A Logistic Regression Model for Consumer Default Risk using the German Dataset

Ed-Joel Odhiambo, Lynn Miano, Kenneth Kiprotich, Christine Keriga, Mercy Maina and Joy Kendi

Strathmore University

Dr. Samuel Chege

1. Introduction

This project aimed to compare the results of the model used in ***A Logistic Regression Model for Consumer Default Risk*** with other models to determine the best model that can be applied for Credit Default Risk. To determine the strength of variables, we used tests such as Pearson's chi test, the T test, and the Mann-Whitney U test.

2. Data Acquisition & Pre-preprocessing

We started by cleaning up the data to allow for proper data manipulation. Then we classify the data into its different variables. We assumed the probabilities of non-default to be 0, and for Default 1.

2.1. Data Preprocessing

This was done to ensure that the cleaned data were indeed legitimate. We checked to see whether the variables identified consisted of null values that would otherwise have affected our results. Having found that the data derived consisted of non-null values, we proceeded with our analysis. We did a basic analysis on the variables to determine their distribution and effects they hold on default probabilities.

2.2. Feature Selection

This section outlines some tests conducted on the datasets to prove their authenticity and interactions with each set of variable. We also classified the data into categorical, ordinal, and continuous types to ensure all data types are correctly analysed through the different techniques.

2.2.1. Chi-Square Test - Categorical/Ordinal Data, Non-parametric

We used the Pearson-Chi square test to understand the significant association between the variables and the target default, where the null hypothesis was: *No association.* this was done to ensure that our variables are directly linked to credit default. Where a low p-value **(p < 0.05)** suggests a statistically significant relationship.

```
Chi-Square Test for Categorical Variables
salary: chi2 = 1.56, p = 0.2114
tax_echelon: chi2 = 36.10, p = 0.0000
```

2.2.2. Mann-Whitney U Test - Continuous Data

This models tests the distribution of continuous variables features that differ between defaulters and non-defaulters. This model is a non-parametric model which assumes no normality. If **p < 0.05** we reject the null hypothesis and conclude that there is significant difference in distributions, when **p >= 0.05** we fail to reject and conclude that there is no significant difference in the models. In our analysis, we found that term, age, and age_credit_cards has strong evidence that their distributions differs between defaulters and non-defaulters.

```
Mann-Whitney U Test for Continuous Variables (non-parametric)
term: U = 77995.50, p = 0.0000
age: U = 119833.00, p = 0.0004
credit_cards: U = 110272.00, p = 0.1348
age_credit_cards: U = 118121.00, p = 0.0017
```

2.2.3. T-Test - Continuous, Parametric

This tests if means differ between default groups, and it assumes normal distribution. It confirms if mean differences in features of the variables have significant predictors. For a T-test, **p < 0.05** proves that there is a significant difference between means and **p >= 0.05** implies that there is no strong evidence for a difference in the means. In our analysis, we found that term, age, and age_credit_cards has strong evidence that their distributions differs between defaulters and non-defaulters.

```
T-Test for Continuous Variables (parametric)
term: t = -6.47, p = 0.0000
age: t = 2.91, p = 0.0038
credit_cards: t = 1.47, p = 0.1416
age_credit_cards: t = 2.86, p = 0.0043
```

2.2.4. Cohen's d - Effect Size

This test measures the standard difference in means. The effect size is measured in three ways **0.8: large effect, 0.5: medium effect, 0.2: small effect**. It quantifies how practically significant a feature is, beyond just p-values, and is useful in prioritizing impactful variables.

```
Cohen's d (effect size) between defaulted vs. non-defaulted:
term: Cohen's d = -0.462
age: Cohen's d = 0.200
credit_cards: Cohen's d = 0.101
age_credit_cards: Cohen's d = 0.192
```

2.2.5. Train-Test Split

The train-test splits datasets into 80% train / 20% test which stratifies on target to preserve class balance. This test prevents data leakage and enables model evaluation on unseen data. Moreover, it maintains target distribution in both splits: defaults and non-defaults. We found that 30% of the samples are defaulters and 70% are non-defaulters in both tests.

```
Training set size: (800, 6)
Test set size: (200, 6)
Default rate in train: 0.3
Default rate in test: 0.3
```

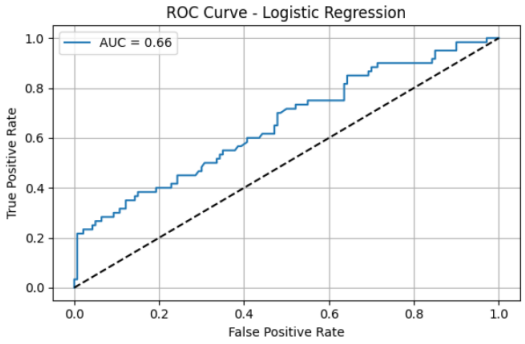
3. Model Training

This section aims to discern different models that can be used to analyse the German datasets.

3.1. Logistic Regression

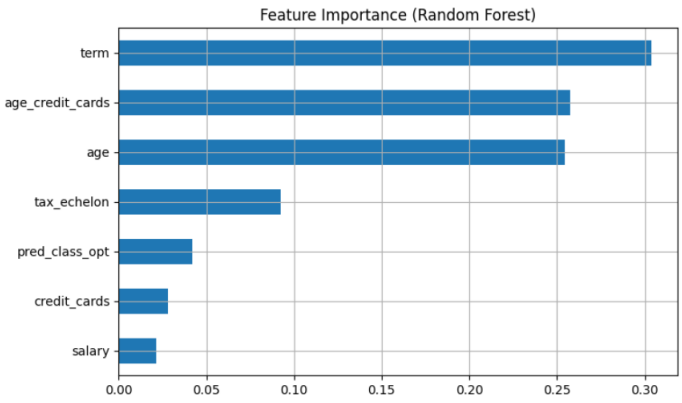
The Logistic Regression models log-odds of defaults as a linear combination of features, where the coefficients indicate direction and strength of association with the target. The goal was to classify the customers as defaulters or non-defaulters based on their credit characteristics. We found 74% accuracy on the model, with a moderate AUC of 0.66. We found our Precision metric to be 70% which implies that 70% of predicted defaults were true defaults, with 27% were Recall defaults. Moreover, the F1-Score of defaults were 39% measuring a weak balance between precision and recall. We also found that recall (non-defaults) were 95% which proves an excellent model at non-defaulters.

```
Confusion Matrix:
[[133   7]
 [ 44  16]]

Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.95      0.84       140
           1       0.70      0.27      0.39        60

    accuracy                           0.74       200
   macro avg       0.72      0.61      0.61       200
weighted avg       0.73      0.74      0.70       200

AUC Score: 0.66
```

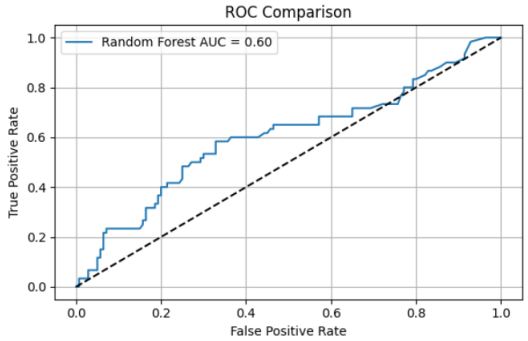


3.2. Random Forest

A random forest builds decision trees to make predictions. We performed multiple random forest models to find the optimal results. The model predicts non-defaulters fairly well with precision of 76% and recall 79%. The model, however struggles for defaulters with Recall lying between 27% - 40%. The model still outlined an AUC of 0.6 which implied low separation power of defaulters and non-defaulters. The model is slightly better than logistic regression with a recall of 40%, but the model is still slightly biased towards non-defaulters.

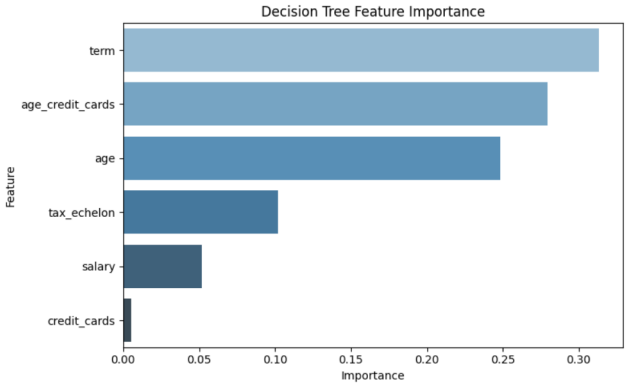



```
Random Forest Confusion Matrix:
[[111  29]
 [ 36  24]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.79      0.77       140
           1       0.45      0.40      0.42        60

    accuracy                           0.68       200
   macro avg       0.60      0.60      0.60       200
weighted avg       0.66      0.68      0.67       200

Random Forest AUC: 0.60
```



3.3. Decision Tree

Like the rest, the model has a fairly good prediction for non-defaulters with 72% precision and 71% recall. Moreover, it struggles for defaulters 34% and 35% recall which further implies a bias on non-defaulters. The tree is large with **depth=21, 297 leaves** which is overfitting to the training data. The AUC in this model is slightly lower with 58% accuracy, indicating poor calibration metrics. Overall, the decision trees performs worse than Random Forest and Logistic Regression. It however, maintains a better recall for defaulters at 35% compared to Logistic Regressions 27%.

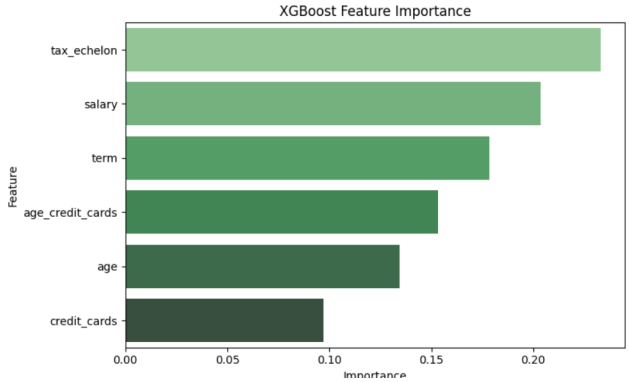



```
Decision Tree Depth: 21, Leaves: 297
Confusion Matrix:
 [[100  40]
 [ 39  21]]
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.71      0.72       140
           1       0.34      0.35      0.35        60

    accuracy                           0.60       200
   macro avg       0.53      0.53      0.53       200
weighted avg       0.61      0.60      0.61       200

AUC: 0.5776190476190475
Accuracy: 0.605
Log Loss: 11.977322700070157
MSE: 0.3602777777777778
Brier Score: 0.3602777777777778
Pearson Corr: 0.12769888365607343
```

3.4. XG-Boost

The XG-Boost improves defaulter recall to 43% which is better than the previous models, with precision 48% which is

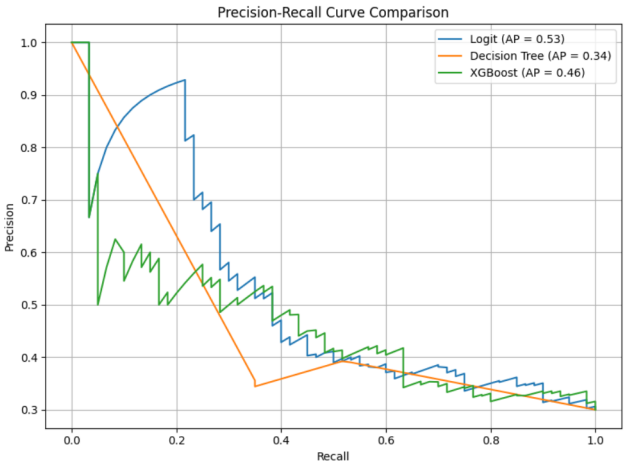
the most balanced detection seen. The F1-score for defaulters is highest among the models at 46% while the AUC is at 64% slightly lower than Logistic but higher than Decision Tree (58%) and Random Forest (60%). It also maintains a strong non-defaulter performance with precision 77% and recall 80%.



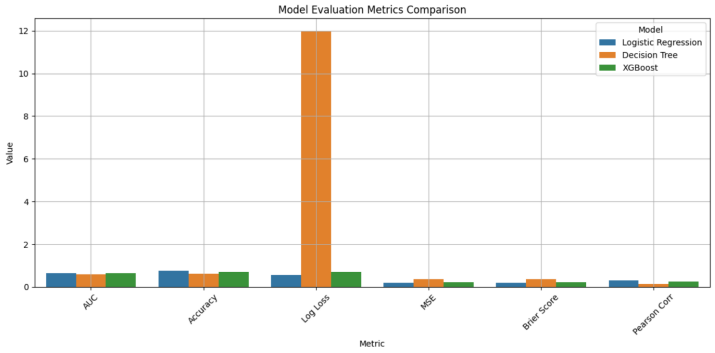

```
XGBoost Feature Importances:
tax_echelon         0.233206
salary              0.203749
term                0.178282
age_credit_cards    0.153129
age                 0.134584
credit_cards        0.097050
dtype: float32
Confusion Matrix:
 [[112  28]
 [ 34  26]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.80      0.78       140
           1       0.48      0.43      0.46        60

    accuracy                           0.69       200
   macro avg       0.62      0.62      0.62       200
weighted avg       0.68      0.69      0.69       200

AUC: 0.6426190476190476
Accuracy: 0.69
Log Loss: 0.715791959969061
MSE: 0.22908149659633636
Brier Score: 0.22908149533990183
Pearson Corr: 0.24275614145349073
```

4. Measurement Metrics

We analysed the different methods to determine the best model which achieves the best precision-recall curve. Logistic regression was found to be the best in and with a mean precision of 53%, maintaining the precision across all recall thresholds.



The decision Tree stands out with a very high log loss (12) which signals overconfidence and incorrect predictions, making it the worst model overall. The models each have their strengths, but XGBoost was found to be the best at defaulter detection, with the highest recall for defaulters at 43% and best precision at 48%. Logistic Regression, moreover, brings about better interpretability and highest accuracy which is easier to explain but weaker at default detection.



5. Analysis and Discussion

5.1. Comparison with the Portuguese Bank Data Analysis

Unlike in the Portuguese bank data, the variable term was found to be relevant in modeling the probability of default for the German dataset. Furthermore, despite the logistic regression we implemented for the German Dataset being weaker at default detection when compared to the machine learning algorithms, at a 70% precision for defaults, and 27% for recall defaults, the dataset performed significantly better than the Portuguese dataset where the logistic regression resulted in 79% precision for defaults, and 15% for recall defaults, demonstrating the impact of different datasets on model performance.

5.2. Challenges Encountered in the Replication

- Target class imbalance which is common in most credit datasets.
- Dataset availability- although a lot of versions of the German dataset were available, they differed in terms of values and structure.
- In the Portuguese dataset, and in common practice, default is assigned the value 1 and non-default is assigned the value 0. However, in our dataset, the assignment was done inversely.
- Inconsistency in feature naming conventions across the datasets.

5.3. How the challenges were addressed

A lot of time was spent in understanding the German dataset, the features and what they represented, and the similarities and differences with those used in the Portuguese dataset. This then greatly contributed to the data pre-processing process which then ensured a smooth analysis.