Report
Student ID:210899247

Report
Student ID:210899247

Section one: Introduction

Within the fast-shifting world of financial technology, one of the most cutting-edge forms of doing business is epitomised by peer-to-peer lending models such as LendingClub, which this study will work with and which has proven to be an effective substitution for traditional banking practices. Apart from providing investors with an opportunity to diversify their investment portfolios and providing access to efficient, low-cost personal loans, LendingClub creates a direct link between investors and borrowers. However, it must be noted that this new system carries enormous risks, particularly in dealing with loan defaults, which LendingClubdefines as 'Charged Off.'

This study implements machine learning to tackle the issue of default prediction using a dataset of loans issued between 2007 and 2015. The dataset has seven predictive variables, including borrowing attributes (annual income), loan characteristics (loan amount, interest rate), and payment behaviours (instalments, last payment amount, total current balance). The target variable, loan_status, indicates whether a loan was fully repaid or charged off.

The study employs four machine learning models to address the challenge of predicting loan defaults. Each model provides distinct strengths and addresses a different aspect of this problem :

1) Logistic Regression – linear model that provides a computationally efficient and interpretable baseline for binary classification tasks
2) Classification Tree(Without Cross-Validation) – A decision tree model that segments the dataset into hierarchical decision rules based on feature importance
3) Classification Tree(With Cross-Validation) – Variation that incorporates hyperparameter tuning and cross-validation techniques to enhance generalisation and prevent overfitting
4) Random-Forest – An ensemble learning method combining decision trees to improve accuracy and robustness.

The Cross-Validated Classification Tree was the optimal model for achieving the highest AUC, indicating it correctly identifies most of the defaults. Random-Forest followed closely. Its ensemble nature made it robust against overfitting and provided reliable predictions across the dataset. The Logistic Regression, on the other hand, while computationally efficient, achieved a moderate AUC but with low precision, producing a high rate of false positives. Lastly, the classification-tree (Without cross-validation) demonstrated the risks of overfitting, achieving the lowest AUC, and underperforming compared to the cross-validated version. It suffered from high variance and highlighted the importance of cross-validation in improving generalisation.

It would be reasonable to assume that among the models tested, the random-forest would be the most optimal model since it automatically merges the best decision trees;

nonetheless, the cross-validated classification tree resulted as the most optimal model.

Section 2: Methodologies:

Before making any analysis, substantial data cleaning had been done to discard inconsistencies and prepare data for modeling to ensure the accuracy of the predictions.

A preliminary examination of the dataset revealed its structure, column names, and missing values. The dataset contained a significant number of missing values. The isna().sum() function indicated 7017 missing entries in the total current balance. The missing entries were removed using dropna(), reducing dataset size while ensuring consistency between variables.

Further, a dummy variable has been created to make categorical variables machine-readable, specifically a binary variable called loan_status, which classifies loans as Charged Off as 1 or Fully Paid(and others) as 0. Then, the standard scaler function has been used to allow the comparability of like-for-like by standardising the data. After the data was cleaned up and prepared, it was split into training-77% of data and testing-27% of data subsets.

Furthermore, exploratory data analysis was conducted to visualise key aspects of the dataset. Afterwards, histograms were implemented to examine loan amount distributions and interest rate distributions, providing insights into typical values. Then, scatterplots explored relationships between variables such as loan amounts and instalments and loan amount and interest rate, while a correlation heatmap summarised the strengths of correlation across all numerical predictors.

The following models were used to analyse the dataset:

1) Logistic Regression:

- Logistic regression is a model used as an alternative to linear regression for binary classification problems. It calculates the probability associated with a binary outcome given some predictors using a logistic function mapping the predictions in a range between 0 and 1. It was used in the project to classify loans as defaulted or non-defaulted

• Advantages:

- 1. Coefficients directly show the influence of predictors on the probability of target class. They provide a meaningful prediction which lies between 0(0%) and 1(100%)
- 2. Easy to implement and computationally efficient for large datasets
- 3. Less prone to overfitting compared to models like decision tree

• Disadvantages:

- 1. A significant limitation of Logistic Regression is that it assumes linearity between the dependent variable and independent variables. They are, therefore, not suitable for capturing non-linear relationships.
- 2. It can only be used to predict discrete functions. Consequently, the dependent variable of Linear Regression is limited to the discrete number set only

2) Classification Tree(Without Cross-Validation)

A classification tree splits the dataset into two branches on a feature threshold(stratification) that maximises purity at each split. As we aim to predict a discrete variable(dummy), we are using a classification tree, not a regression tree. The resulting tree structure represents decision rules for classification.

- Advantages:

- 1. The visual flowchart format of the decision tree facilitates an explanation of how a prediction is made.
- 2. Automatically determines the most important features by splitting the dataset into the variables that provide the highest information gain or Gini index reduction

- Disadvantages:

- 1. Decision trees can grow overly complex and fit noise in the training data, resulting in overfitting and poor generalisation of unseen data
- 2. They suffer from high variance, meaning that small changes in data can lead to a completely different tree structure, making decision trees sensitive to variations in training data

3) Classification Tree (With Cross-Validation):

- Cross-validation involves splitting the dataset into multiple training and testing subsets to evaluate the model's performance across various data subsets. Using cross-validation, three hyperparameters were implemented to control the maximum depth, minimum samples needed to be split, and minimum number of samples required to be at a leaf node.

- Advantages:

- 1. Validation prevents the tree from overfitting the training data since it is trained and validated on different subsets
- 2. Using multiple folds provides a more reliable estimate of model performance than a single train-test split.

- Disadvantages:

- 1. Cross-validation can be computationally expensive, especially with large datasets like ours
- 2. The cross-validation does not fully address the issue of variance, as a small k can result in a high variance in performance estimates, while a large k can be computationally expensive

4) Random-Forest:

- Random-Forest is an ensemble method that combines predictions from multitudinous decision trees. Each tree is trained on a random subset of data and features; this reduces variance and improves generalisation.

- Advantages:

- 1. It achieves high accuracy by reducing overfitting problems in decision trees and helps to improve accuracy. It reduces prediction variance compared to a single decision tree
- 2. Handles both numerical and categorical data. It can handle outliers and missing values.
- 3. Can handle large datasets with high dimensionality.

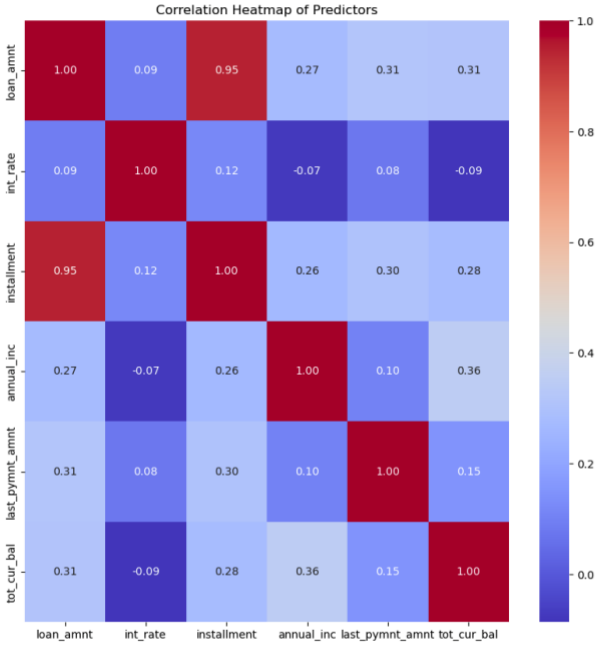
- Disadvantages:

- 1.Less interpretable than a single decision tree since the prediction cannot be explained only by a diagram
- 2. Random-Forest might be computationally expensive, especially when working with large datasets

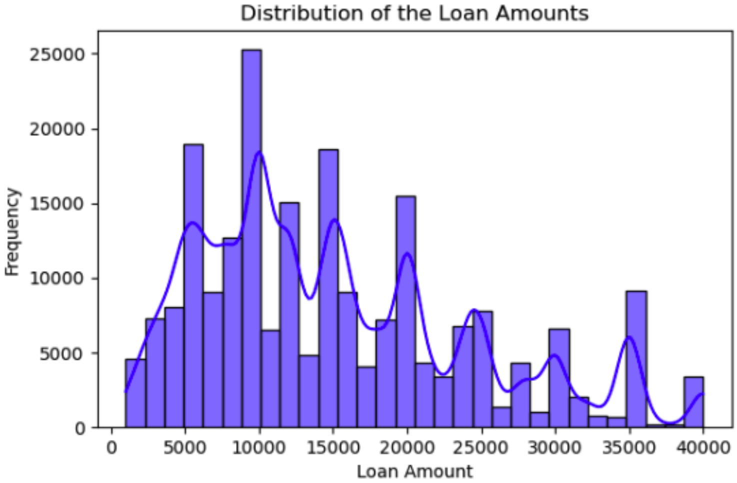
Section 3: Main Findings

Several steps were carried out on the dataset to preprocess it and to ensure its integrity and quality. The missing value problem was handled by deleting rows whose entries were null, hence giving a reduced dataset but with more consistency. . Once the dataset had been processed, the data was visualised as follows:

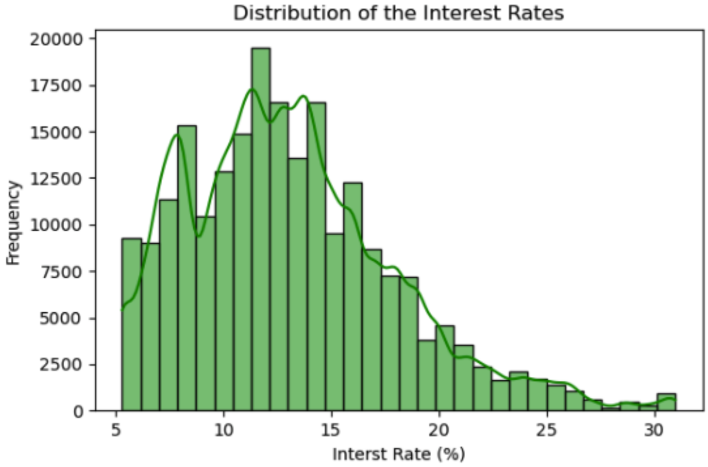
The heatmap of correlations provides a comprehensive view of relationships between predictors. Strong positive correlations were identified between loan amounts and instalments and between these variables and the total current balance. This aligns with the expectation that higher loan amounts require larger periodic repayments. Conversely, there was a negative correlation between annual income and interest rates, as well as interest rates and the total current balance. This reflects logical patterns, where borrowers with higher incomes are more likely to be approved for loans with lower interest rates, while lower balances may result in higher interest rates.



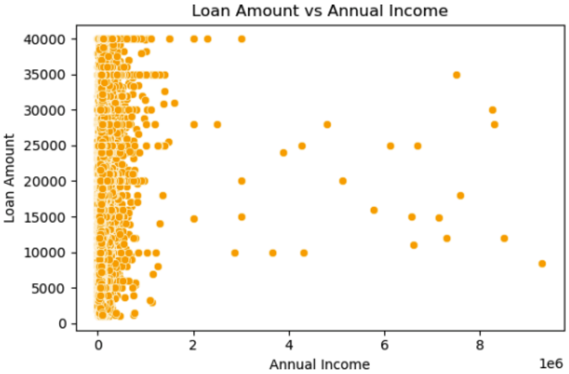
Additionally, figure 2 depicts the loan amounts and reveals normal distribution. Peaks are observed at multiples of 5000, indicating that borrowers tend to round loan amounts to standard increments. A noticeable spike occurs at 10000, representing the most common loan amount. In addition, although loan amounts go up to 40000, most loans are given out at much smaller increments.



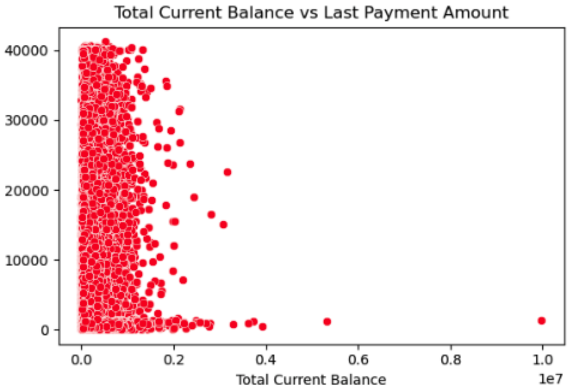
Similarly, the distribution of interest rates show a right-skewed trend, with most loans characterised by interest rates between 10% and 20%. This means that a preference for rates within this range, likely predetermined by the borrower's profile and the platform's risk assessment.



Further, the scatter plot between loan amounts and annual income shows significant clustering at lower income levels, indicating that borrowers with lower incomes dominate the dataset. Nevertheless, it spread to several income levels, hence showing that the platform can accommodate diversified profiles of borrowers.



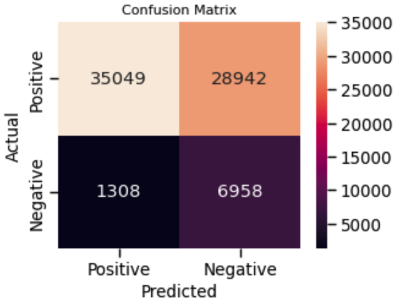
In another scatterplot examining the total current balance vs the last payment amount, a concentration near the origin reflects loans with smaller balances and payments. Outliers with disproportionally high balances or payments indicate atypical cases that might require closer scrutiny.



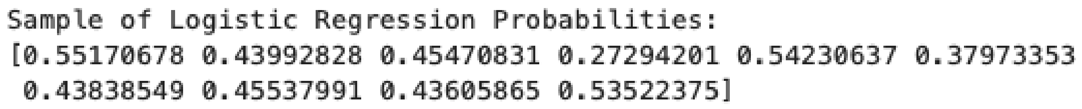
Afterwards, the categorical variable loan_status was transformed into a binary variable (dummy(loan_dummy)), which encoded Fully Paid as 0 and Charged Off as 1. Standardisation was applied using a Standardscaler to ensure that variables with different units of measurement were comparable. Finally, the dataset was split into training and testing subsets.

The first model implemented was Logistic Regression. This model estimates the probability of a loan being Charged off or Fully Paid using a linear combination of predictor variables transformed through a logistic function. To obtain accurate coefficient estimation without penalising larger values, the model was initialised with minimal regularisation (C = 1e6), and the maximum iterations were set to 100000 to ensure convergence.

The Logistic Regression Model delivered moderate predictive performance: Confusion matrix shows that the model predicted most of the Fully-Paid loans correctly. Nonetheless, many Charged-Off loans were misclassified as Fully-Paid, leading to a notable false-negative rate. Additionally, the model produced false-positives, where Fully-Paid loans were misclassified as Charged Off, reducing its precision.

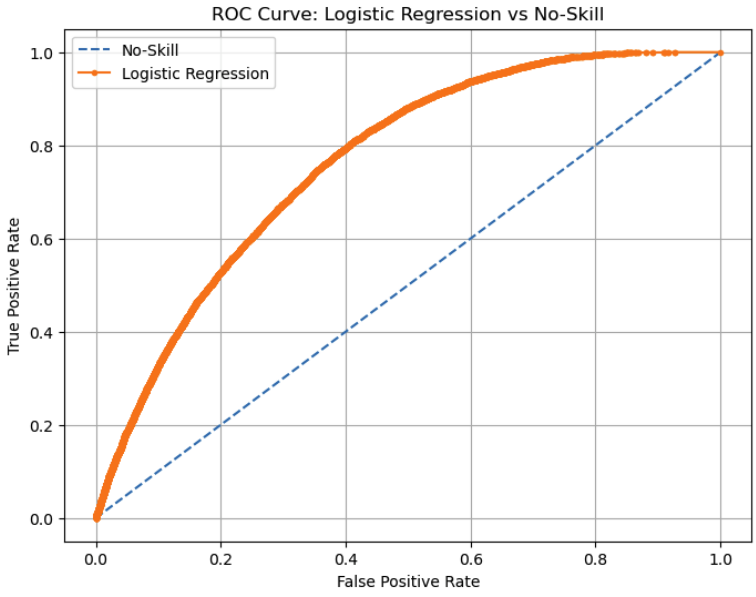


The probabilities predicted by the model showed that the model efficiently divided the data into the two tasks, although the division was not firm, as the likelihood was thrown into the intermediate range.

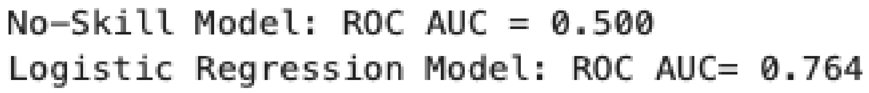


```
Sample of Logistic Regression Probabilities:
[0.55170678 0.43992828 0.45470831 0.27294201 0.54230637 0.37973353
 0.43838549 0.45537991 0.43605865 0.53522375]
```

The ROC curve sits above the diagonal no-skill line; therefore, the model was able to distinguish the two classes more accurately than no-skill. However, the curve's early concave part has such a shallow slope, indicating the difficulty of the model in avoiding the false-positive, which in turn could lose accuracy. Thus, being a Linear model, Logistic Regression can only be barely used with this data set. However, we can see that the relatively gradual slope in the curve's initial segment highlights the model's struggle with false-positives, which diminishes its accuracy. The linear assumption of Logistic Regression limits its ability to capture complex patterns in the dataset fully.

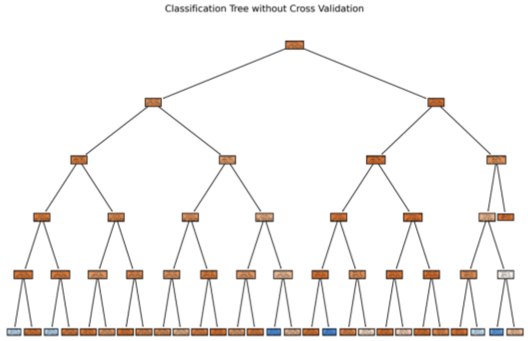


An area under the curve (AUC) value of 0.764 indicates that the model outperformed random guessing (no-skill) but struggled to capture complex(non-linear) relationships between features and the target variable.



```
No-Skill Model: ROC AUC = 0.500
Logistic Regression Model: ROC AUC= 0.764
```

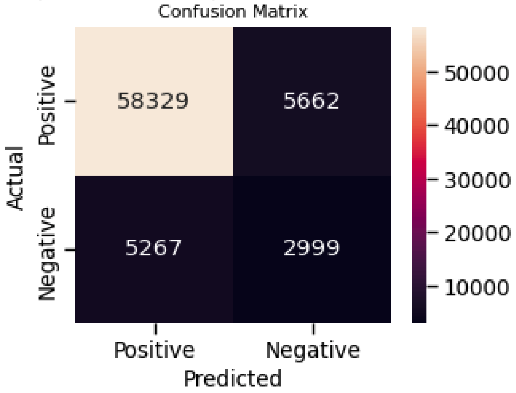
The second model implemented was the Classification Tree(Without Cross-Validation), which utilises a tree structure to classify loans based on a feature threshold.



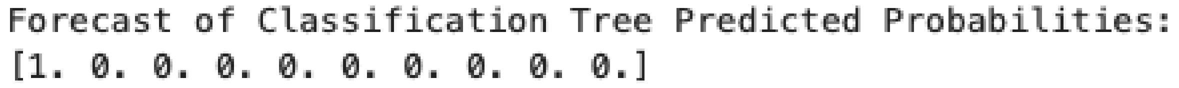
The classification tree was initialised with default parameters(unlimited depth and no constraints on splits), allowing it to fully grow based on the training data. This approach underlined the risks of overfitting in decision trees.

The model exhibited higher sensitivity to the training data as evidenced by its performance:

The confusion matrix revealed that the model performed well on the training data but struggled on the test data. The model misclassified several loans, with a notable number of false-positives (Fully Paid loans misclassified as Charged Off) and false-negatives(Charged Off loans misclassified as Fully Paid)

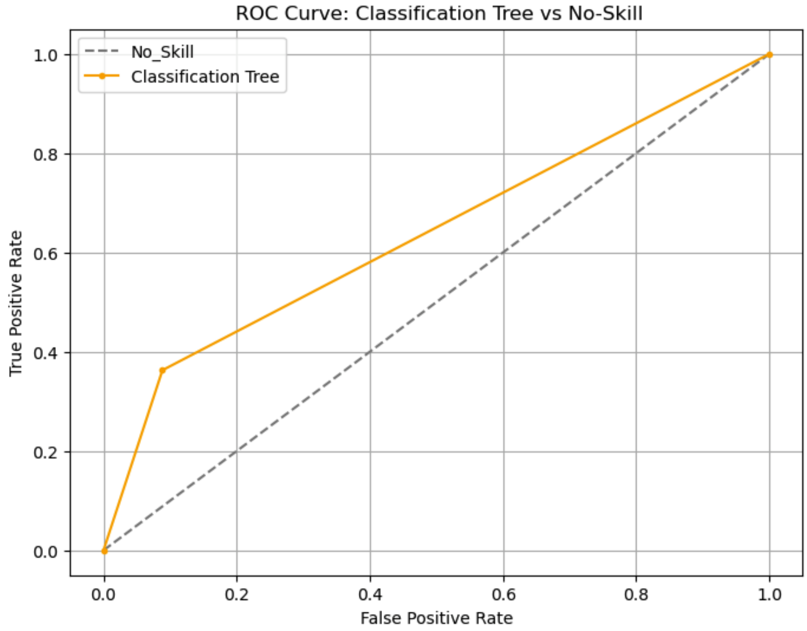


The predicted probabilities were extreme, ranging around 0(0%) or 1(100%), showing how this tree made its split in a deterministic fashion. Nevertheless, these sharp probabilities underscored the model's lack of generalisation on unseen data.

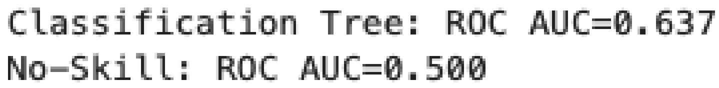


```
Forecast of Classification Tree Predicted Probabilities:
[1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
```

The ROC curve remains close to the no-skill line, indicating limited discriminatory power. This result is consistent with the tree's tendency to overfit the training data and fail to generalise to unseen data.



The model produced an AUC score of 0.637, which reflects the tree's tendency to overfit the training data, capturing noise, therefore highlighting the tree's inability to reliably separate two classes.



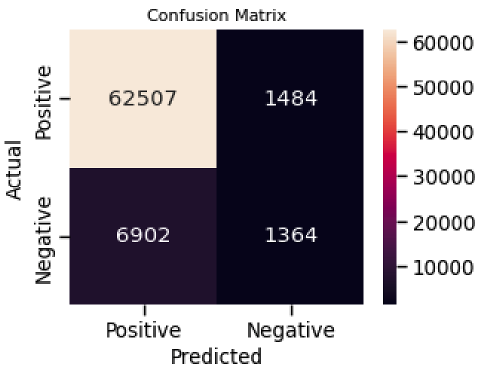
```
Classification Tree: ROC AUC=0.637
No-Skill: ROC AUC=0.500
```

The third model used was Cross-Validated Classification tree. This model uses hyperparameter tuning and validation approaches to improve the the usual classification tree results. Cross-validation allows the model to generalise better by evaluating its performance across multiple data subsets, preventing overfitting and ensuring robustness.

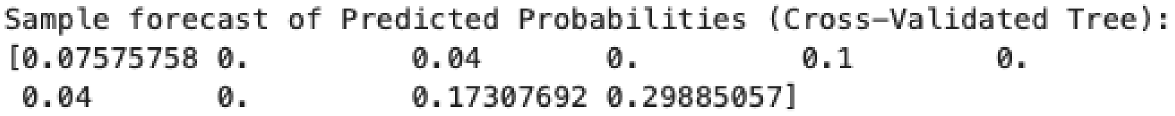
After, a grid search with cross-validation was employed to find the optimal hyperparameters for the tree, including maximum depth(restricted the depth of the tree to prevent overfitting), minimal samples split(defined the minimum number of samples required to split a node) and minimum samples leaf(ensured a minimum number of samples in each leaf node). Afterwards, five-fold cross-validation was carried out to evaluate model performance for different parameter combinations and ultimately select the most suitable configuration(max_depth=23, min_samples_leaf=50, min_samples_split=10)

Overall, the classification tree with Cross-Validation significantly improved upon the earlier tree by balancing flexibility and generalisation:

The confusion Matrix highlighted a reduction in misclassifications compared to the previous tree. False positives and false negatives were notably lower, illustrating a balance in predictions.

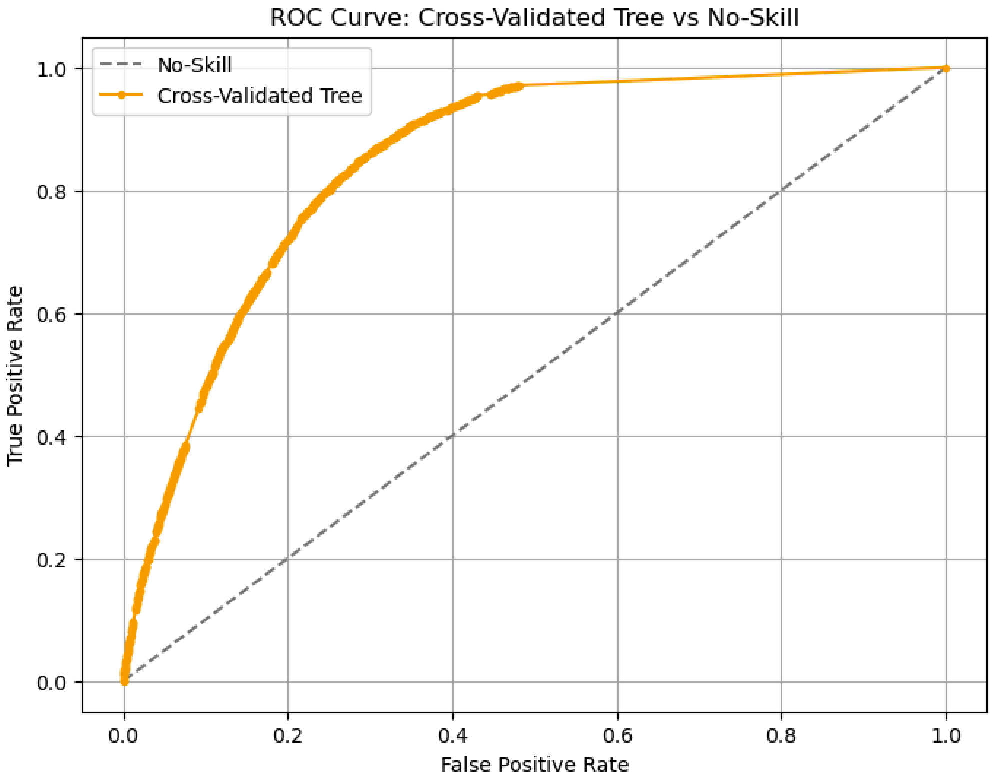


Predicted probabilities were smoother and less extreme than those predicted by the non-validated tree, reflecting the model's improved generalisation.

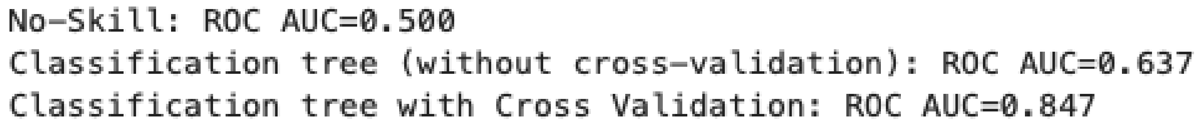


```
Sample forecast of Predicted Probabilities (Cross-Validated Tree):
[0.07575758 0.         0.04       0.         0.1        0.
 0.04       0.         0.17307692 0.29885057]
```

The ROC curve indicates a steep rise at the start, which illustrates a strong ability to correctly classify defaults while maintaining a low false-positive rate. The shape of this curve already reflects the benefits of the tuned hyperparameters and validation , which allowed the model to generalise better and kept sensitivity and specificity high.

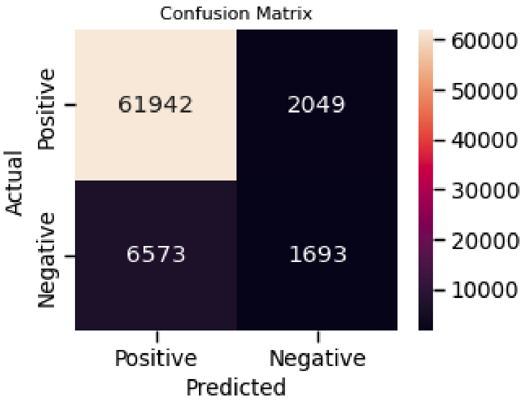


The AUC score of 0.847 demonstrates that the application of hyperparameter tuning, in conjunction with cross-validation, enabled the model to balance bias and variance, thereby indicating its proficiency in identifying intricate relationships and accurately classifying loans across various thresholds.

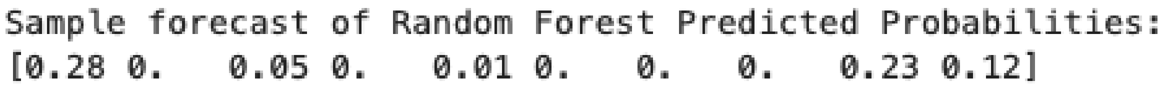


```
No-Skill: ROC AUC=0.500
Classification tree (without cross-validation): ROC AUC=0.637
Classification tree with Cross Validation: ROC AUC=0.847
```

The final model implemented was Random-Forest, an ensemble learning approach combining multiple decision trees to improve accuracy and robustness by aggregating predictions. As a result of introducing randomness in feature selection and data sampling, Random-Forest reduces overfitting and increases generalisation . The Random-Forest classifier was initialised with the random state for reproducibility. Additionally, the model was trained on bootstrapped subsets of the data, creating multiple decision trees, with each tree trained on random subsets of features. The Random-Forest algorithm as a whole demonstrated good performance, indicating that it was able to balance bias and variance well: The confusion matrix showed a dramatic reduction in false positives and false-negatives compared to the previous model. The random forest correctly classified a significant proportion of fully paid and charged-off loans.

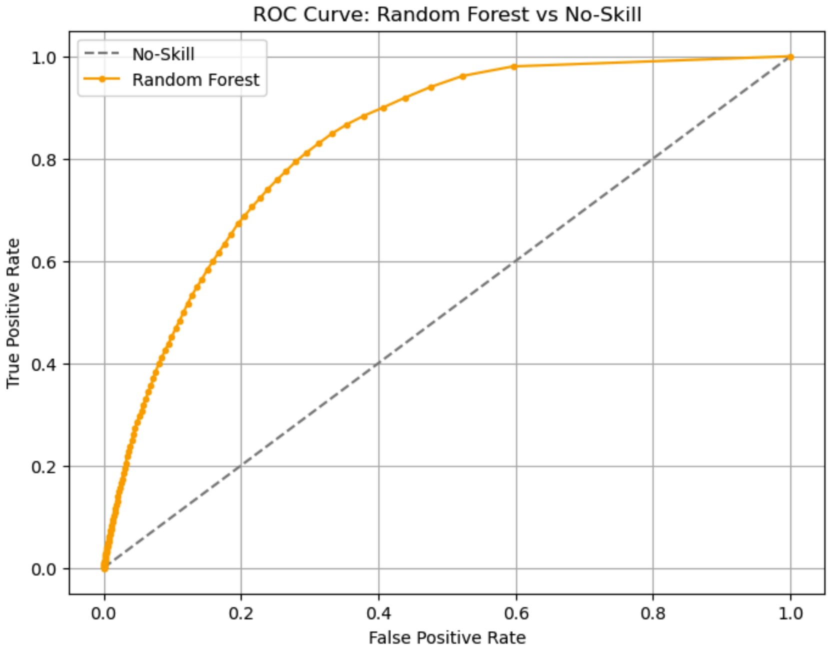


The probabilities generated from 10 samples using Random-Forest were smoother than the non-validated tree, capturing complex relationships in the data.

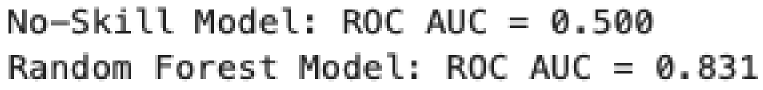


```
Sample forecast of Random Forest Predicted Probabilities:
[0.28 0.   0.05 0.   0.01 0.   0.   0.   0.23 0.12]
```

The steep slope at the beginning as could be seen demonstrates the model's capacity to achieve high sensitivity, especially concerning charged-off loans. Its ensemble method enabled Random-Forest to perform well on the test data and overcome the challenges.



The AUC score was 0.831, slightly lower than the cross-validated tree but significantly higher than logistic regression and the non-validated tree, reflecting its ability to handle complexities in the dataset effectively.

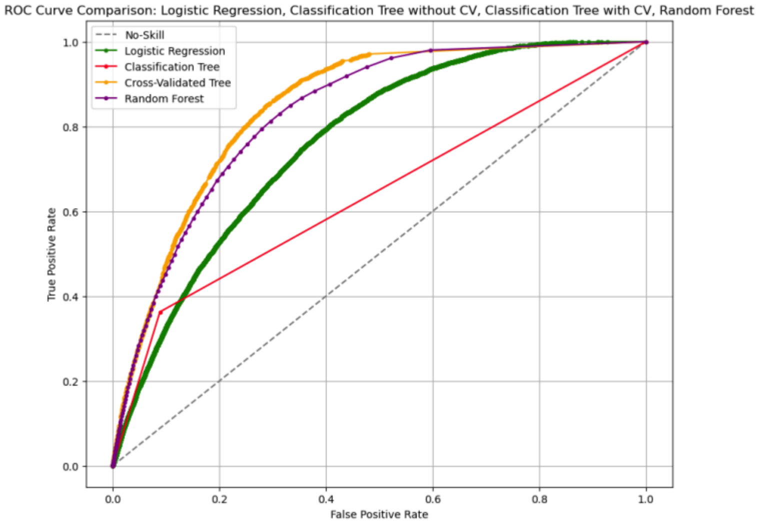


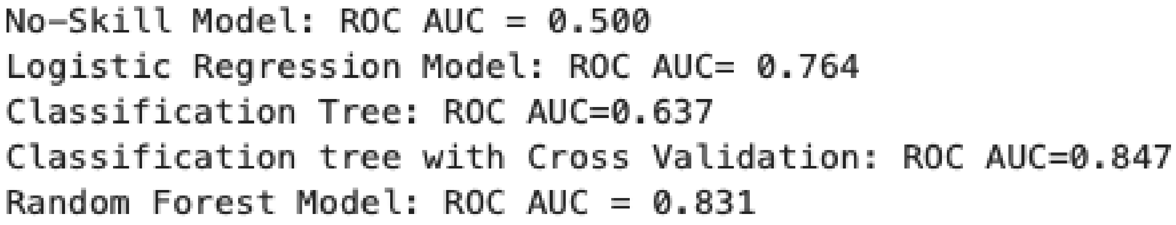
```
No-Skill Model: ROC AUC = 0.500
Random Forest Model: ROC AUC = 0.831
```

The analysis conducted on the models such as Logistic Regression, Classification Tree with cross-validation, classification tree without cross-validation and Random Forest already brought out some differences on what unconstrained assumptions, advantages, disadvantages, and fitting to the LendingClub dataset entails. Each model enabled the decision-making on the problem of forecasting loan defaults from different perspectives, for they were relatively interpretable, accurate, and generalisable. The Logistic Regression model, assumes linear relationships between the predictors and the one variable in the model. Nonetheless, this assumption limits its flexibility, as it may overlook non-linear interactions in the dataset. Despite these constraints, Logistic Regression attained a moderate Area Under the Curve (AUC) score of 0.764. While it is computationally efficient, its high false positive rate diminishes its practical applicability, thus rendering it more appropriate as a baseline model rather than a conclusive solution. Even though the classification tree is easy to comprehend or competitive ‘in the effect of capturing the nonlinear relationships’, it may suffer from overfitting. However, as we can see the training data has not received any cross-validation. An AUC value of 0.637 which means that the system does not reach the desired performance level, indicating the perils linked to the unpruned trees. Hence, the model's poor performance indicated a need for validation and regularisation.

As noted, the Classification Tree with Cross Validation optimised the bias and variance. Hyperparameter Search restricted tree expansion, reducing overfitting, resulting in the AUC score of 0.847, the best performance measure amongst all models tested. A tree-type cross-validation process revealed that the addition of cross-validation successfully achieves its primary aim of improving accuracy without compromising the definition of the model. However, the high price of this model cross-validation process made it more costly than the logistic regression and the non-validated tree case.

The Random Forest combined predictions from several decision trees with a 0.831 AUC score. Its ensemble properties enabled it to capture complex interactions and deal with multi dimensional data suitable for the LendingClub dataset. However, this compromise, in effect, affected interpretability since it became difficult to link decisions to rules. The regression with no adjustment nevertheless proved accurate and robust and was close to the tree cross-validated in terms of accuracy.





```
No-Skill Model: ROC AUC = 0.500
Logistic Regression Model: ROC AUC= 0.764
Classification Tree: ROC AUC=0.637
Classification tree with Cross Validation: ROC AUC=0.847
Random Forest Model: ROC AUC = 0.831
```

Ultimately, the Classification Tree with Cross-Validation is recommended as the optimal model for this dataset based on the results shown in the images above, demonstrating that this model achieved the best balance between performance (ROC and AUC) and interpretability, providing a practical solution for identifying high-risk loans.

Section 4: Conclusion

Finally, we highlight that there is a greater emphasis on automating the entire approach, and therefore, using ensemble techniques alongside K-fold cross-validation would help with model accuracy. They also highlight the importance of dataset properties and the objectives of the study in determining model selection, etc. Classification Tree with Cross-Validation is identified as the most suitable model for this context, providing a robust integration of accuracy and interpretability, thereby making it particularly advantageous for informed decision-making regarding loan default predictions.