# Project 3 - Report

*Phil Bailey*

*November 9, 2018*

## Overview

The goal of this project was to predict the probability of defaulting on a loan using data from Lending Club. The metric used to calculate accuracy was Log-Loss. Initially I tested several models including glm, glmnet, randomForest/ranger, and xgboost. The slow performance and various issues I encountered in some of the other models (due to size of this dataset) compared with xgboost resulted in my relying solely on xgboost for this project. It's performance and accuracy could not be matched - so I focused my efforts using this algorithm only.

## Technical Details

Pre-processing was done using the following methods:

- Impute all continuous variables using mean
- Impute all factor variables using mode
- Drop columns with a large number of factor levels. These included:
  - emp_title
  - title
  - zip_code
  - earliest_cr_line

After pre-processing, cross validation was used to find an optimal number of trees for xgboost. The test-logloss began leveling out early, but I set the number of trees a little higher at 100 to ensure optimal accuracy (while still retaining an acceptable runtime). All other xgboost parameters were left at default settings.

## System Specs & Runtime

System Specs: **Windows 10, 2.7GHz, 8GB Ram**

Processing time to complete the 3 xgboost models over 3 train/test splits: **35 minutes**

Programming Language: **R**

## Final Results

Table 1: Log-Loss

| Model | Test1 | Test2 | Test3 | Average |
|---|---|---|---|---|
| xgboost | 0.4427508 | 0.4441355 | 0.4428931 | 0.4432598 |