Data Glacier

Week 9: Final Project

Bank Marketing Campaign

- Data Science -

Table of Contents

Team member's details:

	Group Name: *Data Science Enthusiasts*				
	Name	**Email**	**Country**	**College/Company**	**Specialization**
1	Amira Asta	amira.asta02@gmail.com	Tunisia	Afrikanda	Data Science
2	Vatsal Vinesh Mandalia	vatsalvm10@outlook.com	Oman	Graduated	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 helps them understand whether a particular customer will buy their product or not. In order to achieve this task, they approached an Analytics company to automate this process of classification. The Analytics company has given responsibility to the **Data Science Enthusiasts** Team and has asked to come up with a ML model to shortlist customers whose chances of buying the product is higher, so that ABC's marketing channel can focus only on those customers.

Business understanding:

There has been a revenue decline for an ABC bank and they would like to know what actions to take. After investigation, they found out that the root cause is that their clients are not depositing as frequently as before. Knowing that term deposits allow banks to hold onto a deposit for a specific amount of time, banks can invest in higher gain financial products to make a profit.

In addition, banks also hold better chances to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. As a result, the ABC bank would like to identify existing clients that have higher chances to subscribe for a term deposit and focus marketing efforts on such clients. The classification goal is to predict if the client will subscribe to a term deposit or not.

Data cleansing and transformation:

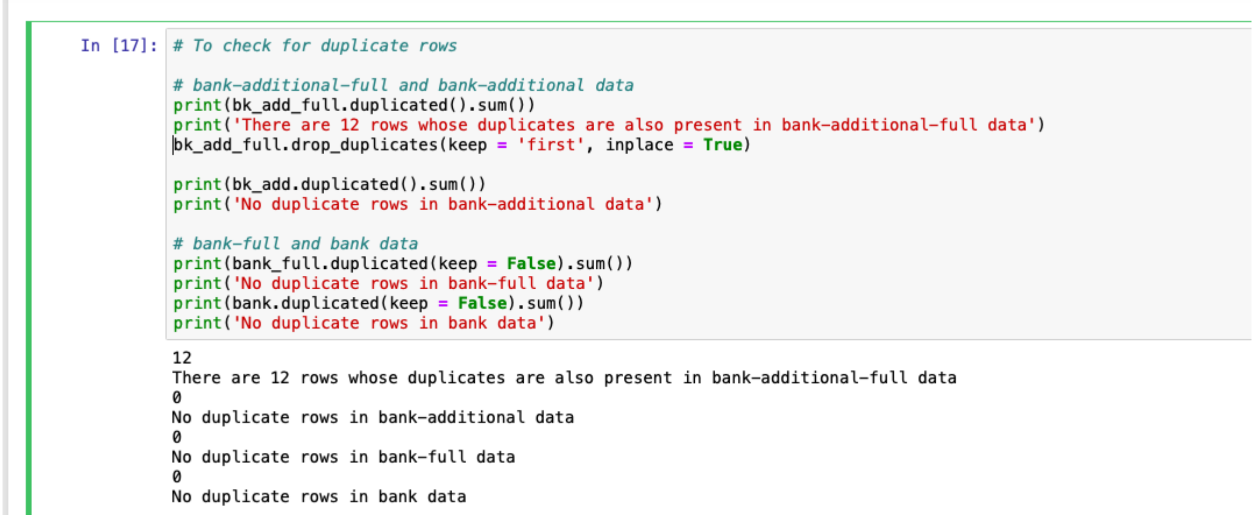
In this part we will mention the techniques used to reach the final data set that we will use for model development. The dataset 'bank-additional-full' is considered for data cleaning.

The dataset in question has multiple issues in it as given below.

- Duplicate observations.
- Missing value detection - Bar plots of categorical features.
- Outlier detection - Histograms and box plots of the features.

Duplicate observations:

The four datasets are checked for duplicate rows. The bank-additional-full data was found to have 12 duplicate rows. Using the drop_duplicates() method, these duplicates are dropped entirely.



```
# To check for duplicate rows

# bank-additional-full and bank-additional data
print(bk_add_full.duplicated().sum())
print('There are 12 rows whose duplicates are also present in bank-additional-full data')
bk_add_full.drop_duplicates(keep = 'first', inplace = True)

print(bk_add.duplicated().sum())
print('No duplicate rows in bank-additional data')

# bank-full and bank data
print(bank_full.duplicated(keep = False).sum())
print('No duplicate rows in bank-full data')
print(bank.duplicated(keep = False).sum())
print('No duplicate rows in bank data')
```

```
12
There are 12 rows whose duplicates are also present in bank-additional-full data
0
No duplicate rows in bank-additional data
0
No duplicate rows in bank-full data
0
No duplicate rows in bank data
```

Missing value detection:

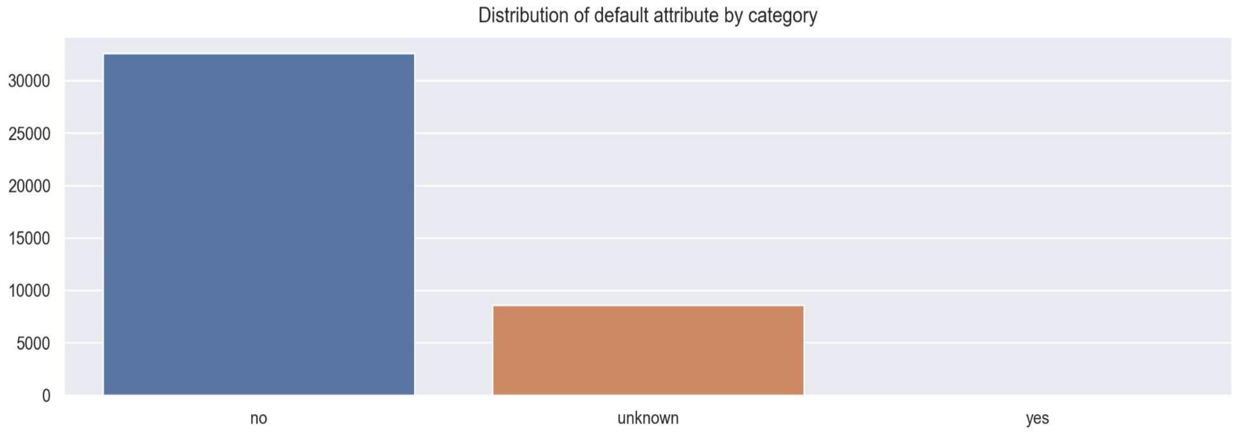
To carry out missing value detection, univariate analysis of the numeric and categorical features is done.

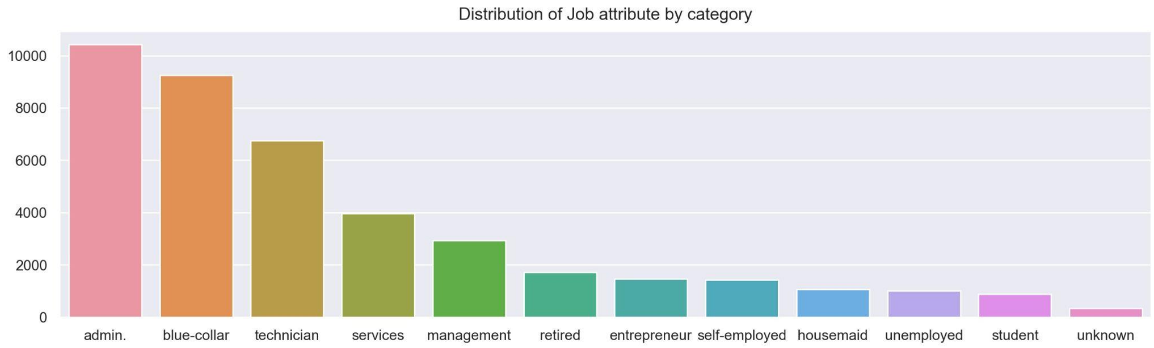
- **Imputation of categorical features:**

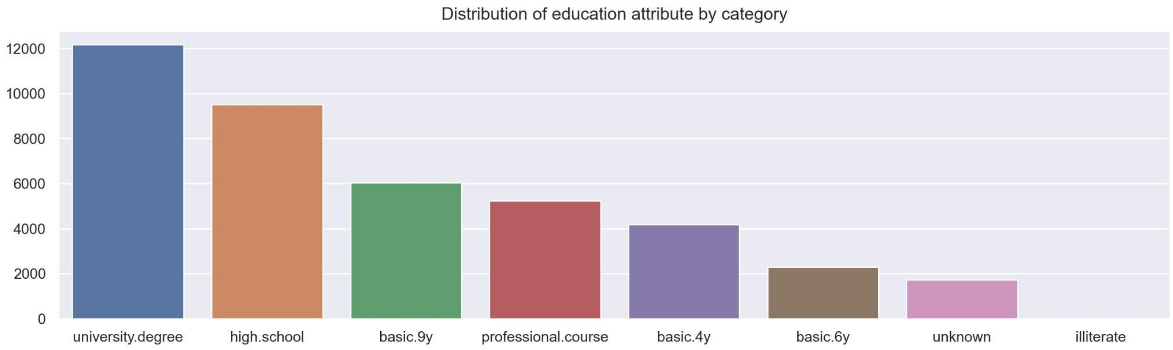
There are ‘unknown’ values for many variables in the data set. There are many ways to handle missing data. One method is to discard the rows but that would lead to reduction of the data set and hence would not serve our purpose of building an accurate and realistic prediction model.

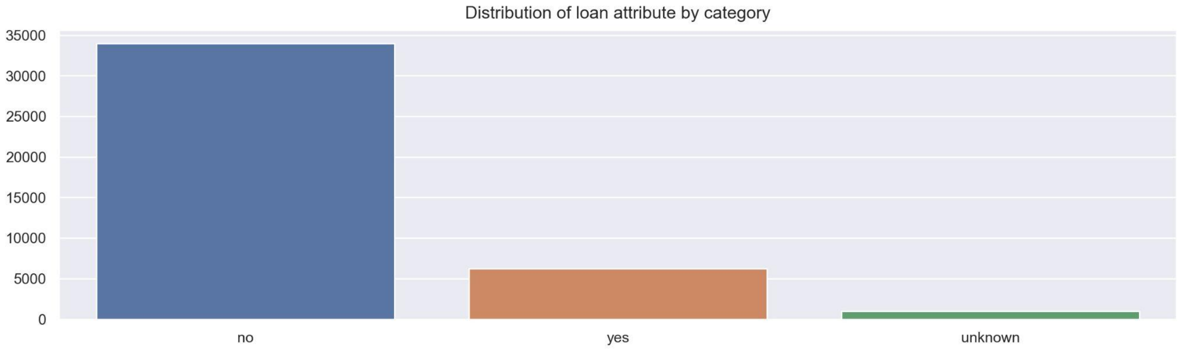
Another method is to infer the value for ‘unknown’ from the particular feature/column. Using the most common category of that categorical variable to fill the missing values is one way of imputation. This doesn't guarantee that all missing values will be addressed but the majority of them will have a reasonable value which can be useful in the prediction. One more issue with this method is this introduces bias in the data.

Variables with ‘unknown’ values are : 'education', 'job', 'loan', 'default', and 'marital'. But the significant ones are 'education', 'job', ‘default’, and 'loan'.







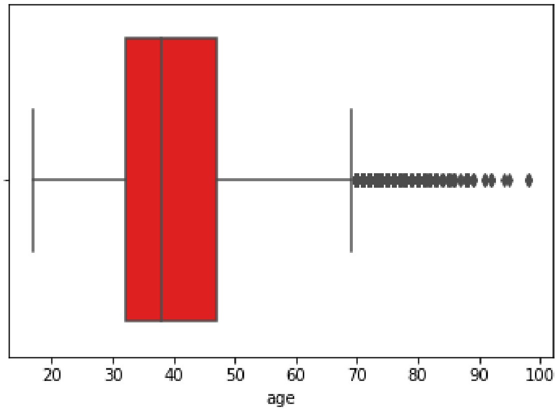


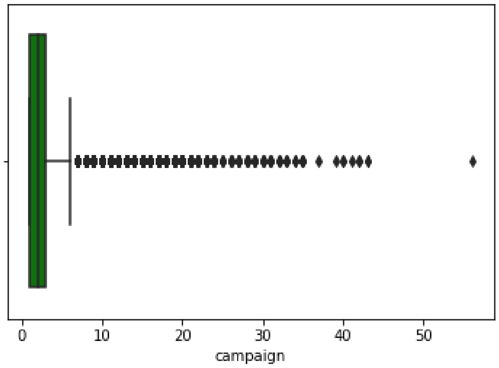
The other way to address the ‘unknown’ values in the categorical columns is by considering them as another category of the features. For example, on looking at the bar plot of ‘default’ feature, there are approximately 9000 clients with a response of ‘unknown’. This means these clients do not want the bank to know their actual default status. Therefore this value can be a good addition in analysing the customer behaviour.

Outlier detection:

Outliers represent those values which are at an abnormal distance from the central distribution of the data points. These values affect the summary statistics of the data, prominently the mean and mode. Hence, it becomes important to deal with outliers in the data cleaning stage so that the performance of the regression model doesn't get compromised.

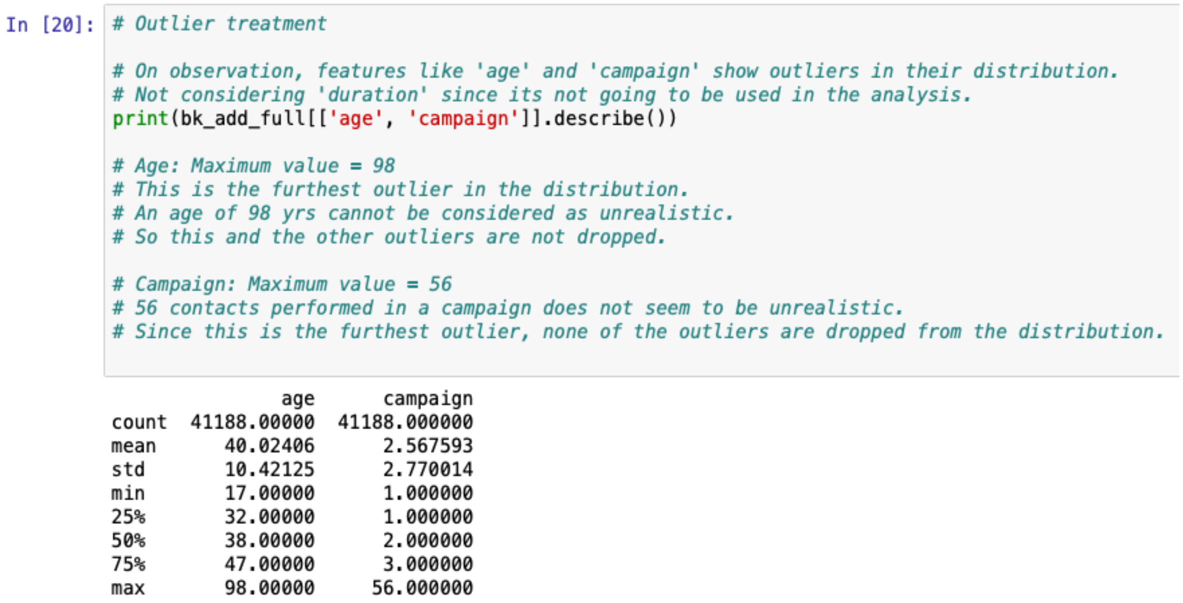
There are few techniques involved in outlier treatment. One of them is totally dropping the outliers from the dataset. This however would lead to loss of important data which can help in giving realistic predictions from the model. The outliers can be detected by creating univariate plots like histogram or box plot of the numeric features. For example, a box plot of the 'age' and 'campaign' features is visualized.





The data points outside the whiskers of the box indicate there are a significant number of outliers in this distribution.

On a detailed examination, we looked at the maximum values of the ‘age’ and ‘campaign’ variables.



```
In [20]: # Outlier treatment

         # On observation, features like 'age' and 'campaign' show outliers in their distribution.
         # Not considering 'duration' since its not going to be used in the analysis.
         print(bk_add_full[['age', 'campaign']].describe())

         # Age: Maximum value = 98
         # This is the furthest outlier in the distribution.
         # An age of 98 yrs cannot be considered as unrealistic.
         # So this and the other outliers are not dropped.

         # Campaign: Maximum value = 56
         # 56 contacts performed in a campaign does not seem to be unrealistic.
         # Since this is the furthest outlier, none of the outliers are dropped from the distribution.
```

```
                age      campaign
count  41188.00000  41188.000000
mean      40.02406      2.567593
std       10.42125      2.770014
min       17.00000      1.000000
25%       32.00000      1.000000
50%       38.00000      2.000000
75%       47.00000      3.000000
max       98.00000     56.000000
```

The maximum values of ‘age’ and ‘campaign’ in bank-additional-full are 98.0 and 56.0 respectively. An age of 98.0 yrs and the number of contacts performed in the campaign being 56.0 don’t seem to be unrealistic. Therefore, the outliers in the data distribution of these features are not dropped.

Github Repo link:

https://github.com/AsAmira02/Bank-Marketing-Campaign-DSEnthusiasts2021

This repository includes the four datasets, model code and necessary files used in this project.