Credit Score Exploration and Prediction

Yichen Yao, Yi Mao, Yuxuan Pan, Meng Li, Yaqi Zheng

1. Introduction

Default is a highly bothersome problem in the financial industry which causes huge losses for the lender companies. Thus, it is essential for the financial institutes to estimate the default risk of its customers to minimize the uncertainty in receiving the loan repayments. In this project, after conducting the data cleaning, we apply different machine learning models (logistic regression, KNN, decision tree, random forest) to the customers' credit-related data to classify them into different credit score categories.

2. Data Description

The data was collected from a finance company (Paris, 2022). The original dataset contains 50,000 observations with 27 features. Some of these features are useless in our machine learning models and thus we will remove them in the feature selection section. We use the extracted features to classify each customer's credit as good, standard, or poor.

3. Data Cleaning

1) Unify the data format

Some data are in invalid format; for example, age is 28_. We define a function to strip away the underscores attached with the values. For some variables that are supposed to be numbers, we converted its format from a string containing other words to a clean number

2) Make the data consistent and modify the unreasonable data

Some features, such as SSN and occupation should be consistent for a certain customer. Thus, we define a function to tackle the inconsistency in these columns. There also exist some entry errors in the dataset (e.g. the number of credit cards for a customer is 1385). We modify these entry errors based on the remainder of the data related to the customer.

3) Impute the missing values

The dataset also contains some missing values. We impute the missing values with the mode or mean of the existing information corresponding to the specific customers. After our processing, the distribution of the data is improved to a great degree (Eg: Figure 5 & 6 in the appendix)

4) Converting category data into numeric values

In the conversion process, we use one_hot_encoder for variables that have no magnitude relationship. On top of that, if there is still a quantity for each split-out variables, the specific value is obtained, instead of just "0-1," eg: "Type of loans" column is converted to 10 columns

4. Exploratory Data Analysis

Firstly, we take a general look at the collated data. In terms of sample diversity, the dataset satisfies diversity requirements well. The distribution of 'Occupation' in the dataset shows that the dataset covers various occupations and the number of people in each industry is not very different, with relatively more lawyers. (Figure 7 in the appendix)

Secondly, as for the main target, credit scores, the percentage/amount distribution is shown as follows. The percentages of Credit Score 1,2 and 3 (poor, standard, good) are 28.98%, 53.19% and 17.83% respectively. (Figure 8 in the appendix)

After knowing about the percentage, we want to know about the geographical distribution of credit scores in the USA. However, there is no lat/lon data in the original dataset. In this case, we adopted the SSN to find out the location.

1) Step1: Use web scraping (requests/bs4/re) to obtain the relationship between states and SSN. Because of the first 3 digits of the SSN assigned throughout the United States and its possessions, we can use SSN to obtain a matched state name. Then match names with abbreviations by 'state-abbrevs.csv'. (URL: https://www.ssa.gov/employer/stateweb.htm)
2) Step 2: Group by the state and calculate the mean of Credit scores.
3) Step 3: Use folium to map credit scores.

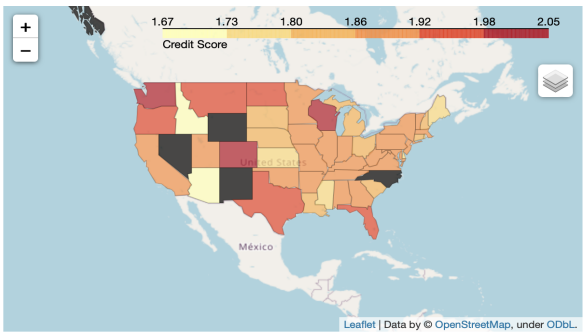

Figure 1. Credit Scores location distribution

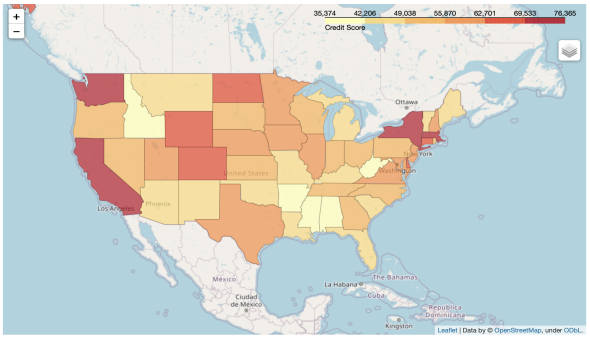

Figure 2. GDP per capita location distribution

From Figure 9, HI (Hawaii) has the highest credit score and ID (Idaho) has the lowest credit score. From Figure 1, it's obvious that credit scores have greater variation in the West and they are more evenly distributed in the East. As for the reason, we think it may be related to the economic level of different regions. From the macroeconomic side, we used GDP per capita for mapping. After mapping the GDP per capita geographical distribution (Figure 2), it's found that it matches with Figure 1 to some extent.

As for the microeconomic level, we visualize the individual variable in our dataset that may affect the Credit Scores by observing the variables in the dataset. It's obvious that the financial level may affect credit scores. Therefore, we take two variables, "Monthly_Inhand_Salary" and "Outstanding_Debt", into consideration.

From the below boxplots, it's obvious that higher monthly in hand salary distribution is accompanied with higher credit scores. As for the outstanding debt, the relationship is the opposite. Then it comes to the quantitative analysis part, which includes Machine learning.

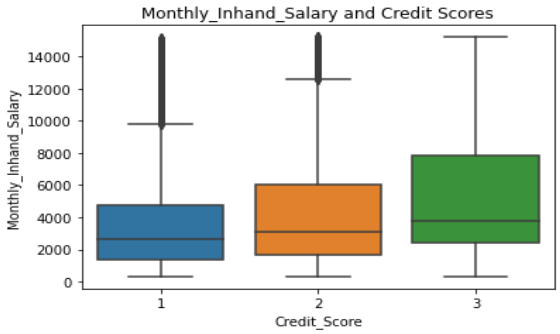

Figure 3. Monthly_Inhand_Salary & Credit Scores

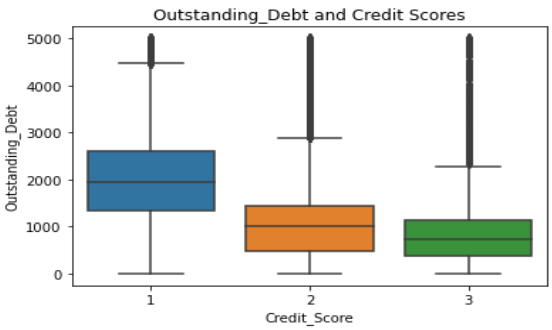

Figure 4. Outstanding_Debt & Credit Scores

5. Machine Learning

5.1 feature selection

We try to figure out what kinds of information may influence the credit score of each person, so that we can know how to get a higher credit score as a customer. To deal with this problem, we use feature selection for four different models, and verify the effect by using the selected features to predict the credit score, and comparing it to the result of using all features.

Logistic Regression: Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. To select features, we first fit the logistic regression model with full features on the training dataset, and then take the coefficients in the model as the importance of the corresponding feature. Then, by sorting the importances, we select the ten most important features to fit the logistic regression/Figure 10/. By comparing the accuracy of prediction between using selected features and all features, we can find that the model using selected features has a higher score/Figure 11/. However, since the overall accuracy score is relatively low, it hints that logistic regression might not be the best model for view prediction due to correlation between our independent variables.

KNN: K- nearest neighbor is a supervised machine learning algorithm that can be used for classification and regression problems. To select features, we use the algorithm called forward stepwise selection to select a feature set with size ten./Figure 12/. By comparing the accuracy of prediction between using selected features and all features, we can find that the model using selected features has a higher score/Figure 13/.

Decision trees: Decision trees are a type of supervised learning algorithm where data will continuously be divided into different categories according to certain parameters. To select features, we first fit the decision trees model with full features on the training dataset, and then take the attributes 'feature_importances_' of the model as the importance of the corresponding feature. Then, by sorting the importances, we select the ten most important features to fit another decision tree model/Figure 14/. By comparing the accuracy of prediction between using selected features and all features, we can find that the model using selected features has a higher score/Figure 15/.

Random Forests:Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. To select features, we first fit the random forests model with full features on the training dataset, and then take the attributes 'feature_importances_' of the model as the importance of the corresponding feature. Then, by sorting the

importances, we select the ten most important features to fit another random forest model/Figure 16/. By comparing the accuracy of prediction between using selected features and all features, we can find that the model using selected features has a higher score/Figure 17/.

Overall, we notice that all of these four models select features (‘Outstanding_Debt’, ‘Changed_Credit_Limit’, and ’Interest_Rate’), which implies that these three features have more influence on the credit score than others. Moreover, we find that random forest has the highest score(0.78) on test data, which means that the features selected in this method are more likely to influence the credit score.

However, the accuracy of the model is also influenced by its parameters. To make it more convincing, we then do the parameters tuning on these four models.

5.2 Model parameters tuning

After doing feature selection, we intend to explore the feature selection's influence on the model and achieve a higher prediction accuracy for every model we chose. We first visualize the impact of the main parameter’s change to get the range for parameters. Then, we tune models through using sklearn.model_selection.GridSearchCV. Detailed visualizations for parameter tuning are shown in the Appendix figure 21-25. We summarize the result in table1 in the appendix. The highest accuracy we can achieve is 80.53% through using a random forest model with all features.

6. Conclusion

From the macroeconomic side, credit scores are related with GDP per capita and 'Interest_Rate'; from the microeconomic side, the most important features are 'Outstanding_Debt', 'Credit_Mix', 'Delay_from_due_date', 'Num_Credit_Inquiries', 'Num_Credit_Card', 'Credit_History_Age', 'Payment_of_Min_Amount', 'Num_Bank_Accounts', 'Changed_Credit_Limit', and 'Num_of_Delayed_Payment' from Random Forest Model (80.53% Precision).

We can see that credit scores have little relationship with the occupation itself (not including salary) and the loan type. Therefore, our prediction models can be used for different loan types (appear in the dataset) and help financial institutes to estimate the default risk of its customers.

7. Limitation and Future Works

1) **Improvements after presentation for feature selection**

We used cross validation to select features for each model above by calculating their R2 before. By sorting the score of R2 for each feature, we only select the features with positive scores. However, we find that some features with negative scores are supposed to be important to the prediction under the guidance of the professor.

Therefore, we try to use other valid methods to select features. For logistic regression, we take the coefficients of features as their importance. For decision trees and random forest, we use the attributes ‘feature_importances_’ of the model to select features. For KNN, we use the algorithm called forward stepwise selection to select a feature set. There are some common features along all selected features by four models, which seems to be more convincing.

2) **Future Improvements for feature selection**

The data after cleaning has too many features to select, especially after cleaning the column ‘Occupation’, which is really time-consuming for KNN to do feature selection. Therefore, we just use the data before cleaning the column ‘Occupation’ to select features on KNN.

Work Cited

Paris, R. (2022, June). *"Credit score classification"*. Kaggle. Retrieved December 17, 2022, from https://www.kaggle.com/datasets/parisrohan/credit-score-classification

SSA.gov *"Social Security Number Allocations"*. Retrieved December 17, 2022, Web scraping URL: https://www.ssa.gov/employer/stateweb.htm

Appendix

	All Features		**Selected Features**	
	Parameters	Test Precision	Parameters	Test Precision
Logistics Regression	C = 0.01 Penalty = l2	0.5384	C = 0.01 Penalty = l2	0.6266
KNN	leaf_size =1 n_neighber=3	0.7675	leaf_size =30 n_neighber=3	0.7765
Decision Tree	max_depth=10	0.7166	max_depth=20	0.7141
Random Forest	min_sample_splite=10 n_estimator=100	0.8053	min_sample_splite=10 n_estimator=100	0.7939

Table 1. ML results after parameter tuning

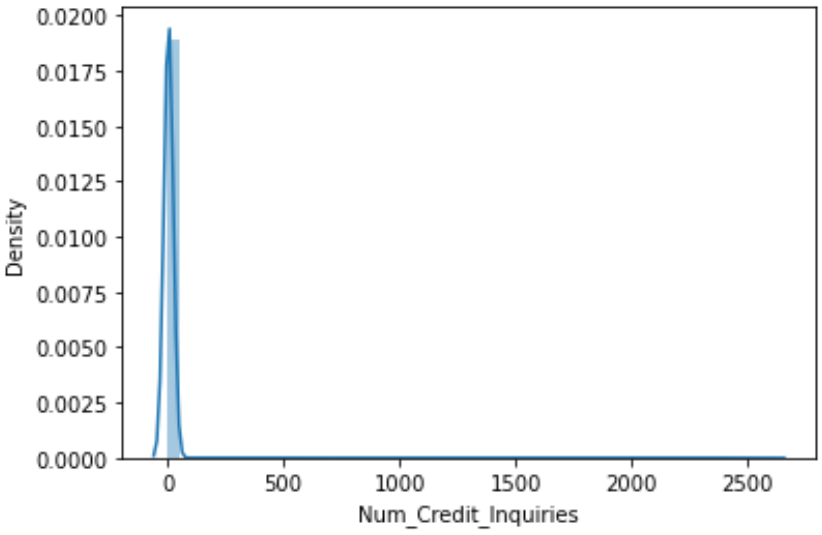

Figure 5. Distribution of “Num_Credit_Inquiries” (before cleaning)

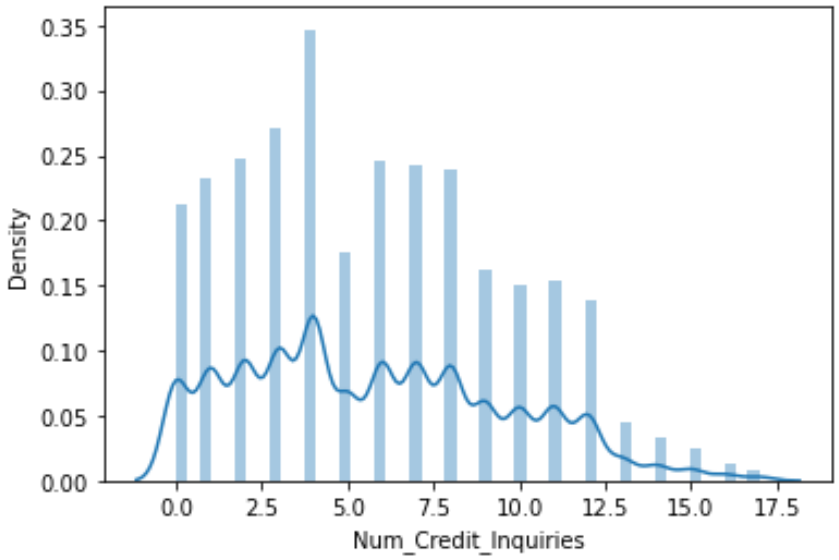

Figure 6. Distribution of "Num_Credit_Inquiries" (after cleaning)

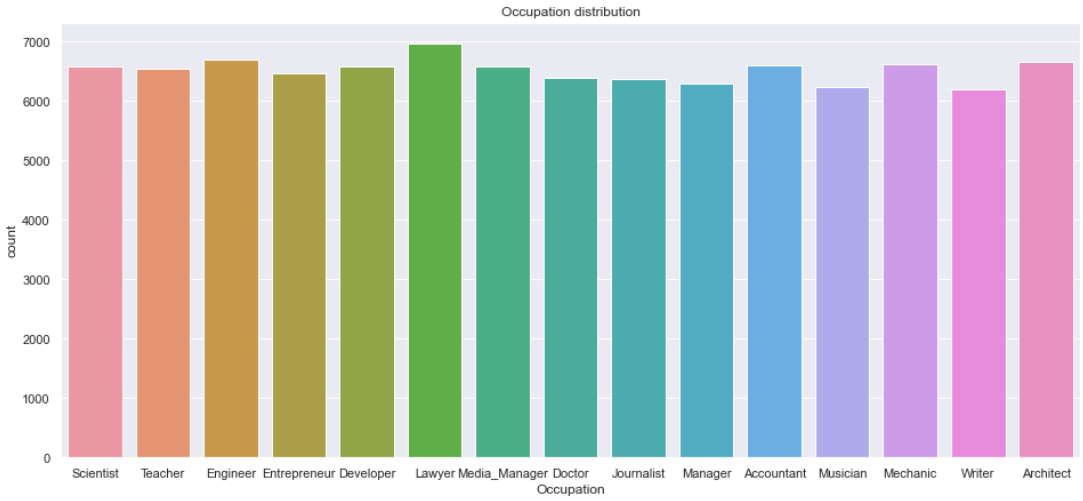

Figure 7. Occupation distribution

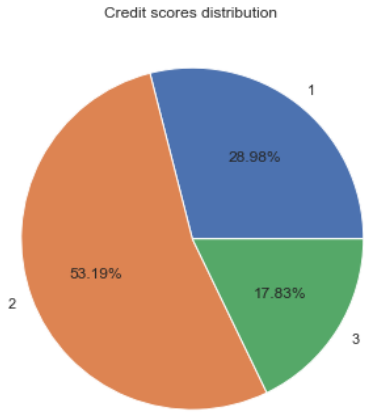

Figure 8. Credit Scores percentage distribution

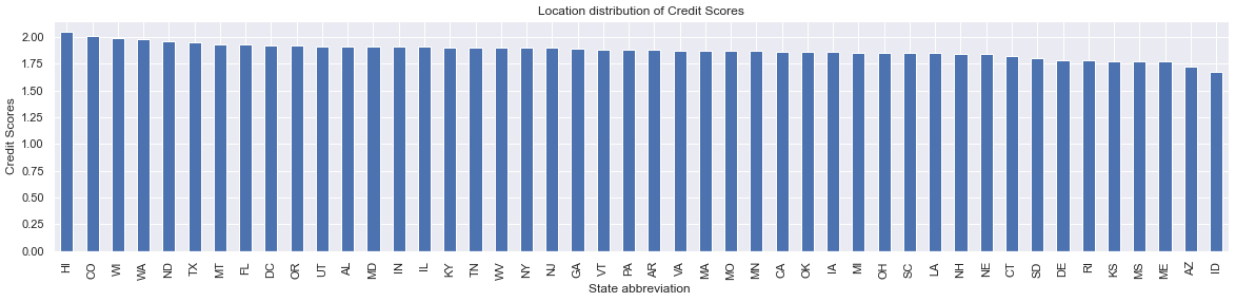

Figure 9. Credit Scores location ranking

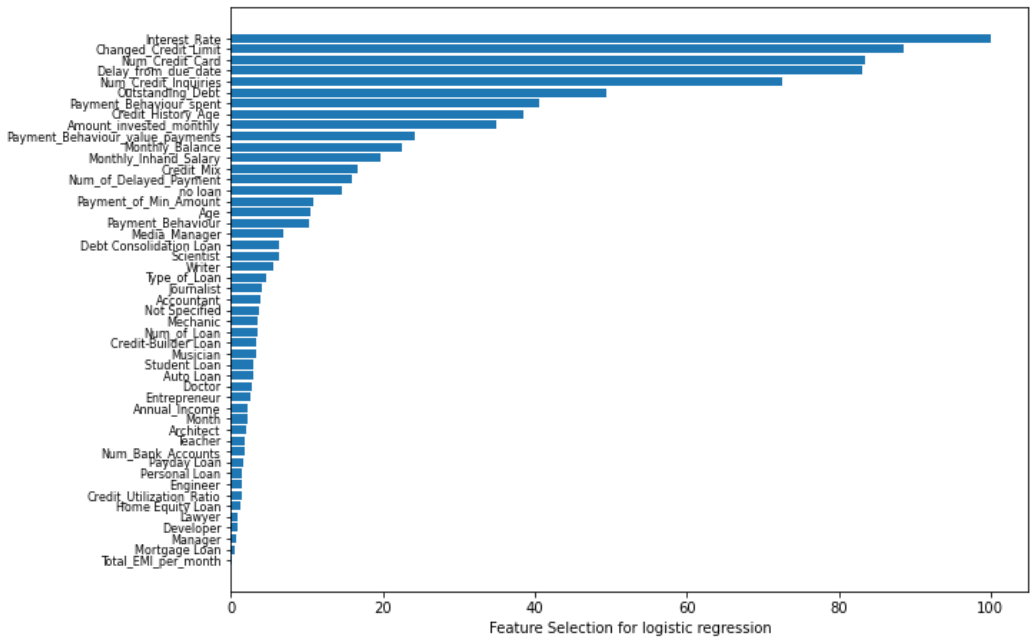

Figure 10. Feature selection for logistic regression

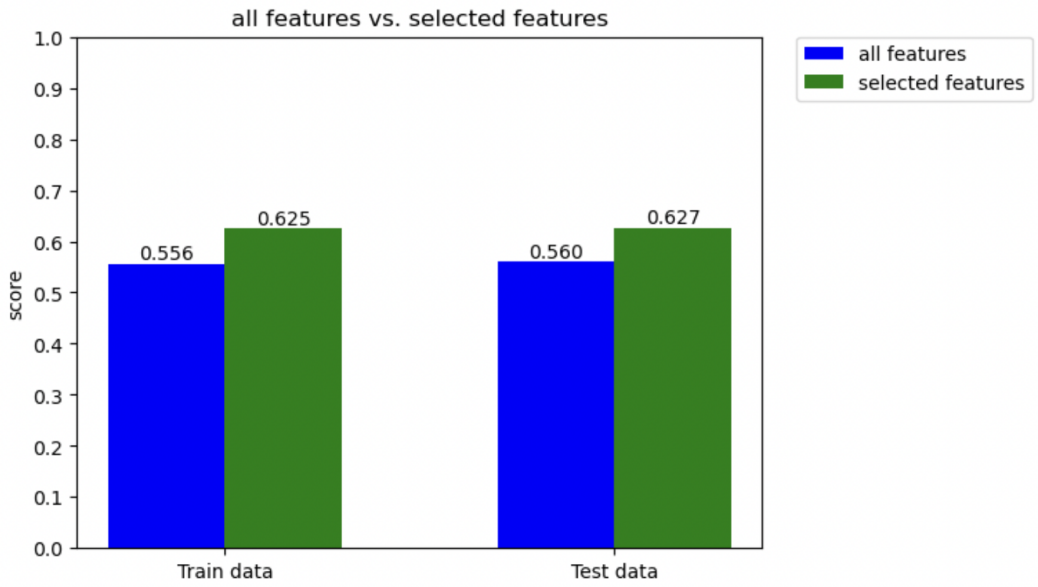

Figure 11. Accuracy of prediction for logistic regression

```
['Annual_Income']
Processed  25 models on 1 predictors in 80.17297720909119 seconds.
['Annual_Income', 'Credit_History_Age']
Processed  24 models on 2 predictors in 25.537167072296143 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan']
Processed  23 models on 3 predictors in 24.981653928756714 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt']
Processed  22 models on 4 predictors in 24.435335159301758 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit']
Processed  21 models on 5 predictors in 24.210216283798218 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit', 'Interest_Rate']
Processed  20 models on 6 predictors in 23.37015986442566 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit', 'Interest_Rate',
'Num_of_Loan']
Processed  19 models on 7 predictors in 22.59090518951416 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit', 'Interest_Rate',
'Num_of_Loan', 'Credit_Mix']
Processed  18 models on 8 predictors in 21.775750875473022 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit', 'Interest_Rate',
'Num_of_Loan', 'Credit_Mix', 'Num_Bank_Accounts']
Processed  17 models on 9 predictors in 20.776670932769775 seconds.
['Annual_Income', 'Credit_History_Age', 'Type_of_Loan', 'Outstanding_Debt', 'Changed_Credit_Limit', 'Interest_Rate',
'Num_of_Loan', 'Credit_Mix', 'Num_Bank_Accounts', 'Num_Credit_Card']
Processed  16 models on 10 predictors in 20.096049785614014 seconds.

In [114]: feature = predictors
          feature

Out[114]: ['Annual_Income',
           'Credit_History_Age',
           'Type_of_Loan',
           'Outstanding_Debt',
           'Changed_Credit_Limit',
           'Interest_Rate',
           'Num_of_Loan',
           'Credit_Mix',
           'Num_Bank_Accounts',
           'Num_Credit_Card']
```

Figure 12. Feature selection for KNN

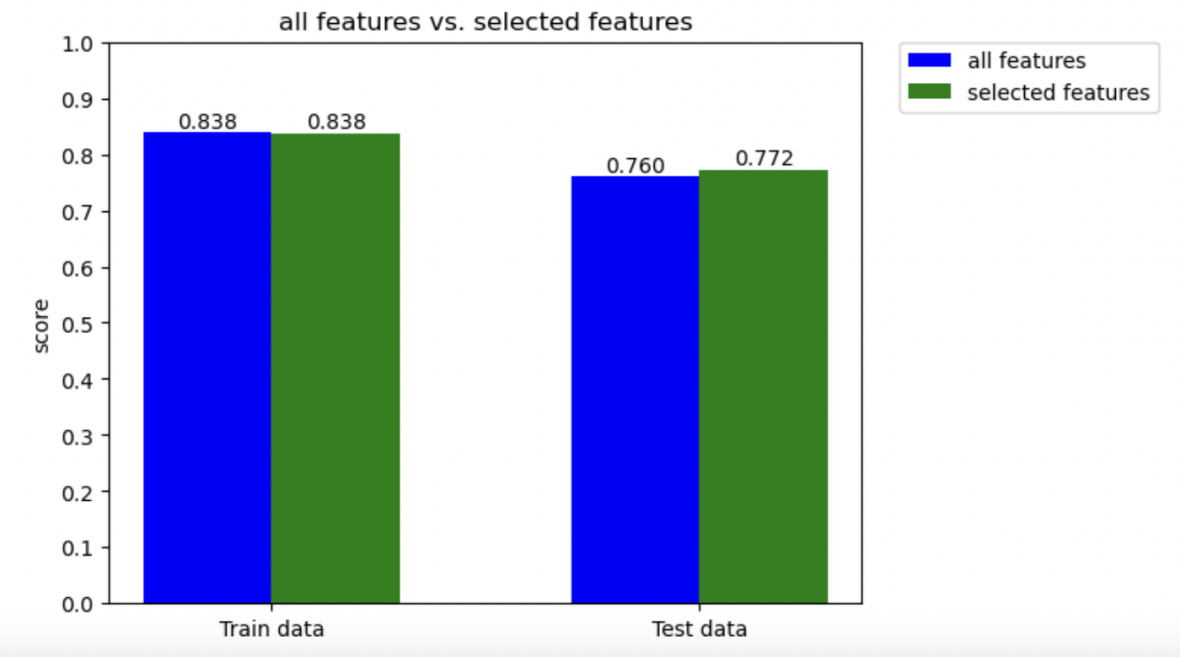

Figure 13. Accuracy of prediction for KNN

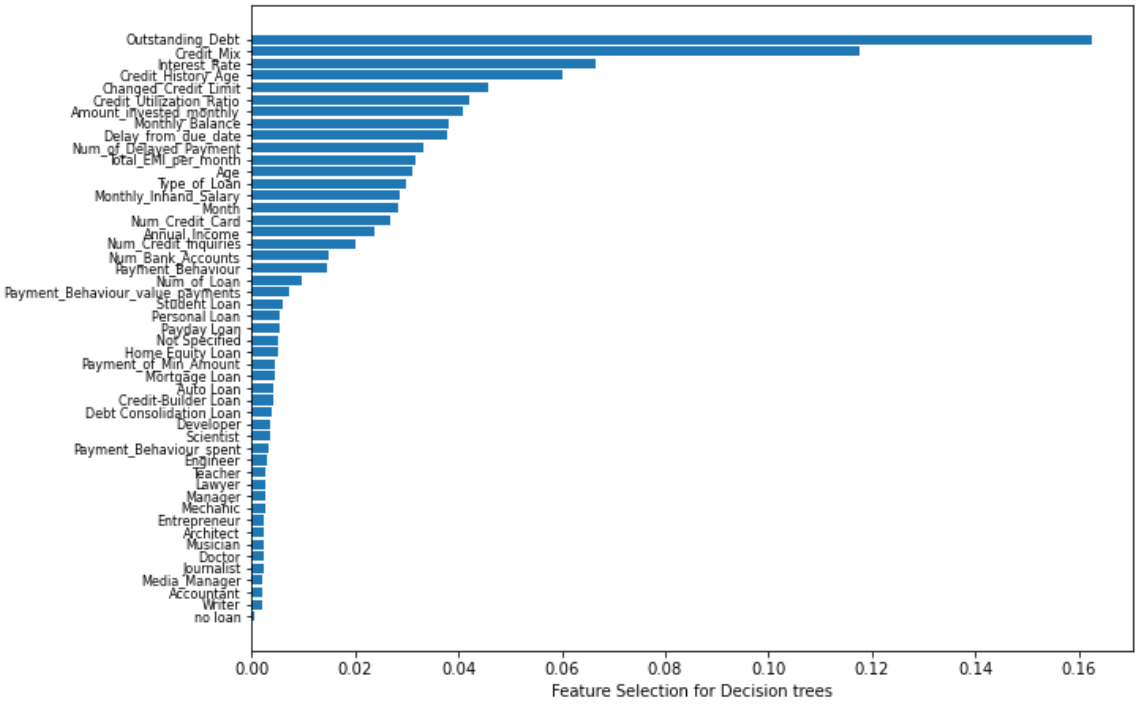

Figure 14. Feature selection for decision trees

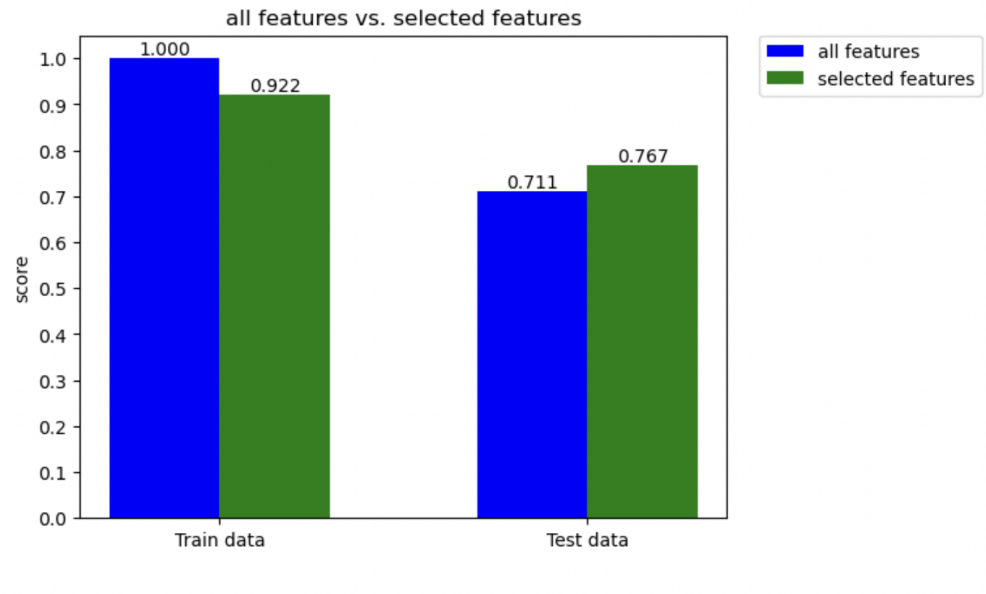

Figure 15. Accuracy of prediction for decision tree

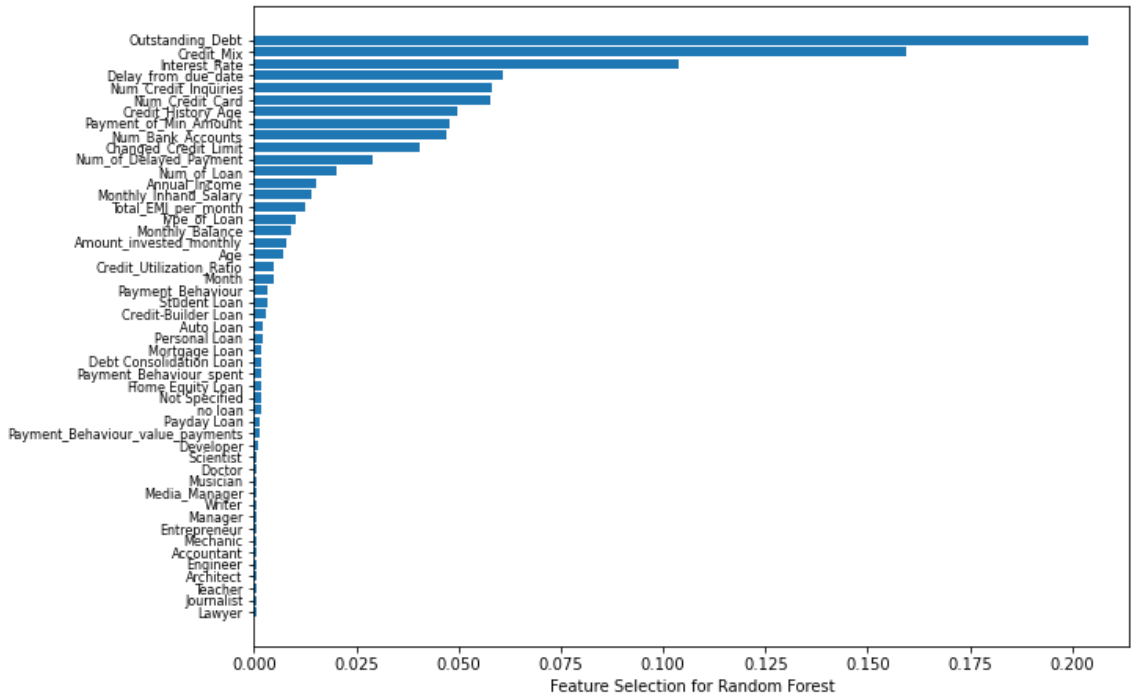

Figure 16. Feature selection for random forest

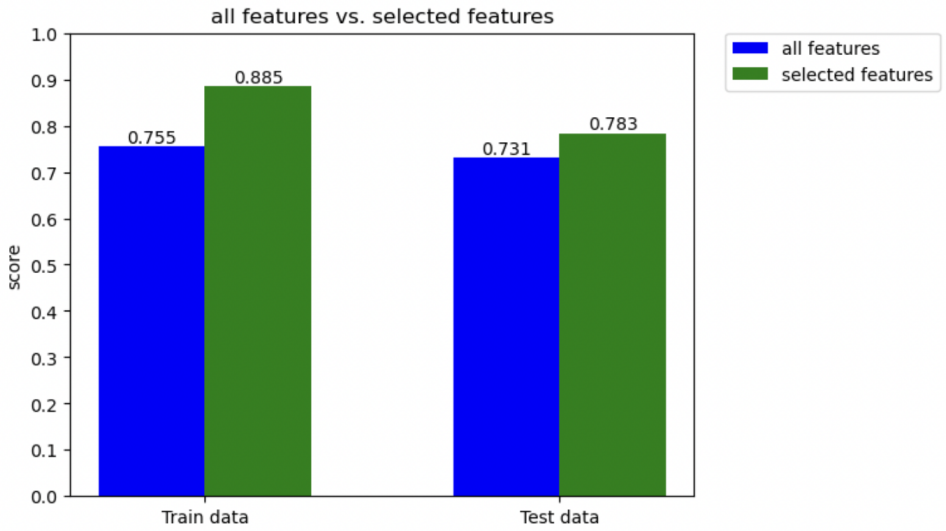

Figure 17. Accuracy of prediction for random forest

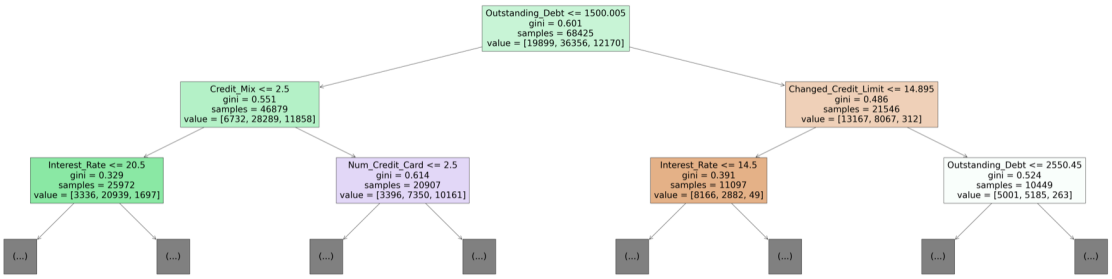

Figure 18. Decision Tree using all features

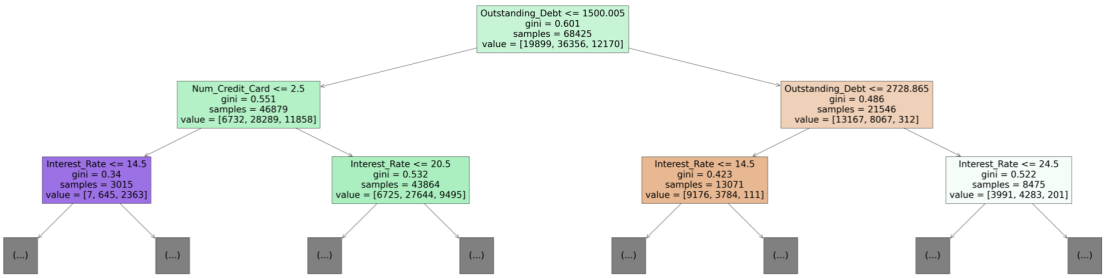

Figure 19. Decision Tree using selected features

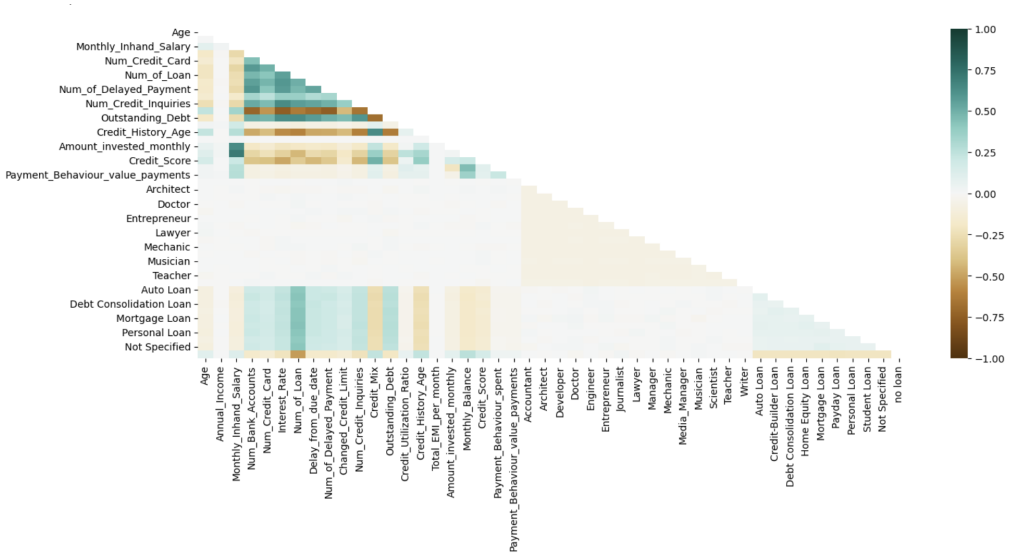

Figure 20. Correlation heatmap

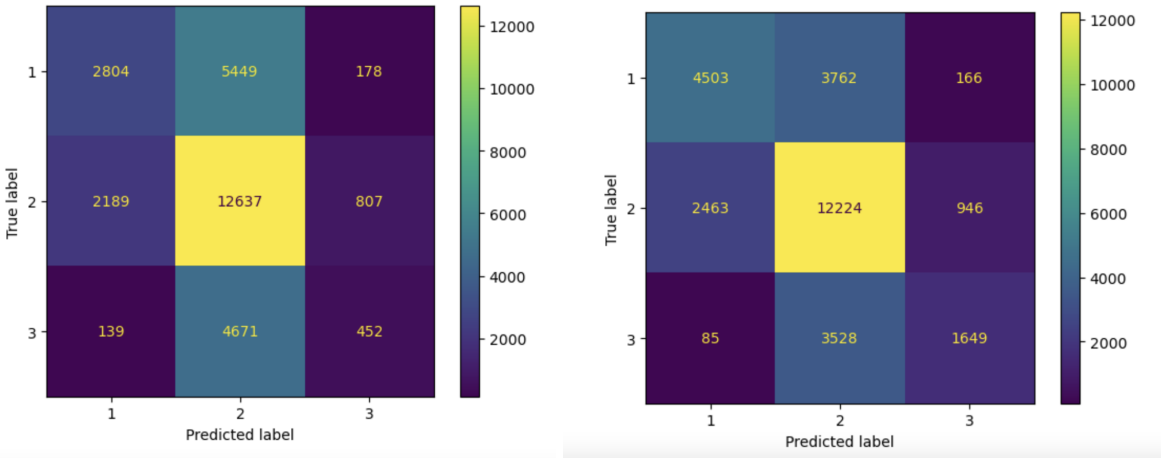

Figure 21. Confusion Matrix for Logistic regression (L: using all features, R: using selected features)

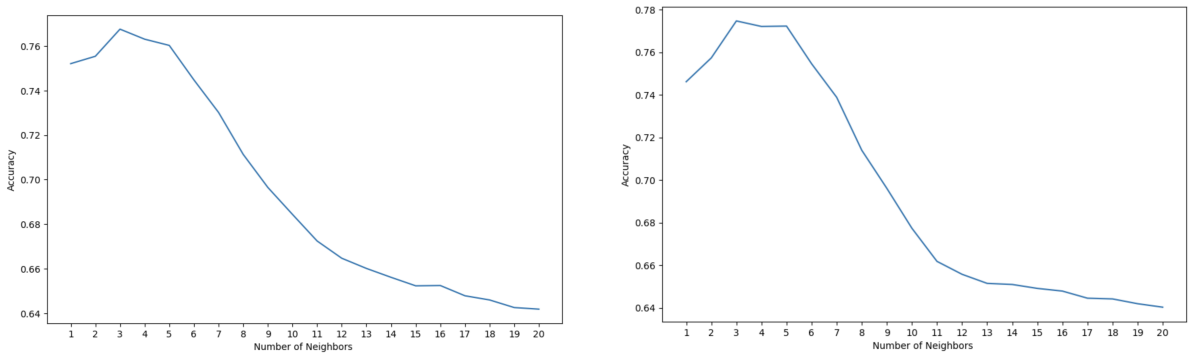

Figure 22. KNN number of neighbors and accuracy (L: using all features, R: using selected features)

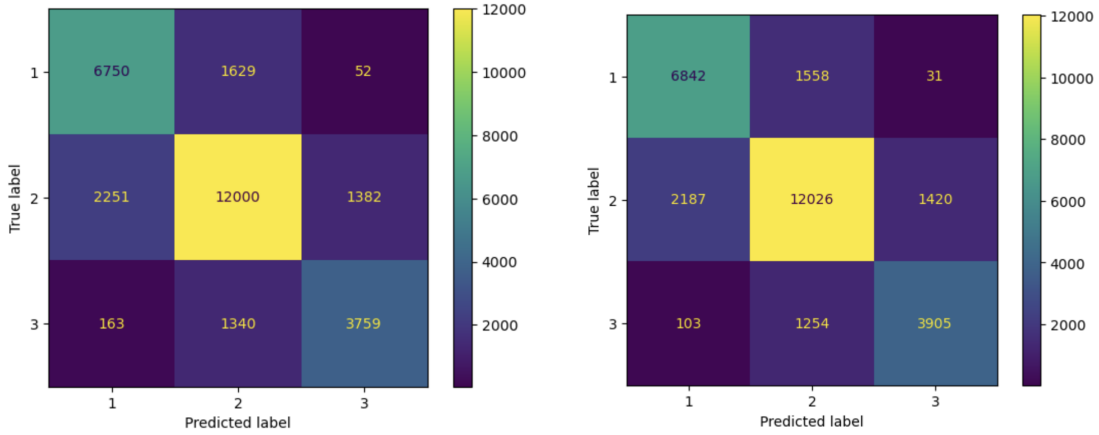

Figure 23. Confusion Matrix for KNN (L: using all features, R: using selected features)

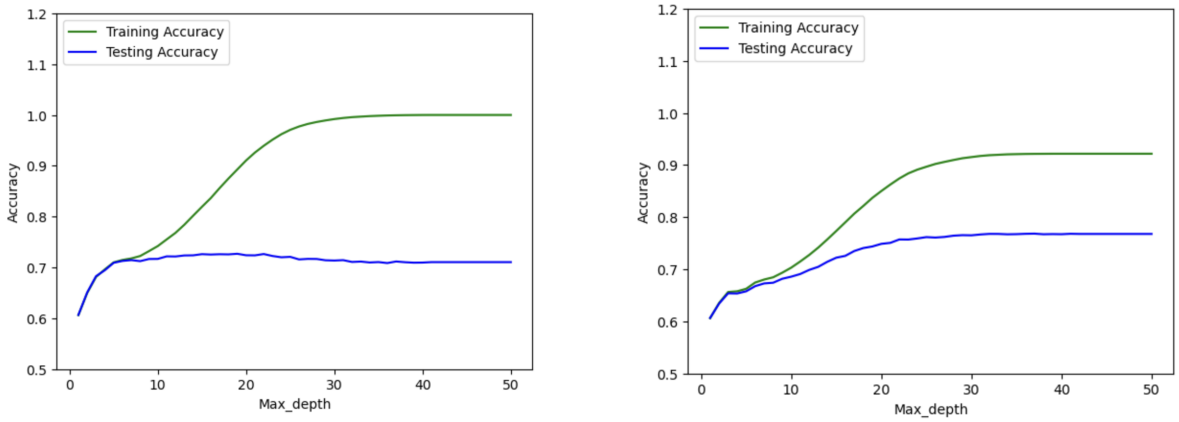

Figure 24. Decision Tree max_depth and accuracy (L: using all features, R: using selected features)

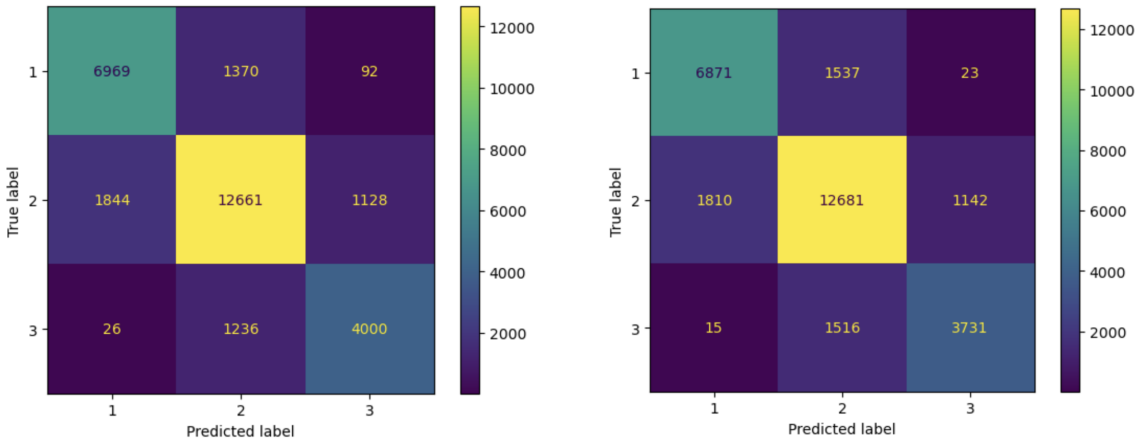

Figure 25. Confusion Matrix for Random Forest (L: using all features, R: using selected features)