ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Αλγόριθμοι μηχανικής μάθησης για την προσέλκυση πελατών: Μια συγκριτική αξιολόγηση στο χώρο των τραπεζικών υπηρεσιών

Υπό
ΤΑΒΕΡΝΑΡΑΚΗ ΜΑΡΙΑ ΖΟΥΖΑΝΝΑ

Χανιά, 2020

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στη Σχολή Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης. Με αφορμή την εργασία αυτή θα ήθελα να ευχαριστήσω θερμά όσους συνέβαλαν στην ολοκλήρωση της αλλά και όσους ήταν δίπλα μου σε όλη τη πορεία φοίτησής μου.

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω τον υπεύθυνο και επιβλέποντα καθηγητή της εργασίας μου κ. Μιχάλη Δούμπο για την πολύτιμη βοήθεια και την καθοδήγησή του κατά την ευκαιρία να εργαστώ πάνω στο θέμα. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου που με στήριξαν με κάθε τρόπο τα χρόνια των σπουδών μου.

Χανιά, 2020

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Σε μια εποχή όπου η τεχνολογία εξελίσσεται ραγδαία, οι ανάγκες των επιχειρήσεων ολοένα και αυξάνονται. Προκειμένου μια επιχείρηση να επιβιώσει θα πρέπει να είναι σε θέση να αντιμετωπίζει την αβεβαιότητα και τη πολυπλοκότητα των διεργασιών της.

Παράλληλα με τις μεταβολές αυτές οι επιχειρήσεις οφείλουν να συμμορφώνονται όχι μόνο με τις τάσεις της αγοράς αλλά και με τις ανάγκες των καταναλωτών. Για το λόγο αυτό χρησιμοποιούν διάφορες μεθόδους και τεχνικές έτσι ώστε να προσελκύσουν πελάτες, να κατανοήσουν τα χαρακτηριστικά τους και ανάλογα να προωθήσουν το προϊόν ή υπηρεσία όπου είναι πιο πιθανόν να ενδιαφέρει τους πελάτες.

Η γνώση των πελατών που διαθέτει μια τράπεζα και κατ' επέκταση κάθε επιχείρηση, είναι σημαντικό βήμα προκειμένου να είναι σε θέση να προβλέψει τη καταναλωτική συμπεριφορά των πελατών που θέλει να προσελκύσει. Δηλαδή, γνωρίζοντας το τρόπο επιλογής προϊόντων και υπηρεσιών που κατέχουν οι πελάτες της είναι δυνατόν να γίνει πρόβλεψη και άλλων με παρόμοια συμπεριφορά.

Από τις πλέον διαδεδομένες μεθόδους που χρησιμοποιούν οι επιχειρήσεις για το σκοπό αυτό είναι η χρήση αλγορίθμων μηχανικής μάθησης. Οι αλγόριθμοι αυτοί ποικίλουν και μπορούν να προσαρμοστούν ανάλογα τα δεδομένα όπου καλούνται να διαχειριστούν. Προαπαιτούμενο για την εφαρμογή των αλγορίθμων αυτών είναι η ύπαρξη βάσεως δεδομένων των πελατών.

Σκοπός της παρούσας διπλωματικής εργασίας είναι, μέσω διερεύνησης της καταναλωτικής συμπεριφοράς ενός δείγματος πελατών μιας τράπεζας, να εξαχθούν συμπεράσματα για την καταναλωτική τους συμπεριφορά το επόμενο διάστημα.

Κατά τα πρώτα κεφάλαια γίνεται μια ανάλυση του θεωρητικού υπόβαθρου του τομέα του CRM, της εξόρυξης δεδομένων και τέλος της μηχανικής μάθησης. Πρόκειται για μια αλυσίδα δραστηριοτήτων που πρέπει να εκτελέσει μια επιχείρηση προκειμένου να εξάγει τα επιθυμητά αποτελέσματα. Έτσι, η ύπαρξη ενός κατάλληλου συστήματος διαχείρισης πελατειακών σχέσεων οδηγεί στην ανάπτυξη καλών σχέσεων με τους

πελάτες της αλλά και τη δυνατότητα ύπαρξης αρχείου καταγραφών των συναλλαγών με την επιχείρηση.

Όπως αναλύεται και στο κεφάλαιο 3, η ύπαρξη των κατάλληλων δεδομένων είναι πολύ σημαντικό βήμα για την εξαγωγή χρήσιμων συμπερασμάτων. Παρόλα αυτά σημαντικό βήμα της διαδικασίας αποτελεί η προ επεξεργασία των δεδομένων έτσι ώστε να γίνει επιλογή μόνο των δεδομένων που είναι απαραίτητων για την ανάλυση.

Έπειτα, στο κεφάλαιο 4, γίνεται η επιλογή των κατάλληλων αλγορίθμων που θα χρησιμοποιηθούν για την ανάλυση. Αναπόσπαστο κομμάτι της εκτέλεσης των αλγορίθμων είναι η αξιολόγησή τους έτσι ώστε να αποσαφηνιστεί η ποιότητα των μοντέλων και εφόσον κρίνεται σκόπιμο να διορθωθούν τα μοντέλα και να επαναληφθεί η διαδικασία.

ΚΕΦΑΛΑΙΟ 2 .ΔΙΑΧΕΙΡΙΣΗ ΠΕΛΑΤΕΙΑΚΩΝ ΣΧΕΣΕΩΝ

Στο παρόν κεφάλαιο αναλύεται η έννοια της διαχείρισης πελατειακών σχέσεων (customer relationship management, CRM) στο σύγχρονο επιχειρησιακό περιβάλλον, ο ρόλος του ανθρώπινου παράγοντα στο σύστημα αυτό καθώς και δείκτες αξιολόγησης του.

2.1 ΟΡΙΣΜΟΣ CRM

Θεμελιώδη μέλημα των επιχειρήσεων έχει καταστεί η εξυπηρέτηση των πελατών τους. Η εξυπηρέτηση αφορά τόσο την παροχή κατάλληλων προϊόντων και υπηρεσιών αλλά και την ύπαρξη ενός περιβάλλοντος κατανόησης των αναγκών των πελατών τους . Για το λόγο αυτό έχουν αναπτυχθεί τα συστήματα CRM, τα οποία επικεντρώνονται στην υποστήριξη των πωλήσεων και στην εξυπηρέτηση των πελατών. Παρακάτω θα δοθούν κάποιοι ορισμοί για την κατανόηση της έννοιας του CRM:

- Σύμφωνα με τους Kumar και Petersen (2012) το CRM ορίζεται ως μια πρακτική συλλογής, αποθήκευσης, ανάλυσης πληροφοριών των πελατών και ενσωμάτωση των αποτελεσμάτων σε μια διαδικασία λήψης αποφάσεων. Το σύστημα αυτό περιλαμβάνει μεταξύ άλλων την ενίσχυση και άλλων επιχειρησιακών διαδικασιών όπως την παραγωγή, τις διεργασίες, τις πωλήσεις, το marketing καθώς και τα οικονομικά.
- CRM (Payne, 2005) είναι μια στρατηγική προσέγγιση που σχετίζεται με την αύξηση της αξίας των μετόχων μέσω της ανάπτυξης κατάλληλων σχέσεων με τους πελάτες-κλειδί και τις συστάδες πελατών. Το CRM ενώνει τις δυνατότητες της τεχνολογίας πληροφοριών(Information Technology, IT) με στρατηγικές σχεσιακού μάρκετινγκ με σκοπό τη δημιουργία επικερδών και μακροχρόνιων σχέσεων με τους πελάτες.
- Το CRM (Tiwana, 2001) είναι ο συνδυασμός επιχειρησιακών διαδικασιών και τεχνολογίας που υιοθετεί μια επιχείρηση επιδιώκοντας, μέσω της

αποκωδικοποίησης της συμπεριφοράς των πελατών της, να διαφοροποιήσει τα προϊόντα και τις υπηρεσίες της για κάθε πελάτη προσβλέποντας στην απόκτηση ανταγωνιστικού πλεονεκτήματος.

Τα δομικά στοιχεία ενός συστήματος CRM (Nelson, 2003) είναι το όραμα, η στρατηγική, η πρόταση αξίας[1], η οργανωτική συνεργασία, οι διεργασίες, οι πληροφορίες, η τεχνολογία καθώς και οι δείκτες επίδοσης του CRM.

Ανάλογα με το στόχο της εκάστοτε επιχείρησης το πρόγραμμα CRM μπορεί να διαφέρει . Γενικά, τα συστήματα CRM κατατάσσονται στις εξής κατηγορίες :

- Στρατηγικό CRM (Strategic CRM)
- Επιχειρησιακό CRM (Operational CRM)
- Αναλυτικό CRM (Analytical CRM)
- Συνεργατικό CRM (Collaborative CRM)

Οι επιχειρήσεις όπου υιοθετούν ένα στρατηγικό CRM έχουν ως προτεραιότητα τις συνεχώς μεταβαλλόμενες ανάγκες των πελατών τους στοχεύοντας κατ' αυτόν τον τρόπο στην προσέλκυση νέων αλλά και στη διατήρηση των ήδη υπαρχόντων. Στο επιχειρησιακό CRM περιλαμβάνονται δραστηριότητες διαχείρισης μιας πώλησης, μεταξύ άλλων και διαχείριση προσφορών, επίλυση προβλημάτων χρηστών και παρακολούθηση συμφωνιών επιπέδου εξυπηρέτησης πελατών. Όσον αφορά το αναλυτικό CRM, η επιχείρηση στοχεύει μέσω της εξόρυξης δεδομένων, να κατανοήσει τις ανάγκες των καταναλωτών και τις τάσεις της αγοράς. Στο συνεργατικό CRM η επιχείρηση προσπαθεί να αυξήσει την αξία των πελατών της. Σύμφωνα με τον Porter (2008) η μεγιστοποίηση της αξίας επιτυγχάνεται μέσω της συνεργασίας των επιχειρήσεων που λαμβάνουν μέρος στην αλυσίδα αξίας. Παρ' όλα αυτά, απαιτείται ύπαρξη διαφορετικής στρατηγικής για την διαχείριση της σχέσης με κάθε μεμονωμένο πελάτη. Στόχος είναι η δημιουργία σωστών σχέσεων με τους σωστούς πελάτες (Kotler et al., 2010).

Αρχική μορφή του επιχειρησιακού συστήματος CRM ήταν η αυτοματοποίηση των πωλήσεων (sales-force automation, SFA). Ο όρος SFA αναφέρεται σε οποιαδήποτε τεχνολογία χρησιμοποιείται για την υποστήριξη της διαδικασίας πωλήσεων. Το

[1] Πρόταση αξίας (value proposition) είναι η σαφής ή έμμεση υπόσχεση που δίνει μια επιχείρηση στους πελάτες της ότι τα προϊόντα ή υπηρεσίες που προσφέρει έχουν περισσότερη αξία από ότι των ανταγωνιστών. Δίνουν δηλαδή στον πελάτη τον λόγο που πρέπει να προτιμήσει το συγκεκριμένο προϊόν ή υπηρεσία από την συγκεκριμένη επιχείρηση.

λογισμικό SFA παρέχει εφαρμογές μορφοποίησης εξατομικευμένων προϊόντων και υπηρεσιών καθώς και την αυτοματοποιημένη κοστολόγηση τους ανάλογα με τα επιλεγμένα χαρακτηριστικά. Η παραπάνω εφαρμογή μειώνει τον κύκλο ζωής της πώλησης με αποτέλεσμα την αύξηση του παραγωγικού χρόνου των πωλήσεων και την ικανοποίηση των πελατών. Μεταξύ άλλων το σύστημα SFA περιλαμβάνει εφαρμογές διαχείρισης λογαριασμών, δραστηριοτήτων, επαφών, παραγγελιών και ευκαιριών. Όσον αφορά τα προϊόντα/υπηρεσίες περιλαμβάνει εφαρμογές διαμόρφωσης τους, αναλυτικές περιγραφές και την απεικόνιση εφόσον είναι εφικτό.

Μια ακόμη εφαρμογή των συστημάτων CRM είναι η διαχείριση σχέσεων εργαζομένων (employee relationship management, ERM). Το ERM (Strohmeier, 2013) αποτελεί μια αναπτυσσόμενη τάση για τη διαχείριση των ανθρωπίνων πόρων μέσω της δημιουργίας και της διατήρησης μοναδικών και αμοιβαίων σχέσεων με τους πελάτες εκμεταλλευόμενη την τεχνολογία πληροφοριών. Η ύπαρξη καλής σχέσης εργαζομένου-επιχείρησης παρουσιάζει επωφελή αποτελέσματα για την επιχείρηση μιας και οι εργαζόμενοι είναι εκείνοι που έρχονται σε άμεση επαφή με τους πελάτες. Κύριος στόχος του ERM είναι η ανάπτυξη μακροχρόνιων σχέσεων με τους εργαζομένους και τους μάνατζερ. Όσον αφορά τους μάνατζερ το ERM παρέχει εργαλεία διαχείρισης προσλήψεων, εκπαίδευσης, επιδόσεων εργαζομένων και αποδοχών. Επιπλέον διευκολύνει την επικοινωνία των μάνατζερ με τις ομάδες τους, παρέχει στους εργαζομένους πληροφορίες σχετικές με τα καθήκοντα τους και ενισχύει την επικοινωνία μεταξύ των μελών της επιχείρησης. Η σύνδεση επιχειρηματικής απόδοσης με ικανοποίησης πελατών και ικανοποίησης εργαζομένων λέγεται αλυσίδα εξυπηρέτησης-κέρδους. Η στρατηγική αυτή στηρίζεται στην πεποίθηση ότι οι εργαζόμενοι που είναι ικανοποιημένοι όσον αφορά την εργασία τους ανταποκρίνονται αποτελεσματικά στα καθήκοντά τους με συνέπεια τη μεγιστοποίηση της εμπειρίας του πελάτη, είτε είναι εσωτερικός, είτε εξωτερικός.

Ωστόσο, πολλές φορές τα σχέδια εφαρμογής συστημάτων CRM αποτυγχάνουν για ποικίλους λόγους. Συχνή είναι η εφαρμογή ενός συστήματος CRM χωρίς να έχει προηγηθεί ορισμός της κατάλληλης στρατηγικής όπως επίσης και η χρήση από εργαζομένους που δεν έχουν εκπαιδευτεί. Σημαντικό ρόλο διαδραματίζουν και τα δεδομένα που θα χρησιμοποιηθούν. Κακής ποιότητας δεδομένα μπορεί να περιλαμβάνουν ελλιπή στοιχεία πελατών, τυπογραφικά λάθη καθώς και πολλαπλές βάσεις δεδομένων διασκορπισμένες σε όλη την επιχείρηση. Επιπλέον, πολλές

επιχειρήσεις αδυνατούν να αξιοποιήσουν τα δεδομένα που κατέχουν με αποτέλεσμα την αδυναμία βελτίωσης των επιχειρησιακών διεργασιών και τη λήψη αποφάσεων. Πιθανόν υπάρχουν περίσσια δεδομένα τα οποία έχουν μικρή ή καθόλου αξία γεγονός που αυξάνει τον όγκο της βάσεως δεδομένων. Πρακτική αξία έχει και η χρονική στιγμή όπου αποκτώνται τα δεδομένα καθώς συχνά χάνονται επιχειρηματικές ευκαιρίες από την καθυστερημένη μάζωξη των κατάλληλων δεδομένων.

Σημαντικό παράγοντα που πρέπει να λαμβάνουν υπόψιν οι επιχειρήσεις αποτελούν τα προσωπικά δεδομένα που κατέχουν. Για το λόγο αυτό θα πρέπει να λαμβάνεται έγκριση από τους πελάτες για την χρήση τους. Την 27 Απριλίου 2016 θεσπίστηκε ο νέος γενικός κανονισμός προστασίας δεδομένων (General Data Protection Regulation, GDPR) της Ευρωπαϊκής Ένωσης. Οι αλλαγές σε σχέση με τον νόμο όπου αντικατέστησε ο GDPR σχετίζονται με τον τρόπο συλλογής, επεξεργασίας, με την αρχή λογοδοσίας αλλά και με το δικαίωμα στη λήθη. Σύμφωνα με τα παραπάνω, ένας πελάτης έχει το δικαίωμα ανά πάσα ώρα να απαιτήσει από τον υπεύθυνο επεξεργασίας χωρίς αδικαιολόγητη καθυστέρηση τη διόρθωση ή την διαγραφή ανακριβών ή μη δεδομένων προσωπικού χαρακτήρα που τον αφορούν. Επιπλέον, ένας πελάτης δικαιούται υπό προϋποθέσεις να εξασφαλίσει από τον υπεύθυνο επεξεργασίας τον περιορισμό της επεξεργασίας. Όσον αφορά τον υπεύθυνο ή τους υπεύθυνους επεξεργασίας πρέπει να μπορούν, ανά πάσα στιγμή, να αποδεικνύουν ότι είναι πλήρως συμμορφωμένοι με το GDPR.

Εν κατακλείδι, τα συστήματα CRM (Chalmeta, 2006) αντιπροσωπεύουν μια πελατοκεντρική επιχειρηματική στρατηγική η οποία ενσωματώνει δυναμικά τις πωλήσεις, το μάρκετινγκ και την εξυπηρέτηση πελατών έτσι ώστε να δημιουργήσει και να προσθέσει αξία στους πελάτες της επιχείρησης. Συγχρόνως, λόγω του γεγονότος ότι τα συστήματα CRM βασίζονται στην αμοιβαία ανταλλαγή πληροφοριών ανάμεσα στην επιχείρηση και στον πελάτη θα πρέπει να λαμβάνεται υπόψη η πιθανή ανησυχία των πελατών όσον αφορά την προστασία των δεδομένων τους και να παρέχεται ενημέρωση σχετικά με τη χρήση τους.

2.2 ΑΞΙΑ ΚΑΙ ΚΥΚΛΟΣ ΖΩΗΣ ΠΕΛΑΤΩΝ

Με το χρόνο η σχέση επιχείρησης - πελάτη εξελίσσεσαι περνώντας από διαδοχικές φάσεις. Ανάλογα τη φάση στην οποία ανήκουν οι πελάτες μπορεί να είναι:

- Προσδοκώμενοι
- Ανταποκρινόμενοι
- Νέοι
- Καθιερωμένοι
- Πρώην

Προσδοκώμενοι είναι οι πελάτες που ανήκουν στην αγορά-στόχο της επιχείρησης, ενώ οι ανταποκρινόμενοι πελάτες είναι εκείνοι όπου έχουν εκδηλώσει το ενδιαφέρον τους για κάποιο προϊόν ή υπηρεσία. Σε αντίθεση με τις προηγούμενες δύο κατηγορίες, οι νέοι πελάτες είναι εκείνοι όπου έχουν κάνει αγορά ενός προϊόντος ή υπηρεσίας και οι καθιερωμένοι ή πιστοί πελάτες εκείνοι όπου συνεχίζουν να προτιμούν και να εμπιστεύονται την επιχείρηση. Στους πρώην πελάτες ανήκουν οι πελάτες οι οποίοι έχουν αποχωρήσει εθελοντικά ή αναγκαστικά, καθώς και οι πελάτες που έχουν πάψει να ανήκουν στην αγορά-στόχο της επιχείρησης. Οι παραπάνω κατηγορίες αποτελούν τα στάδια του κύκλου ζωής ενός πελάτη από την απουσία στην απόκτηση και τέλος στην πιθανή αποχώρησή του.

Ανάλογα με την φάση στην οποία ανήκει ο εκάστοτε πελάτης, προσφέρει και διαφορετική αξία στην επιχείρηση. Με τον όρο αξία νοείται το οικονομικό όφελος όπου αποκομίζει η επιχείρηση από την ύπαρξη του συγκεκριμένου πελάτη. Συγκεκριμένα, χρησιμοποιείται η έννοια της αξίας του κύκλου ζωής ενός πελάτη (Customer Lifetime Cycle- CLC). Ορίζεται ως CLC η παρούσα αξία των καθαρών κερδών που απέκτησε η επιχείρηση από τη σχέση της με τον πελάτη ή της συστάδας πελατών (Buttle, 2009).

Στόχος των επιχειρήσεων είναι, μέσω συστημάτων CRM, να αποκτήσουν και να διατηρήσουν τους πιστούς πελάτες και να απαλλαγούν από εκείνους όπου είναι οικονομικά ασύμφοροι. Με αυτόν τον τρόπο επιτυγχάνουν να μεγιστοποιήσουν τον κύκλο ζωής των πιστών πελατών τους και σαν συνέπεια να αυξήσουν τα κέρδη της επιχείρησης.

Βασικό ρόλο για την προσέλκυση αλλά και την διατήρηση των πελατών διαδραματίζουν τα χαρακτηριστικά των προϊόντων ή υπηρεσιών αλλά και η

ικανοποίηση που αυτά προσφέρουν. Σύμφωνα με τους Kano et al. (1984) τα χαρακτηριστικά μπορούν να διακριθούν στις παρακάτω κατηγορίες :

- Βασικά χαρακτηριστικά (Must-be)
- Επιθυμητά χαρακτηριστικά (One dimensional)
- Ελκυστικά χαρακτηριστικά (Attractive)
- Αδιάφορα χαρακτηριστικά (Indifferent)
- Χαρακτηριστικά αναστροφής (Reverse)

Ένα χαρακτηριστικό με βασική ή αναμενόμενη ποιότητα χαμηλότερη από τις προσδοκίες του πελάτη θα τον δυσαρεστήσει, αλλά η ύπαρξη του δεν θα συμβάλει στην αύξηση της ικανοποίησης του. Όσον αφορά την επιθυμητή ποιότητα η ικανοποίηση του πελάτη είναι ανάλογη του επιπέδου πλήρωσης των απαιτήσεων στο χαρακτηριστικό αυτό. Δηλαδή, όσο μεγαλύτερο είναι το επίπεδο πλήρωσης τόσο πιο πολύ ικανοποιείται ο πελάτης. Ένα χαρακτηριστικό με ελκυστική ποιότητα μπορεί να προκαλέσει ενθουσιασμό και έκπληξη κάνοντας το προϊόν να ξεχωρίσει από άλλα στην ίδια κατηγορία. Αδιάφορο είναι ένα χαρακτηριστικό όταν ο πελάτης αδιαφορεί για την ύπαρξη του ή μη, ενώ η παρουσία χαρακτηριστικού αναστροφής προκαλεί δυσαρέσκεια στον πελάτη. Οι επιχειρήσεις εκμεταλλευόμενες την γνώση των αναγκών των καταναλωτών μπορούν να εστιάσουν σε εκείνα τα οποία προσφέρουν υψηλή ικανοποίηση.

Ακρογωνιαίος λίθος για την απόκτηση μιας πελατοκεντρικής στρατηγικής είναι η κατανόησή της συμπεριφοράς των καταναλωτών και οι πιθανές μεταβολές της. Η αυξημένη χρήση του διαδικτύου έχει οδηγήσει τους ανθρώπους στη χρήση πλατφορμών κοινωνικής δικτύωσης. Για το λόγο αυτό είναι συνετό οι επιχειρήσεις να χρησιμοποιούν νέες μορφές επικοινωνίας προκειμένου να υπάρχει πιο γρήγορη επικοινωνία μεταξύ πελάτη-επιχείρησης αλλά και να επεμβαίνουν αρνητικά σχόλια τα οποία πιθανό να γίνουν γρήγορα γνωστά και από το υπόλοιπο καταναλωτικό κοινό και πιθανόν επηρεάσουν την κρίση τους. Επιπλέον, οι ανάγκες για ευκολία έχουν οδηγήσει στην αύξηση των self-service επιλογών διευκολύνοντας έτσι τον πελάτη να αγοράσει ή να αναζητήσει οποιοδήποτε προϊόν ή υπηρεσία όποια στιγμή θελήσει.

Συνεπώς, όλες οι ενέργειες μιας σύγχρονης επιχείρησης, είτε παραγωγικές, είτε μάρκετινγκ, περιστρέφονται γύρω τους πελάτες. Καθίσταται αδήριτη η ανάγκη για εστίαση του συστήματος CRM της εκάστοτε επιχείρησης στον ανθρώπινο παράγοντα.

2.3 ΔΕΙΚΤΕΣ CRM

Η χρήση ενός συστήματος CRM χωρίς την ύπαρξη μέτρων αξιολόγησής του κρίνεται ατελέσφορη. Είναι συνετό να γίνεται αξιολόγηση για την ορθή ή μη λειτουργία του συστήματος CRM που χρησιμοποιείται, προκειμένου να αποτραπούν δυσμενείς ενέργειες από τη λανθασμένη χρήση και λειτουργία αλλά και η περαιτέρω βελτίωση στους τομείς όπου είναι εφικτό.

Οι δείκτες αυτοί αφορούν το brand είτε τους πελάτες της επιχείρησης δίνοντας με αυτό τον τρόπο μια ξεκάθαρη εικόνα στους manager για την αγορά αλλά και για τον κάθε μεμονωμένο πελάτη. Σύμφωνα με τους Kumar και Petersen (2012) οι δείκτες κατηγοριοποιούνται σε επίπεδο αγοράς, σε επίπεδο πελάτη και σε επίπεδο αγοράς-πελάτη. Ορισμένοι από αυτούς παρουσιάζονται παρακάτω.

Σε επίπεδο αγοράς μερικοί δημοφιλείς δείκτες είναι:

- Μερίδιο αγοράς: το ποσοστό πωλήσεων μιας επιχείρησης προς τις πωλήσεις των υπόλοιπων επιχειρήσεων σε συγκεκριμένο τμήμα της αγοράς.
- Ανάπτυξη πωλήσεων: η αύξηση/μείωση των πωλήσεων σε μια συγκεκριμένη περίοδο σε σχέση με αντίστοιχες προηγούμενες περιόδους.
- Ρυθμός απόκτησης πελατών: ρυθμός προσδοκώμενων πελατών που μετατρέπονται σε πελάτες.
- Ποσοστό επιβίωσης: αναλογία ατόμων που παραμένουν πελάτες μέχρι μια περίοδο t από την έναρξη παρατήρησης αυτών των πελατών.
- Μέση διάρκεια ζωής ενός πελάτη: μέση διάρκεια ζωής των ατόμων όπου παραμένουν πελάτες.
- Δείκτης win-back: ρυθμός προσέλκυσης χαμένων πελατών.

Σε επίπεδο πελάτη, οι ευρέως χρησιμοποιούμενοι δείκτες είναι:

- Πιθανότητα ενεργού πελάτη: πιθανότητα ένας πελάτης να αγοράσει ξανά από συγκεκριμένη επιχείρηση.
- Μερίδιο αγοραστικής δύναμης του πελάτη: χρηματική αξία όπου διαθέτει ένας πελάτης σε μια συγκεκριμένη κατηγορία προϊόντος σε σύγκριση με άλλες επιχειρήσεις.

- Αξία του κύκλου ζωής ενός πελάτη: αναλογία της συνολικής χρηματικής συνεισφοράς συγκεκριμένου πελάτη στα έσοδα της επιχείρησης σε συγκεκριμένη χρονική περίοδο.

Επιπλέον, χρησιμοποιούνται δείκτες επιπέδου αγοράς αλλά και πελάτη. Τέτοιοι μπορεί να είναι:

- Κόστος προσέλκυσης ενός πελάτη: το χρηματικό ποσό που δαπανήθηκε για την προσέλκυση ενός προσδοκώμενου πελάτη.
- Δείκτης διατήρησης ενός πελάτη: η μέση πιθανότητα ένας πελάτης να κάνει ξανά μια αγορά σε μια συγκεκριμένη χρονική περίοδο δεδομένου ότι ο εξεταζόμενος πελάτης έχει πραγματοποιήσει κάποια συναλλαγή τη χρονική περίοδο t-1.
- Δείκτης αγοραστικής δύναμης πελατών: αναλογία των συνολικών πωλήσεων των πελατών σε μια συγκεκριμένη κατηγορία προϊόντων στην συγκεκριμένη επιχείρηση προς τα συνολικά χρήματα όπου ξοδεύουν οι ίδιοι πελάτες στην συγκεκριμένη κατηγορία προϊόντων σε όλες τις επιχειρήσεις.
- Μερίδιο αγοράς συγκεκριμένης κατηγορίας προϊόντων: αναλογία των πωλήσεων συγκεκριμένη κατηγορίας προϊόντων της επιχείρησης προς τις συνολικές πωλήσεις στην αγορά.

Αναμφισβήτητα, για να έχει νόημα η χρήση οποιουδήποτε δείκτη θα πρέπει να έχει ορισθεί εκ των προτέρων η επιθυμητή τιμή του δείκτη έτσι ώστε υπάρχει κάποιο μέτρο σύγκρισης.

ΚΕΦΑΛΑΙΟ 3. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Παλαιότερα, για τους στατιστικολόγους και για τους οικονομολόγους, η έννοια της εξόρυξης δεδομένων ήταν άρρητα συνδεδεμένη με την αναζήτηση δεδομένων τέτοιων ώστε να υποστηρίζουν και να ταιριάζουν σε προκαθορισμένες ιδέες. Σήμερα, η πεποίθηση αυτή έχει αντικατασταθεί με την εύρεση χρήσιμων μοτίβων σε μεγάλους όγκους δεδομένων (Linnof και Berry, 2011). Παρακάτω αναλύεται η έννοια της εξόρυξης δεδομένων (data mining) και η πρακτική αξία των εργαλείων αυτών.

3.1 ΟΡΙΣΜΟΣ

Προκειμένου οι επιχειρήσεις να αναπτύξουν ένα ανταγωνιστικό σύστημα CRM οφείλουν να αναγνωρίζουν ποιους πελάτες θα πρέπει να προσελκύσουν αλλά και ποιους να διατηρήσουν. Για το σκοπό αυτό θα πρέπει να διαθέτουν και να είναι σε θέση να επεξεργαστούν μεγάλους όγκους δεδομένων με στοιχεία και ενέργειες των προσδοκώμενων και ήδη υπαρχόντων πελατών. Προκειμένου να εξάγουν χρήσιμα συμπεράσματα χρησιμοποιούνται εργαλεία εξόρυξης δεδομένων. Σύμφωνα με τον Drucker (1994) κύριος παράγοντας ανταγωνισμού θα καταστεί η επίδοση ενός ατόμου, μιας οργάνωσης, ενός βιομηχανικού κλάδου, ή μιας χώρας στην απόκτηση και εφαρμογή γνώσης.

Οι Linoff και Berry (2011) ορίζουν ως εξόρυξη δεδομένων μια επιχειρηματική διαδικασία για τη διερεύνηση μεγάλων όγκων δεδομένων με σκοπό την αναγνώριση ουσιωδών μοτίβων και κανόνων. Οι Han et al. (2012) αναφέρουν τις εφαρμογές εξόρυξης δεδομένων ως φυσική εξέλιξη της τεχνολογίας πληροφοριών (Information Technology). Οι βάσεις δεδομένων καθώς και η διαχείριση τους εξελίχθηκε σε

πληθώρα κρίσιμων λειτουργιών όπως συλλογή δεδομένων, δημιουργία βάσεων δεδομένων, διαχείριση δεδομένων και ανάλυσή τους.

Η ύπαρξη ενός αποτελεσματικού λειτουργικού συστήματος δεν οδηγεί απαραίτητα σε αποτελεσματικές προσπάθειες εξόρυξης δεδομένων. Σε ένα τυπικό λειτουργικό σύστημα οι διεργασίες και οι αναφορές γίνονται βάσει παλαιότερων δεδομένων σε αντίθεση με τις διαδικασίες εξόρυξης όπου χρησιμοποιούν τα πιο πρόσφατα δεδομένα για τον καθορισμό των μελλοντικών αποφάσεων της επιχείρησης. Το πρώτο επικεντρώνεται στις επιχειρηματικές δραστηριότητες (όπως λογαριασμοί, περιοχή, κωδικός προϊόντος κτλ.) και όχι στον πελάτη, ενώ το δεύτερο εστιάζει στον πελάτη, στα προϊόντα και στις περιοχές πωλήσεων. Επιπλέον ένα τυπικό λειτουργικό σύστημα είναι αναλυτικό, επαναλαμβανόμενο και εστιάζει σε μεμονωμένα πράγματα κάθε φορά ενώ οι διαδικασίες εξόρυξης δεδομένων είναι εφευρετικές και εξετάζουν μεγαλύτερες ομάδες δεδομένων.

Για την χρήση εργασιών εξόρυξης δεδομένων χρησιμοποιούνται ποικίλα υπολογιστικά εργαλεία. Το Rapid miner και το WEKA είναι δύο ευρέως χρησιμοποιούμενα συστήματα λογισμικού τα οποία προσφέρουν ποικιλία αλγορίθμων εξόρυξης δεδομένων. Ακόμα, το SPSS, SAS και STATISTICA είναι εργαλεία στατιστικής ανάλυσης. Για στατιστικούς υπολογισμούς, για ανάλυση και επεξεργασία δεδομένων ορισμένα από τα εργαλεία που χρησιμοποιούνται είναι η R και η MATLAB. Το λογισμικό της Oracle, το Information Harvester και το AlphaMiner είναι μερικά ακόμα πακέτα λογισμικού που χρησιμοποιούνται ευρέως. Για τον προγραμματισμό των αλγορίθμων χρησιμοποιείται συχνά η γλώσσα προγραμματισμού Python. Πρόκειται για μια αντικειμεντοστρεφή γλώσσα προγραμματισμού η οποία είνα γνωστή για την απλότητα της και για την ικανότητα της να συνδέει διαφορετικά τμήματα λογισμικού(glue language).

3.2 ΕΡΓΑΛΕΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η έννοια της εξόρυξης δεδομένων προϋπήρχε για δεκαετίες αλλά παρουσίασε ιδιαίτερη διάδοση στη δεκαετία του 1990. Οι λόγοι που οδήγησαν στη στροφή των επιχειρήσεων στον τομέα αυτό σχετίζονται με τη χρήση δεδομένων αλλά και με τεχνολογικούς περιορισμούς που υπήρχαν. Τα δεδομένα πλέον είναι ευκολότερο τόσο

να αποθηκευτούν όσο και να επεξεργαστούν εξάγοντας πληροφορίες. Επιπλέον, οι υπολογιστές είναι πιο οικονομικοί και εμπορικά λογισμικά εξόρυξης δεδομένων είναι εύκολα διαθέσιμα. Τέλος, οι επιχειρήσεις έχουν στραφεί στην ομαλή διαχείριση των πελατειακών σχέσεων γεγονός που εξαρτάται άμεσα με την εξόρυξη δεδομένων.

Τα εργαλεία εξόρυξης δεδομένων προκειμένου να εξάγουν καινούργια γνώση χρησιμοποιούν μεταξύ άλλων ένα σύνολο από εργασίες (Provost και Fawcett, 2003):

- Ταξινόμηση (Classification): Αποτελεί μια εργασία ανάθεσης ενός νεοεμφανιζόμενου αντικειμένου σε ένα σύνολο από προκαθορισμένες κλάσεις βάσει των χαρακτηριστικών του. Βασίζεται στην ιδέα της εκπαίδευσης ενός μοντέλου ταξινόμησης χρησιμοποιώντας ένα αντιπροσωπευτικό σύνολο δεδομένων (training set) με σκοπό τον ορισμό προτύπων για κάθε κατηγορία δεδομένων.
- Παλινδρόμηση (Regression): Σκοπός είναι, μέσω της εκπαίδευσης του αλγορίθμου, ο υπολογισμός και η πρόβλεψη μιας αριθμητικής τιμής για μια μεταβλητή. Για παράδειγμα, η πρόβλεψη της ζήτησης ενός προϊόντος αποτελεί πρόβλημα παλινδρόμησης.
- Εξαγωγή κανόνων συσχέτισης: Μέσω της συνάρτησης συσχέτισης γίνεται αναγνώριση των σχέσεων των οποίων υπάρχουν σε αντικείμενα τα οποία ανήκουν σε μια δεδομένη συλλογή. Το αποτέλεσμα της ομαδοποίησης αυτού του τύπου είναι η περιγραφή αντικειμένων που εμφανίζονται μαζί γι' αυτό και η μέθοδος καλείται και μέθοδος ανάλυσης καλαθιού (market-basket analysis).
- Ομαδοποίηση (Clustering): Είναι ο καταμερισμός ενός πληθυσμού σε υποομάδες (clusters) βάσει της ομοιότητας τους ώστε τα δεδομένα που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο μεταξύ τους, ενώ δεδομένα που ανήκουν σε διαφορετικές ομάδες να διαφέρουν.
- Πρόβλεψη συνδέσμων (Link prediction): Στόχος είναι η εύρεση συνδέσεων μεταξύ δεδομένων. Εφαρμόζεται στα δίκτυα κοινωνικής δικτύωσης με σκοπό να προτείνει φίλους, προϊόντα και υπηρεσίες σε κάποιο μέλος.
- Ανάλυση συμπεριφοράς (profiling): Στόχος είναι ο χαρακτηρισμός ενός μεμονωμένου ατόμου ή ενός πληθυσμού βάσει της συμπεριφοράς τους. Η μέθοδος αυτή χρησιμοποιείται για την ανίχνευση κάποιας απάτης η οποία ανιχνεύεται μ την απόκλιση της σε σχέση με τις προηγούμενες.

Οι εφαρμογές εξόρυξης δεδομένων βοηθούν στην επιλογή των πελατών-στόχων ή στην αναγνώριση συστάδων πελατών με παρόμοια καταναλωτική συμπεριφορά. Η

αναγνώριση των πελατών που σκοπεύουν να φύγουν αποτελεί μια ακόμα εφαρμογή της εξόρυξης δεδομένων. Στοχεύοντας σε συγκεκριμένους πελάτες, είτε για τη προσέλκυση των επικερδών, είτε για την απομάκρυνση των επιζήμιων, οδηγεί στη μείωση του κόστους λόγω μάρκετινγκ.

Μέσω των εργαλείων εξόρυξης δεδομένων η επιχείρηση στοχεύει στην εύρεση των ιδανικών δράσεων προώθησης των προϊόντων ή των υπηρεσιών της. Ένας ακόμα επιχειρηματικός στόχος είναι η διατήρηση των πιο επικερδών πελατών με συνέπεια την αύξηση του κύκλου ζωής των πελατών και την κερδοφορία της επιχείρησης. Επιπλέον είναι δυνατό να μειωθεί η έκθεση στον κίνδυνο ασυνέπειας των πελατών (default risk). Μια μέθοδος για να επιτευχθεί αυτό είναι η ανάλυση επιβίωσης (survival analysis). Στην ανάλυση αυτή το αποτέλεσμα είναι το αν πραγματοποιήθηκε ή όχι ένα γεγονός αλλά και το πότε πραγματοποιήθηκε. Στην περίπτωση αυτή υπολογίζεται πότε και αν ένας πελάτης πρόκειται να αθετήσει την πληρωμή του προς την επιχείρηση.

Παρ' όλα αυτά ο πιο συχνός λόγος χρησιμοποίησης εργαλείων εξόρυξης δεδομένων είναι η στοχευμένη προώθηση προϊόντων ή υπηρεσιών σε συγκεκριμένη ομάδα ατόμων. Κατ' αυτόν τον τρόπο εισάγεται η έννοια του μάρκετινγκ η οποία κατηγοριοποιείται σε άμεσο και μαζικό μάρκετινγκ. Στο μαζικό μάρκετινγκ χρησιμοποιούνται μαζικά μέσα ενημέρωσης, όπως τηλεόραση, ραδιόφωνο, περιοδικά, κτλ., για την αναπαραγωγή πληροφοριών σχετικών με το προϊόν που προωθείται. Δηλαδή, το μαζικό μάρκετινγκ στοχεύει σε μεγάλο πλήθος ατόμων. Αντιθέτως, το άμεσο μάρκετινγκ στοχεύει σε μεμονωμένα άτομα.

Ως άμεσο μάρκετινγκ ορίζεται η παράδοση ενός μηνύματος ή μιας πρότασης μάρκετινγκ σε ένα συγκεκριμένο πελάτη ή έναν υποψήφιο πελάτη, με μια ευνοϊκή προσφορά για τον πελάτη μέσω τηλεφώνου και άλλων μέσων χωρίς τη μεσολάβηση ενδιάμεσου προσώπου ή έμμεσων μέσων (Roddy, 2002). Οι Fahy και Jobber (2014) χρησιμοποιούν την έννοια του άμεσου μάρκετινγκ για να περιγράψουν την διανομή των προϊόντων, των πληροφοριών και τα οφέλη προώθησης σε καταναλωτές στόχους μέσω διαδραστικής επικοινωνίας με τρόπο που να επιτρέπει τη μέτρηση αντιδράσεων.

Οι υπεύθυνοι του άμεσου μάρκετινγκ προκειμένου να αλληλεπιδράσουν με τους καταναλωτές χρησιμοποιούν ένα μεγάλο εύρος μέσων. Τέτοια μέσα είναι (Fahy & Jobber, 2014. Πανηγυράκης & Κορωνάκη & Μπατσίλα, 2015) :

- Μάρκετινγκ μέσω ταχυδρομείου: Η προώθηση ενός προϊόντος ή/και διατήρηση μιας διαρκούς σχέσης γίνεται με τη στοχευμένη αποστολή διαφημιστικών εντύπων

σε υφιστάμενους ή/και δυνητικούς πελάτες. Κύριο πλεονέκτημα της μεθόδου είναι το κόστος, το οποίο είναι αλληλένδετο με την ποιότητα της ταχυδρομικής λίστας που θα χρησιμοποιηθεί. Καλής ποιότητας λίστα χαμηλώνει το κόστος, ενώ σε αντίθετη περίπτωση όχι μόνο αυξάνεται το κόστος, αλλά και δημιουργεί αρνητική αντίληψη στους αποδέκτες της.

- Τηλεμάρκετινγκ: Βασίζεται στη χρήση τηλεπικοινωνιών και μπορεί να είναι εισερχόμενο αλλά και εξερχόμενο. Στο εξερχόμενο η διαφήμιση γίνεται ανεξάρτητα από το αν ενδιαφέρεται ο πελάτης και η επικοινωνία γίνεται από την επιχείρηση προς τον πελάτη. Αντιθέτως, στο εισερχόμενο τηλεμάρκετινγκ σκοπός είναι η δημιουργία περιεχομένου σχετικού με αυτό που θέλει ο πελάτης έτσι ώστε ο πελάτης να έρθει σε επαφή με την επιχείρηση (π.χ. με το να μπει στην ιστοσελίδα της επιχείρησης). Βασικά πλεονεκτήματα αποτελούν το χαμηλότερο κόστος και ο λιγότερος χρόνος που απαιτείται σε σχέση με τις προσωπικές επισκέψεις..
- Διαφήμιση άμεσης ανταπόκρισης: Σε αυτή τη μέθοδο χρησιμοποιούνται τα μέσα μαζικής ενημέρωσης, όπως τηλεόραση, περιοδικά και ραδιόφωνο. Η άμεση ανταπόκριση των καταναλωτών επιτυγχάνεται είτε με τη χρήση κάποιου τηλεφωνικού νούμερου χωρίς χρέωση, είτε με κάποιου κουπονιού στην περίπτωση έντυπης μορφής διαφήμισης.
- Μάρκετινγκ μέσω καταλόγων: Σε αυτό το κανάλι άμεσου μάρκετινγκ γίνεται διανομή καταλόγων σε αντιπροσώπους ή πελάτες. Με αυτό τον τρόπο διευκολύνεται η διαδικασία επιλογής προϊόντων και διευρύνεται ο δυνατός διαθέσιμος χρόνος παρακολούθησης προϊόντων. Όσον αφορά τις εταιρείες, σημαντικό μειονέκτημα αποτελεί το υψηλό κόστος των καταλόγων και η ανάγκη για συνεχή ενημέρωση τους. Οι καταναλωτές δεν έχουν την δυνατότητα χρήσης ή δοκιμής των προϊόντων πριν προβούν σε αγορά και μπορεί να υπάρχει σημαντική χρωματική διαφορά των απεικονιζόμενων προϊόντων στον κατάλογο σε σχέση με την πραγματικότητα.
- Ψηφιακά μέσα: Πρόκειται για αγορές που πραγματοποιούνται μέσω διαδικτύου. Στα ψηφιακά μέσα οι επιχειρηματικές ευκαιρίες μπορούν να γίνονται αμφίδρομα, δηλαδή από μια εταιρεία προς μια άλλη (business to business, B2B) ή προς κάποιον καταναλωτή (business to consumer, B2C) αλλά και από τον καταναλωτή προς την εταιρεία (consumer to business, C2B) κάποιον άλλον καταναλωτή (consumer to consumer, C2C). Το ψηφιακό μάρκετινγκ μπορεί να είναι διαδικτυακό, διαφημίσεις

που συνδέονται με μηχανή αναζήτησης (search advertising), ηλεκτρονικού ταχυδρομείου, με χρήση μέσων κοινωνικής δικτύωσης, ιογενές (viral marketing), μέσω φορητών συσκευών και με διαδραστική τηλεοπτική διαφήμιση.

- Μάρκετινγκ εκδηλώσεων (event marketing): Σε αυτό το κανάλι επικοινωνίας οι εταιρείες έρχονται κοντά με τους καταναλωτές και αλληλεπιδρούν μαζί τους. Το μάρκετινγκ εκδηλώσεων μπορεί να διακριθεί σε διοργάνωση εκδηλώσεων, σε χορηγία εκδηλώσεων και σε υποστήριξη εκδηλώσεων μέσω διανομής δειγμάτων ή άλλων παροχών.

Απόρροια των παραπάνω αποτελεί το ερώτημα της επιλογής των πελατών που θα επιλεχθούν για την εφαρμογή του άμεσου μάρκετινγκ. Απάντηση στο ερώτημα αυτό δίνεται από τις μεθόδους ανίχνευσης πελατών (detection analysis) όπου δημιουργούν μοντέλα και τα εφαρμόζουν σε ήδη υπάρχοντα δεδομένα πελατών προκειμένου να υπολογίσουν την επιρροή σε κάποια προωθητική ενέργεια.

Σε έρευνα που διεξήχθη (Santos Garcia et al., 2019), ανάμεσα σε 1278 περιπτώσεις άρθρων, έγινε κατηγοριοποίηση σε 572 ομάδες που αφορούσαν μια περιοχή εφαρμογής των εργαλείων εξόρυξης δεδομένων, ενώ 12 ομάδες ήταν σχετικές με περισσότερες περιοχές εφαρμογής. Για κάθε μια από τις 6 μεγαλύτερες ομάδες έγινε εμπειρική επιλογή της κύριας συνεισφοράς της διαδικασίας εξόρυξης δεδομένων στην περιοχή εφαρμογής της ομάδας. Οι κορυφαίες ομάδες αναπαριστούν το 79,41% των συνολικών άρθρων που μελετήθηκαν και σε αυτές ανήκει η ιατροφαρμακευτική περίθαλψη, η τεχνολογία πληροφοριών και επικοινωνίας (information communication technology, ICT), η βιομηχανία, η εκπαίδευση, τα οικονομικά και η λογιστική. Στην ιατροφαρμακευτική περίθαλψη οι κύριες εφαρμογές αφορούν αγωγές των ασθενών, κλινικά μονοπάτια ή τις διαδικασίες ενός νοσοκομείου. Στο χώρο του ICT, τα εργαλεία εξόρυξης χρησιμοποιούνται στην ανάπτυξη λογισμικού, σε υπηρεσίες της τεχνολογίας πληροφοριών και σε εταιρείες τηλεπικοινωνιών. Στην αυτοκινητοβιομηχανία ανήκει ένα μεγάλο κομμάτι εφαρμογής των εργαλείων εξόρυξης δεδομένων. Όσον αφορά την εκπαίδευση, βασικό πεδίο εφαρμογής του DM είναι στην ηλεκτρονική εκπαίδευση (e-learning), στις επιστημονικές εφαρμογές αλλά και στα ερευνητικά κέντρα. Επιπλέον, μεγάλη εφαρμογή έχουν τα εργαλεία DM στις τράπεζες, σε εταιρείες ασφάλισης και σε διαδικασίες σχετικές με επενδύσεις, ανάλυση κινδύνων, καταθέσεις και άλλες οικονομικές συναλλαγές. Τέλος, αναφέρονται εφαρμογές σχετικές με λογιστική, μεταφορές και διαχείρισης αποθεμάτων.

3.3 ΗΘΙΚΟΣ ΚΥΚΛΟΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων είναι μια συνεχής επιχειρησιακή διεργασία και συνεπώς δεν αρκεί η δημιουργία αναλυτικών μοντέλων για την αναγνώριση χρήσιμων για την επιχείρηση μοτίβων. Αντιθέτως, θα πρέπει η επιχείρηση να αναγνωρίζει τα μοτίβα, να ανταποκρίνεται σε αυτά, να μετατρέπει τα δεδομένα σε πληροφορία, την πληροφορία σε δράση, και τη δράση σε αξία για την επιχείρηση. Η διαδικασία αυτή ονομάζεται ηθικός κύκλος εξόρυξης δεδομένων(virtuous cycle of data mining). Πρόκειται δηλαδή για μια επαναλαμβανόμενη διαδικασία η οποία έχει σκοπό την απόκτηση του μεγίστου οφέλους για την επιχείρηση.

Σύμφωνα με τον Rajola (2013), το πρώτο στάδιο για την ανακάλυψη γνώσης είναι ο ορισμός του στόχου της επιχείρησης και του τομέα εφαρμογής του. Στόχος είναι οι ξεκάθαρες απαιτήσεις για τους χρήστες του συστήματος αλλά και οι ξεκάθαρες προδιαγραφές του. Ο στόχος αυτός μπορεί να συνδέεται άμεσα με τις επιχειρηματικές ευκαιρίες που παρουσιάζονται, όπως ο σχεδιασμός ενός νέου προϊόντος και οι καμπάνιες άμεσου μάρκετινγκ. Κατά το στάδιο αυτό πρέπει να γίνεται και ο ορισμός των ορίων των παραμέτρων επιτυχίας. Επιπλέον, πρέπει να γίνεται υπολογισμός του κόστους αλλά και του κέρδους που θα αποφέρει η ανάλυση έτσι ώστε να αποσαφηνιστεί η αξία της ανάλυσης.

Το επόμενο στάδιο αφορά τη συλλογή δεδομένων που θα χρησιμοποιηθούν κατά την ανάλυση. Οι Berson και Thearling (2000) αναφέρουν ότι το στάδιο αυτό περιλαμβάνει τον ακριβή υπολογισμό του όγκου δεδομένων που απαιτείται, τον ορισμό της μεθόδου δειγματοληψίας εφόσον χρησιμοποιηθεί καθώς και μια συνοπτική περιγραφή των σχέσεων μεταξύ των δεδομένων. Στη συνέχεια ακολουθεί ο καθαρισμός των δεδομένων και η προεπεξεργασία τους έτσι ώστε να διασφαλισθεί η καταλληλότητα τους για την χρήση εργασιών εξόρυξης δεδομένων. Σε επίπεδο επεξεργασίας δεδομένων σημαντικό βήμα αποτελεί η αναγνώριση και η επιλογή των συναφών μεταβλητών. Ακολουθεί η εκπαίδευση των αλγορίθμων, η σύγκριση τους και τελικά επιλογή του κατάλληλου με σκοπό την εύρεση των κατάλληλων πελατών που ικανοποιούν τα κριτήρια του στόχου που έχει θέσει η επιχείρηση. Τέλος, ακολουθεί η δράση δηλαδή η αξιοποίηση των μοντέλων, η παρακολούθηση της διαδικασίας καθώς και η βελτίωση τους. Παράλληλα θα πρέπει να ελέγχονται τα αποτελέσματα, να συγκρίνονται με τις παραμέτρους επιτυχίας που έχουν τεθεί και να οδηγούνται σε

επαναπροσδιορισμό του στόχου αλλά και στην επανάληψη της διαδικασίας εφόσον απαιτείται.

Οι Kumar και Reinartz (2006) υποστηρίζουν ότι σημαντικό βήμα στην παραπάνω διαδικασία αποτελεί η αρχειοθέτηση όλων των πληροφοριών που σχετίζονται με την εργασία εξόρυξης δεδομένων. Με αυτόν το τρόπο, οι επιχειρήσεις μπορούν να εντοπίζουν αποκλίσεις κατά την εκτέλεση των μοντέλων καθώς και σε περίπτωση μη αναμενόμενων αποτελεσμάτων τι μπορεί να συνέβη αλλά και γιατί.

Η μέτρηση της ποιότητας του μοντέλου που θα χρησιμοποιηθεί είναι αναπόσπαστο κομμάτι της παραπάνω διαδικασίας. Η ποιότητα ενός μοντέλου (van der Aalst, 2016) περιγράφεται λαμβάνοντας υπόψιν τις τέσσερις διαστάσεις ποιότητας, δηλαδή την καταλληλότητα (fitness), την ακρίβεια (precision), τη γενίκευση (generalization) και την απλότητα (simplicity). Συγκεκριμένα, καταλληλότητα ορίζεται η ικανότητα παρατήρησης της συμπεριφοράς του αρχείου καταγραφής ενός ήδη μελετημένου μοντέλου. Ακρίβεια είναι η ποιότητα αποφυγής συμπεριφορών μη σχετικές με το συνήθη αρχείο καταγραφών, ενώ γενίκευση είναι η δυνατότητα αποδοχής παρόμοιων γεγονότων τα οποία σχετίζονται με προηγούμενα γεγονότα. Τέλος, είναι προτιμότερο το μοντέλο να είναι όσο πιο απλό γίνεται.

ΚΕΦΑΛΑΙΟ 4: ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

4.1 ΟΡΙΣΜΟΣ

Η μηχανική μάθηση (Mohri et al., 2012) είναι υπολογιστικές μέθοδοι, οι οποίες χρησιμοποιώντας δεδομένα μπορούν να παρέχουν προβλέψεις και εκτιμήσεις υψηλής ακρίβειας. Σύμφωνα με τον Mitchell (1997) ένα πρόγραμμα υπολογιστή μπορεί να μάθει από την εμπειρία E (experience) όσον αφορά μια κατηγορία εργασιών T (tasks) και ενός μέτρου απόδοσης P (performance measure), εάν η απόδοση στις εργασίες της T, όπως μετριούνται με την P, βελτιώνονται με την εμπειρία E.

Πολλές φορές απαιτείται ο συνδυασμός αλγορίθμων μηχανικής μάθησης, γεγονός που καθιστά απαραίτητη την ανάγκη για οργάνωση της γνώσης που αποκτήθηκε. Σύμφωνα με τον Langley (1996) η οργάνωση της γνώσης μπορεί να κατηγοριοποιηθεί σε:

- Λίστα απόφασης (decision list): Σε αυτή τη δομή, οι πληροφορίες είναι ταξινομημένες με αποτέλεσμα πληροφορίες που βρίσκονται πιο ψηλά στη λίστα να λαμβάνονται νωρίτερα υπόψιν κατά την εκτέλεση του συστήματος. Μερικές μεταβλητές μπορούν να αποθηκευτούν σε αλληλοαποκλειόμενη σειρά, δηλαδή η ύπαρξη της μιας πληροφορίας να προϋποθέτει την απουσία της άλλης. Αυτή η μέθοδος κρίνεται κατάλληλη για την οργάνωση ανταγωνιστικών στοιχείων.
- Δίκτυο συμπερασμάτων (inference network): Τα δεδομένα οργανώνονται σε μορφή δένδρου ή κατευθυνόμενου γράφου, δηλαδή η προέκταση κάθε κόμβου επηρεάζεται από τις συνδέσεις του με τους κόμβους κάτω από αυτό.
- Ιεραρχία ιδεών (concept hierarchy): Στο δίκτυο αυτό, η οργάνωση των δεδομένων γίνεται με μορφή δένδρου ή κατευθυνόμενου γράφου. Η διαφορά με την οργάνωση σε δίκτυο συμπερασμάτων είναι ότι στην περίπτωση αυτή κάθε κόμβος αντιστοιχεί

με μία ιδέα συνοδευόμενη από μια σχετική περιγραφή της. Κόμβοι με γενικές περιγραφές κατατάσσονται υψηλότερα σε ιεραρχία, ενώ κόμβοι με πιο συγκεκριμένες περιγραφές κατατάσσονται σε χαμηλότερα επίπεδα.

4.2 ΚΑΤΗΓΟΡΙΕΣ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ένα χαρακτηριστικό που περιγράφει τους αλγόριθμους μηχανικής μάθησης είναι ο βαθμός επίβλεψης τους. Τα μοντέλα μηχανικής μάθησης χωρίζονται σε δυο μεγάλες κατηγορίες, στη μάθηση με επίβλεψη (supervised learning) και στη μάθηση χωρίς επίβλεψη (unsupervised learning).

4.2.1 ΜΑΘΗΣΗ ΜΕ ΕΠΙΒΛΕΨΗ

Η μάθηση με επίβλεψη συνιστά ένα είδος μάθησης κατά το οποίο προσεγγίζεται μια άγνωστη συνάρτηση χρησιμοποιώντας δεδομένα τα οποία αποτυπώνουν τα αποτελέσματα της συνάρτησης αυτής. Στους αλγορίθμους δίνονται πληροφορίες εισόδου και γνωστά αποτελέσματα και βάσει των σχέσεων τους καθορίζεται το μοντέλο που τα περιγράφει. Δηλαδή, το μοντέλο πρόβλεψης δημιουργεί μια συνάρτηση που συνδέει τα δεδομένα εισόδου με τον στόχο. Η παραπάνω διαδικασία συνιστά τη φάση της εκπαίδευσης. Στη συνέχεια ένα νέο σύνολο δεδομένων εισάγεται στο μοντέλο που δημιουργήθηκε και έτσι ο αλγόριθμος μπορεί να κάνει προβλέψεις. Ο στόχος είναι δηλαδή η πρόβλεψη των αποτελεσμάτων βάσει μετρήσεων σε ένα πλήθος δεδομένων εισόδου.

4.2.1.1 Μοντέλα ταξινόμησης

Τα μοντέλα ταξινόμησης αφορούν τη δημιουργία μοντέλων τα οποία μπορούν να κατηγοριοποιήσουν δεδομένα τα οποία δεν έχουν ακόμα κατηγοριοποιηθεί. Δηλαδή, στόχος είναι η εύρεση των ορίων απόφασης που καθορίζουν πως θα διαχωρισθούν οι

κλάσεις. Οι Tan et al.(2014) ορίζουν ως ταξινόμηση τη διαδικασία μάθησης μιας συνάρτησης f, η οποία αντιστοιχεί κάθε διάνυσμα μεταβλητών x σε ένα προκαθορισμένο πλήθος κλάσεων y.

Τα μοντέλα ταξινόμησης είναι πιο αποτελεσματικά όταν εφαρμόζονται σε δυαδικά δεδομένα (binary) ή κατηγορικά (nominal) και λιγότερο αποτελεσματικά όταν υπάρχουν ιεραρχικά δεδομένα (ordinal). Αυτό οφείλεται στο γεγονός ότι δε λαμβάνουν υπόψιν την ύπαρξη σειράς ανάμεσα στις κατηγορίες.

Ένα μοντέλο ταξινόμησης (Tan et al., 2014) μπορεί να χρησιμοποιηθεί ως επεξηγηματικό εργαλείο για τη διάκριση αντικειμένων που ανήκουν σε διαφορετικές κλάσεις αλλά και ως εργαλείο πρόβλεψης (descriptive modeling) κλάσεων νέων εγγραφών (predictive modeling). Ορισμένες κατηγορίες αλγορίθμων ταξινόμησης είναι:

- Δέντρα απόφασης (decision trees): Πρόκειται για μια από τις πιο δημοφιλείς μεθόδους ταξινόμησης και αναπαρίσταται με μια δομή δέντρου. Στην περίπτωση που η επιλεγμένη για προσδιορισμό μεταβλητή λαμβάνει διακριτές τιμές το δέντρο απόφασης ονομάζεται δέντρο ταξινόμησης (classification tree), ενώ στην περίπτωση που λαμβάνει συνεχείς τιμές καλείται δέντρο παλινδρόμησης (regression tree). Στο δέντρο υπάρχει ο κόμβος-ρίζα, οι εσωτερικοί κόμβοι και οι εξωτερικοί κόμβοι-φύλλα. Η ρίζα έχει μόνο εξερχόμενες ακμές και συνδέεται με κόμβους χαμηλότερου επιπέδου. Για τη μετάβαση από ένα κόμβο ενός επιπέδου σε κάποιον άλλο χαμηλότερου πραγματοποιείται έλεγχος μιας συνθήκης και στη συνέχεια διάσπαση των δεδομένων σε ένα ή περισσότερα υποσύνολα. Η διαδικασία μετάβασης σε χαμηλότερο επίπεδο επαναλαμβάνεται μέχρι όλα τα χαρακτηριστικά να εισαχθούν στους κόμβους του δέντρου. Τα φύλλα βρίσκονται στο κατώτερο επίπεδο κάθε κλάδου και αντιπροσωπεύουν τις προβλέψεις των τιμών του επιθυμητού χαρακτηριστικού όπως έχουν προκύψει από το μονοπάτι με αφετηρία τη ρίζα και προορισμό το κάθε φύλλο.
- Νευρωνικά δίκτυα (neural networks): Τα νευρωνικά δίκτυα είναι εμπνευσμένα από το βιολογικό νευρικό σύστημα και συγκεκριμένα από τον εγκέφαλο. Ένα νευρωνικό δίκτυο μπορεί να περιγραφεί από έναν κατευθυνόμενο γράφο, του οποίου οι κόμβοι καλούνται νευρώνες και οι συνδέσεις αναπαριστούν τις σχέσεις μεταξύ των νευρώνων. Μέσω των συνδέσεων μεταφέρονται πληροφορίες από τις εισόδους του δικτύου προς τις εξόδους του. Οι συνδέσεις έχουν βάρη (weight), τα οποία καθορίζουν το βαθμό αλληλεπίδρασης για κάθε ζεύγος νευρώνων. Ένα δίκτυο που

περιλαμβάνει αμφίδρομες συνδέσεις χαρακτηρίζεται αναδρομικό (recurrent), ενώ ένα δίκτυο χωρίς αμφίδρομες συνδέσεις χαρακτηρίζεται απλής εμπροσθοτροφοδοτούμενο (feed forward). Τέλος, τα νευρωνικά δίκτυα μπορούν να εφαρμοσθούν και για μη επιβλεπόμενη μάθηση.

- Μπαϋσιανά δίκτυα (Bayesian networks): Στη μέθοδο αυτή χρησιμοποιούνται πιθανοτικά μοντέλα τα οποία βασίζονται στο θεώρημα του Bayes. Βασική δυσκολία της μεθόδου είναι η ανάγκη γνώσης πολλών τιμών πιθανοτήτων η οποία πολλές φορές οδηγεί στην αντικατάστασή τους με εκτιμήσεις από παλαιότερες υποθέσεις ή από εμπειρική γνώση. Επιπλέον, για την αντιμετώπιση της παραπάνω δυσκολίας χρησιμοποιούνται συχνά τα απλά δίκτυα Bayes (simple/naive Bayes classifier) τα οποία θεωρούν ανεξαρτησία ανάμεσα στα χαρακτηριστικά.

4.2.1.2 Παλινδρόμηση

Η παλινδρόμηση (regression) αφορά τη διαδικασία διερεύνησης των σχέσεων μεταξύ της μεταβλητής απόκρισης (response variable) ή εξαρτημένης μεταβλητή (dependent variable) με ένα σύνολο επεξηγηματικών μεταβλητών (explanatory variables) ή ανεξάρτητων μεταβλητών (independent variables). Ανάλογα με τον αριθμό των επεξηγηματικών μεταβλητών με τις οποίες εξαρτάται η μεταβλητή απόκρισης και με την προϋπόθεση ότι η εξαρτημένη μεταβλητή λαμβάνει αριθμητικές τιμές η παλινδρόμηση διακρίνεται στις κατηγορίες:

- Απλή γραμμική παλινδρόμηση: Η εξαρτημένη μεταβλητή εξαρτάται μόνο από μια ανεξάρτητη μεταβλητή.
- Πολλαπλή γραμμική παλινδρόμηση: Στην περίπτωση αυτή η εξαρτημένη μεταβλητή εξαρτάται από περισσότερες μεταβλητές.
- Μη γραμμική παλινδρόμηση: Η σχέση ανάμεσα στην εξαρτημένη μεταβλητή και στις ανεξάρτητες μεταβλητές είναι μη γραμμική.

Στην περίπτωση όπου η τιμή της εξαρτημένης μεταβλητής λαμβάνει ονομαστικές τιμές τότε γίνεται χρήση μοντέλων ταξινόμησης.

4.2.2 ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ

Στη μη επιβλεπόμενη μάθηση (unsupervised models) είναι διαθέσιμα μόνο δεδομένα εισόδου χωρίς καταγραφή των αποτελεσμάτων εξόδου. Δηλαδή, στόχος είναι η αναγνώριση και η ομαδοποίηση των δεδομένων σε ομάδες που δεν είναι γνωστές εκ των προτέρων.

4.2.2.1 Μοντέλα Συσταδοποιησης

Η συσταδοποίηση (cluster analysis) αφορά το διαμερισμό των δεδομένων σε συστάδες με όμοια χαρακτηριστικά. Κατά τη συσταδοποίηση γίνεται προσπάθεια βελτιστοποίησης ενός κριτηρίου διαχωρισμού έτσι ώστε να επιτευχθεί όσο το δυνατόν καλύτερη ομαδοποίηση. Δηλαδή, στόχος είναι η μεγάλη ομοιογένεια ανάμεσα στα δεδομένα μιας συστάδας και μεγάλη διαφοροποίηση των συστάδων μεταξύ τους. Ορισμένοι από τις πιο διαδεδομένες μεθόδους συσταδοποίησης είναι :

- K-means: Πρόκειται για έναν αλγόριθμο που χωρίζει το σύνολο των δεδομένων σε μη επικαλυπτόμενες συστάδες και είναι βασιζόμενο σε πρότυπα (prototype based). Δηλαδή, ένα αντικείμενο μπορεί να αντιστοιχεί σε ακριβώς μια συστάδα και βρίσκεται κοντά με το πρότυπο που καθορίζει τη συστάδα αλλά μακριά με τα πρότυπα των άλλων συστάδων. Αρχικά, ο χρήστης επιλέγει τον επιθυμητό αριθμό των συστάδων K, δηλαδή τα αρχικά κεντροειδή (centroids). Στη συνέχεια γίνεται η ανάθεση των αντικειμένων στις συστάδες και υπολογίζονται τα νέα γεωμετρικά κέντρα κάθε συστάδας βάσει του μέσου όρου όλων των σημείων της κάθε συστάδας. Η διαδικασία επαναλαμβάνεται έως ότου η συσταδοποίηση των δεδομένων σταθεροποιηθεί.
- Συσσωρευτική ιεραρχική συσταδοποίηση (agglomerative hierarchical clustering): Η ομαδοποίηση είναι ιεραρχική, δηλαδή υπάρχουν ένθετες συστάδες οι οποίες οργανώνονται σε μορφή δέντρου (graph based). Αρχικά όλα τα αντικείμενα θεωρούνται μεμονωμένες συστάδες οι οποίες στη συνέχεια συγχωνεύονται σταδιακά, με γνώμονα την ομοιότητα τους. Η διαδικασία συνεχίζεται έως ότου όλα τα αντικείμενα να καταλήξουν σε μια συστάδα.
- DBSCAN: Πρόκειται για ένα μοντέλο όπου ο διαχωρισμός σε συστάδες βασίζεται στην πυκνότητα των αντικειμένων (density based). Μια συστάδα απαρτίζεται από

αντικείμενα πολύ συσσωρευμένα μεταξύ τους η οποία περιβάλλεται από περιοχή με διασκορπισμένα αντικείμενα, δηλαδή περιβάλλεται από περιοχή χαμηλής πυκνότητας. Αρχικά, επιλέγεται τυχαία ένα αρχικό σημείο εκκίνησης και στη συνέχεια δημιουργείται μια γειτονιά βάσει της ακτίνας που έχει ορισθεί θεωρώντας κέντρο το αντικείμενο εκκίνησης. Στην περίπτωση που η γειτονιά περιέχει αρκετά αντικείμενα δημιουργείται μια συστάδα, ενώ σε διαφορετική περίπτωση θεωρείται προσωρινός θόρυβος και επανεξετάζεται όταν επιλεχθεί νέο σημείο εκκίνησης. Εφόσον τηρείται η προϋπόθεση της ακτίνας, κάθε αντικείμενο μέσα στην συστάδα εισάγει αντικείμενα που ανήκουν σε γειτονιές του. Η διαδικασία επαναλαμβάνεται για κάθε αντικείμενο και ολοκληρώνεται όταν δεν μπορούν να προστεθούν άλλα αντικείμενα. Στη συνέχεια επιλέγεται εκ νέου ένα τυχαίο αντικείμενο που δεν έχει ελεγχθεί προηγουμένως και γίνεται επανάληψη της διαδικασίας μέχρι όλα τα αντικείμενα να ανατεθούν σε συστάδες.

4.2.2.2 Αλγόριθμοι εξαγωγής κανόνων συσχέτισης

Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης (association rules) στοχεύουν στην εύρεση κανόνων ή συσχετίσεων ανάμεσα στα δεδομένα μελετώντας την επανεμφάνιση γεγονότων στα δεδομένα. Βασική εφαρμογή των συγκεκριμένων αλγορίθμων είναι η ανάλυση της καταναλωτικής συμπεριφοράς των πελατών βάσει της ταυτόχρονης αγοράς προϊόντων ή υπηρεσιών. Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης μπορεί να αναζητούν είτε θετικές, είτε αρνητικές συσχετίσεις. Για παράδειγμα, η θετική συσχέτιση μπορεί να αφορά αναζήτηση προϊόντων ή υπηρεσιών που αγοράσθηκαν μαζί, ενώ αρνητική συσχέτιση μπορεί να αφορά αντικείμενα αλληλοαποκλειόμενης αγοράς.

Η διαδικασία ξεκινά με τον υπολογισμό των ποσοστών των συναλλαγών που περιέχουν δυο αντικείμενα μαζί, δηλαδή την υποστήριξη (support), καθώς και τον ορισμό της ελάχιστη αποδεκτής τιμής της για κάθε μια από τις συναλλαγές. Επιπλέον υπολογίζεται, για κάθε συναλλαγή που περιέχει ένα αντικείμενο, το ποσοστό εκείνων όπου εμφανίζεται ένα άλλο αντικείμενο και αποτελούν την εμπιστοσύνη κάθε συνδυασμού (confidence). Στη συνέχεια γίνεται εντοπισμός των πιο συχνών συνόλων στοιχείων υπό την προϋπόθεση να έχουν υποστήριξη ίση ή μεγαλύτερη από την

ελάχιστη αποδεκτή. Η διαδικασία ολοκληρώνεται με τη δημιουργία κανόνων συσχέτισης από τα σύνολα στοιχείων με τις περισσότερες επανεμφανίσεις υπό την προϋπόθεση τήρησης της εμπιστοσύνης του συνδυασμού. (Κύρκος, 2015)

Ο πιο συνηθισμένος αλγόριθμος για την εξαγωγή κανόνων συσχέτισης είναι ο APRIORI ο οποίος ονομάστηκε έτσι λόγω της χρήσης προγενέστερης γνώσης (prior knowledge) των χαρακτηριστικών των συνόλων αντικειμένων με τις πιο συχνές επανεμφανίσεις.

Τα πιο συχνά μέτρα για την αξιολόγηση των κανόνων συσχέτισης είναι η υποστήριξη, η εμπιστοσύνη και το ποσοτικό μέτρο Lift. Με το μέτρο Lift (Chorianopoulos, 2016) γίνεται αποτίμηση της ικανότητας πρόβλεψη συγκρίνοντας πόσο καλός ή κακός είναι ο κανόνας που εξήχθη σε σχέση έναν άλλο τυχαίο κανόνα. Σύμφωνα με τον Κύρκο (2015), η ύπαρξη μικρής υποστήριξης στους κανόνες μπορεί αποτελεί ένα τυχαίο γεγονός και αντίθετα η ύπαρξη μεγάλης υποστήριξης συνδυαστικά με μεγάλη εμπιστοσύνη μπορεί να μην ανταποκρίνεται σε πραγματική σχέση.

4.3 ΕΠΙΔΡΑΣΗ ΘΟΡΥΒΟΥ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ

Κατά την εκπαίδευση των αλγορίθμων συχνά παρουσιάζονται σφάλματα. Τα σφάλματα μπορεί να είναι είτε τυχαία, είτε συστηματικά. Το τυχαίο σφάλμα που προκύπτει (Tan et al., 2014) ονομάζεται θόρυβος και αφορά, είτε λανθασμένες τιμές, είτε ακραίες τιμές στα δεδομένα. Πρόβλημα λανθασμένης τιμής μπορεί να οφείλεται στην λανθασμένη καταχώρηση δεδομένων. Αντιθέτως, τα σφάλματα ακραίων τιμών περιγράφουν μεμονωμένες περιπτώσεις που δε συμβαδίζουν με τα συνήθη δεδομένα και δε προσφέρουν χρήσιμη πληροφορία για την εκπαίδευση των αλγορίθμων.

Κατά το στάδιο της προεπεξεργασίας των δεδομένων πρέπει να γίνει προσπάθεια αντιμετώπισης του θορύβου (Κύρκος, 2015). Μια μέθοδος αντιμετώπισης του θορύβου είναι ο κατακερματισμός σε διαστήματα και αντικατάσταση τιμών. Στη μέθοδο αυτή τα δεδομένα χωρίζονται σε ίσα διαστήματα συχνότητας ή πλήθους, υπολογίζεται ο μέσος όρος του διαστήματος και στη συνέχεια γίνεται αντικατάσταση των τιμών του διαστήματος είτε με τον μέσο όρο κάθε διαστήματος, είτε με τις τιμές των οριακών τιμών του κάθε διαστήματος. Προκειμένου να αντιμετωπισθεί η εμφάνιση ακραίων

τιμών στα δεδομένα χρησιμοποιείται ο στατιστικός εντοπισμός υποθέσεων. Ακόμα, μια μέθοδος για τον εντοπισμό αντικειμένων ανόμοιων με τα υπόλοιπα είναι η ανάλυση συστάδων. Έτσι, τα όμοια αντικείμενα ομαδοποιούνται, ενώ τα αντικείμενα που δεν ταιριάζουν σε καμία ομάδα αποτελούν εξαιρέσεις. Τέλος, με την προσαρμογή των δεδομένων με χρήση μοντέλου είναι δυνατόν να γίνει πρόβλεψη των τιμών σε ένα συγκεκριμένο πεδίο με χρήση τιμών σε κάποιο άλλο πεδίο.

4.3.1 ΜΕΡΟΛΗΨΙΑ

Κατά την εκπαίδευση των αλγορίθμων σε διαφορετικά δεδομένα προκύπτει ένα είδος σφάλματος το οποίο καλείται μεροληψία (bias). Στην στατιστική, η μεροληψία προκύπτει από τη διαφορά της μέσης τιμής των επαναλαμβανόμενων μετρήσεων με την πραγματική τιμή. Ορισμένα είδη μεροληψίας είναι (Suresh & Guttag, 2019):

- Ιστορική μεροληψία (historical bias): Εμφανίζεται ακόμα και στην περίπτωση τέλειας δειγματοληψίας και επιλογής χαρακτηριστικών.
- Μεροληψία αναπαράστασης (representation bias): Αυτό το είδος σφάλματος παρουσιάζεται κατά την διαδικασία ορισμού των μεταβλητών και δειγματοληψίας από τον πληθυσμό.
- Μεροληψία μέτρησης (representation bias): Συμβαίνει από τον τρόπο μέτρηση ενός συγκεκριμένου χαρακτηριστικού.
- Μεροληψία συγκέντρωσης (aggregation bias): Εμφανίζεται όταν οι λάθος υποθέσεις για ένα πληθυσμό επηρεάζουν το αποτέλεσμα της εκπαίδευσης του αλγορίθμου.
- Μεροληψία αξιολόγησης (evaluation bias): Συμβαίνει κατά τη διαδικασία αξιολόγησης.
- Μεροληψία παράταξης (deployment bias): Εμφανίζεται όταν υπάρχει ασυμφωνία ανάμεσα στο πρόβλημα που ήταν προγραμματισμένο να επιλύει και στο πρόβλημα που τελικά λύνει.

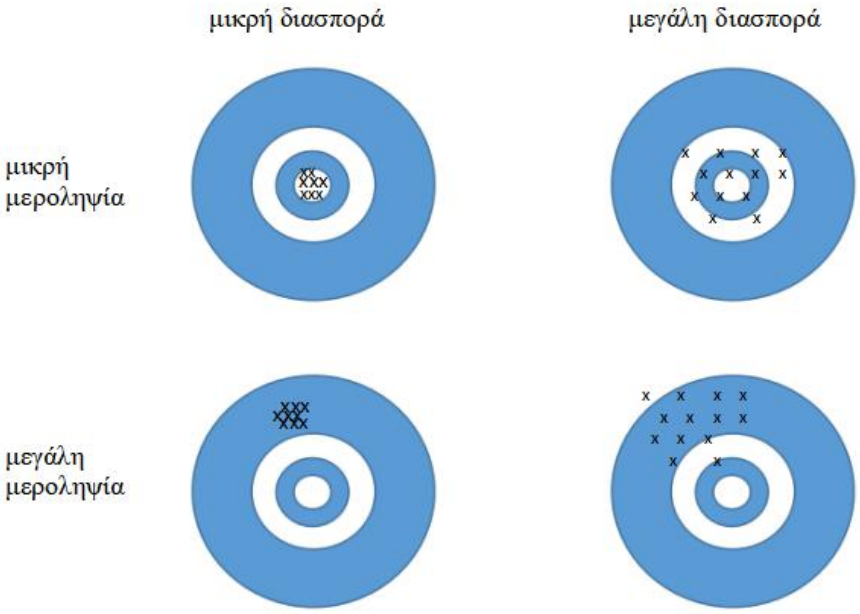
4.3.2 ΔΙΑΣΠΟΡΑ

Το σφάλμα λόγω διασποράς δείχνει πόσο απέχουν οι προβλέψεις μεταξύ τους για ένα συγκεκριμένο πραγματικό σημείο. Για την εκπαίδευση του μοντέλου

χρησιμοποιείται ένα συγκεκριμένο πλήθος δεδομένων. Η ευαισθησία των προβλέψεων αυτών, δηλαδή το πόσο η πρόβλεψη διαφέρει ανάμεσα σε διαφορετικά σύνολα δεδομένων εκπαίδευσης, ονομάζεται διασπορά (Hand et al., 2001).

4.3.3 BIAS-VARIANCE TRADE OFF

Η μεροληψία μετράει το πόσο αποκλίνουν οι προβλέψεις από την πραγματική τιμή, ενώ η διασπορά δείχνει πόσο πολύ διαφέρουν οι προβλέψεις μεταξύ τους. Το κέντρο του στόχου αναπαριστά το μοντέλο όπου περιγράφει τέλεια τις σωστές τιμές, ενώ η απομάκρυνση από το κέντρο του στόχου οδηγεί σε χειρότερες προβλέψεις. Έτσι, στην περίπτωση όπου τα δεδομένα έχουν μικρή διασπορά και μικρή μεροληψία (Σχήμα 4.1, πάνω αριστερά) τα δεδομένα βρίσκονται κοντά στην πραγματική τιμή και κοντά μεταξύ τους, ενώ μικρή μεροληψία συνδυαστικά με μεγάλη διασπορά (Σχήμα 4.1, πάνω αριστερά) οδηγεί σε δεδομένα με κλίση το κέντρο της πραγματικής τιμής αλλά με μεγάλη απόκλιση μεταξύ τους. Αντίθετα, δεδομένα με μεγάλη μεροληψία και μικρή διασπορά (Σχήμα 4.1, κάτω αριστερά) οδηγούν σε αποκεντροποίηση από την πραγματική τιμή αλλά με μικρές αποκλίσεις μεταξύ τους. Τέλος, στην περίπτωση μεγάλης μεροληψίας και μεγάλης διασποράς (Σχήμα 4.1, κάτω δεξιά) τα δεδομένα είναι απομακρυσμένα από την πραγματική τιμή αλλά και μεταξύ τους.



Σχήμα 4.1: Μεροληψία-διασπορά (Gupta, 2017)

Η πολυπλοκότητα εξαρτάται άμεσα με τη διασπορά και τη μεροληψία του μοντέλου που θα χρησιμοποιηθεί. Αύξηση της πολυπλοκότητας (Hastie et al., 2017) της διαδικασίας οδηγεί σε αύξηση της διασποράς και μείωση του τετραγώνου της μεροληψίας και αντιστρόφως. Για το λόγο αυτό πρέπει να γίνεται προσεκτικά η επιλογή της πολυπλοκότητας έτσι ώστε να ελαχιστοποιείται το σφάλμα της δοκιμής. Δηλαδή, θα πρέπει να είναι αρκετά σύνθετο προκειμένου να μπορεί να εκφράσει την δομή των δεδομένων αλλά ταυτόχρονα αρκετά απλό έτσι ώστε να είναι σε θέση να αποφύγει την προσαρμογή σε λανθασμένα μοτίβα.

4.4 ΥΠΕΡΑΠΛΟΥΣΤΕΥΣΗ ΚΑΙ ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ

Στους αλγορίθμους μηχανικής μάθησης βασική αιτία χαμηλής απόδοσης είναι η υπεραπλούστευση (underfitting) και η υπερπροσαρμογή (overfitting).

Υπεραπλούστευση αναφέρεται στην περίπτωση όπου το μοντέλο δεν έχει μάθει αρκετά από το σετ δεδομένα εκπαίδευσης με αποτέλεσμα να προκύπτουν αναξιόπιστες προβλέψεις σε νέα δεδομένα. Στην περίπτωση αυτή, το μοντέλο έχει υψηλή μεροληψία με αποτέλεσμα να υπάρχει χαμηλή ακρίβεια στο σετ εκπαίδευσης και συνεπώς δε περιέχεται η λύση σε αυτό. Έτσι, ένα υπεραπλουστευμένο μοντέλο δεν είναι κατάλληλο για την πραγματοποίηση προβλέψεων, καθώς, αδυνατεί να κάνει γενίκευση σε νέα δεδομένα. Αντιθέτως, ένα μοντέλο που χαρακτηρίζεται από υπερπροσαρμογή έχει υψηλή διασπορά με αποτέλεσμα να έχει μάθει πάρα πολλά από το σετ δεδομένων εκπαίδευσης, μεταξύ άλλων και τον θόρυβο, επηρεάζοντας έτσι αρνητικά την εφαρμογή του μοντέλου σε νέα δεδομένα. (Alpaydin, 2020)

Οι δύο παραπάνω οδηγούν σε αναξιόπιστες προβλέψεις και είναι απαραίτητο να βελτιωθούν. Ορισμένες μέθοδοι για την αντιμετώπιση των μοντέλων υπερπροσαρμογής είναι η πρόωρη διακοπή της εκπαίδευσης του μοντέλου (early stopping), η έγχυση θορύβου (noise injection), η διάσπαση βάρους (weight decay) και οι προσεγγιστικοί αλγόριθμοι βελτιστοποίησης (optimized approximation algorithm). Μια ακόμα μέθοδος για την αντιμετώπιση της υπερπροσαρμογής (Gupta, 2017) είναι η κανονικοποίηση (regularization), η οποία είναι ένα είδος παλινδρόμησης που αποθαρρύνει την

εκπαίδευση ενός σύνθετου ή ευέλικτου μοντέλου. Με τη μέθοδο επιτυγχάνεται σημαντική μείωση της διασποράς του μοντέλου χωρίς όμως να αυξηθεί σημαντικά η μεροληψία.

4.5 ΕΦΑΡΜΟΓΕΣ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Το φάσμα των εφαρμογών της μηχανικής μάθησης είναι πολυδιάστατο και βρίσκει εφαρμογή σε πολλούς κλάδους της σύγχρονης ζωής. Έχει εφαρμογή στην αυτοκίνηση, στην ρομποτική, στην ιατρική διαγνωστική, στις τραπεζικές εφαρμογές και σε πολλά μέρη της παραγωγής.

Στις βιομηχανικές μονάδες, προκειμένου να αυξηθεί η παραγωγικότητα του γίνεται ολοένα και περισσότερο αναγκαία η χρήση των βιομηχανικών ρομπότ. Το γεγονός αυτό έχει δημιουργήσει την ανάγκη για βελτίωση και στον αυτοματισμό των άκαμπτων δυνατοτήτων τους. Χαρακτηριστικό παράδειγμα αποτελεί η Siemens, η οποία χρησιμοποιεί νευρωνικά δίκτυα για την παρακολούθηση των βιομηχανικών δραστηριοτήτων της. Με το συνδυασμό της χρήσης αισθητήρων και της ανάλυσης των μετρήσεων τους (μέσω MindShere) καταφέρνει να εντοπίζει αποτελεσματικά τις δυσλειτουργίες στις εγκαταστάσεις της και να εφαρμόζει συντήρηση. Ακόμη, χρησιμοποιεί εργαλεία αυτοματοποίησης ηλεκτρονικού σχεδιασμού[2] (Solido Design Automation) τα οποία μπορούν να μειώσουν σημαντικά το χρόνο σχεδιασμού κάποιου προϊόντος.

Μια ακόμη σημαντική εφαρμογή βιομηχανικού αυτοματισμού έχει αναπτυχθεί από την εταιρεία GE. Η εταιρεία αυτή έχει καταφέρει μέσω της συγκέντρωσης και ανάλυσης δεδομένων από αισθητήρες, να παρακολουθεί και να ελέγχει την παραγωγή για πιθανά προβλήματα (Briliant Manufacturing Suite)[3].

Σημαντική ρόλο στην βιομηχανία της υγείας διαδραματίζει η μηχανική μάθηση χάρη στην ικανότητα των εφαρμογών τους να διαχειρήζονται μεγάλους όγκους δεδομένων και να εξάγουν πολύ πιο γρήγορα και ακριβή αποτελέσματα σε σχέση με τις συμβατικές μεθόδους. Οι εφαρμογές των αλγορίθμων μηχανικής μάθησης για τη διάγνωση ασθενειών βασίζονται στις συσχετίσεις τους με διάφορα συμπτώματα των

[2] prn.to/3jaQreS
[3] https://bwnews.pr/348BVON

ασθενών αλλά και με διάφορες άλλες ασθένειες που πιθανόν εμφανίζονται από κοινού. Στην Ιατρική, είναι συχνή η χρήση Κλινικών συστημάτων Υποστήριξης Αποφάσεων τα οποία μεταξύ άλλων περιλαμβάνουν δραστηριότητες συνταγογράφισης φαρμάκων για την αποφυγή λαθών και κλινική παρακολούθηση των ασθενών. Παραδείγματος χάριν, με ένα τέτοιο σύστημα μπορεί να επιτευχθεί η ανίχνευση μοτίβων που να υποδεικνύουν την πιθανή υποτροποίαση κάποιου ασθενή.

Ακόμη, είναι δυνατόν μέσω αναγνώρισης εικόνων να ερμηνευτούν ακτινογραφίες, αξονικές και μαγνητικές τομογραφίες και με αυτό το τρόπο να συγκριθούν με τις αντίστοιχες εικόνες υγιών ατόμων. Στην περίπτωση σημαντικών διαφορών μεταξύ των εικόνων γίνεται αναφορά για περαιτέρω διερεύνηση από τους αρμόδιους ιατρούς.

Η χρήση αλγορίθμων μηχανικής μάθησης έχει εφαρμογές και στην ανακάλυψη νέων φαρμάκων. Οι αλγόριθμοι αυτοί λαμβάνουν σαν είσοδο τα φάρμακα με δράσεις παρόμοιες με την επιθυμητή και εξάγουν τις χημικές ενώσεις που είναι υπεύθυνες για αυτές τις δράσεις.

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, η μηχανική μάθηση έχει καταλυτικό ρόλο στην πρόβλεψη της συμπεριφοράς των πελατών. Οι επιχειρήσεις είναι σε θέση να εντοπίζουν ευκολότερα τους πελάτες που είναι πιο πιθανό να προελκυσθούν από μια δράση της επιχείρησης. Ακόμη, στο ηλεκτρονικό εμπόριο η εξυπηρέτηση των πελατών μπορεί να πραγματοποιείται μέσω των Chatbots[4] η χρήση των οποίων επιτρέπει στους πελάτες να επικοινωνούν με την επιχείρηση μέσω ηλεκτρονικών μηνυμάτων οποιαδήποτε στιγμή επιθυμούν. Τα ερωτήματα των πελατών απαντώνται από τα bot τα οποία προσομοιώνουν τον τρόπο με τον οποίο θα απαντούσε και ένας πραγματικός άνθρωπος. Από τις πλέον διαδεδομένες εφαρμογές chatboots χρησιμοποιούν εταιρείες όπως η Apple με την εφαρμογή Siri και η Amazon με την εφαρμογή Alexa.

Στον τραπεζικό τομέα, σημαντικό πρόβλημα, που επιλύεται με χρήση αλγορίθμων μηχανικής μάθησης, αποτελεί η ανίχνευση και η πρόληψη της απάτης. Οι τράπεζες καταγράφουν μεγάλο όγκο δεδομένων των πελατών τους σε καθημερινή βάση και συνεπώς θα ήταν άκαρπο αν δεν υπήρχε η εκμετάλλευσή τους. Τα δεδομένα των συναλλαγών των πελατών τους υποδεικνύουν μοτίβα για την καταναλωτική τους συμπεριφορά. Ακόμη, από την ανάλυση των δεδομένων μιας τράπεζας μπορούν να

[4] https://bit.ly/3mXqYIf

εξαχθούν συμπεράσματα σχετικά με το προφίλ των αφερέγγυων πελατών. Με τον τρόπο είναι δυνατόν να εκτιμηθεί η πιστοληπτική ικανότητα των πελατών και να αποφευχθεί κάποια συνεργασία της τράπεζας με τους εκάστοτε πελάτες.

Τέλος, όλες οι ενέργειες των επιχειρήσεων περιστρέφονται στους πελάτες, από την προσέλκυση στην απόκτηση και στην διατήρησή τους. Οι επιχειρήσεις προκειμένου να ανταπεξέλθουν στο ανταγωνιστικό περιβάλλον αναζητούν συνεχώς λύσεις για την αύξηση της παραγωγικότητας και της αποτελεσματικότητας τους, την εύρεση των πιο επικερδών πελατών και στην προσέλκυση τους. Βασικό κομμάτι είναι η διατήρηση των πελατών τους που επιτυγχάνεται από την ικανοποίηση των αναγκών τους οι οποίες μπορεί να αφορούν κάποιο χαρακτηριστικό ενός προϊόντος ή κάποια υπηρεσία. Κύριος στόχος κάθε επιχείρησης είναι να ανακαλύψει πρώτη τις ανάγκες των πελατών και να τις εκτελέσει με τον πιο αποτελεσματικό τρόπο. Στο σύγχρονο αυτό περιβάλλον, η εκπλήρωση αυτή των αναγκών είναι αλληλένδετη με την εφαρμογή των αλγορίθμων μηχανικής μάθησης.

ΚΕΦΑΛΑΙΟ 5. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Στο παρόν κεφάλαιο αφορά το πειραματικό μέρος της εργασίας. Συγκεκριμένα, στο πρώτο μέρος του κεφαλαίου γίνεται μια παρουσίαση των δεδομένων καθώς και η περιγραφή τους. Στη συνέχεια γίνεται η στατιστική ανάλυση των δεδομένων. Κατά το τρίτο μέρος πραγματοποιείται η ανάλυση των μοντέλων μηχανικής μάθησης που επιλέχθηκαν και τέλος, κατά το τέταρτο μέρος, γίνεται η παρουσίαση και η ανάλυση των αποτελεσμάτων όπως αυτά προέκυψαν από την εφαρμογή των αλγορίθμων.

5.1 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα του προβλήματος συλλέχθηκαν από την ιστοσελίδα kaggle.com και αφορούν δεδομένα πελατών της τράπεζας Santander, θυγατρική της Banco Santander. Πρόκειται για μια κορυφαία τράπεζα λιανεμπορίου καταθέσεων της Ισπανίας η οποία προσφέρει εξατομικευμένες προτάσεις δανειοδότησης στους πελάτες της στοχεύοντας στην υποστήριξη των οικονομικών τους αποφάσεων. Στόχος της παρούσας ανάλυσης είναι, μέσω της καταναλωτικής συμπεριφοράς των πελατών της τράπεζας, η πρόβλεψη των υπηρεσιών που είναι πιο πιθανό να χρησιμοποιηθούν το επόμενο διάστημα.

Το αρχικό δείγμα αποτελούνταν από δεδομένα καταναλωτικής συμπεριφοράς τα οποία συλλέχθηκαν σε διάρκεια 1,5 χρόνου. Συγκεκριμένα, τα δεδομένα αποτελούσαν περίπου 1,3 εκατομμύρια καταχωρήσεις συναλλαγών με 48 διαφορετικά στοιχεία-μεταβλητές για κάθε μια από αυτές. Για την απλοποίηση της ανάλυσης, το δείγμα μειώθηκε σε 133.494 τυχαία επιλεγμένες καταχωρήσεις και χρησιμοποιήθηκαν 13 μεταβλητές. Στον πίνακα 5.1 παρουσιάζεται το όνομα κάθε μεταβλητής καθώς και μια συνοπτική επεξήγηση της κάθε μεταβλητής.

Από αυτές τις μεταβλητές οι ανεξάρτητες συμβολίζονται με Χ και οι εξαρτημένες με Υ (πίνακας 5.1). Οι εξαρτημένες μεταβλητές λαμβάνουν τιμή 1 στην περίπτωση

ύπαρξης της υπηρεσίας που υποδεικνύουν, ενώ μηδενική τιμή σε διαφορετική περίπτωση.

Πίνακας 5.1: Πίνακας μεταβλητών που χρησιμοποιήθηκαν στην ανάλυση

Μεταβλητή	Επεξήγηση
X1	Φύλο πελάτη
X2	Ηλικία πελάτη
X3	Δείκτης νέου πελάτη (1 στην περίπτωση που ο πελάτης έχει καταχωρηθεί τους τελευταίους 6 μήνες)
X4	Μήνες που ένας πελάτης συνεργάζεται με την τράπεζα
X5	Δείκτης καταγωγής (S όταν η χώρα γέννησης του πελάτη ταυτίζεται με την χώρα στην οποία βρίσκεται η τράπεζα και N σε διαφορετική περίπτωση)
X6	Δείκτης κινητικότητας πελάτη (1 για ενεργό πελάτη και 0 για ανενεργό)
X7	Εισόδημα πελάτη
Y1	Ύπαρξη λογαριασμού μισθοδοσίας
Y2	Ύπαρξη ηλεκτρονικού λογαριασμού
Y3	Ύπαρξη επενδύσεων σε αμοιβαία κεφάλαια
Y4	Ύπαρξη συνταξιοδοτικού προγράμματος
Y5	Ύπαρξη πιστωτικής κάρτας

Οι ανεξάρτητες μεταβλητές που επιλέχθηκαν είναι το φύλο, η ηλικία, ο δείκτης νέου πελάτη, η διάρκεια στην οποία η τράπεζα προσφέρει υπηρεσίες σε ένα πελάτη, ο δείκτης καταγωγής του πελάτη, ο δείκτης κινητικότητας του πελάτη και το εισόδημα του πελάτη. Οι εξαρτημένες μεταβλητές είναι ο λογαριασμός μισθοδοσίας, ο ηλεκτρονικός λογαριασμός, οι επενδύσεις σε αμοιβαία κεφάλαια, τα συνταξιοδοτικά προγράμματα και η ύπαρξη πιστωτικής κάρτας.

Από το σύνολο των καταχωρήσεων με ύπαρξη τουλάχιστον μιας υπηρεσίας, περίπου το 70% αφορά κατοχή λογαριασμού μισθοδοσίας και κατοχή ηλεκτρονικού λογαριασμού, ενώ η ύπαρξη πιστωτικής κάρτας αφορά περίπου το 19% των πελατών.

Χαμηλότερα ποσοστά της τάξεως του 11% καταλαμβάνουν πελάτες με χρηματοδοτικό λογαριασμό και συνταξιοδοτικό πρόγραμμα.

5.2 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Από την στατιστική ανάλυση των δεδομένων προέκυψαν τα ποσοστά των πελατών της τράπεζας για κάθε κατηγορία ηλικίας (Πίνακας 5.2). Παρατηρείται ότι οι νέοι πελάτες αφορούν κατά μεγαλύτερο ποσοστό γυναίκες, ενώ για τους πελάτες μέσης ηλικίας οι άνδρες είναι περισσότεροι. Επιπλέον οι ηλικιωμένοι καταλαμβάνουν περίπου ίδια ποσοστά και για τα δύο φύλα. Περίπου οι μισοί πελάτες είναι μεσήλικες ανεξαρτήτως φύλου, αμέσως λιγότεροι είναι οι νέοι ενώ λιγότεροι είναι οι ηλικιωμένοι πελάτες.

Πίνακας 5.2 : Συχνότητες των φύλων των πελατών ανά κατηγορία ηλικίας

Φύλο\Ηλικία	Νέοι	Μεσήλικες	Ηλικιωμένοι
Άνδρες	29,75	56,49	13,75
Γυναίκες	45,34	42,69	11,96
Συνολικά	36,87	50,19	12,94

Προκειμένου να διαπιστωθεί η ύπαρξη κάποιας σχέσης ανάμεσα σε κάποια ανεξάρτητη μεταβλητή και στην κατοχή ενός τραπεζικού προϊόντος έγινε υπολογισμός των σχετικών συχνοτήτων τους. Παρακάτω παρουσιάζεται η ανάλυση της σχέσης του φύλου με τις τραπεζικές υπηρεσίες (Πίνακας 5.3). Παρατηρείται αυξημένη ύπαρξη του λογαριασμού μισθοδοσίας (8,89%) και του ηλεκτρονικού λογαριασμού (8,96%) και για τα δύο φύλα συγκρινόμενο με τα υπόλοιπα τραπεζικά προϊόντα. Ακόμη, σημαντικό ποσοστό των πελατών κατέχουν πιστωτική κάρτα (4,81%) εκ των οποίων οι περισσότεροι είναι άνδρες. Από τους άνδρες που έχουν κάποιο τραπεζικό προϊόν το 10,15% έχουν ηλεκτρονικό λογαριασμό, 9,62% έχουν λογαριασμό μισθοδοσίας, ενώ χαμηλότερα ποσοστά λαμβάνουν στις υπόλοιπες κατηγορίες. Όμοια οι γυναίκες με λογαριασμό μισθοδοσίας και ηλεκτρονικό λογαριασμό λαμβάνουν το 8,02% και 7,55% αντίστοιχα ενώ η κατοχή κάποιας άλλης υπηρεσίας είναι σχετικά μικρή.

Πίνακας 5.3: Σχετικές συχνότητες (ποσοστά %) τραπεζικών προϊόντων ανά φύλο

	Άνδρες	Γυναίκες	Σύνολο
Λογαριασμός μισθοδοσίας	9,62	8,02	8,89
Ηλεκτρονικός λογαριασμός	10,15	7,55	8,96
Αμοιβαία κεφάλαια	2,52	1,38	2,00
Συνταξιοδοτικό πρόγραμμα	1,06	0,84	0,96
Πιστωτική κάρτα	5,74	3,71	4,81

Όμοια ανάλυση με την παραπάνω έγινε και για την σχέση της ύπαρξης κάποιου τραπεζικού προϊόντος με το είδος του πελάτη, δηλαδή αν ένας πελάτης είναι νέος ή παλιός (Πίνακας 5.4). Από το σύνολο των νέων πελατών μόλις το 2,98% έχουν λογαριασμό μισθοδοσίας, 1,88% έχουν ηλεκτρονικό λογαριασμό ενώ η ύπαρξη κάποιου άλλου προϊόντος είναι αρκετά μικρή. Αντιθέτως, από τους παλιούς πελάτες, το 9,05% κατέχει λογαριασμό μισθοδοσίας, το 9,16% ηλεκτρονικό λογαριασμό, το 4,94% πιστωτική κάρτα, ενώ χαμηλότερα ποσοστά πελατών καταλαμβάνει η κατοχή κάποιου άλλου τραπεζικού προϊόντος.

Πίνακας 5.4: Σχετικές συχνότητες (ποσοστά %) τραπεζικών προϊόντων ανά είδος πελάτη

	Νέος πελάτης	Παλιός πελάτης	Σύνολο
Λογαριασμός μισθοδοσίας	2,98	9,05	8,89
Ηλεκτρονικός λογαριασμός	1,88	9,16	8,96
Αμοιβαία κεφάλαια	0,34	2,05	2,00
Συνταξιοδοτικό πρόγραμμα	0,03	0,98	0,96
Πιστωτική κάρτα	0,22	4,94	4,81

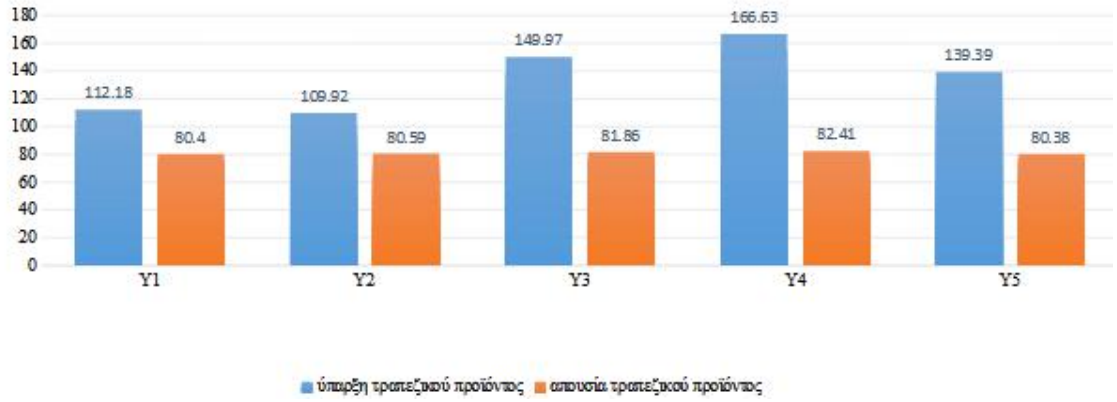
Έπειτα ακολουθεί η ανάλυση της σχέσης της ύπαρξης κάποιας τραπεζικής υπηρεσίας με την καταγωγή των πελατών (Πίνακας 5.5). Από το σύνολο των Ισπανών πελατών της τράπεζας, εκείνοι με λογαριασμό μισθοδοσίας λαμβάνουν το 8,20% των

πελατών. Λίγο χαμηλότερα ποσοστά λαμβάνουν οι Ισπανοί πελάτες με ηλεκτρονικό λογαριασμό (5,18%) και πιστωτική κάρτα (3,80%), ενώ πολύ μικρά ποσοστά αφορούν πελάτες με κατοχή κάποιας άλλης τραπεζικής υπηρεσίας. Παρατηρείται ότι από τους πελάτες με άλλη καταγωγή οι περισσότεροι έχουν ηλεκτρονικό λογαριασμό (9,13%), λογαριασμό μισθοδοσίας (8,92%) και πιστωτική κάρτα (4,85), ενώ πολύ λίγοι έχουν κάποιο άλλο τραπεζικό προϊόν. Συγκρίνοντας τους πελάτες με Ισπανική με εκείνους με άλλη καταγωγή προκύπτει ότι οι περισσότεροι που έχουν ηλεκτρονικό λογαριασμό είναι εκείνη με άλλη καταγωγή.

Πίνακας 5.5: Σχετικές συχνότητες (ποσοστά %) τραπεζικών προϊόντων ανά καταγωγή πελάτη

	Ισπανική καταγωγή	Άλλη καταγωγή	Σύνολο
Λογαριασμός μισθοδοσίας	8,20	8,92	8,89
Ηλεκτρονικός λογαριασμός	5,18	9,13	8,96
Αμοιβαία κεφάλαια	0,57	2,06	2,00
Συνταξιοδοτικό πρόγραμμα	0,97	0,96	0,96
Πιστωτική κάρτα	3,80	4,85	4,81

Ακολουθεί σχηματική απεικόνιση των μηνών που συνεργάζονται οι πελάτες με την τράπεζα (Σχήμα 5.1). Παρατηρείται ότι από τους πελάτες που κατέχουν κάποιο τραπεζικό προϊόν, μεγαλύτερη διάρκεια συνεργασίας έχουν εκείνοι με συνταξιοδοτικό πρόγραμμα (Y4, 166,63 μήνες). Δεύτεροι σε διάρκεια έρχονται οι πελάτες με επενδύσεις σε αμοιβαία κεφάλαια (Y3, 149,97 μήνες) και τρίτοι εκείνοι με κατοχή πιστωτικής κάρτας (Y5, 139,39 μήνες). Ακολουθούν οι πελάτες με κατοχή λογαριασμού μισθοδοσίας (Y1, 112,18 μήνες) και οι πελάτες με κατοχή ηλεκτρονικού λογαριασμού (Y2, 109,92 μήνες). Τέλος, αξίζει να σημειωθεί ότι οι πελάτες χωρίς κάποιο τραπεζικό προϊόν, από τα εξεταζόμενα, έχουν μέση διάρκεια συνεργασίας με την τράπεζα γύρω στους 80 μήνες.



Σχήμα 5.1: Μέση διάρκεια συνεργασίας των πελατών με την τράπεζα

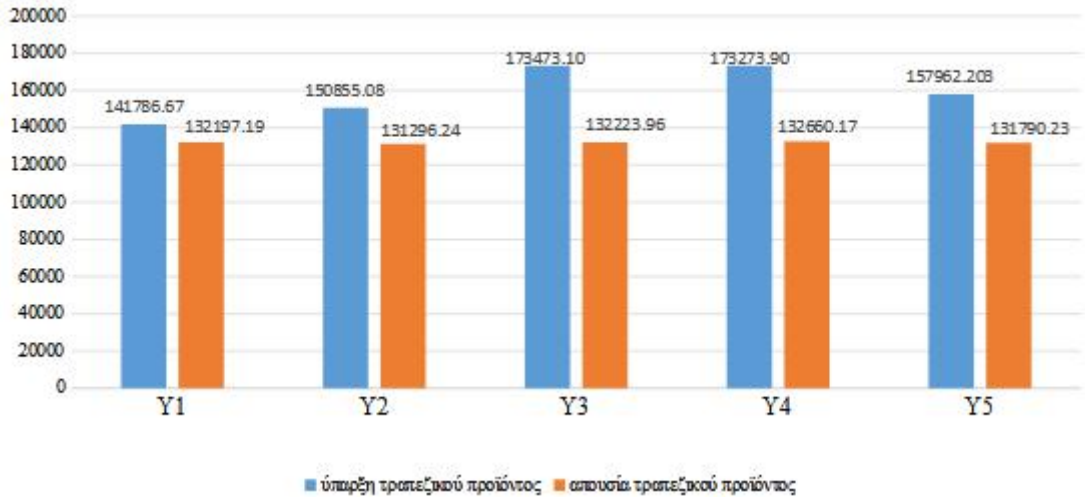
Αναλύοντας τη σχέση της κινητικότητας των πελατών με τα τραπεζικά προϊόντα (Πίνακας 5.6) προκύπτει ότι από το σύνολο των ενεργών πελατών, το 18,69 % κατέχει λογαριασμό μισθοδοσίας και το 17,68% ηλεκτρονικό λογαριασμό. Ακόμη, οι πελάτες που διαθέτουν πιστωτική κάρτα συνιστούν το 10,15 % των ενεργών πελατών, ενώ η κατοχή κάποιου άλλου τραπεζικού προϊόντος είναι συγκριτικά μικρή για αυτή την κατηγορία. Αξιοσημείωτο είναι το γεγονός ότι πολύ μικρά ποσοστά των ανενεργών πελατών διαθέτουν κάποιο τραπεζικό προϊόν.

Πίνακας 5.6: Σχετικές συχνότητες (ποσοστά %) τραπεζικών προϊόντων ανά κινητικότητα πελάτη

	Ενεργός πελάτης	Ανενεργός πελάτης	Σύνολο
Λογαριασμός μισθοδοσίας	18,69	0,18	8,89
Ηλεκτρονικός λογαριασμός	17,68	1,22	8,96
Αμοιβαία κεφάλαια	4,21	0,04	2,00
Συνταξιοδοτικό πρόγραμμα	2,04	0,00	0,96
Πιστωτική κάρτα	10,15	0,07	4,81

Παρακάτω απεικονίζεται το μέσο εισόδημα των πελατών για κάθε τραπεζικό προϊόν (Σχήμα 5.2). Παρατηρείται αυξημένο μέσο εισόδημα των πελατών που έχουν κάποιο τραπεζικό προϊόν σε σχέση με τους πελάτες χωρίς κάποιο από τα εξεταζόμενα τραπεζικά προϊόντα. Υψηλότερο μέσο εισόδημα κατέχουν οι πελάτες με

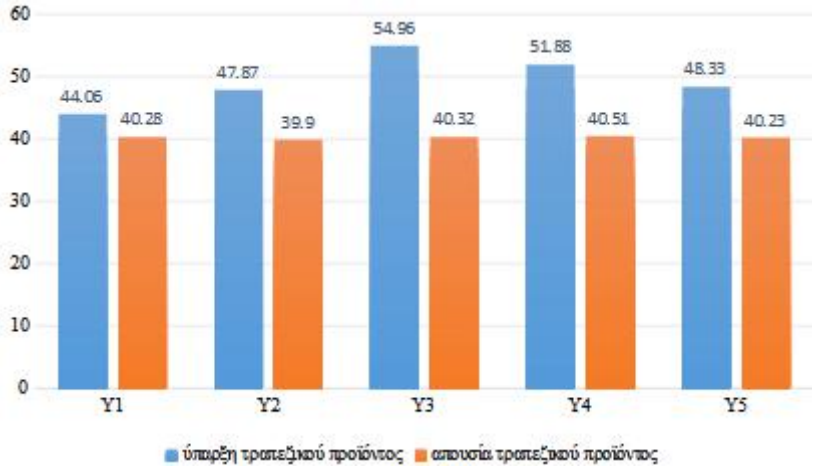
συνταξιοδοτικά προγράμματα (Υ4, 173273,90) ή με επενδύσεις σε αμοιβαία κεφάλαια (Υ3, 173473,10), ενώ λίγο χαμηλότερο μέσο εισόδημα παρουσιάζουν οι πελάτες με κάποιο άλλο τραπεζικό προϊόν.



Σχήμα 5.2: Μέσο εισόδημα πελατών ανά τραπεζικό προϊόν

Από την ανάλυση της ηλικίας των πελατών σε σχέση με την κατοχή ενός προϊόντος (Σχήμα 5.3) προκύπτει ότι οι πελάτες με κατοχή οποιουδήποτε από τα τραπεζικά προϊόντα που αναλύονται έχουν μεγαλύτερη μέση τιμή σε σχέση με εκείνους που δεν έχουν κάποιο τραπεζικό προϊόν. Επίσης, μέγιστη μέση ηλικία έχουν οι πελάτες με ηλεκτρονικό λογαριασμό, ενώ ελάχιστη εκείνοι με λογαριασμό μισθοδοσίας και κυμαίνονται γύρω στα 50 έτη.

Στη συνέχεια αναλύθηκε η συνύπαρξη των τραπεζικών προϊόντων των πελατών (Πίνακας 5.7). Παρατηρείται ότι από τους πελάτες με λογαριασμό μισθοδοσίας το 38,59% έχει και ηλεκτρονικό λογαριασμό, ενώ το 32,78% έχει πιστωτική κάρτα. Ακόμη, το 22,69% των πελατών με ηλεκτρονικό λογαριασμό έχουν και πιστωτική κάρτα, ενώ το 18,68% με επενδύσεις σε αμοιβαία κεφάλαια έχουν και πιστωτική κάρτα. Τέλος, το 26,66% των πελατών με συνταξιοδοτικά προγράμματα έχουν και πιστωτική κάρτα.



Σχήμα 5.3: Μέση ηλικία πελατών ανά τραπεζικό προϊόν

Πίνακας 5.7: Συνύπαρξη τραπεζικών προϊόντων

	Ηλ. λογαρ.	Αμοιβ. κεφ.	Συντ/κά προγρ.	Πιστωτική κάρτα
Λογ. μισθοδοσίας	38,59	5,41	4,19	32,78
Ηλ. λογαριασμός		8,40	3,99	22,69
Αμοιβ. κεφ.			7,79	18,68
Συντ/κά προγράμματα				26,66

5.3 ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Το πρόγραμμα R διαθέτει πληθώρα μεθόδων ταξινόμησης. Για την ανάλυση επιλέχθηκαν οι εξής μέθοδοι : Λογιστική Παλινδρόμηση (Logistic regression), Δέντρα λήψης αποφάσεων ενίσχυσης κλίσης (Extreme Gradient Boosting), τα δέντρα ταξινόμησης και παλινδρόμησης (αλγόριθμος CART), η ενισχυμένη λογιστική παλινδρόμηση (αλγόριθμος LogitBoost) και τα νευρωνικά δίκτυα (μέθοδος MLP). Η εφαρμογή των μοντέλων έγινε μέσω του πακέτου CARET[5].

Λογιστική Παλινδρόμηση (Logistic regression, Hosmer and Lemeshow, 2013): Η μέθοδος αυτή χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών προκειμένου να προβλέψει, μέσω ενός λογιστικού μοντέλου, το αποτέλεσμα μιας

[5]http://topepo.github.io/caret/index.html

εξαρτημένης κατηγορικής μεταβλητής. Η εξαρτημένη μεταβλητή είναι συνήθως δυαδικού χαρακτήρα, ενώ οι ανεξάρτητες μεταβλητές μπορεί να είναι ονομαστικές ή ποσοτικές. Οι παράμετροι του λογιστικού μοντέλου υπολογίζονται με τη χρήση μεθόδων μεγίστης πιθανοφάνειας (Maximum Likelihood Estimation, MLE).

Δέντρα λήψης αποφάσεων ενίσχυσης κλίσης (Extreme Gradient Boosting): Πρόκειται για μια μέθοδο που αναπτύχθηκε από τους Chen and Guestrin (2016) και αποτελεί μια από τις ταχύτερες υλοποιήσεις gradient boosted δέντρων. Εκτός από την ταχύτητά του αλγορίθμου, ένα άλλο χαρακτηριστικό που τον κάνει ιδιαίτερα δημοφιλή είναι η υψηλή απόδοση του.

Συγκεκριμένα, είναι κατάλληλη μέθοδος όταν το σύνολο των δεδομένων εκπαίδευσης αποτελείται κυρίως από αριθμητικά δεδομένα ή και από κατηγορικά, ενώ κρίνεται ακατάλληλη όταν οι μεταβλητές είναι μόνο κατηγορικές.

Χαρακτηριστικό του αλγορίθμου είναι ο συνδυασμός πολλαπλών ταξινομητών για την εξαγωγή της τελικής ταξινόμησης, γεγονός που τον κατατάσσει στην κατηγορία των συνδυασμένων μοντέλων (ensemble models). Στην τεχνική boosting, σκοπός είναι μέσω των δέντρων που δημιουργούνται μεταγενέστερα, να μειωθούν τυχόν σφάλματα από προηγούμενα δέντρα. Αρχικά, οι αδύναμοι ταξινομητές, οι οποίοι έχουν υψηλή μεροληψία, συνεισφέρουν μόνο μερικές απαραίτητες πληροφορίες για την πρόβλεψη επιτρέποντας στην τεχνική boosting να παράγει ισχυρό ταξινομητή. Ο τελικός ταξινομητής καταλήγει με μικρότερη μεροληψία από τους αρχικούς αδύναμους ταξινομητές.

Δέντρα ταξινόμησης και παλινδρόμησης (classification and regression trees, CART): Πρόκειται για έναν αλγόριθμο που δημιουργήθηκε από τους Breiman et al.(1984), ο οποίος επιτρέπει την πρόβλεψη, από ένα πλήθος ανεξάρτητων μεταβλητών, τόσο μιας εξαρτημένης ποσοτικής μεταβλητής όσο και μιας εξαρτημένης κατηγορικής μεταβλητής. Όταν η πρόβλεψη αφορά μια εξαρτημένη ποσοτική μεταβλητή, τότε πρόκειται για λύση με τη μορφή πολλαπλής παλινδρόμησης, ενώ όταν στόχος είναι η πρόβλεψη μιας κατηγορικής μεταβλητής η λύση είναι με μορφή ταξινόμησης.

Με την τεχνική αυτή γίνεται ανάπτυξη ενός δέντρου με τη μορφή διακλαδώσεων με κόμβους σε κάθε μια από αυτές. Σε κάθε ενδιάμεσο κόμβο γίνεται καθορισμός μιας συνθήκης τμήσης η οποία έχει σαν αποτέλεσμα την παραγωγή δυαδικών διαχωρισμών. Όταν η διχοτόμηση έχει σαν αποτέλεσμα την ικανοποιητική πρόβλεψη των

παρατηρήσεων της εξαρτημένης μεταβλητής, τότε η διαδικασία κρίνεται επιτυχημένη. Ο διαχωρισμός γίνεται συνήθως με χρήση του δείκτη Gini:

$$\text{Gini} = 1 - \sum_j p_j^2$$

όπου p_i η πιθανότητα της παρατήρησης να ανήκει στην κλάση j. Όσο χαμηλότερη τιμή Gini έχει μια κλάση, τόσο μεγαλύτερη η ομοιογένεια της. Για τον διαχωρισμό, εκτός του δείκτη Gini, χρησιμοποιείται ο υπολογισμός της εντροπίας (entropy):

$$\text{Entropy} = -\sum_j p_j \log_2 p_j$$

Η εντροπία λαμβάνει μηδενική τιμή στην περίπτωση που το δείγμα που μελετάται είναι τελείως ομοιογενές, ενώ λαμβάνει τιμή 1 σε διαφορετική περίπτωση.

Ενισχυμένη Λογιστική Παλινδρόμηση (Boosted Logistic Regression, Caret method LogitBoost) : Το μοντέλο αυτό δημιουργήθηκε από τους Friedman et al. (2000) και είναι αποτέλεσμα σύνθεσης του μοντέλου AdaBoost με τη συνάρτηση κόστους της λογιστικής παλινδρόμησης. Δηλαδή, πρόκειται για την εφαρμογή της τεχνικής boosting στη δημιουργία ενός λογιστικού μοντέλου.

Αρχικά, ο αλγόριθμος λαμβάνει επαναλαμβανόμενα διαφορετικά σετ εκπαίδευσης με αποτέλεσμα τη δημιουργία ενός αδύναμου κανόνα πρόβλεψης. Στη συνέχεια πραγματοποιείται η ενίσχυση και ο αλγόριθμος μετατρέπει τους αδύναμους κανόνες σε μια ισχυρή πρόβλεψη.

Νευρωνικά δίκτυα (Neural Networks, Ripley, 2014): Για την ανάλυση χρησιμοποιούνται νευρωνικά δίκτυα τύπου Multilayer Perception (MLP). Πρόκειται για δίκτυα απλής προώθησης των οποίων οι νευρώνες οργανώνονται σε στρώματα (layers). Δηλαδή στο ίδιο στρώμα δεν υπάρχουν συνδέσεις μεταξύ των νευρώνων καθώς και δεν υπάρχουν συνδέσεις μεταξύ των νευρώνων που δεν ανήκουν σε διαδοχικά επίπεδα. Επιπλέον, με τα μοντέλα MLP είναι δυνατόν να επιλυθούν προβλήματα ταξινόμησης με μη γραμμικούς κρυμένους νευρώνες. Η εκπαίδευση συνήθως πραγματοποιείται με τη μέθοδο της οπισθοδιάδοση σφάλματος (error backpropagation) και στοχεύει στην ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (mean squared error, MSE).

5.4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Ένα βασικό στάδιο μετά την εκτέλεση των αλγορίθμων μηχανικής μάθησης συνιστά η αξιολόγηση των μοντέλων που δημιουργήθηκαν. Ένας δείκτης που υπολογίζει την απόδοση των μοντέλων είναι ο δείκτης AUROC (Area Under the Receiver Operating Characteristic Curve), ο οποίος λαμβάνει τιμές από 0 έως 1 και υπολογίζει τη ικανότητα των μοντέλων να ταξινομούν.

Όσο μεγαλύτερη τιμή λαμβάνει ο δείκτης AUROC, τόσο καλύτερα το μοντέλο κάνει σωστές προβλέψεις. Έτσι, ένα μοντέλο με δείκτη AUROC ίσο με 0 κάνει προβλέψεις 100% λανθασμένες, ενώ ένα μοντέλο με δείκτη AUROC ίσο με 1 κάνει προβλέψεις 100% σωστές. Βασικό πλεονέκτημα του δείκτη είναι η ικανότητα του να μετράει πόσο καλά ταξινομούνται οι προβλέψεις και όχι τις ακριβείς τιμές αυτών.

Στον πίνακα 5.8 παρουσιάζονται οι τιμές του δείκτη AUROC των μοντέλων που χρησιμοποιήθηκαν σε κάθε επιλογή εξαρτημένης μεταβλητής.

Πίνακας 5.8: πίνακας μέτρου αξιολόγησης AUROC

Εξαρτ. μεταβλ.	LR	LogitBoost	CART	xgbTree	MLP
Y1	0,824	0,791	0,862	0,924	0,816
Y2	0,807	0,796	0,839	0,997	0,747
Y3	0,895	0,864	0,900	1,000	0,895
Y4	0,824	0,791	0,862	0,924	0,824
Y5	0,870	0,828	0,878	0,945	0,860

Στη συνέχεια, για κάθε αλγόριθμο υπολογίζεται η μέση τιμή του δείκτη AUROC (πίνακας 5.9) από όλες τις εξαρτημένες μεταβλητές. Έτσι, η λογιστική παλινδρόμηση λαμβάνει τιμή για τον δείκτη AUROC 0,849, η ενισχυμένη λογιστική παλινδρόμηση 0,809, τα δέντρα ταξινόμησης και παλινδρόμησης 0,870, τα δέντρα λήψης αποφάσεων ενίσχυσης κλίσης 0,934 και τέλος τα νευρωνικά δίκτυα 0,828. Σε κάθε μέθοδο οι τιμές του δείκτη AUROC τείνουν προς τη μονάδα δείχνοντας έτσι την ικανοποιητική ικανότητα πρόβλεψης όλων των αλγορίθμων. Όπως αναφέρθηκε παραπάνω, καλύτερο μοντέλο πρόβλεψης είναι εκείνο που έχει μεγαλύτερη τιμή στον δείκτη AUROC με αποτέλεσμα καλύτερα αποτελέσματα πρόβλεψης να δίνει η μέθοδος Extreme Gradient Boosting.

Πίνακας 5.9: Πίνακας μέσης τιμής AUROC για κάθε μέθοδο

LR	LogitBoost	CART	xgbTree	NNET
0,849	0,809	0,870	0,934	0,828

Από την παραπάνω ανάλυση προκύπτει η κατάταξη των μεθόδων από την καλύτερη στην χειρότερη (Πίνακας 5.10).

Πίνακας 5.10: Πίνακας κατάταξης μεθόδων

Κατάταξη	Μέθοδος
1η	Extreme gradient boosting
2η	CART
3η	Logistic regression
4η	Neural network
5η	Boosted Logistic Regression

Για την πρόβλεψη των εξαρτημένων μεταβλητών χρησιμοποιήθηκαν 7 ανεξάρτητες μεταβλητές που όμως δεν έχουν ισάξια σημαντικότητα. Για το λόγο αυτό έγινε περαιτέρω διερεύνηση της σημαντικότητας των μεταβλητών έτσι ώστε να αποσαφηνιστεί ποιες μεταβλητές είναι σημαντικές για την εξαγωγή των αποτελεσμάτων. Έτσι μεγαλύτερη σημαντικότητα έχει μια μεταβλητή η οποία παρουσιάζει μεγαλύτερη τιμή κατά την ανάλυση σε σχέση με κάποια άλλη.

Όπως προκύπτει από την ανάλυση (πίνακας 5.11), για την μέθοδο extreme gradient boosting σημαντικότερη μεταβλητή για την πρόβλεψη του λογαριασμού μισθοδοσίας (Y4) είναι ο δείκτης ενεργού πελάτη (X6). Για την πρόβλεψη της μεταβλητής ηλεκτρονικού λογαριασμού (Y2) σημαντικότερο χαρακτηριστικό είναι το ενοίκιο (X7). Ακόμη, σημαντικότερο χαρακτηριστικό για την πρόβλεψη της μεταβλητής επενδύσεων σε αμοιβαία κεφάλαια (Y3) είναι το ενοίκιο. Για την πρόβλεψη του συνταξιοδοτικού προγράμματος (Y4) κάθε πελάτη σημαντικότερη είναι η μεταβλητή είναι ο δείκτης ενεργού πελάτη. Τέλος, για την πρόβλεψη της μεταβλητής πιστωτική κάρτα (Y5) σημαντικότερη είναι η μεταβλητή ενεργού πελάτη. Παρατηρείται ότι για την πρόβλεψη αγοράς των υπηρεσιών σημαντικότερες μεταβλητές είναι ο δείκτης ενεργού πελάτη και το ενοίκιο, ενώ ασήμαντες θεωρούνται οι μεταβλητές που καθορίζουν τους νέους πελάτες (X3) και τοποθεσία κατοικίας (X5) αντίστοιχα.

Πίνακας 5.11: Σημαντικότητα μεταβλητών στον αλγόριθμο xgbtree

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	2,97	58,35	0	25,31	0,30	100	41,46
Y2	3,83	55,06	0	55,97	2,02	67,24	100
Y3	2,77	39,87	0	77,10	0,81	37,19	100
Y4	2,97	58,35	0	25,31	0,30	100	41,46
Y5	1,72	59,32	1,49	30,11	0	100	45,29

Πίνακας 5.12: Σημαντικότητα για κάθε εξαρτημένη μεταβλητή

Κατάταξη	Y1	Y2	Y3	Y4	Y5
1	X6	X7	X7	X6	X6
2	X2	X6	X4	X2	X2
3	X7	X4	X2	X7	X7
4	X4	X2	X6	X4	X4
5	X1	X1	X1	X1	X1
6	X5	X5	X5	X5	X3
7	X3	X3	X3	X3	X5

Από την κατάταξη που προέκυψε από την μέτρηση του AUROC (Πίνακας 5.8), η μέθοδος CART είναι δεύτερη σε σειρά πρόβλεψης. Για την μέθοδο αυτή, γίνεται ανάλυση της σημαντικότητας των μεταβλητών εισόδου για την πρόβλεψη των μεταβλητών εξόδου (Πίνακας 5.13). Με κριτήριο τη μέγιστη τιμή της σημαντικότητα προκύπτει η κατάταξη για κάθε εξαρτημένη μεταβλητή (Πίνακας 5.14). Έτσι για την πρόβλεψη του λογαριασμού μισθοδοσίας (Y1), του ηλεκτρονικού λογαριασμού (Y2) και της πιστωτικής κάρτας (Y5) σημαντικότερη μεταβλητή είναι ο δείκτης ενεργού πελάτη (X6). Για την πρόβλεψη των επενδύσεων σε αμοιβαία κεφάλαια (Y3) σημαντικότερη μεταβλητή είναι η ηλικία (X2), ενώ για την πρόβλεψη του συνταξιοδοτικού προγράμματος (Y4) σημαντικότερη μεταβλητή είναι οι μήνες όπου ένας πελάτης συνεργάζεται με την τράπεζα (X4). Δεύτερη σε σημαντικότητα για όλες τις μεταβλητές με εξαίρεση τις επενδύσεις σε αμοιβαία κεφάλαια είναι η ηλικία (X2), ενώ την πρόβλεψη των επενδύσεων είναι οι μήνες όπου ένας πελάτης συνεργάζεται με την τράπεζα (X4). Αντιθέτως, λιγότερο σημαντική μεταβλητή για την πρόβλεψη του λογαριασμού μισθοδοσίας και της πιστωτικής κάρτας είναι ο δείκτης καταγωγής (X5), δηλαδή αν ένας πελάτης κατάγεται από την Ισπανία ή όχι. Για την πρόβλεψη του

ηλεκτρονικού λογαριασμού και του συνταξιοδοτικού προγράμματος η μεταβλητή με τη λιγότερη σημαντικότητα είναι το φύλο (X1). Τέλος, για την πρόβλεψη των επενδύσεων σε αμοιβαία κεφάλαια λιγότερη σημαντική μεταβλητή είναι το εισόδημα (X7) .

Πίνακας 5.13: Σημαντικότητα μεταβλητών στον αλγόριθμο CART

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	1.85	42.13	5.76	25.032	0.82	100	8.47
Y2	1.62	78.98	9.42	40.51	4.02	100	19.26
Y3	10.36	100	2.87	88.48	3.11	84.72	2.87
Y4	1.56	94.65	7.94	100	2.47	67.25	36.02
Y5	2.86	83.85	10.23	71.94	0.14	100	16.03

Πίνακας 5.14: Σημαντικότητα για κάθε εξαρτημένη μεταβλητή

Κατάταξη	Y1	Y2	Y3	Y4	Y5
1	X6	X6	X2	X4	X6
2	X2	X2	X4	X2	X2
3	X4	X4	X6	X6	X4
4	X7	X7	X1	X7	X7
5	X3	X3	X5	X3	X3
6	X1	X5	X3	X5	X1
7	X5	X1	X7	X1	X5

Από την κατάταξη βάσει του δείκτη AUROC (Πίνακας 5.8) προκύπτει ότι η Λογιστική Παλινδρόμηση είναι η τρίτη καλύτερη μέθοδος για την πρόβλεψη των εξαρτημένων μεταβλητών. Κάθε μεταβλητή εισόδου έχει άλλη επίδραση στο μοντέλο πρόβλεψης. Ο παρακάτω πίνακα (Πίνακας 5.15) περιλαμβάνει τις τιμές των συντελεστών των ανεξάρτητων μεταβλητών για την πρόβλεψη κάθε εξαρτημένης μεταβλητής.

Για την ανάλυση των μεταβλητών πρέπει να λαμβάνεται υπόψιν και η στατιστική σημαντικότητα των συντελεστών (Πίνακας 5.17). Η στατιστική σημαντικότητα κυμαίνεται από 0 έως1, ενώ μια μεταβλητή θεωρείται στατιστικά σημαντική όταν η τιμή του είναι μικρότερη από 0,05 και στατιστικά ασήμαντη σε διαφορετική περίπτωση.

Όταν μια μεταβλητή έχει θετικό συντελεστή στο μοντέλο, τότε αυξάνει την πιθανότητα κάποιος πελάτης να ανταποκριθεί θετικά, ενώ όταν ο συντελεστής έχει αρνητική τιμή η μεταβλητή μειώνει την πιθανότητα ένας πελάτης να ανταποκριθεί.

Για την πρόβλεψη της ύπαρξης λογαριασμού μισθοδοσίας (Y1), το φύλο (X1) και το εισόδημα των πελατών (X7) είναι στατιστικά ασήμαντες οπότε εξαιρούνται από την ανάλυση και εξετάζεται η επιρροή των υπολοίπων μεταβλητών. Για την πρόβλεψη του συνταξιοδοτικού προγράμματος προκύπτει ότι ο δείκτης νέου πελάτη (X3) και ο δείκτης κινητικότητας των πελατών (X6) είναι στατιστικά ασήμαντες οπότε δε λαμβάνονται υπόψιν και εξετάζεται η επιρροή των υπολοίπων μεταβλητών. Τέλος, για την ανάλυση της επιρροής των υπολοίπων ανεξάρτητων μεταβλητών, δεδεμένου ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές, εξετάζονται μόνο οι τιμές που λαμβάνουν στους συντελεστές αυξάνοντας ή μειώνοντας την πιθανότητα κάποιου πελάτη να ανταποκριθεί.

Πίνακας 5.15: Συντελεστές των μεταβλητών στο μοντέλο

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	0,01	-0,01	-1,08	0,00	0,30	4,77	0,00
Y2	0,20	0,02	-1,72	0,00	-0,40	2,74	0,00
Y3	0,39	0,05	-0,87	0,01	-1,07	5,19	0,00
Y4	-0,08	0,03	-16,61	0,01	0,66	29,17	0,00
Y5	0,21	0,01	-2,82	0,01	0,25	5,29	0,00

Πίνακας 5.16: Απόλυτες τιμές των συντελεστών των μεταβλητών

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	0,01	0,01	1,08	0,00	0,30	4,77	0,00
Y2	0,20	0,02	1,72	0,00	0,40	2,74	0,00
Y3	0,39	0,05	0,87	0,01	1,07	5,19	0,00
Y4	0,08	0,03	16,61	0,01	0,66	29,17	0,00
Y5	0,21	0,01	2,82	0,01	0,25	5,29	0,00

Πίνακας 5.17: Στατιστική σημαντικότητα των συντελεστών των μεταβλητών

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	0.70	0.00	0.00	0.00	0.00	0.00	0.28
Y2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y4	0.00	0.00	0.94	0.00	0.00	0.58	0.00
Y5	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Η μέθοδος όπου είναι κάνει λιγότερο καλές σε σχέση με την προηγούμενη, όπως υπολογίσθηκε (Πίνακας 5.8) είναι τα Νευρωνικά δίκτυα. Για τη μέθοδο αυτή, όπως και στις δύο πρώτες, η μέτρηση της σημαντικότητας των μεταβλητών πραγματοποιήθηκε με όμοιο τρόπο (Πίνακας 5.18). Η κατάταξη των μεθόδων παρουσιάζεται παρακάτω (Πίνακας 5.19). Έτσι, με εξαίρεση την πρόβλεψη των συνταξιοδοτικών προγραμμάτων (Y4), η πιο σημαντική μεταβλητή για την πρόβλεψη των υπολοίπων είναι πιστωτική κάρτα (X7). Για τα συνταξιοδοτικά προγράμματα σημαντικότερη κρίνεται ο δείκτης κινητικότητας των πελατών (X6). Επόμενη σε σημαντικότητα μεταβλητή για την πρόβλεψη του λογαριασμού μισθοδοσίας (Y1) είναι οι μήνες όπου ένας πελάτης συνεργάζεται με την τράπεζα (X4), ενώ για την πρόβλεψη του συνταξιοδοτικού προγράμματος είναι η ο δείκτης νέου πελάτη (X3). Για τις υπόλοιπες μεταβλητές εξόδου, επόμενη σε σημαντικότητα είναι ο δείκτης κινητικότητας των πελατών (X6). Τέλος, λιγότερο σημαντικές μεταβλητές είναι ο δείκτης νέου πελάτη για την πρόβλεψη του λογαριασμού μισθοδοσίας, η ηλικία (X2) για την πρόβλεψη του ηλεκτρονικού λογαριασμού (Y2), οι μήνες που ένας πελάτης συνεργάζεται με την τράπεζα για την πρόβλεψη των επενδύσεων σε αμοιβαία κεφάλαια (Y3) και των συνταξιοδοτικών προγραμμάτων και τέλος, το φύλο (X1) για την πρόβλεψη της ύπαρξης πιστωτικής κάρτας.

Πίνακας 5.18: Σημαντικότητα μεταβλητών στον αλγόριθμο Nueural Network

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	2,60	18,26	0	22,26	4,37	20,20	100
Y2	0,21	0	2,14	0,74	1,06	10,92	100
Y3	4,01	5,29	2,02	0	0,29	14,02	100
Y4	29,37	34,72	46,47	0	26,76	100	14,31
Y5	0	1,34	17,50	3,74	3,15	81,59	100

Πίνακας 5.19: Σημαντικότητα για κάθε εξαρτημένη μεταβλητή

Κατάταξη	Y1	Y2	Y3	Y4	Y5
1	X7	X7	X7	X6	X7
2	X4	X6	X6	X3	X6
3	X6	X3	X2	X2	X3
4	X2	X5	X1	X1	X4
5	X5	X4	X3	X5	X5
6	X1	X1	X5	X7	X2
7	X3	X2	X4	X4	X1

Από την κατάταξη (Πίνακας 5.8) προέκυψε ότι η χειρότερη μέθοδος για το συγκεκριμένο σύνολο δεδομένων είναι η Ενισχυμένη Λογιστική Παλινδρόμηση. Για τη μέθοδο αυτή η σημαντικότητα των μεταβλητών παρουσιάζεται παρακάτω (Πίνακας 5.20), ενώ η κατάταξη τους, από την καλύτερη στην χειρότερη αναγράφεται στον Πίνακα 5.21. Για την πρόβλεψη του λογαριασμού μισθοδοσίας (Y1), του ηλεκτρονικού λογαριασμού (Y2) και της πιστωτικής κάρτας (Y5) σημαντικότερη μεταβλητή είναι ο δείκτης κινητικότητας του πελάτη (X6). Όμοια, για την πρόβλεψη των επενδύσεων σε αμοιβαία κεφάλαια (Y3) καθώς και για την πρόβλεψη των συνταξιοδοτικών προγραμμάτων (Y4) σημαντικότερο στοιχείο είναι οι μήνες που ένας πελάτης συνεργάζεται με την τράπεζα (X4). Επιπροσθέτως, για την πρόβλεψη του λογαριασμού μισθοδοσίας και της πιστωτικής κάρτας δεύτερο σε κατάταξη έρχεται το στοιχείο των μηνών όπου ένας πελάτης συνεργάζεται με την τράπεζα. Για την πρόβλεψη των επενδύσεων σε αμοιβαία κεφάλαια (Y3) και των συνταξιοδοτικών προγραμμάτων δεύτερη σε σημαντικότητα μεταβλητή είναι δείκτης κινητικότητας των πελατών, ενώ

για την πρόβλεψη του ηλεκτρονικού λογαριασμού είναι η ηλικία (X2). Τέλος οι μεταβλητές εισόδου που είναι λιγότερο σημαντικές είναι εκείνες με τις χαμηλότερες τιμές σημαντικότητας για κάθε μεταβλητή εξόδου. Δηλαδή, για κάθε εξαρτημένη μεταβλητή με εξαίρεση την μεταβλητή επενδύσεων σε αμοιβαία κεφάλαια, λιγότερο σημαντική μεταβλητή εισόδου είναι ο δείκτης καταγωγής, (X5) ενώ για τις επενδύσεις σε αμοιβαία κεφάλαια λιγότερο σημαντικός είναι ο δείκτης νέου πελάτη (X3).

Πίνακας 5.20: Σημαντικότητα μεταβλητών στον αλγόριθμο Boosted Logistic Regression

Εξαρτ. μεταβλ.	X1	X2	X3	X4	X5	X6	X7
Y1	7,56	34,52	2,88	43,07	0	100	19,85
Y2	12,34	62,20	0,25	46,35	0	100	31,66
Y3	21,78	94,57	0	100	1,56	96,90	55,02
Y4	8,18	72,98	3,50	100	0	80,63	41,41
Y5	18,70	64,78	3,33	87,90	0	100	35,55

Πίνακας 5.21: Σημαντικότητα για κάθε εξαρτημένη μεταβλητή

Κατάταξη	Y1	Y2	Y3	Y4	Y5
1	X6	X6	X4	X4	X6
2	X4	X2	X6	X6	X4
3	X2	X4	X2	X2	X2
4	X7	X7	X7	X7	X7
5	X1	X1	X1	X1	X1
6	X3	X3	X5	X3	X3
7	X5	X5	X3	X5	X5

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία πραγματεύεται την επιστήμη της μηχανικής μάθησης στον τραπεζικό τομέα. Κατά τα πρώτα κεφάλαια πραγματοποιείται μια εισαγωγή στη διαχείριση των πελατειακών σχέσεων και στην εξόρυξη δεδομένων. Στη συνέχεια πραγματοποιείται μια εισαγωγή στις βασικές έννοιες και στις κατηγορίες των μοντέλων μηχανικής μάθησης. Η ανάλυση συνεχίζεται με την περιγραφή ορισμένων προβλημάτων που παρουσιάζονται και πρέπει να λαμβάνονται υπόψιν κατά την εφαρμογή των αλγορίθμων. Τέλος, γίνεται μια περιγραφή των εφαρμογών της μηχανικής μάθησης στην σημερινή εποχή.

Σκοπός είναι η καθοδήγηση του αναγνώστη μέσω της ανάλυσης της αξίας των πελατειακών σχέσεων κ της καθοριστικής συμβολής των εργαλείων και τεχνικών εξόρυξης δεδομένων στην εφαρμογή των αλγορίθμων μηχανικής μάθησης. Η εφαρμογή των αλγορίθμων αφορά την εύρεση των τραπεζικών προϊόντων μιας Ισπανικής τράπεζας που είναι πιο πιθανόν να αγορασθούν και από ποιους πελάτες βάσει της καταναλωτικής συμπεριφοράς των ίδιων αλλά και των πελατών που ήδη έχουν αγοράσει αυτά τα προϊόντα.

Για την επίλυση του προβλήματος χρησιμοποιήθηκαν μοντέλα ταξινόμησης, τα οποία εκτελέσθηκαν μέσω του προγράμματος R. Για το σκοπό αυτό χρησιμοποιήθηκαν οι εξής μέθοδοι: Λογιστική Παλινδρόμηση, Δέντρα λήψης αποφάσεων ενίσχυσης κλίσης, Δέντρα ταξινόμησης και παλινδρόμησης, Ενισχυμένη Λογιστική Παλινδρόμηση και Νευρωνικά δίκτυα. Στη συνέχεια έγινε η αξιολόγηση των μοντέλων που δημιουργήθηκαν από τους παραπάνω αλγορίθμους μέσω του δείκτη AUROC και πραγματοποιήθηκε η κατάταξη των μεθόδων βάσει της απόδοσης τους. Για κάθε μοντέλο έγινε υπολογισμός της σημαντικότητας των μεταβλητών που χρησιμοποιήθηκαν για να εξαχθούν τα αποτελέσματα.

Η χρήση των παραπάνω μεθόδων και μεθοδολογιών κρίνεται απαραίτητη για την προσέλκυση, την διατήρηση των πελατών της τράπεζας αλλά και για την ενίσχυση της ανταγωνιστικότητας της στην αγορά.

Τέλος, ορισμένες ενδιαφέρουσες επεκτάσεις της παρούσας εργασίας θα μπορούσε να είναι :

- Η λήψη δεδομένων από κάποια άλλη τράπεζα και εφαρμογή των ίδιων μεθόδων μηχανικής μάθησης με τις ίδιες ή και με παρόμοιες μεταβλητές πρόβλεψης.
- Εφαρμογή των ίδιων αλγορίθμων μηχανικής μάθησης με χρήση κάποιου άλλου λογισμικού όπως το Weka ή με χρήση της γλώσσας προγραμματισμού Python.
- Οποιαδήποτε εταιρεία θα μπορούσε να εφαρμόσει την ίδια ανάλυση για την προσέλκυση των πελατών της. Ακόμη, θα μπορούσε να εξετασθεί και να γίνει πρόβλεψη της πιθανότητας αθέτησης στην περίπτωση πραγματοποίησης μιας συναλλαγής σε δόσεις.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Agrawal, M.-L. (2003). Customer relationship management (CRM) and corporate renaissance, Journal of Service Research, 3(2), 150-171.

Alpaydin, E. (2020). Introduction to Machine Learning (4th ed., 39-44). MIT Press.

Berson, A., Smith, S., & Thearling, K. (2000). Building data mining applications for CRM, New York: McGraw-Hill.

Breiman, L., Friedman, J. -H., Olshen, R.- A., & Stone, C. -J. (1984). Classification and regression trees.

Buttle, F. (2009). Customer Relationship Management: Concepts and Technologies (2nd ed.). USA: Elsevier Ltd.

Chalmeta, R. (2006). Methodology for customer relationship management. Journal of Systems and Software, 79(7), 1015-1024.

Chen, T. & Guestrin, C. (2016). 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. XGBoost: A scalable tree boosting system, August 2016 (pp. 785-794). USA: Association for Computing Machinery.

Chorianopoulos, A. (2016). Effective CRM Using Predictive Analytics. John Wiley & Sons, Ltd.

Drucker, P. (1994). Knowledge, Work and Knowledge Society: The Social Transformations of this Century. The Edwin L. Goldkin Lecture, Harvard University's John F. Kennedy School of Government.

Fahy, J., & Jobber, D. (Επιμ.) (2014). Αρχές μάρκετινγκ. Αθήνα: Εκδόσεις Κριτική.

Freund,Y., & Schapire, R.-E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 55(1), 119-139.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting The Annals of Statistics, 28(2), 337-407. doi:10.1214/aos/1016218223.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). USA: Elsevier, Inc.

Hosmer Jr., D. -W., Lemeshow, S., & Sturdivant, R. -X. (2013). Applied Logistic Regression (3rd ed.). Hoboken, NJ, USA: John Wiley & Sons.

Gupta, P. (2017). Balancing Bias and Variance to Control Errors in Machine Learning. https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db.

Hand, D., Mannila, H. & Smyth, P. (2001). Principles of Data Mining. MIT Press.

Hustie, T., Tibshirani, R., & Friedmann, J. (2017). The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd ed., 37-41).

Kano, N., Seraku, N., Takahashi, F. & Tsuji, S. (1984). Attractive quality and must -be quality. Journal of Japanese Society for Quality Control, 14 (2), 39 -48

Kotler, P. & Armstrong, G. (2010). Principles of Marketing (13th ed.). NJ: Pearson Education, Inc.

Kumar, V., & Petersen, A. -J. (2012). Statistical Methods in Customer Relationship Management. UK: John Willey &Sons, Inc.

Kumar, V., & Reinartz, W. (2006). Customer Relationship Management: Concept, Strategy, and Tools (3rd ed). John Willey & Sons, Inc.

Langley, P. (1996). Elements of Machine Learning. Morgan Kaufmann Publishers, Inc.

Linoff, G., & Berry, M. (Ed.) (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (3rd ed.). Wiley Publishing, Inc.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of Machine Learning. MIT Press.

Nelson, S. -D. (2003). Management Update: The Eight Building Blocks of CRM. Gartner.

Payne, A. (2005). Handbook of CRM: Achieving Excellence in Customer Management (2nd ed.). UK: Butterworth-Heinemann. Springer.

Porter, M.E. (1980). Competitive strategy: techniques for analyzing industries and competitors. New York: Free Press

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. USA: O'Reilly Media, Inc.

Rajola, F. (Ed.) (2013). Customer Relationship Management in the Financial Industry: Organizational Processes and Technology Innovation.

Ripley, B. -D. (1996). Pattern recognition and neural networks. Pattern recognition and neural networks, Cambridge University Press.

Roddy M. (2002). Direct Marketing: A step-by-step Guide to Effective Planning and Targeting. US, Kogan Page Publishers.

Santos Garcia, C., Meincheim, A., Ribeiro Faria, E., Rosano Dallagassa, M., Vecino Sato, D.-M., Carvalho, D.-R., & Portela Santos, E.-A., & Scalabrin, E.-E. (2019). Process mining techniques and applications – A systematic mapping study. Expert Systems with Applications, 133, 260-295.

Strohmeier, S. (2013). Employee relationship management-Realizing competitive advantage through information technology. Human resource management review, 23(1), 93-104.

Suresh, H., & Guttag, J. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. https://arxiv.org/abs/1901.10002

Tan, P.-N., & Steinbach, M., & Kumar, V. (2014). Introduction to Data Mining. Pearson Education Limited.

Therneau, T.M., & Atkinson, E.J., & Foundation, M. (2019). An Introduction to Recursive Partioning Using the RPART Routines. CRAN R-project.

Tiwana, A. (2001). The essential guide to knowledge management: e-business and CRM applications. NJ: Prentice-Hall

van der Aalst, W. (2016). Process mining: Data science in action. Springer Berlin Heidelberg.

Κύρκος, Ε., -Γ. (2015). Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων: Ανακάλυψη Γνώσης για Λήψη Επιχειρηματικών Αποφάσεων. Αθήνα: Κάλλιπος.

Πανηγυράκης, Γ., Κορωνάκη, Ε., & Μπατσίλα, Σ. (2015). Επικοινωνία και Δημόσιες Σχέσεις-Μελέτες περιπτώσεων. Κάλλιπος

Φίτσιλης, Π. (2015). Σύγχρονα Πληροφοριακά Συστήματα Επιχειρήσεων. Αθήνα: Κάλλιπος.