SYRIATEL CUSTOMER CHURN PREDICTION

By: Michael Muthui Gatero

OVERVIEW

What is Customer Churn?

- Customer churn is the percentage of customers who stop using a service over time.
- It is a major challenge in the telecommunications industry, leading to revenue loss and increased customer acquisition costs.

- By leveraging machine learning models, SyriaTel can identify customers at risk of leaving and implement targeted retention strategies to improve customer loyalty and reduce churn.
- This study aims to develop an accurate churn prediction model and provide actionable insights to help SyriaTel retain its customers.

BUSINESS AND DATA UNDERSTANDING

Business problem:

- SyriaTel faces high customer churn, leading to revenue loss and increased customer acquisition costs. Without proactive measures, the company risks losing market share. Predicting churn helps retain customers through targeted interventions.

Dataset:

- The dataset used in this project is the SyriaTel Customer Churn Dataset. It contains approximately 3,333 rows and 21 columns.
- The dataset includes various customer attributes such as: call minutes, service usage, contract type, and customer service interactions.
- Target variable: Churn (1 = customer left, 0 = customer stayed).
- Engineered three new features to enhance model performance:
  - Total Minutes: Sum of day, evening, and night minutes
  - Total Calls: Sum of day, evening, and night calls
  - Total Charges: Sum of day, evening, and night charges
- These features provide a more comprehensive view of customer usage patterns and their relationship with churn.

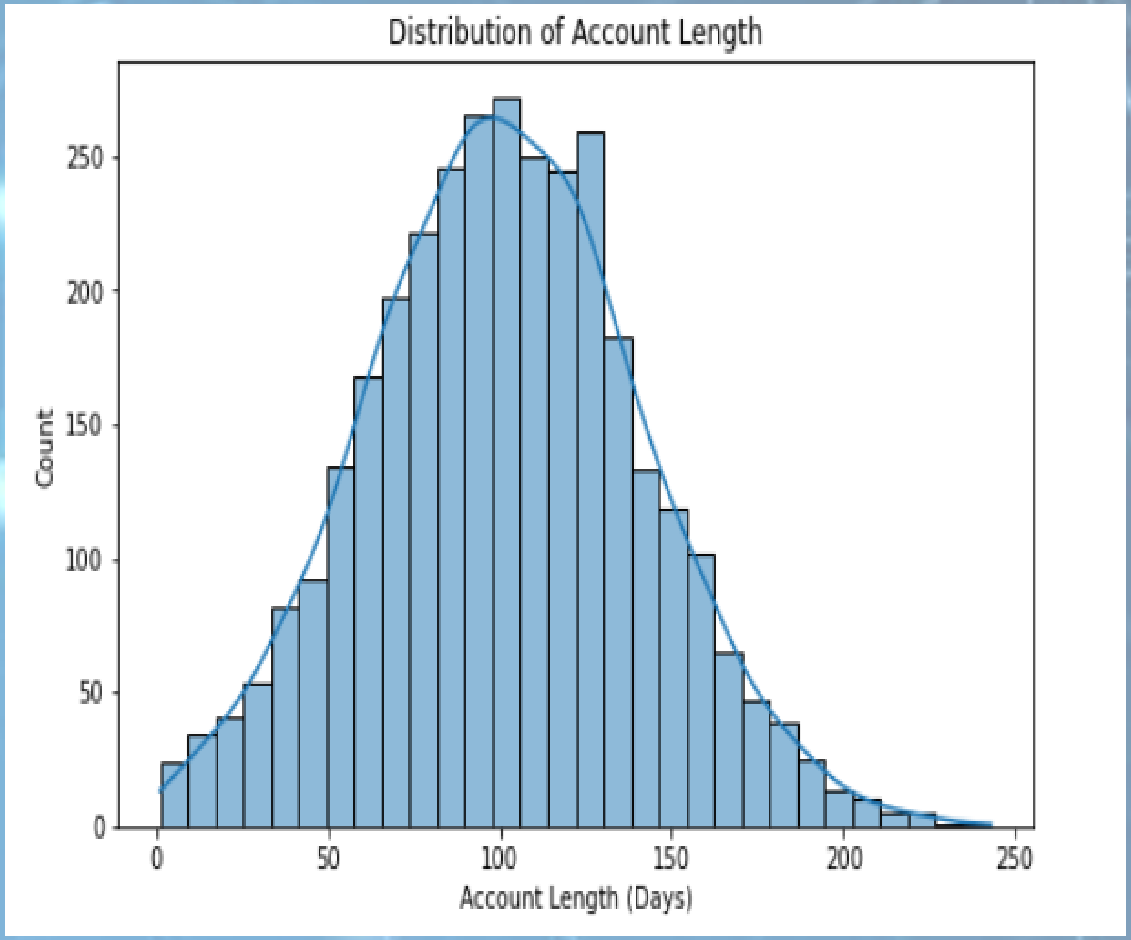
DATA ANALYSIS

- Carried out analysis on the data:
  1. Univariate Analysis
  2. Bivariate Analysis
  3. Multivariate Analysis

I. Univariate Analysis

- Distibution of account length

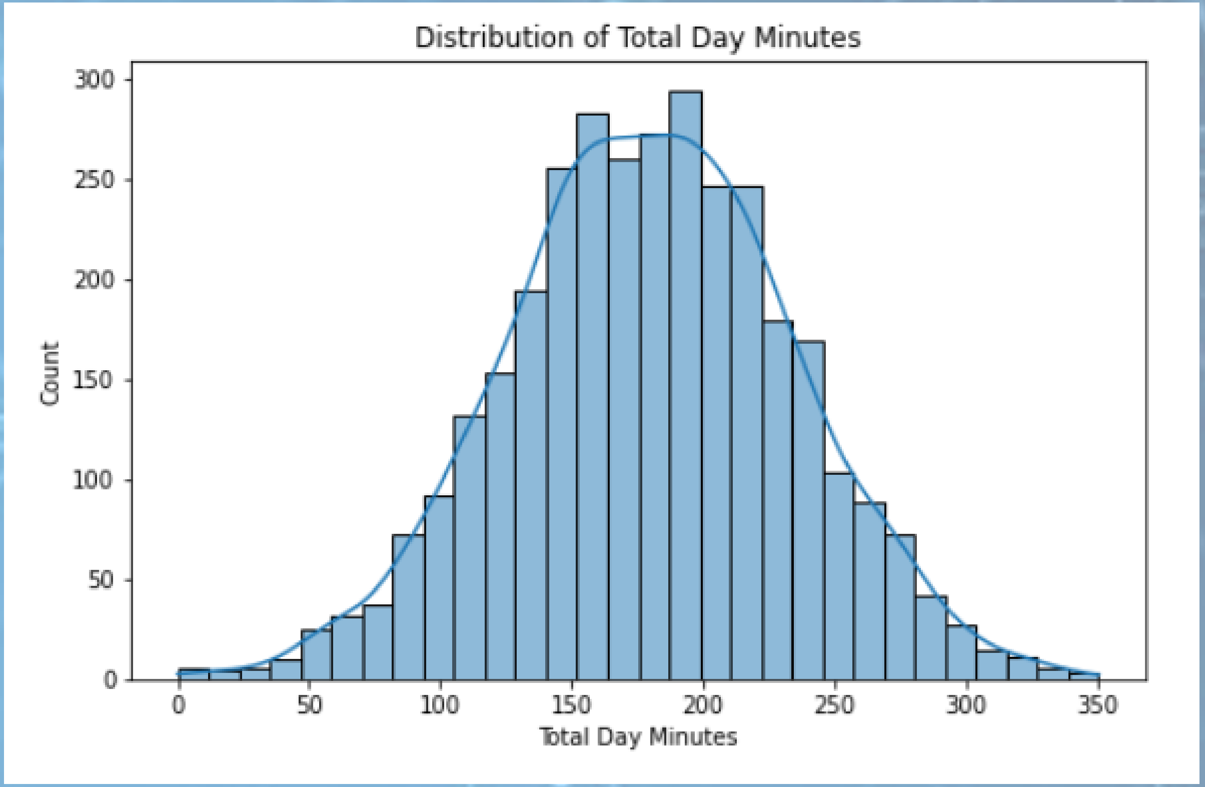
- The histogram shows that the distribution appears to be right-skewed, meaning most customers have relatively short account lengths, but some have significantly longer durations.
- The highest concentration of customers is around 90-110 days.
- There is a gradual decline in the number of customers as the account length increases.



UNIVARIATE ANALYSIS

Distribution of total day minutes

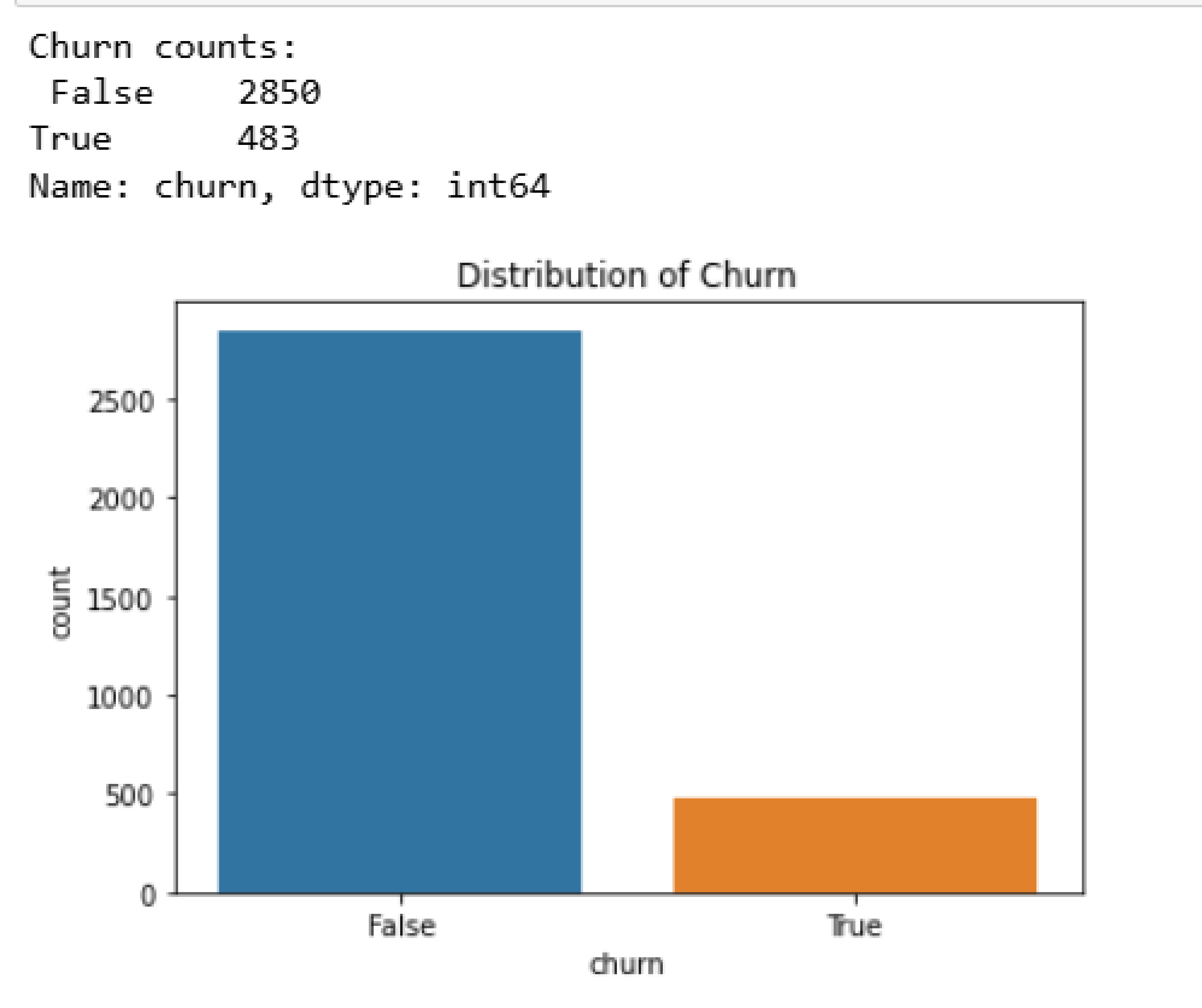
- The distribution appears approximately normal, with most customers having total day minutes between 100 and 250 minutes.
- The peak (mode) occurs around 175-200 minutes, meaning most customers fall within this range.
- There is slight right-skewness, suggesting some customers use significantly more minutes during the day.
- There are no extreme outliers, indicating that day-minute usage is fairly consistent across customers.



UNIVARIATE ANALYSIS

Distribution of Churn

- Majority of customers did not churn (2850), while 483 customers churned.
- Class imbalance is present, which may affect model performance.
- Potential revenue impact due to churned customers.



```
Churn counts:
 False    2850
True      483
Name: churn, dtype: int64
```

II. BIVARIATE ANALYSIS

- Performed several bivariate analysis:

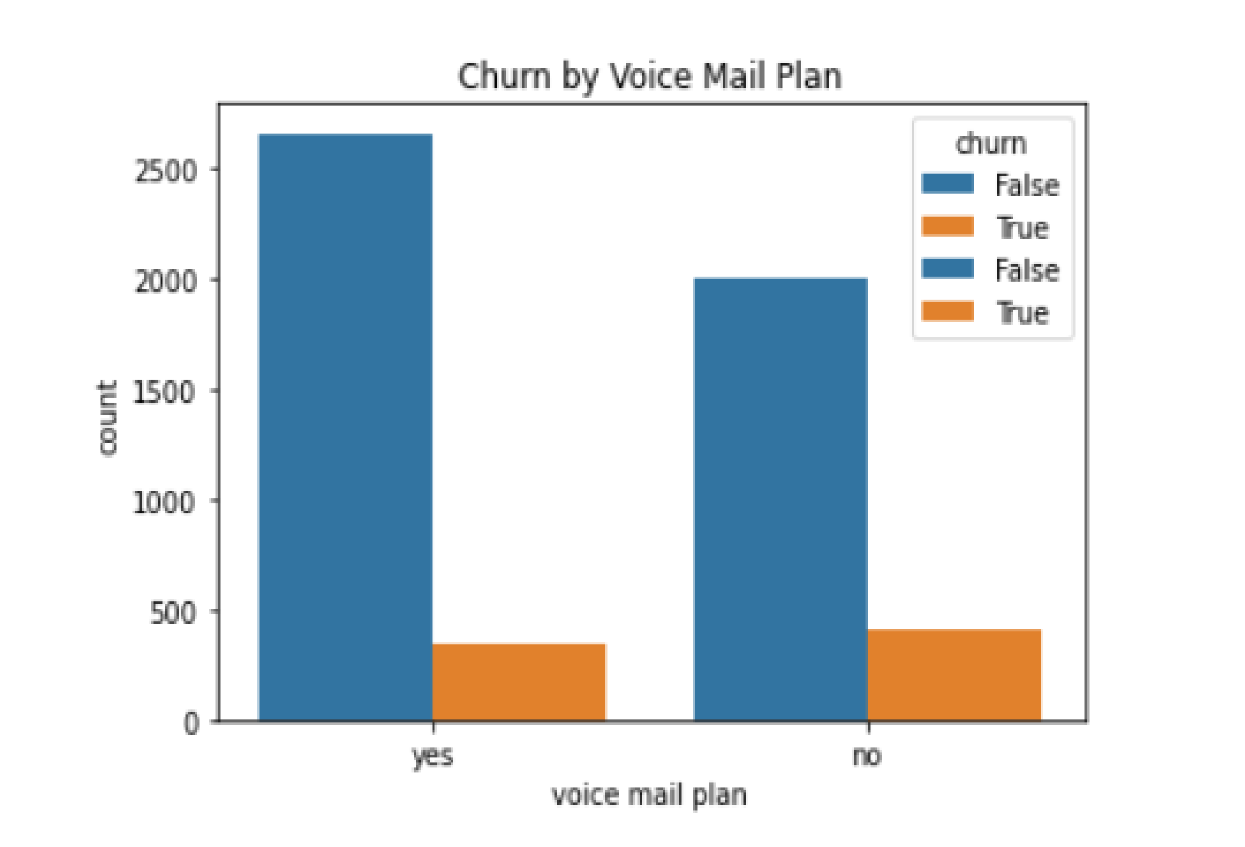
1. Churn distribution by categorical features

i) Churn by International Plan:

- Customers with an international plan have a higher churn rate compared to those without it.
- The proportion of churned customers is noticeably higher among those with an international plan.

ii) Churn by Voice Mail Plan:

- Customers without a voice mail plan have a higher churn rate than those who have it. Having a voice mail plan seems to be associated with lower churn.
- These insights suggest that offering an international plan may be linked to customer dissatisfaction, while a voice mail plan could be a factor in customer retention.



BIVARIATE ANALYSIS

2. Numerical Features vs. Churn

- Each boxplot compares the distribution of a numerical feature between churned and non-churned customers.

1. Churn vs Total Call Minutes (Day, Evening, Night, International):

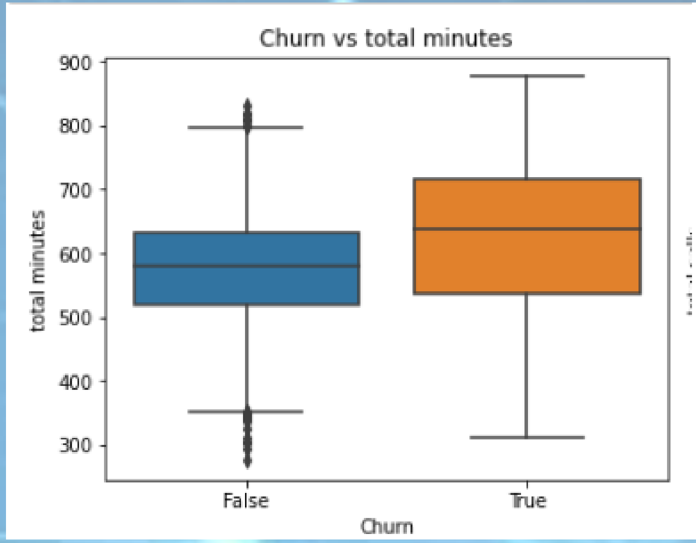
- The boxplot indicates that churned customers tend to have higher total minutes across all time periods (day, evening, night, and international).

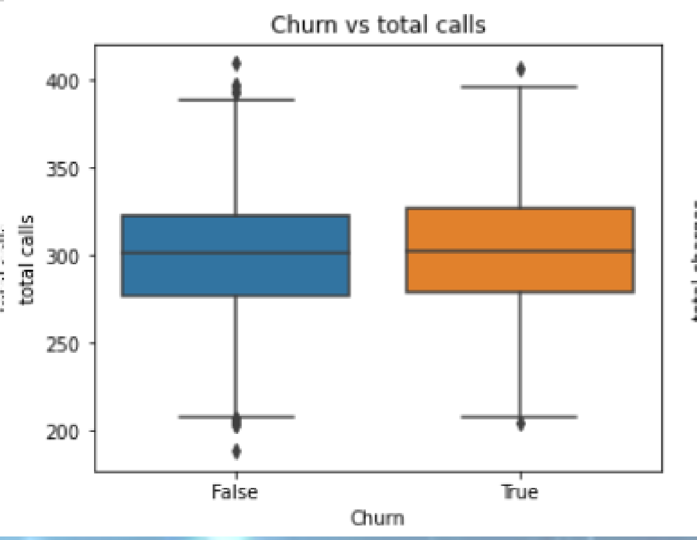
2. Churn vs Total Calls (Day, Evening, Night, International):

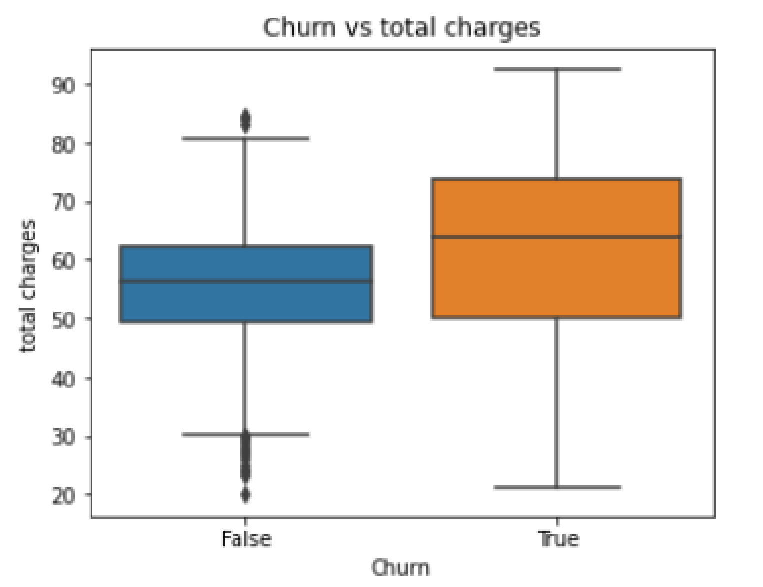
- The distribution of total calls is relatively similar between churned and non-churned customers.

3. Churn vs Total Charges (Day, Evening, Night, International):

- Churned customers generally have higher total charges in all categories.







MULTIVARIATE ANALYSIS

1. **Feature Analysis**

Total Minutes vs. Total Calls:

1. There is no clear relationship between total minutes and total calls.
2. Churned (orange) and non-churned (blue) customers are evenly distributed, suggesting total calls alone may not be a strong predictor of churn.

Total Minutes vs. Total Charges:

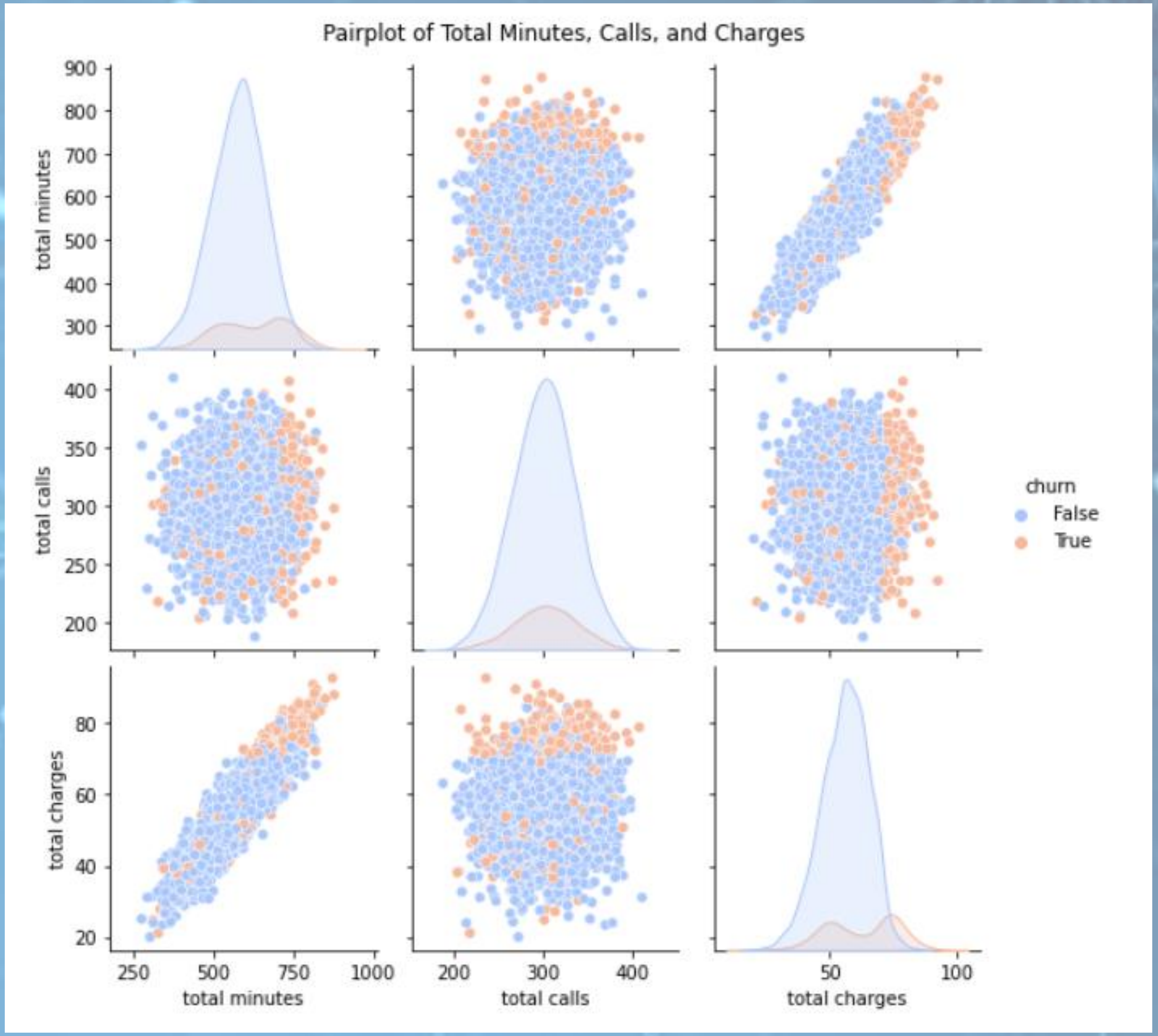
1. A strong positive correlation is observed, as expected (charges increase with minutes).
2. Churned customers seem to have higher total minutes and total charges.

Total Calls vs. Total Charges:

1. No clear trend is visible, implying total calls are not directly linked to charges.
2. Churned customers do not show a distinct pattern in this relationship.

Density Distribution:

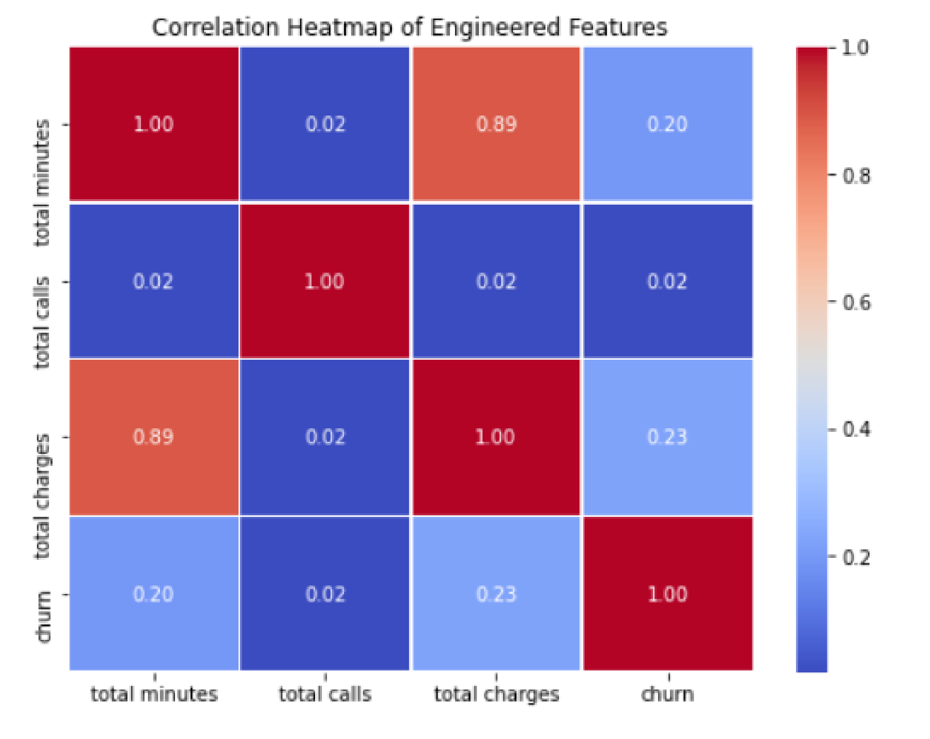
1. Churned customers tend to have a lower density in total minutes and total charges, indicating higher usage might be a risk factor for churn.
2. Most customers (blue) fall within a normal range, but churners (orange) are slightly more spread out.



MULTIVARIATE ANALYSIS

2. **Correlation Heatmap**

- The heatmap shows that total minutes and total charges are highly correlated (0.89), which may cause multicollinearity.
- Churn has a weak positive correlation with total minutes (0.20) and total charges (0.23), suggesting that higher usage slightly increases churn risk.
- Total calls has almost no correlation with churn (0.02), making it a weak predictor.



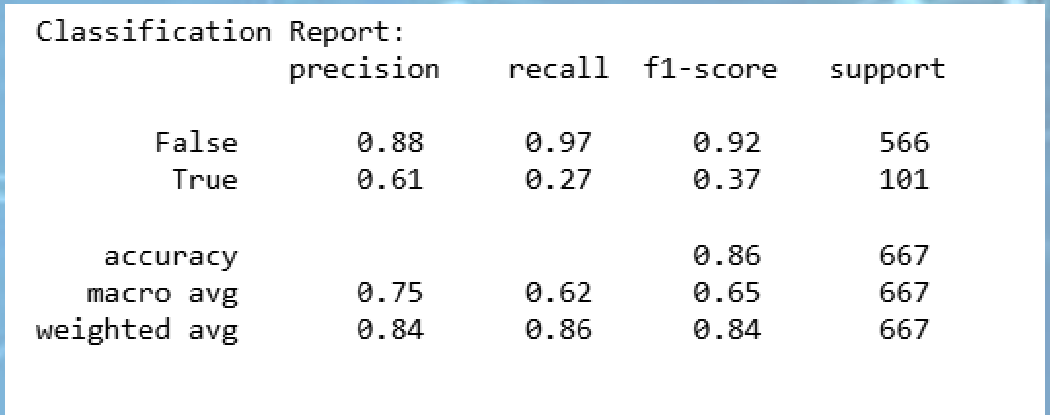
MODELING

Baseline Model - Logistic Regression

- Trained a Logistic Regression model as the baseline.
- Evaluated default performance using accuracy, precision, and recall.
- Tuned the decision threshold to improve recall.

Results:

- **Accuracy:** 86%
- **Precision (Churned Customers):** 61%
- **Recall:** 27%
- **F1-score:** 37%



```
Classification Report:
              precision    recall  f1-score   support

       False       0.88      0.97      0.92       566
        True       0.61      0.27      0.37       101

    accuracy                           0.86       667
   macro avg       0.75      0.62      0.65       667
weighted avg       0.84      0.86      0.84       667
```

Conclusion:

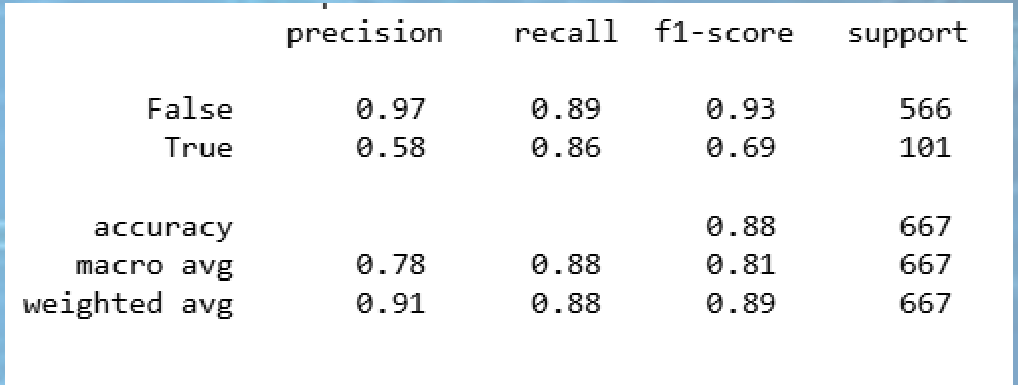
- Poor recall meant that many actual churners were **missed**.
- Adjusting the threshold improved recall but significantly reduced precision.
- **Decision:** Not suitable for churn prediction.

Decision Tree Model

- Trained a Decision Tree classifier to capture non-linear relationships.
- Tuned hyperparameters like max_depth and min_samples_split.

Results:

- **Accuracy:** 88%
- **Precision (Churned Customers):** 58%
- **Recall:** 86%
- **F1-score:** 69%



```
              precision    recall  f1-score   support

       False       0.97      0.89      0.93       566
        True       0.58      0.86      0.69       101

    accuracy                           0.88       667
   macro avg       0.78      0.88      0.81       667
weighted avg       0.91      0.88      0.89       667
```

Conclusion:

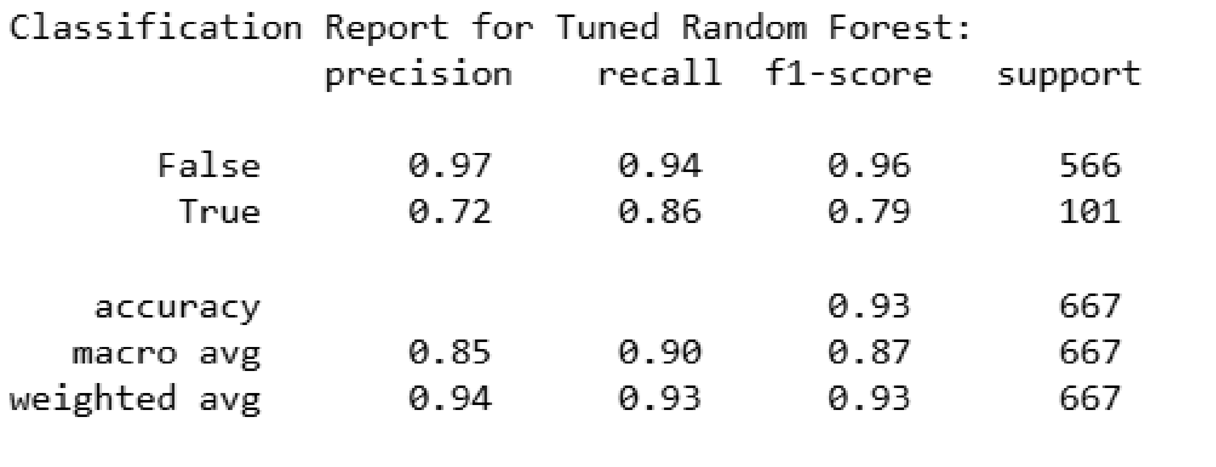
- Recall improved but at the cost of lower precision, leading to more false positives.
- Model was prone to **overfitting** without additional tuning.
- **Decision:** Not selected as the final model.

Random Forest Model (Tuned)

- Trained a Random Forest classifier with 100+ trees.
- Applied hyperparameter tuning to improve recall.

Results:

- **Accuracy:** 93%
- **Precision (Churned Customers):** 72%
- **Recall:** 86%
- **F1-score:** 79%



```
Classification Report for Tuned Random Forest:
              precision    recall  f1-score   support

       False       0.97      0.94      0.96       566
        True       0.72      0.86      0.79       101

    accuracy                           0.93       667
   macro avg       0.85      0.90      0.87       667
weighted avg       0.94      0.93      0.93       667
```

Conclusion:

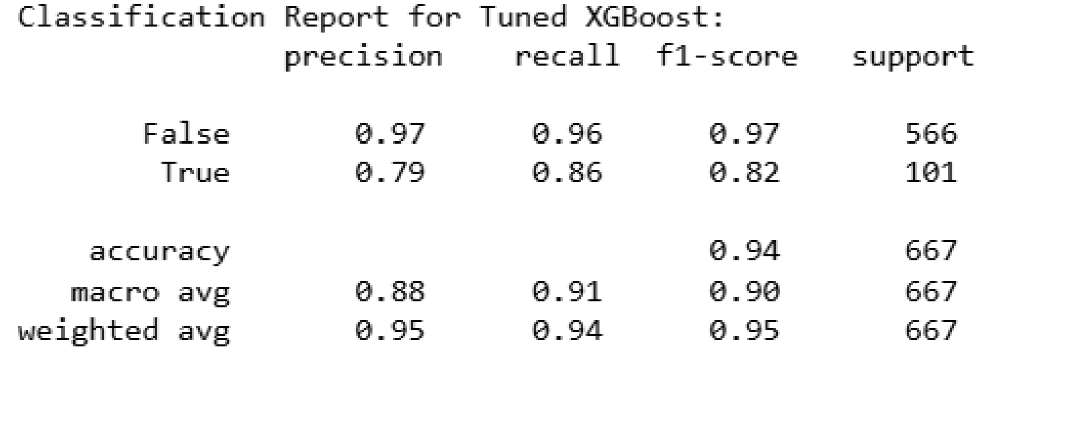
- Performed much better than Decision Tree but still had false positives.
- **Balanced precision and recall**, making it a strong contender.
- **Decision:** Considered but tested against XGBoost for better results.

Final Model - Tuned XGBoost

- Trained an XGBoost classifier with advanced hyperparameter tuning.
- Optimized learning rate, tree depth, and boosting rounds.

Results:

- **Accuracy:** 94%
- **Precision (Churned Customers):** 79%
- **Recall:** 86%
- **F1-score:** 82%



```
Classification Report for Tuned XGBoost:
              precision    recall  f1-score   support

       False       0.97      0.96      0.97       566
        True       0.79      0.86      0.82       101

    accuracy                           0.94       667
   macro avg       0.88      0.91      0.90       667
weighted avg       0.95      0.94      0.95       667
```

Conclusion:

- Outperformed all previous models.
- Best trade-off between false positives and false negatives.
- **Decision:** Selected as the final model for deployment.

EVALUATION

- **ROC-AUC Curve:** Both Random Forest and XGBoost achieved **0.91**, indicating strong classification ability.
- **Precision-Recall Curve:** XGBoost maintained better precision at higher recall values, making it more reliable for churn detection.
- **Confusion Matrix:** XGBoost minimized false negatives while keeping false positives under control, ensuring a balanced model.
- **Feature Importance:** Customer service calls, total charges, and international plans were the strongest indicators of churn.

Final Verdict

- XGBoost was selected as the final model due to its best balance of precision, recall, and overall performance.
- It identified more actual churners (recall: 86%) while keeping false churn predictions lower than Random Forest.
- This model is suitable for real-world deployment to help SyriaTel proactively retain customers.

RECOMMENDATIONS

- Enhance Customer Retention Strategies by proactively identifying high-risk customers with high total charges and frequent customer service calls, offering personalized incentives and discounts to retain them.
- Optimize Service and Pricing Plans by addressing customer dissatisfaction linked to international and voicemail plans, improving service quality, and introducing competitive pricing adjustments.
- Implement Data-Driven Churn Prevention by deploying the XGBoost model for real-time churn prediction, enabling proactive intervention before customers leave.
- Refine and Expand Predictive Capabilities by optimizing the decision threshold for better precision-recall balance and integrating additional data sources like customer satisfaction and competitor offers for improved accuracy.

NEXT STEPS

- Deploy the churn prediction model within SyriaTel’s customer management system.-
- Monitor model performance and adjust the decision threshold as needed.-
- Expand data collection by incorporating customer satisfaction scores and competitor pricing insights
- Refine retention strategies based on ongoing model predictions.

THANK YOU

Any questions?