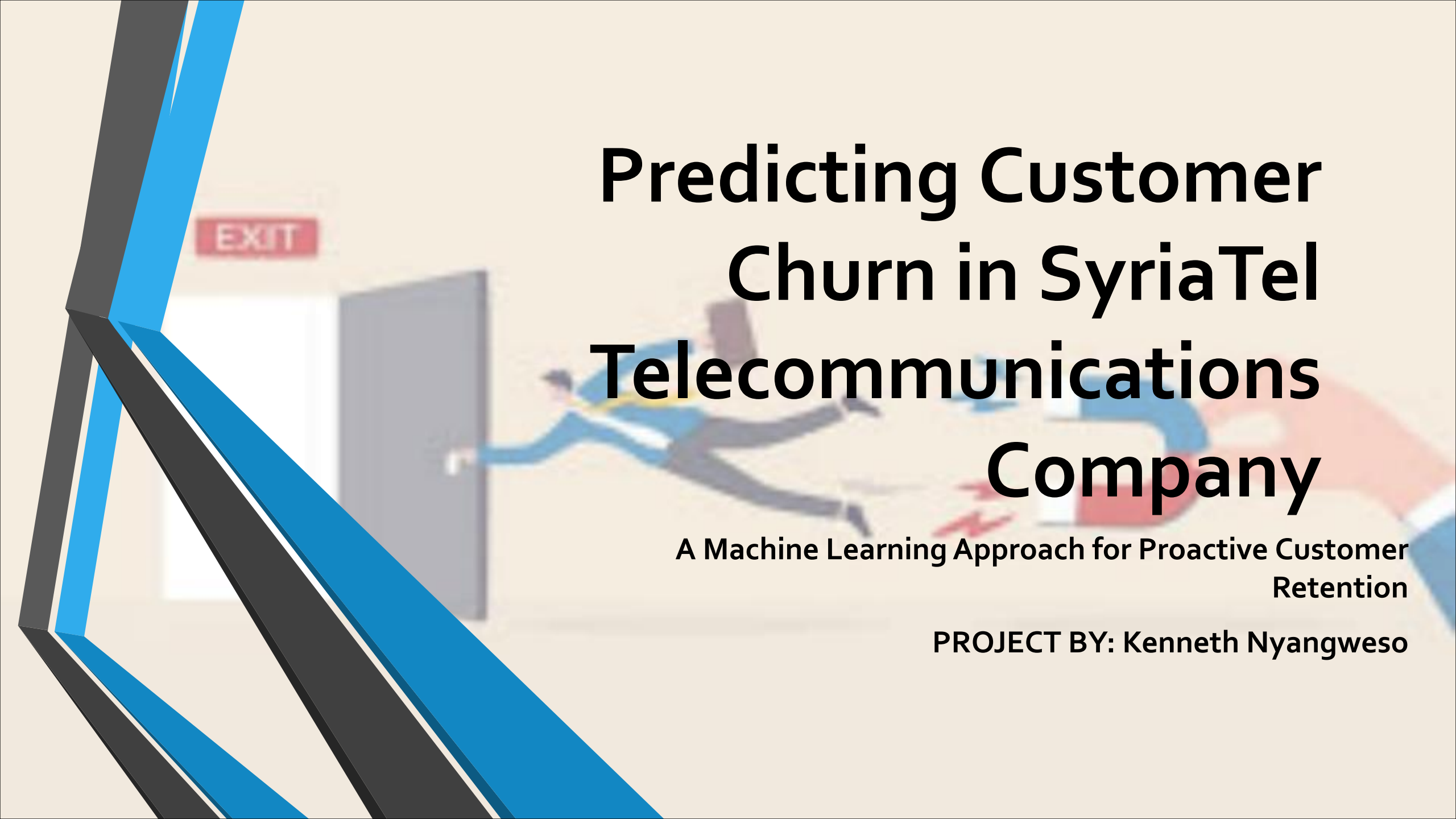


Predicting Customer Churn in SyriaTel Telecommunications Company

A Machine Learning Approach for Proactive Customer Retention

PROJECT BY: Kenneth Nyangweso

1. BUSINESS UNDERSTANDING

PROJECT OVERVIEW:

Customer retention is a critical concern for telecommunications companies, as acquiring new customers is often more expensive than retaining existing ones. With increasing competition, understanding why customers leave (churn) and identifying those at risk can significantly improve business strategies. This project aims to build a binary classification model to predict customer churn based on various usage patterns and customer service interactions.

Problem statement:

SyriaTel, like many telecommunication providers, faces challenges in retaining customers due to factors such service dissatisfaction, competitive offers, shifting customer needs, and many other factors. Identifying customers likely to churn before they actually leave allows the company to take proactive measures such as personalized offers, improved service quality, and better customer engagement. Without a predictive approach, customer retention efforts may be reactive and less effective, leading to revenue losses.

Objectives:

Main Objective:

- To develop machine learning models that accurately predicts customer churn in order to help SyriaTel reduce churn rates and improve customer satisfaction.

Specific Objectives:

- Data Exploration & Cleaning – Understand the dataset structure, handle missing values, and preprocess data for analysis.
- Feature Engineering & Selection – Identify key factors contributing to churn and optimize input variables for modeling.
- Model Development – Train and evaluate multiple classification models to determine the best-performing one.
- Interpretability & Insights – Analyze the key drivers of churn and provide business recommendations.

Why Now?

- **Increased Competition:** Telecommunications companies are experiencing higher competition, making customer retention more crucial than ever.
- **Data Availability:** SyriaTel has rich historical customer data that can be leveraged for predictive modeling.
- **Cost Efficiency:** Predicting churn allows for targeted interventions, reducing marketing costs while improving customer loyalty.
- **Advancements in AI & Machine Learning:** Modern machine learning techniques make churn prediction more accurate and actionable than traditional rule-based approaches.

Metrics Of Success:

- **Accuracy & Precision:** Ensuring the model correctly classifies customers into churn and non-churn categories.
- **Recall (Sensitivity):** High recall ensures that most potential churners are identified.
- **F1-score:** Balancing precision and recall to optimize performance.
- **ROC-AUC Score:** Measuring the model's ability to distinguish between churn and non-churn customers.
- **Business Impact:** Reduction in churn rates and improved customer engagement based on model-driven interventions.

2.DATA UNDERSTANDING:

The SyriaTel Customer Churn Dataset contains information about customers, their usage patterns, service subscriptions, and interactions with customer service. The dataset is structured with 3,333 records and 21 features, including the target variable (Churn), which indicates whether a customer has left the service or not.

3. DATA PREPARATION:

This phase involves transforming raw data into a structured and clean format suitable for modeling. This step ensures that the dataset is free from inconsistencies, missing values, and unnecessary variables while preparing it for machine learning algorithms.

Steps:

- Data cleaning
- Feature Engineering
- Exploratory Data Analysis
- Data preprocessing

Python Libraries used:

- Pandas - Used for data manipulation and analysis
- Numpy- Used for multi-dimensional arrays e.g. matrices
- Matplotlib-Used for interactive visualizations
- Seaborn-Used for advanced visualizations
- Scikit-learn- Used for preprocessing, modeling and evaluation

Data Cleaning:

This stage involved cleaning the dataset by:

- Dropping unnecessary columns
- Checking and dealing with missing values
- Dropping the duplicates
- Changing the columns format
- Checking for outliers

Feature Engineering:

Feature engineering is crucial for improving model accuracy and interpretability. In this project, the goal was to extract meaningful insights from the available features and create new ones that enhance the model's predictive power.

The processing involved adding new columns using different techniques which are:

- Using mathematical operations i.e. addition and division.
- Using the if-else conditional statements.

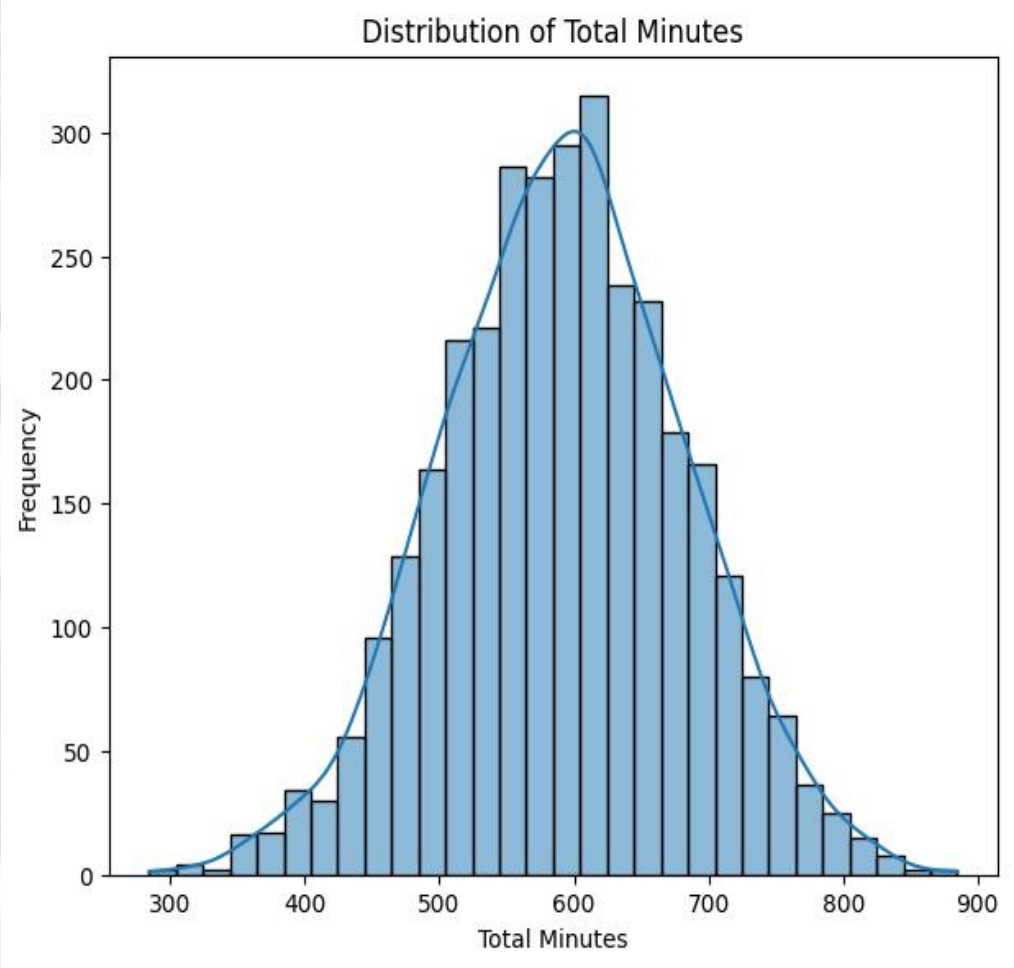
Exploratory Data Analysis (EDA):

This phase involves visualizing the key parameters that will enhance my model performance. These parameters include key features such as; total calls, total minutes, total charges and the target variable which is the churn.

Steps to follow include:

- Univariate Analysis
- Bi-variate Analysis
- Multivariate Analysis

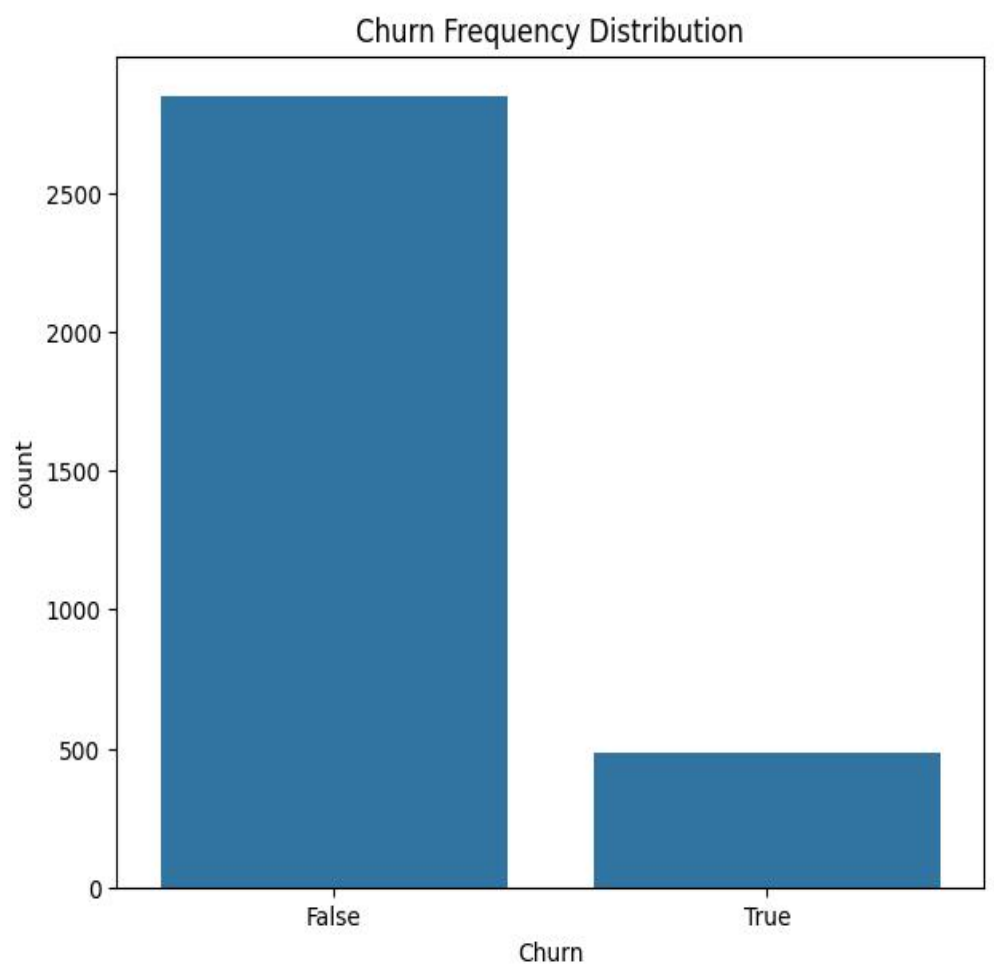
Univariate Analysis: Numeric data



Observations:

- **Normal Distribution:** The histogram and KDE curve suggest that "Total Minutes" is approximately normally distributed. The bell shape is fairly symmetrical, and the KDE curve smooths this out nicely. This is a good sign for many modeling algorithms that assume normality
- **Central Tendency:** The distribution's peak (and thus the mean and median, given the symmetry) appears to be somewhere around 600-650 total minutes. This tells you the "average" total minutes used by your customers.
- **Spread/Variability:** The distribution shows a reasonable amount of spread or variability around the mean. Most customers fall within the range of roughly 400 to 800 total minutes. This spread is important; if the data was too tightly packed, it might mean the feature isn't very discriminative for churn.

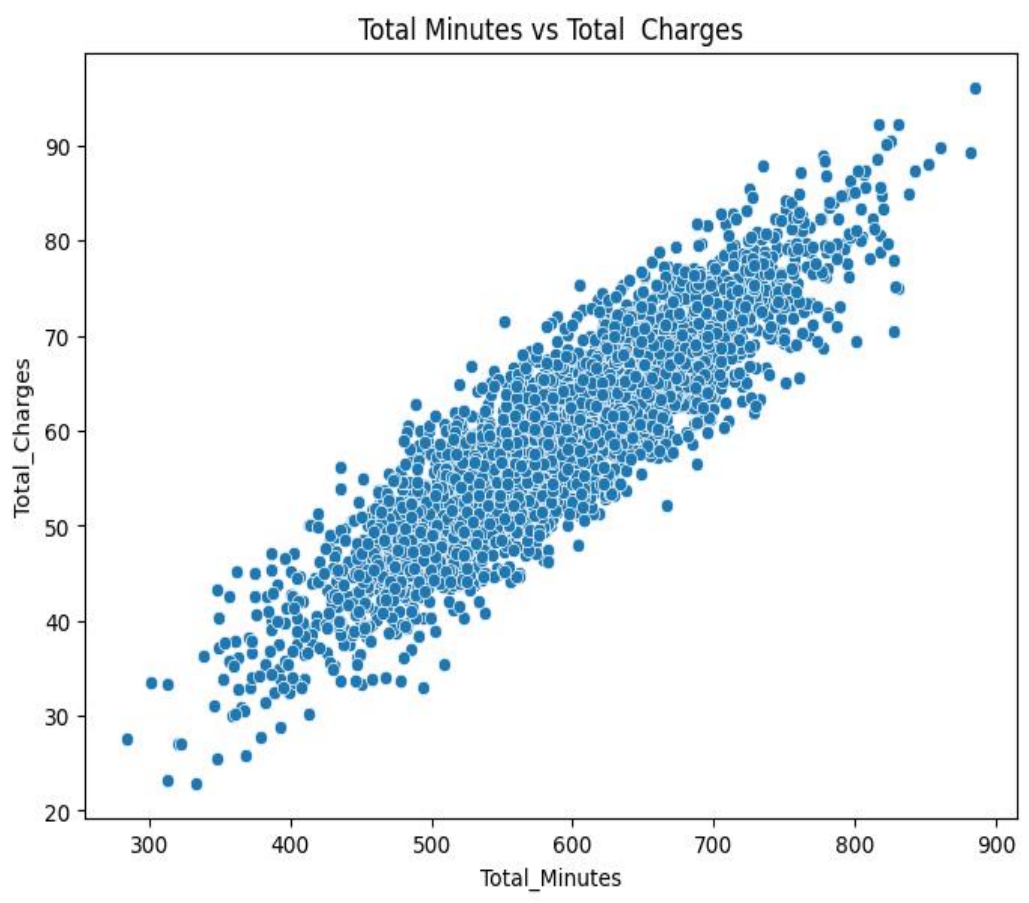
Univariate Analysis: Categorical data



Observations:

- **Class Imbalance :** There is a noticeable class imbalance. The "False" (no churn) category has a substantially higher count than the "True" (churn) category.
- **Majority Class; No Churn:** The taller bar for "False" indicates that the majority of your customers have not churned. This is common in many churn prediction scenarios, as businesses typically have a lower churn rate than their retention rate.
- **Minority Class; Churn:** The shorter bar for "True" shows the proportion of customers who have churned. The exact count isn't explicitly provided, but we can visually infer that it's likely somewhere around 500 out of a total of approximately 3300 customers, suggesting a churn rate in the neighborhood of 15%.

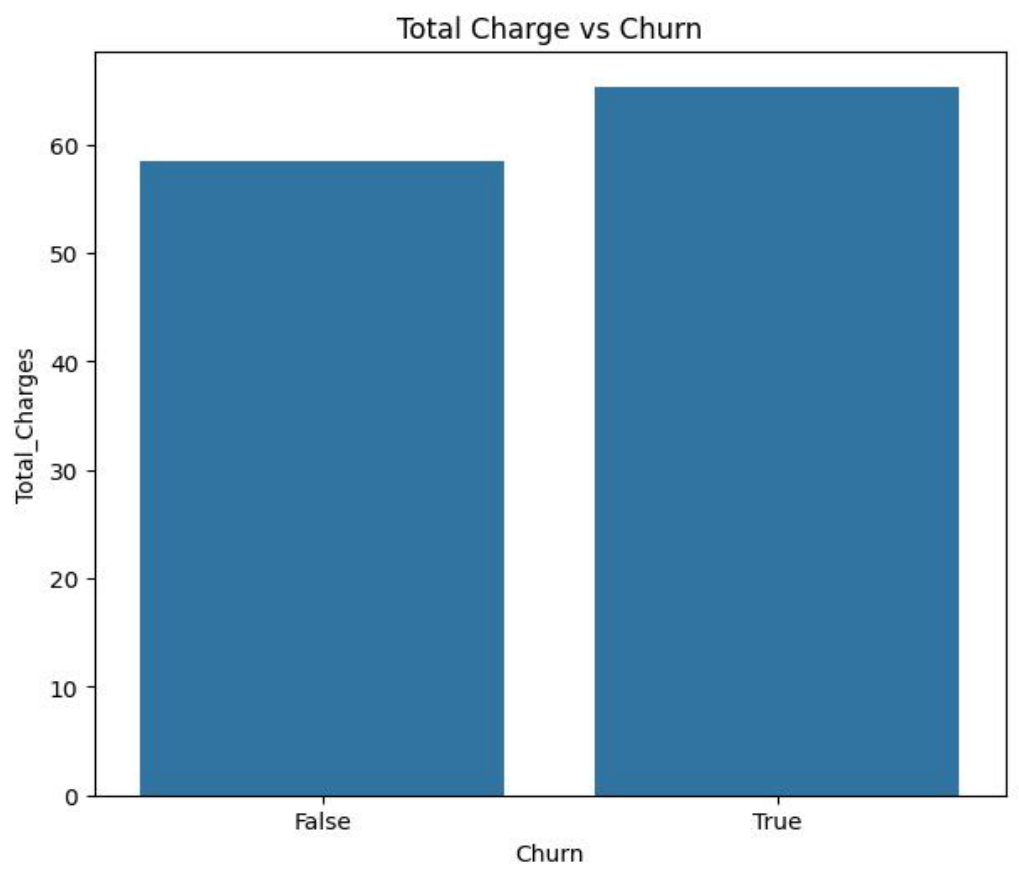
Bi-Variate Analysis: Numeric Vs Numeric



Observations:

- **Strong Positive Correlation:** The scatter plot clearly shows a strong positive correlation between "Total Minutes" and "Total Charges." As "Total Minutes" increases, "Total Charges" also tends to increase. This suggests that customers who use more minutes tend to have higher total charges.
- **Linear Relationship:** The relationship appears to be largely linear. The points roughly fall along a straight line, indicating a consistent proportional increase in charges with increased minutes.
- **No Obvious Clusters:** There aren't any distinct clusters visible in the scatter plot. The points seem to be relatively evenly distributed along the linear trend.

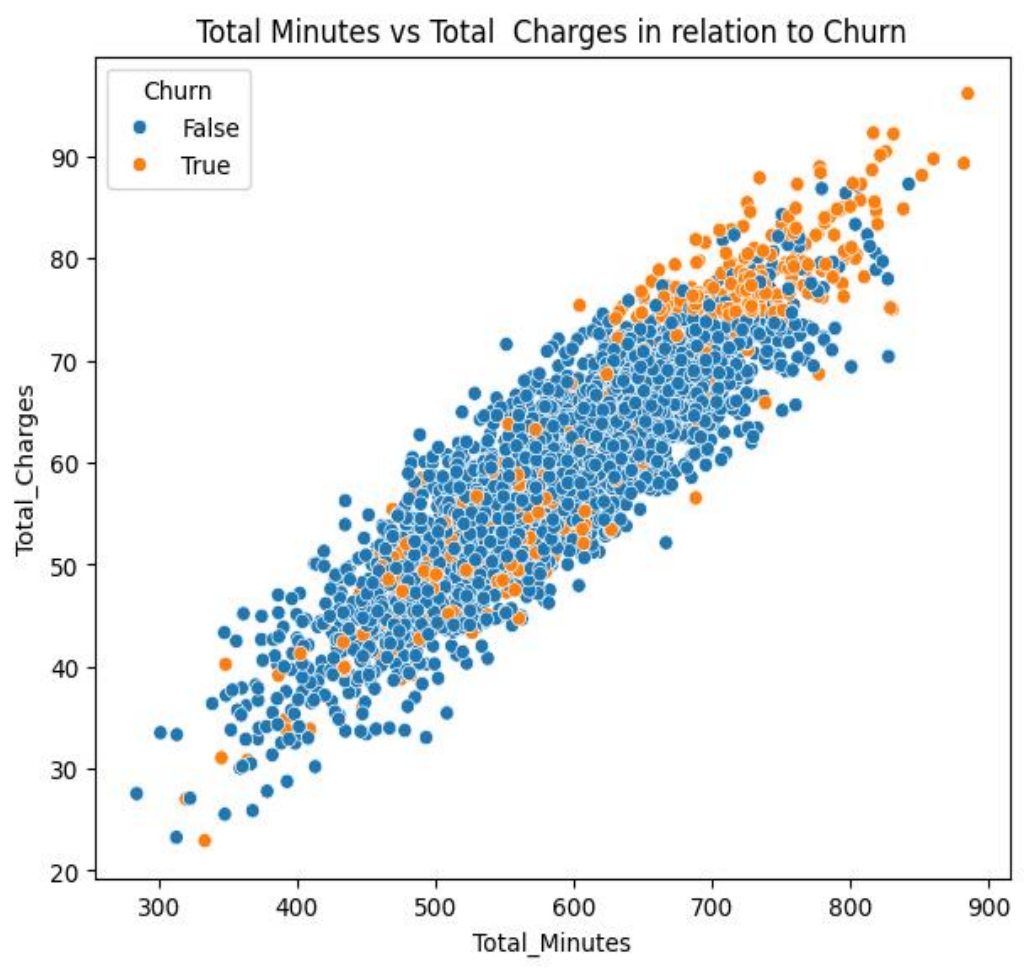
Bi-variate Analysis: Numeric Vs Categorical



Observations:

- **Higher Average Charges for Churned Customers:** The bar chart indicates that, on average, customers who have churned ("True") tend to have higher total charges than customers who haven't churned ("False").
- **Potential for Predictive Power:** This difference in average total charges suggests that "Total Charges" could be a useful feature for predicting churn. Customers with higher total charges might be more likely to churn.

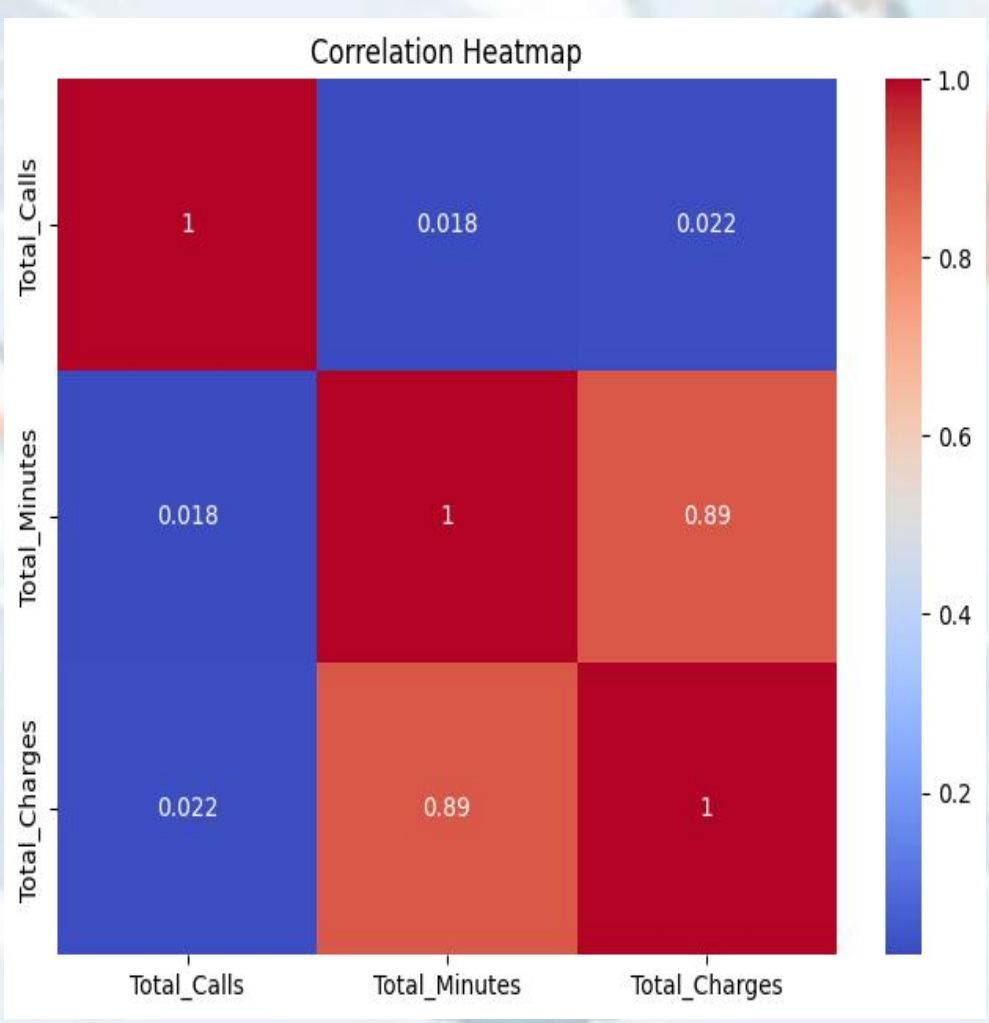
Multivariate Analysis



Observations:

- **Concentration in the Higher Range:** The "True" (churned) points are predominantly concentrated in the upper right region of the scatter plot, corresponding to higher values of both Total Minutes and Total Charges. This indicates that customers with higher minutes and charges are more likely to churn.
- **Overlapping with "False" (Not Churned):** While concentrated in the higher range, the "True" points are not entirely separated from the "False" (not churned) points. There's significant overlap, especially in the middle range of minutes and charges. This suggests that while high minutes and charges increase the likelihood of churn.
- **Spread Across a Range:** The "True" points are spread across a range of minutes and charges, not clustered in a single tight region. This implies that there isn't a single specific threshold of minutes or charges that guarantees churn.

Heat map:



Observations:

- **Strong Positive Correlation Between Total Minutes and Total Charges:** The most prominent feature is the strong positive correlation (0.89) between Total Minutes and Total Charges. This is represented by the intense red color in that cell. It confirms what we saw in the scatter plot: customers who use more minutes tend to have higher charges.
- **Weak Positive Correlation Between Total Calls and Total Minutes/Charges:** The correlations between Total Calls and both Total Minutes (0.018) and Total Charges (0.022) are very weak and positive. The blue color of these cells indicates the near-zero correlation. This aligns with our earlier observation that the number of calls alone doesn't strongly predict total minutes or charges.
- **No Multicollinearity Issues:** While Total Minutes and Total Charges are highly correlated, this is not necessarily a Multicollinearity problem in the typical sense. Multicollinearity usually refers to having multiple predictor variables that are highly correlated. In this case, we are just examining the relationships between key usage metrics, and the high correlation between minutes and charges is expected.

Data Preprocessing:

In this stage, I converted my data into a suitable format for the model. Such as:

- **Encoding Categorical Variables:** Machine learning models typically work with numerical data. This process changes categorical data to numerical.
- **Scaling of the features:** Scaling features is a crucial preprocessing step in many machine learning workflows, ensuring that models perform well, converge faster, and provide meaningful interpretations. However tree-based models (e.g., Decision Trees, Random Forests) are generally less sensitive to the scale of features because they split the data based on feature thresholds rather than distances.

4.MODELING

My main aim was to develop multiple models and assess the best model for this classification problem based on the performance of each model. To achieve this the steps are as follows:

Models to build:

- Logistic regression model
- Decision tree classifier
- Random Forest Classifier
- XGBoost Classifier

To improve some of the models' performance some of the techniques used include:

- Synthetic Minority Oversampling Technique-SMOTE (Handles class imbalance)
- Class weighting with scikit learn(Handles class imbalance)
- Hyper parameter using Grid Search and Randomized Search(Improves the model performance based on the key metrics such as recall)

5. EVALUATION:

For evaluation process, the evaluation metrics are

- **Recall(priority):** It indicates what percentage of the classes to be analyzed were actually captured by the model.
- **F1-score:** Harmonic mean of precision and recall
- **Precision:** Measures how many predicted positives are actually positive
- **Accuracy:** Proportion of correctly classified instances out of the total

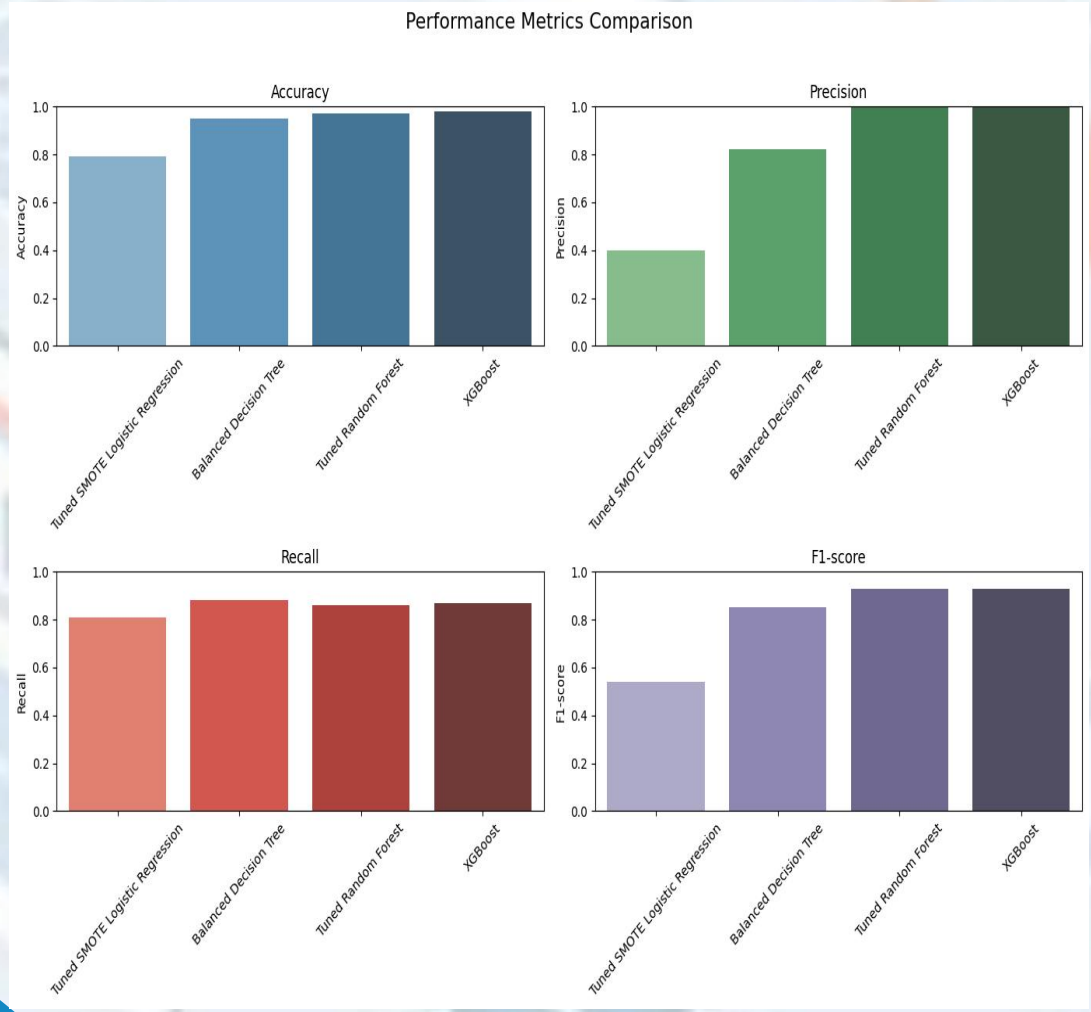
Visualizations to be used for the model analysis:

- **Confusion matrix:** It is a table used to evaluate the performance of a classification model by comparing predicted labels against actual (true) labels. It provides a detailed breakdown of correct and incorrect predictions, making it a valuable tool for understanding model performance.
- **ROC and AUC curves:**

Receiver Operator Characteristic Curve: Illustrates the true positive rate against the false positive rate of the classifier

Area Under the Curve: Singular value summarizing the ROC curve. A higher AUC indicates better performance

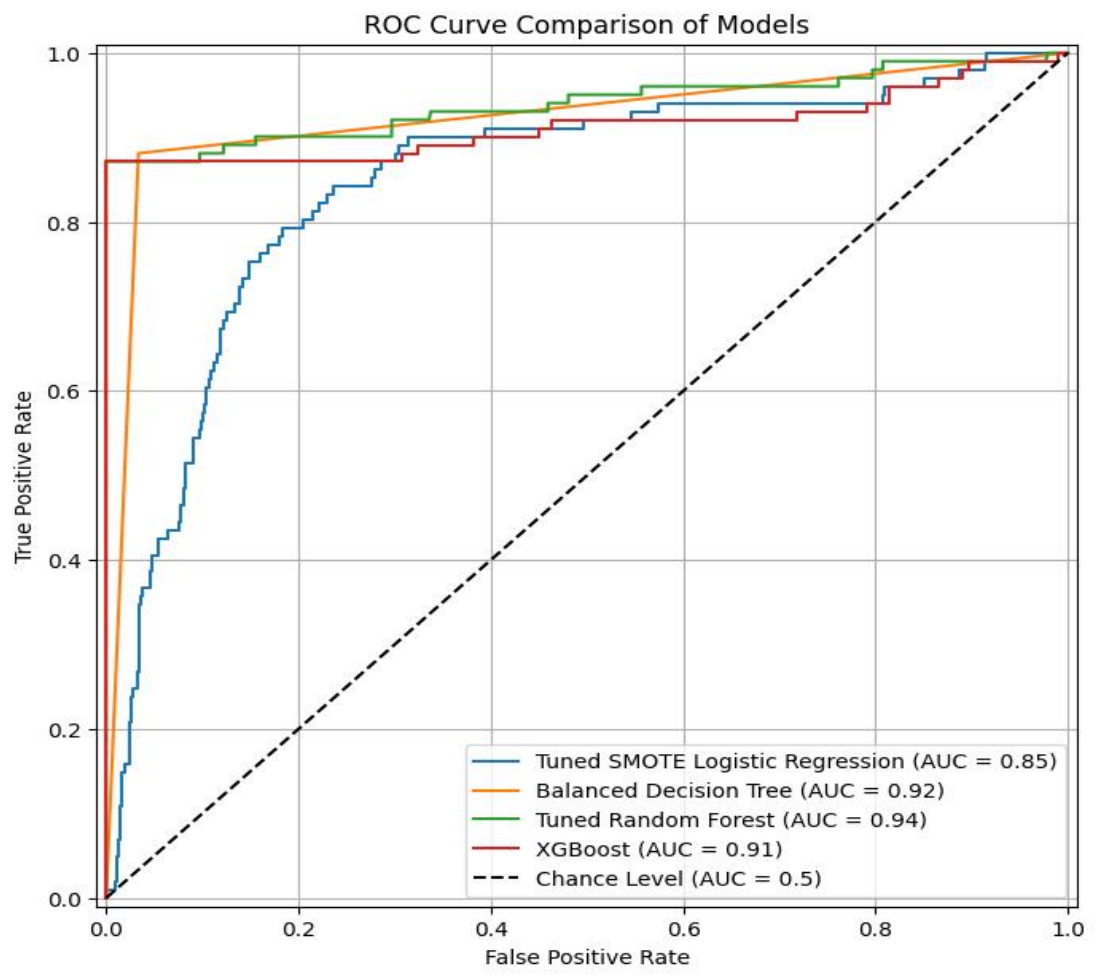
Performance Metrics Comparison:



Analysis of the Performance Metrics Comparison

- **Accuracy:** The Tuned Random Forest and XGBoost models achieve the highest accuracy, significantly outperforming the Tuned SMOTE Logistic Regression model. The Balanced Decision Tree is slightly lower in accuracy but still relatively strong.
- **Precision:** XGBoost and Tuned Random Forest show the highest precision, indicating they make fewer false positive predictions. The Balanced Decision Tree also performs well, but the Tuned SMOTE Logistic Regression lags significantly, meaning it produces more false positives.
- **Recall:** All models have comparable recall, but the Balanced Decision Tree and Tuned Random Forest show a slight edge. High recall means the models are effectively identifying actual churn cases.
- **F1-score:** XGBoost and Tuned Random Forest have the highest F1-scores, balancing both precision and recall. The Balanced Decision Tree follows closely, while the Tuned SMOTE Logistic Regression lags behind due to its poor precision.

ROC-AUC Comparison:



Analysis based on the ROC Curve/AUC curve

Area Under the Curve (AUC) Interpretation:

- **Tuned Random Forest (AUC = 0.94):** Best performing model, indicating a strong ability to distinguish between churners and non-churners.
- **Balanced Decision Tree (AUC = 0.92):** Slightly lower than Random Forest but still very effective.
- **XGBoost (AUC = 0.91):** Close to the Decision Tree and Random Forest, showing strong predictive power.
- **Tuned SMOTE Logistic Regression (AUC = 0.85):** Performs the worst among the models, but still better than random guessing (AUC = 0.5).

Shape of the Curves:

- The Random Forest (green curve) and Balanced Decision Tree (orange curve) maintain the highest true positive rate while minimizing false positives.
- XGBoost (red curve) also follows a strong pattern but is slightly below Random Forest.
- SMOTE Logistic Regression (blue curve) struggles with higher false positive rates compared to the other models.

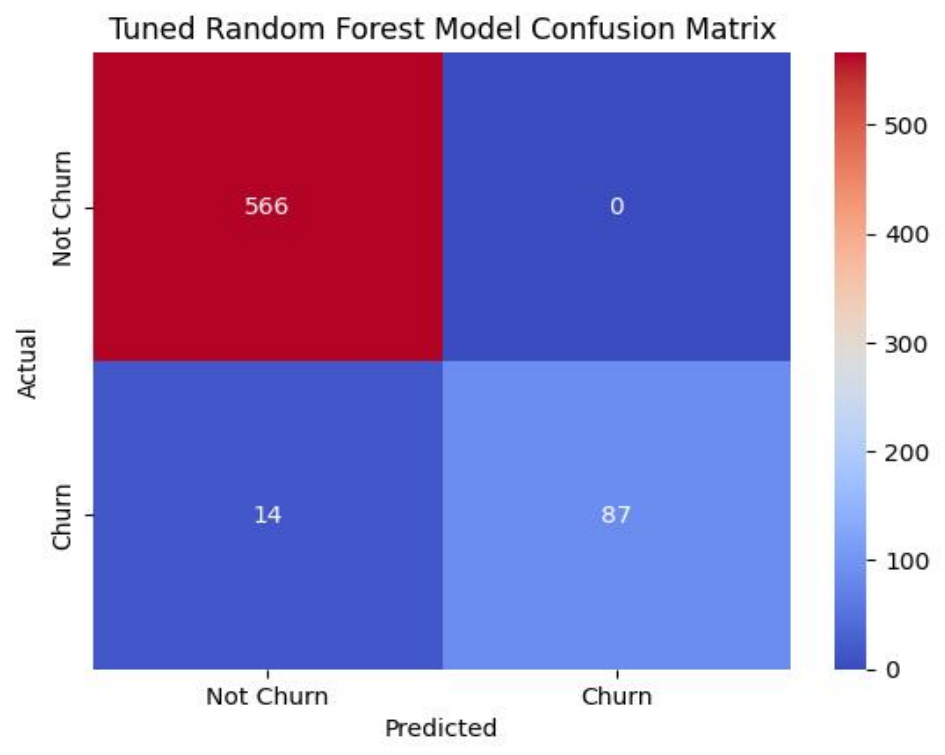
Which model is the most suitable?

- Since our business priority is identifying churned customers, the best model would be the one that maximizes recall while maintaining a good balance with precision. Therefore the best model is the **Tuned Random Forest.**

Reasons for this choice:

- Highest recall (captures the most churned customers) → From the performance metrics bar graphs, Tuned Random Forest has one of the highest recall scores.
- Highest AUC (0.94) → This means it effectively differentiates churners from non-churners.
- Good precision → While recall is the main goal, precision ensures that we don't flood the business with too many false churn predictions.

Confusion matrix of the Tuned Random Forest:



Key Observations

- **Excellent True Negatives (TN):** The model perfectly classified all 566 customers who did not churn.
- **High True Positives (TP):** The model correctly identified 87 out of 101 churned customers.
- **Zero False Positives (FP):** The model did not incorrectly predict any customers as "Churned" when they actually did not churn.
- **Low False Negatives (FN):** The model missed only 14 churned customers.

Tuned Random Forest Performance metric Evaluation:

Class	Accuracy	Recall	Precision	F1-score
Not-churned	97.9%	100%	98%	99%
Churned	97.9%	86%	100%	93%

Observations: (Cont..)

1. Accuracy (97.90%)

- This high accuracy indicates that the model correctly classifies most customers as either **churned (1) or not churned (0)**.
- However, accuracy alone is not enough, as class imbalance can sometimes mask underlying issues.

2. Precision

- **For class 0 (Non-Churned):** 0.98
  - Out of all customers predicted as non-churned, 98% were actually non-churned.
- **For class 1 (Churned):** 1.00
  - Every customer predicted as churned was indeed a churned customer.
  - This is critical in a business setting, as we want to minimize false alarms when flagging churn.

(Cont..)

3. Recall (Sensitivity or True Positive Rate)

- **For class 0 (Non-Churned):** 1.00
  - The model identifies all non-churned customers correctly.
- **For class 1 (Churned):** 0.86
  - The model correctly identifies **86% of actual churned customers**.
  - While high, this suggests that **14% of actual churners were not identified**, meaning some churned customers might still be slipping through.

4. F1-Score (Balance Between Precision & Recall)

- **For class 0:** 0.99
- **For class 1:** 0.93
  - The F1-score for churned customers is high, indicating a strong balance between precision and recall.
  - This confirms that the model is **effective at capturing churn while maintaining minimal false positives**.

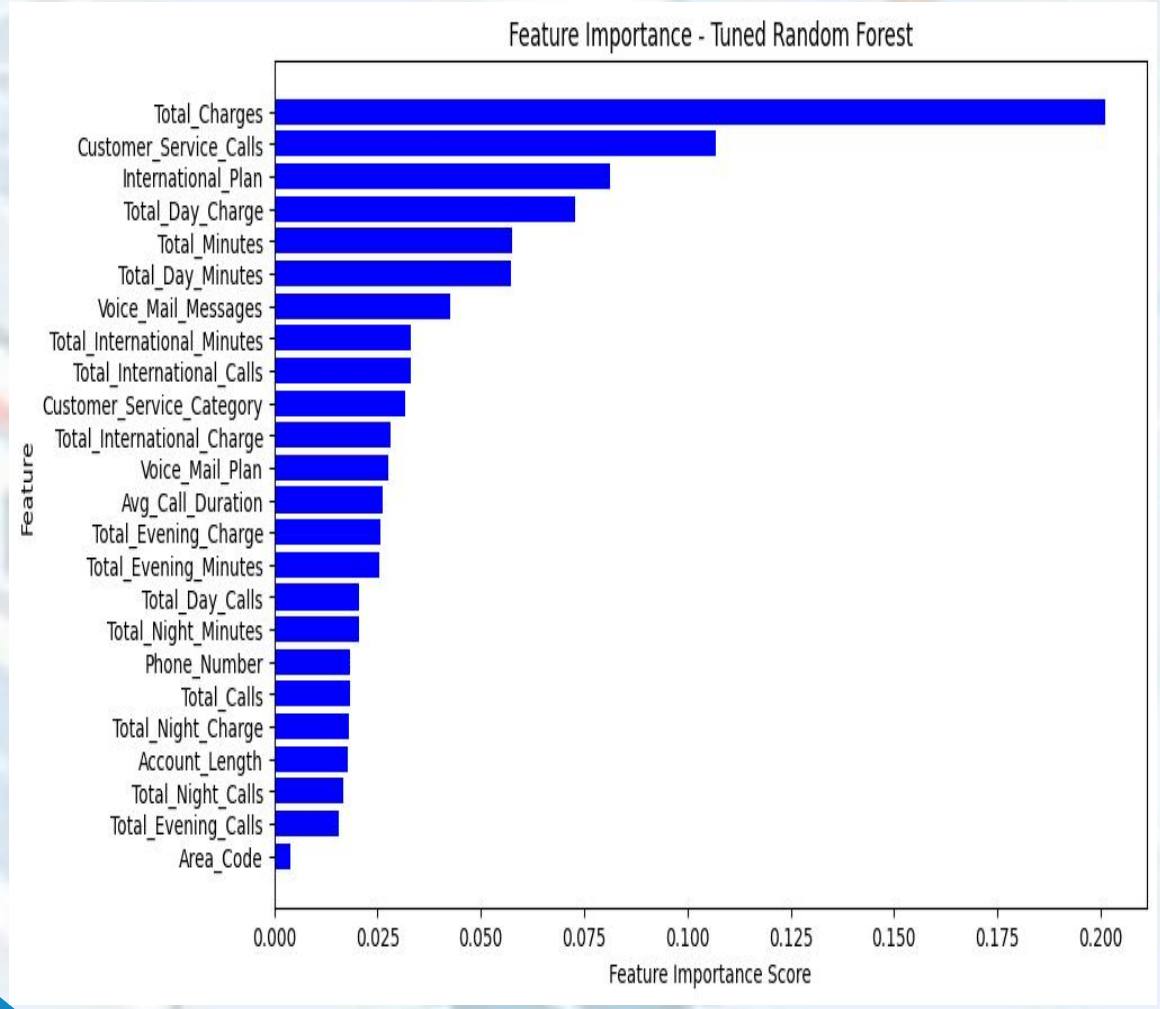
Insights and Business Impact: (Cont..)

- **Reliability:** The high accuracy and precision for both classes make the model dependable for real-world deployment.
- **Churn Detection Strength:** While recall for churned customers is **not perfect (0.86)**, the model is still identifying most churners effectively.
- **Potential Improvement:** To further increase recall for churned customers, additional feature engineering or adjusting the model's decision threshold could be explored.

Conclusion

- The **Tuned Random Forest Model** provides an **excellent balance of precision, recall, and accuracy**, making it a **powerful tool for predicting customer churn**. It is suitable for business deployment, where the goal is to **identify and retain at-risk customers before they leave**.

Random Forest Feature Importance Analysis:



Key Observations

- **Dominant Feature:** Total Charges is overwhelmingly the most important feature, dwarfing all others. This indicates it's the strongest predictor in the model.
- **Moderate Influence:** Customer_Service_Calls, International Plan, Total_Day_Charge, Total Minutes, and Total_Day_Minutes have a moderate level of influence, though considerably less than Total Charges.
- **Low to Negligible Influence:** The remaining features have very low to negligible importance, suggesting minimal impact on the model's predictions.

Business Significance: (Cont..)

1. Total Charges as Primary Driver (Extremely High Business Significance): The overwhelming importance of Total Charges reinforces the understanding that cost is a major factor in churn decisions. This allows the business to:

- **Personalized Retention Offers:** Target high-spending customers with personalized discounts or loyalty programs.
- **Pricing Plan Analysis:** Investigate if specific pricing plans contribute to higher charges and subsequent churn. Restructure plans for better customer satisfaction.

2. Customer Service Interaction (High Business Significance): The influence of Customer_Service_Calls confirms that negative customer service experiences are strongly linked to churn. This enables the business to:

- **Enhance Customer Service Training:** Equip representatives to handle common issues effectively and empathetically.
- **Proactive Customer Service:** Identify customers who frequently contact support and offer proactive assistance.

3. International Plan (High Business Significance): The importance of International Plan suggests specific churn patterns among subscribers. This allows the business to:

- **Optimize International Plan Offerings:** Analyze usage patterns and adjust plans to provide better value and competitiveness.
- **Targeted Promotions:** Offer special promotions or bundles to international plan subscribers.

(Cont..)

3. Total Day Charge and Total Minutes (Moderate Business Significance): The inclusion of Total_Day_Charge and Total Minutes suggests that usage patterns during the day and overall call duration are also relevant churn indicators. This allows the business to:

- **Analyze calling patterns:** Understand if specific usage patterns are associated with churn.
- **Tailor offers based on usage:** Offer packages or promotions that incentivize continued usage.

4.Low Influence Features (Low Business Significance): The low importance of most other features suggests they might not be directly indicative of churn in this model. This could be due to:

- **Redundancy:** Some features might be redundant with Total Charges or other important features.
- **Lack of Direct Impact:** Granular metrics like individual call durations or specific charges might not be as predictive as aggregated measures like Total Charges.

Conclusion: (Cont..)

- **Cost is Key:** The model heavily relies on Total Charges to predict churn, highlighting the importance of cost considerations for customers.
- **Customer Service Matters:** Customer service interactions play a significant role in churn decisions.
- **Usage Patterns are Relevant:** Overall call duration and daytime usage patterns are also indicative of churn.

6.FINAL CONCLUSIONS:

Key Findings:

- **Best Model:** The Tuned Random Forest model achieved the highest recall, ensuring that the majority of churned customers are correctly identified.
- **Feature Importance:** The Total Charges feature, engineered during preprocessing, played a crucial role in predicting churn across all models, highlighting its significance in customer behavior analysis.
- **Evaluation Metrics:** The Tuned Random Forest model demonstrated a strong balance of precision and recall, ensuring effective churn detection while minimizing false positives. It also achieved the highest AUC (0.94) in the ROC curve, confirming its superior ability to distinguish between churners and non-churners.
- **Business Impact:** This model will help SyriaTel identify at-risk customers early, allowing the company to take proactive retention measures such as personalized offers and improved customer service interventions.

7. FINAL RECOMMENDATIONS:

- Based on our findings and the business objective of reducing customer churn, we recommend the following actions:

1. Customer Retention Strategy

- Focus on high total charge customers by offering discounts, loyalty rewards, or personalized service plans to prevent churn. In addition we can also monitor customers with low service usage and engage them through targeted offers and communication.

2. Operational Implementation

- Deploy the Tuned Random Forest Model to predict churn regularly and generate reports for the customer retention team. Automate churn predictions and integrate insights into the company's CRM for proactive interventions.

3. Future Improvements

- Continuously update the model with new data to improve accuracy over time.
- Explore additional customer behavior data (e.g., complaints, network issues) to enhance feature engineering.
- Consider testing hybrid models combining Random Forest and XGBoost for potential performance gains.

- By implementing these recommendations, SyriaTel can leverage machine learning to reduce churn, enhance customer satisfaction, and improve long-term profitability.

NEXT STEPS:

- Deploying the Tuned Random Forest Model to predict churn regularly and generate reports for the customer retention team. Automate churn predictions and integrate insights into the company's CRM for proactive interventions.

QUESTIONS?

ANY QUESTIONS

PERSONAL INFO

For more information reach out to me on:

LinkedIn**: Kenneth Nyangweso**

Email:**kennethnyangweso99@gmail.com**