# SyriaTel Customer Churn Prediction

By: Mary Njeri Kamithi

Team Mentor: William Okomba  Cohort: DSFT 13

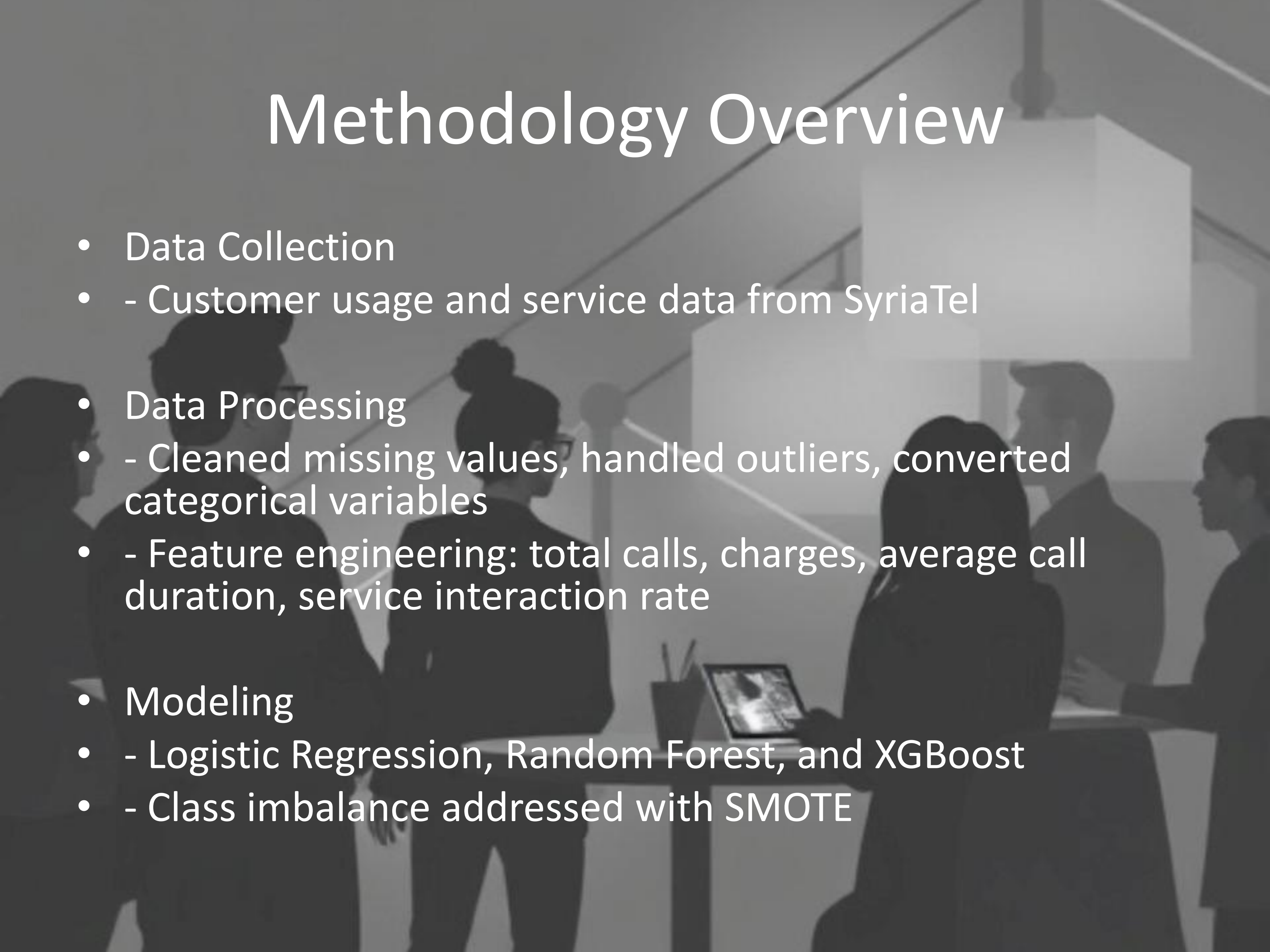
Powered by Twistly

# Project Objectives

- Build a predictive model to identify high-risk customers likely to churn

- Understand the key drivers of customer churn

- Enable SyriaTel to take proactive measures to reduce churn and improve retention



# Methodology Overview

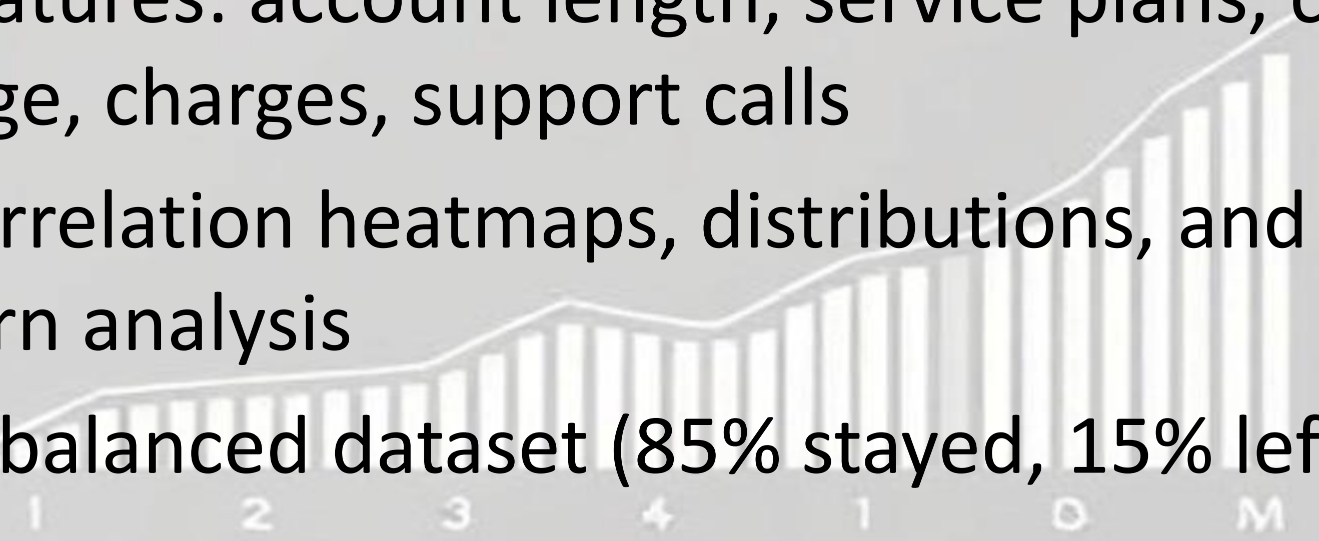
- Data Collection
- - Customer usage and service data from SyriaTel

- Data Processing
- - Cleaned missing values, handled outliers, converted categorical variables
- - Feature engineering: total calls, charges, average call duration, service interaction rate

- Modeling
- - Logistic Regression, Random Forest, and XGBoost
- - Class imbalance addressed with SMOTE

# Data Analysis

- - Dataset: 3,333 customers (≈15% churners)
- - Features: account length, service plans, call usage, charges, support calls
- - Correlation heatmaps, distributions, and churn analysis
- - Imbalanced dataset (85% stayed, 15% left)



# Main Findings

- Churn strongly influenced by total charges, minutes used, and customer service interactions

- XGBoost outperformed other models, achieving the highest predictive power

- Feature importance shows clear business levers (e.g., high charges and frequent support calls signal risk)

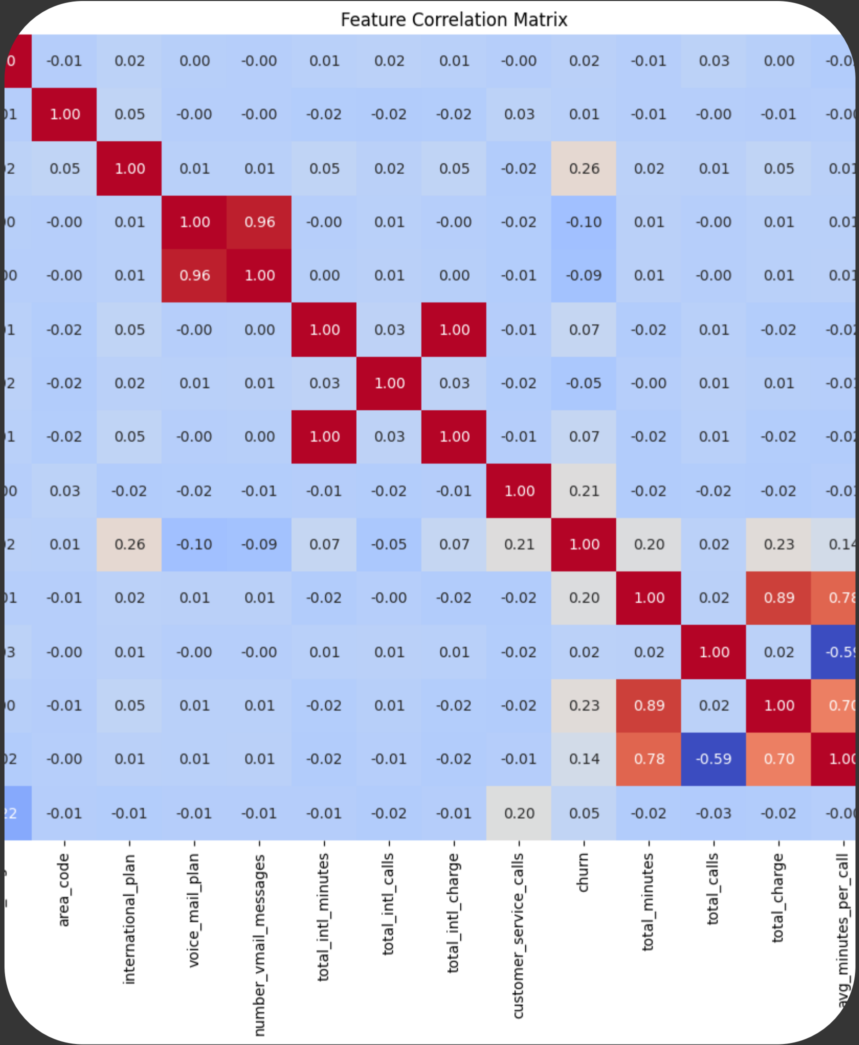
# Results

- Logistic Regression
- - Accuracy: 78% | ROC AUC: 0.84

- XGBoost
- - Accuracy: 94% | Recall: 70% | Precision: 85%

- Random Forest
- - Accuracy: 89% | Key features: total charge, call usage, customer service calls

# Numeric values correlation heatmap



Feature Correlation Matrix

| | area_code | international_plan | voice_mail_plan | number_vmail_messages | total_intl_minutes | total_intl_calls | total_intl_charge | customer_service_calls | churn | total_minutes | total_calls | total_charge | avg_minutes_per_call |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.01 | 0.02 | 0.00 | -0.00 | 0.01 | 0.02 | 0.01 | -0.00 | 0.02 | -0.01 | 0.03 | 0.00 | -0.0 |
| 01 | 1.00 | 0.05 | -0.00 | -0.00 | -0.02 | -0.02 | -0.02 | 0.03 | 0.01 | -0.01 | -0.00 | -0.01 | -0.0 |
| 2 | 0.05 | 1.00 | 0.01 | 0.01 | 0.05 | 0.02 | 0.05 | -0.02 | 0.26 | 0.02 | 0.01 | 0.05 | 0.0 |
| 0 | -0.00 | 0.01 | 1.00 | 0.96 | -0.00 | 0.01 | -0.00 | -0.02 | -0.10 | 0.01 | -0.00 | 0.01 | 0.0 |
| 00 | -0.00 | 0.01 | 0.96 | 1.00 | 0.00 | 0.01 | 0.00 | -0.01 | -0.09 | 0.01 | -0.00 | 0.01 | 0.0 |
| 1 | -0.02 | 0.05 | -0.00 | 0.00 | 1.00 | 0.03 | 1.00 | -0.01 | 0.07 | -0.02 | 0.01 | -0.02 | -0.0 |
| 2 | -0.02 | 0.02 | 0.01 | 0.01 | 0.03 | 1.00 | 0.03 | -0.02 | -0.05 | -0.00 | 0.01 | 0.01 | -0.0 |
| 1 | -0.02 | 0.05 | -0.00 | 0.00 | 1.00 | 0.03 | 1.00 | -0.01 | 0.07 | -0.02 | 0.01 | -0.02 | -0.0 |
| 00 | 0.03 | -0.02 | -0.02 | -0.01 | -0.01 | -0.02 | -0.01 | 1.00 | 0.21 | -0.02 | -0.02 | -0.02 | -0.0 |
| 2 | 0.01 | 0.26 | -0.10 | -0.09 | 0.07 | -0.05 | 0.07 | 0.21 | 1.00 | 0.20 | 0.02 | 0.23 | 0.14 |
| 01 | -0.01 | 0.02 | 0.01 | 0.01 | -0.02 | -0.00 | -0.02 | -0.02 | 0.20 | 1.00 | 0.02 | 0.89 | 0.78 |
| 3 | -0.00 | 0.01 | -0.00 | -0.00 | 0.01 | 0.01 | 0.01 | -0.02 | 0.02 | 0.02 | 1.00 | 0.02 | -0.5 |
| 0 | -0.01 | 0.05 | 0.01 | 0.01 | -0.02 | 0.01 | -0.02 | -0.02 | 0.23 | 0.89 | 0.02 | 1.00 | 0.70 |
| 02 | -0.00 | 0.01 | 0.01 | 0.01 | -0.02 | -0.01 | -0.02 | -0.01 | 0.14 | 0.78 | -0.59 | 0.70 | 1.00 |
| 22 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 | -0.02 | -0.01 | 0.20 | 0.05 | -0.02 | -0.03 | -0.02 | -0.0 |

# Results

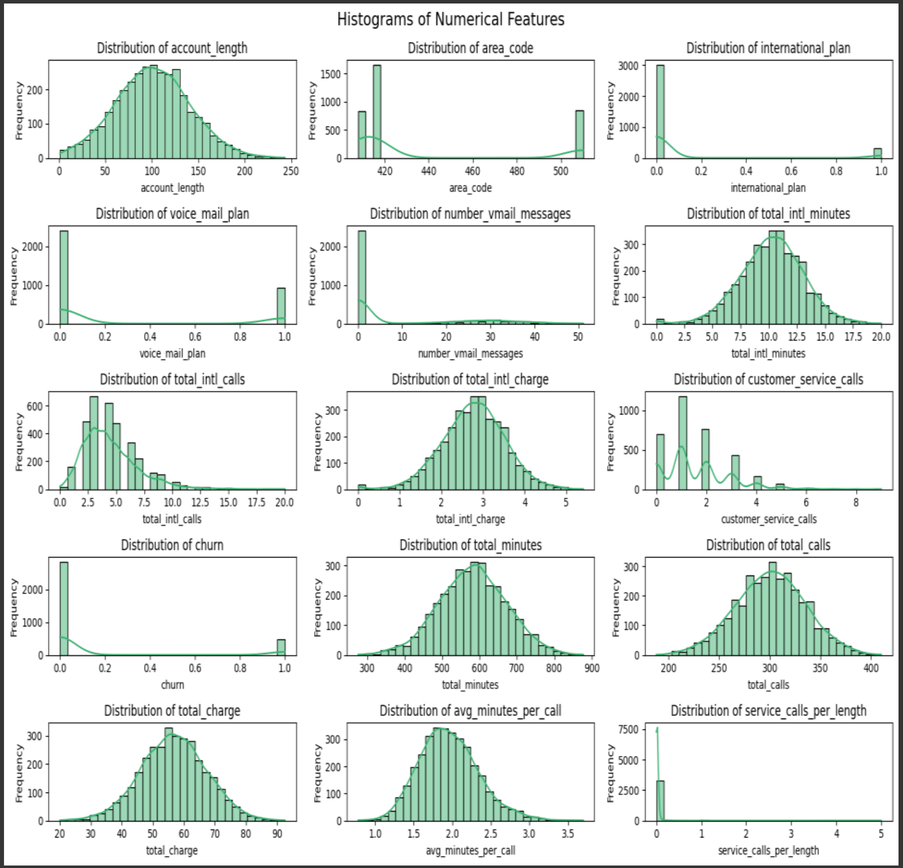
This corellation matrix that shows how different numbers in the dataset are related to each other.

It looks at all the numeric columns, compares them, and shows the results as a heatmap.

Red means a strong positive relationship, blue means a strong negative one, and white means no clear link.

This helps you quickly spot which features are closely connected, which ones might be duplicates, and which ones could be useful for building a model.

# distribution of each numerical feature



Box plots show the spread of each numeric feature, including the median, quartiles, and outliers.

Outliers stand out clearly, helping you spot extreme values that might affect your model

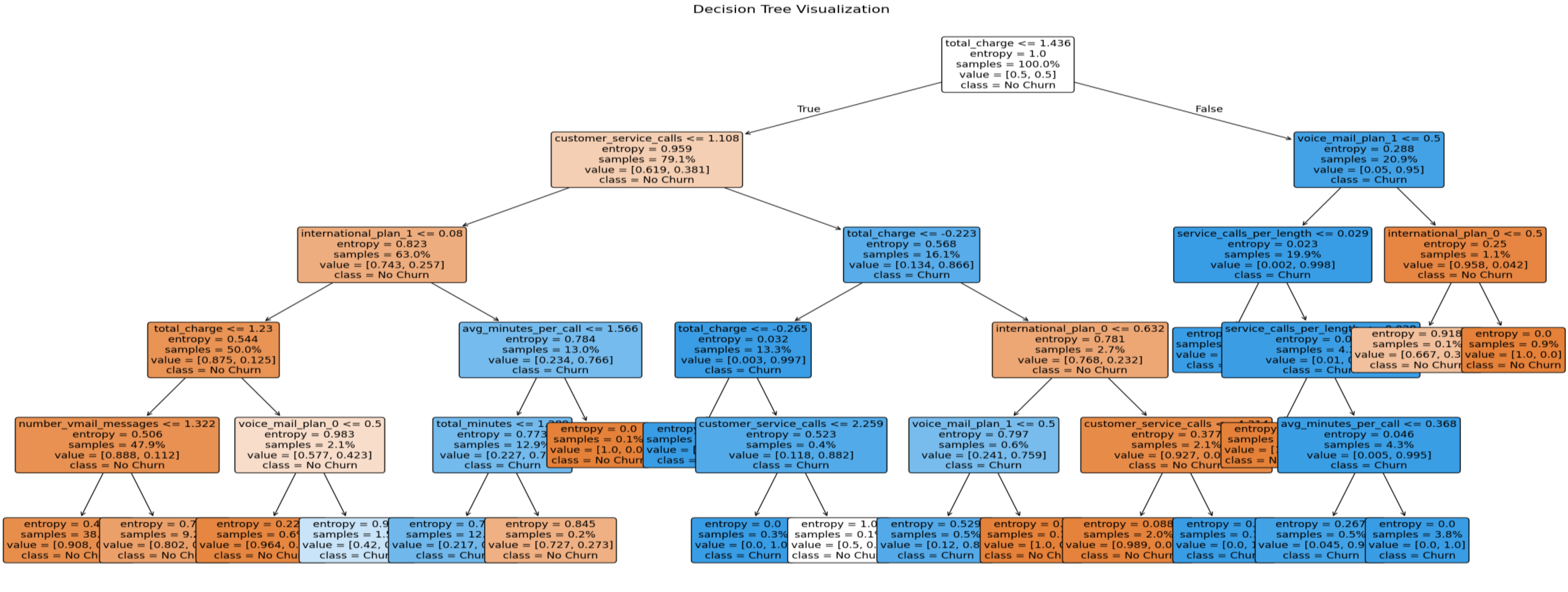
Skewed distributions suggest some features may need transformation.

Features like customer_service_calls and churn may show distinct patterns useful for churn analysis.

If most values are clustered (e.g., many zeros in voice_mail_plan), it hints at class imbalance.

Features with wide ranges (like total_minutes or total_charge) might need scaling for better model performance.

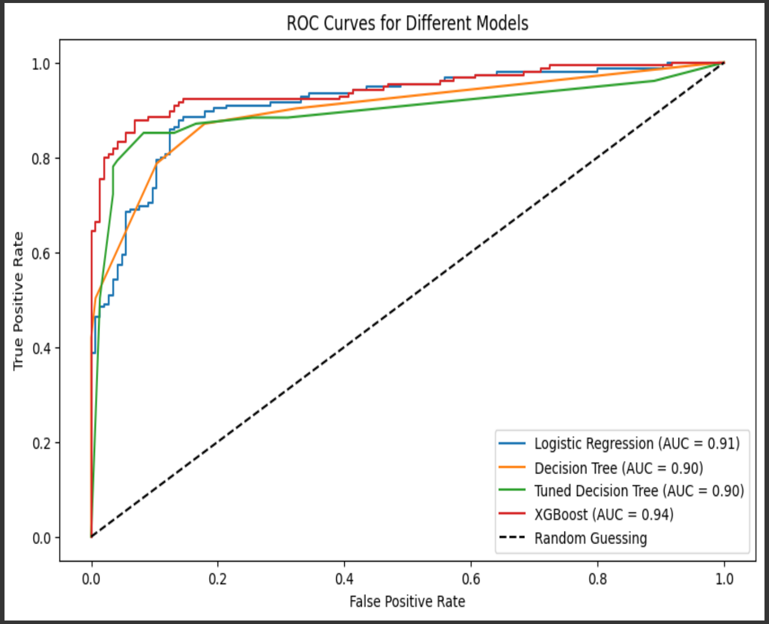
# Decision tree


Decision Tree Visualization

The decision tree shows how customer characteristics combine to predict churn. The most impactful features appear higher in the tree.

Total Charge: This is the initial split, indicating that the overall cost of service is a primary factor. Customers with higher total charges are more likely to be on the 'Churn' side of the tree initially.
Customer Service Calls: For customers on either side of the initial split, the number of customer service calls is a crucial secondary factor. A higher number of calls strongly increases the likelihood of churn, regardless of the total charge.
International Plan: Having an international plan also plays a significant role in predicting churn, especially when combined with other factors.
By following paths down the tree, we can see specific customer segments and their predicted churn probability. For example, customers with high total charges AND multiple customer service calls are very likely to churn. This granular view helps in targeting retention efforts.

# ROC Curve



This plot shows the ROC (Receiver Operating Characteristic) curve for each trained model.

**What it shows:** It plots the True Positive Rate (how many churners were correctly identified) against the False Positive Rate (how many non-churners were incorrectly identified as churners) at various classification thresholds.

**Interpreting the curves:**

A curve that is closer to the top-left corner indicates a better-performing model, as it achieves a higher True Positive Rate for a given False Positive Rate.

The dashed black line represents a random guess (an AUC of 0.5), which is the baseline for performance.

**AUC (Area Under the Curve):** The AUC score is a single number that summarizes the model's overall ability to distinguish between churned and non-churned customers.

An AUC of 1.0 represents a perfect model.

An AUC of 0.5 represents a model that performs no better than random chance.

Higher AUC values indicate better model performance.

**From the plot:** The XGBoost model has the highest AUC (0.94), indicating it is the best at discriminating between churned and non-churned customers among the models tested. The other models also perform reasonably well, with AUCs above 0.7.

# Conclusions

- **Churn is a significant issue:** The initial analysis confirmed that approximately 15% of customers are churning, highlighting the importance of addressing this problem.
- **Key drivers of churn identified:** The feature importance analysis from the models (both Logistic Regression and XGBoost) consistently showed that customer service calls, total charge, and international plan are significant predictors of churn. This aligns with the initial data exploration which also indicated differences in these features between churned and non-churned customers.
- **Model performance:** The XGBoost model generally outperformed the Logistic Regression model, particularly in its ability to identify churners (higher recall) while maintaining reasonable precision. The use of techniques like SMOTE and hyperparameter tuning improved the model's performance on the imbalanced dataset.
- **Threshold matters:** The analysis of precision-recall curves demonstrated the trade-off between identifying more churners (higher recall) and being accurate in those predictions (higher precision). The optimal threshold for prediction depends on the business objective (e.g., minimizing false negatives vs. minimizing false positives).

# RECOMMENDATIONS

**Proactive Customer Service:** Given the strong correlation between customer service calls and churn, SyriaTel should prioritize improving the customer service experience. This could involve:

Training for support staff to handle issues more efficiently and empathetically.

Implementing a system to identify and prioritize customers with a high number of service calls for proactive outreach.

Analyzing the types of issues that lead to multiple service calls to address root causes.

**Review Pricing and Plan Structures:** The importance of "total charge" suggests that pricing and the overall cost of services play a role in churn. SyriaTel should:

Analyze pricing tiers and compare them to competitors.

Consider offering more flexible or cost-effective plans, especially for heavy users or those with international plans.

Clearly communicate billing and usage to customers to avoid unexpected charges.

**Targeted Retention Campaigns:** Using the developed model, SyriaTel can identify customers at high risk of churning. This allows for targeted retention efforts such as:

Offering personalized discounts or incentives to high-risk customers.

Providing tailored service plan recommendations.

Initiating proactive communication to address potential concerns before the customer decides to leave.

**Monitor International Plan Usage:** The "international plan" feature's importance indicates that international usage patterns might be linked to churn. SyriaTel could:

Offer more competitive international calling rates or packages.

Provide alerts or notifications to customers about their international usage and associated costs.

**Continuous Model Monitoring and Improvement:** Customer behavior and market conditions can change over time. It is crucial to:

Regularly monitor the model's performance on new data.

Retrain the model periodically with updated data.

Explore additional features or more advanced modeling techniques to further improve prediction accuracy.

By implementing these recommendations, SyriaTel can leverage the insights from this churn prediction model to reduce customer attrition,