτης (τιμή 0), ενώ οι υπόλοιπες μεταβλητές περιλαμβάνουν δημογραφικά δεδομένα και δεδομένα από την παικτική δραστηριότητα των πελατών και περιγράφονται όλες στο Παράρτημα 1. Τέλος, τα δεδομένα αποτελούνται από 202.666 παρατηρήσεις.

Κάποια συμπεράσματα που προκύπτουν από την αρχική ανάλυση δεδομένων είναι πως η απώλεια πελατών είναι ένα θέμα που αφορά κυρίως στους νεότερους πελάτες, δηλαδή σε ηλικίες μεταξύ είκοσι και σαράντα ετών, καθώς για μεγαλύτερες ηλικίες ο αριθμός των πελατών που αποφασίζουν να διακόψουν την δραστηριότητά τους μειώνεται

δραστικά. Επίσης, χαρακτηριστική είναι και η μεταβλητή που αφορά στις ημέρες που έχουν περάσει από την εγγραφή του πελάτη. Αυτό σημαίνει πως μεγαλύτερη πιθανότητα να αποχωρήσουν έχουν οι πελάτες που είναι για μικρό διάστημα χρήστες της εφαρμογής ή της ιστοσελίδας και πιο συγκεκριμένα οι πρώτες πεντακόσιες μέρες είναι ένα κρίσιμο διάστημα, καθώς φαίνεται πως περνώντας αυτό το όριο είναι πολύ πιο σπάνιο οι πελάτες να στραφούν στον ανταγωνισμό.

Ως προς το μοντέλο πρόβλεψης, η αξιολόγηση του έγινε με τις μετρικές Accuracy που έχει τιμή 0.95, σημαντικά υψηλότερη από την βασική απόδοση του εικονικού αλγορίθμου (Dummy Classifier) η οποία ήταν 0.75. Επίσης, η προβλεπτική ικανότητα του μοντέλου έχει Precision 0.91 και Recall 0.89. Τέλος, οι μετρικές AUROC και AUPRC δείχνουν επίσης την καλή απόδοση στην πρόβλεψη του μοντέλου με τιμές 0.921 και 0.919, αντίστοιχα.

Σημαντικότερα, η εργασία αυτή ασχολήθηκε με την επεξήγηση και την ερμηνεία των αποτελεσμάτων, ως προς το ποιες είναι οι μεταβλητές που επηρεάζουν το μοντέλο αλλά κυρίως, με ποιον τρόπο και σε τι βαθμό επιδρούν στην διαμόρφωσή του, κάτι που η υπάρχουσα για τον συγκεκριμένα κλάδο βιβλιογραφία υστερεί, παρόλο που μπορεί να παρέχει χρήσιμες πληροφορίες για τα στελέχη μιας εταιρίας που έχουν να διαχειριστούν αυτό το φαινόμενο. Πιο συγκεκριμένα, με την χρήση της ανάλυσης SHAP, οι τέσσερις πιο επιδραστικές μεταβλητές για την πρόβλεψη απώλειας πελατών στον κλάδο του διαδικτυακού στοιχηματισμού είναι ο αριθμός των ημερών που παραμένει ενεργός ένας πελάτης και με μεγαλύτερη σαφήνεια, αν ο πελάτης είναι ενεργός έως δύο ημέρες είναι πολύ πιθανό να διακόψει την δραστηριότητα του και να στραφεί στον ανταγωνισμό είτε αυτό ελέγχεται σε επίπεδο μήνα, είτε σε μεγαλύτερο διάστημα όπως είναι το εξάμηνο. Το δεύτερο επιβεβαιώνεται και από το γεγονός πως η αντίστοιχη μεταβλητή είναι μέσα στις τέσσερις πρώτες μεταβλητές που επηρεάζουν την απόφαση του μοντέλου στην κατηγοριοποίηση ενός πελάτη ως κάποιον που θα παραμείνει ή θα αποχωρήσει από την εταιρία. Αντίθετα ενεργός για περισσότερο από έξι ημέρες σε επίπεδο μήνα ή είκοσι ημέρες σε επίπεδο εξαμήνου, έχουμε ισχυρή ένδειξη πως θα παραμείνει πελάτης της εταιρίας. Η επόμενη μεταβλητή που επηρεάζει το μοντέλο είναι το ποσό των καταθέσεων που πραγματοποιεί ένας πελάτης τον τελευταίο μήνα. Ο τρόπος που αναγνωρίζεται να επηρεάζει αυτή η μεταβλητή την πρόβλεψη του μοντέλου είναι θετικός σε χαμηλά ποσά και σε ποσά κοντά στο μηδέν, που σημαίνει πως οι πελάτες με μικρές καταθέσεις είναι αυτοί που κυρίως τείνουν να αποχωρήσουν από την εταιρία,

ενώ για μεγαλύτερα ποσά καταθέσεων η επίδραση μικραίνει και για την ακρίβεια γίνεται αρνητική, δηλαδή οι πελάτες με μεγαλύτερες καταθέσεις ωθούνται να παραμείνουν ενεργοί στην δραστηριότητά τους στην εταιρία. Μία ακόμα μεταβλητή που επιδρά στην προβλεπτική ικανότητα του μοντέλου είναι ο αριθμός των στοιχηματικών δελτίων που τοποθετούν οι πελάτες τον τελευταίο μήνα. Οι πελάτες με μικρό αριθμό στοιχηματικών δελτίων τείνουν να έχουν μεγαλύτερη πιθανότητα να σταματήσουν την παικτική τους δραστηριότητα συγκριτικά με αυτούς που έχουν μεγαλύτερη δραστηριότητα, κάτι που μπορεί να δικαιολογηθεί από το γεγονός ότι πολλοί πελάτες μπορεί να δοκιμάζουν μια νέα πλατφόρμα ώστε να δουν αν το περιβάλλον και οι υπηρεσίες που παρέχονται ταιριάζουν στις δικές του επιλογές, και βάσει αυτού να κρίνουν αν είναι ικανοποιημένοι και θέλουν να συνεχίσουν ή αν στραφούν σε άλλους παρόχους που προσφέρουν αντίστοιχες υπηρεσίες. Στην συνέχεια, η επόμενη μεταβλητή που επιδρά στην ικανότητα πρόβλεψης του μοντέλου είναι τα Gross Gaming Revenue τον τελευταίο μήνα, μια μεταβλητή που όπως μπορεί εύκολα να γίνει αντιληπτό, όσο αυξάνονται τα μικτά κέρδη της εταιρίας που προκύπτουν από την δραστηριότητα των παικτών, δηλαδή οι πελάτες χάνουν, τόσο αυτή η μεταβλητή φαίνεται να δείχνει πως οι πελάτες διακόπτουν την δραστηριότητά τους, και αντίστροφα. Τέλος, μια ακόμη μεταβλητή που φαίνεται να επηρεάζει την απόφαση του μοντέλου είναι το πλήθος των καταθέσεων τον τελευταίο μήνα. Όσο αυξάνεται ο αριθμό των καταθέσεων που πραγματοποιεί ο πελάτης τον τελευταίο μήνα, τόσο φαίνεται να παραμένει ενεργός, αντίθετα για τον μικρό αριθμό καταθέσεων δεν είναι εύκολο να προκύψει ακριβής εικόνα για τον τρόπο που επηρεάζεται η προβλεπτική του ικανότητα.

Παρόλο που η πρόβλεψη απώλειας πελάτη δεν είναι κάτι νέο στην βιβλιογραφία, η εφαρμογή της στον κλάδο του διαδικτυακού στοιχηματισμού είναι κάτι που δεν έχει μελετηθεί σε μεγάλο βαθμό και οι μελέτες που υπάρχουν σταματούν στο σημείο που βρίσκουν ποιες είναι οι μεταβλητές που επηρεάζουν την απόφαση του μοντέλου, χωρίς να ερμηνεύουν τον τρόπο που αυτές επιδρούν στην τελική κατηγοριοποίηση των πελατών από το μοντέλο. Το κομμάτι αυτό είναι κάτι ιδιαίτερα βοηθητικό στα στελέχη που λαμβάνουν αποφάσεις για τον τρόπο που θα διαχειριστούν το πρόβλημα αυτό καθώς επίσης, μπορεί να αποτελέσει ένα αποτελεσματικό οδηγό ως προς τις κινήσεις που μπορούν να κάνουν ώστε να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα,

εστιάζοντας στους παράγοντες που θα επηρεάσουν σε μεγαλύτερο βαθμό την κρίση των πελατών για αυτή τους την απόφαση.

# Βιβλιογραφία


Andrade, E. B. & Iyer, G., 2009. Planned versus actual betting in sequential gambles. *Journal of Marketing Research,* 46(3), p. 372–383.

Anon., χ.χ. *XGBoost documantation.* [Ηλεκτρονικό]
Available at: https://xgboost.readthedocs.io/en/stable/python/python_api.html
[Πρόσβαση 7 6 2023].

Ascarza, E. και συν., 2018. In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions.. *Customer Needs and Solutions,* 5(2), p. 65–81.

Ascarza, E., Netzer, O. & Hardie, B. G. S., 2018. Some customers would rather leave without saying goodbye. *Marketing Science,* 37(1), pp. 54-77.

Ballings, M. & Van den Poel, D., 2022. Customer event history for churn prediction: How long is long enough?. *Expert Systems with Applications,* 39(18), p. 13517–13522.

Bleichrodt, H. & Schmidt, U., 2002. A context-dependent model of the gambling effect. *Management Science,* 48(6), pp. 802-812.

Brandusoiu, I., Toderean, G. & Ha, B., 2016. Methods for churn prediction in the prepaid mobile telecommunications industry. *International conference on communications,* pp. 97-100.

Brownlee, J., 2020. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python.* v1.2 επιμ. s.l.:Machine Learning Mastery.

Buckinx, W. & Van den Poel, D., 2005. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.. *European Journal of Operational Research,* 164(1), p. 252–268.

Chen, T. & Guestrin, C., 2016. *XGBoost: A Scalable Tree Boosting System.* San Francisco California USA, Association for Computing MachineryNew YorkNYUnited States, pp. 785-794.

Clemente-Císcar, M., San Matías, S. & Giner-Bosch, V., 2014. A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings.. *European Journal of Operational Research,* 239(1), pp. 276-285.

Clemente-Císcar, M., San Matías, S. & Giner-Bosch, V., 2014. A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings.. *European Journal of Operational Research,* 239(1), p. 276–285.

Colgate, M., Stewart, K. & Kinsella, R., 1996. Customer defection: A study of the student market in Ireland. *International Journal of Bank Marketing,* 14(3), p. 23–29.

Coussement, K. & De Bock, K. W., 2013. Customer churn prediction in the online gambling industry:The beneficial effect of ensemble learning. *Journal of Business Research,* 66(9), pp. 1629-1636.

Coussement, K., Lessman, S. & Verstraeten, G., 2017. A comparative analysis of data preparation algorithms for customer churnprediction: A case study in telecommunication. *Decision Support Systems,* Τόμος 95, pp. 27-36.

Dawes, J. & Swailes, S., 1999. Retention sans frontieres: Issues for financial service retailers. *International Journal of Bank Marketing,* 17(1), p. 36–43.

De Bock, K. & Van den Poel, D., 2012. Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Apllictaions,* Τόμος 39, pp. 6816-6826.

Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics,* 29(5), pp. 1189-1232.

Ganesh, J., Arnold, M. J. & Reynolds, K. E., 2000. Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing,* 64(3), pp. 65-87.

Gattermann-Itschert, T. & Thonemann, U. W., 2022. Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests. *Industrial Marketing Management,* Τόμος 107, pp. 134-147.

Grant, S. J. & Xie, Y., 2007. Hedging your bets and assessing the outcome. *Journal of Marketing Research,* 44(3), p. 516–524.

Gronros, L. & Janer, I., 2018. *Predicting Customer Churn in the iGmaing Industry using Supervised Machine Learning,* Stockholm, Sweden: Royal Institute of Technology.

Günther, C.-C.και συν., 2011. Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal,* pp. 58-71.

Hung, S.-Y., Yen, D. C. & Wang, H.-Y., 2006. Applying data mining to telecom churn management. *Expert Systems with Applications,* 31(3), pp. 515-524.

Jabeur, S. B., Stef', N. & Carmona, P., 2023. Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. *Computational Economics ,* Τόμος 61, p. 715–741.

Jadhav, R. J. & Pawar, U. T., 2011. Churn Prediction in Telecommunication. *International Journal of Advanced Computer Science and Applications,* Τόμος 2, pp. 17-19.

Kalampokis, E., Karamanou, A. & Tarabanis, K., 2021. *Applying explainable artificial intelligence techniques on linked open government data..* Granada, Spain, Springer.

Karnstedt, M. και συν., 2010. *Handbook of Social Network Technologies and Applications.* Boston: MA: Springer US.

Kaya, E. και συν., 2018. Behavioral attributes and financial churn prediction. *EPJ Data Science,* 7(1), pp. 1-18.

Kim, K., Jun, C.-H. & b, J. L., 2014. Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications,* 41(15), pp. 6575-6584.

Kirui, C., Hong, L., Cheruiyot, W. & Kirui, H., 2013. Predicting Customer Churn in Mobile Telephony Industry. *International Journal of Computer Science Issues,* 10(2), pp. 165-172.

Kraljević, G. & Gotovac, S., 2010. Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services. *Automatika ,* 51(3), pp. 275-283.

Lalwani, P., Mishra, M. K., Chadha, J. S. & Sethi, P., 2021. Customer churn prediction system: a machine learning approach. *Computing,* Τόμος 104, pp. 271-294.

Lam, D., 2006. The influence of religiosity on gambling participation. *Journal of Gambling Studies,* 22(3), p. 305–320.

Lam, D., 2007. An exploratory study of gambling motivations and their impact on the purchase frequencies of various gambling products. *Psychology and Marketing,* 24(9), p. 815–827.

McDaniel, S. R. & Zuckerman, M., 2003. The relationship of impulsive sensation seeking and gender to interest and participation in gambling activities. *Personality and Individual Differences,* 35(6), p. 1385–1400.

Merchie, F. & Ernst, D., 2022. *Churn Prediction in online gambling.* [Ηλεκτρονικό]
Available at:
https://www.researchgate.net/publication/357698886_Churn_prediction_in_online_gambling#fullTextFileContent
[Πρόσβαση November 2022].

Miao, J. & Zhu, W., 2022. Precision–recall curve (PRC) classification trees. *Evolutionary Intelligence,* Τόμος 15, p. 1545–1569.

Moeyersoms, J. & Martens, D., 2015. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems,* Τόμος 72, p. 72–81.

Mowen, J. C., Fang, X. & Scott, K., 2009. A hierarchical model approach for identifying the trait antecedents of general gambling propensity and of four gambling-related genres. *Journal of Business Research,* 62(12), p. 1262–1268.

Nitzan, I. & Libai, B., 2011. Social effects on customer retention. *Journal of Marketing,* Τόμος 75, pp. 24-38.

Prasad, U. D. & Madhavi, S., 2012. Prediction of churn behaviour of bank customers using data mining tools. *Business Intelligence Journal,* 5(1), pp. 25-30.

Qureshi, S. A. και συν., 2013. Telecommunication subscribers' churn prediction model using machine learning. *Eighth International Conference on Digital Information Management,* pp. 131-136.

Rahul, A., 2019 . *https://www.kdnuggets.com/.* [Ηλεκτρονικό]
Available at: https://www.kdnuggets.com/2019/10/5-classification-evaluation-metrics-every-data-scientist-must-know.html
[Πρόσβαση 8 2 2023].

Reinartz, W. J. & Kumar, V., 2003. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing,* 67(1).

Richter, Y., Yom-Tov, E. & Slonim, N., 2010. *Predicting customer churn in mobile networks through analysis of social groups.* Ohio, Society for Industrial and Applied Mathematics, pp. 732-741.

Seybert, N. & Bloomfield, R., 2009. Contagion of wishful thinking in markets. *Management Science,* 55(5), p. 738–751.

Shapley, L., 1953. A General Approach to Computation of Shapley Values. Στο: *Contributions to the Theory of Games* . s.l.:Princeton University Press, pp. 307-317.

Smith, G., Levere, M. & Kurtzman, R., 2009. Poker player behavior after big wins and big losses. *Management Science,* 55(9), p. 1547–1555.

Stekler, H. O., Sendor, D. & Verlander, R., 2010. Issues in sports forecasting. *International Journal of Forecasting,* 26(3), p. 606–621.

Suh, E. & Alhaery, M., 2016. Customer retention:Reducing Online Casino Player Churn Through the Application of Predictive Modeling. *Gaming Research & Review Journal,* 20(2), pp. 63-83.

Tang, L. και συν., 2014. Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operatonal Research,* Τόμος 296, p. 624–633.

Torkzadeh, G., Chang, J. C.-J. & Hansen, G. W., 2006. Identifying issues in customer relationship management at Merck-Medco. *Decision Support Systems,* 42(2).

Tsai, C. & Lu, Y., 2009. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications,* 36(10), pp. 12547-12553.

Verbeke, w. και συν., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research,* 218(1), pp. 211-229.

Xie, Y. & Li, X., 2008. *Churn prediction with Linear Discriminant Boosting algorithm.* Kunming, China, International Conference on Machine Learning and Cybernetics, IEEE.

# 5 Παράρτημα

*Παράρτημα 1 - Περιγραφή μεταβλητών*

| **Α/Α** | **Όνομα μεταβλητής** | **Περιγραφή μεταβλητής** | **Τύπος μεταβλητής** |
|---|---|---|---|
| **0** | Customer_Platform_Name | Πλατφόρμα εγγραφής | object |
| **1** | Customer_Product_Name | Προϊόν επιλογής | object |
| **2** | Customer_Age | Ηλικία πελάτη | int64 |
| **3** | Tenure | Ημέρες από την εγγραφή | int64 |
| **4** | LM_Pre_SB_GGR_Period | Συνολικό ποσό πονταρίσματος - κέρδη του παίκτη από pre-game παιχνίδια τον τελευταίο μήνα | float64 |
| **5** | LM_Live_SB_GGR_Period | Συνολικό ποσό πονταρίσματος - κερδη του παίκτη από live παιχνίδια στοιχηματισμού τον τελευταίο μήνα | float64 |
| **6** | LM_Pre_SB_Turnover_Period | Συνολικό ποσό πονταρίσματος από live παιχνίδια στοιχηματισμού τον τελευταίο μήνα | float64 |
| **7** | LM_Live_SB_Turnover_Period | Τζίρος από live παιχνίδια στοιχηματισμού τον τελευταίο μήνα | float64 |
| **8** | LM_SB_Active_Days_Period | Ημέρες δραστηριότητας στο στοίχημα τον τελευταίο μήνα | int64 |
| **9** | LM_SB_Bets_Period | Αριθμός στοιχηματικών τοποθετήσεων τον τελευταίο μήνα | float64 |
| **10** | L3M_Pre_SB_GGR_Period | Συνολικό ποσό πονταρισματος - κερδη του παίκτη από pre-game παιχνίδια τους τελευταίους τρεις μήνες | float64 |
| **11** | L3M_Live_SB_GGR_Period | Συνολικό ποσό πονταρισματος - κερδη του παίκτη από live παιχνίδια στοιχηματισμού τους τελευταίους τρεις μήνες | float64 |
| **12** | L3M_Pre_SB_Turnover_Period | Συνολικό ποσό πονταρισματος από pre-game παιχνίδια τους τελευταίους τρεις μήνες | float64 |
| **13** | L3M_Live_SB_Turnover_Period | Συνολικό ποσό πονταρισματος από live | float64 |

| | | | |
|---|---|---|---|
| | | παιχνίδια στοιχηματισμού τους τελευταίους τρεις μήνες | |
| **14** | L3M_SB_Active_Days_Period | Ημέρες δραστηριότητας στο στοίχημα τους τελευταίους τρεις μήνες | int64 |
| **15** | L3M_SB_Bets_Period | Αριθμός στοιχηματικών τοποθετήσεων τους τελευταίους τρεις μήνες | float64 |
| **16** | L6M_Pre_SB_GGR_Period | Συνολικό ποσό πονταρισματος - κερδη του παίκτη από pre-game παιχνίδια τους τελευταίους έξι μήνες | float64 |
| **17** | L6M_Live_SB_GGR_Period | Συνολικό ποσό πονταρισματος - κερδη του παίκτη από live παιχνίδια στοιχηματισμού τους τελευταίους έξι μήνες | float64 |
| **18** | L6M_Pre_SB_Turnover_Period | Συνολικό ποσό πονταρισματος από pre-game παιχνίδια τους τελευταίους έξι μήνες | float64 |
| **19** | L6M_Live_SB_Turnover_Period | Συνολικό ποσό πονταρισματος από live παιχνίδια στοιχηματισμού τους τελευταίους έξι μήνες | float64 |
| **20** | L6M_SB_Active_Days_Period | Ημέρες δραστηριότητας στο στοίχημα τους τελευταίους έξι μήνες | int64 |
| **21** | L6M_SB_Bets_Period | Αριθμός στοιχηματικών τοποθετήσεων τους τελευταίους έξι μήνες | float64 |
| **22** | Deposit_Amount_LM | Ποσό καταθέσεων τελευταίου μήνα | float64 |
| **23** | Deposit_No_LM | Αριθμός καταθέσεων τελευταίου μήνα | float64 |
| **24** | Deposit_Amount_L3M | Ποσό καταθέσεων τελευταίων τριών μηνών | float64 |
| **25** | Deposit_No_L3M | Αριθμός καταθέσεων τελευταίων τριών μηνών | float64 |
| **26** | Deposit_Amount_L6M | Ποσό καταθέσεων τελευταίων έξι μηνών | float64 |
| **27** | Deposit_No_L6M | Αριθμός καταθέσεων τελευταίων έξι μηνών | float64 |
| **28** | Withdrawal_Amount_LM | Ποσό καταθέσεων τελευταίου μήνα | float64 |
| **29** | Withdrawal_No_LM | Αριθμός αναλήψεων τελευταίου μήνα | float64 |
| **30** | Withdrawal_Amount_L3M | Ποσό αναλήψεων τελευταίων τριών μηνών | float64 |

| | | | |
|---|---|---|---|
| **31** | Withdrawal_No_L3M | Αριθμός αναλήψεων τελευταίων τριών μηνών | float64 |
| **32** | Withdrawal_Amount_L6M | Ποσό αναλήψεων τελευταίων έξι μηνών | float64 |
| **33** | Withdrawal_No_L6M | Αριθμός αναλήψεων τελευταίων έξι μηνών | float64 |
| **34** | Churned | Απώλεια πελάτη | int64 |

*Παράρτημα 2 – Εισαγωγή βιβλιοθηκών*

```
# Import libraries
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import plotly
import plotly.express as px
! pip install xgboost
import xgboost as xgb
from xgboost import XGBClassifier
from sklearn.model_selection import StratifiedKFold, GridSearchCV
from sklearn import metrics
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report, roc_curve, auc, precision_recall_curve, plot_precision_recall_curve
from sklearn.model_selection import train_test_split
! pip install shap
import shap
```

*Παράρτημα 3 – Grid Search για εύρεση βέλτιστων υπερπαραμέτρων για τη δημιουργία του μοντέλου*

```
# Create a XGBoost model
model = xgb.XGBClassifier()

# Create a grid search object
grid_search = GridSearchCV(estimator=model, param_grid=params, cv=5, verbose=1, n_jobs=-1)
fit = grid_search.fit(X_train, y_train)

# Print the best hyperparameters and score
print("Best hyperparameters:", fit.best_params_)
print("Best score:", fit.best_score_)
```

Παράρτημα 4 – Δημιουργία μοντέλου πρόβλεψης με τον αλγόριθμο XGBoost και την χρήση βέλτιστων υπερπαραμέτρων

```
#Defining our model

xgb_param_grid={"n_estimators":[1000],"max_depth":[7],"learning_rate":[0.50],"min_child_weight":[3],
                'alpha':[0.01],'scale_pos_weight':[3],'colsample_bytree':[0.50],'subsample': [1.0]}   #sale_pos_weight=153144/49522=3.09
xgb = XGBClassifier(objective="binary:logistic", eval_metric="auc", random_state=123)
cv_f=StratifiedKFold(n_splits=3,shuffle=True)
gridsearch = GridSearchCV(estimator=xgb, param_grid=xgb_param_grid, n_jobs=2, cv=cv_f, verbose=1, scoring='f1')
model = gridsearch.fit(X_train,y_train)
preds= model.predict(X_test)
```

*Παράρτημα 5 – Εύρεση ιδανικής τιμής threshold*

```
#Obtain predicted probabilities
preds2=(model.predict_proba(X_test)[:,1])

#Find the best threshold
best_threshold = None
best_f1_score = 0.0

thresholds = np.arange(0.0, 1.01, 0.01)
for threshold in thresholds:
    predicted_labels = (preds2 > threshold).astype(int)
    f1 = f1_score(y_test, predicted_labels)
    if f1 > best_f1_score:
        best_f1_score = f1
        best_threshold = threshold

print("Best Threshold:", best_threshold)
print("Best F1 Score:", best_f1_score)
```

*Παράρτημα 6 – SHAP analysis*

```
#SHAP analysis
explainer = shap.Explainer(xgb_cl)

shap_values = explainer.shap_values(X_test)
fig=plt.figure(figsize=(25,10))
shap.summary_plot(shap_values, X_test, show=False)
```

*Παράρτημα 7 – Dependence plot 1*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("LM_Pre_SB_GGR_Period", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_LM_Pre_SB_GGR.png')
```

*Παράρτημα 8 – Dependence plot 2*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("Deposit_Amount_LM", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_Deposit_Amount_LM.png')
```

*Παράρτημα 9 – Dependence plot 3*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("L6M_SB_Active_Days_Period", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_L6M_Active_Days.png')
```

*Παράρτημα 10 – Dependence plot 4*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("LM_SB_Bets_Period", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_LM_Bets.png')
```

*Παράρτημα 11 – Dependence plot 5*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("LM_Pre_SB_GGR_Period", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_LM_Pre_SB_GGR.png')
```

*Παράρτημα 12 – Dependence plot 6*

```
fig=plt.figure(figsize=(15,5))
shap.dependence_plot("Deposit_No_LM", shap_values, X_test, interaction_index=None, show=False)
plt.savefig('shap_Deposit_No_LM.png')
```