ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. «Πληροφοριακά Συστήματα & Υπηρεσίες»

Ανάλυση τραπεζικών δεδομένων με τη χρήση αλγορίθμων μηχανικής μάθησης και δημιουργία μοντέλου εκτίμησης του δείκτη Customer Lifetime Value

Διπλωματική Εργασία

Μιχαήλ Σιγάλας
Α.Μ.: ΜΕ1824

Επιβλέπων Καθηγητής: Μιχαήλ Φιλιππάκης

Πειραιάς, Σεπτέμβριος 2020

Περίληψη

Βασικός στόχος της παρούσας εργασίας είναι να εξετάσει το καθεστώς λειτουργίας του τραπεζικού κλάδου σε σχέση με την ενσωμάτωση των νέων τεχνολογιών. Θα εκτιμηθούν τα περιθώρια ένταξης νέων τεχνολογιών στην παραγωγική διαδικασία και συγκεκριμένα εργαλείων τα οποία θα εκμεταλλεύονται τη γνώση και τις πρακτικές της μηχανικής μάθησης.

Πιο συγκεκριμένα, θα εξεταστεί η εφαρμογή της έννοιας της **Αξίας του Χρόνου Ζωής του Πελάτη – Customer Lifetime Value** (CLV) στους τραπεζικούς οργανισμούς, σε αντιπαραβολή με άλλους επιχειρηματικούς τομείς στους οποίους ήδη εφαρμόζεται με επιτυχία.

Στο πλαίσιο αυτό, θα επιχειρηθεί η κατανόηση των χαρακτηριστικών τα οποία καθορίζουν τη λειτουργία μίας τράπεζας καθώς και τις σχέσεις της με την αγορά, τόσο σε μακροοικονομικό όσο και σε μικροοικονομικό επίπεδο. Απώτερος σκοπός είναι η δημιουργία προτύπων τα οποία θα επεξηγούν την αλληλεπίδραση μίας τράπεζας με τους ίδιους τους πελάτες. Τα παραπάνω μεγέθη θα μετασχηματιστούν σε μεταβλητές και θα εκφραστούν μέσα από μαθηματικά και στατιστικά μεγέθη ώστε να είναι δυνατή η ανάλυση τους μέσω μοντέλων μηχανικής μάθησης.

Κατόπιν της στοιχειοθετήσεως του προβλήματος μέσω της επιχειρησιακής οπτικής, θα πραγματοποιηθεί καταγραφή και αξιολόγηση μεθόδων ανάλυσης και αλγορίθμων που προέρχονται από τον χώρο της στατιστικής και της μηχανικής μάθησης βάσει των χαρακτηριστικών τα οποία θα έχουν εξαχθεί από την μοντελοποίηση του προβλήματος.

Στη συνέχεια θα συγκεντρωθούν εφαρμογές και υλοποιήσεις από την υφιστάμενη τεχνογνωσία και βιβλιογραφία, οι οποίες θα ληφθούν υπόψη στον μετέπειτα σχηματισμό ενός μοντέλου μηχανικής μάθησης το οποίο θα επικεντρώνεται στη μέτρηση, διαχείριση και τελικώς την πρόβλεψη της Αξίας του Χρόνου Ζωής του Πελάτη σε έναν τραπεζικό οργανισμό.

Θα επιχειρηθεί η δημιουργία πολλαπλών μοντέλων μηχανικής μάθησης τα οποία θα έχουν τη δυνατότητα να εκφράσουν την έννοια του Key Performance Indicator " Customer lifetime value", με σκοπό την τελική αντιπαραβολή και αξιολόγησης της αξιοπιστίας των μοντέλων. Στη συνέχεια θα επιλεγεί το αποδοτικότερο μοντέλο το οποίο και θα αποτελέσει το βασικό μέρος μίας εφαρμογής η οποία θα δύναται να προβλέψει τις αγοραστικές ικανότητες και επιθυμίες ενός πελάτη τραπεζικών προϊόντων και υπηρεσιών.

Ως βασικό εργαλείο υλοποίησης επιλέχθηκε η γλώσσα προγραμματισμού Python καθώς και οι σουίτες οι οποίες την υποστηρίζουν. Σε επίπεδο τελικής εφαρμογής κρίθηκε χρήσιμη για τους σκοπούς της εργασίας η δημιουργία και η διαχείριση Βάσης δεδομένων (SQL) καθώς και επιπλέον εργαλείων ανάλυσης και οπτικοποίησης δεδομένων.

Πίνακας περιεχομένων

1. Εισαγωγή

Το τραπεζικό σύστημα και τα ιδρύματα τα οποία συμμετέχουν σε αυτό αποτελούν κινητήριο μοχλό της παγκόσμιας οικονομίας. Τα τελευταία χρόνια παρατηρείται ένας συνεχής μετασχηματισμός στη δομή των τραπεζών ο οποίος είναι άμεσα συσχετισμένος τόσο με τις πρόσφατες παγκόσμιες οικονομικές κρίσεις όσο και με την εισχώρηση της τεχνολογίας και των συναφών επιστημών στο καθεστώς λειτουργίας όλων των επιχειρήσεων και φορέων.

Η εισαγωγή του όρου Fintech ο οποίος τα τελευταία χρόνια χρησιμοποιείται όλο και περισσότερο όταν γίνεται αναφορά σε καινοτόμες χρηματοοικονομικές υπηρεσίες με την αξιοποίηση των τελευταίων εξελίξεων της τεχνολογίας, σε συνδυασμό με την υιοθέτηση τους από μία μεγάλη μερίδα του πληθυσμού η οποία πλέον στρέφεται σε ανάλογες επιλογές έναντι των παραδοσιακών τραπεζικών υπηρεσιών ως αποτέλεσμα και της διείσδυσης των νέων τεχνολογικών επιτευγμάτων στην καθημερινότητα των πολιτών, οδήγησε στην ανάγκη αναφοράς στην έννοια Fintech.

Τα επιχειρησιακά μοντέλα των χρηματοπιστωτικών οργανισμών αναδιαμορφώνονται και μετασχηματίζουν ανάλογα τις λειτουργίες τους με στόχο να ενσωματώσουν τις καινοτομίες της ψηφιακής επανάστασης.

Οι τράπεζες αποτελούν μία ιδιότυπη μορφή επιχειρήσεων και απλώνονται σε όλα τα στάδια της οικονομίας, ενώ αντλούν πελάτες από όλο το εύρος της κοινωνίας και της επιχειρηματικότητας. Απαιτείται λοιπόν η δημιουργία και η σχεδίαση εξατομικευμένων υπηρεσιών και προϊόντων ώστε να υπάρξει βιωσιμότητα κατά την μετάβαση στην ψηφιακή εποχή.

Το Customer Lifetime Value (CLV), ή αλλιώς, Αξία του Χρόνου Ζωής του Πελάτη, είναι ένα από τα σημαντικότερα KPIs τα οποία πρέπει να παρακολουθούνται και να αναλύονται στον τομέα του Customer Experience. Είναι μια μέτρηση που υποδεικνύει το πόσο πολύτιμος είναι ένας πελάτης για μια εταιρεία σε βάθος χρόνου, σε σχέση με την πρώτη αγορά που πραγματοποίησε. Οι μετρικές αυτές, βοηθούν τους επιχειρηματίες να διαμορφώνουν τις βέλτιστες στρατηγικές.

Πιο συγκεκριμένα, η παραπάνω μέθοδος αντανακλά την αξία που έχει ένας πελάτης για μια επιχείρηση, ακόμα και αν αυτή έχει τις ιδιαιτερότητες μίας τράπεζας, καθ' όλη τη διάρκεια της σχέσης τους. Αναμφισβήτητα, κοστίζει λιγότερο στις επιχειρήσεις να διατηρήσουν έναν πελάτη, από ότι να προσελκύσουν νέους. Επομένως, αυξάνοντας την αξία των υφιστάμενων πελατών, καθώς και των προϊόντων – υπηρεσιών που αυτοί επιλέγουν, οι επιχειρήσεις στοχεύουν στην ανάπτυξη.

2. Χρηματοοικονομική Τεχνολογία (FinTech) στον Τραπεζικό Τομέα

2.1 Η Εξέλιξη της Τραπεζικής

Η εξέλιξη της τεχνολογίας και συνάμα η ανάπτυξη της επιστήμης της πληροφορικής τις τελευταίες δεκαετίες έχει αποτελέσει καθοριστικό ρόλο στην αναδιαμόρφωση των θεσμών που διέπουν τις οργανωμένες κοινωνίες.

Η εξέλιξη αυτή, προκαλεί παράλληλα και τον αναπροσδιορισμό των επαγγελματικών κλάδων και ειδικοτήτων οι οποίες είναι απαραίτητες ώστε να καλυφθούν οι νέες ανάγκες. Ο εκθετικά αυξανόμενος όγκος της πληροφορίας που παράγεται πλέον, τεχνολογία η οποία προσδιορίζεται με τον όρο BIG DATA, απαιτεί τα κατάλληλα εργαλεία και τις τεχνολογίες όπως ψηφιακά συστήματα και αλγορίθμους, τα οποία δίνουν την δυνατότητα διαχείρισης και ανάδειξης των πληροφοριών αυτών. Η ανάπτυξη συνεπώς των τεχνολογιών αυτών, αποτελεί ένα απαραίτητο εργαλείο λήψης επιχειρηματικών και κοινωνικών αποφάσεων.

Αυτή η τεχνολογική ανάπτυξη συνδυάζει τις γνώσεις από διάφορους επιστημονικούς κλάδους, όπως τα μαθηματικά, τη στατιστική, τη γενετική, τη ρομποτική, τη μοριακή βιολογία και τη νανοτεχνολογία με κύριο γνώμονα την βελτίωση της ανθρώπινης σκέψης, δράσης και αντίδρασης έως και την υποκατάστασή της και όχι αντικατάστασή της. (ΟΤΟΕ, 2018)

Η σημερινή εποχή έχει κερδίσει τον όρο ψηφιακή, καθώς οι νέες εφαρμογές και τεχνολογίες είναι εμφανείς σε κάθε έκφανση της κοινωνίας καθώς και στην καθημερινότητα όλων, είτε πρόκειται για οικιακό χώρο είτε για εργασιακό είτε ακόμα και για το φυσικό περιβάλλον.

Σε αυτή τη νέα εποχή, πρωταγωνιστικό ρόλο έχουν τα δεδομένα και οι πληροφορίες. Ως εκ τούτου, μέσω της επιστήμης της πληροφορικής γίνεται η αξιοποίηση των πληροφοριών αυτών σε πρωτοβάθμιο επίπεδο. Ωστόσο, αναλόγως το είδος της πληροφορίας σε δεύτερο βαθμό, χρησιμοποιούνται γνώσεις από όλους τους υπόλοιπους τομείς ώστε τελικώς η ροή των δεδομένων να αξιοποιείται πλήρως.

Το φαινόμενο αυτό, το οποίο δικαίως αποκαλείται ως ψηφιακός μετασχηματισμός, έκανε το ξεκίνημά του την δεκαετία του 1970. Βασικό χαρακτηριστικό του ήταν η εδραίωση της επιστήμης της πληροφορικής καθώς και των τεχνολογιών οι οποίες εφαρμόστηκαν στους διάφορους τομείς παραγωγής και της βιομηχανίας. Απώτερος σκοπός ήταν από τότε η αυτοματοποίηση των διαδικασιών. Το διαδίκτυο όπως και η τεχνολογική εξέλιξη των υπολογιστών, εισήλθαν στις επαγγελματικές διεργασίες και η δυνατότητα της άμεσης επικοινωνίας και διασύνδεσης, παγκοσμιοποίησε την έρευνα και την ανάπτυξη με αποτέλεσμα οι νέες συνεργασίες να οδηγήσουν σε ακόμα πιο γρήγορα και έγκαιρα αποτελέσματα.

Οι νέες τεχνολογίες όπως αυτή της τεχνητής νοημοσύνης και της μηχανικής μάθησης χαρακτηρίστηκαν ως δεύτερος ψηφιακός μηχανισμός. Πρόκειται ουσιαστικά για την τεχνική ενσωμάτωση των φυσικών συστημάτων του κυβερνοχώρου (CPS: cyber physical system) στην παραγωγή, την διοίκηση και την χρήση του Διαδικτύου των Πραγμάτων και των υπηρεσιών στις ίδιες διαδικασίες.

Ως CPS ορίζονται οι διαδικτυακές συνδέσεις μεταξύ ανθρώπων και μηχανών καθώς και τα νέα πληροφοριακά συστήματα και προϊόντα.

Ο δεύτερος ψηφιακός μετασχηματισμός με βασικό πρωταγωνιστή την τεχνητή νοημοσύνη και των εφαρμογών της, θα οδηγήσει σε νέα τάξη πραγμάτων αναφορικά με τη ριζική αναδιοργάνωση και συστηματοποίηση κάθε επιπέδου της παραγωγικής διαδικασίας, της διανομής και διάθεσης κάθε μορφής εργασίας. (ΟΤΟΕ, 2018)

2.2 Ψηφιακός Μετασχηματισμός

Ο ψηφιακός μετασχηματισμός καθώς και όσα αυτός συνεπάγεται, στηρίζεται σε επιστημονικούς τομείς οι οποίοι εξελίσσονται συνεχώς και αποτελούν πεδίο μελέτης για επιχειρηματίες και επιστήμονες. Βάσει έρευνας του Συλλόγου Ελλήνων Βιομηχάνων «Η ψηφιακή Ελλάδα, ο δρόμος προς την ανάπτυξη», το Μάιο 2017, οι τεχνολογίες αιχμής οι οποίες αποτελούν τον ψηφιακό μετασχηματισμό είναι οι παρακάτω:

- **Υπολογιστικό νέφος (cloud)**: η διαδικτυακή κεντρική χρήση υπολογιστικών πόρων με μεγάλη ευελιξία και βαθμό αυτοματοποίησης, η αποθήκευση, χρήση δεδομένων, λογισμικού ή και υπηρεσιών παρέχονται διαδικτυακά.

- **Ανάλυση δεδομένων μεγάλου όγκου (big data analytics)**: αναφέρεται στην εξαγωγή, τη στατιστική επεξεργασία και την τελική ερμηνεία μεγάλου όγκου δεδομένων.

- **Εικονική πραγματικότητα (virtual reality)**: Αναφέρεται σε ένα τρισδιάστατο ψηφιακό περιβάλλον που δίνει τη δυνατότητα αλληλεπίδρασης και επιτρέπει τον απευθείας χειρισμό ή την προσομοίωση αληθοφανών σεναρίων σε πραγματικό χρόνο.

- **Επαυξημένη πραγματικότητα (augmented reality)**: Ο συνδυασμός φυσικών και ψηφιακών πληροφοριών σε πραγματικό χρόνο με χρήση διεπαφών που επιτρέπουν την αντιπαραβολή ψηφιακών πληροφοριών πάνω στο φυσικό περιβάλλον.

- **Τεχνητή νοημοσύνη (virtual reality)**: Αποτελεί την επιστήμη που στόχο έχει την ανάπτυξη συστημάτων που αντιλαμβάνονται, επικοινωνούν και μπορούν να σκεφτούν ορθολογικά, μέσω της χρήσης σύνθετων αλγορίθμων και προηγμένων μεθόδων ανάλυσης.

- **Πλατφόρμες επιχειρησιακής συνεργασίας (enterprise collaboration platforms)**: πλατφόρμες που βελτιστοποιούν την εταιρική συνεργασία μέσω δυνατότητας αποστολής γραπτών μνημάτων ή βίντεο-κλήσεων, ανταλλαγής πληροφοριών, δεδομένων και αρχείων και επιτρέπουν την απρόσκοπτη απομακρυσμένη εργασία, καθιστώντας τις εταιρείες πιο ενεργές και ευέλικτες.

- **Διαδίκτυο των πραγμάτων (internet of things)**: αναφέρεται στη σύνδεση φυσικών αντικειμένων μέσω έξυπνων αισθητήρων με το διαδίκτυο για τη συλλογή δεδομένων και την ανάληψη κάποιας δράσης, όπως αλλαγής παραμέτρων ή προειδοποίησης.

- **Αλυσίδα των μπλοκ (blockchain)**: τεχνολογία που επιτρέπει την καταγραφή μιας μεγάλης λίστας συναλλαγών με ασφάλεια και δίχως τη δυνατότητα παρέμβασης, λόγω της πληροφορίας πάνω στην οποία στηρίζονται τα κρυπτό-συναλλάγματα όπως το bitcoin.

- **Μάθηση μηχανής (machine learning)**: η χρήση προηγμένης τεχνητής νοημοσύνης για την ανάδειξη μοτίβων σε δεδομένα ή συμπεριφορές και τη συνεχή αναπροσαρμογή για την βελτίωση της απόδοσης και της αποτελεσματικότητας της εκτέλεσης μιας εργασίας.

- **Τρισδιάστατη εκτύπωση (3Dprinting)**: κατασκευή αντικειμένων μέσω εναπόθεσης ενός υλικού (συνήθως κάποιου πλαστικού) μέσω κεφαλής εκτύπωσης. Είναι κατάλληλη για γρήγορη και φθηνή κατασκευή σύνθετων αντικειμένων, ειδικά στο στάδιο πειραματισμού.

- **Αυτόνομα ρομπότ**: μηχανές που διαθέτουν τον κατάλληλο βαθμό τεχνητής νοημοσύνης ώστε να μπορούν να εκτελέσουν επαναλαμβανόμενες εργασίες με υψηλό βαθμό αυτονομίας, ανταποκρινόμενες στο περιβάλλον τους δίχως επίβλεψη.

- **Δι-επαφές προγραμματισμού εφαρμογών (APIs)**: ένα επίπεδο ενδιάμεσου λογισμικού μεταξύ δυο ή περισσότερων εφαρμογών που τους επιτρέπει την εκπλήρωση αιτημάτων ή την ανταλλαγή πληροφοριών και δεδομένων με μια συγκεκριμένη δομημένη μορφή και ταξινομία. (ΣΕΒ, 2017)

2.3 Η Τραπεζική στην Ψηφιακή Εποχή

Τα χρηματοπιστωτικά ιδρύματα εκ φύσεως τους αποτελούν οργανισμούς οι οποίοι ανέκαθεν είχαν στη διάθεση τους πληθώρα δεδομένων προερχόμενα από όλους τους τομείς της επιχειρηματικότητας και εν γένει της οικονομικές δραστηριότητες της κοινωνίας. Ωστόσο, τα δεδομένα αυτά και οι πληροφορίες αλλάζουν δομή συνεχώς.

Η παλαιά λογική της τήρησης των δεδομένων αποκλειστικά για ελεγκτικούς σκοπούς και για θέματα ασφαλείας έχει αλλάξει καθώς το νέο δόγμα επιτάσσει την αξιοποίηση των πληροφοριών αυτών με σκοπό την αύξηση του κέρδους, των προσωποποιημένων υπηρεσιών και το πλεονέκτημα έναντι του ανταγωνισμού. Πιο συγκεκριμένα, απαιτούνται νέες εξειδικεύσεις και ειδικότητες οι οποίες θα εξασφαλίσουν την ομαλή

μετάβαση στην ψηφιακή εποχή δίχως να αλλοιωθεί η αξιοπιστία και η ομαλή λειτουργία των ιδρυμάτων.

Επίσης, η επικαιροποίηση της στρατηγικής των τραπεζών καθίσταται απαραίτητη με σκοπό την έγκαιρη λήψη των ορθών αποφάσεων σε ένα συνεχώς μεταβαλλόμενο περιβάλλον με αχαρτογράφητες κατευθύνσεις.

Η νέα οργανωτική μορφή των τραπεζών βάσει των νέων εξελίξεων αποτελεί μεγάλο στοίχημα για τις διοικήσεις αυτών. Η αναδιαμόρφωση αυτή στηρίζεται επί τω πλείστων στη δημιουργία νέων ειδικοτήτων και ρόλων εργασίας καθώς και στην αναθεωρημένη αξιοποίηση του ανθρωπίνου δυναμικού των ιδρυμάτων σε νέους τομείς, καθώς πολλά από τα αντικείμενα της παραδοσιακής τραπεζικής τείνουν να εκλείψουν.

Σημείο μελέτης και αντιπαραθέσεων στον τραπεζικό χώρο αποτελεί ο περιορισμός της φυσικής δραστηριότητας των ιδρυμάτων, δηλαδή των καταστημάτων, καθώς ένα βασικό κομμάτι του ψηφιακού μετασχηματισμού αποτελεί η διεύρυνση των εναλλακτικών δικτύων δραστηριοποίησης. Ωστόσο, η στρατηγική η οποία κυριαρχεί είναι αυτή της παράλληλης λειτουργίας τόσο των παραδοσιακών καναλιών εξυπηρέτησης του πελάτη όσο και η ανάπτυξη ψηφιακών εργαλείων τα οποία αποσκοπούν στην αύξηση της ποιότητας των υπηρεσιών καθώς και την ελαχιστοποίηση του λειτουργικού κόστους. Το αύριο των φυσικών καταστημάτων όπως είναι φυσικό θα είναι εντελώς διαφορετικό, καθώς πολλές από τις παλαιές εργασίες θα εκτελούνται αποκλειστικά από τα εναλλακτικά δίκτυα. Ουσιαστικά το νέο προφίλ των καταστημάτων θα είναι συμβουλευτικό και όχι διεκπεραιοτικό.

Η υφιστάμενη κατάσταση στην Ελλάδα διαμορφώνεται με σημείο αναφοράς τις 4 συστημικές τράπεζες οι οποίες από το 2015 έχουν καταρτίσει τριετές πλάνο μετασχηματισμού. Πρόκειται στην ουσία για ένα νέο σχέδιο δράσης το οποίο δίνει έμφαση στην προώθηση νέων καναλιών εξυπηρέτησης τα οποία θα αναπτυχθούν ραγδαία τα επόμενα χρόνια. Βασικός στόχος είναι να αναπτυχθούν ψηφιακές πλατφόρμες οι οποίες θα είναι ικανές να καλύψουν το μεγαλύτερο εύρος των εργασιών. Επίσης, οι πλατφόρμες αυτές θα διαχειρίζονται έναν τεράστιο όγκο δεδομένων, όπου η αξιοποίηση αυτών θα είναι προς όφελος και των υπηρεσιών προς τους πελάτες αλλά και των ίδιων των τραπεζών. Συνεπώς, η διαδικασία της καταχώρησης, της αποθήκευσης και της ανάλυσης της πληροφορίας είναι αναγκαίο να πραγματοποιείται με αποδοτικό και αποτελεσματικό τρόπο. Στρατηγικός στόχος των τραπεζών της χώρας είναι η δημιουργία σταθερών βάσεων προκειμένου να προβούν σε πιο αποφασιστικά βήματα καθώς οι διαρκώς μεταβαλλόμενες εξελίξεις το απαιτούν. (Μαριόλη, 2016)

Ο ψηφιακός μετασχηματισμός χαρακτηρίζεται ως προτεραιότητα για το παγκόσμιο τραπεζικό σύστημα. Οι τράπεζες θέτουν ως βασικές παραμέτρους για την επιτυχία του μετασχηματισμού την διαχείριση του κινδύνου και την αξιοποίηση νέων ευκαιριών ανάπτυξης. Η ανάπτυξη των κριτηρίων αυτών θα είναι δείκτης επιτυχίας των επενδύσεων στις νέες τεχνολογίες.

Σημαντικό στοιχείο του μετασχηματισμού ωστόσο αποτελεί η ασφάλεια στον κυβερνοχώρο, σε τέτοιο σημείο μάλιστα που θεωρείται πλέον πρώτη προτεραιότητα για τις τράπεζες, αφήνοντας στις επόμενες θέσεις για το έτος 2018 τη διαχείριση κινδύνων, την κανονιστική συμμόρφωση.

Ακόμα ένας σημαντικός παράγοντας με τον οποίο ασχολούνται οι διοικήσεις των τραπεζών είναι αυτός του ανθρώπινου δυναμικού καθώς αποτελεί πρόκληση η εξεύρεση και η διατήρηση του καταλλήλου ανθρώπινου δυναμικού το οποίο θα είναι ικανό να στελεχώσει νέες, νευραλγικές θέσεις εργασίας. Στοχεύονται κατά κύριο λόγο στελέχη με εμπειρία στον κυβερνοχώρο ο οποίος αποτελεί ένα καινούργιο περιβάλλον εξειδίκευσης του εργατικού δυναμικού.

Βάσει έρευνας η οποία συγκέντρωσε τις οπτικές ανωτέρων τραπεζικών στελεχών από 221 τράπεζες της Ευρώπης, της Αμερικής και της Ασίας, αναδεικνύεται η πεποίθηση όλων για την αναγκαιότητα του ψηφιακού μετασχηματισμού.

Ως εκ τούτου, πλέον βασικός οδηγός για την ανάπτυξη του τομέα, είναι η καινοτομία και όχι τόσο το ρυθμιστικό και κανονιστικό πλαίσιο. Πιο συγκεκριμένα, από όσους συμμετείχαν στην εν λόγω έρευνα, το 62% θεωρεί ότι θα επιτύχει τους βασικούς στόχους του ψηφιακού μετασχηματισμού εντός του 2020 ενώ το 19% πιστεύει ότι είναι ήδη πρωτοπόρες στην νέα εποχή.

Αξιοσημείωτο είναι ότι τα αποτελέσματα της έρευνας αναδεικνύουν σε ποσοστό 59% τις τράπεζες εκείνες οι οποίες σκοπεύουν να αυξήσουν τον προϋπολογισμό για τις νέες τεχνολογίες στο 10% τα επόμενα έτη. Η επένδυση στη νέα εποχή βασίζεται σε ποσοστό 44% σε έρευνα και λογισμικά που παράγουν τρίτοι συνεργάτες ενώ το 17% επιδιώκουν να αποκτήσουν έτοιμη τεχνογνωσία μέσω εξαγορών. Ωστόσο, το 70% των τραπεζών θεωρεί ότι μέσω της ενδυνάμωσης της ψηφιακής τους τεχνολογίας, θα αυξηθεί δραστικά το ανταγωνιστικό τους πλεονέκτημα.

Σημαντικός τομέας ο οποίος διαδραματίζει βασικό ρόλο στον μετασχηματισμό είναι η ασφάλεια του κυβερνοχώρου και των δεδομένων καθώς η ενίσχυσή του αποτελεί προτεραιότητα για το 73% των τραπεζών. Η επένδυση συνεπώς στην απαραίτητη τεχνολογία και στο εξειδικευμένο προσωπικό είναι μονόδρομος ώστε τα ιδρύματα να θωρακιστούν απέναντι σε απειλές οι οποίες σχετίζονται με τον κυβερνοχώρο και εν γένει την ψηφιακή ασφάλεια.

Σε αυτό το σημείο αναδεικνύεται και ο πρωταγωνιστικός ρόλος των εταιριών χρηματοοικονομικών τεχνολογιών (fintech) οι οποίες καλούνται είτε να διεκδικήσουν ζωτικό χώρο στην αγορά είτε να λειτουργήσουν σε συνεργασία με τα παραδοσιακά τραπεζικά ιδρύματα. Οι νέες αυτές συνεργασίες οι οποίες θα προκύψουν στο εγγύς μέλλον αναμένεται να δώσουν μεγάλη ώθηση στην ανάπτυξη νέων τεχνολογιών οι οποίες δύναται να εφαρμοστούν άμεσα στο παραγωγικό περιβάλλον. Αυτό θα έχει ως αποτέλεσμα την αναβάθμιση των υπηρεσιών προς τους πελάτες.

Ο τραπεζικός τομέας, αν και παραδοσιακά χαρακτηριζόταν από ήπιο ανταγωνισμό σε σχέση με άλλους πιο δυναμικούς τομείς επιχειρηματικότητας λόγω ρυθμιστικών και κανονιστικών πλαισίων, τα επόμενα χρόνια αναμένεται να αναζωπυρωθεί καθώς η είσοδος των νέων εξω-τραπεζικών οντοτήτων θα αναγκάσει τα παραδοσιακά ιδρύματα να προσφέρουν στους πελάτες προϊόντα και υπηρεσίες υψηλής ποιότητας. Συνάμα, είναι μόνιμη προϋπόθεση η διατήρηση της αξιοπιστίας των υπηρεσιών σε υψηλό επίπεδο καθώς λόγω της φύσεως του κλάδου, δεν είναι εφικτή η υιοθέτηση λύσεων και πρακτικών οι οποίες δε θα καλύπτουν τις αυστηρές εποπτικές απαιτήσεις.

Εστιάζοντας στο ελληνικό τραπεζικό τοπίο, αξίζει να σημειωθεί ότι ήδη έχουν ξεκινήσει να γίνονται σημαντικά βήματα προς τον ψηφιακό μετασχηματισμό. Σημαντικό ρόλο σε αυτόν παίζουν η ανάπτυξη του mobile και του web banking, υπηρεσίες οι οποίες βρίσκονται ήδη σε αρκετά εξελιγμένο επίπεδο, βάζοντας σε επίπεδο σύγκρισης ακόμα και μεγάλα ιδρύματα του εξωτερικού. Απώτερος σκοπός φυσικά είναι η μετατροπή όσο το δυνατόν περισσότερων παραδοσιακών πελατών σε χρήστες των ψηφιακών υπηρεσιών. (EY, 2018)

2.4 Fintech Τεχνολογία και Τραπεζική

Ως FinTech (Financial Technology) ορίζονται οι οργανισμοί εκείνοι οι οποίοι προάγουν τις τεχνολογικές καινοτομίες που εντάσσονται στον τομέα των χρηματοοικονομικών υπηρεσιών. Αυτές οι νέες τεχνολογίες αποτελούν το έναυσμα για τη δημιουργία νέων εφαρμογών, επιχειρησιακών αποφάσεων, υπηρεσιών και προϊόντων. (ECB, 2018)

Οι FinTech επιχειρήσεις έχουν ως προτεραιότητα την πρωτοτυπία μέσω της τεχνολογίας. Οι πρώτες υλοποιήσεις οι οποίες έχουν ήδη ενταχθεί σε παραγωγικό περιβάλλον έχουν να κάνουν κατά κύριο λόγο με νέα συστήματα πληρωμών και με αυτοματοποιημένες επενδυτικές συμβουλές.

2.4.1 Οι Βασικοί πυλώνες του Fintech

Οι υπηρεσίες FinTech όπως είναι προφανές, δεν αποτελούν έναν εξ ολοκλήρου νέο τομέα, εντούτοις έρχονται να εκσυγχρονίσουν τον ήδη υφιστάμενο τομέα της Οικονομίας και των χρηματοοικονομικών υπηρεσιών καθώς και τις τεχνολογίες της πληροφορίας. Ως βασικές κατηγορίες δραστηριοποίησης των υπηρεσιών αυτών, ορίζονται οι κατωτέρω:

1. Χρηματοδότηση και επενδύσεις
2. Διαχείριση κινδύνου
3. Πληρωμές και υποδομές
4. Ασφάλεια δεδομένων και δημιουργία εσόδων
5. Διασύνδεση του πελάτη

Χρηματοδότηση και επενδύσεις: Βασική κατεύθυνση των Fintech είναι οι εναλλακτικές πηγές χρηματοδότησης με πρωταγωνιστή το crowd funding. Η «χρηματοδότηση από το πλήθος» αποτελεί μία παλαιά τεχνική η οποία όμως στις μέρες μας έχει βρει ευρύ πεδίο εφαρμογής. Βασίζεται κατά κύριο λόγο σε συνδρομητικά επιχειρηματικά μοντέλα και η βασική ιδέα είναι αρκετά απλή. Ο κάθε χρήστης μίας υπηρεσίας ή ενός μελλοντικού προϊόντος συνεισφέρει με όποιο ποσό ανταποκρίνεται στις δυνατότητές του παίρνοντας ως αντάλλαγμα κάποιο από τα προκαθορισμένα πακέτα παροχών τα οποία προσφέρει ο χρηματοδοτούμενος. Τα πακέτα αυτά περιλαμβάνουν αποκλειστικά προνόμια με σκοπό την προσέλκυση νέων χρηματοδοτών ακόμα και σε επίπεδο μεμονωμένων ιδιωτών. Τέτοια προνόμια συνήθως είναι μία σημαντική έκπτωση επί του προσφερόμενου προϊόντος ή υπηρεσίας ή κάποια συλλεκτική έκδοση και διαφορά δώρα.

Για την υλοποίηση του όλου εγχειρήματος έχουν σχεδιαστεί ειδικές πλατφόρμες οι οποίες παρέχουν όλες τις απαραίτητες υποδομές καθώς και τα απαιτούμενα επίπεδα ασφαλείας. Δημοφιλείς τέτοιες πλατφόρμες είναι το Kickstarter, το Indiegogo, το RocketHub, το Fundly, το Appsplit , το GoGetFunding καθώς και το Peerbackers.

Διαχείριση κινδύνου: Τα χρηματοπιστωτικά ιδρύματα από καταβολής τους είναι εκτεθειμένα σε διάφορες μορφές κινδύνων. Η εξέλιξη της τεχνολογίας προσδίδει τη δυνατότητα βελτίωσης των υφιστάμενων υπηρεσιών αλλά και των τεχνολογικών υποδομών. Είναι γεγονός ωστόσο ότι παράλληλα αυξάνονται οι διάφοροι κίνδυνοι. Αποτελεί χρέος συνεπώς για τις τράπεζες η εφαρμογή κατάλληλων διαδικασιών και ελέγχων προκειμένου οι κίνδυνοι αυτοί να περιοριστούν στο ελάχιστο δυνατό. Αντίστοιχα και οι νέες εταιρίες Fintech εποπτεύονται από τους σχετικούς εποπτικούς μηχανισμούς. Υπεύθυνη για τον καθορισμού ενός ενιαίου εποπτικού πλαισίου είναι η Ευρωπαϊκή Κεντρική Τράπεζα και τα επιμέρους όργανά της.

Πληρωμές και υποδομές: Ύστερα από την επιβολή των Capital Control στην Ελλάδα, σύμφωνα με στοιχεία της Visa Europe, το πρώτο διάστημα ισχύς των μέτρων παρατηρήθηκε αύξηση 135% στις συναλλαγές των καρτών σε σχέση με το αμέσως προηγούμενο διάστημα.

Το βασικό μερίδιο των συναλλαγών με κάρτες κατείχε ο τομέας των τροφίμων, με αύξηση στο 234% και εν συνεχεία ο τομέας της υγείας με 206% καθώς και τα πρατήρια καυσίμων στο 193%.

Τα Capital Controls του 2015 αποτέλεσαν συνεπώς το εφαλτήριο για την ραγδαία ανάπτυξη των ηλεκτρονικών πληρωμών αφού αναμένεται αύξηση των συσκευών αποδοχής καρτών στην αγορά από 150.000 σε 400.000 τα επόμενα δύο χρόνια. Αξιοσημείωτο είναι επίσης ότι πενταπλασιάστηκαν οι νέοι κωδικοί e- banking. Όλα τα ανωτέρω στοιχεία, υποδεικνύουν συνεπώς μία μόνιμη τάση αύξησης των εναλλακτικών δικτύων.

Η εξέλιξη αυτή συνεισφέρει παράλληλα στην μείωση της φοροδιαφυγής, σε επίπεδο ακόμα και κοντά στο 25%, με αποτέλεσμα η νέα τεχνολογία να δίνει λύσεις σε αρκετά ανεπίλυτα προβλήματα της παγκόσμιας οικονομίας.

Ασφάλεια των δεδομένων και δημιουργία εσόδων: Βασική προτεραιότητα των εταιριών Fintech αποτελεί η συμμόρφωση με τα διεθνή πρότυπα ασφαλείας GDPR όπως αυτά ορίστηκαν από την Ευρωπαϊκή Επιτροπή. Βάσει των κανόνων αυτών οι εταιρίες υπηρεσιών πληρωμών παρέχουν εφαρμογές οι οποίες είναι αποδοτικές, καινοτόμες αλλά συνάμα και ασφαλείς.

Η ανάπτυξη των ψηφιακών υπηρεσιών αναμένεται να δημιουργήσει πιθανούς κινδύνους ασφαλείας για τα τραπεζικά ιδρύματα και τους καταναλωτές. Είναι λογικό όσο πιο πολύ διευρύνεται το φάσμα των εν λόγω υπηρεσιών, τόσο πιο πολύ προσεκτική θα πρέπει να είναι η ανάπτυξη των διάφορων προϊόντων. Κάτι τέτοιο φυσικά απαιτεί και τις κατάλληλες υποδομές, εξειδικεύσεις και εκπαίδευση, ώστε να εξασφαλίζεται η απαιτούμενη ασφάλεια στις συναλλαγές.

Όπως είναι λογικό, τα παραδοσιακά τραπεζικά ιδρύματα έχουν ένα βασικό προβάδισμα σε σχέση με τις Fintech, καθώς σε αυτά υπάρχουν ήδη σχετικές δομές αλλά και τεχνογνωσία ώστε να αντιμετωπίσουν πιο αποτελεσματικά την έκθεση στον οποιαδήποτε νέο κίνδυνο.

Διασύνδεση του πελάτη: Ένα από τα βασικά ζητούμενα του ψηφιακού μετασχηματισμού είναι η βελτίωση της εμπειρίας του πελάτη, καθώς οι υπηρεσίες και οι συναλλαγές οι οποίες ψηφιοποιούνται διακατέχονται αρκετές φορές από μεγάλη πολυπλοκότητα η οποία δυσχεραίνει την καθημερινότητα του τελικού χρήστη. Για τον λόγο αυτό, αρκετά ποσά επενδύονται με σκοπό την βελτίωση των υφιστάμενων πληροφοριών αλλά και την δημιουργία νέων. (ETEN, 2018)

2.4.2 Η Οδηγία PSD2

Σημαντικό σημείο για την ευρύτερη ανάπτυξη των ψηφιακών οικονομικών υπηρεσιών αποτέλεσε η αναθεωρημένη οδηγία «PSD 2» η οποία σχετίζεται με τις υπηρεσίες πληρωμών. Πιο συγκεκριμένα, αφορά στην υποχρέωση των τραπεζών να παρέχουν δεδομένα σε τρίτους ώστε να είναι δυνατή η ανάπτυξη νέου λογισμικού το οποίο θα παρέχει αναβαθμισμένες υπηρεσίες προς τους καταναλωτές. Πρόκειται ουσιαστικά για την αφετηρία του «open banking» το οποίο πέραν από τις παραδοσιακές τράπεζες θα συμπεριλαμβάνει και όλες τις νεοφυείς χρηματοοικονομικές εταιρίες.

Ουσιαστικά, κάθε φυσική ή νομική οντότητα θα έχει την δυνατότητα να τηρεί τραπεζικό λογαριασμό σε ένα τραπεζικό ίδρυμα ενώ ταυτόχρονα θα δύναται να χρησιμοποιεί τρίτες εφαρμογές και συνδυαστικές υπηρεσίες ώστε να ολοκληρώνει τις συναλλαγές της.

Όσον αφορά την Ελλάδα, αξίζει να σημειωθεί ότι η Τράπεζα της Ελλάδος έχει ήδη προχωρήσει στην αδειοδότηση εταιριών πληρωμών και διαχείρισης ηλεκτρονικού χρήματος.

Σε παγκόσμιο επίπεδο, χαρακτηριστικό παράδειγμα αποτελεί η Revolut, η οποία πρόκειται για μία από τις πιο πετυχημένες start-up εταιρίες στον χρηματοοικονομικό τομέα. Η εν λόγω εταιρία έχει χρήστες και στην Ελλάδα από τον Απρίλιο του 2018 όπως και σε όλη την Ευρώπη, απαριθμώντας συνολικά πάνω από 3 εκατομμύρια χρήστες. Ενώ η αξία της υπολογίζεται στα 1,7 δισεκατομμύρια δολάρια. (Ρεντούμης, 2018)

2.4.3 Νέες Χρηματοοικονομικές Τεχνολογίες

Η ανάπτυξη των Fintech εταιριών αλλά και η εξέλιξη της ψηφιακής τραπεζικής δεν θα ήταν εφικτά εάν δεν αναπτύσσονταν παράλληλα και οι νέες τεχνολογικές δυνατότητες όπως οι κατωτέρω.

- **Big Data Analytics**

Ως Big Data ορίζεται ο τεράστιος όγκος δεδομένων ο οποίος δεν είναι εφικτό να επεξεργασθεί και αναλυθεί μέσω των παραδοσιακών εργαλείων.

Όλοι οι Οργανισμοί, συγκεντρώνουν σε καθημερινή βάση τεράστιους όγκους δεδομένων. Βάσει στοιχείων της IBM, περισσότερο από 2,5 τετράκις εκατομμύρια bytes δεδομένων παράγονται κάθε έτος, ενώ το 90% έχει δημιουργηθεί τα τελευταία δύο χρόνια. Κάτι τέτοιο θεωρείται φυσιολογικό καθώς δεδομένα παράγονται συνεχώς σε όλες τις εκφάνσεις της καθημερινότητάς μας, είτε πρόκειται για ψηφιακή αλληλεπίδραση μέσω ηλεκτρονικών αγορών είτε από απλή περιήγηση στο διαδίκτυο. Αξίζει να σημειωθεί ότι βάσει του άρθρου της McKinsey ("Big Data: The next frontier for innovation, competition and productivity", 2011), οι εταιρίες Fintech είναι πρωταγωνίστριες στην διαχείριση και αποθήκευση των μεγάλων δεδομένων.

Η επανάσταση την οποία έχουν προκαλέσει τα Μεγάλα Δεδομένα, μετασχηματίζει συνεχώς της μορφολογία και τη λειτουργία όλων των επιχειρήσεων. Η αξιοποίησή τους ωστόσο δεν είναι εύκολη, καθώς απαιτείται η κατάλληλη τεχνογνωσία ώστε να αφομοιωθεί η σωστή πληροφόρηση και να αποτελέσει εργαλείο λήψης αποφάσεων. Η προσέγγιση των χρηματοοικονομικών εταιριών σε σχέση με την ανάλυση των δεδομένων τους, έγκειται στις εξής κατηγορίες.

- **Κατηγοριοποίηση Πελατών**

Η κατηγοριοποίηση των πελατών είναι σημαντική για τις επιχειρήσεις καθώς βάσει κριτηρίων όπως το φύλλο, την ηλικία, τις καταναλωτικές συνήθειες και άλλα, επιτείνεται η ομαδοποίηση της πελατείας ώστε να είναι πιο αποτελεσματική η εξυπηρέτησή τους.

- **Εξατομικευμένη διαχείριση πελατών**

Ιδιαίτερα στις περιπτώσεις των χρηματοοικονομικών υπηρεσιών, είναι απαραίτητη η εξατομικευμένη προσφορά προϊόντων καθώς υπάρχει μεγάλη απόκλιση στις ανάγκες του κοινού.

- **Marketing**

Ο κόσμος του Μάρκετινγ έχει αλλάξει ολικά ύστερα από την χρήση των δεδομένων με σκοπό την χάραξη νέων στρατηγικών και αποφάσεων. Σύμφωνα με το προαναφερθέν άρθρο της McKinsey, η ανάλυση των πελατειακών δεδομένων αποτελεί την πιο κοινή χρήση των Big Data Analytics.

- **Online Αξιολόγηση Δανειοληπτικής Ικανότητας Πελατών**

Η αξιολόγηση της πιστοληπτικής ικανότητας για τις τράπεζες ήταν ανέκαθεν μία χρονοβόρα και σημαντική εργασία, η οποία ήταν απαραίτητο να αυτοματοποιηθεί ώστε να ακολουθήσεις τις ανάγκες του ψηφιακού μετασχηματισμού. Συνεπώς, αρκετές από τις επιχειρήσεις οι οποίες ασχολούνται με την ψηφιακή δανειοδότηση, αναλύουν τα δεδομένα από το ιστορικό των συναλλαγών, των ηλεκτρονικών αγορών αλλά και τα δεδομένα από τα κοινωνικά δίκτυα ώστε να καταλήγουν σε αυτοματοποιημένα συστήματα αξιολόγησης της ποιότητας των πελατών.

- **Διαχείριση Κινδύνου**

Η διαχείριση κινδύνων καθίσταται πλέον ακόμα πιο ακριβής με την χρήση των Μεγάλων Δεδομένων, καθώς η ανάλυση πραγματοποιείται σε πραγματικό χρόνο. Αυτό έχει ως αποτέλεσμα ακόμα πιο ακριβείς προβλέψεις για τα συστήματα διαχείρισης κινδύνου. Κατά συνέπεια, οι επιχειρήσεις κατορθώνουν την άμεση γνώση των χαρακτηριστικών της πελατείας τους .

- **API (Application Programming Interface)**

Η ανάπτυξη της διαχείρισης των δεδομένων τα οποία προέρχονται από την χρήση των ηλεκτρονικών υπολογιστών δεν θα μπορούσε να είναι δυνατή αν δεν αναπτύσσονταν παράλληλα και τα πληροφοριακά συστήματα εκείνα τα οποία θα μπορούσαν να υποστηρίξουν την εξέλιξη αυτή.

Σημαντικό ρόλο στον ψηφιακό αυτό μετασχηματισμό παίζει η τεχνολογία του API (Application Programming Interface). Πρόκειται για ένα ενδιάμεσο λογισμικό το οποίο επιτρέπει τη διασύνδεση μεταξύ δύο εφαρμογών. Μέσω του λογισμικού αυτού δύναται να δημιουργηθούν νέες εφαρμογές, βασισμένες πάνω σε εξειδικευμένα προγράμματα. Πιο συγκεκριμένα, ένα API έχει τη δυνατότητα να μεταφέρει δεδομένα από ένα σύστημα σε ένα άλλο. Ένα χαρακτηριστικό παράδειγμα είναι η διαδικασία ηλεκτρονικής έκδοσης ενός αεροπορικού εισιτηρίου. Ο τελικός χρήστης δύναται να χρησιμοποιήσει μία ταξιδιωτική πλατφόρμα, μέσω της οποίας θα καταχωρήσει τα στοιχεία του και εν συνεχεία το σύστημα, χρησιμοποιώντας το API της αεροπορικής εταιρίας θα ψάξει στην Βάση Δεδομένων αυτής και θα ολοκληρώσει την κράτηση.

Στον χρηματοοικονομικό τομέα, οι αντίστοιχες επικοινωνίες εστιάζονται μεταξύ των τραπεζικών συστημάτων και των υπηρεσιών που προσφέρουν οι εταιρίες Fintech. Χαρακτηριστικό παράδειγμα είναι τα συστήματα τα οποία έχουν την δυνατότητα να συνδέουν όλους τους λογαριασμούς ενός καταναλωτή, από όλες τις τράπεζες, εντός μίας εφαρμογής.

Η εξέλιξη αυτή πραγματοποιείται σταδιακά, καθώς δεν είναι εύκολη η διαμόρφωση των Host τραπεζικών συστημάτων, τα οποία έχουν δημιουργηθεί πριν αρκετά χρόνια, ώστε να είναι ικανά να επικοινωνήσουν με νέες εφαρμογές. Επίσης, η διασύνδεση αυτή, πέραν από κοστοβόρα είναι αρκετά χρονοβόρα ως προς την τελική υλοποίησή της.

Στον αντίποδα, οι νέες ψηφιακές τράπεζες έχουν ένα συγκριτικό πλεονέκτημα σε αυτόν τον τομέα, καθώς τα συστήματά τους δημιουργούνται με βάση τις νέες τεχνολογίες και απαιτήσεις. Μία τέτοια υλοποίηση είναι και αυτή της τεχνολογίας blockchain, η οποία επιτρέπει την τήρηση όλων των συναλλαγών σε online βάσεις δεδομένων.

- **Τεχνητή Νοημοσύνη**

Ο όρος τεχνητή νοημοσύνη ο οποίος εισχωρεί στην επιχειρηματική καθημερινότητα όλο και πιο συχνά, καταγράφηκε για πρώτη φορά το 1955 από τον John McCarthy, ο οποίος ανέπτυξε την θεωρία ότι η ανθρώπινη νοημοσύνη δύναται να περιγράφει με ακριβή τρόπο μέσω μοντέλων από έναν υπολογιστή.

Ωστόσο, μέχρι της μέρες μας, ο ορισμός αυτός αναπροσαρμόζεται συνεχώς καθώς η ανάπτυξη του εν λόγω τομέα είναι διαρκής. Σε γενικές γραμμές, ο όρος εμπεριέχει όλες τις εργασίες τις οποίες εκτελεί ένας ηλεκτρονικός υπολογιστή και απαιτούν νοημοσύνη. Επίσης, στον ορισμό αυτό συμπεριλαμβάνεται η διεύρυνση της έξυπνης συμπεριφοράς για την επίλυση προβλημάτων και την δημιουργία έξυπνων υπολογιστικών συστημάτων. (Κοντιάδης, Παπαδημητρίου, Γεωργακοπούλου, & Στεφανίδης, 2018).

Στον τραπεζικό τομέα, η τεχνητή νοημοσύνη εισήλθε την δεκαετία του 1980 με τα συστήματα εμπειρογνωμόνων, ενώ τα τελευταία χρόνια εισχωρεί σε όλο και περισσότερες εργασίες ως κατωτέρω.

- **Anti-Money Laundry:** Η κατακόρυφη αύξηση των ηλεκτρονικών συναλλαγών, δημιούργησε την ανάγκη ανάπτυξης συστημάτων τα οποία θα είναι ικανά να διακρίνουν και να εντοπίζουν παράνομες συναλλαγές οι οποίες συνδέονται με το ξέπλυμα χρήματος. Αρκετές τράπεζες σε παγκόσμιο επίπεδο χρησιμοποιούν ήδη λογισμικό βασισμένο στην τεχνητή νοημοσύνη για τους σκοπούς αυτούς.

- **Recommendation Systems:** Πρόκειται για συστήματα τα οποία κάνουν χρήση των μεγάλων δεδομένων ώστε να καταλήγουν σε εξατομικευμένες προτάσεις προς τους πελάτες οι οποίες θα αφορούν σε επενδυτικές συμβουλές ή σε συμβουλές διαχείρισης χρέους.

- **Chat Box:** Τα αυτοματοποιημένα συστήματα συνομιλίας, έχουν ενταχθεί στις υπηρεσίες εξυπηρέτησης πελατών με σκοπό την ολοκλήρωση των σχετικών εργασιών δίχως ανθρώπινη παρέμβαση από πλευράς του προσωπικού της τράπεζας. Με αυτόν τον τρόπο εξοικονομούνται ανθρώπινοι πόροι αλλά παράλληλα συλλέγεται και πλήθος δεδομένων το οποίο αξιοποιείται στην αυτοτροφοδότηση και κατά συνέπεια βελτίωση των συστημάτων αυτών.

- **Fraud Detection:** Δεδομένης της ευαίσθητης φύσης των τραπεζικών υπηρεσιών, η καταπολέμηση της χρηματικής απάτης ήταν ανέκαθεν ένα μεγάλο ζητούμενο για τα χρηματοπιστωτικά ιδρύματα. Την ζήτηση αυτή ήρθε να καλύψει αποτελεσματικά η τεχνητή νοημοσύνη με την δημιουργία εξειδικευμένων συστημάτων.

- **The Internet of Things (IoT)**

Η ορολογία Internet of Things ακούστηκε την πρώτη φορά από τον Kevin Ashton την δεκαετία του 1990. Στις ημέρες μας γνωρίζει μεγάλη άνθηση καθώς ουσιαστικά πρόκειται για την χρήση έξυπνων μηχανημάτων σε όλες τις εκφάνσεις τις καθημερινής δραστηριότητας είτε πρόκειται για το σπίτι, το αυτοκίνητο ή την εργασία. Με πιο απλά λόγια, το IoT είναι η διασύνδεση όλων των ηλεκτρονικών συσκευών τόσο μεταξύ τους όσο και με τον παγκόσμιο ιστό, έτσι ώστε ο χρήστης να έχει τη δυνατότητα ελέγχου τους ακόμα και μέσα από το κινητό του. (Wikipedia, 2019)

- **Blockchain**

Ναυαρχίδα στην εισχώρηση των νέων τεχνολογιών στις χρηματοοικονομικές υπηρεσίες αποτελεί το blockchain. Πρόκειται για μια δημιουργία μίας ομάδας ανθρώπων με το ψευδώνυμο Satoshi Nakamoto. Το αρχικό πλάνο περιελάβανε αποκλειστικά την διαχείριση του κρυπτονομίσματος bitcoin, ωστόσο με το πέρασμα του χρόνου οι λειτουργίες του επεκτείνονται συνεχώς.

«Το blockchain είναι ουσιαστικά μία σειρά καταχωρήσεων που αφορούν συναλλαγές, σε ένα δημόσιο κατάστιχο (ledger). Κάθε καινούρια ομάδα καταχωρήσεων -ένα «block»- συνδέεται με τα προηγούμενα, δημιουργώντας μία «αλυσίδα» καταχωρίσεων, δηλαδή ένα «blockchain». Τα blocks αυτά συνδέονται μονοσήμαντα

μεταξύ τους. Προκύπτουν δε μέσα από μια διαδικασία που ονομάζουμε «proof of work», κατά την οποία επιτυγχάνεται η αλγοριθμική επίλυση ενός «δύσκολου» υπολογιστικού προβλήματος.

Κατ' αυτό τον τρόπο, το blockchain λειτουργεί ως ένα αποκεντρωμένο (decentralized) λογιστικό καθολικό, το οποίο είναι κοινό για όλους τους συμμετέχοντες, μιας και όλοι οι εμπλεκόμενοι αποθηκεύουν ένα αντίγραφό του· κάτι που εξασφαλίζει την ασφάλεια και την διαφάνεια των συναλλαγών. Η ειδοποιός διαφορά -αναφορικά με την προστασία- προκύπτει από το γεγονός ότι δεν είναι πλέον απαραίτητη η ύπαρξη μιας ενδιάμεσης «έμπιστης» αρχής (πχ. μιας τράπεζας - άρα οι συμβαλλόμενοι μπορούν να επιβεβαιώσουν μια συναλλαγή μεταξύ τους, και να θεωρηθεί έγκυρη, χωρίς να υπάρχει υποχρέωση αναμονής για την έγκριση από ένα κεντρικό χρήστη), ενώ η εμπιστοσύνη των συναλλασσόμενων μερών βασίζεται σε αλγοριθμική επιβεβαίωση». (Μάλλας, 2018)

3. Μεγάλα Δεδομένα και Μηχανική Μάθηση

3.1 Ο Ορισμός των Μεγάλων Δεδομένων

Ένας όρος ο οποίος είναι κομβικός πλέον για τις νέες ψηφιακές τεχνολογίες είναι τα "Μεγάλα Δεδομένα" (Big Data), ο οποίος έκανε για πρώτη φορά την εμφάνισή του το 1997 από ερευνητές της NASA. Η θέσπιση μίας νέας ορολογίας ήταν αναγκαία για την περιγραφή των συνόλων δεδομένων τα οποία λόγω του όγκου τους ήταν αδύνατον τόσο να περιγράφουν γραφικά όσο και να αποθηκευτούν σε τοπικούς δίσκους. Αρχικά το φαινόμενο αυτό ονομαζόταν πρόβλημα Μεγάλων Δεδομένων.

Από τότε έως σήμερα, τα μεγάλα δεδομένα εισχώρησαν στην τεχνολογική καθημερινότητα με αποτέλεσμα να αποτελέσουν έναν νέο αυτόνομο τομέα στον οποίο δραστηριοποιούνται ερευνητές και καινοτόμες επιχειρήσεις παγκοσμίως.

Αξιοσημείωτο είναι πως τα δεδομένα παράγονται πλέον με ρυθμό εκθετικής αύξησης δημιουργώντας μία τεχνολογική επανάσταση όσον αφορά την διαχείριση και αξιοποίησή τους. Πλέον, ο προσδιορισμούς ενός συνόλου δεδομένων ως Big Data κυμαίνεται από terabytes έως zetabytes.

3.2 Οι 5 Διαστάσεις των Μεγάλων Δεδομένων

Η γέννηση του τομέα των Μεγάλων Δεδομένων προκάλεσε τον ακόμα πιο λεπτομερή χαρακτηρισμό τους. Έτσι τα σύνολα δεδομένων πλέον χαρακτηρίζονται από την μεγάλη ποικιλία τους (Variety), τον αυξανόμενο όγκο (Volume), καθώς και την μεγάλη ταχύτητα παραγωγής (Velocity). Τα ανωτέρω συγκεντρώνονται στο μοντέλο των 3V's το οποίο εν συνεχεία και ύστερα από περεταίρω μελέτη μετεξελίχθηκε στο μοντέλο των 5 διαστάσεων όπως αυτό περιγράφεται κατωτέρω:

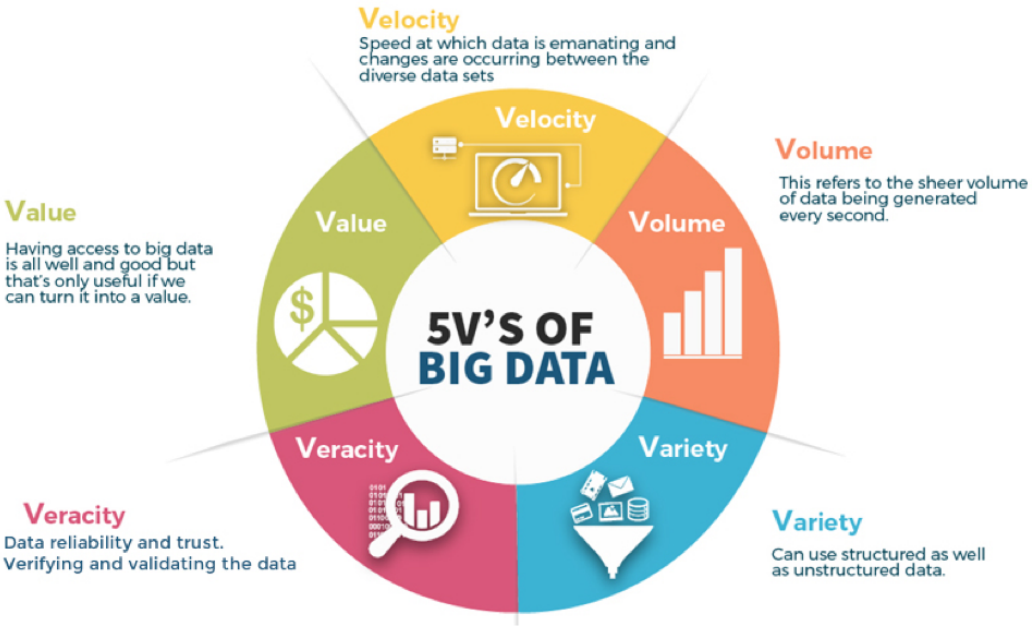
- **Volume**: Με τον όρο αυτό γίνεται αναφορά στις μεγάλες ποσότητες δεδομένων τις οποίες οι διάφοροι οργανισμοί καλούνται να διαχειριστούν και να τις αξιοποιήσουν ως εργαλεία λήψης αποφάσεων. Όπως είναι λογικό ο χαρακτηρισμό «μεγάλος όγκος» ποικίλει αναλόγως του υποκειμένου της ανάλυσης. Υπάρχουν επιχειρήσεις και οργανισμοί που διαχειρίζονται petabytes δεδομένων, ενώ σε άλλες περιπτώσεις τα δεδομένα μετριούνται σε επίπεδα zetabytes. Συνεπώς, ο καθορισμός των μεγεθών είναι δυναμικός, καθώς μέρα με την μέρα οι ανάγκες γίνονται συνεχώς μεγαλύτερες.

- **Variety**: Η ποικιλία έχει να κάνει με την διαχείριση της πολυπλοκότητας των διαφόρων τύπων δεδομένων. Σε αυτήν την διάσταση περιλαμβάνεται ο διαχωρισμός των πληροφοριών σε δομημένες, ημιδομημένες και αδόμητες. Οποιοσδήποτε διαχειρίζεται και αναλύει μεγάλα δεδομένα, χρησιμοποιεί πλήθος πηγών άντλησης, εντός η εκτός ενός οργανισμού. Δεδομένα μπορούν να αντληθούν ακόμα και από αισθητήρες εργαλείων IoT (Internet of Things) καθώς και από διαδίκτυο σε διάφορες μορφές όπως κείμενο, δεδομένα ιστού, αρχεία ήχου, βίντεο και διάφορα άλλα.

- **Velocity**: Πρόκειται για την ταχύτητα με την οποία τα δεδομένα δημιουργούνται, διακινούνται και επεξεργάζονται. Όπως είναι προφανές, η ταχύτητα αυτή αυξάνεται συνεχώς σε όλες τις εκφάνσεις της καθώς τα δεδομένα πλέον παράγονται σε πραγματικό χρόνο, συνεχώς και όλο το 24ωρο. Για αυτόν τον λόγο πλέον η διαχείρισή τους πραγματοποιείται μέσω ροών δεδομένων (data streaming). Βασικό στοιχείο του data streaming είναι ο «νεκρός χρόνος» (lag time) ο οποίος μεσολαβεί ανάμεσα στη λήψη των δεδομένων και στην αξιοποίησή τους. Συνεπώς, ο ρυθμός παραγωγής τους, καθιστά τα παραδοσιακά συστήματα άχρηστα ως προς την διαχείρισή τους, καθώς αρκετές επιχειρήσεις χρειάζονται την ανάλυση των πληροφοριών σε πραγματικό χρόνο ώστε να εκπληρώσουν τους σκοπούς τους.

- **Value**: Ο όρος αυτός σχετίζεται με την αξία, δηλαδή την χρησιμότητα των δεδομένων από τους αναλυτές. Κάποια δεδομένα στην αρχική τους μορφή ενδέχεται να είναι άχρηστα για τους αναλυτές, αλλά με την χρήση των σωστών εργαλείων και μεθόδων, δύναται να αξιοποιηθούν πλήρως από τους διάφορους οργανισμούς.

- **Veracity**: Πρόκειται για την εγκυρότητα και την αξιοπιστία που διαθέτουν τα δεδομένα. Είναι σημαντικό η πληροφορία να χαρακτηρίζεται από υψηλή ποιότητα ώστε να είναι εφικτή η παραγωγή ορθών αποτελεσμάτων. Για τον σκοπό αυτόν, έχουν αναπτυχθεί οι μέθοδοι «καθαρισμού δεδομένων» (data cleansing) οι οποίες συνεισφέρουν σε μία πιο αποδοτική και αξιόπιστη ανάλυση. (Schroeck, Shockley, Smart, Morales, & Tufano, 2012)



Εικόνα 1 5V's of Big Data (Πηγή: (www.techentice.com, 2020)

3.3 Οι Δομές των Μεγάλων Δεδομένων

Τα δεδομένα τα οποία καλούνται να διαχειριστούν οι διάφοροι οργανισμοί αποτελούνται από διάφορους τύπους και μορφές. Ωστόσο ως βασική αρχή κυριαρχεί η εξέταση των πληροφοριών για την εξεύρεση ανακολουθιών (inconsistencies), ελλείψεων (incompleteness), διπλοεγγραφών (duplication) και συγχωνευτικών προβλημάτων (merging). Η διαδικασία της επεξεργασίας αυτής ονομάζεται εκκαθάριση και κατά τη διάρκεια αυτής εφαρμόζονται διάφορες τεχνικές φιλτραρίσματος με απώτερο σκοπό την μείωση του μεγέθους των δεδομένων σε επίπεδα που είναι εφικτό να τα καταστήσουν επεξεργάσιμα.

Σύμφωνα με το αξίωμα (Garbage in Garbage out - GIGO), βάσει του οποίου τα αναξιόπιστα δεδομένα αποδίδουν αναξιόπιστα μοντέλα και κατά συνέπεια λανθασμένα αποτελέσματα. Ως εκ τούτου, το στάδιο της προεπεξεργασίας είναι θεμελιώδες για την οποιαδήποτε ανάλυση ακολουθήσει. Είναι χαρακτηριστικό ότι και το παραμικρό σφάλμα είναι ικανό να επηρεάσει το αποτέλεσμα της ανάλυσης άχρηστο. Συνεπώς, είναι σημαντικό ο αναλυτής να γνωρίζει εξ αρχής κάποιες βασικές πληροφορίες όπως τη μορφή, τις πηγές και τους τύπους των δεδομένων.

- **Δομημένα Δεδομένα (structured)**

Η έννοια των δομημένων δεδομένων τις περισσότερες φορές σχετίζεται με τα δεδομένα τα οποία προκύπτουν από συναλλαγές. Αυτό συμβαίνει διότι οι συναλλαγές αποτελούνται από πληροφορίες σε δομημένη μορφή οι οποίες σχετίζονται με τα στοιχεία του πελάτη και ποσοτικών μεγεθών των συναλλαγών. Κατά κύριο λόγο, αυτού του τύπου τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων επεξεργασίας

συναλλαγών (OLTP). Με αυτόν τον τρόπο, οι συναλλαγές δύναται να συγκεντρωθούν σε εναλλακτικές χρονικές βάσεις, αλλά και να εξαχθούν στατιστικά μοντέλα και μεγέθη.

- **Μη δομημένα (unstructured) δεδομένα**

Τα δεδομένα αυτού του τύπου συναντώνται συνήθως σε έγγραφα κειμένου όπως για παράδειγμα τα μηνύματα ηλεκτρονικού ταχυδρομείου, το διαδίκτυο, γενικά έγγραφα κτλ. Η κατηγορία αυτή γνωρίζει μεγάλη αύξηση δεδομένου ότι επηρεάζεται άμεσα από την ραγδαία αύξηση της χρήσης των πηγών τέτοιων δεδομένων όπως είναι τα διάφορα Social Media. Οι πηγές αυτές χρειάζονται λεπτομερή προεπεξεργασία προτού θεωρηθούν ικανά προς ανάλυση. (Baesens, 2014)

3.4. Τεχνικές Ανάλυσης Μεγάλων Δεδομένων

Όπως έχει αναφερθεί στα ανωτέρω κεφάλαια, η ραγδαία αύξηση του όγκου των δεδομένων αναπροσαρμόζει διαρκώς τις ανάγκες εξεύρεσης νέων λύσεων σχετικά με την αποθήκευση και την επεξεργασία αυτών. Με γνώμονα την ανάγκη εξόρυξης χρήσιμων πληροφοριών οι οποίες θα οδηγήσουν σε στρατηγικές αποφάσεις, δημιουργούνται συνεχώς νέες τεχνικές και εργαλεία ανάλυσης. Τα εργαλεία αυτά προκύπτουν από τις θεμελιώδεις αρχές επιστημών όπως τα μαθηματικά, η στατιστική, οικονομία και η πληροφορική. Σε μια προσπάθεια ομαδοποίησής τους, προκύπτουν οι κατωτέρω κατηγορίες.

- **Βελτιστοποίηση (optimization)**

Πρόκειται για εναλλακτικό τρόπο επίτευξης της υψηλότερης απόδοσης με αποτελεσματικό τρόπο και με το μικρότερο δυνατό υπολογιστικό κόστος. Η πρακτική βελτιστοποίησης περιορίζεται όταν υπάρχει έλλειψη πληροφόρησης και έλλειψη χρόνου για να εκτιμηθεί ποιες πληροφορίες μπορούν να αποκτηθούν.

- **Στατιστική**

Η στατιστική είναι η επιστήμη της συλλογής, της ανάλυσης και της κατανόησης των δεδομένων, υπολογίζοντας όλα τα σχετικά μεγέθη. Ως εκ τούτου, βρίσκει εφαρμογή σε πολλούς τομείς όπως τις κοινωνικές επιστήμες, την οικονομία, την ιατρική και τις επιχειρήσεις.

- **Εξόρυξη δεδομένων (data mining)**

Εξόρυξη δεδομένων είναι η εξεύρεση μιας πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

- **Τεχνικές οπτικοποίησης (visualization)**

Η οπτικοποίηση δεδομένων είναι η γραφική αναπαράσταση πληροφοριών και δεδομένων. Χρησιμοποιώντας οπτικά στοιχεία όπως γραφήματα και χάρτες, τα εργαλεία απεικόνισης δεδομένων παρέχουν έναν προσιτό τρόπο για να δείτε και να κατανοήσετε τις τάσεις, τα υπερβολικά υψηλά και τα πρότυπα στα δεδομένα.

- **Ανάλυση δικτύων (network analysis)**

Πρόκειται για τη μαθηματική ανάλυση σύνθετων διαδικασιών εργασίας από την άποψη ενός δικτύου συναφών δραστηριοτήτων. Αφορά επίσης τον υπολογισμό των ηλεκτρικών ρευμάτων που ρέουν στα διάφορα κανάλια ενός δικτύου.

- **Σημασιολογική ανάλυση (semantic analysis)**

Η διαδικασία της συσχέτισης των συντακτικών δομών, από τα επίπεδα των φράσεων, των προτάσεων και των παραγράφων στο επίπεδο της γραφής στο σύνολό της, στις γλωσσικά ανεξάρτητες έννοιές τους, αφαιρώντας χαρακτηριστικά ειδικά για συγκεκριμένα γλωσσικά και πολιτισμικά πλαίσια, στο βαθμό που είναι εφικτό.

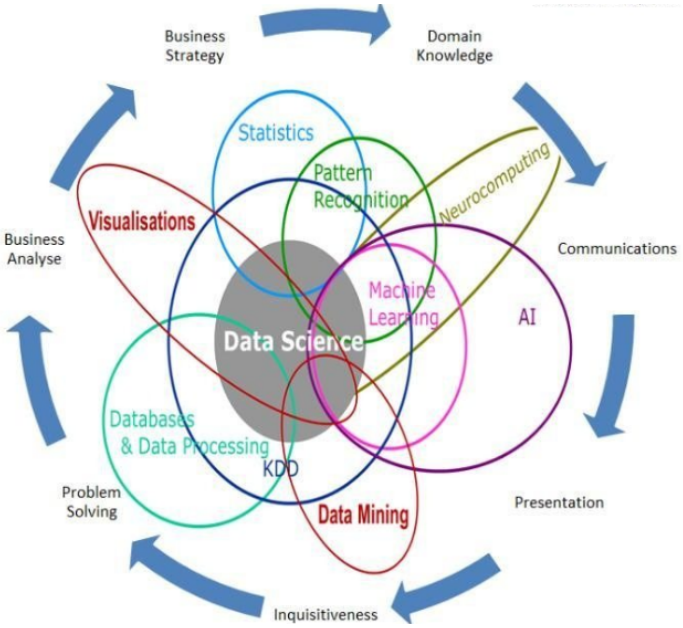
- **Πληθοπορισμός (crowdsourcing)**

Πρόκειται για μία μορφή συλλογικής διαδικτυακής δραστηριότητας, βάσει της οποίας ο οποιοσδήποτε φορέας, είτε σε ατομικό είτε σε συλλογικό επίπεδο, έχει τη δυνατότητα ανάθεσης εργασίας σε ένα σύνολο ατόμων, αποτελούμενο από διαφορετικές ειδικότητες και εξειδικεύσεις. Με αυτόν τον τρόπο πραγματοποιείται μία πολυεπίπεδη συνεργασία των διάφορων γνωστικών αντικειμένων με κοινό σκοπό την πλήρωση των απαιτήσεων της συγκεντρωτικής εργασίας.

Με αυτόν τον τρόπο όλοι οι εμπλεκόμενοι βρίσκονται κερδισμένοι, καθώς οι χρήστες οι οποίοι αναλαμβάνουν επιμέρους εργασίες αναπτύσσουν περεταίρω το γνωσιακό τους αντικείμενο και ενδεχομένως να λαμβάνουν και την αντίστοιχη οικονομική απολαβή ενώ ο αναθέτης της αρχικής εργασίας εξασφαλίζει ένα ποιοτικό αποτέλεσμα το οποίο προκύπτει από την σύνθεση των διαδοχικών εργασιών .

- **Τεχνητή Νοημοσύνη (artificial intelligence)**

Η τεχνητή νοημοσύνη δύναται πλέον να χαρακτηριστεί ως ένα αυτόνομο υποπεδίο της επιστήμης των υπολογιστών το οποίο εξελίχθηκε ύστερα από την μελέτη της δημιουργίας προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Ο τομέας αυτός έχει ως αποστολή την διερεύνηση της έρευνας και τη δημιουργία κατάλληλων μοντέλων αλγορίθμων τα οποία θα είναι ικανά να προβούν σε προβλέψεις έναντι των αρχικών δεδομένων. Οι αλγόριθμοι αυτοί διαθέτουν ως υπόβαθρο της αρχές άλλων παρεμφερών επιστημών όπως τα μαθηματικά και η στατιστική.



Εικόνα 2 Τεχνικές Ανάλυσης Μεγάλων Δεδομένων (Πηγή: (Djamal Abide, 2017)

3.5 Μηχανική Μάθηση

Όπως αναφέρθηκε ανωτέρω, ως Μηχανική μάθηση καλείται το πεδίο της έρευνας το οποίο επικεντρώνεται στην κατανόηση της λειτουργίας των διάφορων συστημάτων μάθησης. Το τομέας αυτός έχει ως βασικό αντικείμενο την κατασκευή προγραμμάτων – αλγορίθμων τα οποία έχουν την δυνατότητα να αυτοβελτιώνονται βάσει της ανατροφοδότησής τους από δεδομένα. Με αυτόν τον τρόπο γίνονται διαρκώς πιο αξιόπιστα, προσφέροντας εν τέλει μία έγκυρη πηγή άντλησης γνώσης. Πρόκειται για ένα επιστημονικό πεδίο το οποίο βρίσκεται υπό συνεχή εξέλιξη και ανάπτυξη καθώς δανείζεται συχνά γνώσεις από άλλους επιστημονικούς τομείς και τις συνδυάζει κατάλληλα ώστε να προκύπτουν επιστημονικά τεκμηριωμένα αποτελέσματα.

Ως βασικός κανόνας των αλγορίθμων μηχανικής μάθησης θεσπίζεται η προσπάθεια εξαγωγής γενικών κανόνων, βάσει των οποίων γίνεται εφικτή η εξεύρεση βέλτιστων λύσεων σε ένα πλήθος προβλημάτων. Αυτό πραγματοποιείται μέσω της χρήσης ενός συνόλου δεδομένων εκπαίδευσης και μίας συνάρτησης στόχου από τα οποία προκύπτουν οι πιθανές υποθέσεις λύσεων. Δεδομένου ότι ως μόνη διαθέσιμη πληροφορία υπάρχει αρχικά μόνο το πεπερασμένο σύνολο της εκπαίδευσης, ανακύπτει μία διαδικασία επαγωγικού συλλογισμού (inductive reasoning) βάσει της οποία προκύπτει μία γενίκευση της αρχικής υποθέσεως.

Ως εκ τούτου, ως θέσφατο θεωρείται η ακρίβεια της συνάρτησης στόχου πάνω στα δεδομένα εκπαίδευσης, καθώς για όλα τα άλλα δεδομένα δημιουργούνται όσο το δυνατόν ασφαλέστερες και τεκμηριωμένες υποθέσεις. Σε αυτό το σημείο

Τα πορίσματα δύναται να ληφθούν από διάφορα παραδείγματα τα οποία αναδεικνύουν τις συσχετίσεις μεταξύ των μελετώμενων μεταβλητών. Βασικό στοιχείο στην μελέτη και στην έρευνα της μηχανικής μάθησης είναι η δυνατότητα των μοντέλων να μαθαίνουν και να αυτοβελτιώνονται μέσω αυτοματοποιημένων διαδικασιών, ερμηνεύοντας περίπλοκα πρότυπα και δημιουργώντας έξυπνες αποφάσεις βασισμένα σε αυτά. (Witten & Frank, 2000)

Σχετικά με την σχεδίαση των διαφόρων συστημάτων μηχανικής μάθησης, σημειώνεται ότι ως δυνατότητα μάθησης προσδιορίζεται η ικανότητα απόκτησης επιπλέον γνώσης η οποία είναι ικανή να επιφέρει μεταβολές στην ήδη δεδομένη. Υπάρχουν περιπτώσεις όπως οι αλγόριθμοι τεχνητών νευρωνικών δικτύων όπου η μάθηση προσδιορίζεται βάσει της δυνατότητας μετασχηματισμού της δομής του ίδιου του αλγορίθμου και όχι της επιπλέον επεξεργασίας των ίδιων των δεδομένων. Με αυτόν τον τρόπο και εξομοιώνοντας τον τρόπο μάθησης του ανθρώπινου εγκεφάλου δημιουργούνται συνεχώς νέα μοντέλα τα οποία καθίστανται ικανά να εξελιχθούν σε εφαρμογές με μεγάλη επιτυχία και εμπορική εφαρμογή.

Μεταξύ των επιστημονικών κλάδων που επωφελούνται από τα επιτεύγματα στον τομέα της Μηχανικής Μάθησης συγκαταλέγονται οι: Εξόρυξη Δεδομένων, Πιθανότητες και Στατιστική, Θεωρία της Πληροφορίας, Αριθμητική Βελτιστοποίηση, Θεωρία της Πολυπλοκότητας, Θεωρία Ελέγχου (προσαρμοστική), Ψυχολογία (εξελικτική, γνωστική), Νευροβιολογία και Γλωσσολογία.

3.5.1 Στάδια Μηχανικής Μάθησης

Ο τομέας της Μηχανικής Μάθησης δύναται να διαχωριστεί στα 8 επιμέρους κατωτέρω στάδια εργασιών τα οποία εκτελούνται στις περισσότερες εφαρμογές μεθόδων.

1. **Συλλογή Δεδομένων**: Η διαδικασία κατά την οποία συλλέγονται τα δεδομένα από μία ή συνήθως περισσότερες πηγές.

2. **Προεπεξεργασία Δεδομένων:** Τα δεδομένα ως επί τω πλείστων συλλέγονται σε μορφή η οποία δεν είναι εφικτό να εφαρμοστούν σε κάποιο δομημένο μοντέλο μηχανικής μάθησης. Ως εκ τούτου, είναι απαραίτητη η επεξεργασία τους μέσω συγκεκριμένων εργαλείων και μεθόδων ώστε να καταλήξουν σε μορφή κατάλληλη για το εκάστοτε μοντέλο. Οι συνήθεις εργασίες οι οποίες εκτελούνται σε αυτό το στάδιο είναι η τακτοποίηση των δεδομένων σε μορφή διακριτών χαρακτηριστικών, η μετατροπή των μεταβλητών του χρόνου σε συνεπή συστήματα μονάδων, η κανονικοποίηση των δεδομένων, η μετατροπή κατηγορικών μεταβλητών σε ποσοτικές μεταβλητές καθώς και η διαχείριση των ελλιπών δεδομένων.

3. **Διερεύνηση των δεδομένων και εξαγωγή σημαντικών χαρακτηριστικών**: Σε αυτό το επίπεδο εντοπίζονται ή εναλλακτικά δημιουργούνται τεχνικώς, τα βασικά χαρακτηριστικά τα οποία χαρακτηρίζουν το σύνολο δεδομένων της ανάλυσης. Εν συνεχεία, πραγματοποιείται βαθμονόμηση των πιο σημαντικών χαρακτηριστικών. Δεδομένου του μοντέλου το οποίο προορίζεται να δεχθεί τα δεδομένα επιλέγονται τα κύρια χαρακτηριστικά τα οποία θα χρησιμοποιηθούν περεταίρω. Η σημασία του συγκεκριμένου σταδίου της ανάλυσης

αναδεικνύεται από το γεγονός ότι υπάρχει κλάδος της μηχανικής μάθησης ο οποίος ασχολείται αποκλειστικά με το feature engineering.

4. **Επιλογή Μοντέλου**: Ύστερα από την διερεύνηση του συνόλου δεδομένων και βάσει της φύσης της ανάλυσης γίνεται η επιλογή του μοντέλου μηχανικής μάθησης το οποίο θα χρησιμοποιηθεί. Στην επιλογή αυτή, καθοριστικό ρόλο παίζει η εμπειρία του αναλυτή και τα αποτελέσματα των δοκιμών που έχει υλοποιήσει. Συνήθως το τελικό μοντέλο δεν θα είναι το μοναδικό το οποίο θα μπορούσε να επιλεγεί, για τον λόγο αυτό τις περισσότερες φορές πραγματοποιείται σύγκριση διαφόρων μοντέλων ώστε να προκριθεί το πιο αποτελεσματικό. Πιθανός είναι επίσης και ο συνδυασμός δύο ή περισσότερων μοντέλων.

5. **Εκπαίδευση Μοντέλου**: Για την εκπαίδευση του μοντέλου το οποίο επιλέχτηκε στο προηγούμενο στάδιο, πραγματοποιείται διαχωρισμός του αρχικού συνόλου δεδομένων σε train set και test set. Το πρώτο υποσύνολο εφαρμόζεται στο αλγοριθμικό μοντέλο έτσι ώστε το μοντέλο να εκπαιδευτεί και να προσαρμόσει κατάλληλα τις δυναμικές του παραμέτρους. Η διαδικασία αυτή πραγματοποιείται μέσω της ελαχιστοποίησης κάποιας συνάρτησης σφάλματος (mse, mae, cost functions, loss functions).

6. **Αξιολόγηση του Μοντέλου**: Για την αξιολόγηση του μοντέλου χρησιμοποιούνται εξειδικευμένες μετρικές προκειμένου να διαπιστωθεί η επάρκεια της ανάλυσης. Σημειώνεται ότι στην περίπτωση αλγορίθμων της κατηγορίας της μη επιτηρούμενης μάθησης, η αξιολόγηση είναι τις περισσότερες φορές εμπειρική καθώς δεν είναι δυνατόν να υπάρξουν οι κατάλληλες μετρικές ώστε εκτιμηθεί απόλυτα η απόδοση και το σφάλμα τους. Ωστόσο, στα μοντέλα συσταδοποίησης υπάρχουν μέτρα ομοιότητας, ενδοσυσταδικής και διασυσταδικής απόστασης τα οποία έχουν την ικανότητα να εκτιμούν αποτελεσματικά την ακρίβεια του μοντέλου. Στην κατηγορία των αλγορίθμων της επιτηρούμενης μάθησης η αξιολόγηση πραγματοποιείται με μετρικές όπως το μέσο τετραγωνικό σφάλμα, το μέσο απόλυτο για τις περιπτώσεις παλινδρόμησης, ο πίνακας σύγχυσης (confusion matrix), τα precision – recall καθώς και η καμπύλη ROC για τα μοντέλα ταξινόμησης.

7. **Βελτιστοποίηση Μοντέλου**: Κατόπιν της αξιολογήσεως του μοντέλου προκύπτουν περιθώρια βελτίωσης των παραμέτρων ώστε τα αποτελέσματα του μοντέλου να βελτιωθούν ακόμα περισσότερο. Οι παράμετροι αυτοί (hyperparameters) αρχικώς καθορίζονται διαισθητικά από τον αναλυτή και δεν μεταβάλλονται κατά την διαδικασία της εκπαίδευσης. Ωστόσο επηρεάζουν την μάθηση του αλγορίθμου συνεπώς χρήζουν παραμετροποίησης.

8. **Πρόβλεψη Μοντέλου**: Απώτερος σκοπός μίας ανάλυσης με χρήση μοντέλου επιτηρούμενης μάθησης είναι οι προβλέψεις μεταβλητών στόχων πάνω σε νέα δεδομένα. Πρόκειται στην ουσία για το στάδιο της παραγωγικής εφαρμογής ενός μοντέλου.

3.5.2 Πρακτικές Μηχανικής Μάθησης

Οι πιο διαδεδομένες διαδικασίες μάθησης που χρησιμοποιούνται στην μηχανική μάθηση είναι οι εξής:

- Επιβλεπόμενη μάθηση (Supervised learning)
- Μη-επιβλεπόμενη μάθηση (Unsupervised learning)
- Ημι-επιβλεπόμενη μάθηση
- Ενισχυτική μάθηση (Reinforcement learning)

3.5.3 Μάθηση με Επίβλεψη

Ως επιβλεπόμενη μάθηση ορίζεται η υποκατηγορία της μηχανικής μάθησης όπου η διαδικασία στηρίζεται σε ζεύγη εισόδου και εξόδου. Το σύνολο δεδομένων αποτελείται από μεταβλητές τα οποία χαρακτηρίζονται ως χαρακτηριστικά εισόδου (features) και από μία ή περισσότερες ετικέτες (labels) οι οποίες αποτελούν τις εξαρτημένες μεταβλητές.

Στην περίπτωση της επιβλεπόμενης μάθησης το σύνολο δεδομένων διαχωρίζεται στις εξής κατηγορίες:

- **Training Set**: Πρόκειται για το υποσύνολο του αρχικού dataset το οποίο τροφοδοτεί το μοντέλο μηχανικής μάθησης προκειμένου να εκπαιδευτούν οι εσωτερικές του παράμετροι.

- **Validation Set**: Το συγκεκριμένο υποσύνολο χρησιμοποιείται κατά την διαδικασία αξιολόγησης του μοντέλου με σκοπό την βελτιστοποίηση των παραμέτρων του αλγορίθμου.

- **Test Set**: Χρησιμοποιείται για την εφαρμογή και επαναξιολόγηση του τελικού μοντέλου, πάνω σε νέα δεδομένα. Μέσω του test set εφαρμόζονται στο μοντέλο νέες είσοδοι για τις οποίες ωστόσο είναι γνωστά τα αποτελέσματα. Με αυτόν τον τρόπο πραγματοποιείται ένας επιπλέον έλεγχος σχετικά με την απόδοση του μοντέλου. Λόγω της συνάφειας του σκοπού των validation και test sets, σε πολλές αναλύσεις αυτά τα δύο συγχέονται. Πιο συγκεκριμένα, σε αυτές τις περιπτώσεις, προσαρμόζονται οι παράμετροι του μοντέλου έτσι ώστε να υπάρξουν θετικά αποτελέσματα στο test set. Εν συνεχεία, ακολουθούνται οι διαδικασίες επικύρωσης όπως για παράδειγμα το cross – validation.

Ύστερα από την ανωτέρω διάκριση του συνόλου δεδομένων πραγματοποιείται η επιλογή του αλγορίθμου και η διαδικασία της επιβλεπόμενης μάθησης συνεχίζεται ως εξής.

Έστω ένα dataset A το οποίο χωρίζεται με βάση την παραπάνω λογική σε training set και test sets A1, A2,… Το A1 αποτελείται από i γραμμές με πεπερασμένες τιμές ανεξάρτητων μεταβλητών X = (x1…xk) καθώς και τις αντίστοιχες τιμές yi της εξαρτημένης μεταβλητής y. Θεωρείται ότι οι X,y συνδέονται μέσω μία συνάρτησης στόχου f για την οποία γνωρίζουμε μόνο τις τιμές που έχουν δοθεί στα y:

$$y = f\ (x_1, \dots x_k) = f(x)$$

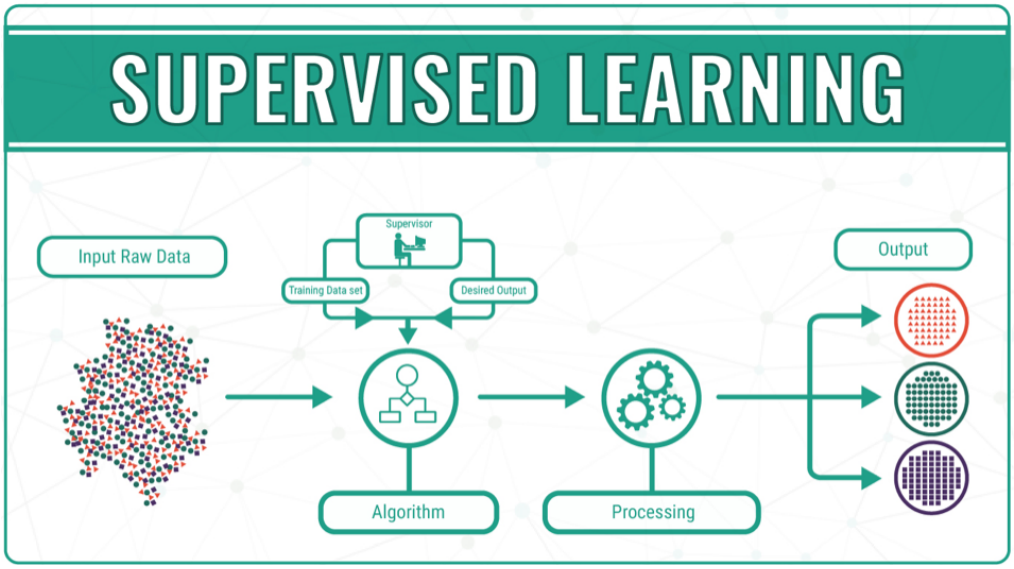
Ως στόχος συνεπώς είναι η δημιουργία μίας γενικευμένης εκτίμησης h της f

$$\hat{y} = h(x_1, \dots x_k) = h(x)$$

Η ανωτέρω σκοπεύει στην ελαχιστοποίηση μιας συνάρτησης σφάλματος

$$E(h) = \sum_x error\big(h(x), f(x)\big), x\ \in A1$$

Η συνάρτηση σφάλματος της οποία γίνεται χρήση κατά την εκπαίδευση εξαρτάται από το είδος και την φύση του εξεταζόμενου προβλήματος. Το μοντέλο h αξιολογείται με τις προαναφερόμενες μεθόδους.

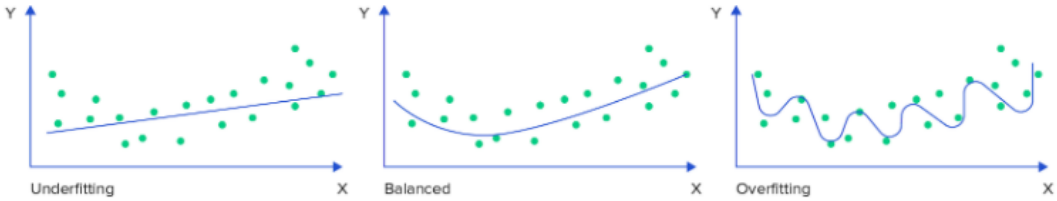


Εικόνα 3 Επιβλεπόμενη Μάθηση (Πηγή: (CHI, 2019))

Ωστόσο, η επιβλεπόμενη μάθηση παρουσιάζει την εξής ιδιαιτερότητα η οποία πρέπει σε κάθε ανάλυση να λαμβάνεται υπόψη και να αντιμετωπίζεται. Πρόκειται για την ισορροπία που αναζητείται μεταξύ της δομής του αλγορίθμου και της φύσης των δεδομένων της ανάλυσης.

Συνεπώς, όταν προκύπτουν καλές αποδόσεις στο training set σε συνδυασμό με κακά αποτελέσματα στο test set, τότε προκύπτει overfitting. Δηλαδή, το μοντέλο εμφανίζει υψηλή μεταβλητότητα (variance) με αποτέλεσμα να προσαρμόζεται υπερβολικά στο σύνολο δεδομένων εκπαίδευσης. Αυτό οδηγεί στην αδυναμία παραγωγής αξιόπιστων αποτελεσμάτων κατά την πρακτική εφαρμογή του μοντέλου. Αυτή η ιδιαιτερότητα προκύπτει συχνά σε πολυπαραμετρικά μοντέλα τα οποία εστιάζουν σε λεπτομέρειες των δεδομένων.

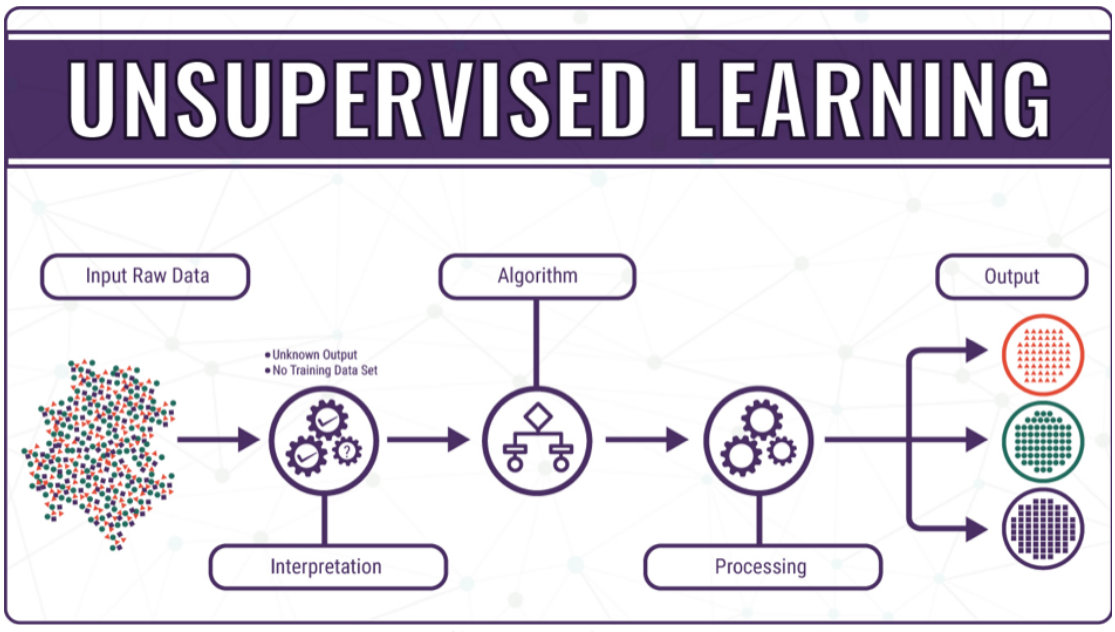
Στην αντίθετη περίπτωση, όταν δηλαδή υπάρχουν κακές αποδόσεις στο training set, τότε προκύπτει underfitting. Στην ουσία αυτό σημαίνει ότι το μοντέλο έχει απλουστευτεί αρκετά με αποτέλεσμα να μην έχει την δυνατότητα να προσαρμόζεται πάνω στα δεδομένα. Πιθανή συνθήκη εμφάνισης underfitting είναι η προσπάθεια πρόβλεψης μη γραμμικών δεδομένων μέσω ενός απλού μοντέλου γραμμικής παλινδρόμησης.



Εικόνα 4 Bias variance tradeoff (Πηγή: (kaggle, 2020))

3.5.4 Μη Επιβλεπόμενη Μάθηση

Ως μη επιβλεπόμενη μάθηση ορίζεται η διαδικασία της μηχανικής μάθησης βάσει της οποίας πραγματοποιείται η δημιουργία μίας συνάρτησης ικανής να περιγράψει επιτυχώς την κρυφή δομή πεπερασμένων αλλά άγνωστων δεδομένων. Αυτό προϋποθέτει την μη επισήμανση του συνόλου των δεδομένων. Από τις πιο κοινές εφαρμογές της μη επιβλεπόμενης μάθησης είναι η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας σε κατανομές αλλά μπορεί να περιλαμβάνει και άλλα προβλήματα όπως η επεξήγηση των βασικών χαρακτηριστικών των δεδομένων. Χαρακτηριστικές εργασίες μη επιτηρούμενης μάθησης αποτελούν η συσταδοποίηση (clustering) και η πλειονότητα των τεχνικών dimensionality reduction όπως η PCA (principal component analysis).



Εικόνα 5 Μη Επιβλεπόμενη Μάθηση (Πηγή: (CHI, 2019))

3.5.5 Ημιεπιβλεπόμενη Μάθηση

Σε αυτήν την κατηγορία μηχανικής μάθησης εντάσσονται προβλήματα στα οποία περιλαμβάνονται δεδομένα τα οποία έχουν μερικώς σημανθέντες ετικέτες εξόδου. Οι κατηγορίες των μοντέλων που απαρτίζουν αυτήν την οικογένεια μοντέλων είναι τα εξής:

- **Generative Μοντέλα**: Τα μοντέλα αυτά βασίζονται κατά κύριο λόγο στην από κοινού συνάρτηση πιθανότητας της εξαρτημένης και ανεξάρτητης μεταβλητής. Μία προσέγγιση των εν λόγω μοντέλων θα μπορούσε να είναι μία μορφή συσταδοποίησης με περισσότερες πληροφορίες ή ταξινόμησης με πληροφορίες οριακής πυκνότητας πιθανοτήτων.

- **Μέθοδοι διαχωρισμού χαμηλής πυκνότητας**: Μοντέλα που ακολουθούν αυτήν την μέθοδο είναι το TSVM (transductive support vector machine), η ταξινόμηση με γκαουσσιανή διαδικασία και προσεγγίσεις μεγιστοποίησης εντροπίας.

- **Μέθοδοι γράφων**: Σε αυτές τις περιπτώσεις τα δεδομένα εμφανίζονται μέσα σε κόμβους γράφων ενώ οι ακμές αυτών είναι σεσημασμένες με βάρη. Μέσω αυτού του δικτύου είναι δυνατή η διάδοση ετικετών από τα σεσημασμένα δεδομένα στα μη μέσω της χρήσης διακριτών μαρκοβιανών πεδίων, τυχαίων γκαουσσιανών πεδίων ή βαθιών συνελικτικών δικτύων.

- **Μέθοδοι 2 βημάτων**: Σε αυτήν την περίπτωση αρχικά πραγματοποιείται μία συσταδοποίηση στο σύνολο δεδομένων και εν συνεχεία μία ταξινόμηση στα σεσημασμένα δεδομένα. Οι μέθοδοι αυτές έρχονται σε συνάφεια με αυτές των γράφων.

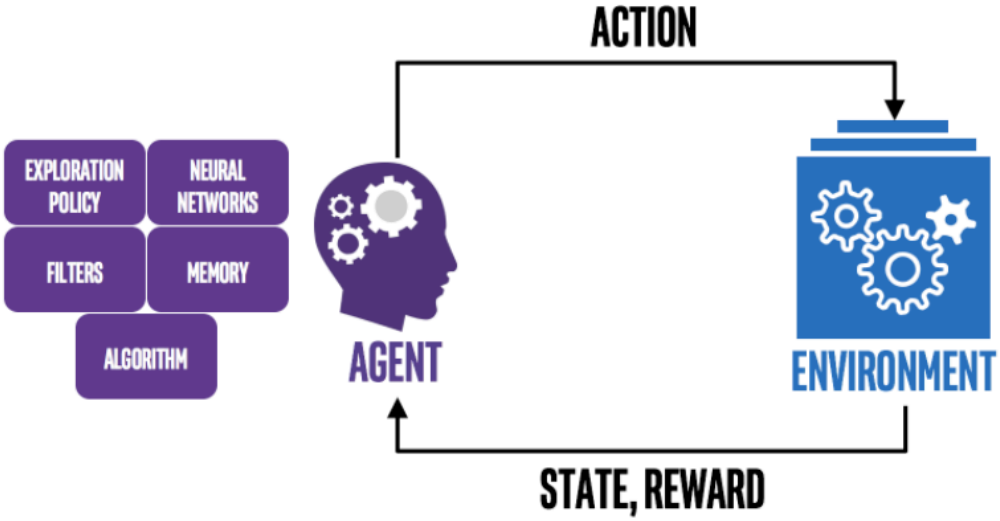
3.5.6 Ενισχυτική Μάθηση

Βάσει της ενισχυτικής μάθησης, οι σχετικοί αλγόριθμοι επιβραβεύονται για τις ορθές τους αποφάσεις και τιμωρούνται για τις λανθασμένες. Έτσι δημιουργείται ένας προσδιορισμός ενός ιδιότυπου σκορ βάσει του οποίου τα μοντέλα εκπαιδεύονται και αυτοβελτιώνονται. Χαρακτηριστικό παράδειγμα προσομοίωσης προβλημάτων που είναι ικανά να αντιμετωπίσουν αυτού του τύπου οι αλγόριθμοι είναι τα προβλήματα περιπλάνησης σε περίπλοκους χώρους, όπως έναν λαβύρινθο.

Σε γενικές γραμμές, αποτελεί μία μορφή μάθησης η οποία περιγράφεται ως εξής: Ένας πράκτορας τοποθετείται σε ένα περιβάλλον και πρέπει να μάθει να συμπεριφέρεται με επιτυχία, μέσα σε αυτό. Παρότι αποτελεί μια πολύ ενδιαφέρουσα μορφή μάθησης, το πεδίο μελέτης της είναι εκτενές και συνεπώς δεν θα επεκταθεί σε αυτή την εργασία, στην οποία και δεν χρησιμοποιείται. (Russel, Norvig, & Davis, 2003)

Σημαντικό παράδειγμα ενισχυτικής μάθησης αποτελεί ένα ρομπότ το οποίο συλλέγει απορρίμματα και καλείται να διακρίνει εάν δύναται να συνεχίζει να καθαρίζει ή πρέπει να μεταβεί στον χώρο φόρτισής του. Η απόφαση αυτή προκύπτει ύστερα από την

στάθμη της μπαταρίας του και βάσει του πόσο γρήγορα έχει εξεύρει σταθμό φόρτισης κατά το παρελθόν δεδομένης της τοποθεσίας του. Σε περίπτωση μηδενισμού της φόρτισης της μπαταρίας τότε προκύπτει ποινή στο σύστημα επιβράβευσης του πράκτορα. (Chapelle , Scholkopf, & Zien, 2006)



Εικόνα 6 Παράδειγμα ενισχυτικής μάθησης (Πηγή: (Lee, 2019))

4. Ο Ρόλος και η Εφαρμογή του Δείκτη Customer Lifetime Value Κατά τη Λήψη Επιχειρησιακών Αποφάσεων

Η εξέλιξη του παγκόσμιου επιχειρείν, δημιούργησε αρκετά νέα ερωτήματα αναφορικά με την βελτιστοποίηση της απόδοσης των επιχειρήσεων. Ένα από τα κυριότερα, τέθηκε προς διαβούλευση στις αρχές της δεκαετίας του 1990 και δημιουργούσε το ερώτημα εάν είναι προτιμότερη η αύξηση των εσόδων ή η μείωση του κόστους διοίκησης.

Το δίλλημα αυτό δημιουργήθηκε διότι επί τω πλείστων όποια επιχείρηση επέλεγε να επικεντρωθεί σε μία από τις δύο τακτικές, υστερούσε σημαντικά στην άλλη. Ως εκ τούτου, θεωρήθηκε ιδανική η εξεύρεση μίας μεθόδου όπου θα μπορούσε να συνδυάζει κατά βέλτιστο τρόπο και τις δύο τακτικές. Με βάση αυτό το σκεπτικό, καλλιεργήθηκε η προσέγγιση της ανάπτυξης βασισμένης στην αγορά με παρακολούθηση της κερδοφορίας και με παράλληλη επιστροφή στην επένδυση του τομέα του marketing.

Με αυτόν τον τρόπον ουσιαστικά γεννήθηκε η έννοια του πελατοκεντρικού marketing βάσει της οποίας ο πελάτης είναι ο κεντρικός άξονας της επιχειρηματικής στρατηγικής. Κατά συνέπεια, η διαχείριση των πελατών αποτέλεσε καίριο ζήτημα για όλες τις επιχειρήσεις. Η χάραξη ωστόσο μίας ορθής στρατηγικής απαιτεί τον εξισορροπισμό των πόρων που δαπανούνται ανά τομέα έτσι ώστε να προκύψει διαχρονικότητα στην αξία που αποδίδει ο εκάστοτε πελάτης. Κάτι τέτοιο φυσικά συνεπάγεται την αποτελεσματική αξιολόγηση της αξίας των πελατών μέσα από σωστές και έγκυρες μετρήσεις.

Ως αξία του πελάτη ορίζεται η συνεισφορά του σε μια επιχείρηση καθ' όλη τη διάρκεια της σχέσης του με αυτήν. Σύμφωνα με τους Reinartz & Kumar, οι παρελθοντικές συνεισφορές ενός πελάτη δεν δύναται να αντανακλούν τη μελλοντική του αξία στην επιχείρηση. Η αξία διάρκειας ζωής ενός καταναλωτή αφορά την πρόβλεψη για το καθαρό κέρδος το οποίο προκύπτει από την μελλοντική συνεισφορά του πελάτη προς την εταιρία. Η πολυπλοκότητα του μοντέλου καθώς και τα επίπεδα ακρίβειας ποικίλουν αναλόγως την μέθοδο και τις τεχνικές της ανάλυσης. Επιπρόσθετα, η αξία διάρκειας ζωής του πελάτη θεωρείται καίριο κομμάτι για τη διαδικασία λήψης αποφάσεων καθώς δίνει το περιθώριο στις επιχειρήσεις να μετατοπίσουν τους στόχους τους από τα άμεσα κέρδη στην μακροπρόθεσμη σχέση με τους πελάτες. Ως βασικό χαρακτηριστικό αυτής της σχέσης ορίζεται η συνολική οικονομική εισφορά (έσοδα μείον κόστος) του καταναλωτή πέραν της συνολικής διάρκειας ζωής του για την εκάστοτε εταιρία. (Kumar V. , 2006)

Πάνω σε αυτήν την λογική δημιουργήθηκε η ανάγκη για τον προσδιορισμό ενός αντικειμενικού μέτρου της μελλοντικής κερδοφορίας ενός οργανισμού από έναν πελάτη. Η σημαντικότητα του μέτρου αντανακλάται στο γεγονός ότι υποδεικνύει ένα ανώτατο όριο για τις δαπάνες προς την διεύρυνση του πελατολογίου. Ο καθορισμός του μέτρου αυτού έχει ιδιαίτερη βαρύτητα κατά τη διαδικασία λήψης αποφάσεων και χάραξης νέων στρατηγικών. (S. Gupta, 2006)

Βάσει των ανωτέρω προκύπτουν οι εξής τελικοί ορισμοί της Αξίας Διάρκειας Ζωής πελάτη:

- Η Αξία διάρκειας ζωής (Lifetime Value, LTV) του πελάτη, συνήθως ορίζεται ως το συνολικό καθαρό εισόδημα που μια επιχείρηση μπορεί να αναμένει από έναν πελάτη (Novo, 2001)

- Η Αξία Διάρκειας Ζωής του πελάτη (Customer Lifetime Value (CLV)) καθορίζεται ως το άθροισμα των συσσωρευμένων ροών μετρητών – μειούμενο χρησιμοποιώντας το Σταθμισμένο Μέσο Όρο Κόστους του Κεφαλαίου (Weighted Average Cost of Capital (WACC))- ενός πελάτη που υπερβαίνει τη συνολική διάρκεια ζωής του στην εταιρεία. (Kumar V. , A Customer Lifetime Value-Based Approach to Marketing in the Multichannel, Multimedia Retailing Environment, 2010)

Οι ανωτέρω θεωρήσεις υιοθετήθηκαν από πλήθος εταιριών και πλέον θεωρείται επιτακτική η ανάγκη χρήσης του συγκεκριμένου μέτρου ειδικά από εταιρίες με περιορισμένους πόρους οι οποίοι πρέπει να διαμοιράζονται με ορθολογικό τρόπο.

4.1 Υπολογισμός Customer Lifetime Value (CLV)

Η αξία διάρκειας ζωής ενός πελάτη δύναται να υπολογιστεί είτε εξατομικευμένα είτε επί του συνόλου ως ο μέσος όρος τον ατομικών υπολογισμών. Για τους σκοπούς της παρούσας εργασίας θα παρουσιαστούν κατωτέρω οι προσεγγίσεις οι οποίες προσδιορίζουν σφαιρικά τον δείκτη για το σύνολο ενός οργανισμού με απώτερο σκοπό την μοντελοποίηση των μεγεθών και την δημιουργία μιας συγκεντρωτικής μεθόδου υπολογισμού του μέτρου.

- **Τυπική Προσέγγιση**:

Στην συγκεκριμένη προσέγγιση, το άθροισμα των αξιών της διάρκειας ζωής όλων των πελατών, ονομάζεται Ίδιο Κεφάλαιο Πελάτη (Customer Equity (CE)) μιας εταιρείας και υπολογίζεται ως εξής:

$$CE = \sum_{i=1}^{I} \sum_{t=1}^{T} CM_{it} \left(\frac{1}{1+\delta}\right)^{t}$$

Όπου,
CE = ίδιο κεφάλαιο πελάτη (άθροισμα των εξατομικευμένων αξιών διάρκειας ζωής)
CM =Περιθώριο συνεισφοράς σε χρονική περίοδο t
δ= προεξοφλητικό επιτόκιο
i= δείκτης πελατών
t= χρονική περίοδος
T= ο αριθμός των χρονικών περιόδων για τις οποίες το CE έχει εκτιμηθεί.

Στον ανωτέρω υπολογισμό, το CE υποδεικνύει την οικονομική αξία μίας επιχείρησης και δύναται να υπολογιστεί ο μέσος όρος του CLV διαιρώντας το CE με τον αριθμό των πελατών.

- **2η Προσέγγιση**:

Βάσει του εν λόγω υπολογισμού, ο μέσος όρος της CLV ενός πελάτη υπολογίζεται από την Αξία Διάρκειας Ζωής ενός όλου ή τμήματος πελατών.

$$CLV = \sum_{t=0}^{T} \left[\frac{(GC - M)}{(1+d)^{t}} r^{t}\right] - A$$

Όπου,
r= ποσοστό διατήρησης
d= προεξοφλητικό επιτόκιο ή κόστος κεφαλαίου για την εταιρεία
t= χρονική περίοδος
T= ο αριθμός των χρονικών περιόδων που ελήφθησαν υπόψη για την κατ' εκτίμηση CE
GC= ο μέσος όρος μικρής εισφοράς
M= το κόστος marketing ανά πελάτη
A= ο μέσος όρος κόστους κτήσης ανά πελάτη

Ο ανωτέρω υπολογισμός συνυπολογίζει μόνο τον μέσο όρο μικρής εισφοράς (GC), τον μέσο όρο κόστους κτήσης ανά πελάτη (A), και το κόστος marketing (M) ανά πελάτη. Το ποσοστό διατήρησης r, (διατήρησης πελάτη), είναι ο μέσος όρος του ποσοστού διατήρησης για ένα όλον (σώμα) και λαμβάνεται ως σταθερή για μία περίοδο. (Kumar & Ramani , Customer lifetime value approaches and best practice applications, 2004)

Αξίζει να σημειωθεί ωστόσο ότι κάτι τέτοιο δεν υλοποιείται σε πραγματικές συνθήκες αγοράς καθώς οι πελάτες συνήθως διακόπτουν την σχέση με μία επιχείρηση σε διαφορετικά χρονικά σημεία. Ως εκ τούτου, οι πιθανότητες διατήρησης ποικίλουν ανά περίπτωση.

- **3η Προσέγγιση**:

Στην προσέγγιση που ακολουθεί, το ιδίο κεφάλαιο των πελατών της επιχείρησης εκφράζεται ως το άθροισμα της απόδοσης κατά την εξαγορά, της απόδοσης για διατήρηση και της απόδοσης για τις επιπρόσθετες πωλήσεις. Η μαθηματική γραφή του μοντέλου είναι η εξής.

$$CE(t) = \sum_{i=0}^{I}\left[N_{i,t}a_{i,t}(S_{i,t} - c_{i,t}) - N_{i,t}B_{i,a,t} + \sum_{k=1}^{\infty} N_{i,t}a_{i,t}\left(\prod_{j=1}^{k}\rho_{j,i+k}\right)(S_{i,t,k} - c_{i,t+k} - B_{i,r,t+k} - B_{i,AO,t+k})\left(\frac{1}{1+d}\right)^{K}\right]$$

Όπου,
CE(t) = η αξία των ίδιων κεφαλαίων πελάτη για πελάτες που αποκτήθηκαν σε χρόνο t
N i,t = ο αριθμός των πιθανών πελατών την στιγμή t για τμήμα i
α i,t = η πιθανότητα απόκτησης την στιγμή t για τμήμα i
ρ i,t = η πιθανότητα διατήρησης την στιγμή t για έναν πελάτη για τμήμα i
B i,a,t = το κόστος marketing ανά προσδοκία (prospect) (N) για απόκτηση πελατών την στιγμή t για τμήμα i
Bi,r,t = το marketing σε χρονική περίοδο t για διατηρούμενους πελάτες για τμήμα i
B i,AO,t = το κόστος marketing την χρονική περίοδο t για πωλήσεις add-on (επιπρόσθετες) για τμήμα i
d =προεξοφλητικό επιτόκιο S i,t = πωλήσεις του προϊόντος / υπηρεσιών που προσφέρονται από την εταιρεία την στιγμή t για τμήμα i
c i,t = κόστος αγαθών την στιγμή t για τμήμα i
I = ο αριθμός των τμημάτων
I = ο χαρακτηρισμός τμήματος
t 0 = η αρχική χρονική περίοδος

Πρόκειται για μία από τις πιο σημαντικές εφαρμογές του μέσου όρου της CLV και χρησιμοποιείται κατά κύριο λόγο για την αξιολόγηση του ανταγωνισμού. Η δυνατότητα αυτή προκύπτει από την μέτρηση του πόσο κερδοφόροι ή μη κερδοφόροι είναι οι ανταγωνιστές μίας επιχείρησης. (Blattberg, Getz, & Thomas, 2001)

Αν και η ανωτέρω προσέγγιση έχει αυξημένη εφαρμογή, ο μέσος όρος του CLV έχει περιορισμένη χρήση στις περιπτώσεις προσδιορισμού των πόρων μεταξύ των πελατών. Αυτό οφείλεται στο γεγονός ότι δεν χρησιμοποιεί μεταβλητές σε επίπεδο πελατών. Ως εκ τούτου είναι αναγκαίο να υπολογίζεται το CLV των εξατομικευμένων πελατών με απώτερο σκοπό την λήψη αποφάσεων και την χάραξη νέων στρατηγικών.

- **Προσέγγιση CLV (Εξατομικευμένου Επιπέδου)**

Στις περιπτώσεις του εξατομικευμένου υπολογισμού του CLV πελάτη, υπολογίζεται το άθροισμα των ροών των μετρητών μέσω της χρήσης του σταθμισμένου μέσου κόστους Κεφαλαίου (Weighted Average Cost of Capital (WACC)) ανά καταναλωτή. Ο εν λόγω υπολογισμός γίνεται για όλη τη διάρκεια όπου ο πελάτης παραμένει ενεργός στην επιχείρηση. Η κατωτέρω εξίσωση εκφράζει το περιθώριο συνεισφοράς ενός πελάτη, την ροπή που έχει αυτός να συνεχίσει να συνεισφέρει, καθώς και τη σχέση με τους πόρους marketing που δαπανώνται.

$$CLV = \sum_{t=1}^{T} \frac{\big((Future\ contribuon\ margin - Future\ Cost)\big)}{(1+d)^{\,t}}$$

Όπου,
t= δείκτης χρόνου
T= ο αριθμός των χρονικών περιόδων για τις οποίες το CLV έχει εκτιμηθεί.
d= προεξοφλητικό επιτόκιο
Future contribution margin= Μελλοντικό περιθώριο συνεισφορών
Future cost= μελλοντικό κόστος

- **2η Εξατομικευμένη Προσέγγιση**:

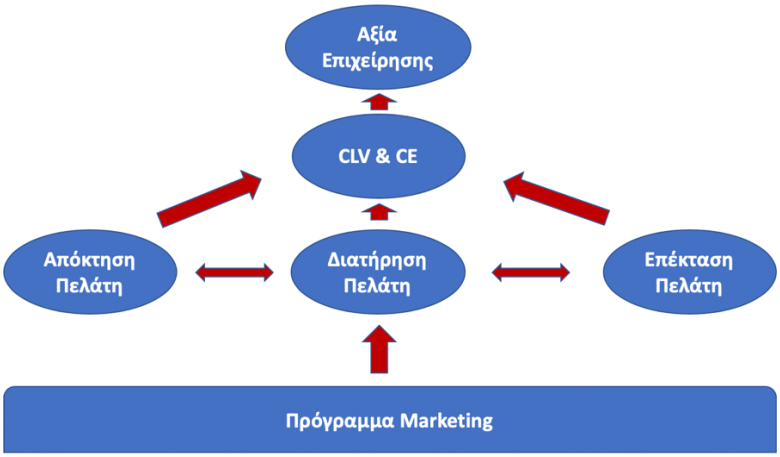
Ακολουθεί μία έτερη προσέγγιση υπολογισμού του CLV βάσει εξατομικευμένης λογικής.

$$LTVc = [Cc + WOMc] * Wc$$

Όπου,
LTVc = η αξία διάρκειας ζωής των πελάτη c,
Cc = το άθροισμα των ταμειακών ροών (προεξοφλούνται σε παρούσες αξίες) του c πελάτη κατά τη διάρκεια ζωής του με την εταιρεία,
WOMc = τα έσοδα από τους πελάτες που αποκτήθηκαν μέσω από στόμα σε στόμα συστάσεων του πελάτη c,
Wc = ένας συντελεστής στάθμισης ανάλογα με τη δυνητική αξία του προϊόντος.

4.2 Μοντελοποίηση του δείκτη Customer Lifetime Value

Για πρώτη φορά το 2006 χαρτογραφήθηκε ένα αρχικό πλαίσιο μοντελοποίησης του δείκτη CLV από τους Gupta και Hanssens. Η αιτία της δημιουργίας του πλαισίου αυτού αλλά και άλλων παρόμοιων είναι η δυσκολία του σαφούς ορισμού του δείκτη καθώς και όλα τα λοιπά προβλήματα τα οποία δημιουργεί η αυθερετότητα αυτή, όπως για παράδειγμα η μη ορθή κατανόηση των βασικών εννοιών και η μη δυνατότητα προσδιορισμού και μέτρησης των πραγματικών οφελών που προκύπτουν εν τέλει για μία επιχείρηση. (Blattberg, Malthouse, & Neslin , 2009)



Εικόνα 7 Πλαίσιο Μοντελοποίησης Αξίας Διάρκειας Ζωής Πελάτη (CLV=Διάρκεια ζωής πελάτη, CE Καθαρή θέση πελάτη)

Το παραπάνω σχήμα αποτελεί μία αποτύπωση των αλληλοεπιδράσεων του CLV σε μία εταιρία με σκοπό την αρχική μοντελοποίησή του. Στην πορεία, ήταν πολλά εκείνα τα μοντέλα τα οποία αναπτύχθηκαν, βασισμένα σε διαφορετικές τεχνικές και προσεγγίσεις.

Μία ομάδα μοντέλων χαρακτηρίζεται από την επικέντρωση τους στην εξεύρεση των επιπτώσεων των στρατηγικών marketing στην απόκτηση και διατήρηση των πελατών. Σε άλλες περιπτώσεις, το επίκεντρο βρίσκεται στις διαφορές των συνιστωσών του CLV. Ωστόσο, υπήρξαν αντικρουόμενες απόψεις ως προς το ποια συνιστώσα είναι η πιο σημαντική. Ως εκ τούτου, σε άλλες περιπτώσεις η διατήρηση του πελάτη κρίνεται ως πιο σημαντική ενώ σε άλλα μοντέλα δεν θεωρείται ότι η διάρκεια είναι αυτή που αποφέρει το μεγαλύτερο κέρδος.

4.2.1 Μοντέλα RFM

Η παλαιότερη οικογένεια μοντέλων είναι αυτή των RFM. Η ονομασία τους έχει προέλθει από τα αρχικά τριών βασικών παραμέτρων.

- **Recency**: Εγγύτητα – το χρονικό διάστημα από την τελευταία αγορά
- **Frequency**: Συχνότητα – η συχνότητα αγορών σε μία συγκεκριμένη χρονική περίοδο.
- **Monetary**: Νομισματική Αξία – Η αξία των αγορών εντός μίας περιόδου

Υπολογίζοντας τα παραπάνω μεγέθη, τα μοντέλα αυτής της κατηγορίας προσπαθούν να προβλέψουν τις κινήσεις είτε μεμονωμένων πελατών είτε ομάδες αυτών. Η χρήση αυτού του είδους των μοντέλων είναι μέρος στρατηγικών που αποβλέπουν στην απόκτηση νέων πελατών αλλά και στη διατήρησή τους.

Ένας βασικός φραγμός που τα χαρακτηρίζει είναι ότι τα μοντέλα αυτά βασίζονται στην έννοια της βαθμολόγησης οπότε θεωρούνται άμεσα εξαρτημένα από τις ανωτέρω βασικές παραμέτρους και τις μετρήσεις τους. Ο συγκεκριμένος περιορισμός ωστόσο δεν τα εμποδίζει να χαρακτηρίζονται ως αποτελεσματικά καθώς τα RFM μοντέλα διαθέτουν σημαντικές μεταβλητές πωλήσεων οι οποίες δύναται να είναι αξιόπιστοι δείκτες τις μελλοντικής συμπεριφοράς των καταναλωτών.

Έχει αποδειχτεί ότι μέσω των μεταβλητών RFM είναι δυνατή η δημιουργία ενός μοντέλου CLV ικανό να ξεπεράσει αρκετούς από τους υπολογιστικού περιορισμούς καθώς τα τρία βασικά μεγέθη έχουν την ικανότητα να προβλέψουν με ασφάλεια την αξία διάρκειας ζωής του πελάτη. Σε αυτό συνεισφέρουν δύο από τα κύρια πλεονεκτήματα τους που είναι η εύκολη εφαρμογή και η ευκολία προσαρμογής στα δεδομένα πληροφόρησης. Επιπρόσθετα, τα μοντέλα της κατηγορίας αυτής κρίνονται ως επαρκή ως προς τα στατισικά στοιχεία που τα διακατέχουν. Αξιοσημείωτο είναι ότι μέσω των iso-CLV καμπυλών είναι εμφανείς οι τιμές R, F και M οι οποίες συνιστούν το CLV ενός πελάτη. (Fader, Hardie, & Lee, 2005)

Ως βασικό μειονέκτημα των μεθόδων αυτών είναι ότι επικεντρώνονται στους πελάτες οι οποίοι αποφέρουν μεγαλύτερο κέρδος, αφήνοντας σε δεύτερη μοίρα τους πελάτες οι οποίοι έχουν λιγότερη αλληλεπίδραση με μία επιχείρηση.

4.2.2 Μοντέλα Πιθανοτήτων (Probability models)

Ως μοντέλο πιθανοτήτων προσδιορίζεται μία αναπαράσταση του κόσμου κατά την οποία κάποια συγκεκριμένη συμπεριφορά θεωρείται ως υλοποίηση μίας βαθύτερης στοχαστικής έννοιας για την οποία υπολογίζονται λανθάντοντα συμπεριφορικά χαρακτηριστικά, τα οποία μάλιστα διαφέρουν μεταξύ τους.

Στη βάση της μελέτης ενός μοντέλου πιθανοτήτων πραγματοποιείται η κεντρική διεργασία η οποία σχετίζεται με μία απλή παραμορφωμένη κατάσταση και περιγράφει την παρατηρούμενη συμπεριφορά, χωρίς προσπαθεί να ερμηνεύσει τις διαφοροποιήσεις μεταξύ του μοντέλου και της αρχικής παρατήρησης. Η μοντελοποίηση αυτή θεωρείται γενικώς αποτελεσματική και έχει ως αρχή το γεγονός ότι η συμπεριφορά των καταναλωτών διαφέρει αναλόγως τον πληθυσμό βάσεις κάποιας στατιστικής κατανομής. (Gupta & Hanssens, 2006)

Στην περίπτωση του CLV, οι αναλυτές επιδιώκουν να είναι ικανοί να κάνουν προβλέψεις σχετικά με το αν ένας καταναλωτής θα συνεχίσει να είναι ενεργός μελλοντικά. Αν ισχύει η πρώτη προϋπόθεση τότε εξετάζουν την αγοραστική του συμπεριφορά. Πάνω σε αυτήν την λογική στηρίχτηκε το μοντέλο Pareto / NBD το οποίο αναπτύχθηκε από τους Schmittlein, Morrison, και Colombo το 1987. Το μοντέλο αυτό περιγράφει την ροή των συναλλαγών σε μη συμβατική ρύθμιση, όταν δηλαδή οι συναλλαγές δύναται να πραγματοποιηθούν σε οποιοδήποτε χρονικό σημείο.

Ως εναλλακτική προσέγγιση σχετικά με το ανωτέρω μοντέλο δημιουργήθηκαν το β-διωνυμικό / β-γεωμετρικό (BG / BB) μοντέλο από τους Fader, Hardie, και Berger το 2004.

Επίσης, έχουν αναπτυχθεί από αρκετούς ερευνητές μοντέλα τα οποία στηρίζονται στις αλυσίδες Markov. Ως βασικό τους χαρακτηριστικό αυτού του τύπου τα μοντέλα έχουν ότι δεν απαιτούν ειδικά εξειδικευμένες προϋποθέσεις για την εφαρμογή τους λόγω της χρήσης των Μαρκοβιανών Αλυσίδων.

4.2.3 Οικονομετρικά μοντέλα (Econometric models)

Τα οικονομετρικά μοντέλα βασίζονται κατά βάση στα μοντέλα πιθανοτήτων. Πιο συγκεκριμένα, γίνεται χρήση των μοντέλων κινδύνου σχετικά με την διατήρηση των πελατών, με αποτέλεσμα να έρχονται σε μεγάλη συνάφεια με τα μοντέλα NBD / Pareto.

Σε γενικές γραμμές αυτή η οικογένεια μοντέλων έχει τα εξής 3 βασικά χαρακτηριστικά - στόχους:

1. Απόκτηση νέων Πελατών (Customer Acquisition)
2. Διατήρηση των υφιστάμενων Πελατών (Customer Retention)
3. Το περιθώριο και την επέκταση του Πελάτη (Customer Expansion and Margin)

Το στάδιο της απόκτησης του νέου πελάτη ασχολείται με την πρώτη αγορά του καταναλωτή. Η ανάλυση σε αυτό το επίπεδο γίνεται κυρίως στους παράγοντες οι οποίοι προσδιορίζουν τις τάσεις αγορών των πελατών.

Η διατήρηση των πελατών στην ουσία αποτελεί την πιθανότητα ενός πελάτη να παραμείνει ενεργός σε έναν οργανισμό. Η πιθανότητα αυτή ελέγχεται μέσω διαφόρων συμβατικών μεθόδων μέσω των οποίων γίνεται αντιληπτή η πρόθεση του καταναλωτή να σταματήσει τις συναλλαγές με μία εταιρία. Τα υπομοντέλα της διατήρησης χωρίζονται σε δύο κατηγορίες. Η πρώτη εστιάζει σε μεθόδους κινδύνου για να υπολογίσει την πιθανότητα απομάκρυνσης του πελάτη από την εταιρία ενώ η δεύτερη κατηγορία επικεντρώνεται σε μεθόδους όπως η μετανάστευση και οι μαρκοβιανές αλυσίδες.

Όσον αφορά το τρίτο πυλώνα του CLV, την επέκταση του πελάτη, αυτό εξαρτάται από την παρελθοντική αγοραστική συμπεριφορά και των μεθόδων cross-selling. Ως εκ τούτου, κυριαρχούν οι μέθοδοι οι οποίες εστιάζουν στο περιθώριο του πελάτη σε συνδυασμό με εκείνες οι οποίες ασχολούνται με το μοντέλο των σταυροειδών πωλήσεων. (Gupta & Hanssens, 2006)

4.2.4 Διαρκώς Επαναλαμβανόμενα Μοντέλα (Persistence models)

Κοινό χαρακτηριστικό των οικονομετρικών μοντέλων και των διαρκών επαναλαμβανόμενων είναι η επικέντρωσή τους στους τρεις βασικούς πυλώνες της CLV (απόκτηση, διατήρηση και ανάπτυξη πελάτη). Ένα από τα βασικά πλεονεκτήματα αυτής της κατηγορίας μοντέλων είναι ότι έχουν την δυνατότητα να προσδιορίζουν την συμπεριφορά του καταναλωτή σε μακροχρόνια βάση.

Τα μοντέλα αυτής της οικογένειας βρίσκουν εφαρμογή σε τομείς όπως είναι η πολιτική εκπτώσεων, οι διαφημίσεις και η ποιότητα των προϊόντων – υπηρεσιών σε σχέση με τα ιδία κεφάλαια του πελάτη. Επίσης, χρησιμοποιούνται και ως εργαλεία σύγκρισης αποτελεσματικότητας άλλων μεθόδων υπολογισμού του CLV. (Villanueva, Yoo, & Hanssens, 2008)

Βασική προϋπόθεση για την εύρυθμη λειτουργία τους είναι η διαθεσιμότητα μεγάλου όγκου δεδομένων, γεγονός ωστόσο το οποίο μπορεί να θεωρηθεί ως μειονέκτημα καθώς η συλλογή όλων των απαραίτητων πληροφοριών ενδεχομένως να μην είναι

πάντοτε εφικτή. Κατά συνέπεια, αν και τα μοντέλα αυτά ενδεχομένως να μην αντιλαμβάνονται την αλληλεπίδραση των επιμέρους χαρακτηριστικών, έχουν αρκετά καλές αποδόσεις σε μακροχρόνιες προβλέψεις.

4.2.5 Μοντέλα Μηχανικής Μάθησης (Machine Learning Models)

Η δημιουργία μοντέλων στα πλαίσια της μηχανικής μάθησης συνεπάγεται την χρήση όλων των αποδοτικών μοντέλων του τομέα αυτού όπως οι τεχνικές ταξινόμησης και κατάταξης, τα νευρωνικά δίκτυα και οι μηχανές υποστήριξης αποφάσεων.

Τα μοντέλα αυτά θεωρούνται αξιόπιστα ως προς την ικανότητά τους να προβλέπουν με ακρίβεια καθώς και ως προς την διαχείριση μεγάλου όγκου πληροφοριών χωρίς να γίνεται απαγορευτικό το υπολογιστικό κόστος.

Πρόκειται για την πιο σύγχρονη κατηγορία μοντέλων η ακόμα βρίσκεται υπό εξερεύνηση και εξέλιξη καθώς η υπάρχουσα βιβλιογραφία του CLV απαρτίζεται κατά κύριο λόγο από τις πιο παραδοσιακές μεθόδους. Στο επόμενο κεφάλαιο θα καταγραφεί η ανάπτυξη μίας νέας μεθόδου η οποία θα συνδυάζει τεχνικές από την κατηγορία των μοντέλων RFM και από την κατηγορία των μοντέλων μηχανικής μάθησης. Συγκεκριμένα, για τους σκοπούς της παρούσας μελέτης έχουν εφαρμοστεί τα εξής μοντέλα μηχανικής μάθησης.

4.2.5.1 Xgboost Classifier

Η ονομασία XGBoost αποτελεί συντομογραφία της φράσης “Extreme Gradient Boosting”. Η μέθοδος Boosting για την εφαρμογή της απαιτεί τον συνδυασμό πολλών μεμονωμένων μοντέλων και έχει αναπτυχθεί την τελευταία δεκαετία. Αρχή της εν λόγω μεθόδου είναι ο συνδυασμός αποτελεσμάτων διαφόρων μη επαρκών αλγορίθμων μάθησης. Πιο συγκεκριμένα, το Boosting χρησιμοποιεί τον συνδυασμό πολλαπλών αλγορίθμων των οποίων το σφάλμα (error rate) θεωρείται καλύτερο από την τυχαία επιλογή.

Μέσω αυτής της λογικής μειώνεται η διακύμανση καθώς και η μεροληψία του μοντέλου. Ως Gradient Boosting ορίζεται η τεχνική βάση της οποίας προστίθενται νέα μοντέλα για την διόρθωση σφαλμάτων μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Τα επιμέρους μοντέλα προστίθενται διαχρονικά έως να μηδενιστεί το περιθώριο βελτιώσεως των αποτελεσμάτων.

Γενικότερα η τεχνική του boosting περιορίζει το φαινόμενο του underfitting στις περιπτώσεις των μοντέλων με μεγάλο bias. Αυτό επιτυγχάνεται με τον συνδυασμό και την κατασκευή ενός μεταβλητού τελικού μοντέλου το οποίο απαρτίζεται από άλλα αδύναμα μοντέλα τα οποία από μόνα τους ενδεχομένως να πετυχαίνουν αποτελέσματα λίγο καλύτερα από τυχαίες επιλογές. Αυτά τα υπομοντέλα είναι συνήθως δέντρα αποφάσεων ενός επιπέδου. Συμμετοχή στην εν λόγω μέθοδο έχει η τεχνική του bootstrapping με τα εξής χαρακτηριστικά:

- Η εκπαίδευση ολοκληρώνεται ύστερα από πεπερασμένο αριθμό επαναλήψεων.
- Ο αλγόριθμος κατά την εκπαίδευσή του συγκρατεί τα δεδομένα εκείνα τα οποία είχαν τα χειρότερα αποτελέσματα και τα αντιστοιχίζει σε μεγαλύτερα βάρη υπολογισμού για την επόμενη επανάληψη.
- Κατά την διαδικασία της πρόβλεψης, ο αλγόριθμος έχει την δυνατότητα να διακρίνει τις επιδόσεις του εκάστοτε υπομοντέλου μέσω της φάσης εκπαιδεύσεως και τελικώς θέτει μεγαλύτερα βάρη στα μοντέλα με τα μικρότερα καταγεγραμμένα σφάλματα.

Συγκεκριμένα στον αλγόριθμο Xgboost οποίος θεωρείται μία εξέλιξη της μεθόδου gradient boosting, τα μεμονωμένα αδύναμα μοντέλα που χρησιμοποιούνται είναι δέντρα αποφάσεων τα οποία αποτελέσματα των οποίων συνεχώς αναπροσαρμόζουν το βασικό αλγόριθμο με σκοπό την πρόβλεψη των σφαλμάτων από τα προηγούμενα δέντρα και εν τέλει την ακριβέστερη τελική πρόβλεψη των εξεταζόμενων τιμών.

Το μοντέλο Xgboost εκπαιδεύεται με έναν προσθετικό ρυθμό. Σε κάθε βήμα t αναπτύσσεται ένα δέντρο με σκοπό την ελαχιστοποίηση το σφάλμα του υφιστάμενου μοντέλου. Η μαθηματική μοντελοποίηση του μοντέλου έχει ως εξής:

1. Προσδιορίζεται το αρχικό μοντέλο F0 με σκοπό την πρόβλεψη της μεταβλητής στόχου y. Από το μοντέλο αυτό προκύπτει το υπόλοιπο (y – F0).
2. Ένα νέο μοντέλο h1 εφαρμόζεται στα υπόλοιπα του προηγούμενου βήματος.

Συνδυάζεται το μοντέλο F0 και h1 με σκοπό την δημιουργία του μοντέλου F1. Η μέση τετραγωνική απόκλιση (mean squared error) του F1 θα είναι μικρότερη από αυτή του F0.

$$F_1(x) < F_0(x) + h_1(x)$$

Για τους σκοπούς βελτιστοποίησης της απόδοσης του μοντέλου επαναλαμβάνεται η εφαρμογή του μοντέλου.

$$F_2(x) < F_1(x) + h_2(x)$$

Η ίδια λογική επαναλαμβάνεται m φορές έως την όσο το δυνατόν ελαχιστοποίηση των υπολοίπων του μοντέλου.

$$F_m(x) < F_{m-1}(x) + h_m(x)$$

Πιο αναλυτικά,

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

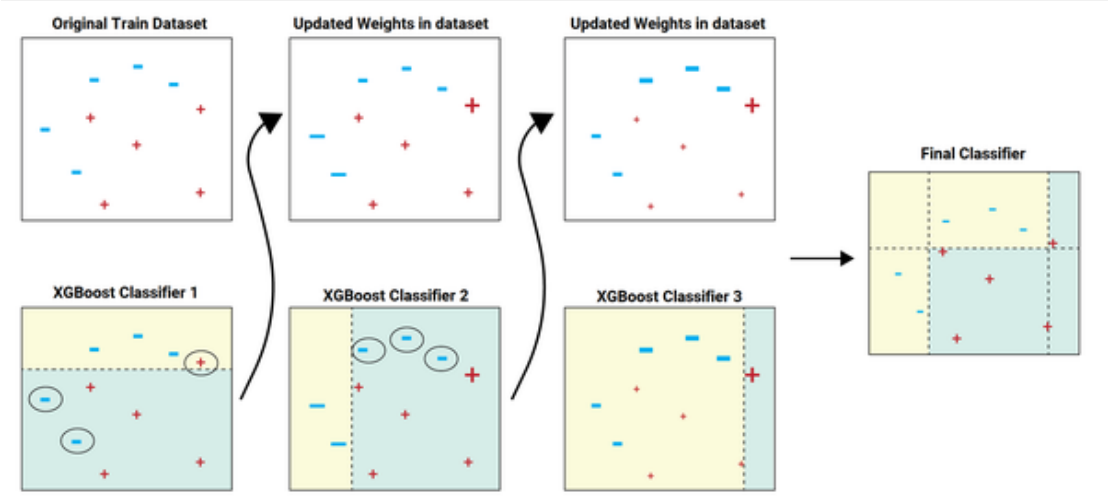
$$argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)^{\,2}$$

$$F_0(x) = \frac{\sum_{i=1}^{n} y_i}{n}$$

Τέλος, η συνάρτηση στόχου ορίζεται ως εξής:

$$L^t = \sum_{i=1}^{n} l\left(y_i, y_i^{t-1} + f_t(x_i)\right) + \Omega(f_t)$$

Όπου xi και yi είναι το σύνολο των features και των labels αντίστοιχα. Ως n ορίζεται ο αριθμός των instances για την διαδικασία της εκπαίδευσης. y_i^{t-1} είναι η πρόβλεψη για το i^{th} instance κατά την $t-1^{th}$ επανάληψη, ft είναι ένα καινούριο δέντρο το οποίο κατηγοριοποιεί το i^{th} instance κάνοντας χρήση του xi, το l δηλώνει την loss function η οποία υπολογίζει την διαφορά μεταξύ του πραγματικού label και της τιμής πρόβλεψης στο τελικό στάδιο συν το αποτέλεσμα του νέου δέντρου το οποίο δημιουργήθηκε. Τέλος, το Ω προσδιορίζει τον όρο κανονικοποίησης ο οποίος στην ουσία «τιμωρεί» την πολυπλοκότητα του δέντρου. (Nishio, Nishizawa , Surigiyama, & others, 2018)



Εικόνα 8 Διαδικασία Κατηγοριοποίησης XGbooster (Πηγή: (Shah, 2020))

Στα δυνατά χαρακτηριστικά της εν λόγω μεθόδου είναι η συμβατότητά του με τους μεγάλους όγκους δεδομένων καθώς και ότι στη γλώσσα προγραμματισμού Python η οποία θα χρησιμοποιηθεί στα πλαίσια της παρούσας εργασίας, δεν είναι απαραίτητη η κανονικοποίηση των δεδομένων (feature scaling), γεγονός το οποίο εξυπηρετεί τη διαισθητική επαφή με τα δεδομένα του προβλήματος.

Πλεονεκτήματα Xgboost:

- **Regularization**: Ο αλγόριθμος Xgboost έχει την δυνατότητα αρνητικής βαθμολόγησης των περίπλοκων μοντέλων με αποτέλεσμα να αποφεύγεται επιτυχώς το overfitting

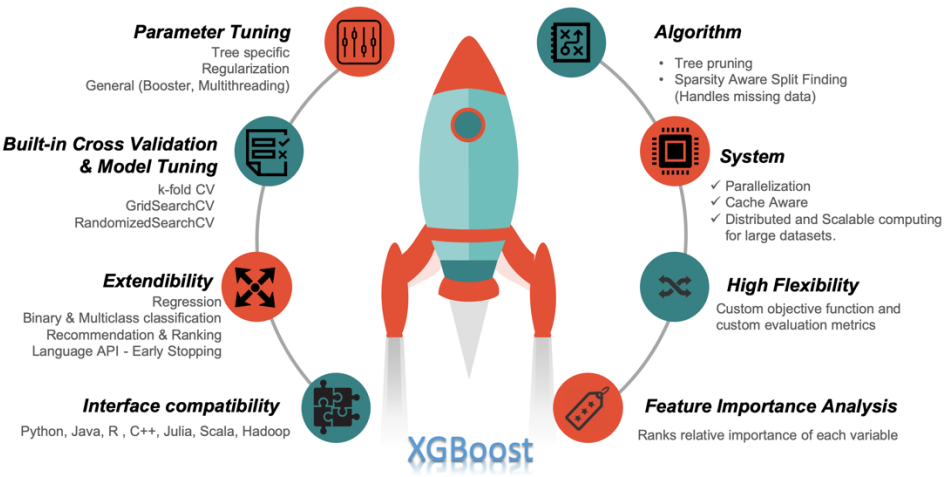
- **Χειρισμός ελλιπών δεδομένων**: Τα εφαρμοσμένα μοντέλα ενσωματώνουν έναν αλγόριθμο εύρεσης μοτίβων ελλιπών δεδομένων με σκοπό τον βέλτιστο χειρισμό των ιδιόμορφων τύπων δεδομένων.

- **Σταθμισμένο ποσοτικό διάγραμμα**: Οι περισσότεροι αλγόριθμοι δέντρων αποφάσεων έχουν την δυνατότητα προσδιορισμού των σημείων διαχωρισμού όταν τα αρχικά δεδομένα έχουν ίσα βάρη. Ωστόσο, δεν μπορούν να διαχειριστούν δεδομένα με άνισα βάρη στις τιμές τους. Ο Xgboost λόγω της δομής του, είναι ιδανικός για σύνολα δεδομένων τα οποία διακρίνονται από διαφορετικά βάρη.

- **Παράλληλη μάθηση**: Με σκοπό τους γρηγορότερους υπολογισμούς, ο Xgboost έχει την δυνατότητα να χρησιμοποιεί πολλαπλούς επεξεργαστές. Αυτό προκύπτει λόγω της δομής Block στον σχεδιασμό του αλγορίθμου. Τα δεδομένα ταξινομούνται και αποθηκεύονται σε in-memory blocks. Αυτό επιτρέπει την διαρκή χρήση των δεδομένων κατά τις επαναλήψεις του αλγορίθμου με το ελάχιστο δυνατό υπολογιστικό κόστος.

- **Cache awareness:** Δεν είναι απαραίτητη η συνεχής χρήση μνήμης λόγω της δομής του αλγορίθμου καθώς τα στατιστικά μεγέθη ανακαλούνται κατά τη χρήση του αλγορίθμου μέσω ευρετηρίων.

- **Out-of-core computing:** Το χαρακτηριστικό αυτό βελτιστοποιεί το διαθέσιμο αποθηκευτικό χώρο και μεγιστοποιεί την χρήση του. Γεγονός το οποίο είναι απαραίτητο στη διαχείριση μεγάλων δεδομένων.



Εικόνα 9 Χαρακτηριστικά Αλγορίθμου XGBoost (Πηγή: (Gowrisankar, 2020))

5. Δημιουργία Μοντέλου Μηχανικής Μάθησης για την πρόβλεψη του δείκτη CLV σε τραπεζικά δεδομένα

Σε αυτό το κεφάλαιο θα περιγράφει η πειραματική διαδικασία δημιουργίας ενός μοντέλου μηχανικής μάθησης το οποίο θα είναι ικανό να τροφοδοτείται με δεδομένα τραπεζικών συναλλαγών μέσω των οποίων θα πραγματοποιεί προβλέψεις σχετικά με τις τιμές του δείκτη Customer Lifetime Value ο οποίος έχει αναλυθεί ενδελεχώς σε προηγούμενη ενότητα.

Θα αναλυθεί όλο το πλαίσιο συλλογής, προεπεξεργασίας και ανάλυσης των δεδομένων που χρησιμοποιήθηκαν καθώς επίσης και η διαδικασία δημιουργίας και εφαρμογής μοντέλων μηχανικής μάθησης.

5.1 Τεχνικό Υπόβαθρο

Η υλοποίηση της παρούσας εργασίας έχει πραγματοποιηθεί με τη χρήση της γλώσσας προγραμματισμού Python. Σε αυτό το πλαίσιο έχει συγγραφεί ο σχετικός κώδικας για την επεξεργασία και την ανάλυση δεδομένων, καθώς και για την εφαρμογή των αλγορίθμων μηχανικής μάθησης. Επιπρόσθετα, η αποθήκευση και η τήρηση των δεδομένων πραγματοποιήθηκε σε Βάση Δεδομένων SQL Server η οποία δημιουργήθηκε για τους σκοπούς της έρευνας.

5.1.1 Η γλώσσα προγραμματισμού Python

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού η οποία δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ (Guido van Rossum) το 1990. Έχει δομηθεί με γνώμονα την αναγνωσιμότητα του κώδικά της καθώς και την ευκολία χρήσης της. Η ευκολία αυτή προκύπτει λόγω του συντακτικού της το οποίο δίνει την δυνατότητα στον προγραμματιστή να ολοκληρώσει τις εργασίες του σε λιγότερες γραμμές κώδικα από ότι σε άλλες παραδοσιακές γλώσσες προγραμματισμού όπως η C++ και η Java.

Στα πλεονεκτήματά της προσμετρούνται οι πολλές διαθέσιμες βιβλιοθήκες της οι οποίες προσφέρουν στους προγραμματιστές επαρκή εργαλεία για όλες τις συνήθεις εργασίες. Ακολουθούν οι βασικές βιβλιοθήκες οι οποίες χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής εργασίας.

- **NumPy**: Η βιβλιοθήκη NumPy εντάσσεται στα πλαίσια του ελεύθερου λογισμικού η οποία δημιουργήθηκε το 2005 από τον Travis Oliphant. Βασική της ιδιότητα είναι η διαχείριση μεγάλων πολυδιάστατων πινάκων και περιλαμβάνει μία σειρά από μαθηματικές συναρτήσεις υψηλού επιπέδου οι οποίες είναι συμβατές με τους πίνακες που προσφέρει η βιβλιοθήκη.

- **Pandas**: Η Pandas αποτελεί μία βιβλιοθήκη ελεύθερου λογισμικού η οποία πρωτοεμφανίστηκε γραμμένη σε γλώσσα προγραμματισμού C και Python από τον Wes McKinney το 2008. Βασική της χρήση είναι τα projects διαχείρισης και ανάλυσης δεδομένων. Κύριο αντικείμενό της είναι τα Data Frames τα οποία

είναι συναφή των πινάκων σε άλλες γλώσσες προγραμματισμού. Η βιβλιοθήκη αυτή επιτρέπει την ανάκτηση δεδομένων από διαφορετικές πηγές, όπως βάσεις δεδομένων.

- **Scikit-learn**: Η βιβλιοθήκη αυτή αποτελεί προϊόν ανοιχτού λογισμικού και προορίζεται για εργασίες μηχανικής μάθησης. Παρέχει απλά και αποδοτικά εργαλεία για εξόρυξη και ανάλυση δεδομένων. Σε αυτήν περιλαμβάνονται πολλές υλοποιήσεις αλγορίθμων οι οποίοι αντιμετωπίζουν προβλήματα ταξινόμησης, ομαδοποίησης και παλινδρόμησης. Επίσης σε αυτήν εμπεριέχονται μέθοδοι αξιολόγησης των αλγορίθμων με σκοπό την αύξηση της αποδοτικότητας.

- **Matplotlib**: Αποτελεί επίσης άλλη μια βιβλιοθήκη ανοιχτού λογισμικού η οποία παράγει υψηλής ποιότητας και λεπτομέρειας γραφήματα. Σκοπός της είναι η δημιουργία γραφημάτων και εν γένει visualizations με τη χρήση λίγων γραμμών κώδικα.

- **Pyodbc**: Η βιβλιοθήκη αυτή χρησιμεύει κατά την διασύνδεση της γλώσσας Python με μία βάση δεδομένων. Αυτό πραγματοποιείται μέσω του API Open Database Connectivity (ODBC). Μέσω του ODBC επιτυγχάνεται η ανεξαρτησία του DBMS συστήματος καθώς όλη η επεξεργασία των δεδομένων πραγματοποιείται σε ένα ενδιάμεσο στάδιο (layer). Το API αυτό διαθέτει λειτουργίες διαχειριστή οι οποίες επιτρέπουν την υλοποίηση ερωτημάτων προς την βάση δεδομένων.

5.1.2. Βάση Δεδομένων και Συστήματα Διαχείρισης (DBMS)

Μία Βάση Δεδομένων αποτελεί έναν αποτελεσματικό τρόπο αποθήκευσης δεδομένων τα οποία είναι δυνατόν να αποθηκευτούν σε ηλεκτρονικό υπολογιστή. Βασική προϋπόθεση για την ύπαρξη μίας σχεσιακής βάσης είναι η συσχέτιση των δεδομένων.

Ως εκ τούτου, μία σωστή βάση δεδομένων θα πρέπει στην ουσία να αντιπροσωπεύει μία κατάσταση του πραγματικού κόσμου. Αυτό προϋποθέτει την ύπαρξη λογικής συνέχειας μεταξύ των πληροφοριών. Βασικός στόχος της δημιουργίας τους είναι η απεικόνιση προβλημάτων μέσω του υπολογιστή και η διευκόλυνση επίλυσής τους. Συνεπώς τα δεδομένα που συμπεριλαμβάνονται σε μία Βάση είναι συνήθως δυναμικά, δηλαδή τροποποιούνται με βάση χρονικό προσδιορισμό και παρέχουν την έννοια της ιστορικότητας.

Όπως είναι φυσικό, η δημιουργία και διαχείριση των βάσεων δεδομένων συνεπάγεται και την ανάγκη ύπαρξης κατάλληλων εργαλείων διαχείρισης τους. Αυτά τα συστήματα ονομάζονται Συστήματα Διαχείρισης Βάσεων Δεδομένων (Database Management Systems ή DBMS). Τα συστήματα αυτά προσφέρουν την δυνατότητα στους χρήστες να κατασκευάζουν και να διαχειρίζονται Βάσεις Δεδομένων.

Ένα τέτοιο σύστημα συνήθως συμπεριλαμβάνει πληθώρα Βάσεων οι οποίες έχουν κατασκευαστεί από διαφορετικούς ενδεχομένως χρήστες. Οι δυνατότητες οι οποίες παρέχονται είναι οι εξής:

- Ορισμός Βάσης Δεδομένων
- Κατασκευή Βάσης Δεδομένων
- Διαγραφή Βάσης Δεδομένων
- Χρήση Βάσης Δεδομένων

Κατά τον ορισμό της Βάσης, ο χρήστης έχει την δυνατότητα να καθορίσει το μοντέλο της και να ορίσει τους τύπους των δεδομένων που θα συμμετάσχουν στη Βάση. Κατά την κατασκευή της τα δεδομένα αποθηκεύονται πλέον στο hardware του υπολογιστή μέσω διαδικασιών που προκύπτουν από το DBMS. Ο χρήστης έχει το περιθώριο να διαγράψει μία βάσης δεδομένων ή ακόμα και μεμονωμένα δεδομένα. Τέλος, κατά τη συνηθέστερη εργασία η οποία είναι η χρήση της Βάσης Δεδομένων, ο χρήστης δύναται να διαχειρίζεται τα δεδομένα, να τα παραμετροποιεί, να προσθέτει νέα ή ακόμα και να θέτει ερωτήσεις προς την Βάση με σκοπό την εξόρυξη πληροφοριών. (Silberschatz, Korth, & Sudarshan)

5.2 Συλλογή και Περιγραφή των Δεδομένων

Στην παρούσα πειραματική μελέτη θα πραγματοποιηθεί η εφαρμογή ανωνυμοποιημένων τραπεζικών δεδομένων τα οποία προέρχονται από συναλλαγές πελατών σε μοντέλα μηχανικής μάθησης τα οποία θα εκτιμούν προβλέψεις για τον δείκτη Customer Lifetime Value. Η χρήση του CLV είθισται να περιορίζεται σε εμπορικές επιχειρήσεις οι οποίες για τους σκοπούς της μέτρησης του χρησιμοποιούν δεδομένα από φυσικές παραγγελίες. Η κατωτέρω μέθοδος επεκτείνει την αρχική ιδέα σε ένα εξελιγμένο πλαίσιο, βάσει του οποίου τη θέση των πωλήσεων εμπορευμάτων παίρνουν οι συναλλαγές οι οποίες πραγματοποιούνται σε ένα τραπεζικό ίδρυμα και ως κέρδος θεωρείται το έσοδο από τις τραπεζικές προμήθειες.

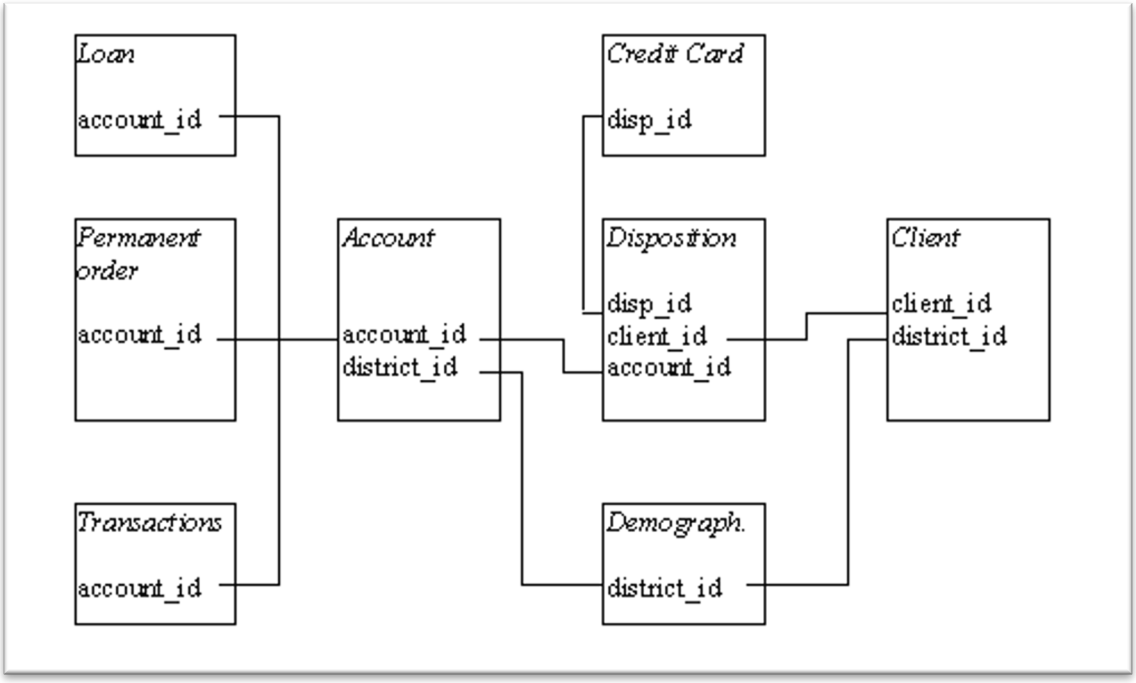
Με γνώμονα το ανωτέρω σκεπτικό, το σύνολο δεδομένων αποτελεί μία συλλογή από οικονομικά δεδομένα μίας τσέχικης τράπεζας το οποίο κυκλοφόρησε το 1999 ανωνυμοποιημένο, για τους σκοπούς του διαγωνισμού PKDD Discovery Challenge ο οποίος είχε ως στόχο την βελτίωση των υπηρεσιών που προσφέρει η συγκεκριμένη τράπεζα. Συνολικά αποτελείται από πληροφορίες οι οποίες σχετίζονται με πάνω από 5.300 πελάτες και 1.000.000 συναλλαγές διαφόρων ειδών. Η τράπεζα είχε προσφέρει περίπου 700 δάνεια ενώ είχε εκδώσει 900 πιστωτικές κάρτες.

Σχετικά με τον κάθε τραπεζικό λογαριασμό υφίστανται τόσο στατικές πληροφορίες όπως η ημερομηνία δημιουργίας, διευθύνσεις πελάτη και καταστήματος δημιουργίας, όσο και δυναμικά χαρακτηριστικά όπως εισερχόμενες και εξερχόμενες συναλλαγές και υπολειπόμενα ποσά.

Ένας πελάτης δύναται να τηρεί άνω του ενός λογαριασμού ενώ πολλοί πελάτες μπορούν να διαχειρίζονται ένα λογαριασμό. Η συσχέτιση μεταξύ πελατών και λογαριασμών πραγματοποιείται μέσω ειδικών αριθμών – κλειδιών. Επίσης, ο κάθε πελάτης μπορεί να έχει στην κατοχή του περισσότερες από μία κάρτες, ενώ για τις περιπτώσεις των δανείων ισχύει η διασύνδεση 1 προς 1 με την κατηγορία των τραπεζικών λογαριασμών.

5.3 Αποθήκευση Δεδομένων

Η αρχική μορφή των δεδομένων ήταν σε επιμέρους αρχεία .CSV (comma-separated values file) τα οποία αντλήθηκαν από την διαδικτυακή τοποθεσία data.world. Για τους σκοπούς επεξεργασίας τους δημιουργήθηκε μία σχεσιακή βάση στο DBMS σύστημα Azure Data Studio.



Εικόνα 10 Σχεδιάγραμμα Σχεσιακής Βάσης

Όπως παρατηρείται στο ανωτέρω διάγραμμα, η βάση αποτελείται από 8 πίνακες οι οποίοι είναι οι εξής:

- **Accounts**: Αποτελείται από 3 μεταβλητές οι οποίες σχετίζονται με τον κωδικό των τραπεζικών λογαριασμών, τον κωδικό της περιοχής του καταστήματος δημιουργίας και τη συχνότητα πάγιων εντολών.

- **Cards**: Περιέχει πληροφορίες σχετικά με τον τύπο της εκάστοτε κάρτας και την ημερομηνία έκδοσης.

- **Clients**: Ο εν λόγω πίνακας περιλαμβάνει όλες τις πελατειακές πληροφορίες όπως τον κωδικός πελάτη, το φύλο, την ηλικία και τον κωδικό που αντιστοιχεί στην περιοχή κατοικίας του πελάτη.

- **Dispositions**: Ο συγκεκριμένος πίνακας λειτουργεί υποστηρικτικά καθώς διασυνδέει τις πληροφορίες για τους τραπεζικούς λογαριασμούς με τα δεδομένα που αφορούν τους πελάτες. Με αυτόν τον τρόπο αντιμετωπίζεται αποτελεσματικά το γεγονός ότι ένας λογαριασμός δύναται να διαχειρίζεται από έναν ή περισσότερους πελάτες.

- **Loans**: Αποτελείται από μεταβλητές οι οποίες τηρούν πληροφορίες για τα ενεργά δάνεια της τράπεζας. Οι πληροφορίες σχετίζονται με τα ποσά των δανείων, τις ημερομηνίες εκταμίευσης, τις μηνιαίες πληρωμές, την διάρκεια του δανείου καθώς και με την βαθμολόγηση της πιστοληπτικής ικανότητας του πελάτη.

- **Orders**: Σε αυτόν τον πίνακα τηρούνται τα στοιχεία με τα εμβάσματα τα οποία πραγματοποιήθηκαν από τους λογαριασμούς των πελατών προς άλλες τράπεζες.

- **Districts**: Αποτελείται από πληροφορίες σχετικά με τα δημογραφικά στοιχεία των πελατών της τράπεζας αλλά και των περιοχών όπου η τράπεζα δραστηριοποιείται, δηλαδή που τηρεί κατάστημα.

- **Transactions**: Στον πίνακα αυτόν συμπεριλαμβάνονται όλες οι πληροφορίες οι οποίες αφορούν τις συναλλαγές που έχουν πραγματοποιηθεί διαχρονικά από τους πελάτες. Είναι εφικτή η διασύνδεση των δεδομένων μέσω των αριθμών λογαριασμών και τους κωδικούς των συναλλαγών. Επιπρόσθετα, τηρούνται στοιχεία σχετικά με τα ποσά των συναλλαγών, τις ημερομηνίες εκτέλεσής τους καθώς και με τις αιτιολογίες των συναλλαγών.

	transaction_id	account_id_	trans_date	type	operation	amount	balance	k_symbol
1	490679	1205	1997-02-13	CREDIT	COLLECTION FROM ANOTHER B…	5239.0000	19516.0000	OLD AGE PENSION
2	490680	1207	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	9252.0000	457759.0000	HOUSEHOLD
3	490681	4186	1997-02-13	CREDIT	CREDIT IN CASH	20729.0000	533078.0000	
4	490682	4259	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	9516.0000	569592.0000	HOUSEHOLD
5	490683	4259	1997-02-13	CREDIT	COLLECTION FROM ANOTHER B…	44110.0000	686752.0000	
6	490684	4190	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	1100.0000	213929.0000	
7	490685	4190	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	2901.0000	224929.0000	HOUSEHOLD
8	490686	4259	1997-02-13	DEBIT	CASH WITHDRAWAL	2200.0000	664752.0000	
9	490687	2932	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	26.0000	380093.0000	
10	490688	2944	1997-02-13	CREDIT	CREDIT IN CASH	5746.0000	168074.0000	
11	490689	2947	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	2228.0000	192419.0000	HOUSEHOLD
12	490690	2936	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	4247.0000	52862.0000	HOUSEHOLD
13	490691	3309	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	2991.0000	225191.0000	
14	490692	3307	1997-02-13	CREDIT	CREDIT IN CASH	1500.0000	457255.0000	
15	490693	3292	1997-02-13	CREDIT	COLLECTION FROM ANOTHER B…	4805.0000	140489.0000	OLD AGE PENSION
16	490694	3948	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	1969.0000	375843.0000	INSURANCE PAYM…
17	490695	3944	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	1165.0000	411508.0000	HOUSEHOLD
18	490696	2676	1997-02-13	CREDIT	CREDIT IN CASH	15263.0000	442875.0000	
19	490697	2676	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	1339.0000	429485.0000	
20	490698	2681	1997-02-13	DEBIT	CASH WITHDRAWAL	2000.0000	515954.0000	
21	490699	2818	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	8947.0000	434444.0000	HOUSEHOLD
22	490700	2810	1997-02-13	DEBIT	REMITTANCE TO ANOTHER BANK	392.0000	292391.0000	HOUSEHOLD
23	490701	2814	1997-02-13	CREDIT	COLLECTION FROM ANOTHER B…	4544.0000	177858.0000	OLD AGE PENSION
24	490702	11111	1997-02-13	DEBIT	CASH WITHDRAWAL	22667.0000	431039.0000	

Εικόνα 11 Απόσπασμα του πίνακα "Transactions" από την Βάση Δεδομένων

Διασύνδεση μεταξύ της βάσης δεδομένων και του λογισμικού ανάπτυξης της γλώσσας Python (IDE), πραγματοποιείται μέσω της βιβλιοθήκης pyodbc. Με αυτόν τον τρόπο, είναι εφικτή η συγγραφή ερωτημάτων σε γλώσσα SQL προς την ανωτέρω βάση δεδομένων βάσει των οποίων αντλούνται δυναμικά τα δεδομένα τα οποία κρίνονται απαραίτητα για τους σκοπούς της ανάλυσης.

```
df_rfm1 = """
            select cl.client_id,
                   case when
                    sum(tr.amount) = 0 then null
                   else
                       case when (floor (max(tr.amount)/sum(tr.amount) * 100)) = 0 then 1
                       else  floor(max(tr.amount)/sum(tr.amount) * 100)
                       end
                   end as trans_index

            from transactions as tr
            left join dbo.dispositions di on tr.account_id_ = di.account_id
            left join dbo.clients cl on di.client_id = cl.client_id --where cl.client_id is null
            left join dbo.accounts ac on di.account_id = ac.account_id
            left join districts ds on ac.district_id = ds.district_id
            where di.type = 'OWNER'
            group by cl.client_id

           """
rfm1 = pd.read_sql(df_rfm1,conn)
```

Εικόνα 12 Παράδειγμα SQL query μέσω του IDE Jupyter της Python

Η αρχιτεκτονική αυτή επιλέχθηκε με γνώμονα το τελικό στάδιο της εφαρμογής υπολογισμού και πρόβλεψης του Customer Lifetime Value, καθώς με βάση την ίδια λογική η εφαρμογή, δοθέντων συγκεκριμένων περιμέτρων, θα δύναται να αντλεί δεδομένα από την Βάση και να πραγματοποιεί προβλέψεις.

5.4 Προεπεξεργασία Δεδομένων

Δεδομένης της ροής των δεδομένων από την Βάση Δεδομένων SQL Server προς το περιβάλλον ανάπτυξης της Python, ένα σημαντικό μέρος της προεπεξεργασίας των δεδομένων πραγματοποιείται μέσω των SQL ερωτημάτων που χρησιμοποιούνται για να καθορίσουν τα διάφορα Dataframes της ανάλυσης.

Ως εκ τούτου πραγματοποιήθηκαν εργασίες ενοποίησης δεδομένων (Data Intergration) από διάφορους πίνακες της Βάσης καθώς και μετασχηματισμού (Data Trasformation). Πιο συγκεκριμένα, ο μετασχηματισμός πραγματοποιήθηκε σε δεδομένα τα οποία εκφράζονταν μέσω ποιοτικών μεταβλητών στην τσέχικη γλώσσα. Αφορούν μεταβλητές όπως τη συχνότητα των πάγιων πληρωμών, τον τύπο των συναλλαγών και την μέθοδο πληρωμής (μετρητά, κάρτα). Επίσης διαμορφώθηκαν ομοίως όλες οι μεταβλητές των ημερομηνιών ώστε να υπάρχει διαρκής συνάφεια αυτής της κατηγορίας των μεταβλητών.

Με σκοπό την εις βάθος ανάλυση των δεδομένων και την καλύτερη κατανόηση του προβλήματος, δημιουργήθηκαν 3 εξής νέες μεταβλητές (feature construction) μέσω αθροιστικών πράξεων σε υφιστάμενες μεταβλητές.

- **Age**: Πρόκειται για συνεχή μεταβλητή η οποία δημιουργήθηκε από τον συνδυασμό της μεταβλητής birthnumber (ημερομηνία γέννησης) και της χρονολογίας του dataste.

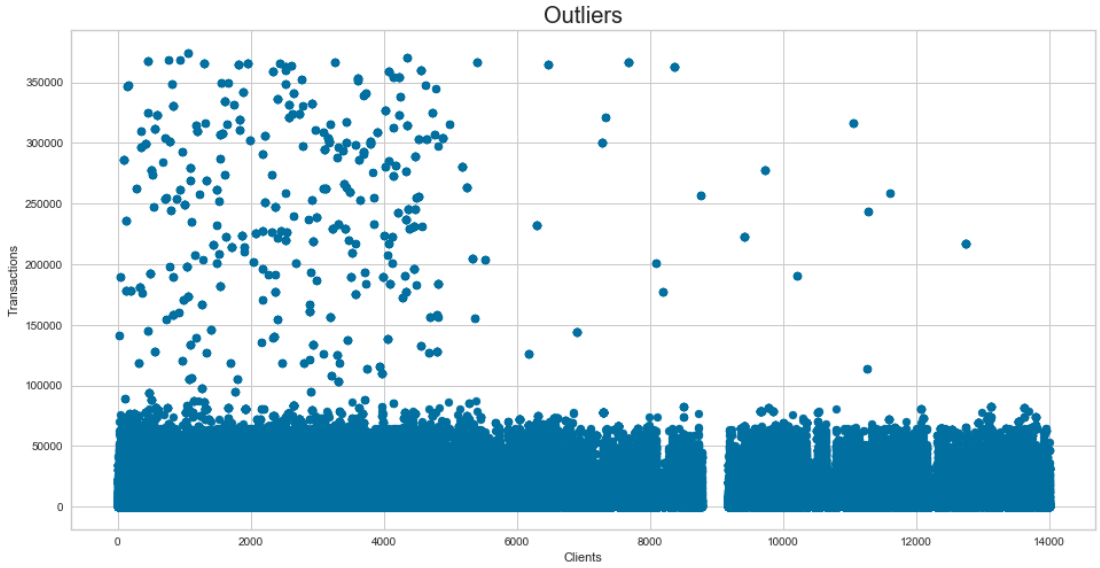
- **Age_levels**: Ποιοτική μεταβλητή η οποία δημιουργήθηκε ύστερα από τμηματοποίηση της μεταβλητής age και δηλώνει το ηλικιακό group των πελατών.

- **Trans_index**: Η συγκεκριμένη μεταβλητή δημιουργήθηκε για τους σκοπούς τροφοδότησης του τελικού μοντέλου καθώς θα αποτελέσει έναν από τους βασικούς παράγοντες στην εκτίμηση της Αξίας Ζωής του Πελάτη και στην πρόβλεψη αυτής. Η συνήθης πρακτική στην θεωρία του CLV είναι να χρησιμοποιείται η ποσότητα των προϊόντων μίας παραγγελίας ως παράγοντα υπολογισμού του δείκτη. Ωστόσο, στα πλαίσια της παρούσας εργασίας και δεδομένης της εφαρμογής της μεθόδου σε δεδομένα τραπεζικών συναλλαγών, η ανωτέρω έννοια αντικαταστάθηκε από μία μετρική η οποία δίνει τα αντίστοιχα βάρη στις συναλλαγές πελατών. Η τιμή καθορίζεται βάσει της μέγιστης συναλλαγής του πελάτη προς το σύνολο των συναλλαγών του. Με αυτόν τον τρόπο καλύφθηκε το κενό της προσαρμογής της μεθόδου σε υπηρεσίες και όχι προϊόντα.

```
df_rfm1 = """
            select cl.client_id,
                   case when
                    sum(tr.amount) = 0 then null
                   else
                       case when (floor (max(tr.amount)/sum(tr.amount) * 100)) = 0 then 1
                       else  floor(max(tr.amount)/sum(tr.amount) * 100)
                       end
                   end as trans_index

            from transactions as tr
            left join dbo.dispositions di on tr.account_id_ = di.account_id
            left join dbo.clients cl on di.client_id = cl.client_id --where cl.client_id is null
            left join dbo.accounts ac on di.account_id = ac.account_id
            left join districts ds on ac.district_id = ds.district_id
            where di.type = 'OWNER'
            group by cl.client_id

           """
rfm1 = pd.read_sql(df_rfm1,conn)
```

Εικόνα 13 Κατασκευή της μεταβλητής trans_index

Τέλος, στα πλαίσια του καθαρισμού των δεδομένων (Data Cleaning), πραγματοποιήθηκε έλεγχος στα δεδομένα σχετικά με την ύπαρξη χαμένων η ακραίων τιμών. Μοναδική μεταβλητή η οποία παρουσίασε ακραίες τιμές ήταν εκείνη των ποσών των συναλλαγών. Ωστόσο, λόγω της ιδιαιτερότητας της μεταβλητής και της βαρύτητάς της στην συγκεκριμένη ανάλυση επιλέχτηκε να μην πραγματοποιηθεί κάποια περεταίρω ενέργεια προς αποφυγής διαστρεβλώσεως των πραγματικών δεδομένων.

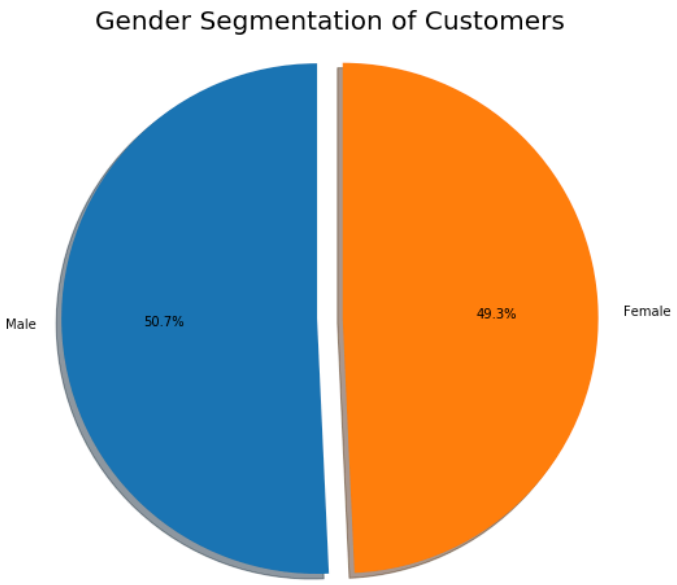


Εικόνα 14 Γράφημα Ακραίων Τιμών

5.5 Διερεύνηση των Δεδομένων

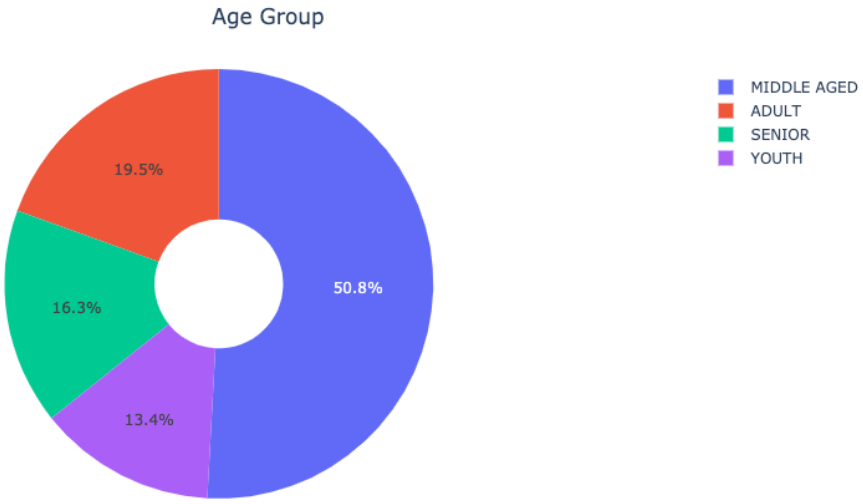
Με σκοπό την αποτελεσματική ανάλυση των δεδομένων της μελέτης πραγματοποιήθηκε μία διερευνητική εργασία πάνω στις πληροφορίες της βάσης δεδομένων ώστε να δημιουργηθεί μία ολοκληρωτική εικόνα σε σχέση με τα μεγέθη και τα χαρακτηριστικά των δεδομένων από τα οποία μετέπειτα θα προκύψει το τελικό Dataset πάνω στο οποίο θα εκπαιδευτούν και θα αξιολογηθούν οι προκριθέντες αλγόριθμοι.

Με βάση τα στοιχεία για το σύνολο των πελατών, παρατηρείται ότι εκείνοι είναι σχεδόν μοιρασμένοι ανάμεσα στα δύο φύλα.



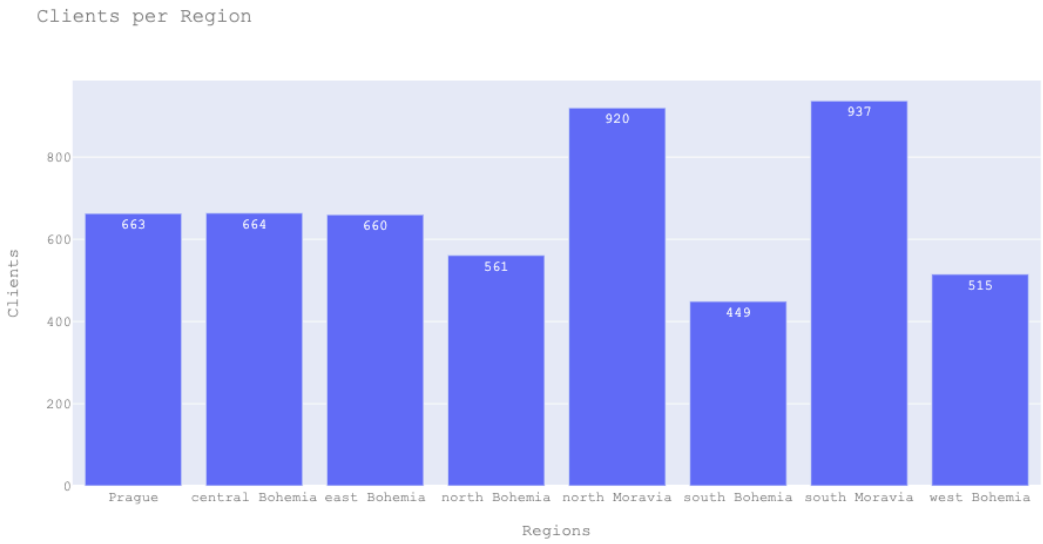
Εικόνα 15 Πελάτες ανά φύλο

Εξετάζοντας ωστόσο τις ηλικιακές ομάδες των πελατών βάσει της νέας μεταβλητής age_levels, είναι εμφανές ότι οι μισοί σχεδόν από τους πελάτες είναι μεσήλικες, γεγονός που σημαίνει ότι ανήκουν σε παραγωγικές ηλικίες και κατά συνέπεια αποτελούν σημαντική πελατεία για μία τράπεζα δεδομένου ότι είναι αυτοί οι οποίοι διακινούν τον περισσότερο πλούτο.



Εικόνα 16 Ηλικιακές Ομάδες

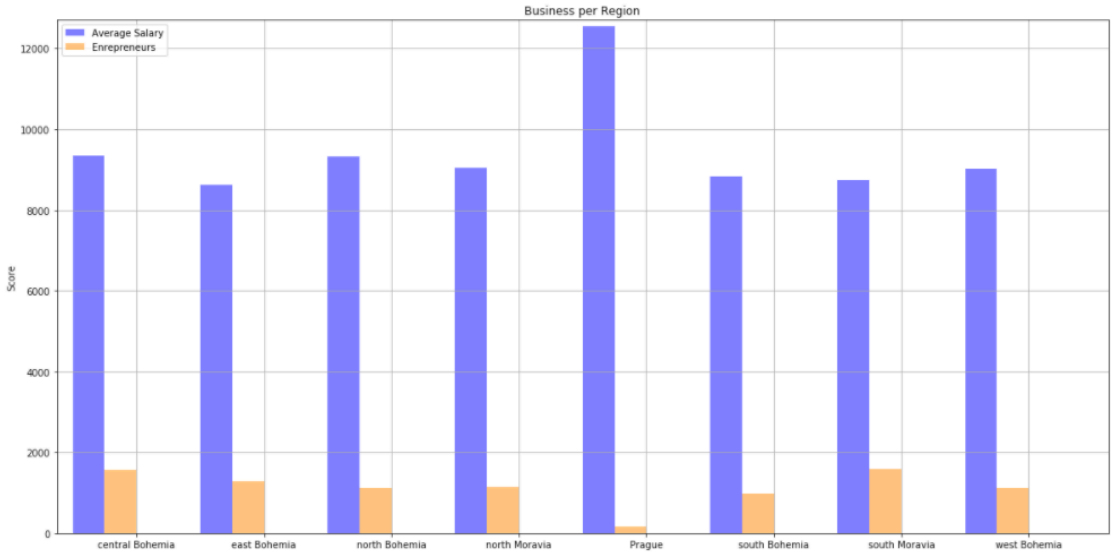
Εξετάζοντας εν συνεχεία ορισμένα γεωγραφικά στοιχεία τα οποία υπάρχουν στην βάση δεδομένων, παρατηρήθηκε ότι οι περισσότεροι πελάτες της τράπεζας δραστηριοποιούνται στις περιοχές North Moravia και South Moravia, ωστόσο η πόλη στην οποία κατοικούν οι πελάτες με τον υψηλότερο μέσο μισθό, είναι η πρωτεύουσα της χώρας, η Πράγα.



Εικόνα 17 Πελάτες ανά Περιοχή

Εντούτοις, η Πράγα παρουσιάζει τα χαμηλότερα ποσοστά επιχειρηματικότητας γεγονός το οποίο υποδεικνύει ότι τα εισοδήματα των κατοίκων της προέρχονται από εξαρτημένη εργασία και όχι από δικές τους επιχειρήσεις. Στοιχείο το οποίο είναι σημαντικό για την διαχρονική πιστοληπτική ικανότητα των πελατών αυτών. Το μεγαλύτερο ποσοστό επιχειρηματικής δραστηριότητας βρίσκεται στις περιοχές South Moravia και Central Bohemia.

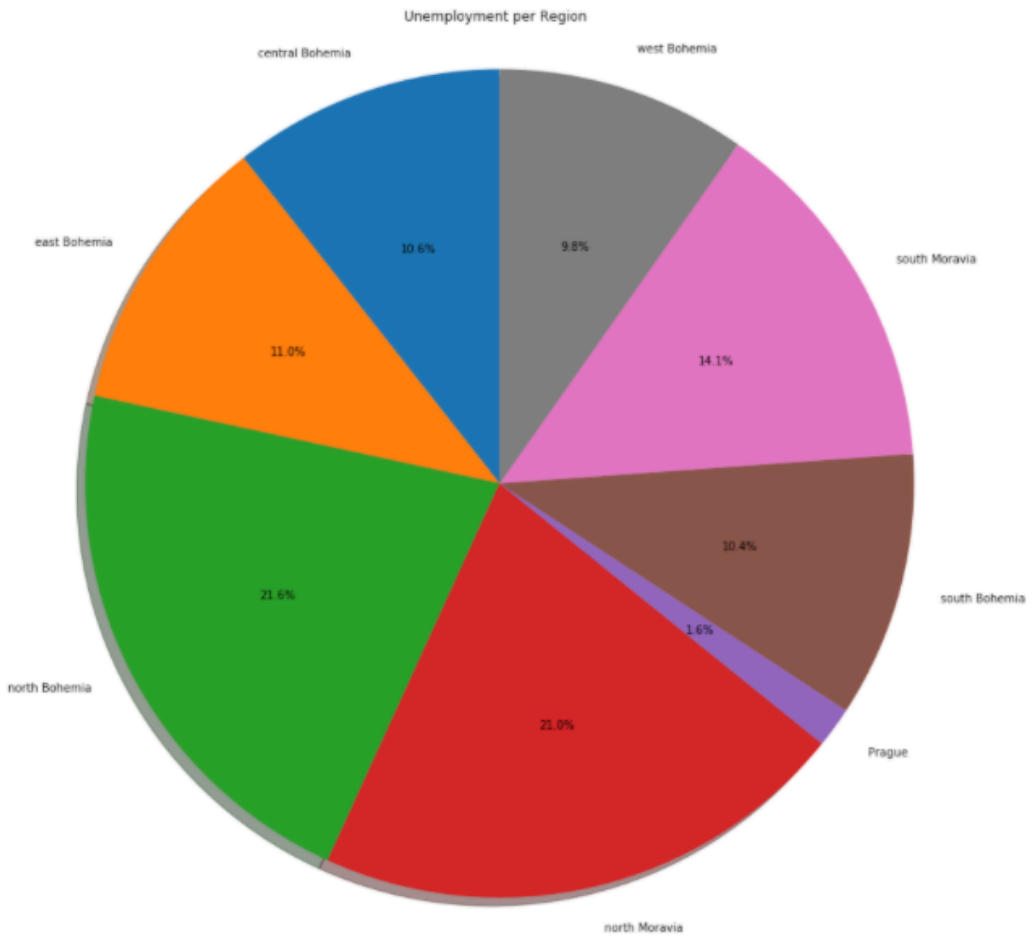
Ως εκ τούτου, για τους σκοπούς μίας στρατηγικής η οποία θα αποσκοπεί σε βελτιστοποίηση των κερδών της και της ποιότητας υπηρεσιών της τράπεζας θα πρέπει να λάβει υπόψη τις διαφοροποιήσεις και τις ανάγκες τόσο του retail όσο και του wholesale πελατολογίου.



Εικόνα 18 Επιχειρηματικότητα ανά περιοχή

Ύστερα από την ολοκλήρωση της αρχικής εξερεύνησης των γεωγραφικών και δημογραφικών στοιχείων των πελατών, η τελευταία σημαντική πληροφορία η οποία προκύπτει από τα στοιχεία της ανεργίας είναι ότι οι περιοχές στις οποίες κατοικούν οι περισσότεροι πελάτες της τράπεζας (North Moravia και South Moravia), βρίσκονται στην 2^{η} και 3^{η} θέση αντίστοιχα όσον αφορά τα ποσοστά ανεργίας της χώρας.

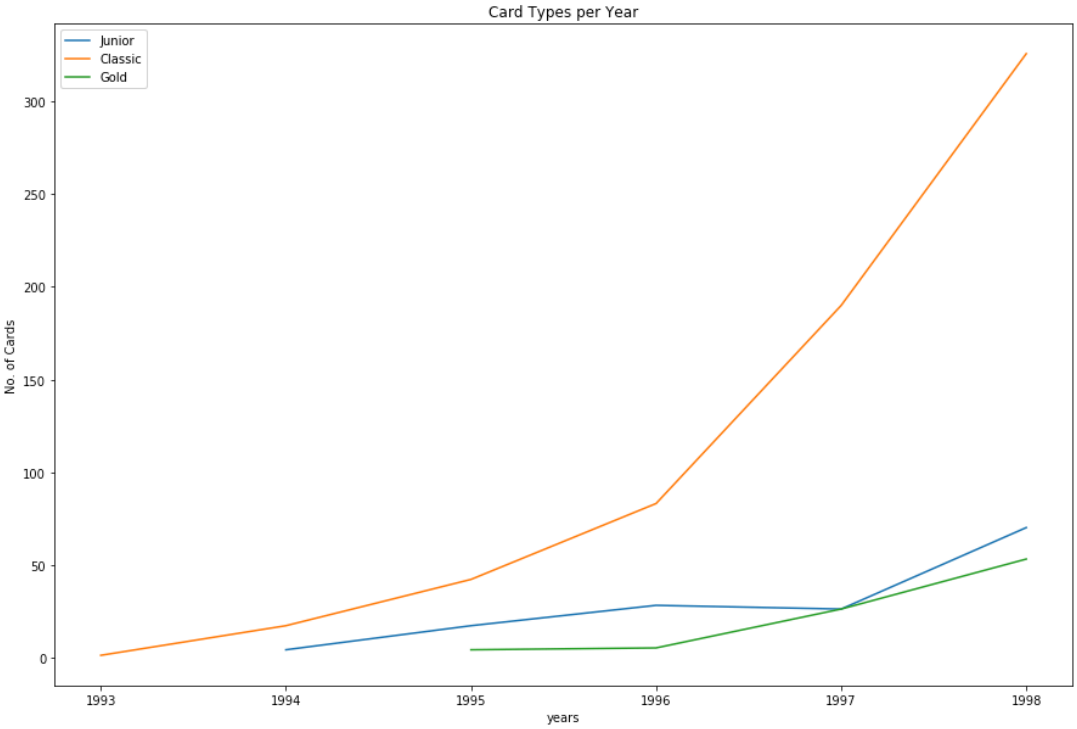
Το συγκεκριμένο στοιχείο σίγουρα είναι δυσχερές δεδομένου του σκοπού της ανάλυσης που είναι η βελτιστοποίηση της αξίας ζωής των πελατών, καθώς όταν το μεγαλύτερο ποσοστό των πελατών της τράπεζας κάτοικοι σε περιοχές με αυξημένη ανεργία, αυξάνεται η δυσχέρεια όσον αφορά τα περιθώρια αποκόμισης μεγαλύτερου κέρδους.



Εικόνα 19 Ποσοστά ανεργίας ανά περιοχή

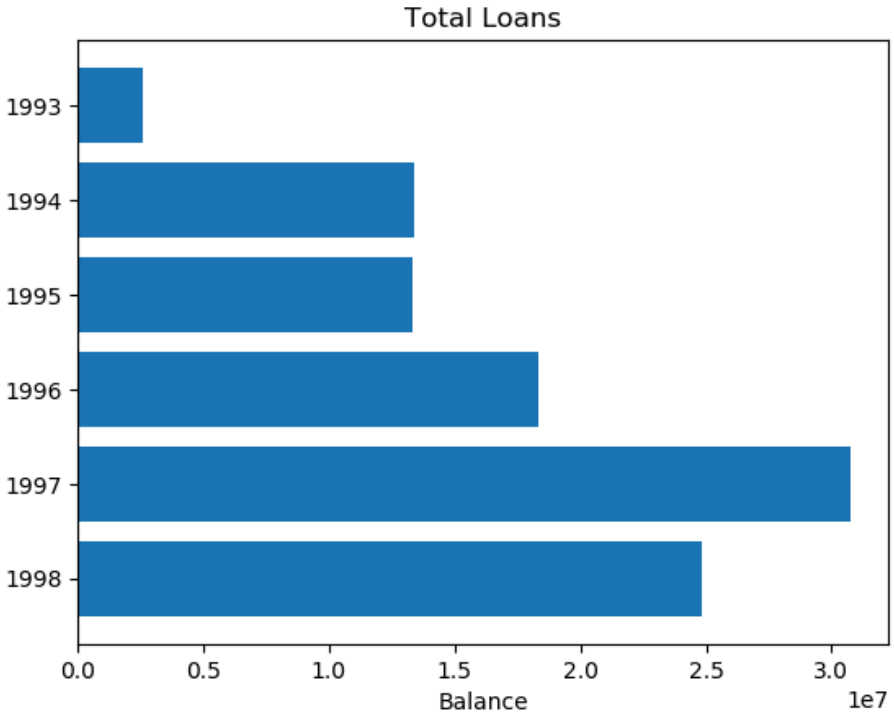
Όσον αφορά τη διάρθρωση των προϊόντων της τράπεζας, αυτά χωρίζονται σε δύο βασικές κατηγορίες, οι οποίες είναι τα δάνεια και οι πιστωτικές κάρτες. Συνεπώς, ένα σημαντικό ποσοστό των συναλλαγών που παράγονται προκύπτουν από τα δύο αυτά προϊόντα με αποτέλεσμα να είναι αυτά τα οποία καθορίζουν και τα κέρδη της τράπεζας.

Παρατηρείται κατά συνέπεια, ότι η τράπεζα προσφέρει τρία προγράμματα πιστωτικών καρτών. Το Junior για νεαρά άτομα, το classic και το gold το οποίο απευθύνεται στους πιο εύπορους πελάτες. Το gold πρόγραμμα είναι το πιο νέο από αυτά που προσφέρονται και εν γένει παρουσιάζεται αύξηση των συνδρομών στις πιστωτικές κάρτες τα τελευταία 3 χρόνια του χρονικού διαστήματος που εξετάζει το σύνολο δεδομένων. Συμπερασματικά, η ανάπτυξη η οποία παρουσιάζεται στον συγκεκριμένο τομέα αποτελεί ευκαιρία αποκόμισης μεγαλύτερου κέρδους για την τράπεζα.



Εικόνα 20 Διάγραμμα ανάπτυξης πιστωτικών καρτών

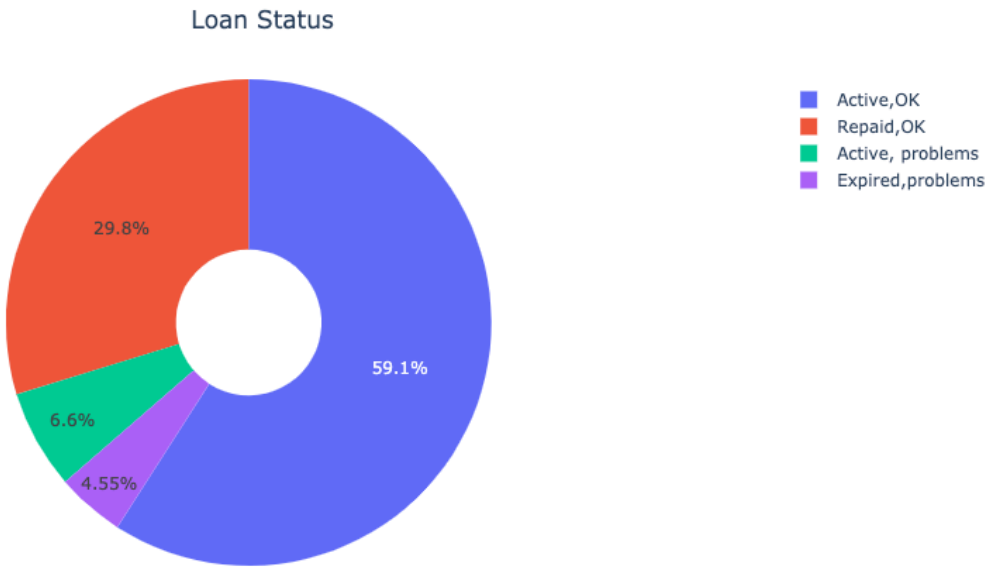
Παρόμοια αυξητική τάση παρουσιάζεται επίσης και στο προϊόν των δανείων καθώς τα τελευταία τρία έτη, η τράπεζα έχει αυξήσει τις χορηγήσεις της σε μεγάλο βαθμό. Κάτι το οποίο υποδηλώνει αυξημένα έσοδα από τόκους και προμήθειες. Η υπόθεση αυτή ενδυναμώνεται από το γεγονός ότι τα ενεργά δάνεια σε ποσοστό 59,1 % είναι πλήρως εξυπηρετήσιμα από τους δανειολήπτες.



Εικόνα 21 Χορηγήσεις ανά έτος

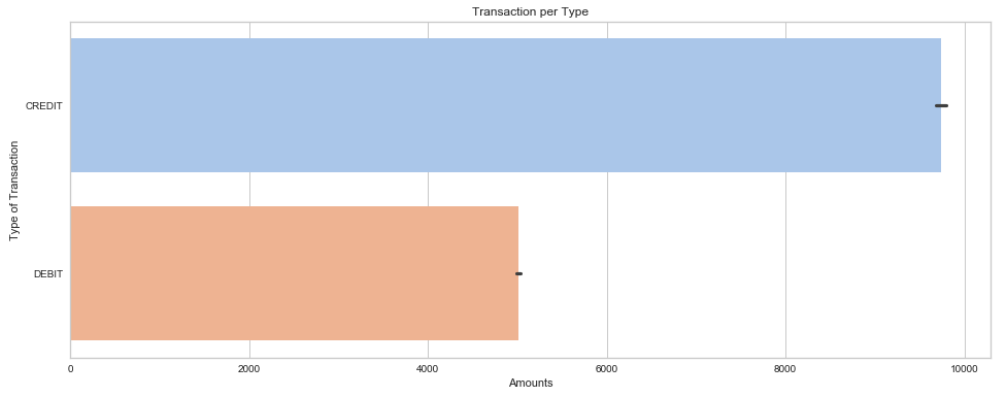
Αξιοσημείωτο είναι ωστόσο το ποσοστό των μη εξυπηρετούμενων δανείων (καθυστερημένα και ληγμένα δάνεια) το οποίο αγγίζει συνολικά το 11,15%. Το νούμερο αυτό αντικατοπτρίζει ένα διαχειρίσιμο ποσό καθυστερημένων χορηγήσεων το οποίο όμως αντιστοιχεί σε έναν σεβαστό αριθμό πελατών για τους οποίους η τράπεζα δεν θα έπρεπε να επενδύσει περεταίρω στην ανάπτυξη της πελατειακής τους σχέσης.

Τέλος, υπολογίζοντας τις συνολικά ποσά των χορηγήσεων που έχει διαθέσει διαχρονικά η τράπεζα, παρατηρείται ότι το 29,8% των συνολικών χορηγήσεων αφορά σε αποπληρωμένες χορηγήσεις. Κατά συνέπεια αυτό το νούμερο αντιστοιχεί σε ένα αξιόλογο αριθμό πελατών οι οποίοι έχουν ολοκληρώσει επιτυχώς τις υποχρεώσεις τους και ενδεχομένως η τράπεζα θα μπορούσε να τους προσελκύσει εκ νέου με νέα χορηγητικά προϊόντα καθώς αποτελούν ένα αξιόπιστο πελατολόγιο.



Εικόνα 22 Ανάλυση ανά κατάσταση δανείου

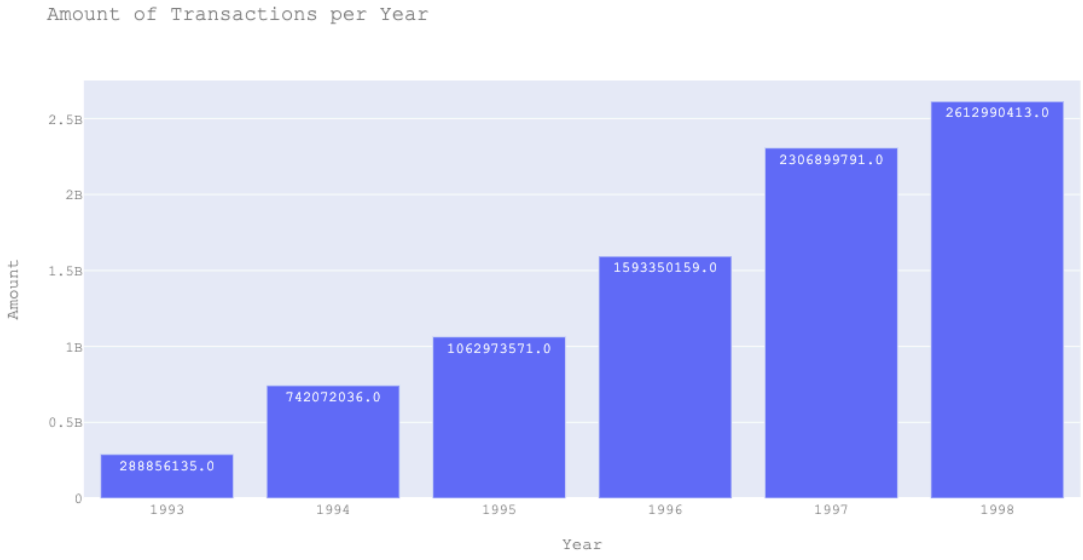
Σχετικά με τις συναλλαγές οι οποίες θα αποτελέσουν και το βασικό στοιχείο της τροφοδοσίας του τελικού μοντέλου, παρατηρείται ότι διαχρονικά οι συναλλαγές οι οποίες αφορούν σε πιστώσεις είναι σχεδόν διπλάσιες από τις χρεώσεις. Αυτό υποδεικνύει, ότι οι πελάτες διαχρονικά αποταμιεύουν στην τράπεζα οπότε μία στρατηγική η οποία θα μπορούσε να ακολουθήσει θα ήταν η προσφορά νέων αποταμιευτικών προγραμμάτων.



Εικόνα 23 Συναλλαγές ανά είδος

Αντίστοιχα, αφού οι αποταμιεύσεις αυξάνονται σημαίνει ότι υπάρχει ακόμα περισσότερο χρήμα διαθέσιμο είτε για νέες επενδύσεις της τράπεζας είτε για την προσφορά ακόμα περισσότερων δανείων προς τους πελάτες με σκοπό την αύξηση των εσόδων της.

Ενθαρρυντικό είναι επίσης το γεγονός ότι το συνολικό ύψος των συναλλαγών ανά έτος αυξάνεται σημαντικά φτάνοντας τα 2,6 δισεκατομμύρια το τελευταίο έτος σε σχέση με 289 εκατομμύρια που ήταν το πρώτο έτος. Τα στατιστικά αυτά υποδηλώνουν πέραν από την προφανή διαπίστωση ότι ο κύκλος εργασιών της τράπεζας έχει αυξηθεί σημαντικά, ότι σχεδόν όλες οι εργασίες της καταλήγουν να ολοκληρώνονται μέσω των συναλλαγών. Κατά συνέπεια, η εξόρυξη συμπερασμάτων από την ανάλυση των συναλλαγών θα μπορούσε να αποτελέσει ένα σημαντικό χαρακτηριστικό ως προς την λήψη αποφάσεων.



Εικόνα 24 Ποσά συναλλαγών ανά έτος

5.6 Εξαγωγή Χαρακτηριστικών

Έχοντας ολοκληρώσει την αρχική αναγνώριση των δεδομένων έχει πλέον παραχθεί μία σφαιρική εικόνα τόσο για την επιχειρησιακή εικόνα της τράπεζας όσο και για τα μεγέθη τα οποία περιέχονται στο σύνολο δεδομένων.

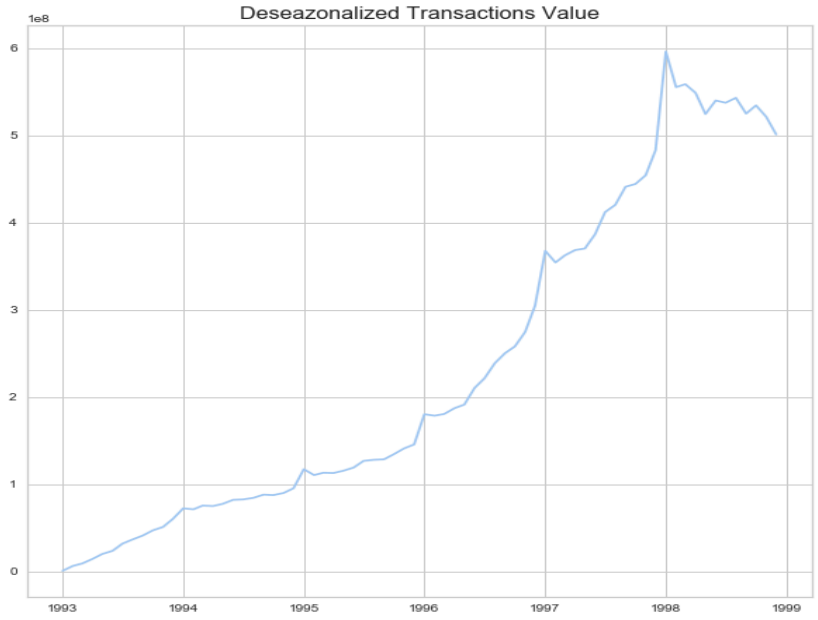
Ως εκ τούτου, είναι πλέον εφικτή η εξαγωγή χρήσιμων χαρακτηριστικών τα οποία είναι απαραίτητα για τη δημιουργία ενός αξιόπιστου και του ορθού μοντέλου. Στην προκειμένη περίπτωση η αρχιτεκτονική της ανάλυσης εστιάζει στις πραγματοποιθείσες συναλλαγές ως μέτρο ανάπτυξης των εσόδων της τράπεζας.

Βάσει αυτής της λογικής, δημιουργήθηκε η πρώτη μετρική συνάρτηση επί των συναλλαγών των πελατών η οποία είναι η εξής:

$$\boldsymbol{Αξία\ συναλλαγής = Ποσό\ συναλλαγής * Βάρος\ συναλλαγής}$$

Ως «αξία» συναλλαγής ορίζεται η βαρύτητα που έχει η εκάστοτε συναλλαγή αναλόγως του πελάτη που την πραγματοποιεί και του ποσού που την διακρίνει. Η βαρύτητα προσδίδεται μέσω της μεταβλητής "trans_index" η οποία υπολογίζει ένα ειδικό βάρος ανά πελάτη αναλόγως των συνολικών ποσών των συναλλαγών του και της συχνότητας αυτών.

Από την ανάγνωση του ανωτέρω γραφήματος, παρατηρείται ότι βάσει της νέας μετρικής «Αξία συναλλαγής» παρατηρείται ότι το κέρδος για την τράπεζα κινείται ανοδικά, ιδιαιτέρως τα τελευταία έτη. Αξίζει να σημειωθεί ότι τα δεδομένα της μετρικής παρουσίαζαν υψηλή περιοδικότητα σε εξαμηνιαία βάση, φαινόμενο το οποίο δικαιολογείται λόγω της φύσης τους. Συνεπώς πραγματοποιήθηκε αφαίρεση της περιοδικότητας αυτής ώστε να προκύψει η ξεκάθαρη τάση της μετρικής.

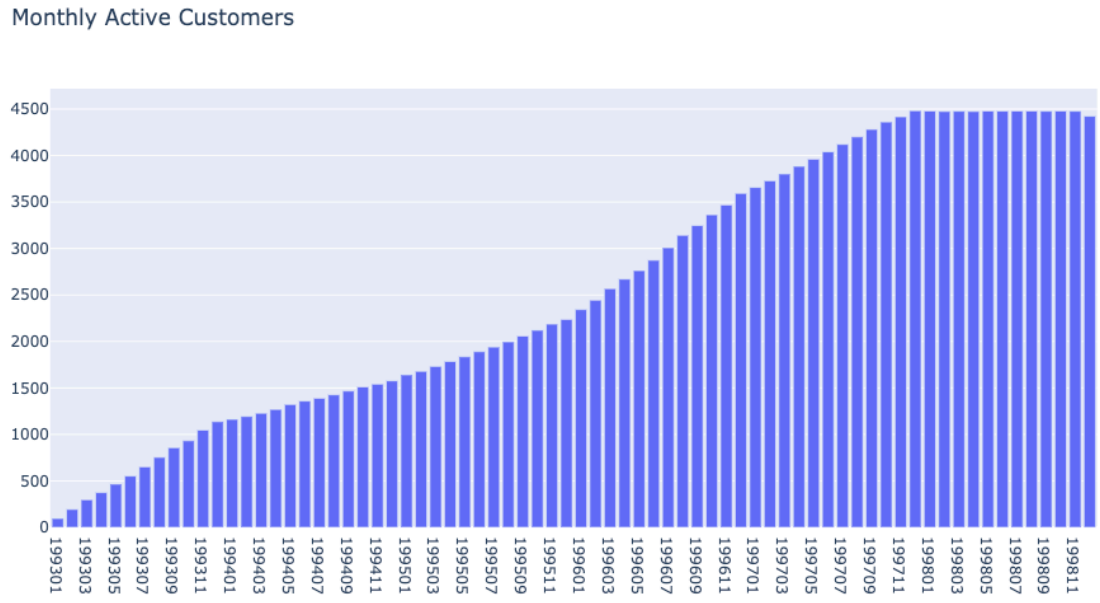


Εικόνα 25 Χρονοσειρά Αξίας συναλλαγών

Χρήσιμο στοιχείο επίσης αποτελεί η πληροφορία για τους πελάτες οι οποίοι ήταν ενεργοί σε μηνιαία βάση. Η μετρική αυτή θα αποτελέσει χρήσιμη ανάγνωση για την κινητικότητα του πελατολογίου της τράπεζας.

$$\textbf{Μηνιαίοι Πελάτες} = \frac{\textbf{Ενεργοί Πελάτες}}{\textbf{Μήνας Συναλλαγής}}$$

Παρατηρείται ότι το πλήθος των ενεργών πελατών ανά μήνα παρουσιάζει παρόμοια αυξητική τάση όπως εκείνη της μετρικής «Αξία συναλλαγών». Ως εκ τούτου, η σταδιακή αύξηση των πελατών ανά τα χρόνια δημιούργησε αύξηση των πραγματοποιθέντων συναλλαγών.

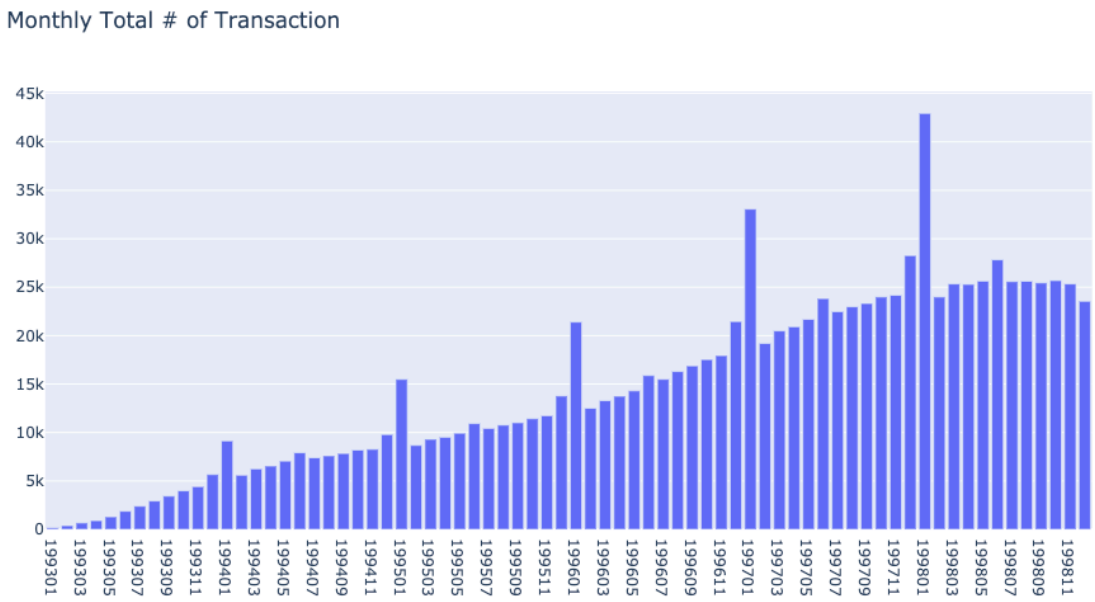


Εικόνα 26 Ενεργοί πελάτες ανά μήνα

Εν συνεχεία, δημιουργήθηκε μία μετρική με σκοπό την παρακολούθηση του κύριου αντικειμένου της ανάλυσης το οποίο είναι οι συναλλαγές. Πρόκειται για τον Αριθμό των μηνιαίων συναλλαγών.

$$\textbf{Μηνιαίες Συναλλαγές} = \frac{\textbf{Αριθμός Συναλλαγών}}{\textbf{Μήνας Συναλλαγής}}$$

Μέσω του πλήθους των συναλλαγών ανά μήνα, σε συνδυασμό με τους πελάτες οι οποίοι ήταν ενεργοί ανά μήνα θα είναι δυνατός στη συνέχεια ο εντοπισμός των πελατών οι οποίοι προσδίδουν μεγαλύτερη αξία στην ίδια την τράπεζα όσον αφορά το κέρδος.

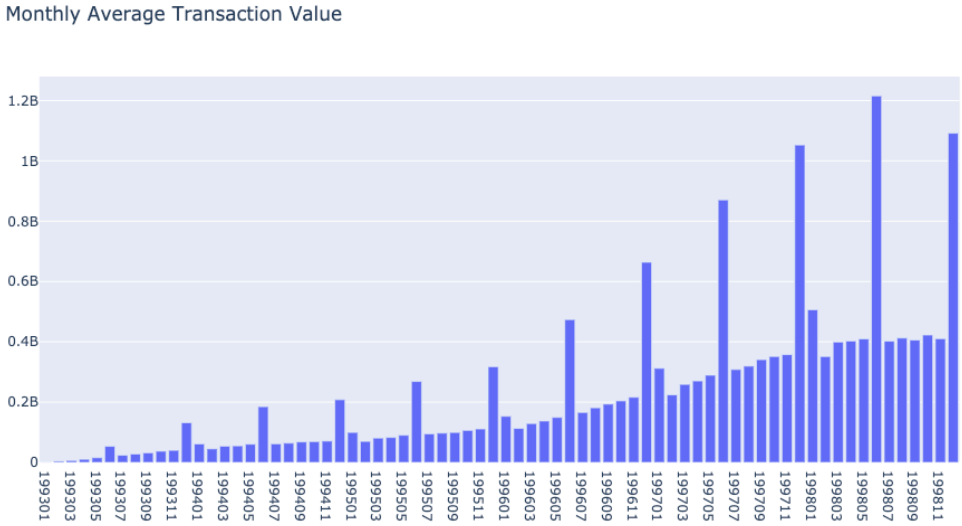


Εικόνα 27 Πλήθος συναλλαγών ανά μήνα

Από το ανωτέρω γράφημα προκύπτει η παρατήρηση ότι τον πρώτο μήνα του εκάστοτε έτους, ο αριθμός των συναλλαγών εκτοξεύεται και εκ των υστέρων ομαλοποιείται. Το φαινόμενο αυτό εξηγείται επιχειρησιακά από το γεγονός ότι ο πρώτος μήνας του έτους είναι αυτός κατά τον οποίο υπάρχει η μεγαλύτερη κίνηση στην αγορά, τόσο λόγω των εορτών όσο και λόγω των οικονομικών υποχρεώσεων που προκύπτουν. Ωστόσο, παρατηρείται μία σταδιακή αυξητική τάση στις μηνιαίες συναλλαγές, δεδομένο το οποίο προκύπτει και από τις 2 προαναφερθείσες μετρικές.

Η επόμενη μετρική είναι εξαρτώμενη από την αρχική (Αξία Συναλλαγής) καθώς υπολογίζει την μέση μηνιαία αξία των συναλλαγών.

$$\textbf{Μέση Αξία Συναλλαγών} = \frac{\textbf{Σύνολο Αξίας Συναλλαγών}}{\textbf{Πλήθος συναλλαγών}}$$
$$\textbf{Μήνας Συναλλαγής}$$



Εικόνα 28 Μέση Αξία Συναλλαγών ανά μήνα

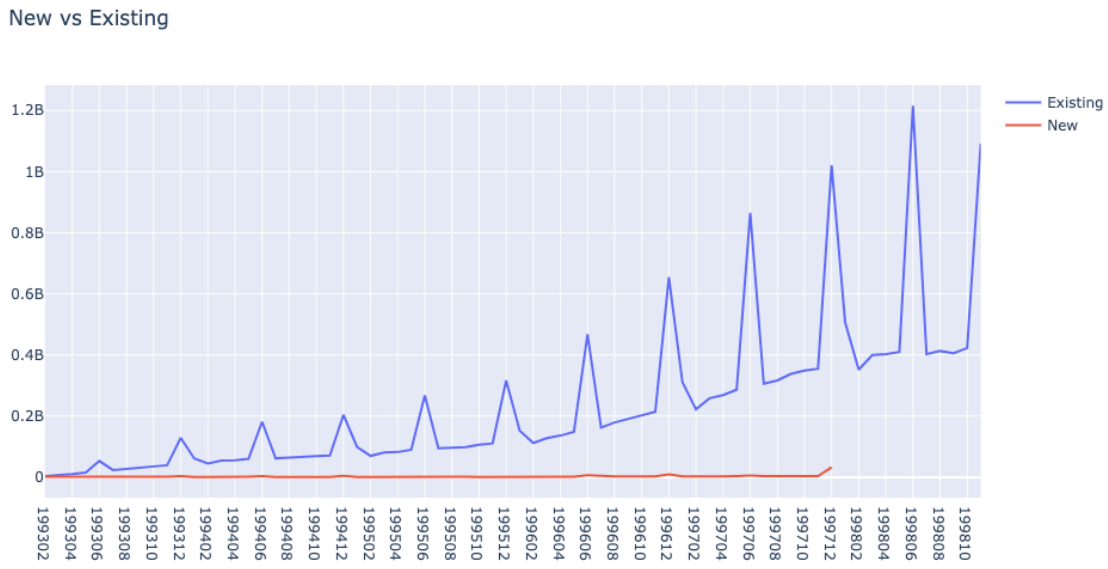
Κατόπιν του υπολογισμού της ανωτέρω μετρικής προκύπτει το εξής ενδιαφέρον εύρημα. Αν και οι περισσότερες συναλλαγές πραγματοποιούνται κατά τους πρώτους μήνες του έτους, οι συναλλαγές οι οποίες προσδίδουν μεγαλύτερη αξία για την τράπεζα είναι αυτές οι οποίες πραγματοποιούνται κατά τα κλεισίματα των εξαμήνων. Γεγονός το οποίο υποδεικνύει ότι οι συναλλαγές οι οποίες σχετίζονται με πάγιες πληρωμές είναι αυτές οι οποίες είναι περισσότερο κερδοφόρες.

Το ανωτέρω συμπέρασμα αποτελεί ένα από τα πρώτα στοιχεία τα οποία θα μπορούσαν αποτελέσουν σημεία καθορισμού της στρατηγικής της τράπεζας καθώς με αυτόν τον τρόπο είναι διακριτή και η ποιότητα των συναλλαγών πέραν των απλών ενδείξεων που προκύπτουν μέσα από τις απλές μετρήσεις.

Η επόμενη μετρική έχει στο επίκεντρο τους ίδιους τους πελάτες και ως σκοπό έχει την μέτρηση των νέων πελατών σε σχέση με τους υφιστάμενους.

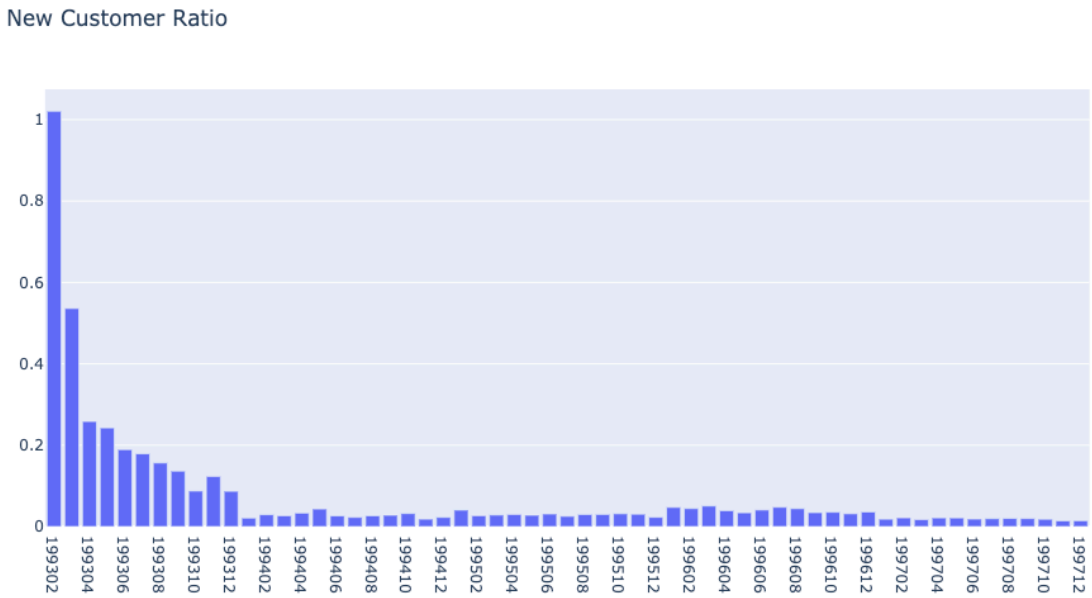
Ως νέος πελάτης κατά συνέπεια ορίζεται αυτός ο οποίος πραγματοποίησε την πρώτη του συναλλαγή στην τράπεζα μέσα σε ένα προκαθορισμένο χρονικό διάστημα.

$$\boldsymbol{Ρυθμός\ Αύξησης\ Πελατών} = \frac{\boldsymbol{Νέοι\ Πελάτες}}{\boldsymbol{Παλαιοί\ Πελάτες}}$$



Εικόνα 29 Αξία Συναλλαγών νέων πελατών vs υφιστάμενων

Το ανωτέρω γράφημα εξετάζει τις δύο παραμέτρους από τις οποίες προκύπτει η νέα μετρική και παρατηρείται ότι διαχρονικά η μεγαλύτερη αξία στις συναλλαγές προέρχεται από τους υφιστάμενους πελάτες και μάλιστα με αυξητική τάση, ενώ η αξία που παράγεται από τους νέους πελάτες του κάθε μήνα είναι σχεδόν αμελητέα. Ωστόσο, η σταθερότητα των συναλλαγών των νέων πελατών υποδεικνύει το σταθερό ρυθμό τους ανά τους μήνες, κάτι το οποίο επεξηγείται καλύτερα από την κατωτέρω εικόνα.



Εικόνα 30 Ρυθμός ανάπτυξης νέων πελατών

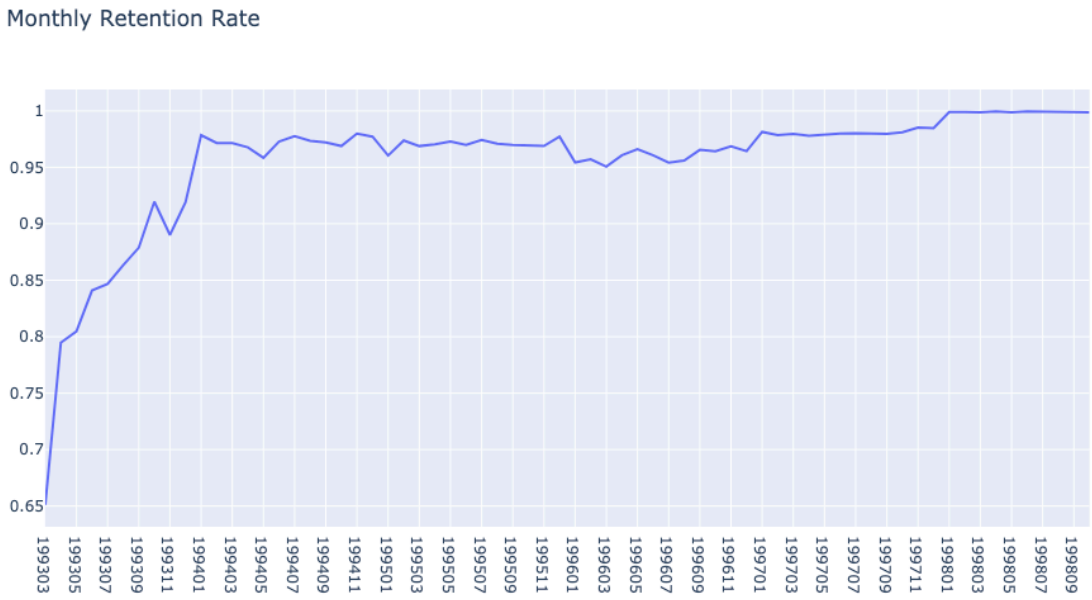
Όπως παρατηρείται, αν εξαιρεθούν οι πρώτοι μήνες όπου ο ρυθμός ήταν υψηλός καθώς όλοι οι πελάτες θεωρούνταν νέοι κατά τις πρώτες τους συναλλαγές, φαίνεται ότι διαχρονικά ο ρυθμός ανάπτυξης του πελατολογίου παραμένει σταθερά μικρός. Γεγονός το οποίο σημαίνει ότι για την τράπεζα είναι σημαντικό να καταφέρνει να διατηρεί τους πελάτες της έναντι του ανταγωνισμού ενώ αντίστοιχα υποδηλώνεται και το πόσο σημαντικός είναι ο υπολογισμός του CLV με σκοπό την εξεύρεση του σημαντικού μεριδίου των πελατών για την επιχείρηση.

Πάνω σε αυτήν την λογική, δημιουργήθηκε η επόμενη μετρική η οποία σχετίζεται με τον ρυθμό διατήρησης των υφιστάμενων πελατών.

$$\boldsymbol{Ρυθμός\ Διατήρησης} = \frac{\boldsymbol{Υφιστάμενοι\ πελάτες\ προηγούμενου\ μήνα}}{\boldsymbol{Συνολικούς\ ενεργούς\ πελάτες}}$$

Ο ρυθμός διατήρησης των πελατών αποτελεί ένα νευραλγικό μέγεθος για όλες τις επιχειρήσεις, πόσο μάλλον για μία τράπεζα η οποία όπως αποδεικνύεται και στην παρούσα ανάλυση, είναι πιο σπάνιο να αποκτά μαζικά νέους πελάτες από το να διατηρεί τους υφιστάμενους.

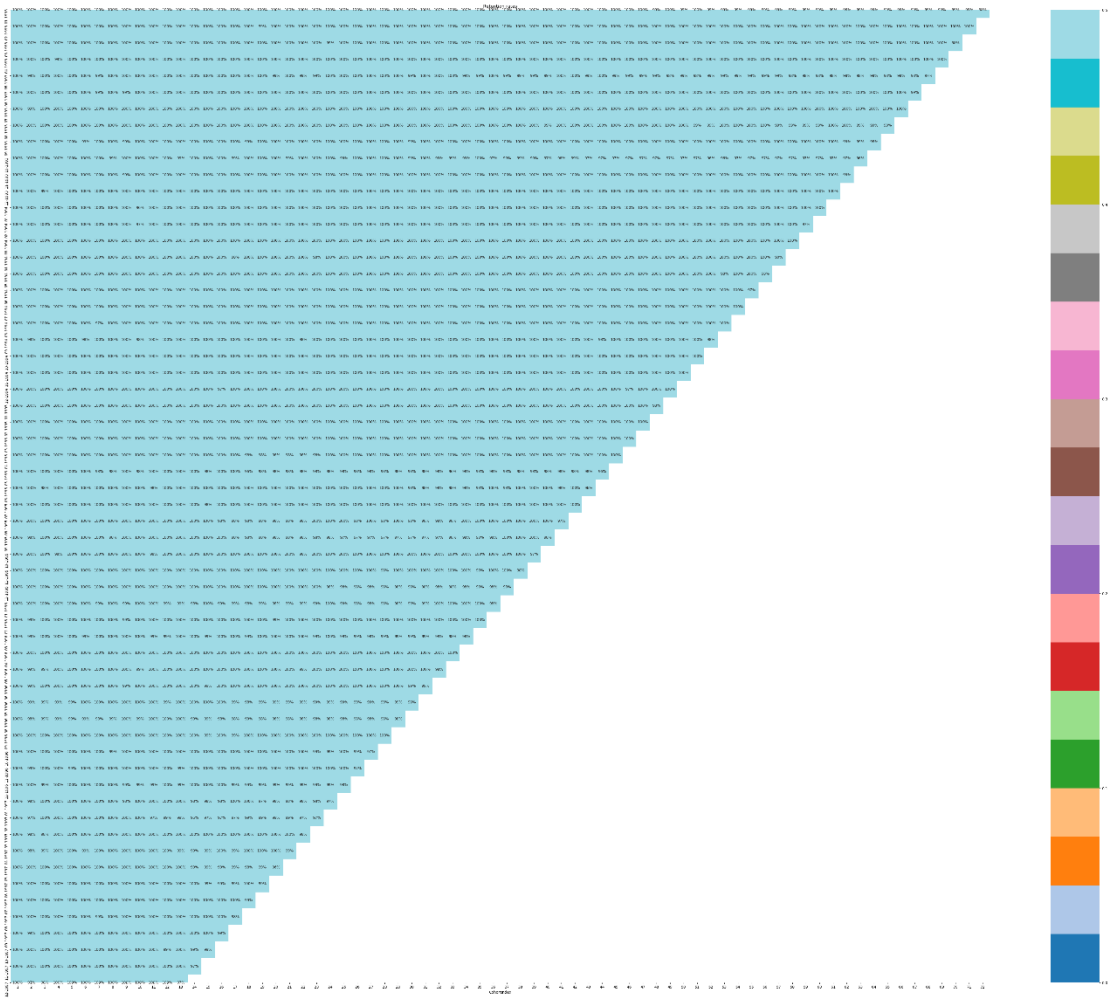
Επίσης, η παρακολούθηση αυτού του μεγέθους είναι απαραίτητη για τις μελλοντικές στρατηγικές αναπτύξεως μιας επιχείρησης.



Εικόνα 31 Ρυθμός ανάπτυξης πελατών

Βάσει του ανωτέρω γραφήματος, ο ρυθμός στην προκειμένη περίπτωση κυμαίνεται σε υψηλά επίπεδα διαχρονικά καθώς τους περισσότερους μήνες πλησιάζει την μέγιστη τιμή που είναι το 1. Ως εκ τούτου, καθίσταται εξαιρετικά σημαντικός στόχος για την τράπεζα να συνεχίζει να διατηρεί τους πελάτες της ώστε εξασφαλίζεται η εύρυθμη λειτουργία της. Το συγκεκριμένο σκεπτικό σε συνδυασμό με την εξεύρεση των σημαντικών πελατών της, βάσει του CLV θα αποτελέσει εφαλτήριο για μία επιτυχημένη εμπορική στρατηγική.

Σε αυτό το πλαίσιο, θα επιχειρηθεί μία δεύτερη προσέγγιση υπολογισμού του ρυθμού διατήρησης των πελατών μέσω της οποίας θα είναι εφικτή η μέτρηση του ποσοστού των πελατών οι οποίοι επανέλαβαν κάποια συναλλαγή τους επόμενους μήνες ύστερα από την αρχική τους (Cohort Based Retention Rate). Ωστόσο, όπως είναι εμφανές και στο παρακάτω γράφημα, τα ποσοστά των πελατών που διατηρούν ενεργή τη σχέση τους με την τράπεζα είναι εξαιρετικά υψηλά, σε βαθμό τέτοιο μάλιστα που επιτρέπεται το συμπέρασμα ότι σχεδόν όλοι οι πελάτες, από την στιγμή που πραγματοποιούν την πρώτη τους συναλλαγή και έπειτα, παραμένουν στην ίδια τράπεζα.



Εικόνα 32 Retention rate Heatmap

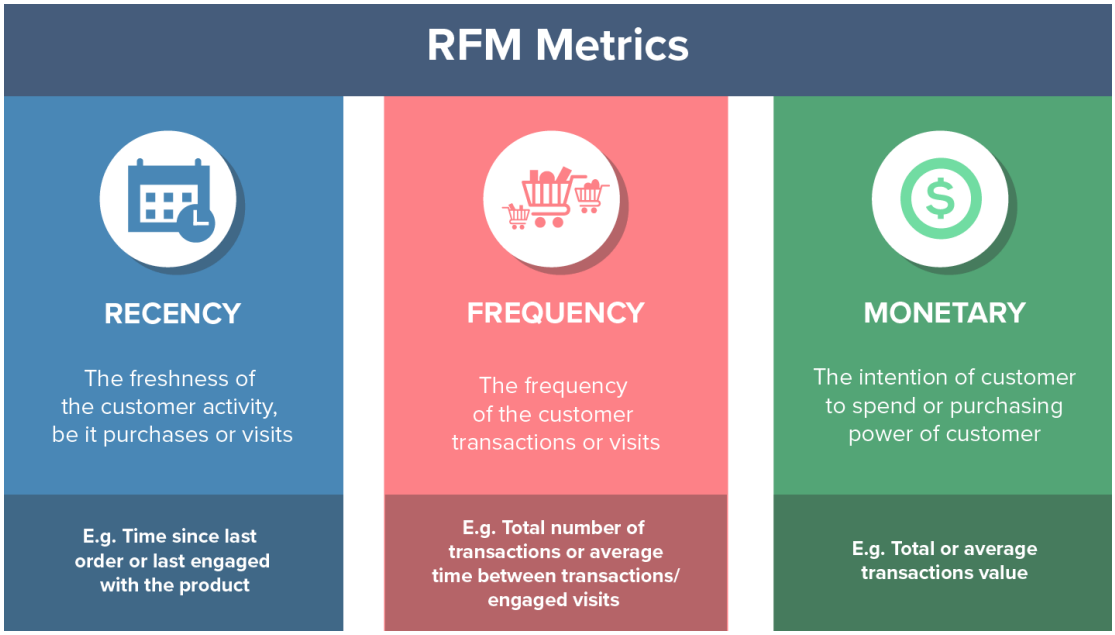
5.7 Κατηγοριοποίηση Πελατών με τη χρήση μεθόδου RFM

Στο προηγούμενο κεφάλαιο πραγματοποιήθηκε η βασική ανάλυση των παραγόντων που επηρεάζουν την κερδοφορία της τράπεζας αλλά και εν γένει την λειτουργία της. Σε αυτό το σημείο συνεπώς, έχει δημιουργηθεί μία εικόνα για τα δεδομένα μέσω της οποίας είναι εφικτή μίας εις βάθος ανάλυση του πελατολογίου της τράπεζας.

Έχοντας ως γνώμονα την πεποίθηση ότι οι πελάτες είναι αυτοί που διαμορφώνουν την εικόνα των κερδών μίας επιχείρησης και προσεγγίζοντας παράλληλα την τράπεζα ως μορφή λειτουργίας μίας εμπορικής επιχείρησης, στο παρόν κεφάλαιο θα επιχειρηθεί η κατηγοριοποίηση με βάση την κερδοφορία που αποφέρουν και την αξία που προσδίδουν στον οργανισμό.

Η πελατοκεντρική αυτή προσέγγιση σχετικά με την μεγιστοποίηση του κέρδους χαρακτηρίζεται από μία αντίστροφή λογική, εξετάζοντας το αίτιο των κερδών το οποίο δεν είναι άλλο από το πελατολόγιο. Με βάση αυτή τη λογική, δεν μπορεί μία επιχείρηση να αντιμετωπίζει όλους του πελάτες με την ίδια νοοτροπία και την ίδια

στρατηγική. Ως εκ τούτου, απαιτείται προσωποποιημένη στρατηγική η οποία θα βασίζεται σε μεθοδολογίες κατηγοριοποίησης.



Εικόνα 33 Βασικές μετρικές μεθόδου RFM (Πηγή: (Makhija, 2020))

Στην συγκεκριμένη ανάλυση θα εφαρμοστεί στα δεδομένα μία υλοποίηση της μεθόδου RFM η οποία αποτελεί μία από τις πιο ασφαλείς και αποδοτικές μεθόδους. Σύμφωνα με την συγκεκριμένη μεθοδολογία, οι πελάτες διαχωρίζονται σε 3 βασικές κατηγορίες οι οποίες ορίζονται ως εξής:

Χαμηλής Αξίας: Συμμετέχουν πελάτες οι οποίοι είναι λιγότερο ενεργοί σε σχέση με τους υπόλοιπους, με χαμηλή συχνότητα συναλλαγών και αμελητέο κέρδος.

Μεσαίας Αξίας: Πρόκειται για την κατηγορία των πελατών η οποία χαρακτηρίζεται από μεσαίο αποδιδόμενο κέρδος και εν γένει μέτριες επιδόσεις όσον αφορά την αλληλεπίδρασή τους με την επιχείρηση.

Υψηλής Αξίας: Ως υψηλής αξίας πελάτες, χαρακτηρίζονται εκείνοι οι οποίοι είναι οι πιο σημαντικοί για μία επιχείρηση. Οι συναλλαγές τους είναι συχνές και επικερδής, ενώ αποτελούν κεφάλαιο για την τράπεζα και την εξέλιξή της.

Όπως έχει αναλυθεί και στο κεφάλαιο 4, η ονομασία της μεθόδου RFM προέρχεται από τα αρχικά των 3 βασικών παραγόντων της τα οποία είναι **Recency** (Εγγύτητα), **Frequency** (Συχνότητα), **Monetary** (Νομισματική Αξία). Ως εκ τούτου είναι απαραίτητος ο υπολογισμών των μεγεθών αυτών. Στην προκειμένη περίπτωση η υλοποίηση της μεθόδου θα πραγματοποιηθεί με τη βοήθεια εργαλείων και τεχνικών της μηχανικής μάθησης. Συγκεκριμένα, θα εφαρμοστούν μοντέλα μη επιβλεπόμενης μάθησης όπως είναι ο αλγόριθμος K-means με σκοπό την αναγνώριση την βασικών ομάδων (clusters) του κάθε μεγέθους.

Για τους σκοπούς της εστιασμένης ανάλυσης που θα ακολουθήσει με σκοπό της εφαρμογής της μεθόδου RFM καθώς και για τον τελικό υπολογισμό του δείκτη Customer Lifetime Value, δημιουργήθηκε ένα στοχευμένο υποσύνολο της Βάσης

Δεδομένων στο οποίο συμμετέχουν οι μεταβλητές εκείνες που αντιπροσωπεύουν τα μεγέθη τα οποία σχετίζονται τόσο με τις απαραίτητες πελατειακές πληροφορίες όσο και με τα ποσοτικά και ποιοτικά μεγέθη των συναλλαγών.

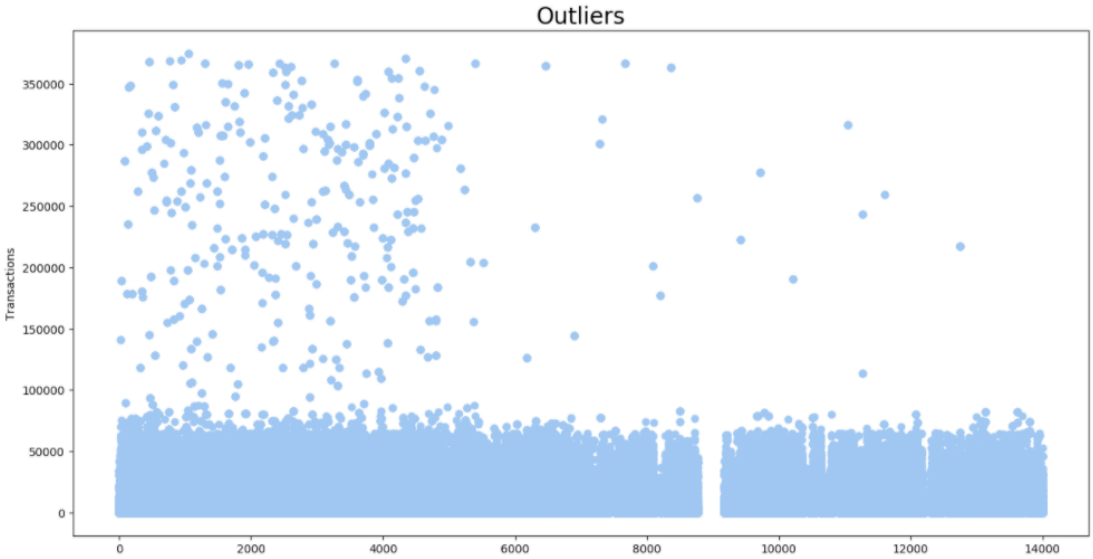
Συγκεκριμένα, το υποσύνολο αυτό δημιουργήθηκε ύστερα από την σύνταξη των κατωτέρω ερωτημάτων SQL Queries μέσω των οποίων δημιουργήθηκε το τελικό Dataset.

```
df_rfm1 = """
             select cl.client_id,
                    case when
                     sum(tr.amount) = 0 then null
                    else
                        case when (floor (max(tr.amount)/sum(tr.amount) * 100)) = 0 then 1
                        else  floor(max(tr.amount)/sum(tr.amount) * 100)
                        end
                    end as trans_index

             from transactions as tr
             left join dbo.dispositions di on tr.account_id_ = di.account_id
             left join dbo.clients cl on di.client_id = cl.client_id --where cl.client_id is null
             left join dbo.accounts ac on di.account_id = ac.account_id
             left join districts ds on ac.district_id = ds.district_id
             where di.type = 'OWNER'
             group by cl.client_id

            """
rfm1 = pd.read_sql(df_rfm1,conn)
```

```
df_rfm2 = """
               select distinct      tr.transaction_id
                                    ,tr.type
                                    ,tr.trans_date
                                    ,tr.amount
                                    ,region
                                    ,cl.client_id

               from transactions as tr
               left join dbo.dispositions di on tr.account_id_ = di.account_id
               left join dbo.clients cl on di.client_id = cl.client_id --where cl.client_id is null
               left join dbo.accounts ac on di.account_id = ac.account_id
               left join districts ds on ac.district_id = ds.district_id
               where di.type = 'OWNER'
            """
rfm2 = pd.read_sql(df_rfm2,conn)
```

Εικόνα 34 SQL Queries

Το τελικό dataset στο οποίο όπως είναι φυσικό συμπεριλαμβάνονται τα μεγέθη των συναλλαγών, πραγματοποιήθηκαν έλεγχοι ως προς την πληρότητά του και συγκεκριμένα για το ενδεχόμενο ελλιπών και ακραίων τιμών. Αξίζει να σημειωθεί ότι το διάγραμμα ακραίων τιμών αποκάλυψε την ύπαρξη πληθώρας αυτών όσον αφορά την μεταβλητή του πλήθους των συναλλαγών. Ωστόσο, κρίθηκε σκόπιμο δεδομένου του σεναρίου της ανάλυσης, να μην πραγματοποιηθεί η αφαίρεσή τους, καθώς τα πρωτότυπα στοιχεία είναι αυτά που θα υπολογίσουν ακριβέστερα τόσο το μοντέλο RFM όσο και τα μοντέλα μηχανικής μάθησης του CLV.



Εικόνα 35 Outliers

5.7.1 Recency (Εγγύτητα)

Για τον υπολογισμό του Recency είναι απαραίτητη η εξεύρεση της πιο πρόσφατης ημερομηνίας συναλλαγής του κάθε πελάτη και εν συνεχεία ο υπολογισμός των ημερών αδράνειας. Ως αδράνεια ορίζεται το χρονικό διάστημα που μεσολάβησε από την πιο πρόσφατη μέχρι την προηγούμενη συναλλαγή. Ύστερα από τον προσδιορισμό του αριθμού αυτού των ημερών, θα εφαρμοστεί ο αλγόριθμος συσταδοποίησης K-means ώστε να δημιουργηθεί μία βαθμολογία βάσει της οποίας θα προκύψει ο υπολογισμός του Recency.

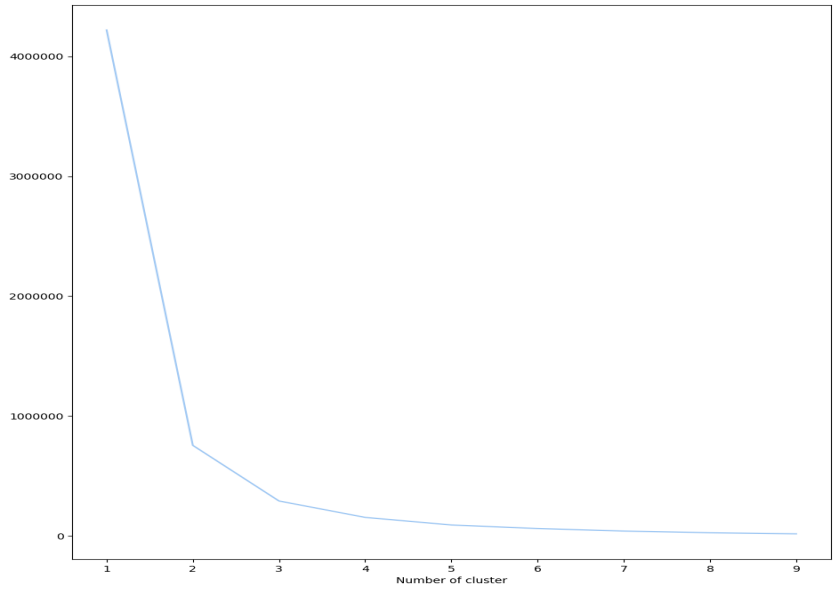
count	4500.000000
mean	2.599111
std	30.625054
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	858.000000

Εικόνα 36 Περιγραφική στατιστική Recency

Στα πλαίσια του υπολογισμού του Recency, προέκυψε ότι η συντριπτική πλειοψηφία των πελάτων διατηρούν πολύ στενή σχέση με την τράπεζα καθώς η μέση τιμή των ημερών που μεσολαβούν μεταξύ των συναλλαγών ενός πελάτη δεν ξεπερνάει τις 3 ημέρες.

Εν συνεχεία, ακολούθησε η εφαρμογή του αλγόριθμού K-means όπως έχει περιγραφεί ανωτέρω έτσι ώστε να δημιουργηθεί μία βαθμολογία Εγγύτητας μέσω της ομαδοποίησης των πελατών. Ωστόσο, είναι απαραίτητος ο προσδιορισμός του αριθμού των συστάδων που θα δημιουργηθούν. Για τον σκοπό αυτό εφαρμόστηκε η μέθοδος Elbow Rule η οποία ανέδειξε τον βέλτιστο προτεινόμενο αριθμό clusters.

Ο «κανόνας του αγκώνα» αποτελεί μία προσέγγιση προσδιορισμού του αριθμού των συστάδων μέσω του κέρδους της πληροφορίας που προστίθεται στις κλάσεις και βάσει του ποσοστού της εξηγούμενης διακύμανσης.



Εικόνα 37 Elbow Method

Βάσει του ανωτέρω γραφήματος, προτείνεται η δημιουργία 3 συστάδων. Ωστόσο βάσει δοκιμών και ελέγχων, προέκυψε ότι η καλύτερη απόδοση του αλγορίθμου επιτυγχάνεται με την χρήση 4 συστάδων.

RecencyCluster	count	mean	std	min	25%	50%	75%	max
0	5.0	757.200000	93.213196	661.0	670.0	752.0	845.0	858.0
1	9.0	347.444444	101.692073	207.0	277.0	334.0	440.0	509.0
2	149.0	30.865772	15.443733	16.0	21.0	28.0	33.0	112.0
3	4337.0	0.042426	0.682915	0.0	0.0	0.0	0.0	15.0

Εικόνα 38 Recency K-Means

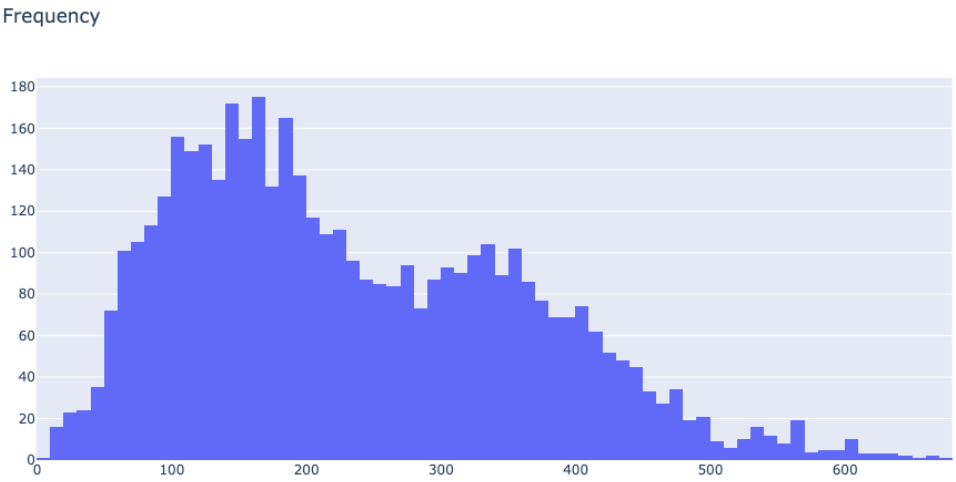
Σημειώνεται ότι ύστερα την διαδικασία εφαρμογής του αλγορίθμου K-means, εφαρμόστηκε η user defined function "**trans_cluster**", δεδομένου ότι ο K-means εξ ορισμού αναθέτει στις συστάδες έναν αριθμό, όχι όμως με διατεταγμένο τρόπο. Με την "**trans_cluster**" ωστόσο, τα clusters ταξινομούνται με βάση την αποδοτικότητά τους.

Κατά συνέπεια, παρατηρείται από τον ανωτέρω πίνακα ότι οι περισσότεροι πελάτες συγκεντρώνονται στο cluster 3 το οποίο μάλιστα είναι αυτό με το οποίο περιέχει τους χαμηλότερους χρόνους αδράνειας όπως αυτοί έχουν οριστεί από την μέθοδο RFM.

Βάσει των συστάδων που δημιουργήθηκαν αποδεικνύεται ότι η συντριπτική πλειοψηφία των πελατών τηρούν εγγύς σχέση με την τράπεζα καθώς ολοκληρώνει συναλλαγές σε τακτά χρονικά διαστήματα ενώ είναι ελάχιστος ο αριθμός εκείνων που αλληλοεπιδρούν λιγότερο με αυτήν.

5.7.2 Frequency (Συχνότητα)

Για τον υπολογισμό της συχνότητας θα ακολουθηθεί η ίδια διαδικασία συσταδοποίησης όπως και για τον προσδιορισμό του Recency. Ως εκ τούτου, θα δημιουργηθούν τα αντίστοιχα σχετικά clusters τα οποία θα συμπεριλαμβάνουν τον συνολικό αριθμό των συναλλαγών του εκάστοτε πελάτη. Σημειώνεται ότι η καταμέτρηση των συναλλαγών θα πραγματοποιηθεί επί των ημερομηνιών της εκάστοτε συναλλαγής, οπότε με αυτόν τον τρόπο θα μετρηθεί σε πόσες διαφορετικές ημερομηνίες ένας πελάτης αλληλοεπίδρασε με την τράπεζα.



Εικόνα 39 Διάγραμμα Frequency

Σε μία πρώτη οπτική απεικόνιση των συνολικών συναλλαγών των πελατών παρατηρείται ότι οι περισσότεροι από αυτούς πραγματοποιούν έναν αρκετά υψηλό αριθμό με αποτέλεσμα η συχνότητά τους να καθορίζεται σε υψηλά επίπεδα. Το γεγονός αυτό θα καταστήσει την διάκριση των επικερδών πελατών λίγο πιο περίπλοκη, ωστόσο αυτό θα πραγματοποιηθεί μέσω της εφαρμογής του αλγορίθμου ομαδοποίησης.

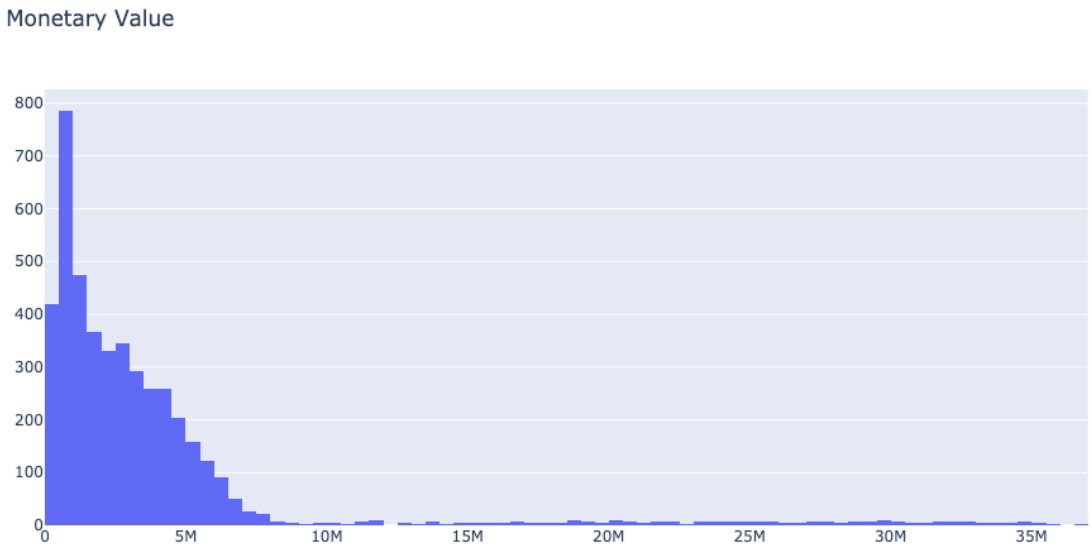
Όπως παρατηρείται από τον κατωτέρω πίνακα τα μεγαλύτερα σκορ συχνοτήτων συγκεντρώνονται στο cluster 3, γεγονός το οποίο υποδεικνύει ότι είναι περιορισμένος ο αριθμός των πελατών που ανήκουν στην υψηλότερη κατηγορία. Αξίζει να σημειωθεί παράλληλα ότι και το cluster 2 αντιπροσωπεύει πελάτες με πολλές συναλλαγές, άρα με υψηλή συχνότητα.

FrequencyCluster	count	mean	std	min	25%	50%	75%	max
0	1365.0	99.703297	32.372770	9.0	76.0	104.0	127.0	148.0
1	1405.0	197.926690	31.800521	149.0	170.0	194.0	224.0	262.0
2	1156.0	326.706747	36.596946	263.0	297.0	327.0	357.0	393.0
3	574.0	460.740418	60.630980	394.0	414.0	441.0	488.0	675.0

Εικόνα 40 K-Means Frequency

5.7.3 Monetary (Νομισματική Αξία)

Για τον υπολογισμό της νομισματικής αξίας του πελάτη, ή αλλιώς το κέρδος που επιφέρει, δημιουργήθηκε μία νέα μεταβλητή στο σύνολο δεδομένων η οποία ονομάστηκε Revenue και προέκυψε από το γινόμενο του ποσού της εκάστοτε συναλλαγής σε συνδυασμό με τον δείκτη "trans_index", η δημιουργία του οποίου περιγράφεται στην παράγραφο 5.4.



Εικόνα 41Διάγραμμα Monetary Value

Ερμηνεύοντας την οπτική απεικόνιση των τιμών της νομισματικής αξίας των πελατών, παρατηρείται ότι σε αυτή την κατηγορία διακρίνεται ο περιορισμένος αριθμός των πελατών οι οποίοι επιφέρουν υψηλό κέρδος. Οι πελάτες αυτοί είναι που στη συνέχεια θα ενταχθούν στην ομάδα της «Υψηλής Αξίας» όπως αυτή έχει οριστεί ανωτέρω.

RevenueCluster	count	mean	std	min	25%	50%	75%	max
0	2695.0	1.325174e+06	7.900045e+05	152349.0	657712.0	1121712.0	1977927.00	2960648.0
1	1533.0	4.601437e+06	1.254316e+06	2961014.0	3595081.0	4351492.0	5311556.00	10782681.0
2	120.0	1.716289e+07	3.501378e+06	11092962.0	14063437.5	17620216.5	20153817.75	22735495.0
3	152.0	2.893175e+07	3.623758e+06	23124300.0	25645253.0	28994680.5	31918159.00	36604575.0

Εικόνα 42 K-Means Revenue

Με την ίδια λογική κατά την οποία έχει εφαρμοστεί ο αλγόριθμος συσταδοποίησης στο Recency και στο Frequency, εφαρμόστηκε και στο Revenue. Ως εκ τούτου προκύπτει ότι το cluster 3 είναι αυτό που συμπεριλαμβάνει τους πελάτες οι οποίοι αποδίδουν το μεγαλύτερο κέρδος. Ωστόσο, ο αριθμός τους είναι περιορισμένος, γεγονός το οποίο είναι αναμενόμενο.

5.7.4 Συνολική Βαθμολόγηση

Ύστερα από τον υπολογισμό των τριών μεγεθών της μεθόδου RFM, είναι δυνατός ο υπολογισμός μίας συνολικής βαθμολογίας για τον εκάστοτε πελάτη. Η διαδικασία τελικής βαθμολόγησης βασίζεται πάνω στις τρεις διακριτές εφαρμογές του αλγορίθμου K-means στα μεγέθη της μεθόδου RFM (Recency – Frequency – Monetary Value).

Η λογική της δημιουργίας αυτού του συστήματος αξιολόγησης χρησιμοποιεί το άθροισμα του αριθμού των clusters στα οποία ανήκει ο εκάστοτε πελάτης. Με αυτόν τον τρόπο, το κάθε άτομο λαμβάνει έναν αριθμό σκορ από το 0 έως το 9.

Δεδομένης της εφαρμογής της ιεραρχημένης δημιουργίας των συστάδων και στις τρεις περιπτώσεις η οποία πραγματοποιείται μέσω της function "**trans_cluster**", όσο πιο μικρός είναι ο αριθμός του cluster τόσο λιγότερο σημαντική είναι η αξία που αποδίδει ο πελάτης στην τράπεζα.

Κατά συνέπεια, το άθροισμα του αριθμού των clusters στα οποία ανήκει το κάθε άτομο, αποτελεί την συνολική απεικόνιση της σημαντικότητας των καταναλωτών. Πιο συγκεκριμένα, στον κατωτέρω πίνακα παρατηρείται η κατανομή των τριών μεγεθών της RFM στις 10 βαθμίδες του νεοσύστατου συστήματος βαθμολόγησης. Όπως είναι αναμενόμενο, οι πελάτες με βαθμολογία 9 είναι αυτοί με την μεγαλύτερη αξία, ενώ αυτοί με βαθμολογία 0 έχουν την χαμηλότερη αξία. Ενδεικτικό είναι ότι στην βαθμολογία 9 συγκεντρώνονται οι πελάτες με μηδενική εγγύτητα, την υψηλότερη συχνότητα συναλλαγών, καθώς και την υψηλότερη απόδοση κέρδους.

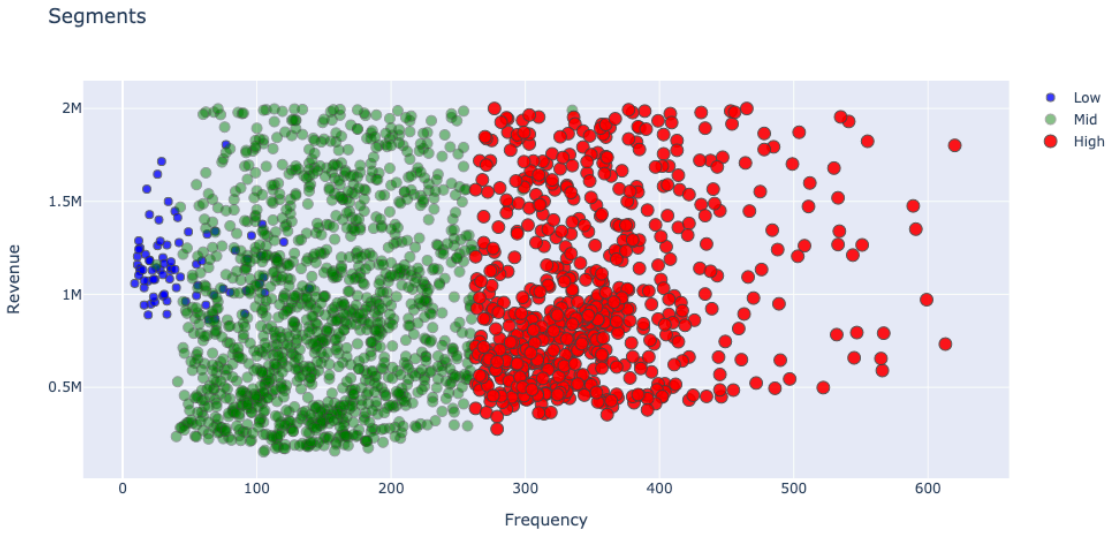
OverallScore	Recency	Frequency	Revenue
0	757.200000	36.200000	1.218581e+06
1	361.375000	20.250000	1.144381e+06
2	37.486111	51.750000	1.271102e+06
3	1.290698	103.247674	1.317443e+06
4	0.496290	170.492993	2.327991e+06
5	0.368421	271.521382	3.006259e+06
6	0.240964	354.317771	5.565804e+06
7	0.000000	418.664894	9.284822e+06
8	0.000000	381.661765	2.458257e+07
9	0.000000	435.666667	2.804916e+07

Εικόνα 43 Πίνακας Βαθμολόγησης RFM

Για σκοπούς καλύτερης μοντελοποίησης του ανωτέρω συστήματος, οι 10 βαθμίδες ομαδοποιήθηκαν σε 3 με την εξής λογική:

- 0 έως 2: Ομάδα χαμηλής Αξίας
- 3 έως 4: Ομάδα μέτριας Αξίας
- 5 έως 9: Ομάδα υψηλής Αξίας

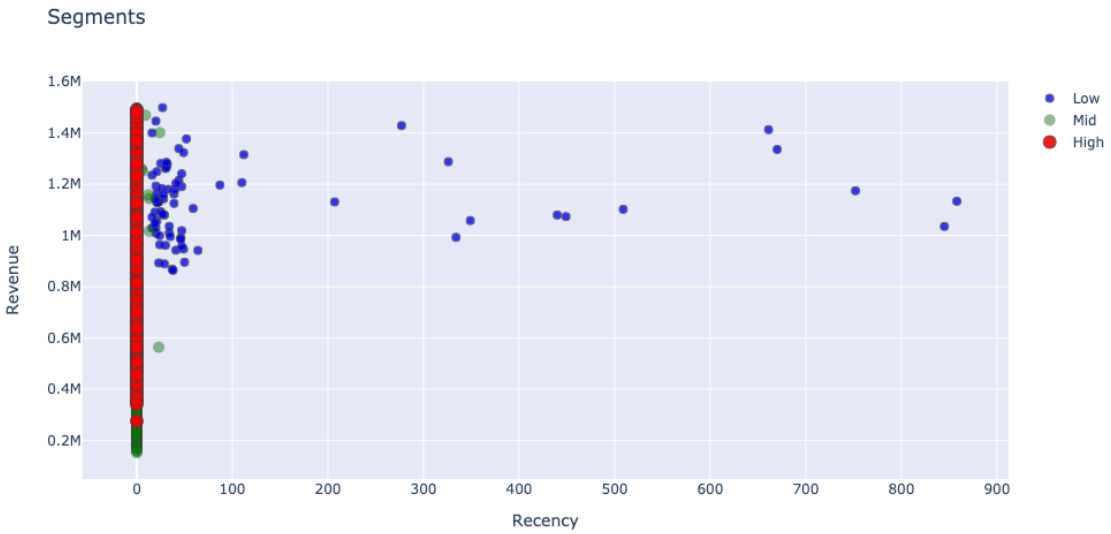
Η ομαδοποιήση αυτή θα βοηθήσει στην τελική τμηματοποίηση των πελατών βάσει της συνολικής τους βαθμολογίας. Τα κατωτέρω γραφήματα είναι αρκετά χρήσιμα στην κατηγοριοποίηση των πελατών με βάση της αξία τους.



Εικόνα 44 Revenue vs Frequency

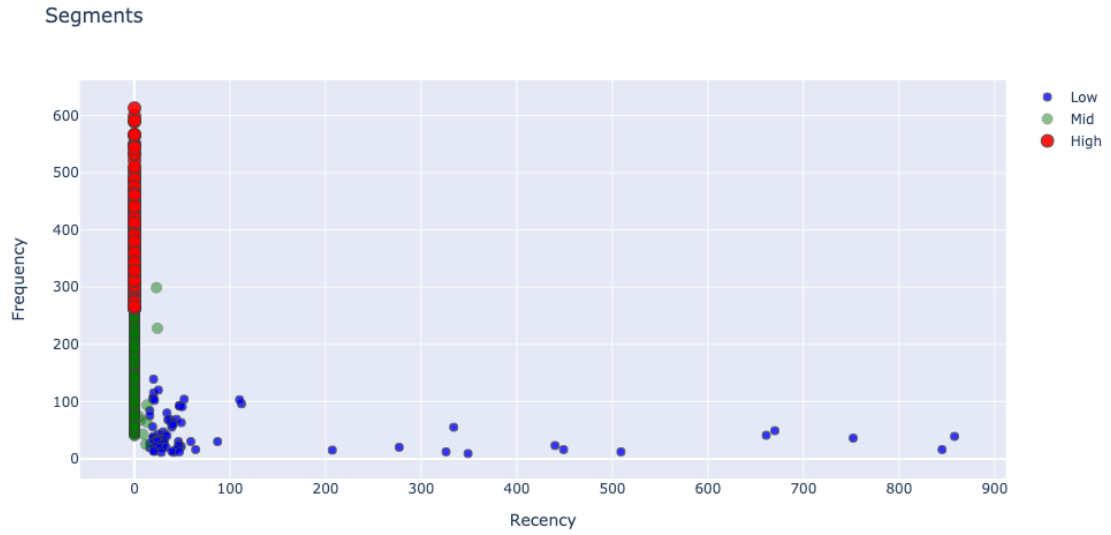
Στην περίπτωση του γραφήματος των μεγεθών Revenue σε συνδυασμό με το Frequency, οι βέλτιστες τιμές είναι εκείνες οι οποίες προσδιορίζονται άνω δεξιά του σχήματος καθώς επιθυμείται υψηλή συχνότητα συναλλαγών παράλληλα με τα υψηλά κέρδη.

Από την αποτύπωση των τριών κλάσεων που δημιουργήθηκαν, παρατηρείται ότι υπάρχει περιθώριο βελτίωσης στους πελάτες «υψηλής αξίας» όσον αφορά την συχνότητα των συναλλαγών τους. Επίσης είναι εμφανές ότι η κλάση των «χαμηλής αξίας» πελατών, αποτελεί την πιο μικρή σε πληθυσμό ομάδα, κατά συνέπεια η τράπεζα έχει το περιθώριο να επικεντρώσει τις δράσεις της για την περεταίρω ανάπτυξη του δυνατού πελατολογίου της.



Εικόνα 45 Revenue vs Recency

Όσον αφορά το ανωτέρω γράφημα, παρατηρείται ότι η σχέση κέρδους – εγγύτητας κυμαίνεται σε αρκετά ικανοποιητικά επίπεδα καθώς το ζητούμενο είναι υψηλές τιμές κέρδους και χαμηλές τιμές εγγύτητας. Οι μόνες αποκλείσεις από αυτό το δόγμα είναι ορισμένες περιπτώσεις «χαμηλής αξίας» οι οποίες όμως είναι μεμονωμένες.



Εικόνα 46 Frequency vs Recency

Η ίδια λογική ακολουθείται και στο γράφημα της συχνότητας σε συνδυασμό με την εγγύτητα, καθώς τόσο το σύνολο των «υψηλής αξίας» πελατών όσο και το σύνολο της «μεσαίας αξίας» έχουν πολύ στενή σχέση με την τράπεζα. Σε αντιδιαστολή, η κλάση των «χαμηλής αξίας» πελατών παρουσιάζει μεμονωμένες περιπτώσεις κατά τις οποίες χρειάζονται δράσεις εκ μέρους της τράπεζας αν αυτό αποτελέσει επιχειρησιακή απόφαση.

Σε γενικές γραμμές, αξίζει να σημειωθεί ότι η εφαρμογή της μεθόδου RFM είναι αρκετά αποτελεσματική ως προς την κατηγοριοποίηση των πελατών καθώς και της συναλλακτικής συμπεριφοράς αυτών, αφού με αυτόν τον τρόπο μία αντίστοιχη επιχείρηση έχει την δυνατότητα να αναπροσαρμόζει τις τακτικές της αναλόγως των μετρήσεων και των μεγεθών που προσφέρει η εν λόγω μέθοδος.

Ως εκ τούτου, με αυτόν τον τρόπο δύναται να δημιουργούνται μοντέλα επιχειρηματικών αποφάσεων ή ακόμα και ενεργειών με σκοπό την βέλτιστη διαχείριση του πελατολογίου είτε συνολικά είτε ανά κατηγορία πελατών με βάση την αξία που προσδίδουν στης τράπεζα.

5.8 Μοντελοποίηση Customer Lifetime Value

Στην προηγούμενη παράγραφο πραγματοποιήθηκε η κατηγοριοποίηση των πελατών σε τρεις βασικές κλάσεις. Σε αυτό το επίπεδο είναι πλέον εφικτή η μέτρηση της συμπεριφοράς της εκάστοτε κλάσης έτσι ώστε να είναι δυνατή η λήψη επιχειρησιακών αποφάσεων.

Δεδομένου ότι ο ορισμός της αξίας του πελάτη δύναται να είναι διαφορετικός ανά περίπτωση και ανά επιχείρηση, ο προσδιορισμός της είναι μια δυναμική διαδικασία η οποία εξαρτάται από την εκάστοτε στρατηγική που υιοθετείται. Παράγοντες αυτής της διαδικασίας είναι οι μέθοδοι marketing και απόκτησης νέων πελατών, οι ενέργειες διατήρησης των υφισταμένων ή ακόμα και της βελτιστοποίησης του κέρδους από μία συγκεκριμένη κατηγορία πελατών.

Σε αυτό το πλαίσιο, η μετρική CLV αποτελεί ένα χρήσιμο εργαλείο αναγνώρισης ης συναλλακτικής συμπεριφοράς των καταναλωτών και εν γένει της αποτύπωσης των χαρακτηριστικών ενός τραπεζικού πελατολογίου.

Βασική προϋπόθεση εφαρμογής της CLV, είναι ο προσδιορισμός μίας χρονικής περιόδου μέτρησής της. Η περίοδος αυτή κυμαίνεται από 3 έως 24 μήνες. Κρίθηκε σκόπιμο για τους σκοπούς της παρούσας ανάλυσης να επιλεγεί μία περίοδος μέτρησης της τάξεως των 6 μηνών καθώς με δεδομένη την συνήθη επιχειρησιακή πρακτική η οποία προσδιορίζει ως μακροσκοπικά τα διαστήματα από 1 έτος και άνω.

Η μέτρηση της CLV θα βασιστεί πάνω στον υπολογισμό της μεθόδου RFM η οποία υλοποιήθηκε στην ανωτέρω παράγραφο. Ως εκ τούτου, είναι απαραίτητη η χρήση της βαθμολογίας η οποία έχει αποδοθεί στον εκάστοτε πελάτη (client_id) για τα τρία μεγέθη της μεθόδου (Recency – Frequency – Monetary Value).

	client_id	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	Segment
0	2700	0	3	102	3	1571947.0	1	7	High-Value
1	1378	0	3	88	3	1208606.0	1	7	High-Value
2	13539	0	3	102	3	1132108.0	1	7	High-Value
3	1216	0	3	104	3	1111099.0	1	7	High-Value
4	3001	0	3	91	3	1055209.0	1	7	High-Value
...	...	...	...	...	...	...	...	...	...
3366	3606	0	3	27	1	5314662.0	3	7	High-Value
3367	2768	0	3	25	1	5309082.0	3	7	High-Value
3368	3735	0	3	28	1	5497569.0	3	7	High-Value
3369	139	0	3	29	1	5762025.0	3	7	High-Value
3370	4137	0	3	28	1	5823333.0	3	7	High-Value

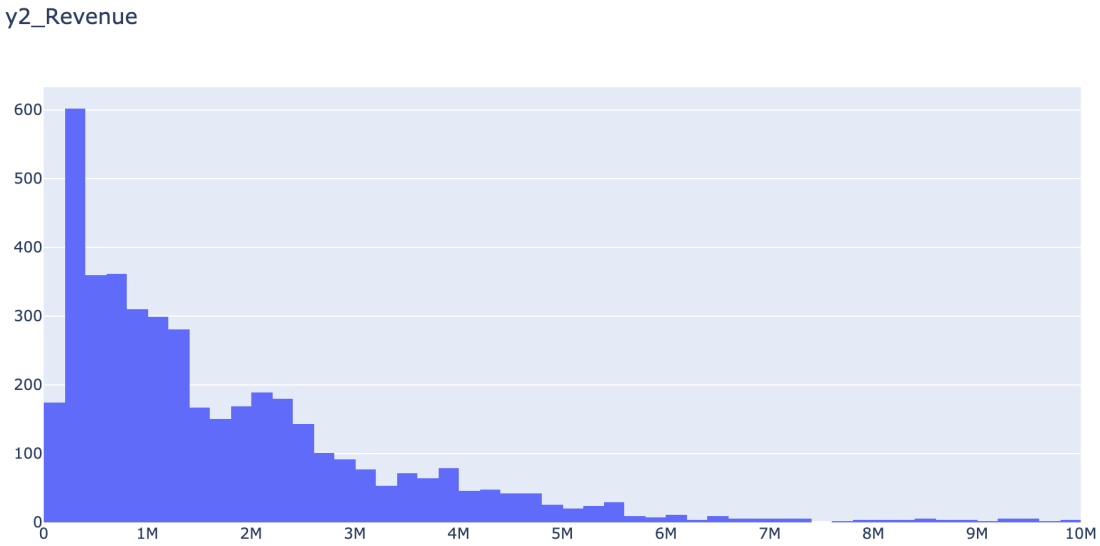
Εικόνα 47 Πίνακας κατηγοριοποίησης

Δεδομένης της κατηγοριοποίησης των πελατών η οποία πραγματοποιήθηκε στην παράγραφο 5.7, έχει δημιουργηθεί ο ανωτέρω πίνακας με την βαθμολογία που προέκυψε από την μέθοδο RFM.

Αυτό είναι και το σύνολο δεδομένων το οποίο θα χρησιμοποιηθεί για τον υπολογισμό του Customer Lifetime Value για το χρονικό διάστημα ενός εξαμήνου. Η μεταβλητή στόχος για τον υπολογισμό του CLV είναι το μέγεθος Lifetime Value (LTV). Ορισμός τους δίνεται από τον κατωτέρω τύπο

$$\boldsymbol{LTV = Total\ Gross\ Revenue - Total\ Cost}$$

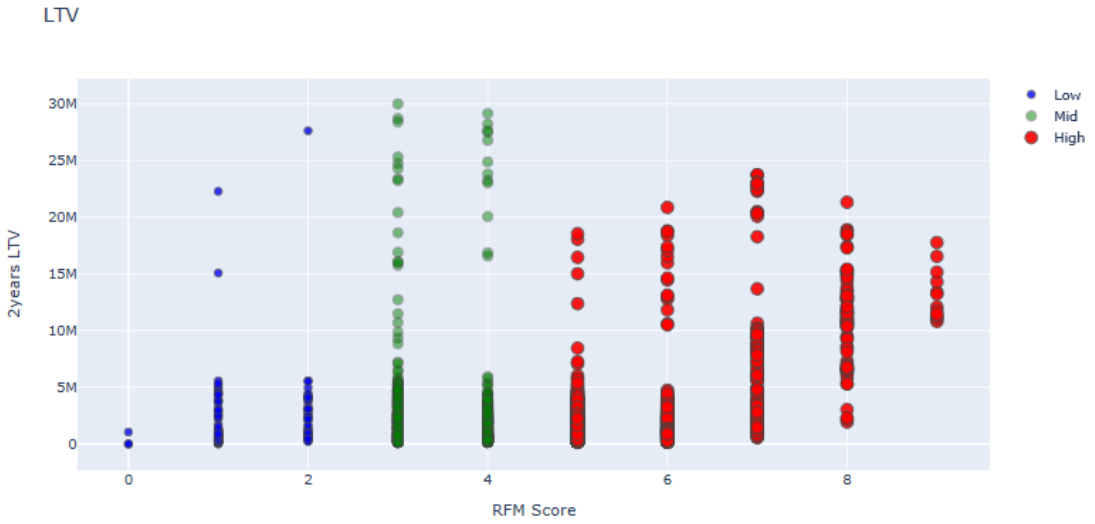
Ωστόσο, στην συγκεκριμένη ανάλυση δεν μελετάται η έννοια του κόστους. Κατά συνέπεια το LTV ισούται με την μεταβλητή του κέρδους (Revenue) η οποία έχει προσδιοριστεί από την μέθοδο RFM και συμπεριλαμβάνεται στο τελικό σύνολο δεδομένων.



Εικόνα 48 Ιστόγραμμα LTV

Από την μελέτη του ιστογράμματος παρατηρείται ότι υπάρχουν πελάτες με μηδενικό LTV, ενώ παρατηρούνται και ορισμένες περιπτώσεις ακραίων τιμών. Ως εκ τούτου, με σκοπό την αρτιότερη τροφοδότηση του τελικού μοντέλου μηχανικής μάθησης με δεδομένα, πραγματοποιήθηκε η αντίστοιχη επεξεργασία στο σύνολο δεδομένων.

Δεδομένης της επικείμενης εφαρμογής τους τελικού μοντέλου μηχανικής μάθησης για τον υπολογισμό του CLV, είναι σημαντικό να εξεταστεί η συσχέτιση μεταξύ των μεγεθών της μεθόδου RFM και της μεταβλητής Revenue.



Εικόνα 49 Διάγραμμα Συσχέτισης

Στο ανωτέρω διάγραμμα αποτυπώνονται τα στοιχεία των 10 βαθμίδων οι οποίες έχουν δημιουργηθεί κατά το στάδιο της κατηγοριοποίησης των πελατών και τον διαχωρισμό τους σε χαμηλής, μέσης και υψηλής αξίας. Οι υψηλές τιμές της RFM συσχετίζονται με τις υψηλές τιμές του κέρδους, ως εκ τούτου, παρατηρείται θετική συσχέτιση μεταξύ των μεγεθών.

Αξίζει να σημειωθεί ότι εν γένει ο υπολογισμός του CLV αποτελεί ένα πρόβλημα παλινδρόμησης. Ωστόσο, για τους σκοπούς της συγκεκριμένης μελέτης, το μοντέλα τα οποία θα δημιουργηθούν, θα εντάσσονται στην οικογένεια των μοντέλων κατηγοριοποίησης (Classification), καθώς στην προκειμένη περίπτωση είναι επιθυμητός ο διαχωρισμός του CLV σε κλάσεις.

5.8.1 Εφαρμογή K-Means

Κατά συνέπεια, θα ακολουθηθεί παρόμοια λογική όπως και στην παράγραφο 5.7, δηλαδή με την βοήθεια του αλγορίθμου K-Means θα αναγνωριστούν οι υφιστάμενες ομάδες CLV.

Απώτερος σκοπός είναι η διάκριση του πελατολογίου στις 3 κατωτέρω κατηγορίες ώστε να είναι πιο ευέλικτη η διαχείρισή του από επιχειρησιακής πλευράς.

- Χαμηλό CLV
- Μεσαίο CLV
- Υψηλό CLV

LTVCluster	count	mean	std	min	25%	50%	75%	max
0	2321.0	8.422255e+05	5.251716e+05	0.0	346744.00	752475.0	1229779.0	2025158.0
1	886.0	3.213846e+06	1.150966e+06	2028234.0	2350306.50	2811022.5	3796598.5	7898675.0
2	130.0	1.263560e+07	3.184315e+06	8048040.0	10147848.75	11809197.0	15073296.5	20054122.0

Εικόνα 50 K-Means LTV

Από τα αποτελέσματα της εφαρμογής του αλγορίθμου της συσταδοποίησης, παρατηρείται ότι το cluster 0 είναι αυτό το οποίο συγκεντρώνει τους πελάτες με το μεγαλύτερη μέση τιμή LTV με 8,4.

5.8.2 Feature Engineering

Κατόπιν της ολοκλήρωσης του αλγορίθμου συσταδοποίησης στο τελικό σύνολο δεδομένων, εφαρμόστηκαν ορισμένες τεχνικές επεξεργασίας χαρακτηριστικών (feature engineering) με σκοπό την προετοιμασία του dataset για την είσοδό του στο τελικό μοντέλο. Σημειώνεται ότι οι κλάσεις του τελικού μοντέλου (Labels) θα προκύψουν από τα clusters του CLV τα οποία δημιουργήθηκαν ανωτέρω. Πιο συγκεκριμένα:

- Πραγματοποιήθηκε μετατροπή της κατηγορικής μεταβλητής "Segment" σε αριθμητικής.
- Υπολογίστηκε η συσχέτιση (correlation) μεταξύ των μεταβλητών που απαρτίζουν το τελικό σύνολο δεδομένων. Από τον κατωτέρω πίνακα παρατηρείται ότι οι μεταβλητές οι οποίες σχετίζονται με το κέρδος, είναι αυτές οι οποίες έχουν μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή του CLV.

```
#calculate and show correlations
corr_matrix = trans_class.corr()
corr_matrix['LTVCluster'].sort_values(ascending=False)
```

```
LTVCluster            1.000000
y2_Revenue            0.826233
RevenueCluster        0.646048
Revenue               0.641102
OverallScore          0.193368
client_id             0.134971
Segment_Mid-Value     0.066172
RecencyCluster        0.050258
Segment_Low-Value    -0.038589
Segment_High-Value   -0.044268
Recency              -0.050504
Frequency            -0.119910
FrequencyCluster     -0.126590
Name: LTVCluster, dtype: float64
```

Εικόνα 51 Data Correlation

- Πραγματοποιήθηκε διαχωρισμός του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης και ελέγχου.

Ύστερα από την υλοποίηση των άνω περιγραφόμενων σταδίων, είναι εφικτή η εφαρμογή των μοντέλων μηχανικής μάθησης μέσω των οποίων θα προκύψει το βέλτιστο.

5.8.3 Μοντέλο Μηχανικής Μάθησης

Απώτερος σκοπός της παρούσας ανάλυσης είναι η δημιουργία ενός μοντέλου το οποίο θα έχει την ικανότητα να κατηγοριοποιεί τους πελάτες της τράπεζας με βάση της επίδοση στην βαθμολογία του δείκτη CLV.

Ύστερα από ενδελεχή έρευνα σε επίπεδο βιβλιογραφίας και δεδομένης της φύσεως της ανάλυσης, ως βασικός αλγόριθμος κατηγοριοποίησης έχει επιλεγεί ο XGBOOST Classifier.

Εικόνα 52 Δενδρόγραμμα XGBoost Classifier

Κατόπιν της εφαρμογής του τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο ελέγχου προέκυψε το κατωτέρω classification report.

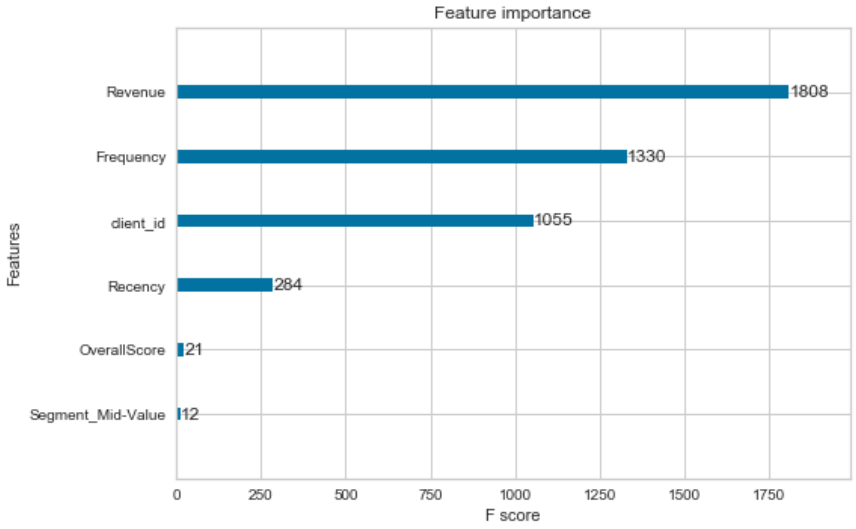
```
Accuracy of XGB classifier on training set: 0.93
Accuracy of XGB classifier on test set: 0.86
              precision    recall  f1-score   support

           0       0.93      0.89      0.91       118
           1       0.67      0.80      0.73        40
           2       1.00      0.67      0.80         9

    accuracy                           0.86       167
   macro avg       0.87      0.79      0.81       167
weighted avg       0.87      0.86      0.86       167
```

Εικόνα 53 Classification Report

- Η ακρίβεια (Accuracy) του μοντέλου είναι αρκετά ικανοποιητική καθώς κατηγοριοποιεί σωστά τους πελάτες στο 93% για το σύνολο εκπαίδευσης και στο 86% για το σύνολο του ελέγχου.

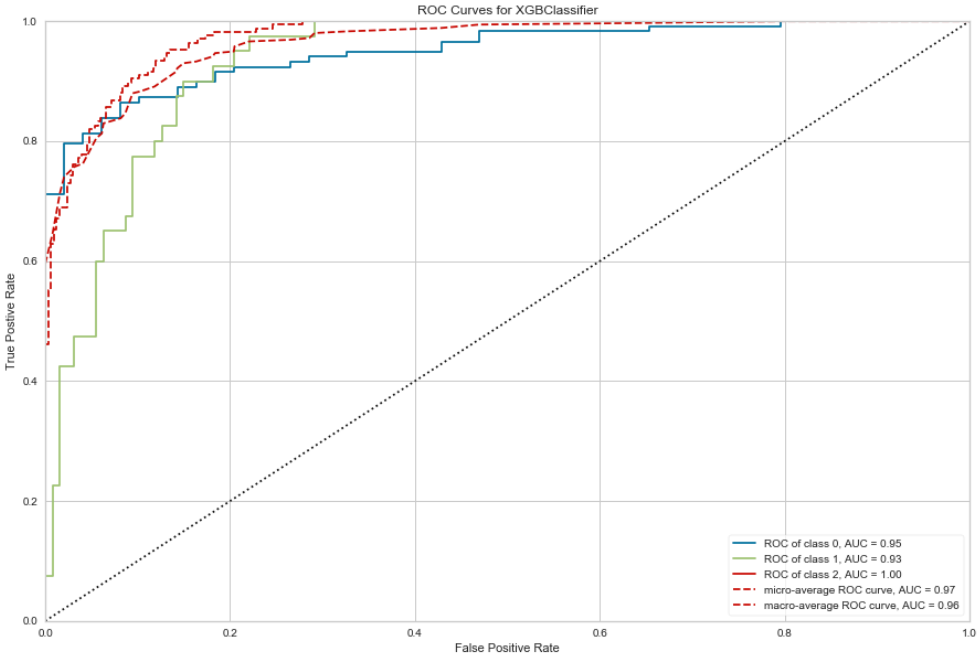
- Όσον αφορά το Precision, το οποίο προκύπτει από την αναλογία (True Positive) / (True Positive + False Positive), προκύπτει πολύ καλή απόδοση στην πρόβλεψη των περιπτώσεων της κλάσης 0 (Χαμηλό CLV) εν αντίθεση με την χαμηλότερη απόδοση στην κλάση 1 (Μεσαίο CLV).

- Αρκετά ικανοποιητικά είναι τα αποτέλεσμα του Recall, δηλαδή της αναλογίας (True Positive) / (True Positive + False Negative) το οποίο κινείται σε υψηλά επίπεδα και για τις τρεις κλάσεις.



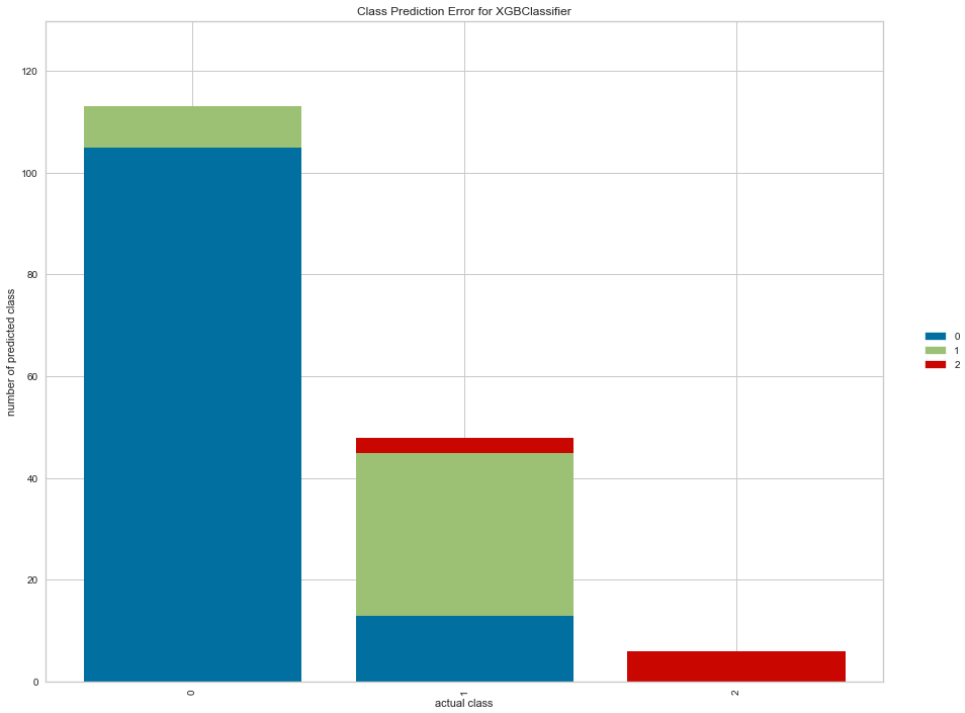
Εικόνα 54 Feature Importance

Παράλληλα, από τον πίνακα 54, αναδεικνύονται οι ανεξάρτητες μεταβλητές του συνόλου δεδομένων οι οποίες επηρεάζουν την απόδοση του αλγορίθμου. Παρατηρείται ότι μεγάλη σημασία έχει η μεταβλητή του κέρδους, καθώς και της συχνότητας.



Εικόνα 55 Καμπύλη ROC

Θετικά είναι τα αποτελέσματα επίσης τα οποία προκύπτουν από την αποτύπωση των καμπυλών ROC τόσο για το σύνολο του μοντέλου όσο και για τις επιμέρους κλάσεις. Οι συγκεκριμένες καμπύλες κινούνται όλες κοντά στο απόλυτο True Positive Rate, κινούμενες ξεκάθαρα πάνω από το μέσο μοντέλο κατηγοριοποίησης.



Εικόνα 56 Class Prediction Error

Τέλος, από το ανωτέρω διάγραμμα είναι διακριτή η κατανομή των εσφαλμένων προβλέψεων ανά κλάση. Παρατηρείται ότι στις 2 πρώτες κλάσεις που είναι και οι πιο πολυπληθείς, το σφάλμα κινείται σε πολύ χαμηλά επίπεδα ενώ στην τρίτη κλάση που είναι αυτή με τα λιγότερα μέλη, δεν υπάρχει καμία εσφαλμένη πρόβλεψη.

5.8.4 Σύγκριση Μοντέλων Κατηγοριοποίησης

Κρίθηκε σκόπιμο στα πλαίσια της αξιολόγησης των επιδόσεων του τελικού μοντέλου να εξεταστεί η εφαρμογή έτερων αλγορίθμων κατηγοριοποίησης στα δεδομένα της ανάλυσης. Η κατωτέρω αναφορά με τα στοιχεία των μετρικών για την εφαρμογή 6 θεμελιωδών μοντέλων μηχανικής μάθησης, παρουσιάζει χρήσιμα συμπεράσματα.

```
LogisticRegression
[[55 63  0]
 [ 5 35  0]
 [ 1  8  0]]
accuracy: 0.5389221556886228
Precision: [0.90163934 0.33018868 0.        ]
Recall: [0.46610169 0.875      0.        ]
RMSE: 0.6921285402059827
############################################################


SVC
[[118   0   0]
 [ 40   0   0]
 [  9   0   0]]
accuracy: 0.7065868263473054
Precision: [0.70658683 0.         0.        ]
Recall: [1. 0. 0.]
RMSE: 0.6746034541560556
############################################################


KNeighborsClassifier
[[106  12   0]
 [ 15  25   0]
 [  4   0   5]]
accuracy: 0.8143712574850299
Precision: [0.848      0.67567568 1.        ]
Recall: [0.89830508 0.625      0.55555556]
RMSE: 0.5074298276019255
############################################################


DecisionTreeClassifier
[[112   6   0]
 [ 12  28   0]
 [  4   0   5]]
accuracy: 0.8682634730538922
Precision: [0.875      0.82352941 1.        ]
Recall: [0.94915254 0.7        0.55555556]
RMSE: 0.4512126044020241
############################################################


RandomForestClassifier
[[104  14   0]
 [ 10  30   0]
 [  3   1   5]]
accuracy: 0.8323353293413174
Precision: [0.88888889 0.66666667 1.        ]
Recall: [0.88135593 0.75       0.55555556]
RMSE: 0.4706982963932894
############################################################


XGBClassifier
[[105  13   0]
 [  6  34   0]
 [  0   3   6]]
accuracy: 0.8682634730538922
Precision: [0.94594595 0.68       1.        ]
Recall: [0.88983051 0.85       0.66666667]
RMSE: 0.3629552685195626
############################################################
```

Εικόνα 57 Σύγκριση Αλγορίθμων Κατηγοριοποίησης

Παρατηρείται ότι ο αλγόριθμος XGBoost Classifier ο οποίος χρησιμοποιήθηκε για την υλοποίηση του βασικού μοντέλου πρόβλεψης παρουσιάζει τις καλύτερες αποδόσεις τόσο σε επίπεδο ακρίβειας (Accuracy), όσο και σε επίπεδο σφάλματος Root Mean Square Error, καθώς ενδεικτικά στον XGBoost Classifier το RMSE αγγίζει το 0.36, τιμή αισθητά χαμηλότερη από τους αλγορίθμους Support Vector Machine και της Λογιστικής Παλινδρόμησης οι οποίοι συγκεντρώνουν τιμές 0.67 και 0.69 αντίστοιχα.

6. Συμπεράσματα & Προοπτικές για Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία προτείνεται μία προσέγγιση μεθόδου υπολογισμού του μεγέθους Customer Lifetime Value μέσα από την οπτική της μηχανικής μάθησης. Ο συγκεκριμένος δείκτης αποτελεί μία σημαντική μετρική για τους μεγάλους οργανισμούς όπως για παράδειγμα οι τράπεζες.

Ένα από τα βασικά προβλήματα ωστόσο τα οποία προκύπτουν κατά την πειραματική διαδικασία είναι τόσο η εξεύρεση όσο και ο χειρισμός των δεδομένων τα οποία σχετίζονται με τραπεζικές εργασίες. Η κωδικοποίηση των μεταβλητών και τα ελλιπή σύνολα δεδομένων δημιουργούν αντικειμενικές δυσκολίες στην παραμετροποίηση ενός μοντέλου το οποίο θα μπορούσε να τεθεί σε παραγωγική διαδικασία υπό ρεαλιστικές συνθήκες.

Στον αντίποδα η μέθοδος RFM παρουσιάζεται αρκετά προσαρμόσιμη όσον αφορά το πεδίο εφαρμογής της. Πιο συγκεκριμένα, ενώ η συγκεκριμένη μέθοδος χτίστηκε πάνω στην λογική του retail εμπορίου καθώς και των αντίστοιχων προφίλ καταναλωτών, στην παρούσα εργασία έδειξε να ανταποκρίνεται σε μεγάλο βαθμό όσον αφορά την προσαρμοστικότητά της στα δεδομένα τραπεζικών συναλλαγών.

Αντίστοιχα, ο αλγόριθμος XGBoost Classifier αποδείχτηκε αποδοτικός και μάλιστα με δεδομένο ότι το αντικείμενο της ανάλυσης ήταν η κατηγοριοποίηση σε πολλαπλές κλάσεις. Ωστόσο, δεδομένων των πειραματικών εφαρμογών στα πλαίσια της παρούσας εργασίας τόσο στον συγκεκριμένο αλγόριθμο όσο και σε άλλους αλγορίθμους κατηγοριοποίησης, διαπιστώθηκε ότι η συγκεκριμένη προσέγγιση έχει περιορισμένα περιθώρια βελτιστοποίησης. Χαρακτηριστικό είναι το γεγονός ότι στην πλειοψηφία της βιβλιογραφίας, οι πιο δημοφιλείς προσεγγίσεις όσον αφορά τον υπολογισμό της CLV ακολουθούν μεθόδους παλινδρόμησης.

6.2 Μελλοντικές Προεκτάσεις

Η υψηλή επίδοση του XGBoost σε συνδυασμό με την δομή του αλγορίθμου η οποία του επιτρέπει να λειτουργεί αποδοτικά σε περιβάλλοντα Parallelization και Distributed Computing, δίνει την προοπτική σε μελλοντικές μελέτες να επεκτείνουν την παρούσα προσέγγιση σε επίπεδο συστήματος και αρχιτεκτονικής μεγάλων δεδομένων.

Σε αυτό το πλαίσιο, θα ήταν ενδιαφέρουσα η πειραματική μελέτη και ανάπτυξη της παρούσας υλοποίησης με την χρήση NoSQL Βάσεων Δεδομένων. Μία τέτοια παραμετροποίηση θα δημιουργούσε τις κατάλληλες προϋποθέσεις για τη δημιουργία ενός μοντέλου το οποίο θα έχει την δυνατότητα να σταθεί σε ένα περιβάλλον εργασίας στο οποίο θα γινόταν κατά κύριο λόγο χρήση αρχιτεκτονικών μεγάλων δεδομένων.

Μία ακόμα μελλοντική επέκταση της συγκεκριμένης εργασίας είναι η δημιουργία ενός front-end περιβάλλοντος το οποίο με την χρήση ενός application framework όπως το Flask για την Python, θα έχει την δυνατότητα να χρησιμοποιεί το ανωτέρω μοντέλο μηχανικής μάθησης για τον υπολογισμό του CLV μέσω μίας εφαρμογής τελικού χρήστη. Μία τέτοια υλοποίηση θα μπορούσε να επεκταθεί σε ένα εργαλείο το οποίο θα ήταν αξιοποιήσιμο σε ένα περιβάλλον εργασίας.

Αξίζει να σημειωθεί επίσης ότι η συγκεκριμένη μελέτη θα μπορούσε να μετεξελιχθεί σε ένα μοντέλο το οποίο ως τελικό σκοπό θα έχει την δημιουργία ενός recommendation system για τους πελάτες μίας τράπεζας.

Πιο συγκεκριμένα, το μοντέλο της κατηγοριοποίησης των πελατών με βάση την Αξία που προσδίδεται από αυτούς στην τράπεζα, θα μπορούσε να εξελιχθεί σε ένα νέο συνδυαστικό μοντέλο το οποίο θα «συνεργάζεται» με το υφιστάμενο και θα έχει την δυνατότητα να αποδίδει προτάσεις πωλήσεων. Με αυτόν τον τρόπο θα ήταν εφικτό να προτείνονται εξατομικευμένες προτάσεις πωλήσεων προϊόντων και υπηρεσιών ανά κατηγορία πελατολογίου.

Τέλος, σημειώνεται ότι στην εν λόγω διπλωματική εργασία, το πρόβλημα του υπολογισμού του δείκτη Customer Lifetime Value υπολογίστηκε με τη λογική ενός προβλήματος κατηγοριοποίησης. Η μηχανική μάθηση, μέσα από την πληθώρα μοντέλων και στατιστικών προσεγγίσεων θα μπορούσε να προσφέρει εναλλακτικές προσεγγίσεις, όπως για παράδειγμα η δημιουργία ενός μοντέλου με τη χρήση νευρωνικών δικτύων. Μία τέτοια λογική θα κινούταν στην κατεύθυνση της βελτιστοποίησης της αξιοπιστίας και της απόδοσης των αποτελεσμάτων.

Βιβλιογραφία

ΟΤΟΕ, Ι. Ε. (2018). *Νέες τεχνολογίες στις τράπεζες και επιπτώσεις στην απασχόληση.*

Κοντιάδης, Ξ., Παπαδημητρίου, Κ., Γεωργακοπούλου, Β., & Στεφανίδης, Δ. (2018). *Νέες τεχνολογίες στις τράπεζες και επιπτώσεις στην απασχόληση.* Ινστιτούτο Εργασίας ΟΤΟΕ.

ΣΕΒ. (2017). *Η ψηφιακή Ελλάδα, ο δρόμος προς την ανάπτυξη.*

ETEN. (2018). *5th digital banking forum.*

EY, Y. E. (2018). *Global Banking Outlook 2018.*

Baesens, B. (2014). *Analytics in a Big Data World - The Essential Guide to Data Science and Its Applications.*

Blattberg, C., Malthouse, E., & Neslin , S. (2009). *Customer Lifetime Value: Empirical Generalizations and Some Conceptual Questions.*

Blattberg, R., Getz, G., & Thomas, J. (2001). *Customer Equity: Building and Managing Relationships As Valuable Assets.*

Chapelle , O., Scholkopf, B., & Zien, A. (2006). *Semi-Supervised Learning* .

CHI, S. (2019). *www.medium.com*. Ανάκτηση από www.medium.com: https://medium.com/@chisoftware/supervised-vs-unsupervised-machine-learning-7f26118d5ee6

Djamal Abide, M. (2017). Application of deep leaning to computer vision.

ECB, E. K. (2018). *Τι είναι οι τράπεζες χρηματοοικονομικής τεχνολογίας (fintech) και πώς επηρεάζουν τις χρηματοπιστωτικές υπηρεσίες;*. Ανάκτηση από https://www.bankingsupervision.europa.eu/about/ssmexplained/html/fintech.el.html.

Fader, P., Hardie, B., & Lee, K. (2005). *"Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model.*

Gowrisankar. (2020). *https://levelup.gitconnected.com/*. Ανάκτηση από https://levelup.gitconnected.com/xgboost-queens-of-boosting-algorithms-f270894c6aa5

Gupta, S., & Hanssens, D. (2006). *Modeling Customer Lifetime Value.*

kaggle. (2020). *https://www.kaggle.com/*.

Kumar, V. (2006). *Customer lifetime Value.*

Kumar, V. (2010). *A Customer Lifetime Value-Based Approach to Marketing in the Multichannel, Multimedia Retailing Environment.*

Kumar, V., & Ramani , G. (2004). *Customer lifetime value approaches and best practice applications.*

Lee, D. (2019). *https://medium.com/*. Ανάκτηση από https://medium.com/ai%C2%B3-theory-practice-business/reinforcement-learning-part-1-a-brief-introduction-a53a849771cf

Makhija, P. (2020). *https://clevertap.com/*. Ανάκτηση από https://clevertap.com/blog/rfm-analysis/

McKinsey. (2016). *Are today's CFOs ready for tomorrow's demands on finance?* Ανάκτηση από https://www.mckinsey.com/business-functions/strategy-andcorporate-finance/our-insights/are-todays-cfos-ready-for-tomorrows-demands-on-finance

Nishio, M., Nishizawa , M., Surigiyama, O., & others. (2018). *Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization.*

Novo, J. (2001). *Maximizing Marketing ROI with Customer Behavior Analysis.*

Oracle. (2019). *The Definition of Big Data*. Ανάκτηση από https://www.oracle.com/big-data/guide/what-is-big-data.html

Reinartz, & Kuzmar. (2003). *The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration.*

Russel, S., Norvig, P., & Davis, E. (2003). *Artificial Intelligence: A modern Approach* (Second εκδ.).

S. Gupta, D. H. (2006). *Modeling Customer Lifetime Value.*

Schroeck, M., Shockley, R., Smart, J., Morales, R., & Tufano, P. (2012). *Analytics: The real-world use of big data.* IBM Report.

Shah, I. (2020). *https://blog.quantinsti.com/*. Ανάκτηση από https://blog.quantinsti.com/xgboost-python/

Silberschatz, Korth, & Sudarshan. (χ.χ.). *Συστήματα Βάσεων Δεδομένων - Η Πλήρης Θεωρία των Βάσεων Δεδομένων.* Εκδόσεις Μ.Γκιούρδας.

Villanueva, J., Yoo, S., & Hanssens, D. (2008). *The Impact of Marketing-Induced versus Word-of-Mouth Customer Acquisition on Customer Equity Growth.*

Wikipedia. (2019). *Διαδίκτυο των πραγμάτων*. Ανάκτηση από https://el.wikipedia.org/wiki/Διαδίκτυο_των_πραγμάτων

Witten, I., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition εκδ.). Morgasn Kayfmann.

www.techentice.com. (2020, 06 20). Ανάκτηση από www.techentice.com: https://www.techentice.com/the-data-veracity-big-data/

Ρεντούμης. (2018). *Απειλή ή ευκαιρία για τις τράπεζες το PSD2;*. Ανάκτηση από https://www.kathimerini.gr/980619/article/oikonomia/die8nhs-oikonomia/apoyh-apeilh-h-eykairia-gia-tis-trapezes-to-psd2

Μαριόλη, Κ. (2016). *Οι τράπεζες μικραίνουν και γίνονται ψηφιακές.* Kathimerini.gr.

Μάλλας, Δ. (2018). *Τι είναι το blockchain και γιατί είναι τόσο σημαντικό*. Ανάκτηση από https://www.cnn.gr/oikonomia/story/117710/ti-einai-to-blockchain-kai-giati-einai-tososimantiko.

Πίνακας Εικόνων