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

Predictive Analytics for Loan Repayment

Leveraging Home Credit Default Risk Data

6th November 2023

ABSTRACT

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 aiding lending decisions and inventory management in both financial and retail sectors.

DATA OVERVIEW

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 the main table, which includes static data for three hundred thousand applications where one row represents one loan in our data sample.

In order to make this dataset more manageable, there are several alterations we have made. Firstly, we will only be using a subset of the columns that we deemed useful to our analysis. The columns from the original application_train.csv dataset we are considering are whether or not the

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

borrower repaid the loan, the borrower's total income, the borrower's credit, the amount of annuity, the amount of goods the borrower owns, the borrower's age, the days the borrower has been employed, the borrower's gender, the borrower's highest education level, the number of children the borrower has, whether or not the borrower owns a car, whether or not the borrower owns real estate, the contract type of the loan, the occupation type of the borrower, the borrower's housing type, and the borrower's marital status. For all the categorical variables, we will apply dummification as part of our feature engineering.

There are variables in this dataset that will impact the fairness of a model which include the borrower's gender, age, and marital status. We plan to use fairness metrics to analyze how well the model performs on the other prediction metrics to see if there is a conflict between fairness and prediction performance.

We took a preliminary look at loan repayment rates across education statuses. We see in Figure 1 that those with lower level education statuses repay their loans at higher rates than those with higher level statuses.

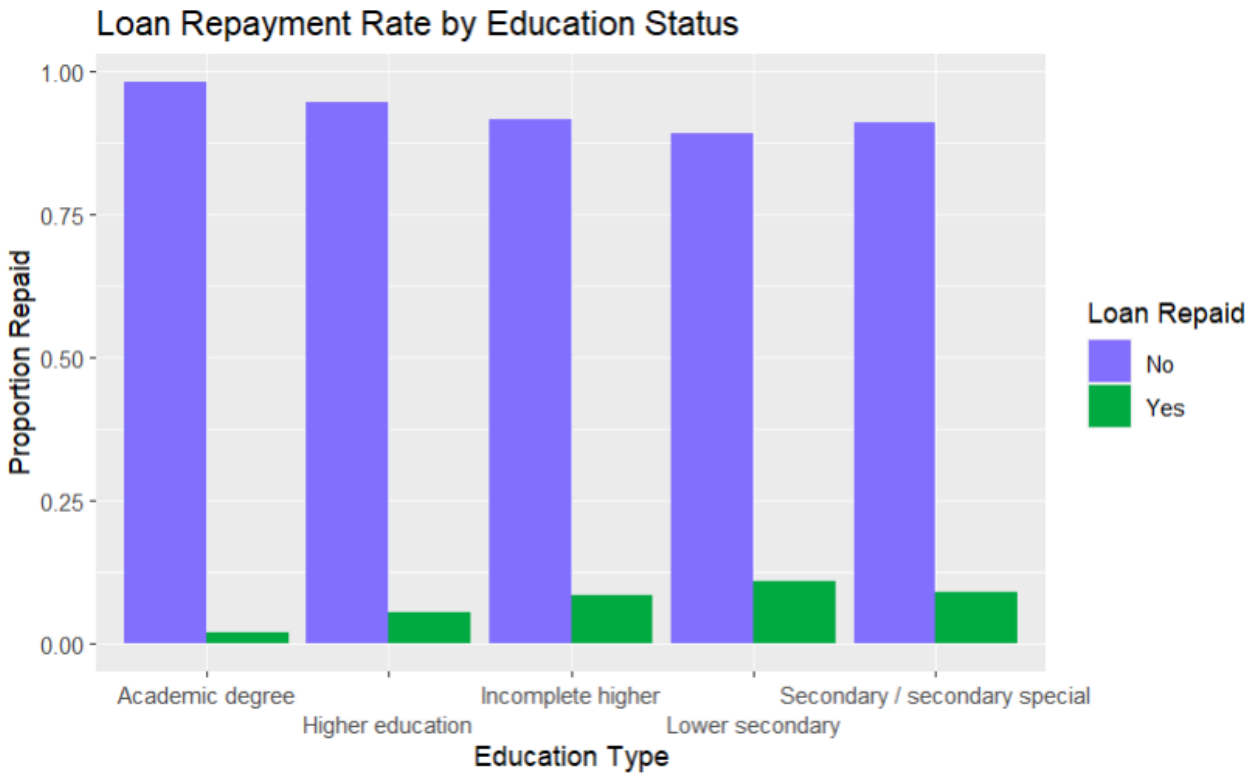

Figure 1: Loan Repayment by Education Status

We also took a preliminary look at loan repayment rates across car ownership status. We see that in Table 1, for those who did not pay back their loan, a majority of those people do not own a car.

	Loan Paid	No	Yes
Own Cars	No	60.3%	5.6%
	Yes	31.7%	2.5%

Table 1: Loan Repayment by Car Ownership

DESCRIPTION OF MODELS

We plan on using three main models to create predictions: logistic regression, support vector machines, and linear discriminant analysis. The utilization of these three classification models will yield optimal results on whether future customers will default on their loans. To validate our model, we will be splitting the data three different ways: (1) a completely randomly sampled split, (2) a stratified split, and (3) a split that is chosen in a non-random way using domain knowledge.

Logistic Regression

Logistic Regression is a statistical model used for binary classification tasks. Logistic regression models the probability of an observation belonging to the positive class. In this case, we will implement logistic regression to find the probability that a customer will default on their loan.

Support Vector Machines

Support Vector Machines (SVM) are a class of supervised used for classification and regression tasks. The algorithm finds the optimal hyperplane that best separates data points into different classes. The goal is to maximize the margin between classes while doing its best to minimize classification errors.

Linear Discriminants

Linear Discriminant Analysis (LDA) is a classifier as well as dimensionality reduction. It projects input data into a lower-dimensional space with low entropy to avoid overfitting to one particular class. We will use LDA to accomplish both tasks.

PROJECT OUTCOMES

In our final product, we aim to produce a classifier that predicts whether or not a loan should be given out. This model will use select features specified in our data overview section. The model shall allow the client to enter features for a loan borrower and return whether or not the loan shall be given out. It will estimate model parameters and validate results using them to return these predictions with a standard of ROC-AUC, accuracy, f1 score, and precision. Precision is important as we want to make sure to not accept applicants that are not able to pay.

The presentation of our findings will be characterized by visually informative charts, tables, and succinct summaries, ensuring accessibility to a wide range of stakeholders, from executives to the analytical team. The results will give the best insight as to what model exceeds most based on our metrics listed above.