### **BANK MARKETING-Fixed Deposit Prediction Model**

**NAME: SIVARAM.R Batch No:DS2309**

### **Business Use Case**

There has been a revenue decline for a Portuguese 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, so banks can invest in higher gain financial products to make a profit. In addition, banks also hold better chance to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. As a result, the Portuguese bank would like to identify existing clients that have higher chance to subscribe for a term deposit and focus marketing efforts on such clients.

**Project Description**

Your client is a retail banking institution. Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing and digital marketing. Telephonic marketing campaigns still remain one of the most effective ways to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

You are provided with the client data such as age of the client, their job type, their marital status, etc. Along with the client data, you are also provided with the information of the call such as the duration of the call, day and month of the call, etc. Given this information, your task is to predict if the client will subscribe to term deposit.

## About The Dataset

The dataset is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal of this dataset is to predict if the client or the customer of polish banking institution will subscribe a term deposit product of the bank or not.

**You are provided with following 2 files:**

1.     **train.csv :** Use this dataset to train the model. This file contains all the client and call details as well as the target variable “subscribed”. You have to train your model using this file.

2.     **test.csv :** Use the trained model to predict whether a new set of clients will subscribe the term deposit.

### **Dataset Attributes**

Here is the description of all the variables:

* Variable: Definition
* ID: Unique client ID
* age: Age of the client
* job: Type of job
* marital: Marital status of the client
* education: Education level
* default: Credit in default.
* housing: Housing loan
* loan: Personal loan
* contact: Type of communication
* month: Contact month
* day\_of\_week: Day of week of contact
* duration: Contact duration
* campaign: number of contacts performed during this campaign to the client
* pdays: number of days that passed by after the client was last contacted
* previous: number of contacts performed before this campaign
* poutcome: outcome of the previous marketing campaign

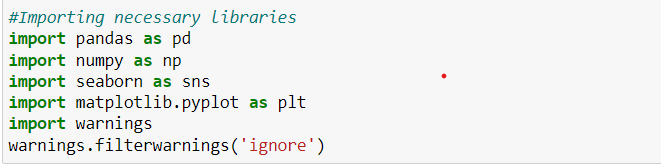
##### **Output variable (desired target):**

* **Subscribed (target):** has the client subscribed a term deposit? (YES/NO)

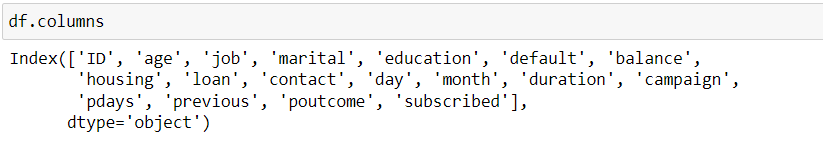
# **Problem Definition:**

All bank marketing campaigns are dependent on customer’s huge electronic data. The size of these data sources is impossible for a human analyst to come up with interesting information that will help in the decision-making process. Data mining models are completely helping in the performance of these campaigns. The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect the success based on a handful of algorithms that we will test (e.g. Logistic Regression, Random Forests, Decision Trees and others). With the experimental results we will demonstrate the performance of the models by statistical metrics like accuracy, sensitivity, precision, recall, etc. With the higher scoring of these metrics, we will be able to judge the success of these models in predicting the best campaign contact with the clients for subscribing deposit. The aim of the marketing campaign was to get customers to subscribe to a bank term deposit product. Whether they did this or not is variable ‘y’ in the data set. The bank in question is considering how to optimize this campaign in future.

**DATA ANALYSIS:**

****

We are importing necessary python libraries such as Pandas, NumPy, Seaborn, matplotlib.pyplot etc. Let’s look at the dataset features:



The features are:

1. ID – Unique no. for each client(numeric)
2. age – Age of the client (numeric).
3. job - Job of the client (Categorical).
4. marital – Martial status of the client (Categorical).
5. education – Education level of the client (Categorical).
6. default – Credit default of the client (Categorical - yes/no).
7. balance – Account balance of the client(numeric)
8. housing – Whether the client has housing loan or not (Categorical - yes/no).
9. loan - Whether the client has personal loan or not (Categorical - yes/no).

10) contact- Client’s type of communication (Categorical).

11) day - Day of week of contact (numeric).

12) month – Contact month (Categorical).

13) duration – Duration of contact with the client(numeric).

14) campaign - Number of contacts performed during this campaign to the client(numeric).

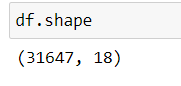
15) pdays - Number of days that passed by after the client was last contacted(numeric).

16) previous - Number of contacts performed before this campaign(numeric).

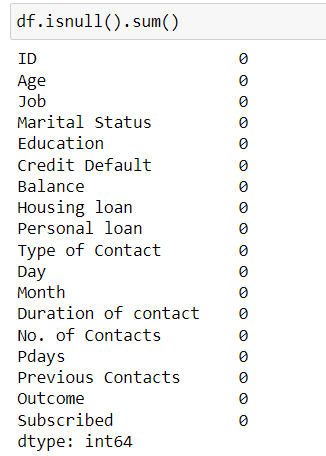
17) poutcome - Outcome of the previous marketing campaign (Categorical)

18) **subscribed(Target variable**) - has the client subscribed to a term deposit? (Categorical YES/NO)

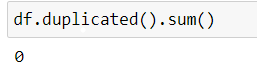
From the given features about our client, we need to build a model to predict whether the client will subscribe to the term deposit or not. For prediction, we are given new set of client data’s i.e. **test.csv.**



This dataset contains 31647 rows and 18 columns.

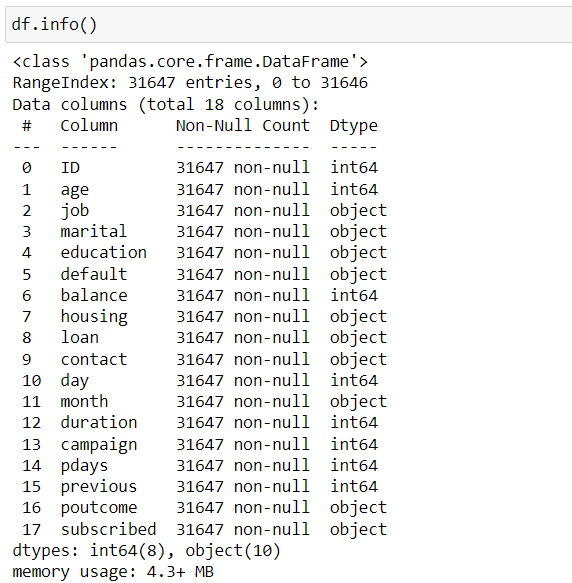


Here, we can see that no null values are present.



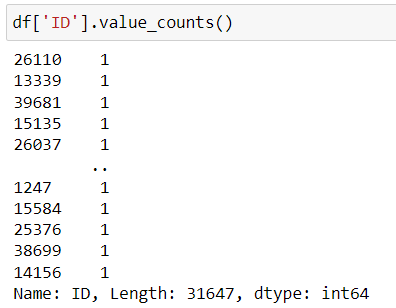
There are no duplicates found in our dataset, now we will go for data exploration.

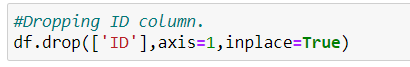
**Data Exploration:**

****

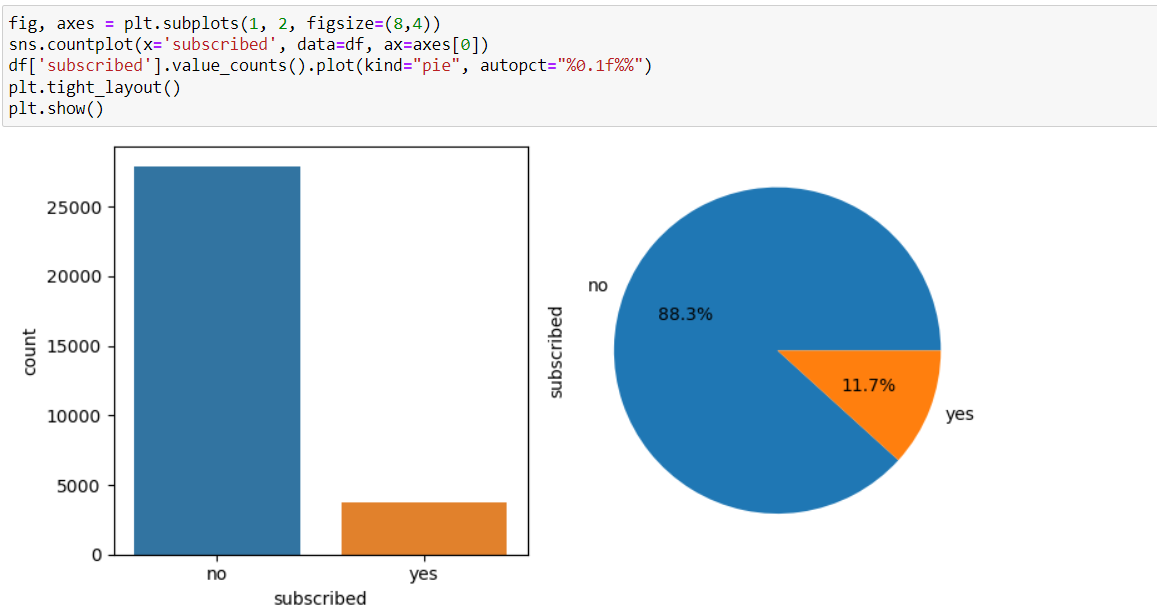
ID COLUMN:

The ‘ID’ column contains unique no. for all clients, so it will make any impact in predicting the target variable, so we will drop this column.



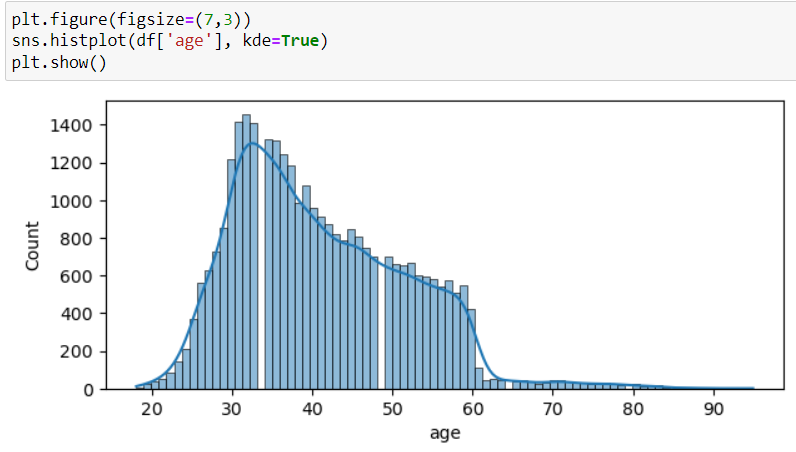


Subscribed (target variable):



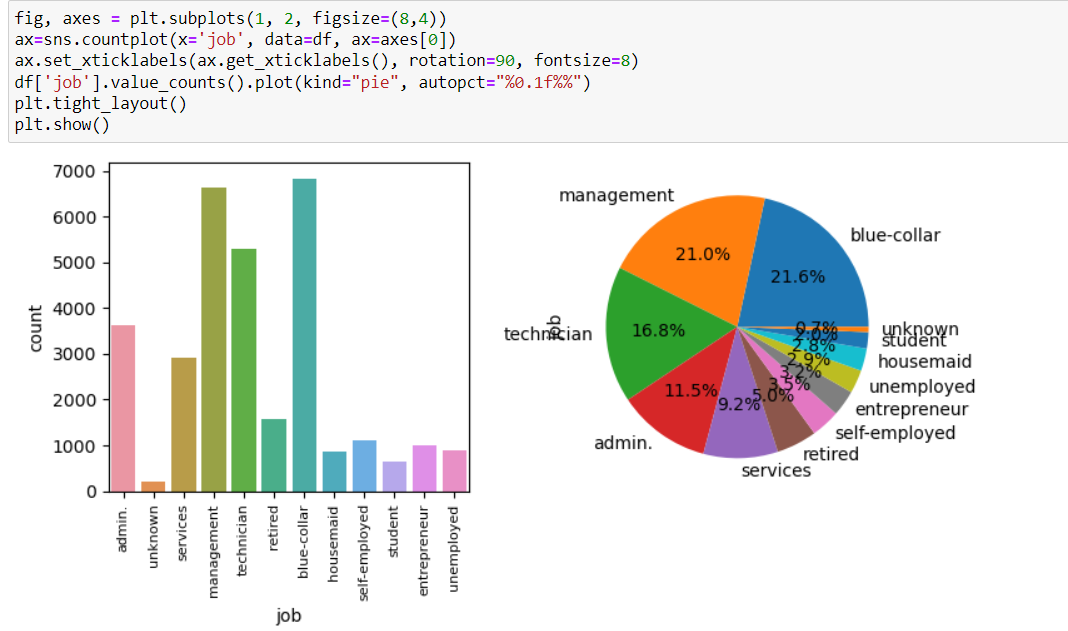
* Majority of the clients did not subscribe for fixed deposit.
* 88.3% of clients did not have subscription, only 11.7% subscribed to fixed deposit.

Age column:

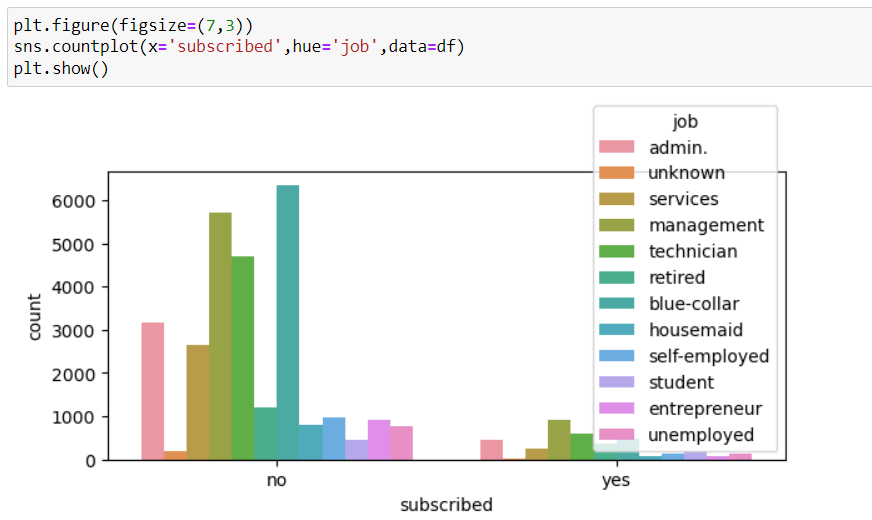


* Larger no. of clients is around the age (25 to 60), compared to the older clients.
* In age column we can see that, right side skewness is present.

Job Column:

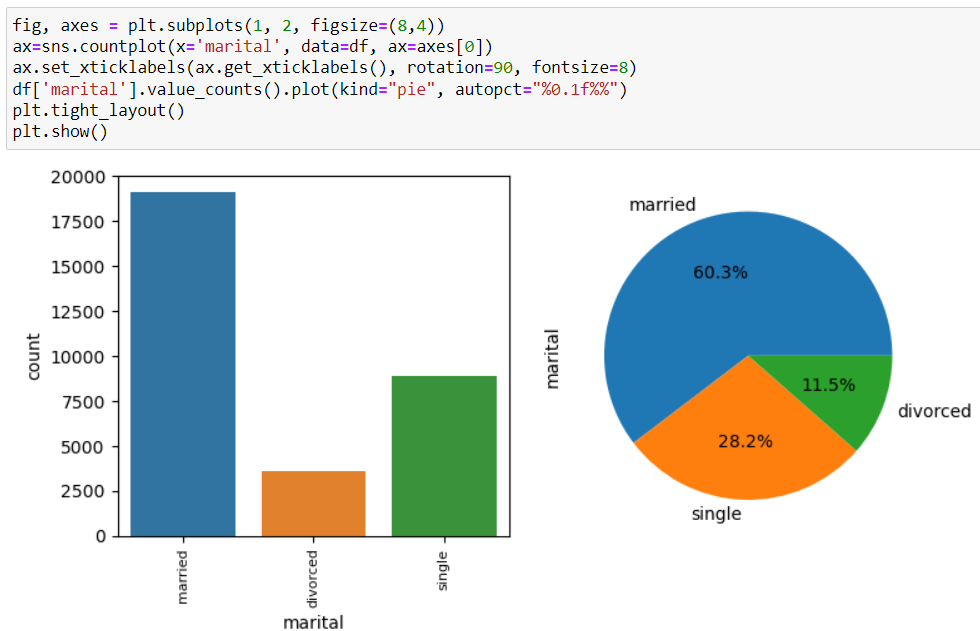


* Larger no. of clients is from blue collar (21.6%) and management (21%) jobs.
* Very few clients are unemployed and jobs unknown.

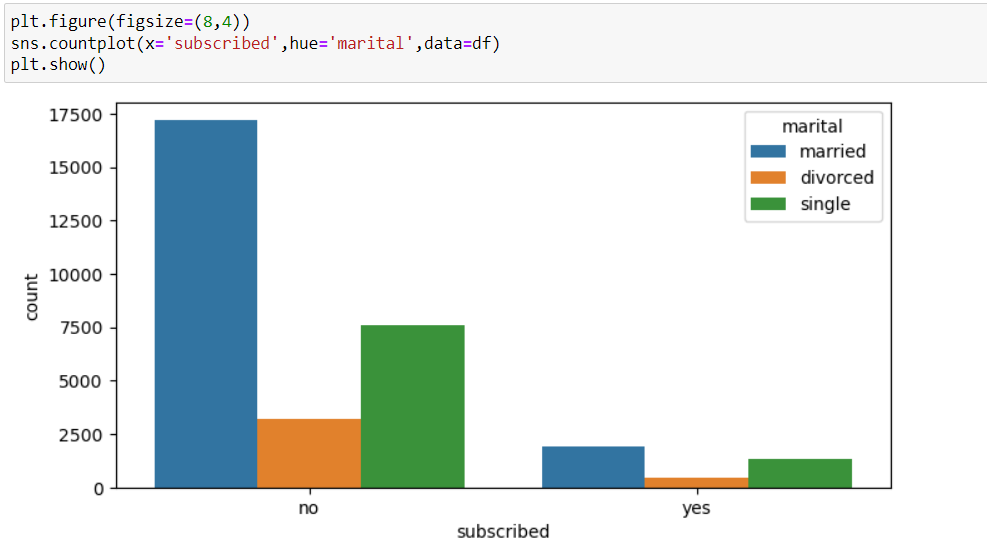


* Clients having job in management have subscribed to fixed deposit more compared to others.
* Clients in jobs like admin, technician and blue collar have subscribed to fixed deposit more.
* Clients whose jobs are unknown did not have any subscriptions.

Martial column:

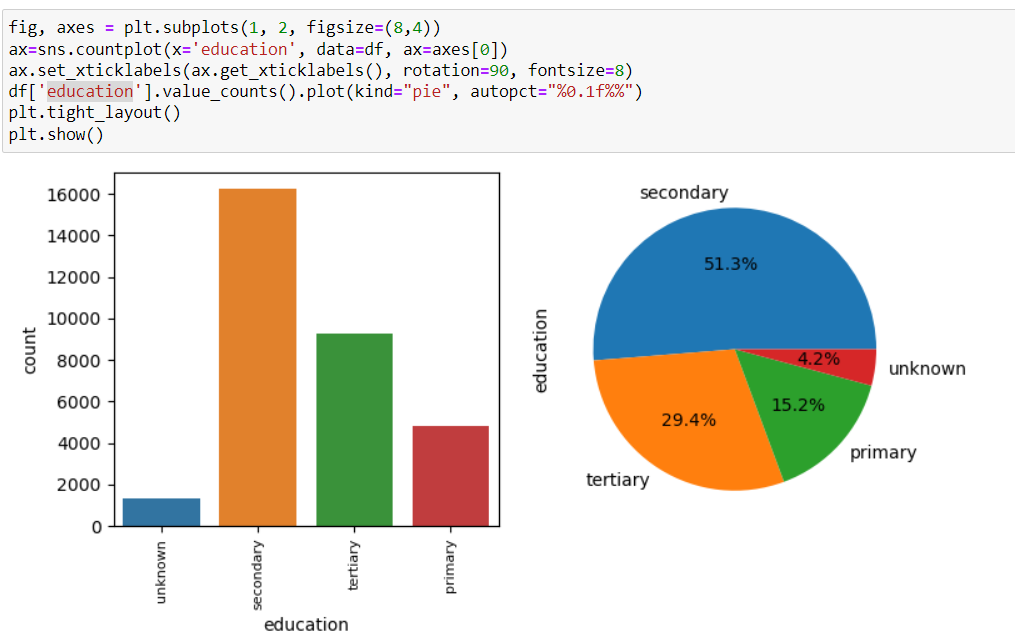


* Larger no. of clients is married nearly 60%.
* 28.2% clients are Single and 11.5% clients are divorced.

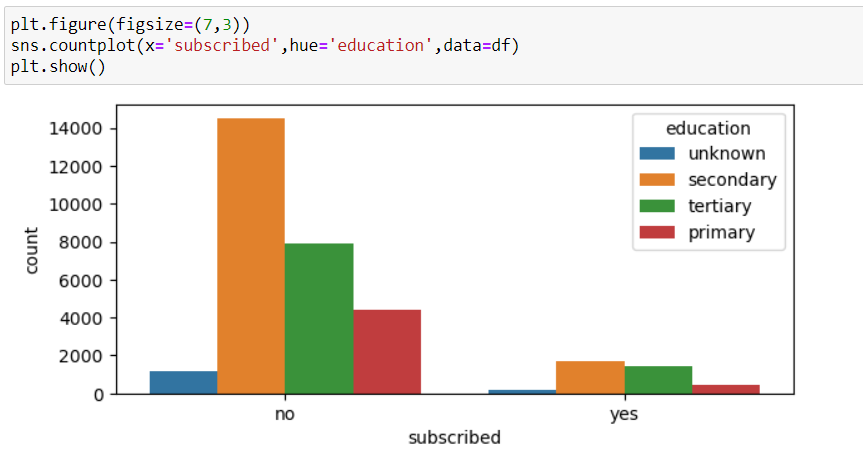


* Married clients have subscribed to fixed deposits more when compared to others.
* Single clients also subscribed to fixed deposits.
* Divorced clients have very less subscription.

Education column:

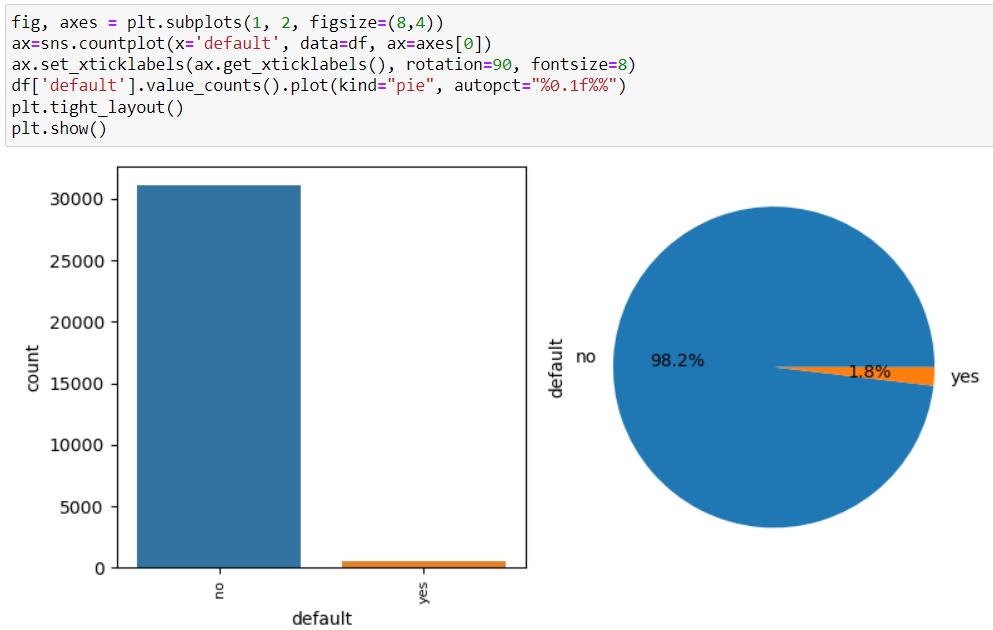


* Most of the clients have secondary level education (51.3%).
* Nearly 30% of clients have tertiary education level, 15% of clients have primary level education.

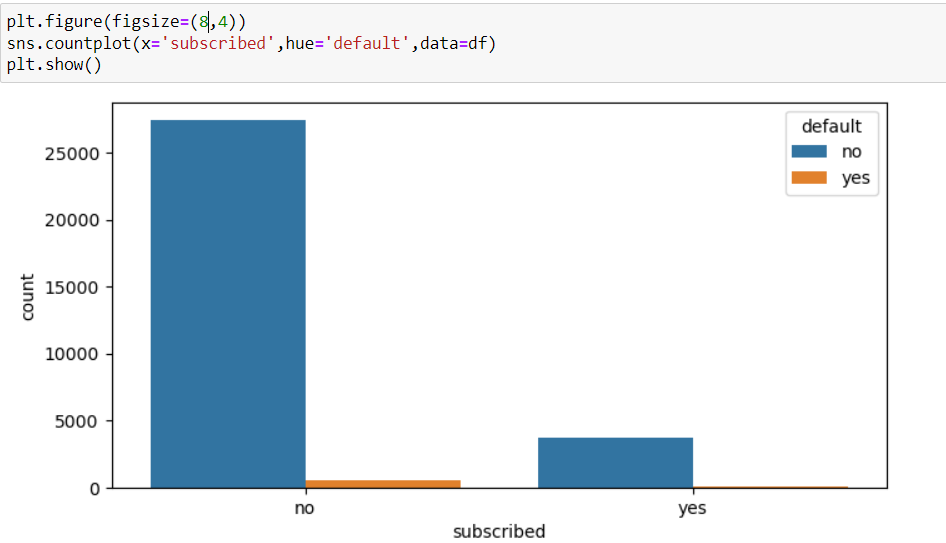


* Clients who have secondary and tertiary level of education have subscribed to fixed deposits more when compared to others.
* Clients who have primary level of education and whose education is unknown have very few subscriptions status.

Default column:

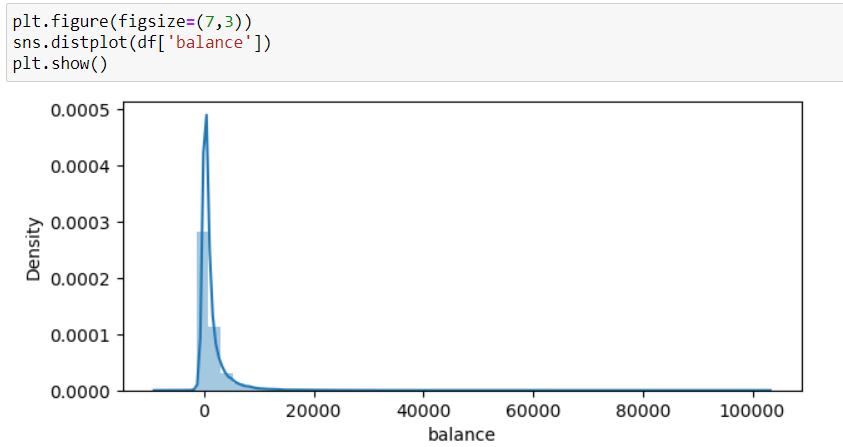


* Larger no. of clients (98.2%) did not have any credit default, only 1.8% of clients have credit default.



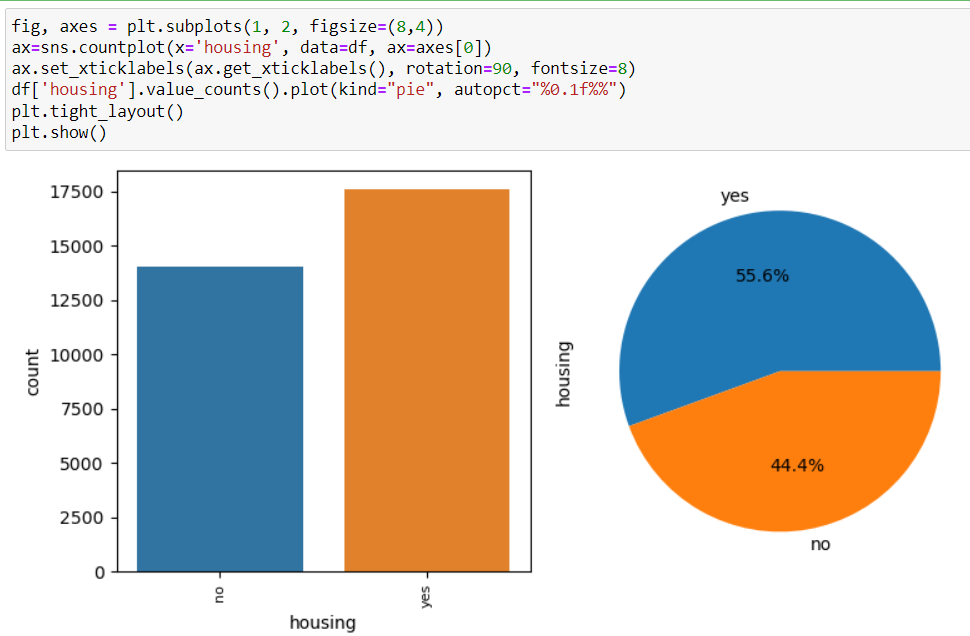
* Only clients who does not have credit default subscribed to fixed deposit.
* Not one client with credit default have fixed deposit subscription.

Balance column:

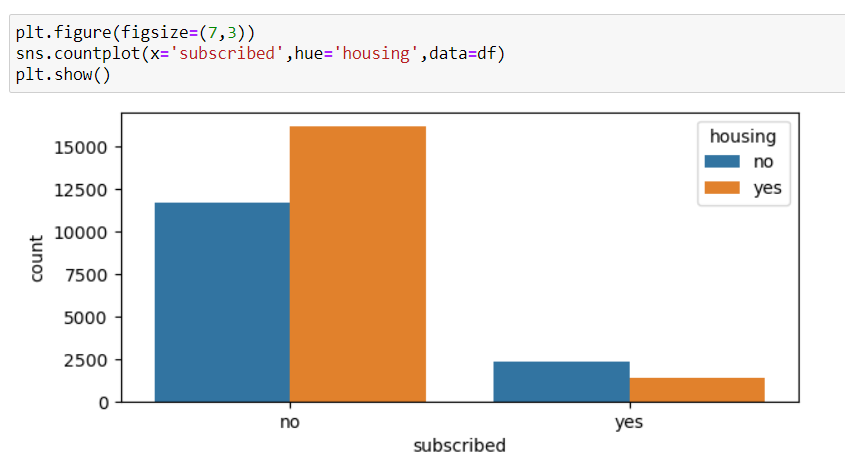


* Most of the clients have Zero balance.
* We can see that right side skewness is present here.

Housing column:

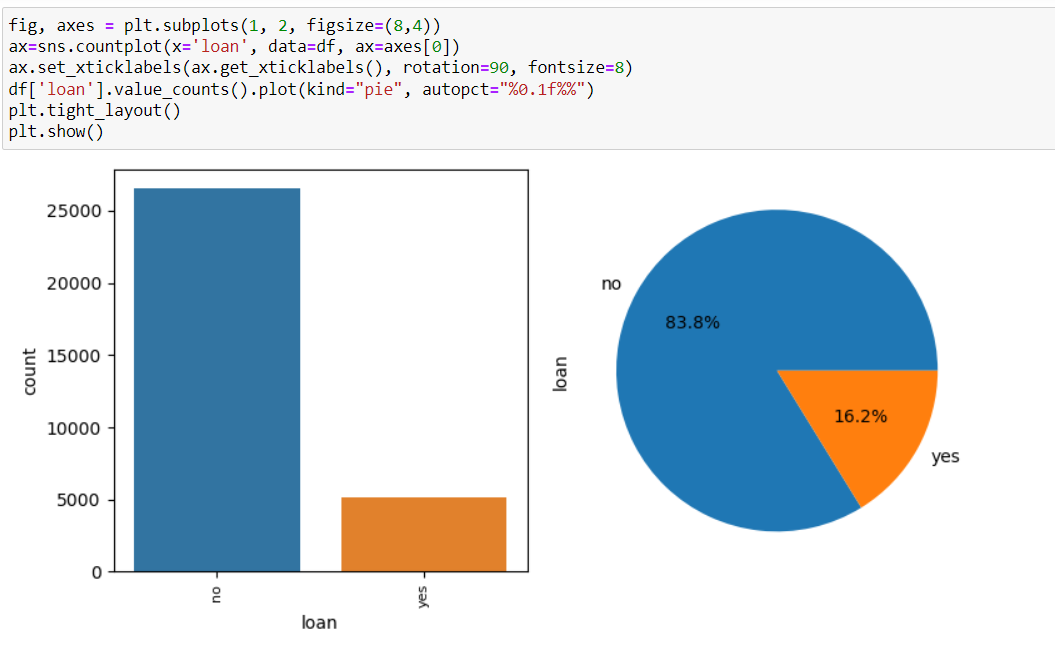


* More than half of the clients acquired the housing loan, nearly 56%.
* 44% of clients did not have housing loans.

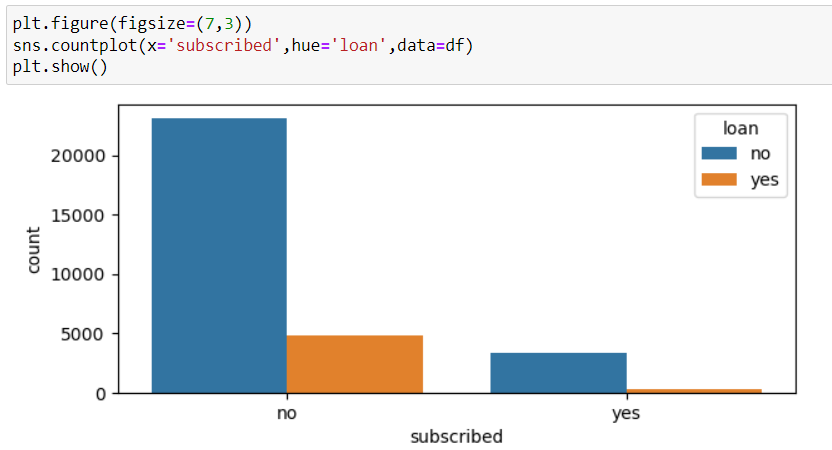


* Clients who did not have housing loan subscribed to fixed deposits more, when compared to clients having housing loan.

Loan column:

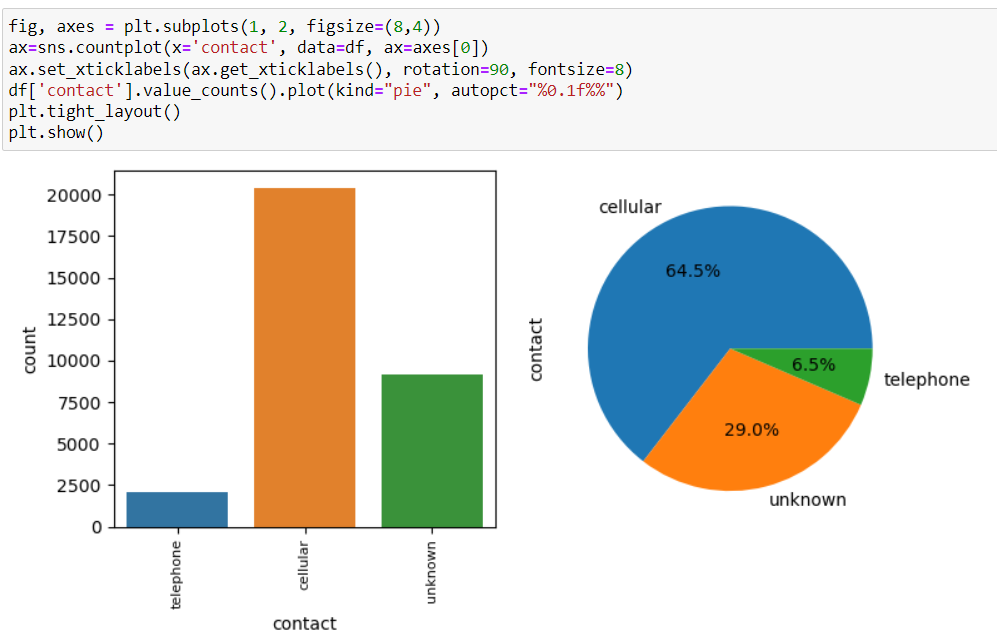


* Larger no. of clients did not have any personal loans, nearly 83.8%
* Only 16% of clients have personal loans.

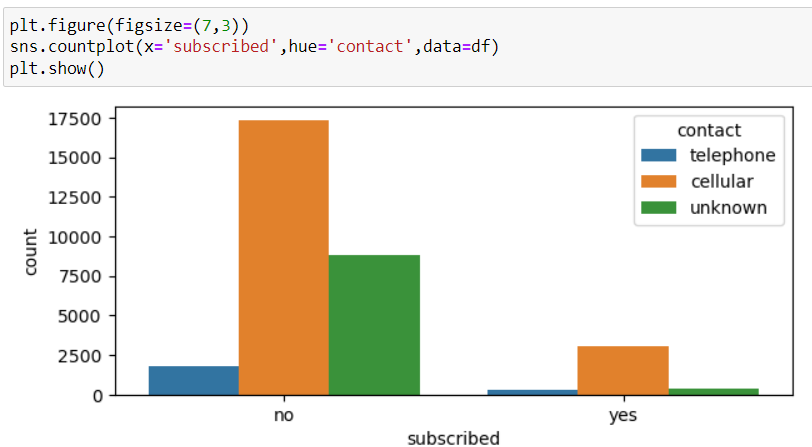


* Majority of the clients who did not have personal loan subscribed to fixed deposits.
* Only very few clients having personal loan subscribed to fixed deposits.

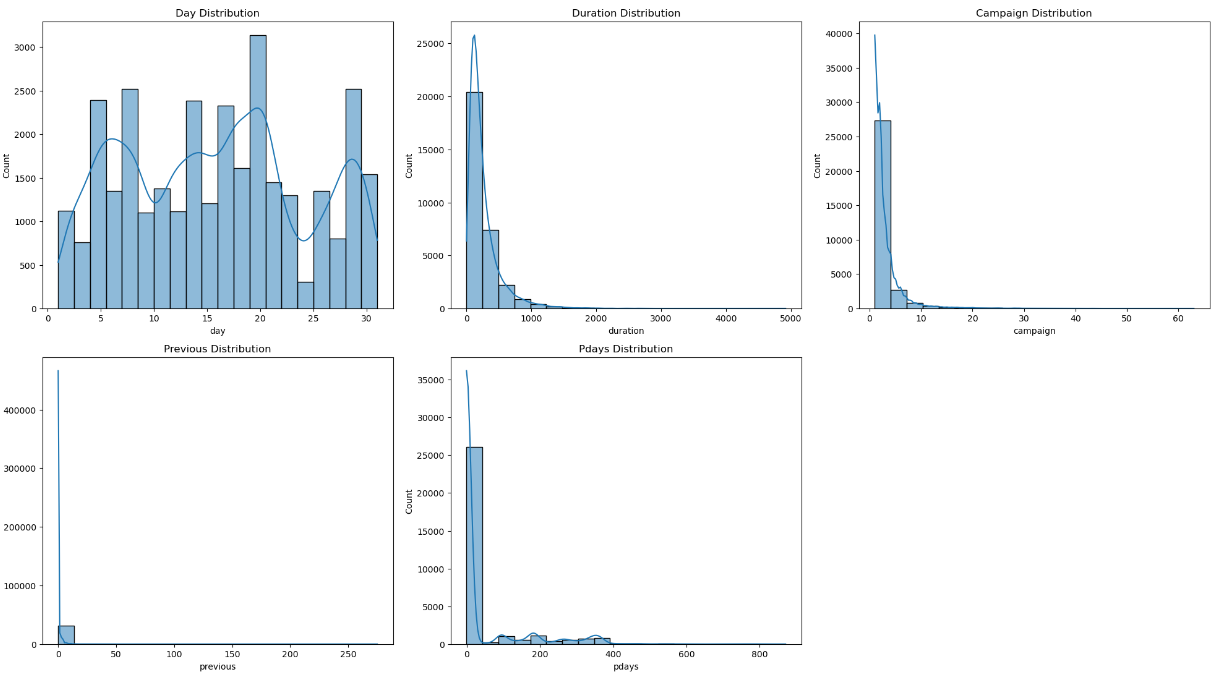
Contact column:



* Most of the clients have cellular type of communication (64.5%).
* Only few clients (6.5%) have telephone communication.
* Nearly 29% of client’s type of communication unknown.

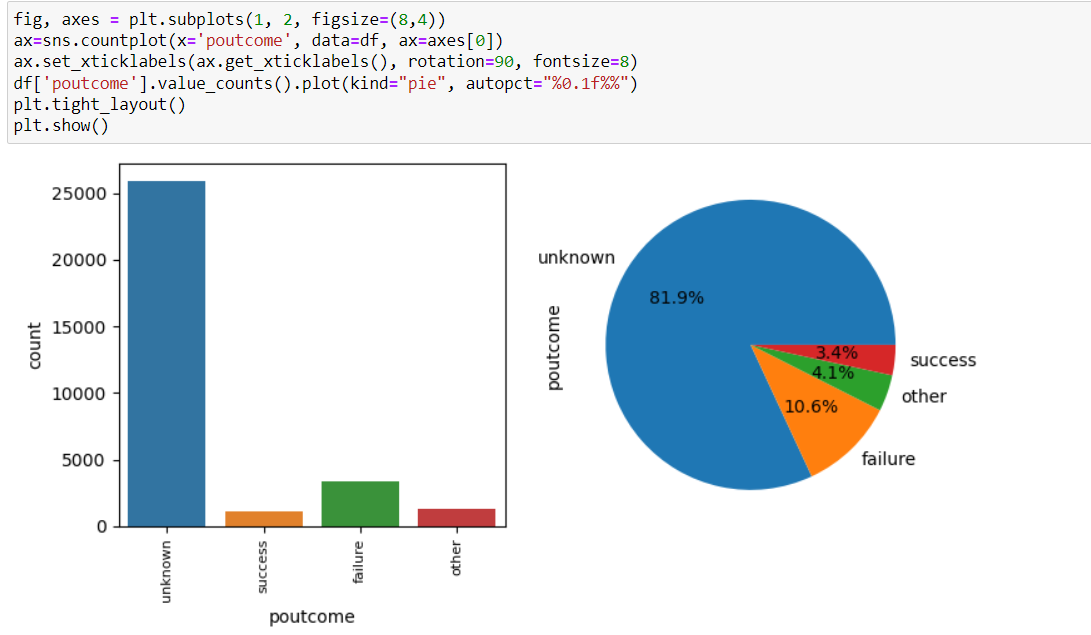


* Majority of clients who has cellular type of communication subscribed to fixed deposits.

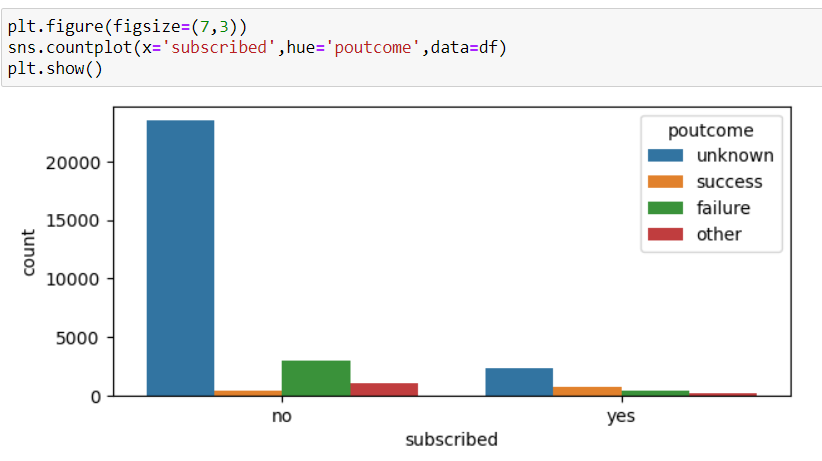


* Day Distribution: Shows which day of the month clients were contacted
* Duration Distribution: Represents the contact duration with clients
* Campaign Distribution: Indicates the number of contacts performed during this campaign for each client.
* Previous Distribution: Reflects the number of contacts before the campaign with each client.
* Pdays Distribution: Shows the number of days passed by, after the client was last contacted from a previous campaign.

Outcome Column:



* Outcome of the most of the client’s previous campaign are unknown (81.9%).
* Only 3.4% are success and 10.6% are failure.



* Most of the clients who subscribed to fixed deposits their previous campaign outcomes are unknown.

**EDA Concluding Remarks:**

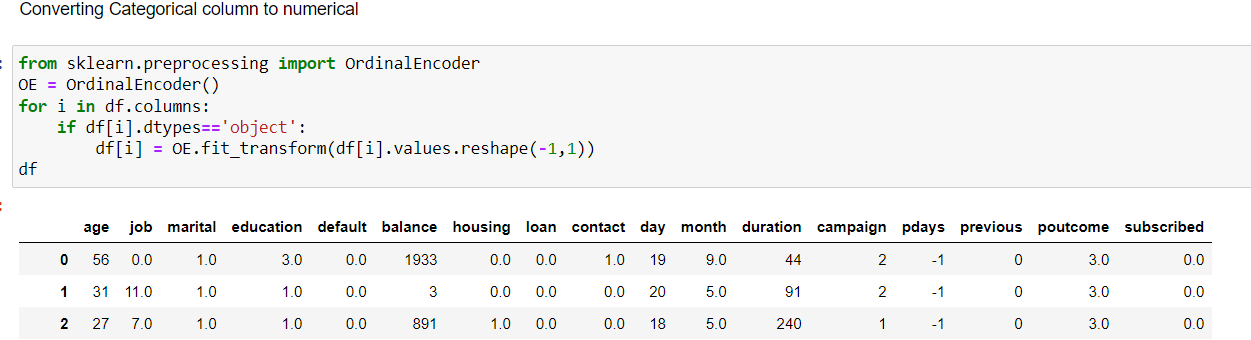
* Clients having job in management have subscribed to fixed deposit more compared to others.
* Married clients have subscribed to fixed deposits more when compared to Single and Divorced.
* Clients who have secondary and tertiary level of education have subscribed to fixed deposits more when compared to others.
* Only clients who did not have any credit defaults, subscribed to fixed deposit.
* Clients who did not have housing loan subscribed to fixed deposits more, when compared to clients having housing loan.
* Majority of the clients who did not have any personal loans, subscribed to fixed deposits.
* Majority of clients who have cellular type of communication subscribed to fixed deposits.

**DATA PREPROCESSING:**

After Data Analysis, we need to do preprocessing before we step in to model building. The data preprocessing includes, encoding categorical column in to numerical, checking correlation between the feature and the target variable, checking skewness and removing it, checking outliers, removing the biasness in the dataset, checking multicollinearity between features and balancing the dataset in case of classification to avoid under/over fitting.

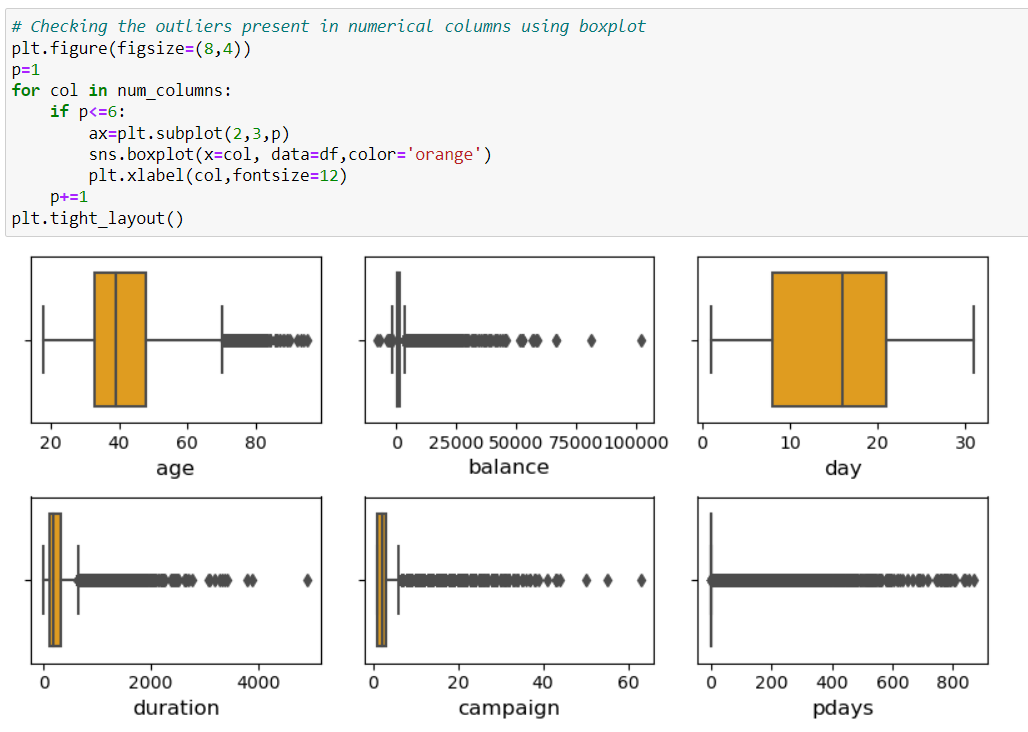
**Encoding categorical column:**

Since, the dataset contains categorical columns, we need to change it to numerical columns, here we use Ordinal Encoder,

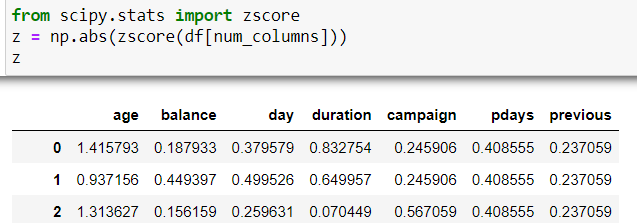


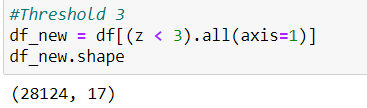
Here, we can see that, all columns are changed to numerical.

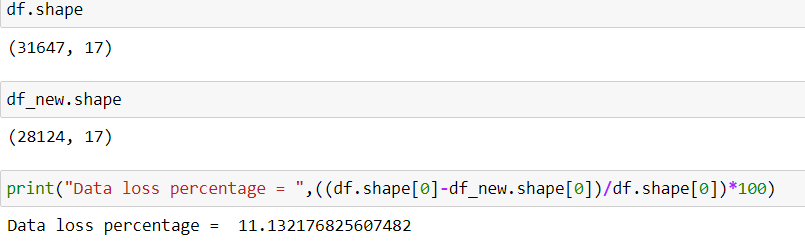
**Outliers Detection:**

****

Here, we can see that outliers are present, we will remove it using z score method.

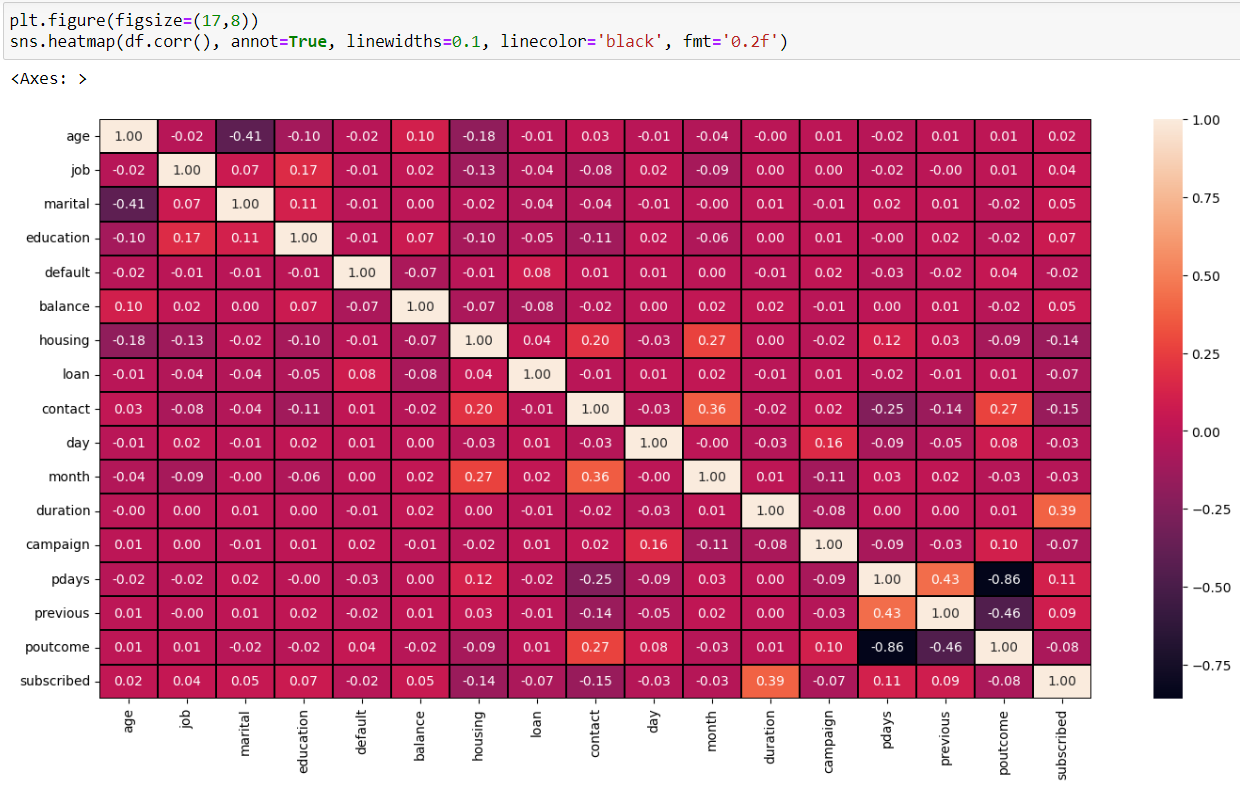






Here, we can see that we are losing more than 11.13% of data. We can’t afford to lose more than 10% of data, so we will not remove the outliers.

**Correlation between features:**

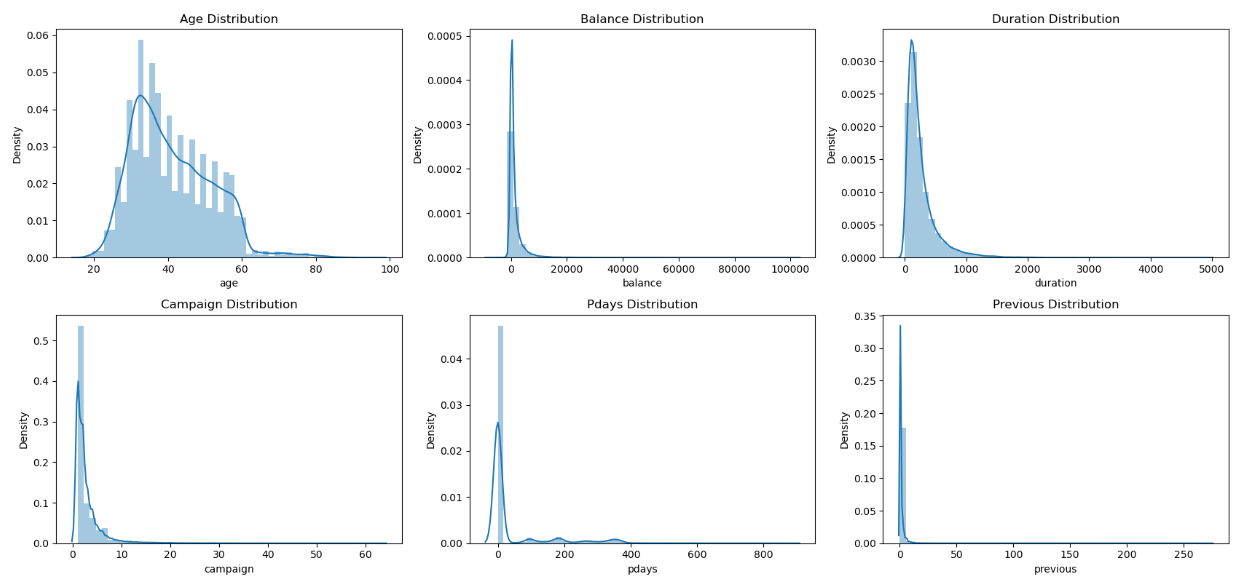


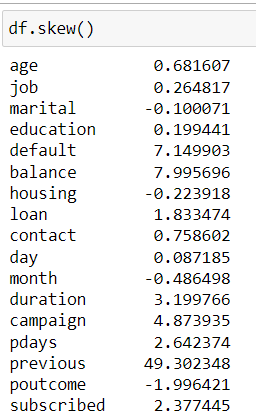
* The heatmap above visualizes the correlation matrix of the dataset.
* This visualization can help identify which features are most related to each other.
* Darker colours indicate stronger correlations i.e. features ‘pdays’ and ‘poutcome’ is strong negatively related and features ‘poutcome’ and ‘previous’ are also negatively and strongly related to each other.
* The annotations on the heatmap provide the exact correlation percentages between the variables.
* The target variable ‘subscribed’ has positive correlation with feature ‘duration’.

**Checking Skewness:**

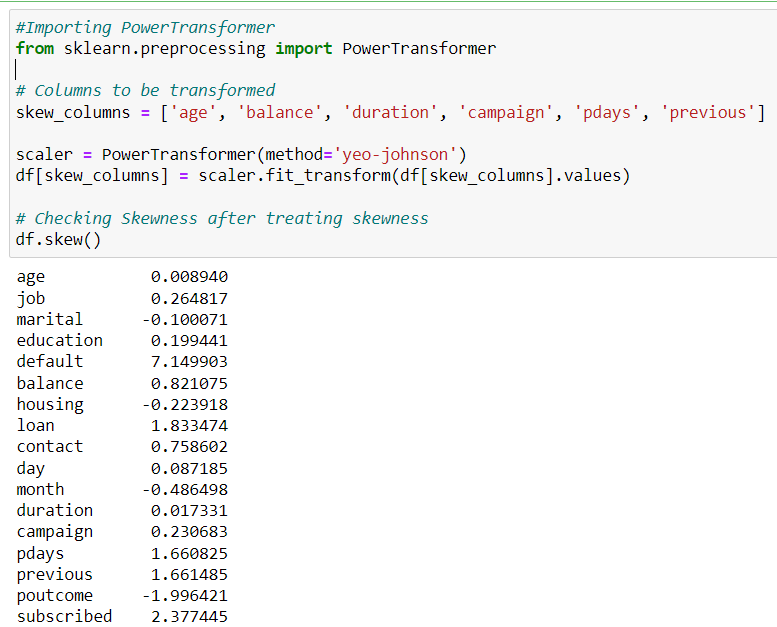
We already know that skewness presents in the numerical columns,

Let’s see that using distribution plots:

****

****

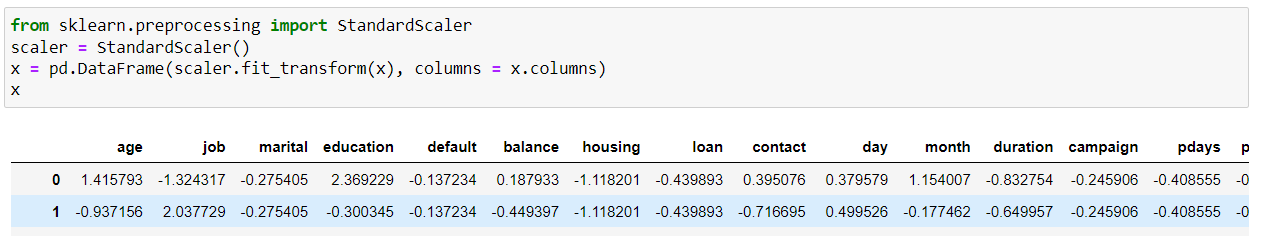
The Threshold for skewness is -0.5 to +0.5. Here, we can see that skewness presents in ‘age’, ‘balance’, ‘duration’, ‘pdays’, ‘previous’ and ‘campaign’. We will remove skewness using Power Transformer using ‘yeo-johnson’ method.



We, can see that skewness has been removed drastically. Also, no need to remove skewness for categorical features and target variables.

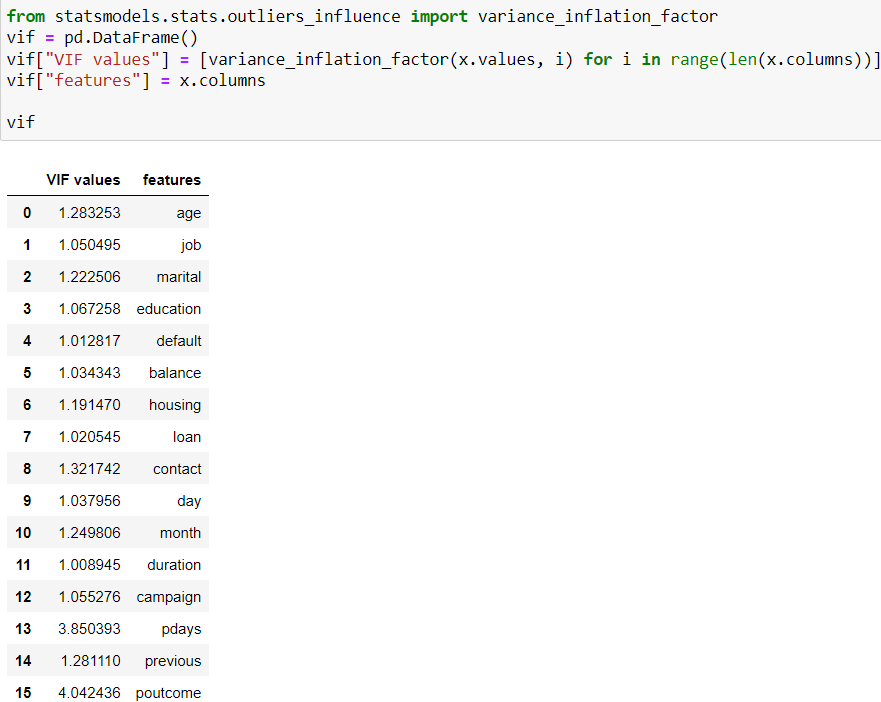
**Scaling the Dataset:**

Feature Scaling is a critical step in building accurate and effective machine learning models. One key aspect of feature engineering is scaling, normalization, and standardization, which involves transforming the data to make it more suitable for modelling. These techniques can help to improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale. Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. We are scaling the dataset using Standard Scaler method.

****

**Checking Multicollinearity:**

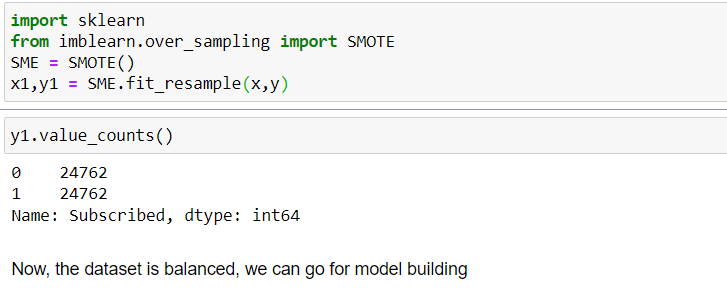
Multicollinearity occurs when independent variables in a model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. Here, we will use Variance Inflation Factor to check multicollinearity.

****

Here, we can see that VIF values less than 5 , so no multicollinearity present between features.

**Balancing the dataset:**

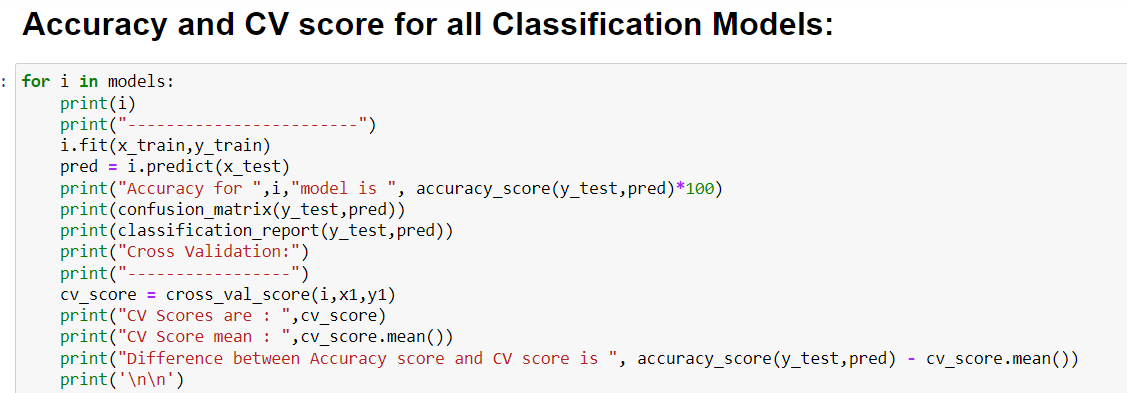
Since it is a classification problem, in order to avoid the over/under sampling issues, we will balance the data using oversampling method.



**Building Machine LEARNING Models:**

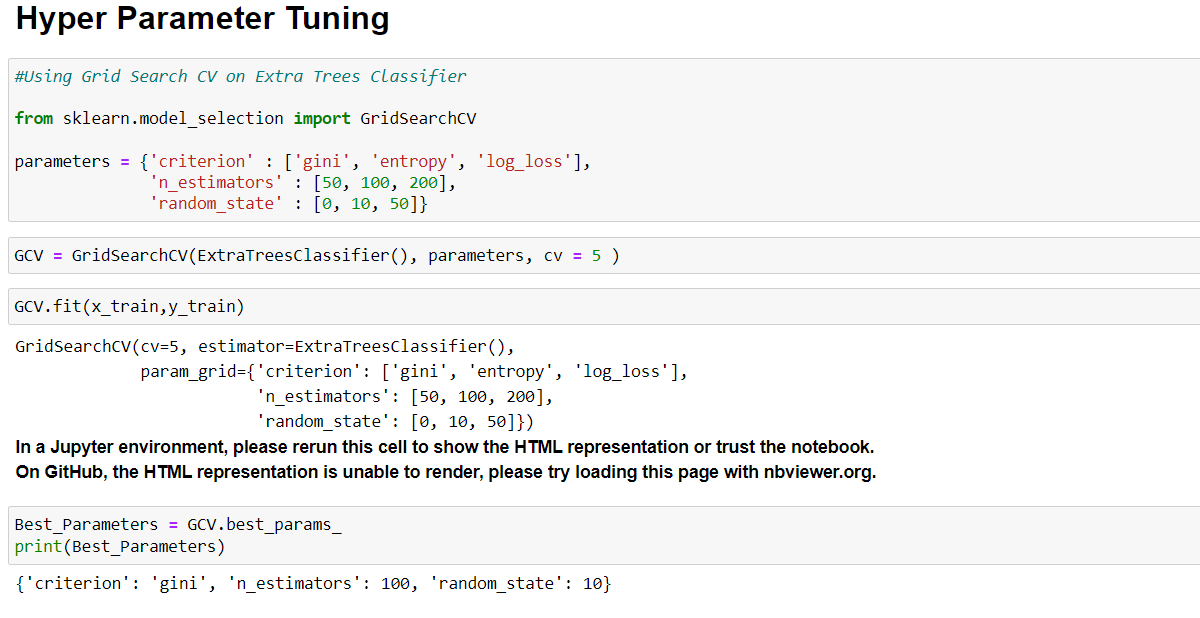
Since it is a classification problem, we will build model using classification algorithms such as Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Gaussian Naive Baies, Extra Tree Classifier, Gradient Boosting Classifier, Ada Boosting Classifier, Bagging Classifier, KNN Neighbors Classifier, Support Vector Classifier. We will use Cross Validation also. The main purpose of cross validation is to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data.

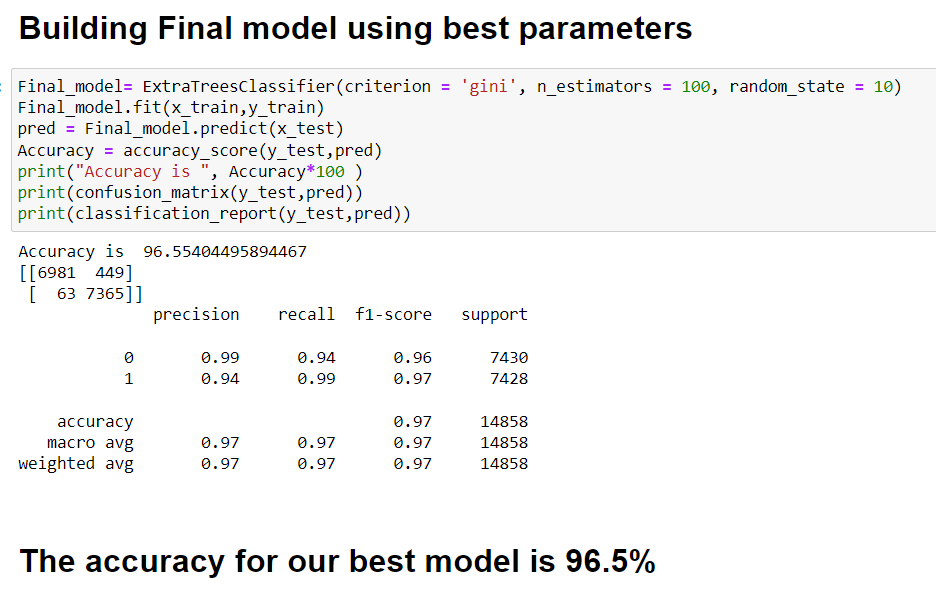


****

After checking all the model’s accuracy and cross validation score, Extra Tree Classifier is our best model because it has highest accuracy and highest cross validation score mean also. We will do Hyper parameter tuning, to get the best parameters for our best model and build the final model for prediction.

**Hyper Parameter Tuning:**

****

****

**Extra Trees Classifier Report:**

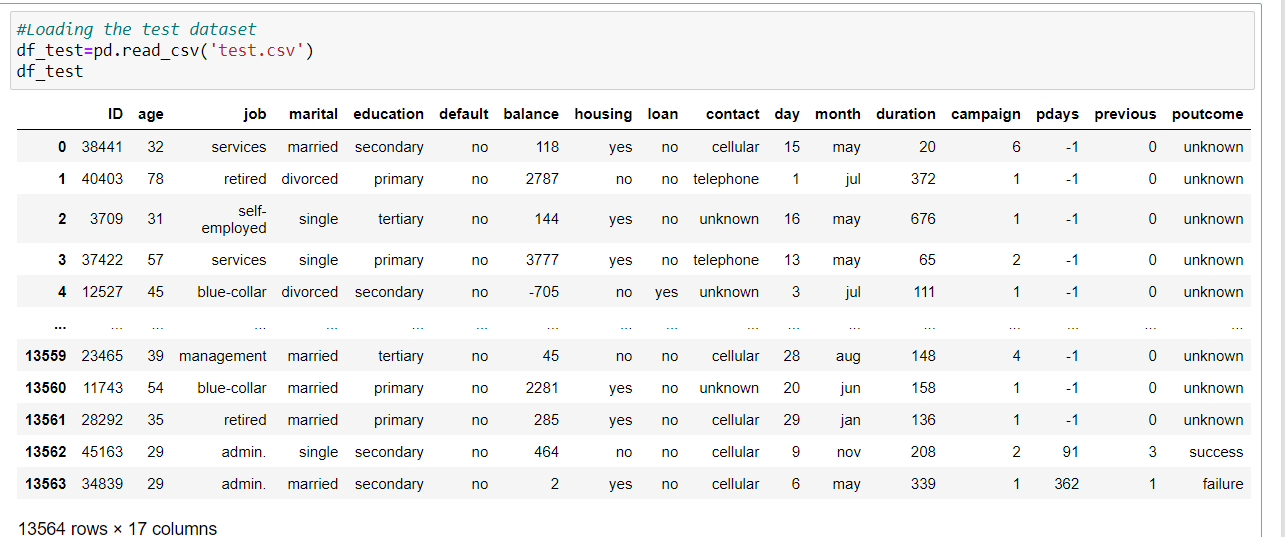
**Accuracy**: 96.5%

**Confusion Matrix:** [6981 449]

[63 7365]

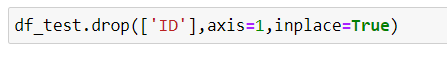
**Prediction in test.csv:**

**Now we will use the trained model to predict whether a new set of clients in test.csv will subscribe to term deposits or not.**

****

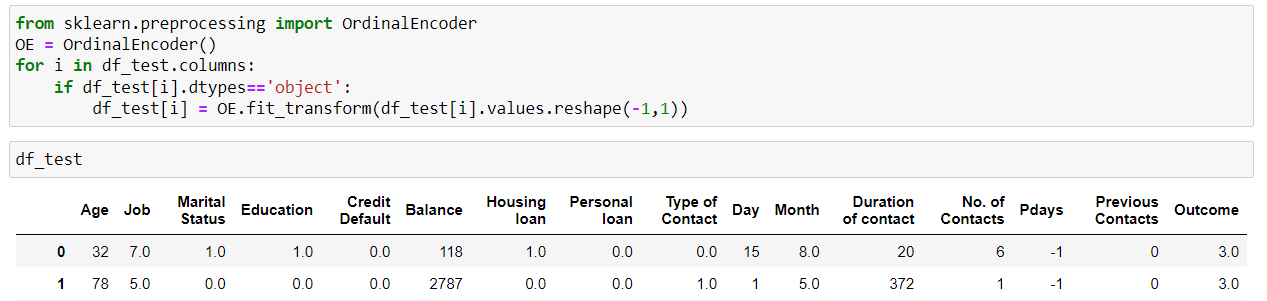
**Dropping Id column:**

We need to drop the ‘ID’ column as per our trained model.

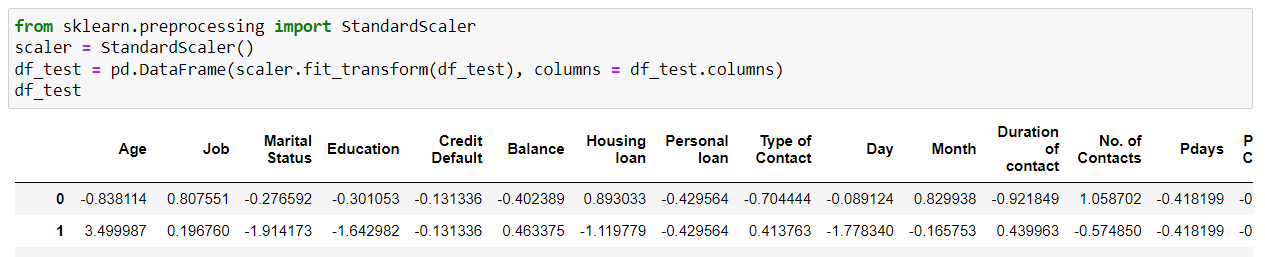
****

**Encoding Categorical column:**

We need to change the categorial columns to numerical for the test datset.

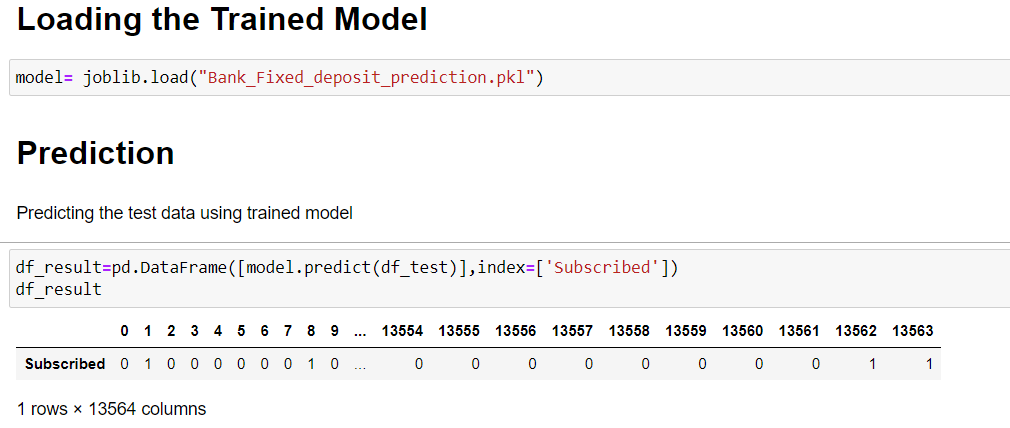
****

**Scaling the test Dataset:**

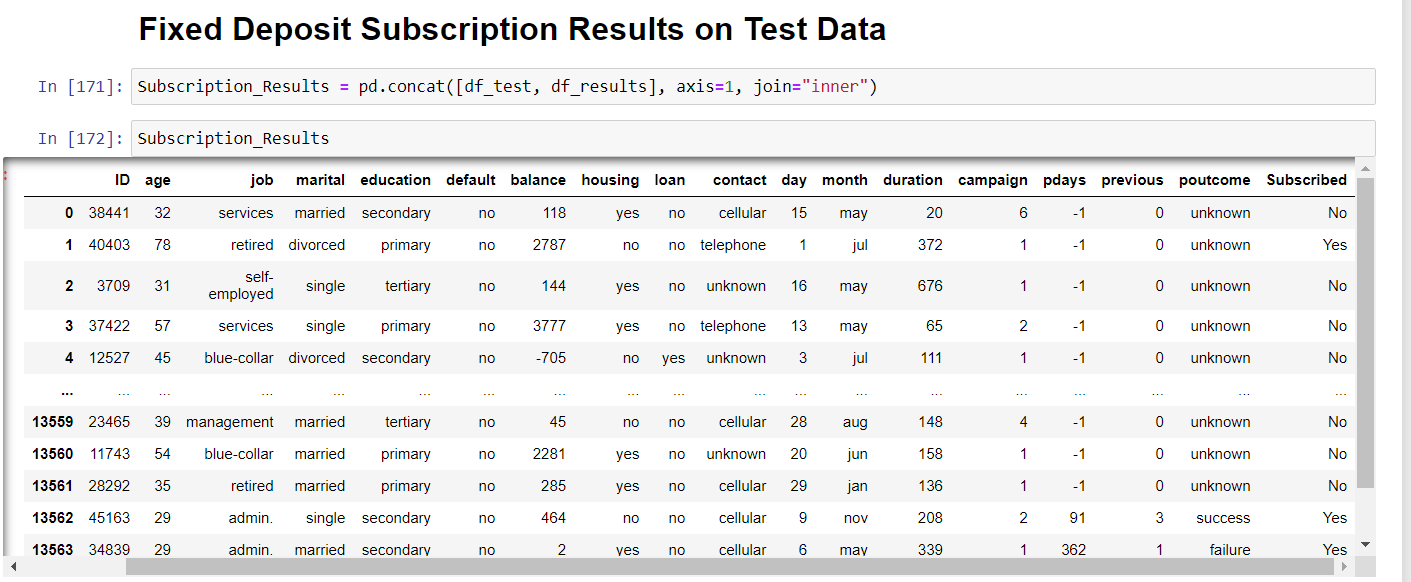
****

**Prediction:**

Now, loading the trained model and making predictions for the test dataset.



**Subscription results for test.scv:**

****

**CONCLUSION:**

The Machine Learning Architecture occupies the major industry interest now as every process is looking out for optimizing the available resources and output based on the historical data available, additionally, machine learning involves major advantages about data forecasting and predictive analytics when coupled with data science technology. The machine learning architecture defines the various layers involved in the machine learning cycle and involves the major steps being carried out in the transformation of raw data into training data sets capable for enabling the decision making of a system. The most important and time-consuming part of the problem was data cleansing and processing. Once the data was prepared and ready, the next challenge was to pick an algorithm that could be best suited for the problem we choose to solve.

In our case, out of all features, Number of contacts made with the clients(duration) has played major role in deciding client’s Fixed Deposit subscription. With experience and prior knowledge and also the outcome of accuracy on training data, we observed that **Extra tree classification** performed the best out of others.