DATA SCIENCE

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

Great Learning
POWER AHEAD

Capstone Project

Prediction of Customers Subscribing to Term Deposits in a Banking Institution

Prepared by:-

Anupam Dash

Ishwarya. B

Ranjith Kumar. A

Suganesh. R

Sundar Rajan Seshadri

Presentation date : 24-Dec'2022

Agenda

1 Problem Definition

- Problem Statement
- Business Understanding
- Explanation of concepts
- Why term deposit is important for bank
- 5W2H Method

2 Suggested Solutions & EDA

- Data Understanding
- Data Exploration
- Data Pre-processing
- Model Building
- Model Evaluation

3 Algorithms Solutions

Model Building

- Logistic Regression
- Decision Tree Classifier
- Random Forest classifier
- KNeighbors Classifier
- Gaussian NB
- AdaBoost Classifier
- Gradient Boosting Classifier
- XGB Classifier
- LGBM Classifier

4 Results & Conclusion

Model Evaluation

- F1 score
- Confusion Matrix
- Classification Report
- Model Conclusions
- Business Suggestion

Problem Definition

Portuguese banking institution

Problem Statements?

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaign was based on phone calls , often more than one contact to the same client was done, in order to access if the product (Bank Term Deposit) would be (“Yes”) or not (“No”) subscribed.

Business Understanding?

It’s all about understanding the overview, the aspects of business activities & the necessary problems faced in this business. To understand how to help the banks to identify its target customers and how to focus and change the method of approach for other customers to involve them into term subscriptions

5W2H – Bank Marketing Term Deposit	
What is the Problem?	Term Deposit Subscribe or not by customer
Where is this Problem?	Portuguese banking institution
When did the data collected?	14-02-2012
Why should this problem be solved?	Portuguese Bank will benefited if more Customers subscription for Term Deposit
Who is our target?	Customers
How to solve the Problem?	Build a Term deposit Subscription Engine
How much customer?	45211 customers

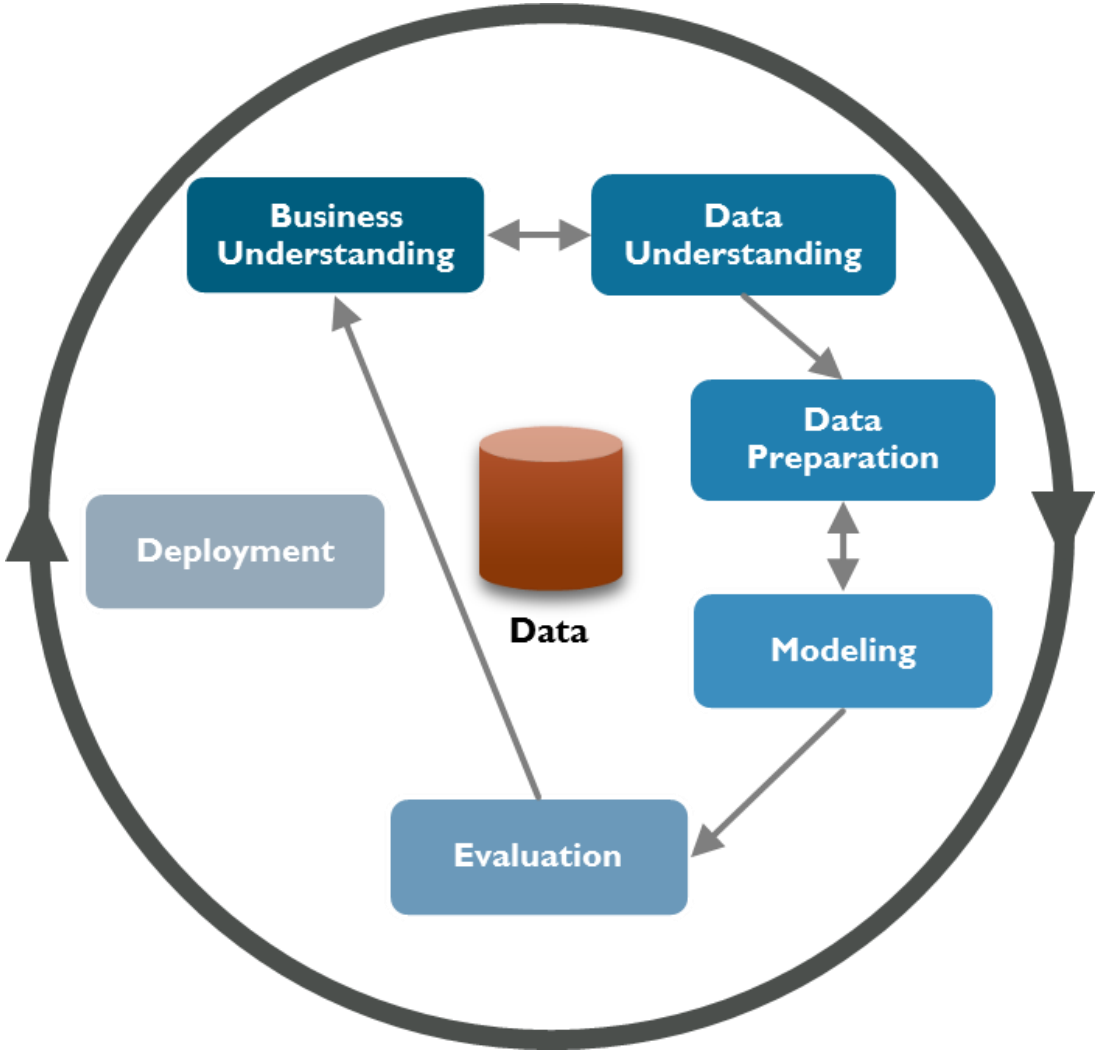
Explanation of Concepts?

Why term deposit is important for bank?

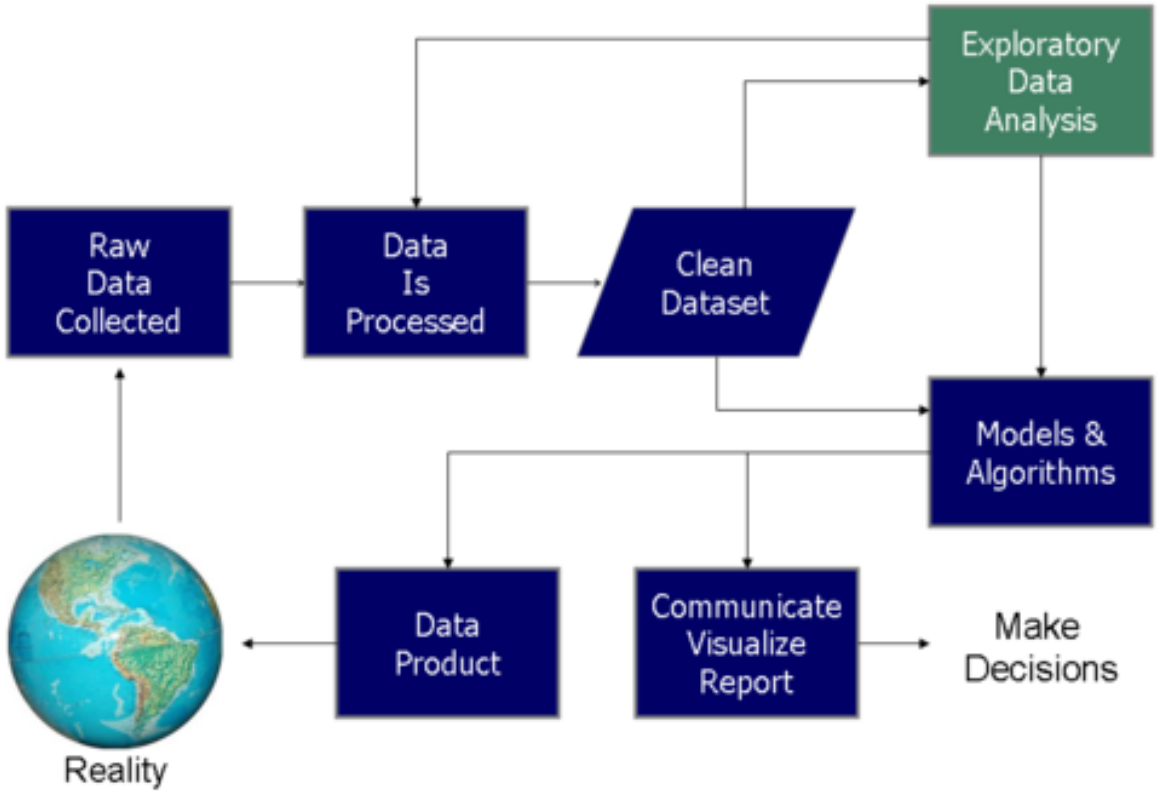
Term Deposits are **one of the best investment options for people who are looking for a stable and safe return on their investments**. In Term Deposits, the sum of money is kept for a fixed maturity and the depositor is not allowed to withdraw this sum till the end of the maturity period.

Problem Definition

CRISP - DM



Approach of Handle



Suggestion Solution & EDA

Raw Data

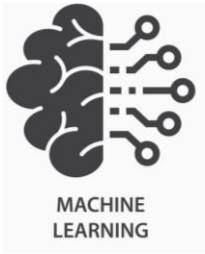
Raw data Understanding

Data Understanding

The main goal of the dataset understanding is to gain general insights about the data that will potentially be helpful for the further steps in the data analysis process

Data Exploration

Data exploration is the first step of data analysis, its used to explore and visualize the data to uncover insights from the start and find patterns to dig into more



Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses

Model Evaluation

Data Preprocessing

Data Preprocessing performed on raw data to prepare it for another data processing procedure below:
1.Data Scaling
2.Data Transformation
3.Data Encoding

Model Building

Machine learning analyzes the data that automates the analytical model building. It tells us that systems if trained to identify patterns, learn the data, and make decisions with little or no human intervention can learn.

A

Dashboard

Portuguese banking institution

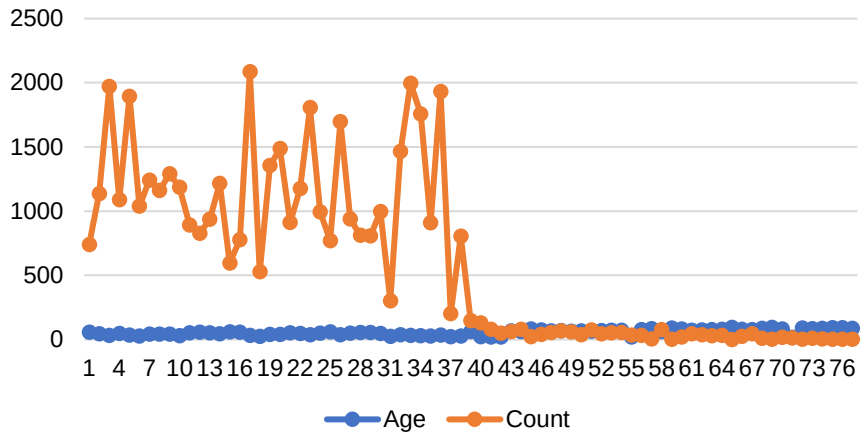
Raw Data

TOTAL CALL : 42511

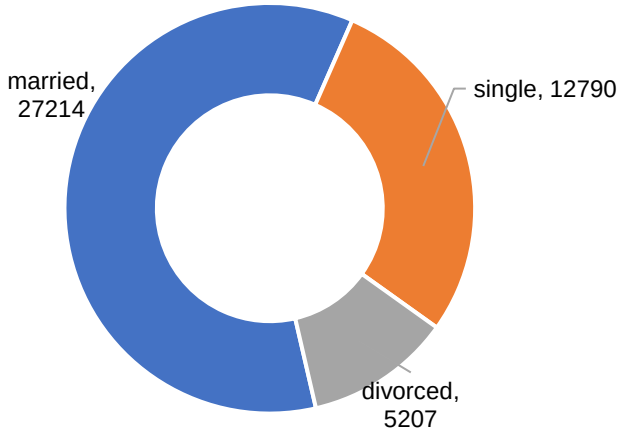
Success Rate

No	Yes
88.30%	11.70%

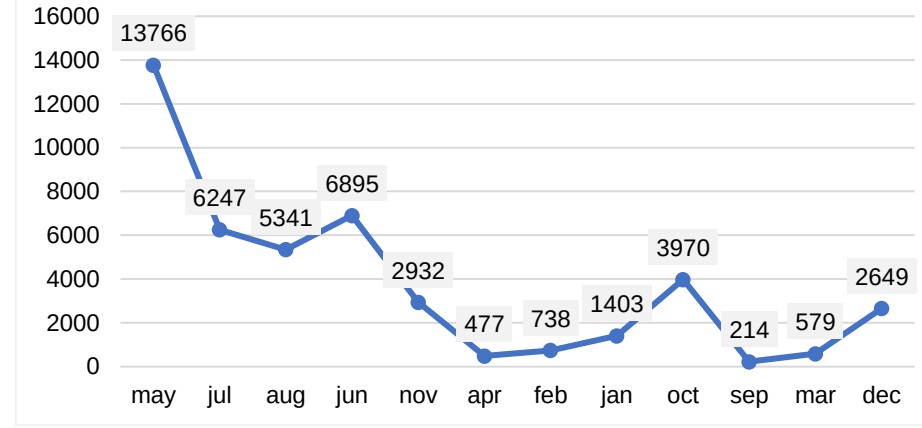
Age Wise



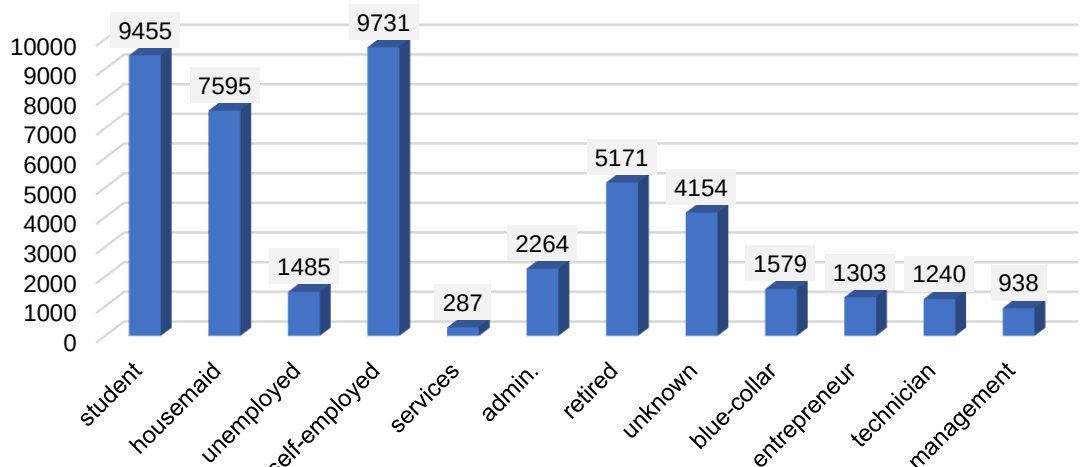
Marital Wise



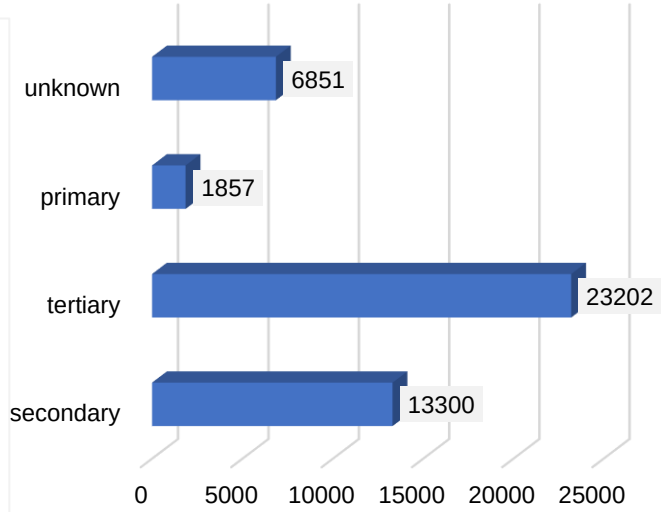
Month Wise



Job Wise



Education



Marital	Bank Balance	Duration call Hours
married	3.88 crore	22988
single	1.66 crore	11362
divorced	61.38 lakhs	4556

	Loan		
	Default	Housing	Personal loan
No	44396	20081	37967
Yes	815	25129	7244

B Data Understanding

- It is a classification problem.
- Data set consists of 17 columns including target variable with 45211 rows.
- It involves study of data, missing value, shape, skewness, datatypes, number of rows and columns, type of columns and univariate analysis – categorical & numerical

Numerical Features in the Dataset

Feature Name	Data Type	Feature Description
Age	Int	Age of the customer.
Balance	Int	The total balance of a person on a yearly basis is given.
Day	Int	The number of days that passed by after the client was last contacted from a previous campaign
Duration	Int	The last contact duration, in seconds(numeric).
Campaign	Int	The number of contacts performed during this campaign and for this client (numeric, includes last contact).
P-days	Int	The number of days that passed by after the client was last contacted from a previous campaign.
Previous	Int	The number of contacts performed before this campaign and for this client.

Categorical Features in the Dataset

Feature Name	Data Type	Feature Description
Job	Object	Describes types of job.
Marital	Object	It contains the marital status of the customer.
Education	Object	It contains the educational qualification of the customer.
Default	Object	It contains the default credit or not in customer's account.
Housing	Object	Contains information about whether the customer has housing loan or not.
Loan	Object	Contains information about whether the customer has personal loan or not.
Contact	Object	It gives the contact communication type.
Month	Object	This feature contains information about the last contact month of year.
P-Outcome	Object	It gives information about the outcome of the previous marketing campaign.
Y	Object	Gives information on whether the client has subscribed to term deposit or not.

B Data Understanding

Great Learning
POWER AHEAD

DATA DIMENSION :

```
print('In this dataset, No.of.rows are :',df.shape[0])
print('In this dataset, No.of.columns are :',df.shape[1])
```

```
In this dataset, No.of.rows are : 45211
In this dataset, No.of.columns are : 17
```

MISSING VALUE ANALYSIS :

```
df.isnull().sum()
```

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
dtype: int64
```

- we dont have any missing values in the dataset

Checking for Duplicates

```
df.duplicated().sum() # there is no duplicates in the data
```

```
0
```

CHECKING FOR SKEWNESS

```
df.skew()
```

```
age          0.684818
balance      8.360308
day          0.093079
duration     3.144318
campaign     4.898650
pdays        2.615715
previous    41.846454
dtype: float64
```

- The balance,duration,campaign,pdays,previous columns are highly skewed

COUNT OF OUTLIERS

```
In [30]:
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).sum()
```

```
Out[30]:
age          487
balance     4729
campaign    3064
contact        0
day            0
default        0
duration    3235
education      0
housing        0
job            0
loan           0
marital        0
month          0
pdays       8257
poutcome       0
previous    8257
y              0
dtype: int64
```

INFO FOR ALL FEATURES :

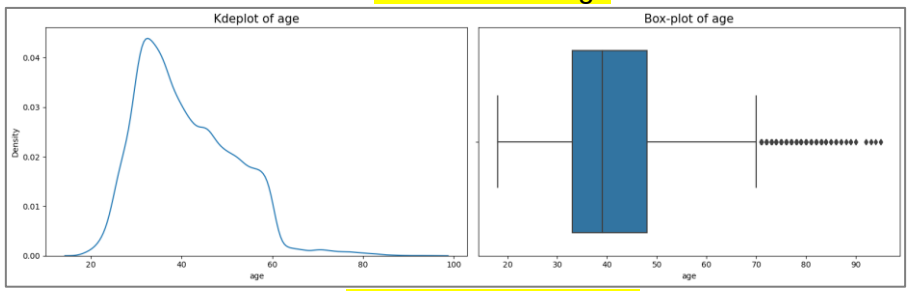
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

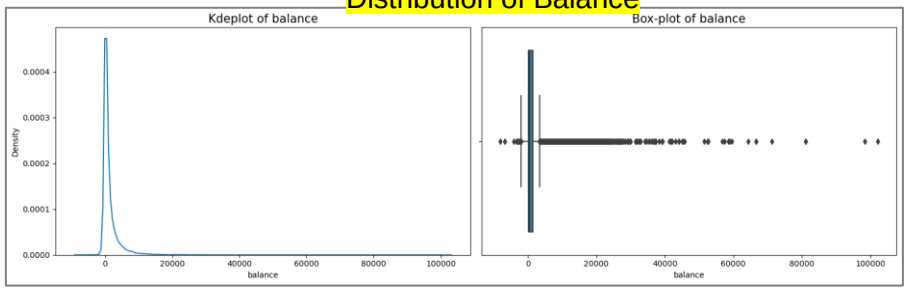
B Data Understanding

Univariate analysis – Numerical

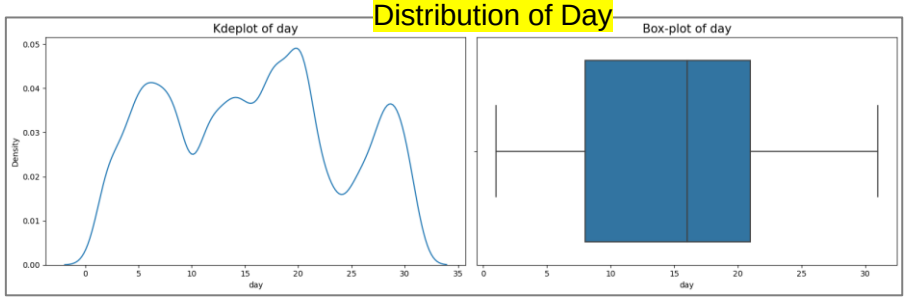
Distribution of age



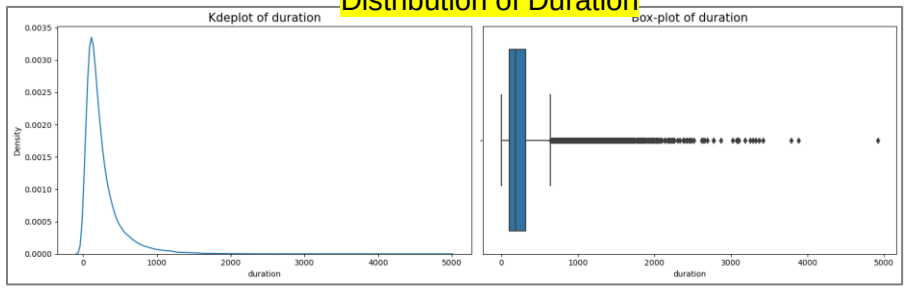
Distribution of Balance



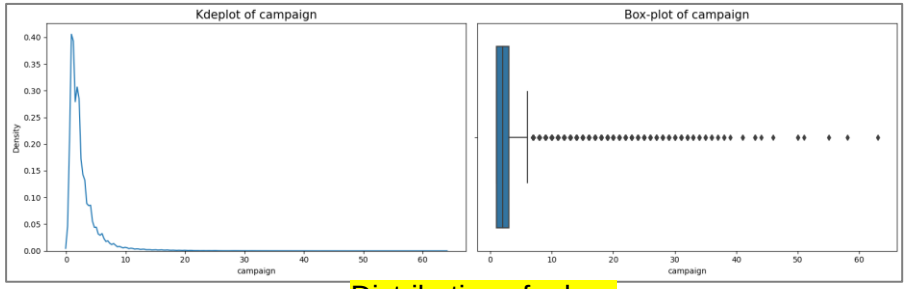
Distribution of Day



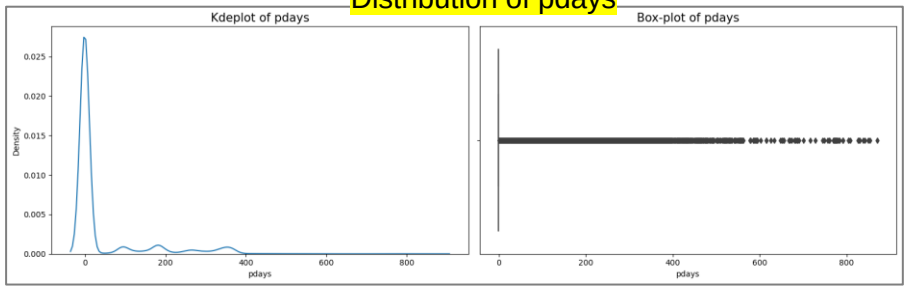
Distribution of Duration



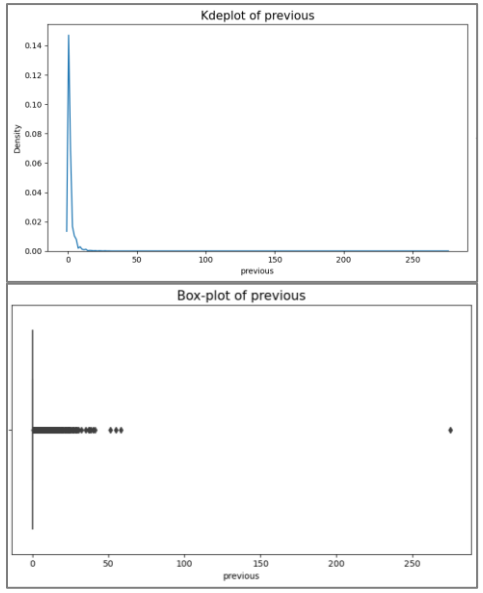
Distribution of Campaign



Distribution of pdays



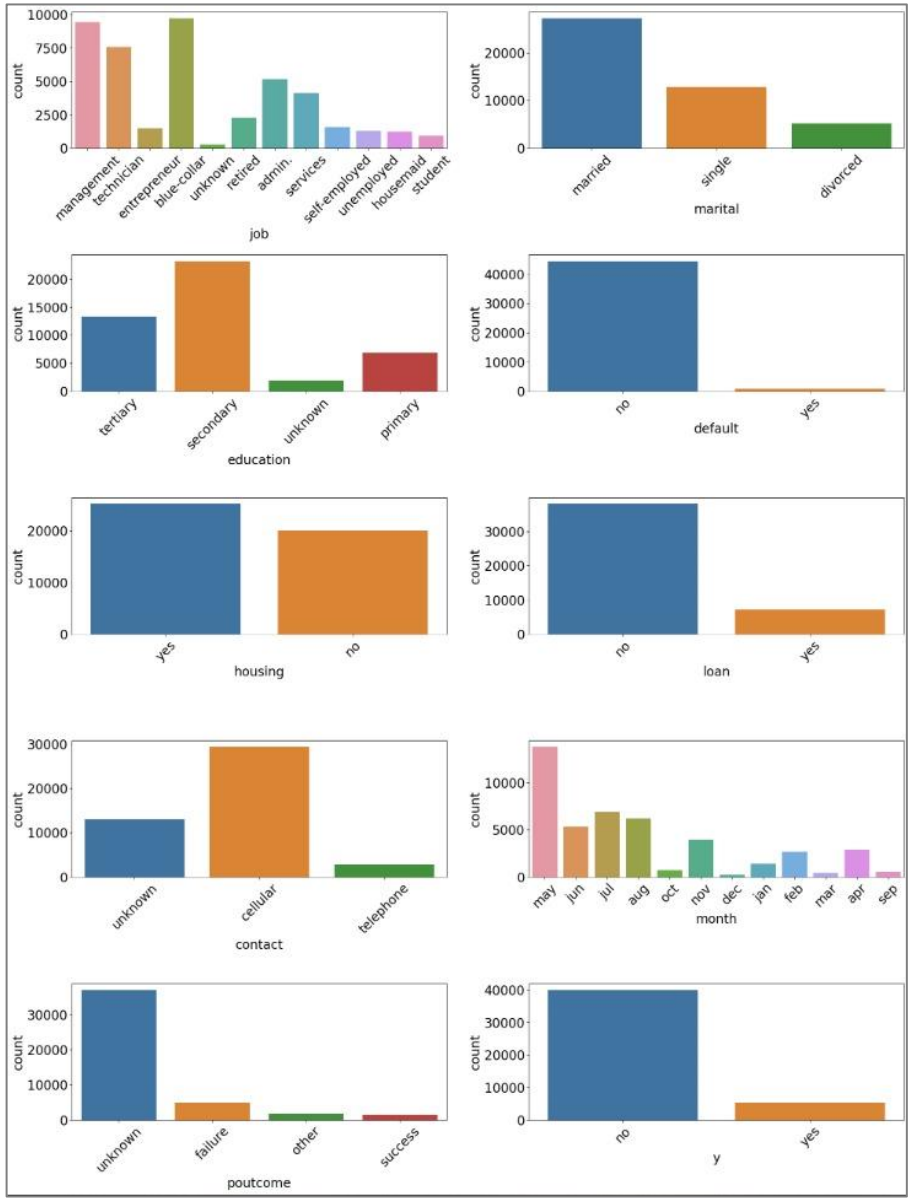
Distribution of Previous



- From the above plot we can identify that the "Balance, Day, Duration, Campaign, Pdays, Previous" is positive and right skewed.
- From the above plot we can say that there are outliers present in the data
- The balance, duration have the high outliers
- Age – 75 % of the people are the age between 20 to 50
- Balance – Average balance in most of the people rs.3000 only
- Day – All the days in the month are equally distributed
- Duration – Average call duration between 400 seconds
- Campaign – Average of 3 calls contacted per person
- Pdays – Most of the people are not contacted
- Previous – Most of the people are not contacted

B Data Understanding

Univariate analysis – Categorical

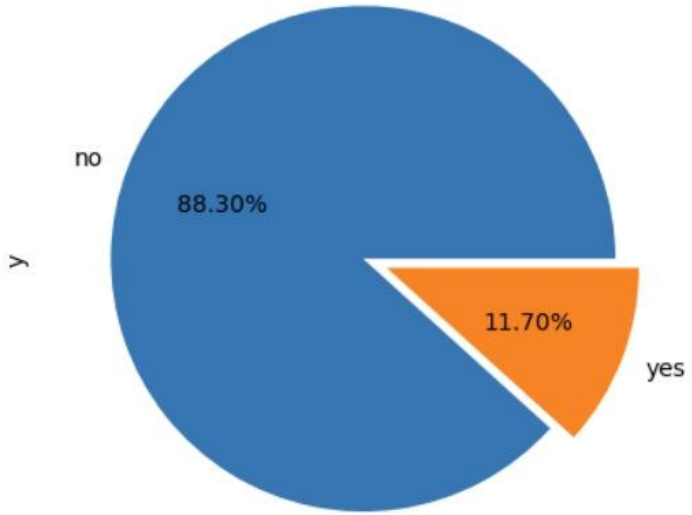


Inference:-

- job - Blue collar and management has high number of working when compared to other jobs
- Marital - Married are high than single and divorced
- Education - Most people are secondary educated
- default - most of the people do not has credit default
- Housing - People partially have less housing loans
- loan -Most of people do not have loans
- contact - Bank has contacted on cellular than other modes
- Month - May month has high contacted no of peoples
- Poutcome - Previous outcome most of the results are unknown
- y - most of the people said no for term deposits

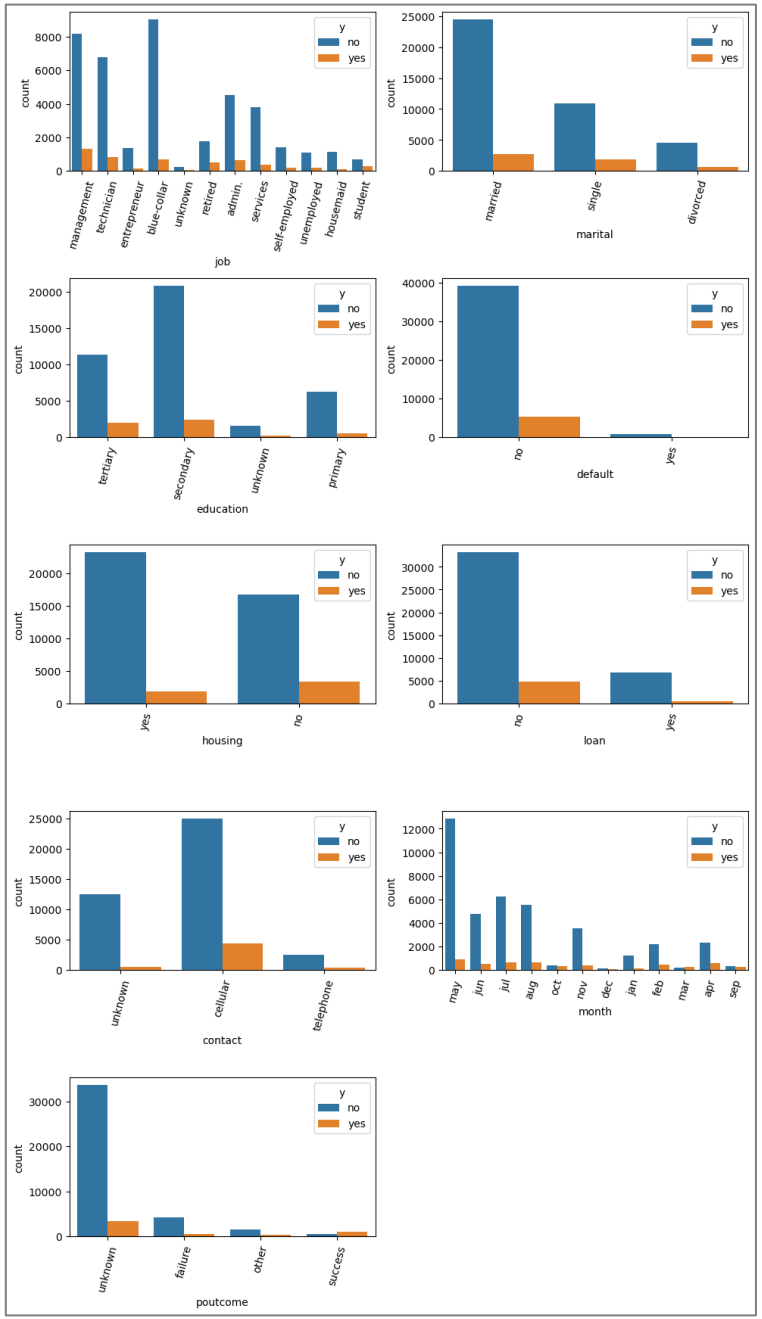
C Data Exploration

Target Variable - Y



From the total of 45211 customers 39922 have said “**no**” , 5289 said “**yes**”

Categorical with respect to target

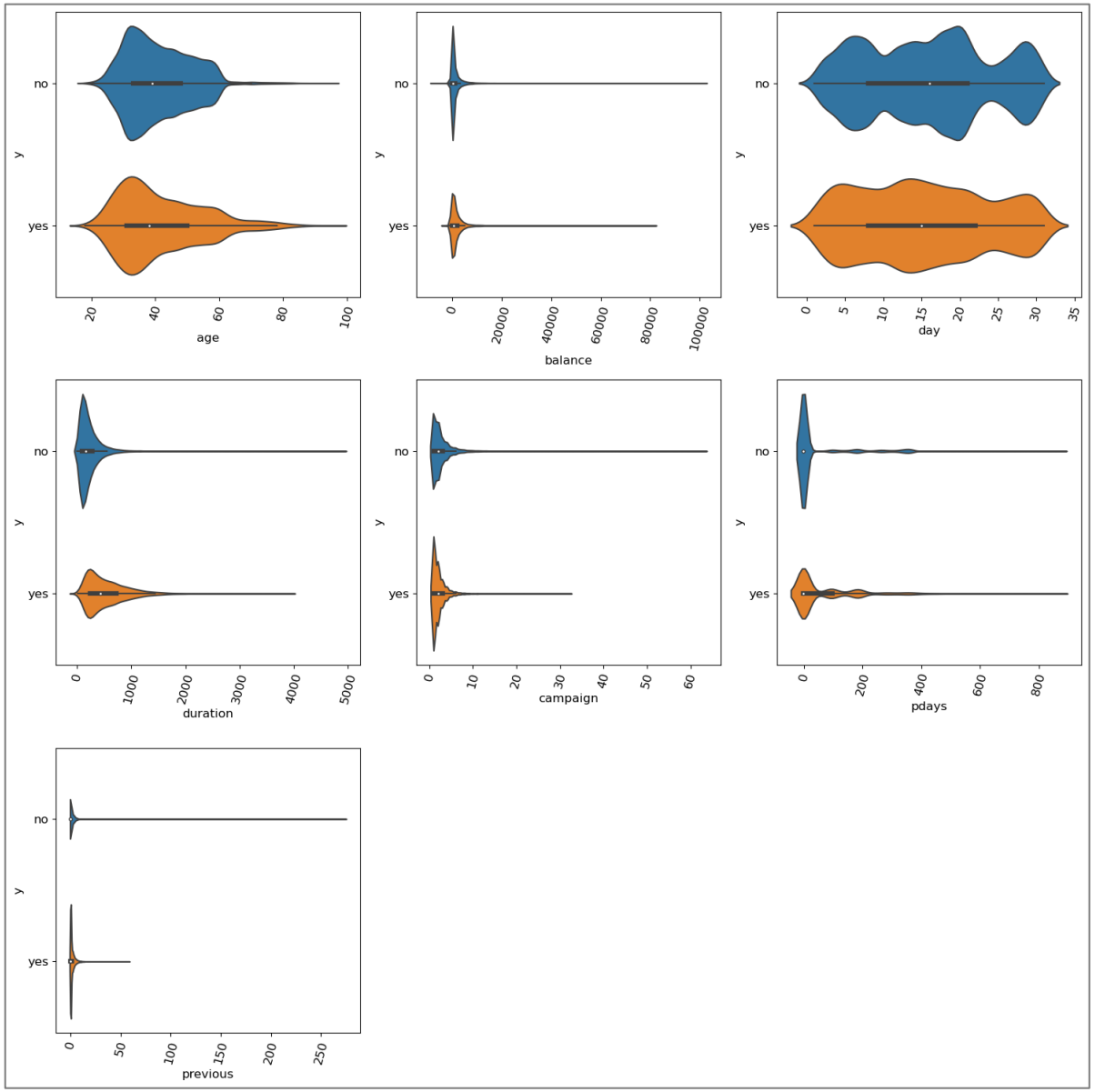


Inference:-

- job - management , tech, Blue-collar have highly subscribed to term deposits
- Marital – Married and single have subscribed more
- In the education category both secondary and tertiary sub-categories have subscribed more in numbers than to those working in management & blue-collar categories.
- In default category most of the customers have not subscribed.
- Housing – People with no housing loan are subscribing more
- loan - People with no loan are subscribing more
- contact - People having cellular phone are subscribing more

C Data Exploration

Numerical with respect to target

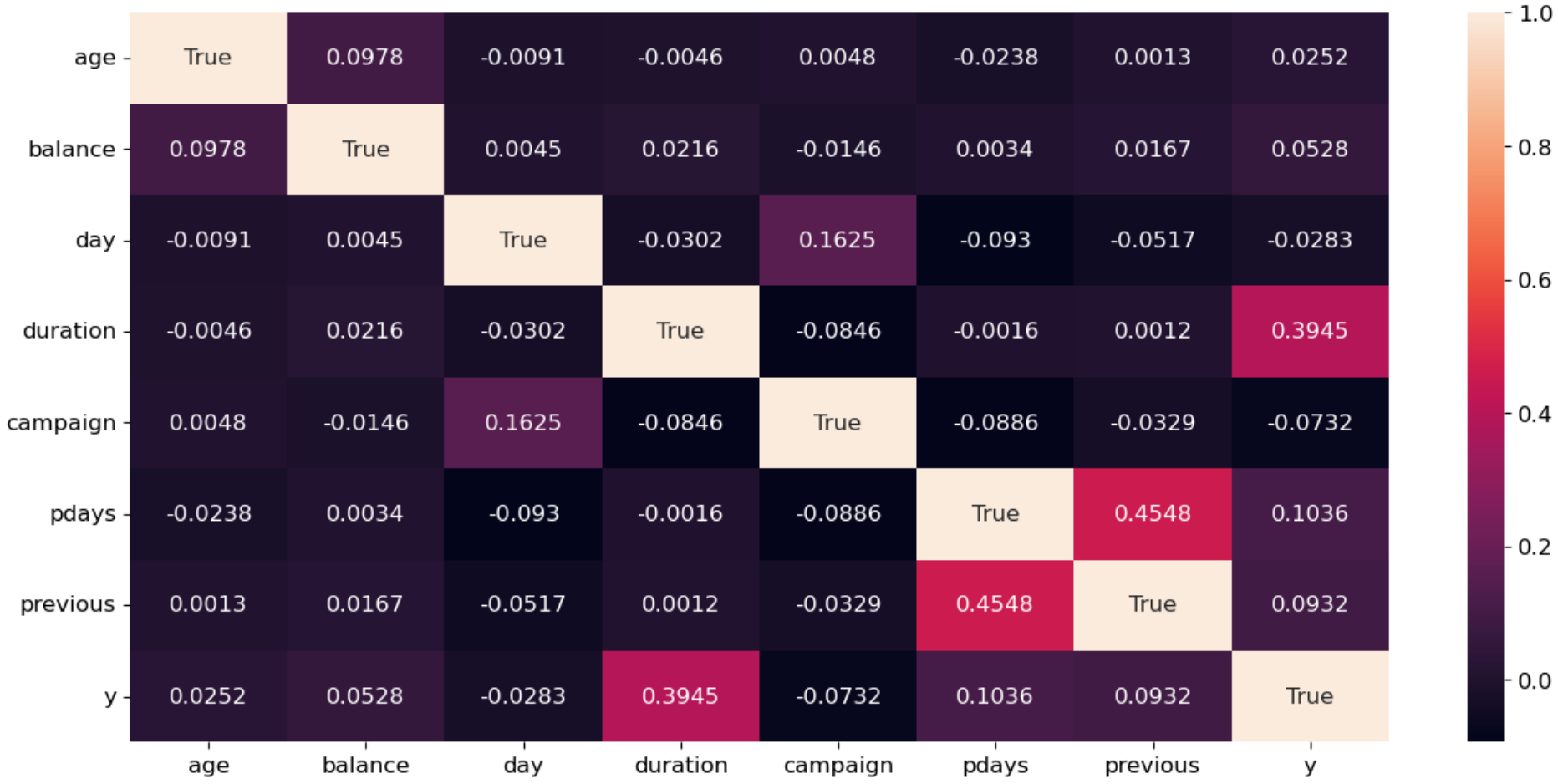


Inference:-

- Average Age between 25 to 40 has more people with yes and no
- Average Balance between 0 to 3000 with yes and no
- The day between 1 to 31 are equally distributed for both ‘yes’ and ‘no’ responses.
- The average duration of calls of customers between 0 to 1000 have said ‘yes’ to the subscription and average duration of calls of customers between 0 to 600 have said ‘no’ for the subscription.
- The average marketing campaign (telephonic marketing) for a person is around 3 calls.
- Most of the people not contacted in previous campaign, only some have been contacted.
- Most of the people have not contacted in previous campaign

C Data Exploration

Correlation Heat map

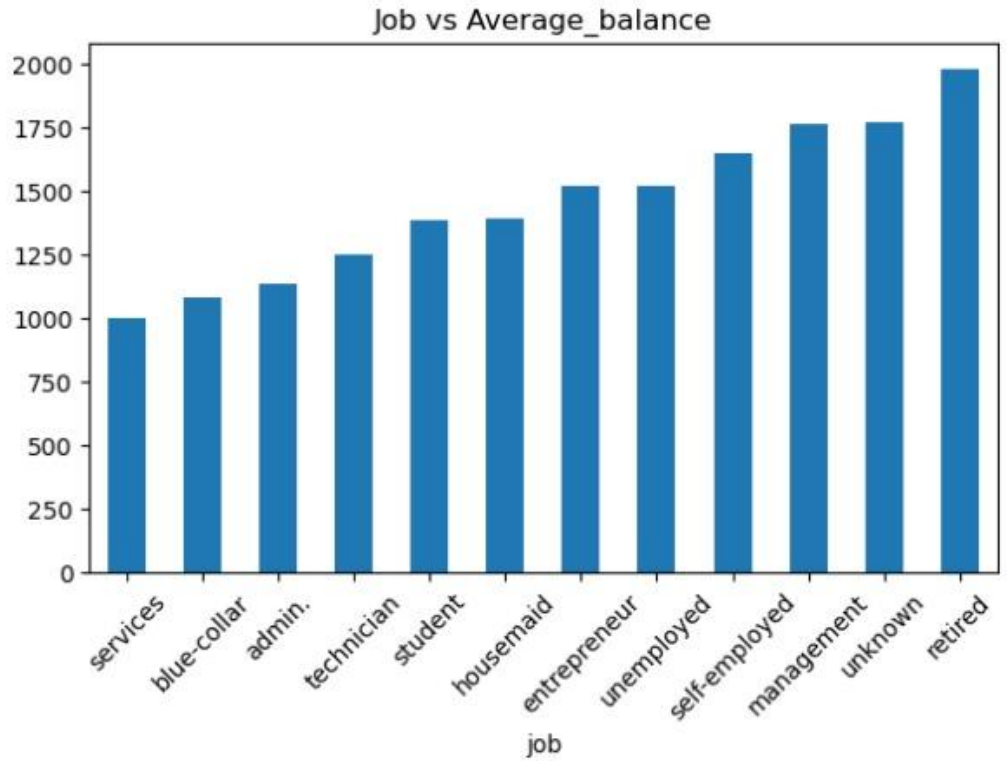


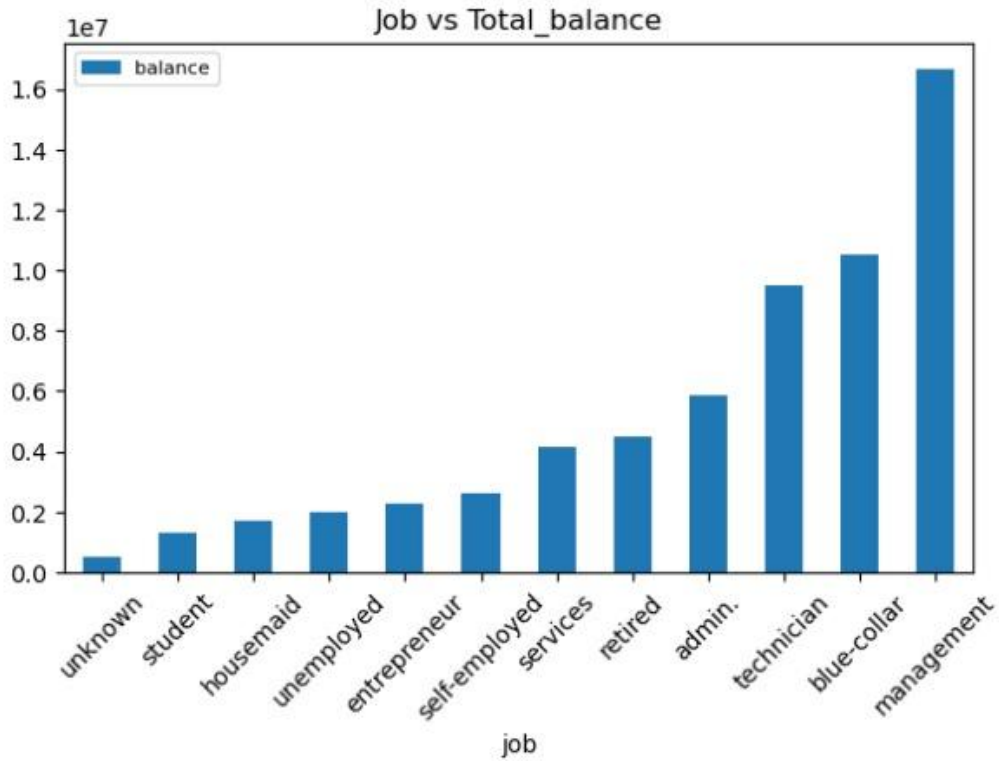
Inference:-

- No feature is highly correlated to each other
- so all the features are equally important to predict the target
- From the above we can say that pdays,previous,duration has some correlation

C Data Exploration

Job & Balance w.r.t to Target



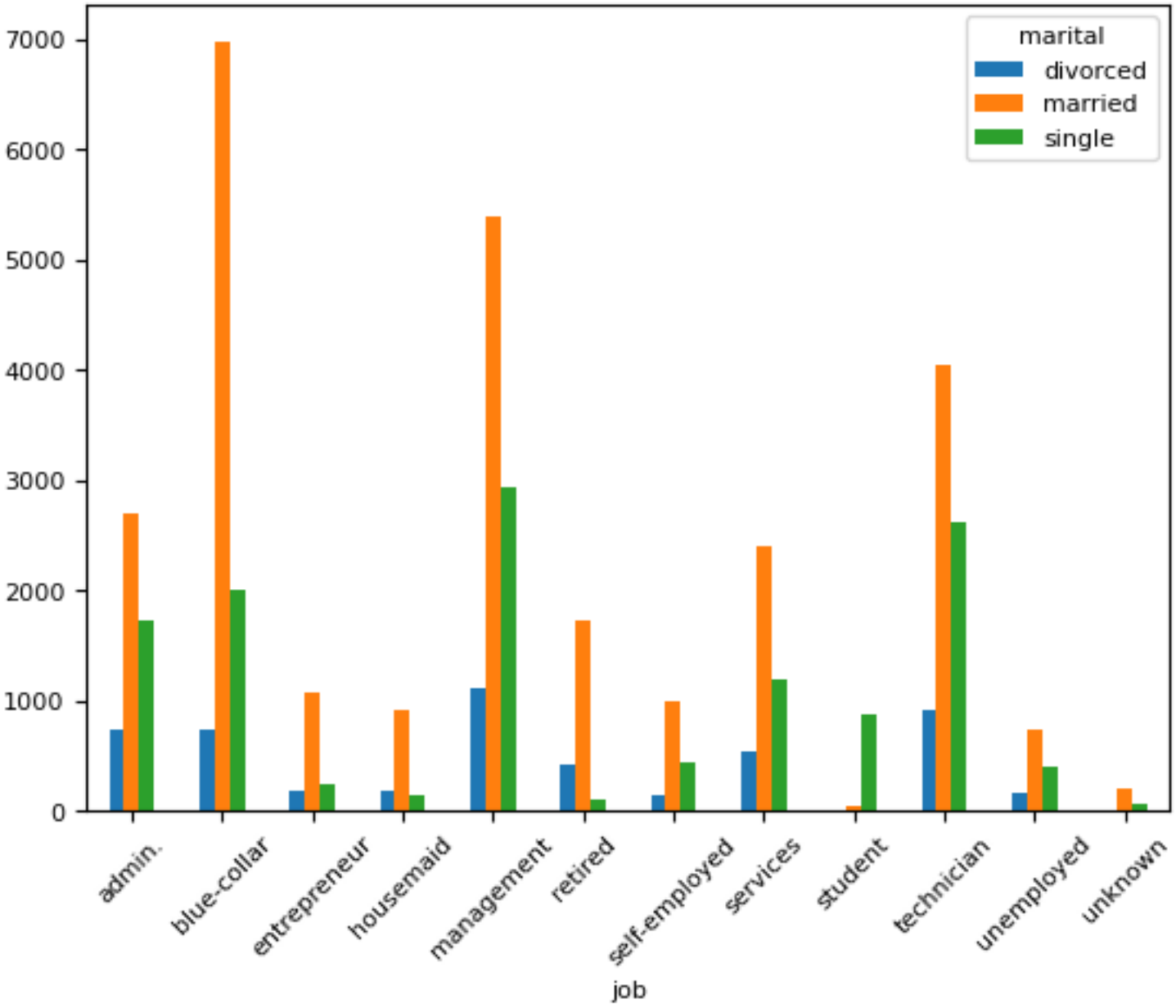


Inference:-

- From the above graph we can say that self employed, retired, management, blue-collar persons has good balance and said yes for term deposits
- People in retired, unknown, management, self-employed has high average balance
- People in management, blue-collar, technician has high total balance
- We should focus on people working in this job categories for subscription

C Data Exploration

Job & Marital

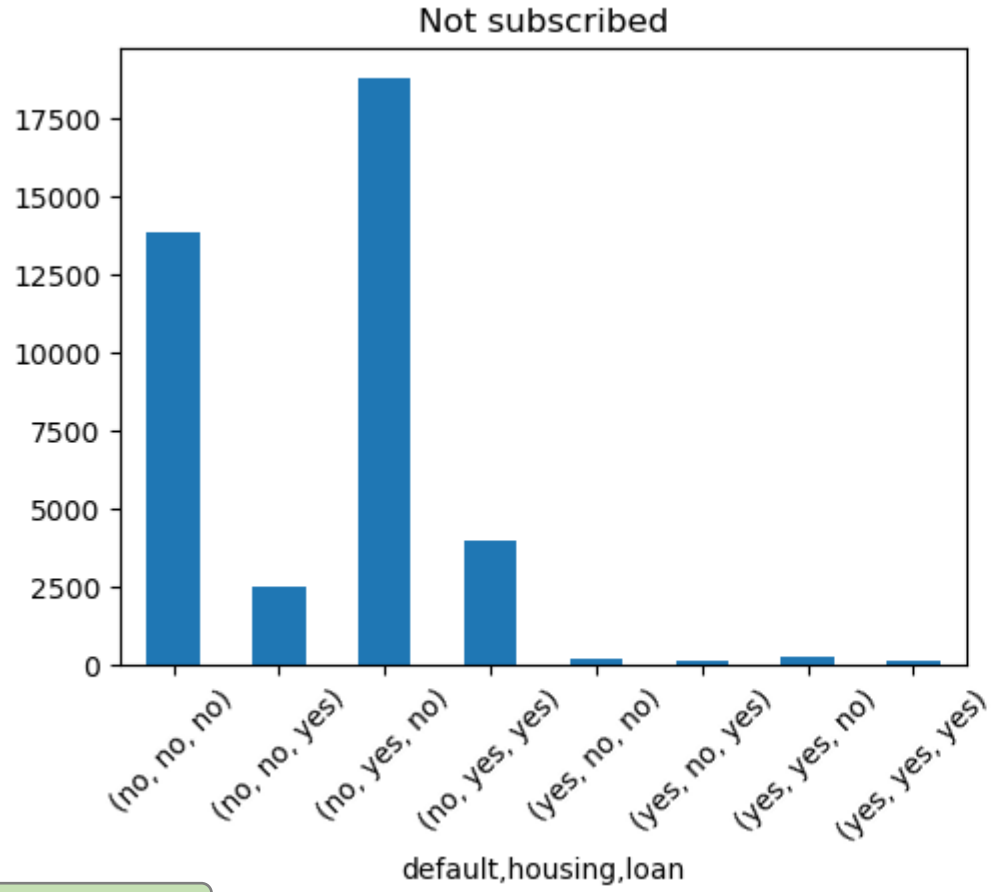


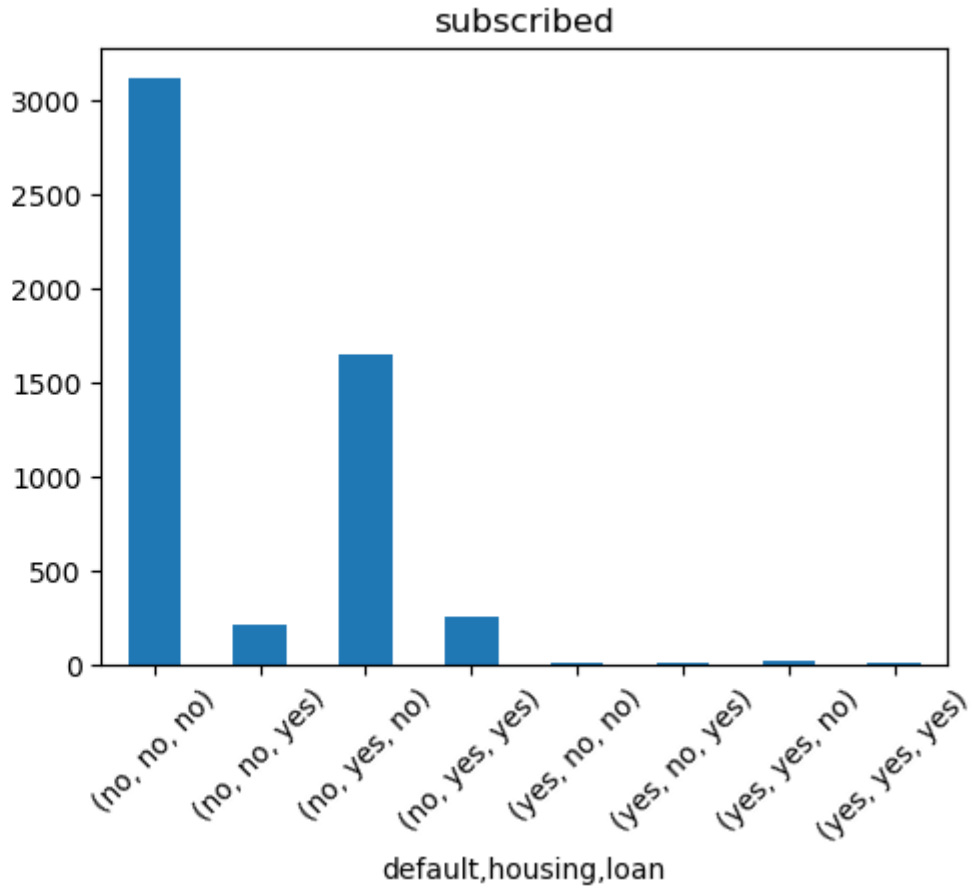
Inference:-

- The married and single people are high in number
- The married and single are highly working in the blue-collar, management, technician, admin jobs
- This answer to the highest bank balance in these kind of jobs
- So if a person is married or single working in blue-collar, management, technician, admin jobs these 4 jobs they will subscribe

C Data Exploration

Loan w.r.t Not subscribed and subscribed



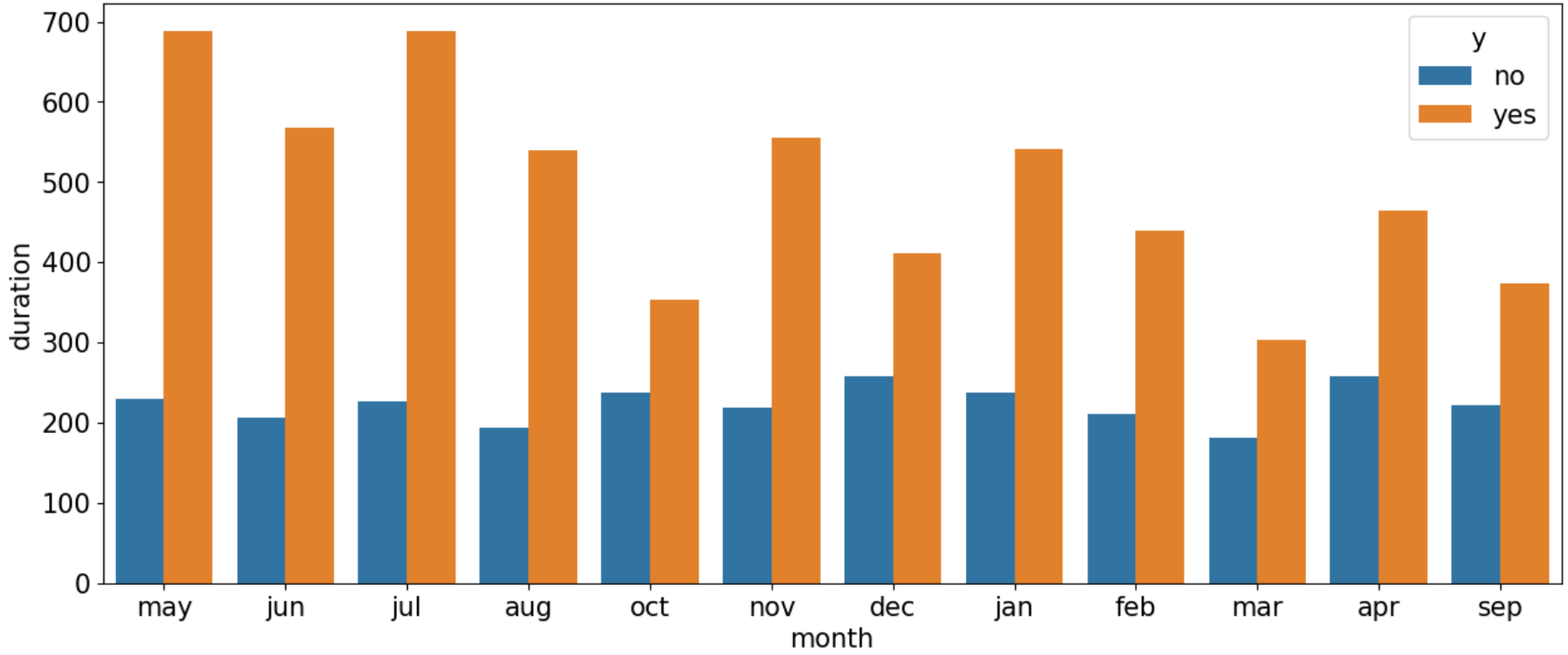


Inference:-

- who don’t have credit, housing and personal loan are more likely to subscribe to term deposits,
- who don’t have credit, personal but have housing are also more to subscribe
- If customer who has credit default, they not subscribing to term deposits, even if they don’t have housing or personal loan.

C Data Exploration

Month w.r.t to duration & Campaign

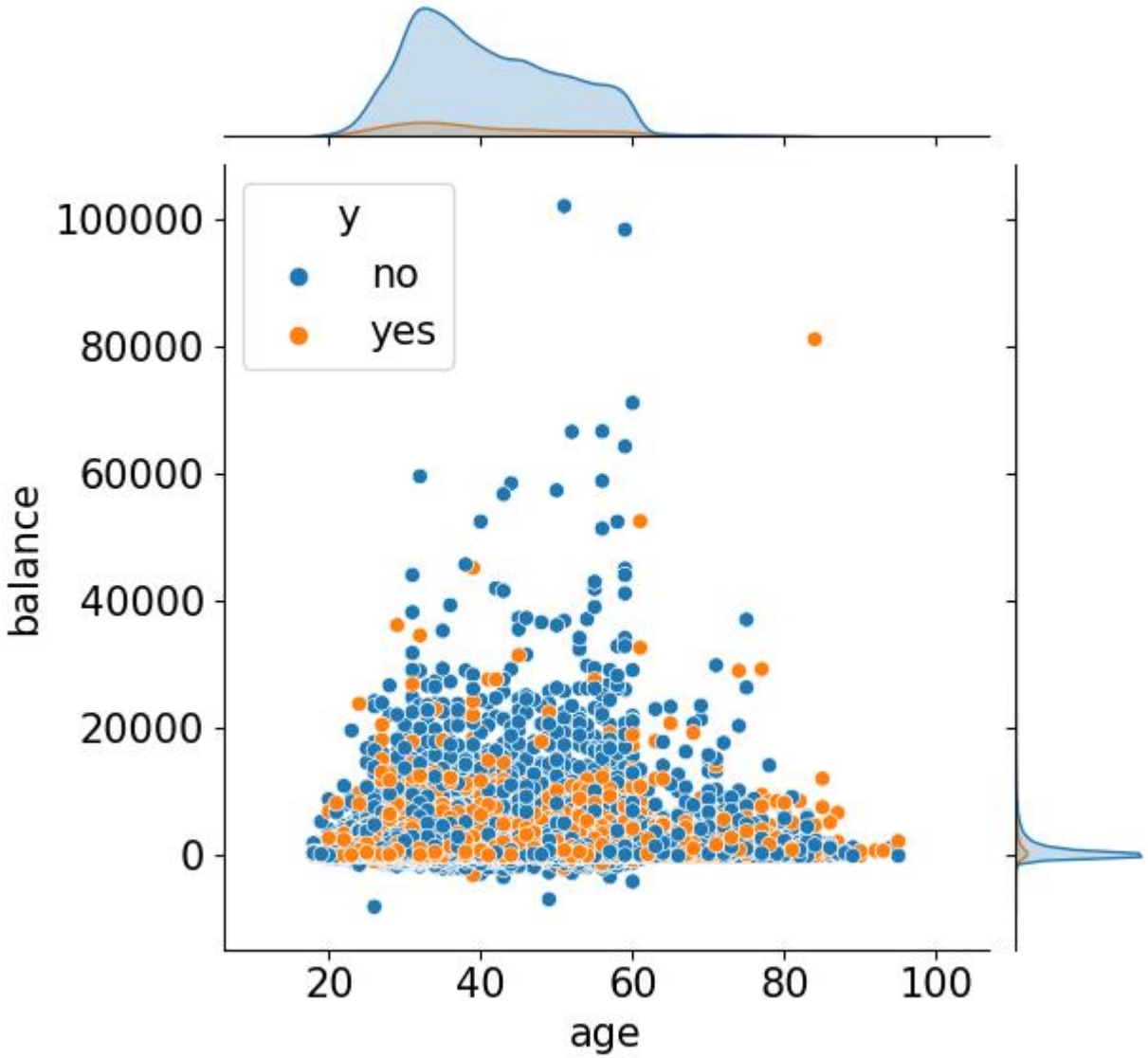


Inference:-

- we infer that increase in number of duration and number of calls have led to more subscriber.
- We can see that people have less call duration and a smaller number of calls have been subscribed less to our term deposits.

C Data Exploration

Age & Balance w.r.t Target

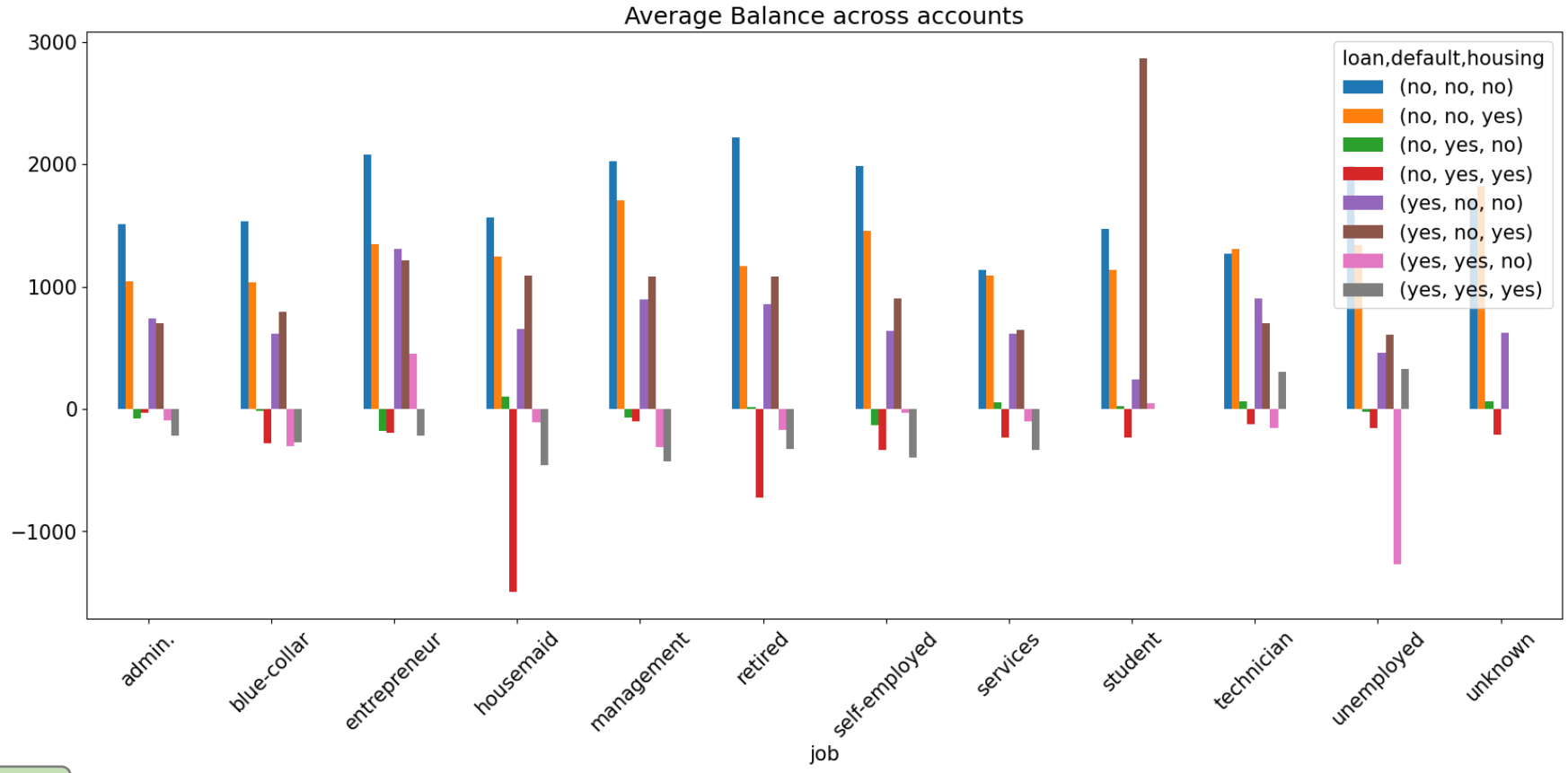


Inference:-

- we can say that customers with age between 20 to 60 with high balance said no for term deposits and we can target those customers for subscribing term deposits.
- Their balance is also very high and they have not been subscribed for term deposits.

C Data Exploration

Balance of each job and Loan categories w.r.t not subscribed



Inference:-

- Irresptive of the type of jobs except student
- When they personal, housing and credit, loan their avg balance falls into negative category
- People who only have housing loan their avg balance is high, does not falls into negative category irrespective of their jobs

C Data Exploration

Inference:-

Job
- ❑ People with management jobs are more attracted towards the policy
- ❑ we should increase out student customer base as the success rate is very high
- ❑ the least subscribed category is unknown and entrepreneur

Marital
- ❑ The married group is subscribing more
- ❑ Our customer base consist of more married people

Education
- ❑ Secondary educated people are subscribing more
- ❑ Most customers are secondary and tertiary educated

Default
- ❑ The blue-collar people are most number of defaulters
- ❑ The non defaulters are subscribing more
- ❑ Defaulters have very low account balance

Month
- ❑ May month has highest said no and yes, because may month has more number of calls .
- ❑ Dec month has less number of calls , so the yes and no are very less
- ❑ The same pattern can be seen for every month

Duration
- ❑ Most of our customers are not attending the calls
- ❑ The average call duration is between 0 to 1000 seconds
- ❑ People not attending the call are un aware of our policy
- ❑ From this we can see the more duration people talked the more people have been subscribed
- ❑ The subscriber count has been increased with the increase in duration
- ❑ The highest subscriber count is call between 300-400s

D Data Preprocessing

Great Learning
POWER AHEAD

Challenges in Data Set

- Imbalance Data Set
- Unknown
- Outliers

Scaling the data

```
from sklearn.preprocessing import StandardScaler
```

```
ss = StandardScaler()
```

```
df1.select_dtypes(exclude = 'object').columns
```

```
Index(['balance', 'day', 'duration', 'campaign', 'pdays', 'previous'], dtype='object')
```

```
col = ['balance', 'day', 'duration', 'campaign', 'pdays', 'previous']
```

```
df1[col]= ss.fit_transform(df1[col])
```

```
df1.head()
```

	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	white-collar	married	tertiary	no	0.414773	yes	no	cellular	-1.298476	second_quater	0.229429	-0.804324	-0.472533	-0.423628	prev_non_contact	no
1	blue-collar	single	secondary	no	-0.410774	yes	no	cellular	-1.298476	second_quater	-0.353322	-0.804324	-0.472533	-0.423628	prev_non_contact	no
2	white-collar	married	secondary	no	-0.431122	yes	yes	cellular	-1.298476	second_quater	-0.891336	-0.804324	-0.472533	-0.423628	prev_non_contact	no
3	blue-collar	married	primary	no	0.197685	yes	no	cellular	-1.298476	second_quater	-0.759656	-0.804324	-0.472533	-0.423628	prev_non_contact	no
4	white-collar	single	tertiary	no	-0.432119	no	no	cellular	-1.298476	second_quater	-0.084640	-0.804324	-0.472533	-0.423628	prev_non_contact	no

In this dataset we have used Scaling (**Standard Scaler**) because to change it into normal distribution

Because the data is skewed we have used Transformation (**Yeo-Johnson**) to transform and reduced the skewness

Check the Skewness

Before Transformation

```
age          0.684818
balance      8.360308
day          0.093079
duration     3.144318
campaign     4.898650
pdays        2.615715
previous    41.846454
dtype: float64
```

After Transformation

```
balance      1.098582
pdays        1.645213
duration     0.018045
campaign     0.230942
previous     1.646051
dtype: float64
```

Transformation

```
df1.head(3)
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	adult	white-collar	married	tertiary	no	2143	yes	no	cellular	5	second_quater	261	1	-1	0	prev_non_contact	0
1	mid	blue-collar	single	secondary	no	29	yes	no	cellular	5	second_quater	151	1	-1	0	prev_non_contact	0
2	mid	white-collar	married	secondary	no	2	yes	yes	cellular	5	second_quater	76	1	-1	0	prev_non_contact	0

- we are not removing the ouliers
- as we consider the every data point for our predictions
- This is bank dataset

D Data Preprocessing

Encoding

```
from sklearn.preprocessing import OrdinalEncoder,StandardScaler
```

Frequency encoding on age,poutcome,job,month columns

```
df1.age.value_counts()
```

```
o = df1.age.value_counts(normalize=True)
df2.age = df2.age.replace(o)
```

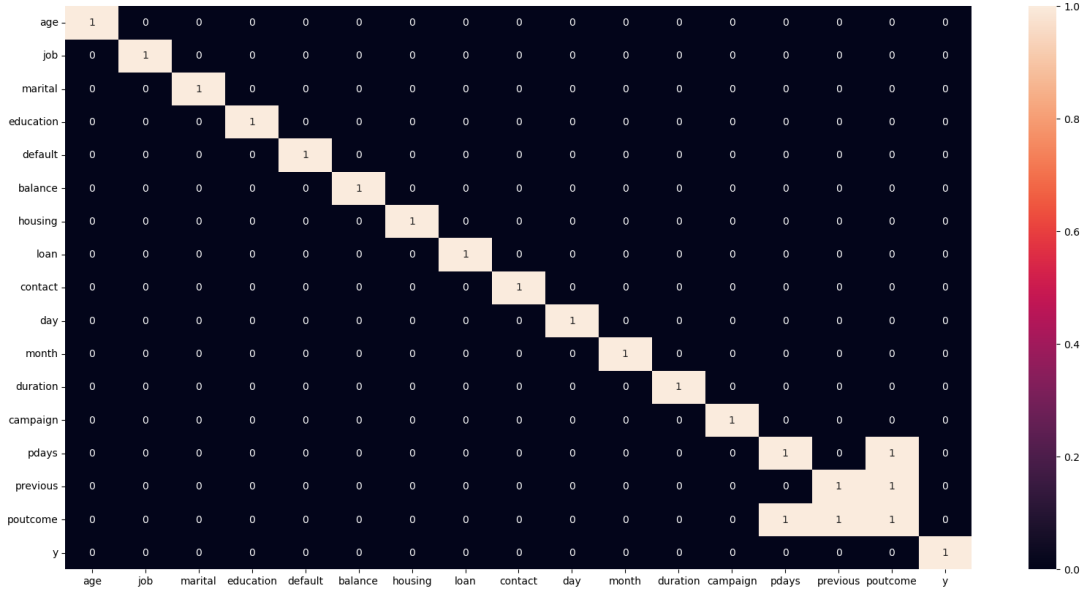
```
# # ordinal encoding
# enc = OrdinalEncoder(categories=[['young','mid','adult','old']])
# df2['age']=enc.fit_transform(df2[['age']])
```

```
df1.poutcome.value_counts()
```

```
c = df2.poutcome.value_counts(normalize=True)
df2.poutcome = df2.poutcome.replace(c)
```

```
df2.head(2)
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	0.293291	white-collar	married	tertiary	no	2143	yes	no	cellular	5	second_quater	261	1	-1	0	0.817478	0
1	0.524939	blue-collar	single	secondary	no	29	yes	no	cellular	5	second_quater	151	1	-1	0	0.817478	0



- Check in for correlation after encoding the data

Label encoding

- Job
- Month
- default
- hosing
- Loan

One hot encoding

- Marital
- Contact

Inference:-

In this data set we have used

Frequency Encoding for P-outcome , Education

E Model Building

LogisticRegression → DecisionTreeClassifier → RandomForestClassifier → KNeighborsClassifier → GaussianNB → AdaBoostClassifier → GradientBoostingClassifier → XGB Classifier → LGBM Classifier

Before Smote

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
7	XGBClassifier	0.953661	0.910096	7705	289	524	525	0.50	0.64	0.56
8	LGBMClassifier	0.929579	0.910870	7720	274	532	517	0.49	0.65	0.56
2	RandomForestClassifier	0.999972	0.905120	7751	243	615	434	0.41	0.64	0.50
6	GradientBoostingClassifier	0.908980	0.905894	7764	230	621	428	0.41	0.65	0.50
1	DecisionTreeClassifier	1.000000	0.875373	7436	558	569	480	0.46	0.46	0.46
5	AdaBoostClassifier	0.900852	0.898043	7745	249	673	376	0.36	0.60	0.45
4	GaussianNB	0.804274	0.806701	6802	1192	556	493	0.47	0.29	0.36
0	LogisticRegression	0.892363	0.889749	7787	207	790	259	0.25	0.56	0.34
3	KNeighborsClassifier	0.913045	0.885768	7776	218	815	234	0.22	0.52	0.31

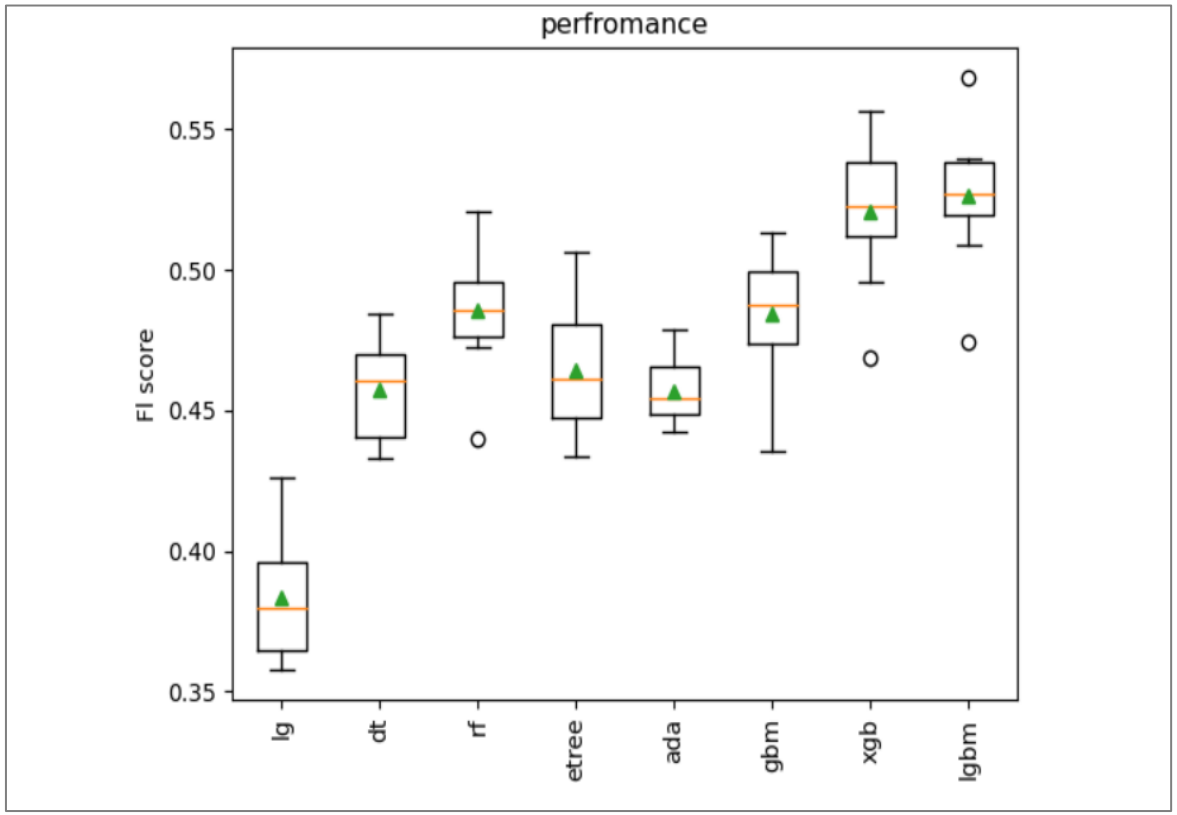
- In Base model without smote the maximum f1 score achived is 0.56 in XGB,LGBM

After Smote

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.948321	0.906115	7560	434	415	634	0.60	0.59	0.60
7	XGBClassifier	0.967192	0.906558	7627	367	478	571	0.54	0.61	0.57
6	GradientBoostingClassifier	0.916515	0.874599	7188	806	328	721	0.69	0.47	0.56
2	RandomForestClassifier	1.000000	0.888643	7433	561	446	603	0.57	0.52	0.54
5	AdaBoostClassifier	0.897050	0.862877	7119	875	365	684	0.65	0.44	0.52
0	LogisticRegression	0.852418	0.815880	6624	1370	295	754	0.72	0.35	0.48
1	DecisionTreeClassifier	1.000000	0.854915	7131	863	449	600	0.57	0.41	0.48
3	KNeighborsClassifier	0.936294	0.817538	6685	1309	341	708	0.67	0.35	0.46
4	GaussianNB	0.752067	0.630764	4942	3052	287	762	0.73	0.20	0.31

- In Base model after smote the maximum f1 score achieved is 0.60,0.57,0.56 by LGBM, XGB,Gradient Boosting
- The highest TP is achieved by Logistic Regression,Gaussian NB Model at a f1 score of 0.47,0.31

F1 Score



E Model Building (Adding Features)

Per call time

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.947429	0.903461	7528	466	407	642	0.61	0.58	0.60
2	RandomForestClassifier	1.000000	0.890302	7426	568	424	625	0.60	0.52	0.56
6	GradientBoostingClassifier	0.916186	0.873383	7179	815	330	719	0.69	0.47	0.56
7	XGBClassifier	0.969823	0.904235	7618	376	490	559	0.53	0.60	0.56
5	AdaBoostClassifier	0.895578	0.860334	7102	892	371	678	0.65	0.43	0.52
0	LogisticRegression	0.848566	0.814221	6612	1382	298	751	0.72	0.35	0.47
3	KNeighborsClassifier	0.938142	0.816654	6662	1332	326	723	0.69	0.35	0.47
1	DecisionTreeClassifier	1.000000	0.850824	7151	843	506	543	0.52	0.39	0.45
4	GaussianNB	0.784860	0.678425	5357	2637	271	778	0.74	0.23	0.35

- In Base model after smote the maximum f1 score achieved is LGBM(0.60),XGB(0.57),Gradient Boosting(0.56)
- Adding per_call_time feature, the f1_score of LGBM(0.60),GBM(0.56),XGB(0.56)

Binning of job

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.948337	0.904567	7551	443	420	629	0.60	0.59	0.59
6	GradientBoostingClassifier	0.919193	0.878359	7224	770	330	719	0.69	0.48	0.57
7	XGBClassifier	0.968210	0.905562	7632	362	492	557	0.53	0.61	0.57
2	RandomForestClassifier	1.000000	0.890634	7470	524	465	584	0.56	0.53	0.54
5	AdaBoostClassifier	0.899524	0.867411	7171	823	376	673	0.64	0.45	0.53
0	LogisticRegression	0.854438	0.817317	6648	1346	306	743	0.71	0.36	0.47
1	DecisionTreeClassifier	1.000000	0.856685	7202	792	504	545	0.52	0.41	0.46
3	KNeighborsClassifier	0.936513	0.813557	6665	1329	357	692	0.66	0.34	0.45
4	GaussianNB	0.756562	0.634082	4972	3022	287	762	0.73	0.20	0.32

- In Base model after smote the maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.56)
- Adding per_call_time feature, the f1_score of LGBM(0.59), GBM(0.56), and Guassian NB has low f1_score
- Adding the job category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)

Binning of Age

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.948713	0.904678	7554	440	422	627	0.60	0.59	0.59
6	GradientBoostingClassifier	0.917001	0.876921	7197	797	316	733	0.70	0.48	0.57
7	XGBClassifier	0.967646	0.903572	7602	392	480	569	0.54	0.59	0.57
2	RandomForestClassifier	1.000000	0.891297	7448	546	437	612	0.58	0.53	0.55
5	AdaBoostClassifier	0.897833	0.861772	7129	865	385	664	0.63	0.43	0.52
0	LogisticRegression	0.854062	0.815659	6630	1364	303	746	0.71	0.35	0.47
1	DecisionTreeClassifier	1.000000	0.858012	7217	777	507	542	0.52	0.41	0.46
3	KNeighborsClassifier	0.936451	0.813889	6667	1327	356	693	0.66	0.34	0.45
4	GaussianNB	0.778752	0.673781	5313	2681	269	780	0.74	0.23	0.35

- In Base model after smote the maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.56)
- Adding per_call_time feature, the f1_score of LGBM(0.59), GBM(0.56), and Guassian NB has low f1_score
- Adding the job category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)
- Adding the age category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)

Binning of month

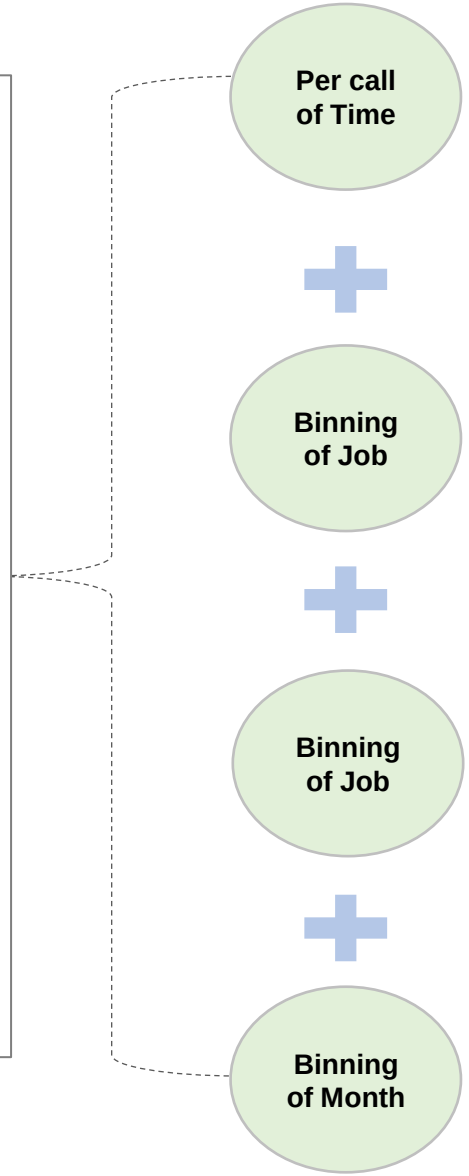
	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.948619	0.905120	7536	458	400	649	0.62	0.59	0.60
2	RandomForestClassifier	1.000000	0.898706	7512	482	434	615	0.59	0.56	0.57
6	GradientBoostingClassifier	0.921433	0.883003	7288	706	352	697	0.66	0.50	0.57
7	XGBClassifier	0.969337	0.904899	7620	374	486	563	0.54	0.60	0.57
5	AdaBoostClassifier	0.900761	0.863651	7157	837	396	653	0.62	0.44	0.51
0	LogisticRegression	0.852669	0.813447	6608	1386	301	748	0.71	0.35	0.47
1	DecisionTreeClassifier	1.000000	0.860113	7211	783	482	567	0.54	0.42	0.47
3	KNeighborsClassifier	0.937469	0.813779	6651	1343	341	708	0.67	0.35	0.46
4	GaussianNB	0.752083	0.627447	4912	3082	287	762	0.73	0.20	0.31

- In Base model after smote the maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.56)
- Adding per_call_time feature, the f1_score of LGBM(0.59), GBM(0.56), and Guassian NB has low f1_score
- Adding the job category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)
- Adding the age category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)
- Adding the month categoty maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.57)

F Model Building (Adding All Features)

	Model	Train_ACC	Test_ACC	TN	FP	FN	TP	recall	Precision	F1 score
8	LGBMClassifier	0.948932	0.906005	7549	445	405	644	0.61	0.59	0.60
7	XGBClassifier	0.970903	0.908548	7640	354	473	576	0.55	0.62	0.58
2	RandomForestClassifier	1.000000	0.893951	7468	526	433	616	0.59	0.54	0.56
6	GradientBoostingClassifier	0.923985	0.880018	7258	736	349	700	0.67	0.49	0.56
5	AdaBoostClassifier	0.904285	0.867080	7168	826	376	673	0.64	0.45	0.53
0	LogisticRegression	0.853765	0.816101	6635	1359	304	745	0.71	0.35	0.47
1	DecisionTreeClassifier	1.000000	0.861661	7240	754	497	552	0.53	0.42	0.47
3	KNeighborsClassifier	0.938158	0.816764	6665	1329	328	721	0.69	0.35	0.47
4	GaussianNB	0.802196	0.715913	5699	2295	274	775	0.74	0.25	0.38

- In Base model after smote the maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.56)
- Adding per_call_time feature, the f1_score of LGBM(0.59), GBM(0.56), and Guassian NB has low f1_score
- Adding the job category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)
- Adding the age category maximum f1 score achieved is LGBM(0.59), XGB(0.57), Gradient Boosting(0.57)
- Adding the month categoty maximum f1 score achieved is LGBM(0.60), XGB(0.57), Gradient Boosting(0.57)
- After adding all the category maximum f1 score achieved is LGBM(0.60), XGB(0.58), Gradient Boosting(0.56)



F Model Evaluation (Hyperparameters Tunning)

- Based on feature engineering we created multiple models, In that feature engineering with month binned has good scores
- From that LGBM,XGB,GB performed good, So were tunning the models to get better scores

LGBM

```
[[7514  480]
 [ 382  667]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.95      7994
           1       0.58      0.64      0.61      1049

    accuracy                           0.90      9043
   macro avg       0.77      0.79      0.78      9043
weighted avg       0.91      0.90      0.91      9043
```

KFold

```
[[7507  487]
 [ 386  663]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.95      7994
           1       0.58      0.63      0.60      1049

    accuracy                           0.90      9043
   macro avg       0.76      0.79      0.77      9043
weighted avg       0.91      0.90      0.91      9043
```

- After doing the hyper parameter tuning the FN is increased 0.61 and the TP is increased.
- But the model predictions have been improved.

XGB

```
[[7613  381]
 [ 463  586]]
              precision    recall  f1-score   support

           0       0.94      0.95      0.95      7994
           1       0.61      0.56      0.58      1049

    accuracy                           0.91      9043
   macro avg       0.77      0.76      0.76      9043
weighted avg       0.90      0.91      0.91      9043
```

KFold

```
[[7632  362]
 [ 464  585]]
              precision    recall  f1-score   support

           0       0.94      0.95      0.95      7994
           1       0.62      0.56      0.59      1049

    accuracy                           0.91      9043
   macro avg       0.78      0.76      0.77      9043
weighted avg       0.90      0.91      0.91      9043
```

- After doing the hyper parameter tuning the TP is increased the f1 score 0.59 also increased for Gradient Boosting

Gradient Boosting

```
[[7424  570]
 [ 372  677]]
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      7994
           1       0.54      0.65      0.59      1049

    accuracy                           0.90      9043
   macro avg       0.75      0.79      0.77      9043
weighted avg       0.90      0.90      0.90      9043
```

KFold

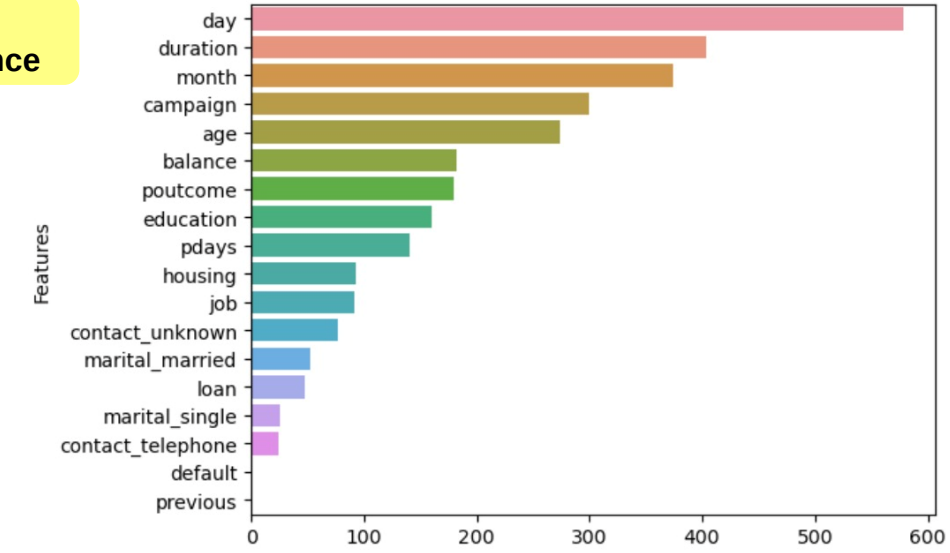
```
[[7270  724]
 [ 343  706]]
              precision    recall  f1-score   support

           0       0.95      0.91      0.93      7994
           1       0.49      0.67      0.57      1049

    accuracy                           0.88      9043
   macro avg       0.72      0.79      0.75      9043
weighted avg       0.90      0.88      0.89      9043
```

- After doing the hyper parameter tuning the TP is increased the f1 score (0.58) also increased for xgb

Feature Importance



- Removing the default and previous from data and checking for improvement

F Model Evaluation (Removing the default and previous from data and checking for improvement)

LGBM After hyperparameter tuning

```
0.9475851916812829
0.9042353201371226
[[7506  488]
 [ 378  671]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.95      7994
           1       0.58      0.64      0.61      1049

    accuracy                           0.90      9043
   macro avg       0.77      0.79      0.78      9043
weighted avg       0.91      0.90      0.91      9043
```

LGBM After feature importance

```
0.9475851916812829
0.9042353201371226
[[7506  488]
 [ 378  671]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.95      7994
           1       0.58      0.64      0.61      1049

    accuracy                           0.90      9043
   macro avg       0.77      0.79      0.78      9043
weighted avg       0.91      0.90      0.91      9043
```

- After hyper parameter tuning the f1 score is increased from 0.60 -0.61
- After doing feature importance (removing default and previous) the f1 score remains the same , so all other features are needed for model building

Conclusions

SUGGESTIONS :

Based on EDA observations the following suggestions have been made:

- Job, Credit default, Housing and Loan significantly affect the customer experience along with the several other variables considered.
- The bank should highly focus on their marketing strategy on promoting their new scheme of term deposits instead of using the old technique of marketing through phone calls as adapting different new strategies may attract customers effectively and can be time and cost reducing.
- Considering the customer's financial liability and stability is important for the business as the customer with huge liabilities and financial risks such as housing loans, personal loans, credit card defaults are prone to risks which may play a major in subscribing to the term deposits or predominantly any other schemes/plans.
- The financial institutions must ensure that their customers have potential qualifications and strong financial background before enrolling them into any kind of policies or schemes introduced such as work background as many of their customers might be students in some cases and they might not have the potential and necessity at present in enrolling/subscribing to any such new schemes.
- As per our findings most of our customers are married and it's also found that married customers have subscribed more than the customers who are single. Concentrating on this category can help increase the favourable conditions for the bank on moving forward with term deposit policy.
- Focusing on job category and especially on customers from work background such as blue collar, management, technician, self-employed and admin as its noticed that these category customers had subscribed more in number than the rest with proper marketing and strategies, we can attract more customers from this category to subscribe to the term deposits as well attract new customers.

Business Interpretation

(Overall Suggestion to the business)

- Through adapting various marketing strategies, the banking institution can probably attract its customers to subscribe on to term deposits as through data analysis we've found that certain job categories such as management, technicians, blue collar, self-employed etc have quite well amount of balance maintained in their accounts and also few customers did not have any loans or credit defaults at present. These categories of customers can be focused more than the rest to assure that the scheme reaches maximum no of customers and making them to subscribe on larger number too.

- The major setback of this banking institution is their marketing method adapted as they've adapted a single strategy for approaching customers i.e., through phone calls. The bank can use either a different strategy or a variety of marketing methods particularly on targeting a set of customers or promoting their new scheme introduced. As their current marketing strategy is barely time and cost effective.

- Major concern of this scheme term deposits is not all customers would have the need to subscribe to term deposits or the idea of investments as the bank has larger categories of people having relationship with it. They can focus on a set of customers to reduce the time of calling each customer and asking for their suggestion. As per EDA we've found that married customers, customers with no personal loan, credit default have said 'yes' to subscribing thus it can particularly attract these categories of customers for a effective reach and also can expand their customer base in these categories in future.