Enhancing the Loan Approval Process

To: Home Credit
From: Aditi Gajjar, Andrew Kerr, Jadyn Ellis, Nathan Hill, Jamie Luna
Date: November 17, 2023

The purpose of this report is to (1) build a predictive model for assessing loan repayment capabilities of applicants, (2) provide a comprehensive evaluation of different machine learning algorithms in this context, and (3) offer actionable insights and recommendations for enhancing loan approval processes. We hope that this information helps you address the research question for the success of your company and its goal:

"Optimizing loan approval processes through predictive analytics to reduce financial risks and strengthen customer relationships."

This report is organized into five sections:

- The first section, "**Data Collection and Preparation,"** includes a description of the data set, cleaning and preprocessing steps, and feature selection rationale. (page 2)
- The second section, "**Model Selection and Validation,"** includes an overview of the models: logistic regression, support vector machines, and linear discriminant analysis, along with criteria for model comparison and selection. (page 4)
- The third section, "**Final Model, "** provides a detailed description and justification of the chosen final model. (page 6)
- The fourth section, "**Ethical Concerns,"** includes a summary of potential biases in the model and recommendations for ethical use. (page 7)
- The fifth section, "**Conclusion,"** summarizes our key findings and their significance. (page 9)

For future meetings or questions regarding our analysis, please contact our lead data scientists at, agajjar@calpoly.edu, adkerr@calpoly.edu, jellis13@calpoly.edu, nhill05@calpoly.edu, jluna28@calpoly.edu

I. Data Collection and Preparation

It is our understanding that you would like a comprehensive analysis of loan repayment capabilities based on applicant data, aimed at an underserved population. Below we have listed the features you have provided[1] to us for use in our analysis process. We will delve more into the feature choices later in this section.

Features:

1. **Total Income**: The amount of monthly total income, self-reported by the applicant
2. **College**: Whether the applicant holds a college degree or not
3. **Total Amount of Loan**: The total credit amount of the loan being applied for
4. **Occupation Type**: The broad category of work done by the applicant
5. **Education Level**: Whether the applicant has a college degree or not
6. **Credit Bureau Enquiries**: The total number of inquiries made about the applicant to the Credit Bureau
7. **Gender**: Gender of the applicant
8. **Commute**: Whether or not the applicant lives in the same city they work in
9. **Application Details**: Time and day the applicant began their process
10. **Marriage Status**: The status of the applicant's relationship
11. **Organization Type**: The type of industry the applicant is employed in
12. **Days Employed**: Number of days the applicant has been with current job
13. **Total Claims**: Number of claims on applicant's history
14. **Loan Type**: Whether the loan is for a lump-sum of cash or revolving credit
15. **Approval**: Previous loan approvals or rejections
16. **Previous Applications**: Number of previous loan applications by the applicant
17. **Amount Credit**: Total amount borrowed on previous loans
18. **Overdue Amount**: Amount overdue on previous loans

From the raw dataset provided to us by Kaggle, we cleaned the application data through a series of steps to ensure accurate and representative analysis. Initially, we dropped columns where 45% or more rows contained missing values to ensure that the features we selected were representative of the entire population. Subsequently, we filled missing numeric values with their average values and categorical values with their most frequent (mode) values. This method was adopted to maintain a robust dataset and to safeguard the integrity and utility of the remaining variables. Furthermore, we computed the correlation between each pair of features using the correlation heatmap (See Appendix A.1). In cases where two features exhibited high correlation, we assessed their Feature Importance Scores through Random Forests and opted for the feature with the higher score. Lastly, we refined the dataset to include only those features deemed

[1] https://www.kaggle.com/competitions/home-credit-default-risk/overview

influential in predicting loan repayment, as identified through exploratory analysis using fundamental insight visualizations and statistics (See Appendix A.2 for a few examples). In the end, we were able to reduce the number of features in the original model from 122 to 64. Of those 64, we researched the features that would result in the best predictions of loan repayment and reached a consensus of the 12 that we have listed above.

We also considered information from supplemental data related to the main application dataset. Specifically, we utilized the loan history and credit history data to create new variables that summarized each applicant's past experience with paying back loans. Over 95% of the applicants in the original dataset had some sort of prior credit history, and the few that did not were assigned mean values, so as to not punish them for their lack of experience.

To address the data imbalance, where only about 10% of the original data resulted in a loan offer, we chose to under-sample from the majority class. This was done to achieve a 70-30 split in our data, where 70% of the cases were those who did not have difficulties repaying their loans and 30% were those who did have some difficulties. We opted for this ratio as it pulls the data towards a balance, leaning away from the overcompensation that a 50-50 split would entail. This approach helps in creating a more stable and representative model, which is especially important in situations where the outcome classes are highly imbalanced.

Furthermore, we created subsets of variables based on similar characteristics, in order to determine which features were most important. We created six of these total variable combinations. The first held columns related to the education and work information from the applicant, including education level, income, and commute type among others. The next consisted of personal data, such as gender and marriage status. Another held information from the loan application itself, encompassing the amount of the loan, when the application was started, and other factors. Finally, the last unique subset contained details about previous loan applications and credit history. We then created an additional subset of features that we predicted would be most relevant to our predictions, based on summary statistics and plots. For completeness, we also included a full model containing all features from all of the subsets of data. Our objective through performing these meticulous steps was to provide you with a detailed understanding of the factors influencing loan repayment abilities, thus enhancing your decision-making process and risk management strategies.

II. Model Selection and Validation

Our Models

In our analysis, we have utilized three distinct models: logistic regression, support vector machines (SVM), and linear discriminant analysis (LDA).

Logistic Regression
Logistic regression is a statistical model used for binary classification. In our implementation, it helped us predict one of two outcomes: whether a person will be able to repay a loan (‘yes’ or ‘no’). This model works by estimating the probability of an event occurring (such as loan repayment) based on the input variables (like income level, employment history, etc.).

Support Vector Machines (SVM)
Support vector machines are a set of supervised learning methods used for classification and regression. In simpler terms, SVMs can be thought of as drawing a line that best separates the data into two categories (loan repayment and non-repayment) and assessing predictions by what side of the line a new point lies in.

Linear Discriminant Analysis (LDA)
Linear discriminant analysis is a method used to find a linear combination of features that optimally separates two or more classes of events or objects, and classifies future observations based on which group the observations are more likely to belong to. In other words, the goal of LDA is to retrieve maximum separation between the means of a given set of classes, while also ensuring minimal variance within classes. In the scope of our project, it attempts to optimally distinguish between those who can repay loans and those who cannot, based on the features provided.

For each of these models, we are using the same subset of determined features to ensure a fair and accurate comparison. By doing so, we can more precisely assess which model performs best under the same conditions. Specifically, our interest lies in exploring different subsets of the features previously listed, such as total income, loan amount, occupation type, and others. Thus, we ran six combinations of features for all three models, resulting in an overall total of fifty-four models after accounting for each of the three split methods we employed for each combination of predictors and model type.

Validation

In order to effectively compare the models and extract the best-performing one, we performed a rigorous validation process described below.

Data Splitting Strategies

In predictive modeling, splitting the data into training and test sets is essential for evaluating the performance of a model. The training set is used to build and train the model, allowing it to learn and identify patterns within the data. The test set, on the other hand, is used to evaluate the model's performance on unseen data. This split helps to ensure that the model can generalize well to new, unseen data, rather than overfitting the model to the data it was trained on. This process is critical for ensuring the accuracy and reliability of our predictive models

1. **Random Split**: We employed a random split as our first strategy, dividing the dataset randomly into training and testing sets. This method provides a baseline for model performance.
2. **Stratified Split**: The stratified split was based on the loan approval status (our target variable). This approach ensures that the proportion of approved and non-approved loans is consistent across both the training and testing sets, providing a more realistic evaluation of model performance.
3. **Income-based Split**: The last strategy involved splitting the data based on income levels, categorizing applicants into low and high-income groups. This split helps us understand how well our models perform across different economic segments, an important consideration in loan approval scenarios.

Metrics Compared

For a comprehensive evaluation of our models, we compared several metrics. These metrics serve as tools to objectively assess different aspects of a model's performance and effectiveness. By comparing these metrics, we can gain insights into how well each model predicts loan repayment capabilities, understand the trade-offs between different types of errors (such as false positives and false negatives), and ensure that the model performs fairly across different demographic groups.

1. **Accuracy**: Measures the proportion of correctly predicted outcomes (both approvals and rejections) out of all predictions made.
2. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: This metric helps in assessing the model's ability to distinguish between the classes (loan repayment and non-repayment) at various threshold settings.
3. **F1-Score**: A balance between precision and recall, the F1-Score is particularly useful in scenarios where equal importance is given to both false positives and false negatives.

4. **False Positive Rate (FPR)**: We included FPR in our metric set because it's essential to minimize the number of loan applications incorrectly predicted to default (i.e. unable to be repaid), since Home Credit's goal is to provide loans to those who are otherwise underserved by loan processes.

5. **Fairness Metric - Demographic Parity**: To ensure our models do not discriminate against any particular gender, we measured demographic parity, which assesses whether decisions (loan approvals) are independent of such a sensitive attribute.

On top of all of these metrics, we created an additional composite metric in order to compare performance across models.

$$combination\ metric\ =\ accuracy\ +\ 2\ *\ ROC\text{–}AUC\ +\ F1\ +\ 2\ *\ (1\ -\ FP\ Rate)$$

The rationale behind this formulation is to place a greater emphasis on the ROC-AUC and False Positive Rate, as these are particularly crucial in the context of loan approvals. By doubling the weight of the ROC-AUC, we prioritize the model's ability to distinguish between those who can and cannot repay a loan. Similarly, by emphasizing the inverse of the False Positive Rate, we aim to reduce the risk associated with incorrectly assessing a loan application as default.

III. Final Model

Model Comparison

We compared the resulting prediction metrics from all of our models, and found that the best-performing set of features was that containing all 18 filtered features. Additionally, LDA was the most reliable model in predicting loan repayments accurately, while being mindful of the factors that are crucial in a loan approval context (See Appendix B.1). The interpretation of the most important metrics of our final model are as follows:

- A ROC-AUC value of 0.65 indicates that our model is moderately effective at distinguishing between the two target classes.
- A false positive rate of 5.71% means that our model will incorrectly label an applicant as likely to have some difficulty repaying their loan when the applicant would not have any difficulties roughly 6% of the time.
- With an accuracy of 68%, we expect our model to correctly classify a majority of applications.

Group Classification

The results our model returns are the probabilities that an applicant will have some difficulty repaying their loan. To further improve the usefulness of our model, we decided to categorize applications into three groups based on the likelihood of the application being classified as having some difficulty:

- Most likely to have difficulties
- Somewhat likely to have difficulties
- Least likely to have difficulties

We decided to classify a plurality of applications into the middle group to strike a balance between identifying potential risky applications and avoiding overly conservative decisions (See Appendix B.2). This approach allows us to identify which applications you should focus additional attention and resources into analyzing, while still providing a fair assessment for those in the other two groups.

IV. Ethical Considerations

Data Concerns

> We would like to highlight several recommendations and considerations regarding the ethical use of the data we have employed in our predictive model.

Firstly, it is crucial to ensure that the data usage aligns with its intended purpose. This dataset, sourced from Kaggle for the explicit use of predicting loan repayment abilities, should be employed strictly within this context. Our team has ensured compliance with the rights and authorities granted for this data usage, adhering to both the terms of the dataset provider and relevant data protection laws, and we strongly encourage all applications of this data to remain consistent with these reliances.

Regarding consent, it is crucial to confirm whether the individuals represented in the dataset were informed about and consented to the use of their data in such analyses. This also applies to data sourced in the future for adding to our model. If these measures are not taken, actions should be taken to anonymize the data to prevent any breach of privacy. While we have taken steps to mitigate biases related to sensitive information such as gender, we encourage this practice to remain in future applications and versions of the model.

The security and confidentiality of the data are also important. Ensuring that the data is stored securely, with access restricted to authorized personnel, is vital for maintaining the trust of those whose information is being used and protecting them from potential data breaches.

Model Concerns

It is necessary that any final decisions made on the basis of our model do not violate the Equal Credit Opportunity Act (ECOA). If not already familiar, the ECOA prohibits loan approval discrimination based on any federally protected classes such as age, gender, religion, race, and marital status (FDIC)[2].

We performed a comprehensive analysis of gender discrimination in our model utilizing the demographic parity methodology. Demographic parity, in terms of gender, is calculated by computing the ratio between the probability that a female receives a class prediction pertaining to the inability to pay back a loan and the probability that a male receives a class prediction pertaining to the inability to pay back a loan. The utilized equation is as follows:

$$Demographic\ Parity\ = \frac{P(C=1|\ x \in G)}{P(C=1|\ x \notin G)}$$

In our final model (Section III), our demographic parity metric turned out to be 0.2747, meaning that on average, the probability of female-identifying loan applicants receiving a predicted target of 1, unable to pay back the loan, is 0.2747 times lower than that of male-identifying applicants. This metric shows a large disparity between the rates of loan approval between the genders.

Moreover, our final model contains variables pertaining to gender. ***If this model were to be used in any real-life applications, it is necessary to remove any variables that would cause underlying model bias between groups of protected classes.***

Our alternative model:

Throughout our model analysis, we additionally selected a model that performed the most fair between the genders. Moreover, we wanted this final "fair" model to not contain any predictors that are associated with any such protected classes. Our final fair model selection contained features pertaining to applicant loan histories, such as the number of past loans, the sum of overdue amounts, the time left on the loan, and the worst-case approval rating. Our computed demographic parity of this fair model is 1.2315, meaning on average, the probability of female-identifying loan applicants receiving a predicted target of 1, unable to pay back the loan,

[2] https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/4/iv-1-1.pdf

is 1.2315 times higher than that of male-identifying applicants. Because this value is close to 1, we can assume that the distribution of predicted class values between males and females is approximately equal.

There are, however, a few downsides to deploying this model. For one, we sacrifice metric values in favor of a more fair model, such as accuracy, roc-auc, f1-score, and false positive rate, implying that our predictive power of this model is lower than the final model (See Appendix C.1).

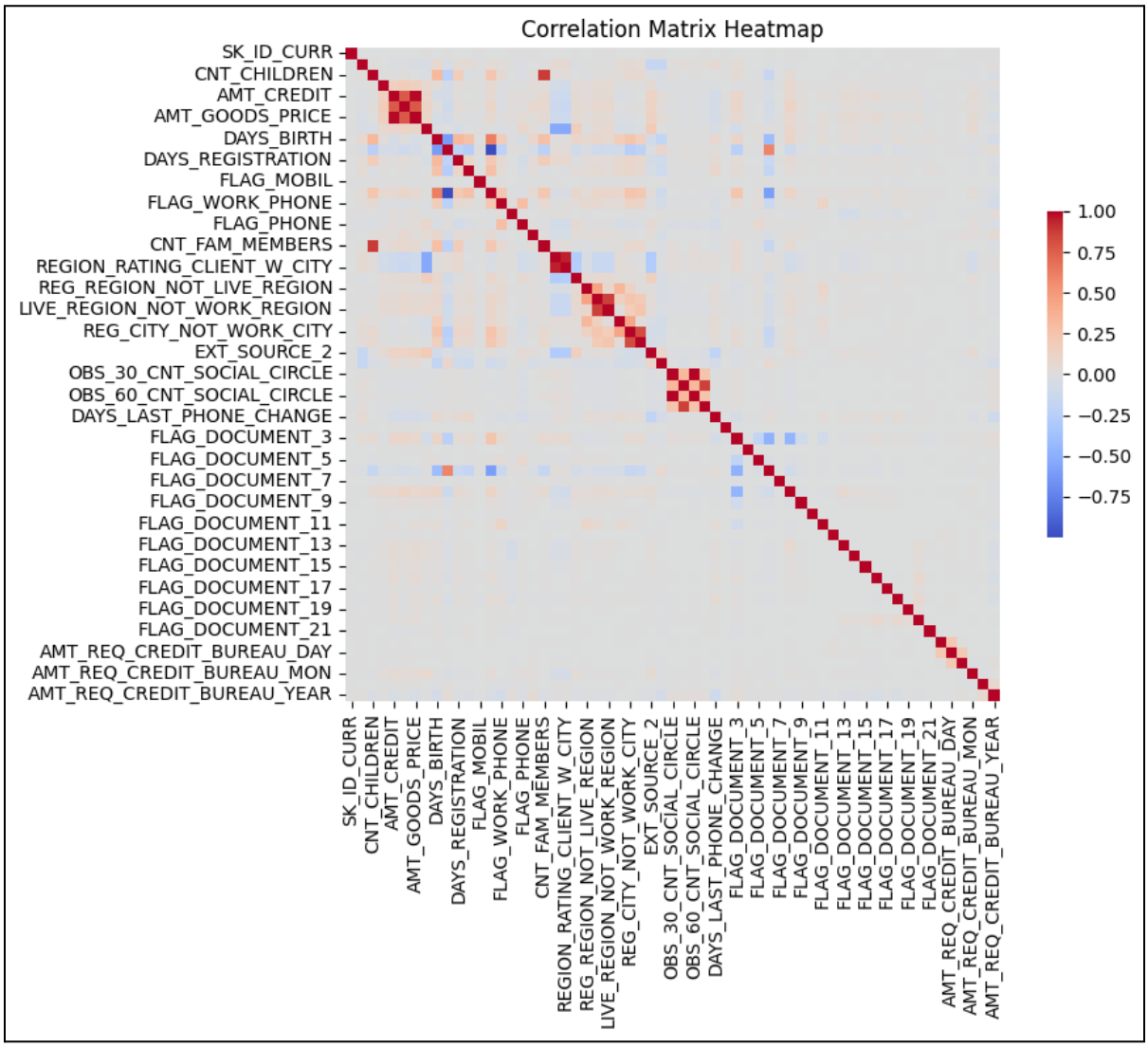
IV. Conclusion and Recommendations

On its own, our model is moderately effective in predicting whether or not an individual should be approved for a loan. However, the accuracy and other metrics highlight the importance of using the model as a tool in collaboration with more traditional methods rather than using it as the exclusive final decision. Thus, it is still important to look over applications manually. We recommend running applications through our model and sorting them by the category that our model outputs. The applications that are classified as least likely to have difficulty repaying their loans can be streamlined to be approved faster, and those classified as most likely to have difficulty can be sorted through for rejection. Those placed in the middle category should face a higher degree of scrutiny to determine the ultimate outcome of the application.

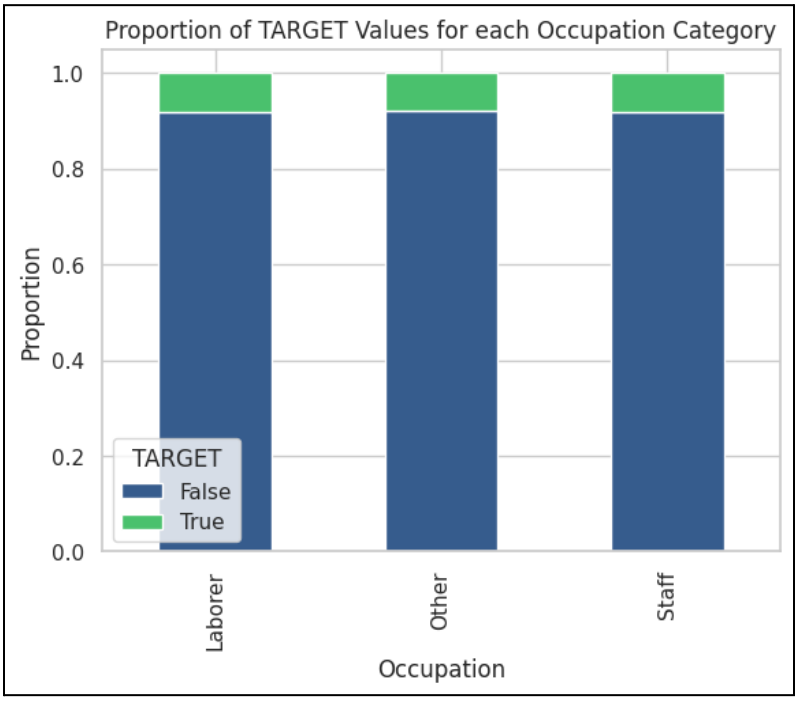
Appendix

Appendix A: Exploratory Data Analysis Visualizations

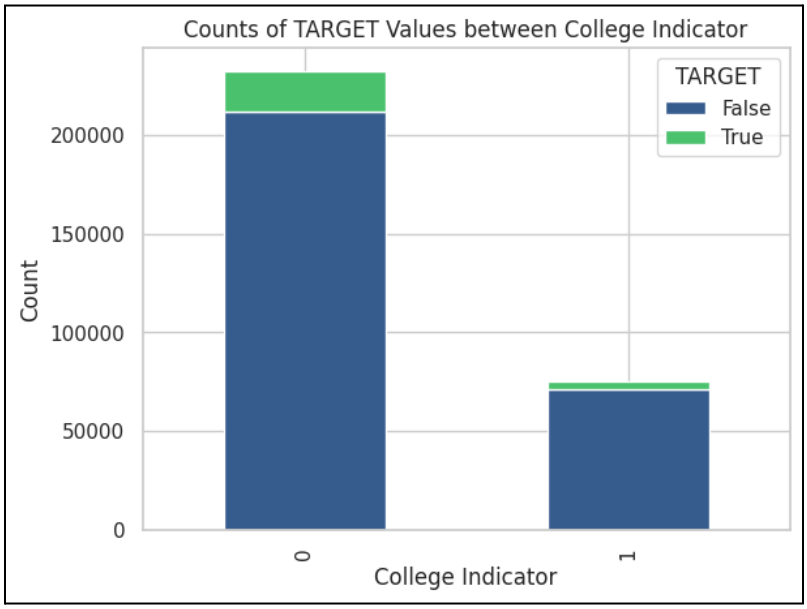
A. 1. Correlation Heatmap of Most Influential Factors



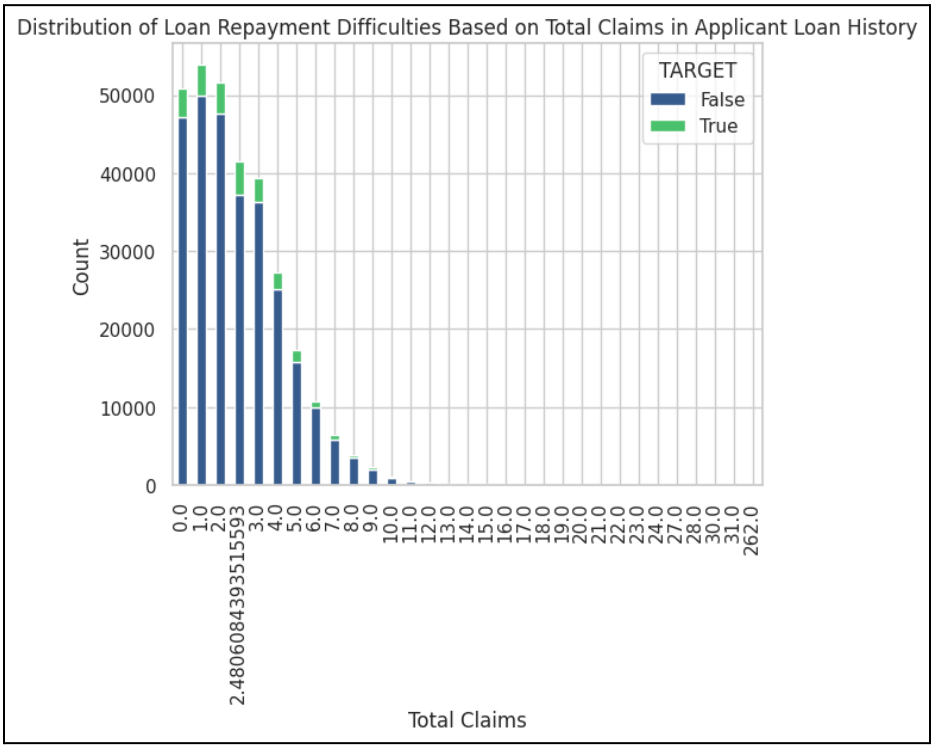
A. 2.1 Bar graph of the distribution of loan repayment difficulties based on occupation type category



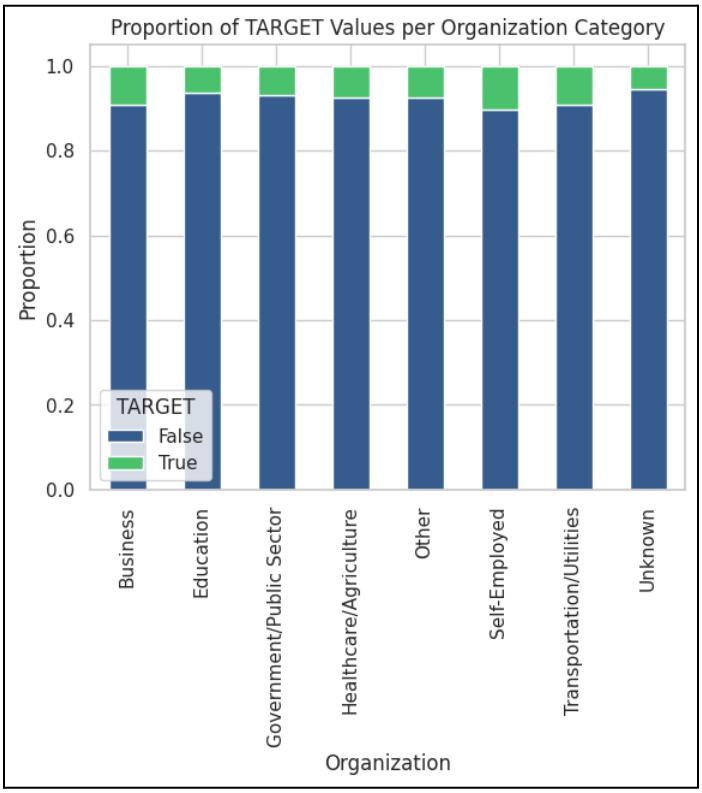
A. 2.2 Bar graph of the distribution of loan repayment difficulties based on the college experience category



A. 2.3 Bar graph of the distribution of loan repayment difficulties based on total number of claims into an applicants history

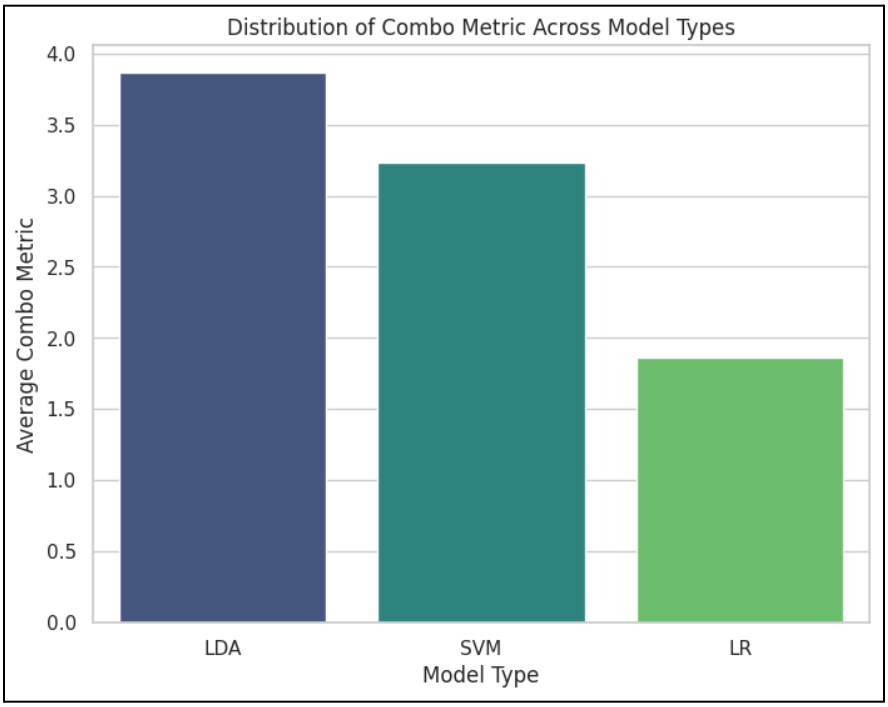


A. 2.3 Bar graph of the distribution of loan repayment difficulties based on employment organization type

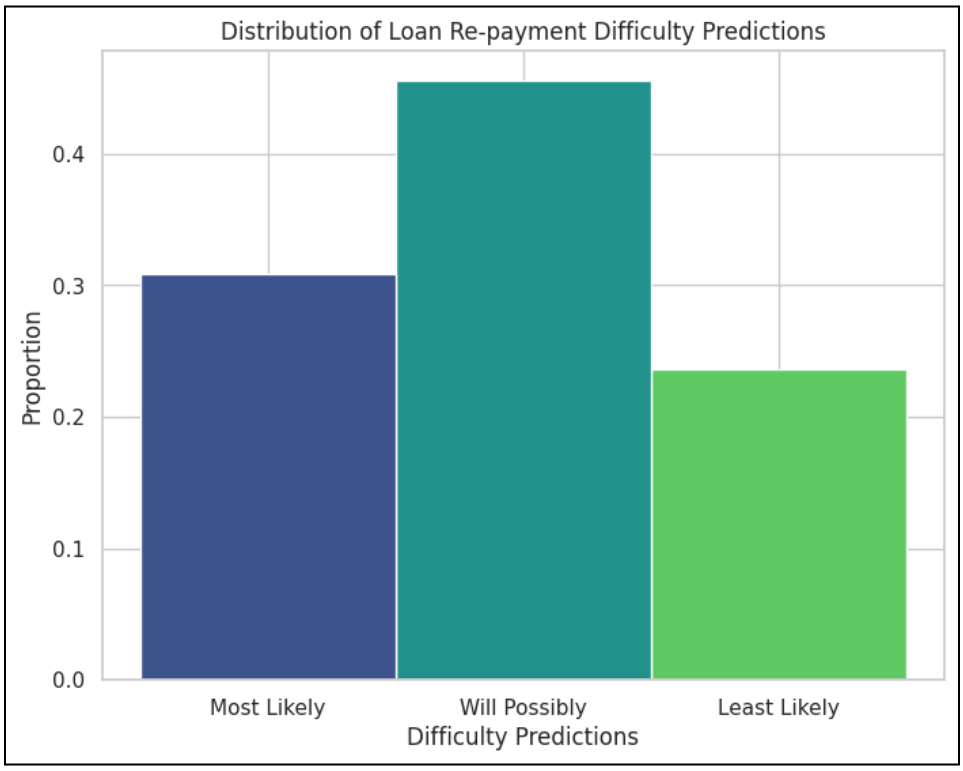


Appendix B: Model Metric and Final Model Visualizations

B. 1. Bar chart of average combination metric score across model types



B. 2. Bar chart of proportion of classified applications in each assigned category



Appendix C: Ethical Considerations

C. 1. Bar chart of distribution of metrics between fair and final models