Home Credit Project Proposal

DATA 403

Daniel Alvarez, Libby Brill, James Irwin

Data Summary

To help Home Credit best predict whether potential clients will repay their loans, we will use an internal application dataset that consists of over 300,000 observations, each representing a loan that has been taken out. There is information about the loan and the loan applicant at the time it was filled, such as loan status, the applicant's occupation, education level, and the size of their living area. The value of the response variable, "TARGET", is a 1 if the client is known to have payment difficulties in the past, and a 0 represents any other case. To ensure Home Credit best predicts the ability of its clients to repay their loans, we will perform a number of tasks. For example, we plan to deal with missing values by encoding them as a 0 if appropriate, such as for the basement area or car age, as well as including a string if possible, like saying "unknown" for an applicant's occupation type. We also plan to create new features that can condense our data as well as provide meaningful insight, such as collapsing income (low income, medium, high) and age (20-30, 30-45, etc) into groups, as well as a variable for the living area size and age. Lastly, some valuable predictors we plan to use include the type of loan taken out, annuity, and whether the applicant owns property or not.

To quickly understand how default rate varies across different categories, we have included several visualizations below:

Default Rate by Number of Children

Default Rate by Education Level
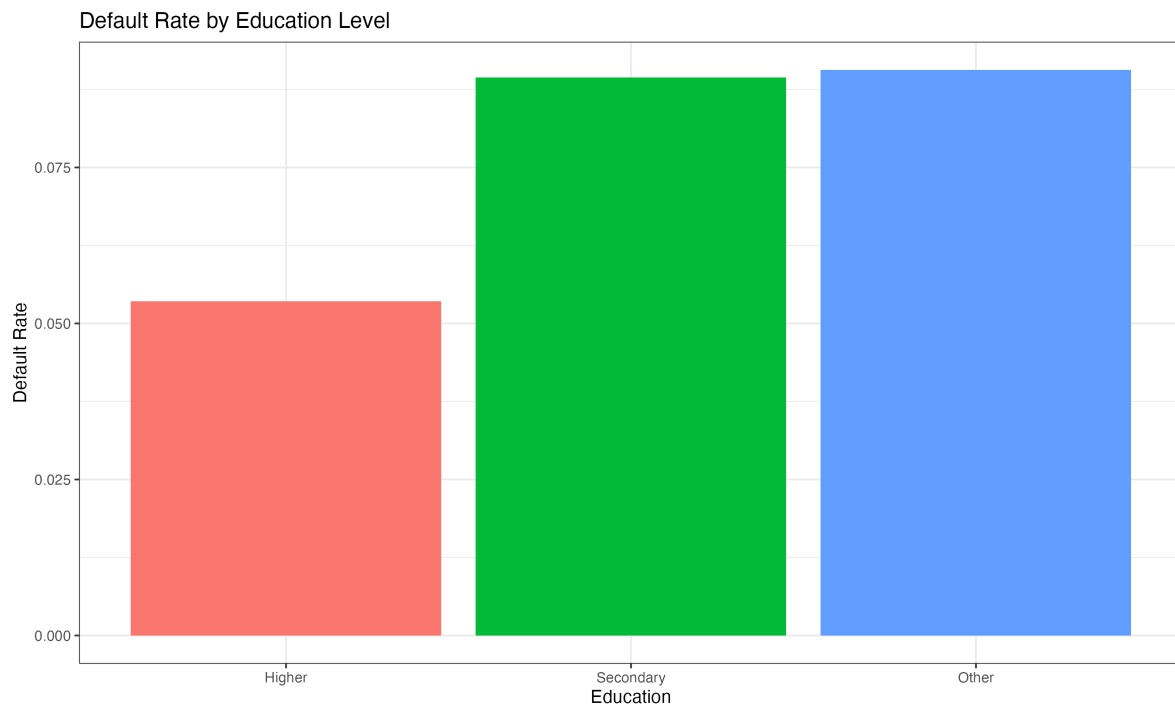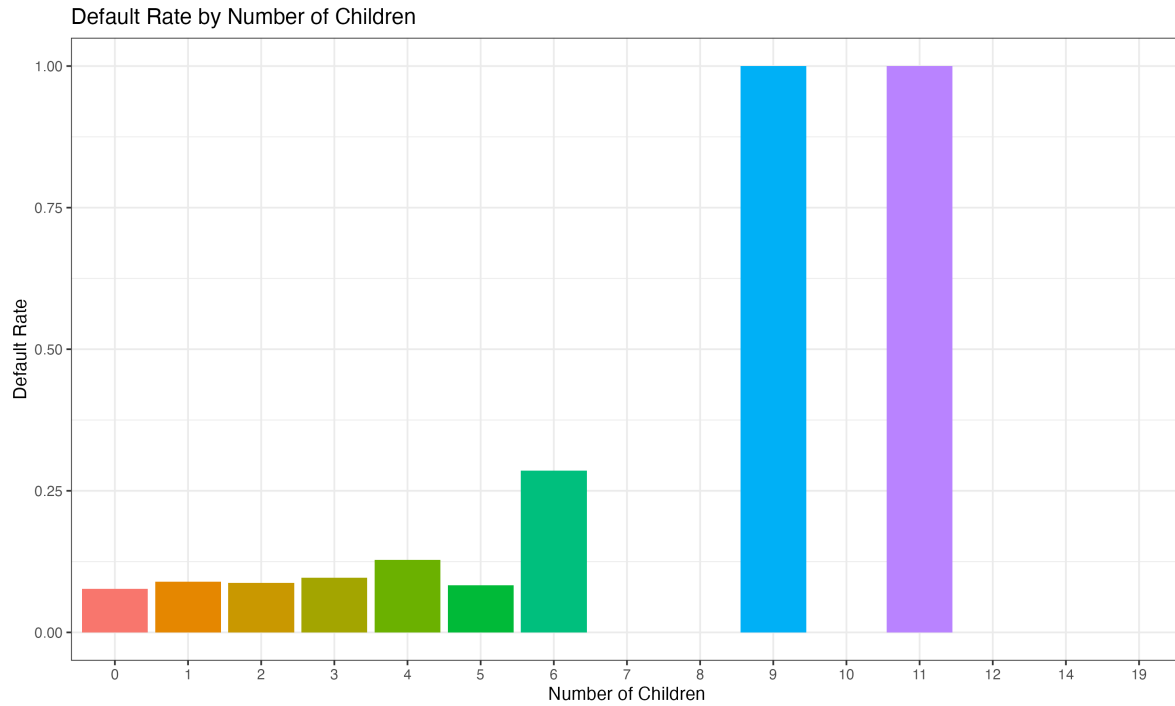
As the visualizations make clear, default rate varies greatly among people with different numbers of children and different education levels. These are just two different attributes within

the application dataset, so we feel confident that we will create a strong model when using all of the data.

Models and Metrics

Models

We will be fitting three different types of models: logistic regression, support vector machines (SVM), and linear discriminant analysis (LDA). While these are all classification models, they each have unique strengths and perform best under different conditions, making it worthwhile to evaluate all three. Logistic regression provides a stable, interpretable baseline for comparison. SVMs offer flexibility for capturing complex, nonlinear relationships. LDA, on the other hand, can be highly efficient when its distributional assumptions are satisfied. Comparing their performance allows us to determine which approach is most effective for our data's structure and patterns.

Testing

For model evaluation, we considered several approaches to splitting the data into training, validation, and testing sets. A completely random split is straightforward to implement and works well when the data are balanced across classes. However, when the data are imbalanced, a stratified split helps prevent bias and ensures more reliable evaluation metrics. In our case, because the dataset only includes individuals who have already received loans, a non-random split may provide a more realistic assessment. By selecting test and validation sets that better reflect the characteristics of all future applicants, rather than just those with approved loans, we can obtain performance estimates that more accurately represent how the models will behave once deployed.

Metrics

To evaluate model performance, we will examine several key metrics, including accuracy, ROC AUC, F1 score, and recall. Accuracy and ROC AUC provide an overall sense of how well the models distinguish between approved and unapproved applicants, while the F1 score balances precision and recall to account for uneven class distributions. However, recall will be a particular focus in this analysis, as it captures the proportion of qualified applicants correctly identified for approval. Because Home Credit strives to broaden financial inclusion for the unbanked population, prioritizing recall aligns with the goal of minimizing the number of creditworthy individuals who are incorrectly denied.

In addition to standard performance metrics, we will evaluate fairness using measures that capture whether the model's predictions are equitable across demographic groups. Specifically, we will consider Equal Opportunity, False Positive Rate (FPR) Parity, and Demographic Parity. Equal Opportunity aligns closely with the company's goal of expanding access to credit among the unbanked population. FPR Parity helps maintain both fairness and financial responsibility by checking that unqualified applicants are not disproportionately approved. Demographic Parity can highlight potential disparities even when differences in qualification are accounted for. Together, these metrics offer a balanced perspective on fairness, capturing both inclusion and consistency in model behavior.

Conclusions and Deliverables

Ultimately, we will produce for Home Credit a model that accurately and fairly predicts whether potential clients will be able to pay back their loans. Because Home Credit is dedicated to providing banking opportunities for the unbanked population, this model will be inclusive of people who may not traditionally receive loan assistance. Additionally, we will provide a discussion about the different factors that are associated with loan repayment, a detailed description of our model selection and validation process, and a report of our calculated metrics. Finally, we will evaluate the fairness of our model with respect to protected attributes like gender and age. We appreciate the opportunity to present our proposal and look forward to working with Home Credit.