# Capstone Proposal – Lending Club Loan Default Prediction

## Domain Background

Lending club is a leading P2P lending company, based out of San Francisco, California with Total disbursed loan amount of USD 47 Billion as of 31-Mar-2019. It offers loan trading on secondary market and allows borrowers to create unsecured personal loans and matches them to Investors, who can decide whether to invest in a particular set of loans based on the supplied information. This helps investors earn higher returns and borrowers' access much needed funding. Lending club provides historical loan information, which can be analysed using machine learning algorithms. This can help to identify portfolios that are likely to default thus helping the investor avoid them.

## Problem Statement

To Identify whether a borrower will default by applying machine learning methods on historical loan data provided by Lending Club. The idea is to help investors identify loans with high chances of default. Loan default prediction is a binary classification problem wherein default event can happen either during the life time of the loan when the borrower defaults on interest payments or it can happen at the end when the borrower defaults on principal payment.

A binary classification problem allows the flexibility to choose from multiple machine learning algorithms, and the usage of quantitative metrics such as accuracy, precision, f1 score, etc to rank classification results.

## Datasets and Inputs

The dataset is from LendingClub: (https://www.lendingclub.com/info/download-data.action).

There are 3 files: LoanStats, browseNotes and RejectStats. Together, these files contain complete loan data for all loans issued from 2007 through 2019 Q1. LoanStats tab consists of 152 columns in the dataset, browseNotes consists of 121 columns, while RejectStats consists of 10 columns. These files contain information on borrowers' credit history, their personal information (such as annual income, years of employment, zip-code, etc.), loan information (description, type, interest rate, grade, etc.), current loan status (Current, Late, Fully Paid, etc.) and latest credit and payment information. DataDictionary

## Solution Statement

The solution is to build a classification algorithm using Machine Learning that predicts whether a loan default or not. The results are to be measured using various metrics defined in Evaluation metrics section.

## Benchmark Model

1) A naïve benchmark model would be to identify most frequent occurrence of loan default vs no default and use that as a naïve model prediction
2) Compare model performance with respect to reference models using the mentioned Evaluation metrics

## Evaluation Metrics

Evaluation metrics consists of Confusion Matrix, Accuracy, Specificity, Sensitivity, Precision

## Project Design

Steps:

1) Data Pre-processing and Cleaning:
    a. Missing value/Null value treatment
    b. Encode Categorical variables
    c. Extract info from text
    d. Outlier handling
    e. Normalizing feature variables
2) Train Classifier:
    a. Split data into Train and Test set
    b. Perform Cross-validation
    c. Compare Training performance vis-à-vis different algorithms
3) Test model performance
    a. Compute test results
    b. Compare test results with Benchmark models

## Reference Models

1. http://www.wujiayu.me/assets/projects/loan-default-prediction-Jiayu-Wu.pdf
2. http://cs229.stanford.edu/proj2015/199_report.pdf