# Phase 3 project

SyriaTel Classification Models

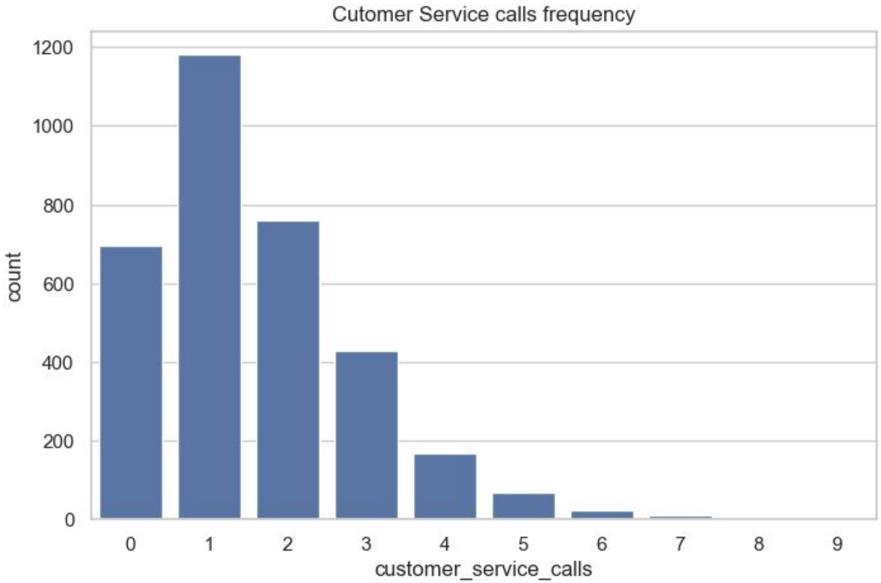
# Business Understanding

I have been commissioned by the Customer Retention Manager at SyriaTel to develop a binary classification model aimed at predicting customer churn. This model is essential for identifying customers at high risk of leaving the company, allowing SyriaTel to proactively address potential attrition. By analyzing the key factors contributing to churn, the Retention Manager can implement targeted strategies to enhance customer retention, thereby minimizing revenue loss and boosting long-term profitability. The insights generated from this model will be instrumental in shaping effective marketing campaigns and optimizing customer engagement efforts.

# Data Understanding

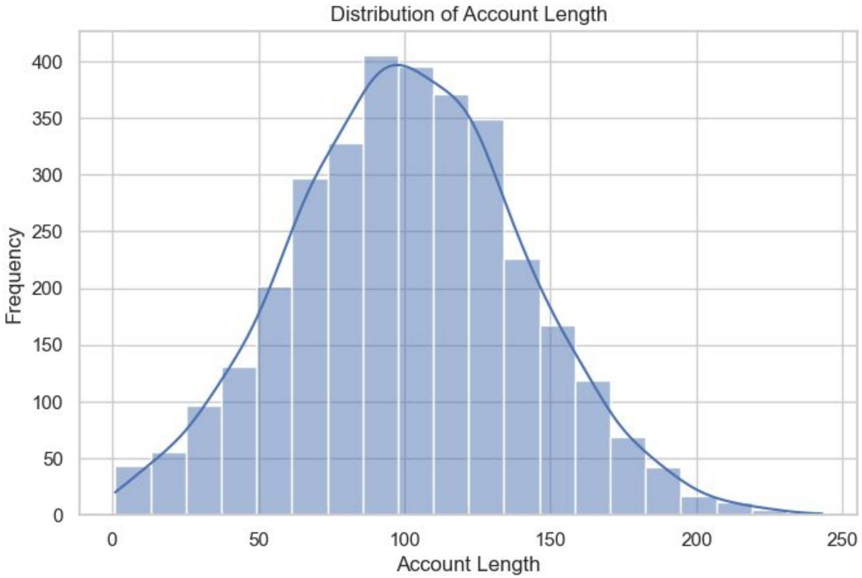
The dataset, sourced from Kaggle, consists of 3,333 records from SyriaTel customers, capturing their service usage and interactions with the company. The primary objective is to use these features to predict customer churn, as indicated by the `churn` column.

# Customer Service calls frequency



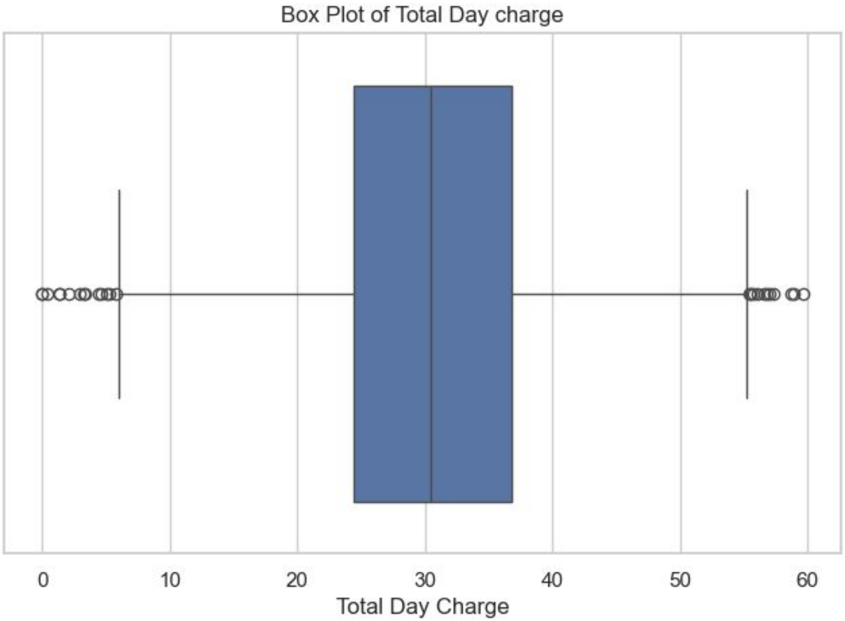
The visual above indicates that the majority of customers make fewer than two service calls. This trend suggests that there are relatively few issues, and when problems do arise, they are typically resolved within two or fewer calls.

# Distribution of Account Length



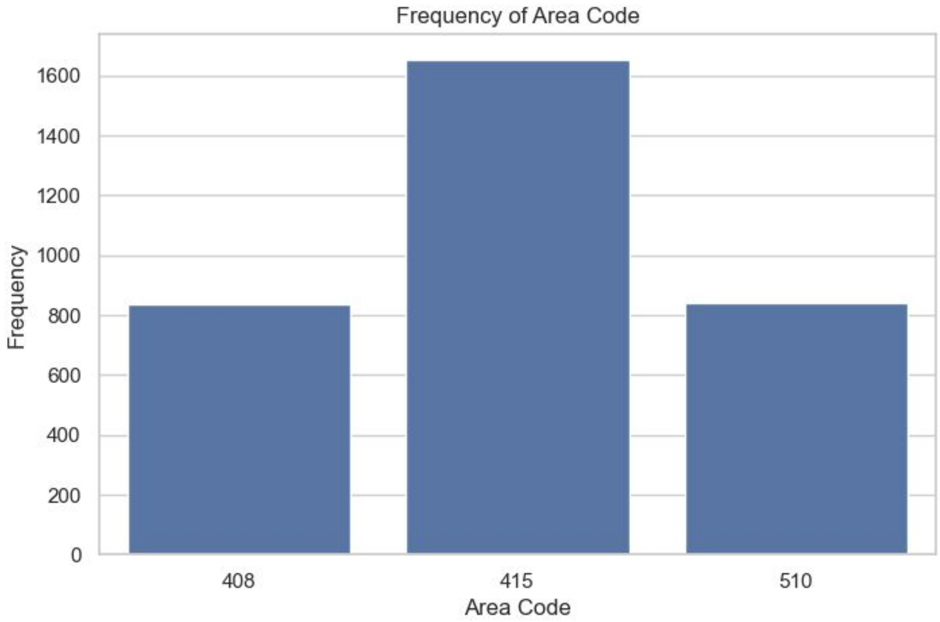
The histogram represents the distribution of account lengths, ranging from 0 to 250. The data follows a normal distribution, peaking at around 90. This suggests that most accounts have a length close to 90, with fewer accounts having significantly shorter or longer durations.

# Box plot of Total Day Charge



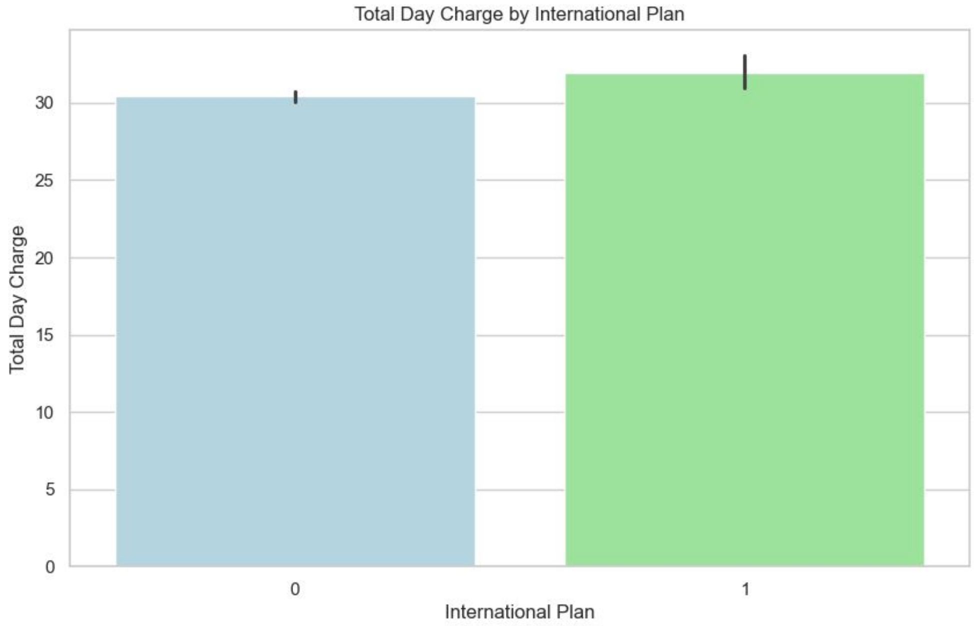
Most customers have total day charges between $25 and $37, with half of them below $31. While the charges are generally balanced, a few customers have unusually low or high charges, which are considered outliers.

# Frequency of Area Code



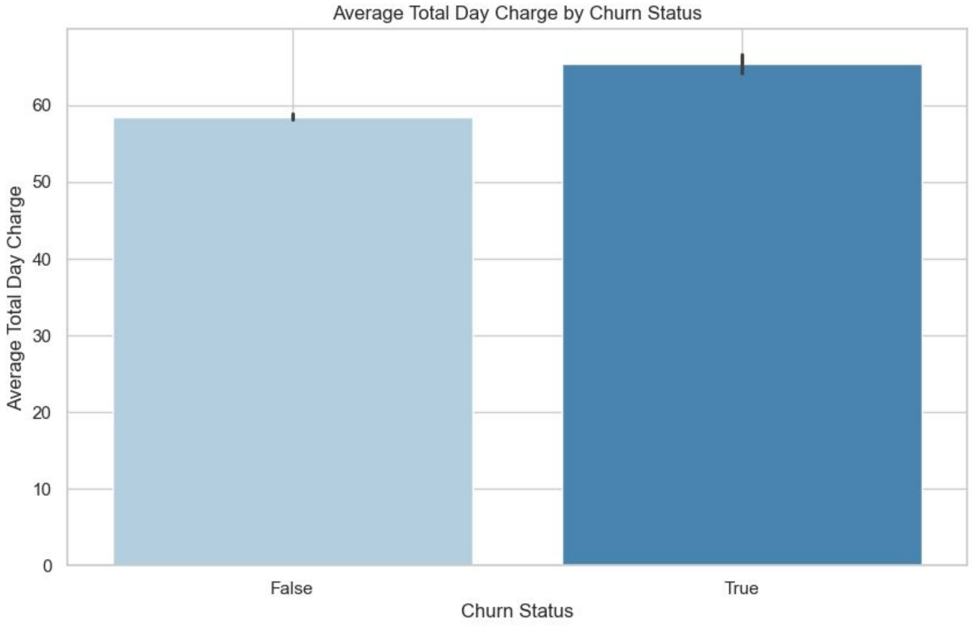
The bar graph shows that customers with area code 415 make the most calls, with an average of about 16.2 calls. In comparison, area codes 408 and 510 each have around 800 calls, indicating a consistent but lower call frequency than 415.

# Total Day charge by international Plan



The bar graph compares the total day charges between customers who use international services and those who don't. Both groups have similar charges, but customers who use international services have a slightly higher average, around 33. This suggests that international service usage slightly increases day charges.

# Total Day charge by international Plan



The bar graph shows that customers who have churned tend to have a higher average total day charge (about 65) compared to those who haven't churned (around 58). This suggests that customers who leave the service usually incur higher day charges.

# Introduction to Modeling

- **Objective:** Predict customer churn using classification models.
- **Models Used:** Logistic Regression & Decision Tree.
- **Data Overview:** Customer information, service usage patterns, and churn status.
- **Goal:** Identify patterns and key factors that contribute to customer churn, enabling proactive retention strategies.

# Best Logistic regression

| Accuracy | | 86 |
|---|---|---|
| **Metric** | **Non-Churned (False)** | **Churned (True)** |
| **Precision** | 87% | 48% |
| **Recall** | 96% | 22% |
| **F1-Score** | 91% | 30% |

The baseline logistic regression model, which achieved an accuracy of 86%, is the best model after multiple iterations. It performs exceptionally well in predicting non-churned customers, with high recall (96%) and precision (87%). However, it struggles with churned customers, showing lower recall (22%) and precision (48%). This imbalance is likely due to the smaller number of churned cases in the dataset.

# Best Decision tree Model

| **Accuracy** | | **97.9%** |
|---|---|---|
| **Metric** | **Non-Churned (False)** | **Churned (True)** |
| **Precision** | 98% | 99% |
| **Recall** | 100% | 87% |
| **F1-Score** | 99% | 93% |

The decision tree model, which underwent hyperparameter tuning and grid search to find the best configuration, achieved outstanding performance. With an accuracy of 98%, it excels in predicting both non-churned and churned customers. It shows exceptional precision (98% for non-churned and 99% for churned), perfect recall (100% for non-churned), and a high recall (87%) for churned customers. The high F1-scores (99% for non-churned and 93% for churned) reflect its strong overall performance, making it the most effective model for this task.

# Evaluation

The decision tree model, which underwent hyperparameter tuning and grid search, significantly outperforms the baseline logistic regression model. With an accuracy of 98%, the decision tree excels in both precision and recall for non-churned customers (98% precision, 100% recall) and churned customers (99% precision, 87% recall). In contrast, the logistic regression model achieved an accuracy of 86%, showing high performance in predicting non-churned customers but struggling with churned customers (48% precision, 22% recall). Given these results, the decision tree model is the superior choice for accurately identifying and targeting at-risk customers, making it the better fit for solving our business needs.

# Limitations

**Class Imbalance**: The dataset is imbalanced, with more non-churned than churned cases. This can lead to a model that favors predicting the majority class and struggles with the minority class.

**Feature Relevance and Data Quality**: Some features like `state` and `area_code` may be irrelevant, introducing noise. Missing or inaccurate data can also impact model performance.

**Overfitting Risk**: The decision tree's high accuracy might suggest overfitting, where it performs well on training data but may not generalize to new data.

**Model Interpretability**: The decision tree's complexity from hyperparameter tuning can make it harder to interpret and explain, which may be a drawback for practical use.

**Performance Variability**: Model performance can vary with different data subsets or changes in data distribution, potentially affecting reliability in real-world applications.

# Findings

**Logistic Regression Models**: Several iterations of logistic regression models, including those with SMOTE and over/under sampling, did not achieve satisfactory performance. These methods did not effectively address the class imbalance, resulting in lower accuracy and performance issues.

**Decision Tree Performance**: A baseline decision tree model, without hyperparameter tuning, performed better than the logistic regression models but was outperformed by a more advanced decision tree model.

**Tuned Decision Tree Model**: The decision tree model with hyperparameter tuning and grid search emerged as the best performer. It significantly outperformed both the logistic regression models and the untuned decision tree, demonstrating its effectiveness in addressing the churn prediction problem.

# Recommendations

**Enhance Customer Retention Strategies:**
Use insights from the decision tree model to target customers at high risk of leaving. Implement strategies like personalized offers, loyalty programs, and better customer support to reduce churn and boost loyalty.

**Optimize Features and Pricing Plans:**
Address class imbalance issues by refining features and adjusting pricing. For example, if high charges are linked to churn, offer discounts or flexible pricing to retain more customers.

**Regularly Update Models:**
Keep your models accurate by updating them regularly. As customer behavior and market conditions change, retrain your models with new data to ensure they stay relevant and effective.