Final Presentation

Bank Marketing Campaign

09.29.2022

Jinwen Li

Agenda

Summary

Data Understanding

EDA

EDA Recommendations

Model Recommendation

Applying Models

Final Model Selection

Data Glacier

Your Deep Learning Partner

Team member's detail

Group Name: Kesimoji
Names: Kemal Cagin Sertkaya, Jinwen Li
Emails: cagin24@gmail.com, jinwen@uw.edu

Colleges: Bogazici University, University of Washington
Specialization: Data Science
Countries: Turkey, US

Summary

- **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 will use the customers' data to build machine learning models and then select customers who most likely buy the product.

- **Analysis:**

The Analysis is divided into the following parts:

- Data Understanding
- EDA: Data analysis
- EDA Recommendations
- Model recommendations

Data Understanding

Data Set Information:

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.

There are four datasets:

1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).

4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Data Understanding

- **Attribute Information:**
- 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.

Data Understanding

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')

EDA- Outlier Detection and Handling

In order to detect and remove outliers, here we use two statistical methods: Interquartile range(IQR) and Standard Deviation.

	age	balance	day	duration	campaign	pdays	previous
count	40,856	40,856	40,856	40,856	40,856	40,856	40,856
mean	41	1,369	15	265	2	42	1
std	11	3,053	8	258	1	102	2
min	18	-8,019	1	0	1	-1	0
25%	33	76	8	109	1	-1	0
50%	39	455	15	187	2	-1	0
75%	48	1,440	21	326	3	-1	0
max	95	102,127	31	4,918	5	871	275

Interquartile range(IQR) statistical method

	age	balance	day	duration	campaign	pdays	previous
count	44,629	44,629	44,629	44,629	44,629	44,629	44,629
mean	41	1,360	16	258	3	38	0
std	11	3,050	8	258	3	98	1
min	18	-8,019	1	0	1	-1	0
25%	33	71	8	103	1	-1	0
50%	39	446	16	180	2	-1	0
75%	48	1,420	21	319	3	-1	0
max	95	102,127	31	4,918	63	871	7

Standard Deviation statistical method

EDA- NA Detection and Handling

In order to detect and handle N/A (unknown) values, here we also use two methods:
Drop N/A and using mode value to fill N/A values.

```
job : 288
marital : 0
education : 1857
default : 0
housing : 0
loan : 0
contact : 13020
month : 0
poutcome : 36959
y : 0
```

Here we can see the original unknown values

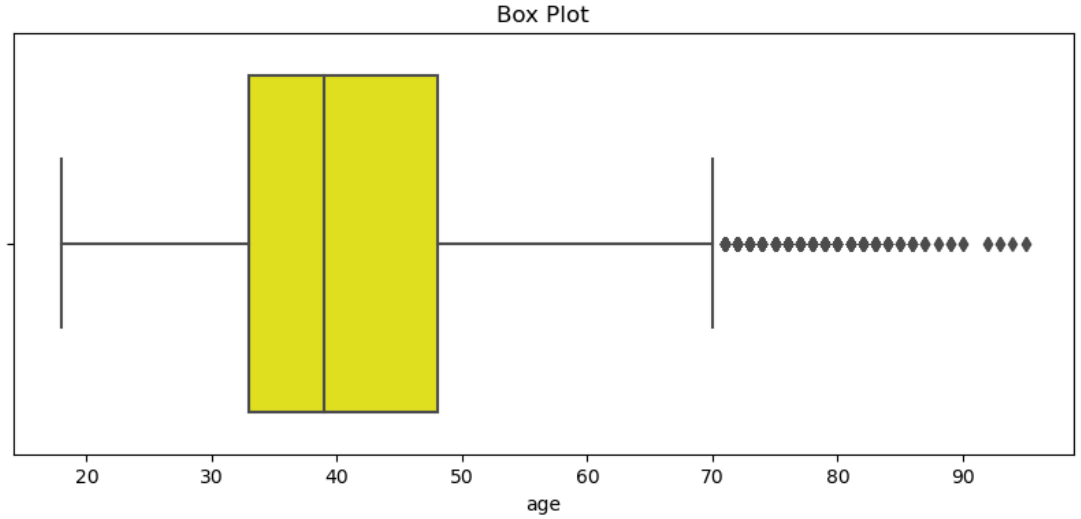
```
0          management
1          technician
2        entrepreneur
3         blue-collar
4                 NaN
             ...
45206      technician
45207         retired
45208         retired
45209     blue-collar
45210    entrepreneur
Name: job, Length: 45211, dtype: object
```

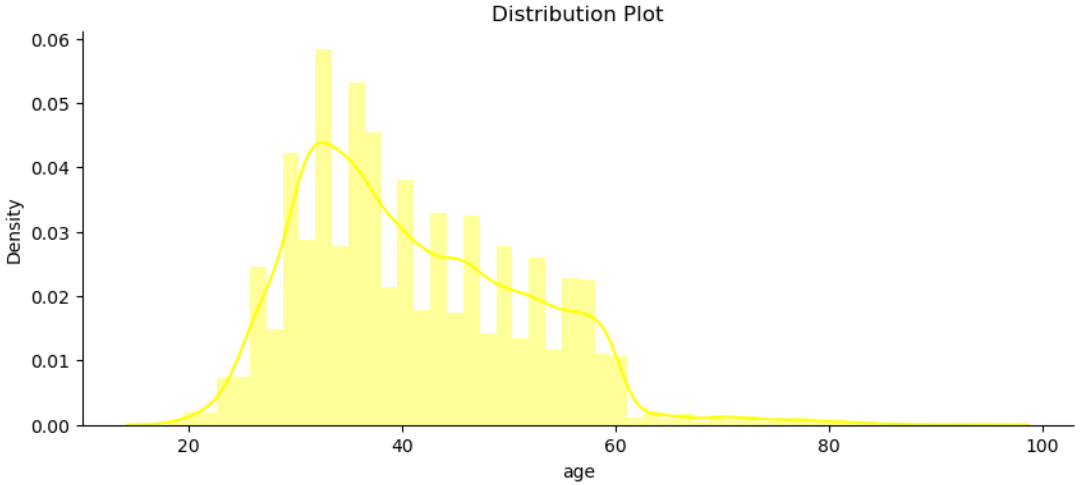
```
0          tertiary
1         secondary
2         secondary
3               NaN
4               NaN
             ...
45206      tertiary
45207       primary
45208     secondary
45209     secondary
45210     secondary
Name: education, Length: 45211,
```

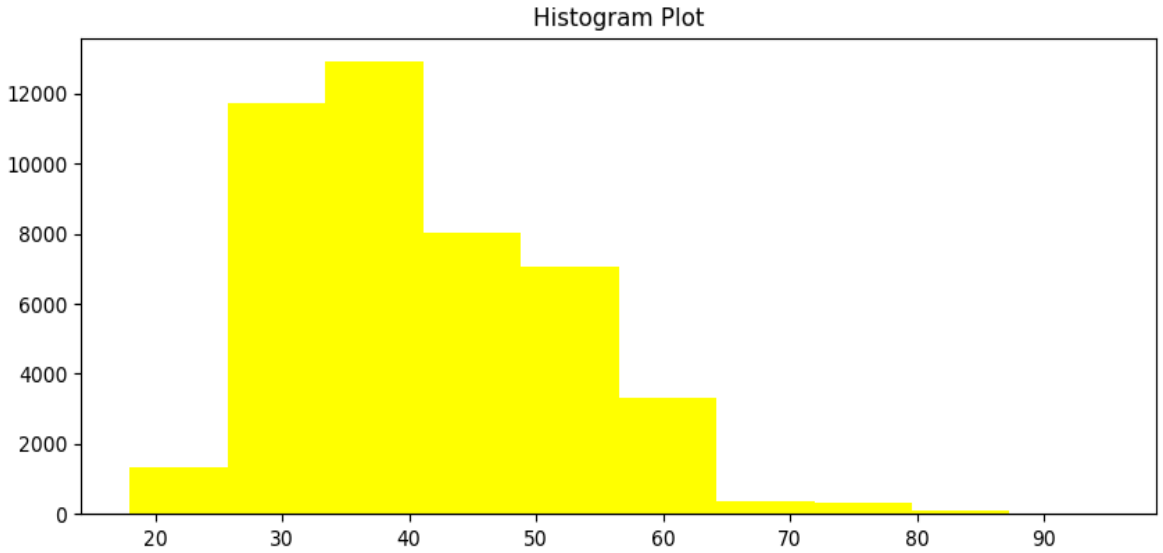
```
0         failure
1             NaN
2             NaN
3             NaN
4             NaN
            ...
45206         NaN
45207         NaN
45208     success
45209         NaN
45210       other
Name: poutcome, Length: 45211, dtype: object
```

Using mode value to fill N/A values

EDA- Categorical Columns-age

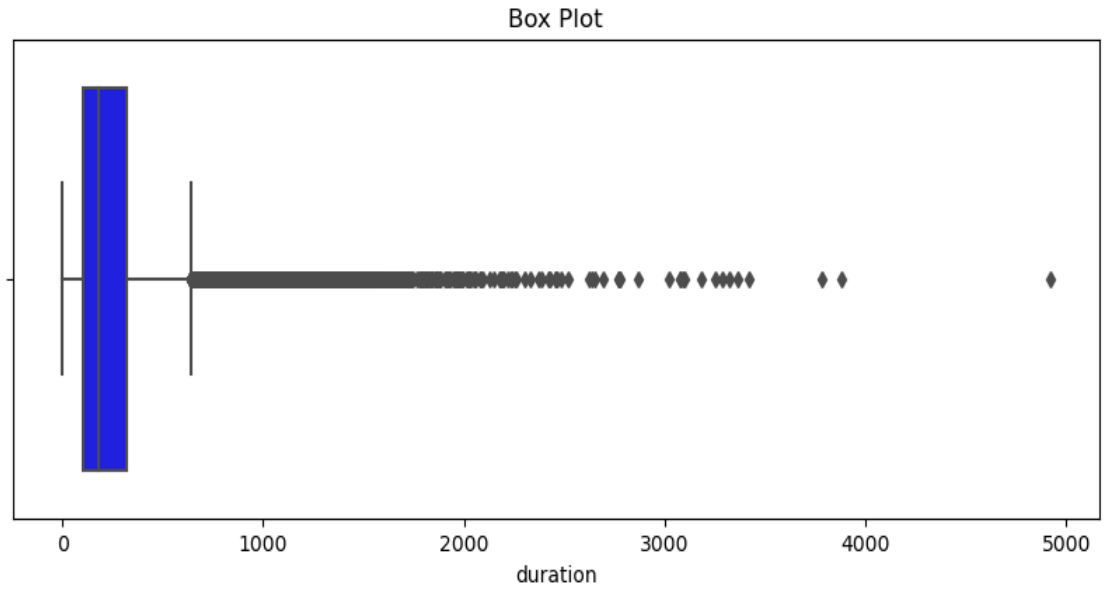


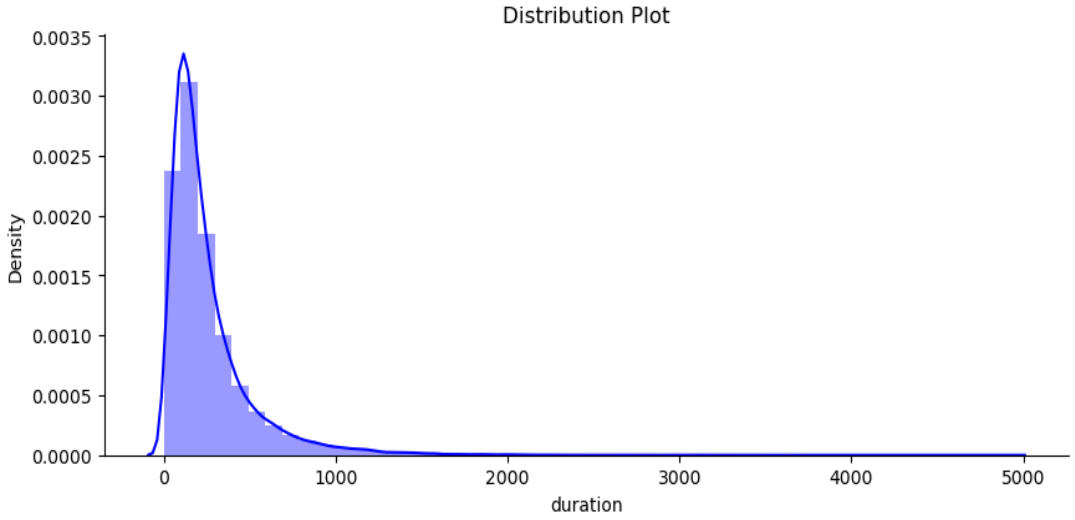


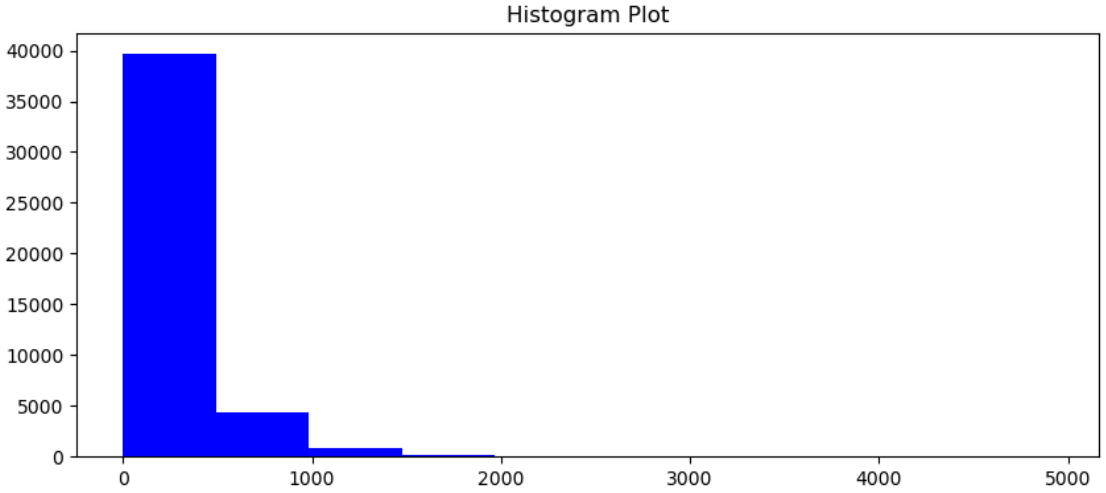


Here, for the age column, we can see the customers are mostly between ages 20-50.

EDA- Categorical Columns-duration

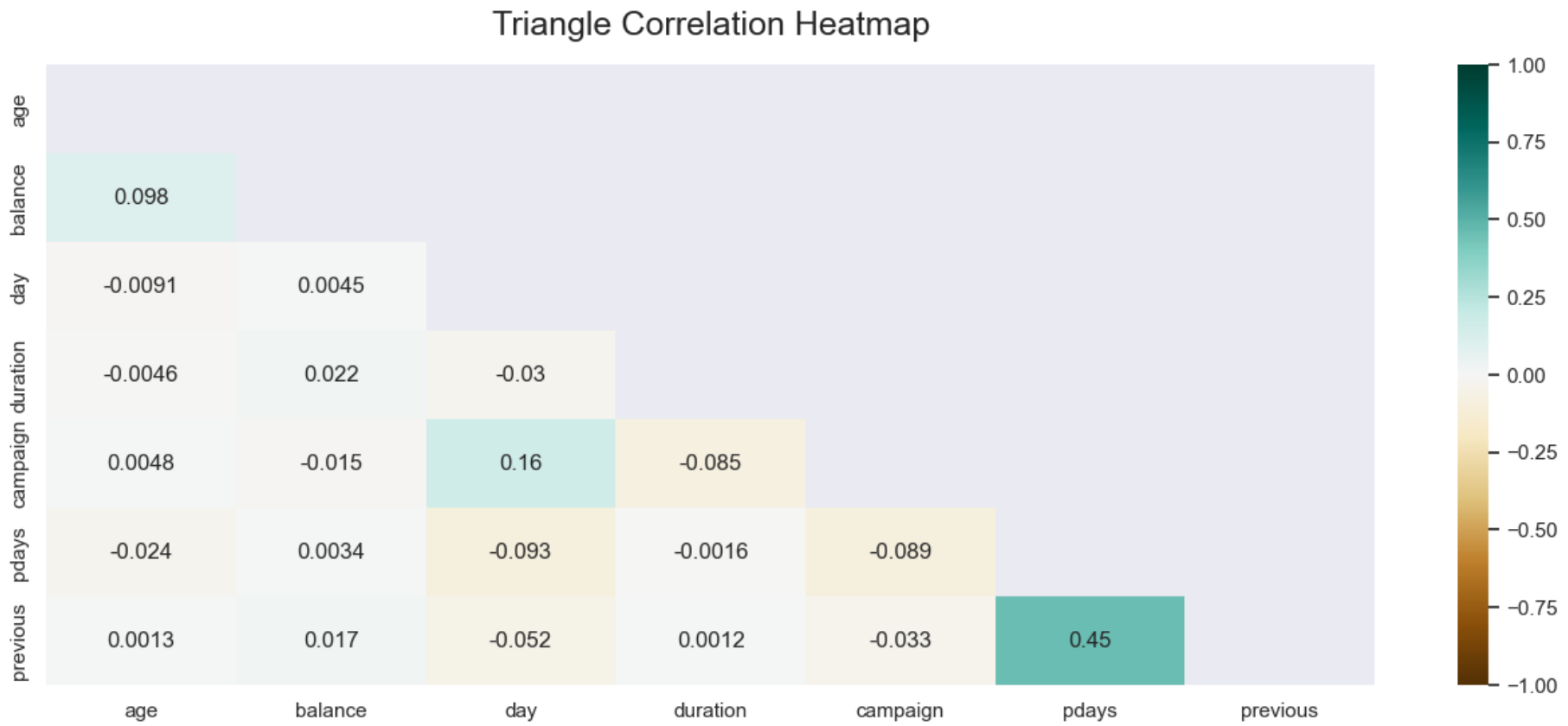






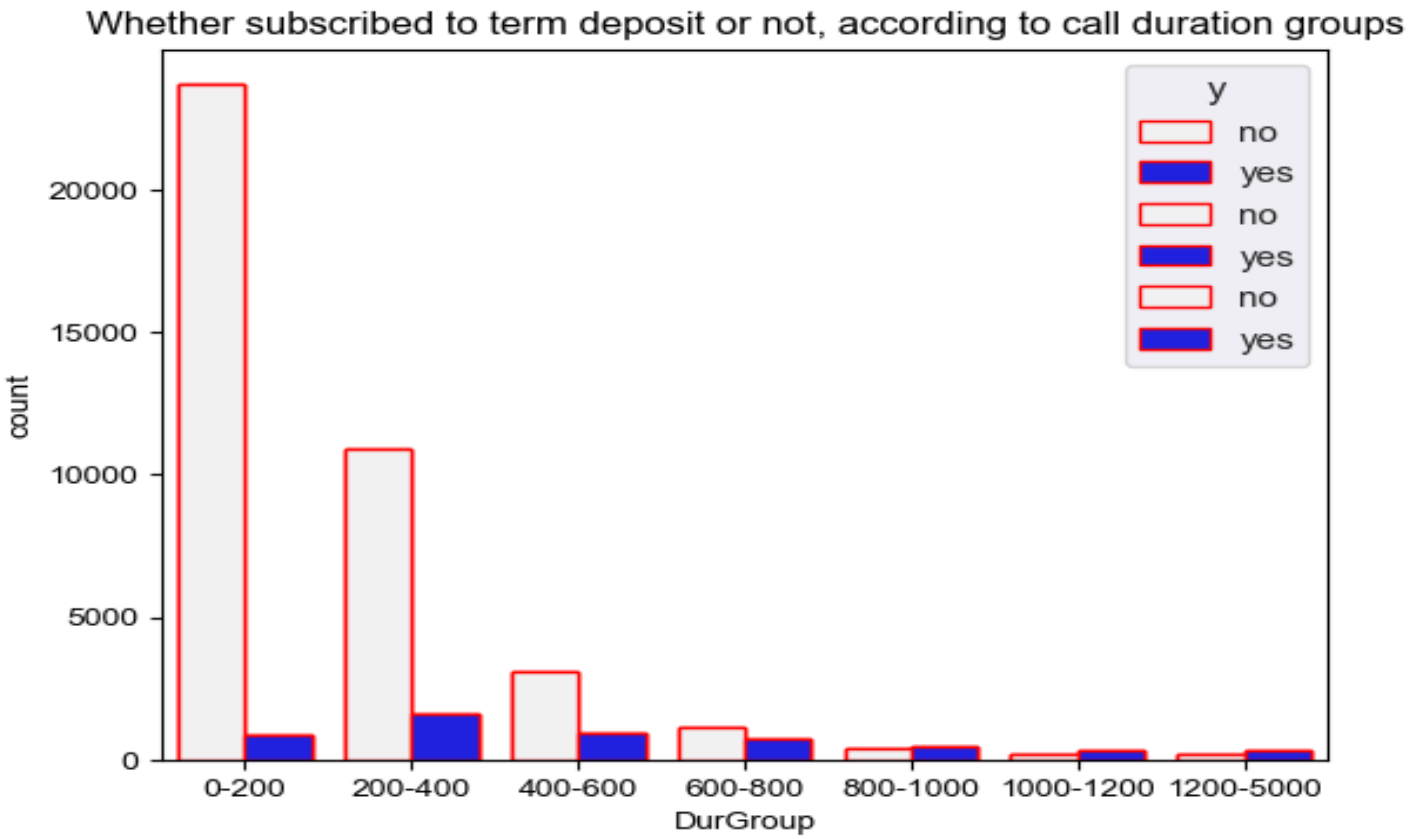
Here, for the duration column, we can see the duration are mostly between the duration 0-500. We can also see we have a lot of unusual big ones.

EDA Recommendations -1



Here, we can see from the correlation heat map, “pdays” and “previous” features have very high correlations between them. One of the two features can be dropped from the data since their existence will not be extra useful-providing good, new information- for the machine learning model which will be deployed in the next steps.

EDA Recommendations -2



X-label represents the duration of calls(in seconds) with customer, Y-label is count of that groups response to this campaign, It can bee seen that when call duration is increased getting “Yes” response probability is clearly increased as well

Model Recommendations

Because We are going to use the customers' data to build some machine learning models and then, select customers who most likely buy the product. Thus, this is a classification problem. Here, we can use the below models to do it.

- K Nearest Neighbors
- Decision Trees
- Random Forest
- Support Vector Machine

Applying Models

	Accuracy of the model (prediction)
1.Logistic Regression	90.03%
2.Support Vector Classification	90.39%
3.Gradient Boosting Classifier	89.62%
4.Random Forest	91.15%
5.Decision Tree	89.16%
6.XGBoost	91.50%

Final Model Selection

Therefore, we would choose the best prediction performance model, which is the XGBoost.

With XGBoost model, we can get 91.50% of prediction accuracy in a fair time.

Thank You