Customer Credit Risk Classification Analysis

Bank of Questrom

Team 1

—

Anshi Mittal - U98195152

Michelle Mengxue Lu - U05037816

Neri Ajiatas Arreaga - U03129799

David Euijoon Kim - U66545284

—

Boston University

Questrom School of Business

Business Decision-Making with Data

QST BA 305

Summer 1 2023

Dr. Huseyin Sami Karaca

—

June 29, 2023

Table of Contents

1 Introduction

1.1 Background

“Bank of Questrom” is a prominent financial institution with numerous products and services available for the general market. The bank has recently launched a revolutionary financial product called ‘BA305’ that provides customers with a new form of credit. Fortunately, the successful product marketing caused an influx of new applications waiting to be revised. However, due to a recent major failure of the credit bureaus in the United States, the bank cannot access the applicants' credit scores. Thus, the ‘Bank of Questrom’ has hired four data scientists to design a system to evaluate the applicant's credit risk. Additionally, the bank has provided historical credit data of its customers.

Following is the provided dataset: https://www.kaggle.com/datasets/parisrohan/credit-score-classification

1.2 Objectives

The credit system of the United States is managed by a complex network of institutions that track and record credit histories. The risk of lending loans or lines of credit to individuals and businesses is assessed through the credit system. Equifax, TransUnion, and Experian are the three major US credit bureaus that organize credit histories and calculate scores to provide a quick summary to lenders. Credit scores are usually based on payment histories, outstanding debts, and the number/amount of open or closed credit.

Since wealth is involved, the credit system is an important and sensitive matter for the majority. Such an influential score is calculated with a complex algorithm. Therefore, it is quite unintelligible without the help of computational tools. This brings into question the influencing factors of credit scores.

The main objective of the credit risk classification study is to identify clients with high credit risk. This will allow the ‘Bank of Questrom’ to provide the financial product to the appropriate clients. Utilizing various methodologies to analyze the critical credit risk variables will help facilitate the banks’ decision-making.

The following questions will be answered in the study of credit risk classification:

1. What are the most important features in classifying someone with high credit risk?
2. What are the characteristics of an average customer with low or high credit risk?
3. What are the optimal benchmarks for the application approval and rejection?

2 Data

2.1 Description

A total of 100,000 observations are available in the dataset. Each observation represents a client's financial data in a particular month, thus considered a ‘pooled data’ or a combination of ‘time series’ and ‘cross-section.’ Specifically, the dataset covers eight months of observation for all 12,500 unique customers. There are 28 variables, which include personal identification information and detailed financial attributes. The dependent variable, ‘Credit_Score’ is segmented into three categories: ‘Good,’ ‘Standard,’ and ‘Poor.’ The observations classified as ‘Poor’ credit scores are the main focus of the analysis.

2.2 Pre-processing

2.2.1 Cross-Sectional Data

Considering the lack of fluctuation in credit risks throughout the available months, a decision is made to transform the pooled data into cross-sectional data. Thus, observations with month values other than ‘August’ are removed. The remaining 12,500 rows are further processed before the analysis.

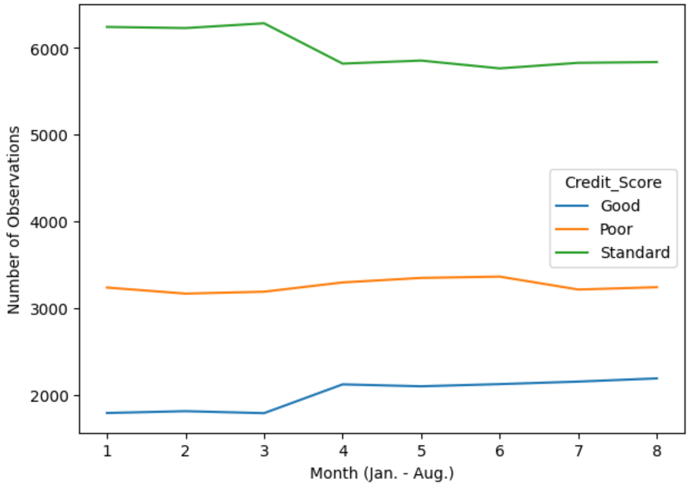


Figure 1. Changes in Credit Score throughout Months

2.2.2 Feature Selection

As the dataset is transformed into a cross-sectional dataset, the ‘Month’ feature and the other seven columns solely relevant to the time series are dropped. Additionally, four identification variables of the clients are removed to prevent overfitting the ML model and, most importantly, to ensure personal data protection. The ‘Occupation’ feature, which contains 15 jobs excluding missing values, is also removed after iterative modification of the ML model as it lacked importance compared to the disadvantages of dummy coding.

2.2.3 Missing Values

Since the values of each observation are crucial in determining the credit risk, an imputation with assumptions or averages should be avoided, as such acts will lead to flaws and misinterpretations. Thus, the rows with missing values are removed entirely.

Different types of missing values are present in the observations. The majority of missing values are already specified as ‘null’ or ‘nan,’ simplifying the removal process. However, the unique values inside each variable are thoroughly examined to find missing values with substituted strings, such as ‘_.’

2.2.4 Reformatting

Multiple numerical variables were assigned string types and had non-numerical characters such as an underscore intertwined with the real numerical value. The non-numerical values were filtered out, while the cleaned variables were converted to integers and floats accordingly. Significant reformatting occurred in ‘Delay_from_due_date,’ ‘Credit_Mix,’ and ‘Credit_History_Age.’ There were negative values in the delayed dates, which indicated early payments. On-time payments should be considered equal and early payments shouldn’t diminish the negative effects of delayed payments. Thus, negative values in ‘Delay_from_due_date’ were transformed into zeros. Credit_Mix was the categorical feature that was dummy-coded to act as a numerical variable throughout the models used in the study. The Credit_History_Age column was originally written in ‘00 years and 00 months’ format, which was recalculated into the number of months ranging from approximately 0 to 400.

2.2.5 Outliers

Outliers usually have a significant impact on statistical analysis and often skew the final results. Thus, each numerical variable is examined to remove extreme outliers. Outliers in the features below were removed as they accounted for less than 1% of the normalized value counts. The results are as follows:

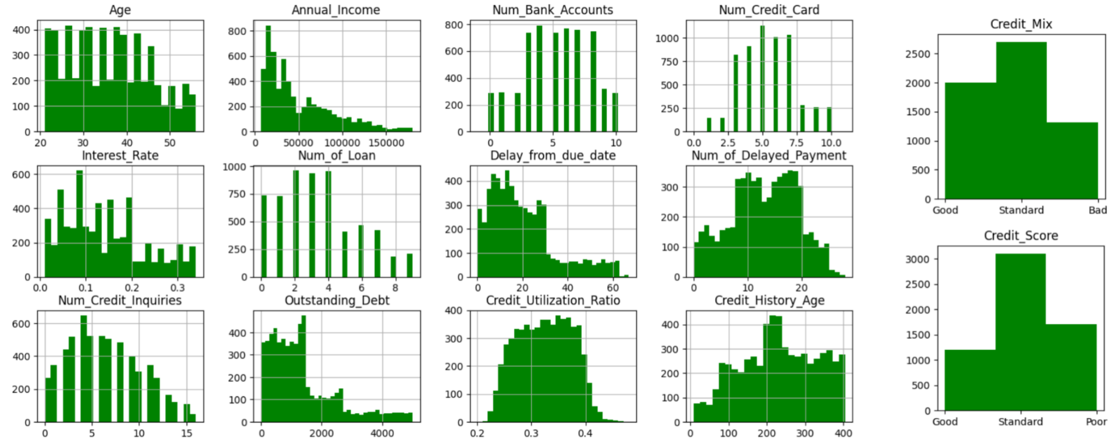
1. The age of customers ranges between 21 and 100.
2. The highest annual income is 200,000 USD, decreased from the prior max of 20 million USD.
3. The number of bank accounts was optimal between 1 and 12.
4. The number of credit cards is below 21, while that of loans is below 22.
5. The interest rate does not go over 41 percent.
6. The number of delayed payments is lower than 29 and lower than 17 for credit inquiries.

2.2.6 Dependent Variable

The study consists of two methodologies, and the dependent variable is the ‘Credit_Score’ feature for both. The k-means clustering will be performed with the original polytomous dependent variable with three unique values. However, the decision tree model will be conducted with the dependent variable transformed into a binary variable with the ‘Standard’ and ‘Good’ set as 0, while ‘Poor’ is specified with 1.

2.2.7 Cleaned Data

Figure 2. Distribution of Variables



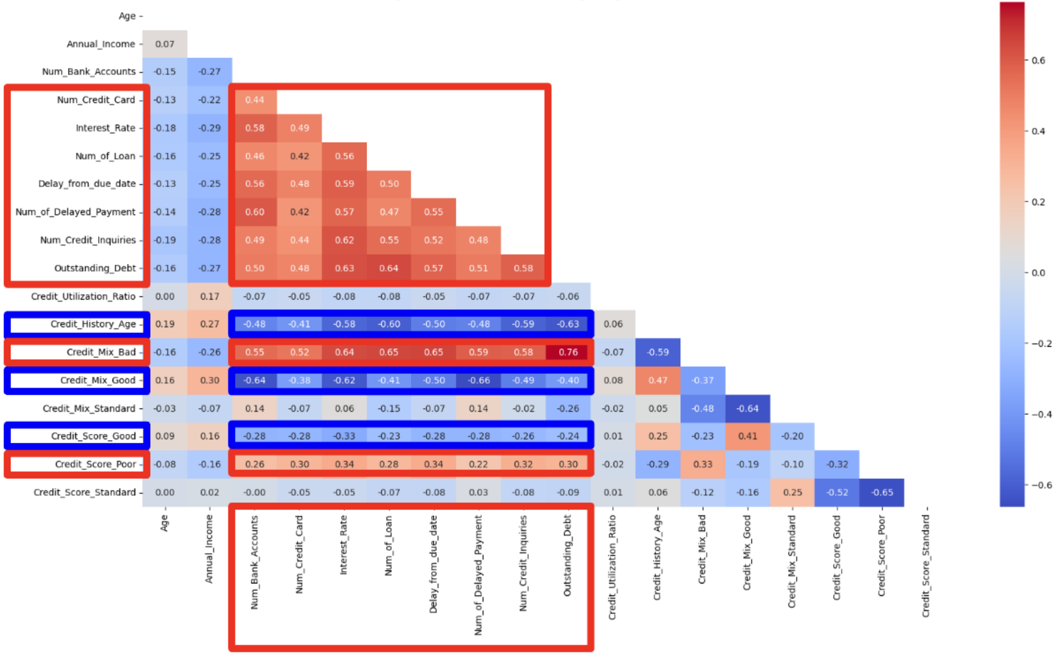
Overall, the cleaned dataset contained 6015 observations with 1 dependent and 13 independent variables, which are 2 categorical and 12 numerical features. Additional pre-processing, such as standardized scaling, is performed before implementing k-means clustering. The histograms above display the distribution of values within the processed features.

2.3 Correlations

2.3.1 Correlation Matrix

Briefly analyzing the correlation matrix in Figure 3, there were two evidently opposing groups.

Figure 3. Correlation Matrix including Categorical Variables



2.3.2 Negative Influence on Credit Score

The first group consists of the variables that are highlighted in red boxes. The features in the group had a moderate positive correlation with each other. Since the variables are also positively correlated with the 'Poor' credit score, an assumption is constructed:

1. As the following features increase, the credit risk is more likely to increase. Thus, applications from individuals with high values in these categories should be flagged for close examination.
   a. Number of credit cards
   b. Interest rate
   c. Number of loans
   d. Delay from the due date
   e. Number of delayed payments
   f. Number of credit inquiries
   g. Outstanding debt
   h. Bad credit mix

2.3.3 Positive Influence on Credit Score

The apparent opposite of the first, the second group includes credit history age and good credit mix. These two variables are positively correlated to each other and also with the 'Good' credit score. Therefore, another assumption is constructed:

2. As the following features increase, the credit risk is more likely to decrease. Thus, individuals with high values in these categories and low values in the variables from 2.3.2 should be considered the standard for individuals with low credit risk.
   a. Credit history age
   b. Good credit mix

3 Methodologies

3.1 Decision Tree

3.1.1 Purpose

Decision tree classifier is a supervised machine learning algorithm with a hierarchical tree-like model of decisions leading to distinct outcomes. The decision tree model enables the interpretation of the features influencing the classification. Therefore, the influential factors are identified in the order of importance, answering the first question.

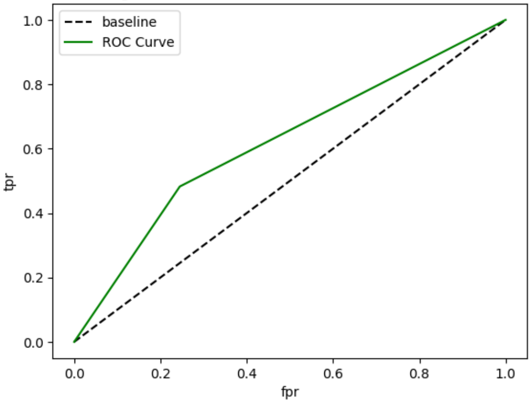
1. What are the most important features in classifying someone with high credit risk?

3.1.2 Implementation

To determine customers with high credit risks, the ‘Good’ and ‘Standard’ values of the ‘Credit_Score’ variable are merged as 0 while the ‘Poor’ value is redefined as 1. As a result, the outcome variable for the decision tree is binary.

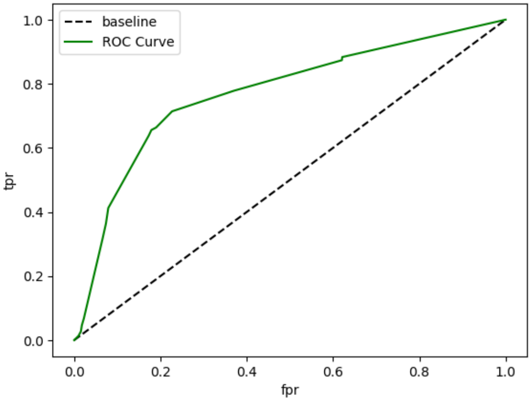
Before implementing the decision tree, categorical variables are dummy-coded into numerical variables. Additionally, the dataset is split into train and test sets to evaluate the model.

Figure 4.1 Full Decision Tree Model ROC curve



Using the Gini index criterion to split the nodes, a full tree had an accuracy of 67.76% and an f1-score of 46.01%. With the optimal hyperparameters tuning with the grid-search tool, the accuracy increased to 77.62% and the f1-score to 51.21%. Furthermore, as shown in Figures 4.1 and 4.2, the area under the Receiver Operating Characteristic (ROC) curve has expanded for the reduced tree, which indicates improvement in the model. However, the model's sensitivity in identifying the high credit risk is extremely low with a value of 41.25%, which imposes evaluation concerns.

Figure 4.2 Reduced Decision Tree Model ROC curve



3.2 K-Means Clustering

3.2.1 Purpose

K-means clustering is an unsupervised machine learning algorithm that classifies unlabeled data into clusters or groups. The clusters have similar characteristics and patterns for every variable in the data. The k-means clustering algorithm is utilized in the study to compute the assumed average features of each credit risk class. Therefore, the second question of the study can be answered.

2. What are the characteristics of an average customer with low or high credit risk?

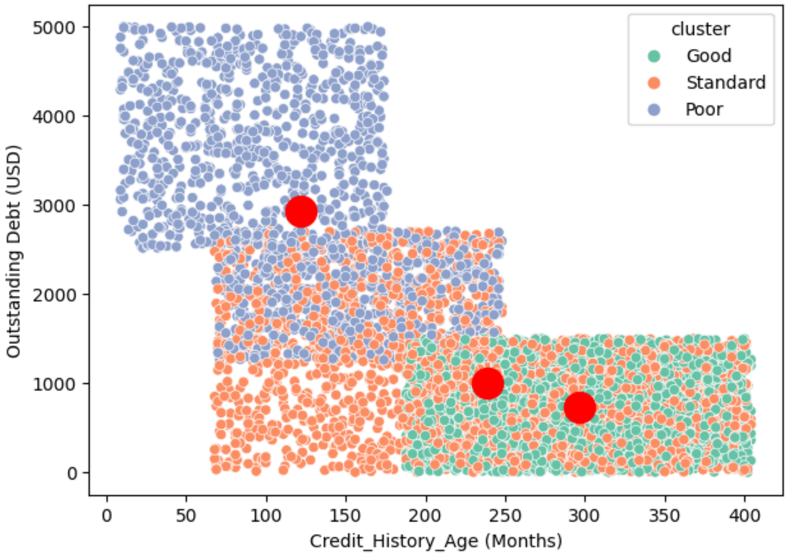
3.2.2 Implementation

The variable k in the k-means algorithm is the number of clusters. In the study, as three unique values are present in the dependent variable 'Credit_Score,' three clusters should be created, and therefore, k should be equal to 3. Since k-means clustering requires all variables to be numerical, the categorical variables are converted to dummy variables. Also, standardization or standard scaling of all values is another crucial step before executing the k-means clustering procedure.

The three clusters hold centroids, the arithmetic means or the centers of the data points assigned to each cluster. Analyzing the feature values within the centroid of each cluster is the approach in the study to discovering the characteristics of customers in each credit risk class.

Figure 5 illustrates the three clusters and their centroids (big red dot) for 'outstanding debt' and 'credit history age.' *The clusters appear to overlap in the two-dimensional figure because the k-means clustering was operated on multiple variables, producing multi-dimensional clusters.

Figure 5. K-Means Clustering Visualization for Two Features



4 Analysis

4.1 Decision Tree

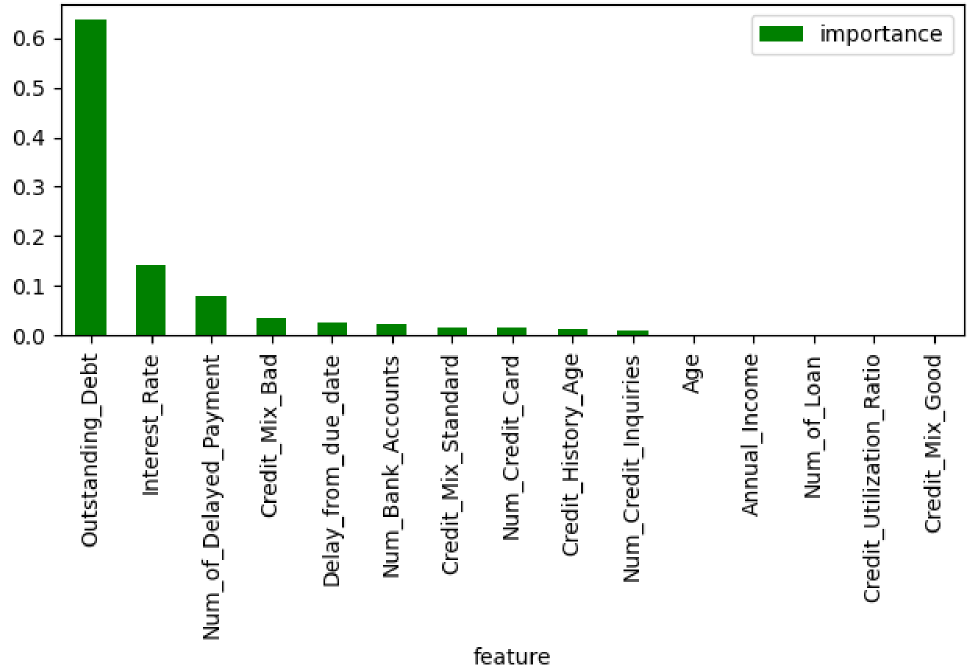
4.1.1 Feature Importance

1. What are the most important features in classifying someone with high credit risk?

Once the decision tree is implemented, the importance of each feature is calculated and utilized in order. The most significant feature is the amount of outstanding debt. After that, the feature order is as follows: Interest Rate, Num_of_Delayed_Payment, Credit_Mix_Bad, Delay_from_due_date, Num_Bank_Accounts, Credit_Mix_Standard, Num_Credit_Card, Credit_History_Age, and Num_Credit_Inquires.

The importance of the features exemplifies which characteristics to examine thoroughly when identifying someone with high credit risk. Also, it specifies the order of variables the machine learning model uses to split decisions before reaching an outcome of rejecting or approving an application. The order is further utilized in section 4.2 of the study when analyzing the attributes of average individuals within the 'Credit_Score' classes.

Figure 6. Decision Tree Feature Importance



4.1.2 Series of Decisions

The decision tree model employs the feature importance to split the observations based on a series of decisions. The reduced decision tree model trained for the study concluded with 29 total nodes and 15 leaf nodes. Out of the 15 leaf nodes, 7 had an outcome of 1 or 'Reject.'

The leaf node with the lowest Gini index of 0.346 or the lowest impurity score for the 'Reject' result had the following decisions:

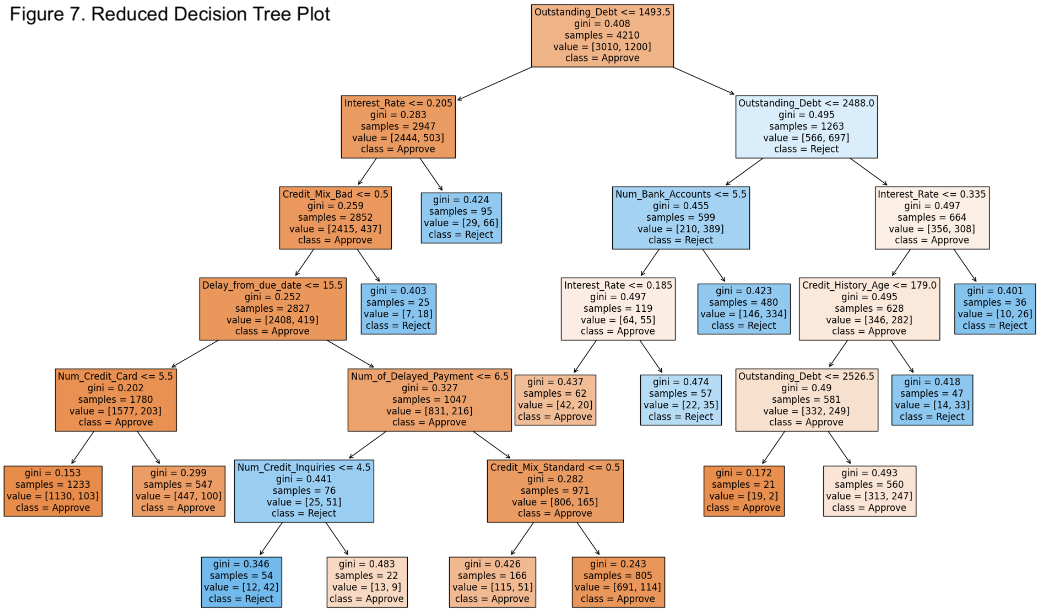
1. Outstanding debt of less than or equal to 1493.5 USD.
2. Interest rate of lower than or equal to 20.5%.
3. Bad credit mix probability smaller than or equal to 0.5.
4. Delays from the due date higher than 15.5 days.
5. Number of delayed payments less than or equal to 6.5.
6. Number of credit inquiries less than or equal to 4.5

Based on the correlation analysis assumptions in section 2.3 of the study, this observation's characteristics seem closer to an individual with low credit risk. Despite the highest purity, the lack of observations and the low sensitivity metric may be causing flaws in the model.

The following series of decisions for the 'Reject' outcome was considered the most intuitive and corresponded to the assumptions made in section 2.3.

- Outstanding debt higher than 2488 USD and interest rate higher than 34%.
- Outstanding debt between 1493.5 and 2488 USD and the number of bank accounts higher than 5.5.

Figure 7. Reduced Decision Tree Plot



4.2 K-Means Clustering

4.2.1 Centroids

Table 1. Centroids of Clusters 1, 2, 3 with Standardized Feature Values

	Num_Bank_Accounts	Num_Credit_Card	Interest_Rate	Delay_from_due_date	Num_of_Delayed_Payment	Num_Credit_Inquiries	Outstanding_Debt	Credit_History_Age
1	-0.90	-0.54	-0.88	-0.71	-0.94	-0.70	-0.56	0.67
2	1.05	1.01	1.22	1.26	1.05	1.11	1.38	-1.09
3	0.12	-0.13	0.01	-0.15	0.15	-0.07	-0.32	0.09

The values in Table 1 represents the standardized feature values of the centroids within each cluster. Eight features are selected to examine deeper based on the feature importance indicated by the decision tree in section 4.1.1. Based on the assumptions made on the correlation analysis (2.3), the three clusters were classified as ‘low credit risk,’ ‘high credit risk,’ and ‘standard credit risk’ in sequential order. The centroids of the three clusters will be the key to answering the second question of the study:

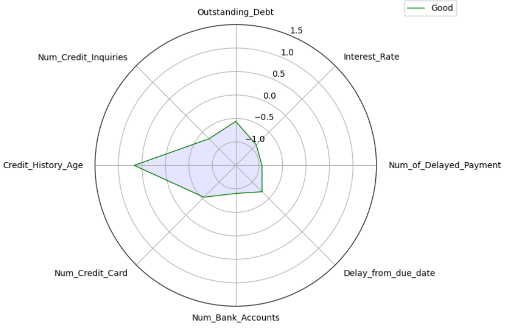
2. What are the characteristics of an average customer with low or high credit risk?

4.2.2 Low Credit Risk

The first cluster is considered as the group of customers with low credit risk or high credit scores. The cluster's centroid indicates that the average customers in the group have the following characteristics:

** The term* ***‘total average’*** *is the arithmetic mean of all values in a specific feature.*

Figure 8.1 Radar Chart of Cluster No. 1



1. **Lower number of bank accounts** than the total average number of bank accounts.
2. **Lower number of credit cards** than the total average number of credit cards.
3. **Lower interest rate** than the total average interest rate.
4. **Lower delayed payment days** than the total average delayed payment days.
5. **Lower number of delayed payments** than the total average number of delayed payments.
6. **Lower number of credit inquiries** than the total average number of credit inquiries.
7. **Lower amount of outstanding debt** than the total average amount of outstanding debt.
8. **Higher credit history age** than the total average amount of credit history age.

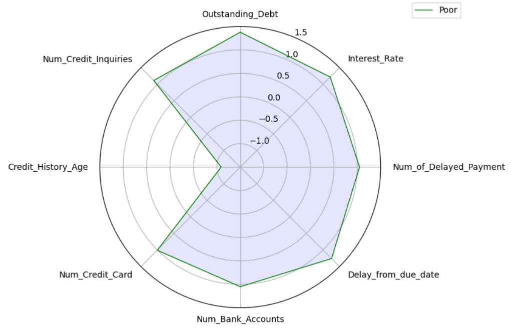
The average people with low credit risk have a more extended credit history with fewer debts, credits, and delays. Moreover, the credit history age seems extremely influential in defining the group.

4.2.3 High Credit Risk

The second cluster is assumed as the group of customers with high credit risk or low credit scores. The cluster's centroid indicates that the average customers in the group have the following characteristics:

** The term* ***'total average'*** *is the arithmetic mean of all values in a specific feature.*

Figure 8.2 Radar Chart of Cluster No. 2



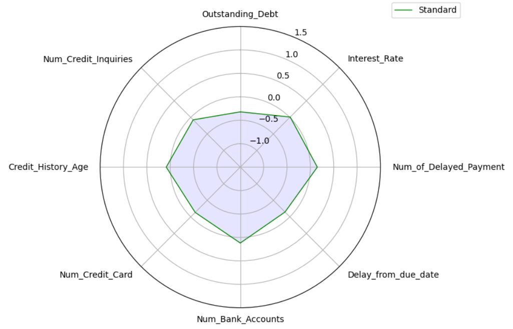
1. **Higher number of bank accounts** than the total average number of bank accounts.
2. **Higher number of credit cards** than the total average number of credit cards.
3. **Higher interest rate** than the total average interest rate.
4. **Higher delayed payment days** than the total average delayed payment days.
5. **Higher number of delayed payments** than the total average number of delayed payments.
6. **Higher number of credit inquiries** than the total average number of credit inquiries.
7. **Higher amount of outstanding debt** than the total average amount of outstanding debt.
8. **Lower credit history age** than the total average amount of credit history age.

The average individuals with high credit risk have a shorter credit history with high numbers of debts, credits, and delays. The group would be considered the complete opposite of the first cluster.

4.2.4 Standard Credit Risk

The third cluster is regarded as the group of customers with standard credit risk or standard credit scores. The cluster's centroid indicates that the average customers in the group have values close to the average of most features. In other words, the cluster's characteristics are positioned between the first two clusters. However, the average amount of outstanding debt was lower than the total average, which would be closer to someone with a low credit risk. This implies that the amount of outstanding debt significantly differentiates someone with high or low credit risk. Thus, it provides evidence for the number one feature importance being the 'Outstanding_Debt' feature from the decision tree model analysis in section 4.1.1.

Figure 8.3 Radar Chart of Cluster No. 3



5 Conclusion

5.1 Result

5.1.1 Analysis

1. The amount of outstanding debt and the interest rate are the most important features which should be prioritized in examining a person's credit risk. Thus, high outstanding debt and high-interest rates are major red flags.
2. Excessive numbers of delayed payments, accounts, credits, and inquiries can also be a red flag.
3. Lengthy credit history age is a good indicator of low credit risk. However, high values in other negative factors may diminish the positive effect of the credit history age.

5.1.2 Optimal Benchmark

When finding the favorable standard for the last question of the study, understanding which factors differentiate a standard credit risk and a high credit risk in the k-means clustering is vital. Additionally, the judgments made in the decision tree should also be scrutinized to determine the benchmarks.

3. What are the optimal benchmarks for the application approval and rejection?

The most notable differences between the standard risk cluster and the high credit risk cluster were in the amount of outstanding debt and the delay from the due date. The most important factors in the decision tree were the outstanding debt and interest rate. Therefore, after examining the destandardized centroids and the decision tree, the optimal benchmark is set to reject applications of individuals who have outstanding debt amount higher than 1900 USD, interest rate higher than 20%, and delay from the due date higher than 28 days.

5.1.3 Testing the Model

The destandardized values (excluding Credit_Score) of the centroids within three clusters are provided as test data to the decision tree model of the study. The model output an array of [0,0,0], which indicates the failure of rejecting the centroid of the high credit risk cluster.

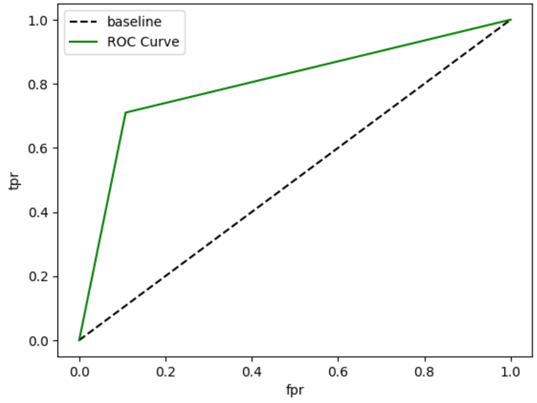
The reasoning behind the failure is the deficient f1 score of 51.21% and sensitivity metric of 41.25%. Sensitivity is the model's ability to predict true positives, in this case, the observations with high credit risk. Only 212 observations out of 514 observations with high risk credit were predicted correctly in this model.

5.2 Improvement

5.2.1 Training with Additional Data

To improve the sensitivity of the decision tree model, an attempt was made to increase the number of data for model training. Instead of including only 'August' for the 'Month' variable at the beginning of the study, data from all eight months are fitted. As a result, the cleaned data advanced from 6015 to 47,900 observations. Once the training was completed, the accuracy of the new model was enhanced from 77.62% to 84.18%. More importantly, the f1 score grew from 51.21% to 71.35%, and the sensitivity increased from 41.25% to 71.01%.

Figure 9. Improved Decision Tree Model ROC curve



5.2.2 Testing the Model

The improved decision tree model successfully identified the cluster centroid with high credit risk - an array output of [0,1,0]. The result implies the model is capable of capturing the average individual with high credit risk. In conclusion, the improved credit evaluation system will be sufficient in rejecting most applications of individuals who may be high risks to the bank.

6 Appendix

6.1 Codes

6.1.1 1_credit_data_cleaning.ipynb

- Data cleaning file with the dependent variable 'Credit_Score' as a polytomous variable
- Unique values of dependent variable: Good, Standard, Poor

6.1.2 1_credit_data_cleaning2.ipynb

- Data cleaning file with the dependent variable 'Credit_Score' as a binary variable
- Unique values of dependent variable: 0 (Includes Good and Standard), 1 (Poor)

6.1.3 2_credit_data_correlation.ipynb

- File for finding correlations between variables via correlation matrix heatmap.

6.1.4 3_credit_knn.ipynb

- KNN classification model with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.5 3_credit_knn2.ipynb

- KNN classification model with cleaned_credit_data2.csv (Binary dependent variable)

6.1.6 4_credit_decision_tree.ipynb

- Decision tree model with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.7 4_credit_decision_tree2.ipynb

- Decision tree model with cleaned_credit_data2.csv (Binary dependent variable)

6.1.8 5_credit_kmeans.ipynb

- K-means clustering with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.9 experiment.ipynb

- File for experimental improvement stage for the study
- Increased number of observations by including all months instead of only 'August'
- Decision tree model with binary dependent variable

6.2 Data

6.2.1 credit_raw.csv

- Original unprocessed data from the following url
- https://www.kaggle.com/datasets/parisrohan/credit-score-classification
- 100,000 Entries with 28 columns

6.2.2 cleaned_credit_data.csv

- Processed data with the dependent variable 'Credit_Score' as a polytomous variable
- 6015 Observations with 2 categorical and 12 numerical variables

6.2.3 cleaned_credit_data2.csv

- Processed data with the dependent variable 'Credit_Score' as a binary variable
- 6015 Observations with 2 categorical and 12 numerical variables

6.2.4 centroids.csv

- Centroids dataframe saved from '5_credit_kmeans.ipynb'
- Standardized feature values of centroids within the three clusters

6.2.5 centroids_test.csv

- Centroids dataframe used for testing the decision tree models
- Unstandardized feature values of centroids within the three clusters