Capstone Project - 3

Bank marketing effectiveness prediction

Team Members

Ganeshkumar Patel

Akanksha Agarwal

Saurabh Funde

Sanjay Kumar

Yaman Saini

Data Science Trainee,

AlmaBetter, Bangalore

Contents:

Problem Statement

- **Aim:- Predicting the effectiveness of bank marketing campaigns**

- Problem Statement: The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.The classification goal is to predict if the client will subscribe a term deposit (variable ‘y’).

Data Summary

Categorical Features

- Marital - (Married , Single , Divorced)
- Job-(Management,BlueCollar,retired etc)
- Contact - (Telephone,Cellular,Unknown)
- Education (Primary,Secondary,Tertiary)
- Month-(Jan,Feb,Mar,Apr,May etc)
- Poutcome - (Success,Failure,Other,Unknown)
- Housing - (Yes/No)
- Loan - (Yes/No)
- Default - (Yes/No)

● Numerical Features

- Age
- Balance
- Day
- Duration
- Campaign
- Pdays

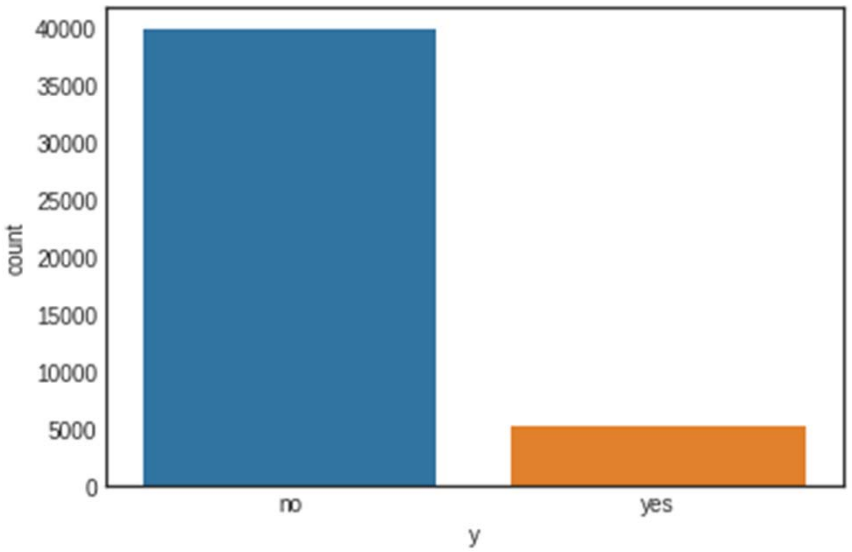
	job	marital	education	default	housing	loan	contact	month	poutcome	y
count	45211	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	12	3	4	2	2	2	3	12	4	2
top	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	13766	36959	39922

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

Exploratory Data Analysis

1. How many people subscribed the Term deposit?

```
no     0.883015
yes    0.116985
Name: y, dtype: float64
```

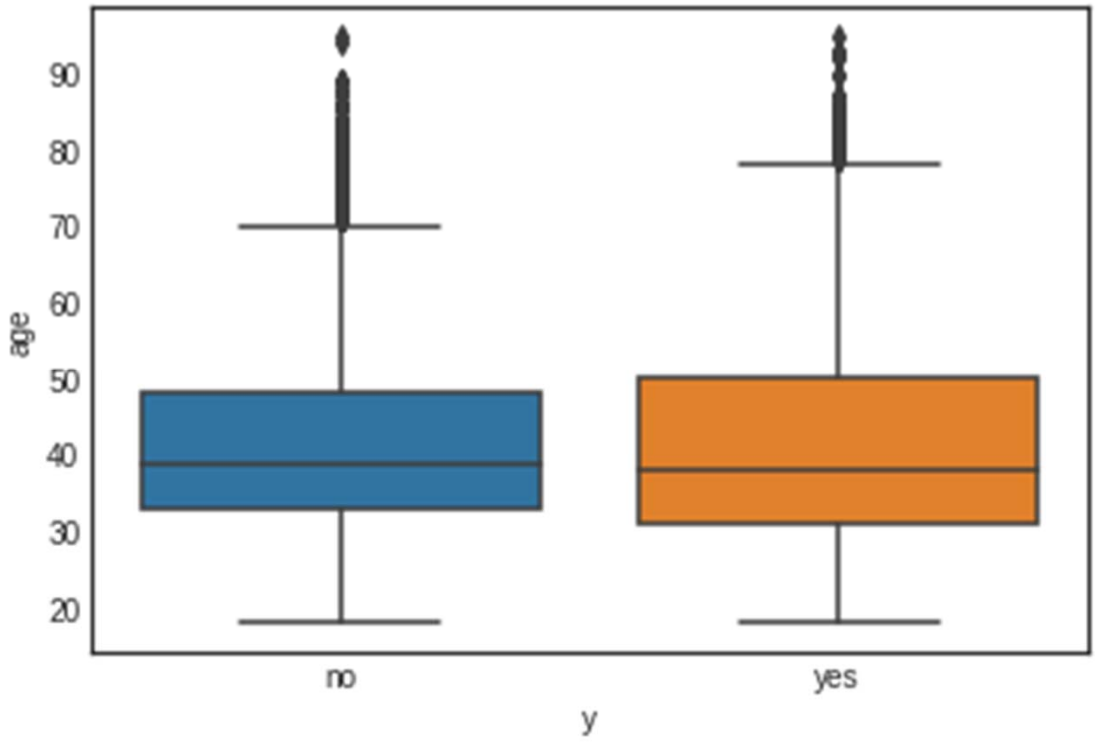


The target variable ‘y’ tells us the outcome of the campaign whether they went ahead for the term deposit or not.

- Out of 100 only 11.70 people subscribed to the term deposit.

EDA(continued)

- **1. Age**

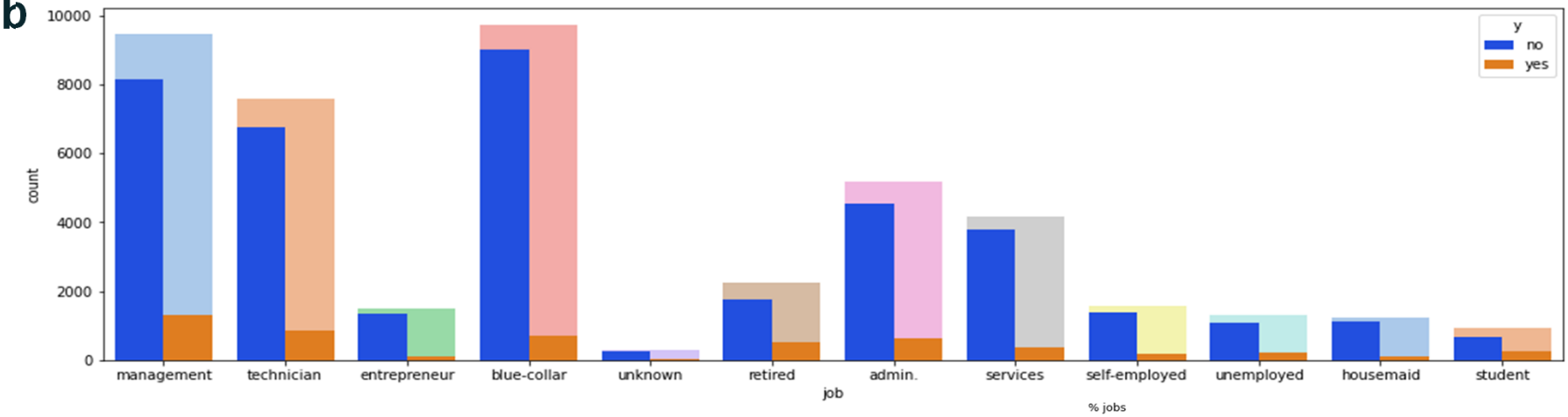


Inference:

In the above plot it is clear that a majority of customers called is in the age of 30s and 40s (33 to 48 years old fall within the 25th to 75th percentiles) and for each of the target variable the age feature is not linearly separable. Thus age will be of less importance to us.

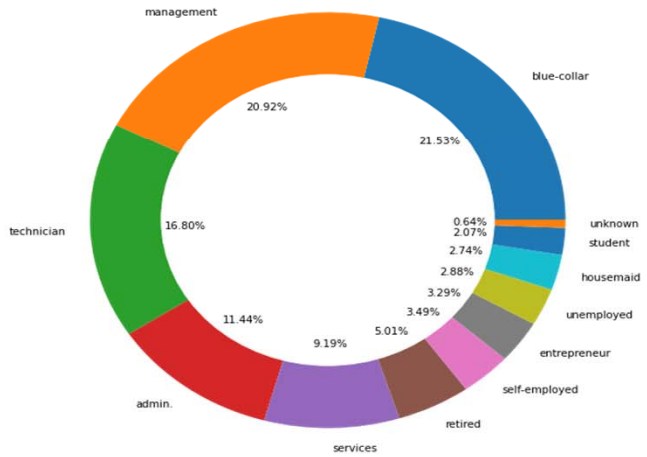
EDA(continued)

- **2. Job**



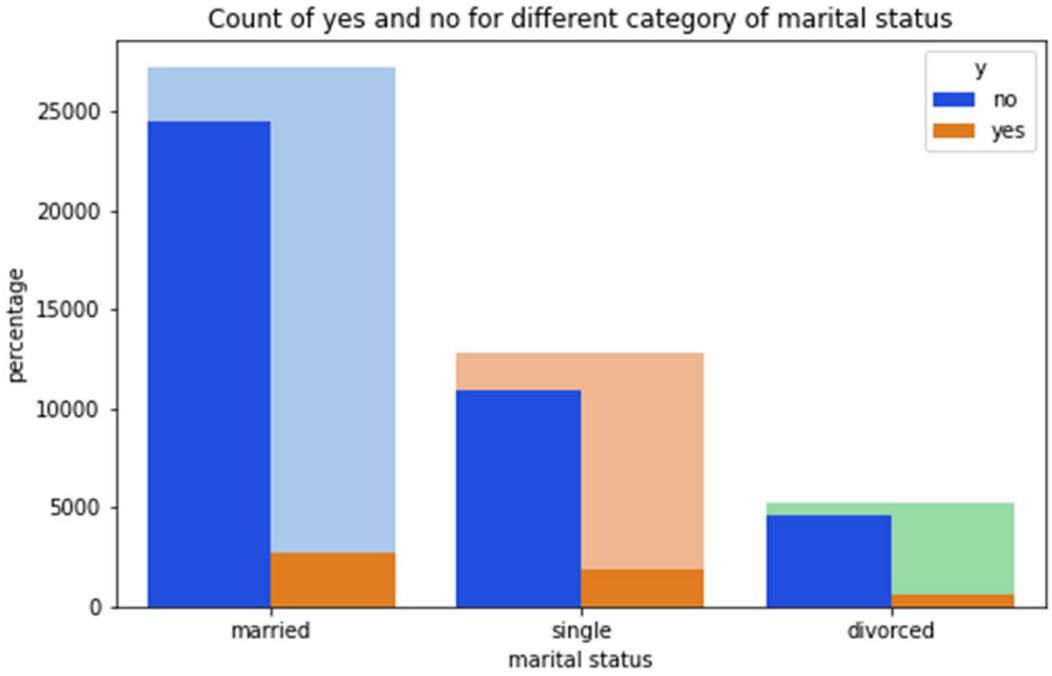
Inference:

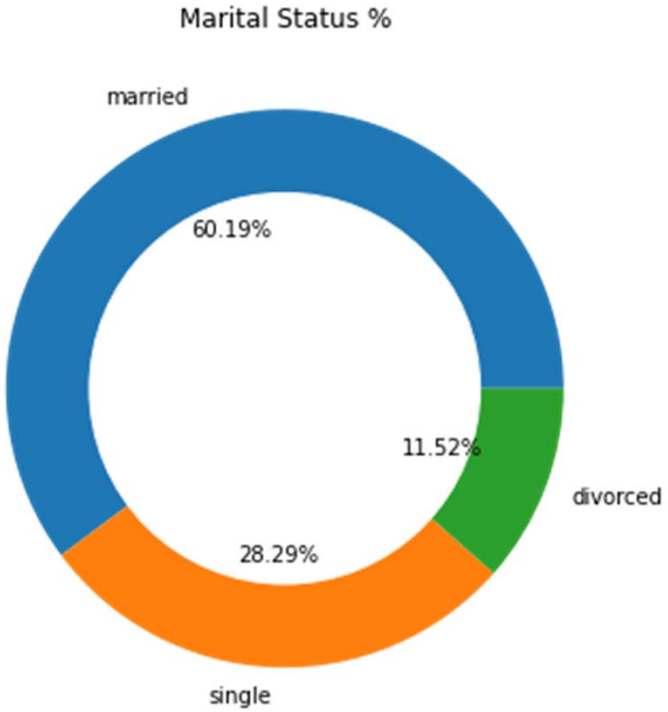
Customers with Blue-collar, management and technician showed maximum interest in subscription. We can also observe the large variance in our data among all categories.



EDA(continued)

2. Marital



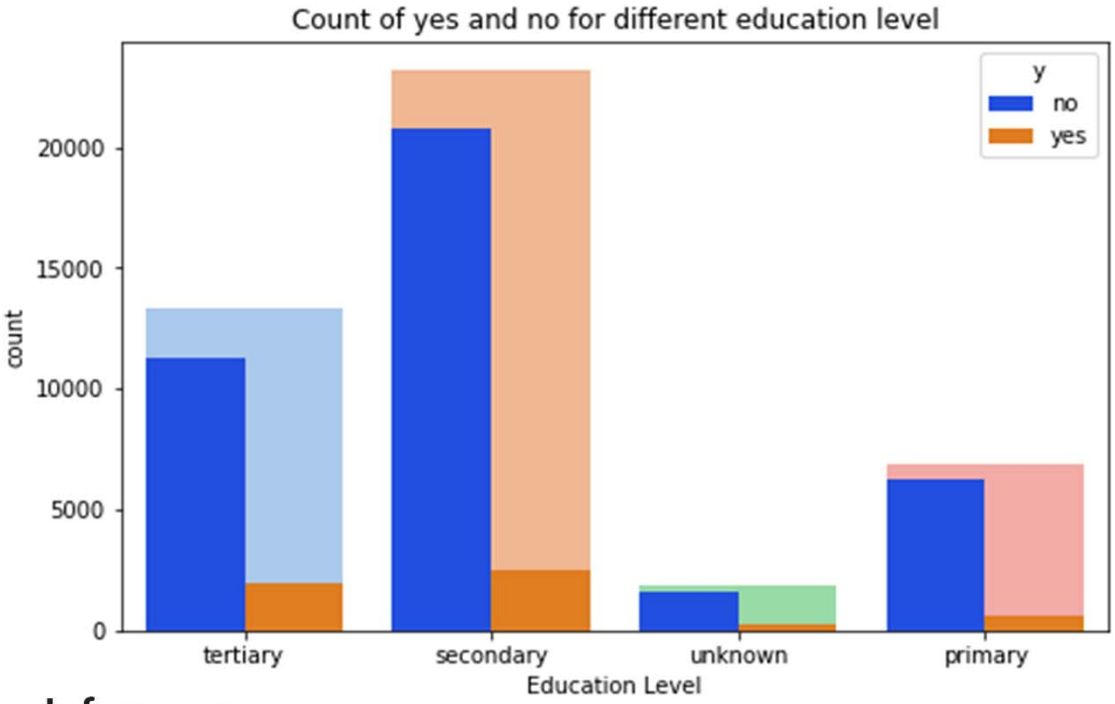


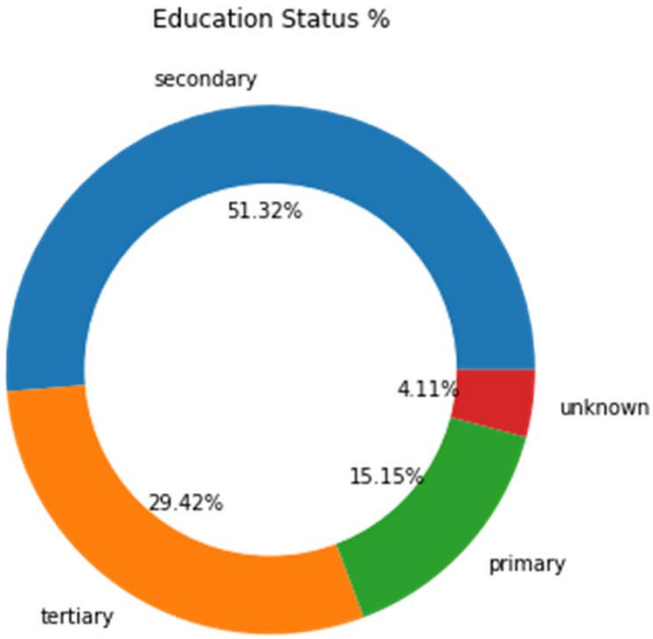
Inference:

Married and single showed more interest in term deposit

EDA(continued)

3. Education



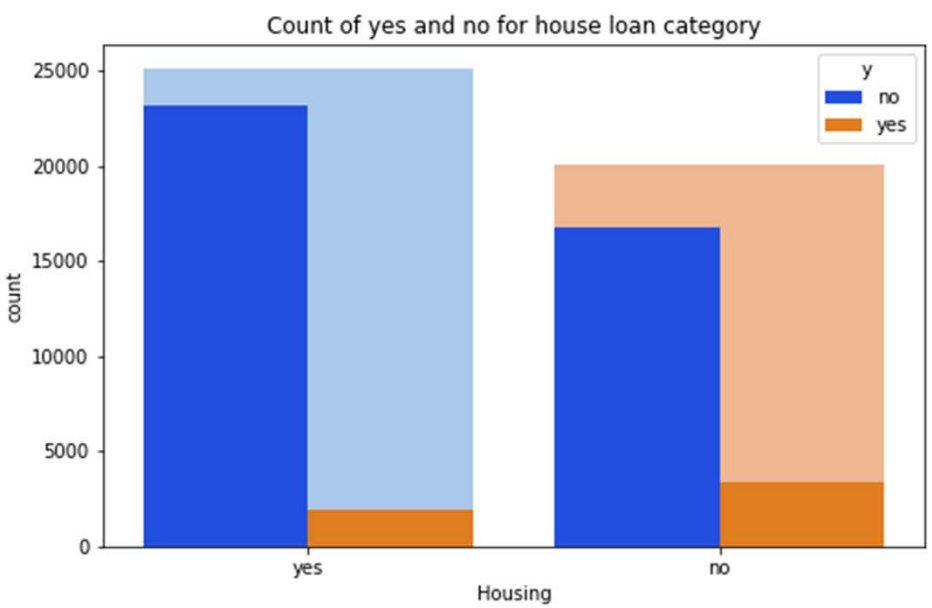


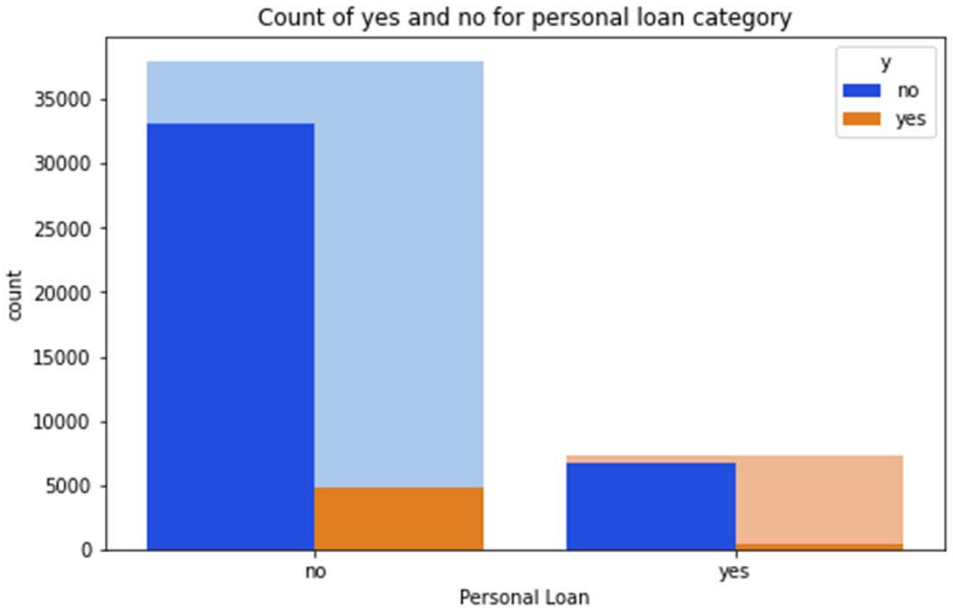
Inference:

In comparison to primary and some unknown education, people with secondary and tertiary education were more driven towards paying term deposit in bank.

EDA(continued)

3. Loans





Inference:

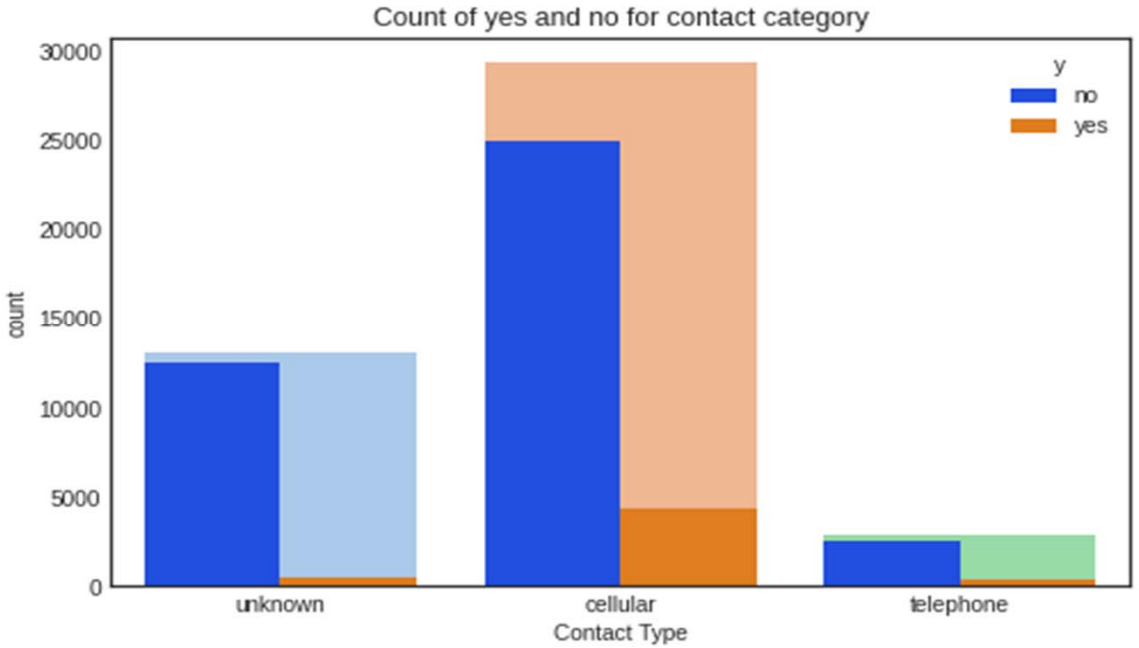
Majority of the people had previous housing loans and thus very few of them opted for term deposit.

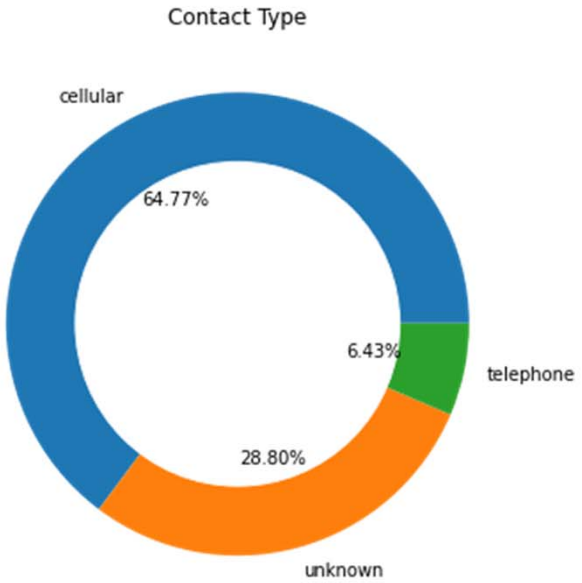
Inference:

Majority of the people had personal loans and thus very few of them opted for term deposit.

EDA(continued)

4. Contact



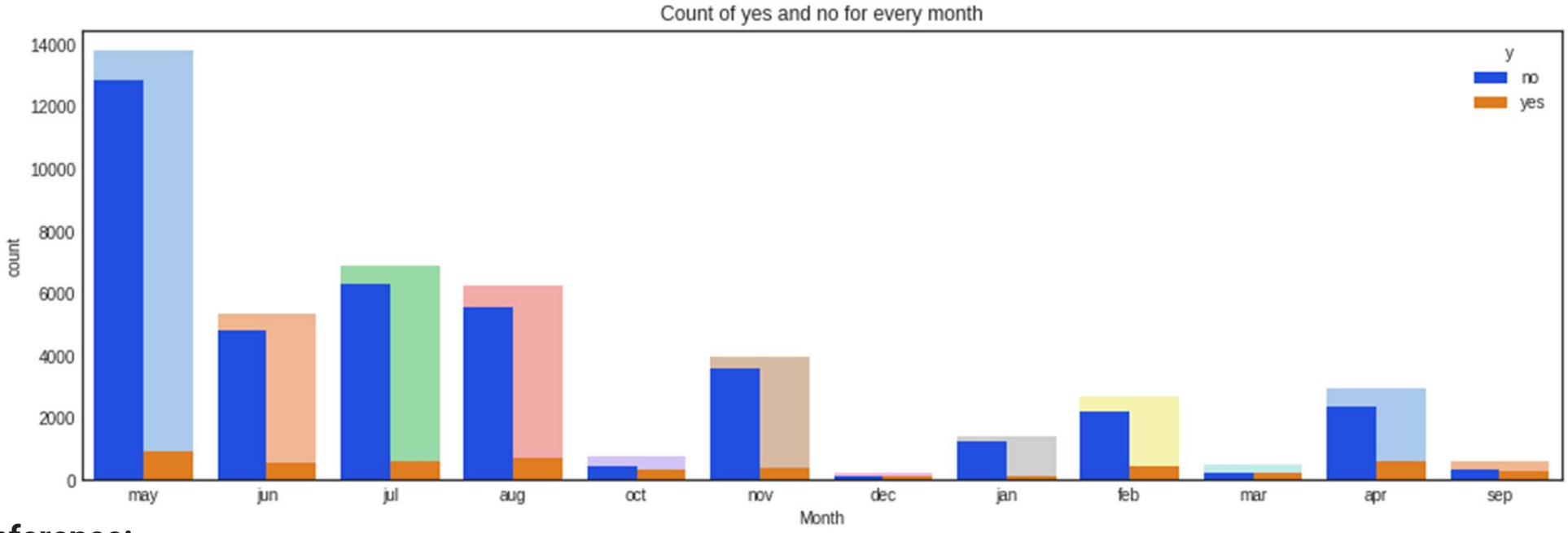


Inference:

Majority of the people were contacted through cellular medium and were converted to the subscription. Thus, cellular medium of contact is more effective in comparison to telephone and other mediums.

EDA(continued)

5. Month

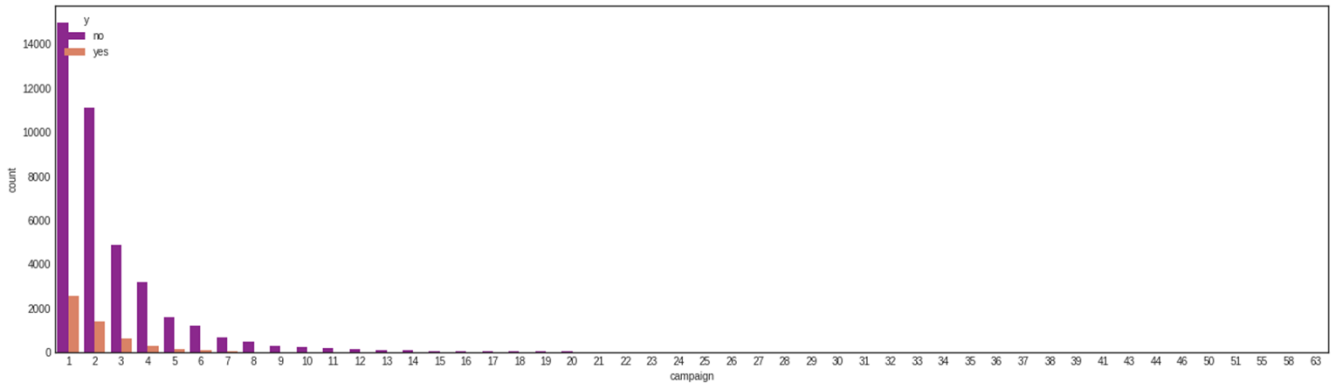


Inference:

During the month of may there were maximum subscriptions with relatively good subscriptions in June, July and august. During other months we can se less subscription and so we can combine few of them attar on

EDA(continued)

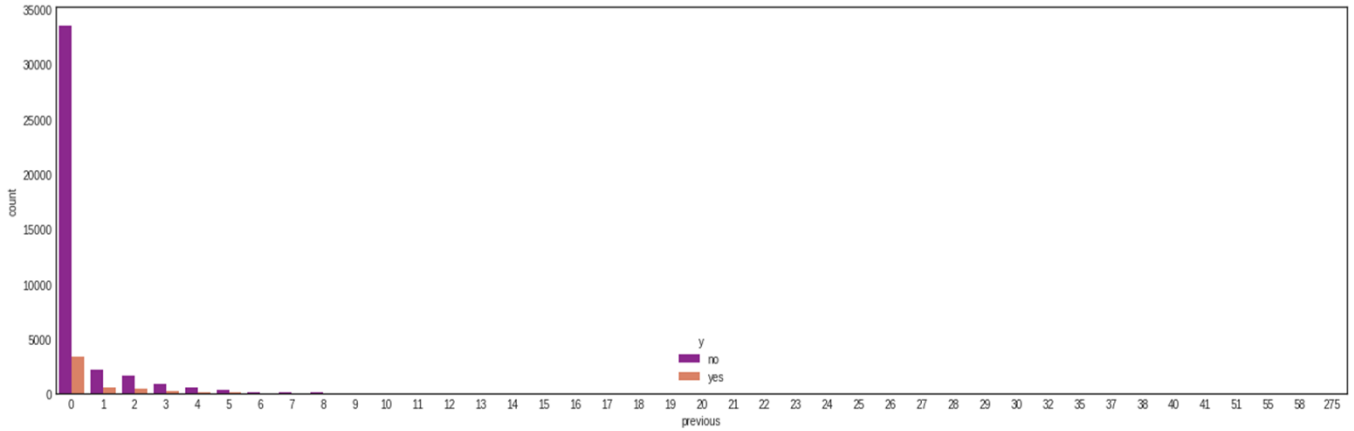
6. Campaign



Inference:

People were mostly contacted once who subscribed to it, while others were contacted more number of times but the conversion rate reduced after 20 we did not see any significant conversions thus we drop those observations later on

7. Previous

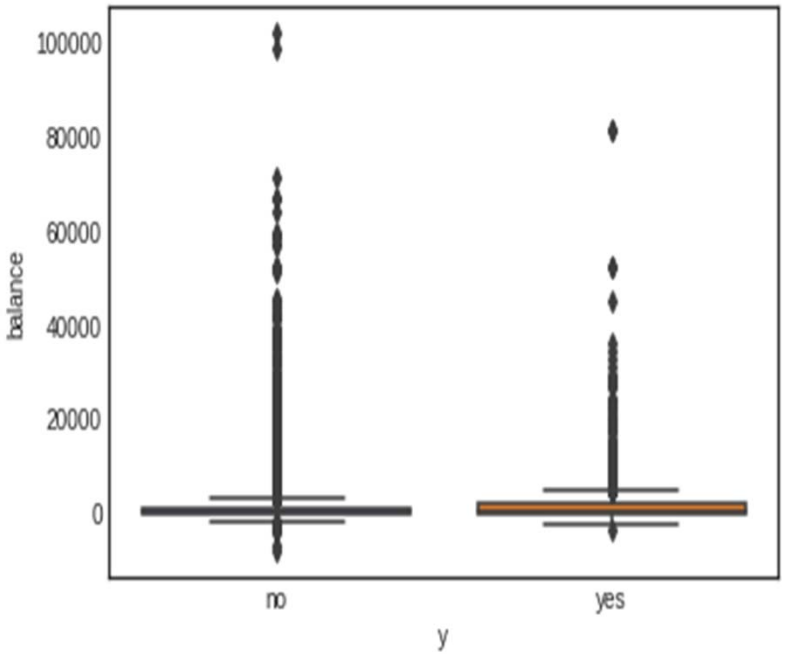


Inference:

We can see above that majority of people were not contacted previously before this campaign and there are no significant contacts after 11 times already done.

EDA(continued)

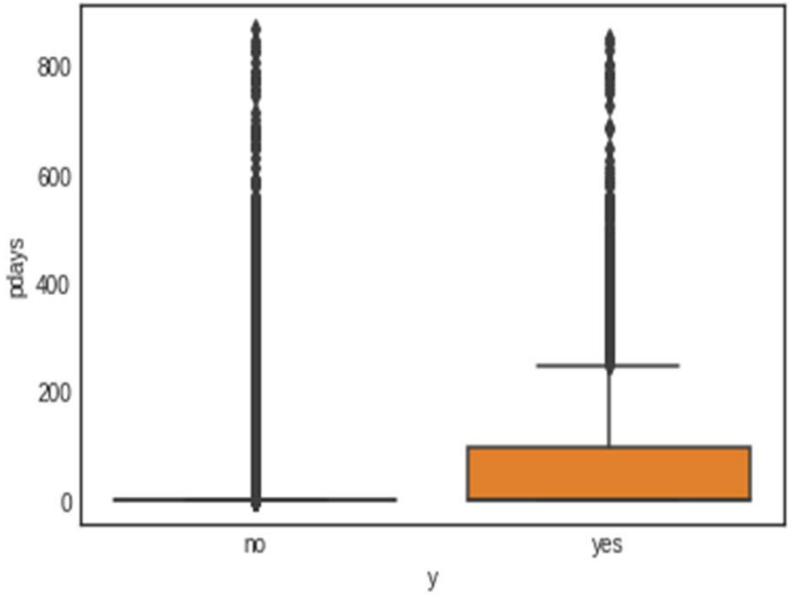
8. Balance



Inference:

Balance of customers is more than 1 lacs and thus we need to remove ouliers as the median is very less near to 450.

9. Pdays

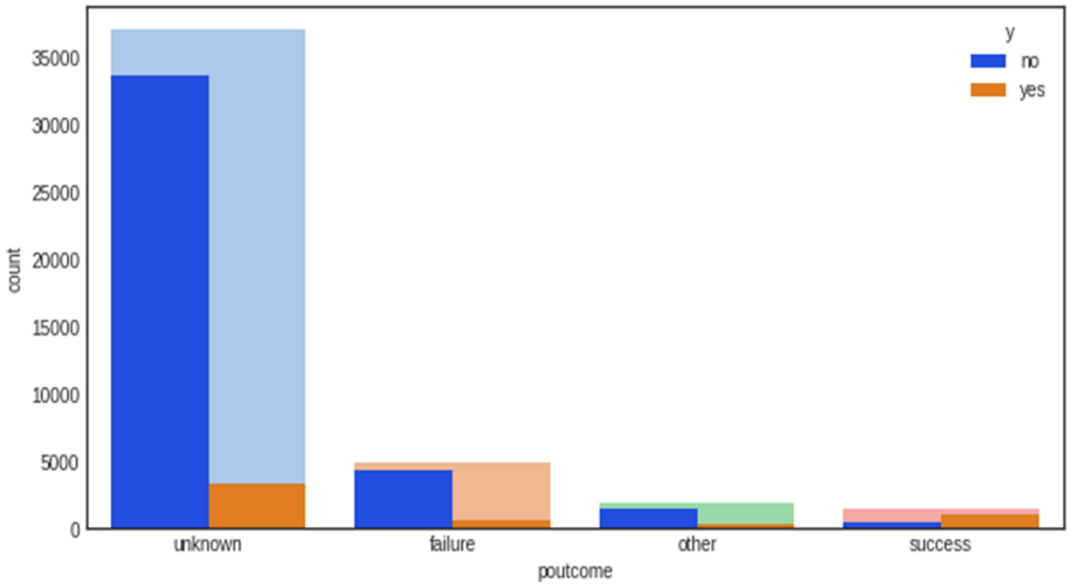


Inference:

pdays have large outliers and will have to look upon more closely.

EDA(continued)

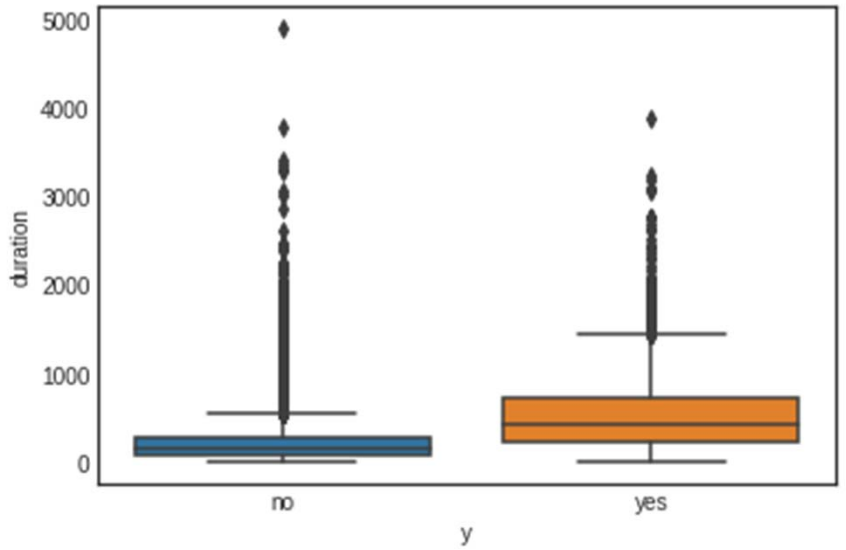
- **Poutcome**



Inference:

Looking at poutcome we can infer that the success rate was high for some unknown category.

- **Duration**

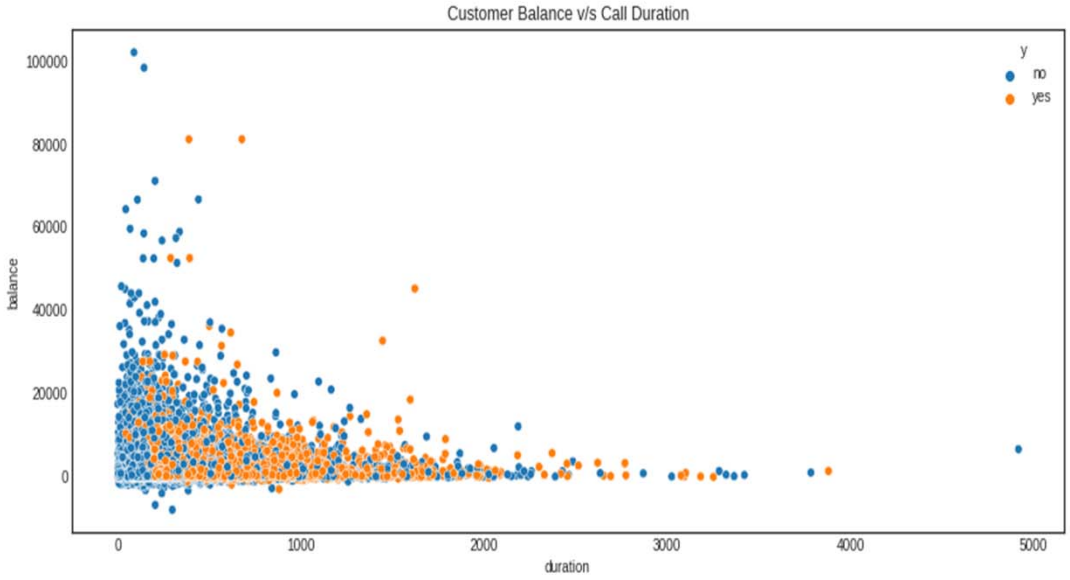


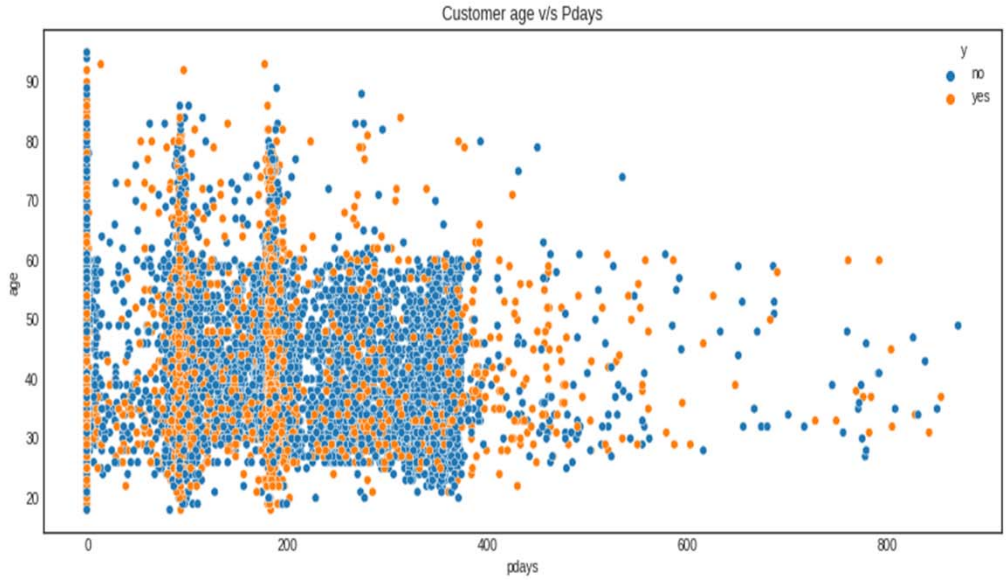
Inference:

The above box plot shows that calls with large duration has more tendency for conversion..

EDA(continued)

- Increase in duration leads to more term deposit.
- We are seeing some pattern of more term with pdays



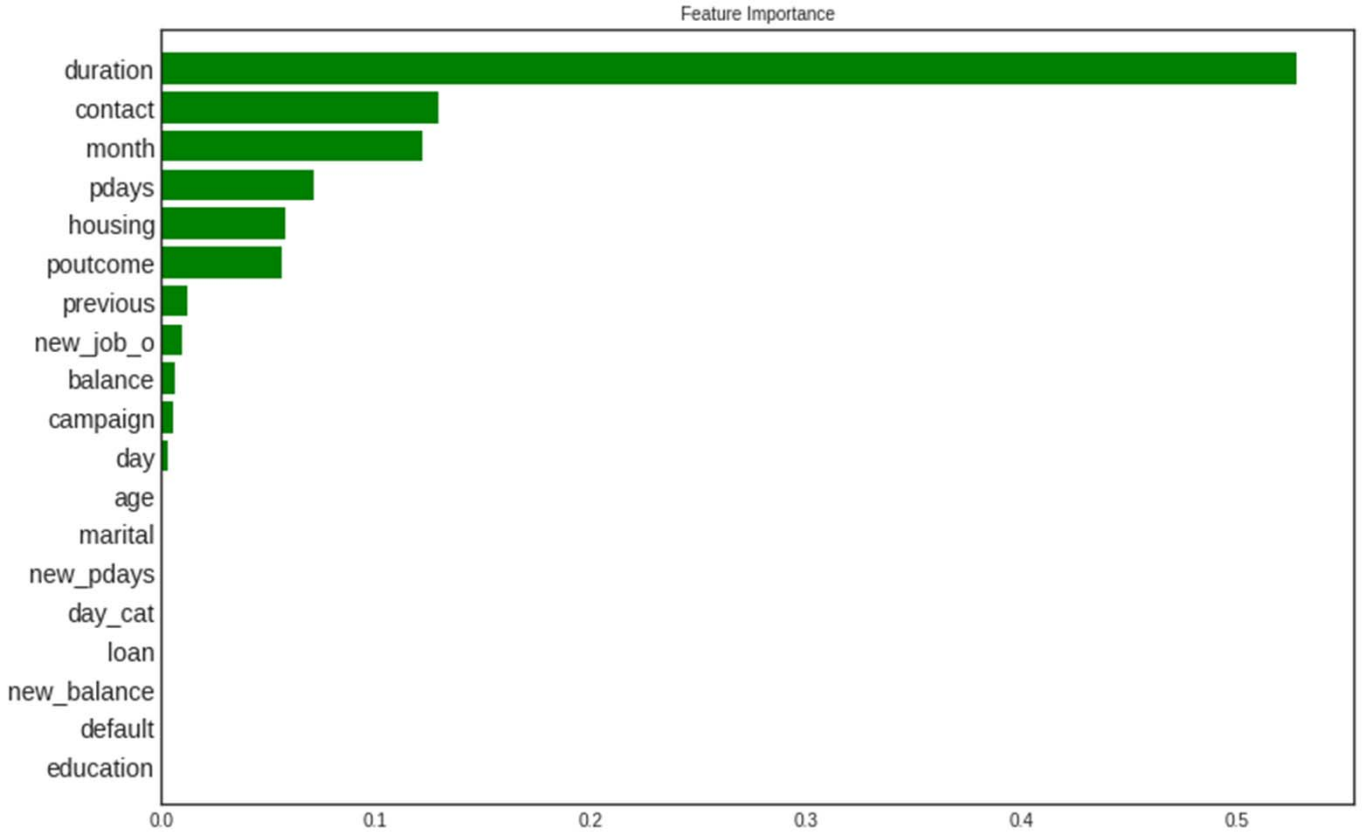


Featuring Engineering

- **As per Exploratory Data Analysis EDA**
- - no missing value found
- - no feature found with one value
- - 9 categorical features
- - not found any significant linear relationship amongst predictors
- - it seems some outliers found (age, balance, duration, campaign, pdays and previous has some
- outliers)

- **Reducing JOB categories**
- - admin+services = adms
- - enterpenure + selfemployed + unemployed + unknown + housemade = others
- - retired + student = rstd

Feature importance for feature selection



Inference:

Using decision tree model we are getting these important features.
Better to go with important features only to make model robust

Handling Imbalance data Using SMOTE

```
Before OverSampling, counts: Features (31296, 29) and Label (31296, 1)
yes
0      27624
1       3672
dtype: int64
yes
0      11839
1       1574
dtype: int64
After OverSampling, counts: Features (55248, 29) and Label (55248, 1)
yes
0      27624
1      27624
dtype: int64
```

- SMOTE library used to generate synthetic data for handling the imbalance in the dataset
- Purpose: to minimize the bias of model towards the class which has higher percent count in our dataset.

Model building

With duration columns:

1. KNN(K Nearest Neighbours)
2. Random Forest
3. Light GBM

Without duration columns:

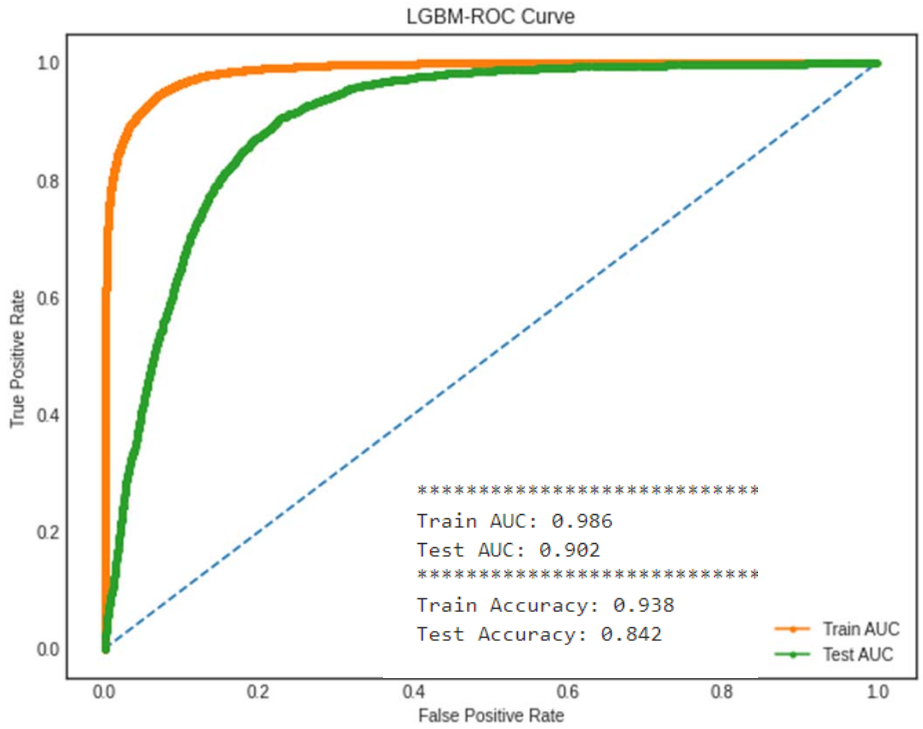
1. Light GBM
2. ANN(Artificial Neural Nets)

**** Comparison of different models for Class 1(Yes) oversampled train data ****

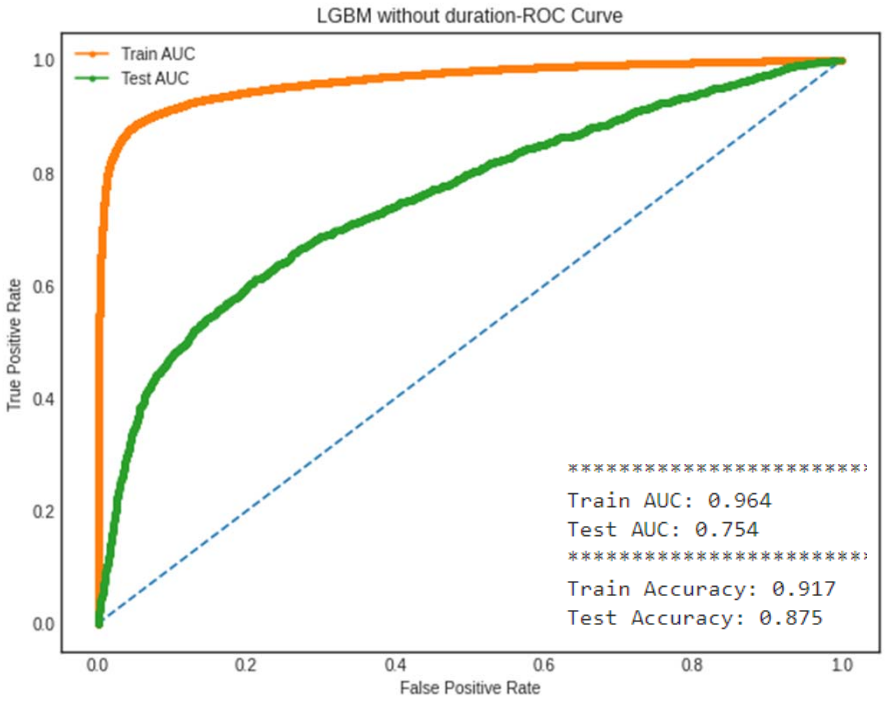
Model	Test Accuracy	Precision	Recall	F1_score
KNN_with_Duration	0.8611	0.44	0.62	0.51
Random Forest_with_Duration	0.7828	0.33	0.85	0.48
LGBM_with_Duration	0.8417	0.41	0.81	0.55
LGBM_without_Duration	0.8744	0.46	0.39	0.42
ANN_without_Duration	0.8777	0.47	0.37	0.41

AUC ROC plot for both selected models

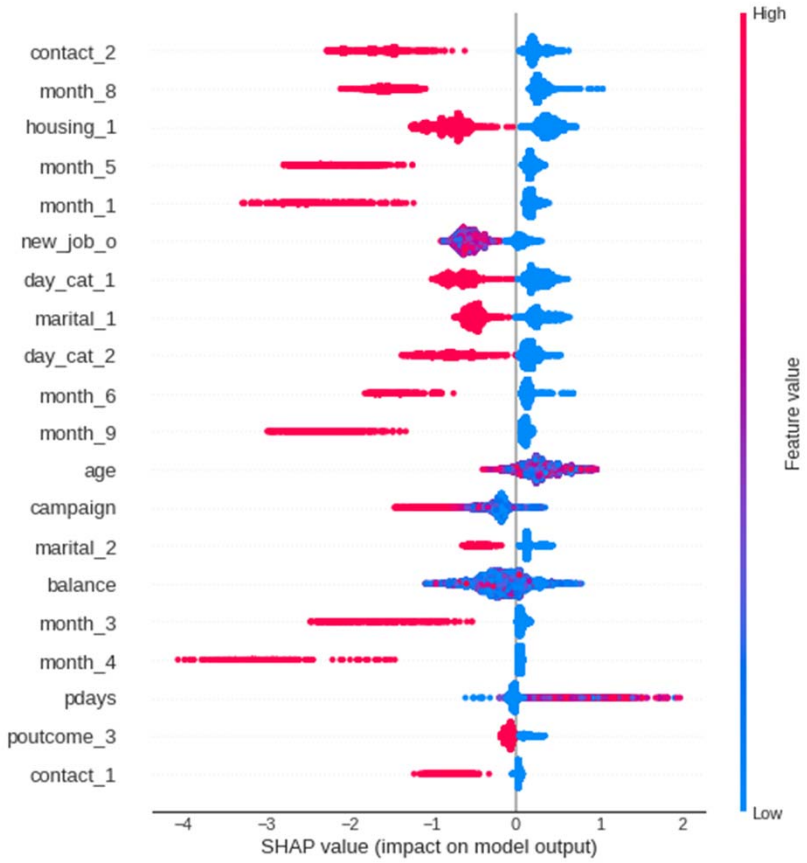
With duration LightGBM

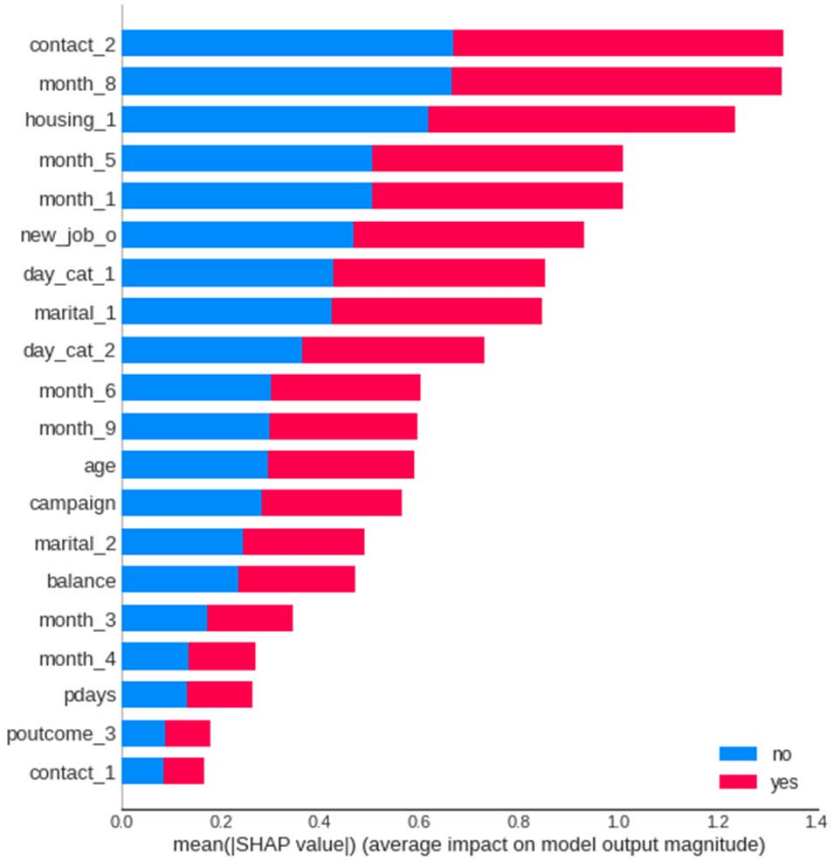


Without duration LightGBM



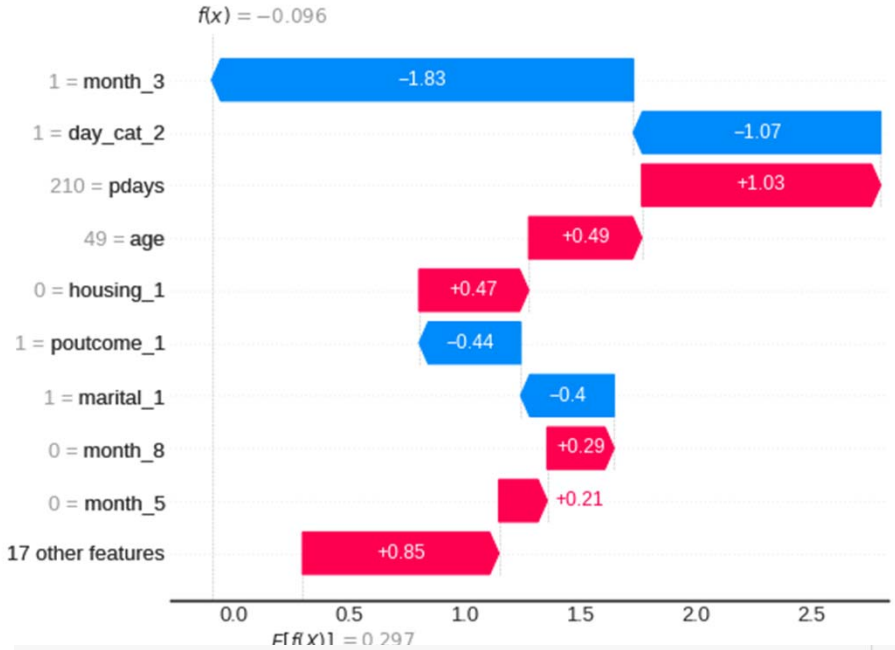
SHAP Summary Plot





SHAP Individual(record explaination) Plot

Water Plot



Force Plot

Conclusion:

Thus we come to an end of our analysis and model building to predict if the client will subscribe a term deposit or not. The most important takeaways are:

- **If the clients are contacted twice in first 10 days of the month, then it is more likely that call will be converted to subscription.**
- **Clients take more subscriptions in a month of January, May or August.**
- **If clients are already paying housing loans, then it is more likely that they will opt for term deposit.**

In our dataset we were provided with call 'duration' but in real world practise that won't be the case. So, to build more realistic model **we trained our models for both the cases i.e with or without 'duration'**. For both the cases **LGBM showed best scores of recall with 'duration' and precision without 'duration'.**

Future Scope-

Our main objective is to get good precision score for without 'duration' models and good recall score for 'duration' included model.
So, we can initially formulate the required time to converge a lead using 'duration' included models and then sort out precise leads for 'duration' excluded models using this formulated time.
Here, the idea is to find out responses for any particular record with varying assumed predefined duration range.
For example let's say, to converge a call, duration ranges between 60 to 2000 sec, then using this range we can predict all responses for each lead while iterating through this duration range. If we get positive response for any value of 'duration' we can assign that duration time to that particular lead.
In this way we can help marketing team to get precise leads along with time required to converge that lead and also, those leads that have least proability to converge (if we get no positive response for any assumed duration). Thus, an effective marketing campaign can be executed with maximum leads converging to term deposit.

Thank you