Bank Loan Defaulter Prediction

Authors: Yahya, Faisal, Ahmad

This data set was collected from kaggle.com website which contains information of funded amount, location, loan, balance, and interest rate. This dataset contains over 67,000 rows and 35 columns.

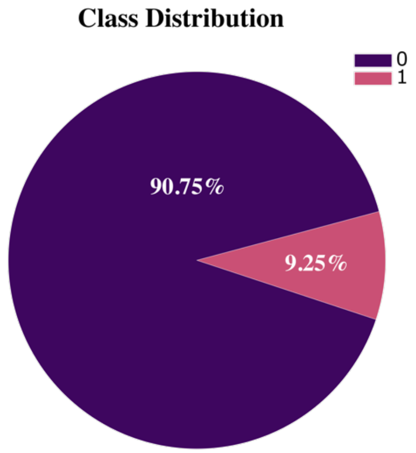

Fig. 1 A plot of imbalanced class distribution

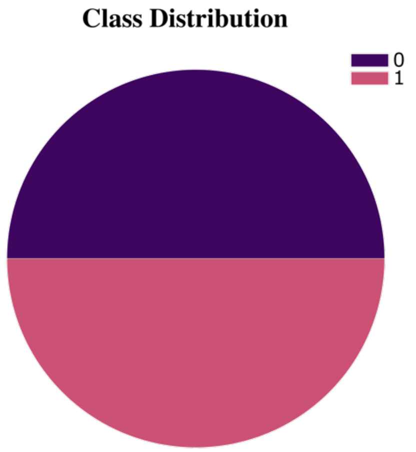

Fig. 2 A plot of a balanced class distribution after applying SMOTE method

From exploratory data analysis process, we noticed that the dataset itself has a severe imbalance in the class distribution as shown in figure 1. To solve this problem, we took a Synthetic Minority Oversampling Technique as an approach to generate new data that balance class distribution. After we apply this method, the dataset has a balance class distribution.

Classification:

We implement a logistic regression model in this dataset to make a prediction of whether a customer could be a loan defaulter or not. The confusion matrix determines the accuracy of this model which is around 51.58%. To improve the accuracy, we try to increase the number of features which helps to improve the score to reach almost 55%. However, when we applied k-nearest neighbors algorithm, the model's accuracy increased from 82% to 97%. Thus, this algorithm is more suitable for this dataset than logistic regression. Next step would be implementing random forest and XGboost algorithms to see if we can make a prediction with a high accuracy.