PRACTICAL DATA SCIENCE

PROJECT REPORT

Parvi Verma
S3744398
Masters of Data Science
Radhika Zawar
S3734939
Masters of Data Science

Contents

EXECUTIVE SUMMARY 2
RESEARCH GOAL 2
DATASET 2
TASK 1: DATA PREPARATION 2
TASK 2: DATA EXPLORATION 3
TASK 3: DATA MODELLING 8
CONCLUSION / RECOMENDATION: 11
REFERNCES: 11

EXECUTIVE SUMMARY

The objective of this Data Science Project is to study and analyse customer tendencies and other descriptive factors to identify potential costumer for the bank. This data set is a marketing data of a Portuguese banking institution. The intension behind investigation is to classify whether the customer will subscribe the term deposit or not. To solve this Binary Classification problem K-NearestNaighbor technique and Decision Tree technique has been used. This project consists of three Tasks: Task 1 focuses on preparation of data by checking for any inconsistencies, missing values and impossible values in data. Task 2 focuses on Visual Exploration of data to understand the features and their relationship with target feature. Task 3 focuses on using machine learning algorithms to classify whether the customer will subscribe the term deposit or not. The visualisations explain that:

- The dataset is imbalanced

RESEARCH GOAL

The goal of this Data Science Project is to select the best model using classification techniques to solve the binary classification problem stating that "classify whether the customer will subscribe the term deposit or not". To solve this Binary Classification problem K-NearestNaighbor technique and Decision Tree technique will be considered. Based on the performance and accuracy of each model, the best model will be chosen for this dataset which will solve the binary classification problem effectively.

DATASET

SOURCE: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

The dataset is Bank Marketing Data of a Portuguese banking institution. It is a direct Banking Campaign which is based on contact made with a customer as required in order to understand whether the product would be subscribed or not. The Target Feature of the dataset is a binary feature where "yes" means subscribing the term deposit and "no" means not subscribing the term deposit.

This dataset originally contains 41188 observations and 21 columns out which 11 columns contains Categorical values and 10 columns contains Numerical values. Only 10% of data is randomly selected for ease of analysis. The sample dataset contains 4119 observations and 21 columns.

Categorical Columns: 'job','marital','education','default,'housing','loan','contact','month','day-of-week','poutcome', `y'

Numerical Columns: 'age,'duration,'campaign','pdays','previous','emp.var.rate','cons.price.idx','cons.conf.idx', 'euribor3m','nr.employed'

TASK 1: DATA PREPARATION

The main objective of this step is to prepare the data for further analysis. The quality of data that we get out of this step will define the accuracy of our further analysis. The raw data can have many anomalies like missing values, typo errors, whitespaces, impossible values etc. This process rectifies all these anomalies and provides us with a good quality data to perform our further analysis on (Ren, 2019)

1.1) Checking the data received:

Sometimes due to error in loading or retrieving the data from the original file, the data that we get is not equivalent to the original data. To check that the loaded data is equivalent to original data or not. Steps performed:

- The first n rows of the data are displayed which confirms that the loaded data is equivalent to the original data.
- The data type for each column is displayed.
- All the column names are displayed in order to check them with the original file.
- The shape that is the structure or size of the dataset is checked which confirms that the data has 238 observations and 26 columns.

1.2) Types of errors:

Following are the types of errors that are taken in account in this investigation.

- **Redundant Whitespace and Typo Error-**

  During the process of data preparation, one can come up with a problem of string mismatch even though the strings look exactly the same. This problem may arise because of presence of redundant white space at start or end of the string whereas Typo error can be induced due to sloppiness of humans or any hardware failure. This type of error includes

lowercase and uppercase mismatch, some letters may switch places, typographical mistakes or an unusual font. (Ren, 2019)
Steps performed:

- First the frequency count table is created for each non numeric column with help of value_counts() function of pandas library to understand the frequency count of data values and redundant whitespaces.
- No Typo error or Whitespaces were found.

- **Missing Values-**

During the process of data preparation, dealing with missing values is an unavoidable step. Almost every data has missing values and there are many ways to deal with them - drop the observation with missing value, imputing the missing value with mean, median or mode for numerical data and padding the missing value with previous or next value for non-numeric data. Following are the steps taken to check Missing Values. (Ren, 2019)

- The isna() and sum() function of pandas library together is used to find out which columns have missing value.
- No Missing Values were found.

- **Impossible Values-**

Sanity Check is one of the important steps in process of data preparation. These checks keep the data from including any garbage or out of range values. The values which cannot be explained theoretically and practically are excluded from the dataset by apply these sanity checks. (Ren, 2019)

- The column age was checked for such values but no such value was found.
- The dataset has no impossible value.

1.3) Renaming Target Feature –

The target feature has been renamed as "target_subscribe" for better understanding and ease of analysis.

1.4) Dropping Feature -

The feature "duration" has been dropped due to the reason that it highly affects the target feature and it is recommended to discard this feature to get realistic predictive model(Ref: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#).

TASK 2: DATA EXPLORATION

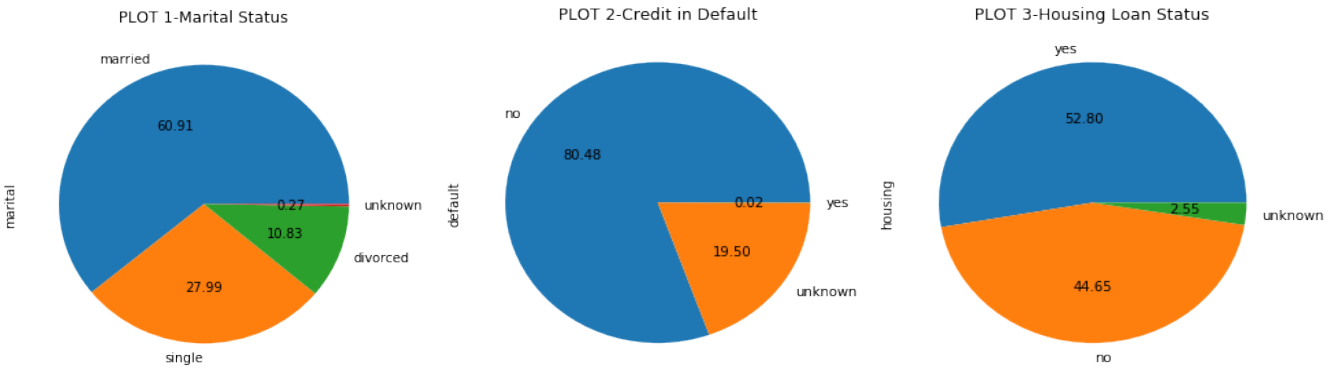
Data Exploration is an important step of in data analysis. The main objective of this step is to get a deep insight about the data. This step helps to get a clear vision of data and understand the main characteristics of the data. The techniques used in this investigation are Graphical techniques. Graphical techniques include scatter plot, histogram, pie-chart, line graph, bar graph etc. (Ren, 2019)

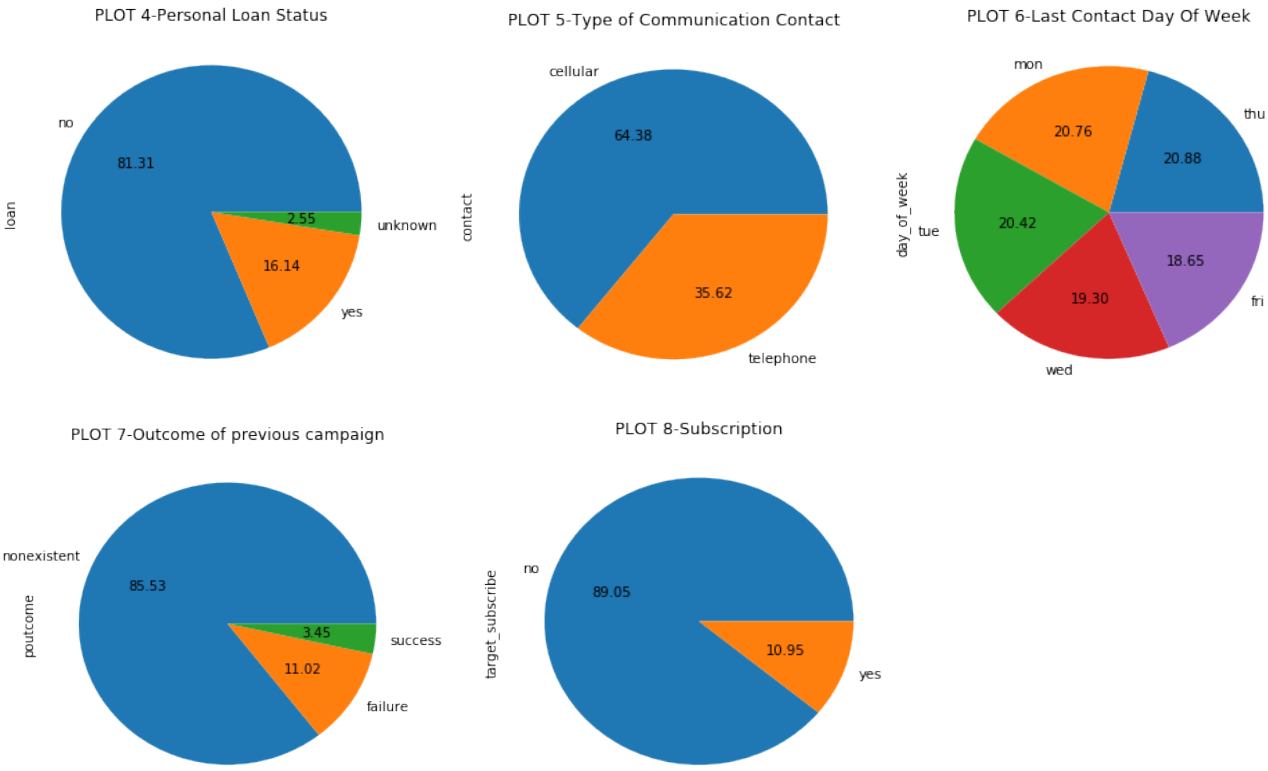
2.1) Individual Visualisation:

- **Categorical Features:**

PIE CHART:

To visualise categorical features, Pie Chart is chosen as it explains these columns perfectly. The Pie Chart represents the percentage distribution of the data. Each segment of Pie Chart explains a Category of Categorical column. The Pie Chart displays all the categories of the Categorical Column as a whole.

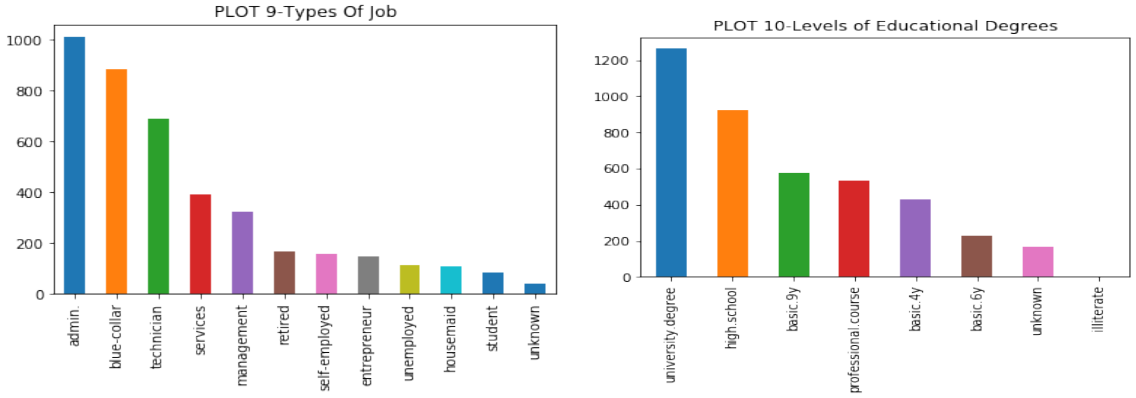


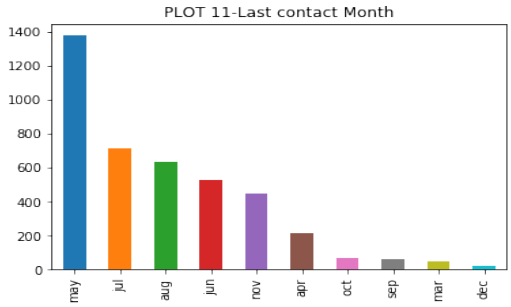


- **PLOT1:** The pie chart explains that majority of customers are married, nearly about 60.91% of all customers. There are less than 0.3% customers with unknown status. About 27.99% customers are single whereas 10.83% of them are divorced.
- **PLOT2:** The pie chart explains that majority of customers do not have credit in default, nearly about 80.48% of all customers whereas 0.02% of customers have credit in default. About 19.50% customers have unknown status.
- **PLOT3:** The pie chart explains that majority of customers have housing loan, nearly about 52.80% of all customers whereas 44.65% of customers do not have housing loan. About 2.55% customers have unknown status.
- **PLOT4:** The pie chart explains that majority of customers do not have personal loan, nearly about 81.31% of all customers whereas 16.14% of customers have personal loan. About 2.55% customers have unknown status.
- **PLOT5:** The pie chart explains that majority of customers have cellular phones, nearly about 64.38% of all customers whereas 35.62% of customers have telephones.
- **PLOT6:** The pie chart explains an almost equal distribution of Last Contact Day with 20.88% for Thursday, 20.76% for Monday, 20.42% for Tuesday, 19.30% for Wednesday and 18.65% for Friday.
- **PLOT7:** The pie chart explains that majority of customers do not have an existing record, nearly about 85.53% of all customers whereas previous campaign outcome for 11.02% customers was failure and 3.45% was success.
- **PLOT8:** Subscription is the target variable for this data set. The pie chart explains that majority of customers did not subscribe, nearly about 89.05% of all customers whereas 10.95% of customers subscribed. That clearly suggest that our dataset is highly imbalanced

BAR CHART:

To visualise these columns, Bar Chart is chosen as it explains these columns perfectly. Bar Chart displays the categories of a categorical attribute in such a way that each rectangular bar represents a particular category with height and length of bar are proportional to the values that they represent.



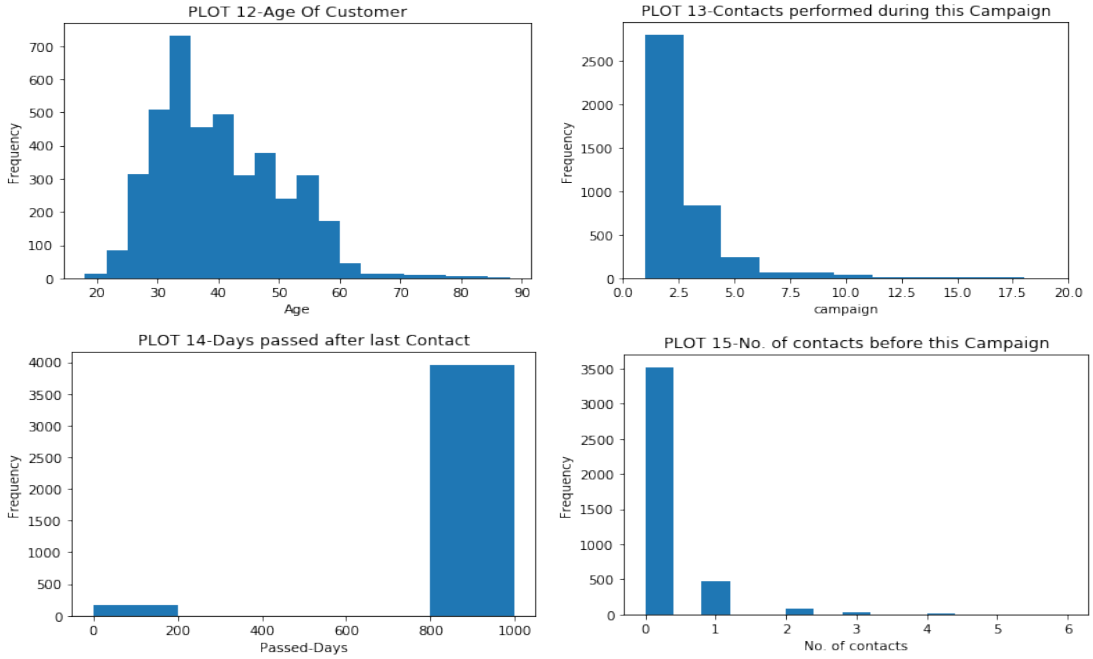


- **PLOT 9:** The bar chart explains that majority of customers are working as admin followed by blue-collar and technician whereas a smaller number of customers are self-employed, entrepreneur, unemployed, housemaids, student or unknown.
- **PLOT 10:** The bar chart explains that larger number of customers have completed University Degree followed by high school whereas only one customer is illiterate.
- **PLOT 11:** The bar chart explains that majority of customers were last contacted in may followed by July, august, June, November and April whereas a smaller number of customers were contacted in October, September, March and December.

- **<u>Numerical Features:</u>**

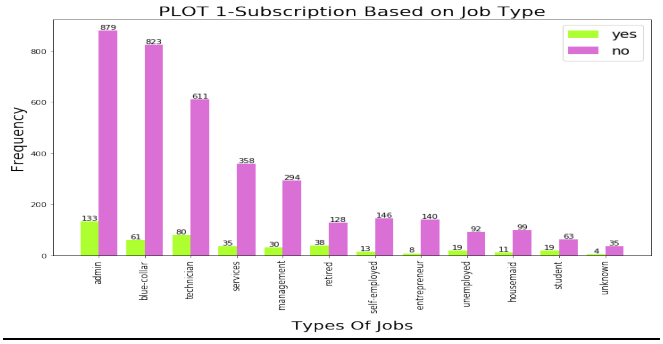
HISTOGRAM:

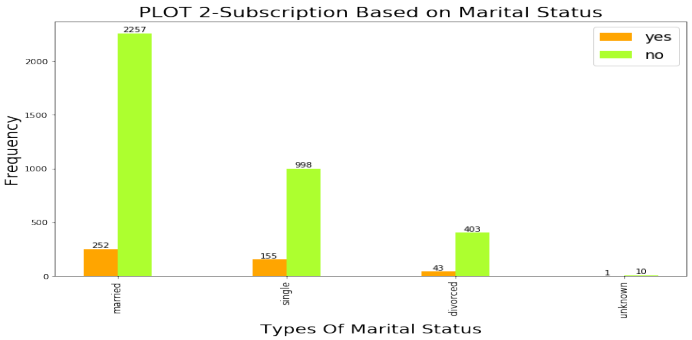
Histogram represents the distribution of data effectively. Histogram uses bars of different heights to display the data. It effectively displays the continuous data values of a numerical attribute. It allows us to understand different properties of distributed data like its skewness, normality, outliers etc.

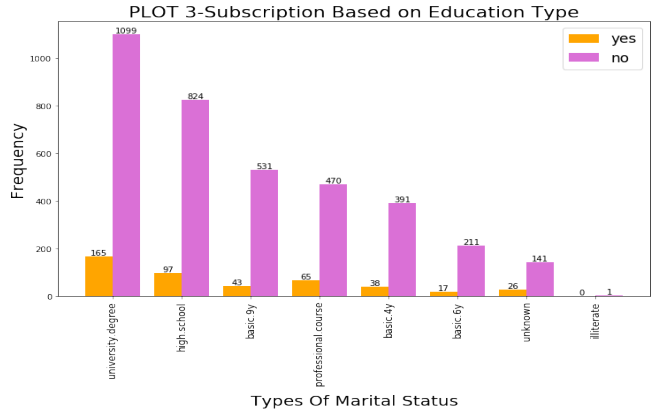


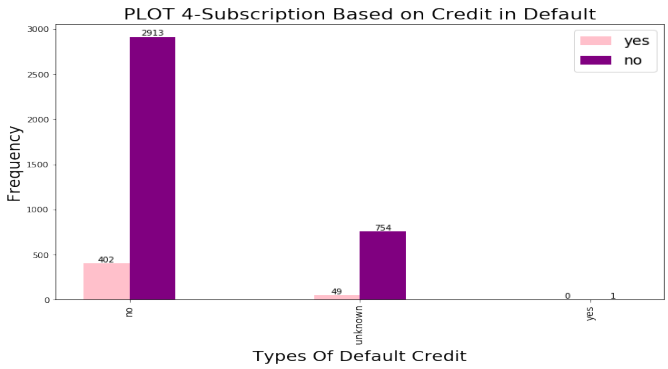
- **PLOT12:** The histogram of the data seems to be symmetrically distributed. The histogram describes that the most common age range lies between 25-60. The plot suggests few outliers around 70-90 age range.
- **PLOT13**: The histogram of data seems to be right skewed. The histogram describes that the most common range of Number of Contacts performed during this campaign lies between 1-6. The plot suggests few outliers around 10-18 range of Number of Contacts performed during this campaign.
- **PLOT14**: The histogram of data seems to be bi-modal. The histogram describes that smaller number of customers were last contacted in last 0-200 days from previous campaign whereas larger number of customers were either not contacted at all or were contacted in last 800-999 days from previous campaign
- **PLOT15**: The histogram of data seems to be right skewed. The histogram describes that the most common range of Number of Contacts performed before this campaign lies between 0-1. The plot suggests few outliers around 2-4 range of Number of Contacts performed before this campaign.

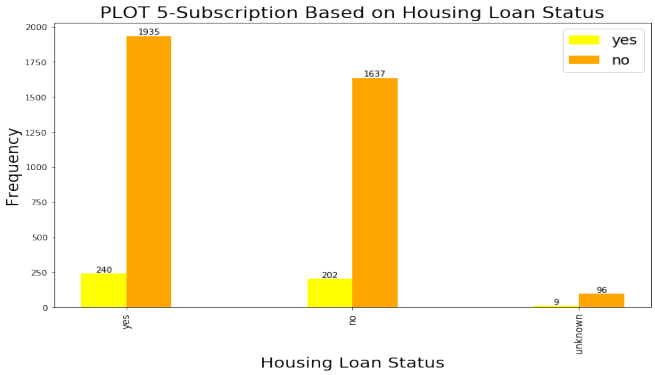
<u>2.2) Bivariate Visualisation:</u>

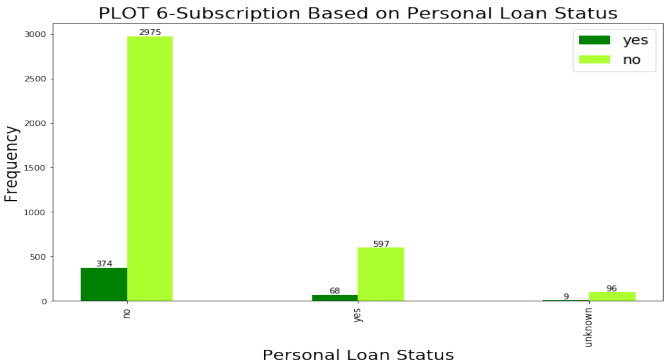

PLOT 1-Subscription Based on Job Type
yes
no
Frequency
admin
blue-collar
technician
services
management
retired
self-employed
entrepreneur
unemployed
housemaid
student
unknown
Types Of Jobs

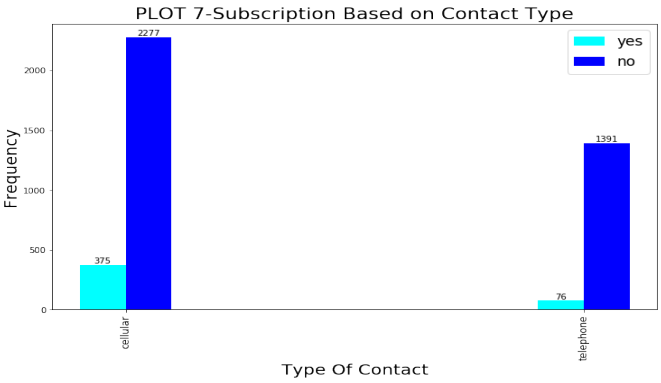

PLOT 2-Subscription Based on Marital Status
yes
no
Frequency
married
single
divorced
unknown
Types Of Marital Status

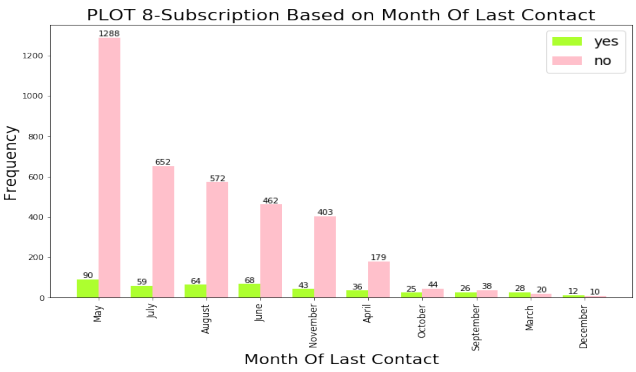

PLOT 3-Subscription Based on Education Type
yes
no
Frequency
university degree
high school
basic.9y
professional course
basic.4y
basic.6y
unknown
illiterate
Types Of Marital Status

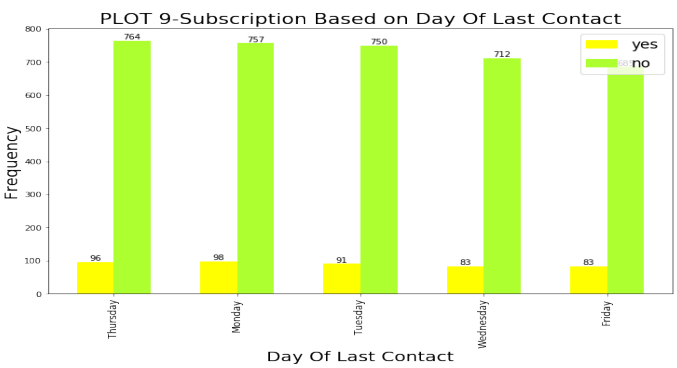

PLOT 4-Subscription Based on Credit in Default
yes
no
Frequency
no
unknown
yes
Types Of Default Credit

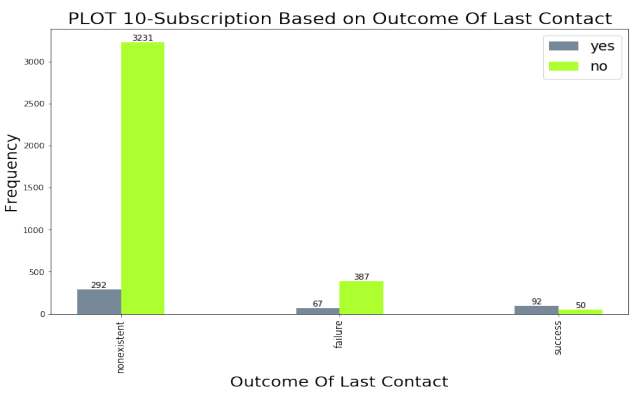

PLOT 5-Subscription Based on Housing Loan Status
yes
no
Frequency
yes
no
unknown
Housing Loan Status


PLOT 6-Subscription Based on Personal Loan Status
yes
no
Frequency
no
yes
unknown
Personal Loan Status


PLOT 7-Subscription Based on Contact Type
yes
no
Frequency
cellular
telephone
Type Of Contact


PLOT 8-Subscription Based on Month Of Last Contact
yes
no
Frequency
May
July
August
June
November
April
October
September
March
December
Month Of Last Contact


PLOT 9-Subscription Based on Day Of Last Contact
yes
no
Frequency
Thursday
Monday
Tuesday
Wednesday
Friday
Day Of Last Contact


PLOT 10-Subscription Based on Outcome Of Last Contact
yes
no
Frequency
nonexistent
failure
success
Outcome Of Last Contact

- **PLOT1:** According to the plausible hypothesis:

Ho (Null Hypothesis) - job and target_subscribe are not independent.
HA (Alternative Hypothesis)- job and target_subscribe are independent.
According to the plot, customers with retired, unemployed, housemaid and student as job type are more likely to subscribe the term deposit than the customers with other job types. This implies that job and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT2:** According to the plausible hypothesis:

Ho (Null Hypothesis) - marital status and target_subscribe are not independent.
HA (Alternative Hypothesis)- marital status and target_subscribe are independent.
According to the plot, customers with divorced and unknown marital status are more likely to subscribe the term deposit than the customers with other marital status. This implies that marital status and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT3:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Education Type and target_subscribe are not independent.
HA (Alternative Hypothesis)- Education Type and target_subscribe are independent.
According to the plot, customers with professional course and unknown education type are more likely to subscribe the term deposit than the customers with other marital status. Customers who have Illiterate education type are not likely to subscribe the term deposit at all. This implies that education type and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT4:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Credit in Default and target_subscribe are not independent.
HA (Alternative Hypothesis)- Credit in Default and target_subscribe are independent.
According to the plot, customers with credit in default are not likely to subscribe the term deposit whereas customers with no credit in default are more likely to subscribe the term deposit. This implies that credit in default and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT5:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Housing Loan Status and target_subscribe are not independent.
HA (Alternative Hypothesis)- Housing Loan Status and target_subscribe are independent.
According to the plot, customers with Housing Loan are more likely to subscribe the term deposit than Customer with no Housing Loan or unknown status. This implies that Housing Loan Status and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT6:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Personal Loan Status and target_subscribe are not independent.
HA (Alternative Hypothesis)- Personal Loan Status and target_subscribe are independent.
According to the plot, customers with no Personal Loan are more likely to subscribe the term deposit than Customer with Personal Loan or unknown status. This implies that Personal Loan Status and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT7:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Contact Type and target_subscribe are not independent.
HA (Alternative Hypothesis)- Contact Type and target_subscribe are independent.
According to the plot, customers with cellular as Contact Type are more likely to subscribe the term deposit than Customer with telephone Contact Type. This implies that Contact Type and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT8:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Month and target_subscribe are not independent.
HA (Alternative Hypothesis)- Month and target_subscribe are independent.
According to the plot, customers last contacted in October, September, March and December are more likely to subscribe the term deposit than customers last contacted in other months. This implies that Month and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

- **PLOT9:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Day_of_week and target_subscribe are not independent.
HA (Alternative Hypothesis)- Day_of_week and target_subscribe are independent.
According to the plot, customers last contacted on any day are equally likely to subscribe and not subscribe the term deposit. This implies that Day_of_week and target_subscribe are independent. Hence, we can say that, we reject Null Hypothesis.

- **PLOT10:** According to the plausible hypothesis:

Ho (Null Hypothesis) - Outcome and target_subscribe are not independent.
HA (Alternative Hypothesis)- Outcome and target_subscribe are independent.

According to the plot, customers with successful last outcome are more likely to subscribe the term deposit than customers with non-existent or failed last outcome. This implies that Outcome and target_subscribe are not independent. Hence, we can say that, we fail to reject Null Hypothesis.

TASK 3: DATA MODELLING

Step by step log:

- **TRANSFORMATION:**

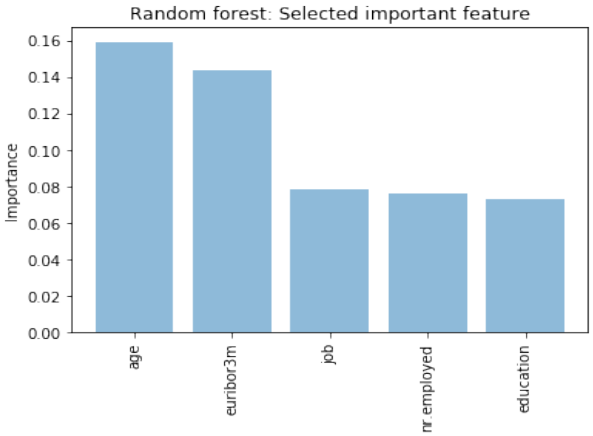
The transformation was very important step for this dataset as some of the descriptive features very categorical and some were numerical. This stage was performed to bring all the descriptive features to one level. To achieve this, the dummy variable target labelEncoder () function has been used. Further, the features of data variable have been encoded to an integer array with each feature corresponding to a single integer array.

- **TRAIN AND TEST SPLIT:**

The Target feature has been dropped from the data and remaining dataset is named as "data". This is done so that model can learn from the train data set later in analysis. A dummy variable named "target" has been created to store the target feature of the data. Both data and target variables have been split up into three pairs of test and train data sets.
First Pair consists of 50% train data and 50% test data.
Second Pair consists of 60% train data and 40% test data.
Third pair consists of 80% train data and 20% test data.

- **FEATURE SELECTION:**

For feature selection Random Forest Technique has been selected because as it selects important features by considering interaction between all features and target variable. For this dataset there were few features which were not having much of the significance with target directly; however, during investigation it came to light that these features were having conditional significance on determining and predicting target feature. For example, descriptive feature nr_employed.



- **Model Selection:**

The Supervised Machine Learning technique requires human interaction to label the data. In this technique, patterns can be found from the previous outcomes of observations which are learnt by the model and are used to find outcome of new observations. Classification is a supervised machine learning problem, which identifies the category or continuous value for outcome of new observation based on the train data whose category or continuous value is known. The dataset chosen for this project is based on a Binary Classification problem. The Target Feature of the dataset is a binary feature where “yes” means subscribing the term deposit and “no” means not subscribing the term deposit. Therefore, applying two modelling techniques K-NearestNeighbor and Decision Tree to solve the Classification problem. (Ren, 2019)

Classification Models

- **K-NearestNeighbor (KNN):**

The K-Nearest Neighbour classification technique classifies the new observation based on the k closest training observation in the model. It is a powerful technique to categorize the observations based on outcomes in training data. KNN model is selected from sklearn and is used to solve the binary classification problem where “yes” means subscribing the term deposit and “no” means not subscribing the term deposit.

- **Decision Tree (DT):**

The decision tree classification technique can work on both continuous or categorical features. Decision tree splits the observations based on most effective splitter which splits the observations in two or more homogeneous groups of observations. Each decision tree consists of three important components**: internal node** which represents a test condition, **leaf node** which represents the class label and **branch** which represents result of the test. To split the population Gini Index is used as it is very effective for binary splits. Decision Tree model is selected from sklearn and is used to solve the binary classification problem where “yes” means subscribing the term deposit and “no” means not subscribing the term deposit.

- **MODEL TRAINING:**

After finding good predictors and a good modelling technique, moving on to training the model seems to be a good idea. In this step, the model is presented with data from which it can learn.

K-NearestNeighbor (KNN):

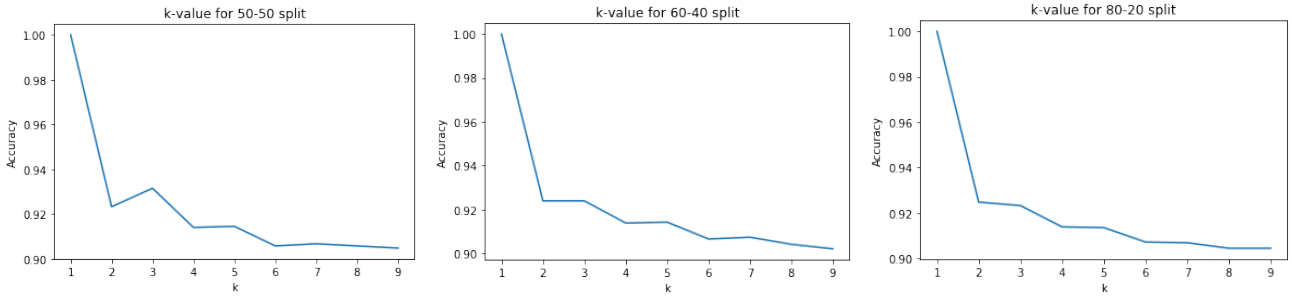
To find the reasonable value of k following steps has been taken:

* For loop has been applied for each value in 1 to 10

*The model has been defined with help KNeighborsClassifier () function for each value of k from 1 to 10.

*The model has been trained with help of fit () function for each value of k from 1 to 10.

*Accuracy score has been calculated for each value of k from 1 to 10.



a) 50% for train data and 50% for test data

*The plot suggests the value of k to be 3 because as the value for k increases its accuracy decreases and for minimum value of k the accuracy is approaching to be 1 which is not possible in real situations.

b) 60% for train data and 40% for test data

*The plot suggests the value of k to be 3 because as the value for k increases its accuracy decreases and for minimum value of k the accuracy is approaching to be 1 which is not possible in real situations.

c) 80% for train data and 20% for test data

*The plot suggests the value of k to be 2 or 3 because as the value for k increases its accuracy decreases and for minimum value of k the accuracy is approaching to be 1 which is not possible in real situations.

To train the model for each pair of train-test the model has been defined with help KNeighborsClassifier () function for k=3 and p=2 where p=2 means Euclidean distance. Results are presented in the table below.

Decision Tree (DT):

To train the model for each pair of train-test datasets following steps has been taken:

The model has been defined with help of DecisionTreeClassfier () function for following parameters.

a) **max_depth:** The value for this parameter is chosen to be 2 as highest accuracy was observed on depth 2.

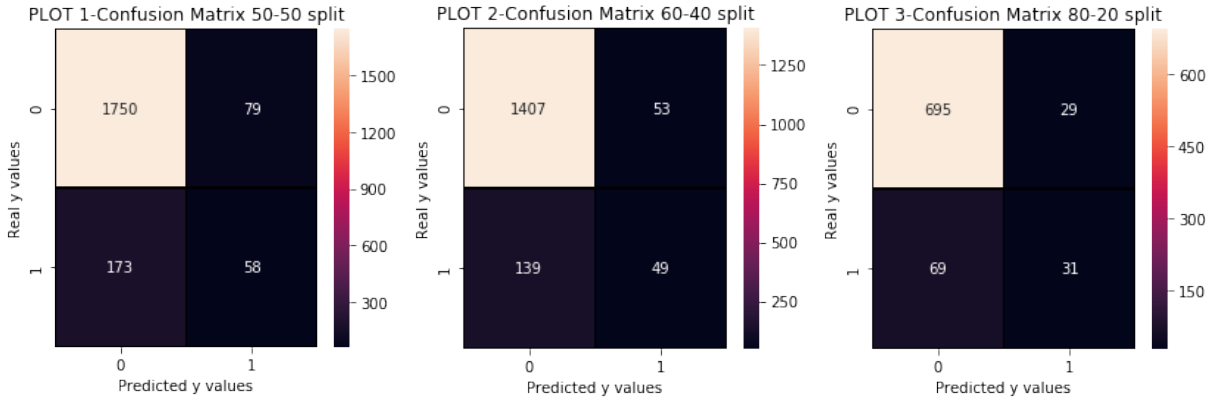
b) **criterion**: To split the population Gini Index is used as it is very effective for binary splits.

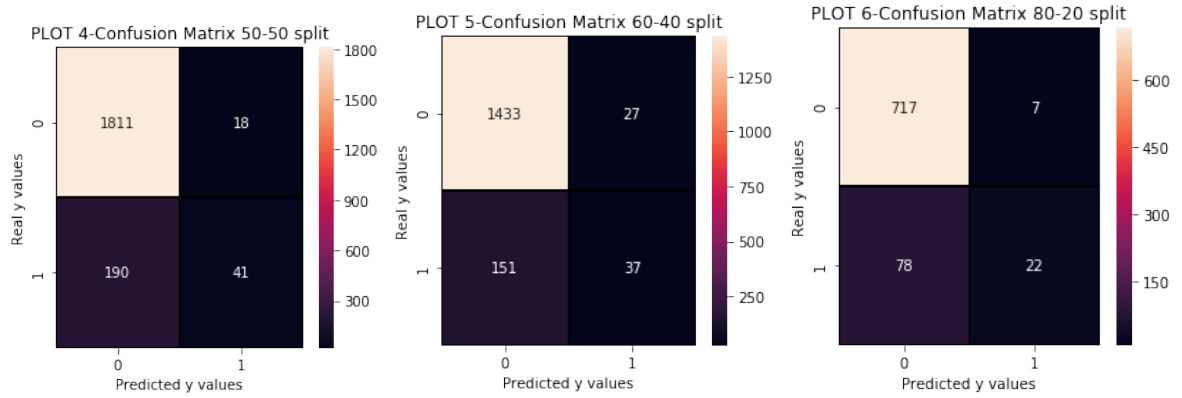
c) **random_state**: this parameter has been set same for target and remaining features to ensure random sampling

d)

3.3) Model Evaluation:

- **K-Nearest Neighbor (KNN):**





a) <u>50% for train data and 50% for test data:</u>

***Confusion Matrix:**

The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot1 suggests that 1750 labels were correctly classified as No Subscription whereas 173 labels were mislabelled as Yes Subscription. The plot1 also suggests that 58 labels were correctly classified as Yes Subscription whereas 79 labels were mislabelled as No Subscription. From this it is clear that model is not identifying potential customers correctly. Thus certainly it is not a good algorithm for this data.

b) <u>60% for train data and 40% for test data</u>

***Confusion Matrix:** The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot2 suggests that 1407 labels were correctly classified as No Subscription whereas 139 labels were mislabelled as Yes Subscription. The plot2 also suggests that 49 labels were correctly classified as Yes Subscription whereas 53 labels were mislabelled as No Subscription.

c) <u>80% for train data and 20% for test data</u>

***Confusion Matrix:** The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot3 suggests that 695 labels were correctly classified as No Subscription whereas 69 labels were mislabelled as Yes Subscription. The plot3 also suggests that 31 labels were correctly classified as Yes Subscription whereas 29 labels were mislabelled as No Subscription.

- **<u>Decision Tree (DT):</u>**

a) <u>50% for train data and 50% for test data</u>

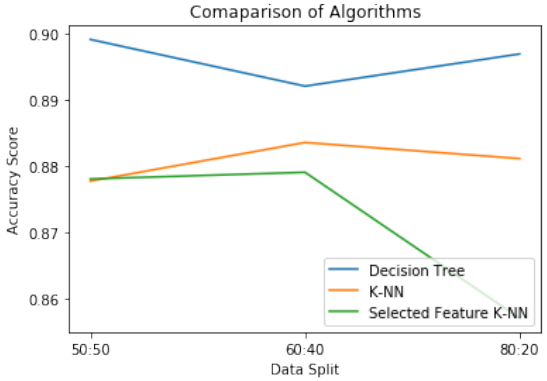
***Confusion Matrix:** The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot4 suggests that 1811 labels were correctly classified as No Subscription whereas 190 labels were mislabelled as Yes Subscription. The plot4 also suggests that 41 labels were correctly classified as Yes Subscription whereas 18 labels were mislabelled as No Subscription.

b) <u>60% for train data and 40% for test data</u>

***Confusion Matrix:** The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot5 suggests that 1433 labels were correctly classified as No Subscription whereas 151 labels were mislabelled as Yes Subscription. The plot5 also suggests that 37 labels were correctly classified as Yes Subscription whereas 27 labels were mislabelled as No Subscription.

c) <u>80% for train data and 20% for test data</u>

***Confusion Matrix:** The Confusion matrix depicts the outcome for test data where 0 is depicted as No Subscription and 1 as Yes Subscription. The plot3 suggests that 717 labels were correctly classified as No Subscription whereas 78 labels were mislabelled as Yes Subscription. The plot3 also suggests that 22 labels were correctly classified as Yes Subscription whereas 7 labels were mislabelled as No Subscription.



Type	KNN	KNN	KNN	DecisionTree	DecisionTree	DecisionTree
Split	50% : 50%	60% : 40%	80% : 20%	50% : 50%	60% : 40%	80% : 20%
Precision	0.4864	0.4509	0.5681	0.6949	0.5781	0.7586
Recall	0.1558	0.1223	0.25	0.1774	0.1968	0.22
F1-Score	0.2360	0.1924	0.3472	0.2827	0.2936	0.3410
Classification Accuracy	0.8868	0.8828	0.8859	0.8990	0.8919	0.8968
Classification Error	0.1131	0.1171	0.1140	0.1009	0.1080	0.1031
Classification Accuracy with Feature Selection	0.884	0.89	0.868	-	-	-

Above table provides values of Precision. Recall, F1-score, classification accuracy, classification Error and Classification accuracy after feature selection for k-Nearest neighbours' model.

ANALYSIS:

- From entire analysis it is stoutly clear that this data set is imbalanced and asymmetric and from confusion matrix it is understood that values of false positive and false negative aren't same. Thus we can't solely rely on accuracy score to determine best model for this dataset.
- As Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, this value has been taken into consideration.
- As Recall is the ratio of correctly predicted positive observations to the all observations which are actually yes, this value is taken in account.
- F1 Score is the weighted average of Precision and Recall which is more useful to identify good model for uneven dataset this would play an important role in decision making.

CONCLUSION:

- From this analysis, it becomes evident that K-Nearest Neighbour model at 80%-20% split of data fits the model best out of all other algorithm observed for following reasons:
    - The precision values that is correctly identifying subscriber from the given instances is 51.66%. As the primary business requirement is to identify subscribers from number of people. However, the not identifying correct subscriber people will certainly not suffice the requirement.
    - The Recall value is that is correctly identifying real subscribers is very good as compared to other models which is exactly what is intended.
    - F1-score that weighted average is always good score to determine efficiency of algorithm for uneven dataset, which is 38.75%.
    - Accuracy is 88.1% and classification error is 11.89% which are also reasonably good.

RECOMENDATION:

* This dataset is highly imbalanced thus some balancing techniques could be applied before running algorithm on this data.
* This data set can also be updated with recent market trend to make it more useful for present time.

REFERNCES:

1)Ren, D. (2019). data curation. [ebook] pp.1-58. Available at: https://rmit.instructure.com/courses/51550/files/6749834?module_item_id=1620445 [Accessed 29 May 2019].

2)Ren, D. (2019). Data Summarisation: Descriptive Statistics and Visualisation. [ebook] pp.1-45. Available at: https://rmit.instructure.com/courses/51550/files/6878828?module_item_id=1631828 [Accessed 29 May 2019].

3)Ren, D. (2019). *Practical Data Science:Course Summary*. [ebook] pp.1-42. Available at: https://rmit.instructure.com/courses/51550/files/7901589?module_item_id=1758754 [Accessed 29 May 2019].