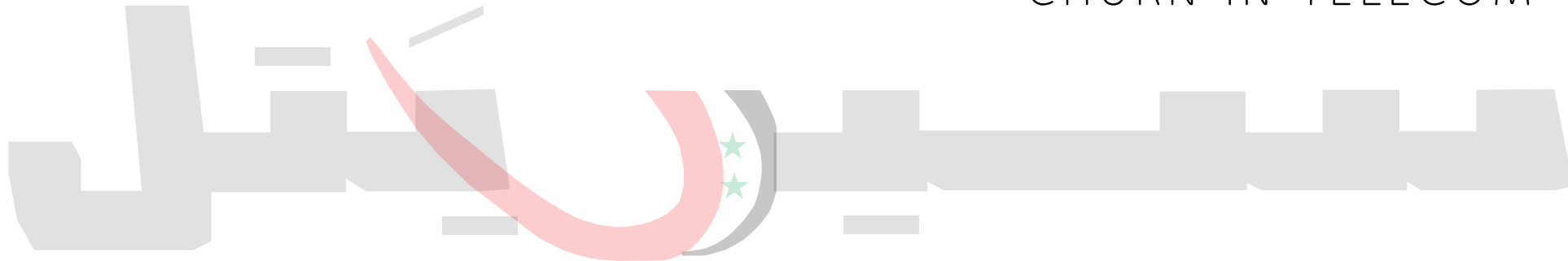


# SyriaTel



---

UNDERSTANDING CUSTOMER  
CHURN IN TELECOM



# Business Understanding

---

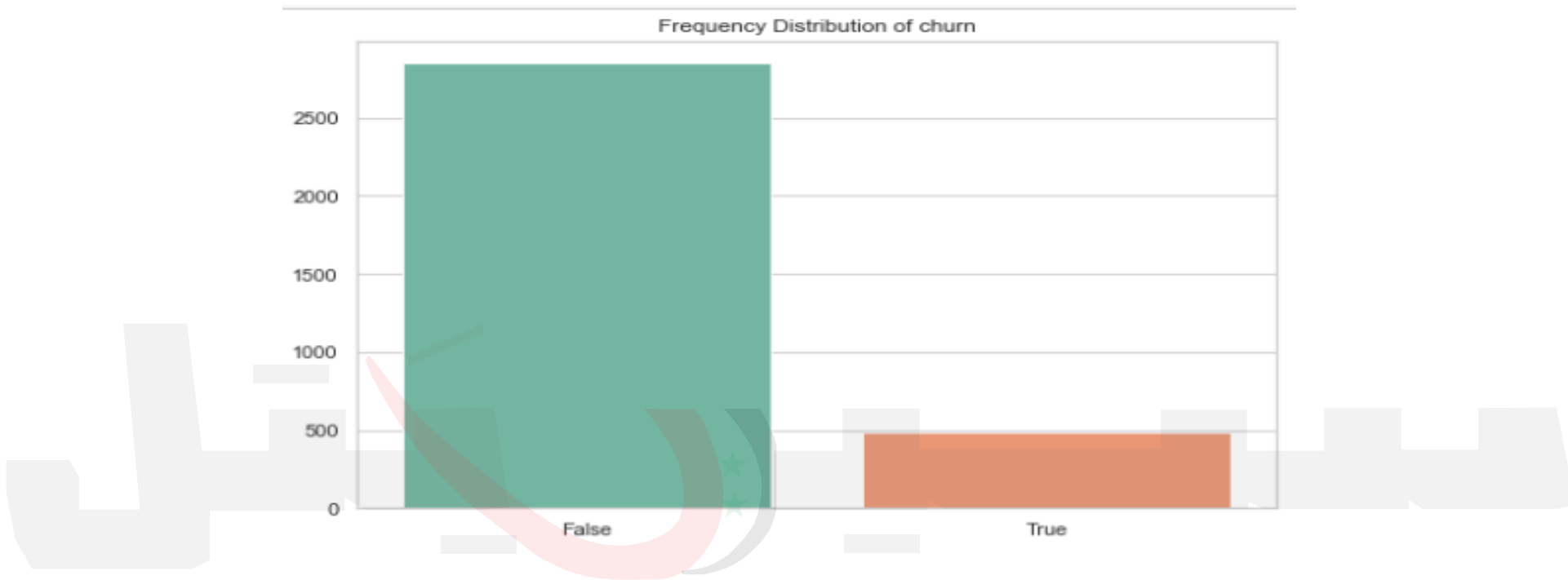
Customer churn is a significant concern for Syriatel, as it directly impacts the company's revenue and growth. Churn occurs when customers stop using Syriatel's services, and addressing this issue is crucial for maintaining a stable customer base. High churn rates mean losing valuable customers, which increases the cost and effort required to acquire new ones and potentially damages the company's reputation. Therefore, understanding the reasons behind customer churn is essential for Syriatel to develop effective retention strategies.

The primary objective of this project is to identify the factors contributing to customer churn at Syriatel and provide actionable insights to reduce churn rates. This involves analyzing customer demographics, usage patterns, service plans, and customer service interactions. By understanding these factors, Syriatel can tailor its services and improve customer satisfaction, leading to higher retention rates.

# Key questions addressed in this project include:

---

1. What are the main factors driving customer churn at Syriatel?
2. How can Syriatel predict which customers are likely to churn?
3. What strategies can be implemented to retain customers and reduce churn?

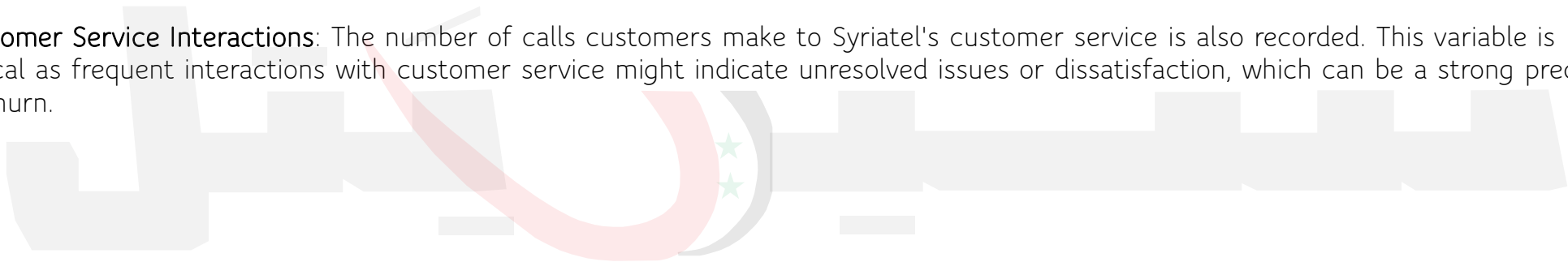


# Data Understanding

---

To effectively address customer churn at Syriatel, it is crucial to thoroughly understand the available data sourced from Kaggle.com. The dataset used for this analysis comprises detailed information on 3,333 customers, encompassing 21 attributes that provide insights into customer behavior, service usage, and interactions with Syriatel. The key variables in this dataset include:

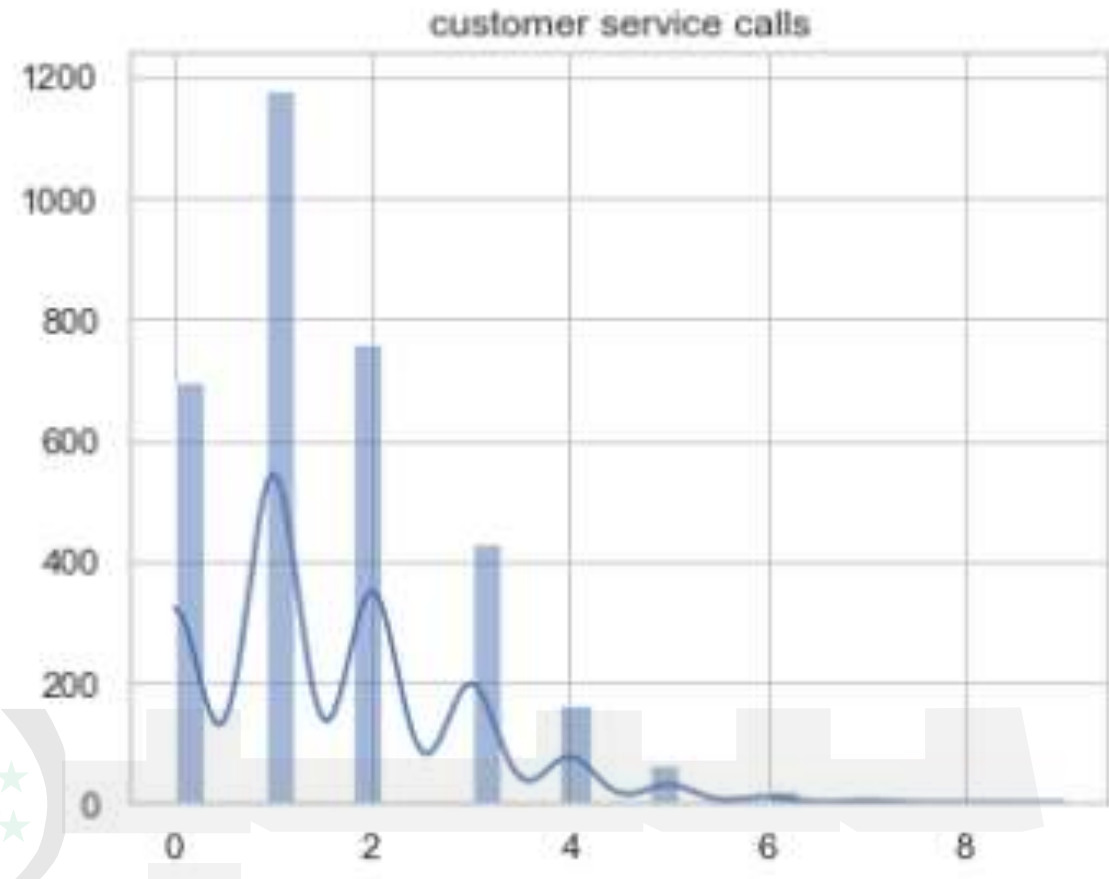
- **Customer Demographics:** Information such as the customer's state and area code helps identify geographic patterns and regional differences in churn behavior. Understanding where customers are located can reveal if certain areas have higher churn rates and need targeted interventions.
- **Service Plans:** Details about whether customers have international plans and voice mail plans are included. These attributes help in understanding the preferences and needs of customers, and how these plans might influence their decision to stay with or leave Syriatel.
- **Usage Metrics:** The dataset captures customers' usage patterns, including total minutes and calls during the day, evening, night, and international calls. Analyzing these metrics helps identify high and low usage customers and how their usage impacts their likelihood of churning.
- **Customer Service Interactions:** The number of calls customers make to Syriatel's customer service is also recorded. This variable is critical as frequent interactions with customer service might indicate unresolved issues or dissatisfaction, which can be a strong predictor of churn.



# Distribution of Customer service calls

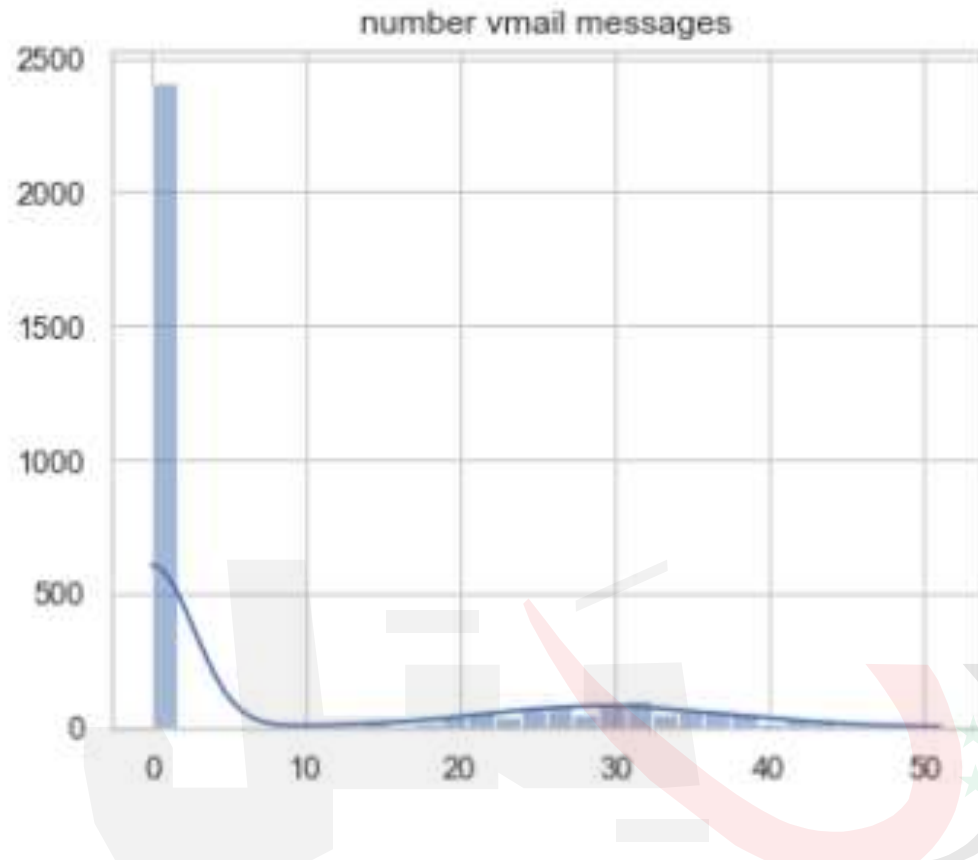
---

**Distribution:** Skewed right with most values at 1. Central Tendency: Mode at 1, with a long tail towards higher values. Spread: High concentration at lower values, indicating infrequent customer service interactions. Relationships: May correlate with customer satisfaction or issues.



# Distribution of number of Voice mail messages

---

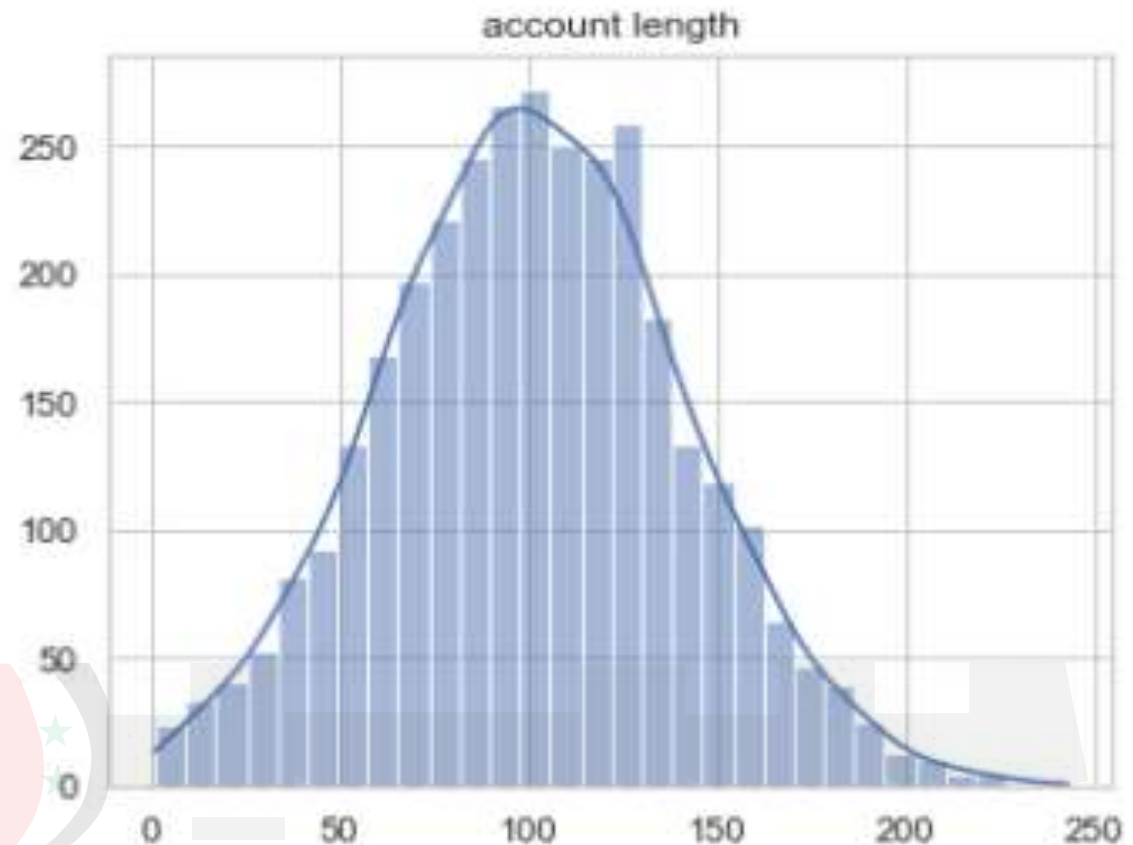


Skewed right with most values at 0. Central Tendency: Mode at 0, with a long tail towards higher values. Spread: High concentration at 0, few users have many voice mail messages. Relationships: May correlate with usage behavior or type of service plan.

# Distribution of Account length

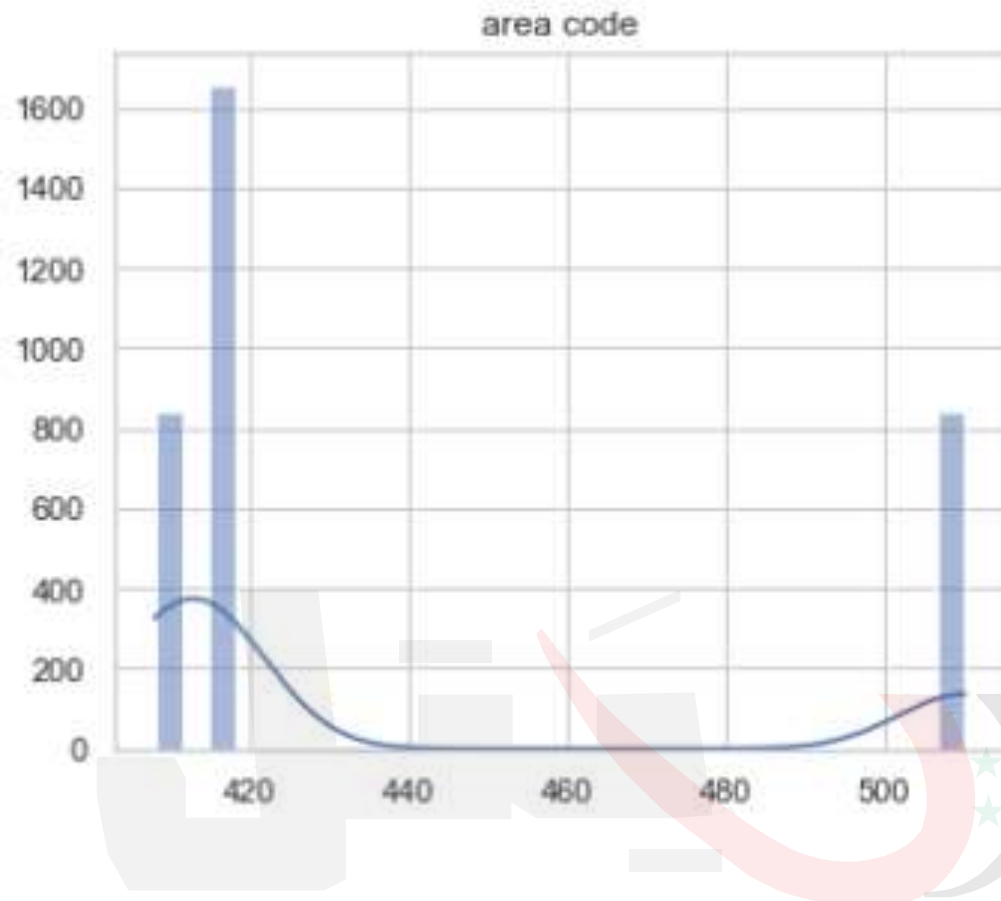
---

Nearly normal distribution. Central Tendency: Mean around 100. Spread: Symmetric, with most values between 75 and 125. Relationships: No immediate relationships suggested by the plot.



# Distribution of Area code

---



Bimodal distribution with spikes at 408, 415, and 510. Central Tendency: Three distinct modes. Spread: Concentrated at specific area codes, indicating categorical data. Relationships: Categorical variable, likely used to segment data by region.



# Data exploration

---

This step involves analyzing the key variables to understand their distribution and relationships, which can help identify factors contributing to customer churn.

**Geographic Distribution:** Customers are distributed across various states and area codes. By mapping this data, we can identify regions with higher or lower churn rates and tailor strategies accordingly.

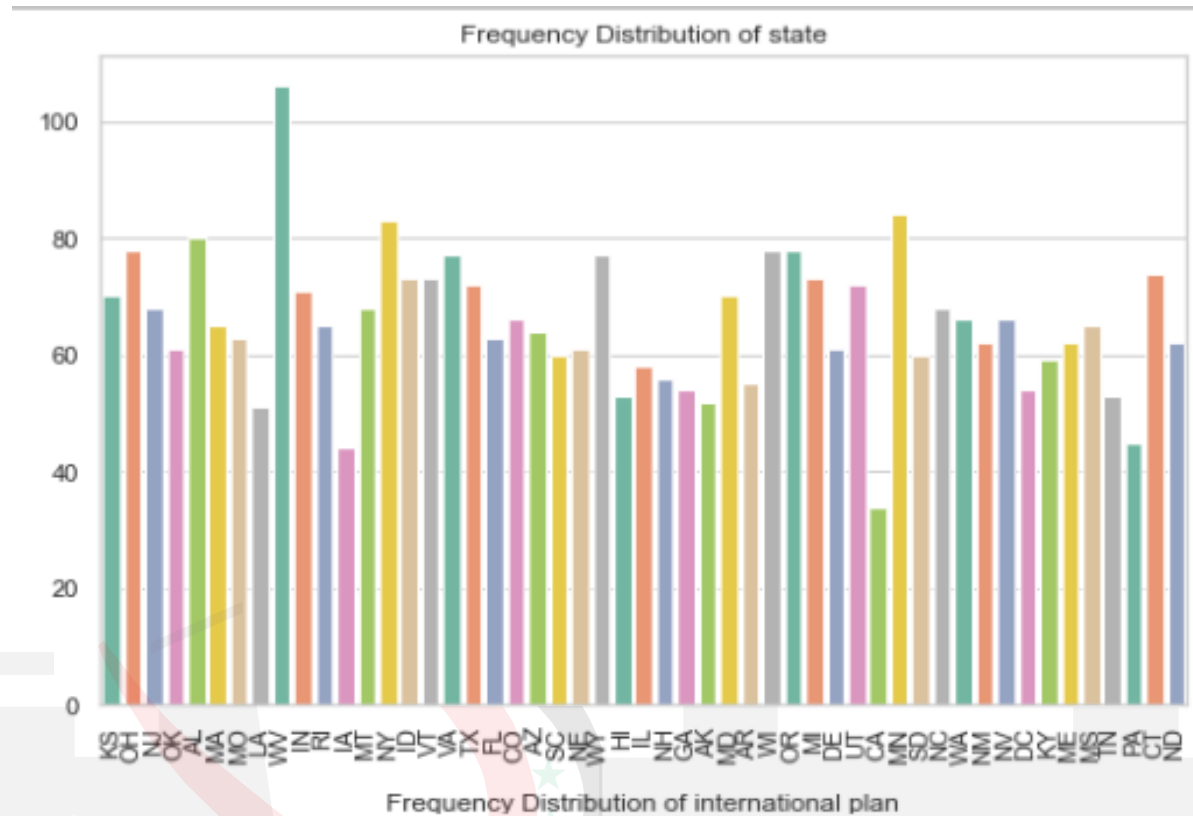
**International and Voice Mail Plans:** Understanding the distribution of customers with these plans helps in identifying if certain plans are associated with higher or lower churn rates.

**Day, Evening, Night, and International Calls:** Analyzing the total minutes and calls during different periods provides insights into customer usage behavior. For example, heavy users might have different churn patterns compared to light users.

**Frequency of Calls to Customer Service:** The number of customer service calls is a critical factor, as frequent interactions may indicate unresolved issues or dissatisfaction. Understanding this distribution can help in identifying at-risk customers.

# International Plans

---



# Features Selection

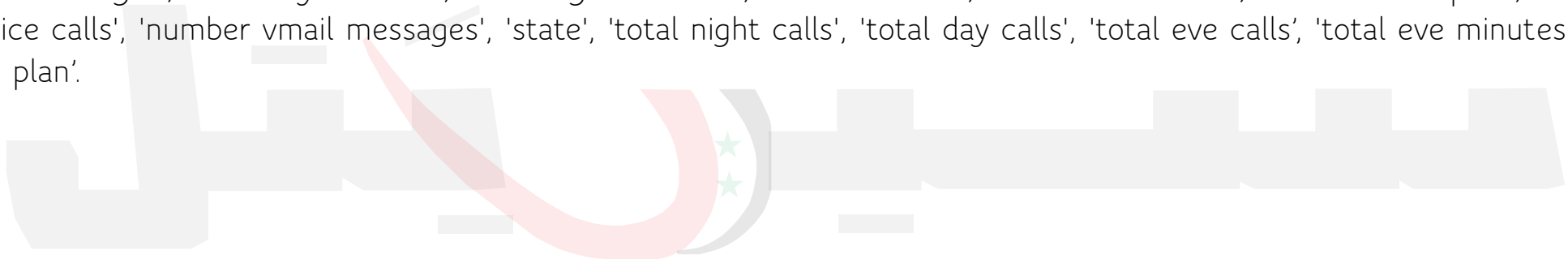
---

## Techniques Used:

- **Multicollinearity Analysis:** This technique identifies pairs of features that are highly correlated with each other. High multicollinearity can distort the model's performance and interpretation.
- **Feature Elimination:** Based on the VIF scores, features with high multicollinearity are reviewed and potentially removed. This step ensures that the remaining features are independent and contribute uniquely to the model's predictive power.

## Features Selected

'account length', 'total day minutes', 'total night minutes', 'total intl calls', 'total intl minutes', 'international plan', 'customer service calls', 'number vmail messages', 'state', 'total night calls', 'total day calls', 'total eve calls', 'total eve minutes', 'voice mail plan'.

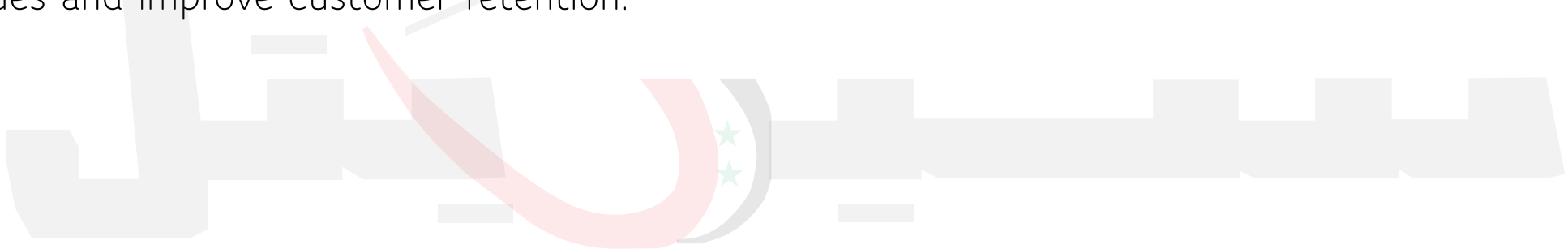


# Modelling

---

To effectively predict customer churn at Syriatel, we employed various modeling techniques that help identify the likelihood of a customer leaving the service. The modeling process involved selecting appropriate algorithms, training the models on historical data, and evaluating their performance to ensure accuracy and reliability.

**Objective:** The primary goal is to develop a predictive model that can accurately identify customers who are at high risk of churning. This allows Syriatel to proactively address potential issues and improve customer retention.



# Model selection

---

We considered three machine learning algorithms to find the best fit for our classification problem:

- **Logistic Regression:** A simple and interpretable model that predicts the probability of churn based on input features. It is effective for binary classification tasks.
- **Decision Tree:** A non-parametric model that splits the data into branches to predict the target variable. It provides clear, visual decision rules.
- **Random Forest:** An ensemble learning method that combines multiple decision trees to improve prediction accuracy and control over-fitting. It provides insights into feature importance.

# Data Preprocessing

---

Before training the models, the data underwent preprocessing steps including:

Scaling Numerical Features: Using `StandardScaler` for features such as account length, total intl calls, and others.

Encoding Categorical Variables: Using `OneHotEncoder` for features like international plan, state, and voice mail plan.

Balancing Data: Applying SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in the training data.



# Training and Validation

---

- **Training Set:** The historical data was split into training and validation sets to train the models and tune their parameters.
- **Cross-Validation:** Cross-validation techniques were used to evaluate the model performance and ensure that the model generalizes well to unseen data.



# Model Performance (Confusion Matrix)

---

Decision Tree:

TP: 70

TN: 531

FP: 35

FN: 31

Logistic  
Regression:

TP: 79

TN: 448

FP: 118

FN: 22

Random  
Forest:

TP: 72

TN: 554

FP: 12

FN: 29

Random Forest has the highest number of true positives (correctly predicted positive cases) and true negatives (correctly predicted negative cases).

It also has the lowest number of false positives (incorrectly predicted positive cases) and false negatives (incorrectly predicted negative cases).

Therefore, I recommend using the Random Forest model for better overall performance.



# Model Performance (Accuracy)

---

The models were evaluated based on metrics such as accuracy, precision, and recall. These metrics help in understanding the effectiveness of the models in predicting churn. Random Forest is the best model here with an accuracy of approximately 94%.

## Logistic Regression:

- Precision: 0.401
- Recall: 0.782
- Accuracy: 0.790

## Decision Tree:

- Precision: 0.667
- Recall: 0.693
- Accuracy: 0.901

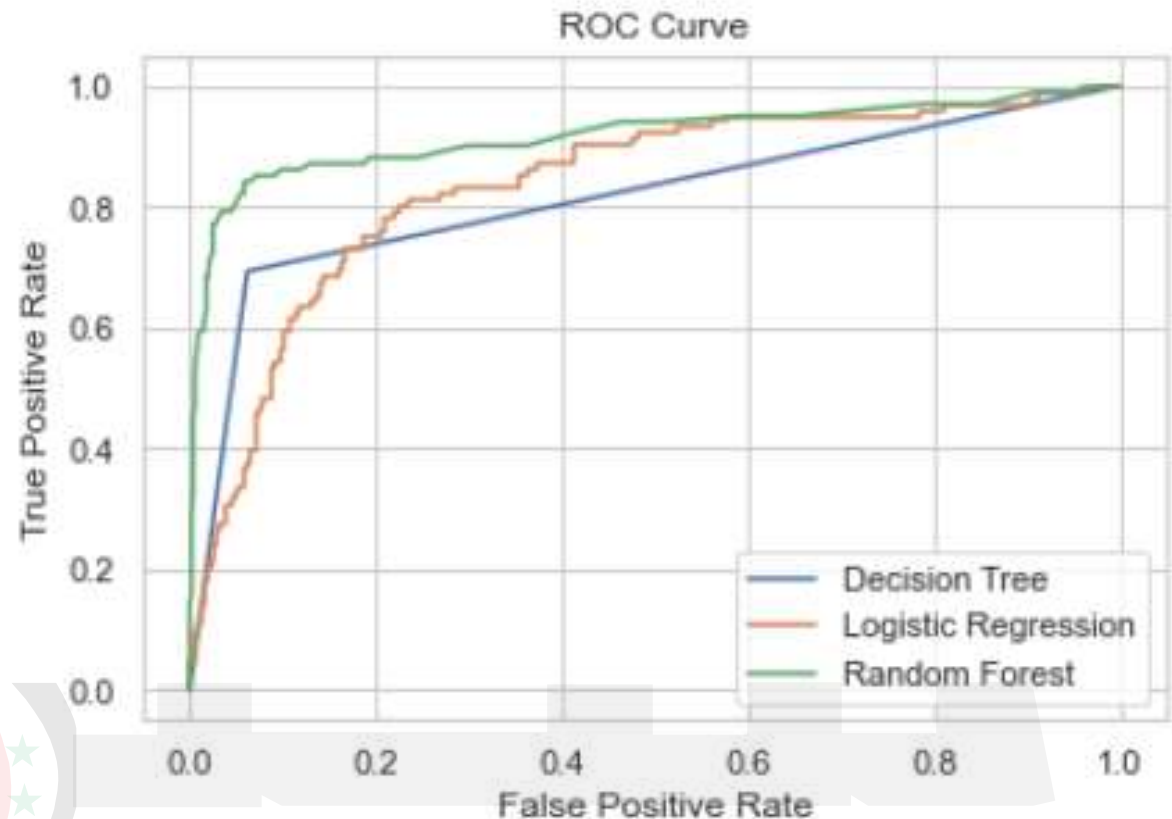
## Random Forest:

- Precision: 0.857
- Recall: 0.713
- Accuracy: 0.939

# Model Performance (ROC curve)

---

Based on this ROC curve alone and without considering other factors like precision-recall trade-off or practical constraints such as computational efficiency or data availability, I recommend using the Random Forest model because its corresponding curve is closest to the top left corner, indicating higher true positive rates for lower false positive rates compared to other models represented in this graph.



# Conclusion

---

The primary business problem Syriatel faces is customer churn. By accurately predicting which customers are likely to churn, Syriatel can implement targeted retention strategies to reduce churn rates, thus improving customer satisfaction and revenue stability.

**Accurate Predictions:** The Random Forest model provides the highest accuracy (93.7%) and precision (0.847), indicating it is the most reliable model for predicting churn. Precision is crucial as it indicates how many of the predicted churners actually churned, helping Syriatel focus its retention efforts efficiently.

**Balanced Trade-offs:** Although Logistic Regression has the highest recall (0.782), its precision is low (0.401), meaning it predicts many non-churners as churners, leading to unnecessary retention efforts. The Decision Tree offers a balanced approach with good precision (0.654) and recall (0.733), making it a viable alternative if simplicity and interpretability are preferred.

**Resource Optimization:** By using the Random Forest model, Syriatel can optimize its resources by accurately identifying high-risk customers and taking proactive measures to retain them, thus minimizing wasted efforts on customers who are not likely to churn.

# Recommendations

---

**Implement Random Forest for Churn Prediction:** Use the Random Forest model for its high accuracy and precision, ensuring reliable identification of potential churners.

**Proactive Retention Strategies:** Develop targeted retention campaigns for high-risk customers identified by the model. These could include personalized offers, improved customer service, and loyalty programs.

**Monitor Model Performance:** Regularly evaluate the model's performance and retrain it with new data to maintain its accuracy and relevance as customer behaviors and market conditions change.

**Analyze Churn Reasons:** Use insights from the model to understand the main drivers of churn. Focus on improving areas like customer service, plan options, and network quality to address the root causes.

**Customer Feedback Loop:** Implement a feedback loop to gather data from customers who were predicted to churn but chose to stay. Use this information to refine retention strategies and improve the prediction model.