Bank Loan Defaulter Prediction

Authors: Yahya, Faisal, Ahmad

Abstract:

The goal of this project is to classify a bank customer who is a loan defaulter from who is not. Due to the nature of this problem, the dataset has a severe class imbalance. After exploring the dataset, we found that only 9% of customers are loan defaulters. The challenge of dealing with such an imbalanced dataset in machine learning field is that mostly will be ignored by AI models which result in poor performance on the minority class even though the focus not on the dominant classes like in our dataset. One way to approach this type of problem is to oversample the minority class. This method can be achieved through synthesising artificial data from existing examples without adding any new information to the model. This technique known as Synthetic Minority Oversampling Technique (SMOTE).

Data Description:

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 features.

Algorithms:

- Feature engineering:
  - Converting categorical variables to binary dummy variables
  - Handling imbalanced classes

- Models
  - Decision Tree Classifier
  - Random Forest Classifier
  - XGBoost Classifier

- Model Evaluation and Selection

The entire dataset of (~67000) observations was split into 80/20 for training and testing. Through the use of cross validation function from sklearn library with 5 folds. We select Random forest classifier as the best model for this dataset as it has a precision score of 92% which much better performance the other models we evaluated.

Tools:

- For data manipulation: Numpy and Pandas
- For data visualization: Matplotlib and Seaborn
- For machine learning algorithm: Scikit-learn