Customer Churn Prediction

Name: Nafiz Mahamud (ID: 5211418)

Introduction: Customer Churn prediction means knowing which customers are likely to leave or unsubscribe from a service. For many companies, this is an important prediction because acquiring new customers often costs more than retaining existing ones. A company with a high churn rate loses many subscribers, resulting in lower growth rates and a greater impact on sales and profits. Companies with low churn rates can retain customers. Once we have identified customers at risk of churn, we need to know exactly what and when marketing efforts we should make with each customer to maximize their likelihood of staying.

For those reasons, I am going to predict whether a customer is likely to Churn (cancel or choose to renew the subscription) within one month based on several factors such as phone service, internet service, gender, tenure, etc, which can help in implementing retention strategies. The dataset I got, provided by Honorable **Prof. Dr. Aysegul Cuhadar** named 'churn.csv', which contains 7043 rows with 21 columns. I have analyzed the dataset carefully like handling missing values, and visualization to get a better understanding. Then, I have split the dataset into training and test sets. After that, I built a couple of classification models and finally evaluated them with the confusion matrix.

Data Analysis:

I look at columns and find that "TotalCharges" column which should be float data type surprisingly was Object data type. So that, first I convert it to the float data type. I fill up the missing values with average.

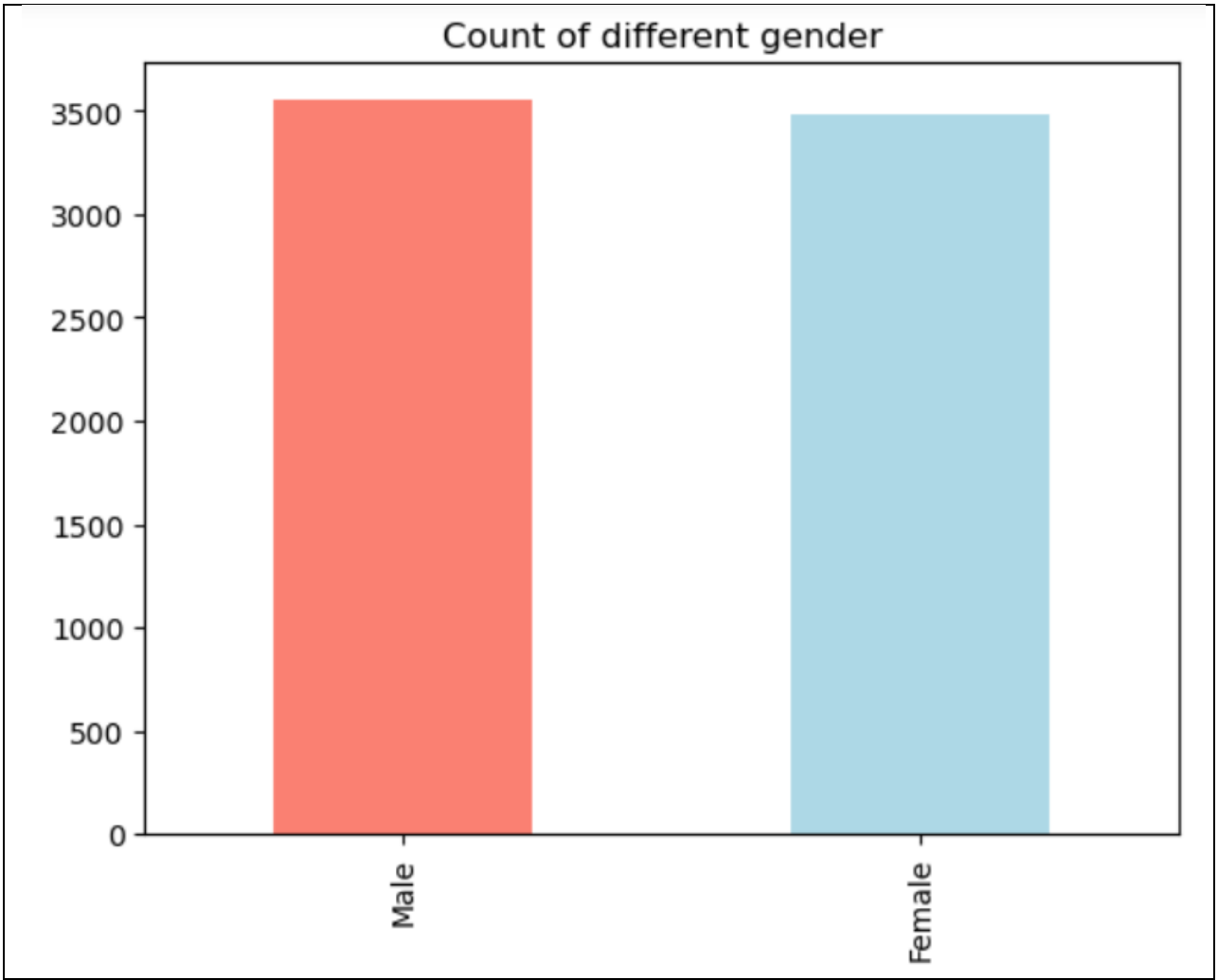
Description of Columns: The data set includes information about:

• Customers who left within the last month – the column is called Churn.

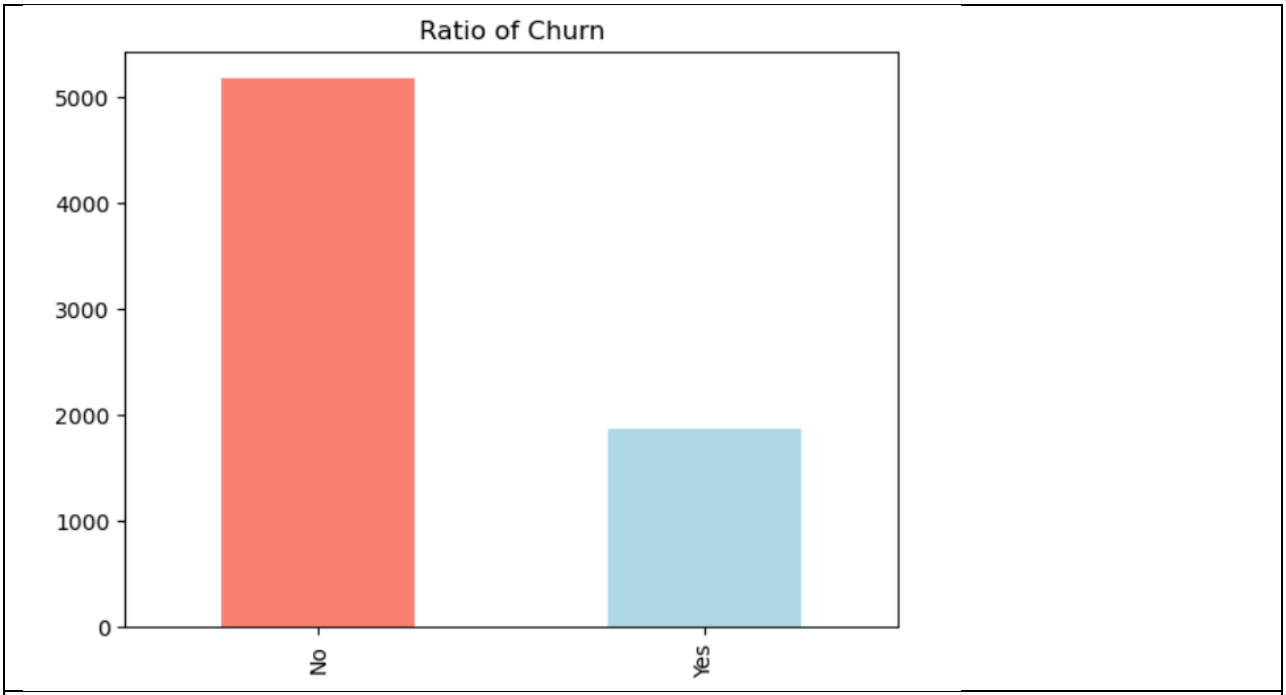
• Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.

• Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.

• Demographic info about customers – gender, age range, and if they have partners and dependents.

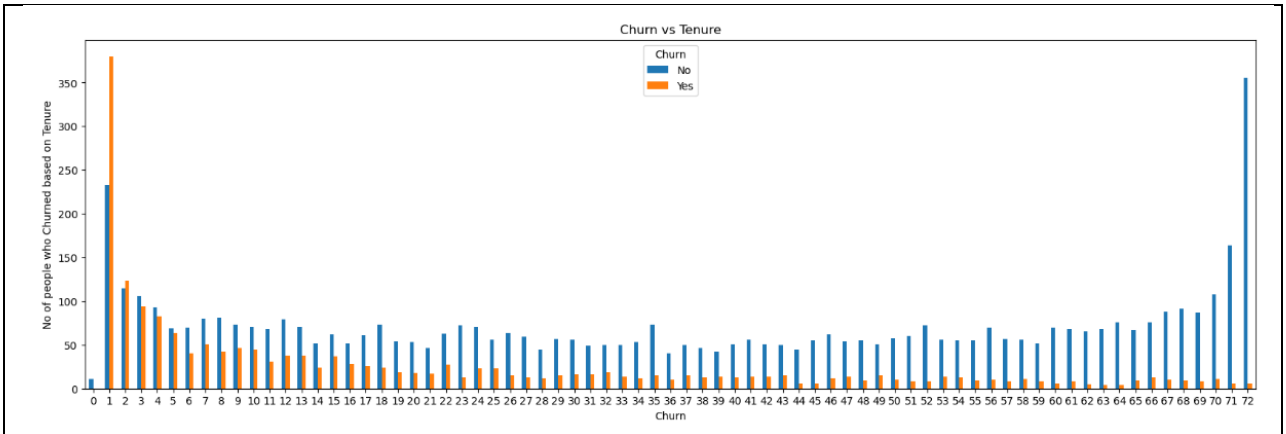


From the ratio, I see that both male and female are interested in taking the service from the company. There is no significant difference in gender.

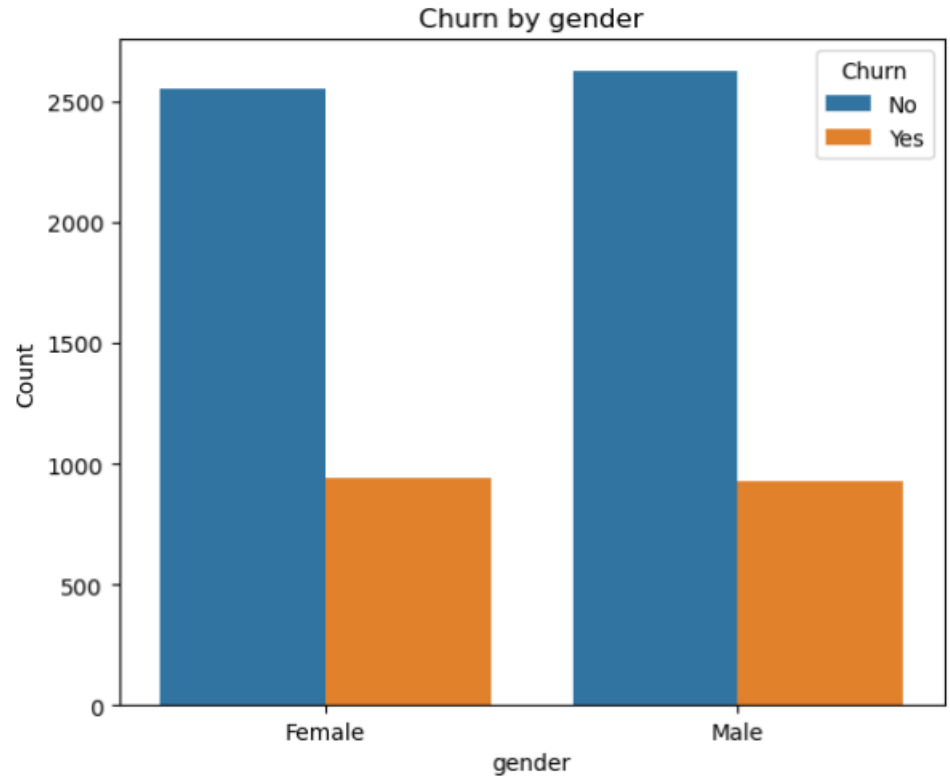


No 5174
Yes 1869

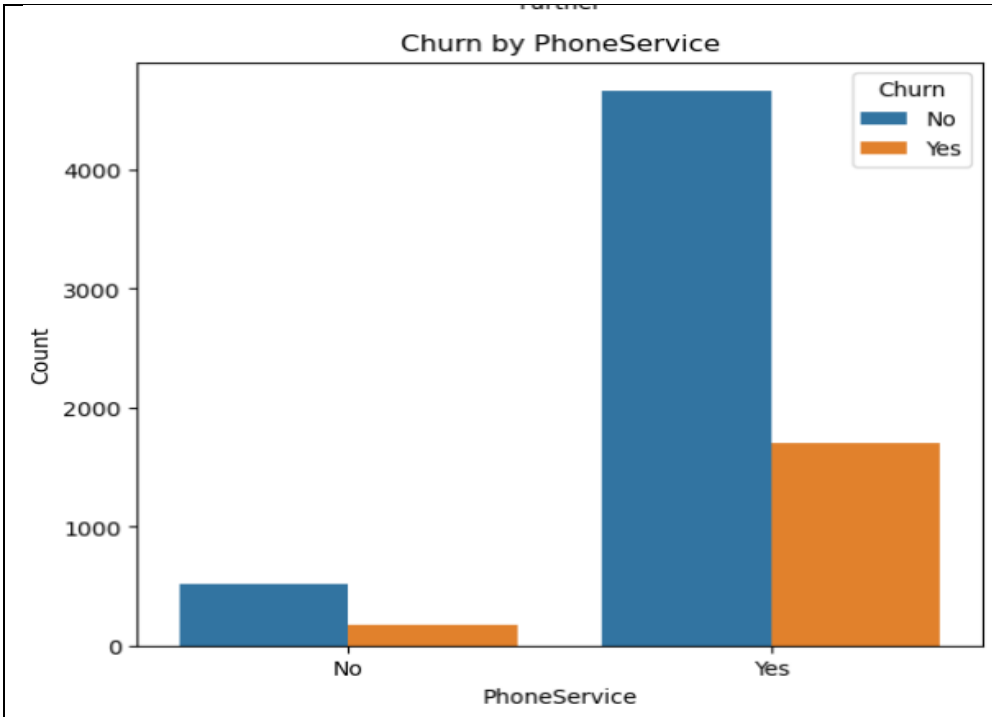
From the ratio of Churn, I can say from a business perspective that many more customers are satisfied with the services the company provides than unsatisfied. However, It would always be better to keep the churn rate as less as possible. On the other hand, for the models, the dataset is not suitable due to both intense imbalances and a pretty much lightweight dataset.



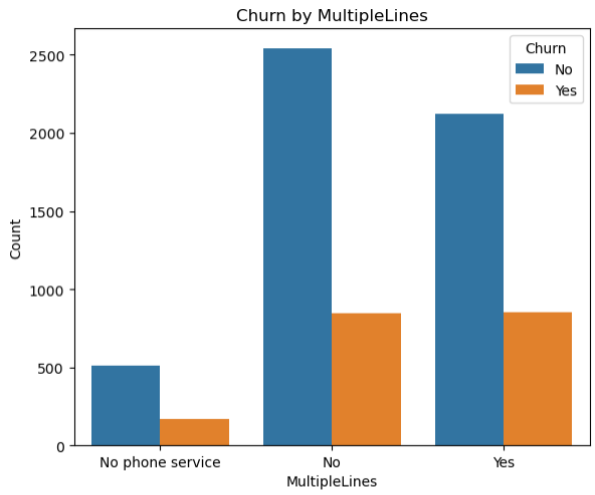
The more the customers get older, the lesser the Churn rate gets reduced.

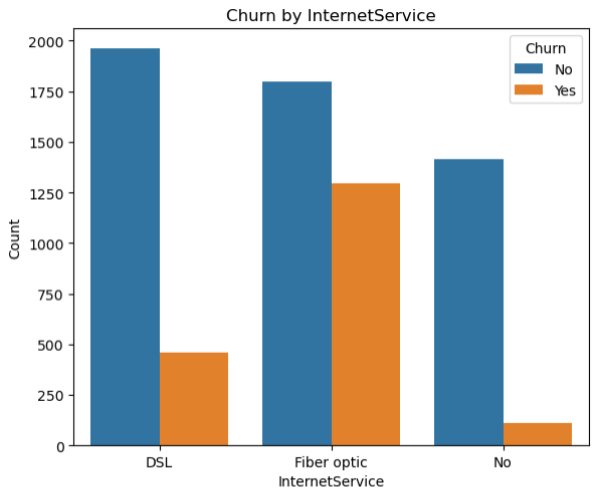


So, gender is not a significant discriminator in the Churn rate

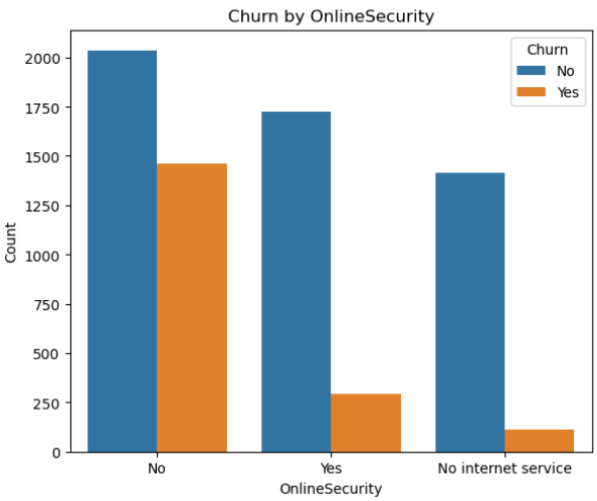


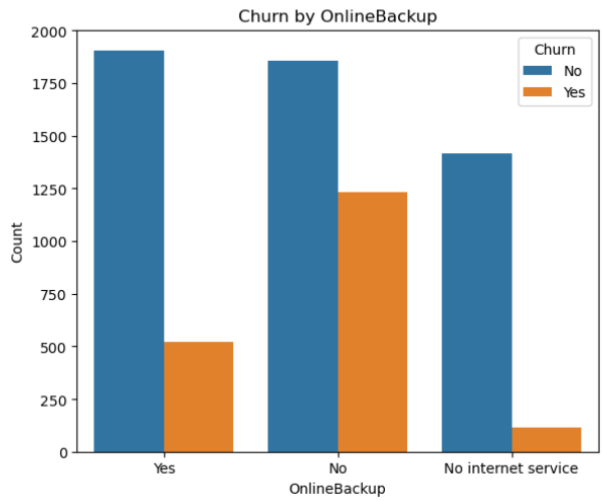
Though it looks like their Phone Service is a highly demanded service and also customers are pretty much satisfied, it needs to be improved.



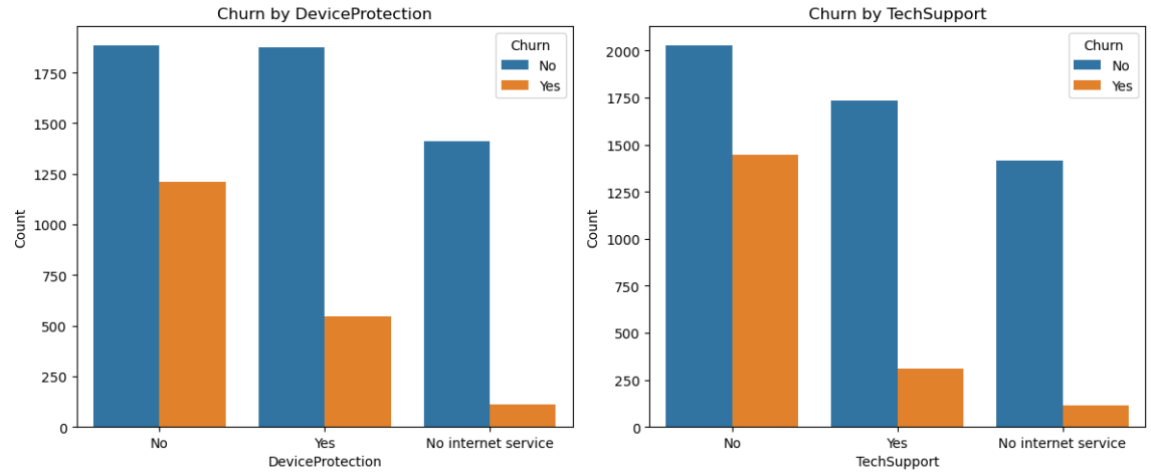


The Multiple Lines service has to be improved	The fiber optic service has to be improved.

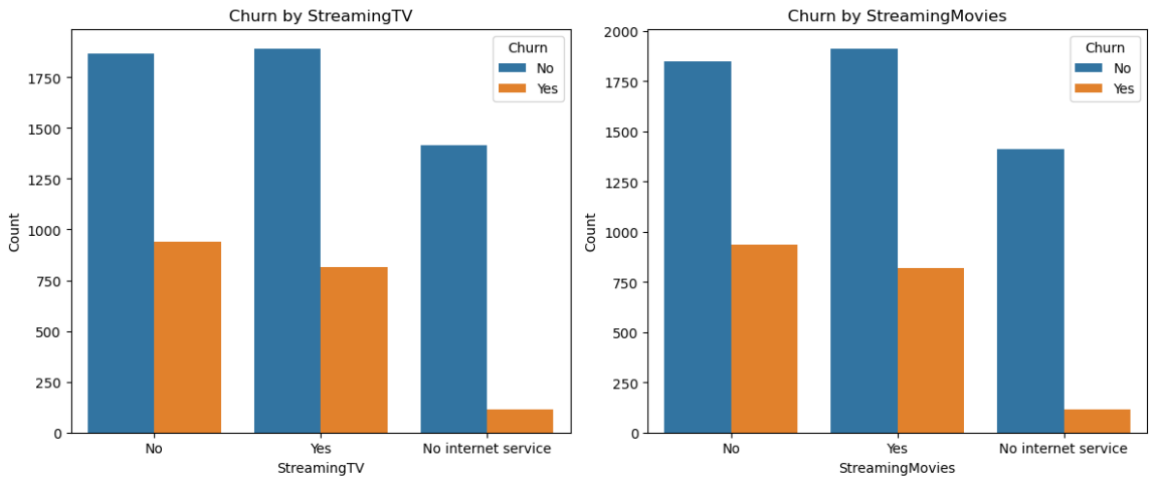




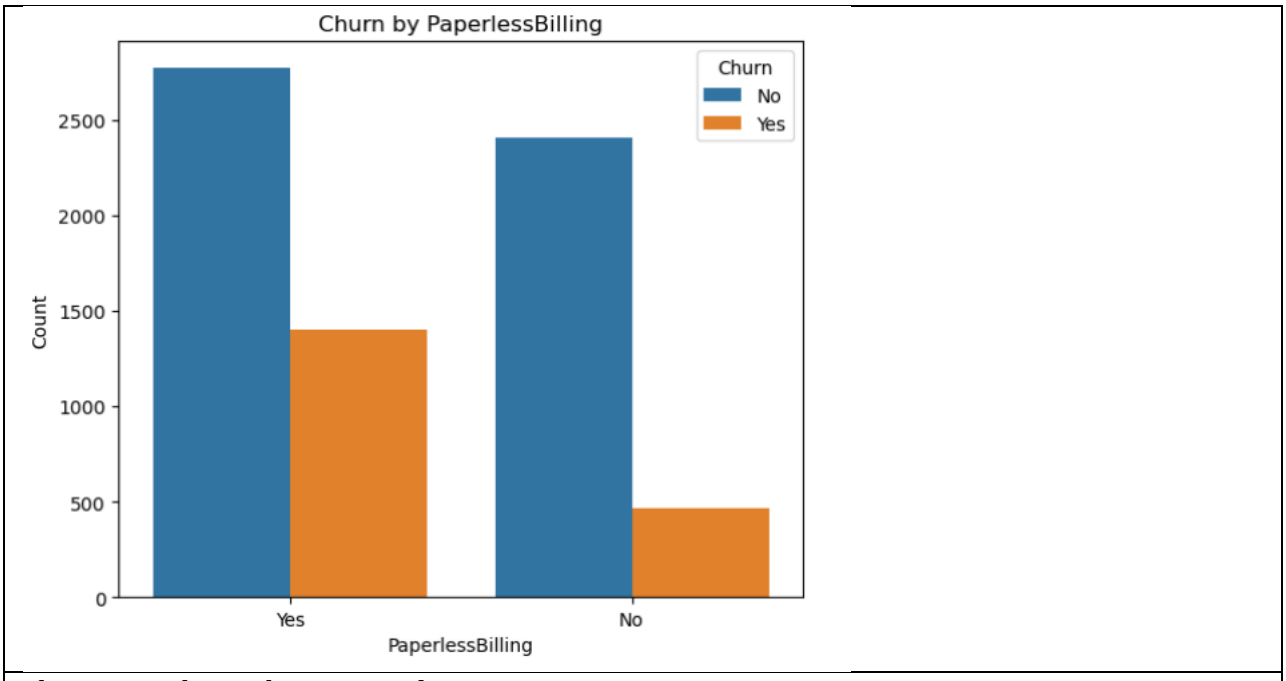
Both the services are fine but not that much demanded



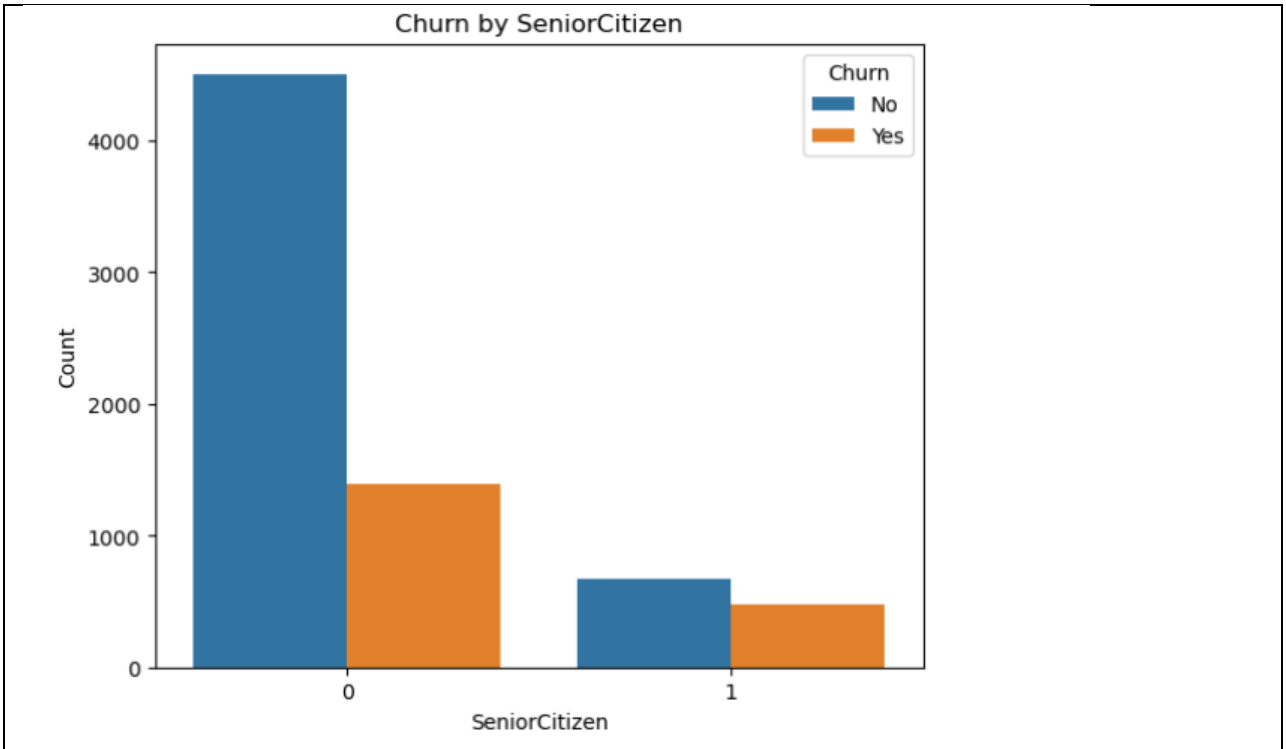
Both the services are fine but the company could do anything else to attract more customer like reducing the cost.



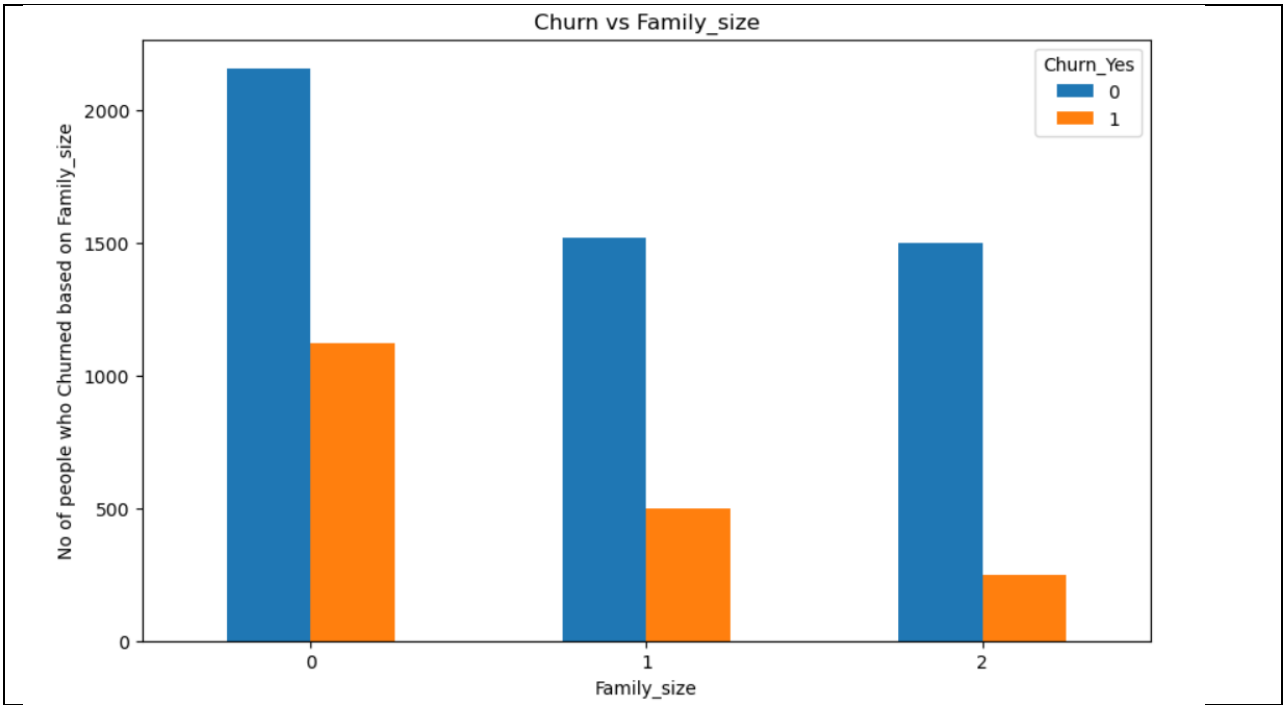
Both the services are not good .Therefore those are have to be improved.



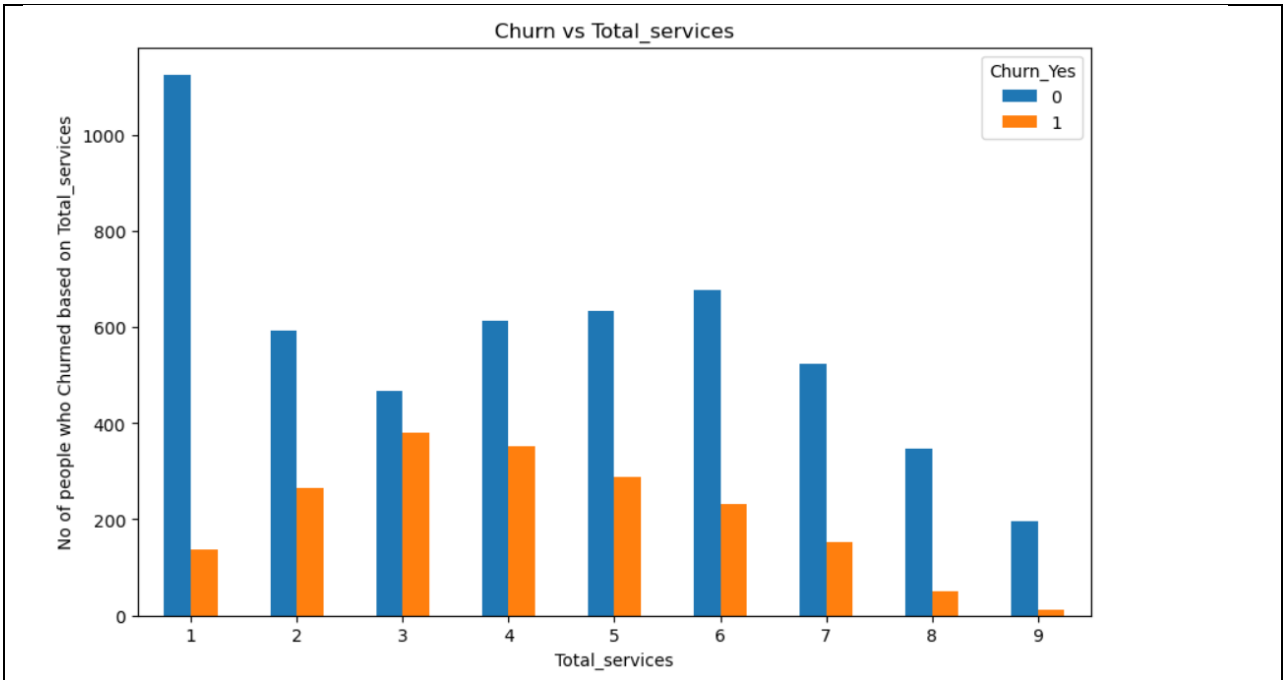
This service has to be improved.



The company should target young citizens.



Family Size: To make it simpler , combined information from 'partner' and 'dependents' columns to create a new column representing the total family size for each customer.



Total Services Used: I created a new column named "Total_services" that sums up

the number of services (phone, internet, streaming TV, etc.) each customer has signed up for. This can give an indication of the level of engagement or dependency a customer has on those services.

Contract Duration Type: I calculated the duration (in months) of the "Contract" column based on the tenure and contract type.

Example: For month-to-month -> tenure (number) in months is the duration

Or, For one year-> 12 months and so on.

Tenure in Years: Then, converted the 'tenure' column (how long they've been a customer) from months to years. This can make the data more intuitive and easier to interpret.

Monthly Charges to Total Charges Ratio: After that, calculated a new column by dividing 'monthly charges' by 'total charges'. This can indicate whether the customer tends to have stable or fluctuating charges over time.

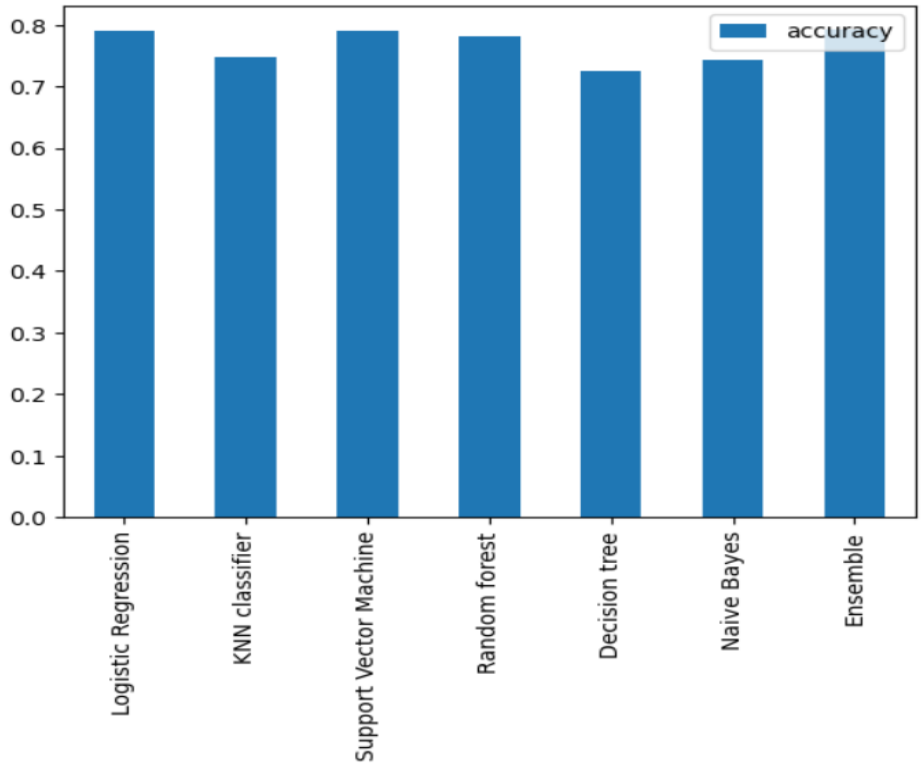
```
Correlation between InternetService_No and StreamingMovies_No internet service: 0.9999999999999999
Correlation between InternetService_No and OnlineSecurity_No internet service: 0.9999999999999999
Correlation between InternetService_No and OnlineBackup_No internet service: 0.9999999999999999
Correlation between InternetService_No and DeviceProtection_No internet service: 0.9999999999999999
Correlation between InternetService_No and TechSupport_No internet service: 0.9999999999999999
Correlation between InternetService_No and StreamingTV_No internet service: 0.9999999999999999
```

I dropped the above columns which were highly correlated.

Methodologies:

I normalized data ranged from 0 to 1 so that my models could treat each column equally and divided into two parts: Training (80%) and Testing (20%).

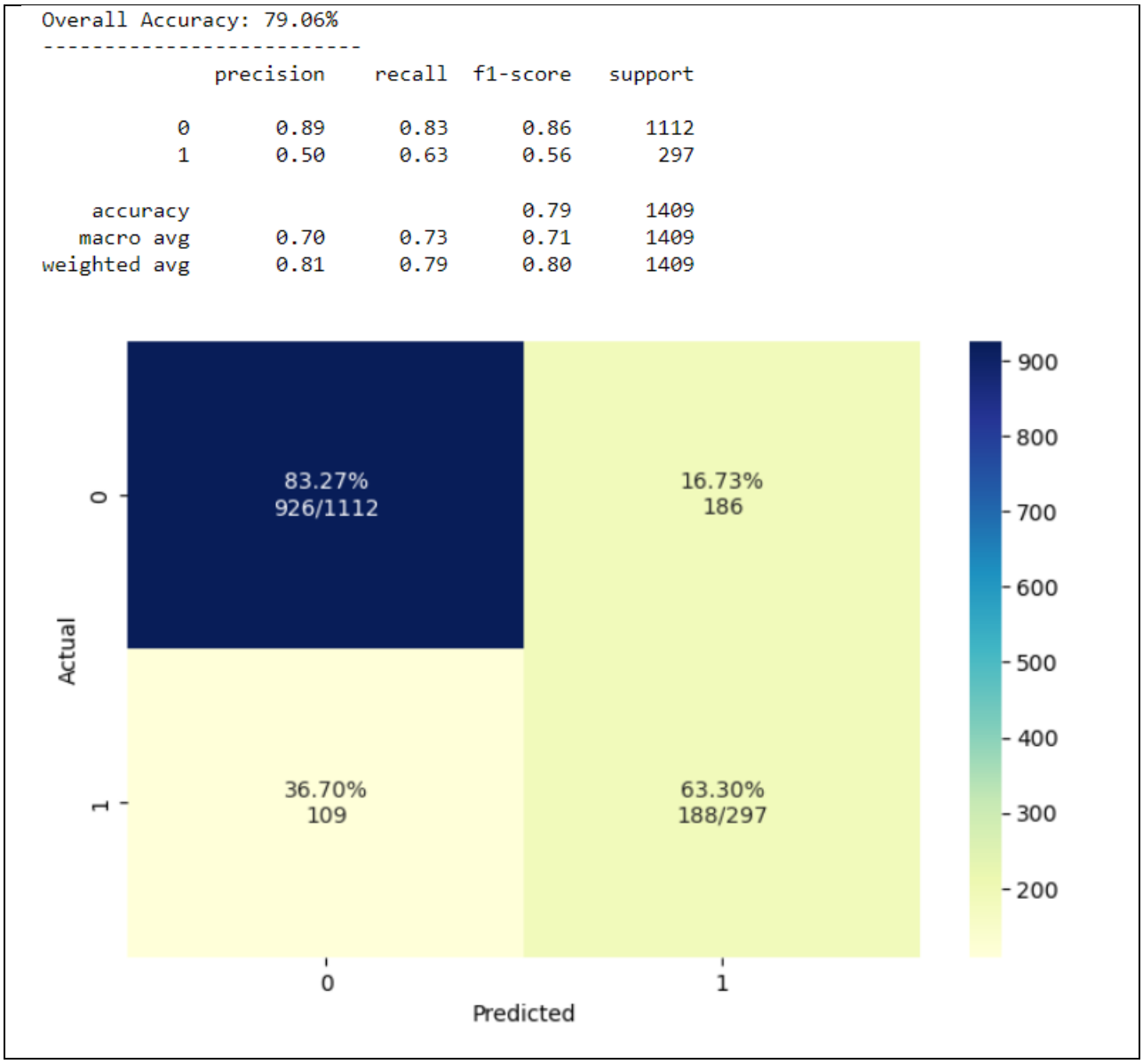
I applied six classifiers are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gaussian NB, K-Neighbors Classifier, and Support Vector Classifier. Churn : Yes->1 and No->0



	Logistic Regression	KNN classifier	Support Vector Machine	Random forest	Decision tree	Naive Bayes	Ensemble
accuracy	0.790632	0.748048	0.789212	0.780696	0.725337	0.74379	0.792051

And, among them I picked up three algorithms which gave three highest accuracies for the further analysis.

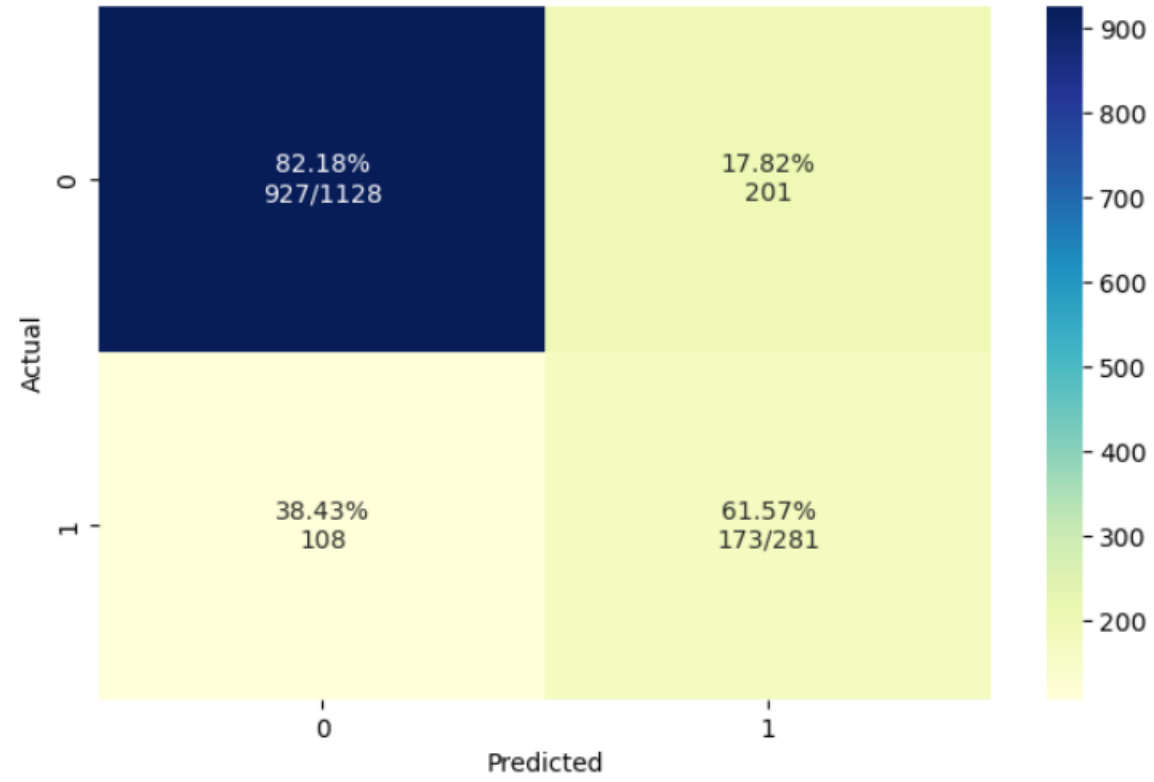
Logistic Regression:



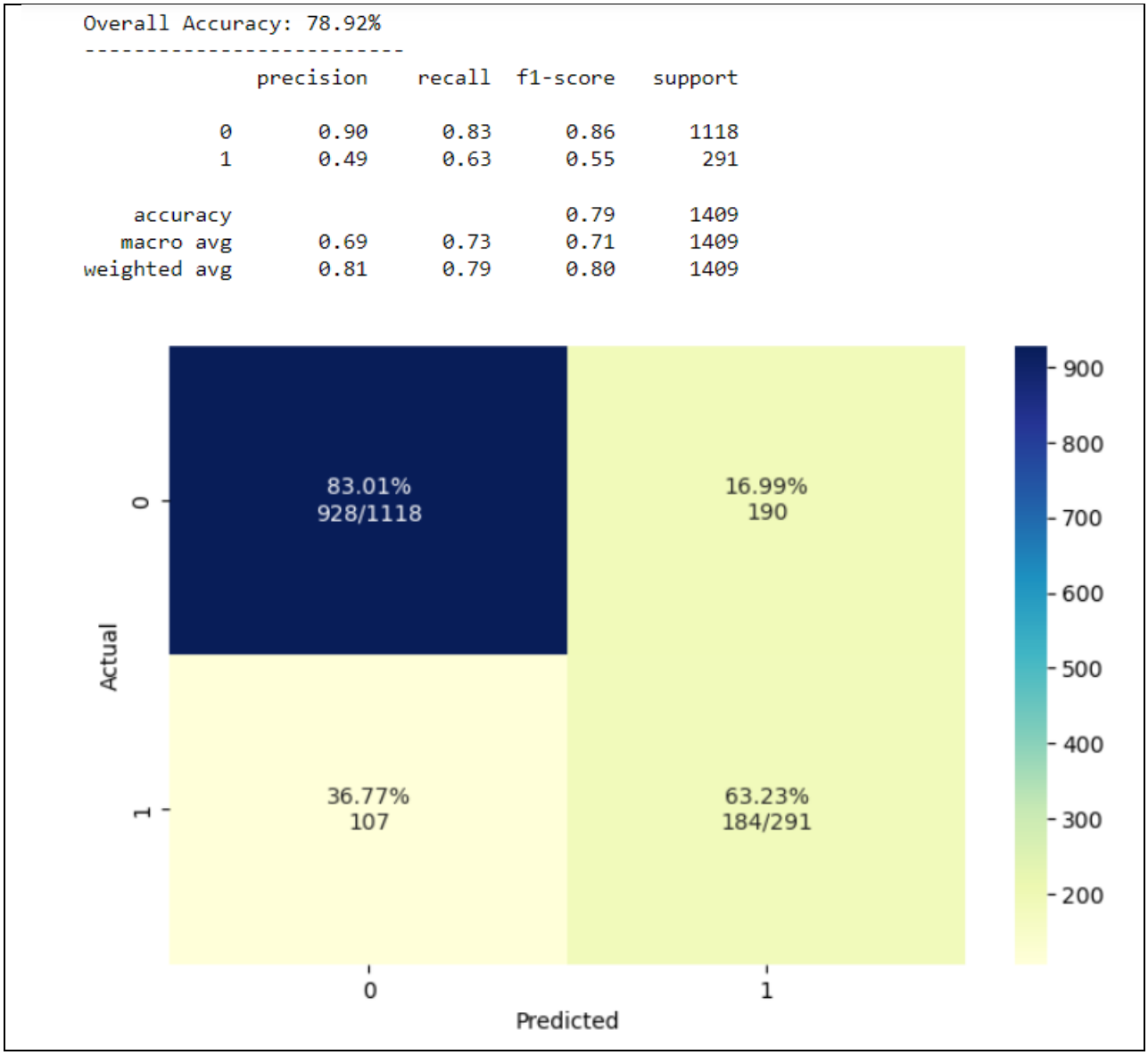
Random Forest Classifier:

Overall Accuracy: 78.07%

	precision	recall	f1-score	support
0	0.90	0.82	0.86	1128
1	0.46	0.62	0.53	281
accuracy			0.78	1409
macro avg	0.68	0.72	0.69	1409
weighted avg	0.81	0.78	0.79	1409



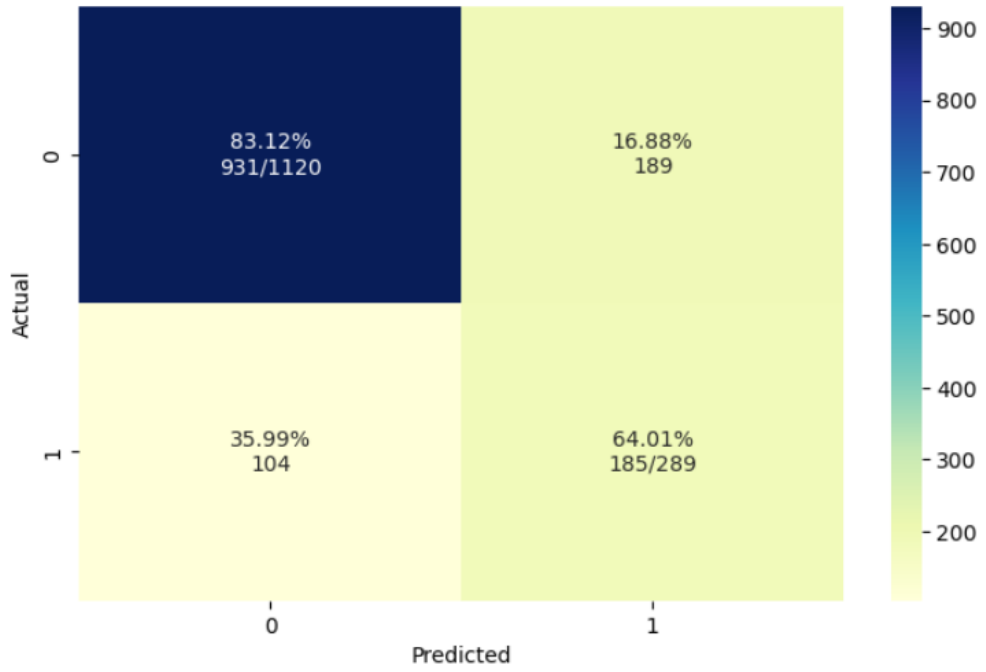
Support Vector Classifier:



Due to very few data of Churn(yes/1), the prediction of positive(1) like precision, recall and f1-oreare quite low for all three algorithms.

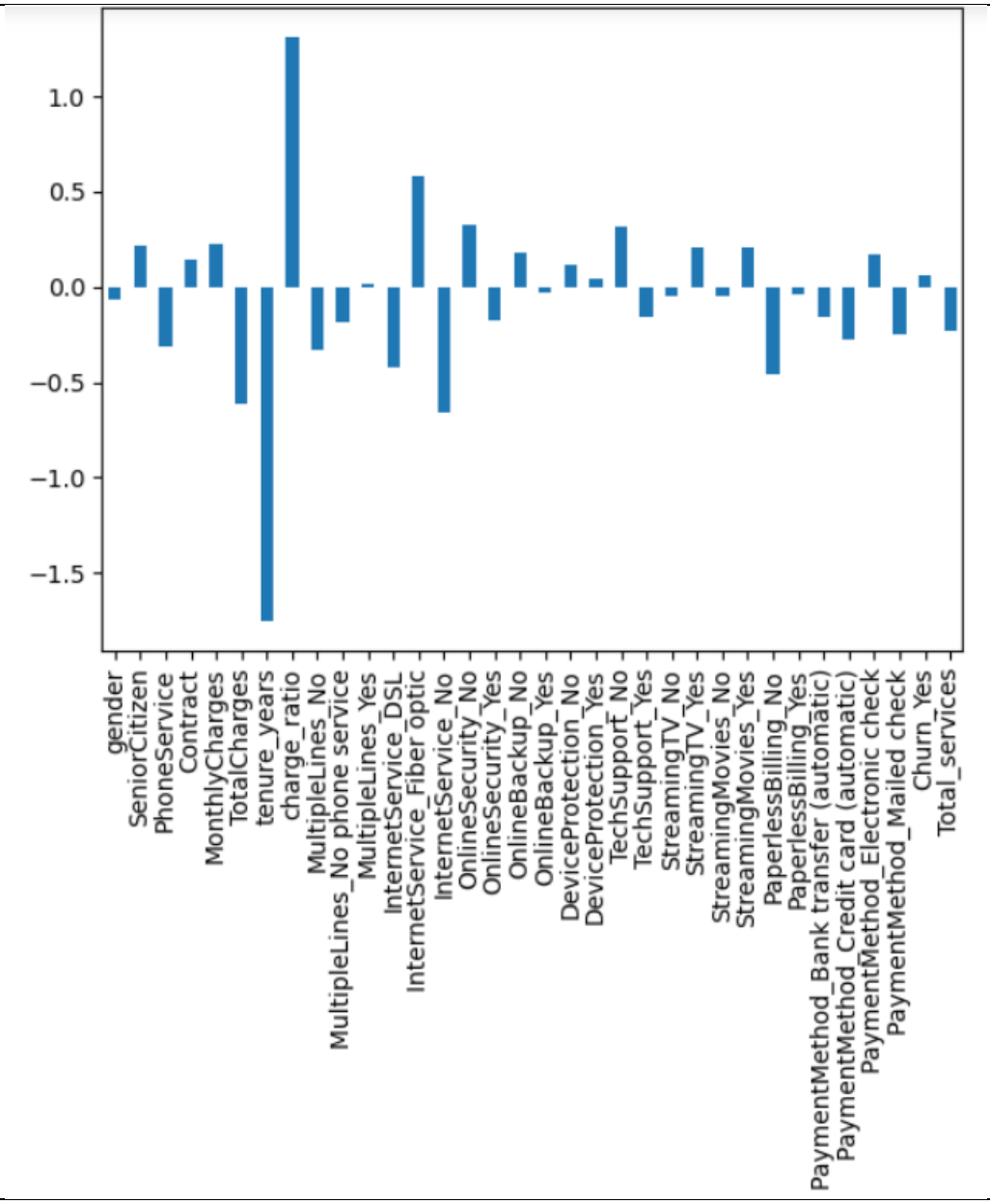
Ensemble: Logistic Regression, Random Forest Classifier, and Support Vector Classifier

```
Overall Accuracy: 79.21%
--------------------------
              precision    recall  f1-score   support

           0       0.90      0.83      0.86      1120
           1       0.49      0.64      0.56       289

    accuracy                           0.79      1409
   macro avg       0.70      0.74      0.71      1409
weighted avg       0.82      0.79      0.80      1409
```



I applied an ensemble technique by combining Logistic Regression, Random Forest Classifier, and Support Vector Classifier for the prediction. However, nothing significantly improved.

Though for the further improvement, I tried Principal Component Analysis with 7 components and Grid Search CV for the parameter tuning, I didn't notice any significant changes in accuracy.



'tenure_years': -1.7540263490880714,
'charge_ratio': 1.309619359910505

Among my four new aggregated columns, two columns were highly impactful. These importance are from Logistic Regression.

Summary of the selected models:

Model	Accuracy	Label	Precision	Recall	F1-Score	False Positive	False Negative
Logistic Regression (C=10)	79.6	0	0.89	0.83	0.86	36.70	16.73
		1	0.50	0.63	0.56		
Random Forest(5)	78.07	0	0.90	0.82	0.86	38.43	17.82
		1	0.46	0.62	0.53		
Support Vector Classifier (linear)	78.92	0	0.90	0.83	0.86	36.77	16.99
		1	0.49	0.63	0.55		
Ensemble	79.21	0	0.90	0.83	0.86	35.99	16.88
		1	0.49	0.64	0.56		

Conclusion: I tried my best to get the best accuracy but to the best knowledge, due to an intense imbalanced dataset, my models were unable to understand the relation among the features or unable to pick a definitive pattern. That's why I didn't get good accuracy. I feel if the dataset was balanced I could get much better accuracy.