SyriaTel Customer Churn Prediction

By Myles Mulusa

22nd May, 2024

Overview & Business Problem

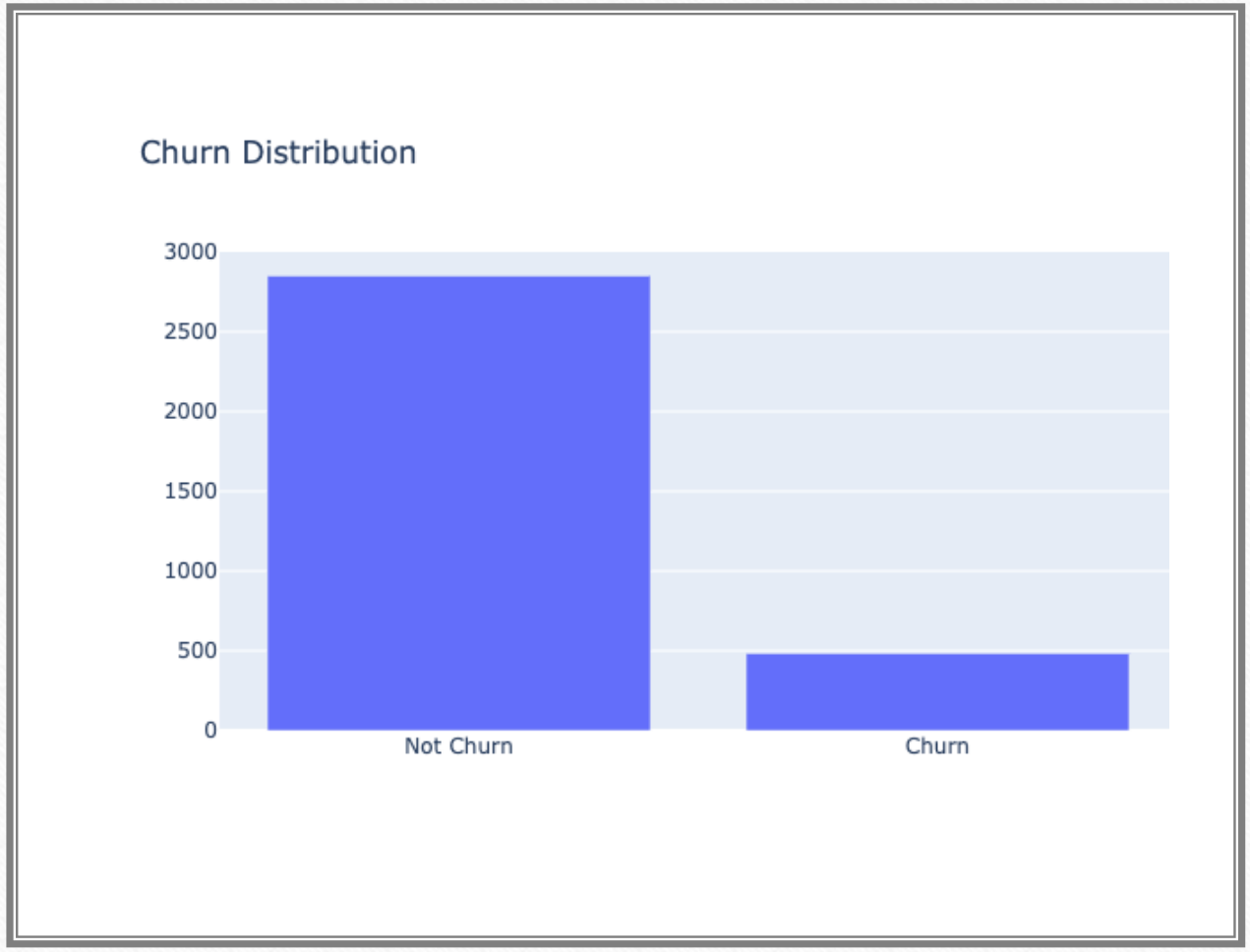
- Customer churn refers to the process where customers stop using a company's services, either due to dissatisfaction or the availability of better alternatives from competitors at lower prices. This can lead to substantial revenue and profit losses.
- Accurately predicting customer churn can enhance retention rates, boost market share, and improve overall business performance.

Specific Objectives

- Conduct exploratory data analysis on the dataset.
- Apply various classification algorithms to identify the most effective model for predicting customer churn.
- Select the optimal model based on performance.
- Use the chosen model to make churn predictions.
- Evaluate the accuracy of the predictions.
- Give recommendations on customer retention.

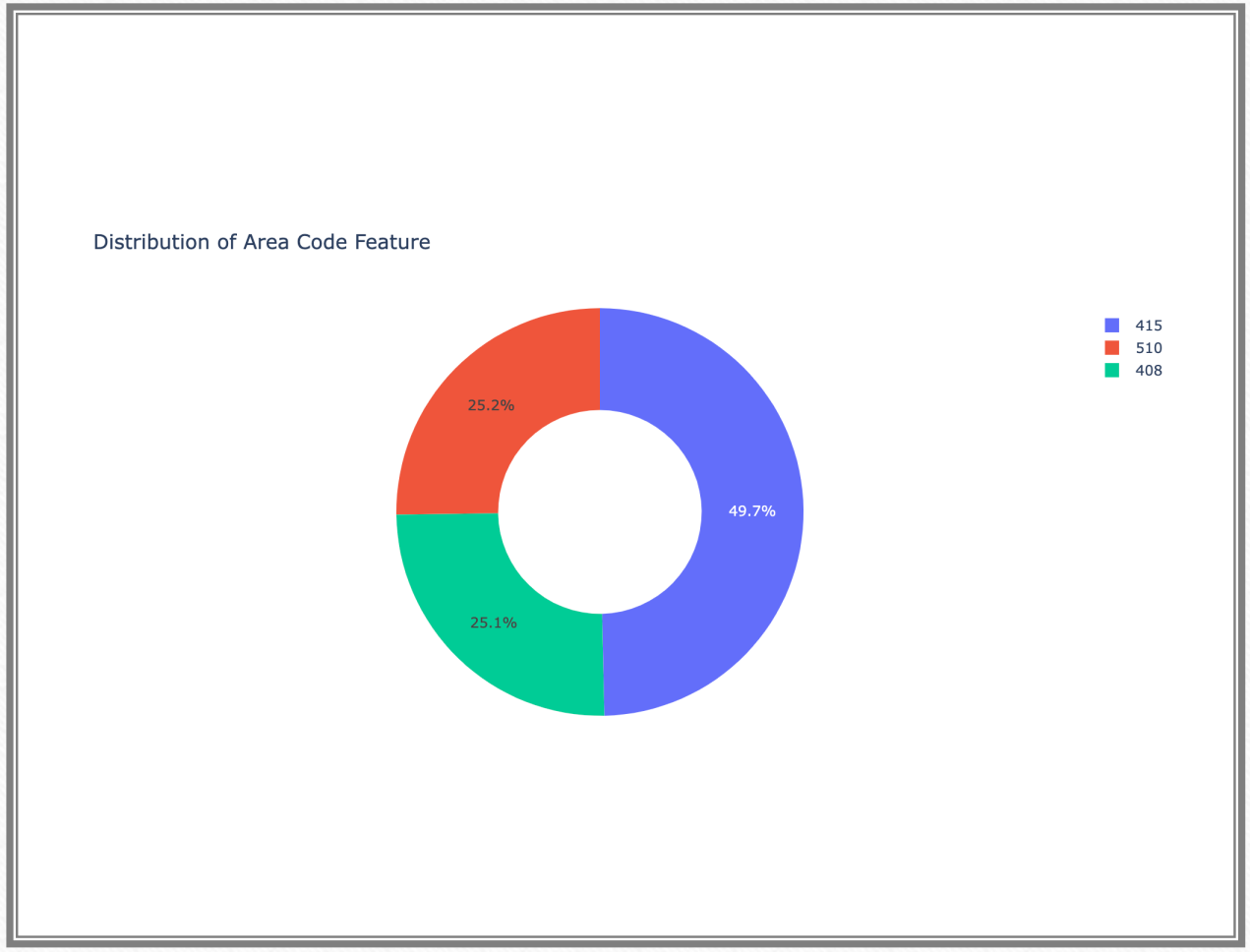
Data Understanding

- We began by examining each predictor variable, focusing on their characteristics and distributions. We then cleaned the data and conducted univariate analysis to understand the distribution of each individual variable, followed by bivariate analysis to compare these variables against our target variable, churn. This helped us understand how the target variable is distributed across the predictor variables.
- We also examined the relationships between variables. Categorical variables were converted to numerical format for modelling purposes. The data was then split into training and testing sets, allowing us to train the model on the training set and evaluate it using the testing set.
- Various classification algorithms were applied to the data, and since many of these algorithms require data scaling, we standardized the predictor variables in both the training and testing sets.



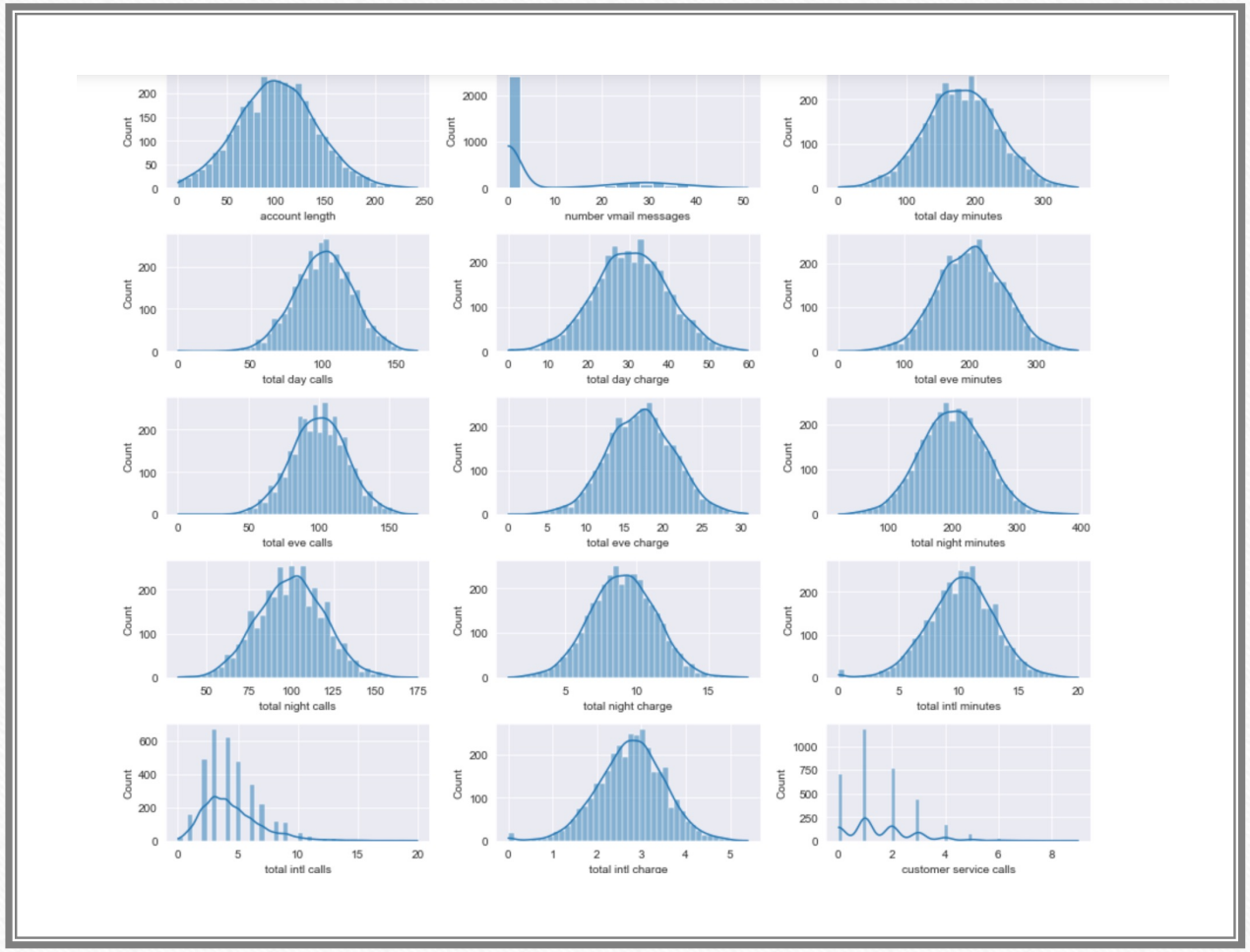
Exploratory Data Analysis(EDA)

- Out of the 3,333 customers included in the dataset, 483 have ended their contracts, representing approximately 14.5% of the total customer base. This disparity in the distribution of the binary classes indicates an imbalance in the data which will be addressed before modeling.



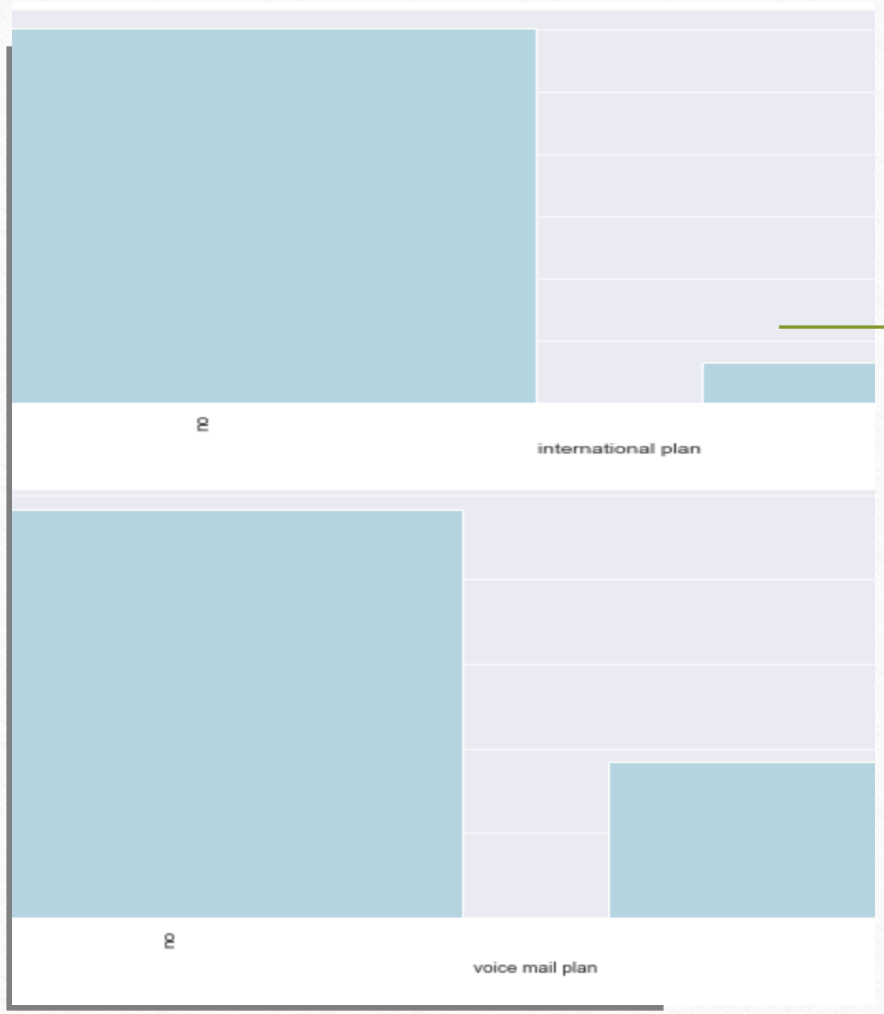
Area Code Feature

- Almost half of the customers are in area code 415, a quarter of customers are in area code 510 and another quarter are in area code 408.



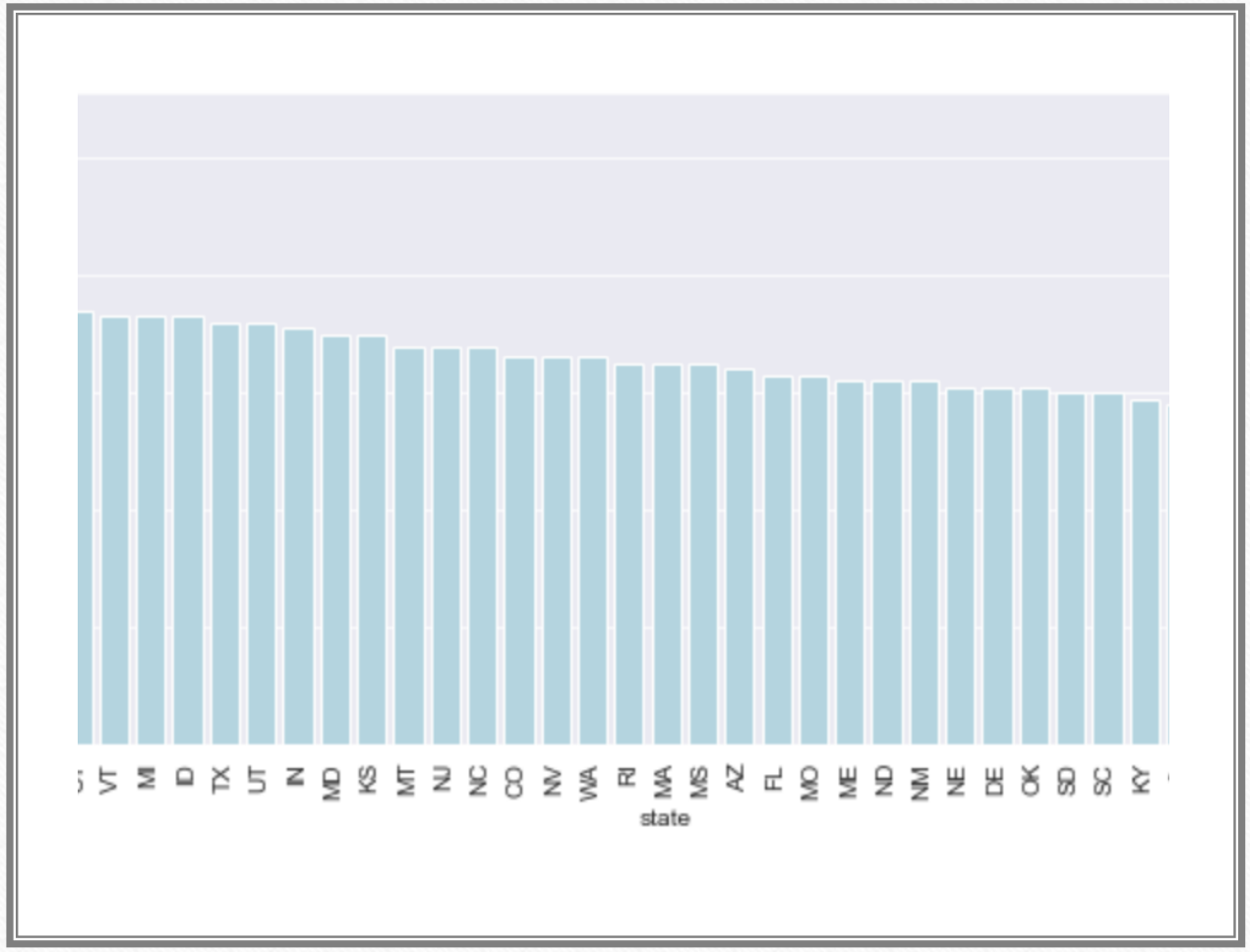
Numerical Features

- All features, with the exception of customer service calls and the number of voicemail messages, exhibit a normal distribution. Although total international calls appear slightly skewed to the right, it still follows a roughly normal distribution.
- However, the distribution of customer service calls displays several peaks, suggesting the presence of multiple modes within the population.



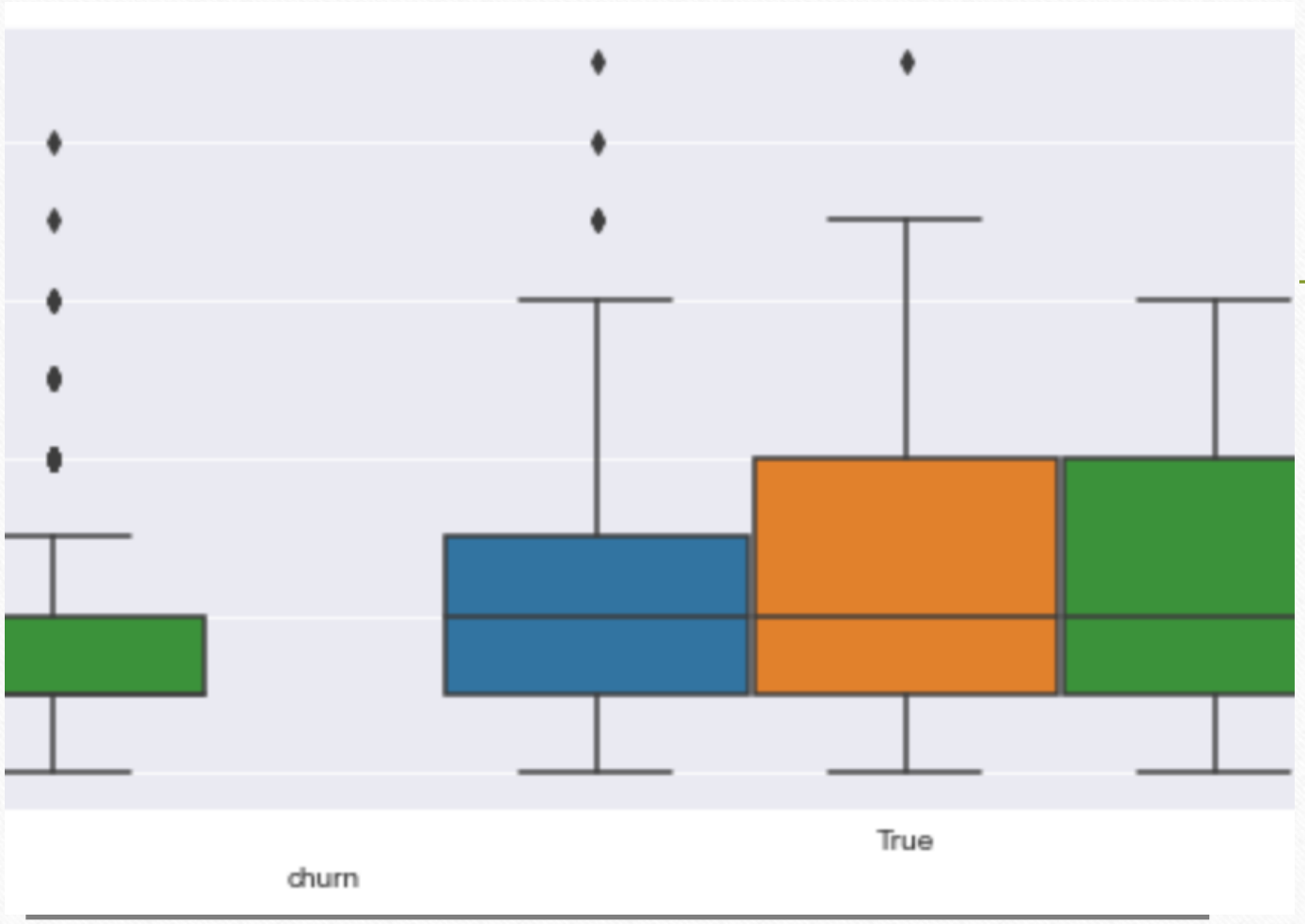
Categorical Features

- Only 323 customers out of 3333 have an international plan which is about 0.1%.
- Only 922 customers out of 3333 have a voicemail plan which is about 0.3%.



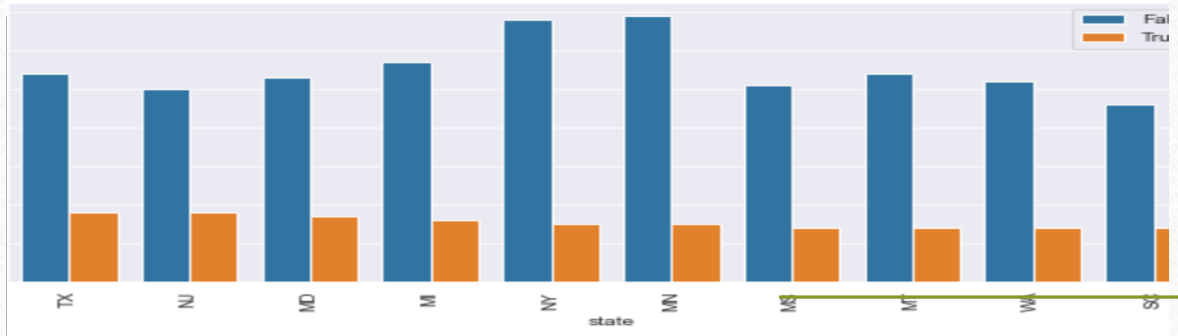
Categorical Features(cont)

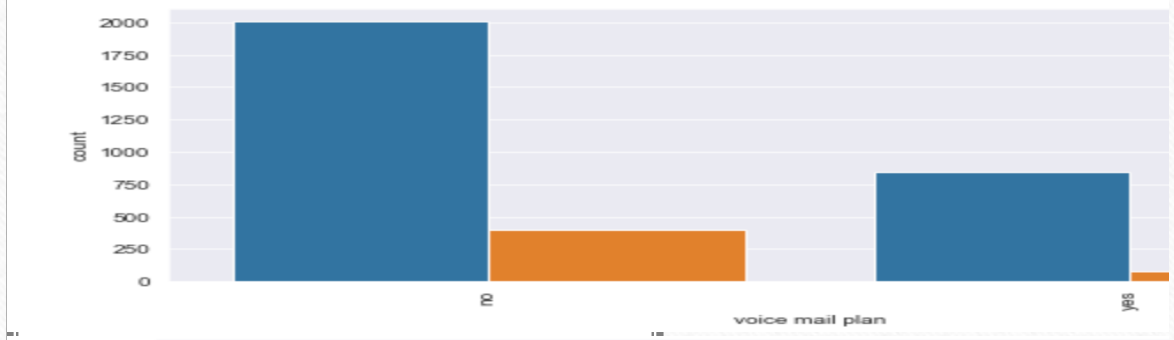
- Most of the customers are from West Virginia, Minnesota, NewYork, Alabama and Wisconsin.

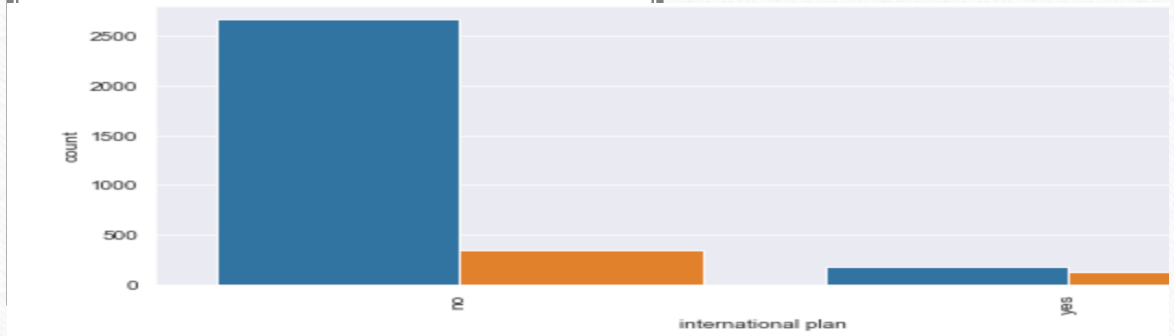


Bivariate Analysis

- The majority of customers who terminated their accounts belong to area codes 415 and 510. Additionally, it is evident that there are several outliers present in the dataset.





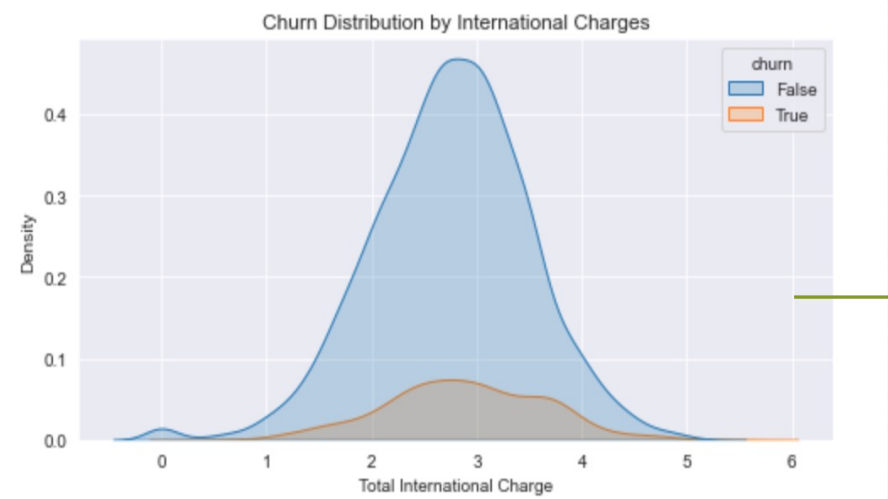


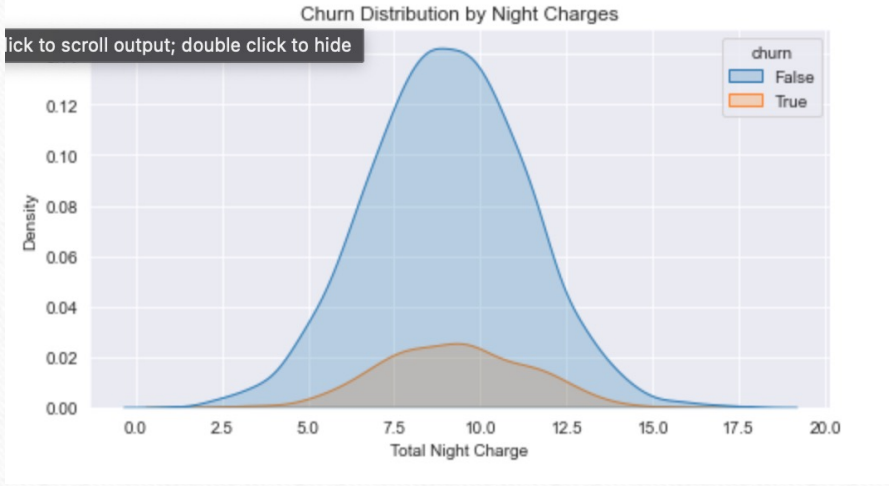
Categorical Features Distribution by churn rate

- Of all the customers that churned, majority are from Texas, New Jersey, Maryland, Miami and NewYork.
- Many customers who churned did not have a voicemail plan.
- Many customers who churned did not have an international plan.

Distributions of churn feature charges

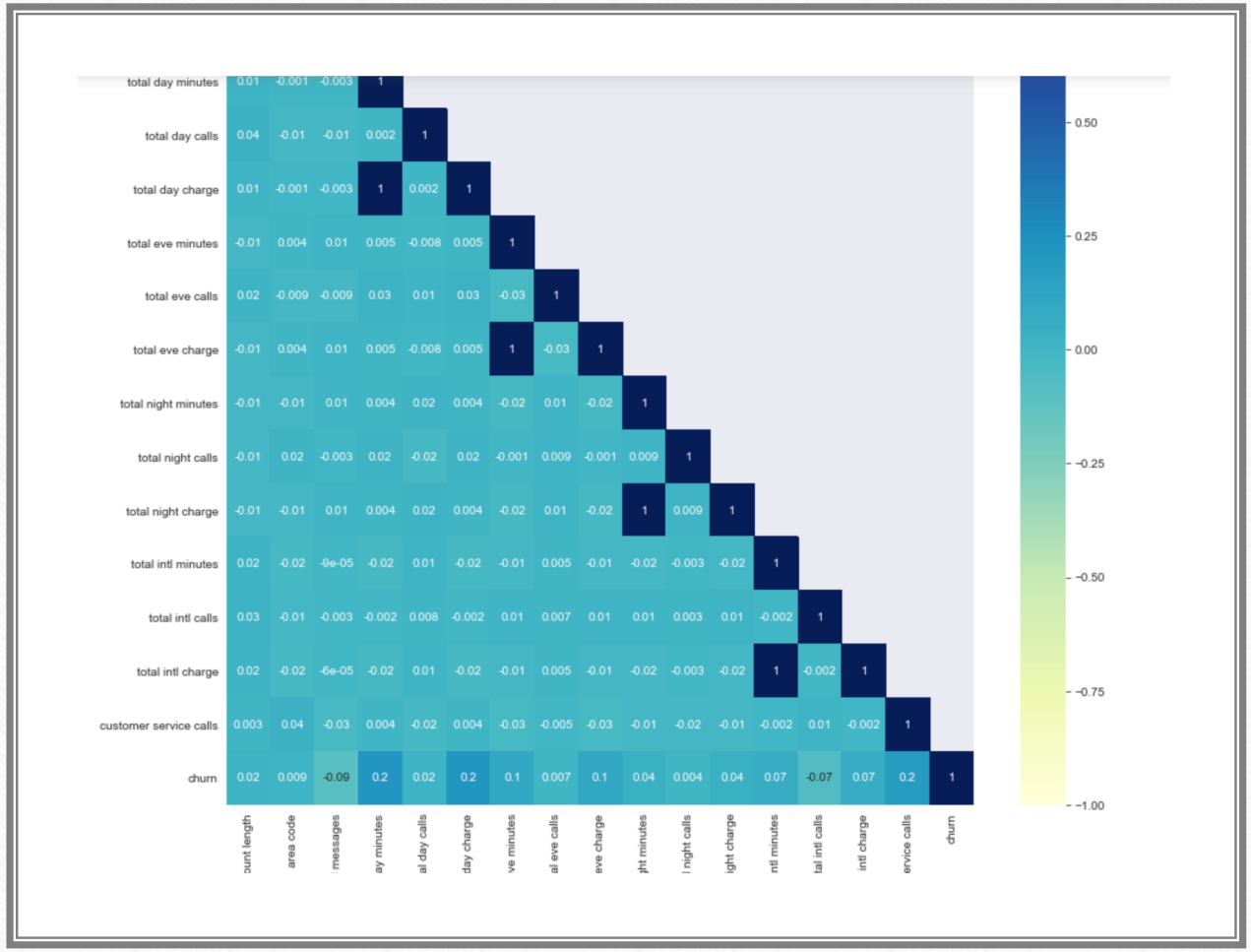
- The KDE plot indicates that customers who have churned tend to have higher total day charges compared to those who have not churned. This observation suggests that dissatisfaction or perceived high expenses during the day may contribute to a higher likelihood of churn among customers.
- Similar to the churn by day charges plot, the KDE plot for churn by evening charges also demonstrates a comparable trend. Customers who have churned exhibit higher total evening charges compared to those who have not churned.





Distributions of churn feature charges(cont)

- The KDE plot for churn by night charges follows a pattern similar to that observed in the churn by day charges and churn by evening charges plots. Customers who have churned typically exhibit higher total night charges compared to those who have not terminated their accounts.
- The plot implies that customers with higher total international charges exhibit a slightly higher probability of churning.

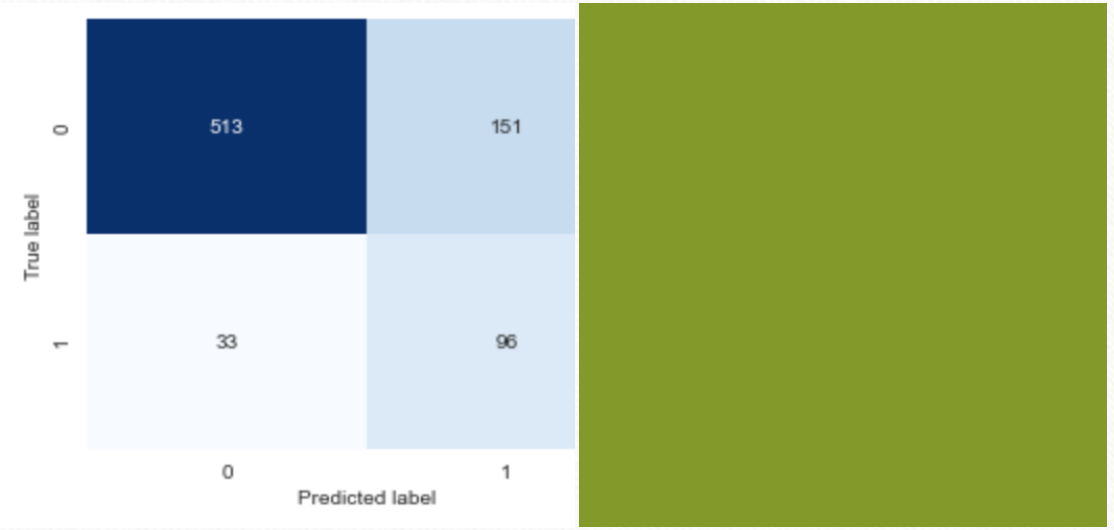


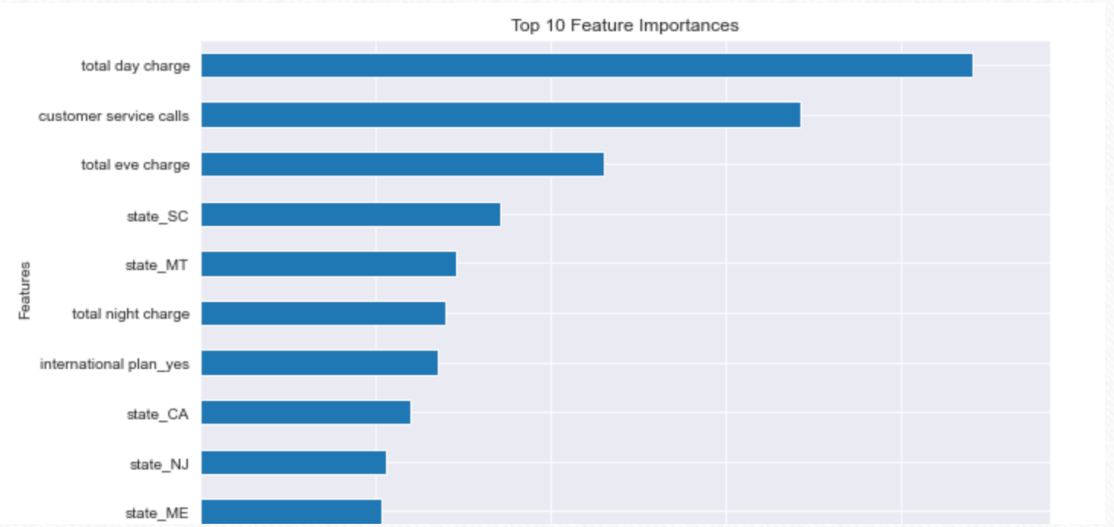
Correlation Heatmap

- The majority of features do not exhibit strong correlations, but some display perfect correlations:
- Total day charge and total day minutes features are fully positively correlated.
- Total eve charge and total eve minutes features are fully positively correlated.
- Total night charge and total night minutes features are fully positively correlated.
- Total int charge and total int minutes features are fully positively correlated.
- This perfect correlation is justifiable because the charge is directly influenced by the minutes used.

Logistic Regression

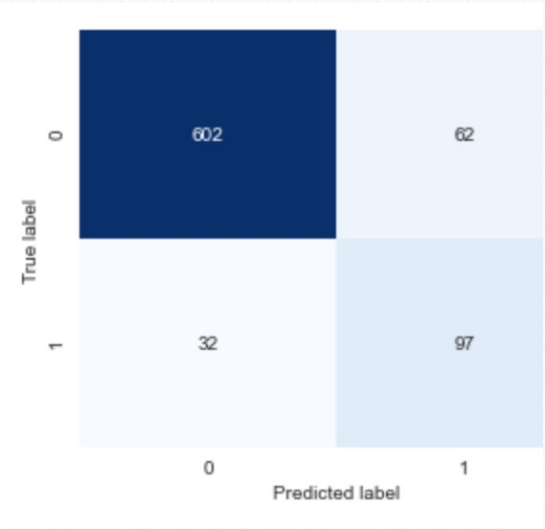
- The logistic regression model achieved a good recall score of 0.74, signifying its effectiveness as a baseline model. This indicates that it accurately identifies approximately 74% of the positive instances.
- Furthermore, an analysis of the confusion matrix revealed a higher count of true positives and true negatives compared to false positives and false negatives. This suggests that the model consistently makes accurate predictions and does not suffer from overfitting.
- In terms of feature importance, the model identifies total day charge, customer service calls, and total eve charge as the three most influential factors.

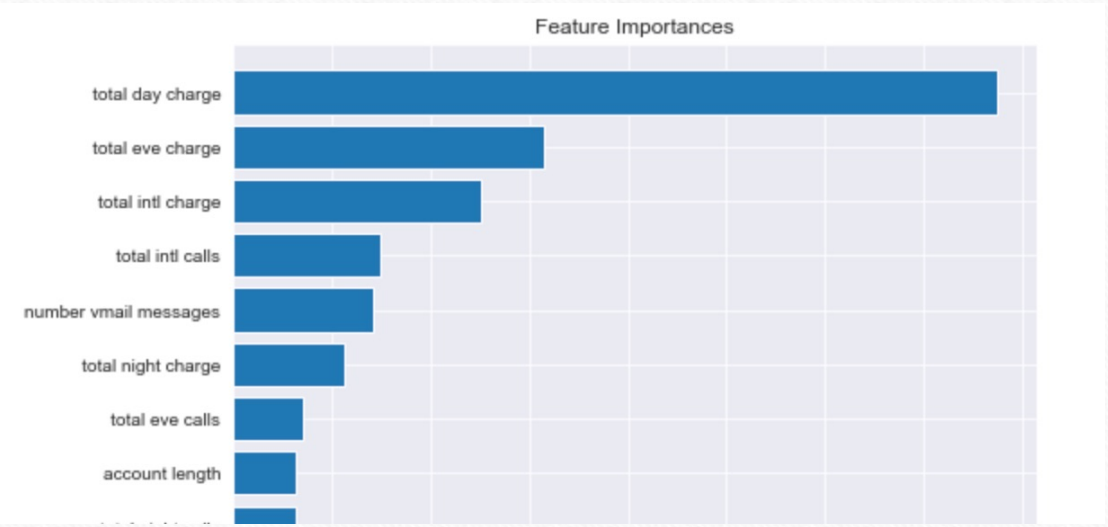




Decision Tree Classifier

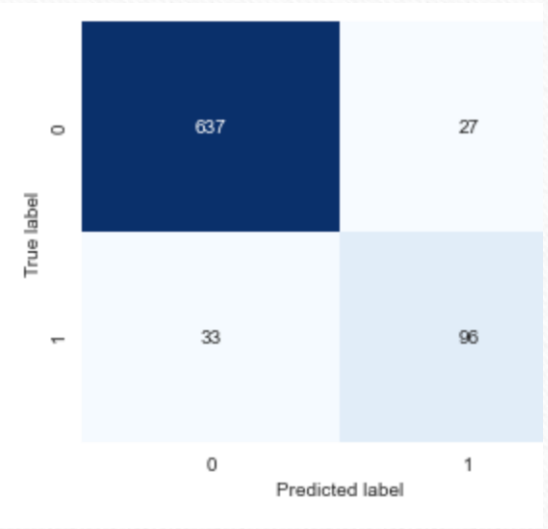
- The decision tree model achieves a recall score of 0.75, which, while respectable, falls slightly short of surpassing our baseline model. This signifies that the model accurately identifies approximately 75% of the true positive instances.
- Analysis of the confusion matrix reveals a notable prevalence of true positives and true negatives over false positives and false negatives. This suggests that the model predominantly makes accurate predictions and does not suffer from overfitting.
- Furthermore, the model identifies total day charge, total eve charge, and total intl charge as the top three most influential features in predicting outcomes.

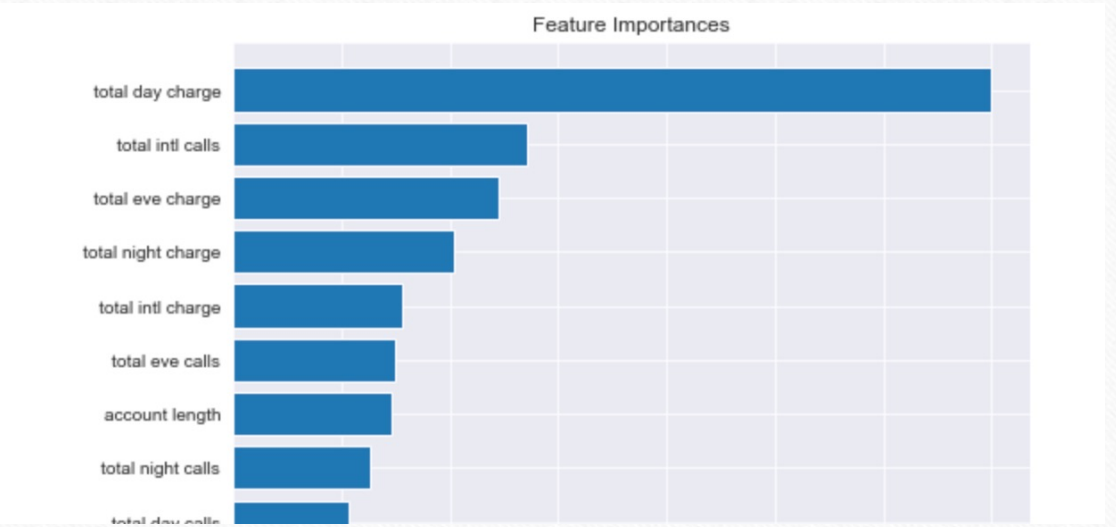




Random Forest Classifier

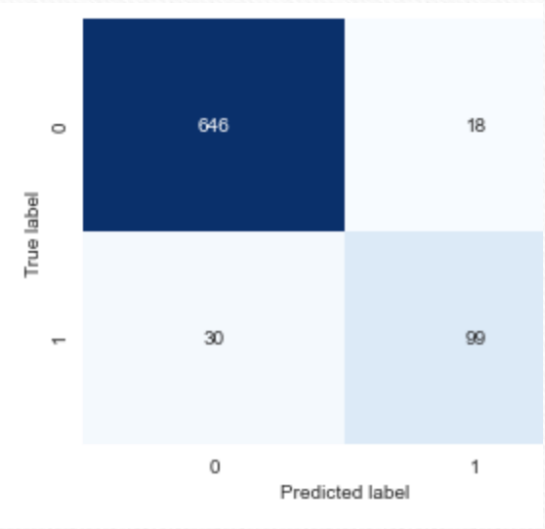
- The random forest classifier model achieves a commendable recall score of 0.74, signifying a notable improvement over the previous model. This implies that the model adeptly identifies approximately 74% of the actual positive instances correctly.
- Upon analyzing the confusion matrix, it becomes evident that the model exhibits a higher count of true positives and true negatives in comparison to false positives and false negatives. This observation suggests that the model consistently makes accurate predictions, indicating it is not overfitting the data.
- As per the model's assessment, the top three most influential features are total day charge, total intl calls, and total eve calls.

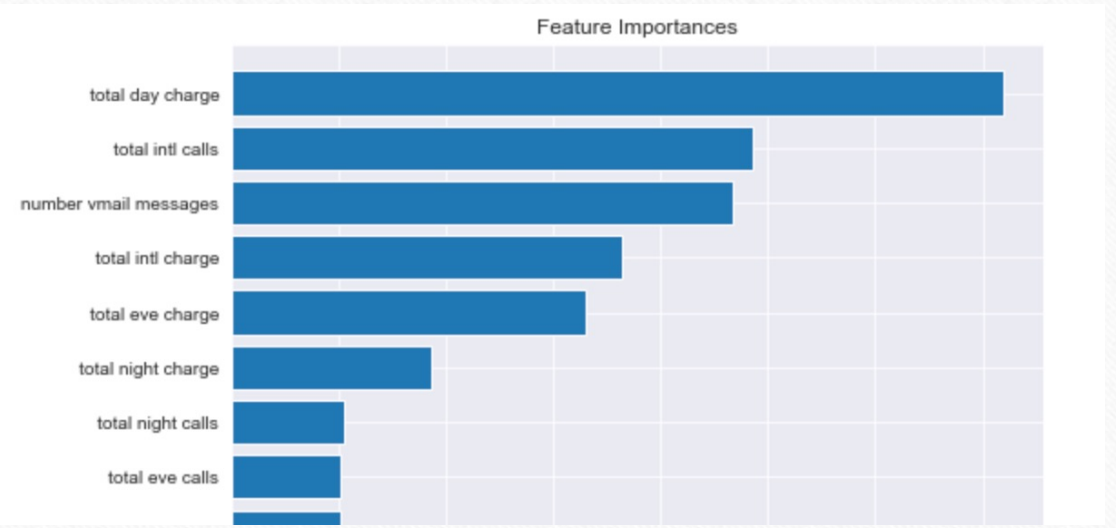




XGBoost Classifier

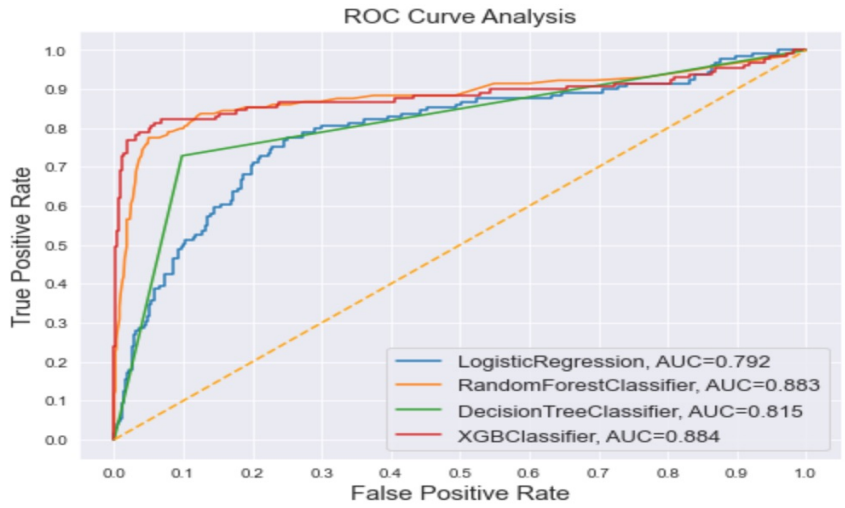
- The XGBoost classifier model showcases a remarkable recall score of 0.77, surpassing the performance of all prior models. This implies that the model adeptly identifies approximately 77% of the actual positive instances.
- Upon evaluating the confusion matrix, it becomes evident that the model exhibits a greater number of true positives and true negatives compared to false positives and false negatives. This suggests that the model consistently makes accurate predictions and avoids overfitting.
- As per the model's analysis, the top three most significant features are total day charge, total intl calls, and number vmail messages.



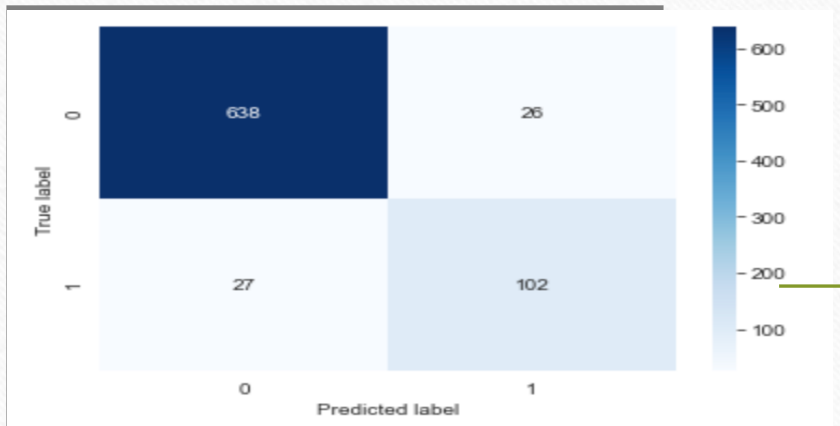


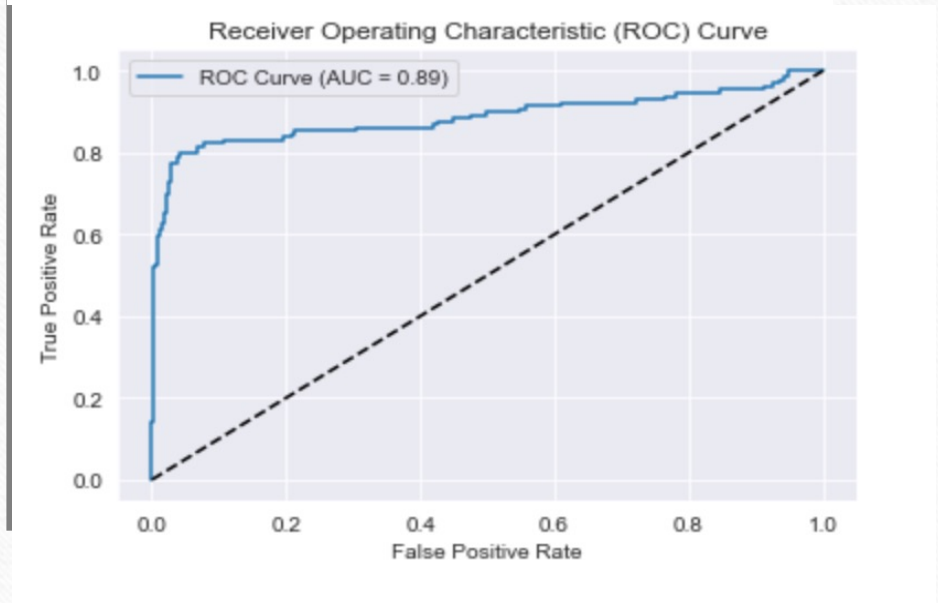
Model Evaluation

classifiers	recall
LogisticRegression	0.744186
RandomForestClassifier	0.744186
DecisionTreeClassifier	0.728682
XGBClassifier	0.767442



- The XGBoostClassifier achieved the highest recall score, followed by the RandomForestClassifier and LogisticRegression. The DecisionTreeClassifier, on the other hand, attained the lowest recall score of 0.73.
- The analysis of the ROC curves reveals that the XGBClassifier exhibits the most favorable performance, followed by the RandomForestClassifier, DecisionTreeClassifier, and LogisticRegression respectively. Specifically, the XGBClassifier achieves the highest AUC score of 0.884, while the LogisticRegression attains the lowest AUC score of 0.79.





Model Tuning

- The analysis of the ROC curve and the recall metric reveals that the tuned XGBoost model outperforms RandomForest slightly in differentiating between positive and negative classes (churned and non-churned customers) and accurately identifying churned customers.
- With a recall score of 0.79, the model successfully captures 79% of the actual churned customers. This result is close to our target recall score of 0.8, demonstrating the effectiveness of the model in identifying churned customers.

Conclusion

- It's commendable that despite achieving a respectable recall score of 79% with our XGB classifier, there's acknowledgment of the potential for further improvement through additional feature engineering. This recognition underscores a commitment to refining the predictive model and maximizing its effectiveness in identifying customers at risk of churn.
- Despite not achieving an ideal recall score, predicting customer churn while maintaining an acceptable level of performance demonstrates the efficiency of the models employed.

Recommendations & Next Steps

- Customer Segmentation: Further analyze customer segments to understand the specific characteristics and behaviors of customers more likely to churn. This understanding can guide targeted marketing strategies and personalized retention efforts.
- Tailored Retention Strategies: Develop tailored retention strategies based on the identified customer segments. These strategies could include personalized offers, loyalty programs, and proactive outreach to at-risk customers to address their concerns and incentivize them to stay.
- Enhanced Customer Experience: Continuously focus on enhancing the overall customer experience across all touchpoints, including customer service, billing processes, and service quality. Satisfied customers are less likely to churn, so investing in customer satisfaction can be a powerful retention strategy.

Recommendations & Next Steps(cont)

- Predictive Analytics Integration: Integrate predictive analytics models, such as the churn prediction model developed in this project, into SyriaTel's operational processes. By leveraging real-time data and predictive insights, the company can identify potential churners early and take proactive measures to retain them.
- Feedback Mechanisms: Implement feedback mechanisms to gather insights from churned customers about their reasons for leaving. Analyzing this feedback can uncover underlying issues or areas for improvement in SyriaTel's products or services, helping to reduce churn in the future.
- Employee Training and Engagement: Invest in employee training and engagement programs to ensure that frontline staff are equipped to address customer concerns and deliver exceptional service. Empowered and motivated employees play a crucial role in retaining customers and fostering long-term relationships.

Thank You

- Q/A Session