**Customer Churn prediction**

**Dhruv kumar Dhandhukiya   
 date: 3/2/2023**

Objective:

To build a predictive model that can predict customer churn for a given company.

Data Description:

1) age: (numeric)

2) job: type of job   
(categorical: “admin”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”, “self-employed”, “retired”, “technician”, “services”)

3) marital: marital status  
(categorical: “married”, “divorced”, “single”)  
divorced means: divorced or widowed

4) education:  
(categorical: “unknown”, “secondary”, “primary”, “tertiary”)

5) default: has credit in default  
(binary: “yes”, “no”)

6) balance: average yearly salary in euros  
(numeric)

7) housing: has housing loan?  
(binary: “yes”, “no”)

8) loan: has personal loan?  
(binary: “yes”, “no”)

9) contact: contact communication type   
(categorical: “unknown”, “telephone”, “cellular”)

10) day: last contact day of the month  
(numeric)

11) month: last contact month of the year  
(categorical: “Jan”, …, “Dec”)

12) duration: last contact duration in seconds   
(numeric)

13) campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

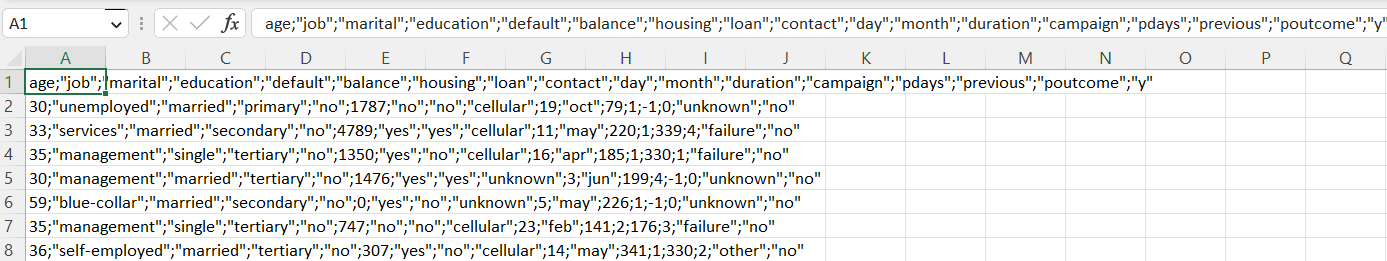
14) pdays: number of days that passed by after the client was last contacted from a previous campaign  
 (numeric, -1 means client was not previously contacted)

15) previous: number of contacts performed before this campaign and for this client (numeric)

16) poutcome: outcome of the previous marketing campaign  
(categorical: “unknown”, “other”, “failure”, “success”)

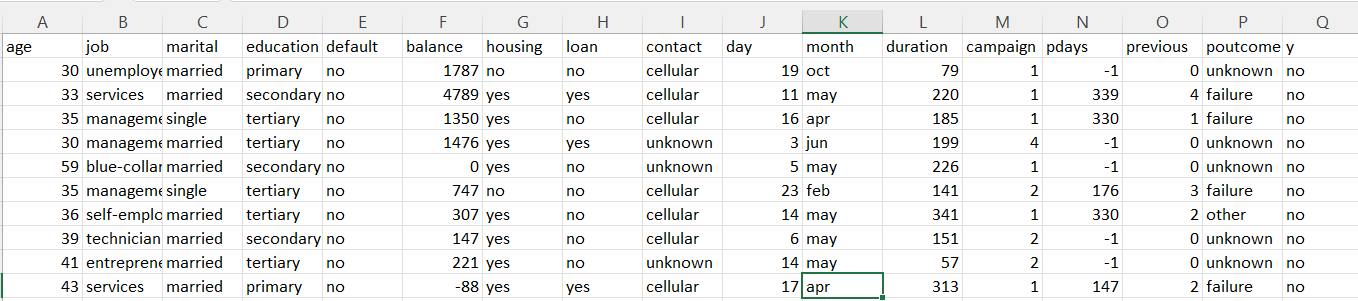
17) y: has the client subscribed a term deposit?  
(binary: “yes”, “no”)

Dataset:

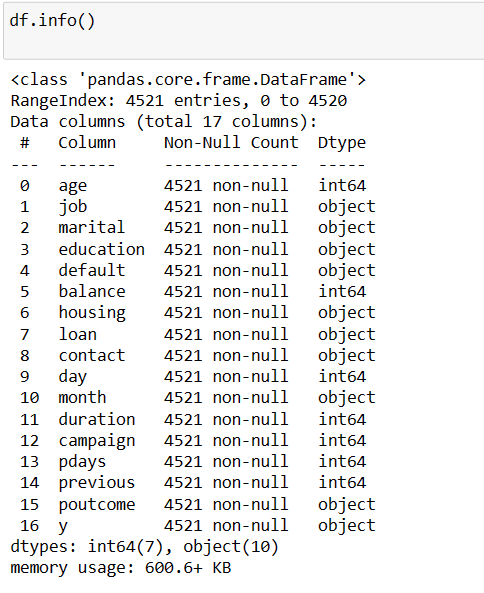


Since the raw data was not proper.  
It was having values separated by semicolon.   
Using excel this semicolon values were converted to excel column.

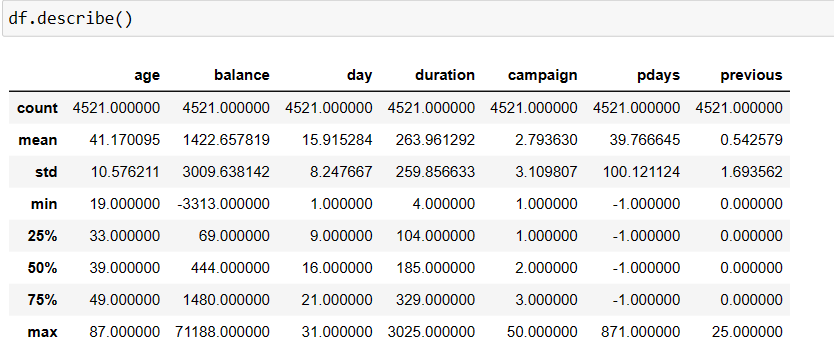
New Dataset:



Data Exploration:



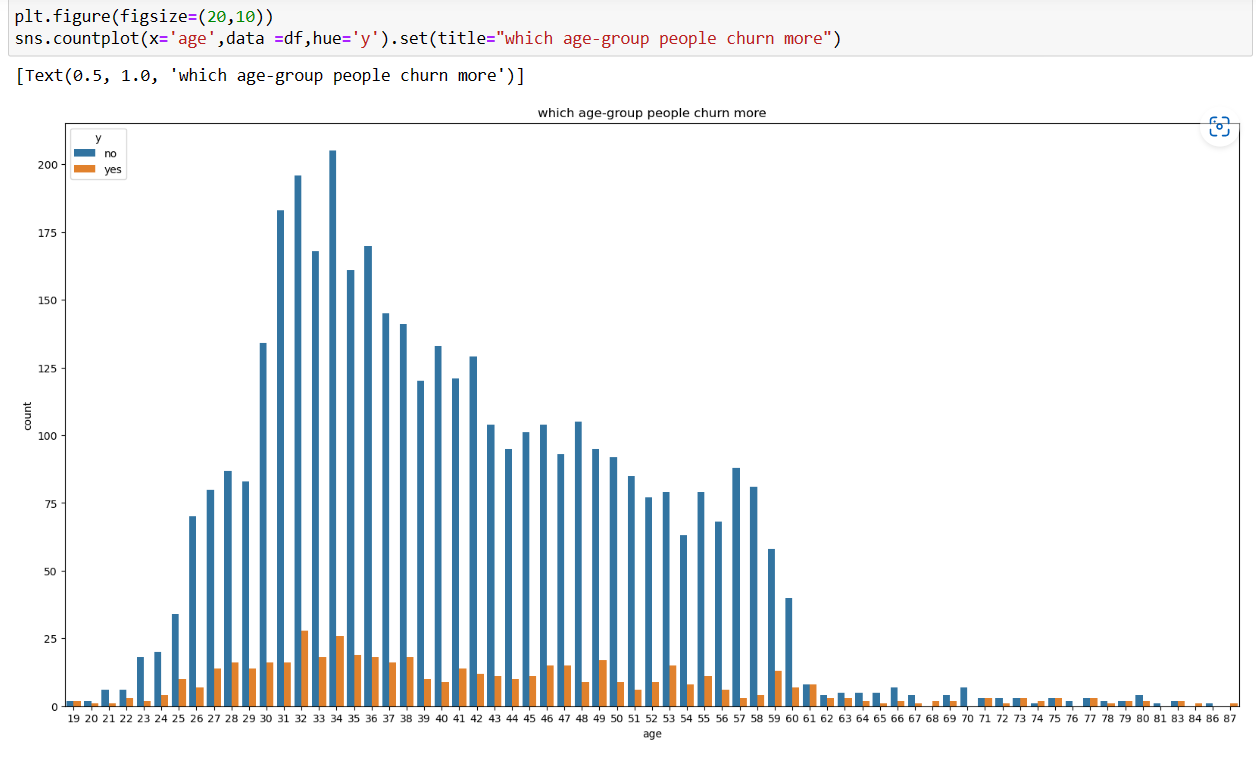
There are no **null value** present in the dataset.



Conclusion:

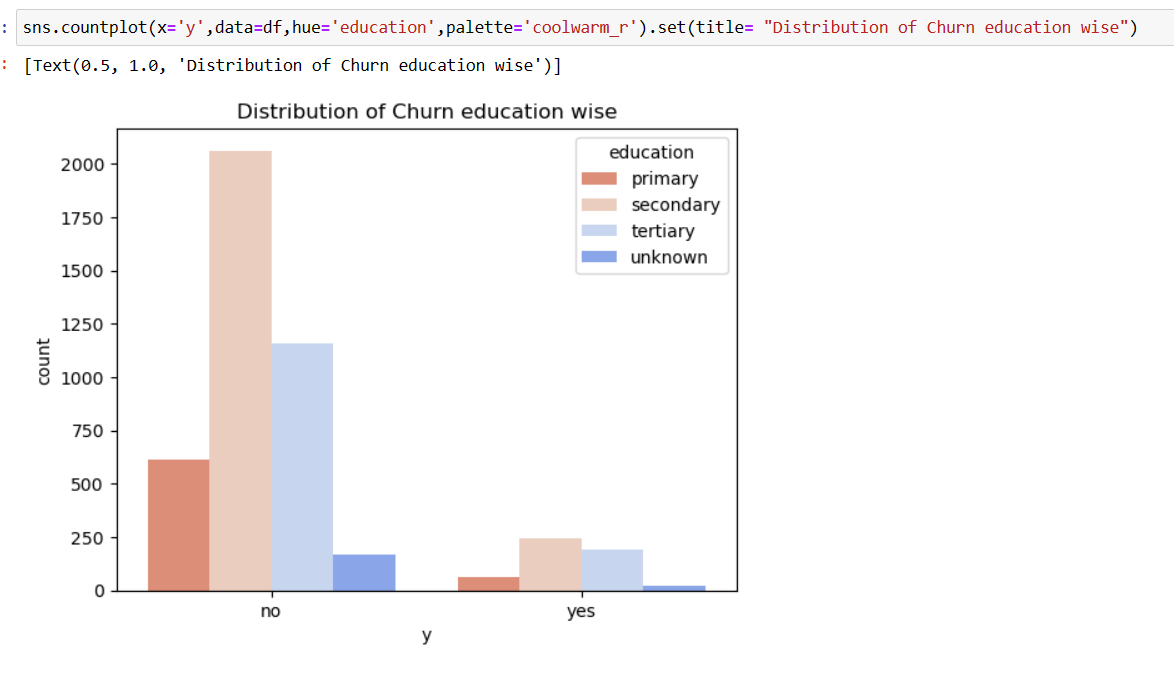
* Mean Age of the customer is 41
* 75% of the customer have balance 1480

Data Visualization:

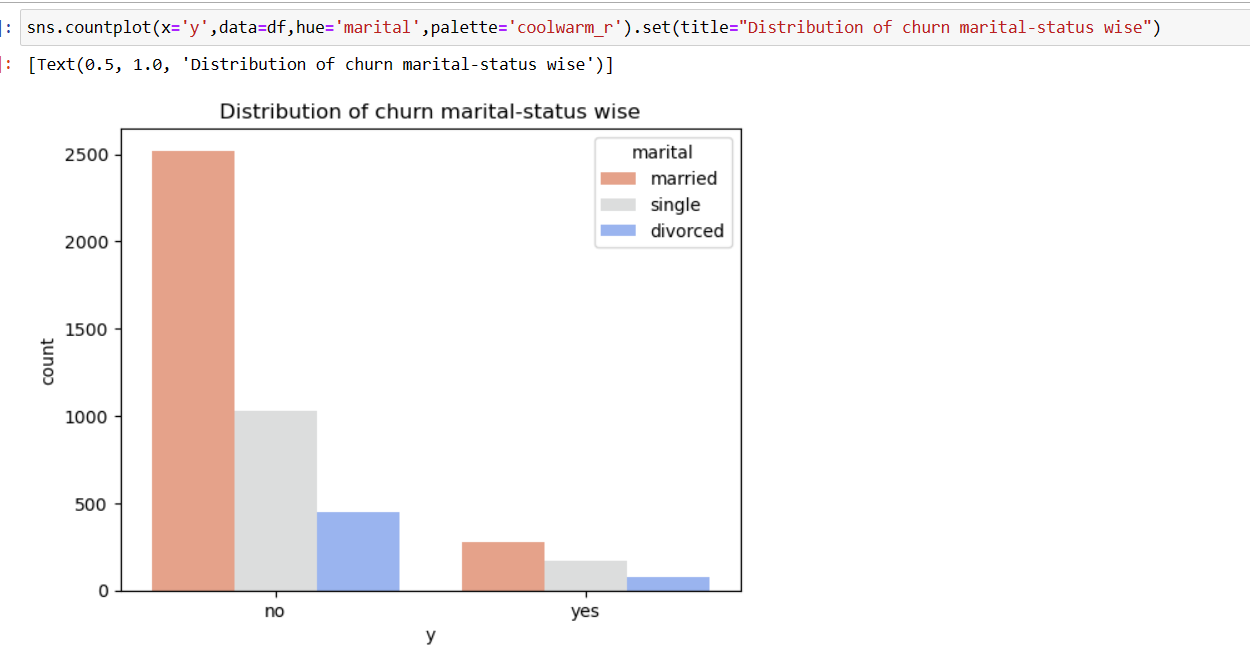


From the above graph, we clearly see that the younger- age group people churn more.

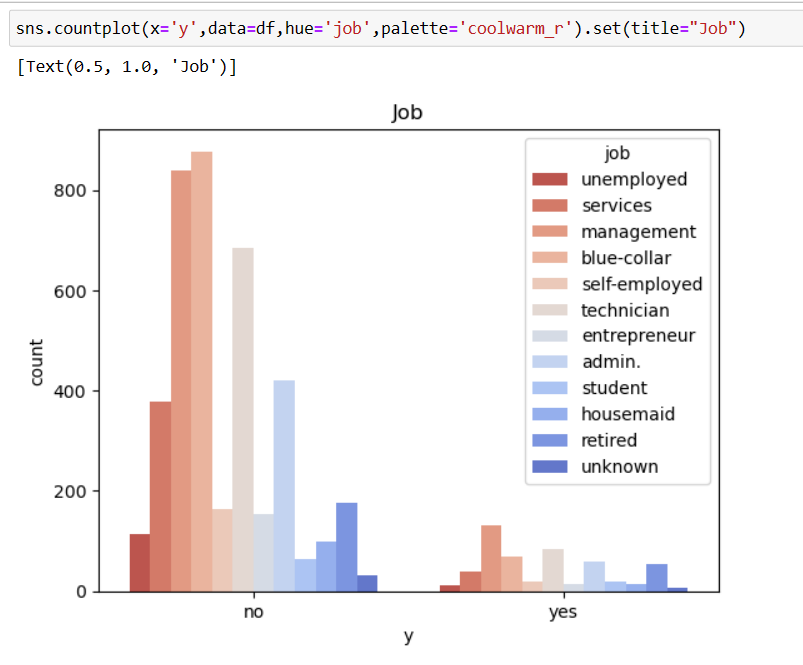
Distribution of Churn according to Variable:



People with Secondary and Tertiary education churn more in comparison with primary and unknown category.



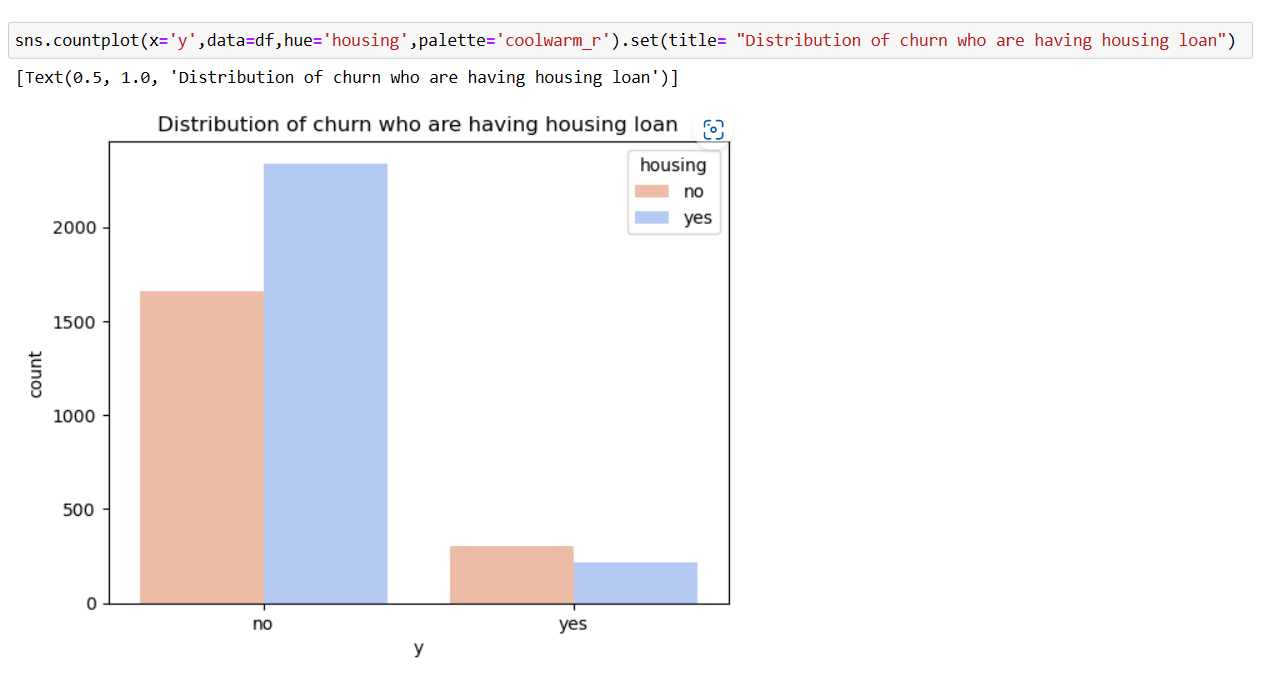
The graph is the evidence that Married people churn more.



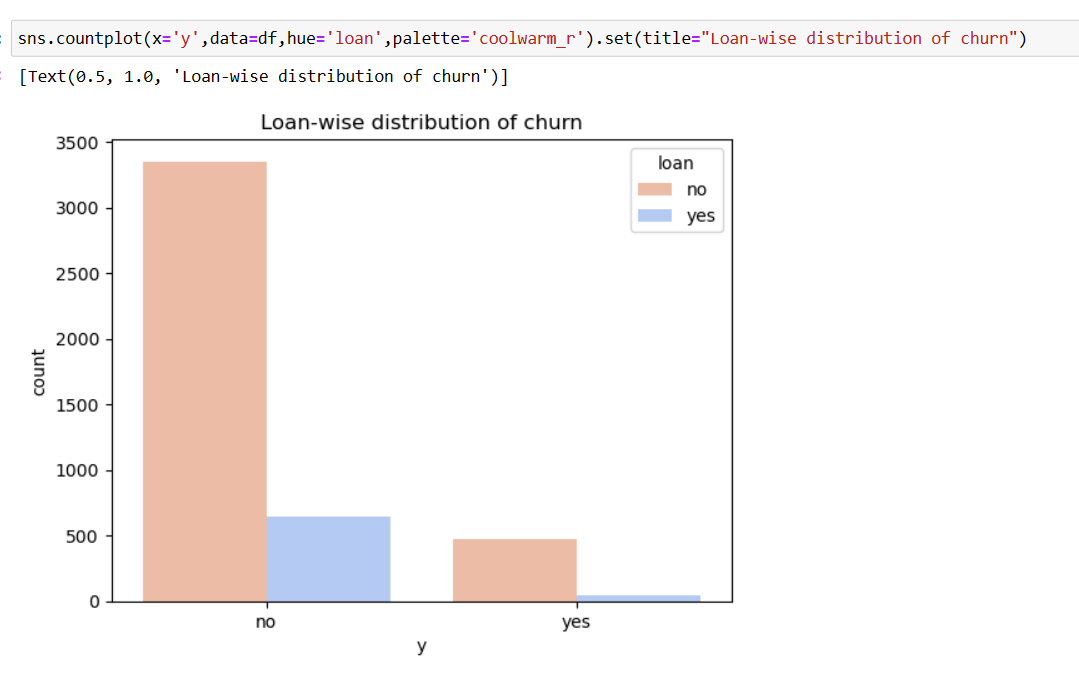
People who are having Management job churn more.

Churning rate of unemployed people is very less.

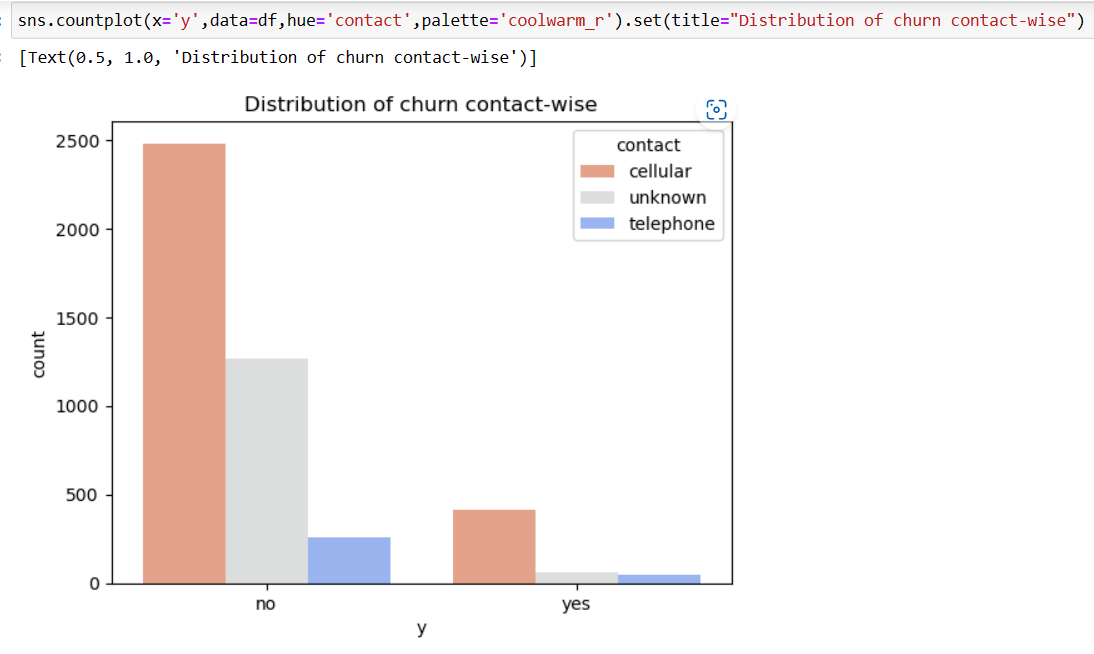
Retired, Admin, Technician have comparable same Churning rate.



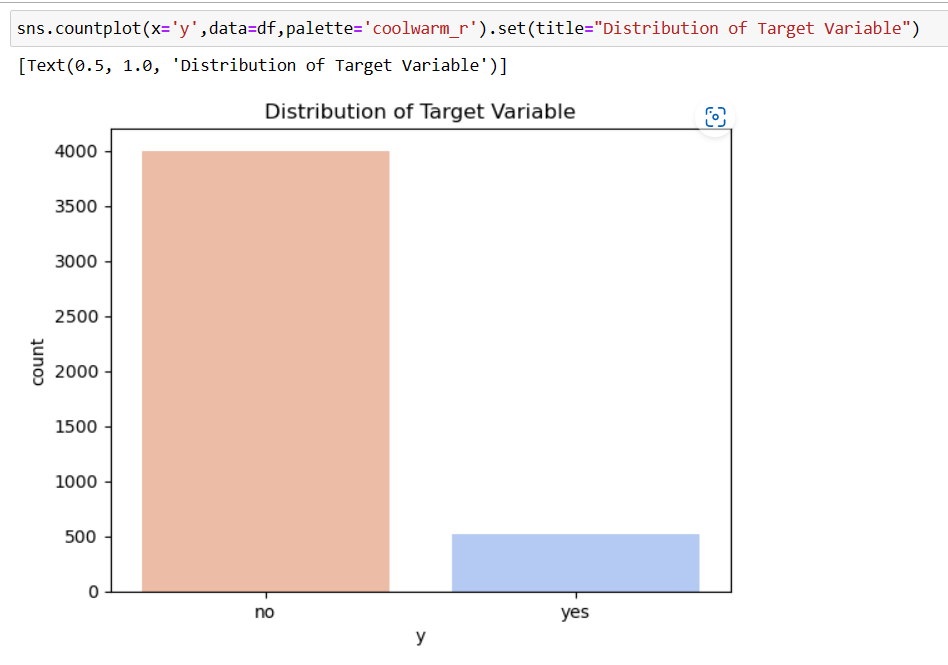
Churning rate of people who are having housing loan is almost same as people who are not having housing loan.



Churning rate is almost 0, for the people who are having loan.

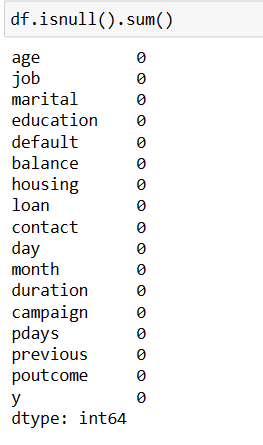
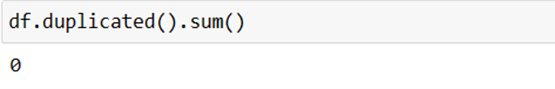


People who are using cell phone churn more.



Here we can see that Target variable is not evenly distributed which conclude that Target Variable is highly imbalanced.

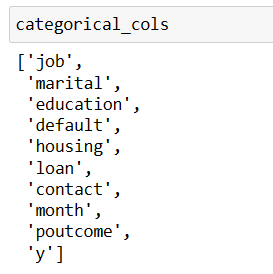
Data Cleaning and Pre-Processing:

As we can see here, our Data set is neither having any Missing Values nor any duplicate values.

Separating Categorical and Numerical Variables from the data:

1) Categorical Variables:



Here Categorical variables are of 3 types:

1. Ordinal:

Education, Month

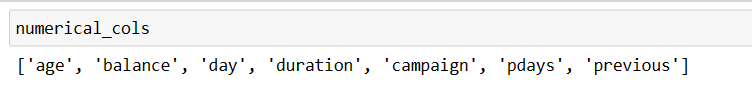
1. Non- Ordinal:

Job, Marital, contact, education

1. Binary:

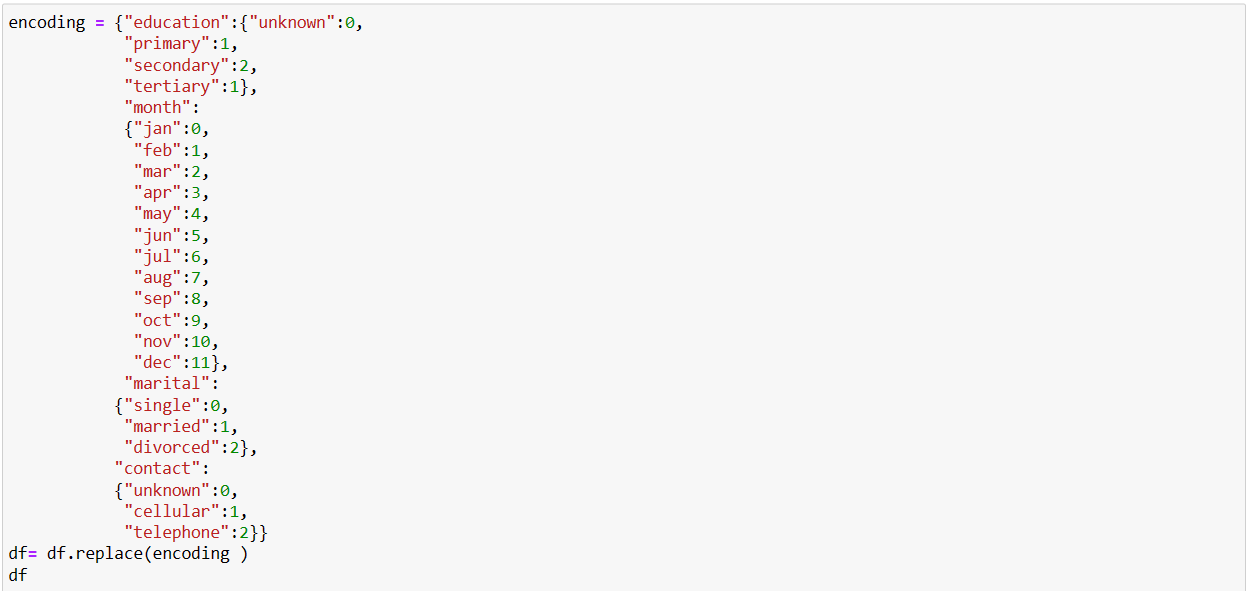
Y (Target variable), loan, default, poutcome

2) Numerical Variables:

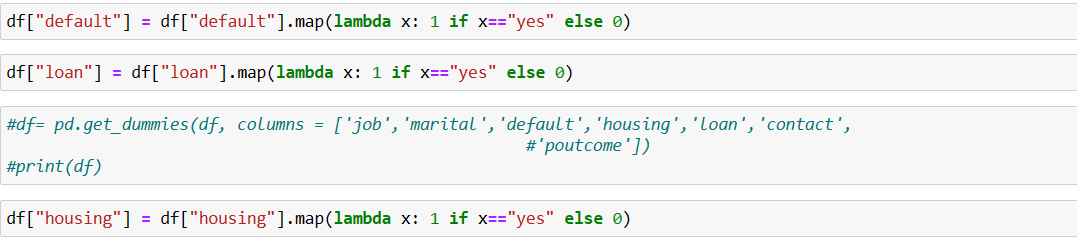


Encoding Techniques:  
Since our Categorical Variables are of 3 types,   
Different types of Encoding techniques will be applied.

Encoding for Education, Month, Marital, contact Variables:



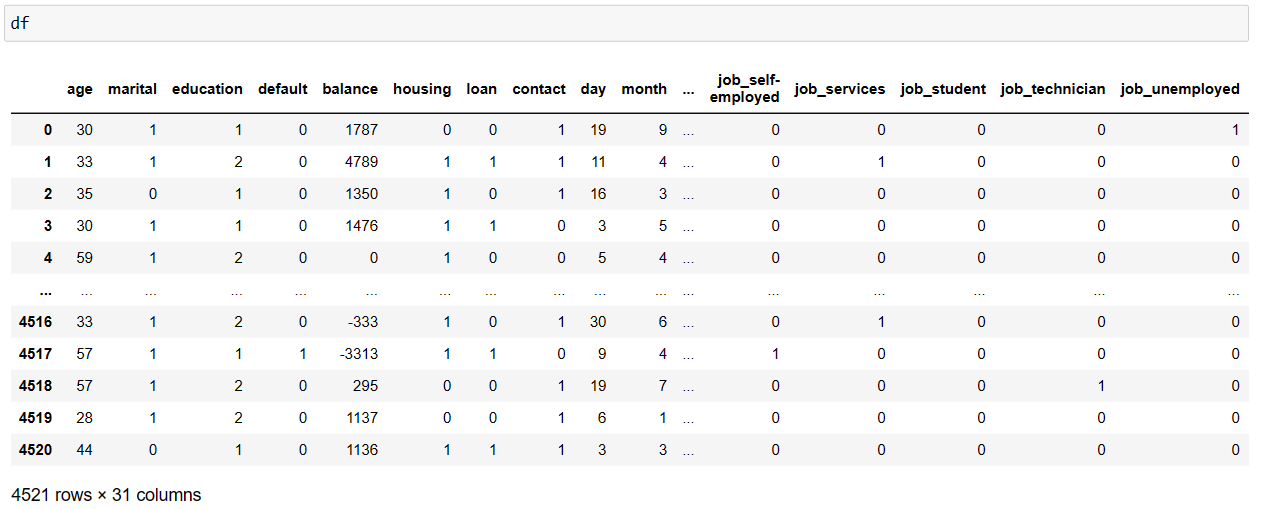
Binary Encoding for Default, housing, loan and Target Variable



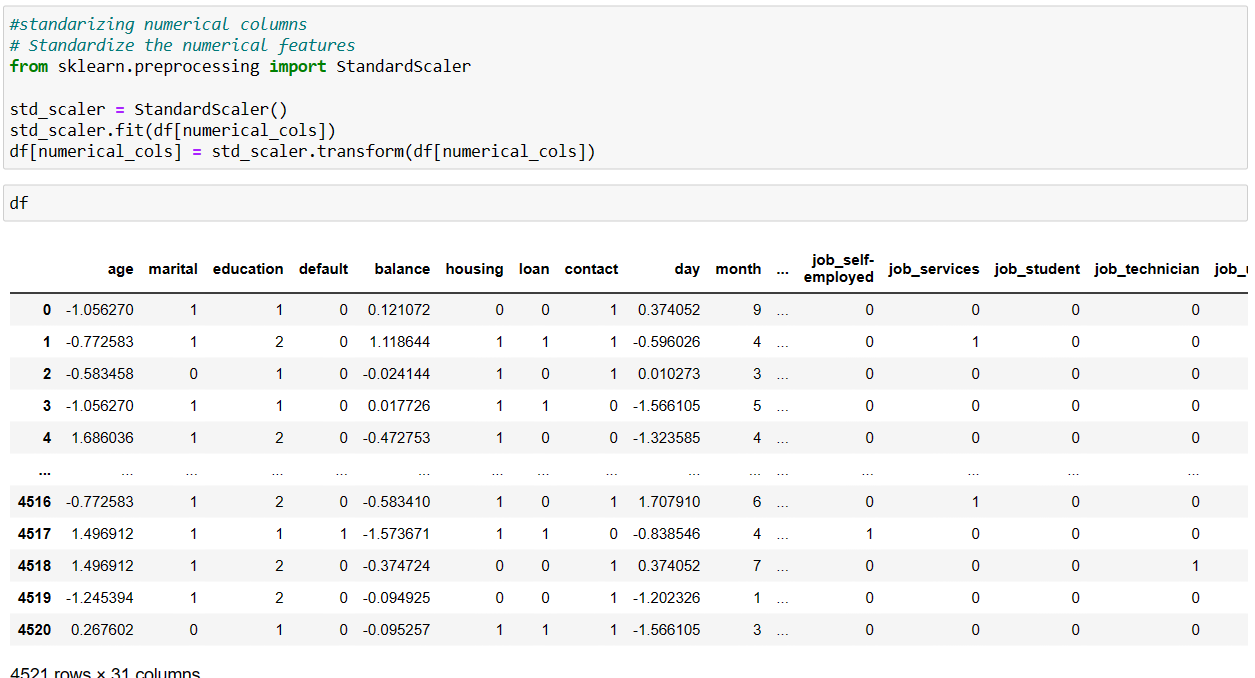
Dummy encoding for job and outcome variables:



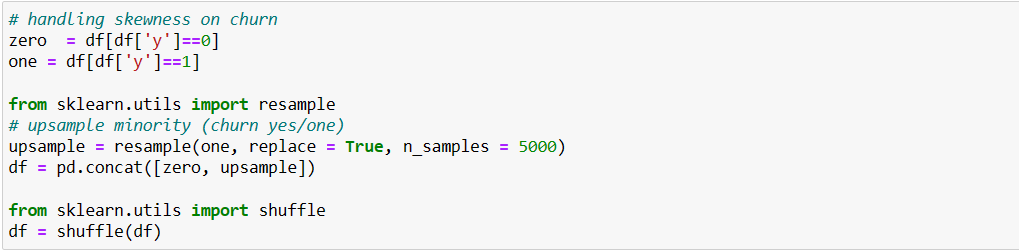
Final Dataset after encoding:



Since there are numeric variables, we are standardizing them:



Target variable was not balanced.   
So oversampling technique is used to handle its skewness

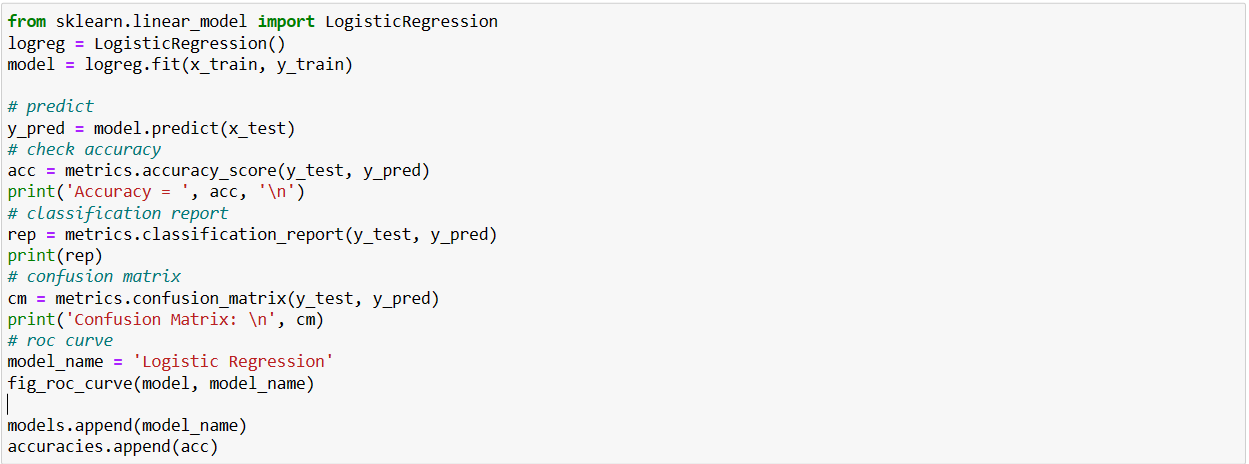


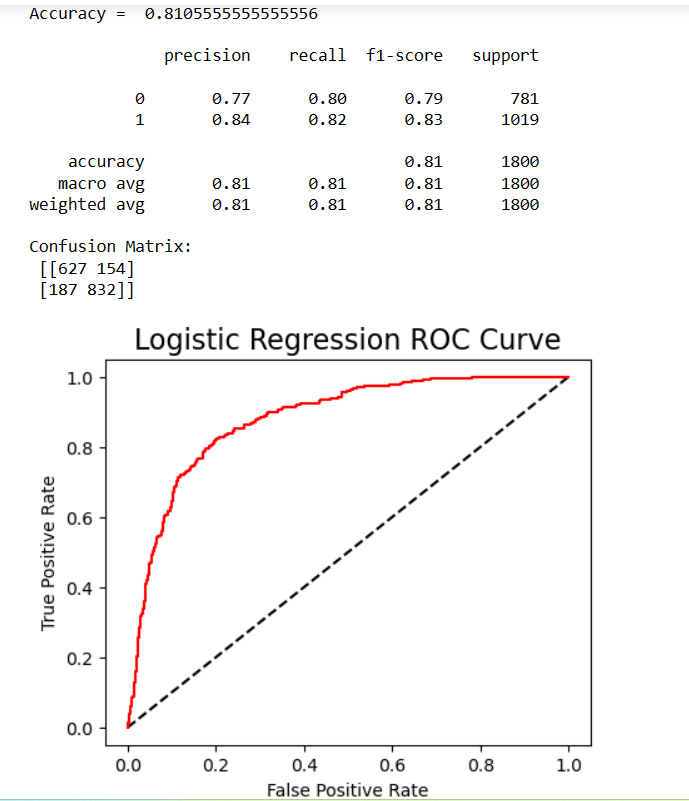
Distribution of Target Variable:



Model Deployment:

1) Logistic Regression:





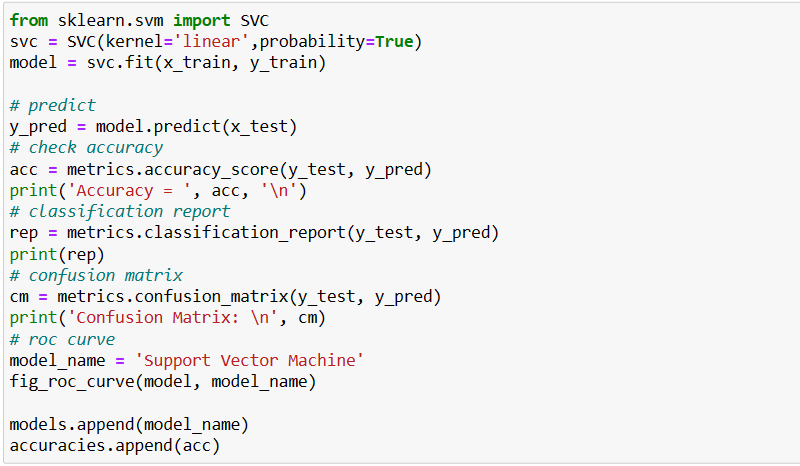
As we can see here, accuracy of Logistic regression is 81.05%

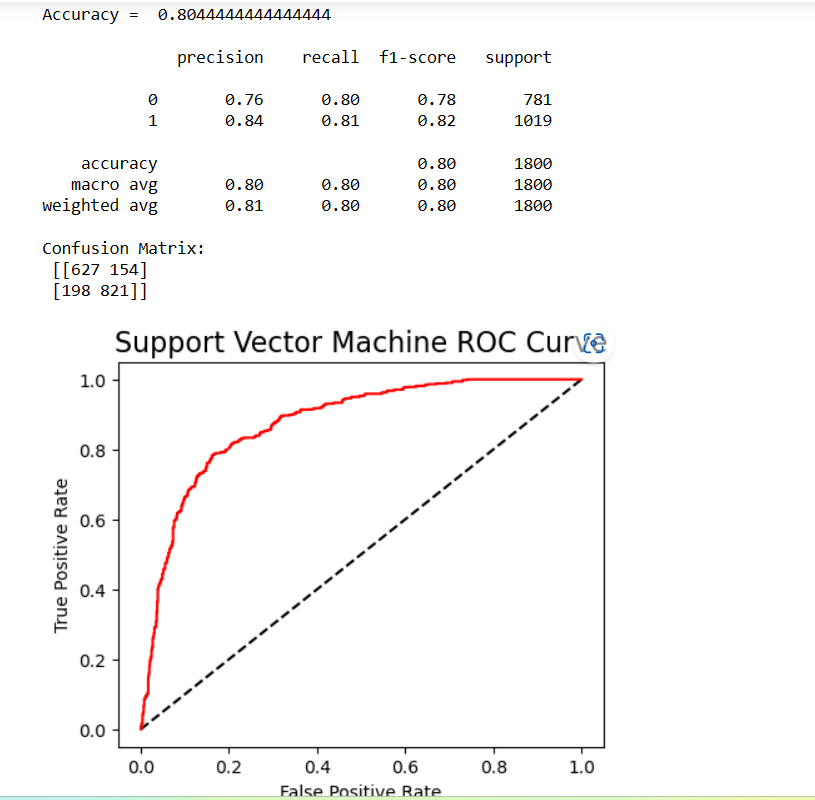
Precision value for Churn is 0.84

Which means False Positive are minimised

This predicts the model is Good Fit.

2) Support Vector Machine:





Accuray is 80.44%

Precision value for Churn is 0.84.   
Since our Goal is to find a model that minimised the False Positve which means to select the model with highest precision value.

Support Vector Machine can be chosen as a Final Model.

Limitations:

1) The above 2 models have been deployed without hyper parameter tuning

2) For Handling skewness of target variable, if different techniques such as under-sampling or SMOTE are applied then result may differ

3) Due to Time constraint only 2, Machine Learning Algorithm are explore.

There are other Machine learning algorithm such as Random Forest, XgBoost which can also be applied.

4) Deep Learning techniques can also be used

5) The results of the algorithms are only limited to this data set

6) For data set containing Million of Rows, different Machine Learning Algorithm may out-perform Logistic Regression and Support Vector Machine