**Syriatel Customer Churn Data Report**

1. **Overview**

Syriatel, a telecommunications company, aims to minimize revenue losses caused by customer churn. The company is focused on identifying the key factors contributing to customer attrition and understanding the reasons behind customers discontinuing their services.

1. **Problem Statement**
   1. **Problem**

The main task of this project is to identify the factors driving customer churn and develop actionable strategies to reduce it. This will help the company to take appropriate actions on time to avoid losing customers and thus revenue at the same time.

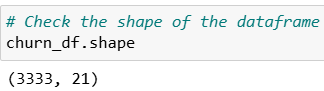
* 1. **Objectives**
* Develop a predictive model to identify customers at risk of churning and any characteristics that are indicative of churn.
* Focus retention efforts on the most at-risk segments to maximize return on investment in customer satisfaction programs.
* Explore patterns and behaviors and use the insights to implement targeted interventions, such as proactive customer support or offering better plans for high usage customers.
  1. **Outcomes**

At the end of the project, we should be able to come up with:

* A predictive model capable of identifying at-risk customers.
* Insights into key factors driving churn.
* A system for proactive customer retention strategies.

1. **Data Understanding**

The dataset contains 3,333 rows and 21 columns. This can be confirmed by checking the dimension of the dataset.



Here's a summary of its structure:

**Summary:**

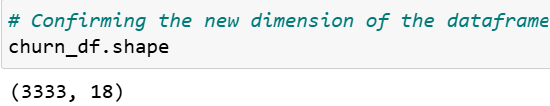
Columns: 21, including numerical, categorical, and boolean data types.

The respective columns are:

* state: Customer state (e.g., KS, OH).
* account length: Length of customer account in days.
* area code: Area code of the customer.
* phone number: Customer phone number.
* international plan: Whether the customer has an international calling plan (yes or no).
* voice mail plan: Whether the customer has a voicemail plan (yes or no).
* number vmail messages: Number of voicemail messages.
* total day minutes, total day calls, total day charge: Metrics for daytime usage.
* total eve minutes, total eve calls, total eve charge: Metrics for evening usage.
* total night minutes, total night calls, total night charge: Metrics for nighttime usage.
* total intl minutes, total intl calls, total intl charge: Metrics for international usage.
* customer service calls: Number of customer service calls made.
* churn: Whether the customer churned (True or False).
  1. **Exploratory Data Analysis (EDA)**

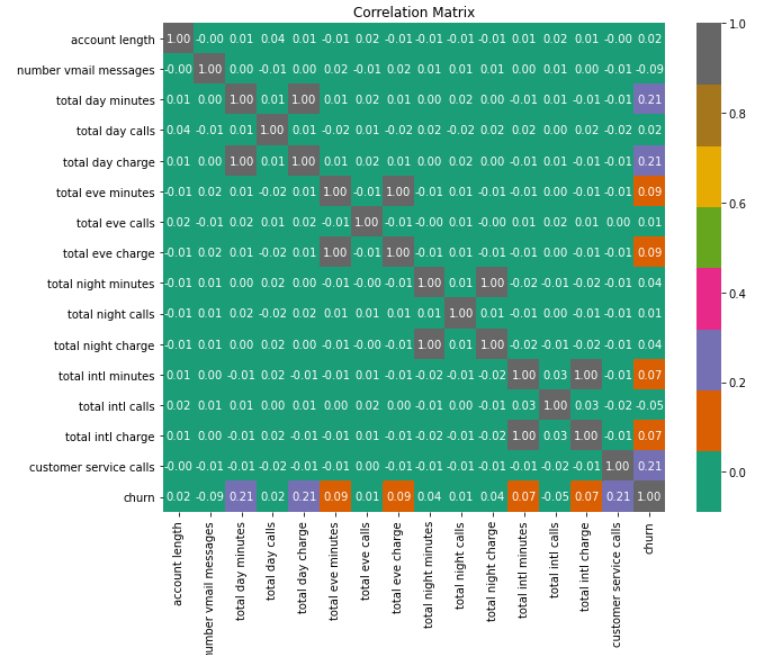
We now do EDA to understand the structure, relationships, and key insights of the data. Before that, drop columns that will not be useful for analysis such as phone number, area code and state.

After dropping the columns above, the dimension of the data frame will have changed as follows:



* 1. **Visualizations to show some key metrics about the data frame**

1. Correlation Matrix

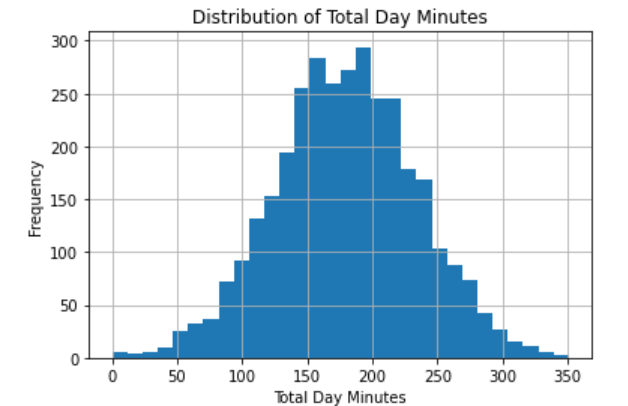
­ 

It is clear from the correlation matrix that the factors that have a high correlation with churn are:

* Total day minutes
* Total day charge
* Customer service calls

1. Histogram for Total Day Minutes

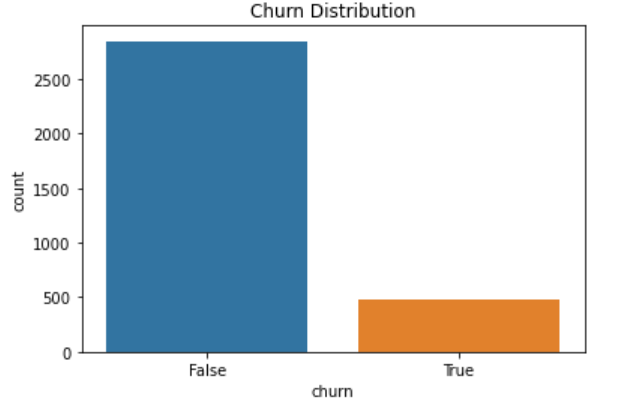
We also want to examine the distribution of minutes utilized by a customer during the day.



The histogram shows a somewhat normal distribution for the amount of minutes utilized by a customer during the day.

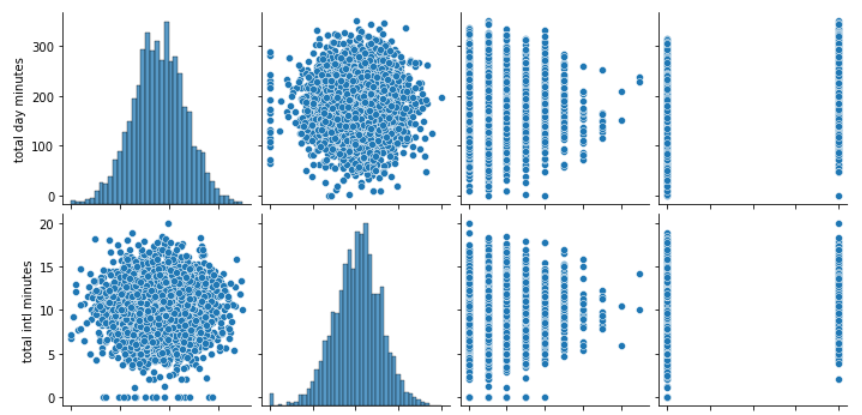
1. Countplot for churn distribution

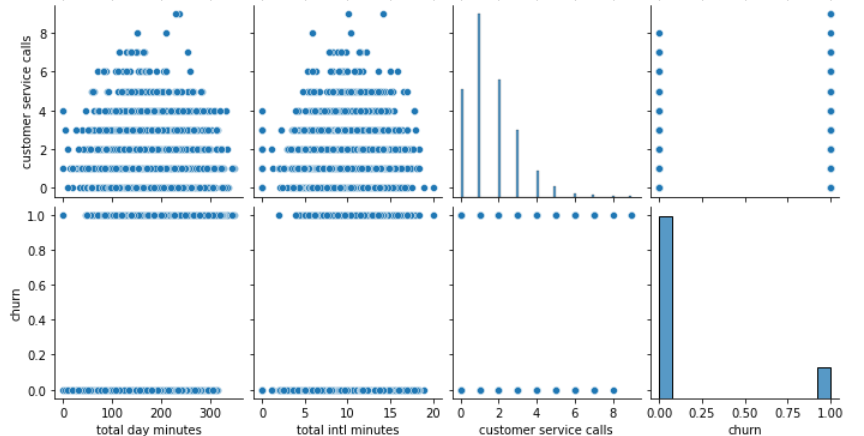
The plot shows the count of customers who have left the business and those who haven’t.



1. Pairplot for pairwise relationship between different variables.

The pairplot shows the relationship between different variables in the dataset and will come in handy in building a model that will help to predict churn.­





From the pairplots above, it is clearly visible that the relationship between most variables do have a somewhat normal distribution.

1. **Modeling**

For the purposes of this project, three Machine Learning Models were used:

* Logistic Regression Model
* Decision Tree
* Decision Tree with Hyperparameter Tuning

Before we start analysis, there’s need to explain the justification for choosing the above models.

**Why Use Logistic Regression**

1. Logistic regression is computationally efficient and less prone to overfitting on smaller datasets compared to more complex models.
2. By examining the coefficients, we can identify which features are most predictive of churn, aiding in actionable insights.
3. Logistic regression is easy to implement and interpret, providing a clear understanding of how each feature impacts the probability of churn.
4. The coefficients directly indicate the direction (positive or negative) and magnitude of the effect of features on the likelihood of churn.

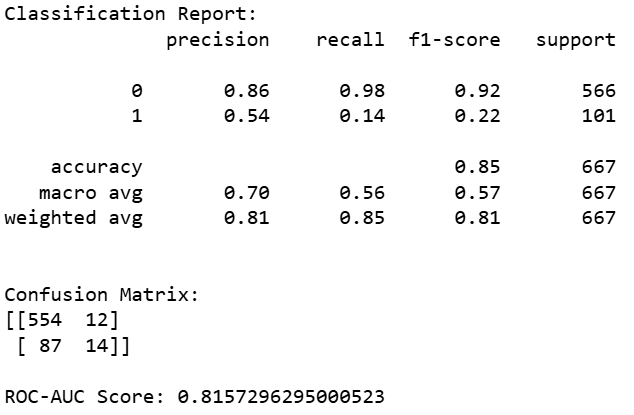
**Why Use Decision Trees with Hyperparameter Tuning?**

1. Handles Mixed Data Types - Decision trees naturally handle both numerical and categorical features without requiring extensive preprocessing (e.g., encoding or normalization).
2. Captures Nonlinear Relationships - Decision trees can model complex, nonlinear interactions between features and the target variable, such as thresholds where churn probability increases sharply (e.g., high customer service calls).
3. Robustness Through Hyperparameter Tuning - Hyperparameter tuning (e.g., adjusting max\_depth, min\_samples\_split, min\_samples\_leaf) improves generalization by preventing overfitting.
   1. **Logistic Regression Model**

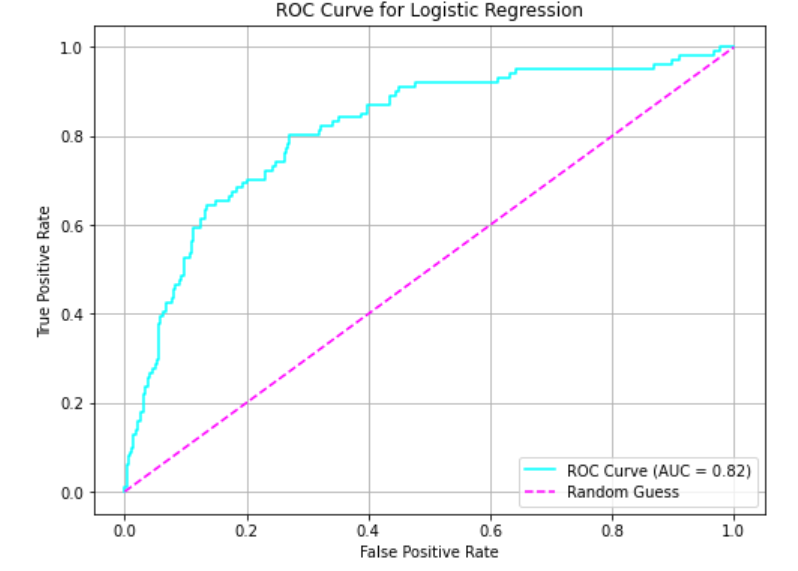
The following steps were followed to create the model

* Preprocess the data: Ensure features and the target variable are appropriately prepared for modeling.
* Train the model: Fit a logistic regression model to the training data.
* Evaluate performance: Use metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess the model.
* Analyze coefficients: Interpret the contribution of each feature to the likelihood of churn.

The model yielded the following results:



The ROC curve for the Logistic Regression Model is shown below.

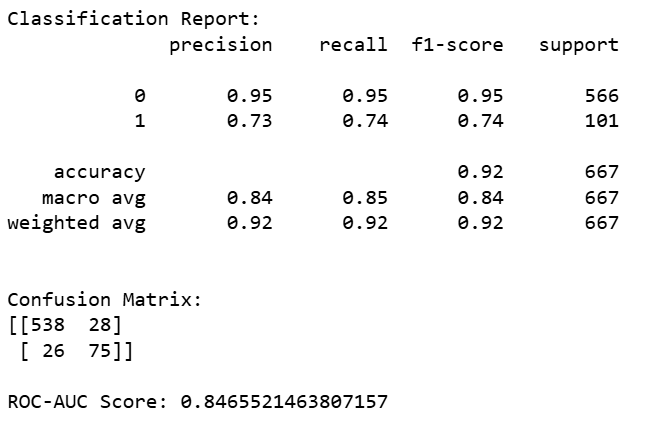


With a ROC score of \*2%, it is quite clear that the Logistic regression model is not accurate enough and thus a second model was put into test.

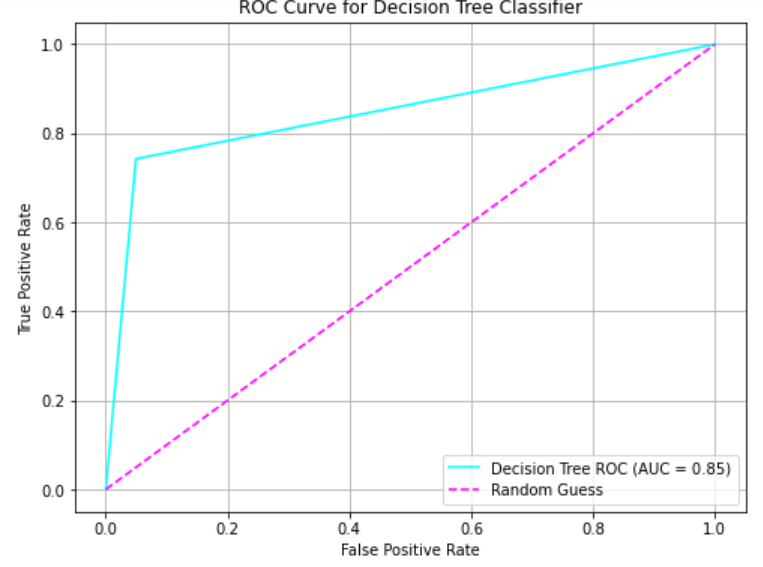
* 1. **Decision Tree Algorithm**

The goal here is to train a Decision Tree without hyperparameter tuning.

The algorithm yielded the below results:



The ROC curve for the Decision Tree is shown below:



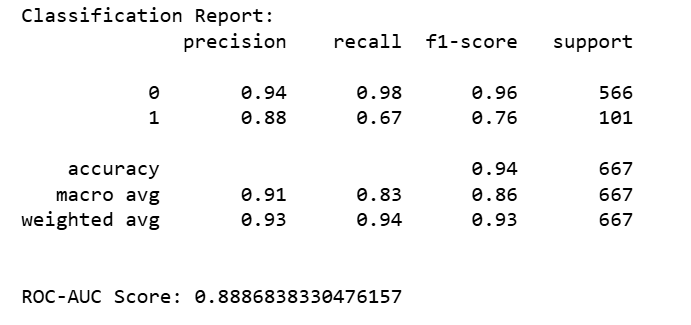
From the graph above, it is evident that the Decision Tree improves a little from the Logistic Regression model with a ROC score of 85%.

At this point, we shall now improve the performance of the Decision Tree with Hyperparameter Tuning

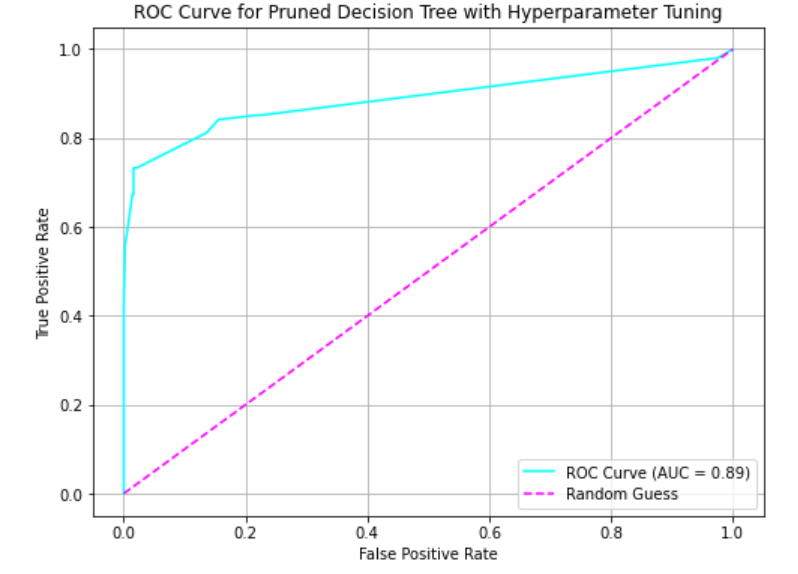
* 1. **Decision Tree with Hyperparameter Tuning**

At this point, we select the best hyperparameters to train the machine learning model. We perform hyperparameter tuning by setting the max\_depth of the tree to 5 and min\_samples\_split of 10 to optimize the model. This will significantly improve the accuracy of the decision tree.

This yielded the below results:



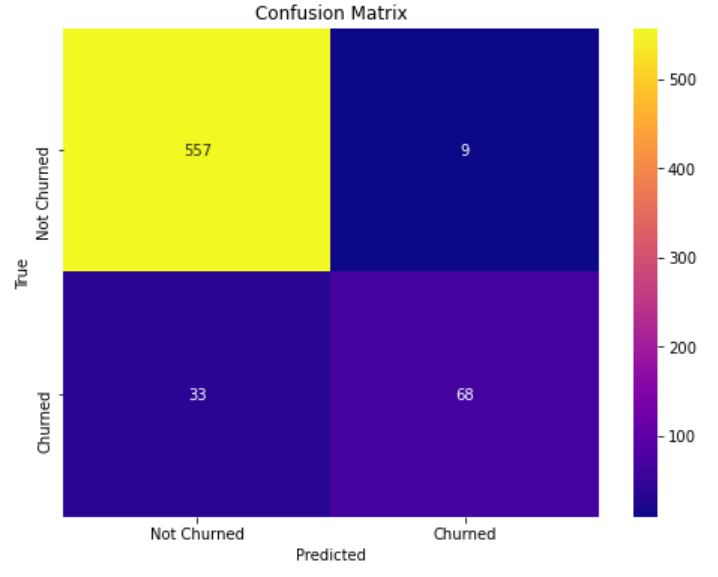
The ROC curve for the Decision Tree is as shown below:



The Decision Tree with Hyperparameter Tuning has achieved an ROC of 89% and thus is much more accurate than the two models above.

**Confusion Matrix**

The confusion matrix helps to understand how well the classifier is performing. The confusion matrix of the model is also shown below:



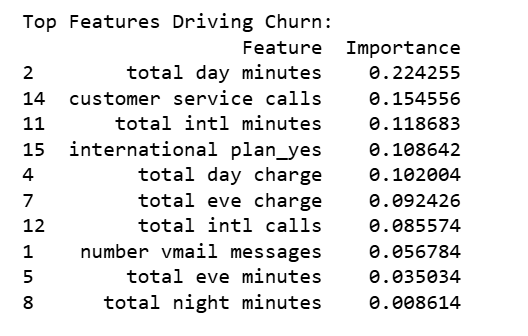
The model was able to correctly predict 68 customers who churned against a total of 667 customers.

It also correctly predicted 557 customers against 667 tested who did not churn.

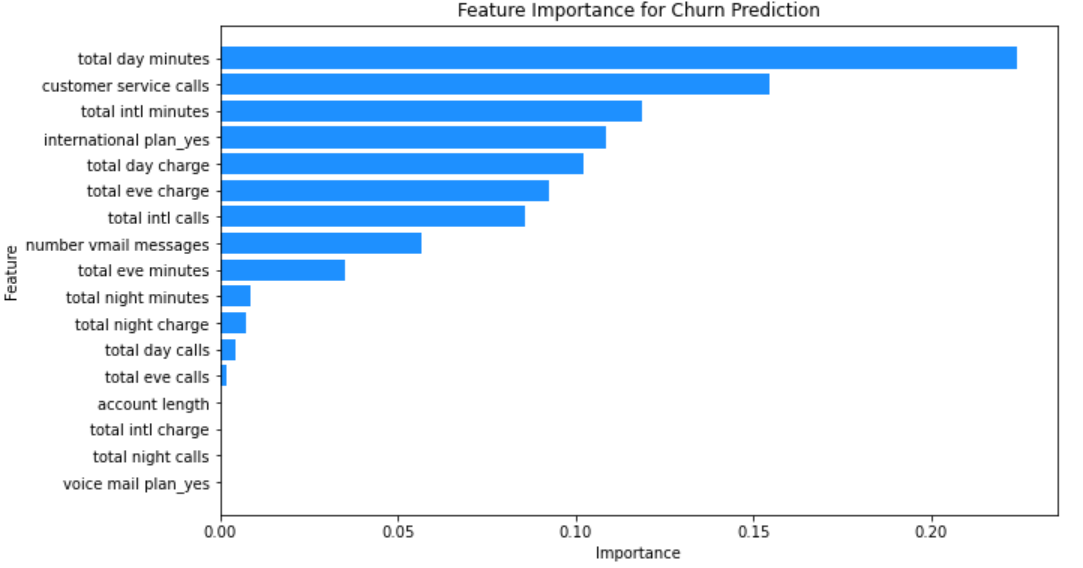
Thus, with more feature training, the model can perform much better and reach higher accuracy levels.

1. **Findings**

Based on the different features under analysis, this were the findings for the different metrics driving churn.



The visual for the above features is as shown below:



1. **Conclusion**

From the analysis

* A high number of customer service calls often indicates dissatisfaction, which and thus may increase churn probability.
* The number one feature that drives customer churn is total number of minutes that customers utilize to make calls during the day.
* Features like international plan and voice mail plan might affect satisfaction and churn likelihood.
* High usage in categories like total day minutes or frequent customer service calls could indicate either high engagement or dissatisfaction.

**6. Recommendations**

* Based on usage patterns (e.g., total day minutes, total intl minutes), offer customized plans to prevent customers from feeling they are overpaying or not getting enough resources for the money spent.
* Focusing on customers with high service usage, frequent complaints, or those on specific plans could help target retention strategies.
* Investigating state-wise trends or customer demographics could offer insights for localized service improvement.
* Monitor Customer Service Calls: Customers making frequent calls to customer service are likely frustrated. Improve resolution rates and provide follow-up after service issues are resolved.
* Use predictive analytics to identify customers at risk of churning and offer personalized discounts or incentives.
* Use machine learning models to predict churn early based on patterns like frequent customer service calls, declining usage, or increased complaints.