Under the microscope: an investigation of Predictive Factors of Loan Acceptance

Hampton Leonard, hll4ce, Andrew Pomykalski, ajp5sb,
Tianye Song, ts7fx, Tyler Worthington, tjw4ry

Summary:

A thought of many Americans today is that if they want to purchase a car or buy a house, they can just go to a loan provider, request a loan, and get that amount right away. But sadly, several hundreds of thousands of people in America cannot and do not get the loan amount they requested. The goal of our project is to investigate what factors are significant in whether or not a person receives a loan for the amount they requested. We will use a large data set outlining several factors collected on people trying to take out a loan. Through this analysis, we will attempt to determine factors that do and do not contribute to the success or failure of receiving a loan of a certain amount. We will be working through several different analyses to determine a model that is accurate and reliable for continuous iteration.

Problem:

When a person applies for a loan, they are required to provide the loan provider with several pieces of information. These pieces are but are not limited to, amount of loan requested, length of employment, credit history, loan purpose, etc. However, many loan applications are rejected, and there can often be little understanding of the metrics behind these rejections and their influence. It would be beneficial to all loan applicants to have a robust understanding of what factors increase their chance of loan application acceptance, as well as have a method to predict the likelihood of their acceptance. Our hypothesis is that employment, debt history, location, and loan request information is related to whether or not a person's loan application is accepted. We will evaluate this hypothesis using logistic regression models and predict which of these factors are significant and important.

Data:

A company known as Lending Club Corporation provides a thorough data set of loan applicants from the first 2 quarters of 2015 and the first two quarters of 2016. Each of the datasets has hundreds of thousands of observations for applicants who were either accepted or denied for a loan. We assert that these two data sets should give us an accurate and representative depiction of the loan environment for our analysis.

The possible data cleaning that will need to be done with this data set will be dealing with missing values in an efficient way that doesn't skew results too much. We will also need to categorize some of our variables due to their format or the number of levels they contain. We will also need to standardize the information in our datasets so we are able to concatenate the two together. For example, those who received loans have their current debt balance and income but those who were rejected just have debt to income ratios. Finally we will need to decide how to incorporate the time variable, beginning of the month vs. end of the month, month 1 vs. month 2, or quarter 1 vs. quarter 2.

Goals:

For our project, we are searching for what factors lead to a success and failure of a loan application. We will look at all the factors provided to us and choose only important and meaningful predictors. By eliminating variables, we should be able to help loan applicants know their chance of getting a loan in advance of requesting said loan. If a person is above our thresholds in say 8 out of 12 predictors, then there is a pretty good chance of that person receiving a loan of some amount. However, if a person is above the thresholds for only 3 out of 12, there is a good chance that person should not be expecting a loan. With this knowledge, we hope to better educate the public about their probability of getting a loan based on their statistics.

Procedure:

First we will begin by cleaning and structuring the data in a more usable format. Next we will look to assign values of whether an applicant was accepted or rejected for their application. After assigning these values, we will combine both the accepted and rejected data sets into one data set that can be used for modeling. We will input all variables into a logistic regression model, and through the process of best subset variable selection, we will be able to determine which variables to keep in the model. Then we plan to apply F-test on our model to identify variable significance. Finally we will use 2016 Q2 and Q1 data to test our model.

Conclusion:

Understanding what leads to a loan being approved and being able to use these factors to predict whether the person is qualified for a loan amount they desire is a valuable piece of knowledge we hope to be able to extract and share.