**PREDICTING CUSTOMER CHURN FOR SYRIATEL**

# **Overview & Business Understanding**

**What is Customer Churn?**

Customer churn refers to the loss of customers or subscribers to a business over a given period. It is a significant concern for any business that relies on recurring subscriptions. Syriatel is a telecommunication provider. In the telecommunications industry, churn occurs when customers switch to competitors or discontinue their service altogether. Therefore,for Syriatel, understanding and predicting churn is crucial to reducing revenue loss and improving customer retention, especially in a highly competitive market where there are other telecom operators such as Verizon Communication, AT&T, T-Mobile US, among others for mobile and internet services. Retaining existing customers is often more profitable than acquiring new ones for telecom providers. According to a 2019 study by Havard Business Review, retaining customers is 5 to 10 times cheaper than acquiring new ones. A strong customer service system and loyalty rewards are often more cost effective. In comparison to acquiring new customers in expensive campaigns for new plans, customer sign-up bonuses and the logistics of onboarding new users. To back up the claim that customer retention is cheaper is a 2020 McKinsey report which estimated that reducing churn by 1% could increase the average Lifetime Value(LTV) of telecom customers by as much as 3 to 5%. This significantly impacts profitability.

How is Churn Typically Measured?

Churn is often measured as the percentage of customers who leave a service provider within a certain timeframe, for example, monthly or annually. It can be represented as:

Challenges Involved in Customer Churn Prediction

1. **Data Quality and Availability**

* **Incomplete or Missing Data**: Inconsistent or incomplete customer data can affect model accuracy. For example, missing customer demographics, usage patterns, or service details can reduce the reliability of predictions.
* **Data Inconsistencies**: Errors or mismatches in customer records, billing data, or service usage information can lead to incorrect analysis.
* **Data Granularity Issues**: The data may not be detailed enough (e.g., monthly instead of weekly usage) or might be too aggregated, which can mask important trends or insights.

1. **Complex Customer Behavior**

* **Multiple Influencing Factors**: Customer churn is driven by a combination of factors such as service quality, customer support, pricing, and external conditions (e.g., economic or political events), making it complex to isolate specific predictors.
* **Non-linear Relationships**: Customer behavior doesn’t always follow simple, linear patterns. Small changes in one factor might trigger a significant impact on churn, which is harder to model accurately.
* **Unpredictable External Factors**: Churn may also be influenced by factors outside the company’s control, such as changes in the market, competitor offers, or regional political and economic instability.

1. **Class Imbalance**

* **Disproportionate Distribution**: Typically, fewer customers churn compared to those who stay, leading to imbalanced datasets. This can cause models to underperform by over-predicting the "staying" class and failing to identify the "churning" class accurately.
* **Model Bias**: Without proper techniques to balance the data, predictive models may favor the majority class, reducing the overall effectiveness in predicting churn.

1. **Customer Segmentation Challenges**

* **Varied Churn Drivers by Segment**: Different customer segments may exhibit distinct churn patterns. For example, price-sensitive customers may churn for financial reasons, while high-value customers may leave due to poor service quality or network issues.
* **Dynamic Profiles**: Customer profiles are not static. Customers may change their usage patterns, service types, or behavior over time, complicating the model’s ability to predict future churn accurately.

1. **Model Complexity and Overfitting**

* **Overfitting**: Predictive models, particularly complex machine learning models, may overfit to the training data, capturing noise rather than genuine patterns. This leads to poor generalization and less reliable predictions when applied to unseen data.
* **Interpretability**: Complex models (e.g., deep learning) can achieve high accuracy but are often not transparent or interpretable, making it difficult for stakeholders to understand why certain customers are predicted to churn and how to act on the insights.

1. **Retention Strategy Effectiveness**

* **Implementing Effective Retention Measures**: Even with accurate churn predictions, the company may lack effective strategies or resources to act on those predictions (e.g., targeted offers, improved customer service). Without a good retention strategy, predicting churn may not significantly reduce churn rates.
* **Impact Assessment**: It is challenging to evaluate the actual impact of retention measures. For example, how well personalized offers or customer outreach actually reduce churn can be hard to measure and assess.

1. **Economic and Political Instability**

* **External Factors**: In regions with economic instability, political unrest, or unpredictable events, customer behavior may change abruptly, making it difficult to predict churn based on historical data.
* **Market Volatility**: Changes in the broader market or government regulations can influence customer decisions, leading to sudden increases or decreases in churn that are hard to predict.

1. **Competitive Dynamics**

* **Competitive Offers**: Customer churn can often be driven by better offers from competitors (e.g., lower prices or superior service), which is difficult to model accurately without detailed competitor data.
* **Price Sensitivity**: Customers may churn based on pricing changes or special promotions offered by competitors, which can be hard to predict in advance.

1. **Technological and Infrastructure Issues**

* **Network and Service Quality**: Variability in service quality (e.g., network outages, slow internet speeds) can drive churn, but predicting this based on customer complaints or usage patterns can be challenging.
* **System Integration**: Integrating churn prediction models into existing customer relationship management (CRM) and operational systems can be a challenge, especially if data is spread across multiple platforms.

Stakeholders

1. **Senior Management and Executives**

* **Role**: Senior leaders, such as the CEO and department heads, are responsible for overseeing the company's overall strategy and financial performance.
* **Interest**: They are primarily concerned with reducing churn to maintain a steady stream of revenue and ensure long-term business sustainability. Insights from churn predictions help in making informed decisions about resource allocation, pricing strategies, and customer retention efforts.

1. **Marketing Team**

* **Role**: The marketing department is responsible for customer acquisition, engagement, and retention campaigns.
* **Interest**: Marketing teams use churn prediction models to target high-risk customers with personalized offers, promotions, or loyalty programs. They also use insights to develop campaigns that are tailored to prevent churn and retain valuable customers.

1. **Customer Service and Support Teams**

* **Role**: Customer service handles customer interactions, complaints, and inquiries, ensuring overall customer satisfaction.
* **Interest**: Proactive interventions can be made by identifying customers who are likely to churn. Customer service teams can focus on resolving complaints, offering incentives, or improving the service experience for high-risk customers to prevent churn.

1. **Sales Team**

* **Role**: The sales department is responsible for acquiring new customers and managing key accounts.
* **Interest**: Sales teams benefit from churn predictions as they can identify at-risk high-value customers and take steps to retain them. They can also refine their sales strategies by offering personalized services or deals to customers on the verge of churning.

1. **Data Science and Analytics Teams**

* **Role**: The data science team is responsible for collecting, cleaning, and analyzing data, as well as building and maintaining predictive models.
* **Interest**: Data scientists and analysts work directly with churn prediction models. They are focused on ensuring the model's accuracy, robustness, and ability to generate actionable insights. They will also monitor model performance over time and retrain it with new data to improve predictions.

1. **IT and Infrastructure Teams**

* **Role**: The IT department ensures the technical infrastructure is in place to collect, store, and process the necessary data for churn prediction.
* **Interest**: IT plays a critical role in integrating churn prediction models with existing systems, ensuring the smooth flow of data from various customer touch points (e.g., CRM, billing system, network usage logs) into the predictive models. They also ensure data security and compliance with privacy regulations.

1. **Product Development**

* **Role**: Product managers and developers work on improving the telecom products and services offered to customers.
* **Interest**: Insights from churn prediction can guide the product team on what features or services to enhance or develop to address customer pain points and reduce churn. For example, if churn is related to poor service quality or specific pricing plans, the product team can adjust these offerings.

1. **Customer Relationship Management (CRM) Team**

* **Role**: CRM specialists manage the company's CRM software, which stores customer interaction data and helps execute personalized marketing campaigns.
* **Interest**: They would benefit from churn prediction insights to identify at-risk customers and automate personalized retention efforts, such as sending offers, reminders, or follow-ups. They also play a role in integrating churn prediction models with CRM systems for real-time action.

1. **Finance and Revenue Team**

* **Role**: The finance team is responsible for managing the company's financial health, including revenue forecasting, profitability, and cost management.
* **Interest**: The finance team uses churn predictions to anticipate revenue losses and develop strategies to offset the impact of churn. By retaining more customers, they can stabilize revenue and avoid unnecessary operational costs tied to acquiring new customers to replace churned ones.

1. **Regulatory and Compliance Team**

* **Role**: Regulatory bodies ensure that the company adheres to the laws and regulations governing customer data and business operations.
* **Interest**: The compliance team is concerned with ensuring that customer data used in churn prediction models is handled securely and in compliance with data protection laws (such as GDPR or local privacy regulations). They ensure that customer information is anonymized and used ethically.

1. **External Shareholders**

* **Role**: External stakeholders may include investors, regulatory bodies, and even competitors.
* **Interest**: Investors are concerned with the company's growth, profitability, and sustainability, so churn rates are key indicators. Regulators focus on ensuring that data usage adheres to privacy and security standards. Competitors may also analyze churn trends to improve their own strategies.

# **Proposed Solution(Analysis & Modeling)**

The goal of this project is to build an accurate predictive model that identifies customers at risk of churning, enabling SyriaTel to take proactive steps to retain them. This process will involve data analysis, feature engineering, model selection, and evaluation, ensuring that the model is both effective and interpretable for various stakeholders. The model will be trained on historical customer data and relevant features, such as usage patterns, customer demographics, and service plan information, to predict the likelihood of a customer leaving SyriaTel. By identifying at-risk customers, the model will empower the business to implement targeted retention measures and reduce churn.

# **Projected Conclusion**

The predictive churn model for SyriaTel will provide a data-driven approach to identifying customers at risk of leaving, based on historical data and key customer features such as usage patterns, demographics, and service plans. By integrating this model into SyriaTel's systems, the company can take proactive measures, such as targeted offers and personalized customer service, to improve retention, reduce churn, and optimize resource allocation. This will not only enhance customer satisfaction and loyalty but also reduce acquisition costs, as retaining existing customers is more cost-effective than acquiring new ones. Ultimately, the model will help SyriaTel maintain a competitive edge, increase profitability, and continuously adapt to changing customer behaviors.

# **Problem Statement**

SyriaTel faces the challenge of customer churn, which leads to lost revenue and higher acquisition costs, yet lacks a data-driven approach to predict and address this issue. This project aims to develop a predictive model that accurately identifies customers at risk of churning, using historical data and key customer features such as usage patterns, demographics, service plans, and interaction history. By leveraging this model, SyriaTel can proactively reduce churn, improve retention, and optimize marketing and customer service efforts, ultimately boosting profitability, customer loyalty, and resource efficiency.

# **Objectives**

1. **Develop a Customer Churn Prediction Model**: Create a machine learning model that can predict which customers are likely to churn based on historical data, including key customer features such as account length, usage patterns, service plans, and customer service interactions.
2. **Identify Key Features Influencing Churn**: Analyze customer attributes such as international plan, voicemail plan, total minutes and charges across different times (day, evening, night, international), and customer service interactions to identify the most influential factors driving churn.
3. **Understand Churn Distribution**: Analyze how churn is distributed across various customer features like state, area code, and service plans, providing insights into which customer segments are most at risk.
4. **Ensure Data Quality and Consistency**: Identify and address missing or inconsistent data within features like phone numbers, messages, or service charges, ensuring the dataset is clean and reliable for building an accurate predictive model.
5. **Examine Relationships Between Features and Churn**: Investigate how usage patterns (e.g., total minutes, calls, charges) and service plans (e.g., international plan, voicemail plan) correlate with customer churn, helping to determine which variables are most predictive of churn behavior.
6. **Handle Data Imbalances**: Assess the balance between churned and non-churned customers in the dataset and implement techniques such as resampling or class weighting to handle any class imbalance that could affect model accuracy.
7. **Provide Actionable Insights for Retention Strategies**: Use the churn prediction model to uncover patterns in customer behavior and identify at-risk segments, allowing SyriaTel to implement targeted retention measures, such as personalized offers, service improvements, or proactive customer support.

# **Metrics of Success**

The success of the customer churn prediction model for SyriaTel will be measured using the following metrics:

1. **Precision**: To measure the proportion of customers predicted to churn who actually do churn. In the context of churn prediction, it helps assess how reliable the model is when it identifies a customer as likely to churn. A higher precision means fewer false positives, which ensures that resources are focused on customers who are truly at risk of leaving.
2. **Recall**: To measure the proportion of actual churned customers that are correctly identified by the model. It tells us how well the model captures customers who are at risk of leaving. A higher recall means that the model successfully identifies more of the customers who would churn, which is crucial for reducing customer attrition.
3. **F1-Score**: It is the harmonic mean of precision and recall, offering a balance between the two metrics. It’s especially valuable when there is a need to minimize both false positives (wrongly predicting churn) and false negatives (failing to predict churn). Since both missed churn predictions and false alarms can be costly, F1-score helps optimize the trade-off.
4. **AUC-ROC**: It evaluates the model's ability to distinguish between churned and non-churned customers across different thresholds. The AUC value ranges from 0 to 1, where a value closer to 1 indicates a better performing model. It helps in understanding the overall performance of the model in terms of its discrimination ability, particularly in cases where the dataset may be imbalanced (more non-churning than churning customers).
5. **Confusion Matrix**: It provides a detailed breakdown of the model's predictions by showing the true positives, false positives, true negatives, and false negatives. This is valuable for understanding the types of errors the model is making and provides insights into how the model is performing with respect to churn prediction. It’s particularly useful for visually assessing how well the model handles churn vs. non-churn predictions.

# **Data Understanding**

The SyriaTel Customer Churn dataset from Kaggle provides data on customer behavior and indicates whether they canceled their subscription with the telecom company. It includes 3,333 rows and 21 columns.

The columns in the SyriaTel Customer Churn dataset are:

1. **State** - The state the customer resides in
2. **Account Length** - The number of days the customer has had an account
3. **Area Code** - The area code of the customer
4. **Phone Number** - The phone number that uniquely identifies each customer
5. **International Plan** - Indicates if the customer has an international plan
6. **Voice Mail Plan** - Indicates if the customer has a voice mail plan
7. **Number Vmail Messages** - The number of voicemail messages the customer has sent
8. **Total Day Minutes** - Total number of minutes the customer has used during the day
9. **Total Day Calls** - Total number of calls the customer has made during the day
10. **Total Day Charge** - Total amount of money the customer was charged for day-time calls
11. **Total Eve Minutes** - Total number of minutes the customer has used during the evening
12. **Total Eve Calls** - Total number of calls the customer has made during the evening
13. **Total Eve Charge** - Total amount of money the customer was charged for evening calls
14. **Total Night Minutes** - Total number of minutes the customer has used during the night
15. **Total Night Calls** - Total number of calls the customer has made during the night
16. **Total Night Charge** - Total amount of money the customer was charged for night-time calls
17. **Total Intl Minutes** - Total number of minutes the customer has used for international calls
18. **Total Intl Calls** - Total number of international calls made by the customer
19. **Total Intl Charge** - Total amount of money the customer was charged for international calls
20. **Customer Service Calls** - Number of times the customer has contacted customer service
21. **Churn** - Target variable indicating whether the customer churned

# **Data Preparation**

In preparation for analysis and modeling, several steps were undertaken to clean, enhance, and transform the dataset to ensure its suitability for predictive modeling:

1. **Column Name Standardization:**
   * Spaces in column names were replaced with underscores to ensure consistency and ease of access during coding and analysis.
2. **Handling Irrelevant Columns:**
   * The phone\_number column was simplified to retain only the area\_code, as the remaining digits provided no meaningful contribution to customer behavior or churn prediction.
   * The column phone\_number was renamed to customer\_id to reflect its role as a unique customer identifier.
3. **Target Variable Transformation:**
   * The churn column was converted from Boolean values ("Yes" and "No") to numeric format, with 1 representing churn and 0 representing no churn. This change simplified model training and evaluation.
4. **Data Cleaning:**
   * No missing values, null values, or duplicate rows were identified in the dataset, ensuring its completeness and consistency for analysis.
   * Outliers were retained in the dataset as they represented valid but rare customer behaviors, such as unusually high service usage or charges. Given that decision tree models are less sensitive to outliers, this decision preserved valuable information.
5. **Feature Engineering:**
   * **New Features Created:**
     + **Total Domestic Metrics:** Summed day, evening, and night metrics (minutes, calls, charges) to create comprehensive total\_domestic\_minutes, total\_domestic\_calls, and total\_domestic\_charge columns.
     + **Charge Per Minute:** Calculated by dividing the total charges (domestic + international) by total minutes, providing a cost-efficiency metric for each customer.
     + **International Call Proportion:** Derived by dividing international calls by the total calls (domestic + international), highlighting the proportion of international usage.
     + **Customer Call Satisfaction Ratio:** Measured as the ratio of customer service calls to total calls, offering insights into customer dissatisfaction levels.
   * These features were designed to capture nuanced customer behaviors that could influence churn.
6. **Regional Categorization:**
   * States were grouped into broader U.S. regions (e.g., Northeast, Midwest, South, and West) using U.S. election regions. This grouping helps capture regional behavior trends, reduce complexity, and improve model interpretability, especially for states with fewer observations.
7. **Outlier Handling:**
   * Outliers were retained to account for rare but valid customer behaviors, such as extreme usage patterns or high service charges, ensuring the dataset reflects real-world diversity.

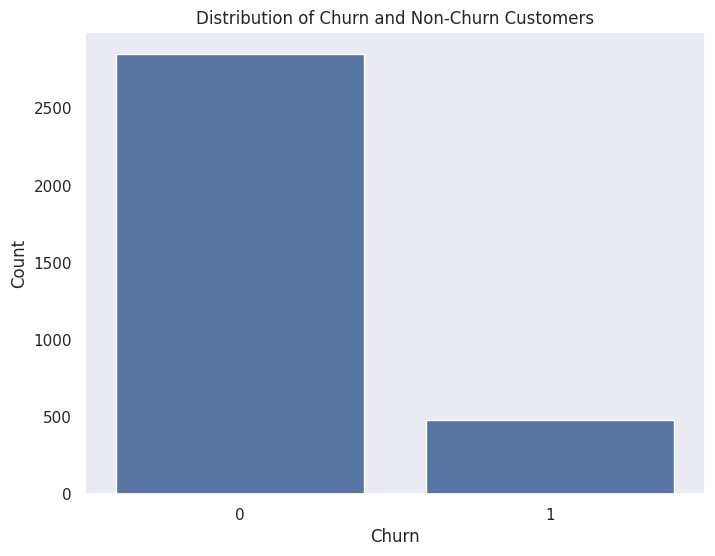
By implementing these data preparation steps, the dataset was transformed into a robust and enriched version, ready for effective exploration and predictive modeling. This comprehensive approach ensures the model captures meaningful insights into customer behavior while maintaining interpretability and accuracy.

# **Data Analysis**

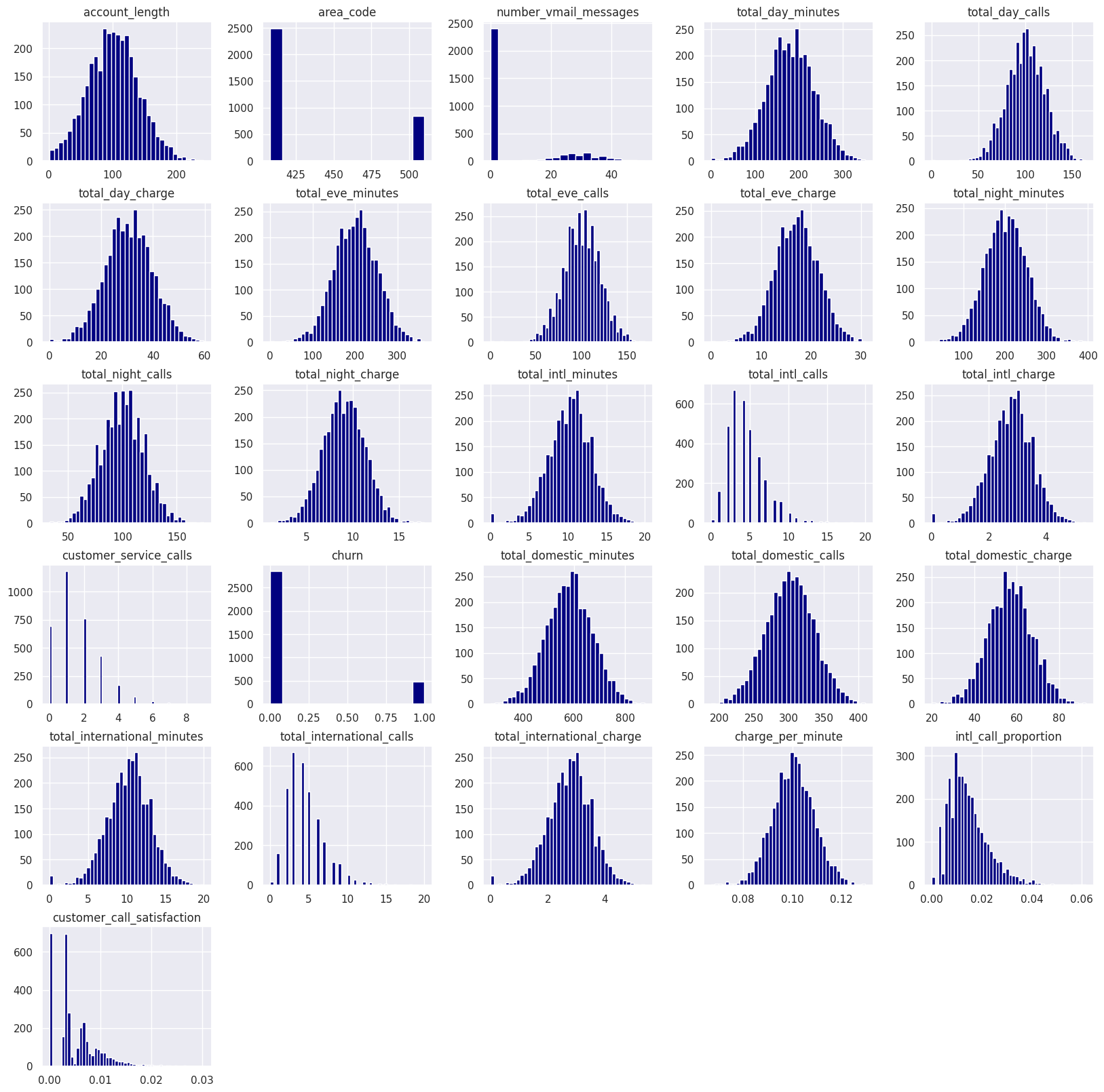
#### **Univariate Analysis**

**Objective:** To examine individual features and their distributions to identify patterns and trends related to churn.

**Churn Distribution:** The dataset exhibits a significant imbalance between churn and non-churn customers. The count plot reveals that a majority of customers belong to the "non-churn" category, which could potentially bias predictive models if not addressed. This imbalance underlines the importance of using techniques like resampling or class weighting to ensure fair model performance.

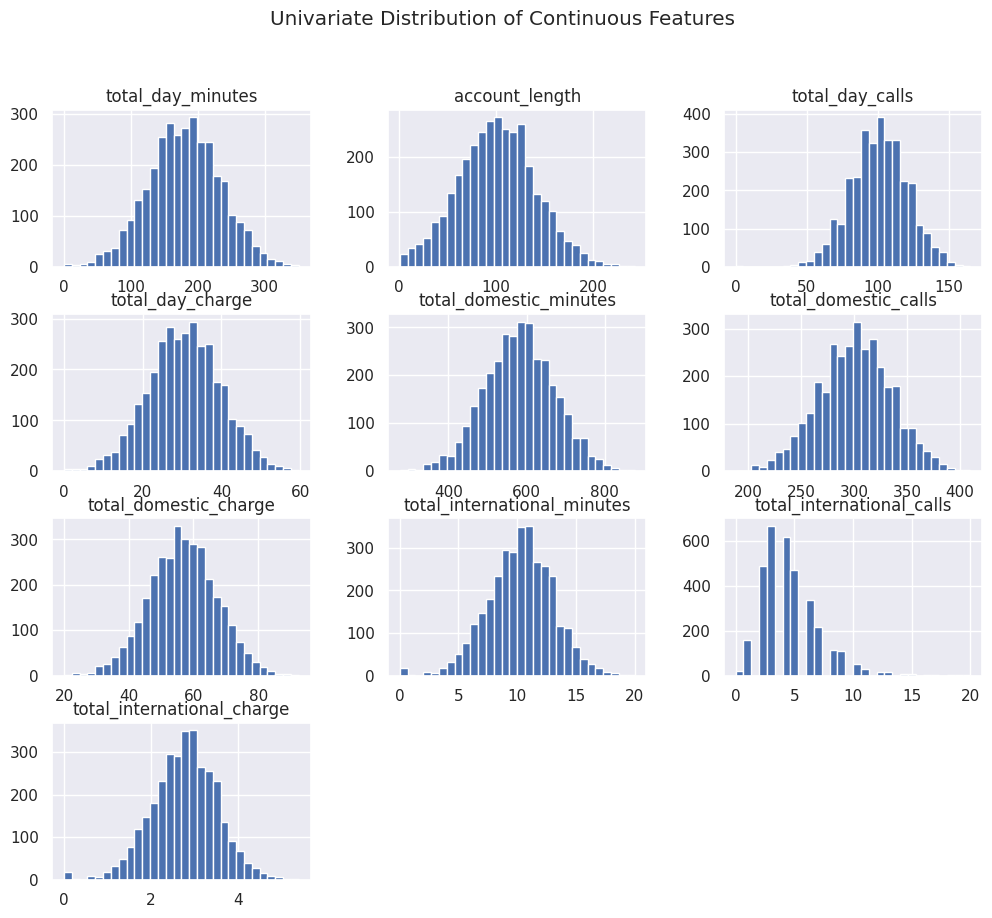


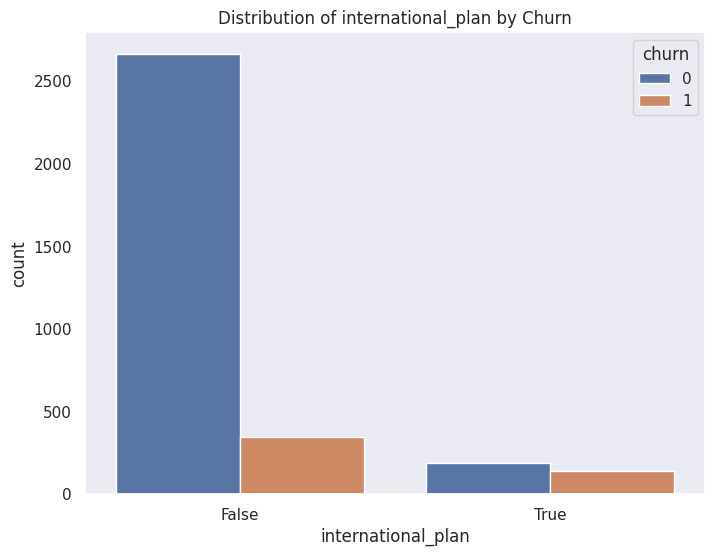
**Distribution of Continuous Features:** The histograms of continuous variables such as total\_day\_minutes, account\_length, and total\_day\_charge indicate relatively normal distributions, with total\_day\_minutes and total\_day\_charge showing positive skews due to a few high-value outliers. Features like total\_domestic\_calls and total\_international\_minutes reveal distinct usage patterns, with international metrics generally lower in volume compared to domestic metrics. These patterns suggest that different customer segments exhibit diverse usage behaviors.

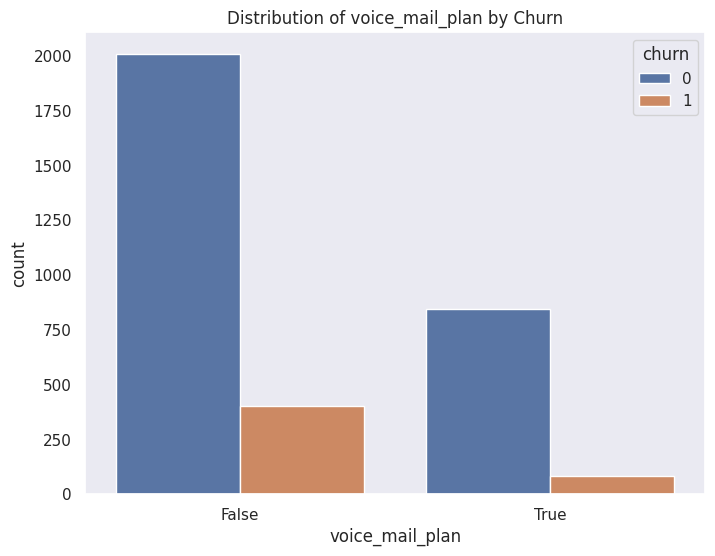


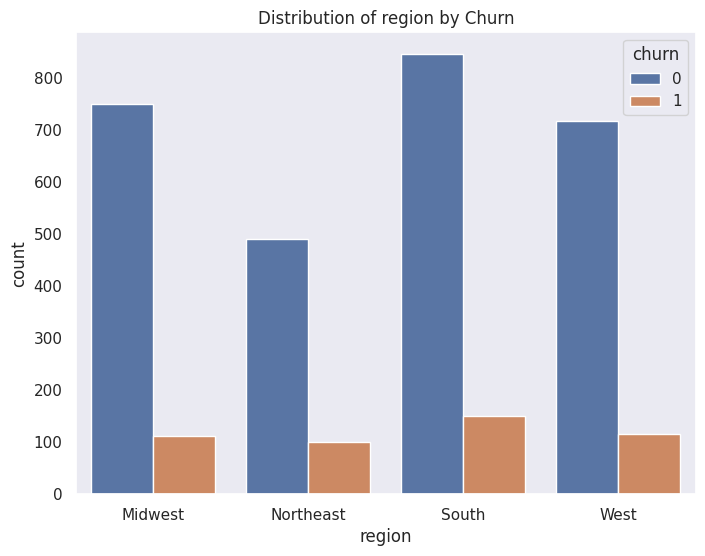
**Categorical Features Analysis:** For categorical variables like international\_plan, voice\_mail\_plan, region, and area\_code, the count plots with churn as a hue highlight significant variations:

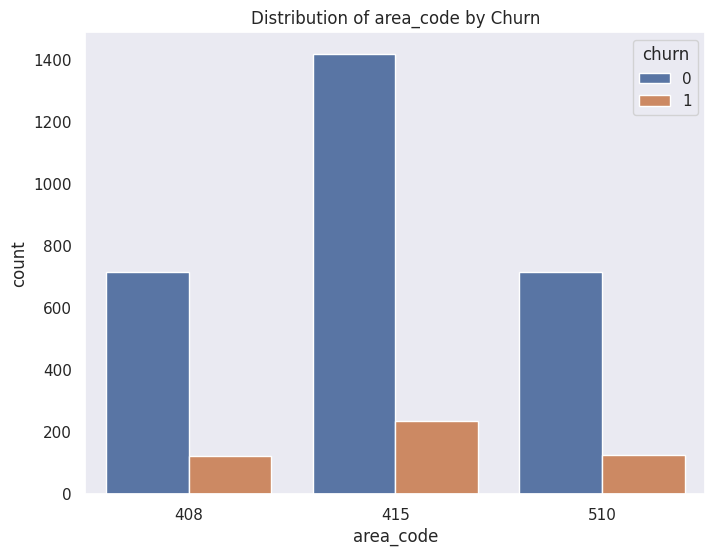
* Customers with an international plan churn more frequently compared to those without.
* Customers with a voicemail plan churn less often, potentially indicating satisfaction with service enhancements.
* Regional differences suggest that customers from certain areas (e.g., Northeast and Midwest) have higher churn rates.
* Area codes reveal no significant churn differences, but their inclusion provides useful customer segmentation.







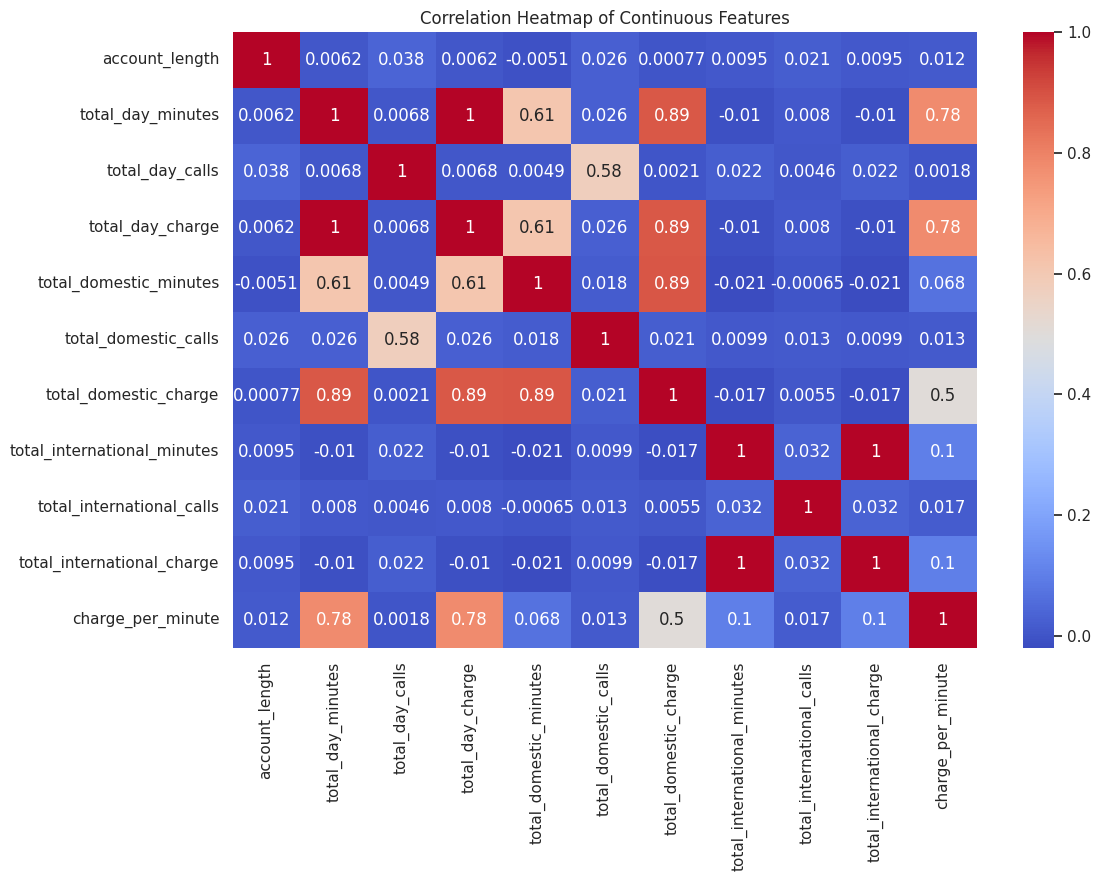




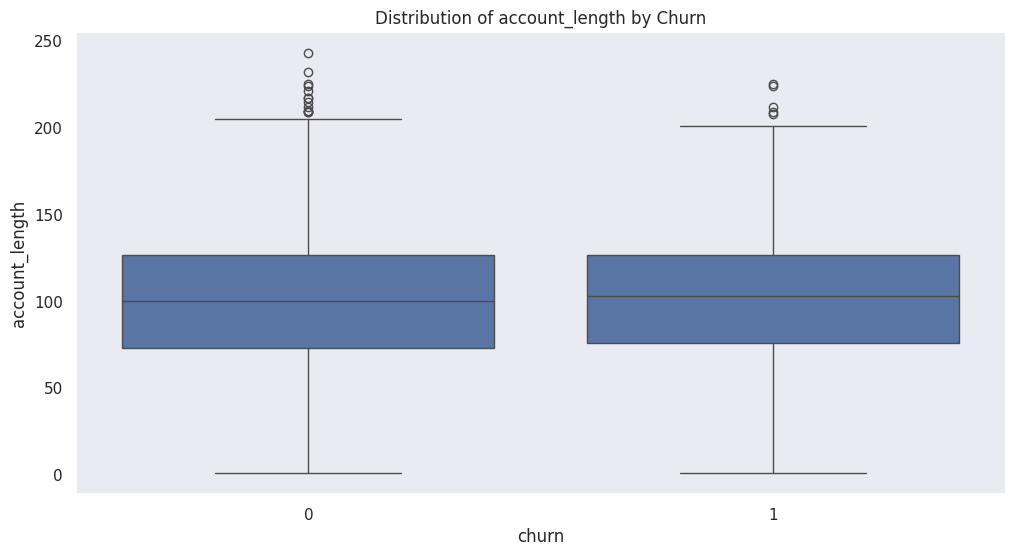
### **Bivariate Analysis**

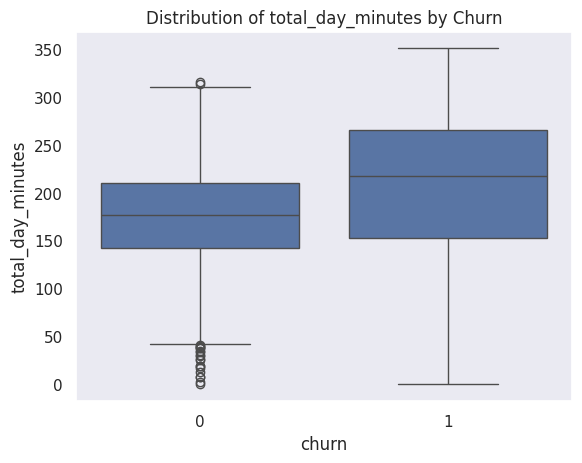
**Objective:** To explore relationships between pairs of features and churn to identify potential correlations and patterns.

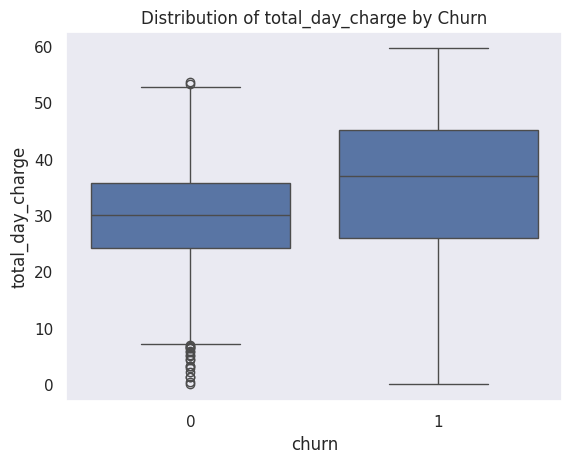
1. **Correlation Heatmap**:  
   Correlation analysis indicates strong relationships between related features such as total\_day\_minutes and total\_day\_charge (r = 1.0). Features like charge\_per\_minute and total\_domestic\_charge also exhibit high correlations, reinforcing their significance for modeling.

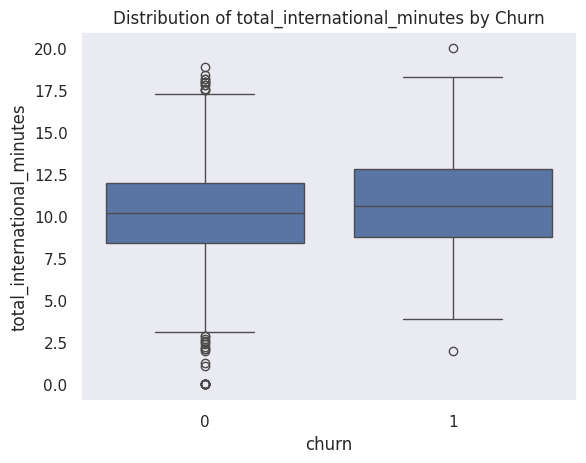


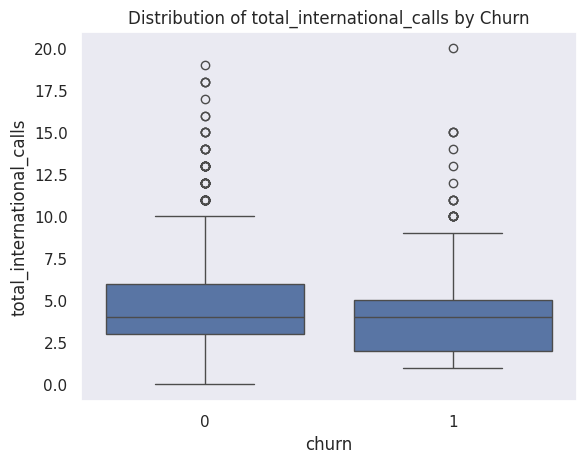
1. **Boxplots by Churn**:
   * **Account Length**: Churners tend to have slightly shorter account lengths, indicating newer customers might be more prone to churn.
   * **Total Day Metrics**: Churners have higher total\_day\_minutes and total\_day\_charge compared to non-churners, suggesting that heavy daytime usage may correlate with dissatisfaction or higher churn rates.
   * **International Metrics**: There is no significant difference in international metrics between churners and non-churners, but churners tend to make slightly fewer international calls and incur lower international charges.

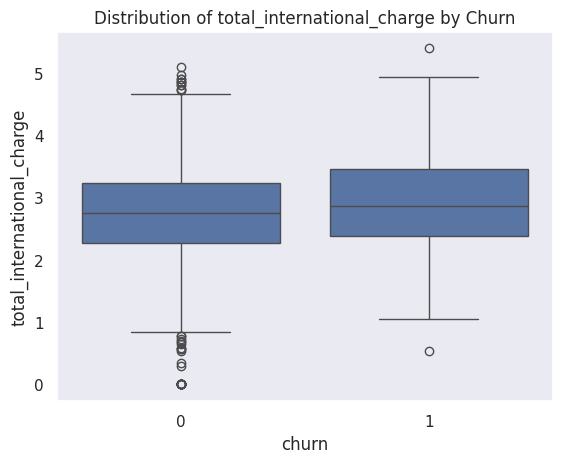












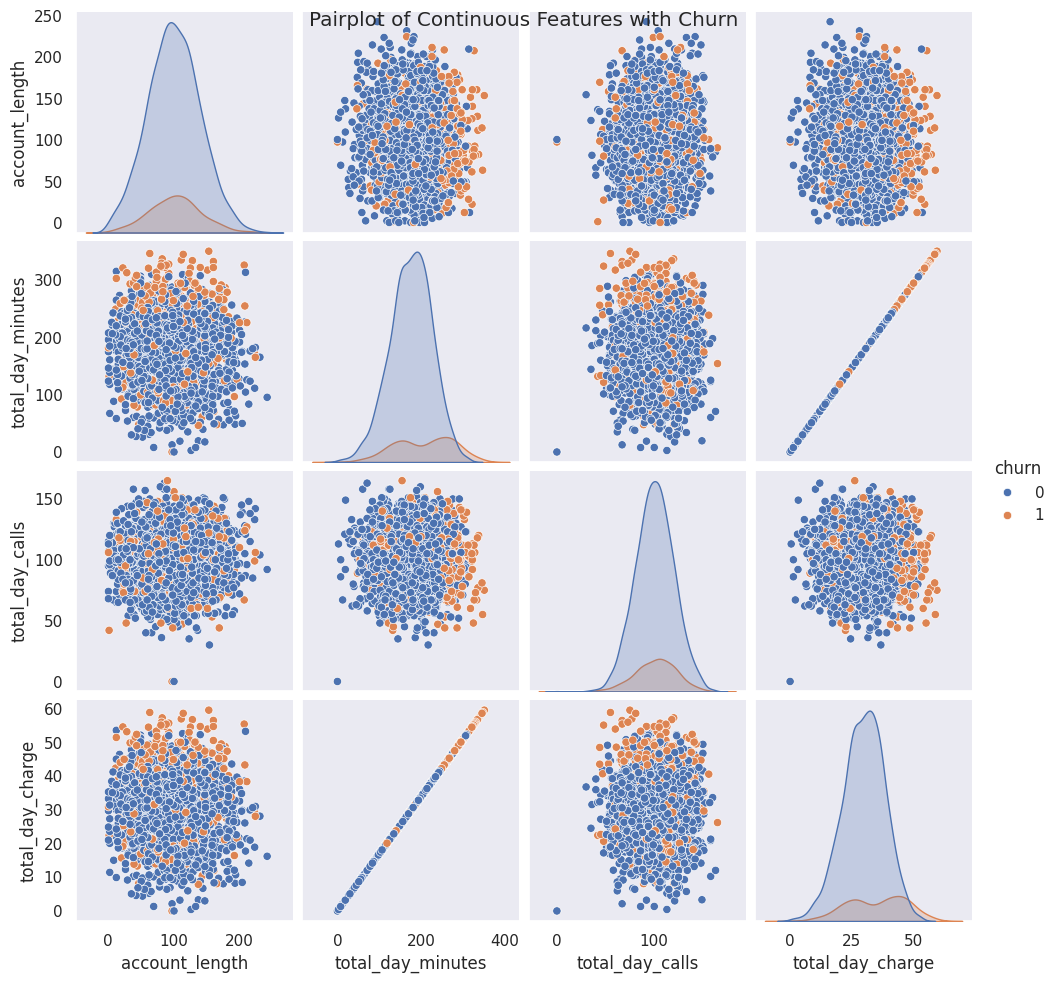
### **Multivariate Analysis**

**Objective:** To explore the relationships between multiple features and churn to uncover key interactions and correlations that may inform model building.

#### **Pairplot Analysis**

A pairplot of continuous features such as account\_length, total\_day\_minutes, total\_day\_calls, and total\_day\_charge with churn reveals several insights:

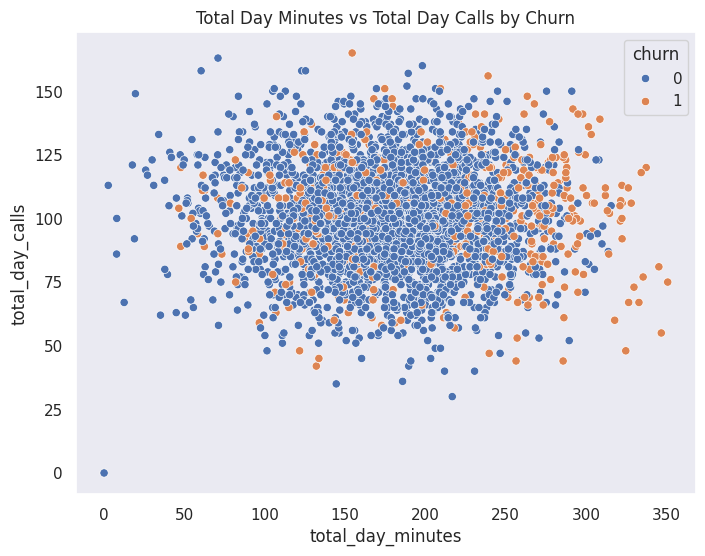
* Churned customers tend to have higher values for total\_day\_minutes and total\_day\_charge, indicating that heavy daytime usage is a potential churn driver.
* Total\_day\_calls shows no strong distinction between churned and non-churned customers, suggesting that call frequency during the day is less indicative of churn.
* There is a strong linear relationship between total\_day\_minutes and total\_day\_charge, confirming the expected proportionality between minutes and charges.



#### **Scatter Plot: Total Day Minutes vs Total Day Calls**

The scatter plot highlights the clustering of churned and non-churned customers:

* Churned customers are more likely to be in the upper ranges of total\_day\_minutes, but the spread of total\_day\_calls remains similar across both groups.
* This indicates that churn is more influenced by the duration of calls rather than their frequency during the day.



#### **Heatmap Analysis**

The heatmap of continuous features reaffirms strong correlations between related variables, such as:

* Total\_day\_minutes, total\_day\_charge, and charge\_per\_minute are closely correlated, indicating redundancy.
* Weak correlations between account\_length and churn suggest that account tenure does not strongly predict churn.

#### **Chi-Square Test Results**

A Chi-Square test was conducted for categorical features against churn:

* **International Plan**: A significant p-value (<0.0001) confirms a strong association between having an international plan and churn, making it a key predictive feature.
* **Voice Mail Plan**: The p-value (<0.0001) indicates that having a voicemail plan is also associated with churn, but to a lesser extent than the international plan.
* **State**: While there is a significant association (p = 0.002), the effect may reflect regional variations in customer behavior or service quality.
* **Area Code**: The p-value (0.91) suggests no significant relationship between area code and churn, indicating that this feature has limited predictive value.

The exploratory data analysis revealed several critical insights into the features and their relationship with customer churn. The dataset shows a significant class imbalance, with approximately 85% of customers not churning and only 15% churning, emphasizing the need for balancing techniques in predictive modeling. Univariate analysis indicated that features related to usage and charges, such as total\_day\_minutes and total\_day\_charge, exhibit higher values for churned customers, suggesting dissatisfaction among heavy users. Conversely, total\_day\_calls showed no significant difference between churned and non-churned customers.

Categorical features provided clear distinctions; customers with an international\_plan have a markedly higher likelihood of churning, while those with a voice\_mail\_plan generally churn less frequently. Regional trends also emerged, with slightly higher churn rates observed in the Midwest and South compared to other regions. Area codes, however, showed no meaningful variation in churn patterns.

Bivariate analysis reinforced the strong correlations between usage metrics and charges, as expected. However, it also highlighted that features like account\_length and area\_code have limited predictive value for churn. Boxplots revealed that churned customers tend to have higher median values for metrics like total\_day\_minutes and total\_day\_charge, aligning with the idea that heavy users may experience dissatisfaction or higher costs.

Multivariate analysis, including scatterplots and pairplots, demonstrated relationships between continuous features, particularly the interplay of high day-time usage and churn. Chi-square tests for categorical variables showed a strong relationship between churn and features like international\_plan and voice\_mail\_plan, while state exhibited moderate significance.

Overall, the EDA indicates that high usage, particularly with international plans, is a significant driver of churn. The findings suggest that retention strategies should focus on these high-risk segments, and the modeling process should emphasize these key features while addressing the class imbalance to improve prediction accuracy.

# **Modeling**

For this analysis, three decision tree models were utilized: a **Baseline Model**, a **Tuned Model**, and a **Further Tuned Model**. The **Baseline Model** serves as a starting point to understand the performance of a simple decision tree without any modifications or class imbalance handling. It provides a reference for evaluating improvements made in subsequent models. The **Tuned Model** incorporates techniques such as hyperparameter optimization (e.g., adjusting depth, splits, and leaf sizes) and class imbalance handling using SMOTE (Synthetic Minority Oversampling Technique) to improve recall and ensure the model performs better on the minority class (churn). The **Further Tuned Model** refines the tuning process with an expanded hyperparameter grid and additional pruning to balance precision and recall, aiming for a model that minimizes both false positives and false negatives. These models were selected to systematically improve upon the baseline by addressing class imbalance, optimizing performance metrics like F1-Score, precision, recall, and AUC-ROC, and ensuring robust and interpretable predictions for SyriaTel.

# **Evaluation**

Below is the comparison table showing the metrics of success for the baseline, tuned, and further tuned decision tree models:

| **Model** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** | **Confusion Matrix (TN, FP, FN, TP)** |
| --- | --- | --- | --- | --- | --- |
| **Baseline Decision Tree** | 0.75 | 0.80 | 0.77 | 0.88 | [[816, 39], [29, 116]] |
| **Tuned Decision Tree** | 0.50 | 0.77 | 0.61 | 0.82 | [[746, 109], [34, 111]] |
| **Further Tuned Decision Tree** | 0.94 | 0.79 | 0.86 | 0.88 | [[848,7], [31,114]] |

**Summary of Models**

1. **Baseline Decision Tree**:
   * The baseline model achieved a **precision of 0.75**, indicating that 75% of predicted churn cases were correct.
   * It achieved a **recall of 0.80**, meaning it successfully identified 80% of actual churn cases.
   * The **F1-Score of 0.77** reflects a good balance between precision and recall.
   * The **AUC-ROC score of 0.88** indicates strong discriminatory power in distinguishing churn and non-churn customers.
   * It had 816 true negatives, 39 false positives, 29 false negatives, and 116 true positives in the confusion matrix.
2. **Tuned Decision Tree (SMOTE)**:
   * This model focused on addressing class imbalance using SMOTE and hyperparameter tuning. It achieved a **precision of 0.50**, meaning half of its churn predictions were correct.
   * With a **recall of 0.77**, it detected 77% of actual churn cases.
   * The **F1-Score of 0.61** reflects a drop in performance compared to the baseline model, as it struggled with precision.
   * The **AUC-ROC score of 0.82** is lower than the baseline but still reflects good model performance.
   * The confusion matrix showed more false positives (109) compared to the baseline.
3. **Further Tuned Decision Tree**:
   * The further tuned model performed the best in terms of precision with a score of **0.94**, meaning 94% of its churn predictions were correct.
   * It had a **recall of 0.79**, slightly lower than the baseline but still high.
   * The **F1-Score of 0.86** reflects a significant improvement in balancing precision and recall.
   * The **AUC-ROC score of 0.88** matches the baseline, indicating equally strong discriminatory power.
   * This model had the lowest false positives (7) and a balanced trade-off in the confusion matrix.

### **Best Model**

The **Further Tuned Decision Tree** is the best model overall. Its **high precision (0.94)** ensures minimal false positives, which is crucial for efficient resource allocation in retention strategies. While its recall (0.79) is slightly lower than the baseline, it still identifies a significant proportion of churn cases. Additionally, the **F1-Score (0.86)** balances precision and recall, making it the most effective model for this problem.

# **Conclusion**

The analysis revealed that customers with international plans and those who frequently contact customer service are significantly more likely to churn. This suggests dissatisfaction with service quality or unmet expectations for international features as potential drivers of customer attrition. Interestingly, while higher usage metrics such as total day minutes and charges indicate active engagement, they do not directly correlate with churn unless coupled with service issues. This implies that dissatisfied high-usage customers, likely paying premium charges, may feel underserved or undervalued. Addressing these issues with targeted interventions, such as improving service quality for high-value customers and offering personalized plans or loyalty rewards for international users, can help mitigate churn. Overall, the findings emphasize the need to address service-related grievances, particularly for segments with high international usage or frequent customer service interactions, to foster customer satisfaction and retention.

# **Recommendations**

1. **Target High-Risk Customers:**
   * Focus retention strategies on customers with international plans, as they are more likely to churn.
   * Proactively address complaints and concerns of customers who frequently contact customer service.
2. **Implement Loyalty Programs:**
   * Introduce tailored loyalty initiatives such as discounts, reward points, or free upgrades for high-risk segments identified by the churn prediction model.
   * Offer exclusive benefits for long-term customers to enhance their commitment to the company.
3. **Enhance Customer Satisfaction:**
   * Conduct regular customer satisfaction surveys to identify key dissatisfaction drivers and unmet needs.
   * Use survey insights to improve service quality, adjust pricing structures, and meet evolving customer expectations.
4. **Improve Customer Service:**
   * Train customer service representatives to handle complaints more effectively and resolve issues promptly.
   * Utilize predictive models to anticipate service issues and address them before they impact customer satisfaction.
5. **Tailored Communication:**
   * Use the churn prediction model to send personalized offers and notifications to customers at risk of churning.
   * Highlight features or services that match the usage patterns and preferences of individual customers.
6. **Monitor and Refine Strategies:**
   * Continuously evaluate the effectiveness of retention strategies and loyalty programs using customer feedback and churn metrics.
   * Regularly update the churn prediction model with new data to improve accuracy and adapt to market changes.

By implementing these strategies, Syriatel can reduce churn, enhance customer satisfaction, and maintain a competitive edge in the telecommunications market.

# **Next Steps**

1. **Deployment of the Churn Prediction Model:**
   * Integrate the churn prediction model into Syriatel's Customer Relationship Management (CRM) system to enable real-time identification of at-risk customers.
   * Automate the process of flagging high-risk customers and triggering retention strategies, such as personalized offers or proactive customer service interventions.
   * Ensure the model is accessible to all relevant teams, including marketing, customer service, and product development, for coordinated retention efforts.
2. **Enhance Data Collection:**
   * Incorporate additional data points such as customer satisfaction scores, social media sentiment analysis, and feedback from surveys to improve the model’s prediction accuracy.
   * Include external factors like competitor pricing, promotional campaigns, and market conditions to better capture the drivers of churn.
   * Set up real-time data pipelines to ensure continuous updates and accuracy in the model’s training dataset.
3. **Actionable Insights for Retention Strategies:**
   * Leverage the model’s outputs to design and implement targeted retention campaigns, such as offering discounts, service upgrades, or loyalty rewards to high-risk customers.
   * Use insights from churn patterns to adjust pricing structures, optimize service plans, and address frequently reported service issues.
   * Develop segmented retention strategies based on customer demographics and usage patterns identified by the model.
4. **Continuous Monitoring and Model Improvement:**
   * Regularly monitor the performance of the deployed model using evaluation metrics such as precision, recall, F1-score, and AUC-ROC to ensure it meets business objectives.
   * Retrain the model periodically with new data to account for evolving customer behaviors and market conditions.
   * Experiment with additional advanced algorithms, such as ensemble methods or deep learning, to further enhance prediction accuracy.
5. **Training and Team Enablement:**
   * Conduct training sessions for marketing, customer service, and sales teams to understand and effectively utilize the model’s predictions.
   * Equip teams with the tools and knowledge needed to interpret the model’s outputs and execute retention strategies efficiently.
6. **Operational Integration and Automation:**
   * Establish workflows for integrating the model’s predictions into day-to-day operations, such as automatically generating prioritized customer lists for retention efforts.
   * Automate feedback loops where outcomes from retention campaigns are fed back into the model to refine predictions and strategies over time.
7. **Expansion of Retention Measures:**
   * Launch experimental programs, such as tiered loyalty programs or exclusive offers for high-value customers, and measure their effectiveness in reducing churn.
   * Use predictive insights to identify opportunities for cross-selling and upselling, enhancing customer lifetime value.
8. **Stakeholder Engagement:**
   * Provide senior management and key stakeholders with regular updates on the model’s performance and its impact on reducing churn rates.
   * Use data-driven insights from the model to guide strategic decisions, such as market positioning, resource allocation, and pricing strategies.

By following these steps, Syriatel can ensure the effective deployment and continuous improvement of the churn prediction model, leading to enhanced customer retention and long-term business success.