ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Πιστωτικού Κινδύνου μέσω Μεθόδου Μηχανικής Μάθησης

Ευφροσύνη Π. Μπαμπούρη

Επιβλέπων Καθηγητής:
Μαγκλογιάννης Ηλίας, Καθηγητής

ΠΕΙΡΑΙΑΣ

Φεβρουάριος 2023

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Πιστωτικού Κινδύνου μέσω Μεθόδου Μηχανικής Μάθησης

Ευφροσύνη Μπαμπούρη
Α.Μ.: ΜΕ2027

ΠΕΡΙΛΗΨΗ

Η ταχεία εξέλιξη της τεχνολογίας, το ανταγωνιστικό περιβάλλον και ο τεράστιος όγκος δεδομένων που είναι διαθέσιμα στην σημερινή εποχή, λαμβάνουν υπόψη την επείγουσα ανάγκη των εταιρειών να μεταβούν σε μια νέα ψηφιακή πραγματικότητα. Η χρήση δεδομένων για την αυτοματοποίηση των διαδικασιών και των αποφάσεων μέσω της χρήσης νέων μεθόδων όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση είναι κύριος στόχος των οργανισμών. Στα πρώτα βήματα της, η τεχνητή νοημοσύνη, αν και προκάλεσε ενθουσιασμό και ενδιαφέρον, δεν μπόρεσε να εφαρμοστεί με την αναμενόμενη επιτυχία. Πρόσφατες τεχνολογικές εξελίξεις και ανακαλύψεις κατέστησαν τις εφαρμογές της Τεχνητής Νοημοσύνης εμπορικά βιώσιμες. Η Τεχνητή Νοημοσύνη καθίσταται μια από τις πιο δημοφιλείς τεχνολογίες η οποία πρόκειται να μετασχηματίσει τον τραπεζικό κλάδο, καθώς οι εφαρμογές της αποδεικνύονται βιώσιμες και η αποδοχή τους από τους πελάτες ικανοποιητική. Στον τραπεζικό κλάδο η Τεχνητή Νοημοσύνη εφαρμόζεται, μεταξύ άλλων, στην εξυπηρέτηση πελατών, στην ανίχνευση απάτης και ξεπλύματος μαύρου χρήματος, στην εφαρμογή κανονιστικής συμμόρφωσης, στην ανάλυση και την αξιολόγηση του πιστωτικού κινδύνου.
Στην παρούσα διπλωματική εργασία αναπτύχθηκαν τρία μοντέλα εποπτευόμενης μηχανικής μάθησης, με τα οποία γίνεται η ταξινόμηση των πελατών μίας τράπεζας σε «καλούς» ή «κακούς» με βάση την πιθανότητα αθέτησης των υποχρεώσεών τους.
Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι Random Forest, KNN και Decision Trees.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Εξόρυξη γνώσης, Μηχανική Μάθηση, Πιστωτικός κίνδυνος, Πρόβλεψη, Πιθανότητα αθέτησης, Αξιολόγηση πιστωτικού κινδύνου, Δέντρα Απόφασης

ABSTRACT

The rapid development of technology, the competitive environment and the huge amount of data available today, take into account the urgent need of companies to move to a new digital reality. The use of data to automate processes and decisions through the use of new methods such as artificial intelligence and machine learning is a major goal of organizations. In its first steps, Artificial Intelligence, although it aroused excitement and interest, could not be implemented with the expected success. Recent technological developments and discoveries have made Artificial Intelligence applications commercially viable. Artificial Intelligence is becoming one of the most popular technologies that is going to transform the banking industry, as its applications prove to be viable and their acceptance by customers satisfactory.

In the banking industry, Artificial Intelligence is applied, among other things, to customer service, fraud detection and money laundering, regulatory compliance, credit risk analysis and assessment. In this dissertation, three models of supervised machine learning were developed, which classify the customers of a bank into ”good” or ”bad” based on the possibility of default.

The algorithms used are Random Forest,KNN and Decision Trees.

SUBJECT AREA: Machine learning

KEYWORDS: Data mining, Machine learning, Credit risk, Probability of default, credit scoring, Random Forest, KNN, Decision Tree

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά τους ανθρώπους που συνέβαλαν στην εκπόνησή της. Ιδιαίτερα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ηλία Μαγκλογιάννη για την εμπιστοσύνη που μου έδειξε με την ανάθεση της διπλωματικής εργασίας αλλά και για το άρτιο κλίμα συνεργασίας που καλλιέργησε καθ' όλη τη διάρκεια εκπόνησής της. Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για τη συμπαράσταση και την κατανόηση που μου πρόσφερε σε όλη τη διάρκεια της ακαδημαϊκής μου πορείας.

Data! Data! Data! I can't make bricks without clay.

Sherlock Holmes

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα Διπλωματική Εργασία εκπονήθηκε κατά την εαρινή περίοδο του Ακαδημαϊκού Έτους 2022 - 2023, στα πλαίσια του Μεταπτυχιακού Προγράμματος “Μεγάλα Δεδομένα και Αναλυτική ” της Σχολής Τεχνολογιών Πληροφορικής και Τηλεπικοινωνιών του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς. Η εργασία πραγματοποιήθηκε υπό την επίβλεψη του κ. Ηλία Μαγκλογιάννη Καθηγητή του Τμήματος. Αντικείμενο της εργασίας αποτελεί . Η ανάλυση και η αξιολόγηση του πιστωτικού κινδύνου.

ΕΙΣΑΓΩΓΗ

Η πρόβλεψη της χρεοκοπίας έχει αποκτήσει ολοένα και αυξανόμενη σημασία στην εταιρική διακυβέρνηση. Οι παγκόσμιες οικονομίες έχουν επιδείξει επιφυλακτικότητα όσον αφορά τους κινδύνους που συνεπάγεται η εταιρική ευθύνη, ιδίως μετά την κατάρρευση των γιγαντιαίων οργανισμών όπως η WorldCom και η Enron, ενώ ένας από τους κύριους στόχους των κανονισμών της Βασιλείας ΙΙ είναι πλέον η ελαχιστοποίηση του πιστωτικού κινδύνου. Πολλά διαφορετικά μοντέλα έχουν χρησιμοποιηθεί για την πρόβλεψη της εταιρικής χρεοκοπίας. Αυτές οι μέθοδοι διαθέτουν πολλά ιδιαίτερα πλεονεκτήματα και αδυναμίες, και η επιλογή τους για εμπειρική εφαρμογή δεν είναι απλό ζήτημα. Υπάρχουν αρκετές αναφορές στη βιβλιογραφία, αλλά αυτές είναι πλέον είτε παρωχημένες (Scott, 1981 [1]; Zavgren, 1983 [2]; Altman, 1984 [3]) ή πολύ περιορισμένες. Οι Zavgren (1983)[2], Altman (1984)[3] και οι Keasey Watson (1991)[4] επικεντρώνονται αποκλειστικά σε στατιστικά μοντέλα ενώ ο Jones (1987) [5] και ο Dimitras et al. (1996) [6] δεν παρέχουν πλήρη κάλυψη των θεωρητικών μοντέλων. Οι Zhang et al. (1999)[7] περιορίζουν την ανασκόπηση τους σε εμπειρικές εφαρμογές μοντέλων νευρωνικών δικτύων ενώ οι Crouhy et al. (2000)[8] καλύπτουν μόνο τα πιο σημαντικά θεωρητικά μοντέλα πιστωτικού κινδύνου.

Για περισσότερα από διακόσια χρόνια, στον τομέα της πρόβλεψης της χρεοκοπίας ενός οργανισμού οι περισσότερες αξιολογήσεις γίνονταν υποκειμενικά. Μόλις τον εικοστό αιώνα ήρθαν περισσότερες ποσοτικές διαδικασίες που κέρδισαν κάποια προσοχή, όπως το έργο βασικής μονομεταβλητής ανάλυσης του Beaver (1966) [9] και το έργο ανάλυσης πολλαπλών διακρίσεων του Altman στη δεκαετία του 1960. Η δουλειά τους έδειξε την ικανότητα να προβλέπουν την αποτυχία ενός οργανισμού ακόμη και πέντε χρόνια πριν αυτή συμβεί. Τέτοια δεδομένα αποτελούν πηγή όχι μόνο για τον οργανισμό ή τους επενδυτές, αλλά και για πολλούς διαφορετικούς συνεργάτες, για παράδειγμα, προμηθευτές και υπαλλήλους. Για να κατανοήσουμε καλύτερα τη σημασία και τις πιθανές επιπτώσεις μιας χρεοκοπίας ενός οργανισμού σε όλους, πρέπει να θυμηθούμε τι συνέβη στη Lehman Brothers Holdings Inc. Αν είχαν έναν αξιόπιστο αλγόριθμο πρόβλεψης, θα μπορούσαν να είχαν δει τα μοτίβα και να αποτρέψουν αυτή την καταστροφή από το να συμβεί. Οι εταιρείες σε όλο τον κόσμο θα ήθελαν να έχουν ένα τέτοιο σύστημα να τους βοηθήσει να προβλέψουν πιθανή κρίση και να λάβουν μέτρα για να την αποφύγουν. Σύμφωνα με την McKinsey Co, οι λειτουργίες κινδύνου στα τραπεζικά ιδρύματα θα πρέπει να είναι πολύ διαφορετικές μέχρι το 2025. Η επέκταση των κανονισμών, οι αυξανόμενες προσδοκίες των πελατών και η εξέλιξη του κινδύνου Οι τύποι αναμένεται να οδηγήσουν σε νέα προϊόντα, υπηρεσίες και τεχνικές διαχείρισης κινδύνου. Η μηχανική μάθηση μπορεί να επιτρέψει τη δημιουργία πιο ακριβών μοντέλων κινδύνου εντοπίζοντας πολύπλοκα, μη γραμμικά μοτίβα σε μεγάλα σύνολα δεδομένων. Αναμένεται ότι η μηχανική εκμάθηση θα εφαρμοστεί σε πολλούς τομείς εντός μιας τράπεζας, ειδικά στον τομέα της λειτουργίας διαχείρισης κινδύνου. Η μελέτη επιδιώκει να εξετάσει τον βαθμό στον οποίο η μηχανική μάθηση, που έχει γίνει πολύ σημαντικός παράγοντας για τις επιχειρήσεις, έχει ερευνηθεί στο πλαίσιο της διαχείρισης κινδύνων στον τραπεζικό κλάδο.

Στόχος της διατριβής είναι να αναλύσει και να αξιολογήσει τις τεχνικές μηχανικής μάθησης που εφαρμόζονται στη διαχείριση κινδύνου σε χρηματοπιστωτικά ιδρύματα και να εξετάσει πιθανά προβλήματα που έχουν εντοπιστεί.

Το σύνολο δεδομένων που χρησιμοποιείται στην εργασία βρίσκεται στο Kaggle, στο οποίο περιλαμβάνονται δεδομένα 32.581 δανειοληπτών. Παρέχει πληροφορίες σχετικά με την ηλικία, το εισόδημα, την κατάσταση του σπιτιού, διάρκεια απασχόλησης, πρόθεση δανείου, ποσό δανείου, βαθμός δανείου, επιτόκιο, αναλογία δανείου προς εισόδημα, ιστορική αθέτηση πληρωμών και κατάσταση δανείου κάθε δανειολήπτη.

1. Βιβλιογραφικές Αναφορές

Εισαγωγή

Σύμφωνα με μελέτες για το θέμα της ανάλυσης πιστωτικού κινδύνου με μηχανική μάθηση,φαίνεται ότι οι τεχνικές μηχανικής μάθησης (Machine Learning, ML) είναι ανώτερες από τα παραδοσιακά στατιστικά μοντέλα [10]. Η μελέτη του Malhotra έδειξε ότι τα νευρωνικά δίκτυα έχουν καλύτερα αποτελέσματα από τις παραδοσιακές στατιστικές τεχνικές. Ωστόσο, οι Huang και Day (2013) έδειξαν ότι τα μοντέλα Support Vector Machine (SVM) έχουν καλύτερα ποσοστά ακρίβειας μεταξύ των 17 μοντέλων ταξινόμησης που εξετάστηκαν, όσον αφορά το πιστωτικό σκορ. Αυτό υποστηρίζεται και από τους Khemakhem και Boujelbene (2017) [11], οι οποίοι διεξήγαγαν μια μελέτη για την αξιολόγηση πιστωτικού κινδύνου στις τράπεζες της Τυνησίας και συνέκρινε τα παραδοσιακά μοντέλα με τα σύγχρονα τεχνητά νευρωνικά δίκτυα (ANN) και SVM. Οι Nwulu και Nnamdi, Shola Oroja Mustafa Ilkan (2011) [12] πραγματοποίησαν μια συγκριτική ανάλυση των SVM και των ANN για την αξιολόγηση της πιστοληπτικής ικανότητας και κατέληξε στο συμπέρασμα ότι τα ANN είχαν ελαφρώς καλύτερη απόδοση από τα SVM. Η αξιολόγηση του πιστωτικού κινδύνου γίνεται μέσω της ανάπτυξης μοντέλων ταξινόμησης, προκειμένου να γίνει διάκριση μεταξύ αξιόπιστων και αναξιόπιστων πελατών [13]. Μια κοινή προσέγγιση για την αξιολόγηση πιστωτικού κινδύνου είναι η εφαρμογή κάποιου είδους τεχνικής ταξινόμησης σε προηγούμενα δεδομένα πελατών ώστε να μπορέσουμε να βρούμε κάποια σχέση μεταξύ των χαρακτηριστικών των πελατών και της αποτυχίας αποπληρωμής του δανείου. Φαίνεται ότι υπάρχει ένα αυξανόμενο ερευνητικό ενδιαφέρον για την αξιολόγηση του πιστωτικού κινδύνου μέσω των τεχνικών της μηχανικής μάθησης.

Πιστωτική Βαθμολογία (credit scoring)

Η βαθμολόγηση με χρήση μηχανικής εκμάθησης γίνεται γενικά με τη χρήση ορισμένων ειδών ταξινομητή που διαφοροποιεί αξιόπιστους και αναξιόπιστους πελάτες και χρησιμοποιεί δεδομένα προηγούμενων πελατών. Οι τεχνικές ML που χρησιμοποιούνται από τους ερευνητές για τη βαθμολογία πιστώσεων είναι τα νευρωνικά δίκτυα, το SVM, το Naive Bayes, Bayesian Networks, Decision Tree, Hybrid models και Ensemble models. Τα νευρωνικά δίκτυα έχουν γίνει ολοένα και πιο δημοφιλή στους ερευνητές τα τελευταία χρόνια.Οι Rong-Zhou Li, Su-Lin Pang, Jian-Min Xu (2002) [14] πρότεινε ένα μοντέλο βασισμένο στον αλγόριθμο Back-Propagation (BP) για τον προσδιορισμό των καλών έναντι των κακών πιστωτών. Οι Xin-yue Hu και Yongli Tang (2006) [15] πρότειναν μια ενημερωμένη αξιολόγηση πιστωτικού κινδύνου με βάση το τεχνητό νευρωνικό δίκτυο (ANN), το οποίο μετρά το πιστωτικό σκορ του αιτούντος. Οι πιο κατάλληλοι υποψήφιοι για αυτό το μοντέλο είναι οι εμπορικές τράπεζες που έχουν ελλιπή στοιχεία.
Ο Dima et al. (2010) [13] πρότεινε ένα μοντέλο ANN για την αξιολόγηση εταιρικού πιστωτικού κινδύνου για την ταξινόμηση των καλών πιστωτών από τους κακούς. Στο έγγραφό τους, αξιολογούν τον κίνδυνο αθέτησης υποχρεώσεων της εταιρείας σε ένα διεθνές δείγμα 3.000 εταιρειών που

υποβάλλουν αίτηση για πίστωση σε διεθνή τράπεζα που λειτουργεί στη Ρουμανία. Το δείγμα περιλαμβάνει τον γενικό πληθυσμό των εταιρειών στη Ρουμανία. Με βάση το παρελθόν πιστωτικό ιστορικό τους, έχουν χωρίσει τις εταιρείες σε επτά κατηγορίες. Έκαναν τις εκτιμήσεις τους αρχικά χρησιμοποιώντας λογιστική παλινδρόμηση και στη συνέχεια ANN (τεχνητά νευρωνικά δίκτυα) και συνέκριναν τα αποτελέσματα με το Standard Poor's.

Οι Tomczak και Zieba (2014) [16] στη μελέτη τους πρότειναν μια νέα τεχνική μηχανικής μάθησης που χρησιμοποιεί την Περιορισμένη Ταξινόμηση Boltzmann Machine (ClassRBM) για την κατασκευή του πίνακα επιδόσεων. Οι πίνακες αποτελεσμάτων είναι τα πιο απλά μοντέλα για ερμηνεία και μπορούν εύκολα να εφαρμοστούν σε οποιοδήποτε τραπεζικό σύστημα. Σε αντίθεση με τις τυπικές μεθόδους, η προσέγγισή τους χρησιμοποιεί τον ισχυρό ταξινομητή (ClassRBM), αντιμετωπίζει το πρόβλημα της άνισης κατανομής κλάσεων και κατασκευάζει ένα εξαιρετικά κατανοητό και εύκολο στην εφαρμογή μοντέλο βαθμολόγησης. Οι B Baesens, T Van Gestel, S Viaene, M Stepanova, J Suykens J Vanthienen (2003) [17] ανέλυσαν τρία σύνολα δεδομένων της πραγματικής ζωής και παρουσίασαν τα αποτελέσματα. Η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας τεχνικές εξαγωγής κανόνων νευρωνικών δικτύων. Συνήχθη το συμπέρασμα ότι οι τεχνικές εξαγωγής νευρωνικών κανόνων μπορούν να χρησιμοποιηθούν για την ανάλυση πιστωτικού κινδύνου. Όπως φαίνεται, οι ερευνητές κινούνται σε υβριδικά συστήματα με νευρωνικά δίκτυα. Οι Huang et al. (2005) πρότεινε την ταξινόμηση των αιτούντων δάνειο από τις κρατικές εμπορικές τράπεζες χρησιμοποιώντας ασαφή νευρωνικά δίκτυα.

Οι Stjepan Oreski, Dijan aOreski και Goran Oreski (2012) [18] πρότειναν ένα υβριδικό σύστημα με Γενετικό Αλγόριθμο (GA) και ANN για την πιστοληπτική ικανότητα των αιτούντων.

Σε αυτό το μοντέλο, η επιλογή των χαρακτηριστικών γίνεται με χρήση GA και η ταξινόμηση με χρήση ANN. Το προτεινόμενο υβριδικό σύστημα βρέθηκε να είναι ανταγωνιστικό με άλλα μοντέλα. Αυτό το άρθρο παρουσιάζει έναν προηγμένο νέο ευρετικό αλγόριθμο, τον υβριδικό γενετικό αλγόριθμο νευρωνικών δικτύων (HGA-NN), ο οποίος χρησιμοποιείται για την αύξηση της ακρίβειας ταξινόμησης στην αξιολόγηση πιστωτικού κινδύνου προσδιορίζοντας ένα βέλτιστο υποσύνολο χαρακτηριστικών. Η απόδοση του προτεινόμενου ταξινομητή αξιολογείται χρησιμοποιώντας ένα σύνολο πιστωτικών δεδομένων που συλλέγονται σε τράπεζα της Κροατίας και τα αποτελέσματα επικυρώνονται περαιτέρω σε ένα άλλο σύνολο πραγματικών δεδομένων πιστώσεων που επιλέγονται από μια βάση δεδομένων UCI. Η ακρίβεια ταξινόμησης συγκρίνεται με αυτή που παρουσιάζεται στη βιβλιογραφία.

Τα ευρήματα ήταν πολλά υποσχόμενα για το χαρακτηριστικό επιλογή και ταξινόμηση στην αξιολόγηση πιστωτικού κινδύνου λιανικής και δείχνουν ότι το HGA-NN είναι μια πολλά υποσχόμενη προσθήκη στις υπάρχουσες τεχνικές εξόρυξης δεδομένων. Οι Yacine Djemaiel, Nadia Labidi Noureddine Boudriga (2016) [19] μελέτησαν ένα μοντέλο υβριδικού νευρωνικού δικτύου που δημιουργήθηκε χρησιμοποιώντας έναν συνδυασμό του νευρωνικού δικτύου της συνάρτησης ακτινικής βάσης (RBF) και του νευρωνικού δικτύου Elman. Το πλαίσιο δεδομένων ορίστηκε χρησιμοποιώντας μεγάλα δεδομένα. Το προτεινόμενο μοντέλο αποδείχθηκε αποτελεσματικό όταν χρησιμοποιήθηκε για την ταξινόμηση των πελατών ως "καλών" ή "κακών" με βάση τα πιστωτικά

τους σκορ. Επομένως, το προτεινόμενο υβριδικό μοντέλο μπορεί να είναι μια καλή επιλογή κατά την επιλογή μιας τεχνικής βαθμολόγησης για τη βαθμολόγηση της πιστοληπτικής ικανότητας.
Το SVM είναι μια ευρέως μελετημένη τεχνική ταξινόμησης της πιστοληπτικής ικανότητας για πολλούς λόγους. Τα SVM παρέχουν μια εξαιρετική ικανότητα γενίκευσης. Είναι επίσης σχετικά εύκολο στην εκπαίδευση. Τα SVM ανταποκρίνονται σχετικά καλά σε πολυδιάστατα δεδομένα.
Πολλοί έχουν χρησιμοποιήσει το SVM για να κερδίσουν πιστώσεις. Οι M. A. H. Farquad, V. Ravi, Sriramjee G. Praveen (2011) [20] στο άρθρο τους, πρότειναν ένα υβριδικό μοντέλο SVM για σκοπούς διαχείρισης πελατειακών σχέσεων (CRM). Η προσέγγιση αποτελείται από τρεις φάσεις. Στην πρώτη φάση το SVM-RFE (SVM-recursive feature elimination) χρησιμοποιείται για τη μείωση του συνόλου χαρακτηριστικών. Το μειωμένο σύνολο δεδομένων χρησιμοποιείται στη συνέχεια στη δεύτερη φάση για να ληφθεί ένα μοντέλο SVM. Στη συνέχεια δημιουργούνται οι κανόνες χρησιμοποιώντας το Naive Bayes Tree (NBTree) στην τελική φάση. Το σύνολο δεδομένων που αναλύθηκε σε αυτή τη μελέτη αφορά την πρόβλεψη του Churn στους πελάτες της τραπεζικής κάρτας (Business Intelligence Cup 2004) και είναι εξαιρετικά ανισόρροπο με 93,24% πιστούς και 6,76% αποσπασμένους πελάτες. Από τα εμπειρικά αποτελέσματα παρατηρείται ότι το προτεινόμενο υβρίδιο ξεπέρασε όλες τις άλλες τεχνικές που δοκιμάστηκαν. Feng et al. (2009) πρότεινε το μοντέλο ταξινόμησης SVM που βασίζεται σε PCA για μείωση διαστάσεων για εμπορικές τράπεζες. Είναι παρόμοιο με το μοντέλο PCA-SVM που προτείνεται από τους Farquad et al (2011). Μια σύγκριση με το νευρωνικό δίκτυο backpropagation (BP) έδειξε ότι η ακατέργαστη μέθοδος SVM είναι πιο ακριβής και αποτελεσματική από αυτήν. Gestel et al. (2003) πρότεινε έναν ταξινομητή SVM Least Squares για μια πιστωτική βαθμολογία που ξεπερνούσε τους παραδοσιακούς ταξινομητές SVM. Αυτή η μέθοδος αποδείχθηκε καλύτερη από τα παραδοσιακά μοντέλα Γραμμικής Διακριτικής Ανάλυσης (LDA) και Λογιστικής Παλινδρόμησης. Εκτός από τις προσεγγίσεις που βασίζονται σε νευρωνικά δίκτυα και SVM, προτείνονται αρκετές άλλες τεχνικές ταξινόμησης για την πιστοληπτική ικανότητα. Αν και δεν είναι ένα δημοφιλές μοντέλο ταξινόμησης για το πιστωτικό αποτέλεσμα, το Η προσέγγιση Naive Bayes έχει επίσης προταθεί. Οι Vedala και Kumar (2012) [21] πρότειναν μια αξιολόγηση Naive Bayes για την πιστοληπτική ικανότητα. Αυτή η αξιολόγηση γίνεται κυρίως σε πλατφόρμες ηλεκτρονικού δανεισμού που χρησιμοποιούν κοινωνικά δίκτυα για να επεκτείνουν τη βάση δεδομένων τους.
Στο άρθρο των Olatunji J. Okesola, Kennedy O. Okokpujie, Adeyinka A. Adewale, Samuel N. John και Osemwegie Omoruyi (2017) [14] μελετήθηκε επίσης ένα μοντέλο ταξινόμησης Naïve Bayes για την πιστοληπτική ικανότητα. Οι μεταβλητές εισόδου σε αυτή τη μέθοδο είναι οι δημογραφικοί και σημαντικοί δείκτες. Μια σύγχρονη προσέγγιση της πιστοληπτικής ικανότητας είναι η μέθοδος του δέντρου αποφάσεων (Hand et al., 1997, [22]). Οι Szwabe και Misiorek (2018) [23] πρότειναν ένα μοντέλο δέντρου αποφάσεων για τη λήψη πιστωτικών αποφάσεων.

Μη εξυπηρετούμενα Στοιχεία - Non-Performing Asset (NPA)

Ένας άλλος τύπος τεχνικής αξιολόγησης πιστωτικού κινδύνου είναι ο NPA. Αυτό έχει να κάνει με την πρόβλεψη ποιο δάνειο είναι πιθανό να αθετηθεί,

ώστε να ληφθούν τα κατάλληλα μέτρα για την αντιμετώπιση της κατάστασης. Ο Baruah (2018) [24] μελέτησε τις εφαρμογές της τεχνητής νοημοσύνης σε 4 κορυφαίες ινδικές τράπεζες και καταλήγει στο συμπέρασμα ότι η χρήση της Μηχανικής Μάθησης στα δεδομένα πελατών οδηγεί σε καλύτερη εξυπηρέτηση και παρέχει στους πελάτες καλύτερη εμπειρία, τόσο από άποψη ταχύτητας όσο και από άποψη ποιότητας. Ο D'Monte (2018) [25] κατέληξε στο συμπέρασμα ότι οι αλγόριθμοι μηχανικής μάθησης μπορούν να συνδέσουν τους χρήστες με διάφορες τραπεζικές υπηρεσίες, παρακολουθεί τη συμπεριφορά δαπανών τους και το πρότυπο συμπεριφοράς τους, ώστε να μπορεί να εντοπίσει οποιαδήποτε συναλλαγή που είναι αμφισβητήσιμη επειδή δεν ταιριάζει με το προφίλ του πελάτη. Οι Ahmad και Ariff (2007) [26] μελέτησαν τους πιστωτικούς κινδύνους στις ανεπτυγμένες οικονομίες της Αυστραλίας, των ΗΠΑ, της Ιαπωνίας, της Γαλλίας και οι αναδυόμενες οικονομίες της Ινδίας, της Μαλαισίας, της Ταϊλάνδης, του Μεξικού. Η μελέτη κατέληξε στο συμπέρασμα ότι ο πιστωτικός κίνδυνος στις τράπεζες των αναδυόμενων οικονομιών είναι υψηλότερος από αυτόν των ανεπτυγμένων. Οι τεχνικές ML που χρησιμοποιούνται για την προεπιλεγμένη πρόβλεψη είναι διαφορετικοί τύποι νευρωνικών δικτύων, SVM και υβριδικά μοντέλα. Ο Zhang (2011) [27] πρότεινε ένα προεπιλεγμένο μοντέλο κινδύνου έγκαιρης προειδοποίησης με βάση τον αλγόριθμος νευρωνικού δικτύου BP. Ένα νευρωνικό δίκτυο BP εκπαιδεύεται σε δείγματα δεδομένων για τον προσδιορισμό του κίνδυνου αθέτησης. Τέλος στο άρθρο τους, οι Asma Feki,Anis BenIshak και Saber Feki (2012) [28] προτείνουν μεθόδους διάκρισης των τραπεζών με βάση το ποσοστό των μη εξυπηρετούμενων δανείων (ΜΕΔ), χρησιμοποιώντας διαφορετικές προσεγγίσεις πολυκλασικών μοντέλων SVM και Gaussian.

Ανίχνευση Απάτης

Η απάτη στις χρηματοοικονομικές συναλλαγές μπορεί να θέσει σε κίνδυνο τη φήμη των χρηματοπιστωτικών ιδρυμάτων καθώς και να προκαλέσει μεγάλες απώλειες. Οι τράπεζες και τα χρηματοπιστωτικά ιδρύματα επενδύουν στους αλγόριθμοι μηχανικής μάθησης και στα συστήματα ανίχνευσης απάτης (Youness Abakarim, Mohamed Lahby, Abdelbaki Attioui, 2018, [29]).
Η ανίχνευση της απάτης είναι ένα πρόβλημα δυαδικής ταξινόμησης. Η ιδέα είναι να εφαρμοστεί ένας κατάλληλος ταξινομητής στο πρόβλημα, το οποίο θα εκπαιδευτεί σε ένα κατάλληλο σύνολο δεδομένων. Οι κύριες προσεγγίσεις για την αντιμετώπιση της απάτης στις πιστωτικές κάρτες είναι τεχνικές ML όπως ο SVM και τα δέντρα απόφασης. Οι Zareapoor and Shamsolmoali (2015) [30] δημοσίευσαν ένα άρθρο για τον εντοπισμό απάτης χρησιμοποιώντας διαφορετικά τεχνικές, όπως Naïve Bayes, KNN και SVM. Στην εργασία τους, εγείρουν ανησυχίες για τη διαθεσιμότητα πραγματικών δεδομένων καθώς τα χρηματοπιστωτικά ιδρύματα δεν αποκαλύπτουν τα δεδομένα τους επειδή είναι εμπιστευτικά και ευαίσθητα. Έτσι, οι έρευνες ολοκληρόνονται με "πλαστά" δεδομένα. Ένα άλλο πρόβλημα είναι ότι, τα περισσότερα από τα όταν τα δεδομένα δεν είναι ισορροπημένα καθώς ο αριθμός των δόλιων συναλλαγών είναι μόνο το 2% ενώ το 98% των συναλλαγών είναι νόμιμες. Αναφέρονται στις ανησυχίες τους για τον μεγάλο όγκο δεδομένων και το χρόνο υπολογισμού που απαιτείται για την εκτέλεση ενός αλγόριθμου σε αυτές τις περιπτώσεις.

Μία από τις μεγαλύτερες προκλήσεις που αναφέρεται συχνά σε πολλές ερευνητικές εργασίες είναι ότι οι αλγόριθμοι μηχανικής μάθησης πρέπει να ενημερώνεται τακτικά, ώστε οι κακόβουλες προσπάθειες να μπορούν να καταγράφονται σε πραγματικό χρόνο. Ο αλγόριθμος AdaBoost χρησιμοποιείται στην εργασία των Randhawa et al. (2018) [31] σχετικά με τον εντοπισμό απάτης σπιστωτικές κάρτες και εξετάζει πολλά διαφορετικά μοντέλα μηχανικής μάθησης, όπως ο Naïve Bayes, το Random Forest κ.λπ. Η μελέτη αναφέρει ότι ο AdaBoost είναι πολύ ευαίσθητος σε ανωμαλίες και ακραίες τιμές. Boltzmann (RBM) μηχανές μπορούν να χρησιμοποιηθούν για την ανακατασκευή δεδομένων σε ένα μη ελεγχόμενο περιβάλλον μάθησης. Μια βάση δεδομένων απο 18.000 τεχνητές εικόνες, που απεικονίζουν 300 δραστηριότητες πελατών για περισσότερες από 60 ημέρες. Ένα CNN (Συνελικτικό Νευρωνικό Δίκτυο) έχει εφαρμοστεί στις εικόνες για τον εντοπισμό δόλιας δραστηριότητας. Οι Gyamfi και Abdulai (2018) [32] χρησιμοποίησαν το SVM with Spark (SVM-S) για να επεξεργαστούν ροές δεδομένων που αντιπροσωπεύουν καλή και κακή συμπεριφορά πελατών και στη συνέχεια χρησιμοποιήστε τα για να αξιολογήσετε την εγκυρότητα των νέων συναλλαγών. Κωτσιάντης κ.ά. (2005) [33] προέβλεψε δόλιες οικονομικές καταστάσεις χρησιμοποιώντας ένα δέντρο αποφάσεων. Το δέντρο απόφασης αποδείχθηκε ότι πέτυχε την καλύτερη απόδοση μεταξύ όλων των ταξινομητών που εξετάστηκαν.

Οι τάσεις της έρευνας στην αξιολόγηση πιστωτικού κινδύνου

Καθώς έχουν σημειωθεί διαρθρωτικές αλλαγές στην παγκόσμια χρηματοπιστωτική αγορά καθώς και αύξηση του κινδύνου, έχει καταστεί επιτακτική ανάγκη να μελετηθεί η αξιολόγηση πιστωτικού κινδύνου. Τα τελευταία 20 χρόνια, έχει σημειωθεί μεγάλη πρόοδος στον τομέα της αξιολόγησης πιστωτικού κινδύνου. Τα μοντέλα πιστοληπτικής αξιολόγησης δημιουργήθηκαν από δύο βασικά και ακόμα δημοφιλή στατιστικά εργαλεία: τη Γραμμική Διακριτική Ανάλυση (LDA) και τη λογιστική παλινδρόμηση (LR). Καθώς οι καιροί αλλάζουν, έχουν φτάσει νέες μέθοδοι όπως τα νευρωνικά δίκτυα,SVM, k-NN και Δέντρα Αποφάσεων. Υπάρχουν πολλές άλλες μέθοδοι όπως περιγράφονται παραπάνω. Ωστόσο, τα υβριδικά μοντέλα και τα μοντέλα συνόλου γίνονται όλο και πιο δημοφιλή. Πρωτογενής έρευνα που διεξάγεται στον τομέα της αξιολόγησης πιστωτικού κινδύνου χρησιμοποιεί μη γραμμικούς αλγόριθμους ταξινόμησης, όπως π.χ νευρωνικά δίκτυα και SVM. Το SVM έχει λάβει μεγάλη προσοχή στη μηχανική μάθηση κοινότητα. Λίγοι προσπάθησαν να κερδίσουν πίστωση χρησιμοποιώντας την ταξινόμηση Naive Bayes (Vedala et al., 2012 [34]). Και για τους τρεις τύπους τεχνικών αξιολόγησης πιστωτικού κινδύνου, οι ερευνητές έχουν επίσης προτείνει πολλά υβριδικά μοντέλα που συνδυάζουν μέρη δύο ή περισσότερων αλγορίθμων.

Μηχανική Μάθηση

Ζούμε σε μια εποχή όπου τα δεδομένα τα οποία παράγονται είναι πάρα πολλά και η αξιοποίησή τους με χρήση διαφόρων αλγορίθμων από το πεδίο της μηχανικής μάθησης μπορεί να μετατρέψει αυτά τα δεδομένα σε σημαντική γνώση. Η μηχανική μάθηση είναι ένα πεδίο έρευνας το οποίο προέρχεται από την διασταύρωση τριών πεδίων: της στατιστικής, της τεχνητής νοημοσύνης και

της επιστήμης των υπολογιστών [35]. Το 1959 ο Arthur Samuel ορίζει τη μηχανική μάθηση ως “Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί.” [36] Ένας άλλος ορισμός που αφορά τη μηχανική μάθηση, αυτή την φορά από τον Tom M. Mitchell το 1997, αναφέρει ότι “ένα πρόγραμμα υπολογιστών λέγεται ότι μαθαίνει από την εμπειρία E, σε σχέση με κάποια τάξη εργασιών T και μέτρηση απόδοσης P (Performance Measure) εάν η απόδοσή του σε εργασίες στο T, όπως μετράτε από το P, βελτιώνεται με την εμπειρία E.” [36], [37]. Με άλλα λόγια, η μηχανική μάθηση μπορεί να θεωρηθεί ως μια συλλογή από μεθόδους που μπορούν αυτόματα να αναγνωρίσουν διάφορα μοτίβα στα δεδομένα και στη συνέχεια με βάση αυτά τα μοτίβα να προβλέψουν μελλοντικά αποτελέσματα ή να πάρουν αποφάσεις κάτω από συγκεκριμένες καταστάσεις. Όλα αυτά έχουν τη δυνατότητα να αξιοποιηθούν χρησιμοποιώντας διάφορους αλγορίθμους που επιτρέπουν στις μηχανές να καταλαβαίνουν διάφορες καταστάσεις και βασισμένες σε αυτές να παίρνονται οι αποφάσεις [38]. Η μηχανική μάθηση είναι χρήσιμη σε πολλούς τους τομείς αφού παρέχει διάφορα πλεονεκτήματα, όπως:

1. Γρήγορη απόφαση: Η μηχανική μάθηση παρέχει γρήγορα τα καλύτερα αποτελέσματα.
2. Ικανότητα προσαρμογής: Έχει τη δυνατότητα να προσαρμόζεται γρήγορα στο νέο περιβάλλον, το οποίο μεταβάλλεται συνεχώς, αφού τα δεδομένα ενημερώνονται συνεχώς.
3. Καινοτομία: Με τη χρησιμοποίηση προηγμένων αλγορίθμων βελτιώνεται η ικανότητα λήψης αποφάσεων, βοηθώντας έτσι στην ανάπτυξη καινοτόμων επιχειρηματικών υπηρεσιών
1. και μοντέλων.
4. Διορατικότητα: Γίνεται η κατανόηση μοναδικών προτύπων δεδομένων και με βάση αυτών
2. βασίζονται στις ενέργειες που μπορούν να παρθούν.
5. Επιχειρηματική ανάπτυξη: Η συνολική επιχειρηματική διαδικασία και η ροή εργασίας είναι ταχύτερες, βοηθώντας έτσι στην επιχειρηματική ανάπτυξη.
6. Καλό αποτέλεσμα: Το αποτέλεσμα θα βελτιώνεται σε αντίθεση με τη πιθανότητα σφάλματος, η οποία θα μειώνεται.

Υπάρχουν τρεις τύποι μηχανικής μάθησης, η επιβλεπόμενη μάθηση (Supervised Learning), η μη επιβλεπόμενη μάθηση (Unsupervised Learning) και η ενισχυτική μάθηση (Reinforcement Learning). Η επιβλεπόμενη μάθηση έχει σαν κύριες μεθόδους την κατηγοριοποίηση και την παλινδρόμηση, ενώ η μη επιβλεπόμενη μάθηση το μετασχηματισμό και τη συσταδοποίηση, έννοιες οι οποίες αναλύονται στη συνέχεια. Και στις δύο περιπτώσεις, τα δεδομένα εισόδου πρέπει να έχουν σωστή αναπαράσταση για να μπορεί να τα καταλάβει ένας υπολογιστής. Η ενισχυτική μάθηση ασχολείται κυρίως με διάφορες οντότητες που ονομάζονται πράκτορες, οι οποίοι παίρνουν τις αποφάσεις τους από το περιβάλλον, με σκοπό να εκτελέσουν κάποια ενέργεια.

Επιβλεπόμενη μάθηση

Η επιβλεπόμενη μάθηση είναι μια από τις πιο κοινές και επιτυχημένες μεθόδους που χρησιμοποιούνται στη μηχανική μάθηση, η οποία χρησιμοποιείται όταν θέλουμε να προβλέψουμε ένα σίγουρο αποτέλεσμα από μία δεδομένη είσοδο [39]. Ονομάζεται επιβλεπόμενη μάθηση γιατί το μοντέλο μας μαθαίνει από ένα σύνολο εκπαίδευσης δημιουργώντας έτσι ένα άλλο μοντέλο, όπου με βάση αυτό εφαρμόζεται στο νέο σύνολο δεδομένων για να προβλέψει τα αποτελέσματα [40]. Υπάρχουν δύο υποκατηγορίες επιβλεπόμενης μάθησης, η κατηγοριοποίηση και η παλινδρόμηση. Αλγόριθμοι οι οποίοι εφαρμόζουν επιβλεπόμενη μάθηση είναι η γραμμική παλινδρόμηση (Linear Regression), τα νευρωνικά δίκτυα (Neural Networks), οι μηχανές διανυσμάτων στήριξης (Support Vector Machines – SVMs), η μάθηση κατά Bayes (Bayesian Learning), τα δένδρα από- φασης (Decision Trees), ο k πλησιέστεροι γείτονες (k Nearest Neighbors – kNN), η λογιστική παλινδρόμηση (Logistic Regression) και τα τυχαία δάση (Random Forests). Πιο συγκεκριμένα, υπάρχει το σύνολο εκπαίδευσης (Training Data) μαζί με την ετικέτα κατηγοριοποίησης (Label) και στη συνέχεια η επιλογή του αλγορίθμου μηχανικής μάθησης (Machine Learning Algorithm), ο οποίος θα χρησιμοποιηθεί για την εκπαίδευση. Αφού ολοκληρωθεί η εκπαίδευση, τα νέα δεδομένα (New Data), εφαρμόζονται στο μοντέλο πρόβλεψης (Predictive Model) το οποίο βασίζεται στον προηγούμενο αλγόριθμο μηχανικής μάθησης και παράγει το αποτέλεσμα πρόβλεψης (Prediction).

Support Vector Machines (SVMs)

Το ”Support Vector Machine” (SVM) είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης. Τα SVM βασίζονται στην ιδέα της εύρεσης ενός υπερεπίπεδου που διαιρεί καλύτερα ένα σύνολο δεδομένων σε δύο κατηγορίες. Ο SVM έχει εφαρμογή στο σχεδιασμό μοντέλων πιστωτικού κινδύνου και πιστοληπτικής αξιολόγησης. Ο Wang et al. (2005) [41] παρουσιάζει ένα σταθμισμένο μοντέλο SVM με πολλά υποσχόμενα αποτελέσματα στην ανάλυση πιστωτικού κινδύνου. Ο Huang et al.(2007) [42] ανέπτυξε ένα μοντέλο αξιολόγησης πιστοληπτικής ικανότητας για να αξιολογήσει το πιστωτικό σκορ ενός αιτούντος από τα χαρακτηριστικά εισροών βασίζεται σε ένα υβριδικό SVM. Οι Yeh και Lien (2009) [39] αναγνώρισαν ότι η πρόβλεψη της πιθανότητας χρεοκοπίας ενός πελάτη είναι πρόκληση που αντιμετωπίζουν οι ερευνητές και χρειάζεται περαιτέρω μελέτη. Οι τράπεζες επιδιώκουν να αναπτύξουν αποτελεσματικά μοντέλα που μπορούν να αξιολογήσουν την πιθανότητα αθέτησης υποχρεώσεων των πελατών τους. (Barboza et al., 2017) [43] δοκίμασαν μοντέλα μηχανικής μάθησης για να προβλέψει τη χρεοκοπία ένα χρόνο πριν από αυτήν συμβεί, συγκρίνοντας την απόδοση με τα αποτελέσματα των παραδοσιακών μεθόδων. Αναφέρουν σημαντική ακρίβεια πρόβλεψης και υποδηλώνουν επίσης ότι οι τεχνικές ML μπορούν εύκολα να εφαρμοστούν με ουσιαστική ακρίβεια ταξινόμησης σε σύγκριση με τις παραδοσιακές μεθόδους. Μια τράπεζα θα μπορούσε να επωφεληθεί από την ικανότητα ενός μοντέλου να επιλέγει τους χρηματοοικονομικούς δείκτες που σχετίζονται περισσότερο με τη διαδικασία πρόβλεψης και επίσης από το υψηλό επίπεδο ακρίβειάς τους.

Δέντρα Απόφασης -Decision trees

Σε αντίθεση με άλλα μοντέλα κατηγοριοποίησης, τα δέντρα απόφασης μπορούν να έχουν συνδυασμό αριθμητικών και κατηγορικών χαρακτηριστικών, αλλά και να κατηγοριοποιήσουν ελλιπή χαρακτηριστικά [44].
Το δέντρο απόφασης είναι ένα σύνολο κόμβων και ακμών - βάσει των οποίων ταξινομείται ο πληθυσμός χρησιμοποιώντας μια επεξηγηματική μεταβλητή (xi) σε κάθε κόμβο και προκύπτει η απόφαση σχετικά με τις διαφορετικές επιλογές. Η κορυφή του δέντρου είναι ο ριζικός κόμβος (root node), τα επόμενα επίπεδα κόμβων είναι τα παιδιά κόμβοι και στο κάτω μέρος του δέντρου βρίσκονται οι τερματικοί κόμβοι που περιγράφουν την τελική ταξινόμηση (Anderson, 2007). Ο τρόπος με τον οποίο λαμβάνεται η απόφαση, καθορίζει τον τύπο και την πολυπλοκότητα του δέντρου αποφάσεων (Anderson, 2007). Στην απλούστερη μορφή του, ειδικά όταν δεν υπάρχουν επαρκή διαθέσιμα δεδομένα για ανάλυση, το σύνολο των αποφάσεων και οι κανόνες καθορίζονται εμπειρικά από μια ομάδα ειδικών. Συνήθως, όμως, τα εργαλεία που χρησιμοποιούνται στην ανάλυση χρησιμοποιούνται για τη λήψη αποφάσεων και τον καθορισμό των κανόνων. Οι πιο γνωστές μέθοδοι για την εφαρμογή πιστωτικού κινδύνου είναι ο CART (Classification and Regression Trees) από τους Breiman et al. (1984) [45]. Ο στόχος είναι να οριστούν ομοιογενείς τάξεις του πληθυσμού, ως προς το επίπεδο κινδύνου, ενώ ταυτόχρονα να μεγιστοποιηθεί η διαφορά στα επίπεδα κινδύνου μεταξύ των τάξεων. Επίσης άλλοι γνωστοί αλγόριθμοι δένδρων απόφασης είναι ο ID3, ο C4.5, ο SLIQ και ο SPRINT.
Τα πιο σημαντικά πλεονεκτήματα της χρήσης δέντρων αποφάσεων αναφέρονται παρακάτω:

1. Ως μη παραμετρική μέθοδος, δεν απαιτούνται πολλές παραδοχές για τη χρήση της.
2. Οι υπολογισμοί είναι συνήθως απλοί και η επιλογή των μεταβλητών και οι διακοπές γίνονται με ένα συγκεκριμένο στατιστικό μέτρο. απλά δέντρα, τα αποτελέσματα είναι διαφανή, ερμηνεύσιμα και εύκολο να εφαρμοστούν.
3. Μπορούν να είναι ένας γρήγορος και εύκολος τρόπος για τον εντοπισμό δανειοληπτών πολύ χαμηλού ή πολύ υψηλού κινδύνου.

Αντίθετα, μερικά κύρια μειονεκτήματα των δέντρων απόφασης είναι:

1. Γενικά δεν δίνουν τόσο καλές εκτιμήσεις όσο τα μοντέλα παλινδρόμησης. Μάλιστα επιλέγονται κυρίως όταν υπάρχει περιορισμένη διαθεσιμότητα δεδομένων.
2. Συχνά σχετίζεται με ζητήματα υπερβολικής τοποθέτησης.
3. Δεν είναι τόσο ευέλικτη όσο άλλες μέθοδοι, για παράδειγμα, τα νευρωνικά δίκτυα.

Τυχαία Δάση -Random forests

Τα τυχαία δάση (Random Forests) είναι ένας αλγόριθμος επέκτασης των δέντρων αποφάσεων ο οποίος χρησιμοποιείται και αυτός για κατηγοριοποίηση ή παλινδρόμηση, αποτελούμενος από μια συλλογή αρκετών δέντρων απόφασης. Σκοπός του αλγορίθμου είναι η δημιουργία ενός δάσους από ένα τυχαίο αριθμό δέντρων αποφάσεων, έτσι προκύπτει και το όνομα του αλγορίθμου. Τα τυχαία δάση είναι η γενίκευση των δέντρων απόφασης, όπου προκύπτει η εκτίμηση για κάθε κόμβο ως ο μέσος όρος των εκτιμήσεων που δίνονται για αυτόν τον κόμβο από ένα μεγάλο σύνολο τυχαίων δέντρων (Breiman, 2001 [46]). Για να αυξηθεί ο βαθμός ακρίβειας της εκτίμησης, τα επιμέρους δέντρα που αποτελούν το το δάσος πρέπει να είναι όσο το δυνατόν άσχετο. Σε τυχαία δάση, οι ακόλουθες τρεις παράμετροι (υπερπαράμετροι) θα πρέπει συνήθως να ορίζονται (Beutel et al., 2019 [47]:

1. Ο αριθμός των δέντρων απόφασης για το δάσος.
2. Ο αριθμός των τυχαία επιλεγμένων μεταβλητών που θα εξεταστούν σε κάθε σπάσιμο.
3. Ν Ο ελάχιστος αριθμός παρατηρήσεων που πρέπει να έχει κάθε τερματικός κόμβος, ο οποίος επίσης καθορίζει την πολυπλοκότητα των δέντρων.

Μερικά από τα κρίσιμα πλεονεκτήματα των τυχαίων δασών είναι:

1. Μείωση της υπερπροσαρμογής (overfitting) σε σύγκριση με τα δέντρα απόφασης.
2. Μοντελοποίηση γραμμικών και μη γραμμικών σχέσεων.
3. Αρκετά καλές και ακριβείς εκτιμήσεις.
4. Υψηλή διάσταση

Από την άλλη πλευρά, υπάρχουν και το μειονεκτήμα τα οποίο είναι:

1. Δεν υπάρχει διαφάνεια και έλεγχος στον τρόπο λειτουργίας του μοντέλου, εκτός από τον ορισμό των παραμέτρων.

K-Nearest neighbors (KNN)

Η μέθοδος K-Nearest γείτονες (KNN) (Cover and Hart, 1967 [48]) είναι μια απλή μη παραμετρική τεχνική που βασίζεται στη μηχανική μάθηση και την εξόρυξη δεδομένων και χρησιμοποιείται για σκοπούς ταξινόμησης και παλινδρόμησης. Στόχος είναι η κατηγοριοποίηση κάθε παρατήρησης σε μια ομάδα. Για να επιτευχθεί αυτό, ο αλγόριθμος εξετάζει τους k πλησιέστερους γείτονες μιας νέας παρατήρησης (test sample) που δεν έχει ακόμη κατηγοριοποιηθεί και στη συνέχεια εκχωρεί αυτήν την παρατήρηση στην κατηγορία που είναι πιο κοινή σε αυτές γείτονες.
Η απόσταση μεταξύ των παρατηρήσεων ορίζεται από μια μετρική (π.χ. Ευκλείδεια απόσταση, Minkowski, Manhattan, Cosine and Hamming). Η μέθοδος είναι απλή και επίσης μια καλή προσέγγιση σε περίπτωση που ο αναλυτής θέλει να προσθέσει νέες παρατηρήσεις στο δείγμα εκπαίδευσης γρήγορα και εύκολα.

Τα κύρια μειονεκτήματα της μεθόδου είναι:

1. Ο αλγόριθμος κάνει μόνο την ταξινόμηση αλλά δεν υπολογίζει τις τελικές πιθανότητες.
2. Δεν υπάρχει διαφάνεια στον τρόπο λήψης των αποφάσεων.
3. Οι υπολογιστικοί χρόνοι μπορεί να είναι μεγάλοι.
4. Τέλος, ο αλγόριθμος KNN υπόκειται σε μεγάλο βαθμό στο φαινόμενο που είναι γνωστό ως 'curse of dimensionality' και όπως αναφέρεται από τους Shalev-Shwartz Ben-David (2014) [49], το μέγεθος του δείγματος είναι απαραίτητο για να επιτευχθεί ένα σχετικά μικρό σφάλμα εκτίμησης αυξάνεται εκθετικά με το αριθμός των προγνωστικών μεταβλητών.

Περιπτώσεις σε Τράπεζες της Αμερικής

Η Τεχνητή Νοημοσύνη είναι κάτι περισσότερο από ένα απλό "τσιτάτο" στον κόσμο των Τραπεζικών και Χρηματοοικονομικών. Οι λύσεις AI και ML ήδη βοηθούν τις τράπεζες σε όλοτον κόσμο να μετατρέψουν τα δεδομένα σε κέρδος παρέχοντας ένα ασφαλέστερο και πιο βολικό περιβάλλον για τις επιχειρήσεις. Ο αριθμός των περιπτώσεων χρήσης μηχανικής μάθησης στις τραπεζικές συναλλαγές παγκοσμίως αυξάνεται συνεχώς. Δεν αποτελεί έκπληξη, γιατί η αυτοματοποιημένη υποστήριξη πελατών, ο εντοπισμός απάτης σε πραγματικό χρόνο, η καλύτερη διαχείριση δεδομένων πελατών, η μοντελοποίηση κινδύνου και ο σχεδιασμός στρατηγικής μάρκετινγκ είναι τα οφέλη που μπορεί να χρησιμοποιήσει κάθε τράπεζα για να βελτιώσει τις διαδικασίες της. Υπάρχουν περιπτώσεις χρήσης Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης στον τραπεζικό κλάδο στις Ηνωμένες Πολιτείες αλλά ωστόσο, υπάρχουν και πολλές άλλες αξιόλογες περιπτώσεις από όλο τον κόσμο.

JP Morgan

Αυτή η κορυφαία τράπεζα στις Ηνωμένες Πολιτείες έχει αναπτύξει ένα έξυπνο σύστημα συμβάσεων που ονομάζεται Contract Intelligence (COiN). Ο αλγόριθμος που βασίζεται σε δεδομένα και τη μηχανική μάθηση βοηθά στη γρήγορη εύρεση των απαραίτητων εγγράφων και των σημαντικών πληροφοριών που περιέχονται σε αυτά. Αυτή τη στιγμή, η τράπεζα συνεργάζεται με περισσότερες από 12.000 δανειακές συμβάσεις και θα χρειαστούν αρκετά χρόνια για να τις αναλύσει χειροκίνητα.

Bank of America

Το chatbot από αυτήν την τράπεζα είναι ένας πραγματικός οικονομικός σύμβουλος και στρατηγός. Το σύστημα αναλύει τα δεδομένα των χρηστών και προειδοποιεί σε περιπτώσεις που ο πελάτης έχει δείξει ελαφρώς διαφορετικές αγοραστικές συνήθειες και του υπενθυμίζει την ανάγκη να πληρώσει τους λογαριασμούς του. Το chatbot της Bank of America γνωρίζει επίσης πώς να εκτελεί απλές λειτουργίες με τραπεζικές κάρτες, όπως κλείδωμα και ξεμπλοκάρισμα καρτών.

Wells Fargo

Αυτή η τράπεζα έχει αναπτύξει ένα έξυπνο chatbot για να μετατρέψει την αλληλεπίδραση με τον ιστότοπο σε μια απλή και βολική διαδικασία. Η τράπεζα Wells Fargo ανέπτυξε το σύστημα ανάλυσης Predictive Banking, το οποίο είναι σε θέση να ειδοποιεί τους πελάτες για ασυνήθιστες καταστάσεις. για παράδειγμα, εάν ο πελάτης έχει ξοδέψει περισσότερο από το μέσο ποσό των επιταγών του. Το σύστημα μπορεί επίσης να προσφέρει εξοικονόμηση ενός συγκεκριμένου ποσού σε μια κατάθεση, εάν ο πελάτης έλαβε μεταφορά χρημάτων που είναι μεγαλύτερη από το χρηματικό ποσό που συνήθως διατηρεί στον λογαριασμό του.

Citibank

Η Citibank έχει αναπτύξει ένα ισχυρό σύστημα πρόληψης απάτης που παρακολουθεί ανωμαλίες στη συμπεριφορά των χρηστών. Συγκεκριμένα, το σύστημα είναι γυαλισμένο για να ανιχνεύει δόλιες συναλλαγές με πιστωτικές κάρτες κατά τις αγορές στο Διαδίκτυο.

Bank of US

Αυτή η τράπεζα έχει αναπτύξει τον Οδηγό Εξόδων , μια εφαρμογή που επιτρέπει στους πελάτες να διαχειρίζονται τους λογαριασμούς τους καθώς και να κάνουν κράτηση αεροπορικών εισιτηρίων και καταλυμάτων στο εξωτερικό. Αυτή η εφαρμογή εστιάζει σε ασφαλείς πληρωμές σε άλλες χώρες. Είναι πολύ βολικό για όσους πηγαίνουν επαγγελματικό ταξίδι χωρίς εταιρική πιστωτική κάρτα, καθώς η εφαρμογή επιτρέπει στον χρήστη να συλλέγει όλα τα οικονομικά δεδομένα σχετικά με το ταξίδι σε ένα μέρος και να δημιουργεί μια αναφορά για το οικονομικό τμήμα της εταιρείας του.

PNC

Αυτή η τραπεζική εταιρεία συμμετοχών και η εταιρεία χρηματοοικονομικών υπηρεσιών επένδυσαν 1,2 δισεκατομμύρια δολάρια από το 2016 έως το 2021 στη Μηχανική Εκμάθηση, με στόχο να αποκτήσουν ταχύτερες, ασφαλέστερες και πιο σταθερές υπηρεσίες και λειτουργίες. Η εταιρεία πόνταρε σε ένα εσωτερικό περιβάλλον cloud, αξιοποιώντας στο έπακρο το AI και το ML. Η PNC ένωσε τις δυνάμεις της με έναν προμηθευτή τεχνητής νοημοσύνης που ονομάζεται Anaconda για αυτό το έργο για να ανανεώσει την υποδομή της επιστήμης δεδομένων και να την προσαρμόσει για R και Python. Αυτή η κίνηση είχε ως αποτέλεσμα η PNC να μπορέσει να δημιουργήσει εσωτερικά μοντέλα Machine Learning και, επιπλέον, να μεταφέρει την υποδομή του PNC στο Anaconda Enterprise 5.2.

Bank of NY Mellon Corp

Μία από τις άλλες εμπνευσμένες περιπτώσεις χρήσης Machine Learning στον τραπεζικό τομέα προέρχεται από έναν οργανισμό με πάνω από 200 χρόνια ιστορίας στον κλάδο. Ποντάρουν στη ρομποτική αυτοματοποίηση διεργασιών

(RPA) για να εξοικονομήσουν κόστος και να ενισχύσουν την αποτελεσματικότητα των λειτουργιών. Θα μπορούσατε να υποστηρίξετε ότι το RPA δεν είναι Τεχνητή Νοημοσύνη ή Μηχανική Μάθηση και θα είχατε δίκιο. Ωστόσο, το RPA ενσωματώνεται με την Τεχνητή Νοημοσύνη και ορισμένες διαδικασίες εκτελούνται από ”ρομπότ” λογισμικού, όχι από πραγματικά μηχανικά ρομπότ.
Πριν από τέσσερα χρόνια, η Bank of NY Mellon Corp κυκλοφόρησε σχεδόν διακόσια bot για να χειρίζονται αυτόματα διάφορες εργασίες, όπως να ζητούν δεδομένα από εξωτερικούς ελεγκτές, να μεταφέρουν κεφάλαια και να διορθώνουν λάθη μορφοποίησης και άλλων δεδομένων. Ως αποτέλεσμα της ενσωμάτωσης του RPA, ο οργανισμός κατάφερε να λάβει 100% ακρίβεια των επικυρώσεων στο κλείσιμο λογαριασμών σε πέντε διαφορετικά συστήματα. Ο χρόνος επεξεργασίας έχει βελτιωθεί σχεδόν κατά 90%, ενώ ο χρόνος διεκπεραίωσης του εμπορίου βελτιώθηκε
σχεδόν κατά 70%.

2. Δεδομένα και Μεθοδολογία

Δεδομένα

Το σύνολο δεδομένων που πρόκειται να χρησιμοποιηθεί για τους σκοπούς αυτής της εργασίας βρίσκεται στην πλατφόρμα του Kaggle (https://www.kaggle.com/laotse/credit-risk-dataset), το οποίο περιέχει δεδομένα από 32.581 δανειολήπτες και 11 μεταβλητές (Age -ηλικία , Annual income -ετήσιο εισόδημα, Home ownership – κατοικία που του ανήκει, Employment length -Διάρκεια απασχόλησης, Loan intent -Πρόθεση δανείου, Loan grade - Βαθμολογία δανείου, Loan amount - Ποσό δανείου, Interest rate - Επιτίκιο, Loan status -Κατάσταση δανείου, Percent income -Ποσοστό εισοδήματος, Historical default -Ιστορική αθέτηση, Credit history -Πιστωτικό Ιστορικό) που σχετίζονται με κάθε δανειολήπτη.

Feature Name	***Description***
person age	Age in years
person income	Annual Income in dollars
person home ownership	Home ownership
person emplength	Employment length (in years)
loan intent	Loan intent
loan grade	Loan grade
loan amnt	Loan amount
loan intrate	Interest rate
loan status	Loan status (0 is non default 1 is default)
loan percent income	Percent income
cb person default on file	Historical default
cb presoncredhist length	Credit history length

Εικόνα 1: Dataset variables

Numerical variable	*Mean*	*Min*	*Max*
Age	27.7 years	20 years	144 years
Annual income	66,074.8 $	4,000$	6,000.000$
Employment length	4.8 years	0 years	123 years
Loan amount	9,589.4 $	500.00 $	35,000 $
Interest rate	11	5.4	23.2

Εικόνα 2: Descriptive Statistics of Numerical variables

Numerical variable	*Non default*	*Default*
Loan status	0	1

Εικόνα 3: Target variable

Categorical variable	*Value*
Home status	Rent, mortgage, own
Loan intent	Education, medical, venture, home, improvement, personal, debt consolidation
Loan grade	A,B,C,D,E,F,G
Historical default	Y, N

Εικόνα 4: categorical variable

Μεθοδολογία

Υπάρχουν πολλοί αλλά και διαφορετικοί τρόποι για να μπορέσει να κατηγοριοποιηθεί ένα μοντέλο μηχανικής μάθησης ανάλογα με τον τύπο των δεδομένων που δίνουν στην έξοδο, τον τύπο των δεδομένων που χρησιμοποιεί το μοντέλο σαν είσοδο ή ακόμα και με το είδος της ίδιας της μάθησης. Όπως έχει αναφερθεί τρείς είναι οι κατηγορίες της μηχανικής μάθησης.

Η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη (unsupervised) και η ενισχυτική (reinforcement) μάθηση. Στην επιβλεπόμενη μάθηση, είναι απαραίτητες μια μεταβλητή εισόδου και εξόδου, ενώ ένα σύνολο εκπαίδευσης που βασίζεται σε προκαθορισμένες εισόδους και εξόδους χρησιμοποιείται για την εκπαίδευση των μοντέλων για την πρόβλεψη του σωστού αποτελέσματος στο μέλλον. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) κ.τ.λ.

Στη μη επιβλεπόμενη μάθηση ο αλγόριθμος κατασκευάζει ένα μοντέλο για ένα σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς όμως να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετίσεων (association analysis), ομαδοποίησης (clustering) κτλ.

Τέλος στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδρασή με το περιβάλλον. Χρησιμοποιείται περισσότερο σε προβλήματα σχεδιασμού (Planning) όπως για παράδειγμα η βελτιστοποίηση εργασιών σε διάφορους εργασιακούς χώρους.

Βήματα

Στην εργασία αυτή, οι αλγόριθμοι που θα χρησιμοποιηθούν είναι επιβλεπόμενης μάθησης. Τα βήματα για αυτό το πρόβλημα ταξινόμησης θα περιγράφουν παρακάτω.

Απόκτηση Δεδομένων -Data Acquisition

Το πιο σημαντικό αλλά και βασικό στοιχείο στην ανάλυση είναι τα δεδομένα που χρησιμοποιούνται. Το πλήθος αλλά και η ποιότητα καθιστούν την ανάλυση μια πρόκληση εφόσον τα αποτελέσματα και η ακρίβεια τους, επηρεάζονται έντονα από αυτά.

Προεπεξεργασία δεδομένων - Data Preprocessing

Η προετοιμασία δεδομένων αναφέρεται στον μετασχηματισμό ακατέργαστων δεδομένων (raw data) ή ακόμα και στην κωδικοποίηση προκειμένου να είναι αναγνώσιμα από τη μηχανή.

Παίζει εξαιρετικά σημαντικό ρόλο στη διαδικασία μάθησης. Ο χειρισμός χαμένων (missing values) και ακραίων τιμών (outliers), κωδικοποίηση κατηγορικών μεταβλητών (categorical variables), μείωση διαστάσεων και επιλογή χαρακτηριστικών είναι μερικές από τις απαιτήσεις. Σε αυτό το βήμα, η δημιουργία πολλών γραφημάτων είναι πολύ χρήσιμη, καθώς μπορούν να δοθούν χρήσιμες πληροφορίες.

Επιλογή μοντέλου - Model selection

Υπάρχει μια ποικιλία αλγορίθμων που χρησιμοποιούνται στη μηχανική μάθηση και καθένας από αυτούς είναι κατάλληλος για διαφορετικά προβλήματα. Το αντικείμενο είναι να προσδιοριστεί το μοντέλο που θα οδηγήσει στην υψηλότερη ακρίβεια. Ένα άλλο βασικό ερώτημα που πρέπει να απαντηθεί πριν από την εφαρμογή ενός αλγόριθμου είναι εάν το πρόβλημα αντιστοιχεί σε ταξινόμηση ή ομαδοποίηση. Σε αυτή τη μελέτη χρησιμοποιήθηκαν τεχνικές ταξινόμησης.

Εκπαίδευση – Training

Το αρχικό σύνολο δεδομένων χωρίζεται σε δεδομένα εκπαίδευσης (training) και δοκιμών (test). Το πρώτο χρησιμοποιείται για την εκπαίδευση ενός μοντέλου, ενώ το δεύτερο χρησιμοποιείται για να αξιολογήσει πόσο ακριβής είναι η πρόβλεψη εξόδου.

Αξιολόγηση - Evaluation

Αυτό το βήμα είναι πολύ κρίσιμο, καθώς απεικονίζει την απόδοση του μοντέλου σε νέες περιπτώσεις που δεν ήταν μέρος της διαδικασίας εκπαίδευσης. Για αυτό το εύρος, χρησιμοποιούνται διάφορες μετρήσεις αξιολόγησης σύμφωνα με την κατηγορία μηχανικής εκμάθησης που χρησιμοποιείται κάθε φορά.

Συντονισμός υπερπαραμέτρων - Hyperparameter Tuning
Αυτή η διαδικασία αναφέρεται στον προσδιορισμό της διαμόρφωσης των υπερπαραμέτρων που οδηγεί στην καλύτερη απόδοση. Εκτός από τις παραμέτρους του μοντέλου, υπάρχει ένα άλλο είδος, που ονομάζεται «υπερπαράμετροι» (hyperparameters), που δεν είναι δυνατό να μαθευτεί κατά τη διάρκεια της εκπαιδευτικής διαδικασίας και περιέχουν χρήσιμες πληροφορίες σχετικά με την πολυπλοκότητα του μοντέλου ή πόσο γρήγορη είναι η ικανότητά του να μαθαίνει. Για παράδειγμα, μια υπερπαράμετρος είναι το k στον αλγόριθμο KNN.

Πρόβλεψη - Prediction
Μετά από όλες τις προαναφερθείσες διαδικασίες, το εκπαιδευμένο μοντέλο είναι τώρα έτοιμο να προβλέψει την πιο πιθανή τιμή εξόδου δεδομένης μιας συγκεκριμένης εισόδου χρησιμοποιώντας δεδομένα σε πραγματικό χρόνο.

Python και βιβλιοθήκη scikit-learn
Στην παρούσα εργασία, η ανάλυση δεδομένων και η ανάπτυξη του μοντέλου γίνονται χρησιμοποιώντας την γλώσσα προγραμματισμού Python και τη βιβλιοθήκη scikit-learn για τα μοντέλα. Η μηχανική μάθηση συνεχίζει να αναπτύσσεται με ταχείς ρυθμούς. Η τεχνητή νοημοσύνη καθιστά δυνατή τη δημιουργία καινοτόμων λύσεων σε πραγματικά προβλήματα, όπως ανίχνευση απάτης (fraud detection), προσωπικούς βοηθούς (personal assistants), φίλτρα ανεπιθύμητης αλληλογραφίας (spam filters). Η ανάγκη για έξυπνες λύσεις σε προβλήματα του πραγματικού κόσμου οδηγεί στην περαιτέρω ανάπτυξη της τεχνητής νοημοσύνης προκειμένου να αυτοματοποιηθούν εργασίες που είναι δύσκολες στον προγραμματισμό χωρίς AI. Η γλώσσα προγραμματισμού Python θεωρείται ο καλύτερος αλγόριθμος για την αυτοματοποίηση τέτοιων εργασιών και προσφέρει μεγαλύτερη απλότητα.

Η βιβλιοθήκη Scikit-learn είναι δωρεάν βιβλιοθήκη μηχανικής εκμάθησης της python που μπορεί να χρησιμοποιηθεί για να απλοποιήσει την εργασία κωδικοποίησης και υλοποίησης αλγορίθμων Μηχανικής Μάθησης. Η Scikit-learn περιλαμβάνει έναν αριθμό διαφορετικών αλγορίθμων μηχανικής μάθησης, όπως Random Forest, SVM και KNN. Είναι μια συλλογή από τα πιο αποτελεσματικά εργαλεία για στατιστική μοντελοποίηση και μηχανική μάθηση. Μερικά από αυτά τα εργαλεία περιλαμβάνουν παλινδρόμηση, ταξινόμηση, ομαδοποίηση, μείωση διαστάσεων. Είναι κυρίως γραμμένο σε Python και βασίζεται σε βιβλιοθήκες SciPy, NumPy και Matplotlib. Αναπτύχθηκε από τον David Cournapeau το 2007 ως μέρος του Google Summer Code. Στη συνέχεια, οι Gael Varoquaux, Fabian Pedregosa, Alexandre Gramfort και Vincent Michel, από το Γαλλικό Ινστιτούτο Έρευνας Επιστήμης Υπολογιστών και Αυτοματισμού, κυκλοφόρησαν μια beta έκδοση του v0.1 το 2010. Από τότε, έχουν κυκλοφορήσει νεότερες εκδόσεις. Η Scikit-learn είναι ένα έργο που βασίζεται στην κοινότητα όπου ο καθένας μπορεί να συνεισφέρει στην ανάπτυξή του. Τα βασικά πλεονεκτήματα που έχει η βιβλιοθήκη είναι τα εξής:

1. Είναι εύκολη στη χρήση.
2. Είναι πολύ ευέλικτη και εξυπηρετεί πραγματικούς σκοπούς όπως η πρόβλεψη της συμπεριφοράς των καταναλωτών.
3. Υποστηρίζεται και ενημερώνεται από πολλούς συνεργάτες στη διεθνή διαδικτυακή κοινότητα.

Προεπεξεργασία δεδομένων -Data preprocessing

Ως πρώτο βήμα της ανάλυσης, είναι να εκλεχθεί το σύνολο δεδομένων εάν έχει τιμές που λείπουν (missing values). Παρατηρείται πως στο σύνολο δεδομένων υπάρχουν δύο πεδία με μηδενικές τιμές. Η διάρκεια απασχόλησης (Employment length) περιέχει 895 μηδενικές τιμές και το επιτόκιο (Interest rate) περιέχει 3116 μηδενικές τιμές.

```
loan_data.isnull().sum()
```

```
person_age                      0
person_income                   0
person_home_ownership           0
person_emp_length             895
loan_intent                     0
loan_grade                      0
loan_amnt                       0
loan_int_rate                3116
loan_status                     0
loan_percent_income             0
cb_person_default_on_file       0
cb_person_cred_hist_length      0
dtype: int64
```

Εικόνα 5: Missing values

Αφού προσδιορίσουμε τις τιμές που λείπουν που περιέχονται στο σύνολο δεδομένων, τις αντικαθιστούμε με τη μέση τιμή κάθε μεταβλητής.

Διερευνητική Ανάλυση Δεδομένων

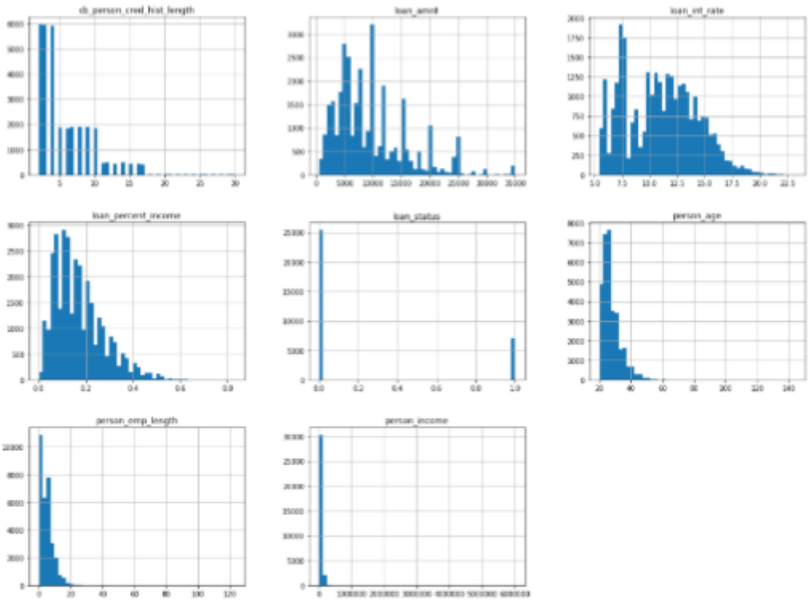
Η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis -EDA) είναι το μέρος της Ανάλυσης Δεδομένων που εξερευνά ακατέργαστα ή δομημένα δεδομένα και τα οπτικοποιεί μέσω πολλαπλών ειδών γραφημάτων, γραφημάτων και τεχνικών. Είναι σημαντικό σε ένα πραγματικό έργο μηχανικής μάθησης ή ανάλυσης δεδομένων να διερευνηθούν τα χαρακτηριστικά και οι πληροφορίες που δίνει ένα σύνολο δεδομένων, με απώτερο στόχο τη γνώση. Τη γνώση των χαρακτηριστικών, της μορφής, του τύπου, της κατανομής, των συσχετίσεων και των βασικών στατιστικών στοιχείων των δεδομένων από ένας αναλυτής.

Αριθμητικά δεδομένα -Numerical Data

Για τα αριθμητικά δεδομένα του συνόλου δεδομένων δημιουργούνται τα ιστογράμματα. Το ιστόγραμμα είναι ένα γράφημα που εμφανίζει αριθμητικά δεδομένα σε ομάδες ράβδων με διαφορετικό ύψος. Η ομαδοποίηση βασίζεται στις διαφορετικές τιμές της μεταβλητής – το χαρακτηριστικό έχει σχεδιαστεί και το εύρος του. Εκτός από το ότι είναι μια μέθοδος για τη γραφική παράσταση των τιμών μιας μεταβλητής, τα ιστογράμματα βοηθούν στην απόκτηση πληροφοριών σχετικά με την κατανομή της, καθώς και τη συχνότητα πολλών ακραίων τιμών.

Παρακάτω εμφανίζονται τα ιστογράμματα των αριθμητικών των παρακάτω χαρακτηριστικών: person age, person income, person emp length, loan amnt, loan int rate, loan status, loanpercent income και cb person cred hist length:

```
loan_data.hist(bins=50, figsize=(20,15))
plt.show()
```

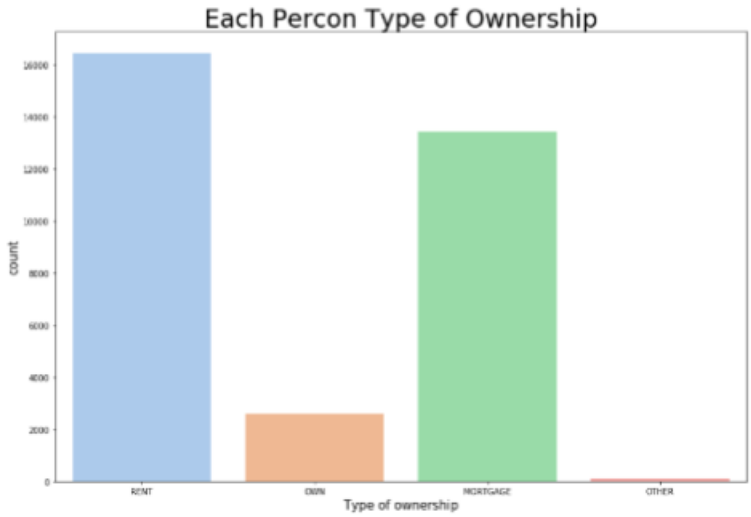


Εικόνα 6: Ιστόγραμμα με αριθμητικά δεδομένα

Κατηγορικά Στοιχεία -Categorical Data

Μετά την ανάλυση των κατηγορικών δεδομένων, δημιουργούμε γραφήματα που έχουν παρόμοια σημασία και πρακτική εφαρμογή με τα ιστογράμματα για αριθμητικά χαρακτηριστικά. Αυτά τα γραφήματα ονομάζονται γραφήματα καταμέτρησης και είναι γραφήματα ράβδων. Βασικά, μετρούν τις τιμές των κατηγοριών για κάθε μία από αυτές, έτσι και πάλι το ύψος μιας ράβδου αποκλίνει όπως στο ιστόγραμμα. Κάθε γραμμή αντιπροσωπεύει μια κατηγορία της κατηγοριοποιημένης μεταβλητής.

Παρακάτω παρουσιάζεται ένα γράφημα καταμέτρησης της κατηγορικής μεταβλητής person home ownership, όπου είναι προφανές ότι οι δύο κύριοι τύποι πελατών που ζητούν δάνειο, έχουν ενοίκιο ή υποθήκη.

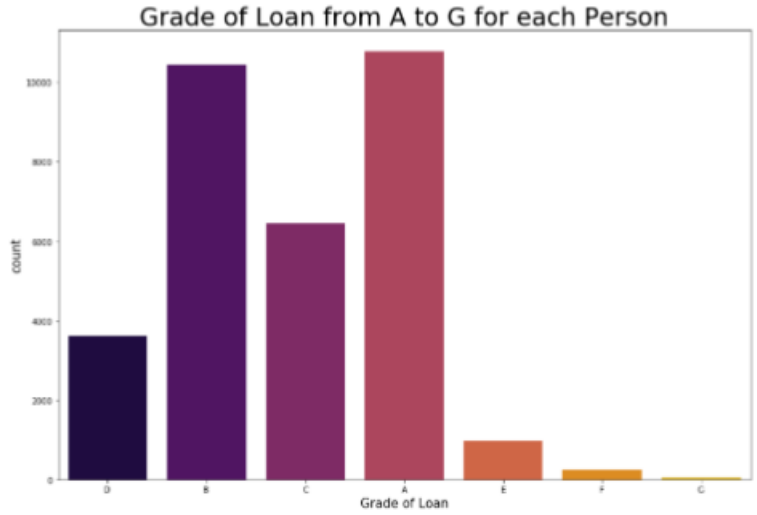


Εικόνα 7: Count plot -Bar graph: Τύπος της μεταβλητής ownership

Επιπλέον, το παρακάτω γράφημα είναι μια γραφική παράσταση μέτρησης της κατηγορικής μεταβλητής loan intent, η οποία αποκαλύπτει ότι οι πελάτες ζητούν δάνειο κυρίως για εκπαιδευτικούς και ιατρικούς σκοπούς, ενώ οι υπόλοιποι λόγοι ακολουθούν χωρίς τόσο μεγάλη διαφορά.

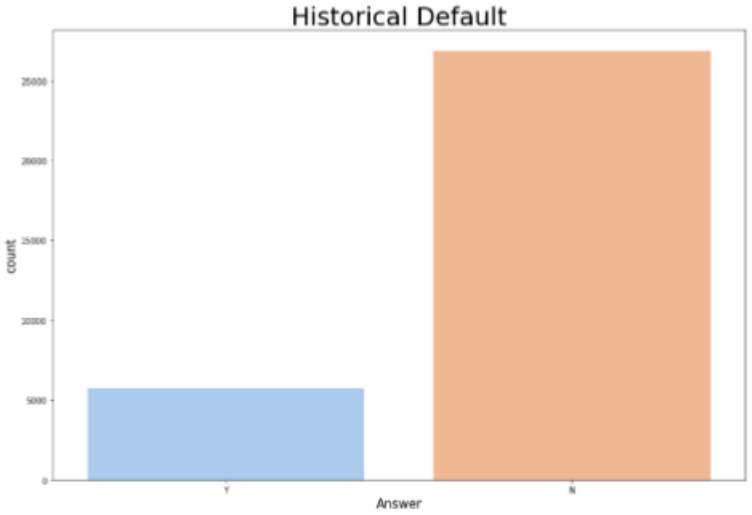
Εικόνα 8: Count plot -Bar graph: Τύποι της μεταβλητής intention

Επιπλέον, ο βαθμός του δανείου ανά άτομο (δάνειο βαθμός) εμφανίζεται με το παρακάτω γράφημα καταμέτρησης, στο οποίο οι πελάτες με βαθμούς Β και Δ είναι μεγαλύτεροι από 20.000 των συνόλων.



Εικόνα 9: Count plot -Bar graph: Βαθμός Δανείων

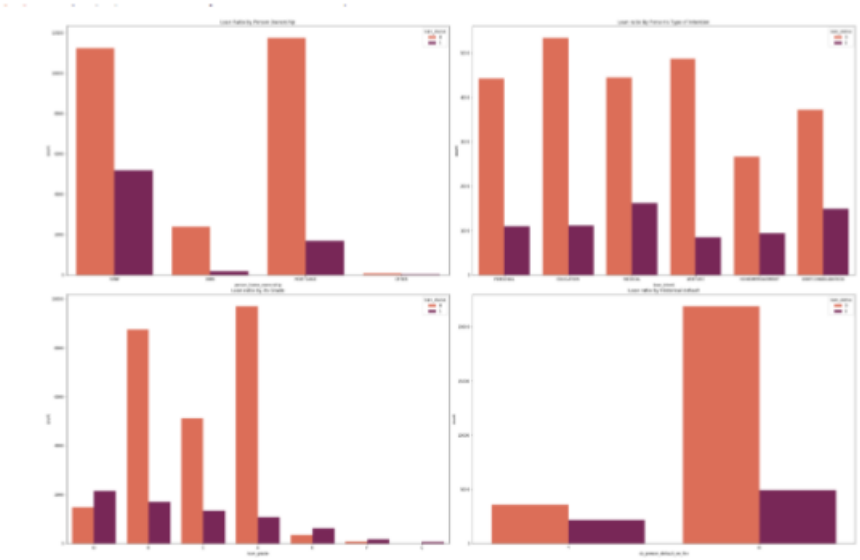
Μια γραφική παράσταση μέτρησης της κατηγορικής μεταβλητής cb person default on file παρουσιάζεται επίσης παρακάτω.



Εικόνα 10: Count plot -Bar graph: Historical default

Μεικτά Γραφήματα -Mixed Plots

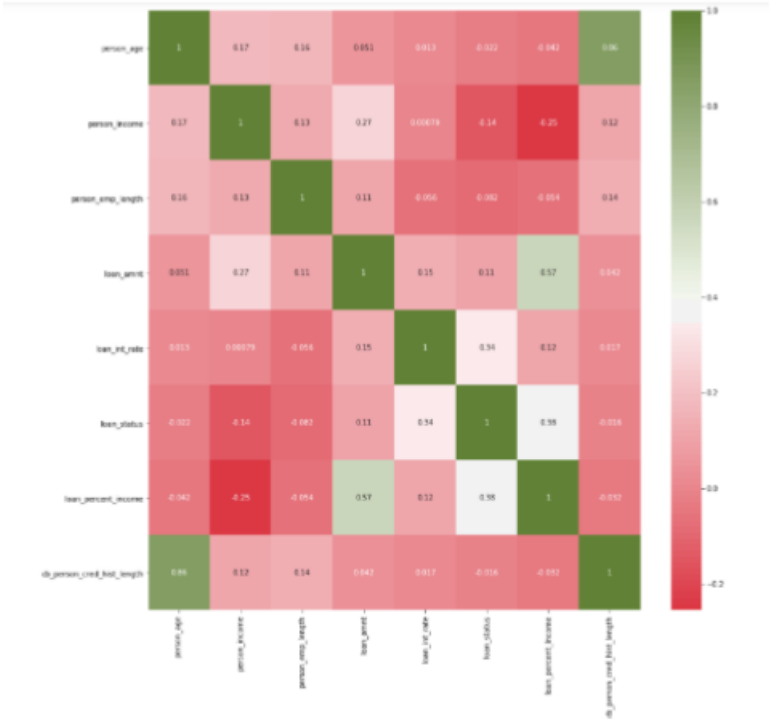
Τα παρακάτω γραφήματα είναι συγκριτικά γραφήματα ράβδων (όπως γραφήματα καταμέτρησης) των κατηγορικών δεδομένων, αλλά αυτή τη φορά συγκρίνονται με τις 2 κατηγορίες (0-1) του στόχου.



Εικόνα 11: Bar graph -Comparison of class variable values in categorical features

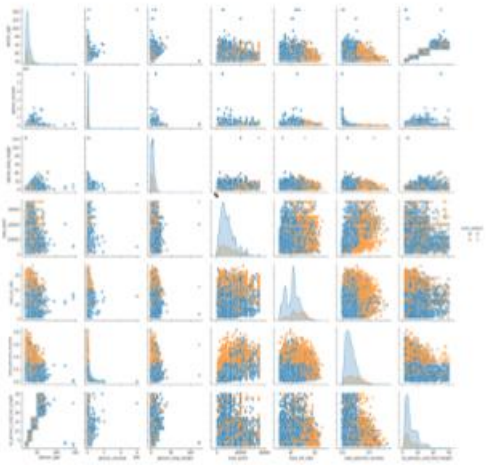
Γραφήματα συσχέτισης -Correlation Plots

Για να δούμε αν οι μεταβλητές στο σύνολο δεδομένων συσχετίζονται μεταξύ τους, είναι απαραίτητη η δημιουργία δύο γραφημάτων συσχέτισης. Το πρώτο αναφέρεται σε έναν πίνακα συσχέτισης που έχει δημιουργηθεί σε ένα heat map γράφημα. Το δεύτερο είναι ένα διάγραμμα ζεύγους, το οποίο είναι ένα πολύ χρήσιμο γράφημα για την παρατήρηση της κατανομής κάθε μεταβλητής στη διαγώνιο και των συσχετισμών σε όλα τα άλλα κελιά, και τα δύο σε σύγκριση με τις 2 ομάδες.



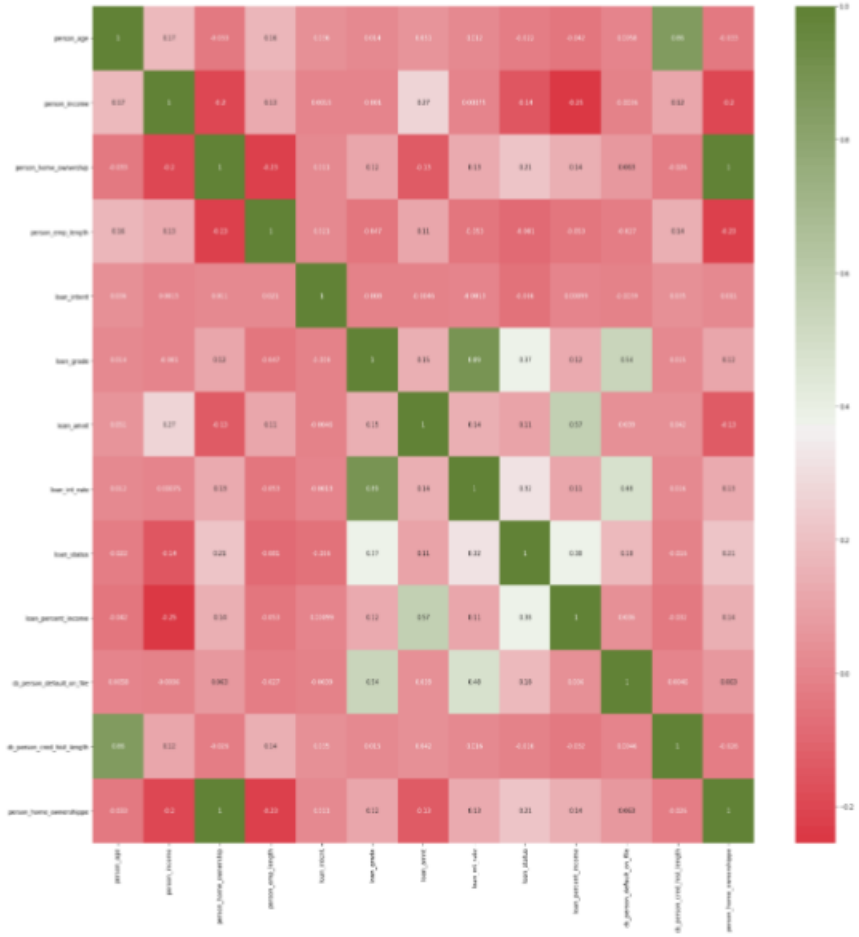
Εικόνα 12: Correlation matrix (Heatmap plot)

Κάθε τετράγωνο δείχνει τη συσχέτιση μεταξύ των μεταβλητών σε κάθε άξονα. Η συσχέτιση κυμαίνεται από -1 έως +1. Οι τιμές πιο κοντά στο μηδέν σημαίνει ότι δεν υπάρχει γραμμική τάση μεταξύ των 2 μεταβλητών. Όσο πιο κοντά στο 1 είναι η συσχέτιση, τόσο πιο θετικά συσχετίζονται και τόσο ισχυρότερη είναι η σχέση τους. Αυτό σημαίνει ότι αν το ένα αυξηθεί, θα αυξηθεί και το άλλο. Μια συσχέτιση πιο κοντά στο -1 είναι παρόμοια, αλλά αντί για τις δύο μεταβλητές, εάν η μία μεταβλητή αυξηθεί, η άλλη θα μειωθεί. Οι διαγώνιοι είναι όλες ίσες με 1 επειδή αυτά τα τετράγωνα συσχετίζουν κάθε μεταβλητή με τον εαυτό της. Η συσχέτιση μεταξύ δύο μεταβλητών είναι μεγαλύτερη όταν ο αριθμός στο τετράγωνο είναι μεγαλύτερος και το χρώμα είναι πιο σκούρο. Η γραφική παράσταση είναι επίσης συμμετρική ως προς τη διαγώνιο αφού οι ίδιες δύο μεταβλητές συνδυάζονται μαζί σε αυτά τα τετράγωνα. Έτσι, ο πίνακας συσχέτισης δείχνει ότι δεν υπάρχει καμία σημαντική συσχέτιση μεταξύ των μεταβλητών. Υπάρχει χαμηλή συσχέτιση μεταξύ του ποσού του δανείου και του προσωπικού εισοδήματος, κάτι που δεν προκαλεί έκπληξη.



Εικόνα 13: Pair plot graph

Ο πιο κρίσιμος παράγοντας που επηρεάζει την επιτυχία της διαδικασίας μηχανικής μάθησης είναι η διαδικασία εκπαίδευσης (training) και δοκιμής (testing). Μια αποτελεσματική διαδικασία εκπαίδευσης βελτιώνει την ποιότητα του μοντέλου που δημιουργείται. Τα σύνολα δεδομένων χωρίζονται σε δύο μέρη για εκπαίδευση αλλά και δοκιμή σύμφωνα με ορισμένους κανόνες. Αφού το μοντέλο μηχανικής εκμάθησης εκπαιδευτεί με βάση τα δεδομένα εκπαίδευσης, πρέπει να ελεγχθεί χρησιμοποιώντας τα δεδομένα δοκιμής. Να επισημανθεί πως στην περίπτωση που εξετάζεται εδώ, πριν από τη διαδικασία εκπαίδευσης, οι κατηγορικές μεταβλητές πρέπει να μετατραπούν, σε αριθμητικές, και αυτό μπορεί να γίνει με τη χρήση της μεθόδου κωδικοποίησης ετικετών (label encoding method). Η κωδικοποίηση ετικετών (label encoding method), είναι μια τεχνική κωδικοποίησης μηχανικής μάθησης που επιτρέπει να μετατρέπονται δεδομένα κατηγοριών και κειμένου σε αριθμητικά, ώστε να μπορούν να χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης. Ο κωδικοποιητής ετικετών απλώς κωδικοποιεί κατηγορικά δεδομένα σε αριθμητικά για καθεμία από αυτές τις τιμές από τις στήλες. Η Python και η scikit-learn έχουν πολλούς εύκολους τρόπους να το κάνουν αυτό αυτόματα σε ένα σύνολο δεδομένων. Ένας άλλος διάσημος κωδικοποιητής είναι ο One-Hot-Encoder. Μετά την τελευταία μετατροπή που πραγματοποιείται, δημιουργείται η τελική μορφή του συνόλου δεδομένων προς χρήση, η οποία περιγράφεται από τα παρακάτω γραφήματα.



Εικόνα 3.14: Final Heatmap (Correlation Matrix)

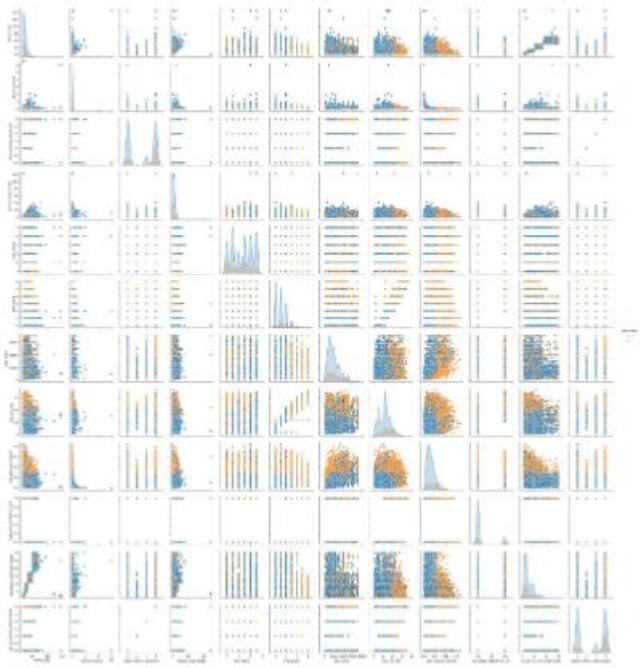
Εικόνα 14:Correlation matrix (Final plot)

Από τον παραπάνω γράφημα συσχέτισης που αντιστοιχεί στο τελικό σύνολο δεδομένων, προκύπτουν συμπεράσματα σχετικά με τις συσχετίσεις μεταξύ των μεταβλητών και των πληροφοριών που λαμβάνονται από αυτές. Μερικά από αυτά τα συμπεράσματα περιγράφονται παρακάτω:

1. Στη διαγώνιο του γραφήματος οι μόνες τιμές που παίρνουμε είναι ίσες με 1. Αυτό συμβαίνει επειδή η συσχέτιση μεταξύ μιας μεταβλητής και της ίδιας (2 φορές) είναι 100.
2. Στα τετράγωνα με ανοιχτό πράσινο χρώμα υπάρχει πολύ καλή σχέση και συσχέτιση μεταξύ των δύο χαρακτηριστικών. Για παράδειγμα, το loan amnt και το loan percent income έχουν υψηλή συσχέτιση ίση με 0,57, που σημαίνει ότι παρέχονται καλές πληροφορίες σχετικά με τα δεδομένα μεταξύ αυτών το δύο μεταβλητών. Ένα άλλο παράδειγμα ίδιου ζεύγους μεταβλητών είναι το loan int rate με loan grade. Υπάρχουν πολλά άλλα ζεύγη καλών συσχετίσεων όπως φαίνεται στο γράφημα.

3. Στα τετράγωνα με τιμές μικρότερες από 0, δεν μπορεί να προκύψει ένα χρήσιμο αποτέλεσμα. Αρνητική συσχέτιση σημαίνει ότι η συσχέτιση μεταξύ των δύο μεταβλητών είναι κακή και δεν υπάρχουν πληροφορίες που να προκύπτουν από αυτές. Για παράδειγμα, το person income με loan percent income έχει τιμή ίση με -0,25.
4. Τέλος, για τα ανοιχτό ροζ τετράγωνα μπορεί να επισημανθεί ότι δεν είναι κακά, αλλά δεν είναι αρκετά καλά για να βγάλουμε ένα συμπέρασμα σχετικά με τις πληροφορίες που μπορεί να δοθούν από αυτές τις μεταβλητές.

Η γραφική παράσταση ζεύγους που αντιστοιχεί στο τελικό σύνολο δεδομένων φαίνεται παρακάτω.



Εικόνα 15: Final Pair plot graph

Τελευταίο αλλά εξίσου σημαντικό, πριν από τη δημιουργία και την αξιολόγηση των μοντέλων μηχανικής μάθησης, το σύνολο δεδομένων χωρίζεται σε δύο μέρη, το σύνολο εκπαίδευσης που περιέχει το 80% του αρχικού συνόλου δεδομένων και το σύνολο δοκιμής που περιέχει το υπόλοιπο 20% των παρατηρήσεων. Χρησιμοποιώντας τη συνάρτηση stratify (stratify = y) για την

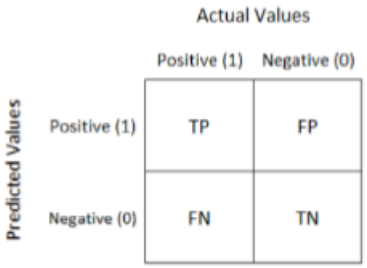
εκπαίδευση, εκμεταλλευόμαστε τη χρήση της διατήρησης του ίδιου ποσοστού για κάθε ομάδα της μεταβλητής κλάσης στα σετ εκπαίδευσης και δοκιμής. Ως αποτέλεσμα, είναι εγγυημένο ότι το ποσοστό των τιμών από τις δύο τάξεις (0-1) της ομάδας-στόχου είναι το ίδιο τόσο στο σύνολο εκπαίδευσης και δοκιμής.

Σε αυτό το κεφάλαιο, αναλύονται τα αποτελέσματα των μοντέλων μηχανικής εκμάθησης που έχουν δημιουργηθεί. Για να γίνει αυτό, αρχικά είναι απαραίτητο να περιγραφούν συνοπτικά οι μετρήσεις αξιολόγησης στις οποίες βασίζεται η ανάλυση.

Το Confusion Matrix είναι μακράν το πιο σημαντικό εργαλείο για την αξιολόγηση μοντέλων μηχανικής μάθησης ταξινόμησης. Είναι ένας πίνακας που αντιπροσωπεύει τις πραγματικές τιμές (τις πραγματικές τιμές του συνόλου δεδομένων) στον άξονα y σε σύγκριση με τις προβλεπόμενες τιμές (τις τιμές που είχαν προβλεφθεί από τα μοντέλα) στον άξονα x. Τόσο οι πραγματικές όσο και οι προβλεπόμενες τιμές διαχωρίζονται σε θετική και αρνητική κατηγορία (1-0). Το νόημα είναι ότι η μεταβλητή κλάσης είναι δυαδική και παίρνει δύο τιμές, 0 για αρνητική κλάση δεδομένων και 1 για θετική κλάση δεδομένων. Ως αποτέλεσμα, ο πίνακας που δημιουργείται έχει τέσσερα τετράγωνα και περιέχει τέσσερις ομάδες τιμών. Αυτές οι τιμές είναι οι ακόλουθες:

1. TP (True positive): Η πραγματική τιμή και η προβλεπόμενη τιμή είναι θετικές (1-1), που σημαίνει ότι η πρόβλεψη του μοντέλου ήταν σωστή (σωστή ταξινόμηση).
2. FP (False Positive): Η πραγματική τιμή είναι αρνητική, αλλά η προβλεπόμενη τιμή είναι θετική (0-1), που σημαίνει ότι η πρόβλεψη του μοντέλου ήταν λάθος (λανθασμένη ταξι- νόμηση).
3. FN (False Negative): Η πραγματική τιμή είναι θετική, αλλά η προβλεπόμενη τιμή είναι αρνητική (1-0), που σημαίνει ότι η πρόβλεψη του μοντέλου ήταν λάθος.
4. TN (True Negative): Η πραγματική τιμή και η προβλεπόμενη τιμή είναι αρνητικές (0-0), που σημαίνει ότι η πρόβλεψη του μοντέλου ήταν σωστή.

Οι τιμές TP, FP, FN και TN μπορούν να γραφτούν είτε ως ένας συγκεκριμένος αριθμός όλων των εισόδων του συνόλου δοκιμής, για παράδειγμα ο αριθμός των προβλέψεων TP στο σύνολο δοκιμής, είτε ως ποσοστό κάθε τετραγώνου.



Εικόνα 16: Confusion Matrix

Ακρίβεια -Precision

Ακρίβεια (Precision) : Είναι το ποσοστό των αληθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} - \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Ανάκληση -Recall

Ανάκληση (Recall): Το ποσοστό των περιπτώσεων που ταξινομούνται ως μια δεδομένη τάξη διαιρούμενο με τον πραγματικό συνολικό σε αυτό τον κλάδο (ισοδύναμο με ποσοστό TP).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} - \frac{True\ Positive}{Total\ Actual\ Positive}$$

Μέσος όρος -F-Measure

Μέσος όρος (F-Measure): Είναι ένα μέτρο απόδοσης που δεν λαμβάνει υπόψη του την απόδοση των αρνητικών κλάσεων.

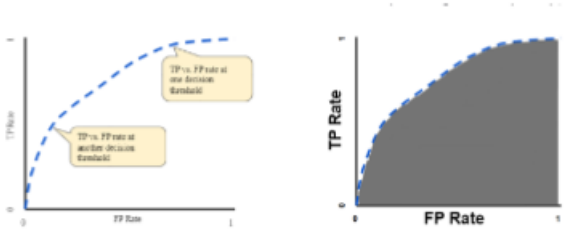
$$F - Measure = \frac{2 + Recall + Precision}{Recall + Precision}$$

Σταθμισμένος μέσος όρος -Weighted Avg.

Σταθμισμένος μέσος όρος (Weighted Avg.): Είναι ένας υπολογισμός που λαμβάνει υπόψη τους διαφορετικούς βαθμούς σπουδαιότητας των αριθμών σε ένα σύνολο δεδομένων. Κατά τον υπολογισμό ενός σταθμισμένου μέσου όρου, κάθε αριθμός στο σύνολο δεδομένων πολλαπλασιάζεται με ένα προκαθορισμένο βάρος πριν από τον τελικό υπολογισμό.

ROC (Area Under the Curve-AUC)

Ένα άλλο δημοφιλές στατιστικό εργαλείο είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curves – ROC curves), οι οποίες εξ ορισμού χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός συστήματος με διχοτομικά εξαγόμενα αποτελέσματα. Παραδοσιακά, το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve-AUC) χρησιμοποιείται ως συνοπτικός δείκτης ακρίβειας ενός τεστ, και είναι χρήσιμο ως περιγραφικό μέτρο της συνολικής απόδοσης ενός τεστ.



Εικόνα 17: ROC curve & ROC -AUC

Με βάση τις μετρήσεις αξιολόγησης που χρησιμοποιήθηκαν σε αυτή την εργασία, το μοντέλο Random Forest αποδίδει καλύτερα από τα άλλα στο δεδομένο μη ισορροπημένο σύνολο δεδομένων. Στην περίπτωση που εξετάζεται, η ανάκληση (recall) είναι η πιο σημαντική μέτρηση που πρέπει να ληφθεί υπόψη, δεδομένου ότι το πρόβλημα περιέχει ψευδείς αρνητικές τιμές, που σημαίνει ότι το μοντέλο προβλέπει ότι κάποιος δεν πρόκειται να προβεί σε προεπιλογή, αλλά το κάνει. Τούτου λεχθέντος, το δέντρο αποφάσεων έχει καλύτερη απόδοση σε αυτήν τη μέτρηση, αλλά το Random Forest έχει συνολικά καλύτερη απόδοση σε όλες τις μετρήσεις.

Συντονισμός Υπερπαραμέτρων -Hyperparameter tuning

Όταν δημιουργείτε ένα μοντέλο μηχανικής μάθησης, παρουσιάζονται επιλογές σχεδίασης ως προς τον τρόπο ορισμού της αρχιτεκτονικής του μοντέλου που θα χρησιμοποιηθεί. Πολλές φορές, δεν γνωρίζεται εξ αρχής ποια θα πρέπει να είναι η βέλτιστη αρχιτεκτονική μοντέλου για ένα δεδομένο μοντέλο, και επομένως καλό θα ήταν να ερευνηθεί μια σειρά από δυνατότητες.
Ιδανικά το μηχάνημα θα εξερευνούσε και θα επέλεγε τη βέλτιστη αρχιτεκτονική μοντέλου αυτόματα. Οι παράμετροι που καθορίζουν την αρχιτεκτονική του μοντέλου αναφέρονται ως υπερπαράμετροι (hyperparameters) και έτσι αυτή η διαδικασία αναζήτησης της ιδανικής αρχιτεκτονικής μοντέλου αναφέρεται ως συντονισμός υπερπαραμέτρων (hyperparameter tuning). Πιο απλά
ο καλύτερος τρόπος για να σκεφτείς τις υπερπαραμέτρους είναι σαν τις ρυθμίσεις ενός αλγορίθμου που μπορεί να προσαρμοστεί για να βελτιστοποιήσει την απόδοση. Ενώ οι παράμετροι του μοντέλου μαθαίνουν κατά τη διάρκεια της εκπαίδευσης, οι υπερπαράμετροι πρέπει να ορίζονται από τον χρήστη πριν από την εκπαίδευση. Για τη βελτίωση του αλγόριθμου KNN, η αναζήτηση πλέγματος απόδοσης, είναι η πιο κοινή προσέγγιση. Ψάχνει εξαντλητικά όλους τους πιθανούς συνδυασμούς υπερπαραμέτρων κατά την φάση της εκπαίδευσης. Για παράδειγμα, καθορίζεται ένα πλέγμα από αριθμού γειτόνων (K=1,2,3) και δύο μητρικές (p=1,2). Η αναζήτηση πλέγματος ξεκινά να εκπαιδεύει το μοντέλο K=1 και p=1 και υπολογίζει το σκορ ακρίβειας. Στην συνέχεια, συνεχίζει να εκπαιδεύει τα μοντέλα για (K=2, p=1), (K=3, p=1), (K=1,p=2),... και (K=3, p=2) και λαμβάνει τα αντίστοιχα σκορ. Με βάση τα σκορ ακρίβειας, η αναζήτηση πλέγματος θα κατατάξει τα μοντέλα και θα

καθορίσει το σύνολο των τιμών των υπερπαραμέτρων που δίνουν την υψηλότερη βαθμολογία ακρίβειας. Χρησιμοποιώντας την αναζήτηση πλέγματος (grid search), παρατηρείται πως ο καλύτερος αριθμός γειτόνων είναι 13, ενώ η βέλτιστη απόσταση είναι η ευκλείδεια και p=2. Για τον ταξινομητή του Random Forest, οι δύο κύριοι υπεπαράμετροι είναι:

1. n estimators:

Η παράμετρος n estimators ελέγχει τον αριθμό των δέντρων μέσα στον ταξινομητή. Να σημειωθεί πως χρησιμοποιώντας πολλά δέντρα για τη δημιουργία του μοντέλου δεν είναι πάντοτε η καλύτερη περίπτωση. Μπορεί να αυξήσει τη χρονική πολυπλοκότητα του μοντέλου. Ο προεπιλεγμένος αριθμός των εκτιμήσεων είναι 100 στη scikit-learn.

2. max depth:

Καθορίζει το μέγιστο ύψος μέχρι το οποίο μπορούν να αναπτυχθούν τα δέντρα μέσα στο δάσος. Να υπογραμμιστεί πως είναι η πιο σημαντική υπερπαράμετρος όσον αφορά την αύξηση της ακρίβεια τ μοντέλου. Η προεπιλεγμένη τιμή είναι κενό (none).
Στο μοντέλο Random Forest, οι προεπιλεγμένες παράμετροι αποδίδουν καλύτερα από οποιαδήποτε άλλη περίπτωση.

Για τον ταξινομητή του Decision Trees, οι κύριοι υπεπαράμετροι είναι:

1. Criterion: Η συνάρτηση για τη μέτρηση της ποιότητας ενός διαχωρισμού. Υποστηριζόμενα κριτήρια είναι το “gini” και η “εντροπία” για το κέρδος πληροφοριών.
2. max depth: Το μέγιστο βάθος του δέντρου. Αν Κανένας (none), τότε οι κόμβοι επεκτείνονται μέχρι να γίνουν όλα τα φύλλα καθαρά ή έως ότου όλα τα φύλλα περιέχουν λιγότερα από min samples split δείγματα.
3. max features: Ο αριθμός των χαρακτηριστικών που πρέπει να ληφθούν υπόψη κατά την αναζήτηση του καλύτερου διαχωρισμού.
4. min samples leaf: Ο ελάχιστος αριθμός δειγμάτων που απαιτείται να βρίσκονται σε έναν κόμβο φύλλου. Ένα σημείο διαίρεσης σε οποιοδήποτε βάθος θα λαμβάνεται υπόψη μόνο εάν αφήνει τουλάχιστον ελάχιστα δείγματα εκπαίδευσης φύλλων σε κάθε έναν από τους αριστερούς και δεξιούς κλάδους. Αυτό μπορεί να έχει ως αποτέλεσμα την εξομάλυνση του μοντέλου.

Synthetic Minority Over-Sampling Technique

Είναι γνωστό πως προσπαθώντας ποτέ να πραγματοποιηθεί οποιοδήποτε έργο ταξινόμησης, υπάρχει μεγάλη πιθανότητα να εμφανιστούν μη ισορροπημένα δεδομένα. Τα μη ισορροπημένα δεδομένα παρουσιάζονται όταν μία (ή περισσότερες) από τις κλάσεις που πρέπει να επισημανθούν έχουν πολλές περισσότερες περιπτώσεις από την άλλη. Για παράδειγμα, στην περίπτωση εντοπισμού απάτης, οι περισσότερες συναλλαγές δεν είναι απάτη, επομένως το τυπικό σύνολο δεδομένων συναλλαγών θα έχει πολλές περιπτώσεις νόμιμων συναλλαγών και μόνο λίγες δόλιες.
Χωρίς να γίνει κάτι για να εξισορροπηθούν τα δεδομένα, οποιοσδήποτε

ταξινομητής χρησιμοποιηθεί πιθανότατα θα χαρακτηρίζει όλες τις συναλλαγές ως νόμιμες. Για να εξισορροπηθούν τα δεδομένα, υπάρχουν πολλές επιλογές. Αρχικά μια λύση είναι η συλλογή περισσότερων δεδομένων . Αν και αυτό είναι πάντα προτιμότερο, συχνά δεν είναι δυνατό. Σε αυτήν την περίπτωση, μπορεί να γίνει εκ νέου δειγματοληψία των δεδομένων, είτε υποδειγματίζοντας την πλειοψηφική κατηγορία (συναλλαγές χωρίς απάτη στο παραπάνω παράδειγμα) είτε υπερδειγματοληψία της κατηγορίας με την μειοψηφία (τις δόλιες συναλλαγές). Η υπερδειγματοληψία συνίσταται είτε στη δειγματοληψία κάθε μέλους της κατηγορίας μειοψηφίας με αντικατάσταση είτε στη δημιουργία συνθετικών μελών με τυχαία δειγματοληψία από το σύνολο χαρακτηριστικών. Αυτό λοιπόν είναι που κάνει το SMOTE. Η τεχνική Synthetic Minority Over-Sampling Technique (SMOTE) ασχολείται στην επίλυση του class imbalance problem. Το βασικό χαρακτηριστικό της SMOTE είναι ότι στην κλάση μειοψηφίας γίνεται oversampling δημιουργώντας «συνθετικά» δείγματα και όχι αναπαράγοντας ήδη υπάρχοντα. Στην κλάση μειοψηφίας γίνεται over-sampling παίρνοντας κάθε δείγμα από την κλάση και δημιουργώντας συνθετικά δείγματα, τα οποία χρησιμοποιούν τους k Nearest Neighbors των δειγμάτων της κλάσης. Η αρχική υλοποίηση της SMOTE χρησιμοποιεί μόνο 5 Nearest Neighbors. Αν το ποσοστό over-sampling που πρέπει να γίνει στην κλάση είναι για παράδειγμα 200%, τότε θα χρησιμοποιηθούν 2 Nearest Neighbors. Η δημιουργία των σύνθετων δειγμάτων γίνεται με την ακόλουθη διαδικασία: Αρχικά υπολογίζεται η διαφορά ανάμεσα στο δείγμα που μας ενδιαφέρει και τον Nearest Neighbor αυτού. Στη συνέχεια πολλαπλασιάζεται αυτή η διαφορά με έναν τυχαίο αριθμό μεταξύ του 0 και το αποτέλεσμα της οποίας προστίθεται στο διάνυσμα το οποίο εξετάζουμε.

Για να γίνει κατανοητό πώς λειτουργεί αυτή η μέθοδος, δίνεται ένα παράδειγμα. Στόχος είναι να ταξινομηθεί σε δύο ομάδες ανθρώπων με βάση το πού βρίσκονται σε ένα δωμάτιο, ένα σύνολο ανθρώπων. Τα άτομα με πράσινα πουκάμισα τείνουν να στέκονται στο νότιο τμήμα του δωματίου, αλλά υπάρχουν πολλά από αυτά (ας πούμε 1000), επομένως γεμίζουν τον νότιο τοίχο μέχρι τη μέση του δωματίου. Τα άτομα με κίτρινα πουκάμισα τείνουν να στέκονται στη μέση και προς τη βορειοδυτική γωνία του δωματίου, αλλά υπάρχουν λίγα από αυτά (ας πούμε 5). Οι περιοχές όπου αυτές οι δύο ομάδες ανθρώπων προτιμούν να στέκονται αλληλεπικαλύπτονται, έτσι υπάρχουν δύο άτομα με κίτρινα πουκάμισα που έχουν γείτονες με πράσινο πουκάμισο γύρω τους. Τώρα, αν εκπαιδεύσουμε τον ταξινομητή σε αυτά τα δεδομένα, απλώς θα μας πει ότι όλοι φορούν πράσινα πουκάμισα και θα είναι σωστό το 99,5% των περιπτώσεων. Οπότε δοκιμάζεται δειγματοληψία με αντικατάσταση έως ότου υπάρχουν 1.000 άτομα στην ομάδα των κίτρινων μπλουζών. Σε αυτήν την περίπτωση, η δειγματοληψία με αντικατάσταση θα κάνει να φαίνεται ότι υπάρχουν πέντε θέσεις όπου στέκονται άτομα με κίτρινα πουκάμισα και όχι αλλού. Αυτό είναι υπερβολικό. Το SMOTE λύνει αυτό το πρόβλημα με τον εξής τρόπο: πάρτε τον πρώτο κίτρινο πουκάμισο και υπολογίστε την απόσταση μεταξύ αυτού και του κοντινότερου φίλου του, ονομάστε το x. Πολλαπλασιάστε αυτή την απόσταση με έναν τυχαίο αριθμό μεταξύ 0 και 1, ονομάστε την α. Βάλτε ένα άλλο κίτρινο πουκάμισο σε εκείνο το σημείο, σε απόσταση από το αρχικό κίτρινο πουκάμισο. Κάντε το ξανά με όλα τα κίτρινα πουκάμισα μέχρι να έχετε 1.000 κίτρινα πουκάμισα να στέκονται στην περιοχή όπου τους αρέσει να στέκονται, αλλά όχι μόνο στα αρχικά 5

σημεία. Τώρα έχετε ένα ισορροπημένο σετ με ένα πιο ρεαλιστικό δείγμα της θέσης των κίτρινων πουκάμισων.

Σημαντικότητα χαρακτηριστικών

Σήμερα, στα έργα μηχανικής μάθησης είναι απαραίτητη η εξαγωγή πληροφοριών για τα χαρακτηριστικά των δεδομένων. Μερικές φορές, ορισμένα χαρακτηριστικά δεν έχουν καμία συμβολή στις προβλέψεις και τη μοντελοποίηση και είναι πρακτικά άχρηστα. Από την άλλη μεριά ορισμένα χαρακτηριστικά μπορεί να περιέχουν τη μεγαλύτερη ποσότητα πληροφοριών για το δεδομένο σύνολο δεδομένων. Σε αυτή την περίπτωση, πρέπει να κατασκευαστούν μοντέλα που εκπαιδεύουν μόνο αυτά τα βασικά χαρακτηριστικά.

Εδώ, χρησιμοποιήθηκε η τεχνική αυτή για τον αλγόριθμο KNN και εφαρμόσαμε τη σημασία του προεπιλεγμένου μοντέλου για το Δέντρο αποφάσεων και το Τυχαίο Δάσος. Ως συμπέρασμα, τα περισσότερα πληροφοριακά χαρακτηριστικά περιγράφονται και συνοψίζονται στον ακόλουθο πίνακα:

Από τον παραπάνω πίνακα συνάγεται το συμπέρασμα ότι οι μεταβλητές – χαρακτηριστικά που είναι οι πιο σημαντικές και φέρουν τη μεγαλύτερη ποσότητα πληροφοριών αυτού του συνόλου δεδομένων είναι το person income, person home owneship, loan grade και loan percent income.

Στη συνέχεια, περιγράφεται η διαδικασία μηχανικής μάθησης που ακολουθείται σε αυτή τη μελέτη.

Τα μοντέλα ταξινόμησης μηχανικής μάθησης που έχουν δημιουργηθεί είναι 12 χρησιμοποιώντας 3 διαφορετικούς αλγόριθμους: Decision Tress, Random Forest και KNN. Σε κάθε αλγόριθμο έχουν αναπτυχθεί τέσσερα διαφορετικά μοντέλα και η σειρά που ακολουθήθηκε, εμφανίζεται παρακάτω.

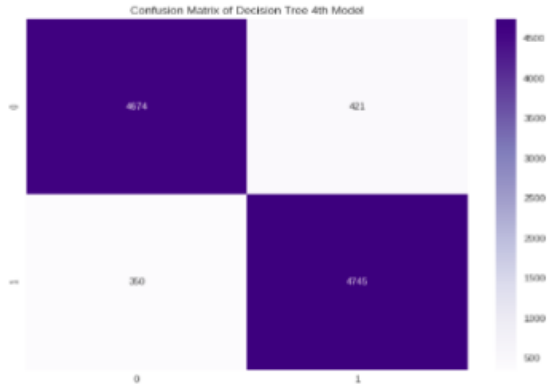
1. Αρχικό μοντέλο -Για κάθε αλγόριθμο μηχανικής μάθησης, δημιουργείται ένα αρχικό μοντέλο και οι παράμετροι που χρησιμοποιούνται είναι μόνο οι προεπιλεγμένες.
2. Μοντέλο με τις καλύτερες παραμέτρους - Μετά τη δημιουργία του αρχικού μοντέλου, το επόμενο βήμα είναι η αναζήτηση των καλύτερων παραμέτρων που μεγιστοποιούν την απόδοσή. Αφού βρεθούν αυτές οι υπερπαράμετροι, χρησιμοποιούνται για τη δημιουργία και την αξιολόγηση ενός νέου δεύτερου μοντέλου.
3. Μοντέλο μετά την χρήση της σηματικότητας χαρακτηριστικών (feature importance implementation)-Το τρίτο μοντέλο κάθε αλγορίθμου είναι ένα μοντέλο που εκπαιδεύεται, δοκιμάζεται και αξιολογείται με αποτέλεσμα ποια χαρακτηριστικά είναι τα καλύτερα για τον αλγόριθμο.
4. Μοντέλο με χρήση της μεθόδου SMOTE - Το τελικό μοντέλο κάθε αλγορίθμου είναι ένα μοντέλο που εκπαιδεύεται, δοκιμάστηκε και αξιολογήθηκε σε SMOTE. Αυτό σημαίνει ότι ο όγκος των σημείων των δεδομένων κάθε κλάσης είναι δεν είναι το ίδιο.

3. Αποτελέσματα

Αποτελέσματα ταξινομητή Δέντρων Απόφασης

Το καλύτερο μοντέλο δέντρου αποφάσεων που προέκυψε από αυτήν την ανάλυση είναι αυτό που χτίστηκε από το σύνολο δεδομένων Smote. Οι καλύτερες παράμετροι που εντοπίστηκαν είναι οι ακόλουθες:

1. criterion= ‘gini’
2. min samples leaf=1
3. min samples split=2



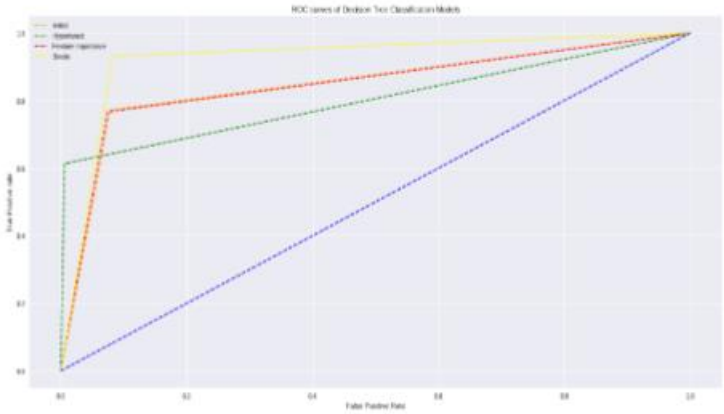
Εικόνα 18: ROC curve & ROC -AUC

Από τον παραπάνω confusion matrix, μπορεί να συναχθεί το συμπέρασμα ότι από τα 10.190 άτομα, 9.419 προβλέπονται στη σωστή κλάση. Επιπλέον, 4.674 άτομα κατατάχθηκαν στην κλάση 0, που σημαίνει ότι θα έπαιρναν δάνειο και αυτό ήταν σωστό καθώς πήραν το δάνειο, ενώ 4.745 ταξινομήθηκαν στην κατηγορίας 1, που σημαίνει ότι δεν θα έπαιρναν δάνειο και στην πραγματικότητα δεν το πήραν.

	Precision	Recall	F1-score	Support
0	0.93	0.92	0.92	5,095
1	0.92	0.93	0.92	5,095
Accuracy			**0.92**	10,190
macro avg	0.92	0.92	0.92	10,190
weighted avg	0.92	0.92	0.92	10,190

Accuracy score	ROC-AUC score
92.4337%	92.4337%

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	88.8752%	**91.1155%**	89.0900%	**92.4337%**
ROC-accuracy	84.6470%	80.3764%	84.6070%	924337%

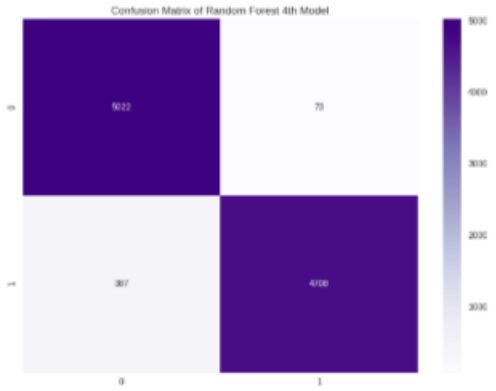


Συνολικά, από τον αλγόριθμο του δέντρου αποφάσεων έχουν δημιουργηθεί δύο μοντέλα με πολύ υψηλή ακρίβεια – πέρα του 90%, που σημαίνει ότι τα μοντέλα ταξινόμησης δένδρων αποφάσεων είναι αρκετά αποτελεσματικά για την ανάλυση πιστωτικού κινδύνου με τη μοντελοποίηση μηχανικής μάθησης.

Αποτελέσματα ταξινομητή Random Forest

Το καλύτερο μοντέλο του Random Forest από τα τέσσερα που έχουν δημιουργηθεί είναι αυτό που κατασκευάστηκε από το σύνολο δεδομένων Smote. Οι παράμετροι που οδήγησαν στην καλύτερη απόδοση ήταν οι εξής:

1. criterion= ‘gini’
2. min samples leaf=1
3. min samples split=2
4. n estimators=100



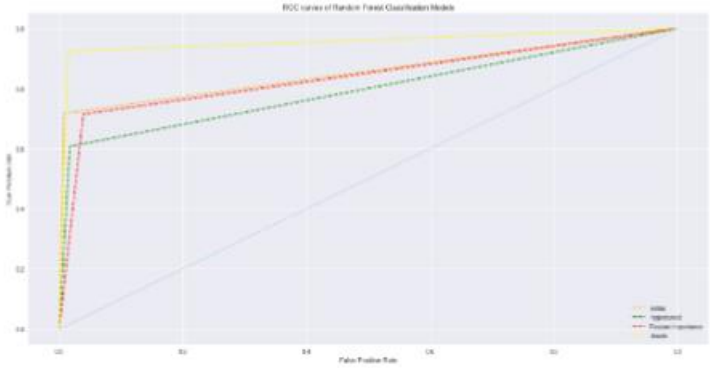
Εικόνα 4.6: ROC curve and ROC-AUC

Από τον παραπάνω confusion matrix, μπορεί να προκύψει ότι από 10.190 άτομα, προβλέφθηκαν 9.730 στη σωστή κλάση. Επιπλέον, στην κατηγορία 0 κατατάχθηκαν 5.022 άτομα, που σημαίνει ότι θα έπαιρναν ένα δάνειο και αυτό ήταν σωστό όπως το πήραν, ενώ 4.708 κατατάχθηκαν στην κατηγορία 1, που σημαίνει ότι δεν θα έπαιρναν δάνειο και στην πραγματικότητα δεν το πήραν.

	Precision	Recall	F1-score	Support
0	0.93	0.99	0.95	5,095
1	0.98	0.92	0.95	5,095
Accuracy			**0.95**	10,190
macro avg	0.96	0.95	0.95	10,190
weighted avg	0.96	0.95	0.95	10,190

Accuracy score	ROC-AUC score
95.4858%	95.4858%

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	93.2484%	**90.1028%**	90.7320%	**95.4858%**
ROC-accuracy	85.5467%	79.5260%	83.7812%	95.4858%



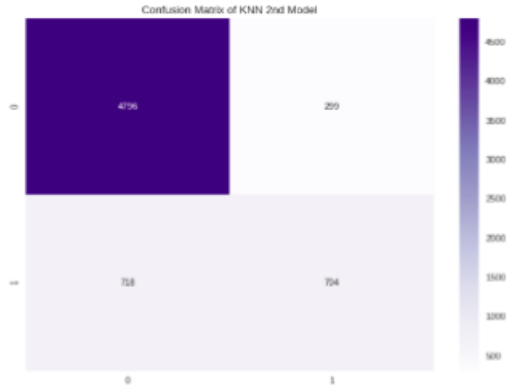
Συνολικά, από αλγόριθμο Random Forest και τα τέσσερα μοντέλα που έχουν δημιουργηθεί, είχαν πολύ υψηλά ακρίβεια – πάνω από 90%, που σημαίνει ότι τα μοντέλα Random Forest είναι τα πιο αποτελεσματικά για την ανάλυση πιστωτικού κίνδυνου με τη χρήση τεχνικών μηχανικής μάθησης.

Αποτελέσματα ταξινομητή KNN

Το καλύτερο μοντέλο KNN από τα τέσσερα που έχουν δημιουργηθεί ήταν αυτό που κατασκευάστηκε με τα καλύτερα παραμέτρους μετά την εφαρμογή υπερσυντονισμού.

Αυτές οι παράμετροι είναι οι εξής:

1.metric=’euclidean’
2.n-neighbors=13
3. p=2



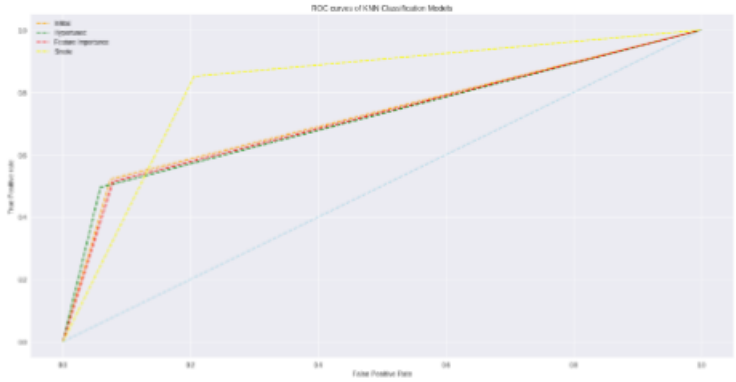
Από τον παραπάνω πίνακα σύγχυσης, συνάγεται το συμπέρασμα ότι από 6.517 άτομα, τα 5.500 είχαν προβλεφθεί σε η σωστή τάξη. Επιπλέον, 4.796 άτομα κατατάχθηκαν στην κατηγορία 0, πράγμα που σημαίνει ότι θα το έκαναν πάρτε ένα δάνειο και αυτό ήταν σωστό όπως το πήραν, ενώ το 704 κατατάχθηκε στην

κατηγορία 1, που σημαίνει ότι δεν θα έπαιρναν δάνειο και στην πραγματικότητα δεν το πήραν.

	Precision	Recall	F1-score	Support
0	0.87	0.94	0.90	5,095
1	0.70	0.50	0.58	1,422
Accuracy			0.84	6,517
macro avg	0.79	0.72	0.74	6,517
weighted avg	0.83	0.84	0.83	6,517

Accuracy score	ROC-AUC score
84.3947%	71.8196%

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	83.7195%	**84.3947%**	83.2592%	**82.2865%**
ROC-accuracy	72.3510%	71.8196%	71.7525%	82.2865%



Συνολικά, και από τα τέσσερα μοντέλα K-Nearest Neighbors που έχουν δημιουργηθεί, έχουν πολύ ελαφρώς καλή ακρίβεια – μεταξύ 80-85%, που σημαίνει ότι το τυχαίο δέντρο δάσους και απόφασης Τα μοντέλα είναι πολύ καλύτερα και αποτελεσματικά για ανάλυση πιστωτικού κινδύνου με μοντελοποίηση μηχανικής μάθησης.

Συνοπτικά Αποτελέσματα

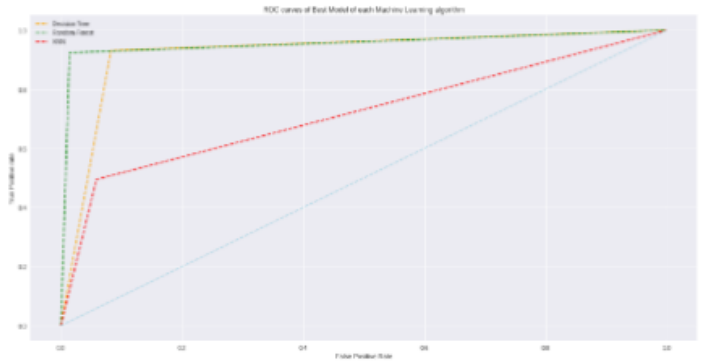
Μετά την ανάλυση ανά μοντέλο χωριστά, μερικές τελικές σημαντικές συγκρίσεις μηχανικής μάθησής παρουσιάζονται. Στους παρακάτω πίνακες, παρουσιάζονται

τα καλύτερα αποτελέσματα κάθε αλγορίθμου μηχανικής μάθησης εκμάθησης σε 3 φάσεις:

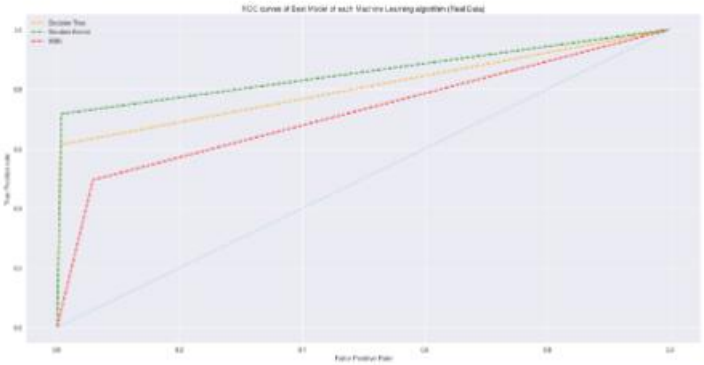
1. Συγκρίνοντας και τα 12 μοντέλα που έχουν αναπτυχθεί,
2. Σύγκριση όλων των μοντέλων που έχουν εκπαιδευτεί στο αρχικό – πραγματικό σύνολο δεδομένων και
3. Σύγκριση όλων των μοντέλων που έχουν εκπαιδευτεί σε σύνολο δεδομένων υπερδειγματοληψίας – smote σύνολο δεδομένων.

Όλες οι συγκρίσεις βασίζονται στις τρεις πιο σημαντικές μετρήσεις αξιολόγησης μηχανικής μάθησης: Ακρίβεια (Accuracy) (βαθμολογία F-1), ακρίβεια ROC (βαθμολογία ROC-AUC) και καμπύλες ROC (ROC curves).

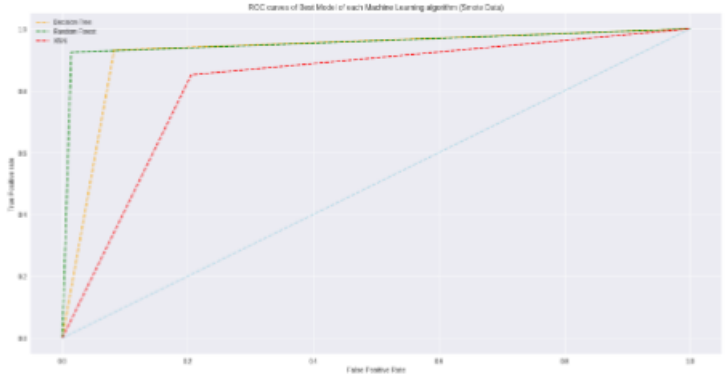
Scores	KNN	Decision Tree	Random Forest
Accuracy	84.3947%	**92.4337%**	**95.4858%**
ROC-accuracy	71.8196%	92.4337%	95.4858%



Model	KNN	Decision Tree	Random Forest
Accuracy	84.3947%	**91.1155%**	**93.2484%**
ROC-accuracy	71.8196%	80.3764%	85.5467%



Model	KNN	Decision Tree	Random Forest
Accuracy	82.2865%	**92.4337%**	**95.4858%**
ROC-accuracy	82.2865%	92.4337%	95.4858%



Ως αποτέλεμσα, τα δέντρα απόφασης, τα random forest και KNN μόντελα ταξινόμησης επιτυγχάνουν υψηλού επιπέδου προβλέψεις και ένα μεγάλο αριθμό σωστών ταξινομήσεων στο πρόβλημα του πιστωτικού κίνδυνου ταξινόμησης μηχανικής μάθησης αλλά και στα προβλήματα ταξινόμησης μηχανικής μάθησης στα οικονομικά γενικά. Ειδικά, ο ταξινομητής random forest είχε την καλύτερη συνολική απόδοση σε αυτή την ανάλυση, εφόσον έχει 95,4858% ακρίβεια και 95,4858% ακρίβεια ROC (ROC-AUC σκορ).

4. Συμπεράσματα

Το μέλλον της μηχανικής μάθησης στον τραπεζικό και χρηματοπιστωτικό τομέα αναμένεται να είναι μεγάλη, ειδικά στον τομέα της διαχείρισης κινδύνου. Θα αναπτυχθούν τεχνικές μηχανικής μάθησης και θα εφαρμοστούν στα τραπεζικά δεδομένα σε μια προσπάθεια βελτίωσης των λειτουργιών τους. Η ικανότητα των μοντέλων μηχανικής μάθησης για την ανάλυση μεγάλου όγκου δεδομένων με σχετικά ευκολότερο τρόπο και με μεγαλύτερη αξιοπιστία είναι πολύ σημαντικό. Η μηχανική μάθηση, έχοντας σημαντικές εφαρμογές στη διαχείριση κινδύνου, μπορεί να ενεργοποιήσει τη δημιουργία πιο ακριβών μοντέλων κινδύνου με τον εντοπισμό πολύπλοκων, μη γραμμικών προτύπων σε μεγάλα δεδομένα. Στην παρούσα διπλωματική εργασία έγινε αξιολόγηση και ανάλυση της βιβλιογραφίας γύρω από την εφαρμογή της μηχανής μάθησης στη διαχείριση κινδύνων στον χρηματοπιστωτικό τομέα. Το μεγαλύτερο μέρος της έρευνας φαίνεται να εστιάζει στη διαχείριση πιστωτικού κινδύνου. Αυτό θα μπορούσε να αποδοθεί στο

γεγονός ότι λαμβάνεται υπόψη ο πιστωτικός κίνδυνος ως ο σημαντικότερος κίνδυνος για έναν χρηματοπιστωτικό οργανισμό. Τα πλεονεκτήματα και τα μειονεκτήματα διαφορετικών τεχνικών μηχανικής μάθησης για την επίλυση συγκεκριμένων προβλημάτων διαχείρισης κινδύνου μελετήθηκαν αλλά και θα συνεχίσουν να μελετώνται και να αξιολογούνται περαιτέρω. Η έρευνα έδειξε ότι με την εφαρμογή της μηχανικής μάθησης στη διαχείριση κινδύνων, όπως ο πιστωτικός κίνδυνος, έχει ερευνηθεί εκτενώς. Ωστόσο, θα μπορούσε να διερευνηθεί περαιτέρω για ορισμένους τομείς όπου απαιτείται ανάλυση δεδομένων μεγάλης κλίμακας με σύνθετους και μη γραμμικούς υπολογισμούς. Καθώς οι τράπεζες και τα χρηματοπιστωτικά ιδρύματα θέλουν να αυξήσουν τις ικανότητές τους στη διαχείριση κινδύνων, θα ήταν χρήσιμο να εξερευνηθεί πώς η μηχανική μάθηση μπορεί να εφαρμοστεί στο συνδυασμό διαφορετικών κινδύνων και βελτίωση της δυνατότητας μείωσης αυτών των κινδύνων. Θα μπορούσαν επίσης να διερευνηθούν τομείς όπως ο κίνδυνος συμπεριφοράς, με άλλα λόγια παρακολούθηση συμπεριφοράς και δραστηριότητας. Αυτό σίγουρα θα μπορέσει να βοηθήσει τη διαχείριση κινδύνων στα χρηματοοικονομικά ιδρύματα. Στην παρούσα εργασία, δώδεκα μοντέλα μηχανικής μάθησης όπως τα δέντρα απόφασης, τα random forest και KNN μελετήθηκαν και η ανάλυση των μετρήσεων των μοντέλων έδειξε ότι ο KNN, τα δέντρα αποφάσεων και τα μοντέλα Random Forest είναι αρκετά ακριβή στην πρόβλεψη για κάθε άτομο στην σωστή κλάση στην κατάταξή για το πρόβλημα του πιστωτικού κινδύνου. Πιο συγκεκριμένα, η καλύτερη απόδοση επιτεύχθηκε με βάση δεδομένα υπερδειγματοληψίας, καθώς το 82,3%, το 92,4% και το 95,5% ήταν ακρίβεια του KNN, του δέντρου απόφασης και του αλγόριθμου Random Forest αντίστοιχα. Ωστόσο, αν πρέπει να επιλεχθεί ένας αλγόριθμος μηχανικής μάθησης είναι απαραίτητο να χρησιμοποιηθεί ο Random Forest, γιατί είναι το πιο ακριβές με ποσοστό 95,Τέλος, είναι σημαντικό να υπογραμμιστεί πως πολλά είδη μετασχηματισμών στο αρχικό σύνολο δεδομένων συνέβαλε στη λήψη της τελευταίας μορφής του συνόλου δεδομένων που τελικά χρησιμοποιήθηκε για εκπαίδευση και δοκιμή στα μοντέλα. Είναι αλήθεια ότι, όταν πρόκειται για μη ισορροπημένα σύνολα δεδομένων, το να εργάζεσαι στα πραγματικά δεδομένα ή στα oversampled data είναι ένα θέμα που προκαλεί μεγάλη συζήτηση μεταξύ ερευνητών και επιστήμονων και εξαρτάται από πολλούς διαφορετικούς παράγοντες. Έχοντας λοιπόν υπόψη τα παραπάνω, χρησιμοποιώντας τη μέθοδο SMOTE μπορεί κανείς να επιτύχει, γενικά, καλύτερη απόδοση του ταξινομητή σε αυτό το πρόβλημα. Ωστόσο, αυτό δεν σημαίνει ότι πάντοτε θα δίνει καλύτερα αποτελέσματα σε κάθε χρήση στη μηχανική. Αν και η μηχανική μάθηση θεωρείται ένα χρήσιμο εργαλείο για την ανάλυση πιστωτικού κινδύνου και την πρόβλεψη αθέτησης, υπάρχουν αρκετοί περιορισμοί που συνδέονται με αυτόν τον τύπο ανάλυσης. Ο πιο σημαντικός περιορισμός είναι η ποιότητα των δεδομένων και η προγνωστική ισχύς. Απαραίτητη προϋπόθεση για να χτίσεις ένα καλό και αξιόπιστο μοντέλο είναι η απόκτηση ενός συνόλου αντιπροσωπευτικών δεδομένων υψηλής ποιότητας. Προσδιορίζοντας τα χαρακτηριστικά που επηρεάζουν σε μεγάλο βαθμό την αδυναμία πληρωμής ή όχι είναι μία από τις μεγαλύτερες προκλήσεις. Όσο οι οικονομικές συνθήκες αλλάζουν συνεχώς και γρήγορα, ενώ νέοι πελάτες, νέα προϊόντα και νέες τάσεις εισάγονται, χρειάζονται νέα χαρακτηριστικά, νέες μεταβλητές και νέοι συσχετισμοί. Όλοι αυτοί οι περιορισμοί υπάρχουν σε αυτήν τη μελέτη, καθώς το σύνολο δεδομένων περιέχει μόνο 32.581 δανειολήπτες με τα χαρακτηριστικά τους, τα οποία είναι περιορισμένα και αποκαλύπτουν μόνο ένα μικρό μέρος

όλων αυτών των παραμέτρων που θα μπορούσαν να συμβάλουν στον δανεισμό ή όχι. Να επισημανθεί επίσης πως τα προσωπικά δεδομένα αλλά και η χρήση τους είναι ένα πολύ ευαίσθητο θέμα που απασχολεί επιχειρήσεις και οικονομικά ιδρύματα. Καθώς ο χρόνος αλλάζει και υπάρχουν περισσότερα δεδομένα διαθέσιμα, μπορεί να γίνει περαιτέρω έρευνα λαμβάνοντας υπόψη πιο σύνθετα σύνολα δεδομένων. Μπορούν να χρησιμοποιηθούν νέες δυνατότητες και νέες τεχνικές στο μέλλον σε έργα, καθώς η μηχανική μάθηση και η τεχνητή νοημοσύνη γενικά αναπτύσσονται ραγδαία.

Κίνδυνοι στην υιοθέτηση της μηχανικής μάθησης για τραπεζικές εργασίες

Φυσικά, η τεχνολογία Τεχνητής Νοημοσύνης μπορεί να φέρει επανάσταση στον τραπεζικό τομέα. Ωστόσο, υπάρχουν ορισμένοι κίνδυνοι — αλλά συνδέονται κυρίως με την καινοτομία των τεχνολογιών και την έλλειψη πλήρους κατανόησης μεταξύ των χρηστών σχετικά με το πώς πραγματικά λειτουργούν.

Περικοπές θέσεων εργασίας

Αυτός είναι ένας από τους πιο συνηθισμένους κινδύνους και φόβους που σχετίζονται με την τεχνητή νοημοσύνη και τη μηχανική μάθηση, ακόμη και ανεξαρτήτως του πεδίου εφαρμογής τους. Ωστόσο, η σύγχρονη έρευνα δείχνει ότι η Τεχνητή Νοημοσύνη στον τραπεζικό τομέα θα προσφέρει πολύ μεγαλύτερο αριθμό νέων θέσεων εργασίας σε σύγκριση με τον αριθμό των επαγγελμάτων που θα γίνουν αζήτητα.

Λιγότερη εμπιστοσύνη λόγω λιγότερης ανθρώπινης επαφής

Υπάρχει επίσης η άποψη ότι οι χρήστες θα αισθάνονται λιγότερη εμπιστοσύνη στα χρηματοπιστωτικά ιδρύματα λόγω των λιγότερων ευκαιριών να συνεργαστούν με ανθρώπινους συμβούλους. Αυτό είναι αλήθεια, αλλά μόνο εν μέρει. Πιθανότατα θα παρατηρήσουμε αυτή την τάση, αλλά μόνο σε σχέση με ανθρώπους που γεννήθηκαν στην προηγούμενη γενιά, οι οποίοι δεν είναι πολύ διατεθειμένοι να πιστεύουν στην τεχνολογία εξ αρχής. Αλλά όσον αφορά τη γενιά των millennials που είναι πρόθυμοι να πληρώσουν περισσότερα για ευκολία και αξιοπιστία, θα χαρούν να έχουν την ευκαιρία να εκτελέσουν οποιαδήποτε λειτουργία με λίγα κλικ.

Ηθικοί κίνδυνοι

Οι ηθικοί κίνδυνοι συνδέονται με το γεγονός ότι ο όγκος των δεδομένων που συλλέγουν, αποθηκεύουν, συστηματοποιούν, αναλύουν και χρησιμοποιούν οι χρηματοοικονομικές εταιρείες προς όφελός τους (καθώς και προς όφελος των πελατών) συνεχίζει να αυξάνεται. Σε ορισμένους χρήστες δεν αρέσει αυτή η τάση, αλλά αυτή τη στιγμή είναι αδύνατο να προβούν σε οποιαδήποτε ενέργεια χωρίς να αφήσουν ίχνη προσωπικών δεδομένων. Αυτό το γεγονός δεν αρέσει περισσότερο στους απατεώνες, καθώς αρχίζουν ήδη να αισθάνονται ότι γίνεται όλο και πιο δύσκολο να εξαπατήσουν τα συστήματα AI. Ταυτόχρονα, αυτό είναι ένα σαφές πλεονέκτημα για τη βελτίωση της εμπειρίας χρήστη και την ενίσχυση του επιπέδου ασφάλειας.

Κίνδυνοι ψευδώς θετικών αποτελεσμάτων

Τα συστήματα μηχανικής μάθησης και η τεχνητή νοημοσύνη παρακολουθούν πρότυπα συμπεριφοράς των χρηστών και τα συγκρίνουν με αποδεκτές εκδόσεις του κανόνα σε σχέση με κάθε χρήστη. Έτσι, για παράδειγμα, εάν ένας χρήστης ολοκληρώσει μια συναλλαγή στο εξωτερικό, αλλά δεν έχει ειδοποιήσει την τράπεζα για το ταξίδι του (ή η τράπεζα για κάποιο λόγο δεν μπόρεσε να καταλάβει αυτές τις πληροφορίες, για παράδειγμα, ο χρήστης δεν αγόρασε το

εισιτήριο από την πιστωτική του κάρτα , αλλά το έλαβε ως δώρο), τότε αυτή η πράξη μπορεί να ερμηνευθεί ως δόλια. Αλλά στην πραγματικότητα, όλα ήταν νόμιμα – μόνο μια μικρή έλλειψη πληροφοριών οδήγησε σε ένα ψευδώς θετικό αποτέλεσμα.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] J. Scott. The probability of bankruptcy: a comparison of empirical predictions and theoretic models. Journal of Banking and Finance, Vol. 5, pp. 317- 44., 1981.

[2] C.V. Zavgren. Tthe prediction of corporate failure: the state of the art. Journal of Accounting Literature, pp. 1-38, 1983.

[3] E. Altman. Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. Journal of Finance 23(4), 589-609., 1968.

[4] K. Keasey and R. Watson. Financial distress prediction models: a review of their usefulness. British Journal of Management, Vol. 2, pp. 89-102, 1991.

[5] F.L. Jones. ‘current techniques in bankruptcy prediction. Journal of Accounting Literature, Vol. 6, pp. 131-64., 1987.

[6] Zanakis S.H. Dimitras, A.I. and C. Zopounidis. A survey of business failure with an emphasis on prediction methods and industrial applications. European Journal of Operational Research, Vol. 90, pp. 487-513., 1996.

[7] Hu M.Y. Patuwo B.E. Zhang, G. and D.C. Indro. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. European Journal of Operational Research, Vol. 116, pp. 16-32, 1999.

[8] Galai D. Crouhy, M. and R. Mark. A comparative analysis of current credit risk models. Journal of Banking and Finance, Vol. 24, pp. 59-117, 2000.

[9] W. H. Beaver. Financial ratios as predictors of failure. Empirical Research in Accounting:Selected Studies 4, 71-102, 1966.

[10] Rashmi Malhotra and D.K. Malhotra. Evaluating consumer loans using neural networks.omega. Omega, 31:83–96, 2003.

[11] Sihem Khemakhem and Younes Boujelbene. Artificial intelligence for credit risk assessment: Artificial neural network and support vector machines. ACRN Oxford Journal of Finance and Risk Perspectives 6.2, 2017.

[12] Shola Oroja Nnamdi I. Nwulu and Mustafa Ilkan. Credit scoring using soft computing schemes: A comparison between support vector machines and artificial neural networks. SpringelLink, 2011.

[13] A. Dima and S Vasilache. Ann model for corporate credit risk assessment. In International Conference on Information and Financial Engineering, pages 94–98.

[14] Jian-Min Xu Rong-Zhou Li, Su-Lin Pang. Neural network credit-risk evaluation model based on back-propagation algorithm. IEEE, 2002.

[15] Yong-li Tang Xin-yue Hu. Ann-based credit risk identificaion and control for commercial banks. IEEE, 2009.

[16] Maciej Zięba Jakub M. Tomczak. Boosted svm with active learning strategy for imbalanced data. SpringerLink, 2014.

[17] S Viaene-M Stepanova J Suykens J Vanthienen B Baesens, T Van Gestel. Benchmarking state-of-the-art classification algorithms for credit scoring. SpringerLink, 2003.

[18] Dijan Oreski Goran Oreski Stjepa nOreski. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Elsevierk, 2012.

[19] Nadia Labidi Noureddine Boudriga Yacine Djemaiel. A dynamic hybrid rbf/elman neural networks for credit scoring using big data. SpringerLink, 2016.

[20] Sriramjee G. Praveen M. A. H. Farquad, V. Ravi. Credit scoring using pca-svm hybrid model. SpringerLink, 2011.

[21] Bandaru Rakesh Kumar Radha Vedala. An application of naive bayes classification for credit scoring in e-lending platform. IEEE, 2012.

[22] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3):523–541, 1997.

[23] Andrzej Szwabe and Pawel Misiorek. Decision trees as interpretable bank credit scoring models: 14th international conference. BDAS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 18-20, 2018.

[24] A. Baruah. “ai applications in the top 4 indian banks”, tech emergence. 2018.

[25] L. D’Monte. Nse bets big on ai, blockchain to mitigate algo-trading risks. 2018.

[26] Ariff M. Ahmad, N. H. Multi-country study of bank credit risk determinants. International journal of banking and finance. 2007.

[27] Zenglian Zhang. Research of default risk of commercial bank’s personal loan based on rough sets and neural network. 2011.

[28] SaberFeki Asma Feki, Anis BenIshak. Feature selection using bayesian and multiclasssupport vector machines approaches: Application to bank risk prediction. 2012.

[29] Abdelbaki Attioui Youness Abakarim, Mohamed Lahby. An efficient real time model for credit card fraud detection based on deep learning. 2018.

[30] Pourya Shamsolmoali Masoumeh Zareapoora. Application of credit card fraud detection: Based on bagging ensemble classifier. 2018.

[31] Manjeevan Seera Chee Peng Lim Asoke K. Nandi Kuldeep Randhawa, Chu Kiong Loo.Credit card fraud detection using adaboost and majority voting. 2018.

[32] Abdulai J.D. Gyamfi, N.K. Bank fraud detection using support vector machine. 2018.

[33] D. Tzelepis V. Tampakas S. Kotsiantis, E. Koumanakos. Forecasting fraudulent financial statements using data mining. 2005.

[34] Radha Vedala; Bandaru Rakesh Kumar. An application of naive bayes classification for credit scoring in e-lending platform. 2012.

[35] GUIDO SARAH MULLER, A. C. INTRODUCTION TO MACHINE LEARNING WITH PYTHON: A GUIDE FOR DATA SCIENTISTS. O'REILLY MEDIA, INC., 1005 GRAVENSTEIN HIGHWAY NORTH, SEBASTOPOL, CA 95472, 2017.

[36] T. M. MITCHELL. MACHINE LEARNING. MCGRAW-HILL

SCIENCE/ENGINEERING/MATH, 1997.

[37]

[38] POOJARA SHIVANANDA R. DHARWADKAR NAGARAJ V. KALYANKAR, G. D.

Predictive analysis of diabetic patient data using machine learning and hadoop. [Online] pp. 619 - 624., 2017.

[39] GUIDO SARAH. MULLER, A. C. INTRODUCTION TO MACHINE LEARNING WITH PYTHON: A GUIDE FOR DATA SCIENTISTS, O'REILLY MEDIA, INC. O' Reilly, 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2017.

[40] POOJARA SHIVANANDA R. DHARWADKAR NAGARAJ V. KALYANKAR, G. D. Predictive analysis of diabetic patient data using machine learning and hadoop. 2017.

[41] Yongqiao Wang; Shouyang Wang; K.K. Lai. A new fuzzy support vector machine to evaluate credit risk. 2005. .

[42] Chieh-Jen Wangc Cheng-Lung Huang, Mu-Chen Chen. Credit scoring with a data mining approach based on support vector machines. 2007.

[43] Edward Altmanc Flavio Barboza, Herbert Kimura. Machine learning models and bankruptcy prediction. 2017.

[44] J. Grus. Data Science from Scrach. O' Reilly Media INC., 1005 Gravenstein Highway North, Sebastopol, 2015.

[45] Richard A. Olshen Charles J. Stone Leo Breiman, Jerome H. Friedman. Classification And Regression Trees. O' Reilly, New York, Routledge, 1984.

[46] Leo Breiman. Random forests. 2001.

[47] Gregorvon Schweinitz Johannes Beutel, Sophia Lista. Does machine learning help us predict banking crises? 2019.

[48] P. Hart T. Cover. Nearest neighbor pattern classification. 1967.

[49] Shalev-Shwartz Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 32 Avenue of Americas, New York, NY 10013- 2473, USA, 2014.