Data Glacier

WEEK 8 DELIVERABLES

Group Name: Data Lover

Name	Email	Country	College/company	Specialization
Safi Cengiz	saficengiz1@gmail.com	Turkiye	Beykoz University	Data science
Mohsen Bahremani	M.Bahremani@gmail.com	Canada	Wilfrid Laurier University	Data science
Batta Liu	liubatta@gmail.com	Canada	University of British Columbia	Data science

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). ABC bank has given responsibility to the Data Science Data Lover Team to develop an automated process of classification and asked to develop a ML model to shortlist customers with higher chances of buying the product, so that ABC's marketing team can focus on them and save the time and money.

2. BUSINESS UNDERSTANDING

There has been a revenue decline for an ABC bank, and they would like to know what actions to take. After investigation, they found out that the root cause is that their clients are not depositing as frequently as before. Knowing that term deposits allow banks to hold onto a deposit for a specific amount of time, banks can invest in higher gain financial products to make a profit. In addition, banks also hold better chances to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. As a result, the ABC bank would like to identify existing clients that have higher chances to subscribe for a term deposit and focus marketing efforts on such clients. The classification goal is to predict if the client will subscribe to a term deposit or not.

3. DATA UNDERSTANDING

The data corresponds to 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.

Four datasets, from May 2008 to November 2010, are provided to be modeled with a classification algorithm, from among which two pairs of train and test data are available for analysis. The 'bank-full.csv' and 'bank.csv' are one of the pairs having less than 20 input features and are an older version of 'bank-additional-full.csv' and 'bank-additional.csv'.

I. Data Description

Dataset	Separation	Description
bank-additional-full.csv	Train	41118 observations and 20 inputs ordered by date (from May 2008 to November 2010)
bank-additional.csv	Test	4118 observations (10% of train data) with 20 inputs
bank-full.csv	Train	45211 observations and 17 inputs ordered by date (older version of bank-additional-full)
bank.csv	Test	4521 observations (10% of train data) and 17 inputs

Input variables:

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 a 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

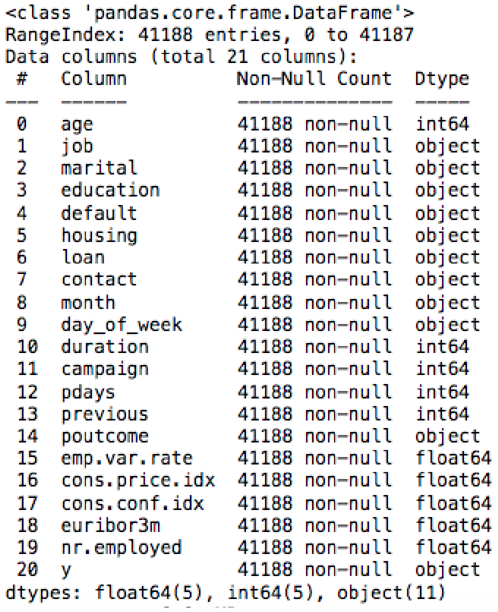
- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

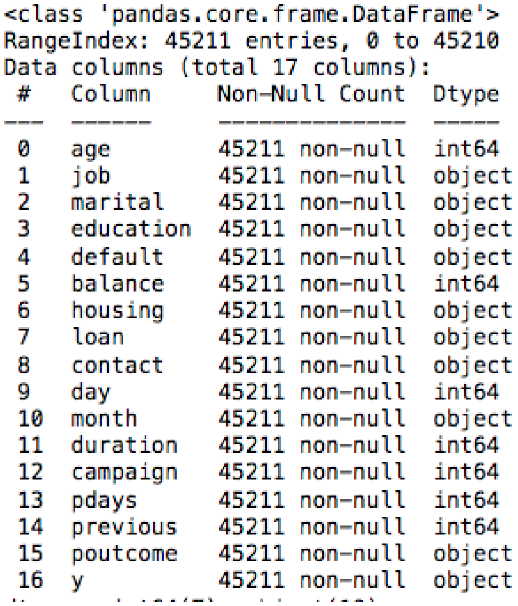
- 21 - y - has the client subscribed a term deposit? (binary: 'yes', 'no')

II. Data type

The features are divided between "object" types, i.e., categorical attributes, and "int64 / float64" types which are numerical attributes. In addition, there is a consistency between bank-additional-full.csv and bank-full.csv data in terms of data type.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
```

III. Missing Values

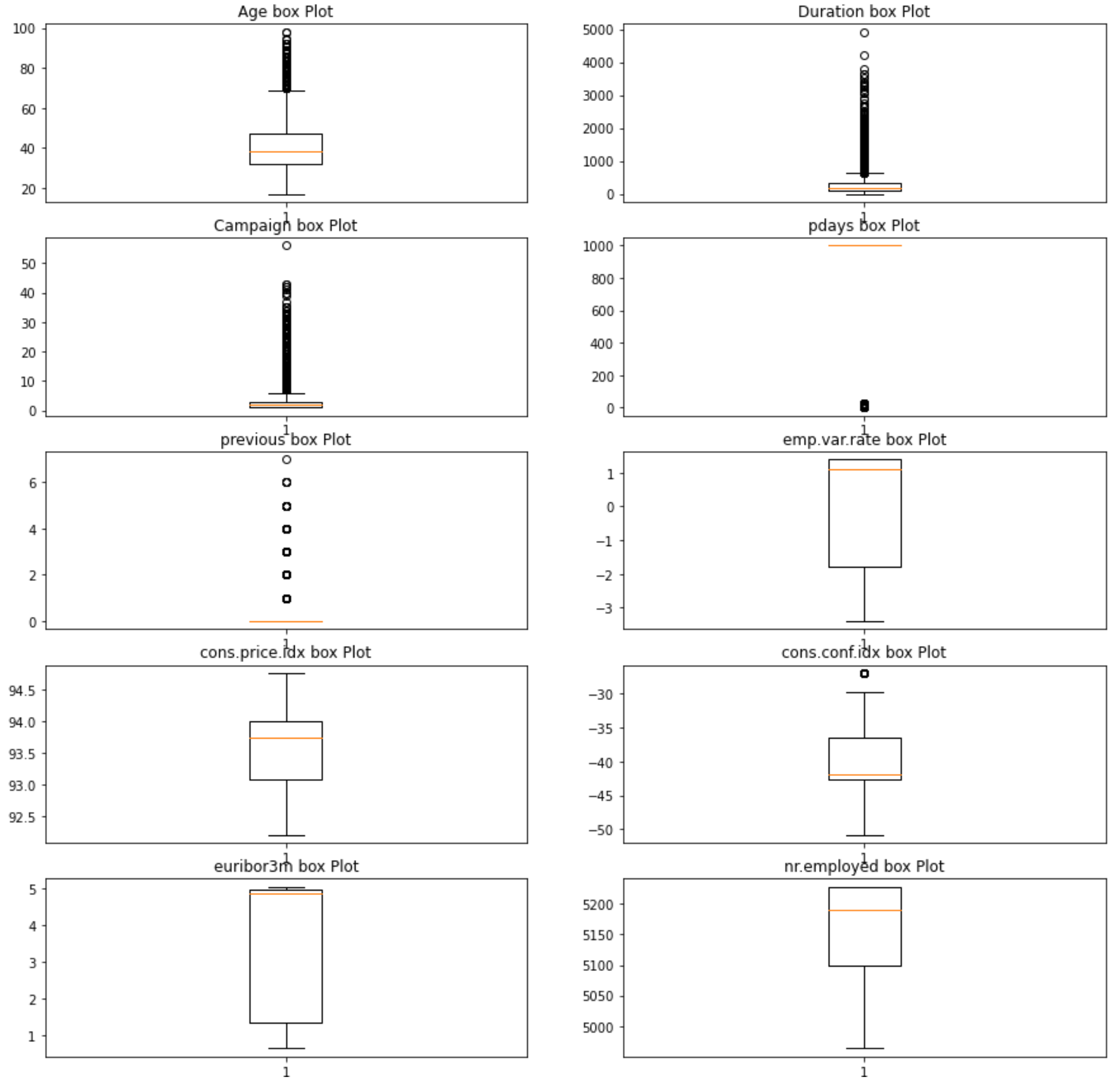
All 4 datasets have no missing values. However, there exist “unknown” value in categorical variables. In some cases, the statisticians leave missing value as a category when they are not random at sampling. At this case, “unknown” belongs to 'job', 'marital', 'education', 'default', 'housing', and 'loan' features, and we will deal with them at data preparation phase.

IV. Duplications

Only ‘bank-additional-full.csv’ has 12 duplicated observations.

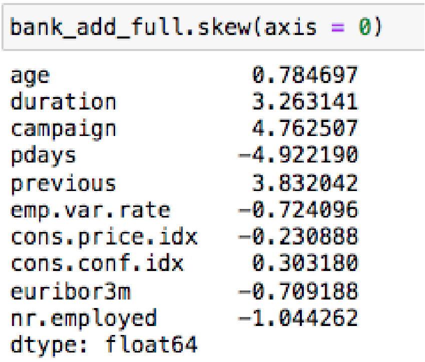
V. Outliers

In ‘age’, ‘campaign’, and ‘previous’ have outliers. The numbers 999 for ‘pdays’, by its definition, should not be considered as outliers.



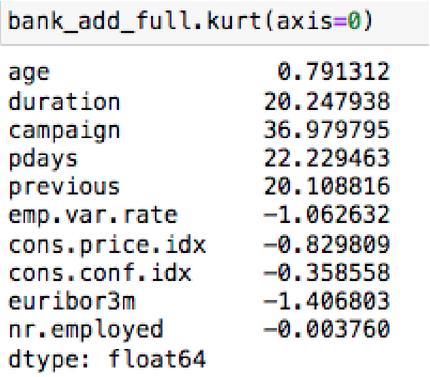
VI. Distributions

The skew result shows a positive (right) or negative (left) skew. Values closer to zero show less skew. Skewness is a measure of asymmetry of the distribution relative to the normal distribution. Positive skewness implies the tail is in the right of the mean of the distribution. Negative skewness implies the tail is in the left of the mean of the distribution.
Kurtosis is the measure of whether or not a distribution has a heavy tail or not relative to the normal distribution. A value >3 means the distribution has a heavy tail. Kurtosis <3 implies the distribution has a lighter tail than the normal distribution.



```
bank_add_full.skew(axis = 0)

age                0.784697
duration           3.263141
campaign           4.762507
pdays             -4.922190
previous           3.832042
emp.var.rate      -0.724096
cons.price.idx    -0.230888
cons.conf.idx      0.303180
euribor3m         -0.709188
nr.employed       -1.044262
dtype: float64
```



```
bank_add_full.kurt(axis=0)

age                0.791312
duration          20.247938
campaign          36.979795
pdays             22.229463
previous          20.108816
emp.var.rate      -1.062632
cons.price.idx    -0.829809
cons.conf.idx     -0.358558
euribor3m         -1.406803
nr.employed       -0.003760
dtype: float64
```

4. REPO

HTTPS://GITHUB.COM/BATTALIU/BANK_MARKETING_GROUP_PROJECT