Data Science Internship 2022

Data Science: Bank Marketing (Campaign)

Submitted by: Big Analytics

Group Members:

Name	Email	College/Company
Taimoor Razi	taimoor.r10@gmail.com	Middle East Technical University, Turkey.
Ogwu Augustine	ogwuaugust@gmail.com	University of Jos, Nigeria.
Akshar Chaklashiya	chaklashiya.akshar@gmail.com	Lambton College, Toronto.

Submitted to: Data Glacier

Due Date: 2nd August 2022

Table of Contents

Project Lifecycle

Tasks

- Business Understanding
- Data understanding
- Exploratory data Analysis
- Data Preparation
- Model Selection & Model Building
- Performance reporting
- Deploy the model
- Converting ML metrics into Business metric and explaining result to business
- Presentation for non-technical persons.

Project Deadline

- 30th September 2022

Business Understanding

Problem Statement

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).

Why ML Model

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

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

Data Understanding

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

Data Intake Report

Group Name: Big Analytics
Report date: 16-08-2022
Internship Batch: LISUM11: 30
Version: 1.0
Data intake by: Taimoor Razi
Data intake reviewer: NA
Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Tabular data details:

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	5.56 MB

Proposed Approach:

- The data is downloaded from the UCI Machine Learning Repository.
- The bank-additional-full has no null values but has 12 duplicates. These 12 duplicates were removed.
- There are some values labelled as “unknown” in categorical variables.

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','service
s','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')

Exploratory Data Analysis (EDA)

What type of data you have got for analysis?

Multivariate dataset with multiple numerical (continuous, discreet and temporal) and categorical variables present.

What are the problems in the data?

Duplicate values: One of the dataset (having 44K rows) has 12 duplicate values which are dropped.

Imbalanced target variable: Dataset is highly imbalanced dataset as the ratio of target variable value is 8:1.

“Unknown” Values seem to appear in some features which is basically a missing value put inside a category.

Duration variable: Duration is obtained after the call is made to the potential client so if the target client has never received calls, this feature is not very useful. Duration variable should be removed during the analysis

Outliers present in some of the variables

What approaches are you trying to apply on your data set to overcome problems and why?

Imbalance dataset: To deal with imbalance target variable undersampling and oversampling methods are being applied.

Missing values: mean/median/mode value imputation for numerical variables. This imputation can also be done together with a groupby function. Most frequent category for categorical variables. Removal of columns with a lot of missing values is also being considered but leads to loss of important information. The use of an ML model to predict the missing values for some columns is also considered. Partial deletion of missing values in the numerical variables which attribute on numerical variables that do not show a strong relationship between the known/unknown status and target response.

Skewness: Transformations of features - log or normalize

Handling Categorical Data: Converting a few categorical values into numerical values by using One hot encoding - (ex. Default, housing, loan, contact). Converting temporal variables from categorical to numeric by using ordinal encoding - Month and week_of_day. Converting categorical target variable into numerical binary variable.

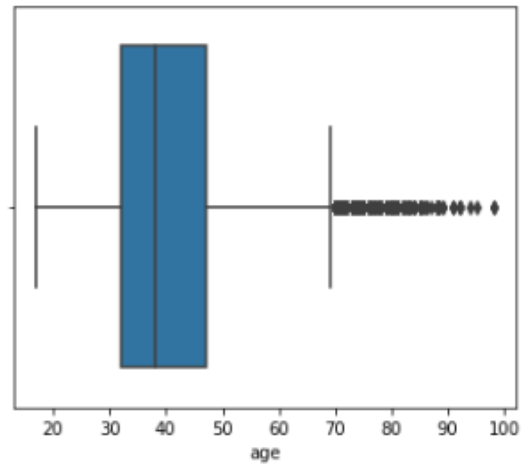
Handle Outlier: We have performed the visualization technique on numerical type variables to determine the outliers. Initially we have plot boxplot graph and then perform our analysis.

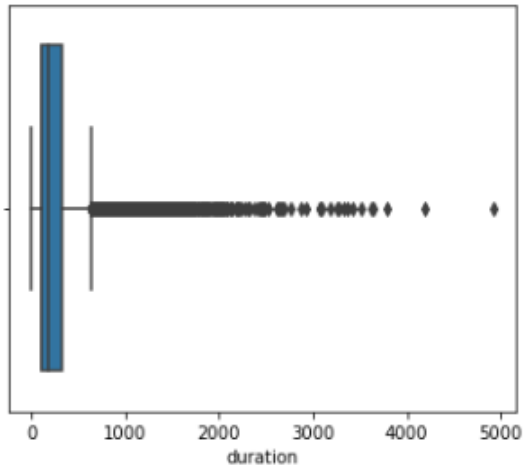
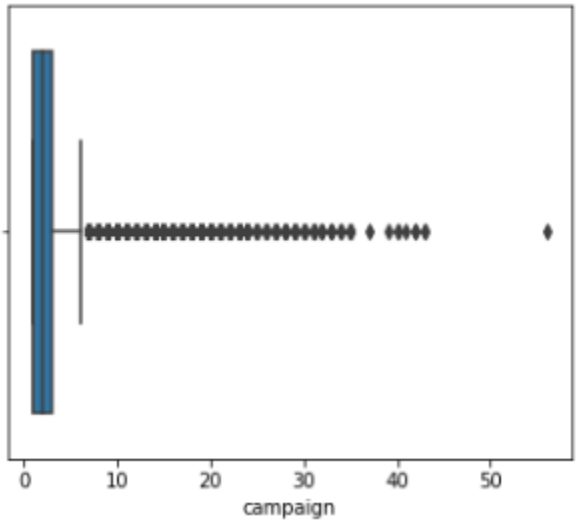
Using visualization (boxplot) we will determine which variables have outliers and then using IQR technique, we will remove/ round those values. (Q1 - 1.5IQR and Q3+1.5IQR)

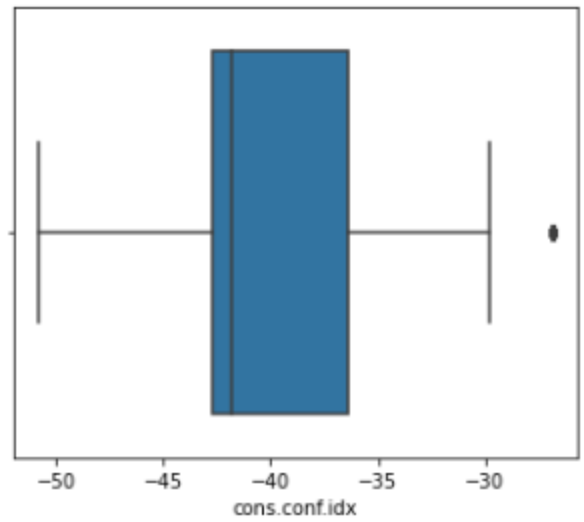
We create a common function to create a box plot graph.

```
fig,axes = plt.subplots(4, 3, figsize=(18, 20))
fig.suptitle('Boxplot graph')
count = 0;
for i in range(4):
    for j in range(3):
        if( count < len(numerical_var)):
            sns.boxplot(ax=axes[i,j], x=numerical_var[count], data=df)
            count = count + 1
```

Using able function we are creating 4 X 3 graph of box plot. And then we take decision based on that graph.

	As we see in the box plot graph of age, we can see there are many values after 70. But we have considered 95% as cutoff to truncate higher value. We used IQR technique to find upper 5% age group people and then we dropped those values.

	For the duration, by looking at boxplot we took 3500 as our threshold value and those values which are more than 3500, we replace it with 3500. This way we have handled duration value.
	For the campaign, we have decided to take 35 as threshold value. Same as duration, here also we are replacing value which is more than 35 to 35.

	For this feature, it is clearly visible that the last point is outlier. So we have replace it with -30 (upper value of upper whisker)

So this way we have handled the outlier in our numerical data sets.

Another method which we can use to handle the outliers in the campaign and age attribute is by transforming the numerical features using the log function.

```
In [269]: # There are several outliers in the age and campaign column. This can be handled by transforming them
          # Transorming the outliers using log
          bank_data['log_age'] = [np.log(x) for x in bank_data['age']]
          bank_data['log_campaign'] = [np.log(x) for x in bank_data['campaign']]
```

```
In [270]: bank_data
```

ult	housing	loan	contact	month	day_of_week	...	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y	log_age	log_campaign
no	no	no	telephone	may	mon	...	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no	4.025352	0.000000
wn	no	no	telephone	may	mon	...	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no	4.043051	0.000000
no	yes	no	telephone	may	mon	...	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no	3.610918	0.000000
no	no	no	telephone	may	mon	...	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no	3.688879	0.000000
no	no	yes	telephone	may	mon	...	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no	4.025352	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
no	yes	no	cellular	nov	fri	...	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	yes	4.290459	0.000000
no	no	no	cellular	nov	fri	...	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no	3.828641	0.000000
no	yes	no	cellular	nov	fri	...	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no	4.025352	0.693147

Although we noticed that the values including the inliers were transformed and this will affect the performance of The Machine Learning model. So it is better to use another method to handle outliers in the numerical variables.

References

1. How to deal with outliers in python -. ProjectPro. (n.d.). Retrieved September 2, 2022, from https://www.projectpro.io/recipes/deal-with-outliers-in-python
2. *Detecting and treating outliers: How to handle outliers*. Analytics Vidhya. (2022, July 21). Retrieved September 2, 2022, from https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/%20
3. Handling outliers in python. Handling Outliers in Python. (n.d.). Retrieved September 2, 2022, from https://www.datasciencesmachinelearning.com/2018/11/handling-outliers-in-python.html