Roxanne Argueta

Capstone Project: Loan Default Prediction

Executive Summary

After having looked at multiple models to address a bank's concern for lending to the appropriate parties, it was found that the Tuned Random Forest was most appropriate for the problem at hand based on its performance. In addition, the use of hyperparameter tuning helped better improve the model. Regardless, there are other factors that should be considered before finalizing the model, such as computational time and resources, interpretability, and ethical considerations.

Problem Summary

Banks often have strict guidelines in determining which applications will be approved for loans versus those who will not. Proper review of a loan application is crucial to a bank's financials as the bank incurs a loss whenever an approved applicant defaults on a loan. As such, it is important to review several factors of the applicant, such as their current employment status and job history, debt-to-income ratio, any derogatory remarks, the amount of the loan looking to be approved, and delinquent credit lines among other important factors. The current bank looks to utilize the Equal Credit Opportunity Act's guidelines that will follow predictive modeling techniques that will help the bank properly justify any adverse actions. As such, the goal is to simplify the decision-making process for credit applications.

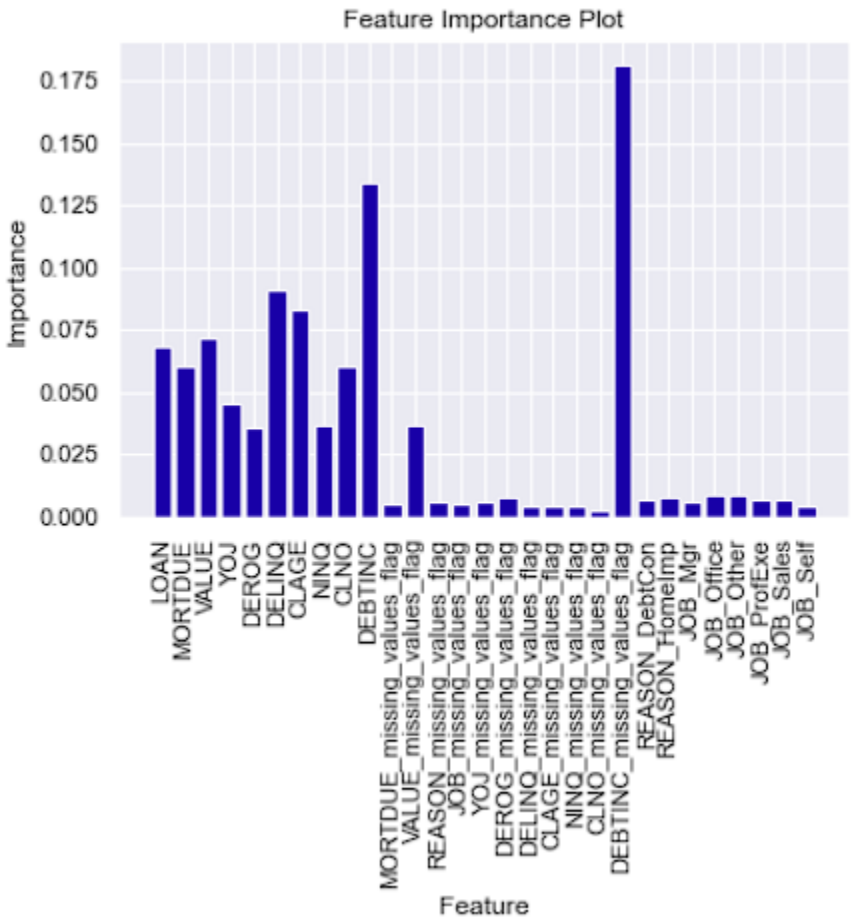
Final Solution Design

The performance metrics of the models provide a summary of their ability to correctly classify the loan applicants as either "approved" or "rejected". In this case, the tuned Random Forest model has the best combination of high-test Accuracy (0.911074), Test Recall (0.697479), and Test Precision (0.83). This means that the model is able to correctly identify a high proportion of loan applicants who are eligible for a loan while also minimizing the number of false positive cases (i.e. applicants who are incorrectly approved).

Precision and recall are important metrics that indicate the trade-off between the ability to detect positive cases (recall) and the ability to minimize false positive cases (precision). A high recall value means that the model is able to identify most of the positive cases, while a high precision value means that the model is able to minimize the number of false positive cases. This means that the bank will not miss any potential loan applications that would be approved, thereby increasing the chances of loan approval and profitability, with regard to recall. With high precision, this means that the bank will minimize the number of loan applications that are rejected but should have been approved, thereby reducing the risk of losing potential business.

In addition to the above, the Feature Importance was observed via bar chart (see **Figure 1**). The Feature (or Gini) Importance was also of use in helping determine that the Tuned Random Forest model was of benefit this measures the total reduction in impurity that a feature provides across all trees in a Random Forest. Essentially, it calculates the average decrease in impurity for each feature in the dataset. The higher the value of the Gini importance for a feature, the more important it is in terms of predicting the target variable. Gini importance can be useful in identifying which features are most important in determining the target variable and can be used to determine which features to include or exclude in the model, reducing overfitting and increasing the interpretability of the model. While Gini Importance was applied to the Tuned Random Forest, it could have also been applied to other tree algorithms. The benefits of overfitting reduction and increased interpretability could help the bank fine-tune their model, as appropriate, to better observe the variables considered when looking at loan applicants.

Figure 1: Feature (Gini) Importance



While the Tuned Random Forest model appears to perform the best based on the performance metrics, it's important to consider other factors before adopting it as the final solution. For example, the computational time and resources required to run the model may be a concern, as Random Forest models can be computationally expensive. Additionally, the interpretability of the model may be important, as some stakeholders may need to understand the factors that influence the loan approval decision. Finally, ethical considerations such as fairness and bias in the model may need to be addressed to ensure that the model is making unbiased decisions as models are still subject to bias.

Overall, based on the performance metrics observed in the data, the proposed solution to the problem is to adopt the tuned Random Forest model for making lending decisions. Based on the model's high test accuracy, test recall, and test precision in comparison to other models, the model has the ability to correctly identify positive cases while minimizing false positives. The use of the Random Forest and its variants have demonstrated better

performance compared to other models, such as Logistic Regression and the Decision Tree, as well as other univariate and bivariate tests. Despite the benefits of the Random Forest, other factors should also be considered as the model is still subject to bias. As such, the bank should consider computational time, resources, interpretability, and ethical considerations. The key next step would be to conduct additional experiments to consider the aforementioned factors in order to confirm the validity of the proposed solution design.

Recommendations for Implementation

To address the bank's concern regarding loan approval and to minimize risk in loan default, the bank is recommended to implement the Random Forest model as the best solution for loan prediction. For actual implementation, it is recommended that the bank conduct additional experimental scenarios prior to fully establishing the model into the bank's official system, as this will help the bank address any issues such as ethical concerns, interpretability, and potential time and resource usage. Furthermore, this additional experimentation will allow for the bank to review additional hyperparameters with additional features, and it will allow for the bank to collect additional data to help fine-tune the model. Once additional experimentation is complete, the bank can implement the model into its existing loan approval process, ensuring that the model's output is integrated with other relevant factors for its clients, such as credit history, income, and employment status to name a few. Finally, the bank should conduct additional and ongoing monitoring of the model's performance to ensure that it continues to accurately predict loan default rates and to continue to minimize false positives.

Although the model works well in comparison to other reviewed models, additional analysis can be done to further improve the solution, such as collecting additional data to further improve the model's performance, conducting further experiments to optimize hyperparameters and improve the model's accuracy, and incorporating other relevant factors that were previously tested, as well as new factors, such as credit history, income, and employment status to improve accuracy.

Considering that the above implementation process will take monetary resources, the use and support from stakeholders is crucial. As such, stakeholders should be concerned with the bank's data privacy and security being maintained throughout the use and implementation of the model. Resources should be provided to the bank, when

possible, to allow for further experimentation, which will allow for further tuning and monitoring of the suggested model. Finally, stakeholders should be concerned that the solution is integrated with other loan approval processes and that other stakeholders are trained on how to use the model effectively, where applicable.

Despite the bank being recommended to continue to tune the model and to use resources for ongoing experimentation and monitoring, there are also risks that should heavily be considered when applying the model. These risks and challenges can include, but are not limited to, ensuring that the model's predictions are consistent with regulatory requirements, maintaining data privacy and security due to the type of information being collected and used in the model, and ensuring that the model's output is properly integrated with other relevant factors in the loan approval process.

All in all, the expected benefit of implementing the solution is an increase in the accuracy of loan predictions, which could result in more appropriate lending decisions and a reduction in loan default rates. There may be cost savings associated with reduced manual labor in the loan approval process as the model may be able to assist the bank in making more decisions, autonomously.

References

Bertani, Alessandro, et al. "How to Describe Bivariate Data." *Journal of Thoracic Disease*, vol. 10, no. 2, Feb. 2018, pp. 1133–37. *PubMed Central*, https://doi.org/10.21037/jtd.2018.01.134.

"Things for a Bank to Consider Before Lending Money to a Business." *Small Business - Chron.Com*, https://smallbusiness.chron.com/things-bank-consider-before-lending-money-business-57341.html. Accessed 26 Jan. 2023.

What Banks Look for When Reviewing a Loan Application. https://www.wolterskluwer.com/en/expert-insights/what-banks-look-for-when-reviewing-a-loan-application. Accessed 26 Jan. 2023.