**Project Capstone 1**

**Lending Club Loan Data: Analyze Lending Club Issued Loan’s**



LendingClub is one of the world’s largest online marketplace that connects borrowers and investors. Lending Club have reinvented the way people access credit. LendingClub makes credit more affordable and investing more rewarding. They strive to provide a better financial future for everyone by building a marketplace that keeps costs low and investor opportunity high.

**Goal: Analyze LendingClub data to predict whether a borrower will default**

**Problem:**

LendingClub performs underwriting in a way that is different from a traditional bank. LendingClub screens the borrowers through the information that they have provided and evaluates the loan decision based on that information. LendingClub then assigns a rating to the loan and determines the interest rate on the loan. LendingClub then puts the loan on the marketplace where investors are able to evaluate the loan and invest. The main problem that we are trying to solve is to improve the credit scoring system of LendingClub using machine learning algorithms to predict whether a borrower will default on the loan. It is essential for LendingClub to access the validity of a loan to assign an accurate rating so that investors are provided with an accurate estimation. Otherwise, investors would be unlikely to invest on the platform.

**Dataset:**

LendingClub provides the ability to download the data from their website which includes complete loan data for all loans issued and latest payment information. However, the same dataset is available in Kaggle which was previously used for a competition that I am planning to use. The data contains loan data for all loans issued through 2007-2015, including the current loan status and latest payment information. It also contains present information on loans issues in last quarter. Aside from LendingClub data, it also contains various other features like credit scores, number of credit inquiries, address, total collections ever owed etc. The csv file contains 890,000 observations with 75 features. Additionally, a data dictionary has been provided to understand the features.

**Approach**:

Our goal is to predict whether the borrower will default on the loan or not. Given that our data is labeled, and our predictions will be placed in various categories. We will be tackling a classification problem.

Exploratory Data Analysis

The data provided in this dataset is fairly clean and formatted, however, we will need to deal with null and empty values. We will also need to look for outliers or transform categorical values. Dates would need to converted into datetime objects as well. The next step would be to do some early data exploration analysis as well data visualizations to understand what a bad borrower looks like and some commonalities shared by them. Furthermore, it will also help to understand how the data is distributed.

Feature Selection

With more than 75 features, it is not necessary to use all of them. Feature selection would need to be done to discover important features that are indicative of a borrower paying or not.

The final steps would involve building a model and evaluation of the models which will be covered in detail in the next report.