**Introduction**

This report presents an analysis of the LendingClub dataset to understand the features and their relationships with loan defaults. The analysis includes data cleaning and visualization, as well as feature engineering to improve the model's predictive power. We have trained and evaluated various models and concluded that the random forest model performed the best. The report also discusses the most important features to consider when predicting loan defaults, identified using the permutation importance algorithm. The findings provide valuable guidance for feature selection and engineering efforts and highlight the usefulness of different techniques and algorithms in predictive modeling.

**Data Exploration**

I have explored the LendingClub dataset to understand the features and their relationships with loan defaults. The dataset consists of 226067 observations and 31 variables, which includes both numeric and categorical variables. After exploring the data, I have selected the following variables for feature engineering:

Notice the variables we want to use for forecasting are:

* Loan amount (**loan\_amnt**)
* Interest rate (**int\_rate**)
* Annual income (**annual\_inc**)
* Total mortgage payment (**total\_pymnt**)
* Mortgage installment (**installment**)
* Total paid in interest rates (**total\_rec\_int**)
* Last payment amount (**last\_pymnt\_amnt**)

Table

Description automatically generated

1. loan\_amnt: The amount of money that a borrower has been approved to borrow from a lender.
2. int\_rate: The interest rate on the loan, expressed as a percentage.
3. annual\_inc: The borrower's annual income, as reported by the borrower at the time of the loan application.
4. total\_pymnt: The total amount that the borrower has paid on the loan, including both principal and interest.
5. installment: The monthly payment that the borrower is required to make on the loan, including both principal and interest.
6. total\_rec\_int: The total amount of interest that the borrower has paid on the loan.
7. last\_pymnt\_amnt: The amount of the borrower's most recent payment on the loan, including both principal and interest.

**Cleaning of data**

To ensure accurate predictions of mortgages that have been charged off, it is crucial to clean the data appropriately. In this scenario, we follow a two-step process to achieve this goal. Firstly, we remove variables that have more than ten missing observations from the dataset to eliminate potentially irrelevant or biased variables. Secondly, we drop only the missing observations for variables with less than ten NaN to retain as much data as possible while ensuring that only reliable and accurate data is used for analysis.

This approach is a common strategy to handle missing data and helps to ensure the accuracy and reliability of the results obtained from data analysis. Overall, data cleaning is an essential process that should not be overlooked in any data analysis project.

**Data Visualisation**

The plots show that the loan amount tends to increase with higher credit grades, while the interest rate tends to be higher for lower credit grades, which suggests that borrowers with lower credit grades may be viewed as higher risk and charged a higher interest rate. These observations are consistent with what we would expect in a lending scenario, where lenders are more cautious about lending to riskier borrowers and may charge higher interest rates to compensate for the increased risk.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

**Feature Engineering**

In feature engineering, I have created new features based on the existing ones to improve the predictive power of the model. For instance, I have created a new feature named “Loan\_dummy” by “Loan\_status” . Loan\_dummy is a binary variable that is created to represent the loan status in a binary form. It takes a value of 1 if the loan status is charged off and 0 otherwise.

Loan\_dummy is an important variable because many machine learning algorithms require the target variable to be a binary variable. By creating loan\_dummy, we have converted the loan status, which is a categorical variable, into a binary variable that we can use in our machine learning models. It helps in simplifying the analysis and improving the performance of the machine learning models. Therefore, the loan\_dummy variable plays a vital role in the analysis and prediction of loan defaults.

A picture containing text, receipt, screenshot

Description automatically generatedTable

Description automatically generated

Model Selection:

After feature engineering, I have trained and evaluated different models on the dataset to find the best predictive model. I have evaluated the following models:

1. Logistic Regression

2. Classification tree

3. Classification tree with cross validation

4. Random Forest

I have evaluated these models using 5-fold cross-validation and measured their performance using Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

1.Logistic Regression

Chart, line chart

Description automatically generated

Graph to compare the logistic regression with the "no-skill" prediction based on the ROC curve

The performance of a logistic regression model can be visualized through a ROC curve that plots the true positive rate against the false positive rate. The AUC of the ROC curve provides an overall measure of the model's performance, with 0.5 indicating random guessing and 1.0 indicating perfect classification. In this case, the AUC of 0.75 suggests that the model has some predictive ability, although there is still room for improvement. The proximity of the ROC curve to the no-skill prediction line indicates that the model is not significantly better than random guessing. To enhance the model's performance, additional feature engineering or the use of more advanced machine learning techniques may be necessary.

**2. Classification tree**

Chart, line chart

Description automatically generated

Graph to compare the classification tree against the logistic regression and the "no-skill" prediction based on the ROC curve.

The AUC values for the logistic regression and classification tree models are 0.75 and 0.79, respectively, indicating that the classification tree model performs slightly better in distinguishing between positive and negative classes. A moderate discriminatory power is suggested by an AUC value of 0.75 for the logistic regression model, while an AUC value of 0.79 for the classification tree model indicates good discriminatory power.

Therefore, the preference for one model over the other will depend on the particular problem and requirements, and a thorough evaluation and comparison of different models should be conducted before a final decision is made.

**3. Classification tree with cross validation**

Chart, line chart

Description automatically generated

A graph comparing classification tree with and without cross validation against the logistic regression and the "no-skill" prediction on the ROC curve.

Among the three models, the classification tree with cross-validation exhibited the best performance, with an AUC value of 0.81. This indicates that this model has the greatest ability to distinguish between the positive and negative classes

The logistic regression model had a lower AUC value of 0.74 compared to both classification tree models, suggesting that it may not be as effective in separating the two classes in the dataset.

The classification tree model without cross-validation had an AUC value of 0.79, which is slightly inferior to the cross-validated classification tree model. The absence of cross-validation in the former model could have contributed to overfitting since it was trained and evaluated on the same data. Conversely, the cross-validated model was assessed on unseen data, which may have helped to reduce overfitting.

**4. Random Forest**

Chart, line chart

Description automatically generated

A graph comparing the random forest against the classification tree, with and without cross validation, the logistic regression and the "no-skill" prediction on the ROC curve.

The best performing model among the four is the random forest model, with an AUC value of 0.93. In comparison, the classification tree model with cross-validation achieved a higher AUC value of 0.81 than the non-cross-validated classification tree model, which obtained an AUC of 0.79. On the other hand, the logistic regression model had the lowest AUC value of all four models, with an AUC of 0.74.

**Evaluation**

The results of the evaluation are shown in the following table:

| Model | AUC-ROC |
| --- | --- |
| Logistic Regression | 0.74 |
| Classification Tree | 0.79 |
| Classification Tree with CV | 0.81 |
| Random Forest | 0.93 |

**Visualizing the Model’s Predictive Power**

To visualize the model’s predictive power, I have plotted the ROC curves of all the models on hand. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds. The following figure shows the ROC curves of all the models:

Chart, line chart

Description automatically generated

Based on the results, the random forest model demonstrated the best performance among all the models tested, achieving an AUC score of 0.93. The classification tree model with cross-validation outperformed the classification tree model without cross-validation, with an AUC score of 0.81 compared to 0.79. In contrast, the logistic regression model had the lowest AUC score of 0.74.

Precision, recall, f1 score and accuracy for "no-skill" prediction

Graphical user interface, text, application

Description automatically generated

Precision, recall, f1 score and accuracy for the classification tree

A picture containing graphical user interface

Description automatically generated

Precision, recall, f1 score and accuracy for the decision tree classifier

A picture containing calendar

Description automatically generated

Precision, recall, f1 score and accuracy for the random forest

Table

Description automatically generated

**CONCLUSION**

The permutation importance algorithm results provide insights into the most important features for predicting loan default, with 'int\_rate', 'total\_pymnt', and installment being the most important features. This suggests that loan interest rates, the total payment received, and the size of the installment payments may be strong predictors of loan default.

Based on the results, it seems that the random forest model is the most effective in predicting loan default, with an impressive AUC of 0.93. This suggests that the model is able to distinguish between the positive and negative classes with high accuracy.This is because random forest has better out-of-sample predictive power than classification tree and logistic regression. This is because it reduces overfitting and variance by constructing multiple decision trees and combining their results. This helps to generalize the model to new data and improves its accuracy. Additionally random forest can handle both categorical and continuous data, which is an advantage in the case of the Lending Club dataset. The dataset contains both types of data, and random forest can handle them without the need for feature engineering.

Overall, these findings may be useful for improving the accuracy of loan default predictions and guiding decision-making for lenders. It is important to note, however, that further analysis and consideration of other factors may be necessary for making informed lending decisions.