University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Predicting Default for Loan Applications Business Brief

Purpose:

The purpose of this project is to utilize data preprocessing, machine learning, and model evaluation skills to correctly predict if customers are high risk at not paying back a loan using Python. This helps the financial institutions to minimize their losses by improving their application process.

Background:

Since the Great Recession of 2008, which resulted in loan default of many financial institutions around the world, the importance of building reliable and well-grounded credit scoring has been increased. Also, the prediction of loan defaults of applicants has become essential due to it being crucial to the profitability of banks. Given the desirable nature of the loan market, a challenging problem for any loan provider is to find suitable loan applicants who are likely to repay the loan.

According to Home Credit Group, "many people struggle to get loans due to insufficient or non-existent credit histories" (2018) which result in some being taken advantage by fraudulent lenders. To avoid this situation from ever occurring, Home Credit Group set up a Kaggle competition for data scientists "to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience" (2018). This way, those without credit histories can make a safe loan that will be able to be paid off.

Current Situation:

Currently, Home Credit Group is "using various statistical and machine learning methods to make these predictions" (2018) but need extra help at uncovering all that the data has to offer for making this type of prediction. With the data set being highly imbalanced (8.1% of the customers will likely not be able to pay off a loan), multiple rebalancing techniques were used like SMOTE and Tomek Links to better increase prediction power. Also, dummy variables were introduced for the categorical columns within the data set plus other data wrangling processes before models for prediction were built.

Conclusion:

This project was a test on how well exploratory data analysis can be used in aid of feature extraction and data imputation. There was also an importance of balancing the data without losing too much information within the dataset. The metric for how well each model performed was with the area under the ROC curve, as suggested by Home Credit Group. For the company, it was better that we were sensitive at predicting if a customer will not be able to pay a loan back vs predicting if they will pay a loan back. This was to protect borrowers as well as to protect the company so that both might not lose money. Random Forest was our predictive model choice achieving an area under the ROC curve of about 0.67.

Reference:

Home Credit Group. (2018, May). *Home Credit Default Risk: Can you predict how capable each applicant is of repaying a loan?* Kaggle. https://www.kaggle.com/c/home-credit-default-risk/rules