Predicting Loan Defaults with Logistic Regression

Josh Jarvey

6/5/2020

## 1.0 Executive Summary

Products within the banking industry are all typically based on a risk assessment, that is to say each offering comes with some sort of risk of “winning” or “losing” on that transaction. For instance, providing investment services to a consumer has an associated risk of increasing or decreasing in value based on numerous factors. When it comes to loans for home, auto, school, personal, etc. these products are no different - they have some associated risk of “paid in full” (with interest collected) or the loan will “default” - the win vs. loss metaphor mentioned earlier.

Historically, the ability to assess risk on loans was something more of an art than a science. Financial bankers would meet with consumers in person, and while it’s true that they would go over various details of finance and stability, the ultimate decision was grounded in intuition and “gut-feel” of that banker (in accordance to bank guidelines of course). Some were very successful at closing the right ones and making a profit for their organization, while others were not as successful. There’s also the ethical dilemma of bias being introduced into gut-feel and intuitive decision making, which can lead to missed revenues, but also potential legal troubles that organizations are wise to mitigate.

In today’s age, the science has caught up with the art - and in many cases has even surpassed it. Where you have had one or two individuals trying to “crunch the numbers” and come up with a decision in the times past, we now have computers to do this work. But where the computers and science really excel at this work is finding the hidden relationships between a consumer’s relevant “parameters” to the loan and aiding in the decision of win vs. loss. It crunches the numbers among thousands upon thousands of historical loan’s parameters, and understanding those outcomes it can provide a sound assessment of risk that’s rooted in science and not intuition.

In this project we will explore the use of one of these scientific tools called Regression. Being that our end goal is to have a decision to “give loan” or “decline loan”, we will use the Logistic version of the regression analysis, which is better suited for a yes/no decision. A more scientific approach to decision making will not only save time and money to improve the bottom line, but will also help to mitigate bias and the associated organizational risk that comes with it.

## 2.0 Introduction

In risk-based decision making, it can be hard for any one human to take into account all the necessary information and variables that go into making the *right* decisions that minimize losses and maximizes revenues. This is where Regression analysis can come into play, because it is designed specifically to tackle the analysis of many variables, at the same time, to predict an outcome. Since the end result of the product in this study is a successful financial loan, we know that the outcome will be one of two choices: either “paid in full” as a success, or “loan default” as a failure. Because of this binary in nature outcome, we will use the Logistic version of the Regression analysis, which is suited well for handling this type of task.

The remainder of this report will outline the steps taken to build, train, and test different candidate models to see which performs the best and will ultimately be released into production to serve its purpose of aiding in the loan decision making process.

As with any data project, we start out by summarizing the dataset of 50,000 observations both numerically and visually to explore it for anomalies - missing data, outliers, etc. - and determine how to best overcome these data quality issues. In the next step we’ll explore the dataset’s 30 unique variables. Some of these variables will be significant toward solving the problem, whereas some will not and will need to be removed from the model or lumped together in an “other” category. Once we’ve landed on the appropriate variables, we’ll take a look at the distribution of its values as data may need to be transformed depending on what is discovered.

In the second half of the project, we will describe fitting the model to the data and accessing its performance. From here we will test at optimizing for prediction accuracy in addition maximizing profit. Finally, we will end with a conclusion and summarize model performance, along with any room for enhancements.

## 3.0 Data Preparation and Cleaning

To get started in this step, first we will load the necessary packages we need in order to perform the data exploration and cleaning.

library(readr)  
library(dplyr)  
library(mice)

Now prepare our response variable based on the status column of the existing dataset. We will discard any “in progress” loans, and only keep those that are “complete” - these three status reflect completed loans, leaving 34,655 observations remaining.

loans = loans[loans$status %in% c("Fully Paid", "Charged Off", "Default"),]

Because our response variable must be binary in order to use logistic regression, we will condense the results into a 2-factor variable called “outcome”. We can now drop “status” from the dataset altogether.

loans = loans %>%  
 mutate(outcome = as.factor(if\_else(status == "Fully Paid","Good","Bad")))

In an effort to keep parsimony, we will a look at each of these variables in the dataset and evaluate if we expect them to add overall value to the model. To get us started, we can drop the “status” variable since we used it already to create our new response variable.

loans = subset(loans, select = -status)

Below are a list of the other variables we’ll remove from the dataset:

* “Employment”: A free form entry field that provides no consistency between titles. Riddled with short hand and spelling mistakes.