Introduction:

Topic: Telecom company's customer churn

Business Problem:

Telecom companies offer many interesting plans for different customers and invest lot of money to attract new customers in the way of special plans and advertisements. After onboarding the new customers, it is very important to retain them for a longer duration of time. The telecom company is concerned about customers leaving the company's services, which means all the campaigning efforts and new enrollment offer spend to attract customers would be in vain. So, to overcome this issue and minimize the impact of customer churn, it would help to identify and proactively reach out to the company's customers who would possibly churn in a given month.

Data:

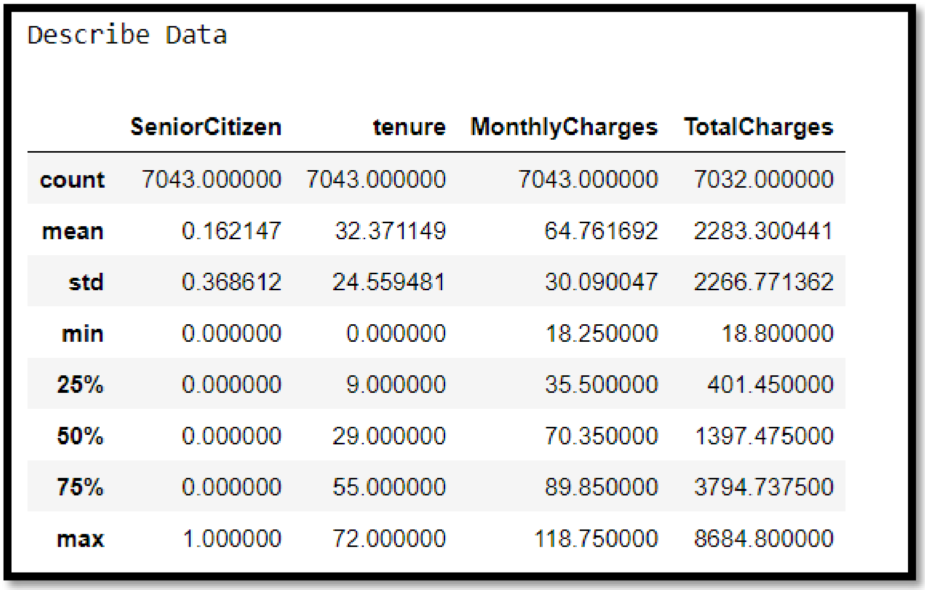
Based on the features such as below, we will be working on identifying the potential customer churn

a) Services signed up such as – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
b) Customer account information – duration of customer relationship, contract, payment method, paperless billing, monthly charges, and total charges.
c) Customer demographic info – gender, age range, and if they have partners or dependents

This will help the telecom company representatives to identify the potential customer churn and rollout any special offers to retain them. To conduct this analysis, I will be using a data set from Kaggle located at https://www.kaggle.com/blastchar/telco-customer-churn. This dataset is shared by "IBM Sample Data Sets".

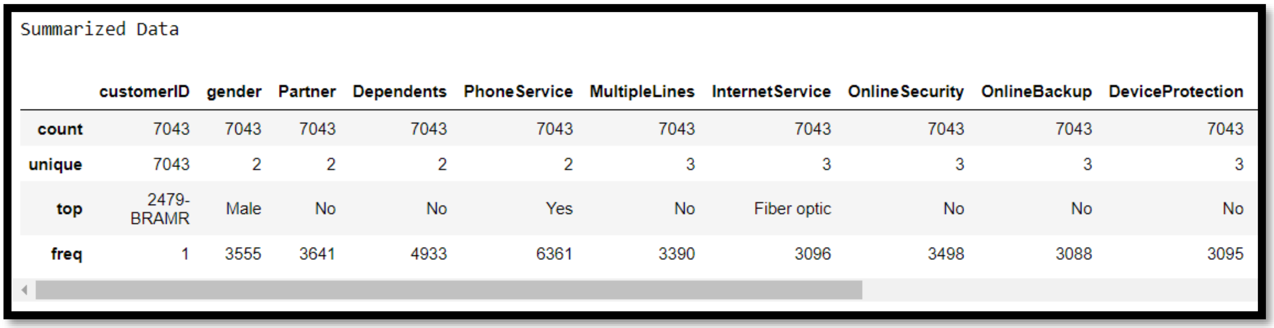
1. Customer ID: This variable is a combination of numeric and alphabetic values
2. gender: Consists of values as Male or Female to indicate the customer gender
3. SeniorCitizen: Indicates whether a customer is a senior citizen or not and holds values as 1 or 0
4. Partner: Whether the customer has a partner or not – values as Yes or No
5. Dependents: Whether customer has a partner or not – values as Yes or No)
6. tenure: Number of months the customer has stayed with the company – integer / numeric value
7. PhoneService: Whether the customer has a phone service or not – values as Yes, No
8. MultipleLines: Whether the customer has multiple lines or not – values as Yes, No, No phone service
9. InternetService: Customer's internet service provider – values as DSL, Fiber optic, No
10. OnlineSecurity: Whether the customer has online security or not – values as Yes, No, No internet service
11. OnlineBackup: Whether the customer has online backup or not – values as Yes, No, No internet service

12. DeviceProtection: Whether the customer has device protection or not – values as Yes, No, No internet service
13. TechSupport: Whether the customer has tech support or not – values as Yes, No, No internet service
14. StreamingTV: Whether the customer has streaming TV or not – values as Yes, No, No internet service
15. StreamingMovies: Whether the customer has streaming movies or not – values as Yes, No, No internet service
16. Contract: The contract term of the customer – values as Month-to-month, One year, Two year
17. PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
18. PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
19. MonthlyCharges: The amount charged to the customer monthly. This is a float / numeric value
20. TotalCharges: The total amount charged to the customer. This is a float / numeric value
21. Churn: Whether customer churned or not



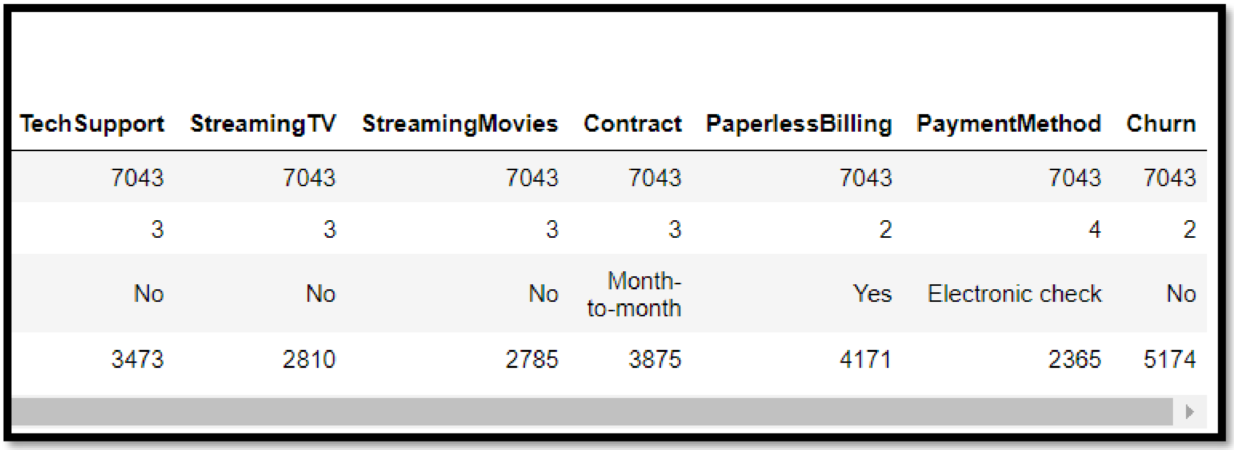
```
Describe Data
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7032.000000
mean	0.162147	32.371149	64.761692	2283.300441
std	0.368612	24.559481	30.090047	2266.771362
min	0.000000	0.000000	18.250000	18.800000
25%	0.000000	9.000000	35.500000	401.450000
50%	0.000000	29.000000	70.350000	1397.475000
75%	0.000000	55.000000	89.850000	3794.737500
max	1.000000	72.000000	118.750000	8684.800000



```
Summarized Data
```

	customerID	gender	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
count	7043	7043	7043	7043	7043	7043	7043	7043	7043	7043
unique	7043	2	2	2	2	3	3	3	3	3
top	2479-BRAMR	Male	No	No	Yes	No	Fiber optic	No	No	No
freq	1	3555	3641	4933	6361	3390	3096	3498	3088	3095

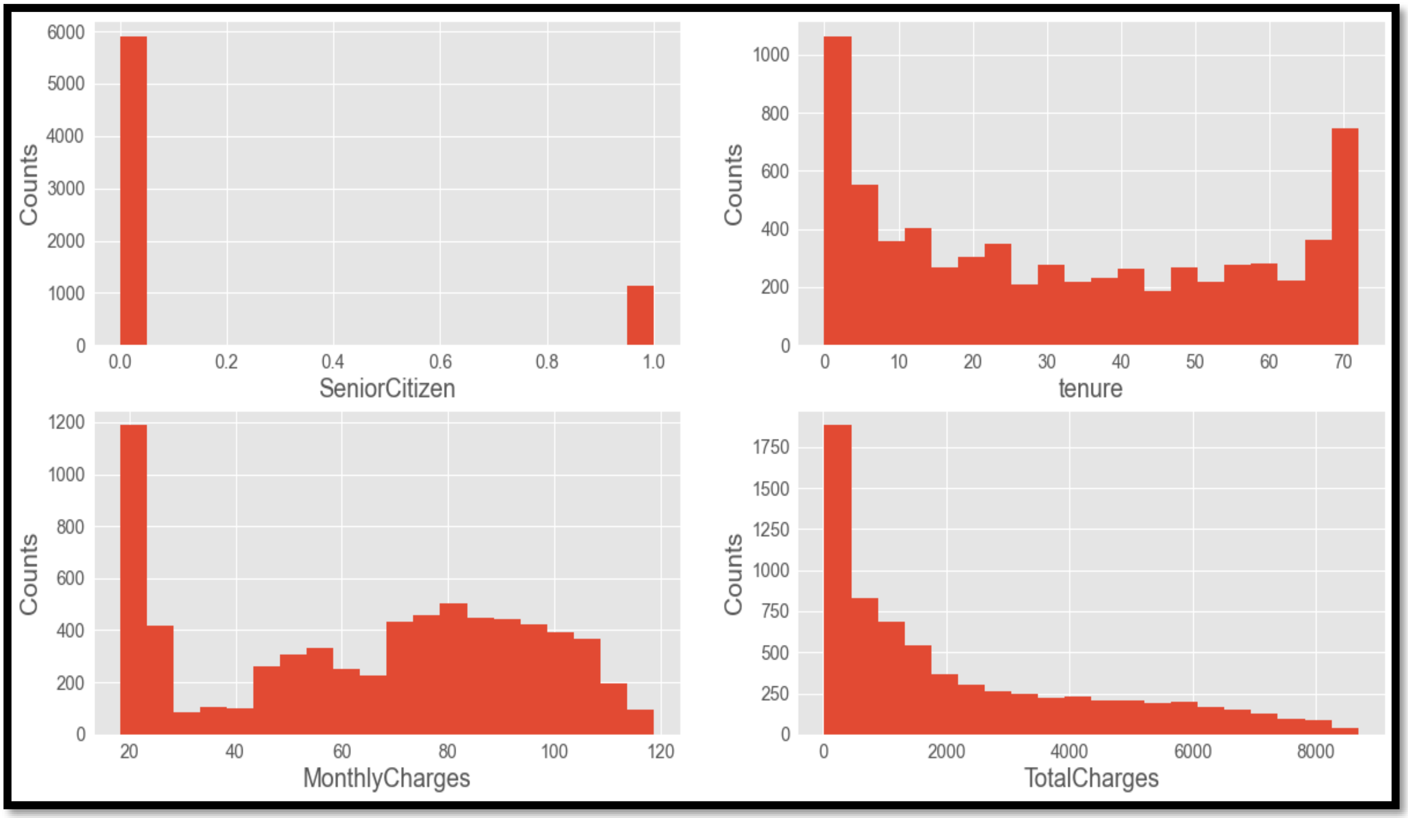


TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	Churn
7043	7043	7043	7043	7043	7043	7043
3	3	3	3	2	4	2
No	No	No	Month-to-month	Yes	Electronic check	No
3473	2810	2785	3875	4171	2365	5174

By using the prediction model, Telecom company can design offers / plans / occasional discounts to existing customers to ensure they remain with the company services for longer duration.

▪ Graphical Analysis:

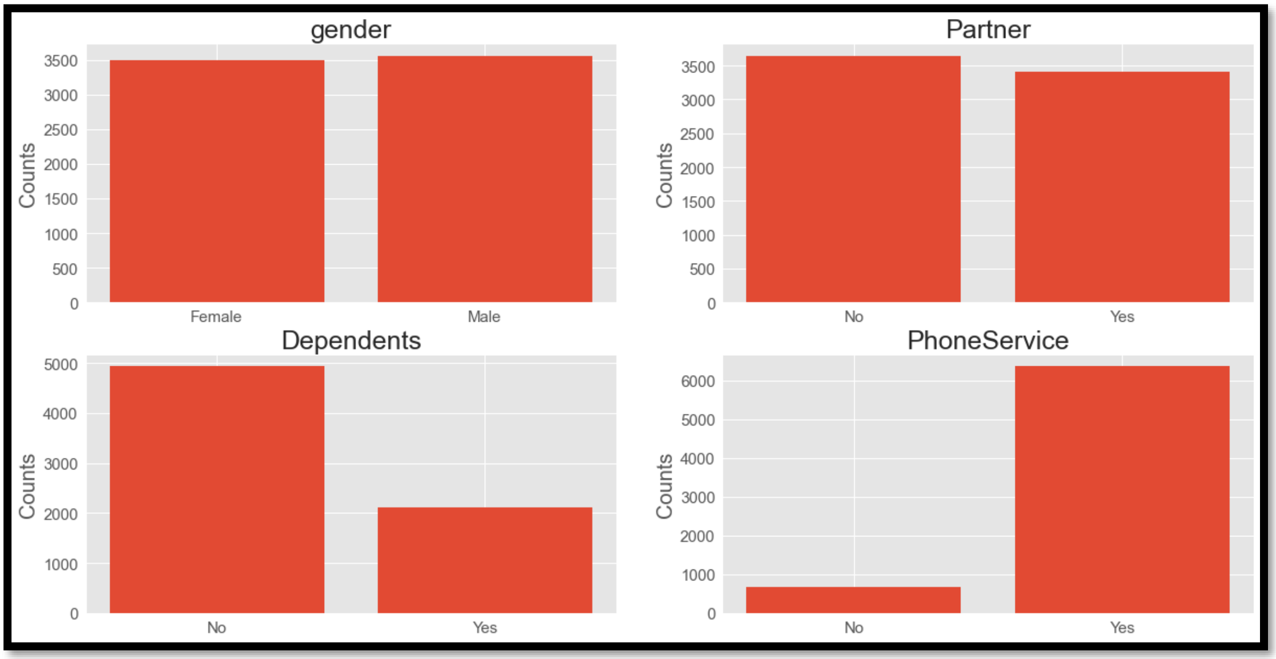
- **Analysis of the numerical features:**



Observations:

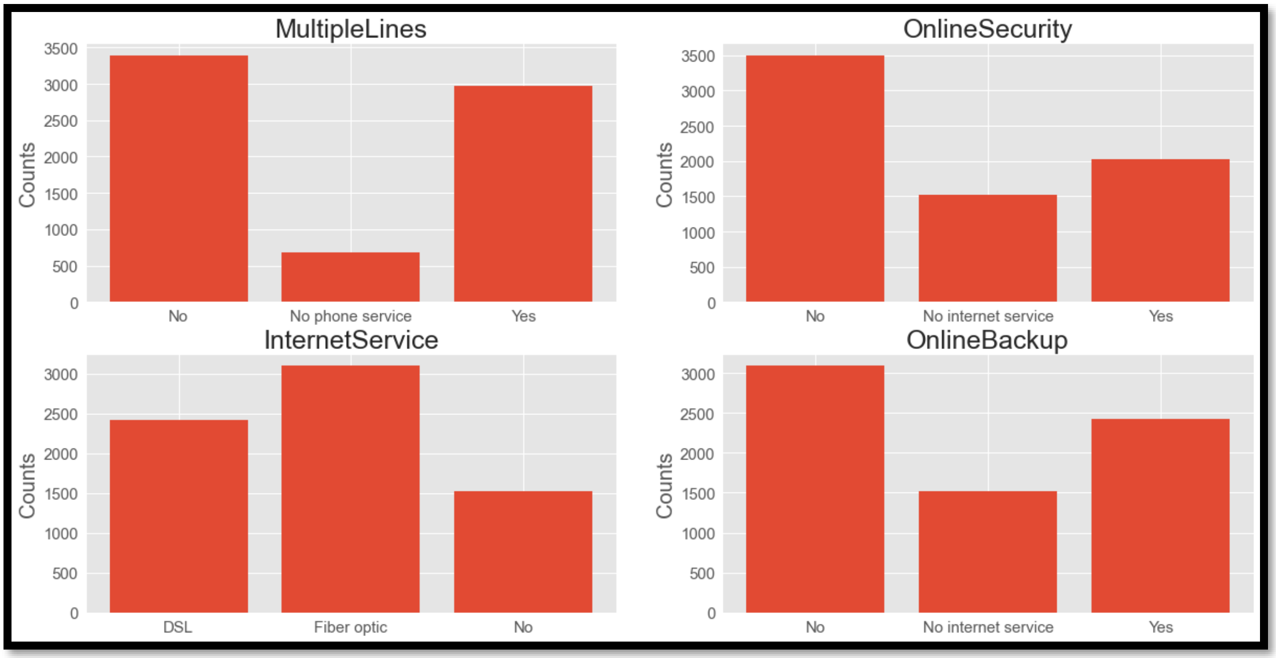
1) Histogram of SeniorCitizen variable shows majority of the customers are non SeniorCitizens
2) Histogram of tenure variable shows majority of the customers are under 10 months and above 65-70 months
3) Histogram of MonthlyCharges variable shows majority of the customers either below 30 dollars or between 70 - 90 dollars
4) Histogram of TotalCharges variable shows majority of the customers are near the lower range of 0 - 2000 dollars (around 500 dollars)

- **Analysis of the Categorical features:**



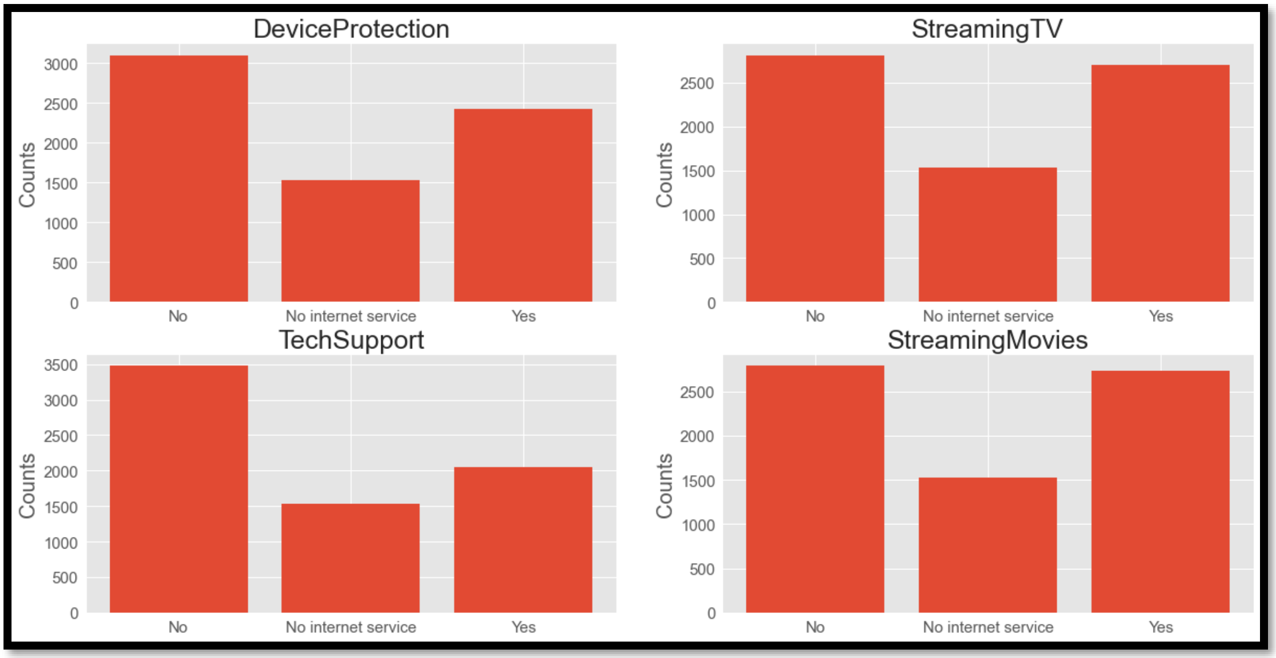
Observations:

1) Almost equal number of Male and Female customers
2) More number of customers with no Partners than customers with Partners
3) More number of customers with no Dependents than customers with Dependents
4) More number of customers with PhoneService as compared to customers with No PhoneService



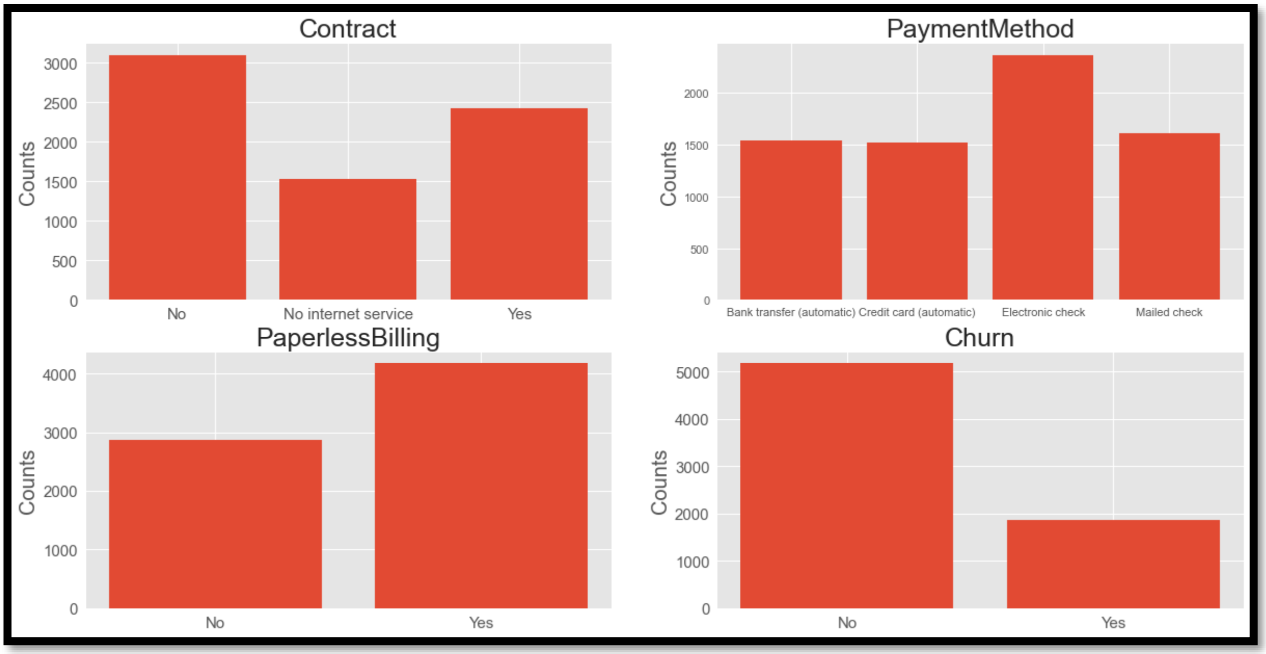
Observations:

1) More customers who do NOT have Multiple lines followed by those with Multiple lines and No Phone service customer at last
2) More customers who do NOT have Online Security followed by those with Online Security and No internet service customer at last
3) More number of customers with Fiber Optic connection followed by those with DSL connection and No InternetService customers at last
4) More customers who do NOT have Online Backup followed by those with Online Backup and No internet service customer at last



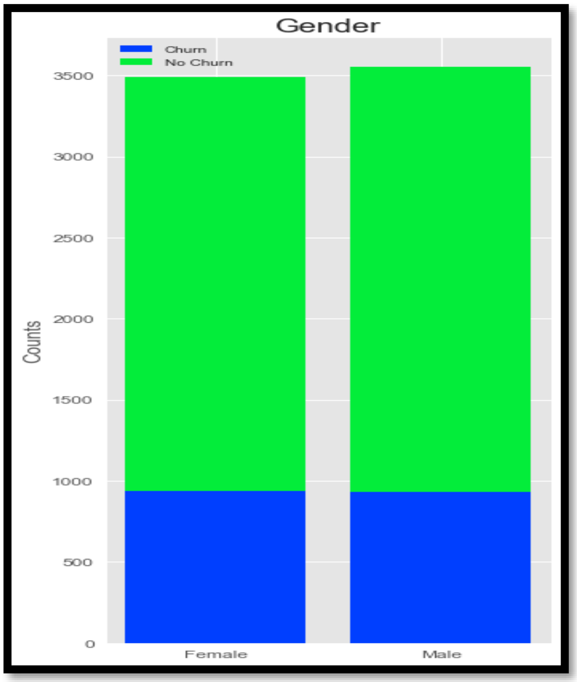
Observations:

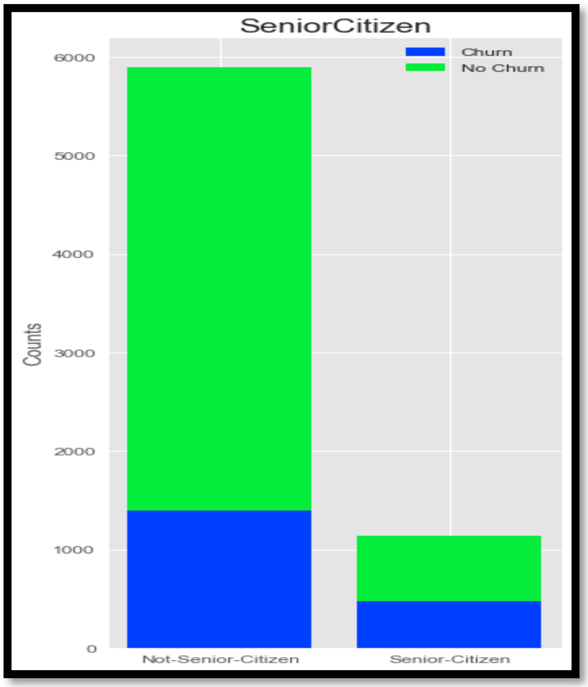
1) More customers who do NOT have Device Protection followed by those with Device Protection and No internet service customer at last
2) More customers who do NOT have Streaming TV followed by those with Streaming TV and No internet service customer at last
3) More number of customers with Tech Support followed by those with Tech Support and No InternetService customers at last
4) More customers who do NOT have Streaming Movies followed by those with Streaming Movies and No internet service customer at last

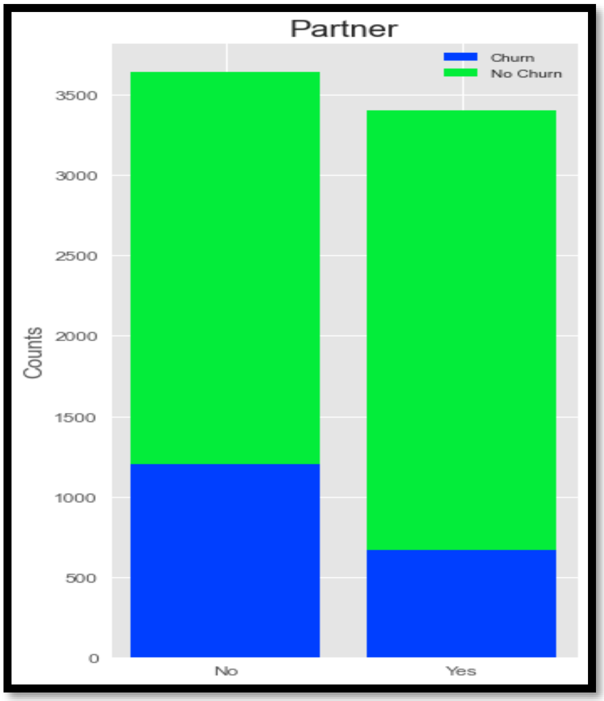


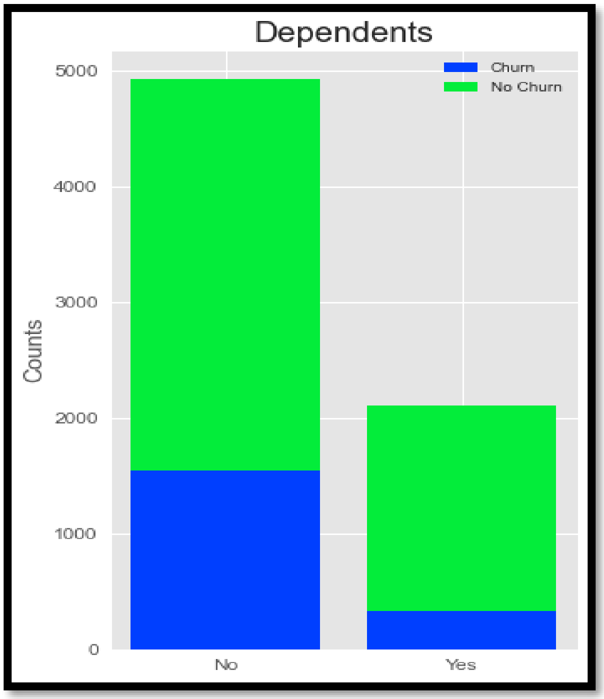
Observations:

1) More number of customers who are NOT under the Contract followed by Customers with Contract
2) More number of customers with Electronic check followed by people using Mailed Check for the payments then payment methods of by Bank Transfer and Credit Card appear
3) number of Customers with Paperless billings is more than number of people who did not have Paperless Billing
4) As we can see, number of Customers Churning is less than number of people who did not Churn





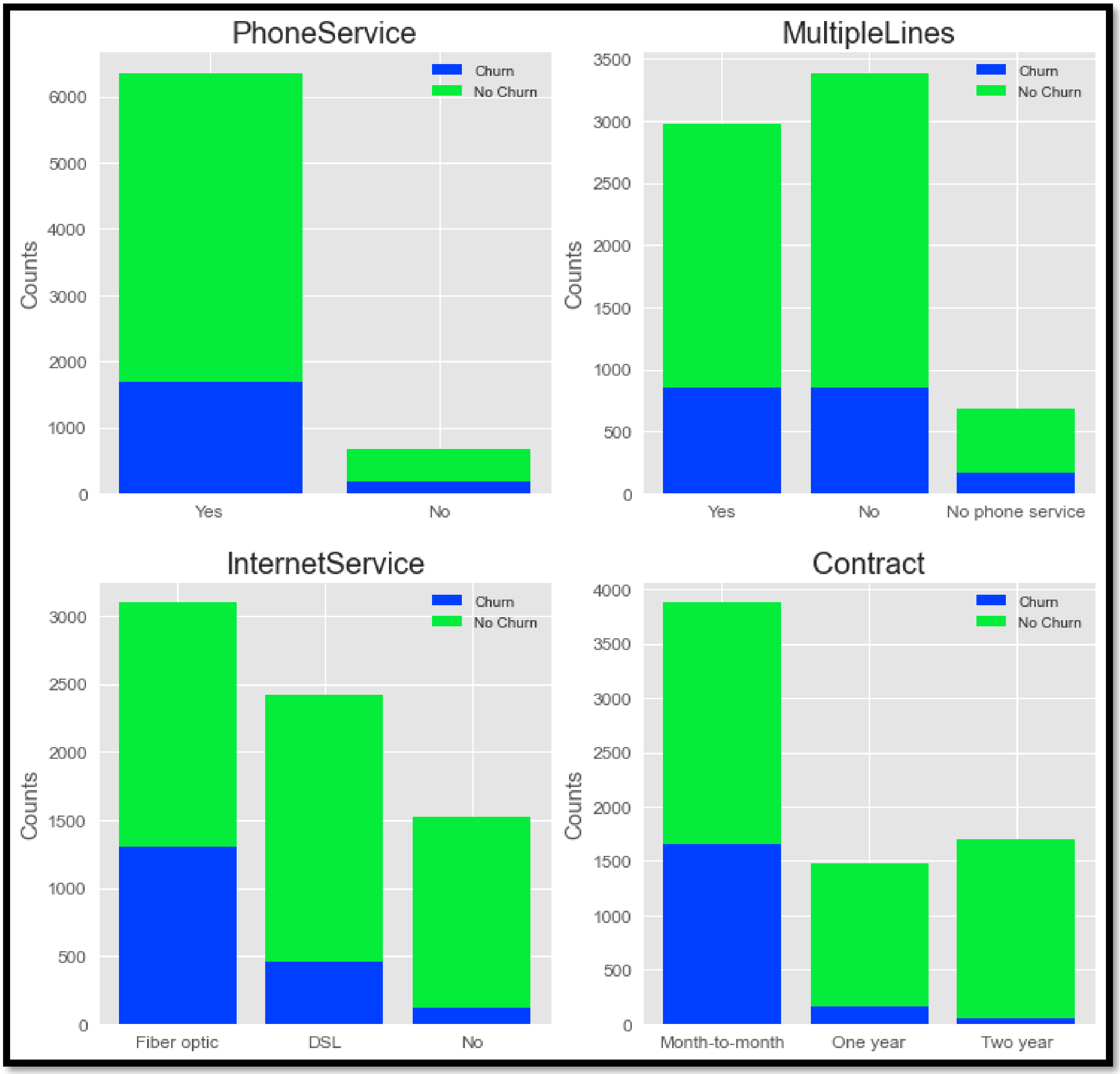




-

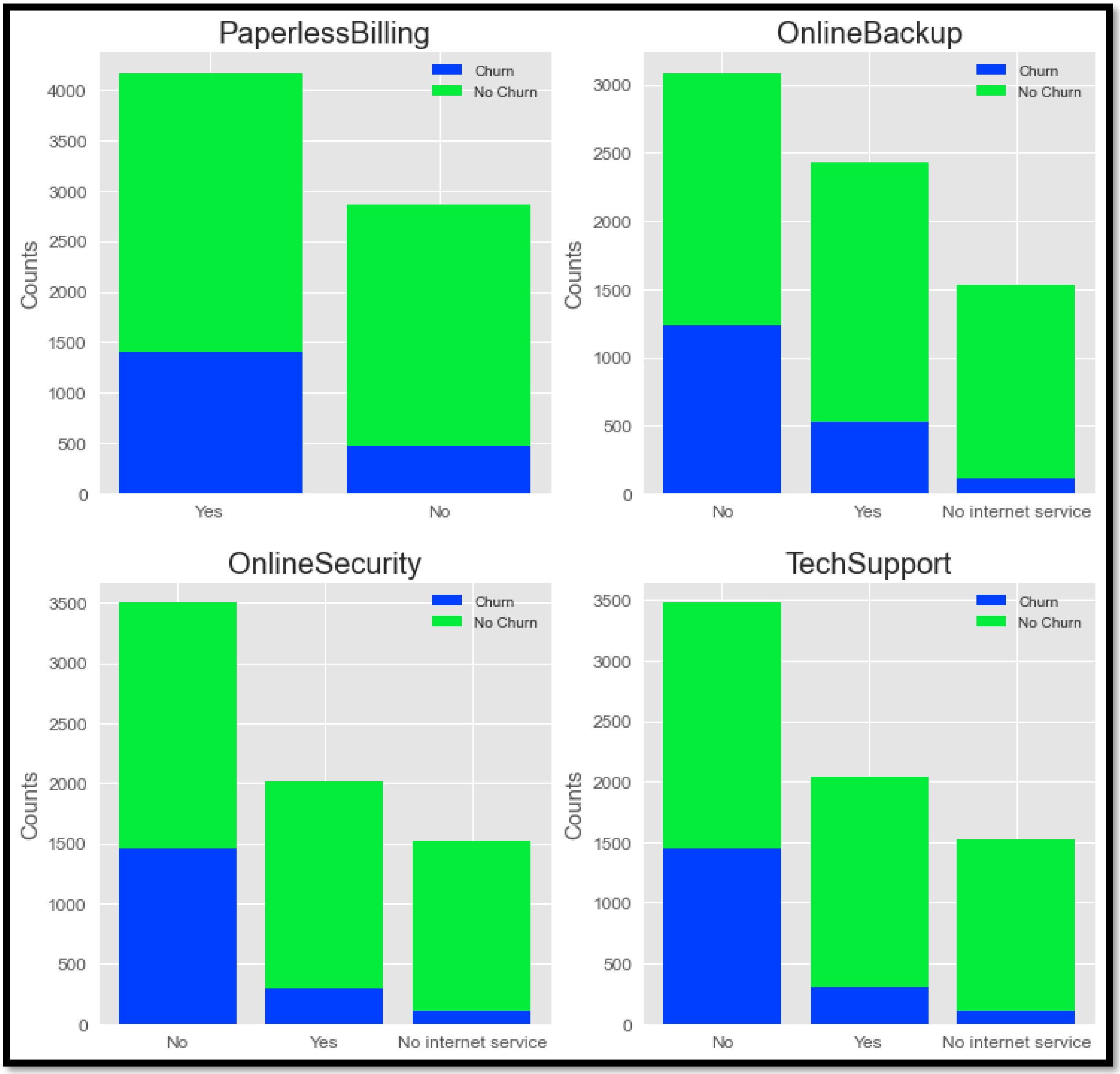
Observations:

1) Slightly higher percentage of Female customers tend to churn more as compared Male customers
2) Higher percentage of Non Senior Citizen customers tend to churn more as compared to Senior Citizen customers
3) Higher percentage of customers Not having Partner tend to churn more
4) Higher percentage of customers with No dependents tend to churn more



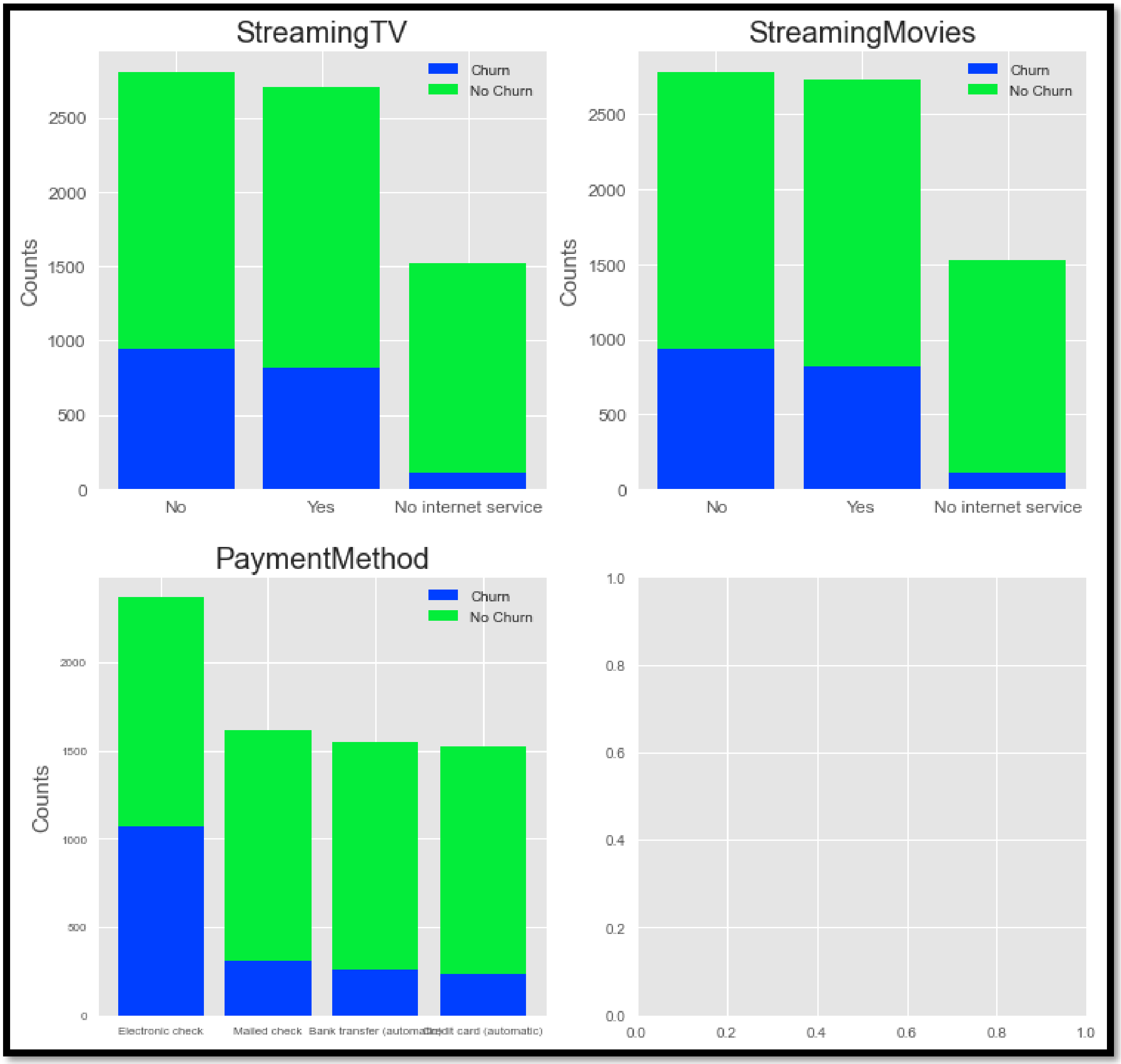
Observations:

1) It appears higher percentage of customers having Phone Services Churn more than customers having Phone Services
2) Higher percentage of having Multiple Lines Churn more than customers not having Multiple Lines
3) Higher ratio of customers having Fiber Optic Internet Service tend to churn more as compared to customers with DSL and no internet service
4) Higher ratio of customers on Month-to-Month contract churn more as compared to customers on One year contract and the least amount of customers under tow year contract churn.



Observations:

1) Customers with Paperless Billing tend to churn more as compared with Paper Billing customers
2) Customers Not using Online Backup services tend to churn more as compared to others using Online Backup / No internet services
3) Customers Not using Online Security services tend to churn more as compared to others using Online Security / No internet services
4) Customers Not using Tech Support services tend to churn more as compared to others using Tech Support / No internet services

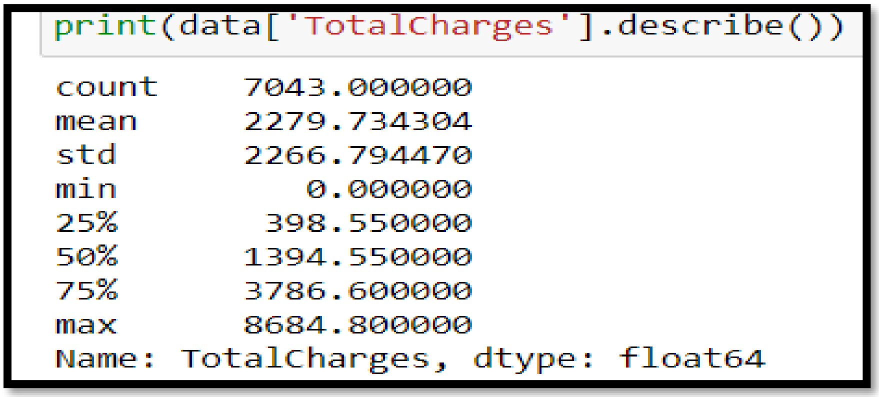


<u>Observations:</u>

1) Customers Not using StreamingTV services tend to churn more as compared to others using StreamingTV / No internet services
2) Customers Not using Streaming Movies services tend to churn more as compared to others using Streaming Movies / No internet services
3) Customers on Electronic Check payment method tend to churn more as compared to others using Mailed Check / Bank Transfer / Credit Card payment methods

Dimensionality, Feature Reduction and Feature Engineering:

- Fill in missing values



```
print(data['TotalCharges'].describe())

count    7043.000000
mean     2279.734304
std      2266.794470
min         0.000000
25%       398.550000
50%      1394.550000
75%      3786.600000
max      8684.800000
Name: TotalCharges, dtype: float64
```

- **Feature selection / Engineering:**

1) customerID field is not useful to determine any impacts as it does not have any logical relation with Customer Churn. So, not using it for next steps.
2) Looking at the visual analysis, gender category has almost similar percentages / ratio amongst the customer Churn / No Churn cases. So, it may not be very helpful for next steps.
3) SeniorCitizen column essentially consists of 1 / 0 values and appears to be categorical in nature. Also, the heatmap indicates that there is no numeric values relation of this variable with Churn / No Churn categories. So, including this variable as Categorical feature below
4) TotalCharges column appears to be related to tenure column i.e. higher the tenure, more amounts are reflected in TotalCharges column, looking at the data. So, we will exclude TotalCharges column as redundant one.
5) From the visual analysis, Rest all categorical columns / values within them, appear to show some relation with Churn / No Churn. So, I will be using them in as input features.

One Hot Encoding:

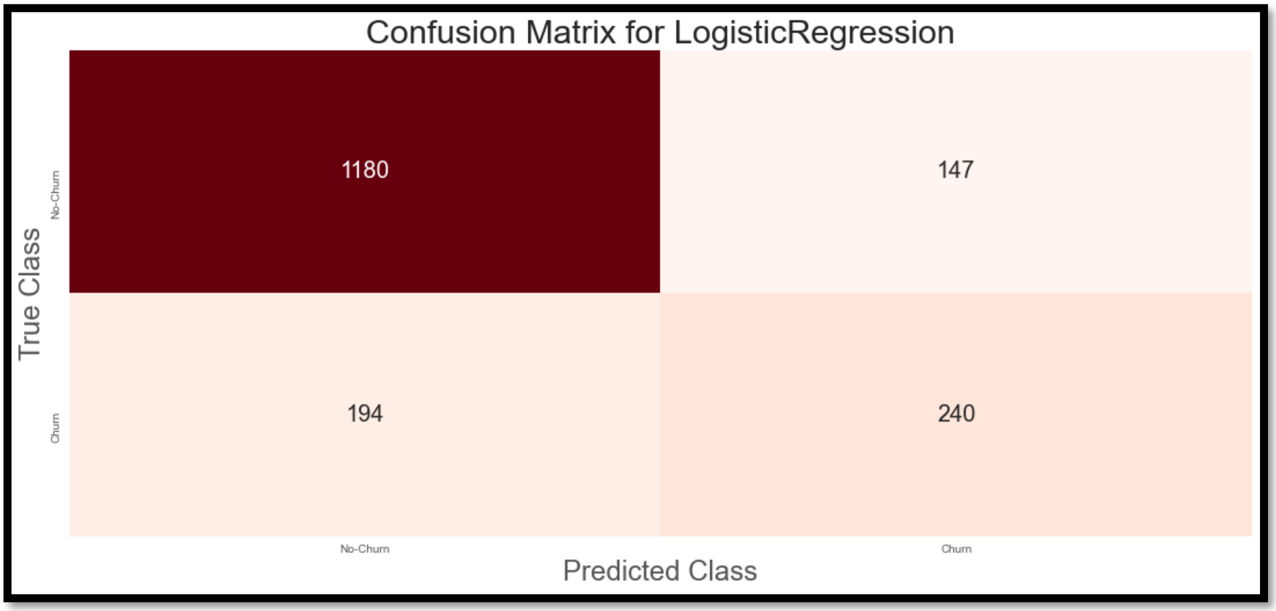
```
   SeniorCitizen  Partner_No  Partner_Yes  Dependents_No  Dependents_Yes  
\
0              0           0            1              1               0
1              0           1            0              1               0
2              0           1            0              1               0
3              0           1            0              1               0
4              0           1            0              1               0
5              0           1            0              1               0
6              0           1            0              0               1
7              0           1            0              1               0
8              0           0            1              1               0
9              0           1            0              0               1
```

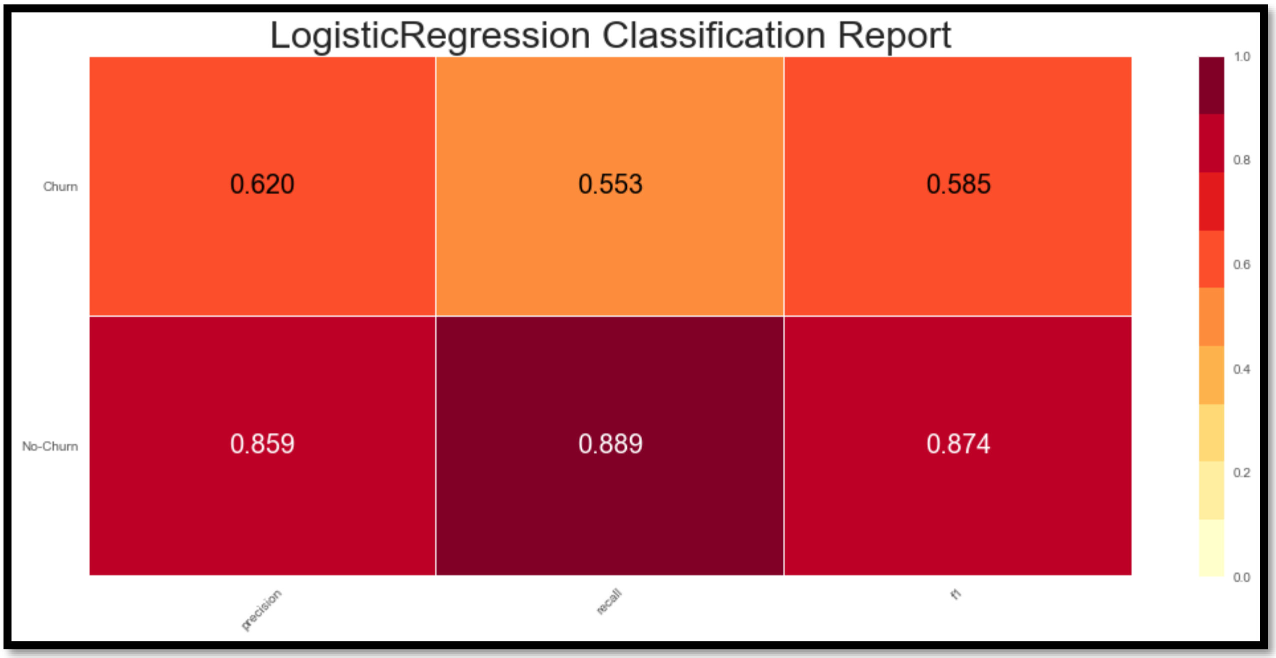
```
   PhoneService_No  PhoneService_Yes  MultipleLines_No  \
0                1                 0                 0   
1                0                 1                 1   
2                0                 1                 1   
3                1                 0                 0   
4                0                 1                 1   
5                0                 1                 0   
6                0                 1                 0   
7                1                 0                 0   
8                0                 1                 0   
9                0                 1                 1   

   MultipleLines_No phone service  MultipleLines_Yes  ...  \
0                               1                  0  ...   
1                               0                  0  ...   
2                               0                  0  ...   
3                               1                  0  ...   
4                               0                  0  ...   
5                               0                  1  ...   
6                               0                  1  ...   
7                               1                  0  ...   
8                               0                  1  ...   
9                               0                  0  ...   

   StreamingMovies_Yes  PaymentMethod_Bank transfer (automatic)  \
0                    0                                        0   
1                    0                                        0   
2                    0                                        0   
3                    0                                        1   
4                    0                                        0   
5                    1                                        0   
6                    0                                        0   
7                    0                                        0   
8                    1                                        0   
9                    0                                        1   

   PaymentMethod_Credit card (automatic)  PaymentMethod_Electronic check  
\
0                                      0                               1   
1                                      0                               0   
2                                      0                               0   
3                                      0                               0   
4                                      0                               1   
5                                      0                               1   
6                                      1                               0   
7                                      0                               0   
8                                      0                               1   
9                                      0                               0   
```

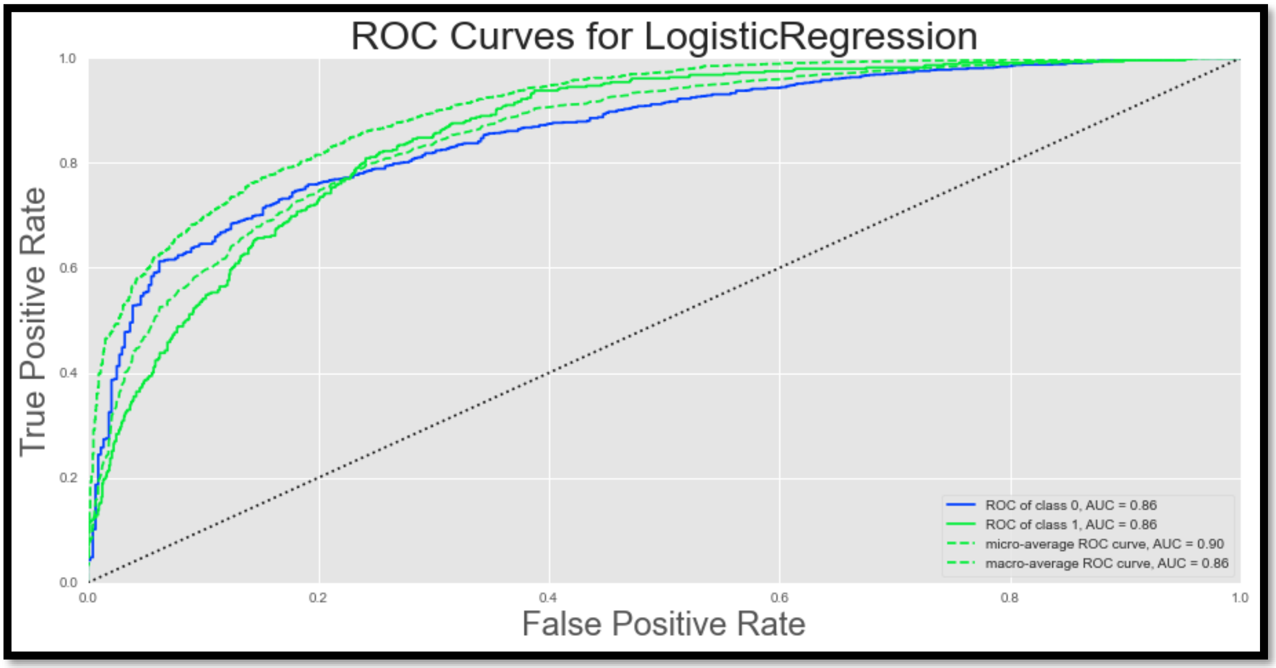
```
   PaymentMethod_Mailed check  PaperlessBilling_No  PaperlessBilling_Yes
\
0                           0                    0                     1
1                           1                    1                     0
2                           1                    0                     1
3                           0                    1                     0
4                           0                    0                     1
5                           0                    0                     1
6                           0                    0                     1
7                           1                    1                     0
8                           0                    0                     1
9                           0                    1                     0

   Contract_Month-to-month  Contract_One year  Contract_Two year
0                        1                  0                  0
1                        0                  1                  0
2                        1                  0                  0
3                        0                  1                  0
4                        1                  0                  0
5                        1                  0                  0
6                        1                  0                  0
7                        1                  0                  0
8                        1                  0                  0
9                        0                  1                  0
```

- **Model Selection and Evaluation:**

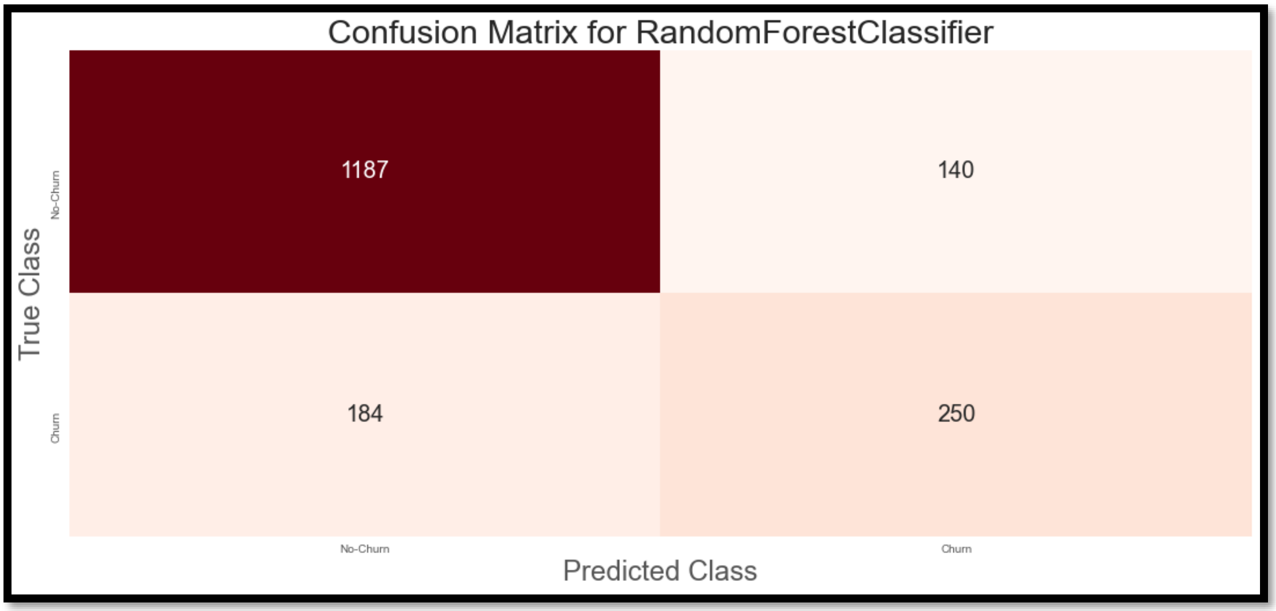
- **Logistic Regression:**

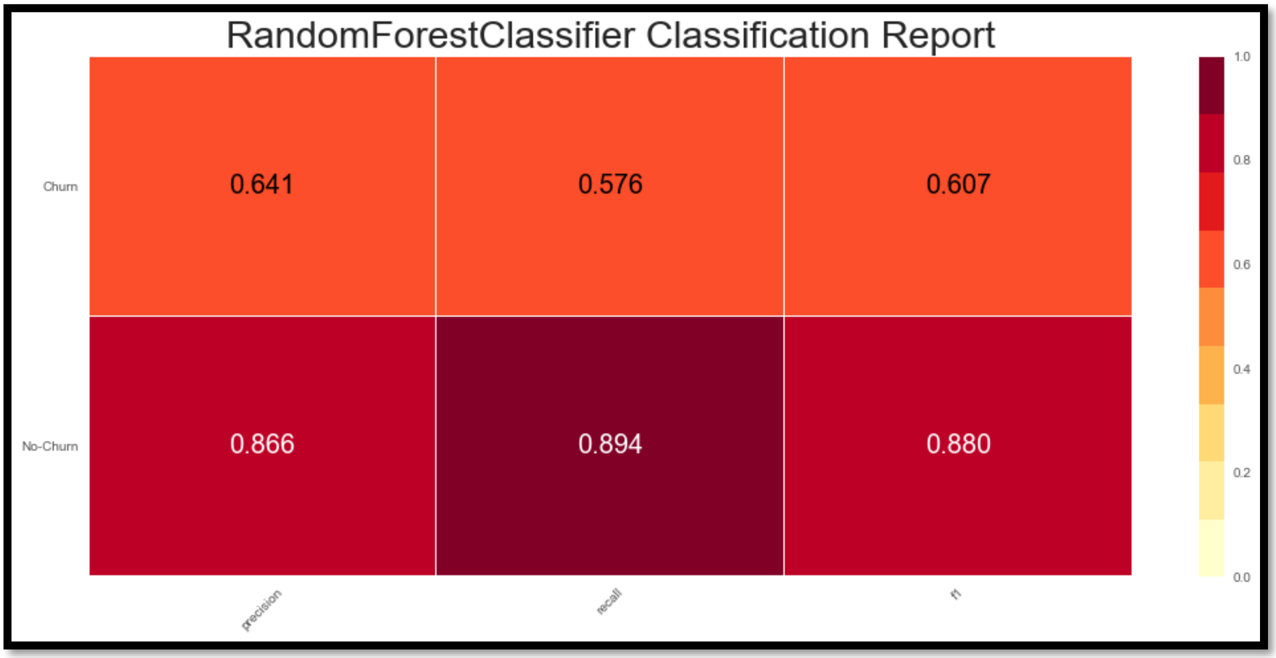


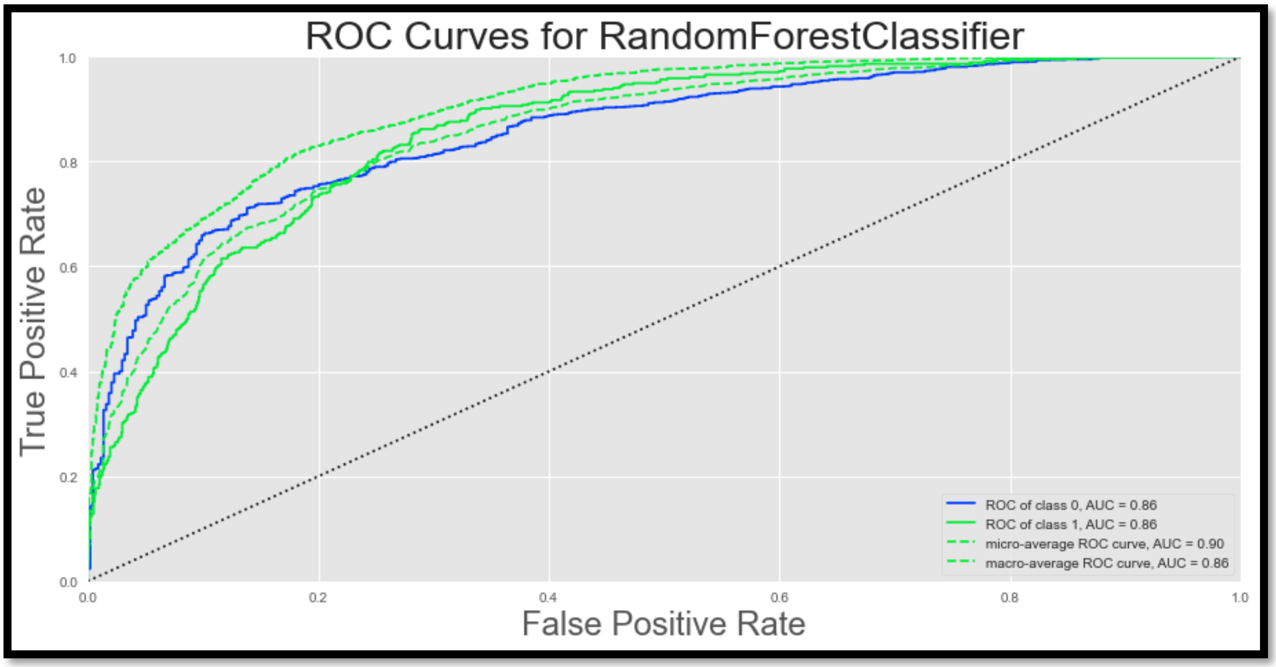

LogisticRegression Classification Report
Churn
No-Churn
0.620
0.553
0.585
0.859
0.889
0.874
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0


ROC Curves for LogisticRegression
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.86
ROC of class 1, AUC = 0.86
micro-average ROC curve, AUC = 0.90
macro-average ROC curve, AUC = 0.86

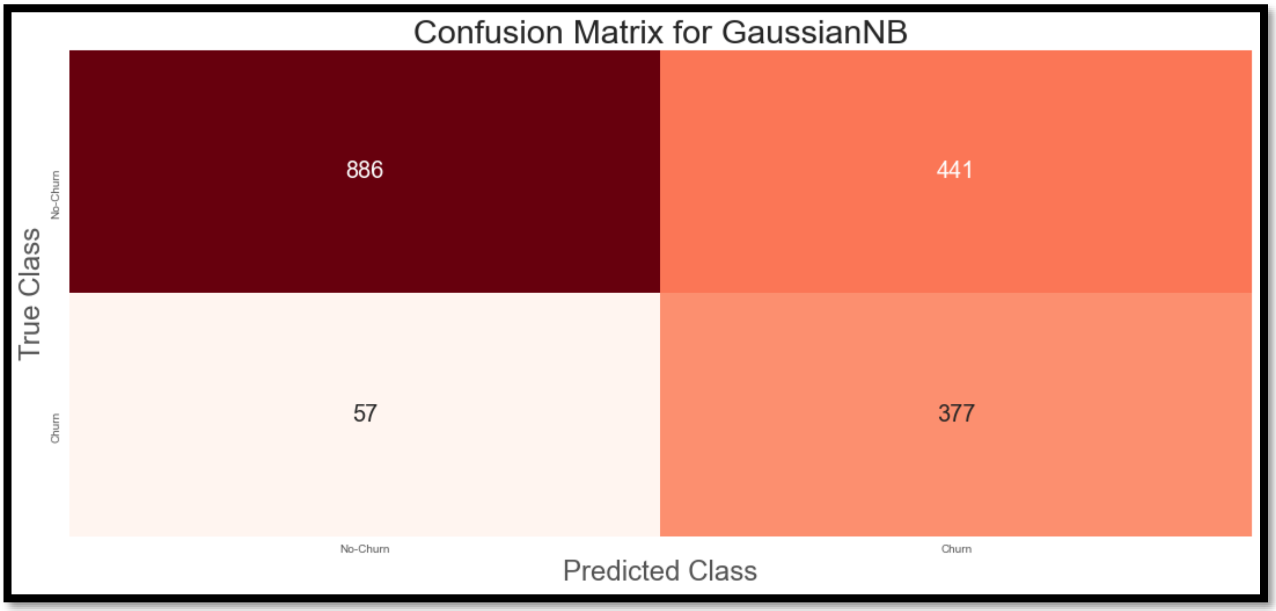
- **Random Forest Classifier:**

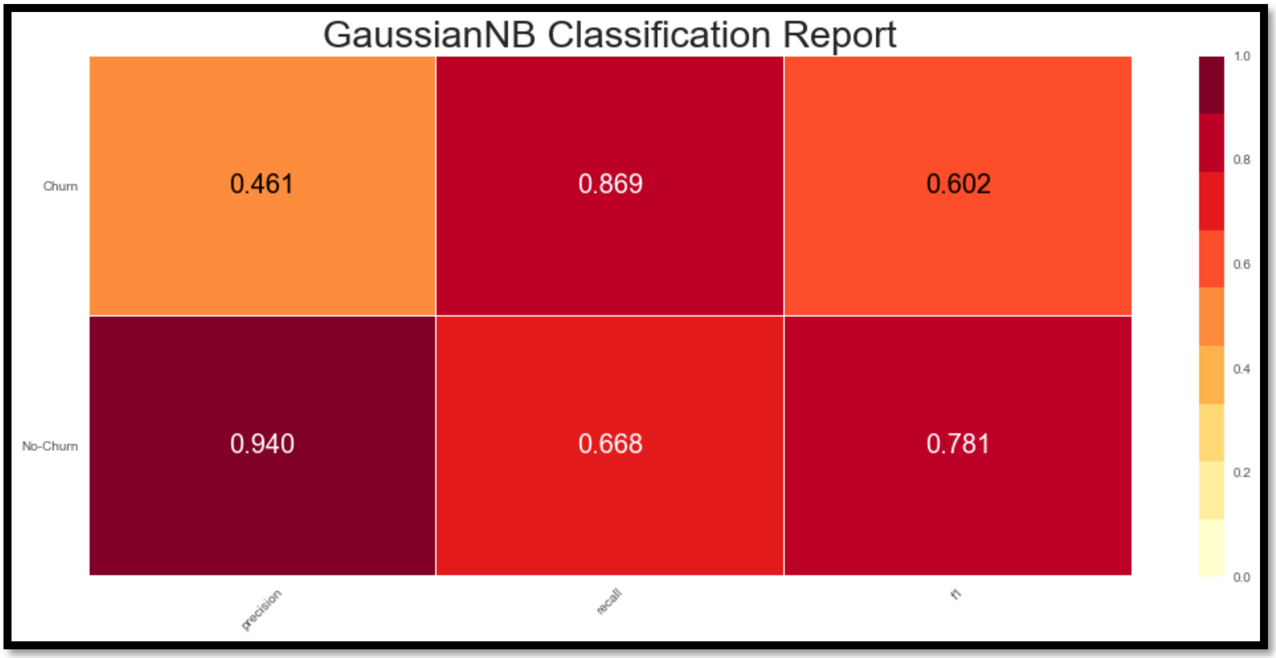


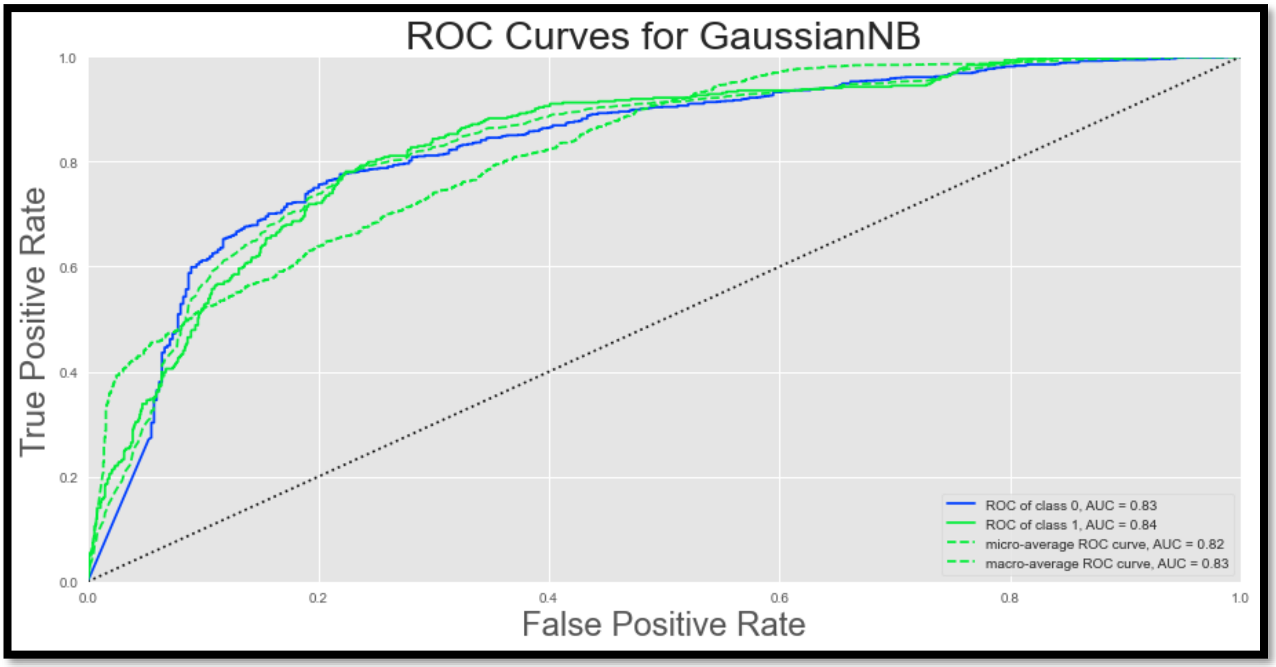




- **Gaussian Naïve Bayes:**




GaussianNB Classification Report
Churn
No-Churn
0.461
0.869
0.602
0.940
0.668
0.781
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0


ROC Curves for GaussianNB
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.83
ROC of class 1, AUC = 0.84
micro-average ROC curve, AUC = 0.82
macro-average ROC curve, AUC = 0.83

Observations:

Looking at the above results, we can interpret following things

1) Logistic Regression Model has around 80.4% and 80.2% accuracy rates on Training and Test datasets respectively
2) Random Forest Tree Model has around 86.2% and 80.0% accuracy rates on Training and Test datasets respectively
3) Gaussian Naive Bayes Model has around 70.0% and 69.9% accuracy rates on Training and Test datasets respectively

So, it appears that both Logistic Regression and Random Forest Tree models have similar accuracy levels on Test datasets - around 80%. So, using either of them would be okay. However, Gaussian Naive Bayes hovers around 70%. So, we may avoid using it.

Also looking at the above Confusion matrices, the numbers for the Random Forest Tree model seems to be slightly better on all fronts as compared to Logistic Regression model.
e.g.
No churn customers are accurately predicted at 1187 vs. 1180 (Log Reg).
Churn customers are accurately predicted at 250 vs. 240 (Log Reg).
Customers who are predicted as No Churn, but belong to Churn category in reality as at 184 vs 194 (Log Reg) - lowered number in inaccuracy.
Customers who are predicted as Churn, but belong to No Churn category in reality as at 140 vs 147 (Log Reg) - lowered number in inaccuracy.

There is room for improvement in the prediction outcomes. Probably with additional Train / Test data, balancing the Churn / No Churn classes data and trying out additional models can help improve the predictions on customer churn.

▪ Conclusion:

Amongst the models implemented and evaluated, Random Forest method appeared to produce better results on all aspects of the predictions for the given data. With False Negative (Churn customers being predicted as potentially No Churn) rate of 10.45%, Random Forest method produces better results as compared with other models. Which means we might still miss out about 10% + potential churn customers, but are better equipped with correctly predicting more than 81% of the customers accurately. It appears that customers who are senior citizens, and those having dependents, partners as well as using internet, online security, multiple lines or streaming services tend stay longer with the company's services. So, potentially we could look to identify the customers using only cell phone / single services and offer additional services like internet service, multiple lines services, online security or streaming services which would benefit them. Especially in the current times of COVID-19, most of the customers are working remotely / from their home. So, offering the multiple product features of internet services, online security / support, streaming services they could enjoy and will encourage them to remain with the company services for the longer term duration.