Group Name: Data Adventurers

Name: Shatabdi Pal

Email: shatabdi@pdx.edu

Country: USA

College/Company: Portland State University

Specialization: Data Science

Problem description:

ABC Bank aims to optimize its marketing efforts to promote a new term deposit product by developing a predictive model. The model will analyze customer data from past interactions to forecast the likelihood of each customer subscribing to the term deposit. By accurately identifying potential buyers, the bank can focus its marketing resources on high-probability customers, enhancing efficiency and reducing costs. This project involves creating and evaluating models with and without the 'duration' feature, a critical element that will ensure practical applicability and performance.

GitHub Repo link: https://bitbucket.org/shatabdi_workpace1/bank-marketing/src/main/

Data Cleansing and Transformation Steps

This part of the project involves cleaning and transforming the dataset and handling NA values and outliers using multiple techniques.

Transformation:

Missing Value: The dataset does not have any null values. Categorical attributes like 'job,' 'marital,' 'education,' 'default,' 'housing,' and 'loan' have 'unknown' values, which can be considered missing. Categorical variables with 'unknown' values were imputed with the most frequent category. The random forest method was also applied to imputing those unknown values. Then, the KS (Kolmogorov-Smirnov) KS statistic and P value were used to compare the two imputation techniques. I found that the KS statistic was 0.0, and the p-value was 1.0 for each column with unknown values. This result suggests that both imputation methods (most frequent and Random Forest Classifier) produce imputed statistically indistinguishable values in their distribution. Therefore, these specific columns and the chosen imputation methods are equally effective in approximating the original data distribution.

Outliers: Certain numeric features like 'age,' 'duration,' 'campaign,' 'days,' and 'previous' exhibit outliers that could impact the model's performance. A comprehensive approach to detect outliers was employed to ensure our data's reliability. Box plots and statistical methods such as the Z-score and IQR (Interquartile Range) were used to identify outliers in numerical features. Then, we implemented the Z-score and IQR methods to deal with these outliers. The comparison of both methods using Cohen's Kappa coefficient, which resulted in 1.0, indicates that both methods are equally effective and reliable in identifying outliers for the specific set of data points or indices being compared.

Skewed Data: Features like 'duration,' 'campaign,' and 'previous' have a skewness greater than 1. A square root transformation is applied to normalize skewed features.

Binning of age features are also done in this part. Besides these three transformation processes, some redundant columns were removed after creating new features such as Total Contacts, Economic Indicators Ratio, and Interaction between features. Finally, the normalization of features was also done.