DATA 622 Assignment

Bank Marketing

Susanna Wong
Professor Joseph Sabeljia
March 2, 2025

Introduction

The Bank Marketing dataset was obtained from UCI Machine Learning Repository. It contains 45,211 client data related to direct marketing campaign conducted by a Portuguese banking institution, which is carried out via phone calls. The client data contains 17 features that includes their demographics, i.e., age, education, job, and martial status, their financial status, i.e., bank balance in euros, personal and housing loan status, and credit default status, and campaign related variables. The target variable indicates whether the client subscribed to term deposit or not. For this assignment, we conducted an EDA on the dataset and analyzed for a suitable machine learning algorithm for the dataset.

Exploratory Analysis

Class Imbalance

Since the subscription rate ("yes" for term deposit) is 11.7%, there is class imbalance. Given this imbalance, proportions normalization the data within each group in necessary to make a fair comparison.

Missing Data and Outliers

About 83% of the clients has at least 1 “unknown” value in their categorical variables. The variable *poutcome* (previous outcome) has the highest percentage (81.7%) of “unknown” values, and the other variables, *job*, *education*, and *contact* have relatively low percentage of “unknown” value. Outliers are present in the data. For example, the variable, *contact*, show extreme outliers where some clients are contacted over 60 times during the campaign. Extreme outliers is present in non-subscribers’ balance.

Correlation

Most numerical variables present weak correlations. However, *pdays* and *previous* show a moderate positive correlation. Duration, campaign, and balance does not show any strong relationships with other variables.

Client Demographic and Financial Status

The age range of the client is between 18 and 95 with a median of 39. There is a higher subscription rate among younger clients (18-25), and older clients (65+), which can indicate that age can be a crucial factor in predicting term deposit subscription.

Most clients are married and have secondary education. The top 3 occupations of the clients are blue-collar workers, management professionals, and technician. The clients' balance ranges from -8,019 and 102,127 euros with a median of 448 euros. About 56% of the clients has housing loan and very few clients (16%) has personal loan. About 1.8% of the clients have credit default.

Subscription Insights

The clients who subscribed to term deposits share the following characteristics. Clients with tertiary education are more likely to subscribe, as well as students and retired clients. Additionally, single clients tend to have higher subscription rates than those in other martial status. Clients with housing loans also have a higher subscription rate.

Campaigns were primarily conducted in May (accounts about 30% of the campaign), July, and August. However, these months has the lowest subscription rates, which suggest an unsuccessful marketing campaign . Subscription rates peak in the following months: March, September, October, and December. Contact duration has the strongest relationship to term deposit subscription compared to the number of campaigns and previous contacts.

Algorithm Selection

Since there are labels (term deposits subscriber: yes or no) in our dataset, I would recommend to use supervised learning algorithm such as Random Forest. About 83% of the data contains at 1 “unknown” value. Random Forest can handle missing values without the step of imputation. The dataset contains only about 11%

subscribers which shows class imbalance. Random Forest is less sensitive to class imbalance. It uses ensemble learning, which reduces overfitting. However, it is computational slow to use as we have a large dataset. Alternatively, if there were fewer than 1,000 data records, we can use Naive Bases as it is simple and fast, and works well with categorical data. However, it may be sensitive to class imbalance.

Pre-Processing Steps

For optimal model performance, we need to perform pre-processing steps such as data cleaning. We need to address the "unknown value" in the categorical variables. The variable *poutcome* (previous outcome) has the highest percentage (81.7%) of "unknown" values. I would remove it as most clients regardless if they subscribed or not have very few previous contact. The number of previous contacts may not have much influence on subscription rates.

Due to the right skewness of all numerical variable except day, we would need to perform data transformation such as box cox transformation. Categorical variables will needed to be converted to numerical format using label encoding. Moreover, we should introduce new features such as age groups and contact duration (short, medium, long). Sampling techniques such as SMOTE should be use to address imbalance class.