Syria Tel Non_Technical Data Report

Project Overview

Customer churn, a significant challenge for telecommunications companies, refers to the number of customers who discontinue their services. For SyriaTel, this involves customers terminating their subscriptions and switching to competitors. This trend directly impacts revenue, customer acquisition costs, and overall market competitiveness.

This project aims to develop a predictive model that identifies customers at high risk of churn. The model is designed as a binary classification task, estimating the likelihood of a customer leaving. By analyzing historical customer behavior data, SyriaTel will gain actionable insights to proactively address churn, improve customer satisfaction, and implement targeted retention strategies.

Business Understanding

Customer retention is a cornerstone of sustainable business growth in the telecommunications industry. Losing customers leads to revenue erosion, increased marketing expenses to acquire new customers, and potential reputational harm.

The goal is to predict churn and identify key drivers, enabling SyriaTel to implement proactive measures that reduce churn rates, optimize resource allocation, and enhance customer satisfaction.

Expected Outcomes:

- **Increased Retention**: Prioritizing interventions for high-risk customers.
- **Efficient Resource Allocation**: Focusing efforts on impactful areas.
- **Improved Customer Experience**: Addressing dissatisfaction to build loyalty.

This initiative positions SyriaTel as a leader in using data analytics for customer retention.

Problem Statement

SyriaTel faces high customer churn, leading to revenue loss and inefficiencies. An inability to accurately predict churn hinders effective intervention. A predictive model is essential to identify at-risk customers, uncover churn drivers, and inform retention strategies.

Objective

The primary objective of this project is to **develop a predictive model that accurately identifies customers likely to churn** and provides insights into the factors driving their decisions. This enables proactive measures to:

1. **Mitigate churn** and protect revenue.
2. **Enhance customer loyalty** through targeted retention strategies.
3. **Optimize resource allocation** for maximum impact on retention efforts.

Ancillary Objectives:

- Analyze customer behavior patterns to inform strategy development.
- Evaluate the impact of customer service interactions on churn.

Metrics of Success

To evaluate the model's effectiveness, the following metrics are used:

1. **Accuracy**: The proportion of correct predictions (both churned and non-churned customers) out of total predictions. A target range of 80–90% indicates the model's ability to generalize well.
2. **Precision**: The proportion of correctly predicted churners out of all customers predicted to churn. A range of 70–90% ensures minimal false positives, critical for focusing retention efforts.
3. **Recall (Sensitivity)**: The proportion of actual churners correctly identified by the model. A range of 60–85% highlights the model's ability to minimize false negatives, ensuring most at-risk customers are identified.
4. **F1-Score**: The harmonic mean of precision and recall. It provides a balanced measure of model performance, particularly useful for imbalanced datasets. Target: 65–87%.
5. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Measures the model's ability to distinguish between churners and non-churners. AUC values range from 0.5 (random guessing) to 1 (perfect prediction). A target AUC of 0.85 or higher indicates strong model discrimination.

These metrics ensure the model is robust, reliable, and effective in predicting churn while minimizing misclassifications.

Data Understanding

The dataset, sourced from Kaggle, contains **3,333 customer records** and **21 features** describing customer demographics, subscription plans, usage patterns, and customer service interactions.

Feature/ Column Descriptions

- State: The U.S. state where the customer resides.

- Account Length: Duration (in days) of the customer's account with the company.
- Area Code: The customer's area code.
- Phone Number: The customer's phone number.
- International Plan: Indicates whether the customer has subscribed to the international calling plan (True/False).
- Voice Mail Plan: Indicates whether the customer has subscribed to the voicemail plan (True/False).
- Number of Voicemail Messages: The total number of voicemail messages sent by the customer.
- Total Day Minutes: Total duration of calls made by the customer during the daytime (in minutes).
- Total Day Calls: Total number of daytime calls made by the customer.
- Total Day Charge: Total charges incurred by the customer for daytime calls.
- Total Evening Minutes: Total duration of calls made by the customer during the evening (in minutes).
- Total Evening Calls: Total number of evening calls made by the customer.
- Total Evening Charge: Total charges incurred by the customer for evening calls.
- Total Night Minutes: Total duration of calls made by the customer during the night (in minutes).
- Total Night Calls: Total number of nighttime calls made by the customer.
- Total Night Charge: Total charges incurred by the customer for nighttime calls.
- Total International Minutes: Total duration of international calls made by the customer (in minutes).
- Total International Calls: Total number of international calls made by the customer.
- Total International Charge: Total charges incurred by the customer for international calls.
- Customer Service Calls: Total number of calls made by the customer to customer service.
- Churn: Indicates whether the customer has terminated their contract (True/False). This is our target feature, the primary focus of this analysis. Predicting churn is essential for understanding customer attrition and implementing effective retention strategies.

Load the data

Importing essential libraries for numerical operations (NumPy), data manipulation (Pandas), data visualization (Matplotlib and Seaborn), and machine learning (Scikit-learn). Additionally, configuring notebook aesthetics ensures that visualizations are clear and easily interpretable.

Data Preparation

I conducted preliminary data understanding by creating a class function in separate python file, then imported it. It loads the dataset, views its contents, the dimension and a concise statistical summary.

Then I proceeded to check for missing values and duplicates and found that it didn't have either.

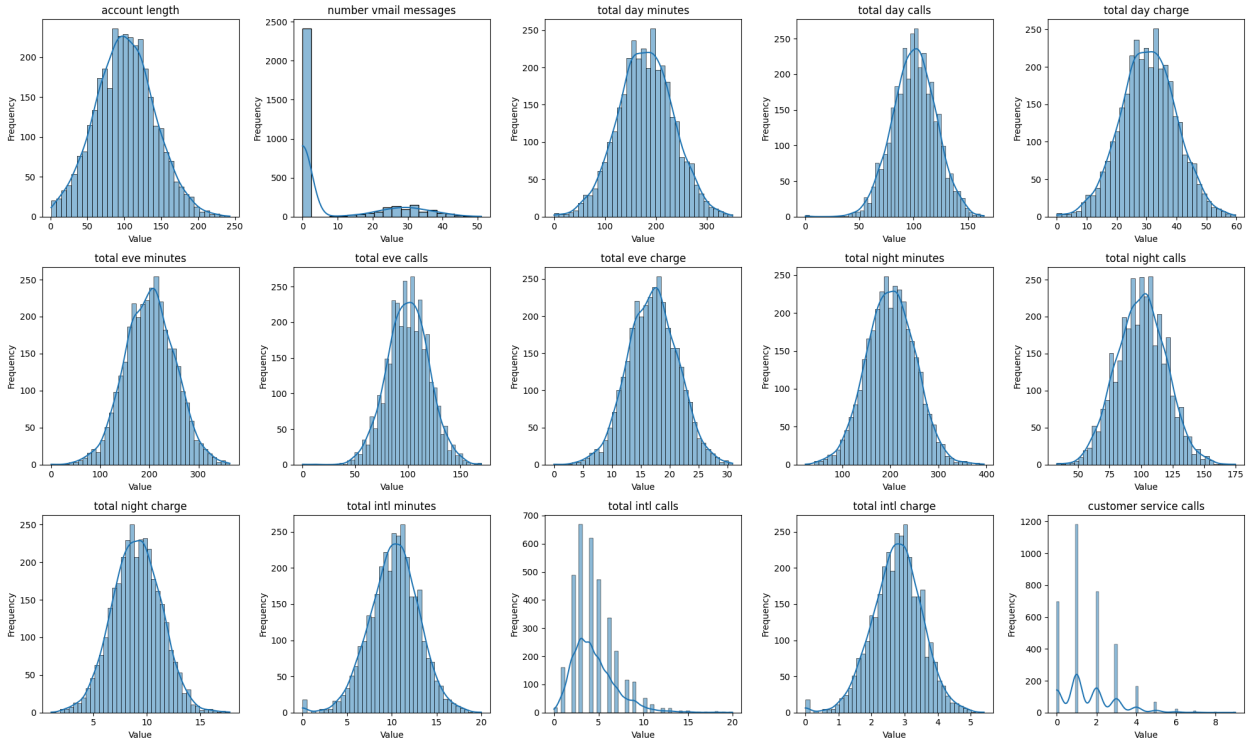
Exploratory Data Analysis (EDA)

This is the process of analyzing and visualizing datasets to uncover patterns, relationships, and anomalies. It aims to understand data structure, and guide preprocessing or modeling decisions.

So here I started by identifying the categorical variables and numerical variables to determine the appropriate statistical techniques and visualizations to use for analyzing each type of data.
At this point I realized that state and telephone columns were irrelevant in my analysis so I drop them.

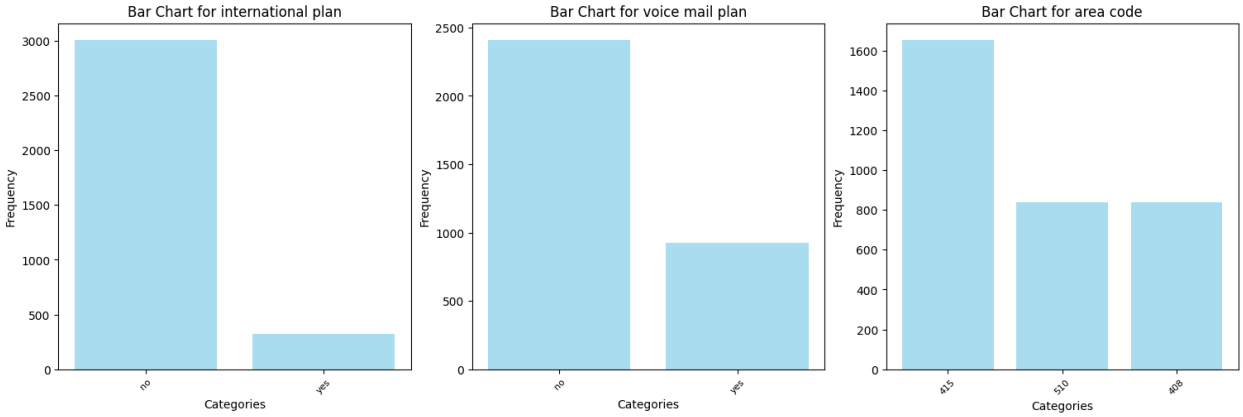
Findings:

1. **Univariate Analysis**:

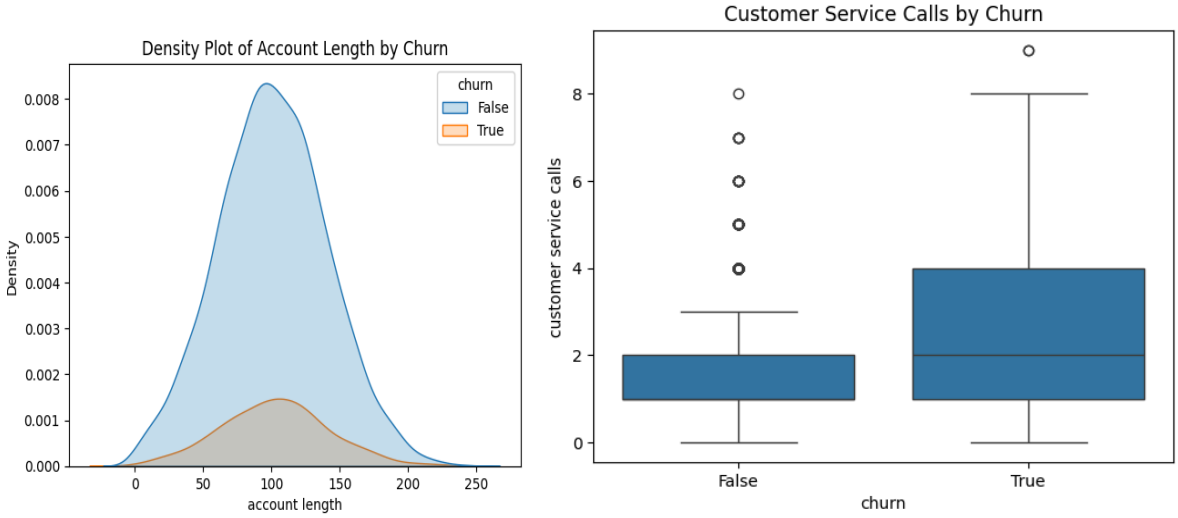


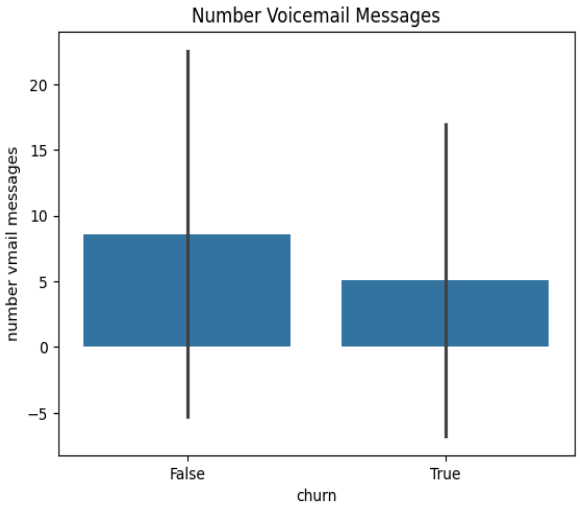
The distribution plots indicate that all features, except for customer service calls, and the number of voicemail messages, follow a normal distribution. The total number of international calls exhibits a slight right skew but remains approximately normal. The number of voicemail messages displays a pronounced peak on the right, suggesting the presence of outliers. Customer

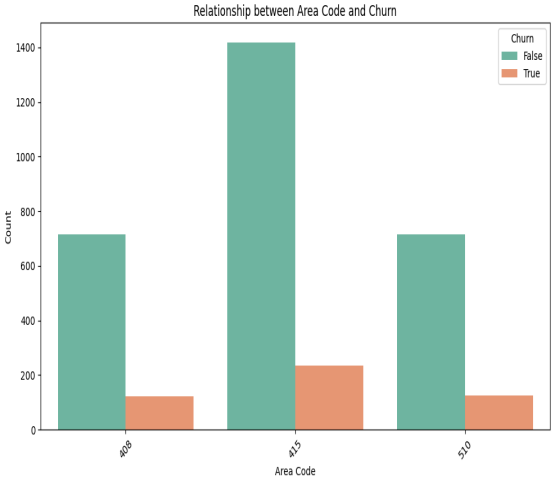
service calls exhibit multiple peaks, indicating a multimodal pattern, consistent with its nature as an integer rather than a continuous variable.



2. Bivariate Analysis







Customers with a voicemail plan exhibit lower churn rates, suggesting it enhances retention. However, most non-churning customers lack a voicemail plan, indicating its appeal to a specific segment rather than the entire customer base.

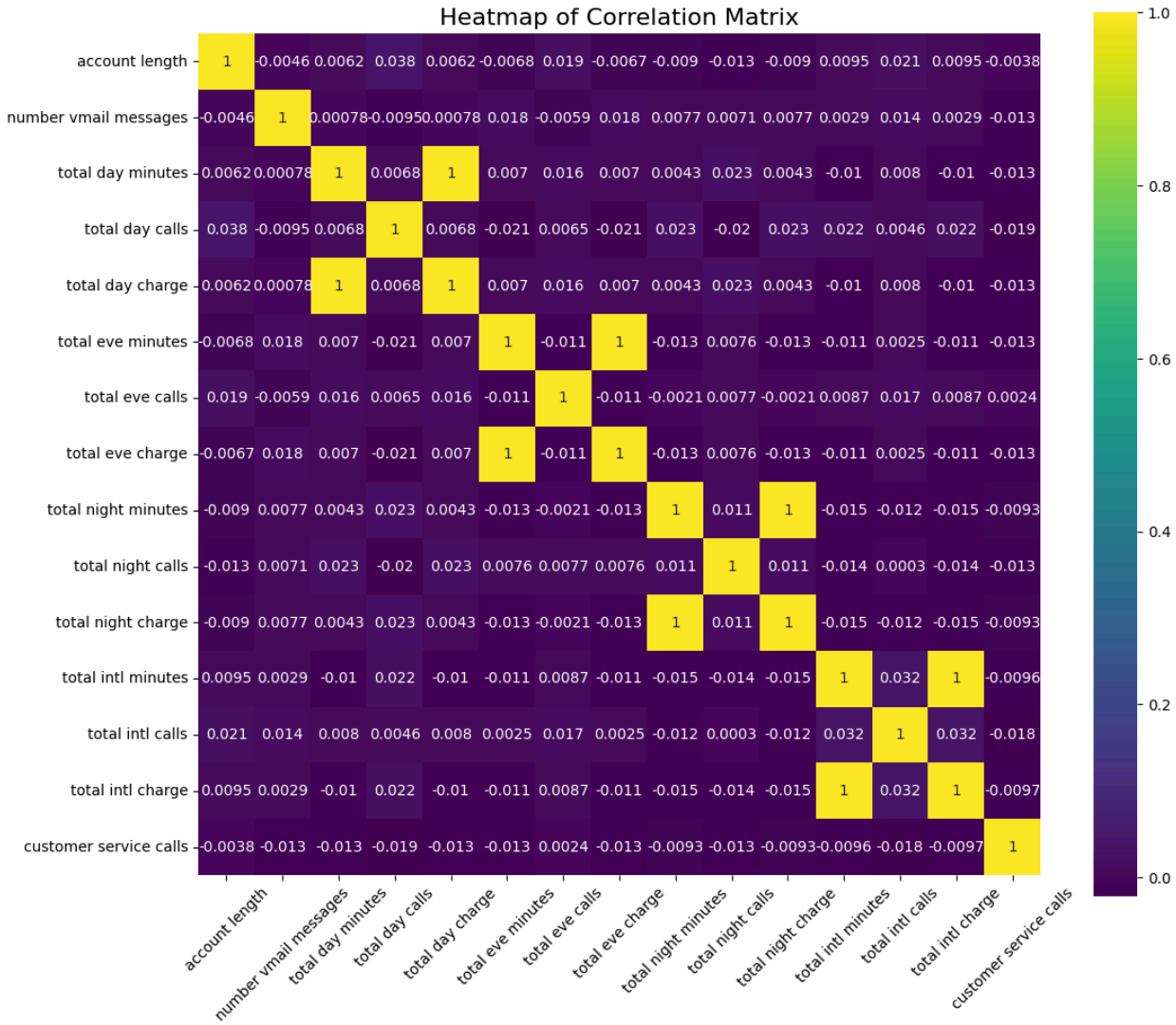
Area codes 408 and 510 show similar churn rates and active customer numbers for Syria Tel, while area code 415 has a larger customer base with slightly higher churn, indicating greater overall activity in that region.

The density plot reveals substantial overlap and similar central tendencies between churn groups, suggesting account length is not a strong churn predictor.

Non-churned customers show higher voicemail usage, while churned customers exhibit lower activity, indicating reduced engagement among those at risk of leaving.

Analysis of customer service calls shows a clear link between increased call frequency (more than four calls) and a higher likelihood of churn.

3. **Multivariate Analysis**:



- Strong correlations between call durations and charges (e.g., day minutes vs. day charges) confirmed consistent billing patterns.
- Customer service interactions showed a clear relationship with churn likelihood, highlighting dissatisfaction.

Data Preparation

1. **Categorical Variable Conversion**: I used one-hot encoding (OHE) for categorical variables, dropping the first category to avoid the dummy variable trap, ensuring compatibility with the model.

2. **Churn Column Mapping**: I mapped the churn column to retain a single, clear column, simplifying interpretation while maintaining its binary nature for the model.
3. **Advantages of One-Hot Encoding**: OHE prevents ordinal misinterpretation, preserves categorical integrity, and enables efficient model processing without introducing biases.
4. **Handling Outliers with Clipping**: I used clipping to handle outliers, capping values within IQR-based bounds to maintain data integrity without loss:
    - **Preserves Data**: Retains all rows, avoiding data loss or bias.
    - **Balances Influence**: Reduces outlier impact while keeping feature relationships intact.
    - **Interpretable**: Maintains the original scale without complex transformations.
    - **Efficient**: Fast and reproducible using IQR thresholds, ensuring data integrity without distorting insights.

Model Development and Evaluation

Several models were developed and compared based on the defined metrics:

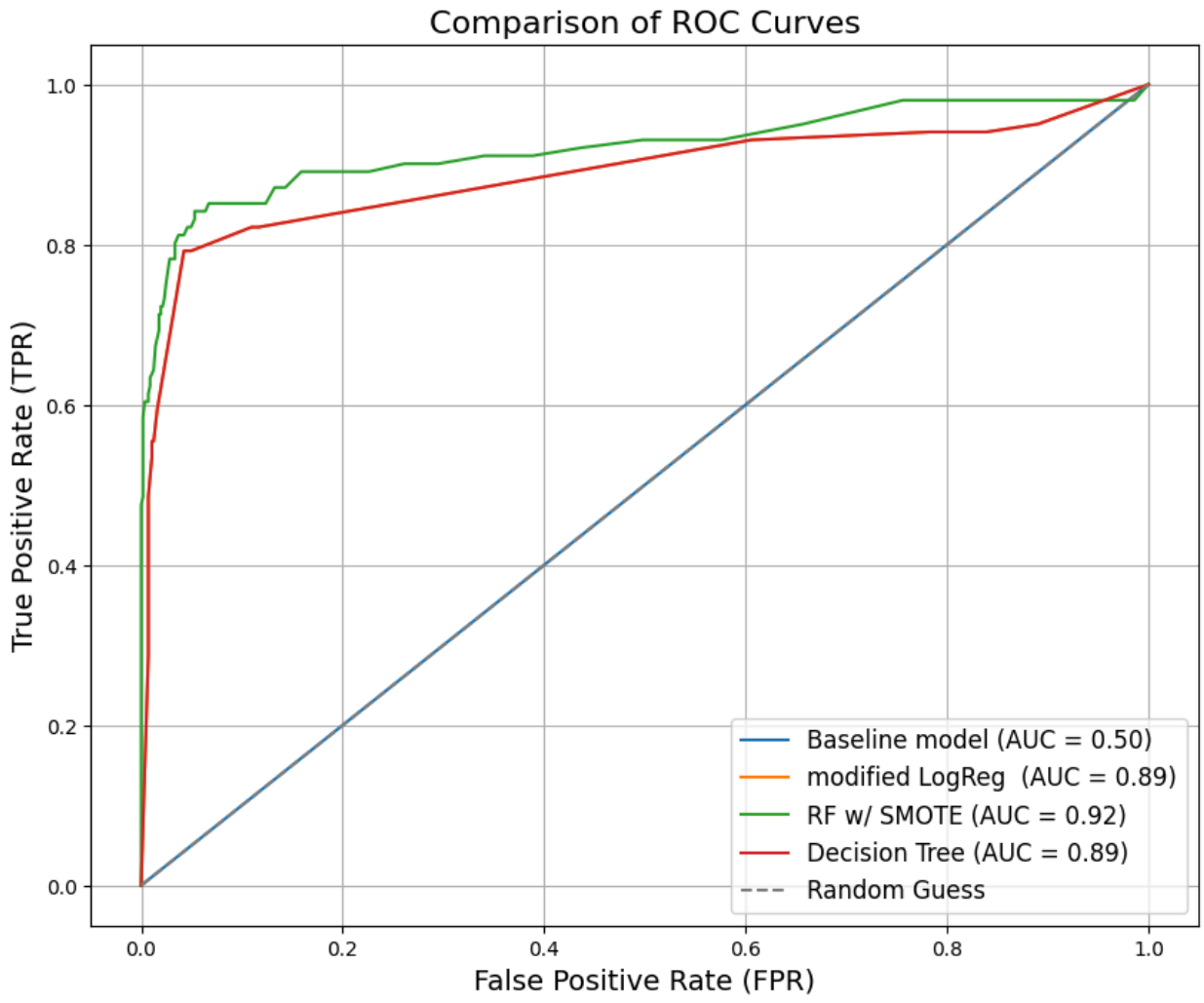
1. **Baseline Logistic Regression**:
    - **Accuracy**: 85%.
    - **Class 0 (non-churn)**: High recall (0.99) but low precision (0.11) for churned customers (Class 1), indicating poor performance due to class imbalance.
2. **Improved Logistic Regression**:
    - Balanced dataset using SMOTE and scaled features.
    - Regularization and optimized parameters improved model performance.
    - **Results**:
        - Precision, Recall, and F1-scores: 0.76.
        - ROC-AUC: 0.82, demonstrating strong discrimination between churners and non-churners.
3. **Random Forest**:
    - **AUC**: 0.92, the best performance among all models.
    - Handled non-linear relationships and feature interactions effectively.
4. **Decision Tree**:
    - **AUC**: 0.89, offering robust performance with higher interpretability.

Recommendations

1. **Customer Service Optimization**:
    - Address dissatisfaction among customers making more than four calls.
    - Invest in training to improve first-call resolution rates.
2. **Retention Campaigns**:

- Promote voicemail and international plans to high-risk customers, as these correlate with lower churn rates.
- Focus on Area Code 415, which has slightly higher churn rates.

3. **Model Deployment**:
   - Deploy the Random Forest model for real-time churn prediction.
   - Regularly retrain the model to adapt to evolving customer behavior.
4. **Monitoring and Insights**:
   - Continuously analyze churn patterns to refine strategies and improve customer satisfaction.



Conclusion

This project demonstrates the critical role of predictive analytics in mitigating churn. Among the evaluated models, the Random Forest model emerged as the most effective, achieving an AUC of 0.92, indicating excellent predictive power.

By implementing these insights and deploying the model, SyriaTel can significantly reduce churn, optimize resource allocation, and enhance its competitive standing. Continuous monitoring and periodic retraining will ensure sustained success.

References

- Productplan.com
- Kaggle
- Worldfolio
- Research on customer churn prediction in telecommunications