Contents

List of Figures

List of Tables

1. Introduction

Credit scores, which represent numerical values of risk associated with an individual on the basis of his/her credit history and financial behavior, are the real drivers of modern finance. Banks and credit card companies use credit scores to ascertain the amount of risk involved in lending money to a borrower. If one has a high credit score, he can expect better interest rates and terms of loans, and it will make obtaining a mortgage, an auto loan, or even a personal loan quite easy. Moreover, credit scores impact insurance premium rates, apartment rentals, and even job applications, since some employers view credit history as a yardstick for one's reliability and character. Understanding how to properly work with credit scoring will allow people more flexibility and stability in this financially driven age.

The credit scores are the results of applying complex mathematical algorithms to data from a person's credit report. Traditionally, lenders used to assess the credit risk manually in the early part of the 20th century because there were no standardized scoring systems then in use. This changed by the late 1950s and 1960s, with the development of mathematical models for predicting credit risk and the need to speed up the process of lending.

An even major breakthrough came about in 1989, introducing the FICO score. It is a three-digit number ranging from 300 to 850 and the first credit risk model developed using statistical analysis. Also, it would prove to be the most widely used credit scoring model in the United States. Its simplicity and reliability gained wide acceptance of the FICO score.

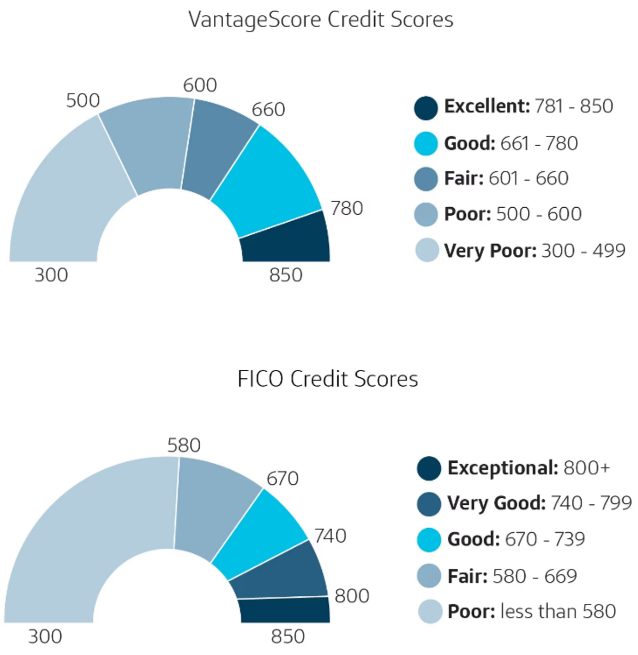


Figure 1The Credit Score Scales

With the growing relevance of credit scores, firms began working on in-house credit scoring models to compete with FICO. One very good example is the VantageScore, introduced in 2006. Nowadays, with technological advancement, models of credit scoring are very sophisticated, able to efficiently process vast volumes of credit data and make very accurate risk assessments that will be better at distinguishing among consumers based on their creditworthiness.

Even after all these developments, most people still have a bad credit score simply because they are uninformed and ill aware of the rules of the game that govern credit scores. Therefore, educating people on the fundamentals that drive credit scores is very relevant to achieving better financial health.

Thorough analyses of data on finance should be done to identify trends and key factors that largely control credit scores.
This analysis will evaluate trends identification and factors associated with credit scores on the Credit Score Classification Dataset taken from Kaggle. We are going to look through this data while trying to answer such questions: what are the chief credit-related factors and which of the personal data elements are heavy on credit scores?.

2. Description of the Question

Our main variable of interest was "Credit Score," which is a qualitative variable users derived from their financial history and personal track record. Today, society is embracing credit in almost all purchases or monetary decisions. For example, if you want to buy a home, the mortgage lending institution will use your credit data to estimate the probability of your defaulting on the loan repayment. You may struggle to get a mortgage with bad credit, or you will with higher interest rates, directly impacting the monthly cost of your mortgage payment. Those with low credit scores may be rejected for the loan or charged higher rates.

Good credit offers a host of advantages that include, among others, the following:

- Lower auto insurance premiums
- Lower premiums on home insurance
- Lower interest rates on credit cards
- Higher credit limits on credit cards
- More likely to get approved for utility services, such as electricity or internet
- You are an attractive candidate for potential employers

Some of our key objectives are:

- Determination of credit-related properties and personal banking data that impact the credit score
- Determination of the optimal conditions of these impactful variables so as to keep a good credit score
- Development of a predictive model in order to predict credit scores based on users' financial history and personal data

Due to the fact that machine learning models can support large data sets, credit score computation is way more accurate as compared to traditional ways by identifying patterns for precise predictions.

3. Description of the Dataset

The dataset used here is the Credit Score Classification Dataset from Kaggle, which includes basic bank details and some credit-related information gathered from 12,500 people in a global finance company. There were, in all, 100,000 observations in this dataset, of monthly data collected from each individual during the January to August period. It contains 29 variables, including one response variable of interest: Credit Score, classifying credit scores into three groups.

Variable	Description of variable	Variable Type
ID	The unique identification of an entry	String
Customer ID	The unique identification of a customer	String
Month	Month of the Year	Categorical(Ordinal)
Name	Name of the person	String
Age	Age of the person	Numerical
SSN	The social security number of a person	String
Occupation	The occupation of the person	Categorical(Nominal)
Annual Income	The annual income of the person	Numerical
Monthly_Inhand_Salary	The monthly salary of a person	Numerical
Num_Bank_Accounts	The number of bank accounts a person holds	Numerical
Num_Credit_Card	The number of other credit cards held by a person	Numerical
Interest_Rate	The interest rate on credit card (percent)	Numerical
Num_of_Loan	The number of loans taken from the bank	Numerical
Type_of_Loan	The type of the loans taken from the bank	Categorical(Nominal)
Delay_from_due_date	The average number of days delayed from the payment date in days	Numerical
Num_of_Delayed_Payment	The average number of payments delayed by a person	Numerical
Changed_Credit_Limit	The percentage change in credit card limit (percent)	Numerical
Num_Credit_Inquiries	The number of credit card inquiries	Numerical
Credit_Mix	The classification of the mix of credits (Bad, Standard, Good)	Categorical(Ordinal)
Outstanding_Debt	The remaining debt to be paid	Numerical
Credit_Utilization_Ratio	The utilization ratio of credit card (percent)	Numerical
Credit_History_Age	The age of credit history of the person (days)	Numerical
Payment_of_Min_Amount	Represents whether only the minimum amount was paid by the person	Categorical(Ordinal)
Total_EMI_per_month	The monthly Equated Monthly Installment made by customer	Numerical
Amount_invested_monthly	The monthly amount invested by the customer	Numerical
Payment_Behaviour	The payment behavior of the customer	Categorical(Nominal)
Monthly_Balance	The monthly balance amount of the customer	Numerical
Credit_Score	Represents the bracket of credit score (Poor, Standard, Good)	Categorical(Ordinal)

Table 1 Dataset Description

4. Data Cleaning and Preprocessing

The original dataset was divided into test and train sets. However, it was noted that the test set was created specifically for competition purposes and did not include credit scores. As a result, this test set was not suitable for future modeling tasks. Hence training set consider as the primary data source. The train dataset contained **100,000 records** with **12,500 unique customers**.

Values of numerical variables such as ***Age, Annual_Income, Num_of_Loan, Num_of_Delayed_Payment, Changed_Credit_Limit, Outstanding_Debt, Amount_invested_monthly,*** and ***Monthly_Balance*** were found to be in object form, with some values starting or ending with string characters. These characters were identified and removed, converting the values into numerical form.

For a particular Customer_ID, variables such as ***Name, Age, SSN, Occupation, Annual_Income, Interest_Rate, Num_of_Loan, Type_of_Loan, Credit_Mix,*** and ***Outstanding_Debt*** typically have unique values for every month. However, occasional discrepancies were observed in one or two months for certain Customer_IDs. These discrepancies were considered human errors and were corrected by assigning the appropriate values.

For a particular Customer_ID, variables such as ***Monthly_Inhand_Salary, Num_Bank_Accounts, Num_Credit_Card, Interest_Rate, Num_of_Loan, Num_of_Delayed_Payment, Changed_Credit_Limit, Num_Credit_Inquiries, Credit_Mix,*** and ***Total_EMI_per_month*** typically follow clear patterns over time. However, occasional unusual values disrupt these patterns. These anomalies were identified and corrected using forward and backward fill methods.

There were 1,426 customers without any Type_of_Loan value. Consequently, it was decided to drop these customers from the dataset. Missing values in ***Amount_invested_monthly, Payment_Behaviour,*** and ***Monthly_Balance*** were then addressed. For each Customer_ID, these missing values were imputed using the **mean and mode** of the respective customer's data.

The Month variable was then converted to a numerical format. Initially, the Type_of_Loan variable had over 6,000 unique categories, making analysis challenging. To address this, the Type_of_Loan variable was transformed into a numeric format using NLP techniques. This approach leverages the potential relationships and sequences among different loan types. By applying NLP techniques, a numeric value was obtained, enabling the identification of similar loan combinations more effectively.

After completing all the cleaning and preprocessing steps, the final dataset contained **88,592 records** with **11,074 unique customers**.

When splitting the dataset into train and test sets, randomly dividing the entire dataset is not meaningful because all instances of a customer must be treated as a single unit. Therefore, the Customer_IDs are first divided randomly, with 80% allocated for training and 20% for testing. Then, all instances belonging to each customer are placed in the correct set accordingly.

5. Results of Descriptive Analysis

5.1 Target variable – Credit Score

Assuming that the observations are independent of one another, a bar plot was created for the response variable, Credit Score. This variable consists of three categories: Poor, Standard, and Good.

Due to the lack of information on the recoding scheme used, the FICO score ranges (300-850) were employed to define these categories:

- **Good Category**: Very Good (740+)
- **Standard Category**: Average (670–739)
- **Poor Category**: Poor (less than 669)

The analysis revealed that the majority of observations fall within the Standard range, with relatively fewer observations in the Good and Poor categories. This indicates that most customers have a Standard Credit Score.

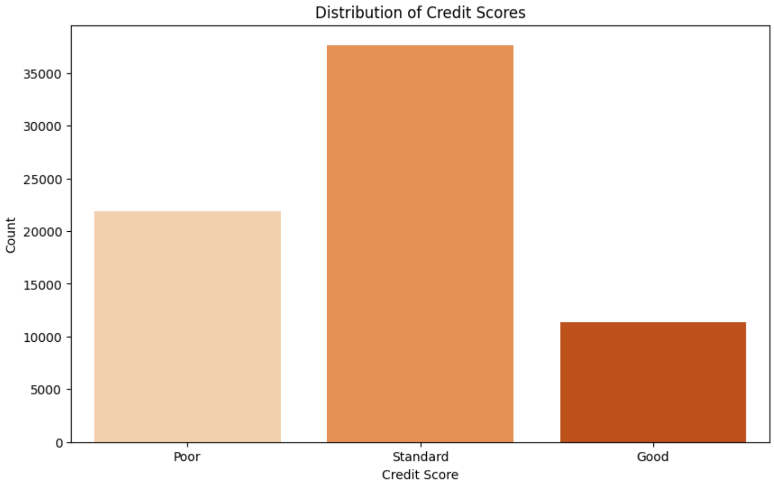


Figure 2 Bar plot for the Credit Scores

5.2 Correlations among the variables

The dataset contained both categorical and numerical predictors. Therefore, three approaches were used to determine the correlations among the variables:

1. **Spearman's Rank Correlation** was employed to assess the association between the numerical predictors and the categorical target variable.
2. **Chi-Square Test** was used to determine the association between the categorical predictors and the categorical target variable.
3. **Pearson Correlation Coefficient** was utilized to calculate the correlations among the numerical variables.

Table 2 Spearman Correlation values of variables

Variable Name	Spearman’s Rank Correlation
Age	-0.0299
Annual Income	-0.0368
Monthly_Inhand_Salary	-0.0356
Num_Bank_Accounts	0.0886
Num_Credit_Card	0.0250
Interest_Rate	0.0486
Num_of_Loan	0.0042
Delay_from_due_date	0.0596
Num_of_Delayed_Payment	0.1192
Changed_Credit_Limit	0.1830
Num_Credit_Inquiries	-0.0073
Outstanding_Debt	-0.0548
Credit_Utilization_Ratio	-0.0055
Credit_History_Age	-0.0445
Total_EMI_per_month	-0.0403
Amount_invested_monthly	-0.0272
Monthly_Balance	-0.0168

The analysis's most important factors were determined and looked at individually based on the results. It is significant to highlight that, for the sake of simplicity, the values were determined under the assumption that each observation is independent. But as was already mentioned, each customer's month-by-month data for the months of January through August was included in the dataset. With 11,074 customers in total, it was a laborious process to analyse the variables for each individual consumer. In order to solve this, the predictors were examined month by month, which made the research simpler while still taking each consumer into account.

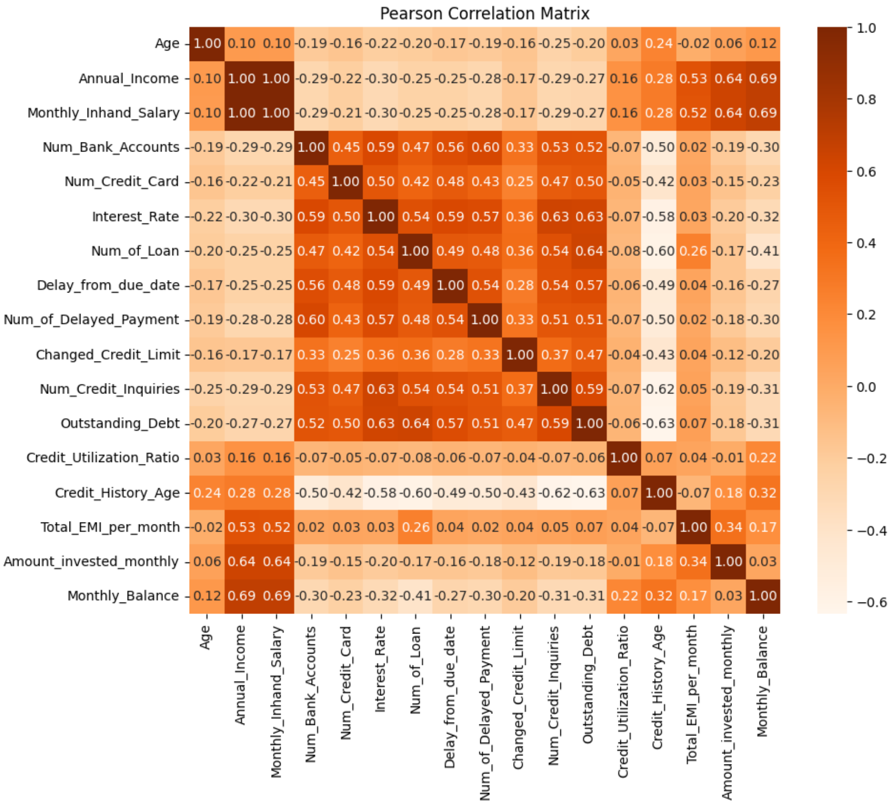


Figure 3 Pearson Correlation Plot

5.3 Variation of the predictor variables with the response variable

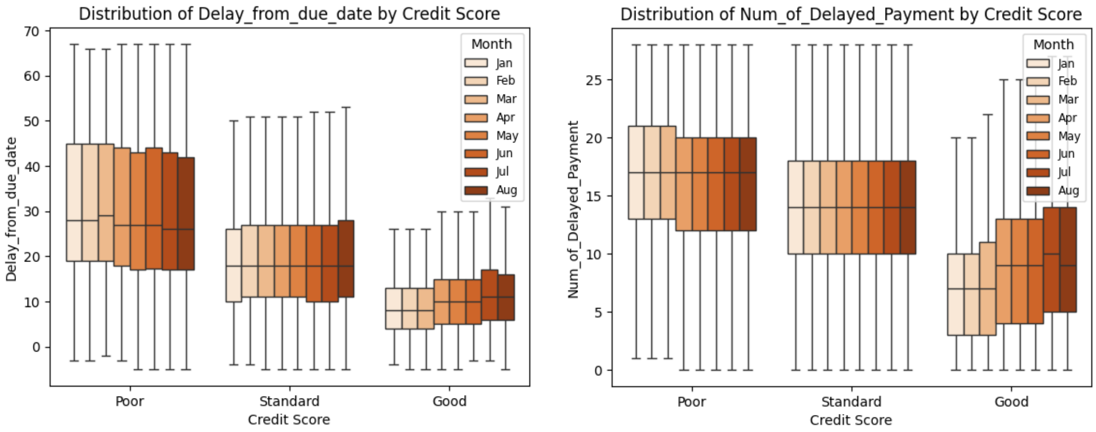


Figure 4 Distribution of Delay from due date and num of delayed payment by Credit Score

According to sources, the most important factor for credit scores is **payment history**, which accounts for 35% of the total importance. This category includes an individual's track record of paying bills on time, as well as any late payments, collections, and bankruptcies. In the dataset, the variables Delay_from_due_date and Num_of_Delayed_Payment are related to this factor. As observed in the plot, the credit score increases with the decreasing Delay_from_due_date, and Num_of_Delayed_Payment

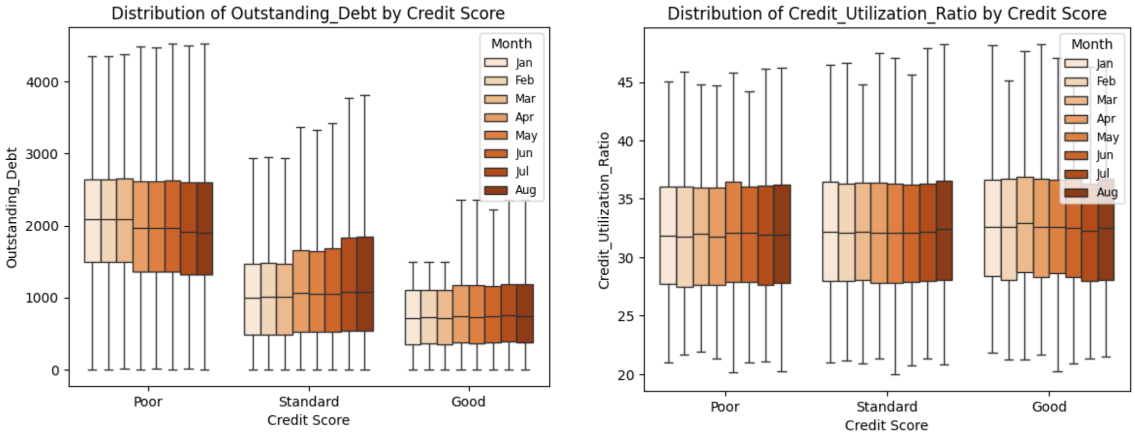


Figure 5 Distribution of Outstanding debt and Credit utilization ration by Credit Score

A**mounts owed** have a 30% importance in the credit score according to sources. This factor considers the total amount of debt an individual has, the number of accounts with balances, and how much of their available credit they are using. Variables such as Outstanding_Debt and Credit_Utilization_Ratio are related to this factor. As observed in the plot, the credit score increases with the decreasing Outstanding_Debt, but there is no significant variation in the Credit_Utilization_Ratio.

According to sources, the **length of credit history** holds a 15% importance in determining a credit score. This factor considers how long an individual's credit accounts have been established, including the age of the oldest account, the average age of all accounts, and the age of specific types of accounts. Generally, a longer credit history can contribute to a higher score. The variable Credit_History_Age is related to this factor.

Credit mix accounts for 10% of a credit score's importance according to sources. FICO scores evaluate the variety of credit accounts an individual has, such as credit cards, retail accounts, installment loans, finance company accounts, and mortgage loans. A diverse mix of different types of credit can positively impact the score. The variable Num_Credit_Card is related to this factor. But as observed in the plot, the credit score decreases with the increasing Num_Credit_Card.

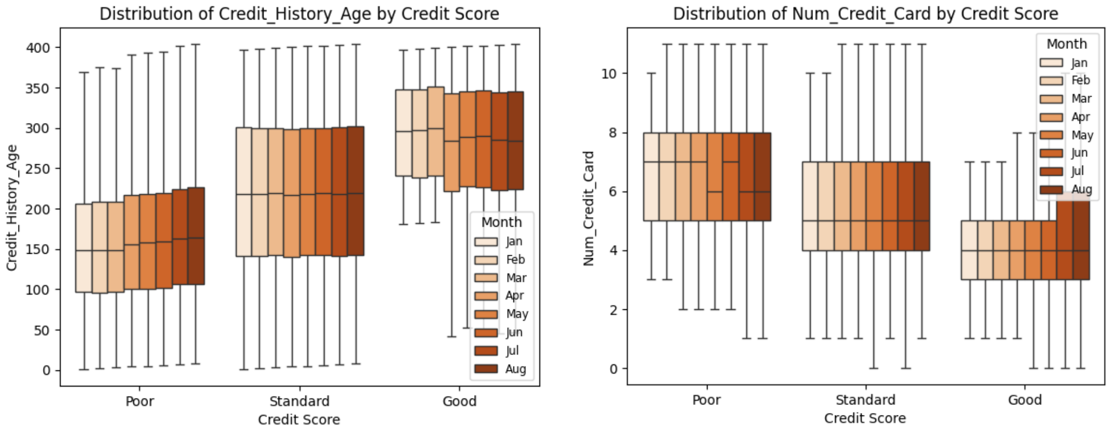


Figure 6 Distribution of Credit history age and Num of credit card by Credit Score

5.4 Bivariate Analysis of Variables

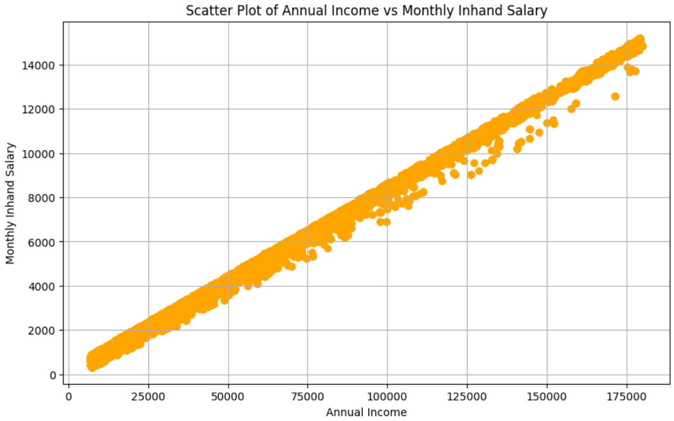


Figure 7 Scatter plot of Annual income vs Monthly inhand salary

There was a perfect correlation found between the year income and the monthly in-hand salary. The link arises from the fact that an increase in a person's monthly wage directly correlates with an increase in their income. While a perfect link was found, there is considerable dispersal in some areas. This could be because those who earn more each month have a tendency to invest, which raises their income.

There were weak negative correlations found between the number of loans, credit enquiries, and outstanding debt and the credit history age, measured in months.

This suggests that the Num_of_Loan reduces or vice versa as Credit_History_Age grows. A number of things could be the cause of this negative correlation, such as the fact that people with longer

Num_of_Loan may have developed a strong credit history and are hence less dependent on taking out multiple loans.

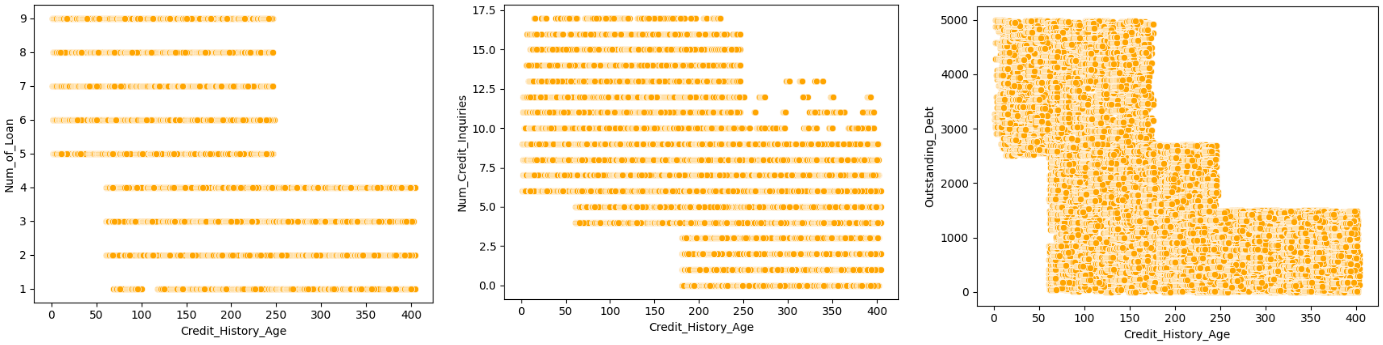


Figure 8 Scatter plots of Credit History Age with No. of loans, No. of credit inquiries and Outstanding Debt

Longer credit histories are generally seen by lenders as more creditworthy, which facilitates the acquisition of larger loans or the consolidation of preexisting debt.
Longer credit histories may indicate that a person has paid off or substantially decreased prior debt, which reduces the amount of current outstanding debt.
Individuals may experience changes in their credit demands and borrowing behaviours as they move through different periods of life. Elderly people may not require as many fresh loans, but younger people may be more inclined to take out several loans for different uses.

Likewise, those with a longer credit history tend to be more established, which lessens the necessity for credit enquiries. Additionally, because they have more expertise handling credit, they would be more circumspect about any potential repercussions while doing credit enquiries. This clarifies the unfavourable correlation.

Additionally, people with longer credit histories may have had more time to develop sound financial practices, which has led to lower levels of outstanding debt. This elucidates the inverse relationship between the age of credit history and outstanding debt.

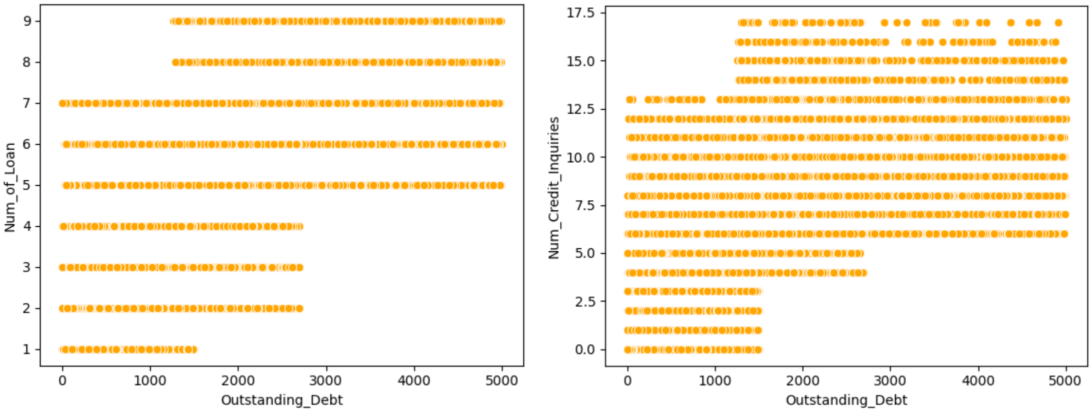


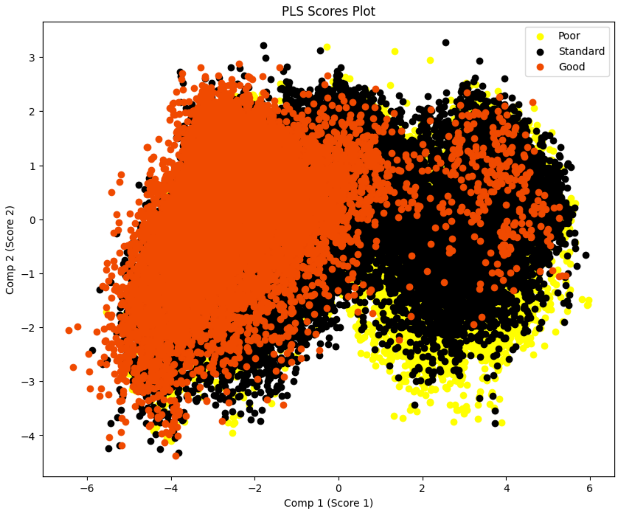
Figure 9 Scatter plots of Outstanding Debt with Credit Inquiries and No. of loans

The amount of outstanding debt positively correlates with both the number of credit enquiries and loan denials.

Higher debt levels may indicate that a person is actively using credit to meet their demands and fund different expenses. There may be a rise in credit enquiries as a result of their increased credit usage and subsequent need to apply for more credit. Additionally, some people may seek for new loans in order to lower their interest rates or consolidate their current obligations. Their goal is to better manage their outstanding debt, thus they do this, which raises the number of credit enquiries.

In the case of loans, people may take out more loans due to their increasing financial needs, which could raise their total amount of outstanding debt. Additionally, company owners and entrepreneurs may take out several loans in order to finance the growth and operation of their enterprises. Their outstanding debt may rise as they make more investments in their company.

5.5 Partial Least Squares Regression



Under the assumption that the observations are independent, Partial Least Squares Regression was performed on the dataset to identify any clusters among the observations and to identify significantly correlated predictors. In the score plot provided, nearly 31% of the variation is explained by the first two components. The plot indicates that there are no significant clusters in the observation set.

Figure 1010 Score Plot

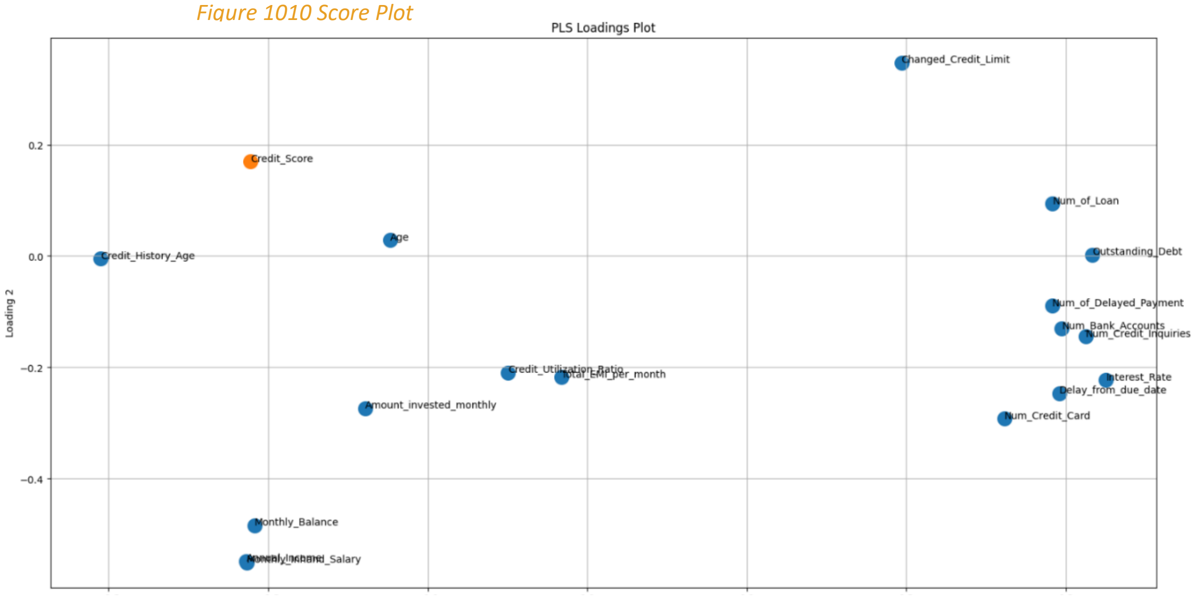


Figure 11 Loading Plot

Moving on to the loadings plot, it was observed that there are strong correlations among some predictors and the response, whereas some predictors are orthogonal to the response. Additionally, some variable clusters were identified.

- Variables such as Age and Credit_History_Age exhibit a positive association with the response. Conversely, variables including Num_of_Loan, Outstanding_Debt, Num_of_Delayed_Payment, Num_Bank_Accounts, Num_Credit_Inquiries, Interest_Rate, Delay_from_due_date, and Num_Credit_Card show a negative association with the response.
- Variables like Changed_Credit_Limit, Credit_Utilization_Ratio, Total_EMI_per_month, Amount_invested_monthly, Monthly_Balance, Annual_Income, and Monthly_Inhand_Salary are almost orthogonal to the response, indicating that these variables might not impact the credit score.
- Most notably, Monthly_Inhand_Salary and Annual_Income seem to overlap, confirming the strong association between the two variables discovered earlier.

6. Important Results of Advanced Analysis

After a thorough discussion and analysis of several related articles, the decision was made to treat each observation as an independent entity. Consequently, the customerID variable was removed from the analysis. However, the month variable was retained, although it will not be treated as time-sensitive data. Instead, the month will be considered as just another ordinary variable.

Given the imbalanced nature of the response variable, with the majority of observations falling into the Standard category, it was determined that an oversampling technique SMOTE, should be applied to the dataset.

The following machine learning models were initially fitted to gain an understanding of the data. Since the classes are not imbalanced, accuracy was used as the primary evaluation metric.

Model	Train Accuracy	CV Accuracy on Training Set	Test Accuracy
Random Forest	1.0000	0.8766	0.8808
Extra Tree	1.0000	0.8737	0.8753
XGBoost	0.8709	0.8374	0.8393
Decision Tree	1.0000	0.7939	0.7997
KNN	0.8914	0.7805	0.7872
Gradient Boost	0.7710	0.7682	0.7650
SVM (RBF)	0.7636	0.7591	0.7571
QDA	0.7381	0.7371	0.7350
Ada Boost	0.7321	0.7326	0.7281
LDA	0.7227	0.7228	0.7218
Logistic Regression	0.7216	0.7215	0.7205
Ridge	0.7170	0.7167	0.7164
SVM (Linear)	0.7164	0.7163	0.7186
Naïve Bayes	0.6972	0.6973	0.6944

Table 3 Results of Base Models

Although models like Random Forest, XGBoost, and KNN have higher training accuracies compared to test accuracies, they are not overfitting because the cross-validation accuracy and test accuracy are almost similar. This suggests that these models achieve high training accuracy due to the nature of the algorithms, which are capable of fitting the training data well. The consistency between the cross-validation and test accuracies indicates that the models are generalizing well.

The top five models, based on the highest training cross-validation accuracy, are selected for further optimization. These models include Random Forest, Extra Trees, XGBoost, Decision Tree, and K-Nearest Neighbors. Subsequently, these selected models are fine-tuned using Bayesian optimization.

Model	Train Accuracy	CV Accuracy on Training Set	Test Accuracy
Random Forest	1.0000	0.8766	0.8808
Extra Tree	1.0000	0.8737	0.8753
XGBoost	0.9986	0.8942	0.9020
Decision Tree	0.8833	0.7950	0.8021
KNN	1.0000	0.8451	0.8568

Table 4 Results of tuned top 5 models

The decision was made to use ensemble techniques on weak learners, including bagging classifiers, voting classifiers, and stacking classifiers. In the bagging classifiers, 100 similar models are employed. The voting classifiers utilizes hard voting, while the stacking classifiers uses logistic regression as the meta-classifier.

Model	Train Accuracy	CV Accuracy on Training Set	Test Accuracy
Bag(Logistic Regression)	0.7229	0.7223	0.7221
Bag(SVM RBF)	0.7687	0.7633	0.7598
Bag(Naïve Baye)	0.6950	0.6938	0.6909
Bag(SVM Linear)	0.7220	0.7215	0.7218
Bag(Ridge)	0.7116	0.7099	0.7123
Bag(LDA)	0.7242	0.7235	0.7235
Bag(QDA)	0.6108	0.5723	0.6066
Vot(Naïve Bayes, SVM, Ridge)	0.7253	0.7251	0.7230
Vot(SVM, Ridge, Logistic Regression)	0.7241	0.7237	0.7200
Vot(Ridge, Logistic Regression, LDA)	0.7229	0.7227	0.7187
Vot(Logistic Regression, LDA, Ada Boost)	0.7295	0.7297	0.7265
Vot(LDA, Ada Boost, QDA)	0.7426	0.7422	0.7372
Vot(Ada Boost, QDA, SVM RBF)	0.7579	0.7553	0.7516
Vot(QDA, SVM RBF, Gradient Boost)	0.7664	0.7629	0.7589

Stack(Naïve Bayes, SVM, Ridge)	0.7277	0.7263	0.7232
Stack(SVM, Ridge, Logistic Regression)	0.7293	0.7245	0.7221
Stack Ridge, Logistic Regression, LDA)	0.7238	0.7237	0.7191
Stack(Logistic Regression, LDA, Ada Boost)	0.7312	0.7322	0.7277
Stack(LDA, Ada Boost, QDA)	0.7434	0.7437	0.7388
Stack(Ada Boost, QDA, SVM RBF)	0.7589	0.7566	0.7529
Stack(QDA, SVM RBF, Gradient Boost)	0.7688	0.7635	0.75

Table 5 Results of Ensemble models of weak learners

7. Issues encountered and proposed solutions

- **Temporal Structure in Data :** In the analysis, the team observed that the dataset contained temporal information through the month variable. Initially, they attempted to use panel data techniques to address the temporal dependencies present in the data. However, these techniques yielded very low accuracies. Consequently, they decided to abandon the focus on temporal structure and chose to treat the month variable as an ordinary variable instead.
- **Class Imbalances: T**he team identified a class imbalance issue, with the majority of observations belonging to the Standard category. To address this imbalance in the credit score variable, they applied **SMOTE**
- **Handling Nominal Data:** In the analysis, the team identified that Occupation, Type_of_Loan, and Payment_Behaviour were the nominal variables. They decided to drop the Occupation variable because the model needed to generalize to occupations not present in the dataset. The Type_of_Loan variable had over 6000 categories. Noting that there were connections and an order among the words in these categories, they chose to encode it using a natural language processing technique called word embedding. For this purpose, they utilized the Sentence Transformer model "all-MiniLM-L6-v2." The Payment_Behaviour variable had only 6 categories, so they decided to apply one-hot encoding to this variable.
- **Computational efficiency:** The team observed extremely prolonged code execution times and time-consuming parameter tuning, even when using Google Colab with a T4 GPU. This slowdown was primarily attributed to the extensive number of observations in the dataset.

8. Discussion and Conclusion

Based on the results obtained from model training during the Advanced Analysis phase, several models stand out. However, after considering accuracy scores, the tuned XGB model is identified as the optimal choice.

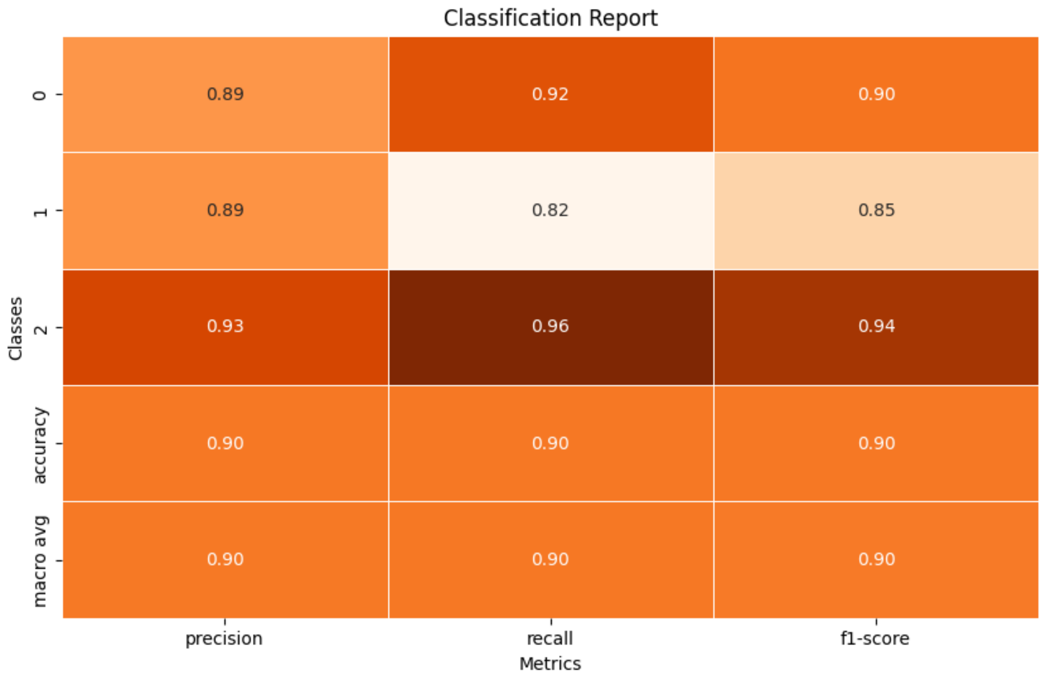


Figure 11 Classification Report for Best Model

According to the feature importance plot, it can be observed that variables such as Credit_History_Age, Credit_Utilization_Ratio, and Outstanding_Debt have the highest impact on the credit score class.

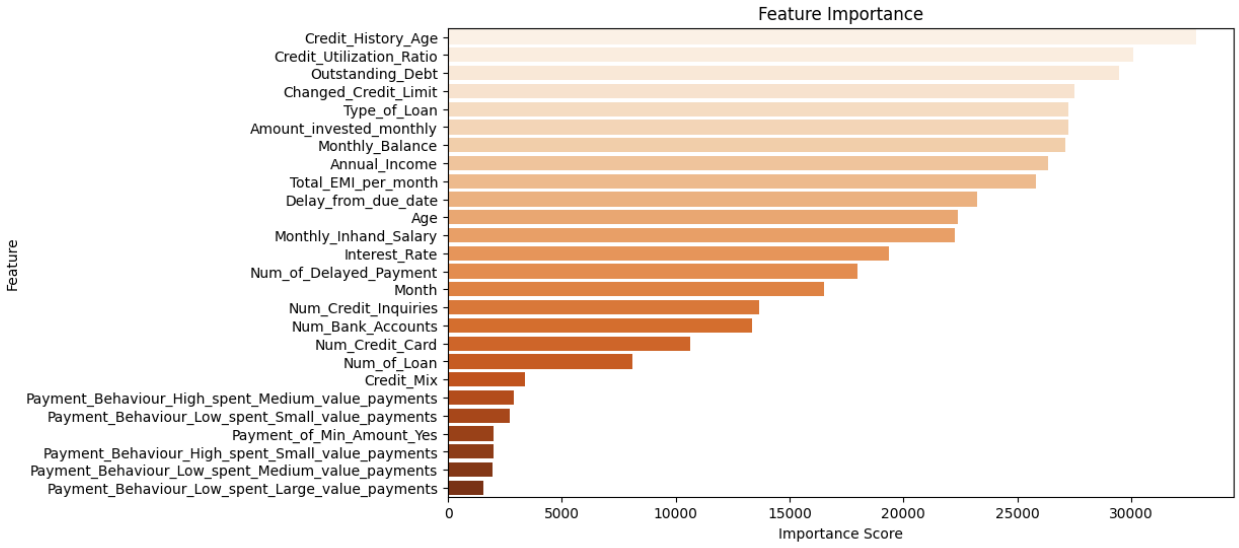


Figure 12 Feature Importance Plot

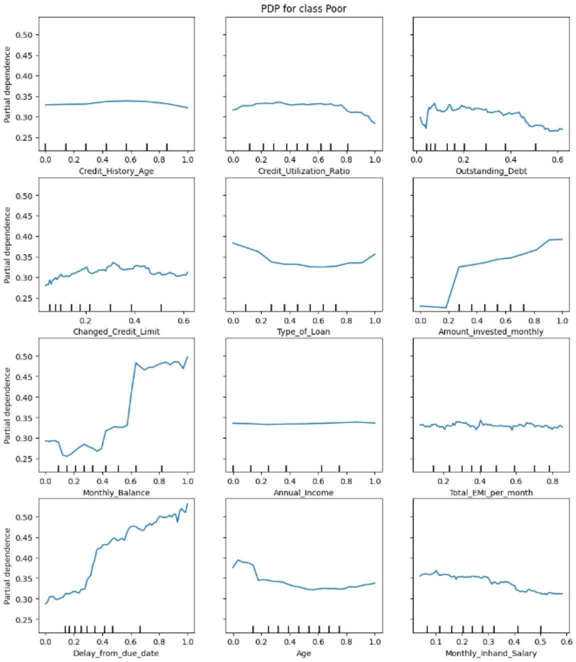


Figure 14 PDP Plot for Poor Class

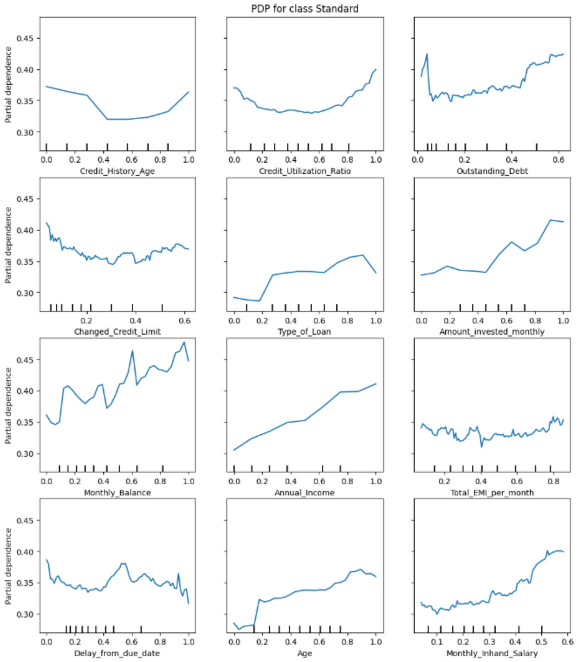


Figure 15 PDP Plot for Standard Class

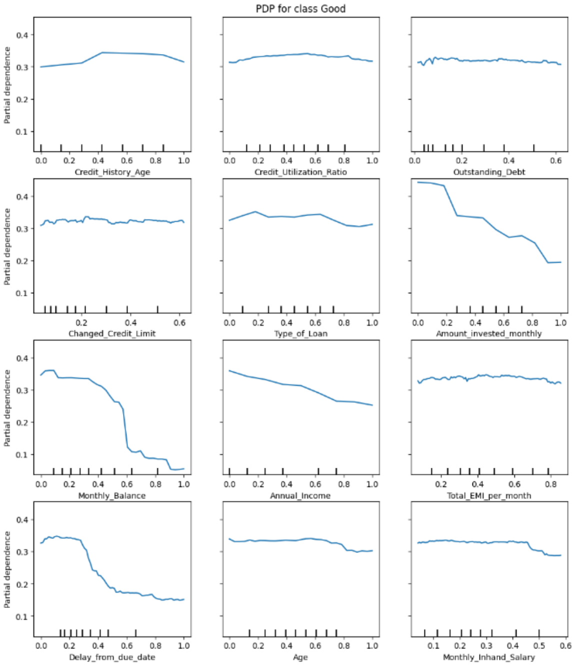


Figure 16 PDP Plot for Good Class

HIGH Monthly_Balance, Delay_from_due_date, could improve the probability of the credit score belonging to the POOR category.

LOW changed_credit_limit and low outstanding debt could increase the probability of the credit score belonging to the STANDARD category.

Along with low Monthly_Balance, Delay_from_due_date, the chances are high that the credit score will belong to the GOOD category.

9. Appendix

All codes and reports at: https://github.com/JanithRavinduRashmika/Credit_Score_Classification