Abstract

The goal of this project is to use classification models to predict the delinquency rate of peer-to-peer ("P2P") personal loans and possible return on investment. The data was directly from LendingClub, which was one of the pioneers and largest P2P platforms. The company allowed retail investors to directly invest in funds that the company would use to originate personal loans for its clients. As of December 31, 2020, the company no longer offers its retail platform as it tries to pivot to a more traditional banking business model. However, the P2P marketplace is still large and growing internationally and this dataset would be helpful in analyzing investment opportunities with other companies.

Design

The data set was provided on Mendeley Data on November 29, 2021. It is an aggregate of macro and micro features for both the specific financial product (revolvers, lines of credits, personal loans, etc.) and borrower credit metrics. The data covers a 12-year period from 2008 – 2019 on loans for all US states.

Data

The raw dataset contains over 2MM observations across 195 features. For this analysis, features pertaining specifically to loan products and borrower credit metrics were pulled. For example, annual income, loan interest rate, loan amount, home ownership status, and employment length.

Algorithms

Feature Engineering

- Converted categorical and string features to binary numerical variables
- Filtered out specific observations that showed high unknown, for example non-verified income status.
- Categorized the target feature into *Default* and *Current* loan status which represented delinquent and non-delinquent loans, respectively. Grouped other categorical observations in the target feature into one of the two above options.

Model Selection

The dataset was split into an 60/20/20 train vs. validation vs. test. The training data was then fitted onto four models and its results validated by the validation set. Finally, the testing set was used to predict outcomes and determine the quality of the final model and make investment returns.

Logistic Regression, XGBoost, kNN, and Random Forest classifiers were used before settling on XGBoost as the predictive model. For due diligence and curiosity purposes, the other three models were also ran to see the investment quality of each. As predicted, XGBoost performed the best.

Tools

The project primarily utilized pandas and Numpy for data handling and wrangling. Matplotlib was utilized to visualize datasets for better decision making. SKlearn was used for splitting the data and modeling it. Tableau was used to create visualizations in the final presentation.

Communication

A presentation was done to communicate the end results and findings.