UNIVERISTY OF PIRAEUS - DEPARTMENT OF INFORMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSC «ADVANCED COMPUTING AND INFORMATICS SYSTEMS»

ΠΜΣ «ΠΡΟΗΓΜΕΝΑ ΣΥΣΤΗΜΑΤΑ ΠΛΗΡΟΦΟΡΙΚΗΣ»

MSc Thesis

Μεταπτυχιακή Διατριβή

Thesis Title: Τίτλος Διατριβής:	**Evolving Measures of Credit Risk** Εξελισσόμενα Μέτρα Πιστωτικού Κινδύνου
Student's name-surname: Ονοματεπώνυμο φοιτητή:	**Michail Papasymeon** Μιχαήλ Παπασυμεών
Father's name: Πατρώνυμο:	**Eleftherios** Ελευθέριος
Student's ID No: Αριθμός Μητρώου:	**ΜΠΣΠ/16024**
Supervisor: Επιβλέπων:	**Dionisios Sotiropoulos, Assistant Professor** Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής

December 2021/Δεκέμβριος 2021

3-Member Examination Committee

Τριμελής Εξεταστική Επιτροπή

Dionisios Sotiropoulos **Assistant Professor**	**Efthimios Alepis** **Associate Professor**	**Evangelos Sakkopoulos** **Assistant Professor**
Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής	Ευθύμιος Αλέπης Αναπληρωτής Καθηγητής	Ευάγγελος Σακκόπουλος Επίκουρος Καθηγητής

Περίληψη

Το πιστωτικό σκορ αποτελεί έναν καθοριστικό παράγοντα για την διαχείριση του οικονομικού ρίσκου δίνοντας τη δυνατότητα στα χρηματοπιστωτικά ιδρύματα να ποσοτικοποιήσουν την πιθανότητα της υγιούς εξόφλησης ενός δανείου. Ωστόσο, τα σύγχρονα και αναγνωρισμένα από την βιομηχανία μοντέλα μέτρησης πιστοληπτικού ρίσκου που παρέχονται από ιδιωτικές εταιρίες μεταξύ άλλων όπως η FICO και η Vantage δεν είναι διαθέσιμα στο κοινό.

Η δυσκολία του προβλήματος αυξάνεται δραστικά λαμβάνοντας υπ' όψη την έλλειψη γνώσης του αρχικού μοντέλου αλλά και των μεταβλητών που αυτό χρησιμοποιεί για να παράξει το πιστοληπτικό σκορ ενός φυσικού προσώπου. Το προτεινόμενο μοντέλο θα παραχθεί αξιοποιώντας ένα υποσύνολο του πλήθους των αρχικών μεταβλητών που έχουν συγκεντρωθεί από τις χιλιάδες αιτήσεις δανεισμού που έχουν πραγματοποιηθεί από φυσικά πρόσωπα στην πλατφόρμα. Το υποσύνολο αυτό αποτελείται από εύκολα προσβάσιμες μεταβλητές δίνοντας τη δυνατότητα στο τελικό μοντέλο να μπορεί να υπολογίσει το πιστοληπτικό ρίσκο ενός φυσικού προσώπου χωρίς την ανάγκη των δύσκολα προσβάσιμων ιστορικών πιστωτικών δεδομένων.

Στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός εναλλακτικού μοντέλου ανάλυσης και μέτρησης του πιστοληπτικού κινδύνου που προσεγγίζει την συμπεριφορά του μοντέλου που έχει αναπτύξει η FICO χρησιμοποιώντας μια μεγάλη συλλογή δεδομένων που σχετίζονται με τον δανεισμό από την πλατφόρμα "Lending Club". Πιο συγκεκριμένα, σκοπός μας είναι να παρουσιάσουμε τη μέτρηση του πιστωτικού ρίσκου με ένα κατανοητό από τον άνθρωπο και παραμετροποιήσιμο σε πολυπλοκότητα μαθηματικό μοντέλο. Για τον λόγο αυτό θα χρησιμοποιήσουμε ένα μοντέλο συμβολικής παλινδρόμησης το οποίο λειτουργεί στο πλαίσιο του γενετικού προγραμματισμού.

Για να πιστοποιήσουμε την εγκυρότητα της προσέγγισης μας συγκρίνουμε τα αποτελέσματα του γενετικού προγραμματισμού με άλλα "state-of-the-art" μοντέλα μηχανικής μάθησης όπως Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ), Πολλαπλών Στρώσεων Perceptron (MLP) και Radial Basis Function Networks (RBFNs). Χρησιμοποιώντας ένα δραστικά μικρότερο σύνολο μεταβλητών αποδεικνύεται ότι τα αποτελέσματα του γενετικού προγραμματισμού είναι συγκρίσιμα με τις παραπάνω αναγνωρισμένες μεθόδους. Ταυτόχρονα, το παραχθέν μοντέλο παρουσιάζει πολύτιμες πληροφορίες για τους μηχανισμούς οι οποίοι συντελούν στον υπολογισμό του FICO σκορ.

Abstract

Credit scoring constitutes a quintessential element of economic risk management allowing financial agencies to quantify the probability of default for a future loan. However, acclaimed contemporary credit risk measures such as the scores provided by FICO or Vantage are not publicly accessible.

The severity of the underlying problem is manifested by the limited amount of knowledge which can be obtained for both the exact analytical formula and the complete set of credit-specific features that underpin the computation of FICO score. The proposed measure will be derived by exploiting a limited amount of entry-level information submitted by each candidate borrower without requiring the accumulation of historical credit data for each consumer over large periods of time.

This thesis addresses the problem of developing an alternative credit scoring measure that approximates the behavior of the original FICO score in a large-scale collection of loan-related data available from Lending Club. We are particularly interested in expressing the acquired credit risk measure in a closed analytical form of adjustable complexity. For this purpose, we utilize a symbolic regression technique which operates within the framework of Genetic Programming (GP). In this context, we harness the notion of Occam's razor to apply evolutionary pressure towards the preservation of models associated with reduced complexity and higher degree of human interpretability.

In order to verify the validity of our approach we compare the approximation ability of the GP-based symbolic regression against state-of-the-art machine learning-based regression methods such as Support Vector Machines (SVMs), Multi-Layer Perceptrons (MLPs) and Radial Basis Function Networks (RBFNs). Our experimentation demonstrates that GP-based symbolic regression achieves comparable accuracy with respect to the aforementioned benchmark techniques. At the same time, the acquired analytical model can provide valuable insights concerning the credit risk assessment mechanisms that underlie the computation of FICO based on a significantly reduced set of credit-related features.

Table of Contents

1. Introduction

Credit risk or credit default risk is a term used to describe the chance that a company or an individual will not be able to pay their debt obligations. In simple terms, credit risk is the probability that a given loan will not be paid back on time. Virtually all forms of credit extensions yield credit risks that the lenders and investors are exposed to. Often, to reduce the impact of default risk, lenders impose additional charges that are proportional to the debtor's level of default risk. The higher the level of risk a debtor has, the higher the required return is asked by the lender. Credit risk is expressed by a number based on an analysis of an individual's credit files called credit score. That score represents how creditworthy an individual is [4][5].

1.1 WHY IS IT SIGNIFICANT TO BE ABLE TO CALCULATE CREDIT SCORE

Credit score is an important factor when it comes to issuing credit because it gives lenders an objective and fast measurement of an individual's creditworthiness [1]. Before credit scoring was used, the process of granting a credit request could be inconsistent, slow, and biased. Using the objective measurement that credit scoring provides, lenders can rely only on the facts that relate directly to credit risk rather than their personal opinion or feeling. Factors that discriminate against individuals such as gender, nationality, religion, race, marital status, and level of education are not taken under consideration by credit scoring [2][3][5].

Credit score helps speed up the decision-making process of a lender. Because a credit score can be issued almost instantaneously in many circumstances the decision can be taken within minutes. This affects decisions that concern mortgages as well, speeding up the process to days instead of what used to be weeks. Scoring also gives retail stores or Internet sites the ability to give "instant credit" to customers accelerating their overall business. These are a considerable advantage especially if an individual has a good credit score since it's more likely to get quick approval and move forward with your plans or purchases [5].

Lenders are confident to issue more loans because of the information given on credit risk by the individual's credit score. Credit scores allow individuals of all levels to access credit and lenders to issue the credit product that aligns with the risk level of the individual instead of automatically rejecting or approving an applicant. Moreover, a credit score allows for lenders to identify individuals with potential, meaning that the score helps the lender understand whether an individual is likely to perform well in the future even if they had problems with credit in the past. Since lending policies vary, another lender might approve the loan application if the first turned it down [5].

1.2 FACTORS THAT AFFECT CREDIT SCORE

In general, the factors that are considered when calculating a credit score are the following:
the individual's payment history, the number of open accounts an individual has, the types of those accounts, the credit the individual has used versus the available credit he/she has,
the length of the individual's credit history (how long has he/she been able to manage credit). The weight of each factor varies depending on the scoring model used to produce the credit score [6][7].

1.3 CREDIT RISK FACTORS BREAKDOWN

Payment history: This factor is heavily influencing the outcome of the score calculation. The question this factor is answering is “if a lender extends the applicant’s credit, will the applicant be able to pay the amount back on time”. Payment history may include installment loans, auto loans, student loans, finance company accounts, home equity loans, mortgage loans, retail department store accounts, and credit cards. Payment history also shows details on missed or late payments and collection information. What a credit score model is generally looking at is how late your payments were, how much did the individual owe, and how recently and how often the individual did miss a payment. Additionally, the credit score model will look at the ratio of delinquent accounts in relation to all of the individual’s accounts in the file. So if an individual had late payments on 5 of 10 of his/her open accounts, that ratio may impact the individual’s credit score. Payment history also includes details on foreclosures, bankruptcies, wage attachments, and the accounts that have been reported to collection agencies. In general, a credit scoring model will consider all of the above, which is why this factor may have a substantial impact on determining some credit scores [6][7].

Credit used vs available credit: This is another factor that creditors and lenders are looking at. It describes how much of your available credit or “credit limit” you are utilizing. What lenders want to see is that an individual is not only able to use credit but pay it off regularly as well. If for example some of the individual’s credit cards are at their limit or “maxed out” this could impact their credit score [6][7].

Type of credit used: Credit scoring models also consider the variety of credit an individual has. This includes revolving debt (such as credit cards) and installment loans (such as mortgages, personal loans, auto loans, student loans, and home equity loans). The count of these accounts is also a factor taken under consideration since the lender or creditor wants to see if the individual can manage a variety of credit types and not just one [6][7].

Length of credit history: This factor details how long an individual’s credit accounts have been active. When calculating a credit score both the age of your oldest and most recent accounts may be taken under consideration. In general, lenders want to see that an individual has a history of responsibly paying their credit accounts. Credit scoring models may also take into consideration the number of credit accounts an individual has opened recently as new accounts might impact the length of the individual’s credit history [6][7].

Hard inquiries: When a lender or creditor checks an individual’s credit report in response to an application for credit a “Hard Inquiry” occurs. Checks made by the individual itself or by a lender or creditor for an already approved loan are considered as “Soft Inquiries” and do not impact score. Having too many hard inquiries may indicate risk and thus hurt an individual’s credit score [6][7].

An individual does not have only one score. Credit scores vary depending on the company that assesses the credit risk, the data each company uses to produce the score, and the method that calculates the score itself. Moreover, the above factors are mainly used in countries like the USA where credit scores for individuals are widely used in everyday life. In other countries or continents such as Europe, credit scores till now are not very popular and are mainly calculated within various financial institutions such as banks on the occurrence of an individual’s application for credit [6][7].

2. Credit Risk assessment models

Several credit score models are being used with their unique characteristics.The FICO Score, the model which is now the industry's most widely accepted credit scoring model was introduced by the Fair Isaac Corporation. The scale of the FICO score is between 300 and 850 points [7].

Vendors like Experian, Equifax, and TransUnion are the ones responsible for selling these scores to their customers as FICO scores are not directly provided to individuals. The vendors keep files and the credit history of their clients and use this information at a given time point to calculate the score.

The PLUS Score is another credit scoring model that was developed by Experian. The PLUS Score ranges between 330 and 830. This model was developed to help customers understand how creditors view their creditworthiness. The higher the number the likelier that the customer will repay their debt promptly thus imposing a lower risk to lenders. As time goes by and the individual's information changes, the score might also change.

The Vantage score was developed by Experian, Equifax, and TransUnion. It's a relatively new model to provide an accurate and consistent approach to credit scoring. With this score, lenders are provided with a nearly identical risk assessment across all three vendors. The Vantage scale ranges from 501 to 990 [6].

According to the Fair Isaac Corporation (FICO) model analysis, most of the population has a credit score between 600 and 800. A score equal or higher to 720 will enable an individual to get a mortgage with the most favorable interest rates. Scores lower than 499 belong to the 4.7% of the total population whereas, scores ranging between 500 and 549 belong to the 6.8%, scores between 550 and 599 belong to the 8.5%, scores between 600 and 649 belong to the 10%, scores between 650 and 699 belong to the 13.2% and, scores between 700 and 749 belong to the 17.1%. The rest of the population has what's called an "Excellent score" which for the 19% and is ranging from 750 to 799 and for the remaining 20.7% above 800 as of 2017. In April of 2017 FICO published a metric that showed the average score being 700 compared to October of 2005 where the average score was at 688 [8][8.5].

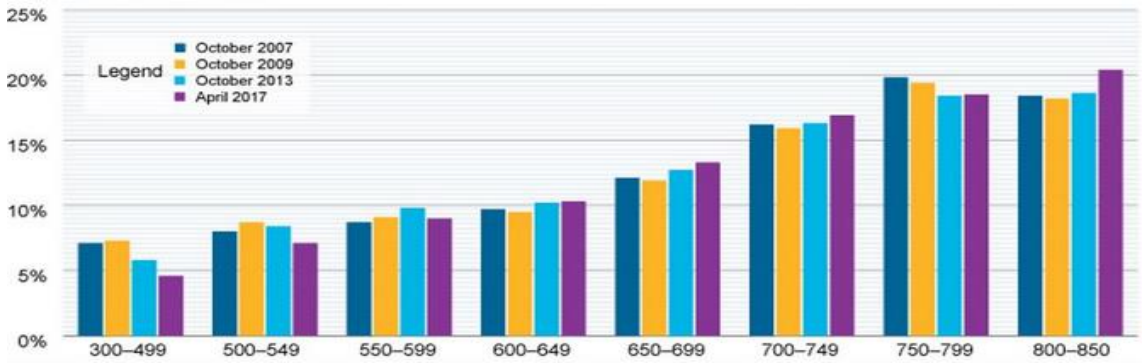


Figure 1 FICO Score Distribution

FICO® Score 8	October 2005	October 2006	October 2007	October 2008	April 2009	April 2010	April 2011	April 2012	April 2013	April 2014	April 2015	April 2016	April 2017
300–499	6.6	6.5	7.1	7.2	7.3	6.9	6.3	5.7	5.6	5.4	4.9	4.6	4.7
500–549	8.0	8.0	8.0	8.2	8.7	9.0	8.7	8.5	8.4	8.1	7.6	7.1	6.8
550–599	9.0	8.8	8.7	8.7	9.1	9.6	9.9	10.0	9.9	9.6	9.4	9.0	8.5
600–649	10.2	10.2	9.7	9.6	9.5	9.5	9.8	10.1	10.1	10.2	10.3	10.3	10
650–699	12.8	12.5	12.1	12.0	12.0	11.9	12.1	12.2	12.2	12.8	13.0	13.3	13.2
700–749	16.4	16.3	16.2	16.0	15.9	15.7	15.5	16.0	16.3	16.4	16.6	16.9	17.1
750–799	20.1	19.8	19.8	19.6	19.3	19.5	19.6	19.0	18.9	18.2	18.2	18.5	19
800–850	16.9	17.9	18.4	18.7	18.2	17.9	18.1	18.5	18.5	19.3	19.9	20.4	20.7
TOTAL*	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Figure 2 FICO Score per population percentage

The exact statistical models that are used on the credit report of an individual to determine their FICO score are kept confidential by the credit scoring agencies. However, as previously shown in this document the five factors that are considered for developing the FICO score are payment history, credit used vs available credit, types of credit used, length of credit history and, the number of the individual's recent applications for a loan (hard inquiries) [2].

2.1 EXISTING METHODS OF ASSESSING RISK

2.1.1 Traditional way of assessing risk

One of the methods lenders and creditors use to assess credit risk is the empirical one. Banks and closed institutions have their credit policies which in the end affect their client portfolio. Many banks, to this day, use empirical models utilizing their staff experience on the subject and create scoring methods that use specific econometric variables along with the data of the loan application. Banks and closed institutions have access to exclusive information about previous loan applications and the individuals who did them as part of their client portfolio so many times the empirical model can be mixed with some statistical data resulting in a hybrid model. These methods are still used inside the bank's ecosystem because they are very specific to that bank's previous data thus creating the bank's face to the public when it comes to lending. Also, these models are highly configurable when a bank is trying to open up to different markets and attract more clients. In the late years, since statistical models combined with data science have skyrocketed, many banks have started using less and less of the empirical or hybrid models and have started relying on what technology has to offer. Many times, of course a bank or institution could make use of both when trying to assess the creditworthiness of an individual.

Nonetheless, the traditional credit risk assessment procedures used by creditors and lenders when evaluating a loan application suffer from being inherently subjective. Moreover, they are heavily dependent on the expert's experience as well as impacted by the fact that the nature of considering the evidence is sequential rather than parallel. Additionally, they are considered to be more costly and time-consuming. Lenders and creditors are always after reducing the delinquency rates and having greater control over credit policies which led them to start experimenting with credit scoring [9].

2.1.2 Credit beaureaus and technology

Credit bureaus on the other hand can use more technology-driven or data-driven techniques since the only job they have is to find out how creditworthy an individual is based on their credit report data. The reason for this is that the data these bureaus rely on are not bank-specific but individual-specific meaning that an individual could be late on their payments on bank A but not late on bank B so these two banks have a different perception of this individual. The main difference here is that even if a bank can see the same data a credit bureau sees for an individual by doing a hard inquiry on the occurrence of a loan application, they simply do not have the same coverage, as they can't just do hard inquiries in everyone, but only those who directly apply for a loan. Contrary to the bank's scope of visible data, credit bureaus, especially the standard ones (for example in the USA the 2 major players in credit scoring are FICO and Vanguard), have a much wider perspective of the population (around 80% of the population of the USA have a credit score). The main tool for scoring these individuals is by the use of Scorecards. Scorecards are something close to a point system that correlates the different values of the credit-oriented variables to points, resulting in the individual's score or creditworthiness. The main target of a scorecard is to distinguish individuals between good and bad debtors. To produce the scorecards the credit bureaus, begin by applying techniques such as Weight of Evidence, Information value (WOE & IV), and then fitting these with a regression model. The result is a look-up table that maps specific attributes of a borrower to specific points [10].

2.1.3 Data driven techniques

These are the two main ways to assess credit risk but other banks that have a lot of data with a wider variety and depth, compared to the credit report data that bureaus have access to, have started using data science and machine learning models to utilize them. These data are often called "Alternative data" because of the differences they have compared to the traditional data used for credit risk assessment. Data like this can be account information and transaction information which results in the creation of spending profiles and the extraction of credit-oriented behavioral patterns. Recently with the introduction of the Payment Services Directive (PSD2) in Europe, even third-party applications can have access to this data if an individual approves of it. The utilization of alternative data can make scores more accurate, diverse, and robust. Although having new data, this doesn't necessarily mean that there are many more new technologies that can have a good result. Scorecards and credit risk assessment must be stable, predictable, and interpretable so not many of the complex machine learning algorithms can yield these results. Still, even with the introduction of the alternative data, logistic regression is a very good solution along with other techniques such as decision trees [10][11].

2.1.4 More diverse and alternative techniques

Lastly, there are also even more diverse techniques for assessing credit risk. Till now we have only discussed data that have to do with how an individual handles and manages credit in numbers. Another important factor when assessing credit risk is the individual's character. It's a common misconception that a wealthy individual will also be a better borrower, but that is not true. An individual having a great amount of wealth doesn't necessarily mean that they are up to pay on time and in full. The exact opposite is true for a not wealthy individual who could be very careful about credit and always pay on time. This is why "Character" is one of the five C's of credit, the other four are Capacity, Condition, Capital, Collateral. To determine the character of an individual many players have started utilizing personality tests and psychometric tests that try to understand how an individual would go on about handling and managing credit. Studies have shown that

tests like these can have an accurate prediction and lately Equifax in the USA has started using them to try and bring in people who haven't got a credit history yet [12].

In this thesis what we are trying to achieve is to try and create an explainable, transparent, and easily accessible way to assess the risk an applicant imposes. To achieve that we will use genetic programming to reverse engineer and describe the underlying mathematical association of the applicant's characteristics compared to their score. The result of the genetic programming model will be an equation that can be then used to calculate the applicant's credit score. Since we want our model to be accessible and easy to use, we are going to focus on credit-related features which can be easily accessed by the industry so this can be used in something as simple as a questionnaire. In addition to being easy to integrate, the resulting equation is also transparent to any form of audit or legislation.

3. Transparency

Before discussing the topic of the techniques creditors and lenders use there is a very important limitation to point out. Credit is a tool with a significant impact on people's lives and can be used to help individuals achieve more in many aspects of life thus it is also a very sensitive one. In most countries around the world creditors and lenders are subject to heavy auditing by higher authorities. Since this domain is such a sensitive and audit heavy one, models that take decisions for people's lives without being transparent, interpretable, and stable are out of scope and cannot stand as a valid solution. At this point, we are not yet able to completely understand how deep machine learning networks work in terms of interpretation and transparency, we do know how to design a neural network but when it comes to explaining why a neural network took a decision, for the most part, we are not able to know [15] so most of the credit risk prediction methods that use neural networks are often considered as supporting models. Also, even at state-of-the-art machine learning models, we can have a small number of outliers which again is unacceptable, imagine a scenario where an individual gets a score that is out of bounds with no solid explanation, rendering them unable to extend their credit. It's nearly impossible to come across a technique that is a "black-box" or a non-interpretable one when it comes to calculating a credit score [15]. Someone must be able to explain why and how this score was calculated at all times. In the newer age of credit, with peer-to-peer lending platforms as well as other fintech companies though there are some exceptions. Fintech companies like these have started relying upon alternative data along with the traditional data and a lot of times they achieve higher accuracy when it comes to measuring creditworthiness.

Regardless of the transparency limitation there are a lot of data-driven techniques that are fairly popular, some of them are Logistic Regression, Neural Networks, Genetic Programming, Gradient Boosting, etc. The main target of the credit scoring model is to predict the probability that an individual will or will not default and regression models are very good in predicting such outcomes.

4 Machine learning models

4.1 LOGISTIC REGRESSION

4.1.1 Logistic Regression Introduction

Logistic regression is one of the very important models when it comes to categorical response data. It is often used when the dependent variable is binary (default, not default). This is a predictive analysis like all regression analyses, it describes data and explains the relationship between one or more ordinal, nominal, ratio-level, or interval independent variables and one dichotomous dependent variable [17]. Logistic regression is used in many applications such as social science research, financial application, biomedical studies as well as marketing. In credit scoring logistic regression is used to model the probability that an individual is creditworthy (i.e. able to meet their financial obligations in time) using several predictors. The predictors can be personal information such as the customer's annual income, other outstanding debts, occupation, credit history, past behavior, and the size of the loan [18]. Logistic regression is considered by many to be the industry standard for credit risk assessment [15].

The data sets on which logistic regression is applied consist of rows of observations (loan applicants). Each observation consists of the same number of covariates (predictors) and the response value which as previously stated is a binary value 0 to for applicants who are "bad" debtors and 1 for applicants who are "good" debtors. Typically, to find out which features should be used as inputs to the logistic regression model we have to compare the predictive power that each feature has. Amongst the various techniques of feature selection one of the most popular in the problem space of credit scoring is "Weight of Evidence & Information Value". These two concepts have evolved from the same logistic regression technique and have been thoroughly tested and used in various credit scoring models.

4.2 WEIGHT OF EVIDENCE AND INFORMATION VALUE

4.2.1 Weight of Evidence

In practice to calculate the Weight of Evidence we start by splitting the data in 10 (or less depending on the distribution) parts or bins. Binning is a categorization process by which we can transform a continuous variable into a set of groups or bins [26]. In credit scoring binning is used at an early stage for feature selection. Values of a feature with high similarity are grouped together to enhance the predictive power.

The most used binning algorithms are: equal-width binning, equal-size binning, optima-binning and multi-interval discretization binning. Equal-width: values of the independent variable are divided into a pre-defined number of intervals with equal width. Equal-size: the attributes are first sorted and then divided into a predefined number of equal-size bins. In case the feature has distinct values all the bins except the last will have the same number of observations. The last bin might have a smaller number of observations. Optimal: the independent variable is divided into a large number of equal width bins (e.g. 50). Afterwards these bins are treated as categories of a nominal variable. These categories are then grouped to the required number of segments in a

tree structure. Multi-interval discretization: is the entropy minimization for binary discretizing the range of a continuous variable into multiple intervals and then recursively defining the best bins.

- The process of "binning" the data has some specific rules [26][27].
- Each bin should contain at least 5% of the observations.
- Each bin should be non-zero for both events and non-events.
- If a WOE value is similar for a number of bins these bins should be aggregated.
- The WOE within the groupings should be monotonic (either increasing or decreasing).
- In case the data set has missing values, they should be binned separately.

A way to validate that the binning algorithm is correct we can run logistic regression with 1 independent variable to a model that has undergone a WOE transformation. If the slope is not 1 or the $Intercept \neq \ln\left(\frac{\%\, non-events}{\%\, events}\right)$ the algorithm is not good.

After binning is done, we can introduce Weight of Evidence (WOE) transformation for continuous variables. Weight of Evidence (WOE) helps us understand the predictive power of an independent variable (an applicant predictor) in relation to the dependent variable [27].

We get WOE from the following calculation:

$$WOE = \ln\left(\frac{\%\ of\ Good\ applicants}{\%\ of\ Bad\ applicants}\right)$$

Or in a more generic format:

$$WOE = \ln\left(\frac{\%\ of\ non_{events}}{\%\ of\ events}\right)\ or\ WOE_{x=i} = \ln\left(\frac{\%\ of\ y = 0\ where\ x = i}{\%\ of\ y = 1\ where\ x = i}\right)$$

In general, a positive WOE value means that: Distribution of Goods > Distribution of Bads and a negative WOE value Distribution of Goods < Distribution of Bads. Based on the -/+ sign of WOE we are able to know the proportion of "good" applicants vs "bad" applicants in the dataset.

The result of the WOE calculation can then be used to determine the information value (IV) of each of the independent variables with the dependent variable. The independent variables with the highest information value will be selected as the final data set upon which we will fit the logistic regression model.

4.2.2 Information Value

Information value is a very useful concept for variable selection. A variable's information value is a way to measure how well it is able to distinguish between a binary response ("good" vs "bad") in some target variable Y (dependent variable). This means that a variable with low information value will not do a sufficient job of classifying the target variable, thus, these variables are removed from the data set leaving us with the variables that have a higher information value.

Information Value (IV) is calculated using the following formula:

$$IV = \sum (\%\ of\ non_{events} - \%\ of\ events) \times WOE$$

Information value result interpretation:

- Less than 0.02: the variable is not useful for modelling. It does not do a sufficient job of separating the "good" from the "bad".
- 0.02 to 0.1: the variable has a weak relationship with the ratio of the "good"/"bad" odds.
- 0.1 to 0.3: the variable has a medium strength relationship with the ratio of the "good"/"bad" odds.
- 0.3 to 0.5: the variable has a strong relationship with the ratio of the "good"/"bad" odds.
- More than 0.5: the variable has a suspiciously strong relationship with the ratio of the "good"/"bad" odds.

Information Value	Variable Predictiveness
IV < 0.02	Not useful for prediction
0.02 > IV 0.1	Weak predictive power
0.1 > IV 0.3	Medium predictive power
0.3 > IV 0.5	Strong predictive power
IV>0.5	Suspiciously strong predictive power

Table 1 Variable predictiveness table

Information value provides a basis for us to drill further down in the relationship analysis between the independent and dependent variables.

As previously mentioned, having calculated the Information Value (IV) of the independent variables we are now in position to reject some of them and end up with a data set with a stronger predictive power. Moreover, the process of binning and calculation of the Weight of Evidence (WoE) values is used for encoding. A value, which belongs to a variable that was found to be a strong predictor, is now changed to the corresponding WoE value of the bin.
As an example suppose we have 5 bins for a continuous variable x 0,100which has been grouped into 5 bins of equal-width and we have calculated the WoE for each bin:

Bins	Weight of Evidence (WoE)
bin1 [1..20]	0.9825
bin2 (20..40]	-1.9645
bin3 (40..60]	0.3835
bin4 (60..80]	0.1256
bin5 (80..100]	0.2425

Table 2 Weight of Evidence per Bin

A value $x_i = y_i, y \in bin_i$ will now have the value of the Weight of Evidence of the bin $x_i = 0.9825$. The same will happen with a categorical variable with Ndifferent classes, instead of performing a one-hot encoding and ending up with Ncolumns with mostly 0 as values we can use WoE technique as means of encoding and replace the classes by their associated WoE values. Encoding values is a very common technique used as a pre-step before training most of the models and this is simply because models and machine learning algorithms primarily take numbers as inputs.

4.2.3 Logistic regression

Generally, logistic regression is one of the most frequently used models when assessing the potential risk of a decision and specifically for credit the probability of default. Using logistic regression, we can find whether based on the values of the attributes (independent variables) the dependent variable takes a value of 0 (default) [30].

The two probabilities, default Y=0 and non-default Y=1 are expressed by equation (1) and (2):

$$P(Y = 0|X) = P = \frac{e^{\beta' X}}{1 + e^{\beta' X}} \ (1)$$

$$P(Y = 1|X) = 1 - P = \frac{1}{1 + e^{\beta' X}} \ (2)$$

Where β' is the vector of coefficients $\beta' = (\beta_0, \beta_1, \dots , \ \beta_n)$ and x is the vector (column) of the attributes (independent variables) $x' = (x_0, x_1, \dots, x_n)$

Equation (3) is equivalent to (1) and (2):

$$\ln\left(\frac{P}{1-P}\right) = \beta' X =: l \quad (3)$$

Where l denotes the logit function of the probability p. Equation (3) shows the linear relation that the attributes (independent variables) have with the logit function of the dependent variable. To estimate the coefficients, we use the maximum likelihood method (4):

$$P(Y = y_i) = P_i^{1-y_i}(1 - P_i)_i^{y_i} \quad (4)$$

Where P_i is the i^{th} observation's probability of default and y_i is the value of random variable $Y \in [0, 1]$. Assuming that our N observations are independent the likelihood of the data can be calculated using equation (5):

$$L = \Pi_{i=10}^{N} P_i^{1-y_i}(1 - P_i)^{y_i} \quad (5)$$

The objective of Maximum Likelihood Estimation is to find the set of beta-coefficients that maximize the likelihood function, e.g., result in the largest likelihood value. The logit variable l is the linear combination of the n independent variables.

$$l = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 \dots\dots + \beta_n * X_n$$

5. Neural Networks

Neural networks consist of layers of interconnected neurons. The simplest form of a neural network has three layers of neurons: input, hidden, and output. The hidden layer forms an internal symbol that represents "concepts". Increasing the complexity of the neural network by adding more layers can make it more robust by avoiding overfitting and increasing the generalization abilities of the network. Neurons process inputs and produce outputs. The connections between the neurons are unique lines of communication between a "sending" and a "receiving" neuron. These connections have an assigned "strength" or "weight" which determine the strength of the incoming signal. To determine how an input signal is an output by a neural network there is a Transfer Function with the most typical one being the Sigmoid. Unlike empirical systems, neural networks do not require "if-then" rules to be specified by the user, they only require specific examples of input values along with the corresponding output values. The network then can determine these "if-then" rules based on the data.

Layers of neurons that are interconnected result in a neural network. The network is capable of learning and storing associations through the connection strengths between layers. As connection strengths get modified, they establish new associations that emulate a rule-like behavior. If a neural network is not predicting the values correctly these weights have to be modified to achieve

higher accuracy. This is done by a supervised learning scheme called Backpropagation wherein feedback of local error signals through the network is used to adjust neurons' connection weights. The values presented to the neurons in the input layer of the network are the input half of the facts then the prediction of the neural network (output values) are compared against the values for the output half of the fact. If the predicted values and the original values match then there is no action taken, however, if they do not match the connection strengths in the network are modified to decrease the error. This process is done repeatedly for every fact in the training set over and over again until the neural network's error is very low or zero, meaning that it is capable of predicting the correct value for all the output halves of the facts presented in the input layer. Training a neural network thus is the process of repeatedly "feeding" related input-output sets so the backpropagation algorithm can adjust the connection strengths or weights for each of the neurons. The variety and size of the initial training set have a major impact on the pattern matching capacity developed by the network.

Once trained, a neural network's response can be, up to a certain degree, insensitive to small variations in the input data. This ability to extract the underlying patterns through noisy and distorted data is a vital component of pattern recognition in real-life problems [20].

6. Genetic Programming

6.1 INTRODUCTION

Genetic Programming (GP) is an optimization method in the field of Evolutionary Computation. This biologically inspired field includes various methods such as Evolution Strategies, Genetic Algorithms, and Evolutionary Programming. Genetic Programming is an evolutionary computation method that is domain-independent and can automatically find a solution in a predefined search space. To achieve that, GP is creating a population of computer programs that are transformed stochastically into a new population of computer programs. This is done by employing evolutionary mechanisms that are inspired by Darwin's theory of evolution. Mechanisms like these are inheritance, crossover, selection, and mutation. To depict a Genetic Programming computer program we often use a tree representation. These trees have arithmetic operators as internal nodes and variables or constants as leaves. In Genetic Programming, the former are called functions and the latter terminals [22]. As an example, to illustrate the program (var1 +var2) * 2, with var1, var2, and 2 as terminals and + and * as functions the tree syntax would be:

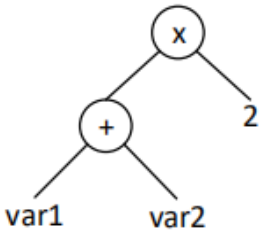


Figure 3 Example genetic programming gene

In credit scoring, these computer programs are discriminative functions that have a goal to assign the associated probability of default to an applicant. Thus, the problem is defined as a symbolic regression problem. Setting an appropriate threshold based on the output value, one can classify applicants into bad and good [21].

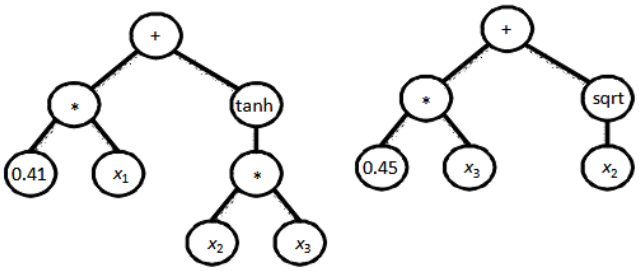
6.2 GPTIPS: GENETIC PROGRAMMING & SYMBOLIC REGRESSION

One of the main features of GPTIPS is that it can be configured to evolve multigene individuals. Each individual consists of one or more traditional GP trees (genes). To improve their fitness (increase prediction accuracy or minimize error), the individuals acquire these genes incrementally. GPTIPS employs a unique type of symbolic regression called multigene symbolic regression. Multigene symbolic regression evolves linear combinations of non-linear transformations of the input variables. By forcing the transformations to be low order (by restricting the GP tree depth), in contrast to standard symbolic regression, we allow the evolution of accurate and relatively compact mathematical models of input - output datasets even if the number of input variables is large.

6.2.1 Standard Symbolic Regression

The “standard” Symbolic regression is performed by using Genetic Programming to evolve a population of GP trees. Each of these GP trees encodes a mathematical equation is using a $(N \times M)$ matrix of inputs X where N is the number of observations of the response variable and Mis the number of input variables to predict a $(N \times 1)$ vector of outputs y.

Conversely, in Multigene symbolic regression, each symbolic model is a weighted linear combination of the outputs from a number of GP trees. Each of these GP trees may be considered to be a gene. Figure G1 shows an example of a multigene model which is using input variables x_1, x_2, x_3 to calculate the output variable y.



$$y = d_0 + d_1(0.41x_1 + \tanh(x_2x_3)) + d_2(0.45x_3 + \text{sqrt}(x_2))$$

Figure 4 Genetic Programming gene

This model structure is linear in the parameters with respect to the coefficients d_0, d_1, d_2 but it contains non-linear terms (e.g. tanh). To control the overall maximum complexity of the evolved models, the user defines the maximum number of genes G_{max}d the maximum tree depth D_{max}at a model and a gene respectively may have. For example, it has been found that changing the tree depth to 4 or 5 nodes often allows relatively compact models, that are linear combinations of low order non-linear transformations of the input variables, to evolve.

The linear coefficients of each model are estimated by using ordinary least squares techniques on the training data. Therefore, multigene GP has a hybrid approach as it can combine linear regression while being able to capture non-linear behavior without the need to specify the structure of the non-linear model beforehand.

6.2.2 GPTIPS

To construct the initial population, we create individuals that contain randomly generated GP trees with gene count between 1 and G_{max}uring a GPTIPS run, a tree crossover operator called two-point high level crossover is used to delete a GP tree or set it as a gene. This happens in GP as it happens to traditional genetic algorithms, and it allows individuals to exchange genes between them in addition to the GP recombination operators. The two-point high level crossover works as follows: two individuals, individual A I_A and individual B I_B both consist of genes with the ith gene being G_i. If we have set, $G_{max} = 5$ can have $I_A -> (G_1, G_5, G_3)$ and $I_B -> (G_4, G_6, G_2, G_7, G_8)$. To denote the genes chosen by the two-point high level crossover we can use $\{..\}$. For I_A the crossover choices are: $(G_1, [\![\{G\}]\!]_5, G_3\})$ and for $I_B$the crossover choices are: $(G_4, [\![\{G\}]\!]_6, G_2, G_7\}, G_8)$. By combining the choices of the crossover, we have these two new individuals: $I_A -> (G_1, G_6, G_2, G_7)$ and $I_B -> (G_4, G_5, G_3, G_8)$. This operation allows both individuals to acquire genes but also "lose" genes as well. For example, let's assume that the selection of genes from the crossover are as follows: $I_A -> (G_1, G_5, \{G_3\})$ and $I_B -> (G_4, [\![\{G\}]\!]_6, G_2, G_7, G_8\})$. This would result to the following individuals: $I_A -> (G_1, G_5, G_6, G_2, G_7, G_8)$ and $I_B -> (G_4, G_3)$. As previously mentioned, we have set $G_{max} = 5$t the count of genes that I_A has acquired, are more than the threshold. When the crossover operation results in gene count greater than the G_{max}eshold then the genes are randomly selected and deleted until the individual reaches a gene count that is equal to G_{max}ther than two-point high level crossover, in GPTIPS, there is also low-level crossover which is a standard GP subtree crossover. In low level crossover, a gene is randomly selected for each parent individual and after standard subtree crossover is performed, the resulting trees replace the parent trees in the next generation [25].

To select the individuals that proceed to the next generations based on their fitness, GPTIPS uses tournament selection. GPTIPS can also be configured to use lexicographic tournament selection but since we are not using it on our approach, we only mention it as a feature. The standard tournament selection starts by taking a random sample of k individuals from the population of size N. These random sampled individuals are then inserted into a tournament of size k and the one with the best fitness is selected. Typically, the tournament selection consists of two steps: sample and select and have sizes of 2, 4 and 7. Generally, the number of tournaments needed to generate all individuals for the next generation is N and this is due to the fact that the standard breeding process in GP produces two offspring by applying crossover to two parents and one offspring by applying mutation to one parent. Tournament selection has easily adjustable selection pressure, is efficient and simple to code and has a low complexity of $O(N)$ since it does not require the whole population to be sorted beforehand. Genetic Programming is very computationally intensive and requires a parallel architecture to improve its efficiency, moreover, since GPTIPS

can be applied to millions of individuals, selection methods that need sorting are nonfit. However, there are two main drawbacks to tournament selection. These drawbacks are: a) Multi-sampled, where an individual is sampled to more than one tournaments due to individuals being sampled with replacement, b) Not-Sampled, when using small tournament sizes, it's possible to have some individuals not sampled at all [31].

6.3 OTHER NOTABLE APPROACHES

Other approaches often researched by scientists include models such as Support Vector Machines, Random Forests, Gradient Boosting, and K-Nearest Neighbors. In the majority of the researched solutions, Logistic Regression is used as the benchmark as it is described as the industry-standard model for credit risk assessment. Financial institutions are reluctant to replace the Logistic Regression models with newer approaches due to regulatory reasons and potential model risks. Most of the techniques researched over the years cannot be recommended as standalone methods for credit risk assessment [15].

7. Dataset and the Lending Club platform

7.1 THE LENDING CLUB PLATFORM

Lending Club is a peer-to-peer lending platform in the USA which releases loan data every quarter. The data we used are from 2007 until the third quarter of 2019. The dataset consists of information on nearly all the loans Lending Club has issued in the aforementioned period. There's information on all the details of the loans at the time of issuance along with information about whether the loan was fully paid back or not as well as information about late payments by the debtor.

7.2 THE DATASET

The dataset is inherently biased on some of its features due to the platform's restrictions. Firstly, the platform does not accept applicants with a debt to income (dti) ratio higher than 40%. Meaning that if you have debts that cover more of the 40% of your income you cannot enter the platform and apply for a loan as an individual. The only way someone can go around this restriction is by adding another individual in his application whose dti is less than 40% so the combined dti is eligible for applying. Another bias is that the fico scores in the platform start from 660 which for Lending Club is considered as a "lending threshold" below which someone is not lendable.

The dataset starts with 150 columns and more than a million rows but ends up smaller after we filter out and only keep the values that we want to work with. Since we have a lot of data in our hands, we can filter specific ranges that result in a more consistent dataset. We filter out annual incomes of less than 10k and more than 700k as the observations outside this range are very few and inconsistent. Since we want to target individuals, we filter out joined applications (applications that were made to overcome the platform's 40% dti threshold) and only keep the individual ones. Also, we keep revolving utilization between 0% and 100% since more than 100% revolving

utilization only applies under certain conditions regarding the applicant's credit card management. To be sure that we are not introducing any unwanted noise in the model we filter out non-verified users. Lastly, we only keep loans with the status 'Fully Paid', 'Charged off', and 'Default' since loans with 'late X days' don't refer to any terminal information about the loan.

The dependent variable, the applicant's score, comes in a range format (score bins) ranging from 660 to 850 increasing in steps of 5, instead of a continuous format. This lowers the resolution of the model's target since the connection of the independent variables with the dependent variable is somewhat blurry and noisy.

7.2.1 Features and Feature preparation

The main features we are going to focus on are: Annual_inc: The annual income of the applicant, Dti: the ratio of overall debt versus annual income, Revol_bal: the total debt amount the applicant continues to owe after the end of each billing cycle and Revol_util: the percentage of the available credit which is currently utilized by the applicant. For the preparation of the features, we apply normalization to the [0, 1] range.

7.2.2 Filtering and data cohesion

A person's credit score is not a static number, it's a number that constantly changes based on their credit behavior. For example, if a person initially has a low score, based on a short series of credit events, as time progresses and if this person is always paying on time, their score will be much higher. On the other hand, if a person initially has been assessed as a low-risk applicant and thus received a high score could overtime start missing payments and building excessive debt, this change of behavior will result in a significant score drop as they now impose a greater risk towards lenders.

This adds a new layer of complexity when trying to develop a model using a static dataset because the applicant's features do not have a direct association with the score. For example, if paying always on time, someone with high credit utilization, which is a dominant indicator of high risk, can have a much higher score compared to someone with the same set of features but a worse credit behavior. The intensity of this phenomenon is of course variable, the higher the intensity the more of an outlier the applicant is. To tackle this problem of the overtime fluctuating risk which inherently fluctuates score as well, we perform a segmentation of the observations based on the previously mentioned distances per score range (bin).

7.2.3 Data segmentation

The segmentation process starts by calculating the euclidean distances of the credit-related features and then creating layers of diversified datasets where the approximation models are applied. Each layer corresponds to a distinct euclidean distance measurement from the centroid of the cluster. As previously mentioned this process is being performed for all the score bins. We can express the layering method using the following illustration (Fig) where C_k is the centroid of the cluster of the k_{th} score bin and $\Lambda = \{\lambda_1, \dots, \lambda_{\max}\}$ the layers of the distinct euclidean distances from the centroid.

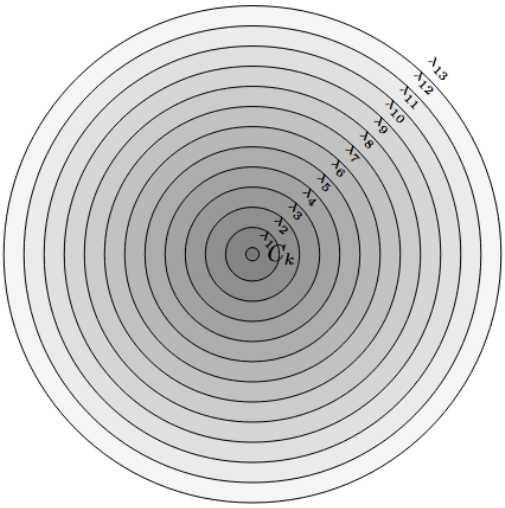


Figure 5 Distance Layers per Score bin

In this approach the regression model is formed by collecting the data patterns a the i_{th} layer from every score bin cluster as shown in Figure 5.

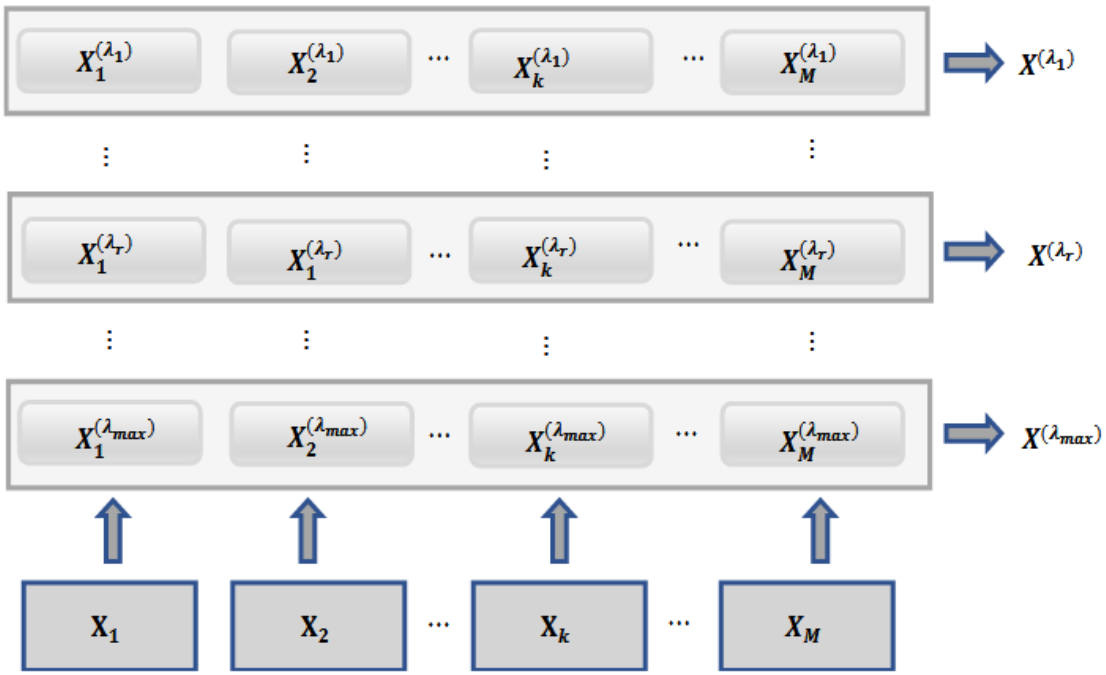


Figure 6 Data Segmentation Process

7.3 GENETIC PROGRAMMING

A partition-wise segmentation of the dataset will be utilized in order to generate appropriate instances of training and testing data subsets. Thus, the GP - based approximation of the FICO score will be assessed for its efficiency in evaluating the empirical probability of default in each bin both during training and testing. The main assumption concerning the partitioning of the dataset relates to the existence of an even number of partitions.

7.3.1 Shuffling

We begin by shuffling our data in order to have a random order. Afterwards we partition the target values (normalized Fico scores) into bins with a uniform width to reveal the underlying shape of the distribution.

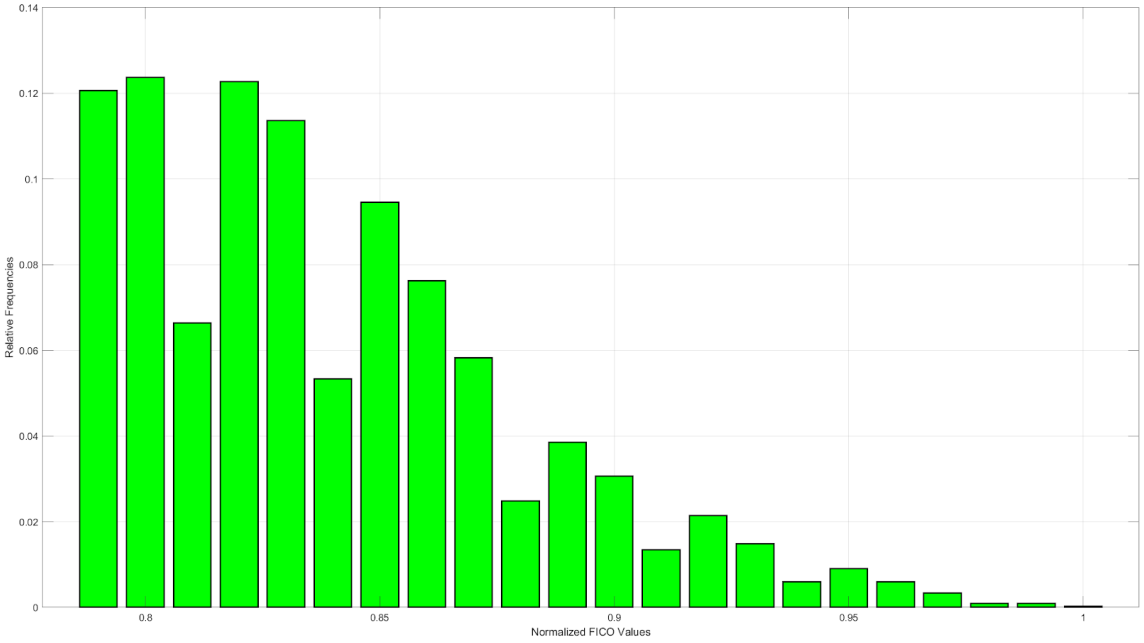


Figure 7 Permutated Count Density Histogram for the average FICO Values

7.3.2 Filtering

For the next step we are filtering the data based on the target values which are associated with every observation. The dataset is assumed to be organized into distinct clusters (i.e. bins) according to the corresponding target values. The primary purpose of this function is to eliminate existing observations within each bin based on the distance from the corresponding centroid point. That is, in each bin, the datapoints to be kept are those whose distances from the corresponding centroid do not exceed a given distance threshold with respect to the centroid of the bin.

The following graphs show the distances and the density of 3 of the generated bins:

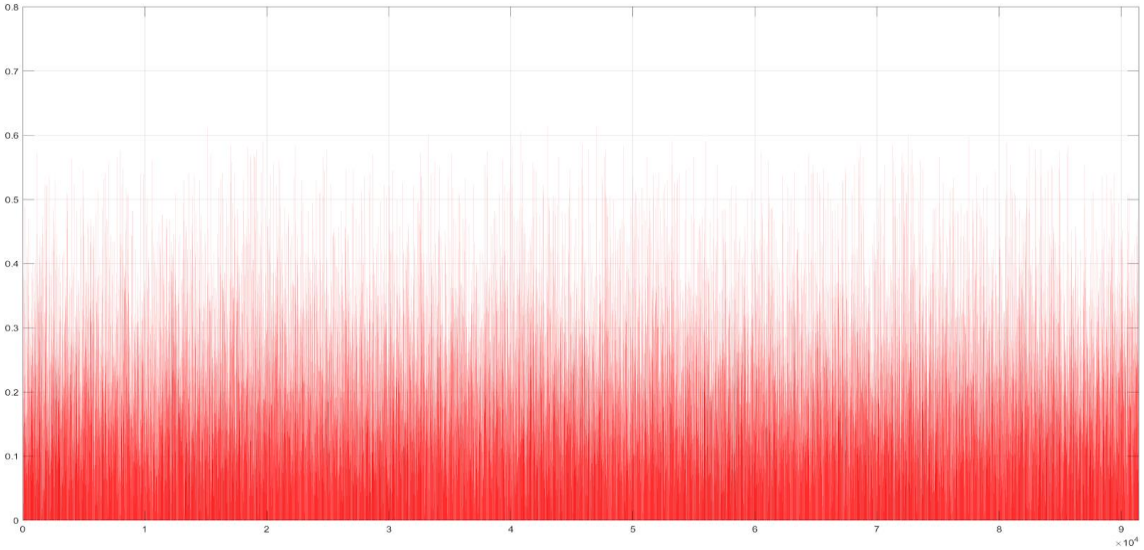


Figure 8 Distances within Bin 1

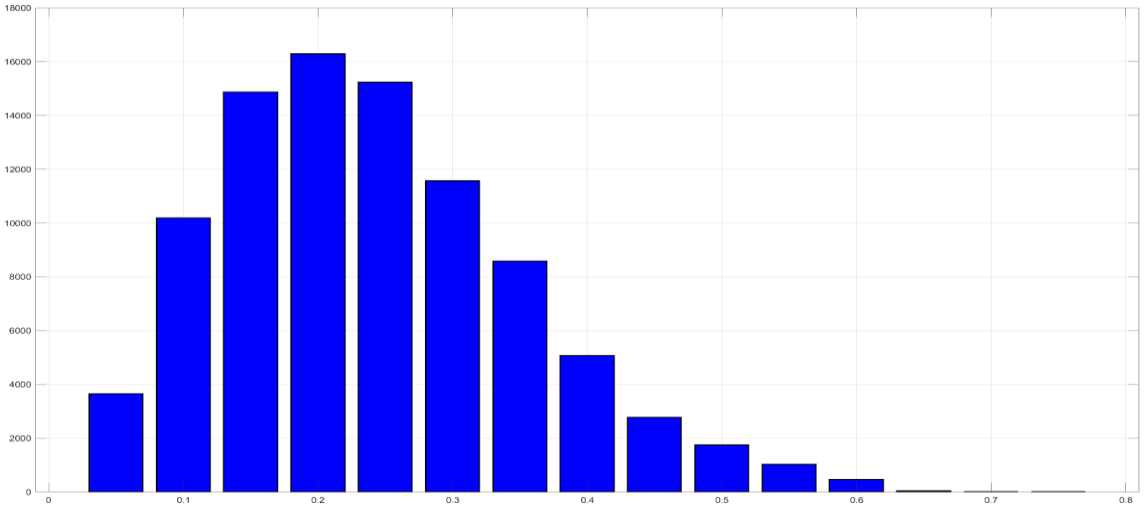


Figure 9 Count Density Histogram of Distances within Bin 1

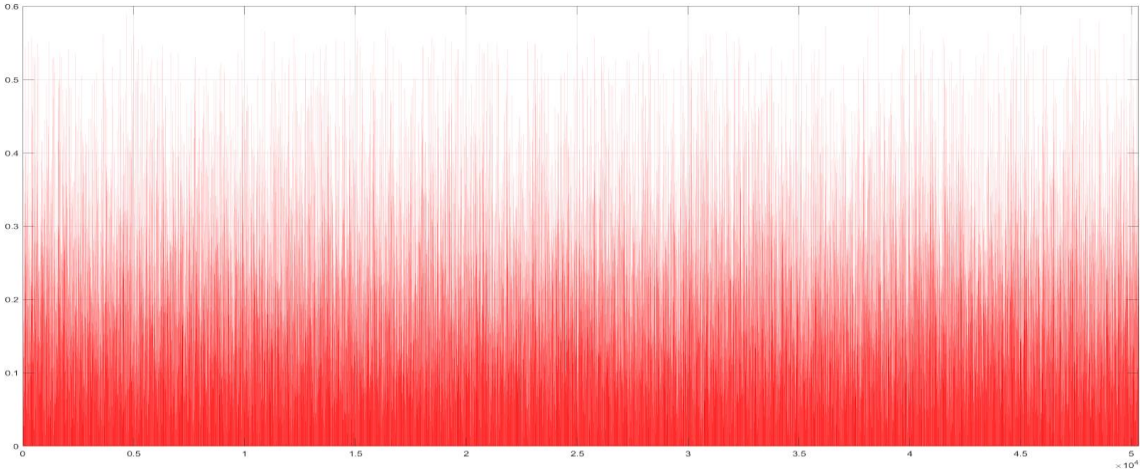


Figure 10 Distances within Bin 3

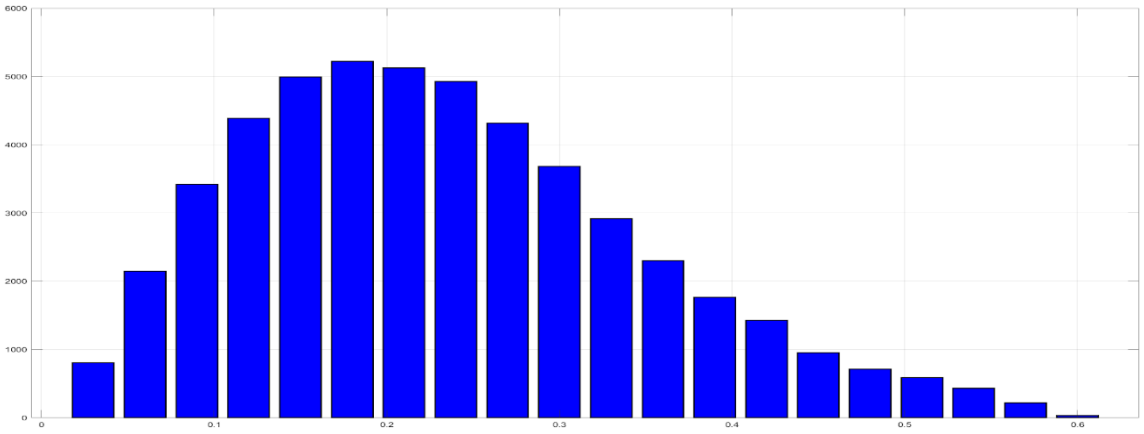


Figure 11 Count Density Histogram of Distances within Bin 3

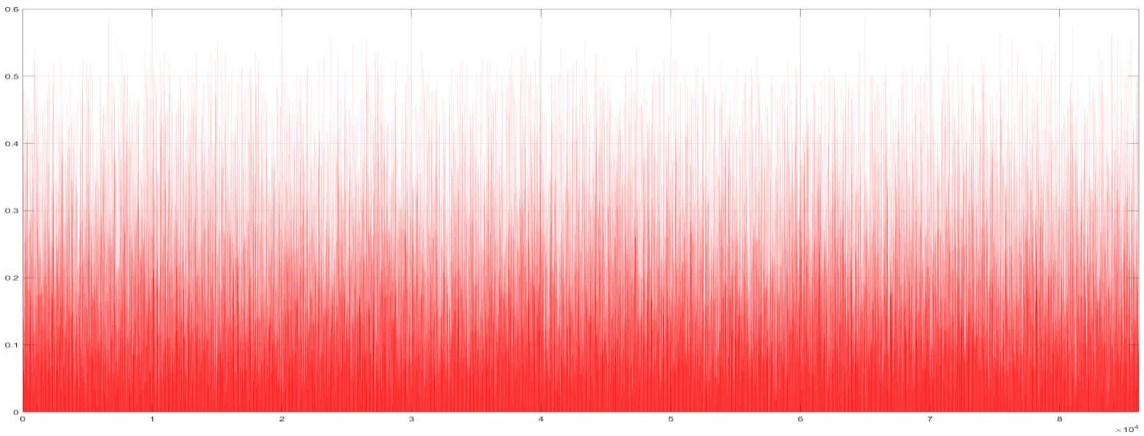


Figure 12 Distances within Bin 5

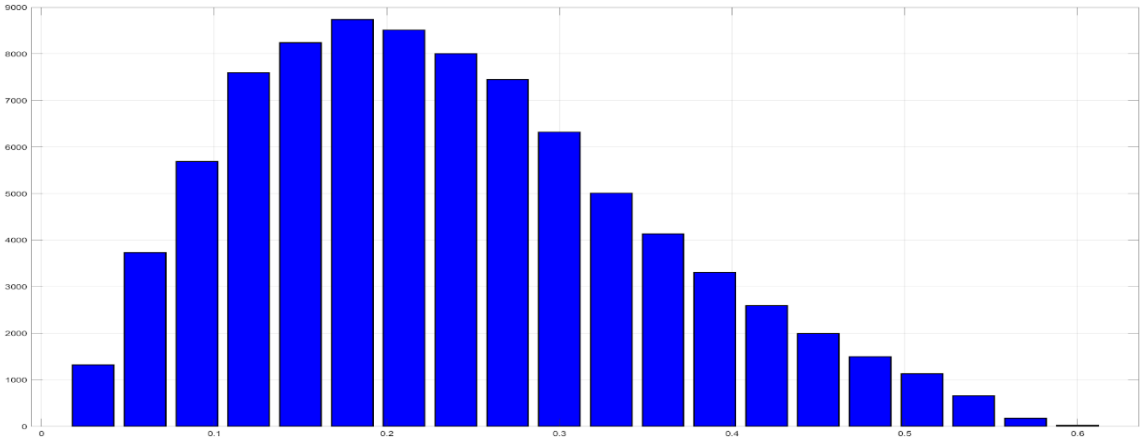


Figure 13 Count Density Histogram of Distances within Bin 5

We calculate the empirical probability of default of each score by also considering the final status of the loan [1 = Fully Paid, 2 = Default] for each fold per bin.

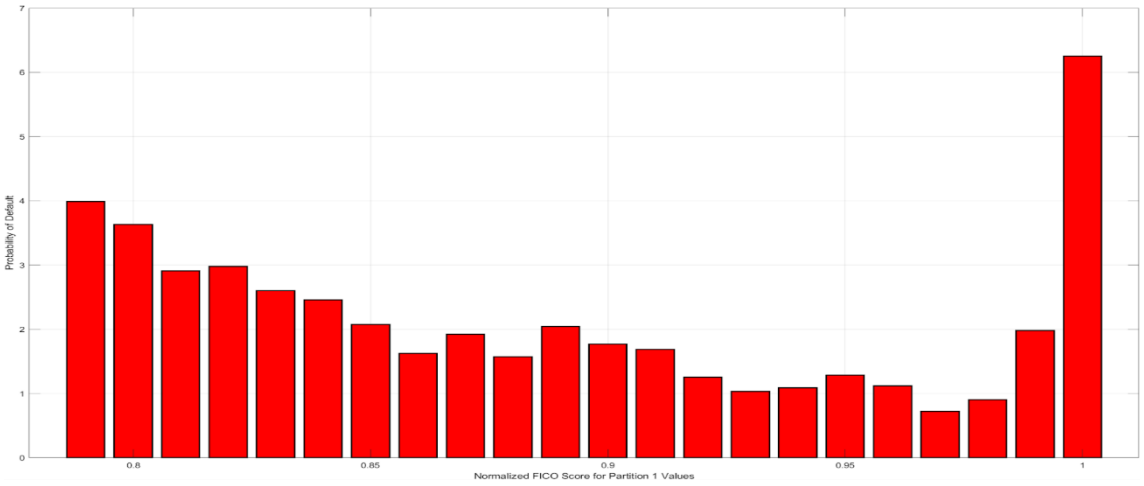


Figure 14 Empirical PD for FICO Score for Partition 1 per Bin

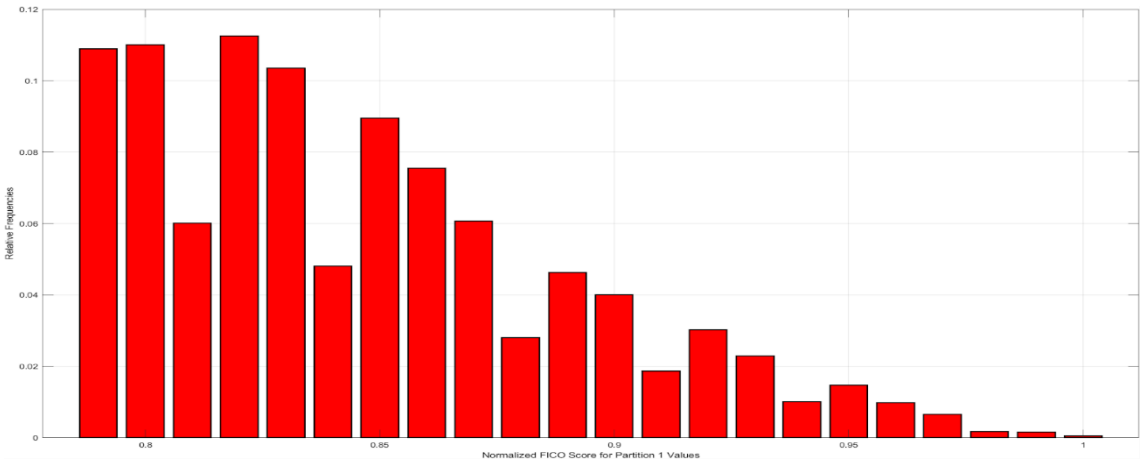


Figure 15 Empirical PD for FICO Score for Partition 1 per Bin

7.3.3 Model Run Configuration

The GP model is trained on the training subsets and validated against the testing subsets of each Partitions.

The Genetic Programming model uses a biologically inspired machine learning method called multigene genetic programming (MGGP) as the Hypothesis-ML engine that drives the automatic model discovery process. As a result it generates rules/models/hypotheses in the form of multiple trees. One of the most common applications of this model is to perform explainable symbolic nonlinear regression.

Symbolic machine learning is the process of extracting hidden, meaningful relationships from data in the form of symbolic equations and the models are 100% transparent in both its mode of operation (the sequence and type of computations) and the features included in the model. This often yields new insight into the physical systems or processes that generated the data.

The model has a Population Size = 500 is trained for 300 generations (Ngenerations = 300). The mutation rate is 0.1 and the crossover 0.7. The tree structure created by the model uses functions as nodes to come up with possible solutions, these functions can be one of ['times', 'minus', 'plus', 'rdivide', 'square', 'tanh', 'exp', 'log', 'mult3', 'add3', 'sqrt', 'cube', 'negexp', 'neg', 'abs']. Each tree can be described as a partial model fragment which has a weighted contribution to the full model. The GP algorithm follows a least squares procedure to minimise the sum of squared errors (SSE) with respect to the training data to determine the weights, based on this procedure the determined weights are guaranteed to be optimal in the least squares sense. More specifically the weights are computed by means of the Moore-Penrose pseudo inverse to mitigate collinearity problems caused by the possible existence of duplicate trees in candidate models. Each tree is represented by a compact coded string. These strings facilitate the machine learning process of simulated evolution. Based on that process new populations of trees are created, using the tree mutation and crossover operations these trees are better from the existing ones [24][25].

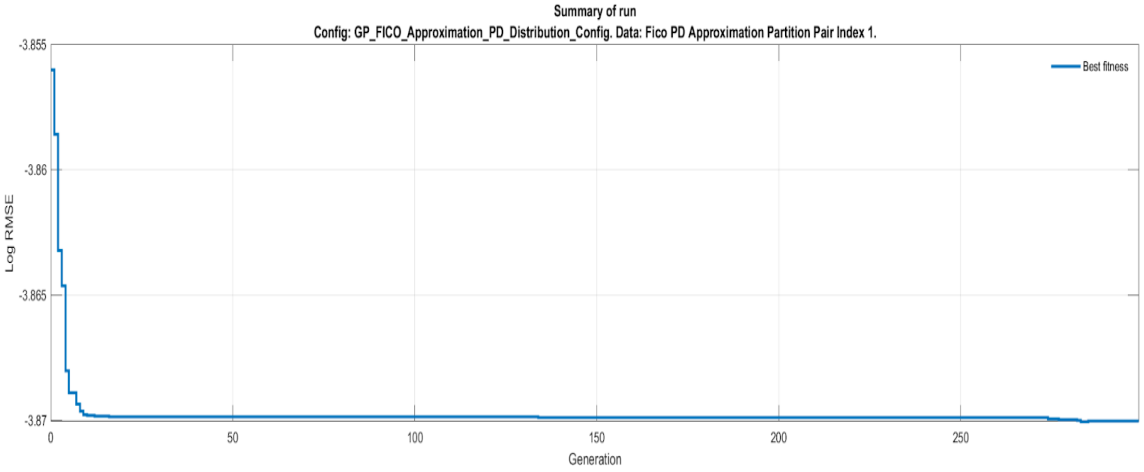


Figure 16 Breakthroughs as generations are created, each generation being better than the previous

The last generation contains 500 trees with the best genes as each generation of trees is better than the previous one. This population has of course better and worse performing individual trees. We gather the 10 best performing trees of the population and rank them based on the performance and simplicity of the overall model so the best performing model could have a slightly lesser accuracy metric but also be a lot shallower and generally easier to interpret. This happens because we haven't restricted the depth of the tree so the model can get quite complex.

7.3.4 Model Results

Overall model after simplification. Numerical precision reduced for display purposes.

$$y = 99.9\tanh(revol\_util^{\frac{1}{2}}) - 33.5\ revol\_util + 39.5\ exp(-1.0\ real(dti))$$
$$+ 102.0\ exp(-1.0\ revol\_util^{\frac{1}{2}})\ + 37.0\ tanh(dti) + 5.85\ (revol\_bal\ revol\_util^2)^{\frac{1}{2}}$$
$$-\ 213.0\ revol\_bal\ revol\_utl^2 - 33.5\ dti^3 + 467.0\ revol\_util^3 + 0.352\ revol\_util^{\frac{1}{4}}$$
$$- 677.0\ revol\_util^{(\frac{7}{2})} - 141.0$$

The following table (Table 3) illustrates the individual genes per model term.

Term	Value
Bias	-141.0
Gene 1	$revol\_util^{\frac{1}{4}}$
Gene 2	$39.5\ exp(-1.0\ real(dti))$
Gene 3	$-\ 213.0\ revol\_bal\ revol\_utl^2 - 33.5\ dti^3$
Gene 4	$102.0\ exp(-1.0\ revol\_util^{\frac{1}{2}})$
Gene 5	$99.9\tanh\ (revol\_util^{\frac{1}{2}})$
Gene 6	$37.0\ tanh(dti)$

Gene 7	$467.0\ revol\_util^3$
Gene 8	$5.85\ (revol\_bal\ revol\_util^2)^{\frac{1}{2}}$
Gene 9	$179.0\ revol\_bal\ +\ 179.0\ dti^3$
Gene 10	$677.0\ revol\_util^{(\frac{7}{2})}$

Table 3 Individual Genes/Model terms

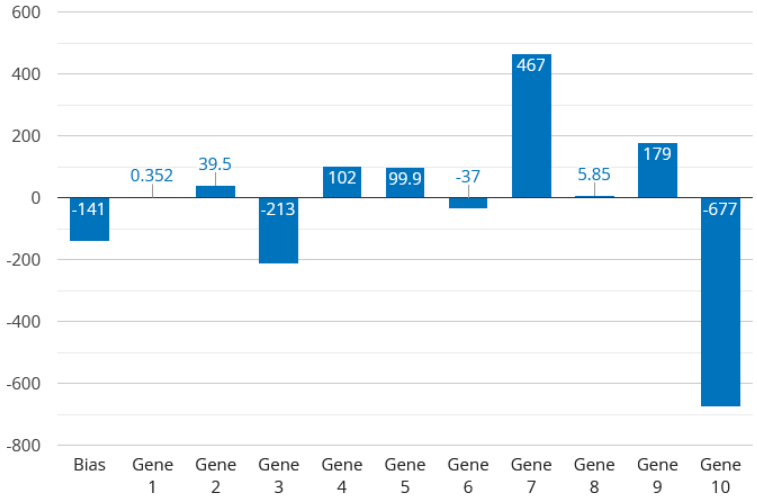


Figure 17 Gene weights

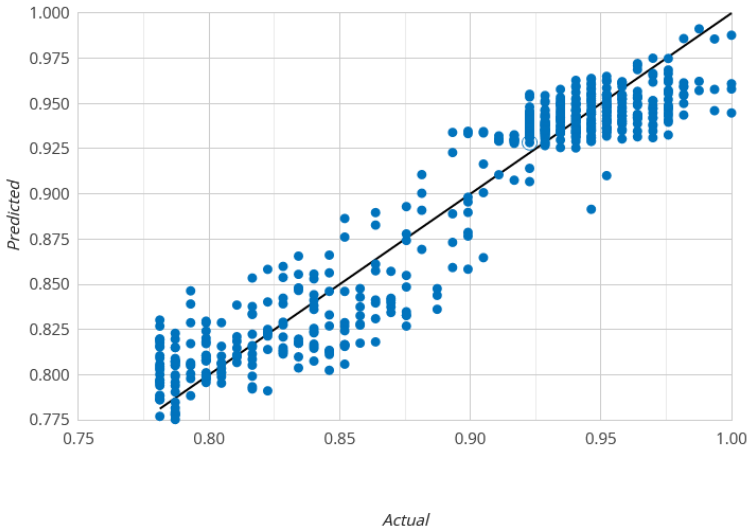


Figure 18 Predicted vs Actual

The above gene was chosen between 10 resulting genes. The way GPTIPS chooses the final model is by waging each model's complexity and accuracy. The following image illustrates the chosen model's gene tree structure (Figure 19):

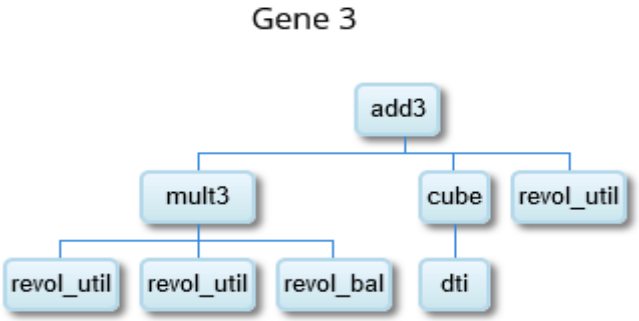


Figure 19 Gene tree structure

8. Conclusions

8.1 MODEL COMPARISONS

Layer	RMSE	MAE	R-Squared
Layer1	0.06541	0.0533	0.82431
Layer2	0.07827	0.06159	0.77041
Layer3	0.08527	0.06539	0.72718
Layer4	0.08787	0.06574	0.71033
Layer5	0.09034	0.06608	0.68656
Layer6	0.09081	0.0656	0.67525
Layer7	0.09224	0.06478	0.66035
Layer8	0.0929	0.06577	0.65126
Layer9	0.09177	0.06443	0.66464
Layer10	0.09286	0.06421	0.65623
Layer11	0.09235	0.06424	0.65299
Layer12	0.09151	0.06457	0.61042
Layer13	0.13172	0.09803	0.32903

Table 4 GP Regression: Training accuracy metrics

Layer	RMSE	MAE	R-Squared
Layer1	0.0655	0.05334	0.8238
Layer2	0.07838	0.06165	0.76973
Layer3	0.08508	0.06529	0.72839
Layer4	0.0877	0.06569	0.71139
Layer5	0.09038	0.06616	0.68602
Layer6	0.09073	0.06551	0.67572
Layer7	0.09254	0.06482	0.65817
Layer8	0.09302	0.06592	0.6504
Layer9	0.09189	0.06451	0.66362
Layer10	0.09277	0.06418	0.65685
Layer11	0.09261	0.06441	0.65112
Layer12	0.09164	0.06462	0.60932
Layer13	0.13174	0.09804	0.32874

Table 5 GP Regression: Testing accuracy metrics

Layer	RMSE	MAE	R-Squared
Layer1	0.06353	0.05075	0.83577
Layer2	0.06718	0.04709	0.83484
Layer3	0.06957	0.04515	0.82322
Layer4	0.07008	0.04326	0.82035
Layer5	0.07213	0.04334	0.80487
Layer6	0.07214	0.04283	0.79932
Layer7	0.07388	0.04159	0.78657
Layer8	0.07311	0.04181	0.78817
Layer9	0.07254	0.04137	0.79425
Layer10	0.07327	0.04277	0.78944
Layer11	0.07253	0.0429	0.78904
Layer12	0.07286	0.04606	0.75617
Layer13	0.11804	0.08292	0.46942

Table 6 Gaussian SVM Regression: Training accuracy metrics

Layer	RMSE	MAE	R-Squared
Layer1	0.0637	0.05095	0.83493
Layer2	0.06733	0.04727	0.83407
Layer3	0.06972	0.04534	0.82239
Layer4	0.07027	0.04345	0.81934
Layer5	0.07231	0.04355	0.80383
Layer6	0.07239	0.04304	0.79792
Layer7	0.07405	0.04178	0.78548
Layer8	0.07332	0.04205	0.78683
Layer9	0.0728	0.0416	0.79275
Layer10	0.07354	0.04301	0.78777
Layer11	0.07283	0.0432	0.78723
Layer12	0.07326	0.04639	0.7534
Layer13	0.11881	0.08361	0.46252

Table 7 Gaussian SVM Regression: Testing accuracy metrics

Layer	RMSE	MAE	R-Squared
Layer1	0.06365	0.05166	0.83365
Layer2	0.06612	0.04958	0.83617
Layer3	0.06872	0.04922	0.82286
Layer4	0.06923	0.0477	0.82032
Layer5	0.07119	0.04813	0.80543
Layer6	0.0712	0.04748	0.80061
Layer7	0.07059	0.04566	0.80111
Layer8	0.07244	0.04767	0.78774
Layer9	0.07085	0.04666	0.80019
Layer10	0.07297	0.04828	0.78779
Layer11	0.07216	0.04815	0.7882
Layer12	0.07391	0.05099	0.74603
Layer13	0.11761	0.08644	0.46511

Table 8 MLP Regression: Training accuracy metrics.

The summary of the accuracy measurements in terms of RMSE, MAE and R-Squared, are given in the tables above for each layer during both training and testing. Tables 4 and 5 show the measurements for our GP-based model and tables 6 to 8 show measurements of various

machine-learning based regression models such as Support Vector Machines (SVMs) and Multi layered Perceptron models (MLPs). It is critical to state the fact that our model, while being highly performant and efficient, remains highly transparent, tunable, and explainable as this has been a requirement and thus a limitation that had to be met and must be met by any industry grade credit risk model. Furthermore, it's crucial to note that all models (including ours) have the same degradation in accuracy as we move further away from the center of the cluster C_k .

8.2 FUTURE WORK

In the future, our work will focus on developing an ensemble of the individual regression models in order to increase the approximation efficiency within the upper layers of the dataset.

Bibliography

1. International Committee on Credit Reporting, and World Bank Group. “CREDIT SCORING APPROACHES GUIDELINES.” The World Bank, 2019, https://pubdocs.worldbank.org/en/935891585869698451/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB.pdf. Accessed 15 April 2021.

2. Investopedia. “Credit Score.” Credit Score, https://www.investopedia.com/terms/c/credit_score.asp.

3. The Street. “A Secret History of Credit Scores: Who Determined What Matters and Why.” https://www.thestreet.com/personal-finance/credit-cards/a-secret-history-of-credit-scores-who-determined-what-matters-and-why-13097739.

4. Wikipedia. “Credit score.” https://en.wikipedia.org/wiki/Credit_score.

5. Experian. “What Information Credit Scores Do Not Consider.” What Is a Good Credit Score?, https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/.

6. “Why There Are Different Credit Scores.” What Is a Good Credit Score?, https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/.

7. “What's in my FICO® Scores?” What's in my FICO® Scores?, https://www.myfico.com/credit-education/whats-in-your-credit-score.

8. FICO. “A Deep Dive into the Distribution of the FICO Score Across the US.” https://www.fico.com/blogs/deep-dive-distribution-fico-score-across-us.

8.5 FICO. US Average FICO Score Hits 700: A Milestone for Consumers
https://www.fico.com/blogs/us-average-fico-score-hits-700-milestone-consumers

9. Lopez, Jose A., and Marc R. Saidenberg. Evaluating Credit Risk Models. https://www.frbsf.org/economic-research/files/wp99-06.pdf.

10. ListenData. “Weight of Evidence (WOE) and Information Value (IV) Explained.” https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

10. Djeundje, Viani B., et al. Enhancing credit scoring with alternative data, https://www.sciencedirect.com/science/article/abs/pii/S095741742030590X.

11. UrbanWire. “Adopting Alternative Data in Credit Scoring Would Allow Millions of Consumers to Access Credit.” https://www.urban.org/urban-wire/adopting-alternative-data-credit-scoring-would-allow-millions-consumers-access-credit.

12. Porche, Brady. “Do you have a creditworthy personality?” https://www.creditcards.com/credit-card-news/psychometric-credit-scoring/.

13. Bachmann, Janio Martinez. “Lending Club || Risk Analysis and Metrics.” https://www.kaggle.com/janiobachmann/lending-club-risk-analysis-and-metrics.

14. Wikipedia. “LendingClub.” LendingClub, https://en.wikipedia.org/wiki/LendingClub.

15. Polena, Michal. “Performance Analysis of Credit ScoringModels on Lending Club Data.” https://is.cuni.cz/webapps/zzp/download/120269679/?lang=cs.

16. Ivestopedia. “Five C's of Credit.” https://www.investopedia.com/terms/f/five-c-credit.asp.

17 .DeepAI. “Logistic Regression.” https://deepai.org/machine-learning-glossary-and-terms/logistic-regression.

18. Deloitte. “Credit scoring Case study in data analytics.” https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-be-aers-fsi-credit-scoring.pdf.

19. West, David. Neural network credit scoring models.

20. Investopedia. “Neural Network.” https://www.investopedia.com/terms/n/neuralnetwork.asp.

21. Genetic Programming. “Tree based Genetic Programming.” http://geneticprogramming.com/about-gp/tree-based-gp/.

22. Vanneschi L., Poli R. (2012) Genetic Programming — Introduction, Applications, Theory and Open Issues. In: Rozenberg G., Bäck T., Kok J.N. (eds) Handbook of Natural Computing. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-92910-9_24

23. arXiv:2012.03749v1 [q-fin.RM]

24. GPTIPS 2: an open-source software platform for symbolic data mining, Searson, D.P. Chapter 22 in Handbook of Genetic Programming Applications, A.H. Gandomi et al., (Eds.), Springer, New York, NY, 2015 .

25. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression, Searson, D.P., Leahy, D.E. & Willis, M.J., Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 (IMECS 2010), Hong Kong, 17-19 March, 2010.

26. https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

27. Applied Mathematical Sciences, Vol. 8, 2014, no. 65, 3229 - 3242 HIKARI Ltd, www.m-hikari.com http://dx.doi.org/10.12988/ams.2014.44300

28. Churn Analysis Using Information Value and Weight of Evidence, https://towardsdatascience.com/churn-analysis-information-value-and-weight-of-evidence-6a35db8b9ec5#9557

29. Koláček, Jan & Rezac, Martin. (2008). Assessment of scoring models using information value. COMPSTAT'2010 Book of Abstracts. 296.

30. Salome Tabagari. (2015). Credit scoring by logistic regression

31. Fang, Yongsheng & li, Jun. (2010). A Review of Tournament Selection in Genetic Programming. 181-192. 10.1007/978-3-642-16493-4_19.