# Springboard Data Science Capstone Project – I Proposal

**By: Ramyata Mehrotra**

**Date: 12/20/2017**

The client for my first Capstone Project is called “Lending Club”, though any financial institution would benefit from the analysis I envision will be part of my project. I will be analyzing loan data and use multiple algorithms, which would estimate the probability that a given loan is good, neutral, or bad (i.e., will default) to briefly summarize it.

Lending has an inherent risk to it and the goal is to minimize that risk. Through this project, Data Science can assist loan analysts in answering the ultimate question of whom could a loan be given, and with whom would the bank need to be more cautious which would help Lending Club take necessary actions when processing loan applications.

I will be using the loan data for Lending Club that is available from Kaggle, through the following link: <https://www.kaggle.com/wendykan/lending-club-loan-data>

This problem can be approached in a number of ways. The current approach I envision is as follows:

1. Do some preliminary exploratory data analysis
   1. Understanding the structure of the data, such as data types of columns, and the number of observations.
   2. Ensuring there are no null values in the columns, which are being used.
   3. This would include doing some counts and running some comparisons, and understanding correlations, if any—in the context of exploring questions of interest, with the goal of documenting the “story” the data is telling, as I see it.
2. Finish cleansing the dataset to make sure it is ready for more detailed analysis as explained below.
3. Considering this is a supervised learning problem, I envision using Logistic Regression with regularization as a base model. More specifically, I currently envision the following in connection with applying this machine learning algorithm.
   1. Split the dataset into testing and training set.
   2. I would create a model which would then classify the loans in one of the two categories: “Good” , “Neutral” and “Bad”
   3. I will then fit the model on my training data making sure I use cross validation and estimate the best regularization parameter.
   4. Once I have the model trained on the training dataset, I will test it against the testing dataset and see how it performed by using the classification report available through scikit-learn
   5. It’s a good practice to analyze the results from the classification report for both test and training datasets to get a deeper insight into whether there is over-fitting, and to estimate how the model would generalize when using unseen data.
   6. Once I have a model ready, I can present my insights with the help of visualizations and draw conclusions.

My deliverables for this capstone project will include—as required:

1. Code to reproduce my analysis
2. A final report
3. A presentation slide deck