Week 10 Deliverables

Group Name: Carpe-Diem group

Specialization: Data Science

Project Name: Bank Marketing (Campaign) -- Group Project

Team Members:

1. **Name**: Mohini Kalbandhe
   - **Email**:amohini099@gmail.com
   - **Country**: Canada
   - **Company**: Happy Orchard
   - **Specialization**: Data science

2. **Name**:Kashish Joshipura
    - **Email**:kashishjoshipura@gmail.com
    - **Country:**Canada
    - **Collage**: The University of British Columbia (UBC)
    - **Specialization**: Data Science

3. **Name**:Amir Shahcheraghian
    - **Email**:Amir.shahcheraghian@gmail.com
    - **Country:**Canada
    - **Collage**: University of Quebec , Canada
    - **Specialization**: Data Science, Energy Management analysis
4. **Name: Mohammed Maqsood**
    - **Email:**mohammedmaqsood48@gmail.com
    - **Country:** Germany
    - **College**: Otto von Guericke University
    - **Specialization:** Chemical and Energy Engineering

Bank Marketing (Campaign)

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).

Project Objective:

By converting this problem into a machine learning classification problem we will build a model to predict whether a client will subscribe a term deposit or not so that the banks can arrange a better management of available resources by focusing on the potential customers “predicted” by the classifier .

Technique to be used: Classification

Data cleansing and transformation:

Dataset has been checked and reviewed by all the peers and concluded that there were no missing values found in the dataset however we found some “unknown” and “other” values which needs to be converted to numerical values or needs to be removed in order to clean the dataset. No “skewness” in dataset was found, data seems symmetrical dataset. Statistics summery such as slandered deviation, distribution and skewness has been checked.

EDA performed on the data:

Exploratory Data Analysis refers to the bank dataset provides critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Descriptive analysis (univariate analysis)

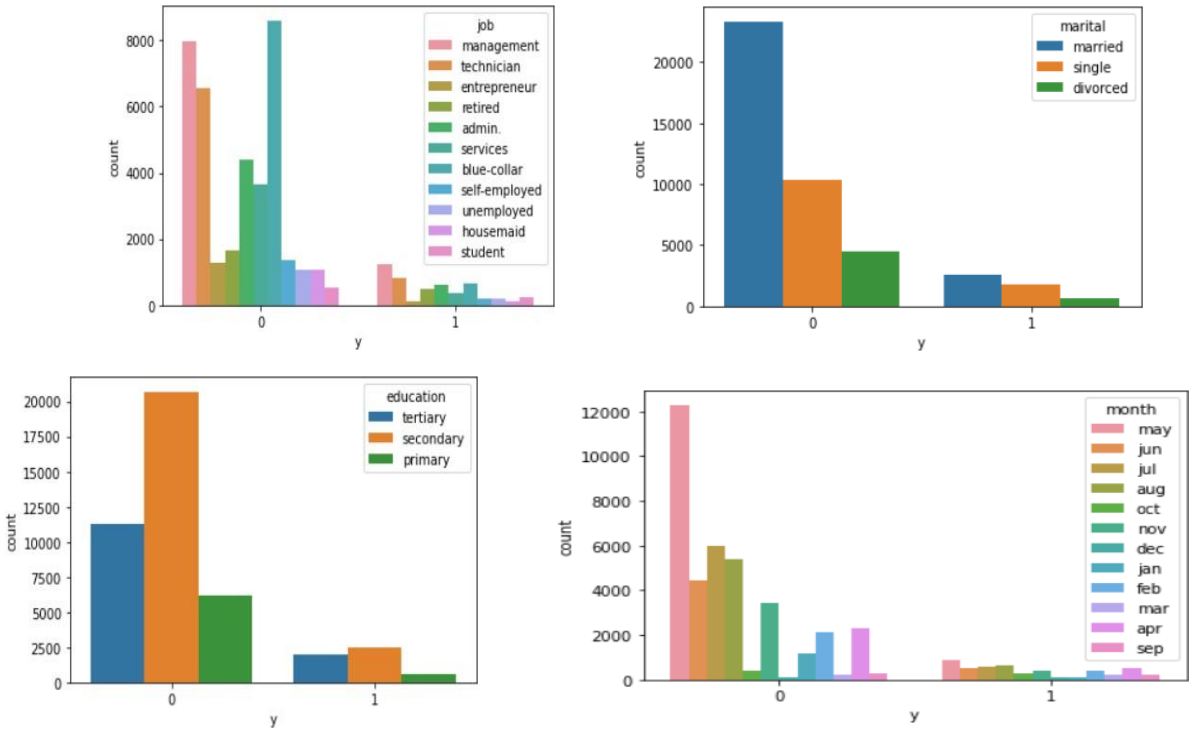
- While performing this analysis we provides an understanding of the characteristics of each attribute of the dataset which is an important evidence for feature selection.
- There were no missing values but found some “unknown” values, we decided to drop the outliers and ambiguous values, such as "others" and "unknown".
- The columns which has two values('yes' and 'no') and slightly imbalanced such as default, loan, y, has been converted to (1,0) numerical values. rest are continuous variable were binned so that outliers value are converted into count values.
- Skewness doesn't provide much insights in data, as values of columns are nearly zero apart from 'previous'. Data seems symmetrical.
- Flooring and clapping using interquintile range(IQR) Outliers are removed by dropping values that is below 25% and 75% percentile.
- We classified dataset into numerical and categorical attributes.

Numerical Attributes:

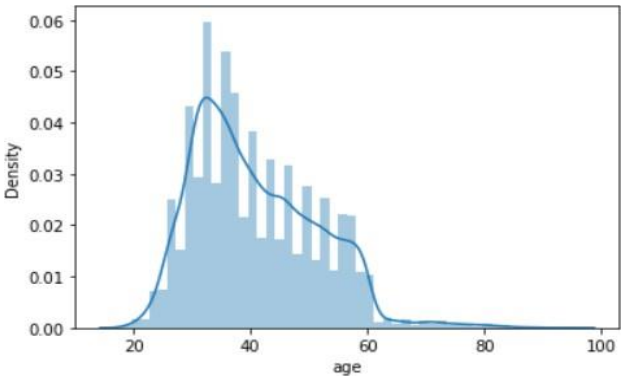
Following table provides statistical information in descriptive analysis.

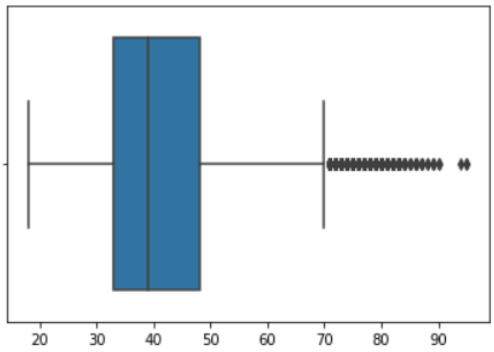
	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

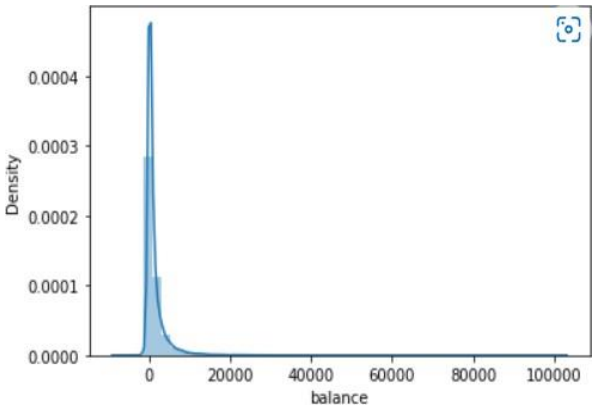
Following are some data visualization performed on dataset which provides the details of Categorical attributes

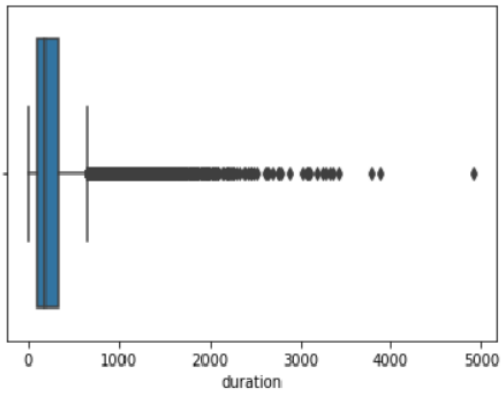


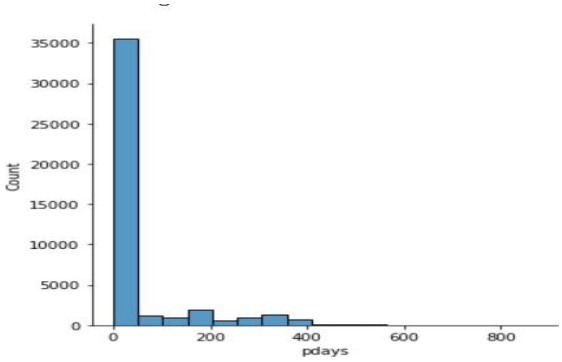
Following histograms display the numerical attributes like “age”, “balance”, ”duration”, “pdays”,” campaign”, etc.

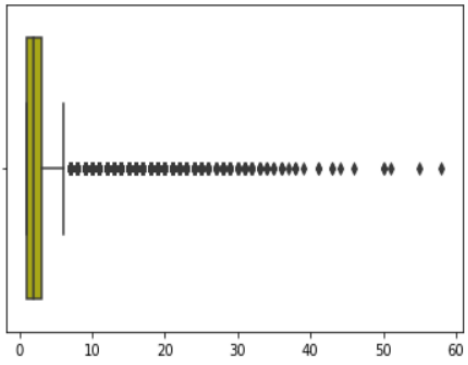












Computed skewness of numerical attributes and data shows symmetrical from following results.

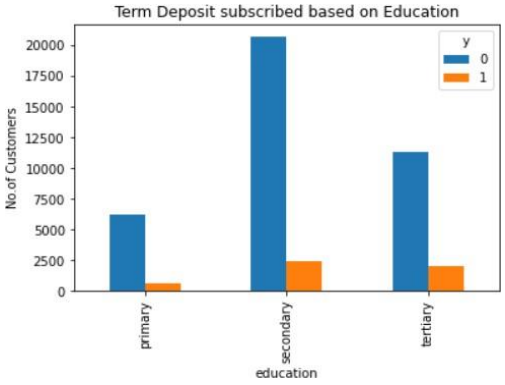
```
cols=["age","duration","campaign","pdays","previous","balance"]
for item in cols:
  print(f'Skewness {item}: {str(bank_df[item].skew())}')
```

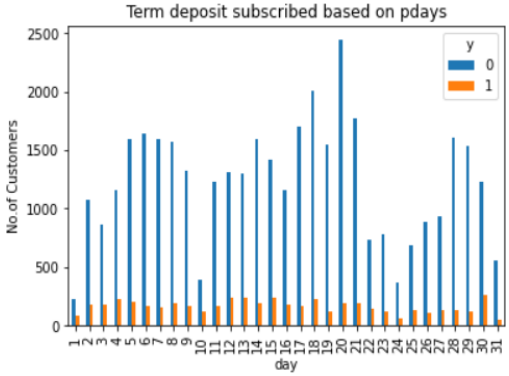
```
Skewness age: 0.6978356364509636
Skewness duration: 3.1701799697784785
Skewness campaign: 4.7924941810208885
Skewness pdays: 2.608337543002269
Skewness previous: 42.08877792244101
Skewness balance: 8.400120937754398
```

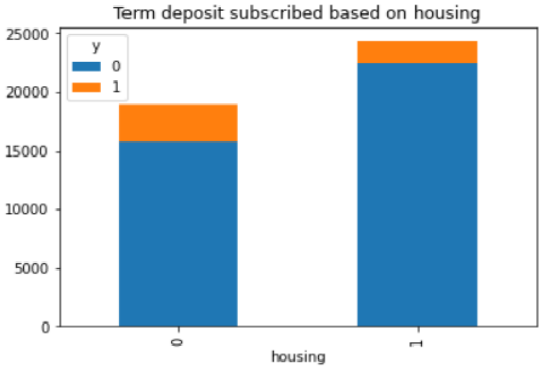
Following data visualization shows that term deposite subscribed and how many customer were approched based on different parameter.

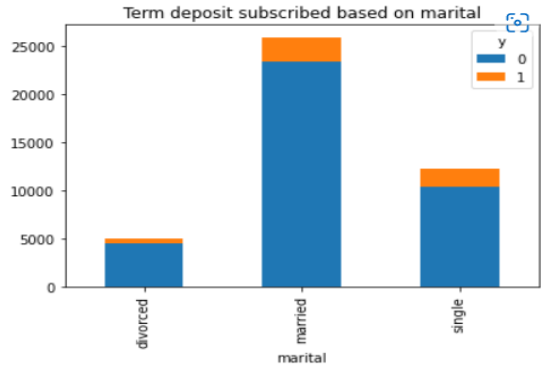
- Job : Highest Number of application are from admin type of job.
- Marital: most of the clients approached were married.
- Education: Client with university degree and high school were approached more as compare to other and they have higher success rate compared to others. default: it doesn't shows much impact.
- Housing: Housing loan does not have much effect on the number of term deposit purchased.
- Loan: most of client with not having personal loan were approched most.
- Contact: Around 64% calls are from cellular.
- Month: Around 33% were approached in may and in January, Febuary we don't have data or no one was approached. Success rate was almost same in june, july and August.

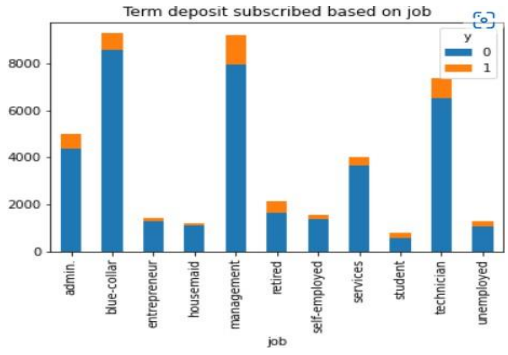
- day_of_week : We have 5 days collected values. There is no significant different in the number of client approached and number of people subscribed. So we will drop this feature.
- poutcome: If a client took the term deposit last time than there is higher chances of that client subscribing to it again.

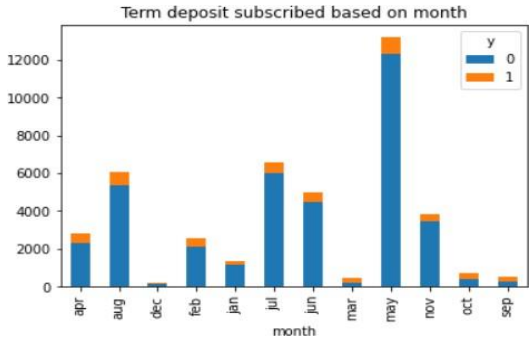


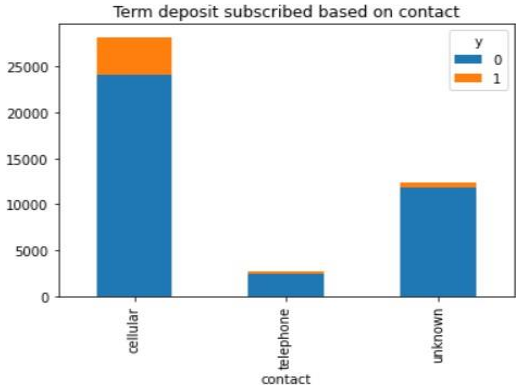


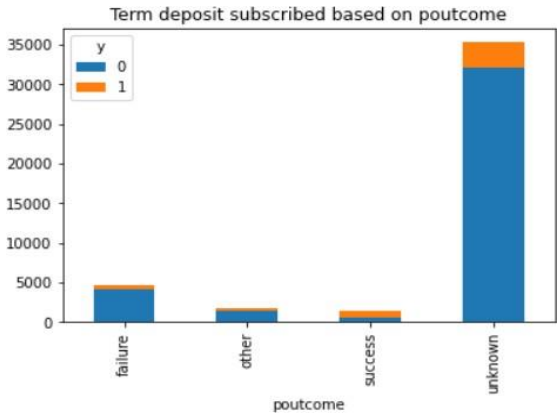


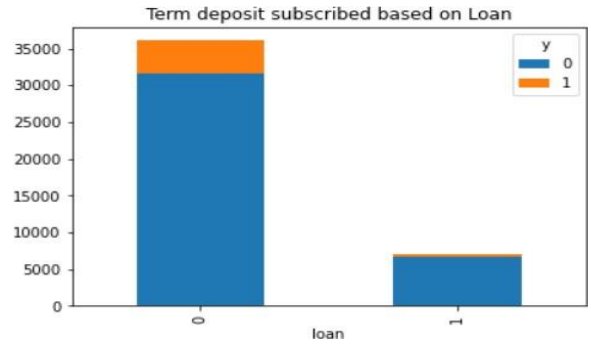


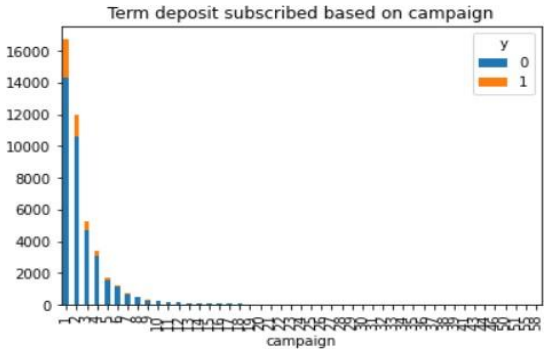




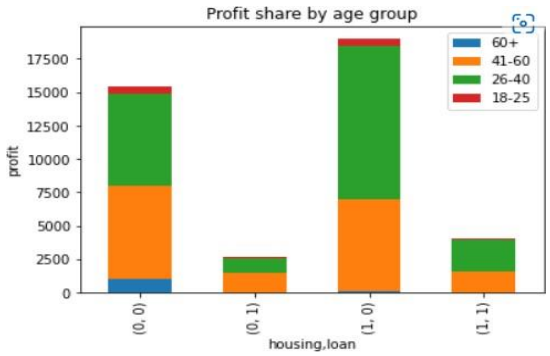


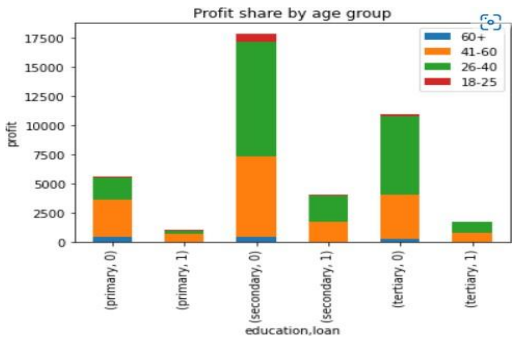






Following bar chart shows in which age range the “education” and “housing” loan is more.

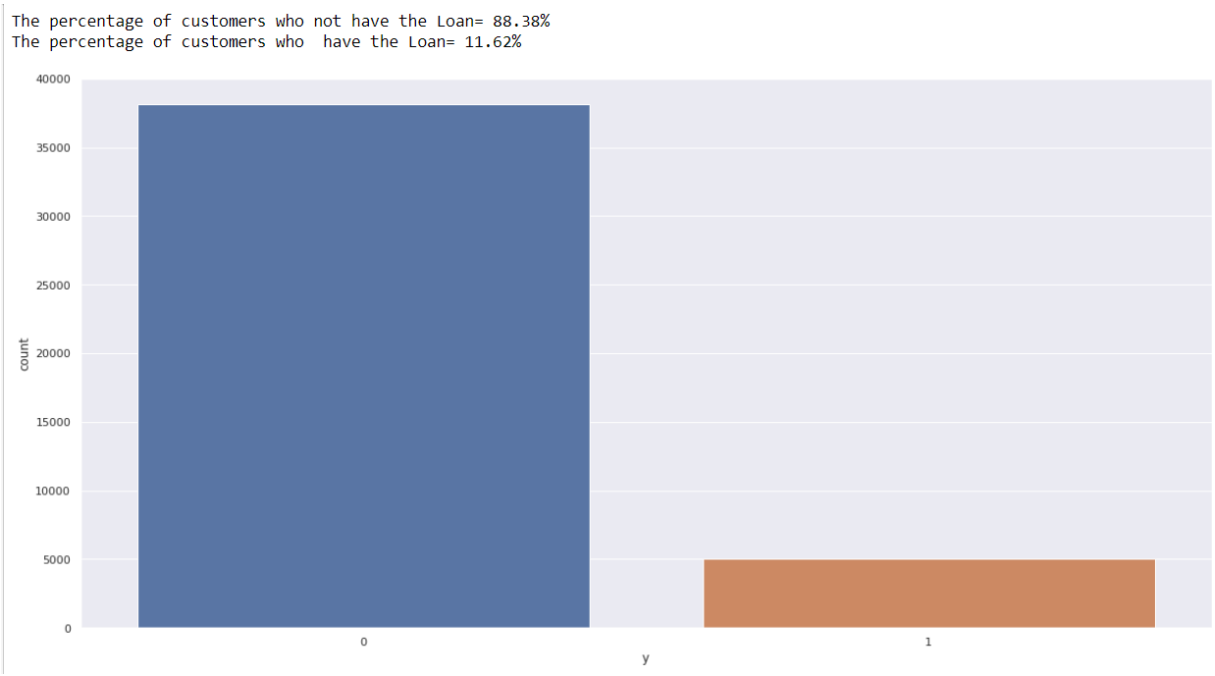




From the bar charts we can conclude that the age range 26-40 has taken more loan than other age groups.

Feature selection based on descriptive analysis:

We have provide basic statistics of each attribute in the dataset, based on this some of problems we have identified such as imbalance of categorical target. In this dataset there are higher number of instances of major class and fewer number of instances of minor class as shown in the bar chart below.



So we decided to use undersampling method to delete the samples in majority class as there is a huge difference between y=1 and y=0, So the ML algorithm will omit the smaller value, which may affect the performance of algorithm.

Feature selection based on correlation analysis (bivariate analysis):

After performing undersampling method and removal of highly positively correlated features we got the features useful in terms of predicting desired target.

Quantitative analysis:

We also tried to perform chi squared test, The comparison is deemed statistically significant if the relationship between the categorical attribute “marital” where we got to know that P-value of 1 depicts that there is no difference in value of the groups other than due to chance.

Whereas in Z-test, we found Z-stat is less than Z-critical we accept the null hypothesis test.

Final Recommendation:

After performing EDA on dataset, we have realized the importance of feature engineering and repeated testing in order to find best parameters; We have understood how hard it is to get an increase of 1% in scores. We learnt how to do feature engineering and efficiently find the best hyperparameters for classifiers. As problem is classification problem we recommend using sklearn.accuracy as the evolution metric. We recommend to use the area under the curve (AUC) and F1 score (average of precision and recall)as a metric.

Github Repo link:

“https://github.com/amohini099/Banco-de-portugal-marketing/tree/main/week10”