Data Glacier week 8

Team member's details:

Project: Bank Marketing (Campaign) -- Group Project

Group Name: Kesimoji

Names: Kemal Cagin Sertkaya, Jinwen Li, Mohamed Elmorsy, Sirui Zhang

Emails: cagin24@gmail.com, jinwen@uw.edu, mmsobhy7@gmail.com, zhangsirui261918@126.com

Colleges: Bogazici University, University of Washington, Mcmaster University, UCL
Specialization: Data Science
Countries: Turkey, US, Canada, UK

Problem description:

One bank wants to sell its term deposit product to customers before launching the product. To save their resource and time, they want to know what kind of customers they should focus on, and then they can put more advertisements to these customers, who have more chances of buying the product. Thus, our problem is to pick up this kind of customer, based on customers' past interaction with this bank or other financial institutions. We are going to use the customers' data to build some machine learning models and then, select customers who most likely buy the product.

Data understanding:

Here the details of the dataset collected (bank_full.csv):

1. It has 45,211 observations
2. It has 17 columns
3. Variables: There are 10 of 17 columns are object data type, and 7 out of 17 columns are int data type.

What type of data you have got for analysis

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

Output:

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

Missing Attribute Values: None

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

1. There are a lot of the columns have 'unknown' values

```
len(bank_full[bank_full['contact']=='unknown'])
```

```
13020
```

```
len(bank_full[bank_full['job']=='unknown'])
```

```
288
```

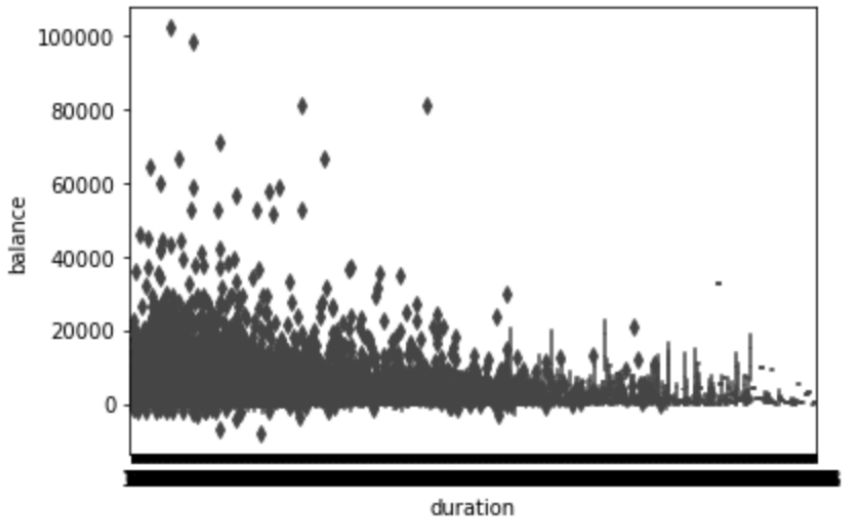
```
len(bank_full[bank_full['poutcome']=='unknown'])
```

```
36959
```

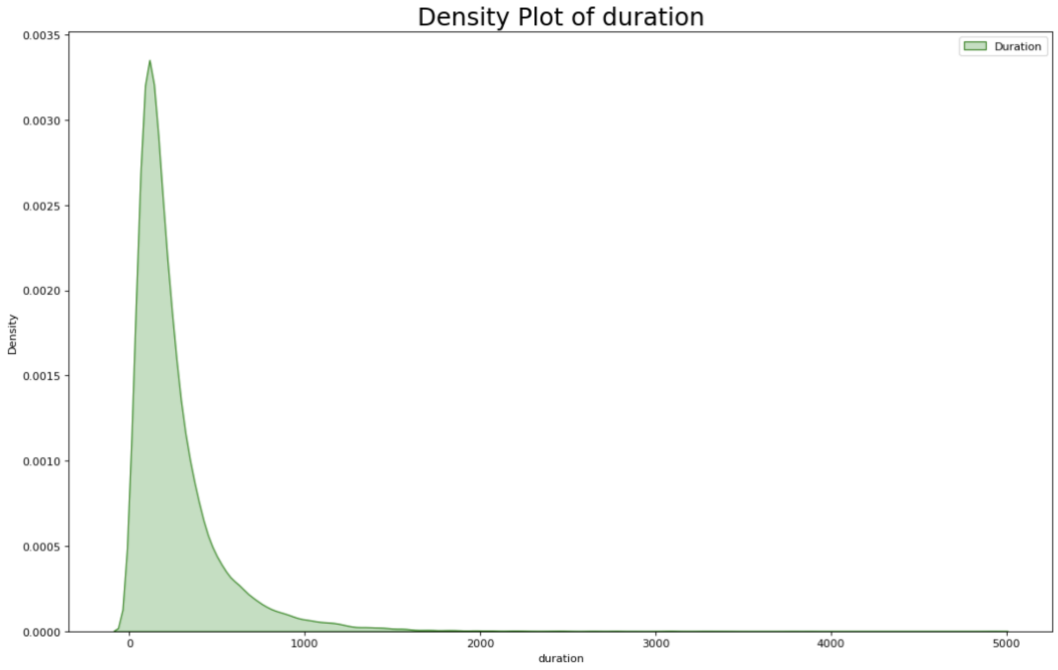
2. We found there are some outliers

```
: sbn.boxplot(x='duration', y='balance', data=bank_full)
```

```
: <AxesSubplot:xlabel='duration', ylabel='balance'>
```



3. We also found the data is skewed



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

1. As for NA value, here is the ‘unknown’ values. We'll use the median or mode value to replace them.
2. Outliers: Since there is enough data, these outliers can be deleted or use interpolation. We can use IQR method to do it.
3. For skewed data, we can do some standardization to make our data is not skewed.