Capstone Project –3

Bank Marketing Effectiveness Prediction

Team Members

Annayan Bose

Prabhat Patel

Content

The following approach was followed in the completion of the project:

- **Business Problem**
- **Data Collection and Preprocessing**
- **Exploratory Data Analysis**
  - Categorical Features
  - Continuous Features
- **Data Manipulation**
  - Outlier Detection and Treatment
  - Feature Engineering and Feature Selection
  - Categorical Data Encoding
- **Modeling**
  - Train Test Split
  - Feature Scaling
  - Logistic Regression
  - KNN
  - Naïve Bayes
  - Random Forest
  - XGBoost
- **Conclusion and Recommendations**

Business Problem

- **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

We have one dataset. Bank dataset consists of 45211 observations and 17 features. There are no null values in the dataset. Below is the breakdown of the features:

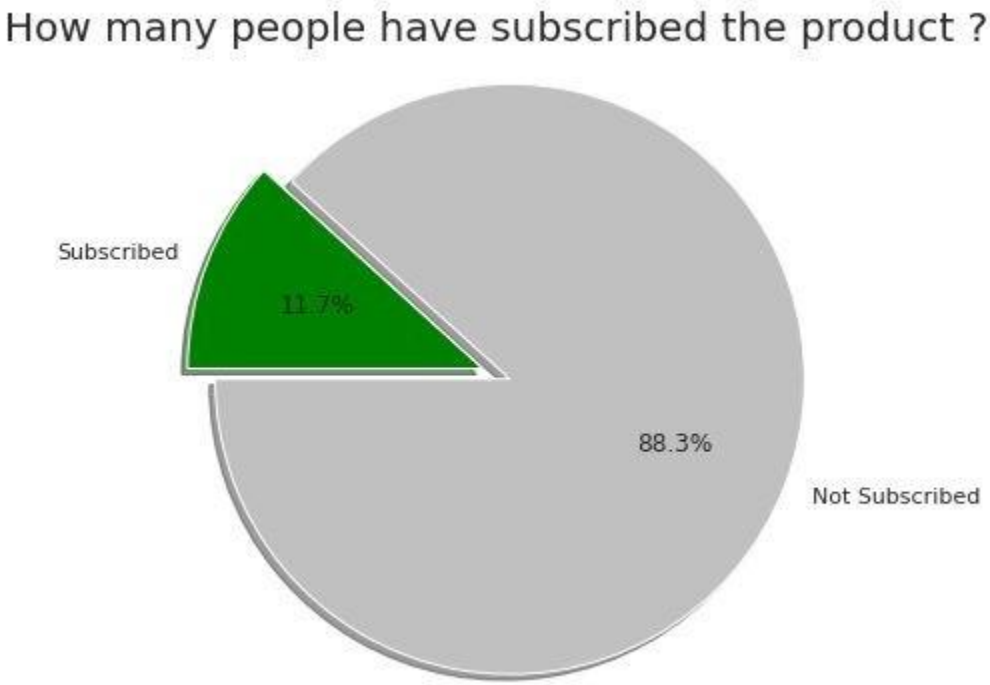
Categorical Features

- 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
- Previous

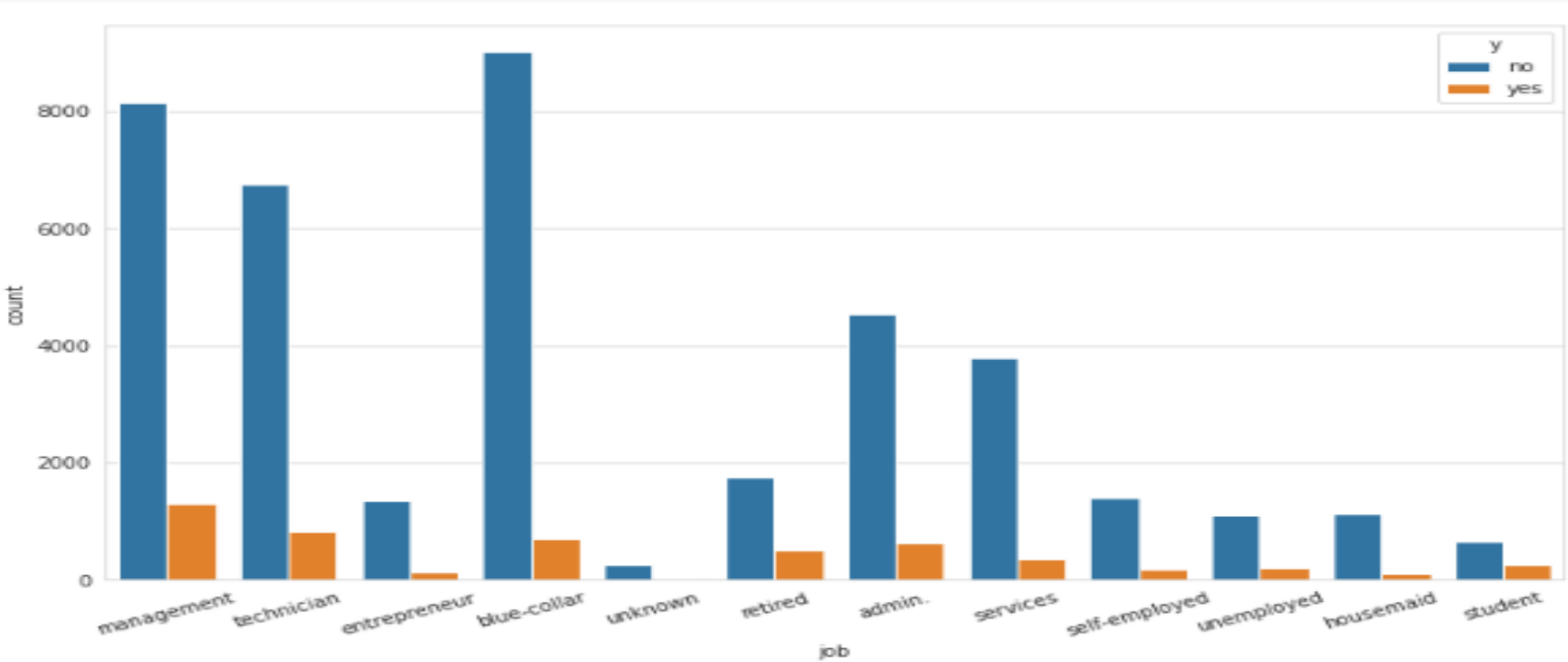
Exploratory Data Analysis



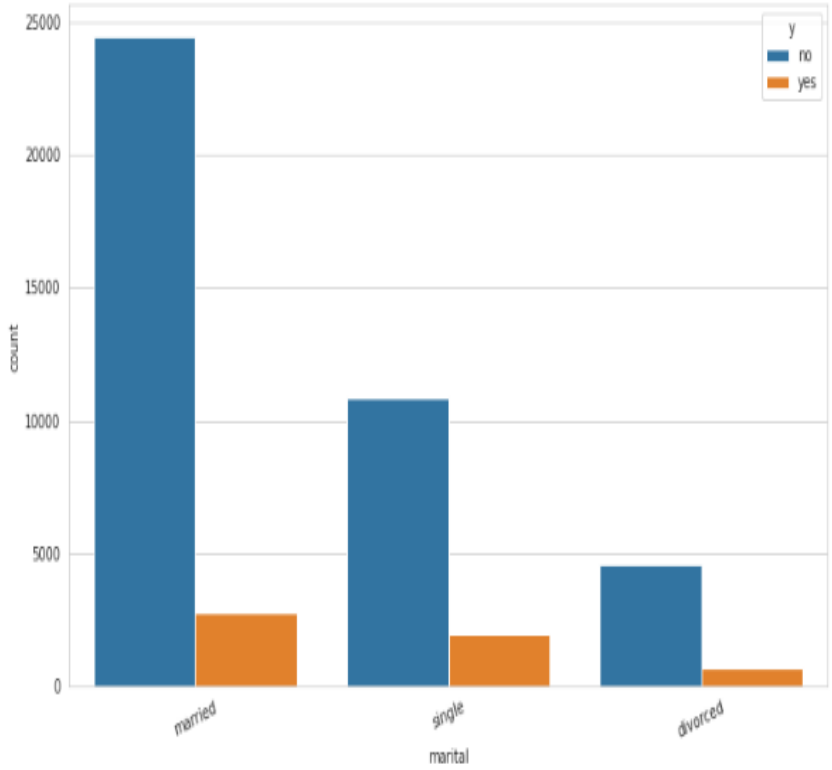
- The target variable ‘y’ tells us the outcome of the campaign whether they went ahead for the term deposit or not.
- Out of 45211 only 5289 people subscribed to the term deposit.

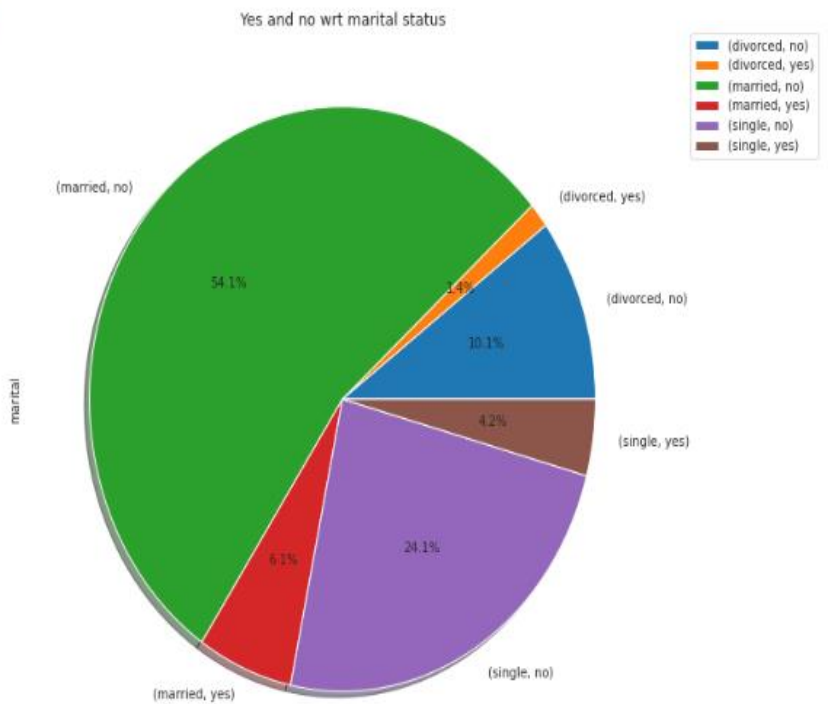
EDA For Categorical Features

Job types v/s target variable - We can say that Retired, Technician, Management, Admin and Blue collar Job person are most likely to opt for the product. Most of the people targeted in our dataset are Management, blue collar, technician, admin and services people. Retired people should be targeted more as they are most likely to opt for the product.

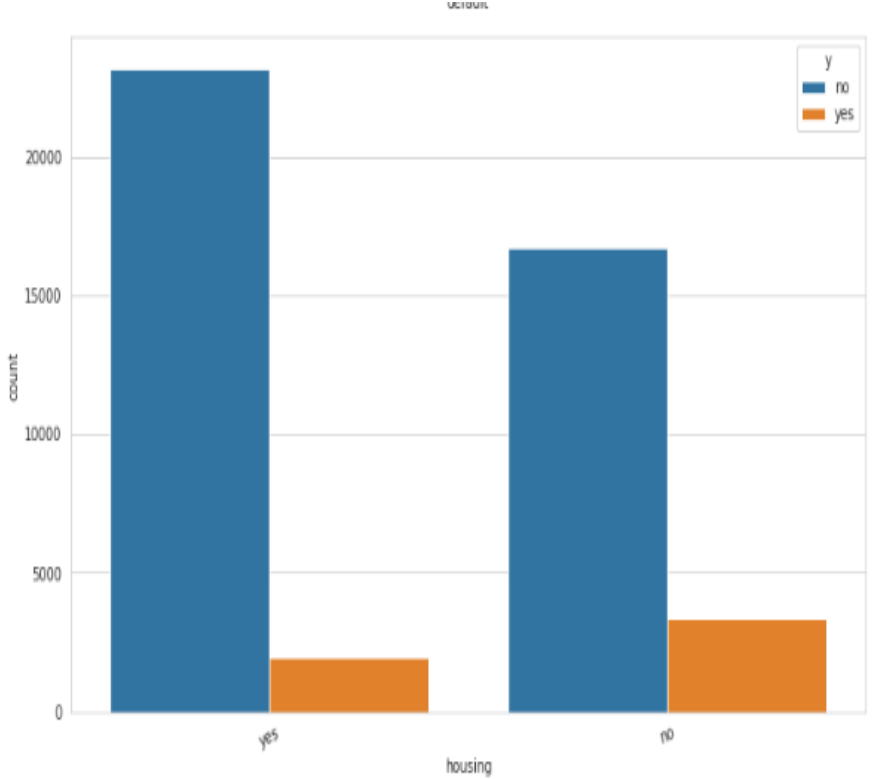


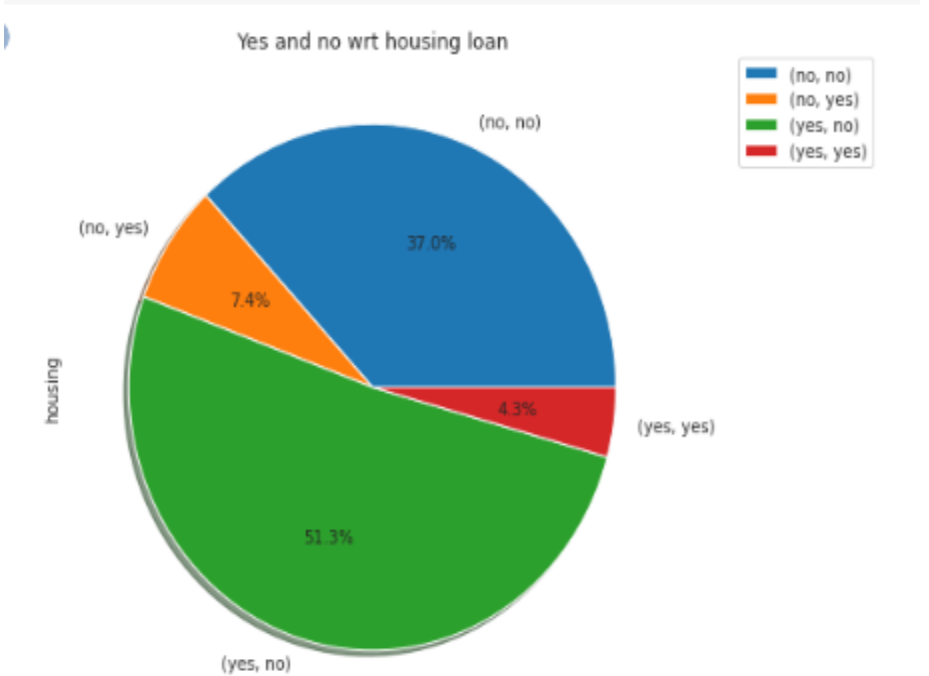
How marital status of a person effects target variable ‘y’ (Subscribed to term deposit yes/no)? -
Married people tend to take term deposit slightly more than the singles.



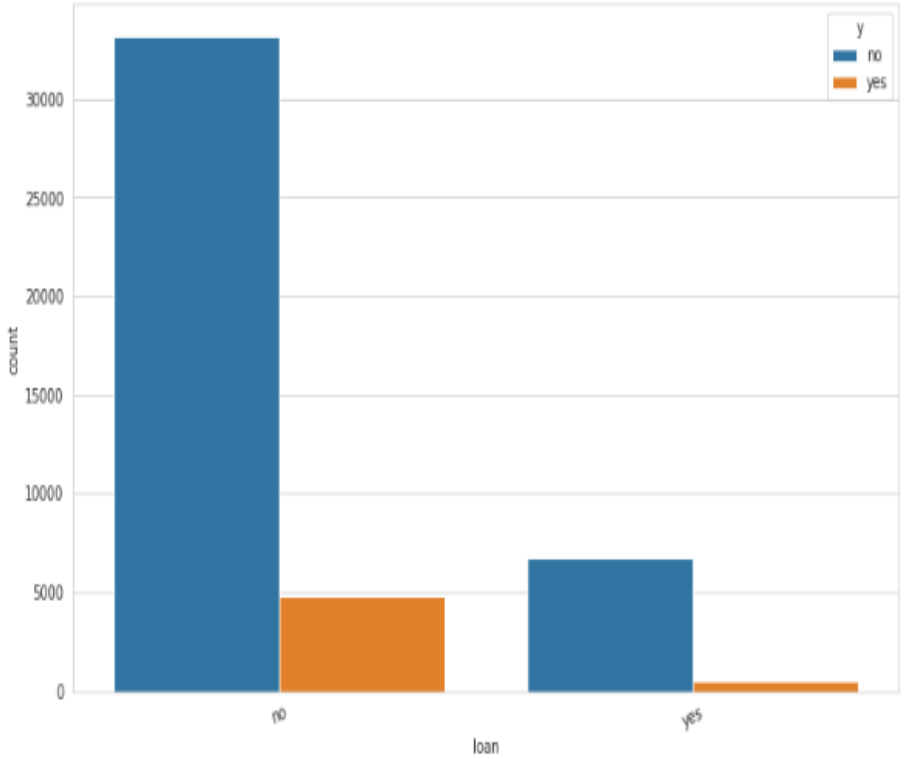


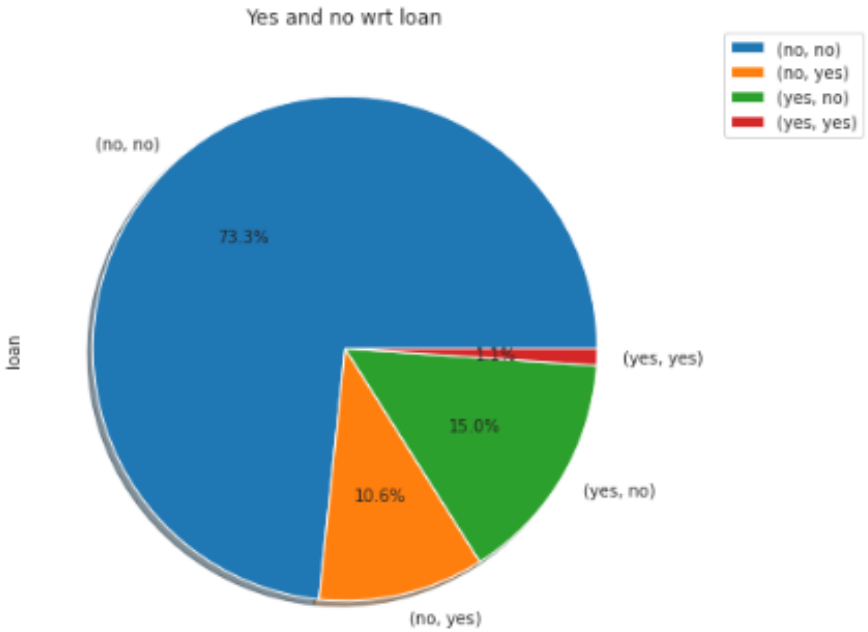
Effect of housing loan on target variable 'y' (Subscribed to term deposit yes/no)? People having housing loan are less likely to opt for term deposit plan as their major chunk of income is consumed by loan.



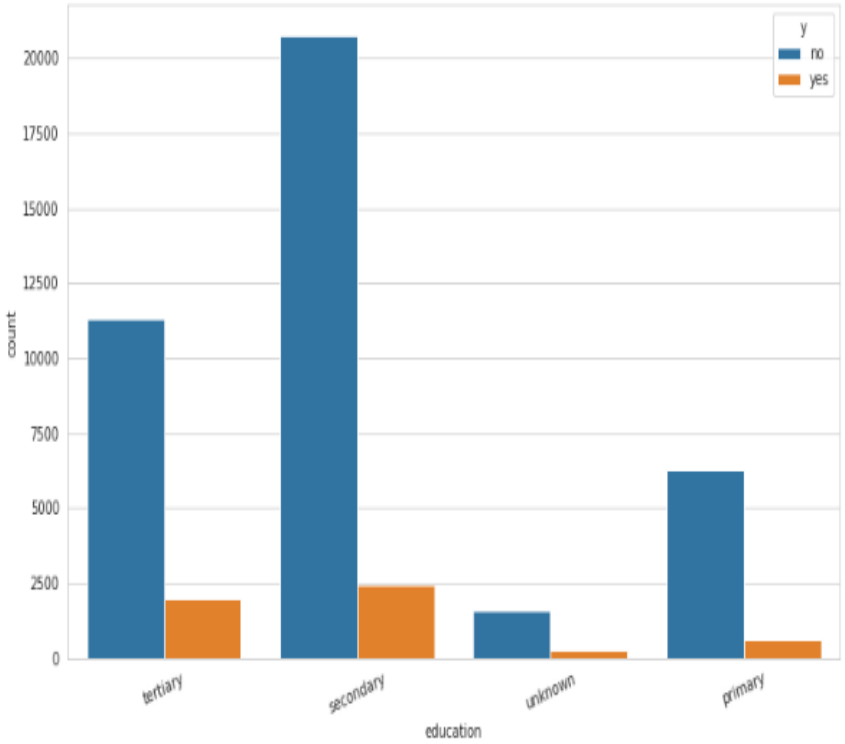


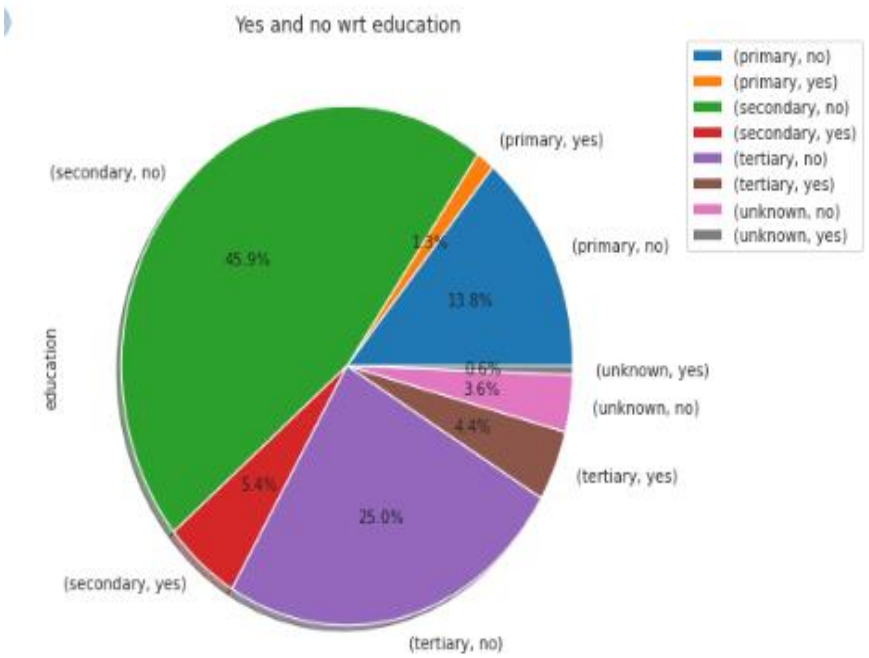
Effect of personal loan on target variable 'y' (Subscribed to term deposit yes/no)? People having no personal loans are opting more for term deposit.



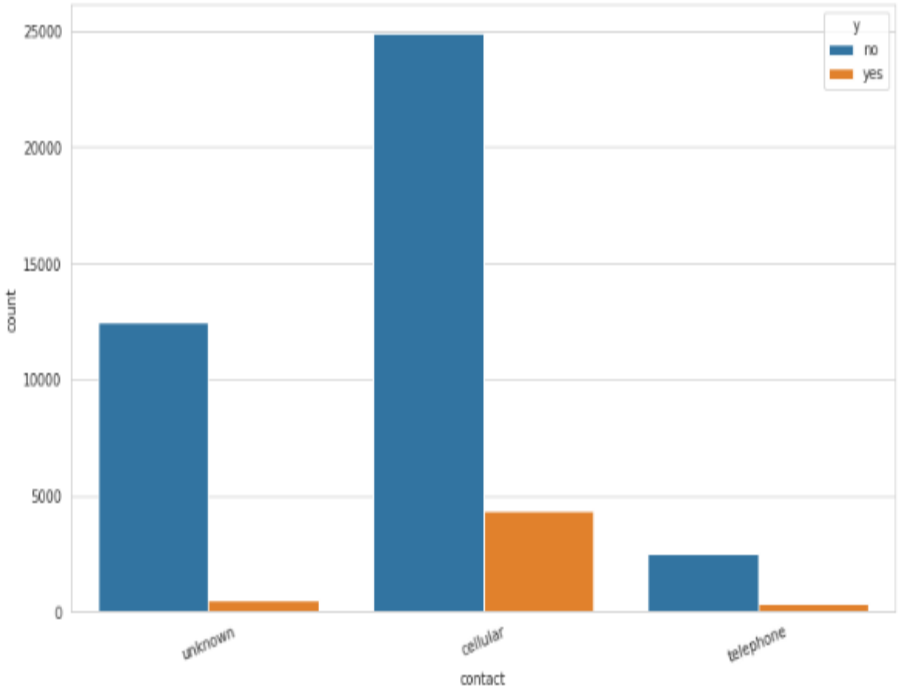


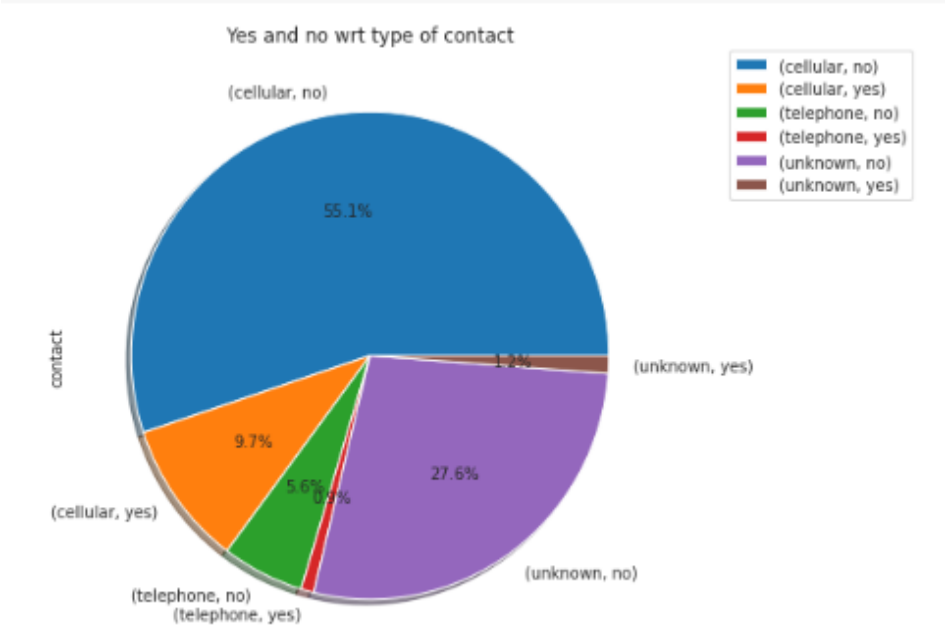
Education v/s target variable 'y'. People with secondary education are more interested in taking term deposit than those who have tertiary level of education but people with tertiary education are good prospect as they are more inclined towards term deposit.



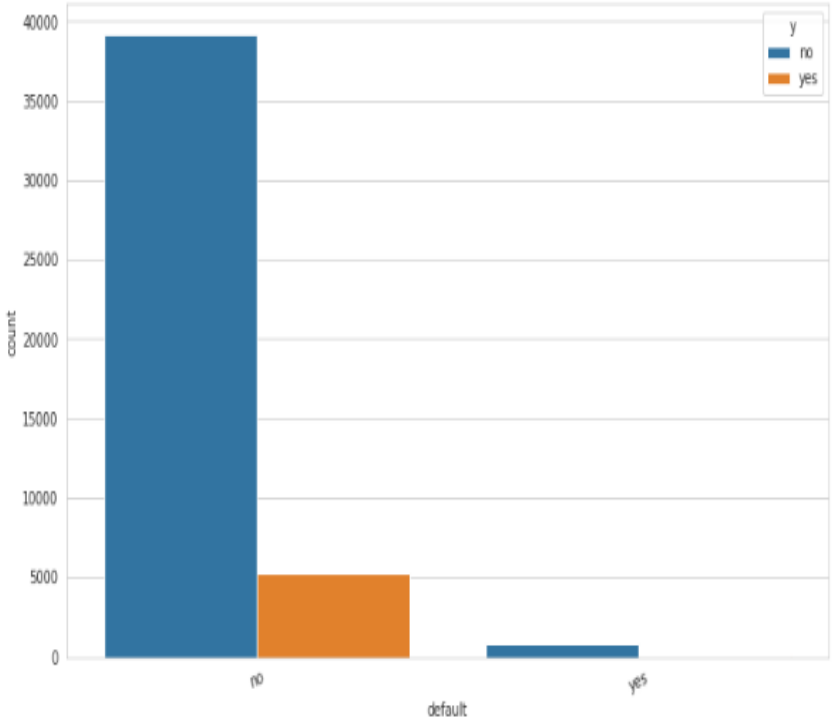


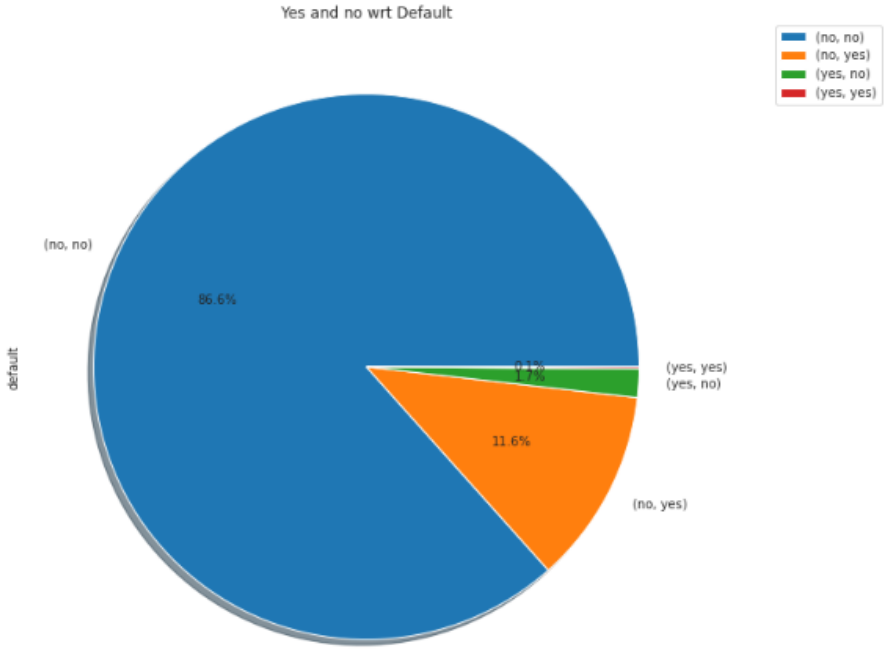
Majority of people who were taking term deposit plan were contacted via cellular mode.



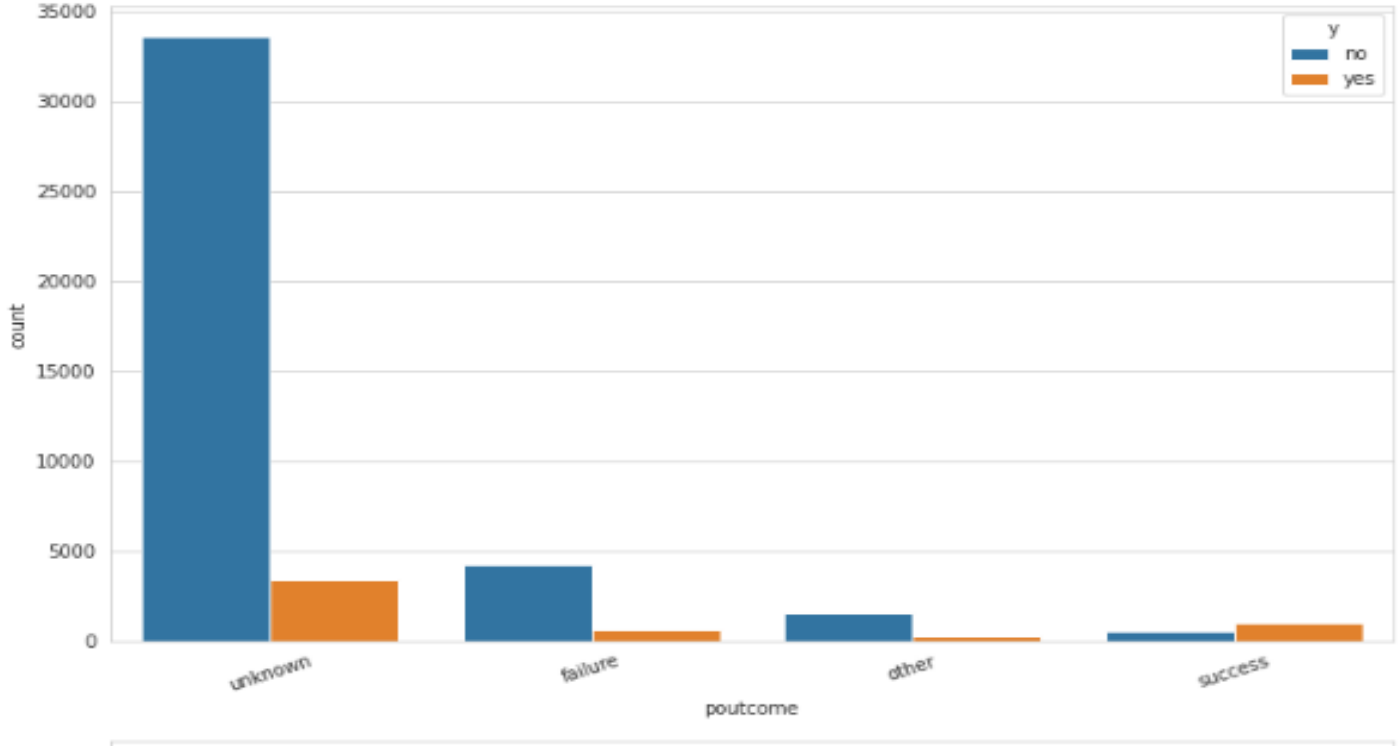


Number of people having no credit in default took more term deposit then those who has credit in default.

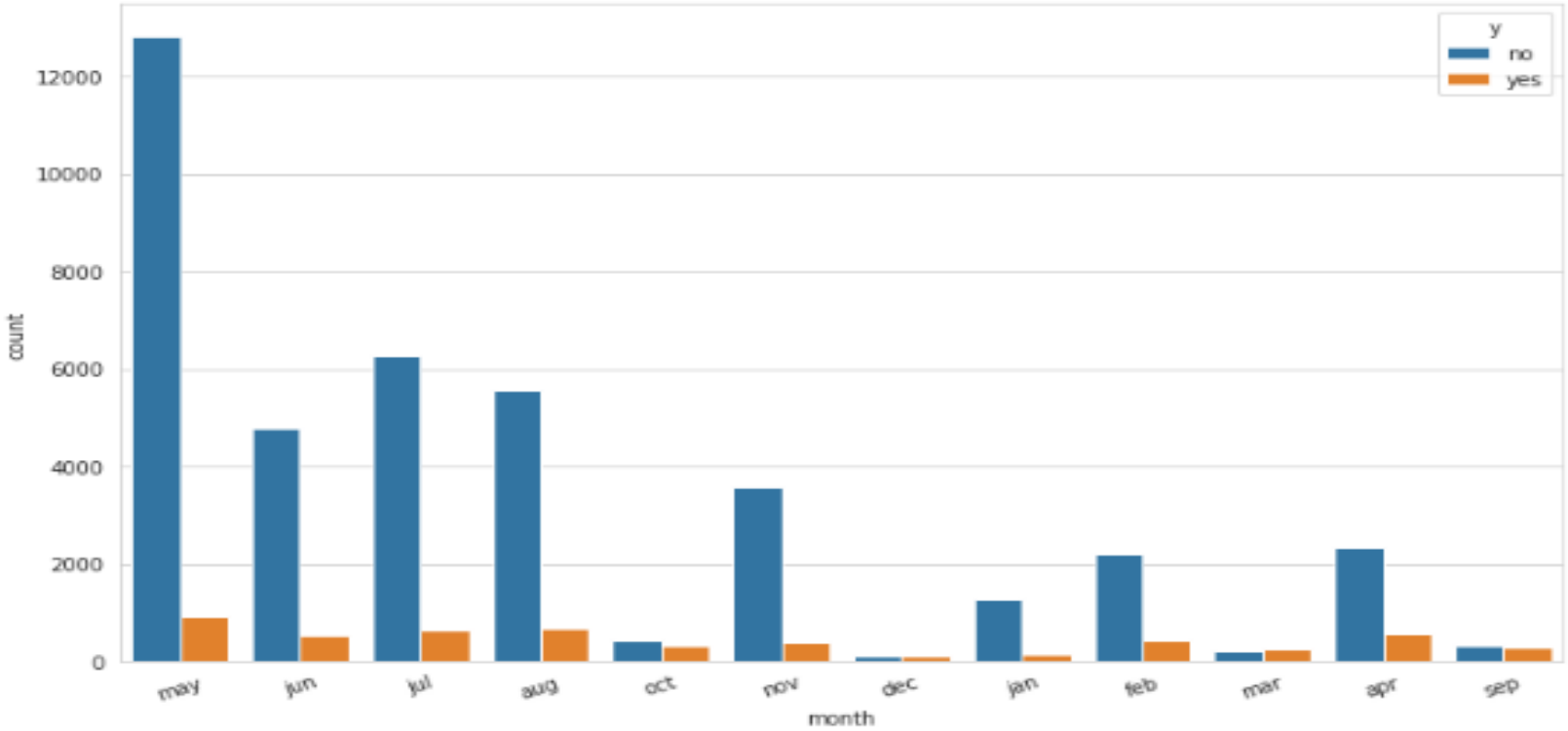




The previous outcome in most of the cases are unknown in given dataset. Outcome of the previous marketing campaign shows that people who already had success with the bank for earlier product are more prone to opt for term deposit.

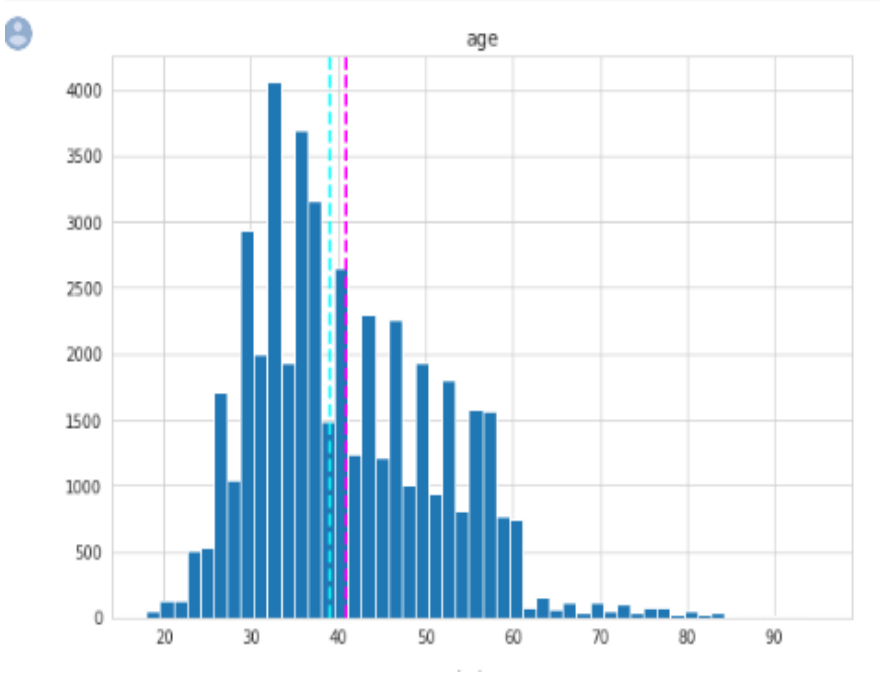


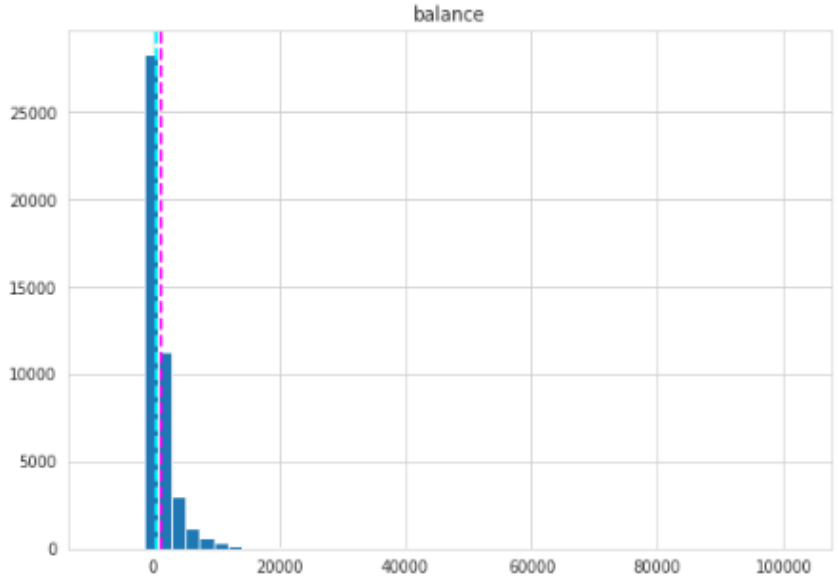
The campaign is more focused on the month of May, June, July, August and Nov. The campaign was more aggressive during 2nd quarter of the year especially in May. March is the best month as it is the only month having higher acceptance rate than rejection.



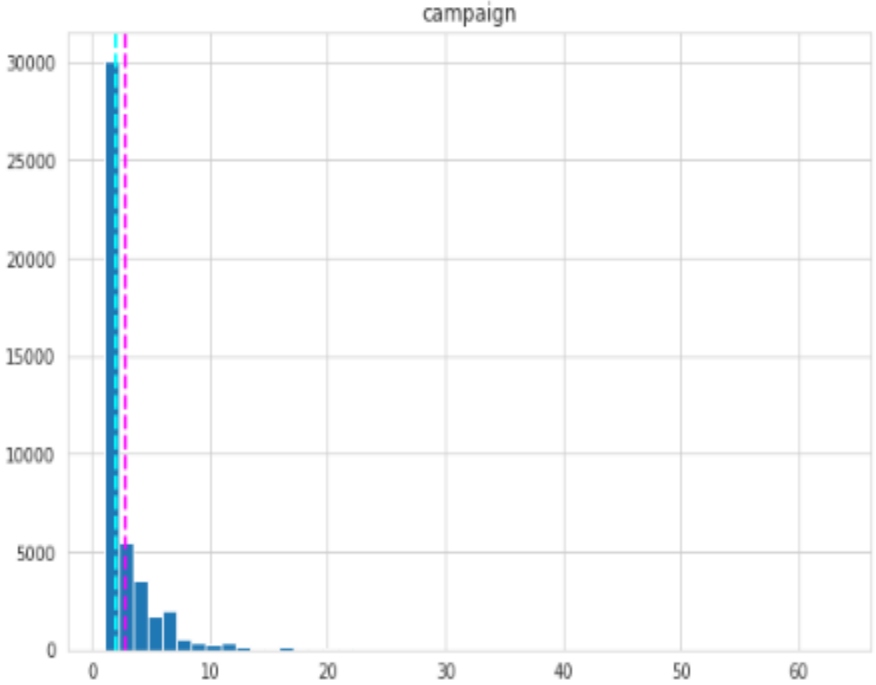
Numerical Features

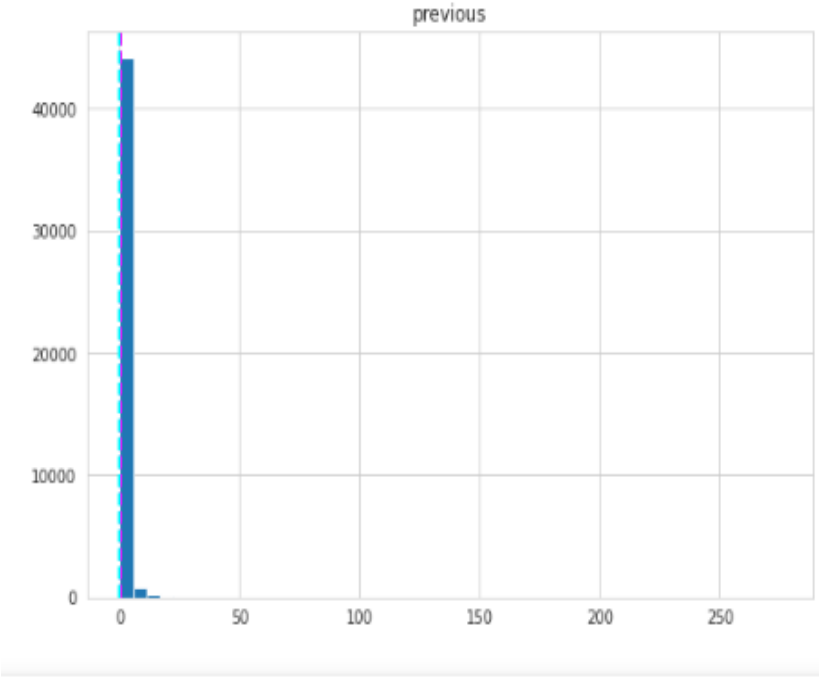
- We have more number of people between falling between the range of 30 and 50.
- Maximum number of people have balance less than 3000 in their account very few people have balance greater than 5000.





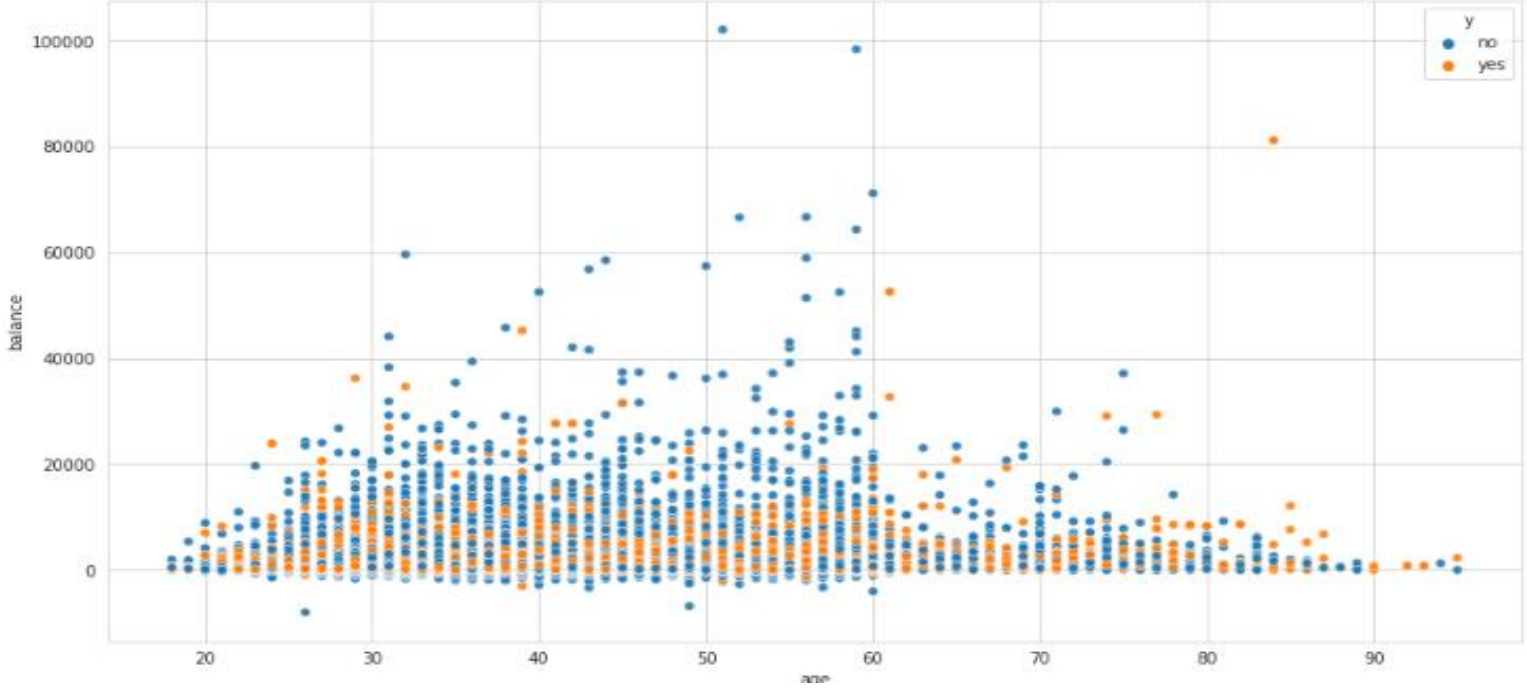
- Majority of people were contacted less than 5 times during this campaign.
- Majority of people were contacted less than 4 times previous to this campaign.





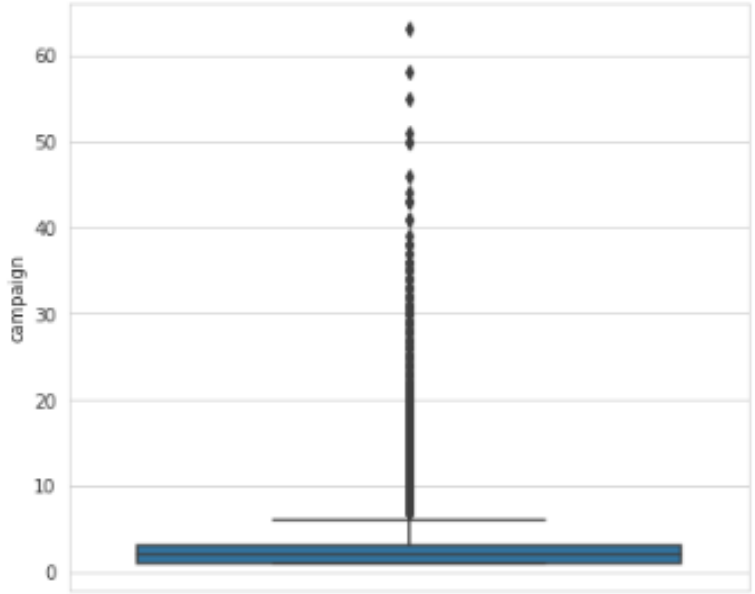
Visualizing age and balance with respect to target variable

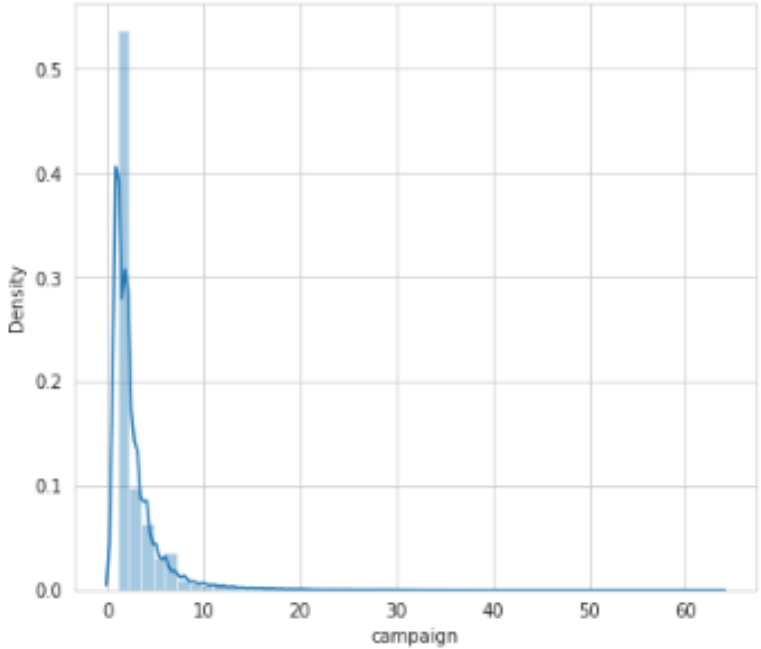
From below scatter plot we can see that 40-60 age group people have subscribed more to the product. People having high balance have very low subscription rate. This also shows that our dataset have some balance below 0 and we can use this scatter plot to remove some outliers from our data set.



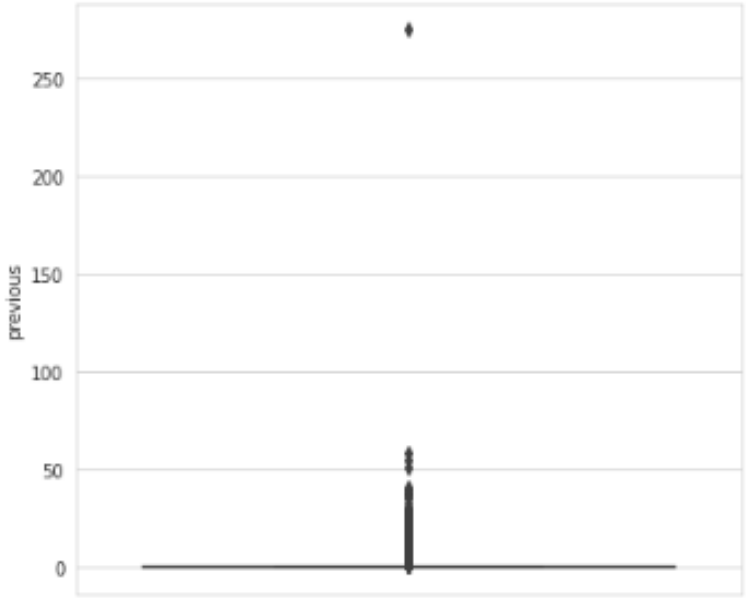
Outlier Detection

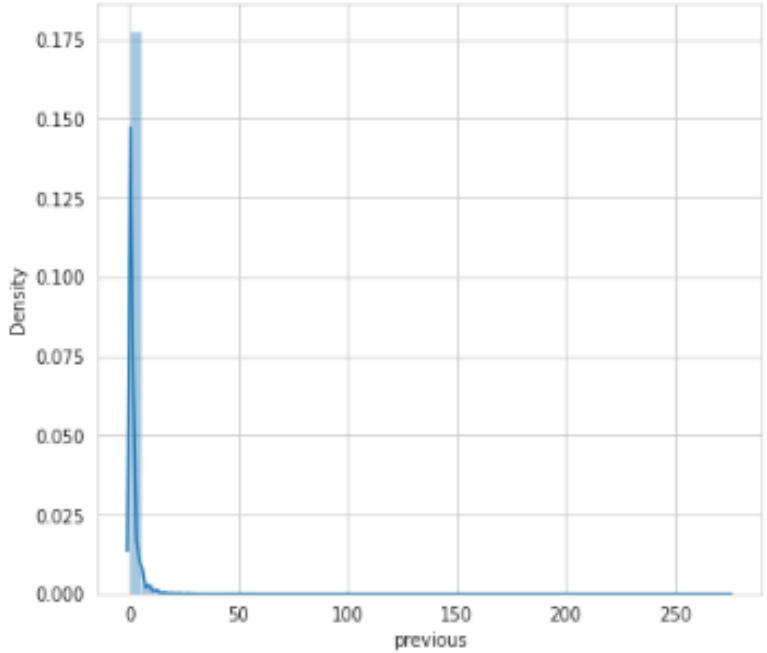
Campaign has a lot of outliers so we'll treat the outliers by capping them between the quantile range of 95% and 5%.





Previous also has a lot of outliers so we'll treat the outliers by capping them between the quantile range of 95% and 5%.





Feature Engineering

- We have dropped the 'duration' variable as the value of the variable will only be known at the end of the call. Hence, at that time we will also know the outcome of the call. The 'duration' variable will lead to leakage in the data and the prediction model will not be realistic.

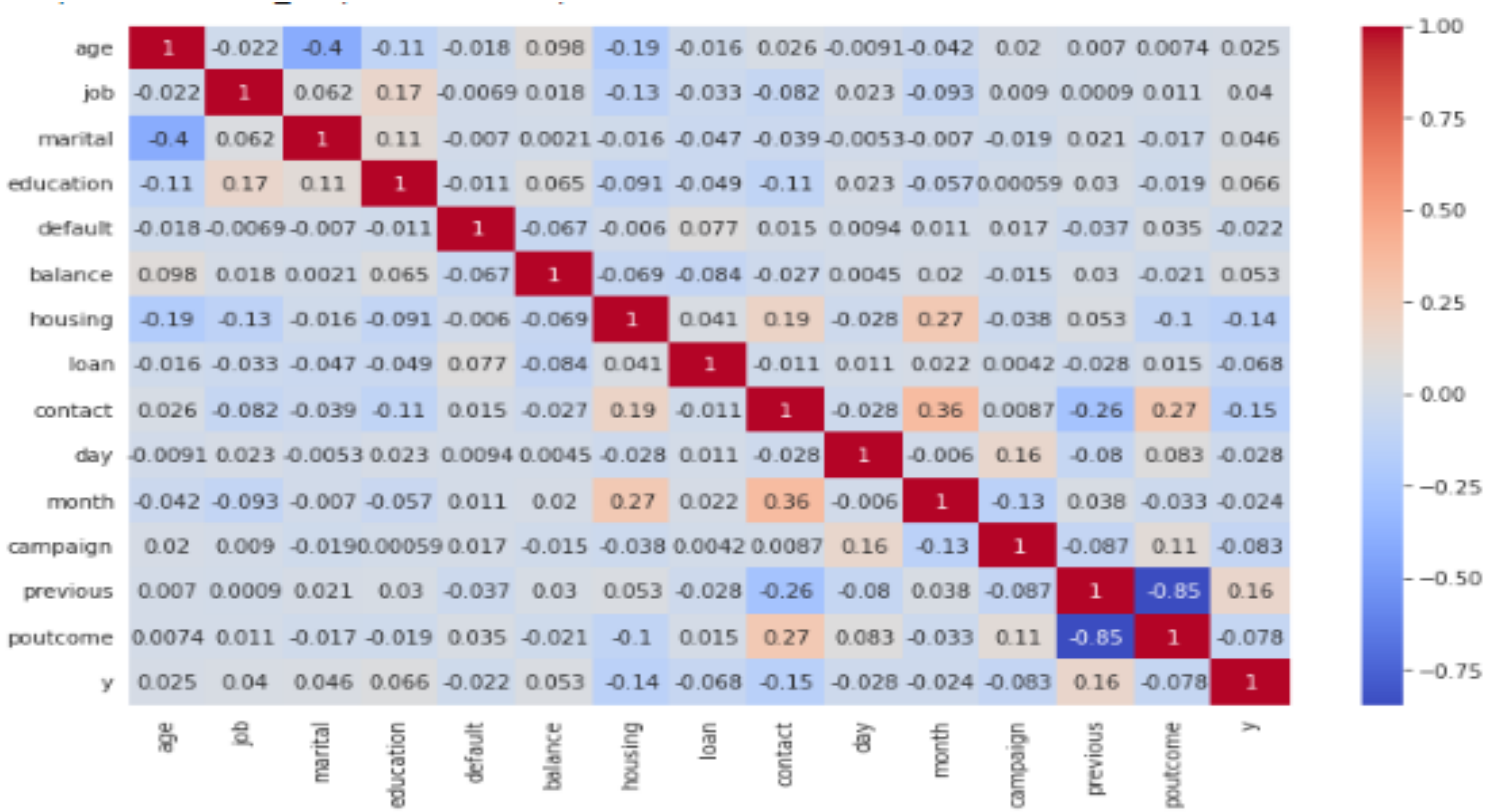
- Also, pdays has more than 80% of its value as -1 which possibly means that the client wasn't contacted before or stands for missing data. So we have dropped pdays too.

Label Encoding

- Columns such as default, loan, housing and y have Boolean values (yes/no) so we have encoded yes as 1 and no as 0.

- All other categorical columns have been numerically encoded.

Heatmap


age 1 -0.022 -0.4 -0.11 -0.018 0.098 -0.19 -0.016 0.026 -0.0091 -0.042 0.02 0.007 0.0074 0.025
job -0.022 1 0.062 0.17 -0.0069 0.018 -0.13 -0.033 -0.082 0.023 -0.093 0.009 0.0009 0.011 0.04
marital -0.4 0.062 1 0.11 -0.007 0.0021 -0.016 -0.047 -0.039 -0.0053 -0.007 -0.019 0.021 -0.017 0.046
education -0.11 0.17 0.11 1 -0.011 0.065 -0.091 -0.049 -0.11 0.023 -0.057 0.00059 0.03 -0.019 0.066
default -0.018 -0.0069 -0.007 -0.011 1 -0.067 -0.006 0.077 0.015 0.0094 0.011 0.017 -0.037 0.035 -0.022
balance 0.098 0.018 0.0021 0.065 -0.067 1 -0.069 -0.084 -0.027 0.0045 0.02 -0.015 0.03 -0.021 0.053
housing -0.19 -0.13 -0.016 -0.091 -0.006 -0.069 1 0.041 0.19 -0.028 0.27 -0.038 0.053 -0.1 -0.14
loan -0.016 -0.033 -0.047 -0.049 0.077 -0.084 0.041 1 -0.011 0.011 0.022 0.0042 -0.028 0.015 -0.068
contact 0.026 -0.082 -0.039 -0.11 0.015 -0.027 0.19 -0.011 1 -0.028 0.36 0.0087 -0.26 0.27 -0.15
day -0.0091 0.023 -0.0053 0.023 0.0094 0.0045 -0.028 0.011 -0.028 1 -0.006 0.16 -0.08 0.083 -0.028
month -0.042 -0.093 -0.007 -0.057 0.011 0.02 0.27 0.022 0.36 -0.006 1 -0.13 0.038 -0.033 -0.024
campaign 0.02 0.009 -0.019 0.00059 0.017 -0.015 -0.038 0.0042 0.0087 0.16 -0.13 1 -0.087 0.11 -0.083
previous 0.007 0.0009 0.021 0.03 -0.037 0.03 0.053 -0.028 -0.26 -0.08 0.038 -0.087 1 -0.85 0.16
poutcome 0.0074 0.011 -0.017 -0.019 0.035 -0.021 -0.1 0.015 0.27 0.083 -0.033 0.11 -0.85 1 -0.078
y 0.025 0.04 0.046 0.066 -0.022 0.053 -0.14 -0.068 -0.15 -0.028 -0.024 -0.083 0.16 -0.078 1
age job marital education default balance housing loan contact day month campaign previous poutcome y
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

Sampling And Scaling

- The given dataset was highly imbalanced, so to balance this we used the technique called SMOTE sampling.

- **SMOTE** - SMOTE(synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class.

- Oversampling and undersampling of classes to fix imbalanced dataset.

- Standard scaling of balanced dataset.

Model implementation

We have implemented the below models On our balanced data set:-

1) **Logistic regression**
2) **KNN**
3) **Naïve Bayes**
4) **Random Forest classifier**
5) **XGBoost**

Model Evaluation

Model	Test AUC	Test Accuracy	F1-score	Precision
Logistic Regression	0.84	0.76	0.76	0.74
KNN	0.91	0.83	0.83	0.85
Naïve Bayes	0.82	0.72	0.76	0.66
Random Forest	0.96	0.91	0.90	0.90
XGBoost	0.95	0.88	0.88	0.89

Random Forest classifier

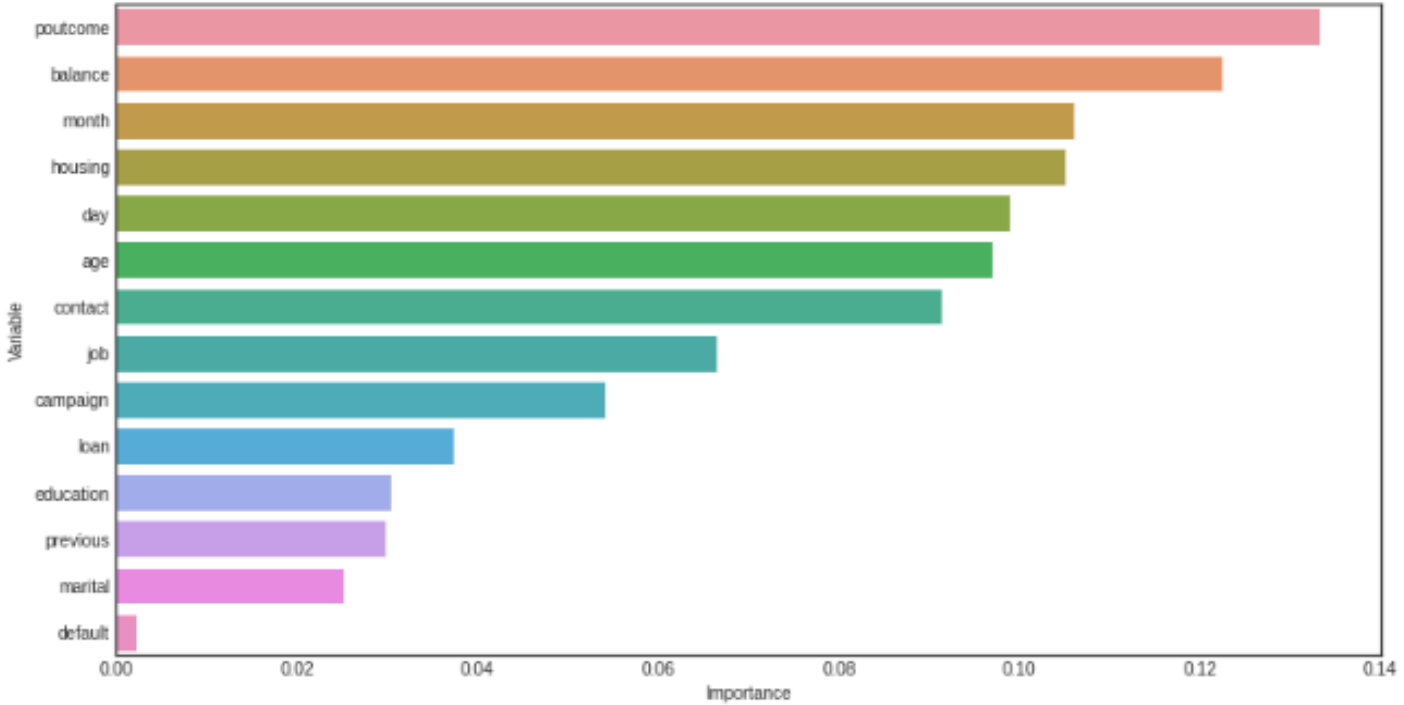
We selected Random Forest Classifier as the best model.

Having-

AUC score = 0.96
F1 score = 0.88
Accuracy = 0.90
Precision = 0.90

Here Precision is important as more number of false positive (type 1 error) can lead to poor marketing campaign because if a prospect is labelled as false positive we are completely losing him/her.

Feature Importance



Conclusion

- Random Forest and XGBoost have shown the best performance.
- The customer's account balance has a huge influence on the campaign's outcome. So we can address those customers having good account balance .
- The customer's age affects campaign outcome as well.
- Number of contacts with the customer during the campaign is also crucial.
- Outcome of previous marketing campaign also plays an important role. So we can focus on previous customers more in order to increase success of the campaign.
- Month of May have seen the highest number of clients contacted but have the least success rate. Highest success rate is observed for end month of the financial year as well as the calendar year. So we can say that our dataset have some kind of seasonality.

Thank You