Team Members: Shuxian Huang, Shravan Vudumu, Jason Duke, Michael Daniels, and Keith Deuser
MGT 6203
April 21st, 2024

Team 004 Final Report: Credit Classification Conundrum

GitHub Link: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-4/tree/main/Code

Background and Overview of the Problem at Hand

In recent years, the field of financial services has undergone significant transformation driven by advancements in computing technology. Increased computing speeds have opened new avenues for applying complex machine-learning algorithms to traditional banking problems. One such enduring challenge is the underwriting process for loan applications, which is both time-consuming and costly. Traditionally, underwriting involves a detailed analysis of the applicant's creditworthiness, requiring substantial human labor introducing significant operational costs.

With the advent of more robust and faster computing solutions, it is now feasible to deploy comprehensive machine learning models that can process large datasets more efficiently and with greater accuracy. This technological shift presents a valuable opportunity to optimize the underwriting process, reducing both time and expense by automating the evaluation of loan applications. Particularly, machine learning models can swiftly identify "safe" claims—those unlikely to result in default—thus allowing human underwriters to focus their efforts on more complex cases.

Recognizing this opportunity, our team aimed to explore and identify the most effective machine learning techniques for enhancing the loan underwriting process. We focused on a variety of models known for their predictive accuracy and efficiency in binary classification problems.

With this backdrop of technological advancement and potential efficiency gains, our project specifically focused on the implementation of a binary classification approach within the machine learning models we selected. The dataset we utilized consisted of loan applications labeled with binary outcomes: ‘0’ for applications where the loan did not default, representing "safe" loans, and ‘1’ for those that did default, categorized as "at risk". This labeling facilitated the application of our selected machine learning models to simulate a real-world underwriting process where the primary task is to predict and flag potential default risks.

The data-driven approach aimed not only to validate the effectiveness of each model by distinguishing between the two classes but also to fine-tune them to maximize predictive accuracy and minimize false negatives—where a risky loan might be mistakenly classified as safe. This focus was critical because the cost of a false negative (approving a risky loan) could be more detrimental to a financial institution than a false positive (rejecting a safe loan).

Our methodology involved training each model on a portion of the dataset and then testing them on separate validation data to objectively assess their performance. By leveraging the predictive power of models like XGBoost, Decision Trees, SVM, and Logistic Regression, our aim was to determine which model could most effectively identify "at risk" loans with high reliability, thus providing a tool that could significantly increase the efficiency of the loan underwriting process by significantly reducing manual review times and operational costs.

Initial Hypotheses

Superior Performance of XGBoost: Based on its robustness and flexibility, our team hypothesized that the XGBoost model would outperform other models in terms of accuracy. XGBoost is renowned for its ability to manage diverse types of data and adapt dynamically, incorporating techniques like ridge and lasso regression to optimize model parameters. This versatility was expected to make it particularly effective in predicting loan default risks within our dataset.

Cost-Optimized Cutoff Threshold: When optimized for minimizing financial losses rather than maximizing accuracy, we posited that the cutoff threshold would be lower than the most accurate cutoff. This hypothesis stems from the assumption that the cost of a false negative (failing to identify a loan likely to default) is significantly higher than that of a false positive (incorrectly flagging a loan as risky). Therefore, a lower threshold might be preferable to reduce the risk of costly defaults, even at the expense of increasing false positives.

Team Members: Shuxian Huang, Shravan Vudumu, Jason Duke, Michael Daniels, and Keith Deuser
MGT 6203
April 21st, 2024

Data Source and Selection:

The data for our project was sourced from Kaggle.com, a platform hosting diverse datasets. We focused on two main sources: application_train, which contains 122 columns of demographic and other factors related to an individual's credit application, and bureau, which includes data on credit bureau-related debts with 17 columns. These datasets are linked by the SK_ID_CURR key, forming a one-to-many relationship that we flattened for simplicity and effectiveness in our analysis.

Data Cleaning and Transformation:

Our initial task was to consolidate the application_train and bureau datasets by joining them on SK_ID_CURR and summing up the AMT_CREDIT_SUM and AMT_CREDIT_SUM_DEBT fields from the bureau data for each applicant. This transformation helped reduce complexity without sacrificing vital information.

The cleaning process was thorough:

> Managing Missing Values: Our dataset initially had 217,150 records, with 209,935 displaying at least one missing value. Missing values in critical fields such as EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 were addressed by setting defaults or imputing with mean values, depending on the nature of the data.
>
> Outliers Management: Extreme values in CNT_CHILDREN and reported income were capped or removed to avoid skewing our models.
>
> Dimensionality Reduction: Attempts to reduce dimensionality using PCA and k-Means Clustering were less fruitful than anticipated, leading us to manually refine the dataset to 49 most relevant fields, based on our expertise and the data's predictive power regarding the outcome.

Key Variables:

The most critical variables in our final model included AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, and the engineered features from the EXT_SOURCE series, which were indicative of an applicant's creditworthiness and risk profile.

Exploratory Data Analysis (EDA) Insights:

EDA revealed interesting patterns, particularly in how different demographic factors correlated with loan defaults. The engineered feature CREDIT_MISSING helped to identify segments of the population that were underrepresented in traditional credit scoring methods but exhibited higher risk levels.

Feature Engineering:

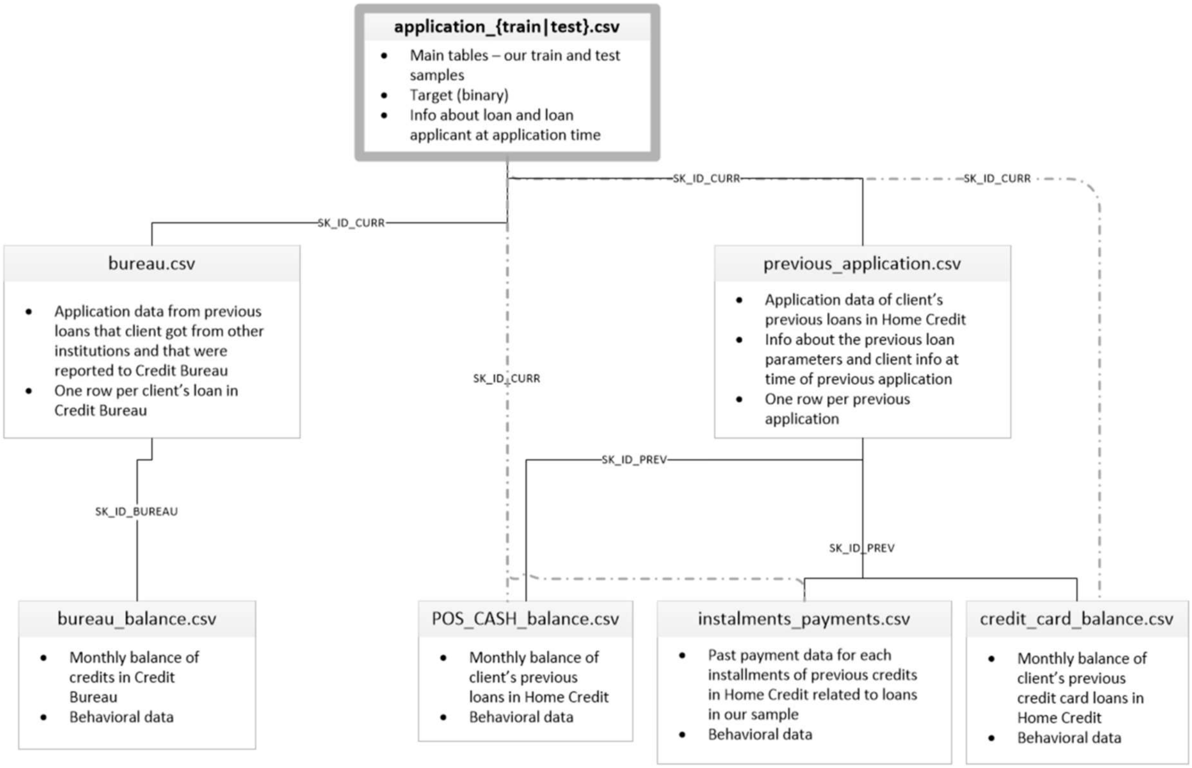
Our feature engineering efforts were pivotal in enhancing model performance. For instance, the creation of the CREDIT_MISSING binary indicator allowed us to differentiate applicants lacking complete credit scores, proving crucial in refining our models' predictive accuracy.

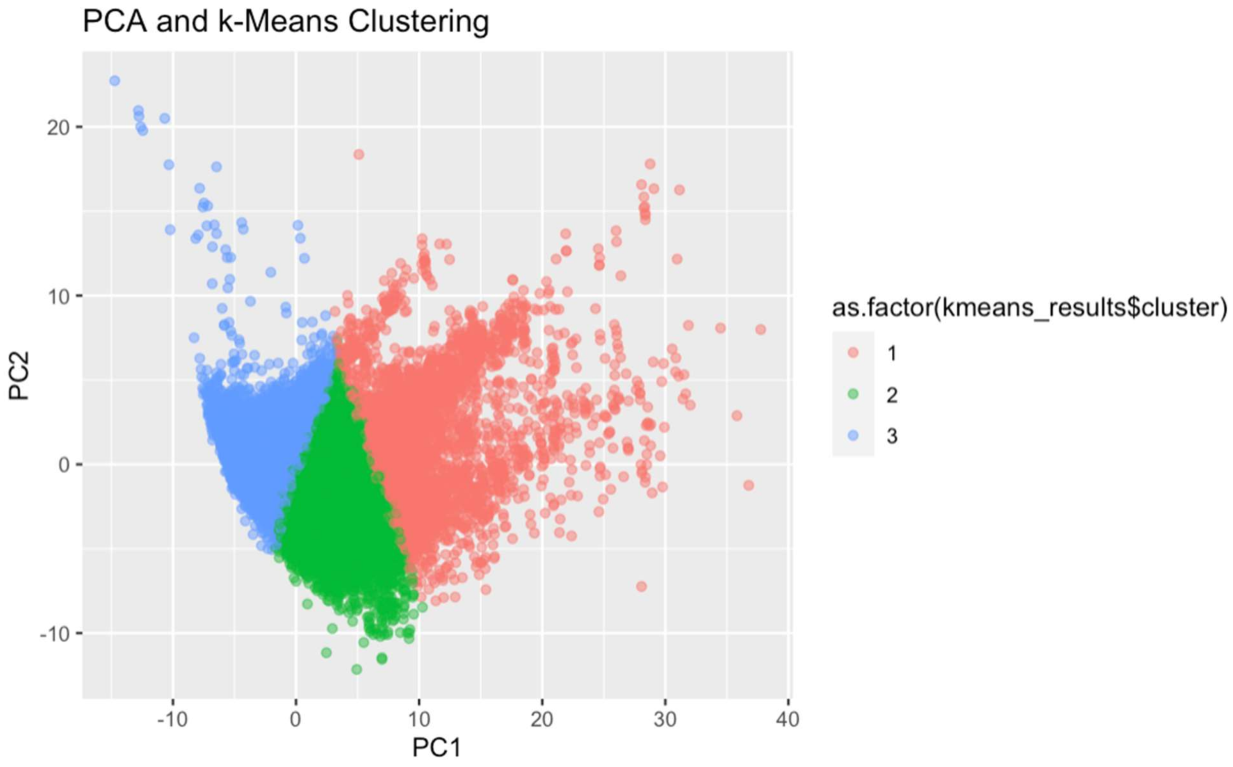
Summary:

This comprehensive data preparation and analysis phase set a solid foundation for the subsequent modeling steps. By meticulously cleaning and transforming the data, we ensured that the insights derived were both dependable and actionable, directly feeding into the strategic objectives of our project.

Shown below are visuals to help the reader visualize the transformation and cleaning processes. A PCA visual was also included to help the reader visualize the feature selection process.

Team Members: Shuxian Huang, Shravan Vudumu, Jason Duke, Michael Daniels, and Keith Deuser
MGT 6203
April 21st, 2024


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data



Team Members: Shuxian Huang, Shravan Vudumu, Jason Duke, Michael Daniels, and Keith Deuser
MGT 6203
April 21st, 2024

Overview of Models Used and Comparative Analysis

In this project, we employed a diverse array of machine learning models, each selected for its strengths and applicability to binary classification challenges. The primary objective of these models was to assess loan applications to determine their risk levels. Specifically, each model classified loans as either '0' for those that are safe and can be approved without further underwriting, or '1' for those that pose a higher risk and therefore require manual underwriting. The models we used included:

1. **Random Forest/Regression Tree:** This ensemble model uses multiple decision trees to make predictions, reducing the risk of overfitting compared to individual decision trees. It is particularly effective for handling nonlinear data with large feature sets.
2. **Support Vector Machine (SVM):** SVM is a robust classifier that works well in high-dimensional spaces, making it suitable for the dataset with numerous variables. It constructs a hyperplane in an n-dimensional space to classify data points distinctly.
3. **K-Nearest Neighbors (KNN):** KNN is a supervised learning classifier which uses proximity to make classifications about the grouping of individual data points. This is achieved through choosing an optimal k-value and calculating the distance to the *k* nearest data points.
4. **XGBoost:** Standing for eXtreme Gradient Boosting, this model has been recognized for its efficiency and performance across a variety of prediction tasks. XGBoost uses gradient boosting frameworks at its core, which is highly effective in handling sparse data and distinct types of predictive modeling.
5. **Logistic Regression:** As a more straightforward approach, logistic regression estimates the probabilities using a logistic function, which is particularly useful for binary classification problems like ours.

Each model was chosen based on its ability to identify and differentiate between low-risk and high-risk loans accurately, optimizing the underwriting process by reducing the need for manual intervention.

Model Evaluation and Accuracy Testing Methodology

Our team employed confusion matrices as a critical tool to rigorously assess and compare each model's performance. A confusion matrix is a table often used in classification problems to visualize an algorithm's performance. Each matrix summarizes the prediction results on a classification problem, breaking down the counts of true positives, false positives, true negatives, and false negatives.

Threshold Adjustments:

We conducted a detailed threshold analysis to fine-tune and determine the optimal cutoff threshold for each model. This threshold is especially critical for achieving the best balance between sensitivity (true positive rate) and specificity (true negative rate).

Accuracy Metrics:

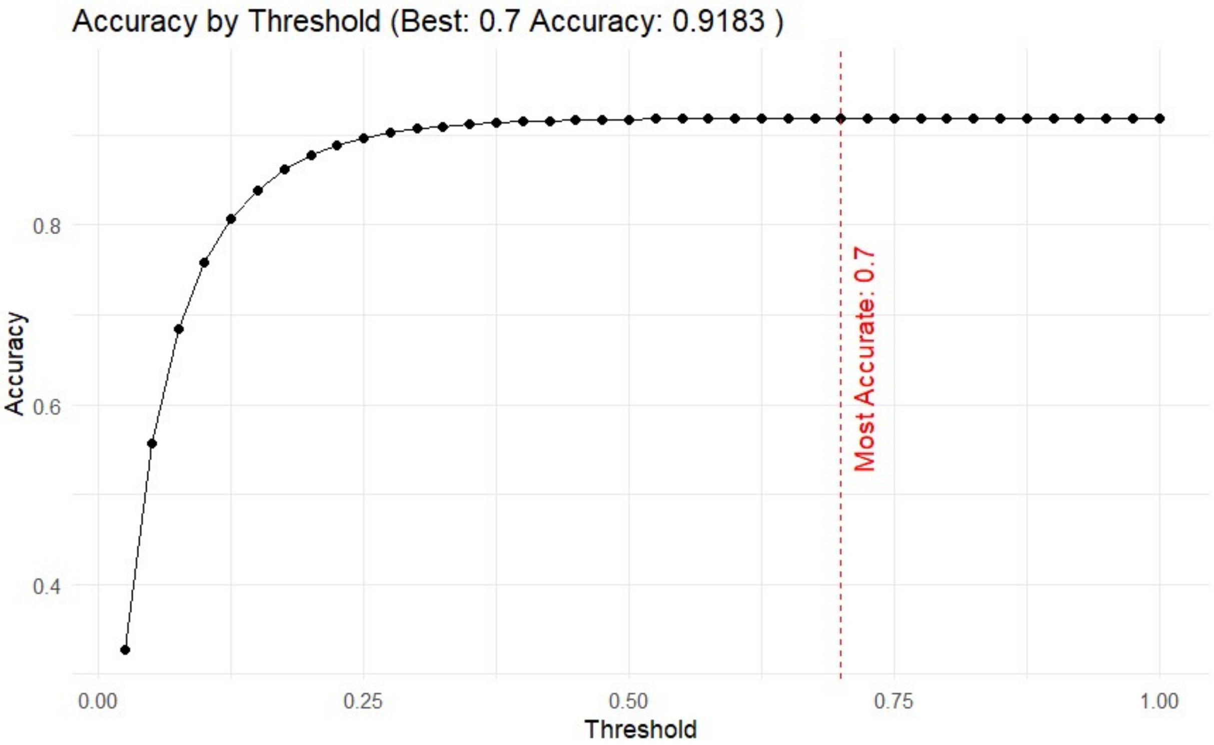
We recalculated the confusion matrix at each threshold increment, which served as the basis for computing the model's accuracy. Accuracy was calculated using the formula:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ Observations}$$

This metric measured how well each model correctly predicted defaults and non-defaults.

Performance Comparison Using Confusion Matrices:

XGBoost: This model showed the highest adaptability and accuracy across different thresholds. At a cutoff of 0.7, XGBoost achieved an accuracy of 91.83%, the best performance among the models evaluated. This prominent level of accuracy at a higher threshold highlighted XGBoost's ability to effectively differentiate between the classes without compromising on predictive reliability. Below is a chart highlighting XGBoost's accuracy at each threshold.

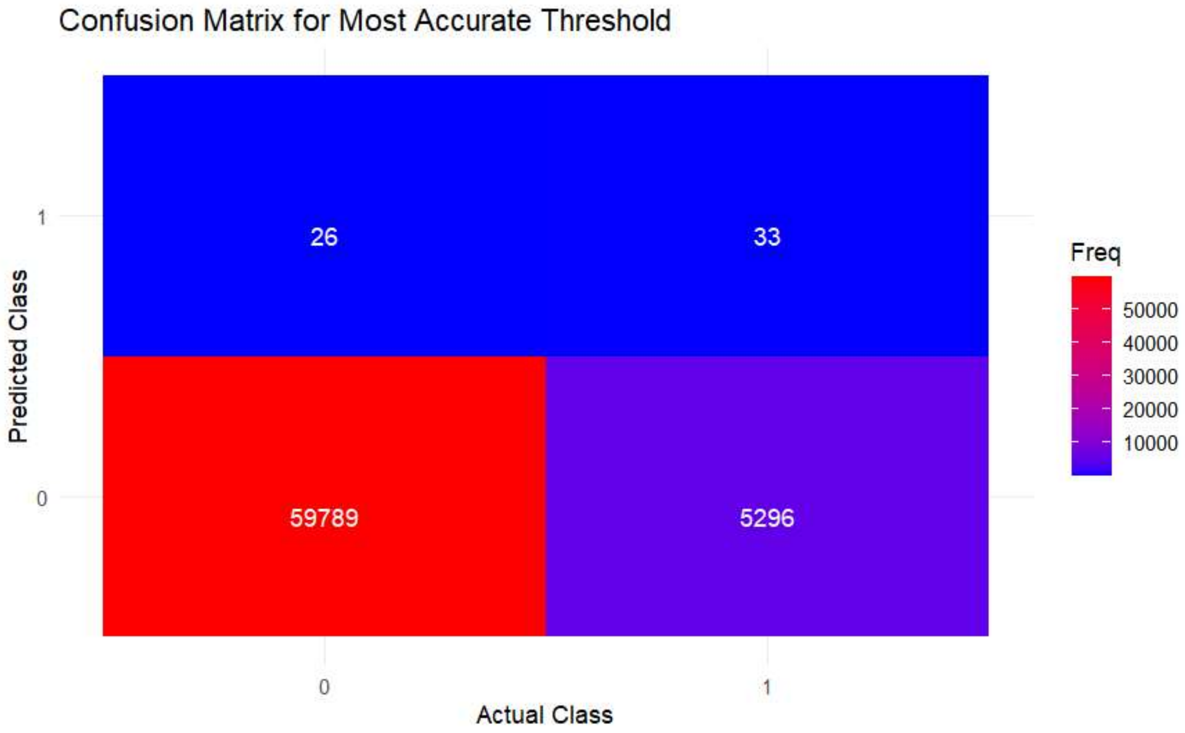


Other Models: For models like Random Forest, SVM, and Logistic Regression, we also adjusted thresholds in similar increments. These models exhibited varying levels of accuracy across different thresholds, but none matched the optimal balance of accuracy and threshold efficiency demonstrated by XGBoost.

By systematically adjusting the threshold and analyzing the outcomes through confusion matrices, we ensured a comprehensive evaluation of each model's capabilities. This approach not only allowed us to confirm XGBoost as the most effective model but also facilitated an understanding of how threshold settings impact the practical deployment of these models in predicting loan defaults.

Further Extrapolation and Analysis

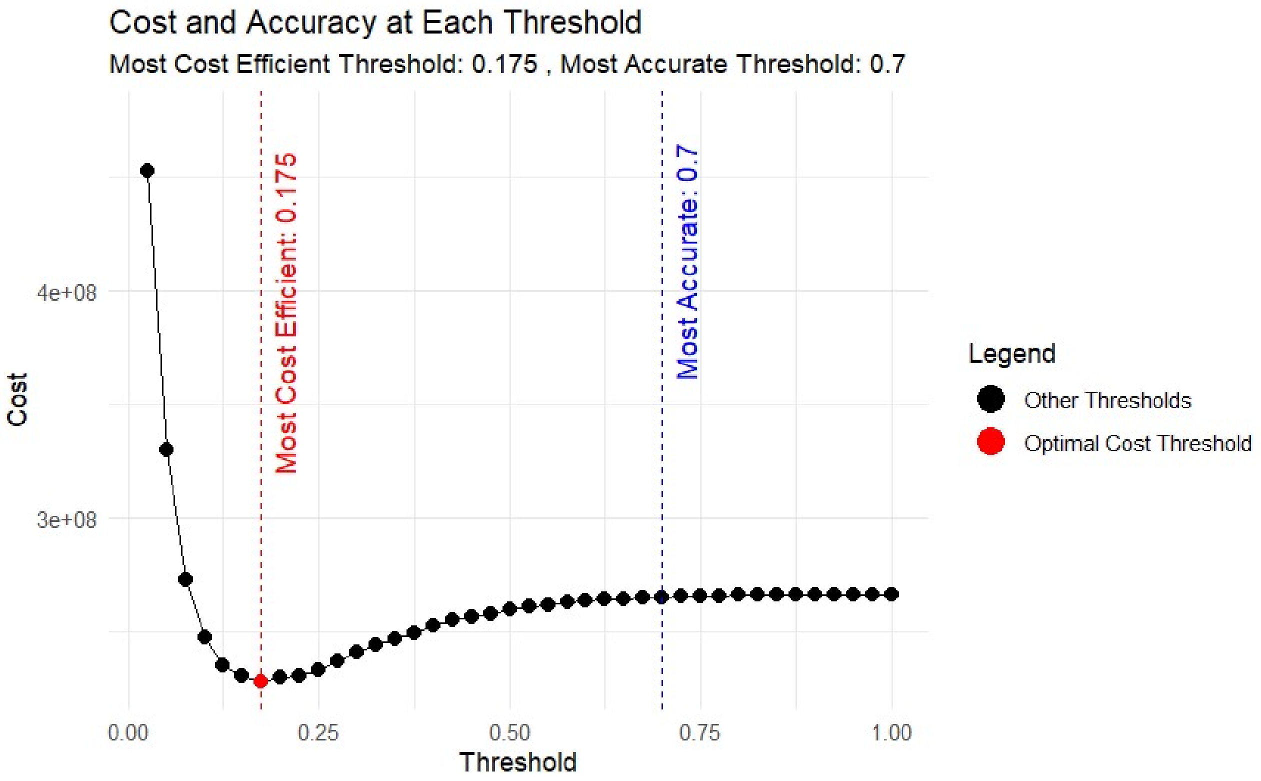
While the XGBoost showed promising accuracy rates, we must examine the confusion matrix associated with the cutoff at 0.7 to examine further the model's practicality in insurance and cost-benefit analysis. Shown below is the confusion matrix:

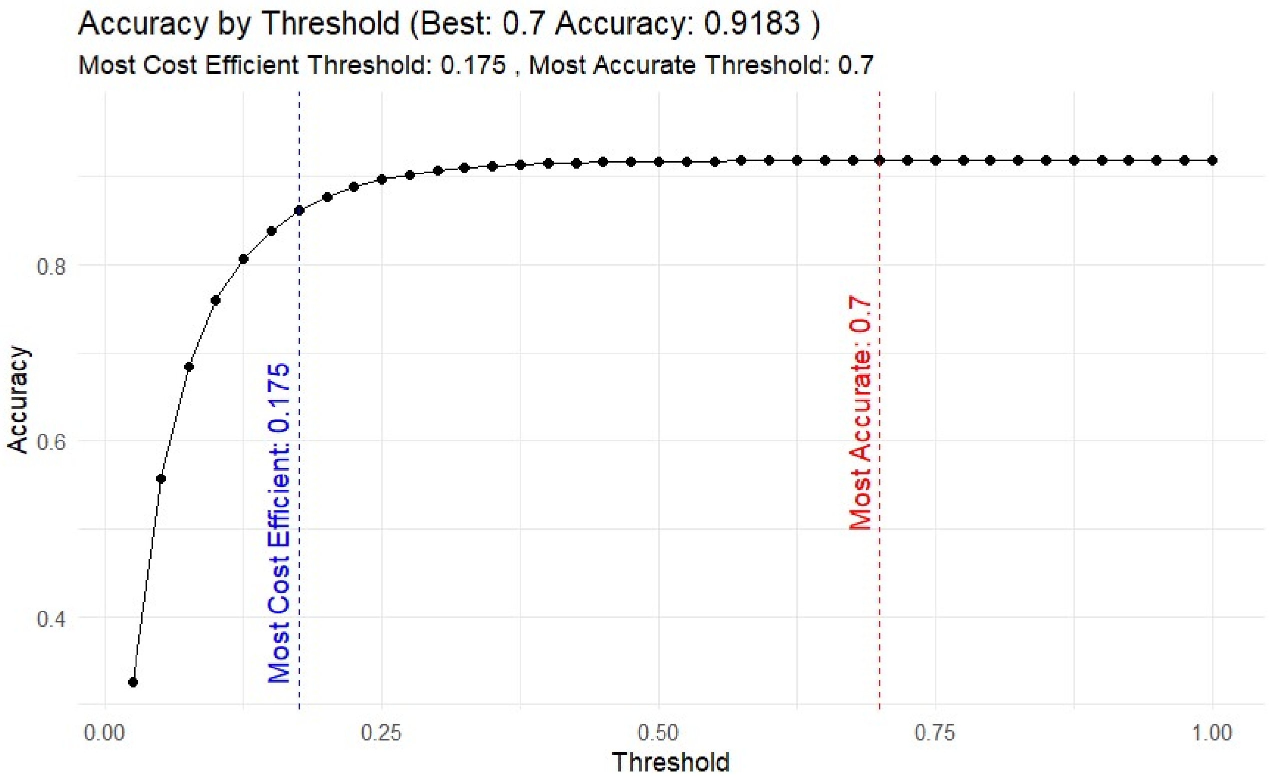


As demonstrated, the most accurate model grappled with a high rate of false negatives. In the context of loan underwriting, the consequences of false negatives are particularly severe. A false negative, where a default is incorrectly predicted as safe by the model, can result in significant financial losses, including the costs associated with default, legal fees related to bankruptcy proceedings, and additional expenses involved in recovering and reselling the property. Conversely, the cost of a false positive—erroneously classifying a dependable borrower as risky, requiring manual underwriting—is primarily limited to the time and resources spent on additional underwriting.

Given these dynamics, our team needed to optimize the model's threshold to minimize the rate of false negatives. However, setting the threshold too conservatively would lead to underwriting all applications, thereby increasing operational costs and negating the efficiency gains offered by the model. To address this challenge, we developed an optimization model to balance these costs effectively. Our analysis determined that the cost of a false negative is approximately five times greater than that of a false positive; as such, the cost of a false positive was set to $10,000.00, and the cost of a false negative was set to $50,000.00. Plugging in these estimated costs into the model, we found that the estimated total cost of wasted expenses at the most accurate threshold is $265,060,000.00. As such, we determined it was necessary to see if a more optimal model could be found relative to the cost.

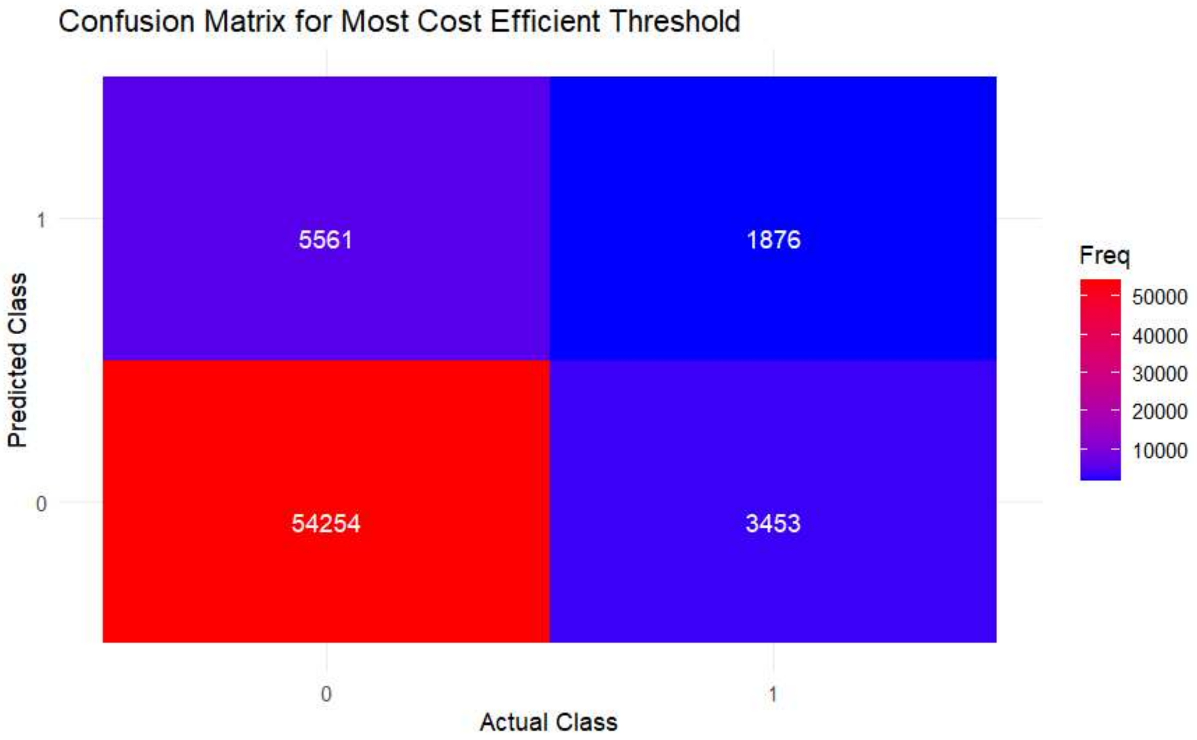
Realistically, achieving zero false negatives is unattainable due to various systematic factors and unpredictable elements, such as health complications, which can result in even the most ideal candidates defaulting on their loans. Consequently, our objective was to find an optimal threshold that minimized overall costs without the unrealistic goal of eliminating all risk. Below is a chart illustrating the cost implications at various threshold levels.





By adjusting the cutoff point to 0.175, the model achieved significant cost savings, albeit with a trade-off in accuracy. The new accuracy rating of the model at a cutoff of .175 was 86.16%. However, the estimated cost of wasted expenses was reduced to $228,260,000.00. Specifically, reducing the cutoff point resulted in a 6.17% decrease in accuracy but led to a 13.88% reduction in wasted expenses, indicating a favorable balance between cost and efficiency. This optimal cutoff was chosen because it minimizes the combined costs associated with false positives (unnecessary underwriting) and false negatives (defaults).

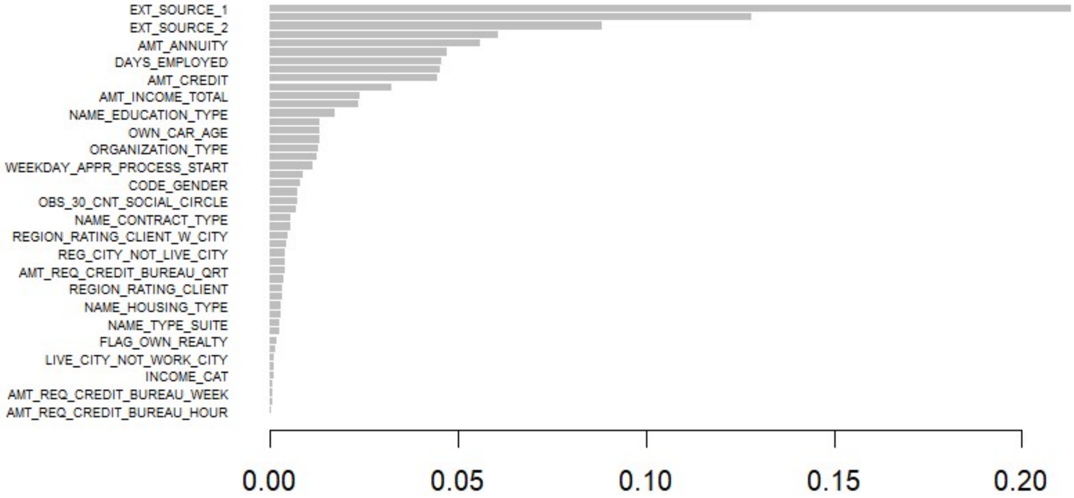
New Confusion Matrix Analysis: The new threshold has shifted the balance in our confusion matrix. Although the false positive rate has increased, this is offset by a substantial decrease in the false negative rate, which is critical given the higher financial stakes associated with defaults. Concurrently, the true negative rate has significantly improved, indicating that the model is effectively identifying a larger number of true non-default cases. This adjustment in the threshold demonstrates a strategic decision to prioritize reducing the costliest errors (false negatives) over a marginal increase in less costly errors (false positives).



Further Analysis and Beyond:

The analysis below illustrates the features that the XGBoost model identifies as most influential. While the model achieves high performance, it raises significant ethical and legal concerns. Notably, the model includes sensitive variables such as sex and certain zip codes, which may inadvertently introduce gender and racial biases. Such biases could affect the fairness of the model's predictions and expose the company to legal risks.

Current legislative trends are increasingly focusing on bias mitigation in machine learning models. Laws are being proposed that would mandate stricter standards to prevent discriminatory practices in automated decision-making systems. Given these developments, it is crucial that our model does not perpetuate or exacerbate existing societal biases.



Proposed Mitigation Steps:

- Reassessment of Feature Selection: The team would need to critically assess and exclude sensitive variables that contribute to biases, such as sex and zip codes.
- Demographic Testing: We recommend rerunning the model across different demographic groups to analyze discrepancies in outcomes. This testing would help identify any hidden biases and ensure that the model performs equitably across all groups.
- Continuous Monitoring: Once adjustments are made, the model should be regularly reviewed and updated in response to new data and evolving legal standards to ensure ongoing compliance and fairness.
- Implementing these measures will help mitigate potential biases and align the model's use with emerging legal frameworks. This would safeguard against ethical oversights and enhance the model's reliability and public trust.

Overall Conclusion

Our project successfully developed and evaluated a binary classification model using XGBoost to predict loan defaults. Through rigorous data preparation, feature selection, and model optimization, we achieved a high degree of accuracy, notably with our XGBoost model outperforming other tested algorithms in terms of predictive power and cost-efficiency.

Key Takeaways

- **Superior Model Performance:** The XGBoost model demonstrated exceptional effectiveness, highlighted by its superior accuracy at an optimal cutoff threshold. This confirms XGBoost's robustness and suitability for complex financial prediction tasks.
- **Importance of Ethical Considerations**: The project underscored the critical importance of integrating ethical considerations in model development, particularly in avoiding biases against protected classes. Adjusting model inputs and continuous monitoring for biases were crucial steps in aligning our practices with emerging legal standards.
- **Strategic Cost Management:** By optimizing the threshold for decision-making, our team significantly reduced wasted expenses without proportionally increasing risk, demonstrating the importance of precision in model calibration.
- **Insightful Data Preparation:** Comprehensive data cleaning and transformation processes were vital in constructing a reliable dataset. These processes ensured that our model was built on a solid foundation, capable of making accurate predictions.
- **Future Applications and Improvements:** The methodologies and insights gained from this project are scalable and can be adapted for broader applications in financial risk assessment. Future work could explore additional data sources and advanced modeling techniques to further enhance model accuracy and bias mitigation.

Closing Message

This project advanced our understanding of machine learning in financial applications and set a precedent for the responsible use of AI in high-stakes environments. Moving forward, the balance between technological innovation and ethical responsibility will remain paramount. We are optimistic about the potential of these tools to transform financial risk management, provided they are used with the utmost care for fairness and accuracy.

Works Cited

1. Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, Lu Peng, Explainable prediction of loan default based on machine learning models, Data Science and Management, Volume 6, Issue 3, 2023, Pages 123-133, ISSN 2666-7649, doi:10.1016/j.dsm.2023.04.003
2. S. K. Shaheen and E. ElFakharany, "Predictive analytics for loan default in banking sector using machine learning techniques," 2018 28th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 2018, pp. 66-71, doi: 10.1109/ICCTA45985.2018.9499147
3. Chen H., “Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model”. Computational Intelligence and Neuroscience. Volume 2022, Article ID: 7907210. doi: 10.1155/2022/7907210