DELIVERABLE WEEK 7

Group Name: Destined Data Team
Specialization: Data Science

Team Members:

1. Name: Praneetha Rajupalepu
Email: Pranitha.724@gmail.com
Country: Canada
Company: Modest Tree
Specialization: Data science

2. Name: Selaelo Ramokgopa
Email: sly.kholo@gmail.com
Country: South Africa
College: University of Johannesburg
Specialization: Chemical Engineering

3. Name: Surya Chandra
Email: ksuryachandra619@gmail.com
Country: Germany
College: Otto von Guericke University
Specialization: Electrical Engineering and Information Technology

Problem description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with the bank or other Financial Institution).

Data Understanding

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. Theclassification goal is to predict if the client will subscribe to a term deposit.

Data downloaded from: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data Table Description:

It consists of four different datasets, but we are using the bank-additional-full.csv dataset that has41188 rows and 21 columns. Each column datatypes are categorized in the table below.

Column names	**Data type**
Age, Balance, Campaign, Day, Duration, Pdays, Previous	Numerical
Job, Marital status, Education, Default, Housing, Loan, Contact, Month, Poutcome	Categorical
y	Binary

Data Analysis Questions:

1. **What type of data you have got for analysis?**
   A clean data and heavily skewed one, most variables have outliers too.

2. **What are the problems in the data (number of NAvalues, outliers, skewed, etc.)?**
   The data has no missing values, there is a certain number of Outliers in 'age', 'duration','campaign' etc., and most columns are skewed.

3. **What approaches you are trying to apply to your dataset to overcome problems like NA value, outlier, etc., and why?**

   a. **Solutions for NA values:**

- For numerical variable: We replace the missing values with the mean or median of thecolumn
- For categorical variable: We replace the missing values with the mode value, the most likelyvalue of the missing

   b. **Solutions for Outliers:**

- Remove outlier in the data, from all the columns because outliers differ significantly fromother observations and change the meaning of a data

   c. **Solutions for skewed:**

- A very skewed column represents a column that does have a normal distribution, it can beright-skewed or left-skewed.
- Asymmetrical distribution will have a skewness of "0".

  There are two types of Skewness:Positive and Negative.

  i. Positive Skewness (similar to our target variable distribution) means the tail on the right side of the distribution is longer and fatter. In positive Skewness, the mean and median will be greater than the mode similar to this dataset. Which means more houses were sold by less than the average price.

  ii. Negative Skewness means the tail on the left side of the distribution is longer and fatter. In negative Skewness, the mean and median will be less than the mode. Skewness differentiates in extreme values in one versus the other tail. Here is a picture to make more sense. You can remove it by performing numpy.log1p on the column

GitHub Repo link

https://github.com/PraneethaRajupalepu/Bank-DataScience-Project