ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ»

«ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ»

Εφαρμογές Εξόρυξης Γνώσης στον Τραπεζικό Κλάδο
Πρόβλεψη Απάτης σε Συναλλαγές Πιστωτικών Καρτών

ΠΛΑΤΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΜΕ1940

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΦΙΛΙΠΠΑΚΗΣ ΜΙΧΑΗΛ

Πειραιάς, Ιούνιος 2021

UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

POSTGRADUATE STUDIES PROGRAMME

«INFORMATION SYSTEMS & SERVICES»

«BIG DATA AND ANALYTICS»

Application of Data Mining in Banking Sector

Credit Card Fraud Prediction

PLATIS KONSTANTINOS

ME1940

SUPERVISOR PROFESSOR: FILIPPAKIS MICHAEL

Piraeus, June 2021

Περίληψη

Η τρέχουσα περίοδος μπορεί να χαρακτηριστεί εύκολα ως η εποχή των «μεγάλων δεδομένων». Η ραγδαία ανάπτυξη της τεχνητής νοημοσύνης, της μηχανικής μάθησης και της εξόρυξης δεδομένων έχει οδηγήσει στην ψηφιοποίηση των περισσοτέρων ενεργειών της καθημερινότητας των ανθρώπων. Ως εκ τούτου, ο κλάδος των χρηματοοικονομικών και ιδιαίτερα ο τραπεζικός, δεν θα μπορούσε να μείνει ανεπηρέαστος από τις μεταβολές αυτές. Τα τελευταία χρόνια παρατηρείται η ολοένα και αυξανόμενη τάση, από πλευράς των διοικήσεων των τραπικών ιδρυμάτων, για ψηφιοποίηση των χρηματοοικονομικών διεργασιών. Η χρήση του πλαστικού χρήματος και συγκεκριμένα πιστωτικών καρτών τείνει να αντικαταστήσει τη χρήση χαρτονομισμάτων στις δοσοληψίες. Ωστόσο, η εκτεταμένη χρήση πιστωτικών καρτών συνοδεύεται από αύξηση των φαινομένων απάτης σε παγκόσμιο επίπεδο.

Η παρούσα διπλωματική εργασία πραγματεύεται το φλέγον ζήτημα των φαινομένων απάτης με χρήση πιστωτικών καρτών, ακολουθώντας ένα επαγωγικό πλάνο ανάπτυξης. Αρχικά, στο πρώτο και δεύτερο κεφάλαιο, γίνεται μια εκτενής αναφορά στα μεγάλα δεδομένα και στην αξία αυτών, στους διάφορους τομείς της καθημερινότητας, εστιάζοντας στον τραπεζικό κλάδο. Έπειτα, στο τρίτο κεφάλαιο λαμβάνει χώρα η περιγραφή του προβλήματος της απάτης στον χρηματοοικονομικό τομέα και στο τραπεζικό σύστημα. Στο τέταρτο κεφάλαιο παρατίθεται το πειραματικό σκέλος της εργασίας, όπου γίνεται μια προσπάθεια ανάπτυξης διαφόρων μοντέλων πρόβλεψης των απατηλών συναλλαγών που έγιναν με πιστωτική κάρτα, κάνοντας χρήση δεδομένων που διατίθενται από την Vesta Corporation. Τέλος, στο πέμπτο κεφάλαιο, αποτυπώνονται τα αποτελέσματα των τελεσθέντων πειραμάτων και τα εξαγόμενα συμπεράσματα.

Abstract

It is commonly considered that we are living the “Big Data” era. The rapid development of the artificial intelligence, machine learning and data mining has led to the digitization of most of the actions of people's everyday lives. Therefore, the financial sector, and in particular the banking sector, could not be unaffected by these changes. In recent years there has been a growing tendency on the part of the administrations of banking institutions, to digitilise financial processes. The use of plastic money and specifically credit cards tends to replace the use of banknotes in transactions. However, the widespread use of credit cards is accompanied by an increase in fraud worldwide.

This dissertation deals with the burning issue of credit card fraud, following an inductive development plan. Initially, in the first and second chapter, an extensive reference is made to the big data and their value, in the various areas of everyday life, focusing on the banking industry. Then, the third chapter describes the problem of fraud in the financial sector and specifically in the banking system. The fourth chapter lists the experimental part of the work, where an attempt is made to develop various models for predicting fraudulent transactions made by credit card, using data made available by Vesta Corporation. Finally, chapter five captures the results of the experiments carried out and the conclusions drawn.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω από καρδιάς τον κύριο Μιχαήλ Φιλιππάκη, καθηγητή του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, για την αμέριστη βοήθεια του, τόσο κατά την εκτέλεση της παρούσας εργασίας, όσο και κατά τη διάρκεια των μαθημάτων του προγράμματος μεταπτυχιακών σπουδών.

Επίσης, θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στους γονείς μου, Χρήστο και Πολυτίμη, για την στήριξη και την εμπιστοσύνη που μου έδειξαν καθόλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

ΚΕΦΑΛΑΙΟ 1: Μεγάλα Δεδομένα και Αρχές Εξόρυξης Γνώσης

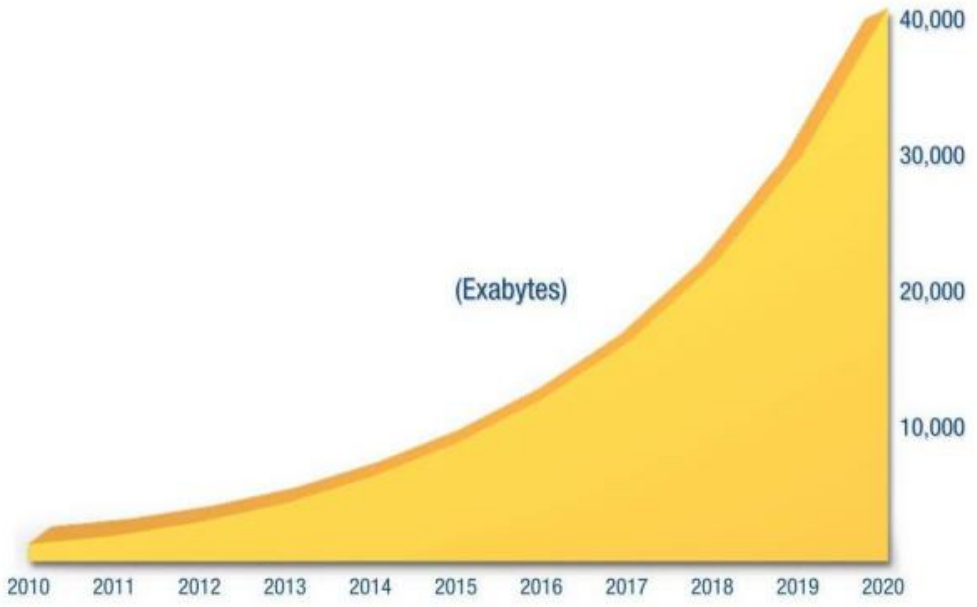
1.1 Μεγάλα Δεδομένα: Ορισμός-Προέλευση, Εφαρμομές, Κατηγορίες

Ο όρος «Μεγάλα Δεδομένα» (Big Data) χρησιμοποιείται για να περιγράψει δεδομένα που είναι δύσκολο έως αδύνατο να υποστούν επεξεργασία με χρήση παραδοσιακών εργαλείων επιχειρηματικής ευφυΐας. Ο επικρατέστερος ορισμός έως και σήμερα είναι αυτός του Gartner από το 2001: « Μεγάλα δεδομένα είναι αυτά που έχουν μεγάλη ποικιλία (Variety), ταχαίως αυξανόμενο όγκο (Volume) και ακόμα μεγαλύτερη ταχύτητα παραγωγής (Velocity).

Το 99% των διαθέσιμων δεδομένων μέχρι σήμερα έχουν δημιουργηθεί τη τελευταία δεκαετία. Ωστόσο, η έννοια των δεδομένων δεν περιορίζεται αποκλειστικά και μόνο στις τελευτάιες δύο δεκαετίες, οπότε και έχει γίνει περισσότερο δημοφιλής ως όρος, αλλά έχει ρίζες δεκάδων χιλιάδων ετών. Κατά το πέρας των αιώνων, οι άνθρωποι διαπίστωσαν ότι η συλλογή και ανάλυση των δεδομένων θα μπορούσε να τους βοηθήσει στην ευκολότερη και ορθότερη λήψη αποφάσεων.

Πιο συγκεκριμένα, το 18.000 πΧ υπάρχουν ενδείξεις ότι οι άνθρωποι ξεκίνησαν να χρησιμοποιούν μπαστούνια για την καταμέτρηση των αποθεμάτων σε τρόφιμα και πρώτες ύλες που εμπορεύονταν. Ήταν η πρώτη προσπάθεια του ανθρώπου να συλλέξει δεδομένα προς επεξεργασία. Ακολούθησε, χιλιάδες χρόνια μετά, η δημιουργία των βιβλιοθηκών της Βαβυλώνας και μετέπειτα της Αλεξάνδρειας, οι οποίες μπορούμε να ισχυριστούμε ότι αποτέλεσαν τις πρώτες αποθήκες δεδομένων της ανθρωπότητας. Αξίζει να τονιστεί ότι η ρωμαϊκή αυτοκρατορία στήριξε πολλούς τομείς της διοικητικής της οργάνωσης στη συλλογή και ανάλυση των δεδομένων, εφαρμόζοντας για πρώτη φορά στην ιστορία ένα εκτεταμένο σύστημα απογραφής των πολιτών της αυτοκράτορίας, σε όλο το μήκος και πλάτος της. Στο πέρας των χρόνων, ο άνθρωπος χρησιμοποίησε τα δεδομένα με σκοπό την εύρεση λύσεων σε θέματα υγείας, οικονομίας και τεχνολογίας. Χαρακτηριστικά, το 1663 ο John Graunt κατέγραψε και επεξεργάστηκε τον αριθμό των θανάτων και τα αίτια αυτών στη περιοχή του Λονδίνου που εκείνη την περίοδο ταλανιζόταν από τον «μάυρο θάνατο», γνωστό και ως βουβωνική πανώλη, με σκοπό να δημιουργήσει ένα σύστημα προφύλαξης του λαού από την εξάπλωση της θανατηφόρου ασθένειας. Τέλος, οι Herman Hollerith και Fritz Pfleumer, το 1881 και 1962 αντίστοιχα, καταπιάστηκαν με τη δημιουργία μηχανών αποθήκευσης δεδομένων, που αποτέλεσαν τη βάση για τη μετέπειτα ψηφιακή αποθήκευση τους.

Τις τελευταίες δύο δεκαετίες παρατηρείται μια ραγδαία αλλαγή στο τρόπο και στη ποσότητα παραγωγής δεδομένων παγκοσμίως. Μόνο τη δεκαετία που μας πέρασε, υπολογίζεται ότι δημιουργήθηκαν περί των 40 zettabytes δεδομένων. Αξίζει να σημειωθεί το γεγονός ότι κάθε άνθρωπος εκτιμάται ότι δημιουργεί περίπου 1,7 MB πληροφορίας κάθε δευτερόλεπτο κάθε ημέρα. Ο τεράστιος αυτός όγκος δεδομένων δημιούργησε, λοιπόν, την ανάγκη εύρεσης εναλλακτικών τρόπων αποθήκευσης και ανάλυσης, καθιστώντας έτσι την εποχή που διανύουμε ως εποχή των «Μεγάλων Δεδομένων».



Εικόνα 1 - Τάση Παραγωγής Δεδομένων

Πηγή: IDC's Digital Universe Study, sponsored by EMC, December 2012

1.1.1 Χαρακτηριστικά Μεγάλων Δεδομένων

Προτού γίνει αναφορά στη σημασία και την αξία των μεγάλων δεδομένων στους διάφορους τομείς της κοινωνίας, πρέπει να τονιστούν τα χαρακτηριστικά που διαθέτουν, τα οποία είναι γνωστά και ως 5 Vs: (Volume, Velocity, Variety, Veracity, Value) και είναι τα ακόλουθα:

I. *Volume (Όγκος)*

   Η ποσότητα των δεδομένων παίζει σημαντικό ρόλο. Η ολοένα και αυξανόμενη χρήση του διαδικτύου και των διαφόρων εφαρμογών (Facebook, Twitter κα) έχει επιφέρει ραγδαία αύξηση στον όγκο των δεδομένων που δημιουργούνται καθημερινά, που στη πλειονότητα τους πρόκεινται περί χαμηλής πυκνότητας, αδόμητα δεδομένα. Οι τάξεις μεγέθους για τις οποίες γίνεται λόγος κυμαίνονται από μερικές δεκάδες terabytes, έως και εκατοντάδες petabytes.

II. *Velocity (Ταχύτητα)*

Ο όρος velocity αναφέρεται στην υψηλή ταχύτητα με την οποία τα δεδομένα εισέρχονται στις βάσεις δεδομένων και εν συνεχεία επεξεργάζονται. Δύο είναι τα σημαντικότερα ζητούμενα στο σημείο αυτό. Πρώτον, το πως ένα σύστημα (μία βάση δεδομένων, ένας σκληρός δίσκος, μια προσωρινή μνήμη) θα υποδεχτεί, φιλτράρει και αποθηκεύσει τα δεδομένα που καταφθάνουν σε αυτό. Δεύτερον, το πως θα επεξεργαστεί τα εισερχόμενα δεδομένα με σκοπό την εκτέλεση μιας λειτουργίας.

III. *Variety (Ποικιλομορφία)*

Ο όρος variety αναφέρεται στη ποικιλομορφία των δεδομένων, δηλαδή στο μεγάλο εύρος διαφορετικών τύπων που μπορούν να έχουν. Τα δεδομένα προέρχονται από διάφορες πηγές και μπορούν να υφίστανται σε διαφορετικούς τύπους, διατάξεις και δομές. Κατά αυτόν τον τρόπο, τα μεγάλα δεδομένα χωρίζονται σε τρείς κύριες υποκατηγορίες (δομημένα, ημιδομημένα και αδόμητα) που θα μελετηθούν εκτενέστερα παρακάτω.

IV. *Veracity (Εγκυρότητα)*

Ο όρος veracity αναφέρεται στην αξιοπιστία, την ποιότητα και την ακρίβεια των δεδομένων. Τα δεδομένα βρίσκονται σε διάφορες μορφές και με διαφόρους τρόπους αποθηκευμένα, έχοντας ως αποτέλεσμα πολλές φορές οι πληροφορίες που περιέχουν να περιλαμβάνουν «θόρυβο», ψευδή στοιχεία και ανακρίβειες. Γα το λόγο αυτό, μεγάλο μέρος της ανάλυσης «μεγάλων δεδομένων» καταλαμβάνει η άρτια διαλογή και ο καθαρισμός αυτών. Χαρακτηριστικό παράδειγμα αποτελούν τα δεδομένα που ανακύπτουν από τα μέσα μαζικής δικτύωσης όπως το Twitter. Κάθε tweet δεν είναι δεδομένο ότι θα περιέχει ορθές πληροφορίες, ή ο χρήστης που θα το συντάσσει θα είναι πραγματικός. Συνεπώς, ένα tweet δεν μπορεί εξαρχής να θεωρηθεί ως αξιόπιστο δεδομένο προτού υποβληθεί στην απαιτούμενη ανάλυση.

V. *Value (Αξία)*

Ο όρος value αναφέρεται στην αξία που μπορούν να έχουν τα δεδομένα ύστερα από τη σωστή διαχείριση τους. Η συλλογή και πρόσβαση σε δεδομένα, χωρίς την ορθή επεξεργασία και μετατροπή σε αξία, δεν έχει την παραμικρή σημασία. Για το λόγο αυτό, ορθά έχει χαρακτηριστεί ως το πιο σημαντικό V των μεγάλων δεδομένων. Η εξαγωγή πληροφορίας από αυτά αποτελεί τη μεγαλύτερη προτεραιότητα των οργανισμών, με

στόχο την όσο τον δυνατόν ορθότερη λήψη στρατηγικών αποφάσεων σε διάφορους τομείς.

Εικόνα 2 - Τα 5 Χαρακτηριστικά (5V΄s) των Μεγάλων Δεδομένων

Πηγή: http://iclerisy.com/what-is-big-data/

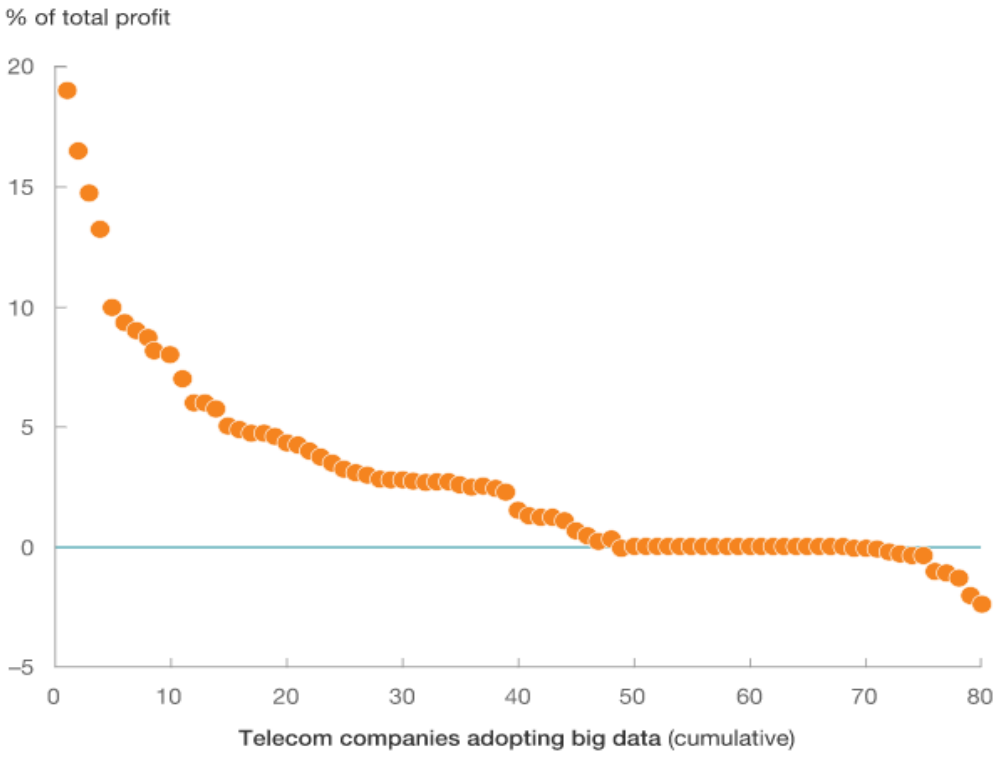
1.1.2 Τομείς Εφαρμογής Μεγάλων Δεδομένων

Την τελευταία δεκαετία τα Μεγάλα δεδομένα χρησιμοποιούνται ευρέως σε πολλαπλούς τομείς της οικονομίας και της κοινωνίας γενικότερα. Ολοένα και περισσότερες εταιρείες συλλέγουν, αποθηκεύουν, επεξεργάζονται και αναλύουν δεδομένα καθημερινά με στόχο την εξαγωγή χρήσιμων πληροφοριών για την ενίσχυση του επιχειρησιακού τους πλάνου. Η ανάλυση της συμπεριφοράς και των προτιμήσεων των πελατών, διαμορφώνει σε μεγάλο βαθμό την πολιτική προώθησης προιόντων και υπηρεσιών από πλευράς των εταιρειών, με σκοπό την αύξηση του κέρδους τους. Επίσης, κατά αυτόν τον τρόπο ενισχύεται η πελατοκεντρική προσέγγιση των επιχειρήσεων, από τη στιγμή που η ανάλυση των Μεγάλων δεδομένων μπορεί να αναδείξει συγκεκριμένες ομάδες ανθρώπων, που χρήζουν διαφορετικής προσέγγισης, τόσο στο τομέα της διαφήμισης όσο και στο τομέα της εξυπηρέτησης.

Παρακάτω παρουσιάζονται συνοπτικά ορισμένοι από τους κυριότερους κλάδους που αξιοποιούν τα μεγάλα δεδομένα σήμερα:

➢ *Τηλεπικοινωνίες*

Στις μέρες μας, οι πάροχοι τηλεπικοινωνιών έχουν πρόσβαση σε τεράστιους όγκους δεδομένων, από τη στιγμή που έχει ενταθεί η χρήση έξυπνων συσκευών (smartphones), φωνητικών μηνυμάτων και βιντεοκλήσεων. Ολοένα και περισσότερες εταιρείες στο χώρο της τηλεπικοινωνίας επενδύουν στην ανάλυση των δεδομένων με σκοπό την αύξηση του κέρδους τους και την ανέλιξη τους στο χώρο. Σύμφωνα με έρευνα της McKinsey το 2016 [1], στην οποία έλαβαν μέρος 273 εταιρείες τηλεπικοινωνιών από όλο το κόσμο, σχεδόν οι μισές εξ΄αυτών αποκρίθηκαν ότι σκοπεύουν να επενδύσουν στην αξιοποίηση των μεγάλων δεδομένων, ενώ το 30% φάνηκε να έχει ήδη προχωρήσει στη κίνηση αυτή. Επίσης, στην έρευνα που διεξήχθη υπολογίστηκε ο βαθμός συσχέτισης του κέρδους που προέκυψε από την αξιοποίηση των δεδόμενων, με το κεφάλαιο και τις εργατοώρες που επενδύθηκαν στη κατεύθυνση αυτή. Τα αποτελέσματα σε δείγμα 80 εταιρειών έδειξαν ότι σε μερικές εξ αυτών τα κέρδη ξεπέρασαν το 10%, στις περισσότερες κυμάνθηκαν από 0-5% και μόνο σε λίγες η επένδυση στα μεγάλα δεδομένα είχε αρνητικό πρόσημο.

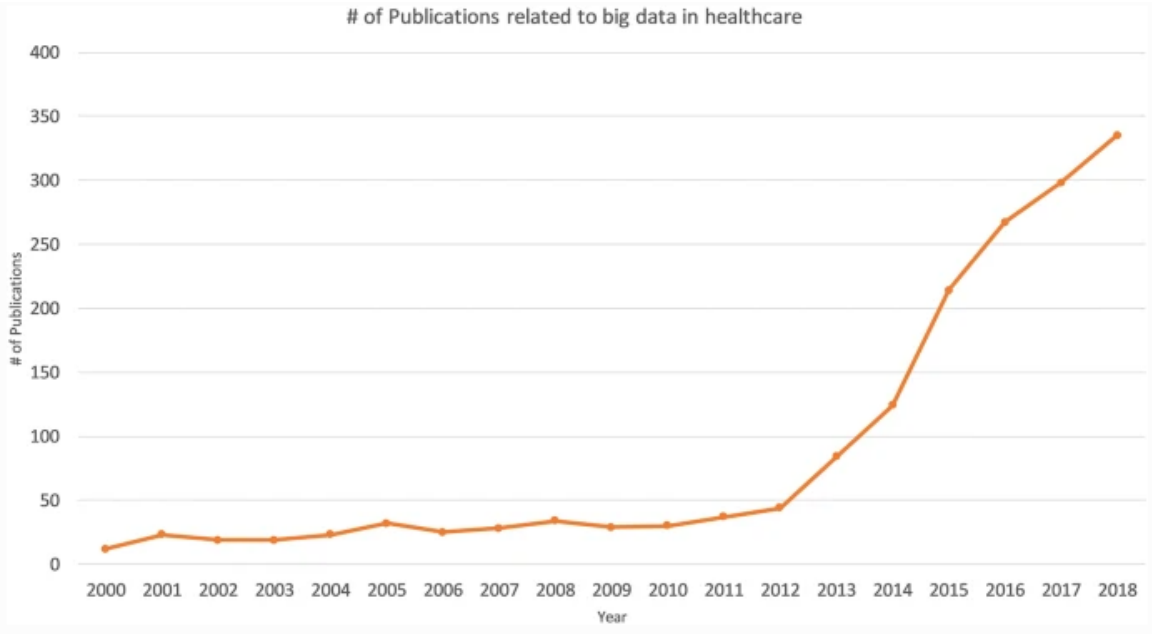


Εικόνα 3 - Επιρροή Μεγάλων Δεδομένων στα Κέρδη των Εταιρειών Τηλεπικοινωνίας

Πηγή: 2015 McKinsey survey of 273 global telecom companies, 80 of which have made big data analytics investments

- *Υγεία*

Τεράστιο ρόλο διαδραματίζουν, επίσης, τα μεγάλα δεδομένα στο χώρο της υγείας. Τα ιατρικά δεδομένα αυξάνονται καθημερινά με εκθετικό ρυθμό, οπότε και η συλλογή και ανάλυση τους αποτελεί μείζον ζήτημα για το χώρο. Η ψηφιοποίηση των συστημάτων καταγραφής ιατρικών δεδομένων έχει συμβάλλει τα μέγιστα στην ταχύτερη συλλογή, διάδοση και επεξεργασία τους. Μέχρι και τις προηγούμενες δεκαετίες οι ιατροί δεν είχαν άμεση πρόσβαση σε ιατρικά στοιχεία σε πραγματικό χρόνο, με αποτέλεσμα να υπάρχει εύλογη καθυστέρηση στη βελτίωση των ιατρικών πρωτοκόλλων και στρατηγικών. Σήμερα, οι ιατροί, από την άνεση του γραφείου τους, μπορούν να έχουν πρόσβαση σε αποθήκες δεδομένων από όλο το κόσμο, να ανταλλάσουν δεδομένα και να λαμβάνουν μέρος σε πειραματικές μελέτες. Η διαδραστικότητα αυτή έχει επιταχύνει σε μεγάλο βαθμό τη λήψη αποφάσεων, τη βελτίωση θεραπευτικών σχημάτων και την εξατομικευμένη ιατρική. Χαρακτηριστική εφαρμογή των μεγάλων δεδομένων στο χώρο της υγείας είναι στην πρόσφατη πανδημία που έπληξε και συνεχίζει να πλήττει την παγκόσμια κοινότητα. Η τάχιστη συλλογή, αποθήκευση, επεξεργασία και ανάλυση των δεδομένων από κάθε μήκος και πλάτος της γης, συνέβαλε στη βαθύτερη κατανόηση της δράσης του ιού, στον έγκαιρο εντοπισμό των φορέων του ιού, στην ιχνηλάτηση των επαφών τους, στην απομόνωση των ευπαθών ομάδων με σκοπό τη προφύλαξή τους και τέλος στην εξεύρεση αποτελεσματικών θεραπευτικών σχημάτων [2].



Εικόνα 4 - Αριθμός Ιατρικών Δημοσιεύσεων Σχετικών με Μεγάλα Δεδομένα

Πηγή: Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. J Big Data 6, 54 (2019)

- *Ενέργεια-Φυσικοί πόροι*

Στο κλάδο της ενέργειας τα μεγάλα δεδομένα χρησιμοποιούνται από τις εταιρείες πετρελαίου και φυσικού αερίου με σκοπό την διευρεύνηση περιοχών που είναι πλούσιες σε κοιτάσματα που μπορούν να καταστούν εκμεταλλεύσιμα. Επίσης, μέσω αυτών, ελέγχονται οι διαδικασίες εγκατάστασης και συντήρησης των αγωγών.

- *Οικονομία*

Ο κλάδος της οικονομίας είναι από τους πρώτους που ξεκίνησαν να αξιοποιούν τα μεγάλα δεδομένα για τη διαμόρφωση οικονομικών στρατηγικών. Σχεδόν όλα τα χρηματοπιστωτικά ιδρύματα δίνουν εξέχουσα σημασία στην ανάλυση των δεδομένων που έχουν στα χέρια τους, με σκοπό αφενός την αποφυγή ζημιογόνων επενδύσεων (κόκκινα δάνεια), αφετέρου τον εντοπισμό των κατάλληλων ομάδων-στόχων (target groups) στα οποία πρέπει να απευθύνουν τα προιόντα και τις υπηρεσίες τους.

- *Μεταφορές*

Τα μεγάλα δεδομένα δεν θα μπορούσαν να λείπουν από ένα τόσο σημαντικό και συνεχώς εξελισσόμενο κλάδο της κοινωνίας. Πολλές μεταφορικές εταιρείες έχουν αναπτύξει αλγορίθμους εύρεσης των συντομότερων διαδρομών με στόχο την μείωση των μεταφορικών εξόδων. Η UPS, μια εταιρεία που έχει εξελιχθεί σε κολοσσό στο κλάδο των μεταφορών και της παράδοσης δεμάτων, έχει αναπτύξει ένα πρωτοπόρο σύστημα βελτιστοποίησης διαδρομής παράδοσης των δεμάτων, βασισμένο στην ανάλυση των δεδομένων που συλλέγονται καθημερινά από τα οχήματα της. Με περισσότερους από 55.000 οδηγούς να καλύπτουν χιλιάδες χιλιόμετρα καθημερινά, η εξεύρεση των συντομότερων διαδρομών ήταν ένα στοίχημα. Το σύστημα Orion που ανέπτυξε η εταιρεία, εντοπίζει τη θέση των οχημάτων κάθε στιγμή, την αιτία που προκαλεί την όποια καθυστέρηση και τις εναλλακτικές-συντομότερες διαδρομές, με τη βοήθεια αισθητήρων και GPS που έχουν ενσωματωθεί σε αυτά. Ενδεικτικά αξίζει να αναφερθεί ότι με το τρόπο αυτό η UPS κατόρθωσε να μειώσει την κατανάλωση καυσίμων κατά 10 εκατομμύρια γαλόνια ετησίως, που μεταφράζεται σε μείωση κόστους περί των 400 εκατομμυρίων δολαρίων.

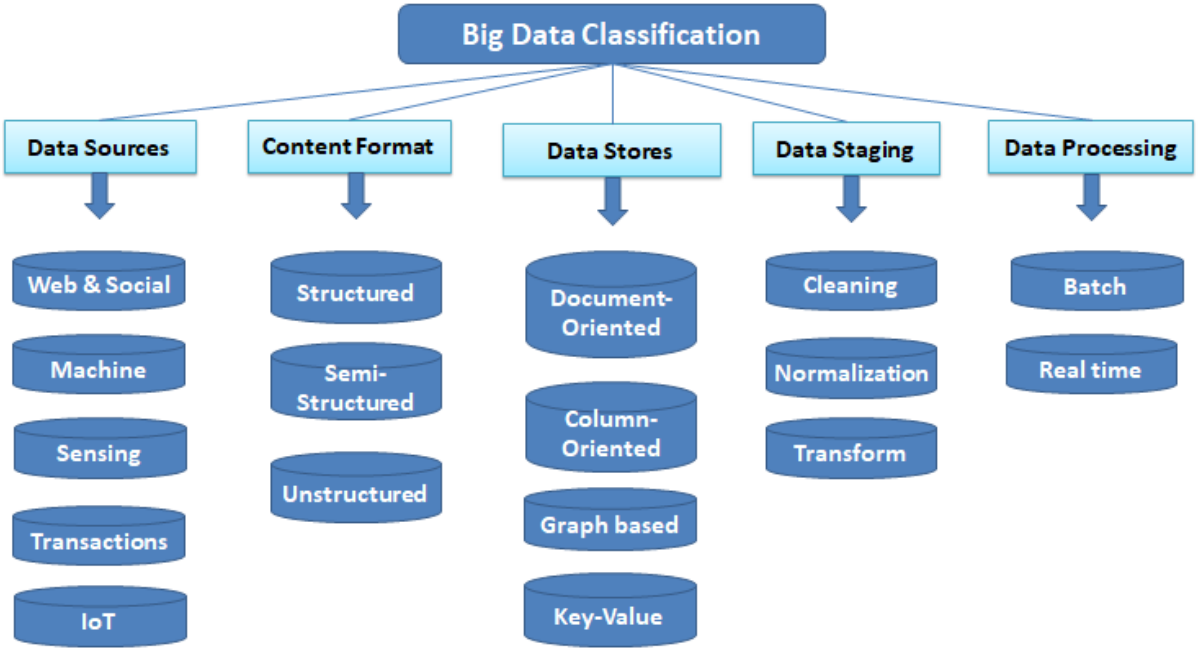
- *Αθλητισμός*

Τα μεγάλα δεδομένα δεν θα μπορούσαν να μη κάνουν αισθητή την παρουσία τους και στο χώρο του αθλητισμού. Υπολογίζεται ότι η μέση αξία της βιομηχανίας του επαγγελματικού αθλητισμού ξεπερνά τα 90 δισεκατομμύρια δολάρια. Όπως γίνεται αντιληπτό οι άνθρωποι του χώρου δίνουν μεγάλη αξία στη σημασία των μεγάλων δεδομένων και στη βοηθεία που μπορούν να τους προσφέρουν στη λήψη αποφάσεων. Παρατηρήθηκε για παράδειγμα, ότι τα προπονητικά επιτελεία θα ήταν ιδαίτερα επωφελές να γνωρίζουν σε πραγματικό χρόνο τη φυσική κατάσταση και τις ενέργειες των αθλητών τους. Κατά αυτό τον τρόπο, θα μπορούν να αποφεύγονται τυχόν τραυματισμοί, υπερκόπωση και λάνθασμένες αποφάσεις. Χαρακτηριστικό παράδειγμα εφαρμογής στο χώρο του ποδοσφάιρου το εργαλείο της STATSports. Πρόκειται για ένα μηχάνημα (Viper Pod), το οποίο φοριέται στο στήθος των αθλητών και στέλνει όλα τα δεδομένα του αθλητή (παλμοί, επίπεδο στρες, ταχύτητα, επιτάχυνση, ισορροπία βημάτων κτλ) σε ένα κεντρικό σύστημα, σε πραγματικό χρόνο. Τα δεδομένα μεταφέρονται μέσω Bluetooth ή μέσω «έξυπνων ρολογιών» σε συσκευές των προπονητικών επιτελείων, δίνοντας τους τη δυνατότητα να γνωρίζουν ανά πάσα στιγμή τη κατάσταση του αθλητή.

1.1.3 Κατηγορίες Μεγάλων Δεδομένων

Στο σημείο αυτό, για τη βαθύτερη κατανόηση των μεγάλων δεδομένων, θα γίνει μια εκτενής αναφορά στις κατηγορίες στις οποίες χωρίζονται. Οι πέντε κυριότερες κατηγορίες είναι οι ακόλουθες όπως φαίνονται και στην εικόνα 5.

i. Πηγές δεδομένων
ii. Δομή περιεχομένων
iii. Αποθήκες δεδομένων
iv. Σταδιοποίηση δεδομένων
v. Επεξεργασία δεδομένων



Εικόνα 5 - Κατηγορίες Μεγάλων Δεδομένων

i. *Πηγές δεδομένων*

- Μέσα κοινωνικής δικτύωσης

Στα μέσα κοινωνικής δικτύωσης (Facebook, Twitter, Instagram) δημιουργούνται εκατοντάδες GB πληροφορίας καθημερινά, που διαδίδονται μέσω του διαδικτύου σε όλο το εύρος του πλανήτη.

- Μηχανικά δεδομένα

Πρόκεινται για πληροφορίες που δημιουργούνται από αυτοματοποιημένα μηχανήματα καταγραφής δεδομένων, όπως υπολογιστές, δορυφόρους κα.

- Δεδομένα αισθητήρων

Δεδομένα που καταγράφονται από αισθητήρες και μηχανές ανίχνευσης για τη καταγραφή μεταβολών φυσικών μεγεθών. Παράδειγμα αισθητήρες κίνησης σε συστήματα συναγερμών ή αισθητήρες πίεσης σε καμπίνες αεροσκαφών.

- Συναλλαγές

Δεδομένα που καταγράφονται από συναλλακτικές κινήσεις, κυρίως στις μέρες μας, μέσω πιστωτικών και χρεωστικών καρτών αλλά και εμβασμάτων.

- Διαδίκτυο των πραγμάτων

Με τον όρο διαδίκυτο των πραγμάτων (Internet of Things) μπορεί να οριστεί η διαδικασία «διασύνδεσης» όλων των αντικειμένων μέσω του διαδικτύου. Βασικός σκοπός είναι η μεταφορά πληροφοριών μεταξύ συσκευών που ανήκουν στο ίδιο δίκτυο, όπως κάμερες, συστήματα ασφαλείας, οικιακές συσκευές (κλιματιστικά-φώτα), ακόμα και αυτοκίνητα. Κύριο χαρακτηριστικό των παραπάνω είναι η σύνδεση τους με σκοπό τη δυνατότητα κάθε χρήστη να τα ελέγχει από τον υπολογιστή του ή ένα «έξυπνο» κινητό [3].

ii. ***Δομή Περιοχομένων***

- Δομημένα (Structured)

Πρόκειται επί της ουσίας για «τακτοποιημένα» δεδομένα, με την έννοια ότι μπορούν να ιεραρχηθούν και να προσπελαστούν με σχετική ευκολία. Χρησιμοποιούν κατά βάση SQL, η οποία είναι μια γλώσσα προγραμματισμού που έχει δημιουργηθεί για τις σχεσιακές βάσεις δεδομένων. Παραδείγματα δομημένων δεδομένων είναι τα στοιχεία πελατών και συναλλαγών μιας επιχείρησης.

- Ημι-δομημένα (Semi-structured)

Τα ημι-δομημένα δεδομένα διαθέτουν κάποιου είδους δομή, ωστόσο δεν ακολουθούν μια συμβατική βάση δεδομένων, από τη στιγμή που οι πληροφορίες που συλλέγονται δεν έχουν όλες ταυτόσημη δομή. Χαρακτηριστικό παράδειγμα ημι-δομημένων δεδομένων είναι οι ιστοσελίδες και οι πίνακες σε υπολογιστικά φύλλα.

- Αδόμητα (Unstructured)

Τα αδόμητα δεδομένα δεν ακολουθούν καμία προκαθορισμένη μορφή, και ως εκ τούτου η διαχείρισή τους καθίσταται εξαιρετικά περίπλοκη. Αντιπροσωπεύουν σχεδόν το 90% του συνόλου δεδομένων, για το λόγο αυτό η ανάλυση τους κρίνεται εξαιρετικά σημαντική. Εικόνες, βίντεο, δεδομένα κινητής τηλεφωνίας και αισθητήρων αποτελούν μερικές ενδεικτικές μορφές αδόμητων δεδομένων.

iii. Αποθήκες δεδομένων

Πέρα από τη κλασσική μορφή αποθήκευσης δεδομένων σε μορφή πινάκων (γραμμές και στήλες), που συναντάται στο σχεσιακό μοντέλο, τα δεδομένα αποθηκεύονται και στις ακόλουθες μορφές (μή σχεσιακό μοντέλο):

- Προσανατολισμένες σε έγγραφα (Document-oriented)

Προτιμάται για δεδομένα που ακολουθούν τη μορφή εγγράφων (παραστατικά, συνταγές φαρμάκων, παραπεμπτικά κτλ). Υποστηρίζει δεδομένα σε διάφορους τύπους όπως JSON, XML, PDF, MS Word. Αυτός ο τρόπος αποθήκευσης προσφέρει ευελιξία και ταχύτερη ανάκτηση περιεχομένου. Οι πιο γνωστές αποθήκες είναι οι ακόλουθες: MongoDB, CouchDB, Couchbase, RethinkDB, Terrastore, Elasticsearch.

- Προσανατολισμένες σε στήλες (Column-oriented)

Προτιμάται για πίνακες δεδομένων, πολύ μεγάλου μεγέθους, στους οποίους συνήθως ανατίθενται ερωτήματα που αφορούν μόνο λίγες από τις διαθέσιμες στήλες. Εδώ τα δεδομένα αποθηκεύονται εκτός από γραμμές και σε στήλες. Ενδεικτικές αποθήκες τέτοιου είδους είναι οι BigTable, Hbase, Hypertable και Cassandra.

- Βάση δεδομένων γραφημάτων (Graph)

Προτιμάται για δεδομένα που έχουν μεγάλη συσχέτιση μεταξύ τους και η συσχέτιση αυτή πρέπει να αποτυπωθεί όπως συμβαίνει με τα δεδομένα κοινωνικού δικτύου. Γνωστότερες βάσεις γραφημάτων θεωρούνται οι Neo4j, FlockDB, OrientDB, AllegroGraph και GraphDB.

- Βάση δεδομένων κλειδιού-τιμής (Key-Value)

Προτιμάται για δεδομένα απλής φύσης, που μπορούν να αναπαρασταθούν με τη μορφή κλειδιού – τιμής, όπως ένα καλάθι αγορών για παράδειγμα. Οι Memcached, membase, Voldemort και Redis είναι μερικές από τις χαρακτηριστικότερες βάσεις αυτού του είδους.

iv. ***Σταδιοποίηση δεδομένων***

- Καθαρισμός (Cleaning)

Εντοπισμός και αφαίρεση/αντικατάσταση ελλειπουσών τιμών. Αφαίρεση θορύβου και παράλογων στοιχείων.

- Κανονικοποίηση (Normalization)

Διαδικασία μετατροπής των δεδομένων σε μία ακολουθία κανονικών μορφών, οι οποίες αποτελούνται από απλές και σαφείς σχέσεις που δεν περιέχουν επαναλήψεις.

- Μετασχηματισμός (Transformation)

Μετατροπή των δεδομένων σε μορφή ευνοική προς επεξεργασία και ανάλυση.

v. ***Επεξεργασία***

- Κατά δεδμίδες (Batch)

Στην επεξεργασία κατά δεσμίδες, τα δεδομένα αποθηκεύονται σε συστάδες για ένα χρονικό διάστημα (από 1ώρα-μερικές ημέρες), μέχρι να υποστούν ανάλυση. Εναλλακτικά, τίθεται ένα ανώτατο όριο (threshold), το οποίο όταν επιτευχθεί στέλνει τα συγκεντρωμένα δεδομένα προς επεξεργασία [4].

- Σε πραγματικό χρόνο (Stream)

Στην επεξεργασία σε πραγματικό χρόνο, τα δεδομένα αναλύονται με το που φτάσουν στο χώρο αποθήκευσης, σχεδόν αμέσως μετά τη δημιουργία τους. Τα γνωστότερα εργαλεία επεξεργασίας σε πραγματικό χρόνο είναι η KSQL. Αυτό που στην ουσία κάνει η KSQL είναι συνεχείς μετασχηματισμούς στα δεδομένα μέσω επερωτήσεων, ενώ παράλληλα τα δεδομένα ανανεώνονται συνεχώς [5].

1.2 Μεγάλα Δεδομένα: Συστήματα αποθήκευσης και επεξεργασίας

Η αποθήκευση των μεγάλων δεδομένων αποτελεί πρόκληση για τις μεγάλες εταιρείας επεξεργασίας δεδομένων, λόγω του μεγάλου όγκου και της ταχύτητας με την οποία παράγονται. Η επένδυση σε τοπικούς server και συμπλέγματα διακομιστών (cluster) αποτελεί ρίσκο, λόγω του πεπερασμένου σε χωρητικότητα διαθέσιμου χώρου αλλά και του πιθανού αυξημένου χρόνου επεξεργασίας των δεδομένων. Για το λόγο αυτό, ολοένα και περισσότερες εταιρείες επιλέγουν στις μέρες μας να επενδύουν σε συστήματα αποθήκευσης δεδομένων το υπολογιστικό νέφος (cloud).

Με τον όρο υπολογιστικό νέφος περιγράφεται η διάθεση υπολογιστικών πόρων μέσω διαδικτύου από κεντρικά συστήματα τα οποία βρίσκονται απομακρυσμένα από τους τελικούς χρήστες. Για τους επαγγελματίες χρήστες της τεχνολογίας των πληροφοριών το υπολογιστικός νέφος σημαίνει μεγάλη ευελιξία όσον αφορά τις ανάγκες υπολογιστικής ισχύος, από τη στιγμή που σε περίπτωση αυξημένης χρήσης μιας υπηρεσίας είναι εύκολο να προστεθεί επιπλέον δυναμικό. Η Microsoft, η Amazon και η Google είναι από τους βασικότερους παρόχους υπολογογιστικού νέφους που παρέχουν τέτοιου είδους υπηρεσίες σε μεγάλη κλίμακα μέσω των πλατφορμών:

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)

CLOUD COMPUTING	
ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
1. Χαμηλό κόστος απόδοσης	1. Τεχνικά προβλήματα
2. Σχεδόν απεριόριστη αποθήκευση	2. Ασφάλεια
3. Δημιουργία αντιγράφων ασφαλείας	3. Επιρρρέπεια σε επιθέσεις
4. Αυτόματη ενσωμάτωση λογισμικού	4. Πιθανός χρόνος διακοπής
5. Εύκολη πρόσβαση	5. Έλλειψη υποστήριξης
6. Εύκολη ανάπτυξη	

Πίνακας 1 – Πλεονεκτήματα-Μειονεκτήματα Υπολογιστικού Νέφους

Τα δεδομένα αποθηκεύονται στο υπολογιστικό νέφος στα ακόλουθα συστήματα:

- *Σχεσιακές βάσεις δεδομένων (RDBMS)*
- *Μη σχεσιακές βάσεις δεδομένων*
- *Κατανεμημένα συστήματα αρχείων:*

  Τα κατανεμημένα συστήματα αρχείων, όπως το σύστημα αρχείων Hadoop (HDFS), προσφέρουν τη δυνατότητα αποθήκευσης μεγάλων ποσοτήτων μη δομημένων δεδομένων με αξιόπιστο τρόπο. Επειδή τα δεδομένα γράφονται μία φορά και στη συνέχεια διαβάζονται πολλές φορές, αντί για τις συνεχείς αναγνώσεις άλλων συστημάτων αρχείων, το HDFS (Hadoop Distributed File System) παρουσιάζεται ως μια εξαιρετική επιλογή για την υποστήριξη της μεγάλης ανάλυσης δεδομένων. Τα μεγάλα δεδομένα φέρνουν τις μεγάλες προκλήσεις του όγκου, της ταχύτητας και της ποικιλίας στις βάσεις δεδομένων. Το HDFS αντιμετωπίζει αυτές τις προκλήσεις με το «σπάσιμο» των αρχείων αυτών σε μια σχετική συλλογή μικρότερων ομάδων που αποτελούνται από δεδομένα.

- *Μηχανές αναζήτησης (Query platforms):*

  Η **Hive** (κυψέλη) αποτελεί είδος αποθήκευσης δεδομένων βασισμένο στον πυρήνα του Hadoop. Η Hive επιτρέπει σε δομημένα αρχεία να ερωτηθούν από μια απλή εφαρμογή SQL-lite. Επίσης παρέχει μεγάλη ευελιξία, καθώς τα σχήματα μπορούν να μεταβάλλονται εύκολα από τη στιγμή που αποθηκεύονται ανεξάρτητα από τα δεδομένα.Τα δεδομένα με τη σειρά τους επικυρώνονται μόνο κατά τη περίοδο του ερωτήματος. Η προσέγγιση αυτή είναι γνωστή και ως schema-on-read. Η μηχανή αναζήτησης **Impala** έχει, επίσης, σχεδιαστεί για την εκτέλεση ερωτημάτων με χαμηλότερους χρόνους καθυστέρησης. Η **Spark SQL** αποτελεί μια πλατφόρμα με αξιοσημείωτα χαμηλό χρόνο απόκρισης και υποστηρίζει τη διασύνδεση με τη Hive. Μπορεί να εκτελέσει ερωτήματα ακόμα και 100 φορές ταχύτερα από τη μηχανή αναζήτησης Hive, χωρίς καμιά τροποποίηση στα υπάρχοντα δεδομένα. Αυτό

επιτυγχάνεται με την εκτέλεση των ερωτημάτων χρησιμοποιώντας το πλαίσιο Spark και όχι το πλαίσιο MapReduce του Hadoop. Τέλος, το **Drill** είναι μια υλοποίηση ανοικτού κώδικα του Dremel της Google, που παρόμοια με την Impala έχει σχεδιαστεί ως ένα επεκτάσιμο, διαδραστικό ad-hoc σύστημα ερωτημάτων για εμφωλευμένα δεδομένα. Το Drill παρέχει τη δική του γλώσσα ερωτήματος που μοιάζει με SQL και έχει σχεδιαστεί για να υποστηρίζει άλλες γλώσσες ερωτήματος, όπως η γλώσσα ερωτήματος Mongo. Σε αντίθεση με την Hive και την Impala, υποστηρίζει μια σειρά από πηγές δεδομένων χωρίς σαφές σχήμα, όπως HDFS, HBase, Κασσάνδρα και MongoDB [6].

1.3 Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης

1.3.1 Ορισμός Εξόρυξης Δεδομένων - Ανακάλυψης Γνώσης

Μέχρι στιγμής έχει γίνει λόγος για την εκθετική αύξηση που σημειώνει ο ρυθμός καταγραφής των δεδομένων. Πέρα, λοιπόν, από την συλλογή και αποθήκευση τους, ένα άλλο μείζον ζητήμα είναι η ανάλυση αυτών. Οι τεράστιοι όγκοι δεδομένων που συλλέγουν καθημερινά οι επιχειρήσεις στις βάσεις τους, από μόνοι τους δεν έχουν απολύτως καμία αξία. Στόχος είναι η μετατροπή των δεδομένων αυτών σε χρήσιμες γνώσεις προσανατολισμένες στις εργασίες της εκάστοτε επιχείρησης. Η επεξεργασία των δεδομένων χωρίς εξειδικευμένα εργαλεία είναι μια διαδικασία ακριβή, αργή και πολλές φορές μη αντικειμενική. Οι εώς σήμερα υφιστάμενες στατιστικές αναλύσεις δίνουν λύση στην επεξεργασία των δεδομένων, ωστόσο δεν επιλύουν το ζήτημα του μεγάλου όγκου αυτών. Η μηχανική μάθηση δεν αντιμετωπίζει, επίσης, το πρόβλημα αυτό. Οι βάσεις δεδομένων, με τη σειρά τους, είναι υπεύθυνες για τη συλλογή, αποθήκευση και ανάκτηση των δεδομένων, όχι όμως και για την ανάλυση τους. Η εξόρυξη δεδομένων ήρθε, λοιπόν, να δώσει λύση στο ζήτημα αυτό. Βασισμένη στους επιστημονικούς κλάδους που προαναφέρθηκαν και σε συνδυασμό με την οπτικοποίηση, η εξόρυξη δεδομένων παρέχει χρήσιμη γνώση υπό τη μορφή συσχετίσεων , προτύπων και τάσεων.

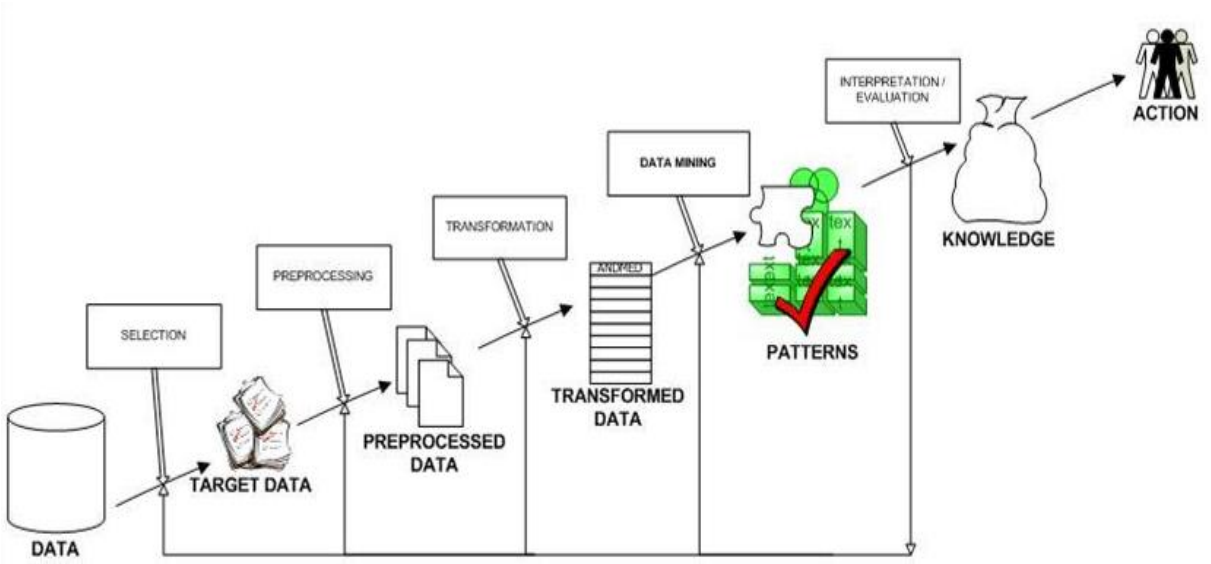
Εικόνα 6 - Επιστημονικοί Κλάδοι που Συμβάλλουν στην Εξόρυξη Δεδομένων

Το 2000 οι Witten και Frank όρισαν την Εξόρυξη Δεδομένων (Data Mining) ως τη διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, τονίζοντας τη διάσταση του όγκου τους. Το 2005 οι Maimon και Rokack εισήγαγαν μία νέα έννοια, γνωστή και ως Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD). Με τον όρο αυτό περιγράφεται η συνολικότερη διαδικασία ανακάλυψης προτύπων μέσα από περίπλοκα και μεγάλα σε όγκο σύνολα δεδομένων. Η διαδικασία αυτή ξεκινάει από τη συλλογή των αποθηκευμένων δεδομένων και καταλήγει στη τελική διατύπωση συμπερασμάτων και παρατηρήσεων με σκοπό τη λήψη ορθών αποφάσεων. Σημαντικό κομμάτι της ανακάλυψης γνώσης αποτελεί και η εξόρυξη δεδομένων, η οποία περιλαμβάνει την εφαρμογή αλγορίθμων μηχανικής μάθησης και στατιστικών αναλύσεων με σκοπό την εξαγωγή προτύπων. Τέλος, η αποθήκευση και η ανάκληση των δεδομένων, η κλιμάκωση των αλγορίθμων και η οπτικοποίηση των αποτελεσμάτων αποτελούν επιμέρους κομμάτια των ενεργειών που πραγματεύεται η ανακάλυψη γνώσης.

1.3.2 Στάδια Ανακάλυψης Γνώσης

Η διαδικασία ανεύρεσης γνώσης αποτελεί μια επαναληπτική διαδικασία που απαρτίζεται από μια σειρά βημάτων που ξεκινούν από τη συλλογή των δεδομένων και φτάνουν έως την εξαγωγή και οπτικοποίηση της χρήσιμης πληροφορίας. Παρακάτω περιγράφονται συνοπτικά τα επιμέρους βήματα της διαδικασίας ανεύρεσης γνώσης.

1. **Επιλογή δεδομένων (Data selection)**: Από τις διάφορες ετερογενείς πηγές συγκεντρώνονται και επιλέγονται τα δεδομένα εκείνα που είναι σχετικά με την επικείμενη ανάλυση.
2. **Προεπεξεργασία (Data preprocessing)**: Το βήμα αυτό περιλαμβάνει τον καθαρισμό και την ενσωμάτωση των δεδομένων. Αρχικά αφαιρούνται τα δεδομένα που παράγουν θόρυβο, έτσι ώστε να μην αλλοιώσουν το αποτέλεσμα και εν συνεχεία ενσωματώνονται σε κοινά σύνολα-βάσεις δεδομένων.
3. **Μετασχηματισμός (Data transformation)**: Στο στάδιο αυτό τα δεδομένα υφίστανται τις απαραίτητες τροποποιήσεις έτσι ώστε να λάβουν τη κατάλληλη μορφή προς επεξεργασία και να προσαρμοστούν στις απαιτήσεις των μεθόδων ανάλυσης.
4. **Εξόρυξη δεδομένων (Data Mining)**: Πρόκειται για το σημαντικότερο βήμα της ανακάλυψης γνώσης, κατά το οποίο αναπτύσσονται εξελιγμένες τεχνικές (περιγραφικές ή προβλεπτικές) για την εξαγωγή χρήσιμων προτύπων.
5. **Αξιολόγηση προτύπων (Pattern Evaluation)**: Ο αναλυτής με χρήση συγκεκριμένων μέτρων αξιολόγησης αναγνωρίζει τα εξαγόμενα πρότυπα και επανέρχεται σε προηγούμενα στάδια, αν κρίνει ότι τα αποτελέσματα δεν είναι ικανοποιητικά.
6. **Αναπαράσταση γνώσης (Knowledge Representation)**: Λαμβάνει χώρα η οπτικοποίηση των αποτελεσμάτων και η παρουσίασή τους στον χρήστη, με τελικό στόχο την άρτια λήψη αποφάσεων.



Εικόνα 7 - Στάδια Ανακάλυψης Γνώσης

1.3.3 Μέθοδοι Εξόρυξης Δεδομένων

Στόχος της εξόρυξης δεδομένων είναι η δημιουργία προτύπων διαφόρων μορφών, με βάση πάντοτε τα διαθέσιμα δεδομένα. Οι εργασίες που λαμβάνουν χώρα κατά τη διάρκεια της εξόρυξης δεδομένων μπορούν να χωριστούν στις δύο κύριες κατηγορίες της επιβλεπόμενης (supervised) και μή επιβλεπόμενης μάθησης (unsupervised) μάθησης. Η επιβλεπόμενη μάθηση μοντελοποιεί τη συσχέτιση ενός εξαρτημένου γνωρίσματος με άλλα ανεξάρτητα. Το μοντέλο της επιβλεπόμενης μάθησης μπορεί, επομένως, να χαρακτηριστεί ως ένα προβλεπτικό μοντέλο. Αντιθέτως, στη μη επιβλεπόμενη μάθηση δεν υπάρχει κάποιο γνώρισμα στόχος και οι αλγόριθμοι επικεντρώνονται στην ομαδοποίηση των δεδομένων που έχουν κοινά γνωρίσματα. Θεωρείται, λοιπόν, ένα περιγραφικό μοντέλο εκμάθησης.

ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ	ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ
1. Ταξινόμηση	1. Συσταδοποίηση
2. Παλινδρόμηση	2. Ανάλυση Κανόνων Συσχέτισης
3. Ανάλυση Εξαιρέσεων	3. Εντοπισμός προτύπων
4. Ανάλυση Χρονοσειρών	4. Διαδοχικά μοτίβα

Πίνακας 2 – Τεχνικές Επιβλεπόμενης και Μη Επιβλεπόμενης Μάθησης

Η **ταξινόμηση** ή αλλιώς **κατηγοριοποίηση** πρόκειται για μία από τις βασικότερες μεθόδους επιβλεπόμενης μάθησης στον τομέα της εξόρυξης δεδομένων. Σε τέτοιου είδους προβλήματα είναι γνωστό εξ αρχής ότι τα δεδομένα ανήκουν σε κατηγορίες και υπάρχουν μία ή (και μερικές φορές) περισσότερες στήλες-στόχοι. Η ανάλυση επικεντρώνεται στη τυποποίηση των σχέσεων μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών, έτσι ώστε να προβλέπεται με όσο το δυνατόν μεγαλύτερη ακρίβεια η ορθή κλάση-κατηγορία του αντικειμένου. Χαρακτηριστικό παράδειγμα είναι η εξέταση των χαρακτηριστικών των πελατών μιας τράπεζας (ηλικία, φύλο, εισόδημα, περιουσιακή κατάσταση) ώστε να προβλεφθεί η συνέπεια τους στην αποπληρωμή ενός δανείου.

Η **παλινδρόμηση** αποτελεί σημαντική μέθοδο επιβλεπόμενης μάθησης και εμφανίζει μεγάλη ομοιότητα με τη ταξινόμηση. Στην παλινδρόμηση, όπως ακριβώς και στη κατηγοριοποίηση, υπάρχει ένα χαρακτηριστικό-στόχος που θεωρείται εξαρτημένο από μια σειρά άλλων χαρακτηριστικών και γίνεται προσπάθεια να αναπτυχθεί ένας μηχανισμός πρόβλεψης βάση αυτών. Η διαφορά έγκειται στο γεγονός ότι στη περίπτωση της παλινδρόμησης υπολογίζονται αριθμητικά μεγέθη. Ένα παράδειγμα είναι οι τεχνικές παλινδρόμησης για τη πρόβεψη αριθμητικών μεγεθών, όπως το ισοζύγιο εσόδων-εξόδων (τζίρος) μιας επιχείρησης.

Η **ανάλυση εξαιρέσεων** βασίζεται κυρίως στις εργασίες που έχουν σπάνια εμφάνιση και ιδιότυπα χαρακτηριστικά. Οι τεχνικές κατηγοριοποίησης που αναφέρθηκαν έως τώρα ασχολούνται με τη διατύπωση προτύπων βασισμένα σε γεγονότα που συμβαίνουν κατά πλειοψηφία και αποτελούν σχεδόν τον κανόνα. Ορισμένες φορές, ωστόσο, το ενδιαφέρον εστιάζεται στην εξαίρεση του κανόνα. Για να γίνει πιο κατανοητό, ένα απλό παράδειγμα είναι ο εντοπισμός των κινήσεων με κλεμμένες πιστωτικές κάρτες. Στη συντριπτική πλειοψηφία, καθημερινά, το σύνολο των κινήσεων με πιστωτικές κάρτες πρόκειται για φυσιολογικές συμπεριφορές. Υπάρχει όμως μεγάλη ανάγκη να εντοπίζονται άμεσα και με ακρίβεια οι ύποπτες κινήσεις με πιστωτικές κάρτες, ακόμα κι αν αποτελούν πολύ μικρό ποσοστό των κινήσεων, από τη στιγμή που ο χρηματοικονομικός κλάδος είναι ευάισθητος σε τέτοια ζητήματα και δεν επιτρέπει απώλειες.

Η **ανάλυση χρονοσειρών** εστιάζει στα μεγέθη που εμφανίζουν μια χρονική εξάρτηση. Ουσιαστικά πρόκειται για μια σειρά από παρατηρήρεις που λαμβάνονται σε ορισμένες χρονικές στιγμές και επεξηγούν τη κίνηση μιας μεταβλητής συναρτήσει του χρόνου. Επίσης η ανάλυση χρονοσειρών εξετάζει αν μια χρονοσειρά μπορεί να εμφανίζει τάση (αργές μεταβολές στη μέση τιμή με το χρόνο), περιοδικότητα (μεταβολές σε συγκεκριμένες περιόδους: μήνα-τρίμηνο-τετράμηνο) ή και εποχικότητα (μεταβολές σε μεγάλα χρονικά διαστήματα πέραν του έτους). Ενδεικτικό παράδειγμα είναι η κίνηση των δεικτών τιμών του χρηματιστηρίου.

Η **συσταδοποίηση** είναι η βασικότερη μέθοδος μή επιβλεπόμενης μάθησης. Στόχος της συσταδοποίησης είναι ο επιμερισμός ενός συνόλου αντικειμένων σε επιμέρους ομάδες με κοινά χαρακτηριστικά. Γίνεται προσάθεια, δηλαδή, να αυξηθεί η ομοιότητα εντός τον ομάδων και η ανομοιότητα μεταξύ αυτών. Αξίζει να σημειωθεί ότι, σε αντίθεση με την κατηγοριοποίηση, δεν είναι γνωστή εκ των προτέρων η κατηγορία στην οποία ανήκει κάθε αντικείμενο.

Η **ανάλυση κανόνων συσχέτισης** αναφέρεται στην ανακάλυψη σχέσεων μεταξύ των διαφόρων γνωρισμάτων και παίζει εξέχοντα ρόλο κυρίως στους τομείς του ηλεκτρονικού εμπορίου και του μάρκετινγκ. Ένα παράδειγμα για να γίνει πιο κατανοητή η έννοια της ανάλυσης συσχετίσεων είναι οι πωλήσεις ενός σούπερ μάρκετ. Συχνά, κατόπιν αναλύσεων, παρατηρείται ότι οι καταναλωτές αγοράζουν προιόντα συνδυαστικά, δηλαδή καφέ-ζάχαρη, γάλα-ψωμί κα. Κατά αυτό τον τρόπο, οι κανόνες συσχέτισης επιτρέπουν τον εντοπισμό καταναλωτικών και μη προτύπων με στόχο την προσωποποιημένη στόχευση των πελατών.

Τέλος, ο **εντοπισμός προτύπων** και τα **διαδοχικά μοτίβα,** που είναι έννοιες πολύ κοντινές μεταξύ τους, αποτελούν μεθόδους εξόρυξης δεδομένων. Ο εντοπισμός προτύπων εστιάζει στην

παρακολούθηση των τάσεων της αγοράς. Χρησιμοποείται ευρέως από τις επιχειρήσεις για την αναγνώριση ομάδων καταναλωτών που επιλέγουν συγκεκριμένα προιόντα, με στόχο την βελτίωση των ήδη παρεχόμενων υπηρεσιών ή τη δημιουργία νέων βασισμένων στην υπάρχουσα γνώση. Τα διαδοχικά μοτίβα επικεντρώνονται στα γεγονότα που παρουσιάζουν μια ακολουθία και είναι ιδιαίτερα χρήσιμη στον τομέα των εμπορικών συναλλαγών.

ΚΕΦΑΛΑΙΟ 2: Εφαρμογή Εξόρυξης Δεδομένων στον Τραπεζικό Τομέα

2.1 Τομείς Εφαρμογής Εξόρυξης Δεδομένων

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο, η Εξόρυξη Δεδομένων είναι ένας νεοσύστατος επιστημονικός κλάδος που στοχεύει στην ανακάλυψη γνώσης και πληροφορίας από τα δεδομένα. Οι επιχειρήσεις κατακλίζονται από τεράστιους όγκους δεδομένων, χωρίς ωστόσο, να έχουν κάποια αξία, αν προηγουμένως δεν έχουν επεργαστεί και αναλυθεί. Η άρτια και ενδελεχής ανάλυση των δεδομένων μπορεί να οδηγήσει τις επιχειρήσεις στη βαθύτερη κατανόηση των τάσεων της αγοράς, τη μείωση του ρίσκου κατά τη λήψη αποφάσεων και στη βελτίωση της ποιότητας των υπηρεσιών που παρέχουν. Η γνώση που αποκομίζεται από τα δεδομένα μπορεί να αξιοποιηθεί με τελικό στόχο την αύξηση της κερδοφορίας, τη μείωση του κόστους παραγωγής και τέλος την αύξηση της καινοτομίας. Η Εξόρυξη Δεδομένων, συνεπώς, έχει γίνει αναπόσπαστο κομμάτι των επιχειρησιακών διεργασιών στις μέρες μας και βρίσκει πεδίο εφαρμογής σε πολλαπλούς τομείς, όπως στο εμπόριο-πωλήσεις, στις τράπεζες, στην υγεία και τις τηλεπικοινωνίες.

Πωλήσεις-Διαφημίσεις

Ο τομέας των πωλήσεων αποτελεί έναν από τα πιο δημοφιλή πεδία εφαρμογής της Εξόρυξης Δεδομένων. Η αύξηση της κερδοφορίας των επιχειρήσεων στηρίζεται σε δύο κύριους άξονες που είναι η διεύρυνση του πελατολογίου και η διατήρηση των υπαρχόντων πελατών. Η εξόρυξη δεδομένων χρησιμοποιείται σε τεχνικές όπως οι διασταυρούμενες πωλήσεις, η αναγνώριση της καταναλωτικής συμπεριφοράς των πελατών και η στοχευμένη διαφήμιση. Οι διασταυρούμενες πωλήσεις είναι μία τεχνική αύξηση των πωλήσεων προιόντων ή υπηρεσιών που λειτουργούν συνδυαστικά. Δηλαδή, η αγορά μιας συσκευής κινητής τηλεφωνίας έχει παρατηρηθεί ότι συνοδεύεται στην πλειονότητα των περιπτώσεων από την αγορά προστατευτικών θηκών ή ακουστικών. Ωστόσο, υπάρχουν και συνδυασμοί που δεν είναι τόσο έκδηλοι και κρίνεται αναγκαία η ενδελεχής ανάλυση των προς διάθεση δεδομένων. Η αναγνώριση της καταναλωτικής συμπεριφοράς είναι εξίσου σημαντική στον τομέα των πωλήσεων. Βασίζεται εν πολλοίς στην ανάλυση κανόνων συσχέτισης με σκοπό τη βαθύτερη κατανόηση του τί και πότε αγοράζεται από τους πελάτες και παίζει σημαντικό ρόλο στον τρόπο προβολής και διαφήμισης των προιόντων. Ο χώρος των σουπερ μάρκετ στηρίζεται σε μεγάλο βαθμό σε τέτοιου είδους τεχνικές εξόρυξης γνώσης, από τον τρόπο τοποθέτησης των

προιόντων στα ράφια έως και τις προσφορές συνδυαστικής αγοράς προιόντων. Τέλος, η στοχευμένη διαφήμιση βασίζεται στη τμηματοποίηση του αγοραστικού κοινού σε επιμέρους ομάδες με ομοειδή χαρακτηριστικά. Έχοντας γνώση των χαρακτηριστικών κάθε ομάδας, οι επιχειρήσεις μπορούν να προκρίνουν στρατηγικές στοχευμένης διάθεσης και διαφήμισης προιόντων, βασισμένες στις ανάγκες κάθε ομάδας.

Ηλεκτρονικό εμπόριο

Οι τεχνικές εξόρυξης γνώσης βρίσκουν σημαντικότατο πεδίο εφαρμογής και στο χώρο του ηλεκτρονικού εμπορίου. Το γεγονός ότι οι αγοραπωλησίες διεξάγονται ηλεκτρονικά, επιτρέπει σε μεγάλο βαθμό την καταγραφή ενδελεχών λεπτομερειών των συναλλαγών, όπως προηγούμενες αγορές, κατηγορίες προιόντων που επελέγησαν κα. Η καταγραφή της διαδρομής και περιήγησης των χρηστών στους ιστοχώρους προσφέρει σημαντικές πληροφορίες σχετικά με το προφίλ του χρήστη-αγοραστή, και σκιαγραφεί λεπτομερώς την καταναλωτική συμπεριφορά και τις προτιμήσεις του. Στόχος και εδώ, όπως και στο λιανικό εμπόριο, είναι η εξατομικευμένη προβολή και διαφήμιση προιόντων για την αύξηση της κερδοφορίας. Οι διασταυρούμενες πωλήσεις είναι μια ευρέως διαδεδομένη τεχνική στο χώρο του ηλεκτρονικού εμπορίου. Συχνά, επισκέπτες ηλεκτρονικών καταστημάτων λαμβάνουν μηνύματα προωθητικού χαρακτήρα που σχετίζονται με αγαθά στα οποία προηγουμένως έχουν δείξει ενδιαφέρον ή και έχουν αγοράσει. Πίσω από αυτές τις κινήσεις «κρύβονται» αλγόριθμοι ανάλυσης κανόνων συσχέτισης και συσταδοποίησης, με σκοπό τον εντοπισμό ομάδων προιόντων με πανομοιότυπα χαρακτηριστικά.

Χρηματιστήριο

Το χρηματιστήριο αξιών πρόκειται για ένα σύνθετο, ευμετάβλητο, δυναμικό και μη γραμμικό σύστημα. Τα δεδομένα που γεννώνται επί καθημερινής βάσης από τις αγοραπωλησίες μετοχών αυξάνονται εκθετικά και η αξιοποίηση τους είναι μονόδρομος για τους αναλυτές.. Για τη πρόβλεψη της διακύμανσης των τιμών των μετοχών και των δεικτών, που είναι και το ζητούμενο της επενδυτικής διαδικασίας, έχει επιστρατευτεί πληθώρα στρατηγικών κατηγοριοποίησης, συσταδοποίησης και στατιστικών μεθόδων όπως η ανάλυση των τάσεων και εν γένει των χρονοσειρών. Πρόκειται, ωστόσο, για μια δύσκολη διαδικασία πρόβλεψης καθότι η κίνηση των χρηματιστηριακών δεικτών επηρεάζεται σημαντικά από απρόβλεπτα γεγονότα της καθημερινότητας [7].

Ασφάλειες

Οι τεχνικές εξόρυξης δεδομένων βρίσκουν εφαρμογή και στον κλάδο των ασφαλειών. Πληθώρα ασφαλιστικών εταιρειών χρησιμοποιούν τέτοιου είδους τεχνικές με σκοπό τη μείωση του κόστους, την αύξηση των κερδών, τη διαχείριση και πρόβλεψη του ρίσκου, την εξεύρεση νέων πελατών και τη διατήρηση των υπαρχόντων. Η ανάλυση κανόνων συσχετισης, η κατηγοριοποίηση και η ομαδοποίηση είναι οι τρεις κύριες μέθοδοι που χρησιμοποιούνται κατά κόρον στον ασφαλιστικό τομέα. Αρχικά, αξίζει να αναφερθεί ότι οι ασφαλιστικές εταιρείες εφαρμόζουν ενδελεχείς αναλύσεις κανόνων συσχέτισης στα δεδομένα που έχουν, αποσκοπώντας στην προώθηση ομαδοποιημένων πακέτων ασφάλειας (σπιτιού-αυτοκινήτου) σε προσιτές τιμές. Επίσης, με χρήση μεθόδων συσταδοποίησης εντοπίζουν ομάδες πιθανών πελατών και με χρήση προσωποποιημένων, πελατοκεντρικών τεχνικών (διαφήμιση) προτείνουν τα κατάλληλα πακέτα στην κάθε ομάδα. Τελευταίο και σημαντικότερο σημείο εφαρμογής τεχνικών εξόδυξης δεδομένων για τις ασφαλιστικές εταιρείες, αποτελεί η διαχείριση του κινδύνου. Αρκεί μόνον να αναλογιστεί κανείς πως λειτουργέι μια ασφαλιστική εταιρεία. Η κερδοφορία αυξάνεται όταν ολοένα και περισσότεροι πελάτες συνάπτουν συμβόλαια ασφάλισης και η πλειοψία αυτών καταλήγει να μή κάνει ποτέ χρήση. Επομένως, πέρα από την αύξηση του πελατολογίου, κρίνεται σημαντικό να εντοπίζονται οι πελάτες εκείνοι, που έχουν μεγάλες πιθανότητες να μην κάνουν ποτέ χρήση της ασφάλειας τους και συνεπώς να διεκδικήσουν κάποιου είδους αποζημίωση. Κάτι τέτοιο δεν συνεπάγεται απαράιτητα ότι μια ασφαλιστική εταιρεία αποκλείει ομάδες υψηλού κινδύνου πελατών. Ωστόσο, επηρεάζει σε μεγάλο βαθμό το ύψος των ασφαλίστρων του πελάτη. Παραδείγματος χάρη, τα ασφάλιστρα ενός νέου και εν συνεπεία άπειρου οδηγού θα είναι εμφανώς υψηλότερα από εκείνα ενός έμπειρου ενήλικα που έχει συνάψει διαδοχικά συμβόλαια με την εταιρεία.

Εκπαίδευση

Τα τελευταία χρόνια οι τεχνικές εξόρυξης δεδομένων ξεκίνησαν να εφαρμόζονται και στο χώρο της εκπαίδευσης. Αρκετές μελέτες έχουν δείξει ότι η επεξεργασία και ανάλυση των δεδομένων που ανακύπτουν από τον κλάδο της εκπαίδευσης, μπορούν να οδηγήσουν στη λήψη ασφαλών συμπερασμάτων σε μέιζοντα ζητήματα. Αλγόριθμοι κατηγοριοποίησης μπορούν να προβλέψουν την επιτυχία και αποτυχία των μαθητών σε εξετάσεις, την εισαγωγή τους στο πανεπιστήμιο, ακόμα και τους λόγους της επικείμενης αποτυχίας-αποτυχίας. Επιπρόσθετα, οι μέθοδοι συσταδοποίησης επιτρέπουν την τμηματοποίηση των μαθητών σύμφωνα με το εκπαιδευτικό τους υπόβαθρο, διαδικασία που βοηθά τον εκπαιδευτικό στην κατανόηση των αναγκών των μαθητών της εκάστοτε ομάδας και στην διαμόρφωση της διδασκαλίας του

σύμφωνα με αυτές. Τέλος, με χρήση κανόνων συσχέτισης μπορούν να εντοπιστούν οι συσχετίσεις των μαθητών που εγκατέλειψαν το σχολείο-πανεπιστήμιο με τα κοινωνικο-παιδαγωγικά τους χαρακτηριστικά. Έχει παρατηρηθεί ότι παιδιά οικογενειών με οικονομικά προβλήματα, εμφανίζουν τη τάση να εγκαταλείπουν με μεγαλυτερη συχνότητα τη φοίτησης τους. Με χρήση, λοιπόν, τέτοιων μεθόδων μπορεί να κατανοηθούν βαθύτερα οι αιτίες του φαινομένου και να αντιμετωπιστεί αποτελεσματικότερα το ζήτημα της εκπαιδευτικής διαρροής [8].

Υγεία

Κύρια επιδίωξη κάθε συστήματος υγείας είναι η παροχή ιατροφαρμακευτικής περίθαλψης σε όλους τους πολίτες χωρίς διακρίσεις, σε υψηλή ποιότητα και χαμηλό κόστος. Αν και το επίπεδο των ερευνών στο χώρο της υγείας έχει σημειώσει αλματώδη πρόοδο τα τελευταία χρόνια, παραμένει μια εγγενής αδυναμία επερξασίας των δεδομένων μεγάλης κλίμακας, Η ταχεία και αποτελεσματική ανάλυση συνόλων δεδομένων από ετερογενείς πηγές αναμένεται να συμβάλει σημαντικά σε μια σειρά διεργασιών στο χώρο της ιατρικής. Συγκεκριμένα η ιατρική θα είναι σε θέση να:

- Αποτυπώνει με ταχύτητα και ακρίβεια τις διεργασίες που πραγματοποιούνται στον οργανισμό, εφόσον υπάρχει η δυνατότητα άμεσης ανάλυσης του ανθρώπινου γονιδιώματος. Επομένως, επιταχύνεται η διάγνωση της νόσου.
- Προκρίνει τις βέλτιστες θεραπευτικές μεθόδους μιας νόσου, προσαρμοσμένες στον εκάστοτε ασθενή.
- Προσδιορίζει τους παράγοντες υψηλού κινδύνου σε χειρουγικές επεμβάσεις
- Κατανείμει τους προς διάθεση χρηματικούς πόρους όπου κρίνεται απαραίτητο ανάλογα τις περιστάσεις.

Τηλεπικοινωνίες

Οι εταιρείες τηλεπικοινωνίας είναι από τις πρώτες που ενσωμάτωσε τις τεχνικές εξόρυξης δεδομένων στα επιχειρησιακά πλάνα τους. Καθημερινά καταγράφουν τεράστιους όγκους δεδομένων που περιλαμβάνουν τόσο στοιχεία πελατών (όνομα, διεύθυνση, τύπος συνδρομής κτλ), όσο και στοιχεία κλήσεων (ονόματα καλούντων, διάρκεια κλήσης, δίκτυο κλήσεων, κόστος). Τα δεδομένα αυτά οι εταιρείες τα χρησιμοποιούν για σκοπούς διαφήμισης, διασταυρούμενων πωλήσεων και αποτροπής-πρόβλεψης απάτης. Η απάτη ήταν και παραμένει μείζον ζήτημα για τις εταιρείες τηλεπικοινωνίας, καθώς συμβάλλει διττά στη μείωση της κερδοφορίας. Υπάρχουν δύο είδη απάτης, η απάτη συνδρομής και η απάτη υπέρθεσης (Weiss

[9]. Η απάτη συνδρομής πρόκειται για περιπτώσεις πελατών που δεν σκοπεύουν να εξοφλήσουν τους λογαριασμούς τους και περιλαμβάνει ιδώτες αλλά και επιχειρήσεις. Έχει παρατηρηθεί έντονα το φαινόμενο διαφημιστικές εταιρείες να αναλαμβάνουν έργα τηλεφωνικής προώθησης προιόντων και οι αριθμοί που καλούν να μην ανταποκρίνονται σε πραγματικούς πελάτες, αλλά εικονικούς. Κατά αυτόν τον τρόπο, πετυχάινουν γρήγορα και χώρις μεγάλα έξοδα τους στόχους της συμφωνίας, εις βάρος πάντοτε της εταιρείας που τους εμπιστεύτηκε την προώθηση των προιόντων της. Η απάτη υπέρθεσης, από την άλλη, συμβαίνει στις περιπτώσεις που δράστες αποκτούν πρόσβαση στο λογαριασμό ενός χρήστη-νόμιμου πελάτη και κάνει χρήση των υπηρεσιών του. Τέτοιου είδους απάτες δεν έχουν άμεσο οικονομικό αντίκτυπο, ωστόσο προκαλούν τη δυσαρέσκεια των πελατών με συνέπεια την πιθανή διακοπή του συμβολαίου τους.

Λογιστική-Ελεγκτική

Οι τομείς της λογιστικής και κατ'επέκταση της ελεγκτικής δεν θα μπορούσε να μην αποτελεί προνομιακό πεδίο εφαρμογής των μεθοδολογιών εξόρυξης δεδομένων. Αντικείμενο της ελεγκτικής είναι ο έλεγχος των επιχειρησιακών-λογιστικών διεργασιών των εταιρειών, με σκοπό τη διασφάλιση της ορθότητας των οικονομικών δεδομένων τους. Πρόκειται για μια διαδικασία ιδιαιτέρως δύσκολη με υψηλό βαθμό αβεβαιότητας, από τη στιγμή που ο ελεγκτής-αναλυτής καλείται να εξάγει συμπεράσματα σχετικά με την ακεραιότητα των οικονομικών στοιχείων μιας εταιρείας, έχοντας στη διάθεση του μια σειρά από αδόμητα δεδομένα που πολλές φορές έχουν υποστεί και αλλοιώσεις. Οι τεχνικές κατηγοριοποίησης, που χρησιμοποιούνται κατά κόρον στον τομέα της ελεγκτικής [10], δίνουν λύση σε δύο κύρια προβλήματα, αυτό της πρόβλεψης της χρεοκοπίας και της παραποίησης των χρηματοοικονομικών καταστάσεων. Οι χρεοκοπίες επιχειρήσεων αποβαίνουν επιζήμιες για ένα μεγάλο μέρος μετόχων, πιστωτών, επενδυτών ακόμα και κρατών. Οι εξωτερικοί ελεγκτές καλούνται, λοιπόν, να καταγράψουν και να αποτυπώσουν με πλήρη ακρίβεια την οικονομική κατάσταση των εταιρειών για τη διασφάλιση της οικονομικής ευημερίας του επενδυτικού χώρου. Σε όλες αυτές τις δυσκολίες που έρχονται αντιμέτωπες καθημερινά οι ελεγκτικές αρχές, έρχεται να προστεθεί η παραποίηση των χρηματοοικονομικών καταστάσεων από τα διοικητικά στελέχη των επιχειρήσεων που συχνά εμπλέκονται σε οικονομικές απάτες και σκάνδαλα.

Τράπεζες

Τα τελευταία χρόνια, ο χρηματοοικονομικός κλάδος και ιδαίτερα οι τράπεζες εφαρμόζουν ευρέως τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης. Ο μεγάλος όγκος δεδομένων που συλλέγουν καθημερινά από τις συναλλαγές των πελατών, τους επιτρέπει τη διεξαγωγή αναλύσεων που βοηθούν στην προώθηση υπηρεσιών (δάνεια, πιστωτικές κάρτες, επενδυτικά προγράμματα), στη διαχείριση του πιστοληπτικού ρίσκου, στην πρόληψη και πρόβλεψη της απάτης με πιστωτικές κάρτες και τέλος στην πάταξη του ξεπλύματος χρήματος.

2.2 Εξόρυξη Δεδομένων στην Τραπεζική

Η εξόρυξη δεδομένων, όπως αναφέρθηκε παραπάνω, έχει βρεί εφαρμογή σε αρκετές διεργασίες των τραπεζικών ιδρυμάτων την τελευταία δεκαετία. Οι βάσεις δεδομένων των τραπεζών συλλέγουν καθημέρινα terabytes δεδομένων, η ανάλυση των οποίων βοηθά στην παραγωγή πολύτιμης πληροφορίας. Οι τεχνικές εξόρυξης δεδομένων σε συνδυασμό με μαθηματικές και στατιστικές αναλύσεις οδηγούν στη δημιουργία προτύπων που περιγράφουν τόσο την συμπεριφορά των πελατών, όσο και τους χρηματοπιστωτικούς κινδύνους που υφέρπουν.

Πωλήσεις-Διαφήμιση

Οι πωλήσεις και η διαφήμιση είναι εργασίες που λαμβάνουν χώρα στο τραπεζικό σύστημα όπως ακριβώς συμβαίνει και σε κάθε άλλη επιχείρηση. Οι τράπεζες έχουν κατανοήσει εις βάθος την αξία των πελατών, ιδιαίτερα από τη στιγμή που ο τραπεζικός κλάδος είναι ένας κλειστός κλάδος, με την έννοια ότι το μέγεθος του πελατολογίου είναι καθορισμένο και δεν υφίσταται σημαντικές μεταβολές με την πάροδο των ετών. Η διαχείριση των σχέσεων των πελατών, γνωστή και ως CRM, δίνει τη δυνατότητα στους τραπεζικούς αναλυτές να αντιλαμβάνονται γρήγορα τις ανάγκες των πελατών τους, βάζοντας έτσι τα θεμέλια για μακροχρόνιες συνεργασίες. Επίσης, συμβάλλει στη τμηματοποίηση της αγοράς, με σκοπό την στοχευμένη διαφήμιση προιόντων και υπηρεσιών σε νέους υποψήφιους πελάτες. Για παράδειγμα ένας νέος ηλικιακά πελάτης της τράπεζας ενδιαφέρεται περισσότερο για τις ψηφιακές υπηρεσίες που παρέχει η τράπεζα (e-banking, εικονικές προπληρωμένες κάρτες κα) σε αντίθεση με τους μεγαλύτερους ηλικιάκα που μπορεί να ενδιαφέρονται για επενδυτικά - αποταμιευτικά προγράμματα. Τέλος, οι τεχνικές εξόρυξης δεδομένων μπορούν με σχετικά μεγάλη ακρίβεια να προβλέπουν το ρυθμό μετακίνησης των πελατών ανά χρηματοπιστωτικό ίδρυμα και να

επικεντρώνονται στη διατήρηση των «επίφοβων» προς αποχώρηση πελατών με προσωποποιημένα προνομιακά προγράμματα.

Διαχείριση ρίσκου

Ο τραπεζικός κλάδος είναι ένας από τους κατεξοχήν κλάδους που εμπεριέχουν το στοιχείο του κινδύνου, συνεπώς η διαχείριση ρίσκου βρίσκεται στην κορυφή των τραπεζικών διεργασιών. Τα τραπεζικά ιδρύματα καλούνται να αντιμετωπίσουν μια σειρά από κινδύνους, οι οποίοι έχουν να κάνουν με τον πιστωτικό κίνδυνο, τον κίνδυνο ελλιπούς ρευστότητας, τον συστηματικό και λειτουργικό κίνδυνο. Με τον όρο πιστωτικό κίνδυνο (credit risk) περιγράφεται η αδυναμία τους δανειολήπτη (φυσικό πρόσωπο, επιχείρηση, κράτος) να αποπληρώσει τις συμβατικές υποχρεώσεις του βάσει των προσυμφωνημένων όρων. Στόχος των τραπεζών είναι η χορήγηση όλο και περισσότερων δανείων, με τη βασική προυπόθεση ωστόσο αυτά να μην είναι επισφαλή. Γίνεται κατανοητό, λοιπόν, ότι οι τεχνικές πρόβλεψης πιστοληπτικού κινδύνου έχουν κομβική σημασία στον χρηματοοικονομικό κλάδο.

Ο κίνδυνος ελλιπούς ρευστότητας είναι ένα ζήτημα που απασχολεί πάντοτε τα τραπεζικά ιδρύματα. Κάθε τράπεζα θα πρέπει να είναι σε θέση να υποστηρίξει αναλήψεις μεγάλων ποσών σε περίπτωση ενός απροσδόκητου γεγονότος, το λεγόμενο bank run, έτσι ώστε να διασφαλίζει την αξιοπιστία της προς τους πελάτες. Χαρακτηριστικό παράδειγμα, ο σχηματισμός ουρών στα μηχανήματα αυτόματης ανάληψης μετρητών στην Ελλάδα μετά την ανακοίνωση του δημοψηφισματος το 2015, οπότε και οι τράπεζες αναγκάστηκαν να θέσουν πλαφόν στο όριο αναλήψεων για να αποφευχθεί η κατάρρευση του εγχώριου τραπεζικού συστήματος.

Ο συστηματικός κίνδυνος αφορά τις ζημίες που μπορεί να υποστεί ένα χρηματοπιστωτικό ίδρυμα εξαιτίας παραγόντων της αγοράς, όπως η ύφεση, οι φυσικές καταστροφές και οι κοινωνικοπολιτικές αναταραχές. Ο λειτουργικός κίνδυνος, από την άλλη πλευρά, προκύπτει από πιθανές αδυναμίες στις εσωτερικές διεργασίες ενός ιδρύματος

Ξέπλυμα Χρήματος

Τα τραπεζικά ιδρύματα, στα πλαίσια της κανονιστικής συμμόρφωσης, οφείλουν σε συνεργασία με την οικονομική αστυνομία να ελέγχουν συναλλαγές πελατών τους που αποτελούν πιθανές περιπτώσεις ξεπλύματος χρήματος. Με τον όρο ξέπλυμα χρήματος περιγράφεται η διαδικασία νομιμοποίησης παράνομων εσόδων και είναι άρρηκτα συνυφασμένη με το εμπόριο όπλων και ναρκωτικών. Από μελέτες που έχουν πραγματοποιηθεί [11], διαπιστώθηκε ότι οι περισσότερες περιπτώσεις νομιμοποίησης παράνομων εσόδων αφορούσαν πολλά μικρά και διάσπαρτα χρηματικά ποσά και όχι μεγάλα όπως θα ανέμενε κανείς. Οι τεχνικές εξόρυξης δεδομένων που

χρησιμοποιούνται για τον εντοπισμό φαινομένων ξεπλύματος χρήματος βασίζονται στην αναγνώριστη προτύπων και στον εντοπισμό αποκλινουσών συμπεριφορών. Επίσης χρησιμοποιούνται μέθοδοι μέθοδοι ανάλυσης συστάδων, κατηγοριοποίησης και δικτύων [12].

Απάτη Πιστωτικών Καρτών

Στην εποχή του ψηφιακού μετασχηματισμού που διανύουμε, η χρήση του λεγόμενου «πλαστικού» χρήματος, δηλαδή των πιστωτικών-χρεωστικών καρτών, τείνει να αποτελέσει τη βασική μέθοδο συναλλαγματικών διεργασιών. Παρά την ευκολία που παρουσιάζουν οι πιστωτικές κάρτες στις συναλλαγές (ηλεκτρονικές αγορές, ανέπαφες συναλλαγές), δεν είναι λίγα τα φαινόμενα εξαπάτησης. Τέτοιου είδους πρακτικές εξαπάτησης στηρίζονται είτε στη φυσική απώλεια-κλοπή μιας κάρτας από τον κάτοχό της, είτε στην υπεξαίρεση στοιχείων της. Για τον περιορισμό του φαινομένου, οι τράπεζες έχουν διαμορφώσει ολόκληρα τμήματα που απασχολούνται με την πρόβλεψη και τον έγκαιρο εντοπισμό περιπτώσεων απάτης. Η ανάλυση προτύπων, συστάδων και εξαιρέσεων χρησιμοποιούνται ευρέως για να εντοπιστούν ενδεχόμενες αποκλίσεις από συνήθεις πρακτικές και να διερευνηθούν πιθανές περιπτώσεις απάτης.

Εικόνα 8 - Τομείς Εφαρμογής Εξόρυξης Δεδομένων στην Τραπεζική

Πηγή: Introduction to Banking Technology and ManagementVadlamani RaviInstitute for Development and Research in Banking Technology, India [13]

2.3 «Έξυπνες» Εφαρμογές Εξόρυξης Δεδομένων στο Τραπεζικό Σύστημα

Τα μεγάλα δεδομένα σε συνδυασμό με τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης αποτελούν τη ραχοκοκαλιά του ψηφιακού μετασχηματισμού των τραπεζών. Η συνέργεια των παραπάνω τομέων έχει οδηγήσει σε μια σειρά βελτιωμένων παρεχόμενων πελατοκεντρικών υπηρεσιών και στην ορθότερη λήψη στρατηγικών αποφάσεων, γρήγορα και αποτελεσματικά. Παρακάτω αναφέρονται ενδεικτικά ορισμένες από τις «έξυπνες» υπηρεσίες-εφαρμογές που έχουν αναπύξει τα χρηματοπιστωτικά ιδρύματα, βασισμένα στα μεγάλα δεδομένα και την εξόρυξη γνώσης από αυτά.

- ***Έξυπνος λογαριασμός:*** Η ευρεία χρήση του e-banking επιτρέπει στους πελάτες να έχουν άμεση εικόνα του λογαριασμού τους ανά πάσα στιγμή, να ελέγχουν τις κινήσεις τους, να προγραμματίζουν μελλοντικές συναλλαγές, να ορίζουν το όριο των δαπανών τους και να κατηγοριοποιούν τα έξοδά τους. Επίσης, παρέχεται η δυνατότητα να λαμβάνουν προβλέψεις των μελλοντικών τους εξόδων, έτσι ώστε να διαμορφώνουν τον προυπολογισμό τους αντίστοιχα.
- ***Προσωποποιημένες υπηρεσίες στους πελάτες:*** Η περιήγηση των πελατών στις ψηφιακές υπηρεσίες των τραπεζών (internet banking) «αφήνει» πίσω της μια σειρά από δεδομένα που καταγράφονται στις βάσεις των τραπεζών. Με την ανάλυση των δεδομένων αυτών, οι τράπεζες μπορούν να κατανοήσουν και να διαμορφώσουν το προφίλ του πελάτη και να στοχεύσουν ατομικά στις ανάγκες του. Για παράδειγμα ένας χρήστης που φαίνεται να αφιερώνει αρκετό χρόνο στο μενού των πιστωτικών καρτών, είναι πολύ πιο πιθανό να δεχτεί την έκδοση μιας πιστωτικής κάρτας από ένα συμβατικό πελάτη. Επίσης, έχει παρατηρηθεί ότι γεγονότα στην προσωπική ζωή των πελατών επηρεάζουν και την οικονομική τους συμπεριφορά. Για παράδειγμα όταν ένας πελάτης μιας τράπεζας προσθέτει ένα νέο μέλος στην προσώπικη του ασφάλιση (γέννηση ενός παιδιού), τότε η τράπεζα μπορεί να εκμεταλλευτεί αυτό το στοιχείο και να του προσφέρει έναν προνομιακό αποταμιευτικό λογαριασμό παίδων. Ωστόσο, οι τράπεζες, λόγω του αυξημένου ανταγωνισμού, δεν έχουν την πολυτέλεια να περιμένουν τους πελάτες να ενημερώσουν τα προσωπικά τους στοιχεία στους καταλόγους της τράπεζας. Τα κοινωνικά μέσα και η συμπεριφορά στο διαδίκτυο έρχονται να δώσουν λύση στο ζήτημα αυτό. Πρόσφατα, καθηγητές του πανεπιστημίου του Stanford δημοσιοποίησαν ένα προγνωστικό μοντέλο των επερχόμενων γεγονότων στην προσωπική ζωή των ατόμων. Με χρήση μεθόδων ανάλυσης κειμένου σε αναρτήσεις σε κοινωνικά μέσα (Twitter, Facebook) μπορούν πια να προβλεφθούν γεγονότα όπως γάμοι, αποφοιτήσεις,

γεννήσεις παιδιών κ.α. Έχοντας γνώση αυτών των στοιχείων οι τράπεζες μπορούν να στοχεύουν συγκεκριμένα πελάτες και να τους κάνουν προσωποποιημένες προσφορές με βάση τις ανάγκες τους [14].

- ***Προσέλκυση νέων πελατών:*** Η ανάλυση των μέσεων κοινωνικής δικτύωσης μπορεί να εντοπίσει, επίσης, πελάτες που είναι δυσαρεστημένοι με τις υπηρεσίες άλλων τραπεζών και συνεπώς να τους πλησιάσει με δελεαστικότερα πακέτα. Για παράδειγμα αν ένας χρήστης που ακολουθούσε ένα τραπεζικό ίδρυμα στα κοινωνικά μέσα, έπαψε να το ακολουθεί, έχει μεγάλες πιθανότητες να αποτελέσει μελλοντικό πελάτη άλλης τράπεζας. Έτσι με τεχνικές στοχευμένης διαφήμισης, μπορούν να εμφανιστούν κατά την περιήγηση του στο διαδίκτυο ενημερωτικές καμπάνιες, σχετικές με προιόντα και υπηρεσίες της έτερης τράπεζας. Τέτοιες περιπτώσεις εντοπίζονται άμεσα με τεχνικές εξόρυξης δεδομένων.
- ***Βέλτιστη ανάπτυξη δικτύου καταστημάτων και ΑΤΜ:*** Η ίδρυση μιας νέα μονάδας τραπεζικού υποκαταστήματος αποτελεί σημαντικό γεγονός στον προγραμματισμό ενός τραπεζικού ομίλου. Για τη λήψη τέτοιων αποφάσεων λαμβάνονται υπόψη μια σειρά από παράγοντες. Αρχικά, ελέγχονται τα οικονομικά στοιχεία των κατοίκων της περιοχής (ετήσιο εισόδημα, καταθέσεις), καθώς και τα δημογραφικά (πληθυσμός, ηλικία). Επίσης, συνυπολογίζονται παράγοντες όπως η παρουσία ανταγωνισμού στην περιοχή (υποκαταστήματα άλλων τραπεζών, ΑΤΜ) και η φυσική θέση των περιοχών. Σημεία αυξημένου ενδιαφέροντος είναι περιοχές ευπρόσιτες στο ευρύ κοινό, κοντά σε συγκοινωνιακούς κόμβους και εμπορικά κέντρα. Ο έλεγχος των παραπάνω παραγόντων γίνεται με χρήση τεχνικών εξόρυξης δεδομένων, με σκοπό την εύρεση της βέλτιστης θέσης των καταστημάτων για την αύξηση της κερδοφορίας. Αντίστοιχα, με την ανάλυση των στοιχείων από τα Αυτόματα Ταμειακά Μηχανήματα (ΑΤΜ), η τράπεζα μπορεί να βγάλει χρήσιμα συμπεράσματα για τις περιοχές που χρησιμοποιούν περισσότερο οι πελάτες της, τις ενέργειες που προβαίνουν, και τις περιοχές που επιλέγουν συχνότερα ΑΤΜ άλλων τραπεζών [15].
- ***Προγραμματισμός συχνότητας και ποσότητας τροφοδοσίας ΑΤΜ***: Με τεχνικές εξόρυξης δεδομένων και μαθηματικές αναλύσεις οι υπεύθυνοι ενός τραπεζικού ιδρύματος μπορούν να προγραμματίσουν τη ροή τροφοδοσίας των ΑΤΜ με μετρητά για τη βέλτιστη εξυπηρέτηση των πελατών. Για παράδειγμα η συχνότητα τροφοδοσίας ενός μηχανήματος αυτόματης ανάληψης μετρητών σε μια ορεινή κωμόπολη θα διαφέρει σημαντικα με αυτόν ενός αντίστοιχο σε ένα πολυσύχναστο δρόμο ενός μεγάλου αστικού κέντρου. Επίσης, λαμβάνονται υπόψη τα γεγονότα που πρόκεινται να

ακολουθήσουν, όπως εκδηλώσεις, καιρικά φαινόμενα, έτσι ώστε να αποφεύγεται η έλλειψη μετρητών από ορισμένα μηχανήματα και συνεπώς η δυσαρέσκεια της πελατείας.

- ***Βελτιστοποίηση μέσων επικοινωνίας με τους πελάτες***: Οι τράπεζες καλούνται να βρούν εναλλακτικούς τρόπους επικοινωνίας με τους πελάτες που θα είναι πιο άμεσοι και εξυπηρετικοί. Η εγκατάλειψη της έντυπης αλληλογραφίας είναι μονόδρομος τόσο για περιβαλλοντικούς λόγους, καθώς εξοικονομούνται τόνοι χαρτιού ετησίως από την παύση των έντυπων ενημερώσεων, όσο και για λόγους αμεσότητας. Ενδεικτικά ένα συχνό φαινόμενο που προκαλεί δυσαρέσκεια στις επιχειρήσεις είναι η ανανέωση των νομιμοποιητικών εγγράφων τους, διαδικασία που απαιτεί χρόνο, ειδικά αν η επιχείρηση έχει αμελήσει να προβεί έγκαιρα σε αυτή. Τώρα πια, η όλη διαδικασία μπορεί να γίνει ηλεκτρονικά, μέσω εναλλακτικών καναλιών (e-banking), ενώ η αποστολή των απαραίτητων εγγράφων μπορεί να πραγματωθεί και μέσω ηλεκτρονικού ταχυδρομείου (e-mail), αποφεύγοντας έτσι την καθυστέρηση της αποστολής τους μέσω ταχυδρομείου ή ταχυμεταφορών.
- ***Βελτιστοποίηση και απλοποίηση συχνά χρησιμοποιούμενων λειτουργιών***: Η ανάλυση των κινήσεων των πελατών στα κανάλια εξυπηρέτησης των τραπεζών (e-banking, ATM, apps) μπορεί να προσδιορίσει πρότυπα συμπεριφορών. Για παράδειγμα, αν εντοπίζεται ότι οι χρήστες της εφαρμογής μιας τράπεζας δαπανούν αρκετό χρόνο στο κομμάτι των πληρωμών, αυτό πιθανώς να συνεπάγεται δυσκολία στην κατανόηση και χρήση του συγκεκριμένου πεδίου, οπότε και χρήζει βελτιώσεων από το αρμόδιο τμήμα πληροφορικής. Αντίστοιχα, μια τακτική που υιοθετήθηκε τους τελευταίους μήνες από το σύνολο των συστημικών τραπεζών της Ελλάδας είναι η αποστολή των μηνυμάτων επιβεβαίωσης συναλλαγών (i-code) σε πλατφόρμες επικοινωνίας (πχ Viber), που δεν απαιτούν την παρουσία τηλεπικοινωνιακού σήματος. Τα ταξίδια στο εξωτερικό για επαγγελματικούς λόγους ή λόγους αναψυχής αποτελούν συχνό φαινόμενο στην καθημερινότητα των πολιτών. Παρατηρήθηκε, λοιπόν, το φαινόμενο αδυναμίας χρήσης των ηλεκτρονικών υπηρεσιών των τραπεζών, λόγω έλλειψης σήματος από τους παρόχους τηλεπικοινωνίας. Για την επίλυση, λοιπόν, τέτοιου είδους ζητημάτων οι διευθύνσεις των τραπεζών σκέφτηκαν ότι η αποστολή των μηνυμάτων επιβεβαίωσης (SMS i-code) μπορεί να πραγματοποιείται μέσω εφαρμογών επικοινωνίας που χρησιμοποιούνται ευρέως από τους πελάτες και απαιτούν μόνο την σύνδεση του στο διαδίκτυο (Wi-Fi), λαμβάνονοντας πάντοτε υπόψη τους κανονισμούς ασφαλείας για την προστασία των ευαίσθητων δεδομένων.

2.4 Πλεονεκτήματα και Μειονεκτήματα Βαθειάς Μάθησης στις Τράπεζες

Η χρήση τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης στον τραπεζικό κλάδο χαρακτηρίζεται εν γένει από μια σειρά θετικών παρεμβάσεων. Εγείρονται, ωστόσο, ορισμένες ανησυχίες που οφείλονται κυρίως στην ορθή χρήση τέτοιων τεχνικών και στα μελλοντικά προβλήματα που μπορούν να ανακύψουν.

2.4.1 Πλεονεκτήματα

Βελτιωμένη εξυπηρέτηση πελατών

Η χρήση των μεθόδων εξόρυξης δεδομένων και μηχανικής μάθησης έχει επιφέρει μια σειρά από θετικές αλλαγές στην αλληλεπίδραση των πελατών με τις τράπεζες. Πλέον, ο κάθε πελάτης δύναται να αντιμετωπίζεται εξατομικευμένα, σύμφωνα με τις ανάγκες του. Επίσης, οι εναλλακτικοί τρόποι εξυπηρέτησης μέσω εικονικών βοηθών (chatbots) λειτουργούν συμβουλευτικά και καθοδηγητικά, προσφέροντας έτσι μια καινοτόμα εμπειρία χρήσης στους πελάτες [16]. Τεχνικές που εκμηδενίζουν το χρόνο αναμονής και είναι διαθέσιμες καθόλη τη διάρκεια του εικοσιτετραώρου συμβάλλουν στην αύξηση του αισθήματος ικανοποίησης του πελάτη και στην προσέλκυση νέων.

Αυξημένη ασφάλεια συναλλαγών

Λόγω της ευρείας χρήσης ηλεκτρονικών υπηρεσιών και καρτών, έχει παρατηρηθεί ραγδαία αύξηση φαινομένων τραπεζικής απάτης. Η αναγνώριση και η μείωση τέτοιων φαινομένων αποτελεί πρόκληση για τον τραπεζικό τομέα. Η τεχνητή νοημοσύνη διευκολύνει τον εντοπισμό των παραγόντων που εμπλέκονται σε απάτες και την υποστήριξη των ερευνητών. Βελτιώνει, έτσι, την οικονομική ασφάλεια με προηγμένες τακτικές πρόληψης της απάτης. Η τεχνητή νοημοσύνη προσφέρει λύσεις σε πραγματικό χρόνο για τον τραπεζικό τομέα, ενώ χειρίζεται πολύπλοκες καταστάσεις, εντοπίζοντας και επισημαίνοντας ασυνήθιστες συναλλαγές. Τροφοδοτεί επίσης το προφίλ του καταναλωτή με δεδομένα με σκοπό τη διαμόρφωση ενός ασφαλούς περιβάλλοντος τόσο για τον ίδιο, όσο και για την τράπεζα.

Συμμόρφωση και ενσωμάτωση κανονιστικών αλλαγών/εποπτικών πλαισίων

Η ταχύτητα και η ευελιξία που παρέχουν οι τεχνικές βαθειάς μάθησης στον τραπεζικό τομέα αποτελούν σημαντικά όπλα στον αγώνα των χρηματοπιστωτικών ιδρυμάτων προς συμμόρφωση με τα νέα κανονιστικά πλαίσια που επιβάλλονται. Οι αλλαγές στους κανόνες εποπτείας σε θέματα επάρκειας κεφαλαίων και ρευστότητας αποτελούν βασικό μέρος της ατζέντας των κανονιστικών αλλαγών, ως επακόλουθο της τελευταίας διεθνούς

χρηματοπιστωτικής κρίσης. Πολλά από τα στοιχεία ισολογισμού που σταθμίζονται με τους νέους κανόνες κεφαλαιακής επάρκειας δημιουργούνται μέσα από τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης. Για παράδειγμα, η δομή χρηματοδότησης μιας τράπεζας, η έμφαση σε συγκεκριμένα δανειακά προϊόντα έναντι άλλων, ή τυχόν προσαρμογές στα χαρακτηριστικά κινδύνου των στεγαστικών δανείων που εκταμιεύονται, στηρίζονται σε τέτοιες μεθόδους.

2.4.2 Μειονεκτήματα

Κόστος

Η παραγωγή και η συντήρηση της τεχνητής νοημοσύνης απαιτούν τεράστιο κόστος, δεδομένου ότι αποτελείται από εξαιρετικής πολύπλοκότητας μηχανήματα. Επίσης, τα προηγμένα προγράμματα λογισμικού απαιτούν τακτικές ενημερώσεις για την κάλυψη των αναγκών ενός συνεχώς μεταβαλλόμενου περιβάλλοντος, όπως αυτό του τραπεζικού τομέα. Σε περίπτωση κατάρρευσης των συστημάτων, η διαδικασία επαναφοράς τους και ανάκτησης των χαμένων κωδικών μπορεί να απαιτήσει τεράστιο χρόνο και κόστος, από πλευράς απασχόλησης ανθρωπίνου δυναμικού. Τέλος, θα πρέπει να συνυπολογιστούν τα κόστη που απαιτούνται για την εξασφάλιση του απαιτούμενου αποθηκευτικού χώρου.

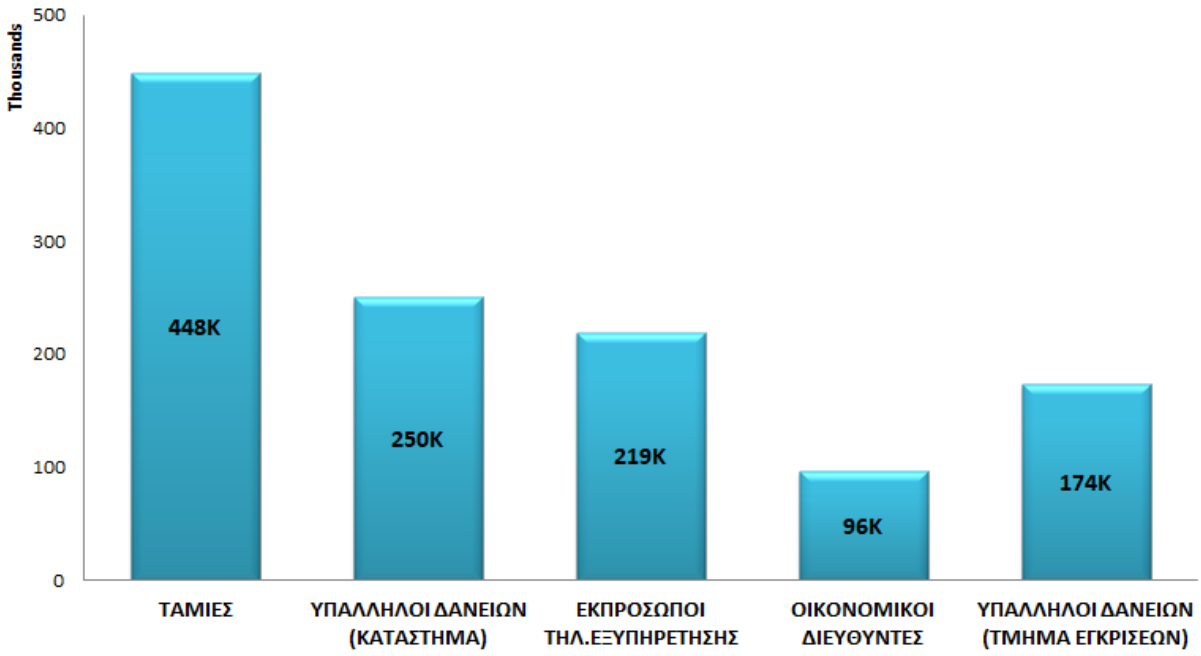
Ηθικά διλήμματα

Όπως έχει αναφερθεί και παραπάνω, η ραγδαία τεχνολογική πρόοδος και η ευρεία χρήση του διαδικτύου έχει ως αποτέλεσμα να συγκεντρώνονται καθημερινά εκατοντάδες GB δεδομένων. Ωστόσο, η συγκέντρωση και η ανάλυση των δεδομένων αυτών θα πρέπει να διέπεται από αρχές και αξίες, χωρίς να παραβιάζονται προσωπικά ευαίσθητα δεδομένα. Επίσης, οι τράπεζες οφείλουν να διαμορφώνουν αυστηρά πλαίσια διαχείρισης των δεδομένων και να ελέγχουν τα εμπλεκόμενα μέλη που έχουν πρόσβαση σε αυτά.

Μείωση εργατικού δυναμικού

Δεν είναι λίγοι εκείνοι που αντιμετωπίζουν με σκεπτικισμό την είσοδο της τεχνητής νοημοσύνης στο τραπεζικό κλάδο. Είναι γεγονός, ότι με σε λίγα χρόνια αρκετές από τις αυτοματοποιημένες διεργασίες που εκτελούν μέχρι σήμερα υπάλληλοι, θα αντικατασταθούν με παράγωγα της μηχανικής μάθησης (υπάλληλοι εξυπηρέτηση πελατών, ταμίες κα). Ως εκ τούτου αναμένεται εντός των επόμενων δεκαετιών να καταργηθούν αρκετές από αυτές τις θέσεις. Ωστόσο, θα πρέπει να ληφθεί υπόψη ότι η υιοθέτηση τεχνικών μηχανικής μάθησης και εξόρυξης δεδομένων θα επιφέρει τη δημιουργία μιας σειράς νέων - και την αύξηση ορισμένων

υπαρχόντων (αναλυτές δεδομένων, διαχειριστές βάσεων δεδομένων)- ειδικοτήτων στο χώρο της τραπεζικής.



Εικόνα 9 - Θέσεις εργασίας που αναμένεται να χαθούν εξαιτίας της τεχνητής νοημοσύνης μέχρι το 2030

Πηγή: https://www.americanbanker.com/news/how-artificial-intelligence-is-reshaping-jobs-in-banking

ΚΕΦΑΛΑΙΟ 3: Ανίχνευση Απάτης και Πιστωτικές Κάρτες

3.1 Μορφές Απάτης στον Τραπεζικό Τομέα

3.1.1 Εισαγωγή

Ένα από τα βασικά προβλήματα που αντιμετωπίζουν τα τραπεζικά ιδρύματα είναι τα φαινόμενα απάτης, τα οποία αποτελούν κίνδυνο, τόσο για την οικονομική ευημερία των ιδρυμάτων, όσο και για τη φήμη τους. Ως απάτη μπορεί να οριστεί η εκ προθέσεως πράξη από ένα ή περισσότερα πρόσωπα της οικονομικής μονάδας ή από τη διοίκηση ή από εκείνους που είναι επιφορτισμένοι με τη διακυβέρνηση, από εργαζόμενους ή από τρίτα μέρη, η οποία ενέχει παραπλάνηση για την απόκτηση ενός μη δίκαιου ή παράνομου πλεονεκτήματος [17].

Ιδιαίτερα, όταν γίνεται λόγος για τραπεζική απάτη, ο αντίκτυπος στη φήμη και την αξιοπιστία του ιδρύματος είναι μεγαλύτερος από την οικονομική απώλεια αυτή καθ' αυτή. Είναι γνωστό, ότι οι τράπεζες στηρίζουν μεγάλο μέρος της δράσης τους στην οικοδόμηση σχέσεων αξιοπιστίας και ασφάλειας με τους πελάτες τους. Όταν μια τράπεζα πέσει θύμα φαινομένων απάτης κλονίζεται σημαντικά το προφίλ της, από τη στιγμή που δημιουργούνται αισθήματα φόβου και αμφιβολίας στους πελάτες σχετικά με την ασφάλεια των συναλλαγών και των αποταμιεύσεών τους. Τέτοια φαινόμενα, οδηγούν στη φυγή σημαντικού αριθμού πελατών και στην δυσκολία προσέλκυσης νέων. Γίνεται αντιληπτό, λοιπόν, ότι η δίωξη και η παραδειγματική τιμωρία τέτοιων φαινομένων από τις τράπεζες, πηγάζει από την ανάγκη ενίσχυσης και διατήρησης της εικόνας τους στην αγορά, από ενδεχόμενες φήμες που θα μπορούσαν να επηρεάσουν την αξιοπιστία τους. Αρωγός στην ελαχιστοποίηση του φαινομένου της τραπεζικής απάτης είναι:

i. Τα Συστήματα Εσωτερικού Ελέγχου που εξασφαλίζουν την απρόσκοπτη λειτουργία των τραπεζικών διεργασιών.
ii. Οι λειτουργοί του Εσωτερικού Ελέγχου που εξετάζουν την άρτια λειτουργία των συστημάτων ελέγχου, αξιολογούν την αποτελεσματικότητα τους, διενεργούν ελέγχους σχετικά με την εφαρμογή των κανονισμών και συντάσσουν εκθέσεις αξιολόγησης που υποβάλλονται στη διοίκηση της τράπεζας.

3.1.2 Κατηγορίες και Μορφές Τραπεζικής Απάτης

Σύμφωνα με τα διεθνή κανονιστικά πλαίσια «Βασιλεία I, II, III», τα φαινόμενα τραπεζικής απάτης που υφίστανται χωρίζονται σε δύο κύριες κατηγορίες, τις απάτες στο εσωτερικό και

στο εξωτερικό περιβάλλον της τράπεζας [18]. Όπως γίνεται εύκολα αντιληπτό, στην πρώτη κατηγορία ανήκουν οι απάτες που τελώνται από το έμψυχο δυναμικό της τράπεζας ξεκινώντας από τους απλούς ταμίες και φτάνοντας ως τις διοικήσεις των τραπεζών. Από την άλλη, στις εξωτερικές απάτες ανήκουν τα αδικήματα που γίνονται από οντότητες εκτός του τραπεζικού περιβάλλοντος δηλαδή κοινούς επιτήδειους ακόμα και χάκερ. Παρακάτω γίνεται μια σύντομη αναφορά στις μορφές με τις οποίες συναντώνται τα φαινόμενα τραπεζικής απάτης, στο εσωτερικό και στο εξωτερικό αυτής.

- ***Δάνεια σε εικονικούς δανειολήπτες (εσωτερικό)***:

  Σε χρηματοπιστωτικά ιδρύματα με ανεπαρκή εσωτερικό έλεγχο, ο υπάλληλος μπορεί να υποβάλει μια φόρμα δανείου με πλαστά στοιχεία (διεύθυνση, ονοματεπώνυμο, ηλεκτρονικό ταχυδρομείο), να την προωθήσει στο τμήμα εγκρίσεων δανείων, όπου σε συνεργασία με τον αρμόδιο υπάλληλο, να εγκρίνουν την αίτηση και να προχωρήσουν στην εκταμίευση των χρημάτων του δανείου. Μάλιστα, σε ορισμένες περιπτώσεις έχει παρατηρηθεί το φαινόμενο οι υπάλληλοι (δράστες) να πληρώνουν κανονικά τις πρώτες δόσεις του δανείου, ώστε να μη κινηθούν άμεσα υποψίες. Όταν πλέον γίνεται αντιληπτό από τις υπηρεσίες της τράπεζας, τα ίχνη των δραστών έχουν ήδη χαθεί.

- ***Απάτη αδρανών λογαριασμών (εσωτερικό)***:

  Όταν σε έναν τραπεζικό λογαριασμό δεν πραγματοποιούνται δοσοληψίες για ένα μεγάλο χρονικό διάστημα, συνήθως ένα έτος ή και περισσότερο, τότε αυτός θεωρείται αδρανής. Τέτοιου είδους λογαρισμοί συχά γίνονται στόχος από ανέντιμους υπαλλήλους που προχωρούν σε αναλήψεις των ποσών που είναι αποταμιευμένα. Σε περιπτώσεις που ο κάτοχος δεν υφίσταται πλέον ή δεν ελέγχει το λογαριασμό του, τέτοιου είδους απάτη είναι αρκετά δύσκολο να εντοπιστεί.

- ***Παραποίηση μέσων πληρωμής (εσωτερικό)***:

  Συνήθως πρόκειται για υπαλλήλους που είναι αρμόδιοι για τις πληρωμές της τράπεζας και έχουν πρόσβαση στα αξιόγραφα του χρηματοπιστωτικού ιδρύματος. Κύριες απάτες στη κατηγορία αυτή είναι η έδκοση εικονικών τιμολογίων, η συνεργασία με εταιρείες «φαντάσματα» κ.α.

- ***Απάτη στη χρηματοοικονομική αναφορά (εσωτερικό)***:

  Πρόκειται ίσως για τη συχνότερη μορφή εσωτερικής απάτης εντός των τραπεζικών ιδρυμάτων διαχρονικά. Στελέχη, τις περισσότερες φορές υψηλόβαθμα, προβαίνουν σε

κινήσεις παραποίησης λογιστικών εκτιμήσεων, πλαστογραφίας εγγράφων και απόκρυψης οικονομικών στοχείων με στόχο το προσωπικό όφελος.

- ***Ηλεκτρονικό ψάρεμα (εξωτερικό)***:

Δεν είναι λίγες οι περιπτώσεις υποκλοπής δεδομένων μέσω ηλεκτρονικού ταχυδρομείου (phishing). Το υποψήφιο θύμα λαμβάνει μέσω ηλεκτρονικού ταχυδρομείου ένα μήνυμα, το οποίο εμφανίζει ως αποστολέα την τράπεζα και του ζητείται να πληκτρολογήσει προσωπικά του στοιχεία με σκοπό την επικαιροποίηση του λογαριασμού του.

- ***Διαδικτυακό έγκλημα και Παγίδευση ATM (εξωτερικό)***:

Συχνό φαινόμενο, επίσης, είναι η παράνομη δημιουργία ιστοσελίδων που με πρόσχημα την παροχή υπηρεσιών, υποκλέπτουν προσωπικά δεδομένα πελατών (κωδικούς internet banking, κωδικούς ασφάλισης, στοιχεία καρτών) και αποκτούν πρόσβαση στο λογαριασμό τους. Παράλληλα, έχουν παρατηρηθεί και φαινόμενα ηλεκτρονικής παγίδευσης μηχανημάτων αυτόματης ανάληψης μετρητών (ATM skimming).

- ***Υποκλοπή καρτών (εξωτερικό)*** :

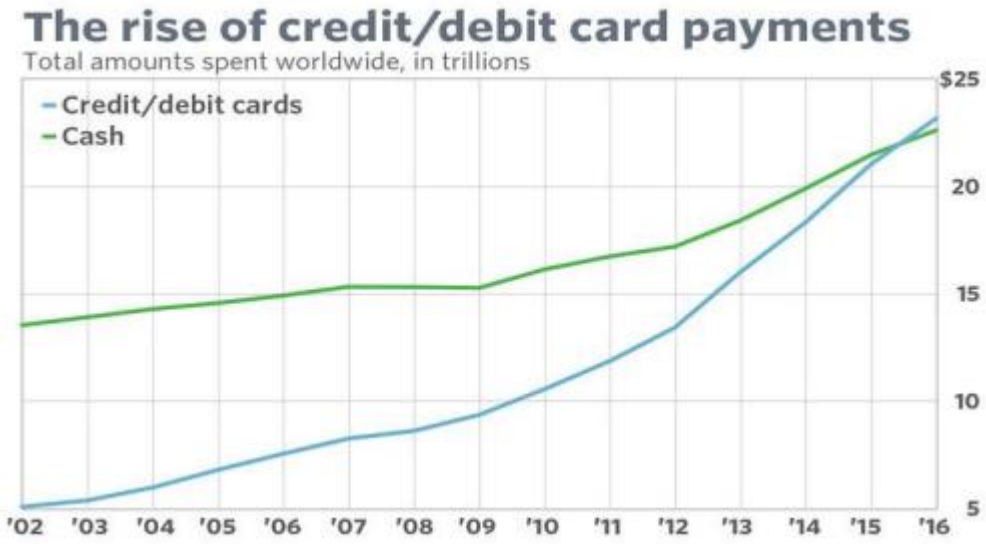
Αποτελεί τη συνηθέστερη περίπτωση απάτης στο τραπεζικό σύστημα. Επιτήδειοι αποσπούν τις κάρτες ανυποψίαστων πελατών και τις χρησιμοποιούν σε ανέπαφες συναλλαγές που δεν απαιτούν τη χρήση προσωπικού αριθμού αναγνώρισης (PIN). Όταν το αντιλαμβάνονται οι νόμιμοι κάτοχοι, πλέον είναι αργά, αφού έχουν αφαιρεθεί σημαντικά ποσά από το λογαριασμό τους. Για το λόγο αυτό, συστήνεται η άμεση δήλωση της απώλειας μιας κάρτας στο αρμόδιο τμήμα της τράπεζας.

3.2 Πλαστικό χρήμα και Πιστωτικές Κάρτες

3.2.1 Πλαστικό χρήμα: Ορισμός και ιστορική αναδρομή

Ο όρος πλαστικό χρήμα χρησιμοποιείται για να περιγράψει κατά κύριο λόγο τις συναλλαγές που εκτελούνται με πλαστικοποιημένες κάρτες πληρωμής εικονικών χρημάτων, αντικαθιστώντας τη χρήση μετρητών χαρτονομισμάτων. Το πλαστικό χρήμα δύναται να χρησιμοποιηθεί για μια σειρά διαφόρων συναλλαγών, από την πληρωμή λογαριασμών έως και την αγορά προιόντων. Τέτοιου είδους κάρτες (χρεωστικές, προπληρωμένες, πιστωτικές) εκδίδουν τα χρηματοπιστωτικά ιδρύματα σε πελάτες που διαθέτουν τραπεζικούς λογαριασμούς (αποταμιευτικούς, μισθοδοσίας, όψεως) για να πραγματοποιούν τις συναλλαγές τους με εναλλακτικό τρόπο, πέρα από τα μετρητά.

Σαν όρος εμφανίστηκε για πρώτη φορά στις αρχές του προηγούμενου αιώνα και πιο συγκεκριμένα το 1920, όταν και εισήχθη η πρώτη κάρτα πληρωμής στις ΗΠΑ. Λίγες δεκαετίες αργότερα, το 1950, η Diners Club και η American Express ξεκίνησαν τη χρήση των πρώτων πλαστικών καρτών. Στη σημερινή εποχή, η χρήση του πλαστικού χρήματος είναι ευρέως διαδεδομένη αφενός λόγω της ευκολίας στη χρήση που παρέχει, αφετέρου λόγω συμμόρφωσης με τις απαιτήσεις των φορολογικών συστημάτων. Ύστερα από την οικονομική κρίση που έπληξε την παγκόσμια κοινότητα την περίοδο 2008-2009, η χρήση της πλαστικής κάρτας ώς μέσο πληρωμής σημείωσε ραγδαία αύξηση. Ενδεικτικά, αξίζει να αναφερθεί ότι μόνο κατά το έτος 2011 ο όγκος των συναλλαγών με κάρτες αυξήθηκε κατά 13,5%, με τις κάρτες να αποτελούν την πλεόν προτιμώμενη μέθοδο πληρωμής πλήν των μετρητών.



Εικόνα 10 - Συγκριτικό Διάγραμμα Συναλλαγών με Χρήση Μετρητών και Πλαστικού Χρήματος

Πηγή: Euromonitor International

Από τη μελέτη του παραπάνω διαγράμματος, παρατηρείται ότι το έτος 2015 ήταν κομβικής σημασίας, όσον αφορά τις μεθόδους διενέργειας συναλλαγών παγκοσμίως. Ενώ η χρήση μετρητών τις τελευταίες δύο δεκαετίες παρουσιάζει μια σταθεροποιητική πορεία με ελαφρώς ανοδική τάση, κάτι τέτοιο δεν συμβαίνει και με τις συναλλαγές με χρήση πλαστικού χρήματος. Από τις αρχές του 2000 παρατηρείται μια σημαντική ετήσια αύξηση του ποσού των συναλλαγών που πραματοποιούνται με πιστωτικές/χρεωστικές κάρτες, η οποία δείχνει να κορυφώνεται μετά το 2012 και να αγγίζει το 2015, σε αξία, το ποσό των συναλλαγών που διενεργούνται παγκοσμίως με μετρητά.

3.2.2 Πιστωτικές Κάρτες: Ορισμός και ιστορική αναδρομή

Πριν αναλυθεί το κύριο αντικείμενο της παρούσας εργασίας - το οποίο είναι ο εντοπισμός και πρόβλεψη φαινομένων απάτης με χρήση πιστωτικών καρτών (credit card fraud detection) - παρατίθενται ορισμένα στοιχεία σχετικά με τις πιστωτικές κάρτες και τη σημασία τους στην τραπεζική.

Πιο αναλυτικά, η πιστωτική κάρτα αποτελεί μια υποτυπώδη μορφή δανείου από πλευράς της τράπεζας προς τον πελάτη. Όταν πραγματοποιείται η αγορά ενός καταναλωτικού αγαθού από έναν πελάτη, δεν είναι απαραίτητο η κάρτα να διαθέτει το επαρκές υπόλοιπο χρημάτων έτσι ώστε να ολοκληρωθεί η αγορά. Κάθε συναλλαγή με χρήση πιστωτικής κάρτας δημιουργεί μια οφειλή του συναλλασσομένου απέναντι στο χρηματοπιστωτικό ιδρυμά του, η οποία είναι ισάξια με την αξία του αγαθού προσαυξημένη με τους αντίστοιχους τόκους που έχουν συμφωνηθεί μεταξύ πελάτη-τράπεζας. Ο κάτοχος της πιστωτικής κάρτας αποπληρώνει την οφειλές του σε μηνιαίες δόσεις που έχουν προκαθοριστεί κατά τη σύναψη του συμβολαίου έκδοσης της κάρτας, έχοντας τις ανάλογες επιβαρύνσεις σε περίπτωση αθέτησης των όρων αποπληρωμής. Φυσικά, η δυνατότητα πίστωσης δύναται μέχρι ενός ορίου που προκαθορίζεται εξαρχής από την τράπεζα, πέραν του οποίου ουδεμία συναλλαγή μπορεί να εκτελεστεί. Αυτό είναι το γνωστό στην τραπεζική ορολογία πιστωτικό όριο.

Σύμφωνα με τον Bellis [19], η πρώτη πιστωτική κάρτα ήταν η κάρτα Diners Club και εκδόθηκε πρώτη φορά το 1950 στις Η.Π.Α από τον ιδρυτή του Diners Club, Frank McNamara . Η ιδέα της δημιουργίας μιας τέτοιας κάρτας συνελήφθη όταν ο McNamara κατά τη διάρκεια ενός επαγγελματικού δείπνου που είχε στη Νέα Υόρκη διαπίστωσε ότι είχε ξεχάσει το πορτοφόλι του. Έτσι, σκέφτηκε τη δημιουργία μιας κάρτας «προνομίων» για αξιόπιστα πρόσωπα που επισκέπτονται συχνά κέντρα ψυχαγωγίας για επαγγελματικούς λόγους. Οι κάτοχοι της κάρτας αυτής χρεώνονταν με μια ετήσια συνδρομή 3 δολαρίων και υποχρεούνταν να πληρώσουν άτοκα το ποσό που δαπάνησαν εντός 60 ημερών. Η Dinners Club πλήρωνε τα εστιατόρια που συμμετείχαν στο πρόγραμμα, παρακρατώντας προμήθεια 7% ανά συναλλαγή.

Η πρώτη πιστωτική κάρτα χρηματοπιστωτικού ιδρύματος εκδόθηκε από την Αμερικανική Τράπεζα το 1958 με την επονομασία BankAmericard, η οποία και μετονομάστηκε το 1976 στη γνωστή έως σήμερα Visa. Οι πιο γνωστές πιστωτικές κάρτες αυτή τη στιγμή είναι η MasterCard, η Visa International και η American Express.

3.2.3 Πλεονεκτήματα και Μειονεκτήματα Πιστωτικών Καρτών

Η χρήση των πιστωτικών καρτών είναι γεγονός ότι προσφέρει μια σειρά από σημαντικά πλεονεκτήματα για τους κατόχους αυτών. Παρακάτω αναφέρονται ενδεικτικά ορισμένα από αυτά:

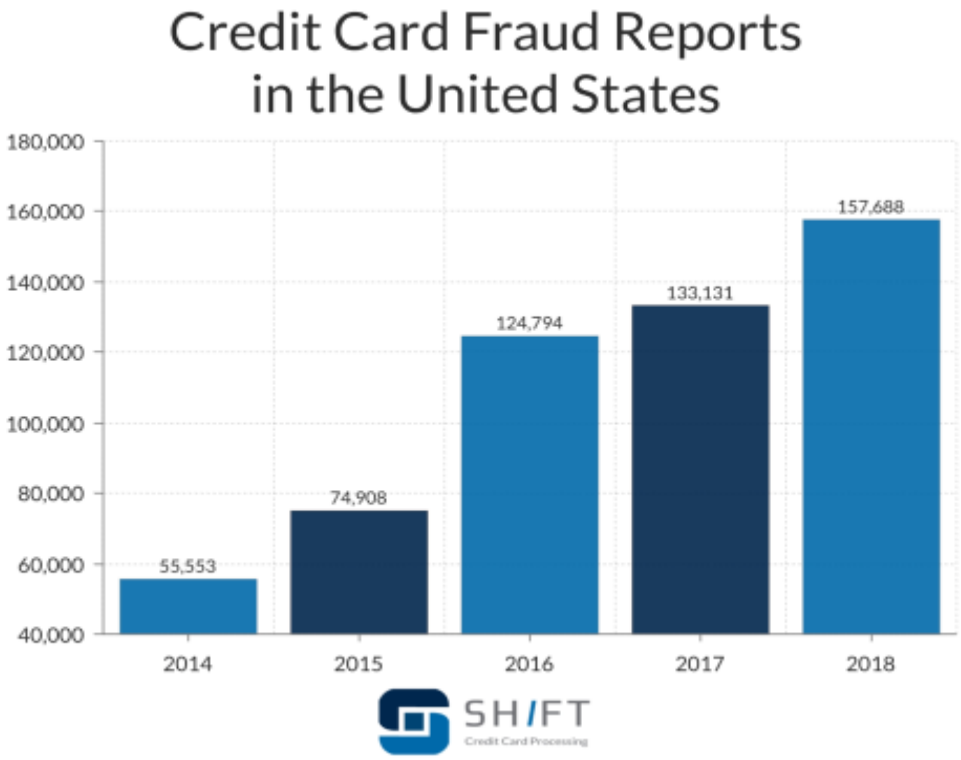
- ***Ασφάλεια συναλλαγών*:** Ο κάτοχος της κάρτας μπορεί να πραγματοποιεί συναλλαγές υψηλής αξίας χωρίς να χρειάζεται να μεταφέρει μαζί του μετρητά, περιορίζοντας έτσι τον κίνδυνο κλοπής ή απώλειας τους.
- ***Ευκολία και ταχύτητα συναλλαγών:*** Οι πιστωτικές κάρτες είναι εύχρηστες και προσφέρουν ταχύτητα στις καθημερινές συναλλαγές. Η ολοκλήρωση της συναλλαγής απαιτεί μόνο την εισαγωγή του PIN για την αυθεντικοποίηση του πελάτη, διαδικασία που διαρκεί ελάχιστα μόνο δευτερόλεπτα. Έτσι, ο πελάτης εξοικονομεί χρόνο και κόπο από το να αναζητά το ακριβές ποσό της αγοράς του σε χαρτονομίσματα και κέρματα.
- ***Συναλλαγές μέσω διαδικτύου/τηλεφώνου:*** Η πιστωτική κάρτα παρέχει τη δυνατότητα αγοράς αγαθών και υπηρεσιών από το διαδίκτυο με λίγες μόνο κινήσεις. Ωστόσο, επειδή πάντοτε ενέχει ο κίνδυνος απάτης, θα πρέπει οι χρήστες να είναι ιδιαίτερα προσεκτικοί σχετικά με την αξιοπιστία της εταιρείας που επιλέγουν για την αγορά τους.
- ***Αγορές με δόσεις:*** Με τη χρήση πιστωτικής κάρτας οι καταναλωτές έχουν τη δυνατότητα αγοράς αγαθών σε δόσεις. Κατά αυτό τον τρόπο, μπορούν να πραγματοποιούν αγορές ακόμα και αν εκείνη τη χρονική στιγμή δεν διαθέτουν το απαιτούμενο ποσό, έχοντας τη δυνατότητα αποπληρωμής σε μηνιαίες δόσεις. Το χαρακτηριστικό αυτό είναι εξίσου σημαντικό, τόσο για τους καταναλωτές, όσο και για τις επιχειρήσεις.
- ***Αγορές στο εξωτερικό:*** Οι πιστωτικές κάρτες αποτελούν έναν ασφαλή τρόπο συναλλαγών παγκοσμίως, ενώ λύνουν ταυτόχρονα το συχνό πρόβλημα μετατροπής των νομισμάτων, από τη στιγμή που η διαδικασία αυτή γίνεται αυτόματα από τους χρηματοπιστωτικούς φορείς.
- ***Έλεγχος μηνιαίων εξόδων:*** Ο καταναλωτής έχει τη δυνατότητα να ενημερώνεται σε πραγματικό χρόνο για το ακριβές ποσό των αγορών του, λαμβάνοντας είτε τον ενημερωτικό λογαριασμό από την τράπεζα, είτε παρακολουθώντας τις κινήσεις του μέσω e-banking. Έτσι, έχει απόλυτο έλεγχο των εξόδων του. Αντιθέτως, με τη χρήση μετρητών στις αγορές, η παραπάνω διαδικασία είναι σίγουρα πιο χρονοβόρα.

Παρά το σύνολο των πλεονεκτημάτων που αναφέρθηκαν παραπάνω, η χρήση πιστωτικών καρτών συνοδεύεται από ορισμένα μειονεκτήματα και κινδύνους.

- ***Υψηλά επιτόκια*:** Η κερδοφορία των χρηματοπιστωτικών ιδρυμάτων και των εταιρειών από τις πιστωτικές κάρτες προέρχεται από τα υψηλά επιτόκια που επιβάλλονται στα ανεξόφλητα υπόλοιπα. Συνυπολογίζοντας τις ετήσιες συνδρομές και τα έξοδα ανάληψης μετρητών, δικαίως, η πιστωτική κάρτα θεωρείται μια από τις πιο κοστοβόρες μεθόδους δανεισμού.
- ***Πιστωτικό όριο*:** Το χρηματικό όριο των συναλλαγών που μπορούν να εκτελεστούν δεν είναι απεριόριστο. Συνεπώς, ο κάτοχος οφείλει να είναι διαρκώς ενήμερος για το υπόλοιπο του.
- ***Κίνδυνος απάτης*:** Ο κίνδυνος υποκλοπής προσωπικών στοιχείων του κατόχου της κάρτας είναι πάντοτε υπαρκτός. Οι χρήστες θα πρέπει να είναι ιδιαίτερα προσεκτικοί ως προς τη χρήση των προσωπικών τους δεδομένων σε συναλλαγές που πραγματοποιούν ηλεκτρονικά με χρήση της κάρτας τους, καθώς έχουν σημειωθεί αρκετά περιστατικά ηλεκτρονικής απάτης. Επίσης, η φυσική απώλεια από τον κάτοχο ή η κλοπής της κάρτας από επιτήδειους μπορεί να οδηγήσει σε απώλεια σημαντικών ποσών.

3.3 Φαινόμενα απάτης με χρήση πιστωτικών καρτών

Οι απάτες μέσω πιστωτικών καρτών αποτελούσαν πάντοτε σοβαρό πρόβλημα, τόσο για τους κατόχους των καρτών, όσο και για τα χρηματοπιστωτικά ιδρύματα. Τα τελευταία χρόνια λόγω και της ευρείας εξάπλωσης των διαδικτυακών συναλλαγών, το πρόβλημα αυτό έχει ενταθεί. Υπολογίζεται ότι μόνο κατά το έτος 2018, σύμφωνα με έρευνες της αμερικανικής εταιρείας Shift Processing [20], έχουν χαθεί λόγω απάτης με πιστωτικές κάρτες περίπου 24,26 δισεκατομμύρια δολάρια παγκοσμίως, σημειώνοντας ετήσια αύξηση της τάξεως του 18,4%. Οι Η.Π.Α ηγούνται της σχετικής λίστας των χωρών που πλήττονται από αυτές τις απάτες, με το 38,6% των περιστατικών να λαμβάνουν χώρα εντός αυτής..



Εικόνα 11 - Ετήσια Περιστατικά Απάτης με Πιστωτικές Κάρτες στις Η.Π.Α

Πηγή: Shift Processing

Η πρόβλεψη και έγκαιρη ανίχνευση τέτοιων περιστατικών κρίνονται ιδιαίτερα σημαντικά, για τον περιορισμό του φαινομένου και τη διατήρηση του αισθήματος εμπιστοσύνης και ασφάλειας μεταξύ πελάτη-τράπεζας. Οι συνηθέστεροι τρόποι απάτης μέσω πιστωτικών καρτών είναι οι ακόλουθοι:

- Συναλλαγές μέσω κλεμμένων/χαμένων καρτών.
- Πλαστογραφία, δηλαδή υποκλοπή στοιχείων της κάρτας είτε κατά τη διάρκεια συναλλαγής στο ATM (skimming), είτε με φορητές συσκευές αντιγραφής της μαγνητικής λωρίδας. Κατά αυτόν τον τρόπο δημιουργούνται πανομοιότυπα αντίγραφα των καρτών.
- Έκδοση πιστωτικής κάρτας, δίνοντας ψευδή προσωπικά και οικονομικά στοιχεία στις αρχές έκδοσης.
- Ηλεκτρονική απάτη, με υποκλοπή και παράνομη χρήση στοιχείων της κάρτας για χρήση σε ηλεκτρονικές συναλλαγές.

Σύμφωνα με αναφορές της Ευρωπαικής Κεντρικής Τράπεζας (Ε.Κ.Τ) [21], το 2012, χανόταν λόγω απάτης 1€ για κάθε 2.635€ που δαπανούνταν σε αγορές με πιστωτικές ή χρεωστικές

κάρτες σε κράτη του Ενιαίου Χώρου Πληρωμών σε Ευρώ (SEPA). Το συνολικό ποσό των χρήματων που χάθηκαν εξαιτίας φαινομένων απάτης με χρήση κάρτας το 2012 ανήλθε στα 1,33 δισεκατομμύρια, σημειώνοντας αύξηση της τάξεως του 14,8% σε σύγκριση με το 2011. Πιο συγκεκριμένα, 60% των περιπτώσεων προήλθαν από συναλλαγές απουσία κάρτας (μέσω τηλεφώνου, διαδικτύου), 23% από συναλλαγές σε τερματικά POS και 17% από ATM. Η υποχρεωτική χρήση του προσωπικού αριθμού αναγνώρισης (PIN) σε ορισμένες συναλλαγές που ξεπερνούν ένα συγκεκριμένο όριο, είχε ως αποτέλεσμα τη μείωση των παράνομων συναλλαγών σε σχέση με συνολικό αριθμό τους, από 0,048% το 2008 σε 0,038% το 2012. Ωστόσο, λόγω της αυξησης των πληρωμών απουσία κάρτας κατά 15-20% στο ίδιο διάστημα (2008-2012), η απάτες με τον τρόπο αυτό σημείωσαν αύξηση 21%. Ένα ακόμα ενδιαφέρον χαρακτηριστικό είναι το γεγονός ότι περιστατικά απάτης απουσία κάρτας λαμβάνουν χώρα πιο συχνά στις λεγόμενες «αναπτυγμένες» χώρες, ενώ στις λιγότερο αναπτυγμένες οικονομίες συχνότερα φαινόμενα απάτης συναντώνται σε αγορές σε τερματικά μηχανήματα. Τέλος, αξίζει να αναφερθεί ότι οι πιστωτικές κάρτες είναι πιο επιρρεπείς σε περιστατικά απάτης. Υπολογίζεται ότι για συναλλαγές αξίας 1000€ με πιστωτικές κάρτες χάνεται 1€, ενώ σε αγορές με χρεωστικές κάρτες παρόμοιες απώλειες συναντώνται για αγορές αξίας 5.400€.

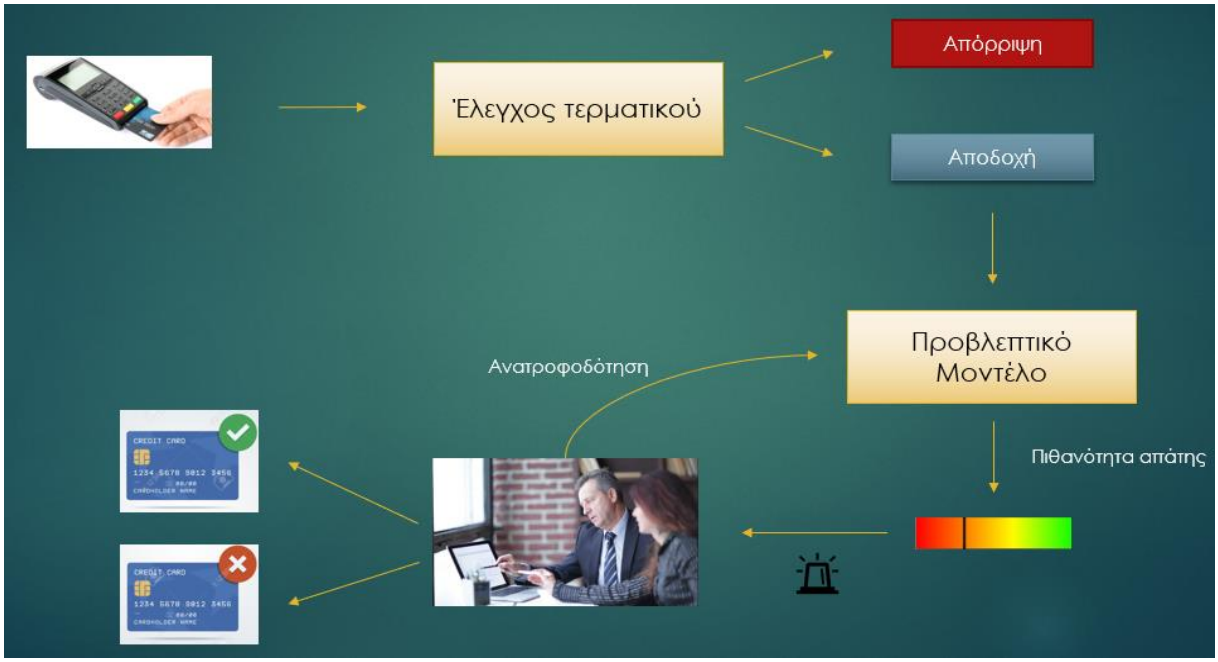
3.4 Μηχανισμοί πρόβλεψης και ανίχνευσης απάτης

Ο εντοπισμός απάτης είναι η διαδικασία χαρακτηρισμού μια συναλλαγής ως νόμιμη ή παράνομη. Ένα σύστημα εντοπισμού φαινομένων απάτης θα πρέπει όχι μόνο να εντοπίζει με ακρίβεια τις παράνομες κινήσεις, αλλά να είναι και οικονομικά ανεκτό από τα χρηματοπιστωτικά ιδρύματα. Σύμφωνα με τον Bhatla [22], ο έλεγχος του 2% των συναλλαγών, θα μπορούσε να επιφέρει μείωση των απωλειών λόγω απάτης στο 1% της συνολικής αξίας τους. Ωστόσο, ο έλεγχος του 30% των αγορών με κάρτα, θα μείωνε το αντίστοιχο ποσοστό στο 0,06%. Γίνεται αντιληπτό, λοιπόν, ότι η κατά το δυνατόν εκτενέστερη μελέτη όλο και περισσότερων συναλλαγών, συμβάλει σε μεγάλο βαθμό στην εξάλειψη του φαινομένου. Για τη μείωση του κόστους εντοπισμού παράνομων συναλλαγών με κάρτες σημαντικό ρόλο διαδραματίζουν τόσο τα στατιστικά μοντέλα, όσο και οι τομείς της εξόρυξης δεδομένων και της μηχανικής μάθησης.

Οι τεχνικές εξόρυξης δεδομένων επιτρέπουν την ανακάλυψη προτύπων από πολυδιάστατες ροές δεδομένων σε εξαιρετικά μικρά χρονικά διαστήματα. Τα στοιχεία που προέρχονται από συναλλαγές ενημερώνονται συνεχώς, οπότε η επεργασία τους σε πραγματικό χρόνο και όχι

κατά δεσμίδες κρίνεται επιβεβλημένη. Επίσης, έχουν τη δυνατότητα να εντοπίζουν έγκαιρα τις νέες στρατηγικές που ακολουθούν οι επιτήδειοι, παρακολουθώντας αιφνίδιες μεταβολές στην καταναλωτική συμπεριφορά του κατόχου της κάρτας.

Τα χρηματοπιστωτικά ιδρύματα διαθέτουν στις βάσεις δεδομένων τους τεράστιο όγκο δεδομένων που έχουν να κάνουν με τα στοιχεία των συναλλαγών με χρήση πιστωτικών καρτών. Κατά τη διάρκεια οποιασδήποτε συναλλαγής καταγράφονται σημαντικά στοιχεία όπως τα προσωπικά στοιχεία του κατόχου της κάρτας, το ύψος του ποσού της συναλλαγής, ο τύπος της κάρτας, η τοποθεσία που έλαβε χώρα η συναλλαγή, οι λογαριασμοί του αποδέκτη της αγοράς. Τα μοντέλα μηχανικής μάθησης έχουν τη δυνατότητα να επεξεργάζονται τον μεγάλο αυτό όγκο δεδομένων και να προβλέπουν την πιθανότητα να είναι μια συναλλαγή προιόν απάτης. Στη συνέχεια, οι αρμόδιοι των τμημάτων ασφάλειας συναλλαγών των χρηματοπιστωτικών ιδρυμάτων ελέγχουν τις ύποπτες κινήσεις και εντοπίζουν ποιές από αυτές ήταν όντως παράνομες. Ακολούθως το σύστημα ανατροφοδοτείται με τα νέα πλέον δεδομένα, διαδικασία που συμβάλει στη βελτίωση της ακρίβειας πρόβλεψης.



Εικόνα 12 – Λειτουργία μηχανισμών πρόβλεψης ύποπτων συναλλαγών

ΚΕΦΑΛΑΙΟ 4: Πειραματικό μέρος – Υλοποίηση Αλγορίθμων

Στο σημείο αυτό, κρίνεται επιβεβλημένο, πέρα από το θεωρητικό κομμάτι της σημασίας των τεχνικών εξόρυξης δεδομένων στην πρόβλεψη της τραπεζικής απάτης, να αναδειχθεί στην πράξη και ο τρόπος λειτουργίας τους. Για το λόγο αυτό, διεξήχθη μια σειρά πειραματικών δοκιμών πάνω σε πραγματικά δεδομένα συναλλαγών με χρήση καρτών, με σκοπό την όσο το δυνατόν ακριβέστερη πρόβλεψη για το αν μια συναλλαγή είναι νόμιμη ή παράνομη. Συγκεκριμένα υλοποιήθηκαν οι αλγόριθμοι μηχανικής μάθησης **LightGBM, XGBoost** και **Random Forest**, με χρήση διαφόρων τεχνικών επανα-δειγματοληψίας (Random undersampling, Random Oversampling, SMOTE)**.** Εν συνεχεία, έγινε εφαρμογή της μεθόδου συσσωρευμένης γενίκευσης (Stacked Generalization) ή αλλιώς Stacking, όπου τα αποτελέσματα των παραπάνω αλγορίθμων χρησιμοποιήθηκαν ως είσοδος σε έναν μετα-ταξινομήτη (**Logistic Regression**). Η τεχνική αυτή έδειξε να επιφέρει καλύτερα αποτελέσματα πρόβλεψης αλλά ταυτόχρονα και σημαντικά χαμηλότερο χρόνο εκτέλεσης.

4.1 Περιγραφή Δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν για τους σκοπούς της εργασίας είναι δύο αρχεία τύπου csv, τα οποία περιέχουν πληροφορίες για συναλλαγές που πραγματοποιήθηκαν ηλεκτρονικά με χρήση πλαστικού χρήματος και συγκεκριμένα πλαστικής κάρτας. Πρόκειται για πραγματικά δεδομένα που διατίθενται από την, παγκοσμίου φήμης, εταιρεία παροχής υπηρεσιών πληρωμών, Vesta Corporation. Τα δύο αρχεία έχουν μια κοινή μεταβλητή "TransactionID" βάση της οποίας θα γίνει η συνένωση τους, με σκοπό τη δημιουργία ενός τελικού ενοποιημένου αρχείου, πάνω στο οποίο θα πραγματοποιηθεί η εφαρμογή των επιλεχθέντων αλγορίθμων.

4.1.1 Σύνολο Δεδομένων "train_transaction"

Το πρώτο σύνολο δεδομένων είναι το "train_transaction", το οποίο περιέχει πληροφορίες σχετικά με συναλλαγές που πραγματοποιήθηκαν με χρήση πλαστικής κάρτας και καλύπτουν ένα χρονικό διάστημα περίπου 6 μηνών. Ορισμένες από τις πληροφορίες που παρέχονται είναι το χρονικό στίγμα και το ποσό της συναλλαγής, το αναγνωριστικό του προιόντος που αγοράστηκε, τα ακριβή στοιχεία της κάρτας, το ιστορικό προηγούμενων συναλλαγών κα. Επίσης συμπεριλαμβάνεται η μεταβλητή "isFraud" που υποδηλώνει την ιδιότητα της

συναλλαγής, λαμβάνοντας την τιμή «0» όταν εκείνη είναι νόμιμη και την τιμή «1» όταν πρόκειται για προιόν απάτης.

Αριθμεί περί των 590.540 εγγραφών και 394 μεταβλητών, εκ των οποίων οι 380 είναι αριθμητικές και οι 14 κατηγορικές. Προτού γίνει περαιτέρω ανάλυση του συνόλου "train_transaction", θα φιλτραριστούν, με την ακόλουθη εντολή, οι συναλλαγές εκείνες που έχουν να κάνουν μόνο με χρήση πιστωτικής κάρτας, δημιουργώντας παράλληλα ένα νέο αρχείο csv, με την ονομασία "cc_train_transaction".

```
## φιλτράρισμα μόνο των εγγραφών που αφορούν πιστωτικές κάρτες

cc_train_transaction = train_transaction[train_transaction["card6"]=="credit"]
cc_train_transaction = pd.read_csv(r"C:\Users\Kostas\Desktop\cc_train_transaction.csv", index_col=0)
```

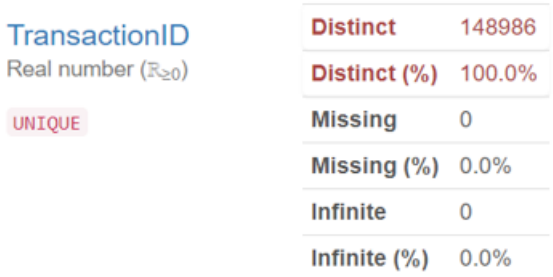
Το νέο πλέον αρχείο, "cc_train_transaction", αριθμεί 148.986 εγγραφές και 394 στήλες (380 αριθμητικές, 14 κατηγορικές).

Dataset statistics

Number of variables	394
Number of observations	148986
Missing cells	21266381
Missing cells (%)	36.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	447.8 MiB
Average record size in memory	3.1 KiB

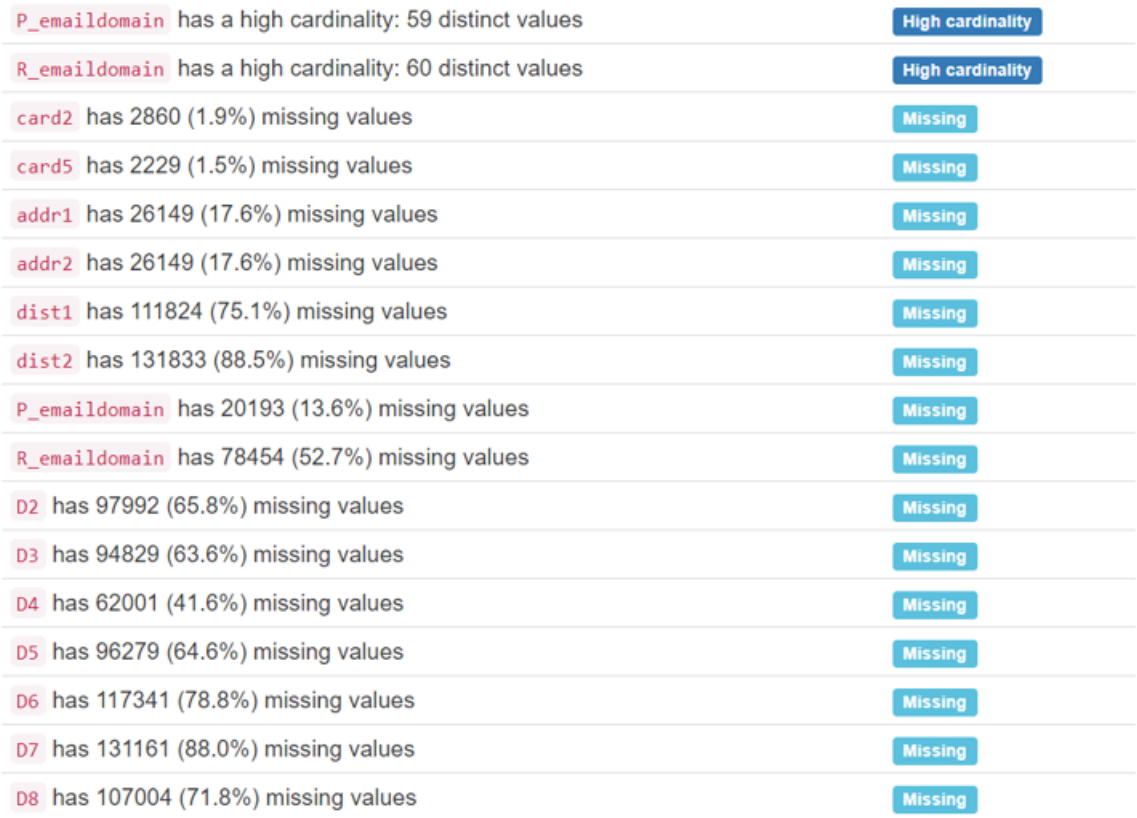
Εικόνα 13 - Στατιστικά Στοιχεία Συνόλου "cc_train_transaction"

Κατόπιν θα διαγραφεί η στήλη "card6" που περιείχε πληροφορίες σχετικά με τον τύπο της κάρτας (πιστωτική, χρεωστική, προπληρωμένη), αφού πλέον δεν έχει κάποια χρησιμότητα. Η στήλη "TransactionID" αποτελεί το αναγνωριστικό κάθε συναλλαγής και περιέχει μοναδικές τιμές (unique values).



Εικόνα 14 - Στατιστικά Στοιχεία Μεταβλητής "TransactionID"

Με μια πρώτη ανάγνωση του αρχείου, παρατηρείται ότι πολλές μεταβλητές έχουν πληθώρα ελλειπουσών τιμών (missing values). Ενδεικτικά, παρατίθενται στην ακόλουθη εικόνα μερικές από τις μεταβλητές αυτές.



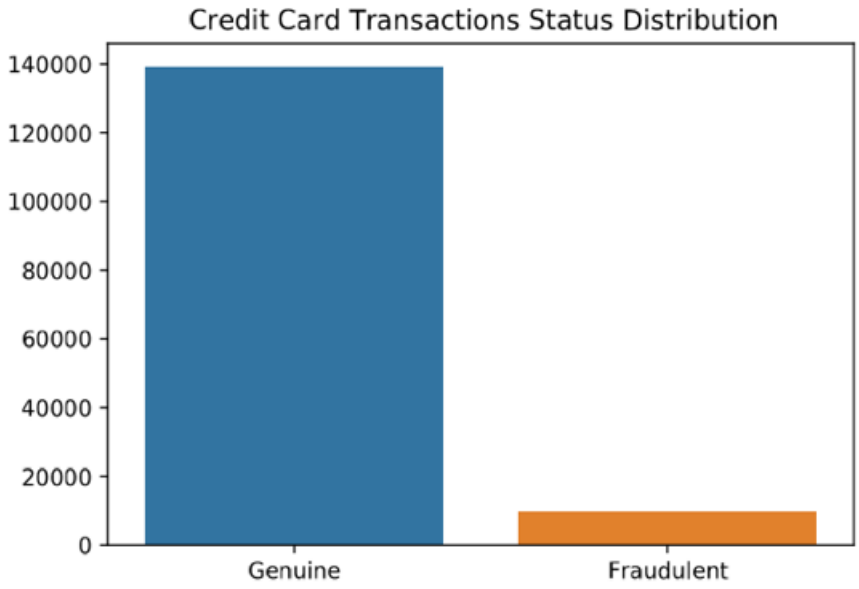
Εικόνα 15 - Δείγμα Μεταβλητών με Ελλείπουσες Τιμές Συνόλου "cc_train_transaction"

Ιδιαίτερο ενδιαφέρον παρουσιάζει η μεταβλητή "isFraud", από την οποία μπορεί να αποτυπωθεί το ποσοστό των συναλλαγών που είναι προιόν απάτης.

```
## ποσοστό φαινομένων απάτης με χρήση πιστωτικών καρτών

print(cc_train_transaction["isFraud"].value_counts())
print("percent of fraud trans: ", len(cc_train_transaction.loc[cc_train_transaction.isFraud == 1])*100/len(cc_train_transaction)
```

```
0    139036
1      9950
Name: isFraud, dtype: int64
percent of fraud trans:  6.678479857167788
```

Παρατηρείται ότι το 6,68% των συναλλαγών είναι παράνομες, συνεπώς τα δεδομένα βρίσκονται σε ανισορροπία ως προς την κλάση τους (γνήσιες/δόλιες).



Εικόνα 16 - Κατανομή Συναλλαγών ως προς την Κλάση τους (Γνήσιες/Δόλιες)

4.1.2 Σύνολο Δεδομένων "train_identity"

Το δεύτερο σύνολο δεδομένων είναι το "train_identity" που περιέχει πληροφορίες σχετικά με την ταυτότητα του κατόχου της κάρτας, το δίκτυο που χρησιμοποιήθηκε (ιστοσελίδες, IP, Proxy) και στοιχεία περιηγητών, λειτουργικών συστημάτων κ.α. Τα δεδομένα αυτά έχουν συλλεχθεί από τα τμήματα ασφαλείας συναλλαγών της Vesta Corporation και εξωτερικούς της συνεργάτες. Αξίζει να τονιστεί ότι για λόγους προστασίας προσωπικών δεδομένων, έχουν υποστεί επεξεργασία, διασφαλίζοντας έτσι την ανωνυμία των χρηστών.

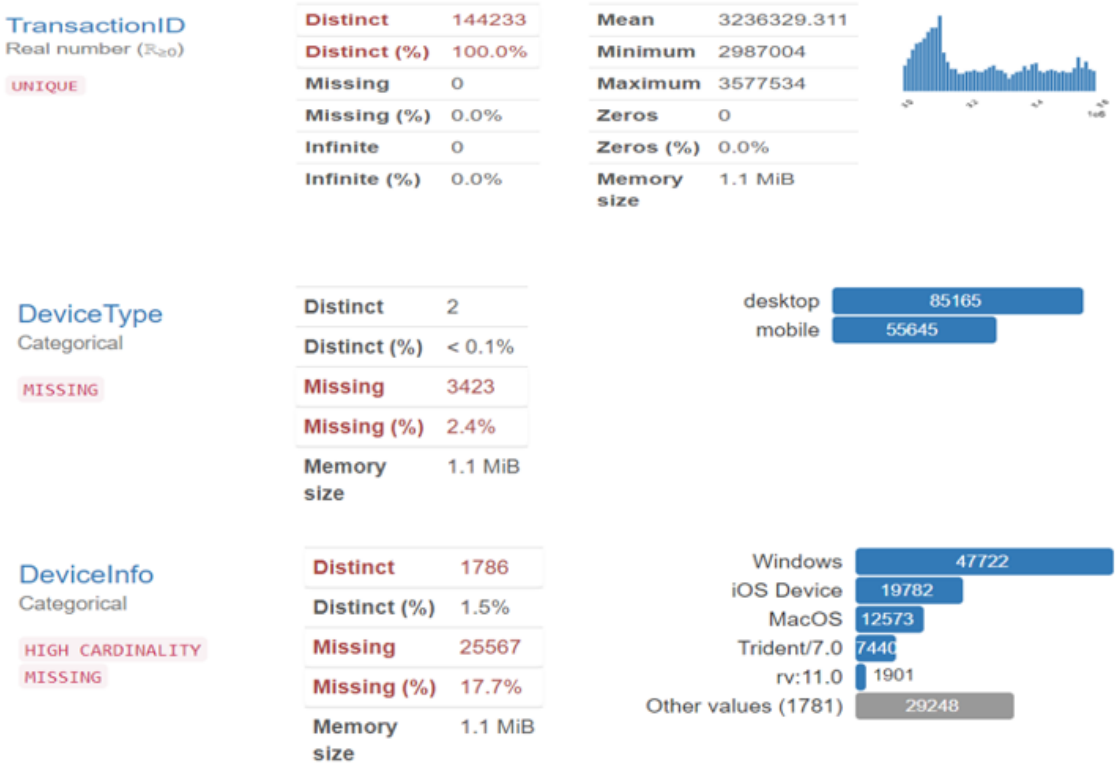
Αριθμεί 144.233 εγγραφές και 41 μεταβλητές, εκ των οποίων οι 24 είναι αριθμητικές και οι 17 κατηγορικές. Όπως και στο σύνολο δεδομένων "cc_train_transaction", έτσι και εδώ, κάθε εγγραφή χαρακτηρίζεται από έναν αναγνωριστικό αριθμό, τη στήλη "TransactionID", η οποία είναι μοναδικοποιημένη, ενώ παράλληλα το 35,6% των τιμών είναι μηδενικές ή κενές.

Dataset statistics

Number of variables	41
Number of observations	144233
Missing cells	2104107
Missing cells (%)	35.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	45.1 MiB
Average record size in memory	328.0 B

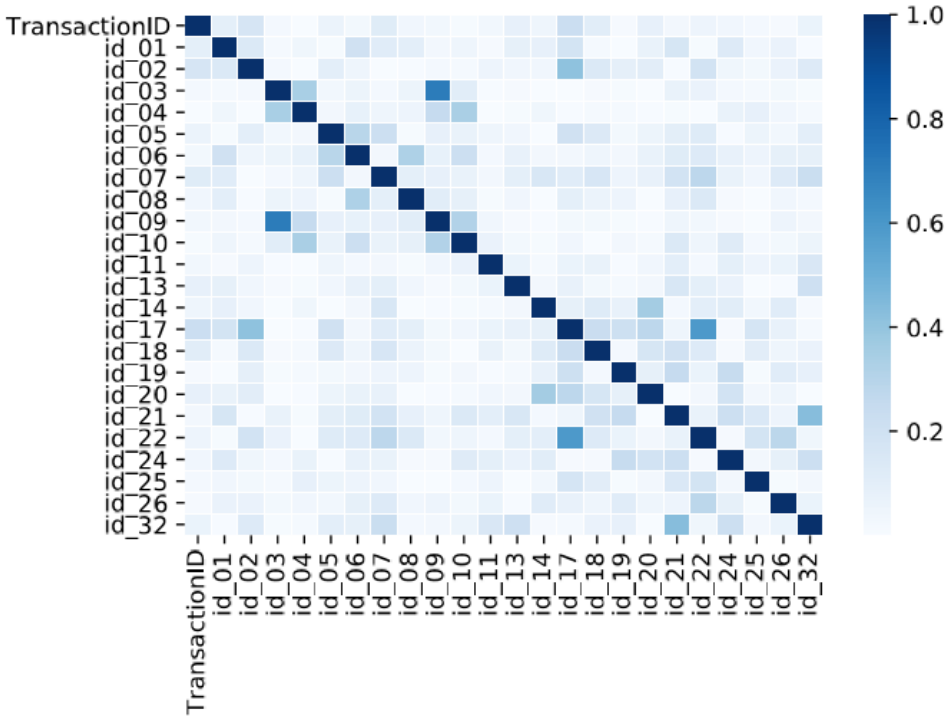
Εικόνα 17 - Στατιστικά Στοιχεία Συνόλου "train_identity"

Επίσης, αξίζει να προσεχθούν οι μεταβλητές "DeviceType" και "DeviceInfo" που προσδιορίζει, η μεν τη συσκευή από την οποία πραγματοποιήθηκε η ηλεκτρονική συναλλαγή, η δε πληροφορίες σχετικά με το λειτουργικό σύστημα της συσκευής.



Εικόνα 18 - Στατιστικά Στοιχεία Μεταβλητών "TransactionID", "DeviceType", "DeviceInfo"

Για να εξεταστεί κατά πόσον οι μεταβλητές του συνόλου “train_identity” έχουν υψηλή συσχέτιση μεταξύ τους, δημιουργήθηκε το γράφημα συσχετίσεων με χρήση του συντελεστή “Pearson” [23].



Εικόνα 19 - Γραφική Απεικόνιση Συσχέτισης Μεταβλητών Συνόλου “train_identity”

Μεταβλητές με υψηλό βαθμό συσχέτισης μεταξύ τους είναι περισσότερο γραμμικά εξαρτημένες και επομένως έχουν την ίδια επίδραση στην εξαρτημένη μεταβλητή-στόχο. Για το λόγο αυτό, όταν δύο χαρακτηριστικά έχουν υψηλό βαθμό συσχέτισης, μπορεί να εξαιρεθεί από το μοντέλο πρόβλεψης, ένα από τα δύο. Παρατηρείται ότι οι μεταβλητές του συνόλου “train_identity” είναι σε μεγάλο βαθμό στατιστικά ανεξάρτητες η μία από την άλλη, συνεπώς θα πρέπει να ληφθούν όλες υπόψη.

4.2 Προεπεξεργασία Δεδομένων

Σκοπός στο σημείο αυτό είναι η δημιουργία όσο το δυνατόν περισσότερων μεταβλητών που θα περιέχουν σημαντική πληροφορία για την πρόβλεψη της μεταβλητής στόχου “isFraud”. Όπως αναφέρθηκε και παραπάνω τα δεδομένα περιέχουν αρκετές ελλείπουσες και κενές τιμές. Πρωταρχικό μέλημα, λοιπόν, είναι η διαχείριση αυτών των μεταβλητών.

Ξεκινώντας από το σύνολο δεδομένων “cc_train_transaction” εντοπίζονται σε πρώτη φάση οι μεταβλητές αριθμητικού και κατηγορικού τύπου.

```
['ProductCD', 'card4', 'P_emaildomain', 'R_emaildomain', 'M1', 'M2', 'M3', 'M4', 'M5', 'M6', 'M7', 'M8', 'M9']

number categorical transaction features:  13


['TransactionID', 'isFraud', 'TransactionDT', 'TransactionAmt', 'card1', 'card2', 'card3', 'card5', 'addr1', 'addr2', 'dist1',
'dist2', 'C1', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8', 'C9', 'C10', 'C11', 'C12', 'C13', 'C14', 'D1', 'D2', 'D3', 'D4', 'D5',
'D6', 'D7', 'D8', 'D9', 'D10', 'D11', 'D12', 'D13', 'D14', 'D15', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28',
'V29', 'V30', 'V31', 'V32', 'V33', 'V34', 'V35', 'V36', 'V37', 'V38', 'V39', 'V40', 'V41', 'V42', 'V43', 'V44', 'V45', 'V46',
'V47', 'V48', 'V49', 'V50', 'V51', 'V52', 'V53', 'V54', 'V55', 'V56', 'V57', 'V58', 'V59', 'V60', 'V61', 'V62', 'V63', 'V64',
'V65', 'V66', 'V67', 'V68', 'V69', 'V70', 'V71', 'V72', 'V73', 'V74', 'V75', 'V76', 'V77', 'V78', 'V79', 'V80', 'V81', 'V82',
'V83', 'V84', 'V85', 'V86', 'V87', 'V88', 'V89', 'V90', 'V91', 'V92', 'V93', 'V94', 'V95', 'V96', 'V97', 'V98', 'V99', 'V100',
'V101', 'V102', 'V103', 'V104', 'V105', 'V106', 'V107', 'V108', 'V109', 'V110', 'V111', 'V112', 'V113', 'V114', 'V115', 'V116',
'V117', 'V118', 'V119', 'V120', 'V121', 'V122', 'V123', 'V124', 'V125', 'V126', 'V127', 'V128', 'V129', 'V130', 'V131', 'V132',
'V133', 'V134', 'V135', 'V136', 'V137', 'V138', 'V139', 'V140', 'V141', 'V142', 'V143', 'V144', 'V145', 'V146', 'V147', 'V148',
'V149', 'V150', 'V151', 'V152', 'V153', 'V154', 'V155', 'V156', 'V157', 'V158', 'V159', 'V160', 'V161', 'V162', 'V163', 'V164',
'V165', 'V166', 'V167', 'V168', 'V169', 'V170', 'V171', 'V172', 'V173', 'V174', 'V175', 'V176', 'V177', 'V178', 'V179', 'V180',
'V181', 'V182', 'V183', 'V184', 'V185', 'V186', 'V187', 'V188', 'V189', 'V190', 'V191', 'V192', 'V193', 'V194', 'V195', 'V196',
'V197', 'V198', 'V199', 'V200', 'V201', 'V202', 'V203', 'V204', 'V205', 'V206', 'V207', 'V208', 'V209', 'V210', 'V211', 'V212',
'V213', 'V214', 'V215', 'V216', 'V217', 'V218', 'V219', 'V220', 'V221', 'V222', 'V223', 'V224', 'V225', 'V226', 'V227', 'V228',
'V229', 'V230', 'V231', 'V232', 'V233', 'V234', 'V235', 'V236', 'V237', 'V238', 'V239', 'V240', 'V241', 'V242', 'V243', 'V244',
'V245', 'V246', 'V247', 'V248', 'V249', 'V250', 'V251', 'V252', 'V253', 'V254', 'V255', 'V256', 'V257', 'V258', 'V259', 'V260',
'V261', 'V262', 'V263', 'V264', 'V265', 'V266', 'V267', 'V268', 'V269', 'V270', 'V271', 'V272', 'V273', 'V274', 'V275', 'V276',
'V277', 'V278', 'V279', 'V280', 'V281', 'V282', 'V283', 'V284', 'V285', 'V286', 'V287', 'V288', 'V289', 'V290', 'V291', 'V292',
'V293', 'V294', 'V295', 'V296', 'V297', 'V298', 'V299', 'V300', 'V301', 'V302', 'V303', 'V304', 'V305', 'V306', 'V307', 'V308',
'V309', 'V310', 'V311', 'V312', 'V313', 'V314', 'V315', 'V316', 'V317', 'V318', 'V319', 'V320', 'V321', 'V322', 'V323', 'V324',
'V325', 'V326', 'V327', 'V328', 'V329', 'V330', 'V331', 'V332', 'V333', 'V334', 'V335', 'V336', 'V337', 'V338', 'V339']

number numerical transaction features:  380
```

Εικόνα 20 - Μεταβλητές Συνόλου "cc_train_transaction"

Αντίστοιχη διαδικασία ακολουθείται για τις μεταβλητές του συνόλου "train_identity".

```
['id_12', 'id_15', 'id_16', 'id_23', 'id_27', 'id_28', 'id_29', 'id_30', 'id_31', 'id_33', 'id_34', 'id_35', 'id_36', 'id_37',
'id_38', 'DeviceType', 'DeviceInfo']

number categorical identity features:  17


['TransactionID', 'id_01', 'id_02', 'id_03', 'id_04', 'id_05', 'id_06', 'id_07', 'id_08', 'id_09', 'id_10', 'id_11', 'id_13',
'id_14', 'id_17', 'id_18', 'id_19', 'id_20', 'id_21', 'id_22', 'id_24', 'id_25', 'id_26', 'id_32']

number numerical identity features:  24
```

Εικόνα 21 - Μεταβλητές Συνόλου "train_identity"

Στη συνέχεια ακολουθεί η διαχείριση των ελλειπουσών τιμών. Το πλάνο που θα ακολουθηθεί έχει ως εξής. Οι αριθμητικές μεταβλητές που έχουν αριθμό ελλειπουσών τιμών πάνω από το 50% του συνόλου τους, θα διαγραφούν, καθώς αν συμπεριληφθούν στην πρόβλεψη θα χάνεται σημαντική πληροφορία για πολλές συναλλαγές. Επίσης, οι ελλείπουσες τιμές των μεταβλητών, που αποτελούν ποσοστό μεταξύ 15% - 50%, θα αντικατασταθούν με την διάμεση τιμή της στήλης. Τέλος, οι ελλείπουσες τιμές των μεταβλητών με ποσοστό μικρότερο του 15% missing values, θα αντικατασταθούν με τη μέση τιμή των υπολοίπων τιμών της στήλης.

Όσον αφορά τις κατηγορικές μεταβλητές, η στρατηγική που θα ακολουθηθεί έχει ώς εξής. Θα διαγραφούν οι μεταβλητές που έχουν ποσοστό άνω του 50% missing values, για τους ίδιους

λόγους που αναφέρθηκαν παραπάνω. Για τις υπόλοιπες, θα αντικατασταθούν οι ελλείπουσες τιμές με την πιο συνά εμφανιζόμενη τιμή της εκάστοτε στήλης. Ακολούθως, θα χρησιμοποιηθεί η μέθοδος Label Encoder για τις στήλες με αρκετές μοναδικές τιμές, ενώ για αυτές με λίγες μοναδικές τιμές θα χρησιμοποιηθεί η μέθοδος One Hot Encoder. Προφανώς αυτό συμβαίνει διότι με τη μέθοδο One Hot Encoder δημιουργούνται τόσες στήλες όσες και οι μοναδικές τιμές της στήλης που εφαρμόζεται. Συνεπώς, σε περίπτωση που εφαρμοζόταν σε όλες τις κατηγορικές μεταβλητές, θα δημιουργούνταν πληθώρα μεταβλητών που θα επηρέαζε αρνητικά τόσο την πρόβλεψη, όσο και τον χρόνο εκτέλεσης των αλγορίθμων.

Στην ακόλουθη εικόνα παρουσιάζονται οι κατηγορικές μεταβλητές με μικρό (low_card_cols) και μεγάλο (high_card_cols) αριθμό μοναδικών τιμών.

cc_train_transaction

```
low_card_trans_cols = ["ProductCD", "card4"]
high_card_trans_cols = ["P_emaildomain"]
```

train_identity

```
low_card_id_cols =  ["id_12", "id_15", "id_16", "id_28", "id_29", "id_34", "id_35", "id_36", "id_37", "id_38", "DeviceType"]
high_card_id_cols = ["id_30", "id_31", "id_33", "DeviceInfo"]
```

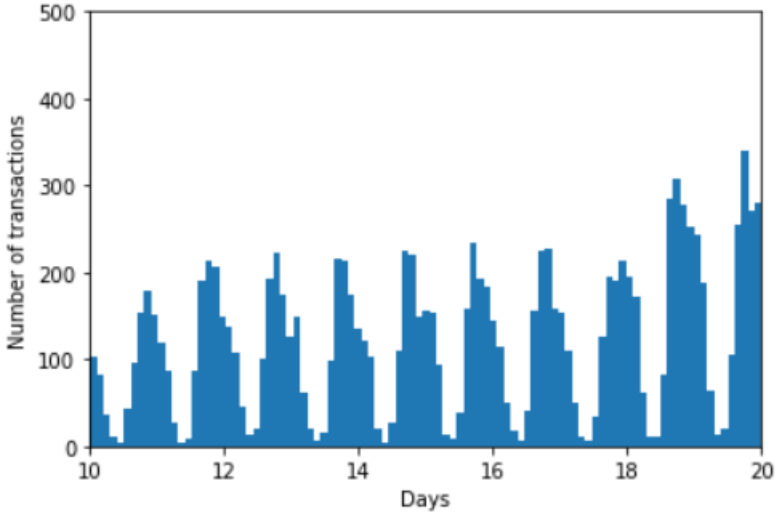
Εικόνα 22 - Κατηγορικές Μεταβλητές Συνόλων Συναρτήσει των Μοναδικών Τιμών τους

Στο σημείο αυτό, θα γίνει η ένωση των δύο συνόλων δεδομένων ως προς την κοινή μεταβλητή “TransactionID”, διαγράφοντας στη συνέχεια τη στήλη, από τη στιγμή που δεν προσφέρει πλέον κάποια αξία στην μετέπειτα ανάλυση που θα ακολουθήσει.

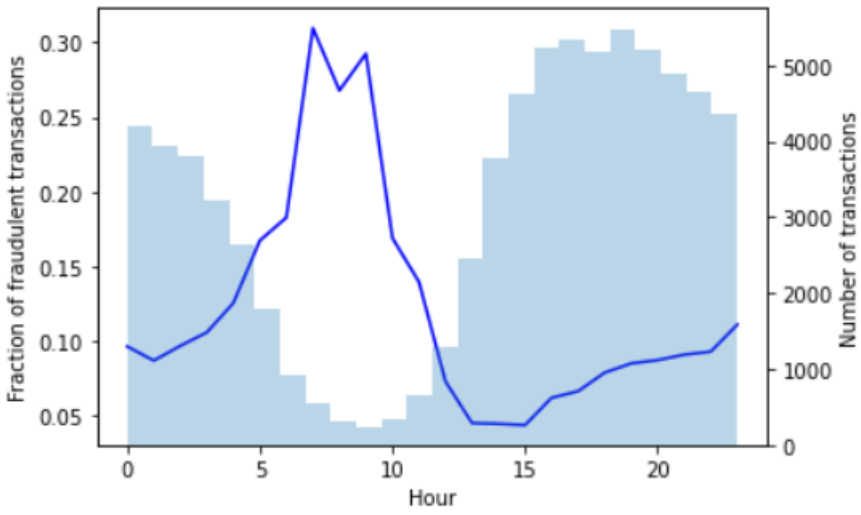
Η δημιουργία νέων μεταβλητών από τις ήδη υπάρχουσες έχει δείξει να βελτιώνει την απόδοση των προβλεπτικών μεθόδων. Για το λόγο αυτό, ύστερα από προσεκτική μελέτη των μεταβλητών, δημιουργήθηκαν δύο νέες στήλες, ονόματι “Trans_min_mean” και “Trans_min_std”, εκμεταλλευόμενοι πληροφορίες από τη στήλη “TransactionAmt” και τα στατιστικά δεδομένα της (μέση τιμή, τυπική απόκλιση).

Επίσης, παρατηρήθηκε ότι η στήλη “TransactionDT”, που περιλαμβάνει δεδομένα σχετικά με το χρονικό στίγμα της συναλλαγής, μπορεί να περιέχει χρήσιμη πληροφορία. Η ελάχιστη τιμή της μεταβλητής είναι 86.400 που αντικατοπτρίζει τον αριθμό των δευτερολέπτων μίας ημέρας (60*60*24 = 86400), ενώ ο μέγιστη είναι 15.811.131 που αντιστοιχεί στην ημέρα 183. Θα ακολουθήσει μια προσπάθεια να εντοπιστούν ορισμένες συσχετίσεις μεταξύ του πλήθους και

του είδους των συναλλαγών ανά ημέρες και ώρες. Παρατηρείται ότι το πλήθος των συναλλαγών που έχουμε στη διάθεση μας είναι ισοκατανεμημένο ανά τις ημέρες (εικόνα 23). Ωστόσο, πραγματοποιώντας το διάγραμμα του είδους των συναλλαγών κατά τη διάρκεια των ωρών μιας ημέρας, παρατηρείται μια ξεκάθαρη διαφοροποίηση. Συγκεκριμένα, εντοπίζεται ότι τις πρώτες πρωινές ώρες (5-8 π.μ) που δεν πραματοποιούνται αρκετές συναλλαγές, σχεδόν το 25-30% εξ αυτών είναι απατηλές (εικόνα 24). Το γεγονός αυτό πιθανότατα οφείλεται σε περιστατικά διεθνούς απάτης, που λαμβάνει χώρα την ώρα που οι κάτοχοι δεν κάνουν χρήση της κάρτας τους ή κοιμούνται.



Εικόνα 23 - Πλήθος Συναλλαγών Ανά Ημέρα



Εικόνα 24 - Πιθανότητα Απάτης Ανά Ώρα Συναλλαγής

Εκμεταλλευόμενοι, λοιπόν, τις παραπάνω λεπτομέρειες, θα δημιουργηθεί μια νέα στήλη με την ονομασία “Hour”, που θα υποδεικνύει την ώρα που έλαβε χώρα κάθε συναλλαγή. Μετά την

διαδικασία αυτή, θα διαγραφεί η στήλη “TransactionDT” από τη στιγμή που περιέχει μοναδικοποιημένες τιμές και συνεπώς θα επηρεάζει σε μεγάλο βαθμό το αποτέλεσμα της πρόβλεψης. Τέλος, θα διαγραφούν από το τελικό σύνολο (“final_dataset”) που δημιουργήθηκε, οι στήλες με μηδενική τυπική απόκλιση αφού δεν συνεισφέρουν στην πρόβλεψη.

Το τελικό σύνολο, με ονομασία “final_dataset”, που προκύπτει κατόπιν των παραπάνω διαδικασιών αποτελείται πλέον από 75.090 εγγραφές και 230 μεταβλητές.

```
['isFraud', 'TransactionAmt', 'card1', 'card2', 'card3', 'card5',
 'addr1', 'addr2', 'P_emaildomain', 'C1', 'C2', 'C3', 'C4', 'C6',
 'C7', 'C8', 'C10', 'C11', 'C12', 'C13', 'C14', 'D1', 'D4', 'D10',
 'D15', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21',
 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'V31', 'V32',
 'V33', 'V34', 'V37', 'V38', 'V39', 'V40', 'V41', 'V42', 'V43',
 'V44', 'V45', 'V46', 'V47', 'V50', 'V51', 'V52', 'V55', 'V56',
 'V57', 'V58', 'V59', 'V60', 'V61', 'V62', 'V63', 'V64', 'V65',
 'V66', 'V67', 'V68', 'V71', 'V72', 'V73', 'V74', 'V77', 'V78',
 'V79', 'V80', 'V81', 'V82', 'V83', 'V84', 'V85', 'V86', 'V87',
 'V88', 'V89', 'V92', 'V93', 'V94', 'V95', 'V96', 'V97', 'V98',
 'V99', 'V100', 'V101', 'V102', 'V103', 'V104', 'V105', 'V106',
 'V107', 'V108', 'V109', 'V110', 'V111', 'V112', 'V113', 'V114',
 'V115', 'V116', 'V117', 'V118', 'V119', 'V120', 'V121', 'V122',
 'V123', 'V124', 'V125', 'V126', 'V127', 'V128', 'V129', 'V130',
 'V131', 'V132', 'V133', 'V134', 'V135', 'V136', 'V137', 'V279',
 'V280', 'V281', 'V282', 'V283', 'V284', 'V285', 'V286', 'V287',
 'V288', 'V289', 'V290', 'V291', 'V292', 'V293', 'V294', 'V295',
 'V296', 'V297', 'V298', 'V299', 'V300', 'V301', 'V302', 'V303',
 'V304', 'V306', 'V307', 'V308', 'V309', 'V310', 'V311', 'V312',
 'V313', 'V314', 'V315', 'V316', 'V317', 'V318', 'V319', 'V320',
 'V321', 'ProductCD_C', 'ProductCD_H', 'ProductCD_R', 'ProductCD_S',
 'card4_american express', 'card4_discover', 'card4_mastercard',
 'card4_visa', 'id_01', 'id_02', 'id_05', 'id_06', 'id_09', 'id_10',
 'id_11', 'id_13', 'id_14', 'id_17', 'id_19', 'id_20', 'id_30',
 'id_31', 'id_32', 'id_33', 'DeviceInfo', 'id_12_Found',
 'id_12_NotFound', 'id_15_Found', 'id_15_New', 'id_15_Unknown',
 'id_16_Found', 'id_16_NotFound', 'id_28_Found', 'id_28_New',
 'id_29_Found', 'id_29_NotFound', 'id_34_match_status:-1',
 'id_34_match_status:0', 'id_34_match_status:1',
 'id_34_match_status:2', 'id_35_F', 'id_35_T', 'id_36_F', 'id_36_T',
 'id_37_F', 'id_37_T', 'id_38_F', 'id_38_T', 'DeviceType_desktop',
 'DeviceType_mobile', 'Trans_min_mean', 'Trans_min_std', 'Hour'],
```

Εικόνα 25 - Μεταβλητές Συνόλου “final_dataset”

Στην ακόλουθη εικόνα παρατίθενται οι τύποι των μεταβλητών του τελικού συνόλου δεδομένων.

```
 #   Column                Dtype
---  ------                -----
 0   isFraud               float64
 1   TransactionAmt        float64
 2   card1                 float64
 3   card2                 float64
 4   card3                 float64
 5   card5                 float64
 6   addr1                 float64
 7   addr2                 float64
 8   P_emaildomain         int64
 9   C1                    float64
 10  C2                    float64
 11  C3                    float64
 12  C4                    float64
 13  C6                    float64
 14  C7                    float64
 15  C8                    float64
 16  C10                   float64
 17  C11                   float64
 18  C12                   float64
 19  C13                   float64
 20  C14                   float64
 21  D1                    float64
 22  D4                    float64
 23  D10                   float64
 24  D15                   float64
 25  V14                   float64
 26  V15                   float64
 27  V16                   float64
 28  V17                   float64
 29  V18                   float64
 30  V19                   float64

 217 id_35_F               int64
 218 id_35_T               int64
 219 id_36_F               int64
 220 id_36_T               int64
 221 id_37_F               int64
 222 id_37_T               int64
 223 id_38_F               int64
 224 id_38_T               int64
 225 DeviceType_desktop    int64
 226 DeviceType_mobile     int64
 227 Trans_min_mean        float64
 228 Trans_min_std         float64
 229 Hour                  float64
dtypes: float64(192), int64(38)
```

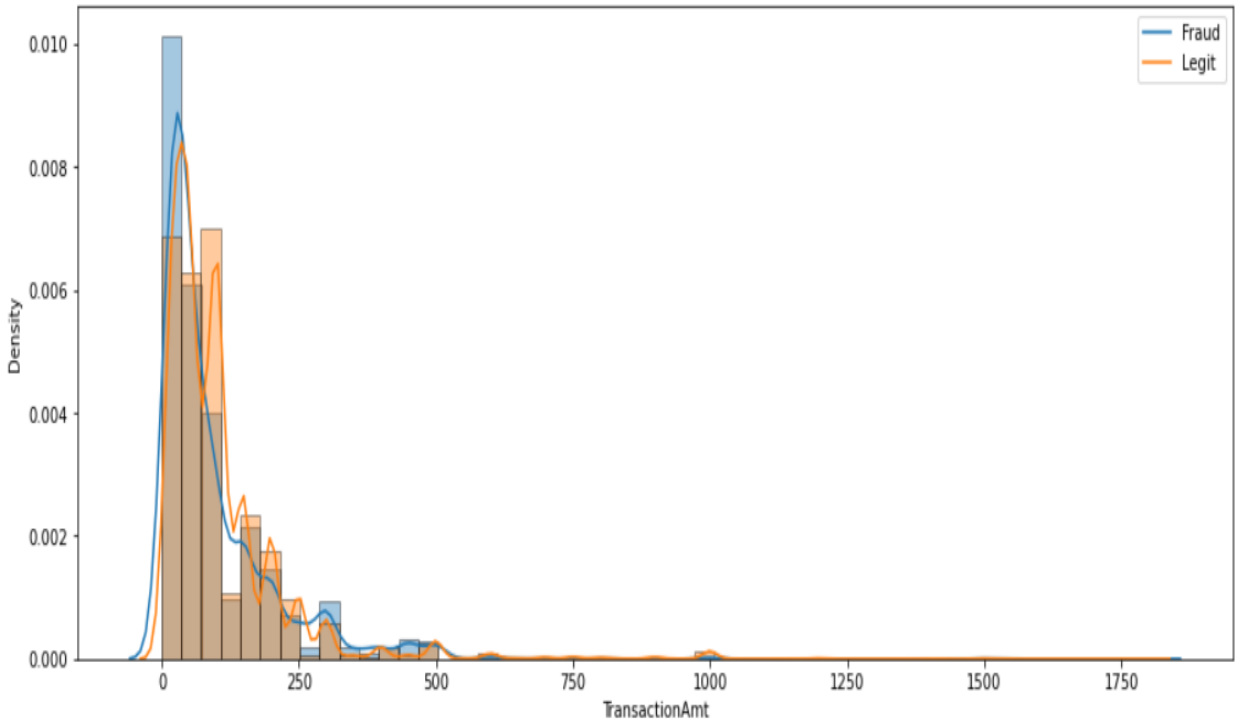
Εικόνα 26 - Τύποι Μεταβλητών Συνόλου "final_dataset"

Κρίσιμο είναι να διαπιστωθεί εάν οι μεταβλητές των δεδομένων που περιγράφουν τις νόμιμες και τις απατηλές συναλλαγές ακολουθούν τις ίδες κατανομές. Για το σκοπό αυτό θα γίνει χρήση ενός «ελέγχου καλής προσαρμογής» πάνω στα δεδομένα και συγκεκριμένα του τεστ Kolmogorov-Smirnov [24]. Το κριτήριο Kolmogorov-Smirnov είναι ένας μη παραμετρικός

έλεγχος που χρησιμοποιείται για να εξετάσει την καλή προσαρμογή ενός τυχαίου δείγματος σε μία δεδομένη κατανομή (one-sample K-S test) ή για τη σύγκριση δύο τυχαίων δειγμάτων (two-sample K-S test). Πήρε την ονομασία του από τους Andrey Kolmogorov και Nikolai Smirnov.

Παρακάτω παρατίθεται ενδεικτικά, το αποτέλεσμα του ελέγχου για τη μεταβλητή "TransactionAmt", για τα δύο δείγματα που αντιστοιχούν στις νόμιμες και δόλιες συναλλαγές ("isFraud" = 0 και "isFraud" = 1). Η μετρική K-S d-statistic παίρνει τιμές από 0, όταν τα δύο δείγματα ακολουθούν την ίδια κατανομή, έως 1 όταν οι κατανομές τους διαφέρουν κατά πολύ. Στην περίπτωτη της μεταβλητής "TransactionAmt" η μετρική d-statistic βρέθηκε να έχει τιμή ίση με 0,155, γεγονός που υποδηλώνει ότι τα δύο δείγματα εμφανίζουν καλή προσαμογή.

```
KS D statistic for TransactionAmt: 0.15551078394078205
```



Εικόνα 27 - Τέστ Kolmogorov-Smirnov

Τελευταίο βήμα αποτελεί ο διαχωρισμός του τελικού συνόλου δεδομένων ("final_dataset") σε σύνολο εκπαίδευσης (X_train, y_train) και ελέγχου (X_test, y_test). Ο διαχωρισμός γίνεται σε αναλογία 4:1, δηλαδή το 80% των δεδομένων χρησιμοποιείται για την εκπαίδευση των αλγορίθμων και το 20% για τον έλεγχο της απόδοσης τους. Το σύνολο εκπαίδευσης, πλέον, αριθμεί 60.072, με το σύνολο ελέγχου να απαρτίζεται από 15.018 εγγραφές. Είναι σημαντικό, ο μέσος όρος της μεταβλήτης "isFraud" να είναι αντιπροσωπευτικός και στα δύο δείγματα.

Πράγματι, παρατηρώντας τη μέση τιμή της μεταβλητής για τα δύο σύνολα (train και test), διακρίνεται ότι οι τιμές είναι πολύ κοντά.

```
y_train_mean: 0.0885603941936343
y_test.mean: 0.09142362498335331
```

4.3 Διαχείριση Μη Ισορροπημένων Συνόλων Δεδομένων

Η πλειονότητα των πραγματικών προβλημάτων ταξινόμησης στον τομέα της εξόρυξης γνώσης έχει να κάνει με ανομοιογενή δεδομένα, όπου τα περιστατικά μιας συγκεκριμένης κλάσης αποτελούν το μεγαλύτερο ποσοστό επί του συνόλου. Ένα τέτοιο πρόβλημα είναι και η πρόβλεψη και εντοπισμός φαινομένων απάτης με χρήση πιστωτικών καρτών. Όπως έχει προαναφερθεί, τα περιστατικά απάτης αποτελούν τη συντριπτική μειοψηφία των συναλλαγών που γίνονται με χρήση πλαστικού χρήματος. Συγκεκριμένα, ακόμα και στο σύνολο δεδομένων που εξετάζεται στην παρούσα εργασία, παρατηρείται ότι η αναλογία των νόμιμων προς τις παράνομες συναλλαγές είναι σχεδον 9:1 (93,32% - 6,68%).

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης εμφανίζουν μειωμένη απόδοση σε τέτοιου είδους προβλήματα ταξινόμησης, καθότι τείνουν να ταξινομούν τις περιπτώσεις της μειοψηφίας (περιπτώσεις απάτης) στην κλάση της πλειοψηφίας (νόμιμες συναλλαγές). Ανακύπτουν, λοιπόν, ορισμένα πρόδηλα ερωτήματα ως προς:

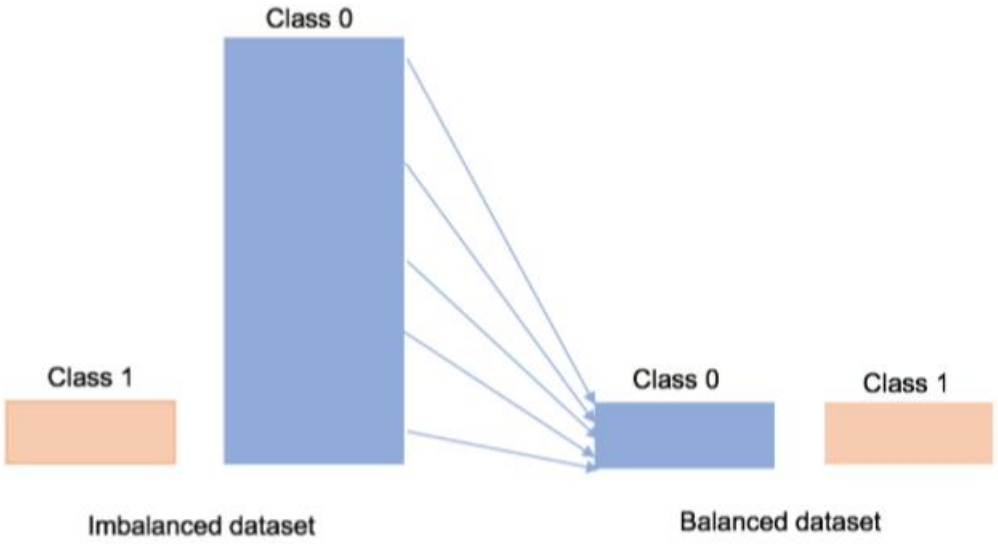
i. Τη διαχείριση των μη ισορροπημένων (ανομοιογενών) δεδομένων.
ii. Την επιλογή των μέτρων εκτίμησης που θα αξιολογηθούν.

4.3.1 Μέθοδοι επαναδειγματοληψίας

Για την εξισορρόπηση της αναλογίας των ανομοιγενών συνόλων δεδομένων έχει αναπτυχθεί μια σειρά από τεχνικές επαναδειγματοληψίας. Οι τρεις γνωστότερες τεχνικές που θα εξετασθούν στην παρούσα εργασία είναι η τυχαία υποδειγματοληψία (Random Undersampling), η τυχαία υπερδειγματοληψία (Random Oversampling) και η τεχνική υπερ-δειγματοληψίας με συνθετική μειονότητα (Synthetic Minority Oversampling Technique [SMOTE]). Μετέπειτα θα ακολουθήσει η σύγκριση των αποτελεσμάτων τους με αυτά που προέκυψαν χωρίς τη διαδικασία επαναδειγματοληψίας.

Τυχαία υποδειγματοληψία

Η τεχνική της τυχαίας υποδειγματοληψίας περιορίζει το μέγεθος των δειγμάτων της τάξης πλειοψηφίας στο μέγεθος των δειγμάτων μειοψηφίας. Η διαδικασία αφαίρεσης παρατηρήσεων από το μέγεθος πλειοψηφίας λαμβάνει χώρα με τυχαίο τρόπο και είναι κατάλληλη για σύνολα δεδομένων πολύ μεγάλου όγκου, αφού περιορίζει δραστικά τον χρόνο εκτέλεσης των αλγορίθμων. Ωστόσο, η τυχαία αυτή αφαίρεση παρατηρήσεων έχει ως συνέπεια να χάνεται χρήσιμη πληροφορία, επηρεάζοντας κατά αυτόν τον τρόπο αρνητικά την απόδοση των αλγορίθμων.



Εικόνα 28 - Τυχαία Υποδειγματοληψία

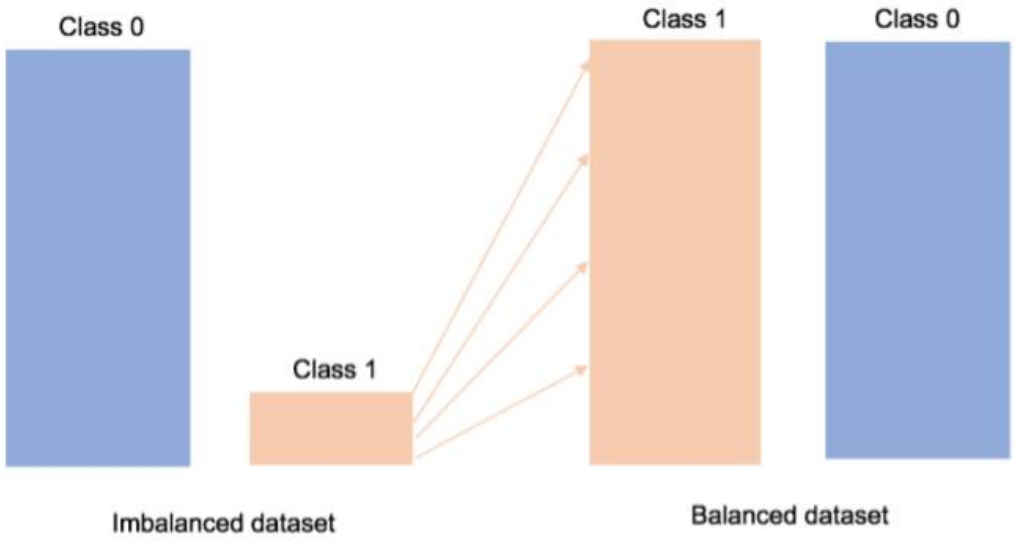
Εκ των 60.072 συναλλαγών που απαρτίζουν το σύνολο εκπαίδευσης, οι 5.320 αφορούν συναλλαγές που είναι προιόν απάτης, με τις υπόλοιπες 54.752 να έχουν να κάνουν με νόμιμες. Κατόπιν εφαρμογής της μεθόδου τυχαίας δειγματοληψίας, το νέο σύνολο εκπαίδευσης αριθμεί συνολικά 10.640 συναλλαγές, με την αναλογία νόμιμων-δόλιων συναλλαγών να είναι πλέον 1:1.

```
y_rs.value_counts()
```

```
1.0    5320
0.0    5320
Name: isFraud, dtype: int64
```

Τυχαία υπερδειγματοληψία

Η τεχνική της τυχαίας υπερδειγματοληψίας έχει τον ακριβώς αντίθετο τρόπο λειτουργίας από την παραπάνω μέθοδο. Χρησιμοποιεί τα δεδομένα της τάξης μειοψηφίας, επαναλαμβάνοντας με τυχαίο τρόπο τις παρατηρήσεις, έτσι ώστε ο αριθμός των παρατηρήσεων της μειοψηφικής κλάσης να ισούται με τον αριθμό των παρατηρήσεων της πλειοψηφικής κλάσης. Το βασικό μειονέκτημα της μεθόδου είναι η πιθανή εμφάνιση του φαινομένου της υπερπροσαρμογής (overfitting), από τη στιγμή που επαναλαμβάνει τα δεδομένα της μειοψηφικής κλάσης.



Εικόνα 29 - Τυχαία Υπερδειγματοληψία

Με την εφαρμογή της μεθόδου τυχαίας υπερδειγματοληψίας στο τελικό σύνολο δεδομένων ο αριθμός των δόλιων συναλλαγών αυξάνεται από 5.320 σε 54.752, όσο δηλαδή και το πλήθος των νόμιμων. Το τελικό, πλέον, σύνολο εκπαίδευσης ανέρχεται σε 109.504 εγγραφές με την αναλογία νόμιμων δόλιων να είναι και εδώ 1:1.

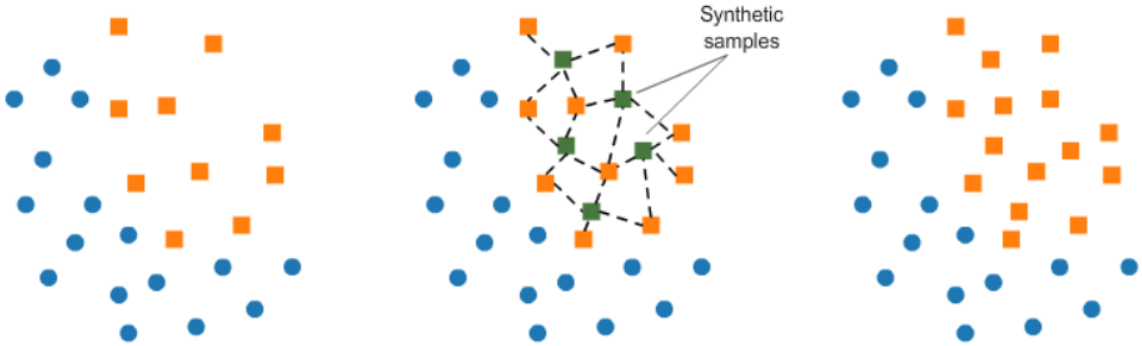
```
y_ran.value_counts()
```

```
1.0    54752
0.0    54752
Name: isFraud, dtype: int64
```

Υπερδειγματοληψία με Συνθετική Μειονότητα (SMOTE)

Η τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας (Synthetic Minority Oversampling Technique [SMOTE]) πρόκειται για μια μέθοδο υπερδειγματοληψίας που προσθέτει νέα τεχνητά μειονοτικά δεδομένα, κατόπιν κατάλληλης επεξεργασίας των υπαρχόντων (Chawla et

al., 2002) [24]. Η τεχνική αυτή υπολογίζει τους K-πλησιέστερους γείτονες της μειονοτικής τάξης για κάθε μειονοτική εγγραφή και δημιουργεί νέες τεχνητές εγγραφές με τα στοιχεία των ανωτέρω. Συνεπώς αυξάνει, όχι μόνο το μέγεθος του συνόλου εκπαίδευσης, αλλά και την ποικιλία των δεδομένων, επιλύοντας το πιθανό πρόβλημα της υπερπροσαρμογής που παρουσιάζεται με την τυχαία υπερδειγματοληψία.



Εικόνα 30 - Υπερδειγματοληψία με Συνθετική Μειονότητα

Όπως και στην περίπτωση της τυχαίας υπερδειγματοληψίας, έτσι και εδώ ο αριθμός των δόλιων συναλλαγών αυξάνεται σε 54.752, περιλαμβάνοντας ωστόσο ορισμένα νέα συνθετικά δεδομένα της μειονοτικής τάξης. Το τελικό σύνολο εκπαίδευσης αριθμεί 109.504 εγγραφές με την αναλογία νόμιμων-παράνομων να είναι και πάλι 1:1.

```
y_sm.value_counts()
```

```
1.0    54752
0.0    54752
Name: isFraud, dtype: int64
```

4.3.2 Μετρικές αξιολόγησης

Εξίσου σημαντική με την εξισορρόπηση των δεδομένων, είναι η επιλογή των μετρικών που πρέπει να ελεγχθούν, ώστε να εκτιμηθεί η απόδοση των αλγορίθμων. Ένα βασικό κριτήριο στα προβλήματα ταξινόμησης είναι η ακρίβεια (**accuracy**). Η ακρίβεια ορίζεται ως το ποσοστό των σωστά ταξινομημένων παρατηρήσεων στο σύνολο της πρόβλεψης. Αξίζει να τονιστεί ότι δεν αποτελεί πάντοτε χρήσιμο κριτήριο για την αξιολόγηση της απόδοσης ενός αλγορίθμου, και ιδιαίτερα σε προβλήματα με δεδομένα που βρίσκονται σε ανισορροπία. Ενδεικτικό παράδειγμα είναι η πρόβλεψη των φαινομένων απάτης, που αποτελεί και το ζητούμενο της παρούσας εργασίας. Έστω ότι πραγματοποιούνται 1000 συναλλαγές με τις 10 εκ των οποίων να είναι

προιόν απάτης. Ένας ταξινομητής που θα προβλέπει όλες τις συναλλαγές ως νόμιμες, θα εμφανίζεται να έχει ακρίβεια (accuracy) 99%. Δεν θα έχει καταφέρει, όμως, να προβλέψει ούτε μία από τις 10 συναλλαγές που ήταν παράνομες. Συνεπώς, παρά την ικανοποιητική ακρίβεια που έχει επιτύχει, ο αλγόριθμος αυτός θα είναι αναποτελεσματικός.

Δύο εξαιρετικά χρήσιμες μετρικές για προβλήματα μη ισορροπημένων δεδομένων δυαδικής ταξινόμησης είναι η ευαισθησία ή αλλιώς ανάκληση (**recall**) και η ακρίβεια πρόβλεψης (**precision**). Η ευαισθησία εκφράζει το ποσοστό των πραγματικά θετικών που έχουν αναγνωριστεί σωστά ως θετικά ενώ η ακρίβεια (precision) εκφράζει το ποσοστό των ορθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών.

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

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

Στον τραπεζικό κλάδο, και ιδιαίτερα σε ένα σύστημα πρόβλεψης και εντοπισμού ύποπτων συναλλαγών, τα δύο παραπάνω μεγέθη έχουν εξέχουσα σημασία. Τα χρηματοπιστωτικά ιδρύματα έχουν ως πρωταρχικό στόχο των ακριβή εντοπισμό όλων των συναλλαγών που είναι προιόν απάτης. Συνεπώς, παρακολουθώντας την ευαισθησία των αλγορίθμων, εντοπίζουν το ποσοστό των, εντοπισμένων από το σύστημα, παράνομων συναλλαγών, επί του συνόλου αυτών. Όσο μεγαλύτερη η τιμή της ευαισθησίας, τόσο πιο αξιόπιστο είναι το σύστημα εντοπισμού της απάτης.

Ταυτόχρονα, άλλο ένα μείζον ζήτημα στον τομέα της τραπεζικής απάτης είναι η εξυπηρέτηση/ικανοποίηση των πελατών (customer experience). Ως γνωστόν, όταν μια συναλλαγή υποδεικνύεται ως ύποπτη από τα συστήματα εντοπισμού απάτης, ακολουθεί μια σειρά διαδικασιών. Αρχικά, παγώνει η συναλλαγή και οι υπεύθυνοι των τμημάτων ασφαλείας της τράπεζας έρχονται σε επαφή με τον κάτοχο της κάρτας για την εξακρίβωση των στοιχείων τους. Μόνο ένα μικρό ποσοστό των συναλλαγών αυτών πρόκειται πραγματικά περί απάτης. Είναι εύλογο, λοιπόν, πως για την πλειονότητα των πελατών η διαδικασία αυτή επιφέρει δικαιολογημένα δυσαρέσκεια. Η ακρίβεια πρόβλεψης έρχεται να δώσει λύση στο ζήτημα αυτό, καθώς εκφράζει το ποσοστό των ορθώς θετικών προβλέψεων (απάτης) επί του συνόλου των προβλέψεων που θεωρήθηκαν παράνομες. Όσο μεγαλύτερη τιμή λαμβάνει η ακρίβεια πρόβλεψης, τόσο πιο στοχευμένα εντοπίζονται οι περιπτώσεις απάτης, αποφεύγοντας έτσι την άσκοπη αναστάτωση των πελατών.

Επιπροσθέτως, ένα ακόμα στατιστικό εργαλείο είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curves – **ROC** curves), οι οποίες χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός συστήματος δυαδικής ταξινόμησης. Η καμπύλη ROC εμφανίζεται ως το γράφημα της ευαισθησίας έναντι του 1-ειδικότητα για όλες τις πιθανές τιμές αποκοπής. Το εμβαδόν κάτω από την καμπύλη ROC (**Area Under the Curve**) χρησιμοποιείται ως δείκτης ακρίβειας ενός ελέγχου, και είναι χρήσιμο ως περιγραφικό μέτρο της συνολικής απόδοσης ενός ταξινομητή. Το εμβαδόν κάτω από την καμπύλη ROC (AUC) συμβολίζει την πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό υπόδειγμα υψηλότερα από ένα τυχαίο επιλεγμένο αρνητικό υπόδειγμα.

Τέλος, το εμβαδόν κάτω από την καμπύλη ανάκλησης-ακρίβειας (**Area Under Precision Recall Curve**) αποτελεί ένα ακόμα κρίσιμο μέτρο αξιολόγησης μοντέλων πρόβλεψης, ειδικά όταν υπάρχει μη ισορροπημένη κατανομή κλάσης.

4.4 Αλγόριθμος LightGBM

4.4.1 Εισαγωγή στον LightGBM

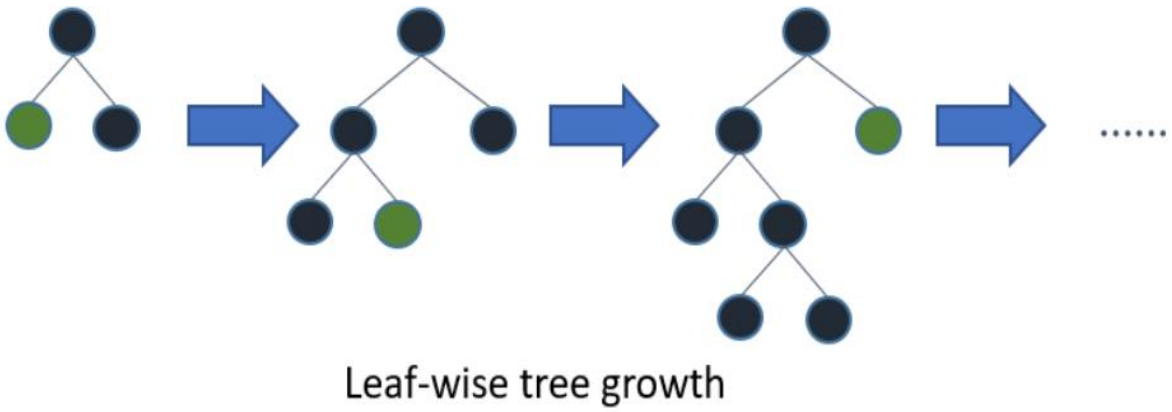
Ο αλγόριθμος LightGBM ανήκει στην οικογένεια των Gradient Boosting αλγορίθμων, όπου η εκμάθηση βασίζεται σε δέντρα. Είναι ιδαίτερα δημοφιλής λόγω της ταχύτητας και της ακρίβειας του.

Το Gradient boosting αποτελεί μια τεχνική μηχανικής μάθησης για προβλήματα παλινδρόμησης και κατηγοριοποίησης, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου «αδύναμων» μοντέλων πρόβλεψης, ως επί το πλείστον δέντρων αποφάσεων. Χτίζει το μοντέλο με έναν εξελικτικό τρόπο, όπως και άλλες μέθοδοι boosting, και το γενικεύει επιτρέποντας τη βελτιστοποίηση μιας αυθαίρετης διαφοροποιήσιμης συνάρτησης απώλειας. [26].

Η τεχνική Gradient Boosting προτάθηκε από τον Friedman (2001) και αποτελεί εξέλιξη της τεχνικής Boosting [27]. Αποτελεί συνδυασμό του αλγορίθμου Gradient Descent και της τεχνικής Boosting. Ο Friedman πρότεινε η εκπαίδευση των δέντρων, να γίνεται στην αρνητική παράγωγο της συνάρτησης κόστους και όχι στα κατάλοιπα των προηγούμενων γύρων, όπως συνέβαινε στην τεχνική Boosting. Ουσιαστικά, η ειδοποιός διαφορά έγκειται στο γεγονός ότι στην τεχνική Gradient Boosting παρέχεται η δυνατότητα επιλογής διαφορετικών συναρτήσεων κόστους. Ανάλογα με τη φύση των δεδομένων επιλέγονται και διαφορετικές.

Ο αλγόριθμος LightGBM δεν χρησιμοποιεί το σύνολο των δεδομένων εκπαίδευσης, παρά μόνο ένα δείγμα του, που προκύπτει βάσει της Gradient-One-Sampling (GOSS) μεθόδου. Η ιδέα πίσω από αυτό το σκεπτικό είναι πως όλες οι παρατηρήσεις δεν συνεισφέρουν το ίδιο στην εκπαίδευση του αλγορίθμου, καθώς όσες έχουν μικρή πρώτη παράγωγο συνάρτησης κόστους είναι καλύτερα εκπαιδευμένες από εκείνες που έχουν μεγάλη.

Ένα επιπλέον σημαντικό χαρακτηριστικό είναι ο τρόπος ανάπτυξης των δέντρων. Ο LightGBM αναπτύσσεται δενδροειδώς κατακόρυφα, δηλαδή σε επίπεδο φύλλων-κόμβων (leaf-wise). Επιλέγει το φύλλο με τη μέγιστη διαφορά συναρτήσεως κόστους για να αναπτυχθεί, αναζητώντας ουσιαστικά τον βέλτιστο διαχωρισμό σε κάθε κόμβο ανεξαρτήτως επιπέδου.



Εικόνα 31 - Ανάπτυξη Δέντρου ανά Τερματικό Κόμβο

Εν γένει, ο LightGBM χαρακτηρίζεται από μια σειρά πλεονεκτημάτων έναντι των κλασικών αλγορίθμων ταξινόμησης [28]. Παρακάτω παρατίθενται ορισμένα εξ΄αυτών:

- Υψηλή ταχύτητα εκπαίδευσης και απόδοσης. Οφείλεται κυρίως στη δημιουργία ιστογραμμάτων και στη χρήση των παραγόμενων κλάσεων, αντί όλου του εύρους τιμών κάθε μεταβλητής.
- Χαμηλή χρήση μνήμης. Οφείλεται στην αντικατάσταση των συνεχών τιμών με διακριτά πεδία.
- Υψηλότερη ακρίβεια από οποιονδήποτε άλλο αλγόριθμο gradient boosting. Είναι αποτέλεσμα της κατακόρυφα δενδροειδούς ανάπτυξης (leaf-wise). Χρήζει προσοχής το γεγονός ότι πολλές φορές μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting).

Ωστόσο, κατόπιν κατάλληλης ρύθμισης ορισμένων παραμέτρων (num_leaves, max_depth) μπορεί να αποφευχθεί.

- Συμβατότητα με μεγάλα σύνολα δεδομένων.
- Υποστήριξη παράλληλης μάθησης.

Αν και η εφαρμογή του αλγορίθμου LightGBM είναι σχετικά απλή, χρείαζεται μεγάλη προσοχή στην κατάλληλη ρύθμιση των παραμέτρων του, ώστε να αποφεύγονται προβήματα όπως αυτό της υπερπροσαρμογής που αναφέρθηκε παραπάνω. Το πλήθος των παραμέτρων του LightGBM ξεπερνά τους 100. Παρακάτω ακολουθεί μια συνοπτική περιγραφή των βασικότερων εξ αυτών.

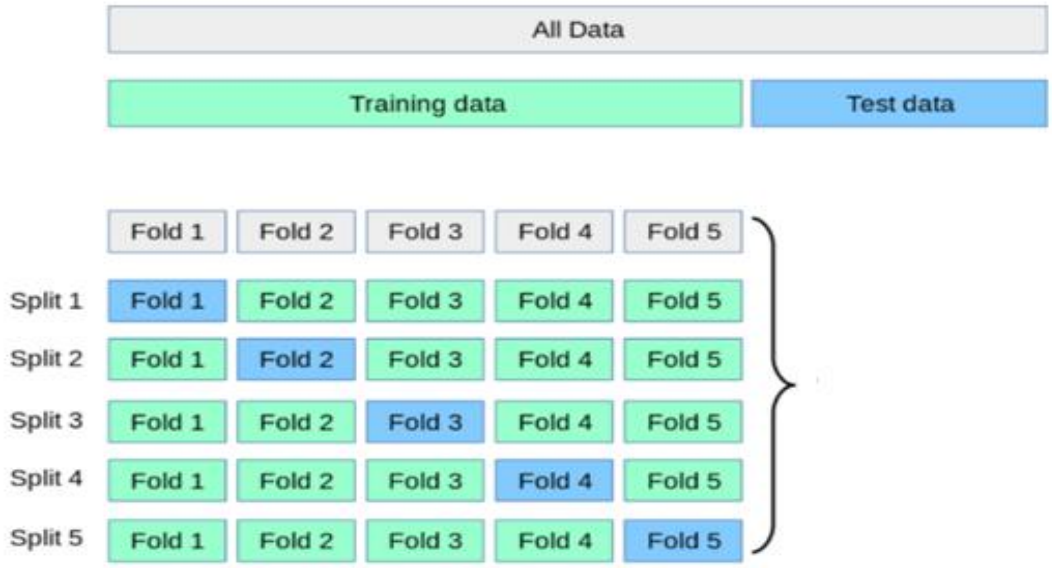
Βασικές Παράμετροι:

- *Task*: Καθορίζει την εργασία που θα εκτελεστεί στα δεδομένα. Μπορεί να είναι είτε εκπαίδευση είτε πρόβλεψη.
- *Application*: Καθορίζει πότε θα γίνει ταξινόμηση και πότε παλινδρόμηση. Ο LightGBM έχει ως προεπιλογή το μοντέλο παλινδρόμησης. Οι τιμές που παίρνει είναι:
  • binary: για δυαδική ταξινόμηση
  • multiclass: για πρόβληματα ταξινόμησης με πολλαπλών κλάσεων
  • regression: για παλινδρόμηση
- *Boosting*: Ορίζει τον τύπο του αλγόριθμου που θα εκτελεστεί. Οι τιμές που παίρνει είναι:
  • gbdt: Gradient Boosting Decision Tree
  • rf: random forest
  • goss: Gradient-based One-Side sampling
- *Data*: Τα δεδομένα από τα οποία θα εκπαιδευτεί ο αλγόριθμος.
- *Num_iterations*: Ορίζει τον αριθμό επαναλήψεων. Προκαθορισμένη τιμή: 100 επαναλήψεις
- *num_leaves*: Ορίζει τον αριθμό φύλλων-κόμβων ολόκληρου του δέντρου. Όσο μεγαλύτερο είναι το πλήθος των φύλλων του δέντρου, τόσο βελτιώνεται η ακρίβεια. Ελλοχεύει, ωστόσο, ο κίνδυνος της υπερπροσαρμογής.
- *learning_rate*: Προσδιορίζει τον ρυθμό εκμάθησης. Ανάλογα με την έξοδο του κάθε δέντρου, μεταβάλλεται η αρχική εκτίμηση. Το μέγεθος της μεταβολής αυτής εκφράζει το ρυθμό εκμάθησης του μοντέλου. Συνήθεις τιμές: 0.1, 0.001, 0.003.

- *Device*: Συσκευή στην οποία επιθυμεί ο χρήστης να «τρέξει» τον αλγόριθμο. Επιλογές: cpu, gpu. Προκαθορισμένη τιμή: cpu.
- *max_depth*: Περιγράφει το μέγιστο βάθος του δέντρου και ελέγχει την υπερπροσαρμογή του μοντέλου. Όσο μικρότερες τιμές παίρνει, τόσο μειώνεται η υπερπροσαρμογή.
- *min_data_in_leaf*: Είναι ο ελάχιστος αριθμός των εγγραφών που μπορεί να έχει ένα φύλλο και χρησιμοποιείται επίσης για την αντιμετώπιση της υπερπροσαρμογής.
- *feature_fraction*: Ελέγχει την υποδειγματοληψία των χαρακτηριστικών που χρησιμοποιούνται στην εκπαίδευση του μοντέλου. Για παράδειγμα, αν λαμβάνει τιμή ίση με 0.7, σημαίνει ότι ο LightGBM θα επιλέξει με τυχαίο τρόπο το 70% των μεταβλητών σε κάθε επανάληψη.
- *bagging_fraction*: Ορίζει το ποσοστό των δεδομένων που χρησιμοποιούνται σε κάθε επανάληψη. Συχνά, χρησιμοποιείται για την επιτάχυνση της εκπαίδευσης και την αποφυγή ενδεχόμενης υπερπροσαρμογής.
- *early_stopping_round*: Χρησιμοποιείται για τον πρόωρο τερματισμό της ανάλυσης. Το μοντέλο σταματά την εκπαίδευση εάν μια μέτρηση ενός συνόλου δεδομένων δεν βελτιώνεται στις τελευταίες επαναλήψεις που ορίζονται ως “early_stopping_round”.
- *Num_threads:* Ορίζει το πλήθος των threads που θα χρησιμοποιηθούν για την εκτέλεση του αλγορίθμου.
- *min_gain_to_split*: Χρησιμοποιείται για τον έλεγχο του αριθμού των διαχωρισμάτων του δέντρου.

4.4.2 Εφαρμογή LightGBM

Για την καλύτερη εκπαίδευση των αλγορίθμων στα δεδομένα της εργασίας, χρησιμοποείται η μέθοδος της k-διασταυρούμενης επικύρωσης. Με τη μέθοδο επικύρωσης k τμημάτων, το σύνολο εκπαίδευσης (“X_train”, “y_train”) διαιρείται σε k υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις και η επιλογή τους είναι τυχαία. Ο αλγόριθμος εκπαιδεύεται με τα δεδομένα των k-1 υποσυνόλων, ενώ το ένα χρησιμοποείται ως σύνολο επικύρωσης (validation). Η διαδικασία επαναλαμβάνεται k φορές, θέτοντας κάθε φορά διαφορετικό υποσύνολο επικύρωσης. Στο τέλος υπολογίζεται μέση επίδοση του μοντέλου. Για τα δεδομένα της εργασίας επιλέγεται ο αριθμός k των υποσυνόλων να είναι τα πέντε.



Εικόνα 32 - 5-Πτυχη Διασταυρούμενη Επικύρωση

Κατόπιν εκτελείται η εφαρμογή του αλγορίθμου LightGBM με την ακόλουθη εντολή:

```
clf = LGBMClassifier(boosting = 'gbdt',
                     #nthread = 4,
                     n_estimators = 500,
                     silent = -1,
                     verbose = -1,
                     max_debth = -1,
                     learning_rate = 0.016,
                     colsample_bytree = 0.25,
                     subsaple = 0.7,
                     reg_alpha = 0.2041,
                     min_split_gain = 0.2453,
                     random_seed = 40,
                     bagging_seed = 40,
                     num_leaves = 60)

clf.fit(train_x, train_y, eval_set = [(train_x,train_y),(valid_x, valid_y)],

eval_metric = 'auc', verbose = 100, early_stopping_rounds = 200)

y_preds[valid_idx] = clf.predict_proba(valid_x, num_iteration = clf.best_iteration_)[:,1]

y_oof += clf.predict_proba(X_test, num_iteration = clf.best_iteration_)[:,1]/folds.n_splits
```

Οι τιμές των παραμέτρων αποφασίστηκαν κατόπιν επαναλαμβανόμενων δοκιμών. Οι παραπάνω τιμές βρέθηκε να συμβάλλουν στην βέλτιστη απόδοση του μοντέλου. Επίσης τα αποτελέσματα των προβλέψεων αποθηκεύτηκαν στους πίνακες y_preds και y_oof. Ο πίνακας

“y_preds” (“Predictions” – 80%) περιέχει τις προβλέψεις του συνόλου επικύρωσης, ενώ ο πίνακας “y_oof” (“Out Of Folder” – 20%) περιέχει τις προβλέψεις για το σύνολο ελέγχου.

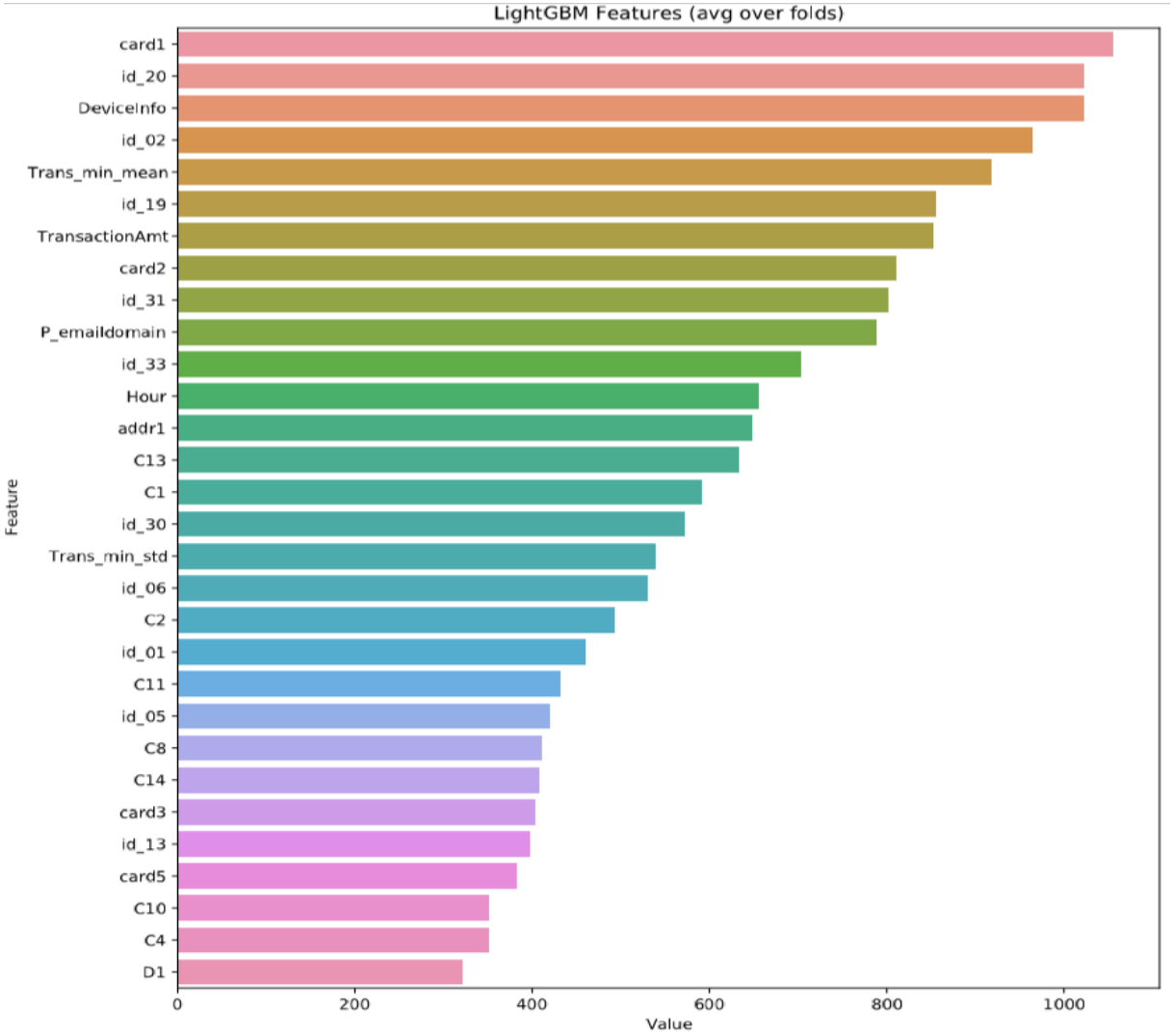
Ένα μέρος του αποτελέσματος της εκτέλεσης του αλγορίθμου παρουσιάζεται στην ακόλουθη εικόνα.

```
Training until validation scores don't improve for 200 rounds
[100]   training's auc: 0.966277        training's binary_logloss: 0.126583     valid_1's auc: 0.950449 valid_1's binary_logl
oss: 0.144446
[200]   training's auc: 0.978576        training's binary_logloss: 0.0949564    valid_1's auc: 0.958795 valid_1's binary_logl
oss: 0.120896
[300]   training's auc: 0.986342        training's binary_logloss: 0.0778101    valid_1's auc: 0.963866 valid_1's binary_logl
oss: 0.110811
[400]   training's auc: 0.991407        training's binary_logloss: 0.0655548    valid_1's auc: 0.967715 valid_1's binary_logl
oss: 0.104012
[500]   training's auc: 0.994364        training's binary_logloss: 0.0567506    valid_1's auc: 0.97007  valid_1's binary_logl
oss: 0.0994514
Did not meet early stopping. Best iteration is:
[500]   training's auc: 0.994364        training's binary_logloss: 0.0567506    valid_1's auc: 0.97007  valid_1's binary_logl
oss: 0.0994514
```

Εικόνα 33 - Τμήμα της Εξόδου του LightGBM

Επίσης μεγάλη αξία έχει να εντοπιστεί η σημαντικότητα των μεταβλητών. Με χρήση της παρακάτω εντολής, εντοπίζονται οι 30 πρώτες μεταβλητές με τη μεγαλύτερη συνεισφορά στον διαχωρισμό των δέντρων του αλγορίθμου.

```
feature_imp = pd.DataFrame(sorted(zip(clf.feature_importances_,X.columns)), columns = ['Value','Feature'])

plt.figure(figsize = (10, 10))
sns.barplot(x ="Value", y ="Feature", data = feature_imp.sort_values(by = "Value", ascending=False).head(30))
plt.title('LightGBM Features (avg over folds)')
plt.tight_layout()
plt.savefig('LGBM feature import.pdf', bbox_inches='tight')
plt.show()
```



Εικόνα 34 - Μέσες Τιμές 30 Σημαντικότερων Μεταβλητών LightGBM

Για την εξέταση της απόδοσης του αλγορίθμου ελέγχθηκαν τρείς μετρικές. Η περιοχή κάτω από την καμπύλη ROC (AUC), η ευαισθησία (Recall) και η ακρίβεια πρόβλεψης (Precision). Με χρήση των κάτωθι εντολών εμφανίζονται οι τιμές των παραπάνω μέτρων αξιολόγησης.

```
valid_auc_score = ('Valid AUC score %.6f' % roc_auc_score(y_train,y_preds))
print(valid_auc_score)
test_auc_score = ('Test AUC score %.6f' % roc_auc_score(y_test,y_oof))
print(test_auc_score)
valid_recall_score = ('Valid recall score %.6f' % recall_score(y_train,y_preds.round()))
print(valid_recall_score)
test_recall_score = ('Test recall score %.6f' % recall_score(y_test,y_oof.round()))
print(test_recall_score)
```

```
prec = precision_score(y_test,y_oof.round())
print("The precision is {}".format(prec))
```

```
Valid AUC score 0.971878
Test AUC score 0.971033
Valid recall score 0.703571
Test recall score 0.680991

The precision is 0.9060077519379846
```

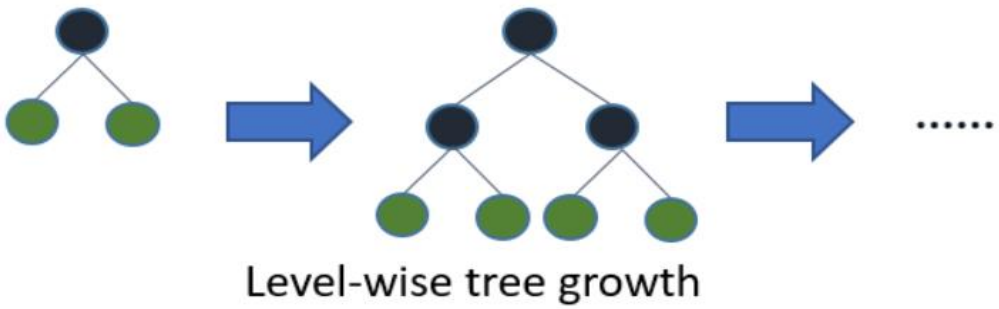
Από την παρατήρηση των αποτελεσμάτων προκύπτει ότι η τιμή της μετρικής AUC λαμβάνει αρκετά κοντινές τιμές για την πρόβλεψη τόσο στο σύνολο επικύρωσης (validation set), όσο και στο σύνολο ελέγχου (test set), σημειώνοντας τιμές 0,9719 και 0,9710 αντίστοιχα. Λαμβάνοντας υπόψη ότι ο αλγόριθμος έχει αρκετά υψηλή απόδοση και στα νέα δεδομένα (test set), συμπεραίνεται ότι αποφεύγεται η υπερπροσαρμογή.

4.5 Αλγόριθμος XGBoost

4.5.1 Εισαγωγή στον XGBoost

Ο XGBoost αποτελεί έναν επιπλέον αλγόριθμο που ανήκει στην οικογένεια των Gradient Boosting αλγορίθμων, όπου η εκμάθηση βασίζεται σε δέντρα. Είναι ιδιαίτερα δημοφιλής στο χώρο της εξόρυξης δεδομένων λόγω της υψηλής απόδοσης του, έναντι των κλασσικών αλγορίθμων.

Η τεχνική eXtreme Gradient Boosting προτάθηκε από τον Chen (2014). Όπως προαναφέρθηκε, η τεχνική Gradient Boosting προσθέτει διαδοχικά δέντρα σε κάθε χρονική στιγμή t στην αρνητική παράγωγο της συνάρτησης κόστους. Η σημαντική διαφορά σε σχέση με τον LightGBM είναι ο τρόπος της δενδροειδούς ανάπτυξης, ο οποίος είναι οριζόντιος. Πρακτικά αυτό σημαίνει ότι ο XGBoost αναπτύσσεται σε επίπεδα βάθους (level-wise). Αναζητά τον βέλτιστο διαχωρισμό, σαρώνοντας το σύνολο των δεδομένων και διατηρώντας τον αριθμό των φύλλων σε κάθε επίπεδο του δέντρου σταθερό.



Εικόνα 35 - Ανάπτυξη Δέντρου ανά Επίπεδα

Παρακάτω παρατίθενται ορισμένα από τα βασικότερα πλεονεκτήματα του αλγορίθμου XGBoost:

- Κανονικοποίηση δεδομένων.
- Παράλληλη επεξεργασία δεδομένων. Υποστηρίζει την εφαρμογή του σε κατανεμημένα περιβάλλοντα, όπως αυτό του Hadoop.
- Υψηλή ευελιξία. Επιτρέπει στον χρήστη να εφαρμόζει τεχνικές βελτιστοποίησης και κριτηρίων αξιολόγησης.
- Διαχείριση ελλειπουσών τιμών. Ο χρήστης αντικαθιστά τις ελλείπουσες τιμές με μια συγκεκριμένη τιμή, θέτοντας την ταυτόχρονα ως παράμετρο στο μοντέλο. Ο XGBoost εντοπίζει την τιμή αυτή και μαθαίνει πως να διαχειρίζεται τις επόμενες ελλείπουσες τιμές που θα εντοπίσει.
- Ενσωματωμένη διασταυρούμενη επικύρωση. Επιτρέπει στο χρήστη να τρέξει μια διασταυρούμενη επικύρωση σε κάθε επανάληψη και έτσι είναι εύκολο να πάρει τον ακριβή βέλτιστο αριθμό των επαναλήψεων σε ένα μόνο «τρέξιμο» του μοντέλου.

Η άρτια εφαρμογή του XGBoost συνοδεύεται από την κατάλληλη ρύθμιση των παραμέτρων του, έτσι ώστε να επιτευχθεί η βέλτιστη απόδοση. Οι παράμετροι του αλγορίθμου χωρίζονται σε τρείς κατηγορίες: τις γενικές, τις μετρικές και τις παραμέτρους ενίσχυσης.

Γενικές Παράμετροι:

- *Booster*:
  • gbtree: για δενδροειδή μοντέλα
  • gblinear: για γραμμικά μοντέλα
- *Silent*: Παίρνει τιμή 1 όταν ο χρήστης δεν επιθυμεί να εμφανίζονται μηνύματα κατά την εκτέλεση της εντολής.
- *nthread*: Χρησιμοποιείται όταν λαμβάνει χώρα παράλληλη επεξεργασία και ορίζει των αριθμό των πυρήνων που θα λάβουν μέρος στην εκτέλεση της εντολής. Αν δεν λάβει τιμή, τότε εξ ορισμού θα χρησιμοποιήθούν όλοι οι διαθέσιμοι πυρήνες του υπολογιστικού συστήματος.

Παράμετροι Ενίσχυσης:

- *eta*: Είναι μέγεθος αντίστοιχο του ρυθμόυ εκμάθησης των αλγορίθμωνς gradient boosting. Οι τυπικές τιμές του κυμάινονται από 0,01 έως 0,2.

- *min_child_weight*: Ορίζει τον ελάχιστο άθροισμα των βαρών όλων των παρατηρήσεων κάθε φύλλου. Χρησιμοποιείται στον έλεγχο της υπερπροσαρμογής. Οι υψηλές τιμές εμποδίζουν τον αλγόριθμο να απομνημονεύει συσχετίσεις που πολλές φορές οδηγούν σε υπερπροσαρμογή.
- *max_leaf_nodes*: Ορίζει τον μέγιστο αριθμό τερματικών κόμβων ή φύλλων ενός δέντρου.
- *gamma*: Καθορίζει την ελάχιστη μείωση της απώλειας που απαιτείται για να γίνει ένας διαχωρισμός. Επί της ουσίας διαμορφώνει την πολυπλοκότητα των δέντρων. Μεγάλες τιμές του γ οδηγούν σε μικρά δέντρα, αντίστοιχα μικρές τιμές οδηγούν σε μεγάλα δέντρα.
- *lambda*: Ρυθμίζει κατά πόσο θα συρρικνώνονται τα βάρη του δέντρου. Όσο αυξάνεται η τιμή της, τα βάρη του δέντρου συρρικώνονται.

Μετρικές Παράμετροι:

- *Mae*: μέσο απόλυτο σφάλμα (mean absolute error)
- *Rmse*: τετραγωνική ρίζα μέσου τετραγωνικό σφάλμα (root mean squared error)
- *logloss*: απώλεια για δυαδική ταξινόμηση
- *mlogloss*: απώλεια για πολλαπλή ταξινόμηση
- *error*: Σφάλμα δυαδικής ταξινόμησης (όριο: 0,5)

4.5.2 Εφαρμογή XGBoost

Για την εφαρμογή του αλγορίθμου XGBoost στα δεδομένα της εργασίας ακολουθείται παρόμοια διαδικασία με αυτή του LightGBM. Δημιουργούνται και πάλι δύο πίνακες με ονομασία “y_preds1” και “y_oof1” που αποθηκεύουν τις προβλέψεις των συνόλων επικύρωσης (validation) και ελέγχου (test) αντίστοιχα. Επίσης, η τιμές των παραμέτρων ανατέθηκαν κατόπιν της διαδικασίας πειραματισμού και σφάλματος (trial and error). Η εντολή που εκτελέστηκε για την εφαρμογή του XGBoost φαίνεται παρακάτω.

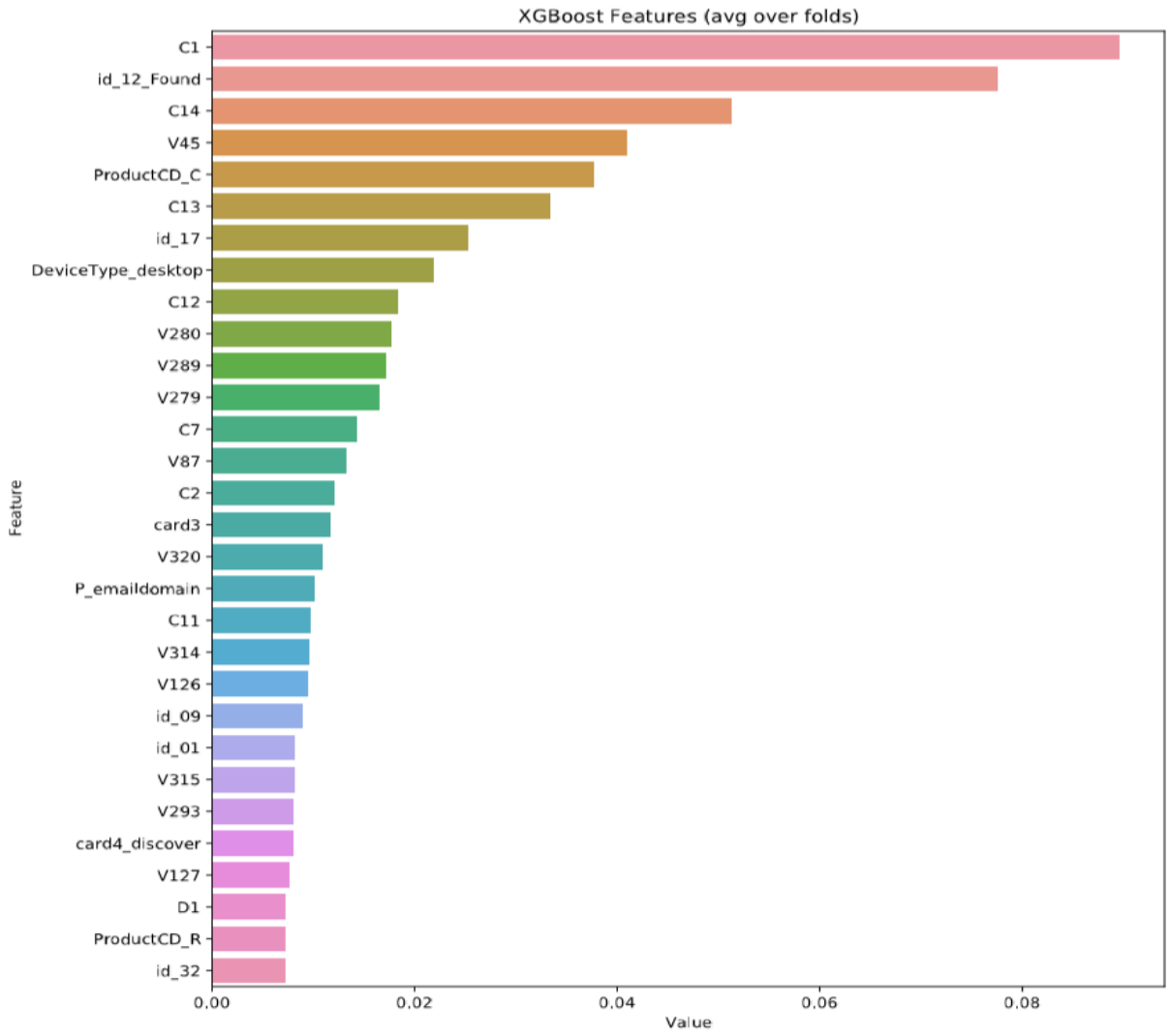
```
clf1 = XGBClassifier( base_score=0.5,
                      booster='gbtree',
                      colsample_bylevel=1,
                      colsample_bynode=1,
                      colsample_bytree=1,
                      gamma=0,
                      learning_rate=0.1,
                      max_delta_step=0,
                      max_depth=3,
                      min_child_weight=1,
                      missing=None,
                      n_estimators=500,
                      n_jobs=1,
                      nthread=4,
                      objective='binary:logistic',
                      random_state=0,
                      reg_alpha=0,
                      reg_lambda=1,
                      scale_pos_weight=1,
                      seed=None,
                      silent=None,
                      subsample=1,
                      verbosity=1
                    )

clf1.fit(train_x, train_y, eval_set = [(train_x,train_y),(valid_x, valid_y)],

eval_metric = 'auc', verbose = 100, early_stopping_rounds = 200)

y_preds1[valid_idx] = clf1.predict_proba(valid_x)[:,1]

y_oof1 += clf1.predict_proba(X_test)[:,1]/folds.n_splits
```

Ένα μέρος του αποτελέσματος της εκτέλεσης του αλγορίθμου XGBoost παρουσιάζεται στην ακόλουθη εικόνα.

```
Will train until validation_1-auc hasn't improved in 200 rounds.
[100]   validation_0-auc:0.95275        validation_1-auc:0.95250
[200]   validation_0-auc:0.96371        validation_1-auc:0.96184
[300]   validation_0-auc:0.97007        validation_1-auc:0.96583
[400]   validation_0-auc:0.97446        validation_1-auc:0.96839
[499]   validation_0-auc:0.97702        validation_1-auc:0.96911
[0]     validation_0-auc:0.82776        validation_1-auc:0.81081
Multiple eval metrics have been passed: 'validation_1-auc' will be used for early stopping.
```

Εικόνα 36 - Τμήμα της Εξόδου του XGBoost

Οι 30 μεταβλητές που συνεισφέρουν περισσότερο στο διαχωρισμό των δέντρων και συνεπώς στην πρόβλεψη, με χρήση του XGBoost, παρουσίαζονται στην ακόλουθη εικόνα.



Εικόνα 37 - Μέσες Τιμές 30 Σημαντικότερων Μεταβλητών XGBoost

Και εδώ, όπως και στα αποτελέσματα του αλγορίθμου LightGBM, ελέγχονται οι μετρικές αξιολόγησης AUC, η ευαισθησία (Recall), και η ακρίβεια πρόβλεψης (Precision).

```
Valid AUC score 0.965818
Test AUC score 0.963264
Valid recall score 0.678759
Test recall score 0.656956

The precision is 0.9010989010989011
```

Παρατηρείται ότι η περιοχή κάτω από την καμπύλη ROC σημειώνει, και με τον αλγόριθμο XGBoost, υψηλή τιμή, ελαφρώς χαμηλότερη από αυτή που σημείωσε με τον LightGBM. Επίσης, αποφεύγεται το φαινόμενο της υπερπροσαρμογής, από τη στιγμή που οι τιμές της AUC τόσο για το σύνολο επικύρωσης (Valid), όσο και για το σύνολο ελέγχου (Test) είναι αρκετά κοντινές.

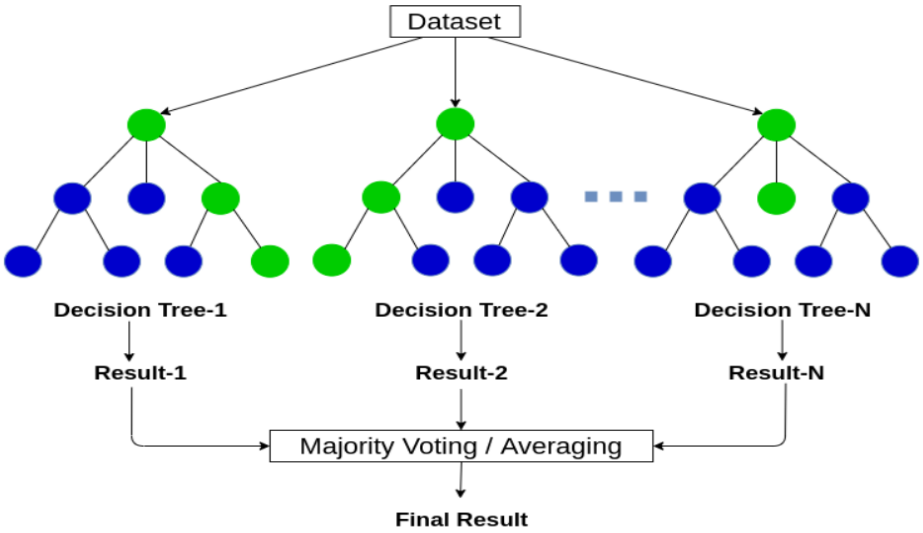
4.6 Αλγόριθμος Random Forest

4.6.1 Εισαγωγή στον Random Forest

Τα τυχαία δάση αποφάσεων, μέθοδος γνωστή με την ονομασία “Random Forest”, αποτελούν μια μέθοδο κατηγοριοποίησης συνεχών και μη συνεχών μεταβλητών, με βασικό άξονα λειτουργίας τους να είναι η κατασκευή πολλαπλών δέντρων αποφάσεων. Ο πρώτος αλγόριθμος τύπου “Random Forest” δημιουργήθηκε από τον T.Ho, ενώ η τελική τους ονομασία δόθηκε από τους Leo Breiman και Adele Cutler [29].

Αρχικά δημιουργούνται επιμέρους δέντρα απόφασης, το καθένα εκ των οποίων ψηφίζει για την κλάση που ανήκει η παρατήρηση προς ταξινόμηση. Η κλάση με τους περισσότερους ψήφους, είναι και αυτή που εν τέλει επικρατεί στην πρόβλεψη. Ο συνδυασμός των επιμέρους δέντρων απόφασης λειτουργεί ως ένα είδος boosting, επιφέροντας πιο ακριβείς κατηγοριοποιήσεις σε σύγκριση με τους μεμονωμένους ταξινομητές.

Τα δέντρα απόφασης που απαρτίζουν ένα τυχαίο δάσος δεν προσομοιάζουν στα τυπικά δέντρα απόφασης. Στη μέθοδο των τυχαίων δασών, ανατίθενται στη ρίζα των δέντρων το σύνολο των δειγμάτων εκπαίδευσης. Σε κάθε κόμβο χρησιμοποιείται ένα τυχαίο υποσύνολο των δειγμάτων, έως ότου κατασκευαστεί το δέντρο. Έπειτα, λαμβάνει χώρα ο έλεγχος, ο οποίος αφορά ένα υποσύνολο των χαρακτηριστικών των δειγμάτων εκπαίδευσης. Η διαδικασία αυτή επαναλαμβάνεται για τον αριθμό των δέντρων που έχουν οριστεί εξ αρχής. Αξίζει επίσης να τονιστεί ότι τα δέντρα δεν υφίστανται αποκοπή κόμβων - κλάδεμα (prooning), ενώ για την επιλογή του καλύτερου διαχωρισμού χρησιμοποιούνται κατάλληλα μέτρα όπως ο δείκτης Gini ή η εντροπία.



Εικόνα 38 - Τρόπος Ανάπτυξης «Τυχαίων Δασών»

Παρακάτω αναφέρονται τα κύρια πλεονεκτήματα των “Random Forest” αλγορίθμων:

- Αποτελεσματική λειτουγία σε μεγάλες βάσεις δεδομένων και δεδομένα υψηλών διαστάσεων.
- Δημιουργία αμερόληπτης εκτίμησης σφάλματος γενίκευσης κατά τη διαδικασία κατασκευής του δέντρου.
- Ανεκτικότητα στο θόρυβο και στα αριθμητικά σφάλματα (ακραίες - ελλιπείς τιμές).
- Περιορισμένο σφάλμα, που οφείλεται στην ανάπτυξη πολύ μεγάλου αριθμού δέντρων Ως αποτέλεσμα η εμφάνιση του φαινομένου υπερ-εκπαίδευσης (over fitting) είναι σχεδόν απίθανη.
- Χρήση πρωτοτύπων και μετρικών που εκφράζουν συσχετίσεις μεταξύ των μεταβλητών και της κατηγοριοποίησης.
- Πεπερασμένη διαδικασία εκπαίδευσης, από τη στιγμή που ο αλγόριθμος ολοκληρώνεται σε σταθερό αριθμό βημάτων.
- Ελάττωση της συσχέτισης ανάμεσα στα δέντρα και διατήρηση της πόλωσης (bias) σε σχετικά χαμηλά επίπεδα.

Σημαντικό ρόλο διαδραματίζει η ανάθεση τιμών στις παραμέτρους του αλγορίθμου, με σκοπό την επίτευξη της βέλτιστης ακρίβειας στην πρόβλεψη. Ενδεικτικά, ακολουθεί η περιγραφή των σημαντικότερων παραμέτρων του αλγόριθμου “Random Forest”:

- *n_estimators*: Ορίζει το πλήθος των δέντρων που απαρτίζουν το δάσος
- *criterion*: Εκφράζει τη μετρική βάση της οποίας θα ελεγχθεί η ποιότητα του διαχωρισμού. Ενδεικτικά μεγέθη είναι το κριτήριο gini, ή η εντροπία.
- *max_depth*: Εκφράζει το μέγιστο «βάθος» των επιμέρους δέντρων.
- *min_samples_split*: Ορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για να γίνει ένας διαχωρισμός.
- *min_samples_leaf*: Καθορίζει τον ελάχιστο αριθμό δειγμάτων που χρειάζεται να υπάρχουν σε κάθε φύλλο.
- *min_weight_fraction_leaf*: Αναφέρεται στο ελάχιστο σταθμισμένο κλάσμα του συνόλου των βαρών (όλων των δειγμάτων εισόδου) που απαιτείται να βρίσκεται σε κάθε κόμβο φύλλων. Τα δείγματα έχουν το ίδιο βάρος όταν δεν δίνεται τιμή για την παράμετρο.
- *max_features*: Εκφράζει τον αριθμό των μεταβλητών που πρέπει να λαμβάνονται υπόψη κατά της αναζήτηση του βέλτιστου διαχωρισμού.

4.6.2 Εφαρμογή Random Forest

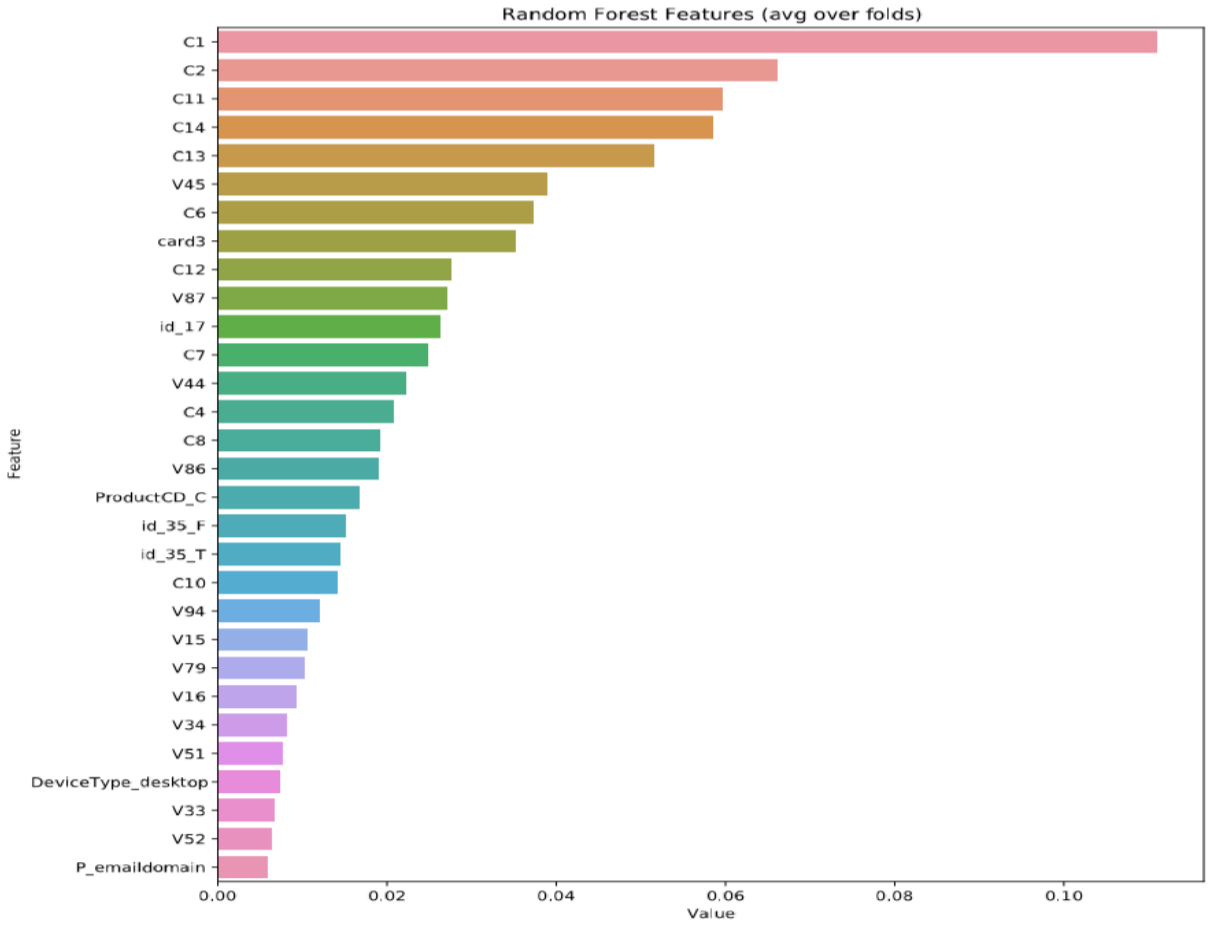
Για την εφαρμογή του "Random Forest" στα δεδομένα της εργασίας χρησιμοποιήθηκε και πάλι η μέθοδος της διασταυρούμενης επικύρωσης, ενώ οι τιμές των παραμέτρων προέκυψαν κατόπιν συνεχών δοκιμών και ελέγχων. Τα αποτελέσματα των προβλέψεων, τόσο του συνόλου επικύρωσης (validation set), όσο και του συνόλου ελέγχου (test set), αποθηκεύτηκαν στους κενούς πίνακες "y_preds2" και "y_oof2" αντίστοιχα, με τις εντολές που φαίνονται παρακάτω.

```
y_oof2 = np.zeros(X_test.shape[0])
y_preds2 = np.zeros(X_train.shape[0])

start_time = time.time()

folds = KFold(n_splits = 5, shuffle = True, random_state = 2)
for n_fold, (train_idx, valid_idx) in enumerate (folds.split(X_train, y_train)):
    train_x, train_y = X_train.iloc[train_idx], y_train.iloc[train_idx]
    valid_x, valid_y = X_train.iloc[valid_idx], y_train.iloc[valid_idx]
```

Η τελική εφαρμογή του Random Forest στα δεδομένα έγινε με την ακόλουθη εντολή:

```
clf2 = RandomForestClassifier( n_estimators=500,
                               criterion='gini',
                               max_depth=5,
                               min_samples_split=2,
                               min_samples_leaf=1,
                               min_weight_fraction_leaf=0.0,
                               max_features='auto',
                               max_leaf_nodes=None,
                               min_impurity_decrease=0.0,
                               min_impurity_split=None,
                               bootstrap=True,
                               oob_score=False,
                               n_jobs=-1,
                               random_state=0,
                               verbose=0,
                               warm_start=False,
                               class_weight='balanced'
                             )

clf2.fit(train_x, train_y)
y_preds2[valid_idx] = clf2.predict_proba(valid_x)[:,1]
y_oof2 += clf2.predict_proba(X_test)[:,1]/folds.n_splits
```

Παρατηρείται ότι οι μεταβλητές που συνεισφέρουν περισσότερο στον διαχωρισμό των επιμέρους δέντρων αποφάσεων του δάσους είναι εκείνες της κατηγορίας C (C1-C15).



Εικόνα 39 - Μέσες Τιμές 30 Σημαντικότερων Μεταβλητών Random Forest

Τα αποτελέσματα της εφαρμογής του αλγορίθμου των τυχαίων δασών ελέγχονται βάση των μετρικών AUC, της ευαισθησίας (Recall) και της ακρίβειας πρόβλεψης (Precision). Παρατηρώντας την τιμή της μετρικής AUC, ο Random Forest σημείωσε χαμηλότερη απόδοση σε σύγκριση με τους προηγούμενους δύο που δοκιμάστηκαν. Επίσης, αν και η ευαισθησία βρίσκεται να σημειώνει τιμή 0.78, η ακρίβεια πρόβλεψης λαμβάνει αρκετά χαμηλή τιμή, ίση με 0.41.

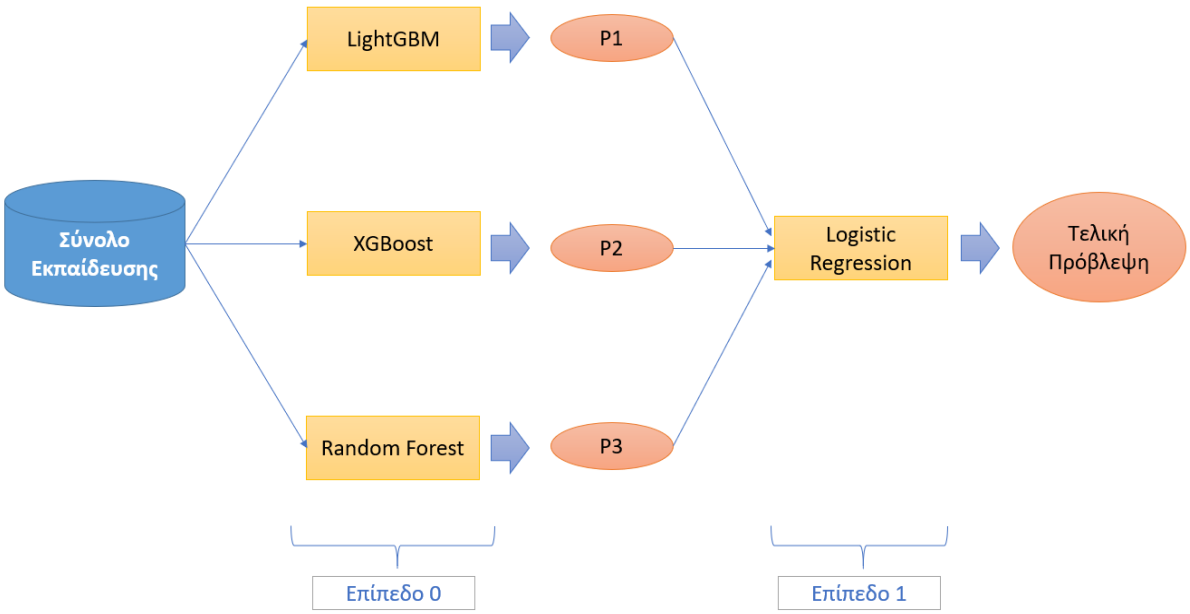
```
Valid AUC score 0.920036
Test AUC score 0.914754
Valid recall score 0.796992
Test recall score 0.778587

The precision is 0.4062262718299165
```

4.7 Συσσωρευμένη Γενίκευση με Χρήση Λογιστικής Παλινδρόμησης

Η συσσωρευμένη γενίκευση (stacked generalization), γνωστή και με την ονομασία stacking, αποτελεί μια μέθοδο συνδυασμού ταξινομητών με στόχο την επίτευξη μεγαλύτερης ακρίβειας πρόβλεψης. Διατυπώθηκε από τον Wolpert το 1992 και έκτοτε έχει εφαρμοστεί τόσο για προβλήματα κατηγοριοποίησης [30], όσο και για προβλήματα αριθμητικής πρόβλεψης-παλινδρόμησης (Breiman,1996) [31] και μη επιβλεπόμενης μάθησης (Smyth & Wolpert, 1997) [32].

Σε πρώτο στάδιο, ένα σύνολο βασικών ταξινομητών εκπαιδεύονται σε ένα πλήθος από υποσύνολα των αρχικών δεδομένων, παράγοντας αντίστοιχο αριθμό από μοντέλα. Οι ταξινομητές αυτοί ονομάζονται πρώτου επιπέδου (επίπεδο-0). Κάθε στιγμιότυπο αντιστοιχίζεται σε ένα νέο, που αναπαριστά την πρόβλεψη του εκάστοτε μοντέλου για το αρχικό στιγμιότυπο μαζί με την πραγματική τιμή. Στο σημείο αυτό πρέπει να διασφαλίζεται ότι τα μοντέλα που διαμορφώνονται δεν εμπεριέχουν το στιγμιότυπο για το οποίο κάνουν πρόβλεψη. Σε δεύτερο στάδιο, χρησιμοποιείται ένας έτερος ταξινομητής σε μετα-επίπεδο (επίπεδο-1), γνωστός και ως μετα-ταξινομητής, ο οποίος δέχεται ως είσοδο τις προβλέψεις των ταξινομητών του πρώτου επιπέδου.

Στην παρούσα εργασία θα χρησιμοποιηθούν ως ταξινομητές πρώτου επιπέδου οι αλγόριθμοι που αναφέρθηκαν στις προηγούμενες ενότητες (LightGBM, XGBoost, Random Forest), ενώ ως ταξινομητής δεύτερου επιπέδου θα χρησιμοποιηθεί η λογιστική πανλινδρόμηση.



Εικόνα 40 - Μέθοδος Συσσωρευμένης Γενίκευσης - Stacking

Το λογιστικό μοντέλο, ή αλλιώς η λογιστική παλινδρόμηση, χρησιμοποιείται συχνά, όταν η εξαρτημένη μεταβλητή ενός προβλήματος ταξινόμησης επιδέχεται μόνο δύο τιμές και οι ανεξάρτητες μεταβλητές είναι συνεχείς, κατηγορικές ή και τα δύο. Αποτελεί ιδανικό αλγόριθμο κατηγοριοποίησης όταν το αποτέλεσμα της πρόβλεψης λαμβάνει δύο τιμές, καθότι η λογιστική καμπύλη κινείται μεταξύ των τιμών 0 και 1.

Στις μεθόδους πρόβλεψης ύποπτων συναλλαγών η εξαρτημένη μεταβλητή, όπως έχει αναφερθεί και έως τώρα, λαμβάνει την τιμή 0 όταν η συναλλαγή είναι νόμιμη και την τιμή 1, όταν είναι προιόν απάτης. Σε αντίθεση με τη συνήθη γραμμική παλινδρόμηση, το λογιστικό μοντέλο δεν θεωρεί ως γραμμική τη συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτώμενη μεταβλητή.

Το λογιστικό μοντέλο ορίζεται ως εξής:

$$\log(\frac{p}{1-p}) = a_0 \ + \ a_1 X_1 \ + a_2 X_2 \ + \ a_n X_n$$

Όπου X_1, X_2, X_n είναι οι ανεξάρτητες μεταβλητές, και p η πιθανότητα η εξαρτημένη μεταβλητή να παίρνει τιμή 1, να πρόκειται δηλαδή για απάτη. Η a_0 πρόκειται για μια σταθερά, ενώ τα a_1, a_2, a_n είναι οι συντελεστές συσχέτισης των ανεξάρτητων μεταβλητών.

Τα αποτελέσματα των προβλέψεων των ταξινομητών του πρώτου επιπέδου (επίπεδο-0) αποθηκεύονται σε δύο πίνακες, αφού πρώτα μετασχηματιστούν ώστε να έχουν τιμές από -1 έως 1. Η διαδικασία αυτή πραγματοποιείται με τις παρακάτω εντολές.

```
## LightGBM
cl1_y_preds = y_preds.reshape(-1,1)
cl1_y_oof = y_oof.reshape(-1,1)
## XGBoost
cl2_y_preds = y_preds1.reshape(-1,1)
cl2_y_oof = y_oof1.reshape(-1,1)
## Random Forest
cl3_y_preds = y_preds2.reshape(-1,1)
cl3_y_oof = y_oof2.reshape(-1,1)
```

```
x_train = np.concatenate((cl1_y_preds, cl2_y_preds, cl3_y_preds), axis=1)
x_test = np.concatenate((cl1_y_oof, cl2_y_oof, cl3_y_oof), axis=1)
```

Στη συνέχεια ακολουθεί η εκτέλεση του αλγορίθμου της λογιστικής παλινδρόμησης, με ταυτόχρονη εφαρμογή 5-πτυχης διασταυρούμενης επικύρωσης.

```
start_time = time.time()
logistic_regression = LogisticRegressionCV(n_jobs=-1,cv=5)
logistic_regression.fit(x_train,y_train)

y_oof3 = logistic_regression.predict_proba(x_test)[:,1]
```

Η τιμή της μετρικής AUC βρίσκεται να σημειώνει τιμή 0.9721, με τις μετρικές της ευαισθησίας και της ακρίβειας πρόβλεψης να φτάνουν το 0.72 και 0.90 αντίστοιχα.

```
Test AUC score 0.972135
Test recall score 0.725419

The precision is 0.8966756513926325
```

ΚΕΦΑΛΑΙΟ 5: Αποτελέσματα, Συμπεράσματα και Μελλοντική Εργασία

5.1 Αποτελέσματα Ανά Μέθοδο Επαναδειγματοληψίας

Μετά και την ολοκλήρωση της εφαρμογής των 4 αλγορίθμων στα δεδομένα της εργασίας, στην παρούσα ενότητα παρουσιάζονται ορισμένες αναλύσεις επί των αποτελεσμάτων. Τα αποτελέσματα αφορούν τις τιμές των μετρικών AUC, Recall, Precision και Time για τις τέσσερις μεθόδους επαναδειγματοληψίας (Without Resampling, Random Undersampling, Random Oversampling, SMOTE) που εφαρμόστηκαν.

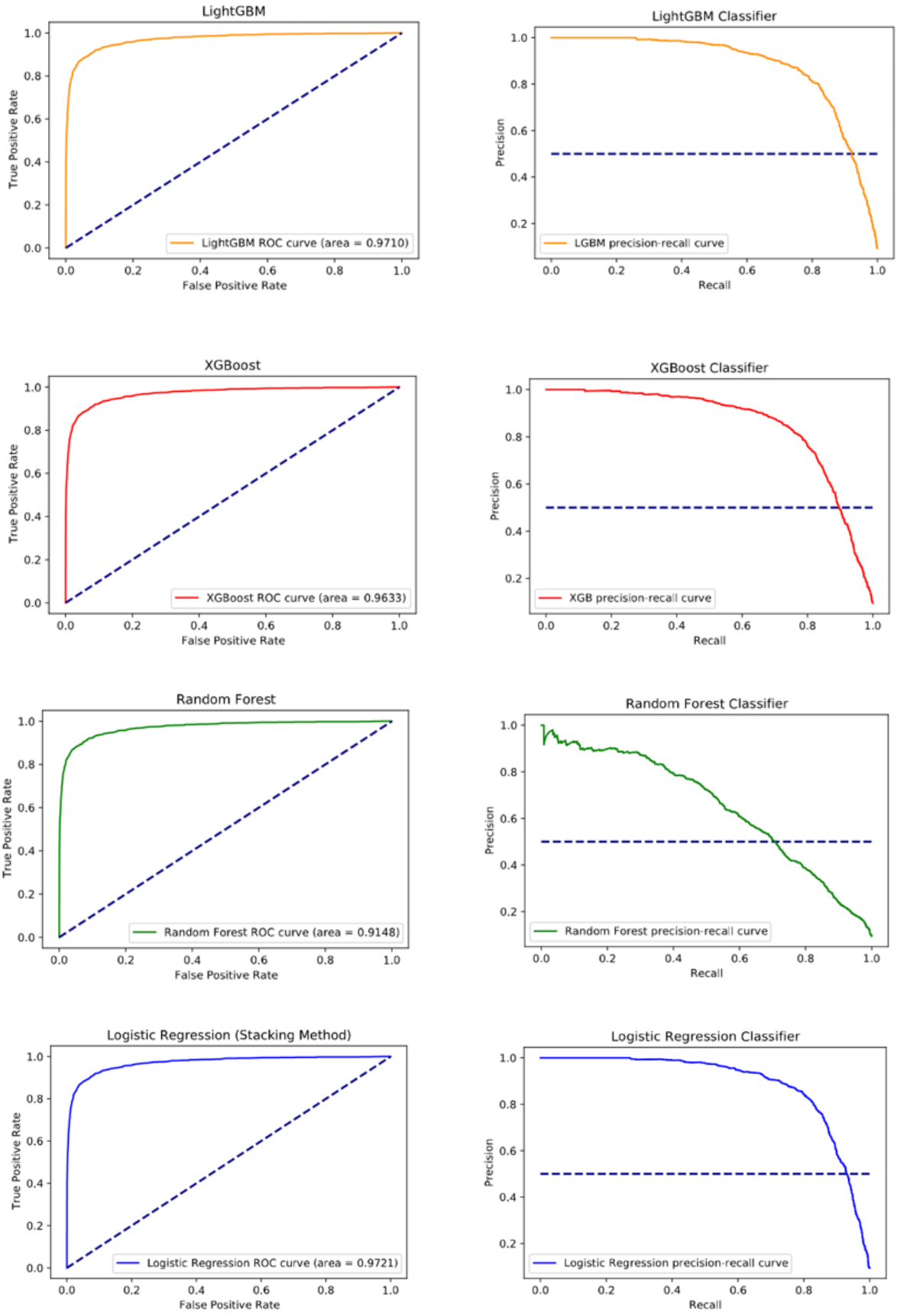
5.1.1 Χωρίς Επαναδειγματοληψία

	AUC	Recall	Precision	Time (sec)
LightGBM	0.9710	0.6810	0.9060	50.26
XGBoost	0.9633	0.6570	0.9010	230.05
Random Forest	0.9148	0.7786	0.4104	35.23
Log.Regression	0.9721	0.7254	0.9005	2.31

Πίνακας 3 – Αποτελέσματα Αλγορίθμων Χωρίς Επαναδειγματοληψία

Παρατηρείται ότι η μέθοδος της γενίκευσης (Stacking) επιτυγχάνει υψηλότερες τιμές για τις μετρικές τόσο της περιοχής κάτω από την καμπύλη ROC (AUC), που υποδηλώνει και την γενική επίδοση του αλγορίθμου, όσο και της ευαισθησίας. Ταυτόχρονα ο χρόνος εκτέλεσης του λογιστικού μοντέλου, που είναι απόρροια συνδυασμού των 3 πρώτων ταξινομητών, είναι σημαντικά μικρότερος και διαμορφώνεται στα 2.31 δευτερόλεπτα. Αξίζει να σημειωθεί ότι η ακρίβειας πρόβλεψης είναι ελάχιστα πιο χαμηλή σε σχέση με αυτή του αλγορίθμου LightGBM σημειώνοντας τιμή 0.9005 ένταντι τιμής 0.9060 αντίστοιχα.

Παρακάτω αποκολουθούν ορισμένα ενδεικτικά διαγράμματα των καμπύλων ROC και ακρίβειας-ευαισθησίας των 4 αλγορίθμων.



Εικόνα 41 - Καμπύλες ROC και Precision-Recall αλγορίθμων

5.1.2 Με Τυχαία Υποδειγματοληψία

	AUC	Recall	Precision	Time (sec)
LightGBM	0.9655	0.8820	0.5450	23.31
XGBoost	0.9618	0.8798	0.5315	51.21
Random Forest	0.9132	0.7873	0.3937	7.00
Log.Regression	0.9686	0.8995	0.5426	1.25

Πίνακας 4 – Αποτελέσματα Αλγορίθμων με Τυχαία Υποδειγματοληψία

Κάνοντας χρήση της μεθόδου τυχαίας υποδειγματοληψίας παρατηρείται ότι η απόδοση των αλγορίθμων μειώνεται ελάχιστα. Η διαπίστωση αυτή αποτυπώνεται στην μετρική AUC και Precision. Πιο συγκεκριμένα η περιοχή κάτω από την καμπύλη ROC ελαττώνεται για όλους τους αλγορίθμους κατά περίπου μισή με μία μονάδα. Ενδεικτικά παρατηρείται ότι η τιμή της AUC, για το λογιστικό μοντέλο, μειώνεται από το 0.9721 χωρίς επαναδειγματοληψία, στο 0.9686 με τη μέθοδο της τυχαίας υποδειγματοληψίας. Αξίζει, επίσης, να τονιστεί ότι η μετρική της ακρίβειας, με την τεχνική της τυχαίας υποδειγματοληψίας, σημειώνει σημαντική πτώση κατά περίπου 35 με 40 μονάδες. Βασική αιτία της μείωσης αυτής είναι το γεγονός, ότι κατά τη μείωση των δεδομένων χάνεται σημαντική πληροφορία. Τέλος, οι χρόνοι εκτέλεσης είναι εμφανώς μικρότεροι λόγω του μικρότερου όγκου δεδομένων προς επεξεργασία.

5.1.3 Με Τυχαία Υπερδειγματοληψία

	AUC	Recall	Precision	Time (sec)
LightGBM	0.9706	0.8762	0.6409	71.66
XGBoost	0.9653	0.8711	0.5900	579.09
Random Forest	0.9154	0.7837	0.4077	86.74
Log.Regression	0.9620	0.8689	0.7261	3.6

Πίνακας 5 – Αποτελέσματα Αλγορίθμων με Τυχαία Υπερδειγματοληψία

Με χρήση της τεχνικής της τυχαίας υπερδειγματοληψίας, παρατηρείται ότι ο αλγόριθμος με τη βέλτιστη απόδοση είναι πλέον ο LightGBM με τιμή AUC ίση με 0.9706. Φυσικά η τιμή αυτή δεν υπερβαίνει την απόδοση του λογιστικού μοντέλου χωρίς επαναδειγματοληψία, ωστόσο επιτυγχάνει ταυτόχρονα υψηλή ευαισθησία της τάξης του 0.8762. Μειονέκτημα της μεθόδου είναι ο αυξημένος χρόνος εκτέλεσης συγκριτικά με τις προηγούμενες δύο, που οφείλεται κυρίως στην αύξηση του όγκου των δεδομένων και συγκεκριμένα των περιπτώσεων απάτης.

5.1.4 Με Υπερδειγματοληψία Συνθετικής Μειονότητας (SMOTE)

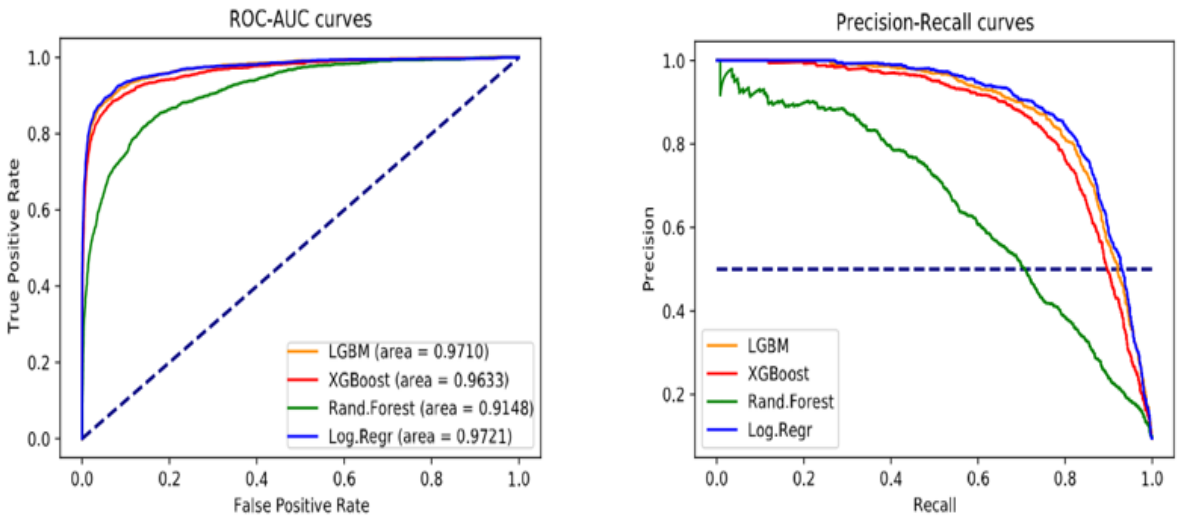
	AUC	Recall	Precision	Time (sec)
LightGBM	0.9647	0.6657	0.8788	114.92
XGBoost	0.9603	0.6708	0.8763	502.57
Random Forest	0.8941	0.6089	0.5157	79.24
Log.Regression	0.9559	0.6766	0.8772	3.41

Πίνακας 6 – Αποτελέσματα Αλγορίθμων με Υπερδειγματοληψία Συνθετικής Μειονότητας (SMOTE)

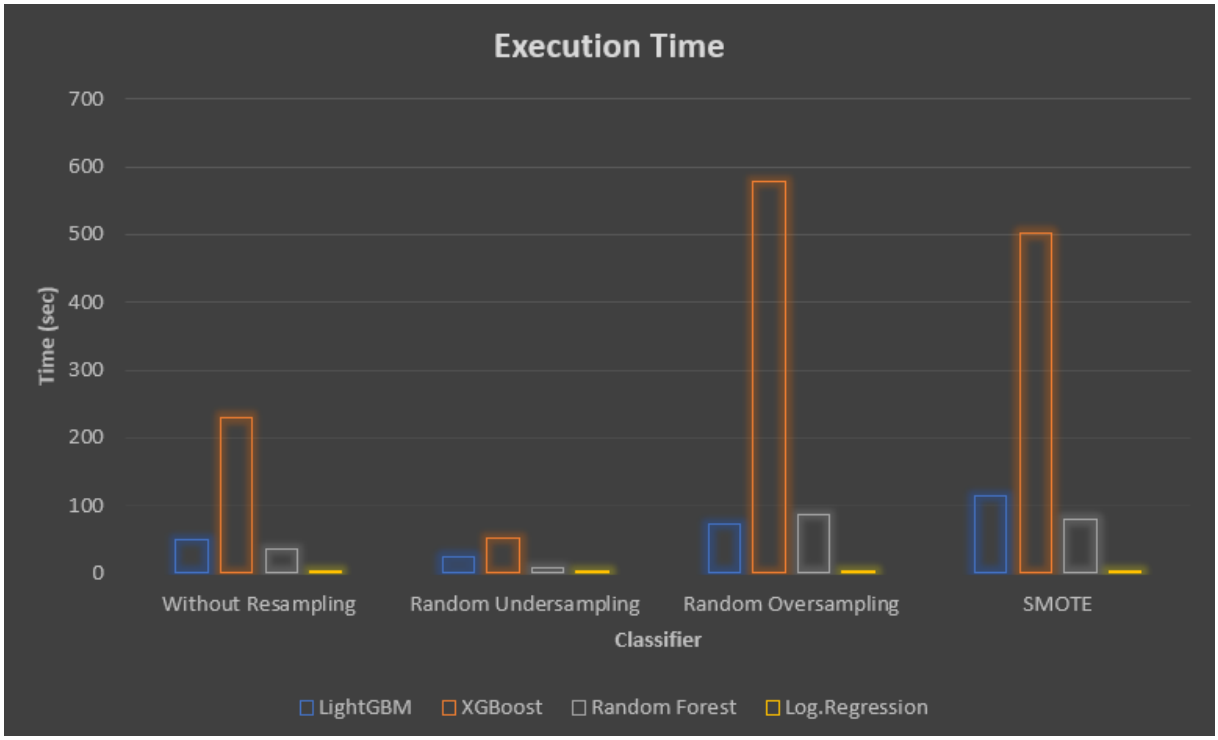
Και εδώ όπως και στη μέθοδο τυχαίας υπερδειγματοληψίας, ο αλγόριθμος LightGBM επιτυγχάνει τη βέλτιστη απόδοση έναντι των ανταγωνιστών ταξινομητών, ωστόσο η τιμή της AUC (0.9647) δεν υπερβαίνει την αντίστοιχη του αλγορίθμου Log.Regression με τη μέθοδο χωρίς επαναδειγματοληψία (0.9721). Επιπροσθέτως, ο χρόνος εκτέλεσης βρίσκεται σε αρκετά υψηλά επίπεδα, με την ταξινόμηση να ολοκληρώνεται σε 114.92 δευτερόλεπτα.

Εν τέλει, από την μελέτη των παραπάνω στοιχείων διαπιστώνεται ότι για τα υπάρχοντα δεδομένα της εργασίας, η μέθοδος χωρίς επαναδειγματοληψία πέτυχε τον καλύτερο συνδυασμό AUC, ευαισθησίας και ακρίβειας πρόβλεψης, σε συνάρτηση πάντα με τον χρόνο εκτέλεσης των αλγορίθμων. Το μοντέλο της λογιστικής παλινδρόμησης σημείωσε τον ταχύτερο χρόνο εκτέλεσης (2.31 sec) έναντι των υπολοίπων ταξινομητών, επιτυγχάνοντας παράλληλα τη βέλτιστη πρόβλεψη (0.9721) για το ποια συναλλαγή είναι προιόν απάτης και ποια όχι. Η εξήγηση έγκειται στο γεγονός ότι η ταξινόμηση με λογιστική παλινδρόμηση συνδυάζει τις πιθανότητες από τρία μοντέλα (LightGBM, XGBoost και Random Forest), μειώνοντας κατά αυτόν τον τρόπο τη μεροληψία στα δεδομένα και τη διακύμανση των προβλέψεων. Τέλος, ακολουθούν τα συγκεντρωτικά διαγράμματα:

- των καμπύλων ROC και ακρίβειας-ευαισθησίας για τους 4 αλγορίθμους, με τη μέθοδο χωρίς επαναδειγματοληψία (εικόνα 42).
- του χρόνου εκτέλεσης των 4 αλγορίθμων με τις διαφορετικές τεχνικές δειγματοληψίας που έλαβαν χώρα (εικόνα 43).



Εικόνα 42 - Συγκεντρωτικές Καμπύλες ROC, Precision-Recall με Μέθοδο Χωρίς Επαναδειγματοληψία



Εικόνα 43 - Συγκριτικό Διάγραμμα Χρόνου Εκτέλεσης Αλγορίθμων

5.2 Συμπεράσματα

Σκοπός της παρούσας εργασίας είναι η ενδελεχής μελέτη και ανάλυση της συνεισφοράς των επιστημών της μηχανικής μάθησης και της εξόρυξης δεδομένων στον χρηματοοικονομικό κλάδο και συγκεκριμένα στα τραπεζικά ιδρύματα. Η ραγδαία ανάπτυξη των τεχνικών εξόρυξης δεδομένων και της τεχνητής νοημοσύνης τα τελευταία χρόνια, δεν θα μπορούσε να αφήσει ανεπηρέαστο έναν από τους κινητήριους κλάδους της οικονομικής ζωής, όπως είναι αυτός των χρηματοοικονομικών. Η εργασία αυτή πραγματεύεται αρχικά τα πλεονεκτήματα καθώς και τα μειονεκτήματα που διέπουν την επιστήμη των μεγάλων δεδομένων. Εν συνεχεία, γίνεται αναφορά στους κινδύνους που υφέρπουν για τα χρηματοπιστωτικά ιδρύματα και συγκεκριμένα στον κίνδυνο της απάτης.

Πιο αναλυτικά, στο κύριο μέρος της εργασίας, πραγματοποιείται μια εκτενής αναφορά στον κίνδυνο της απάτης στις συναλλαγές που πραματοποιούνται με χρήση πιστωτικής κάρτας. Στο πρώτο μέρος, ο αναγνώστης έρχεται σε επαφή με το θεωρητικό υπόβαθρο που υπάρχει πίσω από τα φαινόμενα απάτης με χρήση πιστωτικής κάρτας στον τραπεζικό τομέα. Στο δεύτερο μέρος, παρατίθεται το πειραματικό σκέλος της εργασίας, όπου λαμβάνει χώρα η εφαρμογή αλγορίθμων εξόρυξης δεδομένων με σκοπό τη δημιουργία ενός μοντέλου που θα προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια το είδος κάθε συναλλαγής (νόμιμη/απατηλή). Αποδεικνύεται ότι η η ανάπτυξη μεθόδων πρόβλεψης βασισμένες στην συσσωρευμένη γενίκευση ταξινομητών μπορεί να επιφέρει τις βέλτιστες προβλέψεις, έχοντας παράλληλα μικρό χρόνο εκτέλεσης.

Η ανάπτυξη τέτοιων μοντέλων πρόβλεψης απάτης είναι πλέον μείζονος σημασίας για τα χρηματοπιστωτικά ιδρύματα, από τη στιγμή που ο όγκος των συναλλαγών με χρήση πλαστικού χρήματος αυξάνεται διαρκώς και τείνει να αποτελέσει το βασικό τρόπο συναλλαγματικών διεργασιών. Η άρτια εφαρμογή τέτοιων μεθόδων θα μπορέσει να περιορίσει στο ελάχιστον τις ζημίες των ιδρυμάτων από φαινόμενα απάτης, ενισχύοντας παράλληλα το πνεύμα ασφάλειας προς τους πελάτες.

Τέλος, αξίζει να τονιστεί ότι η εισαγωγή της τεχνητής νοημοσύνης στον τραπεζικό κλάδο θα πρέπει να συνοδεύεται από την κατάλληλη εκπαίδευση του προσωπικού, έτσι ώστε να διασφαλίζεται η απρόσκοπτη λειτουργία των συστημάτων. Κατά αυτόν τον τρόπο, η μετάβαση στην νέα εποχή του ψηφιακού μετασχηματισμού θα μπορέσει να γίνει ομαλά, ανταποκρινόμενοι πάντοτε στις ανάγκες και τις απαιτήσεις των πελατών.

5.3 Μελλοντικές Κατευθύνσεις

Στην παρούσα εργασία παρουσιάζονται ορισμένες αναλύσεις επί των διαθέσιμων δεδομένων, έτσι ώστε να επιτευχθεί η βέλτιστη απόδοση των αλγορίθμων μηχανικής μάθησης για πρόβλεψη αν μια συναλλαγή είναι προιόν απάτης. Ως μελλοντική εργασία, θα μπορούσε να ακολουθηθούν διαφορετικές τεχνικές ανάλυσης δεδομένων, όπως η δημιουργία νέων μεταβλητών, από τις ήδη υπάρχουσες, που θα συνεισφέρουν περισσότερο στην πρόβλεψη. Τέλος, θα εμφάνιζε ενδιαφέρον η εφαρμογή εναλλακτικών, μεμονωμένων ταξινομητών, διαφορετικών χαρακτηριστικών, από τη στιγμή που έχει παρατηρηθεί ότι κατά την τεχνική της συσσωρευμένης γενίκευσης, η διαφορετικότητα των αλγορίθμων συμβάλλει στην μέγιστη απόδοση των μετα-ταξινομητών.

Βιβλιογραφικές Αναφορές

1. Jacques Bughin. Telcos: The untapped promise of big data. McKinsey Quarterly, 2016
https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/telcos-the-untapped-promise-of-big-data

2. Indian J Orthop. Significant Applications of Big Data in COVID-19 Pandemic
2020 Jul; 54(4): 526–528.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/

3. Antonio Marcos Alberti, Dhananjay Singh. Internet of Things: Perspectives, Challenges and Opportunities, 2013
https://www.researchgate.net/publication/236656851_Internet_of_Things_Perspectives_Challenges_and_Opportunities

4. Niels Martin, Marijke Swenne, Benoit Depaire, Mieke Julie Jans, An Caris, Koen Vanhoof. Batch processing: definition and event log identification, 2015
https://www.researchgate.net/publication/287198292_Batch_processing_definition_and_event_log_identification

5. Adeyinka K.Akanbi, Muthoni Masinde. A Distributed Stream Processing Middleware Framework for Real-Time Analysis of Heterogeneous Data on Big Data Platform: Case of Environmental Monitoring, 2020
https://www.researchgate.net/publication/341904461_A_Distributed_Stream_Processing_Middleware_Framework_for_Real-Time_Analysis_of_Heterogeneous_Data_on_Big_Data_Platform_Case_of_Environmental_Monitoring#pf18

6. Martin Strohbach, Jörg Daubert, Herman Ravkin, Mario Lischka. Big Data Storage
https://www.researchgate.net/publication/299617100_Big_Data_Storage

7. Ehsan Hajizadeh, Hamed Davari-Ardakani, Jamal Shahrabi. Application of data mining techniques in stock markets: A survey, 2010
https://www.researchgate.net/publication/228664309_Application_of_data_mining_techniques_in_stock_markets_A_survey

8. Abdulmohsen Algarni. Data Mining in Education, 2016
https://www.researchgate.net/publication/304808426_Data_Mining_in_Education

9. Gary M.Weiss. Data Mining and Knowledge Discovery Handbook pp 1189-1201
https://link.springer.com/chapter/10.1007/0-387-25465-X_56

10. Efstathios Kirkos, Charalambos Spathis, Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32(4), 995-1003, 2007
https://www.researchgate.net/publication/222581013_Data_mining_techniques_for_the_detection_of_fraudulent_financial_statements_Expert_Systems_with_Applications_324_995-1003

11. Mary DeRosa. Data Mining and Data Analysis for Counterterrorism, 2004

12. Zengan Gao, Mao Ye. A framework for data mining-based anti-money laundering research, 2007
https://www.researchgate.net/publication/228353074_A_framework_for_data_mining-based_anti-money_laundering_research

13. Ravi Vadlamani. Chapter I Introduction to Banking Technology and Management, 2011
https://www.researchgate.net/publication/237383828_Chapter_I_Introduction_to_Banking_Technology_and_Management

14. https://www.infosys.com/industries/financial-services/white-papers/Documents/social-media-analytics.pdf

15. Simon C.K. Shiu, James N.K.Liu, Jennie L.C.Lam, Bo Feng. A Data Mining Approach for Branch and ATM Site Evaluation, 2006
https://www.researchgate.net/publication/221337952_A_Data_Mining_Approach_for_Branch_and_ATM_Site_Evaluation

16. Netra Pal Singh, Devender Singh. Chatbots and Virtual Assistant in Indian Banks, 2019
https://www.researchgate.net/publication/338659152_Chatbots_and_Virtual_Assistant_in_Indian_Banks

17. Λουμιώτης, Β.Ι. & Τζίφας, Β.Ν. (2018), Βασικές Οδηγίες Εφαρμογής Διεθνών Προτύπων Ελέγχου, Εκδόσεις Σταμούλη, Αθήνα

18. Basel Committee on Banking Supervision. "Basel Accords II." Basel, Switzerland: Bank for International Settlements Press & Communications, June 2006

19. M. Bellis. Invention of Credit Cards
http://inventors.about.com/od/cstartinventions/a/credit_cards.htm

20. https://shiftprocessing.com/credit-card-fraud-statistics/

21. European Central Bank. Report on card fraud available:
https://www.ecb.europa.eu/press/pr/date/2014/html/pr140225.en.html

22. Tej Paul Bhatla, V.Prabhu, A.dua. Understanding Credit Card Frauds, 2003
https://api.semanticscholar.org/CorpusID:168440535

23. Pearson Correlation Coefficient.
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

24. Kolmogorov-Smirnov test.
https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

25. N.V. Chawla, K.W.Bowyer, L.O.Hall, W.P.Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, 2011
https://arxiv.org/abs/1106.1813

26. Gradient Booosting.
https://en.wikipedia.org/wiki/Gradient_boosting

27. Jerome H.Friedman. Stochastic Gradient Boosting (Computational Statistics & Data Analysis, 2002, vol. 38, issue 4, 367-378)
https://www.researchgate.net/publication/222573328_Stochastic_Gradient_Boosting

28. Which algorithm takes the crown: Light GBM vs XGBOOST?
https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

29. Breiman. L. Random Forests, Machine Learning 45, 5–32, 2001
https://doi.org/10.1023/A:1010933404324

30. Wolpert, D. 1992. Stacked generalization. Neural Networks, 5 (2), 241-260

31. Breiman, L. 1996. Stacked regressions. Machine Learning, 24, 49-64

32. Smyth, P., and Wolpert D. 1997. Stacked density estimation. Advances in Neural Information Processing systems