Week 8

Data Science Intern at Data Glacier
Project: Bank Marketing (Campaign)

Name: Adrian Baysa
Email: adrianbaysa2@gmail.com
Country: Philippines
Batch Code: LISUM16

1. **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).

2. **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 classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). The bank-full.csv dataset has all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).

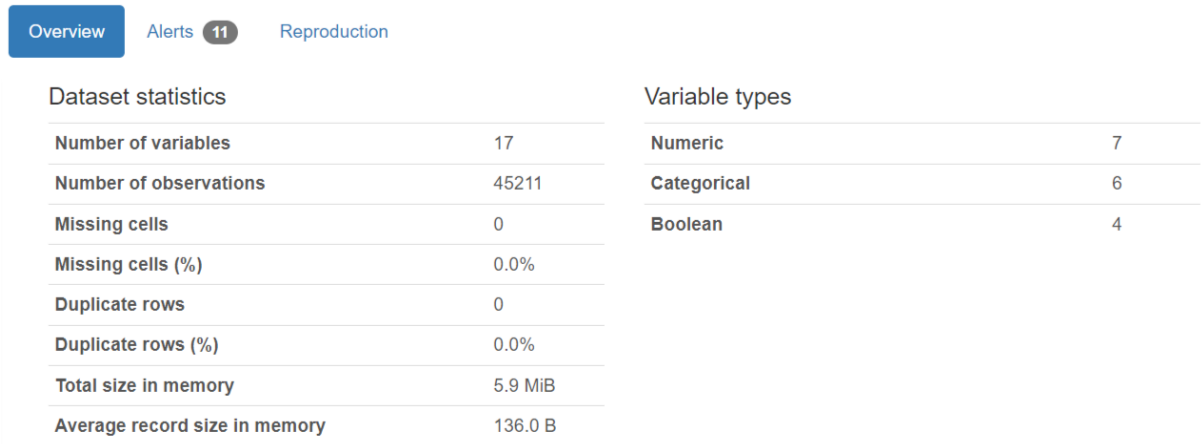
3. **Data Types**

   The following are the details of the features and target variable:

   a. 1 - age (numeric)
   b. 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
   c. 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
   d. 4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
   e. 5 - default: has credit in default? (categorical: 'no','yes','unknown')
   f. 6 - housing: has housing loan? (categorical: 'no','yes','unknown')
   g. 7 - loan: has personal loan? (categorical: 'no','yes','unknown')
      # related with the last contact of the current campaign:
   h. 8 - contact: contact communication type (categorical: 'cellular','telephone')
   i. 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
   j. 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
   k. 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:

l. 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
m. 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)
n. 14 - previous: number of contacts performed before this campaign and for this client (numeric)
o. 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
p. 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
q. 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
r. 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
s. 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
t. 20 - nr.employed: number of employees - quarterly indicator (numeric)
u. Output variable (desired target): 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Further details of the data types and the sample of the dataframe are presented below:

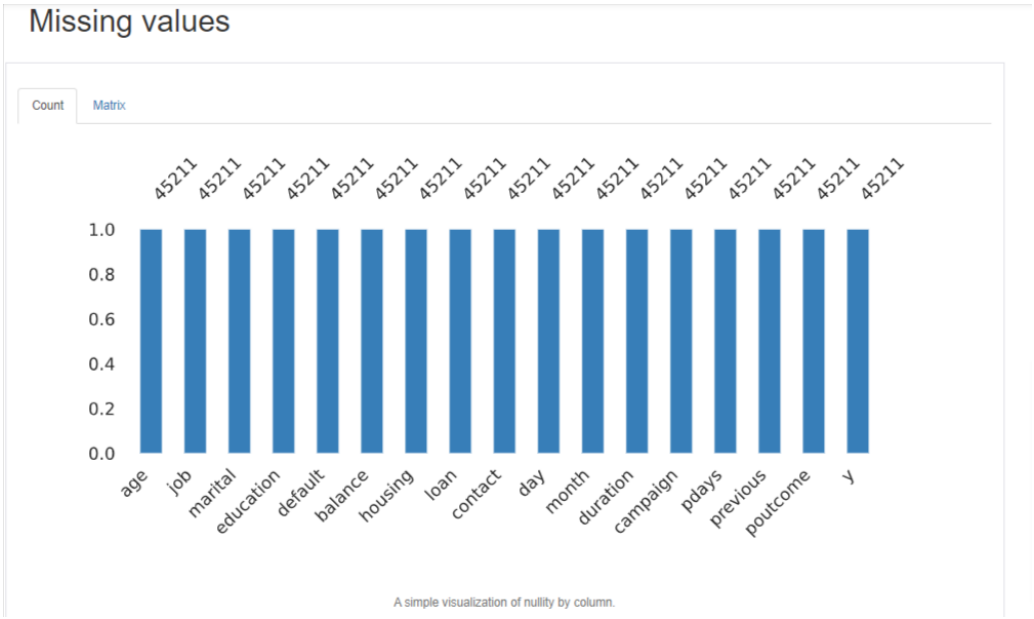


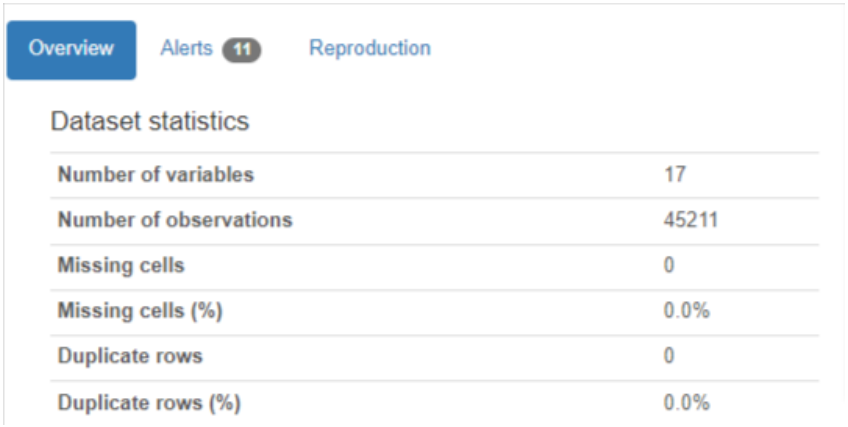
Sample

First rows | Last rows

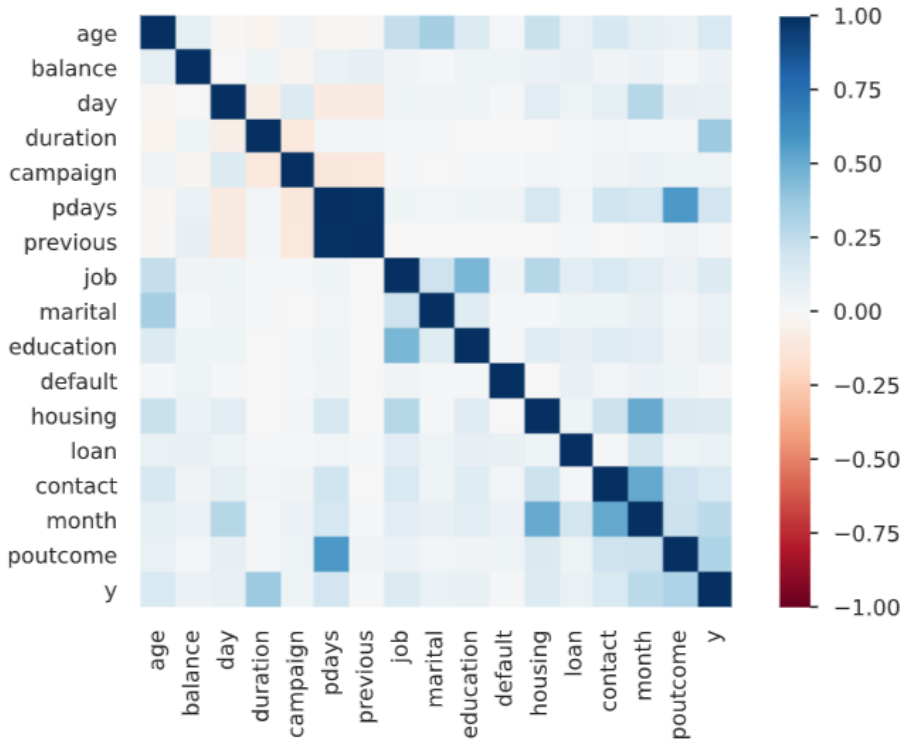
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcom
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown
5	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown
6	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown
7	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown
8	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown
9	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown

4. **Problems with the data**
    a. For the data set used, no missing data or duplicate rows were encountered.





b. Some features are strongly correlated, but can easily be managed.



Overview | Alerts 11 | Reproduction

Alerts

pdays is highly overall correlated with previous and 1 other fields	High correlation
previous is highly overall correlated with pdays	High correlation
housing is highly overall correlated with month	High correlation
contact is highly overall correlated with month	High correlation
month is highly overall correlated with housing and 1 other fields	High correlation
poutcome is highly overall correlated with pdays	High correlation
default is highly imbalanced (87.0%)	Imbalance
poutcome is highly imbalanced (53.1%)	Imbalance
previous is highly skewed (γ1 = 41.84645447)	Skewed
balance has 3514 (7.8%) zeros	Zeros
previous has 36954 (81.7%) zeros	Zeros

5. **Approaches**
    a. Do features engineering for continuous data
    b. Check the impact of Principal Component analysis to reduce the dimensionality of the data and address highly correlated features
    c. Do features engineering to reduce the cardinality of categorical data

6. **Github Repolink:** https://github.com/agbaysa/dataglacier_week8