DELIVERABLE WEEK 8

Group Name: The Powerpuff Girls

Specialization: Data Science

Team Members:

1. **Name: Chaithanya Shivakumar Ittamadu**
   Email: sichaithanya889@gmail.com

   Country: Ireland

   College: Dublin Business School

   Specialization: Data science

2. **Name: Memudu Alimatou Sadia Anike**

   Email: anikesadia01@gmail.com

   Country: Nigeria

   College: University of Ilorin

   Specialization: Data science

3. **Name:** Lakshmi Chandana Vupputuri

   Email: vupputuri.chandana@gmail.com

   Country: Ireland

   Specialization: Data Science

Problem description

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

Data Understanding

The data is related with direct marketing campaigns of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit.

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

It consists of four different datasets, but we are using the bank-additional-full.csv dataset that has 41188 rows and 21 columns. Each column datatypes is categorized in the table below.

Columns names	**Datatype**
Age, duration, campaign, pdays, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed	Numerical
Job, Marital, education, default, housing, loan, contact, month, day_of_week, poutcome	Categorical
y	Binary

What type of data you have got for analysis?

A clean data and heavily skewed one, most variables has outliers too.

What are the problems in the data (number of NA values, 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

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

Solutions for NA values:

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

Solutions for Ouliers:

- Remove outlier in the data, from all the column because outliers differ dignificantly from other observation and change the meaning of a data

Solutions for skewed:

- A very skewed column represent a column that does have a normal distribution, it can be right-skewed or left skewed
- A symmetrical distribution will have a skewness of "0". There are two types of Skewness: Positive and Negative. 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. 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/chai1122/VC