# Mortgages Division Proposal: Credit Applicant Risk Prediction

Consultants:
Alexander Arrieta - ajarriet@calpoly.edu
Joshua Walter 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 propose a machine learning approach fitting a classification model to determine the risk level of loan applicants defaulting on a loan. The model will be developed and fit upon historical applicant-level loan repayment data. Furthermore, the variables included in the model will reflect factors that we determine to be key identifiers of an applicant's predicted risk classification, allowing for simple and efficient interpretation and planning.

**Background**

One of the most critical and proactive ways to mitigate the risk associated with applicants defaulting on loans is to create a rigorous process with which to determine who is qualified for a mortgage loan. Properly classifying risk for repaying loans and using them to make discretionary evaluations of applicants can avoid loan repayment packages that lead to higher rates of default. HIgh rates of defaulting are dangerous and, such as in the case of subprime loans in the 2008 housing crisis, lead to critical declines in banking confidence[1].

We believe that a compelling solution to this problem can be addressed by fitting a machine learning model that can quantify or classify the likelihood of a loan recipient paying their loans in a timely manner.

[1] https://www.investopedia.com/articles/economics/09/subprime-market-2008.asp

**Data Collection**

As a case study, the data that will be used to train our delivered models comes from historical loan repayment data for individual applicants from Kaggle[2]. The main table included in the data that we plan to analyze is:

> ***Loan Application Dataset*** - A table containing information for individual loan applications. The table contains information relating to accepted loan applications such as the credit amount of the loan, and a column indicating whether the applicant had difficulties paying off the loan or not. The dataset also includes personal information about each applicant. Some variables of interest that we plan to draw from include client demographics, education, income, assets, provided documentation, residence, and social connections. An exploration of distributions of select variables of interest as they relate to repayment risk can be found in Appendix A.

With the remaining tables, we plan to investigate and iterate between two approaches. The first approach involves combining the application data table with federal credit bureau data. Some variables of interest we plan to use from the federal credit bureau data include those regarding federal credit principal amount and repayment information. The second approach involves combining the application data table with the applicant's previous loan history. Some variables of interest we plan to use from previous loan and credit history include those regarding repayment, credit amount, down payment, and contract status. We will leverage the additional data in each case, both separately and combined, to draw more information about each applicant and increase our prediction accuracy.

**Modeling Process**

We will inspect three different model specifications, evaluating the performance, assumptions, and interpretability of each type of model. The goal of each model is to provide a score for each applicant. This score relates to the risk of an applicant having difficulties paying back a loan. This score will then be evaluated against a threshold. Applicants that meet or do not meet the threshold will be classified into a high or low risk category accordingly. The models we plan to fit to the data include the following:

1. ***Logistic Regression*** - Logistic regression will produce probabilities relating to applicant payment difficulties based on the input variables. This is interpreted into an actual decision based on a threshold probability. Individuals whose probability is below the threshold will be categorized as a low-risk loan candidate.

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

2. ***Support Vector Machine (SVM)*** - SVM models segment the data into groups. It maximizes the distance between low-risk and high risk applicant groups. Any new individuals will be processed to see which group they fall in.
3. ***Linear Discriminant Analysis (LDA)*** - LDA segments the data by inspecting the distribution differences between low-risk and high-risk individuals when trying to classify a new individual application. The model then decides which category the new application is more likely to belong to, and assigns the individual to that group.

To decide on the best model to accurately assess payment difficulty risk , we will analyze several metrics that summarize the model's predictive ability. This includes metrics such as the accuracy, precision, and scores that evaluate a balanced combination of the two such as F1 score and ROC-AUC. We will apply the selection process to each model type separately first to determine the best version of each model type before applying the process across all three model types.

**Project Outcomes**

In our final product, we aim to develop a model that predicts whether an applicant will be able to repay their mortgage loan based on both applicant profiles and historical repayment data. Through experimenting with various test, training, and validation splits along with various linear classification algorithms, our models shall use a subset of features to determine an applicant's predicted risk category. Our final model will be selected based on accuracy, precision, and fairness.

The presentation of our findings will be characterized by engaging and informative visuals as well as direct action points that stakeholders of all backgrounds can easily understand. To be as transparent with the client, we will incorporate high-level summaries of our data processes that uncover our final product.

## Appendix A: Exploratory Data Analysis

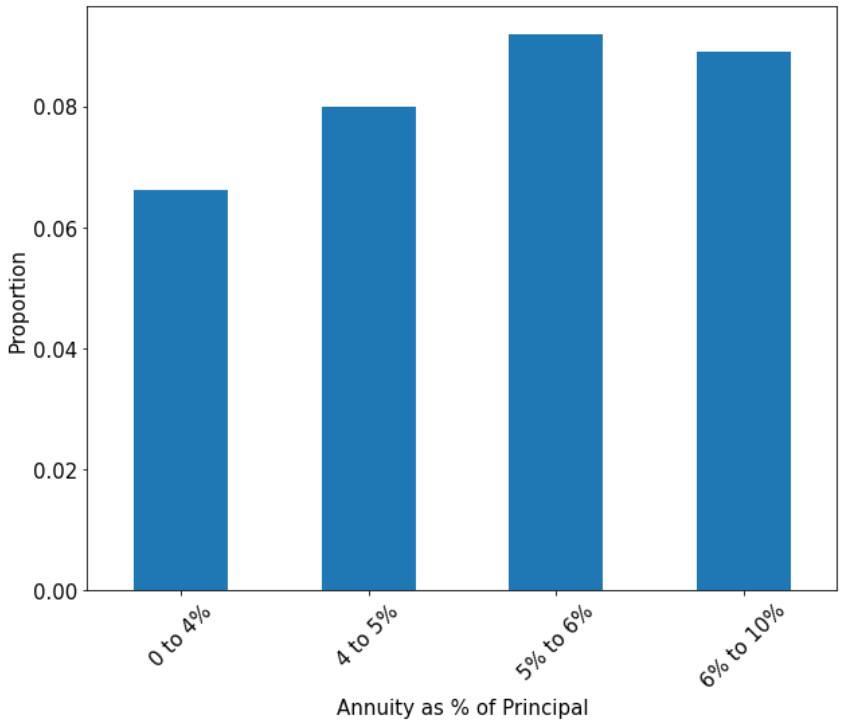


*Figure A.1: Proportion of Applicants Defaulted, by Annuity Payment Percent Quartile*

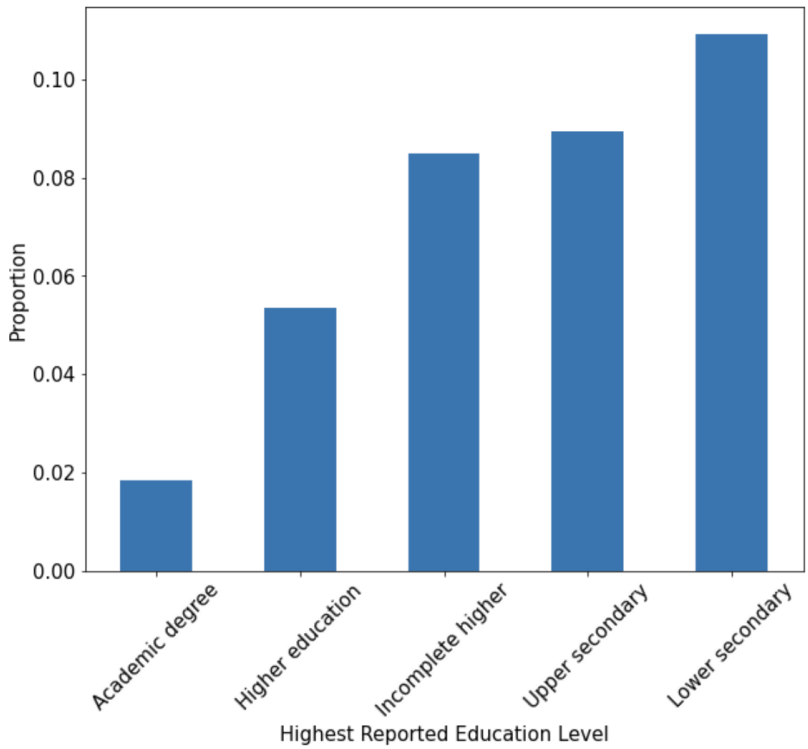


*Figure A.2: Proportion of Applicants Defaulted, by Education Level*

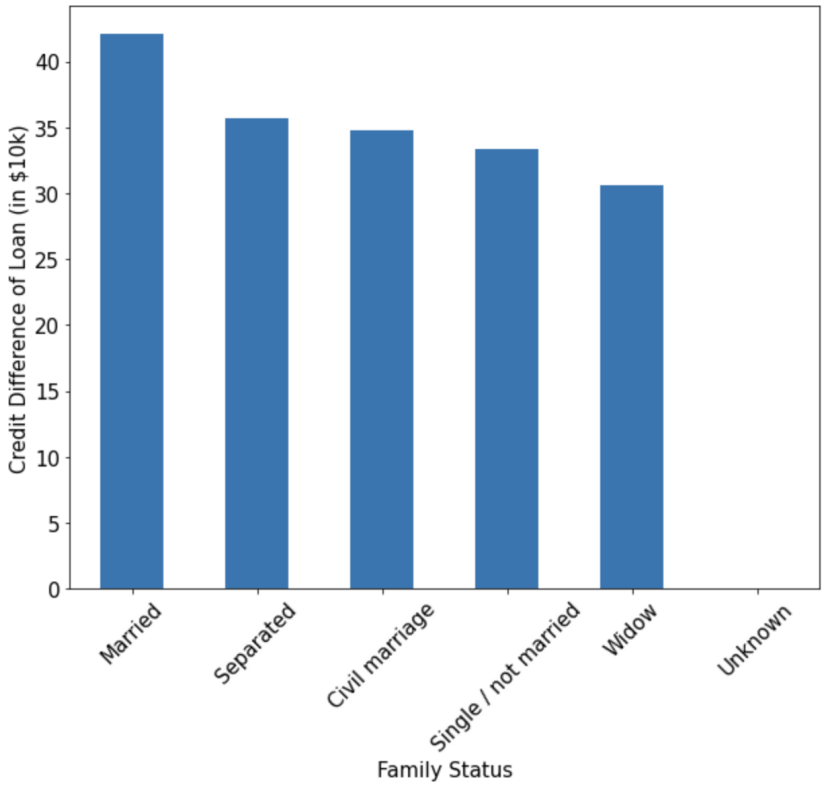


*Figure A.3: Average Credit Difference of Loan if Priors (in $10k), by Family Status*