Final Report: German Bank Loan Default Prediction

Introduction:

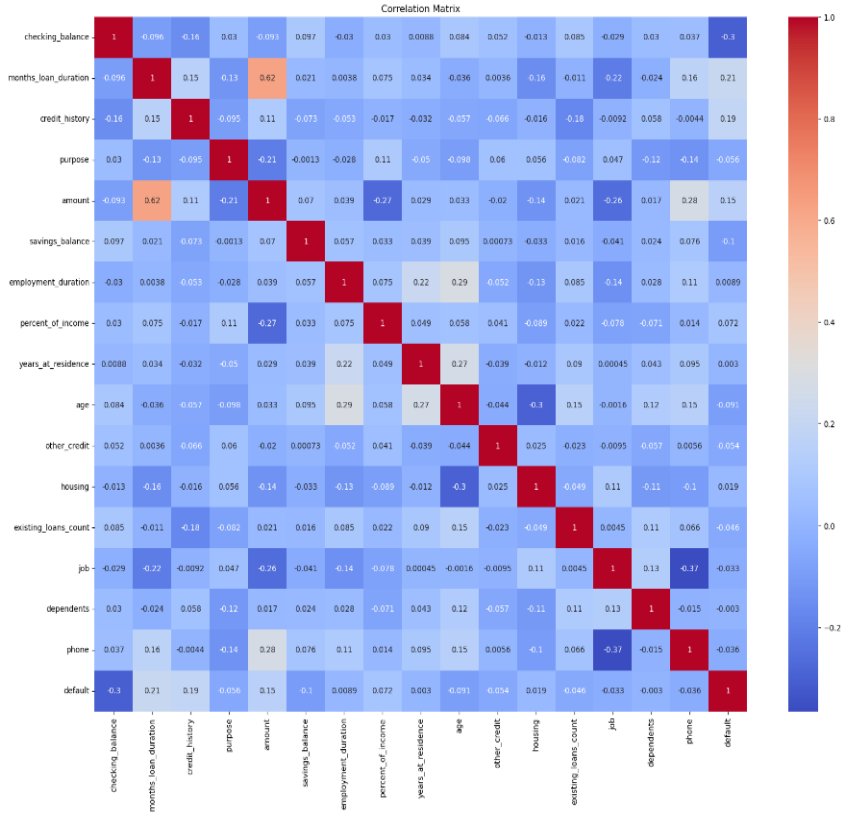
Financial institutions are very concerned with loan default prediction since it aids in determining the lending risk. A bank's decision-making process can be greatly enhanced by the capacity to forecast whether a borrower is likely to default on a loan, enabling better risk management and resource allocation. In this project, our goal is to develop a machine learning model that can accurately forecast loan default by analysing a dataset that a German bank gave. The dataset contains numerous variables linked to consumers' financial background and demographics.

In order to identify which machine learning model has the best predictive performance, this analysis examines a number of models, including Support Vector Machine (SVM), Random Forest, Decision Tree, and Logistic Regression. This project's main questions are: Which characteristics are most suggestive of loan default? Which artificial intelligence model yields the most precise forecasts? And how may the outcomes improve the bank's approaches to risk management?

Methods and Materials:

Data Exploration and Preprocessing

In order to comprehend the structure of the dataset and find any missing values, the analysis started with an exploratory data analysis (EDA). Next, the dataset underwent preprocessing to include missing value fills and label encoding for categorical variables. The data is in a format that is appropriate for model training and evaluation thanks to these preprocessing procedures.

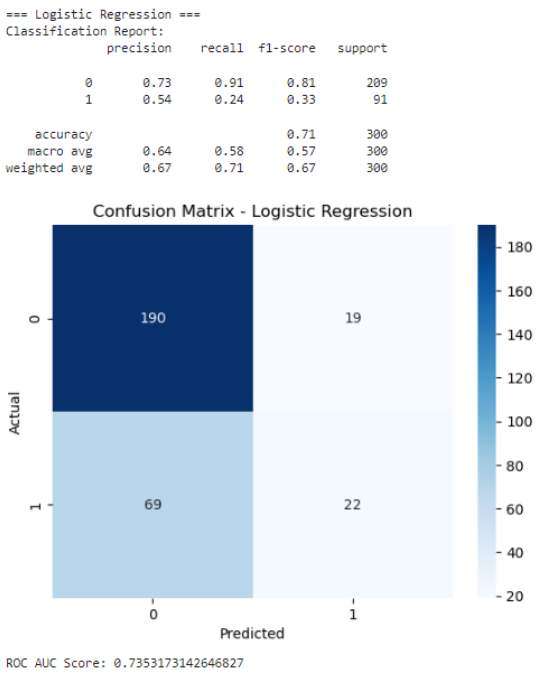


Model Selection and Training

For this investigation, we used four machine learning models: Support Vector Machine (SVM), Random Forest, Decision Tree, and Logistic Regression. After the data was processed, each model was trained, and its performance was assessed using critical performance indicators like accuracy, precision, recall, and ROC AUC score. To guarantee that the models were tested on untested data and that the evaluation of their performance was realistic, the data was divided into training and testing sets.

1) Logistic Regression

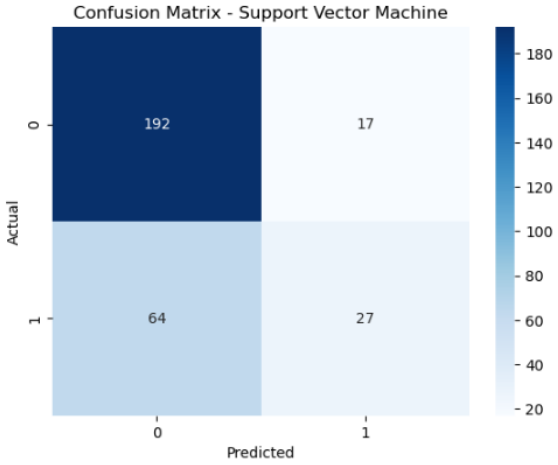
With an overall score of 71%, the Logistic Regression model showed a respectable level of accuracy. With a precision of 0.73 and a high recall of 0.91, it demonstrated good performance in detecting non-defaulters—that is, most non-defaulters were properly identified. Its recall of only 0.24, however, indicates that it performed much worse in detecting defaulters, missing a large number of real defaulters. Defaulters performed less evenly, as indicated by their F1-score of 0.33. Although the model's ROC AUC score of 0.7353 indicates a moderate ability to differentiate between defaulters and non-defaulters, it may underestimate the risk of defaults, which is important in financial applications, due to its propensity to misclassify a sizable number of defaulters (as shown in the confusion matrix).



2) Support Vector Machine (SVM)

The total accuracy of the Support Vector Machine (SVM) model's performance was 73%. For defaulters, it displayed a precision of 0.61, which means that 61% of the time it correctly anticipated a default. On the other hand, the recall for defaulters was only 0.30, suggesting that a significant proportion of real defaulters were difficult for the model to identify. This imbalance is reflected in the default class's F1-score of 0.40. The ability to differentiate between the two groups is reasonable but not outstanding, as indicated by the ROC AUC score of 0.7439. The confusion matrix reveals that although the SVM model was somewhat successful in identifying non-defaulters, many real defaulters were overlooked due to its high false negative rate.

```
=== Support Vector Machine ===
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.92      0.83       209
           1       0.61      0.30      0.40        91

    accuracy                           0.73       300
   macro avg       0.68      0.61      0.61       300
weighted avg       0.71      0.73      0.70       300
```

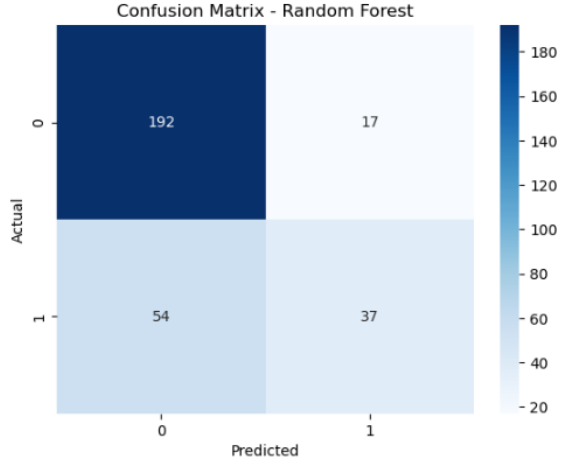



```
ROC AUC Score: 0.7438876912561122
```

3) Random Forest

In comparison to SVM and Logistic Regression, the Random Forest model performed better, predicting defaulters with a precision of 0.69 and a recall of 0.41, indicating a better balance in defaulter identification. Its F1-score for defaulters was 0.51—a higher degree of dependability in this domain. Although the model's ROC AUC score of 0.7805 indicates a higher ability to identify between defaulters and non-defaulters, the total accuracy stayed at 73%. Compared to SVM and Logistic Regression, Random Forest has less false negatives and a more balanced and effective model for forecasting loan defaults, as evidenced by its confusion matrix, which also exhibits greater recall for defaulters.

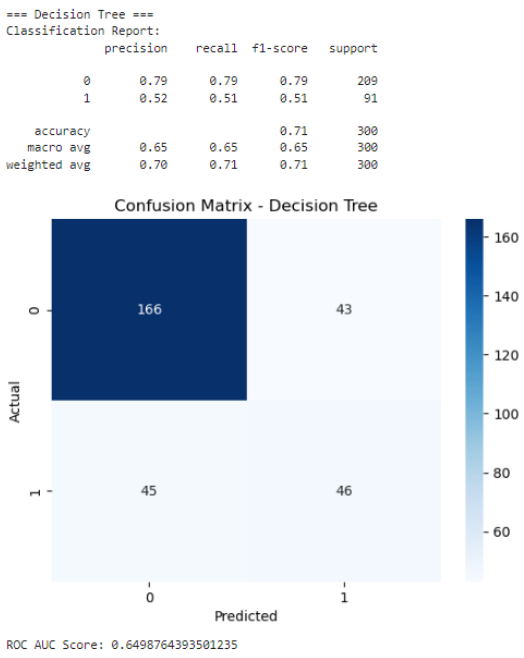
```
=== Random Forest ===
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.92      0.84       209
           1       0.69      0.41      0.51        91

    accuracy                           0.76       300
   macro avg       0.73      0.66      0.68       300
weighted avg       0.75      0.76      0.74       300
```




```
ROC AUC Score: 0.7804826752195173
```

4) Decision Tree

Despite being easier to understand and more straightforward, the Decision Tree model had the lowest overall accuracy (71%). For defaulter prediction, it obtained a precision of 0.52 and a recall of 0.51, indicating a steady but less successful performance. According to the F1-score of 0.51, the model did not do particularly well in either precision or recall, but rather was balanced in both areas. With the lowest ROC AUC score of 0.6499 out of all the models, it was clear that the model was having trouble differentiating between defaulters and non-defaulters. As can be seen from the confusion matrix, the Decision Tree model performed less well overall than the other models since it incorrectly classified a sizable portion of both defaulters and non-defaulters.



Evaluation Metrics

This analysis's main assessment metric was the ROC AUC score, which gauges how well the model can discriminate between positive and negative classifications. In order to shed light on the models' performance in terms of categorisation and emphasise the harmony between recall and precision, confusion matrices were also produced.

Results:

The results of our investigation showed that there were notable differences in how well the four models predicted loan defaults. The Logistic Regression model had a strong ability to predict non-defaulters (class 0) with a precision of 0.73 and a recall of 0.91, resulting in a f1-score of 0.81, even if it only achieved an accuracy of 71% and a ROC AUC score of 0.7353. With a precision of 0.54 and a recall of 0.24, it performed worse in class 1 defaulter prediction, resulting in a lower f1-score of 0.33. The interpretable coefficients produced by this model highlight the importance of individual traits, but its lower recall for defaulters raises the possibility that it may understate risk in crucial circumstances.

With a 71% accuracy rate, the Decision Tree model had the lowest ROC AUC score of 0.6499, suggesting that it was less successful in differentiating between defaulters and non-defaulters. The model achieved a balanced f1-score of 0.79 with precision and recall of 0.79, which is an adequate performance for non-defaulters. With precision and recall both at 0.52, it predicted defaulters less accurately, yielding a f1-score of 0.51 despite this.

The Decision Tree model is simpler and easier to understand, but it is less useful for this task due to its somewhat poorer accuracy and less capacity to balance precision and memory when compared to other models.

The Random Forest model, which had the best accuracy at 76% and the highest ROC AUC score of 0.7805, proved to be the most successful in forecasting loan defaults. With a precision of 0.78 and a recall of 0.92, this model demonstrated remarkable performance in predicting non-defaulters, resulting in an exceptional f1-score of 0.84. The model's recall of 0.41 produced a f1-score of 0.51 despite its somewhat lower accuracy for defaulters (0.69), indicating its improved ability to balance recall and precision. The Random Forest is a dependable and efficient method of controlling financial risk, as evidenced by its great overall performance and improved capacity to differentiate between defaulters and non-defaulters.

With a ROC AUC score of 0.7439 and an accuracy of 73%, the Support Vector Machine (SVM) model performed competitively, but marginally less well than the Random Forest model. SVM produced a strong f1-score of 0.83 with a precision of 0.75 and a good recall of 0.92 for non-defaulters. It had more difficulty, though, in predicting defaulters, as it only managed a f1-score of 0.40 despite achieving a precision of 0.61 and a recall of 0.30. SVM performed similarly to the Decision Tree model, although it was less effective at reliably forecasting loan defaults due to its slightly weaker ability to balance recall and precision, especially for defaulters.

Discussion:

According to the analysis, this dataset's Random Forest model is the best method for forecasting loan defaults. It is a useful tool for the bank's risk management procedures because to its strong ROC AUC score and capacity to capture intricate feature interactions. Even though it is not quite as accurate, the Logistic Regression model is still useful since it may be used to understand the significance of particular features.

The size of the dataset is one of the study's limitations, which can restrict how broadly the results can be applied. Furthermore, typical preprocessing methods were used for the analysis; extra data or feature engineering could potentially enhance model performance. Subsequent research endeavours may investigate more sophisticated models or broaden the dataset in order to improve loan default prediction.

Conclusion:

In summary, the Random Forest model proved to be the most successful and well-rounded method for forecasting loan defaults, despite the fact that each machine learning model displayed distinct advantages. It is a strong option for financial institutions looking to improve their risk management procedures because of its exceptional precision, recall, and ROC AUC score. The Logistic Regression and SVM models are applicable in high-stakes financial decision-making, but they are limited in their capacity to reliably identify defaulters, despite their overall accuracy being respectable. The Decision Tree model performs worse than the other models, which makes it less appropriate for this purpose, even if it is easy to understand and straightforward.

This study highlights how machine learning models, especially the Random Forest, can increase the accuracy of loan default prediction. Financial organisations can successfully reduce risk and make better loan selections by utilising these insights. Subsequent investigations ought to concentrate on broadening the dataset and investigating sophisticated modelling methodologies to further hone and improve these models' prediction capacities concerning loan defaults.