Applied Machine Learning Project Proposal
Team 4: William Gu, Angad Nandwani, Jiaxi Zhou, Yue Wang, Yuxin Qian

Problem Statement

Financial institutions generally face a typical challenge to decide whether they should lend money or approve credit cards of an applicant while mitigating risks at the same time. Lending Club is a large lending institution where registrants are allowed to do peer-to-peer lending with lower interest rates and faster processing compared to local banks. Though Lending Club provides an easier way for borrowers, potential risks also originate from some borrowers who refuse to repay their loans, therefore causing credit loss or financial loss to lenders and the company in the end. These applicants are considered 'risky'.

Problem statement: The ultimate goal of this project is to help the company, Lending club, reduce the amount of credit loss through deep analysis of historical data and discovery of strong driving factors that can be used to best assess the risk level of applicants and identify risky applicants.

Datasets

The Lending Club dataset consists of data for ~890k loans issued between 2007-2015. Each loan is described by up to 75 features. The features include examples of many data types: both discrete and continuous numerical data of varying scales in features such as credit score and annual income; both ordered and unordered categorical data in features such as employment length and employment title; and time/date data in loan issuance date. The target variable to predict is loan status, which indicates whether an applicant has fully paid the loan, is in the process of paying it, or has defaulted on the loan.

ML Techniques

Since our response variable, loan status, is a categorical variable with two factor levels, we would apply classification algorithms and select the model that gives the best prediction result. We will start with basic models and adjust along the way. Our tentative roadmap looks like:

1) Exploratory analysis: This step serves to help us understand the structure of the data and observe underlying patterns.

2) Preprocessing: Numerical data will be scaled where appropriate. We will explore multiple methods, such as imputation or dropping, for handling missing data and encoding categorical data. Dates can be decomposed into season, month, and day of week.

3) Modeling: Models we would like to try include logistic regression, SVM, and boosted trees. We may also explore ensemble methods that utilize some combination of these models.

4) Cross-Validation & Hyperparameter Tuning: We will use k-fold cross validation because it usually gives a more stable result and apply a grid search method to tune our hyperparameters according to the score given by the validation set. This step will also incorporate some regularization techniques such as Lasso and Ridge.

5) Evaluation: As a classification problem, we plan to use metrics such as precision, recall, accuracy, F1 score, and AUC to evaluate the performance of our model and its ability to handle potential not-ideal situations.