DATA SCIENCE INTERNSHIP - DATA GLACIER

Project: Bank Marketing (Campaign) -- Group Project

Group Name: **Datazoids**

Name: **Efe KARASIL - Sefa Sözer**

E-mail: **ekarasil@sabanciniv.edu - sefasozer9@gmail.com**

Country: **Turkey - Turkey**

College: **Sabancı University - Trakya University**

Specialization : **Data Science**

<u>Problem Description:</u>

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps 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).

<u>Business Understanding:</u>

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

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

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 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) to a term deposit (variable y).

Citation Request:
This dataset is publicly available for research. The details are described in [Moro et al., 2014].
Please include this citation if you plan to use this database:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

Available at: [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001
[bib] http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt

1. Title: Bank Marketing (with social/economic context)

2. Sources
Created by: Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL) @ 2014

3. Past Usage:

The full dataset (bank-additional-full.csv) was described and analyzed in:

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

4. Relevant Information:

This dataset is based on the "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing).
The data is enriched by the addition of five new social and economic features/attributes (nationwide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb.
This dataset is almost identical to the one used in [Moro et al., 2014] (it does not include all attributes due to privacy concerns).
Using the rminer package and R tool (http://cran.r-project.org/web/packages/rminer/), we found that the addition of the five new social and economic attributes (made available here) lead to substantial improvement in the prediction of a success, even when the duration of the call is not included. Note: the file can be read in R using:
d=read.table("bank-additional-full.csv",header=TRUE,sep=";")

The binary classification goal is to predict if the client will subscribe to a bank term deposit (variable y).

5. Number of Instances: 41188 for bank-additional-full.csv

6. Number of Attributes: 20 + output attribute.

7. Attribute information:

For more information, read [Moro et al., 2014].

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")

8. Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

The problems in the data and their envisaged solutions:

```
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   age             30113 non-null  int64  
 1   job             30113 non-null  object 
 2   marital         30113 non-null  object 
 3   education       30113 non-null  object 
 4   default         30113 non-null  object 
 5   housing         30113 non-null  object 
 6   loan            30113 non-null  object 
 7   contact         30113 non-null  object 
 8   month           30113 non-null  object 
 9   day_of_week     30113 non-null  object 
 10  duration        30113 non-null  int64  
 11  campaign        30113 non-null  int64  
 12  pdays           30113 non-null  int64  
 13  previous        30113 non-null  int64  
 14  poutcome        30113 non-null  object 
 15  emp.var.rate    30113 non-null  float64
 16  cons.price.idx  30113 non-null  float64
 17  cons.conf.idx   30113 non-null  float64
 18  euribor3m       30113 non-null  float64
 19  nr.employed     30113 non-null  float64
 20  y               30113 non-null  object 
dtypes: float64(5), int64(5), object(11)
```

“NA” value is not observed in any of the data, which is both numeric and object type. However, in object types, as stated in the 8th item above, "unknown" statement-string is used to point out the missing value.

```
df2.isnull().sum()
```

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
```

When it is checked with python code it can be seen that there is no NA value in the dataset.

```
job
330
marital
80
education
1731
default
8597
housing
990
loan
990
contact
0
month
0
day_of_week
0
poutcome
0
y
0
```

When object typed columns are checked, the "unknown" string is searched, the number of NA, "unknown" rows from 41188 rows in total in each column are stated with their column names above.

```
df2['job'].value_counts()
```

admin.	10422
blue-collar	9254
technician	6743
services	3969
management	2924
retired	1720
entrepreneur	1456
self-employed	1421
housemaid	1060
unemployed	1014
student	875
unknown	330

When the "job" column is handled, "unknown" is found in 330 rows, as it is pointed with a red rectangle on the left. So, it is planned to be removed from the data because it is seen as an insignificant amount.

```
df2['marital'].value_counts()
```

married	24928
single	11568
divorced	4612
unknown	80

When the "marital" column is handled, "unknown" is found in 80 rows, as it is pointed with a red rectangle on the left. So, it is planned to be removed from the data because it is seen as an insignificant amount.

```
df2['education'].value_counts()
```

university.degree	12168
high.school	9515
basic.9y	6045
professional.course	5243
basic.4y	4176
basic.6y	2292
unknown	1731
illiterate	18

When the "education" column is handled, "unknown" is found in 1731 rows, as it is pointed with a red rectangle on the left. So, it is planned to be handled as a different class since it has a significant number of members.

```
df2['default'].value_counts()

no         32588
unknown     8597
yes            3
```

When the “default” column is handled, “unknown” is found in 8597 rows, as it is pointed with a red rectangle on the left. So, it is planned to be handled as a “no” since “no” is the median(most and even very frequently used value).

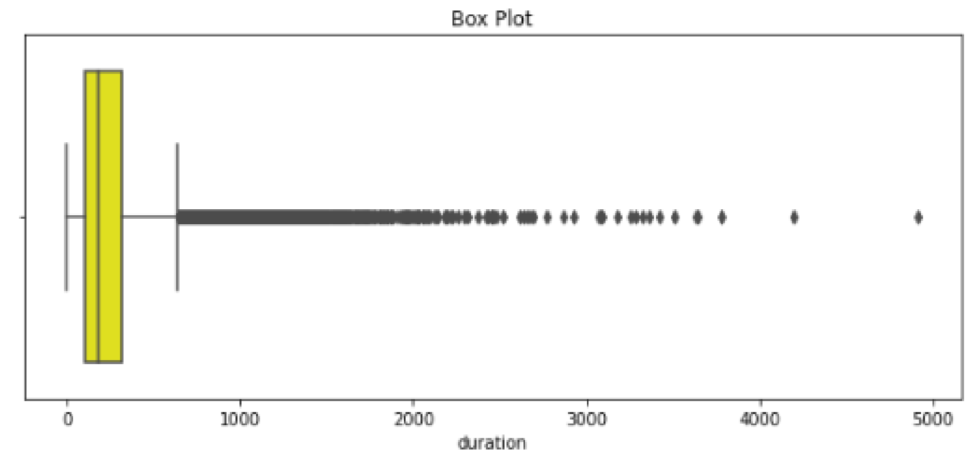
```
df2['housing'].value_counts()

yes        21576
no         18622
unknown      990
```

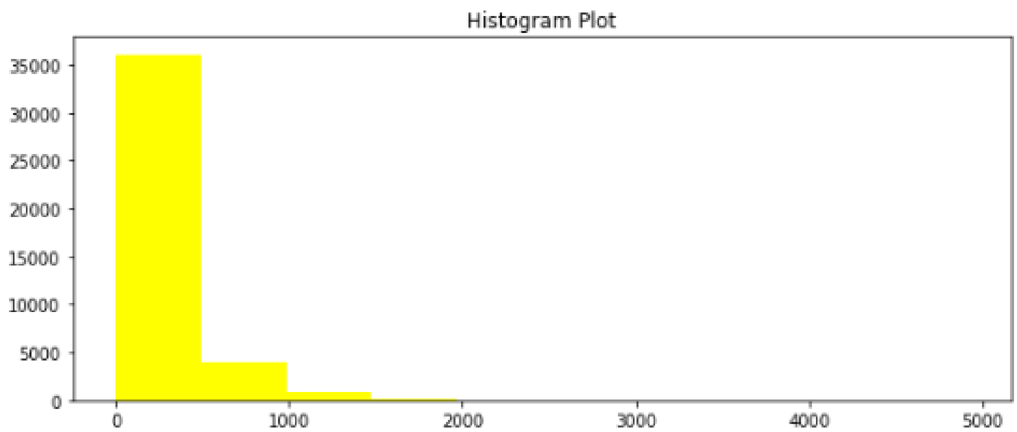
When the “housing” column is handled, “unknown” is found in 990 rows, as it is pointed with a red rectangle on the left. So, it is planned to be removed from the data because it is seen as an insignificant amount.

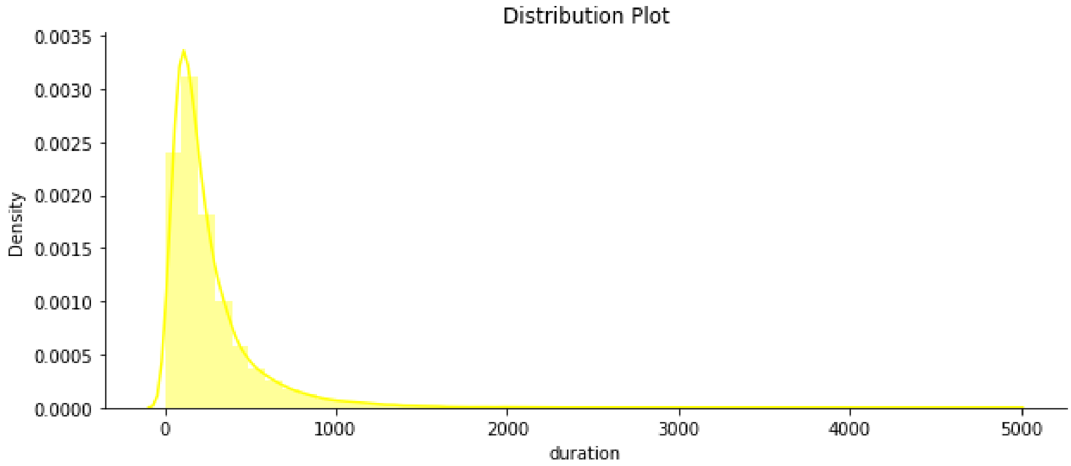
```
df2['loan'].value_counts()

no         33950
yes         6248
unknown      990
```

When the “loan” column is handled, “unknown” is found in 990 rows, as it is pointed with a red rectangle on the left. So, it is planned to be removed from the data because it is seen as an insignificant amount.

Outliers - skewness and their envisaged solutions:







To see the outliers and skewness, it is planned to use Boxplot, Histogram and Distribution plots. Examples of the “duration” column and its plots are given above.

Since the dataset is not that small and there is enough data, there can be both deletion process of the outliers or imputation process, it is planned to use **IQR method** for both processes.

IQR Method

In this method by using InterQuartile Range(IQR), we detect outliers. IQR tells us the variation in the data set. Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR treated as outliers.

- Q1 represents the 1st quartile/25th percentile of the data.
- Q2 represents the 2nd quartile/median/50th percentile of the data.
- Q3 represents the 3rd quartile/75th percentile of the data.
- (Q1–1.5IQR) represent the smallest value in the data set and (Q3+1.5IQR) represent the largest value in the data set