SYRIA TEL CHURN ANALYSIS

GENERAL OVERVIEW

- SyriaTel is a telecommunications company that provides mobile and data services to thousands of customers. Like many telecom providers, SyriaTel faces a recurring challenge: customer churn — when customers stop using the company's services and switch to a competitor.
- Churn results in significant losses because acquiring new customers is far more expensive than retaining existing ones. This project aims to use machine learning classification models to help SyriaTel identify customers who are at high risk of leaving.

OUTLINE

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Models Evaluation
- Conclusion
- Recommendation
- Next steps

BUSINESS UNDERSTANDING

Problem Statement

SyriaTel currently does not have an effective system for predicting churn early enough to intervene. Many customers who leave show warning signs — such as increased customer service calls or specific plan choices — but these signals are not consistently tracked or analyzed.

Because of this:

- SyriaTel loses revenue when valuable customers leave.
- The marketing team cannot target retention campaigns effectively.
- Customer service cannot focus on high-risk customers who need support.
- The company lacks data-driven insight into the main reasons behind churn.

Objectives

- Build a classification model to predict customer churn.
- Evaluate model performance using metrics like AUC, recall, and F1-score.
- Identify key factors that contribute to churn.
- Provide actionable recommendations for reducing customer churn.
- Support SyriaTel in improving retention and reducing revenue loss.

DATA UNDERSTANDING

- The dataset consists of customer-level data from SyriaTel, with each record representing an individual customer and their service usage.
- Key variables include call usage patterns, customer service interactions, subscription plans, and churn status.
- Initial exploration focused on understanding the data structure, feature types, and the target variable.
- The analysis revealed class imbalance in churn, with fewer customers churning than staying, which informed later modeling and evaluation decisions.

DATA PREPARATION

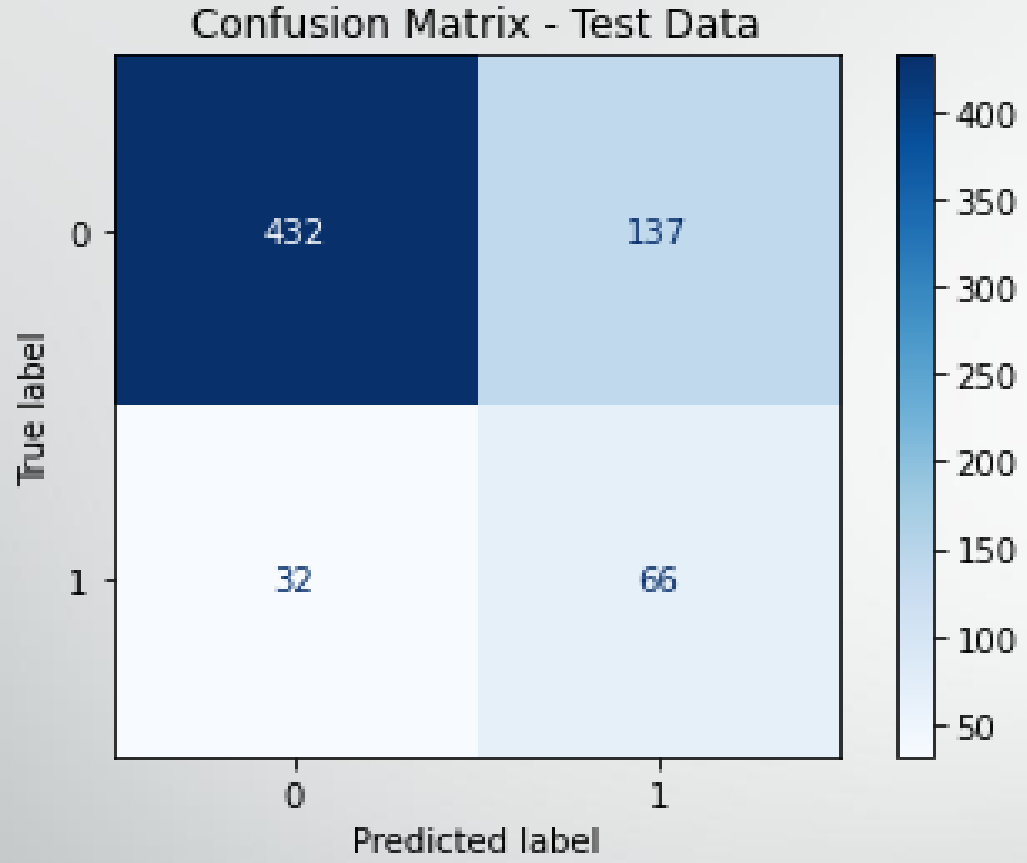
- Data preparation involved cleaning and transforming the dataset to make it suitable for modeling.
- This included checking for missing values and duplicates, encoding categorical variables into numerical form, and scaling numerical features to ensure they were on a comparable range.
- The dataset was then split into training and testing sets. To address class imbalance in the churn variable, resampling techniques were applied to the training data to create a more balanced dataset, improving the model's ability to learn patterns associated with customer churn.

MODELING

In this stage, machine learning classification models were built to predict customer churn based on customer behavior and service usage patterns. Multiple models were trained and evaluated to compare their performance and identify the most effective approach for accurately distinguishing between churners and non-churners.

1.LOGISTICS REGRESSION MODEL

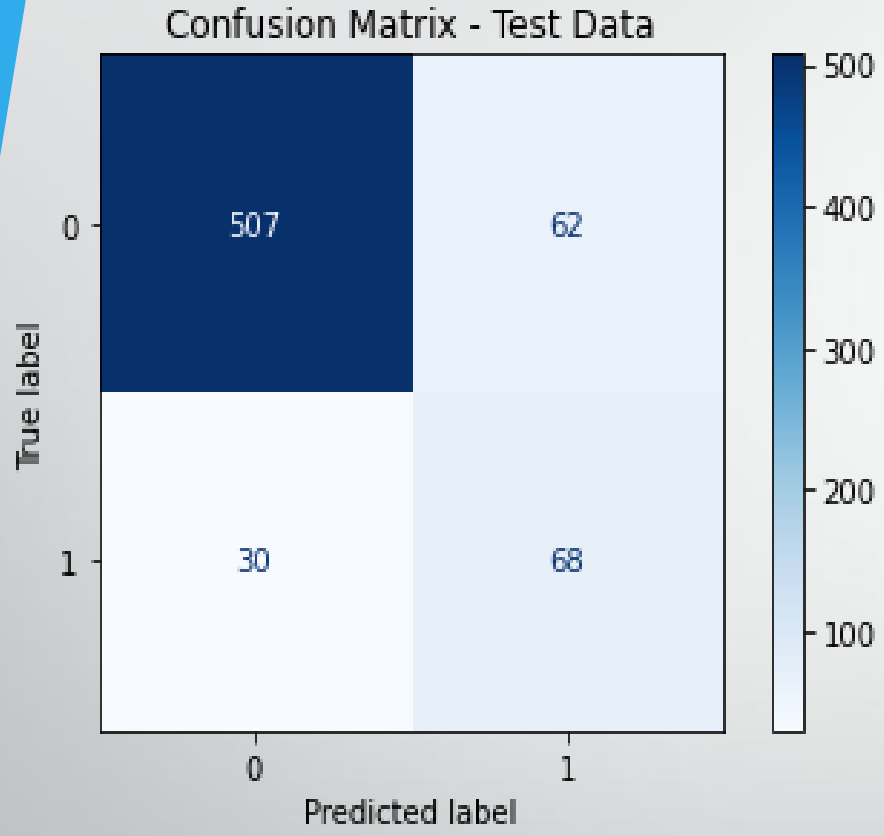
The model demonstrated strong learning performance on the training data, achieving balanced precision, recall, and F1-score of approximately 0.77. When evaluated on unseen test data, recall remains relatively high at 0.67, indicating the model's ability to identify customers at risk of churning. However, precision drops to 0.33, suggesting an increase in false positive churn predictions. This trade-off reflects a recall-focused model that prioritizes identifying potential churners over minimizing false alarms, which may be suitable depending on the cost of customer retention strategies.



The confusion matrix shows that the model performs well at identifying non-churners, correctly classifying most of them (432 true negatives), but it struggles with precision when predicting churners. Although it successfully detects a good proportion of actual churners (66 true positives, about 67% recall), it also incorrectly labels many non-churners as churners (137 false positives), which lowers precision. Overall, the model prioritizes catching churners rather than being very accurate about who will churn, making it useful for retention strategies where missing a churner is more costly than contacting a customer who would not churn.

2.DECISION TREE MODEL

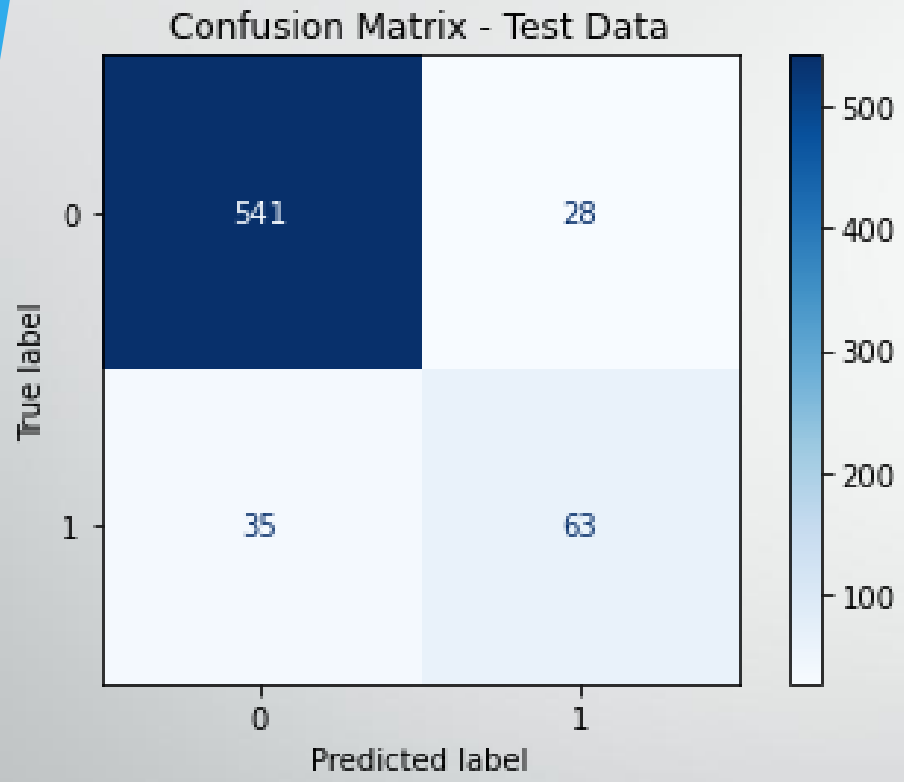
The Decision Tree model demonstrated strong predictive performance, achieving high accuracy and substantially improved churn detection compared to previous models. It correctly identifies most non-churners and captures approximately 69% of actual churners, making it effective for customer retention efforts. However, the model exhibits signs of overfitting, as evidenced by perfect training scores, suggesting that pruning or limiting tree depth would be necessary to improve generalization to unseen data



This confusion matrix shows that the model performs very well at identifying non-churners, correctly classifying 507 customers who did not churn, with relatively few false alarms (62 false positives). It also identifies a good proportion of actual churners, correctly predicting 68 churn cases, while missing 30 churners. Overall, the model achieves a strong balance between catching customers at risk of churning and minimizing unnecessary churn predictions, making it more reliable and cost-effective for churn prediction than the earlier model.

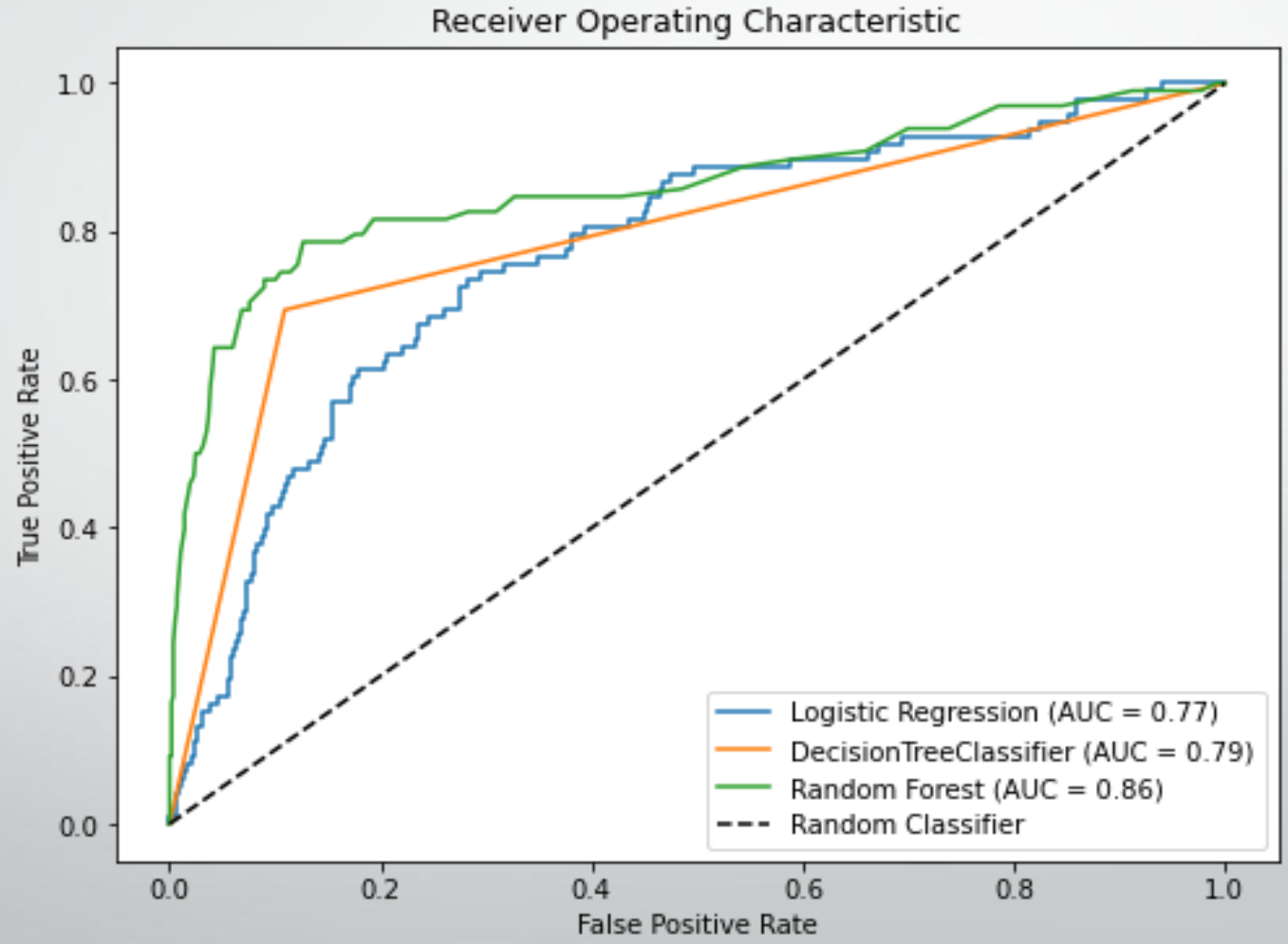
3.RANDOM FOREST MODEL

The Random Forest model demonstrated strong and well-balanced performance, achieving a high test accuracy of 91%. It effectively identifies non-churners while also improving churn detection, correctly capturing 64% of actual churners with a precision of 69%. Although the model fits the training data perfectly, its strong test results indicate good generalization and reduced overfitting compared to simpler tree-based models. Overall, this model provides the most reliable and business-useful churn predictions among the models evaluated.



This confusion matrix shows that the model performs very strongly overall, especially in identifying non-churners. It correctly classifies 541 non-churn customers with very few false positives (28), indicating high precision for the non-churn class. The model also identifies a good number of churners (63 true positives), though it still misses some (35 false negatives). Overall, the model achieves a strong balance between minimizing false alarms and correctly detecting customers at risk of churning, making it well-suited for churn prediction tasks.

MODELS EVALUATION

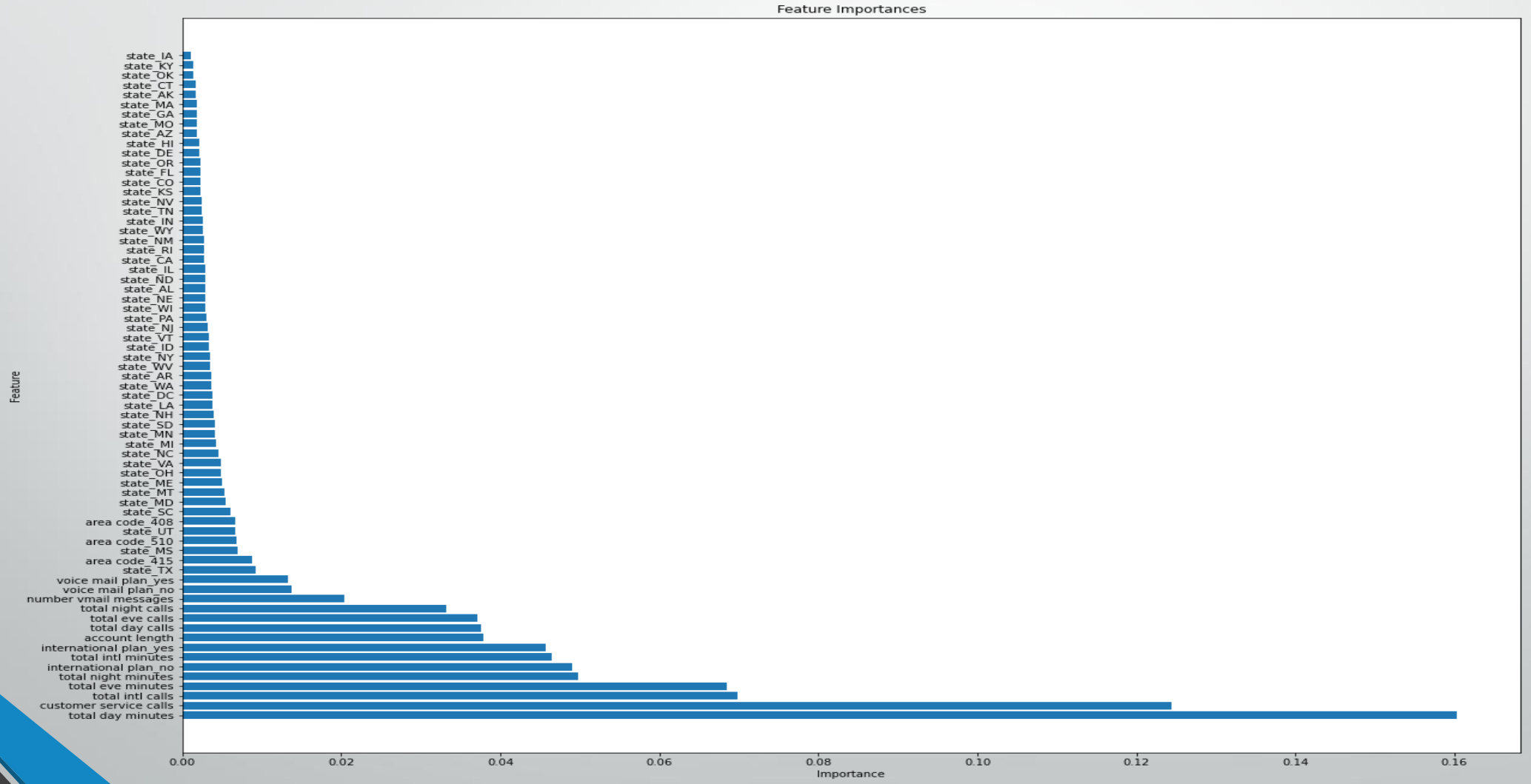


- The ROC curve shows that the Random Forest model performs best, with the highest AUC (0.86), indicating the strongest ability to distinguish between churners and non-churners. The Decision Tree performs moderately well (AUC = 0.79), while Logistic Regression has the lowest performance (AUC = 0.77). Overall, Random Forest is the most reliable model among the three.
- From further evaluations, the Random Forest model is found the most appropriate choice because of its high accuracy, F1 score, It obtained an accuracy of 94.685% and an F1 score of 0.841, demonstrating its ability to effectively identify occurrences while maintaining a balance of precision and recall.

Hyperparameter tuning

- GridSearchCV was used to tune the Random Forest model by testing **16 parameter combinations** with **5-fold cross-validation** (80 total fits).
- The best model selected uses **100 trees** with no depth restriction and minimal split constraints.
- After tuning, the model achieved **91% test accuracy** with improved churn detection (**63% recall, 70% precision** for churners).
- Although the model fits the training data perfectly, its strong test performance indicates **good generalization**.
- The tuned Random Forest provides the **best balance between accuracy and churn prediction**, making it suitable for deployment.

Important features



- Based on the Random Forest feature importance analysis, customer churn is mainly influenced by usage intensity and service experience.
- **Total day, evening, and night minutes** indicate highly engaged customers who are sensitive to pricing and service quality.
- A high number of **customer service calls** strongly signals dissatisfaction and unresolved issues, increasing churn risk.
- Additionally, customers with many **international calls** are more likely to churn due to higher costs and sensitivity to international pricing.

CONCLUSION

- This project successfully developed a classification model to predict customer churn for SyriaTel and evaluated multiple models using accuracy, recall, F1-score, and ROC–AUC.
- The Random Forest model emerged as the best-performing model, demonstrating strong predictive power and good generalization.
- Feature importance analysis revealed that customer churn is largely driven by usage intensity and service experience, particularly total day, evening, and night minutes, the number of customer service calls, and international call activity.
- These findings indicate that highly active customers and those frequently interacting with customer support are more sensitive to pricing, service quality, and unresolved issues, making them more likely to churn.

RECOMMENDATIONS

- Improve Customer Service Resolution: Prioritize resolving customer issues on the first interaction, especially for customers with frequent service calls, as repeated contact is a strong churn indicator.
- Target High-Usage Customers: Monitor customers with high day, evening, and night call minutes and proactively offer loyalty incentives, improved service quality, or personalized plans to reduce churn risk.
- Review International Call Pricing: Reassess international call rates and consider bundled or discounted international plans to retain customers with high international call usage.
- Proactive Retention Strategies: Use the churn model to identify high-risk customers early and engage them with targeted retention campaigns before they decide to leave.
- Continuous Model Monitoring: Regularly retrain and evaluate the model using new data to ensure it remains accurate as customer behavior and market conditions change.

NEXT STEPS

- Deploy the churn prediction model to regularly identify customers at high risk of churning.
- Flag high-risk customers and prioritize them for proactive retention efforts.
- Implement targeted retention campaigns for customers with high usage and frequent customer service interactions.
- Monitor the impact of these interventions on churn rates and customer satisfaction.
- Continuously update and retrain the model to maintain accuracy as customer behavior changes.

THANK YOU

Email: nancymuriithi925@gmail.com

Github: @Nancy_Muriithi