Group 18: Ji Qi(jq2365), Wen Song(ws2685), Unal Yigit Ozulku(uyo2000), Yizhi Liu(yl4993)

Project Deliverable #1 - Project Proposal (due 02/23/2023)

Backgrounds

In light of the changing consumer attitudes and preferences, personal loan has become an increasingly popular topic in recent years. Personal loan related topics can imply the business decisions of banks, playing an important role in the banking industry. This type of problem is related to various features, such as income, credit card usage, education and so on. The objective of this project is to utilize various machine learning techniques to figure out an optimal model that can predict whether a customer will take a personal loan or not (binary classification problem). The major steps that are potentially included are data preprocessing, exploratory data analysis, selecting appropriate features, and training and evaluating the models. This project will explore the relationships between different factors and customers' decision to take a personal loan. It will be a valuable exercise in applying machine learning techniques to a real-world banking problem by helping increase customer satisfaction, improve business performance, and reduce risks for the bank.

Dataset

The dataset from Kaggle includes 5000 observations with 13 features classified into two categories (numerical or categorical features). The numerical features contain 5 variables: age, experience, income, CC AVG(Avg. spending on credit cards per month) and mortgage. For categorical variables, there are 2 ordinal variables including family and education and 5 boolean variables including securities account, CD account, online banking and credit card. In addition, Zipcode and ID are also classified as categorical features, but the variable ID only acts as a customer identifier without adding any useful information to the dataset. Thus, we may drop this feature during the data cleaning process. Our target label is personal loan, indicating whether the potential customers have a higher probability of purchasing the loan based on their demographic characteristics. The '0' in the personal loan means the customers didn't accept the personal loan offered in the last campaign and '1' means the customers accept the personal loan.
Link to the dataset: https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling?resource=download

Analysis of Methods

Before applying any machine learning algorithms, it is important to explore the data and understand its characteristics. The dataset has 5,000 observations and 13 features, with the target variable being 'Personal Loan,' which has binary values of 1 for customers who accepted the loan and 0 for those who did not. The dataset is devoid of missing values, obviating the need for methods such as kNN and regression models to handle them. However, the dataset is imbalanced, with around 10% of customers accepting the personal loan offered in the last campaign. This means we will need to use techniques like oversampling or under-sampling to balance the dataset. Additionally, some features, like 'Zip Code,' may not be relevant for the classification task and can be removed or processed using related methods. We will also analyze the correlation of each feature. If some feature pairs are highly related, we will drop one from this pair. After exploring the dataset, we will evaluate several classification algorithms for this dataset. We plan to implement a series of machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, SVM (with different kernels), Ensemble learning, Gradient Boosting, XGBoost, neural network, and some advanced neural networks (e.g., LSTM). Cross-validation will be used to evaluate the performance of each model. We will use appropriate evaluation metrics such as F1-score, recall, precision, and AUC (area under the curve) well-suited for imbalanced datasets. After this, we will further analyze the feature importance of the dataset and use the findings to process the dataset and retrain the machine learning model. Finally, we will compare the results to determine if we can improve the classification performance of these models.