# Comprehensive Report on Credit Risk Classification using Logistic Regression and Random Forest

## 1.0 Introduction

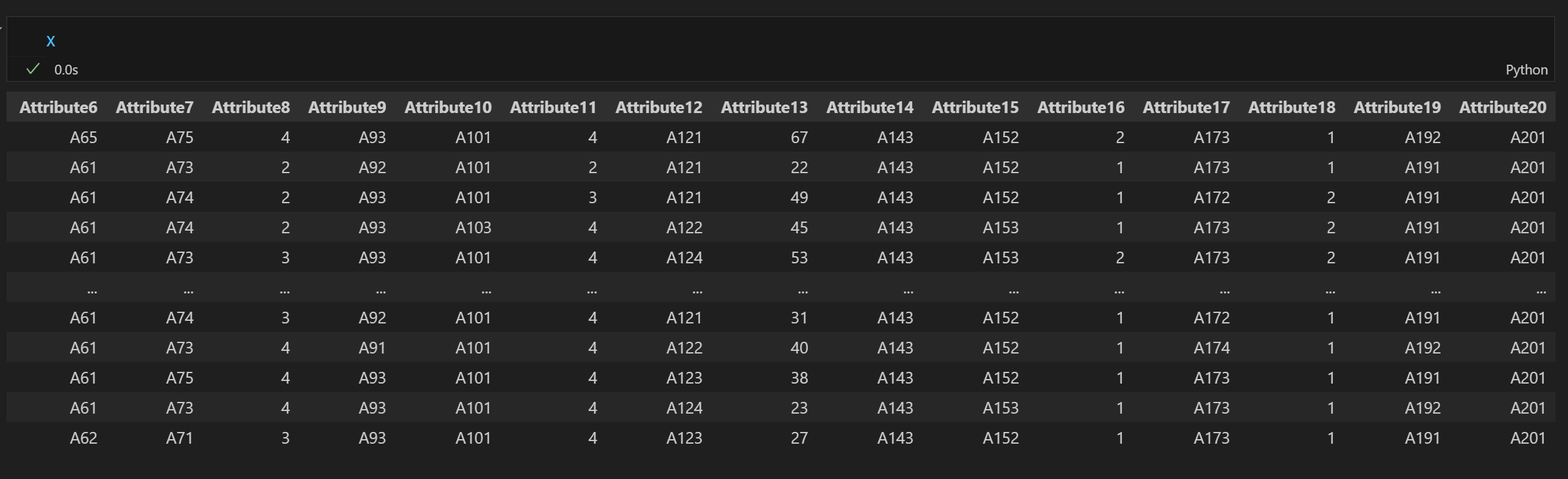
Logistic Regression is a statistical strategy used to solve binary problem classification in which one of two probable outcomes is predicted (Rogers, 2007).It simulates the likelihood of a binary response given one or more predictor variables. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities using a logistic function (sigmoid function) with output values ranging from 0 to 1 (Jain, 2010). Random Forest is an ensemble learning algorithm that may be used for both classification and regression applications. During training, it generates numerous decision trees and returns the mode (for classification) or mean (for regression) of their predictions (Cunningham et al, 2007)

The classification dataset was credit risk, which was obtained from the UCI Machine Learning Repository via Statlog (German Credit Data). The dataset includes 1,000 instances and 20 attributes regarding people's credit histories and personal information, which are critical for determining if a person's credit risk is "good" (1) or "bad"(2). In this research, Logistic Regression and Random Forest Classifier were used to predict the target class, with the models' performance measured using measures such as accuracy, precision, recall, F1-score, AUC, and precision-recall curves.

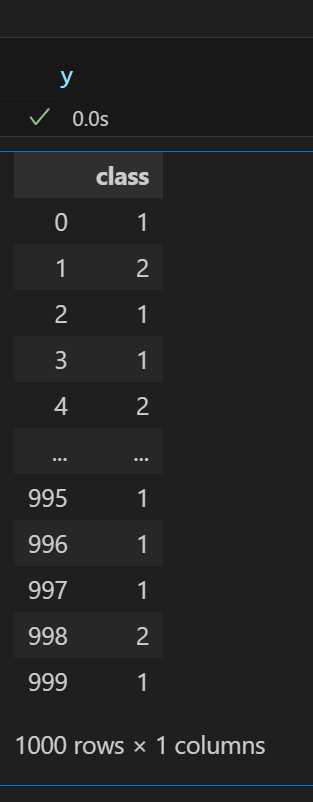
## 2.0 Dataset Loading and Exploration

Loading the dataset and examining its structure was the first step. The ucimlrepo library, which makes it easier to download datasets straight into a pandas DataFrame, was used to retrieve the dataset from the UCI repository. The Statlog German Credit Data was loaded using the fetch\_ucirepo(id=144) function. It included 20 features, which were renamed as Checking Account Status, Duration (months), Credit History, Purpose, Credit Amount, Savings Account/Bonds, Employment Duration, Installment Rate (%), Personal Status and Sex, Other Debtors/Guarantors, Residence Duration, Property, Age (years), Other Installment Plans, Housing, Existing Credits Count, Job, Dependents Count, Telephone Availability, Foreign Worker, and y-'Credit Risk (the target variable indicating whether a credit is good or bad).

In order to better comprehend the dataset, the metadata and variable information were then displayed.

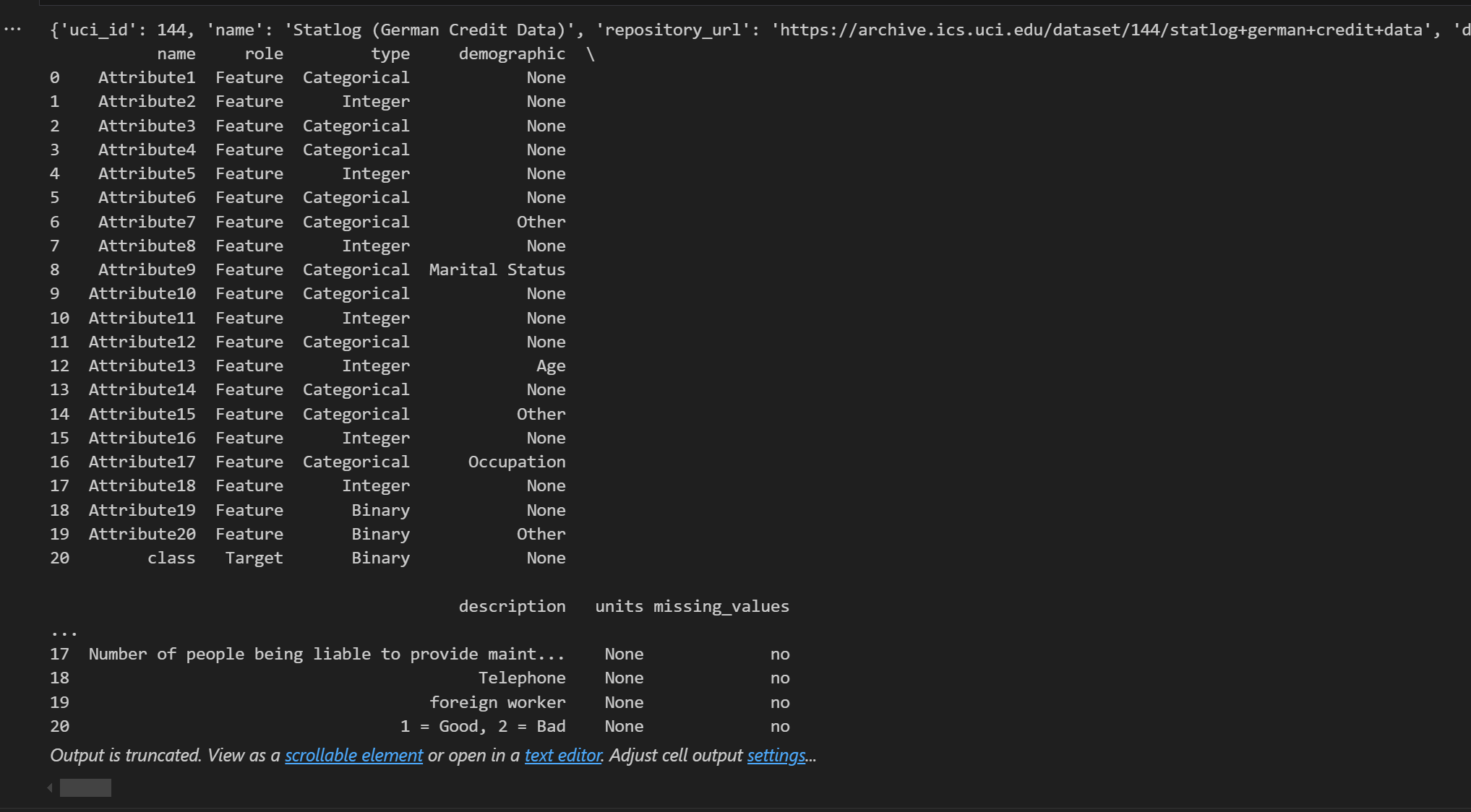


*Dataset Features*



*Dataset target*

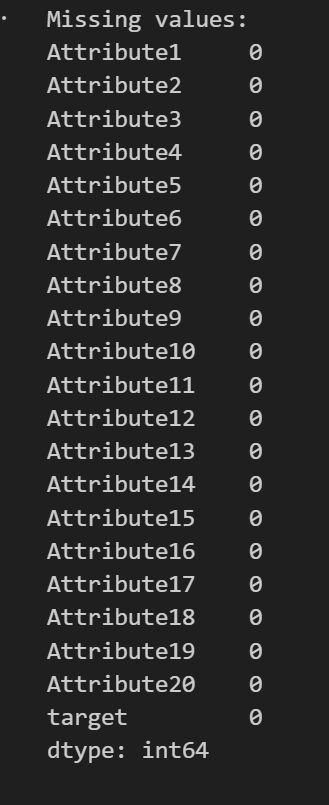
The target class, class, was binary (1 = Good, 2 = Bad), according to the metadata, which also showed that the dataset was utilised for classification. The variables discussed included factors that are important in assessing credit risk, such as checking account status, credit history, credit purpose, savings, length of employment, and others.



*Dataset metadata*

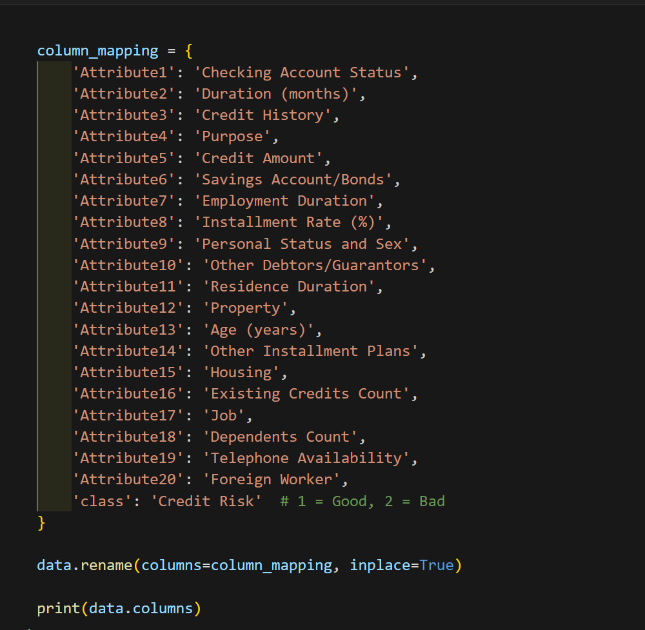
## 3. Data Preprocessing and Data Exploratory Analysis

A few preprocessing processes were carried out prior to model training(Hastie et al, 2009). Although there were no missing values in the dataset, it was cleaned by looking for them. To examine the distribution of numerical data and comprehend its range, mean, and standard deviation, summary statistics were printed.



*No missing value in the data set*

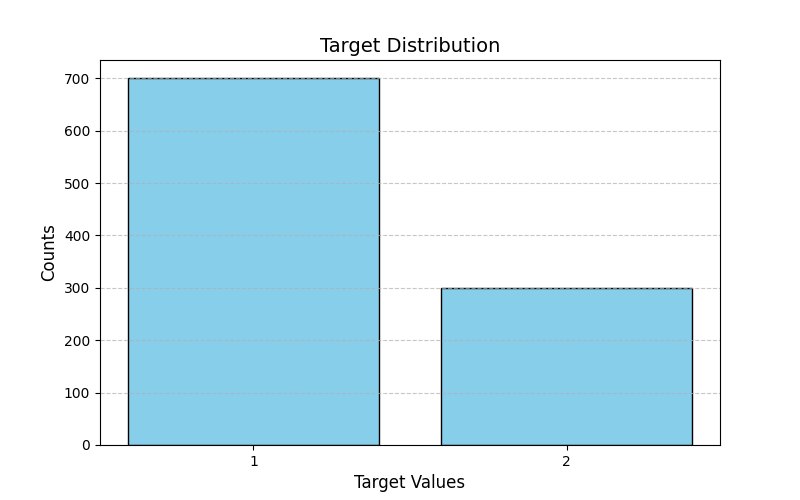
The dataset was made easier to deal with by renaming the columns to more accessible names. This renaming made the dataset more readable and clear.



*Renaming the columns.*

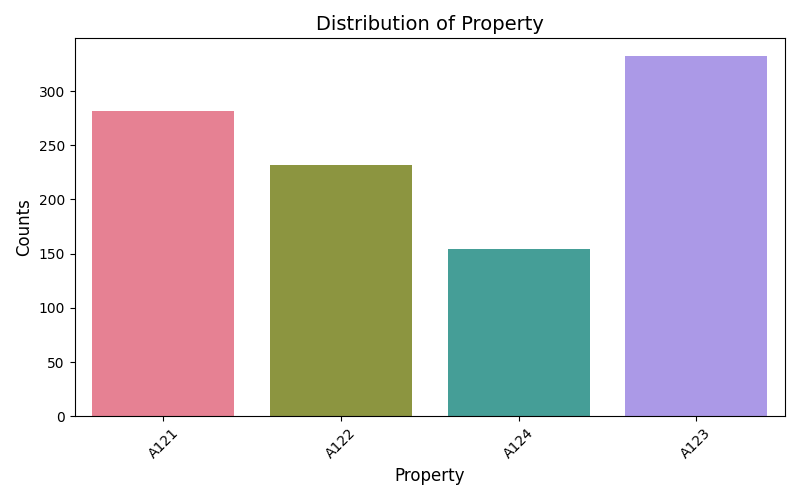
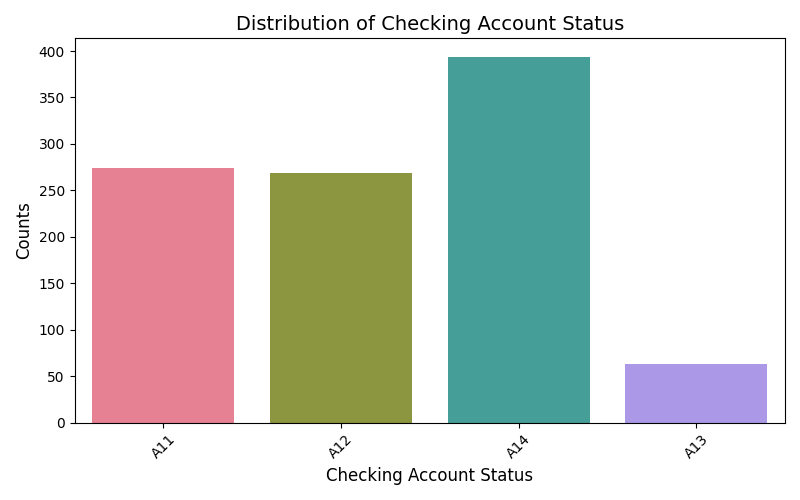
## 3.1 Data Exploratory Analysis Visualization

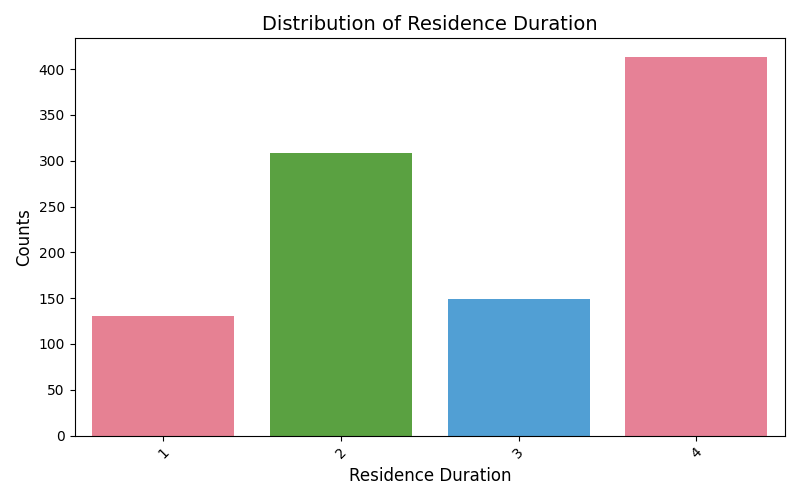
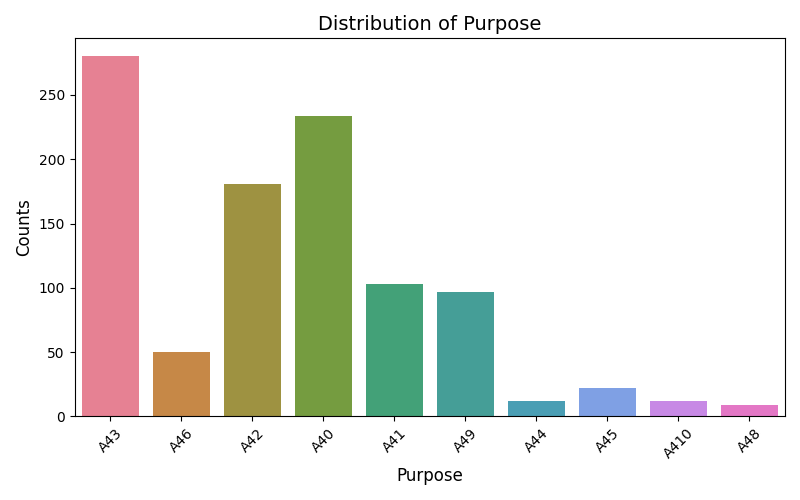
EDA, or exploratory data analysis, was done to learn more about the dataset(Bishop, 2006). The target variable's (Credit Risk) distribution was shown. For the "Good" and "Bad" credit classes, the counts were shown as a bar plot. Because it indicated whether the dataset's class distribution was balanced or unbalanced, this phase was essential ( Kuhn et al 2013).

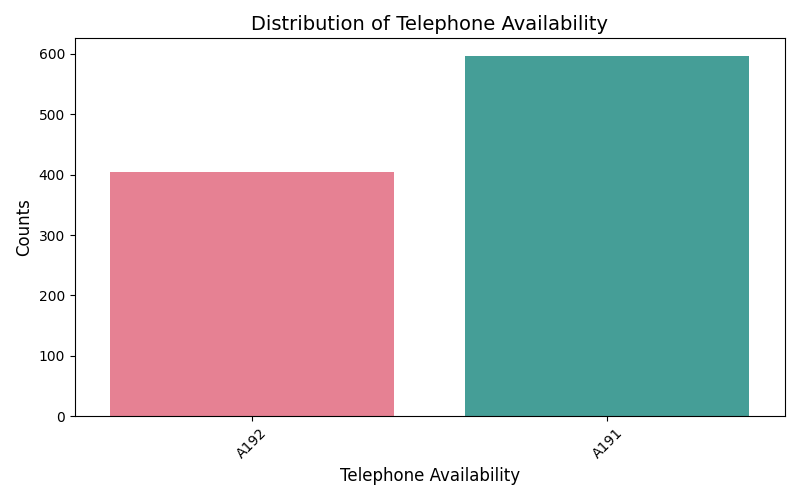
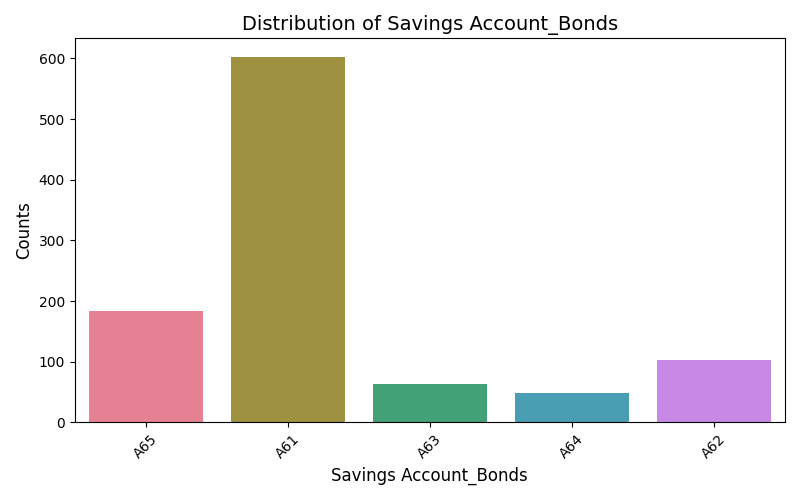


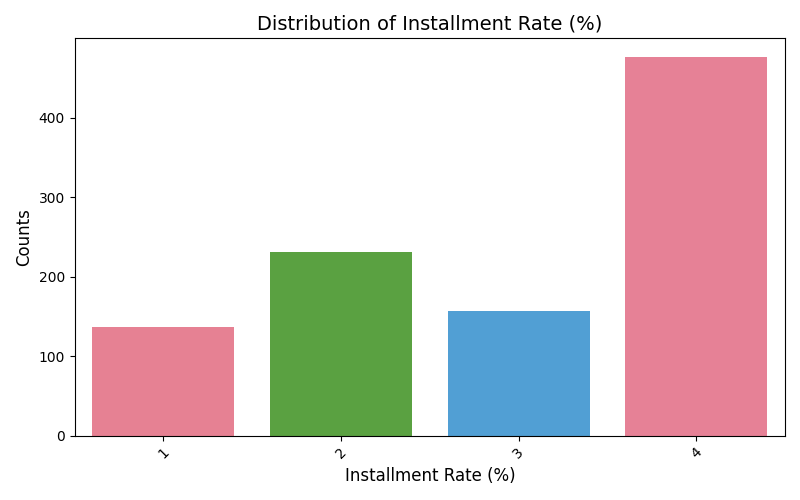
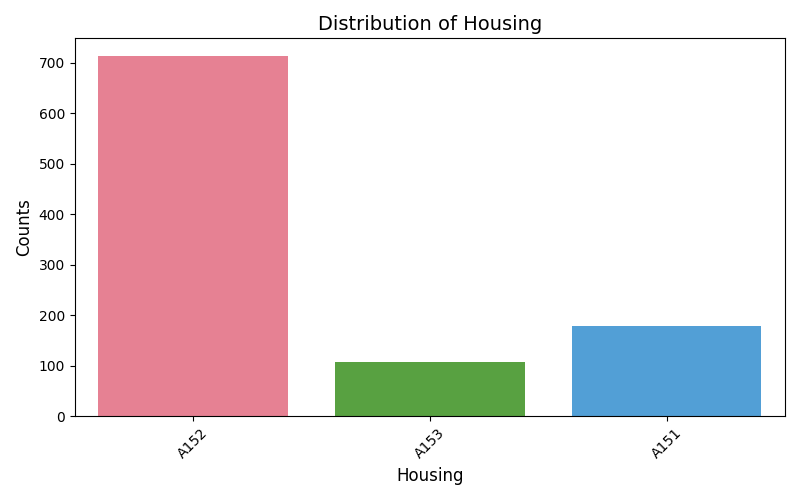
*Target or class distribution*

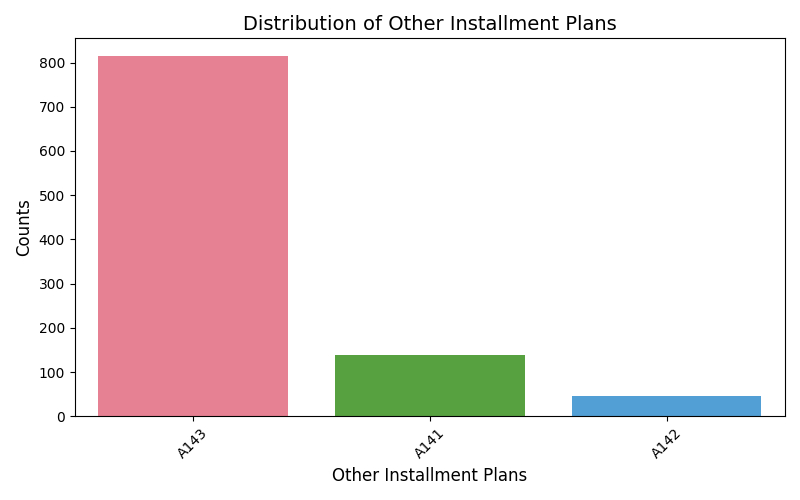
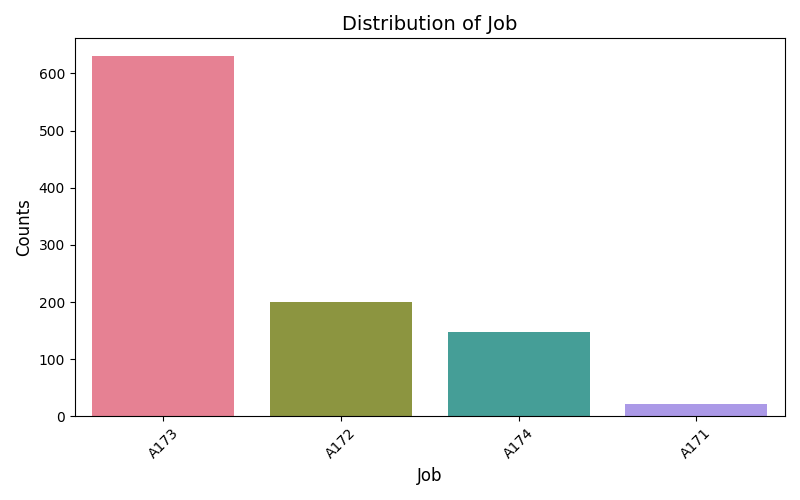
Using countplots, feature distributions for every column were analyzed (Van et aj, 2009). "Checking Account Status," "Credit History," and "Purpose" were examples of categorical elements that were visualized in order to comprehend their distribution. It made the categorical data distribution easier to see. For instance, "Personal Status and Sex" displayed the applicants' gender and marital status distribution, while "Purpose" displayed the loan types that were most prevalent in the dataset.

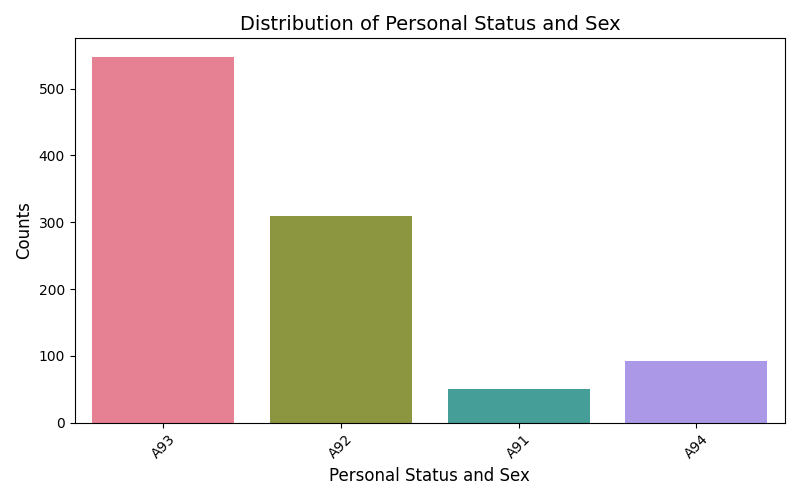
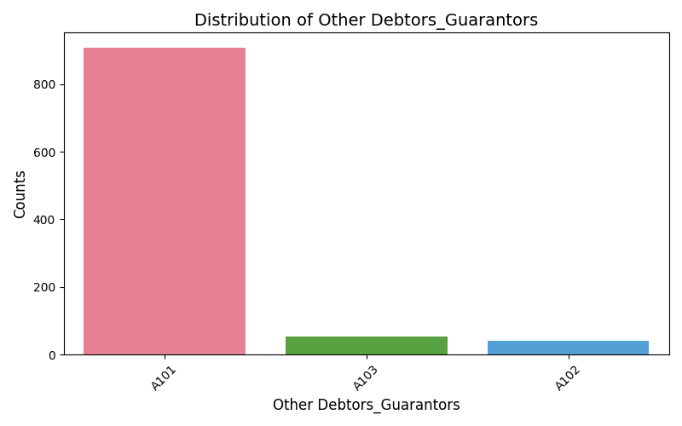


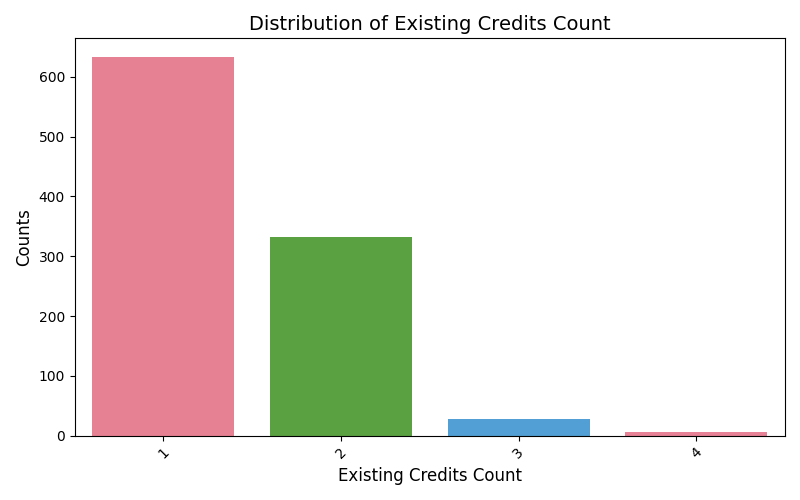
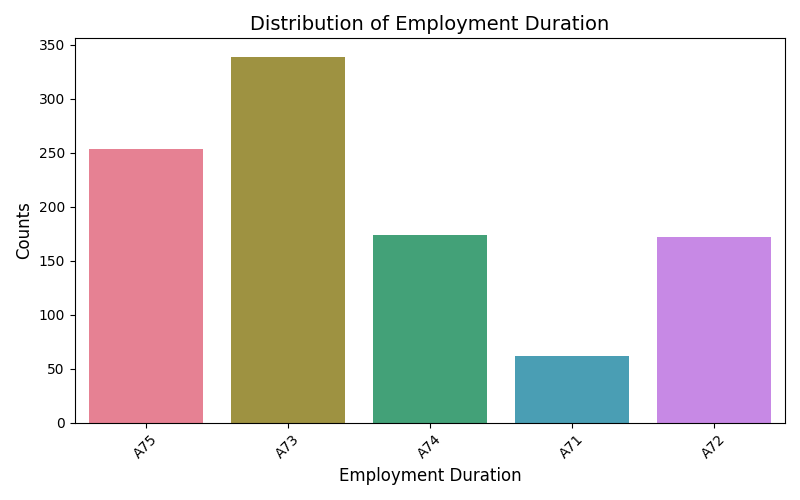
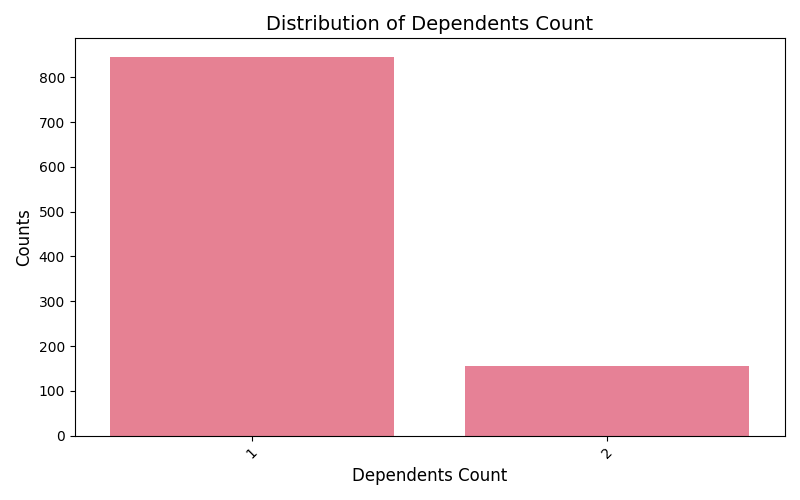
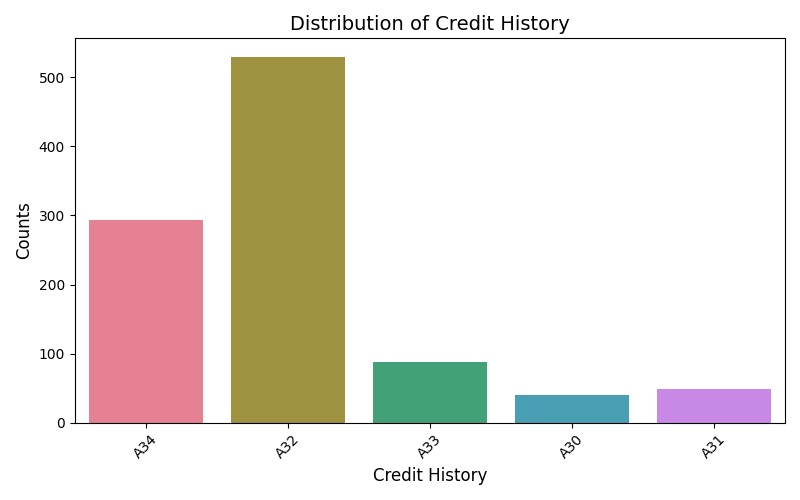










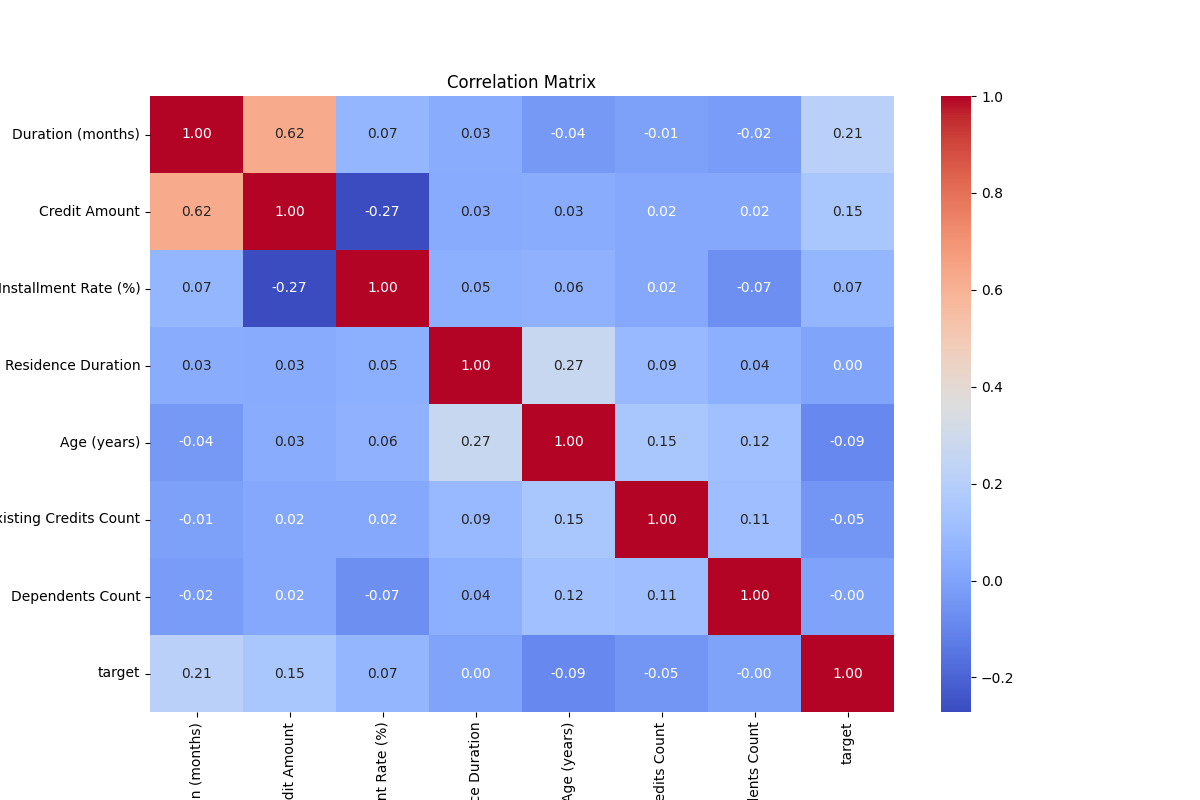


*Features distribution*

## 3.2 Correlation Analysis

To investigate correlations between numerical features, a correlation matrix was generated and visualised as a heatmap. This matrix highlighted the strength and direction of correlations among several qualities. For example, it revealed that "Credit Amount" and "Duration (months)" had moderate associations, implying that larger loan amounts associated with longer durations.

The correlation matrix was an effective tool for identifying how different qualities in the dataset were related to one another, which is critical for feature engineering and picking meaningful predictors for the models (Cunningham et al, 2007)..



*features distribution*

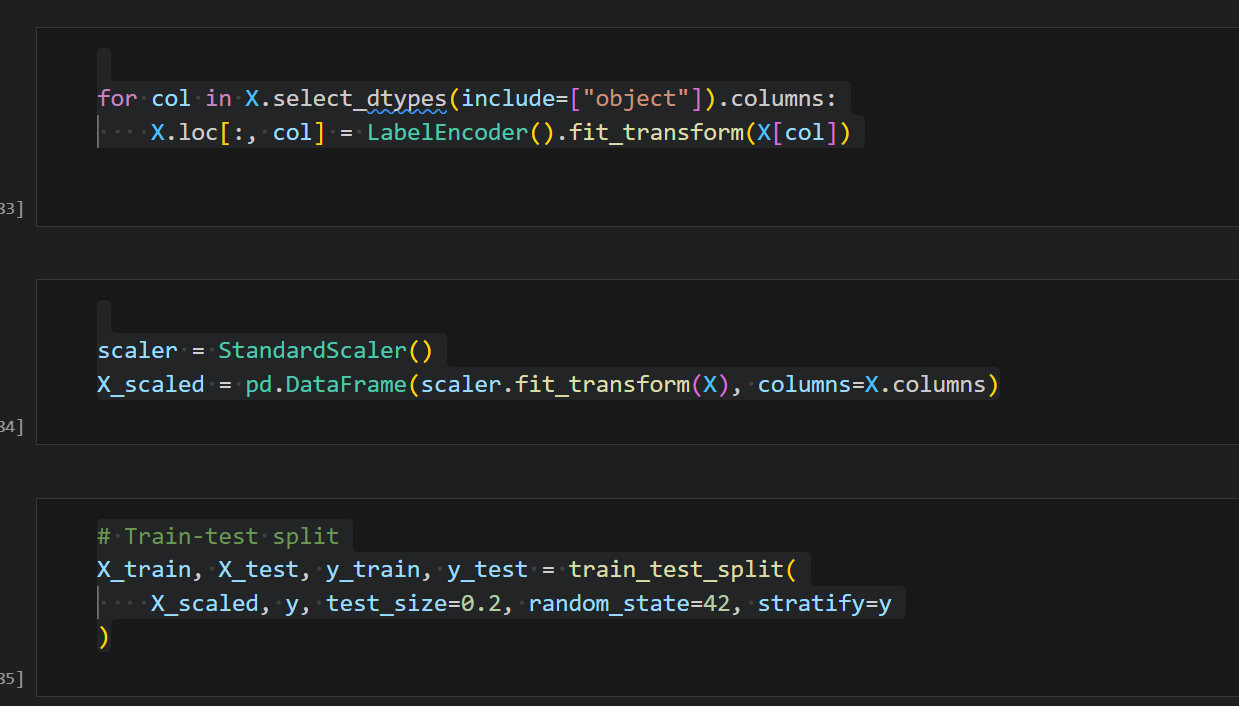
Also, a pairplot was used. The pairplot function was used to visualise the correlations between the dataset's many numerical features, as well as their link to the target variable(Hastie et al, 2009). The figure included scatterplots for each pair of features, as well as histograms or density plots for individual features on the diagonal. The pairplot highlighted the distribution of both "Good" and "Bad" credit risk categories by utilising target variable-specific colour coding. This gave useful information on potential correlations between features and the target variable, as well as how well the features distinguished the two classes (Rogers, 2007) .



*Pairplot visualization*

## 4. Feature Encoding and Scaling

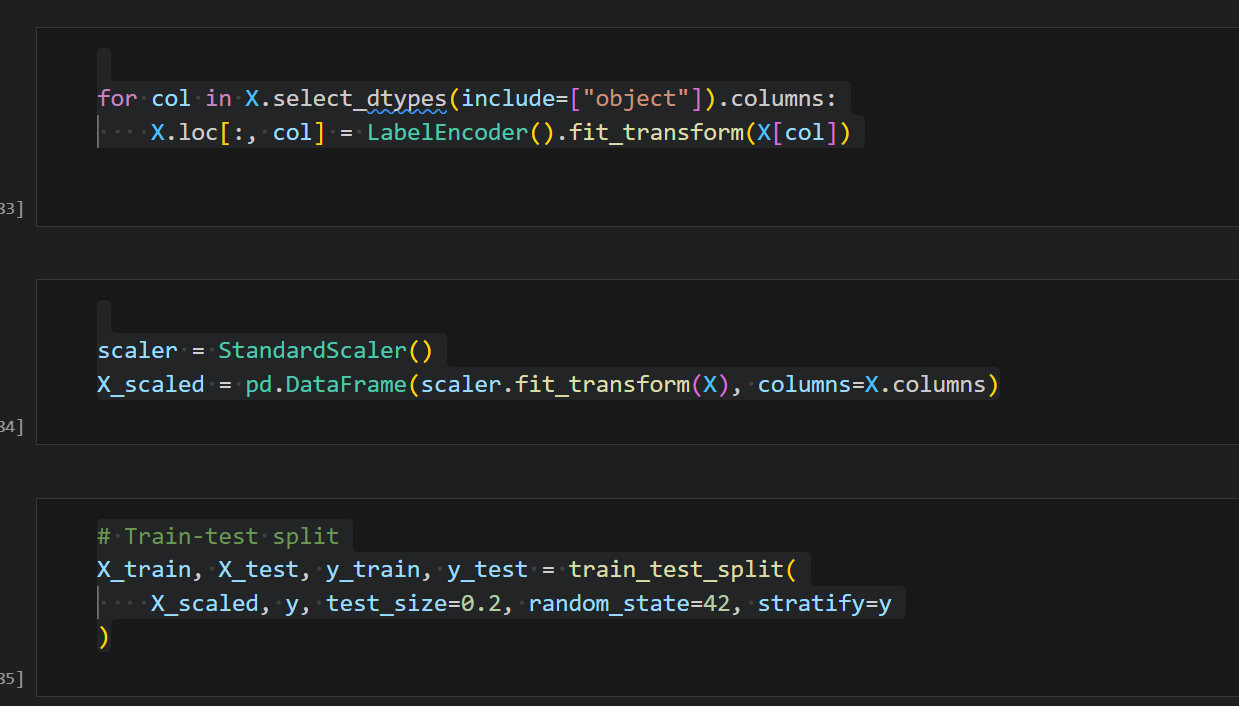
The dataset included both categorical and numerical characteristics; hence, feature encoding for categorical variables and scaling for numerical features were required. To convert categorical variables to numeric representations, label encoding was utilized. Numerical characteristics were then scaled using StandardScaler to ensure a consistent range. These pretreatment methods guaranteed that the models received data in an appropriate manner, boosting their performance and stability (Kuhn et al 2013).



*Code for label encoding, standardization*

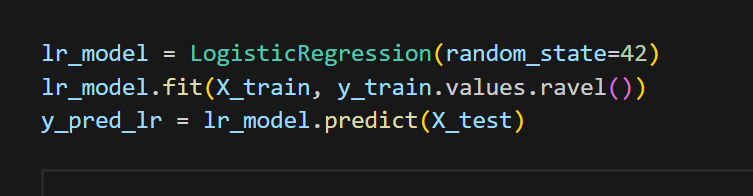
## 5. Logistic Regression

The preprocessed data was analysed using Logistic Regression. This model was chosen because it is a straightforward, understandable model often used for binary classification tasks (Bishop, 2006). The model was trained using an 80% training split and tested using a 20% test split.



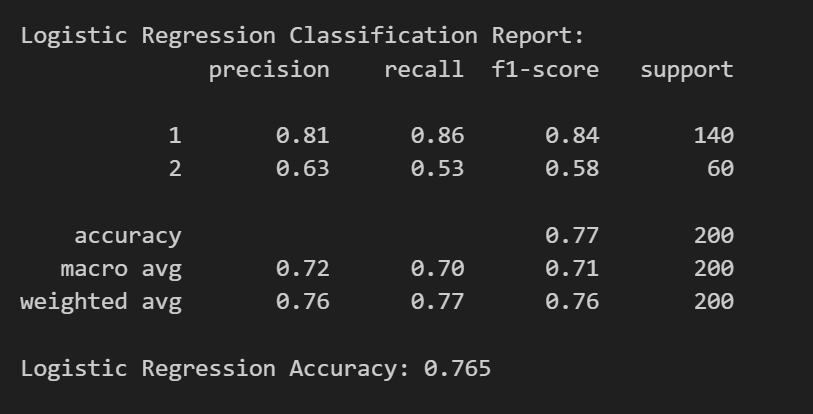
*train-test split*

The logistic regression model was then assessed on the test set using a variety of measures, including accuracy, precision, recall, and F1-score. The confusion matrix was visualized to show how effectively the model predicted both "good" and "bad" credit cases (Jain, 2010).



*Logistic Regression code*

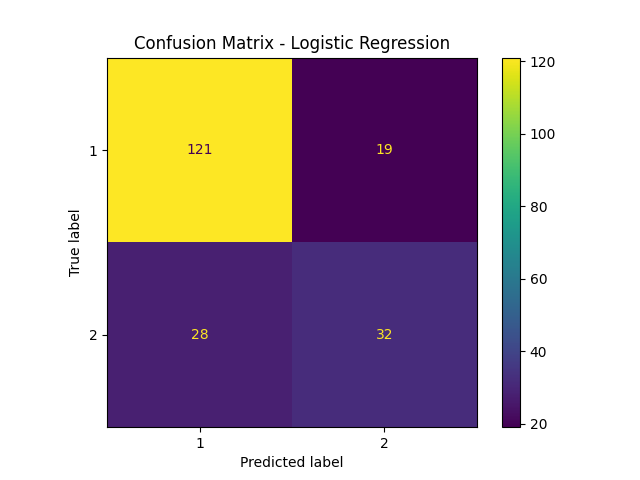
The logistic regression model's accuracy score of 76.5% suggested that it had acceptable predictive capacity, but there was still potential for improvement, particularly in the "Bad" credit situations (precision for class 2) (Van et aj, 2009).



*Logistic Regression classification report*

## 5.1. Confusion Matrix

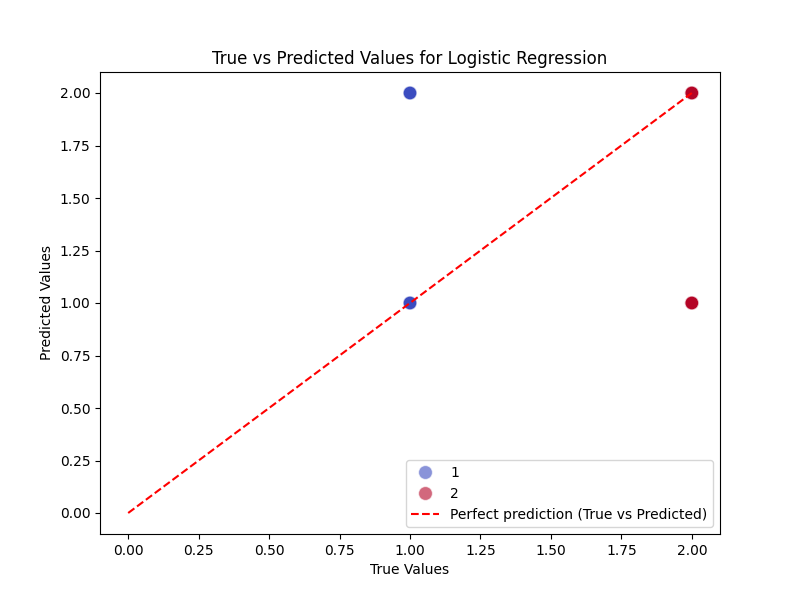
The Logistic Regression model's performance was assessed using the Confusion Matrix. (Hastie et al, 2009). It displayed the true positive, true negative, false positive, and false negative counts, giving a complete picture of the model's ability to categorise occurrences of good and bad credit risk (Rogers, 2007) The confusion matrix assisted in identifying regions where the model made errors, such as misclassifying "bad" credit as "good," which is critical for assessing its precision and recall (Cunningham et al, 2007)..



*Confusion Matrix for Logistic Regression*

## 5.2. True vs Predicted Values

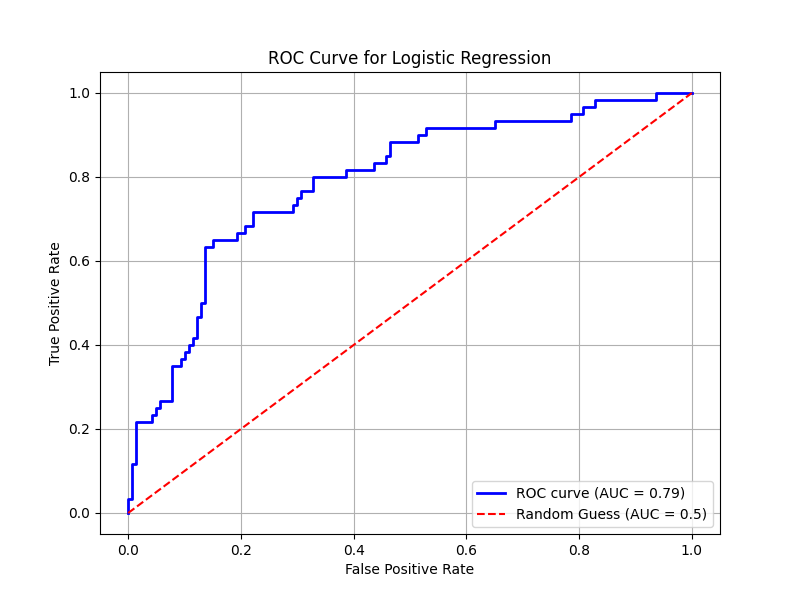
A scatterplot comparing the true and predicted values for the Logistic Regression model demonstrated how closely the model's predictions matched the actual labels (Jain, 2010). A red dashed line representing flawless predictions was added, with spots spread around it indicating departures from the ideal forecasts. This visualisation demonstrated the model's accuracy and ability to accurately estimate credit risk, with closer alignment to the red dashed line indicating improved performance (Kuhn et al 2013).



*True vs predicted*

## 5.3. **ROC Curve for Logistic Regression**

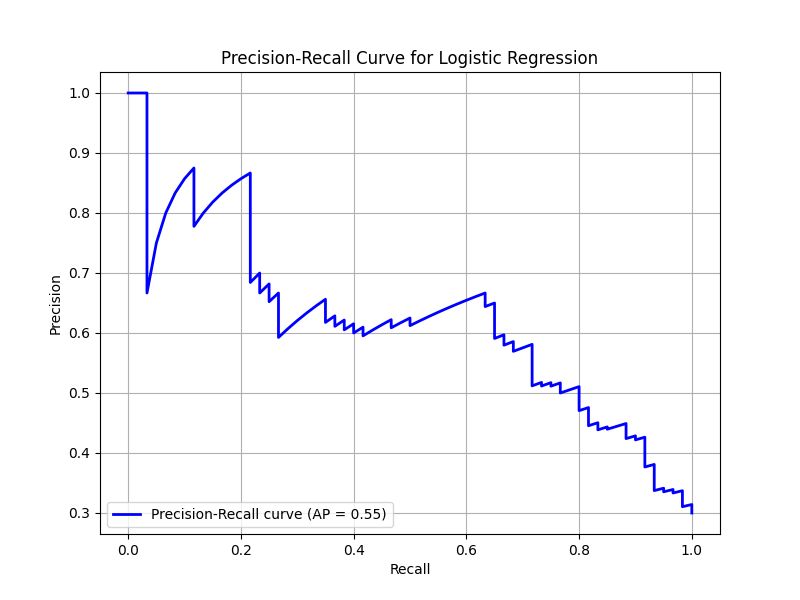
The ROC Curve for the Logistic Regression model(Van et aj, 2009). The Receiver Operating Characteristic (ROC) curve depicts the model's ability to discriminate between positive and negative classes as the decision threshold varies (Bishop, 2006). The curve represented the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. The area under the curve (AUC), which was generated as a specific value, represented the classifier's overall performance. A higher AUC indicates greater model performance, with values closer to 1 indicating good classification skill and 0.5 implying random guessing. The graphic also contained a red dashed line indicating random guessing and a blue line showing Logistic Regression model’s performance (Hastie et al, 2009).



*ROC CURVE for Logistic classification*

## 5.4. **Precision - Recall Curve**

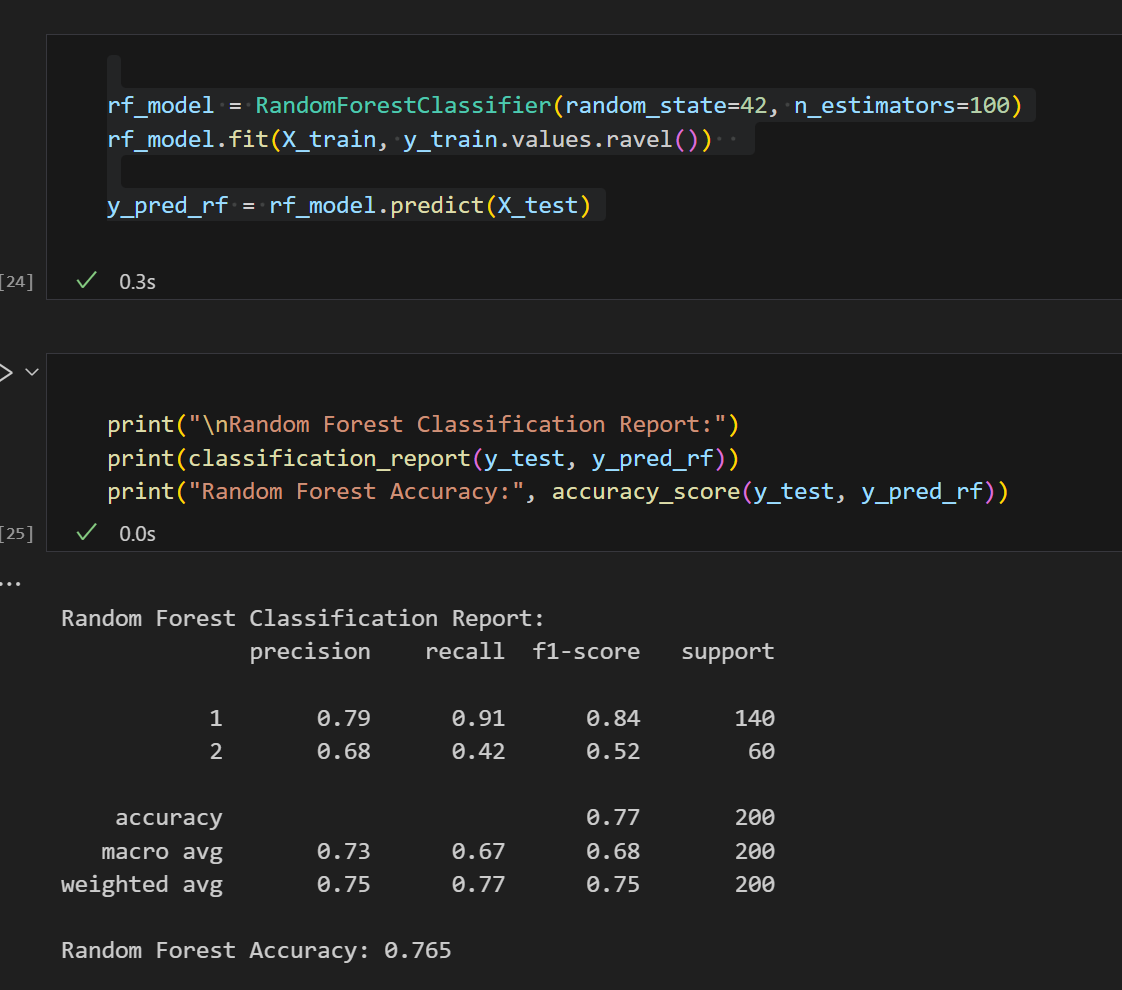
The Precision-Recall Curve, which is especially useful for imbalanced datasets, depicts the trade-off between precision and recall as the decision threshold changes( Kuhn et al 2013) Precision represents the percentage of true positive predictions out of all positive forecasts, whereas recall represents the percentage of true positives accurately identified. The Average Precision (AP) score was produced and shown in the legend as a single number that summarized the model's precision-recall performance (Rogers, 2007). A higher AP suggests that the model maintained good precision and recall at various thresholds. This curve is very useful when evaluating models for imbalanced classification issues because it focusses on the performance of the positive class (Bishop, 2006)



*Precision-Recall Curve*

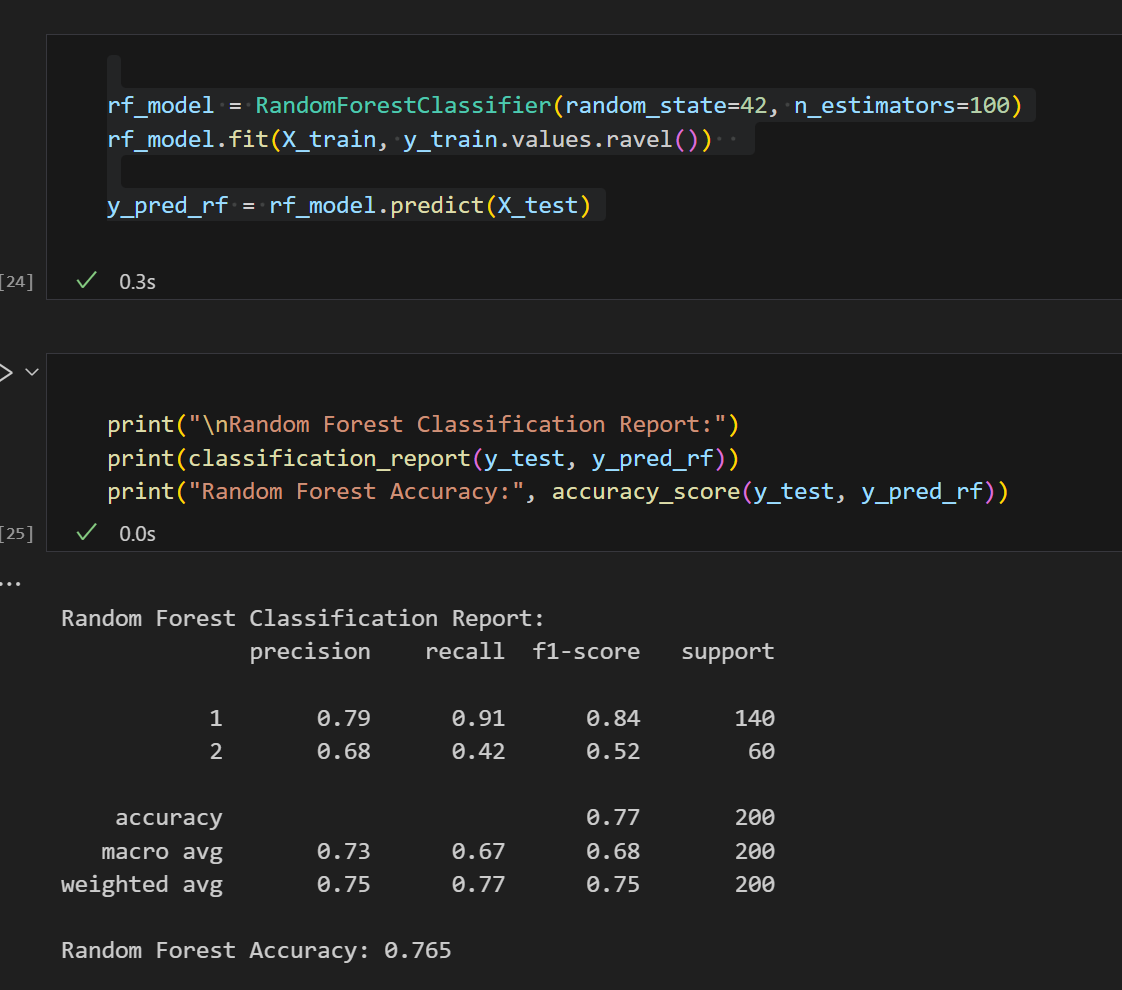
## 6. Random Forest Classifier

A Random Forest Classifier was also used to analyse the dataset. Random Forest is an ensemble method that creates numerous decision trees during training and returns the mode of the classes as the final forecast. This approach is well-known for its reliability and ability to handle complex, nonlinear interactions (Cunningham et al, 2007).



*Random Forest Classifier code*

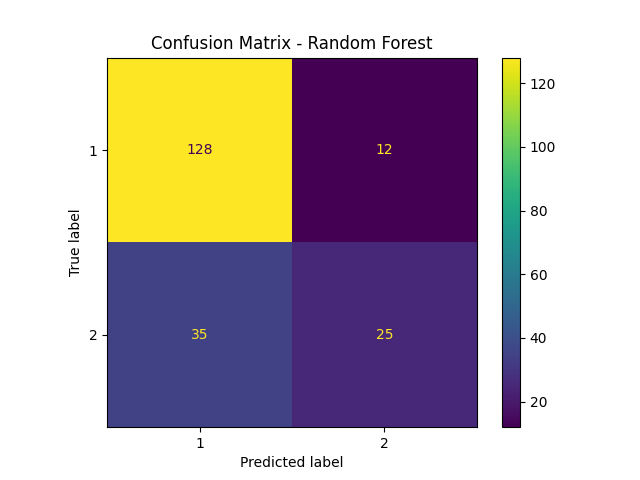
The Random Forest model outperformed Logistic Regression, with a 76.5% accuracy score. Feature significance was also computed to discover which features made the most significant contributions to the model's predictions. "Credit History" and "Credit Amount" appeared as the most influential factors in forecasting credit risk.



*Random Forest Classification report*

## 6. 1 Random Forest Confusion Matrix

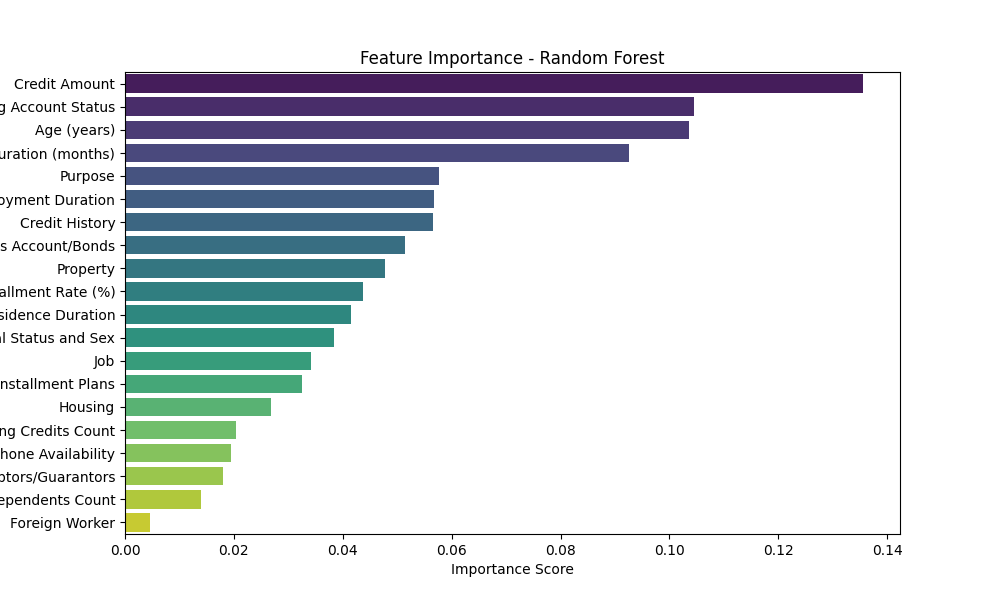
The Random Forest Confusion Matrix gave a detailed description of the model's predictions, including the counts of true positives, true negatives, false positives, and false negatives (Van et aj, 2009). This graph demonstrated how well the Random Forest classifier identified between the two credit risk classes, "Good" and "Bad". The matrix showed how many cases were correctly identified and where the model made mistakes, with a special emphasis on misclassifications between the two classes.



*Random Forest Confusion matrix*

## 6. 2 Random Forest Feature Importance

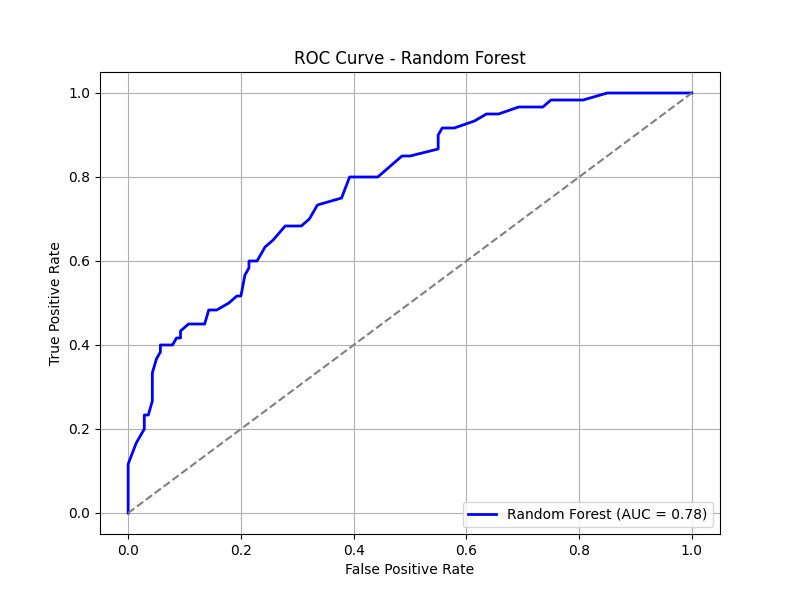
The Random Forest is well-known for its capacity to determine the value of each feature in making predictions (Rogers, 2007). The feature importance plot displayed the relative importance scores of each feature, allowing us to see which variables were most influential in forecasting credit risk. Features with higher significance ratings were thought to play a bigger impact in the model's decision-making process(Cunningham et al, 2007). The plot made it easier to identify the most relevant features, such as "Credit Amount" and "account status," which contributed the most to credit risk rating. This type of analysis is critical for feature selection and can drive future model development or assist in identifying the underlying trends in the data.



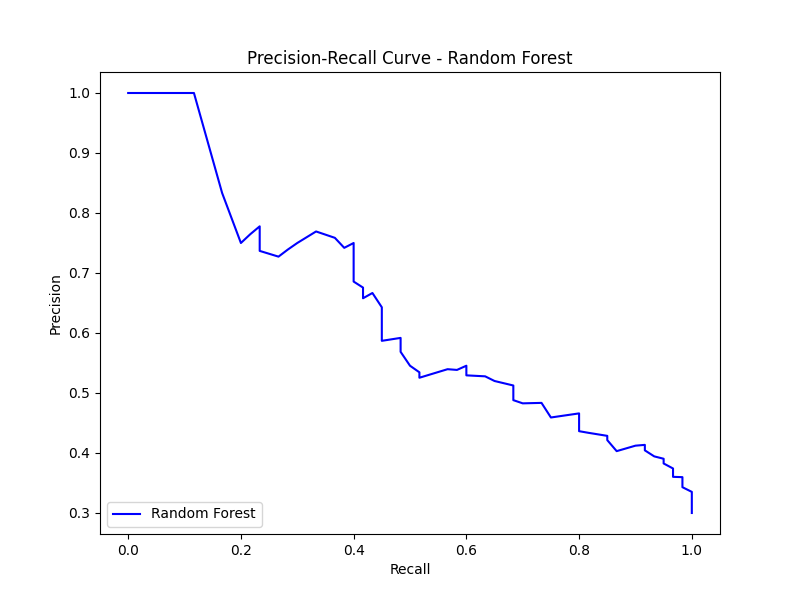
*Random Forest’s Feature importance*

## 6.3 ROC AND Precision-Recall Curve

Various measures, such as the ROC Curve and Precision-Recall Curve, were used to evaluate the models (Hastie et al, 2009). The ROC Curves for Logistic Regression and Random Forest indicated that both models outperformed random guessing, with Random Forest having a higher AUC (Jain, 2010).   
These evaluations shed light on the models' ability to distinguish between "Good" and "Bad" credit instances under various threshold circumstances.



*ROC CURVE*



*Precision-Recall Curve for Random Forest*

## Conclusion

The investigation of the German Credit dataset yielded useful insights into the elements that influence credit risk and demonstrated the efficacy of machine learning approaches. We investigated prediction patterns using Logistic Regression and Random Forest models, with each strategy providing distinct advantages. Logistic Regression established a baseline with interpretable results, whereas Random Forest provided a more nuanced knowledge of feature relevance, indicating "Credit History" and "Credit Amount" as crucial predictors (Jain, 2010).

The use of visualizations, such as confusion matrices and feature significance plots, revealed the models' performance and places for development (Bishop, 2006). The ROC and accuracy-Recall curves further proved the models' capacity to balance true positive rates and false positives, as well as accuracy and recall, providing key insights for evaluating predictive quality (Van et aj, 2009).

This study emphasized the need of data preparation, careful model selection, and thorough evaluation in generating dependable classification results. These findings not only improve our understanding of credit risk assessment, but also pave the way for the development of predictive models to aid in real-world financial decision-making ( Kuhn et al 2013).

## References

1. UCI Machine Learning Repository. (n.d.). *Statlog (German Credit Data)*. Retrieved from <https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)>
2. Jain, A. K. (2010). Data clustering: 50 years beyond K-means.  
   *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(7), 645-666.*
3. Cunningham, P., & Delany, S. J. (2007). k-Nearest Neighbor Classifiers.  
   *Multiple Classifier Systems, 1-17.*
4. Rogers, S., & Girolami, M. (2007). A tutorial on kernel methods for pattern analysis.  
   *International Journal of Neural Systems, 17(04), 217-234.*
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.  
   *Springer Science & Business Media.*  
   [ISBN: 978-0387848570]
6. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling.  
   *Springer Science & Business Media.*
7. Bishop, C. M. (2006). Pattern Recognition and Machine Learning.  
   *Springer.*
8. Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review.  
   *Journal of Machine Learning Research, 10, 66-71.*

(Van et aj, 2009) (Bishop, 2006)( Kuhn et al 2013) (Hastie et al, 2009) (Rogers, 2007) (Cunningham et al, 2007) (Jain, 2010)