Week #8 Deliverables

Team member details:

Group Name: Intern_Project

Sophonie Sidrac, sophiesidrac@gmail.com, USA, Tennessee State University, Data Science

Vijayarajan Vijaya Jothi, vijayajothi23s@gmail.com, UK, Data Glacier, Data Science

Yusuf Aisha

Bristy

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 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 marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The data set involves customer profiles showing their age, job, marital status, education, housing, loan, and more.

What type of data you have got for analysis?

- **age** (**integer**)
- **job** : type of job (**string**)
- **marital** : marital status (**string**)
- **education** (**string**)
- **default**: has credit in default? (**Boolean**)
- **housing**: has housing loan? (**Boolean**)
- **loan**: has personal loan? (**Boolean**)
- **contact**: contact communication type (**string**)
- **month**: last contact month of year (**string**)
- **day_of_week**: last contact day of the week (**string**)
- **duration**: last contact duration, in seconds (**numeric**).
- **campaign**: number of contacts performed during this campaign and for this client (**numeric**)
- **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 (**string**)
- **emp.var.rate**: employment variation rate - quarterly indicator (**numeric**)
- **cons.price.idx:** consumer price index - monthly indicator (**numeric**)
- **cons.conf.idx:** consumer confidence index - monthly indicator (**numeric**)
- **euribor3m**: euribor 3 month rate - daily indicator (**numeric**)
- **nr.employed:** number of employees - quarterly indicator (**numeric**)
- **y** - has the client subscribed a term deposit? (**Boolean**)

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

- The data is semicolon-delimited instead of comma-delimited.
- There are no null values in the data.
- The Booleans in the data are represented as y or n
- There is a skewed distribution of data which is causing the data to be imbalanced. This could make the model inaccurate.
- There are also outliers present in one of the feature columns.

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

- When reading csv file, (delimiter=';' and quotechar= ' " ') must be applied to reload the data with the correct delimiter.
- To resolve the boolean issue, the data needs to be mapped to binary data as '1' and '0'.
- To balance the data and resolve the skewness, transformative methods need to be applied to the data.
- Lastly, I could remove the outliers altogether, use the means of the columns or use kmeans technique to predict an estimated value.

Data intake Report:
Name: Bank Marketing Campaign (Data Science)
Report date: 10/26/2024
Internship Batch: LISUM37
Version:1.0
Data intake by: Sophonie Sidrac
Data intake reviewer: Sophonie Sidrac
Data storage location: https://github.com/1Sophani/DataGlacier-Internship/tree/main/Week%208

Tabular data details:

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	csv file
Size of the data	5.6 MB

Submitted by: Sophonie Sidrac
Submitted to: Data Glacier