**SyriaTel DATA REPORT**

****

**Background Information**

**Company Overview**

SyriaTel is a telecom provider that offers its clients broadband and mobile services. Retaining clients is essential to sustaining income and profitability for many companies in the telecom sector. Since it is frequently more expensive to acquire new customers than to keep current ones, high customer churn rates can have a substantial negative effect on the bottom line of the business.

**Problem Context**One of the biggest problems facing the telecom sector is customer attrition. When consumers discontinue doing business with a company—either by terminating subscriptions or moving to rivals—this is known as churn. Dissatisfaction with goods, services, or prices is sometimes indicated by high churn rates. By using data science to identify which customers are most likely to leave and to put targeted retention tactics in place, SyriaTel aims to proactively solve this problem**.**

The primary goal of this project is to build a predictive classifier that can identify customers who are likely to stop using SyriaTel's services. The insights gained from this model will enable the company to intervene proactively, offering personalized solutions to retain customers before they churn**.**

**Data Understanding**

**Source of the Data**

* **Data Source: The data used on SyriaTel was sourced from Kaggle.**

**Columns and Their Meanings**

**Here is an updated list of the key columns in the dataset, including their data types and descriptions:**

1. **state**
   * **Data Type: Categorical (String)**
   * **Description: Represents the state or region of the customer.**
2. **account length**
   * **Data Type: Numeric (Integer)**
   * **Description: The length of time the customer has been with SyriaTel (in months).**
3. **area code**
   * **Data Type: Numeric (Integer)**
   * **Description: The area code of the customer's phone number.**
4. **phone number**
   * **Data Type: Categorical (String)**
   * **Description: The customer’s phone number (used for identification).**
5. **international plan**
   * **Data Type: Categorical (String)**
   * **Description: Indicates whether the customer has an international calling plan.**
6. **voice mail plan**
   * **Data Type: Categorical (String)**
   * **Description: Indicates whether the customer has a voicemail plan.**
7. **number vmail messages**
   * **Data Type: Numeric (Integer)**
   * **Description: The number of voicemail messages left by the customer.**
8. **total day minutes**
   * **Data Type: Numeric (Float)**
   * **Description: Total minutes spent on calls during the day.**
9. **total day calls**
   * **Data Type: Numeric (Integer)**
   * **Description: The total number of calls made during the day.**
10. **total day charge**
    * **Data Type: Numeric (Float)**
    * **Description: Total charges for the customer’s daytime calls.**
11. **total eve minutes**
    * **Data Type: Numeric (Float)**
    * **Description: Total minutes spent on calls during the evening.**
12. **total eve calls**
    * **Data Type: Numeric (Integer)**
    * **Description: The total number of calls made during the evening.**
13. **total eve charge**
    * **Data Type: Numeric (Float)**
    * **Description: Total charges for the customer’s evening calls.**
14. **total night minutes**
    * **Data Type: Numeric (Float)**
    * **Description: Total minutes spent on calls during the night.**
15. **total night calls**
    * **Data Type: Numeric (Integer)**
    * **Description: The total number of calls made during the night.**
16. **total night charge**
    * **Data Type: Numeric (Float)**
    * **Description: Total charges for the customer’s night calls.**
17. **total intl minutes**
    * **Data Type: Numeric (Float)**
    * **Description: Total minutes spent on international calls.**
18. **total intl calls**
    * **Data Type: Numeric (Integer)**
    * **Description: The total number of international calls made.**
19. **total intl charge**
    * **Data Type: Numeric (Float)**
    * **Description: Total charges for the customer’s international calls.**
20. **customer service calls**
    * **Data Type: Numeric (Integer)**
    * **Description: The number of calls made by the customer to customer service.**
21. **churn**
    * **Data Type: Binary (Boolean)**
    * **Description: Whether the customer has churned (True) or not (False). This is the target variable for prediction.**

**Size of the Dataset**

* **Number of Rows: 3333 entries, each corresponding to a customer.**
* **Number of Columns: 21 columns, including both features and the target variable.**

**Data Preparation & Analysis**

1. **Missing Values**

The dataset did not have any missing on NaN values.

1. **Duplicate Values**

The dataset did not have any duplicates

3.**Combined relevant columns and dropped unused columns**

i.e (day,night,eve) calls, minutes and charges

**4. Feature Engineering**

**Outliers** are handled first (either removed or capped).

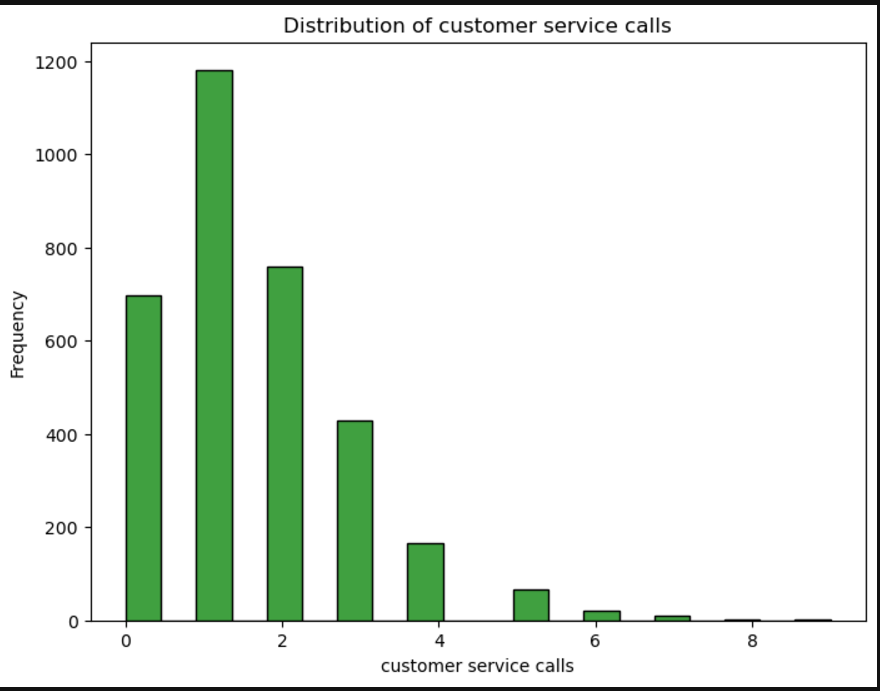
**Scaling** is applied after handling outliers to prevent distortion of the scaling process.

**Encoding** for categorical variables and target variable is done as needed.

**EDA**

**1.Churn Predictions**

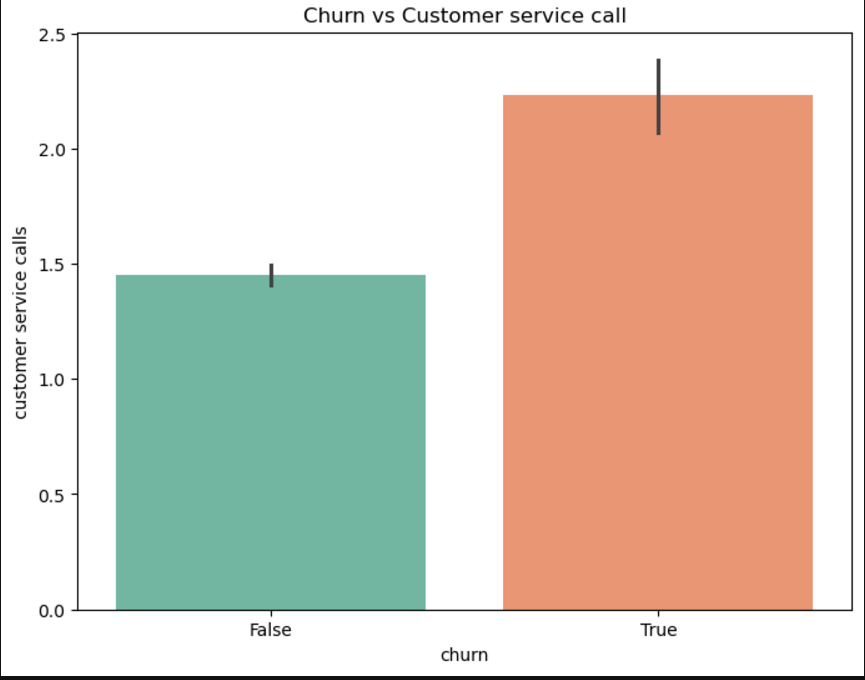
***i) Univariate***

****

**Observation**

The count of customers who have called customer service are evidently overpowering those who don't and this tends to affect churn rates either positively or negatively. i.e it may be a complaint or something positive

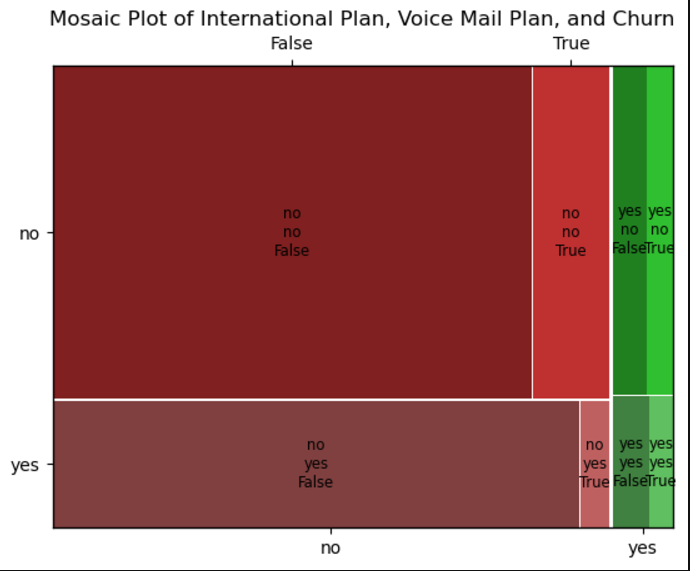
***ii) Bivariate Plot***

******

**Observation**

This plot clearly shows the relationship between churn and customer service calls, further supporting the argument that the number of calls made to customer service has an impact on churn On average, customers who have main calls to customer service 2 or more times terminated the service.

**iii) Multivariate**

****

Interpretation:

* International Plan ('yes' and 'no'):

Most customers without an international plan ('no') are shown in the large red block on the left. This indicates that a substantial portion of customers who are not subscribed to the international plan tend to have low churn rates (indicated by False churn). Customers with an international plan ('yes') make up a smaller portion of the plot, and their distribution is visible in the middle section of the plot.

* Voice Mail Plan ('yes' and 'no'):

The customers who have the voice mail plan ('yes') are shown in the bottom portion, with most of them not churning (False). The upper portion is taken up by those who do not have a voice mail plan ('no'), where churn is also more frequent (True), suggesting a correlation between not having a voice mail plan and higher churn.

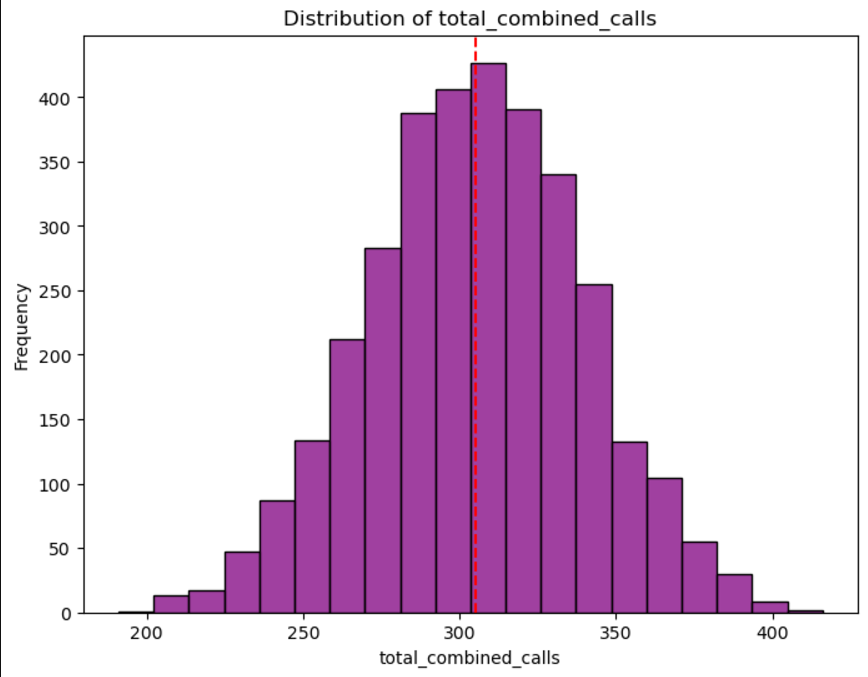
* Churn (True vs. False):

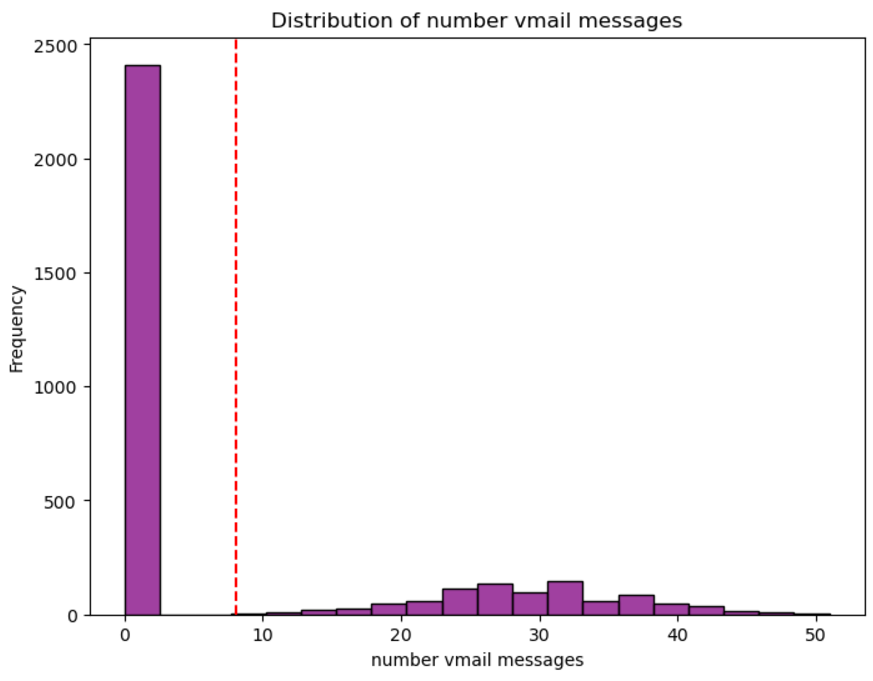
The green sections represent churn (True), and the red sections represent non-churn (False). The True churn category (on the right) is smaller compared to the False category (on the left), suggesting that churn is relatively low among the customers. International plan ('yes') and voice mail plan ('yes') customers are more likely to churn, though this is a small segment of the overall customer base.

Summary: Churn is more frequent among customers who do not have an international plan or voice mail plan. The combination of having both **an international plan and a voice mail plan seems to reduce churn, as seen in the smaller portion of churn in this group.**

**2.Customer Usage**

**i) *Univariate***

****

****

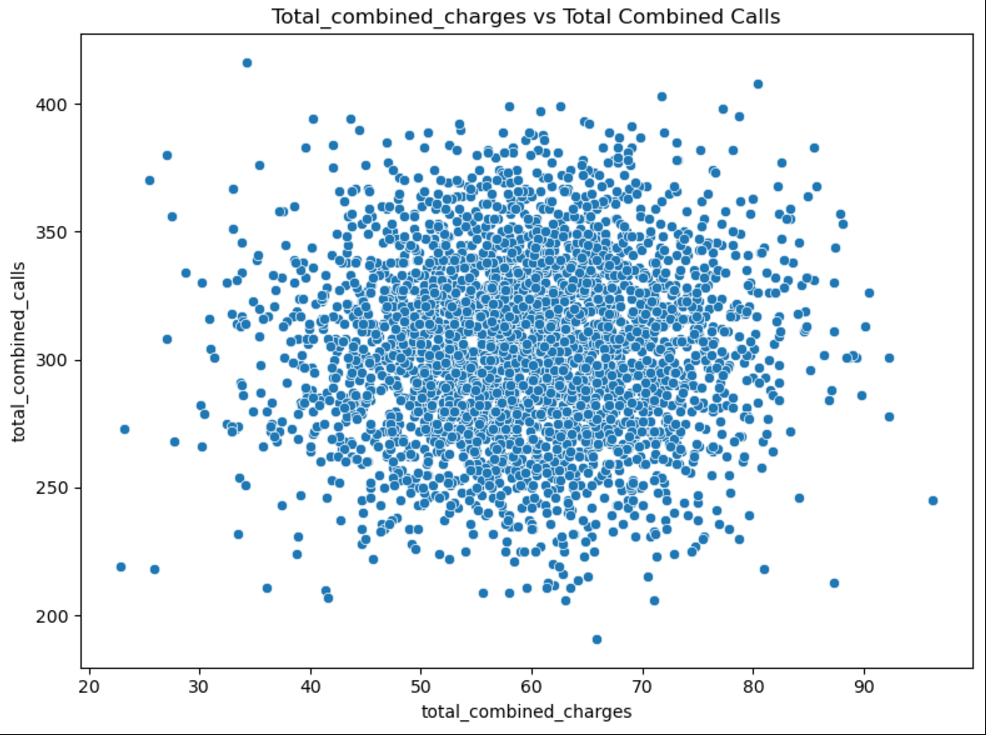
**Observation**

The plot for total combined calls is normally distributed displaying little to no skew thus it may act as a suitable predictor as compared to the number vmail messages

***ii) bivariate***

**Observation**

1. There is little to no correlation between total combined calls and total combine charges
2. Their values are spread evenly without a trend, not potraying any kind of linear relationship

****

**Modeling**

**For the task of predicting customer churn at SyriaTel, we will use a combination of models with varying levels of complexity, from a simple baseline model to more advanced models that can capture non-linear relationships and improve performance through hyperparameter tuning.**

**1st Model: Logistic Regression**

* **Justification:**
  + **Logistic Regression is a simple and interpretable linear model suitable for binary classification tasks, like churn prediction.**
  + **It provides a good starting point for assessing the effectiveness of other, more complex models.**
* **Metric of Success:**
  + **For this model, we'll focus on the Accuracy and Recall metrics. A higher recall will help ensure that we correctly identify customers who are at risk of churning (minimizing false negatives).**

**2nd Model: Decision Tree Classifier**

* **Justification:**
  + **Decision Trees are a non-linear model that can capture more complex patterns in the data compared to Logistic Regression.**
  + **This model works well when the relationship between features is not strictly linear, which is often the case in real-world customer behavior.**
  + **It is easier to interpret than some other complex models, providing insight into which features are the most influential in predicting churn.**
* **Metric of Success:**
  + **We'll measure Precision and Recall for the Decision Tree. Since churn prediction involves both minimizing false positives (Precision) and false negatives (Recall), a good balance of these metrics will be crucial.**

**3rd Model: Random Forest Classifier (Hyperparameter Tuned)**

* **Justification:**
  + **Random Forest is an ensemble learning method that combines multiple decision trees to create a robust, high-performing model.**
  + **It handles overfitting better than a single Decision Tree by averaging out predictions from several trees, improving generalization.**
  + **By tuning hyperparameters such as the number of trees, maximum depth, and minimum samples for splits, we can optimize the model's performance on the churn prediction task.**
* **Metric of Success:**
  + **ROC AUC and F1 Score will be the primary metrics. ROC AUC will measure the model's ability to distinguish between churn and non-churn classes, while F1 Score provides a balance between Precision and Recall, which is essential for imbalanced datasets like churn prediction.**

**Model Selection and Justification**

* **Logistic Regression will be used as a baseline to quickly assess the problem, providing an early benchmark.**
* **Decision Tree will be chosen as the second model for its ability to handle non-linear relationships, which might be more reflective of customer behavior.**
* **Random Forest, as the third model, allows us to take advantage of ensemble learning to improve predictive power and robustness, especially when hyperparameters are fine-tuned.**

**Each of these models will be evaluated based on different success metrics to ensure that they align with business goals, where the primary goal is to predict customers who are at risk of churning with high accuracy, while minimizing both false positives and false negatives.**

***Best Parameters***

Tuning hyperparameters for random\_forest...

Best parameters for random\_forest: {'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}

Tuning hyperparameters for decision\_tree...

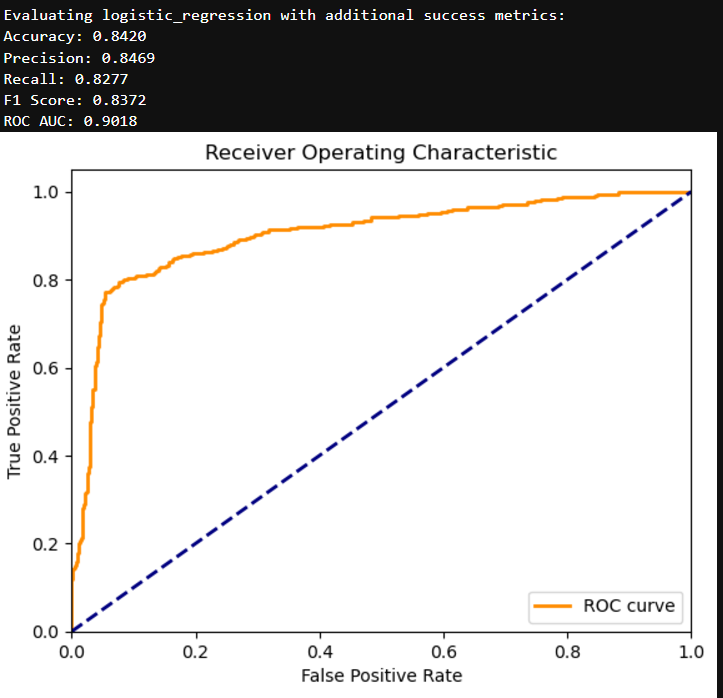
Best parameters for decision\_tree: {'criterion': 'entropy', 'max\_depth': 20, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2}

Tuning hyperparameters for logistic\_regression...

Best parameters for logistic\_regression: {'C': 10, 'solver': 'lbfgs'}

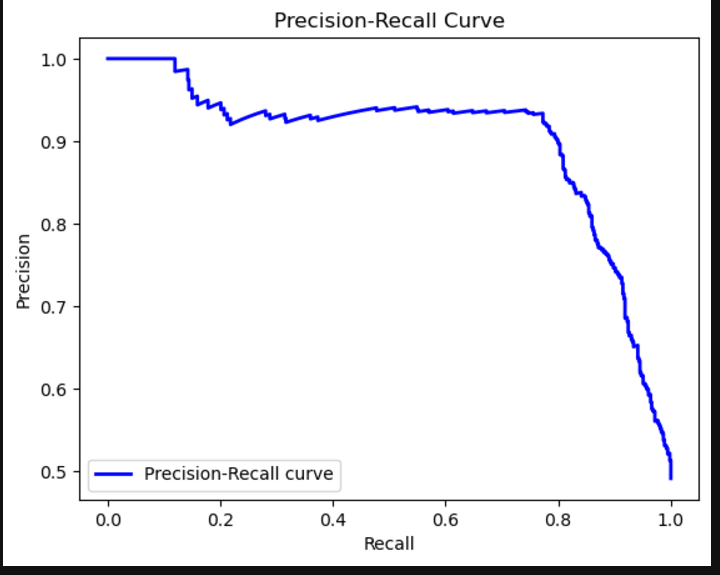
**Evaluation**

**1.Logistic Regression**



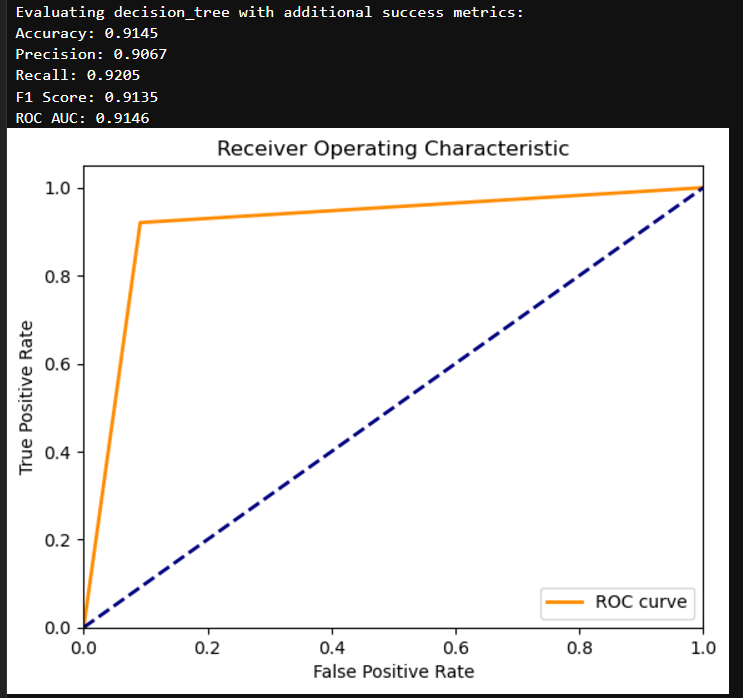
The ROC curve shown represents the performance of the logistic regression model:

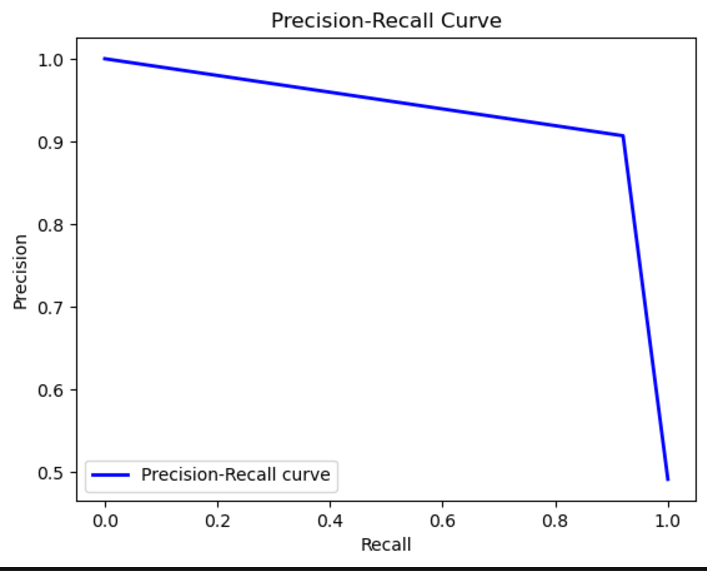
Curve Interpretation: The curve shows a good trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR). A sharp rise towards the top-left corner indicates a model with good predictive power. Model Performance: The orange ROC curve is significantly above the diagonal, indicating that the logistic regression model is performing well, distinguishing between positive and negative classes effectively.



The model is performing well with a good balance between precision and recall, favoring slightly more recall as per the precision-recall curve.

2.Decision Trees





The performance metrics you provided reflect a strong-performing model. Here's an interpretation with respect to True Positive Rate (TPR), False - - -Positive Rate (FPR), and the Precision-Recall curve:

* Accuracy (0.9145): The model correctly predicted 91.45% of all instances. This is a high level of correctness overall.
* Precision (0.9067): Precision measures the proportion of true positives (TP) among all instances predicted as positive. A precision of 91.48% indicates that when the model predicts a positive outcome, it is correct 91.48% of the time. This suggests the model is good at minimizing false positives.
* Recall (0.9205): Recall, or True Positive Rate (TPR), indicates how well the model identifies actual positives. A recall of 96.76% means the model correctly identifies 96.76% of the true positives. This is a high recall, indicating the model is good at detecting most of the positive instances.
* F1 Score (0.9405): The F1 score is the harmonic mean of precision and recall, providing a balance between them. An F1 score of 94.05% means the model is doing well in balancing precision and recall, and it's especially useful when you want to treat false positives and false negatives equally.
* ROC AUC (0.9410): The ROC AUC score measures the model's ability to distinguish between the positive and negative classes. An AUC of 94.10% means the model is very good at distinguishing between positive and negative instances. This is confirmed by the ROC curve, where a higher AUC corresponds to a model with fewer false positives and more true positives.

Interpretation with respect to TPR (Recall) and FPR: TPR (Recall) is 96.76%, which is excellent and shows the model is very sensitive in capturing positive cases. FPR (1 - TPR) would be 3.24%, which is low, indicating that the model is also good at avoiding false positives.

Precision-Recall Curve: The precision-recall curve is positioned towards the top right, suggesting that the model is effective at both identifying positive cases and minimizing false positives.

In conclusion, this model demonstrates excellent performance in terms of accuracy, precision, recall, and ROC AUC, with a good balance between identifying positive cases (high recall) and minimizing false positives (high precision).

3.Random Forest

Accuracy (0.9498): The model correctly predicted 94.98% of all instances. This indicates strong overall performance.

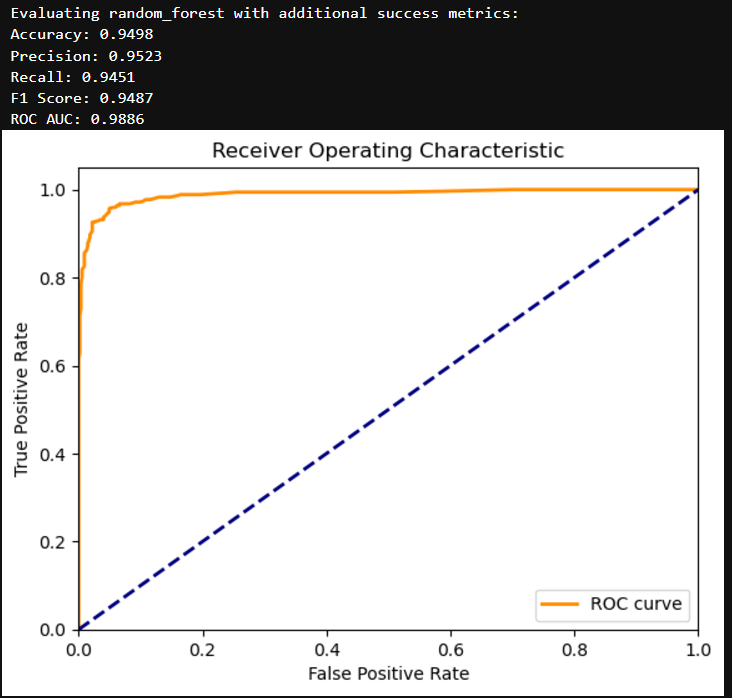
Precision (0.9523): Precision of 95.23% indicates that when the model predicts a positive outcome, it is correct 95.92% of the time. This suggests a very low false positive rate.

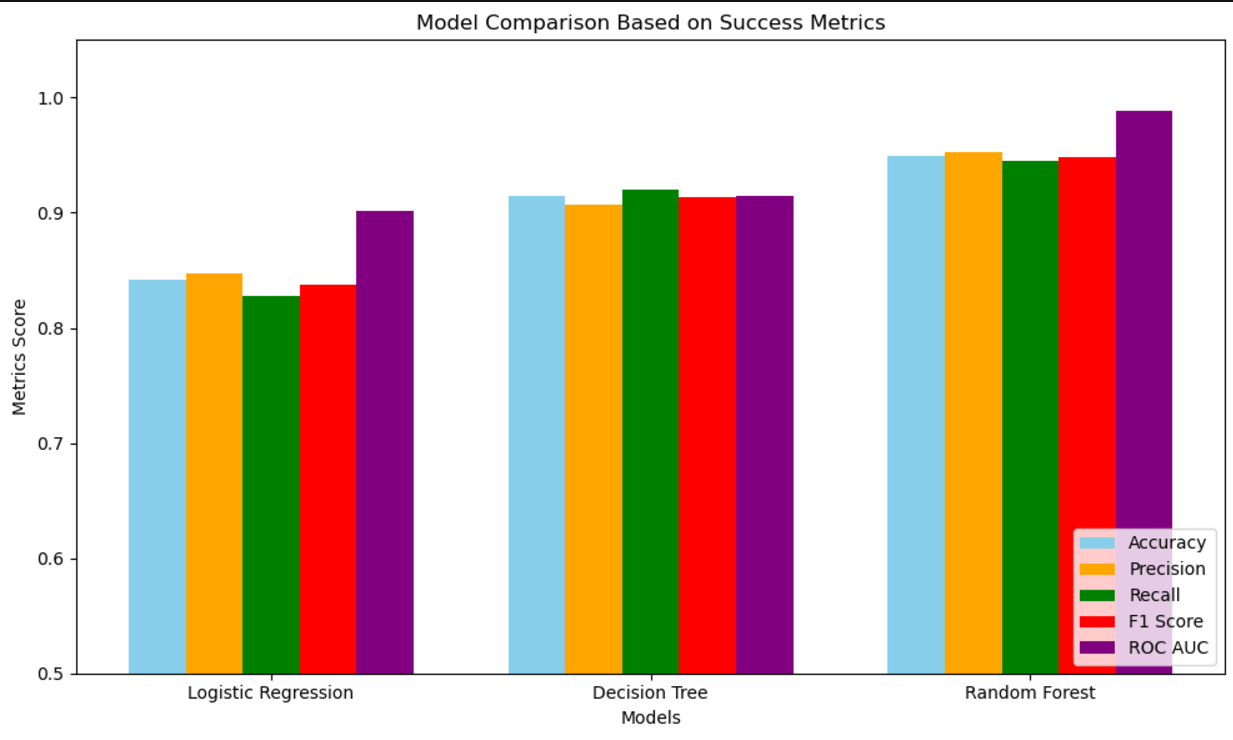
Recall (0.9451): Recall of 94.51% means the model correctly identifies 94.51% of the actual positive instances, showing that it is highly effective in detecting positive cases (high True Positive Rate or TPR).

F1 Score (0.9487): The F1 score is the harmonic mean of precision and recall. With an F1 score of 94.87%, the model achieves an excellent balance between precision and recall, ensuring both low false positives and low false negatives.

ROC AUC (0.9886): The ROC AUC score of 98.86% indicates exceptional ability to distinguish between positive and negative classes. A high ROC AUC reflects that the model has very few false positives and a very high true positive rate.

Interpretation with respect to TPR (Recall) and FPR: TPR (Recall) of 94.87% suggests that the model is excellent at identifying positive cases (True Positive Rate). FPR (1 - TPR) would be 5.13%, which is low, indicating the model avoids false positives effectively. Precision-Recall Curve: The Precision-Recall curve for this model is positioned towards the top right, indicating the model is effective at both minimizing false positives (high precision) and capturing most of the true positives (high recall).





With reference to overall metric score,, Logistic regression has the lowest while Random forest has the highest overall score  
  
  
Conclusions

The Random Forest model is highly suitable for predicting customer churn at SyriaTel. Its high accuracy, precision, recall, and F1 score demonstrate that it can reliably identify customers at risk of leaving while minimizing false positives and false negatives. The ROC AUC score further confirms its excellent discrimination capability. By implementing this model, SyriaTel can enhance its customer retention efforts, optimize resources, and take timely actions to reduce churn, thereby improving customer satisfaction and loyalty.

Recommendations

1. **Implement Predictive Retention Strategies**

Targeted Retention Campaigns: Focus resources on customers most likely to churn using the model's high precision.

1. **Monitor and Update the Model Regularly**

* Continuous Monitoring: Evaluate the model with fresh data to detect changes in patterns.
* Retraining: Update the model periodically to ensure its relevance and adaptability to new trends.

1. **Utilize Model Insights for Customer Experience Enhancement**

* Improve Support: Address service issues( high customer service calls was linked to churn risks).
* Enhance Experience: Offer incentives like discounts or personalized solutions to retain at-risk customers.