# CS109A Data Science Final project

# Milestone # 3 (27-July-2018)

**Lending Club**

Prepared by Group # 8

Naval Goel and Vinay Nihalani

## Project Background

Lending club is a peer-to-peer platform that matches people looking to borrow with people willing to lend.

## Objective

Our goal remains to build a strategy that will enable an investor to select and build a portfolio of loan notes, say 5 to 10, that maximizes the rate of return of the portfolio given a risk profile of the investor. We will continue to work to achieve this objective as outline in the plan below.

## Description of the Data

We located a file name “LoanStats3a.csv” on the lending club website and imported the data to ipython notebook. We found that there are 42,539 observations and 164 columns. The Lending Club website listed that the data was from 2007 to 2011.

The data for loan applications from individuals. The loan amount requested ranged from $ 500 to $ 35,000 for two terms: 36 months and 60 months. The borrowings were at different interest rates from a low of 5.4% to 24.6% depending on the risk profile of the borrower, here called note grade. Most of the loans requested are for the purpose of personal use: credit cards payment, car purchase, home improvement, medical bills, vacation, wedding etc. And there are about 2,000 applications for small business and renewal energy purposes. One interesting observation in the dataset is that there are about 15 individuals listing annual income of over $ 1 million but requesting loans for less than $ 25,000.

## Data Cleaning Here are the steps we have taken so far to clean up the data:

## First we reviewed all 164 columns of data from which would have to identify response variable or variables and predictor variables. We looked for those columns that had no data and removed them from the dataset. With this cleaning, we were left with 64 columns of data. A list of the column removed is given in Appendix A containing the ipython notebook.

## Second, we went through each of the 64 columns and reviewed columns with majority of the data missing. We found nine columns with more than 90% of data missing, except for one with about 63% data missing. We dropped these columns also and were left with 55 variables: either predictor or response variables.

## We also dropped about observations which did not have any values for some of the columns. Out of a total of 42,539 loan applications in the data set, we were left with 41,084 observations.

## Then, we looked at the correlation between these variables and found significant correlations between several variables. The correlation matrix is shown below.

## 

## Next, we reviewed our strategy to come up with a response variable. We decided to evaluate an investor’s rate of return on the amount funded for the loan.

## To calculate the return, we calculated the total net amount investor received from this investment. We calculate the total net payment received as follows

## The amount invested was given as amount committed to the loan

P

## The duration for this loan was calculated using the month when loan was funded and the month when the last payment was received

## With this information, we had the calculations for the response variable: rate of return

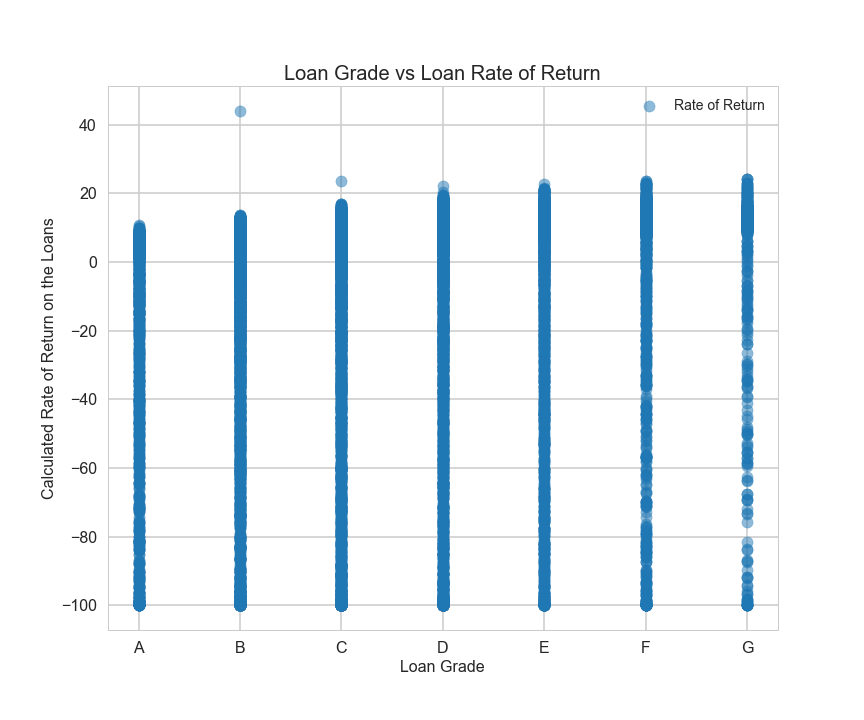
## Below is a plot of distribution of rate of return for all loans, 42,452. As we see, there are several loans where an investor lost all of the funding amount

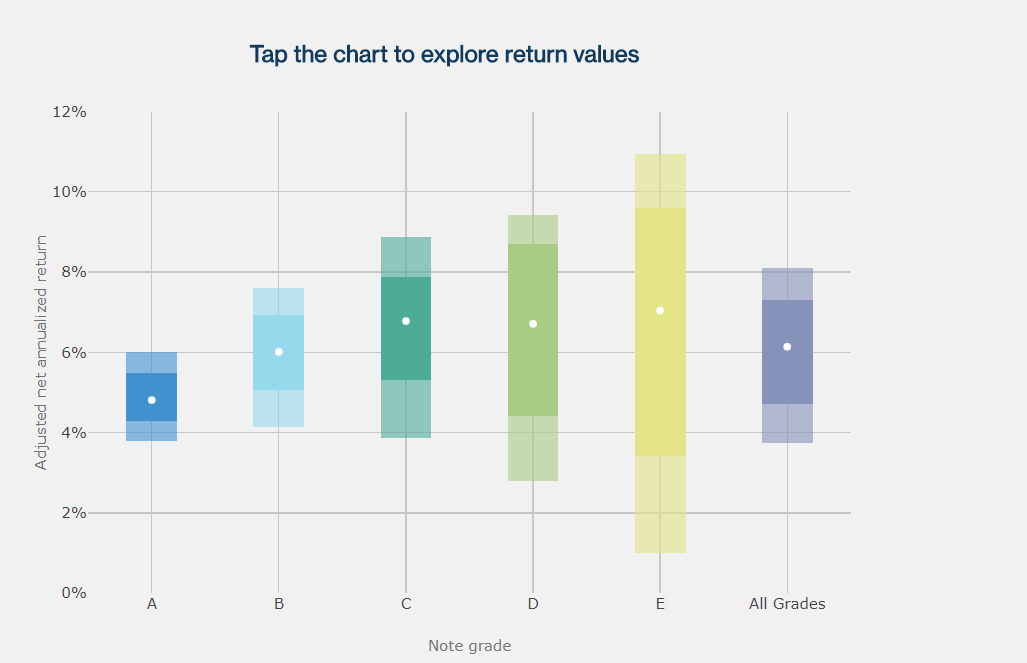
## 

For our initial simplified modeling for Milestone # 3, we assumed all of the payments: principal, interest, fees are received at the end of the period.

## Based on correlation matrix and our definition of the response variables, we reviewed the variables in the data. We decided on some variables as not relevant to our analysis. These included identification of borrower’s state, zip code and other columns that were descriptive, neither categorical nor numerical features. We were thus left with 27 predictor variables.

1. We also plotted calculated rate of return against the Grade assigned to the loan and found the trend similar to what is posted on Lending Club website. That is the rate of return increases as the Loan grade increases from “A” to “G” as shown below. Also shown is the plot from the lending club website. One thing different our plot and that of the lending club is that we show negative returns but the website does not show negative return. We know there are several borrowers who did not repay their loans so the data shown in the chart at the website is missing data of the delinquent borrowers.





## Baseline Model Summary Results

## 

## Project Outline

1. Extract, and visualize data on loan notes. Do a preliminary analysis to understand which variables seem to be dependent.
2. Clean data as necessary
3. Determine a methodology to accurately calculate a rate of return on individual loan notes based on loan amount, loan status, interest rate, service charges etc.
4. Group data by loan grade and calculate loan data statistics.
5. Compare results with rate of returns advertised by the site and explain potential reasons why results may be different.
6. Determine optimal number of significant predictor variables and build a model to predict rate of return after splitting the data in training set, validation and testing set.
7. Prepare report/presentations etc.

**Appendix A:**

**ipython Notebook for the Final Project Work So Far**