SYRIATEL CUSTOMER CHURN PREDICTION

AUTHOR: NASHON OKUMU

Project Overview

SyriaTel is confronting a critical issue with high customer churn rates, where numerous customers are discontinuing their services in favor of competitors.

To address this, SyriaTel plans to develop a customer churn prediction model.

This model aims to identify customers at risk of leaving, uncover the factors contributing to churn, and provide actionable insights to enhance customer retention and maximize profitability.

Business Understanding

SyriaTel is a telecommunications company that is currently facing a very high churn rate as many customers tend to discontinue and leave their services and replacing them with their competitors.

The company wants to develop a customer churn prediction model that would address this issue.

SyriaTel aims to gain insights on contributing factors to the high churn rate in order to reduce it, increase customer retention and maximize the company's profits.

Data Understanding

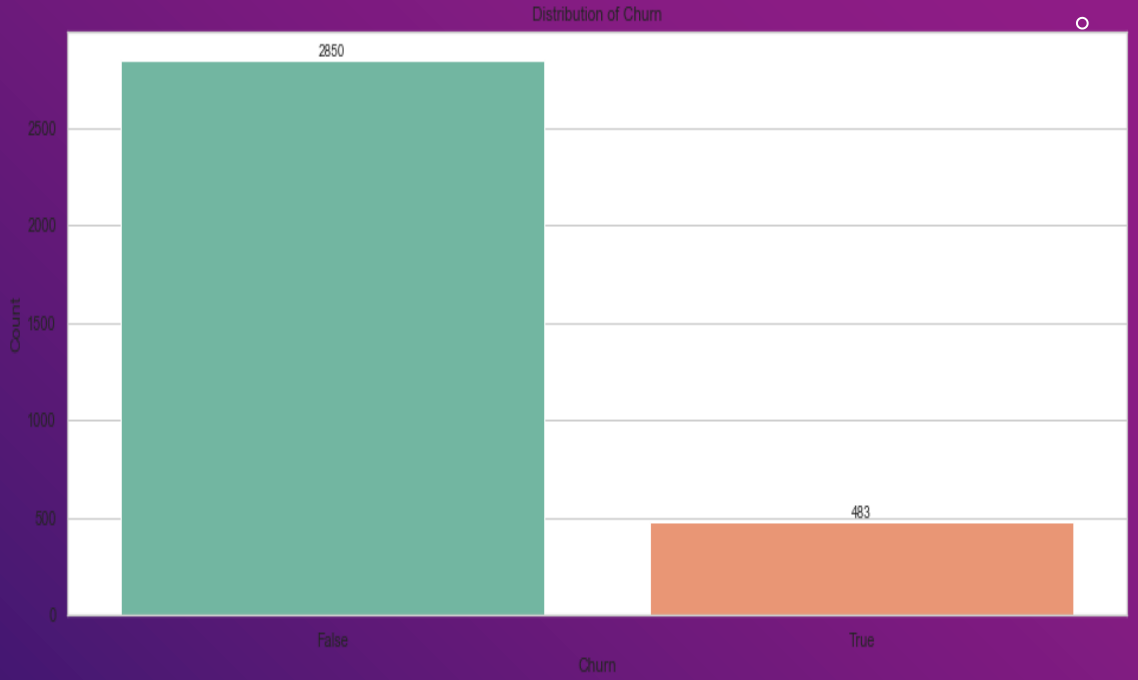
The dataset for this project is sourced from Kaggle (https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset).

Our target features is churn, however, the dataset has a total of 20 columns and 3333 rows.

By analyzing these features, we can build predictive models to identify customers who are likely to churn, enabling the company to take proactive measures to retain them.

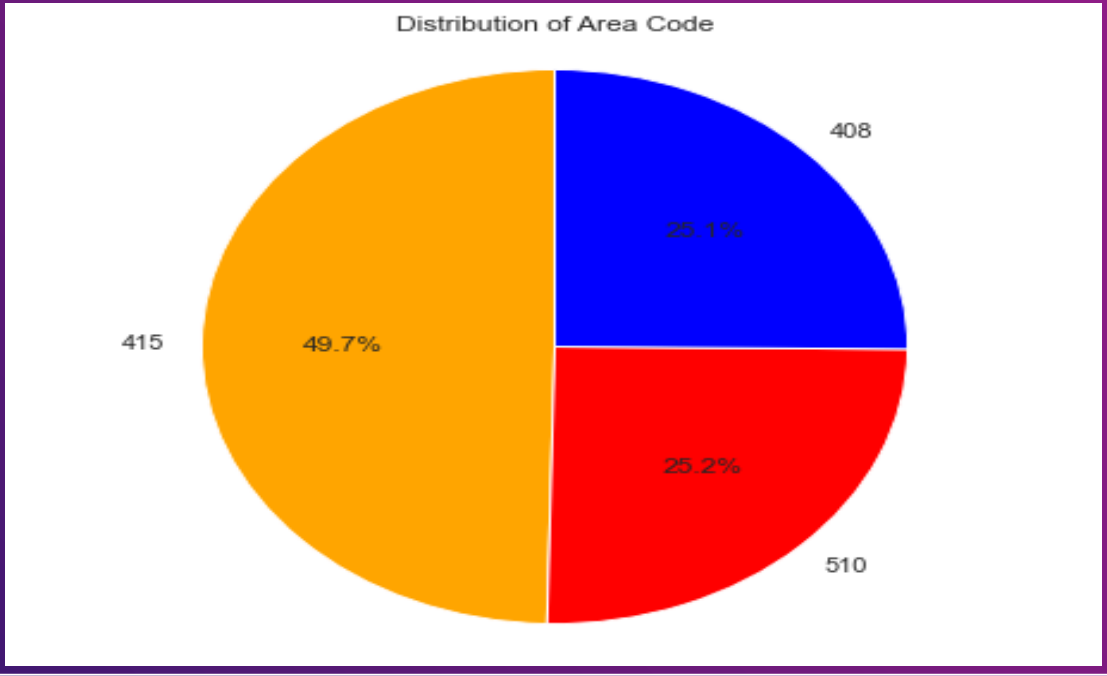
Explanatory Data Analysis

Distribution of Churn
Out of the 3333 customers, 483 have churned, which means they have terminated their contracts with the company.



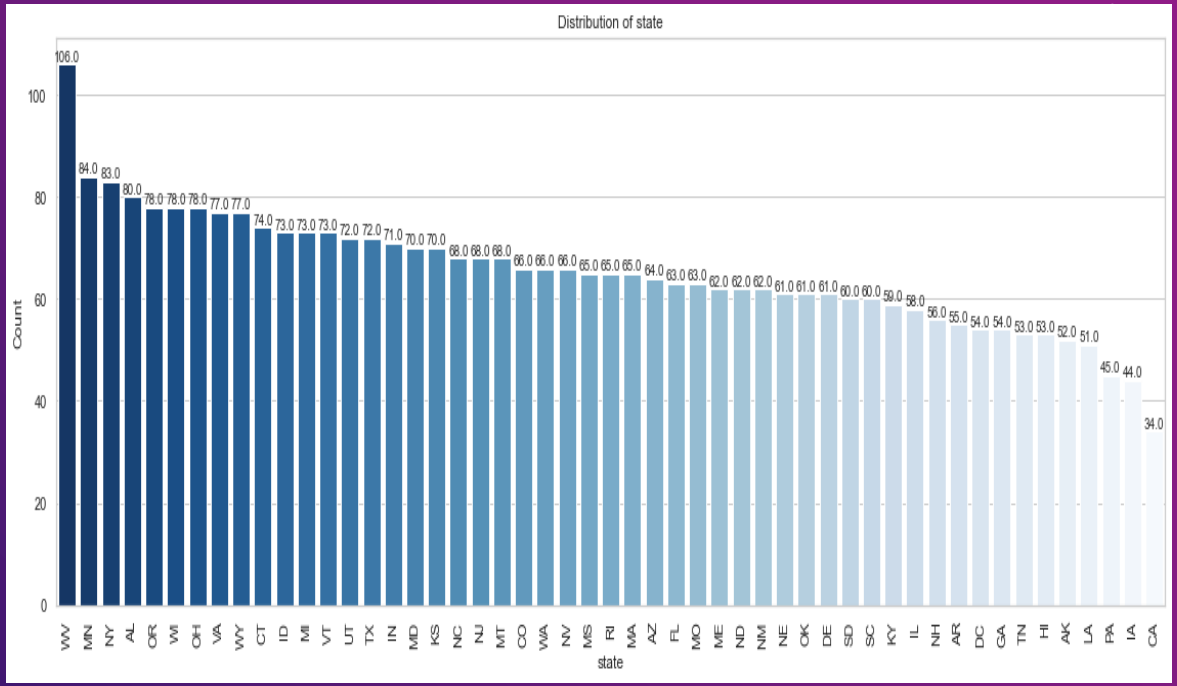
Distribution of Area Code

Almost half (49.2%) of customers come from 415. About a fourth of customers come from 510 and another fourth from 408.



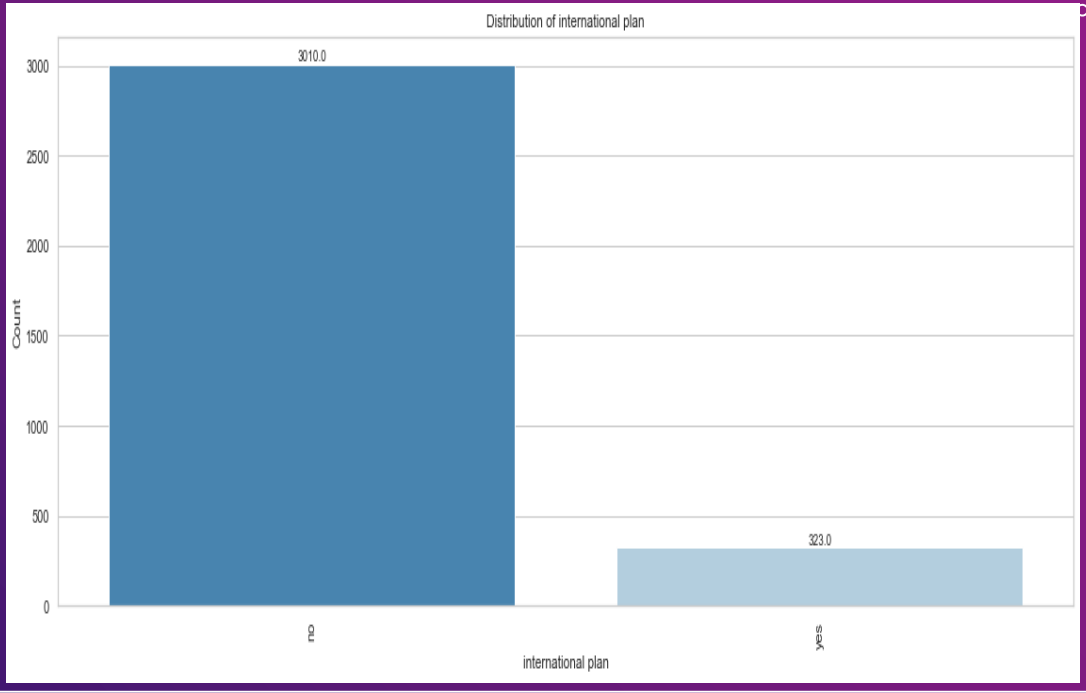
Distribution of Categorical Features

This distribution shows that most customers come from West Virginia WV, Minnesota MN, New York NY, and Ohio OH.



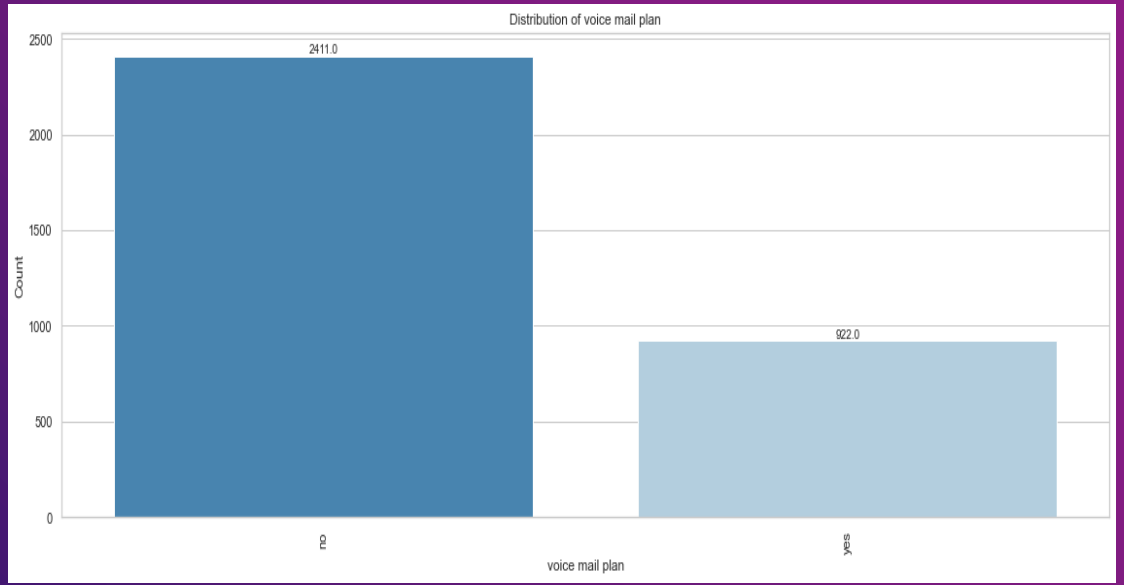
Distribution of International Plan

Out of the 3333 customers, only 323 have international plan.



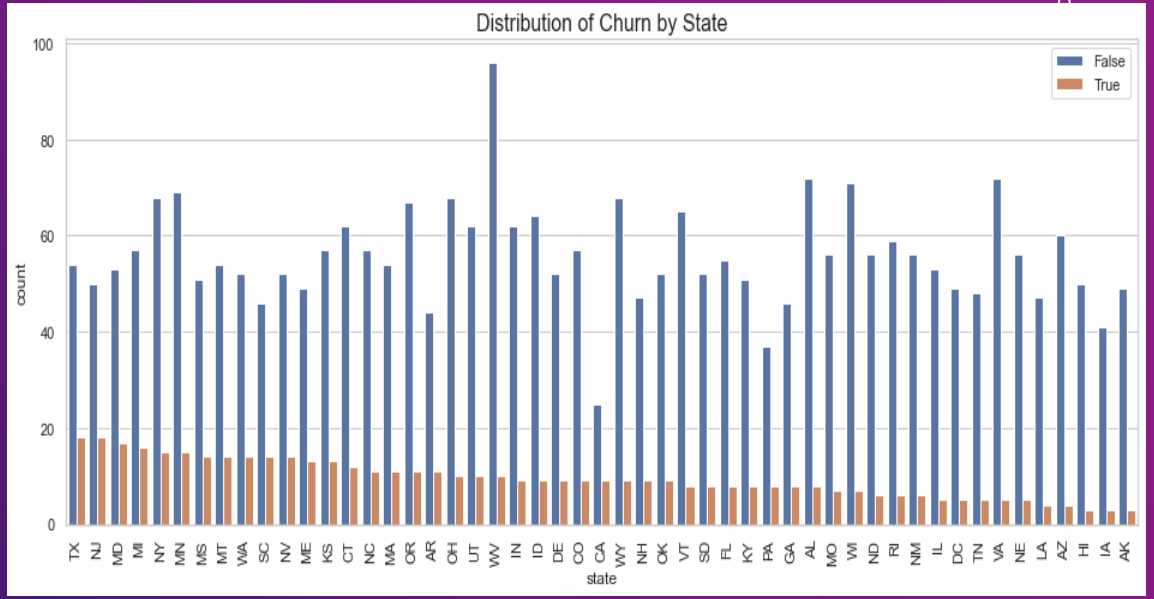
Distribution of Voice Mail Plan

Out of the 3333 customers, only 922 have voice mail plan.



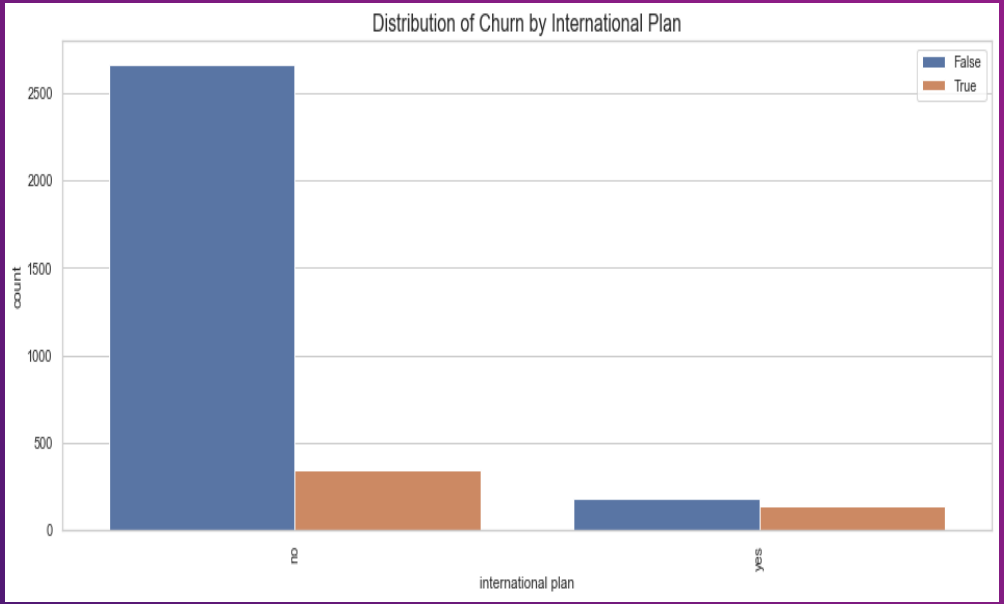
Distribution of Churn by Categorical Features

Majority of customers who churn come from Texas TX, New Jersey NJ , Maryland MD, Miami MI and New York NY.



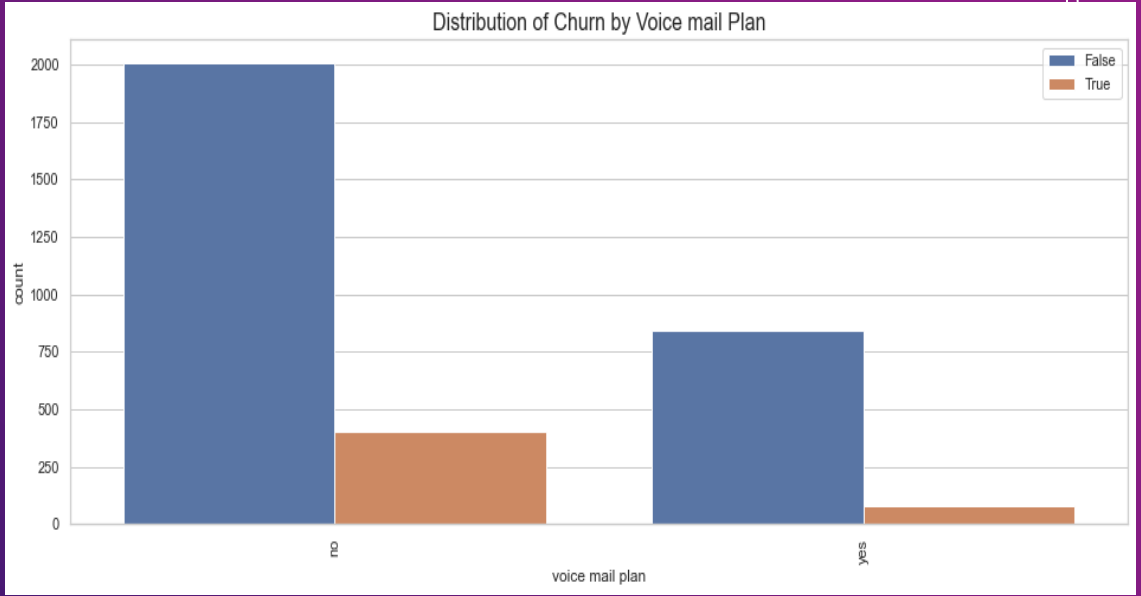
Distribution of Churn International Plan

Majority of customers who churned did not subscribe to an international plan while those who retained their accounts had very few with international plans but quite higher than those who churned..



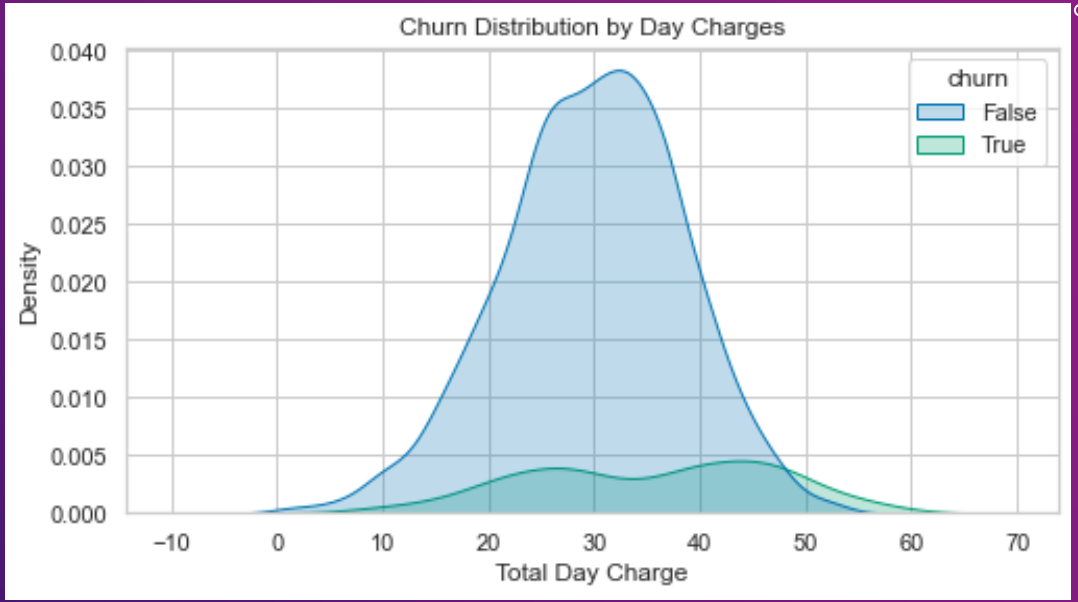
Distribution of Churn by Voice Mail Plan

Majority of customers who churned did not subscribe to a voice mail plan while those who retained their accounts had very few with voice mail plans but quite higher than those who churned.



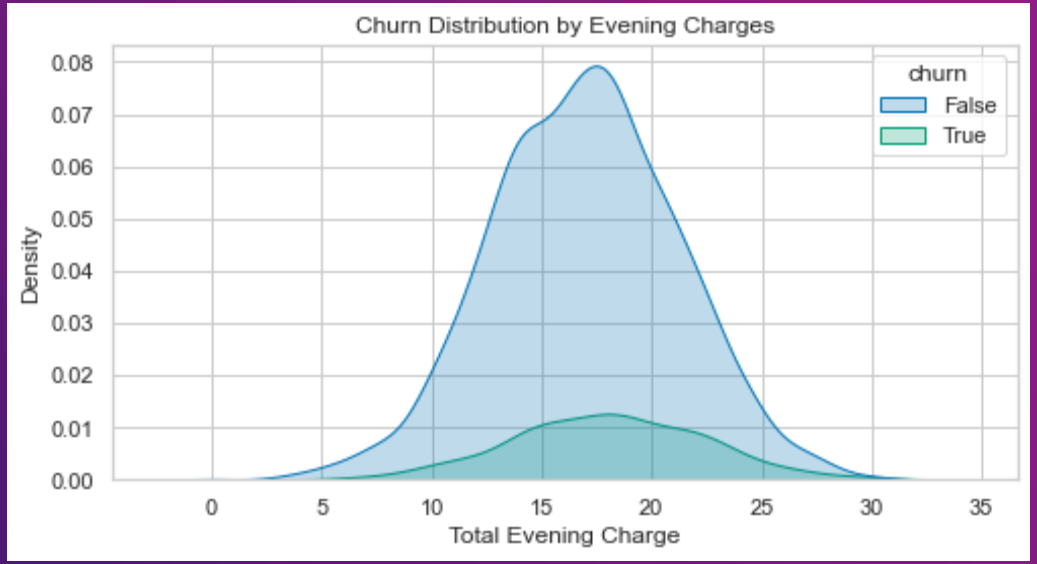
Churn Distribution by Day Charges

Customers who terminated their accounts incurred high charges during the day implying that unhappy customers find the day charges to be too expensive.



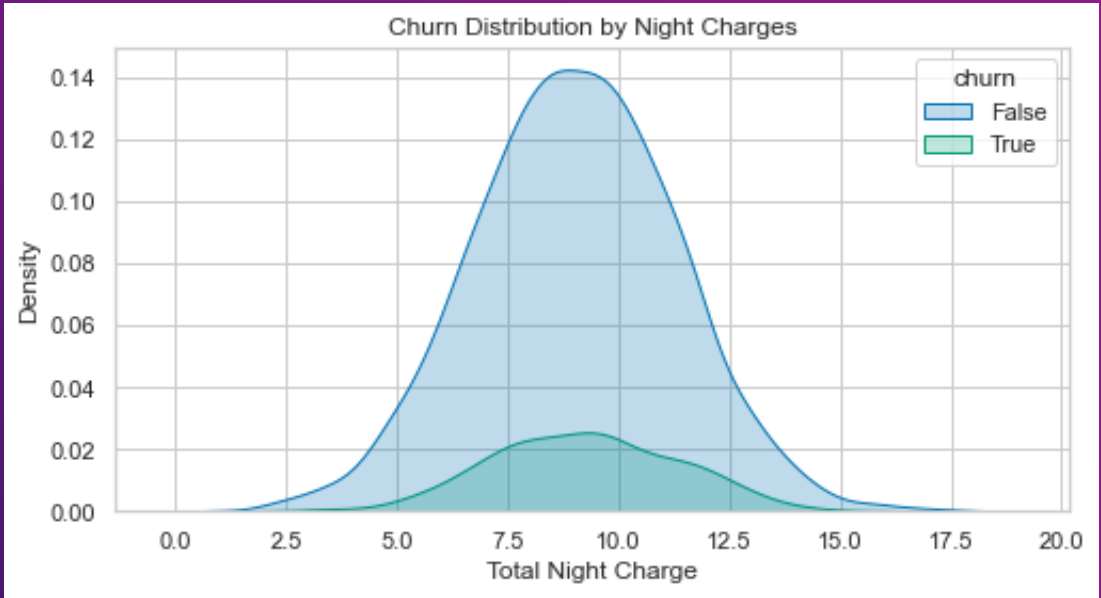
Distribution Churn by evening charges

This patterns is also similar to day charges hence those who terminated their accounts tend to incur higher charges in the evening.



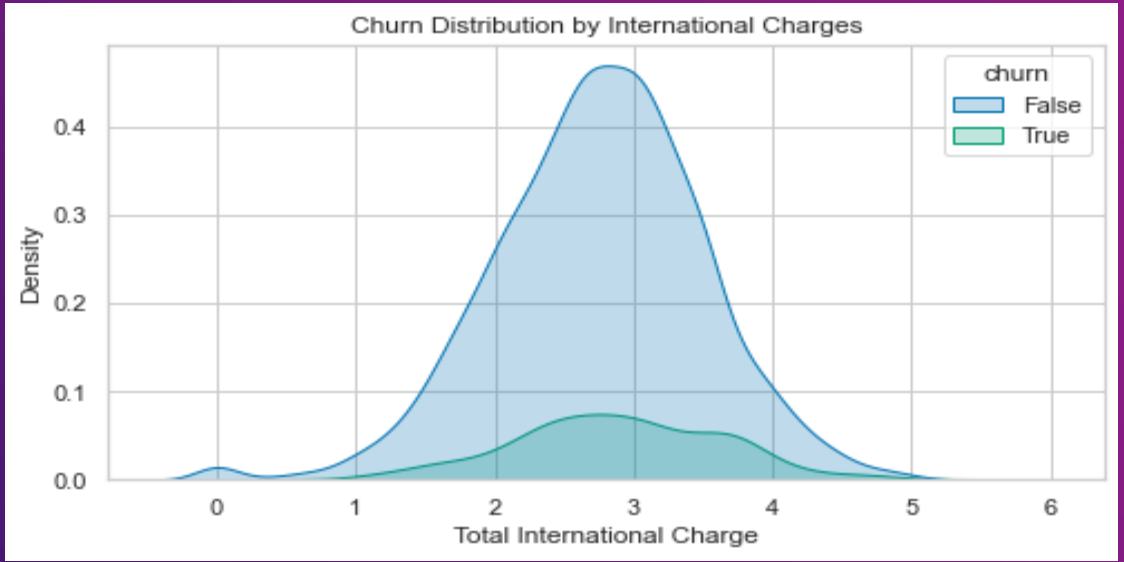
Distribution Churn by night charges

Customers who have terminated their account incur high charges at night compare to those who have retained their account.



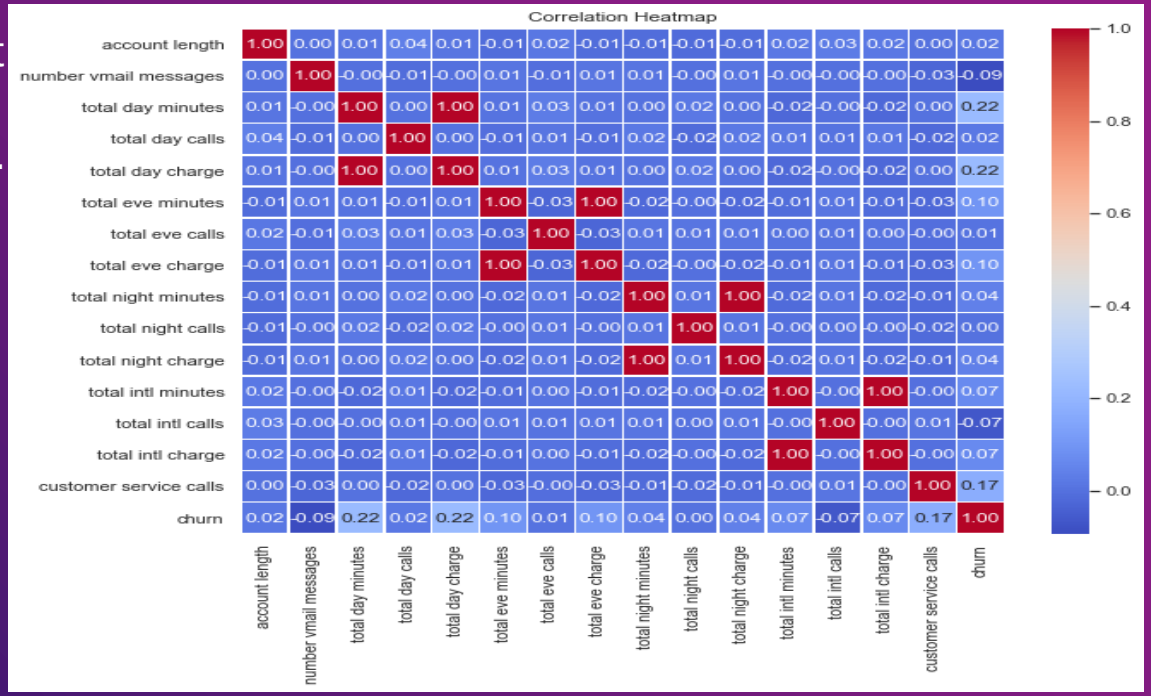
Distribution Churn by International Charges

Customers with international charges are less likely to churn compared to those with no international charges.



Feature Correlation

Clearly, most features are not correlate as the correlation value is too low or a negative. However, Total day charge and total day minutes, Total eve charge and total eve minutes , Total night charge and total night minutes, Total int charge and total int minutes share a strong positive correlation which indicates a perfect correlation between the selected pair of features..



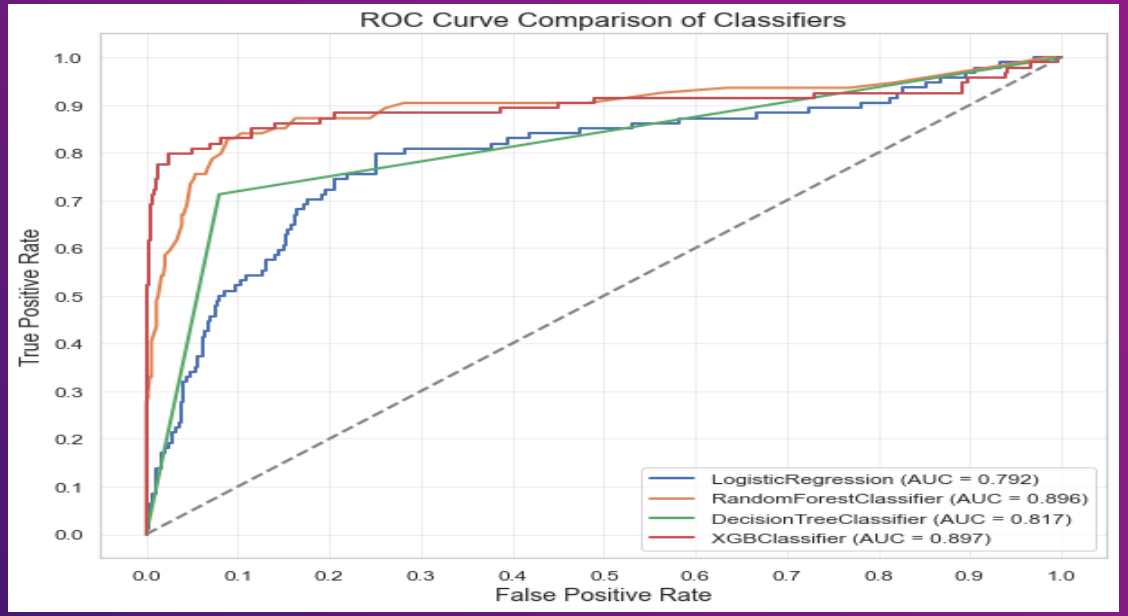
Modeling

Comparing the models, XG boost had a recall score of 80% which achieved a recall score of 80%. The models used were Logistic Regression, Random Forest, Decision Tree and XG Boost.

classifier	recall
LogisticRegression	0.755319
RandomForestClassifier	0.670213
DecisionTreeClassifier	0.712766
XGBClassifier	0.797872

Model Evaluation

Comparing the models, XG boost had a recall score of 80% which achieved a recall score of 80%. The model performance was also the best compared to other models in the ROC curve.



Conclusion

The XG Boost Classifier has the highest recall score of 0.8 that is 80%. It is also the best performing model according to the ROC curve. We achieve the objective as the model is able to predict customer churn at the expected recall score.

Recommendations

1. Provide discounts and promotions in areas 415 and 510 since the churn rate here is too high. This would make them stay with the company.
2. Review pricing structure at daytime,eveing and nights as well as international chares. Discounted prices plans will help reduce churn rate.
3. Improve the quality of customer service calls by providing training to the customer representatives as this would create high customer retention.
4. Driving compaighns in states with high churn rate such as Texas, New Jersey and Maryland by improving marketing strageies for customer retention.