Project Documentation

Project Name: **Portuguese Bank Marketing(PRCP-1000)**

Project Scope Statement:

The Project is related with direct marketing campaigns via phone calls of Portuguese Banking Institution.

Business Requirement:

The classification goal is to predict, if the client will subscribe to term deposit (Yes/No).

Raw Data:

The given raw data contains following features, as given below:

Numerical Data: 10 Features

Age, Duration, Campaign, Pdays, Previous, Emp.var.rate, Cons.price.idx, Cons.conf.idx, Nr.employed, euribor 3m.

Categorical Data: 10 Features

Job, Marital, Education, Default, Housing, Loan, Contact, Month, Day_of_week, Poutcome.

Target variable: 1 Feature

Y – The client subscribed a term deposit binary (Yes/No).

This is about the raw data of the project and the next step is to do insight the raw data. The step as follows,

1) Data Cleaning

2) Exploratory Data Analysis (EDA)

3) Data Pre-processing

4) Modelling

Data Cleaning

In Data cleaning, what we do is just checking the null values or Nan values, missing values, duplicate values, skewness, checking Outliers using the codes.

And what we concluded in the data cleaning step,

1) There is no null/nan values

2) There is no missing data

3) There is 12 duplicate values are there and we removed it.

4) The target variable (Y) can either be ‘yes’ or ‘no’. In the given dataset 87% of tuples belong to the class ‘no’. Therefore the data is **highly skewed**. To make the data balanced, we took the tuples belonging to minority class (i.e. Class ‘yes’) and duplicated them until both the classes have an equal number of tuples in them.

5) And finally checking outliers to the columns – “age” & “duration” using IQR method and we are not removed the outliers due to loss of data. We just impute the upper fence value in place of outliers.

Finally we completed data cleaning section successfully.

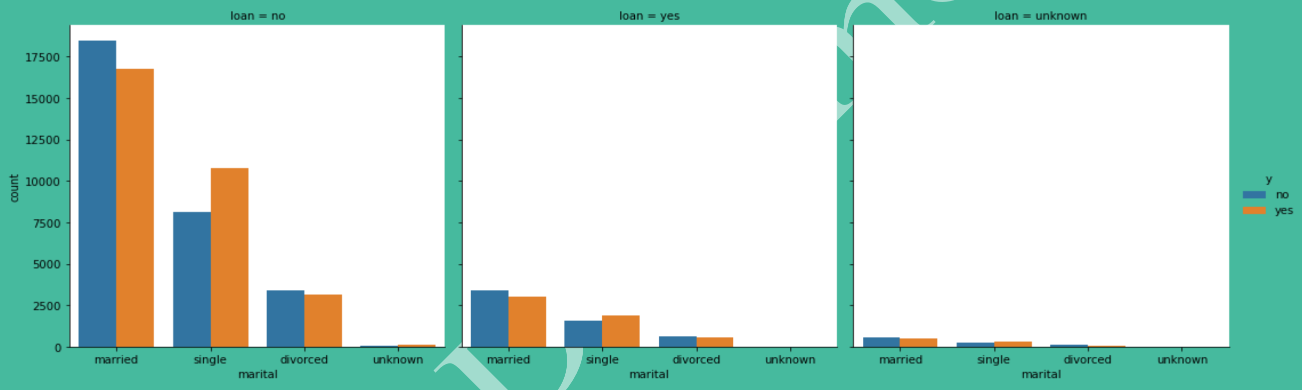
Note: Outliers treatment must be done on continuous data not to the categorical data.

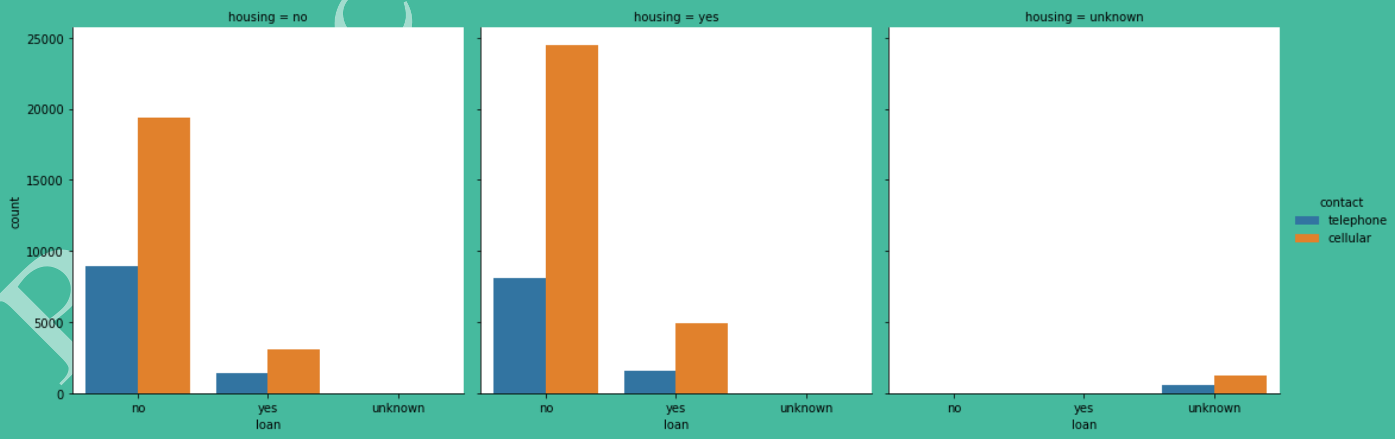
Exploratory Data Analysis (EDA)

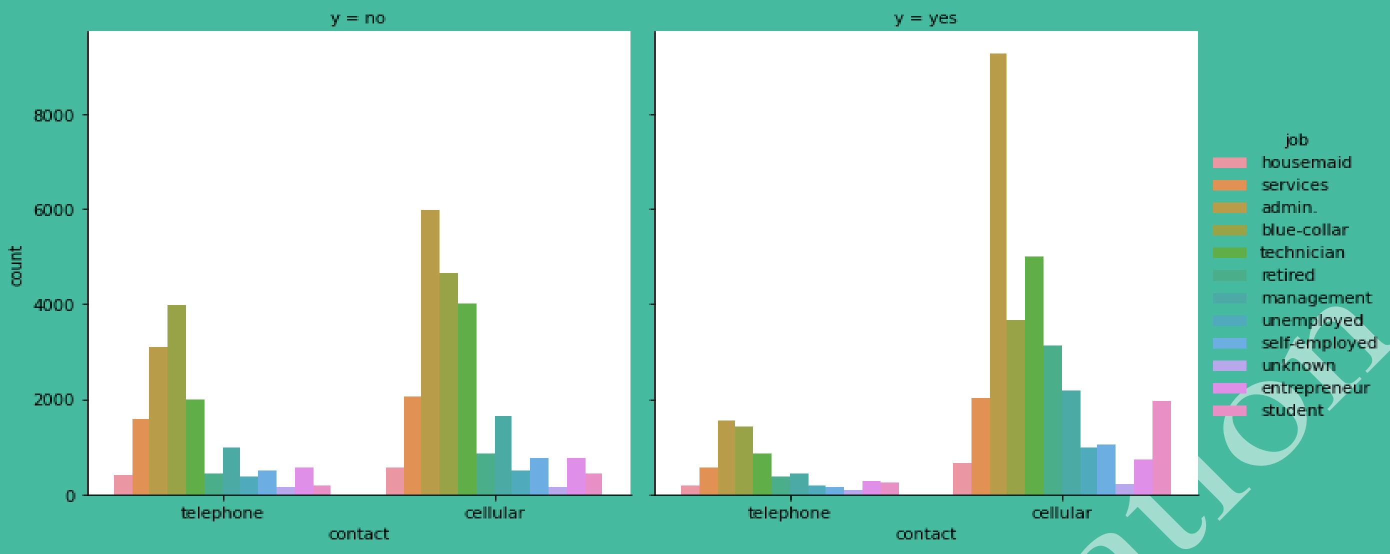
EDA is one of the crucial step in data science that allows us to achieve certain insights and statistical measure that is essential for the business continuity and data scientists. It performs to define and refine our important features variables selection that will be used in our model.

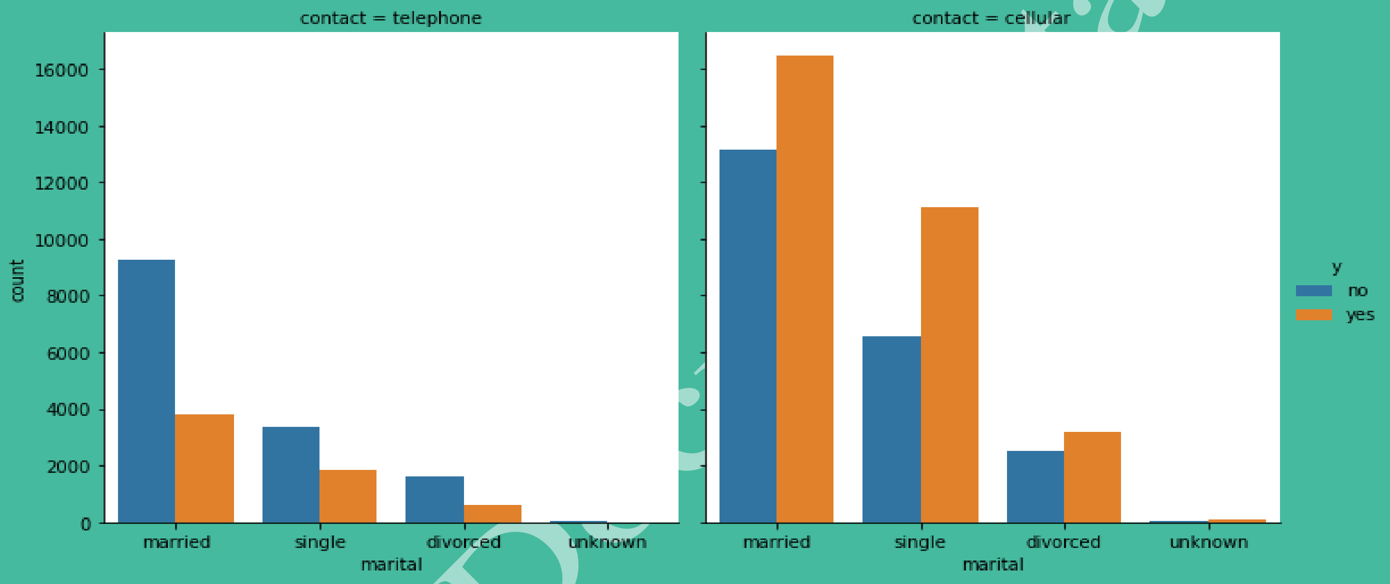
Once EDA is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modelling.

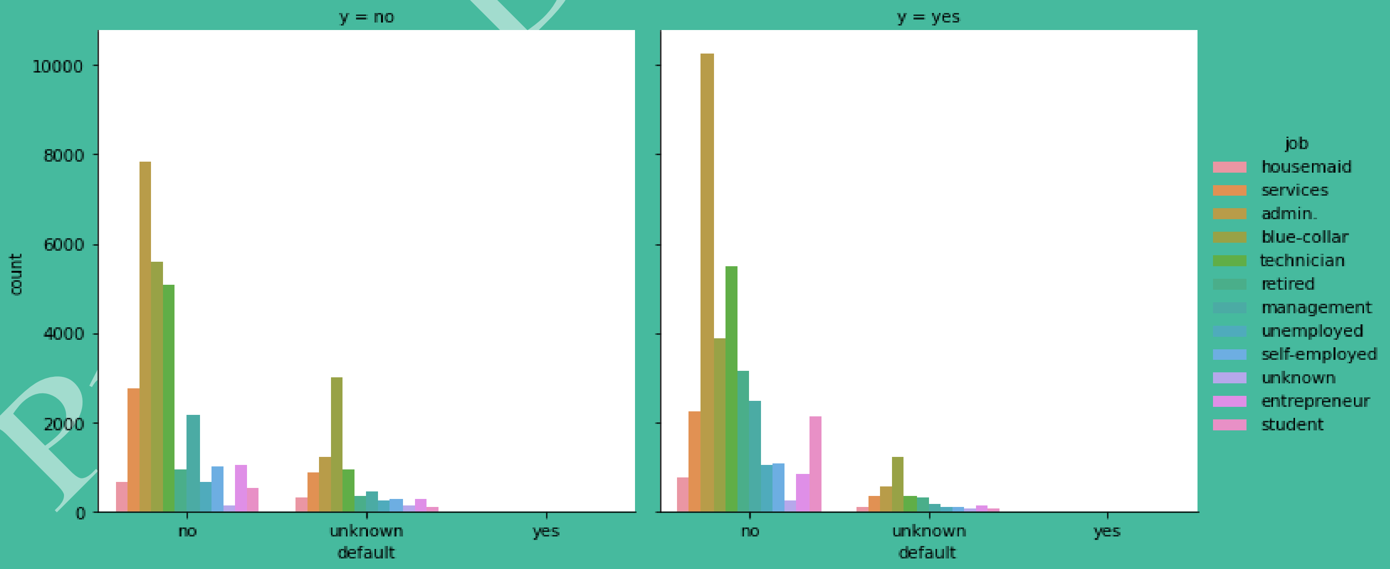
For this model we did few plots to the features, Job, marital, education, contact, Y (target variable), Defaulters, Housing, and Loan. The plots are shown below,

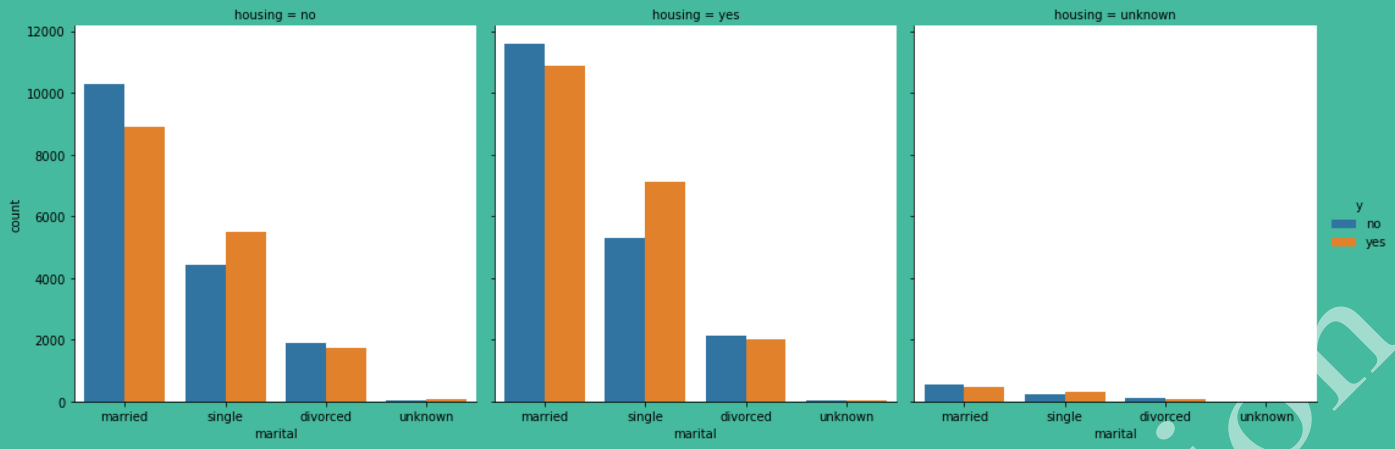


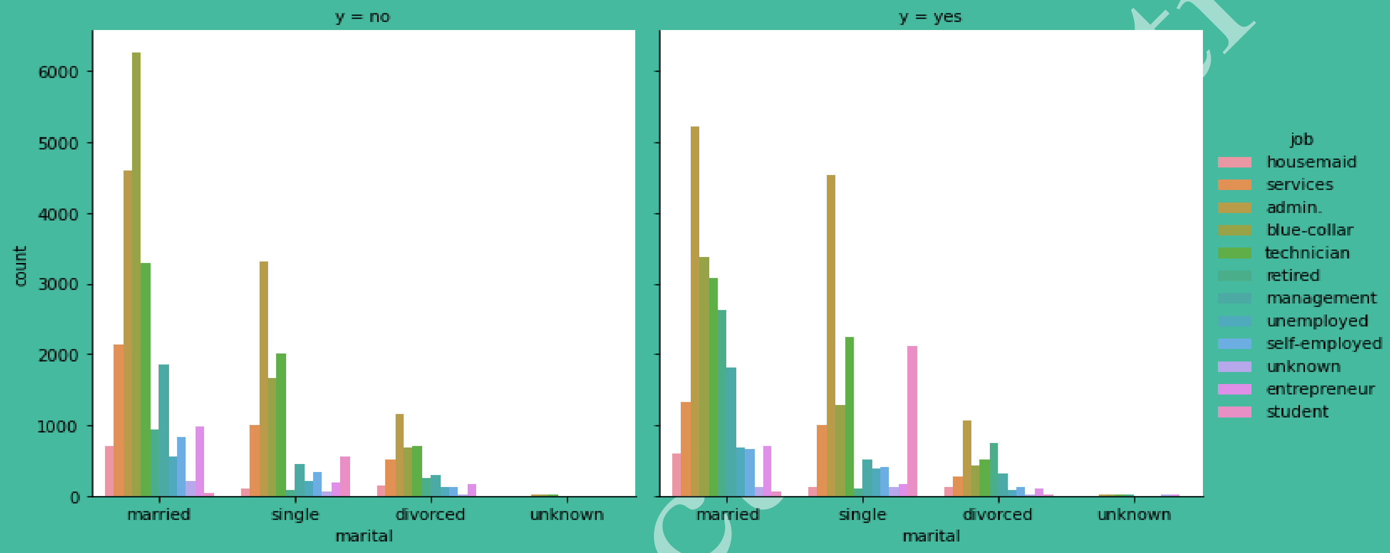



y = no
y = yes
count
8000
6000
4000
2000
0
telephone
cellular
contact
job
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student


contact = telephone
contact = cellular
count
16000
14000
12000
10000
8000
6000
4000
2000
0
married
single
divorced
unknown
marital
y
no
yes


y = no
y = yes
count
10000
8000
6000
4000
2000
0
no
unknown
yes
default
job
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student





These are the plots drawn in EDA part and what we conclude that,

Married/single persons with no loan, who are blue-collar (job) with housing and no credit values has more subscribed clients. And also cellular campaign got huge clients to subscribers compared to telephone.

Data Pre-processing

Generally a raw data contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is requires tasks for cleaning the data and making it suitable for machine learning model which also increases the accuracy and efficiency of a machine learning model.

For this dataset we did MinMax scalar, the data feature in the range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset. This scaling compresses all the inliers I the narrow range [0, 0.005]. We did scaling to few feature to get the data in limit.

Features: ‘age’, ‘pdays’, ‘previous’, ‘emp.var.rate’, cons.price.idx’, ‘cons.conf.idx’, euribor3m’, ‘nr.employed’.

Label Encoding:

Label encoding is also one of the step in data pre-processing, it assigns a unique numbers (starting from 0) to each class of data. This may lead to the generation of priority issue in training of data sets. A label with high value may be considered to have high priority than label having lower value. Label encoding is done to categorical variables.

Features: Job, housing, default, education, loan, contact, month, day_of_week, poutcome, y(Clients to depositors Y/N).

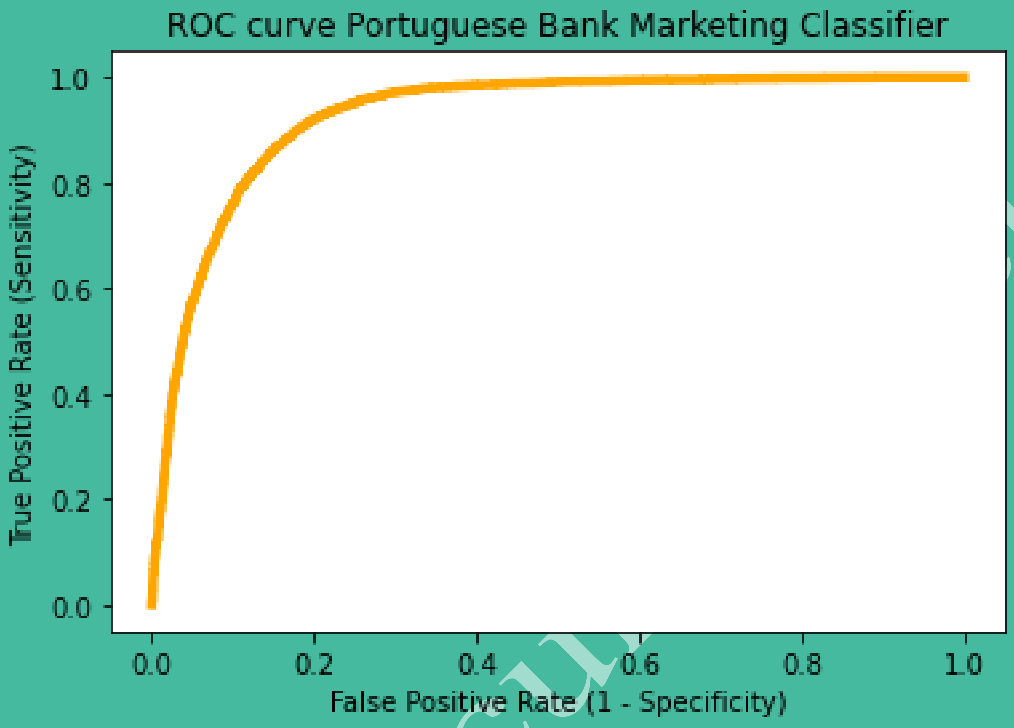
Modelling

The given dataset is a classification model, so for classification model we preferred “**Logistic Regression Model**”. It is the go to method for binary classification model. The given dataset target variable is binary(Y/N).The metrics as follows,

S.No	Metrics	Logistic Regression Model
0	Accuracy	0.857932
1	Precision	0.877477
2	Recall	0.845560
3	F1-Score	0.861223

Roc Curve and ROC_AUC_Score:

Roc curve is used to visualize how good your model at the final step is although it works for only binary classification problems.



The **roc_auc_score** is **0.9271353809471136** and it denotes that the model is good.

Project Members

1. **Kovuru Harun Rasheed**
2. **Laxman Singh**
3. **Pulugu Vamsi**
4. **Mogallapalli Manibhargav**