SyriaTel Customer Churn Prediction

Presentation to the stakeholders.

Project Overview

- In the highly competitive telecom industry, customer churn—where customers discontinue their services—poses a significant challenge, with average churn rates ranging from 30% to 35%. This is especially critical for Syriatel, where customer retention is not only cost-effective but essential for maintaining market position and driving profitability. Given that acquiring a new customer is 5-10 times more expensive than retaining an existing one, reducing churn has become a top priority.

- Syriatel is one of the leading telecommunications companies in Syria, offering a wide range of mobile and data services to millions of customers across the country. As a key player in the telecom sector, Syriatel faces significant challenges in retaining customers due to the competitive nature of the industry. Understanding and reducing customer churn is vital for Syriatel to maintain its market position and continue providing high-quality services to its customers.

Business Understanding

- The objective of this project is to analyze customer behavior to predict and mitigate churn. By identifying high-risk customers through predictive modeling, Syriatel can focus its retention strategies more effectively, thereby enhancing customer loyalty and ensuring long-term business success. The insights gained from this analysis will enable Syriatel to implement targeted interventions that not only reduce churn but also optimize customer engagement, ultimately supporting the company's growth and profitability goals.

Stakeholders:

Syriatel Management: To strategize and implement customer retention programs.

Customer Service Teams: To identify and engage with high-risk customers effectively.

Data Analysts: To continuously monitor and refine the model for better accuracy.

Data Understanding

The dataset utilized in this project originates from Syriatel, containing detailed information about the company's customers. Each record corresponds to a unique customer, with attributes that provide insights into their interaction with Syriatel's services.

- The dataset includes 3333 rows and 21 columns, representing a substantial amount of data for robust analysis.

Key Features and Descriptive Statistics:

- State: The state where the customer resides. This categorical variable can help identify geographic patterns in churn.

- Account Length: The number of days the customer has had this account. Longer account durations might correlate with customer loyalty.

- Area Code: The area code associated with the customer's phone number. This categorical variable could influence service usage patterns.

Data Analysis

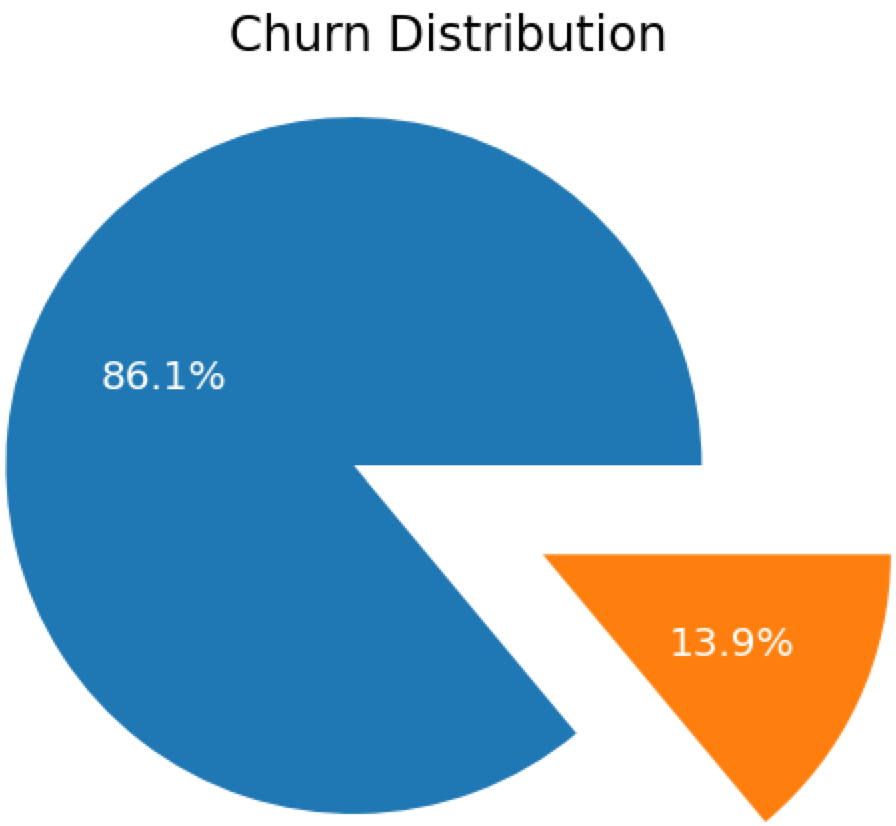
- **Summary of the analysis process.**
- **Data Collection:** Gathered data from various sources, including CSV and SQLite databases.
- **Data Cleaning:** Processed and cleaned data to ensure consistency and accuracy.
- **Exploratory Data Analysis:** Conducted analysis to identify key trends and patterns.
- **Modelling** : Creating different machine learning models to predict whether the customer will churn or not.

Key insights obtained from the data.

- Features like customer service calls, total day minutes, and total day charge show a positive correlation with churn, indicating that higher usage in these areas is associated with a higher likelihood of customer churn.

Data Analysis

- Techniques and tools used for analysis.
- **Techniques:**
  - Data Cleaning
  - Exploratory Data Analysis
  - Data Visualization
  - Modelling
- **Tools:**
  - Python (Pandas, Matplotlib, Numpy , Plotly, Seaborn, Scikit-learn)
  - Jupyter Notebook



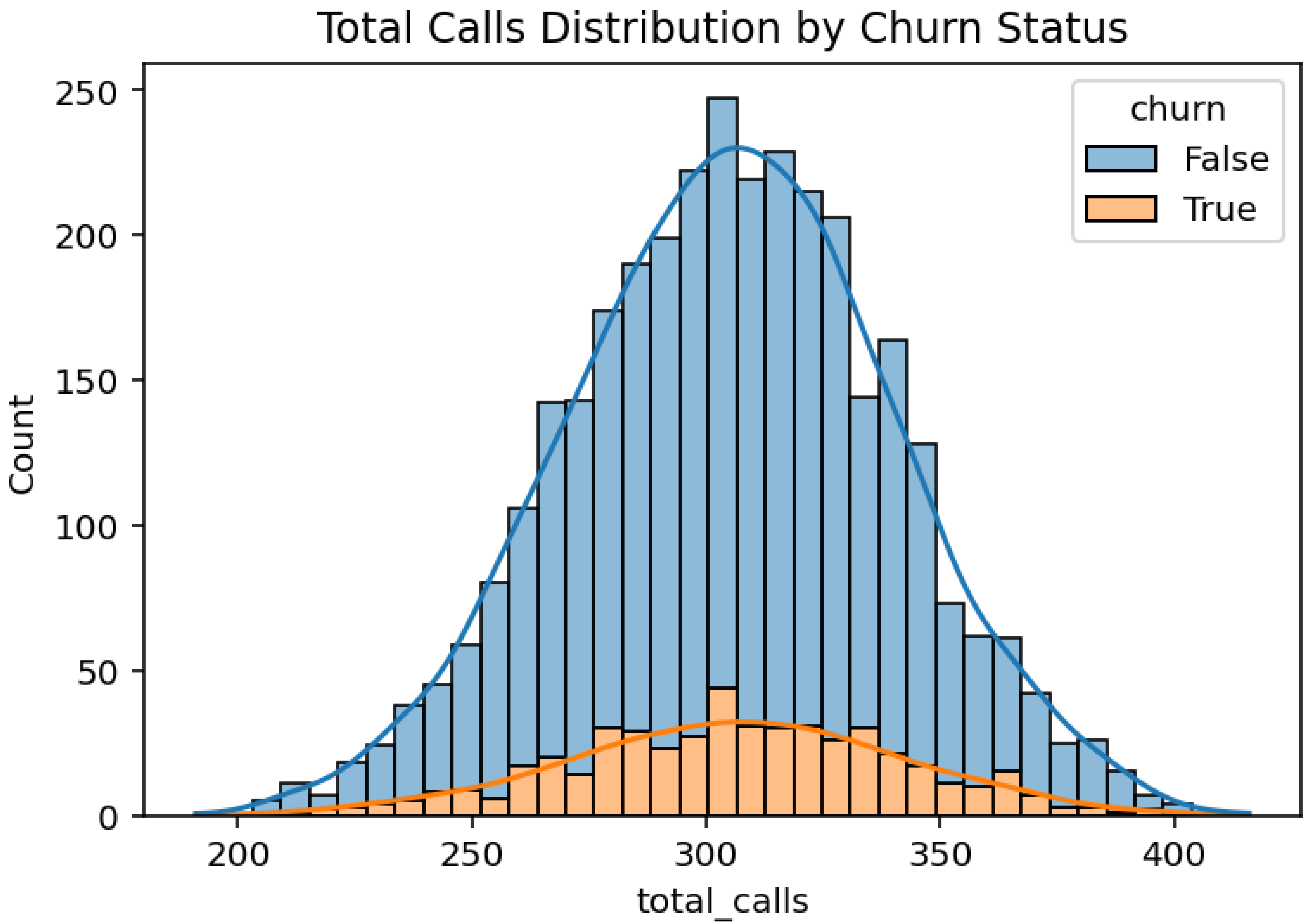
Findings

Churn Distribution:

85.5% of the people did not churn (i.e., their churn value is False).

14.49% of the people did churn (i.e., their churn value is True).

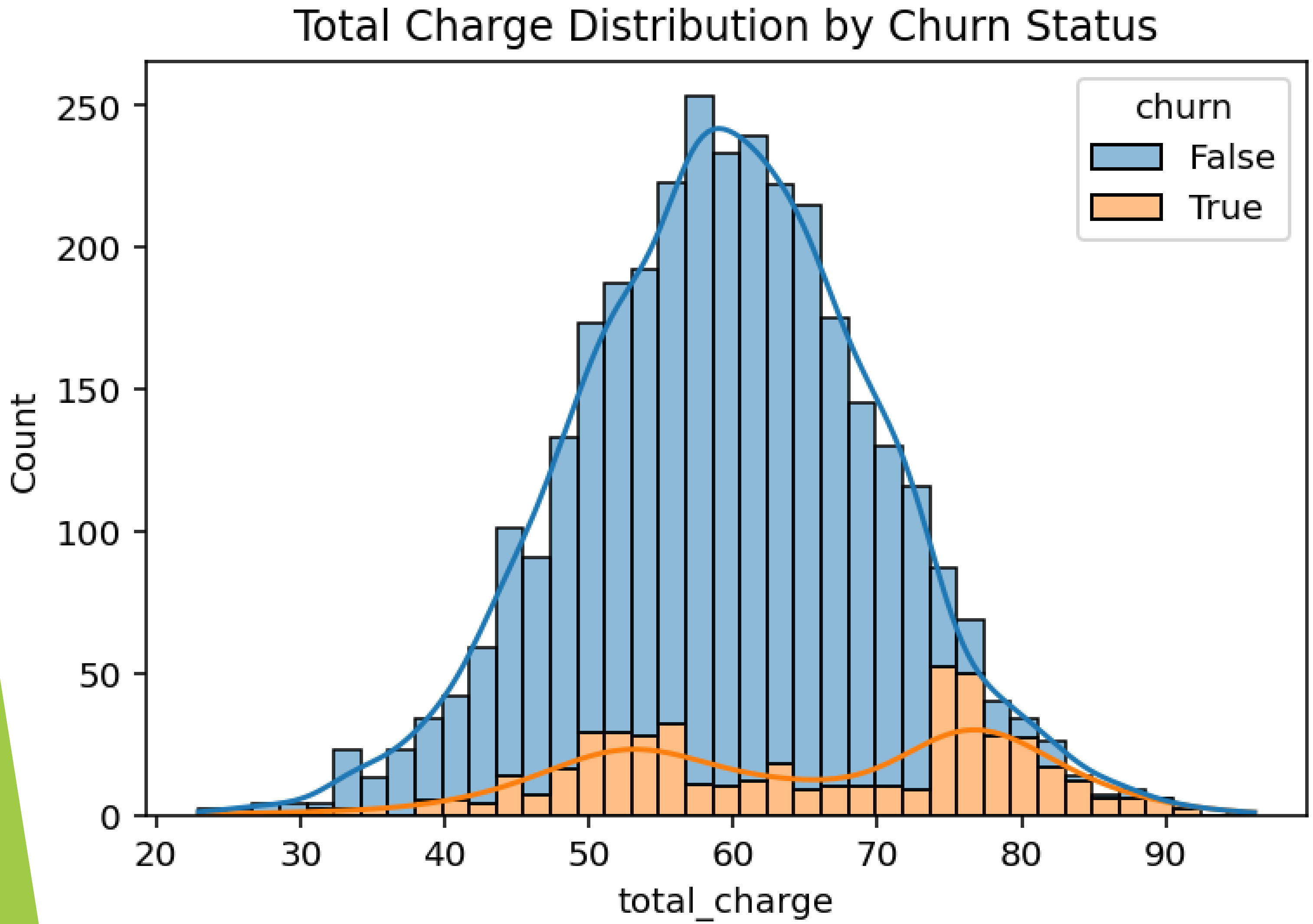
Total Calls Distribution by Churn Status



Non-Churned Users:
The distribution of total calls for users who did not churn (churn = False) is symmetric and follows a normal distribution centered around 300 calls.
The spread of calls is quite wide, ranging from about 200 to 400 calls.

Churned Users:
For users who churned (churn = True), the distribution is also symmetric but with a lower center around 250 calls.
The distribution is narrower, suggesting that churned users tend to have fewer total calls compared to non-churned users.

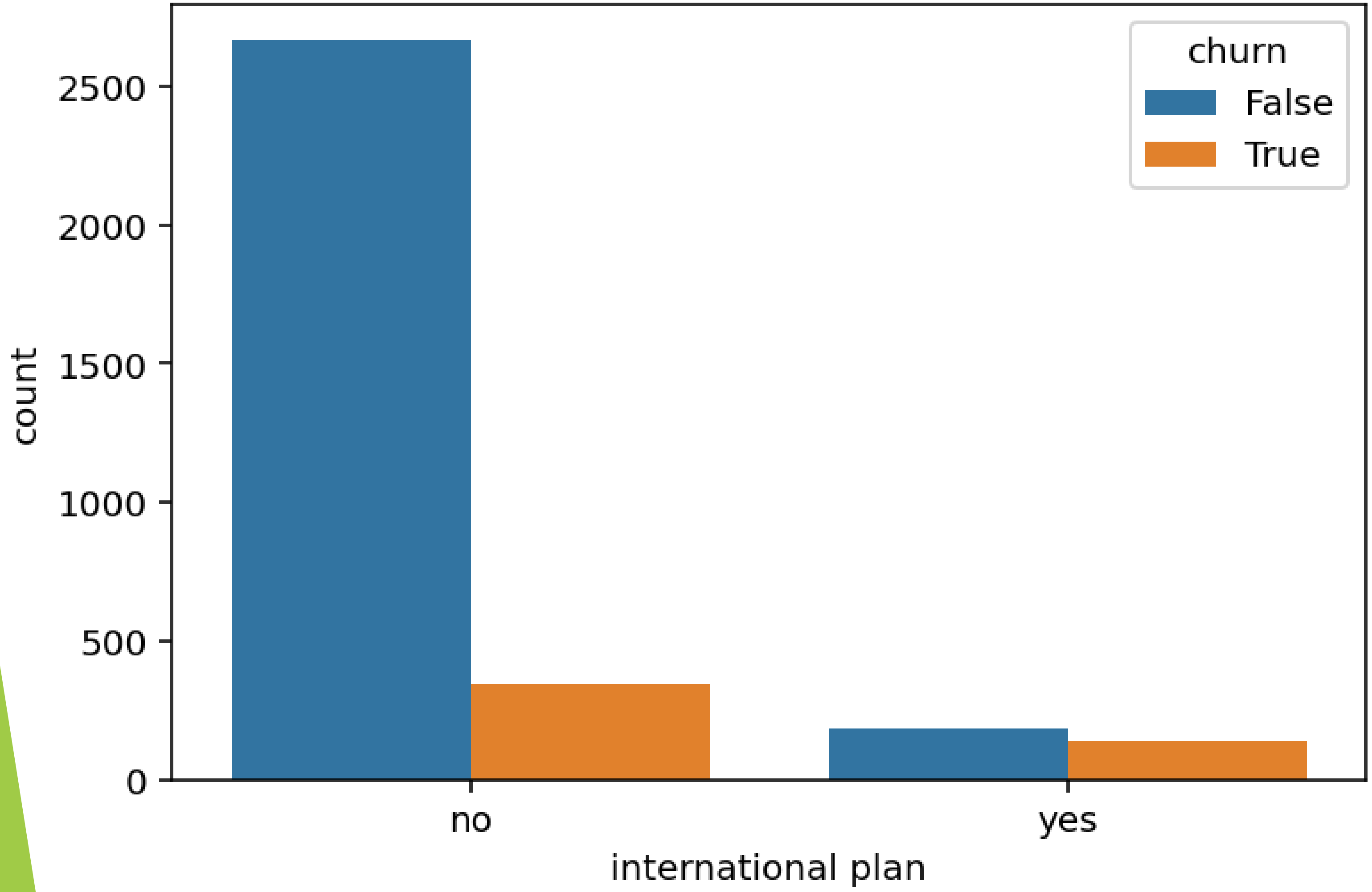
Total Charge Distribution by Churn Status



Non-Churned Users:
The total charge for users who did not churn follows a fairly normal distribution, centered around 60. The spread of charges is moderate, ranging from about 30 to 90.

Churned Users:
The distribution of total charges for churned users is bimodal, indicating two distinct groups within the churned users.
One group has a lower charge, and another has a higher charge. This suggests variability in the billing amounts among churned users.

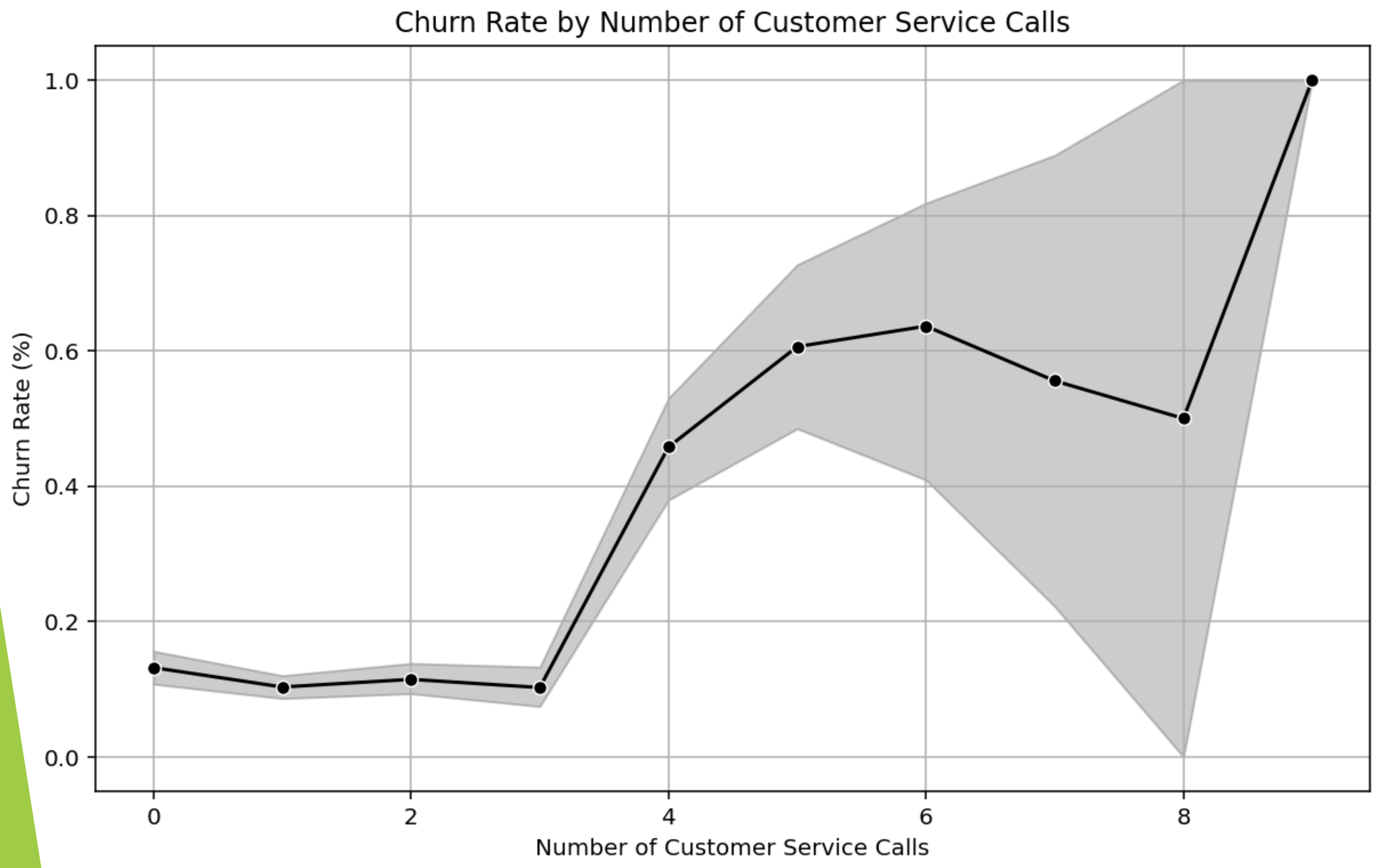
International plan vs churn



Conclusion:

Customers without an international plan are much more likely to stay (not churn) compared to those with an international plan. Additionally, the proportion of customers who churn is relatively higher among those with an international plan compared to those without one.

Churn vs Customer service calls



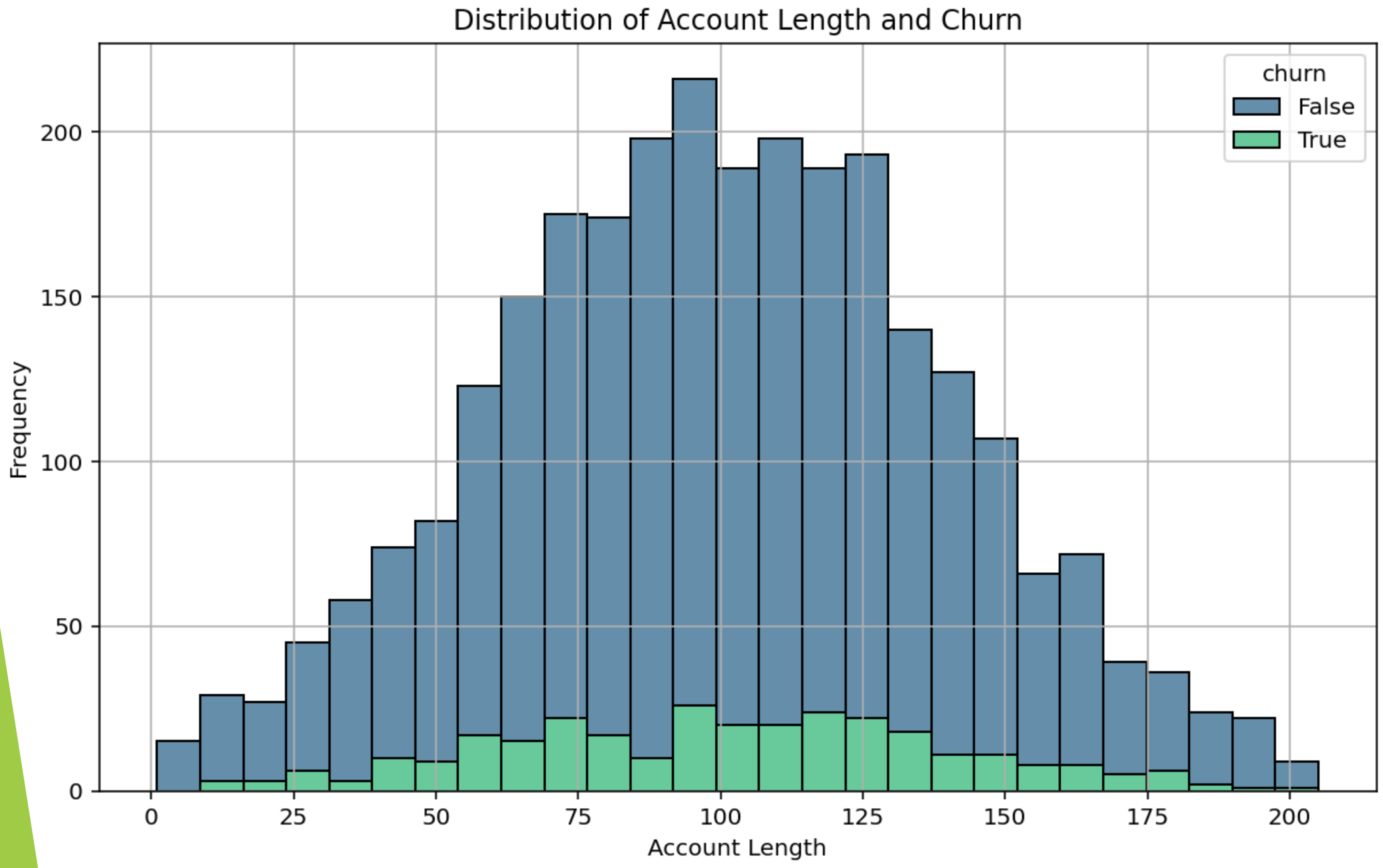
Conclusion:

Low Customer Service Calls (0-3): Customers making fewer customer service calls have a relatively low churn rate, suggesting that they are generally more satisfied or have fewer issues that require frequent support interactions.

Moderate Calls (4-6): A noticeable increase in churn rate occurs when customers make between 4 to 6 calls, indicating potential dissatisfaction or unresolved issues that might lead them to consider leaving.

High Calls (9): A 100% churn rate at 9 calls strongly suggests that customers who need to make many service calls are likely highly dissatisfied, facing significant unresolved issues, or have exhausted all avenues for resolving their concerns through customer service.

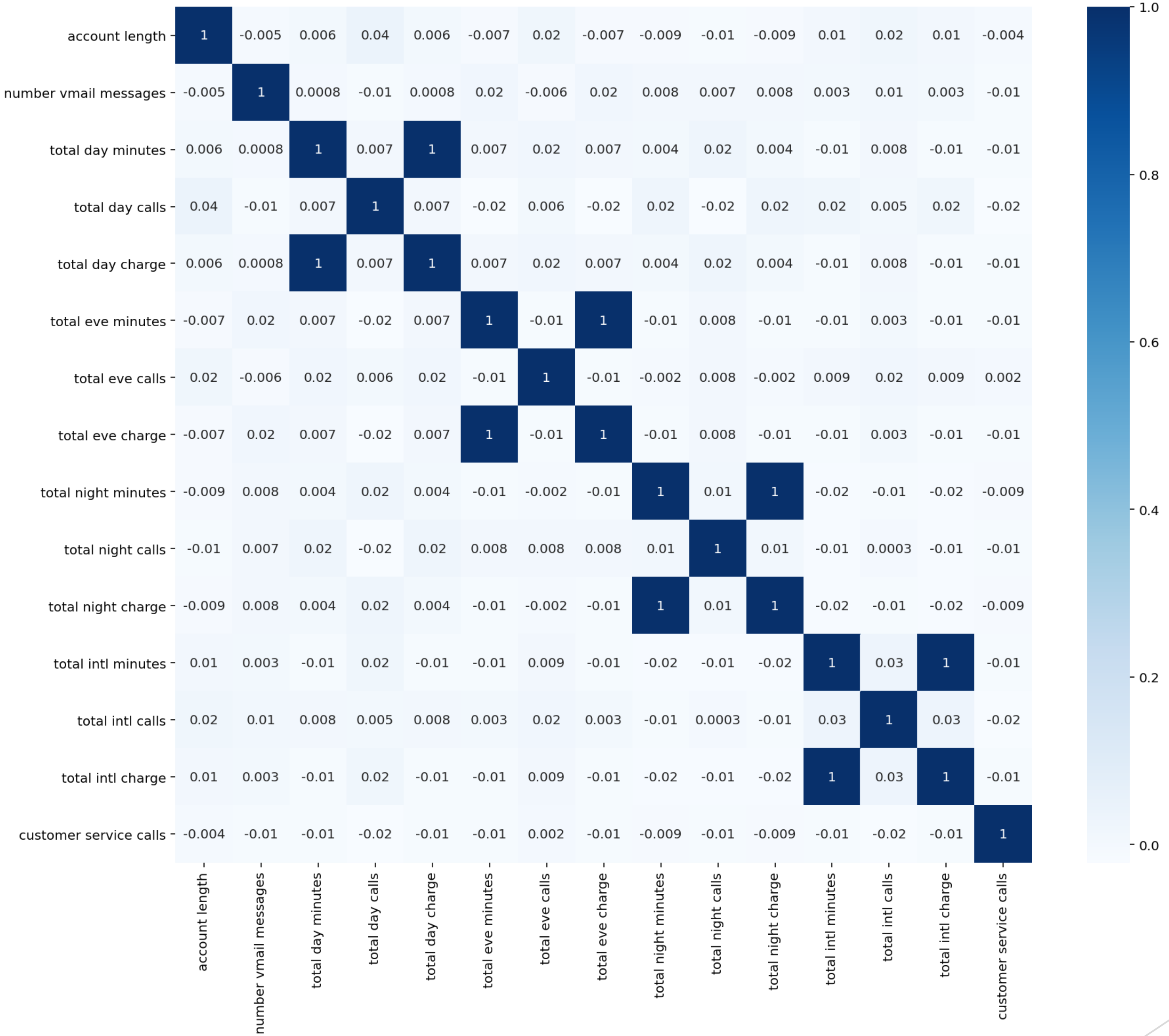
Distribution of Account Length and Churn



Conclusion:

The histogram shows that Account Length has a normal distribution, centered around 100 days, but it doesn't strongly differentiate between customers who churn and those who don't. The proportion of churned customers remains relatively consistent across all account lengths, suggesting that Account Length alone is not a significant predictor of churn.

Feature Correlations with Churn

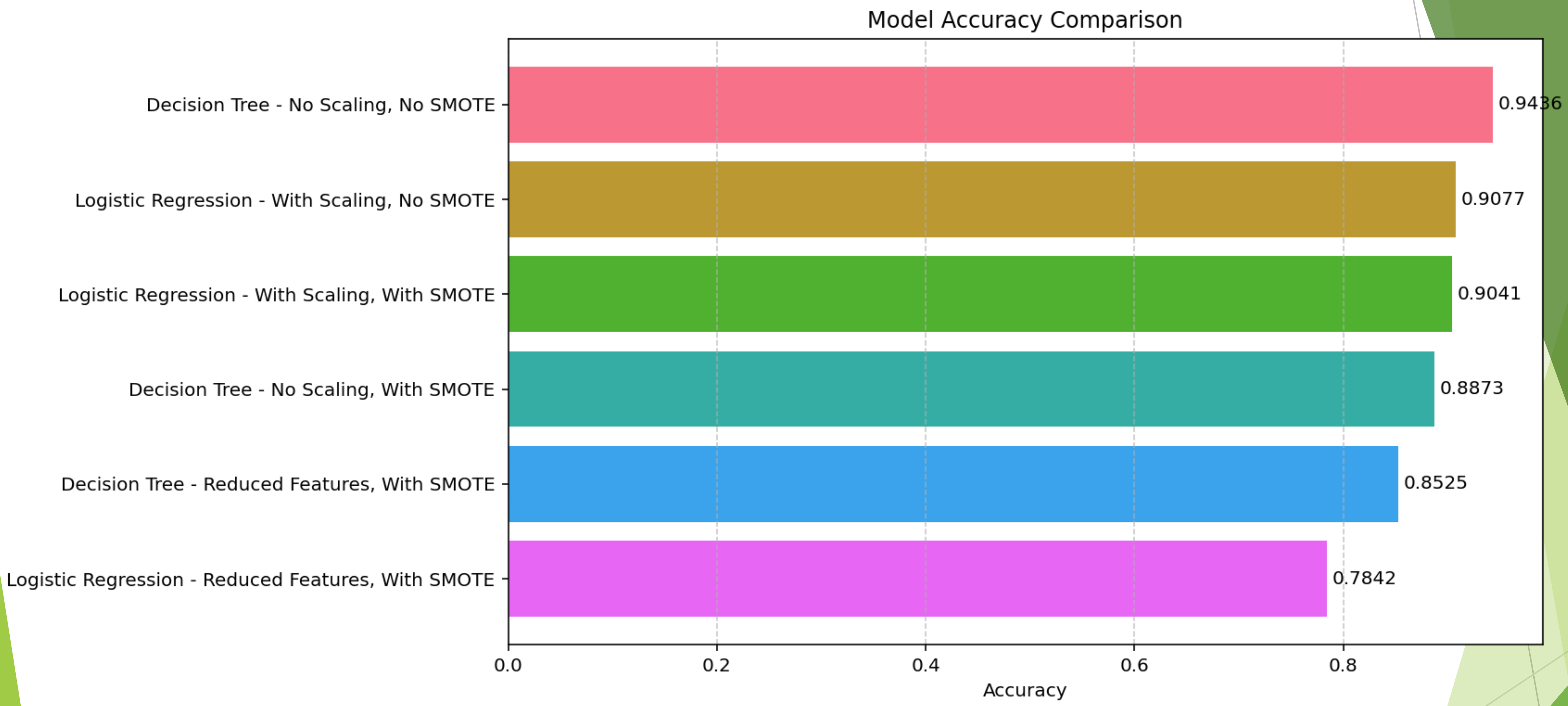


Positive Correlations: Features like customer service calls, total day minutes, and total day charge show a positive correlation with churn, indicating that higher usage in these areas is associated with a higher likelihood of customer churn.

Negative Correlations: Features such as number vmail messages and total intl calls show a negative correlation with churn, suggesting that higher usage in these areas might reduce the likelihood of churn.

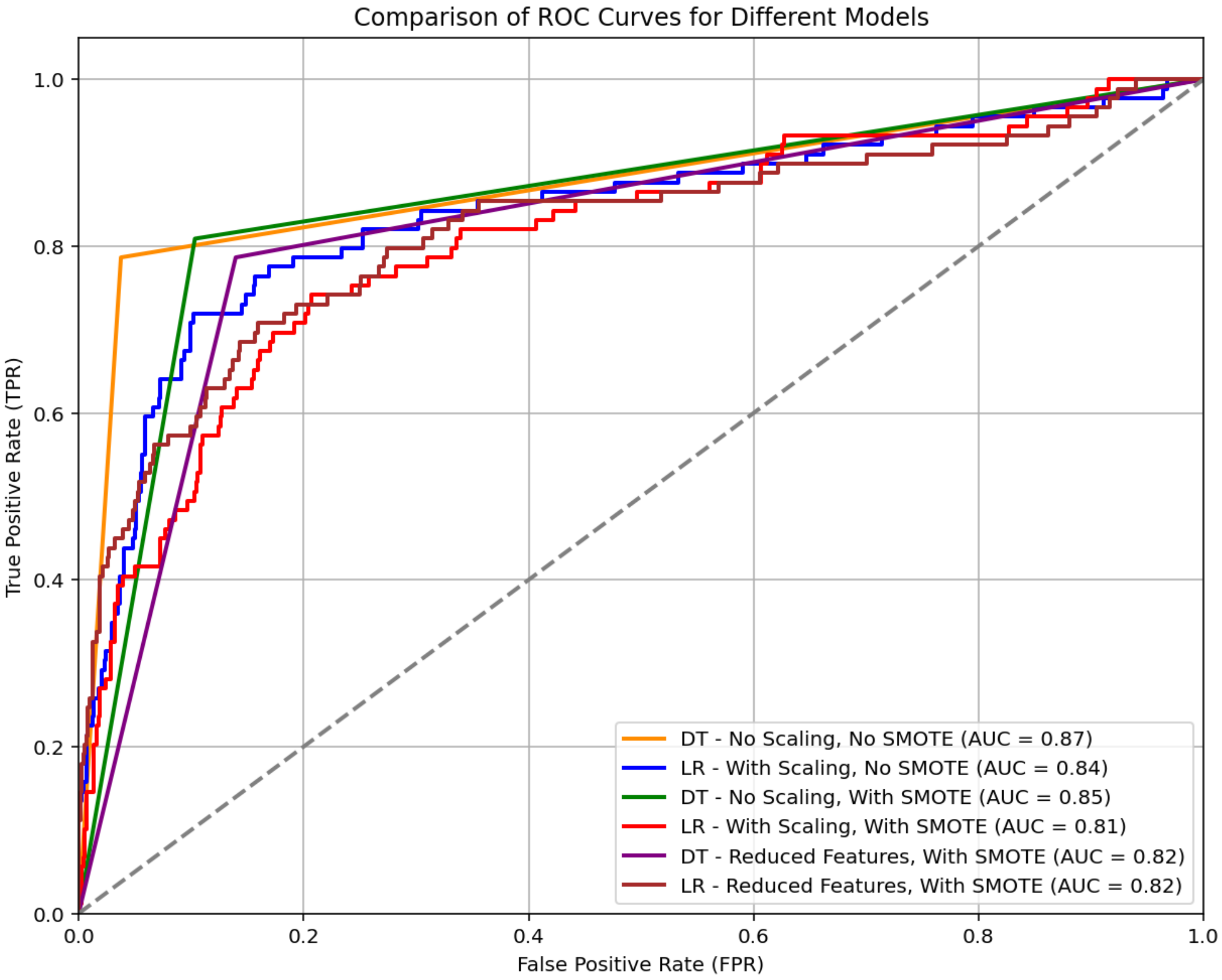
Weak Correlations: Several features like total night calls and account length have little to no correlation with churn, indicating minimal influence on the likelihood of a customer churning.

Model accuracies comparison



Highest Accuracy: The Decision Tree model without scaling and SMOTE achieved the highest accuracy of 94.36%.
Lowest Accuracy: The Logistic Regression model with reduced features and SMOTE had the lowest accuracy at 78.42%.

Comparison of ROC Curves for Different Models



Models with both scaling and SMOTE generally show better ROC-AUC performance, indicating better ability to differentiate between classes.

Logistic regression models tend to benefit more from scaling, while decision trees may show significant improvements with SMOTE due to their nature in handling imbalanced datasets.

Recommendations

1. **Prioritize High-Recall Models:**

Given that the Decision Tree model with SMOTE exhibits the highest recall (80.90%), I recommend using this model as the primary tool for identifying customers at risk of churn. The high recall ensures that we are identifying a substantial portion of potential churners, which is crucial for effective retention strategies.

2. Balance Precision and Recall:

Although this model has lower precision, the trade-off between precision and recall should be considered. While the cost of false positives (i.e., offering incentives to customers who might not churn) is manageable, the cost of false negatives (i.e., missing customers who actually churn) could be more detrimental. The priority should be on minimizing churn rather than solely focusing on precision.

Recommendations

- 3. Leverage SMOTE for Better Performance:
- The application of SMOTE has proven effective in enhancing the model’s performance by addressing class imbalance. I recommend continuing to use SMOTE in future iterations of the model to maintain and possibly improve its ability to detect churn.

- 4 Maintain Comprehensive Feature Sets:
- The Decision Tree model that used all features (without reduction) performed better than models with reduced features. I suggest keeping the full set of features for training the model, as each feature contributes valuable information for predicting churn.

Recommendations

5. Implement Targeted Retention Strategies:

With the selected model's insights, we should develop and implement targeted retention strategies. These could include personalized offers, special promotions, or customer engagement programs aimed at high-risk customers identified by the model.

Next Steps

1. **Deploy the Model:**

The next steps would involve proceeding with deploying the Decision Tree model with SMOTE into our production environment. This will involve integrating the model with our current customer management systems to enable real-time prediction and flagging of high-risk customers.

2. **Design and Test Retention Strategies:**

Collaborating with marketing and customer service teams, I will develop and test various retention strategies based on the model's predictions. The focus will be on creating personalized engagement plans and offers tailored to the needs and behaviors of high-risk customers.

Shortcomings to Address:

Effectiveness of Strategies: Evaluate the effectiveness of retention strategies and refine them based on actual customer responses and feedback. Personalization Challenges: Ensure that engagement plans are genuinely personalized and not perceived as generic or irrelevant by customers.

Next Steps

- 3. **Monitor and Evaluate Performance:**
- I will closely monitor the model's performance post-deployment, including metrics such as accuracy, recall, and precision. Regular evaluations will be conducted to ensure the model's effectiveness in detecting churn and to make necessary adjustments to strategies.
- **Shortcomings to Address:**
- **False Positives/Negatives:**
- Address issues with false positives or negatives that might impact the overall effectiveness of churn detection.

Next Steps

4. **Explore and Implement Advanced Techniques:**

Exploring and implementing advanced machine learning techniques to further enhance the approach. This may include evaluating ensemble other classification machine learning models for improved performance and deeper insights into customer churn.

Shortcomings to Address:

Complexity vs. Usability: Balance the complexity of advanced techniques with their practical usability and interpretability.

Next Steps

5. **Refine and Enhance the Model:**

Based on performance feedback and retention strategy outcomes, I will iteratively refine the model. This may involve tuning hyperparameters, experimenting with different preprocessing methods, or incorporating additional features to boost predictive accuracy.

Shortcomings to Address:

Model Overfitting/Underfitting:

Continuously evaluate and adjust the model to avoid overfitting or underfitting issues.

Feature Relevance: Assess and include relevant features that might have been previously overlooked or underutilized.

Thank You!
Q&A

- Open floor for questions

GEOFFREY MWANGI

Linkedin link.