Data Glacier

Your Deep Learning Partner

FA

Week 8 – Deliverables Document

Name: Fatimah Asiri
Username: alassirifatima@gmail.com
Batch code: LISUM14 30 Sep – 30 Dec 2022
Submission Date: 11/26/2022
Submitted to: Data Glacier.

This document contains basic information on the project

Data Science

Bank Marketing (Campaign)

Group Project

Prepared by:
Fatimah Asiri

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 in understanding whether a particular customer will buy its product or not (based on the customer's past interaction with the bank or other financial institutions).

Data understanding

Bank wants to use the ML model to shortlist customer whose chances of buying the product is more so that their marketing channels marketing SMS/email marketing, etc. can focus only on those customers whose chances of buying the product is more.

This will save resources and time (which is directly involved in the cost (of resource billing).

Develop a model with Duration and without duration features and report the performance of the model.

The duration feature is not recommended as this will be difficult to explain the result to the business and also it will be difficult for businesses to campaign based on duration.

What type of data you have got for analysis

Data Set Information:

The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Attribute Information:

Input variables: bank client data:

1. age (numeric)
2. job: type of job (categorical)

admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown

3. marital: marital status (categorical)

divorced, married, single, unknown note: 'divorced' means divorced or widowed)

4. education (categorical)

basic.4y, basic.6y, basic.9y, high school, illiterate, professional. course, university degree, unknown

5. default (categorical)

Has credit in default? No, yes, unknown

6. Housing (categorical)

Has a housing loan? no, yes, unknown

7. Loan (categorical)

Has a personal loan? no, yes, unknown - related with the last contact of the current campaign

8. contact (categorical)

contact communication type: cellular, telephone

9. month (categorical)

last contact month of the year Jan, Feb, Mar, ..., Nov, Dec

10. day_of_week (categorical)

last contact day of the week Mon, Tue, Wed, Thus, Fri

11. duration (numeric)

last contact duration, in seconds

Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

12. campaign (numeric)

number of contacts performed during this campaign and for this client (numeric, includes the last contact)

13. pdays (numeric)

number of days that passed by after the client was last contacted from a previous campaign; 999 means the client was not previously contacted)

14. previous (numeric)

number of contacts performed before this campaign and for this client

15. poutcome (categorical)

outcome of the previous marketing campaign: (failure, nonexistent, success) # social and economic context attributes

16. emp.var.rate (numeric)

employment variation rate - quarterly indicator

17. cons.price.idx (numeric)

consumer price index - monthly indicator

18. cons.conf.idx (numeric)

consumer confidence index - monthly indicator

19. euribor3m (numeric)

Euribor 3-month rate - daily indicator (numeric)

20. nr.employed (numeric)

number of employees - quarterly indicator

Output variable (desired target):

21. y - (binary)

has the client subscribed to a term deposit? (Yes, No)

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

1. Handling the Outliers.
2. Duplicate Rows.
3. Correlation.

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

1- To handle the null values if data categorical handling missing data by mode and if data numeric handling missing value by mean or median.
2- Dealing with Outliers

1. Deleting the values: I can delete the outliers if you know that the outliers are wrong or if the reason the outlier was created is never going to happen in the future.
2. Changing the values: We can also change the values in the cases when we know the reason for the outliers. Consider the previous example for measurement or instrument errors where we had 10 voltmeters out of which one voltmeter was faulty.
3. Data transformation: Data transformation is useful when we are dealing with highly skewed data sets. By transforming the variables, we can eliminate the outliers. This can also be done for data sets that do not have negative values.
4. Using different analysis methods: I could also use different statistical tests that are not as much impacted by the presence of outliers.
5. Valuing the outliers: In case there is a valid reason for the outlier to exist and it is a part of our natural process, we should investigate the cause of the outlier as it can provide valuable clues that can help you better understand your process performance. Outliers may be hiding precious information that could be invaluable to improving your process performance.

GitHub Repo link:

https://github.com/ASfatima/DataGlacierVirtualInternship/tree/main/Week%208%20-%20Deliverables