Week 8 Deliverable

Group Name: Data Go

Team member's details :

Name:Yuheng Chen

Email: yuh363@gmail.com

Country: USA

College/Company:N/A

Specialization: Data Science

Name: Terry Chou

Email: techch26@gmail.com

Country: USA

College/Company: N/A

Specialization: Data Science

Name: Rishi Aluri

Email: rishialuri@gmail.com

Country: UK

College/Company: N/A

Specialization: Data Science

Name: Justine Pile

Email: justypile@gmail.com

Country: US

College/Company: N/A

Specialization: Data Science

Problem description

The objective of this project is to develop a predictive model for ABC Bank to determine the likelihood of customers subscribing to their term deposit product. By analyzing customer interactions with the bank and other financial institutions, the machine learning model will identify potential clients who are more likely to purchase the product. This will enable the bank to optimize their marketing efforts by focusing resources on customers with a higher probability of conversion, thus enhancing campaign efficiency and reducing costs. The project will assess model performance with and without using the "duration" feature while also addressing any data imbalance through suitable techniques.

Data understanding

The key independent variables from these data are:
age, job, marital, education, default, balance, housing loan, contact, day, month, duration, campaign, pdays, previous, poutcome

- **age:** the age of the person
- **job:** the person's occupation (categorical: occupation name)
- **marital:** the person's marital status (categorical: 'married', 'single', 'divorced')
- **default:** whether or not the person has credit in default (yes/no)
- **balance:** average yearly balance (numeric)
- **housing loan:** whether or not the person has housing loan (yes/no)
- **contact:** communication type (categorical: 'unknown', 'cellular', 'telephone')
- **day:** last contact day of the month (numeric)
- **month:** last contact month of year (categorical: 'Jan', 'Feb', … , 'Dec')
- **duration:** last contact duration in seconds (numeric)
- **campaign:** number of contacts performed during this campaign and for this client
- **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric)
- **previous:** number of contacts performed before this campaign and for this client (numeric)
- **poutcome:** outcome of the previous marketing campaign (categorical: 'unknown', 'failure', 'other', 'success')

Columns for **bank-additional.csv** and **bank-additional-full.csv** only:

- **emp.var.rate:** employment variation rate (numeric)
- **cons.price.idx:** consumer price index (numeric)
- **cons.conf.idx:** consumer confidence index (numeric)
- **euribor3m:** 3 month Euribor interest rate (numeric)
- **nr.employed:** the number of employees (numeric)

Output variable:

- The output variable is column **y**, which is a binary data of either 'yes' or 'no'

What type of data you have got for analysis

- Integer
- Categorical
- Binary
- Date

What are the problems in the data (number of NA values, outliers , skewed etc)

- No duplicate rows (rows that are exactly the same across all columns) appears in **bank.csv**, **bank-full.csv**, and **bank-additional.csv**
- 12 duplicate rows appear in **bank-additional-full.csv**
- The age count graph (count vs age) is **right skewed** – the majority are between age 30 to 40. The number of people over 60 years old is very small
- The campaign count graph (count vs campaign) is right skewed – the higher the total number of campaigns, the smaller the total count.

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

- EDA to determine NA values or other data that may be an issue
- Duplicate data can be removed from the data set
- Right skewed data in numeric columns such as age, balance, duration, campaign, previous will be fixed to normal distribution using log-transformation.
- Outliers can be identified through EDA methods including:
    - Statistical methods
    - Visualizations
- Machine learning algorithms can also be used to predict missing values

Data Source: https://archive.ics.uci.edu/dataset/222/bank+marketing

Github Repo link

https://github.com/yuh39/Bank_Marketing_Group_Project/tree/main