Data Report: Business Understanding and Approach to Reducing Customer Churn

1. Business Problem Overview

a.Problem Statement

Customer churn is a critical issue for many businesses, especially in the telecommunications industry. High churn rates not only lead to lost revenue but also incur high costs related to acquiring new customers. For SyriaTel, understanding which customers are at risk of leaving and taking measures to retain them can significantly reduce churn, improve customer loyalty, and boost overall profitability.By doing so, SyriaTel can improve customer retention and ensure sustainable growth.

b. Objective of the Analysis

The primary objectives of this analysis are:

- Develop a Predictive Model: I aim to develop a model that predicts the likelihood of churn for each customer, with a focus on achieving a recall score of 0.7. This ensures that 70% of customers who are likely to churn are correctly identified.
- Feature Analysis: We will analyze the key features influencing customer churn, including customer service interactions, usage patterns, and account details.
- Provide Retention Recommendations: Based on the model's predictions, we will offer actionable recommendations for reducing churn, such as targeted promotions or improving customer support.

c. Key Metrics of Success

To measure the success of this project, I will focus on the following metrics:

- Recall Score: A recall score of 0.7,ensuring that 70% of at-risk customers are correctly identified.
- Precision and F1 Score: Balancing recall with precision to minimize false positives, ensuring that the churn predictions are both accurate and actionable.
- Churn Reduction: Evaluating the reduction in churn rate after implementing retention strategies based on our model's predictions.

2. Data Understanding

The following dataset (SyriaTel telecom dataset) is typically available in online data science and machine learning platform (Kaggle). It is frequently used in projects focused on predicting customer churn and contains anonymized records of customer information.SyriaTel is a telecommunication company based in Syria, whose dataset of 20 columns and 3,333 rows.The columns and rows represented the following:

Categorical Features:

- state: The state where the customer lives.
- phone number: The customer's phone number.

- international plan: Does the customer have an international plan?(Yes or No).
- voice mail plan: Whether the customer has a voice mail plan (Yes or No).

Numeric Features:

- area code: The area code associated with the customer's phone number.
- account length: The number of days the customer has been an account holder.
- number vmail messages: The number of voice mail messages received by the customer.
- total day minutes:The total number of minutes the customer used during the day.
- total day calls: The total number of calls made by the customer during the day.
- total day charge:The total charges paid by the customer for daytime usage.
- total eve minutes: The total number of minutes the customer used during the evening.
- total eve calls: The total number of calls made by the customer during the evening.
- total eve charge: The total charges paid by the customer for evening usage.
- total night minutes: The total number of minutes the customer used during the night.
- total night calls: The total number of calls made by the customer during the night.
- total night charge:The total charges paid by the customer for nighttime usage.
- total intl minutes: The total number of international minutes used by the customer.
- total intl calls: The total number of international calls made by the customer.
- total intl charge: The total charges incurred by the customer for international usage.
- customer service calls:The number of customer service calls made by the customer.

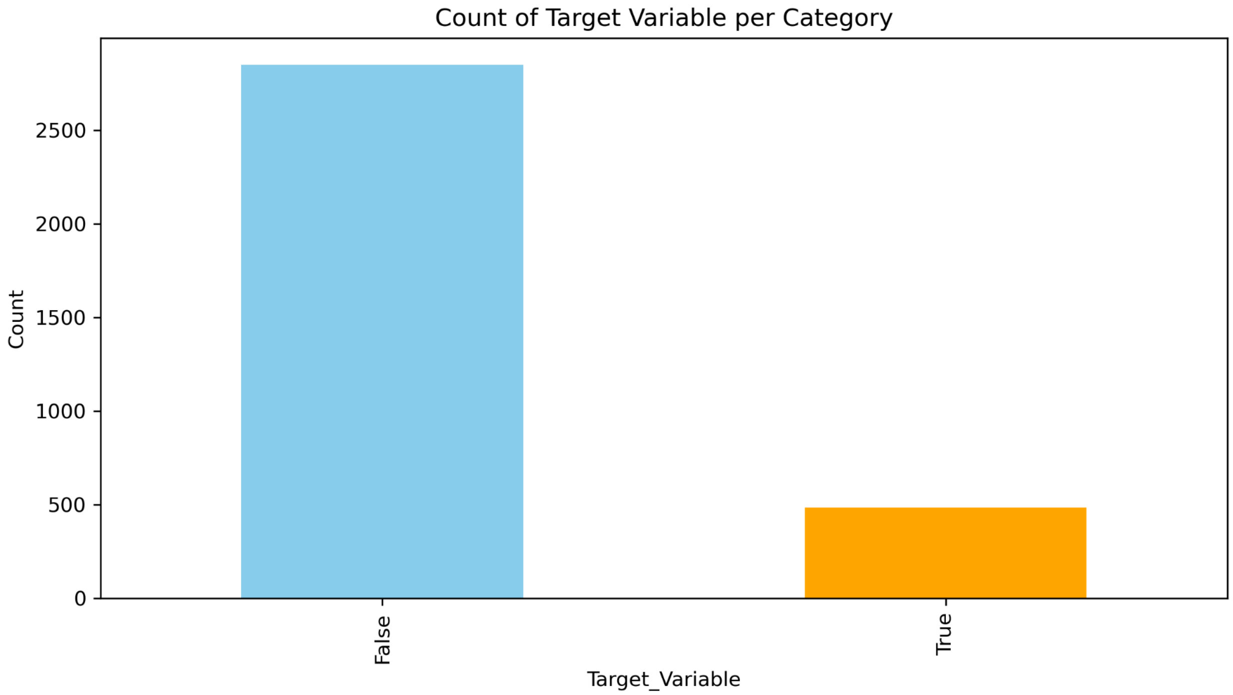
3. DATA PREPARATION

I checked the dataset for missing values, unique values, and duplicates, and found no issues that required attention. Additionally, I identified and removed the phone number column, as it does not influence whether a customer churns or not.I also renamed the states to provide a clearer view of which state each refers to. Furthermore, I binned the account length to prepare it for data analysis, allowing for the differentiation of short-term, medium-term, and long-term customers.

4. DATA ANALYSIS

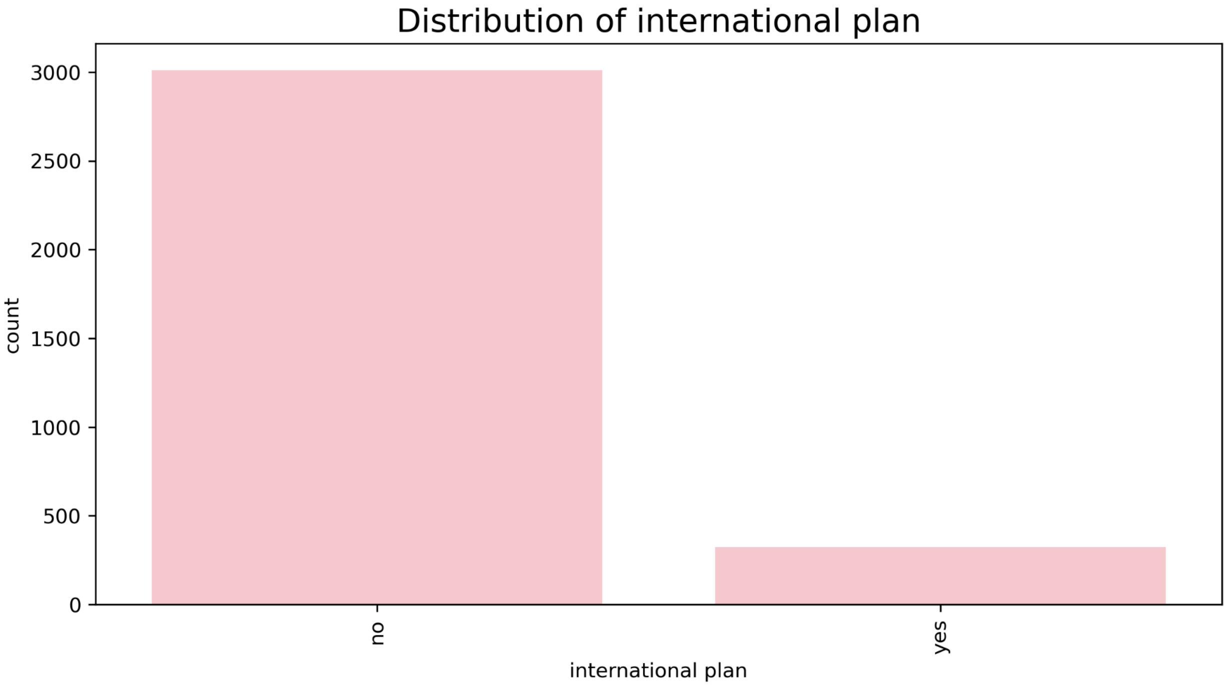
4.1 Univariate Analysis

- **Checked the normalized distribution of my target variable (‘Churn’)**



- The data is highly imbalanced
- Most customers stay with the company (85.5%), while only a small group (14.5%) leave. This difference means we need to ensure the model can still focus on predicting those who leave.

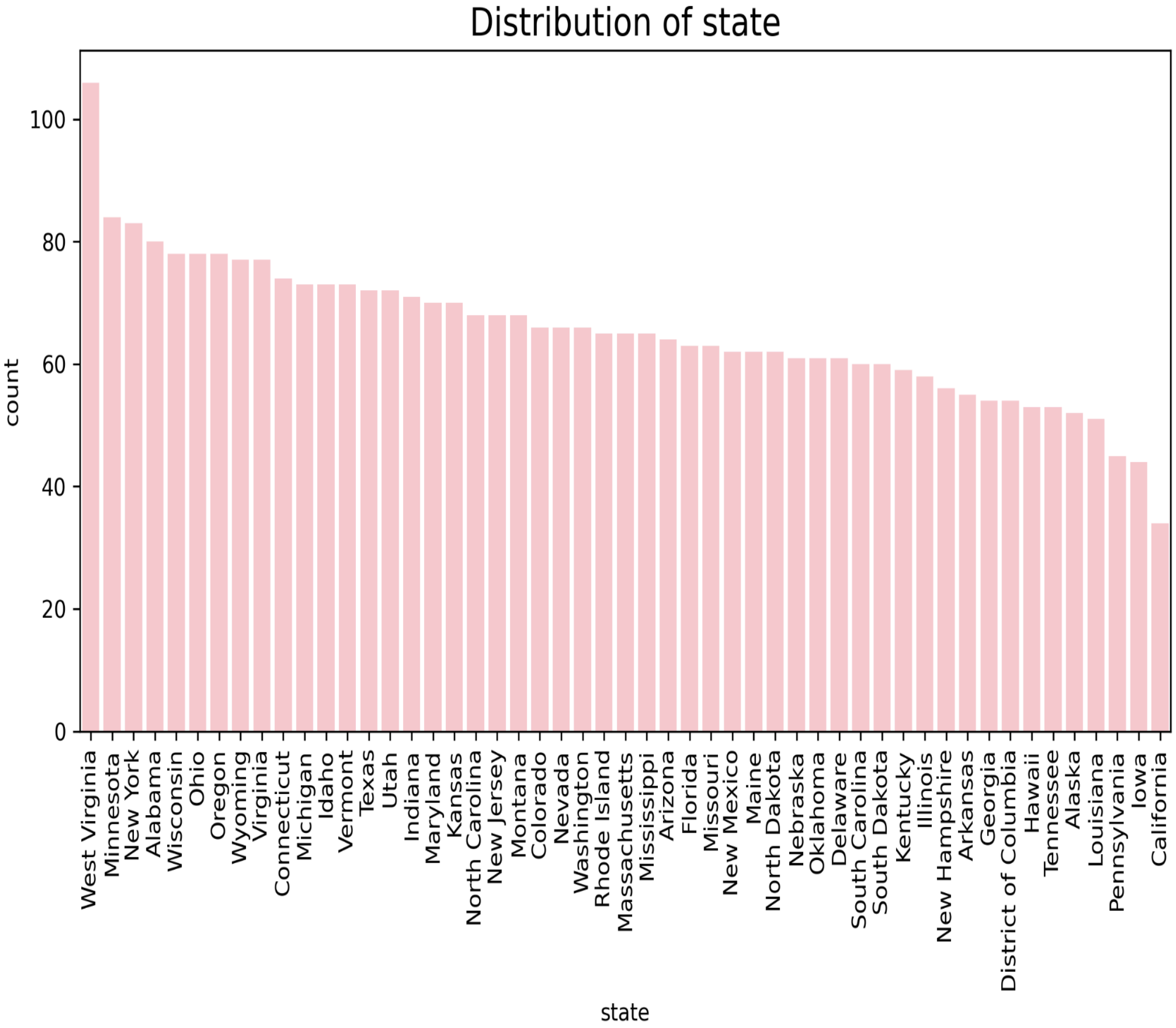
INTERNATIONAL PLAN



No active International Plan -
323 customers out of 3333 customers have an international plan.

- **DISTRIBUTION OF STATE**

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



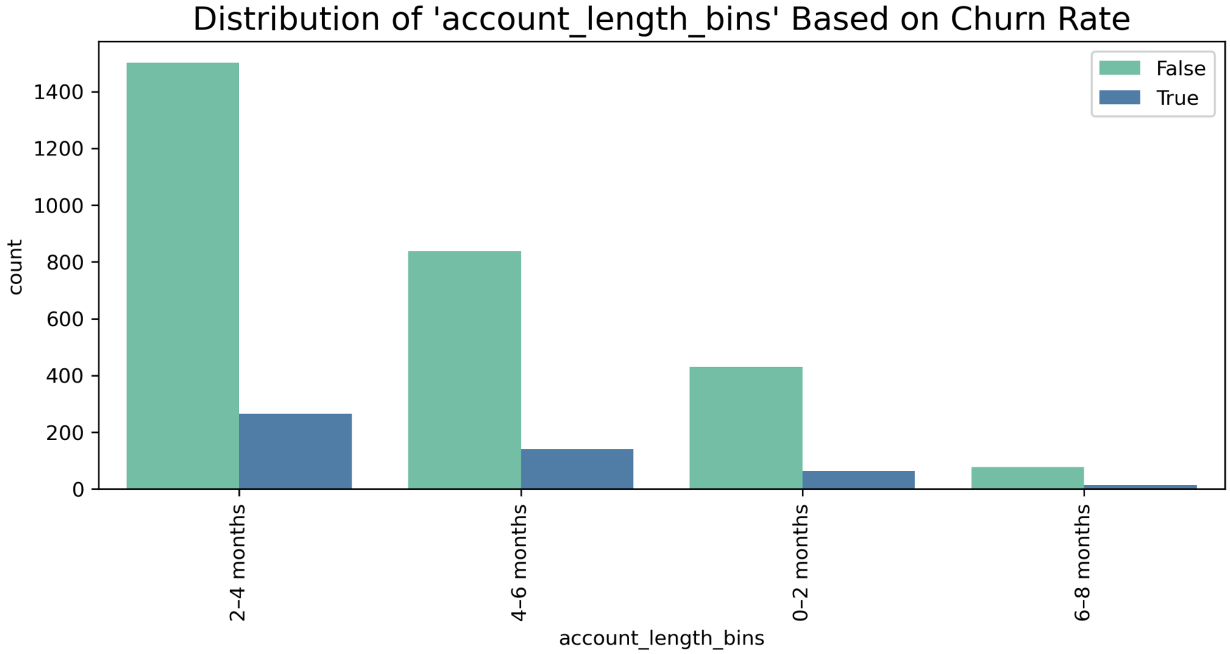
Most of the customers are from West Virginia,Minnesota,NewYork,Alabama and Wisconsin.

4.2 BIVARIATE ANALYSIS

- Analysis of how the target values interact with other features in the dataset to uncover useful patterns and insights.

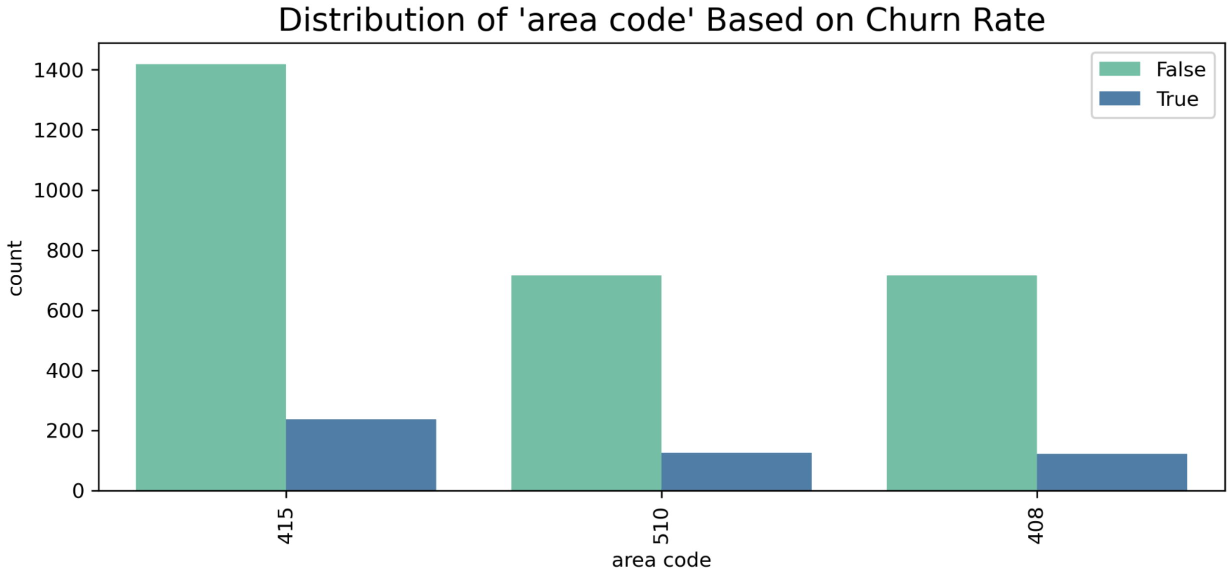
Account Length Bins

- The number of customers churning decreases as theaccount length (in months) increases suggests that customer loyalty strengthens over time. However, also for the first few months customers can also churn.This implies that customers who stay with the company for longer periods are less likely to churn.



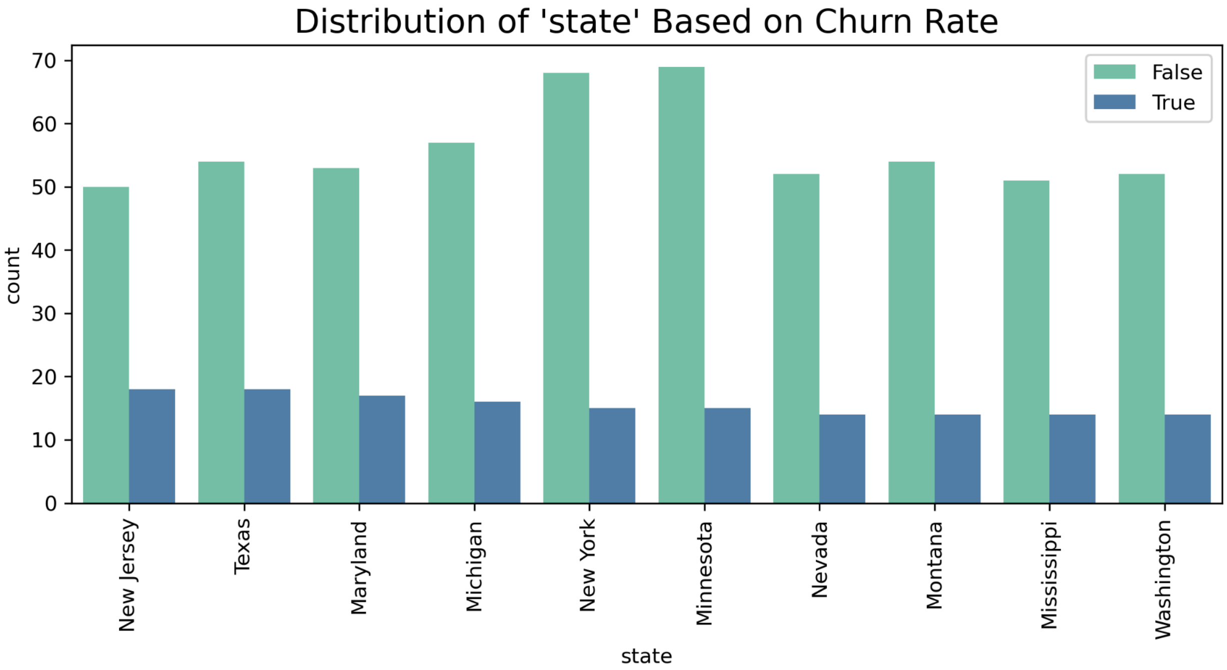
Area Code

- Of the customers who have terminated their account, most of them are from area codes 415 and 510



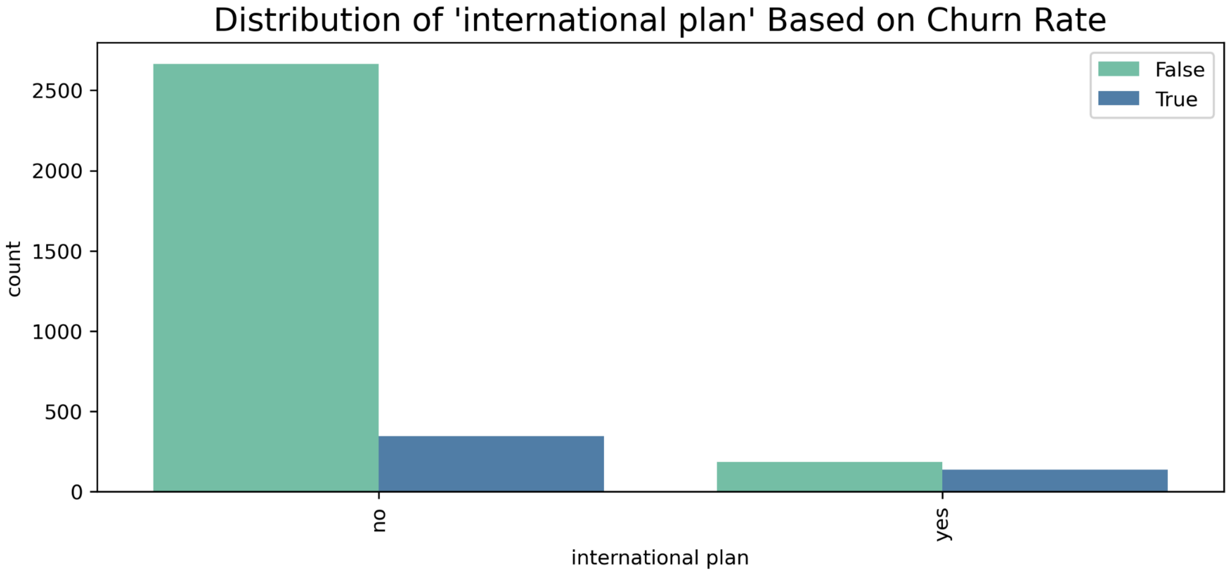
State

- Most of the customers who churned were from New Jersey,Texas,Maryland and Michigan.



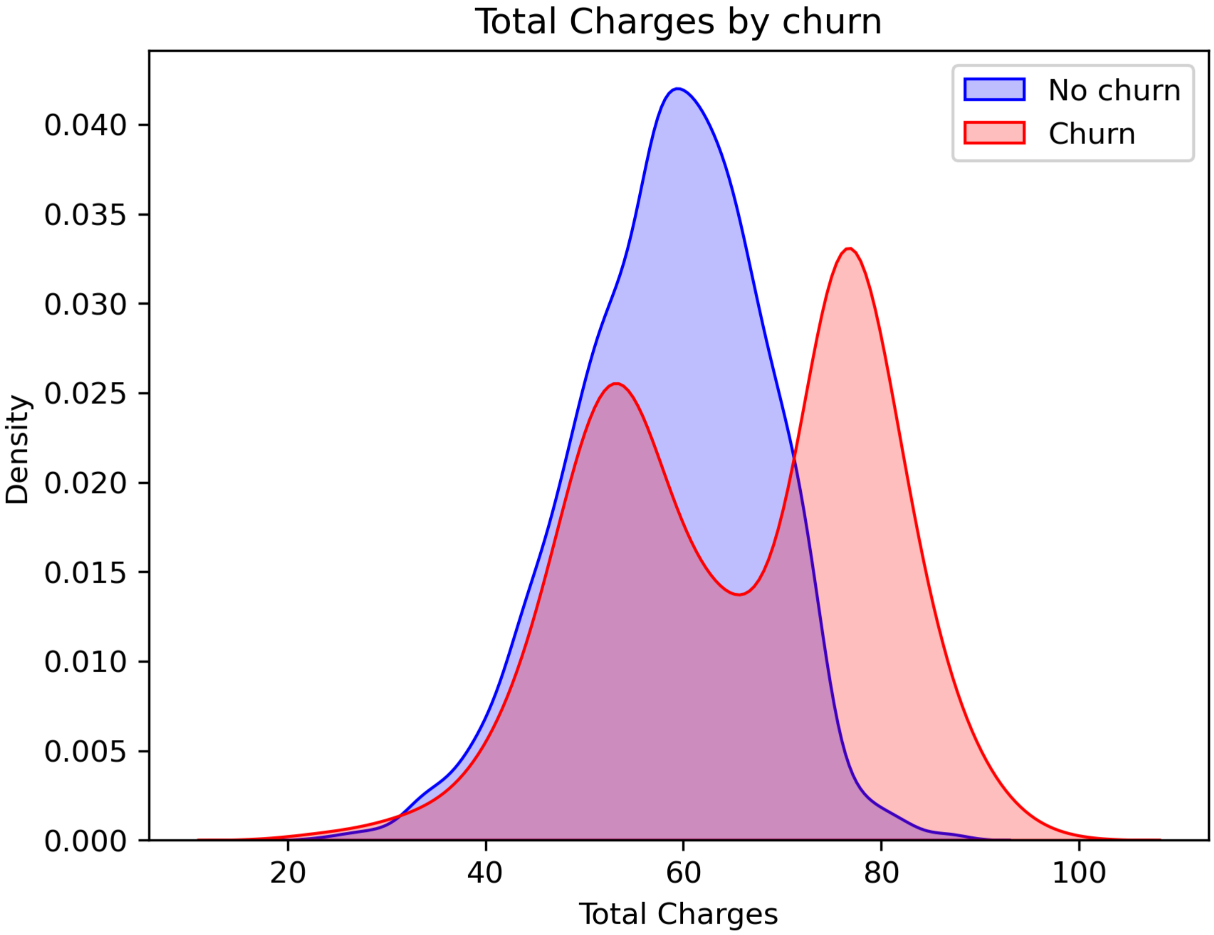
International Plan

- Customers with an international plan are less likely to churn.



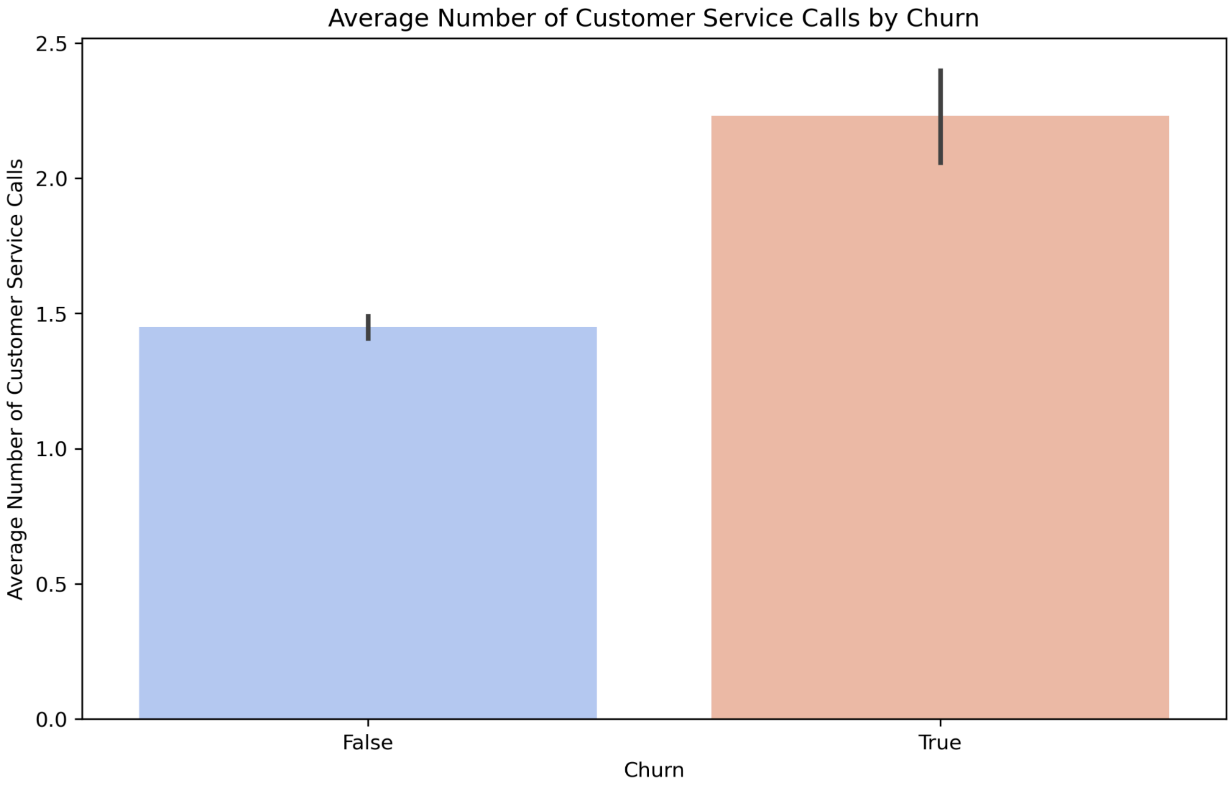
Further analysis revealed that individuals with an active international plan have a churn rate of 42%. This is a concerning statistic, highlighting the need for a strategic restructuring plan to address and mitigate churn among these customers.

Total Charge



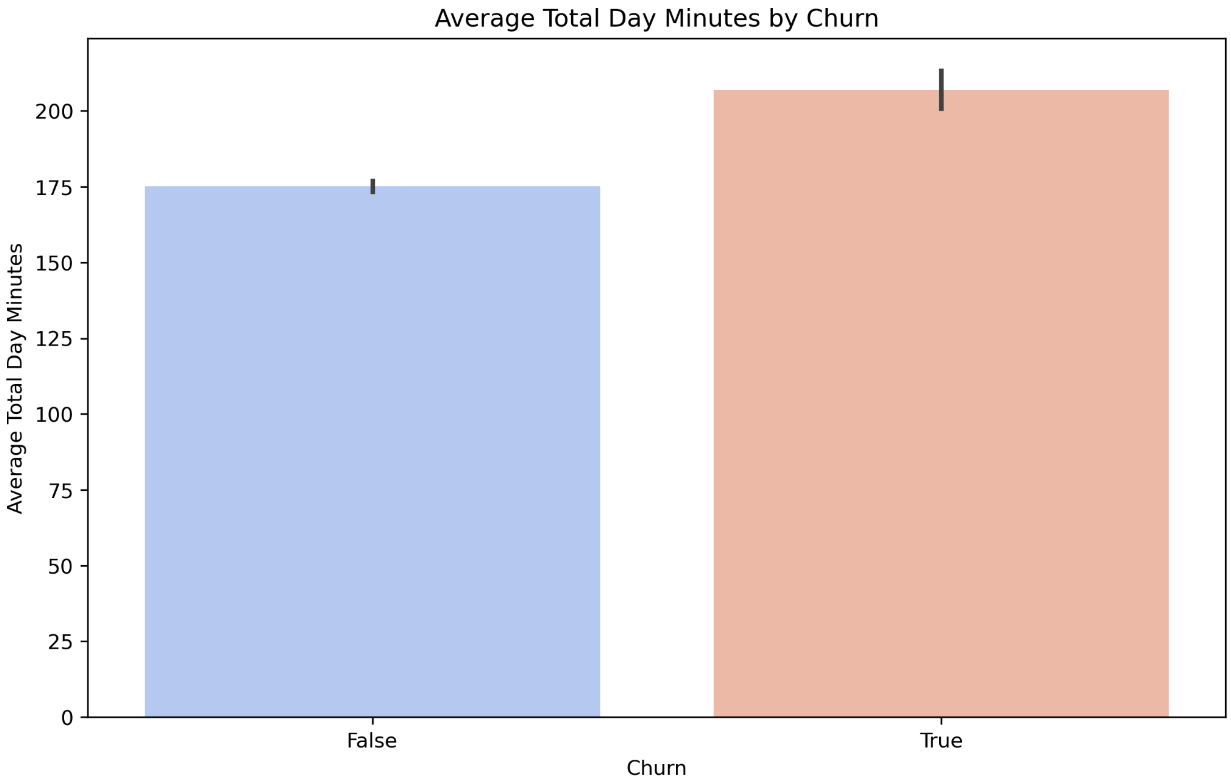
- This suggests that customers who stay with the company tend to have moderate total charges compared to those who churned.
- Customers who churn tend to have higher total charges. This might indicate that higher charges could be correlated with dissatisfaction or the likelihood of leaving.

Customer Service Call



On average, customers who churned tend to make more customer service calls, which could suggest that they experienced issues that led to them leaving the service. This could be important for retention strategies, as addressing customer concerns may reduce churn.

Total Day Minutes

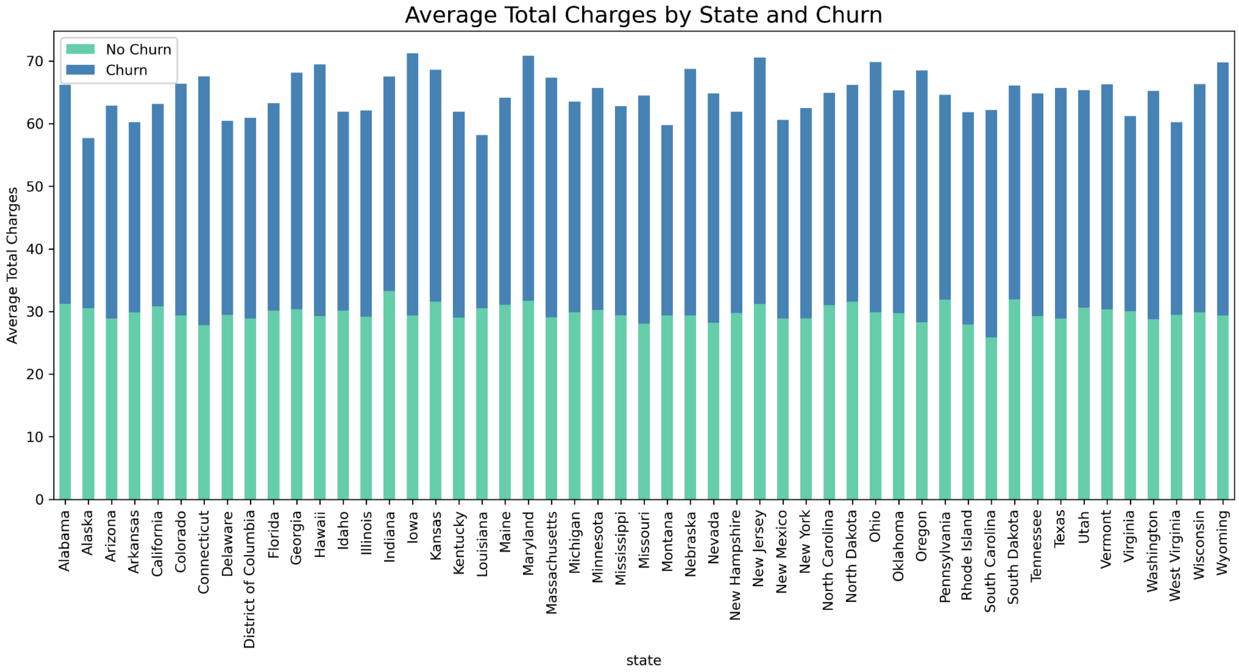


- Day Minutes appear to be the most informative among the three features that have minutes details for churn prediction, assuming it shows a meaningful difference between churned and non-churned customers.
- International and Night Minutes may have lower predictive power for churn.

4.3 MULTIVARIATE ANALYSIS

- Analyze the relationships between multiple variables in the dataset to understand how they interact and influence the target variable (Churn). This includes identifying potential correlations, dependencies, and trends among features**.**

- **Total Charge**



Texas,Maryland,New Jersey have one of the highest churn rates and this shows that that the total day charges have an impact to why most people must be leaving the company from those states.

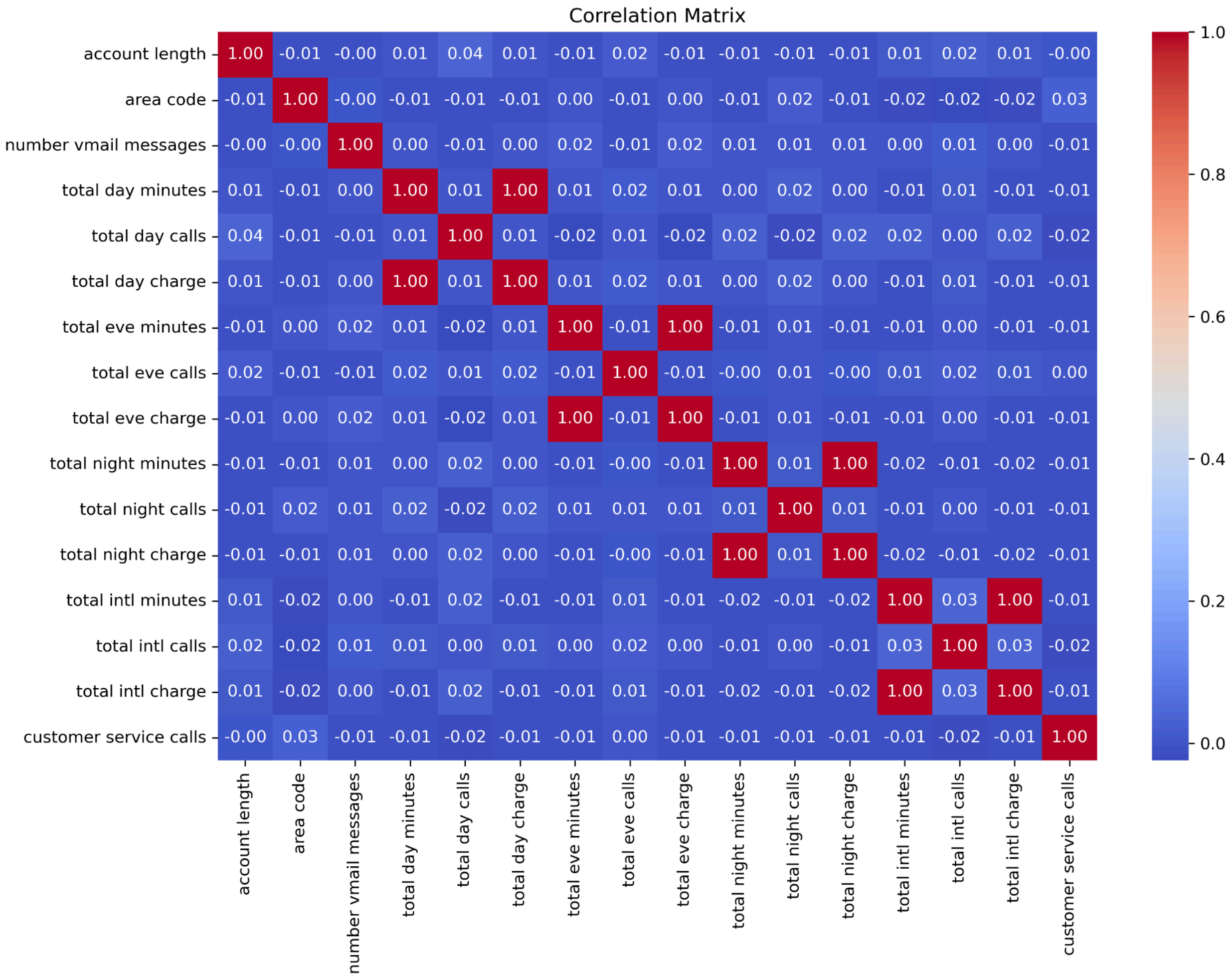
Area code

The most areas that people churn are 415 and 510 but the question is why I did a comparison of high churn areas and other areas on average with total day charge,customer service,account length and total international charge.

Customer Service Calls: Higher calls might point to dissatisfaction of services in area 415 and 510 leading to high churn_rate.

Total Day Charge: The slightly higher charges might lead to perceived overpricing, contributing to churn.

CORRELATION ANALYSIS



- * total day charge, total eve charge, total night charge, total intl charge, and total day minutes indicate high multicollinearity.
- * Drop total day charge, total eve charge, total night charge, and total intl charge because they are directly derived from the "minutes" .The more minutes someone has the more charge they incur.

5. MODELING

- **Models Used**

I used Logistic Regression and Decision Tree Classifier.

I chose Logistic Regression for its simplicity in predicting churn. It gives clear and interpretable probabilities, helping us identify customers at risk. I focused on **recall** to catch as many churned customers as possible.

I used Decision Tree to capture more complex patterns and interactions that logistic regression might miss. It’s easy to understand and shows which features matter most. Like logistic regression, **recall** was the priority to minimize missed churn predictions.

Both models help us make sure we’re not overlooking customers who might leave.

EVALUATION

Baseline Model

- Confusion Matrix Analysis: The confusion matrix shows that the model predicted 448 true negatives and 75 true positives, demonstrating its strength in identifying non-churning customers. However, the 118 false positives and 26 false negatives highlight the trade-off between precision and recall.

- For the "No Churn" class (False): The model exhibits a solid precision of 0.95, indicating a high likelihood that non-churning predictions are correct. However, the recall of 0.79 suggests room for improvement in identifying all non-churning customers.

- For the "Churn" class (True): The precision of 0.39 indicates that when the model predicts churn, it is correct approximately one-third of the time. The recall of 0.74, while moderately high, points towards the potential to capture more churn cases.

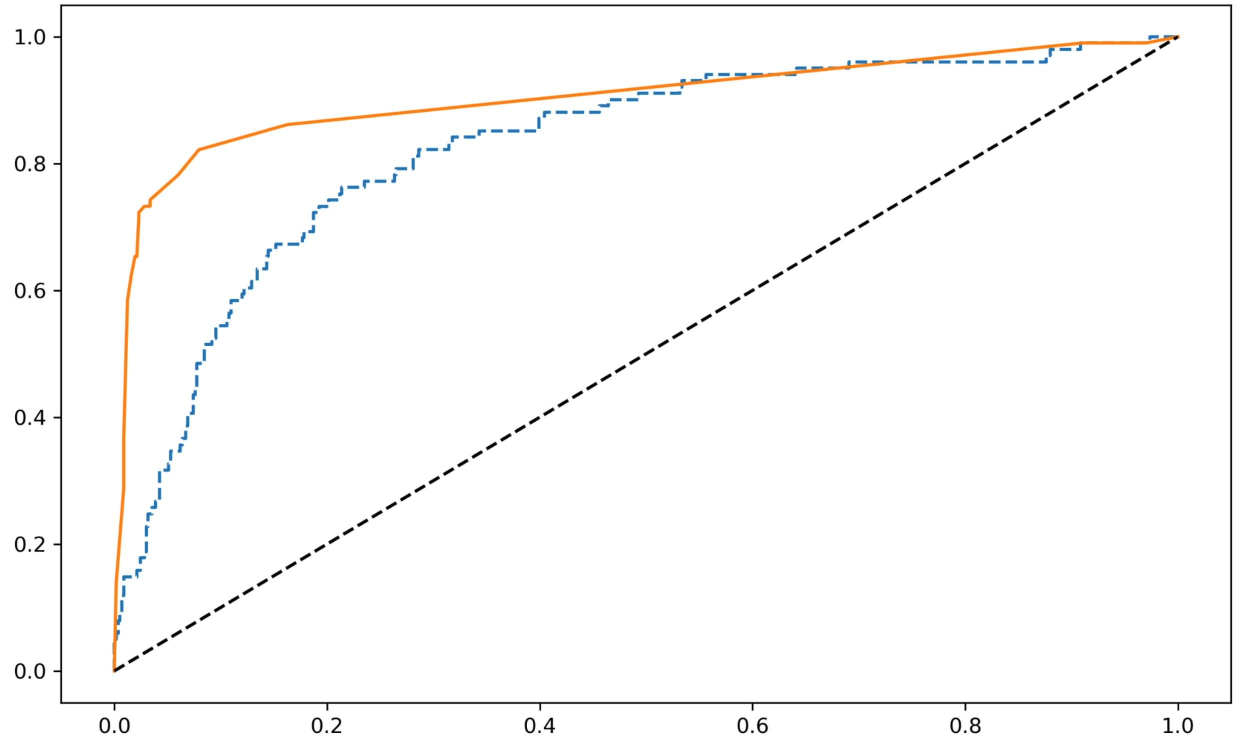
Decision Tree Classifier Model

- Accuracy: The decision tree model is more accurate overall, improving from 78.4% to 87.9%.
- Recall: Both models have good recall for churn, but the decision tree performs slightly better, with 75.2% versus the baseline's 74.3%.
- Precision: The decision tree's precision is higher, making it more reliable in correctly predicting who will churn.

Tuned Decision Tree Classifier Model

- The tuned Decision Tree model shows an improvement in accuracy (0.90 vs. 0.88) and precision (0.63 vs. 0.58) compared to the baseline model. Recall remains the same (0.75), which is crucial for correctly identifying churned customers. The F1-score also improves (0.69 vs. 0.65), indicating better balance between precision and recall. The ROC-AUC of 0.88 highlights the model's ability to distinguish between churned and non-churned customers. Overall, the tuned model is more effective in targeting churned customers with better precision while maintaining strong recall.

ROC COMPARISON CURVE MODEL



- The ROC curve shows that the tuned Decision Tree (orange) outperforms the baseline Logistic Regression (blue). The Decision Tree has a higher true positive rate, meaning it's better at predicting customers who will churn. This is important for retention strategies, as it helps identify customers to focus on before they leave.

CONCLUSIONS

- Charges are directly derived from the "minutes" .The more minutes someone has the more charge they incur. This makes them have a linear dependency.

- Key features like total day minutes, customer service calls, and international plan_yes have the highest importance, meaning they are key drivers in predicting churn.

- Customers who have opted for the International Plan are observed to have a higher likelihood of churning compared to those who haven't chosen this plan.

- Customers who make more customer service calls exhibit a significantly higher churn rate.
- Some areas have a higher churn rate due to charges incurred.

- The tuned Decision Tree model has a higher ROC-AUC score (0.88) than the baseline model (0.83), demonstrating that it better distinguishes between churned and non-churned customers.

RECOMMENDATIONS

- Since customers who make more customer service calls are more likely to churn, it would be necessary to improve the customer service experience. Training staff to resolve issues more effectively or providing self-service options could reduce the need for frequent calls and improve retention.

- Customers who opt for the International Plan have a higher likelihood of churning.Analyzing why these customers are dissatisfied would reduce the churn rate.and consider offering special loyalty incentives or alternative plans to keep them engaged.

- Use the insights from the model to proactively reach out to customers who are likely to churn, especially those with high service usage or international plans.Offering them tailored deals and loyalty programs, or even educational content about optimizing their plans can increase their satisfaction and reduce churn.

- Certain regions show higher churn rates, you may want to investigate the local service quality or charge-related factors. Offering localized promotions or improving network coverage in those areas could help reduce churn.

- Customers may churn due to unexpected or high charges. Offering more transparency about how charges are calculated could help build trust and reduce churn.
- Focus on customers with high charges due to long call durations (high "total day minutes"). These customers are more likely to churn, so offering discounts or personalized plans could help keep them satisfied.

NEXT STEPS

- Experiment with different machine learning models (e.g., Random Forest, Gradient Boosting) to potentially improve churn prediction accuracy.
- Use churned customer feedback to improve customer support processes, such as faster response times or improved issue resolution.
- Collect more information of the customers.

○