Submit a pdf document which should contain following details:

Team member's details : Group Name (give a name to your group), Name, Email, Country, College/Company, Specialization (Data Science, NLP, Data Analyst)

Problem description

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

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

GitHub Repo link

MuhammedZek/MBA-Data-Science-Project (github.com)

Group Name : MBA

Members

Name	Email	Country	College
Alhamza Ibrahim	Hamzai7brahim@gmail.com	Turkey	Turkish-German University
Muhammed Zekeriya	mohammed.mbz.96@gmail.com	Turkey	Turkish-German University
Bilal Yildiz	bilalo.gg@gmail.com	Turkey	Turkish-German University

Specialization : Data Science

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)

Data understanding

we have here 41188 Row and 21 Columns without Null values including bank client data, social and economic context attributes and some other attributes. We can see that the data have three types integer data type, float data type and categorical data type including target variable which is named as ”y” in the dataset.

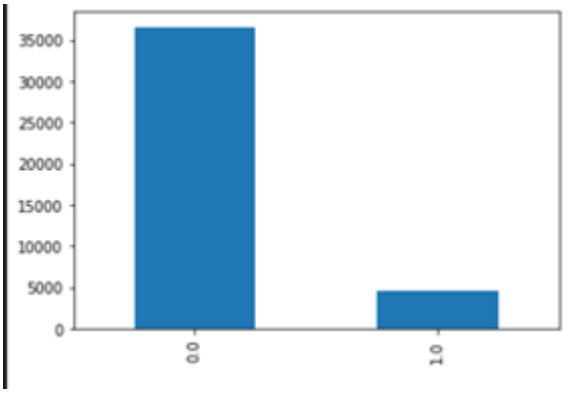
What type of data you have got for analysis

We most likely have Predictive data analytics here because we seek to predict what is likely to happen in the future. Based on past patterns and trends, and that match the data that we have here(customers information) and that is especially useful as it enables businesses to plan ahead like selling a deposit product as in our case.

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

One of the main problems that we encountered while analyzing the data set is that the data is imbalanced (89% No and only 11 Yes)

	count	percentage
0.0	36548	0.887346
1.0	4640	0.112654



To overcome this problem we tried three of the most common algorithms to handle imbalanced Data (undersampling , oversampling and oversampling with SMOTE)

And the final data set is now shaped as follow:

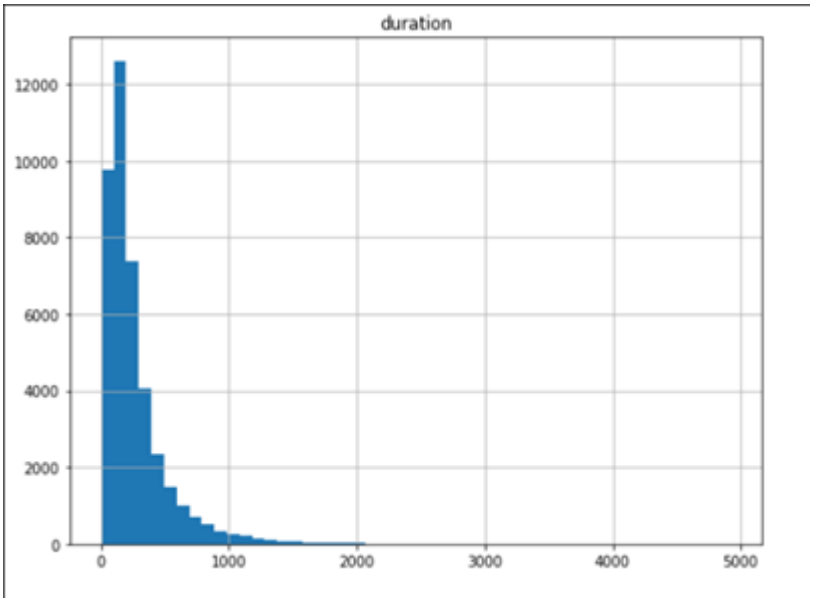
	method	X-shape	y_shape	y_yes	y_no
0	no method applied	41188	41188	4640	36548
1	underSampling with NearMiss	9280	9280	4640	4640
2	oversampling with RandomOverSampler	73096	73096	36548	36548
3	SMOTE	73096	73096	36548	36548

We will ignore the undersampling methods for now because the bank wants to maintain as much data as we can so for the next steps(training the model.. etc.) we will be dealing with the oversampled data and we will compare the results and choose the best data.

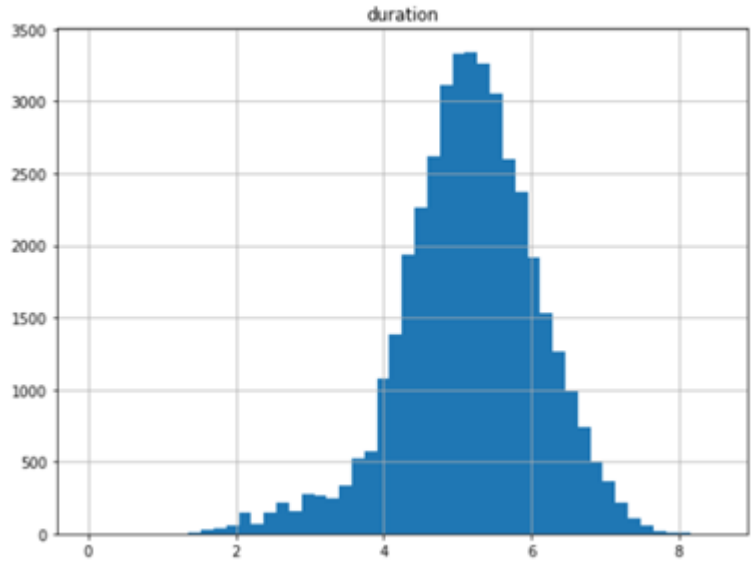
Another problem is the skewness of data especially in the Duration, Campaign and Pdays columns (where the skewness is bigger than 1 or smaller the -1)

```
df.iloc[: , 10:13].skew()
[181] ✓ 0.1s
... duration    3.263141
    campaign    4.762507
    pdays      -4.922190
    dtype: float64
```

For example here is the graph that shows the distribution of data in the duration columns before applying any method:

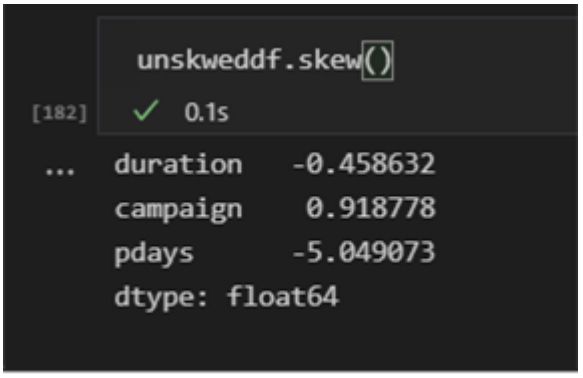


And here's after applying Log Transform method :



We can clearly see how the data is now almost normally distributed.

and now the final results :



The Log Transform methos did a great job dealing with Duration and Campaign columns but the Pdays columns got worse than before! Thus we will apply another methods like Square root and coxbox methods to deal with it.