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 outliers has an impact on the models.
- **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 scores.
- **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 credit risks by considering many factors like checking 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 and decision tree models to test the validity of the risks. I am also checking if removing the outliers from checking and savings account has an impact on the models. The main objective is to develop a model with a higher accuracy score that would exceed 70% as that is our dummy accuracy ratio or baseline ratio to check for imbalance between two variables.
- **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.

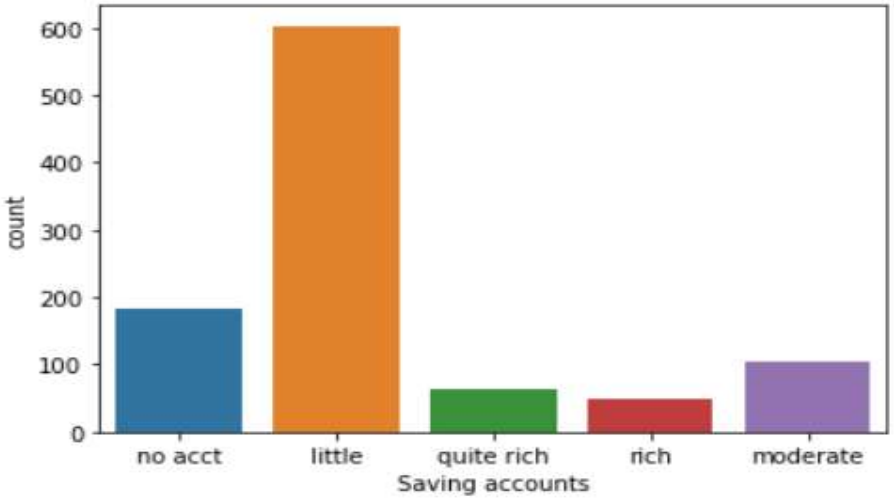
- The dataset has: **"100 rows and 10 columns.“**

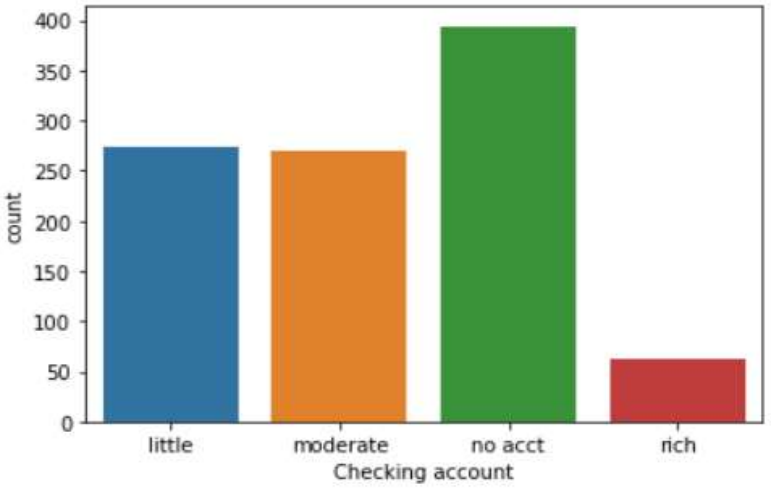
Techniques used and steps

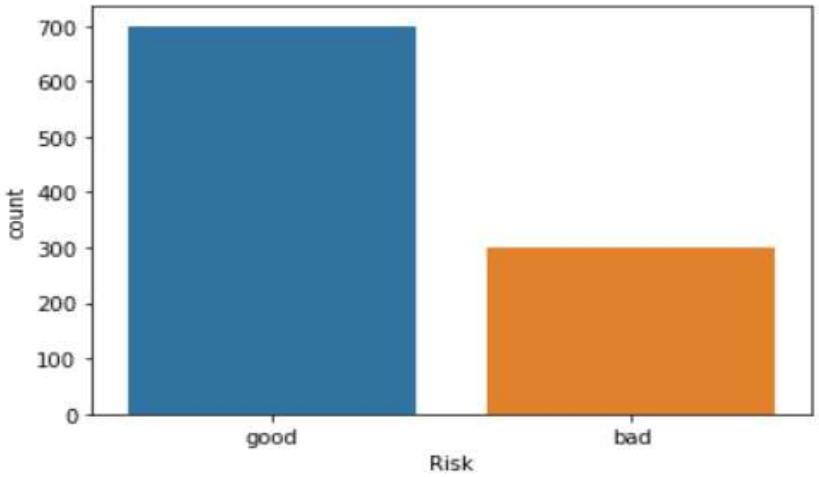
- We used logistic regression and decision tree as the target is a classification column and I also tested both models after removing outliers from checking and savings account columns.

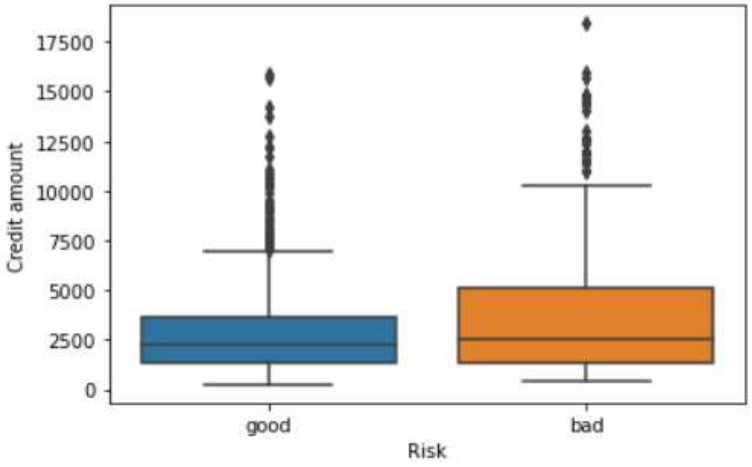
- First, I split the dataset into a train and test set.
- Next, I used scaling to make sure that all datasets are scaled into a specific range.
- Next, I used scikit learn to develop a prediction using logistic regression model for the training datasets and testing datasets and also found accuracy scores
- 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. **(Formula mentioned below for those).**
- Next, I used cross validation to check the mean and standard deviation and verify the model.
- 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



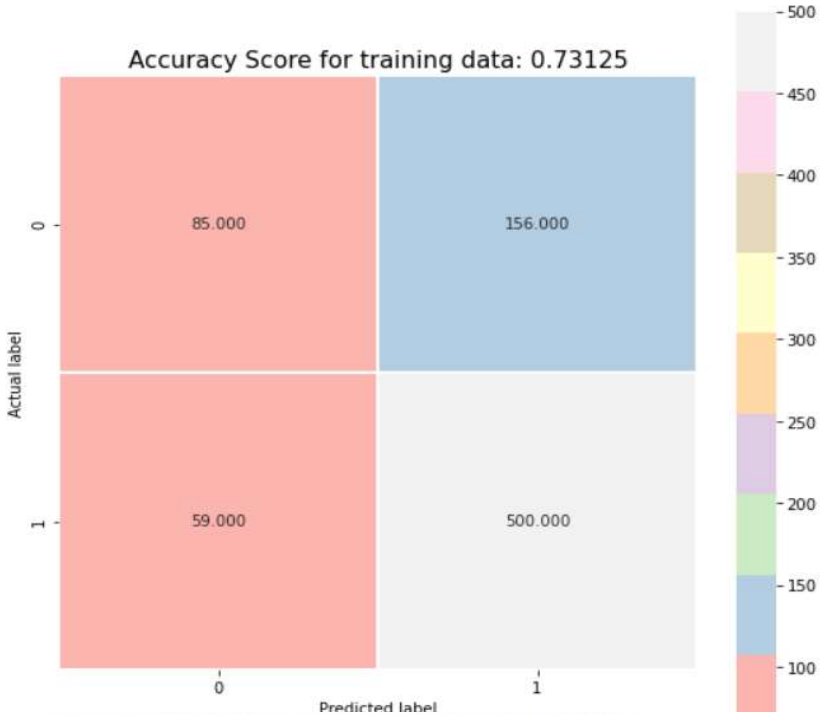






Logistic Regression vs Decision Trees

Logistic Regression



Log reg accruacy score: 0.73125
Log reg precision score: 0.7621951219512195
Log reg recall score: 0.8944543828264758

Logistic Regression 5-fold cv results (Accuracy) 0.730 =/- 0.028

Decision Tree

Dec tree accruacy score: 0.7475
Dec tree precision score: 0.8534653465346534
Dec tree recall score: 0.7710196779964222

Decision tree 5-fold cv results (Accuracy) 0.729 =/- 0.026

Both models – without outliers

Logistic Regression comparison after removing outliers

```
array([[ 85, 156],
       [ 59, 500]], dtype=int64)
```

```
Log reg accruacy score: 0.73125
Log reg precision score: 0.7621951219512195
Log reg recall score: 0.8944543828264758
```

```
array([[ 82, 123],
       [ 54, 412]], dtype=int64)
```

```
Log reg accruacy score without outliers: 0.736214605067064
Log reg precision score without outliers: 0.7700934579439253
Log reg recall score without outliers: 0.8841201716738197
```

Decision Tree after removing outliers

```
array([[167,  74],
       [128, 431]], dtype=int64)
```

```
Dec tree accruacy score: 0.7475
Dec tree precision score: 0.8534653465346534
Dec tree recall score: 0.7710196779964222
```

```
array([[106,  99],
       [ 46, 420]], dtype=int64)
```

```
Dec tree accruacy score without outliers: 0.7839046199701938
Dec tree precision score without outliers: 0.8092485549132948
Dec tree Recall score without outliers: 0.9012875536480687
```

- As we can see above, the scores do not have much of a difference when we remove the outliers from the savings and checking's account for logistics regression, but we see a significant difference when we look at the scores for the decision tree.

Model Selection and thoughts

- I decided to go with the decision tree model as it had a higher accuracy score and there were not as many false positives and false negatives as the logistic regression model.
- The logistic regression and the decision tree had almost the same mean (logistic regression was slightly higher with the mean of 73% while decision tree was 72.9% and the standard deviation was 0.28 in logistics while it was 0.26 in decision tree.
- Even though, the cross validation showed higher data for logistics regression, I feel there is not much difference in the numbers and the accuracy scores and other scores like precision seemed to be overall higher when we used the decision tree model.
- When we removed the outliers, it seemed that we saw an improvement in the overall scores for the decision tree model.
- Due to these reason, I feel decision tree would be the best option.

Conclusions and Limitations

- While we noticed that removing the outlier was helpful for one model and not helpful for other, that may not always be a feasible option depending on business rules. In our case, we would be excluding some individuals in the calculations if we remove outliers from two columns only.

- We noticed that when we used decision tree matrix, we get relatively higher scores in terms of accuracy and the precision value but not recall. I had also set max leaf nodes in decision trees to 10 and max leaf node means the root tree can split until they have 10 leaf nodes.

- 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 other types of classification models, see how an individuals job title or role like engineer, manager would play a role in the classification model. I would also like to remove outliers from the overall dataset instead of just two columns. Though, we had set our aim higher than 70% for the accuracy scores, I believe that a 90-95% accuracy would be ideal for customers as it is not good when someone is in risk and they see that they are not in risk.

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.