Πανεπιστήμιο Πειραιώς - Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Μεθοδολογίες μηχανικής μάθησης για την πρόγνωση πιστωτικού κινδύνου** **Machine Learning Methodologies For Credit Risk Prediction**
Ονοματεπώνυμο Φοιτητή	**Γεώργιος Χίνης**
Πατρώνυμο	**Παναγιώτης**
Αριθμός Μητρώου	**ΜΠΠΛ/17065**
Επιβλέπων	**Σωτηρόπουλος Διονύσιος Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης **Ιανουάριος 2021**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γ. Τσιχριντζής, Καθηγητής	Δ. Σωτηρόπουλος Επίκουρος Καθηγητής	Ε. Σακκόπουλος Επίκουρος Καθηγητής

Περίληψη

Με βάση την εξέλιξη της τεχνολογίας τις τελευταίες δεκαετίες σε ότι αφορά την μηχανική μάθηση, δίνεται η ευκαιρία στα χρηματοπιστωτικά ιδρύματα να την χρησιμοποιήσουν προς όφελός τους. Ένα σύγχρονο και πλέον διαδεδομένο πρόβλημα στις επιχειρήσεις αυτού του είδους είναι ο πιστωτικός κίνδυνος. Η επιστήμη των υπολογιστών λοιπόν καλείται να δώσει λύση σε ένα πρόβλημα το οποίο φαινόταν αδύνατον να λυθεί. Με την εξόρυξη των δεδομένων που ο όγκος τους φαντάζει απαγορευτικός για να τον διαχειριστεί ο άνθρωπος, αλλά και με ειδικά μοντέλα μηχανικής μάθησης δόθηκε η δυνατότητα της αύξησης της πιθανότητας να αποφευχθεί ο πιστωτικός κίνδυνος. Στην συγκεκριμένη εργασία λοιπόν αναλύονται όλες αυτές οι έννοιες και δίνεται λύση σε ένα ειδικό πρόβλημα μιας εταιρείας που χορηγεί δάνεια και προσπαθεί να επιτύχει όσον το δυνατόν πιο εύκολη και άμεση επιστροφή των κεφαλαίων.

Abstract

Based on the evolution of technology in recent decades in terms of machine learning, financial institutions are given the opportunity to use it to their advantage. A modern and widespread problem in this type of business is credit risk. Computer science is therefore called upon to provide a solution to a problem that seemed impossible to solve. With the extraction of data whose volume seems prohibitive for human management, but also with special models of machine learning, it was possible to increase the probability of avoiding credit risk. In this paper, therefore, all these concepts are analyzed and a special problem is solved for a company that grants loans and tries to achieve as easy and immediate return of funds as possible.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα καθηγητή κ. Διονύσιο Σωτηρόπουλο για την εμπιστοσύνη, την επιμονή αλλά και την καθοδήγηση που μου πρόσφερε καθ' όλη την διάρκεια της πορείας των σπουδών μου αλλά και πιο συγκεκριμένα μέχρι την περάτωση της διατριβής μου.

Επιπλέον δεν θα μπορούσα να μην ευχαριστήσω όλους τους κοντινούς μου ανθρώπους που ήταν δίπλα μου, με στήριξαν και μου έδωσαν ώθηση να συνεχίσω μέχρι να πετύχω τον στόχο μου. Ένα μεγάλο ευχαριστώ λοιπόν στους γονείς, τους φίλους και την κοπέλα μου Κατσούλη Μαρία για την υπομονή που έδειξαν και ελπίζω να συνεχίσουν να δείχνουν.

Περιεχόμενα

Περίληψη 3
Abstract 4
Ευχαριστίες 5
Πίνακας Εικόνων 7
1. Εισαγωγή 8
2. Μεθοδολογίες Μηχανικής Μάθησης στην πρόβλεψη πιστωτικού κινδύνου 12
2.1. Αλγόριθμοι Παλινδρόμησης (Regression Algorithms) 19
2.2. Αλγόριθμοι Κανονικοποίησης (Regularization Algorithms) 19
2.2 Αλγόριθμοι Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Network Algorithms) 20
2.4. Εκμάθηση με Δέντρο Αποφάσεων (Decision Tree Learning) 21
2.5. Ο Αλγόριθμος Catboost 23
3. Dataset 26
3.1. Επεξεργασία των δεδομένων 26
3.2 Επιλογή χαρακτηριστικών 30
4. Πειραματικά Αποτελέσματα 41
Συμπεράσματα και μελλοντική έρευνα 50
Αναφορές 52

Πίνακας Εικόνων

Εικόνα 1: Πιστωτικός κίνδυνος 9
Εικόνα 2: Διαδικασία αλγορίθμου πρόβλεψης πιστωτικού κινδύνου [10] 15
Εικόνα 3: Αξιολόγηση αλγορίθμου 16
Εικόνα 4: Πίνακας σύγχυσης [12] 17
Εικόνα 5: accuracy, precision,recall 18
Εικόνα 6: Στρώματα νευρωνικού δικτύου [15] 20
Εικόνα 7: Δέντρο αποφάσεων[16] 21
Εικόνα 8: Ο αλγόριθμος Catboost [18] 24
Εικόνα 9: Σύγκριση Catboost με άλλους αλγορίθμους ίδας κατηγορίας [19] 25
Εικόνα 10: φόρτωμα των δεδομένων από το αρχείο csv 27
Εικόνα 11: Αλλαγή του issue_d σε format ημερομηνίας 27
Εικόνα 12: Διαφορές στις στήλες των δεδομένων 28
Εικόνα 13: Αλλαγή πεδίων στην κατάλληλη μορφή προς επεξεργασία 29
Εικόνα 14: Αλλαγή χαρακτηριστικών του πεδίου emp_length 30
Εικόνα 15: Πεδία με κενές τιμές 31
Εικόνα 16: Παράδειγμα πεδίου μέγιστης τιμής 32
Εικόνα 17: Πεδία με τιμές σε όλα τα κελιά 33
Εικόνα 18: Απαλοιφή στήλης με ένα μονο δεδομένο 33
Εικόνα 19: Πεδία με ομοιότητες 34
Εικόνα 20: Ζευγάρια με μεγάλη ομοιότητα μεταξύ τους 35
Εικόνα 21: Περιεχόμενα αλφαριθμητικών πεδίων 36
Εικόνα 22: Ομοιότητα ανάμεσα σε purpose και title 37
Εικόνα 23: Εξάρτηση application_type με verification_status_joint 38
Εικόνα 24: Πλήθος δανείων ανά κατηγορία 38
Εικόνα 25:Διαχωρισμός ανάλογα με τον τύπο του δανείου 39
Εικόνα 26: Πλήθος δανείων μετά τον διαχωρισμό 40
Εικόνα 27: Στατιστική εικόνα 40
Εικόνα 28: Πλήθος δανείων ανά βαθμό 41
Εικόνα 29: Εκπαίδευση αλγορίθμου 42
Εικόνα 30: Διάγραμμα αποτελεσμάτων 44
Εικόνα 31: Σημαντικά χαρακτηριστικά 45
Εικόνα 32: Σχέση καλών-κακών δανείων για το loan_amnt 46
Εικόνα 33: Σχέση καλών κακών δανείων με βάση την τελευταία έρευνα 47
Εικόνα 34: Καμπύλη precision με recall 48
Εικόνα 35: Διάγραμμα αποτελεσμάτων μετά την προσαρμογή των δεδομένων 49

1. Εισαγωγή

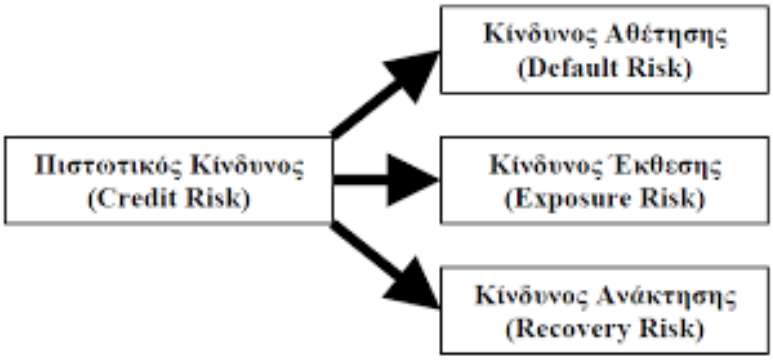
Οι σύγχρονες επιχειρήσεις έχοντας ως βασικό γνώμονα την ανάπτυξη πρέπει να σκέφτονται πάντα τον τρόπο με τον οποίο θα κάνουν νέες επενδύσεις. Χωρίς επενδύσεις μια εταιρεία μπορεί εύκολα να φτάσει σε σημείο να μείνει πίσω σε ότι αφορά τον ανταγωνισμό και να μην μπορεί να ανταπεξέλθει στις απαιτήσεις της ταχύρυθμης ανάπτυξης. Έτσι θεωρείται αναγκαίο, μια επιχείρηση να βρίσκει τους πόρους για να επενδύσει σε κάτι καινοτόμο το οποίο σίγουρα θα επιφέρει περισσότερα κέρδη εάν σαν επένδυση πετύχει.

Μια επιχείρηση είναι μια οικονομική μονάδα που ως στόχο έχει την μεγιστοποίηση της αξίας, με σκοπό την επιβίωση αλλά και την μελλοντική επέκταση [1]. Για την βιωσιμότητα λοιπόν χρειάζεται σίγουρα μεγάλο μέρος των πόρων να διατεθεί στις πωλήσεις και στο πως αυτές θα κρατήσουν σταθερή την ανάπτυξη της επιχείρησης, αλλά ένα μέρος πρέπει να διατίθεται και στην επένδυση. Μια επένδυση μπορεί να είναι σίγουρα απαραίτητη αλλά εγκυμονεί κινδύνους. Οι πόροι οι οποίοι δαπανούνται για αυτήν σίγουρα θα αναγκάσει ένα άλλο κομμάτι της επιχείρησης να μείνει πίσω. Για παράδειγμα αν ένα μεγάλο μέρος του κεφαλαίου της επιχείρησης χρησιμοποιηθεί για μια μακροπρόθεσμη επένδυση τότε υπάρχει κίνδυνος να μην μπορεί να αποπληρωθεί κάποια άλλη υποχρέωση μείζονος σημασίας. Ο κίνδυνος αυτός μπορεί να επιφέρει οικονομική δυσχέρεια σε μια επιχείρηση, σε περίπτωση που μια επένδυση δεν καρποφορήσει σε ότι αφορά τα κέρδη στα χρονικά περιθώρια που είχαν προγραμματιστεί.

Από την άλλη πλευρά υπάρχει και ο πιστωτικός κίνδυνος που αποτελεί την πιθανότητα που έχει μια εταιρεία να εισπράξει αργοπορημένα ή ακόμα και να μην εισπράξει τις οφειλές από τρίτους [2]. Ο πιστωτικός κίνδυνος είναι ένας από τους πιο συχνούς πλέον λόγους που οδηγούν μία επιχείρηση σε πτώχευση. Είναι πολύ σημαντικό για μια εταιρεία να μπορεί να προϋπολογίσει τα κέρδη και τις ζημιές που θα έχει μέσα σε μία χρονιά. Έτσι θα μπορεί να είναι σε θέση να συνεχίζει χωρίς να φτάσει σε σημείο να μην μπορεί να ανταπεξέλθει σε υποχρεώσεις. Το πρόβλημα είναι ότι σε επιχειρήσεις που οι δοσοληψίες δεν γίνονται άμεσα και από τις δύο πλευρές,

υπάρχει πάντα ο πιστωτικός κίνδυνος. Καμία επιχείρηση δεν μπορεί να γνωρίζει προκαταβολικά αν οι πιστωτές θα αποπληρώσουν τα χρέη τους.

Παραδείγματα εταιρειών που λειτουργούν μόνο με τον πιστωτικό κίνδυνο είναι αυτές της τηλεπικοινωνίας. Ποτέ δεν συμβαίνει κάποιος να προπληρώνει την συνδρομή του σε μια τέτοιου είδους εταιρεία και έτσι εγκυμονεί ο κίνδυνος να μην αποπληρώσει τον λογαριασμό ο πελάτης πάρα την χρήση των υπηρεσιών. Βέβαια ο κίνδυνος στην συγκεκριμένη περίπτωση δεν μπορεί να θεωρηθεί μεγάλος καθώς ο πελάτης γνωρίζει ότι θα έχει κυρώσεις σε περίπτωση που δεν ανταποκριθεί στην ώρα του. Επίσης συνήθως τα ποσά ανά άτομο είναι πολύ μικρά οπότε αυτό κάνει ακόμα πιο αμελητέο τον κίνδυνο. Σε περίπτωση όμως που έχουμε μεγάλο πιστωτικό κίνδυνο οι επιπτώσεις για μια εταιρεία είναι τεράστιες. Εάν μια εταιρεία δεν μπορεί να ανταπεξέλθει στις υποχρεώσεις της τότε χάνει την αξιοπιστία της απέναντι στους δικούς της προμηθευτές ή σε αυτούς που της παρέχουν υπηρεσίες.



Εικόνα 1: Πιστωτικός κίνδυνος

Αναλυτικότερα ο πιστωτικός κίνδυνος μπορεί να λάβει τρεις διαστάσεις. Η πρώτη είναι ο κίνδυνος αθέτησης που αφορά την μη πραγματοποίηση μιας προγραμματισμένης οφειλής σε ένα προσυμφωνημένο χρονικό διάστημα. Η δεύτερη είναι ο κίνδυνος της έκθεσης που αφορά την μη τήρηση των υποχρεώσεων του δανειστή λόγω της υποχρέωσης που δεν καλύφθηκε από τον δανειολήπτη. Η Τρίτη είναι ο κίνδυνος της ανάκτησης που έχει να κάνει με το ποσό που μπορεί να ανακτηθεί σε περίπτωση που ο δανειολήπτης δεν μπορεί τελικά να ανταπεξέλθει στην υποχρέωσή του [3].

Έτσι ο πιστωτικός κίνδυνος μπορεί να επιφέρει τον κίνδυνο του να εκτεθεί μια εταιρεία ή ακόμα και να μην καταφέρει να ανταπεξέλθει καν στις

υποχρεώσεις της. Αυτά με την σειρά τους οδηγούν σε συσσώρευση χρεών που μπορεί σε συνδυασμό με κάποια οφειλή σε τράπεζα ή στο κράτος να θέσουν την εταιρεία σε κίνδυνο πτώχευσης. Συνήθως βέβαια όταν μιλάμε για μεγάλο πιστωτικό κίνδυνο αναφερόμαστε στον τραπεζικό τομέα. Είναι ο τομέας που αν δεν γίνει σωστή πρόβλεψη και διαχείριση με βάση τον δανειολήπτη, μπορεί να χαθεί το κεφάλαιο του δανεισμού [4].

Έχοντας λοιπόν κατανοήσει την σπουδαιότητα του πιστωτικού κινδύνου μία επιχείρηση πρέπει να βρει και τρόπο για την πρόβλεψη του [5]. Ο πιο ασφαλής τρόπος να προβλεφτεί ο πιστωτικός κίνδυνος είναι η χρήση του ιστορικού του δανειολήπτη. Είναι πολύ σημαντικό για το πιστωτικό ίδρυμα να είναι σε θέση να γνωρίζει κάποια στοιχεία για τον δανειολήπτη με τα οποία θα μπορούν να τον αξιολογήσουν. Το πρόβλημα που παρουσιάζεται είναι ότι έχουμε δύο ειδών δανειολήπτες. Η πρώτη κατηγορία αφορά τις επιχειρήσεις. Είναι λογικό ότι σε ότι αφορά τις επιχειρήσεις - νομικά πρόσωπα, η συλλογή των στοιχείων είναι αρκετά απλή. Το πρόβλημα που παρουσιάζεται όμως έχει να κάνει με τα ποσά που συνήθως είναι πολύ μεγάλα σε αυτήν την κατηγορία. Η δεύτερη περίπτωση αφορά ιδιώτες όπου έχουμε συνήθως αρκετά μικρότερα ποσά. Το πρόβλημα όμως στην δανειοδότηση ενός ιδιώτη είναι το ότι δεν είναι σε θέση ένα ίδρυμα να γνωρίζει πολλά για αυτόν. Όποτε πρέπει να τον αξιολογήσει με διαφορετικά κριτήρια, όπως είναι η περιουσιακή κατάσταση, το αν χρωστούσε κάποια στιγμή στην τράπεζα, ο χρόνος συνεργασίας κλπ.

Υπάρχουν τεχνικές τις οποίες εφαρμόζουν οι φορείς πίστωσης με σκοπό την εξάλειψη του πιστωτικού κινδύνου. Η σωστή δομή ενός δανείου είναι πολύ σημαντική για την αποπληρωμή του. Ένα δάνειο θα πρέπει να δίνει τέτοιους όρους στον δανειολήπτη που σε συνδυασμό με τους τόκους να είναι σε θέση να καταβάλει εμπρόθεσμα την υποχρέωση του. Πολύ σημαντικό επίσης είναι όταν ένας δανειολήπτης είναι ήδη σε κατάσταση όπου έχει εν ενεργεία δάνειο, να μην του χορηγηθεί περαιτέρω πίστωση. Συνήθως κάθε νέα χρηματοδότηση καθιστά δυσκολότερη την αποπληρωμή των χρεών. Τέλος η τιτλοποίηση των περιουσιακών στοιχείων είναι απαραίτητη προϋπόθεση για την χορήγηση ενός δανείου. Πέρα από αυτές τις τεχνικές όμως η πρόβλεψη του πιστωτικού κινδύνου μπορεί να πάρει άλλες διαστάσεις. Σίγουρα το να είναι σε θέση ένα ίδρυμα να υπολογίσει σε γενικές

περιπτώσεις το τι μπορεί να κάνει για να λειτουργήσει σωστά η διαδικασία της αποπληρωμής, θα του δώσει τις βάσεις και μόνο. Το πιο σημαντικό κομμάτι στην πρόληψη του πιστωτικού κινδύνου είναι να εξειδικεύει τις περιπτώσεις των δανειοληπτών. Πιο συγκεκριμένα θα πρέπει να έχει την δυνατότητα να βαθμολογεί την πιστοληπτική ικανότητα του δανειολήπτη. Μόνο έτσι μπορεί να καταφέρει την πρόβλεψη του πιστωτικού κινδύνου και να την περιορίσει σημαντικά ώστε να μην χαθούν τα κεφάλαια.

2. Μεθοδολογίες Μηχανικής Μάθησης στην πρόβλεψη πιστωτικού κινδύνου

Με την πάροδο των χρόνων και την εξέλιξη της τεχνολογίας δημιουργήθηκε και η ανάγκη της ψηφιοποίησης των δεδομένων. Αυτό με την σειρά του έχει ως αποτέλεσμα να είναι τόσο ογκώδη τα δεδομένα που συχνά δεν είναι εύκολα να διαχειριστούν. Στο σημείο εκείνο έρχεται η εξόρυξη των δεδομένων η οποία μπορεί και δίνει την δυνατότητα σε τέτοιες βάσεις δεδομένων, να μπορούν να είναι επεξεργάσιμες και να δώσουν χρήσιμα συμπεράσματα [6]. Όπως αναφέρθηκε νωρίτερα ένας από τους πιο σημαντικούς τρόπους να προβλεφτεί ο πιστωτικός κίνδυνος είναι μέσω της αξιολόγησης ενός δανειολήπτη. Ένας δανειολήπτης μπορεί να αξιολογηθεί έχοντας κάποια δεδομένα που έχουν κομβικό χαρακτήρα για το εάν είναι σε θέση να αποπληρώσει ένα δάνειο. Έτσι καθίσταται πολύ σημαντικό γεγονός η ψηφιοποίηση των δεδομένων αυτών με σκοπό την άμεση ή μελλοντική χρήση τους.

Οι τράπεζες αναγκάστηκαν να αναπτύξουν ένα σύστημα μέτρησης κινδύνου για τα δάνεια που είναι να δώσουν στους πελάτες τους, χωρίς όμως να είναι εύκολο να αξιολογηθούν σωστά από τις πληροφορίες που παίρνουν από τον δανειολήπτη και να βγάλουν συμπεράσματα για τον κίνδυνο από αυτές [7]. Το τραπεζικό σύστημα είναι ένας από τους μεγαλύτερους πυλώνες ενός κράτους και έτσι έπρεπε να βρεθεί ένα μοντέλο το οποίο θα ήταν σε θέση να βαθμολογήσει τον κάθε δανειολήπτη με βάση τα βασικά του χαρακτηριστικά και στη συνέχεια να βγουν τα ανάλογα συμπεράσματα για το αν πρέπει να δανειστεί και το ύψος του ποσού που πρέπει να δοθεί με σκοπό πάντα την επιστροφή του ποσού με τους τόκους στην τράπεζα. Το μοντέλο αυτό θα πρέπει να είναι σε θέση να διαχειριστεί βασικές μεταβλητές όπως είναι το ετήσιο εισόδημα του ενδιαφερόμενου ή το ιστορικό του στην συγκεκριμένη τράπεζα. Με αυτόν τον τρόπο θα μπορεί να εξάγει δεδομένα τα οποία θα κατηγοριοποιήσουν τους δανειολήπτες σε καλές και κακές περιπτώσεις ανάλογα με τον κίνδυνο τον οποίο συντρέχει η αθέτηση και η μη αποπληρωμή του δανείου. Σε αυτό το σημείο είναι που δημιουργούνται και τα

μεγαλύτερα προβλήματα σε μοντέλα τέτοιου είδους καθώς φαντάζει δύσκολο να μπορεί να προβλεφτεί η αθέτηση ενός δανείου.

Οι εξελίξεις στην τεχνολογία σε ότι αφορά τον πιστωτικό κίνδυνο έχουν αυξηθεί ραγδαία καθώς θεωρήθηκε ένα από τα μεγαλύτερα προβλήματα των τελευταίων δεκαετιών. Σε θεωρητικό επίπεδο η διαδικασία ενός μοντέλου πρόβλεψης έχει να κάνει με την συλλογή σημαντικών πληροφοριών για τον ενδιαφερόμενο και στην συνέχεια με την χρήση της μηχανικής μάθησης να επέλθει ένα συμπέρασμα σε ότι αφορά την στάση που θα πρέπει να κρατήσει το πιστωτικό ίδρυμα με βάση την βαθμολογία που έδωσε το μοντέλο. Συνοπτικά η διαδικασία του μοντέλου βαθμολόγησης πίστωσης είναι η ανάπτυξη ενός στατιστικού μοντέλου με βάση τα δεδομένα που έχει στην κατοχή της η τράπεζα από παλιά δάνεια, η εφαρμογή του μοντέλου που εξήχθη με σκοπό την βαθμολόγηση του ενδιαφερόμενου και στο τέλος ο υπολογισμός της ακρίβειας του μοντέλου. Όπως είναι λογικό το να γίνει κάποιο λάθος στο στατιστικό μοντέλο μπορεί να αποβεί μοιραίο στην ακρίβεια και κατά συνέπεια στον κίνδυνο για αθέτηση του δανείου.

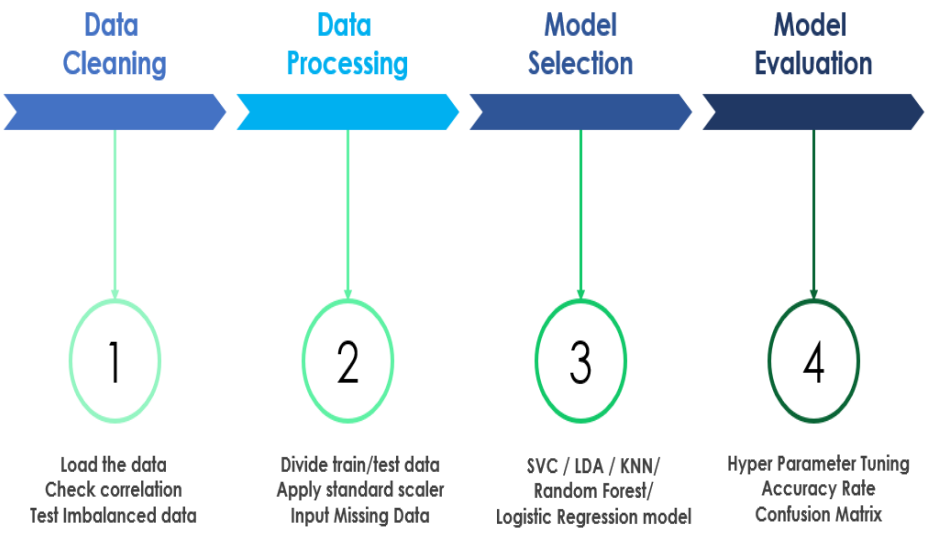
Έτσι λοιπόν χρησιμοποιήθηκε μια προσέγγιση γνωστή ως «μηχανική μάθηση» που αποτελείται από ένα σύνολο αλγορίθμων που έχουν σχεδιαστεί έτσι ώστε να είναι σε θέση να δώσουν λύση σε θέματα που δημιουργούν οι υπολογισμοί δεδομένων με μεγάλο όγκο [8]. Οι αλγόριθμοι αυτοί έχουν την δυνατότητα να ταξινομούν τα δεδομένα και να τα δομούν έτσι ώστε να μπορούν να εξάγουν χρήσιμα συμπεράσματα σχετικά με τον πιστωτικό κίνδυνο. Με την ταχύρυθμη εξέλιξη του τρόπου λειτουργίας του τραπεζικού συστήματος πλέον είναι σαφές ότι θα ήταν ανέφικτο να υπάρχει οποιοδήποτε άλλο μέτρο πρόληψης που να μπορεί να φέρει εις πέρας ένα τόσο δύσκολο έργο. Γι αυτό η μηχανική μάθηση έχει τόσο σημαντική επιρροή στον τραπεζικό τομέα και στον τρόπο λειτουργίας του με πιο πρόσφατο παράδειγμα αυτό της αντιμετώπισης του πιστωτικού κινδύνου.

Ως μηχανική μάθηση ορίζεται η επιστήμη που ασχολείται με την κατανόηση, την επεξεργασία και την μοντελοποίηση δεδομένων με την χρήση της τεχνητής νοημοσύνης. Ουσιαστικά ως στόχο η μηχανική μάθηση έχει να πάρει μια σειρά από δεδομένα, να τα φέρει σε μια μορφή με την οποία θα μπορεί να τα επεξεργαστεί ένας αλγόριθμος και ο αλγόριθμος αυτός να δώσει τα επιθυμητά αποτελέσματα. Βασική προϋπόθεση είναι αυτός ο αλγόριθμος

να έχει δημιουργηθεί και εκπαιδευτεί από τον άνθρωπο έτσι ώστε να είναι σε θέση να ανταπεξέλθει στις απαιτήσεις του ζητούμενου προβλήματος. Ένας αλγόριθμος μηχανικής μάθησης μπορεί να διαφέρει ανάλογα με τον τρόπο με τον οποίο προσεγγίζει το πρόβλημα [9].

Αναλυτικότερα στόχος της μηχανικής μάθησης είναι η υλοποίηση μιας μηχανής η οποία θα είναι σε θέση να κατανοήσει και να αποστηθίσει μια σειρά από δεδομένα που αναφέρονται σε γεγονότα που ήδη υπάρχουν. Σκοπός αυτής της εκμάθησης της μηχανής είναι να καταφέρει να εξάγει συμπεράσματα από μια σειρά δεδομένων τα οποία θα ήταν αδύνατον να διαβαστούν χωρίς κάποιον αυτοματοποιημένο τρόπο. Σε αυτό το σημείο είναι που η μηχανική μάθηση μπορεί να συμβάλει αποτελεσματικά στην λειτουργία μιας επιχείρησης χρηματοπιστωτικού χαρακτήρα και να καταφέρει να προβλέψει τυχόν πιστωτικό κίνδυνο.

Γεννούνται δύο κατηγορίες μηχανικής μάθησης ανάλογα με την επιθυμητή έξοδο. Η πρώτη είναι η μάθηση με επίβλεψη, η οποία θεωρείται πολύ ειδικευμένη σαν περίπτωση καθώς λειτουργεί με τέτοιο τρόπο ώστε μια είσοδος στον αλγόριθμο να δίνει μία έξοδο. Η δεύτερη είναι η μάθηση χωρίς επίβλεψη. Σε αυτήν την περίπτωση η μέθοδος που ακολουθείται είναι αρκετά πιο γενικευμένη καθώς πρέπει να μπει σε διαδικασία δημιουργίας συνάρτησης από το σύνολο δεδομένων εκπαίδευσης. Αυτό στη συνέχεια προκαλεί και το πρόβλημα της ορθής ταξινόμησης των δεδομένων σε υποομάδες και συνήθως τελικός στόχος των αλγορίθμων τέτοιου τύπου είναι η αριθμητική πρόβλεψη.



Εικόνα 2: Διαδικασία αλγορίθμου πρόβλεψης πιστωτικού κινδύνου [10]

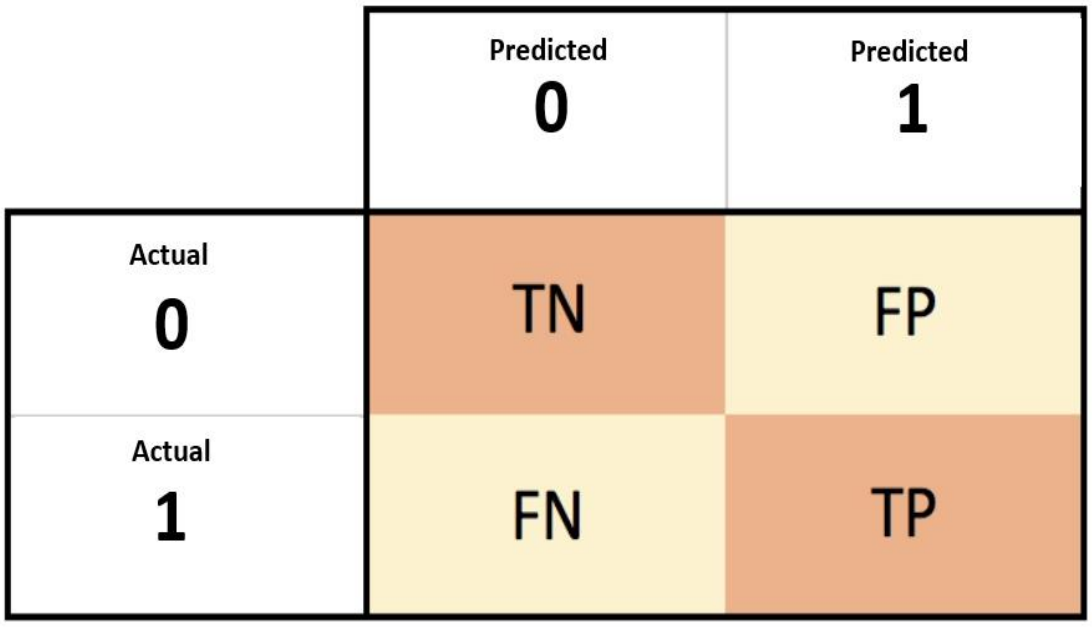
Συνοπτικά η διαδικασία η οποία ακολουθεί η μηχανική μάθηση είναι αρχικά η επεξεργασία των δεδομένων έτσι ώστε να είναι δομημένα σωστά. Έπειτα ακολουθεί η επιλογή των χαρακτηριστικών των δεδομένων που θα παίξουν καθοριστικό ρόλο στην εξαγωγή των αποτελεσμάτων. Στη συνέχεια γίνεται η εκμάθηση του αλγορίθμου από ένα σύνολο δεδομένων και τέλος η δοκιμή του αλγορίθμου για το αν μπορεί τελικά να καταφέρει να προβλέψει τον κίνδυνο να μην αποπληρωθεί κάποιο δάνειο. Όλη αυτή η διαδικασία έχει ως στόχο στο υποθετικό σενάριο που ένας καινούριος δανειολήπτης αιτηθεί κάποιο δάνειο να είναι σε θέση αυτός ο αυτοματοποιημένος αλγόριθμος να προβλέψει το αν θα υπάρξει πιστωτικός κίνδυνος αξιολογώντας τα δεδομένα του.

Εικόνα 3: Αξιολόγηση αλγορίθμου

Πιο συγκεκριμένα στην αποθήκη δεδομένων βρίσκονται μια σειρά από χαρακτηριστικά τα οποία περιέχουν δεδομένα δανείων. Τα δεδομένα αυτά αφού έρθουν σε επεξεργάσιμη μορφή για τον αλγόριθμο, μπορούν να μετατραπούν σε γνώση. Επίσης όλα αυτά τα δεδομένα μπορούν να περιέχουν τιμές αριθμητικές αλλά και αλφαριθμητικές. Οι αριθμητικές τιμές είναι εύκολο να επεξεργαστούν από τον αλγόριθμο, ενώ οι τιμές όπου περιέχουν αλφαριθμητικά θα πρέπει να υποστούν μια προεργασία. Η προεργασία αυτή συνήθως αφορά την μετατροπή των μεταβλητών αυτών σε κάποιον μοναδικό αριθμό. Έτσι ο αλγόριθμος κερδίζει χρόνο κατά την εκπαίδευση καθώς έχει να επεξεργαστεί τιμές που καταλαμβάνουν λιγότερη μνήμη από ότι μια ολόκληρη πρόταση. Μετά το πέρας της εκπαίδευσης και αφού δοθούν από τον χρήστη οι απαραίτητοι παράμετροι για την σωστή λειτουργία του αλγορίθμου, χρησιμοποιώντας ένα μέρος των δεδομένων που πλέον θεωρείται το γνωστό μέρος, είναι σε θέση να ξεκινήσει η αναγνώριση και για τα υπόλοιπα δεδομένα που δεν χρησιμοποιήθηκαν για την εκπαίδευση.

Αφού λοιπόν ο αλγόριθμος μπορεί να αναγνωρίσει τα δεδομένα δοκιμής, μπορεί και να εξάγει αποτελέσματα τα οποία θα αξιολογήσουν το κατά πόσο είναι ορθός και αξιόπιστος. Η αξιοπιστία του αλγορίθμου αφορά μια σειρά από επιτυχημένες προβλέψεις σε αντιδιαστολή με τις συνολικές που έγιναν. Αυτή η

διαδικασία μπορεί να απεικονιστεί μέσω του πίνακα σύγχυσης [11]. Είναι ένας πίνακας στον οποίο καταχωρούνται οι ορθές αλλά και οι λανθασμένες προβλέψεις του αλγορίθμου. Απεικονίζονται ουσιαστικά στον πίνακα οι προβλέψεις σε συνάρτηση με την πραγματική τιμή. Έτσι μπορεί να γίνει μια κατηγοριοποίηση των αποτελεσμάτων σε τέσσερις διαφορετικές καταστάσεις. Αρχικά έχουμε τα αληθώς θετικά που αφορούν τις προβλέψεις αυτές που ο αλγόριθμος προέβλεψε ορθά ότι το αποτέλεσμα είναι θετικό. Στην συνέχεια έχουμε τα αληθώς αρνητικά όπου αντίστοιχα πάλι ο αλγόριθμος προβλέπει σωστά αλλά αυτή τη φορά ένα αρνητικό αποτέλεσμα. Τρίτη και τέταρτη κατηγορία αφορούν τα ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα, τα οποία αφορούν τα λάθη που έκανε ο αλγόριθμος σε θετικά και αρνητικά αποτελέσματα αντίστοιχα. Αυτές οι τέσσερις τιμές είναι που αποδίδουν τελικά στον αλγόριθμο ένα ποσοστό επιτυχίας το οποίο θα αξιολογήσει τον αλγόριθμο για το αν έχει την δυνατότητα να προβλέψει σωστά τον πιστωτικό κίνδυνο.



Εικόνα 4: Πίνακας σύγχυσης [12]

Μέσα από τα αποτελέσματα του πίνακα σύγχυσης υπάρχει η δυνατότητα να εξαχθούν ποσοστά τα οποία δίνουν τα τελικά συμπεράσματα για τον αλγόριθμο. Η γενική ακρίβεια (accuracy) είναι το πιο απλό χαρακτηριστικό καθώς αναφέρεται στην ακρίβεια του αλγορίθμου αναφορικά με όλα το αποτελέσματα. Ο υπολογισμός της ακρίβειας αυτής είναι ο αριθμός των σωστών προβλέψεων του αλγορίθμου διά τον συνολικό αριθμό των

προβλέψεων. Η ακρίβεια (precision) είναι κάπως διαφορετική από την προηγούμενη. Αφορά συγκεκριμένη πιθανότητα που έχει να κάνει με την ορθότητα ένταξης μιας πρόβλεψης στην θετική κλάση. Συγκεκριμένα είναι ο λόγος των σωστών θετικών αποτελεσμάτων με το άθροισμα των σωστών θετικών με των λανθασμένων θετικών. Αυτή η ακρίβεια επηρεάζεται μονάχα από τις θετικές προβλέψεις είτε αυτές είναι ορθές είτε όχι. Τέλος υπάρχει και η ανάκληση (recall) η οποία αντιστρόφως με την ακρίβεια επηρεάζεται από τα λανθασμένα αρνητικά αποτελέσματα. Η ανάκληση δίνεται από το πηλίκο των σωστών θετικών δια το άθροισμα των σωστών θετικών με τα ψευδώς αρνητικά [13]. Η τιμή που επηρεάζει κατά πολύ την πρόβλεψη του πιστωτικού κινδύνου είναι η ακρίβεια (precision) καθώς μέσω αυτής μπορούν να αποφευχθούν τα ψευδώς θετικά αποτελέσματα τα οποία δίνουν άμεση ζημία στον χρηματοπιστωτικό φορέα. Έτσι είναι πολύ σημαντικό ο αλγόριθμος να είναι σε θέση να εξάγει όσο το δυνατόν πιο ακριβή αυτά τα αποτελέσματα.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Εικόνα 5: accuracy, precision,recall

Υπάρχουν διάφορες προσεγγίσεις για τους αλγορίθμους πρόβλεψης. Οι αλγόριθμοι μεταξύ τους διαφοροποιούνται ανάλογα με τον τρόπο τον οποίο θα διαχειριστούν ένα πρόβλημα [14]. Επιθυμητό σε έναν αλγόριθμο είναι να μπορεί να εξάγει αποτελέσματα σε μικρό χρονικό διάστημα και με την λιγότερη δυνατή επεξεργαστική δύναμη. Επίσης πρέπει να είναι σε θέση να επεξεργάζονται δεδομένα μεγάλου όγκου και να μπορούν να δώσουν μεγάλη ακρίβεια στα αποτελέσματα. Τέλος ο αλγόριθμος πρέπει να είναι εύκολα

παραμετροποιήσιμος από τον εκπαιδευτή του έτσι ώστε να κάνει όσο το δυνατόν πιο αυτόματη την διαδικασία. Μερικοί από τους βασικούς αλγορίθμους μηχανικής μάθησης αναφέρονται παρακάτω.

2.1. Αλγόριθμοι Παλινδρόμησης (Regression Algorithms)

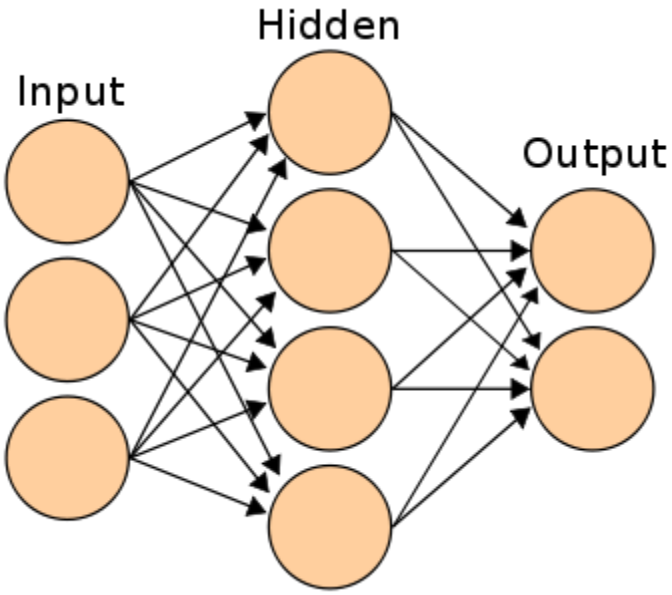
Παλινδρόμηση είναι μια σειρά από διαδικασίες οι οποίες σκοπό έχουν να αναδείξουν μέσα από τις μεταβλητές και τα δεδομένα την αξιολόγηση αυτών. Ουσιαστικά αφορά την σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών και το πώς αλληλεπιδρούν μεταξύ τους. Στόχος λοιπόν είναι να προβλεφθεί η τιμή της εξαρτημένης μεταβλητής σε συνάρτηση με τις τιμές των ανεξάρτητων μεταβλητών. Ανάλογα με τις μεταβλητές μπορεί ένα μοντέλο παλινδρόμησης να πάρει την μορφή γραμμικής ή πολλαπλής γραμμικής παλινδρόμησης. Ζητούμενα είναι η μεγαλύτερη δυνατή ακρίβεια αλλά και η ερμηνεία της διακύμανσης της τιμής των μεταβλητών με σκοπό τα πιο ασφαλή συμπεράσματα. Από τους πιο γνωστούς αλγορίθμους παλινδρόμησης είναι οι Logistic Regression και ο Linear Regression [9].

2.2. Αλγόριθμοι Κανονικοποίησης (Regularization Algorithms)

Η κανονικοποίηση έχει να κάνει με την επιλογή δεδομένων τα οποία δεν επηρεάζονται από πειραματικές συνθήκες σε συνάρτηση με μια σταθερά. Στις πιο απλές και σύνηθες περιπτώσεις, η κανονικοποίηση αφορά την προσαρμογή των τιμών που μετρώνται σε διαφορετικές κλίμακες σε μια θεωρητικά κοινή κλίμακα. Σε πιο δύσκολές και σπάνιες περιπτώσεις, η κανονικοποίηση μπορεί να αναφέρεται σε πιο εξελιγμένες καταστάσεις όπου σκοπός πλέον είναι η ολική ενσωμάτωση και κατανομή των πιθανοτήτων για τις προσαρμοσμένες τιμές. Ένας από τους πιο γνωστούς αλγορίθμους κανονικοποίησης είναι αυτός ο Tikhonov Regularization ο οποίος είναι γνωστός ως παλινδρόμηση κορυφής και χρησιμεύει για τον μετριασμό του προβλήματος της πολυγραμμικότητας στη γραμμική παλινδρόμηση που συναντάμε κυρίως σε προβλήματα με μεγάλο όγκο σε παραμέτρους [9].

2.2 Αλγόριθμοι Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Network Algorithms)

Η μέθοδος των τεχνητών νευρωνικών δικτύων έχει ονομαστεί έτσι καθώς εμπνεύστηκε από τον τρόπο λειτουργίας του εγκεφάλου. Δέχονται λοιπόν μια σειρά από σήματα τα οποία μπαίνουν σε μια διαδικασία επεξεργασίας με σκοπό την ομαδοποίησή τους σε νευρώνες που έχουν κοινά στοιχεία με αυτά. Σαν είσοδο μπορεί να έχουμε ένα σύνολο από αριθμούς οι οποίοι μετά από απαραίτητη διαδικασία φέρνει την έξοδο η οποία επήλθε αφού πέρασε από μια μη γραμμική συνάρτηση. Έχουμε λοιπόν ένα πλήθος στρωμάτων τα οποία απαρτίζουν έναν νευρώνα και επεξεργάζονται κάθε σήμα χωριστά, πάντα χρησιμοποιώντας το πρώτο στρώμα σαν είσοδο και το τελευταίο σαν έξοδο.



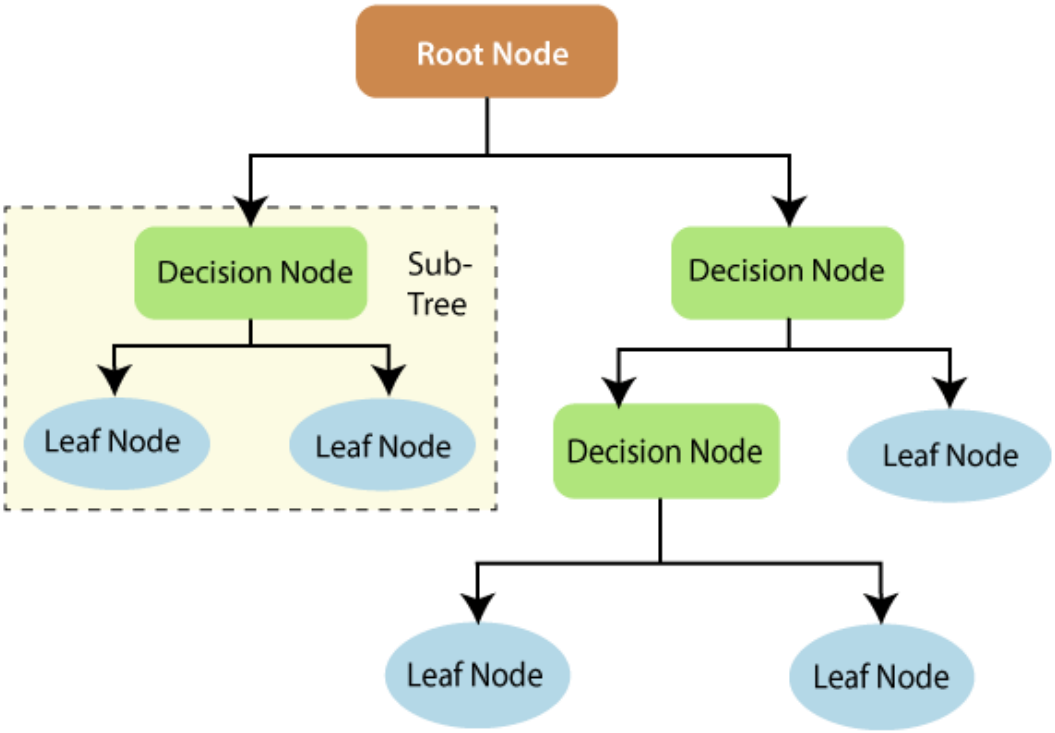
Εικόνα 6: Στρώματα νευρωνικού δικτύου [15]

Η βασική λειτουργία που κάνει ένα νευρωνικό δίκτυο αξιόπιστο είναι η εκπαίδευση. Το δίκτυο αποθηκεύει όλα τα δεδομένα από την είσοδο τους μέχρι την έξοδο προσπαθώντας να βρει κοινά στοιχεία σε κάθε προσπέλαση. Έτσι λοιπόν για να υπάρχει αξιοπιστία στα αποτελέσματα, στόχος είναι το νευρωνικό δίκτυο να είναι εκπαιδευμένο από τον δημιουργό του ώστε να έχει σαν δεδομένο ένα ή περισσότερα παραδείγματα που σχετίζονται με το δείγμα

και έτσι να είναι σε θέση να αντιληφθεί ποια είναι η πιθανή έξοδος, σε μία είσοδο δεδομένου που δεν υπήρξε ξανά. Ένας από τους πιο γνωστούς αλγορίθμους που χρησιμοποιούν την μέθοδο των τεχνιτών νευρωνικών δικτύων είναι ο Perceptron [9].

2.4. Εκμάθηση με Δέντρο Αποφάσεων (Decision Tree Learning)

Στην περίπτωση των δέντρων αποφάσεων έχουμε μια από τις πιο διαδεδομένες περιπτώσεις από αλγορίθμους καθώς θεωρείται ένας από τους καλύτερους τρόπους να κάνει κάποιος εξόρυξη δεδομένων. Επίσης χρησιμοποιούνται με μεγάλη επιτυχία και σε περιπτώσεις μηχανικής μάθησης αλλά και για στατιστικές αναλύσεις. Ο τρόπος που λειτουργεί έχει να κάνει με την δομή ενός δέντρου. Δηλαδή ξεκινώντας από την ρίζα του δέντρου με βάση κάποιες αποφάσεις περνάμε στα φύλλα όπου βρίσκονται τα συμπεράσματα.



Εικόνα 7: Δέντρο αποφάσεων[16]

Τα φύλλα λοιπόν είναι οι ετικέτες όπου φτάνουμε μετά από ένα ερώτημα το οποίο συνδέει τα φύλλα με την ρίζα. Γνωστή κατηγορία δέντρων είναι αυτή των δέντρων παλινδρόμησης. Είναι τα δέντρα που σαν μεταβλητές παίρνουν πραγματικούς αριθμούς. Σαν κατηγορία αλγορίθμων είναι αρκετά διαδεδομένοι κυρίως λόγω της απλότητας αλλά και για την επιτυχία που έχουν στην εξόρυξη δεδομένων. Η μέθοδος που χρησιμοποιείται είναι του «διαίρει και βασίλευε» γνωστή για την ταχύτητα που προσφέρει στους αλγορίθμους. Με αυτήν την τεχνική ένα δέντρο ξεκινάει να είναι μια σειρά από αποφάσεις που πρέπει να παρθούν και καταλήγει σε μια σειρά από συμπεράσματα.

Αναλυτικότερα δέντρο απόφασης ονομάζεται ένας ταξινομητής ο οποίος αποτελείται από μια σειρά από ετικέτες που δημιουργούν ένα δέντρο το οποίο έχει μία μοναδική ρίζα. Η ρίζα έχει μόνο εξόδους και καμία είσοδο ενώ οι ετικέτες έχουν μόνο μία είσοδο. Υπάρχουν βέβαια και οι ενδιάμεσες ετικέτες οι οποίες μπορούν να έχουν μία είσοδο αλλά παραπάνω από μία έξοδο. Έχουμε λοιπόν μια ταξινόμηση των γεγονότων με βάση την διαδρομή που ακολουθεί κάθε τερματική ετικέτα από την αρχική ρίζα. Ένα δέντρο απόφασης έχει την ευελιξία να μπορεί να ενσωματώνει και χαρακτήρες αλλά και αριθμούς. Έτσι ο αλγόριθμος μπορεί και επισημαίνει το κάθε χαρακτηριστικό μιας ετικέτας το οποίο το βοηθάει για να πάρει τις σωστές αποφάσεις.

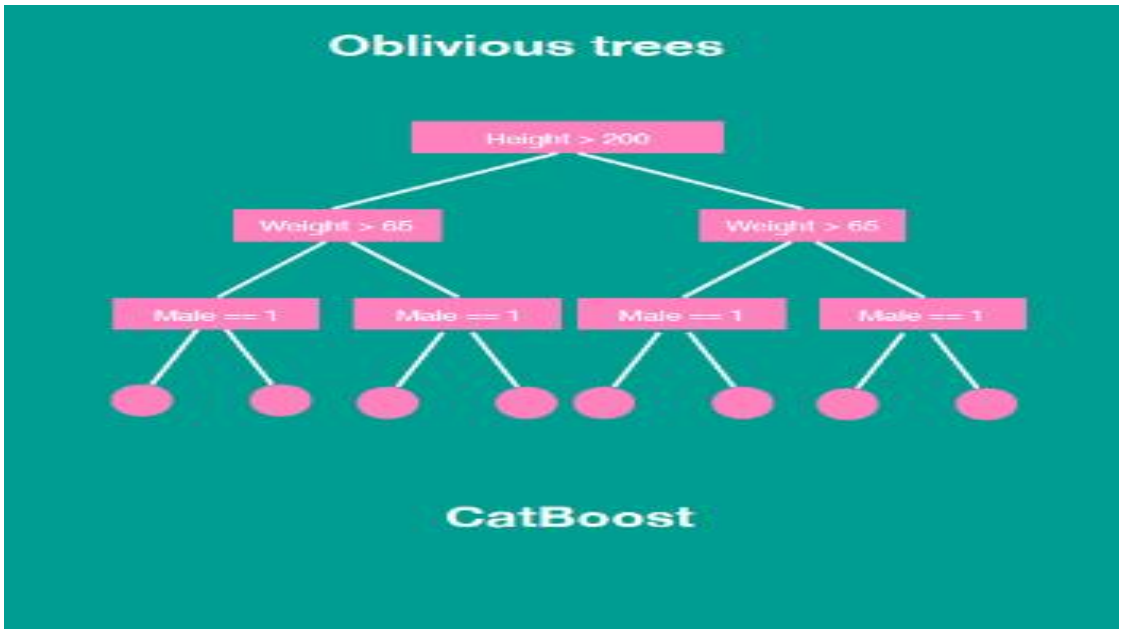
Ένας αλγόριθμος δέντρου απόφασης έχει την δυνατότητα έπειτα από επεξεργασία δεδομένων που μπορεί να βρίσκονται σε μία βάση, να εξάγει ένα ή περισσότερα συμπεράσματα. Χρησιμοποιώντας λοιπόν μια σειρά από αλγορίθμους οι οποίοι θα ταξινομήσουν και θα ομαδοποιήσουν τα δεδομένα μας αλλά και με την βοήθεια στατιστικών στοιχείων, έχουμε μία πληροφορία η οποία έχει δομή και υπόσταση τέτοια ώστε να είναι άμεσα κατανοητή και αξιόπιστη για να πάρουμε κάποιες αποφάσεις. Η διαδικασία που περιγράφεται παραπάνω ονομάζεται εξόρυξη δεδομένων. Με την εξόρυξη δεδομένων προσδοκούμε μια επεξεργασία μεγάλου όγκου δεδομένων τα οποία θα μας δώσουν πληροφορίες που δεν μπορούσαμε να έχουμε λόγω του μεγέθους της αποθήκης δεδομένων. Έτσι πέρα από συμπεράσματα με την βοήθεια της μηχανικής μάθησης υπάρχει η δυνατότητα της εκμάθησης. Η εκμάθηση γίνεται πάνω σε έναν αλγόριθμο ο οποίος με την σειρά του μας δίνει κάποιο αποτέλεσμα ή κάποια πρόβλεψη. Αυτή η διαδικασία μπορεί να

δώσει μια σχετική ασφάλεια στον κάτοχο της βάσης για μελλοντικές κινήσεις που θα θέλει να κάνει καθώς θα μπορεί να έχει μια πρόγνωση σχετικά με το αποτέλεσμα.

Έχοντας λοιπόν περάσει στην εποχή που η συλλογή δεδομένων αυξάνεται όλο και περισσότερο δημιουργείται η ανάγκη για χρήση τέτοιων τακτικών καθώς έχοντας μια δυσανάγνωστη και δυσπρόσιτη σε συμπεράσματα βάση δεδομένων το μόνο που μπορεί να την κάνει χρήσιμη είναι μια αυτόματη διαδικασία εξαγωγή αποτελεσμάτων. Η τεχνική της μηχανικής μάθησης για την εξόρυξη δεδομένων λειτουργεί με την επαγωγή. Χρησιμοποιείται ένα μέρος των δεδομένων για την εκπαίδευση του αλγορίθμου και στη συνέχεια τα υπόλοιπα δεδομένα έρχονται για να μας δώσουν την πρόβλεψη έχοντας ως κανόνα αυτά της εκπαίδευσης [14].

2.5. Ο Αλγόριθμος Catboost

Ένας από τους αλγορίθμους που χρησιμοποιούν την μέθοδο των δέντρων αποφάσεων είναι ο Catboost. Ο αλγόριθμος Catboost είναι σχετικά νέος στον χώρο της μηχανικής μάθησης και δημιουργήθηκε από την Yandex. [17] Χρησιμοποιείται αποκλειστικά για περιπτώσεις μηχανικής μάθησης και ξεχωρίζει λόγω της ταχύτητας που προσφέρει σε σχέση με άλλους αλγορίθμους ίδιου τύπου. Επίσης πέρα από ταχύτητα προσφέρει και ακρίβεια καθώς ο Catboost μπορεί να δώσει τα επιθυμητά αποτελέσματα τρέχοντάς τον μία φορά και όχι πολλές όπως απαιτούν άλλοι αλγόριθμοι. Γενικότερα μπορεί να χρησιμοποιηθεί για ταξινόμηση και πρόβλεψη, χαρακτηριστικά τα οποία χειρίζονται με επιτυχία αλγόριθμοι που χρησιμοποιούν τα δέντρα αποφάσεων.

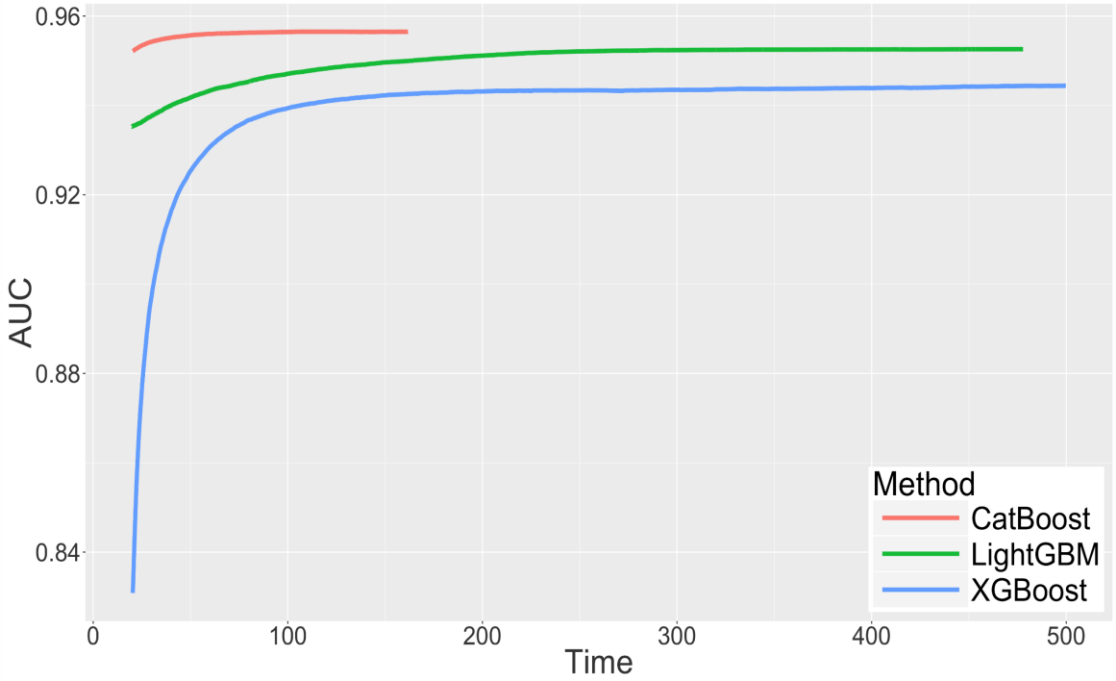


Εικόνα 8: Ο αλγόριθμος Catboost [18]

Ο Catboost πήρε το όνομά του από το “category” και το “boosting”. Ένα από τα χαρακτηριστικά που κάνουν τον αλγόριθμο αυτό να ξεχωρίζει είναι ότι μπορεί να χειριστεί με ευκολία τόσο αριθμητικές αλλά και κατηγορηματικές μεταβλητές. Ουσιαστικά ο τρόπος που λειτουργεί είναι να μετατρέπει αυτές τις μεταβλητές σε αριθμητικές και έτσι δεν χρειάζεται να γίνει κάποια ιδιαίτερη επεξεργασία για την μετατροπή τους από τον χρήστη. Για να γίνει αυτό, ο αλγόριθμος χρησιμοποιεί μια σειρά από στατιστικά στοιχεία. Επίσης έχει την ικανότητα να δίνει αξιόπιστα αποτελέσματα με πολύ λιγότερα δεδομένα από άλλους αλγορίθμους. Προσφέρει δηλαδή μεγάλη ποιότητα σε μικρότερο όγκο δεδομένων αλλά και μεγαλύτερη ταχύτητα σε μεγαλύτερο όγκο δεδομένων. Τέλος ένα μεγάλο θετικό του Catboost είναι ότι σαν αποτέλεσμα εξάγει γενικευμένα μοντέλα πράγμα το οποίο μειώνει την περίπτωση να εμφανίσει overfitting και έτσι τον κάνει ευέλικτο σε σχέση με άλλους. Όταν εμπλέκονται στατιστικά στοιχεία, το overfitting ή αλλιώς υπερβολική προσαρμογή μπορεί να επιφέρει αρνητικά αποτελέσματα στην εκπαίδευση του αλγορίθμου καθώς έτσι παράγει ένα αποτέλεσμα το οποίο μπορεί να ταιριάζει με ένα σύνολο δεδομένων, με αποτέλεσμα να μην είναι σε θέση πλέον να ξεχωρίζει ορθά όλα τα δεδομένα και να προβλέψει καινούριες και διαφορετικές παρατηρήσεις.

Με άλλα λόγια, το μοντέλο θυμάται έναν τεράστιο αριθμό παραδειγμάτων αντί να μαθαίνει να παρατηρεί χαρακτηριστικά. Έτσι λοιπόν ο χρήστης μπορεί

να προσαρμόσει τον αλγόριθμο για το πόσες επαναλήψεις θα κάνει, με την βοήθεια ενός ειδικού ανιχνευτή του overfitting που τον προειδοποιεί όταν αυτό ξεπερνάει τα όρια. Μέσα στα αρνητικά του Catboost είναι ότι έχει πολλές παραμέτρους που πρέπει να συντονιστούν από τον χρήστη όπως είναι ο ρυθμός της εκμάθησης, η κανονικοποίηση και ο αριθμός των δέντρων.



Εικόνα 9: Σύγκριση Catboost με άλλους αλγορίθμους ίδας κατηγορίας [19]

Στην παραπάνω εικόνα φαίνεται η αισθητή διαφορά του Catboost από τους LightGBM και XGBoost, αλγόριθμοι οι οποίοι είναι παρόμοιοι με τον Catboost. Ο αλγόριθμος Catboost προσφέρει μεγάλη διαφορά στην ταχύτητα εκμάθησης καθώς έχει μεγάλη ταχύτητα υποστήριξης από την GPU γεγονός που τον καθιστά από 2 έως 20 φορές πιο γρήγορο στην κατηγορία του. Αυτά λοιπόν τα χαρακτηριστικά τον καθιστούν έναν από τους πιο αξιόπιστους αλγορίθμους για να διαχειριστεί προβλήματα τα οποία αφορούν μεγάλο όγκο δεδομένων και χρίζουν ανάγκης για ασφαλή και αξιόπιστα αποτελέσματα.

3. Dataset

Το dataset που χρησιμοποιήθηκε δίνεται από τον ιστότοπο www.kaggle.com και αφορά τα δάνεια μιας μεγάλης εταιρείας δανεισμού με όνομα LendingClub. Είναι μια εταιρεία με έδρα στην Καλιφόρνια και έχει μέχρι σήμερα δώσει δάνεια δισεκατομμυρίων. Ο τρόπος που λειτουργεί η εταιρεία είναι μέσω μιας διαφορετικής οπτικής στα δάνεια. Πέρα από τον δανειολήπτη που μπορεί να πάρει ένα δάνειο όχι υπέρογκου ποσού, στην συναλλαγή μπορούν να επέμβουν και επενδυτές. Δίνεται λοιπόν η ευκαιρία σε αυτούς να επενδύσουν σε άτομα που πιστεύουν ότι θα αποπληρώσουν το δάνειο τους με κέρδος ένα μέρος από τους τόκους που θα εισπράξει η εταιρεία. Αυτή η κίνηση βέβαια χρεώνεται από την εταιρεία με ένα τέλος συμμετοχής. Η εταιρεία Lending club έφτασε να συγκεντρώσει πάνω από ένα δισεκατομμύριο δολάριο μέχρι το 2014, όμως τα πράγματα άλλαξαν 2 χρόνια μετά αφού παρατηρήθηκε μεγάλη μείωση στην επενδυτική κίνηση.

Έτσι λοιπόν με την βοήθεια της τεχνολογίας οι επενδύσεις σε αυτά τα δάνεια μπορούν μειώσουν αισθητά το ρίσκο στο να χάσει τα λεφτά του ο επενδυτής. Συγκεκριμένα με την μηχανική μάθηση και την χρήση του dataset των δανείων που έχουν ήδη προχωρήσει στην εταιρεία μέχρι σήμερα, υπάρχει η δυνατότητα της μελέτης και πρόβλεψης του ρίσκου να μην αποπληρωθεί ένα δάνειο.

3.1. Επεξεργασία των δεδομένων

Για την μελέτη των δανείων χρησιμοποιήθηκε το αρχείο database.sqlite που επεξεργάστηκε από έναν db browser για βάσεις δεδομένων τύπου sqlite. Επίσης για την επεξεργασία των δεδομένων χρησιμοποιήθηκε το αρχείο loan.csv το οποίο με την βοήθεια της γλώσσας προγραμματισμού python ήρθε στην επιθυμητή μορφή. Τα αρχεία αυτά περιέχουν πάνω από δύο εκατομμύρια δάνεια και τα στοιχεία των δανειοληπτών. Τέλος η χρήση των εντολών έγινε μέσω του jupyter notebook του anaconda.

Το πρώτο βήμα που πρέπει να γίνει είναι η ανάγνωση αλλά και η επεξεργασία των δεδομένων. Το αρχείο τύπου csv είναι εύκολο να

αναγνωριστεί και να επεξεργαστεί, όμως υπάρχουν αρκετά δεδομένα τα οποία χρειάζονται επεξεργασία για να είναι σε μορφή η οποία θα μπορεί να θέσει τον αλγόριθμο να κάνει υπολογισμούς.

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	hardship_payoff_balance_amount	hardshi
0	NaN	NaN	2500	2500	2500.0	36 months	13.56	84.92	C	C1	...	NaN	
1	NaN	NaN	30000	30000	30000.0	60 months	18.94	777.23	D	D2	...	NaN	
2	NaN	NaN	5000	5000	5000.0	36 months	17.97	180.69	D	D1	...	NaN	
3	NaN	NaN	4000	4000	4000.0	36 months	18.94	146.51	D	D2	...	NaN	
4	NaN	NaN	30000	30000	30000.0	60 months	16.14	731.78	C	C4	...	NaN	

5 rows × 145 columns

Εικόνα 10: φόρτωμα των δεδομένων από το αρχείο csv

Για να αποφευχθεί ο μεγάλος όγκος των δεδομένων και η αύξηση της ταχύτητας προσπέλασης γίνεται μετατροπή του πεδίου issue_d από πεδίο τύπου αλφαριθμητικό σε πεδίο τύπου ημερομηνίας και γίνεται διαχωρισμός σε συγκεκριμένο χρονικό διάστημα που θα χρησιμοποιηθεί.

	issue_d
1	Dec-2018
2	Dec-2018
3	Dec-2018
4	Dec-2018
5	Dec-2018
6	Dec-2018
7	Dec-2018
8	Dec-2018
9	Dec-2018
10	Dec-2018

Εικόνα 11: Αλλαγή του issue_d σε format ημερομηνίας

Επόμενο βήμα είναι να φορτώσουμε τις επικεφαλίδες των πεδίων όπως δίνονται από το excel LCDataDictionary. Συγκρίνοντας τα πεδία της βάσης με τα πεδία του αρχείου βλέπουμε αρκετές διαφορές οι οποίες πρέπει να έρθουν ακριβώς στην ίδια μορφή για να μπορεί να γίνει σωστά η εκπαίδευση. Πρώτη κίνηση είναι να φέρουμε τον τύπο των γραμμάτων στην ίδια μορφή σε ότι αφορά το αν είναι κεφαλαία ή πεζά. Στη συνέχεια βλέπουμε ότι ακόμα υπάρχουν διαφορές ανάμεσα στα πεδία τα οποία και πρέπει να διορθωθούν.

```
np.setdiff1d(dict_final, my_data_cols)
```

```
Out[4]: array(['accept_d', 'credit_pull_d', 'effective_int_rate', 'exp_d',
       'exp_default_rate', 'fico_range_high', 'fico_range_low',
       'ils_exp_d', 'is_inc_v', 'list_d', 'msa',
       'mths_since_most_recent_inq', 'mths_since_oldest_il_open',
       'mths_since_recent_loan_delinq', 'review_status',
       'review_status_d', 'sec_app_fico_range_high',
       'sec_app_fico_range_low', 'service_fee_rate',
       'verified_status_joint'], dtype='<U35')
```

```
In [5]: np.setdiff1d(my_data_cols, dict_final)
```

```
Out[5]: array(['collection_recovery_fee', 'debt_settlement_flag',
       'debt_settlement_flag_date', 'deferral_term', 'funded_amnt_inv',
       'hardship_amount', 'hardship_dpd', 'hardship_end_date',
       'hardship_flag', 'hardship_last_payment_amount', 'hardship_length',
       'hardship_loan_status', 'hardship_payoff_balance_amount',
       'hardship_reason', 'hardship_start_date', 'hardship_status',
       'hardship_type', 'issue_d', 'last_credit_pull_d',
       'last_pymnt_amnt', 'last_pymnt_d', 'loan_status',
       'mo_sin_old_il_acct', 'mths_since_recent_bc_dlq',
       'mths_since_recent_inq', 'next_pymnt_d',
       'orig_projected_additional_accrued_interest', 'out_prncp',
       'out_prncp_inv', 'payment_plan_start_date', 'policy_code',
       'pymnt_plan', 'recoveries', 'settlement_amount', 'settlement_date',
       'settlement_percentage', 'settlement_status', 'settlement_term',
       'total_pymnt', 'total_pymnt_inv', 'total_rec_int',
       'total_rec_late_fee', 'total_rec_prncp', 'verification_status',
       'verification_status_joint'], dtype=object)
```

Εικόνα 12: Διαφορές στις στήλες των δεδομένων

Αφού φέρουμε τις στήλες στην μορφή που θέλουμε μπορούμε να ασχοληθούμε με τα δεδομένα που μας έχουν απομείνει. Είναι σημαντικό να έχουμε όλα τα δεδομένα σε δομή επεξεργάσιμη από τον αλγόριθμο. Για να το επιτύχουμε αυτό θέλουμε όλα τα χαρακτηριστικά που έχουμε μέσα στο αρχείο να είναι σε κατάσταση τέτοια ώστε να είναι εύκολα συγκρίσιμα μεταξύ τους. Για παράδειγμα όταν έχουμε σαν χαρακτηριστικό σε ένα πεδίο, μια σειρά από ημερομηνίες τότε είναι σημαντικό και ο τύπος που χρησιμοποιείται στο πεδίο

να είναι ημερομηνία και όχι αλφαριθμητικό. Έτσι πρέπει να μεταβάλουμε στη σωστή μορφή τα πεδία earliest_cr_line και sec_app_earliest_cr_line που πρέπει να είναι τύπου ημερομηνίας.

	earliest_cr_line	sec_app_earliest_cr_line
13	Jun-2003	
14	Oct-2008	
15	Jul-1990	
16	Dec-1988	May-2002
17	Dec-2002	
18	Oct-2010	Aug-2003
19	Jul-2005	
20	Feb-2001	
21	Dec-2004	
22	Oct-2001	

Εικόνα 13: Αλλαγή πεδίων στην κατάλληλη μορφή προς επεξεργασία

Επόμενη δυσλειτουργία στην εκπαίδευση του αλγορίθμου βλέπουμε ότι θα φέρει το πεδίο emp_length. Στην βάση βρίσκουμε τις τιμές <1 year και 10+ years τα οποία είναι λογικό ότι δεν γίνεται να χρησιμοποιηθούν. Για να μπορεί ο αλγόριθμος να συγκρίνει τις τιμές θα πρέπει αυτές να είναι ίδιου τύπου. Συγκεκριμένα δεν είναι δυνατόν να συγκρίνει έναν ακέραιο αριθμό με ένα αλφαριθμητικό γι αυτό θα πρέπει να έρθουν σε μορφή ακεραίων όλα τα δεδομένα. Όποτε πρέπει να αλλαχτούν και να γίνουν αριθμοί ώστε να ταιριάζουν με τα υπόλοιπα πεδία της βάσης. Τα αλλάζουμε σε 0 και 11 αντίστοιχα.

	emp_length
1	10+ years
2	10+ years
3	6 years
4	10+ years
5	10+ years
6	10+ years
7	4 years
8	10+ years
9	10+ years
10	< 1 year

Εικόνα 14: Αλλαγή χαρακτηριστικών του πεδίου emp_length

3.2 Επιλογή χαρακτηριστικών

Αφού έχει γίνει η απαραίτητη προεπεξεργασία των δεδομένων επόμενη διαδικασία είναι να επιλέξουμε τα χαρακτηριστικά εκείνα τα οποία μας είναι απαραίτητα για την σωστή εκπαίδευση του αλγορίθμου. Ξεκινώντας με την επιλογή των πεδίων που θα κρατήσουμε είναι πολύ σημαντικό να μην έχουμε δεδομένα τα οποία να παραπέμπουν σε κενά από τιμές. Αυτά τα πεδία δεν χρειάζονται από την στιγμή που δεν θα επιφέρουν καμία απολύτως αλλαγή. Οπότε εξαιρούμε τα πεδία desc, member_id, id, url που φαίνεται παρακάτω ότι είναι τελείως κενά από δεδομένα.

```
pos_data = X.isna().mean()
pos_data = pos_data[pos_data != 0].sort_values()
pos_data
```

Out[9]:

```
pct_tl_nvr_dlq                  0.000004
avg_cur_bal                     0.000081
all_util                        0.000260
revol_util                      0.001195
dti                             0.002286
mths_since_recent_bc            0.012515
bc_open_to_buy                  0.013303
percent_bc_gt_75                0.013319
bc_util                         0.013737
num_tl_120dpd_2m                0.025046
mths_since_rcnt_il              0.037174
mo_sin_old_il_acct              0.037174
emp_length                      0.084781
emp_title                       0.110368
mths_since_recent_inq           0.123788
il_util                         0.163201
mths_since_last_delinq          0.558620
mths_since_recent_revol_delinq  0.711878
mths_since_last_major_derog     0.768128
mths_since_recent_bc_dlq        0.801895
sec_app_earliest_cr_line        0.860704
sec_app_open_act_il             0.860704
dti_joint                       0.860704
sec_app_open_acc                0.860704
revol_bal_joint                 0.860704
annual_inc_joint                0.860704
sec_app_chargeoff_within_12_mths 0.860704
sec_app_collections_12_mths_ex_med 0.860704
sec_app_num_rev_accts           0.860704
sec_app_inq_last_6mths          0.860704
sec_app_mort_acc                0.860704
sec_app_revol_util              0.863121
verification_status_joint       0.870748
mths_since_last_record          0.872822
sec_app_mths_since_last_major_derog 0.954816
url                             1.000000
desc                            1.000000
id                              1.000000
member_id                       1.000000
dtype: float64
```

Εικόνα 15: Πεδία με κενές τιμές

Πέρα από τις περιπτώσεις όπου είχαμε τελείως κενές τις τιμές, θα πρέπει να διορθωθούν και όλες οι υπόλοιπες περιπτώσεις με τιμές. Έχουμε δύο διαφορετικές εκδοχές. Η πρώτη είναι η περίπτωση του αλφαριθμητικού η

οποία είναι και η πιο εύκολη καθώς την γεμίζουμε με ένα κενό χωρίς να επηρεάζει τα αποτελέσματα άμεσα. Η δεύτερη περίπτωση είναι να έχουμε κενά σε πεδία αριθμητικού τύπου. Εδώ δημιουργείται το θέμα ότι οι τιμές που θα επιλέξουμε θα επηρεάσουν σημαντικά τα αποτελέσματα οπότε δεν μπορούμε να βάλουμε τυχαίες τιμές. Έτσι τα πεδία αυτά χωρίζονται σε δύο υποπεριπτώσεις. Είναι τα πεδία τα οποία θα πάρουν την μέγιστη τιμή και τα πεδία που θα πάρουν την ελάχιστη τιμή. Μέγιστη τιμή θα πάρουν τα πεδία τα οποία μπορούμε να υποθέσουμε ότι λείπουν διότι δεν έχουν συμπληρωθεί ακόμα επειδή δεν έχει πραγματοποιηθεί η συνθήκη αλλά όταν πραγματοποιηθεί θα είναι ο μέγιστος δυνατός αριθμός.

	mths_since_last_delinq
1	
2	71
3	
4	
5	
6	
7	
8	
9	32
10	17

Εικόνα 16: Παράδειγμα πεδίου μέγιστης τιμής

Για παράδειγμα βλέπουμε πολλά πεδία που αναφέρουν τους μήνες τους οποίους δεν έχει πραγματοποιηθεί κάτι. Είναι κενά γιατί δεν έχει έρθει ακόμα η στιγμή της πραγματοποίησης της αναμενόμενης συνθήκης. Οπότε σε αυτές τις περιπτώσεις η τιμή θα είναι η μέγιστη δυνατή. Στην αντίθετη περίπτωση παίρνουμε την ελάχιστη τιμή. Σε πεδία όπως η εργασιακή εμπειρία που έχουμε κενό κελί τότε καταλαβαίνουμε ότι ο δανειολήπτης δεν έχει εργασιακή εμπειρία άρα γι αυτό και είναι κενό το κελί. Σε τέτοιες περιπτώσεις χρησιμοποιούμε την ελάχιστη τιμή.

```
pos_data = X.isna().mean()
pos_data = pos_data[pos_data != 0].sort_values()
pos_data
```

```
Out[12]: Series([], dtype: float64)
```

Εικόνα 17: Πεδία με τιμές σε όλα τα κελιά

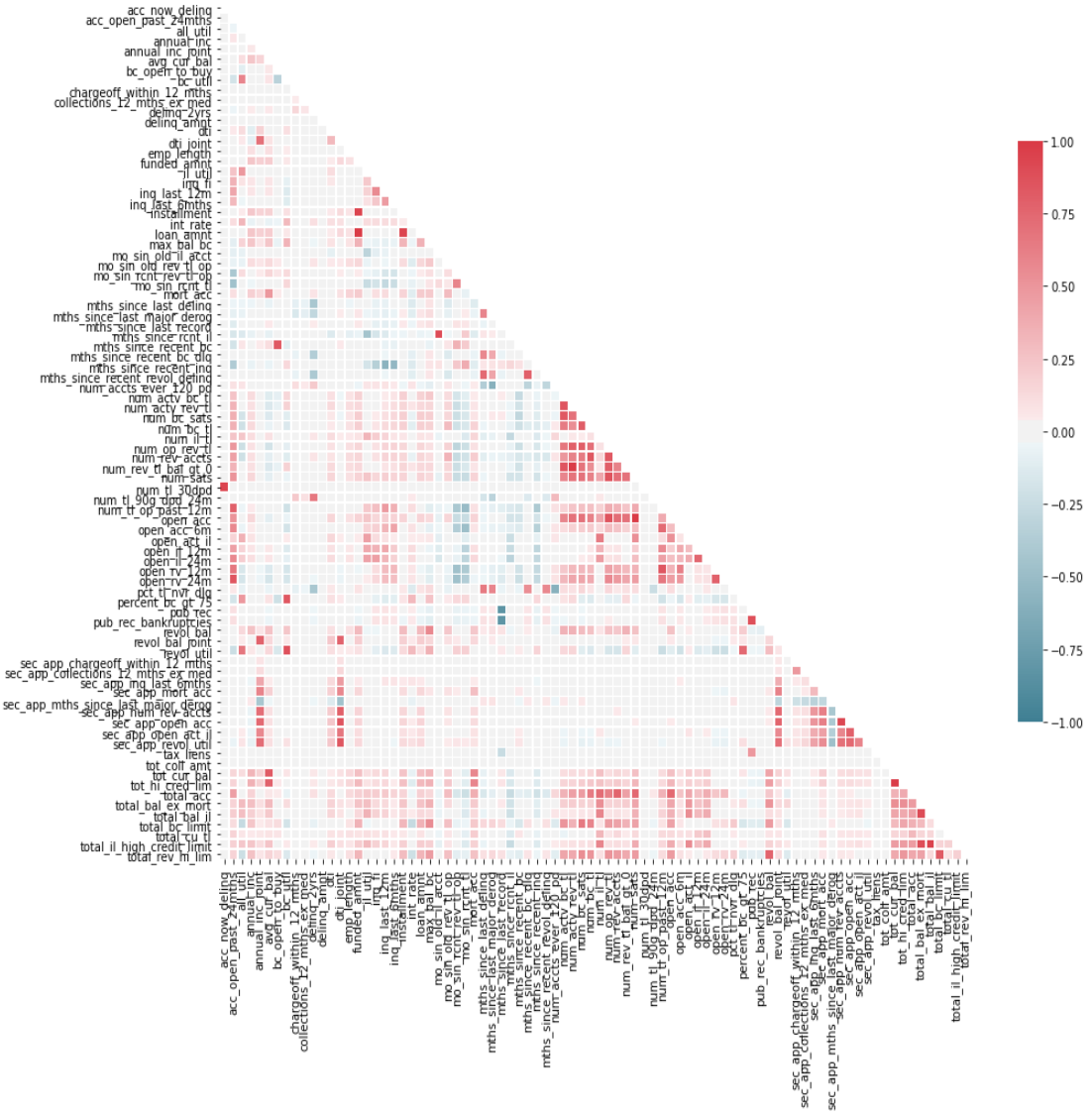
Πολύ σημαντικό για την ορθότητα των αποτελεσμάτων είναι να μην υπάρχουν στήλες με παρόμοια δεδομένα με άλλες στήλες. Είναι πολύ σημαντικό κάθε στήλη να αντιπροσωπεύει κάτι που δεν σχετίζεται ή ταυτίζεται με άλλες στήλες. Αν δεν αποφευχθεί αυτό τότε ο αλγόριθμος μπορεί εύκολα να βγάλει λάθος συμπεράσματα καθώς οι τιμές που θα βλέπει σε διαφορετικές στήλες θα παραπέμπουν στο ίδιο ακριβώς συμπέρασμα. Όποτε είναι σημαντικό κάθε στήλη με δεδομένα να προσφέρει πληροφορία χρήσιμη και να μην είναι βασιζόμενη σε άλλες. Αρχικά όπως φαίνεται στην παρακάτω εικόνα το στοιχείο num_tl_120dpd_2m έχει μία και μοναδική τιμή. Αυτό σαν δεδομένο δεν μας χρειάζεται καθώς δεν μας δίνει καμία διαφορετικότητα άρα το αφαιρούμε.

```
numbers = X.select_dtypes('number').columns.values
X[numbers].nunique().sort_values()
```

```
num_tl_120dpd_2m                1
acc_now_delinq                  2
num_tl_30dpd                    2
inq_last_6mths                  6
sec_app_inq_last_6mths          7
                              ...
total_bal_il                115553
total_il_high_credit_limit  128691
total_bal_ex_mort           139900
tot_cur_bal                 254794
tot_hi_cred_lim             266282
Length: 85, dtype: int64
```

Εικόνα 18: Απαλοιφή στήλης με ένα μονο δεδομένο

Στη συνέχεια πρέπει να εντοπιστούν τα πεδία εκείνα που παρουσιάζουν ομοιότητες και να μείνει ένα μόνο πεδίο από κάθε κατηγορία. Η διαδικασία αυτή γίνεται και για στήλες που περιέχουν αριθμούς αλλά και για αυτές που περιέχουν αλφαριθμητικά. Είναι πολύ σημαντικό να μην υπάρχουν κοινά στοιχεία στις στήλες της αποθήκης δεδομένων καθώς έτσι υπάρχει κίνδυνος ο αλγόριθμος να μην φέρει σωστά αποτελέσματα.



Εικόνα 19: Πεδία με ομοιότητες

Παραπάνω έχουμε μια απεικόνιση των στηλών οι οποίες περιέχουν αριθμούς. Δημιουργείται μια κλίμακα η οποία δείχνει πόσο πολύ μοιάζουν τα πεδία της βάσης σε ότι αφορά το περιεχόμενό τους. Όπως φαίνεται τα σημεία όπου συναντιούνται δύο πεδία έχουν διάφορα χρώματα. Όσο πιο κόκκινο

είναι ένα σημείο τόσο πιο πολλές ομοιότητες συναντάμε ανάμεσα στα δύο αυτά πεδία. Για να πετύχουμε το καλύτερο δυνατό αποτέλεσμα χρησιμοποιούμε το 0.9 σαν αριθμό από την κλίμακα ομοιότητας. Αυτό πρακτικά σημαίνει ότι όσα πεδία είναι όμοια κατά 90% τότε το ένα από τα δύο πεδία θα αγνοηθούν από τον αλγόριθμο.

```
zeugaria = idioi_arithmoi[np.abs(sostoi_ar) >= 0.9]
zeugaria
```

```
array([['acc_now_delinq', 'num_tl_30dpd'],
       ['funded_amnt', 'installment'],
       ['funded_amnt', 'loan_amnt'],
       ['installment', 'loan_amnt'],
       ['mo_sin_old_il_acct', 'mths_since_rcnt_il'],
       ['num_actv_rev_tl', 'num_rev_tl_bal_gt_0'],
       ['num_sats', 'open_acc'],
       ['sec_app_num_rev_accts', 'sec_app_open_acc'],
       ['tot_cur_bal', 'tot_hi_cred_lim'],
       ['total_bal_ex_mort', 'total_bal_il'],
       ['total_bal_il', 'total_il_high_credit_limit']], dtype='<U35')
```

Εικόνα 20: Ζευγάρια με μεγάλη ομοιότητα μεταξύ τους

Η ίδια διαδικασία πρέπει να γίνει και τα μη αριθμητικά πεδία. Βλέποντας αυτά τα πεδία φαίνεται το emp_title να φέρει πάρα πολλά αποτελέσματα. Πέρα από την περίπτωση όπου ένα πεδίο εμφανίζει μόνο μία τιμή, αρνητική επιρροή για τον αλγόριθμο έχουν και οι πολλές διαφορετικές τιμές. Όταν ένα πεδίο δεν παρουσιάζει ομοιότητες έτσι ώστε να συγκριθούν τα περιεχόμενα μεταξύ τους, τότε δεν μπορεί να βοηθήσει στην διαδικασία οπότε και θα αγνοηθεί και αυτό.

```
strings = X.select_dtypes('object').columns.values
X[strings].nunique().sort_values()
```

```
application_type              2
disbursement_method           2
initial_list_status           2
term                          2
verification_status           3
home_ownership                4
verification_status_joint     4
grade                         7
title                        12
purpose                      13
sub_grade                    35
addr_state                   50
zip_code                    897
emp_title                129450
dtype: int64
```

Εικόνα 21: Περιεχόμενα αλφαριθμητικών πεδίων

Αντίστοιχα με τους αριθμούς θα πρέπει να έχουμε μοναδικές στήλες που αφορούν ίδιο περιεχόμενο. Στην περίπτωση των αλφαριθμητικών πεδίων όμως η διαδικασία είναι χειροκίνητη καθώς οι ομοιότητες δεν έχουν να κάνουν μόνο με το περιεχόμενο των πεδίων αλλά και με το τι αντιπροσωπεύουν αυτά. Έτσι σχηματίζονται τα εξής ζευγάρια: zip_code με addr_state, title με purpose, grade με sub_grade, application_type με verification_status_joint. Το πρώτο ζευγάρι ουσιαστικά αφορά το ίδιο ακριβώς περιεχόμενο. Δεν θέλουμε να έχουμε πεδία που θα μας δώσουν την ίδια πληροφορία δύο φορές. Ίδια περίπτωση είναι και το τρίτο ζευγάρι καθώς και το δεύτερο που εκτός από την ίδια πληροφορία έχει και ίδιο ακριβώς περιεχόμενο.

```
In [47]: pd.value_counts(my_data.title).to_frame()
```

Out[47]:

	title
Debt consolidation	259642
Credit card refinancing	127702
Other	35023
Home improvement	32748
Major purchase	11622
Medical expenses	6622
Home buying	5430
Car financing	4979
Business	4583
Vacation	3501
Moving and relocation	3115
Green loan	275

```
In [46]: pd.value_counts(my_data.purpose).to_frame()
```

Out[46]:

	purpose
debt_consolidation	259642
credit_card	127702
other	35018
home_improvement	32748
major_purchase	11622
medical	6622
house	5430
car	4979
small_business	4583
vacation	3501
moving	3115
renewable_energy	275
wedding	5

Εικόνα 22: Ομοιότητα ανάμεσα σε purpose και title

Στο τέταρτο ζευγάρι η τιμή joint app είναι αυτή που καθορίζει τις τιμές του verification_status_joint. Όπως φαίνεται στην εικόνα παρακάτω υπάρχει εξάρτηση μεταξύ των δύο πεδίων η οποία δεν πρέπει να υφίσταται καθώς έτσι ο αλγόριθμος θα λαμβάνει τιμές και από τις δύο περιπτώσεις χαρακτηριστικών, χωρίς όμως να είναι δύο ανεξάρτητες τιμές. Έτσι είναι σημαντικό η μία από τις δύο στήλες να μην ληφθεί υπ' όψιν. Όποτε από κάθε ζευγάρι θα αγνοηθεί το ένα από τα δύο μέλη. Συγκεκριμένα μένουν εκτός τα: zip_code, verification_status_joint, sub_grade και title.

	application_type	verification_status_joint
13	Individual	
14	Individual	
15	Individual	
16	Joint App	Verified
17	Individual	
18	Joint App	Not Verified
19	Individual	
20	Individual	
21	Individual	
22	Individual	

Εικόνα 23: Εξάρτηση application_type με verification_status_joint

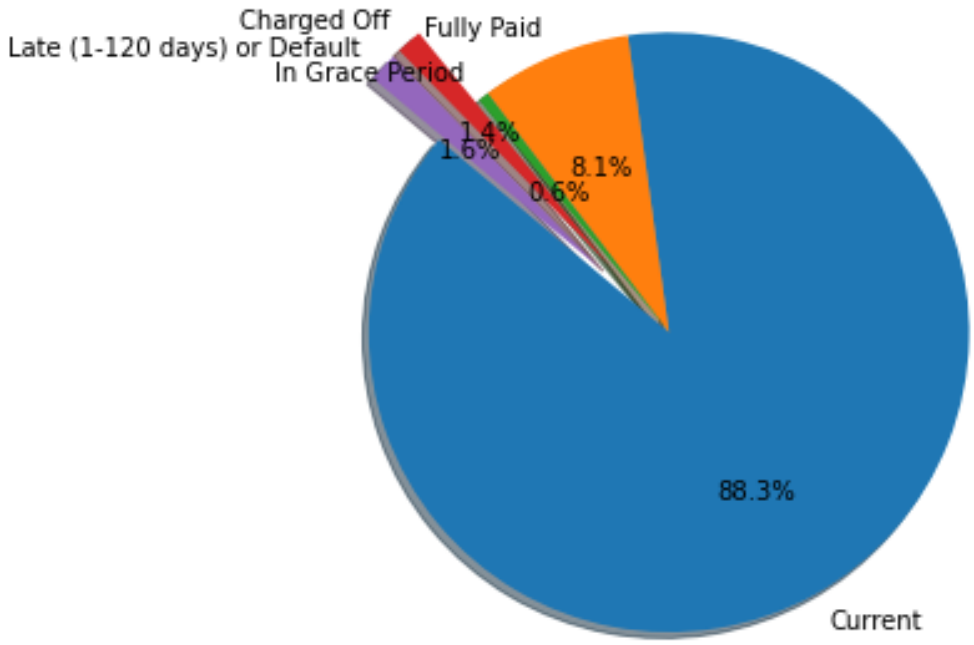
Αφού λοιπόν τελειώσει η διαδικασία της επεξεργασίας των χαρακτηριστικών αλλά και της επιλογής αυτών που θα χρησιμοποιηθούν, πρέπει να γίνει η ομαδοποίηση για τον τρόπο προσέγγισης των δανείων. Έχοντας εφτά διαφορετικές καταστάσεις δανείων πρέπει να χωριστούν ανάλογα με το αν θεωρούνται καλά δάνεια, δηλαδή αυτά που έχουν αποπληρωθεί ή που έχουν καλές ελπίδες να αποπληρωθούν, αλλά και τα κακά δάνεια που αφορούν τις υπόλοιπες κατηγορίες.

```
my_data['loan_status'].value_counts()
```

```
Current                437318
Fully Paid              40240
Charged Off              6942
Late (31-120 days)       6509
In Grace Period          2901
Late (16-30 days)        1323
Default                     9
Name: loan_status, dtype: int64
```

Εικόνα 24: Πλήθος δανείων ανά κατηγορία

Σύμφωνα με τις κατηγορίες που φαίνονται παραπάνω η ομαδοποίηση είναι λογικό να φέρει στα καλά δάνεια αυτά που έχουν αποπληρωθεί, αυτά που είναι τρέχοντα και αυτά που είναι σε περίοδο χάριτος. Όλα τα υπόλοιπα ανήκουν στα κακά δάνεια, δηλαδή αυτά που είτε δεν αποπληρώθηκαν ή έχουν καθυστερήσει να αποπληρωθούν.



Εικόνα 25:Διαχωρισμός ανάλογα με τον τύπο του δανείου

Όπως φαίνεται και στο γράφημα παραπάνω ο διαχωρισμός των δανείων γίνεται με βάση τον τύπο τους. Ως κακά δάνεια επιλέγονται αυτά που έχουν κάποια καθυστέρηση στην αποπληρωμή τους. Τα default είναι τα δάνεια που έχουν αργήσει να κινηθούν από 121 έως 150 μέρες, ενώ τα charged off είναι αυτά τα οποία αφορούν πάνω από 150 μέρες τα οποία δεν υπολογίζονται πλέον σαν δάνεια που μπορούν να αποπληρωθούν.

```
y = my_data['loan_status'].copy()
y = y.isin(['Current', 'Fully Paid', 'In Grace Period']).astype('int')
y.value_counts()
```

```
1    480459
0     14783
Name: loan_status, dtype: int64
```

Εικόνα 26: Πλήθος δανείων μετά τον διαχωρισμό

Έχοντας πλέον ολοκληρώσει και την ομαδοποίηση των δανείων, υπάρχει η τελική εικόνα των δεδομένων. Στη φάση που έχουμε φτάσει τα δεδομένα μας αποτελούνται από 495.242 δάνεια. Όπως φαίνεται και στην εικόνα παρακάτω, η επεξεργασία των δεδομένων έφερε το τελικό αποτέλεσμα και έχουμε τα πρώτα στατιστικά στοιχεία των χαρακτηριστικών μας.

	acc_open_past_24mths	all_util	annual_inc	annual_inc_joint	avg_cur_bal	bc_open_to_buy	bc_util	chargeoff_within_12_mths	collec
count	495242.000000	495242.000000	4.952420e+05	4.952420e+05	495242.000000	495242.000000	495242.000000	495242.000000	
mean	4.426622	54.073689	8.009399e+04	2.277788e+04	13708.241403	22918.296134	49.182116	0.006823	
std	3.234548	21.042735	8.887161e+04	5.285770e+04	17229.296185	70431.862928	29.384073	0.093093	
min	0.000000	0.000000	0.000000e+00	5.693510e+03	0.000000	0.000000	0.000000	0.000000	
25%	2.000000	40.000000	4.600000e+04	5.693510e+03	2904.000000	2924.000000	25.000000	0.000000	
50%	4.000000	55.000000	6.600000e+04	5.693510e+03	7032.000000	8500.000000	48.100000	0.000000	
75%	6.000000	69.000000	9.600000e+04	5.693510e+03	19029.750000	20410.000000	74.000000	0.000000	
max	54.000000	239.000000	9.930475e+06	7.874821e+06	623229.000000	605996.000000	201.600000	9.000000	

8 rows × 74 columns

Εικόνα 27: Στατιστική εικόνα

4. Πειραματικά Αποτελέσματα

Για την προσέγγιση του αλγορίθμου θα χρησιμοποιήσουμε όχι όλα τα δεδομένα μας αλλά μια ομάδα αυτών. Όπως φαίνεται από το χαρακτηριστικό grade τα δάνεια χωρίζονται από βαθμίδες οι οποίες τα κατατάσσουν ανάλογα με την πιθανότητα που έχουν να αποπληρωθούν. Ουσιαστικά ο βαθμός έχει να κάνει με την επικινδυνότητα στην αποπληρωμή του δανείου. Οι βαθμοί είναι από τα A έως το G όπου A είναι ο μικρότερος κίνδυνος και G ο μεγαλύτερος.

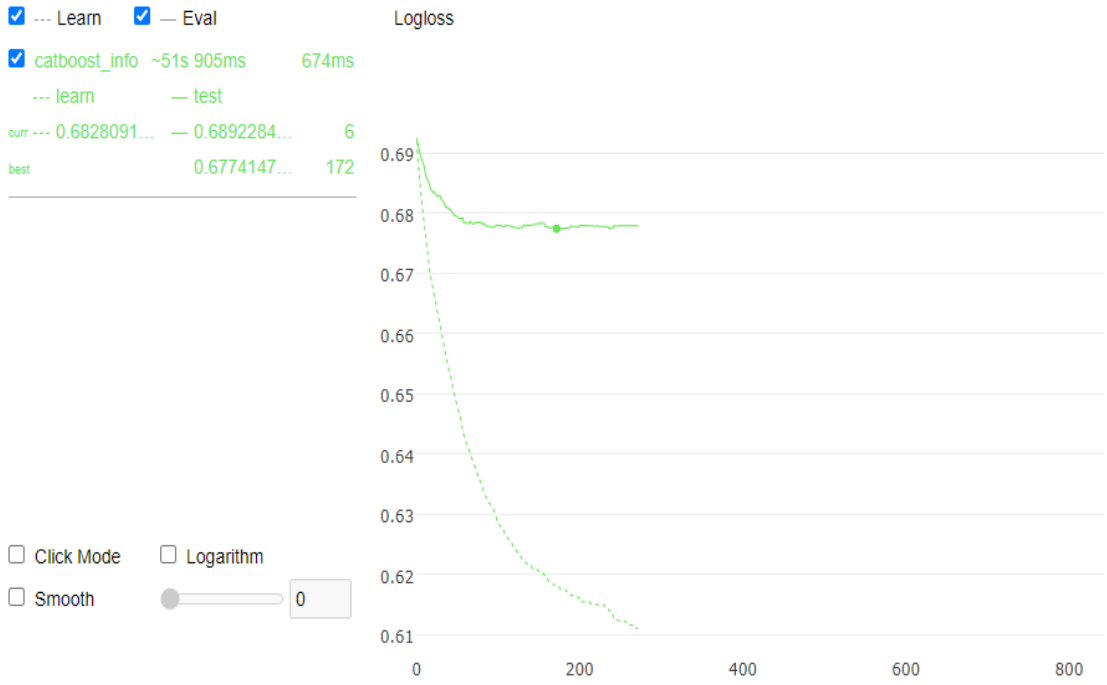
grade loan_status	A	B	C	D	E	F	G
Charged Off	518	1400	2126	1971	706	175	46
Current	123733	126244	110948	58061	15440	2399	493
Default	1	2	2	3	1	0	0
Fully Paid	10071	11413	10237	6206	1870	370	73
In Grace Period	304	655	999	688	202	39	14
Late (16-30 days)	120	310	411	353	98	21	10
Late (31-120 days)	430	1341	2127	1764	641	171	35

Εικόνα 28: Πλήθος δανείων ανά βαθμό

Σκοπός μας είναι να προβλέψουμε την πιθανότητα να αποπληρωθεί ένα δάνειο ή ακόμα καλύτερα να μην αποπληρωθεί κάποιο. Έτσι είναι λογικό ότι η κάθε κατηγορία δανείου μας φέρει διαφορετικό κίνδυνο και δεν θα ήταν λογικό να βγάλουμε συμπεράσματα για όλα τα δάνεια γιατί δεν έχουν τις ίδιες πιθανότητες να αποπληρωθούν. Είναι σημαντικό επίσης να αποφύγουμε την πιθανότητα ένα δάνειο που είναι επικίνδυνο να αποπληρωθεί να έρθει ως καλή επένδυση και να φέρει άμεση απώλεια. Για το καλύτερο δυνατό αποτέλεσμα λοιπόν θα επιλέξουμε δάνεια τα οποία έχουν μεγάλη επικινδυνότητα. Οι δύο τελευταίες κατηγορίες έχουν πολύ μικρό πλήθος δανείων οπότε επιλέγουμε την αμέσως επόμενη.

Στη συνέχει θα πρέπει να γίνει ο διαχωρισμός των δεδομένων σε αυτά της εκπαίδευσης, της επικύρωσης και της δοκιμής. Τα δεδομένα εκπαίδευσης είναι αυτά που θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου. Αποτελούν τον μεγαλύτερο όγκο των δεδομένων και είναι αυτά που θα δώσουν συμπεράσματα σε σχέση με αυτά της δοκιμής. Τα δεδομένα επικύρωσης βοηθούν στην παραμετροποίηση του αλγορίθμου και σε κάποια στατιστικά αποτελέσματα που θα βγουν από τον αυτόν. Είναι σημαντικό η διάσπαση των δεδομένων σε εκπαίδευσης-δοκιμής και σε εκπαίδευσης-επικύρωσης να είναι αριθμητικά παρόμοια για την ακρίβεια των αποτελεσμάτων.

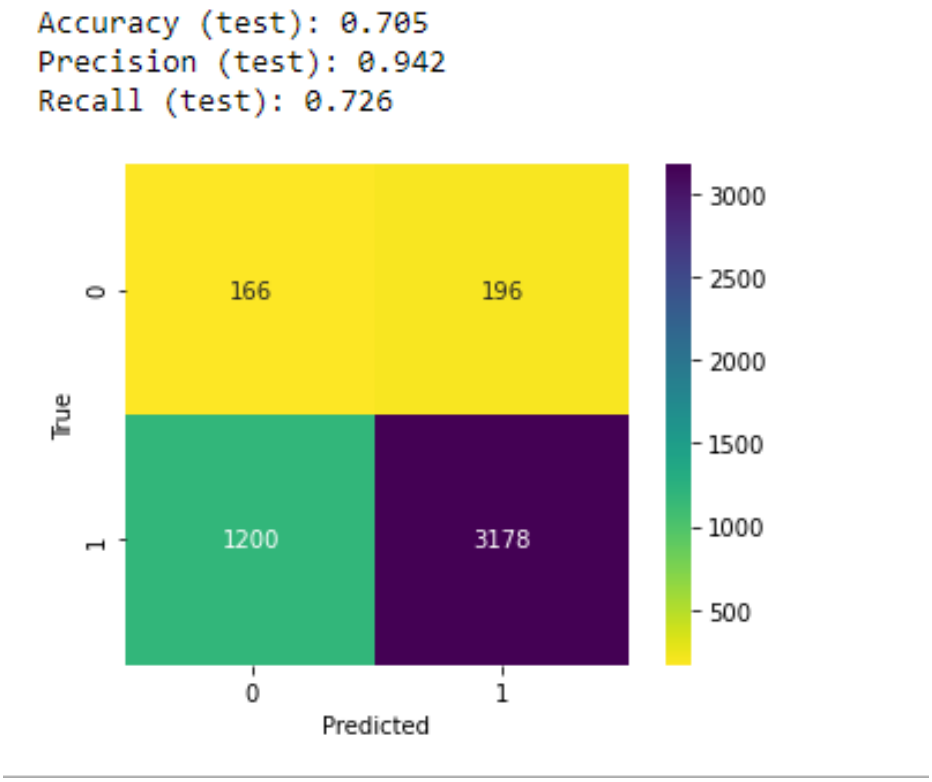
Όπως αναφέρθηκε ήδη ο Catboost είναι ένας αλγόριθμος ο οποίος βγάζει τα επιθυμητά αποτελέσματα με την πρώτη προσπέλαση. Δεν είναι αναγκαίο να τρέξει ο αλγόριθμος πολλές φορές και να γίνει σύγκριση αποτελεσμάτων. Έτσι αφού δώσουμε τις παραμέτρους που χρειάζονται για τον αλγόριθμο που αφορούν τις επαναλήψεις, τον ρυθμό εκμάθησης και τις τάξεις βάρους τότε γίνεται η σχετική εκπαίδευση του αλγορίθμου. Είναι σημαντικό να χρησιμοποιηθούν σωστά οι παράμετροι καθώς υπάρχει ανισορροπία στο δείγμα της αποθήκης δεδομένων. Σημαντική συμβολή του Catboost στα δεδομένα είναι ο χειρισμός των χαρακτηριστικών που δεν αφορούν αριθμούς.



Εικόνα 29: Εκπαίδευση αλγορίθμου

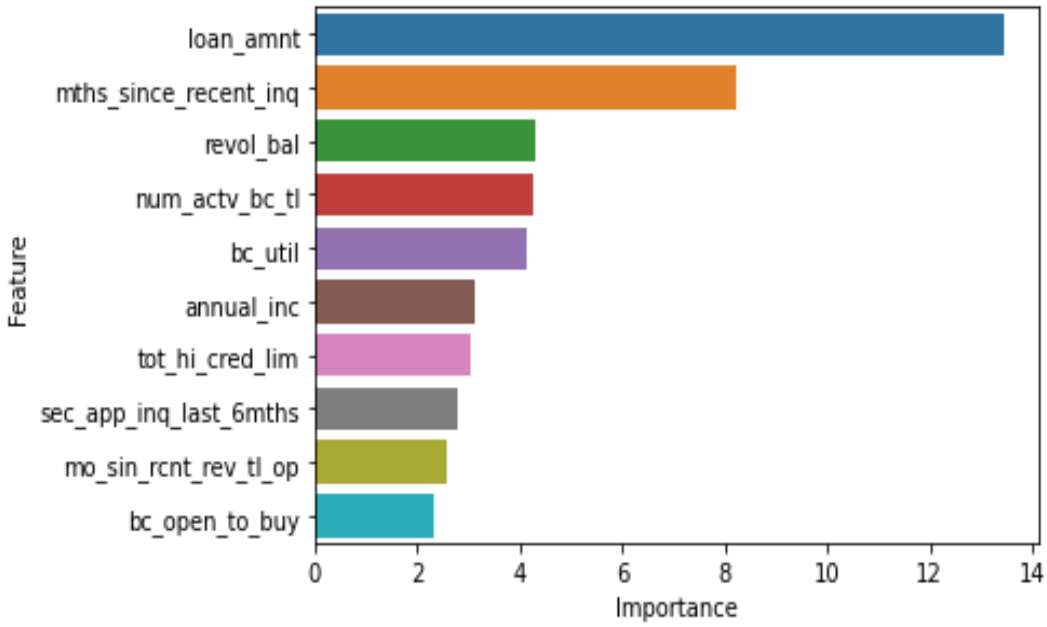
Έπειτα από την εκπαίδευση του αλγορίθμου θα γίνει η σύγκριση με τα δεδομένα της δοκιμής καθώς αυτά της επικύρωσης έχουν ήδη χρησιμοποιηθεί. Η σύγκριση αυτή θα δώσει τρεις τιμές οι οποίες θα εξεταστούν και θα μας δώσουν συμπεράσματα για το αν τελικά έχουμε τα επιθυμητά αποτελέσματα. Όπως αναφέρθηκε παραπάνω οι τιμές αυτές έχουν να κάνουν με την ακρίβεια ανάμεσα σε θετικά και αρνητικά αποτελέσματα και είναι το accuracy, το recall και το precision. Το accuracy είναι η γενική ακρίβεια που έχουμε ανάμεσα σε σωστές προβλέψεις και σε λάθος προβλέψεις. Precision είναι η ακρίβεια που έχουμε ανάμεσα σε σωστά αποτελέσματα και λάθος θετικά αποτελέσματα ενώ το recall αφορά την ανάκληση, δηλαδή τον λόγο ανάμεσα σε θετικά και σε λάθος αρνητικά αποτελέσματα.

Μαζί με τις τιμές αυτές εξάγουμε και τον πίνακα σύγχυσης ο οποίος μας δείχνει την δομή των αποτελεσμάτων. Σε αυτόν μπορούμε να δούμε αναλυτικά τις προβλέψεις που ορθώς ή λανθασμένα έγιναν και να υπολογίσουμε τις επιθυμητές τιμές αποτελεσμάτων. Όπως φαίνεται στο παρακάτω διάγραμμα οι τιμές που πέτυχε ο αλγόριθμος είναι ικανοποιητικές καθώς σε γενική ακρίβεια έχουμε 70% επιτυχία ενώ το precision είναι στο 94%. Ο αλγόριθμος κατάφερε να προβλέψει 3.344 από τα 4.740 δάνεια που υπήρχαν στο δείγμα δοκιμής, το οποίο είναι πολύ σημαντικό σε αντιδιαστολή με το συνολικό δείγμα που επιλέξαμε που ήταν κάτω από 20.000. Αυτό είναι ένα πολύ σημαντικό χαρακτηριστικό του Catboost που τον καθιστά στους καλύτερους αλγορίθμους της κατηγορίας του.



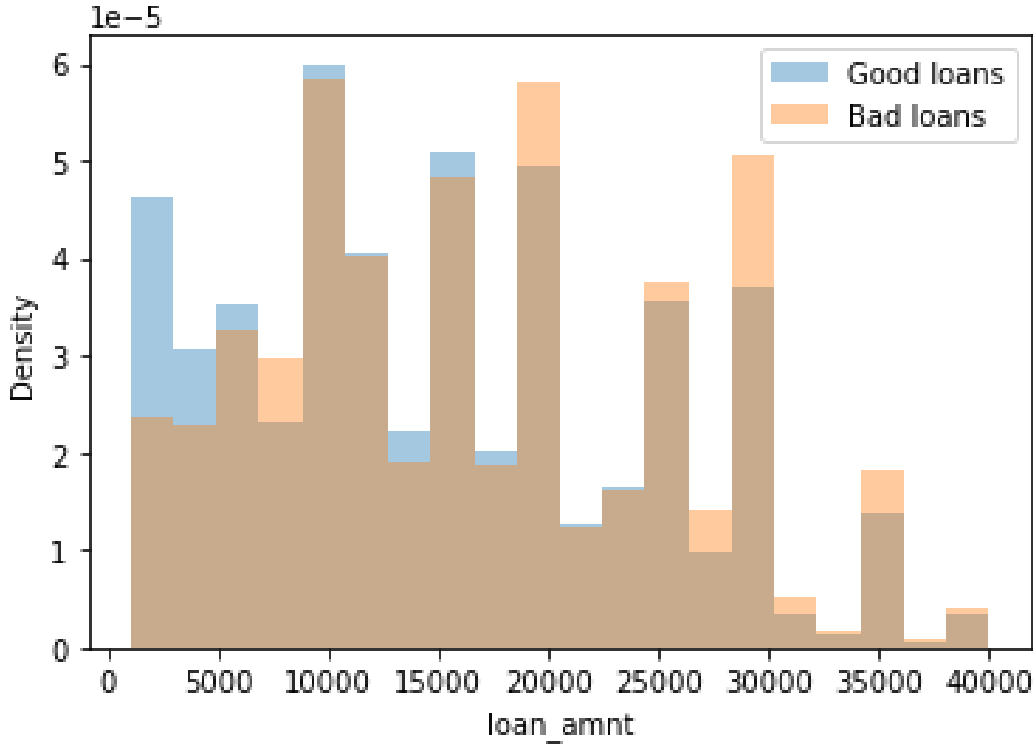
Εικόνα 30: Διάγραμμα αποτελεσμάτων

Χρήσιμα συμπεράσματα για τα αποτελέσματα του αλγορίθμου μπορούμε να βγάλουμε από τα δεδομένα τα οποία είχαν καθοριστικό ρόλο στην εξαγωγή του αποτελέσματος. Είναι δηλαδή οι στήλες αυτές της αποθήκης δεδομένων που καθορίζουν κατά μεγάλος βαθμό την πρόβλεψη. Μέσα από αυτά μπορούμε να αξιολογήσουμε τον αλγόριθμο και την ακεραιότητα αυτού.



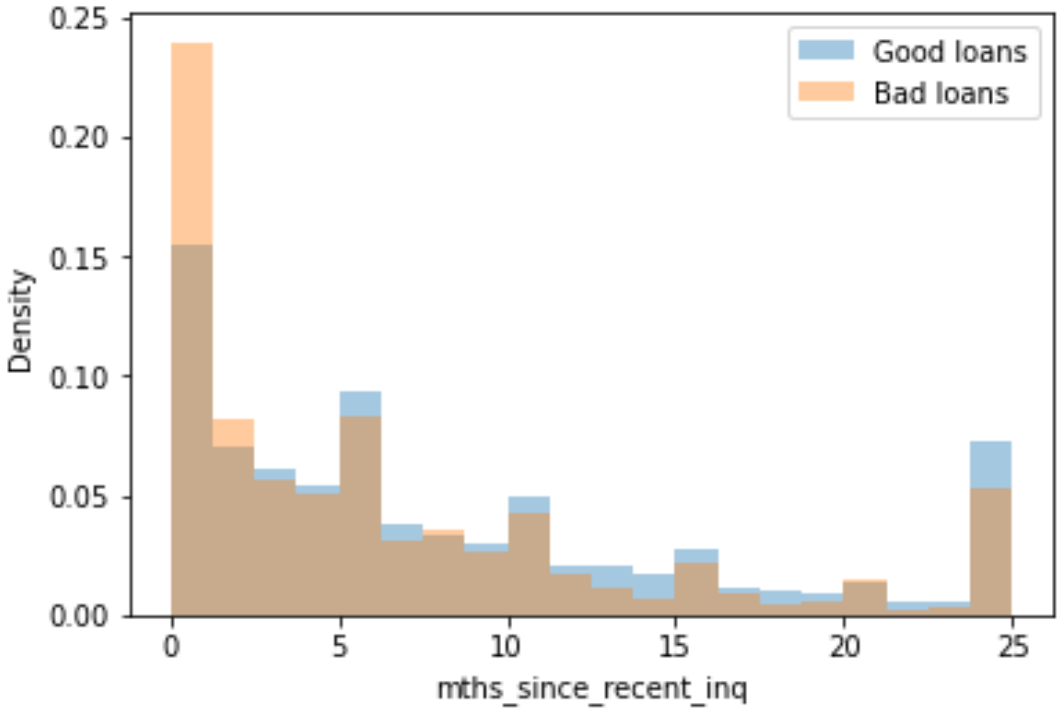
Εικόνα 31: Σημαντικά χαρακτηριστικά

Όπως φαίνεται παραπάνω καθοριστικό ρόλο έχουν το loan_amnt και το mths_since_recent_inq. Το πρώτο έχει να κάνει με το ποσό του δανείου. Είναι προφανές ότι το ποσό του δανείου θα έχει πρωταρχικό ρόλο στο αν κάποιο θα αποπληρώσει το δάνειο του. Το δεύτερο χαρακτηριστικό αφορά τους μήνες που έχουν παρέλθει από την τελευταία έρευνα στον δανειολήπτη.



Εικόνα 32: Σχέση καλών-κακών δανείων για το loan_amnt

Όπως είναι λογικό το ποσό του δανείου έχει καθοριστικό ρόλο στο αν θα αποπληρώσει ο δανειολήπτης το δάνειο ή όχι. Όσο πιο μικρό είναι το πόσο, τόσο πιο εύκολο φαντάζει το γεγονός ότι κάποιος θα το αποπληρώσει. Χαρακτηριστικά φαίνεται μέχρι τις 17.000 τα καλά δάνεια είναι παραπάνω από τα κακά. Όσο τα ποσά μεγαλώνουν τόσο πιο πολλά είναι τα κακά δάνεια. Όποτε φαίνεται λογικό να επενδύσει κάποιος σε δάνειο μικρότερου ποσού.

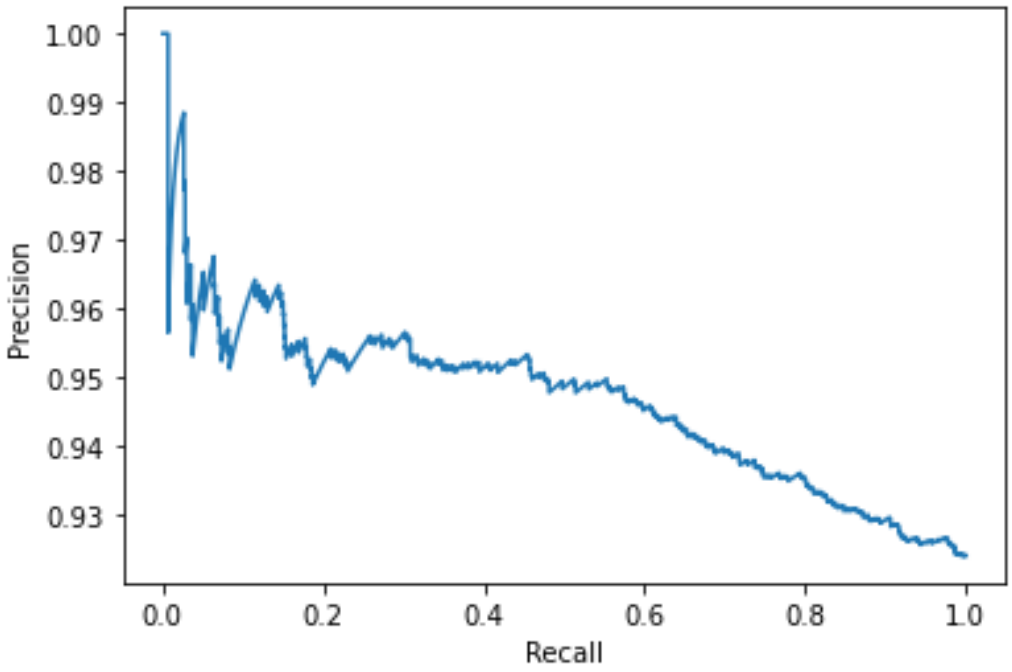


Εικόνα 33: Σχέση καλών κακών δανείων με βάση την τελευταία έρευνα

Στο παραπάνω γράφημα φαίνεται η διαφορά των δανείων με βάση την τελευταία έρευνα που έγινε στον δανειολήπτη. Εδώ τα αποτελέσματα φαντάζουν και πάλι λογικά καθώς μία έρευνα γίνεται όταν γίνει αίτηση για νέο δάνειο ή για έκδοση πιστωτικής κάρτας. Πράγμα το οποίο σημαίνει ότι τα οικονομικά του δανειολήπτη δεν είναι και τόσο καλά άρα με την σειρά του αυτό θα επιφέρει καθυστέρηση ή και καθόλου αποπληρωμή του δανείου. Έτσι μπορούμε να συμπεράνουμε ότι τα καλά δάνεια βρίσκονται στα σημεία όπου δεν έχει γίνει έρευνα σε μικρό χρονικό διάστημα.

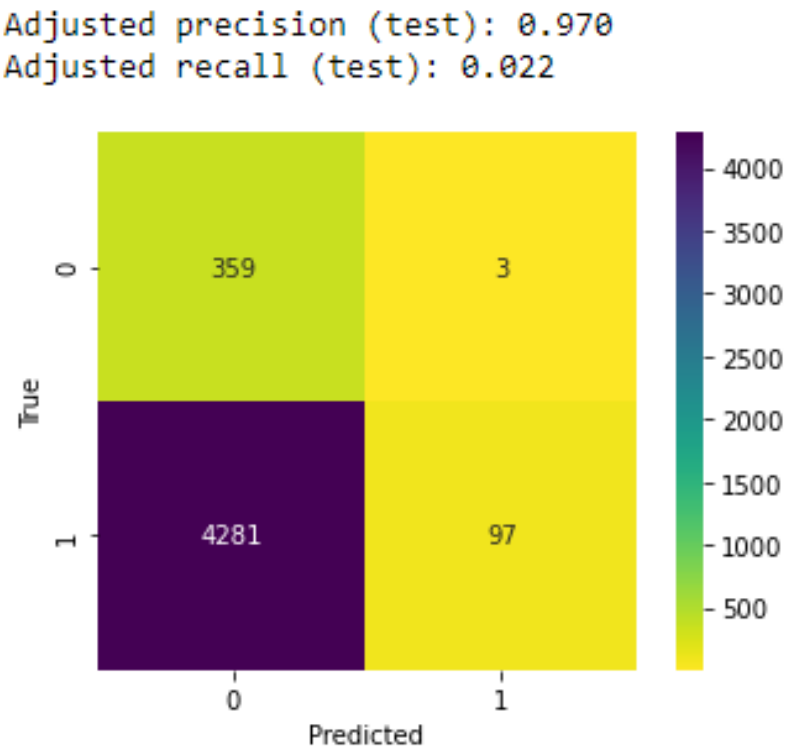
Τα παραπάνω αποτελέσματα αφορούν μια ισορροπημένη και σωστή απεικόνιση σε ότι αφορά την πρόβλεψη του κινδύνου. Σε γενικές γραμμές έχουμε νούμερα που μπορούν να φέρουν κέρδος στον επενδυτή καθώς κατά μεγάλο ποσοστό ο αλγόριθμος μπορεί να πέσει σωστά στην πρόβλεψη του. Αυτό όμως σαν γεγονός δεν αναιρεί το ότι υπάρχει κίνδυνος ο επενδυτής να έχει άμεση απώλεια χρημάτων. Παρά το πολύ μεγάλο ποσοστό που έφερε το precision που ήταν πάνω από 94%, είναι ταυτόχρονα και πολύ επικίνδυνο. Το precision επηρεάζεται από τα δάνεια που προβλέφτηκαν ως σωστά αλλά εσφαλμένα. Αυτό σημαίνει ότι ο επενδυτής χάνει το κεφάλαιό του. Το recall από την άλλη επηρεάζεται από τα δάνεια που έγινε πρόβλεψη ότι θα είναι

κακά δάνεια αλλά τελικά δεν είναι. Εδώ έχουμε ακριβώς το αντίθετο, δηλαδή χαμένες ευκαιρίες για επένδυση. Σίγουρα και αυτό σαν αποτέλεσμα θα φέρει λιγότερα κέρδη αλλά σε ότι αφορά μία επένδυση πάντα πρέπει να μειώνεται το ρίσκο απώλειας του κεφαλαίου. Έτσι σε μια προσπάθεια να γίνει αυτό θα πρέπει να περιορίσουμε την περίπτωση λάθους επένδυσης ακόμα και αν χαθεί κάθε ευκαιρία επένδυσης σε δάνεια που κακώς απορρίφθηκαν. Έτσι θα δημιουργήσουμε μια σχέση ανάμεσα στα δύο αυτά στοιχεία με σκοπό τα ακόμα καλύτερα αποτελέσματα.



Εικόνα 34: Καμπύλη precision με recall

Για την παραπάνω καμπύλη χρησιμοποιήθηκαν τα δεδομένα της επικύρωσης. Όπως φαίνεται όταν η τιμή του precision παίρνει την τιμή 1 τότε το recall είναι μηδέν. Πράγμα το οποίο δεν το θέλουμε καθώς μετά δεν θα μπορεί ο αλγόριθμος να προβλέψει καλά δάνεια καθόλου. Έτσι αφαιρείται το πάνω όριο του precision. Έπειτα λοιπόν από τον υπολογισμό των ορίων για το νέο precision μπορούμε να δούμε μια σημαντική αύξηση σε αυτό, αλλά και μια δραματική μείωση στο recall.



Εικόνα 35: Διάγραμμα αποτελεσμάτων μετά την προσαρμογή των δεδομένων

Η προσαρμογή που έγινε στα δεδομένα μας φέρνει μια αύξηση στο precision κατά 3%. Από το 94% έχουμε πλέον ένα πιο ασφαλές ποσοστό στο 97%. Από την άλλη πλευρά όμως το recall έπεσε στο 2%. Αυτό σημαίνει ότι πλέον υπάρχει πολύ μεγάλη πιθανότητα να προβλεφθεί κάποιο δάνειο ως κακό ενώ στην πραγματικότητα είναι καλό. Σκοπός όμως στην επένδυση είναι να μειωθεί το κόστος όσο το δυνατόν πιο πολύ. Πράγμα το οποίο αποτυπώνεται στο παραπάνω διάγραμμα καθώς από τα 100 δάνεια βλέπουμε μόλις τρία να έχουν την ένδειξη του λανθασμένα σωστά.

Συμπεράσματα και μελλοντική έρευνα

Η παρούσα διπλωματική αναφέρεται στην εξέλιξη της τεχνολογίας και πως αυτή έχει επηρεάσει τα χρηματοπιστωτικά ιδρύματα αλλά και άλλου φορείς οι οποίοι επηρεάζονται άμεσα από τον πιστωτικό κίνδυνο. Περιέχει πληροφορίες σχετικές με την μηχανική μάθηση καθώς και για αλγορίθμους και πώς αυτοί συμβάλουν στην πρόβλεψη του πιστωτικού κινδύνου. Γίνεται μια σχετική ανάλυση λοιπόν στο τι είναι πιστωτικός κίνδυνος και ποιες είναι οι επιπτώσεις για μια επιχείρηση σε περίπτωση που δεν τον λάβει σοβαρά υπ όψιν. Επίσης υπάρχουν σημαντικές αναφορές για την εξόρυξη των δεδομένων και το πώς αυτή συνδέεται με την μηχανική μάθηση για να φέρουν το επιθυμητό αποτέλεσμα. Σε θεωρητικό επίπεδο θα μπορούσε να εμβαθύνει κάποιος στην έννοια της μηχανικής μάθησης και το πώς αυτή έχει βοηθήσει γενικότερα την επιστήμη των υπολογιστών και όχι απλά για τον πιστωτικό κίνδυνο. Επίσης οι αλγόριθμοι οι οποίοι αναφέρθηκαν είναι αυτοί που είναι κοντά στο πρακτικό κομμάτι και η αναφορά έγινε για να δώσει μια γενική εικόνα στον αναγνώστη. Υπάρχουν αλγόριθμοι με μεγάλη απήχηση στην πληροφορική και όχι μόνο που μπορούν να αναλυθούν.

Από την άλλη πλευρά στο πρακτικό κομμάτι έχει γίνει επιλογή του Catboost. Με βάση τα κριτήρια που αναλύθηκαν επιλέχτηκε ως ο πλέον κατάλληλος για να ανταπεξέλθει σύμφωνα με τις ανάγκες της συγκεκριμένης αποθήκης δεδομένων. Τα αποτελέσματα που βγήκαν από τον αλγόριθμο ήταν ικανοποιητικά και αποδίδουν επιθυμητά ποσοστά αναφορικά με τις απαιτήσεις ενός επενδυτή. Όμως αυτό δεν σημαίνει ότι δεν αξίζει να γίνει μια ανάλογη προσπάθεια με άλλους αλγορίθμους και μοντέλα όπως είναι ο XGBoost ή ο LightGBM. Γενικότερη έρευνα για την ακρίβεια του μοντέλου, την ταχύτητα απόδοσης αλλά και τις παραμέτρους που χρησιμοποιήθηκαν, θα μπορούσαν να επιφέρουν ακόμα καλύτερα αποτελέσματα. Επιπρόσθετα ίσως μία διαφορετική προσέγγιση σε ότι αφορά την επεξεργασία των δεδομένων αλλά και την επιλογή των χαρακτηριστικών μπορεί να φέρει περισσότερα συμπεράσματα αναφορικά με την γενικότερη λειτουργία του μοντέλου. Τέλος όπως αναφέρθηκε παραπάνω έγινε χρήση ενός μέρους των δεδομένων για να αποφευχθεί η επιβράδυνση της διαδικασίας. Με γνώμονα αυτό θα

μπορούσε να χρησιμοποιηθεί άλλο υποσύνολο δεδομένων και να γίνει μια γενικότερη σύγκριση των αποτελεσμάτων. Με αυτήν την τακτική θα μπορούσε να επιτευχθεί ακόμα περισσότερη ακρίβεια στην αποφυγή του πιστωτικού κινδύνου.

Αναφορές

[1] Ζοπουνίδης, Κ. (2001). Ανάλυση και διαχείριση χρηματοοικονομικών κινδύνων. Αθήνα: Εκδόσεις Κλειδάριθμος.

[2] Qi Zhang , Jue Wang, Aiguo Lu , Shouyang Wang, Jian Ma, (2017) An improved SMO algorithm for financial credit risk assessment –Evidence from China's banking

[3] Καλφάογλου Φ. (1999). Υποδείγματα μέτρησης πιστωτικού κινδύνου.

[4] Hellwig M. (1995). Systemic Aspects of Risk Management in Banking and Finance. Vol. 131 (4/2), 723-737.

[5] Machauer, A., Weber, M., Bank behavior based on internal credit ratings of borrowers. Journal of Banking and Finance 22, 1355-1383.1998.

[6] Felipe Alonso Arias-Arbelaez, Juan Sebastian Bravo-Valbuenay Francisco Ivan Zuluaga-Dıaz, (2015) Estimation of a credit scoring model for lenders company

[7] Fayyad, Usama (1996). «From Data Mining to Knowledge Discovery in Databases», AAAI 97, Providence Rhode Island, July 27 – 31, 1997, Conference

[8] Thuraisingham, B. (1999). Data Mining, Technologies, Techniques, Tools and Trends

[9] https://en.wikipedia.org/

[10] https://ann-chung-porfolio.webflow.io/work/predictive-analysis

[11] Bishop, C. M., (2006) "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer.

[12]https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781838555078/6/ch06lvl1sec34/confusion-matrix

[13] Witten I.H. and Frank E., (2005) "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", second eds., Morgan Kaufmann

[14] Artem Bequé, Stefan Lessmann, (2017) Extreme learning machines for credit scoring: An empirical evaluation

[15]https://www.datascienceexamples.com/artificial-neural-networks-an-introduction-for-data-science-beginners

[16] https://www.tutorialandexample.com/decision-trees/

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin. (2018). CatBoost: unbiased boosting with categorical features.

[18]https://towardsdatascience.com/introduction-to-gradient-boosting-on-decision-trees-with-catboost-d511a9ccbd14

[19] https://catboost.ai/news/extremely-fast-learning-on-gpu-has-arrived