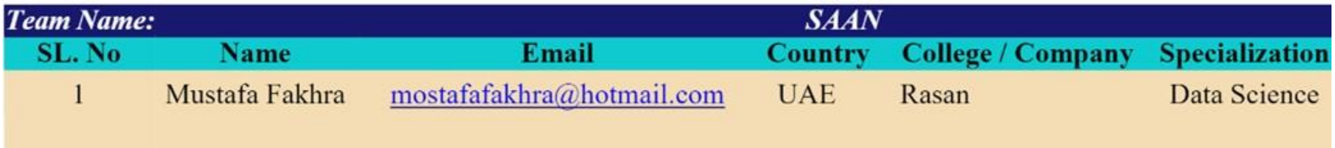
Data Glacier



Team Name:			SAAN		
SL. No	Name	Email	Country	College / Company	Specialization
1	Mustafa Fakhra	mostafafakhra@hotmail.com	UAE	Rasan	Data Science

Final Project Report

Table of Contents

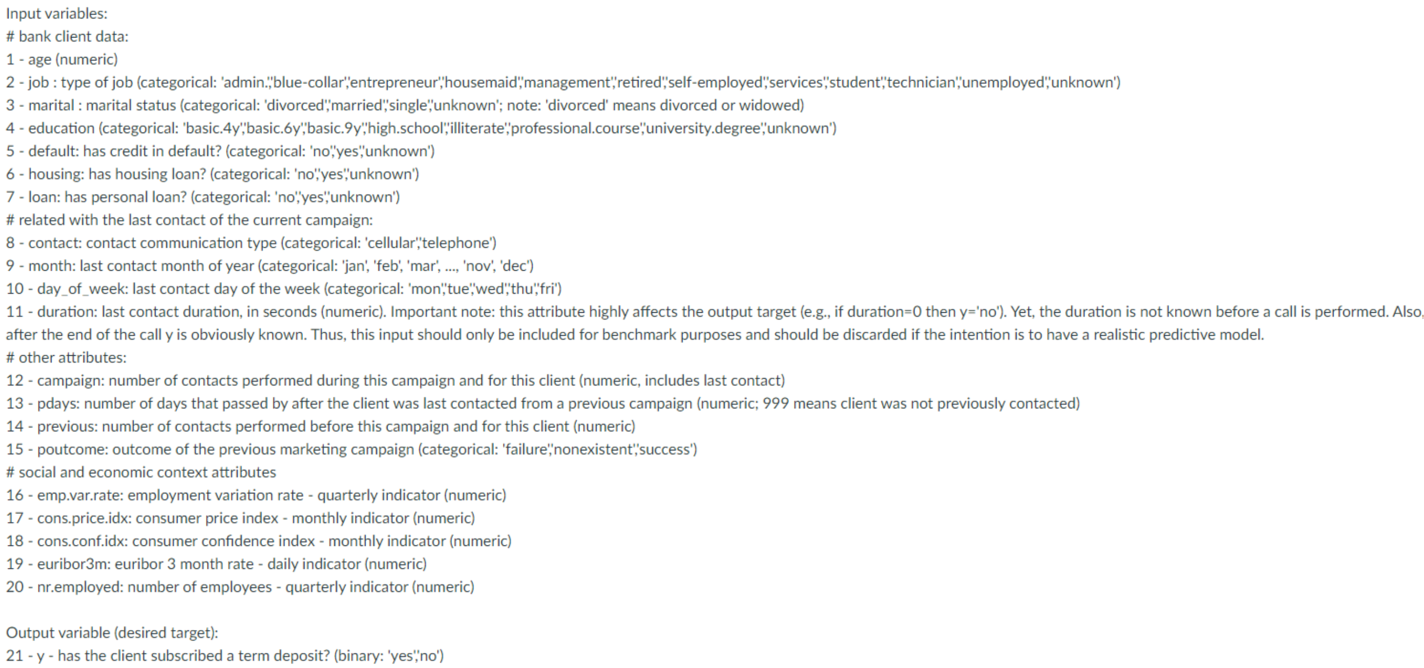
Problem Description

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).

Business Understanding

Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more. This will save resource and their time (which is directly involved in the cost (resource billing)). Develop model with Duration and without duration feature and report the performance of the model. Duration feature is not recommended as this will be difficult to explain the result to business and also it will be difficult for business to campaign based on duration.

Dataset



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: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). 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: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

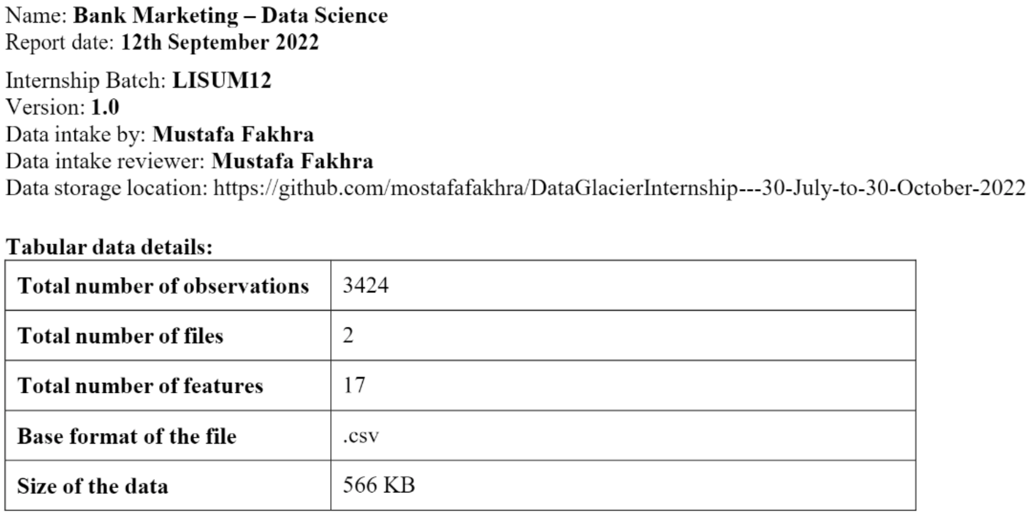
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Project Lifecycle

TASKS	19th Sep Week 0	26th Sep Week 1	2nd Oct Week 2	9th Oct Week 3	16t W
Week 7	🟩				
Week 8		🟩			
Week 9		🟩			
Week 10			🟩		
Week 11				🟩	
Week 12				🟩	🟩

Data Intake Report

Data Intake Report:



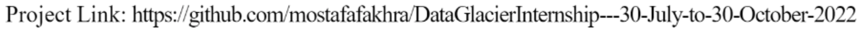
Name: **Bank Marketing – Data Science**
Report date: **12th September 2022**

Internship Batch: **LISUM12**
Version: **1.0**
Data intake by: **Mustafa Fakhra**
Data intake reviewer: **Mustafa Fakhra**
Data storage location: https://github.com/mostafafakhra/DataGlacierInternship---30-July-to-30-October-2022

Tabular data details:

Total number of observations	3424
Total number of files	2
Total number of features	17
Base format of the file	.csv
Size of the data	566 KB

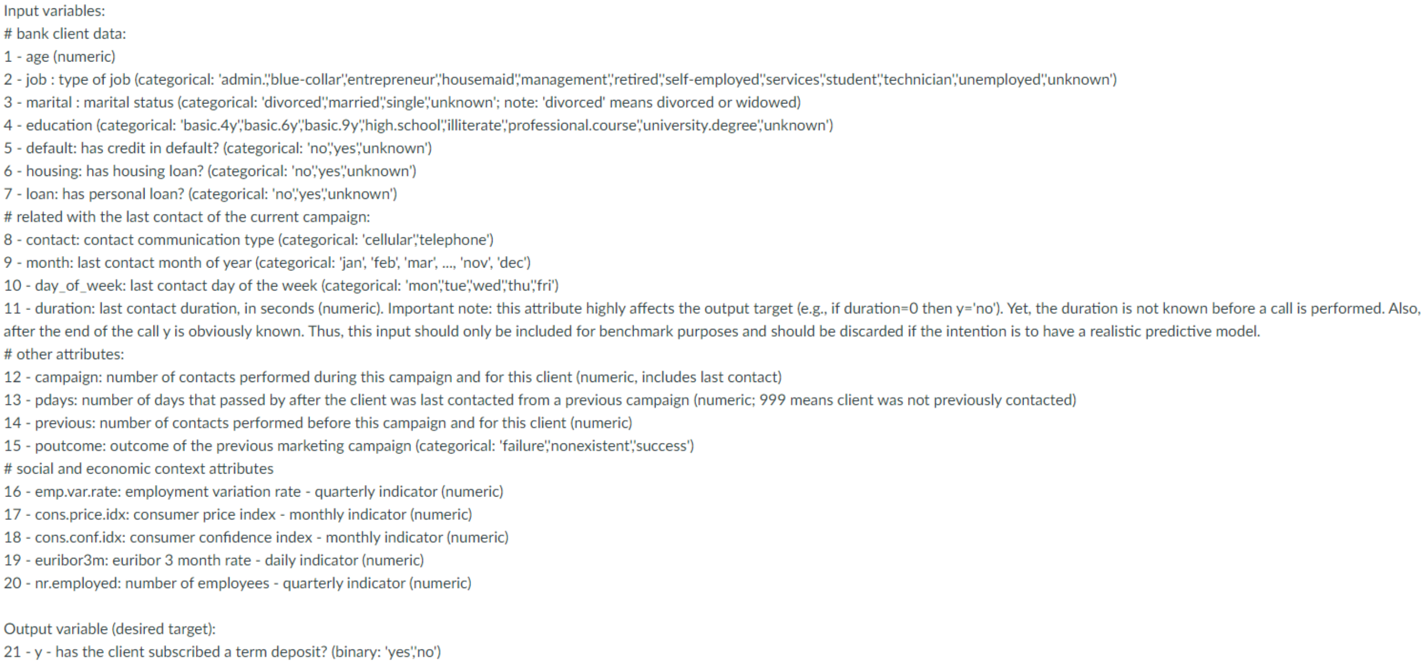
GitHub Repository:



Project Link: https://github.com/mostafafakhra/DataGlacierInternship---30-July-to-30-October-2022

Data Types

In this dataset as you can find in data intake report, we have dataset with following datatypes, “object” types mean categorical columns:



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: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). 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: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Data Problems

- Null Values: This dataset has Null values
- Outliers: We have only two numerical columns and both of them have some outliers.

- Skewness and Kurtosis: We have only two numerical columns and both of them have some outliers.

Transformation

As we did not have any Null values, so we have nothing to do in this regard. We have some skewness and Kurtosis in our two numerical features, so we will scaled their values by RobustScaler() and after that remove their outliers by calculating IQR and remove data smaller/greater than two whiskers. After removing outliers from “dexa_freq_during_rx” .

in this way :

Data

The other thing that we performed on the dataset is “one hot encoding”, For using classifiers we need numerical values, to do this I used One Hot Encoding that implemented by “get_dummies()” function from Pandas library.

Final Recommendation

Now we can perform classifiers models on the train set which we get it by splitting whole dataset to train and test sets in the way 70% for train set and 30% test set.