Predictive Analytics for Loan Repayment

Rachel Roggenkemper: rroggenk@calpoly.edu
Matteo Shafer: mshafe01@calpoly.edu
Anagha Sikha: arsikha@calpoly.edu
Cameron Stivers: ctstiver@calpoly.edu
Sucheen Sundaram: sssundar@calpoly.edu

California Polytechnic State University - San Luis Obispo
November 17, 2023

Introduction

Home Credit Group requests a method to predict whether or not a loan shall be given. We propose a machine learning approach using multiple classification models to give us the most accurate and precise results. The model will fit training data and accurately predict whether or not the applicant receives a loan. Our models will be primarily evaluated based on ROC-AUC, accuracy, F1 score, and precision. These results will be presented using visual aids to cater to diverse stakeholders, ultimately assisting lending decisions and inventory management in both financial and retail sectors.

Background

For this project, we will be working with the data provided by the Home Credit Group[1] about loan applications and credit default. The main dataset we will be utilizing is the application_train.csv[2] data file. This is our table of interest, which includes static data for three hundred thousand applications where one row represents one loan in our data sample.

Table 1. Dataset Snippet

TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...	CODE_GENDER
1	202500.0	406597.5	24700.5	...	M
0	270000.0	1293502.5	35698.5	...	F
0	67500.0	135000.0	6750.0	...	M
0	135000.0	3126282.5	29686.5	...	F

The target variable was coded as 0 if the client paid back their loan with no difficulties and 1 if the client had difficulties paying back their loan.

[1] Retreived from https://www.kaggle.com/competitions/home-credit-default-risk/data
[2] Retreived from https://www.kaggle.com/competitions/home-credit-default-risk/data?select=application_train.csv

Data Preparation

Before our model tuning process, we prepared our data by selecting potential explanatory variables that we thought may be important predictors on whether or not somebody defaulted on their loan. The variables selected fell under the six categories listed below.

Client Financial Profiles
Here we looked at variables focusing on the client's total income, credit amount, amount of goods the client owns, and the amount of annuity.

Personal Demographic Information
Here we looked at variables focusing on the client's age, gender, highest education level, marital status, and the number of children the client has.

Employment and Occupation
Here we looked at variables focusing on the days the client has been employed and the occupation type of the client.

Housing Situation
Here we looked at variables focusing on the client's housing type.

Asset Ownership
Here we looked at variables focusing on whether or not the client owns real estate.

Loan Specifics
Here we looked at variables focusing on the contract type of the loan.

We created a new variable to group occupations into frequent and infrequent. The most common occupations of our data consisted of Laborers, Sales staff, Core staff, Managers, and Drivers. We also converted the age variable from days to years for simplicity. We made sure to dummify the categorical variables, so we could treat them as numeric and use them in our model. Lastly, we removed any observations that contained missing data to ensure our models were trained on consistent data.

After we completed our data preparation and feature engineering, we were interested in exploring the trends in our data so we could have an intuition of what predictors to include in our model.

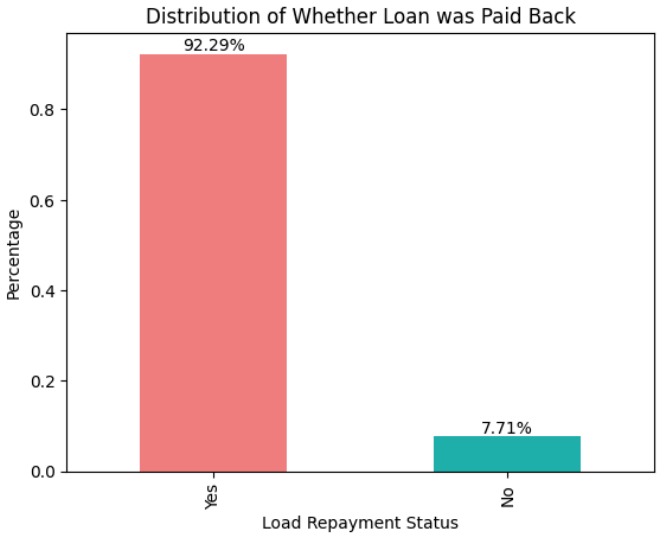

Figure 1: Distribution of Whether Loan was Paid Back

We first looked into the overall distribution of loan repayment in our data. As we can see in Figure 1, the majority of clients (92.29%) had no payment difficulties, while only 7.71% of clients had payment difficulties. This data imbalance will be important to keep in mind when fitting our model.

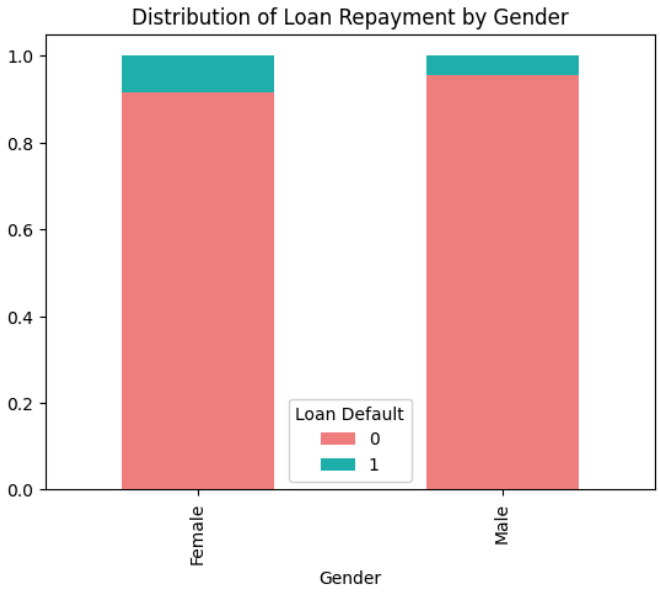

Figure 2: Distribution of Loan Repayment by Gender

We then looked into the distribution of loan repayment by gender to see if loan repayment varied. In Figure 2, we see that female clients default their loans at a slightly higher rate compared to male clients. This is important to investigate when looking into fairness metrics of our model and to see potential gender-specific factors influencing loan defaulting outcomes.

Model Selection

We used three main models to create predictions: logistic regression, support vector machine, and linear discriminant analysis. These models output classifications on whether future customers will default on their loans. To validate our models, we split the data three different ways: a completely randomly sampled split, a stratified split on gender, and a split based on whether the observed client had a common or infrequent occupation. To evaluate how powerful our predictive model is, we looked at the following metrics: accuracy, precision, recall, f1, misclassification rate, specificity, ROC-AUC, and fairness metrics on gender.

The models using a completely random sample had similar performances to the models using a stratified sample, and this could be because gender isn't as influential as some of the other features we used. But, because the models using stratified samples were similarly performing and gave us assurance that it would work similarly on future data, assuming a roughly 50/50 split in gender, we chose to apply the methods using a stratified sample over a completely random split.

When running our models using our non-random splits, we found that our models not only performed worse in accuracy and f1 score, but also produced very concerning fairness scores. What that told us is that our models would not have done well on people whose occupations are not very well represented in the data, because they were trained on frequently occurring occupations and tested on less frequently occurring occupations.

We adjusted our probability thresholds to select our final model, as the data only had people failing to pay back their loan as intended 8% of the time. Thus, with a base threshold of 0.5, we were almost never getting positive predictions. With a lowered probability threshold, we were able to more frequently predict applicants as potentially

unable to pay off their loan reliably. We took SVM out of consideration, as it was very computationally intensive, as well as producing poorer metric scores in comparison to the other two models, leaving us with logistic regression and linear discriminant analysis. Our accuracy, f1, ROC-AUC, and recall scores were slightly higher for linear discriminant analysis, however our fairness gender metric was very low. Our fairness metric for logistic regression was near-perfect at a parity of about 1.04, and we wanted to ensure our model classifies males and females at similar rates. Because of our obtained fairness metric, even though our accuracy and other metrics were not quite as high, we still decided that this was a superior model due to it being much more fair at only a small performance cost.

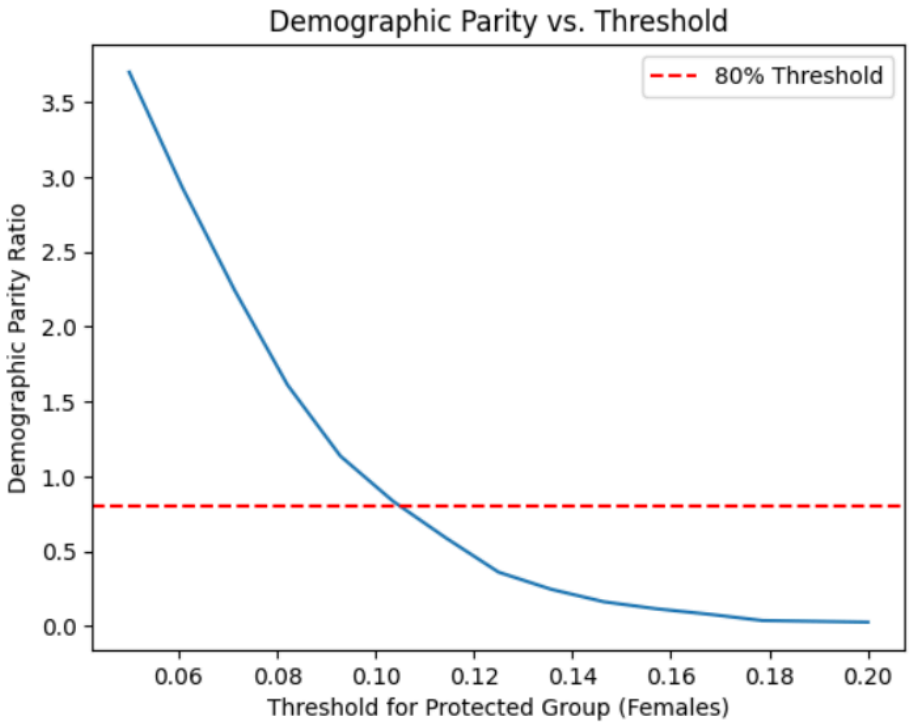

Figure 3: Demographic Parity vs. Threshold

Looking at Figure 3, which plots the trend between demographic parity and probability threshold, we see the demographic parity values that our model produces at different thresholds for our protected group (females) while keeping the male probability threshold at a constant. The threshold value is the probability cutoff that the model uses to classify applicants on whether or not they will have difficulty paying back their loan.

Final Model

After testing and performing cross validation on each of our candidate models, we have chosen our final model to be a Logistic Regression model with no penalty applied, which is trained and tested on a stratified split such that males and females are equally represented. In this model, we used a probability threshold of 0.15 to classify an applicant as potentially unable to pay out the loan. Logistic Regression, in general, performed only slightly worse than Linear Discriminant Analysis in terms of accuracy, f1 score, and ROC-AUC, but because of its superior fairness score, we believe that this model is worth deploying over the better performing one. By being based on a stratified sample, we can also conclude that this model will be applicable to a population that is roughly equal in males and females, which we believe to be a very plausible population, while having similar performances on those two categories.

Our accuracy for our final model was 0.756, meaning our model correctly classified whether or not the client defaulted 75.6% of the time. The ROC-AUC of our final model was 0.597. Our recall was 0.276, meaning the model predicts 27.6% of applicants who struggle to pay off their loans as so. Our f1-score was 17.1% which is a balance of precision and recall. Given that only 8% of the data had people who struggled to pay off their loans on time, these scores indicate that the model may provide some help in classifying future applicants. Having reviewed the key performance metrics of our model, it's clear that our model demonstrates a reasonable capability in classifying loan defaults.

We'll explore how some of these factors contribute to our predictions and which of them stand out as the most influential in determining whether an applicant may struggle to pay off their loans.

We chose to analyze the most influential variables from our model since our final model contains a large list of predictors after dummification. We found that occupation type (specifically whether or not they were a manager), age, whether or not they completed higher education, and whether or not they completed secondary education were particularly influential in our model.

A sample of some of our most influential model coefficients can be found below:

Table 2. Model Coefficients Sample

Occupation Type: Managers	Age	Higher Education	Secondary Education
-0.299	-0.297	-0.238	0.233

It is important to note that negative coefficients in this scenario indicate trends towards a client that will pay their loan back on time and without complications. So, when controlling for other variables, a manager is more likely to pay off their loan without any issues. Also, older people tend to pay off their loans more reliably than younger people. People with a higher education status tend to be more reliable, and people with a secondary education status tend to be less reliable.

Project Takeaways

We developed a machine-learning model to aid in recommending customers for loans. It is designed to assist, rather than take over, in the decision process. The features of the data provide insight into the likelihood of loan repayment. However, it is crucial to consider each application individually due to customer-to-customer variation. We also listed a set of critical predictors that influence the probability of loan repayment. Brokers can use these insights to scrutinize these aspects more closely in the application process.

While doing this, we took into account fairness. We believe it is essential for the client to understand the value of fairness. It ensures that the loan approval process is equitable across different genders. This also prevents unintentional model “discrimination” due to category imbalances in the provided training data. With the deployment of our model, we encourage training for mortgage brokers to understand how to interpret and use the model’s predictions effectively. This ensures that the model is used as intended and adds value to the decision-making process.

This model can have broader applications in the credit industry, in addition to mortgage loan repayment prediction. This model can become a general tool to aid in financial risk-taking endeavors.