Loan Prediction Analysis

Abstract:

The loan is one of the most important schemes of banks and it can be short term or long term, depending on the purpose of the loan. For example, if the loan is for buying a house, then most probably it will be a long term loan. To be more specific, short term loans are called such because of how quickly the loan needs to be paid off. In most cases, it must be paid off within two years. Any longer loan than that is considered a long term loan.

The goal of this project is to build multiple machine learning classification models and calculate the accuracy of each model, then perform a comparison between all built models to decide which one is the best. Therefore, the project will predict the type of loan whether it is short term or long term based on many features. This project, will help bankers to determine the type of loan term that the customer needs.

Data Description:

The data for this project will be read into a CSV file using SQL (Find the dataset on the following link https://www.kaggle.com/panamby/bank-loan-status-dataset/data). The obtained dataset consists of over 110,000 loan records with 18 features. Some of them may effect on our target (type of loan term) such as: the loan amount, the purpose of the loan, and the customer's annual income…etc.

The taget column has two classes (binary class) Short term 70.58% and Long term 29.42%.

After cleaning the data by dropping null rows and outliers, the new data set has 43957 rows. Then, the data has been splited into 80% train, 10% validation, 10% test.

Algorithms:

The classification algorithms that has been used in this project:

- Logistic Regression:
  - Logistic Regression
  - Logistic Regression Scaled
  - LogisticRegression (class weight {Long Term : 2 , Short term : 1})
  - LogisticRegression (class weight : balanced)

- Naive Bayes
    - Gaussian NB
    - Bernoulli NB
    - Multinomial NB

- K-Neatest Neighbors (3)
- Decision Tree
- Random Forest
- Extra Tree
- Ada Boost
- Stochastic Gradient Descent
- XGBoost

Best Algorithm was XGBoost with testing score 0.8812.

Tools:

- Python
- Numpy
- Pandas
- Sklearn
- Seaborn
- Matplotlib
- XGBoost
- Pickle