Customer Conversion Prediction

Problem Statement

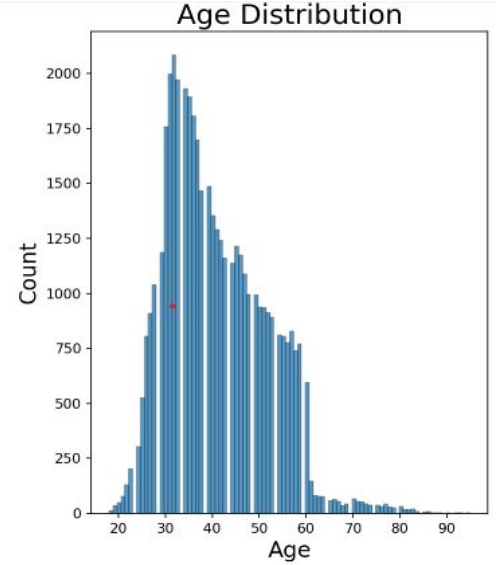
Problem statement is to build a ML model that predict if a client will subscribe to insurance or not.We need to uses the historical data to reduce the cost of telephonic marketing,still it effective but Incur lot of cost.It is also important to identify the customers that are most likely to convert beforehand so that they can be specifically targeted via call

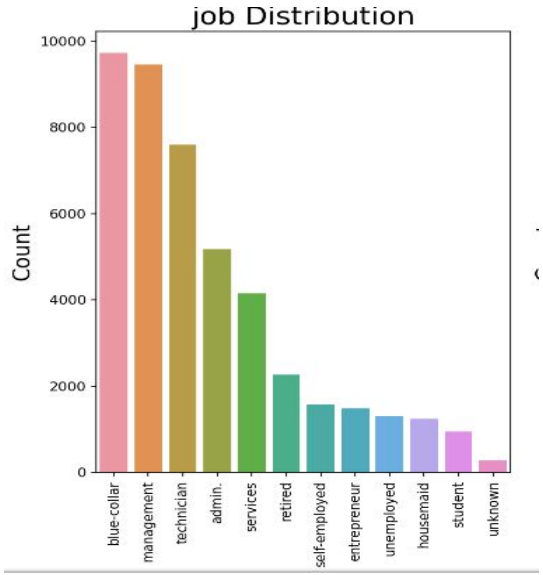
Data Cleaning

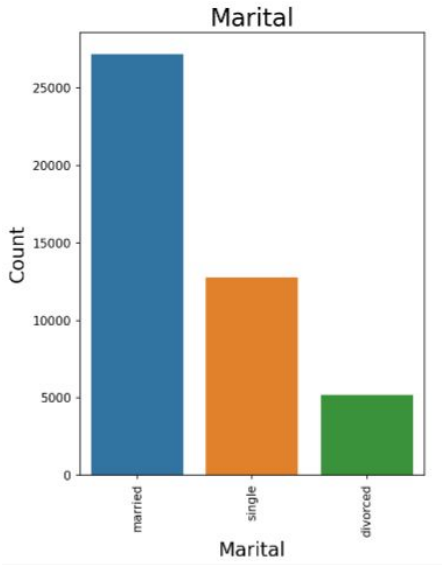
1. Checked for null values but found no null values
2. Checked for duplicate values , we got some duplicate values and we removed them
3. Checked for data types but all the data types were found correct
4. Checked outliers by describe function and we got some outliers but we have imbalance data so we are not able to determine this is outliers or not

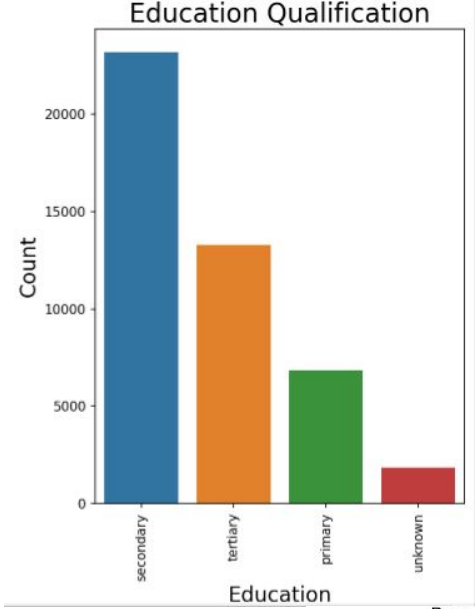
Important insights of EDA

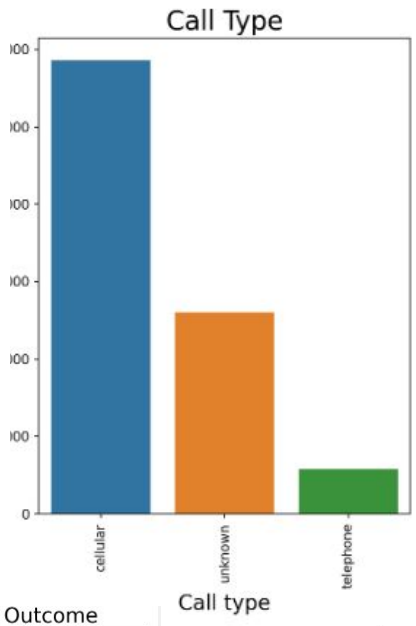
1. Checked linear correlation between the features but we didn't get any relationships between the data
2. Interpretations of feature distribution

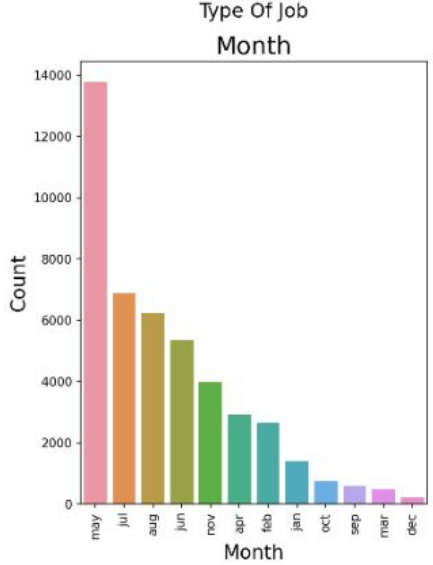


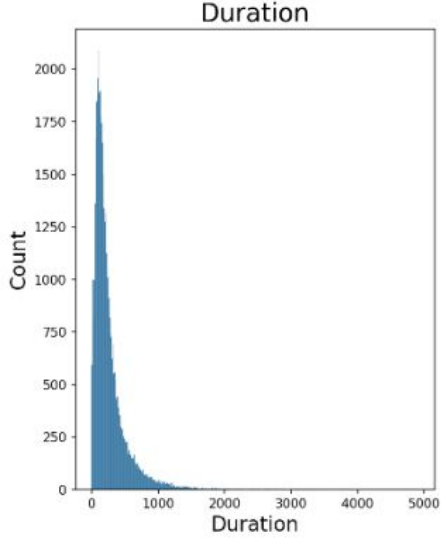


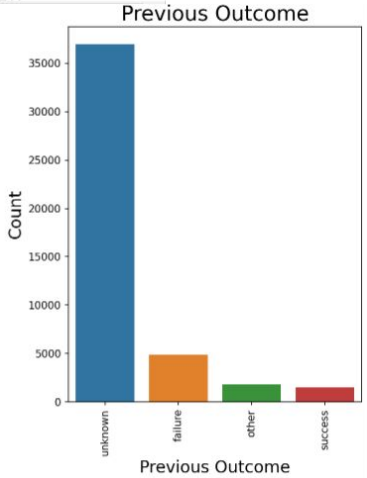


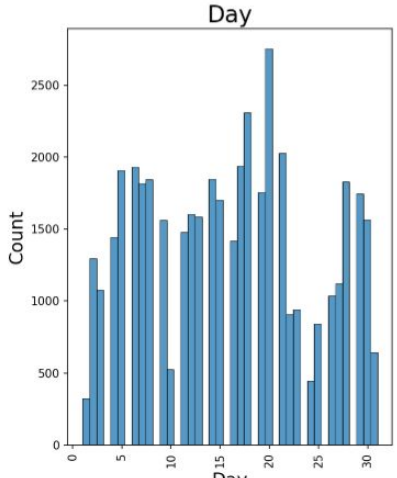

Education Qualification
Count
20000
15000
10000
5000
0
secondary
tertiary
primary
unknown
Education

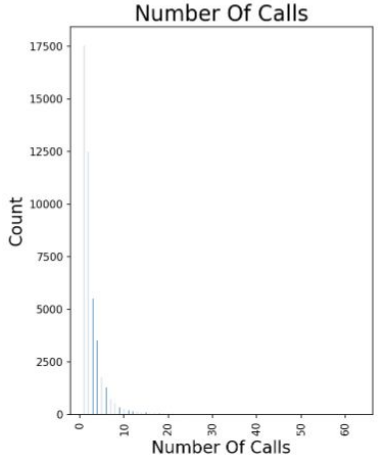

Call Type
cellular
unknown
telephone
Call type


Type Of Job
Month
Count
14000
12000
10000
8000
6000
4000
2000
0
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec
Month


Duration
Count
2000
1750
1500
1250
1000
750
500
250
0
0
1000
2000
3000
4000
5000
Duration


Previous Outcome
Count
35000
30000
25000
20000
15000
10000
5000
0
unknown
failure
other
success
Previous Outcome


Day
Count
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30

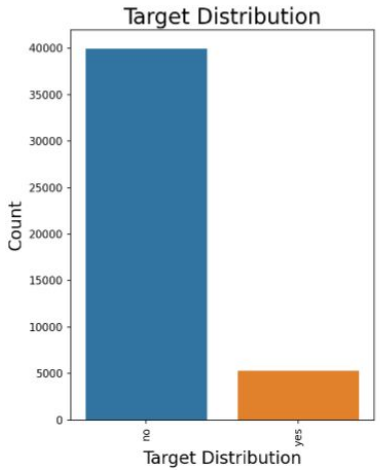

Number Of Calls
Count
17500
15000
12500
10000
7500
5000
2500
0
0
10
20
30
40
50
60
Number Of Calls

1. .AGE
    - Company targets 30 to 40 ages people to sell their insurance plan
    - People aged less than 25 years and more than 60 years are less targeted by the company for insurance
2. JOB
    - Blue collar and management people are targeted most by insurance company
    - Students are least targeted for insurance plan

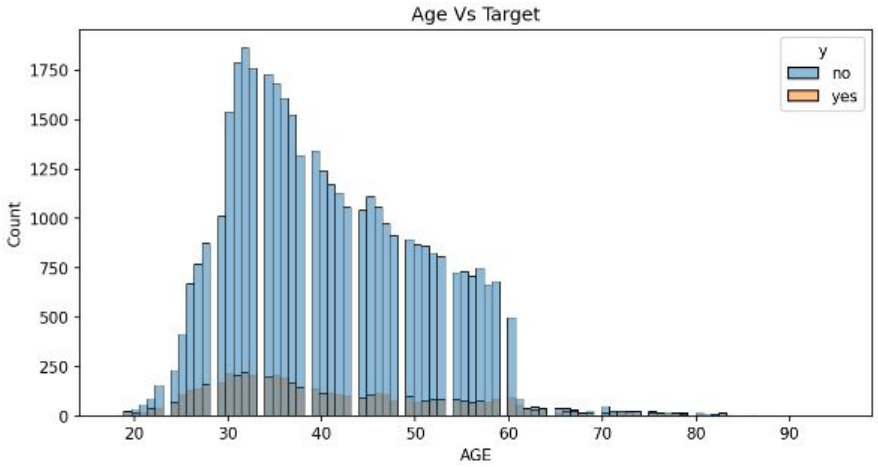
3. MARITAL STATUS
    - Married people are targeted the most and divorced people are least targeted

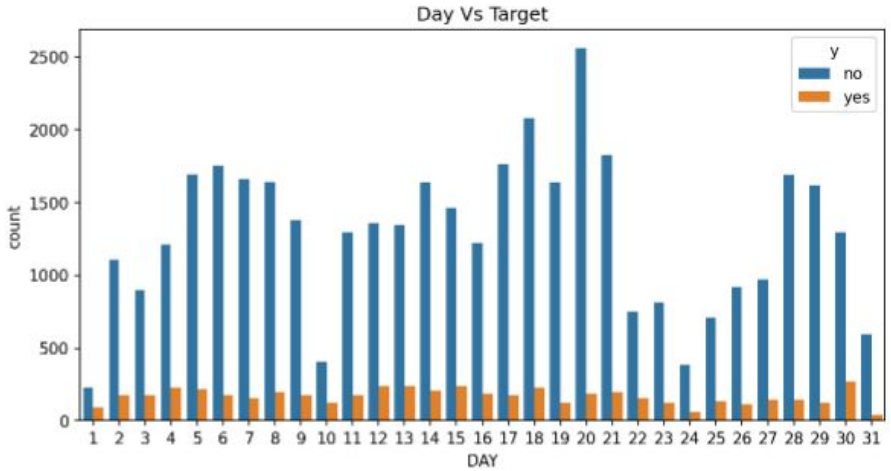
4. EDUCATIONAL QUALIFICATION
    - Secondary qualified people have been targeted thye most for insurance followed by tertiary and primary
5. CALL_TYPE
    - Cellur mode is most used for contacting the people for campaign

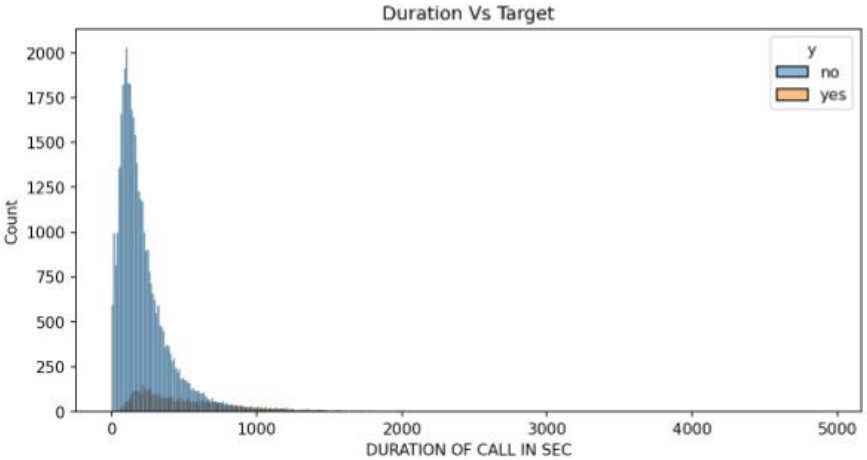
6. Month
    - May month is when maxmium calls are generated for new insurance customers,december is the least no of calls placed by the company
7. Duration
    - Maximum number of calls lasted less than 1000 seconds
8. Previous Outcome
    - More unknown as status than failure and success
9. Day
    - Most number of calls are between in the mid of month
10. Number of Calls
    - Most People are contacted only 1 or 2 times by the company
11. Target
    - No of people subscribed is very less compared to Unsubscribed people

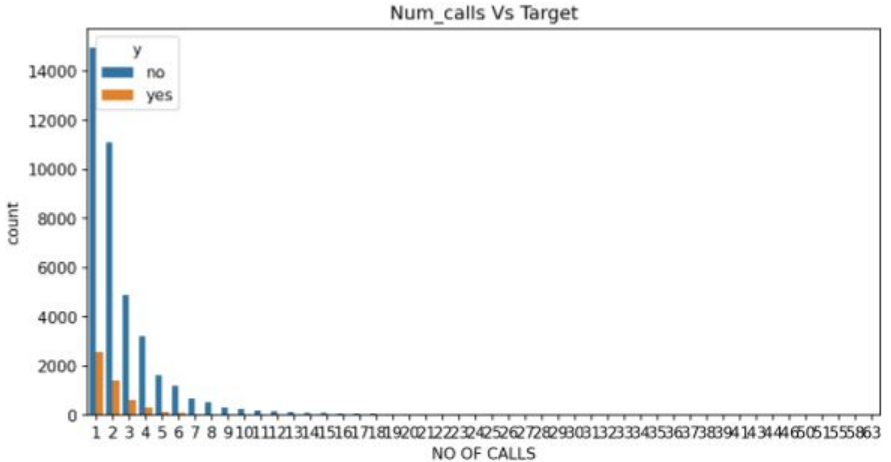


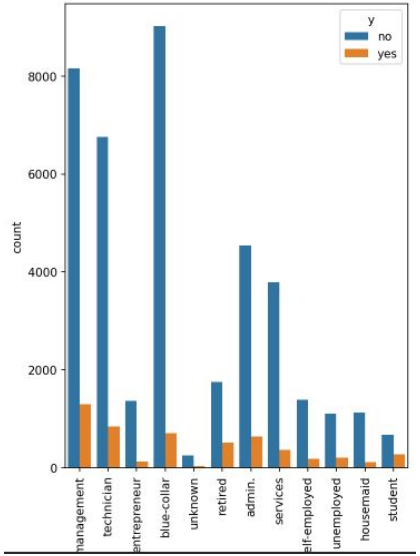
EDA on feature vs Target

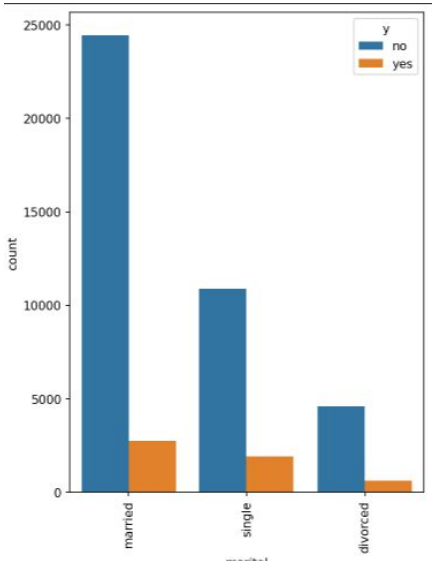


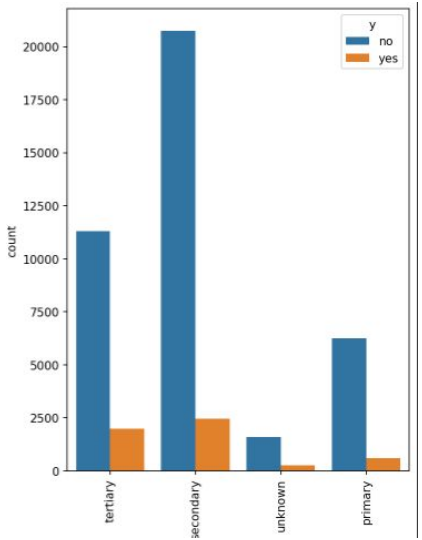


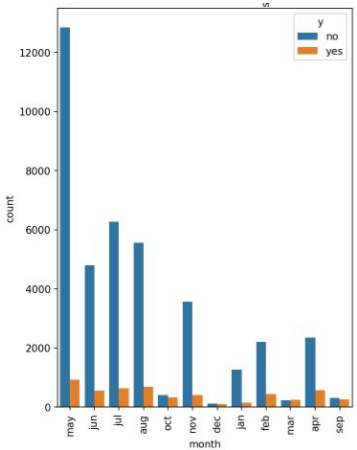


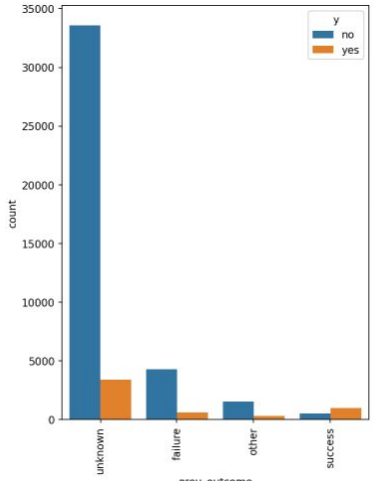


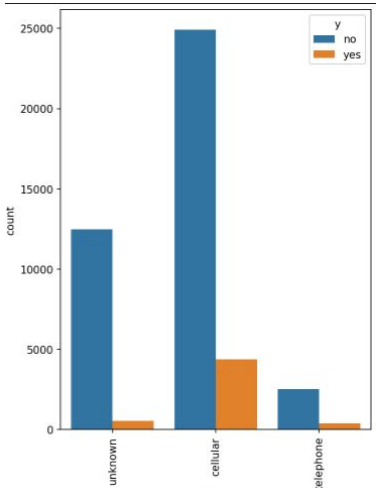

y
no
yes
count
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student


y
no
yes
count
married
single
divorced


y
no
yes
count
tertiary
secondary
unknown
primary


y
no
yes
count
may
jun
jul
aug
oct
nov
dec
jan
feb
mar
apr
sep
month


y
no
yes
count
unknown
failure
other
success


y
no
yes
count
unknown
cellular
telephone

Built Models

- We Encode our data by Label and One-hot encoding
- We scaled our data by standard scalar and Split our data
- We imported SMOTETomek module from imblearn.combine to balance our trained data
- We built Logistic regression , KNN, Decision Tree, Random Forest, XG Boost
- With all the features we got AUROC score of 0.77
- From the feature selector we found age,education qualification, duration , day, month to have high feature importance and model was built on these features
- We got a 0.79 AUROC Score, by Random Forest. So we can consider Random Forest is best suitable model