Exploratory Data Analysis

Insight for Bank Marketing Campaign Data

12/8/22

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Data Glacier

Your Deep Learning Partner

Executive Summary

Client:

ABC Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc) can focus only on those customers whose chances of buying the product is more.

Objective:

Use Machine Learning to shortlist customer whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more. This will save resource and their time (which is directly involved in the cost (resource billing)).

Problem Statement

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

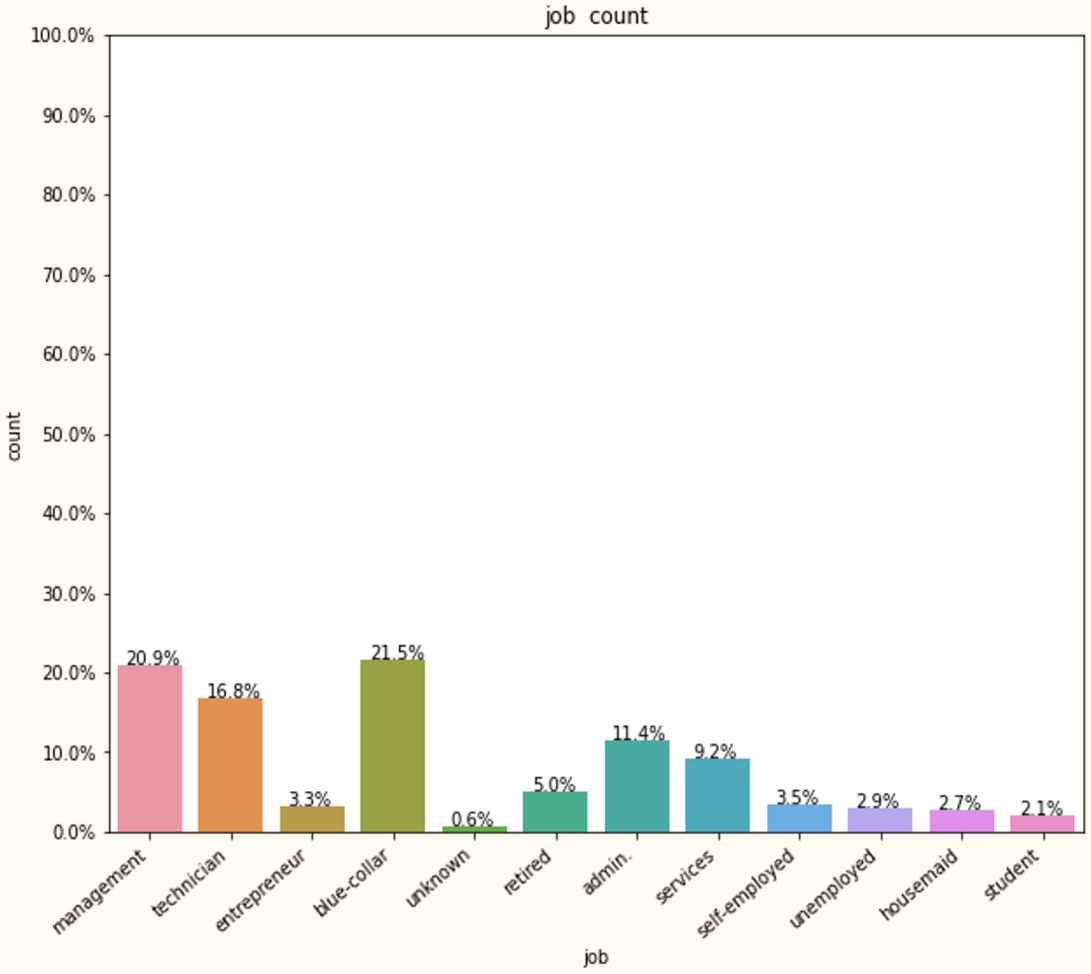
Problem Statement

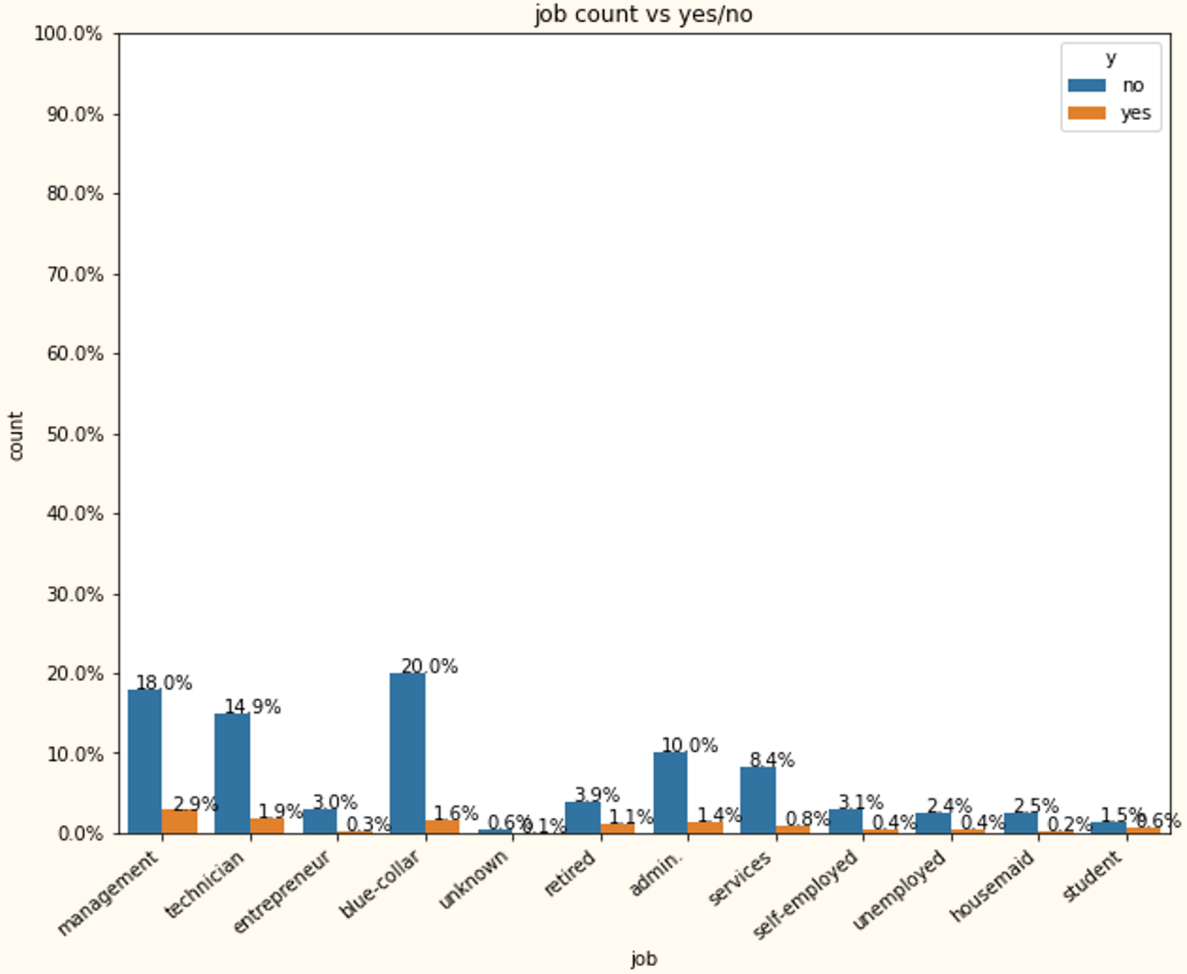
- Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.
- This will save resource and their time (which is directly involved in the cost (resource billing)).
- Develop model with Duration and without duration feature and report the performance of the model.
- Duration feature is not recommended as this will be difficult to explain the result to business and also it will be difficult for business to campaign based on duration.

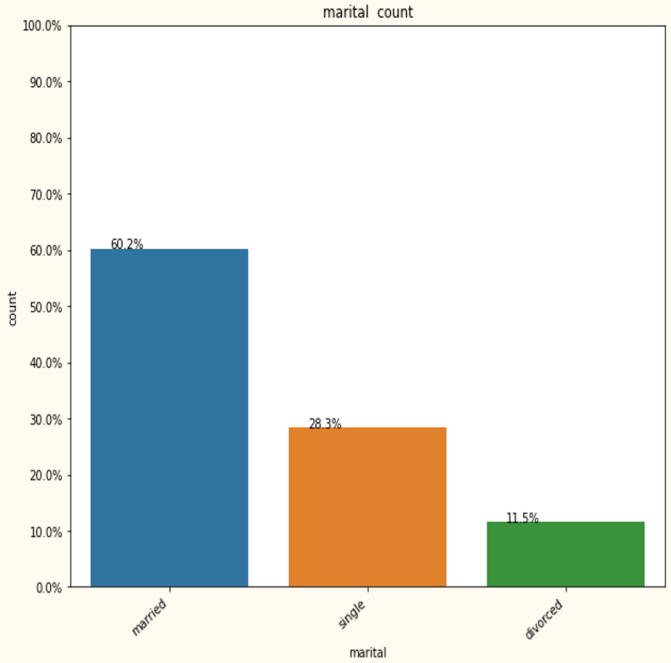
Approach

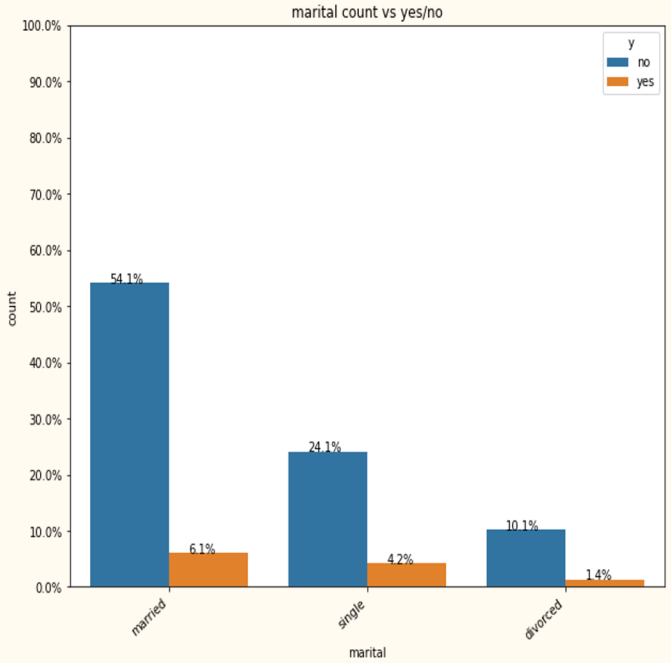
- The data set given is bank-full.csv
- It contains 16 features.
- The data had no replications
- The data had no missing values.
- The data did have many outliers.
- Since there were many outliers I considered the log of the numerical data and ways of incorporating the median in some calculations so that outliers do not affect the problem as much.

Customers with the job blue-collar were more likely to subscribe

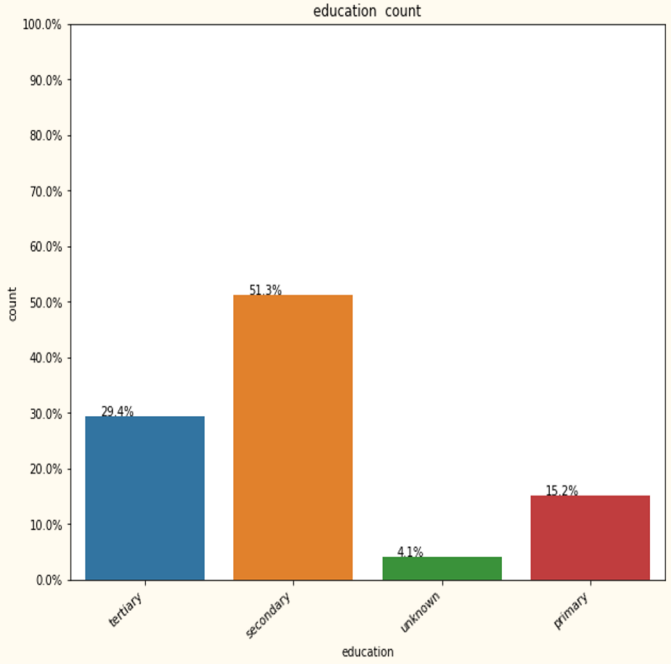


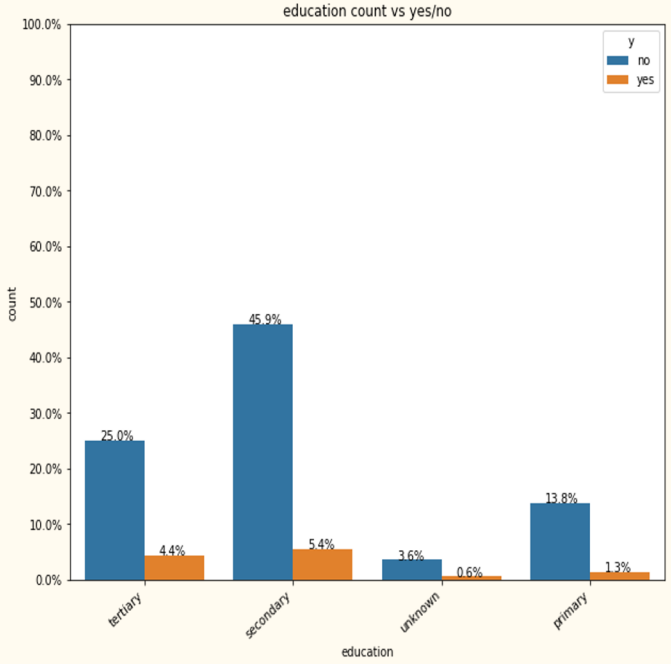




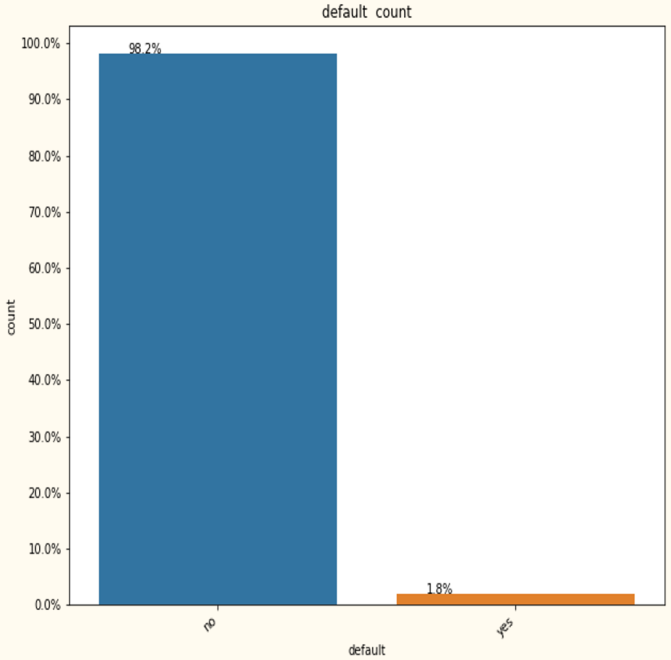


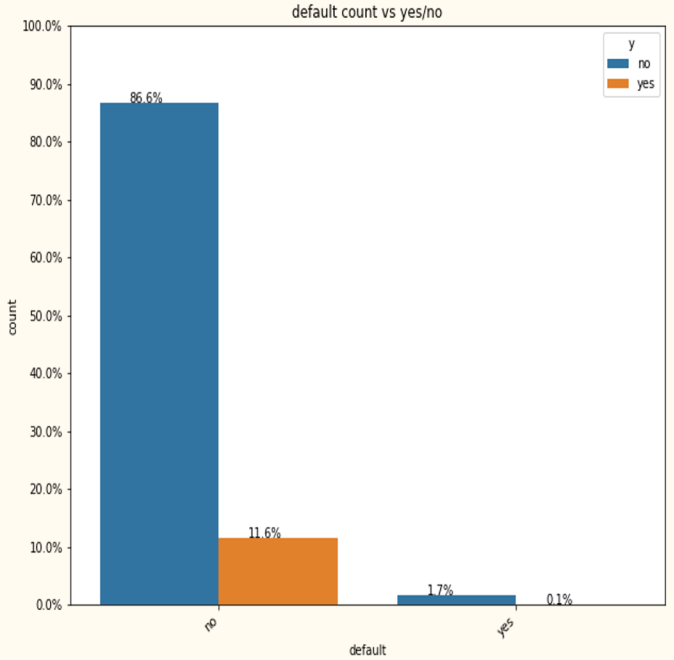
Customers with the marital status "married" are more likely to subscribe and not to subscribe. This is because most customers are married.



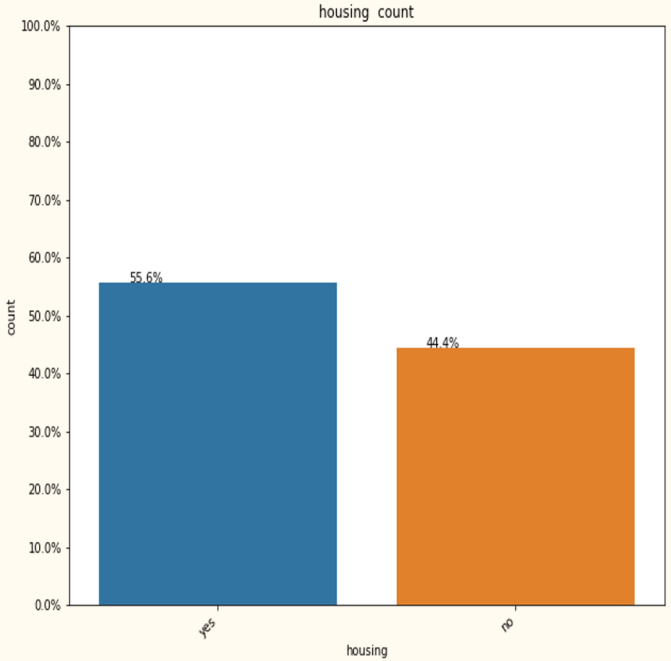


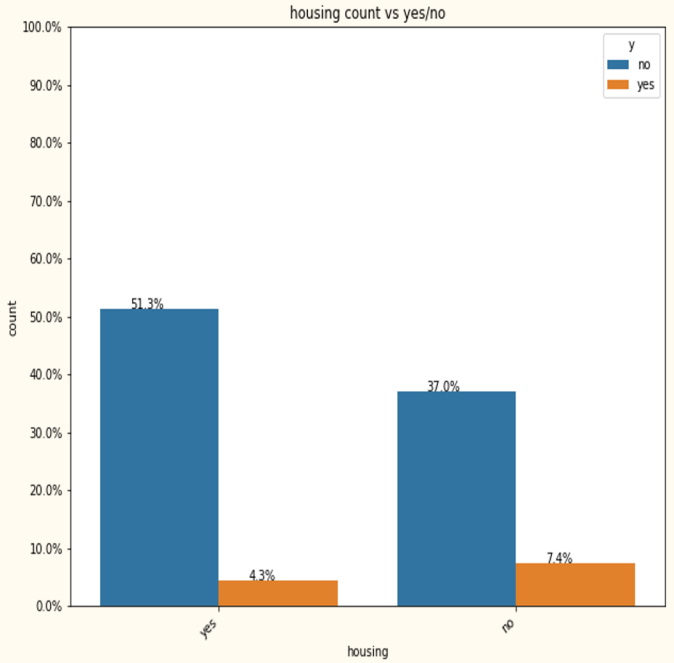
Customers with a secondary education are more likely to suscribe and not subscribe. Most customers have a secondary education.



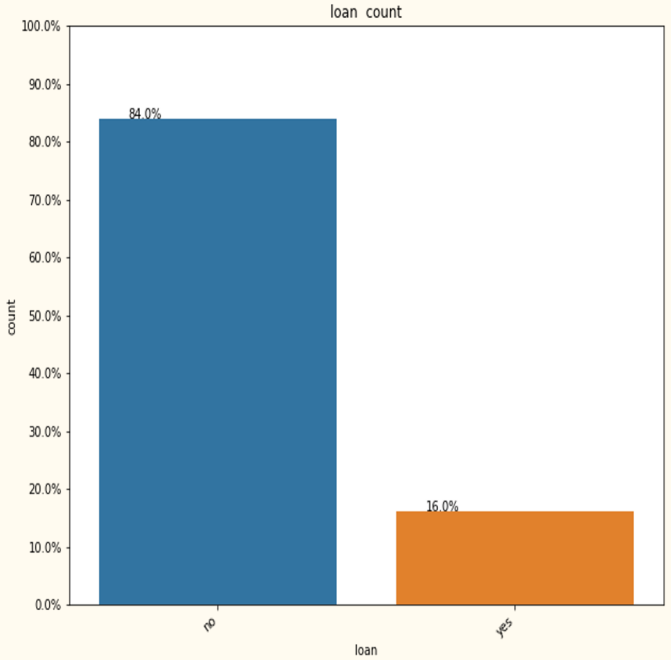


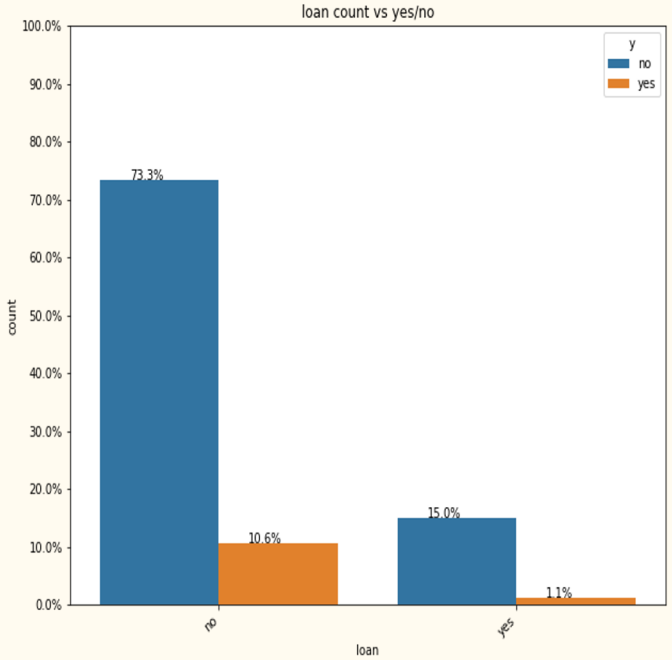
Most customers do not have defaulted credit so they are more likely to subscribe as well as not subscribe. But most of them subscribe.



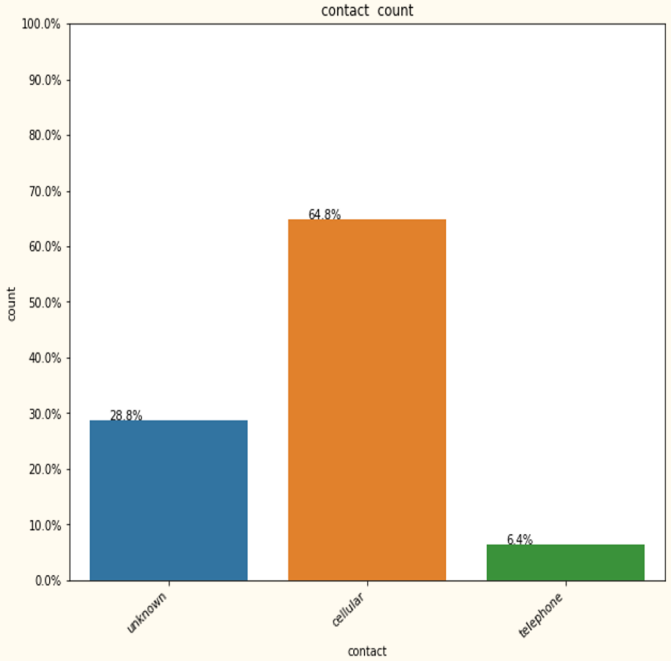


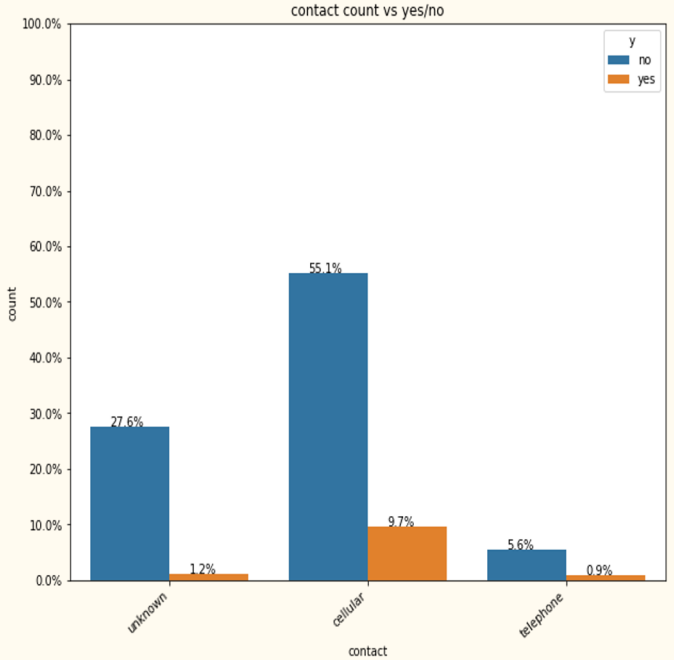
Those with a housing loan are more likely to subscribe.



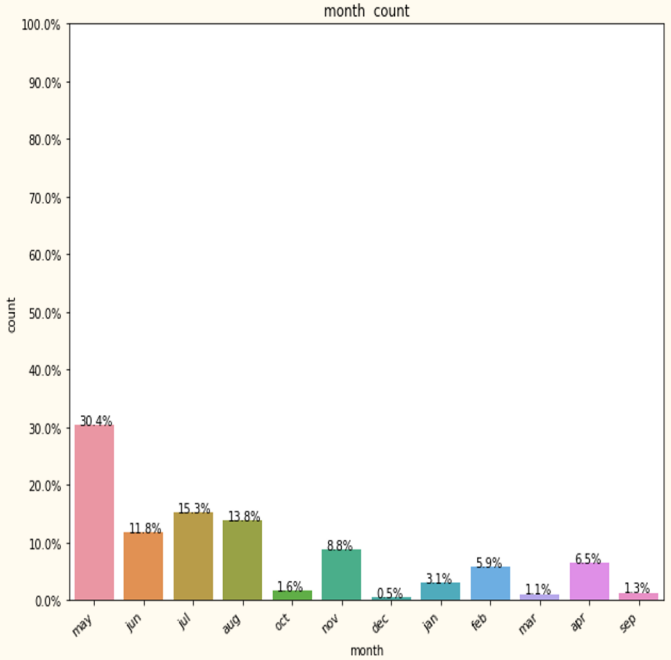


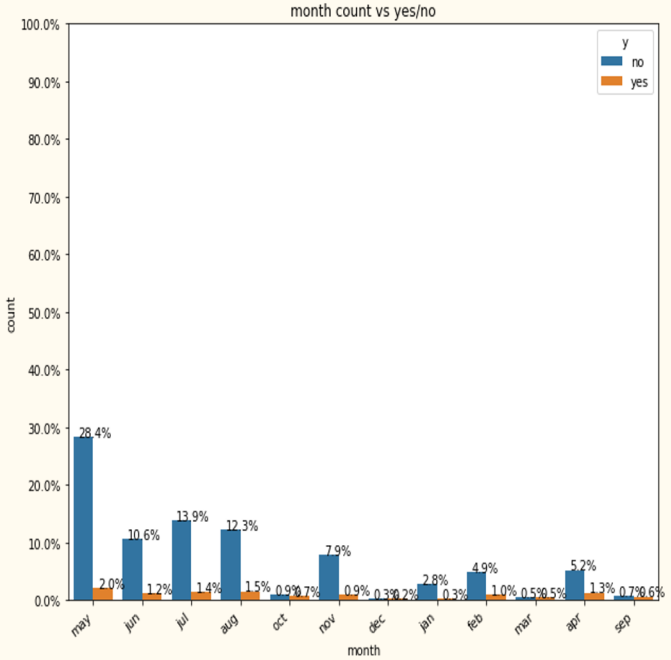
Most customers do not have a personal loan so they are more likely to subscribe if they do not have a personal loan.



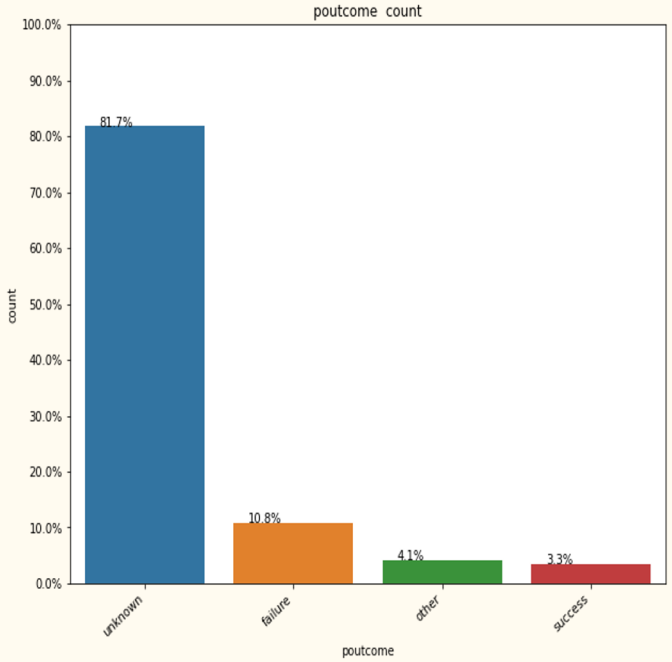


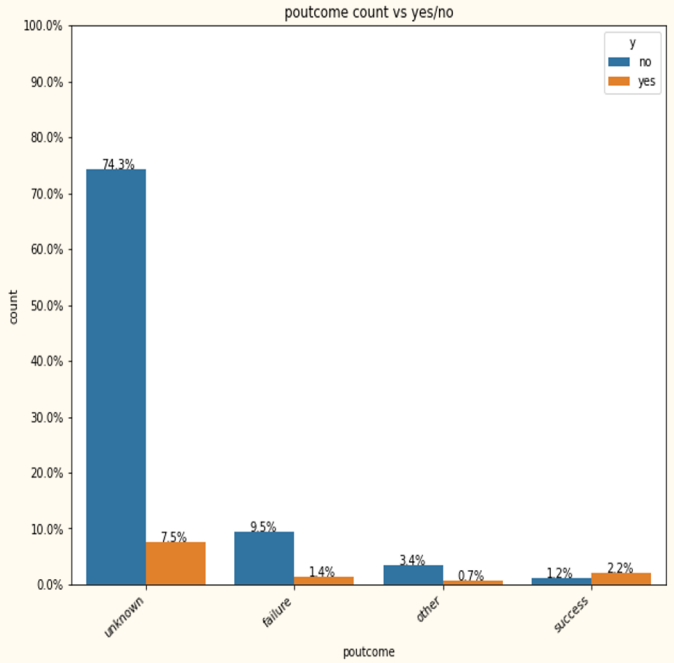
Most customers have cell phones.



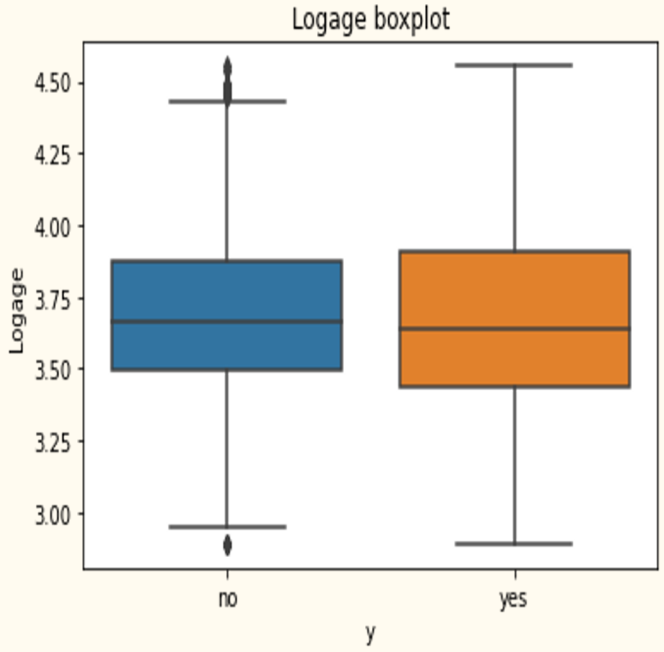


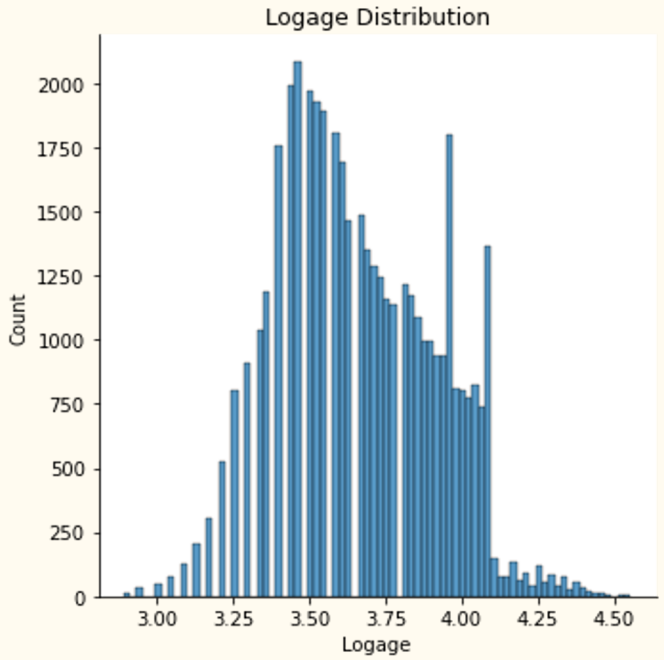
Most customers were contacted last in May.

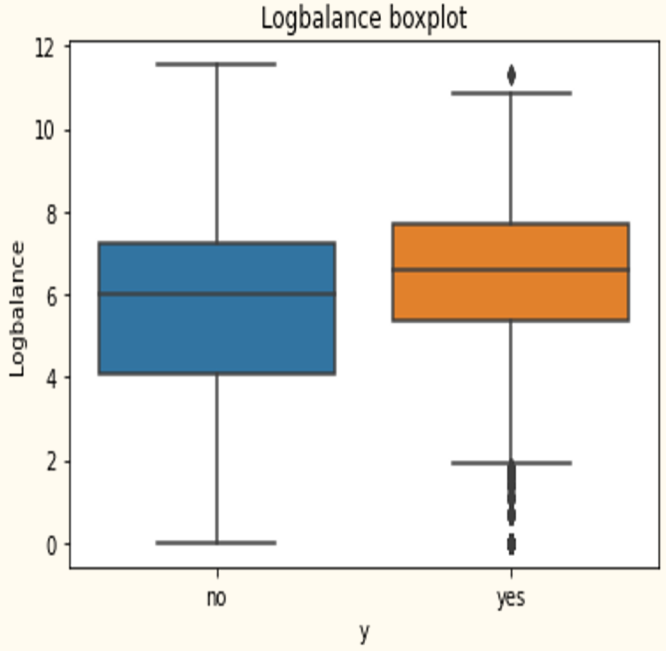


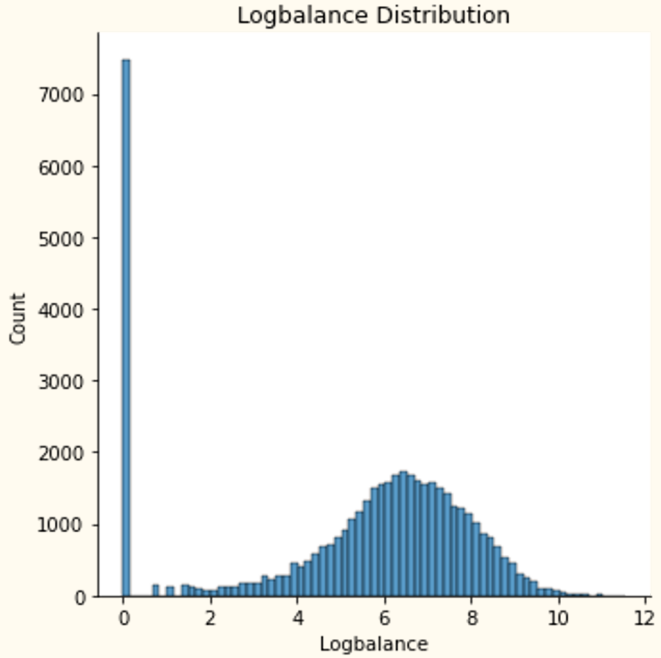


Most customers have a previous outcome of unknown.



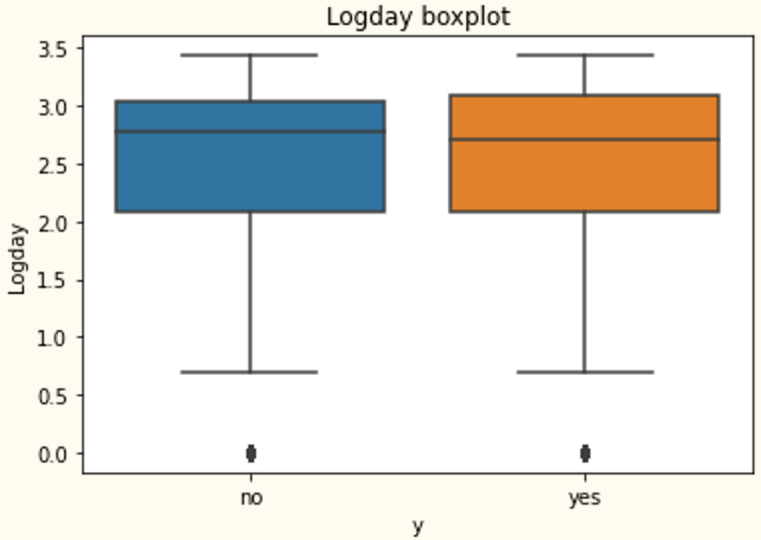


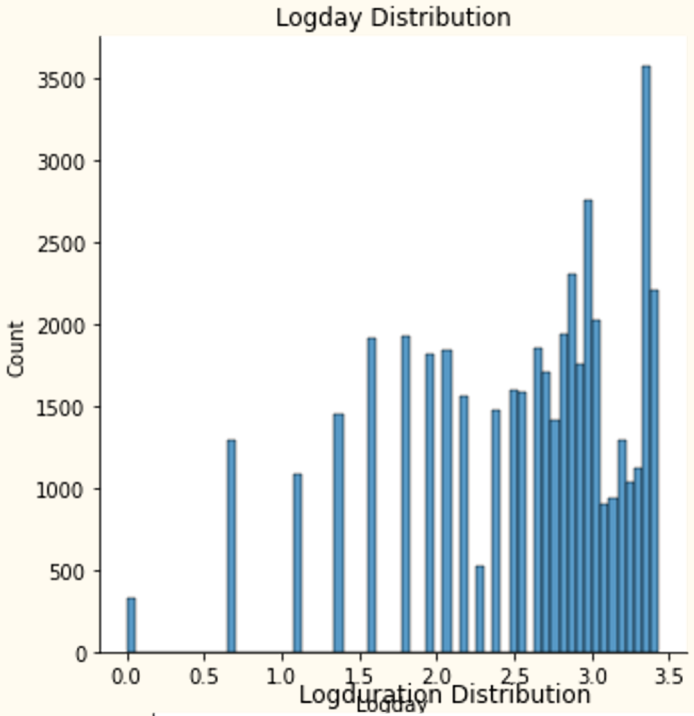




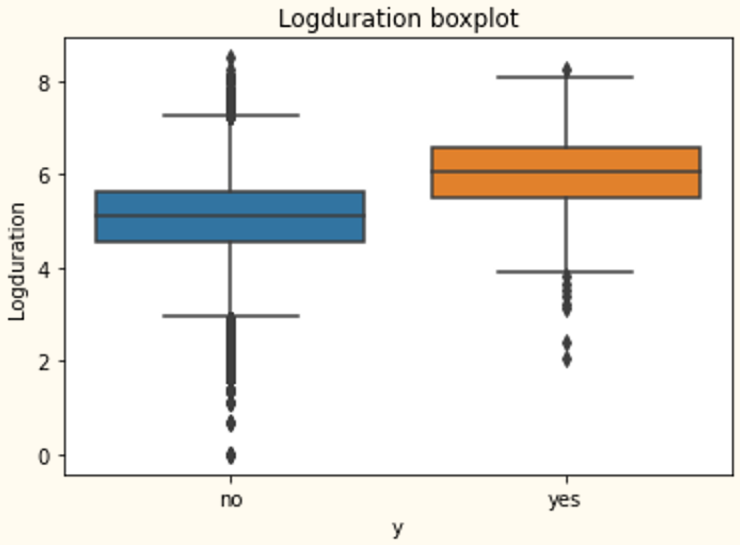
From the boxplot we see the median age of the customer who did and did not subscribe is 37-39. With so much overlap age is not necessarily a good indicator of customers and their subscription choices.
From the distribution we see that most of the customers are between 30-45.

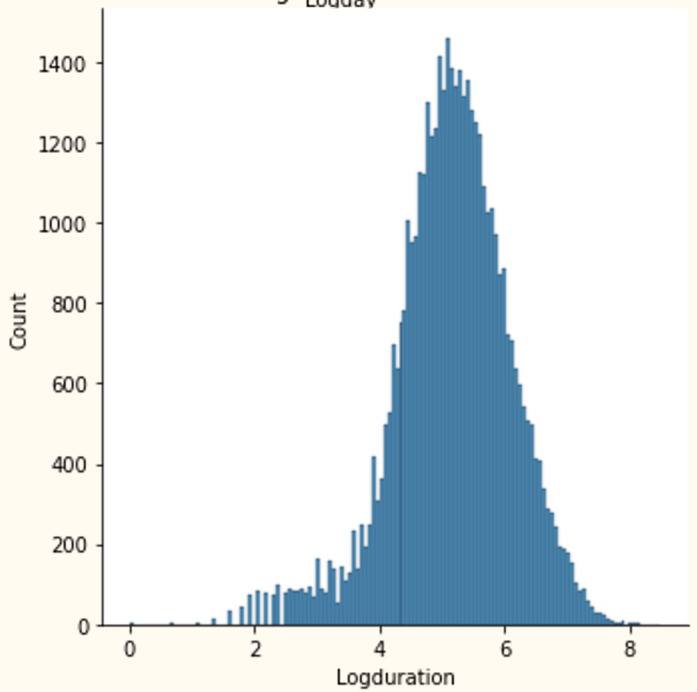
The median for balance is about the same for customers who subscribe as well as those who do not subscribe.
There is a lot of overlap so balance does not seem useful.



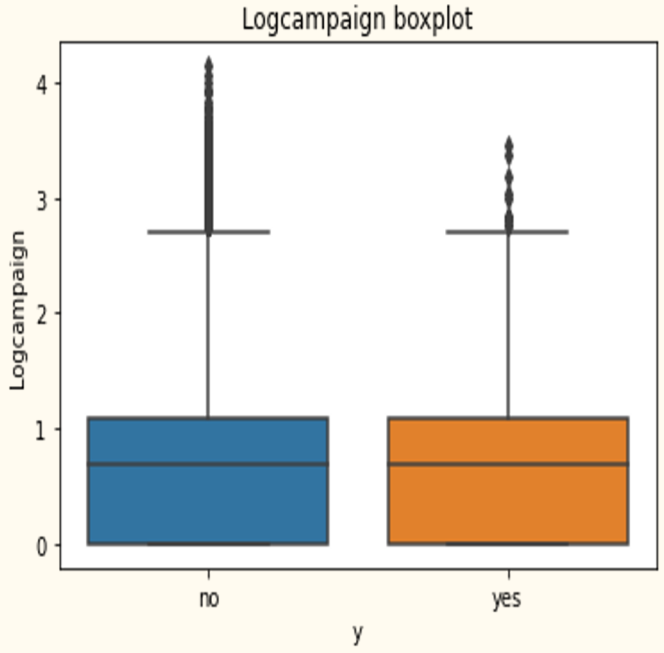


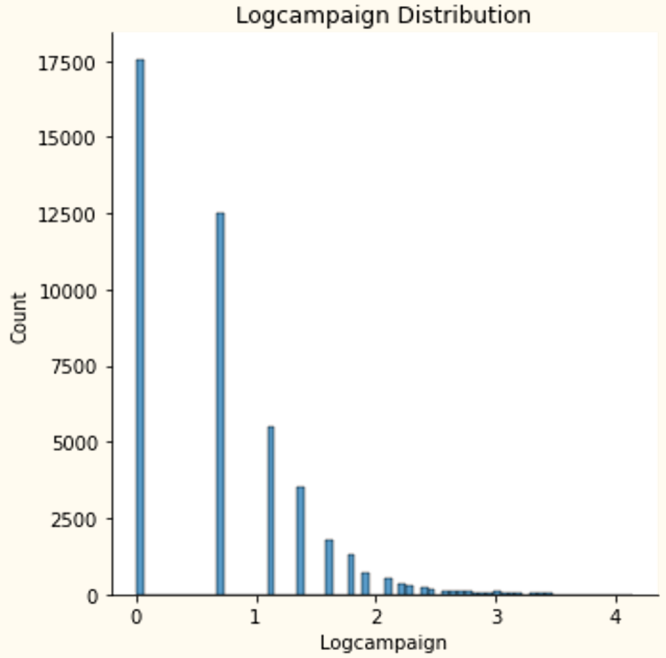
The median for the day customers subscribe or do not subscribe is between 15-16.
The distribution shows the most customers are between days 5 and 20.
With so much overlap the day does not seem useful.

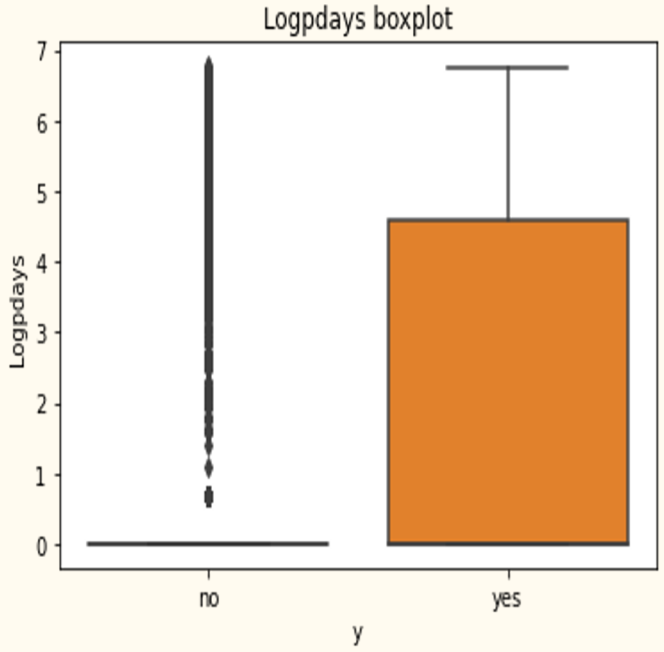


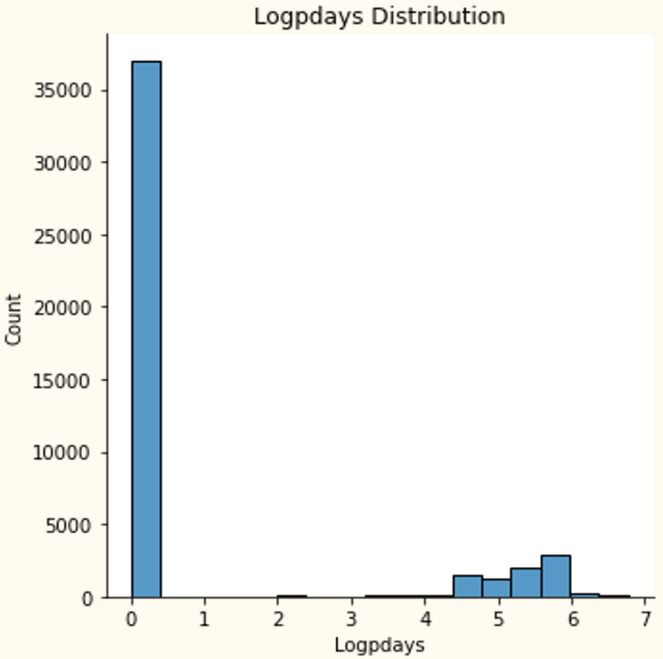


The duration has very little overlap so it will be useful in indicating subscription choices

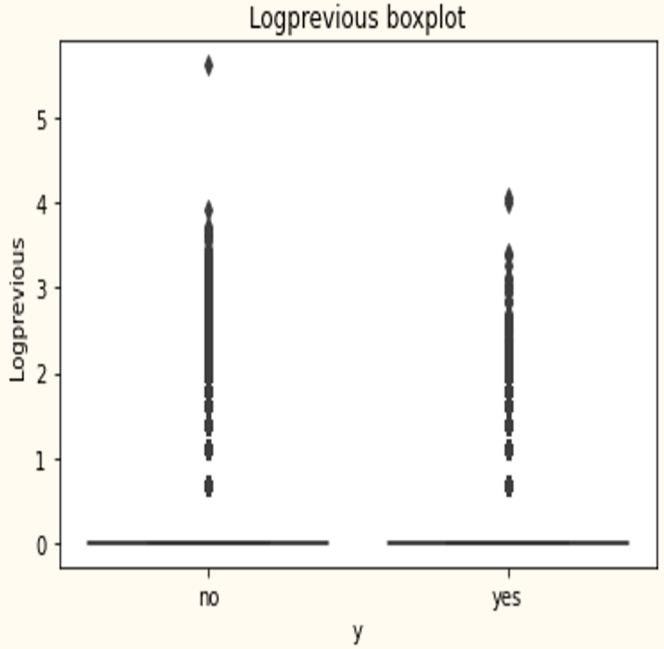


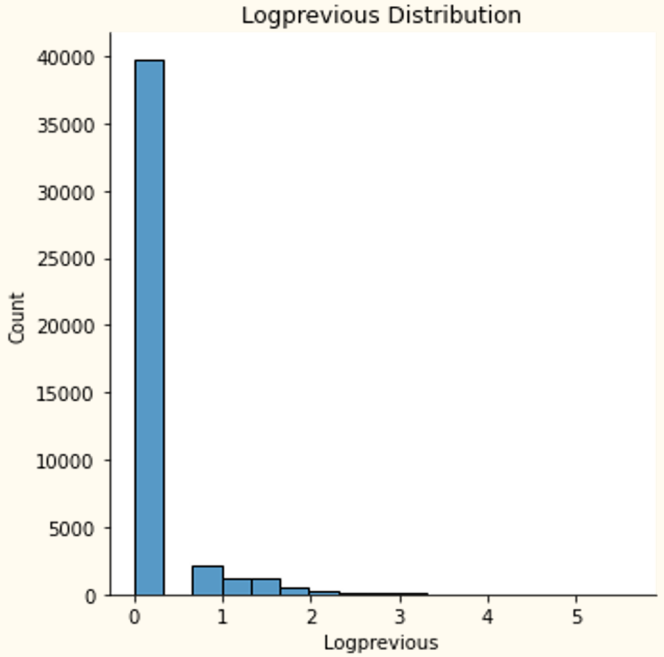




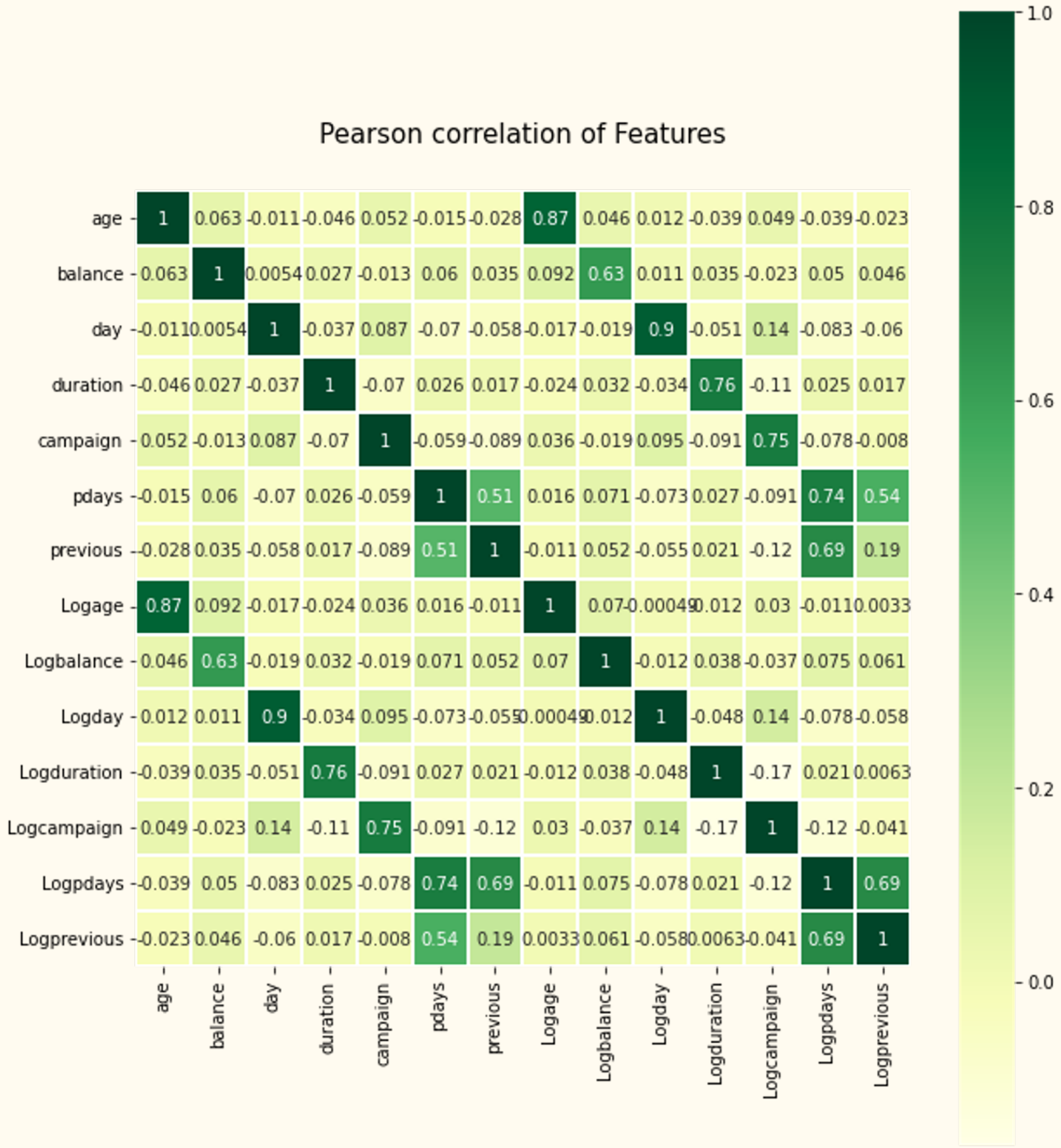


The campaign and pdays has very little overlap so they will be useful in indicating subscription choices





The customers fall between 0-0.25 for previous.

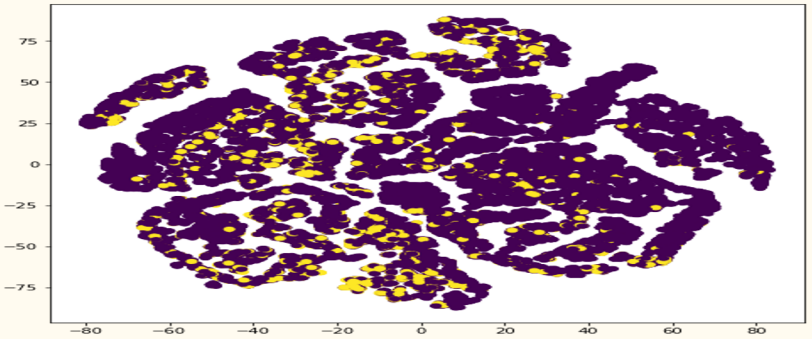


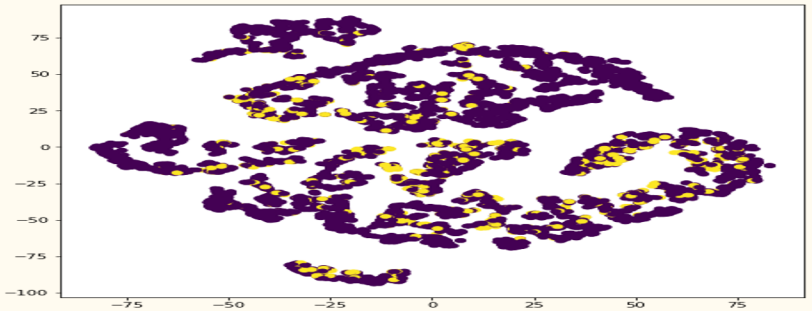
Previous and pdays have the highest correlation. They have a positive correlation of 0.69.

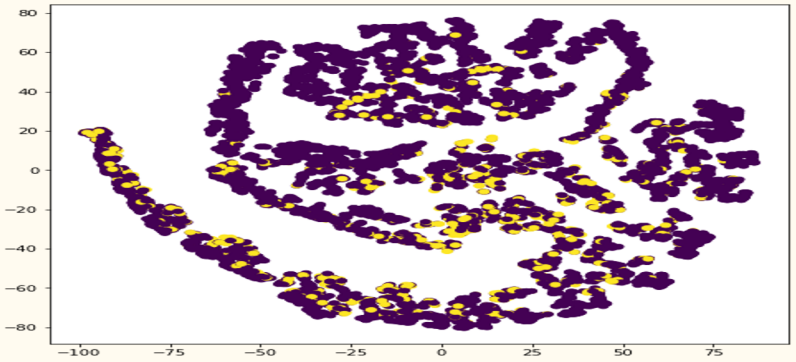
Recommendations

- Previous and pdays have the highest correlation.
- The campaign and pdays has very little overlap so they will be useful in indicating subscription choices.
- The duration has very little overlap so it will be useful in indicating subscription choices.

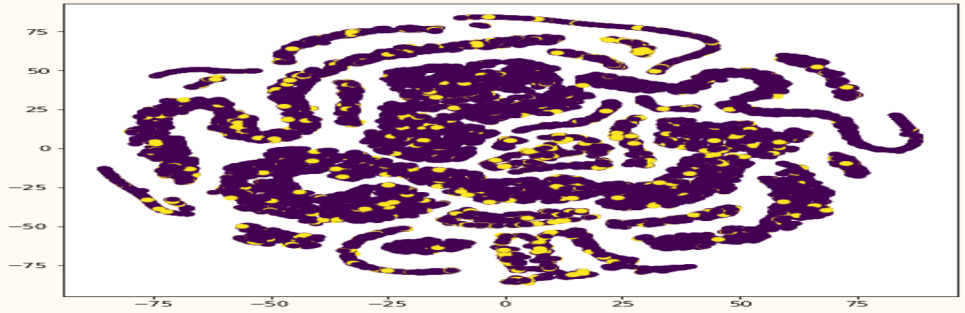
Visualization with Duration

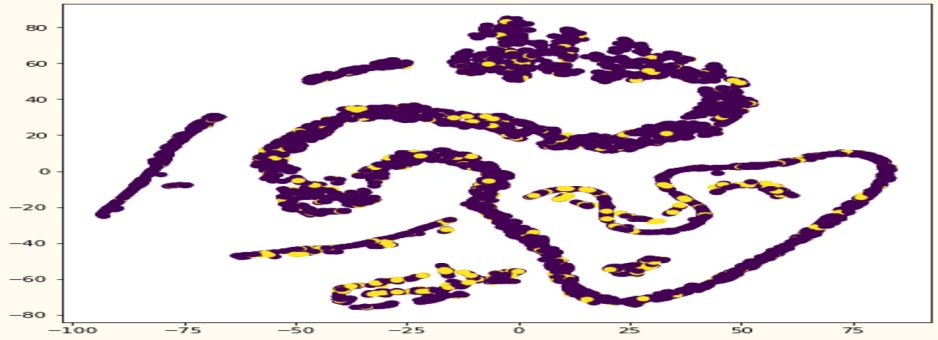


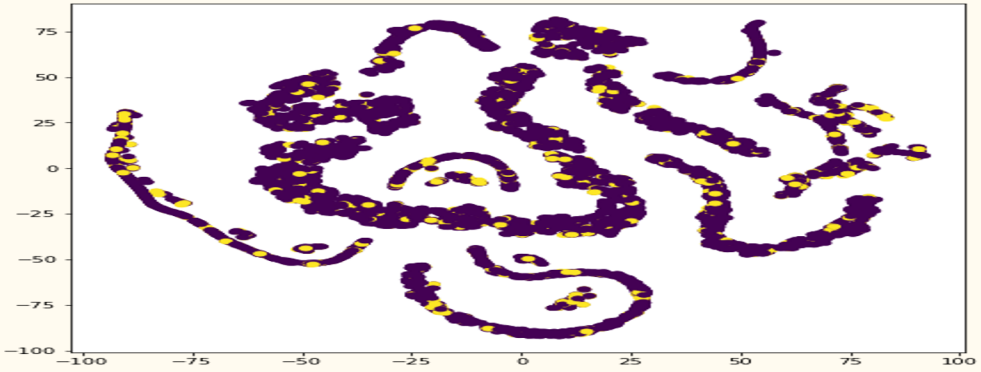




Visualization without Duration







Comparing the Model with and without Duration

- The "duration" is highly correlated with the targeted feature so the final model will not include the duration feature with "duration" column.
- Next we were able to check the model with a ROC AUC score of 1 without duration and class balancing.
- The ROC AUC score was 1 with duration.

Logistic Regression Results

AUC for k = 1e-05 is 0.791782429371228

AUC for k = 0.0001 is 0.971879036568271 8

AUC for k = 0.001 is 0.999960763066678

AUC for k = 0.01 is 0.999999083249221 5

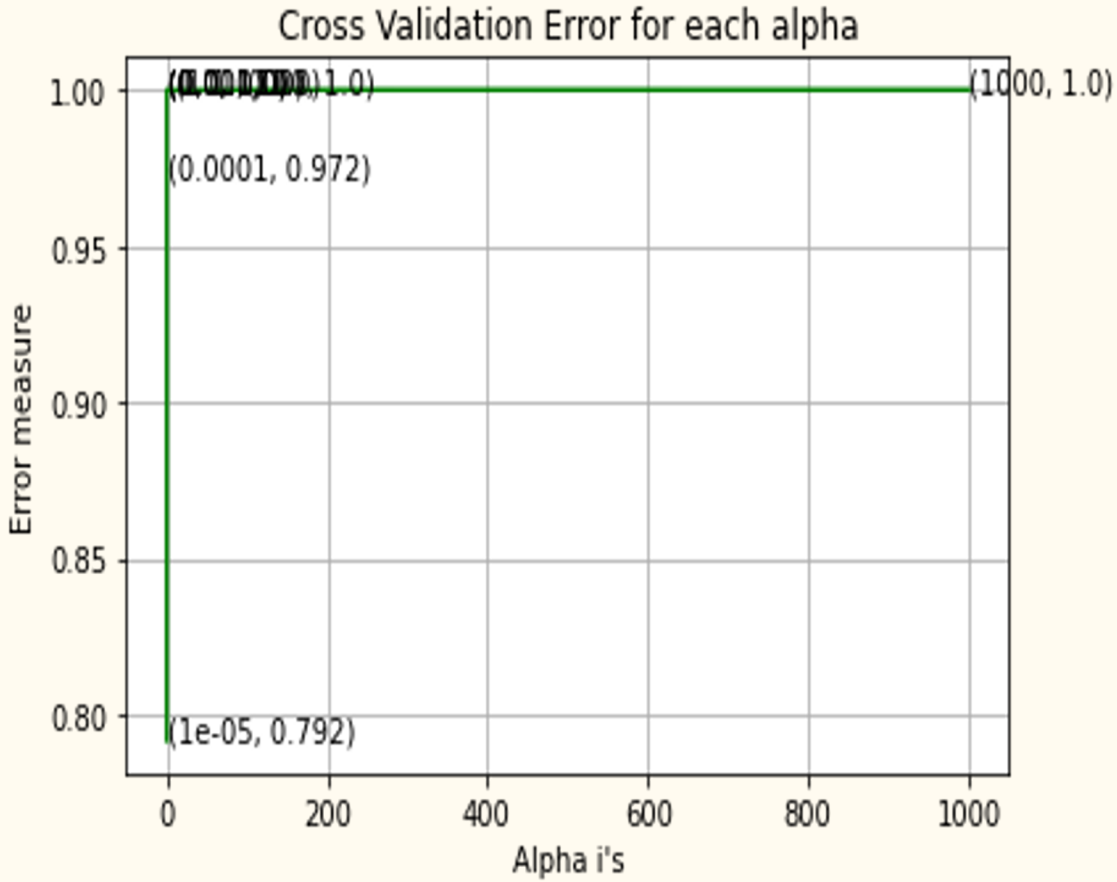
AUC for k = 0.1 is 0.999998349848598 6

AUC for k = 1 is 1.0

AUC for k = 10 is 0.999890906657352 5

AUC for k = 100 is 0.999988998990657 4

AUC for k = 1000 is 0.999969747224307 9



For values of best alpha = 1 The train AUC is: 1.0

For values of best alpha = 1 The cross validation AUC is: 1.0

For values of best alpha = 1 The test AUC is: 1.0

Linear SVM Model

AUC for alpha = 1e-05 is 0.5

AUC for alpha = 0.0001 is 0.9763815663420451

AUC for alpha = 0.001 is 0.9670252078961579

AUC for alpha = 0.01 is 0.9876326652970409

AUC for alpha = 0.1 is 0.9047264919889733

AUC for alpha = 1 is 0.8357076261747015

AUC for alpha = 10 is 0.7374968486691987

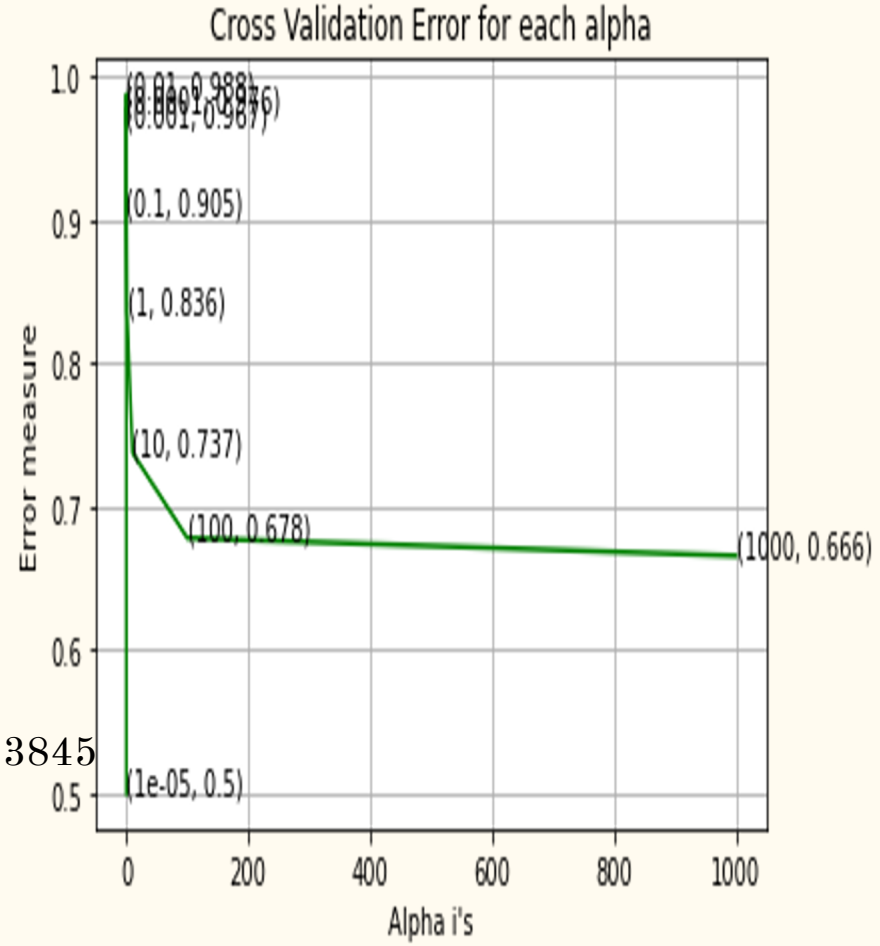
AUC for alpha = 100 is 0.6780008232421991

AUC for alpha = 1000 is 0.6656153368738248

For values of best alpha = 0.01 The train AUC is: 0.9763558135102671

For values of best alpha = 0.01 The cross validation AUC is: 0.9772851159093845

For values of best alpha = 0.01 The test AUC is: 0.9774027766682715



Random Forest Model

AUC for number of estimators = 10 is 1.0

AUC for number of estimators = 50 is 1.0

AUC for number of estimators = 100 is 0.9999999999999999

AUC for number of estimators = 500 is 1.0

AUC for number of estimators = 1000 is 1.0

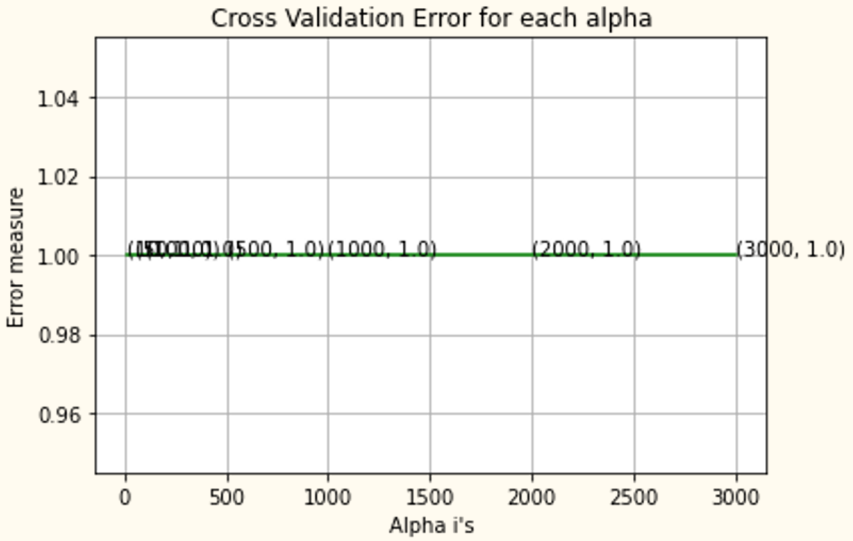
AUC for number of estimators = 2000 is 1.0

AUC for number of estimators = 3000 is 1.0

For values of best alpha = 10 The train AUC is: 1.0

For values of best alpha = 10 The cross validation AUC is: 1.0

For values of best alpha = 10 The test AUC is: 1.0



Conclusion

- We wanted to predict whether a customer will subscribe to a term deposit.
- We did EDA on the numerical and categorical data.
- After doing data visualization with T-SNE the data shows the two classes are overlapping.
- Using one-hot encoding we were able to encode the categorical variables.
- After testing various models the Random Forest Model was the most accurate with a test AUC score of 1.0.