Customer Churn Prediction for SyriaTel

Business understanding

- SyriaTel, like many telecom providers, operates in a highly competitive market where customer loyalty is crucial. Retaining existing subscribers is significantly more cost-effective than acquiring new ones—making churn a serious business concern.

- Churn, or the loss of customers, directly reduces revenue and threatens long-term profitability. However, churn rarely happens at random; it's often triggered by issues like poor service quality, pricing dissatisfaction, inadequate customer support, or more attractive competitor offers.

- By predicting which customers are most likely to leave, SyriaTel can take proactive, data-driven steps—such as personalized offers or improved service—to reduce churn and strengthen customer retention.

Business Goals

i) **Build a predictive model** to identify customers likely to churn

ii) Gain insights into key factors driving churn

iii) Provide strategic recommendations to reduce churn and improve retention

Data Understanding

- The project used the "Churn in Telecoms" dataset from Kaggle.
- Dataset contains 3,333 entries and 21 columns.
- There were no missing values hence no need of further cleaning.
- Dataset included both categorical and numerical variables.
- Key categorical features: state, international plan, voice mail plan, churn.
- Key numerical features: total day minutes, customer service calls, total day charge etc

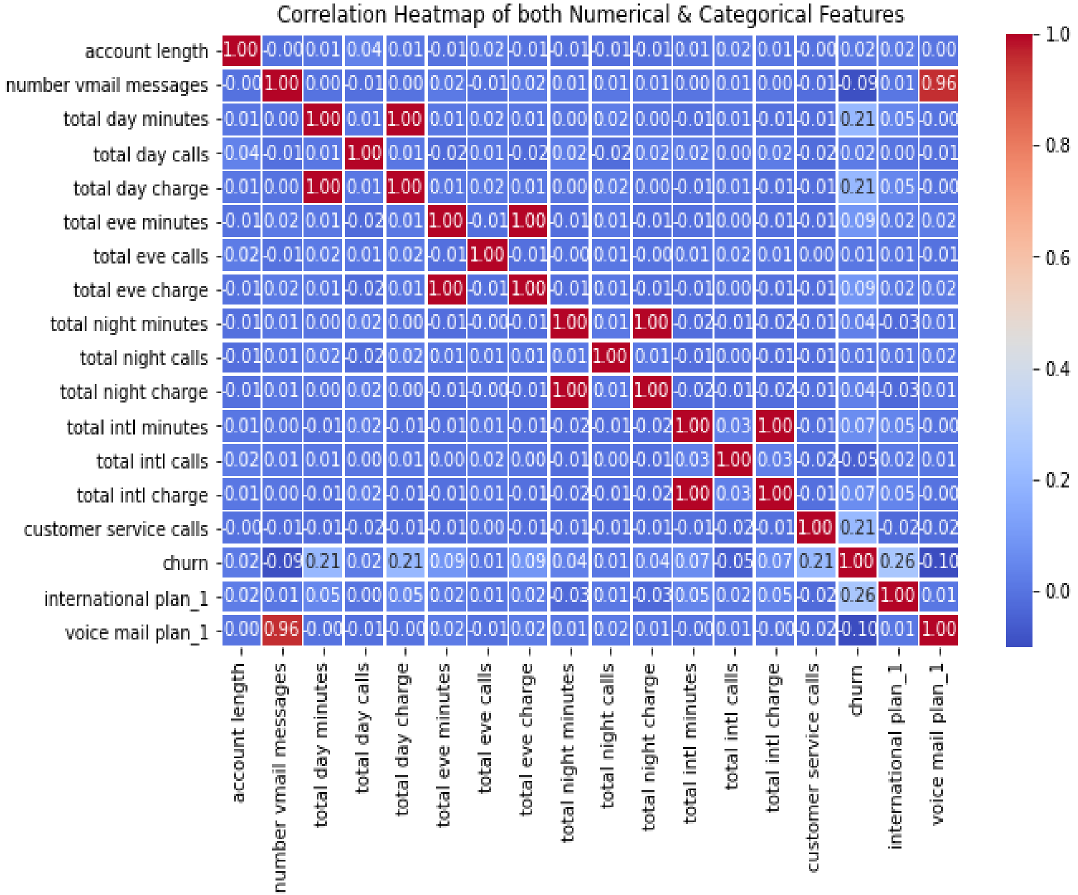
Steps followed in Exploratory Data Analysis

- Steps followed in exploratory data analysis
- Dropped low-value columns like phone number.
- Performed descriptive analysis on numerical features.
- Checked for class imbalance in churn variable.
- Applied one-hot encoding to categorical variables international plan &voice mail plan.
- Converted all categorical data to numeric format.
- Conducted correlation analysis to explore relationships and multicollinearity.

Correlation matrix

From the heatmap, here are the key findings:

- **International_plan_Yes** has the highest positive correlation with churn (0.26), suggesting those on the plan are more likely to churn (though still a weak correlation).
- **Customer service calls**, **total day charge**, and **total day minutes** follow with ~0.21 correlation.
- All other features have very weak correlations (-0.02 to 0.09).
- This indicates **no strong linear relationships** and **low risk of multicollinearity**.



Correlation Heatmap of both Numerical & Categorical Features

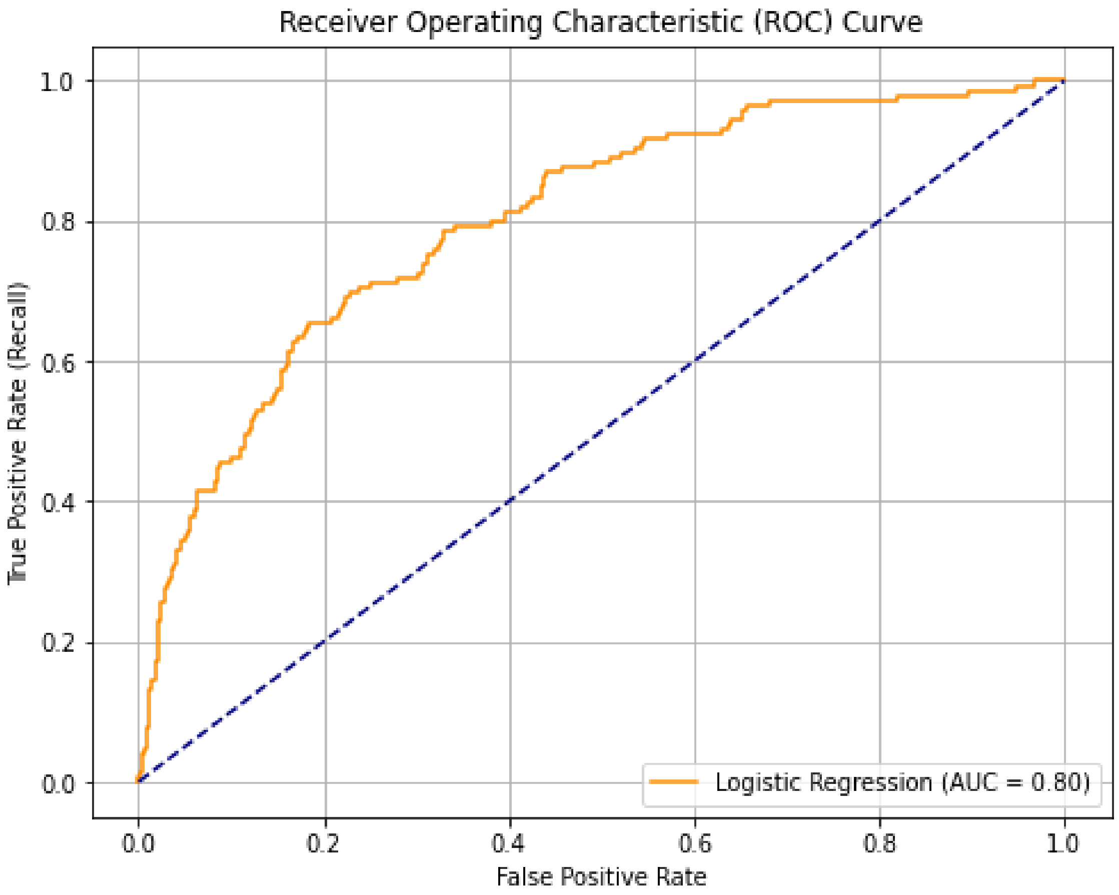
	account length	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn	international plan_1	voice mail plan_1
account length	1.00	-0.00	0.01	0.04	0.01	-0.01	0.02	-0.01	-0.01	-0.01	-0.01	0.01	0.02	0.01	-0.00	0.02	0.02	0.00
number vmail messages	-0.00	1.00	0.00	-0.01	0.00	0.02	-0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.00	-0.01	-0.09	0.01	0.96
total day minutes	0.01	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01	0.21	0.05	-0.00
total day calls	0.04	-0.01	0.01	1.00	0.01	-0.02	0.01	-0.02	0.02	-0.02	0.02	0.02	0.00	0.02	-0.02	0.02	0.00	-0.01
total day charge	0.01	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01	0.21	0.05	-0.00
total eve minutes	-0.01	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.09	0.02	0.02
total eve calls	0.02	-0.01	0.02	0.01	0.02	-0.01	1.00	-0.01	-0.00	0.01	-0.00	0.01	0.02	0.01	0.00	0.01	0.01	-0.01
total eve charge	-0.01	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.09	0.02	0.02
total night minutes	-0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01	0.04	-0.03	0.01
total night calls	-0.01	0.01	0.02	-0.02	0.02	0.01	0.01	0.01	0.01	1.00	0.01	-0.01	0.00	-0.01	-0.01	0.01	0.01	0.02
total night charge	-0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01	0.04	-0.03	0.01
total intl minutes	0.01	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01	0.07	0.05	-0.00
total intl calls	0.02	0.01	0.01	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	-0.01	0.03	1.00	0.03	-0.02	-0.05	0.02	0.01
total intl charge	0.01	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01	0.07	0.05	-0.00
customer service calls	-0.00	-0.01	-0.01	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	1.00	0.21	-0.02	-0.02
churn	0.02	-0.09	0.21	0.02	0.21	0.09	0.01	0.09	0.04	0.01	0.04	0.07	-0.05	0.07	0.21	1.00	0.26	-0.10
international plan_1	0.02	0.01	0.05	0.00	0.05	0.02	0.01	0.02	-0.03	0.01	-0.03	0.05	0.02	0.05	-0.02	0.26	1.00	0.01
voice mail plan_1	0.00	0.96	-0.00	-0.01	-0.00	0.02	-0.01	0.02	0.01	0.02	0.01	-0.00	0.01	-0.00	-0.02	-0.10	0.01	1.00

Modelling

- **Train-test split** was used to separate data for learning and performance evaluation.
- **Standard scaling** was then applied to numerical features in the training set.
- **SMOTE** method was used to balance class distribution (50% churn / 50% no churn).
- **Logistic Regression** is then trained on resampled data and evaluated using:
  - Accuracy, Precision, Recall, Confusion Matrix, ROC Curve.
- **Random Forest** and **Decision Tree** models also trained and evaluated similarly.
- **Hyperparameter tuning** applied to RF and DT to improve performance.

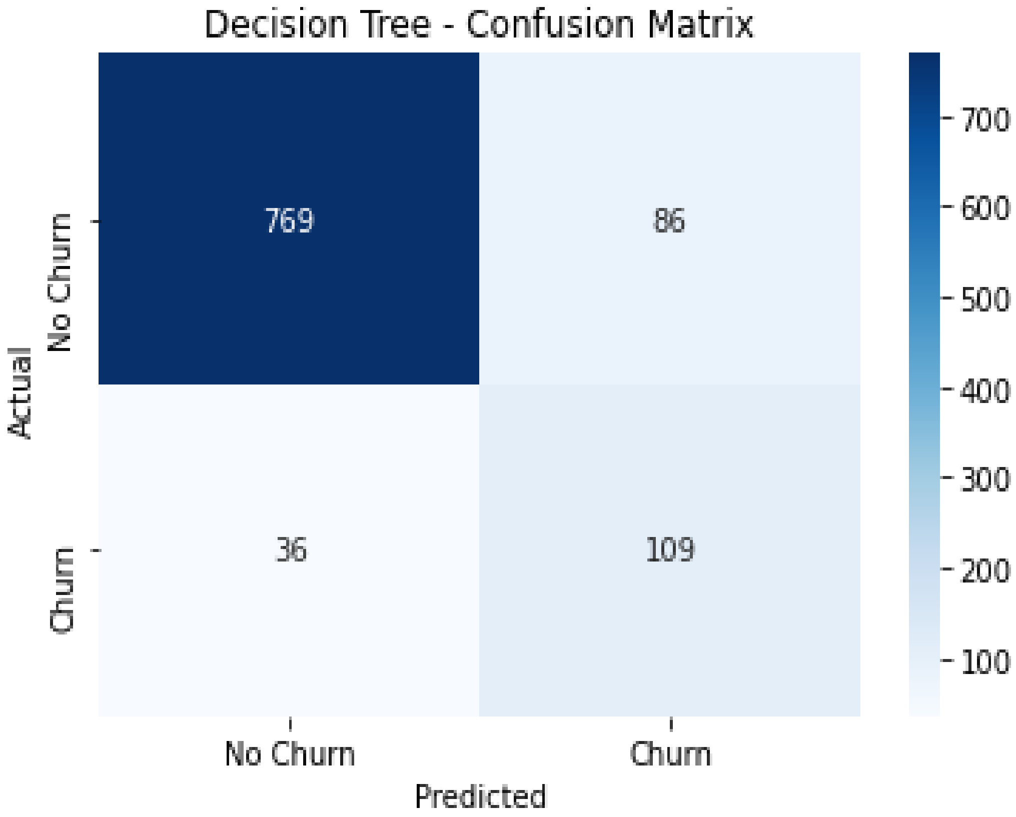
Evaluation-Logistic Model

- The Logistic Regression model achieved an accuracy of 0.72 with a precision of 0.30, recall of 0.71, and F1 score of 0.42.
- The confusion matrix revealed 103 true positives, 42 false negatives, 239 false positives, and 616 true negatives.
- In terms of cross-validation, the model achieved an average accuracy of 75% across 5 folds, indicating a reasonably good ability to distinguish between churned and non-churned customers.
- The AUC score of 0.8015 suggests that the model has an 80.15% chance of correctly ranking a randomly chosen churned customer above a non-churned one, highlighting decent discriminative power despite its lower precision.



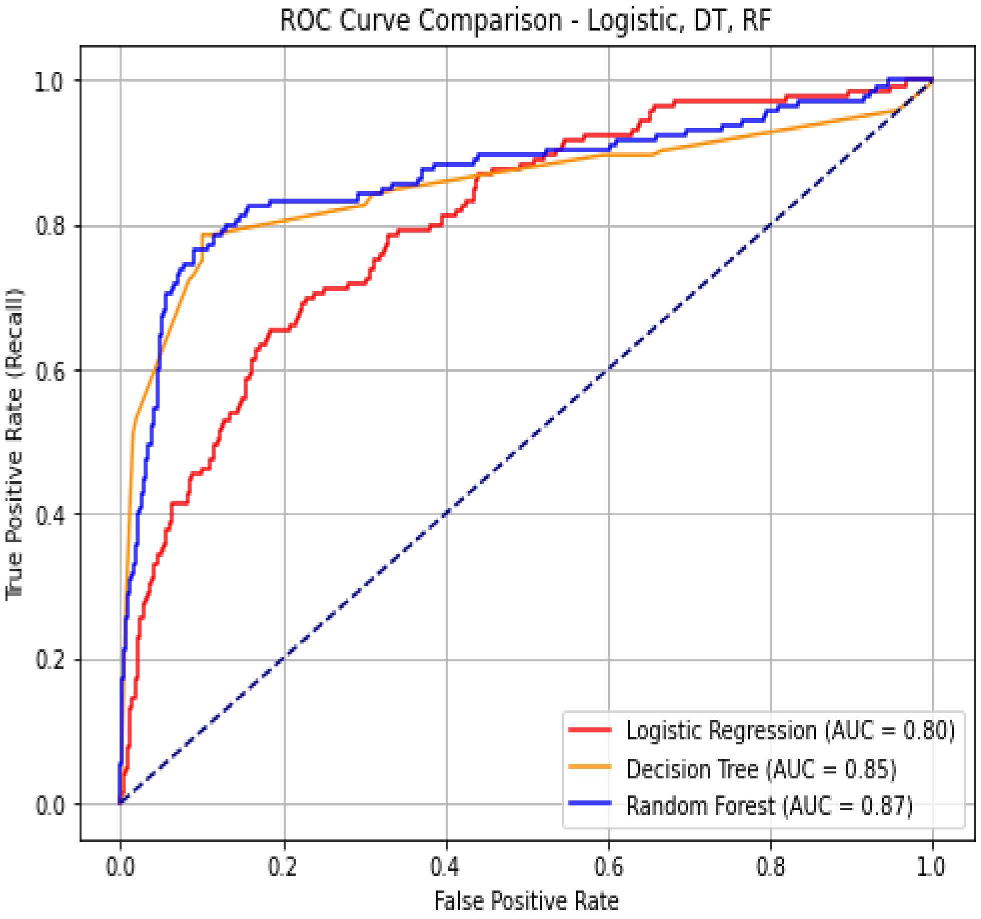
Evaluation – Decision Trees

- The **Decision Tree** model achieved a strong accuracy of 0.88, indicating it correctly classified 88% of the customers.
- It recorded a **precision of 0.56** and a **recall of 0.75**, showing it was fairly effective at identifying churned customers while maintaining a moderate rate of false positives.
- The **confusion matrix** revealed 769 true negatives, 86 false positives, 36 false negatives, and 109 true positives, demonstrating the model's strength in correctly identifying both churned and non-churned customers.
- With a **cross-validation accuracy of 0.86,** the model showed consistent performance across different data folds, reinforcing its reliability.
- Overall, the Decision Tree offered balanced performance with better recall and accuracy than the logistic model, making it a strong candidate for churn prediction.



Evaluation- Random Forest

- The **Random Forest** model delivered the best overall performance, achieving an **accuracy of 0.90**, with **precision of 0.64** and **recall of 0.73**.
- It correctly identified most non-churned customers with **795 true negatives** and detected **106 true churn cases**, while keeping **false positives (60)** and **false negatives (39)** relatively low.
- The model demonstrated strong generalization, with a **cross-validation accuracy of 0.87**, consistently classifying 87% of customers correctly across different splits.
- This makes Random Forest the most reliable and balanced model for predicting churn in this study.

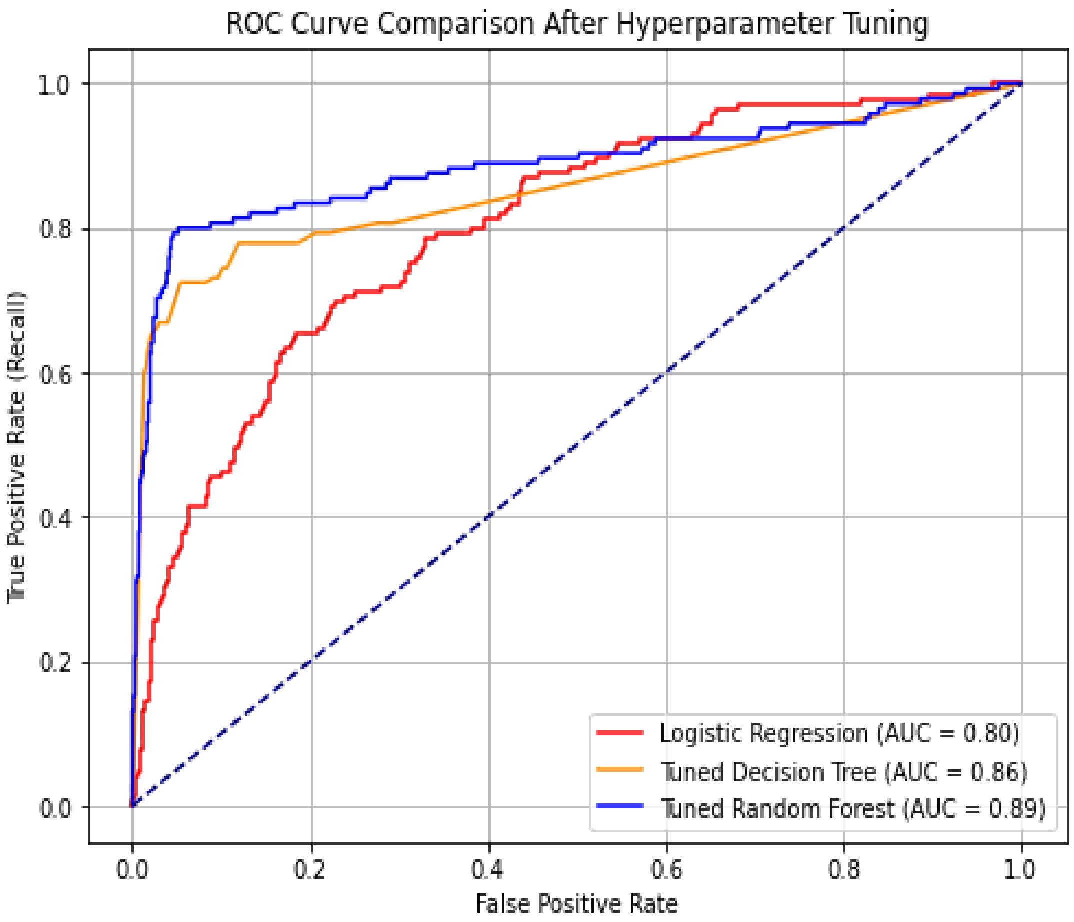


Hyperparameter Tuning

- **After hyperparameter tuning**, both models showed notable improvement. The **Decision Tree** reached an **accuracy of 90%**, up from ~84%, with more consistent results across folds (±0.01 std). However, it remains somewhat prone to overfitting on complex patterns.

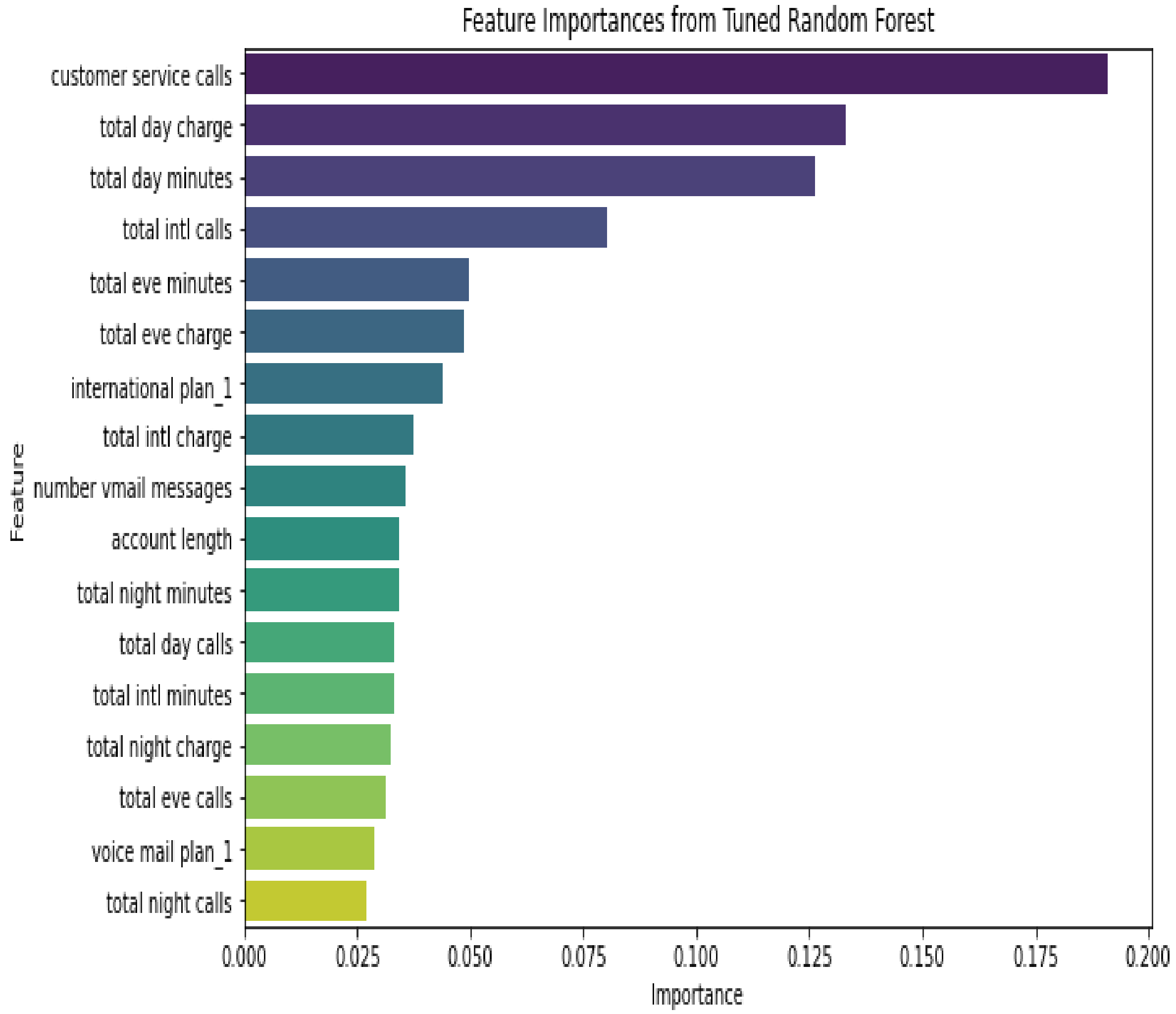
- The **Random Forest** model achieved the **highest accuracy of 95%**, with excellent stability (±0.01 std). Tuning significantly enhanced its generalization and reduced overfitting, thanks to the ensemble approach.

- In terms of **ROC scores**, the **tuned Random Forest** led with an **AUC of 0.89**, followed by the **tuned Decision Tree** at **0.86**. **Logistic Regression** remained a solid baseline with an **AUC of 0.80**, though it lagged behind the other models



Feature Selection

- **Feature selection** was performed using the best-performing model — Random Forest — to identify the key drivers of customer churn. This approach leveraged the model's ability to rank features based on their importance in predicting churn, ensuring more reliable insights.
- **Customer service calls** emerged as the most important feature for predicting churn, suggesting that frequent contact with support may indicate dissatisfaction or unresolved issues.
- This is closely followed by **total day charge** and **total day minutes**, implying that higher daytime usage and related costs could signal a higher likelihood of churn.



Recommendations & Next Steps

- **Enhance Customer Support:** Service calls are a top churn driver—improve satisfaction through feedback surveys, sentiment analysis, and agent training.
- **Engage High-Usage Customers:** Customers with high daytime usage and charges are at higher risk. Offer personalized plans, loyalty rewards, and proactive retention strategies.
- **Embed Insights into Strategy:** Incorporate churn drivers into regular business reviews and continuously retrain models to adapt to evolving customer behavior.

- **Embed Insights into Strategy:** Incorporate churn drivers into regular business reviews and continuously retrain models to adapt to evolving customer behavior

Next Steps

- **Model Deployment:** Integrate the Random Forest model into SyriaTel's customer management system for real-time churn prediction.
- **Monitor Model Performance:** Track prediction accuracy and update the model regularly with new data to maintain relevance.
- **Operationalize Interventions:** Develop automated workflows for retention offers, triggered by churn risk scores.
- **Cross-Functional Collaboration:** Work with customer service, marketing, and product teams to act on churn insights.

THANK YOU