ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ»
«ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ»

Τεχνικές Εξόρυξης Γνώσης στον Τραπεζικό Τομέα

Μελέτη Περίπτωσης: Πρόβλεψη Πιθανότητας Αθέτησης στις Πιστωτικές Κάρτες

Καρζής Αναστάσιος, Α.Μ: ΜΕ1812
Επιβλέπων Καθηγητής: Φιλιππάκης Μιχαήλ

Διπλωματική Εργασία υποβληθείσα στο Τμήμα Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς ως μέρους των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης στα Πληροφοριακά Συστήματα και Υπηρεσίες

Πειραιάς, Φεβρουάριος 2020

UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

POSTGRADUATE STUDIES PROGRAMME

« INFORMATION SYSTEMS & SERVICES»

«BIG DATA AND ANALYTICS»

Data Mining Techniques in the Banking Sector

Case Study: Predicting Credit Card Default Probability

Karzis Anastasios, A.M: ME1812

Supervisor Professor: Filippakis Michael

Master Thesis submitted to the Department of Digital Systems of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Information Systems and Services

Piraeus, February 2020

Περίληψη

Στη σημερινή εποχή η τεχνολογία έχει αναπτυχθεί περισσότερο από κάθε άλλη επιστήμη. Σε αντίθεση με παλαιότερα, τις τελευταίες δεκαετίες, επικρατεί η τάση η έννοια τεχνολογία να αναφέρεται κυρίως στην τεχνολογία τον υπολογιστών. Στη σύγχρονη πραγματικότητα, κυριότερες ανάγκες και προκλήσεις στον τομέα αυτό αποτελούν, τόσο η διαχείριση του τεράστιου όγκου δεδομένων, τα οποία αυξήθηκαν εκθετικά – και εξακολουθούν να αυξάνονται – τα τελευταία χρόνια, όσο και η αξιοποίηση των δεδομένων αυτών με τον βέλτιστο τρόπο ώστε να παραχθούν πολύτιμες, αξιόπιστες και ενδεχομένως άγνωστες έως τώρα πληροφορίες. Τα πλεονεκτήματα από την επίτευξη των παραπάνω στόχων είναι ποικίλα και δύνανται να επηρεάσουν καταλυτικά πλήθος επιχειρηματικών κλάδων. Ο κλάδος των χρηματοοικονομικών είναι ένας από τους κυριότερους, και μάλιστα σε μια εποχή όπου η παγκόσμια και ευρωπαϊκή οικονομία βρίσκονται σε φάση αξιοσημείωτης επιβράδυνσης, προσπαθώντας παράλληλα να ανακάμψουν από την παγκόσμια οικονομική ύφεση του 2008 (μια από τις μεγαλύτερες στην ιστορία). Γίνεται εύκολα επομένως κατανοητό ότι ιδιαίτερα για τον τραπεζικό κλάδο, που αποτελεί την καρδιά της εκάστοτε οικονομίας, η χρήση εξελιγμένων μορφών τεχνολογίας μπορεί να βοηθήσει ουσιαστικά στην αυτοματοποίηση και βελτιστοποίηση των διαδικασιών με αποτέλεσμα αφενός τον περιορισμό των κινδύνων και αφετέρου την κερδοφορία.

Οι προαναφερθείσες έννοιες και οι μεταξύ τους σχέσεις παρουσιάζονται και αναλύονται στην παρούσα διπλωματική εργασία, η οποία στοχεύει στο να αποτυπώσει βασικές θεωρητικές γνώσεις επί των νέων τεχνολογιών και καινοτομιών και να τις συνδυάσει με τη σημασία τους για τον τραπεζικό κλάδο. Πιο συγκεκριμένα, η εργασία αρχικά (1ο και 2ο Κεφάλαια) πραγματεύεται τις έννοιες των Μεγάλων Δεδομένων, της Εξόρυξης Δεδομένων, της Μηχανικής Μάθησης και της χρήσης λύσεων Τεχνητής Νοημοσύνης στον επιχειρηματικό κόσμο. Έπειτα (3ο Κεφάλαιο) ο συντάκτης εστιάζει στη συσχέτιση των ανωτέρω με τον χρηματοπιστωτικό τομέα και ιδιαίτερα στη σχέση τους με την εκτίμηση και τον περιορισμό ενός από τους σημαντικότερους κινδύνους που αντιμετωπίζουν οι τράπεζες, τον πιστωτικό κίνδυνο. Η ανάλυση μάλιστα κατευθύνεται σταδιακά σε μια πιο εξειδικευμένη περιοχή, αυτή της εκτίμησης της πιθανότητας αθετήσεως πιστούχου για το χαρτοφυλάκιο των πιστωτικών καρτών, μιας κατηγορίας δανείων με ιδιαίτερα χαρακτηριστικά. Στα επόμενα δυο κεφάλαια που έπονται, πραγματοποιείται η πρακτική εφαρμογή αλγορίθμων μηχανικής μάθησης επί συνόλου δεδομένων τραπέζης εξωτερικού για την πρόβλεψη της πιθανότητας αθετήσεως πιστούχου και η αξιολόγηση των αποτελεσμάτων.

Η διαδρομή που καλείται ο αναγνώστης να ακολουθήσει στα πλαίσια της συγκεκριμένης εργασίας, περνώντας από το «γενικό» στο «ειδικό» και συσχετίζοντας διαφορετικές έννοιες - αλληλένδετες ωστόσο – είναι απόρροια τόσο του ενδιαφέροντος του συντάκτη επί των σχετικών αντικειμένων όσο και του ακαδημαϊκού και επαγγελματικού υπόβαθρού του.

Σημαντικοί Όροι: Εξόρυξη Δεδομένων, Τραπεζικός Τομέας, Πιστωτικές Κάρτες, Πιθανότητα Αθέτησης Πιστούχου

Abstract

Nowadays, technology is one of the most developed sciences. In recent decades, there is a tendency by term technology to refer mainly to computer technology. In today's reality, the major needs and challenges in this area are both the management of huge volumes of data, which have grown exponentially and continue to grow, and the utilization of this data in an optimal way to produce valuable and reliable data and explore new areas and information yet unknown. The advantages of achieving the above goals are various and can have a catalytic effect on many business sectors. The financial sector is one of the foremost, especially at a time when the global and european economies are in a phase of significant slowdown, while trying to recover from the 2008 global economic downturn. It is therefore easily understood that particularly for the banking sector, which is the heart of the economy, the use of sophisticated technologies can help in automation and optimization of processes and thereby in reducing risk and in the increase of profitability.

The aforementioned areas of interest and the relationship between them are presented and analyzed in this thesis, which aims to capture and combine basic theoretical knowledge on new technologies and innovations and relate them to the banking industry. In particular, the thesis initially (Chapters 1 and 2) addresses the concepts of Big Data, Data Mining, Machine Learning and the use of Artificial Intelligence solutions in the business world. Subsequently (Chapter 3) the author focuses on the relationship between the above technologies and the financial sector, and in particular on their relationship with the assessment and mitigation of one of the major risks facing banks, credit risk. The analysis is progressively moving to a more specialized area, that of assessing the probability of default on the credit card portfolio, a category of loans with particular characteristics. In the next two chapters, machine learning algorithms are applied on an external bank’s data set to predict the probability of default and evaluate the results.

The path which the reader is invited to follow in this particular thesis, in which different - interlinked however - meanings are presented, is a result of both the author's interest in these subject matters and his academic and professional background.

Keywords: Data Mining, Banking Field, Credit Cards, Probability of Default (PD)

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Φιλιππάκη Μιχαήλ για την πολύτιμη βοήθεια και καθοδήγησή του καθ’ όλη τη διάρκεια της διπλωματικής μου εργασίας και μέχρι την ολοκλήρωσή της.

Επίσης, θέλω να πω ένα μεγάλο ευχαριστώ σε όλους τους κοντινούς και αγαπημένους μου ανθρώπους για τη στήριξη και τη βοήθεια που μου παρείχαν και ειδικότερα στην Αρχοντία και στον Δήμο.

Περιεχόμενα

Λίστα Εικόνων

Κεφάλαιο 1: Γενικές αρχές Εξόρυξης Γνώσης και Μεγάλων Δεδομένων

1.1 Μεγάλα Δεδομένα: Χαρακτηριστικά, Σημασία, Κατηγορίες

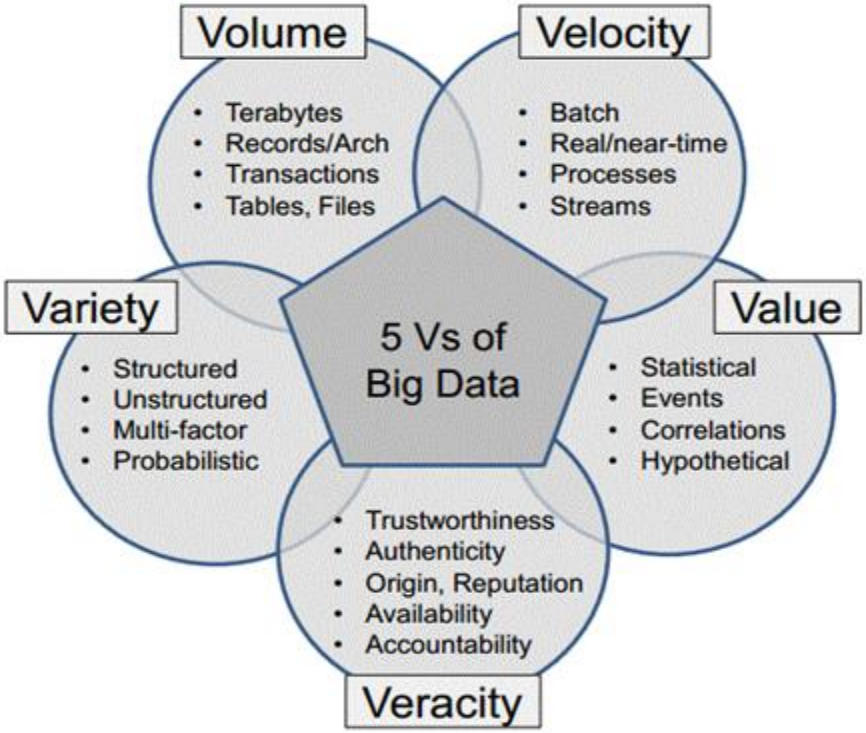
Ο όρος "μεγάλα δεδομένα" (Big Data) χρησιμοποιείται από τις αρχές της δεκαετίας του 1990. Αν και δεν είναι ακριβώς γνωστό ποιος χρησιμοποίησε για πρώτη φορά τον όρο, οι περισσότεροι άνθρωποι πιστώνουν τον John R. Mashey (ο οποίος τότε εργάστηκε στη Silicon Graphics) ότι έκανε τον όρο δημοφιλή.

Στην πραγματικότητα, τα μεγάλα δεδομένα δεν είναι μια έννοια εντελώς νέα ή παρατηρείται μόνο τις τελευταίες δύο δεκαετίες. Κατά τη διάρκεια των αιώνων, οι άνθρωποι προσπαθούν να χρησιμοποιήσουν τεχνικές επεξεργασίας δεδομένων και ανάλυσης για να στηρίξουν τη διαδικασία λήψης αποφάσεων. Οι αρχαίοι Αιγύπτιοι γύρω στα 300 π.Χ. προσπάθησαν ήδη να συλλάβουν όλα τα υπάρχοντα «δεδομένα» στη βιβλιοθήκη της Αλεξάνδρειας. Επιπλέον, η Ρωμαϊκή Αυτοκρατορία συνήθιζε να αναλύει προσεκτικά τα στατιστικά του στρατού της για να καθορίσει τη βέλτιστη κατανομή για τον στρατό της.

Ωστόσο, τις τελευταίες δύο δεκαετίες, ο όγκος και η ταχύτητα με την οποία παράγονται τα δεδομένα άλλαξαν. Ο συνολικό όγκος δεδομένων στον κόσμο ήταν 4,4 zettabytes το 2013 και έως το τέλος του έτους (2020) προβλέπεται να έχει φτάσει τα 44 zettabytes. Για να γίνει κατανοητό, 44 zettabytes ισοδυναμούν σε 44 τρισεκατομμύρια gigabytes. Ακόμη και με τις πλέον προηγμένες τεχνολογίες σήμερα, είναι αδύνατο να αναλυθούν όλα αυτά τα δεδομένα. Η ανάγκη επεξεργασίας αυτών των όλο και μεγαλύτερων (και μη δομημένων) συνόλων δεδομένων γέννησε λοιπόν την ανάγκη μετατροπής της παραδοσιακής ανάλυσης δεδομένων σε «Μεγάλα Δεδομένα», την τελευταία δεκαετία.

Ψηφιακά δεδομένα συναντώνται πλέον παντού: σε κάθε τομέα, σε κάθε οικονομία, σε κάθε οργανισμό και χρήστη της ψηφιακής τεχνολογίας. Τα μεγάλα δεδομένα έλκουν όλο και περισσότερο το ενδιαφέρον των ηγετών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται να ωφεληθούν από την αξιοποίησή τους. Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή.

Ο όρος Μεγάλα Δεδομένα (Big Data) χρησιμοποιείται κατά κόρον τα τελευταία χρόνια, χωρίς να υπάρχει κάποιος σαφής ορισμός του. Ο ορισμός που έχει επικρατήσει είναι αυτός της Gartner από το 2001: “Τα μεγάλα δεδομένα είναι δεδομένα που περιέχουν μεγάλη ποικιλία (Variety), πολύ αυξανόμενο όγκο (Volume) και μεγάλη ταχύτητα παραγωγής (Velocity). Αυτά είναι γνωστά ως τα 3V’s”. Τα τελευταία χρόνια όμως εμφανίζονται και άλλες δύο διαστάσεις: εγκυρότητα (Veracity) και αξία (Value).



Εικόνα 1 - Τα 5 V's των Μεγάλων Δεδομένων

1.1.1 Χαρακτηριστικά Μεγάλων Δεδομένων

(i) Volume (Όγκος)

Το όνομα "Big Data" το ίδιο υποδηλώνει το μεγάλο μέγεθος της πληροφορίας. Το "volume" ουσιαστικά αντιπροσωπεύει ένα τεράστιο σύνολο δεδομένων.

Για τον προσδιορισμό της αξίας των δεδομένων, το μέγεθος αυτών διαδραματίζει πολύ σημαντικό ρόλο. Εάν ο όγκος των δεδομένων είναι πολύ μεγάλος, τότε στην πραγματικότητα θεωρούνται ως "Μεγάλα Δεδομένα". Αυτό σημαίνει ότι εξαρτάται από τον όγκο των δεδομένων το αν κάποια δεδομένα μπορούν να θεωρηθούν ως μεγάλα δεδομένα ή όχι.

- Παράδειγμα Όγκου: Το έτος 2016, η εκτιμώμενη παγκόσμια κινητή επισκεψιμότητα ήταν 6,2 Exabytes (6,2 δισ. GB) το μήνα και μέχρι το έτος 2020 θα έχουμε σχεδόν 40.000 ExaBytes.

(ii) Velocity (Ταχύτητα)

Ο όρος velocity αναφέρεται στην υψηλή ταχύτητα συσσώρευσης δεδομένων και στην ταχύτητα ροής μεγάλων δεδομένων από πηγές όπως μηχανές, δίκτυα, κοινωνικά μέσα, κινητά τηλέφωνα κ.λπ. Υπάρχει μια μαζική και συνεχής ροή δεδομένων. Αυτό καθορίζει το δυναμικό των δεδομένων και το πόσο γρήγορα παράγονται και επεξεργάζονται τα δεδομένα για να ικανοποιήσουν τις απαιτήσεις.

- Παράδειγμα Ταχύτητας: Υπάρχουν περισσότερες από 3,5 δισεκατομμύρια αναζητήσεις την ημέρα που πραγματοποιούνται στη μηχανή της Google. Επίσης, οι χρήστες του Facebook αυξάνονται κατά 22% (περίπου) κάθε χρόνο.

(iii) Variety (Ποικιλία)

Ο όρος αναφέρεται στη φύση των δεδομένων και στη διασπορά τους στις διάφορες πηγές.

Η ποικιλία είναι βασικά η άφιξη δεδομένων από ποικίλες/ νέες πηγές που βρίσκονται μέσα και έξω από μια επιχείρηση. Μπορεί να είναι δομημένα, ημι-δομημένα αλλά και αδόμητα.

Δομημένα δεδομένα: Τα δεδομένα αυτά είναι βασικά οργανωμένα δεδομένα.
Γενικά αναφέρεται σε δεδομένα με καθορισμένο μήκος και μορφή.

Ημι-δομημένα δεδομένα: Τα δεδομένα αυτά είναι κατά βάση ημι-οργανωμένα δεδομένα. Είναι γενικά μια μορφή δεδομένων που δεν συμμορφώνονται με την επίσημη δομή των δεδομένων.

Μη δομημένα δεδομένα: Σε αυτή την κατηγορία αναφερόμαστε σε μη οργανωμένα δεδομένα, δηλαδή δεδομένα που δεν ταιριάζουν με την παραδοσιακή δομή γραμμών και στηλών της σχεσιακής βάσης δεδομένων. Τα κείμενα, οι εικόνες, τα βίντεο κ.λπ. είναι τα παραδείγματα μη δομημένων δεδομένων που δεν μπορούν να αποθηκευτούν με τη μορφή σειρών και στηλών.

(vi) Veracity (Μεταβλητότητα)

Ο όρος αναφέρεται στις ασυνέπειες των δεδομένων, δηλαδή τα δεδομένα που είναι διαθέσιμα μπορεί μερικές φορές να είναι ακατάστατα και κατά συνέπεια η ποιότητα και η ακρίβεια είναι δύσκολο να ελεγχθούν.

Επιπλέον, τα μεγάλα δεδομένα είναι συνήθως μεταβλητά λόγω της πληθώρας των διαστάσεων που προκύπτουν από πολλαπλούς και διαφορετικούς τύπους δεδομένων και πηγές.

- Παράδειγμα: Τα δεδομένα μεγάλου όγκου θα μπορούσαν να δημιουργήσουν σύγχυση, από την άλλη όμως, μικρότερη ποσότητα δεδομένων θα μπορούσε να μεταφέρει μισή ή ελλιπή πληροφόρηση.

(v) Value (Αξία)

Αφού λάβουμε υπόψιν τα 4V's, έρχεται ένα ακόμα V που σημαίνει αξία (Value) .
Ο μεγάλος όγκος δεδομένων που δεν έχει αξία δεν προσφέρει τίποτα στην εκάστοτε εταιρεία, εκτός αν μετατραπεί σε κάτι χρήσιμο.

Τα δεδομένα από μόνα τους δεν έχουν καμία χρησιμότητα ή σημασία και για να αξιοποιηθούν πρέπει να μετατραπούν σε κάτι πολύτιμο για την εξαγωγή πληροφοριών. Ως εκ τούτου, μπορεί να ειπωθεί ότι η έννοια Value είναι το πιο σημαντικό V όλων των 5V's.

Εικόνα 2 - Χαρακτηριστικά των Μεγάλων Δεδομένων

1.1.2 Σημασία Μεγάλων Δεδομένων

Οι εταιρείες χρησιμοποιούν τα μεγάλα δεδομένα που συσσωρεύονται στα συστήματά τους για να βελτιώσουν τις λειτουργίες τους, να προσφέρουν καλύτερη εξυπηρέτηση πελατών, να δημιουργήσουν εξατομικευμένες καμπάνιες μάρκετινγκ με βάση συγκεκριμένες προτιμήσεις των πελατών και, τελικά, να αυξήσουν την κερδοφορία. Οι επιχειρήσεις που χρησιμοποιούν τα μεγάλα δεδομένα κατέχουν ένα δυνητικό ανταγωνιστικό πλεονέκτημα έναντι εκείνων που δεν χρησιμοποιούν, καθώς έχουν τη δυνατότητα να παίρνουν ταχύτερες και πιο επίκαιρες επιχειρηματικές αποφάσεις, υπό την προϋπόθεση ότι χρησιμοποιούν τα δεδομένα αποτελεσματικά.

Για παράδειγμα, τα μεγάλα δεδομένα μπορούν να παρέχουν στις επιχειρήσεις πολύτιμες γνώσεις για τους πελάτες τους, οι οποίες μπορούν να χρησιμοποιηθούν για να βελτιώσουν τις εκστρατείες και τις τεχνικές μάρκετινγκ, προκειμένου να αυξήσουν τα ποσοστά επαφής και διατήρησης των πελατών.

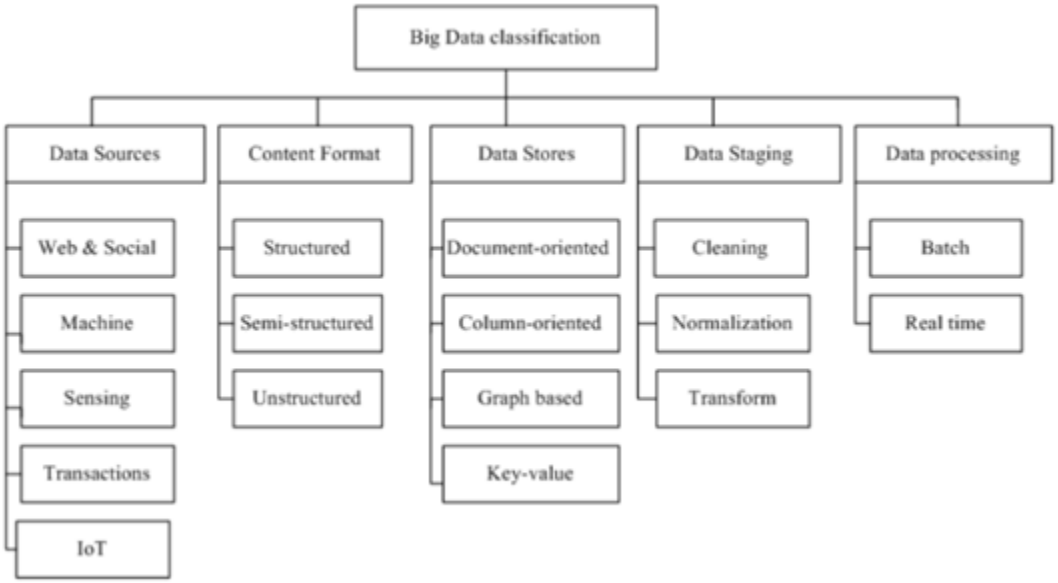
Επιπλέον, η αξιοποίηση των μεγάλων δεδομένων επιτρέπει στις εταιρείες να γίνονται όλο και περισσότερο πελατοκεντρικές. Τα ιστορικά δεδομένα και τα δεδομένα σε πραγματικό χρόνο μπορούν να χρησιμοποιηθούν για να εκτιμηθούν οι εξελισσόμενες προτιμήσεις των καταναλωτών, επιτρέποντας έτσι στις επιχειρήσεις να ενημερώσουν και να βελτιώσουν τις στρατηγικές μάρκετινγκ τους και να ανταποκριθούν καλύτερα στις επιθυμίες και τις ανάγκες των πελατών.

Τα μεγάλα δεδομένα χρησιμοποιούνται επίσης από ιατρικούς ερευνητές για τον εντοπισμό πιθανών αιτιών ασθενειών και από τους γιατρούς για τη διάγνωση των ασθενειών και των συμπτωμάτων σε μεμονωμένους ασθενείς. Επιπλέον, δεδομένα που προέρχονται από ηλεκτρονικά αρχεία υγείας, κοινωνικά μέσα ενημέρωσης, ίντερνετ και άλλες πηγές παρέχουν στους οργανισμούς υγειονομικής περίθαλψης και στις κυβερνητικές υπηρεσίες ενημερωμένες πληροφορίες σχετικά με απειλές ή εκδηλώσεις μολυσματικών ασθενειών.

Στον ενεργειακό κλάδο, τα μεγάλα δεδομένα βοηθούν τις εταιρείες πετρελαίου και φυσικού αερίου να εντοπίζουν πιθανές τοποθεσίες γεωτρήσεων και να παρακολουθούν τις εργασίες των αγωγών. Ομοίως, οι επιχειρήσεις κοινής ωφελείας τα χρησιμοποιούν για την παρακολούθηση των ηλεκτρικών δικτύων. Οι επιχειρήσεις χρηματοπιστωτικών υπηρεσιών χρησιμοποιούν συστήματα μεγάλων δεδομένων για τη διαχείριση του κινδύνου και την ανάλυση σε πραγματικό χρόνο των δεδομένων της αγοράς. Οι κατασκευαστές και οι εταιρείες μεταφορών βασίζονται σε μεγάλα δεδομένα για να διαχειρίζονται τις αλυσίδες εφοδιασμού τους και να βελτιστοποιούν τις διαδρομές παράδοσης. Άλλες κυβερνητικές χρήσεις περιλαμβάνουν την αντιμετώπιση έκτακτης ανάγκης, την πρόληψη του εγκλήματος και τις πρωτοβουλίες για τη δημιουργία έξυπνων πόλεων.

1.1.3 Κατηγοριοποίηση Μεγάλων Δεδομένων

Τα μεγάλα δεδομένα ταξινομούνται σε διάφορες κατηγορίες για την καλύτερη κατανόηση των χαρακτηριστικών τους. Το παρακάτω σχήμα δείχνει τις πολυάριθμες κατηγορίες μεγάλων δεδομένων. Η ταξινόμηση βασίζεται σε πέντε πτυχές: (i) πηγές δεδομένων, (ii) μορφή περιεχομένου, (iii) αποθήκες δεδομένων, (iv) ιεράρχηση δεδομένων και (v) επεξεργασία δεδομένων.



Εικόνα 3 - Κατηγοριοποίηση Μεγάλων Δεδομένων

(i) Πηγές δεδομένων

- **Κοινωνικά μέσα (Social Media)**

Τα κοινωνικά μέσα είναι μια πηγή πληροφοριών που δημιουργείται μέσω μιας διεύθυνσης URL για να μοιράζονται ή να ανταλλάσσονται πληροφορίες και ιδέες σε εικονικές κοινότητες και δίκτυα, όπως Facebook και Twitter.

- **Μηχανικά Δεδομένα (Machine-generated data)**

Τα μηχανικά δεδομένα είναι πληροφορίες που παράγονται αυτόματα από υλικό ή λογισμικό, όπως υπολογιστές, ιατρικές συσκευές ή άλλα μηχανήματα, χωρίς ανθρώπινη παρέμβαση.

- **Δεδομένα Ανίχνευσης (Sensing)**

Υπάρχουν αρκετές μηχανές ανίχνευσης για την μέτρηση φυσικών μεγεθών και για την μετατροπή τους σε σήματα.

- **Συναλλαγές (Transactions)**

Τα δεδομένα συναλλαγών, όπως είναι τα οικονομικά δεδομένα ή τα δεδομένα μιας εργασίας, περιέχουν μια χρονική διάσταση για την περιγραφή των δεδομένων.

- **Διαδίκτυο των πραγμάτων (IoT)**

Το IoT αντιπροσωπεύει ένα σύνολο αντικειμένων που αναγνωρίζονται με μοναδικό τρόπο ως μέρος του Διαδικτύου. Αυτά τα αντικείμενα περιλαμβάνουν smartphones, ψηφιακές φωτογραφικές μηχανές και tablet. Όταν αυτές οι συσκευές συνδέονται μεταξύ τους μέσω του Διαδικτύου, επιτρέπουν πιο έξυπνες διαδικασίες και υπηρεσίες που υποστηρίζουν οικονομικές, περιβαλλοντικές και υγειονομικές ανάγκες. Ένας μεγάλος αριθμός συσκευών που συνδέονται με το Διαδίκτυο παρέχει πολλούς τύπους υπηρεσιών και παράγει τεράστιες ποσότητες δεδομένων και πληροφοριών [1].

(ii) Μορφή περιεχομένου

- **Δομημένο (Structured)**

Τα δομημένα δεδομένα χρησιμοποιούν κυρίως SQL, μια γλώσσα προγραμματισμού που δημιουργήθηκε για να διαχειρίζεται και να εκτελεί ερωτήματα προς τα δεδομένα σε σχεσιακές βάσεις δεδομένων. Τα δομημένα δεδομένα είναι εύκολο να εισαχθούν, να ερωτηθούν, να αποθηκευτούν και να αναλυθούν. Παραδείγματα δομημένων δεδομένων είναι οι αριθμοί, οι λέξεις και οι ημερομηνίες.

- **Ημι-δομημένο (Semi-structured)**

Τα ημιδομημένα δεδομένα είναι δεδομένα που δεν ακολουθούν μια συμβατική βάση δεδομένων. Τα ημιδομημένα δεδομένα μπορεί να έχουν τη μορφή δομημένων δεδομένων που δεν είναι οργανωμένα σε μοντέλα σχεσιακών βάσεων, όπως είναι οι πίνακες. Η καταγραφή ημιδομημένων δεδομένων για ανάλυση είναι διαφορετική από τη λήψη μιας σταθερής μορφής αρχείου. Επομένως, η λήψη ημιδομημένων δεδομένων απαιτεί τη χρήση περίπλοκων κανόνων που αποφασίζουν δυναμικά για την επόμενη διαδικασία μετά την λήψη των δεδομένων [2].

- **Αδόμητο (Unstructured)**

Τα μη δομημένα δεδομένα, όπως είναι τα μηνύματα κειμένου, οι πληροφορίες τοποθεσίας, τα βίντεο και τα δεδομένα κοινωνικών μέσων, είναι δεδομένα που δεν ακολουθούν μια καθορισμένη μορφή. Λαμβάνοντας υπόψη ότι το μέγεθος αυτού του τύπου δεδομένων συνεχίζει να αυξάνεται μέσω της χρήσης smartphones, η ανάγκη να αναλυθούν και να κατανοηθούν τέτοια δεδομένα έχει γίνει μια πρόκληση.

(iii) Αποθήκες δεδομένων

- **Προσανατολισμένες σε έγγραφα (Document-oriented)**

Οι αποθήκες δεδομένων που είναι προσανατολισμένες σε έγγραφα είναι κυρίως σχεδιασμένες για να αποθηκεύουν και να ανακτούν συλλογές εγγράφων ή πληροφοριών και να υποστηρίζουν σύνθετες μορφές δεδομένων σε διάφορους τύπους, όπως JSON, XML και δυαδικές μορφές (π.χ. PDF και MS Word). Οι αποθήκες δεδομένων που είναι προσανατολισμένες σε έγγραφα είναι παρόμοιες με μια εγγραφή ή μια γραμμή σε μια σχεσιακή βάση δεδομένων, αλλά είναι πιο ευέλικτες και μπορούν να ανακτήσουν έγγραφα βάσει του περιεχομένου τους (π.χ. MongoDB, SimpleDB και CouchDB).

- **Προσανατολισμένες σε στήλες (Column-oriented)**

Μια βάση δεδομένων με προσανατολισμό στις στήλες αποθηκεύει το περιεχόμενό της σε στήλες εκτός από τις γραμμές, με τις τιμές χαρακτηριστικών που ανήκουν στην ίδια στήλη να αποθηκεύονται συνεχόμενα. Η προσανατολισμένη στη στήλη είναι διαφορετική από τα κλασικά συστήματα βάσεων δεδομένων που αποθηκεύουν ολόκληρες σειρές μία μετά την άλλη, όπως το BigTable [3].

- **Βάση δεδομένων γραφημάτων (Graph database)**

Μια βάση δεδομένων γραφημάτων, όπως είναι η Neo4j, είναι σχεδιασμένη να αποθηκεύει και να αντιπροσωπεύει δεδομένα που χρησιμοποιούν ένα μοντέλο γραφημάτων με κόμβους, άκρες και ιδιότητες που σχετίζονται μεταξύ τους μέσω συσχετίσεων [3].

- **Βάση δεδομένων κλειδιών-τιμών (Key-value)**

Μια βάση δεδομένων κλειδιών-τιμών είναι ένα εναλλακτικό σύστημα σχεσιακής βάσης δεδομένων που αποθηκεύει και αποκτά πρόσβαση σε δεδομένα που έχουν σχεδιαστεί να φτάσουν πολύ μεγάλο μέγεθος [3]. Το Dynamo [3] αποτελεί καλό παράδειγμα μιας βάσης δεδομένων κλειδιών-τιμών.

Χρησιμοποιείται από το amazon σε ορισμένες από τις υπηρεσίες του. Άλλα παραδείγματα είναι τα Apache Hbase , Apache Cassandra και Voldemort. Το Hbase χρησιμοποιεί το HDFS, μια έκδοση ανοιχτού κώδικα του BigTable της Google που είναι χτισμένη στην Cassandra. Το Hbase αποθηκεύει δεδομένα σε πίνακες, σειρές και κελιά. Οι σειρές ταξινομούνται ανά κλειδί γραμμής και κάθε κελί σε έναν πίνακα καθορίζεται από ένα κλειδί γραμμής, ένα κλειδί στηλών και μια έκδοση, με το περιεχόμενο να περιέχεται ως μια μη ερμηνευμένη σειρά από bytes.

(vi) Ιεράρχηση δεδομένων

- **Καθαρισμός (Cleaning)**

Ο καθαρισμός είναι η διαδικασία αναγνώρισης ελλιπών και παράλογων δεδομένων.

- **Μετασχηματισμός (Transform)**

Ο μετασχηματισμός είναι η διαδικασία μετατροπής των δεδομένων σε μορφή κατάλληλη για ανάλυση.

- **Κανονικοποίηση (Normalization)**

Η κανονικοποίηση είναι η μέθοδος δόμησης της διάταξης βάσης δεδομένων για την ελαχιστοποίηση του πλεονάσματος.

(v) Επεξεργασία δεδομένων

- **Κατά δεσμίδες (Batch)**

Τα συστήματα βασισμένα στο MapReduce έχουν υιοθετηθεί από πολλούς οργανισμούς τα τελευταία χρόνια για εργασίες μακράς διάρκειας επεξεργασίας δεδομένων κατά δεσμίδες [4]. Ένα τέτοιο σύστημα επιτρέπει την κλιμάκωση των εφαρμογών σε μεγάλες συστάδες μηχανών που περιλαμβάνουν χιλιάδες κόμβους.

- **Σε πραγματικό χρόνο (Real Time)**

Ένα από τα πιο διάσημα και ισχυρά εργαλεία επεξεργασίας μεγάλων δεδομένων σε πραγματικό χρόνο είναι το απλό κλιμακωτό σύστημα ροής δεδομένων S4 [5]. Το S4 είναι μια κατανεμημένη πλατφόρμα υπολογιστών που επιτρέπει στους προγραμματιστές να αναπτύσσουν εύκολα εφαρμογές για την επεξεργασία συνεχόμενων απεριόριστων ροών δεδομένων.

1.2 Μεγάλα Δεδομένα: Αποθήκευση, Επεξεργασία και Παραδείγματα

1.2.1 Πως αποθηκεύονται και επεξεργάζονται τα Μεγάλα Δεδομένα

Ο χειρισμός της μεγάλης ταχύτητας και του όγκου των δεδομένων προϋποθέτει να πληρούνται κάποιες απαιτήσεις στην υπολογιστική υποδομή. Η υπολογιστική ισχύς που απαιτείται για την γρήγορη επεξεργασία τεράστιων όγκων και ποκίλων δεδομένων μπορεί να μην υποστηρίζεται από ένα μόνο διακομιστή ή ένα σύμπλεγμα διακομιστών. Οι οργανισμοί πρέπει να έχουν επαρκή χωρητικότητα για εργασίες μεγάλων δεδομένων, προκειμένου να επιτευχθεί η απαιτούμενη ταχύτητα. Αυτό μπορεί δυνητικά να απαιτήσει εκατοντάδες ή χιλιάδες διακομιστές που να μπορούν να διανείμουν τις εργασίες και να λειτουργούν σε συνεργασία σε μια αρχιτεκτονική cluster, συχνά βασισμένη σε τεχνολογίες όπως Hadoop και Apache Spark.

Η επίτευξη τέτοιων ταχυτήτων με οικονομικά αποδοτικό τρόπο είναι επίσης μια πρόκληση. Πολλοί ηγέτες επιχειρήσεων είναι πρόθυμοι να επενδύσουν σε ένα server και μια υποδομή αποθήκευσης που θα υποστηρίζουν τον μεγάλο φόρτο εργασίας από μεγάλα δεδομένα,

ιδιαίτερα εκείνων που δεν λειτουργούν 24/7. Ως αποτέλεσμα, το υπολογιστικό νέφος (cloud) είναι πλέον ένα βασικό μέσο για τη φιλοξενία συστημάτων μεγάλων δεδομένων. Ένας πάροχος cloud μπορεί να αποθηκεύσει petabytes δεδομένων και να αυξήσει τον απαιτούμενο αριθμό διακομιστών για αρκετό καιρό τόσο ώστε να ολοκληρωθεί ένα έργο ανάλυσης μεγάλων δεδομένων. Η επιχείρηση πληρώνει μόνο για τον χώρο αποθήκευσης με βάση το χρόνο που πραγματικά χρησιμοποιείται και οι λειτουργίες cloud μπορούν να απενεργοποιηθούν μέχρι να χρειαστούν ξανά.

Για την περαιτέρω βελτίωση των υπηρεσιών, οι πάροχοι cloud προσφέρουν δυνατότητες μεγάλων δεδομένων μέσω διαχειριζόμενων υπηρεσιών που περιλαμβάνουν τα εξής:

- Amazon EMR (πρώην Elastic MapReduce)
- Microsoft Azure HDInsight
- Google Cloud Dataproc

Σε περιβάλλον cloud, τα μεγάλα δεδομένα μπορούν να αποθηκευτούν στα εξής:

- Διανεμημένο σύστημα αρχείων Hadoop (HDFS)
- Χαμηλού κόστους αποθήκες cloud, όπως η υπηρεσία Amazon Simple Storage Service (S3)
- NoSQL βάσεις δεδομένων και
- Σχεσιακές βάσεις δεδομένων

Για τους οργανισμούς που επιθυμούν να αναπτύξουν συστήματα μεγάλων δεδομένων στο χώρο, οι κοινές τεχνολογίες ανοιχτού κώδικα της Apache εκτός από το Hadoop και το Spark περιλαμβάνουν ακόμη το Yet Another Resource Negotiator (YARN) - τον ενσωματωμένο διαχειριστή πόρων και τον προγραμματιστή εργασίας του Hadoop-, το πλαίσιο προγραμματισμού MapReduce, το Kafka - πλατφόρμα μηνυμάτων και ροής δεδομένων, τη βάση δεδομένων HBase. και μηχανές αναζήτησης SQL-on-Hadoop όπως Drill, Hive, Impala και Presto.

Οι χρήστες μπορούν να εγκαταστήσουν τις εκδόσεις ανοικτού κώδικα των τεχνολογιών τους ή να στραφούν σε εμπορικές πλατφόρμες μεγάλων δεδομένων που προσφέρονται από τη Cloudera, η οποία συγχωνεύθηκε με τον πρώην αντίπαλο Hortonworks τον Ιανουάριο του 2019, ή τη Hewlett Packard Enterprise (HPE), η οποία αγόρασε τα περιουσιακά στοιχεία του μεγάλου προμηθευτή δεδομένων MapR Technologies τον Αύγουστο του 2019. Οι πλατφόρμες Cloudera και MapR υποστηρίζονται επίσης στο cloud.

1.2.2 Παραδείγματα Μεγάλων Δεδομένων

Τα μεγάλα δεδομένα προέρχονται από πληθώρα διαφορετικών πηγών, όπως συστήματα επιχειρηματικών συναλλαγών, βάσεις δεδομένων πελατών, ιατρικά αρχεία, αρχεία καταγραφής clickstream στο διαδίκτυο, κινητές εφαρμογές, κοινωνικά δίκτυα, αποθετήρια επιστημονικής έρευνας, μηχανοποιημένα δεδομένα και αισθητήρες δεδομένων σε πραγματικό χρόνο που χρησιμοποιούνται στο IoT (internet of things) περιβάλλον. Τα δεδομένα μπορεί να παραμείνουν σε ακατέργαστη μορφή σε συστήματα μεγάλων δεδομένων ή προεπεξεργασμένα χρησιμοποιώντας εργαλεία εξόρυξης δεδομένων ή λογισμικό επεξεργασίας δεδομένων, ώστε να είναι έτοιμα για να κάνουν αναλύσεις οι χρήστες.

Χρησιμοποιώντας ως παράδειγμα δεδομένα πελατών, οι διάφοροι κλάδοι ανάλυσης που μπορούν να πραγματοποιηθούν με τις πληροφορίες που βρίσκονται σε σύνολα μεγάλων δεδομένων είναι οι εξής:

Συγκριτική ανάλυση

Αυτό περιλαμβάνει την εξέταση δεικτών συμπεριφοράς των χρηστών και την παρατήρηση της εμπιστοσύνης των πελατών σε πραγματικό χρόνο προκειμένου να συγκριθούν τα προϊόντα και οι υπηρεσίες μιας εταιρείας με εκείνες του ανταγωνισμού.

Παρατήρηση κοινωνικών μέσων ενημέρωσης

Αυτές είναι πληροφορίες σχετικά με το τι λένε οι άνθρωποι στα κοινωνικά μέσα ενημέρωσης αναφορικά με μια συγκεκριμένη επιχείρηση ή προϊόν, κάτι που δεν μπορεί να αποτυπωθεί σε μια δημοσκόπηση ή μια έρευνα. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για να βοηθήσουν στον εντοπισμό των πελατών για στοχευμένες καμπάνιες μάρκετινγκ, παρατηρώντας τη δραστηριότητα τους σε σχέση με συγκεκριμένα θέματα από διάφορες πηγές.

Ανάλυση μάρκετινγκ

Αυτό περιλαμβάνει πληροφορίες που μπορούν να χρησιμοποιηθούν για την καλύτερη ενημέρωση και καινοτομία στην προώθηση νέων προϊόντων, υπηρεσιών και πρωτοβουλιών.

Ανάλυση ικανοποίησης και συναισθήματος πελατών

Όλες οι πληροφορίες που συλλέγονται μπορούν να αποκαλύψουν πως αισθάνονται οι πελάτες για μια εταιρία ή μάρκα, εάν προκύψουν πιθανά ζητήματα, πώς θα μπορούσε να διατηρηθεί η εμπιστοσύνη στη συγκεκριμένη μάρκα και πώς θα μπορούσαν να βελτιωθούν οι προσπάθειες εξυπηρέτησης πελατών.

1.3 Εισαγωγή στην Εξόρυξη Δεδομένων και Τομείς Επιρροής

Ο τεράστιος πολλαπλασιασμός των βάσεων δεδομένων σε όλες σχεδόν τις περιοχές της ανθρώπινης εργασίας δημιούργησε μεγάλη ζήτηση για νέα, ισχυρά εργαλεία για τη μετατροπή δεδομένων σε χρήσιμες γνώσεις προσανατολισμένες στις εργασίες. Στις προσπάθειες για την ικανοποίηση αυτής της ανάγκης, οι ερευνητές εξετάζουν ιδέες και μεθόδους που αναπτύσσονται στη μηχανική μάθηση, στην αναγνώριση προτύπων, στην ανάλυση στατιστικών δεδομένων, στην οπτικοποίηση δεδομένων, στα νευρωνικά δίκτυα, κ.λ.π. Οι προσπάθειες αυτές έχουν οδηγήσει στην εμφάνιση ενός νέου χώρου έρευνας, που συχνά ονομάζεται εξόρυξη δεδομένων και ανακάλυψη γνώσης.

Η τρέχουσα εποχή της πληροφορίας χαρακτηρίζεται από ραγδαία ανάπτυξη των δεδομένων που παράγονται και αποθηκεύονται για κάθε είδους ανθρώπινη ανάγκη. Ένα μεγάλο ποσοστό αυτών των δεδομένων καταγράφεται με τη μορφή βάσεων δεδομένων στους υπολογιστές ώστε η τεχνολογία των υπολογιστών να μπορεί να έχει εύκολη πρόσβαση σε αυτές. Η διαθεσιμότητα πολύ μεγάλων όγκων τέτοιων δεδομένων δημιούργησε το ερώτημα πώς να εξαχθεί χρήσιμη γνώση με γνώμονα τις εργασίες.

Τεχνικές ανάλυσης δεδομένων που έχουν παραδοσιακά χρησιμοποιηθεί για τέτοιες εργασίες περιλαμβάνουν αναλύσεις παλινδρόμησης, ανάλυση cluster, αριθμητική ταξινόμηση, πολυδιάστατη ανάλυση, άλλες πολυμεταβλητές στατιστικές μεθόδους, στοχαστικά μοντέλα, ανάλυση χρονοσειρών, τεχνικές μη γραμμικής εκτίμησης και άλλες. Αυτές οι τεχνικές έχουν χρησιμοποιηθεί ευρέως για την επίλυση πολλών πρακτικών προβλημάτων. Ωστόσο, κατά κύριο λόγο προσανατολίζονται προς την εξαγωγή ποσοτικών και στατιστικών χαρακτηριστικών των δεδομένων και, ως εκ τούτου, έχουν εγγενείς περιορισμούς.

Για παράδειγμα, μια στατιστική ανάλυση μπορεί να καθορίσει αλληλεπιδράσεις και συσχετισμούς μεταξύ των μεταβλητών στα δεδομένα. Δεν μπορεί, ωστόσο, να εντοπίσει τυχόν εξαρτήσεις σε ένα αφηρημένο, εννοιολογικό επίπεδο και να αποδώσει μια περιστασιακή εξήγηση των λόγων για τους οποίους υπάρχουν αυτές οι εξαρτήσεις, ούτε μπορεί να αναπτύξει μια δικαιολογία για αυτές τις σχέσεις με τη μορφή υψηλού επιπέδου λογικών περιγραφών και νόμων. Μια στατιστική ανάλυση δεδομένων μπορεί να καθορίσει την κεντρική τάση και τη διακύμανση των δοσμένων παραγόντων και μια ανάλυση παλινδρόμησης μπορεί να χωρέσει μια καμπύλη σε ένα σύνολο σημείων. Αυτές οι τεχνικές δεν μπορούν, ωστόσο, να παράγουν μια ποιοτική περιγραφή της κανονικότητας και να καθορίσουν την εξάρτησή τους από παράγοντες που δεν παρέχονται ρητά στα δεδομένα, ούτε μπορούν να συντάξουν μια αναλογία μεταξύ της ανακαλυφθείσας κανονικότητας και της κανονικότητας σε έναν άλλο τομέα.

Μια τεχνική αριθμητικής ταξινόμησης μπορεί να δημιουργήσει μια ταξινόμηση των οντοτήτων και να εντοπίσει μια αριθμητική ομοιότητα μεταξύ των οντοτήτων που υπάρχουν στην ίδια ή σε διαφορετικές κατηγορίες. Ωστόσο, δεν μπορεί να δημιουργήσει ποιοτική περιγραφή των τάξεων (classes) που δημιουργήθηκαν και των λόγων για τους οποίους οι οντότητες είναι στην ίδια κατηγορία. Χαρακτηριστικά που ορίζουν την ομοιότητα, καθώς και τα μέτρα ομοιότητας, πρέπει να ορίζονται από έναν αναλυτή εκ των προτέρων.

Για την αντιμετώπιση τέτοιων εργασιών όπως αυτές που αναφέρονται παραπάνω, πρέπει ένα σύστημα ανάλυσης δεδομένων να είναι εξοπλισμένο με σημαντικό υπόβαθρο και να είναι σε θέση να εκτελέσει καθήκοντα που περιλαμβάνουν αυτές τις γνώσεις και τα δεδομένα. Συνοψίζοντας, οι παραδοσιακές τεχνικές ανάλυσης δεδομένων διευκολύνουν τις χρήσιμες ερμηνείες των δεδομένων και μπορούν να συμβάλλουν στη δημιουργία σημαντικών ανακαλύψεων για τις διαδικασίες πίσω από τα δεδομένα. Αυτές οι ερμηνείες και οι ανακαλύψεις είναι οι ύστατες γνώσεις που αναζητούν αυτοί που κατασκευάζουν βάσεις δεδομένων. Ωστόσο, οι γνώσεις δεν δημιουργούνται από αυτά τα εργαλεία, αλλά αντίθετα, πρέπει να προκύψουν από την ανθρώπινη ανάλυση των δεδομένων.

Στις προσπάθειες να ικανοποιηθεί η αυξανόμενη ανάγκη για νέα εργαλεία ανάλυσης δεδομένων που θα μπορούν να ξεπεράσουν τους παραπάνω περιορισμούς, οι ερευνητές στράφηκαν σε ιδέες και μεθόδους που αναπτύχθηκαν στη μηχανική μάθηση. Το πεδίο της μηχανικής μάθησης είναι μια φυσική πηγή ιδεών για το σκοπό αυτό, διότι η ουσία της έρευνας στον τομέα αυτό είναι η ανάπτυξη υπολογιστικών μοντέλων για την απόκτηση γνώσεων από γεγονότα και γνώση στο background. Αυτές οι προσπάθειες έχουν οδηγήσει στην εμφάνιση ενός νέου χώρου έρευνας, που ονομάζεται ***εξόρυξη δεδομένων (data mining)*** και ***ανακάλυψη γνώσεων (knowledge discovery-KDD)***.

Υπάρχει σύγχυση σχετικά με την ακριβή έννοια των όρων "εξόρυξη δεδομένων" και "KDD." Το KDD προτάθηκε το 1995 για να περιγράψει την όλη διαδικασία εξόρυξης της γνώσης από τα δεδομένα [6]. Σε αυτό το πλαίσιο, η γνώση σημαίνει σχέσεις και πρότυπα μεταξύ των δεδομένων. Η "εξόρυξη δεδομένων" πρέπει να χρησιμοποιείται αποκλειστικά για το στάδιο ανακάλυψης της διαδικασίας KDD.

1.3.1 Τι είναι η Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων ή η ανακάλυψη της γνώσης είναι η διαδικασία, όπου με τη βοήθεια υπολογιστή, πραγματοποιείται η αναζήτηση και ανάλυση τεράστιων συνόλων δεδομένων και στη συνέχεια η εξαγωγή της σημασίας των δεδομένων αυτών. Τα εργαλεία εξόρυξης δεδομένων προβλέπουν συμπεριφορές και μελλοντικές τάσεις, επιτρέποντας στις επιχειρήσεις να παίρνουν προληπτικές αποφάσεις που βασίζονται στη γνώση. Τα εργαλεία εξόρυξης δεδομένων μπορούν να απαντήσουν σε επιχειρηματικές ερωτήσεις που ήταν παραδοσιακά πολύ χρονοβόρες ώστε να επιλυθούν. Αναζητούν τις βάσεις δεδομένων για

κρυμμένα μοτίβα, βρίσκοντας πληροφορίες από προβλέψεις που οι εμπειρογνώμονες μπορεί να χάσουν επειδή βρίσκονται εκτός των προσδοκιών τους.

Η εξόρυξη δεδομένων πήρε το όνομά της από τις ομοιότητες μεταξύ της αναζήτησης πολύτιμων πληροφοριών σε μια μεγάλη βάση δεδομένων και της εξόρυξης ενός βουνού για ένα πολύτιμο αγαθό. Και οι δύο διαδικασίες απαιτούν είτε ξεκαθάρισμα μέσα από μια τεράστια ποσότητα υλικού, είτε ανίχνευση της αξίας με έξυπνο τρόπο.

1.3.2 Τομείς Επιρροής των Τεχνολογιών Εξόρυξης Δεδομένων

Οι τεχνικές εξόρυξης δεδομένων έχουν εφαρμοστεί με επιτυχία σε πολλούς τομείς, από την επιχειρηματικότητα μέχρι την επιστήμη και τον αθλητισμό. Η εξόρυξη δεδομένων έχει χρησιμοποιηθεί σε βάσεις δεδομένων μάρκετινγκ, σε αναλύσεις δεδομένων λιανικής, σε εγκρίσεις πιστώσεων κ.λπ. Τεχνικές εξόρυξης δεδομένων έχουν χρησιμοποιηθεί στην αστρονομία, τη μοριακή βιολογία, την ιατρική, τη γεωλογία, και πολλά άλλα πεδία. Έχει επίσης χρησιμοποιηθεί στη διαχείριση της υγειονομικής περίθαλψης, στην ανίχνευση φορολογικής απάτης, στην παρακολούθηση της νομιμοποίησης εσόδων από παράνομες δραστηριότητες και στον αθλητισμό.

Διαχείριση της αγοράς (Market Management)

Το στοχευμένο μάρκετινγκ, η διαχείριση των σχέσεων με τους πελάτες, η ανάλυση του καλαθιού αγοράς, οι διασταυρούμενες πωλήσεις, η κατάτμηση της αγοράς.

Διαχείριση κινδύνου (Risk Management)

Η πρόβλεψη, η διατήρηση πελατών, η βελτιωμένη ασφάλεια, ο ποιοτικός έλεγχος, η ανταγωνιστική ανάλυση.

Διαχείριση της απάτης (Fraud management)

Ανίχνευση απάτης.

Βιομηχανικές εφαρμογές (Industrial-specific applications)

Ανάλυση κέρδους (για κάθε εργαζόμενο υποκαταστήματος, προϊόν, ομάδα προϊόντων, παρακολούθηση προγραμμάτων και καναλιών μάρκετινγκ, ανάλυση δεδομένων πελάτη και ανάλυση προφίλ πελάτη).

Τηλεπικοινωνίες και μέσα ενημέρωσης (Telecommunications and media)

Βαθμολογία απόκρισης, διαχείριση εκστρατειών μάρκετινγκ, ανάλυση κερδοφορίας και διαχωρισμός πελατών.

Υγειονομική περίθαλψη (Health care)

Ανάπτυξη του Συστήματος Διαχείρισης Απάτης και Κατάχρησης (FAMS) το οποίο βοηθά τους ασφαλιστικούς οργανισμούς υγείας που ασχολούνται με την απάτη και την κατάχρηση: ανίχνευση, έρευνα, διευθέτηση, πρόληψη της υποτροπής.

Νέες εφαρμογές

Η πειθαρχία της εξόρυξης δεδομένων οδηγείται εν μέρει από νέες εφαρμογές, οι οποίες απαιτούν νέες δυνατότητες που δεν παρέχονται επί του παρόντος από τη σημερινή τεχνολογία. Αυτές οι νέες εφαρμογές μπορούν φυσικά να χωριστούν σε τρεις ευρείες κατηγορίες:

Δεδομένα επιχειρήσεων και ηλεκτρονικού εμπορίου (Business & E-commerce Data)

Το back office, το front office και οι εφαρμογές δικτύου παράγουν μεγάλα ποσά δεδομένων για τους σκοπούς των επιχειρήσεων. Η χρησιμοποίηση αυτών των δεδομένων για αποτελεσματική λήψη αποφάσεων παραμένει μια θεμελιώδης πρόκληση.

Επιστημονικά, Μηχανικά και Υγειονομικά Στοιχεία (Scientific, Engineering, and Health Care Data)

Επιστημονικά δεδομένα και μετα-δεδομένα τείνουν να είναι πιο σύνθετα στη δομή από τα δεδομένα των επιχειρήσεων. Επιπλέον, οι επιστήμονες και οι μηχανικοί χρησιμοποιούν όλο και περισσότερο την προσομοίωση και τα συστήματα με γνώση στον τομέα των εφαρμογών.

Δεδομένα Ιστού (Web data)

Τα δεδομένα στο διαδίκτυο αυξάνονται όχι μόνο σε ένταση, αλλά επίσης σε πολυπλοκότητα. Τα δεδομένα Ιστού περιλαμβάνουν τώρα όχι μόνο κείμενο και εικόνα, αλλά επίσης και ροή δεδομένων και αριθμητικά δεδομένα.

Στη παρακάτω ενότητα περιγράφουμε πολλές από αυτές τις εφαρμογές από κάθε κατηγορία.

Εμπορικές συναλλαγές (Business transactions)

Σήμερα, οι επιχειρήσεις εξαπλώνονται και όλο και περισσότερες επιχειρήσεις έχουν εκατομμύρια πελάτες και δισεκατομμύρια συναλλαγές.

Πρέπει να κατανοήσουν τους κινδύνους (είναι κάποια συναλλαγή ύποπτη; Θα πληρώσουν οι πελάτες τους λογαριασμούς;) και τις ευκαιρίες (ποιο είναι το αναμενόμενο κέρδος από αυτούς τους πελάτες? Ποιο προϊόν είναι πιο πιθανό να αγοράσει αυτός ο πελάτης).

Ηλεκτρονικό εμπόριο (Electronic commerce)

Το ηλεκτρονικό εμπόριο όχι μόνο παράγει σύνολα μεγάλων δεδομένων στα οποία η ανάλυση των προτύπων μάρκετινγκ και των προτύπων κινδύνων είναι κρίσιμη, αλλά σε αντίθεση με ορισμένες από τις παραπάνω εφαρμογές, είναι επίσης σημαντικό να συμβαίνει αυτό σε πραγματικό ή σχεδόν πραγματικό χρόνο, προκειμένου να ικανοποιηθεί η ζήτηση των συναλλαγών on-line.

Γονιδιωματικά δεδομένα (Genomic data)

Η γονιδιωματική αλληλουχία και οι προσπάθειες χαρτογράφησης έχουν δημιουργήσει μια σειρά βάσεων δεδομένων, οι οποίες είναι προσβάσιμες μέσω του Διαδικτύου. Επιπλέον, υπάρχει επίσης μια ευρεία ποικιλία άλλων on-line βάσεων δεδομένων, συμπεριλαμβανομένων εκείνων που περιέχουν πληροφορίες σχετικά με τη νόσο, την κυτταρική λειτουργία και τα φάρμακα. Εύρεση σχέσης μεταξύ αυτών των πηγών δεδομένων, οι οποίες είναι σε μεγάλο βαθμό ανεξερεύνητες, είναι ένα άλλο θεμελιώδες στοιχείο στη πρόκληση της εξόρυξης γνώσης. Πρόσφατα αναπτύχθηκαν κλιμακούμενες τεχνικές για τη σύγκριση ολόκληρων γονιδιωμάτων.

Δεδομένα αισθητήρα (Sensor data)

Δορυφόροι, σημαδούρες, μπαλόνια και διάφοροι άλλοι αισθητήρες παράγουν ογκώδη ποσά δεδομένων σχετικά με την ατμόσφαιρα της γης, τους ωκεανούς και τα εδάφη. Μια βασική πρόκληση είναι να κατανοήσουμε τις σχέσεις μεταξύ αυτών των δεδομένων. Για παράδειγμα, επηρεάζουν οι βιομηχανικοί ρύποι την παγκόσμια υπερθέρμανση; Υπάρχουν επίσης σύνολα μεγάλων δεδομένων από terabyte μέχρι petabyte που παράγονται από

αισθητήρες και όργανα σε άλλους κλάδους, όπως η αστρονομία, η φυσική και η πυρηνική φυσική.

Δεδομένα προσομοίωσης (Simulation Data)

Η προσομοίωση είναι πλέον αποδεκτή ως ένας τρίτος τρόπος επιστήμης, συμπληρώνοντας τη θεωρία και το πείραμα. Σήμερα, δεν παράγουν μόνο τα πειράματα τεράστια σύνολα δεδομένων, αλλά και οι προσομοιώσεις. Η εξόρυξη δεδομένων και γενικότερα τα δεδομένα των υπολογιστών αποδεικνύεται ότι είναι ένας κρίσιμος σύνδεσμος μεταξύ θεωρίας, προσομοίωσης και πειραματισμού.

Δεδομένα υγειονομικής περίθαλψης (Health care Data)

Η υγειονομική περίθαλψη υπήρξε το ταχύτερα αναπτυσσόμενο κομμάτι του εθνικού ακαθάριστου εγχώριου προϊόντος (ΑΕΠ) για κάποιο χρονικό διάστημα. Τα νοσοκομεία, οι υγειονομικές οργανώσεις, οι ασφαλιστικές εταιρείες και η ομοσπονδιακή κυβέρνηση έχουν τεράστιες συλλογές δεδομένων για τους ασθενείς, τα προβλήματα υγείας τους, τα κλινικά κόστη τους και τα κλινικά αποτελέσματα. Η κατανόηση των συσχετίσεων σε αυτά τα δεδομένα είναι κρίσιμη για μια ευρεία ποικιλία προβλημάτων, που κυμαίνονται από τον προσδιορισμό του ποιες διαδικασίες και ποια κλινικά πρωτόκολλα είναι πιο αποτελεσματικά μέχρι του πως θα παρέχεται με τον καλύτερο δυνατό τρόπο υγειονομική περίθαλψη στους περισσότερους ανθρώπους, σε μια εποχή μείωσης των πόρων.

Έγγραφα πολυμέσων (Multimedia Documents)

Λίγοι άνθρωποι είναι ικανοποιημένοι με τη σημερινή τεχνολογία για την ανάκτηση εγγράφων στον Ιστό, αλλά ο αριθμός των εγγράφων και ο αριθμός των ατόμων που έχουν πρόσβαση σε αυτά τα έγγραφα αυξάνεται εκρηκτικά. Επιπλέον, γίνεται ολοένα και πιο εύκολη η αρχειοθέτηση δεδομένων πολυμέσων, συμπεριλαμβανομένων των δεδομένων ήχου, εικόνων και βίντεο, αλλά πιο δύσκολο να εξαχθεί σημαντική πληροφορία από τα αρχεία καθώς ο όγκος αυξάνεται.

1.4 Εξόρυξη Δεδομένων: Διαδικασία και Τεχνικές

1.4.1 Διαδικασία Εξόρυξης Δεδομένων

Ένα έργο εξόρυξης δεδομένων αποτελείται από έναν κύκλο ζωής που έχει έξι φάσεις. Δεν είναι υποχρεωτικό ότι αυτές οι φάσεις πρέπει να είναι στη σειρά. Μπορεί το έργο είτε να ακολουθήσει την κανονική σειρά των φάσεων είτε να πάει ανάποδα. Η διαδικασία της εξόρυξης δεδομένων συνεχίζεται μέχρι την επίτευξη μιας λύσης.

Τα παρακάτω βήματα βοηθούν να εξηγηθεί κάθε στάδιο της διαδικασίας εξόρυξης δεδομένων:

Επιχειρηματική κατανόηση

Εστιάζει στην κατανόηση των στόχων και των απαιτήσεων του έργου από την επιχειρηματική σκοπιά και στη συνέχεια εντοπίζει τους παράγοντες εκείνους που βοηθούν στην επίτευξη των στόχων.

Κατανόηση δεδομένων

Αυτό το στάδιο εστιάζει στη συλλογή των δεδομένων και την πληθύσμωση εργαλείου που τυχόν χρησιμοποιείται. Τα δεδομένα καταχωρούνται μαζί με τις πληροφορίες σχετικά με την πηγή τους, την τοποθεσία, τον τρόπο με τον οποίο αποκτήθηκαν αλλά και αν αντιμετωπίσθηκε οποιοδήποτε πρόβλημα κατά την συλλογή τους. Τα

δεδομένα οπτικοποιούνται και εκτελούντεαι τα απαραίτητα ερωτήματα (queries) για να ελεγχθεί η πληρότητά τους.

Προετοιμασία δεδομένων

Η φάση της προετοιμασίας των δεδομένων καλύπτει όλες τις δραστηριότητες για να συγκεντρωθούν τα απαραίτητα δεδομένα που θα τροφοδοτούνται σε εργαλεία μοντελοποίησης, από τα αρχικά ακατέργαστα δεδομένα. Σε αυτό το στάδιο οι εργασίες είναι πιθανό να εκτελούνται πολλές φορές και δεν απαιτείται να γίνεται τίποτα κατά σειρά. Οι εργασίες περιλαμβάνουν την επιλογή των κατάλληλων δεδομένων για τα εργαλεία μοντελοποίησης, τον καθαρισμό τους, το μετασχηματισμό τους και την ενσωμάτωση πινάκων και εγγραφών από διάφορες βάσεις δεδομένων.

Μοντελοποίηση

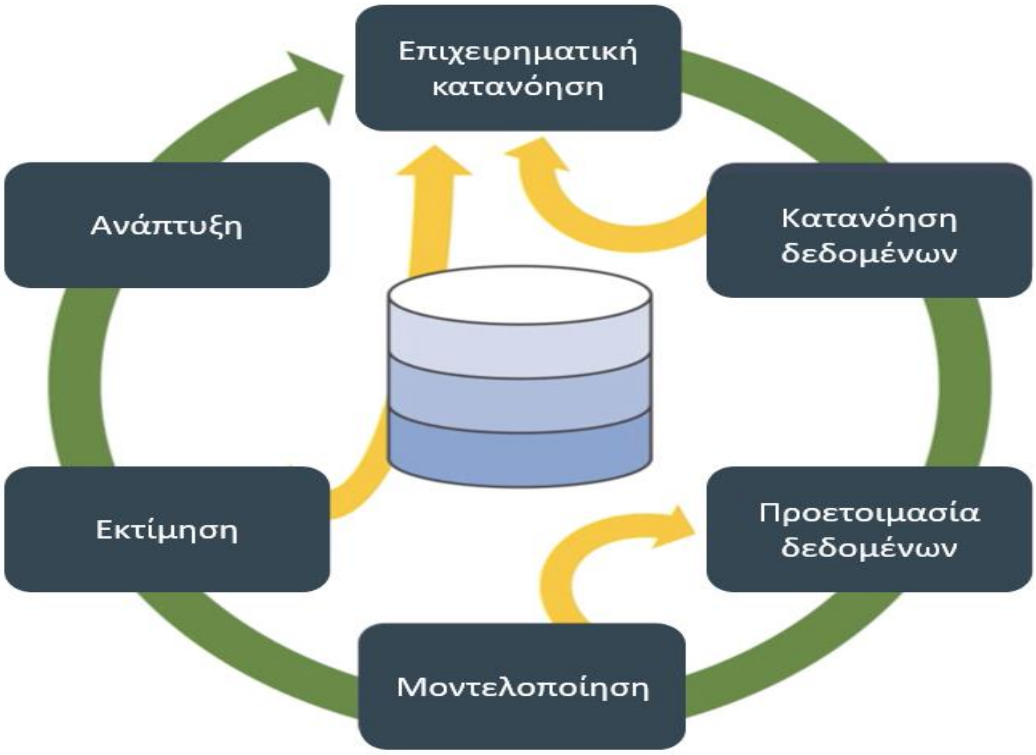
Η επιλογή της τεχνικής εξόρυξης δεδομένων, η δημιουργία δοκιμαστικού σχεδίου για την αξιολόγηση του επιλεγμένου μοντέλου, η δημιουργία μοντέλων από το σύνολο δεδομένων και η αξιολόγηση του κατασκευαζόμενου μοντέλου από εμπειρογνώμονες για συζήτηση του αποτελέσματος, πραγματοποιούνται σε αυτό το βήμα.

Εκτίμηση

Μέσα από αυτό το στάδιο δημιουργείται ένα μοντέλο για το εκάστοτε έργο. Είναι σημαντικό να διεξαχθεί προσεκτικά αξιολόγηση του μοντέλου και να εξετασθούν τα βήματα μέχρι το τέλος για την κατασκευή του μοντέλου. Αυτό μπορεί να βοηθήσει στην ορθή επίτευξη των επιχειρηματικών στόχων. Αυτό είναι ένα αντικειμενικό κλειδί να προσδιοριστεί εάν υπάρχουν κάποια σημαντικά επιχειρηματικά ζητήματα που χρειάζονται να ληφθούν υπόψη. Στο τέλος αυτής της φάσης, αυτή η απόφαση θα βρει τη χρησιμότητα των αποτελεσμάτων εξόρυξης δεδομένων.

Ανάπτυξη

Η δημιουργία του μοντέλου δεν είναι το τέλος του έργου. Ακόμη και αν ο σκοπός του μοντέλου είναι να αυξηθεί η γνώση των δεδομένων, οι γνώσεις που θα αποκτηθούν θα χρειαστεί να οργανωθούν και να παρουσιαστούν με τρόπο που να μπορεί να τις χρησιμοποιήσει ο καταναλωτής. Αυτή η φάση μπορεί να είναι τόσο απλή όσο η δημιουργία μιας έκθεσης ή τόσο σύνθετη όσο η εφαρμογή μιας επαναληπτικής διαδικασίας εξόρυξης δεδομένων. Σε πολλές περιπτώσεις, θα είναι ο πελάτης και όχι ο αναλυτής δεδομένων, αυτός ο οποίος θα εκτελέσει τα βήματα ανάπτυξης. Επομένως, εάν ο αναλυτής δεν πραγματοποιήσει αυτή την ανάπτυξη, είναι σημαντικό να καταβάλει προσπάθεια ώστε οι πελάτες να κατανοήσουν τη δράση που πρόκειται να γίνει με σκοπό την πραγματική χρήση των μοντέλων που δημιουργήθηκαν.



Εικόνα 4 - Διαδικασία Εξόρυξης Δεδομένων

1.4.2 Τεχνικές Εξόρυξης Δεδομένων

Υπάρχουν πολλές τεχνικές εξόρυξης δεδομένων που μπορούν να χρησιμοποιήσουν οι οργανισμοί για να μετατρέψουν τα ανεπεξέργαστα δεδομένα σε ενεργές γνώσεις. Οι τεχνικές αυτές περιλαμβάνουν τα πάντα, από τεχνητή νοημοσύνη αιχμής μέχρι τα βασικά στοιχεία της προετοιμασίας των δεδομένων, τα οποία είναι και τα δύο βασικά για τη μεγιστοποίηση της αξίας των δεδομένων. Ακολούθως θα αναπτυχθούν κάποιες από τις τεχνικές εξόρυξης δεδομένων.

Καθαρισμός και προετοιμασία δεδομένων (Data cleansing and preparation)

Ο καθαρισμός και η προετοιμασία δεδομένων αποτελούν ζωτικό μέρος της διαδικασίας εξόρυξης δεδομένων. Τα ακατέργαστα δεδομένα πρέπει να καθαρίζονται και να μορφοποιούνται ώστε να είναι χρήσιμα σε διάφορες μεθόδους ανάλυσης. Ο καθαρισμός και η προετοιμασία δεδομένων περιλαμβάνουν διάφορα στοιχεία της μοντελοποίησης δεδομένων, του μετασχηματισμού, της μεταφοράς δεδομένων, της ETL, της ELT, της οριστικοποίησης των δεδομένων και της ενσωμάτωσης αυτών. Είναι ένα απαραίτητο βήμα για την κατανόηση των βασικών χαρακτηριστικών και ιδιοτήτων των δεδομένων με σκοπό τον προσδιορισμό της καλύτερης χρήσης τους.

Η επιχειρηματική αξία του καθαρισμού και της προετοιμασίας των δεδομένων είναι αυτονόητη. Χωρίς αυτό το πρώτο βήμα, τα δεδομένα είτε δεν έχουν νόημα σε έναν οργανισμό είτε είναι αναξιόπιστα λόγω της ποιότητάς τους. Οι εταιρείες πρέπει να έχουν τη δυνατότητα να εμπιστεύονται τα δεδομένα τους, τα αποτελέσματα των αναλύσεών τους και τη δράση που δημιουργείται από αυτά τα αποτελέσματα.

Αυτά τα βήματα είναι επίσης απαραίτητα για την ποιότητα των δεδομένων και τη σωστή διαχείρισή τους.

Εντοπισμός προτύπων (Tracking patterns)

Ο εντοπισμός προτύπων είναι μια βασική τεχνική εξόρυξης δεδομένων. Περιλαμβάνει τον εντοπισμό και την παρακολούθηση των τάσεων ή των προτύπων στα δεδομένα, ώστε να βγουν ευφυείς συμπεράσματα σχετικά με τα αποτελέσματα των επιχειρήσεων. Μόλις ένας οργανισμός εντοπίσει μια τάση στα δεδομένα πωλήσεων, για παράδειγμα, υπάρχει μια βάση ώστε να ληφθούν μέτρα και να κεφαλαιοποιηθεί αυτή η γνώση. Εάν διαπιστωθεί ότι ένα συγκεκριμένο προϊόν πουλάει περισσότερο από άλλα για μια συγκεκριμένη δημογραφική ομάδα, ένας οργανισμός μπορεί να χρησιμοποιήσει αυτή τη γνώση για να δημιουργήσει παρόμοια προϊόντα ή υπηρεσίες ή απλά να δημιουργήσει μεγαλύτερο απόθεμα από το αρχικό προϊόν για αυτή την ομάδα.

Ταξινόμηση (Classification)

Οι τεχνικές εξόρυξης δεδομένων ταξινόμησης περιλαμβάνουν την ανάλυση των διαφόρων χαρακτηριστικών που σχετίζονται με διαφορετικούς τύπους δεδομένων. Όταν οι οργανισμοί εντοπίσουν τα κύρια χαρακτηριστικά αυτών των τύπων δεδομένων μπορούν να κατηγοριοποιήσουν ή να ταξινομήσουν τα δεδομένα αυτά. Κάτι τέτοιο είναι κρίσιμο, για παράδειγμα, για τον εντοπισμό προσωπικής πληροφόρησης που οι οργανισμοί μπορεί να θέλουν να προστατεύσουν ή να επεξεργαστούν από έγγραφα.

Σύνδεση (Association)

Η σύνδεση είναι μια τεχνική εξόρυξης δεδομένων που σχετίζεται με τη στατιστική. Υποδεικνύει ότι ορισμένα δεδομένα (ή γεγονότα που βασίζονται στα δεδομένα) συνδέονται με άλλα δεδομένα ή γεγονότα που βασίζονται σε δεδομένα. Είναι παρόμοια με την έννοια της συνύπαρξης στη μηχανική μάθηση, στην οποία η πιθανότητα ενός γεγονότος που βασίζεται σε δεδομένα υποδεικνύεται από την παρουσία άλλου.

Η στατιστική έννοια της σύνδεσης είναι επίσης παρόμοια με την έννοια της συσχέτισης. Αυτό σημαίνει ότι η ανάλυση των δεδομένων δείχνει ότι υπάρχει σχέση μεταξύ δύο γεγονότων που εντοπίζονται στα δεδομένα.

Ανίχνευση ακραίων τιμών (Outlier detection)

Η ανίχνευση των ακραίων τιμών προσδιορίζει τυχόν ανωμαλίες στα σύνολα δεδομένων. Μόλις οι οργανισμοί βρουν αποκλίσεις στα δεδομένα τους, γίνεται ευκολότερο να κατανοήσουν γιατί συμβαίνουν αυτές οι ανωμαλίες και προετοιμάζονται για τυχόν μελλοντικά περιστατικά για την καλύτερη επίτευξη επιχειρηματικών στόχων. Για παράδειγμα, εάν υπάρχει έντονη χρήση των συναλλακτικών συστημάτων για πιστωτικές κάρτες σε μια συγκεκριμένη ώρα της ημέρας, οι οργανισμοί μπορούν να επωφεληθούν από αυτές τις πληροφορίες, υπολογίζοντας γιατί συμβαίνει με σκοπό να βελτιστοποιούν τις πωλήσεις τους κατά τη διάρκεια της υπόλοιπης ημέρας.

Ομαδοποίηση (Clustering)

Η ομαδοποίηση είναι μια τεχνική ανάλυσης που βασίζεται σε οπτικές προσεγγίσεις για την κατανόηση δεδομένων. Οι μηχανισμοί ομαδοποίησης χρησιμοποιούν γραφήματα για να δείξουν πού η κατανομή των δεδομένων σχετίζεται με διαφορετικούς τύπους μετρήσεων. Οι τεχνικές ομαδοποίησης χρησιμοποιούν επίσης διαφορετικά χρώματα για να δείξουν τη διασπορά των δεδομένων.

Με γραφήματα και ομαδοποίηση συγκεκριμένα, οι χρήστες μπορούν οπτικά να δουν πώς διανέμονται τα δεδομένα για να εντοπίζουν τις τάσεις που σχετίζονται με τους επιχειρηματικούς τους στόχους.

Παλινδρόμηση (Regression)

Οι τεχνικές παλινδρόμησης είναι χρήσιμες για τον προσδιορισμό της φύσης της σχέσης μεταξύ των μεταβλητών σε μια ομάδα δεδομένων. Αυτές οι σχέσεις θα μπορούσαν να είναι αιτιώδεις σε ορισμένες περιπτώσεις, ή απλώς να συσχετίζονται με άλλες. Η παλινδρόμηση είναι μια απλή τεχνική που αποκαλύπτει με σαφήνεια τον τρόπο με τον οποίο σχετίζονται οι μεταβλητές. Οι τεχνικές παλινδρόμησης χρησιμοποιούνται σε θέματα πρόβλεψης και μοντελοποίησης δεδομένων.

Πρόβλεψη (Prediction)

Η πρόβλεψη είναι μια πολύ ισχυρή πτυχή της εξόρυξης δεδομένων που αντιπροσωπεύει έναν από τους τέσσερις κλάδους της ανάλυσης. Οι προγνωστικές αναλύσεις χρησιμοποιούν πρότυπα που βρίσκονται σε τρέχοντα ή ιστορικά δεδομένα για να τα επεκτείνουν στο μέλλον. Έτσι, δίνει στους οργανισμούς τη δυνατότητα να δουν ποιες τάσεις θα ακολουθήσουν στη συνέχεια στα δεδομένα τους. Υπάρχουν πολλές διαφορετικές προσεγγίσεις για τη χρήση των προγνωστικών αναλύσεων. Ορισμένες από τις πιο προηγμένες περιλαμβάνουν πτυχές της μηχανικής μάθησης και της τεχνητής νοημοσύνης. Ωστόσο, οι προγνωστικές αναλύσεις δεν εξαρτώνται υποχρεωτικά από αυτές τις τεχνικές και μπορούν επίσης να διευκολυνθούν με πιο απλούς αλγόριθμους.

Διαδοχικά μοτίβα (Sequential patterns)

Αυτή η τεχνική εξόρυξης δεδομένων επικεντρώνεται στην αποκάλυψη γεγονότων που πραγματοποιούνται σε σειρά. Είναι ιδιαίτερα χρήσιμη για εξόρυξη δεδομένων συναλλαγών. Για παράδειγμα, αυτή η τεχνική μπορεί να αποκαλύψει ποια είδη ένδυσης είναι πιο πιθανό να αγοράσουν οι πελάτες μετά από μια αρχική αγορά, για παράδειγμα, κοστουμιού. Η κατανόηση των διαδοχικών προτύπων μπορεί να βοηθήσει τους οργανισμούς να προτείνουν επιπλέον στοιχεία στους πελάτες για να ωθήσουν τις πωλήσεις.

Δέντρα αποφάσεων (Decision trees)

Τα δέντρα αποφάσεων είναι ένας συγκεκριμένος τύπος προγνωστικού μοντέλου που επιτρέπει στους οργανισμούς να διοχετεύουν δεδομένα αποτελεσματικά. Από τεχνική άποψη, ένα δέντρο αποφάσεων αποτελεί μέρος της μηχανικής μάθησης, αλλά είναι ευρύτερα γνωστό ως τεχνική μηχανικής μάθησης λευκού κουτιού (white box machine learning technique) λόγω της εξαιρετικά απλής φύσης της.

Ένα δέντρο αποφάσεων επιτρέπει στους χρήστες να κατανοούν με σαφήνεια τον τρόπο με τον οποίο οι εισροές δεδομένων επηρεάζουν τις εξόδους. Όταν συνδυάζονται διάφορα μοντέλα δέντρων αποφάσεων, δημιουργούν προγνωστικά μοντέλα ανάλυσης γνωστά ως τυχαία δάση (random forest). Τα περίπλοκα τυχαία μοντέλα δάσους θεωρούνται τεχνικές μηχανικής μάθησης μαύρου κουτιού (black box machine learning techniques), επειδή δεν είναι πάντα εύκολο να κατανοηθούν οι έξοδοί τους με βάση τις εισόδους τους. Στις περισσότερες περιπτώσεις, ωστόσο, αυτή η βασική μορφή της μοντελοποίησης του συνόλου είναι ακριβέστερη από τη χρήση δέντρων αποφάσεων από μόνα τους.

Στατιστικές τεχνικές (Statistical techniques)

Οι στατιστικές τεχνικές βρίσκονται στον πυρήνα των περισσότερων αναλυτών που εμπλέκονται στη διαδικασία εξόρυξης δεδομένων. Τα διαφορετικά μοντέλα ανάλυσης βασίζονται σε στατιστικές έννοιες, οι οποίες εξάγουν αριθμητικές τιμές που εφαρμόζονται σε συγκεκριμένους επιχειρηματικούς στόχους. Για παράδειγμα, τα νευρωνικά δίκτυα χρησιμοποιούν σύνθετες στατιστικές βασισμένες σε διαφορετικά βάρη και μέτρα για να καθορίσουν αν μια εικόνα είναι σκύλος ή γάτα σε συστήματα αναγνώρισης εικόνων.

Τα στατιστικά μοντέλα αντιπροσωπεύουν έναν από τους δύο κύριους κλάδους της τεχνητής νοημοσύνης. Τα μοντέλα για ορισμένες στατιστικές τεχνικές είναι στατικά, ενώ άλλα που αφορούν τη μηχανική μάθηση βελτιώνονται με το χρόνο.

Οπτικοποίηση (Visualization)

Οι οπτικοποιήσεις δεδομένων είναι ένα άλλο σημαντικό στοιχείο της εξόρυξης δεδομένων. Παρέχουν στους χρήστες πληροφορίες για τα δεδομένα που γίνονται κατανοητές και με τη χρήση της αισθήσεως της όρασης από το δέκτη. Οι οπτικοποιήσεις των σημερινών δεδομένων είναι δυναμικές, χρήσιμες για τη ροή δεδομένων σε πραγματικό χρόνο και χαρακτηρίζονται από διαφορετικά χρώματα που αποκαλύπτουν διαφορετικές τάσεις και πρότυπα στα δεδομένα.

Τα Dashboards είναι ένας ισχυρός τρόπος για να χρησιμοποιηθούν οι οπτικοποιήσεις δεδομένων για να αποκαλυφθούν οι πληροφορίες εξόρυξης δεδομένων. Οι οργανισμοί μπορούν να βασίζουν τα dashboards σε διαφορετικές μετρήσεις και να χρησιμοποιήσουν οπτικοποιήσεις για να αποδώσουν οπτικά τα μοτίβα στα δεδομένα, αντί να χρησιμοποιούν απλώς αριθμητικές εξόδους στατιστικών μοντέλων.

Νευρωνικά δίκτυα (Neural networks)

Ένα νευρωνικό δίκτυο είναι ένας συγκεκριμένος τύπος μοντέλου μηχανικής μάθησης που χρησιμοποιείται συχνά με τη τεχνητή νοημοσύνη και τη βαθιά εκμάθηση. Πήραν το όνομά τους από το γεγονός ότι έχουν διαφορετικά στρώματα που μοιάζουν με τον τρόπο που εργάζονται οι νευρώνες στον ανθρώπινο εγκέφαλο. Τα νευρωνικά δίκτυα είναι από τα πιο ακριβή μοντέλα μηχανικής μάθησης που χρησιμοποιούνται σήμερα.

Αν και ένα νευρωνικό δίκτυο μπορεί να είναι ένα ισχυρό εργαλείο στην εξόρυξη δεδομένων, οι οργανισμοί θα πρέπει να προσέχουν όταν το χρησιμοποιούν, γιατί ορισμένα από αυτά τα μοντέλα νευρωνικών δικτύων είναι εξαιρετικά πολύπλοκα, γεγονός που δυσχεραίνει την κατανόηση του τρόπου με τον οποίο ένα νευρωνικό δίκτυο καθορίζει μια έξοδο - αποτέλεσμα.

Αποθήκευση δεδομένων (Data warehousing)

Η αποθήκευση δεδομένων αποτελεί σημαντικό μέρος της διαδικασίας εξόρυξης δεδομένων. Παραδοσιακά, η αποθήκευση δεδομένων περιλάμβανε την αποθήκευση δομημένων δεδομένων σε συστήματα διαχείρισης σχεσιακών βάσεων, έτσι ώστε να μπορούν να αναλυθούν για την επιχειρησιακή νοημοσύνη, για αναφορές και για βασικά dashboards. Σήμερα, υπάρχουν αποθήκες δεδομένων στο cloud και αποθήκες δεδομένων σε ημι-δομημένα και αδόμητα εργαλεία δεδομένων όπως το Hadoop. Ενώ οι αποθήκες δεδομένων χρησιμοποιούνταν παραδοσιακά για ιστορικά δεδομένα, πολλές σύγχρονες προσεγγίσεις μπορούν να παρέχουν μια εις βάθος ανάλυση δεδομένων σε πραγματικό χρόνο.

Μακροπρόθεσμη επεξεργασία μνήμης (Long-term memory processing)

Η μακροπρόθεσμη επεξεργασία μνήμης αναφέρεται στην ικανότητα ανάλυσης δεδομένων σε παρατεταμένες χρονικές περιόδους. Τα ιστορικά δεδομένα που είναι αποθηκευμένα σε αποθήκες δεδομένων είναι χρήσιμα για το σκοπό αυτό. Όταν ένας οργανισμός μπορεί να εκτελέσει αναλύσεις για μια εκτεταμένη χρονική περίοδο, είναι σε θέση να εντοπίσει μοτίβα τα οποία διαφορετικά μπορεί να είναι πολύ ανεπαίσθητα για να ανιχνευτούν. Για παράδειγμα, αν αναλυθεί η φθορά σε μια περίοδο αρκετών ετών, ένας οργανισμός μπορεί να βρει ενδείξεις που θα μπορούσαν να οδηγήσουν στη μείωση του χρέους στη χρηματοδότηση.

Μηχανική μάθηση και τεχνητή νοημοσύνη (Machine learning and artificial intelligence)

Η μηχανική μάθηση και η τεχνητή νοημοσύνη (AI) αποτελούν μερικές από τις πιο προηγμένες εξελίξεις στην εξόρυξη δεδομένων. Οι προηγμένες μορφές μηχανικής μάθησης, όπως η βαθιά εκμάθηση (deep learning), προσφέρουν εξαιρετικά ακριβείς προβλέψεις όταν εργάζονται με δεδομένα σε κλίμακα. Ως εκ τούτου, είναι χρήσιμες για την επεξεργασία δεδομένων σε εφαρμογές AI όπως η όραση του υπολογιστή, η αναγνώριση ομιλίας ή εξελιγμένες αναλύσεις κειμένων χρησιμοποιώντας την επεξεργασία φυσικής γλώσσας (Natural Language Processing). Αυτές οι τεχνικές εξόρυξης δεδομένων είναι καλές για τον προσδιορισμό της αξίας από ημι-δομημένα και μη δομημένα δεδομένα.

Κεφάλαιο 2: Εφαρμογές Εξόρυξης Δεδομένων- Πλεονεκτήματα και Μειονεκτήματα στον Τραπεζικό Τομέα

Εισαγωγή

Η εξόρυξη δεδομένων εξελίσσεται σε μια στρατηγικά σημαντική περιοχή δράσης για πολλές επιχειρήσεις, συμπεριλαμβανομένου του τραπεζικού τομέα. Πρόκειται για μια διαδικασία ανάλυσης των δεδομένων από διάφορες πλευρές και τη σύνοψή τους σε πολύτιμες πληροφορίες. Η εξόρυξη δεδομένων βοηθά τις τράπεζες να αναζητήσουν και εντοπίσουν κρυφά μοτίβα σε ομάδες δεδομένων, καθώς και στην ανακάλυψη άγνωστων σχέσεων μέσα στα δεδομένα αυτά. Σήμερα, οι πελάτες έχουν πολλές επιλογές σχετικά με το που μπορούν να στρέψουν την επιχειρηματική τους δραστηριότητα. Προγενέστερες τεχνικές ανάλυσης δεδομένων ήταν προσανατολισμένες προς την εξαγωγή ποσοτικών και στατιστικών χαρακτηριστικών δεδομένων. Οι νέες τεχνικές συμβάλλουν στην εξαγωγή χρήσιμων συμπερασμάτων στον τραπεζικό τομέα με σκοπό την έγκαιρη και αποτελεσματική προσφορά εύστοχων λύσεων, επενδυτικών προτάσεων και διευκολύνσεων στον πελάτη, και επομένως στην αποφυγή απώλειας πελατών. Η διατήρηση του πελάτη είναι ο σημαντικότερος παράγοντας προς ανάλυση στο σημερινό ανταγωνιστικό επιχειρηματικό περιβάλλον. Επίσης, η απάτη αποτελεί σημαντικό πρόβλημα στον τραπεζικό τομέα. Η ανίχνευση και η πρόληψη της απάτης είναι δύσκολη, διότι οι εγκληματίες αναπτύσσουν νέα συστήματα συνεχώς, και τα μέτρα τους αυξάνονται και εξελίσσονται ολοένα και περισσότερο με σκοπό να αποφεύγουν την εύκολη ανίχνευση.

Ξεκινώντας από τη σχέση πιστωτικού ιδρύματος – πελάτη, οι τεχνολογικές καινοτομίες επέτρεψαν στον τραπεζικό κλάδο να ανοίξει αποτελεσματικά κανάλια παράδοσης και να ασχοληθεί με τις προκλήσεις που θέτει η νέα οικονομία. Σήμερα, οι τράπεζες έχουν συνειδητοποιήσει ότι οι σχέσεις με τους πελάτες είναι πολύ σημαντικός παράγοντας για την επιτυχία τους. Η διαχείριση των σχέσεων των πελατών (CRM) είναι μια στρατηγική που μπορεί να βοηθήσει τις τράπεζες να χτίσουν μακροχρόνιες σχέσεις με τους πελάτες τους και να αυξήσουν τα έσοδα και τα κέρδη τους. Το CRM στον τραπεζικό τομέα έχει ιδιαίτερη αξία. Η εστίαση του CRM μετατοπίζεται από την απόκτηση πελατών στην παραμονή των ήδη υπαρχόντων πελατών και στη διασφάλιση του απαραιτήτου χρόνου, χρημάτων και διευθυντικών πόρων. Η πρόκληση που αντιμετωπίζουν οι τράπεζες είναι πώς να διατηρήσουν τους πιο κερδοφόρους πελάτες με το χαμηλότερο κόστος. Την ίδια στιγμή, για το παραπάνω ζητούμενο, πρέπει να βρίσκεται η βέλτιστη λύση, με ταχύτητα και ευελιξία.

Περνώντας στο πρόβλημα της απάτης, σημειώνεται ότι χρησιμοποιούνται από παλιά παραδοσιακές μέθοδοι ανάλυσης δεδομένων. Αυτός το πεδίο δράσης απαιτεί πολύπλοκες και χρονοβόρες έρευνες που σχετίζονται με διαφορετικούς τομείς γνώσης, όπως η χρηματοοικονομική, η οικονομία, οι επιχειρηματικές πρακτικές και το δίκαιο. Οι περιπτώσεις απάτης μπορούν να είναι παρόμοιες σε περιεχόμενο και εμφάνιση, αλλά συνήθως δεν είναι πανομοιότυπες. Σε αναπτυσσόμενες χώρες όπως η Ινδία, οι τράπεζες αντιμετωπίζουν περισσότερα προβλήματα με κακόβουλες ενέργειες. Χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων, είναι πιο εύκολο να δημιουργηθεί ένα επιτυχημένο μοντέλο πρόβλεψης και αποφυγής της απάτης, που θα παρέχει σημαντικές πληροφορίες στο χρήστη.

Τα εργαλεία εξόρυξης δεδομένων, που χρησιμοποιούν βάσεις μεγάλων δεδομένων, μπορούν να διευκολύνουν τόσο την αυτόματη πρόβλεψη των μελλοντικών τάσεων και συμπεριφορών, όσο και την αυτόματη ανακάλυψη προηγούμενα άγνωστων προτύπων.

Οι σύγχρονες τράπεζες πρέπει να ανταποκριθούν σε προκλήσεις όπως η αυτοματοποίηση των διαδικασιών τους, οι αυξημένες προσδοκίες των πελατών, ο αυξανόμενος και επιθετικός ανταγωνισμός, οι συγχωνεύσεις και εξαγορές, η ανάπτυξη νέων προϊόντων και ο διαχωρισμός της αγοράς. Ταυτόχρονα, οι τράπεζες πρέπει επίσης να διαχειριστούν τους κινδύνους και να εναρμονίσουν τις επιχειρηματικές δραστηριότητές με τους εθνικούς και διεθνείς κανονισμούς λειτουργίας, όπως τα IFRS 9, AML, BASEL III κλπ. Η διοίκηση έχει να κάνει με τη λήψη αποφάσεων και οι αποφάσεις πρέπει να είναι έγκαιρες, επαρκείς και βασισμένες σε ακριβείς και αξιόπιστες πληροφορίες που προέρχονται από δεδομένα. Οι τράπεζες καταγράφουν μεγάλο αριθμό δεδομένων ημερησίως. Τα δεδομένα καταγράφονται για όλους τους πελάτες με βάση προσωπικά και ψυχοκοινωνικά τους χαρακτηριστικά, την περιουσία και τα οικονομικά χαρακτηριστικά τους, καθώς και όλους τους λογαριασμούς τους, τις συναλλαγές ανά λογαριασμό, τις πιστωτικές υποχρεώσεις κλπ. Αυτά τα στοιχεία παράγονται στο σύστημα της τράπεζας και αποθηκεύονται σε βάσεις δεδομένων συναλλαγών. Η εμπειρία έχει δείξει ότι οι βάσεις δεδομένων συναλλαγών είναι μια πλούσια πηγή πληροφοριών που μπορεί να χρησιμοποιηθεί για την ενίσχυση της επιχειρηματικής δραστηριότητας οποιασδήποτε επιχείρησης, ιδιαίτερα μιας τράπεζας, λόγω των προαναφερθέντων στοιχείων σχετικά με τη διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων. Έγινε σαφές πριν από πολύ καιρό ότι οι τράπεζες έχουν πολλά στοιχεία (δεδομένα) αλλά λίγες πληροφορίες και πολύ λίγες γνώσεις σε πολλές πτυχές των λειτουργιών τους. Οι βάσεις δεδομένων των συναλλαγών ωστόσο, είναι τεράστιες.

Έστω ότι η διοίκηση των τραπεζών θέλει να καθορίζονται τα χαρακτηριστικά των πελατών που ήταν αφερέγγυοι στο παρελθόν. Τέτοιες πληροφορίες μπορούν να ζητηθούν από το προσωπικό του τμήματος πληροφορικής της τράπεζας, το οποίο πρέπει να δαπανήσει πολύ χρόνο για την παραγωγή της ζητηθείσας πληροφορίας, πέραν του αναμενόμενου φόρτου εργασίας τους. Επιπλέον, τη στιγμή που η τελική αναφορά θα φτάσει στο γραφείο του διευθυντή, μπορεί να είναι ήδη πολύ αργά για τη λήψη αποφάσεων.

Η ανάπτυξη της τεχνολογίας πληροφοριών και επικοινωνιών (Information and communications technology - ICT) προσφέρει μια επιτυχημένη λύση στα προαναφερθέντα προβλήματα. Πρώτο βήμα από έναν οργανισμό αποτελεί η εκπαίδευση του προσωπικού σε ένα σύνολο μεθόδων, εργαλείων και εφαρμογών που σημειώνονται με τον όρο "Επιχειρηματική ευφυΐα" (Business Intelligence-BI). Σήμερα, το BI θεωρείται ως ένας ξεχωριστός κλάδος που συνδυάζει στοιχεία της τεχνολογίας των πληροφοριών, της στρατηγικής, της λογιστικής, της εταιρικής ανάλυσης και του μάρκετινγκ. Επιτρέπει τη συλλογή, την ανάλυση, τη μετάδοση και τη αξιοποίηση της επιχειρηματικής πληροφορίας, με στόχο τη διευκόλυνση της επίλυσης των προβλημάτων διαχείρισης και τη λήψη καλύτερων επιχειρηματικών αποφάσεων. Ένα σύστημα επιχειρηματικής ευφυΐας δεν αποτελεί τελικό προϊόν. Οι παραγωγοί του προσφέρουν πλατφόρμες τεχνολογίας και γνώσεις για την εφαρμογή και χρήση τους.

Οι νέες τράπεζες είναι ευρέως γνωστό πως βρίσκονται ανάμεσα στους πρωτοπόρους στο τομέα της ίδρυσης νέων τεχνολογιών και γνώσης, κάτι το οποίο αιτιολογεί το γεγονός ότι αποτελούν γόνιμο έδαφος για την υλοποίηση τέτοιων υποδομών. Ένας τύπος βάσεων δεδομένων, γνωστός ως αποθήκες δεδομένων (data warehouses-DW), δημιουργήθηκαν για να καλύψουν τις ανάγκες τέτοιων συστημάτων, όπου τα δεδομένα είναι οργανωμένα με πρακτικό τρόπο για τη διεξαγωγή διαδικασιών ανάλυσης σε σύνολα μεγάλων δεδομένων. Μία αποθήκη δεδομένων περιέχει αντίγραφα δεδομένων που απομονώνονται από επιχειρησιακές βάσεις δεδομένων και δομούνται ειδικά για αναφορές (reports) και αναλύσεις. Οι αποθήκες δεδομένων (DW) και η ηλεκτρονική αναλυτική επεξεργασία (OnLine Analytical Processing-OLAP) αποτελούν τη βάση για την εφαρμογή της επιχειρηματικής ευφυΐας. Η εξόρυξη δεδομένων αποτελεί επίσης σημαντικό παράγοντα της επιχειρηματικής ευφυΐας καθώς ασχολείται με πολύπλοκες στατιστικές αναλύσεις και με την ανακάλυψη “κρυμμένων” σχέσεων μεταξύ δεδομένων και πρόβλεψης συμπεριφορών των επιχειρηματικών συστημάτων.

2.1 Που εφαρμόζεται η Εξόρυξη Γνώσης

Τα παραδείγματα εξόρυξης γνώσης από δεδομένα ποικίλουν ανάλογα με τον τομέα στον οποίο εφαρμόζονται. Στη σημερινή εποχή όπου τα δεδομένα υπάρχουν σχεδόν παντού και τις περισσότερες φορές βρίσκονται σε ηλεκτρονική μορφή, η σωστή ανάλυση τους οδηγεί πάντα στην ανάδειξη και οργάνωση της πληροφορίας, η γνώση της οποίας είναι ο σημαντικότερος παράγοντας για την εύρεση μιας στρατηγικής και την ορθολογική λήψη αποφάσεων.

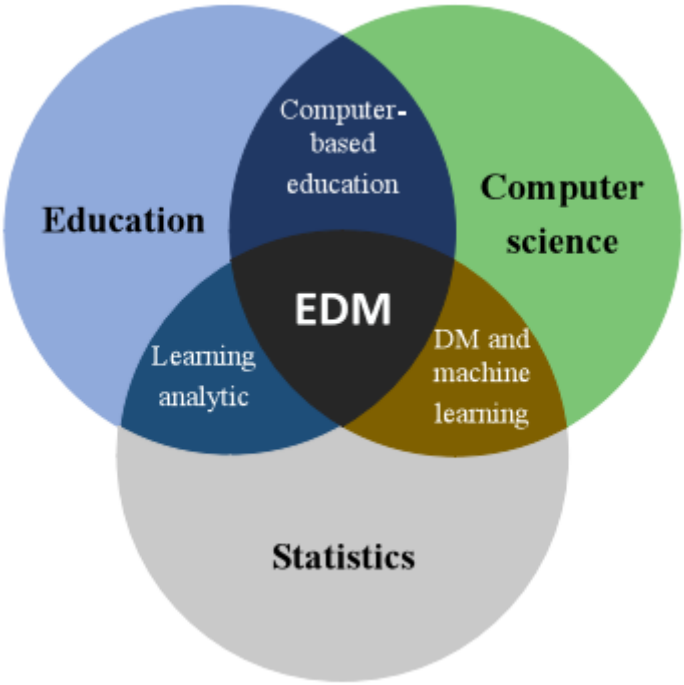
Ο χρηματοοικονομικός τομέας, ο τομέας των τηλεπικοινωνιών, της υγείας και της εκπαίδευσης, ο δημόσιος τομέας καθώς επίσης και αυτός της βιομηχανίας και της έρευνας, αποτελούν ίσως το μεγαλύτερο δείγμα εφαρμογών των τεχνολογιών εξόρυξης γνώσης από δεδομένα.

Χρηματοοικονομικός τομέας

Στον χρηματοοικονομικό τομέα η παρουσία αρκετών τραπεζικών και ασφαλιστικών ιδρυμάτων, σε συνδυασμό με το υπάρχον οικονομικό κλίμα, έχει αυξήσει κατά πολύ τον ανταγωνισμό μεταξύ των επιχειρήσεων. Πολλά από τα επιχειρηματικά ζητήματα του κλάδου, όπως η προσέλκυση νέων πελατών που θα αποφέρουν κέρδος, η προώθηση και πώληση επιπρόσθετων προϊόντων ή παροχή υπηρεσιών, η διατήρηση πελατών, ο εντοπισμός οικονομικού δόλου, η ανάλυση του πιστωτικού κινδύνου ενδεχόμενων πελατών, μπορούν να αντιμετωπιστούν με τα κατάλληλα εργαλεία εξόρυξης γνώσης.

Εκπαίδευση

Παρατηρήθηκε ότι η εξόρυξη δεδομένων έχει χρησιμοποιηθεί σε πολλές μελέτες και στον τομέα της εκπαίδευσης. Για παράδειγμα στον καθορισμό της κατάστασης επιτυχίας ή αποτυχίας των μαθητών, σε παράγοντες που επηρεάζουν την επιτυχία των φοιτητών που εγγράφονται στο πανεπιστήμιο, στη δημιουργία της προτίμησης του πανεπιστημιακού τμήματος και τον προσδιορισμό των παραγόντων που επηρεάζουν την προτίμηση των νέων εγγεγραμμένων φοιτητών στο τμήμα αυτό [7]. Επιπλέον, στην επιλογή ενός επαγγέλματος σύμφωνα με τα δημογραφικά και προσωπικά χαρακτηριστικά, στην πρόληψη των μαθητών από την αποτυχία και στον προσδιορισμό των παραγόντων που επηρεάζουν την επιτυχία, στο να προσδιοριστούν οι σχέσεις ανάμεσα στο είδος του σχολείου από το οποίο οι σπουδαστές αποφοιτούν σε σχέση με τα πανεπιστημιακά τους τμήματα αξιολογώντας τις δραστηριότητες σπουδών των φοιτητών από απόσταση και στον προσδιορισμό των προφίλ και των προτιμήσεων των φοιτητών για την είσοδο στο πανεπιστήμιο. Τέλος, η εξόρυξη δεδομένων χρησιμοποιείται στον προσδιορισμό της σχέσης μεταξύ της ακαδημαϊκής επιτυχίας και της συμμετοχής σε εξωπανεπιστημιακές δραστηριότητες των φοιτητών, στον προσδιορισμό των σχέσεων μεταξύ του κοινωνικοοικονομικού επιπέδου των μαθητών και του επιπέδου ακαδημαϊκής κατάρτισης και στον προσδιορισμό του αν υπάρχει σχέση μεταξύ βαθμολογίας εισόδου στο σχολείο και σχολικού επιπέδου. Αυτοί οι τομείς χρήσης στον εκπαιδευτικό τομέα βοηθούν τους εκπαιδευτικούς να κατανοούν τα μαθήματά και τους μαθητές τους, να κατανοούν την επίδοση των μαθητών τους και να παρέχουν προληπτική πληροφόρηση στους εκπαιδευόμενους.



Εικόνα 5 - Κύρια Επιστημονικά Πεδία της Εξόρυξης Δεδομένων Εκπαίδευσης.
Πηγή: Cristobal Romero & Ventura, (2013)

Τομέας Τηλεπικοινωνιών

Στον κλάδο των τηλεπικοινωνιών το διαρκώς μεταβαλλόμενο επιχειρηματικό περιβάλλον με τον αστείρευτο ανταγωνισμό, αναγκάζουν τις εταιρείες να προβούν στην αναζήτηση νέων τρόπων ενίσχυσης της θέσης τους έναντι άλλων. Έχοντας γίνει πλήρως αντιληπτό από τη πλευρά τους το συγκριτικό πλεονέκτημα που τους προσφέρει η εξόρυξη γνώσης από δεδομένα, εφαρμόζουν μεταξύ άλλων τεχνικές έγκαιρης πρόβλεψης διακοπής υπηρεσιών από πελάτες, κατηγοριοποίηση των απαιτήσεων τους, ομαδοποίηση των συνηθειών τους και όλα αυτά με τελικό σκοπό την συγκράτηση των πελατών που ήδη έχουν αλλά και την προσέλκυση νέων. Πιο συγκεκριμένα η εξόρυξη δεδομένων στις εταιρείες τηλεπικοινωνιών μπορεί να προβλέψει και να καθορίσει τυχόν μετακινήσεις στον τομέα των επικοινωνιών από τους χρήστες κινητών τηλεφώνων. Επιπλέον, μπορεί να ανιχνεύσει απάτες, να μειώσει μεγάλο μέρος της ανθρώπινης ανάλυσης, εξοικονομώντας χρήμα και χρόνο και να προσδιορίσει τους παράγοντες εκείνους που επηρεάζουν τους πελάτες ώστε να πραγματοποιούν τις περισσότερες κλήσεις σε συγκεκριμένα χρονικά διαστήματα. Τέλος, η εξόρυξη δεδομένων μπορεί να ανακαλύψει νέες προοπτικές χρησιμοποιώντας δημογραφικά στοιχεία και να προσδιορίσει τα χαρακτηριστικά των πελατών που χρειάζονται ειδικές ενέργειες για αναστολή ή απενεργοποίηση λογαριασμών [8].

Λιανικό εμπόριο

Ο τομέας του λιανικού εμπορίου είναι ένας άλλος κλάδος ιδιαίτερα ανταγωνιστικός, όπου οι εφαρμογές εξόρυξης γνώσης βρίσκουν μεγάλη ανταπόκριση. Οι συνεχείς αλλαγές των καταναλωτικών προτιμήσεων και οι τεράστιοι όγκοι δεδομένων πωλήσεων, κρύβουν πολύτιμα στοιχεία εκ των οποίων ελάχιστα μπορούν να αξιοποιηθούν από τα συμβατικά συστήματα ανάλυσης πληροφορίας. Αντιθέτως οι εφαρμογές εξόρυξης γνώσης δίνουν μια νέα διάσταση στην παλαιότερη επιχειρηματική διαδικασία, βασιζόμενες στην αρχή ότι «αναλύοντας ότι έγινε στο παρελθόν και κατανοώντας τα αποτελέσματα μπορούμε να γίνουμε αποτελεσματικότεροι στο μέλλον». Επιπλέον, οι εφαρμογές εξόρυξης γνώσης

κάνουν εφικτή μια προσωποποιημένη σχέση με κάθε ένα πελάτη χωριστά, κάτι που εξασφαλίζει την διαχρονική σχέση και την μεγιστοποίηση του κέρδους ανά πελάτη.

Κλάδος Υγείας

Οι επαγγελματίες στο χώρο της υγείας πάντα αντιμετωπίζουν την ανάγκη να συλλέγουν, να αποθηκεύουν και να αναλύουν μεγάλες ποσότητες δεδομένων που μπορεί να περιλαμβάνουν καρτέλες ασθενών, δοκιμές νέων φαρμάκων, εξάρσεις ασθενειών και πολλά άλλα. Οι τεχνικές εξόρυξης δεδομένων και τα εργαλεία εφαρμογής είναι πιο πολύτιμα για τον τομέα της υγείας και χρησιμοποιούνται ευρέως για να μειωθεί η πολυπλοκότητα της μελέτης των δεδομένων. Η εξόρυξη δεδομένων χρησιμοποιείται στον τομέα της υγείας για τη διάγνωση κάποιας νόσου, για να προσδιοριστεί η μέθοδος θεραπείας που θα εφαρμοστεί σε κάποια νόσο, για να εκτιμηθεί η χρήση των πόρων που θα χρησιμοποιηθούν και ο αριθμός των ασθενών στα νοσοκομεία, για να καθοριστεί η επιτυχία των μεθόδων θεραπείας που εφαρμόζονται στα νοσοκομεία, για να ταξινομηθούν τα δεδομένα των ασθενών σύμφωνα με παράγοντες όπως ηλικία, φύλο, φυλή και θεραπεία, για να προσδιοριστούν οι παράγοντες υψηλού κινδύνου στις χειρουργικές επεμβάσεις και τέλος για να αποφευχθεί η διαφθορά στις δαπάνες των νοσοκομείων [9], [10].

Δημόσιος τομέας

Η εξόρυξη δεδομένων χρησιμοποιείται συχνά για την πρόβλεψη θεμάτων δημόσιας ασφάλειας στον κόσμο και στον δημόσιο τομέα. Οι τεχνικές εξόρυξης δεδομένων προσφέρουν ανοικτές ευκαιρίες στον δημόσιο τομέα για τη βελτιστοποίηση των αποφάσεων. Αυτές οι αποφάσεις βασίζονται σε γενικές τάσεις που εξάγονται από προηγούμενες εμπειρίες και ιστορικά δεδομένα. Εκτός αυτού, η εξόρυξη γνώσης είναι απαραίτητη ώστε να εντοπιστεί η φοροδιαφυγή, να προβλεφθεί ο αντίκτυπος των αλλαγών του φορολογικού συστήματος στον προϋπολογισμό, να καθοριστούν οι δαπάνες και να αποτραπούν οι ζημίες που προκαλούνται από τις δαπάνες, να προβλεφθεί ο καιρός, να καθοριστούν νέες θέσεις εργασίας, να μετρηθεί η απόδοση των εργαζομένων, να βελτιστοποιηθούν οι επιχειρηματικές διαδικασίες, να ταξινομηθούν οι δημόσιες δαπάνες, να σχεδιαστεί η σωστή χρήση των διαθέσιμων πόρων, να προβλεφθεί το μέλλον των δημοσίων επενδύσεων, να αναλυθούν τα δεδομένα της αμυντικής βιομηχανίας και να εντοπισθούν οι παραβάτες που είναι πιθανό να διαπράξουν έγκλημα από την άποψη της ασφάλειας [11].

Κατασκευαστικός κλάδος

Η εξόρυξη δεδομένων χρησιμοποιείται στον κατασκευαστικό κλάδο στις κατασκευές, στη διαχείριση έργων, στην υδραυλική, στις εφαρμογές για την υγεία και την ασφάλεια στην εργασία, στην ανάλυση των σεισμικών δεδομένων, στις μελέτες εδάφους και σε πολλούς άλλους τομείς. Λαμβάνοντας υπόψη τις έρευνες που διεξήχθησαν στο πλαίσιο αυτό, διαπιστώθηκε ότι έχουν γίνει μελέτες για να δημιουργηθεί σύστημα ταξινόμησης πληροφοριών στα έγγραφα των έργων [12], να προβλεφθεί ο αντίκτυπος των αλλαγών του φορολογικού συστήματος στον προϋπολογισμό μιας κατασκευαστικής εταιρείας, να καθοριστούν οι δαπάνες και να αποτραπούν οι ζημίες που προκαλούνται από αυτές, να εκτιμηθεί το κόστος κατασκευής αυτοκινητοδρόμων, να προσδιοριστούν νέες ευκαιρίες απασχόλησης, να εκτιμηθεί η αντοχή του προϊόντος (τσιμέντου), να μετρηθεί η απόδοση των εργαζομένων, να βελτιστοποιηθούν οι επιχειρηματικές διαδικασίες, να σχεδιαστεί η σωστή χρήση των πόρων, η μέτρηση της παραγωγικότητας των εργαζομένων, να προσδιοριστεί η αντοχή του σκυροδέματος, να γίνει η πρόβλεψη των μελλοντικών επενδύσεων μιας εταιρείας και να προσδιοριστεί η σχέση ηγεσίας-κινήτρου μεταξύ επικεφαλής και εργαζομένου.

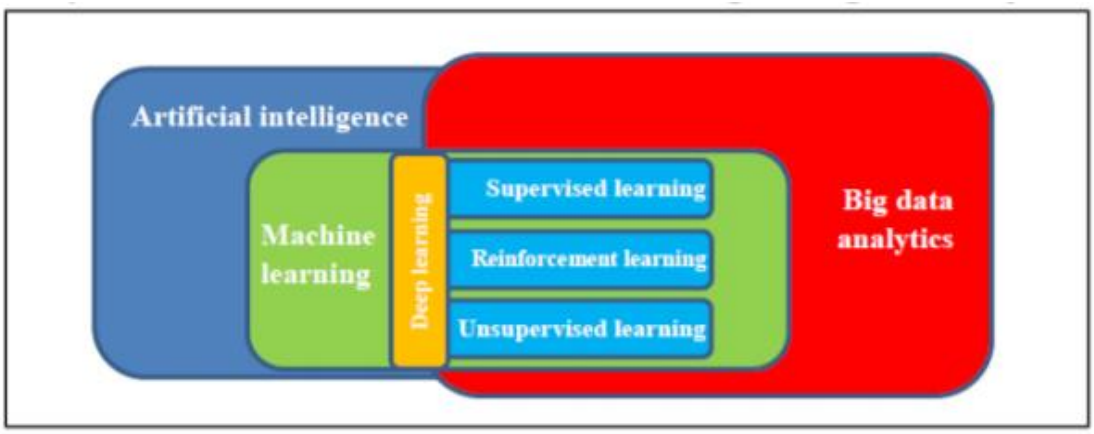
Τομέας Μηχανικής και Επιστημών

Μεγάλες ποσότητες δεδομένων συλλέγονται από επιστημονικούς κλάδους όπως η αστρονομία, η βιοπληροφορική, η πληροφορική, η εγκληματολογία, η μηχανική, οι γεωεπιστήμες, τα μαθηματικά κλπ. Η εξόρυξη δεδομένων παρέχει πολλά οφέλη στον τομέα της μηχανικής και της επιστήμης, όπως είναι η διαχείριση του λογισμικού που χρησιμοποιείται στις επιχειρήσεις, η μείωση του αριθμού των καθηκόντων, η εξοικονόμηση χρόνου και προσπάθειας, η παροχή ανταγωνιστικού πλεονεκτήματος στους οργανισμούς με την προβλεπόμενη ανάλυση, η βελτίωση της παραγωγικής διαδικασίας, η ανάλυση της ποιότητας του εδάφους, η εξαγωγή χαρακτηριστικών σε πραγματικό χρόνο για την ανάλυση τυρβώδους ροής, η απόκτηση νέων σπόρων καλής ποιότητας, η ανάπτυξη νέων φυτών καλλιέργειας, η ταξινόμηση των αστρονομικών αντικειμένων, η μοντελοποίηση του οικοσυστήματος, η ανακάλυψη των συσχετίσεων για την καλύτερη αξιοποίηση των χώρων αποθήκευσης και χρήσης του νερού των ποταμών και η ταξινόμηση των ακολουθιών στη βιοπληροφορική [8]. Όλα τα ανωτέρω βοηθούν τους χρήστες να βελτιώσουν την απόδοση του συστήματος του χρησιμοποιούμενου λογισμικού, να εξάγουν γνώση σε πολλά τμήματα των διαδικασιών ανάπτυξης λογισμικού και να προγραμματίσουν τη μελλοντική διαδικασία λήψης αποφάσεων.

Από όλα τα παραπάνω, συμπεραίνεται ότι η εξόρυξη γνώσης είναι ένα απαραίτητο εργαλείο σε πολλούς τομείς της σύγχρονης κοινωνίας. Η ραγδαία αύξηση του όγκου δεδομένων έχει καταστήσει σαφές ότι οι παλιές, παραδοσιακές τεχνικές και μέθοδοι δε μπορούν πλέον να βοηθήσουν στην ανάλυση και οργάνωση της πληροφορίας. Πολύ περισσότερο, δε μπορούν να φέρουν στην επιφάνεια γνώση που τα δεδομένα περιέχουν καλά κρυμμένη και η οποία απαιτεί εμπλοκή ειδικών για να αποκαλυφθεί.

2.2 Η Σημασία των Επιστημονικών Δεδομένων στην Τραπεζική

Οι πελάτες απαιτούν ψηφιακό μετασχηματισμό σε όλους τους τομείς και οι εταιρείες έχουν συνειδητοποιήσει τις ευεργετικές επιπτώσεις που αυτός έχει στην κατάσταση αποτελεσμάτων τους για βελτιστοποίηση των εσωτερικών διαδικασιών και επίτευξη λήψης των βέλτιστων αποφάσεων. Η συνέργεια μεταξύ της ανάλυσης δεδομένων, της τεχνητής νοημοσύνης και των μεγάλων δεδομένων αποτελεί τη βάση αυτού του ψηφιακού μετασχηματισμού [13].



Εικόνα 6 - Συσχέτιση Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και Μεγάλων Δεδομένων
Πηγή: FSB (2017)

Η Τεχνητή Νοημοσύνη στο σύνολό της και, ειδικότερα, με τη Μηχανική Μάθηση, δίνει τη δυνατότητα στα λογισμικά των εταιρειών να μάθουν μοντέλα και συμπεριφορές πελατών και να λαμβάνουν αποφάσεις αυτόνομα. Ωστόσο, για να γίνει αυτό, απαιτείται ανθρώπινη

καθοδήγηση. Αυτή τη στιγμή, η βαθιά μηχανική μάθηση, ένας κλάδος της μηχανικής μάθησης, εργάζεται για να επιτύχει την πλήρως ανεξάρτητη μάθηση λογισμικού, χωρίς ανθρώπινη παρέμβαση, και κατά συνέπεια να ξεπεράσει την πρόκληση της προσομοίωσης του τρόπου με τον οποίο μαθαίνει το ανθρώπινο μυαλό. Οι αναλύσεις δεδομένων και οι τεχνικές τεχνητής νοημοσύνης υπάρχουν εδώ και πολύ καιρό. Ωστόσο, αυτή τη στιγμή βρίσκονται σε εξέλιξη στην ανάπτυξη με βάση τα μεγάλα δεδομένα, καθώς είναι δυνατή η διαχείριση πολύ μεγάλου όγκου πληροφοριών και η επεξεργασία τους γρήγορα και αποτελεσματικά. Όσο μεγαλύτερος είναι ο όγκος των πληροφοριών, τόσο πιο ακριβείς είναι τα εντοπισμένα πρότυπα και οι συμπεριφορές. Για το λόγο αυτό, είναι σημαντικό να μπορούμε να βασιζόμαστε σε τεράστιο όγκο δεδομένων και να μπορούμε να τον επεξεργαζόμαστε γρήγορα ακόμη και σε πραγματικό χρόνο.

Η συνέργεια μεταξύ Data Analytics, Machine Learning (ή Deep Learning) και Big Data δίνει τη δυνατότητα σε εταιρείες να καινοτομούν σε όλες τις δομές τους και να προσφέρουν στους πελάτες τους πλήρως προσαρμοσμένες και εξατομικευμένες υπηρεσίες. Ο χρηματοπιστωτικός τομέας έχει συνειδητοποιήσει ότι κάθε ένας από τους πελάτες του παράγει τεράστιες ποσότητες δεδομένων ο καθένας και αναμορφώνεται πλήρως για να εξάγει όσο το δυνατόν περισσότερες κρυφές γνώσεις από αυτές τις πληροφορίες. Μέχρι στιγμής, τα δεδομένα αυτά δεν προσέφεραν αξία, αλλά τώρα ο στόχος είναι να βρίσκεται ο πελάτης και τα δεδομένα του στον πυρήνα της επιχείρησης.

13 Συνέργειες όπου ο χρηματοδοτικός τομέας μπορεί να επωφεληθεί:

1. Έξυπνος λογαριασμός: Τα χρηματοπιστωτικά ιδρύματα έχουν ως στόχο να προσφέρουν στους πελάτες τους μια νέα αντίληψη του τραπεζικού λογαριασμού αντικαθιστώντας τους παραδοσιακούς λογαριασμούς με μια νέα υπηρεσία προστιθέμενης αξίας που επιτρέπει στους πελάτες να λαμβάνουν προβλέψεις δαπανών και πιθανή βραχυπρόθεσμη και μακροπρόθεσμη υπερανάληψη. Επιτρέπει να αναλύσουν τη συμπεριφορά τους βάσει των δαπανών που δημιουργούνται, να κατηγοριοποιήσουν αυτόματα τις συναλλαγές για να τις βλέπουν ανά ομάδα και να συγκρίνουν τα έξοδα με ανώνυμους πελάτες με το ίδιο προφίλ ή να ελέγχουν συστάσεις προϊόντων που ανταποκρίνονται στις συγκεκριμένες ανάγκες τους.
2. Προσωποποιημένα χρηματοοικονομικά προϊόντα: Κάθε πελάτης έχει τη δική του οικονομική δραστηριότητα και, με την ανάλυση δεδομένων, είναι δυνατό να εντοπιστούν τα πρότυπα και οι συμπεριφορές για να του προσφέρουν εξατομικευμένα χρηματοοικονομικά προϊόντα για καλύτερη εμπειρία πελάτη και μεγαλύτερη ικανοποίηση [14].
3. Νέες επιχειρηματικές ευκαιρίες με τους πελάτες του ιδρύματος: Εκτός από τις πληροφορίες σχετικά με την οικονομική δραστηριότητα των πελατών τους, οι τράπεζες μπορούν πλέον να έχουν πρόσβαση σε εξωτερικές πληροφορίες, όπως δεδομένα από κοινωνικά μέσα ή συμπεριφορά στο διαδίκτυο, για να τα προσθέσουν στο οικοσύστημα δεδομένων που περιβάλλει κάθε πελάτη [14] . Αναλύοντας εξωτερικές πληροφορίες, νέες επιχειρηματικές ευκαιρίες ανοίγουν στις τράπεζες. Εάν ο πελάτης τους ανεβάζει φωτογραφίες συγκεκριμένου τύπου αυτοκινήτου και κεντρίσει το ενδιαφέρον τους, η τράπεζα μπορεί να δημιουργήσει μια προσφορά ενός προϊόντος τη συγκεκριμένη στιγμή, που ανταποκρίνεται στις ιδιαίτερες ανάγκες του πελάτη, η οποία αποστέλλεται μέσω του κοινωνικού δικτύου και μπορεί να υλοποιηθεί γρήγορα και με "πολύ λίγα κλικ".

4. Νέες επιχειρηματικές ευκαιρίες για μη πελάτες: Η ανάλυση εξωτερικών δεδομένων μπορεί να δημιουργήσει νέες επιχειρηματικές ευκαιρίες, ακόμη και για τους νέους πελάτες του χρηματοπιστωτικού ιδρύματος. Οι οικονομικές ανάγκες ενός ατόμου μπορούν να εντοπιστούν και η τράπεζα μπορεί να τους προσφέρει ένα προϊόν για τις ιδιαίτερες περιστάσεις τους και αυτό ίσως οδηγήσει ενδεχομένως σε μια συμφωνία στο μέλλον, ώστε να γίνουν και αυτοί πελάτες της τράπεζας.
5. Διαχείριση κινδύνων και πρόληψη απάτης: Υπάρχουν δύο περιπτώσεις πρωτοποριακής χρήσης της ανάλυσης δεδομένων, της μηχανικής μάθησης και των μεγάλων δεδομένων στα τραπεζικά ιδρύματα [15]: η διαχείριση του κινδύνου και η πρόληψη της απάτης είναι δύο από τα σημαντικότερα ζητήματα για τις τράπεζες αυτή τη στιγμή και γι' αυτό το λόγο, είναι τα πρώτα έργα που έχουν αντιμετωπιστεί με αυτές τις τεχνολογίες.
6. Εσωτερικός μηχανισμός σύστασης για τη φυσική θέση των υποκαταστημάτων: Το χρηματοπιστωτικό ίδρυμα πρέπει να συλλέγει στοιχεία σχετικά με τις περιοχές της πόλης που επισκέπτονται πιο συχνά οι πελάτες του, πότε πηγαίνουν εκεί, που ψωνίζουν, τι είδους πελάτες είναι και που υπάρχουν οι λιγότεροι πελάτες [16]. Με την εφαρμογή τεχνικών εξόρυξης δεδομένων, τα ιδρύματα είναι σε θέση να προσδιορίσουν την περιοχή που θα αποφέρει τα περισσότερα οφέλη ως αποτέλεσμα της τοποθέτησης των υποκαταστημάτων.
7. Εσωτερικός μηχανισμός συστάσεων για τη φυσική θέση των ATM: Όπως και παραπάνω, η τράπεζα πρέπει να αναλύσει τις περιοχές της πόλης όπου οι πελάτες της δαπανούν περισσότερο, πώς το κάνουν και τις περιοχές της πόλης όπου οι πελάτες της χρησιμοποιούν ATM από διαφορετικό χρηματοπιστωτικό ίδρυμα [17].
8. Μηχανισμός σύστασης σχετικά με τα χρήματα που πρέπει να προστεθούν στα ATM τα Σαββατοκύριακα και τις αργίες: Με βάση το ημερολόγιο της θέσης του ATM, τις καιρικές συνθήκες και τα γεγονότα - εκδηλώσεις της πόλης και τη τοποθεσία διεξαγωγής τους, είναι δυνατό να υπολογιστεί ακριβώς το σωστό ποσό που πρέπει να προστεθεί στα ATM [17]. Με αυτόν τον τρόπο, οι τράπεζες αποφεύγουν να βάζουν παραπάνω μετρητά στο ATM από όσα χρειάζονται ή από την άλλη αποφεύγουν τις διακοπές λειτουργίας λόγω έλλειψης μετρητών.
9. Πρόβλεψη πότε ένας πελάτης θα εγκαταλείψει το ίδρυμα: Με την ανάλυση της δραστηριότητας του λογαριασμού ενός πελάτη και τη συνδυασμένη πληροφορία με εσωτερικά δεδομένα από άλλες πηγές (υποκατάστημα ή ηλεκτρονικά) καθώς και με εξωτερικά δεδομένα (κοινωνικά μέσα), είναι δυνατό να προσδιοριστεί αν ο πελάτης θα εγκαταλείψει τη τράπεζα. Αν δεν υπάρχει κάποια δραστηριότητα στον λογαριασμό για κάποιο χρονικό διάστημα, ο πελάτης δεν επισκέπτεται τον ιστότοπο ή δεν πηγαίνει στο υποκατάστημα και ακολουθεί άλλες τράπεζες στα κοινωνικά μέσα ενημέρωσης, τότε είναι δυνατόν να προβλεφθεί πότε ο πελάτης θα εγκαταλείψει την τράπεζα . Όταν εντοπιστεί αυτή η κατάσταση, είναι σημαντικό να μπορούν να προταθούν προϊόντα ή βελτιώσεις για τη διατήρηση του πελάτη (τη δραστηριότητά του και τι μπορεί η τράπεζα να προσφέρει). Είναι πάντα πιο οικονομικό να διατηρείται ένας υφιστάμενος πελάτης από το να προσελκύεται ένας νέος.
10. Συχνά χρησιμοποιούμενες λειτουργίες ATM: Όταν χρησιμοποιούνται τα ATM, πολλοί από τους πελάτες της τράπεζας εκτελούν την ίδια λειτουργία (π.χ ερώτηση υπολοίπου, ανάληψη). Ο στόχος είναι να προσδιοριστεί το πρότυπο και η συμπεριφορά τους και να προσφερθεί αυτή η λειτουργία άμεσα, χωρίς ερωτήσεις ή περιήγηση. Για παράδειγμα, ένας πελάτης που επιλέγει συνήθως την ίδια ενέργεια θα χρειαστεί μόνο να πιέσει ένα κουμπί με τις συχνές εντολές του όταν εισάγει την κάρτα του στο μηχάνημα και το ATM θα παραδώσει 50 ευρώ χωρίς απόδειξη. Ως

αποτέλεσμα, ο πελάτης χρειάζεται πολύ λιγότερο χρόνο για να ολοκληρώσει μια λειτουργία και επομένως υπάρχει μεγαλύτερη ικανοποίηση του πελάτη.

11. Ανάλυση και προσδιορισμός των βέλτιστων μέσων επικοινωνίας με τον πελάτη: Οι πελάτες ζητούν από την τράπεζά τους να επικοινωνούν μέσω νέων καναλιών επικοινωνίας, τα οποία χρησιμοποιούν στην καθημερινή τους ζωή: κοινωνικά μέσα, ηλεκτρονικό ταχυδρομείο ή ανταλλαγή άμεσων μηνυμάτων. Τα χρηματοπιστωτικά ιδρύματα πρέπει να αναλύσουν και να προσδιορίσουν το κανάλι προτεραιότητας εκείνο, το οποίο κάνει τους πελάτες τους να αισθάνονται πιο άνετα με το να λαμβάνουν ειδοποιήσεις και να τις στείλουν με αυτό το μέσο. Τα χρηματοπιστωτικά ιδρύματα πρέπει να εγκαταλείψουν την παραδοσιακή πολιτική αποστολής ειδοποιήσεων με κανάλια που δεν χρησιμοποιούνται από τον πελάτη (π.χ αλληλογραφία μέσω ταχυδρομείου) και, ως εκ τούτου, έχουν ως αποτέλεσμα άσκοπες δαπάνες.
12. Νέα κανάλια επιχειρήσεων για τη δημιουργία εσόδων από συγκεντρωτικά και ανώνυμα δεδομένα: Τα δεδομένα πελατών αποτελούν το σημαντικότερο περιουσιακό στοιχείο των χρηματοπιστωτικών ιδρυμάτων. Ωστόσο, αυτές οι συγκεντρωτικές και ανώνυμες πληροφορίες μπορεί να έχουν μεγάλη αξία σε μια άλλη τράπεζα ή εταιρεία για εκμετάλλευση. Για παράδειγμα, οι τράπεζες μπορούν να έχουν τη δυνατότητα να επωφεληθούν από μια σημαντική επιχειρηματική ευκαιρία, διαθέτοντας ανώνυμα στοιχεία σχετικά με τις συχνές δαπάνες που σχετίζονται με την συμπεριφορά ενός συγκεκριμένου πληθυσμιακού προφίλ, ώστε κάποιες άλλες εταιρείες του κλάδου να είναι σε θέση να σχεδιάζουν ελκυστικές προσφορές και να αποκομίσουν κέρδος από αυτό.
13. Βελτιστοποίηση των διαδικασιών και των πόρων της τράπεζας: Συγκεντρώνοντας στοιχεία σχετικά με τις διαδικασίες και τους πόρους του ιδρύματος και στη συνέχεια αναλύοντάς τες, είναι δυνατόν να ανακαλύψουμε μέχρι τώρα άγνωστα πρότυπα και συμπεριφορές για τη μεγιστοποίηση των οφειλών και τη μείωση των δαπανών.

2.3 Πλεονεκτήματα και Μειονεκτήματα της Εξόρυξης Γνώσης

Η εξόρυξη δεδομένων αποτελεί ένα σημαντικό μέρος στη διαδικασία ανακάλυψης γνώσης με αποτέλεσμα να μπορούμε να αναλύσουμε ένα τεράστιο σύνολο δεδομένων και να αποκτήσουμε κρυφές και χρήσιμες γνώσεις. Η εξόρυξη δεδομένων εφαρμόζεται αποτελεσματικά όχι μόνο στο επιχειρηματικό περιβάλλον αλλά και σε άλλους τομείς όπως η πρόβλεψη του καιρού, η ιατρική, οι μεταφορές, η υγειονομική περίθαλψη, η ασφάλιση, η κυβέρνηση και πολλά άλλα. Η εξόρυξη δεδομένων έχει πολλά πλεονεκτήματα όταν χρησιμοποιείται σε μια συγκεκριμένη βιομηχανία. Εκτός από αυτά τα πλεονεκτήματα, η εξόρυξη δεδομένων έχει επίσης τα δικά της μειονεκτήματα, π.χ. ιδιωτικότητα, ασφάλεια και κακή χρήση πληροφοριών. Θα εξετάσουμε τα πλεονεκτήματα και τα μειονεκτήματα της εξόρυξης δεδομένων σε διάφορες βιομηχανίες με μεγαλύτερη λεπτομέρεια.

2.3.1 Πλεονεκτήματα

Μάρκετινγκ / Λιανική

Η εξόρυξη δεδομένων επιτρέπει στις εταιρείες μάρκετινγκ να φτιάξουν μοντέλα που στηρίζονται σε ιστορικά δεδομένα για να προβλέψουν ποιες από τις νέες εκστρατείες μάρκετινγκ θα επιτύχουν στον κόσμο, όπως το email, η εκστρατεία μάρκετινγκ στο internet

κλπ. Αυτά τα μοντέλα επιτρέπουν στους πωλητές να σχηματίζουν μια ολοκληρωμένη εικόνα για την πώληση κερδοφόρων προϊόντων σε στοχευμένους πελάτες.

Η εξόρυξη δεδομένων παρέχει πολλά πλεονεκτήματα και στις εταιρείες λιανικής πώλησης με τον ίδιο τρόπο όπως και το μάρκετινγκ. Δίνεται η δυνατότητα στα καταστήματα να παρέχουν προσωποποιημένες υπηρεσίες στους καταναλωτές και μέσα από την ανάλυση του καλαθιού αγοράς να μπορούν να διαθέτουν τα δημοφιλή προϊόντα με την κατάλληλη διάταξη για να τα εντοπίζουν πιο εύκολα. Επιπλέον, βοηθά τις εταιρείες λιανικής να παίρνουν αποφάσεις αναφορικά με προσφορές και εκπτώσεις για συγκεκριμένα προϊόντα που θα προσελκύσουν περισσότερους πελάτες.

Χρηματοοικονομικά / Τραπεζικά

Σημαντική είναι η συνεισφορά της εξόρυξης δεδομένων και στα χρηματοπιστωτικά ιδρύματα, όπως έχουμε αναφέρει παραπάνω, καθώς παρέχει πληροφορίες σχετικά με τα δάνεια και τις πιστώσεις. Χρησιμοποιώντας μοντέλα από ιστορικά δεδομένα πελατών, η τράπεζα μπορεί να πάρει αποφάσεις για αν θα πρέπει να χορηγήσει ένα δάνειο σε κάποιον αιτούντα με βάση το ρίσκο που έχει αυτός και μπορεί να καθορίσει τα καλά και τα επισφαλή δάνεια. Επιπλέον, η εξόρυξη δεδομένων βοηθά τις τράπεζες να ανιχνεύουν ύποπτες συναλλαγές πιστωτικών καρτών για την προστασία του κατόχου της πιστωτικής κάρτας.

Μεταποιητική βιομηχανία

Με την εφαρμογή εξόρυξης δεδομένων σε δεδομένα μηχανικής, οι κατασκευαστές μπορούν να εντοπίσουν ελαττωματικό εξοπλισμό και να καθορίσουν τις βέλτιστες παραμέτρους ελέγχου. Παραδείγματος χάριν, οι κατασκευαστές ημιαγωγών έχουν την πρόκληση ότι ακόμη και αν οι βιομηχανικές συνθήκες παραγωγής σε διαφορετικές εγκαταστάσεις παραγωγής πλακιδίων είναι παρόμοιες, η ποιότητα των ημιαγωγών είναι πολύ ίδια και κάποιες φορές για άγνωστους λόγους έχουν ακόμα και ελαττώματα. Η εξόρυξη δεδομένων εφαρμόζεται για τον προσδιορισμό των παραμέτρων ελέγχου που οδηγούν στην παραγωγή του καλύτερου ημιαγωγού. Στη συνέχεια, αυτές οι βέλτιστες παράμετροι ελέγχου χρησιμοποιούνται για την κατασκευή ημιαγωγών με την επιθυμητή ποιότητα.

Κυβερνήσεις

Η εξόρυξη δεδομένων βοηθάει τις κυβερνητικές υπηρεσίες να βρίσκουν και να αναλύουν τα αρχεία των χρηματοοικονομικών συναλλαγών για να χτίσουν πρότυπα που μπορούν να ανιχνεύσουν το ξέπλυμα χρημάτων ή τις εγκληματικές δραστηριότητες.

2.3.2 Μειονεκτήματα

Προβλήματα απορρήτου

Τα τελευταία χρόνια εξαιτίας της ραγδαίας αύξησης των χρηστών του internet σε υπηρεσίες όπως τα κοινωνικά δίκτυα, το ηλεκτρονικό εμπόριο, τα φόρουμ και τα blogs, έχει αυξηθεί παράλληλα και η ανησυχία πολλών για το ιδιωτικό απόρρητο. Οι άνθρωποι φοβούνται ότι τα προσωπικά τους στοιχεία συλλέγονται και χρησιμοποιούνται με τρόπο ανήθικο. Προκειμένου οι επιχειρήσεις να είναι ανταγωνιστικές και να κερδίσουν όσο το δυνατόν μεγαλύτερο μερίδιο αγοράς, συλλέγουν πληροφορίες σχετικά με τους πελάτες για να

μπορέσουν να κατανοήσουν τις τάσεις συμπεριφοράς των αγορών τους. Ακόμα και αν κλείσει μια επιχείρηση ή πωληθεί σε κάποιον τρίτο, οι προσωπικές πληροφορίες των πελατών που έχουν στην κατοχή τους μεταβιβάζονται σε άλλες επιχειρήσεις ή διαρρέουν.

Θέματα ασφάλειας

Η ασφάλεια αποτελεί ένα μείζον ζήτημα. Ο μεγάλος όγκος προσωπικών και ευαίσθητων πληροφοριών που διαθέτουν στην κατοχή τους οι επιχειρήσεις, συμπεριλαμβανομένου του αριθμού κοινωνικής ασφάλισης, του αριθμού ταυτότητας, της ημερομηνίας γέννησης, του αριθμού λογαριασμού, της μισθοδοσίας και πολλών άλλων, εγείρουν ερωτηματικά πόσο σωστά γίνεται η ενημέρωση αυτών των πληροφοριών και ποιος έχει πρόσβαση σε αυτές. Υπήρξαν πολλές περιπτώσεις που χάκερ είχαν πρόσβαση και έκλεψαν δεδομένα πελατών από μεγάλες εταιρείες όπως η Ford και η Sony. Με τόσο πολλά προσωπικά και οικονομικά στοιχεία διαθέσιμα, η κλοπή της πιστωτικής κάρτας και η κλοπή της ταυτότητας γίνονται αυτομάτως μεγάλα προβλήματα [18].

Κατάχρηση πληροφοριών / ανακριβείς πληροφορίες

Οι πληροφορίες που συλλέγονται μέσω της εξόρυξης δεδομένων και προορίζονται για ηθικούς σκοπούς μπορούν να χρησιμοποιηθούν κατά παράβαση της νομοθεσίας. Αυτές οι πληροφορίες μπορούν να αξιοποιηθούν από ανήθικους ανθρώπους ή επιχειρήσεις για να επωφεληθούν ευάλωτα άτομα ή να κάνουν διακρίσεις ενάντια σε μια ομάδα ανθρώπων. Επιπλέον, η τεχνική εξόρυξης δεδομένων δεν είναι απόλυτα ακριβής. Επομένως, αν χρησιμοποιούνται ανακριβείς πληροφορίες για τη λήψη αποφάσεων, αυτό θα έχει σοβαρές συνέπειες.

Η εξόρυξη δεδομένων προσφέρει πολλά οφέλη στις επιχειρήσεις, την κοινωνία, τις κυβερνήσεις καθώς και τον κάθε άνθρωπο ξεχωριστά. Ωστόσο, η προστασία της ιδιωτικής ζωής, η ασφάλεια και η κακή χρήση των πληροφοριών είναι τα μεγάλα προβλήματα εάν δεν αντιμετωπιστούν και επιλυθούν σωστά.

2.4 Πλεονεκτήματα και Μειονεκτήματα Τεχνητής Νοημοσύνης στις Τράπεζες

Η χρήση τεχνολογιών τεχνητής νοημοσύνης από τις τράπεζες προσφέρει πληθώρα πλεονεκτημάτων σε ότι αφορά την αποδοτικότητα και την ταχύτητα εκτέλεσης των επιχειρησιακών εργασιών, συμβάλλοντας καταλυτικά στις σχέσεις των χρηματοπιστωτικών ιδρυμάτων με τους πελάτες / καταναλωτές παρέχοντας στους τελευταίους προτάσεις και λύσεις διαμέσου εξατομικευμένων προϊόντων, καινοτόμων εμπειριών και μεγαλύτερης ευκολίας χρήσης των μέσων συναλλαγής.

2.4.1 Πλεονεκτήματα

Η τεχνητή νοημοσύνη μπορεί να ωφελήσει σε πολλές περιοχές του κοινωνικο - πολιτικο - οικονομικού συστήματος. Σημαντικότερες εξ αυτών μπορούν να θεωρηθούν η οικονομία, η ασφάλεια, η νομική, οι σχέσεις αντισυμβαλλομένων κ.α.

Παρακάτω παρατίθενται ορισμένα καίρια σημεία, στα οποία η χρήση τεχνητής νοημοσύνης στις εργασίες των τραπεζών μπορεί και λειτουργεί προς όφελός τους.

Αύξηση ασφάλειας πληροφοριών και πρόληψη απάτης

Οι τράπεζες στα ανάλογα τμήματα ψηφιακής καινοτομίας που διαθέτουν, μπορούν να αξιοποιήσουν την τεχνητή νοημοσύνη για να δημιουργήσουν έξυπνες μηχανές οι οποίες κατά τη χρήση τους από τους συναλλασσόμενους είναι ικανές να αποθηκεύουν και να επεξεργάζονται με ασφάλεια τεράστιο όγκο πληροφοριών. Αυτή η διαδικασία μπορεί να συμβάλλει στην άμεση και αποτελεσματική δημιουργία του προσωπικού προφίλ του κάθε χρήστη μέσω συγκεκριμένων μοτίβων, συνθέτοντας τη συνήθη συμπεριφορά τους, γεγονός που δίνει αφενός τη δυνατότητα στα χρηματοπιστωτικά ιδρύματα να παρέχουν εύστοχες εξατομικευμένες επενδυτικές συμβουλές και αφετέρου να ανιχνεύσουν τυχόν ύποπτες συμπεριφορές και να ενημερώνουν το ταχύτερο δυνατό τους αρμόδιους.

Κανονιστικές Αρχές / Απαιτήσεις - Συμμόρφωση και ανταπόκριση

Οι διαρκώς αυξανόμενες κανονιστικές απαιτήσεις στις οποίες καλούνται τα χρηματοπιστωτικά ιδρύματα να ανταποκριθούν, εντείνουν την ανάγκη αξιοποίησης με τον πλέον αποδοτικό τρόπο των δεδομένων που συνδέονται με τους πελάτες. Η τεχνητή νοημοσύνη παρέχει αυτή τη δυνατότητα, μέσω της επεξεργασίας των δεδομένων με ταχύτητα και ευελιξία, την εκπόνηση αξιόπιστων αποτελεσμάτων και την επίλυση προβλημάτων. Είναι σημαντικό να σημειωθεί ότι ήδη υφίσταται η απαραίτητη γνώση διασύνδεσης των υπαρχόντων γνώσεων και αυτοματοποιημένων διαδικασιών με την τεχνητή νοημοσύνη. Ως εκ τούτου, η χρήση νέων τεχνολογιών όπως η μηχανική μάθηση και η εξόρυξη δεδομένων μπορούν να συμβάλλουν ουσιαστικά στην δημιουργία διαδικασιών από τα τραπεζικά ιδρύματα που προσδίδουν αξία και προσφέρουν ευέλικτες λύσεις καθώς και στο σχεδιασμό αυτοματοποιήσεων που δεν απαιτούν επίβλεψη και εναπόκεινται στο κανονιστικό πλαίσιο.

Μείωση κόστους τραπεζών / Βελτίωση εμπειρίας πελάτη

Η χρήση ψηφιακών μεθόδων προσφέρει πλέον των άλλων τη δυνατότητα μείωσης του λειτουργικού κόστους, καθώς η διαμόρφωση αυτοματοποιημένων διαδικασιών εξασφαλίζει τον περιορισμό των απαιτούμενων χρόνων επεξεργασίας και μετασχηματισμού των δεδομένων, της ανάγκης δυναμικών ελέγχων των διαδικασιών σύμφωνα με το κανονιστικό πλαίσιο καθώς και του κόστους παροχής υπηρεσιών και εξυπηρέτησης πελατών. Ως αποτέλεσμα των ανωτέρω οι τράπεζες οδηγούνται σε αύξηση των καθαρών εσόδων τους [19].

Ταυτόχρονα, μέσω της χρήσης μεθόδων τεχνητής νοημοσύνης και εφαρμογής καινοτόμων διαδικασιών στον τομέα της εξυπηρέτησης, ο εκάστοτε πελάτης δύναται να αντιμετωπίζεται από το χρηματοπιστωτικό ίδρυμα διακριτά, μέσω της παροχής προσωποποιημένων προϊόντων και διευκολύνσεων ανάλογα με τις ανάγκες του [19].

2.4.2 Μειονεκτήματα

Καθώς τα συστήματα τεχνητής νοημοσύνης καταλαμβάνουν ολοένα και μεγαλύτερο έδαφος στην ζωή των σύγχρονων ανθρώπων, εγείρονται αντίστοιχα και προβληματισμοί αναφορικά με τη χρήση τους και τους κινδύνους και προβλήματα που ενδεχομένως μπορεί να προκύψουν.

Εστιάζοντας στην ανάπτυξη συστημάτων επιχειρησιακής ευφυΐας στον τραπεζικό κλάδο περιγράφονται παρακάτω ορισμένες από τις υπάρχουσες ανησυχίες, υπό το πρίσμα του ηθικού, κανονιστικού και λειτουργικού πλαισίου.

Ηθικές Ανησυχίες

Η χρήση της τεχνητής νοημοσύνης από τα πιστωτικά ιδρύματα πρέπει να διέπεται από ένα σύνολο ηθικών αρχών και αξιών, έτσι ώστε να διασφαλίζεται η ακεραιότητα και αμεροληψία απέναντι στον πελάτη. Ως εκ τούτου οι τράπεζες, όπως και το σύνολο των επιχειρήσεων, προκειμένου να αξιοποιήσουν τις νέες τεχνολογίες χωρίς να θέσουν σε κίνδυνο εξαπάτησης τον καταναλωτή, οφείλουν να ενσωματώσουν στις δομές και διαδικασίες τους συγκεκριμένα και αυστηρά πλαίσια διαχείρισης του τεράστιου όγκου δεδομένων, διαφάνειας και ασφάλειας. Επιπλέον, είναι σημαντικό να παρέχονται , με τρόπο σαφή και ξεκάθαρο, αναλυτικές πληροφορίες σχετικά με τις δυνατότητες και τους περιορισμούς των τεχνολογιών τεχνητής νοημοσύνης που αναπτύσσονται και χρησιμοποιούνται εντός του εκάστοτε οργανισμού, σε όλα τα εμπλεκόμενα μέρη και χρήστες.

Στις βασικότερες ηθικές ανησυχίες κατατάσσεται επίσης και ο βαθμός διασφάλισης της ιδιωτικότητας των πελατών, τα στοιχεία των οποίων χρησιμοποιούν τα συστήματα τεχνητής νοημοσύνης. Το συγκεκριμένο αποτελεί ένα ζήτημα που θα πρέπει να αντιμετωπίζεται με ιδιαίτερη προσοχή κάθε φορά που εξετάζεται μια λύση τεχνητής νοημοσύνης από τα πιστωτικά ιδρύματα.

Επιπλέον προβληματισμοί γεννώνται σε ότι αφορά τη πλήρη ή μη διαφάνεια σχετικά με τους αλγορίθμους που χρησιμοποιούνται από την τεχνητή νοημοσύνη. Πιο συγκεκριμένα, κάθε ανάπτυξη ενός τέτοιου αλγορίθμου από τις αρμόδιες μονάδες της τραπέζης θα αποτελεί φυσικά περιουσία της. Ταυτόχρονα όμως, στα πλαίσια ενός άρτιου ηθικού πλαισίου και με σκοπό την ανάπτυξη σχέσεων εμπιστοσύνης μεταξύ τράπεζας – πελάτη, θα μπορούσε να κριθεί σκόπιμος ο ακριβής προσδιορισμός των αλγορίθμων στον πελάτη, καθώς και των παραμέτρων και υπολογισμών που αυτοί χρησιμοποιούν ώστε να εξάγουν τα αποτελέσματα πάνω στα οποία εν τέλει βασίζεται η συναλλαγματική του σχέση με την τράπεζα.

Τέλος, εξαιρετικά έντονη παραμένει η ανησυχία που σχετίζεται με την ανάληψη ευθύνης για πράξεις που στηρίζονται σε αποτελέσματα τεχνικής νοημοσύνης. Αξίζει να σημειωθεί ότι έχει ανακοινωθεί από την πρόεδρο της Ευρωπαϊκής Επιτροπής πως έως τη λήξη του 2020 θα καταρτισθεί νομοθεσία για συντονισμένη πανευρωπαϊκή προσέγγιση των ανθρωπίνων και ηθικών συνεπειών της τεχνητής νοημοσύνης.

Ποιότητα και Κόστος

Ένα από τα κύρια προβλήματα στην ανάπτυξη συστημάτων τεχνητής νοημοσύνης και επιχειρησιακής ευφυΐας, είναι η ποιότητα των δεδομένων που είναι διαθέσιμα και θα χρησιμοποιηθούν από αυτά. Πολύ συχνά τα πρωτογενή δεδομένα είναι διάσπαρτα και χαρακτηρίζονται από έλλειψη πληρότητας, ομοιογένειας και ενδεχομένως ορθότητας. Αυτό, στα πλαίσια της αυτοματοποιήσεως διαδικασιών με τεχνητή νοημοσύνη μπορεί να οδηγήσει σε εσφαλμένη τελική πληροφόρηση και αποτέλεσμα, ενδεχόμενο ιδιαίτερα επικίνδυνο για τον επιχειρησιακό κόσμο. Πλέον του ανωτέρω, πρέπει να αναφερθεί και το ζήτημα συμβατότητας των συστημάτων νέων τεχνολογιών με τα υπάρχοντα. Λαμβάνοντας υπόψη ότι τα τελευταία είναι αυτά που ουσιαστικά θα τροφοδοτήσουν με δεδομένα τα νέα συστήματα που αναπτύσσονται, ενέχει ο κίνδυνος να προκύψουν θέματα ελλιπούς συμβατότητας τόσο μεταξύ κύριων συστημάτων όσο και μεταξύ συνδέσεων αυτών με τα συστήματα επιχειρησιακής ευφυΐας.

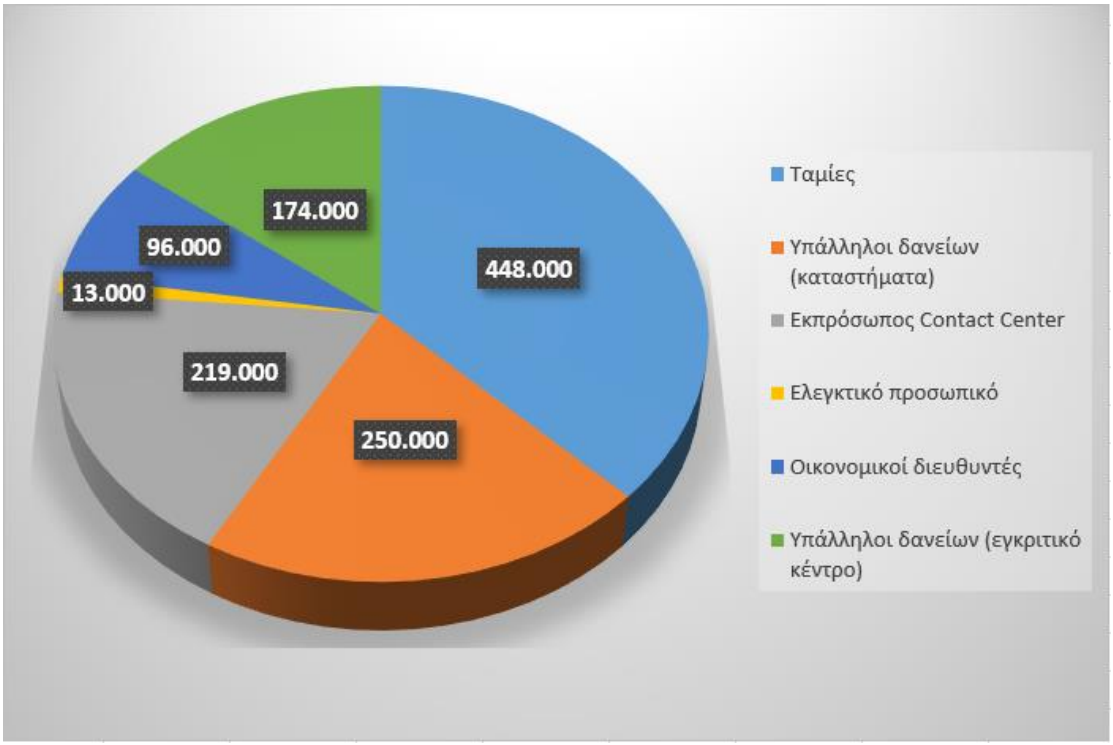
Επιπροσθέτως, ακόμα και στην περίπτωση που τα παραπάνω προβλήματα δύνανται να αντιμετωπιστούν επιτυχώς από τα πιστωτικά ιδρύματα, πρέπει να τονιστεί ότι το κόστος που πρέπει να επωμιστούν, τόσο για την απόκτηση χώρου αποθήκευσης μεγάλου όγκου δεδομένων που διαχειρίζονται τα συστήματα τεχνητής νοημοσύνης όσο και των εργαλείων επιχειρησιακής ευφυΐας, μπορεί να είναι ιδιαίτερα μεγάλο. Σε αυτό θα πρέπει να συνυπολογιστεί επίσης και το κόστος πρόσληψης ειδικά εκπαιδευμένου προσωπικού με

σκοπό την ομαλή εναρμόνιση του συνόλου των στελεχών στη νέα πραγματικότητα και την βελτιστοποίηση του τρόπου χρήσης των νέων τεχνολογιών.

Επιπτώσεις στο ανθρώπινο εργατικό δυναμικό

Η προοπτική της υιοθέτησης τεχνητής νοημοσύνης σε ένα οργανισμό ενδέχεται να αντιμετωπίζεται με αρνητισμό από μέρος του εργατικού δυναμικού του.

Είναι σαφές ότι με την ανάπτυξη των νέων τεχνολογιών και μεθόδων η ανάγκη ύπαρξης ορισμένων θέσεων εργασίας θα εκλείψει, κυρίως αυτών που αποτελούνται από επαναλαμβανόμενες κινήσεις χωρίς να απαιτούν οποιαδήποτε εφαρμογή κριτικής σκέψης ή πρωτοβουλίας από τον άνθρωπο. Αυτό το γεγονός, όπως είναι αναμενόμενο, εγείρει ανησυχίες και αντιδράσεις για την τύχη των εργαζομένων που μέχρι στιγμής καταλαμβάνουν τις εν λόγω θέσεις εργασίας. Παρ' όλα αυτά πρέπει να τονιστεί ότι από τη μία πλευρά κάποιες θέσεις εργασίας θα καταργηθούν, αντίστοιχα όμως κάποιες θα μετασχηματιστούν ώστε να ανταποκρίνονται στις νέες απαιτήσεις και κάποιες ακόμα θα δημιουργηθούν για να καλύπτουν νέες ανάγκες που προκύπτουν στον τραπεζικό κλάδο από την ενσωμάτωση της τεχνητής νοημοσύνης στις διαδικασίες του.



Εικόνα 7 - 1,2 εκατομμύρια εργαζόμενοι θα αντικατασταθούν από την τεχνητή νοημοσύνη έως το 2030.
Πηγή: https://www.americanbanker.com/news/how-artificialintelligence-is-reshaping-jobs-in-banking

Κεφάλαιο 3: Πιστωτικές Κάρτες και Διαχείριση Πιστωτικού Κινδύνου

Όπως περιγράφεται εκτενέστερα στο ανωτέρω κεφάλαιο, καθοριστική παράμετρο για την εξέλιξη του τραπεζικού κλάδου αποτελεί η επιτυχής διαχείριση μεγάλων δεδομένων μέσω της τεχνολογίας εξόρυξης δεδομένων καθώς και της χρήσης λύσεων επιχειρηματικής ευφυίας, οι οποίες μπορούν να συμβάλλουν καταλυτικά στην εκτίμηση των πιθανών κινδύνων, στο στρατηγικό προγραμματισμό, στην κερδοφορία αλλά και στην προστασία από κακόβουλες ενέργειες που μπορούν να κλονίσουν την εύρυθμη λειτουργία του εκάστοτε χρηματοπιστωτικού ιδρύματος.

Πριν αναλυθεί το κύριο αντικείμενο/ άσκηση της αυτής εργασίας – το οποίο είναι ο υπολογισμός της πιθανότητας αθέτησης πιστούχου (Probability of Default – PD) κατά την εκτίμηση του κόστους πιστωτικού κινδύνου για τα προϊόντα πιστωτικών καρτών - παρατίθενται κάποιες αναφορές στα βασικά είδη κινδύνων, στις μεθόδους εκτιμήσεως πιστωτικού κινδύνου αλλά και στις πιστωτικές κάρτες ως ξεχωριστό προϊόν δανειακού χαρτοφυλακίου.

Οι τραπεζικοί κίνδυνοι μπορούν να χωριστούν σε τέσσερεις κύριες κατηγορίες, τον πιστωτικό κίνδυνο, τον κίνδυνο αγοράς, τον λειτουργικό κίνδυνο και τον κίνδυνο ρευστότητας. Κάθε ένας από τους κινδύνους αυτούς γνωστοποιείται μέσα από τον ισολογισμό της τράπεζας.

Πιο συγκεκριμένα:

Πιστωτικός κίνδυνος (επίπεδο 1 κατηγοριοποίησης / παρακολούθησης κινδύνων) εμφανίζεται όταν ο οφειλέτης δεν είναι σε θέση να εκπληρώσει τις συμβατικές του υποχρεώσεις, δηλαδή να επιστρέψει το δάνειο (κεφάλαιο και τόκοι) που του έχει χορηγηθεί. Οι τρείς βασικές συνιστώσες του πιστωτικού κινδύνου (επίπεδο 2 κατηγοριοποίησης / παρακολούθησης κινδύνων) είναι ο κίνδυνος αθέτησης (default risk), ο κίνδυνος έκθεσης (exposure risk) και ο κίνδυνος ανάκτησης (recovery risk). Επιπλέον όμως , χαμηλότερου επιπέδου κίνδυνοι (επίπεδο 3 / drivers) της κατηγορίας πιστωτικού κινδύνου χαρακτηρίζονται ο κίνδυνος χώρας (sovereign risk), που συνδέεται με την αδυναμία αποπληρωμής του δημόσιου χρέους και ο κίνδυνος αντισυμβαλλομένων (counterparty risk), που ορίζεται ως η αθέτηση των υποχρεώσεων, γενικά, του αντισυμβαλλομένου.

Ο **κίνδυνος αγοράς** αφορά την πιθανότητα ενός επενδυτή να βιώνει ζημίες εξαιτίας παραγόντων που επηρεάζουν τη συνολική απόδοση των χρηματοπιστωτικών αγορών στις οποίες εμπλέκεται. Ο κίνδυνος αγοράς, που ονομάζεται επίσης "συστηματικός κίνδυνος", δεν μπορεί να εξαλειφθεί μέσω διαφοροποίησης, αν και μπορεί να αντισταθμιστεί με άλλους τρόπους. Πηγές κινδύνου αγοράς αποτελούν η ύφεση, η πολιτική αναταραχή, οι μεταβολές των επιτοκίων, οι φυσικές καταστροφές και οι τρομοκρατικές επιθέσεις.

Ο συστηματικός κίνδυνος ή κίνδυνος αγοράς τείνει να επηρεάζει ταυτόχρονα ολόκληρη την αγορά. Αυτό μπορεί εύκολα να γίνει κατανοητό εάν συγκριθεί με τον μη συστηματικό κίνδυνο, ο οποίος είναι μοναδικός για μια συγκεκριμένη εταιρεία ή βιομηχανία.

Ο **λειτουργικός κίνδυνος** συνοψίζει τις αβεβαιότητες και τους κινδύνους που αντιμετωπίζει μια εταιρεία όταν επιχειρεί να κάνει τις καθημερινές επιχειρηματικές της δραστηριότητες σε ένα δεδομένο τομέα ή βιομηχανία. Ένα είδος επιχειρηματικού κινδύνου μπορεί να προκύψει από τις καταστροφές στις εσωτερικές διαδικασίες, τους ανθρώπους και τα συστήματα - σε αντίθεση με τα προβλήματα που προκύπτουν από εξωτερικές δυνάμεις, όπως πολιτικά ή οικονομικά γεγονότα, ή συνυφασμένα με ολόκληρη την αγορά ή το τμήμα της αγοράς, γνωστό ως συστηματικός κίνδυνος. Ο λειτουργικός κίνδυνος μπορεί

επίσης να ταξινομηθεί ως μια ποικιλία ασταθούς κινδύνου, ο οποίος είναι μοναδικός για μια συγκεκριμένη εταιρεία ή βιομηχανία.

Κίνδυνος ρευστότητας συντελείται όταν μια τράπεζα αντιμετωπίσει απροσδόκητες αναλήψεις μετρητών σε λογαριασμούς καταθέσεων. Σημειώνεται ότι μια κατάθεση είναι ένα ρευστό χρηματοπιστωτικό προϊόν, στο οποίο ο πελάτης έχει τη δυνατότητα να απαιτήσει τα χρήματά του ανά πάσα στιγμή. Συνακόλουθα ένας μεγάλος αριθμός αναλήψεων πιθανόν να οδηγούσε σε έλλειψη ρευστότητας για την τράπεζα.

3.1 Εισαγωγή στη Διαχείριση Πιστωτικού Κινδύνου

Προκειμένου να ελαχιστοποιηθεί ο πιστωτικός κίνδυνος τα χρηματοπιστωτικά ιδρύματα πρέπει να δημιουργήσουν διαδικασίες διαχείρισης κινδύνου χορήγησης δανείων, τόσο λιανικής τραπεζικής όσο και επιχειρηματικών δανείων, αξιοποιώντας τα διαθέσιμα ιστορικά δεδομένα κατά το βέλτιστο τρόπο. Ωστόσο, η διαχείριση του πιστωτικού κινδύνου στο πλαίσιο των δανείων λιανικής, διαφέρει σημαντικά από την διαχείριση του πιστωτικού κινδύνου στο πλαίσιο του εταιρικού δανεισμού. Κατά την τελευταία εξετάζεται ένα μεγάλο πλήθος δεικτών και μεταβλητών που συχνά δεν είναι κατάλληλες για την αντίστοιχη εκτίμηση του πιστωτικού κινδύνου σε μικρές επιχειρήσεις ή μεμονωμένα άτομα [20]. Στον τομέα των δανείων προς ιδιώτες, η πρόσβαση στα δεδομένα είναι μάλλον περιορισμένη. Παρ 'όλα αυτά, ιστορικά, ο εταιρικός δανεισμός αποδείχθηκε ιδιαίτερα εκτεθειμένος στον πιστωτικό κίνδυνο λόγω του μεγέθους των εν λόγω δανείων.

Τα χρηματοπιστωτικά ιδρύματα μπορούν να ελαχιστοποιήσουν τον πιστωτικό κίνδυνο με πέντε διαφορετικούς τρόπους: τη σωστή και ακριβή τιμολόγηση δανείων (Pricing) ,τον περιορισμό πρόσθετων χρηματοδοτήσεων (Credit Rationing), τη χρήση εξασφαλίσεων (Collaterals) , την κατανομή των πιστώσεων σε διαφορετικές χρηματοπιστωτικές περιοχές οι οποίες αντιδρούν διαφορετικά στο ίδιο γεγονός (Diversification) και την τιτλοποίηση περιουσιακών στοιχείων ή / και χρήση πιστωτικών παραγώγων.

3.1.1 Μέθοδοι Αξιολόγησης Πιστωτικού Κινδύνου

Η παρούσα διπλωματική εργασία επικεντρώνεται στην κατανομή των πιστώσεων στον τομέα των πιστωτικών καρτών και ειδικότερα στη δανειοδότηση με βάση τα χαρακτηριστικά πιστωτικού κινδύνου που αξιολογούνται από την τράπεζα. Στη λιανική τραπεζική, ο πιστωτικός κίνδυνος μπορεί να εκτιμηθεί με δύο διαφορετικούς τρόπους, χρησιμοποιώντας ποιοτικές ή ποσοτικές μεθόδους. Ιδανική θα ήταν η χρήση συνδυασμού μεθόδων (ποιοτικής και ποσοτικής) από την τράπεζα ούτος ώστε να αποφασιστεί η χορήγηση ή μη δανείου. Η πιο γνωστή και αναγνωρισμένη ποσοτική μέθοδος είναι η μέθοδος βαθμολόγησης της πιστοληπτικής ικανότητας του πιστούχου (credit scoring[1]). Ωστόσο, υπό συγκεκριμένες συνθήκες, ενδέχεται να μην είναι διαθέσιμες όλες οι απαραίτητες πληροφορίες για την εν λόγω ποσοτική αξιολόγηση. Για παράδειγμα, στην περίπτωση όπου ο πιστωτής (τράπεζα) δεν είναι σε θέση να βρει μια έκθεση για έναν συγκεκριμένο αιτούντα, η τράπεζα τότε είναι πιθανό να μεταβεί σε ποιοτική προσέγγιση και να αξιολογήσει χειροκίνητα την αίτηση του δανειολήπτη. Αυτό συνεπάγεται τον έλεγχο μιας λίστας στοιχείων που συμβάλλουν στον καθορισμό του προφίλ κινδύνου του δανειολήπτη, όπως ο χρόνος συνεργασίας με την τράπεζα, το ιστορικό απασχόλησης και τα περιουσιακά στοιχεία του οφειλέτη.

Προκειμένου να παρακολουθούνται τα ανοίγματα με πιστωτικό κίνδυνο, τα χρηματοπιστωτικά ιδρύματα χρησιμοποιούν και βασίζουν την αξιολόγηση του πιστωτικού κινδύνου σε ένα σύνολο δεικτών (Key Performance Indicators / Key Risk Indicators) όπως η

[1] https://en.wikipedia.org/wiki/Credit_score

πιθανότητα αθέτησης (PD), τα ανοίγματα σε αθέτηση (EAD) και η ζημιά λόγω αθέτησης (LGD) .Τα τραπεζικά ιδρύματα σε ότι αφορά τους προαναφερθέντες δείκτες , διέπονται από κανονισμούς που ενδέχεται να διαφέρουν από χώρα σε χώρα και επομένως, σε ορισμένες χώρες οι κανονισμοί αυτοί αποτελούν ανασταλτικό παράγοντα για την καθιέρωση τραπεζικών δραστηριοτήτων ενώ σε άλλες χώρες λειτουργούν βοηθητικά προς αυτή την κατεύθυνση.

3.1.2 Τραπεζική Εποπτεία – Ρυθμιστικές Αρχές

Σε ολόκληρο τον κόσμο, όπου υφίσταται τραπεζικό σύστημα, οι τράπεζες συμμορφώνονται σε εποπτικές απαιτήσεις και κανονισμούς. Οι κανονισμοί αυτοί επηρεάζουν όχι μόνο τη διαχείριση της εκάστοτε τράπεζας αλλά και ολόκληρο τον κλάδο συνολικά. Οι βασικές κανονιστικές απαιτήσεις εμπίπτουν σε έξι κύριες κατηγορίες: τα ανώτατα όρια επιτοκίων καταθέσεων, τους περιορισμούς εισόδου στις χρηματαγορές, τους περιορισμούς στη χρήση διαδικασιών διακλάδωσης (branching processes) στην μοντελοποίηση συστημάτων, τους περιορισμούς ιδρύσεως καταστημάτων και συγχωνεύσεως καταστημάτων (network restrictions) και τέλος τους περιορισμούς χαρτοφυλακίου συμπεριλαμβανομένων των υποχρεωτικών ελάχιστων αποθεματικών, της ασφάλισης καταθέσεων και της εποπτικής παρακολούθησης και ελέγχου (συμπεριλαμβανομένης της πολιτικής κλεισίματος) [21].

Η ρυθμιστική και εποπτική λειτουργία σε ότι αφορά το τραπεζικό σύστημα, ασκείται από τις εθνικές ρυθμιστικές αρχές. Στην Ευρώπη, η εποπτεία των τραπεζών θα ακολουθεί τα διεθνώς αποδεκτά πρότυπα τραπεζικής εποπτείας που καθορίζονται από την επιτροπή τραπεζικής εποπτείας της Βασιλείας (Basel Committee for Banking Supervision) . Οι εθνικές ρυθμιστικές αρχές είναι εξουσιοδοτημένες να ασκούν τις ακόλουθες αρμοδιότητες:

- Εποπτεία του χρηματοπιστωτικού τομέα, η οποία περιλαμβάνει τον επιτόπου ή προγραμματισμένο έλεγχο, καθώς και την επιβολή κανονιστικών μέτρων.

- Χορήγηση / Ανάκληση τραπεζικών αδειών.

- Παρακολούθηση / Καθορισμός ρυθμιστικού πλαισίου προληπτικής εποπτείας.

3.1.3 Βασιλεία 3 (Basel III)

Η Βασιλεία ΙΙΙ είναι ένα διεθνές κανονιστικό πλαίσιο που δημοσιεύθηκε στις 16 Δεκεμβρίου 2010 το οποίο περιλαμβάνει ένα σύνολο μεταρρυθμιστικών μέτρων που συμφωνήθηκαν από την Επιτροπή της Βασιλείας σχετικά με την Τραπεζική Εποπτεία. Κύρια επιδίωξη της εφαρμογής του συγκεκριμένου πλαισίου είναι η ενίσχυση των κεφαλαιακών απαιτήσεων των τραπεζών μέσω της αύξησης της ρευστότητας και τη μείωση της μόχλευσης. Πιο συγκεκριμένα, οι μεταρρυθμίσεις αποσκοπούν στην ενίσχυση του κανονισμού για την αποφυγή ή έστω περιορισμό των συστημικών κινδύνων, της εποπτεία του τραπεζικού τομέα απαιτώντας περισσότερη διαφάνεια και γνωστοποιήσεις, τη διαχείριση κινδύνων και τη διακυβέρνηση του τραπεζικού τομέα γενικότερα. Κύριοι στόχοι των μέτρων αυτών είναι να καθιερώσουν αφενός έναν πλαίσιο ελέγχου της ικανότητας των τραπεζών να αντιμετωπίζουν περιόδους χρηματοοικονομικού στρες, και αφετέρου ένα ρυθμιστικό πλαίσιο που συνίσταται στη μείωση των συστημικών κινδύνων και στην ενίσχυση της κεφαλαιακής επάρκειας των τραπεζών.

Γίνεται λοιπόν κατανοητό ότι οι τράπεζες εποπτεύονται σε μεγάλο βαθμό και η ανάγκη για την εφαρμογή αυστηρών ρυθμιστικών πλαισίων είναι άμεσα συνδεδεμένη και με την πρόσφατη εμφάνιση νέων υπηρεσιών πληρωμών. Όλα αυτά τα συστήματα πληρωμών έχουν δημιουργήσει ανησυχίες μεταξύ των κυβερνήσεων, παρά τα σαφή πλεονεκτήματα αυτών των νέων μεθόδων πληρωμής, όπως η ασφάλεια. Το επόμενο τμήμα στοχεύει στην περιγραφή των διαφορετικών υπηρεσιών πληρωμών και ειδικότερα των πιστωτικών καρτών.

3.2 Πιστωτικές Κάρτες

Ο σκοπός αυτής της ενότητας είναι να υπενθυμίσει στον αναγνώστη τι είναι μια πιστωτική κάρτα και πού / πότε / γιατί εμφανίστηκε στην αγορά.

3.2.1 Ορισμός Πιστωτικής Κάρτας

Η πίστωση είναι μια μέθοδος πώλησης αγαθών ή υπηρεσιών χωρίς ο αγοραστής να κάνει χρήση μετρητών. Μια πιστωτική κάρτα είναι ένας αυτόματος τρόπος προσφοράς πίστωσης στον καταναλωτή κάτοχο της κάρτας. Σήμερα, κάθε πιστωτική κάρτα φέρει έναν αριθμό αναγνώρισης που επιταχύνει τις συναλλαγές αγορών.

3.2.2 Ιστορική Αναδρομή

Σύμφωνα με την Εγκυκλοπαίδεια Britannica, "η χρήση πιστωτικών καρτών προήλθε από τις Ηνωμένες Πολιτείες κατά τη διάρκεια της δεκαετίας του 1920, όταν μεμονωμένες επιχειρήσεις, όπως πετρελαϊκές εταιρείες και ξενοδοχειακές αλυσίδες, άρχισαν να τις εκδίδουν στους πελάτες". Ωστόσο, οι αναφορές σε πιστωτικές κάρτες έγιναν ήδη από το 1890 στην Ευρώπη. Οι πρώτες πιστωτικές κάρτες αφορούσαν πωλήσεις απευθείας μεταξύ του εμπόρου που προσφέρει την κάρτα και του πελάτη. Περίπου το 1938, οι εταιρείες άρχισαν μεταξύ τους να δέχονται πληρωμές με τις κάρτες των άλλων εταιρειών. Σήμερα, οι πιστωτικές κάρτες επιτρέπουν την πραγματοποίηση αγορών με αμέτρητα τρίτα μέρη.

Οι πιστωτικές κάρτες , ανά τα χρόνια, έχουν αλλάξει σχήμα. Αρχικά, οι πιστωτικές κάρτες ήταν μάρκες από μεταλλικά νομίσματα, ύστερα έγιναν μάρκες από μεταλλικές πλάκες, ίνες ή χαρτί. Σήμερα είναι κατασκευασμένες από πλαστικό.

Σύμφωνα με τον Bellis [22], ο εφευρέτης της πρώτης πιστωτικής κάρτας που εκδόθηκε από τράπεζα ήταν ο John Biggins της Flatbush National Bank of Brooklyn στη Νέα Υόρκη. Το 1946, ο Biggins εφηύρε το πρόγραμμα "Charge-It" μεταξύ πελατών, τραπεζών και τοπικών εμπόρων. Οι έμποροι κατέθεταν τα δελτία πωλήσεων στην τράπεζα και η τράπεζα χρέωνε τον πελάτη που χρησιμοποίησε την κάρτα. Μία από τις πρώτες πιστωτικές κάρτες ήταν η πιστωτική κάρτα Diners Club. Εμφανίστηκε το 1950 στις Ηνωμένες Πολιτείες και εφευρέθηκε από τον ιδρυτή του Diners Club, Frank McNamara. Ο σκοπός της κάρτας ήταν να μπορεί κάποιος να πληρώσει τους λογαριασμούς εστιατορίων. Ένας πελάτης μπορούσε να φάει χωρίς μετρητά σε οποιοδήποτε εστιατόριο που δεχόταν τις πιστωτικές κάρτες Diners Club. Το Diners Club πλήρωνε το εστιατόριο και ο κάτοχος της πιστωτικής κάρτας αποπλήρωνε την Diners Club. Η κάρτα Diners Club ήταν ουσιαστικά, αρχικά, μια χρεωστική κάρτα και όχι μια πιστωτική, δεδομένου ότι ο πελάτης έπρεπε να επιστρέψει ολόκληρο το ποσό όταν χρεωνόταν από τη Diners Club .

Το 1958, η American Express εξέδωσε την πρώτη της πιστωτική κάρτα και η Τράπεζα της Αμερικής (Bank of America) εξέδωσε την πιστωτική κάρτα της Americard (τη σημερινή Visa).Ο κύριος στόχος ήταν η χρήση της από πωλητές που ταξιδεύουν.

Στη δεκαετία του 1960, οι πιστωτικές κάρτες προωθήθηκαν ως μέσο εξοικονόμησης χρόνου και όχι ως μορφή πίστωσης. Η American Express και η MasterCard έγιναν τεράστιες επιτυχίες πολύ γρήγορα.

Ωστόσο, καθώς η επιτυχία συνδέεται επίσης με καταχρήσεις, στα μέσα της δεκαετίας του 70, το Κογκρέσο των ΗΠΑ άρχισε να ρυθμίζει τη βιομηχανία των πιστωτικών καρτών. Ένα παράδειγμα είναι ότι απαγορεύτηκε η μαζική αποστολή ενεργών πιστωτικών καρτών σε όσους δεν αιτήθηκαν σχετικά. Ωστόσο, δεν ήταν όλες οι κανονιστικές ρυθμίσεις φιλικές προς τον καταναλωτή, καθώς επίσης η απελευθέρωση στη συνέχεια επέτρεψε την επιβολή πολύ υψηλών επιτοκίων.

Η επόμενη ενότητα στοχεύει στην περιγραφή της διαφορετικής χρήσης του προϊόντος.

3.2.3 Η Πιστωτική Κάρτα ως Μέσο Δανεισμού

Στην σημερινή κοινωνία, υπάρχουν δύο τύποι αναγκών ρευστότητας, οι μακροπρόθεσμες ανάγκες που συνήθως συνδέονται με τις αγορές ακινήτων, αυτοκινήτων ή τις ιδιωτικές επιχειρηματικές πρωτοβουλίες και οι βραχυπρόθεσμες που συνήθως αντικατοπτρίζουν μια πολύ συγκεκριμένη ανάγκη. Κάποιος που αντιμετωπίζει μια επείγουσα ανάγκη μετρητών σε μια συγκεκριμένη χρονική στιγμή και έχοντας έλλειψη ρευστότητας την ίδια ακριβώς στιγμή θα τείνει να ψάξει για τις διάφορες διαθέσιμες δυνατότητες δανεισμού.

Οι ακόλουθες επιλογές είναι συνήθως διαθέσιμες [23]:

Δάνειο με υποθήκη ακινήτου

Αυτή η επιλογή είναι διαθέσιμη μόνο για το τμήμα του πληθυσμού που είναι ιδιοκτήτες σπιτιού. Ανάλογα με την αξία του ακινήτου ιδιοκτησίας, ορισμένοι θα είναι σε θέση να αντλήσουν χρήματα με το εν λόγω δάνειο. Ωστόσο, η ιδιοκτησία σπιτιού δεν είναι προσβάσιμη σε όλους.

Πώληση περιουσιακών στοιχείων

Μια λύση είναι να πουληθούν τα υπάρχοντά τους για να καλυφθούν οι επόμενες δαπάνες.

Περιουσιακά στοιχεία πελατείας

Μια άλλη λύση είναι η διάθεση περιουσιακών στοιχείων πελάτη για την κάλυψη των επόμενων δαπανών.

Τραπεζικά δάνεια

Η πιο συνηθισμένη επιλογή είναι ο πελάτης να πάει στην τράπεζα και να ζητήσει δάνειο. Ωστόσο, είναι σημαντικό να σημειωθεί ότι οι τράπεζες είναι μάλλον επιλεκτικές και δεν χορηγούν δάνεια σε όλους. Επιπλέον, τα μικρά βραχυπρόθεσμα τραπεζικά δάνεια είναι συνήθως αρκετά δαπανηρά λόγω του κόστους των συναλλαγών.

Δάνεια των λιανοπωλητών

Σε συγκεκριμένες περιπτώσεις, είναι πιθανό να δοθεί ένα δάνειο απευθείας από τον έμπορο λιανικής πώλησης. Ωστόσο, οι περισσότεροι έμποροι δεν παρέχουν αυτή τη δυνατότητα.

Πιστωτικές κάρτες

Εκτός από το μέσο συναλλαγής, οι πιστωτικές κάρτες προσφέρουν επίσης ανακυκλούμενες διευκολύνσεις. Σε περίπτωση επείγουσας βραχυπρόθεσμης ανάγκης ρευστότητας, θα πρέπει να υπάρχει πιστωτική κάρτα.

Η τελευταία θεωρείται και από τις πιο δημοφιλείς επιλογές. Πράγματι, οι άλλες επιλογές που εξετάστηκαν παραπάνω δεν είναι συγκριτικά ελκυστικότερες. Ακόμη και σε ότι αφορά το κόστος, τα επιτόκια που χρεώνουν οι εταιρείες πιστωτικών καρτών δεν είναι εξωφρενικά σε σχέση με το κόστος που συνδέεται με τις άλλες επιλογές.

Η πιστωτική κάρτα χρησιμοποιείται κυρίως ως συσκευή συναλλαγών. Παρόλα αυτά, επιτρέπει επίσης στους χρήστες να δανείζονται χρήματα από ένα μήνα σε άλλο. Ο οφειλέτης μπορεί να εξοφλήσει πλήρως το υπόλοιπο δανεισμού ή να καταβάλει μερικές πληρωμές ίσες ή ανώτερες από την ελάχιστη πληρωμή που απαιτείται από το δανειοδοτικό ίδρυμα.

Εμπειρικές ενδείξεις υποστηρίζουν ένθερμα την άποψη ότι η αύξηση της χρήσης πιστωτικών καρτών από κατόχους χαμηλού εισοδήματος αποτελεί μια ορθολογική υποκατάσταση άλλων μορφών, λιγότερο ελκυστικών, καταναλωτικής πίστης [23]. Πράγματι, σε σύγκριση με τις άλλες επιλογές για μικρά βραχυπρόθεσμα δάνεια, οι πιστωτικές κάρτες είναι σαφώς ένα ελκυστικό μέσο δανεισμού που διατίθεται ακόμη και σε άτομα με περιορισμένο πιστωτικό όριο.

3.3 Διαχείριση Πιστωτικού Κινδύνου στη Βιομηχανία Πιστωτικών Καρτών

Τα ανωτέρω υποκεφάλαια 3.1 και 3.2 είναι απολύτως θεωρητικά, με το πρώτο να αφορά στην εισαγωγή στη διαχείριση πιστωτικού κινδύνου στον τραπεζικό κλάδο και το δεύτερο στις τραπεζικές συναλλαγές που συνδέονται με προϊόντα πιστωτικών καρτών ως μέσο δανεισμού.

Σε αυτό το σημείο η ανάλυσή μας θα μπορούσε να ακολουθήσει δύο διαφορετικές προσεγγίσεις, η μεν μια προσανατολισμένη στη σκοπιά του δανειστή και η άλλη στη σκοπιά του καταναλωτή / οφειλέτη. Η διαφορά ανάμεσα στις δυο προσεγγίσεις έγκειται κυρίως στο στόχο που επιδιώκεται να επιτευχθεί κατά την εκάστοτε περίπτωση. Δηλαδή, μπορεί κανείς να αποφασίσει να επικεντρωθεί σε ένα σύστημα πιστωτικού κινδύνου για χρηματοπιστωτικά ιδρύματα που στοχεύει στη μεγιστοποίηση του κέρδους τους, αλλά θα μπορούσε επίσης να επικεντρωθεί σε ένα σύστημα πιστωτικού κινδύνου που θα ευνοούσε την ευημερία των καταναλωτών.

Η παρούσα εργασία είναι προσανατολισμένη σε ανάλυση μεθόδων διαχείρισης του πιστωτικού κινδύνου πιστωτικών καρτών με σκοπό την κερδοφορία των πιστωτικών ιδρυμάτων. Ως εκ τούτου, μετά την εισαγωγή του αναγνώστη στην αγορά των πιστωτικών καρτών, το επόμενο βήμα είναι η εισαγωγή του αναγνώστη στην περιοχή κινδύνου όπως ορίζεται από τα χρηματοπιστωτικά ιδρύματα. Κατωτέρω περιγράφονται οι κύριοι δείκτες που χρησιμοποιούνται στον τομέα του πιστωτικού κινδύνου / χρηματοπιστωτικό τομέα καθώς και πως θα πρέπει να είναι η δομή διαχείρισης του πιστωτικού κινδύνου. Επιπλέον περιγράφεται ποιος είναι ο βαθμός (score) πιστοληπτικής ικανότητας και από ποια διαδικασία αξιολόγησης προκύπτει. Η βαθμολόγηση της πίστωσης (credit scoring) μπορεί να έχει διαφορετικούς στόχους και έτσι μπορούν ανάλογα να εφαρμοστούν και διαφορετικοί δείκτες κατά τον πιστωτικό έλεγχο (scorecards).

Συνοπτικά λοιπόν σημειώνεται ότι, ακολούθως θα γίνει αναφορά στον τρόπο διαχείρισης πιστωτικού κινδύνου, στους δείκτες που χρησιμοποιούνται κατά τον πιστωτικό έλεγχο και πως αυτοί μπορούν να συμβάλλουν στη μείωση του πιστωτικού κινδύνου και τελικά των πιστωτικών ζημιών, καθώς και ποιοι είναι οι διαφορετικές δείκτες που μπορούν να χρησιμοποιηθούν κατά τη διαδικασία αξιολογήσεως του πιστωτικού κινδύνου.

3.3.1 Εισαγωγή στους Χρηματοοικονομικούς Κινδύνους

Η διαχείριση των κινδύνων στοχεύει στον έλεγχο όλων των κινδύνων που μπορεί να αντιμετωπίσει μια τράπεζα. Παρακάτω παρατίθενται βασικές πληροφορίες σχετικά με τους χρηματοοικονομικούς κινδύνους. Δεδομένου ότι η εργασία επικεντρώνεται στον πιστωτικό κίνδυνο, ορισμένες έννοιες σχετικές με αυτόν περιγράφονται λεπτομερέστερα έναντι άλλων.

Κατά ISO 31000, κίνδυνος (risk) ορίζεται ως το αποτέλεσμα της αβεβαιότητας σχετικά με συγκεκριμένους στόχους, είτε θετικό είτε αρνητικό.

Αντίστοιχα, κίνδυνος (risk) στον κλάδο των οικονομικών ορίζεται η μεταβλητότητα των δυνητικών αποτελεσμάτων μιας επένδυσης γύρω από την αναμενόμενη τιμή ή τον αριθμητικό τους μέσο. Πρόκειται για ένα άγνωστο στοιχείο της μελλοντικής αξίας ενός

χρηματοοικονομικού περιουσιακού στοιχείου. Μια τράπεζα θα υπολογίσει την αναμενόμενη απόδοση για ένα περιουσιακό στοιχείο και ο σχετικός κίνδυνος αντανακλάται στην προβλεπόμενη μεταβλητότητα της αναμενόμενης απόδοσης. Ο κίνδυνος δηλαδή είναι η διακύμανση γύρω από την αναμενόμενη απόδοση. Μπορεί να υποδηλώνει τη μεταβλητότητα ενός περιουσιακού στοιχείου, αλλά και τη μεταβλητότητα του ποσοστού αθέτησης για ένα πιστωτικό χαρτοφυλάκιο. Η πρόβλεψη της πιθανότητας αθέτησης είναι ένα χαρακτηριστικό της αποτίμησης κινδύνου, και οι πιθανές διακυμάνσεις γύρω από την πρόβλεψη αυτή ένα άλλο.

Στον τραπεζικό κλάδο, ο κίνδυνος διαιρείται συνήθως σε 4 πυλώνες. Ο πιστωτικός κίνδυνος ειδικά αποτελεί τον μεγαλύτερο κίνδυνο που αντιμετωπίζει μια τράπεζα. Το επόμενο τμήμα της εργασίας επικεντρώνεται στην αξιολόγηση του εν λόγω κινδύνου (credit risk assessment).

3.3.2 Αρχές Πιστωτικού Κινδύνου

Αυτή η διπλωματική εργασία εξετάζει τις πιστωτικές κάρτες και τον τρόπο κατασκευής ενός συστήματος για την ανάθεση κατάλληλων πιστωτικών γραμμών. Επομένως, η κατανόηση του εννοιολογικού και πιστωτικού κινδύνου αποτελεί απαραίτητη βάση για αυτό.

3.3.2.1 Ορισμός της Αθέτησης

Μια αθέτηση συμβαίνει όταν ένα συμβαλλόμενο μέρος δεν συμμορφώνεται με τις οικονομικές δεσμεύσεις του. Συνήθως, ο παραβάτης ορίζεται ως τέτοιος όταν χάσει την πρώτη πληρωμή για οποιεσδήποτε οικονομικές υποχρεώσεις. Στα περισσότερα χρηματοπιστωτικά ιδρύματα, μια αθέτηση είναι η αποτυχία να πραγματοποιηθούν οι απαιτούμενες πληρωμές χρέους εντός του προβλεπόμενου χρόνου. Ο οργανισμός αξιολόγησης Standard & Poor's (2003) ορίζει έναν οφειλέτη σε αθέτηση (defaulter) όταν δεν μπορεί να εκπληρώσει τις συμβατικές του υποχρεώσεις και να πληρώσει εγκαίρως.

Στις τράπεζες, ένας κάτοχος πιστωτικής κάρτας λαμβάνει ένα λογαριασμό κάθε μήνα, ο λογαριασμός αναφέρει το ποσό που θα πρέπει να πληρωθεί έως την καταληκτική ημερομηνία, την ημερομηνία λήξης. Ανάλογα με την εκάστοτε πολιτική προϊόντος, μπορεί να απαιτηθεί είτε το πλήρες ποσό είτε ένα μέρος αυτού. Εάν ο κάτοχος της κάρτας δεν εκπληρώσει τις συμβατικές του υποχρεώσεις, πληρώνοντας το απαιτούμενο ποσό έως την ημερομηνία λήξης, θα λάβει μια υπενθύμιση που θα περιλαμβάνει τα τέλη καθυστέρησης πληρωμής και τους τόκους που προκύπτουν από αυτά. Για να ξαναγίνει ο λογαριασμός του ενήμερος, ο κάτοχος της κάρτας πρέπει να πληρώσει το χρέος του το συντομότερο δυνατό.

Στον κλάδο των πιστωτικών ιδρυμάτων, ως άνοιγμα σε αθέτηση - σύμφωνα με το ισχύον κανονιστικό πλαίσιο (CRR Default) – θεωρείται κάθε άνοιγμα σε καθυστέρηση άνω των 90 ημερών (υπερημερία).

Συνήθως, ένας λογαριασμός αποστέλλεται στο τμήμα των εισπράξεων (collections) μετά από 180-210 ημέρες καθυστέρησης. Οι εισπράκτορες (collectors) είναι υπεύθυνοι να υπενθυμίσουν στον οφειλέτη ότι ο λογαριασμός είναι καθυστερημένος και να ζητήσουν την άμεση αποπληρωμή της καθυστερημένης οφειλής. Ο ορισμός της αθέτησης που χρησιμοποιείται σε αυτή την εργασία είναι αυτός που περιγράφηκε παραπάνω.

3.3.2.2 Ορισμός Πιστωτικού Κινδύνου

Ο πιστωτικός κίνδυνος είναι εγγενής στην τραπεζική δραστηριότητα. Το κύριο χρηματοοικονομικό περιουσιακό στοιχείο που σχετίζεται περισσότερο με τον πιστωτικό κίνδυνο είναι το δάνειο και ακολουθείται από τα ομόλογα αλλά σε μικρότερο βαθμό. Ωστόσο, όλο και περισσότερο επηρεάζεται ο πιστωτικός κίνδυνος από άλλα προϊόντα, όπως

τα εξωχρηματιστηριακά παράγωγα, τα ομολογιακά δάνεια, τις διατραπεζικές συναλλαγές, τις δεσμεύσεις και τις εγγυήσεις. Ο πιστωτικός κίνδυνος υποδηλώνει τον κίνδυνο το ένα μέρος που οριοθετείται από μια χρηματοοικονομική σύμβαση να μην είναι σε θέση ή να μην επιθυμεί να εκπληρώσει τις υποχρεώσεις του σε εύθετο χρόνο, προκαλώντας οικονομική ζημία στο άλλο μέρος. Όταν ο δανειολήπτης αθετήσει, η επόμενη έκθεση για τον δανειστή είναι το ποσό που οφείλεται από τον οφειλέτη. Ωστόσο, η τελική ζημία που προκύπτει ανέρχεται στο καθαρό άνοιγμα (συμπεριλαμβανομένης της προστασίας που κατέχει ο πιστωτικός φορέας, όπως οι εγγυήσεις τρίτων, οι εξασφαλίσεις κα) μείον το ποσό που μπορεί να εισπραχθεί από τους οργανισμούς είσπραξης (ή εσωτερικά μέσω διαπραγματεύσεων πτώχευσης).

3.3.3 Στοιχεία Πιστωτικού Κινδύνου

Τα τρία βασικά στοιχεία του Πιστωτικού Κινδύνου είναι η πιθανότητα αθέτησης, η έκθεση σε αθέτηση και η ζημία που οφείλεται σε αθέτηση υποχρεώσεων. Ο πιστωτικός κίνδυνος μπορεί να εκφραστεί ως συνάρτηση αυτών των παραμέτρων:

Πιστωτικός κίνδυνος = f (PD, EaD, LGD) όπου

PD: πιθανότητα αθέτησης

EaD: έκθεση σε αθέτηση

LGD: ζημιά λόγω αθέτησης

Το πιστωτικό VaR είναι ένα άλλο βασικό στοιχείο του πιστωτικού κινδύνου.

3.3.3.1 Πιθανότητα Αθέτησης

Η πιθανότητα αθέτησης, γνωστή και ως επιτόκιο υπερημερίας, κακό επιτόκιο ή αναμενόμενη συχνότητα αθέτησης, είναι η πιθανότητα ο δανειολήπτης να αθετήσει σε ένα συγκεκριμένο χρονικό ορίζοντα.

Τα πιστωτικά ιδρύματα χρησιμοποιούν την πιθανότητα αθέτησης σε διάφορα οικονομικά σενάρια και με ποικίλους τρόπους. Όταν ο χρονικός ορίζοντας είναι ένα έτος, εκφράζει την πιθανότητα ο δανειολήπτης να χρεοκοπήσει τους επόμενους δώδεκα μήνες. Για τα περισσότερα χρηματοπιστωτικά ιδρύματα, το ένα έτος για τον υπολογισμό της πιθανότητας αθέτησης θεωρείται μια λογική εύλογη περίοδος για την εκτίμηση της συνολικής έκθεσης σε κινδύνους. Λαμβάνοντας υπόψη το ανωτέρω γεγονός, μια πιθανότητα αθέτησης ενός έτους ικανοποιεί τις απαιτήσεις της Επιτροπής της Βασιλείας για τον υπολογισμό των κανονιστικών κεφαλαιακών απαιτήσεων. Αντίθετα, για μια τράπεζα όπου η πιθανότητα αθέτησης απαιτεί δύο χρόνια για να σταθεροποιηθεί συνίσταται αυτή να χρησιμοποιήσει μια πιθανότητα αθέτησης δύο ετών. Όταν ο χρονικός ορίζοντας είναι σωρευτικός για έτη t, εκφράζει την πιθανότητα ότι ένας δανειολήπτης θα αθετήσει τα επόμενα έτη. Συνήθως για να ληφθεί υπόψη η πιθανότητα διακύμανσης της βαθμολογίας (score) του δανειολήπτη κατά τη διάρκεια των ετών t, χρησιμοποιείται ένας «πίνακας μετάβασης» (transition matrix). Η σωρευτική πιθανότητα αθέτησης χρησιμοποιείται για την εσωτερική έγκριση πίστωσης και για σκοπούς τιμολόγησης δανείων. Για ένα συγκεκριμένο χρονικό σημείο, η πιθανότητα ο δανειολήπτης να αθετήσει ακριβώς μέσα στο έτος Τ είναι η πιθανότητα πρόωρης αθέτησης.

3.3.3.2 Έκθεση σε Αθέτηση

Το άνοιγμα σε αθέτηση είναι το ποσό που οφείλει ο δανειολήπτης στον δανειστή κατά τη στιγμή της αθέτησης, δηλαδή όταν παύει να εκπληρώνει τις υποχρεώσεις του. Αυτό είναι το οφειλόμενο ποσό ή απαίτηση κατά την ημερομηνία της αθέτησης του δανειολήπτη που θα

σταλεί στους οργανισμούς είσπραξης. Πρόκειται για το σύνολο του ανεξόφλητου χρέους και όχι μόνο για την εγγενή πληρωμή που δεν έχει καταβληθεί.

Το ποσοστό ανάκτησης είναι το ποσοστό της απαίτησης του οφειλέτη που αθέτησε το οποίο θα ανακτηθεί από τον δανειστή. Η απώλεια λόγω αθέτησης συνδέεται συνήθως με το ποσοστό ανάκτησης: LGD = 1-RR.

3.3.3.3 Ζημιά λόγω Αθέτησης

Η ζημία που προκύπτει από την αθέτηση είναι η αναμενόμενη πραγματική απώλεια που απορρέει από έναν δανειολήπτη με καθυστερημένες οφειλές που δεν θα ανακτηθούν. Η διαφορά μεταξύ του ανοίγματος σε αθέτηση και της ζημίας που προκύπτει από την αθέτηση υποδεικνύει πόσα από αυτά που οφείλει ο δανειολήπτης στον δανειστή έχουν ανακτηθεί από το πιστωτικό ίδρυμα.

Σημειώνεται ότι σε αντίθεση με το ποσοστό αθέτησης, η ζημία που προκύπτει από αθέτηση υποχρεώσεων εκφράζει μια διευκόλυνση, καθώς η ζημία που οφείλεται σε αδυναμία πληρωμής εξαρτάται από ειδικούς παράγοντες για την ασφάλεια αθέτησης, όπως η θέση, οι εξασφαλίσεις ή οι συμβατικές ρήτρες.

3.3.3.4 Αξία σε Κίνδυνο

Με τη μέθοδο VaR μετριέται, κάτω από κανονικές συνθήκες στην αγορά, η μέγιστη πιθανή ζημία και κατά συνέπεια η μείωση της αξίας ενός χαρτοφυλακίου ή η μείωση της αξίας (καθαρής θέσης) ενός χρηματοοικονομικού οργανισμού, για δεδομένο χρονικό ορίζοντα ή για μία συγκεκριμένη χρονική περίοδο και εντός συγκεκριμένου διαστήματος στατιστικής εμπιστοσύνης (δηλαδή με προεπιλεγμένη πιθανότητα) . Η μεθοδολογία του VaR συμπληρώνεται από δύο επιπλέον διαδικασίες, το StressTesting και το BackTesting. Με την πρώτη διαδικασία ελέγχουμε τη συμπεριφορά του υπό εξέταση χαρτοφυλακίου κάτω από ακραία και δυσμενή μακροοικονομικά σενάρια, ενώ με την δεύτερη επαληθεύεται η ορθότητα του VaR η οποία υπολογίζεται.

Παρακάτω παρατίθενται οι τρεις βασικοί παράμετροι της εκτίμησης της αξίας σε κίνδυνο:

Χρονικός Ορίζοντας

Η επιλογή του χρονικού ορίζοντα εξαρτάται από το είδος και τους στόχους της επενδυτικής θέσης, καθώς και τη ρευστότητα των τίτλων αυτής. Εξαρτάται πρακτικά από τη συχνότητα αναπροσαρμογών της θέσης. Συνήθως, το VaR υπολογίζεται για μία ημέρα.

Επίπεδο Εμπιστοσύνης

Το επίπεδο εμπιστοσύνης συνήθως λαμβάνει τιμές στατιστικής σημαντικότητας 90%, 95%, 98% και 99%. Καθορίζει το ποσοστό των περιπτώσεων κατά τις οποίες δεν θα έχουμε ζημίες πάνω από το ποσό που καταδεικνύει το VaR. Η επιλογή του διαστήματος εμπιστοσύνης είναι ενδεικτική της στάσης κάθε οργανισμού απέναντι στον κίνδυνο. Όσο μεγαλύτερο επίπεδο εμπιστοσύνης επιλέξουμε, τόσο ελαττώνεται η πιθανότητα το VaR να αποτύχει να προβλέψει ακραία φαινόμενα.

«Παράθυρο Δεδομένων»

Το «παράθυρο δεδομένων» αφορά την χρονική περίοδο που καλύπτει το δείγμα των δεδομένων. Ο υπολογισμός του VaR αποτελεί σημαντική υπόθεση και απαιτεί αρκετό όγκο δεδομένων, ιστορικών ή πραγματικών. Η χρήση ιστορικών δεδομένων είναι περισσότερο δημοφιλής γιατί τα πραγματικά στοιχεία, παρόλο που δίνουν σαφώς καλύτερες εκτιμήσεις, είναι περιορισμένα σε διαθεσιμότητα.

Κεφάλαιο 4: Υλοποίηση Αλγορίθμων - Τεχνικών Εξόρυξης Δεδομένων στα δεδομένα

Εκτός από το θεωρητικό κομμάτι για την εφαρμογή τεχνικών εξόρυξης δεδομένων στον τραπεζικό τομέα, κρίθηκε σκόπιμο να ενσωματωθεί ένα κεφάλαιο σε αυτή την εργασία που πραγματεύεται την αξιοποίηση πραγματικών δεδομένων για τον υπολογισμό της πιθανότητας αθέτησης πιστούχου. Συγκεκριμένα, θα γίνει υλοποίηση αλγορίθμων μηχανικής μάθησης σε ένα σύνολο δεδομένων πιστωτικών καρτών τράπεζας εξωτερικού με σκοπό να προβλεφθεί η πιθανότητα αθέτησης (Probability of Default - PD) του κάθε πελάτη της τράπεζας. Οι αλγόριθμοι που θα χρησιμοποιηθούν είναι ο **LightGBM,** ο **Catboost** (δύο αλγόριθμοι gradient boosting) και τα **Τεχνητά Νευρωνικά Δίκτυα**. Στην πορεία θα γίνει σύγκριση των αποτελεσμάτων τους και θα εφαρμοστεί μία τεχνική ενοποίησης των αποτελεσμάτων των τριών αλγορίθμων με τη χρήση της μεθόδου Stacking και της λογιστικής παλινδρόμησης. Αυτή η τεχνική δίνει τη δυνατότητα να υπάρξει ακόμα καλύτερη πρόβλεψη μιας και αποτελεί συνδυασμό όλων των παραπάνω αλγορίθμων.

4.1 Σύνολα δεδομένων

Στην παρούσα ενότητα θα παρουσιαστούν τα σύνολα δεδομένων που θα χρησιμοποιηθούν στην ανάλυση. Τα αρχεία αυτά είναι τέσσερα αρχεία τύπου csv, τα οποία θα επεξεργαστούν κατάλληλα ώστε να παραχθεί το τελικό ενοποιημένο αρχείο επί του οποίου θα γίνει η εφαρμογή των αλγορίθμων.

4.1.1 Σύνολο δεδομένων “application_train”

Το πρώτο dataset είναι το αρχείο με όνομα “application_train”. Το συγκεκριμένο σύνολο δεδομένων αποτελεί το κύριο dataset και περιέχει πληροφορίες σχετικά με τον πελάτη που έκανε αίτηση για δάνειο, πληροφορίες σχετικά με τα δάνεια, καθώς επίσης και την μεταβλητή “TARGET” που είναι η μεταβλητή που θα χρησιμοποιηθεί για την πρόβλεψη στη συνέχεια και αφορά το PD. Αριθμεί 307.511 εγγραφές και αποτελείται από 122 μεταβλητές, εκ των οποίων οι 71 είναι αριθμητικού τύπου (numeric), οι 36 είναι δυαδικού τύπου (boolean) και τέλος οι 15 αποτελούν κατηγορικές μεταβλητές (categories).

Dataset info

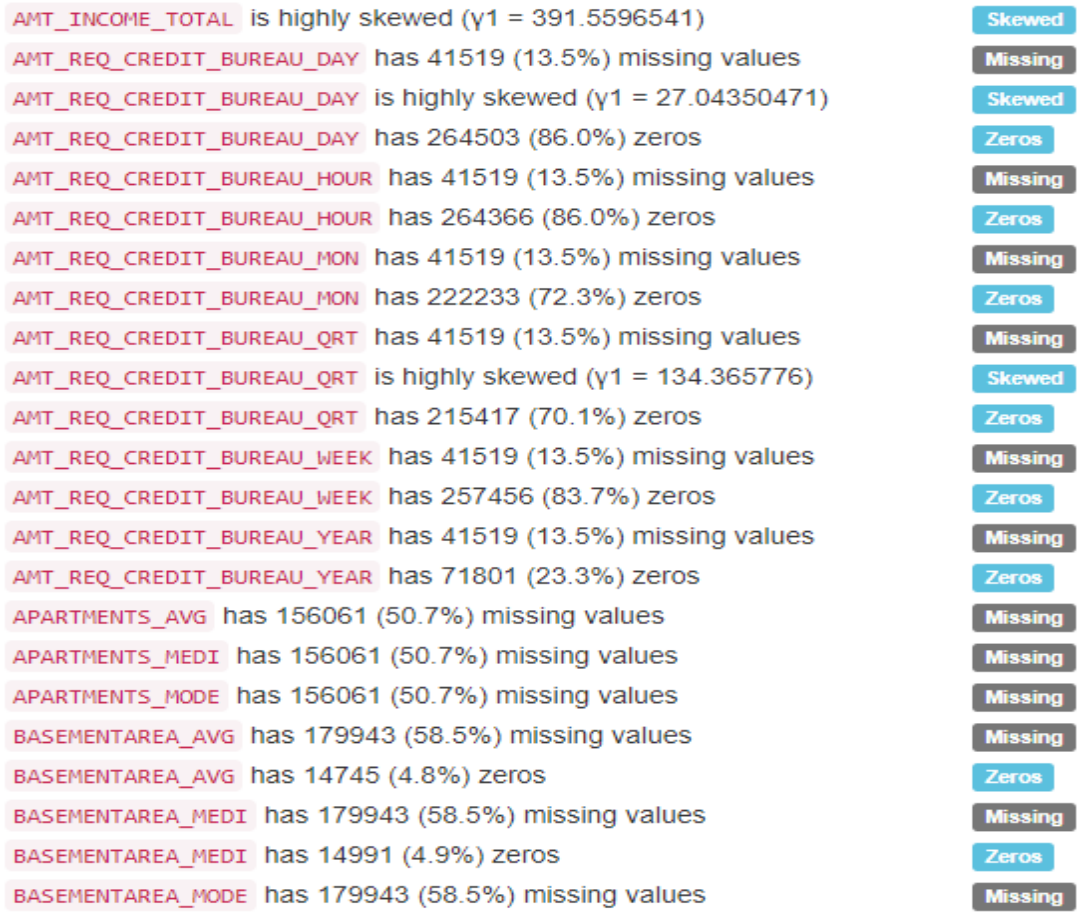
Number of variables	122
Number of observations	307511
Missing cells	9152465 (24.4%)
Duplicate rows	0 (0.0%)
Total size in memory	540.2 MiB
Average record size in memory	1.8 KiB

Variables types

NUM	71
BOOL	36
CAT	15

Εικόνα 8 - Σύνοψη περιεχομένου συνόλου application_train

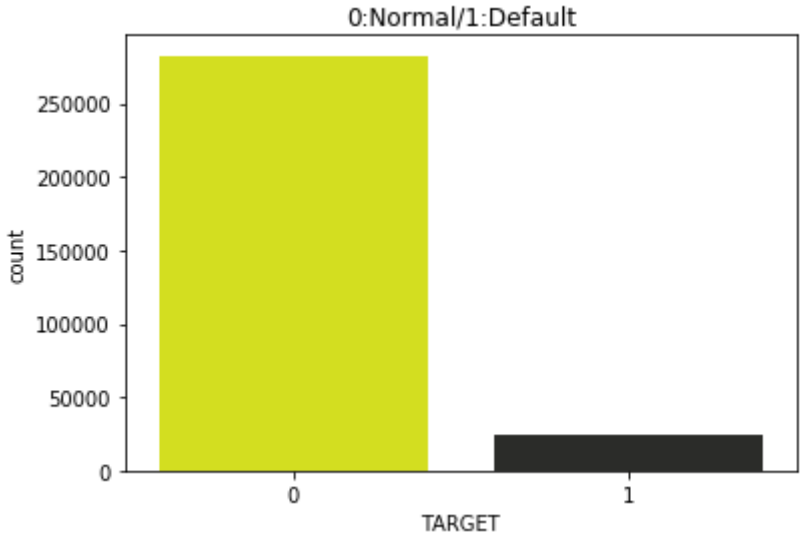
Παρατηρείται ότι πολλές από τις μεταβλητές περιέχουν πολλά κενά (missing values), είτε πολλά μηδενικά. Ακολουθεί παρακάτω ένα τυχαίο δείγμα μεταβλητών.



Εικόνα 9 - Δείγμα μεταβλητών με κενές τιμές

Μεταβλητή κλειδί στο αρχείο αυτό αποτελεί η μεταβλητή “SK_ID_CURR”, που υποδηλώνει το ID για κάθε δάνειο και περιέχει μοναδικές τιμές (unique). Η μεταβλητή “TARGET” υποδηλώνει την πιθανότητα κάποιος πιστούχος να μην καταβάλλει τη δόση του δανείου του και παίρνει την τιμή ‘1’ όταν κάποιος έχει δυσκολίες στην αποπληρωμή και ‘0’ όταν πληρώνει κανονικά, σύμφωνα με το πλάνο αποπληρωμής του. Ακολουθούν εικόνες που αποτυπώνουν τον αριθμό των παρατηρήσεων για τις τιμές της μεταβλητής “TARGET”:

```
In [84]: df.TARGET.value_counts()
Out[84]:
0    282686
1     24825
Name: TARGET, dtype: int64
```



Εικόνα 10 - Αριθμός παρατηρήσεων της μεταβλητής TARGET

Φαίνεται ότι το μεγαλύτερο ποσοστό των παρατηρήσεων κατηγοριοποιείται στην κλάση '0', ήτοι 282.686 παρατηρήσεις από το σύνολο 307.511 παρατηρήσεων.

Αξίζει να γίνει αναφορά σε τρεις από τις μεταβλητές του dataset που είναι ιδιαίτερα σημαντικές και δημοφιλείς στο εξωτερικό. Οι τρεις αυτές μεταβλητές είναι οι: "EXT_SOURCE_1", "EXT_SOURCE_2", "EXT_SOURCE_3". Ουσιαστικά αποτελούν τις σκοροκάρτες (scorecards[2]) των πελατών μιας τράπεζας. Οι σκοροκάρτες είναι μαθηματικά μοντέλα που επιχειρούν να παρέχουν μια ποσοτική εκτίμηση της πιθανότητας ότι ένας πελάτης θα εμφανίσει μια καθορισμένη συμπεριφορά (π.χ. αθέτηση υποχρεώσεων, πτώχευση ή χαμηλότερο επίπεδο παραβατικότητας) σε σχέση με την τρέχουσα ή προτεινόμενη πιστωτική τους θέση με έναν δανειστή. Τα περισσότερα εμπειρικά συστήματα πιστωτικής βαθμολόγησης έχουν από 10 έως και 20 μεταβλητές [24].

Η βαθμολογία πίστωσης συνήθως χρησιμοποιεί παρατηρήσεις ή δεδομένα από πελάτες που αθετούν τα δάνειά τους, καθώς και παρατηρήσεις σε μεγάλο αριθμό πελατών που δεν έχουν αθετήσει. Στατιστικά, τεχνικές εκτίμησης όπως η λογιστική παλινδρόμηση ή η probit χρησιμοποιούνται για να δημιουργηθούν εκτιμήσεις της πιθανότητας αθέτησης για παρατηρήσεις βάσει των ιστορικών δεδομένων. Αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη της πιθανότητας αθέτησης για νέους πελάτες χρησιμοποιώντας τα ίδια χαρακτηριστικά παρατήρησης (π.χ. ηλικία, εισόδημα, ιδιοκτήτης σπιτιού). Οι πιθανότητες αθέτησης στη συνέχεια δημιουργούν το "πιστωτικό σκορ". Αυτή η βαθμολογία κατατάσσει τους πελάτες με βάση τον κίνδυνο, χωρίς να προσδιορίζει ρητά την πιθανότητα αθέτησης.

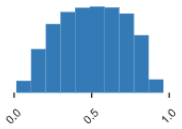
Η καθεμιά από τις 3 μεταβλητές του αρχείου προέρχεται από διαφορετική εξωτερική πηγή και έχει υποστεί κανονικοποίηση.

[2] https://en.wikipedia.org/wiki/Credit_scorecards

EXT_SOURCE_1
Real number (ℝ≤0)

MISSING

Distinct count	114585	Mean	0.5021298057
Unique (%)	37.3%	Minimum	0.01456813241
Missing	173378	Maximum	0.9626927706
Missing (%)	56.4%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	2.3 MiB



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Quantile statistics

Minimum	0.01456813241
5-th percentile	0.1580212307
Q1	0.3340072887
median	0.5059979305
Q3	0.6750525982
95-th percentile	0.8322601466
Maximum	0.9626927706
Range	0.9481246381
Interquartile range (IQR)	0.3410453096

Descriptive statistics

Standard deviation	0.2110622493
Coefficient of variation (CV)	0.4203340389
Kurtosis	-0.9651552848
Mean	0.5021298057
Median Absolute Deviation (MAD)	0.1791559886
Skewness	-0.0687550587
Sum	67352.17722
Variance	0.04454727307

Εικόνα 11 - Στατιστικά στοιχεία μεταβλητής EXT_SOURCE_1

Για να καλύτερη κατανόηση του αποτελέσματος των τριών αυτών μεταβλητών στη μεταβλητή “TARGET”, η εικόνα 12 παρουσιάζει την κατανομή για τις τρεις μεταβλητές για κάθε τιμή της “TARGET” (0 ή 1). Ο σκοπός αυτού του γραφήματος είναι να παρατηρηθούν πιθανές διαφορές μεταξύ των τιμών για τους στόχους. Σε αντίθεση με την “EXT_SOURCE_1”, οι “EXT_SOURCE_2” και “EXT_SOURCE_3” παρουσιάζουν τη σημαντικότερη διαφορά μεταξύ των δύο τιμών της “TARGET”, καθιστώντας τες χρήσιμες για την πρόβλεψη της πιθανότητας αθέτησης σε ένα μοντέλο.

Εικόνα 12 - Κατανομή μεταβλητών

4.1.2 Σύνολο δεδομένων “credit_card_balance”

Στη συνέχεια, θα γίνει αναφορά στο δεύτερο dataset που θα χρησιμοποιηθεί και αυτό είναι το αρχείο με όνομα “credit_card_balance”. Το αρχείο αυτό περιλαμβάνει λογαριασμούς πιστωτικών καρτών πελατών καθώς επίσης και τα μηνιαία υπόλοιπα των πιστωτικών καρτών. Αριθμεί 3.840.312 εγγραφές και αποτελείται από 23 μεταβλητές, εκ των οποίων οι 22 είναι αριθμητικού τύπου (numeric), και η 1 αποτελεί κατηγορική μεταβλητή (categories). Ακολουθεί περιγραφή των μεταβλητών.

Dataset info

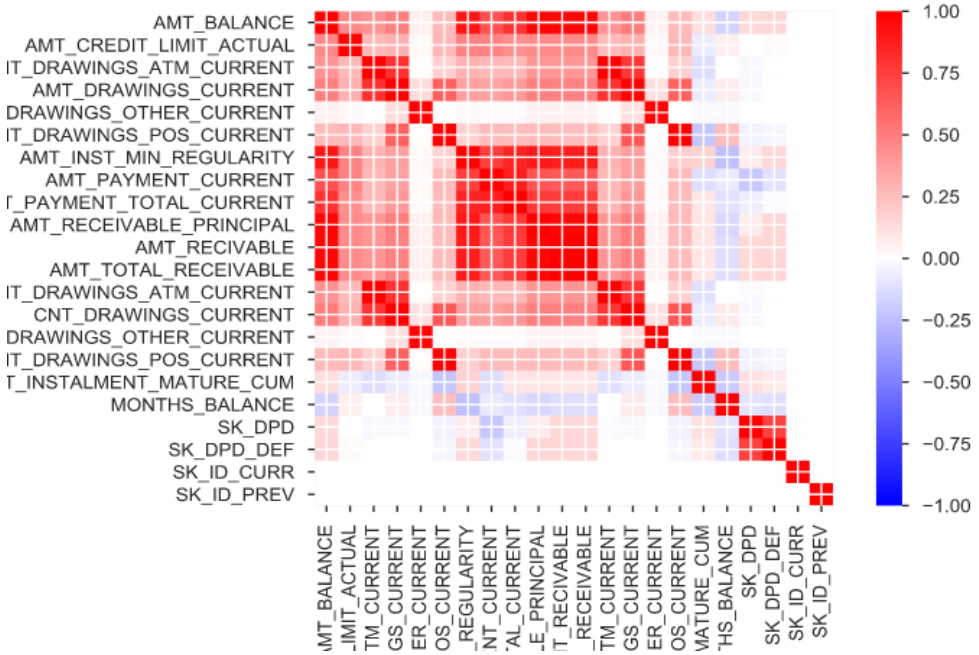
Number of variables	23
Number of observations	3840312
Missing cells	5877356 (6.7%)
Duplicate rows	0 (0.0%)
Total size in memory	875.7 MiB
Average record size in memory	239.1 B

Variables types

NUM	22
CAT	1

Εικόνα 13 - Σύνοψη περιεχομένου συνόλου credit_card_balance

Παρατηρείται και σε αυτό το σύνολο δεδομένων ότι περιέχει μεταβλητές με κενές τιμές ή μηδενικά. Ακολούθως θα απεικονιστεί ο συντελεστής συσχέτισης Spearman για τα δεδομένα, για να περιγραφεί η στατιστική συνάρτηση μεταξύ των μεταβλητών.

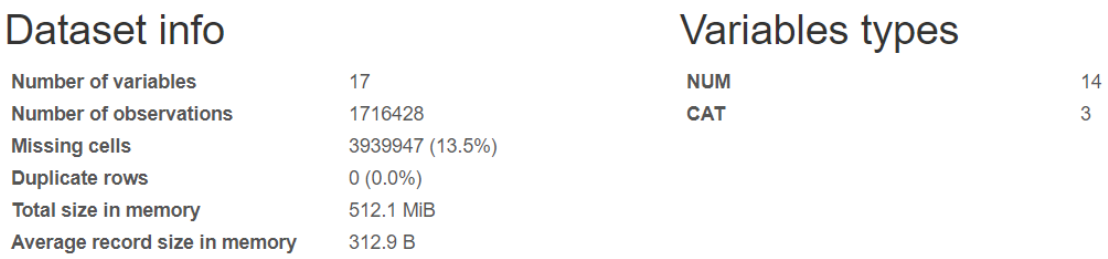


Εικόνα 14 - Συντελεστής συσχέτισης Spearman

Σε αυτό το dataset ως μεταβλητή κλειδί θα χρησιμοποιηθεί η “SK_ID_CURR”, αφού μοναδικοποιηθεί καθώς πρωτογενώς οι τιμές της επαναλαμβάνονται. Αυτό συμβαίνει διότι στο αρχείο αποτυπώνονται στοιχεία του ίδιου ID για διαφορετικά χρονικά στιγμιότυπα (snapshots) που διακρίνονται από τη στήλη “MONTHS_BALANCE”. Επαναλαμβάνεται δηλαδή το κάθε δάνειο και τα αντίστοιχα υπόλοιπα του πχ ένα μήνα πριν την αίτηση για δάνειο (“MONTHS_BALANCE”=-1), δύο μήνες πριν την αίτηση για δάνειο (“MONTHS_BALANCE”=-2) και ούτω καθεξής.

4.1.3 Σύνολο δεδομένων “bureau”

Το αρχείο “bureau” περιλαμβάνει όλες τις προηγούμενες πιστώσεις του πελάτη που παρέχονται από άλλα χρηματοπιστωτικά ιδρύματα και τηρούνται στο Credit Bureau[3] (αφορά μόνο τους πελάτες που έχουν δάνειο στο δείγμα της Τράπεζας εξωτερικού που εξετάζεται και όχι όλους τους πελάτες του Credit Bureau). Για κάθε δάνειο στο δείγμα, υπάρχουν πολλές σειρές με τον αριθμό των πιστώσεων που είχε ο πελάτης στο Credit Bureau πριν από την ημερομηνία υποβολής της αίτησης. Ουσιαστικά, το αρχείο αυτό αποτελεί το αντίστοιχο ελληνικό σύστημα Τειρεσίας[4]. Αριθμεί 1.716.428 εγγραφές και περιέχει 17 στήλες, εκ των οποίων οι 14 είναι αριθμητικού τύπου (numeric), και οι 3 αποτελούν κατηγορικές μεταβλητές (categories).



Dataset info

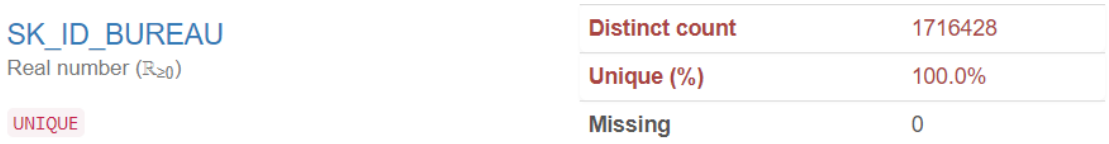
Number of variables	17
Number of observations	1716428
Missing cells	3939947 (13.5%)
Duplicate rows	0 (0.0%)
Total size in memory	512.1 MiB
Average record size in memory	312.9 B

Variables types

NUM	14
CAT	3

Εικόνα 15 - Σύνοψη περιεχομένου συνόλου bureau

Πρώτη είναι η στήλη “SK_ID_CURR” που χρησιμοποιείται επίσης στο σύνολο δεδομένων “application_training”. Επίσης, υπάρχει η “SK_ID_BUREAU” στήλη που δίνει σε κάθε δάνειο σε αυτό το σύνολο δεδομένων ένα μοναδικό id και αφορά τα δάνεια που έχει ο πελάτης στα άλλα χρηματοπιστωτικά ιδρύματα. Αποτελεί τη μεταβλητή κλειδί του αρχείου.



SK_ID_BUREAU

Real number ($\mathbb{R}_{\geq 0}$)

UNIQUE

Distinct count	1716428
Unique (%)	100.0%
Missing	0

Εικόνα 16 - Μεταβλητή-Κλειδί αρχείου bureau

Επιπλέον, το σύνολο δεδομένων περιέχει πληροφορίες σχετικά με τον τύπο του δανείου, το ποσό και τις πιθανές ημέρες καθυστέρησης κλπ. Τέλος, είναι σημαντικό να αναφερθεί ότι σε αυτό το σύνολο δεδομένων δεν περιέχεται η μεταβλητή “TARGET”.

Το “bureau” χρησιμοποιείται σαν ενδιάμεσο αρχείο κλειδί για να γίνει η ένωση των αρχείων “application_train” και “credit_card_balance” με το αρχείο “bureau_balance” που θα αναλυθεί παρακάτω.

4.1.4 Σύνολο δεδομένων “bureau_balance”

Το αρχείο “bureau_balance” περιλαμβάνει και αυτό πιστώσεις του πελάτη που παρέχονται από άλλα χρηματοπιστωτικά ιδρύματα και τηρούνται στο Credit Bureau και συγκεκριμένα περιέχει χρονικά στιγμιότυπα σε μήνες με το status του κάθε δανείου. Δηλαδή, έχει μια

[3] https://en.wikipedia.org/wiki/Credit_bureau

[4] http://www.tiresias.gr/

σειρά για κάθε ιστορικό μήνα κάθε προηγούμενης πίστωσης που τηρείται στο Credit Bureau. Αριθμεί 27.299.925 εγγραφές και 3 στήλες, εκ των οποίων οι 2 είναι αριθμητικού τύπου (numeric) και 1 είναι κατηγορική (categories).

Παρακάτω απεικονίζονται οι τιμές της μεταβλητής “STATUS” καθώς και οι ερμηνείες τους:

STATUS	**Ερμηνεία**
C	Κλειστός (Closed)
X	Άγνωστο (Unknown status)
0	0 μέρες καθυστέρησης
1	1-30 μέρες καθυστέρησης
2	30-60 μέρες καθυστέρησης
3	60-90 μέρες καθυστέρησης
4	90-120 μέρες καθυστέρησης
5	120+ μέρες καθυστέρησης ή πώληση ή διαγραφή

Όπως ακριβώς και στο αρχείο “credit_card_balance”, έτσι και εδώ η αντίστοιχη μεταβλητή κλειδί “SK_ID_BUREAU” επαναλαμβάνεται και δεν είναι μοναδικοποιημένη. Αυτό συμβαίνει διότι στο αρχείο αποτυπώνονται στοιχεία του ίδιου ID για διαφορετικά χρονικά στιγμιότυπα (snapshots) που διακρίνονται από τη στήλη “MONTHS_BALANCE”.

4.2 Προ-επεξεργασία δεδομένων

Σκοπός στο πρακτικό κομμάτι της εργασίας αυτής είναι να δημιουργηθούν όσο το δυνατόν περισσότερες μεταβλητές , που θα έχουν όμως υψηλή συσχέτιση μεταξύ τους ώστε να υπάρξει το καλύτερο δυνατό αποτέλεσμα στην πρόβλεψη που θα γίνει. Επομένως, στη συνέχεια θα γίνει μια παραγωγή νέων μεταβλητών από τις ήδη υπάρχουσες. Επίσης, θα εξαιρεθούν οι τιμές ‘XNA’ από τη μεταβλητή ‘CODE_GENDER’, καθώς δεν νοείται να υπάρχουν τιμές εκτός του F (female) και M (Male) και θα αντικατασταθούν οι τιμές ‘365243’ από τη μεταβλητή “DAYS_EMPLOYED” με ‘Nan’, γιατί 365.243 ημέρες απασχόλησης ισοδυναμούν με 1000 χρόνια και όπως είναι κατανοητό αυτό δεν είναι εφικτό. Οι τιμές αυτές είναι “ακραίες τιμές” και επηρεάζουν τα αποτελέσματά, εάν δεν εξαιρεθούν.

Ακολούθως, θα γίνει επεξεργασία των αρχείων “bureau” και “bureau_balance”. Αρχικά, γίνεται ένωση του “bureau” με το “credit_card_balance” στη κοινή μεταβλητή κλειδί των δύο αρχείων, την “SK_ID_CURR”, ώστε να μείνουν στο “bureau” μόνο τα δάνεια που αφορούν πιστωτικές κάρτες. Έπειτα, γίνεται ένωση του “bureau” με το “bureau_balance” στη μεταβλητή “SK_ID_BUREAU” και θα φιλτραριστούν οι εγγραφές έως και 4 χρόνια πριν. Επιπλέον, διαγράφονται όσοι λογαριασμοί είναι κλειστοί ή βρίσκονται σε άγνωστο στάτους. Επειδή το ID επαναλαμβάνεται θα μετατραπούν οι τιμές της μεταβλητής “STATUS” σε στήλες (εκτός από αυτές που διαγράφηκαν προηγουμένως), κρατώντας τον μέσο και την τυπική απόκλιση για καθεμία από αυτές ως προς το ID. Έτσι επιτυγχάνεται η μοναδικοποίηση της μεταβλητής στο αρχείο, που ονομάζεται “final”, για να μπορεί να γίνει στη συνέχεια η ένωση με τα άλλα αρχεία.

Στη συνέχεια, θα γίνει επεξεργασία του αρχείου “credit_card_balance”, όπου θα φιλτραριστούν και θα απομονωθούν οι εγγραφές έως και 4 χρόνια πριν. Επιπλέον, λόγω της επανάληψης τιμών στο ID για διαφορετικές χρονικές περιόδους όπως προαναφέρθηκε, θα παραχθούν νέες μεταβλητές όπως είναι το max, το mean, το sum και το var διαφόρων μετρικών ανά ID, με σκοπό τη μοναδικοποίηση αργότερα της μεταβλητής κλειδί.

Τα παραπάνω υλοποιούνται με τις κάτωθι εντολές:

```
# Preprocess credit_card_balance.csv

cc=cc[cc.MONTHS_BALANCE>=-48]

cc.drop(['SK_ID_PREV'], axis= 1, inplace = True)

cc_agg = cc.groupby('SK_ID_CURR')['AMT_BALANCE','CNT_DRAWINGS_ATM_CURRENT','AMT_PAYMENT_TOTAL_CURRENT','SK_DPD',

       'SK_DPD_DEF'].agg([ 'max', 'mean', 'sum', 'var'])
print(cc_agg)

cc_agg.columns = pd.Index(['CC_' + e[0] + "_" + e[1].upper() for e in cc_agg.columns.tolist()]) #enwnw tis epikefalides tou dataset

df=df.join(cc_agg,on='SK_ID_CURR',how='left',rsuffix='_agg')
```

Ακόμα, για σκοπούς ευκολίας της χρήσης των δεδομένων θα τροποποιηθεί η μεταβλητή “MONTHS_BALANCE” ώστε να παρουσιάζονται οι αρχικές τιμές της ομαδοποιημένες σε διαστήματα. Έτσι, από τέσσερα χρόνια πίσω έως δύο χρόνια πίσω θα λαμβάνει την τιμή ‘4’, από δύο χρόνια πίσω έως ένα χρόνο πίσω θα λαμβάνει την τιμή ‘3’, από ένα χρόνο πίσω έως έξι μήνες πίσω θα λαμβάνει την τιμή ‘2’ και από έξι μήνες πίσω έως σήμερα την τιμή ‘1’. Αντιστοίχως θα πραγματοποιηθεί το ίδιο και επί της μεταβλητής “SK_DPD_DEF”, όπου θα κατηγοριοποιηθούν σε εύρη οι ημέρες καθυστέρησης των δανείων.

Για να ολοκληρωθεί η προ-επεξεργασία στα σύνολα δεδομένων ώστε να ενοποιηθούν, θα πρέπει να υπάρχει η μεταβλητή κλειδί σε κάθε ένα από αυτά. Η μεταβλητή αυτή είναι η “SK_ID_CURR” σε όλα τα σύνολα δεδομένων, μόνο που στο αρχείο “credit_card_balance” δεν είναι ακόμα μοναδικοποιημένη και αυτό δεν θα επιτρέψει την ένωση. Για το λόγο αυτό θα χρησιμοποιηθεί η δυνατότητα της python να πραγματοποιηθεί pivot των δεδομένων, ουσιαστικά, μέσω αυτής της διαδικασίας, γίνεται ομαδοποίηση των δεδομένων ως προς την “SK_ID_CURR”, με αποτέλεσμα να επιτευχθεί η μοναδικοποίηση της εν λόγω μεταβλητής.

Εφόσον υλοποιηθούν όλα τα παραπάνω μπορεί να γίνει η ένωση των αρχείων “application_train” και “credit_card_balance” εκεί όπου η μεταβλητή “SK_ID_CURR” είναι κοινή στα δύο αρχεία, ώστε να παραχθεί ενοποιημένο αρχείο μόνο για τα δάνεια που αφορούν τις πιστωτικές κάρτες. Για να είναι βέβαιο ότι στο αρχείο δεν επαναλαμβάνεται καμία εγγραφή και επομένως ότι το ID είναι μοναδικοποιημένο, εκτελείται η εντολή:

```
assert df['SK_ID_CURR'].unique().shape[0]==df.shape[0]
```

Δεδομένου ότι δεν επιστρέφεται κάποιο σφάλμα, επιβεβαιώνεται ότι η μεταβλητή έχει μοναδικές τιμές. Στην πορεία, γίνεται η ένωση του αρχείου “final” me to “df”, ώστε να σχηματιστεί το τελικό ενοποιημένο αρχείο.

Γίνεται εύκολα αντιληπτό ότι λόγω του μεγάλου μεγέθους του ενοποιημένου αρχείου και των πολλών μεταβλητών που έχει (201 στήλες), θα πρέπει να καταργηθούν όσες δεν έχουν καμία χρησιμότητα στην ανάλυση. Παρακάτω οι μεταβλητές που θα διαγραφούν.

```
['SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'NAME_TYPE_SUITE', 'DAYS_ID_PUBLISH', 'REGION_RATING_CLIENT',
'REGION_RATING_CLIENT_W_CITY', 'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START',
'REG_REGION_NOT_LIVE_REGION', 'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION',
'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE',
'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG',
'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG',
'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG',
'APARTMENTS_MODE', 'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE',
'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE', 'FLOORSMIN_MODE',
'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE',
'NONLIVINGAREA_MODE', 'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI',
'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRANCES_MEDI', 'FLOORSMAX_MEDI',
'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI',
'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE',
'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
'DAYS_LAST_PHONE_CHANGE', 'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5',
'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9', 'FLAG_DOCUMENT_10',
'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15',
'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY',
'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
'AMT_REQ_CREDIT_BUREAU_YEAR']
```

Εικόνα 17 - Μεταβλητές προς διαγραφή, μη χρήσιμες για την ανάλυση

Πλέον στο dataset έχουν απομείνει οι ακόλουθες στήλες:

```
['TARGET' 'CODE_GENDER' 'FLAG_OWN_CAR' 'FLAG_OWN_REALTY' 'CNT_CHILDREN'
 'AMT_INCOME_TOTAL' 'AMT_CREDIT' 'AMT_ANNUITY' 'AMT_GOODS_PRICE'
 'NAME_INCOME_TYPE' 'NAME_EDUCATION_TYPE' 'NAME_FAMILY_STATUS'
 'NAME_HOUSING_TYPE' 'REGION_POPULATION_RELATIVE' 'DAYS_BIRTH'
 'DAYS_EMPLOYED' 'DAYS_REGISTRATION' 'OWN_CAR_AGE' 'FLAG_MOBIL'
 'FLAG_EMP_PHONE' 'FLAG_WORK_PHONE' 'FLAG_CONT_MOBILE' 'FLAG_PHONE'
 'FLAG_EMAIL' 'OCCUPATION_TYPE' 'CNT_FAM_MEMBERS' 'EXT_SOURCE_1'
 'EXT_SOURCE_2' 'EXT_SOURCE_3' 'NEW_CREDIT_TO_ANNUITY_RATIO'
 'NEW_CREDIT_TO_GOODS_RATIO' 'NEW_INC_PER_CHLD' 'NEW_INC_BY_ORG'
 'NEW_EMPLOY_TO_BIRTH_RATIO' 'NEW_ANNUITY_TO_INCOME_RATIO'
 'NEW_SOURCES_PROD' 'NEW_EXT_SOURCES_MEAN' 'NEW_SCORES_STD'
 'NEW_CAR_TO_BIRTH_RATIO' 'NEW_CAR_TO_EMPLOY_RATIO'
 'NEW_PHONE_TO_BIRTH_RATIO' 'NEW_PHONE_TO_BIRTH_RATIO_EMPLOYER'
 'NEW_CREDIT_TO_INCOME_RATIO' 'YEARLY_SUR' 'CC_AMT_BALANCE_MAX'
 'CC_AMT_BALANCE_MEAN' 'CC_AMT_BALANCE_SUM' 'CC_AMT_BALANCE_VAR'
 'CC_CNT_DRAWINGS_ATM_CURRENT_MAX' 'CC_CNT_DRAWINGS_ATM_CURRENT_MEAN'
 'CC_CNT_DRAWINGS_ATM_CURRENT_SUM' 'CC_CNT_DRAWINGS_ATM_CURRENT_VAR'
 'CC_AMT_PAYMENT_TOTAL_CURRENT_MAX' 'CC_AMT_PAYMENT_TOTAL_CURRENT_MEAN'
 'CC_AMT_PAYMENT_TOTAL_CURRENT_SUM' 'CC_AMT_PAYMENT_TOTAL_CURRENT_VAR'
 'CC_SK_DPD_MAX' 'CC_SK_DPD_MEAN' 'CC_SK_DPD_SUM' 'CC_SK_DPD_VAR'
 'CC_SK_DPD_DEF_MAX' 'CC_SK_DPD_DEF_MEAN' 'CC_SK_DPD_DEF_SUM'
 'CC_SK_DPD_DEF_VAR' 'amax_1.0' 'amax_2.0' 'amax_3.0' 'amax_4.0'
 'amin_1.0' 'amin_2.0' 'amin_3.0' 'amin_4.0' 'mean_-24' 'mean_-23'
 'mean_-22' 'mean_-21' 'mean_-20' 'mean_-19' 'mean_-18' 'mean_-17'
 'mean_-16' 'mean_-15' 'mean_-14' 'mean_-13' 'mean_-12' 'mean_-11'
 'mean_-10' 'mean_-9' 'mean_-8' 'mean_-7' 'mean_-6' 'mean_-5' 'mean_-4'
 'mean_-3' 'mean_-2' 'mean_-1' 'CC_STATUS_0_MEAN' 'CC_STATUS_0_STD'
 'CC_STATUS_1_MEAN' 'CC_STATUS_1_STD' 'CC_STATUS_2_MEAN' 'CC_STATUS_2_STD'
 'CC_STATUS_3_MEAN' 'CC_STATUS_3_STD' 'CC_STATUS_4_MEAN' 'CC_STATUS_4_STD'
 'CC_STATUS_5_MEAN' 'CC_STATUS_5_STD']
```

Εικόνα 18 - Στήλες ενοποιημένου συνόλου δεδομένων

Παρακάτω εμφανίζονται οι τύποι κάθε μιας από τις ανωτέρω στήλες:

```
TARGET                          int64
CODE_GENDER                    object
FLAG_OWN_CAR                   object
FLAG_OWN_REALTY                object
CNT_CHILDREN                    int64
AMT_INCOME_TOTAL              float64
AMT_CREDIT                    float64
AMT_ANNUITY                   float64
AMT_GOODS_PRICE               float64
NAME_INCOME_TYPE               object
NAME_EDUCATION_TYPE            object
NAME_FAMILY_STATUS             object
NAME_HOUSING_TYPE              object
REGION_POPULATION_RELATIVE    float64
DAYS_BIRTH                      int64
DAYS_EMPLOYED                 float64
DAYS_REGISTRATION             float64
OWN_CAR_AGE                   float64
FLAG_MOBIL                      int64
FLAG_EMP_PHONE                  int64
FLAG_WORK_PHONE                 int64
FLAG_CONT_MOBILE                int64
FLAG_PHONE                      int64
FLAG_EMAIL                      int64
OCCUPATION_TYPE                object
CNT_FAM_MEMBERS               float64
EXT_SOURCE_1                  float64
EXT_SOURCE_2                  float64
EXT_SOURCE_3                  float64
NEW_CREDIT_TO_ANNUITY_RATIO   float64
                                ...
amax_3.0                      float64
amax_4.0                      float64
amin_1.0                      float64
amin_2.0                      float64
amin_3.0                      float64
amin_4.0                      float64
mean_-24                      float64
mean_-23                      float64
mean_-22                      float64
mean_-21                      float64
mean_-20                      float64
mean_-19                      float64
mean_-18                      float64
mean_-17                      float64
mean_-16                      float64
mean_-15                      float64
mean_-14                      float64
mean_-13                      float64
mean_-12                      float64
mean_-11                      float64
mean_-10                      float64
mean_-9                       float64
mean_-8                       float64
mean_-7                       float64
mean_-6                       float64
mean_-5                       float64
mean_-4                       float64
mean_-3                       float64
mean_-2                       float64
mean_-1                       float64
```

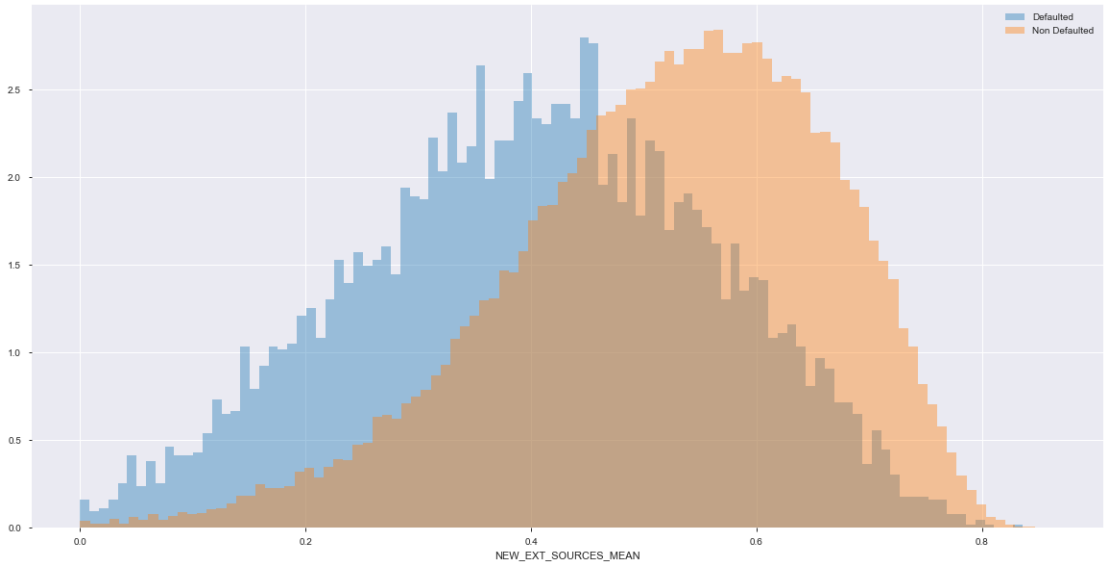
Εικόνα 19 - Τύποι πεδίων συνόλου δεδομένων

Παρατηρείται ότι κάποιες από τις μεταβλητές είναι τύπου object και θα γίνει η μετατροπή τους σε binary για να μην υπάρχει μεροληψία στα δεδομένα. Αφού διαγραφούν και οι μεταβλητές που έχουν τυπική απόκλιση ίση με '0', καθώς δεν συνεισφέρουν στη πρόβλεψη, απομένει το τελικό αρχείο με 86.905 γραμμές και 144 στήλες.

Κατόπιν εκτελείται στο αρχείο το τεστ Kolmogorov-Smirnov ώστε να γίνει η συσχέτιση μεταξύ των κατανομών των μεταβλητών. Στη στατιστική, το τεστ Kolmogorov-Smirnov[5] είναι ένα μη παραμετρικό τεστ της ισότητας των συνεχόμενων (ή μη συνεχών), μονοδιάστατων κατανομών πιθανοτήτων που μπορούν να χρησιμοποιηθούν για τη σύγκριση ενός δείγματος με μια κατανομή πιθανότητας (one-sample K–S test) ή για τη σύγκριση δύο δειγμάτων (two-sample K–S test). Ονομάστηκε έτσι από τον Andrey Kolmogorov και τον Nikolai Smirnov.

Ελέγχοντας τυχαία το αποτέλεσμα μιας μεταβλητής, εν προκειμένω τη μεταβλητή "NEW_EXT_SOURCES_MEAN", παρατηρείται ότι όσο πιο κοντά στο '1' είναι η τιμή στο Kolmogorov-Smirnov τεστ, τόσο μεγαλώνει η διασπορά των κατανομών για τις τιμές της "TARGET" (0,1) στο ιστόγραμμα.

```
Name: NEW_EXT_SOURCES_MEAN, dtype: float64
KS for NEW_EXT_SOURCES_MEAN : 0.32016594925437736
count   86905.000
mean        0.150
std         0.093
min         0.000
25%         0.081
50%         0.150
75%         0.202
max         0.619
```



Εικόνα 20 - Kolmogorov-Smirnov

Στην πορεία, χωρίζεται το αρχείο σε σύνολο εκπαίδευσης (train) (80%) και σύνολο ελέγχου (valid) (20%). Θα εξεταστεί ο μέσος όρος της μεταβλητής 'TARGET' στο train και το valid και το αποτέλεσμα είναι:

```
train mean: 0.08755249985616478, test mean: 0.08325182670732409
```

[5] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Αυτό δείχνει ότι ο μέσος όρος είναι πολύ κοντά στο train και στο valid, άρα είναι αντιπροσωπευτικό το δείγμα, δεδομένου ότι το sampling έχει γίνει με τυχαίο τρόπο.

Στο επόμενο κεφάλαιο και έπειτα από την προ-επεξεργασία των δεδομένων και την παραγωγή του τελικού συνόλου δεδομένων, θα ακολουθήσει η περιγραφή και η εφαρμογή των αλγορίθμων υπολογισμού της πιθανότητας αθέτησης πιστούχου.

4.3 Αλγόριθμος LightGBM

4.3.1 Εισαγωγή στον LightGBM

Η πρώτη πρόβλεψη για το PD του dataset θα γίνει με τον αλγόριθμο LightGBM.

Τι είναι όμως ο αλγόριθμος LightGBM και η τεχνική gradient boosting;

Ο LightGBM[6] ανήκει στην κατηγορία των gradient boosting αλγορίθμων που χρησιμοποιεί αλγόριθμο μάθησης βασισμένο σε δέντρα. Είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης, λόγω της αποτελεσματικότητάς, της ακρίβειας και της ερμηνείας του [25].

Το Gradient boosting[7] είναι μια τεχνική μηχανικής μάθησης για προβλήματα παλινδρόμησης και ταξινόμησης, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης, συνήθως δέντρων αποφάσεων. Χτίζει το μοντέλο με ένα εξελιγμένο τρόπο όπως και άλλες μέθοδοι boosting, και το γενικεύει επιτρέποντας τη βελτιστοποίηση μιας αυθαίρετης διαφοροποιήσιμης συνάρτησης απώλειας.

Η ιδέα του Gradient boosting προέκυψε από την παρατήρηση του Leo Breiman ότι το boosting μπορεί να ερμηνευθεί ως ένας αλγόριθμος βελτιστοποίησης σε μια κατάλληλη συνάρτηση κόστους. Οι gradient boosting αλγόριθμοι παλινδρόμησης αναπτύχθηκαν στη συνέχεια από τον Jerome H. Friedman [26], ταυτόχρονα με την γενικότερη λειτουργική θεώρηση του gradient boosting που ειπώθηκε από τους Llew Mason, Jonathan Baxter, Peter Bartlett και Marcus Frean [27]. Οι αλγόριθμοι που βελτιστοποιούν μια συνάρτηση κόστους σε σχέση με το χώρο της συνάρτησης, επιλέγουν με επαναληπτικό τρόπο μία αδύναμη υπόθεση που δείχνει την αρνητική gradient κατεύθυνση. Αυτή η λειτουργική όψη της gradient μεθόδου, στο πλαίσιο του boosting, έχει οδηγήσει στην ανάπτυξη boosting αλγορίθμων σε πολλούς τομείς της μηχανικής μάθησης και των στατιστικών πέρα από την παλινδρόμηση και την ταξινόμηση.

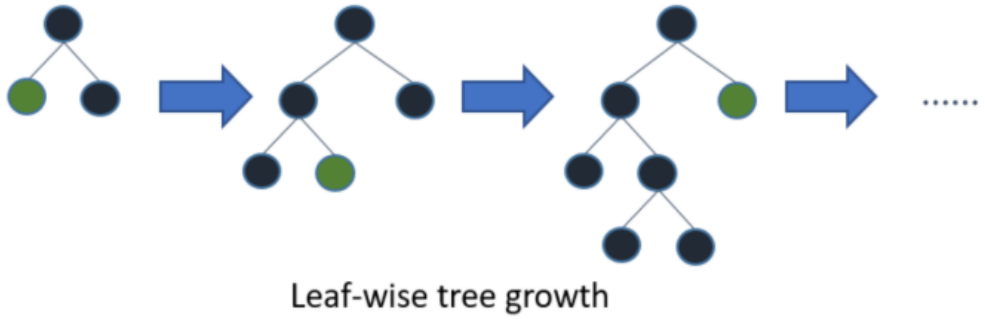
Πώς διαφέρει από άλλους αλγόριθμους που βασίζονται σε δέντρα;

Ο LightGBM μεγαλώνει δενδροειδώς κατακόρυφα, ενώ άλλοι αλγόριθμοι μεγαλώνουν δενδροειδώς οριζόντια που σημαίνει ότι ο LightGBM αναπτύσσεται σε επίπεδο φύλλου (leaf-wise) ενώ άλλοι αλγόριθμοι αναπτύσσονται σε επίπεδα βάθους (level-wise). Θα επιλέξει το φύλλο με τη μέγιστη διαφορά συναρτήσεως κόστους για να αναπτυχθεί. Όταν αναπτύσσεται το φύλλο, ο αλγόριθμος Leaf-wise μπορεί να μειώσει περισσότερο την απώλεια από έναν αλγόριθμο Level-wise, σε μικρότερο χρόνο.

Τα παρακάτω διαγράμματα εξηγούν την εφαρμογή του LightGBM και άλλων boosting αλγορίθμων.

[6] https://lightgbm.readthedocs.io/en/latest/

[7] https://en.wikipedia.org/wiki/Gradient_boosting



Explains how LightGBM works

How other boosting algorithm works

Εικόνα 21 - Ανάπτυξη LightGBM και λοιπών boosting αλγορίθμων

Γιατί ο LightGBM κερδίζει μεγάλη δημοτικότητα;

Το μέγεθος των δεδομένων αυξάνεται καθημερινά και καθίσταται δύσκολο για τους παραδοσιακούς αλγόριθμους επιστήμης δεδομένων να δίνουν ταχύτερα αποτελέσματα. Ο LightGBM έχει ονομαστεί ως «Light» λόγω της υψηλής ταχύτητας του και των μικρότερων απαιτήσεων σε μέγεθος μνήμης. Ένας άλλος λόγος για τον οποίο ο LightGBM είναι δημοφιλής είναι επειδή επικεντρώνεται στην ακρίβεια των αποτελεσμάτων. Ο LGBM υποστηρίζει επίσης την εκμάθηση της GPU και συνεπώς οι επιστήμονες δεδομένων χρησιμοποιούν ευρέως τον LGBM για την ανάπτυξη εφαρμογών επιστήμης δεδομένων.

Μπορούμε να χρησιμοποιήσουμε τον LightGBM παντού;

Όχι, δεν συνιστάται η χρήση του LGBM σε μικρά σύνολα δεδομένων. Ο LGBM είναι ευαίσθητος στην υπερπροσαρμογή (overfitting) και μπορεί εύκολα να υπερκεράσει μικρά δεδομένα.

Συζητήθηκε εν συντομία η έννοια του LightGBM, τώρα τι γίνεται με την εφαρμογή του;

Η υλοποίηση του LightGBM είναι εύκολη, το μόνο περίπλοκο πράγμα είναι η ρύθμιση των παραμέτρων. Ο LGBM καλύπτει περισσότερες από 100 παραμέτρους. Είναι πολύ σημαντικό για έναν αναλυτή να γνωρίζει τουλάχιστον κάποιες βασικές παραμέτρους του LightGBM. Θα γίνει παρακάτω μια ανάλυση των παραμέτρων.

Παράμετροι

(i) Παράμετροι ελέγχου

max_depth: Περιγράφει το μέγιστο βάθος του δέντρου. Αυτή η παράμετρος χρησιμοποιείται για την αντιμετώπιση της υπερπροσαρμογής του μοντέλου. Μικρότερα δέντρα μειώνουν την υπερπροσαρμογή.

min_data_in_leaf: Είναι ο ελάχιστος αριθμός των εγγραφών που μπορεί να έχει ένα φύλλο. Η προεπιλεγμένη τιμή είναι 20, η βέλτιστη τιμή. Χρησιμοποιείται επίσης για την αντιμετώπιση της υπερπροσαρμογής.

feature_fraction: Ελέγχει την υποδειγματοληψία των χαρακτηριστικών που χρησιμοποιούνται για την εκπαίδευση. Για παράδειγμα τιμή ίση με 0.8 σημαίνει ότι ο LightGBM θα επιλέξει το 80% των μεταβλητών τυχαία σε κάθε επανάληψη για την ανάπτυξη των δένδρων.

bagging_fraction: Καθορίζει το ποσοστό των δεδομένων που πρέπει να χρησιμοποιούνται για κάθε επανάληψη και χρησιμοποιείται γενικά για την επιτάχυνση της εκπαίδευσης και την αποφυγή overfitting.

early_stopping_round: Αυτή η παράμετρος μπορεί να βοηθήσει να επισπευστεί η ανάλυση. Το μοντέλο θα σταματήσει την εκπαίδευση εάν μια μέτρηση ενός συνόλου δεδομένων για test δεν βελτιωθεί στις τελευταίες επαναλήψεις “early_stopping_round”. Αυτό διασφαλίζει ότι δεν γίνεται overfitting.

lambda: Το lambda ορίζει το συντελεστή κανονικοποίησης. Η τυπική τιμή κυμαίνεται από 0 έως 1.

min_gain_to_split: Αυτή η παράμετρος περιγράφει το ελάχιστο όφελος για να κάνει ένα διαχωρισμό. Μπορεί να χρησιμοποιηθεί για τον έλεγχο του αριθμού των χρήσιμων χωρισμάτων στο δέντρο.

max_cat_group: Όταν ο αριθμός των κατηγορικών μεταβλητών είναι μεγάλος, η εύρεση του σημείου διαχωρισμού, ενδέχεται να οδηγήσει σε overfitting. Έτσι, ο LightGBM τους συγχωνεύει σε ομάδες 'max_cat_group' και βρίσκει τα σημεία διαχωρισμού στα όρια της ομάδας, default τιμή 64.

(ii) Βασικές παράμετροι

Task: Προσδιορίζει την εργασία που θα εκτελεστεί στα δεδομένα. Μπορεί να είναι είτε εκπαίδευση είτε πρόβλεψη.

Application: Αυτή είναι η πιο σημαντική παράμετρος και καθορίζει την εφαρμογή του μοντέλου, είτε πρόκειται για πρόβλημα παλινδρόμησης είτε για πρόβλημα ταξινόμησης. Ο LightGBM θα θεωρήσει από προεπιλογή το μοντέλο ως μοντέλο παλινδρόμησης. Οι τιμές που παίρνει η παράμετρος αυτή είναι:

- regression: για παλινδρόμηση

- binary: για δυαδική ταξινόμηση
- multiclass: για πρόβλημα ταξινόμησης με πολλές κλάσεις

boosting: ορίζει τον τύπο του αλγόριθμου που θα εκτελεστεί, default = gdbt. Οι τιμές που παίρνει η παράμετρος αυτή είναι:

- gbdt: Gradient Boosting Decision Tree
- rf: random forest
- dart: Dropouts meet Multiple Additive Regression Trees
- goss: Gradient-based One-Side sampling

num_boost_round: Αριθμός επαναλήψεων boosting, τυπικά 100+.

learning_rate: Αυτό καθορίζει τον αντίκτυπο κάθε δέντρου στο τελικό αποτέλεσμα. Ο GBM λειτουργεί ξεκινώντας με μια αρχική εκτίμηση, η οποία ενημερώνεται χρησιμοποιώντας την έξοδο κάθε δέντρου. Η παράμετρος μάθησης ελέγχει το μέγεθος αυτής της αλλαγής στις εκτιμήσεις. Τυπικές τιμές: 0.1, 0.001, 0.003.

num_leaves: αριθμός φύλλων- κόμβων σε ολόκληρο το δέντρο. Η κατοχή μεγάλου αριθμού φύλλων θα βελτιώσει την ακρίβεια, αλλά θα οδηγήσει επίσης σε υπερπροσαρμογή.

device: προεπιλογή είναι η cpu, μπορεί επίσης να περάσει στη gpu.

Μετρικές παράμετροι

metric: μια από τις σημαντικές παραμέτρους, καθώς καθορίζει το μέτρο με το οποίο γίνεται η αξιολόγηση του μοντέλου. Παρακάτω υπάρχουν μερικά μέτρα για παλινδρόμηση και ταξινόμηση.

mae: μέσο απόλυτο σφάλμα (mean absolute error)

mse: μέσο τετραγωνικό σφάλμα (mean squared error)

binary_logloss: απώλεια για δυαδική ταξινόμηση

multi_logloss: απώλεια για πολλαπλή ταξινόμηση

Πλεονεκτήματα αλγορίθμου LightGBM

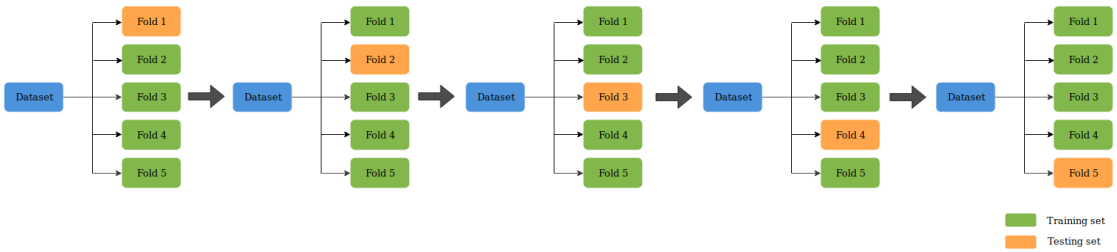
- Μεγάλη ταχύτητα εκπαίδευσης και υψηλή απόδοση.
- Χαμηλή χρήση μνήμης.
- Μεγαλύτερη ακρίβεια από οποιονδήποτε άλλον αλγόριθμο gradient boosting: Παράγει πολύ πιο πολύπλοκα δέντρα καθώς αναπτύσσεται σαν φύλλο (leaf-wise) ενώ άλλοι αλγόριθμοι αναπτύσσονται σε επίπεδα (level-wise), που είναι ο βασικός παράγοντας για την επίτευξη μεγαλύτερης ακρίβειας. Ωστόσο, μερικές φορές μπορεί να οδηγήσει σε overfitting που μπορεί να αποφευχθεί με τη ρύθμιση της παραμέτρου max_depth.
- Συμβατότητα με σύνολα μεγάλων δεδομένων: Είναι ικανός να αποδίδει εξίσου καλά με σύνολα μεγάλων δεδομένων με σημαντική μείωση του χρόνου εκπαίδευσης σε σύγκριση με τον XGBOOST αλγόριθμο.
- Υποστήριξη παράλληλης μάθησης και μάθησης GPU.

4.3.2 Εφαρμογή LightGBM στα δεδομένα της εργασίας

Περνώντας στην υλοποίηση του αλγορίθμου στα δεδομένα της εργασία, αρχικά θα δημιουργηθεί ένας κενός πίνακας με αριθμό γραμμών όσες και οι εγγραφές του συνόλου train (69.524) και άλλος ένας κενός πίνακας με αριθμό γραμμών όσες είναι οι εγγραφές του συνόλου test (17.381). Αυτοί οι δύο πίνακες δημιουργούνται με σκοπό να συμπληρωθούν εκεί τα αποτελέσματα της πρόβλεψης του αλγορίθμου. Στη συνέχεια, χωρίζεται το σύνολο του train (το οποίο αναφέρεται στο 80% του συνολικού πληθυσμού) σε 5 folds με την κάτωθι εντολή:

```
folds = KFold(n_splits= 5, shuffle=True, random_state=5)

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 παραπάνω εντολή έχει ως αποτέλεσμα, στην πρώτη επανάληψη το πρώτο fold να χρησιμοποιηθεί για τη δοκιμή (valid) του μοντέλου και τα υπόλοιπα τέσσερα να χρησιμοποιηθούν για την εκπαίδευση (train) του μοντέλου. Στη δεύτερη επανάληψη θα χρησιμοποιηθεί το δεύτερο fold ως σύνολο δοκιμών ενώ τα υπόλοιπα θα χρησιμεύσουν ως σύνολο εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται έως ότου κάθε fold από τα πέντε να έχει χρησιμοποιηθεί ως σύνολο δοκιμών. Με λίγα λόγια, χωρίζεται έτσι το αρχικό σύνολο εκπαίδευσης (train) σε επιμέρους train (80%) και valid (20%).



Εικόνα 22 - 5 Fold Cross Validation

Ακολούθως, εφαρμόζεται ο αλγόριθμος LightGBM με την κάτωθι εντολή:

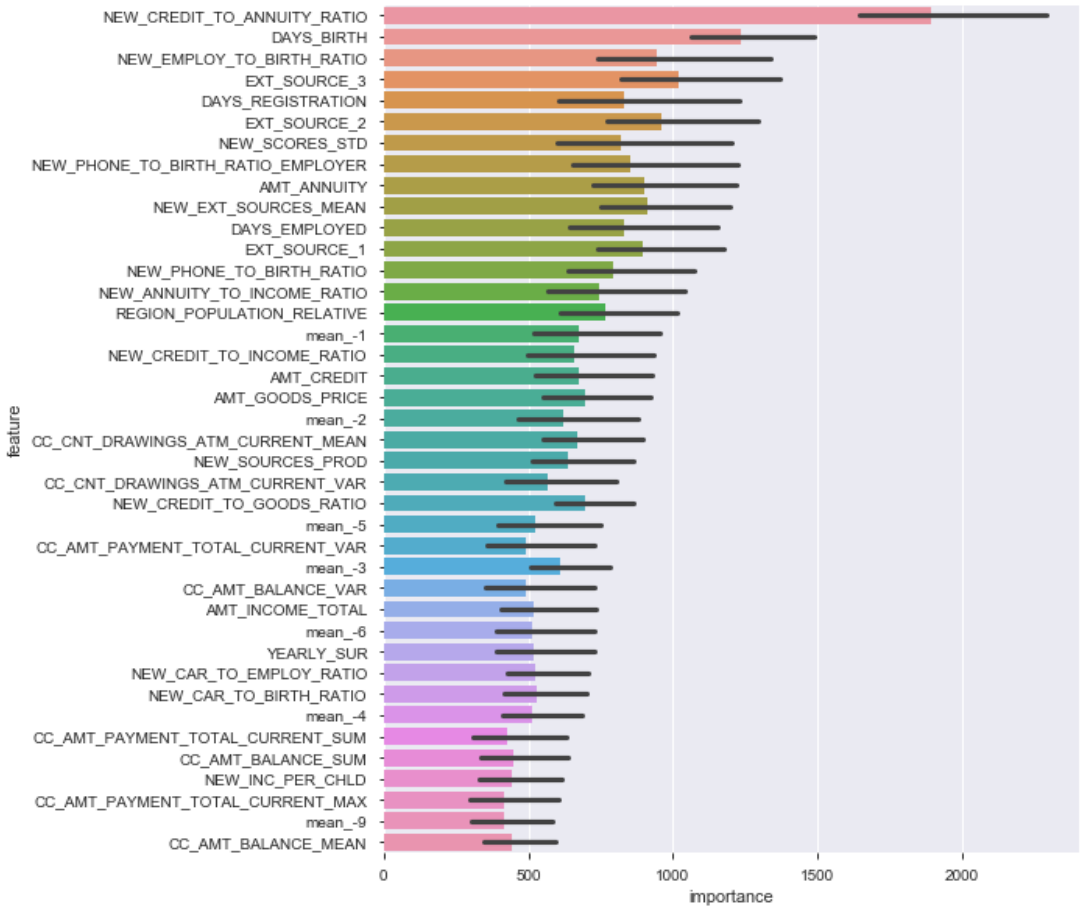
```
clf = LGBMClassifier(boosting= 'gbdt',
    nthread=4,
    n_estimators=10000,
    learning_rate=0.0137,
    num_leaves=36,#32
    colsample_bytree=0.2484,
    subsample=0.6386,
    max_depth=-1,#8
    reg_alpha=0.2041,#=0.2041
    reg_lambda=44.006,
    min_split_gain=0.2453,
    silent=-1,
    verbose=-1,
    random_seed=42,
    bagging_seed=42
    )

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)#na mhn kanei overfitting
oof_preds[valid_idx] = clf.predict_proba(valid_x, num_iteration=clf.best_iteration_)[:, 1]
sub_preds += clf.predict_proba(x_valid, num_iteration=clf.best_iteration_)[:, 1] / folds.n_splits
```

Η επεξήγηση των παραμέτρων έχει γίνει νωρίτερα σε αυτήν την ενότητα και οι τιμές τους επελέγησαν μετά από δοκιμές (trial and error). Επιπροσθέτως, δημιουργείται παράλληλα σε κάθε επανάληψη ένα κενό Dataframe, όπου περιλαμβάνει τις μεταβλητές του αρχικού συνόλου train (εκτός της μεταβλητής “TARGET”) και τη σημαντικότητά τους. Με την έννοια σημαντικότητα νοείται το πόσες φορές έχει χρησιμοποιηθεί η κάθε μεταβλητή για να διαχωρίσει τα δέντρα του αλγορίθμου. Αυτή η διαδικασία επαναλαμβάνεται και για τα 5 folds. Έτσι, μετά το πέρας των επαναλήψεων το Dataframe αυτό παρουσιάζει τη συνολική σημαντικότητα των μεταβλητών. Αυτό γίνεται με την κάτωθι εντολή:

```
fold_importance_df = pd.DataFrame()
fold_importance_df["feature"] = x_train.columns
fold_importance_df["importance"] = clf.feature_importances_
fold_importance_df
fold_importance_df["fold"] = n_fold + 1
feature_importance_df = pd.concat([feature_importance_df, fold_importance_df], axis=0)
```

Ακολουθεί η εντολή και το διάγραμμα από τους μέσους των 40 πιο σημαντικών μεταβλητών:

```
cols = feature_importance_df[["feature", "importance"]].groupby("feature").mean().sort_values(by="importance", ascending=False)[:40].index
best_features = feature_importance_df.loc[feature_importance_df.feature.isin(cols)]
plt.figure(figsize=(8, 10))
sns.barplot(x="importance", y="feature", data=best_features.sort_values(by="importance", ascending=False))
```



Εικόνα 23 - Μέσες τιμές σημαντικότερων μεταβλητών

Το αποτέλεσμα (ένα μέρος) της εκτέλεσης του αλγορίθμου είναι το κάτωθι:

```
Fold  1 AUC : 0.773895
Training until validation scores don't improve for 200 rounds
[100]   training's auc: 0.777077        training's binary_logloss: 0.264528     valid_1's auc: 0.759282 valid_1's binary_logloss: 0.271998
[200]   training's auc: 0.791306        training's binary_logloss: 0.253139     valid_1's auc: 0.765001 valid_1's binary_logloss: 0.264205
[300]   training's auc: 0.803726        training's binary_logloss: 0.246271     valid_1's auc: 0.768766 valid_1's binary_logloss: 0.261043
[400]   training's auc: 0.814604        training's binary_logloss: 0.241077     valid_1's auc: 0.770888 valid_1's binary_logloss: 0.259553
[500]   training's auc: 0.824401        training's binary_logloss: 0.236639     valid_1's auc: 0.772022 valid_1's binary_logloss: 0.258841
[600]   training's auc: 0.833337        training's binary_logloss: 0.232578     valid_1's auc: 0.772369 valid_1's binary_logloss: 0.258491
[700]   training's auc: 0.841713        training's binary_logloss: 0.228876     valid_1's auc: 0.772403 valid_1's binary_logloss: 0.258355
[800]   training's auc: 0.84951 training's binary_logloss: 0.225328     valid_1's auc: 0.772541 valid_1's binary_logloss: 0.258256
[900]   training's auc: 0.856779        training's binary_logloss: 0.222043     valid_1's auc: 0.772641 valid_1's binary_logloss: 0.258238
[1000]  training's auc: 0.863459        training's binary_logloss: 0.21897      valid_1's auc: 0.772698 valid_1's binary_logloss: 0.258254
Early stopping, best iteration is:
[890]   training's auc: 0.856124        training's binary_logloss: 0.222343     valid_1's auc: 0.772708 valid_1's binary_logloss: 0.258222
Fold  2 AUC : 0.772708
Training until validation scores don't improve for 200 rounds
[100]   training's auc: 0.780046        training's binary_logloss: 0.264062     valid_1's auc: 0.744127 valid_1's binary_logloss: 0.273156
[200]   training's auc: 0.793667        training's binary_logloss: 0.252589     valid_1's auc: 0.751521 valid_1's binary_logloss: 0.266135
[300]   training's auc: 0.805726        training's binary_logloss: 0.245766     valid_1's auc: 0.756689 valid_1's binary_logloss: 0.263271
[400]   training's auc: 0.816445        training's binary_logloss: 0.240567     valid_1's auc: 0.759627 valid_1's binary_logloss: 0.261833
[500]   training's auc: 0.825887        training's binary_logloss: 0.23619      valid_1's auc: 0.761557 valid_1's binary_logloss: 0.261008
[600]   training's auc: 0.83457 training's binary_logloss: 0.232223     valid_1's auc: 0.76271  valid_1's binary_logloss: 0.26054
[700]   training's auc: 0.842676        training's binary_logloss: 0.228556     valid_1's auc: 0.763559 valid_1's binary_logloss: 0.260218
[800]   training's auc: 0.850311        training's binary_logloss: 0.225092     valid_1's auc: 0.763859 valid_1's binary_logloss: 0.260117
[900]   training's auc: 0.857292        training's binary_logloss: 0.221848     valid_1's auc: 0.764096 valid_1's binary_logloss: 0.260038
[1000]  training's auc: 0.86396 training's binary_logloss: 0.218779     valid_1's auc: 0.764163 valid_1's binary_logloss: 0.260041
[1100]  training's auc: 0.869971        training's binary_logloss: 0.21593      valid_1's auc: 0.764018 valid_1's binary_logloss: 0.260098
Early stopping, best iteration is:
[945]   training's auc: 0.860375        training's binary_logloss: 0.220452     valid_1's auc: 0.764301 valid_1's binary_logloss: 0.259988
Fold  3 AUC : 0.764301
Training until validation scores don't improve for 200 rounds
[100]   training's auc: 0.780098        training's binary_logloss: 0.266258     valid_1's auc: 0.745417 valid_1's binary_logloss: 0.264385
[200]   training's auc: 0.793864        training's binary_logloss: 0.254624     valid_1's auc: 0.7518   valid_1's binary_logloss: 0.257543
[300]   training's auc: 0.80611 training's binary_logloss: 0.247666     valid_1's auc: 0.756233 valid_1's binary_logloss: 0.254811
[400]   training's auc: 0.816732        training's binary_logloss: 0.242423     valid_1's auc: 0.758495 valid_1's binary_logloss: 0.253563
[500]   training's auc: 0.82588 training's binary_logloss: 0.237997     valid_1's auc: 0.759535 valid_1's binary_logloss: 0.252977
[600]   training's auc: 0.834431        training's binary_logloss: 0.234031     valid_1's auc: 0.760414 valid_1's binary_logloss: 0.252587
[700]   training's auc: 0.842285        training's binary_logloss: 0.230366     valid_1's auc: 0.760876 valid_1's binary_logloss: 0.25239
```

Εικόνα 24 - Αποτέλεσμα αλγορίθμου LightGBM

Παρατηρείται ότι η εκπαίδευση συνεχίζεται σε κάθε επανάληψη μέχρις ότου το αποτέλεσμα του validation να μην βελτιώνεται παραπάνω για 200 γύρους. Αυτό επιλέχθηκε όταν στην παράμετρο 'early_stopping_rounds' ετέθη τιμή ίση με 200.

Τέλος, εκτελέστηκε η παρακάτω εντολή για να αποτυπωθεί το τελικό AUC σκορ πρόβλεψης του μοντέλου:

```
full_auc_score_LIGHTgbm = ('Full AUC score %.6f' % roc_auc_score(y_train, oof_preds))
print(full_auc_score_LIGHTgbm)
test_auc_score_LIGHTgbm = ('Test AUC score %.6f' % roc_auc_score(y_valid, sub_preds))
print(test_auc_score_LIGHTgbm)
```

```
Full AUC score 0.768204
Test AUC score 0.765927
```

Όπως φαίνεται και από τα παραπάνω, ο αλγόριθμος πέτυχε πρόβλεψη με AUC score = 0.765927. Συγκρίνοντάς το με την πρόβλεψη στο αρχικό 80% του πληθυσμού (AUC score = 0.768204), παρατηρείται ότι οι δύο τιμές είναι πολύ κοντά και επομένως το μοντέλο δεν έκανε overfitting, ενώ μπορεί να ειπωθεί ότι γενικεύει επιτυχώς σε νέα δεδομένα.

4.4 Αλγόριθμος Catboost

4.4.1 Εισαγωγή στον Catboost

Ο CatBoost[8] είναι ένας αλγόριθμος τεχνικής gradient boosting στα δέντρα αποφάσεων. Αναπτύχθηκε από τους ερευνητές και τους μηχανικούς της Yandex [28], είναι ο διάδοχος του αλγορίθμου MatrixNet που χρησιμοποιείται ευρέως για την ταξινόμηση εργασιών, την πρόβλεψη και τη διατύπωση συστάσεων. Είναι παγκόσμιος και μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα τομέων και σε διάφορα προβλήματα.

Είναι ένας σχετικά πρόσφατος αλγόριθμος μηχανικής μάθησης που μπορεί να λειτουργήσει με εργαλεία βαθιάς μάθησης όπως το Google TensorFlow και το Apple Core ML. Είναι ιδιαίτερα ισχυρός αλγόριθμος με δύο τρόπους:

- Παρέχει τα καλύτερα αποτελέσματα τελευταίας τεχνολογίας χωρίς εκτεταμένη εκπαίδευση δεδομένων που απαιτείται συνήθως από άλλες μεθόδους μηχανικής μάθησης και
- Παρέχει ισχυρή υποστήριξη για τις πιο περιγραφικές μορφές δεδομένων που συνοδεύουν πολλά επιχειρηματικά προβλήματα

Το όνομα "CatBoost" προέρχεται από τις δύο λέξεις "Category (Κατηγορία)" και "Boosting (Ενίσχυση)". Το "Boost" προέρχεται από gradient boosting αλγόριθμο μηχανικής μάθησης, καθώς αυτή η βιβλιοθήκη βασίζεται στη τεχνική gradient boosting. Ο gradient boosting αποτελεί έναν ισχυρό αλγόριθμο μηχανικής μάθησης που εφαρμόζεται ευρέως σε πολλαπλούς τύπους επιχειρησιακών προκλήσεων όπως ανίχνευση απάτης, συστάσεις και προβλέψεις και αποδίδει καλά. Μπορεί επίσης να επιστρέψει πολύ καλό αποτέλεσμα με σχετικά λιγότερα δεδομένα, σε αντίθεση με τα μοντέλα DL που πρέπει να μάθουν από ένα τεράστιο όγκο δεδομένων.

Ο Catboost εισάγει δύο κρίσιμες αλγοριθμικές εξελίξεις - από τη μια την εφαρμογή του ordered boosting (boosting στη σειρά), μιας εναλλαγής που οδηγείται από τον κλασικό αλγόριθμο και από την άλλη ενός καινοτόμου αλγορίθμου για την επεξεργασία κατηγορικών χαρακτηριστικών [28]. Και οι δύο τεχνικές χρησιμοποιούν τυχαίες μεταβολές

[8] https://catboost.ai/docs/

των δειγμάτων εκπαίδευσης για την καταπολέμηση της μετατόπισης πρόβλεψης που προκαλείται από ένα είδος διαρροής στόχου που υπάρχει σε όλες τις υπάρχουσες υλοποιήσεις αλγορίθμων gradient boosting.

Οι περισσότεροι από τους αλγορίθμους GBDT είναι ήδη εξοικειωμένοι με τη χρήση της στατιστικής Target (ή της μέσης κωδικοποίησης στόχου). Είναι μια απλή αλλά αποτελεσματική προσέγγιση στην οποία κωδικοποιείται κάθε κατηγορηματικό χαρακτηριστικό με την εκτίμηση του αναμενόμενου στόχου y που καθορίζεται από την κατηγορία. Αποδεικνύεται ότι η εφαρμογή αυτής της κωδικοποίησης απρόσεκτα (μέση τιμή του y πάνω από τα παραδείγματα εκπαίδευσης με την ίδια κατηγορία) οδηγεί σε διαρροή στόχου. Για την καταπολέμηση αυτής της μετατόπισης πρόβλεψης, ο CatBoost χρησιμοποιεί μια πιο αποτελεσματική στρατηγική [28]. Στηρίζεται στην αρχή του ordering και εμπνέεται από τους αλγόριθμους εκμάθησης που παραδίδουν τα παραδείγματα εκπαίδευσης διαδοχικά στο χρόνο. Σε αυτή τη ρύθμιση, οι τιμές της στατιστικής Target για κάθε παράδειγμα βασίζονται μόνο στο ιστορικό που παρατηρείται. Για να προσαρμοστεί αυτή η ιδέα σε μια τυπική ρύθμιση εκτός σύνδεσης, ο Catboost εισάγει έναν τεχνητό "χρόνο" - μια τυχαία μετατόπιση σ1 των παραδειγμάτων εκπαίδευσης. Στη συνέχεια, για κάθε παράδειγμα, χρησιμοποιεί όλη τη διαθέσιμη "ιστορία" για να υπολογίσει την στατιστική Target. Να σημειωθεί ότι, χρησιμοποιώντας μόνο μία τυχαία μετάθεση, οδηγεί σε προηγούμενα παραδείγματα με μεγαλύτερη διακύμανση στη στατιστική Target από τα επόμενα. Για το σκοπό αυτό, ο CatBoost χρησιμοποιεί διαφορετικές παραλλαγές για διαφορετικά βήματα gradient boosting.

Ο CatBoost έχει δύο τρόπους για να επιλέξει τη δομή του δέντρου, την Ordered και την Plain. Η Plain λειτουργία αντιστοιχεί σε συνδυασμό του τυπικού αλγορίθμου GBDT με μια ταξινομημένη στατιστική Target. Στην Ordered λειτουργία boosting εκτελείται μια τυχαία μετάθεση των παραδειγμάτων εκπαίδευσης - σ2, και διατηρούνται n διαφορετικά μοντέλα υποστήριξης -M1,. . ., Mn έτσι ώστε το μοντέλο Mi να εκπαιδεύεται χρησιμοποιώντας μόνο τα πρώτα δείγματα i στη μετάθεση [28]. Δυστυχώς, αυτός ο αλγόριθμος δεν είναι εφικτός στις περισσότερες πρακτικές εργασίες λόγω της ανάγκης διατήρησης n διαφορετικών μοντέλων, που αυξάνουν τις απαιτήσεις πολυπλοκότητας και μνήμης κατά n φορές. Ο Catboost εφαρμόζει μια τροποποίηση αυτού του αλγορίθμου, με βάση τον αλγόριθμο gradient boosting,, χρησιμοποιώντας μια δομή δέντρου που μοιράζονται όλα τα μοντέλα που πρόκειται να κατασκευαστούν.

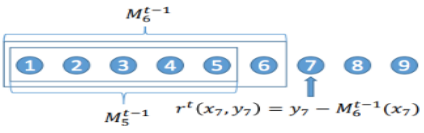


Figure 1: Ordered boosting principle.

Algorithm 1: Ordered boosting

input : $\{(\mathbf{x}_k, y_k)\}_{k=1}^n$, I;
$\sigma \leftarrow$ random permutation of $[1, n]$;
$M_i \leftarrow 0$ for $i = 1..n$;
for $t \leftarrow 1$ **to** I **do**
  for $i \leftarrow 1$ **to** n **do**
    $r_i \leftarrow y_i - M_{\sigma(i)-1}(i)$;
  for $i \leftarrow 1$ **to** n **do**
    $\Delta M \leftarrow$ $LearnModel((\mathbf{x}_j, r_j) : \sigma(j) \leq i)$;
    $M_i \leftarrow M_i + \Delta M$;
return M_n

Algorithm 2: Building a tree in CatBoost

input : $M, \{y_i\}_{i=1}^n, \alpha, L, \{\sigma_i\}_{i=1}^s, Mode$
$grad \leftarrow CaclGradient(L, M, y)$;
$r \leftarrow random(1, s)$;
$G \leftarrow (grad_r(1), \ldots, grad_r(n))$ for $Plain$;
$G \leftarrow (grad_{r,\sigma_r(1)-1}(i)$ for $i = 1$ to $n)$ for $Ordered$;
$T \leftarrow$ empty tree;
foreach *step of top-down procedure* **do**
  foreach *candidate split* c **do**
    $T_c \leftarrow$ add split c to T;
    if $Mode == Plain$ **then**
      $\Delta(i) \leftarrow \text{avg}(grad_r(p)$ for $p : leaf(p) = leaf(i))$ for all i;
    if $Mode == Ordered$ **then**
      $\Delta(i) \leftarrow \text{avg}(grad_{r,\sigma_r(i)-1}(p)$ for $p : leaf(p) = leaf(i), \sigma_r(p) < \sigma_r(i)) \ \forall i$;
    $loss(T_c) \leftarrow ||\Delta - G||_2$
  $T \leftarrow \text{argmin}_{T_c}(loss(T_c))$
if $Mode == Plain$ **then**
  $M_{r'}(i) \leftarrow M_{r'}(i) - \alpha \,\text{avg}(grad_{r'}(p)$ for $p : leaf(p) = leaf(i))$ for all r', i;
if $Mode == Ordered$ **then**
  $M_{r',j}(i) \leftarrow M_{r',j}(i) - \alpha \,\text{avg}(grad_{r',j}(p)$ for $p : leaf(p) = leaf(i), \sigma_{r'}(p) \leq j$ for all r', j, i;
return T, M

Εικόνα 25 - Catboost Ordered Boosting and Tree Building

Προκειμένου να αποφευχθεί η μεταβολή πρόβλεψης, ο Catboost χρησιμοποιεί μεταθέσεις τέτοιες ώστε σ1 = σ2. Αυτό εγγυάται ότι το target-Yi δεν χρησιμοποιείται για την κατάρτιση Mi, ούτε για τον υπολογισμό της Target Statistic και ούτε για την εκτίμηση κλίσης.

Σημαντικές Παράμετροι

cat_features: Αυτή η παράμετρος είναι απαραίτητη προκειμένου να επιτευχθεί η προεπεξεργασία του Catboost για τα κατηγορικά χαρακτηριστικά.

one_hot_max_size: Ο αλγόριθμος Catboost χρησιμοποιεί την κωδικοποίηση one-hot-encoding (τεχνική για κατηγορικά χαρακτηριστικά) για όλα τα χαρακτηριστικά με μοναδικές τιμές one_hot_max_size.

learning_rate & n_estimators: Όσο μικρότερος είναι ο ρυθμός μάθησης (learning rate), τόσο περισσότερα n_stimators πρέπει να χρησιμοποιηθούν στο μοντέλο. Συνήθως, η προσέγγιση είναι να ξεκινάει κάποιος με ένα σχετικά υψηλό ρυθμό μάθησης, μετά να συντονίζονται άλλες παράμετροι και στη συνέχεια να μειώνεται ο ρυθμός μάθησης ενώ αυξάνονται τα n_estimators.

max_depth: Βάθος δέντρου. Αυτή η παράμετρος έχει μεγάλη επίδραση στον χρόνο εκπαίδευσης.

subsample: Το ποσοστό των επιμέρους δειγμάτων. Δεν μπορεί να χρησιμοποιηθεί σε μια ρύθμιση Bayesian boost type.

colsample_bylevel, colsample_bytree, colsample_bynode: Ο ρυθμός δειγματοληψίας των στηλών ανά δέντρο, επίπεδο και node.

l2_leaf_reg: L2 συντελεστής κανονικοποίησης.

random_strength: Κάθε χώρισμα (split) παίρνει ένα σκορ και η παράμετρος random_strength προσθέτει κάποια τυχαιότητα στο σκορ. Βοηθά να μειωθεί το overfitting.

iterations: Μέγιστος αριθμός δέντρων που μπορούν να δημιουργηθούν. Υψηλές τιμές εδώ μπορούν να οδηγήσουν σε overfitting.

Πλεονεκτήματα αλγόριθμου Catboost

- Εξαιρετική ποιότητα σε σύγκριση με άλλες βιβλιοθήκες GBDT σε πολλά σύνολα δεδομένων.
- Πολύ καλή ταχύτητα στην πρόβλεψη.
- Υποστήριξη τόσο για αριθμητικά όσο και για κατηγορικά χαρακτηριστικά. Ο Catboost χειρίζεται αυτόματα τις κατηγορικές μεταβλητές. Μετατρέπει μη αριθμητικές μεταβλητές σε αριθμητικές με τη χρήση διαφόρων στατιστικών στοιχείων που βοηθούν να ξεπεραστεί η προ-επεξεργασία των δεδομένων για να μετατραπούν οι κατηγορίες σε αριθμητικές.
- Γρήγορη υποστήριξη από GPU και multi-GPU για την εκπαίδευση.
- Περιλαμβάνονται visualization εργαλεία.
- Μειώνει το overfitting: Ο Catboost αποφεύγει την υπερπροσαρμογή του μοντέλου με τη βοήθεια ανιχνευτή υπερπροσαρμογής που οδηγεί σε πιο γενικευμένα μοντέλα. Βασίζεται σε έναν αποκλειστικό αλγόριθμο για την κατασκευή μοντέλων που διαφέρουν από τους τυπικούς gradient boosting.

4.4.2 Εφαρμογή Catboost στα δεδομένα της εργασίας

Η υλοποίηση του αλγορίθμου Catboost γίνεται με τον ίδιο τρόπο και την ίδια λογική όπως με τον LightGBM. Αρχικά δημιουργούνται δύο κενοί πίνακες, ένας για το σύνολο train και ένας για το σύνολο test, με αριθμό γραμμών όσο και το κάθε σύνολο αντίστοιχα για να συμπληρωθούν εκεί οι προβλέψεις του μοντέλου. Επιπλέον, χωρίζεται και εδώ το σύνολο του train σε πέντε folds και ύστερα υλοποιείται ο αλγόριθμος Catboost με την εξής εντολη:

```
clf = CatBoostClassifier(iterations=7000,
                         learning_rate=0.02,
                         depth=5,
                         loss_function='Logloss',
                         l2_leaf_reg=50,
                         eval_metric='AUC',
                         random_seed = 12,
                         subsample=0.65,
                         nan_mode='Min',
                         bootstrap_type='Bernoulli',
                          random_state=12,
                         od_type='Iter',
                         metric_period = 200,
                         od_wait=150,
                         use_best_model=False)

clf.fit(train_x, train_y, eval_set=[(train_x, train_y), (valid_x, valid_y)])#na mhn kanei overfitting

oof_preds2[valid_idx] = clf.predict_proba(valid_x)[:, 1]

sub_preds2 += clf.predict_proba(x_valid)[:, 1] / folds.n_splits


print('Fold %2d AUC : %.6f' % (n_fold + 1, roc_auc_score(valid_y, oof_preds2[valid_idx])))

del clf, train_x, train_y, valid_x, valid_y
```

Ένα μέρος του αποτελέσματος εκτέλεσης του κώδικα είναι το κάτωθι:

```
Warning: Overfitting detector is active, thus evaluation metric is calculated on every iteration. 'metric_period' is ignored for evaluation metric.
0:      test: 0.5991832 test1: 0.5988545        best: 0.5988545 (0)     total: 50.7ms   remaining: 5m 54s
200:    test: 0.7571448 test1: 0.7597850        best: 0.7597850 (200)   total: 6.04s    remaining: 3m 24s
400:    test: 0.7707679 test1: 0.7663489        best: 0.7663489 (400)   total: 11.8s    remaining: 3m 14s
600:    test: 0.7785850 test1: 0.7684919        best: 0.7684919 (600)   total: 17.5s    remaining: 3m 5s
800:    test: 0.7834824 test1: 0.7698256        best: 0.7698256 (800)   total: 23.3s    remaining: 3m
1000:   test: 0.7870920 test1: 0.7704805        best: 0.7704845 (996)   total: 28.9s    remaining: 2m 53s
1200:   test: 0.7908760 test1: 0.7709226        best: 0.7709226 (1199)  total: 34.6s    remaining: 2m 46s
1400:   test: 0.7942843 test1: 0.7714133        best: 0.7714133 (1400)  total: 40.4s    remaining: 2m 41s
1600:   test: 0.7979956 test1: 0.7717517        best: 0.7717634 (1582)  total: 46s      remaining: 2m 35s
1800:   test: 0.8020845 test1: 0.7722889        best: 0.7722889 (1800)  total: 51.7s    remaining: 2m 29s
2000:   test: 0.8055834 test1: 0.7725884        best: 0.7725884 (2000)  total: 57.6s    remaining: 2m 23s
2200:   test: 0.8085712 test1: 0.7728210        best: 0.7728321 (2189)  total: 1m 3s    remaining: 2m 17s
2400:   test: 0.8122587 test1: 0.7729501        best: 0.7730468 (2313)  total: 1m 9s    remaining: 2m 12s
2600:   test: 0.8152749 test1: 0.7730825        best: 0.7731542 (2552)  total: 1m 14s   remaining: 2m 6s
Stopped by overfitting detector  (150 iterations wait)

bestTest = 0.7731542138
bestIteration = 2552

Fold  2 AUC : 0.773064
Warning: Overfitting detector is active, thus evaluation metric is calculated on every iteration. 'metric_period' is ignored for evaluation metric.
0:      test: 0.6034667 test1: 0.5999472        best: 0.5999472 (0)     total: 33.5ms   remaining: 3m 54s
200:    test: 0.7614124 test1: 0.7451340        best: 0.7451340 (200)   total: 5.85s    remaining: 3m 17s
400:    test: 0.7744870 test1: 0.7519903        best: 0.7519903 (400)   total: 11.7s    remaining: 3m 11s
600:    test: 0.7832074 test1: 0.7553336        best: 0.7553336 (600)   total: 17.3s    remaining: 3m 4s
800:    test: 0.7890422 test1: 0.7568736        best: 0.7568737 (799)   total: 23.2s    remaining: 2m 59s
1000:   test: 0.7940207 test1: 0.7579687        best: 0.7579687 (1000)  total: 28.9s    remaining: 2m 53s
1200:   test: 0.7982888 test1: 0.7586824        best: 0.7586824 (1200)  total: 34.5s    remaining: 2m 46s
1400:   test: 0.8024316 test1: 0.7591610        best: 0.7592086 (1382)  total: 40.5s    remaining: 2m 41s
1600:   test: 0.8067867 test1: 0.7596362        best: 0.7596464 (1558)  total: 46.2s    remaining: 2m 35s
1800:   test: 0.8107682 test1: 0.7600505        best: 0.7600505 (1800)  total: 51.9s    remaining: 2m 29s
2000:   test: 0.8142998 test1: 0.7603103        best: 0.7603358 (1986)  total: 57.8s    remaining: 2m 24s
2200:   test: 0.8179660 test1: 0.7602487        best: 0.7604372 (2114)  total: 1m 3s    remaining: 2m 18s
2400:   test: 0.8218573 test1: 0.7604819        best: 0.7605246 (2368)  total: 1m 8s    remaining: 2m 12s
Stopped by overfitting detector  (150 iterations wait)

bestTest = 0.7605245943
bestIteration = 2368

Fold  3 AUC : 0.760520
```

Εικόνα 26 - Μέρος αποτελέσματος υλοποίησης Catboost

Όπως αναφέρθηκε ανωτέρω, ο Catboost είναι ένας αλγόριθμος που δίνει τη δυνατότητα να μην γίνεται overfitting στα δεδομένα. Συγκεκριμένα, παρατηρείται και από την παραπάνω εικόνα ότι υπάρχει προειδοποίηση σε κάθε επανάληψη ότι ο ανιχνευτής υπερπροσαρμογής του μοντέλου είναι ενεργός και σταματάει τις επαναλήψεις όποτε χρειαστεί. Η ενεργοποίηση του ανιχνευτή γίνεται από την παράμετρο ‘od_type=Iter’ και ο αριθμός των

επαναλήψεων που θα γίνει μετά από την πιο πρόσφατη καλύτερη επίδοση γίνεται με την παράμετρο ‘od_wait=150’.

Το αποτέλεσμα του Catboost έχει ως εξής:

```
Full AUC score 0.765620
Test AUC score 0.765113
```

Αυτό δείχνει ότι και εδώ το αποτέλεσμα της πρόβλεψης μεταξύ του συνόλου test και του συνόλου train είναι πολύ κοντά και συνεπώς μπορεί να ειπωθεί ότι δεν γίνεται overfitting στα δεδομένα.

4.5 Τεχνητά Νευρωνικά Δίκτυα

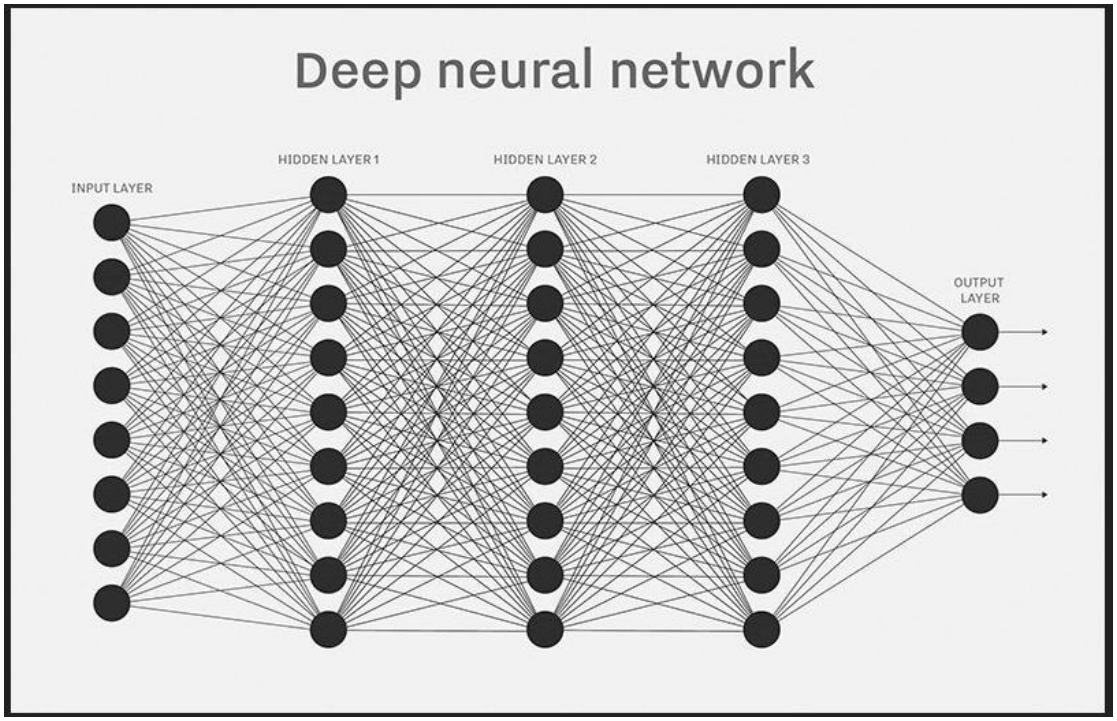
4.5.1 Εισαγωγή στα Τεχνητά Νευρωνικά Δίκτυα

Πώς λειτουργούν τα τεχνητά νευρωνικά δίκτυα

Ένα ANN (Artificial Neural Network) συνήθως περιλαμβάνει έναν μεγάλο αριθμό επεξεργαστών που λειτουργούν παράλληλα και είναι διατεταγμένοι σε επίπεδα. Η πρώτη βαθμίδα λαμβάνει τις πρώτες πληροφορίες εισόδου - ανάλογα με τα οπτικά νεύρα στην ανθρώπινη οπτική επεξεργασία. Κάθε διαδοχική βαθμίδα λαμβάνει την έξοδο από τη βαθμίδα που προηγείται - με τον ίδιο τρόπο οι νευρώνες από το οπτικό νεύρο λαμβάνουν σήματα από εκείνα που βρίσκονται πιο κοντά σε αυτό. Το τελευταίο επίπεδο παράγει την έξοδο - αποτέλεσμα του συστήματος.

Κάθε κόμβος επεξεργασίας έχει τη δική του μικρή σφαίρα γνώσης (κυρίως αναφέρεται σε μη γραμμικές συσχετίσεις), συμπεριλαμβανομένων όσων έχει δει και τυχόν κανόνες που είχε αρχικά προγραμματιστεί ή αναπτυχθεί για τον εαυτό του. Οι βαθμίδες είναι πολύ διασυνδεδεμένες, πράγμα που σημαίνει ότι κάθε κόμβος της βαθμίδας n θα συνδεθεί με πολλούς κόμβους της βαθμίδας n-1 - τις εισόδους του - και με τη βαθμίδα n + 1, στην οποία παρέχει δεδομένα εισόδου για τους εκεί κόμβους. Μπορεί να υπάρχει ένας ή περισσότεροι κόμβοι στο επίπεδο εξόδου, από το οποίο μπορεί να διαβαστεί η απάντηση που παράγει.

Τα τεχνητά νευρωνικά δίκτυα είναι αξιοσημείωτα για το γεγονός ότι είναι προσαρμοσμένα, πράγμα που σημαίνει ότι τροποποιούν τον εαυτό τους καθώς μαθαίνουν από την αρχική εκπαίδευση και τα επακόλουθα τρεξίματα παρέχουν περισσότερες πληροφορίες. Το πιο βασικό μαθησιακό μοντέλο επικεντρώνεται στη στάθμιση των ροών εισόδου, που είναι ο τρόπος με τον οποίο κάθε κόμβος ζυγίζει τη σημασία των δεδομένων εισόδου από κάθε έναν από τους προκατόχους του. Οι εισροές που συμβάλλουν στη λήψη σωστών απαντήσεων είναι υψηλότερα σταθμισμένες.



Εικόνα 27 – Αρχιτεκτονική Τεχνητών Νευρωνικών Δικτύων

Πώς εκπαιδεύονται τα νευρωνικά δίκτυα

Συνήθως, ένα ANN είναι αρχικά εκπαιδευμένο ή τροφοδοτείται με μεγάλα ποσά δεδομένων. Η εκπαίδευση συνίσταται στην παροχή εισόδου και στην ενημέρωση του δικτύου για το ποια θα είναι η έξοδος. Για παράδειγμα, για να οικοδομηθεί ένα δίκτυο που να προσδιορίζει τα πρόσωπα των ηθοποιών, η αρχική εκπαίδευση μπορεί να είναι μια σειρά από εικόνες με ηθοποιούς, με μη ηθοποιούς, με μάσκες, με πρόσωπα από αγάλματα και ζώα. Κάθε είσοδος συνοδεύεται από ταυτοποίηση ταυτότητας, όπως τα ονόματα των ηθοποιών και πληροφορίες "όχι ηθοποιός" ή "όχι άνθρωπος". Η παροχή των απαντήσεων επιτρέπει στο μοντέλο να προσαρμόζει τις εσωτερικές του βαρύτητες για να μάθει πώς να κάνει καλύτερη τη δουλειά του.

Για τον καθορισμό των κανόνων και τη λήψη αποφάσεων σχετικά με το τι να σταλεί στην επόμενη βαθμίδα με βάση τα δεδομένα εισόδου από τα προηγούμενα επίπεδα, τα νευρωνικά δίκτυα βασίζονται σε ποικίλες αρχές. Αυτές περιλαμβάνουν gradient-based εκπαίδευση, την fuzzy logic, τους γενετικούς αλγόριθμους και τις Bayesian μεθόδους [29], [30], [31], [32].

Για παράδειγμα, μπορεί να δοθεί εντολή στο σύστημα αναγνώρισης προσώπου, "Τα φρύδια βρίσκονται πάνω από τα μάτια" ή "Το μουστάκι είναι κάτω από τη μύτη ή "Το μουστάκι είναι πάνω και / ή δίπλα στο στόμα". Οι κανόνες που μπαίνουν στην αρχή μπορούν να κάνουν την εκπαίδευση πιο γρήγορη και να κάνουν το μοντέλο πιο δυναμικό νωρίτερα. Αλλά βασίζεται επίσης σε υποθέσεις σχετικά με τη φύση του προβληματικού χώρου, η οποία μπορεί να αποδειχθεί είτε άσχετη και άχρηστη είτε λανθασμένη και αντιπαραγωγική, κάνοντας τη λήψη αποφάσεων σχετικά με το ποιοι, εάν υπάρχουν, κανόνες πρέπει να οικοδομηθούν πολύ σημαντικούς.

Επιπλέον, οι υποθέσεις που κάνουν οι άνθρωποι όταν εκπαιδεύουν αλγορίθμους κάνουν τα νευρωνικά δίκτυα να ενισχύσουν τις προκαταλήψεις. Τα προκατειλημμένα σύνολα δεδομένων αποτελούν μια πρόκληση στα συστήματα εκπαίδευσης που βρίσκουν μόνα τους απαντήσεις αναγνωρίζοντας πρότυπα στα δεδομένα.

Τύποι νευρωνικών δικτύων

Τα νευρικά δίκτυα περιγράφονται μερικές φορές από την οπτική του βάθους τους, συμπεριλαμβανομένου του αριθμού των βαθμίδων που έχουν μεταξύ εισόδου και εξόδου ή των λεγόμενων κρυφών επιπέδων του μοντέλου. Αυτός είναι ο λόγος που ο όρος νευρωνικό δίκτυο χρησιμοποιείται σχεδόν συνώνυμα με τη βαθιά μάθηση. Μπορούν επίσης να περιγραφούν από τον αριθμό των κρυφών κόμβων του μοντέλου ή από τον αριθμό των εισόδων και εξόδων κάθε κόμβου. Οι παραλλαγές στο σχεδιασμό του κλασσικού νευρικού δικτύου επιτρέπουν διάφορες μορφές εμπρόσθιας και οπίσθιας διάδοσης πληροφοριών μεταξύ των επιπέδων.

Ειδικοί τύποι τεχνητών νευρωνικών δικτύων περιλαμβάνουν:

- Feed-forward neural networks (Νευρωνικά δίκτυα πρόσθιας τροφοδότησης)
- Recurrent neural networks (Αναδρομικά νευρωνικά δίκτυα)
- Convolutional neural networks (Συνελικτικά νευρωνικά δίκτυα)
- Deconvolutional neural networks (Αποσυνελικτικά νευρωνικά δίκτυα)
- Modular neural networks (Αρθρωτά νευρωνικά δίκτυα)

Πλεονεκτήματα τεχνητών νευρωνικών δικτύων

Τα πλεονεκτήματα των τεχνητών νευρωνικών δικτύων περιλαμβάνουν:

- Δυνατότητες παράλληλης επεξεργασίας που επιτρέπουν στο δίκτυο να μπορεί να εκτελεί περισσότερες από μία εργασίες ταυτόχρονα.
- Οι πληροφορίες αποθηκεύονται σε ολόκληρο το δίκτυο και όχι μόνο σε μια βάση δεδομένων.
- Ικανότητα να μαθαίνουν και να μοντελοποιούν μη γραμμικές, πολύπλοκες σχέσεις που βοηθάει στη μοντελοποίηση των σχέσεων πραγματικής ζωής μεταξύ εισόδων και εξόδων.
- Ανεκτικότητα σφάλματος που σημαίνει ότι το λάθος ενός ή περισσοτέρων κόμβων ANN δεν θα σταματήσει την παραγωγή του αποτελέσματος.
- Σταδιακή φθορά που σημαίνει ότι το δίκτυο θα υποβαθμιστεί αργά με την πάροδο του χρόνου, αντί να καταστραφεί αμέσως.
- Ικανότητα παραγωγής αποτελεσμάτων με ελλιπή γνώση, με την απώλεια απόδοσης να βασίζεται στο πόσο σημαντικές είναι οι ελλειπούσες πληροφορίες.
- Δεν υπάρχουν περιορισμοί στις μεταβλητές εισόδου, όπως είναι ο τρόπος με τον οποίο πρέπει να διανεμηθούν.
- Μηχανική μάθηση σημαίνει ότι το ANN μπορεί να μάθει από γεγονότα και να λαμβάνει αποφάσεις βάσει παρατηρήσεων.
- Ικανότητα μάθησης κρυφών σχέσεων στα δεδομένα που επιτρέπει σε ένα ANN να μοντελοποιήσει καλύτερα δεδομένα με υψηλή μεταβλητότητα και μη σταθερή διακύμανση.
- Δυνατότητα γενίκευσης και συμπερίληψης κρυφών σχέσεων σε κρυφά δεδομένα που δίνει τη δυνατότητα σε ένα ANN να μπορεί να προβλέψει την έξοδο των κρυφών δεδομένων.

Μειονεκτήματα των τεχνητών νευρωνικών δικτύων

Τα μειονεκτήματα των τεχνητών νευρωνικών δικτύων περιλαμβάνουν:

- Έλλειψη κανόνων για τον προσδιορισμό της σωστής δομής δικτύου που σημαίνει ότι η κατάλληλη αρχιτεκτονική του τεχνητού νευρωνικού δικτύου μπορεί να βρεθεί μόνο μέσω δοκιμών και σφαλμάτων (trial and error) και εμπειρίας.

- Απαίτηση για επεξεργαστές με παράλληλες δυνατότητες επεξεργασίας που καθιστούν το υλικό των νευρωνικών δικτύων εξαρτημένο.
- Το δίκτυο λειτουργεί με αριθμητικές πληροφορίες, γι' αυτό όλα τα προβλήματα πρέπει να μεταφραστούν σε αριθμητικές τιμές πριν μπορέσουν να παρουσιαστούν στο ANN.
- Έλλειψη εξηγήσεων πίσω από τις λύσεις που αποτελεί ένα από τα μεγαλύτερα μειονεκτήματα των ANN. Η αδυναμία να εξηγηθεί το γιατί ή το πώς πίσω από τη λύση, δημιουργεί έλλειψη εμπιστοσύνης στο δίκτυο.

4.5.2 Εφαρμογή Τεχνητών Νευρωνικών Δικτύων στα δεδομένα της εργασίας

Σε αντίθεση με τους προηγούμενους αλγορίθμους, το NN μοντέλο χρειάζεται να αναπτυχθεί. Ξεκινώντας, θα σχηματιστεί ο αριθμός των κόμβων σε κάθε επίπεδο όπου επιλέγονται στο πρώτο επίπεδο 120 κόμβοι, στο δεύτερο επίπεδο 80 κόμβοι, στο τρίτο επίπεδο 20 κόμβοι και στο τέλος είναι η έξοδος με ένα κόμβο. Το κάθε επίπεδο έχει συνάρτηση ενεργοποίησης ‘activation=relu’, ενώ στην έξοδο η συνάρτηση ενεργοποίησης είναι ‘activation=sigmoid’. Παράλληλα μετά από κάθε επίπεδο επιλέγεται ένα ποσοστό κόμβων που θα απορριφθεί από την εκπαίδευση. Αυτό γίνεται με την παράμετρο ‘Dropout’, όπου μετά το πρώτο και δεύτερο επίπεδο κόβεται το 20% και μετά το τρίτο επίπεδο κόβεται το 50%. Η απόρριψη (dropout[9]) είναι μια μέθοδος κανονικοποίησης για την αποφυγή της υπερπροσαρμογής, όπου κατά τη διάρκεια της εκπαίδευσης, κάποιοι τυχαίοι νευρώνες εξόδου του επιπέδου αγνοούνται ή "εγκαταλείπονται (dropout)" σε κάθε επανάληψη. Αυτό έχει ως αποτέλεσμα να κάνει το κάθε επίπεδο να μοιάζει και να αντιμετωπίζεται σαν επίπεδο με διαφορετικό αριθμό κόμβων και συνδεσιμότητα με το προηγούμενο επίπεδο. Στην πραγματικότητα, κάθε ενημέρωση σε ένα επίπεδο κατά τη διάρκεια της εκπαίδευσης γίνεται με μια διαφορετική "προβολή" του διαμορφωμένου επιπέδου. Όλα τα παραπάνω γίνονται με τις εξής εντολές:

```
#Neural Networks
start_time = time.time()
def build_nn():
    inputs = Input(shape=(x_train.shape[1],))
    x=BatchNormalization()(inputs)
    x = Dense(120, activation='relu')(x)
    x = Dropout(0.2)(x)
    x = Dense(80, activation='relu')(x)
    x = Dropout(0.2)(x)
    x = Dense(20, activation='relu')(x)
    x = Dropout(0.5)(x)
    out = Dense(1, activation='sigmoid')(x)
    model = Model(inputs=inputs, outputs=out)

    return model
```

Στη συνέχεια, όπως και με τα προηγούμενα μοντέλα, θα δημιουργηθούν δύο κενοί πίνακες, ένας για το σύνολο train και ένας για το σύνολο test, με αριθμό γραμμών όσο και το κάθε σύνολο αντίστοιχα για να συμπληρωθούν εκεί οι προβλέψεις του μοντέλου. Επιπλέον, χωρίζεται και εδώ το σύνολο του train σε πέντε folds και ύστερα υλοποιείται ο αλγόριθμος NN με τις εξής εντολές:

[9] https://en.wikipedia.org/wiki/Dropout_(neural_networks)

```
oof_preds3 = np.zeros(x_train.shape[0]) #dhmiourgoume pinaka me mhdenika oses einai oi eggrafes tou train
sub_preds3 = np.zeros(x_valid.shape[0])#dhmiourgoume pinaka me mhdenika oses einai oi eggrafes tou test

x_valid=x_valid.replace([-np.inf,np.inf],0)
folds = KFold(n_splits= 5, shuffle=True, random_state=5)

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]

    train_x=train_x.fillna(0)
    valid_x=valid_x.fillna(0)

    train_x=train_x.replace([-np.inf,np.inf],0)
    valid_x=valid_x.replace([-np.inf,np.inf],0)

    model=build_nn()
    model.compile(optimizer='nadam',
                  loss='binary_crossentropy')

    checkpoint1=ModelCheckpoint('model.h5', monitor='val_loss',
                                verbose=0, save_best_only=True, save_weights_only=True, mode='min', period=1)
    checkpoint2=EarlyStopping(monitor='val_loss', min_delta=0.000, patience=15,verbose=0, mode='min')

    os.environ['PYTHONHASHSEED'] = '0'
    np.random.seed(42)
    tf.random.set_seed(42)
    rn.seed(42)

    callbacks=[checkpoint1,checkpoint2]

    model.fit(train_x, train_y,callbacks=callbacks,batch_size=256,
            epochs=850,verbose=1, shuffle=True,validation_data=(valid_x,valid_y))

    model.load_weights('model.h5')

    oof_preds3[valid_idx] = model.predict(valid_x).ravel()
    sub_preds3 += model.predict(x_valid.fillna(0)).ravel() / folds.n_splits



    print('Fold %2d AUC : %.6f' % (n_fold + 1, roc_auc_score(valid_y, oof_preds3[valid_idx])))

    del model, train_x, train_y, valid_x, valid_y
```

Όταν εκτελείται ο αλγόριθμος δημιουργείται το παρακάτω (ένα μέρος):

```
Fold  1 AUC : 0.751514
Train on 55619 samples, validate on 13905 samples
Epoch 1/850
55619/55619 [==============================] - 1s 23us/step - loss: 0.3214 - val_loss: 0.2761
Epoch 2/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2902 - val_loss: 0.2669
Epoch 3/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2830 - val_loss: 0.2636
Epoch 4/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2762 - val_loss: 0.2643
Epoch 5/850
55619/55619 [==============================] - 1s 16us/step - loss: 0.2721 - val_loss: 0.2641
Epoch 6/850
55619/55619 [==============================] - 1s 15us/step - loss: 0.2691 - val_loss: 0.2636
Epoch 7/850
55619/55619 [==============================] - 1s 16us/step - loss: 0.2669 - val_loss: 0.2639
Epoch 8/850
55619/55619 [==============================] - 1s 16us/step - loss: 0.2626 - val_loss: 0.2644
Epoch 9/850
55619/55619 [==============================] - 1s 16us/step - loss: 0.2616 - val_loss: 0.2648
Epoch 10/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2605 - val_loss: 0.2637
Epoch 11/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2573 - val_loss: 0.2654
Epoch 12/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2561 - val_loss: 0.2652
Epoch 13/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2546 - val_loss: 0.2658
Epoch 14/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2532 - val_loss: 0.2658
Epoch 15/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2506 - val_loss: 0.2681
Epoch 16/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2498 - val_loss: 0.2667
Epoch 17/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2472 - val_loss: 0.2686
Epoch 18/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2469 - val_loss: 0.2700
Epoch 19/850
55619/55619 [==============================] - 1s 17us/step - loss: 0.2451 - val_loss: 0.2712
```

Εικόνα 28 - Μέρος αποτελέσματος υλοποίησης Neural Network

Το αποτέλεσμα του NN έχει ως εξής:

```
Full AUC score 0.748233
Test AUC score 0.750197
```

Όπως φαίνεται ανωτέρω, ο αλγόριθμος πέτυχε πρόβλεψη με AUC score = 0.750197. Συγκρίνοντάς το με την πρόβλεψη στο αρχικό 80% του πληθυσμού (AUC score = 0.748233), παρατηρείται ότι οι δύο τιμές είναι πολύ κοντά και επομένως το μοντέλο δεν έκανε overfitting, ενώ μπορεί να ειπωθεί ότι γενικεύει επιτυχώς σε νέα δεδομένα.

4.6 Συνδυασμός Αλγορίθμων με Λογιστική Παλινδρόμηση

Στη στατιστική, το λογιστικό μοντέλο χρησιμοποιείται για να μοντελοποιήσει την πιθανότητα μιας συγκεκριμένης κατηγορίας ή συμβάντος. Μπορεί να μοντελοποιήσει αρκετές κατηγορίες συμβάντων όπως καθορισμός αν μια εικόνα περιέχει μια γάτα, σκύλο, λιοντάρι κλπ. Κάθε αντικείμενο που ανιχνεύεται στην εικόνα θα έχει μια πιθανότητα μεταξύ 0 και 1.

Η λογιστική παλινδρόμηση[10] είναι ένα στατιστικό μοντέλο που στη βασική της μορφή χρησιμοποιεί μια λογιστική συνάρτηση για να μοντελοποιήσει μια δυαδική εξαρτημένη μεταβλητή, αν και υπάρχουν πολύ πιο πολύπλοκες επεκτάσεις. Στην ανάλυση παλινδρόμησης, η λογιστική παλινδρόμηση εκτιμά τις παραμέτρους ενός λογιστικού μοντέλου (μορφή δυαδικής παλινδρόμησης). Μαθηματικά, ένα δυαδικό λογιστικό μοντέλο έχει μια εξαρτημένη μεταβλητή με δύο πιθανές τιμές, όπως π.χ pass / fail που αντιπροσωπεύεται από μια μεταβλητή δείκτη, όπου οι δύο τιμές είναι ‘0’ και ‘1’. Η αντίστοιχη πιθανότητα της τιμής ‘1’ μπορεί να κυμαίνεται μεταξύ ‘0’ (βεβαίως η τιμή ‘0’) και ‘1’ (βεβαίως η τιμή ‘1’), εξ ου και η επισήμανση. Το καθοριστικό χαρακτηριστικό του λογιστικού μοντέλου είναι ότι η αύξηση μιας από τις ανεξάρτητες μεταβλητές πολλαπλασιάζει τις πιθανότητες του δεδομένου αποτελέσματος με σταθερό ρυθμό, με κάθε ανεξάρτητη μεταβλητή να έχει τη δική της παράμετρο, για μια δυαδική εξαρτημένη μεταβλητή αυτό γενικεύει τον λόγο πιθανότητας.

Σε ένα μοντέλο δυαδικής λογιστικής παλινδρόμησης, η εξαρτημένη μεταβλητή έχει δύο επίπεδα (κατηγορική). Οι έξοδοι με περισσότερες από δύο τιμές μοντελοποιούνται με multinomial λογιστική παλινδρόμηση. Το μοντέλο λογιστικής παλινδρόμησης απλώς μοντελοποιεί την πιθανότητα εξόδου ως προς την είσοδο και δεν εκτελεί στατιστική ταξινόμηση (δεν είναι ταξινομητής), αν και μπορεί να χρησιμοποιηθεί για να κάνει έναν ταξινομητή.

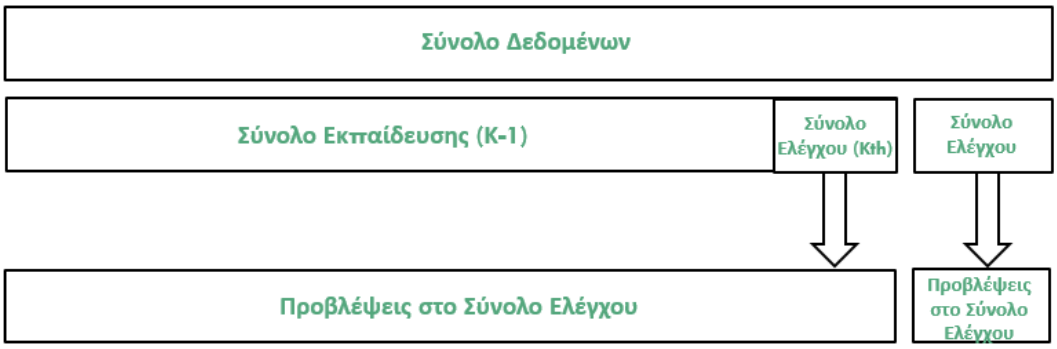
Αφού έχουν πραγματοποιηθεί και οι τρεις αλγόριθμοι στα δεδομένα, θα χρησιμοποιηθεί στη συνέχεια μια τεχνική συνδυασμού των αλγορίθμων με τη χρήση της λογιστικής παλινδρόμησης. Θα εφαρμοστεί η μέθοδος Stacking στα δεδομένα. Το Stacking ή αλλιώς Stacked Generalization έχει ως στόχο τη μελέτη πολλών διαφορετικών αλγορίθμων για το ίδιο πρόβλημα. Η ιδέα είναι ότι μπορεί να γίνει ανάλυση σε ένα μαθησιακό πρόβλημα με διαφορετικούς τύπους μοντέλων που είναι σε θέση να μάθουν κάποιο μέρος του προβλήματος, αλλά όχι ολόκληρο το πρόβλημα. Έτσι, μπορούν να δημιουργηθούν πολλαπλά διαφορετικά μοντέλα εκπαίδευσης και να χρησιμοποιηθούν για να δημιουργηθεί μια ενδιάμεση πρόβλεψη, μια πρόβλεψη για κάθε μοντέλο. Στη συνέχεια, προστίθεται ένα νέο μοντέλο που μαθαίνει από τις ενδιάμεσες προβλέψεις τον ίδιο στόχο.

[10] https://en.wikipedia.org/wiki/Logistic_regression

Αυτό το τελικό μοντέλο λέγεται ότι στοιβάζεται (stacked) στην κορυφή των άλλων, εξ ου και το όνομα. Αυτό, επιτρέπει να βελτιωθεί η συνολική απόδοση και τελικά να καταλήγουμε σε ένα μοντέλο που είναι καλύτερο από οποιοδήποτε μεμονωμένο ενδιάμεσο μοντέλο.

Εικόνα 29 - Μέθοδος Stacking

Για να γίνει καλύτερα κατανοητό πως δουλεύει το Stacking, παρακάτω αναλύεται η μεθοδολογία που ακολουθήθηκε. Αρχικά, χωρίζεται το σύνολο εκπαίδευσης σε 5 folds με τη K-fold cross-validation. Τα αρχικό μοντέλο εκπαιδεύεται στα τέσσερα folds (K-1) και γίνονται προβλέψεις για το πέμπτο (Kth). Αυτό επαναλαμβάνεται για κάθε fold του συνόλου εκπαίδευσης, όπως έχει αναλυθεί λεπτομερώς στην ενότητα 4.3.2. Ύστερα, το μοντέλο εκπαιδεύεται σε όλο το σύνολο εκπαίδευσης και υπολογίζεται η πιθανότητα στο σύνολο test. Αυτό επαναλαμβάνεται και για τα τρία μοντέλα (Light-GBM, Catboost, Neural Networks), που αποτελούν τα μοντέλα πρώτου επιπέδου. Στη συνέχεια, οι σχηματισμένες προβλέψεις από το σύνολο εκπαίδευσης χρησιμοποιούνται ως μεταβλητές για το μοντέλο δεύτερου επιπέδου, την λογιστική παλινδρόμηση. Τέλος, η λογιστική παλινδρόμηση κάνει πρόβλεψη για το σύνολο test.



Εικόνα 30 - Υλοποίηση Stacking

Θα ακολουθήσει η εφαρμογή των παραπάνω. Αρχικά, δημιουργούνται δύο πίνακες με όλες τις προβλέψεις των προηγούμενων τριών εφαρμοσμένων αλγορίθμων. Στη συνέχεια, όπως και στα προηγούμενα μοντέλα θα δημιουργηθούν δύο κενοί πίνακες, ένας για το σύνολο train και ένας για το σύνολο test, με αριθμό γραμμών όσο και το κάθε σύνολο αντίστοιχα για να συμπληρωθούν εκεί οι προβλέψεις του μοντέλου. Επιπλέον, χωρίζεται και εδώ το σύνολο του train σε πέντε folds. Έπειτα, γίνεται μετασχηματισμός των χαρακτηριστικών χρησιμοποιώντας τη μέθοδο QuantileTransformer. Αυτή η μέθοδος μετατρέπει τα χαρακτηριστικά σε μια ομοιόμορφη ή κανονική κατανομή, εν προκειμένω σε κανονική κατανομή. Επομένως, για ένα δοσμένο χαρακτηριστικό, αυτός ο μετασχηματισμός τείνει να

κατανείμει τις πιο συχνές τιμές και μειώνει επίσης τον αντίκτυπο των ακραίων τιμών. Ο μετασχηματισμός εφαρμόζεται ανεξάρτητα σε κάθε χαρακτηριστικό. Επιπλέον, με τη χρήση της μεθόδου PolynomialFeatures, θα δημιουργηθούν πολυώνυμα. Με τη μέθοδο αυτή δημιουργείται μια νέα μήτρα χαρακτηριστικών που αποτελείται από όλους τους πολυωνυμικούς συνδυασμούς των χαρακτηριστικών με βαθμό μικρότερο ή ίσο από τον καθορισμένο βαθμό. Για παράδειγμα, αν ένα δείγμα εισόδου είναι δύο διαστάσεων και της μορφής [a, b], τα χαρακτηριστικά πολυώνυμα βαθμού 2 είναι [1, a, b, a ^ 2, ab, b ^ 2]. Ουσιαστικά από τις πιθανότητες κάθε μοντέλου υπολογίζονται τα γινόμενά τους, έτσι ώστε να δημιουργηθεί μια νέα πιθανότητα. Τέλος, εκπαιδεύονται εκ νέου τα δεδομένα με τη βοήθεια της λογιστικής παλινδρόμησης.

Το αποτέλεσμα της λογιστικής παλινδρόμησης είναι το παρακάτω:

```
Fold  1 AUC : 0.773737
Fold  2 AUC : 0.774360
Fold  3 AUC : 0.764545
Fold  4 AUC : 0.762358
Fold  5 AUC : 0.768265
--- 3.884338140487671 seconds ---
Full AUC score 0.768573
Test AUC score 0.766156
```

Κεφάλαιο 5: Αποτελέσματα, Συμπεράσματα και Μελλοντική Εργασία

5.1 Αποτελέσματα

Έχοντας πλέον υλοποιήσει και τα τέσσερα μοντέλα στα σύνολα δεδομένων, θα γίνει μια σύγκριση των αποτελεσμάτων τους. Παρατηρείται ότι ο αλγόριθμος με το καλύτερο αποτέλεσμα πρόβλεψης της μεταβλητής “TARGET” είναι η τεχνική που εφαρμόστηκε με τη λογιστική παλινδρόμηση με σκορ 0.766156. Δεύτερος έρχεται ο αλγόριθμος LightGBM με σκορ 0.765927, τρίτος ο Catboost με σκορ 0.765113 και τελευταία τα τεχνητά νευρωνικά δίκτυα με σκορ 0.750197. Ακόμα και από την μεριά του χρόνου εκτέλεσης, η λογιστική παλινδρόμηση έχει τον πιο γρήγορο χρόνο, μόλις 4 δευτερόλεπτα. Δεύτερο πιο γρήγορο χρόνο έκανε ο LightGBM, καθώς χρειάστηκαν 94 δευτερόλεπτα, τρίτο καλύτερο χρόνο έκαναν τα νευρωνικά δίκτυα με 121 δευτερόλεπτα και πιο αργός ήταν ο Catboost με 298 δευτερόλεπτα.

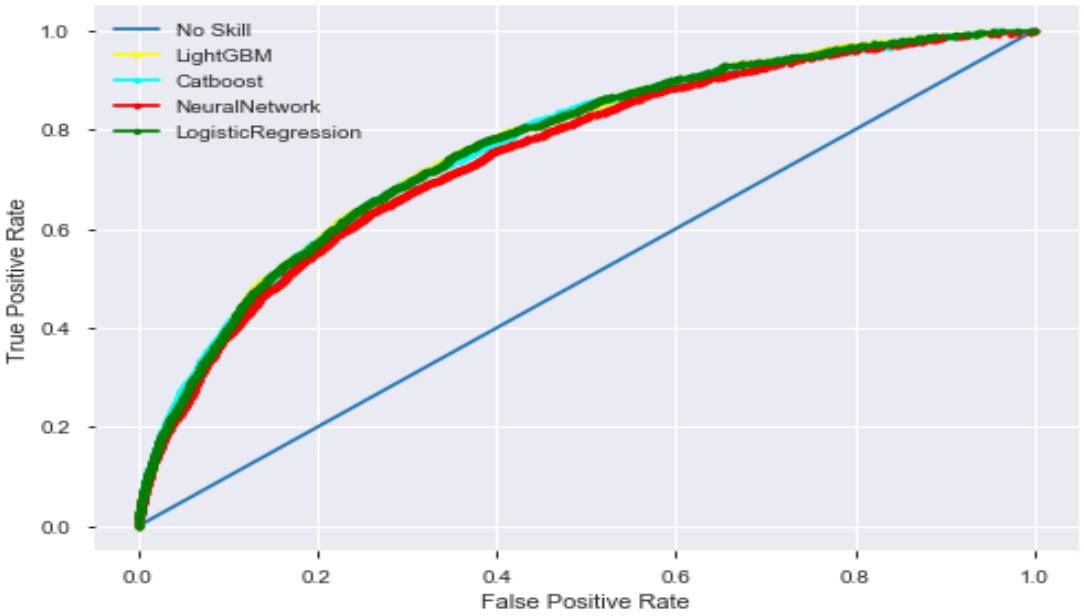
Στη παρακάτω εικόνα, εμφανίζονται τα AUC σκορ καθώς επίσης και οι χρόνοι εκτέλεσης του κάθε αλγορίθμου.

	auc score	execution time
LightGBM	Test AUC score 0.765927	--- 94.3875982761383 seconds ---
Catboost	Test AUC score 0.765113	--- 298.0066936016083 seconds ---
Neural_Networks	Test AUC score 0.750197	--- 121.11705255508423 seconds ---
Logistic_Regression	Test AUC score 0.766156	--- 3.884338140487671 seconds ---

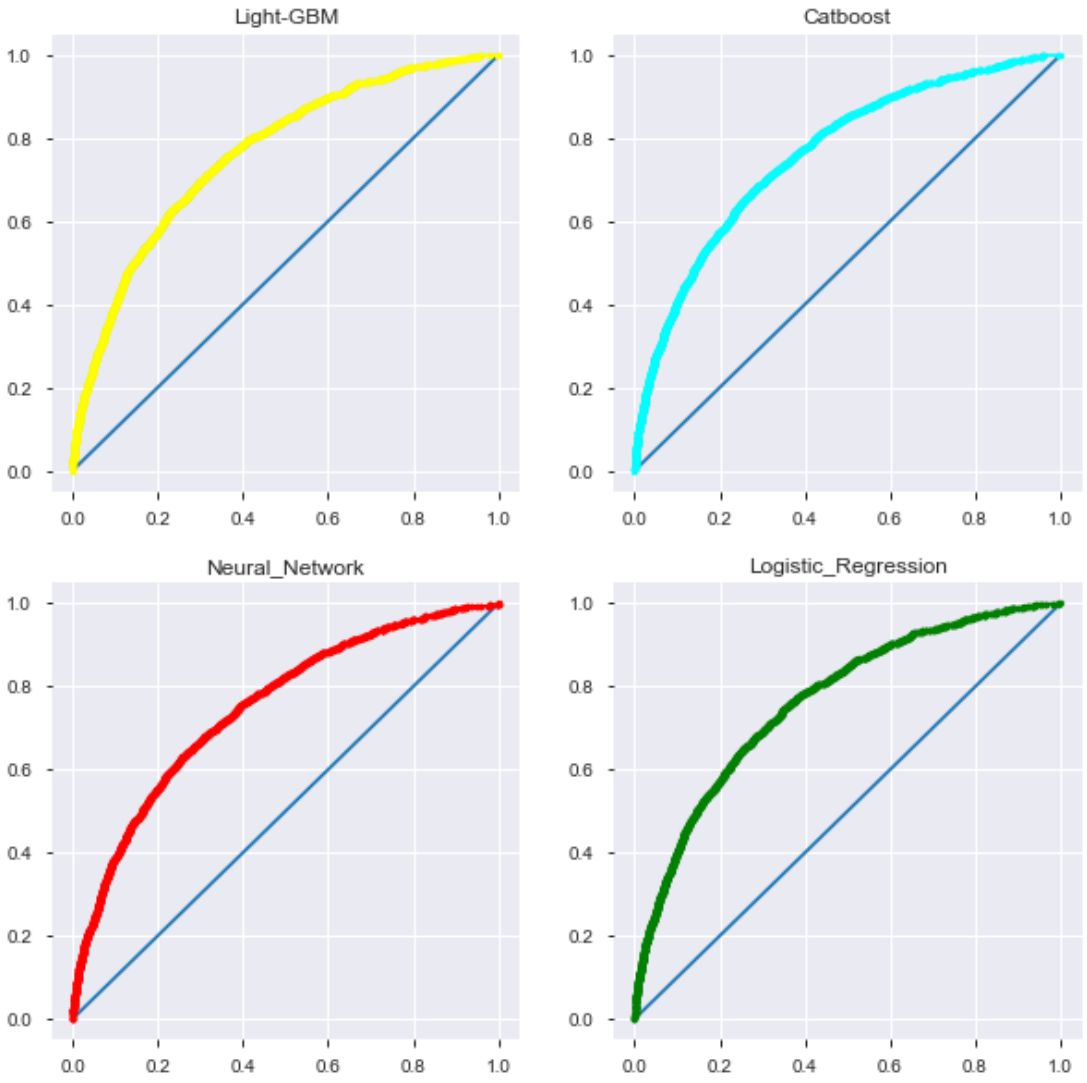
Εικόνα 31 - Σύγκριση και Αποτελέσματα Αλγορίθμων

Συμπεραίνεται λοιπόν, ότι η τεχνική που εφαρμόστηκε με τη λογιστική παλινδρόμηση κάνει την καλύτερη πρόβλεψη όσον αφορά ποιος πιστούχος θα αποπληρώσει τη πιστωτική του κάρτα ή όχι. Η τεχνική αυτή απέδωσε τα καλύτερα αποτελέσματα γιατί συνδυάζει τις πιθανότητες από τρία άλλα μοντέλα, με αποτέλεσμα να μειώνεται η μεροληψία (bias) στα δεδομένα και η διακύμανση (variance) των προβλέψεων, σχηματίζοντας ένα μοντέλο που είναι περισσότερο ανεκτικό στις μεταβολές των πιθανοκατανομών των μεταβλητών που χρησιμοποιούν οι αλγόριθμοι LightGBM, Catboost και Neural Networks. Συνεπώς, οι νέες αυτές προβλέψεις είναι και καλύτερες και πιο αποτελεσματικές.

Ακολουθούν διαγράμματα που απεικονίζουν τις καμπύλες ROC των μοντέλων:



Εικόνα 32 - Καμπύλες ROC αλγορίθμων (1)



Εικόνα 33 - Καμπύλες ROC αλγορίθμων (2)

5.2 Συμπεράσματα

Η παρούσα εργασία πραγματεύεται την επιστήμη των μεγάλων δεδομένων, της μηχανικής μάθησης και της εξόρυξης γνώσης στον τραπεζικό τομέα. Πραγματοποιείται μια εισαγωγή στις βασικές έννοιες, στους στόχους, τα στάδια, και τις κατηγορίες των επιστημών αυτών, καθώς και τα πλεονεκτήματα και τα μειονεκτήματα που απορρέουν από τις συγκεκριμένες επιστήμες. Σημαντική είναι επίσης και η αναφορά στους κινδύνους που αντιμετωπίζει ένα χρηματοπιστωτικό ίδρυμα και ειδικότερα στη διαχείριση του πιστωτικού κινδύνου των πιστωτικών καρτών.

Σκοπός αυτής της διπλωματικής εργασίας είναι να καθοδηγήσει τον αναγνώστη μέσω ενός προβλήματος εξόρυξης δεδομένων στη πρόβλεψη της ικανότητας των πιστούχων να αποπληρώσουν ένα τραπεζικό δάνειο, εν προκειμένω την πιστωτική τους κάρτα. Η καθοδήγηση του αναγνώστη γίνεται επεξηγώντας με σαφήνεια όλα τα βήματα της διαδικασίας. Ξεκινώντας από τη θεωρητική προσέγγιση, καταλήγει στο πρακτικό κομμάτι, όπου γίνεται η εφαρμογή αλγορίθμων μηχανικής μάθησης, καλύπτοντας έτσι όλο το φάσμα του προβλήματος.

Η εξόρυξη δεδομένων μπορεί να είναι χρήσιμη για οποιοδήποτε χρηματοπιστωτικό ίδρυμα. Οι τεχνικές εξόρυξης δεδομένων μπορούν να βοηθήσουν τις τράπεζες για καλύτερη στόχευση και απόκτηση νέων πελατών, ανίχνευση απάτης σε πραγματικό χρόνο, παροχή εξατομικευμένων προϊόντων και ανάλυση των προτύπων αγοράς των πελατών με την πάροδο του χρόνου. Εάν ένα χρηματοπιστωτικό ίδρυμα μπορεί να κάνει ακριβείς προβλέψεις των δυνατοτήτων των αιτούντων δανείων, τότε μπορεί να ενεργήσει βάσει των πληροφοριών αυτών. Η όλη διαδικασία της αίτησης γίνεται πιο αποτελεσματική καθώς δεν σπαταλάται χρόνος σε αμφίβολους και κακούς αιτούντες. Επιπλέον, τόσο ο χρηματοπιστωτικός οργανισμός όσο και ο αιτών λαμβάνουν μικρότερο κίνδυνο. Αυτή η μείωση του κινδύνου είναι σημαντικό γεγονός αν ληφθούν υπόψη τα αίτια και οι συνέπειες της οικονομική κρίσης του 2008.

Συνοψίζοντας, η χρήση των προαναφερθέντων τεχνολογιών είναι απολύτως αναγκαία ώστε οι τράπεζες να διατηρήσουν αλλά και να ενισχύσουν την ανταγωνιστικότητά τους στην αγορά. Απαραίτητη προϋπόθεση για να επιτευχθεί ο εν λόγω στόχος είναι ο μετασχηματισμός τόσο των δομών και διαδικασιών τους όσο και του προσωπικού τους, ώστε το τελευταίο να εκπαιδευτεί ανάλογα στις νέες τεχνολογίες και να εντρυφήσει στη τεχνολογική σκέψη και καινοτομία. Με αυτό τον τρόπο θα καταφέρουν να εξασφαλίσουν την βέλτιστη αλληλεπίδραση ανθρώπινου δυναμικού και μηχανής, αλλά και να ανταποκριθούν στη νέα γενιά πελατών.

5.3 Μελλοντική Εργασία

Η διπλωματική αυτή εργασία, όπως έχει ήδη αναφερθεί νωρίτερα, περιέχει αναφορές σε πλήθος εννοιών σχετικών με την εξέλιξη της τεχνολογίας των υπολογιστών, την σύνδεση αυτής με τον τραπεζικό κλάδο αλλά και την εφαρμογή αλγορίθμων μηχανικής μάθησης. Ως εκ τούτου, μπορεί να αποτελέσει τον κορμό για ποικίλες μελλοντικές αναλύσεις προς διάφορες κατευθύνσεις άλλες θεωρητικού και άλλες περισσότερο πρακτικού περιεχομένου.

Στο θεωρητικό μέρος της υφιστάμενης εργασίας έχουν πραγματοποιηθεί αναφορές σε σημαντικές περιοχές της επιστήμης των υπολογιστών αλλά και της χρηματοοικονομικής επιστήμης στα πλαίσια των χρηματοπιστωτικών ιδρυμάτων, παραθέτοντας εισαγωγικές έννοιες και βασικές αρχές των περιοχών αυτών. Αναλύθηκε εκτενέστερα ο τομέας της εξόρυξης γνώσης αλλά και του πιστωτικού κινδύνου τραπεζών κυριότερα στις πιστωτικές κάρτες. Συνεπεία των ανωτέρω, γίνεται αντιληπτό ότι πολλές περιοχές ενδιαφέροντος, όπως η Τεχνητή Νοημοσύνη, οι λύσεις Επιχειρηματικής Ευφυΐας, η Μηχανική Μάθηση κλπ ,

χρίζουν μεγαλύτερης και βαθύτερης ενασχόλησης και ανάλυσης, έως ότου θεωρηθεί ότι έχουν παρουσιαστεί επαρκώς στον αναγνώστη συμπυκνωμένες μεν, ολοκληρωμένες δε οι γνώσεις γύρω από τα αντίστοιχα ζητήματα. Οι περιοχές αυτές αναπτύχθηκαν ως ένα βαθμό ώστε να παρέχουν βασικές εισαγωγικές γνώσεις στον αναγνώστη, αποφεύγοντας τις μακροσκελείς αναλύσεις, ώστε το νόημα και οι έννοιες να είναι εύκολα κατανοητές και όχι περίπλοκες ή σύνθετες, δεδομένου ότι κάθε ένα από τα θέματα που αναφέρονται σε αυτή την εργασία θα μπορούσαν να αποτελέσουν και θέματα αυτόνομων διπλωματικών εργασιών.

Από την πρακτική μεριά της εργασίας, δηλαδή στο κομμάτι εφαρμογής αλγορίθμων μηχανικής μάθησης για την πρόβλεψη της πιθανότητας αθέτησης, μπορούν επίσης να τεθούν στόχοι για μελλοντική επιπρόσθετη έρευνα. Ένας από αυτούς μπορεί να είναι σαφώς η υλοποίηση περισσότερων αλγορίθμων που θα επέτρεπε τη σύγκριση μεμονωμένα περισσότερων αποτελεσμάτων και θα μπορούσε να προσφέρει τυχόν καλύτερες επιδόσεις στην πρόβλεψη της αθέτησης. Συνάμα, η μέθοδος Stacking που ακολουθήθηκε στην υφιστάμενη ανάλυση, θα μπορούσε να πραγματοποιηθεί και με επιπλέον μοντέλα όπως XGBoost, KNN, SVM και Random Forest. Επιπλέον, ένας ακόμη στόχος, θα μπορούσε να είναι η ανάπτυξη ακόμα περισσότερων μεταβλητών από τις ήδη υπάρχουσες, γεγονός που θα βελτίωνε τις αποδόσεις των αλγορίθμων, μιας και οι συσχετίσεις των μεταβλητών θα ήταν μεγαλύτερες. Τέλος, συναρτήσει του χρόνου που πραγματοποιήθηκε η αυτή εργασία, δηλαδή λίγους μήνες (1.1.2021) πριν την εφαρμογή του νέου ορισμού αθέτησης (EBA – New Definition of Default) όπως απαιτείται από τις κανονιστικές αρχές, κρίνεται ίσως εξαιρετικά σκόπιμο στο μέλλον να επαναληφθεί η άσκηση με προσαρμογή των αλγορίθμων στις προδιαγραφές και κανόνες που ορίζονται στο νέο πλαίσιο. Κατόπιν αυτού θα μπορούν να συγκριθούν τα αποτελέσματα και να ποσοτικοποιηθεί η επίδραση του νέου ορισμού στο μοντέλο υπολογισμού της πιθανότητας αθέτησης πιστούχων πιστωτικών καρτών.

Βιβλιογραφία

[1] B.P. Rao, P. Saluia, N. Sharma, A. Mittal, S.V. Sharma. Cloud computing for Internet of Things & sensing based applications. (2012).
https://www.semanticscholar.org/paper/Cloud-computing-for-Internet-of-Things-%26-sensing-Rao-Saluia/cfb0b88089a3b6b1307977fe069f2288b0012149

[2] M. Sukanya, S. Biruntha. Techniques on Text Mining. (2012).
https://www.researchgate.net/publication/261317364_Techniques_on_text_mining

[3] G. Bathla, R. Rani, H. Aggarwal. Comparative study of NoSQL databases for big data storage. (2018).
https://www.researchgate.net/publication/323726150_Comparative_study_of_NoSQL_databases_for_big_data_storage

[4] M. Sharma, V. Chauhan, K. Kishore. A review: Mapreduce and Spark for Big Data Analytics. (2016).
https://www.researchgate.net/publication/325527960_A_REVIEW_MAPREDUCE_AND_SPARK_FOR_BIG_DATA_ANALYTICS

[5] F. Xhafa, V. Naranjo, S Caballe. Processing and Analytics of Big Data Streams with Yahoo!S4. (2015).
https://www.researchgate.net/publication/277139578_Processing_and_Analytics_of_Big_Data_Streams_with_YahooS4

[6] J. F. Elder IV, D. Pregibon. A Statistical Perspective on KDD. (1995).

[7] B. Namratha, N. Sharma. Educational Data Mining – Applications and Techniques. (2016).
https://www.ijltet.org/journal/147021291075.pdf

[8] M. Shahbaz, M. Rahman. Data mining for engineering sector in pakistan: Issues and implications. (2008).
https://pdfs.semanticscholar.org/7aca/916400272c9f84bb92b04a44aa345ffb2f6f.pdf

[9] M. Durairaj, V. Ranjani. Data Mining Applications In Healthcare Sector: A Study. (2013).
https://www.ijstr.org/final-print/oct2013/Data-Mining-Applications-In-Healthcare-Sector-A-Study.pdf

[10] R. D.Canlas Jr. . Data mining in Healthcare: Current applications and issues. (2009).
http://mines.humanoriented.com/classes/2010/fall/csci568/papers/Data_Mining_Health.pdf

[11] M. P. Bach. Data Mining Applications in Public Organizations. (2003).
https://www.researchgate.net/publication/4031742_Data_mining_applications_in_public_organizations

[12] C. H. Caldas, L. Soibelman. Automating hierarchical document classification for construction management information systems. (2003).
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.1015&rep=rep1&type=pdf

[13] O. Wibisono, H. D. Ari, A. Widjanarti, A. A. Zulen, B. Tissot. The use of big data analytics and artificial intelligence in central banking. (no date).
https://www.bis.org/ifc/publ/ifcb50_01.pdf

[14] S. Kappagantula, A. Kulkarni. Social Media Analytics – Personalize Product and Service Offerings
https://www.infosys.com/industries/financial-services/white-papers/Documents/social-media-analytics.pdf

[15] B. C. Amanze, C. G. Onukwugha. Data Mining Application in Credit Card Fraud Detection System. (2018).
https://www.researchgate.net/publication/328478013_Data_Mining_Application_in_Credit_Card_Fraud_Detection_System

[16] I. Met, G. Tunali, A. Erkoc, S. Tanrikulu, M. O. Dolgun. Branch Efficiency and Location Forecasting: Application of Ziraat Bank. (2017).
https://www.researchgate.net/publication/316846895_Branch_Efficiency_and_Location_Forecasting_Application_of_Ziraat_Bank

[17] S. M. H. Hasheminejad, Z. Reisjafari. ATM management prediction using Artificial Intelligence techniques: A survey. (2017).
https://www.researchgate.net/publication/318666269_ATM_management_prediction_using_Artificial_Intelligence_techniques_A_survey

[18] B. R. Sharmaa, D. Kaura, Manjub. A Review on Data Mining: Its Challenges, Issues and Applications. (2013).
https://inpressco.com/wp-content/uploads/2013/06/Paper85695-700.pdf

[19] Citi GPS (2018). Bank of the Future: The ABCs of digital disruption in Finance
http://www.smallake.kr/wp-content/uploads/2018/05/AHDX6.pdf

[20] S. Heffernan. Modern Banking. (2005).
https://www.academia.edu/35734316/Heffernan_-_Modern_Banking

[21] X. Freixas & J.C. Rochet. Microeconomics of Banking. (2008).
https://www.academia.edu/25443293/The_microeconomics_of_banking_-_Xavier_Freiax

[22] M. Bellis. Invention of Credit Cards. (no date).
http://inventors.about.com/od/cstartinventions/a/credit_cards.htm

[23] J. Z. Zywicki (2000). The economics of credit cards. (2000). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=229356

[24] M. Bailey. Practical Credit Scoring: Issues and Techniques. (2006).

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. (2017). https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[26] J. H. Friedman. Greedy function approximation: a gradient boosting machine. (1999). https://statweb.stanford.edu/~jhf/ftp/trebst.pdf

[27] L. Mason, J. Baxter, P. Bartlett, M. Frean. Boosting Algorithms as Gradient Descent. (1999). https://papers.nips.cc/paper/1766-boosting-algorithms-as-gradient-descent.pdf

[28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin. CatBoost: unbiased boosting with categorical features. (2018). https://arxiv.org/abs/1706.09516

[29] G. Wang, L. Huang, C. Zhang. Study of artificial neural network model based on fuzzy clustering. (2006). https://ieeexplore.ieee.org/abstract/document/1712857

[30] G. F. Miller, P. M. Todd, S. U. Hedge. Designing Neural Networks using Genetic Algorithms. (1989). https://www.researchgate.net/publication/220885651_Designing_Neural_Networks_using_Genetic_Algorithms

[31] V. Mullachery, A. Khera, A. Husain. Bayesian Neural Networks. (2018). https://arxiv.org/ftp/arxiv/papers/1801/1801.07710.pdf

[32] Y. Bengio. Practical Recommendations for Gradient-Based Training of Deep Architectures. (2012). https://link.springer.com/chapter/10.1007/978-3-642-35289-8_26

Παράρτημα

Μεταβλητές συνόλου δεδομένων “application_train”

Row	Description
SK_ID_CURR	ID of loan in our sample
TARGET	Target variable (1 - client with payment difficulties, 0 - all other cases)
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_INCOME_TOTAL	Income of the client
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_REGISTRATION	How many days before the application did client change his registration
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
OWN_CAR_AGE	Age of client's car
FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL	Did client provide email (1=YES, 0=NO)
OCCUPATION_TYPE	What kind of occupation does the client have
CNT_FAM_MEMBERS	How many family members does client have
REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
ORGANIZATION_TYPE	Type of organization where client works
EXT_SOURCE_1	Normalized score from external data source

EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
APARTMENTS_AVG	Normalized information about building where the client lives
BASEMENTAREA_AVG	Normalized information about building where the client lives
YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives
YEARS_BUILD_AVG	Normalized information about building where the client lives
COMMONAREA_AVG	Normalized information about building where the client lives
ELEVATORS_AVG	Normalized information about building where the client lives
ENTRANCES_AVG	Normalized information about building where the client lives
FLOORSMAX_AVG	Normalized information about building where the client lives
FLOORSMIN_AVG	Normalized information about building where the client lives
LANDAREA_AVG	Normalized information about building where the client lives
LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives
LIVINGAREA_AVG	Normalized information about building where the client lives
NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives
NONLIVINGAREA_AVG	Normalized information about building where the client lives
APARTMENTS_MODE	Normalized information about building where the client lives
BASEMENTAREA_MODE	Normalized information about building where the client lives
YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives
YEARS_BUILD_MODE	Normalized information about building where the client lives
COMMONAREA_MODE	Normalized information about building where the client lives
ELEVATORS_MODE	Normalized information about building where the client lives
ENTRANCES_MODE	Normalized information about building where the client lives
FLOORSMAX_MODE	Normalized information about building where the client lives
FLOORSMIN_MODE	Normalized information about building where the client lives
LANDAREA_MODE	Normalized information about building where the client lives
LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives
LIVINGAREA_MODE	Normalized information about building where the client lives
NONLIVINGAPARTMENTS_MODE	Normalized information about building where the client lives
NONLIVINGAREA_MODE	Normalized information about building where the client lives
APARTMENTS_MEDI	Normalized information about building where the client lives
BASEMENTAREA_MEDI	Normalized information about building where the client lives
YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives
YEARS_BUILD_MEDI	Normalized information about building where the client lives
COMMONAREA_MEDI	Normalized information about building where the client lives
ELEVATORS_MEDI	Normalized information about building where the client lives
ENTRANCES_MEDI	Normalized information about building where the client lives
FLOORSMAX_MEDI	Normalized information about building where the client lives
FLOORSMIN_MEDI	Normalized information about building where the client lives
LANDAREA_MEDI	Normalized information about building where the client lives
LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives
LIVINGAREA_MEDI	Normalized information about building where the client lives
NONLIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives
NONLIVINGAREA_MEDI	Normalized information about building where the client lives
FONDKAPREMONT_MODE	Normalized information about building where the client lives
HOUSETYPE_MODE	Normalized information about building where the client lives
TOTALAREA_MODE	Normalized information about building where the client lives

WALLSMATERIAL_MODE	Normalized information about building where the client lives
EMERGENCYSTATE_MODE	Normalized information about building where the client lives
OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
FLAG_DOCUMENT_2	Did client provide document 2
FLAG_DOCUMENT_3	Did client provide document 3
FLAG_DOCUMENT_4	Did client provide document 4
FLAG_DOCUMENT_5	Did client provide document 5
FLAG_DOCUMENT_6	Did client provide document 6
FLAG_DOCUMENT_7	Did client provide document 7
FLAG_DOCUMENT_8	Did client provide document 8
FLAG_DOCUMENT_9	Did client provide document 9
FLAG_DOCUMENT_10	Did client provide document 10
FLAG_DOCUMENT_11	Did client provide document 11
FLAG_DOCUMENT_12	Did client provide document 12
FLAG_DOCUMENT_13	Did client provide document 13
FLAG_DOCUMENT_14	Did client provide document 14
FLAG_DOCUMENT_15	Did client provide document 15
FLAG_DOCUMENT_16	Did client provide document 16
FLAG_DOCUMENT_17	Did client provide document 17
FLAG_DOCUMENT_18	Did client provide document 18
FLAG_DOCUMENT_19	Did client provide document 19
FLAG_DOCUMENT_20	Did client provide document 20
FLAG_DOCUMENT_21	Did client provide document 21
AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

Μεταβλητές συνόλου δεδομένων " credit_card_balance"

Row	Description
SK_ID_PREV	ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit)
SK_ID_CURR	ID of loan in our sample
MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
AMT_BALANCE	Balance during the month of previous credit
AMT_CREDIT_LIMIT_ACTUAL	Credit card limit during the month of the previous credit
AMT_DRAWINGS_ATM_CURRENT	Amount drawing at ATM during the month of the previous credit
AMT_DRAWINGS_CURRENT	Amount drawing during the month of the previous credit
AMT_DRAWINGS_OTHER_CURRENT	Amount of other drawings during the month of the previous credit
AMT_DRAWINGS_POS_CURRENT	Amount drawing or buying goods during the month of the previous credit
AMT_INST_MIN_REGULARITY	Minimal installment for this month of the previous credit
AMT_PAYMENT_CURRENT	How much did the client pay during the month on the previous credit
AMT_PAYMENT_TOTAL_CURRENT	How much did the client pay during the month in total on the previous credit
AMT_RECEIVABLE_PRINCIPAL	Amount receivable for principal on the previous credit
AMT_RECIVABLE	Amount receivable on the previous credit
AMT_TOTAL_RECEIVABLE	Total amount receivable on the previous credit
CNT_DRAWINGS_ATM_CURRENT	Number of drawings at ATM during this month on the previous credit
CNT_DRAWINGS_CURRENT	Number of drawings during this month on the previous credit
CNT_DRAWINGS_OTHER_CURRENT	Number of other drawings during this month on the previous credit
CNT_DRAWINGS_POS_CURRENT	Number of drawings for goods during this month on the previous credit
CNT_INSTALMENT_MATURE_CUM	Number of paid installments on the previous credit
NAME_CONTRACT_STATUS	Contract status (active signed,...) on the previous credit
SK_DPD	DPD (Days past due) during the month on the previous credit
SK_DPD_DEF	DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

Μεταβλητές συνόλου δεδομένων " bureau"

Row	Description
SK_ID_CURR	ID of loan in our sample - one loan in our sample can have 0,1,2 or more related previous credits in credit bureau
SK_BUREAU_ID	Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application)
CREDIT_ACTIVE	Status of the Credit Bureau (CB) reported credits
CREDIT_CURRENCY	Recoded currency of the Credit Bureau credit
DAYS_CREDIT	How many days before current application did client apply for Credit Bureau credit
CREDIT_DAY_OVERDUE	Number of days past due on CB credit at the time of application for related loan in our sample
DAYS_CREDIT_ENDDATE	Remaining duration of CB credit (in days) at the time of application in Home Credit
DAYS_ENDDATE_FACT	Days since CB credit ended at the time of application in Home Credit (only for closed credit)
AMT_CREDIT_MAX_OVERDUE	Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)
CNT_CREDIT_PROLONG	How many times was the Credit Bureau credit prolonged
AMT_CREDIT_SUM	Current credit amount for the Credit Bureau credit
AMT_CREDIT_SUM_DEBT	Current debt on Credit Bureau credit
AMT_CREDIT_SUM_LIMIT	Current credit limit of credit card reported in Credit Bureau
AMT_CREDIT_SUM_OVERDUE	Current amount overdue on Credit Bureau credit
CREDIT_TYPE	Type of Credit Bureau credit (Car, cash,...)
DAYS_CREDIT_UPDATE	How many days before loan application did last information about the Credit Bureau credit come
AMT_ANNUITY	Annuity of the Credit Bureau credit

Μεταβλητές συνόλου δεδομένων " bureau_balance"

Row	Description
SK_BUREAU_ID	Recoded ID of Credit Bureau credit (unique coding for each application) - use this to join to CREDIT_BUREAU table
MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
STATUS	Status of Credit Bureau loan during the month (active, closed, DPD0-30,… [C means closed, X means status unknown, 0 means no DPD, 1 means maximal did during month between 1-30, 2 means DPD 31-60,… 5 means DPD 120+ or sold or written off])