GERMAN CREDIT RISK CLASSIFICATION: ARE YOU AT RISK?

Data 602

Debanjan Chowdhury

Introduction

- **Overview**: For this project, we are taking a dataset that has the credit details of individuals in Germany by considering important factors like bank (savings and checking) account details, age, job, purpose of credit (purchase), etc. The dataset also has a risk column as the target where the credit risk is evaluated considering all features. My role is to evaluate if the risks are accurate or not by using logistic regression and comparing it with decision tree models. I am also evaluating if removing the null values or outliers has an impact on the logistics regression model results.

- **Motivation**: A CNBC article mentions an interview where they understood that 37% of the 1,000 people they interviewed do not have any idea on how their credit scores are calculated. When I was young, I was also unsure as I got a credit card which showed FICO score, but only for that card not overall. I wondered if many individuals have different credit cards, then how would they know the overall .

- **Goals**: For this project, I am a Data Scientist in Frankfurt Germany for a financial company where they take provided information from customers and analyze their credit risk by considering many factors like checking account, savings account, income, etc. My role is to verify whether the risk analysis was accurate or not. In order to do that I use logistic regression models to test the validity of the risks and also use decision tree for a comparison and do feature engineering where I remove the null values and some outliers to test if they have any impact on the logistic regression.

- **Research Question**: When we use logistic regression, will the accuracy and overall scores of credit risk evaluation be larger or will it be larger when we use other classification models and will it play a role when we modify the dataset like remove outliers?

Dataset

- Dataset: https://www.kaggle.com/kabure/german-credit-data-with-risk This dataset is in Kaggle where it contains all the features with targets to evaluate the credit risk

- We referred to this following dataset to gain an understanding of the features in more details: https://www.kaggle.com/uciml/german-credit The overall dataset was inspired by a dataset in the UCI repo: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.

- This dataset contains information about individuals in Germany who provided information about their bank account details, jobs, etc and according to UCI, the initial dataset was published by Professor Hoffman in the University of Hamburg.

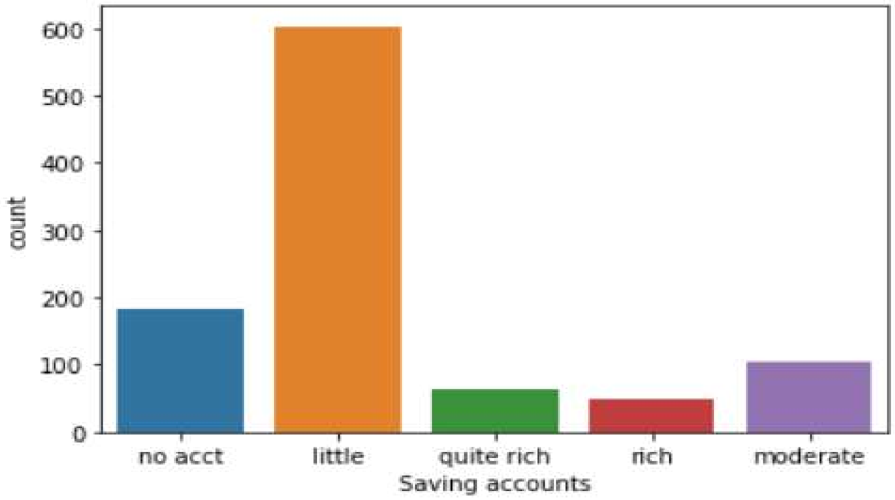
- Columns: Age, Sex, Job, Housing, Saving accounts, Checking account, Credit amount, Duration, Purpose, Risk
  - Target column is Risk.

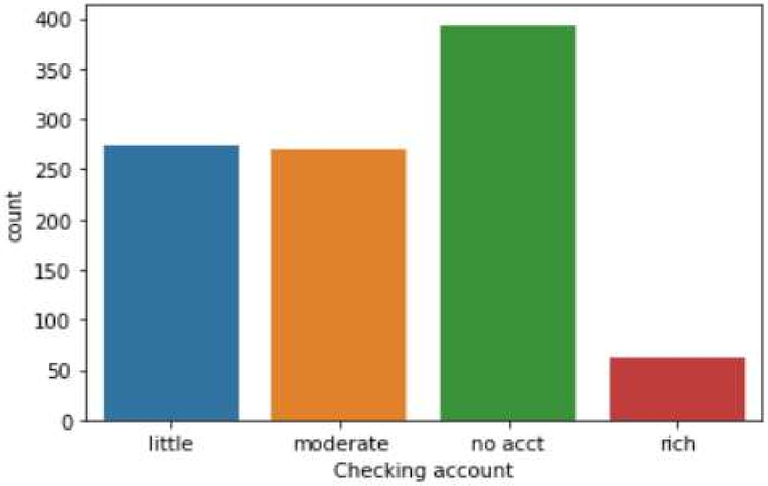
- The dataset has: **"100 rows and 10 columns.“**

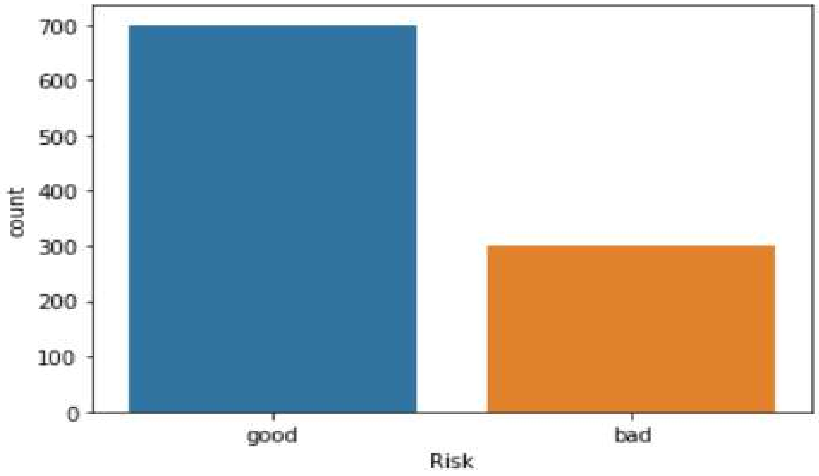
Techniques used and steps

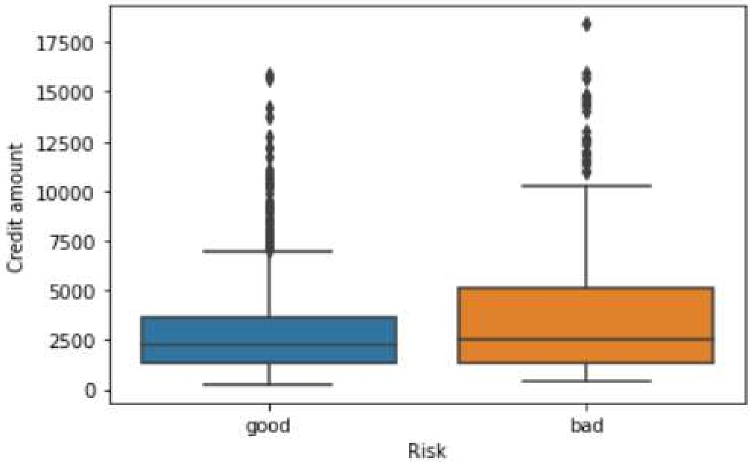
- We used logistic regression and decision tree as the target is a classification column and I also tested the logistic regression after removing the outliers from dataset and tested with removing null values.
- First, we divided the dataset into training and testing datasets.
- Next, I used scikit learn to develop a prediction using logistic regression model for the training datasets and testing datasets.
- I developed a confusion matrix for the training and testing model that would show us the correlation between how many predictions vs accurate data matched and did not match.
- Following that step, I found the accuracy, precision and the recall scores.
- Formulas below:
  - Accuracy = correct prediction amount / number of the total cases
  - Precision = number of true positives / (true positive + false negative) or actual positives
  - Recall = number of true positives / (true positive + false positive) or predicted positives

Visualizations to understand the data



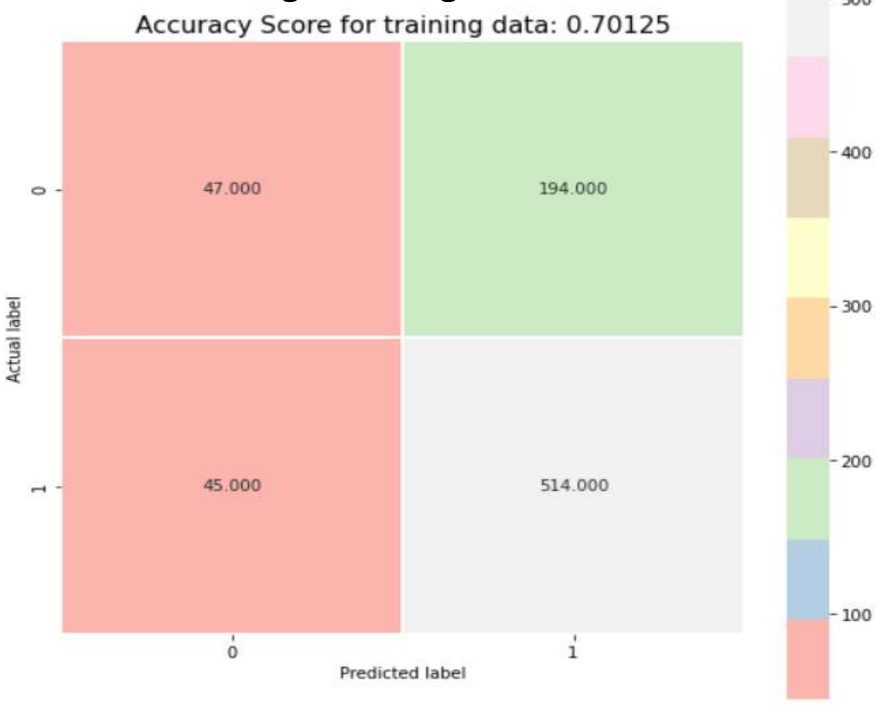




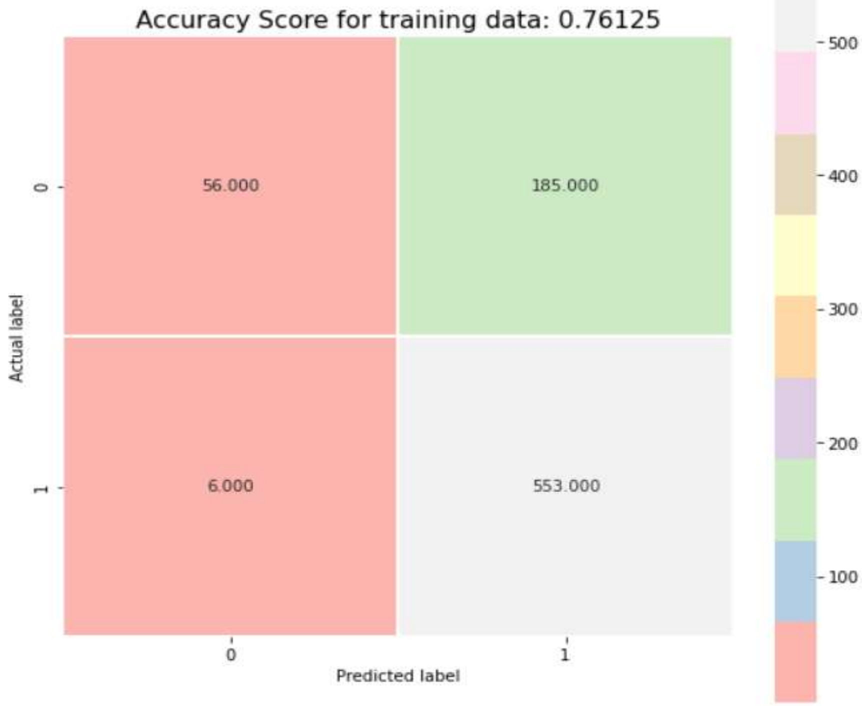


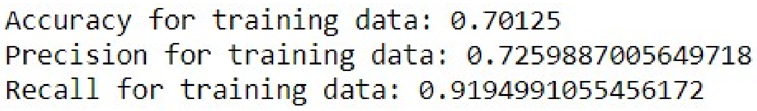
Logistic Regression vs Decision Trees

Logistic Regression

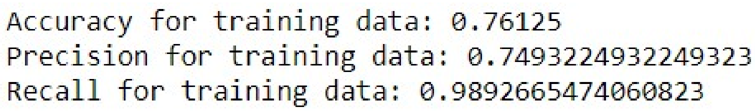


Decision Tree





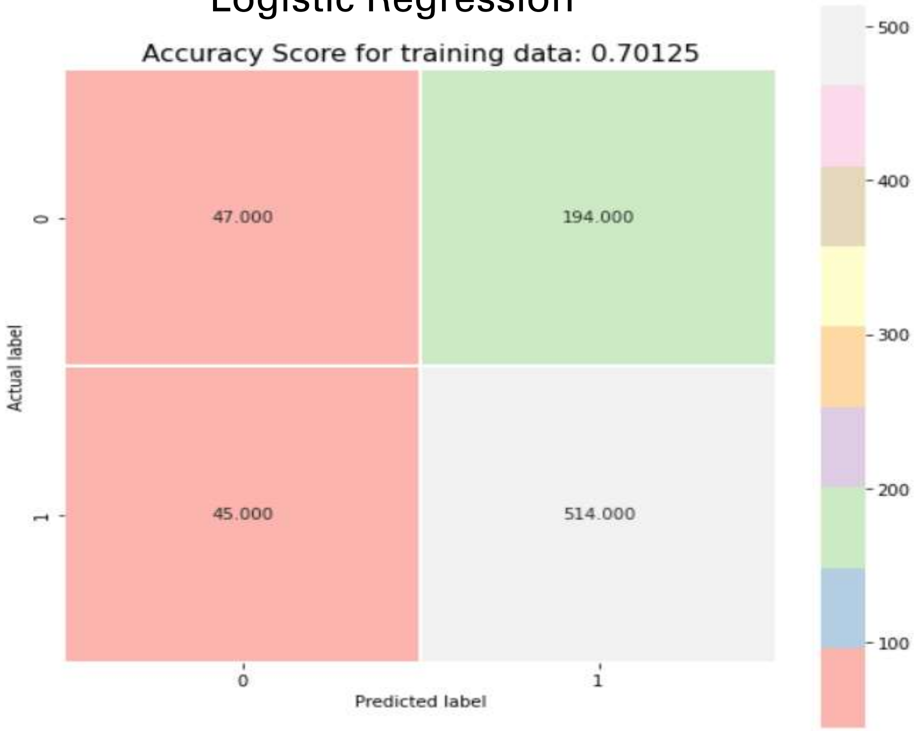
```
Accuracy for training data: 0.70125
Precision for training data: 0.7259887005649718
Recall for training data: 0.9194991055456172
```

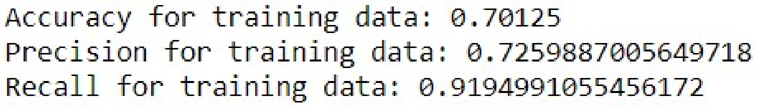


```
Accuracy for training data: 0.76125
Precision for training data: 0.7493224932249323
Recall for training data: 0.9892665474060823
```

Logistic Regression – after removing null values

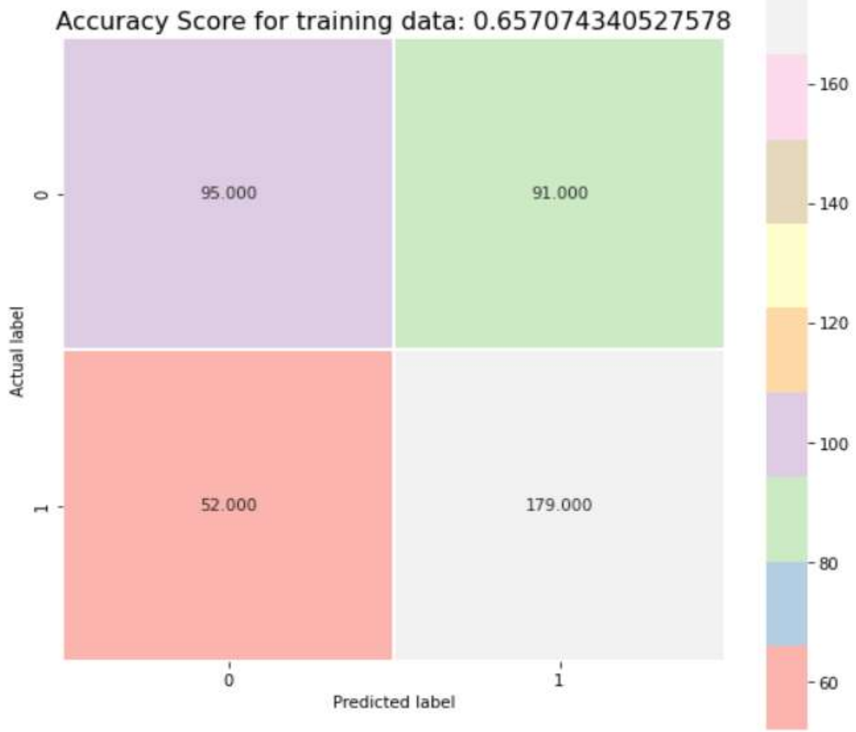
Logistic Regression

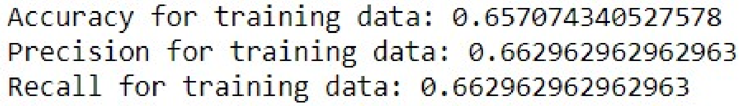




```
Accuracy for training data: 0.70125
Precision for training data: 0.7259887005649718
Recall for training data: 0.9194991055456172
```

Logistic Regression after removing NaN values

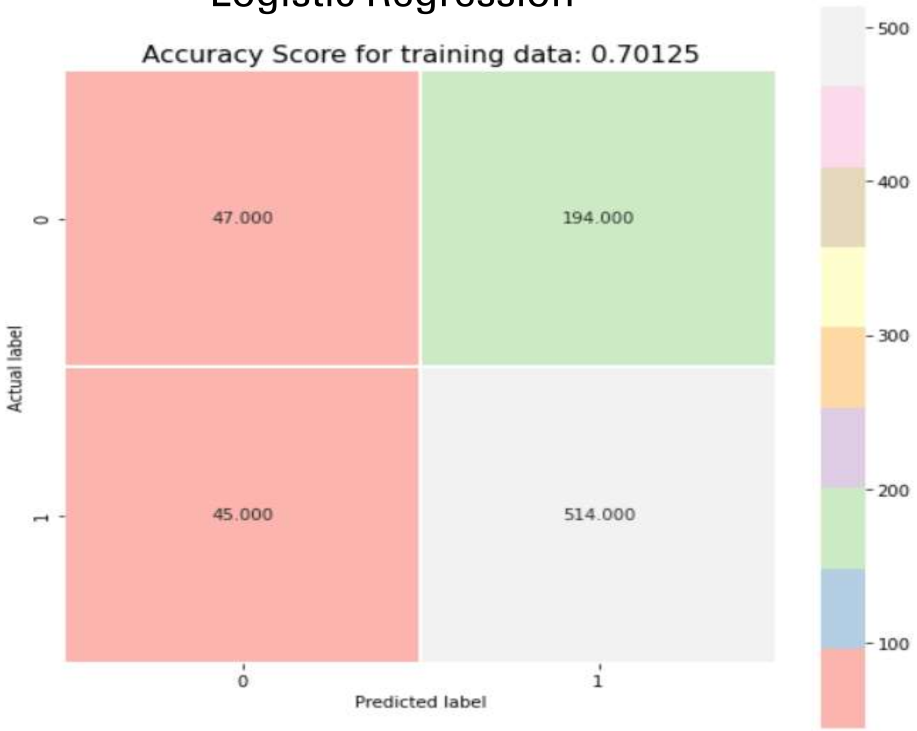




```
Accuracy for training data: 0.657074340527578
Precision for training data: 0.662962962962963
Recall for training data: 0.662962962962963
```

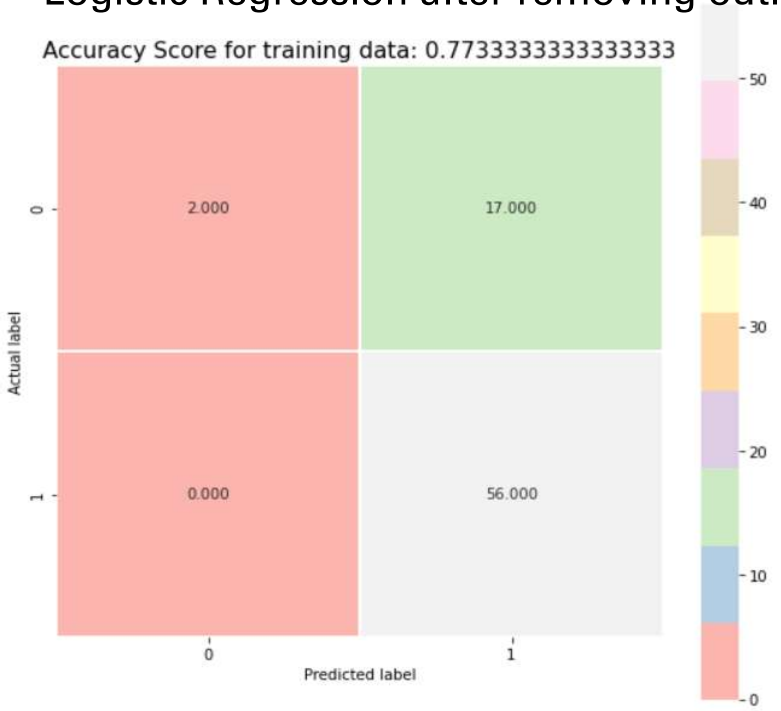
Logistic Regression – without outliers

Logistic Regression



```
Accuracy for training data: 0.70125
Precision for training data: 0.7259887005649718
Recall for training data: 0.9194991055456172
```

Logistic Regression after removing outliers



```
Accuracy for training data: 0.79
Precision for training data: 0.8041237113402062
Recall for training data: 0.975
```

Evaluation and Conclusions

- While we noticed that removing the outliers from two columns helped improve the accuracy scores, in a busines setting removing outliers may not always be the solution to verifying the accuracy as each customer would matter.

- I noticed that removing the rows with null values also did not help us as we had a lower dataset and the overall scores were lower in that case.

- We noticed that when we used decision tree matrix, we get relatively higher scores in terms of accuracy, precision and recall. However, I also tried testing the test variables that we had split in the beginning and it seemed that logistic regression seemed to work better there. I feel that maybe the test size we took played a role and if we took a larger test size we could possibly see different results. Also, I set max leaf nodes in decision trees to 10 and a higher node may have shown different results as well.

- I feel that in terms of performance the Decision Tree model is a better option for classification models and as a Data Scientist in the financial company, I would like to recommend that model.

- In the future, I would like to try increasing the max leaf node in the decision tree model and possibly try to remove the main outliers from the entire dataset instead of just of a few. I would also like to see if an individuals job title or position also plays a role in the classification of the credit risks.

Work Cited

- Dataset: https://www.kaggle.com/kabure/german-credit-data-with-risk.
- Referred to: https://www.kaggle.com/uciml/german-credit
- Dataset inspired by: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.
- https://www.cnbc.com/2019/01/04/about-4-in-10-americans-have-no-idea-how-credit-scores-are-determined.html
- https://www.investopedia.com/the-side-effects-of-bad-credit-4769783#:~:text=Poor%20credit%20can%20make%20it,%2C%20renter's%2C%20and%20homeowner's%20insurance.