Logistic Regression With Lending Club

### Introduction

For the machine learning portion of this project, I elected to build a logistic regression model that could be used to predict the likelihood of a personal loan going into default. After tuning the threshold, this model will separate loans into two classes: low-risk and high-risk. We can then compare the average profit margin among the two classes of loans to see how avoiding high-risk loans impacts return on investment.

### Building the Model

During exploratory data analysis, I was able to find the variables that had the most significant impact on a loan's successful repayment. These will be the predictors that are used for the initial logistic regression model:

* Loan Amount
* Term
* Employment Length
* Homeownership Type
* Annual Income
* Loan Purpose
* Debt-to-income ratio

Prior to creating the model, the downsampled dataset of 65,000 rows will be split into a training dataset (with 60% of the data) and a test dataset (with the other 40%). The training dataset will then be used to create the model using the features mentioned above, at which point the coefficient and significance will be reviewed for each feature. Next, a Receiver Operator Characteristic (ROC) curve will be created to compare the true positive rate and false positive rate across all possible thresholds. After selecting a threshold, a confusion matrix will be created to test the the out-of-sample accuracy of the model.

Finally, in order to optimize the logistic regression model, features will be added and removed from the initial model. The process of creating an ROC curve and confusion matrix will be repeated in order to determine whether an improvement in accuracy was observed.

Once the logistic regression modelling is complete, we will be able to compare the return on investment from high and low-risk loans by predicting payout. This will use the following logic:

* If loan defaulted, payout = total payment received to date.
* If loan did not default, payout = monthly installment x loan term.

Once the expected payout for each loan is determined, the sum will be taken and divided by the sum of loan amounts for each group of loans to calculate return on investment.