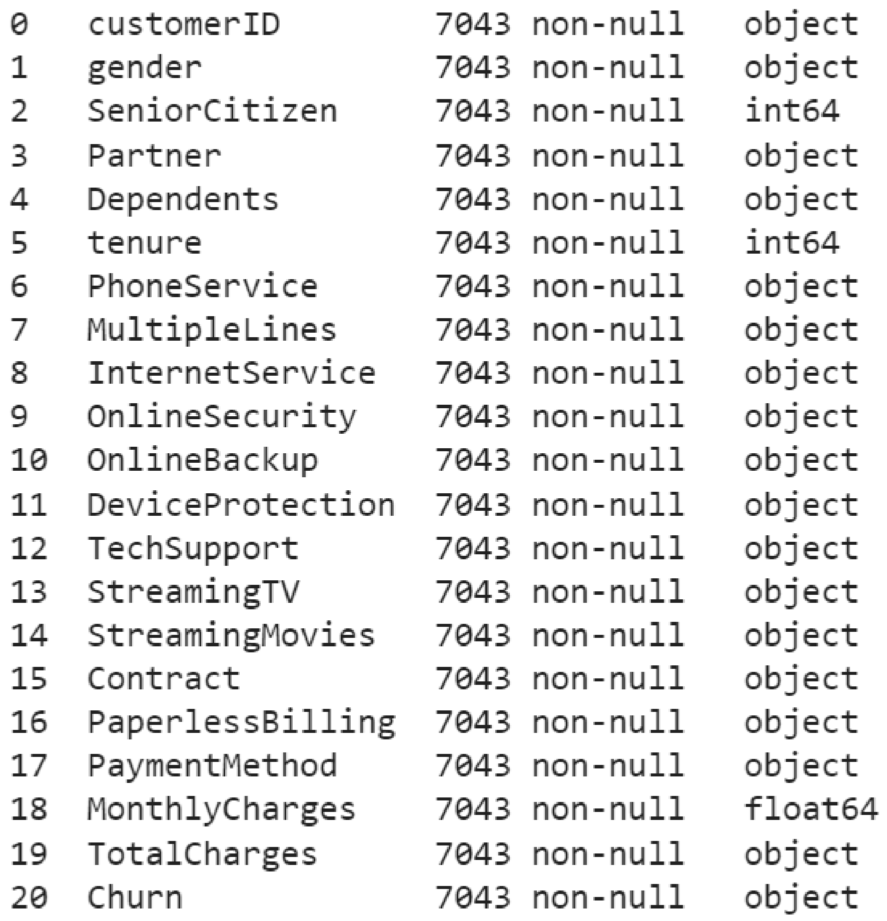
Task: Customer Churn Prediction

Submitted by:

Muhammad Abdullah Asif Khan

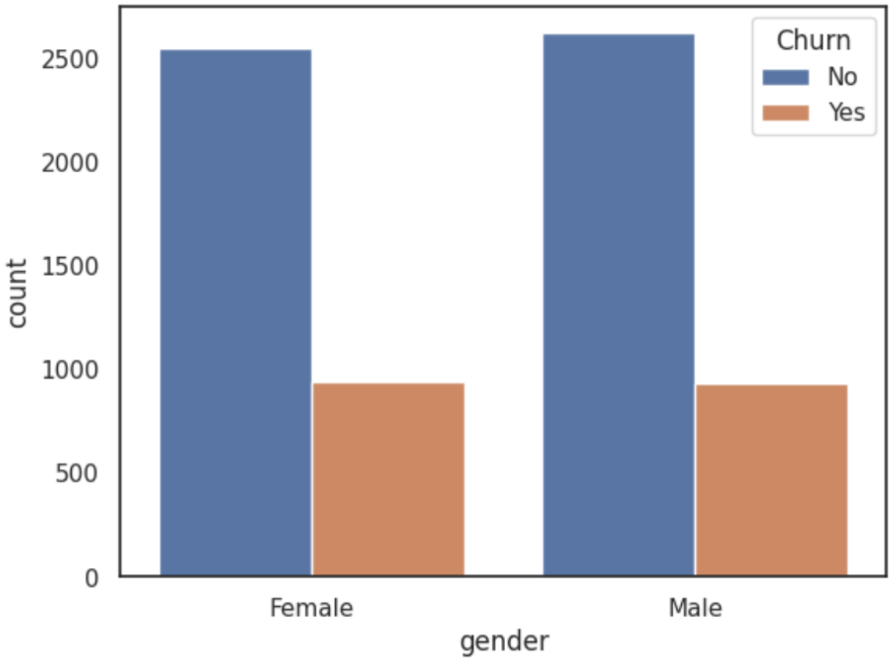
First we checked the data types and found the column of DataCharges has data type object, so coverted it to numerical

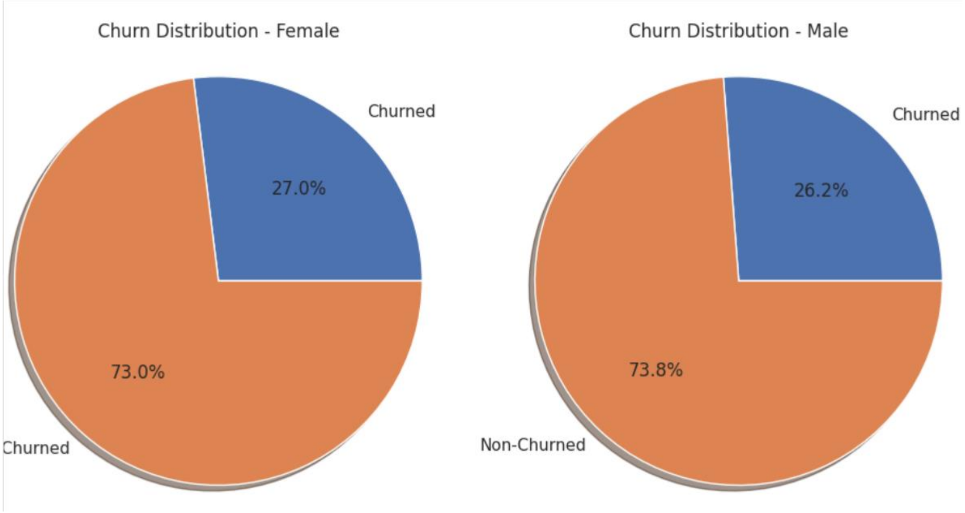


```
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
```

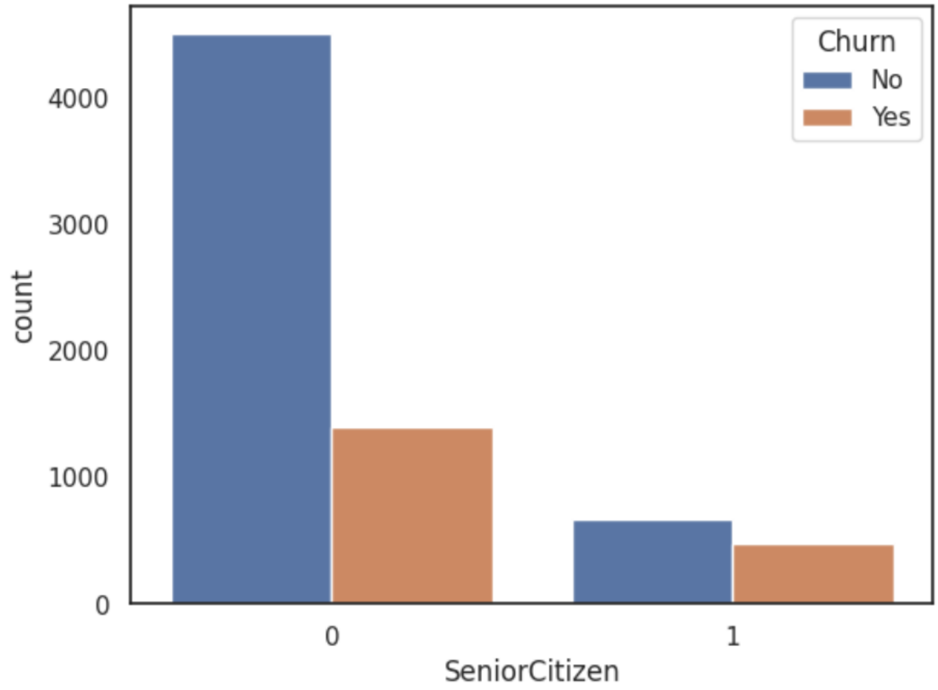
EDA :

Visualization Based on Demographics of customer

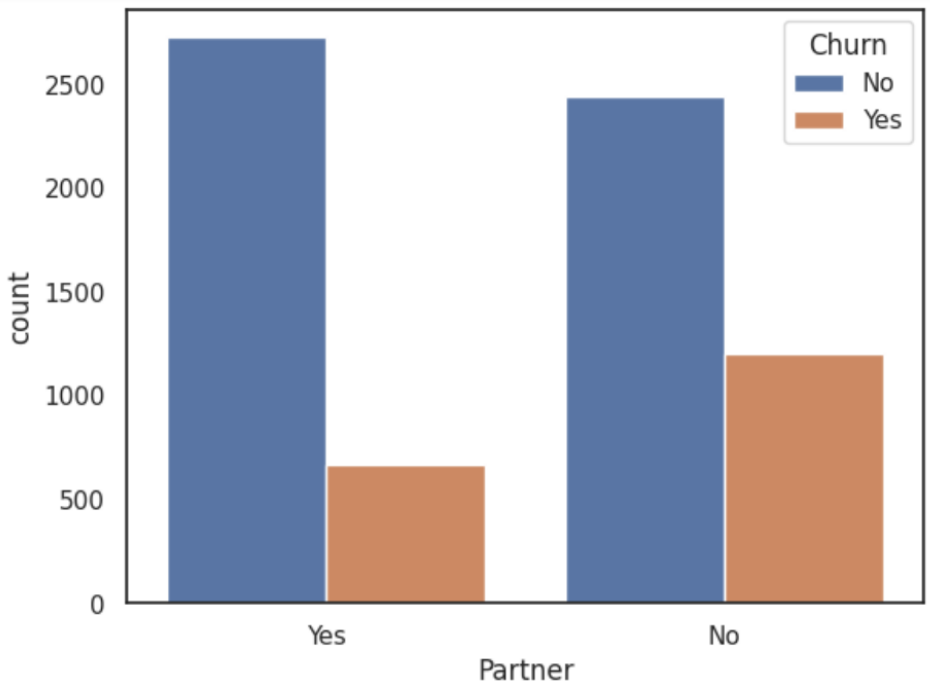




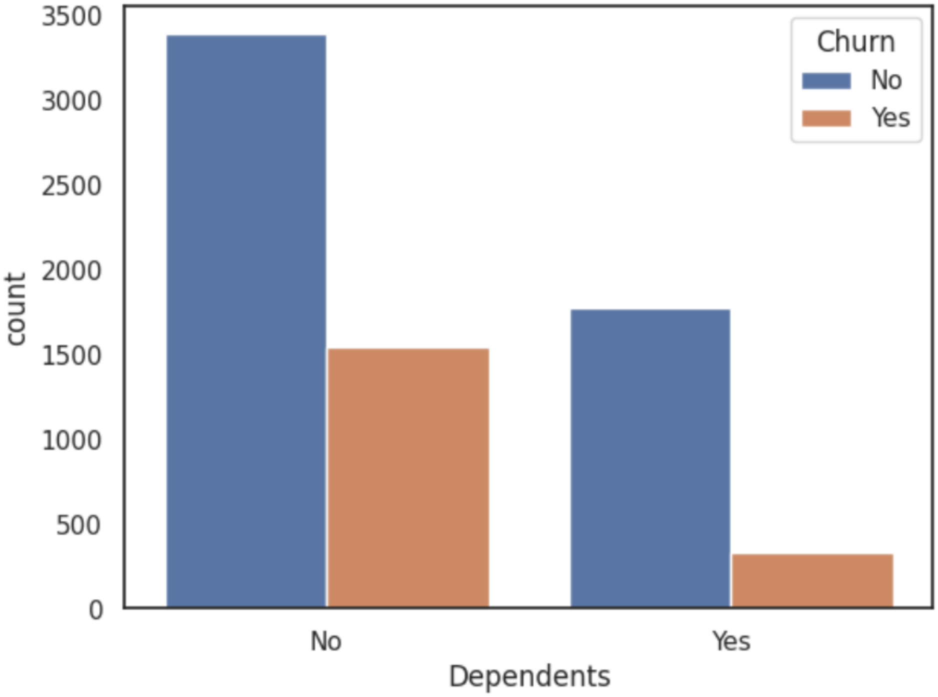
- Plot and pie chart show that the amount of churned customers within the two genders present are quite balanced, hence we can infer that gender won't be able to provide much information regarding the churn.



It can be inferred that the percentage of churned customer is more in case if the customer belongs to the senior citizen category

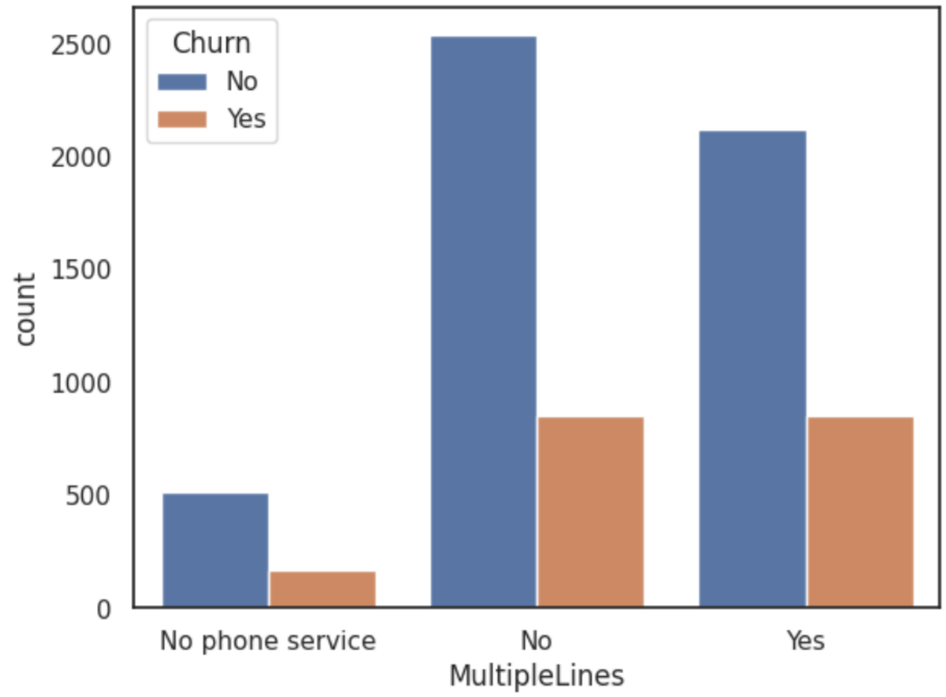


Customers who don’t have partners are more likely to churn.

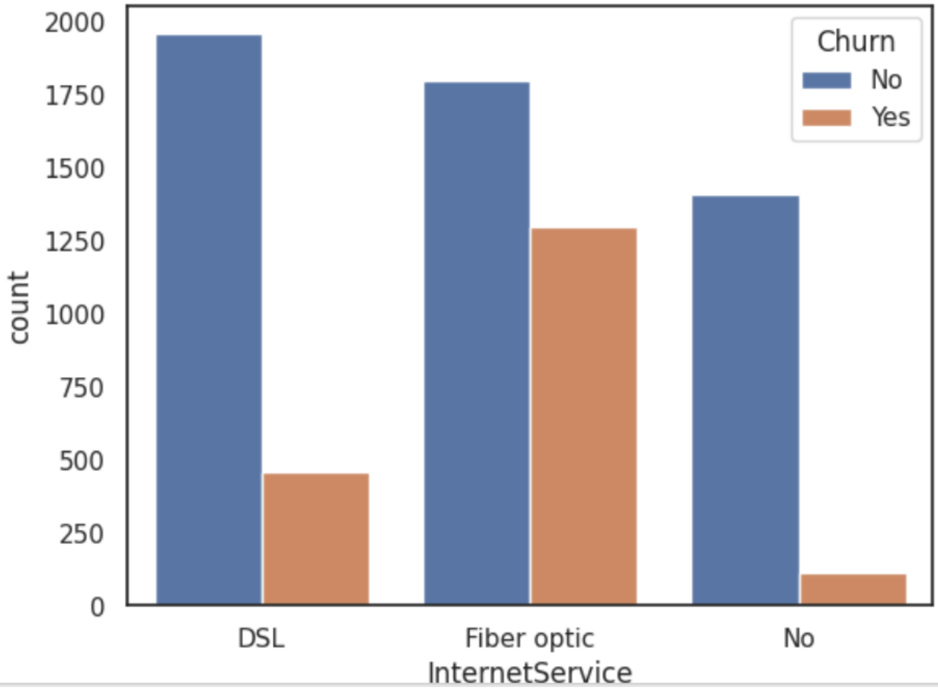


Customers who don’t have dependents are more likely to churn.

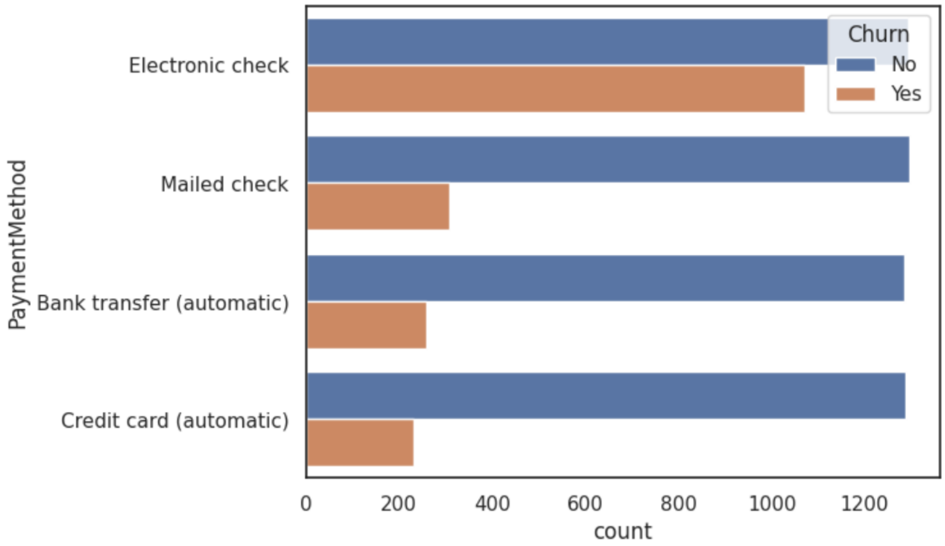
Visualization based on services



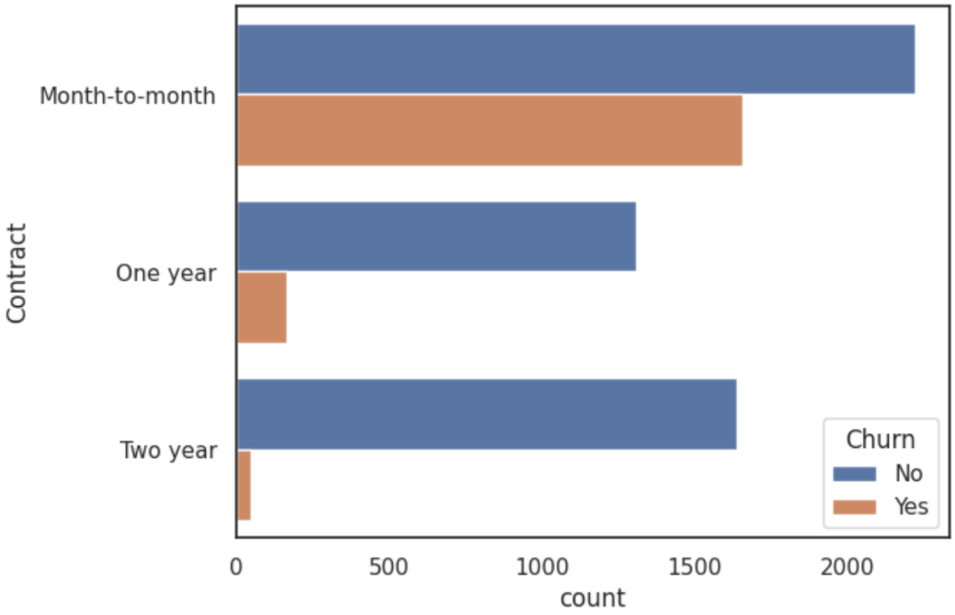
Customers having multiple lines are more likely to churn.



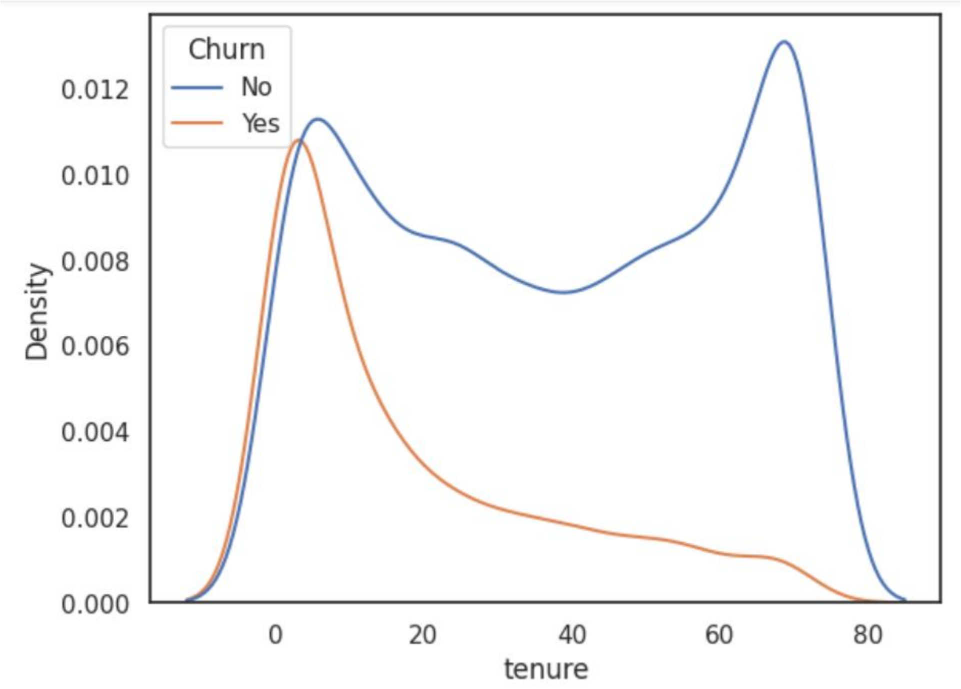
Customers having Fiber Optic Internet service are more likely to churn.



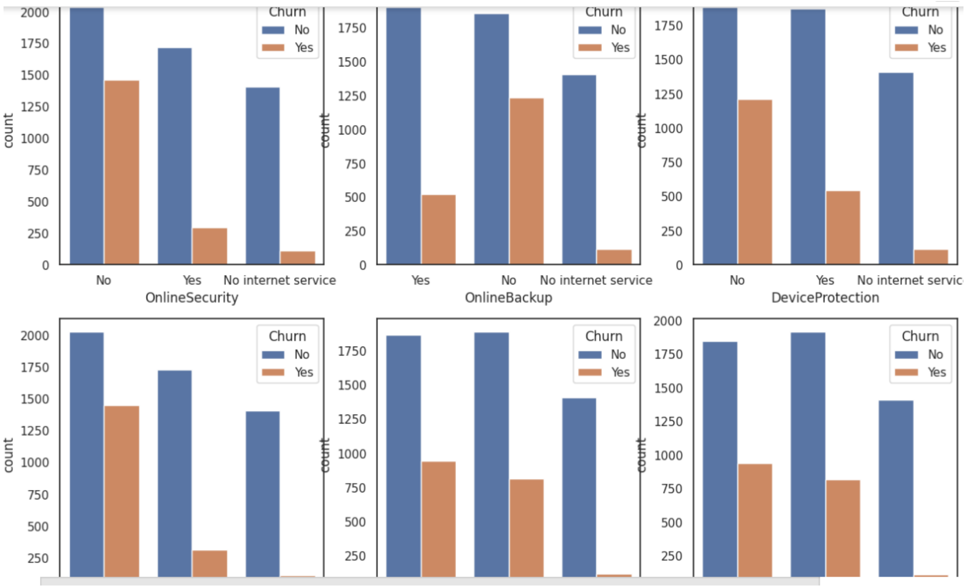
Customers having Electronic check as Payment Method are more likely to churn.



Customers having month-to-month contract are more likely to churn.



We can see from the tenure graph, that most customer churn at 0-20 months of tenure, hence more attention needs to be given there.



Insights from EDA graphs

Looking at the different variables and their distribution through the graph we can take some initial insights as

Some of the important variables which can affect the churn probability are: Seniorcitizen, having parter/dependents, Phone service, Internet service, Payment method, Tenure, monthlycharges, total_charges, contract.

Customers on month-to-month contract have more probability of churning

Customers paying through electronic check shows more possibility of churning.

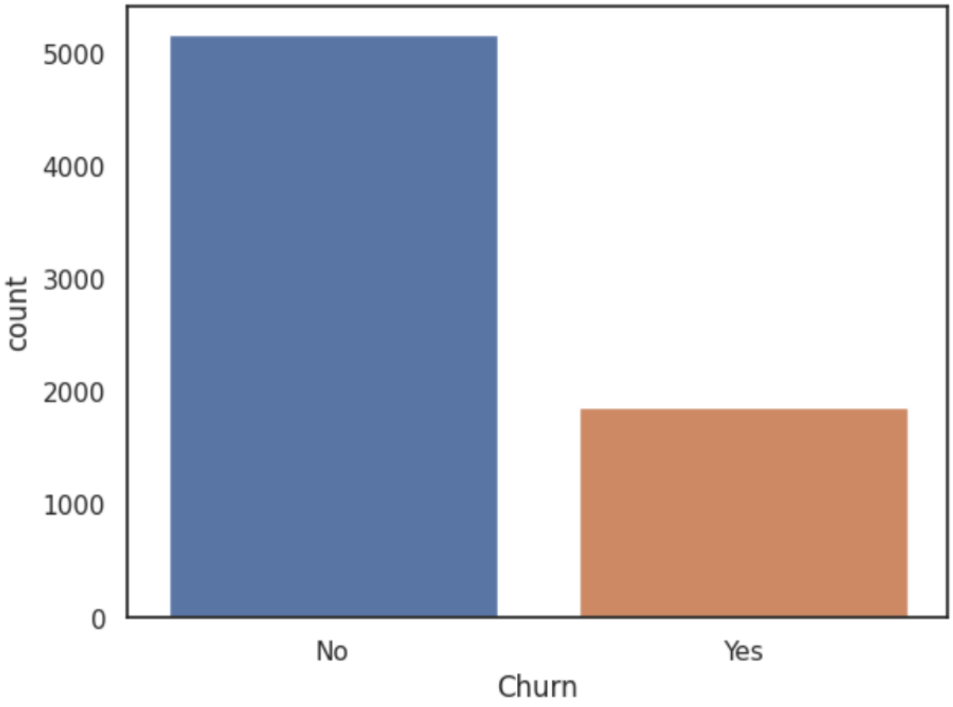
Percentage of churn is more in customers who have 'Fiber optic' as their internet service compared to others.

Majority amount of churned customers are churning within first 30 months of tenure.

Customers with no Online security, no Online Backup, no Device protection and no Tech support have more probability of churning.

Since the customers streamingTV, streamingMovies and similar to not doing them respectively, hence won't be much useful for churn determination.

Imbalanced Data:



As we can see the data is imbalanced.

There are couple of basic method that we can follow to balance the data like

Oversampling (duplicating the datapoints from minority class)

Undersampling (removing the datapoints from majority class)

Introducing class weights (weights given to each class while training, inversely proportional to the no. of datapoints present)

We'll follow the class weights approach to balance the dataset

Feature Engineering

Dropping the customerID column

Transforming the columns with YES/NO values to 1 and 0 respectively

Transforming gender column with 1 representing Female and 0 representing Male

Getting dummy variable for columns having more than 2 categories(/values)

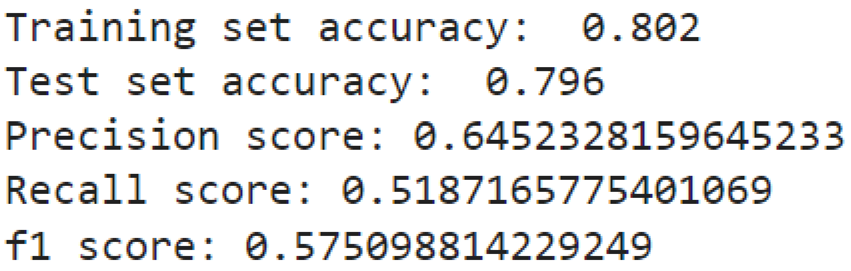
Getting the correlation dataframe and selecting columns with greater than correlation coefficient

Correlation coefficient threshold calculated as median value of the list

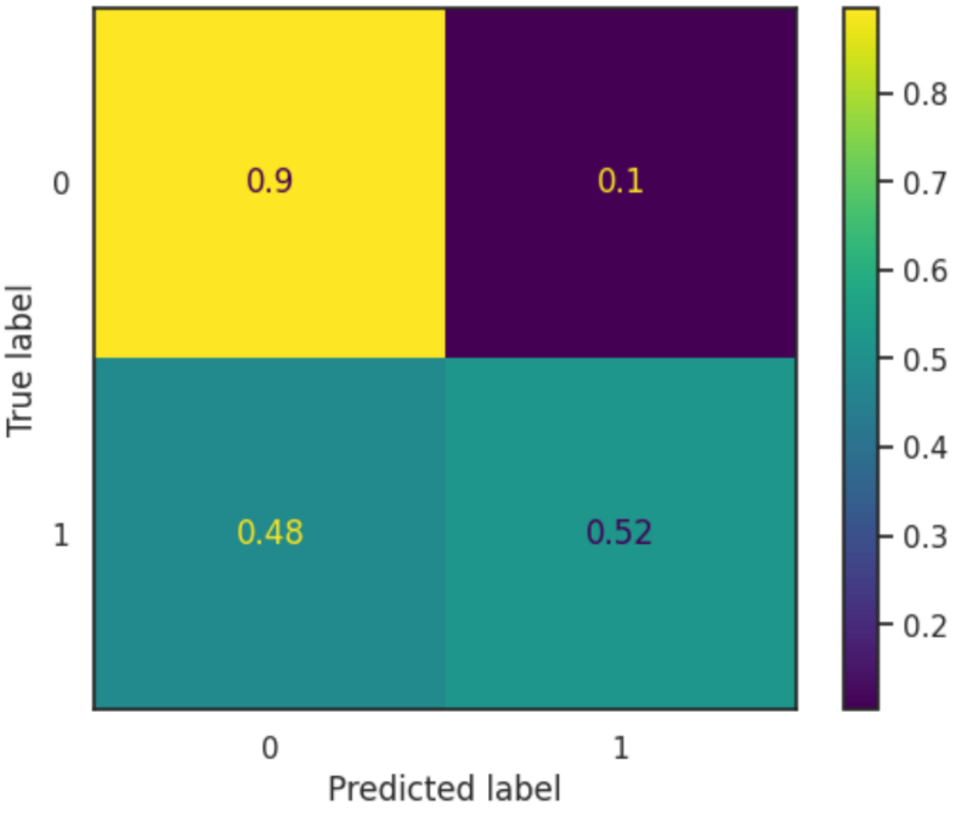
Now after Test Train split we apply models :

LOGISTIC REGRESSION :

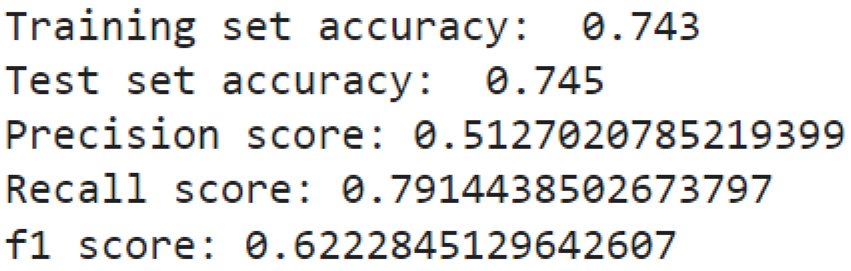
Without Balanced classes :



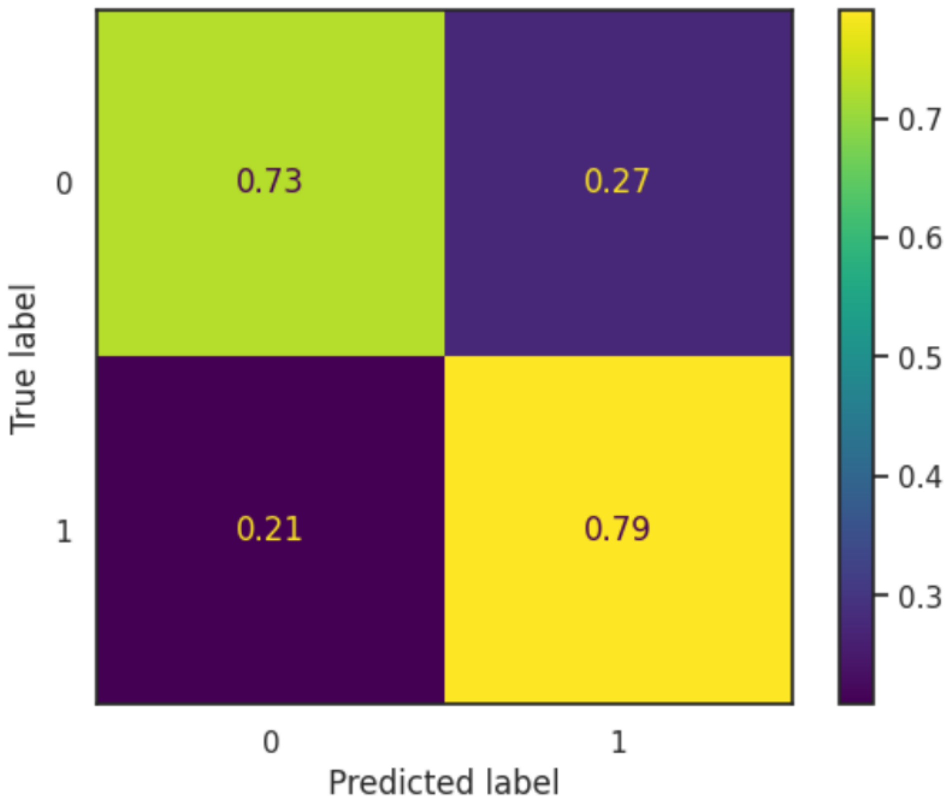
```
Training set accuracy:  0.802
Test set accuracy:  0.796
Precision score: 0.6452328159645233
Recall score: 0.5187165775401069
f1 score: 0.575098814229249
```



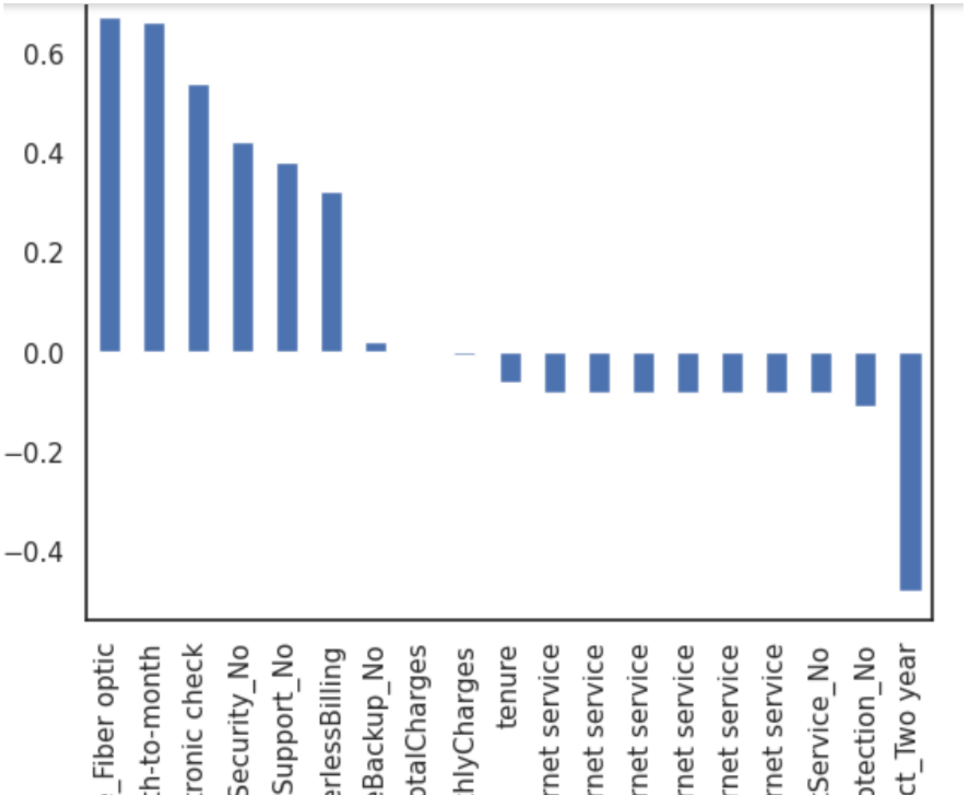
With Balanced Classes :



```
Training set accuracy:  0.743
Test set accuracy:  0.745
Precision score: 0.5127020785219399
Recall score: 0.7914438502673797
f1 score: 0.6222845129642607
```



0.73
0.27
0.21
0.79
0
1
True label
0
1
Predicted label
0.7
0.6
0.5
0.4
0.3



0.6
0.4
0.2
0.0
−0.2
−0.4
_Fiber optic
h-to-month
ronic check
Security_No
Support_No
erlessBilling
eBackup_No
otalCharges
hlyCharges
tenure
rnet service
rnet service
rnet service
rnet service
rnet service
rnet service
Service_No
otection_No
ct_Two year

Insights from Logistic Regression Results

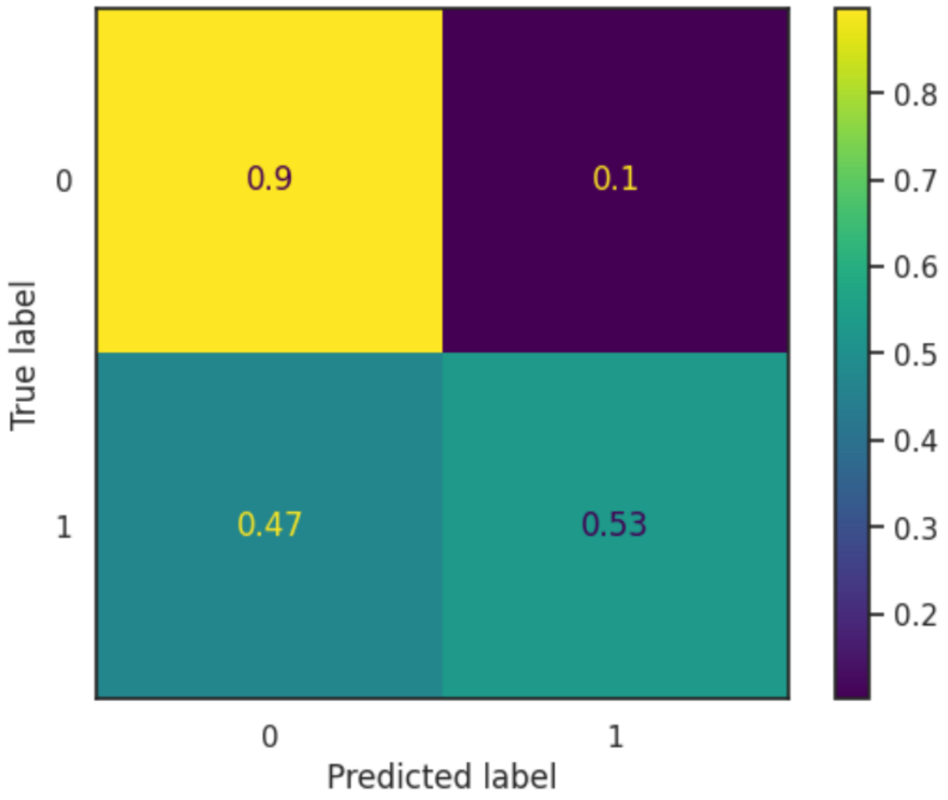
We can notice that balancing in the dataset results in significant increase in recall, while decrease in precision and accuracy.

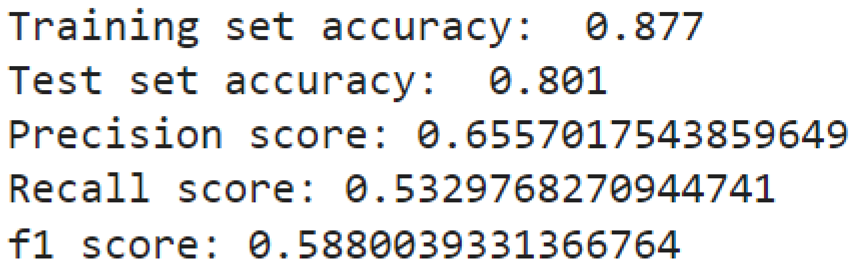
In my opinion, we can live with that trade. As if we have weak recall we might end up predicting someone who'll quit as no churn, which will affect the company more.

We can see from the coefficient plot the affect of different variables on our target variable i.e. 'Churn' which is in line with what insights we had after the EDA.

Random Forest

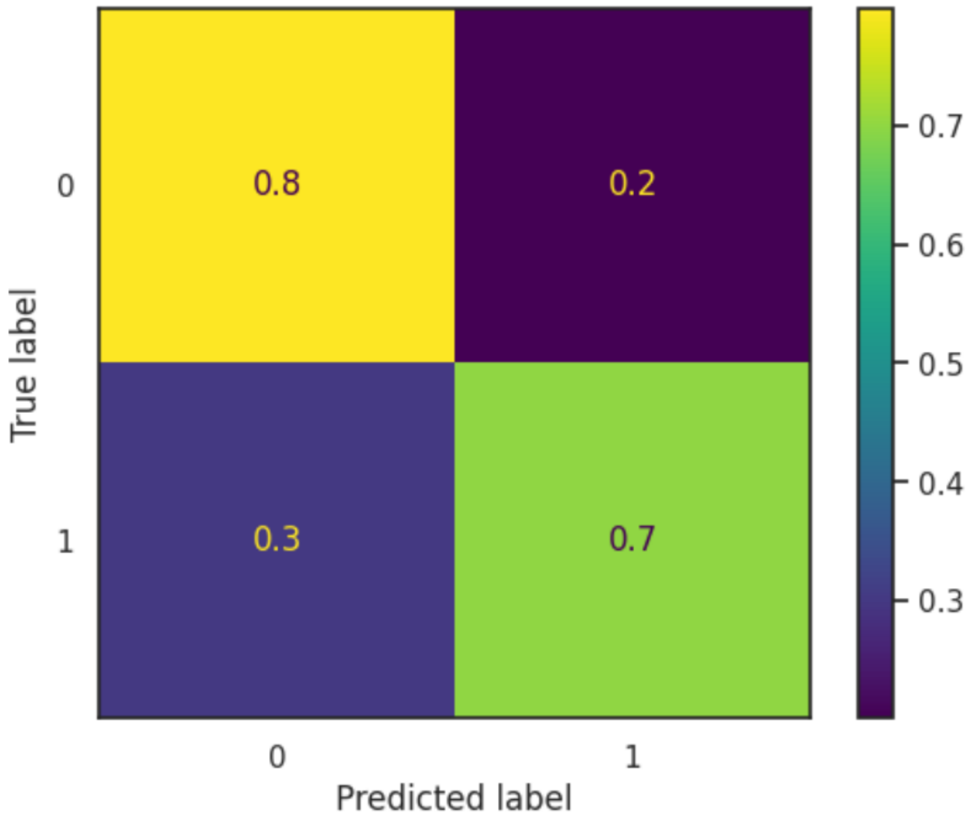
Without Balanced Data

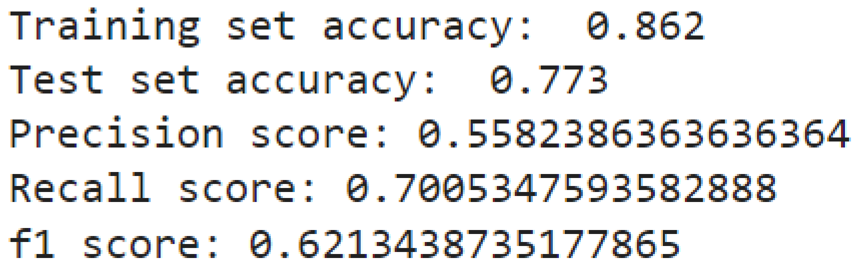




```
Training set accuracy:  0.877
Test set accuracy:  0.801
Precision score: 0.6557017543859649
Recall score: 0.5329768270944741
f1 score: 0.5880039331366764
```

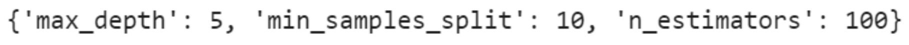
With Balanced Data



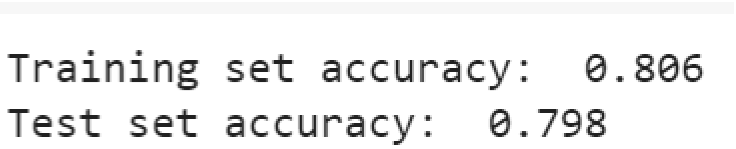


```
Training set accuracy:  0.862
Test set accuracy:  0.773
Precision score: 0.5582386363636364
Recall score: 0.7005347593582888
f1 score: 0.6213438735177865
```

Used Grid Search to find the best Parameters



```
{'max_depth': 5, 'min_samples_split': 10, 'n_estimators': 100}
```



```
Training set accuracy:  0.806
Test set accuracy:  0.798
```

Insights from Random Forest

- Random forest provided a better accuracy compared to the logistic regression model.
- Here also balancing the dataset results in significant increase of the recall score while a slight decrease in accuracy, and significant decrease in precision.

Conclusion

Balancing of the data results in significant improvement in the model.

Considering all the results, random forest algorithm seems the best fit for the data looking at the recall value obtained there.

Tenure, Contract(month-to-month), Internet service are some of the variable that effect out target variable 'Churn' the most, so focus can be laid around them to make the strategies to reduce the churn.