Drawing the Line:
Classifying Risky Mortgage Applicants with Machine Learning

Consultants:
Alexander Arrieta - ajarriet@calpoly.edu
Joshua Blank - jwblank@calpoly.edu
Brendan Callender - bscallen@calpoly.edu
Sophia Chung - spchung@calpoly.edu
Martin Hsu - mshsu@calpoly.edu

The bank mortgage division requests a method by which to predict a mortgage loan applicant's ability to repay credit. We have taken a machine learning approach to fit a classification model on historical applicant-level loan repayment data to determine the risk level of new applicants defaulting on a loan. The classification methods that we explored were logistic regression, support vector machines, and linear discriminant analysis with the goal of creating a final predictive model that best predicts whether applicants will struggle to repay credit.

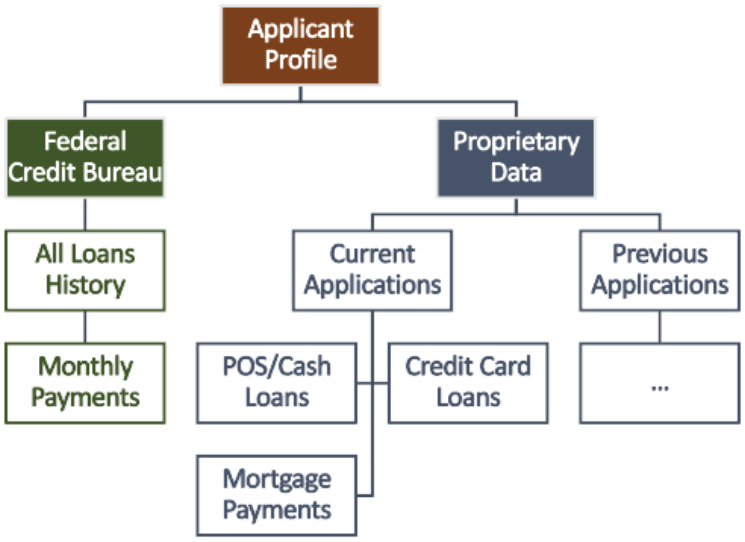
I. Introduction

In the mortgages division, accepting loan applicants who default on their payments burdens the bank financially, and in the worst case, may result in bankruptcy. Therefore, it is vital to develop a system that allows the bank to identify whether a new loan applicant may be at high risk for defaulting on their payments. To achieve this goal, we will be taking a machine learning approach to explore and fit three different types of classification models: logistic regression, support vector machine, and linear discriminant analysis. These models will be fit on historical loan application data. We will then choose our final model by fitting and testing each model using several different model validation metrics. Using the significant variables from our final model, we can then identify qualities frequently present in ideal loan candidates, and qualities that we may want to look out for as red flags.

II. Data Preparation

As a case study, the data we have used to train our delivered models comes from historical loan repayment data for individual applicants from the Home Credit Group, sourced from Kaggle[1]. The primary table contains information for past accepted loan applications. The data includes information about each applicant, the details of their loan application, and whether or not the applicant defaulted on the loan. The data also included additional tables relating to previous loans reported to the Credit Bureau for individual applicants as well as tables containing applicant payment history for home credit and credit card loans.

Figure 2.1: Home Credit/Kaggle Database Relational Structure



Feature Engineering

Certain aspects of the data relating to loan history were aggregated to get a better overall sense of the applicant's prior loan history. This included things like totaling how many late payments the applicant had made across all known prior loans. We did this to simplify how we look at an applicant and make our model more interpretable when studying how it interacts with the applicant's prior loan history. It also allowed each observation in our data to be a single set of information known about an applicant, rather than have multiple observations be related to a single applicant. We believed this would improve the effectiveness of our model due to it being able to access all relevant information about an applicant at the moment of classification.

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

Joining Datasets

We were able to combine all these different data sources to create a more complete profile for each applicant to better predict their ability to repay credit. *Figure 2.2* below shows an example observation, a single loan applicant, from our final data set broken up by the different sources. The first section, in red, contains the applicant profile which details personal information about the applicant including their demographics, assets, education, and income. This also includes the response variable in our model, whether or not they defaulted on their loan. The second section, in blue, includes proprietary data regarding previous loan applications from the Home Credit Group. The third section, in green, indicates variables regarding information on previous loans, obtained from the Federal Credit Bureau. In our example, we can see that the applicant is a male high school graduate who owns a car, makes over $200k a year, and has a strong credit history.

Figure 2.2 Final Training Data

Prediction Target / Applicant Profile

Application ID	Defaulted?	Education	Gender	Owns Car	Income	...	Credit Amount
100002	1	Secondary	M	1 (Yes)	$202,500	...	$406,598

Proprietary Mortgage History

Prev Approved Loans	Prev Canceled Loans	No. Late Payments	...	No. Completed Contracts
1	0	0	...	0

Federal Credit Bureau Loan History

Mortgage Total Days Overdue	Other Loan Total Days Overdue	Mortgage Total Amount Overdue	Other Loan Total Amount Overdue
0	0	0	0

III. Modeling Process

Because our final training dataset was so comprehensive, the modeling process first began by attempting to find which subset of variables in the data were most important for determining whether an applicant will default on a loan. Additionally, we also needed to determine which classification model would be used in the final model of the three proposed models. We were able to address both of these tasks together by running several different subsets of variables for each classification method.

The different subsets of variables were determined by picking groups of variables to cover important attributes for each applicant when determining their risk of struggling to repay credit. The different facets covered by each variable subset include information for

an applicant's education, assets, job, income, history of past payments, and history of previous loan applications. We used five different variable subsets on the three different classification methods leading to fifteen different model specifications. We were then able to test and compare the results for each of these models by evaluating different metrics using cross-validation.

Sampling and Cross-Validation

To prepare the data for the training and testing process, we first investigated the distribution of the target variable in the data in advance of the splitting process of cross-validation. We discovered that approximately 10% of the data had difficulty repaying credit whereas 90% had no problems. Due to the small proportion of applications which resulted in the applicant struggling to repay credit, it became important to ensure these observations were equally spread across the training and testing data. Since differing distributions of the target variable in the training and testing data may affect accuracy, we applied different levels of oversampling of the target variable to promote more balanced predictions, in addition to examining a simple random sampling approach. Additionally, to ensure fairness, we examined stratified sampling approaches on both the target variable and gender. For our final training and testing approach, we opted for stratification upon the target variable, whether or not an applicant defaulted on their loan. Each training and testing set's target distribution is therefore representative of the complete population in the dataset.

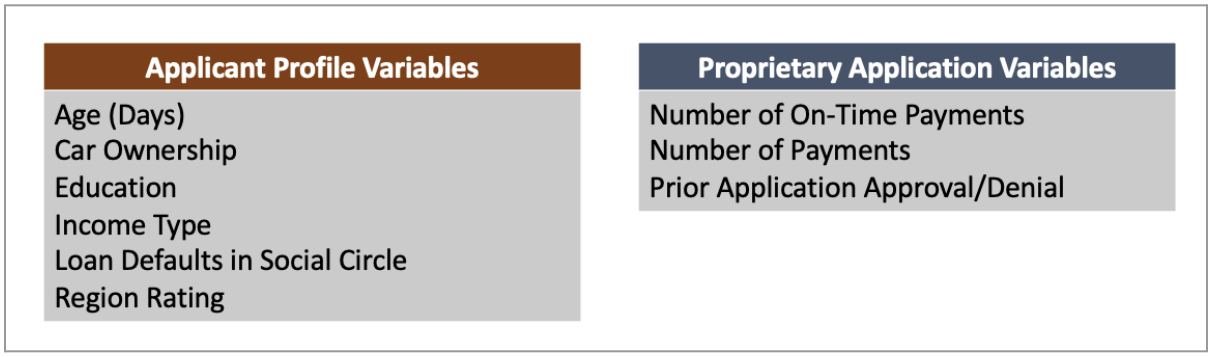
Model Selection Metrics

When determining the effectiveness of each classification method and variable set, we used the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) Score, which tests the overall effectiveness of a model by testing a model's predictive power across varying threshold values. Using ROC-AUC allowed us to first tune the variable set and model specification before choosing a desired threshold for the final model. Our final model was chosen by picking the model specification which produced the largest ROC-AUC score using cross validation.

IV. Final Model

Our final predictive model used Linear Discriminant Analysis with variables shown in *Figure 3.1* below. Additionally, we adjusted the threshold for determining high risk applicants to 0.2 from the standard value of 0.5 to increase the predictive accuracy of our final model.

Figure 3.1 Final Model Variable Set



V. Model Performance

Model performance metrics were determined by using cross validation to train and test our model on different sets of data to measure the predictive power of our model for future applications. When creating training and testing splits, we used stratified sampling on the target variable to ensure an equal distribution of our target variable across the training and testing splits. Our final model metrics were calculated on the dataset containing roughly a 90/10 split of the target variable.

Accuracy = 0.61
Our final model was able to correctly predict whether or not an applicant would have difficulty repaying credit for 61% of observations in the validation set.

Precision = 0.47
Of the applicants we predicted to have difficulty repaying credit, 47% of these individuals actually had difficulty repaying credit.

Recall = 0.62
Of the applicants who actually had difficulty repaying credit, our model correctly predicted 62% of these individuals to have difficulty repaying credit

F1-Score = 0.54
F1-score is a metric which combines precision and recall. Since our F1-score is close to both our precision and recall, our model is effective at balancing the precision and recall to produce balanced model predictions.

ROC-AUC = 0.599

On the unseen application testing dataset, our model was able to produce a ROC-AUC of 0.599 indicating our model is able to predict whether an applicant will struggle to repay credit more effectively than using random guesses.

While these metrics are clear in their interpretation, it can be difficult to gauge what represents good values for each metric. So as a baseline, we can compare the observed metrics of our model to the metrics of a model which randomly guesses whether someone will struggle to repay credit. If we were to use random classification, we would expect an accuracy of 0.50, a precision of 0.10, a recall of 0.50, a F1-score of 0.16, and a ROC-AUC of 0.50. Comparing the metrics of our final model to each of these, we see that our model produces predictions leading to more desirable results for each of the five metrics. Therefore, our model is effective at gaining insight into what makes an applicant more or less likely to default leading to more accurate predictions.

VI. Model Fairness

According to the Fair Housing and Equal Credit Opportunity Acts, it is illegal to discriminate according to age, gender, race, and marital status when approving mortgage applications. As a result, we opted to examine the fairness of our models under one of these features, gender. We strived to produce a model that would not reflect a discriminatory bias, and would correctly determine default risk for both male and female applicants.

We evaluated our model on its ability to make equally correct predictions regardless of the gender of the applicant. Women were 1.01 times more likely than men to be correctly predicted to default on their loans, while women were 0.95 times less likely to be correctly predicted to not default. These values are both very close to 1 and indicate that the effectiveness of our model predictions is almost the same for men and women indicating fair model predictions across gender.

Other Ethical Considerations

When taking a data driven approach to solve problems, it can be easy to get lost in the mathematics and methods, and to disregard the ethical implications of a machine learning model. The data set we used to fit our model used some predictor variables from demographic data, so it is important to ensure that when we apply the results from our model to make decisions, we do not discriminate against disadvantaged individuals. For example, our initial model flagged maternity as a key predictor for defaulting on a loan payment. As one can imagine, intentionally rejecting applicants who are providing

childcare is not only immoral, but may place the mortgages division at risk of legal action.

As a result, we have preemptively removed significant variables that may impact fairness, but we encourage the client to critically evaluate variables in the model for ethicality and legality, and take steps to remove them from both the model and the decision making process accordingly.

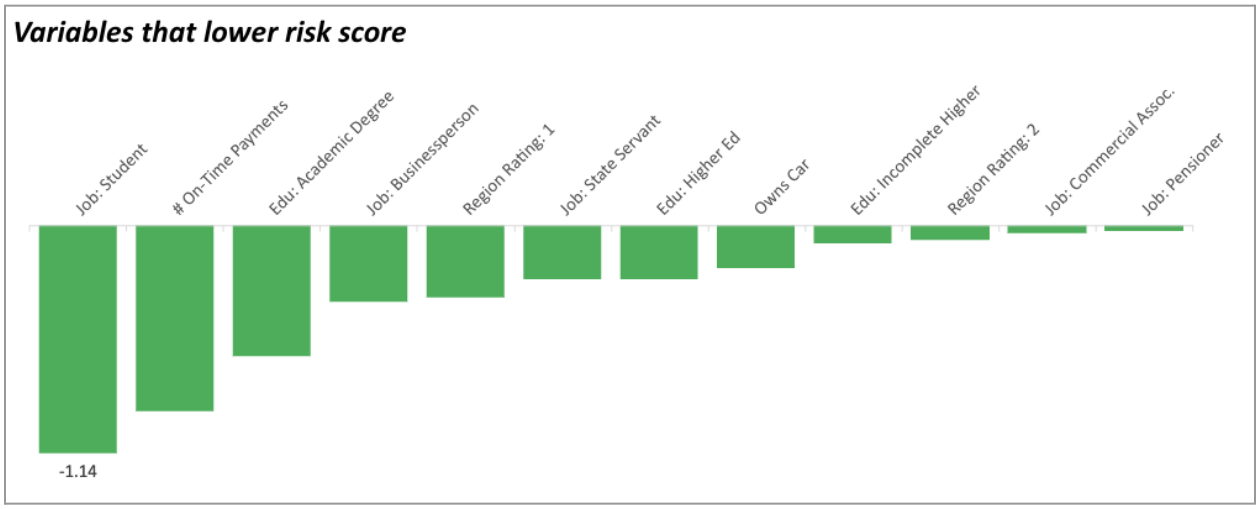
VII. Project Takeaways

Because Linear Discriminant Analysis is a linear classifier, we can examine the relationships between the values of model features, and increases or decreases in the resulting score. This can be accomplished by examining the model coefficients.

Factors Decreasing Risk

As seen in Figure 7.1, the feature leading to the greatest decrease in the model's risk score output is an indicator showing if the applicant's occupation is "student." In fact, many of these desirable features are related to post-secondary education and employment. Overall, we can conclude that having completed or being in the process of completing higher education, as well as being actively employed in a management or government capacity can positively impact risk by lowering it. Some other factors that decrease risk by decreasing the model score include region rating, having a high number of on-time payments in an applicant's credit history, and fully owning assets such as cars.

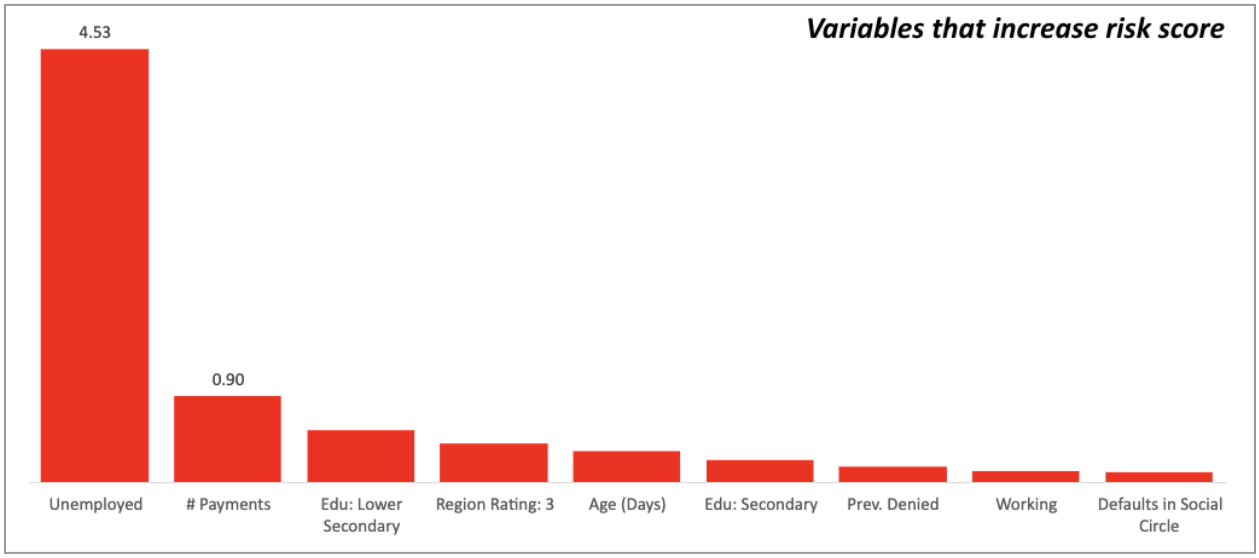
Figure 7.1: Model Features with Negative Coefficients



Factors Increasing Risk

As seen in Figure 7.2, the feature that has the greatest impact overall on the model, and negatively impacts risk by increasing it the most, is unemployment.

Figure 7.2: Model Features with Positive Coefficients



Recommendations

We believe that our model can provide valuable insight into what makes an applicant high risk and low risk. Using our model can both quantify and help prioritize which aspects of an applicant's profile and credit history should be first examined. For instance, when seeking for indicators of an applicant with low risk of default that is a good candidate for loan approval, it may be wise to focus on the applicant's educational history. Completing or being in the process of completing a post-secondary education may be a good indicator that an applicant is low risk. Furthermore, if an applicant is employed in business, management, or government roles, this may also be a good indicator of a candidate for approval. Applicants should be examined for timely payments and asset ownership as well.

On the other hand, when seeking red flags that may lead to a rejection decision, we first and foremost recommend examining if the applicant is unemployed. Taking a high number of payments to repay loans, not yet completing a high-school education, having a high number of post-30-day defaults in an applicant's social circle, and having a high number of previously denied applications may also be specific features to underscore in the decision making process that should lead to application rejection.

Conclusion

While we stand by the accuracy and fairness of our risk classification model, we would like to emphasize that this model should not be used on its own. It instead should be used to guide and introduce systematic insight to a rigorous, multi-level decision making process.

Many of the insights provided by this model are intuitive and may even be obvious to mortgage professionals. However, we believe our model can provide value by guiding the decision making process through systematic and rigorous quantification of risk. Furthermore, we hope that our model's focus on fairness leads to a more equitable outcome for mortgage applicants.