Πανεπιστήμιο Πειραιώς - Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Κατανεμημένα Συστήματα, Ασφάλεια και Αναδυόμενες Τεχνολογίες Πληροφορίας»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Αξιολόγηση Προσωπικών Δανείων** **Evaluation of Personal Loans**
Ονοματεπώνυμο Φοιτητή	**Ιωάννης Καπήλου**
Πατρώνυμο	**Κωνσταντίνος**
Αριθμός Μητρώου	**ΜΠΚΣΑ 18011**
Επιβλέπων	**Δρ. Γρηγόριος Κορωνάκος**

Ημερομηνία Παράδοσης **Νοέμβριος 2022**

Τριμελής Εξεταστική Επιτροπή

Δρ. Γρηγόριος Κορωνάκος Διδάσκων ΠΜΣ	Δημήτριος Αποστόλου Καθηγητής	Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής μου διατριβής, ευχαριστώ τον καθηγητή Αποστόλου Δημήτριο που μου έδωσε την δυνατότητα να ασχοληθώ με τον τομέα της Αξιολόγησης Δανειοληπτών.

Τέλος θέλω να ευχαριστήσω τα μέλη της οικογένειάς μου για την συνεχόμενη υποστήριξή τους καθ' όλη την πορεία μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη

Στην εποχή της Πληροφορίας οι χρηματοπιστωτικοί οργανισμοί αναζητούν τον βέλτιστο τρόπο να αξιολογήσουν τους δανειολήπτες τους χρησιμοποιώντας αυτοματοποιημένες διαδικασίες ανάλυσης δεδομένων. Οι εν λόγω διαδικασίες προστατεύουν και βοηθούν τους χρηματοπιστωτικούς οργανισμούς να είναι αποδοτικότεροι. Ως εκ τούτου τα δεδομένα είναι ζωτικής σημασίας για την επιβίωσή και την ανταγωνιστικότητά τους. Τα δεδομένα που λαμβάνουν οι οργανισμοί αντλούνται από δημόσιους φορείς εισοδήματος, όπου ο πελάτης έχει πρόσβαση, και από ιδιωτικούς ανεξάρτητους φορείς μελετών οι οποίοι, συνήθως με κάποια προμήθεια, δίνουν στις τράπεζες ή άλλους οργανισμούς δεδομένα συναλλακτικής συμπεριφοράς και βαθμολογίες αξιολόγησης των δανειοληπτών. Παρόλο που το τραπεζικό και ιδιωτικό δίκαιο προστατεύουν και περιφράττουν αυτά τα συστήματα, διάφοροι μελετητές έχουν αποσαφηνίσει τις έννοιες των χαρακτηριστικών που οδηγούν τους οργανισμούς να λάβουν αποφάσεις δίνοντάς έτσι το έναυσμα να μελετήσουμε ποιες είναι εννοιολογικά αυτές οι μεταβλητές αλλά και τι εργαλεία μπορούμε να χρησιμοποιήσουμε ως αναπαράσταση της πραγματικής διαδικασίας μιας δανειοδότησης ώστε να εντοπίσουμε μοτίβα και αρχές που αξιοποιούν αυτοί οι οργανισμοί στην παραγωγική τους διαδικασία. Στην εργασία γίνεται μελέτη πάνω στην αξιολόγηση των δανειοληπτών, στις τεχνικές που μπορούν να αξιοποιηθούν από την επιστήμη της Πληροφορικής για την ανάλυση δεδομένων με απώτερο σκοπό να εντοπιστούν τα μεγέθη τα οποία επηρεάζουν άμεσα την απόφαση δανειοδότησης.

Abstract

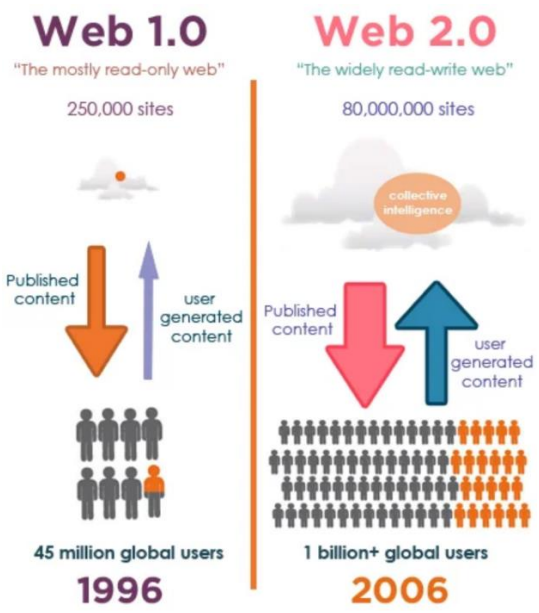
In the Information Age financial institutions are looking for the best way to evaluate their borrowers using automated data analysis processes. These processes protect and help financial institutions to be more efficient. Data is therefore vital to their survival and competitiveness. The data the agencies receive is drawn from public revenue agencies, where the customer has access, and from private independent survey agencies who, usually for a fee, provide banks or other agencies with transactional data and borrower ratings. Although banking and private law protect and fence these systems, various scholars have clarified the concepts of the characteristics that lead organizations to make decisions, thus giving the impetus to study what these variables are conceptually and what tools we can use as a representation of the actual lending process to identify patterns and principles that these organizations use in their production process. In this thesis, a study is made on the evaluation of borrowers regarding their personal loan applications, on the techniques that can be used by the science of Information Technology for data analysis with the goal of identifying the figures that directly influence the lending decision.

1. ΕΙΣΑΓΩΓΗ

1.1. ΙΣΤΟΡΙΚΟ ΜΕΛΕΤΗΣ

Με τον ρυθμό ανάπτυξης της επιστήμης της πληροφορικής καθώς και τις δυναμικά αναπτυσσόμενες ανάγκες των πληροφοριακών συστημάτων μικρών και μεγάλων επιχειρήσεων, δημιουργείται και η ανάγκη για ποιοτικό έλεγχο των προαναφερθέντων συστημάτων καθώς και για αναπροσαρμογή και βελτιστοποίηση διαφόρων διαδικασιών που αφορούν τους τομείς παραγωγής και μελέτης καθώς οι διαδικασίες αξιολόγησης και πρόβλεψης αυτοματοποιούνται και χρίζουν παραμετροποίησης ανάλογα με τις ανάγκες και στρατηγικές του οργανισμού.

Το λεγόμενο Web 2.0 ορίζει το διαδίκτυο ως μια ‘πλατφόρμα’ η οποία συνδέει τις συνδεδεμένες συσκευές μεταξύ τους. Οι εφαρμογές που φτιάχνονται πάνω σε αυτό είναι το λογισμικό ως υπηρεσία (Software as a Service) που ενημερώνεται συνεχώς και το οποίο αναδραστικά δημιουργεί τις απαιτήσεις, όσο περισσότεροι χρήστες το χρησιμοποιούν, καταναλώνοντας και αναμιγνύοντας δεδομένα από πολλαπλές πηγές, συμπεριλαμβανομένων μεμονωμένων χρηστών, ενώ παρέχουν τα δικά τους δεδομένα και υπηρεσίες σε μορφή που επιτρέπει την ανάμιξη από άλλους, δημιουργώντας μια «αρχιτεκτονική συμμετοχής», υπερβαίνοντας έτσι τη μεταφορά της σελίδας η οποία παρέχει στον χρήστη πληροφορία και εμπειρία στο διαδίκτυο, web 1.0 (O' Reilly, 2005). Ως φυσικό επακόλουθο η εμφάνιση της τεχνολογίας web 2.0 έχει προκαλέσει την ταχεία καθιέρωση των διαδικτυακών αγορών και της εικονικής κοινότητας (Virtual Community) όπου ένα άτομο μπορεί να αλληλοεπιδράσει εικονικά για να καλύψει τις ανάγκες του σε ότι αφορά τον οικονομικό προγραμματισμό του και σε ένα ευρύτερο φάσμα την γρήγορη μετάδοση την πληροφορίας σε όποιον κλάδο ή/και υπηρεσία μπορεί να χρειαστεί.



ΕΙΚΟΝΑ 1. ΔΙΑΚΡΙΣΗ ΟΡΙΩΝ ΜΕΤΑΞΥ WEB 1.0 & WEB 2.0 (TSAKIRIDIS, 2020)

Με τα εργαλεία που δημιουργήθηκαν πάνω στο Web 2.0 εμφανίστηκε και η φιλοσοφία του τρόπου δανεισμού Peer to Peer Lending ο οποίος με την σειρά του αποδίδει έναν εναλλακτικό τρόπο δανειοδότησης σε άτομα τα οποία ψάχνουν να δανειστούν γρήγορα και συνήθως όχι υπέρογκα ποσά.

Για την αποσαφήνιση της έννοιας peer-to-peer (P2P) lending, αυτό το είδος δανεισμού επιτρέπει στα άτομα να λαμβάνουν δάνεια απευθείας από άλλα άτομα, αποκλείοντας τα παραδοσιακά χρηματοπιστωτικά ιδρύματα ως την μόνη επιλογή. Οι ιστοσελίδες που διευκολύνουν τον δανεισμό P2P έχουν αυξήσει σημαντικά την υιοθέτησή του ως εναλλακτική μέθοδο χρηματοδότησης. Ο δανεισμός P2P είναι επίσης γνωστός ως «κοινωνικός δανεισμός» (Communal Lending) ή «δανεισμός πλήθους» (Crowd Lending). Υπάρχει από το 2005 και ένα ενδεικτικό πλήθος ανταγωνιστών περιλαμβάνει το Prosper, το Lending Club, το Peerform, το Upstart και το StreetShares (Kagan, 2020). Είναι σημαντικό να αναφερθεί πως τέτοιοι οργανισμοί δεν έχουν αναπτυχθεί ακόμα στην ελληνική αγορά, παρόλα αυτά υπάρχει η δυνατότητα πρόσβασης σε τέτοιες υπηρεσίες μέσω του διαδικτύου οι οποίες αναγνωρίζουν ελληνικά πιστοποιητικά όπως τα Δελτία Ταυτότητας, καταθέσεις κεφαλαίων σε τρίτους και φορολογικές δηλώσεις. Χαρακτηριστικά παραδείγματα είναι οι η Mintos η οποία παρέχει την δυνατότητα P2P δανεισμού μέσω της υπηρεσίας της Invest & Access.

Όλοι οι χρηματιστικοί οργανισμοί αναλαμβάνουν ρίσκο όταν προχωρούν σε πιστώσεις χρημάτων σε τρίτους, πελάτες, είτε μέσω καταναλωτικών δανείων, στεγαστικών δανείων ή και επιχειρηματικών δανείων μικρών επιχειρήσεων. Οι P2P πλατφόρμες θεωρούνται Επιχειρήσεις Κερδοσκοπικού χαρακτήρα που αναπτύσσονται στον τομέα της Τραπεζικής και των Επενδύσεων. Λαμβάνεται ως Δεδομένο πως υπάρχει αντίστοιχη Διεύθυνση ή Τμήμα ή τρίτη εταιρεία (outsourced) η οποία είναι υπεύθυνη να διαχειρίζεται το ρίσκο καθώς και να ελέγχει τα μετρικά (Credit Metrics) προ απαιτούμενα όπως για παράδειγμα η χρηματοπιστωτική ικανότητα του δανειολήπτη (Credit Score). Επίσης θεωρείτε δεδομένο πως οι διαδικασίες έχουν αυτοματοποιηθεί σε κάποιο πρόγραμμα εσωτερικής χρήσης που αφορά την δανειοδότηση.

Προκειμένου να γίνει πραγματική αξιοποίηση των παραπάνω, είναι ζωτικής σημασίας να υπάρχει πληρότητα των δεδομένων, ομοιότητα, ακεραιότητα αλλά και κατηγοριοποίηση πρωτοφανών περιπτώσεων (Aylin Hejazi, 2017). Η ίδια αρχή ακολουθείται γενικά στην Ανάλυση Δεδομένων όπου θα πρέπει όσο είναι δυνατό να εκμηδενίζεται η μεροληψία δεδομένων. Το λεγόμενο Data Bias ή αλλιώς η μεροληψία δεδομένων είναι ότι τα διαθέσιμα δεδομένα δεν είναι αντιπροσωπευτικά του πληθυσμού ή του φαινομένου μιας μελέτης ή ανάλυσης. Αυτό μπορεί να αποφευχθεί με την επικαιροποίηση και επεξεργασία έγκυρων δεδομένων ώστε να μην φτάνει κανείς σε εσφαλμένα αποτελέσματα από εσφαλμένες υποθέσεις λόγω της πρώτης εικόνας των δεδομένων. Η πρόσβαση στα εσωτερικά συστήματα είναι απίθανη, συνεπώς η μελέτη θα γίνει καθαρά από datasets και άλλες μελέτες ερευνητών. Είναι σημαντικό να αναφερθεί πως η τραπεζική θεωρία δείχνει ότι ενδιαφερόμενα μέρη όπως τράπεζες, πιστωτικές ενώσεις κ.λπ. μπορούν να μειώσουν ορισμένα από τα δυσμενή προβλήματα επιλογής μέσω της πρόσληψης ειδικών στελεχών ελέγχου, την λήψη εγγυήσεων και εξασφαλίσεων καθώς και την παρακολούθηση των δανείων μετά την εκταμίευσή τους (Akerlof, 1970).

1.2. ΟΡΙΣΜΟΣ ΑΞΙΟΛΟΓΗΣΗΣ ΔΑΝΕΙΟΛΗΠΤΩΝ

Η Αξιολόγηση των Δανειοληπτών, σε ότι αφορά την διαδικασία δανειοδότησης, γίνεται βάση κάποιων στοιχείων που αρχικά προμηθεύει τον εκάστοτε οργανισμό ο πελάτης/δανειζόμενος. Αξιολογείται δηλαδή η ικανότητά του να αποπληρώσει ένα δάνειο καθώς αυτά τα στοιχεία αποτελούν την αρχική οικονομική του εικόνα. Στο πλαίσιο μιας τέτοιας αξιολόγησης, ο οργανισμός εξετάζει αυτά τα στοιχεία, τόσο ξεχωριστά όσο και συνολικά, για τον προσδιορισμό της καταλληλότερης λύσης προϊόντος για τον πελάτη. Πιο συγκεκριμένα στη διαδικασία της αξιολόγησης η Ευρωπαϊκή Κεντρική Τράπεζα αναφέρει (Ευρωπαϊκή Κεντρική Τράπεζα, 2017), σε μικρή απόκλιση με την Επιτροπή Κυβερνητών της Ομοσπονδιακής Τράπεζας των Η.Π.Α (Board of Governors of the Federal Reserve System, 2021), πως λαμβάνονται υπόψιν οι εξής εμπλεκόμενες πτυχές:

- Εξέταση των προσωπικών οικονομικών και μη οικονομικών στοιχείων του δανειολήπτη.
- Εξέταση της συνολικής δανειακής επιβάρυνσης του δανειολήπτη, ιδίως ότι αφορά δεσμεύσεις αποπληρωμής μη εξαφανιζόμενου χρέους.
- Το συμφωνηθέν πρόγραμμα αποπληρωμών του δανείου θα πρέπει να ισούται με το εναπομένον διαθέσιμο εισόδημα ή να είναι μικρότερο αυτού μετά από την αφαίρεση όλων των δαπανών και υποχρεώσεων.
- Ανάλυση/αξιολόγηση των ιστορικών δεδομένων για να εντοπιστεί η χρονική στιγμή και οι λόγοι των οικονομικών δυσχερειών του δανειολήπτη και παροχή ενδείξεων σχετικά με τη βιωσιμότητα της προσφερθείσας λύσης αναδιάρθρωσης.

Η αξιολόγηση του επιπέδου δαπανών του δανειολήπτη θα πρέπει να λαμβάνει υπόψη πιθανές μελλοντικές αυξήσεις των δαπανών. Κατ' ελάχιστο, οι τράπεζες θα πρέπει να είναι σε θέση να καταδεικνύουν ότι έχουν ληφθεί υπόψη τυχόν αυξήσεις σύμφωνα με τον πληθωρισμό, αλλά και ότι έχουν ληφθεί υπόψη τυχόν αυξήσεις που αφορούν συγκεκριμένα τον δανειολήπτη και την ιδιαίτερη κατάστασή του (π.χ. αύξηση των εξαρτώμενων μελών ή μελλοντικό κόστος σπουδών κ.λπ.).

Όσον αφορά στην τρέχουσα, από την στιγμή της αξιολόγησης, ικανότητα αποπληρωμής αναφέρεται πως ενδεικτικά θα πρέπει να λαμβάνονται υπόψιν τα ακόλουθα:

- Προσωπικά οικονομικά και μη οικονομικά στοιχεία (π.χ. εξαρτώμενα μέλη, ανάγκες νοικοκυριού, απασχόληση, εισόδημα, δαπάνες, κ.λπ.)·
- Συνολική δανειακή επιβάρυνση
- Τρέχουσα ικανότητα αποπληρωμής
- Ιστορικό προηγούμενων αποπληρωμών
- Λόγοι που οδήγησαν σε ληξιπρόθεσμες οφειλές
- Ωρίμανση και επίπεδο ληξιπρόθεσμων οφειλών
- Καταλληλόλητα του μεγέθους ακινήτου περί στεγαστικών αναγκών του δανειολήπτη

Σχετικά με τη μελλοντική ικανότητα αποπληρωμής δανείου/λογαριασμού, ενδεικτικά λαμβάνονται υπόψη τα ακόλουθα, έχοντας ο κάθε οργανισμός την δυνατότητα να προσαρμόζει τα απαιτούμενα στις ανάγκες του και ακολουθώντας πάντα τη νομοθεσία αυστηρά:

- Εισόδημα.
- Έτη έως την Συνταξιοδότηση του δανειολήπτη σε σχέση με τη Διάρκεια Δανείου.
- Στάδιο Κύκλου Ζωής.
- Εξαρτώμενα μέλη και η ηλικία τους.
- Καθεστώς αλλά και προοπτικές απασχόλησης.
- Τομέας του Κλάδου εργασίας/Δανειοδότησης (εάν αφορά δάνεια Μικρών Επιχειρήσεων).
- Αποταμιεύσεις και Περιουσιακά στοιχεία.
- Δάνεια και άλλες υποχρεώσεις σε Τρίτους.
- Μελλοντική ικανότητα πληρωμής.
- Ελάχιστο επίπεδο διαβίωσης.
- Σχετικοί Δείκτες της αγοράς εργασίας.
- Γνωστές μελλοντικές μεταβολές της κατάστασης του δανειολήπτη.

1.3. ΣΚΟΠΟΣ ΤΗΣ ΜΕΛΕΤΗΣ

Σκοπός της εργασίας είναι να εντοπισθούν τα μεγέθη τα οποία είναι άμεσα συμβαλλόμενα στην απόφαση των οργανισμών να δώσουν δάνειο σε έναν πελάτη. Συγκεκριμένα για την εύρεση αυτών και ώστε να παραχθεί ένα αποτέλεσμα που θα ανταποκρίνεται στην πραγματικότητα πρέπει αρχικά να γίνει αποσαφήνιση της σημασίας των μεγεθών που επηρεάζουν την λήψη αποφάσεων των χρηματοπιστωτικών οργανισμών, όπου στην δικιά μας περίπτωση είναι το Lenders Club καθώς κάποια στοιχεία πελατών είναι open-source και εξυπηρετούν τον σκοπό της εργασίας.

Για να επιτευχθεί ο στόχος πρέπει να είναι σαφής η ερμηνεία των διαθέσιμων χαρακτηριστικών, οι μεταβλητές του Συνόλου Δεδομένων ή Μεγέθη των Υποσυνόλων (Εγκεκριμένα Δάνεια & Απορριφθέντα Δάνεια), πρέπει να διαχωριστούν ώστε να μελετηθεί ποιες αντιπροσωπεύουν θετικά χαρακτηριστικά και ποιες αρνητικά αλλά και να γίνει μείωση των διαστάσεων του Συνόλου για καλύτερη ανάλυση. Τέλος η ομαδοποίηση όμοιων χαρακτηριστικών, πράξεις οι οποίες θα πραγματοποιηθούν με τεχνικές που αναφέρονται στο Κεφάλαιο 3 της Μεθοδολογίας. Ο έλεγχος συσχέτισης των μεταβλητών που θα γίνει βοηθά να εντοπισθούν οι μεταβλητές οι οποίες αντιπροσωπεύουν παρόμοια μεγέθη ώστε να μπορέσουμε να μειώσουμε τις διαστάσεις του Συνόλου δεδομένων για περαιτέρω έλεγχο.

Το πρώτο και κύριο ερώτημα της εργασίας είναι αρχικά ο προσδιορισμός των χαρακτηριστικών που ορίζουν έναν πιθανό δανειολήπτη ως ‘καλό’ ή ‘κακό’ για την χορήγηση ενός δανείου βάση την ένδειξη αποπληρωμής που έχει μέσα στο σύνολο δεδομένων. Τέλος θα συγκριθεί η συνάφεια των αποτελεσμάτων με υπάρχουσα έρευνα προκειμένου να κατοχυρωθεί εάν τα μεγέθη που η ανάλυση μας δείχνει πως είναι τα πιο ‘σχετικά’ έχουν όντως συνάφεια με το απτό αποτέλεσμα της αποπληρωμής του δανείου.

1.4. ΠΕΔΙΟ ΕΦΑΡΜΟΓΗΣ

Για τους σκοπούς της εργασίας χρησιμοποιούνται δημόσια διαθέσιμα δεδομένα που συλλέχθηκαν από τον ιστότοπο Kaggle.com και αφορά την περίοδο 01/2007 - 12/2018 απορριφθέντων και εγκεκριμένων δανείων του P2P Δανειστικού Οργανισμού Lenders Club. Οι μεταβλητές που χρησιμοποιούνται στην ανάλυση είναι εκείνες που αναφέρονται ως χαρακτηριστικά των πελατών στον ιστότοπο του οργανισμού, τα οποία θα αναλυθούν αργότερα στην εργασία σε ξεχωριστή ενότητα.

1.5. ΠΕΡΙΟΡΙΣΜΟΙ

Κύριος περιορισμός της εργασίας είναι η λογική ‘Black Box’ που υπάρχει σε ότι αφορά την πραγματική εσωτερική αξιολόγηση δανειοληπτών που γίνεται από τους δανειστές P2P και πιο συγκεκριμένα εντός του Lenders Club.

1.6. Οργάνωση Κειμένου

Η εργασία περιέχει πέντε κεφάλαια:

Το κεφάλαιο 1 με τίτλο «Εισαγωγή» εξετάζει το γενικό υπόβαθρο, τους στόχους, τα ερευνητικά ερωτήματα, τη σημασία και τους περιορισμούς της μελέτης.

Το κεφάλαιο 2 με τίτλο «Ανασκόπηση Βιβλιογραφίας» που περιλαμβάνει την βιβλιογραφία που χρησιμοποιείται για την εργασία σχετικά με την παραβατικότητα των δανείων, την ανάλυση δεδομένων σχέσεων των μεταβλητών και γενικές τάσεις πάνω στον τομέα.

Το κεφάλαιο 3 με τίτλο «Μεθοδολογία Έρευνας» συζητά τον σχεδιασμό, τη δειγματοληψία, τη συλλογή και την προετοιμασία δεδομένων της έρευνας μαζί με τις τεχνικές ανάλυσης δεδομένων.

Το κεφάλαιο 4 με τίτλο «Αποτελέσματα Μελέτης» παραθέτει τα ευρήματα της ανάλυσης.

Το κεφάλαιο 5 με τίτλο «Συμπεράσματα και Προτάσεις» συνοψίζει τα ευρήματα της εργασίας καθώς και συμπεριλαμβάνει συστάσεις για μελλοντική έρευνα.

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η ανάγκη για την Ανάλυση Δεδομένων σε σύνολα που αφορούν δάνεια γεννήθηκε αρχικά από την ραγδαία ανάπτυξη των δυνατοτήτων των υπολογιστών μετά το 2006 γεγονός το οποίο έδωσε στους χρηματοπιστωτικούς οργανισμούς την δυνατότητα να βελτιστοποιήσουν την βραχυπρόθεσμη ανάλυσή τους σε ότι αφορά την αποδοτικότητα των συστημάτων τους (Capanella, 2017). Ως φυσικό ακόλουθο επηρεάστηκε και ο σχεδιασμός των προϊόντων έτσι ώστε να αποδίδεται η βέλτιστη λύση σε κάθε πελάτη ανάλογα με τις ανάγκες και τις ικανότητές του.

Από τη στιγμή που οι δυνατότητες της τεχνολογίας βελτιώνονται (μεγαλύτεροι αποθηκευτικοί χώροι, γρηγορότερες ταχύτητες σύνδεσης και μεταφοράς στο διαδίκτυο, αναμετάδοση γνώσης on-demand κλπ.), αναπτύχθηκαν παράλληλα και κοινότητες όπως το Git, HacktheCamp, Kaggle οι οποίες συντονίζονται και αποτελούνται από αναλυτές διαφόρων επιπέδων οι οποίοι/οποίες ασχολούνται καθαρά με την επιστήμη των δεδομένων.

Η τραπεζική αγορά και οι καταναλωτές που χρησιμοποιούν χρηματοοικονομικά προϊόντα παράγουν τεράστιο όγκο δεδομένων σε καθημερινή βάση. Το λογισμικό που έχει αναπτυχθεί με βάση την ανάλυση δεδομένων έχει αλλάξει τον τρόπο επεξεργασίας αυτών των πληροφοριών, καθιστώντας δυνατό τον εντοπισμό τάσεων και προτύπων που μπορούν στη συνέχεια να χρησιμοποιηθούν για την ενημέρωση επιχειρηματικών αποφάσεων σε κλίμακα. Αυτό μπορεί να μεταφραστεί σε περιορισμό ρίσκου για τις τράπεζες ή και ευκαιρία για παραγωγή κέρδους έναντι του αυξημένου ρίσκου αλλά και για τους πελάτες από την άποψη ότι μπορούν να ακολουθήσουν μια συντηρητική φιλοσοφία σχετικά με τα δάνεια από την στιγμή που μπορούν ανά πάσα στιγμή να ερευνήσουν διαφορετικές τάσεις του τραπεζικού τομέα (Nobanee, 2021) ή στην περίπτωση της εργασίας, την προσέγγιση στον δανεισμό P2P.

2.1. ΜΕΛΕΤΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΔΑΝΕΙΟΔΟΤΗΣΗ

Αναφερόμενοι στις δανειοδοτήσεις P2P και με γνώμονα την υπάρχουσα έρευνα πάνω σε αυτόν τον τομέα διατυπώνεται (Gonzalez, Gonzalez, & Komarova, 2014) πως οι χρηματοοικονομικές, δημογραφικές και κοινωνικές μεταβλητές είναι οι καθοριστικοί παράγοντες του επιτυχημένου δανεισμού στην αγορά δανεισμού. Ορισμένες από αυτές τις μεταβλητές όχι μόνο επηρεάζουν την πιθανότητα χρηματοδότησης αλλά επηρεάζουν και τους τόκους του δανείου που δημιουργείται μέσω των πλατφορμών δανεισμού P2P (Bachman, και συν., 2011).

Υπάρχει διχασμός μεταξύ των αποτελεσμάτων των ερευνών σχετικά με το αν υπάρχει ρατσιστική προκατάληψη, το λεγόμενο Lender Profiling, σχετικά με την καταγωγή, την υγεία, την επαγγελματική κατάσταση. Συγκεκριμένα ο Μίκαλ Χερζενστάιν, καθηγητής του Πανεπιστημίου του Ντέλαγουερ (Herzenstein, Andrews, Dholakia, & Lyandres, 2008) συμπερασματολογεί πως η επιτυχία χρηματοδότησης στην Online αγορά P2P επηρεάζεται από την οικονομική ευρωστία των δανειοληπτών, τον βαθμό κινήτρων για εγγραφή και δημοσιοποίηση και δημογραφικές μεταβλητές, ωστόσο οι δημογραφικές μεταβλητές όπως η φυλή και το φύλο έχουν πολύ μικρή επίδραση στην επιτυχία χρηματοδότησης σε σύγκριση με τις άλλες μεταβλητές όπως η οικονομική ευρωστία των δανειοληπτών. Οι χρηματοοικονομικές μεταβλητές, τα προσωπικά χαρακτηριστικά και η σύσταση των πλατφορμών με τη μορφή της απόδοσης βαθμών πίστωσης στους δανειολήπτες λειτουργούν ως μεσολαβητές μεταξύ των δανειοληπτών και των δανειστών για μια επιτυχημένη συναλλαγή δανεισμού. Υποστηρίζει επίσης ότι οι δανειστές στην διαδικτυακή αγορά P2P λαμβάνουν την απόφαση χρηματοδότησης πιο δίκαια σε σχέση με την συμβατική μέθοδο δανεισμού στις ΗΠΑ όπου οι διακρίσεις τεκμηριώνονται καλά από τους μελετητές. Τέλος τεκμηριώνεται πως η αγορά P2P έχει ρόλο στη μείωση της πρακτικής διακρίσεων στον χρηματοπιστωτικό τομέα. Αντίθετα η μελέτη (Pope & Syndor, 2010) δείχνει την ύπαρξη διακρίσεων στη διαδικασία υποβολής προσφορών που αποδεικνύεται από τα στοιχεία του prosper.com. Η διατριβή αποκαλύπτει ότι οι Αμερικάνοι Αφρικανικής Γενεαλογίας, υπέρβαροι και ηλικιωμένοι υποψήφιοι υφίστανται διακρίσεις με υψηλότερα επιτόκια σε σύγκριση με αυτά των λευκών και των νέων. Τα άτομα με τη στρατιωτική ένωση ευνοούνται με καλύτερους όρους δανεισμού από τους άλλους αιτούντες.

Η οικονομική δυνατότητα ενός δανειολήπτη μπορεί να περιοριστεί σε τρεις κύριους παράγοντες οι οποίοι είναι επικρατέστεροι και ισχυρότεροι σε σχέση με τα υπόλοιπα χαρακτηριστικά που μπορεί κάποιος να ελέγξει προκειμένου να λάβει την απόφαση δανειοδότησης –η αναλογία χρέους προς εισόδημα (DTI, Debt to Income Ratio) ,αν το άτομο έχει σπίτι (House Ownership) καθώς και η σχέση εργασίας με τον Ενώ οι δύο πρώτες από αυτές τις μεταβλητές είναι άμεσοι δείκτες της πιστοληπτικής ικανότητας ενός δανειολήπτη, η τρίτη, η ιδιοκτησία σπιτιού, είναι ενδεικτική της σταθερότητας και της προηγούμενης δυνατότητας πρόσβασης σε πίστωση για την απόκτηση υποθήκης. Ένας μεγάλος αριθμός εμπειρικών μελετών που εξετάζουν τον καταναλωτικό δανεισμό έχουν δείξει ότι η οικονομική ευρωστία των δανειοληπτών παίζει σημαντικό ρόλο στην ικανότητά τους να λαμβάνουν εξασφαλισμένα και μη εξασφαλισμένα δάνεια από συμβατικά χρηματοπιστωτικά ιδρύματα (Herzenstein, Andrews, Dholakia, & Lyandres, 2008).

2.3. ΜΕΛΕΤΗ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΤΟΥ LENDERSCLUB

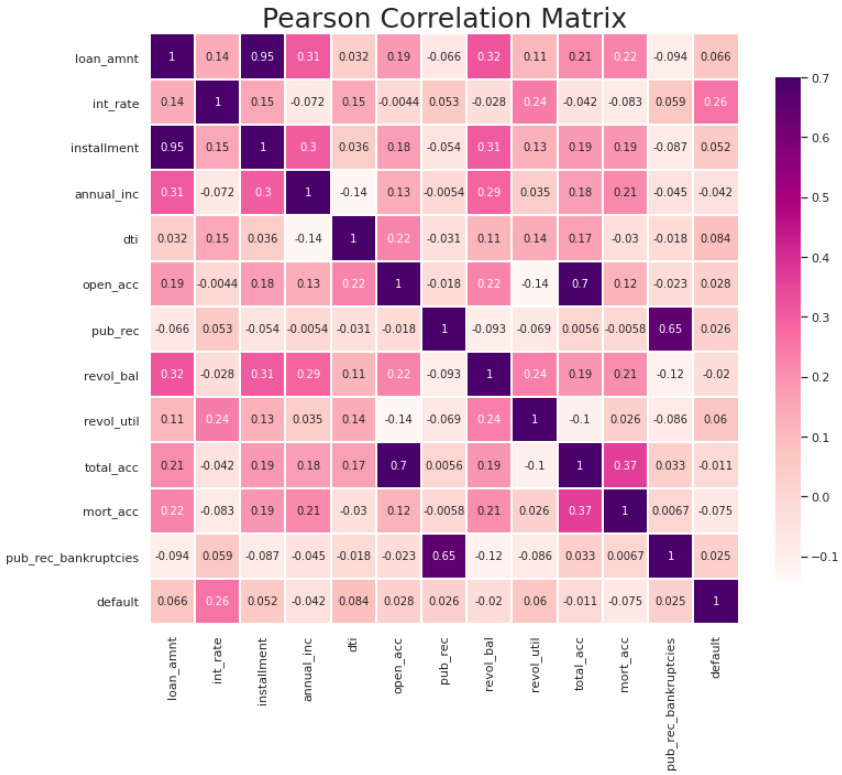
Το Σύνολο Δεδομένων του Lenders Club (George, 2019) αποτελεί ιδανικό παράδειγμα συνόλου δεδομένων καθώς διαθέτει και εμπεριέχει δεινά δεδομένα για παραμετροποίηση αλλά και δυσκολίες εκ ’φύσεως, με αποτέλεσμα να δημιουργείται η ανάγκη για ανάδραση με τα δεδομένο ώστε να γίνει σωστή εκροή της πληροφορίας, αλλά και ευκαιρία αντίκρουσης και εκμηδένισης δυσκολιών της τεχνολογίας. Σε αυτά τα δεδομένα υπάρχει ήδη μελέτη συσχετίσεων μεταβλητών καθώς και πρόβλεψης αθέτησης δανείων. Συνεπώς αντλούνται αποτελέσματα για αποσαφήνιση της πληροφορίας που χρειαζόμαστε αλλά είναι και εμφανείς οι τεχνικές στις γλώσσες Python και R ώστε να γίνει η σωστή επικύρωση των αποτελεσμάτων με σκοπό να χρησιμοποιηθούν για τους σκοπούς της εργασίας.

Η Kaggle, θυγατρική της Google LLC, είναι μια διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών μηχανικής μάθησης. Επιτρέπει στους χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να εξερευνούν και να δημιουργούν μοντέλα σε ένα διαδικτυακό περιβάλλον επιστήμης δεδομένων, να συνεργάζονται με άλλους επιστήμονες δεδομένων και μηχανικούς μηχανικής μάθησης και να συμμετέχουν σε διαγωνισμούς για την επίλυση προκλήσεων της επιστήμης δεδομένων. Στις παροχές της συγκεκριμένης κοινότητας παρατίθεται και το σύνολο δεδομένων που χρησιμοποιούμε στην εργασία.

2.3.1. ΤΕΧΝΙΚΗ ΠΡΟΒΛΕΨΗΣ ΤΟΥ ΠΟΣΟΣΤΟΥ ΑΘΕΤΗΣΗΣ ΣΕ ΟΜΟΤΙΜΑ ΔΑΝΕΙΑ

Σε αυτή την ενότητα γίνεται μελέτη των αποτελεσμάτων και των τεχνικών που χρησιμοποίησε ο Alwin Fassbender (Fassbender, 2021) σε ότι αφορά την Αθέτηση Δανείων. Μας ενδιαφέρουν τα μεγέθη που καταλήγει να μελετήσει. Στην συγκεκριμένη εργασία, ο Αναλυτής παραθέτει τα προβλήματα που οι διάφορες μέθοδοι αποθήκευσης δεδομένων κληροδοτούν στην εκάστοτε ανάλυση. Συγκεκριμένα αναφέρει πως ορισμένες εταιρείες δεν διαθέτουν σύγχρονη υποδομή πληροφορικής, όπου πολλά δεδομένα είτε δεν είναι διαθέσιμα σε ψηφιακή μορφή, είτε τα διαφορετικά τμήματα έχουν διαφορετικά συστήματα πληροφορικής (πίνακες Excel, βάσεις δεδομένων SQL) και μη τυποποιημένα ονόματα μεταβλητών (Fassbender, 2021). Η πρώτη ενέργεια σε αυτή την εργασία ήταν να ελεγχθούν αρχικά οι μεταβλητές που εμπεριέχονται στο υποσύνολο δεδομένων των απορριφθέντων. Χρησιμοποιεί τον τρόπο της Επεξηγηματικής

Ανάλυσης και εκρέει το συμπέρασμα πως οι μεταβλητές και οι τιμές τους στα Rejected δάνεια, απορριφθέντα, δεν έχουν καμία αξία στην πρόβλεψη ποσοστού αθέτησης. Ως εκ ’τούτου τα απορρίπτει και προχωράει σε έλεγχος συσχέτισης των Approved Loans και βρίσκει τα παρακάτω σε μορφή Πίνακα συσχέτισης Pearson, όπου ο συντελεστής συσχέτισης κυμαίνεται από −1 έως 1. Μια απόλυτη τιμή ακριβώς 1 υποδηλώνει ότι μια γραμμική εξίσωση περιγράφει τέλεια τη σχέση μεταξύ X και Y, με όλα τα σημεία δεδομένων να βρίσκονται σε μια ευθεία. μια τιμή +1 υποδηλώνει ότι όλα τα σημεία δεδομένων βρίσκονται σε μια γραμμή για την οποία το Y αυξάνεται καθώς το X αυξάνεται και αντίστροφα για το -1.



ΕΙΚΟΝΑ 2. ΣΥΣΧΕΤΙΣΗ ΜΕΓΕΘΩΝ (FASSBENDER, 2021)

Δεν θα γίνει περαιτέρω αναφορά στην παραπάνω εργασία που παρατέθηκε καθώς ο ρόλος που εξυπηρετεί είναι η άντληση δεδομένων και η σχέση μεταξύ των μεταβλητών.

2.3.2. ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (EDA)

Σε αυτή την εργασία, το ερώτημα που θέτει ο ερευνητής (Peterson, 2020) αφορά ποια είναι τα κριτήρια και τι επιλογή γίνεται σχετικά με τα δεδομένα και τις μεταβλητές προκειμένου ένας χρηματοπιστωτικός οργανισμός να εγκρίνει ή να απορρίψει ένα δάνειο. Αναφέρει πως για προφανείς λόγους οι πρώτες ενέργειες που αφορούν πιστωτικούς ελέγχους είναι η αίτηση λίστας εισοδημάτων και περιουσιακών στοιχείων, έλεγχος οικονομικού ιστορικού και του προηγούμενου ιστορικού δανείων του πιθανού δανειολήπτη.

Η λεγόμενη διαχείριση πιστωτικού κινδύνου εντός των τραπεζών πραγματοποιείται μέσω της θέσπισης πολιτικών χορήγησης πιστώσεων, οδηγιών και συντονισμού μεταξύ των διαφόρων τμημάτων της τράπεζας, όπως η πλήρης εποπτεία και έλεγχος της πιστωτικής έρευνας των πελατών, οι επιλογές τρόπων πληρωμής, η αναπροσαρμογή των πιστωτικών ορίων. Οι τράπεζες έχουν εγγύηση ότι θα ανακτήσουν τις απαιτήσεις με ασφάλεια στο χρόνο (Aebi V, 2012), οι οποίες ορίζονται και διασφαλίζονται από τα παραπάνω.

Ωστόσο, υπάρχει το φαινόμενο του «πιστωτικού παράδοξου» στην πρακτική της διαχείρισης πιστωτικού κινδύνου. Αυτό το επονομαζόμενο «πιστωτικό παράδοξο» είναι, η θεωρία διαχείρισης κινδύνου που απαιτεί από τις τράπεζες να ακολουθούν τις αρχές της αποκέντρωσης και διαφοροποίησης των επενδύσεων στη διαχείριση του τραπεζικού πιστωτικού κινδύνου, για να αποτρέψουν τη συγκέντρωση της άδειας πίστωσης (Reinhart, Levich, & Majoni, 2002). Η διαφοροποίηση είναι ακόμη πιο σημαντική και ο χρυσός κανόνας που πρέπει να υπακούει κανείς, καθώς ιδίως το παραδοσιακό μοντέλο διαχείρισης πιστωτικού κινδύνου στερείται αποτελεσματικής αντιστάθμισης πιστωτικού κινδύνου (Reinhart, Levich, & Majoni, 2002). Από την άλλη πλευρά, στην πράξη, η δραστηριότητα των τραπεζικών δανείων δείχνει συχνά ότι η αρχή της διαφοροποίησης δεν είναι εύκολο να εφαρμοστεί, επειδή πολλές τράπεζες δεν τηρούν πολύ τον κανόνα διαφοροποίησης στις δανειακές τους δραστηριότητες (Aebi V, 2012).

Ο Peterson χρησιμοποίησε ένα μικρότερο υποσύνολο των εγκεκριμένων δεδομένων δανείου, το οποίο συμπεριλαμβάνει το πεδίο Ποσό δανείου (loan_amnt), Λόγος χρέους προς εισόδημα (dti), Επιτόκιο (int_rate) και Μήκος απασχόλησης (emp_length), τα οποία περιγράφονται παρακάτω (Peterson, 2020).

Field	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
int_rate	Interest Rate on the loan
FICO_Average	The average of the fico_range_high and fico_range_low fields
fico_range_high	The upper boundary ranges the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary ranges the borrower's FICO at loan origination belongs to.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

ΕΙΚΟΝΑ 3. ΜΕΓΕΘΗ ΓΙΑ ΤΗΝ EDA (PETERSON, 2020)

Επίσης χρησιμοποίησε ορισμένα ισοδύναμα πεδία στα απορριφθέντα δεδομένα δανείου, συμπεριλαμβανομένων των πεδίων Αίτηση ποσού, Αναλογία χρέους προς εισόδημα, Διάρκεια απασχόλησης και Βαθμολογία κινδύνου, τα οποία περιγράφονται επίσης παρακάτω (Peterson, 2020).

Field	Description
Amount Requested	The total amount requested by the borrower
Risk_Score	For applications prior to November 5, 2013 the risk score is the borrower's FICO score. For applications after November 5, 2013 the risk score is the borrower's Vantage score.
Debt-To-Income Ratio	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
Employment Length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

ΕΙΚΟΝΑ 4. ΜΕΓΕΘΗ ΑΠΟΡΡΙΦΘΕΝΤΩΝ ΔΑΝΕΙΩΝ (PETERSON, 2020)

2.3.2.1. ΕΠΙΛΟΓΗ ΜΟΝΟΜΕΤΑΒΛΗΤΩΝ (**UNIVARIATE SELECTION)**

Η επιλογή μονομεταβλητών περιλαμβάνει τη χρήση στατιστικών μετρήσεων, όπως η δοκιμή τετραγώνου χ για να καθοριστεί ποιο χαρακτηριστικό ενός συνόλου δεδομένων εισόδου είναι πιο προβλέψιμο για μια συγκεκριμένη έξοδο γνωστή ως ετικέτα. Το τεστ Χ-τετράγωνο (chi-square test) του Pearson χρησιμοποιείται για να προσδιοριστεί εάν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των αναμενόμενων συχνοτήτων και των παρατηρούμενων συχνοτήτων σε μία ή περισσότερες κατηγορίες σε ένα διασταυρωμένο πίνακα (Cross Tabulation). Σε αυτή την περίπτωση, όσο χαμηλότερη είναι η βαθμολογία του πεδίου εισαγωγής, τόσο πιο προβλέψιμη είναι η ετικέτα.

Chi-Square Test for Rejected Loans

	Field	Score
0	amount_requested	2.059962e+06
1	debt_to_income_ratio	1.970787e+07
2	employment_length	6.990179e+02

ΕΙΚΟΝΑ 5. ΈΛΕΓΧΟΣ ΤΕΤΡΑΓΩΝΟΥ Χ ΓΙΑ ΤΑ ΑΠΟΡΡΙΦΘΕΝΤΑ ΔΑΝΕΙΑ (PETERSON, 2020)

Ο παραπάνω πίνακας υποδεικνύει ότι η μεταβλητή emp_length είναι πιο προειδοποιητική για τη βαθμολογία κινδύνου (Risk Score) για έναν υποψήφιο και οι μεταβλητές debt_to_income_ratio καθώς και η μεταβλητή amount_requested είναι ομοίως προγνωστικές της βαθμολογίας κινδύνου (Peterson, 2020).

Ο πίνακας για τα αποδεκτά δάνεια δείχνει ότι το emp_length είναι το πιο προβλέψιμο για το επιτόκιο, ακολουθούμενο από το λόγο χρέους προς εισόδημα, Μέση βαθμολογία FICO και η λιγότερο προβλέψιμη μεταβλητή είναι το ποσό του δανείου (loan_amnt) (Peterson, 2020).

Chi-Square Test for Accepted Loans

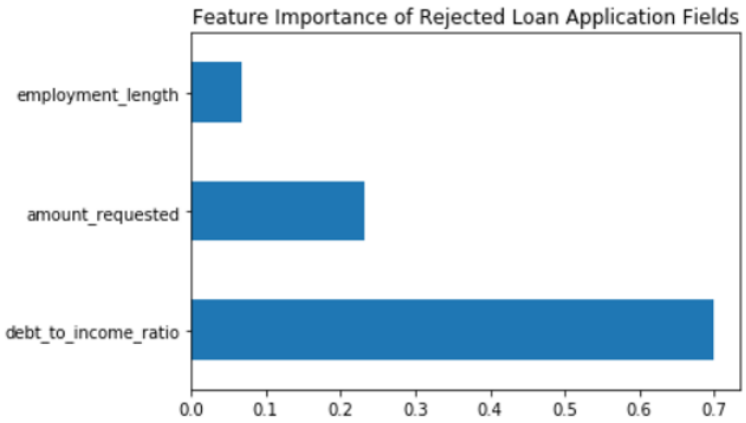
	Field	Score
0	loan_amnt	251698.386560
1	dti	297.226321
2	emp_length	54.408844
3	FICO_Average	430.047898

ΕΙΚΟΝΑ 6. ΈΛΕΓΧΟΣ ΤΕΤΡΑΓΩΝΟΥ Χ ΓΙΑ ΤΑ ΕΓΚΕΚΡΙΜΕΝΑ ΔΑΝΕΙΑ (PETERSON, 2020)

2.3.2.2. ΣΗΜΑΣΙΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Μια εναλλακτική λύση για αντί της επιλογής Univariate είναι η χρήση μετρήσεων βάσει σημασίας χαρακτηριστικών που χρησιμοποιεί έναν ταξινομητή για να καθορίσει τον βαθμό της σχέσης μεταξύ ενός πεδίου εισαγωγής και της ετικέτας (Peterson, 2020).

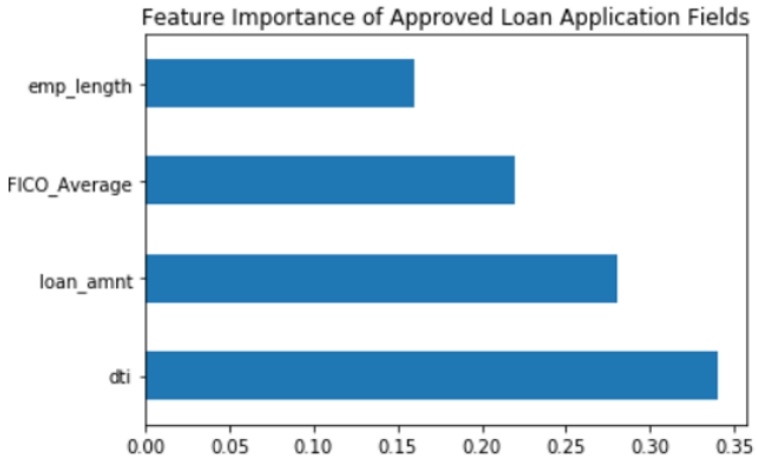
Χρησιμοποιώντας έναν ταξινομητή «Extra Trees Classifier», κλάση η οποία υλοποιεί έναν μετα-εκτιμητή που ταιριάζει σε έναν πλήθος τυχαιοποιημένων δέντρων αποφάσεων (extra trees) σε διάφορα υποδείγματα του συνόλου δεδομένων και χρησιμοποιεί τη μέση τιμή για τη βελτίωση της προγνωστικής ακρίβειας και του ελέγχου της υπερπροσαρμογή. Χρησιμοποιώντας αυτόν τον τύπο ταξινομητή στα δεδομένα των απορριφθέντων δανείων το ακόλουθο σύνολο αποτελεσμάτων:



ΕΙΚΟΝΑ 7. ΣΗΜΑΣΙΑ ΜΕΓΕΘΩΝ ΤΩΝ ΑΠΟΡΡΙΦΘΕΝΤΩΝ ΔΑΝΕΙΩΝ (PETERSON, 2020)

Σε αυτήν την περίπτωση, όσο υψηλότερη είναι η βαθμολογία, τόσο πιο προγνωστική είναι η μεταβλητή εισόδου της ετικέτας. Έτσι, χρησιμοποιώντας τον πρόσθετο ταξινομητή δέντρων, το dti είναι πολύ πιο προβλέψιμο για το δείκτη κινδύνου και ακόμη και το amnt_requested είναι πιο προβλέψιμο από το emp_length (Peterson, 2020).

Η κατανομή της σημασίας των χαρακτηριστικών για τις μεταβλητές που σχετίζονται με τα εγκεκριμένα δάνεια είναι πιο στενά ομαδοποιημένη, με το dti να είναι ελαφρώς πιο προβλέψιμο από το ποσό (amount requested) και οι υπόλοιπες μεταβλητές να ακολουθούν παρόμοια τάση στο βαθμό της προβλεψιμότητάς τους (Peterson, 2020).

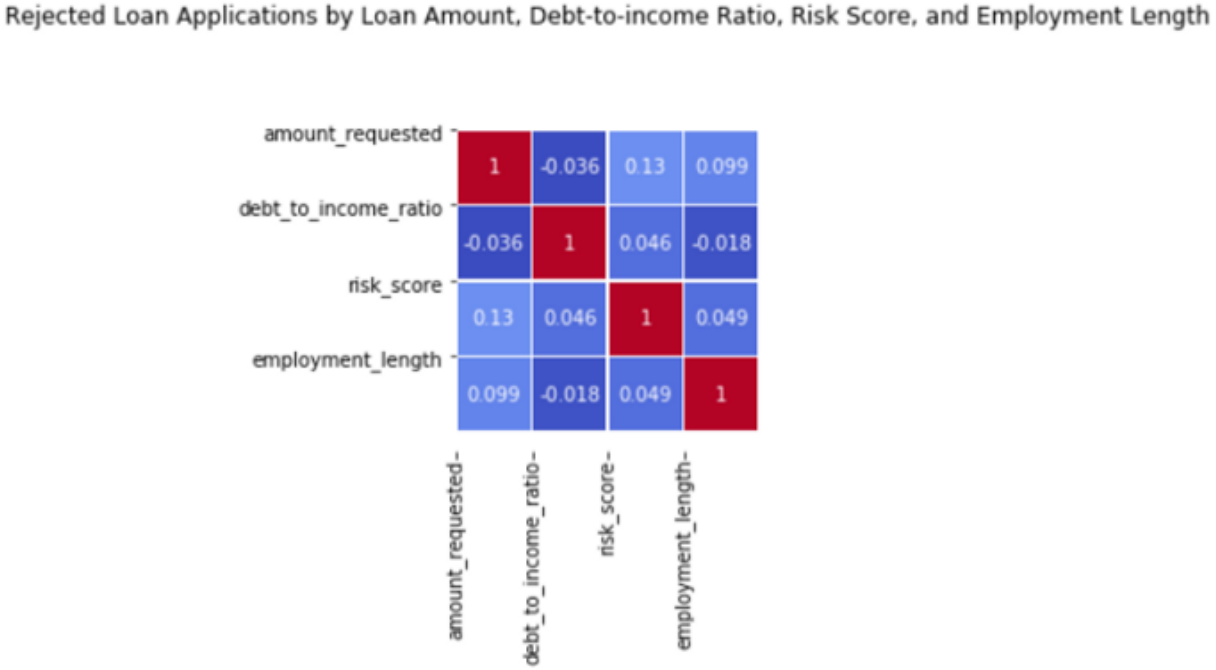


ΕΙΚΟΝΑ 8. ΣΗΜΑΣΙΑ ΜΕΓΕΘΩΝ ΤΩΝ ΕΓΚΕΚΡΙΜΕΝΩΝ ΔΑΝΕΙΩΝ (PETERSON, 2020)

2.3.2.3. Συντελεστης Συσχετισης pearson με Heatmap

Χρησιμοποιώντας βασικές περιγραφικές στατιστικές όπως ο συντελεστής συσχέτισης Pearson, μπορούμε να πάρουμε μια βασική εικόνα της σχέσης όλων των μεταβλητών σε ένα σύνολο δεδομένων.

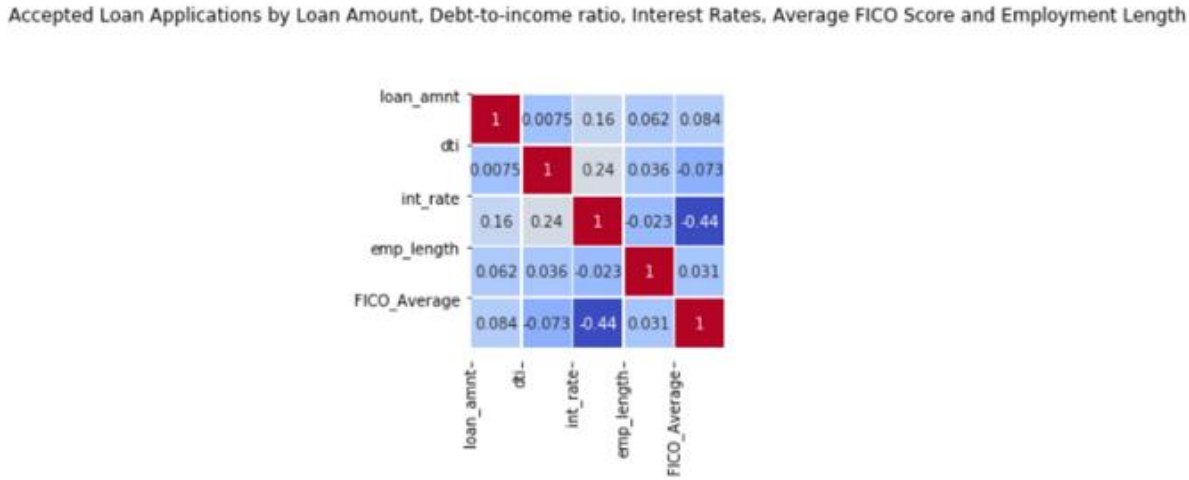
Ένας από τους ευκολότερους τρόπους να αποτυπωθεί η συσχέτιση μεταξύ δύο μεταβλητών είναι να σχεδιαστεί ένα Heatmap plot (Peterson, 2020). Σε αυτήν την περίπτωση, τόσο τα πεδία εισόδου όσο και η ετικέτα σχεδιάζονται κατά μήκος του άξονα X και του άξονα Y, με μια αριθμητική βαθμολογία μεταξύ [-1,00 και 1,00] που αντιπροσωπεύει τη σχέση μεταξύ των μεταβλητών κατά μήκος κάθε άξονα.



Εικονα 9. Σχετικοτητα Μεγεθων απορριφθεντων δανειων (Peterson, 2020)

Οι μεταβλητές με τον υψηλότερο βαθμό θετικής σχετικότητας είναι το amount_requested και ο δείκτης κινδύνου (risk_score) και οι μεταβλητές με τον υψηλότερο βαθμό αρνητικής σχετικότητας είναι το amnt_req με το dti (Peterson, 2020).

Σε ότι αφορά τα εγκεκριμένα δάνεια οι μεταβλητές με τον υψηλότερο βαθμό θετικής συγγένειας είναι dti και int_rate και οι μεταβλητές με τον υψηλότερο βαθμό αρνητικής συγγένειας είναι FICO_Average και int_rate (Peterson, 2020).



ΕΙΚΟΝΑ 10. ΒΑΘΜΟΣ ΣΥΣΧΕΤΙΣΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΤΑ ΕΓΚΕΚΡΙΜΕΝΑ ΔΑΝΕΙΑ

Η συγκεκριμένη ανάλυση καταλήγει στο συμπέρασμα πως δεν υπάρχει καμία μεταβλητή που να μπορεί να προβλέψει εάν μια αίτηση δανείου θα γίνει αποδεκτή ή θα απορριφθεί με βάση τη συνολική βαθμολογία κινδύνου του αιτούντος, εκτός από τους φυσικούς κανόνες που μπορεί να έχει θέσει ο οργανισμός. Ωστόσο, χρησιμοποιώντας τεχνικές Univariate, Feature Selection και Correlation σε ένα υποσύνολο εγκεκριμένων και απορριφθέντων δεδομένων αίτησης δανείου από το Lending Club, μπορούμε να έχουμε μια βασική αίσθηση για το σχετικό βαθμό στον οποίο συγκεκριμένες μεταβλητές μπορεί να επηρεάσουν τις ετικέτες εντός της αίτησης δανείου και εάν αυτές οι τάσεις είναι συνεπείς χρησιμοποιώντας διαφορετικές τεχνικές μοντελοποίησης (Peterson, 2020).

3. ΜΕΘΟΔΟΛΟΓΙΑ ΈΡΕΥΝΑΣ

Αυτή η ενότητα περιγράφει τη διαδικασία με τον οποίο γίνεται η εργασία για τους σκοπούς της επίτευξης των ερευνητικών στόχων. Η ενότητα περιλαμβάνει θέματα όπως ο σχεδιασμός της έρευνας, την συλλογή δεδομένων, την επεξεργασία δεδομένων και τις τεχνικές ανάλυσης δεδομένων.

3.1. ΣΧΕΔΙΑΣΜΟΣ ΜΕΛΕΤΗΣ

Το χρονικό αποτύπωμα των δεδομένων αφορά τα έτη 2007-2018 (Q4). Η περιγραφική στατιστική ανάλυση γίνεται για να διερευνηθούν οι ομοιότητες και οι διαφορές, όσον αφορά τα χαρακτηριστικά του δανείου, καθώς και τα μοτίβα που σχηματίζονται όταν ορισμένα συνδυάζονται ώστε να επιλεχθούν οι κατάλληλες μεταβλητές για την δημιουργία του δείκτη.

Σχετικά με το σύνολο δεδομένων και συγκεκριμένα τα αποδεκτά δάνεια, περιλαμβάνει 2260701 αντικείμενα που εμφανίζονται ως «γραμμές» από τα οποία τρείς (3) δεν επικυρώνονται ως σειρές και έτσι αφαιρούνται στην αρχή. Φαίνεται ότι υπάρχουν εκατό πενήντα μία (151) μεταβλητές, που εμφανίζονται ως 'στήλες', οι οποίες αντιπροσωπεύουν διαφορετικά χαρακτηριστικά από τα οποία αφαιρούνται οι πενήντα οκτώ (58) λόγω τιμών N/A. Επιπλέον, τα περισσότερα από αυτά δεν είναι σχετικά με την αξιολόγηση δανειοληπτών, ωστόσο δεν επηρεάζουν τις μεθόδους εκκαθάρισης. Αυτή η εξαίρεση είναι ασήμαντη λαμβάνοντας υπόψη τη συνολική αναλογία τους στο σύνολο δεδομένων. Πολλά από τα στατιστικά μοντέλα δεν λειτουργούν καλά με μεταβλητές υψηλής συσχέτισης, ταύτισης, και για να αποφευχθεί μια τέτοια κατάσταση αυτές οι μεταβλητές δεν λαμβάνονται υπόψη.

3.2. ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Το Lenders Club, βάση του συνόλου δεδομένων, φέρει 2260698 δάνεια ως εγκεκριμένα για την περίοδο 2007-2018(Q4). Από αυτά το 78% αφορά πιστωτικές κάρτες και ενοποιήσεις χρεών. Το Σύνολο Δεδομένων χρειάζεται εκκαθάριση προκειμένου να γίνει μελέτη η οποία θα γίνει βάση κάποιων αναλυτικών μεθόδων που θα περιγράφουν παρακάτω προκειμένου να καταστεί χρήσιμο για τους σκοπούς της εργασίας.

3.2.1. ΠΕΡΙΓΡΑΦΗ ΜΕΓΕΘΩΝ

Στην πορεία της εργασίας θα γίνεται αποτύπωση των αποτελεσμάτων φρασεολογικά με τις ονομασίες των μεταβλητών, καθώς θα χρειαστεί για την σύγκριση τους. Οι παρακάτω πίνακες αφορούν τα μεγέθη που βρίσκονται στο Σύνολο των Δεδομένων τα οποία μπορούν να ανακτηθούν από τον ιστότοπο Kaggle.

ΠΙΝΑΚΑΣ 1. ΟΝΟΜΑΣΙΕΣ ΜΕΓΕΘΩΝ ΤΩΝ ΕΓΚΕΚΡΙΜΕΝΩΝ ΔΑΝΕΙΩΝ

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of

	delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range of the borrower’s FICO at loan origination belongs to.
fico_range_low	The lower boundary range of the borrower’s FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are RENT, OWN, MORTGAGE, OTHER
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.

int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The upper boundary range of the borrower's last FICO pulled belongs to.
last_fico_range_low	The lower boundary range of the borrower's last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts
member_id	A unique LC assigned Id for the borrower member.
mo_sin_old_il_acct	Months since oldest bank installment account opened
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
mths_since_rcnt_il	Months since most recent installment accounts opened
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
next_pymnt_d	Next scheduled payment date
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)

num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc	The number of open credit lines in the borrower's credit file.
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_act_il	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
policy_code	Publicly available policy_code=1 new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file

total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
total_rev_hi_lim	Total revolving high credit/credit limit
url	URL for the LC page with listing data.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
sec_app_fico_range_low	FICO range (high) for the secondary applicant
sec_app_fico_range_high	FICO range (low) for the secondary applicant
sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
sec_app_open_acc	Number of open trades at time of application for the secondary applicant
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant

hardship_flag	Flags whether the borrower is on a hardship plan
hardship_type	Describes the hardship plan offering
hardship_reason	Describes the reason the hardship plan was offered
hardship_status	Describes if the hardship plan is active, pending, canceled, completed, or broken
deferral_term	Number of months that the borrower is expected to pay less than the contractual monthly payment amounts due to a hardship plan
hardship_amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan
hardship_start_date	The start date of the hardship plan period
hardship_end_date	The end date of the hardship plan period
payment_plan_start_date	The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is the start of the three-month period in which the borrower is allowed to make interest-only payments.
hardship_length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan
hardship_dpd	Account days past due as of the hardship plan start date
hardship_loan_status	Loan Status as of the hardship plan start date
orig_projected_additional_accrued_interest	The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.
hardship_payoff_balance_amount	The payoff balance amount as of the hardship plan start date
hardship_last_payment_amount	The last payment amount as of the hardship plan start date
disbursement_method	The method by which the borrower receives their loan. Possible values are CASH, DIRECT_PAY
debt_settlement_flag	Flags whether the borrower, who has charged-off, is working with a debt-settlement company.
debt_settlement_flag_date	The most recent date that the Debt_Settlement_Flag has been set
settlement_status	The status of the borrower’s settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT
settlement_date	The date that the borrower agrees to the settlement plan
settlement_amount	The loan amount that the borrower has agreed to settle for
settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan
settlement_term	The number of months that the borrower settles for

ΠΙΝΑΚΑΣ 2. ΟΝΟΜΑΣΙΕΣ ΜΕΓΕΘΩΝ ΑΠΟΡΡΙΦΘΕΝΤΩΝ ΔΑΝΕΙΩΝ

RejectStats File	Description
Amount Requested	The total amount requested by the borrower
Application Date	The date which the borrower applied
Loan Title	The loan title provided by the borrower
Risk_Score	For applications prior to November 5, 2013, the risk score is the borrower's FICO score. For applications after November 5, 2013, the risk score is the borrower's Vantage score.
Debt-To-Income Ratio	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
Zip Code	The first 3 numbers of the zip code provided by the borrower in the loan application.
State	The state provided by the borrower in the loan application
Employment Length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Policy Code	publicly available policy_code=1 new products not publicly available policy_code=2

3.3. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Η Python είναι μια ερμηνευμένη, διαδραστική, αντικειμενοστραφής γλώσσα προγραμματισμού. Ενσωματώνει ενότητες, εξαιρέσεις, δυναμική πληκτρολόγηση, δυναμικούς τύπους δεδομένων πολύ υψηλού επιπέδου και κλάσεις. Υποστηρίζει πολλαπλά παραδείγματα προγραμματισμού πέρα από αντικειμενοστραφή προγραμματισμό, όπως διαδικαστικός και λειτουργικός προγραμματισμός. Η Python συνδυάζει αξιοσημείωτη ισχύ με πολύ σαφή σύνταξη. Διαθέτει διεπαφές σε πολλές συστημικές κλήσεις και βιβλιοθήκες συστήματος, καθώς και σε διάφορα συστήματα παραθύρων και είναι επεκτάσιμη σε C ή C ++. Μπορεί επίσης να χρησιμοποιηθεί ως γλώσσα επέκτασης για εφαρμογές που χρειάζονται προγραμματιζόμενη διεπαφή. Τέλος, η Python είναι φορητή καθώς έχει την δυνατότητα να τρέχει σε πολλές παραλλαγές Unix, συμπεριλαμβανομένου Linux και macOS, και σε Windows (Python Software Foundation, 2021).

Για την ανάλυση γίνεται χρήση της Python καθώς μέσα από την Dask μας δίνεται η δυνατότητα να ξεπεράσουμε περιορισμούς υπολογιστικής δύναμης και να γίνεται η ανάλυση γρηγορότερα. Φυσικά για το την χρήση της Python χρησιμοποιούμε την σουίτα PyCharm (Jet Brains Inc., 2021) καθώς είναι ανοιχτού κώδικα και μπορούμε να κάνουμε παραμετροποιήσεις αρκετά γρήγορα.

Τέλος τα διάφορα υποσύνολα που προκύπτουν από την Διερευνητική ανάλυση που πραγματοποιείται, εισάγονται και αναλύονται στο IBM SPSS προκειμένου να γίνει χρήση των εργαλείων του.

3.3.1. ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (EDA)

Η διερευνητική ανάλυση δεδομένων είναι μια προσέγγιση ανάλυσης συνόλων δεδομένων για τη σύνοψη των κύριων χαρακτηριστικών τους, χρησιμοποιώντας συχνά στατιστικά γραφικά και άλλες μεθόδους οπτικοποίησης δεδομένων. Ένα στατιστικό μοντέλο μπορεί να χρησιμοποιηθεί ή όχι, αλλά πρωτίστως το EDA είναι για να δούμε τι μπορούν να μας πουν τα δεδομένα πέρα από την τυπική εργασία μοντελοποίησης ή δοκιμής υποθέσεων. Χρησιμοποιείται προσαρμοσμένος κώδικας της κοινότητας Jovian για την ανάλυση με αλλαγές για την συμβατότητα των εντολών σε Python 3.9 και PyCharm 2020 Win.

Στις περισσότερες περιπτώσεις, το EDA αποτελείται από:

Την φυσική δομή του μεγέθους των δεδομένων, ο αριθμός των δειγμάτων και των χαρακτηριστικών, την παρουσία/απουσία τιμών που λείπουν, είτε τα χαρακτηριστικά είναι αριθμητικά είτε αντικείμενα συμβολοσειράς, τα οποία όλα έχουν σχέση με την περαιτέρω ανάλυση.

Την κατανομή των δεδομένων σε μεμονωμένα χαρακτηριστικά (κανονική, διωνυμική κ.λ.π.).

Την συσχέτιση των χαρακτηριστικών μεταξύ τους Σημαντικά συμπεράσματα που μπορούν να εξαχθούν από τα δεδομένα σε σχέση με τους επιχειρηματικούς στόχους για τη διεξαγωγή της ανάλυσης.

3.3.2 Λογιστικη Παλινδρομηση

Η δυαδική λογιστική παλινδρόμηση (binary Logistic Regression) προβλέπει την πιθανότητα ότι μια παρατήρηση εμπίπτει σε μία από τις δύο κατηγορίες μιας διχοτόμης εξαρτημένης μεταβλητής που βασίζεται σε μία ή περισσότερες ανεξάρτητες μεταβλητές που μπορεί να είναι είτε συνεχείς είτε κατηγορικές.

Ο λόγος που πραγματοποιείται η συγκεκριμένη ανάλυση γίνεται διότι καλύπτει τις ανάγκες ως προς τους κυριότερους στόχους της εργασίας. Το υποσύνολο είναι εκείνο που προκύπτει από την EDA, μεταξύ των απορριφθέντων και των εγκεκριμένων δανείων, με εξαρτημένη μεταβλητή την Έγκριση ή την Απόρριψη, χωρίς να λαμβάνεται υπόψιν το αν στα εγκεκριμένα δάνεια έγινε αποπληρωμή μετά την εκταμίευση του δανείου ή υπάρχει αθέτηση.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1. ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΙΕΡΕΥΝΗΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ (EDA)

4.1.1. ΕΥΡΕΣΗ ΚΟΙΝΩΝ ΜΕΓΕΘΩΝ ΜΕΤΑΞΥ ΤΩΝ ΔΥΟ ΥΠΟΣΥΝΟΛΩΝ (ΕΓΚΕΚΡΙΜΕΝΑ ΔΑΝΕΙΑ & ΑΠΟΡΡΙΦΘΕΝΤΑ ΔΑΝΕΙΑ)

ΠΙΝΑΚΑΣ 3. ΚΟΙΝΑ ΜΕΓΕΘΗ ΜΕΤΑΞΥ ΕΓΚΕΚΡΙΜΕΝΩΝ ΚΑΙ ΑΠΟΡΡΙΦΘΕΝΤΩΝ ΔΑΝΕΙΩΝ

	accepted_headings	**rejected_headings**
0	loan_amnt	Amount Requested
1	title	Loan Title
2	dti	Debt-To-Income Ratio
3	zip_code	Zip Code
4	addr_state	State
5	emp_length	Employment Length

4.1.2. ΠΛΗΘΟΣ ΕΓΚΕΚΡΙΜΕΝΩΝ ΚΑΙ ΑΠΟΡΡΙΦΘΕΝΤΩΝ

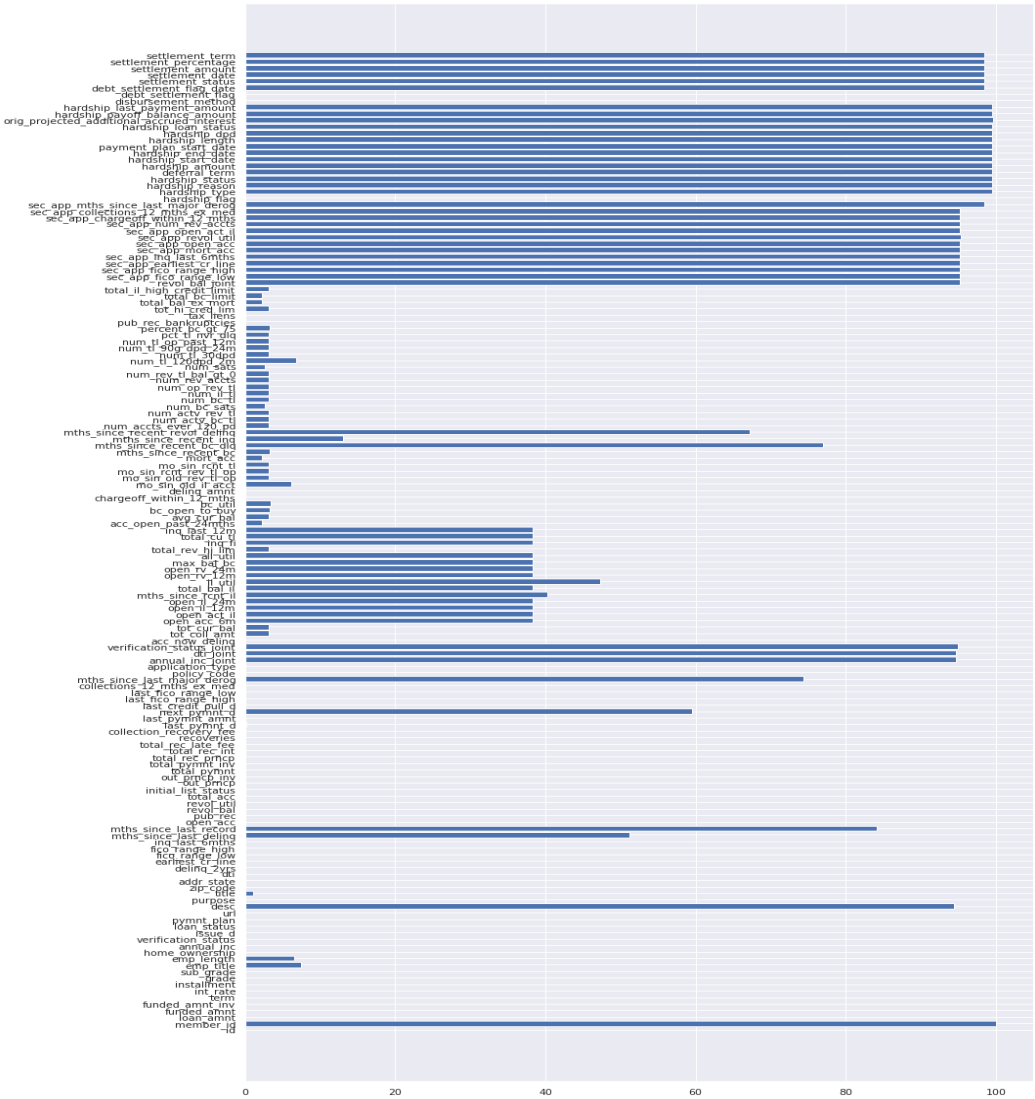
Το παραπάνω μας επιστρέφει 2260701 Εγκεκριμένα Δάνεια και 27648741 Απορριφθέντα. Οι ονομασίες των ‘κεφαλίδων’ μέσα στο Σύνολο Διαφέρει ενώ αναφέρεται στο ίδιο μέγεθος. Τα απορριφθέντα δάνεια υπερτερούν αριθμητικά σε σχέση με τα εγκεκριμένα 10:1.

Δειγματοληψία:

Διότι το Σύνολο Δεδομένων είναι μεγάλο και δεν θα χρησιμοποιηθεί ακόμα συγκεκριμένη βιβλιοθήκη για παράλληλη υπολογιστική γίνεται δειγματοληψία 20% του συνόλου για γρηγορότερους χρόνους λαμβάνοντας υπόψη την πιθανότητα ύπαρξης του bias αλλά και την πιθανότητα εσφαλμένης δειγματοληψίας αν υπάρχουν ανωμαλίες στο σύνολο.

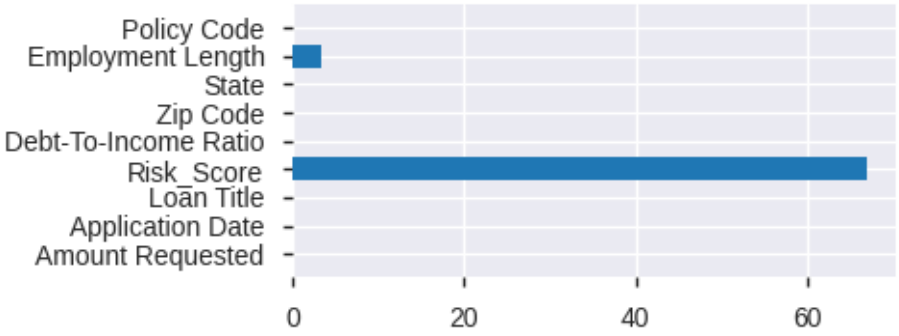
4.1.3. Εύρεση και Έλεγχος N/A τιμών στο Σύνολο Δεδομένων

Σχεδόν 44 χαρακτηριστικά στα Εγκεκριμένα Δάνεια λείπουν από το Dataset, πάνω από το 50% των εγγραφών, καθιστώντας έτσι αυτά τα χαρακτηριστικά αβάσιμα για περαιτέρω ανάλυση αλλά και 58 από αυτά έχουν pitfall στο 30%. Αυτά παρουσιάζονται παρακάτω με 'μπλε' χρώμα.



Εικονα 11. N/A τιμες στο συνολο εγκεκριμενων δανειων

Παρατηρούμε πως στα απορριφθέντα δάνεια απουσιάζει το 66% του μεγέθους Risk_Score. Αυτό ουσιαστικά αντιπροσωπεύει την βαθμολόγηση ρίσκου που έχει ο συγκεκριμένος πελάτης εντός του οργανισμού. Η απουσία τόσο μεγάλου ποσοστού είτε σημαίνει πως τα δεδομένα έχουν διασκευαστεί ή δεκαοχτώ εκατομμύρια (18000000) αιτήσεις δανείου απορρίφθηκαν χωρίς να περάσουν διαδικασία αξιολόγησης και υπολογισμού ρίσκου. Όπως αναφέρεται σε πανομοιότυπη ανάλυση στο Jovian (Ratadas, 2020) εικάζεται πώς ο λόγος χρέους προς εισόδημα ,Debt-To-Income Ratio (DTI), και η διάρκεια απασχόλησης ,Employment Length, είναι βασικοί παράγοντες για απόρριψη.



Εικονα 12. . N/A Τιμες στο Συνολο των Απορριφθεντων Δανειων

4.1.4. Αφαιρεση τιμων N/A και διπλοτυπων απο τα υποσυνολα

Κατά αρχήν όταν οι τιμές N/A ξεπερνάνε το 30% του συνόλου της κάθε στήλης (column) τότε αυτό το μέγεθος θα πρέπει να αφαιρείται από το σύνολο. Σε περίπτωση που τα μεγέθη που θα αφαιρεθούν αναπαριστούν το 30% του συνόλου των μεγεθών τότε είναι θεμιτό να γίνει μελέτη και ανάλυση για την αναγκαιότητά τους ως προς το σύνολο, για να αποφευχθούν εσφαλμένα συμπεράσματα.

Στα Εγκεκριμένα Δάνεια, 58 μεγέθη είναι προς αφαίρεση βάση των παραπάνω. Ο παρακάτω πίνακας αναφέρει τα μεγέθη που αφαιρέθηκαν, το ποσοστό N/A τιμών που είναι πάνω από 70%.

Πινακας 4. Απουσιαζοντα δεδομενα πανω του 70%

Ονομασία Στήλης	N/A >70%
all_util	ΟΧΙ
annual_inc_joint	ΝΑΙ
debt_settlement_flag_date	ΝΑΙ
deferral_term	ΝΑΙ
desc	ΝΑΙ
dti_joint	ΝΑΙ
hardship_amount	ΝΑΙ
hardship_dpd	ΝΑΙ
hardship_end_date	ΝΑΙ
hardship_last_payment_amount	ΝΑΙ
hardship_length	ΝΑΙ
hardship_loan_status	ΝΑΙ
hardship_payoff_balance_amount	ΝΑΙ
hardship_reason	ΝΑΙ
hardship_start_date	ΝΑΙ
hardship_status	ΝΑΙ
hardship_type	ΝΑΙ
il_util	ΟΧΙ
inq_fi	ΟΧΙ
inq_last_12m	ΌΧΙ
max_bal_bc	ΌΧΙ
member_id	ΝΑΙ
mths_since_last_delinq	ΟΧΙ
mths_since_last_major_derog	ΝΑΙ
mths_since_last_record	ΝΑΙ
mths_since_rcnt_il	ΟΧΙ
mths_since_recent_bc_dlq	ΝΑΙ

mths_since_recent_revol_delinq	OXI
next_pymnt_d	OXI
open_acc_6m	OXI
open_act_il	OXI
open_il_12m	OXI
open_il_24m	OXI
open_rv_12m	OXI
open_rv_24m	OXI
orig_projected_additional_accrued_interest	NAI
payment_plan_start_date	NAI
revol_bal_joint	NAI
sec_app_chargeoff_within_12_mths	NAI
sec_app_collections_12_mths_ex_med	NAI
sec_app_earliest_cr_line	NAI
sec_app_fico_range_high	NAI
sec_app_fico_range_low	NAI
sec_app_inq_last_6mths	NAI
sec_app_mort_acc	NAI
sec_app_mths_since_last_major_derog	NAI
sec_app_num_rev_accts	NAI
sec_app_open_acc	NAI
sec_app_open_act_il	NAI
sec_app_revol_util	NAI
settlement_amount	NAI
settlement_date	NAI
settlement_percentage	NAI
settlement_status	NAI
settlement_term	NAI
total_bal_il	OXI
total_cu_tl	OXI

Τα παραπάνω μεγέθη θα αφαιρεθούν λόγω του πλήθους των N/A αλλά και διότι, αφορούν κυρίως μεγέθη που είτε αντιπροσωπεύονται ατομικά και όχι σε ομάδες λογαριασμών σε ίδιες συμβάσεις, π.χ. dti με dti-joint όπου η εγγραφή της ομαδοποίησης των dti δεν έχει νόημα εάν υπάρχει μόνο ένας ενεχόμενος.

4.1.5. ΕΞΟΜΑΛΥΝΣΗ ΔΕΔΟΜΕΝΩΝ, ΥΠΕΡ ΑΡΙΘΜΩΝ

Τα περισσότερα χαρακτηριστικά στο σύνολο δεδομένων αποδεκτών υποθέσεων δεν δείχνουν καμία συσχέτιση, ορισμένα χαρακτηριστικά εμφανίζουν κυρίως θετική συσχέτιση που σημειώνεται με κόκκινο χρώμα. Τα χαρακτηριστικά που εμφανίζουν θετική συσχέτιση είναι τα εξής:

1. Αριθμός λογαριασμών με καθυστέρηση 120 ή περισσότερες ημέρες (num_accts_ever_120_pd)
2. Αριθμός ενεργών ανακυκλούμενων συναλλαγών (num_actv_rev_tl)
3. Αριθμός λογαριασμών τραπεζικής κάρτας (num_bc_tl)
4. Αριθμός ανοιχτών ανακυκλούμενων λογαριασμών (num_op_rev_tl)

Τα παραπάνω χαρακτηριστικά σχετίζονται με τη δραστηριότητα των πιστωτικών προϊόντων που χρησιμοποιούν οι πιστωτές (Choudhry & Darell, 2011) και ως εκ τούτου υπάρχει πλεονασμός σε αυτήν την επιλογή χαρακτηριστικών σε αυτό το σύνολο δεδομένων.

ΠΙΝΑΚΑΣ 5. ΣΥΣΧΕΤΙΣΗ ΠΟΣΟΥ, ΒΑΘΜΟΛΟΓΙΑΣ ΡΙΣΚΟΥ ΚΑΙ DTI

	Amount Requested	**Risk_Score**	**Debt-To-Income Ratio**
Amount Requested	1.000000	0.246186	-0.000139
Risk_Score	0.246186	1.000000	-0.001841
Debt-To-Income Ratio	-0.000139	-0.001841	1.000000

Στο σύνολο δεδομένων των απορριφθέντων δανείων, υπάρχει κακή συσχέτιση μεταξύ του Risk_Score και των άλλων αριθμητικών χαρακτηριστικών. Αυτό υπογραμμίζει την ανάγκη για προσεκτική αξιολόγηση των δεδομένων καθώς δεν μπορεί να αφαιρεθεί σαν μέγεθος το Risk_Score ενώ ταυτόχρονα απουσιάζει περίπου το 66% των εγγραφών του.

4.1.6. Σύγκριση και Ανάλυση κοινών μεγεθών Εγκεκριμένων και Απορριφθέντων Δανείων:

Προκειμένου να γίνεται σύγκριση και ενοποίηση των 2 υποσυνόλων αφαιρούνται πρώτα τα outliers.

Πίνακας 6. Πίνακας Έλεγχου επιτυχους διαγραφης των outliers των Εγκεκριμενων Δανειων & Απορριφθεντων

Accepted	**loan_amnt**	**dti**	**emp_length**
count	222488.000000	224002.000000	226063.000000
mean	14645.999784	17.774493	5.645400
std	8685.804556	8.574079	3.770817
min	500.000000	0.000000	0.000000
25%	8000.000000	11.000000	2.000000
50%	12500.000000	17.000000	5.000000
75%	20000.000000	24.000000	10.000000
max	37950.000000	43.000000	10.000000

Rejected	**loan_amnt**	**dti**	**emp_length**
count	251861.000000	251861.000000	251861.000000
mean	12502.079369	20.282203	1.483711
std	10663.678381	16.682394	1.629744
min	0.000000	0.000000	0.000000
25%	5000.000000	6.000000	1.000000
50%	10000.000000	17.000000	1.000000
75%	20000.000000	31.000000	1.000000
max	42475.000000	77.000000	10.000000

Με στόχο την ενοποίηση των δύο υποσυνόλων, θα πρέπει να δημιουργηθεί το Πλαίσιο Δεδομένων (Dataframe) το οποίο πρέπει πρώτα να οριστεί ώστε τα δεδομένα να μην χάσουν την κληρονομικότητά τους ως αντικείμενα αλλά και να γίνει ομαδοποίηση και ενοποίηση στοιχείων βάση των χαρακτηριστικών τους όπως ως προς την Διάρκεια απασχόλησής τους.

ΠΙΝΑΚΑΣ 7. ΠΙΝΑΚΑΣ ΜΕ ΤΑ ΟΜΑΔΟΠΟΙΗΜΕΝΑ ΕΓΚΕΚΡΙΜΕΝΑ ΔΑΝΕΙΑ

Accepted	**emp_length**	**dti**	**loan_amnt**	**emp_count**	**status**
0	0.0	17.828431	14122.072484	14210	accepted
1	1.0	17.779014	14518.269260	33165	accepted
2	2.0	17.682888	14546.211857	20056	accepted
3	3.0	17.768015	14557.585745	17902	accepted
4	4.0	17.880638	14679.930586	13614	accepted
5	5.0	17.653979	14647.088272	13583	accepted
6	6.0	17.646447	14584.371215	9908	accepted
7	7.0	17.867380	14696.414136	9154	accepted
8	8.0	17.623198	14662.544356	9018	accepted
9	9.0	17.803005	14849.398319	7853	accepted
10	10.0	17.812219	14824.283273	74017	accepted

ΠΙΝΑΚΑΣ 8. ΠΙΝΑΚΑΣ ΜΕ ΤΑ ΟΜΑΔΟΠΟΙΗΜΕΝΑ ΑΠΟΡΡΙΦΘΕΝΤΑ ΔΑΝΕΙΑ

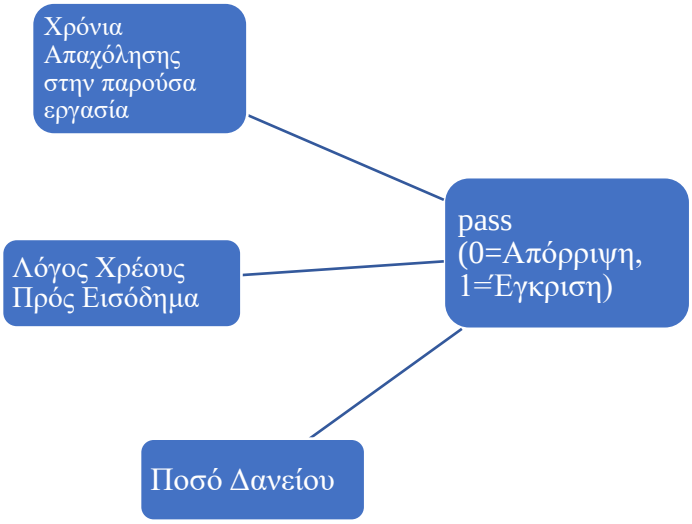
Rejected	**emp_length**	**dti**	**loan_amnt**	**emp_count**	**status**
0	0	20.87648773	12492.38242	9493	rejected
1	1	20.26735118	12483.86779	232689	rejected
2	2	19.88528529	12361.27346	1870	rejected
3	3	19.90086741	12655.46875	1733	rejected
4	4	20.85905442	12626.45254	1223	rejected
5	5	20.44292694	12347.62717	22911	rejected
6	6	18.51140065	12647.77607	656	rejected
7	7	19.27219	12941.09	534	rejected
8	8	18.39462	12469.86	701	rejected
9	9	20.72632	12252.26	498	rejected
10	10	19.94599	12758.11	4181	rejected

4.2. ΑΠΟΤΕΛΕΣΜΑΤΑ ΛΟΓΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (BINARY LOGISTIC REGRESSION)

Στην λογιστική παλινδρόμηση (binary Logistic Regression), στα εγκεκριμένα και απορριφθέντα δάνεια με εξαρτημένη μεταβλητή το STATUS όπου η τιμή 1 αναφέρεται σε Εγκεκριμένο Δάνειο και η τιμή 0 αναφέρεται σε Απορριφθέν Δάνειο, χρησιμοποιούμε τους τρείς πλέον προλέγοντες (predictors) που είναι οι μεταβλητές:

1. Emp_length
2. DTI
3. Loan_Amnt

Στην συγκεκριμένη έχει γίνει επιλογή και προλέγουσας ομάδας στην οποία θα ανήκουν οι περιπτώσεις της παλινδρόμησης προκειμένου να δούμε την ακρίβεια σε σχέση με το σύνολο δεδομένων. Χρησιμοποιούνται 250.000 εγκεκριμένα και 250.000 απορριφθέντα δάνεια ως 100% των περιπτώσεων για μεγαλύτερη ακρίβεια.



ΕΙΚΟΝΑ 13. ΕΝΝΟΙΟΛΟΓΙΚΟ ΜΟΝΤΕΛΟ

Classification Table[a]

Observed			Predicted Status 0	Predicted Status 1	Percentage Correct
Step 1	Status	0	233671	30203	88.6
		1	68913	130765	65.5
	Overall Percentage				78.6

a. The cut value is .500

Η ανάλυση έχει επιστρέψει ικανοποιητικό ποσοστό ολικής πρόβλεψης αποτελέσματος το οποίο είναι 78.6%. Δηλώνει πως η λογιστική παλινδρόμηση είναι ένα εύστοχο μοντέλο ανάλυσης δεδομένων με τους predictors που ορίσαμε παραπάνω για το συγκεκριμένο σύνολο δεδομένων. Το συνολικό ποσοστό υποδεικνύει το ποσοστό των περιπτώσεων με ένα παρατηρούμενο αποτέλεσμα που είχαν προβλεφθεί σωστά (ως προς το αποτέλεσμα) από το μοντέλο.

Στην συγκεκριμένη περίπτωση το **συνολικό ποσοστό** είναι 78,6%, υπολογιζόμενο ως:

$$\frac{233671+130765}{233671+30203+68913+130765} = 0{,}786 = 78{,}6\%$$

Τα ποσοστά στις δύο πρώτες σειρές παρέχουν πληροφορίες σχετικά με την ευαισθησία και την ειδικότητα του μοντέλο όσον αφορά την πρόβλεψη της ιδιότητας μέλους της ομάδας (εγκεκριμένα/απορριφθέντα) στην εξαρτημένη μεταβλητή. Η ευαισθησία αναφέρεται στο ποσοστό των περιπτώσεων που παρατηρήθηκε ότι εμπίπτουν στην ομάδα στόχο (Y=1, π.χ. Εγκεκριμένο Δάνειο) οι οποίοι είχαν προβλεφθεί σωστά από το μοντέλο να εμπίπτουν σε αυτήν την ομάδα (π.χ. προβλεπόμενη Έγκριση).

Η **ευαισθησία** για το μοντέλο υπολογίζεται ως:

$$\frac{130765}{68913 + 130765} = 0{,}655 = 65{,}5\%$$

Η ειδικότητα (specificity) αναφέρεται στο ποσοστό των περιπτώσεων που παρατηρήθηκε ότι εμπίπτουν στην κατηγορία μη-στόχου (ή αναφοράς) (π.χ. το δάνειο απορρίφθηκε) που είχε προβλεφθεί σωστά από το μοντέλο ότι εμπίπτουν σε αυτήν την ομάδα (π.χ., προβλέφθηκε πως θα απορριφθεί το δάνειο).

Η ειδικότητα για αυτό το μοντέλο υπολογίζεται ως:

$$\frac{233671}{233671 + 30203} = 0{,}886 = 88{,}6\%$$

Συνολικά, το ποσοστό ακρίβειας είναι αρκετά καλό 78,6%. Το μοντέλο παρουσιάζει μέτρια προς καλή ευαισθησία αφού μεταξύ εκείνων των δανειοληπτών για τους οποίους τα δάνεια Εγκριθήκανε, το 65,5% έχει προβλεφθεί σωστά ως υπόθεση Εγκεκριμένου Δανείου.

Το μοντέλο πολύ καλή εξειδίκευση, καθώς μεταξύ των δανειοληπτών για τους οποίους απορρίφθηκε το δάνειο, το 88,6% έχει προβλεφθεί πως θα απορριφθούν.

		B	Exp(B)
Step 1[a]	LOAN_AMOUNT	.000	1.000
	DTI	-.017	.983
	emp_length	.533	1.705
	Constant	-1.726	.178

Variables in the Equation

		95% C.I.for EXP(B)	
		Lower	Upper
Step 1[a]	LOAN_AMOUNT	1.000	1.000
	DTI	.983	.984
	emp_length	1.699	1.710
	Constant		

Η στήλη Exp(B) περιέχει αναλογίες πιθανοτήτων, οι οποίες είναι γενικά πιο εύκολο να κατανοήσουν την ερμηνεία από ό,τι τα logit. Οι δύο τελευταίες στήλες περιέχουν το διάστημα εμπιστοσύνης 95% για τις αναλογίες πιθανοτήτων. Σε γενικές γραμμές ισχύει η ερμηνεία ότι:

Εάν ένας λόγος πιθανοτήτων είναι 1, τότε υποδηλώνει ότι δεν υπάρχει αλλαγή στις πιθανότητες ανά μονάδα αύξησης στον προγνωστικό παράγοντα.

Εάν ένας λόγος πιθανοτήτων είναι > 1, τότε υποδηλώνει ότι οι πιθανότητες που σχετίζονται με τη συμμετοχή στην ομάδα-στόχο αυξάνονται με αυξήσεις στον προγνωστικό παράγοντα.

Εάν μια αναλογία πιθανοτήτων είναι < 1, τότε υποδηλώνει ότι οι πιθανότητες συμμετοχής στην ομάδα στόχο μειώνονται με αύξηση του δείκτη πρόβλεψης.

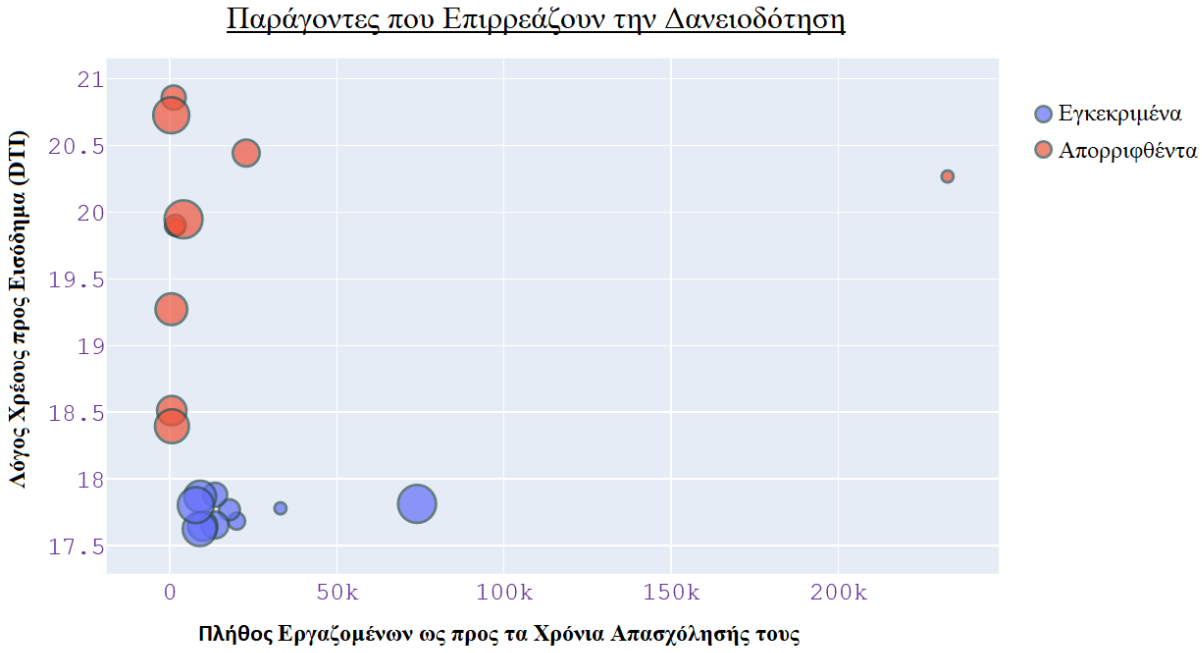
Κάθε λόγος πιθανοτήτων σε αυτόν τον πίνακα υποδεικνύει την πολλαπλασιαστική αλλαγή στις πιθανότητες (μιας περίπτωσης που εμπίπτει στην ομάδα-στόχο, ή Y=1) ανά μονάδα αύξησης σε έναν δεδομένο προγνωστικό παράγοντα, ελεγκτής (controller) για τους άλλους στο μοντέλο.

Το μοντέλο της λογιστικής παλινδρόμησης ενώ φαίνεται πως ανταποκρίνεται σωστά στην πρόγνωση του αποτελέσματος σχετικά με τα Εγκεκριμένα και τα Απορριφθέντα Δάνεια μας επιβεβαιώνει την σχέση μεταξύ των μεταβλητών όπου ουσιαστικά όσο αυξάνεται το DTI τόσο μειώνεται και

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρατηρείται πως οι δανειολήπτες οι οποίοι φέρουν χαμηλό Debt-to-Income ratio και περισσότερα χρόνια Απασχόλησης είναι αυτοί που κατά κύριο λόγο αποτελούν το σύνολο των Εγκεκριμένων Δανείων. Σε αντίθεση εκείνοι οι οποίοι έχουν είτε DTI μέτριο προς υψηλό ή λιγότερα χρόνια Απασχόλησης, συγκεκριμένα λιγότερα από 2 αλλά και συνδυασμός αυτών είναι εκείνοι οι οποίο δεν έλαβαν δανειοδότηση παρόλο που τα ποσά που ζήτησαν ήταν σχετικά χαμηλά σε σχέση με το Σύνολο.

Κατά την κατηγοριοποίηση των Δανειοληπτών σε σχέση με την Διάρκεια Απασχόλησής τους, παρατηρούμε πως υπάρχει εμφανής σχέση στην τάση δανειοδότησης βάση αυτού το οποίο επιβεβαιώνεται και από την αντιπαράθεση των αποτελεσμάτων με την βιβλιογραφία.



ΕΙΚΟΝΑ 14. DTI ΠΡΟΣ EMP_LENGTH

Ως εναρκτήριος άξονας χρησιμοποιηθήκαν τα απορριφθέντα δάνεια καθώς δεν θα υπήρχε τρόπος σύγκρισης των μεγεθών των δύο υποσυνόλων εάν δεν υπήρχε συνοχή και αντιστοιχία στα χαρακτηριστικά του κάθε υποσυνόλου. Η ομαδοποίηση των μεταβλητών μας έδωσε ξεκάθαρη εικόνα για το ποιες μεταβλητές είναι αυτές οι οποίες επηρεάζουν κατά προσέγγιση την δανειοδότηση σαν απόφαση βάση της μελέτης.

Είναι φυσικό ένας από τους κύριους παράγοντες δανειοδότησης να είναι ο λόγος χρέους ως προς το εισόδημα καθώς ο τρόπος αποπληρωμής του δανείου βασίζεται σε αυτό. Το γεγονός πως οι δανειολήπτες οι οποίοι φαίνονται να έχουν χαμηλό DTI και πάνω από 2 χρόνια σταθερής σχέσης απασχόλησης μπορεί να ερμηνευθεί ως βιωσιμότητα από πλευράς του οργανισμού. Να μπορούν δηλαδή να αποπληρώσουν το δάνειο μαζί με τους τόκους σε μια υγιή σχέση σε βάθος χρόνου.

Ενδιαφέρον φέρει η σύσταση του Credit Score καθώς και του FICO. Είναι έννοιες οι οποίες σίγουρα συμπεριλαμβάνουν τα μεγέθη που μελετήθηκαν στην εργασία. Παρόλα αυτά διότι δεν υπάρχουν διαθέσιμα τα βάρη του καθενός σε αυτούς τους δείκτες για τα απορριφθέντα δάνεια, μελλοντική μελέτη στην περίπτωση που υπάρχουν δεδομένα μπορεί να μας δώσει μια καλύτερη εικόνα σχετικά με την δανειοδότηση. Συγκεκριμένα μια εναρκτήρια οδός θα ήταν να μελετηθούν τα δεδομένα μόνο των εγκεκριμένων δανείων ώστε να το ποσοστό αθέτησης και να τρέξει η Binary Logistic, συμπεριλαμβανομένου της μεταβλητής Risk_Score & FICO score αντίστοιχα ή να τρέξει μεταξύ των εγκεκριμένων και απορριφθέντων, συμπεριλαμβανομένου όμως του FICO score.

Τέλος η πανδημία του COVID-19 αμβλύνει σίγουρα το πρόβλημα αθέτησης των δανειοληπτών, σε ότι αφορά αποπληρωμές δανείων και απάτη πιστωτικών καρτών, καθώς προβλήματα ρευστότητας και προβλήματα υγείας μπορεί να έχουν φέρει τους οργανισμούς σε μία αμυντική στάση σε σχέση με τις δανειοδοτήσεις, όπως έγινε και άλλωστε με το Lending Club το 2020, και είναι άξιο μελέτης να δούμε παρόμοια έρευνα σε δεδομένα από το Q4 2019 έως και το Q4 2021.

6. ΒΙΒΛΙΟΓΡΑΦΙΑ

Aebi V, S. G. (2012). Risk management, corporate governance, and bank performance in the financial crisis. *Journal of Banking and Finanace*, σσ. 3213-3226.

Akerlof, G. (1970, Αύγουστος). The Market of Lemons. *The Quarterly Journal of Economics* , σσ. 488-00.

Aylin Hejazi, N. A. (2017, Απρίλιος). Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area. *International Journal of Trade, Economics and Finance*.

Bachman, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., . . . Funk, B. (2011). Online Peer-to-Peer Lending-A Literature. *Journal of Internet Banking and Commerce, 16*(1).

Board of Governors of the Federal Reserve System. (2021). *federalreserve.gov*. Ανάκτηση Οκτώβριος 31, 2021, από https://www.federalreserve.gov/supervisionreg/reglisting.htm

Capanella, F. (2017). The Effects of Technological Innovation on the Banking Sector. *Journal of the Knowledge Economy, 8*(10.1007/s13132-015-0326-8).

Choudhry, M., & Darell, D. (2011). *Structured Credit Products: Credit Derivatives and Synthetic Securitisation* (2nd εκδ.). Wiley Finance.

Fassbender, A. (2021). *kaggle.com*. Ανάκτηση 7 28, 2021, από https://www.kaggle.com/alwinfassbender/predicting-the-default-rate-for-peer-to-peer-loans

George, N. (2019, Απρίλιος 10). *Kaggle Inc.* Ανάκτηση Απρίλιος 6, 2021, από https://www.kaggle.com/wordsforthewise/lending-club

Gonzalez, L., Gonzalez, L., & Komarova, Y. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance, 2*(C), 44-58.

Herzenstein, M., Andrews, R., Dholakia, U., & Lyandres, E. (2008). *The Democratization of Personal Consumer Loans? Derminants of Success in Online Peer-to-Peer Loan Auctions.* Universtity of Delaware; Rice University.

Jet Brains Inc. (2021). *JetBrains*. Ανάκτηση 9 12, 2021, από https://www.jetbrains.com/pycharm/download/#section=windows

Kagan, J. (2020). *Investopedia*. Ανάκτηση Οκτώβριος 10, 2021, από https://www.investopedia.com/terms/p/peer-to-peer-lending.asp

Nobanee, H. (2021). Big Data Applications the Banking Sector: A Bibliometric Analysis Approach. *Sage Open, 11*(4).

O' Reilly, T. (2005). *O Reilly Radar*. Ανάκτηση 10 28, 2021, από https://d1wqtxts1xzle7.cloudfront.net/38552727/OReilly_Radar_-_Web_2.0_Compact_Definition-with-cover-page-v2.pdf?Expires=1635420918&Signature=CBZyIFfCNzyYWNkphbO3dzcv5z6wOp7piprg~SdBP3GeelIlFqbLRrmrrIYJz~SXBl9N38j7HlsSJf5CAyiTs0nVGXXp9cQG3Mvq07VVhTk9rrJ-0

Peterson, J. (2020). *petersoninquires.medium*. Ανάκτηση 4 16, 2021, από https://petersoninquiries.medium.com/loan-analysis-using-python-and-lending-club-data-5475e1a9844

Pope, D., & Syndor, J. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives, 24*(2), 95-108.

Python Software Foundation. (2021). *python.org*. Ανάκτηση 6 12, 2021, από https://docs.python.org/3/faq/general.html#what-is-python

Ratadas, A. (2020). *Jovian*. Ανάκτηση 8 28, 2021, από https://jovian.ai/anubratadas/all-lending-club-loan

Reinhart, K., Levich, R., & Majoni, G. (2002). Ratings, rating agencies and the Global Financial System. Munich: Kluwer Academic Publishers, Boston.

Tsakiridis, T. (2020). *Web 2.0 Ορισμός και Δυνατότητες*. Ανάκτηση April 23, 2022, από https://thetsakiridis.sites.sch.gr/2020/web-2-0-%CE%BF%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82-%CE%BA%CE%B1%CE%B9-%CE%B4%CF%85%CE%BD%CE%B1%CF%84%CF%8C%CF%84%CE%B7%CF%84%CE%B5%CF%82-%CE%B1%CE%BE%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%AE/

Ευρωπαϊκή Κεντρική Τράπεζα. (2017). *bankingsupervision.europa*. Ανάκτηση Οκτώβριος 31, 2021, από https://www.bankingsupervision.europa.eu/ecb/pub/pdf/guidance_on_npl.el.pdf