Group 3
IE 7300 20838

Project Proposal

Overview:

This project focuses on direct marketing campaigns conducted by a Portuguese banking institution. The marketing campaigns were executed via phone calls. In many scenarios, the same client was contacted multiple times to conclude whether the bank term deposit would be a “yes” or “no”.

Objective:

This is a classification problem, and the goal of this project is to determine whether a client will subscribe to a term deposit or not.

Data Description:

The dataset has 45211 instances and 17 attributes. The meaning of each attribute present in the dataset and its datatype is described below:

- Age-> *Numeric*
- Job->*Categorical*-> Type of job (admin, blue collar, entrepreneur, housemaid, etc)
- Marital Status: *Categorical*
- Education->*Categorical* (basic, high school, graduate, unknown, etc)
- Default->*Categorical*, Do the customers have credit in default (No,Yes,Unknown)
- Housing->*Categorical,* Do the customers have housing loan?
- Loan-> *Categorical,* Do the customers have personal loan?
- The other 11 attributes are related with last contact of current campaign.

The target variable of the dataset is in the form of yes/no and tells us whether the client has subscribed in the term deposit or not.

Plan of Action

1. Data exploration and preprocessing:
    - Check for missing values and outliers present in the dataset
    - Check for the distribution of each column
    - Appropriate preprocessing steps following dataset exploration
2. Modeling (including but not limited to):
    - Algorithm selection (Further details in below section)
    - Model building
    - Train and test the model
3. Analysis and Evaluation

Algorithm Selection

The first algorithm that will be implemented is Logistic Regression. Based on the accuracy score of the algorithm, other classification algorithms such as K-nearest neighbor and Decision Tree will be further explored. The Bias and Variance of executed algorithms will be examined, and errors of a learned classifier will be decomposed into the two terms. To measure this, many variants of the dataset will be approximated as $E_D\{h_D(x)\}$. Bias-Variance will be measured with Bootstrap sampling. Using these results, the model can be adjusted to improve its performance; if the Bias is high, the model will need additional complexity and if the variance is high the model will need some simplification.

Based on performance, other alternative models that can be implemented are:

- Random Forest Classifier
- Naïve Bayes
- Support Vector Machine
- Gradient Boosting

The selection of the algorithms may vary, and the focus of this project will be on getting the most optimal model (one which won't lead to overfitting or underfitting).

Group 3 Members:

- Ankita Goyal
- Simran Bhatia
- Venkata Sai Tarun Reddy Pongulaty
- Ahmed Alsaadi