



Project Title:

**Customer Churn Rate Prediction Model for SyriaTel
Company Using Random Forests, Gradient Boosting
Model**



Outline of the Presentation

- Project Title
- Business Understanding
 - Problem Statement
 - Stakeholder Understanding
- Data Understanding
 - Data Insights
- Data Preparation
- Data Visualization
- Modelling
- Recommendations
- Next Steps



Problem Statement:

- SyriaTel, a leading telecommunications provider, faces significant financial challenges due to customer churn, where customers discontinue their services. Identifying customers at risk of churn is critical for implementing proactive retention strategies and maintaining revenue growth.
- The goal of this project is to develop a predictive classification model that analyzes customer data to determine the likelihood of churn.
- By leveraging historical data, including customer demographics, usage patterns, and service-related metrics, this model aims to uncover actionable insights and predictable patterns of customer behavior. This will enable SyriaTel to focus retention efforts on at-risk customers, minimize churn-related losses, and enhance customer satisfaction, ultimately driving business sustainability.



Stakeholder Understanding

- The Key Stakeholders interest in this project include:
 1. **Senior Management Team:** The CEO and COO, as a stakeholder will benefit from understanding how churn affects revenue and in identifying strategies to retain customers.
 2. **Marketing and Business Development Teams:** Will benefit from the insights of this predictive model by targeting customers at risk of with personalized campaigns to reduce churn.
 3. **Client Service & Support Team:** In Understanding churn trends will enable the team to proactively address customer concerns before they decide to leave, through customer complaints and service issues.
 4. **Customers:** Improved customer retention strategies will lead to better service and tailored offers, for the members satisfaction

Data Understanding: Data Insights

- The SyriaTel dataset contains 3333 rows and 21 columns comprising of; 16 Numerical columns and 4 Categorical columns and 1 Boolean column. The dataset has four data types (integer, float, object and boolean)
- **The Categorical columns includes** ; State, Phone number, International plan, & voice mail plan
- **The Numerical columns includes** ; account length, area code, 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
- **Boolean Target:** churn. Given we are solving classification problem, our target variable is churn- since it is a binary variable.



Data Preparation:

- **No Missing Values and No Duplicate Values:** All columns have complete data.
- The dataset has four data types (integer, float, object and boolean)
- **Boolean Target:** churn. Given we are solving classification problem, our target variable is churn- since it is a binary variable.

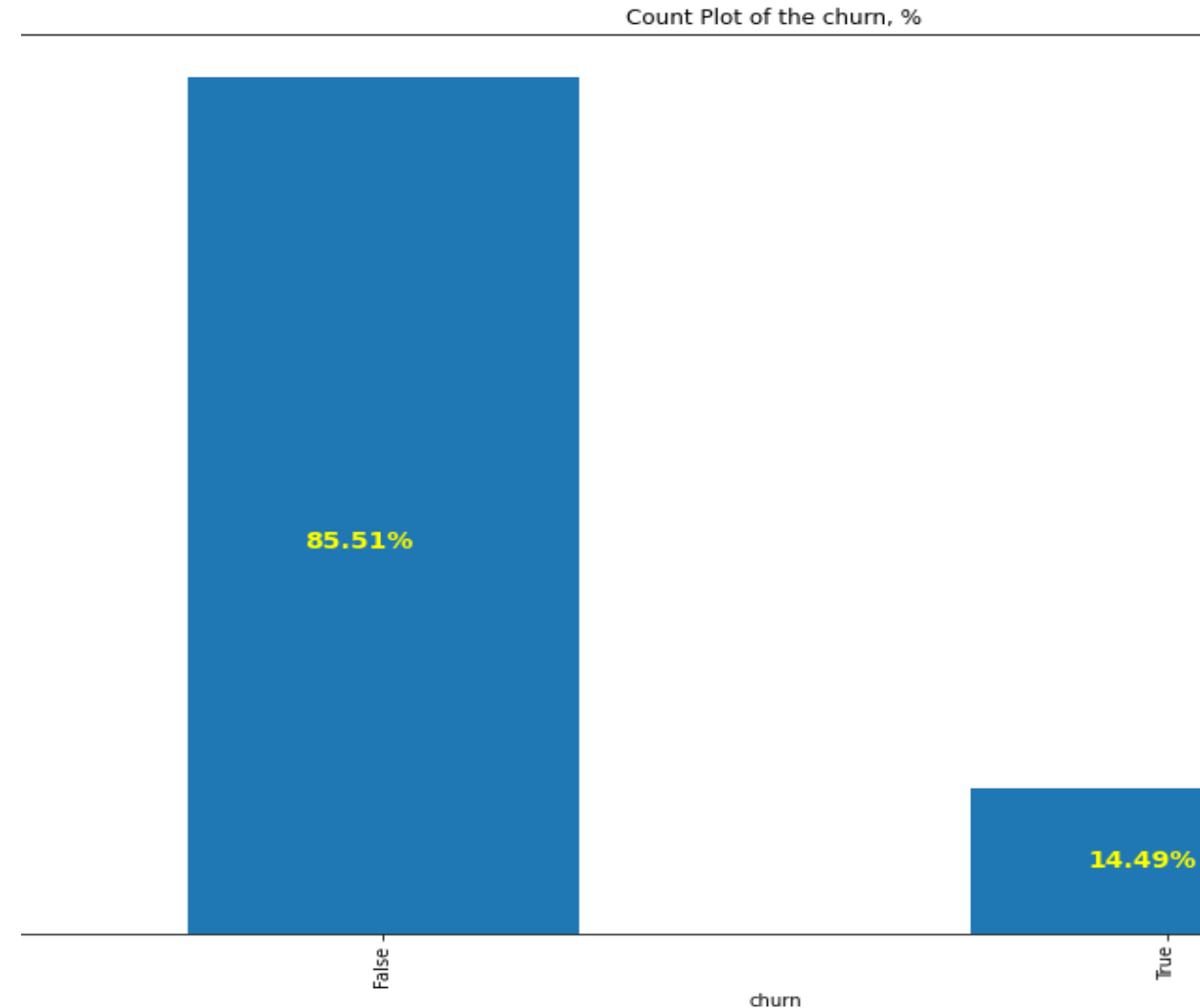


Data Visualization: Categorical Columns Of The Dataset

- **Interpretation :**

- Target Variable Distribution:

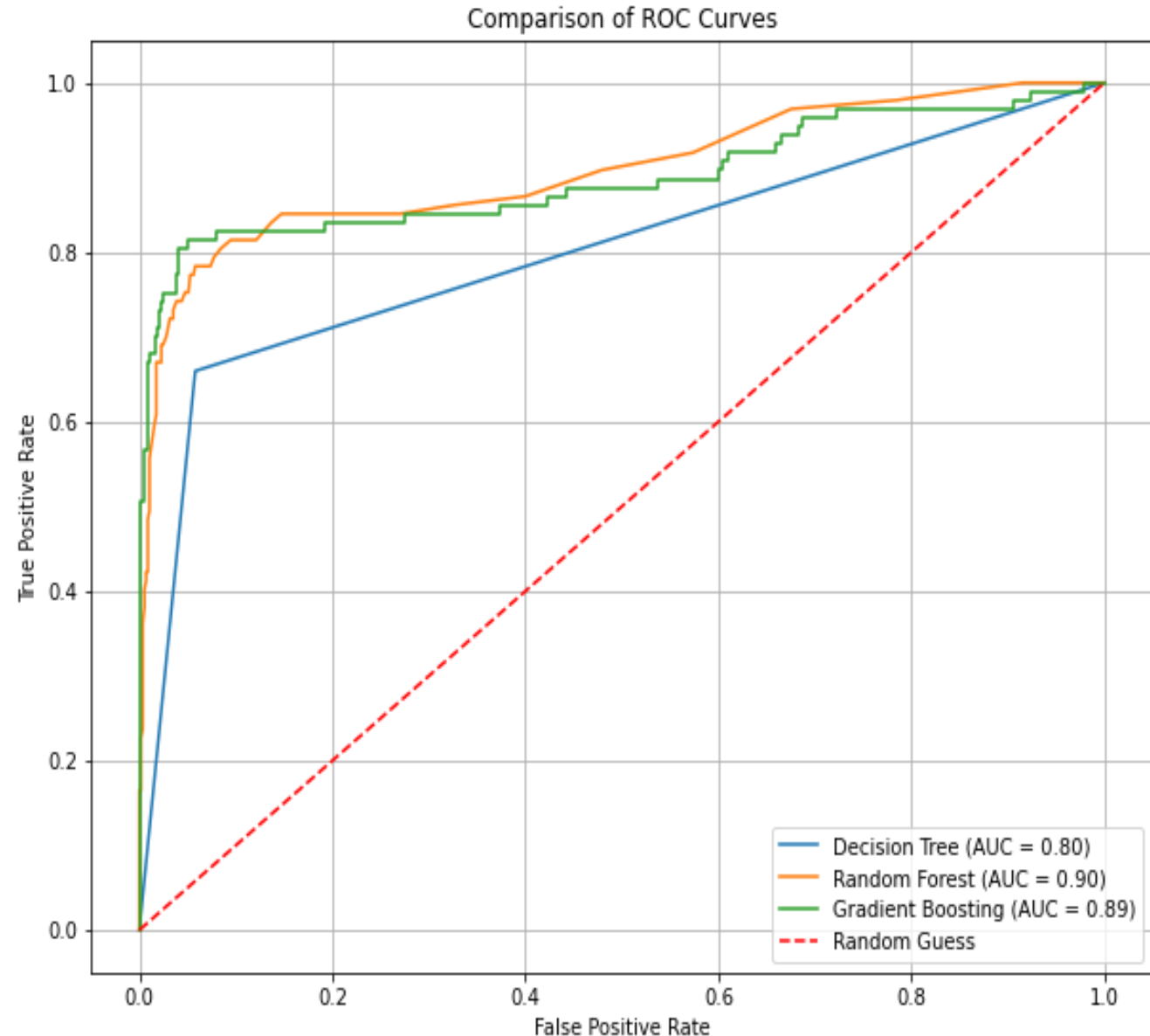
- 85.5% of customers did not churn (False)
WHILE 14.5% of customers churned (True).
- The Count plot of Churn rate of SyriaTel shows dataset is unbalanced with only 14.49% of objects belonging to class 1 (True)



Data Visualization: ROC Curves

- **Insights and Observations:**

- Random Forest performs slightly better than Gradient Boosting for this dataset, reflected by its higher AUC score and its curve being closer to the ideal top-left corner.
- Decision Tree, while simpler, has significantly lower performance due to its inability to generalize well without ensembling techniques.
- Both ensemble models (Random Forest and Gradient Boosting) handle the class imbalance better, with higher recall for the minority class (churn).





Modelling:

- Given the project was to solve a classification problem, the following models were selected for this problem:-
 1. Logistic Regression Model – This was used to build a Baseline model which further Tuned to generate using scaling method to generate Best model
 2. Decision Tree Model -
 3. Random Forest Model – (This was selected as preferred model)
 4. Gradient Boosting Model -

Model Evaluation



Model Evaluation: Interpretation

- **Baseline Logistic Regression Model Performance:**

- Reported overall accuracy of 75%.
- The model achieved an ROC-AUC score of 81%, indicating decent discrimination between churn and non-churn customers.

- **Decision Tree Model:**

- Overall Model Accuracy: 90%
- Macro Average F1-Score: 0.80 (balanced performance across classes).
- The Decision Tree achieved an ROC-AUC score of 0.80, reflecting its ability to differentiate between churn and non-churn customers.

- **Observations:**

- The model performs well for the majority class (False) but struggles with the minority class (True), which is expected due to the tree's tendency to overfit small datasets

- **Random Forest Model:-**

- Overall Accuracy: 94%
- Macro Average F1-Score: 0.87 (better balance across classes compared to the Decision Tree).
- The Random Forest model achieved an ROC-AUC score of 0.90, significantly better than the Decision Tree.

- **Observations:**

- Random Forest improves both precision and recall for the churn class compared to the Decision Tree. The higher ROC-AUC indicates better discrimination between classes.
 - Random Forest benefits from ensembling, which reduces overfitting and enhances minority class performance.



Recommendation :

1. Random Forest Model is the preferred model for predicting churn rate by SyrialTel company based on performance.
2. Thus, SyrialTel management can use this model to predict Churn rate of the customers since it gives highest Overall Accuracy: 94%, F1-Score: 0.87 and a ROC-AUC score of 0.90, significantly better than the Decision Tree.
3. Management should choose this model, because the model improves both precision and recall for the churn class compared to the Decision Tree. The higher ROC-AUC indicates better discrimination between classes.



Next Steps:

- The organization can further improve the model performance by including hyperparameter tuning for both Random Forest and Gradient Boosting to further optimize performance.

Thank You

