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.

Data understanding

The data is taken from the UCI machine learning data repository. The dataset has 21 features, including the label. It is related to a Portuguese banking institution's direct marketing campaigns (phone calls). The classification goal is to predict whether the client will subscribe to a term deposit (variable y). The features are both categorical and numerical.

Type of data

Numerical Datatype	Categorical Datatype
Age	Job
Duration (in seconds)	Marital status
Campaign (number of contacts performed	Education

during this campaign)	
Pdays (days since the client was last contacted from a previous campaign; 999 means the client was not previously contacted)	Default (has credit in default?)
Previous (number of contacts before this campaign for this client)	Housing (has housing loan?)
Employment variation rate (quarterly indicator)	Loan (has personal loan?)
Consumer price index (monthly indicator)	Contact (communication type)
Consumer confidence index (monthly indicator)	Month (last contact month of the year)
Euribor 3-month rate (daily indicator)	Day of the week (last contact day of the week)
Number of employees (quarterly indicator)	Poutcome (outcome of the previous marketing campaign)
	Y (target variable - has the client subscribed to a term deposit?)

.

Problems in the data

1. **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.

2. **Outliers:** Numeric features like 'age,' 'duration,' 'campaign,' 'days,' and 'previous' have outliers that could affect the model performance.

3. **Skewed Data:** Features like 'duration,' 'campaign,' and 'previous' have skewness greater than 1.

4. **Imbalanced Data**: The target variable 'y' is imbalanced, meaning there are significantly more 'no' than 'yes' responses.

Approaches to Handle Data Problems

Handling Missing Values:

Categorical variables with 'unknown' values can be imputed with the most frequent category.

Dealing with Outliers:

Outliers can be detected using box plots and statistical methods such as the Z-score and IQR (Interquartile Range). It can be treated by removing or capping and flooring them at a certain threshold to reduce their impact.

Addressing Skewed Data:

Square root transformation is applied to normalize skewed features. Binning the data into categories can also help in dealing with skewness.

Handling Imbalanced Data:

SMOTE (Synthetic Minority Over-sampling Technique) can be used. Ensemble methods like Random Forest and Gradient Boosting or algorithms robust to class imbalance, like ADA Boost, can also help.

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