Bank Marketing Campaign

Synopsis

We are going to do data analysis on the bank marketing campaign. This data is related to direct marketing campaigns of a Portuguese institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. A Portuguese bank has been struggling with its telemarketing strategy of promoting subscriptions to its long-term deposit accounts. So here we are going to build an effective strategy to promote this product to its potential subscriber based on the data the bank has previously collected. For this purpose, we must have to look at the characteristics of customers, and to get the best prediction, we must have to preprocess the data.

Data Preprocessing

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data

to make it suitable for building and training Machine Learning models. The dataset contains 17 columns out of 16 are features and 1 is a target column. important features columns:

- **pdays: is the number of days that passed by after the client was last contacted from a previous**
- **Education: is the highest education level completed by the customer**
- **Default: indicates whether the customer has credit in default or not**
- **Duration: measures the last contact duration in seconds**

There are some of the columns that are strongly correlated to the target column. so to determine the likelihood of customers who are willing to subscribe to the long-term deposit accounts, we have to focus on those columns which strongly correlated to the target column. This dataset does not have any null values but contains many outliers. Outliers will affect prediction so we have to get rid of them.

We are going to perform the following steps:

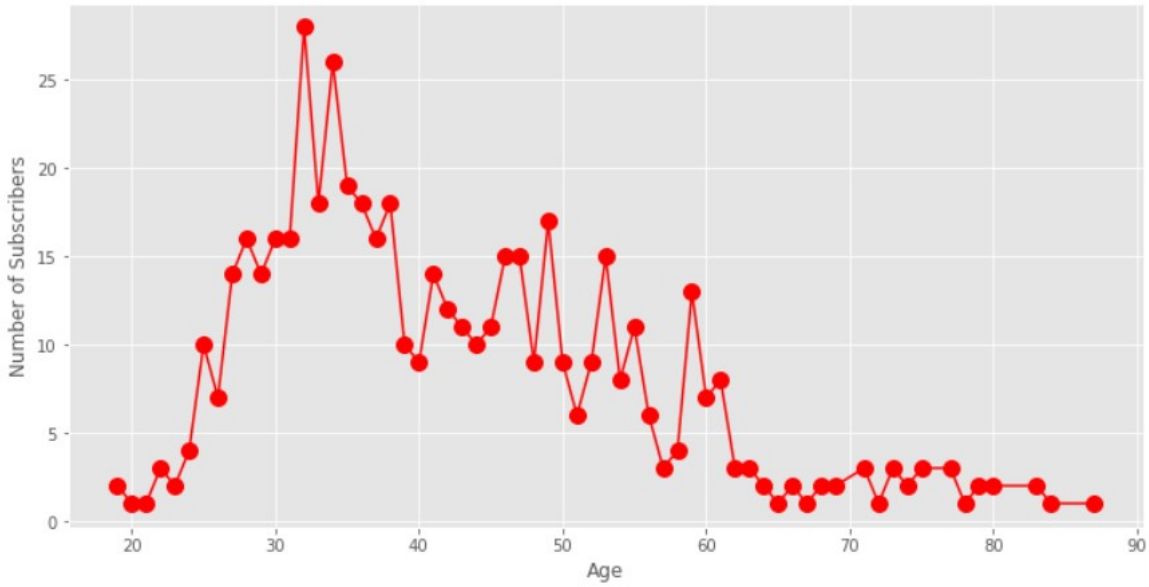
Step#1: Exploratory Data Analysis.
Step#2: Find the Correlations between the columns.
Step#3: Normalize the Date.
Step#4: Scaling the data.
Step#5: Balance the Data.

Step#1: Exploratory Data Analysis:

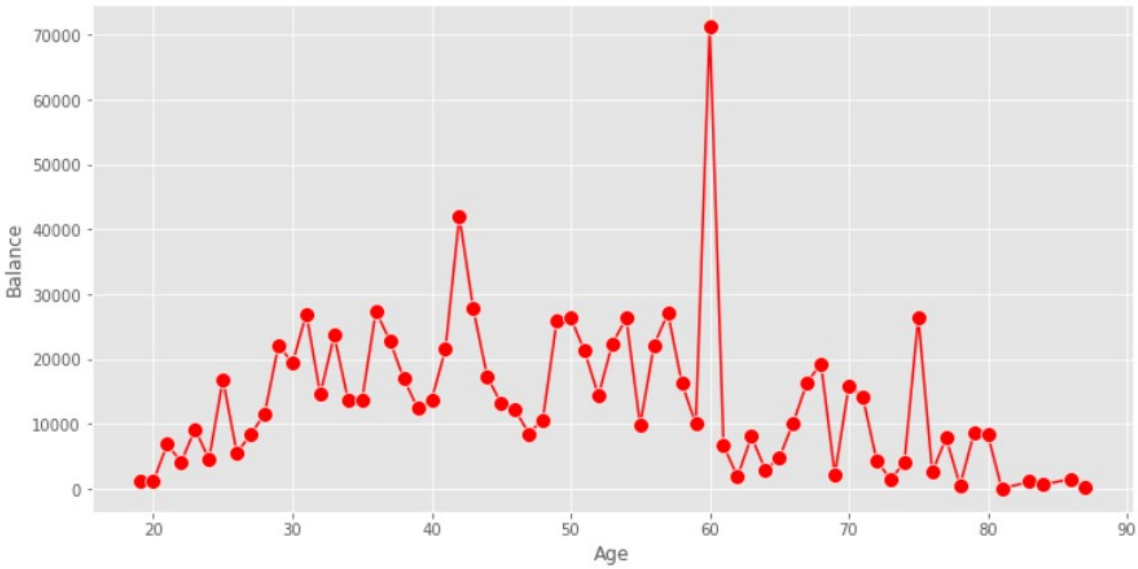
Question#1:Which Age Group people subscribed to the most?
Question#2:What is the education level of most of the subscribers?
Question#3:What is the bank balance of most of the subscribers?
Question#4:What is the loan applying percentage of the subscribers?
Question#5:What is the job type of most subscribers?
Question#6:What is the marital status of most subscribers?
Question#7:Those who subscribed the most has credit in default or not?
Question#8:On which day and month do people subscribe the most?
Question#9:What were the outcomes of the previous marketing campaign, if it failed then why did it happen?

To create good marketing strategies. we have to find the answers to these questions. So I am going to plot many graphs and charts with the help of matplotlib and seaborn. Now let's start.

Question#1:Which Age Group people subscribed to the most?

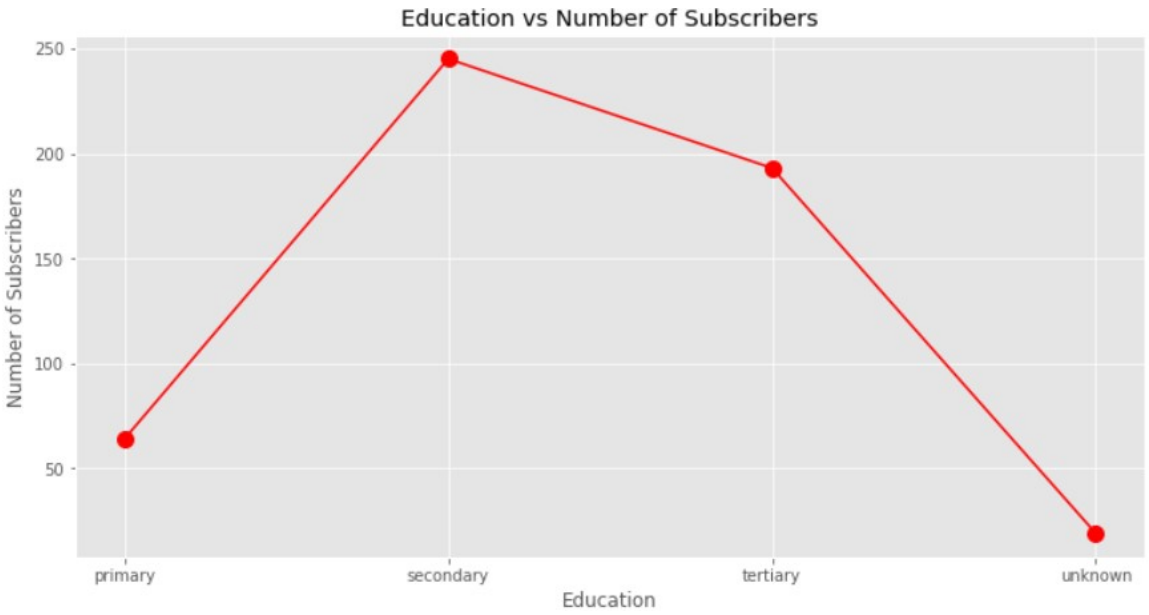


Answer: People who have ages between 30 to 40 subscribed the most.



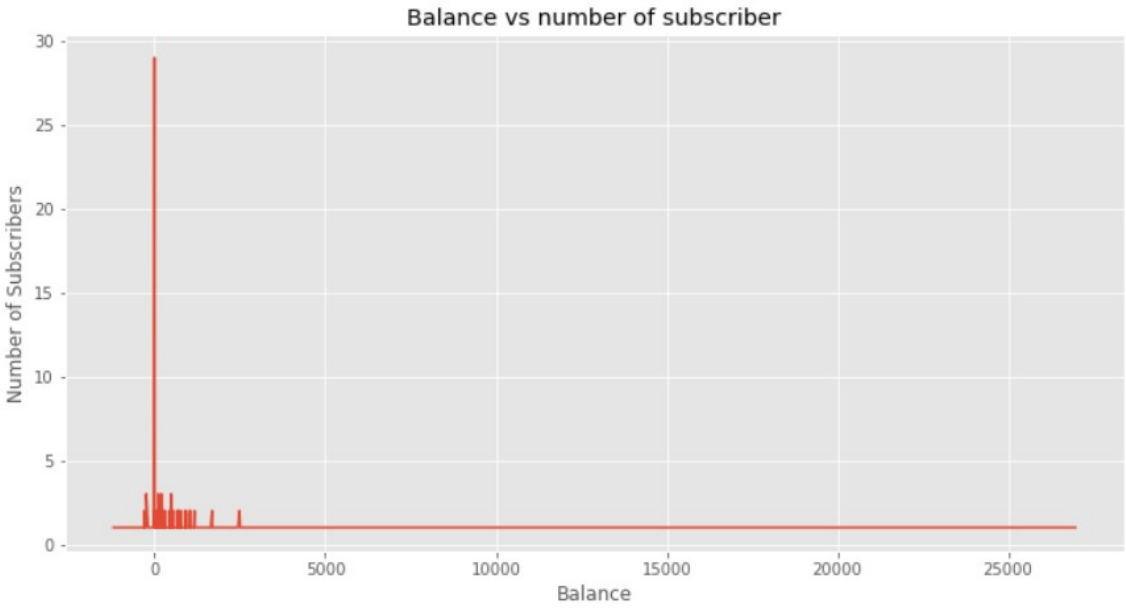
People who have ages between 40 to 60 have the high balance.

Question#2:What is the education level of most of the subscribers?



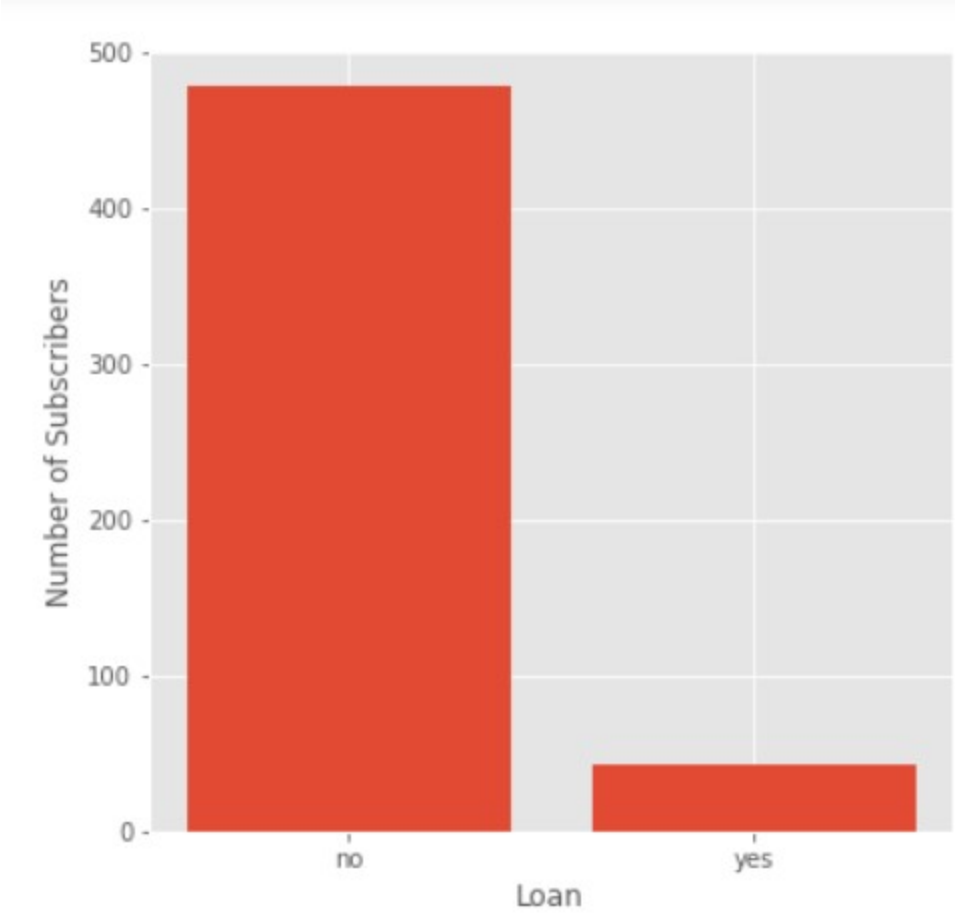
Answer:
people who have secondary education level subscribed the most.

Question#3:What is the bank balance of most of the subscribers?



Answer:
people who have bank balance is zero subscribed the most.

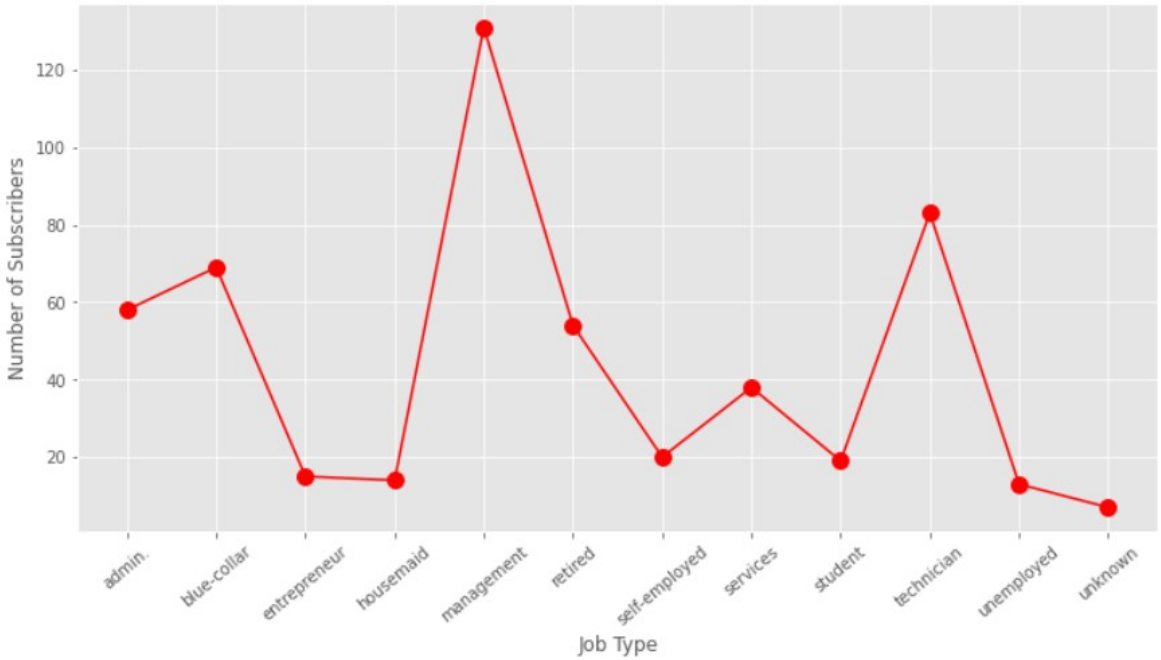
Questtion#4:What is the loan applying percentage of the subscribers?



Answer:

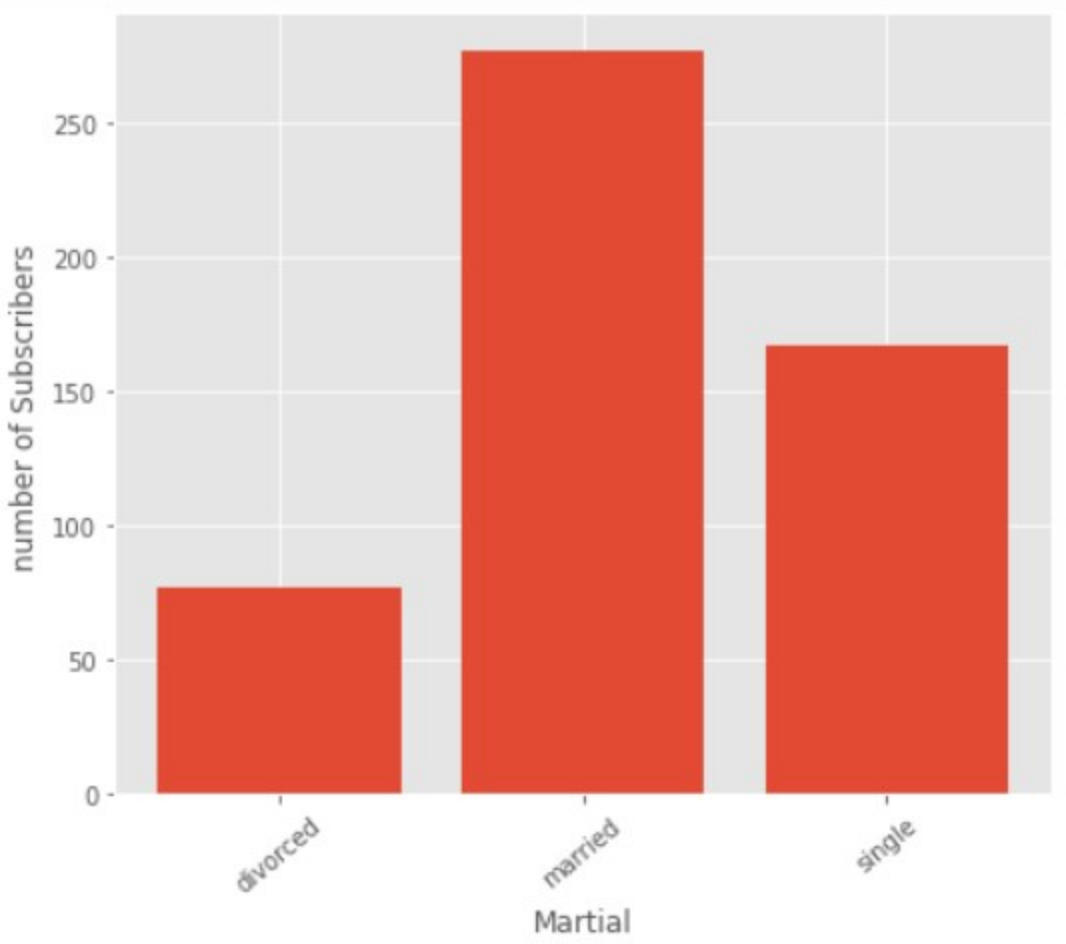
Most subscribers did not apply for the loan

Question#5:What is the job type of most subscribers?



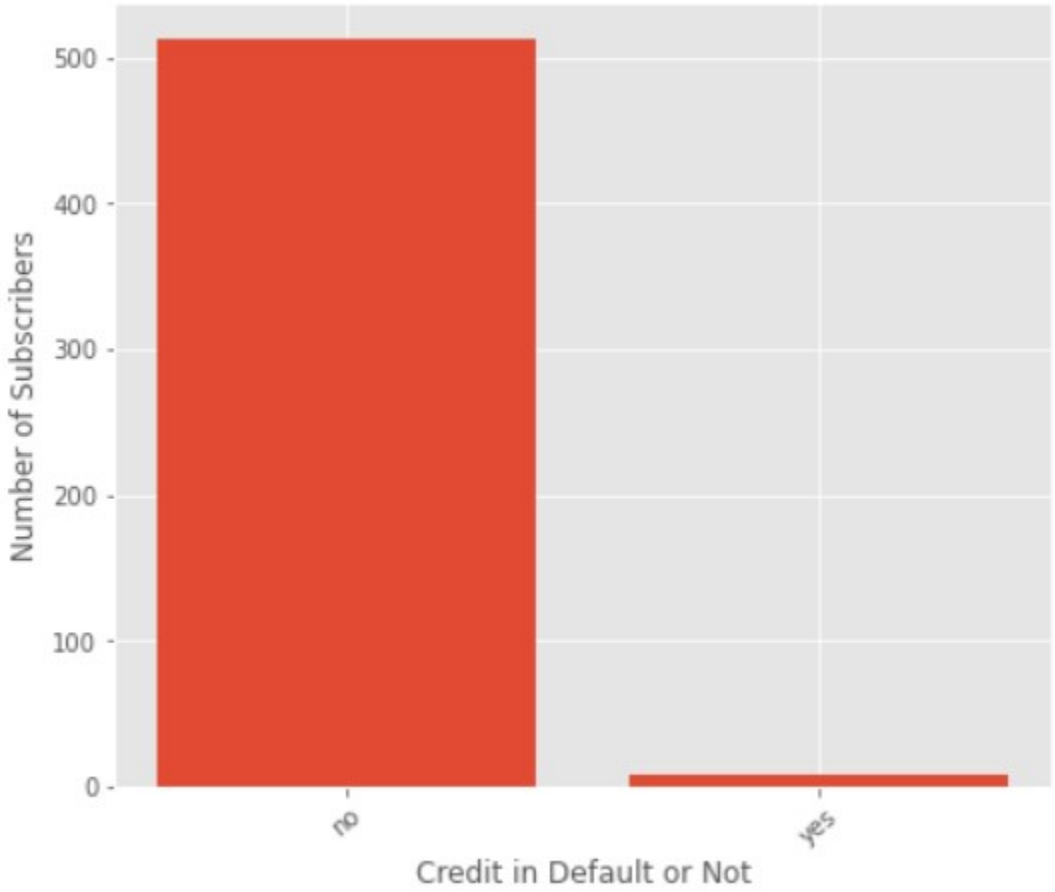
Answer: people who have job type is management, subscribed the most.

Question#6:What is the marital status of most subscribers?



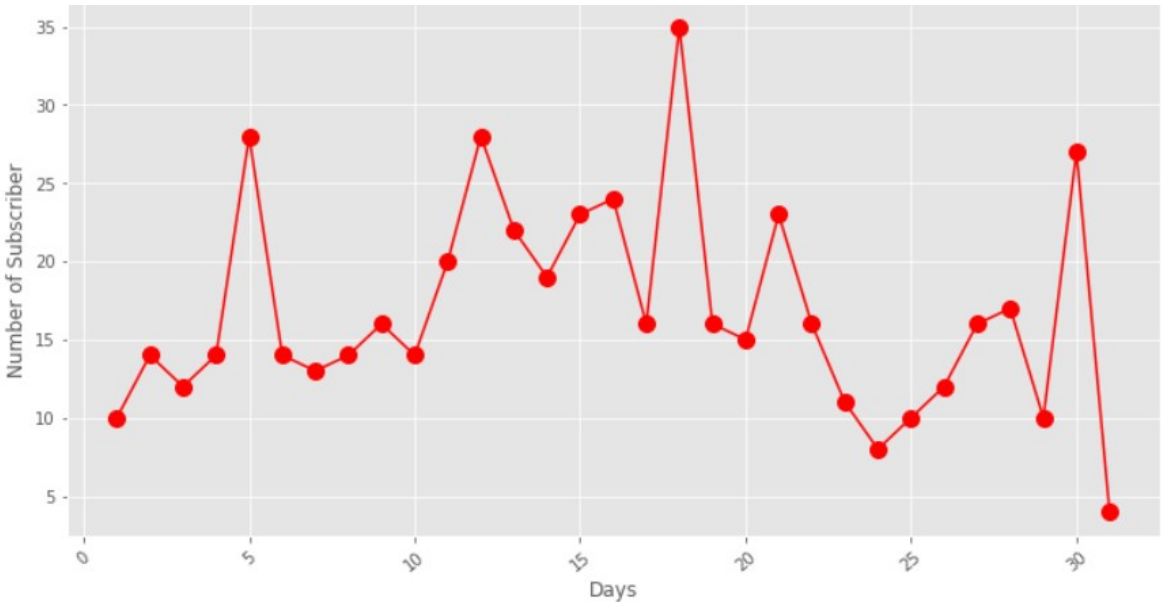
Answer: Married people subscribed the most

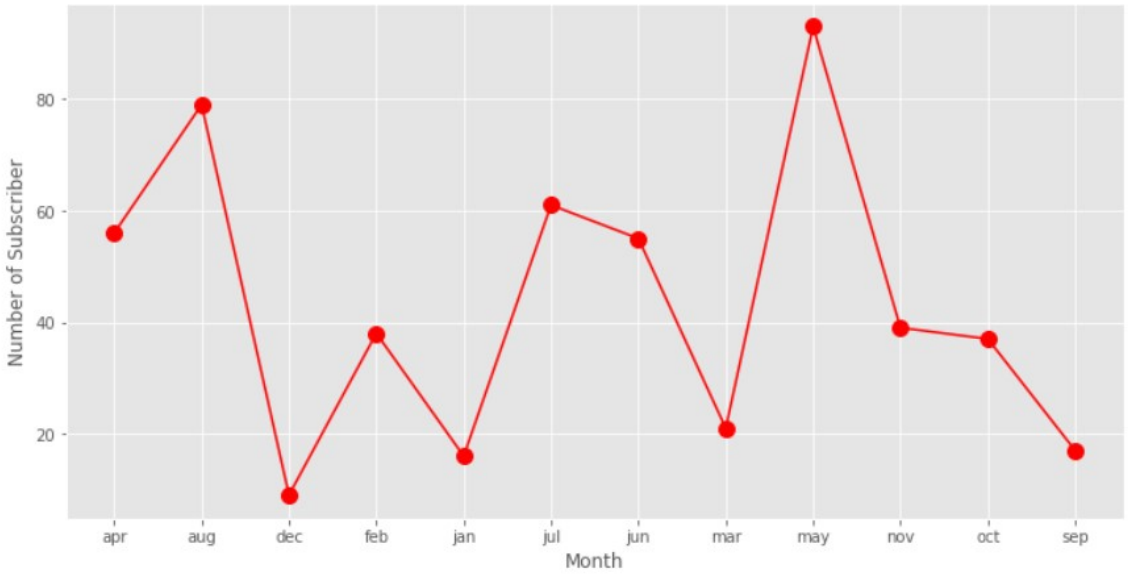
Question#7:Those who subscribed the most has credit in default or not?



Result: people who have not credit in default subscribed th most

Question#8:On which day and month do people subscribe the most?

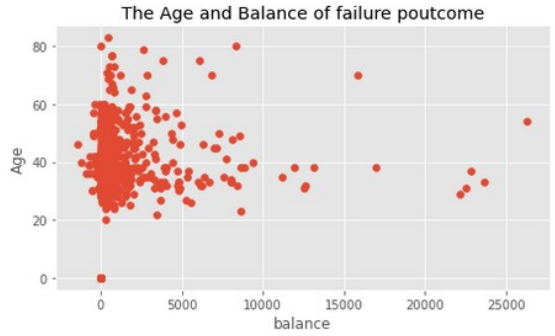


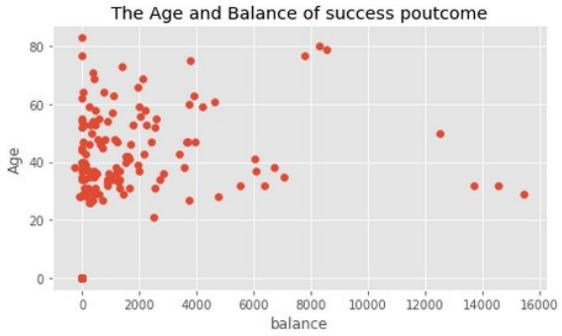


Answer: The day between 16 to 19 and in month may got most subscribers.

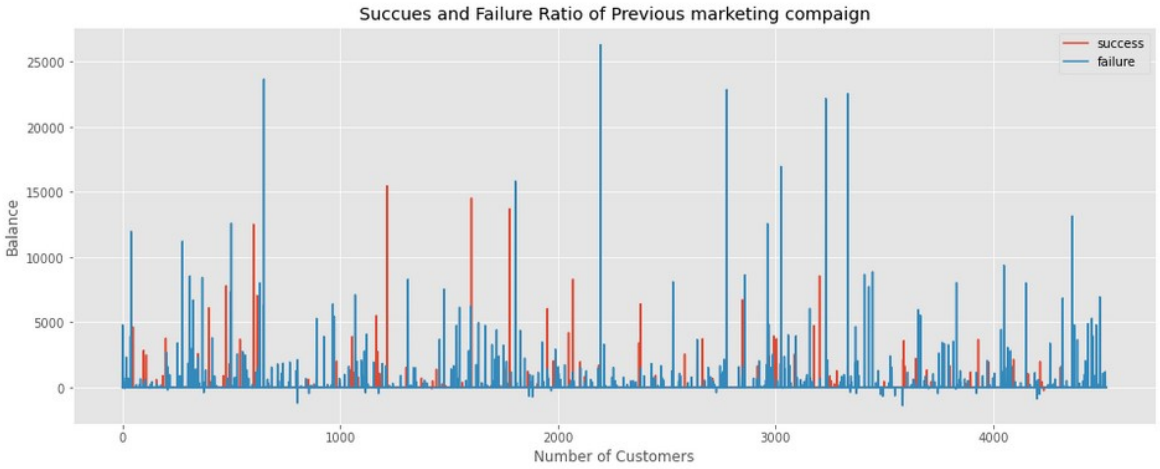
Question#9:What were the outcomes of the previous

marketing campaign, if it failed then why did it happen?





Answer: I think compaigns failed because most subscribers had zero balance

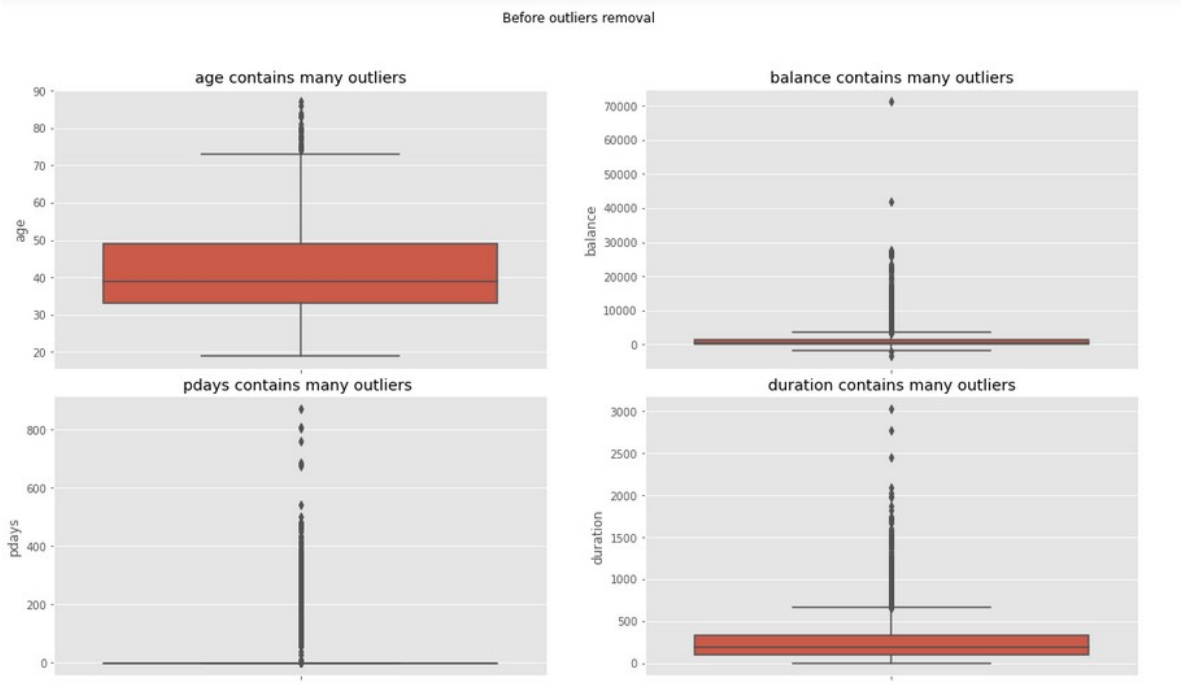


Step#1: Outliers Detection:

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate the experimental error; the latter are sometimes excluded from

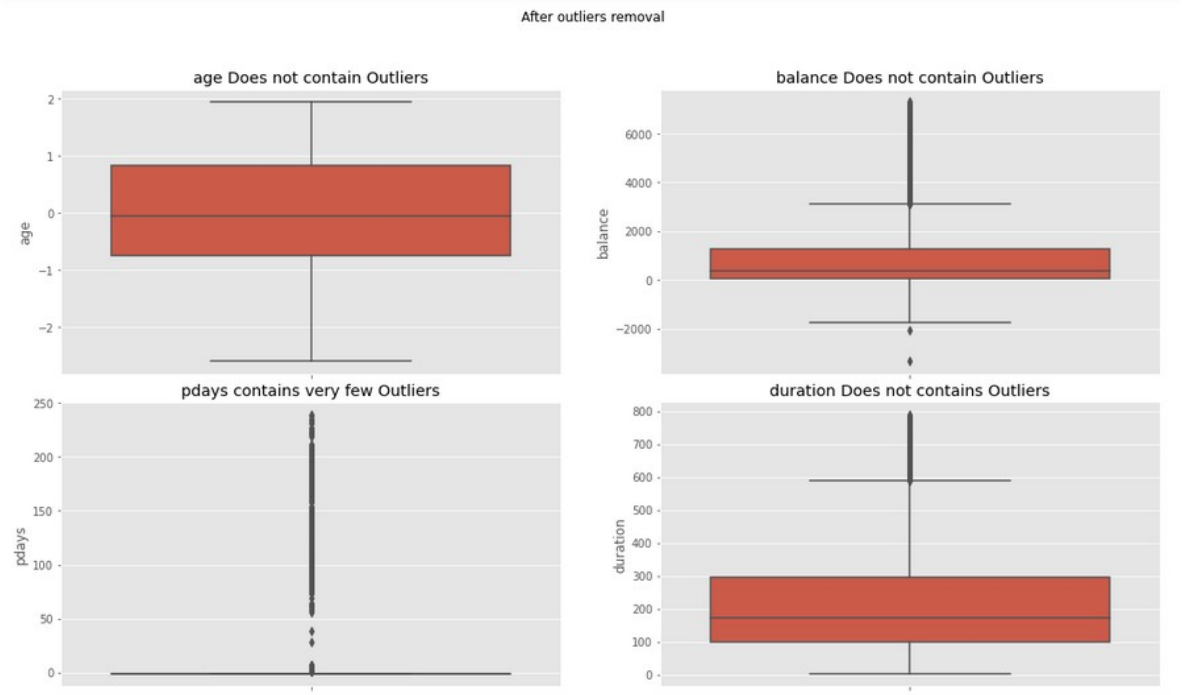
the data set. An outlier can cause serious problems in statistical analyses. Some of the most popular methods for outlier detection are: Z-Score or Extreme Value Analysis (parametric) Probabilistic and Statistical Modeling (parametric) Linear Regression Models (PCA, LMS) Proximity Based Models (non-parametric)

Columns Before Outliers Removal



As you can see dataset contains many outliers

Columns After Outiliers Removal

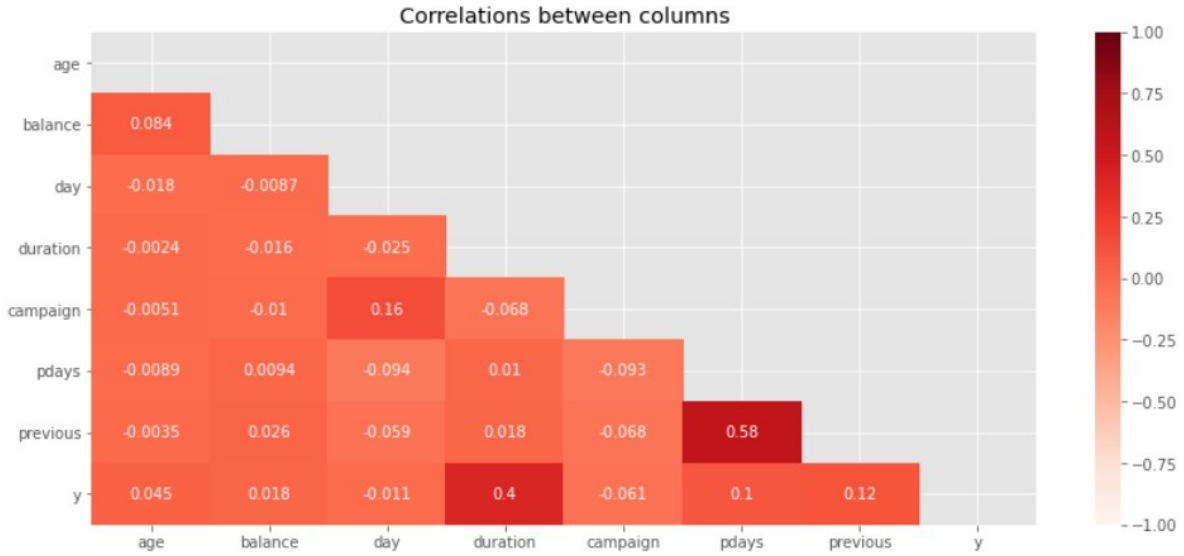


Now we have been removed the outiers

5

Step#1:Find the Correlation between the columns.

If Correlation is near to 1 then it means columns have Strong Positve Correlation with eachother.
If Correlation is near to -1 then it means columns have Strong Negative Correlation with each other.
If Correlation is near to 0 then it means Columns have very Weak correlation with each other.

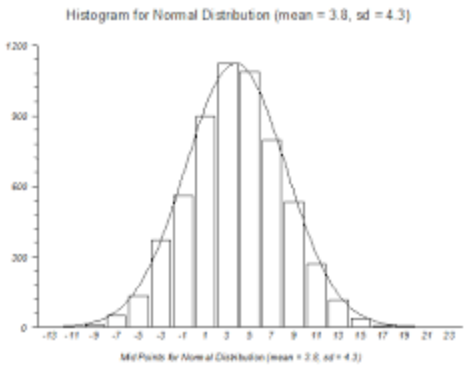


Strong Postive Correlations:

pdays column have Strong Postive correlation with previous.

note: Most of the columns have Weak correlation with each other

Step#2: Do Normal Distribution



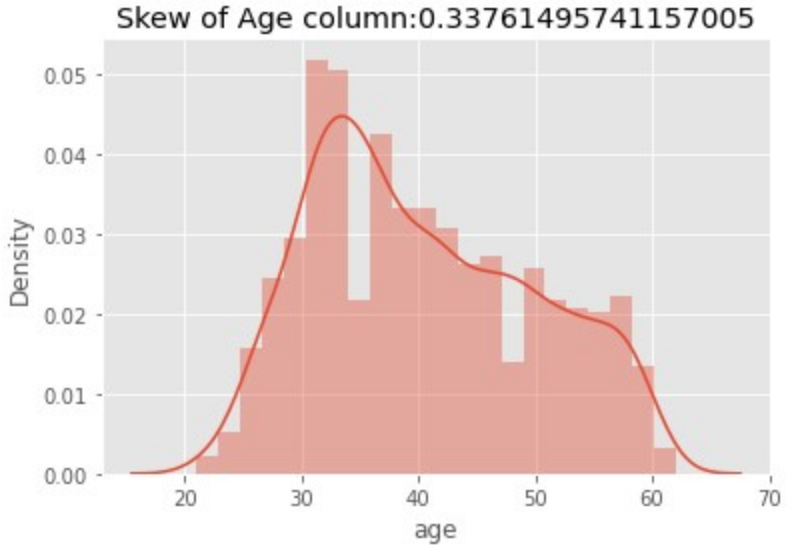
A normal distribution is a bell-shaped curve, and it is assumed that during any measurement, Values will follow a normal distribution with

an equal number of measurements above and below the mean value. This Bell curve shows the normal distribution. We are going to check whether features columns are normally distributed or not, If not then we normalize them with Boxcox Transformation. for normal distribution curve should be equal to zero

Age column before Transformation:

```
Before Boxcox Transformation

Text(0.5, 1.0, 'Skew of Age column:0.33761495741157005')
```

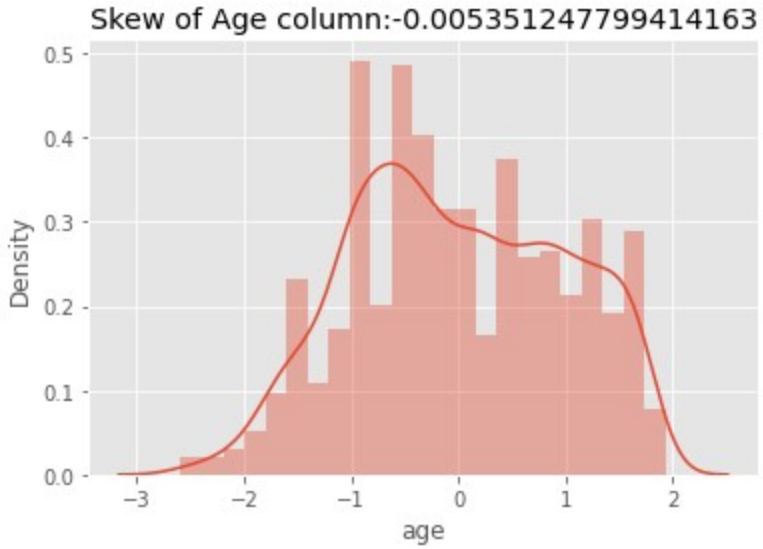


NOTE: ***This graph clearly shows that age column is not normaly distributed because skew is very high***

Age Column After transfromation

```
After Boxcox Transformation

Text(0.5, 1.0, 'Skew of Age column:-0.005351247799414163')
```



NOTE: ***Now the data is normaly distributed because skew is near to zero***

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	-1.139332	10	1	0	0	1787	0	0	0	19	10	79	1	-1	0	3	0
3	-1.139332	4	1	2	0	1476	1	1	2	3	6	199	4	-1	0	3	0
4	1.720514	1	1	1	0	0	1	0	2	5	8	226	1	-1	0	3	0
5	-0.500398	4	2	2	0	747	0	0	0	23	3	141	2	176	3	0	0
7	-0.047372	9	1	1	0	147	1	0	0	6	8	151	2	-1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4515	-0.872739	7	2	1	0	473	1	0	0	7	5	624	5	-1	0	3	0
4516	-0.745168	7	1	1	0	-333	1	0	0	30	5	329	5	-1	0	3	0
4517	1.571107	6	1	2	1	-3313	1	1	2	9	8	153	1	-1	0	3	0
4518	1.571107	9	1	1	0	295	0	0	0	19	1	151	11	-1	0	3	0
4519	-1.422881	1	1	1	0	1137	0	0	0	6	3	129	4	211	3	1	0

3732 rows × 17 columns

Now our data is pretty clean and we are ready to train our data,fit into Classifiers and predict the outcomes

Step#6: Balance the data

Imbalanced data sets are a special case for classification problems where the class distribution is not uniform among the classes. Typically, they are composed of two classes: The majority (negative) class and the minority (positive) class. So we have to balance the minority class with the majority class

After Balancing the data

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	-1.139332	10	1	0	0	1787	0	0	0	19	10	79	1	-1	0	3
1	-1.139332	4	1	2	0	1476	1	1	2	3	6	199	4	-1	0	3
2	1.720514	1	1	1	0	0	1	0	2	5	8	226	1	-1	0	3
3	-0.500398	4	2	2	0	747	0	0	0	23	3	141	2	176	3	0
4	-0.047372	9	1	1	0	147	1	0	0	6	8	151	2	-1	0	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6819	-1.506528	6	1	1	0	1342	0	0	0	21	4	285	4	90	4	1
6820	1.408468	2	0	1	0	460	0	0	0	8	3	533	1	35	0	2
6821	0.024780	6	1	1	0	18	0	0	0	8	7	240	1	-1	0	3
6822	-0.759042	4	1	1	0	161	0	0	0	20	7	166	1	-1	0	3
6823	-0.556760	6	1	0	0	138	0	0	2	2	5	668	1	-1	0	3

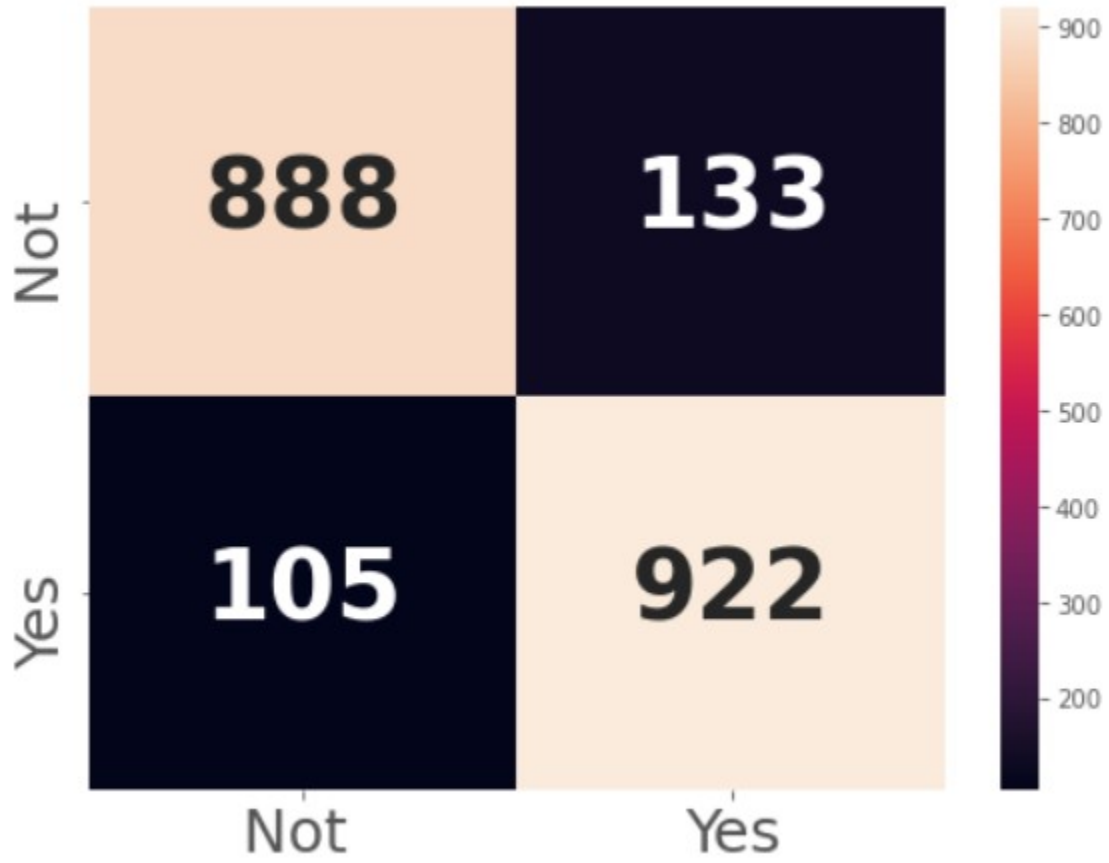
6824 rows × 16 columns

Our dataset became a large dataset.Now it contains 6824 rows × 16 columns.

Predictive Model/Classification

So what are classification models? A classification model attempts to draw some conclusions from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset.

DecisionTreeClassifier Prediction:



This Confusion metricx shows that the prediction rate is very good

Best Hyperparameters:

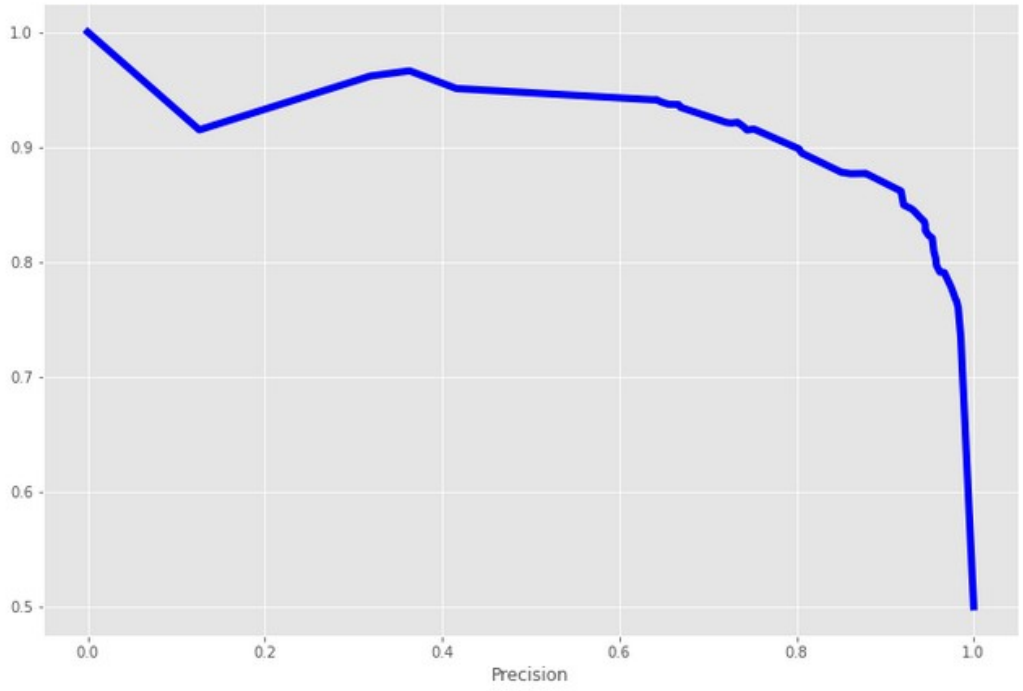
Best criterion : gini
Best max_depth Value : 9
Best min_samples_split Value : 5
Best min_samples_leaf Value : 9
Mean cross-validated accuracy score of the best_estimator: 0.878

Classification Report:

	precision	recall	f1-score	support
Not Deposit	0.92	0.83	0.87	1024
Deposit	0.84	0.93	0.89	1024
accuracy			0.88	2048
macro avg	0.88	0.88	0.88	2048
weighted avg	0.88	0.88	0.88	2048

The accuracy and precision of DecisionTreeClassifier is 88% and 84%

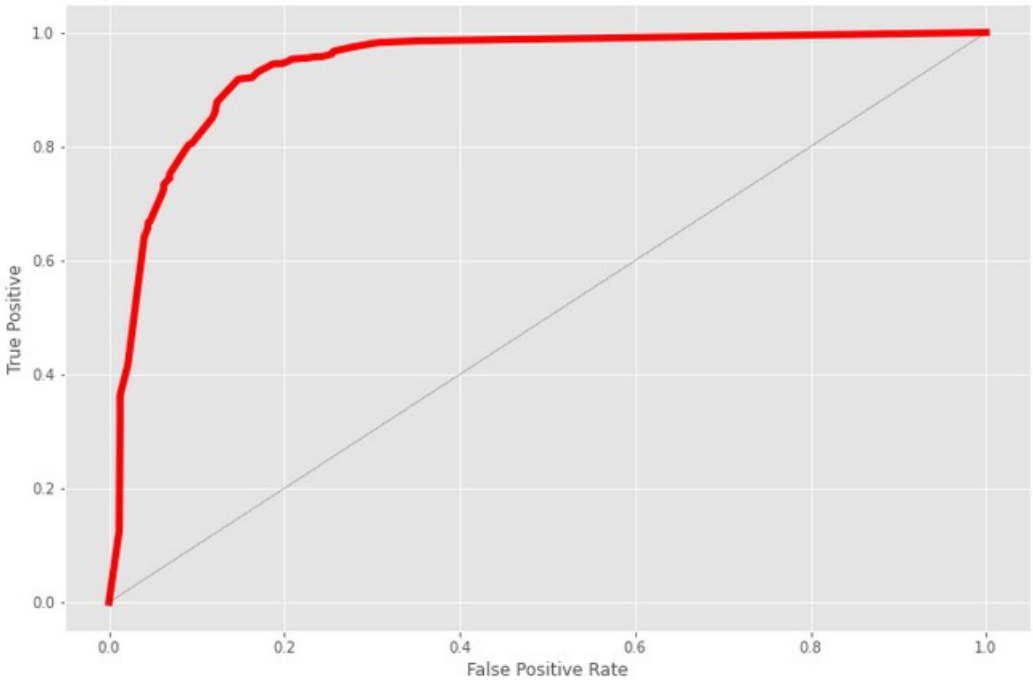
Precision_Recall_curve of Decision Tree:



Precision-Recall curves summarize the trade-off between the

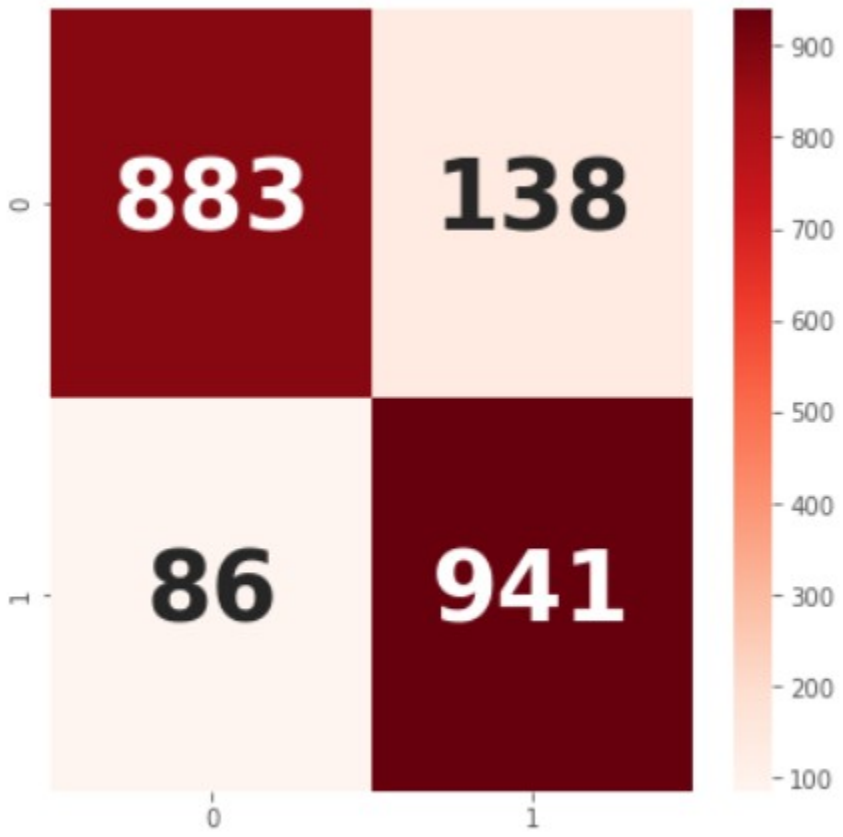
true positive rate and the positive predictive value for a predictive model using different probability thresholds. so as you can see the true positive rate is very of the decision tree.

Roc_Auc_curve of Decision Tree:



AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. so decision tree classifier is proved to be very good for prediction quality

RandomForestClassifier Prediction:



The confusion shows a very high prediction rate.

Best Hyperparameters:

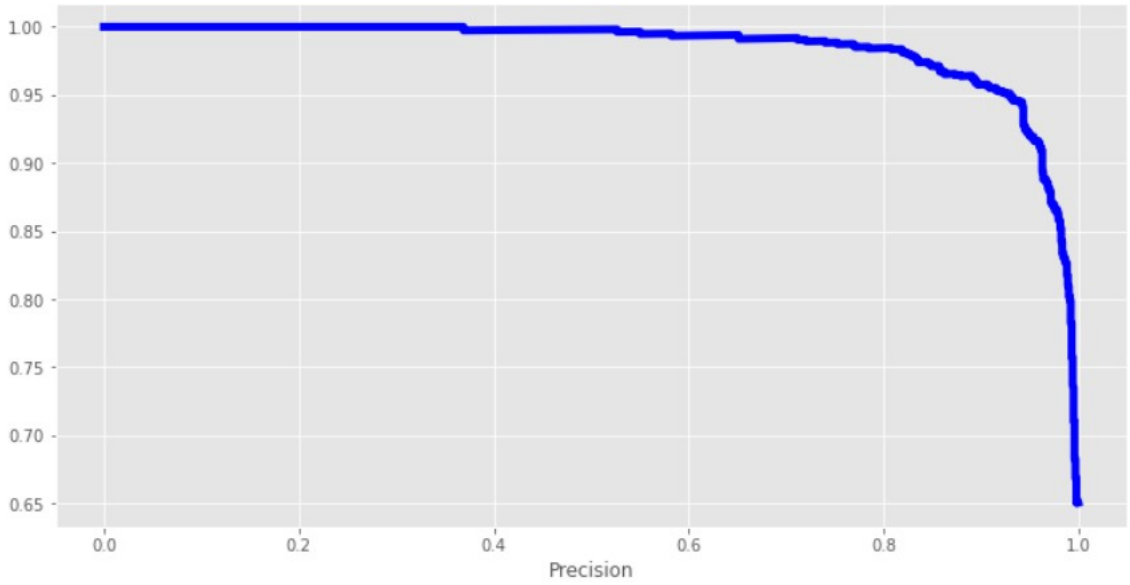
Best max_depth: 16
Best n_estimators: 200
Mean cross-validated accuracy score of the best_estimator: 0.941

Classification Report of RandomForest:

	precision	recall	f1-score	support
Not subscribed	0.91	0.86	0.89	1021
subscribed	0.87	0.92	0.89	1027
accuracy			0.89	2048
macro avg	0.89	0.89	0.89	2048
weighted avg	0.89	0.89	0.89	2048

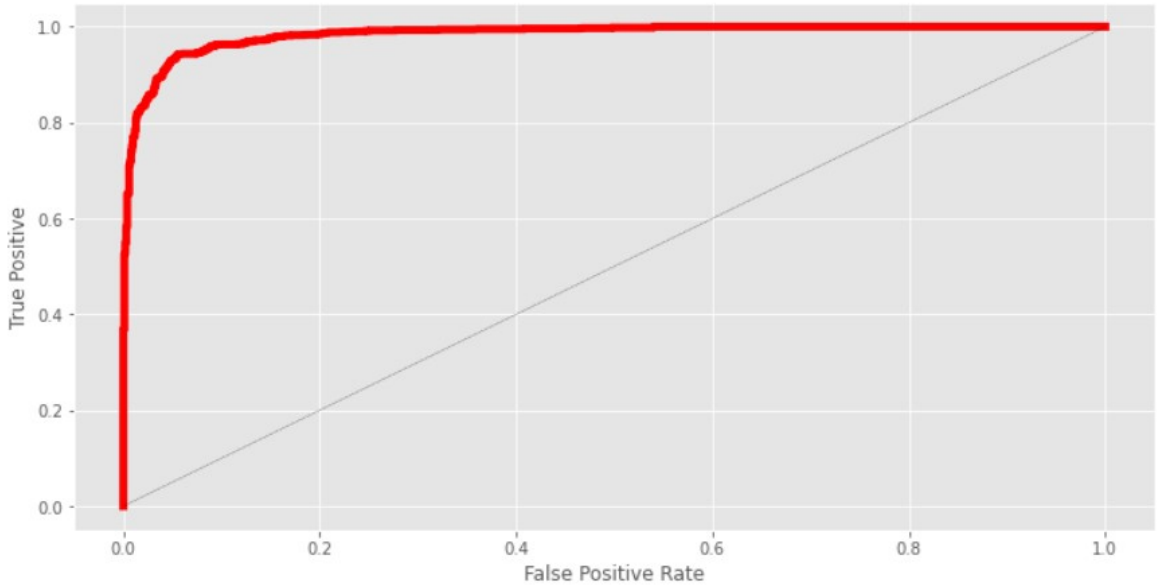
The accuracy and precision of RandomForestClassifier is 90% and 89% which is low as compared to DecisionTreeClassifier

Precision_recall_curve RandomForest :



Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. so as you can see the true positive rate and Positive prediction rate is very high of RandomForest as compared to the decision tree.

Roc_Auc_Curve oF RandomForest:



AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. This model predicts very well.

Conclusion/Recommendation:

Now we have done data visualization, preprocessing, and predictive analysis. In the predictive analysis, we found out that the person who has contact cellular, default no, married, and has secondary education are more likely to subscribe. The clustering shows that the candidates who are in management, with no housing and loan are more likely to subscribe. If we want to get more subscribers then we have to focus on the following candidates

1. **The candidates who have ages between 30 to 40.**
2. **The candidates who have a bank balance is equal to or greater than 0 euro**
3. **The candidates who belong to the management department and have secondary education.**
4. **The candidates have no personal loan and have no credit. The best month for subscription is march.**

which is the best predicted model?

DecisionTreeClassifier:

accuracy_score is 88%
precision_score is 84%
Recall_score is 90%

RandomForestClassifier:

accuracy_score is 91%
precision_score is 89%
Recall_score is 91%

So the best model is the Randomforest classifier. because it has high accuracy, precision, and recall.