Proposal Report

Team 7

1. Introduction

This report explains the process of predicting customers' subscription to term deposits based on given customer information, campaign data, and socio-economic indicators. Term deposits are one of the key financial products offered by banks, and predicting customers' financial behaviors plays a crucial role in enhancing the bank's marketing strategies and operational profits.

2. Data visualization

First of all, we attempted data visualization to quickly identify patterns and outliers in the data and to gain a deeper understanding of the data. By doing so, we can optimize the input variables of the model and improve predictive performance.

3. Data pre-processing

3-1. Missing values

We will measure the number of missing values for each variable and use different handling methods depending on the amount of missing data. If a variable has a large number of missing values, we will remove the variable. If the number of missing values is not significant, we will use a missing data prediction model to replace the missing values by leveraging the other features of the data.

3-2. Outliers

Box plot is a useful tool for visually identifying outliers in data. Therefore, we plan to use box plots to assess the presence of outliers and then use the mean or IQR to handle them.

3-3. Multicollinearity

We will visually inspect the correlation between numerical variables by plotting scatter plots. If there is a strong linear relationship between two variables, it indicates the possibility of multicollinearity. In the presence of multicollinearity, it becomes difficult to accurately assess the impact of each independent variable, leading to a potential degradation in the predictive performance of the model. To obtain reliable results, multicollinearity needs to be addressed. While one approach is to remove variables, we will also employ ridge and lasso regression methods to mitigate multicollinearity while preserving the data.

3-4. Normalization

Normalization includes standardization and normalization. Standardization adjusts the distribution of data by setting the mean to 0 and the standard deviation to 1, whereas normalization adjusts the values of data to a specific range. In the case of min-max normalization, data is adjusted to the [0,1] range. If the data distribution follows a normal distribution and outliers exist, standardization is adopted. If the data distribution is not normal, or if the model is nonlinear, or if the scales of the data vary, min-max normalization is adopted. In cases where the criteria are not clear, the method that improves the final F1 score will be adopted.

4. Modeling

4-1. Logistic Regression

Our goal is to predict whether a client will subscribe to a term deposit or not using a model. In other words, since we can classify the outcome into "yes" or "no", it's suitable for a binary classification problem, and logistic regression can be applied effectively for this purpose. Logistic regression is a simple yet effective model. Moreover, it's suitable for problems assuming linear relationships between variables. Therefore, if there are high correlations among independent variables, it will reduce one of the variables. Additionally, we plan to prevent overfitting and enhance the performance and stability of the model using regularization techniques such as Ridge or Lasso.

4-2. Decision Tree

There are some advantages to using Decision Tree to predict whether clients will subscribe to bank term deposit. Decision Tree is easy to understand visually, as they depict how data is split based on certain criteria. Additionally, they require minimal computational effort and can quickly make predictions. And in our dataset, we deal with both numerical and categorical data, and Decision Tree can handle both data. Moreover, Decision Tree splits based on information gain at each node, which means the model indicates which features are most influential in prediction, helping to identify key variables. In addition, Decision Tree is not significantly affected by missing or outliers. To prevent overfitting in decision trees, we limit the depth of the tree or perform pruning. Additionally, we handle outliers to prevent performance degradation.

4-3. additional methods

If there are models covered in the remaining lectures that are suitable to the problem, we plan to consider and apply them together.