Capstone 2 Project Proposal: LendingTree Loan

Step 1: Problem Identification

I'm interested in getting a personal loan to pay off some educational expenses! LendingTree, an online middleman, connects consumers with multiple lenders for loans varying purposes (mortgages, auto loans, or personal loans). One of the perks of LendingTree is that it centralizes one application from a consumer to multiple lenders in LendingTree's network, versus having a consumer apply to each individual bank or financial institution for a loan. I am looking to see if it makes financial sense to take on a loan from LendingTree, however I don't know what I would get for an interest rate. The kaggle data set has data on multiple observations with varying backgrounds with the following features: credit_policy, purpose (of loan), interest rate, installments, log of annual income, debt to income ratio, fico credit score, days with a credit line, revolving balances, revolving utilization rates, inquiries in the last 6 months, delinquencies in the past 2 years, public record, any non fully paid debts. Utilizing the various features available, I plan to find which features have the biggest impact on deciding the interest rate and build a model to determine what I could expect as an interest rate.

Step 2: Wrangling

Review the dataset, ensure that each variable/feature is an individual column and each row is an individual observation.

Step 3: EDA

Examine for any skews in the data, make sure that each feature has values that make sense, i.e. credit score is between 300-850, income looks normal, etc. Get the shape of the data, min, max values of features and total number of observations. Look for missing values, outliers and correlations between features. Generate a heat map to show the correlation relationship between features.

Step 4: Pre-processing Data & Training Data

Fill in missing values where needed with median or mean value. If data is very skewed for particular features from "Step 3: EDA", will need to scale for a zero mean. Utilize cross validation technique to determine how to divide dataset into a train and test split for linear regression model. Generate a linear and random forest regressor model and examine the Mean absolute error, standard deviation and other metrics to see which one is better for our scenario.

Step 5: Modeling

Utilize the better model for predicting interest rate on my personal scenario to determine my interest rate for a personal loan. Create plots to see if the generated value for interest rate makes sense in comparison to other data. Lastly, see if there are any features where I could improve to lower my potential interest rate.

Step 6: Documentation

Generate a report and slide deck to summarize my findings for stakeholders(myself and others interested in utilizing the models I built).