Capstone 1: Project Proposal
Matthew Gudorf
April 20th 2020

The financial loss associated with loans being charged-off is incredibly large. According to the FDIC the amount of loans and leases merely past due in 2018 is on the order of 95 billion dollars. The goal of this project is to reduce capital loss; in both a proactive and reactive manner. This project utilizes Lending Club loan data (https://www.kaggle.com/wendykan/lending-club-loan-data), which tracks information pertaining to the borrower as well as the loan itself. The proactive measure will be to find an accurate model which determines whether a loan should be provided to potential borrowers. This proactive nature of these predictions presents a natural metric for the model of loan outcome; namely, comparison of the model predictions versus future observations. Unfortunately, there is a large time delay due to the length of each loan. In this dataset all of the loans have a maturity date of either three or five years. This not only makes the accuracy critically important but also motivates the formulation of an additional model. That is, a model which can predict the amount of money to be recovered from charged off loans. This would allow for a more targeted pursuit of lost capital through collection agencies before deciding to sell the bad debt. The combination of these two models: prediction of the loan outcome and then prediction of the amount of recoverable money provides a two step process recommendation system for present and future decisions. The deliverables for this project are then the code notebooks wherein the data is cleaned and analyzed and the models are created, as well as a written report which describes the process which describes this process. This system would be of use to institutions which offer large quantities of general purpose loans.

The prediction method of loan outcome can be formulated as a classification problem, specifically a binary classification as good or bad. Creating an unbiased model with a specific dataset is hard to do and strict discipline is required to prevent data snooping (the process of accidentally creating a biased model). There are many ways in which to succumb to this pitfall, especially in the presence of time dependent data such as in this project. A subtle form of data snooping is tuning the hyperparameters repeatedly to get better performance on this dataset; this can create “better” results but also can have undesirable consequences for generalization.

The methodology and process for the creation of the target models are as follows. First, the data must be cleaned and processed. with actions such as determining how to handle missing values, imbalanced data, nearly collinear data, etc. Once the data is clean I will train the machine learning models on the data. Specifically, for the binary classification task a comparison between logistic regression and random forest classification will be made. Likewise, for the regression of the amount of money recovered from charged off loans a comparison between stochastic gradient descent and ridge regression will be performed. The predictions and efficiency of the binary classification can be measured with tools such as the plotting of the precision recall curve, the receiver operating characteristic, accuracy, etc.
For the regression of the continuous variable, the regression can be graded with numerical scores such as the mean squared error and explained variance.