**Data Wrangling Report**

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



The Lending Club dataset contains 74 features and 887378 instances in a csv file. The target variable is loan\_status. The two goals of the project are to be able to predict whether or not an individual will default and also to understand the features of a person that defaults.

The dataset is derived from Lending Club a subset of which was posted on Kaggle for competitions. The dataset itself is relatively clean and has good quality thus, external sources will not be required to achieve the objectives.

The first step in my process of cleaning the dataset was find missing points and then drop those points from the dataset to not affect my analysis with bias. Looking at the data there were 15 features that contained a significant amount of missing data. If the feature is missing more than 60% of the data, the feature was removed as it did not hold any value towards the model.

The next step was to look at the NaN’s. Some features contained about 0.5%-2% NaN values and since we have a dataset of 887378 instances, we can remove the instances containing NaNs without it affecting our model and losing any information.

Our target feature is a categorical variable that contains 8 categories [Current, Fully Paid, Charged Off, Late (31-120 days), In grace period, Late (16-30 days), Default, Issued]. We are only interested in Fully Paid vs Default/Charged Off because our objective is to predict whether or not an individual will default. In order to do so we need to compare individuals that have fully paid the loan vs people that have defaulted. Therefore, we remove all rows not related to our target objective.

The next objective is to look at the datatype of the remaining features. We have two datatypes in our datatype: float32 and object. Few features contained just comments and text data that could be converted into usable form using NLP but not needed for the scope of the project to increase the accuracy by 1-2% and thus can be removed. Rest of the categories can be converted using one hot encoding including our target variable so that the data is in a usable format.

Finally, the last thing was to look at outliers in the continuous data. Building a scatter plot of the default data with loan amount and interest rate. It was easy to see that most annual amount of for the defaulted category are in the range of 0-40k however a single instance of 120k stands out which is an outlier and can add noise in the data. Outlier detection is tricky in a credit risk model because outliers contain data that is needed to predict whether or not an individual will default.