ECSE 411

Semester project

Final Report

Student

Ilke Kas - ixk238

Supervisor

Evren Gurkan Cavusoglu

Teaching Assistant

Shroog Alshamari

Table of Contents

1. Abstract

This study aims to find the most effective and optimal statistical and machine learning model for classifying credit scores among linear regression, logistic regression, decision tree, and random forest. Evaluating their performances, the random forest model emerged as the most accurate, achieving an impressive 85% classification accuracy. Notably, linear regression, while the least suitable for classification, still attained a respectable 67% accuracy. In order to improve these models, feature importance and feature extraction method PCA (Principal Component Analysis) is applied to the models. However, given the underperformance of PCA-applied models, feature importance extracted from the random forest model showed better performance. Outstanding debt emerged as the most influential feature, followed by credit history age and interest rate. Additional moderately important features included monthly balance, delay from due date, and credit utilization ratio. This study also analyzes the relationship between the features in a given credit score dataset to understand the data better and find a strong positive linear relationship between monthly balance, annual income, and monthly in-hand salary, further enhancing our understanding of credit score determinants.

Through meticulous statistical analysis, this study not only identified the optimal classification model but also shed light on crucial features impacting credit score classification and their interrelations, providing valuable insights for financial decision-making.

2. Introduction and Background

Credit scoring indicates a prediction of the customer's credit behaviors [1]. How likely to pay debt on time, what amount of the available credit used, how many times one applied for any credit or loan can be given as an example to these credit behaviors [2]. Credit scoring is used by companies to decide whether to offer credit products such as credit card, auto loan, mortgage [1]. These companies also use the credit scores when they accept or deny the requests of these credit products made by the customers.

There are many factors that affect the credit score such as bill-paying history, current debt, credit inquiry number etc. However, there is not one way to calculate the credit score [1]. The most used two scoring models are FICO and Vantage scores [3]. However, these scores also can be varied for specific industries, the time they calculated and the data used [1][3]. Most of the time, the companies that develop these models do not reveal the details of these models since they make money from them [3].

This study uses and compares statistical and machine learning models to predict the credit score of the customers with given data. The data is from one of the challenges in Kaggle [4]. It is related to the finance field as mentioned. This dataset contains the basic bank information of the customers related to their credit-based information. The aim of the company that extracted this data is to create an intelligent system to classify the people according to their credit-scores to reduce manual effort to do that. This dataset has 20 features varying from age of the customer to the monthly balance of the customers and these customer's credit scores are grouped under three different classes: "Good", "Standard" and "Poor".
There are 3 goals of this study:

- **Main goal:** Find the best statistical model between these four: linear regression, logistic regression, decision tree and random forest to classify the credit scores of the customers
- Find the features that have more effect on the classification of the credit scores
- The relation between the features

These goals can be answered by using some statistical analysis methods. Firstly, for the main goal, the regression models are used as statistical models and to compare them with other models, decision trees and the random forest model are used. Besides that, since the dataset has many features, in order to learn the correlation between the features and which has the highest importance on the classification of the credit score prediction model, I used statistical analysis to

find the correlations and covariances. The feature importance is calculated after model training. In addition to these, principal component analysis is used when training the models and comparing the results. As known, principal component analysis is a statistical method and it is a linear dimensionality reduction technique, especially for the datasets chosen in this study (high number of features/ dimensions). The larger absolute value coefficient the more important the corresponding variable is in calculating the component (in terms of variance) [5]. Finally, the results found are shown visually by using box and scatter plots. The results of the models are shown by the confusion matrices. By analyzing the various statistical measures, this study checks whether the models trained are generalized enough to classify any customer's credit scores.

This study will be descriptive (for quantitative insights to large datasets as in this study), predictive (statistical models such as regression to predict future events) and exploratory (analyzing datasets through box plots, PCA etc.)

The goals of this project can be affected by some design options of the experiment. For the models used, the mathematical background of the models remain unchanged since the models used are from currently available libraries. However, these models can be initialized with different hyperparameters and these hyperparameters can affect the main goal of this study. These hyperparameters are used to tune the model in a way that it works in the intended way on the given data. Besides that, the study can be influenced by how the data is given to these models by using statistical analysis. For example, the number of used principal components before training the data can affect the goals or the number of important features used can affect the goals. Through this report, the design options chosen will be explained.

In this report, the experiment design, dataset, results and methodology will be explained. Firstly, In **section 3,** the dataset will be described by explaining the dependent and independent variables. In **section 4,** the design of the study, statistical methods used and the models will be explained, including the analysis of any potential sources of bias in the design. In **section 5,** the results of the statistical analysis with the interpretation will be given. Finally, in **section 6,** the report will be concluded.

3. The Sample Study

The dataset used in this study is from one of the challenges on Kaggle and belongs to the finance domain [4]. This dataset contains the basic bank information of the customers related to their credit-based information. The aim of the company that extracted this data is to create an intelligent system to classify the people according to their credit-scores to reduce manual effort to do that. This dataset has 20 features varying from age of the customer to the monthly balance of the customers. There are 99961 observations with 20 features and that makes this dataset size as 999961x21. To make sure that the dataset is clear, I double checked whether there is any duplicated data or not. I did not find any duplicated data, so there is 999961 unique data.

For this study, the features are assumed as the independent variables since by considering these features we are trying to find the credit score. Therefore, the credit score is considered as the dependent variable. Here is the list of features in the dataset:

- Average number of delayed days from due date
- Number of delayed payments
- Number of credit inquiries
- Credit utilization ratio
- Credit history age
- Amount invested monthly
- Monthly balance
- Age
- Annual income
- Number of bank accounts
- Number of credit cards
- Interest rate
- Number of loan
- Monthly inhand salary
- Changed credit limit
- Outstanding debt
- Total EMI per month
- Credit Mix
- Payment of minimum amount monthly
- Payment behavior

There are 3 categorical data in this dataset besides the Credit Score feature. All other features other than these three will have numerical value. Here is the categorical data: Payment of minimum amount can take values Yes, No and NM (Not Matter). Credit Mix can take values Good, Standard and Bad. Payment Behavior can take values “High spent Large value payments”, “High spent Medium value payments”, “High spent Small value payments”, “Low spent Large value payments”, “Low spent Medium value payments”, “Low spent Small value payments”.

Even though these features are considered as independent variables, there may be a relation between them. This will be analyzed in the results section. Before that, knowing the description of each variable in the finance domain plays an important role for the goals of this study and for a descriptive statistical analysis.
The descriptions of features (finance domain review):

- **Delay from due date** gives the average number of delayed days from the due date of the payments. It is known that if the late payment day is before 30 days, the lenders and the creditors may not report it to credit bureaus as a late payment [6]. These credit bureaus “collect information relevant to your credit and financial history” and calculate your credit scores according to the model they use [7]. However, late fee payments can be applied depending on the number of delays of the payment [7]. High number of delay days from due date may not be considered as good credit behavior and can cause lower credit score.
- **Number of Delayed Payments** gives the number of late payments for each customer. It is known that multiple late payments will have a negative impact on the credit score since it is not a good credit behavior [8].
- **Number of Credit Inquiries** refers to the number of requests to look at someone’s credit file from credit bureaus [9]. These can be either hard or soft inquiries [9]. Soft inquiries such as annual credit report requests and credit file reviews do not impact the credit scores while hard inquiries can affect the credit scores since they indicate the recent and frequent credit-seeking behaviors [9].
- **Credit Utilization Ratio** shows the amount of revolving credit used divided by the total credit available [10]. It is calculated by summing up the outstanding debt over all credits and then dividing it to the total credit available for the customer over all credit [10]. It is one of the methods that enables lenders to assess the ability to manage the debts of the customer [10]. It is generally preferred to be less than 30-40%.

- **Credit History Age** is calculated by taking the average of all accounts age in months [11]. The higher credit age is more preferable by the lenders since it is better to collect the data over long duration compared to the short duration [11].
- **The amount invested monthly** represents the monthly amount invested by the customer in USD. According to many sources, the amount invested monthly does not affect the credit score directly [12][13]. However, it can affect it indirectly [12]. For example, instead of paying the debts if one invests his/her income mostly, this may affect the credit score due to delayed/missed payments [12]. It is one of the skills that shows the budgeting skills of the individuals.
- **Monthly Balance** refers to the amount one needs to pay at the end of each month. Having a balance on your credit card at all times results in interest costs and increases your credit utilization ratio, which is taken into account when determining your credit scores [14].
- **The age** of the people does not affect the credit score of the individuals according to many sources [15]. However, there are statistics that show that the average credit score increases when the age increases due to the probability of having better credit history age increases [15].
- **Annual Income** refers to the total amount one earned in a year. According to external sources, it does not affect one's credit score directly [16]. However, it may affect it indirectly since the annual income can affect one's ability to pay their debts and loans [16].
- **Number of Bank Accounts** refers to the number of bank accounts owned by the individuals. According to Experian [17], the number of bank accounts does not affect the credit score except in one situation. If the balance on a checking account becomes negative and never paid, the bank may report it to the bureaus and therefore it can affect the credit score of the individuals [17].
- **The number of Credit Cards** may affect the credit score in many direct and indirect ways [18].
    - Firstly, it has an impact on the credit utilization ratio. Increasing the number of accounts means increasing the total credit available to individuals. That means decreasing credit utilization ratio, which is desirable by the lenders [10][18].
    - Secondly, every time one applies for a new credit or loan, the lender requests a new hard inquiry. As we mentioned before, a high number of hard inquiries is not a preferred situation for the credit score and by the lender.

    - Thirdly, overspending and difficulty managing payments, potentially harming your credit score due to charge offs, late payments, and high utilization rates [18].
- **Interest Rate** represents the interest rate of the credit card. While rising interest rates don't directly affect credit scores, they can have an impact on a number of the elements that do. For example, they could impact measures such as outstanding debt and the monthly balance for credit cards and loans [19].

- **The number of loans** can both help or hurt the credit score [20]:
    - It can boost the credit mix which is important for credit score. (help)
    - It establishes a positive payment history (help)
    - It can reduce the credit utilization ratio (help)
    - It requires hard inquiry (hurt)
    - It may get you deeper in debt (hurt)
    - Some additional fees such as late and origination fees (hurt)
- **Monthly Inhand Salary** is the net salary, which is derived from gross income after deductions like taxes, is the amount available for living expenses and other costs [21]. Like any kind of income, it does not directly affect the credit score but it can affect one's ability to pay their debts and loans.
- **Changed Credit Limit** refers to the percentage change in credit card limit. Changed credit limit may affect the credit score in two ways:
    - If the credit score is increased, it will affect the credit utilization ratio since it affects the total credit available.
    - If one requests the increase for the credit limit, it may hurt the credit score temporarily since the credit card issuers can request hard inquiry to verify the cardholder meets the requirements for the increase [22].
- **Outstanding Debt** refers to the total debt one owes to the creditors. Outstanding debt affects the credit score since it also affects the credit utilization ratio [23]. It is best to put one's outstanding debt 25-30% of their credit limits [23].
- **Total EMI Per Month** refers to the fixed monthly payments that borrowers make to lenders on a regular basis [24]. They cover principal as well as interest and guarantee full loan repayment over a predetermined period of time, one month in this scenario [24]. It affects the credit score since it affects the credit utilization ratio.
- **Credit Mix** is one of the categorical data. It refers to the type of different credit accounts one has such as mortgages, loans, credit cards, etc [25]. Even though the used credit score

calculations change the effect of credit mix on one's credit score, it is known that having a diverse mix of credit types may positively affect the credit score [25]. It can take three different values in this dataset: Credit Mix Bad, Credit Mix Good and Credit Mix Standard.

- **Payment of Minimum Amount** is one of the categorical data that exists in the dataset. Paying the minimum amount due each month guarantees on-time payment to prevent penalties; nevertheless, carrying a balance results in interest charges [26]. It emphasizes the benefit of paying the entire balance each month to completely avoid interest [26]. It can take three different values in this dataset: Payment of Min Amount No, Payment of Min Amount Yes and Payment of Min amount NM. Even though there is no explanation about what Payment of Min amount NM refers to in the dataset explanation in Kaggle, I guess it refers to "Not Matter".
- **Payment Behavior** is one of the categorical data that exists in the dataset. It can take six different values in this dataset and they are called:
    - High Spent Large Value Payments
    - High Spent Medium Value Payments
    - High Spent Small Value Payments
    - Low Spent Large Value Payments
    - Low Spent Medium Value Payments
    - Low Spent Small Value Payments

  Payment behavior affects the credit score directly.

As seen from the definitions, not all features are independent from each other. According to the definitions and articles I found, while some features directly affect the credit score, some might affect it indirectly.

4. Statistical Analysis

In order to complete the goals, I performed 5 different statistical analysis parts for this project. Firstly, I analyzed the dataset by calculating the mean, variance and standard deviation of the features. Secondly, I calculated the covariances and correlations between the features. Thirdly, in order to observe the characteristics of the data, I visualized the dataset by using histograms, box plots and scatter plots. Thanks to these three analysis parts, I gained important information about the data used such as their distributions and the later parts of the project gained insights about the features. After that, I performed some statistical analyses to answer the main goal of this project. I

performed some statistical and machine learning models such as linear and logistic regression, decision trees and random forest model on the data to predict the credit score of the customers and classify their credit score under three different categories: Good, Standard and Bad. However, since there are lots of features in this dataset, I also tried to train the models by using dimensionality reduction techniques. I used two different techniques for dimensionality reduction: Principal Component Analysis and Selecting Features according to their importance.

As mentioned in the sample study part, this dataset has some categorical data. In order to take these categorical data into consideration for both dimensionality reduction via PCA (Principal Component Analysis) and regression models, I converted these data into numerical values. Because of that, I used a one-hot encoding method. In one hot-encoding, for every value of the categorical feature, a new dummy variable will be created. For example, according to these features, the credit score of the customers are classified under 3 groups as poor, standard, and good. The one-hot encoding method creates three different variables for the credit score feature called “Credit_Score_Bad”, “Creedit_Score_Good” and “Credit_Score_Standard”. After creating these variables, it will give value 1 to the variable that corresponds to the value in the credit score feature. Other dummy variables will be assigned to 0. An example of the one-hot encoding of the credit score feature can be seen in **Table 1** below.

Credit_Score (Original Categorical Feature)	Credit_Score_Bad	Credit_Score_Good	Credit_Score_Standard
Bad	1	0	0
Standard	0	0	1
Good	0	1	0

Table 1

4.1. Statistical Analysis of Dataset

Statistical analysis of dataset can help us to gain insights about the characteristics of data such as their distributions, variability and shape. I calculated the mean, variance and standard deviation of each feature.

4.1.1. Sample Mean

As we are familiar from the lectures, mean value gives us the average of the data. It is calculated by summing up all of the values and dividing them to the number of observations as in the following equation (i).

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(i)}$$

The resulting mean values of the features in the dataset can be found in **Table 4** in Section 5.1.

4.1.2. Sample Variance

As we are familiar from the lectures, sample variance shows us dispersion of the data by calculating how far each data is from the mean. It is calculated as in the following equation (ii).

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}{n-1} \qquad \text{(ii)}$$

If the value of σ^2 is small, it shows less variability of the data. The resulting variance values of the features in the dataset can be found in **Table 4** in Section 5.1.

4.1.3. Standard Deviation

Standard deviation is calculated by taking the square root of the variance as in equation (iii). It shows how dispersed the data is from the mean. Higher standard deviation indicates that data has more sparse distribution compared to the lower standard deviation. Lower standard deviations mean the data is tightly clustered around the mean and the distribution is not so dispersed.

$$\sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}{n-1}} \qquad \text{(iii)}$$

The resulting standard deviation of the features in the dataset can be found in **Table 4** in Section 5.1.

4.2. Statistical Analysis for Relation Between Features

In order to understand the relationship between features, I calculated the covariance and correlations of the features by using cov() and corr() functions of the pandas library. One can see the calculations and results of the correlations and covariances in the corresponding Colab file given in reference [27]. Both of these measures show the linear relationship between the features.

4.2.1. Covariances

Covariance measures the relationship and the dependency between two variables. It is calculated as in equation (iv).

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_i)\ (y_i-\bar{y}_i)}{n-1} \quad \text{(iv)}$$

It is important to note that the covariance only measures the direction between two variables. The strength of their dependency or the relationship is not calculated by the covariance since the values are not scaled. Therefore, the covariance values can be so big or low and this does not show us the strength of their relation.

4.2.2. Correlations

Correlation measures both the direction and the strength of the relationship and the dependency between two variables. It is calculated as in equation (v).

$$corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_i)\ (y_i-\bar{y}_i)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2(y_i-\bar{y}_i)^2}} \quad \text{(iv)}$$

As one can notice, the correlation is scaled according to the values of the variables. Therefore, in addition to the direction of the relation between two variables, it can also give the strength of the relationship. The correlation of two variables can take values between -1 and 1. -1 shows the perfect negative linear

relation between two variables while 1 shows perfect positive linear relation between variables.

4.3. Visualization of Data

4.3.1. Histograms

Histograms are used in this project to show the number of occurrences of one data value for a specific feature. The y-axis shows the number of occurrences of the data value, while x-axis represents the data values that feature can take. It is especially used to visualize categorical variables for this study. It is important for the dataset to be balanced in order to classify any data with high accuracy. Otherwise, the statistical or the machine learning models have a tendency to classify the test data as the most occurring class. Histograms can help us to decide whether this dataset is balanced or not in terms of the labels. In addition to this, it can help us to visualize the distribution of the data. The resulting histograms can be found under Section 5.3.1 for Figures between 2-6.

4.3.2. Box Plots

Box plots are used to compare the distribution of the different data groups in general. In this study, box plots are used to compare the distributions of features in three different credit score classes “Good”, “Standard” and “Bad”. In this way, we can analyze whether the value of any feature can affect the credit score class significantly by looking at the box plots. We can interpret some feature importance results by looking at these plots. In order to draw the boxplots, I used the boxplot function of the pandas library. The resulting box plots can be found under Section 5.3.2 for Figures between 7-9.

4.3.3. Scatter Plots

Scatter plots are used to observe the relationship between two numeric variables. In this study, scatter plots are used to visualize the relationship between variables who have strong correlation values. In order to draw the scatter plots, I used the

plot.scatter function of the pandas library. The resulting scatter plots can be found under Section 5.3.3 for Figures between 10-12.

4.4. Statistical Analysis for Main Goal

4.4.1. Linear Regression

Linear regression tries to find an optimal line between the dependent and independent variables by assuming that there is a linear relationship between them [28]. While seeking the optimal line, the linear regression uses the sum of the squared distances between predicted and actual values and tries to minimize it [28].

If we formalize this where y is the dependent variable, α is constant, β is the slope of the feature and x is the independent feature variable and ε_i is the error term.

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{(v) [28]}$$

The aim of the cost function of linear regression is to minimize the square of the error term.The cost function of the Linear Regression function provided by sklearn library is calculated as in equation by using Ordinary Least Square Model [29] (vi and vii).

$$\widehat{\alpha} = min_{\alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = min_{\alpha} \sum_{i=1}^{n} (\varepsilon_i)^2 \quad \text{(vi) [29]}$$

$$\widehat{\beta} = min_{\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = min_{\beta} \sum_{i=1}^{n} (\varepsilon_i)^2 \quad \text{(vii) [29]}$$

The optimization is calculated both for term $\widehat{\alpha}$ and $\widehat{\beta}$ by taking the derivative of equation in (viii) in terms of both $\widehat{\alpha}$ and $\widehat{\beta}$ and make the derivatives equal to 0 before solving them.

The linear regression is generally used to predict the continuous values from the given data. It is not generally used for classification. However, in this study, I tried to use it for classification of the data for three different categories. Firstly, I used the LinearRegression function of the sklearn library and fit the model to train on the train dataset. In this train dataset, the classes for classification can take values 0 or 1 since I applied one-hot-encoding as mentioned before. After fit

the model, I predict the continuous values for the classes “Credit_Score_Good”, “Credit_Score_Poor” and “Credit_Score_Standard” of each data. However, since the linear regression model gives continuous values for the predictions, I needed to classify these continuous values. Therefore, for each data, I took the maximum predicted values between these three classes. Then, classify the data as it belongs to that class. For example, let’s assume that the result of linear regression for data as in the follows:

	Credit_Score_Goood	Credit_Score_Poor	Credit_Score_Standard
Data	0.052	0.57	0.37

Table 2

I classified the data given in **Table 2** in a way that it belongs to the class whose result of linear regression is the maximum. For example for the data in **Table 2,** the classification result will be as in **Table 3.**

	Credit_Score_Goood	Credit_Score_Poor	Credit_Score_Standard
Data	0	1	0

Table 3

It may not be the best way to classify the data like this. However, in order to classify the result of a linear regression, an additional way like this is necessary.

The resulting coefficients for the linear regression of this dataset can be seen in **Figure 13.** The resulting prediction scores and metrics for the linear regression can also be seen in **Figure 13.**

4.4.2. Logistic Regression [30][31][32]

Logistic regression is generally used for classification by finding a relationship between features and the probability that the data belongs to any class by looking at this relationship. In contrast to the linear regression which is used to predict a linear outcome, the logistic regression uses logistic functions to find the

probability of the outcome. First, let's look at how logistic regression computes its output. The logistic function used is a sigmoid curve whose value can be in range from 0 to 1 as seen in **Figure 1.** The function of the sigmoid curve is given in (ix).

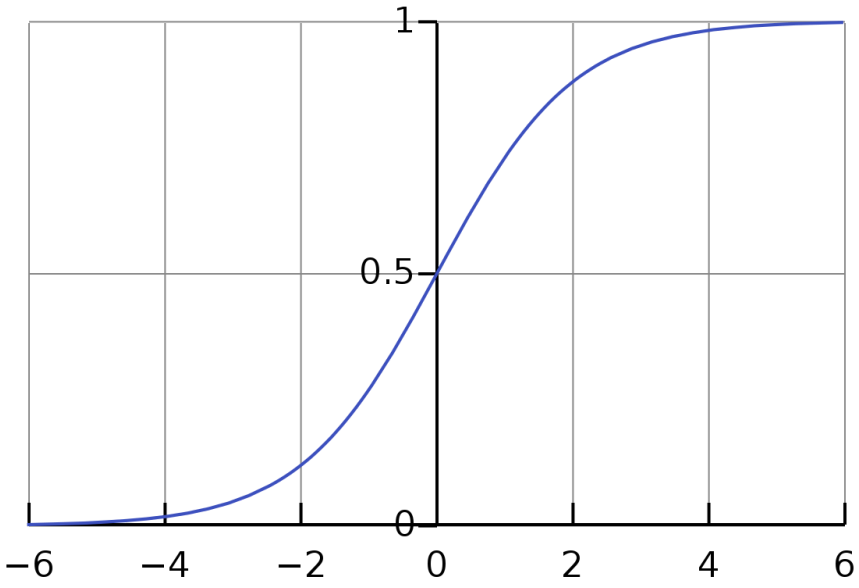


Figure 1 [33]

$$S(x) = \frac{1}{1+e^{-x}} \quad \textbf{(ix)}$$

As you can see from **Figure 1** too, the sigmoid function will result between 0 and 1 for the value of the given x. Logistic regression using the linear regression as the input value (ix). So, when linear regression calculate the output of the data by given $x_1, x_2, \ldots, x_3$ as the features and $b_1, b_2, \ldots, b_3$ as the coefficients and a as the constant term:

$$y = a + b_1 . x_1 + b_2 . x_2 + \ldots + b_n . b_n$$

Therefore, the logistic regression function calculates its outputs as in equation(x).

$$P(y = 1|x_1, x_2, \ldots, x_3) = \frac{1}{1+e^{-(a+b_1 . x_1 + b_2 x_2 + \ldots + b_n b_n)}} \quad \textbf{(x)}$$

Secondly, how we will compute the coefficient values $b_1, b_2, \ldots, b_3$ in logistic regression? In order to solve this, logistic regression is using the Maximum Likelihood Method. Initially let's think about the binomial classification case. Logistic regression aims to maximize the log likelihood of the samples labeled as 1 in a way that the probability is as close to 1 as possible while for the other samples (labeled as 0), it wants to maximize the log likelihood of them in a way that 1-probability is as close to 0 [30]. We can formalize this in the following equation (xi):

$$l(\beta) = \sum_{i=1}^{n} (y_i).\, log(p(x_i)) + (1 - y_i).\, log(1 - p(x_i)) \quad \text{(xi)}$$

For multinomial classification with logistic regression, we can generalize this function as in the following equation (xii) where n shows the number of samples and k shows the number of classes.

$$l(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} (y_{ik}).\, log(p(x_{ik})) \quad \text{(xii)}$$

Instead of maximizing this function, multinomial logistic regression function of the sklearn library minimizes the cross-entropy loss which can be understood by comparing the equation with (xii)[34]:

$$\text{MCE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (y_{ik}).\, log(p(x_{ik})) \text{ (xiii)}$$

The optimizer used in the logistic regression tries to minimize this loss function. In this study, I tried all of the optimizers used for multinomial classification of the logistic regression [34]. However, I got the best result with the "newton-cg" solver, probably because it has better parameter updates thanks to second-order derivation utilization and quadratic approximation [34]. The resulting scores and prediction metrics of the logistic regression can be seen in **Section 5.4, Figure 13**.

4.4.3. Decision Trees

Decision trees are another supervised learning technique used to classify the data. They are easy to interpret since it has a similar way of thinking with human

beings. Besides that, in addition to the numerical data, it can also take categorical data into consideration without preprocessing such as one-hot-encoding. It consists of root, decision and leaf nodes. Decision nodes are the nodes that can be splitted while leaf nodes cannot be splitted. The decision tree can be build by calculating the entropies of the branches. Firstly, the gini impurity of each feature is calculated to select the root node of the tree. Gini index is a measure of impurity calculated by the following formula:

$$Gini = \sum_{i=1}^{n} (p_i)^2 \text{ (xiv) [35]}$$

Where pi shows the probability of a particular element of being a member of a specific class. Other impurity measures can also be used. However, for this study, I used the default gini index. After the weighted average of the gini impurities of each child nodes for each feature is calculated, the feature with the lowest weighted gini impurity is selected as the decision node. The tree is splitting recursively like that until one of the nodes has 0 gini impurity or the tree reaches the maximum depth defined by the user. In order to avoid overfitting, choosing the or restricting the max_depth variable is generally important for the decision trees. For this study, I tried different max_depths for the decision trees such as 10,15 and 20. However, I got the best result when I used the default number of max_depth which is the number that until all leaves are pure. The prediction results and metrics results can be found in then **Section 5.4, Figure 13**.

4.4.4. Random Forest

Random Forest uses many decision trees to make more robust decisions. For the random forest model, some data points and features are used to construct the trees [36]. After each decision tree generates their outputs, by majority voting or the averaging, the output of the classification is selected [36]. I did not restrict the number of trees in the forest. It uses the default which is 100 trees. The prediction results and metrics results can be found in then **Section 5.4, Figure 13**.

4.5. Statistical Analysis for Dimensionality Reduction and Feature Importance

4.5.1. Principal Component Analysis (PCA)

Principal component analysis is a statistical method and it is a linear dimensionality reduction technique, especially for the datasets like I chose (high number of features/ dimensions). PCA algorithm applies these step by step to reduce extract the principal components [37]:

1. It standardizes the data
2. It calculates the covariance matrix of the standardized data
3. It calculates the eigenvectors and eigenvalues of the covariance matrix extracted. While eigenvectors indicate the directions in which the data vary the most, eigenvalues indicate the amount of variation along each eigenvector.
4. The eigenvectors with highest eigenvalues are chosen as the principal components and data are reformed according to these principal components.

In order to calculate the principal components, I used the sklearn library and used the first four principal components for the models with feature reduction. Finding the principal components can be useful for the dataset with many features since it helps reducts the dimensionality of the dataset. It can decrease the computational time significantly and sometimes can result in better performance.

One can find the resulting principal component analysis in **Section 5.5.1.1, Figure 14**. Besides that, the models that use the extracted features by using the principal components can also be found in n **Section 5.5.3, Figure 17**.

.

4.5.2. Feature Importance

In addition to Principal Component Analysis, another most commonly used way to extract the feature importance is to use the results of the random forest model. The importance factor is given here is the impurity based as mentioned before [38]. Even though it may be misleading for features that have many unique values, it can still provide us an idea for the importance of the features [38]. Finding the feature importance

can be useful for the dataset with many features. It can decrease the computational time significantly and sometimes can result in better performance.

One can find the resulting feature importance in section xx and figure xx. Besides that, the models that use the extracted features by using the feature importance can also be found in **Section 5.5.2.1, Figure 15-16**.

5. Results and Discussions of Findings

5.1. Results of Statistical Analysis of Dataset

In order to understand the distribution of the dataset and gain insights into data's variability for each feature, the mean, variance and the standard deviation of the dataset is calculated. As seen in **Table 1**, the mean, variance and the standard deviation of the features can be observed.

	mean	std	var
Delay_from_due_date	21.095718287314900	14.827413836428500	219.85220107671100
Num_of_Delayed_Payment	13.336344537815100	6.269963772278150	39.3124457056805
Num_Credit_Inquiries	5.774569827931170	3.8622244165478000	14.916777443778000
Credit_Utilization_Ratio	32.28454382959670	5.116888311134370	26.182545988623500
Credit_History_Age	221.12281912765100	99.6960835149188	9939.309068213660
Amount_invested_monthly	193.66521850787900	194.782737379757	37940.31478115140
Monthly_Balance	403.44509967154800	214.38709150239500	45961.825002856400
Age	33.26922769107640	10.76237126214310	115.82863518420300
Annual_Income	50498.704152861100	38294.24316119290	1466449059.2885700
Num_Bank_Accounts	5.368867547018810	2.5916683861024500	6.716745023522870
Num_Credit_Card	5.532853141256500	2.067697755036670	4.2753740061836700
Interest_Rate	14.535174069627900	8.741047403068470	76.40590970269000
Num_of_Loan	3.5339735894357700	2.4461560795510600	5.9836795655246300
Monthly_Inhand_Salary	4196.814288236280	3186.5181631114100	10153898.003838900
Changed_Credit_Limit	10.39695318127250	6.510846252782280	42.39111892736910
Outstanding_Debt	1426.5149659863900	1155.2525038176900	1334608.3475770500
Total_EMI_per_month	105.56580228082600	125.8209830637110	15830.919779118500

Credit_Mix_Bad	0.2377751100440180	0.42572282087295300	0.18123992021202500
Credit_Mix_Good	0.3037214885954380	0.45986613440834200	0.2114768615756710
Credit_Mix_Standard	0.45850340136054400	0.49827754926349700	0.24828051610003600
Payment_of_Min_Amount_NM	0.12010804321728700	0.3250894621889290	0.10568315842628700
Payment_of_Min_Amount_No	0.35642256902761100	0.47894448124124800	0.22938781611144800
Payment_of_Min_Amount_Yes	0.523469387755102	0.49945138237213900	0.24945168335344000
Payment_Behaviour_High_spent_Large_value_payments	0.16750700280112000	0.3734297817123680	0.13944980186974700
Payment_Behaviour_High_spent_Medium_value_payments	0.26738695478191300	0.4425981596319210	0.19589313090956300
Payment_Behaviour_High_spent_Small_value_payments	0.06514605842336940	0.2467846404571490	0.06090265876556410
Payment_Behaviour_Low_spent_Large_value_payments	0.057623049219687900	0.233030420047211	0.054303176667379600
Payment_Behaviour_Low_spent_Medium_value_payments	0.1030812324929970	0.3040664679553880	0.09245641693486490
Payment_Behaviour_Low_spent_Small_value_payments	0.33925570228091200	0.4734590935687510	0.22416351328294300

Table 4

As seen from **Table 4**, categorical data have smaller means, variances and standard deviation compared to the other features since one-hot-encoding is applied to them. This means that the categorical data can only take value 0 or 1. Therefore, their mean, standard deviation and the variance are in between 0 and 1, as seen from the red part of **Table 4**.

When we look at the highest value of these values belongs to the Annual Income feature (in green row). Monthly Inhand Salary and Outstanding Debt features are also higher values of mean, standard deviation and variance compared to other features. However, without understanding of the meanings of these features, the mean, standard variation and variance quantities are hard to interpret. Therefore, it is important to grasp the meaning of the terms used in this dataset first.

Interpretation of Some Features' Mean, Variance and Standard Deviation Values:

- On average, payments are delayed 21 days with high standard deviation 14.82. This high standard deviation shows significant variability in payment delays. This kind of variability is expected among 999961 people.

- The individuals in this dataset have 13 delayed payments on average with a moderate standard deviation 6.27. Even though the variation seems high (39.3), I was expecting higher variability for 999961 people. This moderate variability shows consistency among the individuals.

- In this dataset, even though there is no explicit information which type of inquiries are considered (soft, hard or both), I assumed that the credit inquiries taken into account are hard inquiries since the mean is 5.77 and the standard deviation and the mean are not high values, approximately 4 and 15, respectively.

- In this dataset, the average credit utilization ratio of the individuals is 32 with standard deviation 5.11. I can say that the mean credit utilization ratio is pretty preferable for the lenders.

- The average credit history age is 221 months for this population with high variance 9939 and high standard deviation.

- According to this dataset, the monthly amount invested is 193.66 USD while it has high variance (194.78) and standard deviation (37940). This variability for this feature is expected due to the diverse investment behavior in the population.

- The average age of an individual in this dataset is 33 and has predictable standard deviation.

- As mentioned before, annual income has the highest mean, variance and standard deviation compared to the other features in the dataset.

- For this dataset, the average number of accounts is 5.36 white the standard deviation is 2.59. It is moderately small.

- Average number of loans for this dataset is 3.53 while standard deviation and variance are 2.44 and 6, respectively. Therefore, saying that there is a low variability in the number of loans for this dataset would not be wrong.

- Average monthly in hand salary for this dataset is 4189 USD. It has higher variance and standard deviation compared to the other features except the annual income and monthly balance.

- In this dataset, the changed credit limit has mean 10, standard deviation 6.51 and variance is 42 approximately.

- The average outstanding debt for individuals in this dataset is approximately 1427. It is a pretty moderate mean value, while the standard deviation from the mean and the variance have high values, 1155 and 1334608 USD, respectively. However, it is normal for that kind of high variability for 999961 people.

- In this dataset, the average value for total EMI per month is 106 approximately. The standard deviation and variance values approximately are 126 and 15830. I think this is a high variation but expected among 999961 people.

- Credit Mix is one of the categorical data that exists in the dataset. However, to perform PCA and machine learning algorithms on the categorical data, as I mentioned, I applied one hot encoding. This results in three different variables in this dataset called: Credit Mix Bad, Credit Mix Good and Credit Mix Standard. Here are the values for credit mix from Table 1:

	Mean	Standard Variation	Variance
Credit_Mix_Bad	0.2377751100440180	0.4257228208729530 0	0.18123992021202500
Credit_Mix_Good	0.3037214885954380	0.4598661344083420 0	0.2114768615756710
Credit_Mix_Standard	0.45850340136054400	0.4982775492634970 0	0.24828051610003600

Table 5

According to these values, it is easy to notice that the most frequent types of credit mix is standard since it has the highest mean value. We can also see that fact by looking at the histogram values of the credit mix feature in **Figure 2**. After the standard credit mix, the good credit mix is more frequent than the bad credit mix according to the mean values. However, since the data is binary here (can take values 0 or 1), the standard deviation and variance might not provide particularly meaningful insights in this case.

- Payment of Minimum Amount is one of the categorical data that exists in the dataset. However, to perform PCA and machine learning algorithms on the categorical data, as I mentioned, I applied one hot encoding. This results in three different variables in this dataset called: Payment of Min Amount No, Payment of Min Amount Yes and Payment of Min amount NM. Even though there is no explanation about what Payment of Min amount NM refers to in the dataset explanation in

Kaggle, I guess it refers to “Not Matter”. Here are the values for payment of minimum amount from **Table 4**:

	Mean	Standard Deviation	Variance
Payment_of_Min_Amount_NM	0.12010804321728700	0.3250894621889290	0.10568315842628700
Payment_of_Min_Amount_No	0.35642256902761100	0.47894448124124800	0.22938781611144800
Payment_of_Min_Amount_Yes	0.523469387755102	0.49945138237213900	0.24945168335344000

Table 6

According to these values, it is easy to notice that most of the individuals are paying the minimum amount since it has the highest mean value (Payment_of_Min_Amount_Yes). We can also see that fact by looking at the histogram values of the payment of Minimum Amount feature in **Figure 3**. However, since the data is binary here (can take values 0 or 1), the standard deviation and variance might not provide particularly meaningful insights in this case.

- Payment Behavior is one of the categorical data that exists in the dataset. However, to perform PCA and machine learning algorithms on the categorical data, as I mentioned, I applied one hot encoding. This results in six different variables in this dataset called:
    - High Spent Large Value Payments
    - High Spent Medium Value Payments
    - High Spent Small Value Payments
    - Low Spent Large Value Payments
    - Low Spent Medium Value Payments
    - Low Spent Small Value Payments

Payment behavior affects the credit score directly .Here are the values for payment behavior from **Table 4**:

	Mean	Standard Deviation	Variances
Payment_Behaviour_High_spent_Large_value_payments	0.16750700280112000	0.3734297817123680	0.13944980186974700

Payment_Behaviour_High_spent_Medium_value_payments	0.26738695478191300	0.4425981596319210	0.19589313090956300
Payment_Behaviour_High_spent_Small_value_payments	0.06514605842336940	0.2467846404571490	0.06090265876556410
Payment_Behaviour_Low_spent_Large_value_payments	0.057623049219687900	0.233030420047211	0.054303176667379600
Payment_Behaviour_Low_spent_Medium_value_payments	0.1030812324929970	0.3040664679553880	0.09245641693486490
Payment_Behaviour_Low_spent_Small_value_payments	0.33925570228091200	0.4734590935687510	0.22416351328294300

Table 7

According to these values, it is easy to notice that most of the individuals are spending high and making medium value payments since it has the highest mean value (Payment_Behaviour_High_Spent_ Medium_value_Payments). We can also see that fact by looking at the histogram values of the payment behavior feature in **Figure 4**. However, since the data is binary here (can take values 0 or 1), the standard deviation and variance might not provide particularly meaningful insights in this case.

Now, let's look at the covariances and correlations between these features to understand the relationship between features.

5.2. Results of Statistical Analysis for Relation Between Features

5.2.1. Covariances

The one part of the calculated covariances are given below. Since the spreadsheet is too big, one can also find the whole table in the spreadsheet given in reference [39]. The **Table 8** given below shows only part of the covariances.

	Delay_from_due_date	**Num_of_Delayed_Payme**	**Num_Credit_Inquiries**	**Credit_Utilization_Ratio**	**Credit_History_Age**	**Amount_invested_monthly**

		nt				
Delay_from_due_date	219.8522011	50.32818311	30.96914651	-4.846468808	-726.1248825	-468.9982826
Num_of_Delayed_Payment	50.32818311	39.31244571	12.11323831	-2.364894909	-301.0624686	-222.4113818
Num_Credit_Inquiries	30.96914651	12.11323831	14.91677744	-1.524644349	-235.1501455	-132.7640573
Credit_Utilization_Ratio	-4.846468808	-2.364894909	-1.524644349	26.18254599	37.3490125	-5.884157619
Credit_History_Age	-726.1248825	-301.0624686	-235.1501455	37.3490125	9939.309068	3412.679118
Amount_invested_monthly	-468.9982826	-222.4113818	-132.7640573	-5.884157619	3412.679118	37940.31478
Monthly_Balance	-895.9315892	-409.4833143	-266.7919602	268.8739568	6963.930214	1949.741422
Age	-27.76800459	-12.38861637	-10.53677231	1.397086539	251.3950089	119.8870954
Annual_Income	-141949.1692	-68447.17816	-41531.43788	34455.37143	1039940.694	4680039.06
Num_Bank_Accounts	21.51394747	9.755282708	5.196811306	-0.9508437387	-125.3918565	-91.94716711
Num_Credit_Card	14.69237048	5.484288003	3.671768552	-0.5864133257	-85.96584438	-57.69734543
Interest_Rate	76.32858927	31.27160048	21.3845838	-3.382573509	-502.1579481	-331.794185
Num_of_Loan	18.18262428	7.275485689	5.346567701	-1.256238401	-147.7402494	-78.33289607
Monthly_Inhand_Salary	-11778.75415	-5674.549794	-3446.445262	2869.947584	86245.1774	389832.7601
Changed_Credit_Limit	29.63152314	13.88991163	9.960881591	-1.685420471	-287.7445066	-152.7605216
Outstanding_Debt	9797.170707	3654.833891	2667.384502	-420.4240772	-72482.32462	-39115.89154
Total_EMI_per_month	168.716373	48.83649005	50.50333245	13.35814875	-1575.667117	6982.670725
Credit_Mix_Bad	4.143528524	1.57936517	0.9759103665	-0.1494725572	-24.83381351	-13.67055354
Credit_Mix_Good	-3.345661793	-1.896691845	-0.9022593261	0.1884707912	22.06019679	17.55866654
Credit_Mix_Standard	-0.7978667308	0.3173266754	-0.07365104046	-0.03899823393	2.77361672	-3.888113004
Payment_of_Min_Amount_NM	0.005730411894	0.002569528297	0.002616219108	-0.004519184425	0.1116501121	0.08493698552

Payment_of_Min_Amount_No	-3.25733802	-1.609882884	-1.05651691	0.1861637137	25.40905972	16.61512512
Payment_of_Min_Amount_Yes	3.251607608	1.607313356	1.053900691	-0.1816445293	-25.52070983	-16.7000621
Payment_Behaviour_High_spent_Large_value_payments	-0.5772237318	-0.2691178677	-0.1678027711	0.1667220562	4.311923054	13.11199419

Table 8

As we know from our lectures, since covariances are not standardized they are scale dependent. Therefore, we cannot assess the direct relationship between two variables by looking at covariance values only. We can assess the direction of their relationship though. The reason behind this is the fact that these features can take larger values compared to other ones and covariance does not scale it. Besides that, it is not easy to analyze the relationship between independent and dependent variables by looking at the large tables above. Therefore, I grouped the positive and negative covariances for each feature. These values can be seen in **Table 9** [40].

	Positive Cov	Negative Cov

Delay_from_due_date	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Num_of_Delayed_Payment	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments

		Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good
Num_Credit_Inquiries	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard

Credit_Utilization_Ratio	Credit_Utilization_Ratio Credit_History_Age Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Amount_invested_monthly Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Credit_History_Age	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	
Amount_invested_monthly	Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Credit_Utilization_Ratio Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard
Monthly_Balance	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Age	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Credit_Score_Good Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Medium_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Annual_Income	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Num_Bank_Accounts	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good
Num_Credit_Card	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments

	Credit_Score_Poor	Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Interest_Rate	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard

Num_of_Loan	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Monthly_Inhand_Salary	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Changed_Credit_Limit	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good
Outstanding_Debt	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments

		Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Total_EMI_per_month	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Credit_Utilization_Ratio Amount_invested_monthly Monthly_Balance Annual_Income Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Monthly_Inhand_Salary Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Poor	Credit_History_Age Age Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Standard

Credit_Mix_Bad	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Credit_Mix_Good	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Large_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Small_value_payments

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good	Credit_Score_Poor Credit_Score_Standard
Credit_Mix_Standard	Num_of_Delayed_Payment Credit_History_Age Num_Bank_Accounts Interest_Rate Changed_Credit_Limit Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Standard	Delay_from_due_date Num_Credit_Inquiries Credit_Utilization_Ratio Amount_invested_monthly Monthly_Balance Age Annual_Income Num_Credit_Card Num_of_Loan Monthly_Inhand_Salary Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Poor
Payment_of_Min_Amount_NM	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Credit_History_Age Amount_invested_monthly Monthly_Balance Annual_Income Interest_Rate Monthly_Inhand_Salary Outstanding_Debt Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_Behaviour_High_spent_Medium_value_payments	Credit_Utilization_Ratio Age Num_Bank_Accounts Num_Credit_Card Num_of_Loan Changed_Credit_Limit Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_No Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Small_value_payments

	Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Poor	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Standard
Payment_of_Min_Amount_No	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard
Payment_of_Min_Amount_Yes	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments

	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard	Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good
Payment_Behaviour_High_spent_Large_value_payments	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Credit_Score_Good Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor

Payment_Behaviour_High_spent_Medium_value_payments	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor
Payment_Behaviour_High_spent_Small_value_payments	Credit_History_Age Age Changed_Credit_Limit Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Small_value_payments Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Credit_Utilization_Ratio Amount_invested_monthly Monthly_Balance Annual_Income Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Monthly_Inhand_Salary Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes

		Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Good Credit_Score_Poor
Payment_Behaviour_Low_spent_Large_value_payments	Credit_History_Age Amount_invested_monthly Age Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_Low_spent_Large_value_payments Credit_Score_Standard	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Credit_Utilization_Ratio Monthly_Balance Annual_Income Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Monthly_Inhand_Salary Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Medium_value_payments

		Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Good Credit_Score_Poor
Payment_Behaviour_Low_spent_Medium_value_payments	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Poor	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Age Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Standard

Payment_Behaviour_Low_spent_Small_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good Credit_Score_Standard
Credit_Score_Good	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Total_EMI_per_month Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Credit_Mix_Bad Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments

	Payment_Behaviour_Low_spent_Medium_value_payments Credit_Score_Good	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor Credit_Score_Standard
Credit_Score_Poor	Delay_from_due_date Num_of_Delayed_Payment Num_Credit_Inquiries Num_Bank_Accounts Num_Credit_Card Interest_Rate Num_of_Loan Changed_Credit_Limit Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Payment_of_Min_Amount_NM Payment_of_Min_Amount_Yes Payment_Behaviour_Low_spent_Medium_value_payments Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Poor	Credit_Utilization_Ratio Credit_History_Age Amount_invested_monthly Monthly_Balance Age Annual_Income Monthly_Inhand_Salary Credit_Mix_Good Credit_Mix_Standard Payment_of_Min_Amount_No Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Credit_Score_Good Credit_Score_Standard
Credit_Score_Standard	Num_of_Delayed_Payment Credit_Utilization_Ratio Credit_History_Age Monthly_Balance Age Annual_Income Num_Bank_Accounts Monthly_Inhand_Salary Changed_Credit_Limit Credit_Mix_Standard Payment_of_Min_Amount_Yes Payment_Behaviour_High_spent_Large_value_payments Payment_Behaviour_High_spent_Medium_value_payments	Delay_from_due_date Num_Credit_Inquiries Amount_invested_monthly Num_Credit_Card Interest_Rate Num_of_Loan Outstanding_Debt Total_EMI_per_month Credit_Mix_Bad Credit_Mix_Good Payment_of_Min_Amount_NM Payment_of_Min_Amount_No Payment_Behaviour_Low_spent_Medium_value_payments

	Payment_Behaviour_High_spent_Small_value_payments Payment_Behaviour_Low_spent_Large_value_payments Credit_Score_Standard	Payment_Behaviour_Low_spent_Small_value_payments Credit_Score_Good Credit_Score_Poor

Table 9

Negative covariance values implies that the higher values in one variable tend to correspond to lower values in another variable while positive covariance values implies higher covariance value in one variable tends to correspond to higher values. Let's analyze the covariance values between the features and credit scores (the green part of table 9). As one can see, the good credit score values correspond to the higher values of Credit Utilization Ratio, Credit History Age, Amount invested monthly, Monthly Balance, Age, Annual Income, Monthly Inhand Salary, Total EMI per month, Credit Mix Good, Payment of Min Amount NM, Payment of Min Amount No, Payment Behaviour High spent Large value payments, Payment Behaviour High spent Medium value payments, Payment Behaviour Low spent Medium value payments. All of these match with our knowledge from the finance domain except credit utilization ratio and payment of minimum amount. As we know, high credit utilization ratio hurts one's credit score.

When we look at the poor credit score covariance, the poor credit score values correspond to higher values of Delay from due date, Number of Delayed Payment Num Credit Inquiries, Num Bank Accounts, Num Credit Card, Interest Rate, Num of Loan, Changed Credit Limit, Outstanding Debt, Total EMI per month, Credit Mix Bad, Payment of Min Amount NM, Payment of Min Amount Yes, Payment Behaviour Low spent Medium value payments, and Payment Behaviour Low spent Small value payments. All of these match with our knowledge from the finance domain.

From these covariances, we can notice that Total EMI per month may not be a differentiating factor for the credit score since it both occurs in good and bad credit scores' positive covariance part. However, saying that it is not a differentiating factor for the credit score by only looking at the covariances would not be true.

5.2.2. Correlations

The one part of the calculated correlations are given below. Since the spreadsheet is too big, one can find the whole table in the spreadsheet given in reference [41]. The **Table 10** given below shows only part of the covariances.

	Delay_from_due_date	Num_of_Delayed_Payment	Num_Credit_Inquiries	Credit_Utilization_Ratio	Credit_History_Age	Amount_invested_monthly
Delay_from_due_date	1	0.541353329	0.540787099	-0.06387840675	-0.4912106933	-0.1623885405
Num_of_Delayed_Payment	0.541353329	1	0.5002161677	-0.07371245528	-0.481629954	-0.1821132362
Num_Credit_Inquiries	0.540787099	0.5002161677	1	-0.07714807711	-0.6107024313	-0.1764788016
Credit_Utilization_Ratio	-0.06387840675	-0.07371245528	-0.07714807711	1	0.07321416075	-0.005903749139
Credit_History_Age	-0.4912106933	-0.481629954	-0.6107024313	0.07321416075	1	0.1757384908
Amount_invested_monthly	-0.1623885405	-0.1821132362	-0.1764788016	-0.005903749139	0.1757384908	1
Monthly_Balance	-0.2818453071	-0.3046299126	-0.3222082087	0.2451004838	0.3258199556	0.0466904351
Age	-0.1740088328	-0.1835903306	-0.2534907526	0.02536935357	0.2342990803	0.05718919771
Annual_Income	-0.2499965155	-0.2850736171	-0.2808057339	0.1758399323	0.2723936533	0.6274303495
Num_Bank_Accounts	0.5598546167	0.6003373961	0.5191824599	-0.07170076385	-0.48530169	-0.1821413106
Num_Credit_Card	0.4792249335	0.4230270762	0.4597806994	-0.05542565536	-0.4170237405	-0.1432578112
Interest_Rate	0.5889227769	0.5705866445	0.6334317408	-0.07562716672	-0.5762338516	-0.1948744175
Num_of_Loan	0.5013107199	0.4743651715	0.5659178613	-0.1003649247	-0.6058101787	-0.1644029303

Monthly_Inhand_Salary	-0.2492972856	-0.2840206922	-0.2800383318	0.17601579	0.2714815523	0.6280749459
Changed_Credit_Limit	0.3069383308	0.3402491205	0.3961164082	-0.05059002419	-0.443293643	-0.120454555
Outstanding_Debt	0.5719503734	0.5045749588	0.5978210459	-0.07112212493	-0.6293280615	-0.1738304482

Table 10

Correlation gives more information to us between the relationship of the variables since it does not depend on the scale of the features. In other words, it uses standardization. Since there are many features, I grouped them for each feature that has correlation with other features which have correlation values higher than 0.7 and smaller than -0.7. The result for threshold 0.7 and -0.7 can be seen in **Table 11**.

	Strong Positive Corr >0.7	**Strong Negative Corr <-0.7**
Delay_from_due_date	Delay_from_due_date: 1.0	
Num_of_Delayed_Payment	Num_of_Delayed_Payment: 1.0	
Num_Credit_Inquiries	Num_Credit_Inquiries: 1.0	
Credit_Utilization_Ratio	Credit_Utilization_Ratio: 1.0	
Credit_History_Age	Credit_History_Age: 1.0	
Amount_invested_monthly	Amount_invested_monthly: 1.0	
Monthly_Balance	Monthly_Balance: 1.0 Annual_Income: 0.7054464068905248 Monthly_Inhand_Salary: 0.7065162345530189	
Age	Age: 1.0	
Annual_Income	Monthly_Balance: 0.7054464068905248 Annual_Income: 1.0 Monthly_Inhand_Salary: 0.9981578364124154	
Num_Bank_Accounts	Num_Bank_Accounts: 1.0	

Num_Credit_Card	Num_Credit_Card: 1.0	
Interest_Rate	Interest_Rate: 1.0	
Num_of_Loan	Num_of_Loan: 1.0	
Monthly_Inhand_Salary	Monthly_Balance: 0.7065162345530189 Annual_Income: 0.9981578364124154 Monthly_Inhand_Salary: 1.0	
Changed_Credit_Limit	Changed_Credit_Limit: 1.0	
Outstanding_Debt	Outstanding_Debt: 1.0 Credit_Mix_Bad: 0.7609083085762309	
Total_EMI_per_month	Total_EMI_per_month: 1.0	
Credit_Mix_Bad	Outstanding_Debt: 0.7609083085762309 Credit_Mix_Bad: 1.0	
Credit_Mix_Good	Credit_Mix_Good: 1.0 Payment_of_Min_Amount_No: 0.7211122634594723	
Credit_Mix_Standard	Credit_Mix_Standard: 1.0	
Payment_of_Min_Amount_NM	Payment_of_Min_Amount_NM: 1.0	
Payment_of_Min_Amount_No	Credit_Mix_Good: 0.7211122634594723 Payment_of_Min_Amount_No: 1.0	Payment_of_Min_Amount_Yes: -0.7799781513496837
Payment_of_Min_Amount_Yes	Payment_of_Min_Amount_Yes: 1.0	Payment_of_Min_Amount_No: -0.7799781513496837
Payment_Behaviour_High_spent_Large_value_payments	Payment_Behaviour_High_spent_Large_value_payments: 1.0	
Payment_Behaviour_High_spent_Medium_value_payments	Payment_Behaviour_High_spent_Medium_value_payments: 1.0	
Payment_Behaviour_High_spent_Small_value_payments	Payment_Behaviour_High_spent_Small_value_payments: 1.0	
Payment_Behaviour_Low_spent_Large_value_payments	Payment_Behaviour_Low_spent_Large_value_payments: 1.0	
Payment_Behaviour_Low_spent_Medium_value_payments	Payment_Behaviour_Low_spent_Medium_value_payments: 1.0	
Payment_Behaviour_Low_spent_Small_value_payments	Payment_Behaviour_Low_spent_Small_value_payments: 1.0	

Credit_Score_Good	Credit_Score_Good: 1.0	
Credit_Score_Poor	Credit_Score_Poor: 1.0	
Credit_Score_Standard	Credit_Score_Standard: 1.0	

Table 11

According to that result, monthly balance , annual income and monthly in hand salary have strong positive linear relation. Outstanding debt and credit mix bad have also strong positive linear relation. Even though these two variables have a relationship according to the given dataset, according to the base knowledge of the finance domain, there should not be a direct relationship between the outstanding debt and credit mix values. The strong negative linear relationship found only between the payment of min amount yes and no features. That makes sense since they are opposite categorical data. One can find these results also in reference [42].

5.3. Results of Visualization of Data

5.3.1. Histograms

For the categorical data in the dataset, I created histograms to observe their distribution.

a-) Credit Mix:

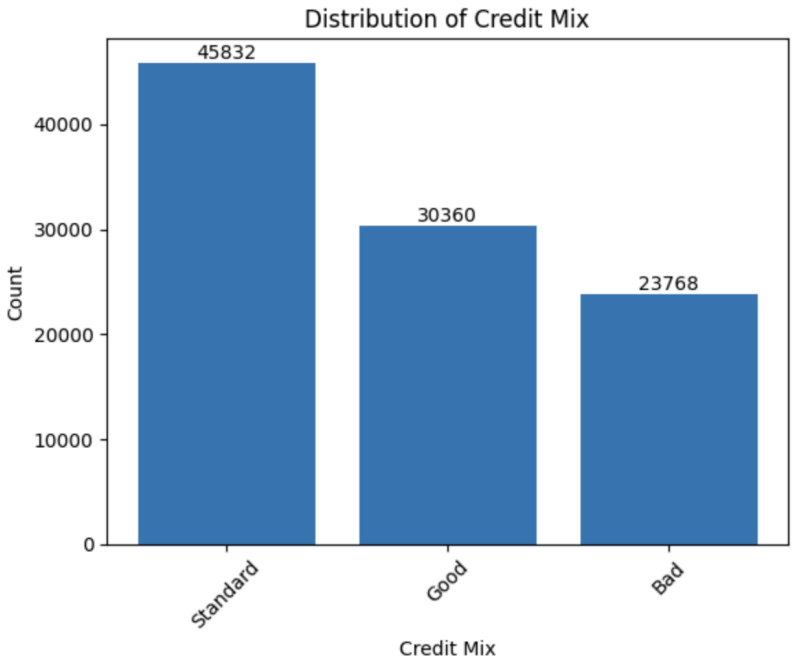


Figure 2

In **Figure 2,** the distribution of the dataset for three different credit mix values is given. As one can see, the standard value of the credit mix value is more frequent in the dataset.

b-) Payment of Minimum Amount:

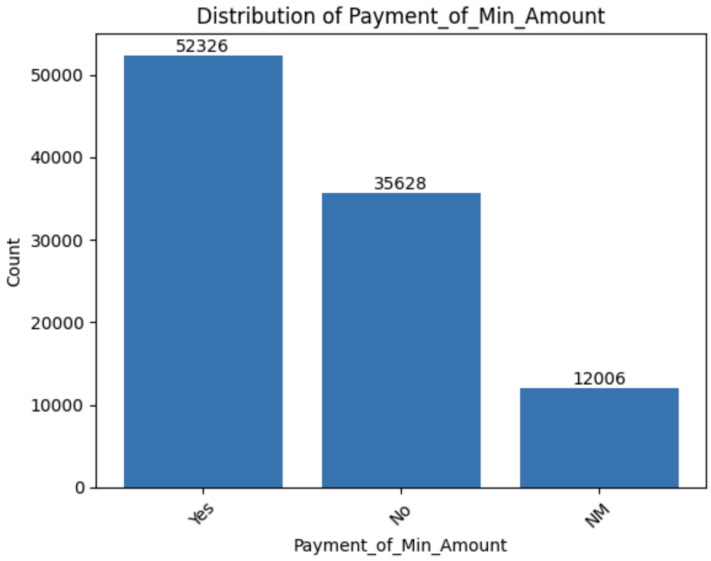


Figure 3

In **Figure 3,** the distribution of the dataset for three different behaviors for payment of minimum amount values is given. As one can see, the yes value of the minimum payment amount value is more frequent in the dataset.

c-) Payment Behavior:

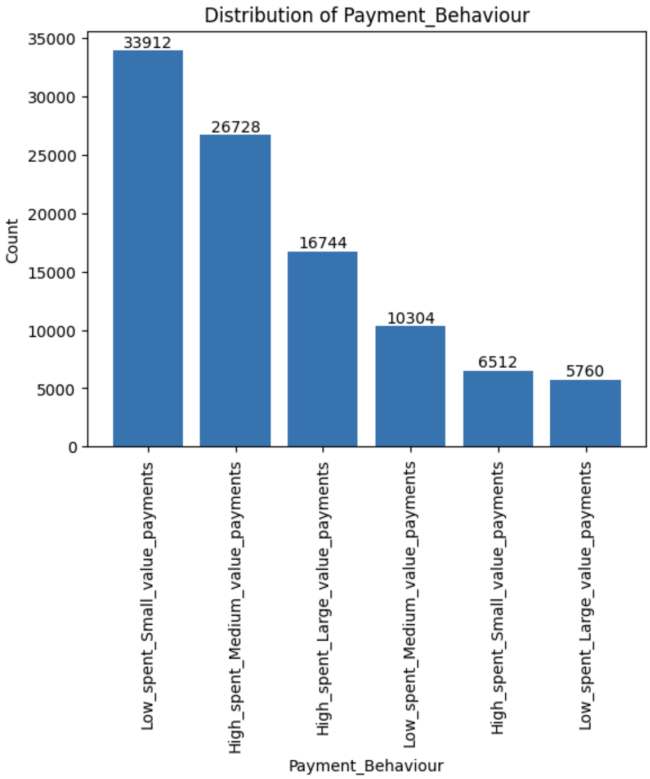


Figure 4

In **Figure 4,** the distribution of the dataset for six different payment behaviors is given. As one can see, the yes value of the minimum payment amount value is more frequent in the dataset. The most common behavior for the individuals in this dataset is low spent small value payments. The second most common payment behavior is high spent medium value payments. The lowest payment behavior is low spent large value payments.

d-) Credit Score:

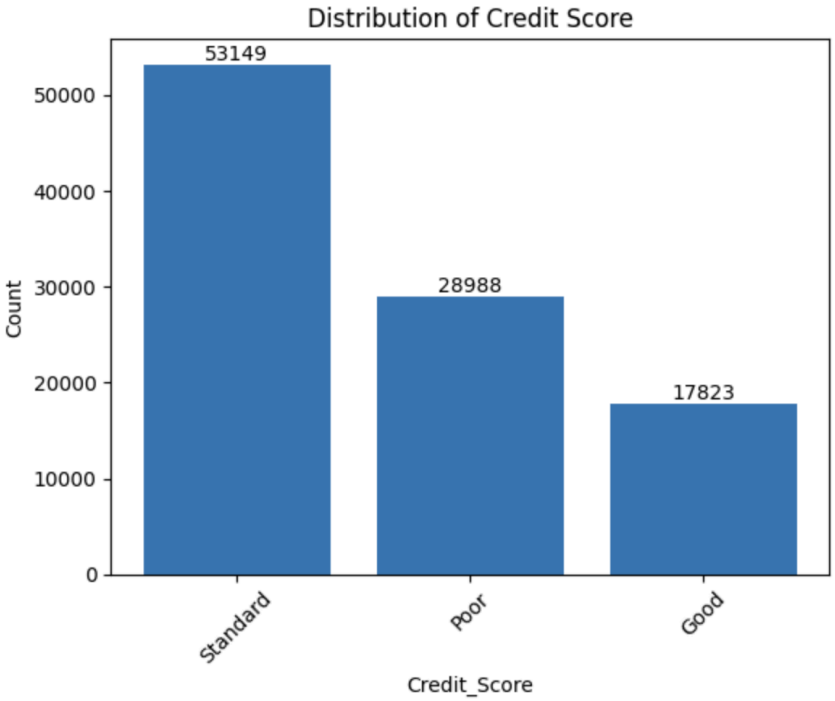


Figure 5

In **Figure 5,** the distribution of the dataset for three different credit score values is given. This histogram is important because it enables us to see whether the dataset is balanced or unbalanced. The most common credit score classified in this dataset is standard type. After the standard type, the most common one is poor credit scores followed by a good credit score. The number difference between these values are considerably high. In order to make this dataset more balanced, one could randomly select the standard and poor credit score data with the same count of the good credit score data. However, to avoid the effect of an unbalanced dataset for the models I trained, I used a random train-test split method.

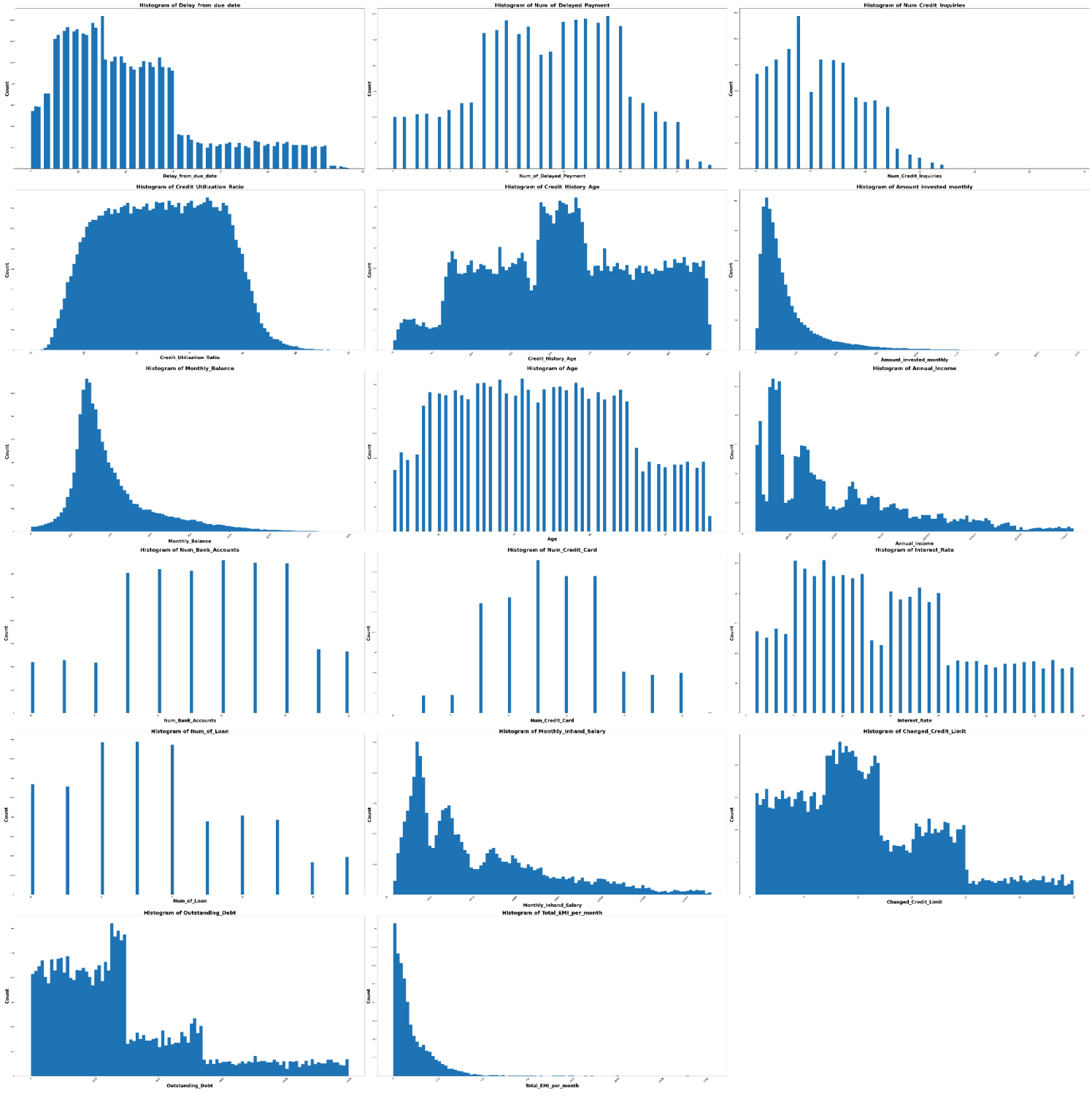

Histogram of Delay_from_due_date
Count
Delay_from_due_date
Histogram of Num_of_Delayed_Payment
Count
Num_of_Delayed_Payment
Histogram of Num_Credit_Inquiries
Count
Num_Credit_Inquiries
Histogram of Credit_Utilization_Ratio
Count
Credit_Utilization_Ratio
Histogram of Credit_History_Age
Count
Credit_History_Age
Histogram of Amount_invested_monthly
Count
Amount_invested_monthly
Histogram of Monthly_Balance
Count
Monthly_Balance
Histogram of Age
Count
Age
Histogram of Annual_Income
Count
Annual_Income
Histogram of Num_Bank_Accounts
Count
Num_Bank_Accounts
Histogram of Num_Credit_Card
Count
Num_Credit_Card
Histogram of Interest_Rate
Count
Interest_Rate
Histogram of Num_of_Loan
Count
Num_of_Loan
Histogram of Monthly_Inhand_Salary
Count
Monthly_Inhand_Salary
Histogram of Changed_Credit_Limit
Count
Changed_Credit_Limit
Histogram of Outstanding_Debt
Count
Outstanding_Debt
Histogram of Total_EMI_per_month
Count
Total_EMI_per_month

Figure 6

In **Figure 6,** the distributions of the dataset for not categorical features are given. As one can see, for some of the features histogram takes discrete values since the features can take integer values such as number of delayed payments. On the other hand, some features can take continuous values such as credit utilization ratio. The histogram of these kinds of features seems continuous too. Among these histograms only the monthly balance histogram is similar to the normal distribution.

5.3.2. Box Plots

To understand the relation between the features and the credit score, I also created the box plots as seen in **Figure 7.**

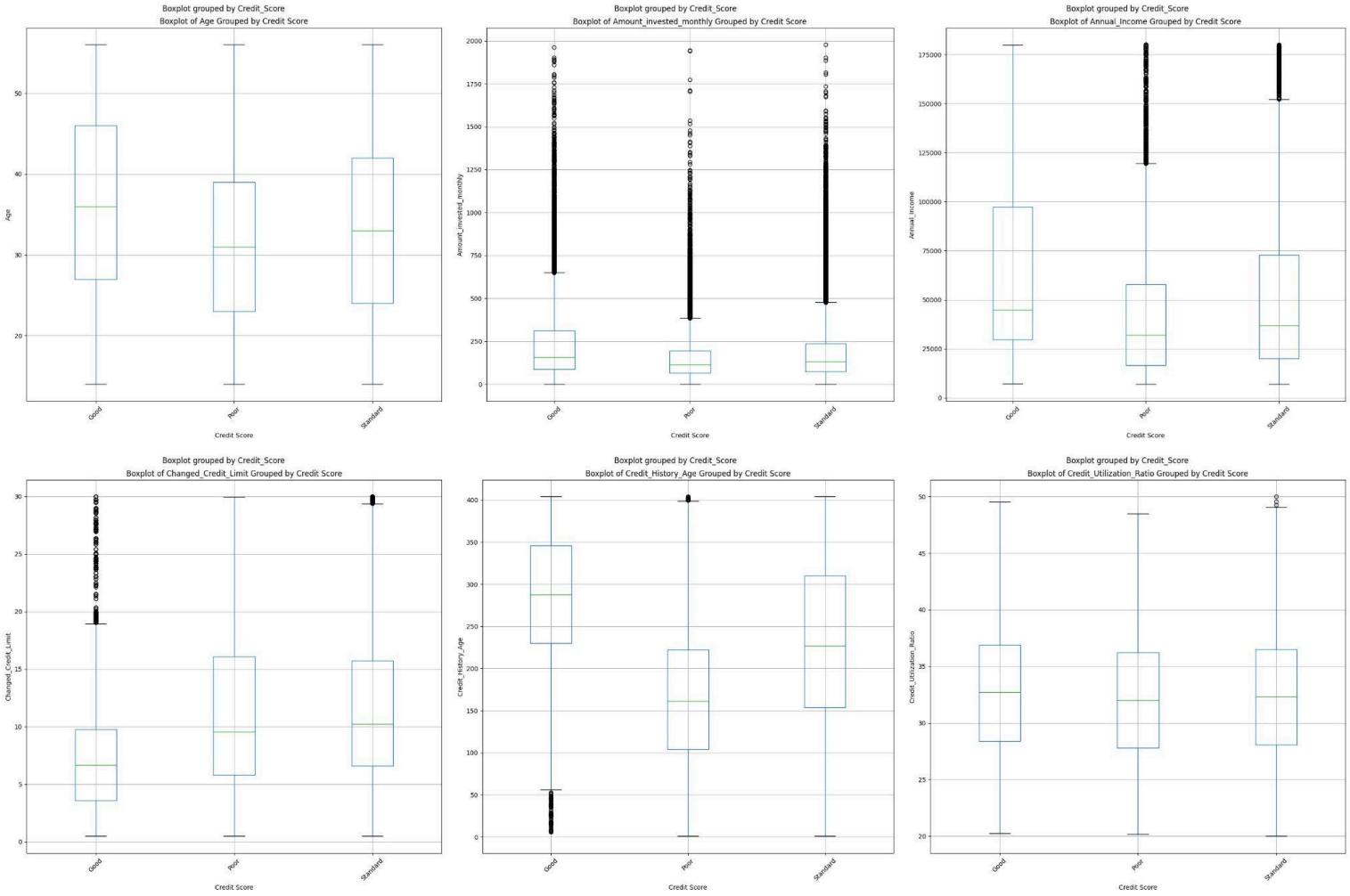


Figure 7

Interpretation of the plots in Figure 7:

- The upper left plot is the boxplot of age grouped by the credit score. Medians of the age for each credit score group are different. For older people, having a good or standard credit score seems more common from the box plot since the median of good and standard credit score is higher than the poor credit score in terms of age. The interquartile range of the boxes seems not so different, so the data for all groups seems equally dispersed. According to this box plot, there is not an obvious difference between the different age distributions for credit scores. Since the median is in the middle of the box, and the whiskers are about the same on both sides of the box, the distribution for all credit score types seems symmetrical. There are no outliers for this boxplot.

- The upper middle plot is the boxplot of amount invested monthly grouped by credit score. Medians for each group seem so close to each other for this plot. The interquartile range for each credit score group seems similar and short. This means the data is less dispersed for this feature. Since the median seems closer to the bottom of the box, the distribution seems positively skewed for all of three boxes. That means the median is lower than the mean. According to this box plot, there is not an obvious difference between the different amount investment distributions for credit scores. There are many outliers that show there are many data which are more extreme than the expected variation.

- The upper right plot is the boxplot of annual income grouped by the credit score. Medians of the annual income for a good credit score group seems different. For people who have higher annual income, having a good credit score seems more common from the box plot since the median of a good credit score is higher than the poor credit score. Besides that, the interquartile range of the good credit score group seems higher than the other two groups as in the range of the annual income for the good credit score groups. That means, regardless of how much people earn, they may have good credit scores. The absence of outliers for this group also proves that.

- The lower left plot is the boxplot of the changed credit limit grouped by the credit score. Medians of the changed credit limit number for a good credit score group seem different. For people whose credit limit changed less, having a good credit score seems more probable from the box plot since the median of a good credit score is lower than the poor and standard credit score. Besides that, the interquartile range is lower for the good credit score group compared to the standard and poor credit score group and there are more outliers that show there are many data which are more extreme than the expected variation.

- The lower middle plot is the boxplot of the credit history age grouped by the credit score. There are significant differences between the medians of credit history age for each group. Especially, the good credit

score group has a higher median while the bad credit score group has lower median. For people who have longer credit history, having a good credit score seems more probable from the box plot since the box and median of a good credit score is higher than the poor credit score. Besides that, there is not an obvious difference between the interquartile range of the groups.

- The lower right plot is the boxplot of credit utilization ratio grouped by the credit score. There are no significant differences between the medians of credit utilization ratio between the groups.Their interquartile ranges are also similar and moderate. Since the median is in the middle of the box, and the whiskers are about the same on both sides of the box, the distribution for all credit score types seems symmetrical.

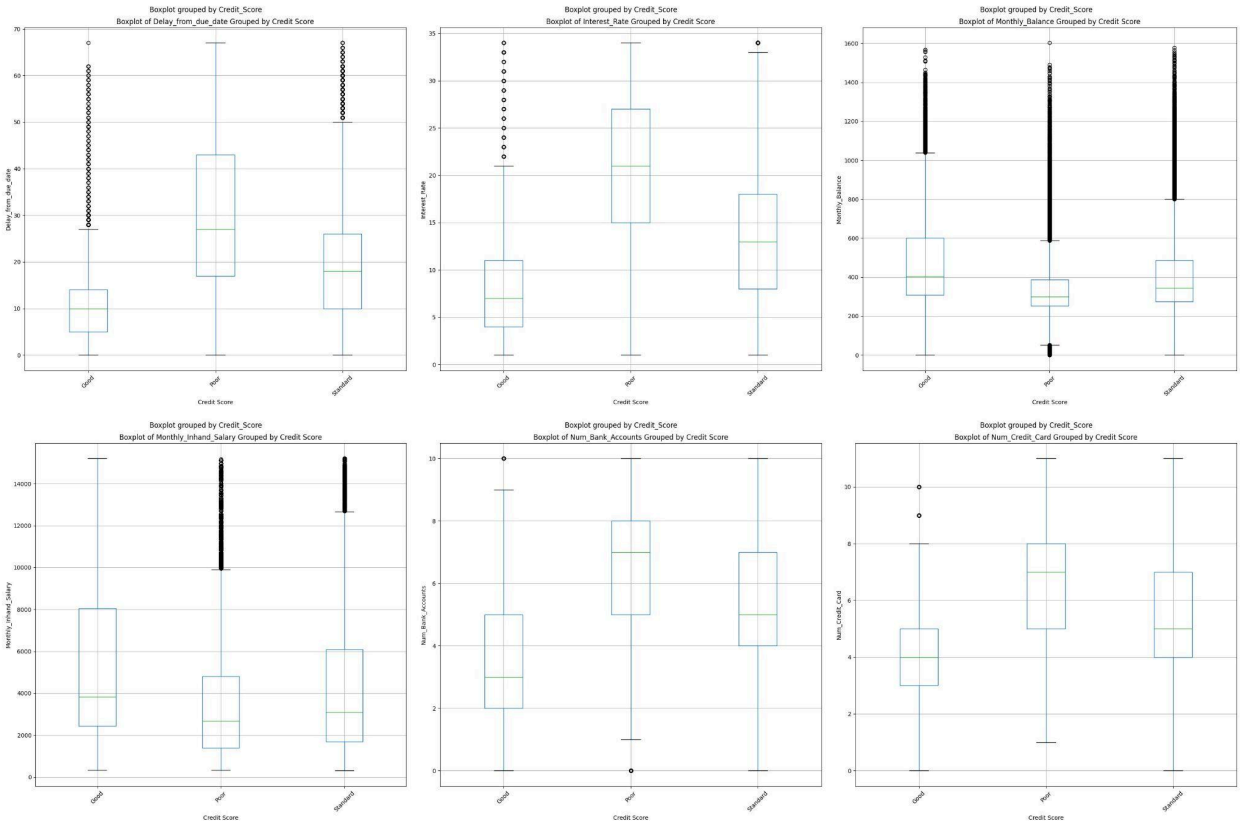


Figure 8

Interpretation of the plots in Figure 8:

- The upper left plot is the boxplot of delay from due date grouped by the credit score. There are significant differences between the medians of delay from due date for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher median. For people who have less

delayed payments from due date, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, there is an obvious difference between the interquartile range of the groups. The poor credit score group has higher interquartile range compared to the others. So, it has more dispersed data.Except the poor credit score groups, since the median is in the middle of the box, and the whiskers are about the same on both sides of the box, the distribution seems symmetrical. There are moderately many outliers in the good and standard credit score groups.

- The upper middle plot is the boxplot of interest rate grouped by the credit score. There are significant differences between the medians of interest rate for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher median. For people who have less interest rate, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, there is an obvious difference between the interquartile range of the groups. The poor credit score group has higher interquartile range compared to the others. So, it has more dispersed data. There are moderately many outliers in the good and standard credit score groups.

- The upper right plot is the boxplot of monthly balance grouped by the credit score. There are not so significant differences between the medians of interest rate for each group. The good credit score has a little bit higher median value. Since the medians are in the middle of the box, and the whiskers are about the same on both sides of the box, the distribution for all credit score types seems symmetrical. The distribution seems slightly positively skewed for all of three boxes. That means the median is lower than the mean. One thing that we can notice from this box plot is that there are many outliers. This shows the variability in the data is high.

- The lower left plot is the boxplot of monthly-inhand salary grouped by the credit score. It is possible to notice that it has similar distribution with the annual income boxplot in **Figure 7.** Medians of the monthly-inhand salary for a good credit score group seems different. For people who have higher monthly-inhand salary, having a good credit score seems more common from the box plot since the median of a good credit score is higher than the poor credit score. Besides that, the interquartile range of the good credit score group seems higher than the other two groups as in the range of the monthly-inhand salary for the good credit score groups. That means, regardless of how much people earn, they may have good credit scores. The absence of outliers for this group also proves that.

- The lower middle plot is the boxplot of the number of bank accounts grouped by the credit score. There are significant differences between the medians of the number of bank accounts for each group. Especially, the good credit score group has a lower median while the bad credit score group has higher median. For people

who have less bank accounts, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, there is not an obvious difference between the interquartile range of the groups. The distribution seems slightly positively skewed for the good and standard credit score groups. That means the median is lower than the mean. The distribution seems slightly negatively skewed for the poor credit score groups. That means the median is higher than the mean.

- The lower right plot is the boxplot of the number of credit cards grouped by the credit score. There are significant differences between the medians of the number of credit cards for each group. Especially, the good credit score group has a lower median while the bad credit score group has higher median. For people who have less credit cards, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, the interquartile range of the good credit score is lower than the others. It means it has less dispersed data. The distribution seems slightly positively skewed for standard credit score groups. That means the median is lower than the mean. The distribution seems slightly negatively skewed for the poor credit score groups. That means the median is higher than the mean. For the good credit score group, the distribution seems symmetrical.

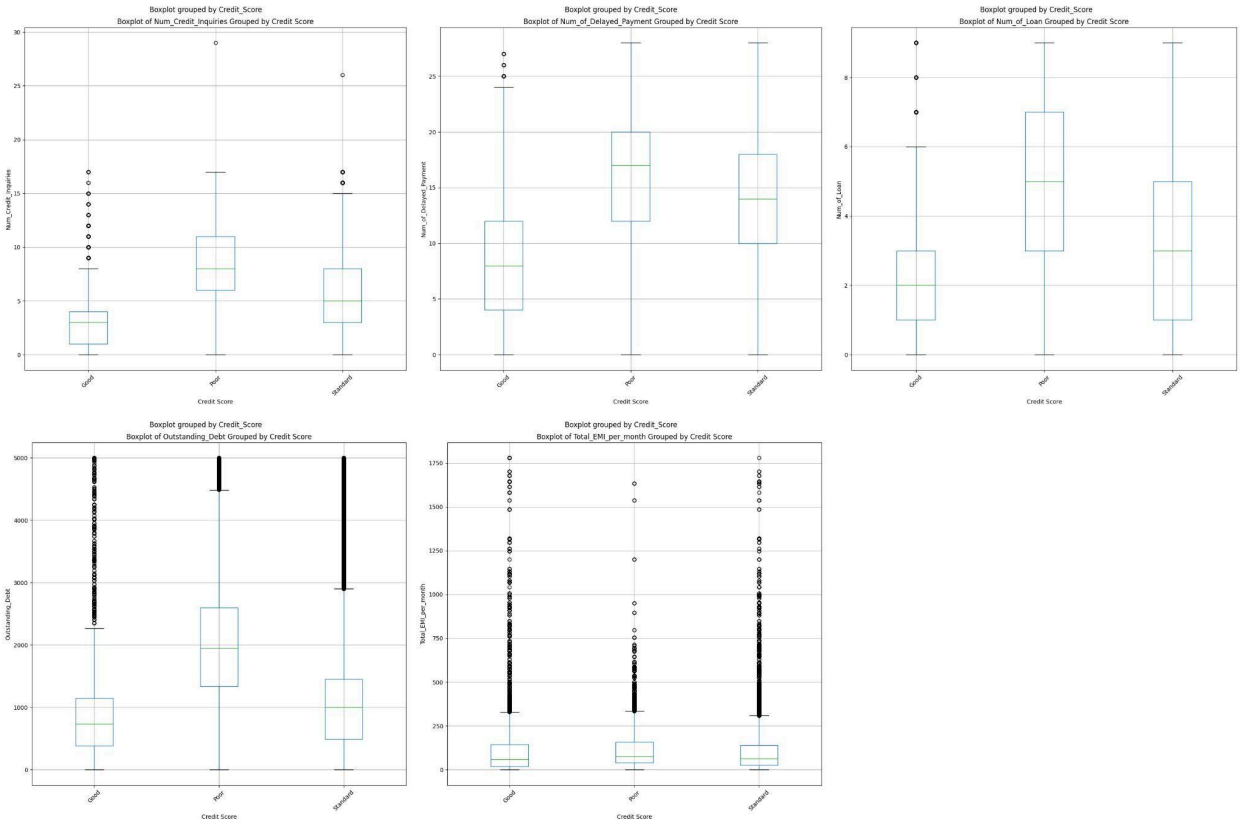


Figure 9

Interpretation of the plots in Figure 9:

- The upper left plot is the boxplot of the number of credit inquiries grouped by credit score. Medians for each group seem so close to each other for this plot. There are significant differences between the medians of the number of credit inquiries for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher median. For people who have less number of inquiries, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. The interquartile range for the good credit score group seems short. This means the data is less dispersed for this feature. Since the median seems closer to the top of the box, the distribution seems negatively skewed for the good credit score group. That means the median is higher than the mean. There are few outliers that show there is some data which are more extreme than the expected variation for the standard and the poor credit score groups.

- The upper middle plot is the boxplot of the number of delayed payments grouped by the credit score. There are significant differences between the medians of the number of delayed payments for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher

median. For people who have less number of delayed payments, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, there is not a significant difference between the interquartile range for all three groups. There are only three outliers in the good and credit score groups.

- The upper right plot is the boxplot of the number of loans grouped by the credit score. There are significant differences between the medians of the number of delayed payments for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher median. For people who have less number of loans, having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, the interquartile range of the good credit score is lower than the others. It means it has less dispersed data. For all credit score groups, the distributions seem symmetrical.

- The lower left plot is the boxplot of the outstanding debt grouped by the credit score. There are significant differences between the medians of the outstanding debt for each group. Especially, the good credit score group has a lower median while the poor credit score group has higher median. For people who have less outstanding debt , having a good credit score seems more probable from the box plot since the box and median of a good credit score is lower than the poor credit score. Besides that, the interquartile range of the good credit score is lower than the others. It means it has less dispersed data. For all credit score groups, the distributions seem symmetrical.

- The lower middle plot is the boxplot of total EMI per month monthly grouped by credit score. Medians for each group seem so close to each other for this plot. The interquartile range for each credit score group seems similar and short. This means the data is less dispersed for this feature. Since the median seems closer to the bottom of the box, the distribution seems positively skewed for all of three boxes. That means the median is lower than the mean. According to this box plot, there is not an obvious difference between the different total EMI per month for credit scores. There are many outliers that show there are many data which are more extreme than the expected variation.

5.3.3. Scatter Plots

As we know from our lectures, scatter diagrams are useful to visually evaluate the relationship between pairs of variables. Since in this project there are nearly 20 features, in this part of the report, I will not put all of the scatter plots. Instead of that, I will put the ones, which have strong correlation between them as given in the correlation part. One can find all of the scatter plots in the file “scatterplots.zip” file that I submitted.

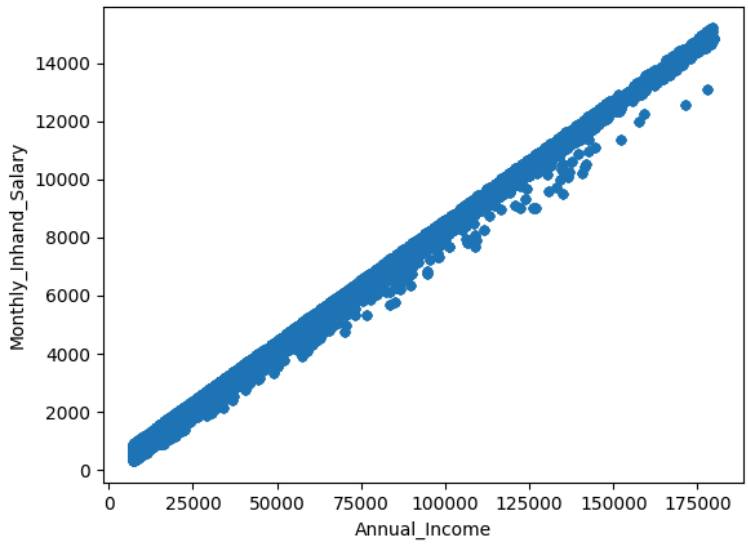


Figure 10

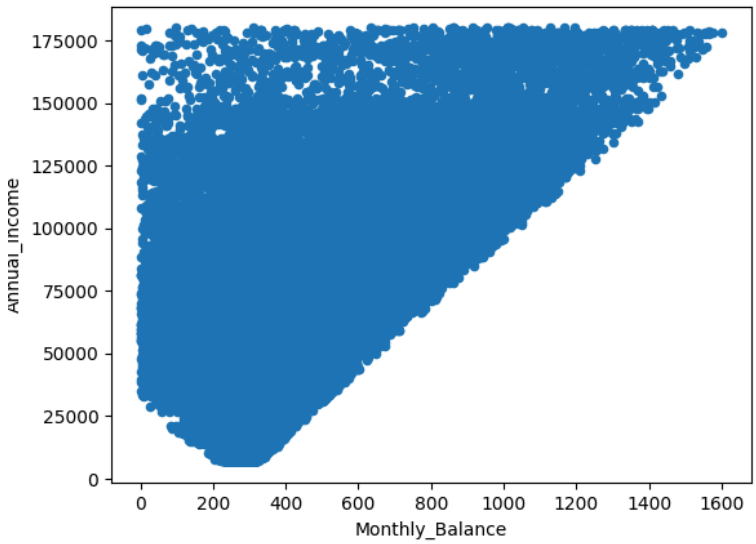


Figure 11

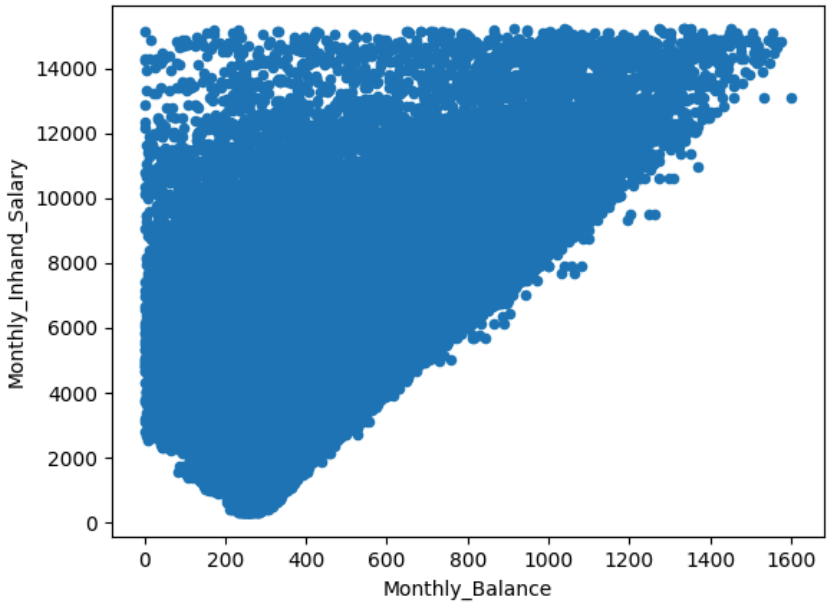
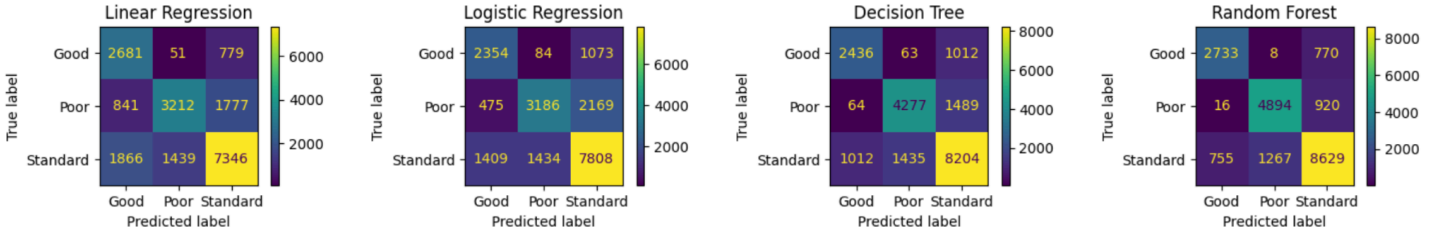


Figure 12

As we mentioned in the correlation part, Monthly balance , annual income and monthly in hand salary have strong positive linear relation (> 0.7). **Figure 10, 11** and **12** also show correlation between these variables. When annual income increases, the monthly income salary also increases linearly as seen from **Figure 11.** When we look at **Figure 11,** we can see that the people who have less annual income also have less maximum monthly balance compared to the people who have higher annual income. **Figure 12** also shows the similar structure with **Figure 11.**

5.4. Results of Statistical Analysis for Main Goal



	precision	recall	f1-score	support
0	0.50	0.76	0.60	3511
1	0.68	0.55	0.61	5830
2	0.74	0.69	0.71	10651
micro avg	0.66	0.66	0.66	19992
macro avg	0.64	0.67	0.64	19992
weighted avg	0.68	0.66	0.66	19992
samples avg	0.66	0.66	0.66	19992

	precision	recall	f1-score	support
0	0.56	0.67	0.61	3511
1	0.68	0.55	0.60	5830
2	0.71	0.73	0.72	10651
accuracy			0.67	19992
macro avg	0.65	0.65	0.64	19992
weighted avg	0.67	0.67	0.67	19992

	precision	recall	f1-score	support
0	0.69	0.69	0.69	3511
1	0.74	0.73	0.74	5830
2	0.77	0.77	0.77	10651
accuracy			0.75	19992
macro avg	0.73	0.73	0.73	19992
weighted avg	0.75	0.75	0.75	19992

	precision	recall	f1-score	support
0	0.78	0.78	0.78	3511
1	0.79	0.84	0.82	5830
2	0.84	0.81	0.82	10651
accuracy			0.81	19992
macro avg	0.80	0.81	0.81	19992
weighted avg	0.81	0.81	0.81	19992

Figure 13

In **Figure 13,** the class 0 in the classification reports corresponds to "Credit_Score_Good", the class 1 corresponds to "Credit_Score_Poor" and the class 2 corresponds to the "Credit_Score_Standard". The upper colored squares are the confusion matrices while the bottom written parts shows their classification reports.

Metrics Used To Evaluate

There are different evaluation metrics to evaluate the performance of the model. Their general rule is given below for the reader. "TP" means "True Positives", "TN" means "True Negatives", "FP" means "False Positives" and "FN" means "False Negatives".

- Accuracy is the ratio of number of correct predictions to size of the dataset.

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

- Precision indicates how many positive predictions made are correct .

$$Precision = \frac{TP}{TP+FP}$$

- Recall indicates how many of the originally positive labeled data are predicted correctly by the model .

$$Recall = \frac{TP}{TP+FN}$$

- F-1 Score combines the precision and the recall. It is the harmonic mean of these two metrics.

$$F1 = 2 * \frac{Precision.Recall}{Precision+Recall}$$

The support shown in the classification reports in **Figure 13** gives the number of data whose ground truth is Good, Standard and Poor. In the classification reports in **Figure 13,** one can also see that these values are calculated for each class. In addition to this, it is possible to see the macro average and weighted average calculation for metrics precision, recall and f-1. The macro average score for a metric is calculated by summing up the values of the metric for each class and dividing it to the number of classes.

- $MA_{Precision} = \frac{Precision_{Good}+Precision_{Poor}+Precision_{Standard}}{3}$
- $MA_{Recall} = \frac{Recall_{Good}+Recall_{Poor}+Recall_{Standard}}{3}$
- $MA_{F1} = \frac{F1_{Good}+F1_{Poor}+F1_{Standard}}{3}$

Finally, the weighted average is calculated by multiplying the metric of a specific class with the support proportion of that class and summing up all of them.

- $WA_{Precision} = \frac{Precision_{Good}Support_{Good}+Precision_{Poor}*Support_{Poor}+Precision_{Standard}*Support_{Standard}}{Support_0+Support_1}$
- $WA_{Recall} = \frac{Recall_{Good}*Support_{Good}+Recall_{Poor}*Support_{Poor}+Recall_{Standard}*Support_{Standard}}{Support_{Good}+Support_{Poor}+Support_{Standard}}$

$$WA_{F1} = \frac{F1_{Good}*Support_{Good}+F1_{Poor}*Support_{Poor}+F1_{Standard}*Support_{Standard}}{Support_{Good}+Support_{Poor}+Support_{Standard}}$$

Under the lights of the meanings of the metrics, we can discuss the results of the model.

- According to **Figure 13,** the linear regression model has the worst performance among the four models. Even though few of the values in its classification report are better than

the logistic regression model (such as the recall value for class 0 (Good) and precision score for class 2 (Standard)), the overall model does not have a good performance. There can be many reasons for that. Firstly, linear regression is not generally used for the classification as we mentioned since they output the continuous variables, not any probabilities etc. Secondly, the extra step we performed after the linear regression (taking maximum among the scores) may not be the most appropriate way to classify the data according to the result of the linear model. However, in addition to these, since we have many features with different effects to the credit score, expecting for them to have a linear relationship between the credit score value seems also improbable. To conclude, linear regression model is not suitable to classify this dataset in general.

- According to **Figure 13,** as mentioned above we can see that the logistic regression model performs better than the linear regression model. However, when we look at the evaluation metrics and the confusion matrix, we can easily see that it performs worse than both the decision tree and random forest model. Its accuracy is 0.67 which is good but not enough. As we mentioned before, the logistic regression model is suitable for the classification. However, it is still a linear model. We can easily conclude for the dependent variable (credit score) does not have a linear relationship with the features used after considering the evaluation metrics of both linear and logistic regression models.
- According to **Figure 13,** we can easily say that the random forest model outperforms all of the models. It has accuracy 0.85 on the test data. It means even though the depth of the trees are too much (around 42), it did not overfit. This result is expected since a random forest model uses 100 decision trees, it should outperform the decision model.
- When we look at the confusion matrix of each four of the models, we can see that all of the models are not confused about good and poor credit scores in general. What I mean is that data whose ground truth is poor are rarely labeled as good or data whose ground truth is good are rarely labeled as poor according to all confusion matrices. Most of the time, models are confused in a way that data whose ground truth is poor are labeled as standard or data whose ground truth is good are labeled as standard.

5.5. Results of Statistical Analysis for Feature Importance

5.5.1. Principal Component Analysis

5.5.1.1. Features Extracted

```
[0.29681447 0.10659768 0.07301723 0.04655482]
    0                                                  1
0  PC1                                     Interest_Rate
1  PC2                                     Annual_Income
2  PC3                                Credit_Mix_Standard
3  PC4  Payment_Behaviour_High_spent_Medium_value_paym...
```

Figure 14

In **Figure 14,** one can find the results related to the principal components. In order to reduce the dimensionality of the data, I extract 4 principal components from 20 features. As a result, I got the 4 coefficients for each component that can be seen on top of **Figure 14.** These coefficients indicate the percentage of variance explained by each of the principal components. As one can see from Figure 14, Principal component 1 has the highest explained variance 0.29 while principal component 2 has 0.106 and principal component 3 has 0.07. Principal component 4 has a very small explained variance 0.04. We know that the larger absolute value coefficient the more important the corresponding variable is in calculating the component (in terms of variance). However, these values are too low and the cumulative explained variance of these 4 principal components are approximately 0.521. This value is pretty low and shows us that extracting features by using PCA for this dataset may not be a good idea since it discards nearly half of the information. Therefore, applying PCA to the dataset, in this situation may cause underfitting and bad accuracy results contrary to the expectations.

5.5.2. Feature Importance through Random Forest

5.5.2.1. Features Used

```
Credit_Mix_Good                                          0.044949
Num_Credit_Inquiries                                     0.042944
Num_of_Delayed_Payment                                   0.040697
Credit_Mix_Standard                                      0.039210
Annual_Income                                            0.038708
Monthly_Inhand_Salary                                    0.038553
Num_Credit_Card                                          0.037225
Total_EMI_per_month                                      0.037196
Age                                                      0.033724
Num_Bank_Accounts                                        0.026397
Num_of_Loan                                              0.022132
Credit_Mix_Bad                                           0.016802
Payment_of_Min_Amount_No                                 0.012018
Payment_of_Min_Amount_Yes                                0.010899
Payment_Behaviour_High_spent_Medium_value_payments       0.004969
Payment_of_Min_Amount_NM                                 0.004898
Payment_Behaviour_Low_spent_Small_value_payments         0.004711
Payment_Behaviour_High_spent_Large_value_payments        0.003972
Payment_Behaviour_Low_spent_Medium_value_payments        0.003695
Payment_Behaviour_High_spent_Small_value_payments        0.003054
Payment_Behaviour_Low_spent_Large_value_payments         0.002645
dtype: float64
```

Figure 15

In **Figure 15,** we sorted the features whose feature importance is less than 0.05. These features are not important according to the random forest algorithm. Therefore, we are going to remove them from the feature list and train the models according to that.

- Even though there are some expected features considered as unimportant for the credit score such as age, most of the features considered as unimportant are not expected according to the descriptions of these features. For example, I expected the number of credit inquiries and the payment behavior variables to be important features. This shows that these variables' loss-cardinality means they have less unique values since the random forest gives importance to the features based on their impurities.

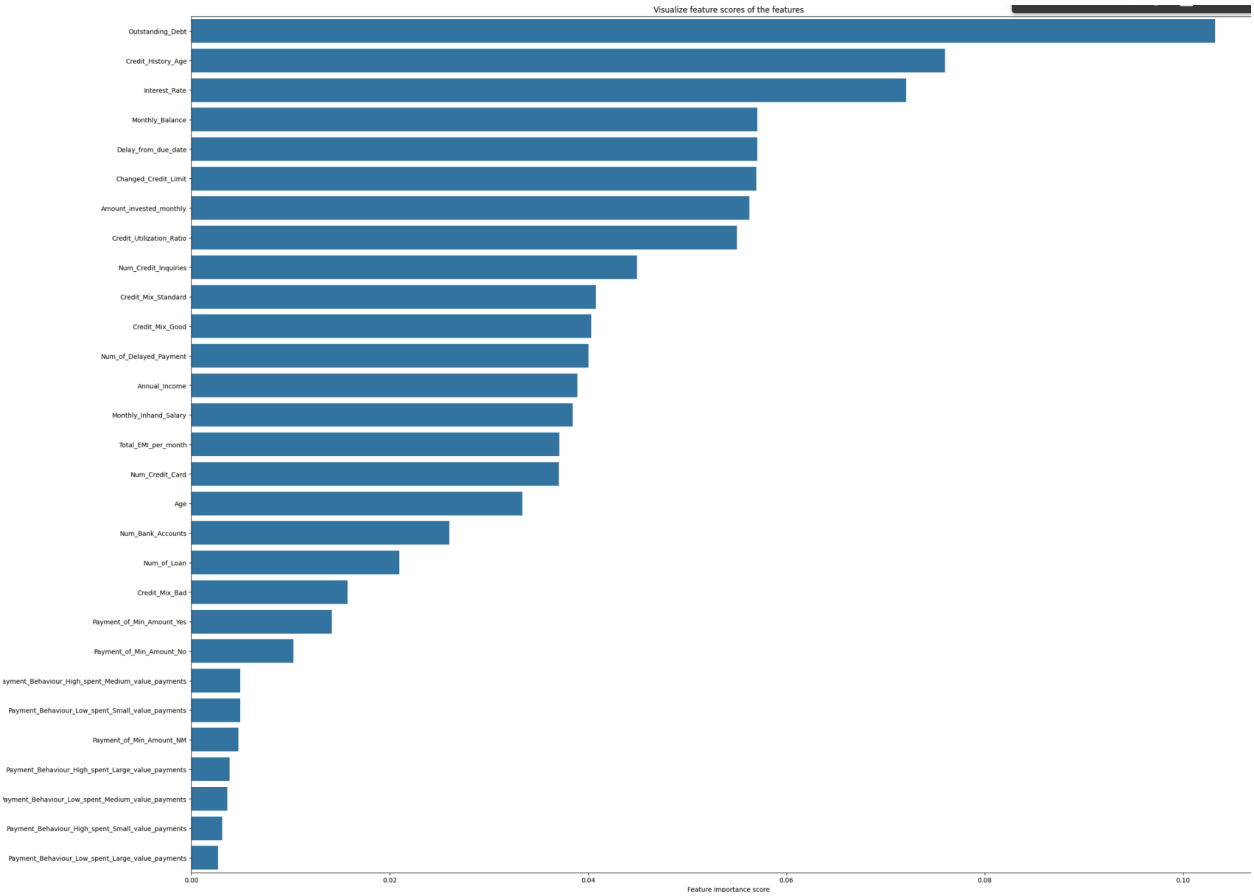


Figure 16

In **Figure 16,** we can easily observe that outstanding debt is the most important feature together with credit history age and interest rate. Interest rate is also considered as one of the important features for PCA as we can remember from **Figure 14.** In addition to these features, monthly balance, delay from due date, changed credit limit, amount invested monthly and credit utilization ratio also are moderately important features. These all are expected important features. These

features have higher cardinality according to the random forest model. In other words, they have more unique values than other features. This makes these features more important in the eye of random forest algorithms. However, since the random forest algorithm has the best accuracy when classifying the credit scores, it also makes these features more important according to us.

5.5.3. Overall Comparisons

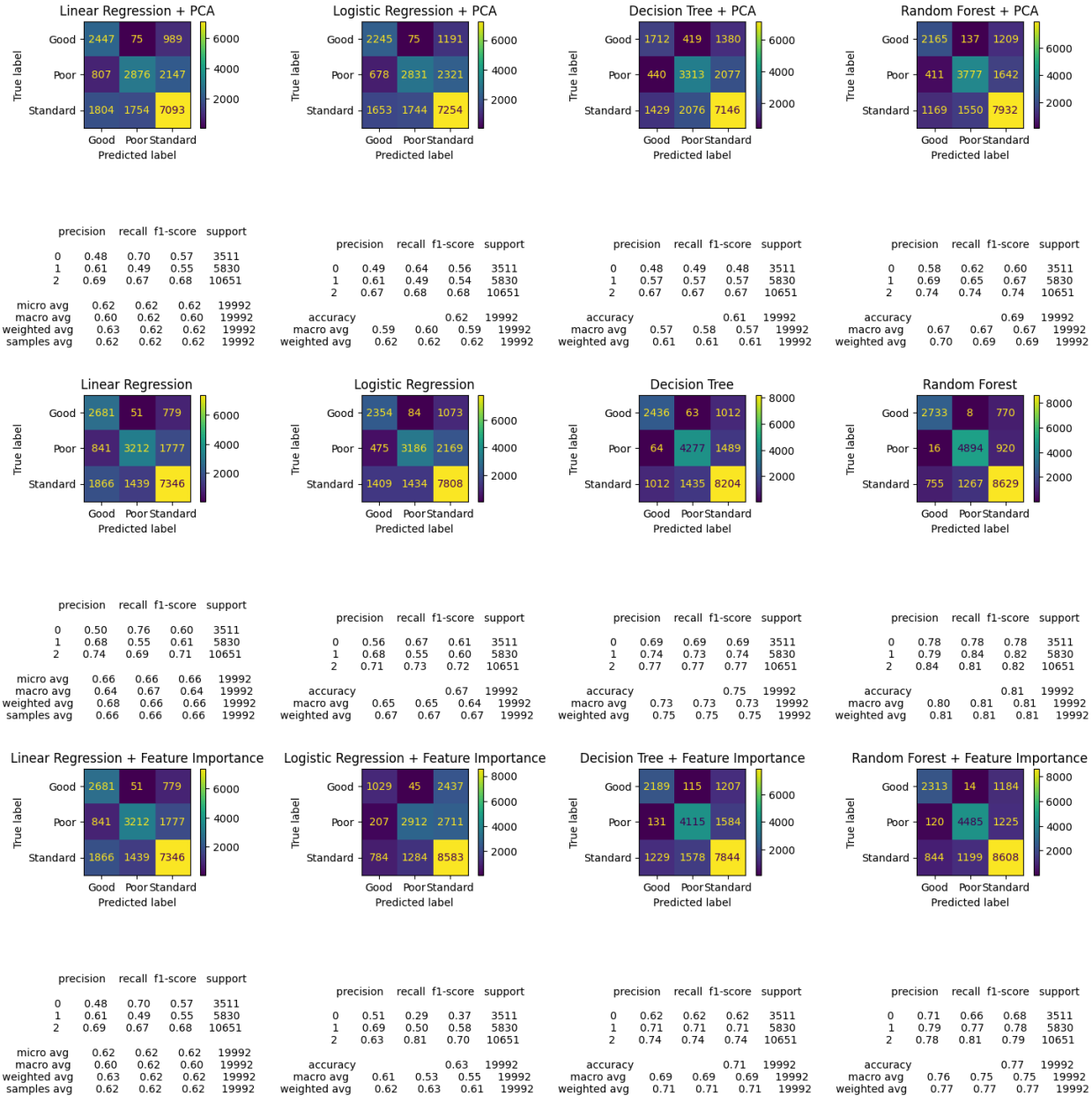


Figure 17

Figure 17 shows the performance of each model with PCA, feature importance and without any feature reduction techniques.

- For linear regression, we can see that both pca and the feature importance nearly has the same effect on the evaluation metrics. However, one can easily notice they reduce the performance of the model.
- For logistic regression, as in linear regression, both pca and the feature importance selection reduces the performance of the model. Especially recall of class good credit score decreased significantly for feature importance.
- For decision tree models and random forest models, both pca and the feature importance selection reduces the performance of the mode, again.
- When we compare the PCA and feature importance selection, we can easily see that the feature selection with feature importance outperforms the pca applied models. The reason can be the number of features used in the feature importance method is higher than the PCA. However, selecting more principal components whose coefficients are so small would not close the gap between PCA and Feature importance applied model.
- We can say that reducing features for this dataset did not improve the performance of the models. However, I can say that they improved the execution time.

6. Conclusions

In conclusion, this study achieved all of the goals defined in the introduction part. There are 3 goals of this study:

1. **Main goal: Find the best statistical model between these four: linear regression, logistic regression, decision tree and random forest to classify the credit scores of the customers**

   According to the performances of the four models, the best classification model of this dataset is the random forest model without any dimensionality method. It has accuracy 0.85 for the classifying the credit scores. The worst model is the linear regression model since it is not suitable for classification. Even though it is the worst model, it has accuracy around 0.67. The feature reduction methods applied (PCA and selection through feature importance) did not perform well but decreased the execution time.

2. **Find the features that have more effect on the classification of the credit scores**

   For Principal Component Analysis the features that have more effect on the classification of the credit score are interest rate, annual income, credit mix standard and payment behavior high spent medium value payment. However, since the models that applied PCA did not perform well, instead of PCA we should take feature importance extracted from random forest into consideration. According to feature importance found in random forests according to the impurity base, outstanding debt is the most important feature together with credit history age and interest rate.In addition to these features, monthly balance, delay from due date, changed credit limit, amount invested monthly and credit utilization ratio also are moderately important features. These all are expected important features from the definitions too.

3. T**he relation between the features**

   As we mentioned in the correlation part, Monthly balance , annual income and monthly in hand salary have strong positive linear relation (> 0.7). Scatter plots we provided also show correlation between these variables. When annual income increases, the monthly income salary also increases linearly**.** We can also observe that the people who have less annual income also have less maximum monthly balance compared to the people who have higher annual income.

   For further improvements, new models can be tried with better feature selection or extraction methods. In addition to that, execution times can be calculated and given numerically as the improvement of the feature dimensionality reduction methods' advantages.

7. References

[1]*What is a credit score?*. Consumer Financial Protection Bureau. (n.d.-a). https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-score-en-315/#:~:text=Companies%20use%20credit%20scores%20to,and%20credit%20limit%20you%20receive.

[2]*What is credit behaviour?: 2 answers from research papers*. SciSpace - Question. (n.d.). https://typeset.io/questions/what-is-credit-behaviour-c9e02e90-0ef4-de4d-d0a9-10cef8efc5de

[3]*Credit scoring models: FICO, VantageScore & more*. Debt.org. (2023, September 1). https://www.debt.org/credit/report/scoring-models/#:~:text=Experian%20and%20Equifax%20provide%2016,auto%20loans%20or%20credit%20cards.

[4]Rastogi, S. (2023, April 14). *Multi-class classification problem*. Kaggle. https://www.kaggle.com/datasets/sudhanshu2198/processed-data-credit-score?resource=download

[5]Avcontentteam. (2024, April 4). *PCA: What is Principal Component Analysis & How It Works? (updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/#:~:text=First%20Principal%20Component,-The%20first%20principal&text=It%20determines%20the%20direction%20of,higher%20than%20first%20principal%20component.

[6]*Articles*. Equifax. (n.d.-e). https://www.equifax.com/personal/education/credit-cards/articles/-/learn/when-late-credit-card-payments-post/#:~:text=of%20grace%20period.-,Even%20a%20single%20late%20or%20missed%20payment%20may%20impact%20credit,may%20still%20incur%20late%20fees.

[7]*How is your credit score calculated?*. LendingTree. (n.d.). https://www.lendingtree.com/credit-repair/how-is-my-credit-score-calculated/#:~:text=It%20all%20starts%20with%20your,%3A%20Experian%2C%20Equifax%20and%20TransUnion.

[8]Lake, R. (2024, January 29). *How long do late payments stay on a credit report? | time stamped*. Time. https://time.com/personal-finance/article/late-payments-on-credit-report/#:~:text=It's%20all%20about%20your%20overall,two%2C%20three%2C%20or%20more.

[9]*What's a credit inquiry?* Consumer Financial Protection Bureau. (n.d.-b). https://www.consumerfinance.gov/ask-cfpb/whats-a-credit-inquiry-en-1317/

[10]*Articles*. Equifax. (n.d.-d). https://www.equifax.com/personal/education/debt-management/articles/-/learn/credit-utilization-ratio/#:~:text=credit%20utilization%20ratio%3F-,Your%20credit%20utilization%20ratio%2C%20generally%20expressed%20as%20a%20percentage%2C%20represents,and%20paid%20back%20multiple%20times.

[11]It, A. (2023, May 10). *Length of credit history: An in-detail guide*. Credit Strong. https://www.creditstrong.com/length-of-credit-history/

[12]*Does investing affect your credit score*. Chase. (n.d.-a). https://www.chase.com/personal/credit-cards/education/credit-score/do-investments-affect-your-credit-score

[13]DeNicola, L. (2023, December 22). *Does buying stocks affect my credit score?*. Experian. https://www.experian.com/blogs/ask-experian/does-buying-stock-affect-credit-score/

[14]*Articles*. Equifax. (n.d.-c). https://www.equifax.com/personal/education/credit-cards/articles/-/learn/should-i-pay-off-my-credit-card-in-full-each-month/#:~:text=Carrying%20a%20monthly%20credit%20card,as%20you%20can%20each%20month.

[15]*What credit score do you need for a credit card*. Chase. (n.d.-b). https://www.chase.com/personal/credit-cards/education/build-credit/what-credit-score-is-needed-for-a-credit-card

[16]Does your income affect your credit score? | Chase. (n.d.-a). https://www.chase.com/personal/credit-cards/education/credit-score/does-income-affect-score

[17]Luthi, B. (2024, March 21). *How many bank accounts should I have?*. Experian. https://www.experian.com/blogs/ask-experian/should-i-have-multiple-bank-accounts/#:~:text=In%20general%2C%20bank%20accounts%20don,up%20on%20your%20credit%20report.

[18]*Articles*. Equifax. (n.d.-b). https://www.equifax.com/personal/education/credit-cards/articles/-/learn/how-many-credit-cards-should-i-have/#:~:text=Two%20factors%20that%20contribute%20to,of%20credit%2C%20are%20generally%20recommended.

[19]*Articles*. Equifax. (n.d.-a). https://www.equifax.com/personal/education/credit-cards/articles/-/learn/should-i-pay-off-my-credit-card-in-full-each-month/#:~:text=Carrying%20a%20monthly%20credit%20card,as%20you%20can%20each%20month.

[20]Axelton, K. (2024, January 30). *How does a personal loan impact your credit?*. Experian. https://www.experian.com/blogs/ask-experian/how-does-a-personal-loan-impact-your-credit/

[21]David, W. (2023, February 9). *Day 15: What is meant by CTC, gross and in hand salary?*. LinkedIn. https://www.linkedin.com/pulse/day-15-what-meant-ctc-gross-hand-salary-wicky-david#:~:text=It%20is%20also%20known%20as,contributions%2C%20and%20other%20mandatory%20deductions

[22]How to increase your credit limit (without harming your score) | credit cards | U.S. news. (n.d.-b).

https://money.usnews.com/credit-cards/articles/how-to-increase-your-credit-limit-without-harming-your-score

[23]READ, 6 MIN. (n.d.). *Understanding debt & credit scores*. American Medical Association. https://www.ama-assn.org/medical-residents/medical-residency-personal-finance/understanding-debt-credit-scores#:~:text=A%20credit%20score%20can%20range,less%20of%20their%20credit%20limits.

[24]Kagan, J. (n.d.). *Equated monthly installment (EMI): How it works, formula, examples*. Investopedia. https://www.investopedia.com/terms/e/equated_monthly_installment.asp

[25]*Articles*. What is a Credit Mix? - Benefits of Credit Diversity | Equifax®. (n.d.). https://www.equifax.com/personal/education/credit/score/articles/-/learn/what-is-a-credit-mix/#:~:text=Simply%20put%2C%20a%20credit%20mix,of%20calculating%20credit%20scores)%20used.

[26]*Credit card minimum payments: What to know*. Capital One. (n.d.). https://www.capitalone.com/learn-grow/money-management/credit-card-minimum-pay-explained/

[27]Google. (n.d.). *Google colaboratory*. Google Colab. https://colab.research.google.com/drive/1ZspCgINEKnRVD0UqIijF8I-_WZbR7y1i?usp=sharing

[28]Mali, K. (2024, January 23). *Everything you need to know about linear regression!*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

[29]*Sklearn.linear_model.linearregression*. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[30] Demir, N. (2021, December 28). *Understanding logistic regression*. Medium. https://blog.demir.io/understanding-logistic-regression-26802c0da856

[31]*Sklearn.linear_model.logisticregression*. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[32]YouTube. (2023, January 8). *Logistic regression [simply explained]*. YouTube. https://www.youtube.com/watch?v=C5268D9t9Ak

[33]Wikimedia Foundation. (2024, April 25). *Sigmoid function*. Wikipedia. https://en.wikipedia.org/wiki/Sigmoid_function#/media/File:Logistic-curve.svg

[34]Saxena, S. (2023, September 13). *Binary cross entropy/log loss for binary classification*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/

[35]Koli, S. (2023, February 28). *Decision trees: A complete introduction with examples*. Medium. https://medium.com/@MrBam44/decision-trees-91f61a42c724

[36] R, S. E. (2024, April 19). *Understand random forest algorithms with examples (updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[37] Biswal, A. (2023, November 7). *What is principal component analysis?*. Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis

[38] *Feature importances with a forest of trees*. scikit. (n.d.-a). https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

[39] Google. (n.d.). *Styled_cov_matrix .CSV*. Google Drive. https://drive.google.com/file/d/1U1KJbwYDuE5OwS1i6VOHuU7aj1ifTjCP/view?usp=sharing

[40] Google. (n.d.-a). *Cov_positive_negative.CSV*. Google Drive. https://drive.google.com/file/d/1XP58J8qRW--CwUzw28ER5-PL2W3YmQPm/view?usp=sharing

[41] Google. (n.d.-b). *Styled_corr_matrix .CSV*. Google Drive. https://drive.google.com/file/d/1Zs3UUExwrsLUjfhTyIzistIpDLMF0srG/view?usp=sharing

[42] Google. (n.d.-a). *COV_POSITIVE_NEGATIVE_CORR.CSV*. Google Drive. https://drive.google.com/file/d/133P0yif00UE5Yz94TyZ7MkBYJLVswfaj/view?usp=sharing