Taylor Cina

Data Analytics Capstone

January 14, 2025

Random Forest Classification Analysis on Loan Approval Dataset Summary Report

The research question for this analysis was if a random forest classification model could be used to generate a predictive model capable of correctly predicting loan approval statuses with an 80% or higher accuracy. For this question, two hypotheses were considered.

***Null Hypothesis H*0:** A random forest classification model cannot be constructed from the research dataset with an accuracy of > 80%.

***Alternative Hypothesis H*a:** A random forest classification model can be constructed from the research dataset with an accuracy of > 80%.

**Data Analysis Summary**

The dataset used for this analysis is a synthetic dataset derived from real loan approval data without any identifying user information created specifically for the purpose of model training. The dataset is hosted [publicly on Kaggle (Lo, 2024)](https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data). The cursory data exploration found no duplicate, missing, or infinite values. However, some preprocessing was required to assist in model training. Ordinal string values were converted to integers. The data was then split 70/30 for training and testing data respectively; the data was stratified on the target variable “loan\_status”.

The underlying model is a random forest classifier. Random forest classification is especially well suited to this analysis as it has been used by other researchers for a similar problem with a high degree of accuracy (Saini, 2023). The model-building process consisted of two major phases. The first phase was the initial build of the model and analysis of its performance and most important features. The second phase was dimensionality reduction using the first model’s list of feature importances and testing how fewer features impacted the performance of the model. The hyperparameters for all models were tuned using randomized search cross-validation and optimized for F1 score. Scikit-learn’s *RandomForestClassifier* package was used to generate and train the model. A classification report and confusion matrix for each model was generated using the relevant Scikit-learn packages.

**Analysis Conclusions**

According to the results of the analysis, it is conclusive that random forest classification is capable of producing a model with an accuracy greater than 80%. Even the initial unoptimized model had an accuracy of greater than 80%, though this metric is not as impressive when one realizes that only 22% of the dataset contained approved loans. The initial model has a terrible recall of 56%, which is barely better than random guessing. However, it did guide the creation of the subsequent models. The second round of models performed much better, with the 5-feature model having the most desirable traits for this specific business case. This model had the highest F1 score of 78%. Although it does not have the highest recall at 71%, it doesn’t score much lower than the model that does. An F1 score of 78% is significantly more indicative of a functional model than accuracy is in this case as the dataset is skewed heavily toward unapproved loans.

**Analysis Limitations**

As with any synthetic dataset, there is always the possibility that the synthetic dataset does not perfectly accurately represent a real-world set with identifying user information.

With regards to the model, random forests are computationally expensive, and each build of a new model took two or more minutes even with a dataset of medium size on a higher end machine. Additionally, while it is possible that the best hyperparameters could be missed using this algorithm, the given hyperparameters still perform well. This analysis found that a 5-feature model had the most desirable traits, what those 5 features are could be different than the ones used in the second round of model testing as they were based on an initial model that did not perform as well.

**Proposed Actions**

There are other analyses that are worth performing on this dataset besides the one shown here. Separating the data based on age groups, education level, or loan purpose for example could reveal patterns that aren’t present within a full range analysis. Another example is the fact that loanee credit score appeared to play a very minimal role in predicting loan defaults according to the initial model’s feature importance list, which doesn’t initially appear to make intuitive sense. It would also be worthwhile to analyze a real-world dataset with similar fields and see if the findings of this analysis correspond with the findings of an analysis conducted on non-synthetic data. Finally, a more thorough analysis of correlated features could be performed to create an initial model that may have different feature importances and perform better.

**Expected Outcomes of this Analysis**

The most obvious benefit to be sought from this analysis is increased profit due to better judgement of loan approval. However, understanding customers based on a few key traits is always beneficial. This information can be used for marketing purposes, such as offering deals to clients that are not expected to default in order to gain preference over other loan services.

**Sources**

[Lo, T. (2024, October 29). *Loan approval classification dataset*. Kaggle. https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data](https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data)

[P. S. Saini, A. Bhatnagar and L. Rani, "Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms," *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2023](https://ieeexplore.ieee.org/document/10182799/metrics#metrics)