Can Machine Learning Predict Loan Approvals?

Jude Wallace

1 Introduction

In the complex landscape of financial transactions, loan approval stands as a crucial gateway influencing economic activities on individual and societal scales. Loans, in their various forms, are indispensable for accessing capital, managing financial goals, and fostering economic growth. The need for efficient loan approval systems arises from the intrinsic nature of modern economies, where individuals, businesses, and governments often require financial assistance. Loans empower individuals to make significant life investments, while businesses leverage them for growth and innovation. Governments rely on loans for infrastructure projects. In contemporary financial systems, virtually all money is debt-based, underscoring the critical role loans play in sustaining economies worldwide. The ability to predict loan approvals is strategic, ensuring efficient resource allocation and minimising financial risks for lenders, contributing to overall financial stability[1].

2 Dataset

The dataset for our analysis has been been collected from Kaggle[3]. It is a collection of financial records and associated information used to determine the eligibility of individuals or organisations for obtaining loans from a lending institution. It includes various features such as CIBIL score, income, employment status, loan term, loan amount, assets value, and loan status. With over 4000 loan applications and 11 features, the dataset provides enough data for a comprehensive analysis using machine learning. With the dataset being labelled it tends itself well to a classification problem.

3 Random Forest Classifier

Random Forest is a supervised machine learning algorithm, where there is a labelled target variable. It is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classes for classification tasks or the mean prediction for regression tasks. In the training phase, subsets of the training data are created through bootstrapped sampling, and each tree is trained on a different subset with randomly selected features at each node. This process introduces diversity among the trees and helps prevent overfitting. During prediction, each tree makes a prediction, and the final prediction is determined by the majority vote (classification) or the average (regression) of all the individual tree predictions[2].

In the context of loan approval, where the goal is to categorise applications into approved or denied classes, the ensemble of decision trees generated by Random Forest proves advantageous. By building multiple trees during training, each on a different subset of the data with random feature selection at each node, Random Forest introduces diversity crucial for capturing the multifaceted nature of loan approval factors. The inherent mechanism of bootstrapped sampling mitigates overfitting concerns, ensuring that the model generalises well to new loan applications. During prediction, the amalgamation of individual tree predictions through majority voting yields a robust final decision, providing a reliable foundation for the nuanced and critical task of loan approval.

4 Experimental Design and Results

After analysing the dataset, it was clear it contains several multi-correlated features maintaining and correctly modelling these is a key task, furthering our reasoning to use a Random Forest classifier. The main preprocessing for the dataset was converting non-numerical data to numerical, such information requiring preprocessing was loan status, self-employment and education. Using a label encoder we converted the relevant data into a binary representation, allowing them to be used in our model. Our dataset contains 2656 approved applications and 1613 rejected applications, to reduce any bias when training our model, we performed stratified sampling. This allows the model to be trained on an equal amount of approvals and rejections reducing the potential bias of the model.

After preprocessing the data we trained our first model. Utilising all 11 features of the dataset our model has an accuracy of 98%, showing significant results. The classification report indicated a precision for correctly classifying approved loans at 99%, similarly 98% for rejected applications. Metrics for the recall and f1-score for both classes are equally as high as precision. The macro and weighted averages, both at 98%, affirm the overall robustness of the model across the different evaluation metrics. With such a high accuracy it is important we check for overfitting, to do this we used K-folds cross-validation. Using 5 folds the model consistently performed well on the training folds and on the validation folds with an average accuracy of 97%, indicating overfitting may not be occurring. Extracting the models feature importance the CIBIL score had an importance of 0.82 followed by loan term with 0.06 and with self employment and education both having negligible importance. The CIBIL score being such a dominant importance is inline with the nature of the problem, CIBIL score being a metric for the reliability of the candidate paying back the loan based on previous loans, it is not a surprise it has such a high importance in our model. In Figure 1 the ROC curve shows a near perfect classification with an AUC of 0.982, indicating strong discriminatory power in distinguishing between positive and negative instances. Looking at the confusion matrix for the model we discover it classified 9 True Negatives and 6 False Positives. To further the accuracy of our model, with the use of GridSearchCV, we explored the possible parameter combinations that yielded in the best model. Using the best estimators found by the grid search, training and testing a model with these parameters also led to a high accuracy of 0.98%.

To reduce the dataset we explored Principle Concept Analysis (PCA). Initially we ran PCA on our dataset with the default configuration. Doing so PCA ended up not reducing the feature size of the dataset. Plotting the Cumulative Explained Variance, on a scree plot, it was clear after 4 principle concepts the cumulative variance tended towards 100 for the remaindering principle concepts demonstrating no major information gain of including more than 4. Another trail of PCA was conducted this time setting PCA to retain 95% of the data's information, resulting in the dataset being reduced to 4 features, which is inline with the findings based off the scree plot. For these 4 principle components, the first principle component had an explained variance of 0.79 suggesting most of the information within the dataset can be reduced into one principle concept.

With the PCA reduced dataset we trained a Random Forest classifier, which led to a sub optimal accuracy of 58% prompting reflection on the model's sensitivity to the selected reduced feature set. With an AUC value of 0.48 it suggests the model does not have strong discriminatory power, and the performance is similar to a random guess. A potential reason for the model having such a poor accuracy could be because when PCA reduced the dataset it lost crucial non-linear relationships that was present within the data. While PCA can handle linear relationships, it is not so good at encapsulating non-linear relationships when reducing the dataset.

5 Conclusion

From our experimental approach we can conclude the best approach for loan approval was to keep all the features of the dataset. Using the Random Forest Classifier we were able to achieve a classification average accuracy of 97%. K-folds cross-validation demonstrated that the model exhibited robust generalisation and was not prone to overfitting, even with a high accuracy. Attempts to enhance model efficiency by reducing the dataset through PCA yielded sub-optimal results. The reduced dataset, comprising of only four features, led to a weakened classification accuracy of 58%. A limitation of our study to consider is that each lender may have different specific factors they favor. For instance, some lenders may restrict candidates to a loan amount equivalent to five times their income. Future research that takes these lender-specific criteria into account could yield different results. In summary our model is able to correctly generalise the dataset and accurately predict the classification of new loan applications, with a favoured importance of CIBIL score.

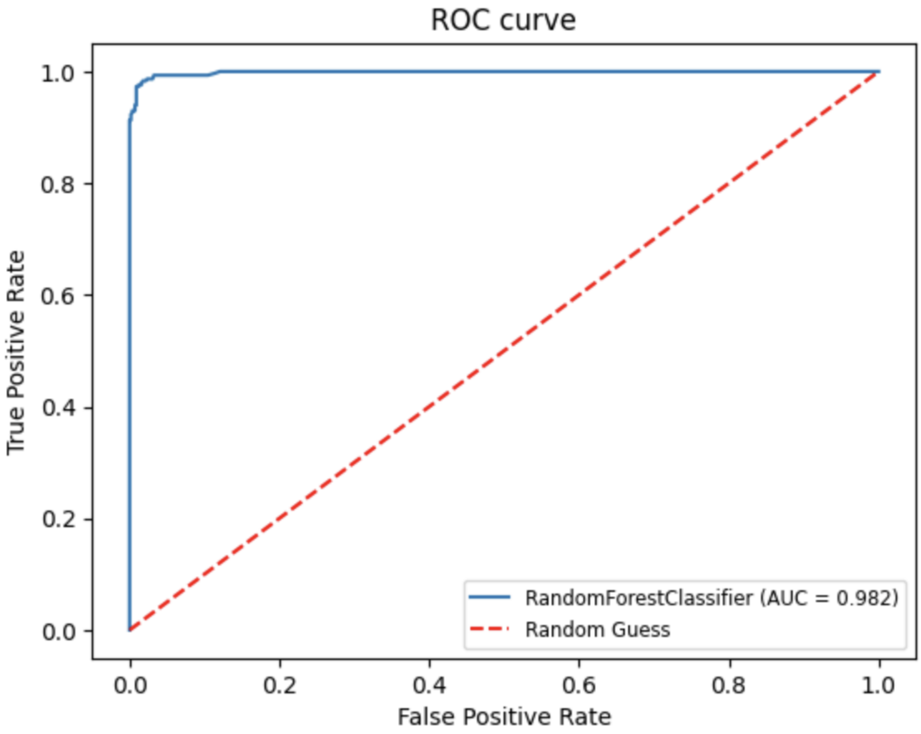

Figure 1: The plot displays the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (Specificity) across various classification thresholds.

References

[1] Julia Kagan. What Is a Loan, How Does It Work, Types, and Tips on Getting One. https://www.investopedia.com/terms/l/loan.asp, 2023. Accessed: November 24, 2023.

[2] IBM. Random Forest. https://www.ibm.com/topics/random-forest, 2023. Accessed: December . 1, 2023.

[3] Archit Sharma. Loan approval prediction dataset. https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset, 2023. Accessed: Novemeber 25, 2023.