Final Submission – Loan Default Prediction

MIT-PE ADSP Sep23

12/13/23

Andrey Drozdov

Executive Summary

In the world where banks need to produce interest income to continue to serve their customers, issuance of quality loans is of utmost importance. The main goal of this project was to create a credit scoring model for the home equity department of the bank, specifically to improve the decision-making process and reduce loan default rate that is currently at 20%. With the advent of data science and machine learning models, it is imperative to build a predictive model to leverage data from an existing loan application process.

The objective is to construct a classification model that could accurately predict clients who are more likely to default on their loans, while also providing the bank with recommendations on key features to consider during the loan approval process. Additionally, it was crucial to ensure the model's interpretability and prevent biases that percolated from the human-centered approval process in the past.

After a thorough analysis of the available data and extensive experimentation with various machine learning models (see Appendix 1), it has been determined that the Random Forest Classifier with Hyperparameter Tuning is the most suitable model in achieving the project's objectives. This model has exhibited high accuracy and precision on both the training and test datasets, achieving accuracy of 89% on the test data set with 87% precision which means the model can predict defaults accurately 89% of the time with 87% precision rate. Furthermore, the model offers a relatively high degree of interpretability compared to more complex models like neural networks, aligning well with the project requirements. The most significant features identified by the model for predicting the target variable include applicant's missing debt-to-income ratio data, debt-to-income ratio (DEBTINC), age of the oldest credit line in months (CLAGE), the number of delinquent credit lines (DELINQ), current value of the property (VALUE), and the amount of loan approved (LOAN) (see Appendix 2).

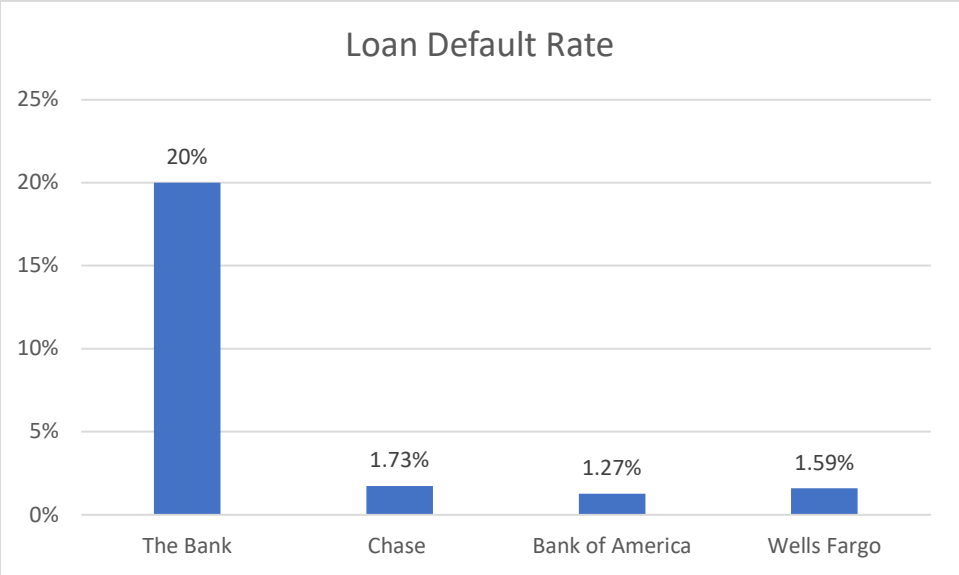
The Random Forest Classifier tuned with hyperparameter model stands out as the most robust and accurate solution for the given problem. By leveraging multiple decision trees and selecting optimal features for predictions, it demonstrates superior generalization on unseen data. The feature importance analysis also offers insights into the critical factors influencing loan approval.

Consequently, the adoption of the Random Forest Classifier with hyperparameter tuning is recommended for constructing a dependable and interpretable model for credit scoring in this specific context. It is believed that a combination of data collection improvement strategy along with Random Forest Classifier with Hyperparameter tuning model implementation will reduce default rate for the bank and significantly reduce regulatory risk of biased decisions as the model will continuously make approval and rejection decision based solely on data and not from preconceived biases. It is also safe to assume that once debt-to-income ratio data is complete (not missing), the model selected for production implementation will show improved accuracy, recall, and precision scores.

Problem and solution summary

Summary of the problem:

A bank's consumer credit department aims to simplify the decision-making process for home equity lines of credit applications. To accomplish that, they will adopt the Equal Credit Opportunity Act's guidelines to establish an empirically derived and statistically sound model for credit approval scoring. The model will be based on the data obtained via the existing loan underwriting process from recent 5,960 applicants who have been approved for a home equity line/loan of credit. 13 input variables were registered for each applicant with 20% (1189 cases) of cases resulting in adverse or defaulted outcome. A comfortable range of default for banks is 3 to 5 percent based on national average.



The objective is to build a classification model using predictive modeling techniques to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan. The model created must be interpretable enough to provide a justification for any adverse behavior (rejections) in compliance with ECOA.

The end goal of the predictive model is to reduce default rate of 20% to an industry standard of 3% and reduce company losses in the process.

Reasons for solution design:

Leveraging data science and machine learning methods enables the automation of the loan approval process, reducing susceptibility to biases and inaccuracies. Developing an easily understandable model ensures transparency in the loan approval process, allowing for clear explanations to customers. Furthermore, adherence to the guidelines outlined in the Equal

Credit Opportunity Act ensures fairness in the loan approval and denial process, preventing discrimination based on race, color, religion, national origin, sex, marital status, or age. The balance achieved between accuracy, precision and recall for Random Forest Classifier with Hyperparameter Tuning model makes it the best solution to reduce loan defaults and improve adherence to ECOA regulation.

Solution design's effects on the business problem:

The suggested design solution has the potential to enhance the bank's loan approval process significantly. Through automation and the reduction of human biases and errors, the bank stands to save valuable time and resources while ensuring a fair and transparent loan approval system. Leveraging data science and machine learning, the bank can construct a precise and dependable model capable of predicting creditworthiness and pinpointing borrowers at risk of default. This approach has the potential to decrease loan repurchase and enhance the overall profitability of the bank. In summary, the proposed solution design offers an opportunity for the bank to enhance customer satisfaction, mitigate regulatory and default risks, as well as streamline its loan approval process.

Recommendations for implementation

Key recommendations to implement the solution for stakeholders:

- **Improve loan application system**

Based on the evaluation of performance metrics and the importance of features in the models, the improvement of loan application system that eliminates missing data and allows for complete data entry is recommended. All models used showed a great degree of importance for missing debt-to-income ratio. Debt-to-income ratio is one of the most important factors in determining whether a customer has the ability to repay a loan hence if the process to eliminate the lack of data entry is too expensive for the bank, a less expensive option of implementing only DEBTINC information completion is suggested.

- **Implement Random Forest Classifier with Hyperparameter Tuning model**

The model showed a good balance between accuracy, precision and recall and is relatively easy to interpret for regulators and stakeholders. Implementation of the model should be accompanied by robust training of loan officers to ensure understanding of how to use the model effectively and interpret the results accurately.

- **Automate the loan approval process**

Utilizing the model's output in an automated loan approval system can help mitigate human biases and errors, enhancing the speed and efficiency of loan approval processes while

alleviating the workload on loan officers and underwriters, hance reducing the cost of loan approval.

- **Monitor performance of the model**

Continuously observe the model's performance and, when needed, retrain it to uphold accuracy and stay current with evolving trends and patterns in the data related to loan applications as well as everchanging regulations.

Benefits of implementing the model:

As far as model implementation goes, the adoption of the Random Forest Classifier with Hyperparameter Tuning model is recommended due to its ability to exhibit notable accuracy and precision across both the training and test datasets. Furthermore, its analysis of feature importance has successfully identified crucial variables for predicting the target variable, enhancing the decision-making process. Additionally, compared to more intricate models such as neural networks, the Random Forest model is relatively easier to interpret.

Moreover, the Random Forest model provides interpretability through its feature importance metrics, facilitating an explanation of the decision-making process. It enables the identification of significant variables contributing to the model's predictions, a valuable asset in understanding the factors influencing loan approval.

Random Forest Classifier with Hyperparameter Tuning model exhibit resilience to outliers, non-linearity, and multicollinearity in the data, making them particularly well-suited for this problem where human biases could impact the data.

In summary, the model stands out as the most robust and accurate solution for the given problem. By leveraging multiple decision trees and selecting optimal features for predictions, it demonstrates superior generalization on unseen data. The feature importance analysis also offers insights into the critical factors influencing loan approval.

Consequently, the adoption of the Random Forest Classifier with hyperparameter tuning is recommended for constructing a dependable and interpretable model for credit scoring in this specific context. It is believed that a combination of data collection improvement and Random Forest Classifier with Hyperparameter tuning model implementation will reduce default rate for the bank and significantly reduce regulatory risk of biased decisions as the model will continuously make approval and rejection decision based solely on data and not on preconceived biases.

Potential risks and challenges:

Potential risks to implementation of the model selected include but not limited to:

- Model has a small chance to misclassify certain applicants resulting in unfair and prejudiced loan denial. The chance of that happening with missing debt-to-income values is around 13% on a training set and 28% for a test set.
- Random Forest model selected can be computationally intensive which may lead to increased model run time impacting the real-time performance of the production system.
- Changes in customer behavior, economic conditions, or internal income and debt verification processes can impact the data distribution over time and significantly reduce performance of the model.
- Possibility of the model compromise by fraudulent actors to manipulate loan approval process for personal gain.

To address these risks, it is recommended to regularly monitor and update the model while implementing appropriate security measures to safeguard against potential cyber threats and to adhere to regulatory requirements. It is recommended to implement a more robust loan application system that would have hard stops and not allow application to go to production without all the data be completed and verified. It is essential to conduct thorough testing, validation, and ongoing monitoring of the model's performance in the production environment. Regular updates and re-evaluation of the model should be part of the deployment into production environment strategy to address changing conditions and maintain model effectiveness.

Appendix

Appendix 1: List of models used

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
1	Logistic Regression	0.80	0.82	0.05	0.04	0.8	0.69
2	Decision Tree	1.00	0.87	1.00	0.64	1.00	0.71
3	Decision Tree Tuned	0.89	0.87	0.69	0.65	0.74	0.69
4	Random Forest	1.00	0.90	1.00	0.67	1.00	0.82
5	Random Forest Tuned	0.97	0.89	0.99	0.76	0.87	0.72

Appendix 2: Feature importance