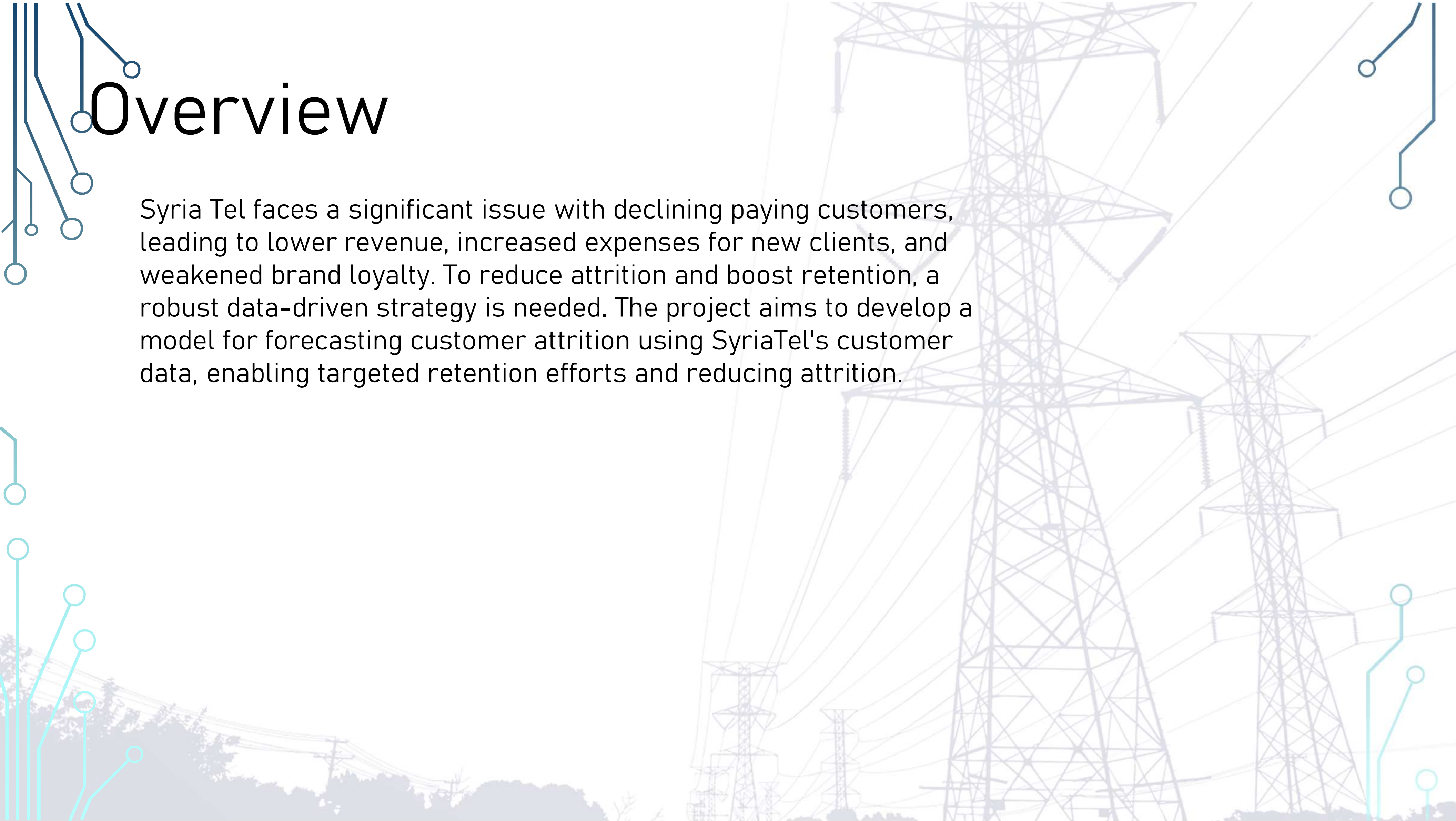
Syria Tel : consumer churn data analysis

BY: EUGENE MARIUS

Overview

Syria Tel faces a significant issue with declining paying customers, leading to lower revenue, increased expenses for new clients, and weakened brand loyalty. To reduce attrition and boost retention, a robust data-driven strategy is needed. The project aims to develop a model for forecasting customer attrition using SyriaTel's customer data, enabling targeted retention efforts and reducing attrition.



BUSINESS & DATA UNDERSTANDING

The project aims to forecast customer churn, focus on high-risk consumers for retention, increase customer lifetime value, enhance client experience, and use model insights for customized marketing campaigns. SyriaTel's CRM system will be used to gather data on user demographics, account information, call usage trends, and customer service interactions. By anticipating problems and understanding customer behavior, businesses can build loyalty and improve overall customer experience. Data-driven decision making guides strategic retention initiatives and resource allocation..

BUSINESS & DATA UNDERSTANDING

SyriaTel CRM Dataset Analysis:

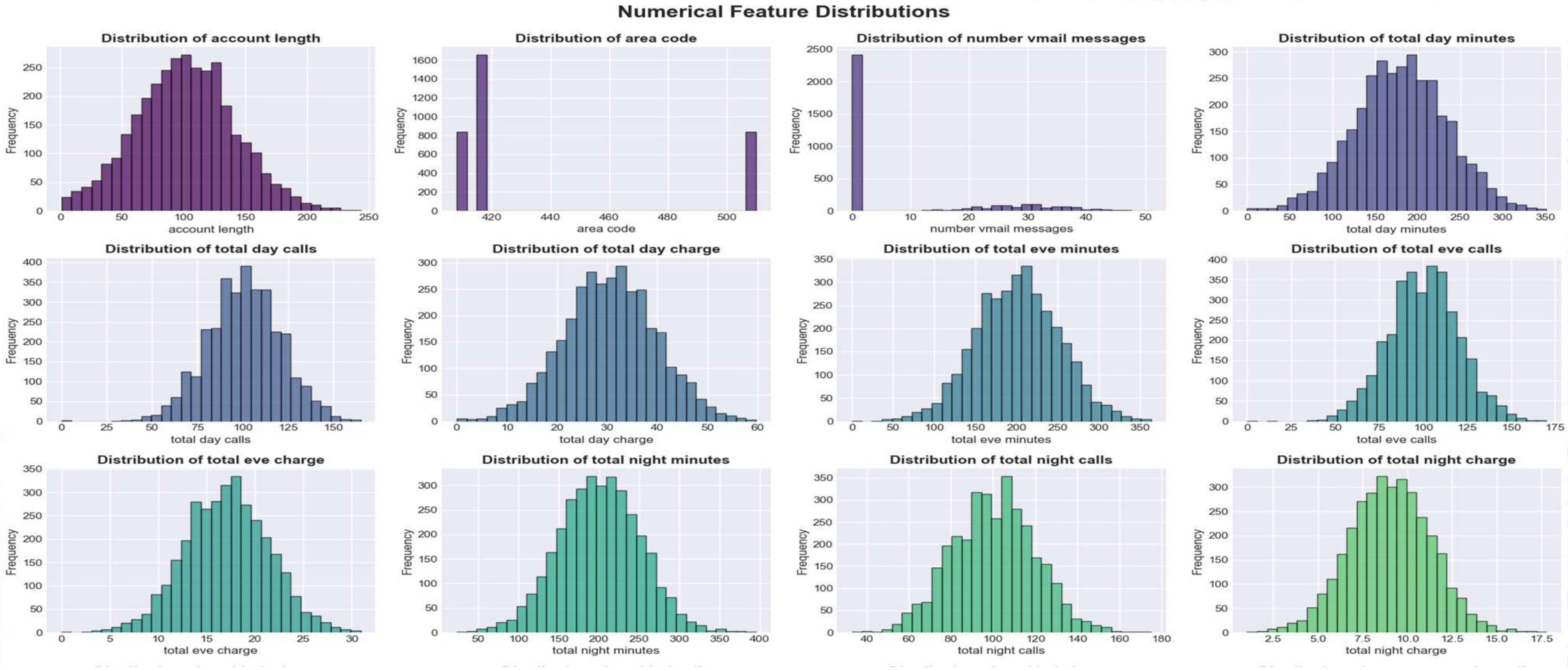
· The dataset provides historical customer data including demographiiinformationcs, account, phone usage trends, and customer service exchanges.

· Features include state of residence, account length, service plans, call usage metrics, and customer service interactions.

· Potential issues include missing values, outliers, and data discrepancies.

BUSINESS & DATA UNDERSTANDING

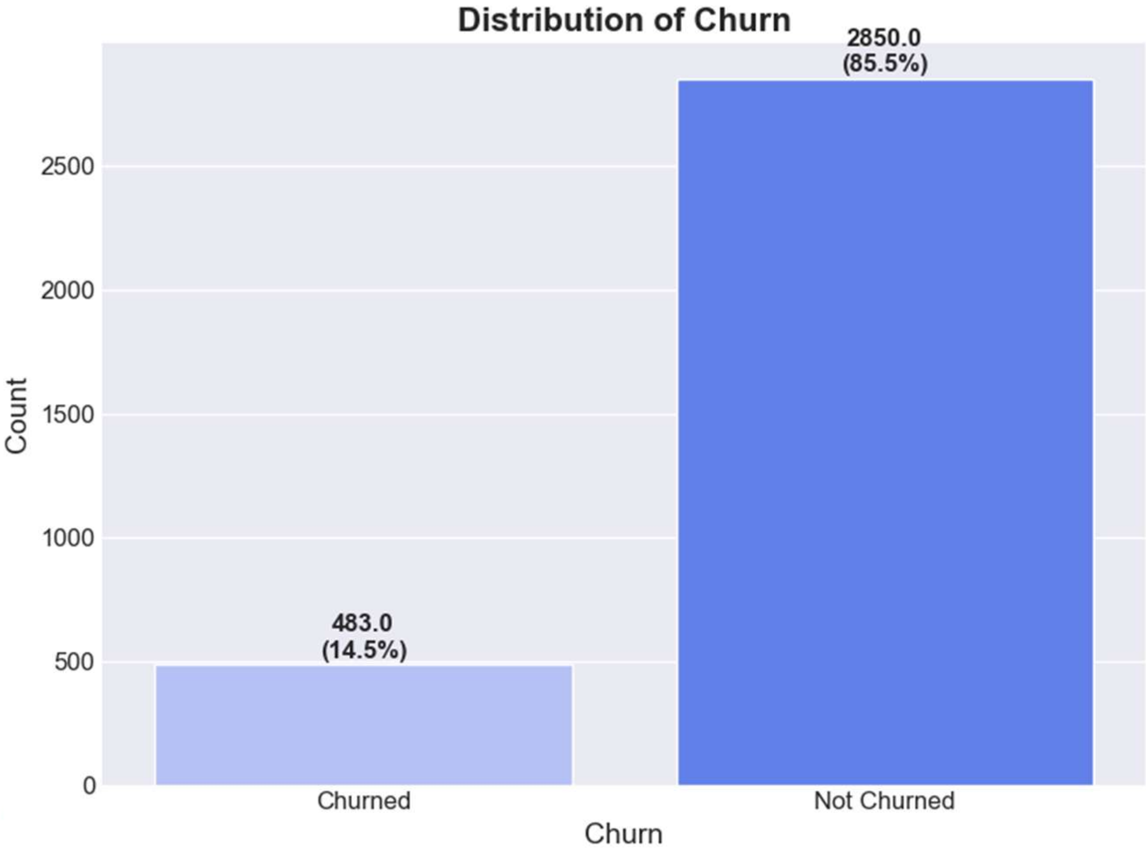
Univariate Data Analysis



BUSINESS & DATA UNDERSTANDING

- The numerical data showed a right-skewed pattern in customer tenure and call durations across different time periods, suggesting lower usage patterns. International calls may have higher volumes for a smaller customer segment. Customer service calls showed a multimodal distribution, indicating distinct groups with varying frequency of contacting customer service. Understanding these relationships and potential outliers can help predict churn.

BUSINESS & DATA UNDERSTANDING

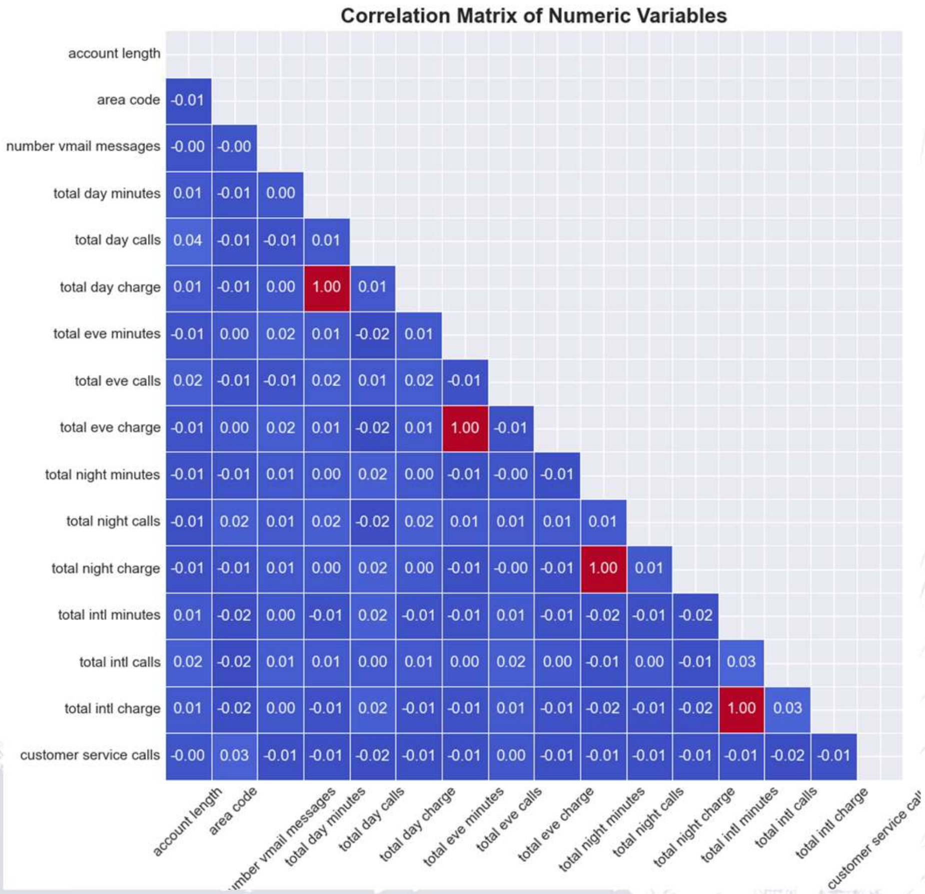


- A detailed analysis of churn was conducted by counting the number of consumers who stopped using the service and the number of customers who continued. The following are the outcomes:
- Total number of churned customers: 483 There were 2850 clients that remained loyal.

BUSINESS & DATA UNDERSTANDING

- **Multivariate Data Analysis**: This entailed looking at several variables' relationships at once. This study looked at the relationship between several features and the aim variable (customer churn) when taken into account collectively.
- The correlations between the various variables in the dataset were found using a correlation matrix.

BUSINESS & DATA UNDERSTANDING



Correlation Matrix of Numeric Variables

	account length	area code	umber 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 cal
account length																
area code	-0.01															
number vmail messages	-0.00	-0.00														
total day minutes	0.01	-0.01	0.00													
total day calls	0.04	-0.01	-0.01	0.01												
total day charge	0.01	-0.01	0.00	1.00	0.01											
total eve minutes	-0.01	0.00	0.02	0.01	-0.02	0.01										
total eve calls	0.02	-0.01	-0.01	0.02	0.01	0.02	-0.01									
total eve charge	-0.01	0.00	0.02	0.01	-0.02	0.01	1.00	-0.01								
total night minutes	-0.01	-0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01							
total night calls	-0.01	0.02	0.01	0.02	-0.02	0.02	0.01	0.01	0.01	0.01						
total night charge	-0.01	-0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01					
total intl minutes	0.01	-0.02	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02				
total intl calls	0.02	-0.02	0.01	0.01	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	-0.01	0.03			
total intl charge	0.01	-0.02	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03		
customer service calls	-0.00	0.03	-0.01	-0.01	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	

- Correlation Matrix Analysis: Call Numbers, Minutes, Charges
- Significant positive relationships between call numbers, minutes, and charges.
- Higher call spending leads to more calls.
- Weak positive link between call minutes and account length.
- Customer service calls show weak correlations, unrelated to calling habits or account tenure.
- International calls show favorable link, leading to increased spending.

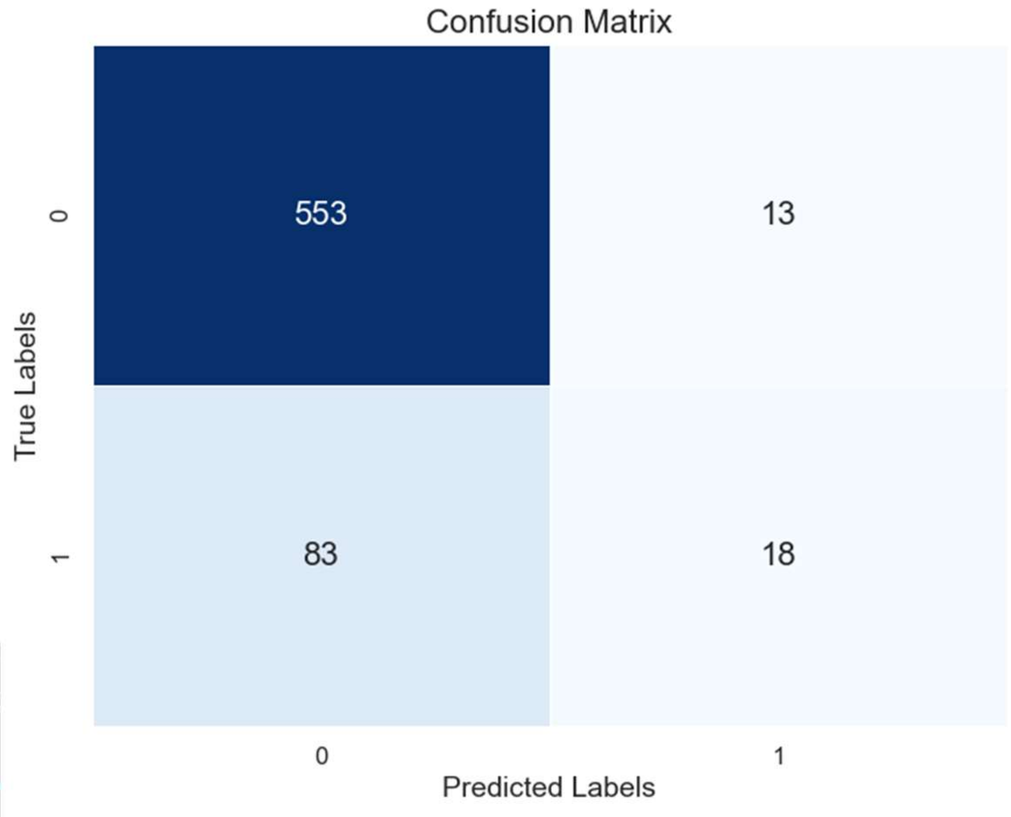
DATA PROCESSING & MODELING

- -Label Encoding: converts label variables in "international plan", "voice mail plan", and "churn" columns to numeric form. The "Yes" and "No" categories are converted to 1 and 0, respectively, representing presence or absence. In the "churn" column, "False" and "True" categories are converted to 0 and 1, respectively, representing customer churn.
- -One-hot encoding: this converts categorical variables into numerical format for machine learning algorithms, especially useful for dealing with variables with multiple categories.
- -Resample function: ensures a fairer representation of data and potentially enhancing the model's performance in handling imbalanced datasets

DATA PROCESSING & MODELING

- **Modeling:**The SyriaTel Customer Churn project aims to develop a classifier that predicts customer churn based on call usage, account details, and customer service interactions, using machine learning techniques to accurately predict outcomes for unseen data.

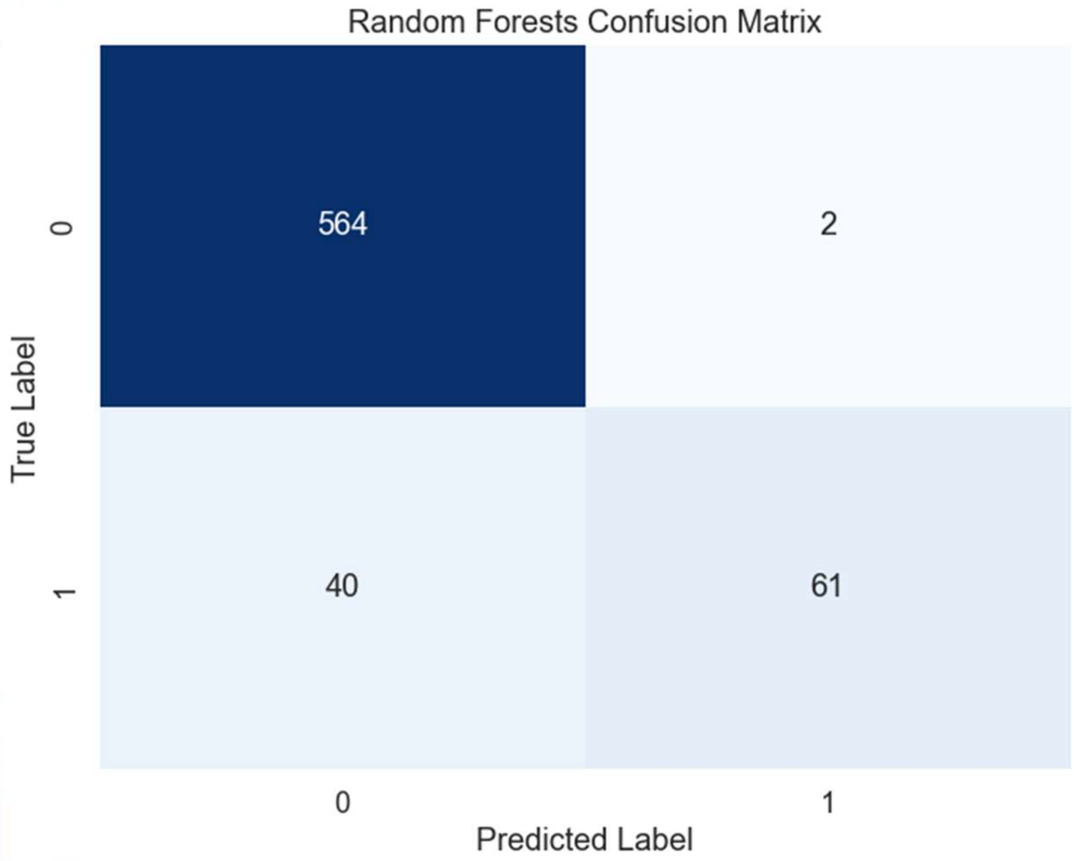
DATA PROCESSING & MODELING



Logistic Regression:

- Despite having an accuracy of 85%, the logistic regression model's recall and precision for predicting churn are both poor at 18% and 58%, respectively, suggesting a large miss of actual churn cases. The model might not be the best option for this classification assignment, according to these results..

DATA PROCESSING & MODELING



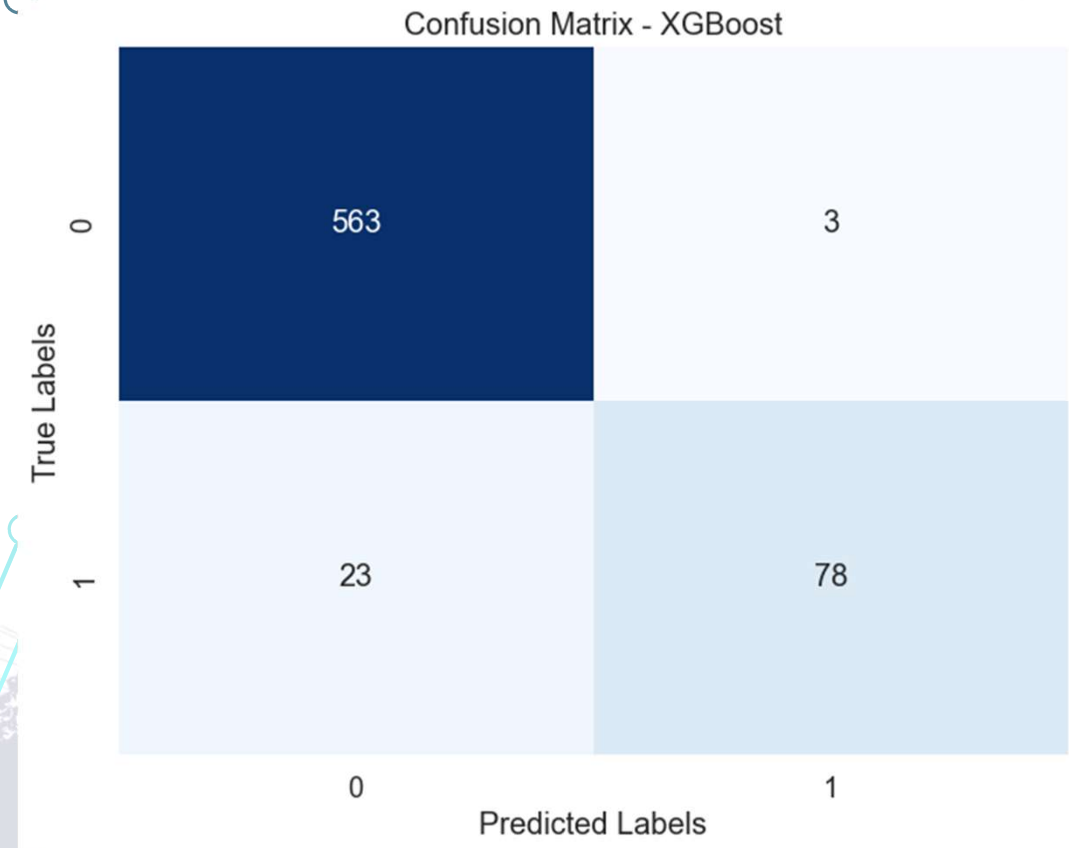
Random Forests Confusion Matrix

	0	1
0	564	2
1	40	61

- **Random Forests**

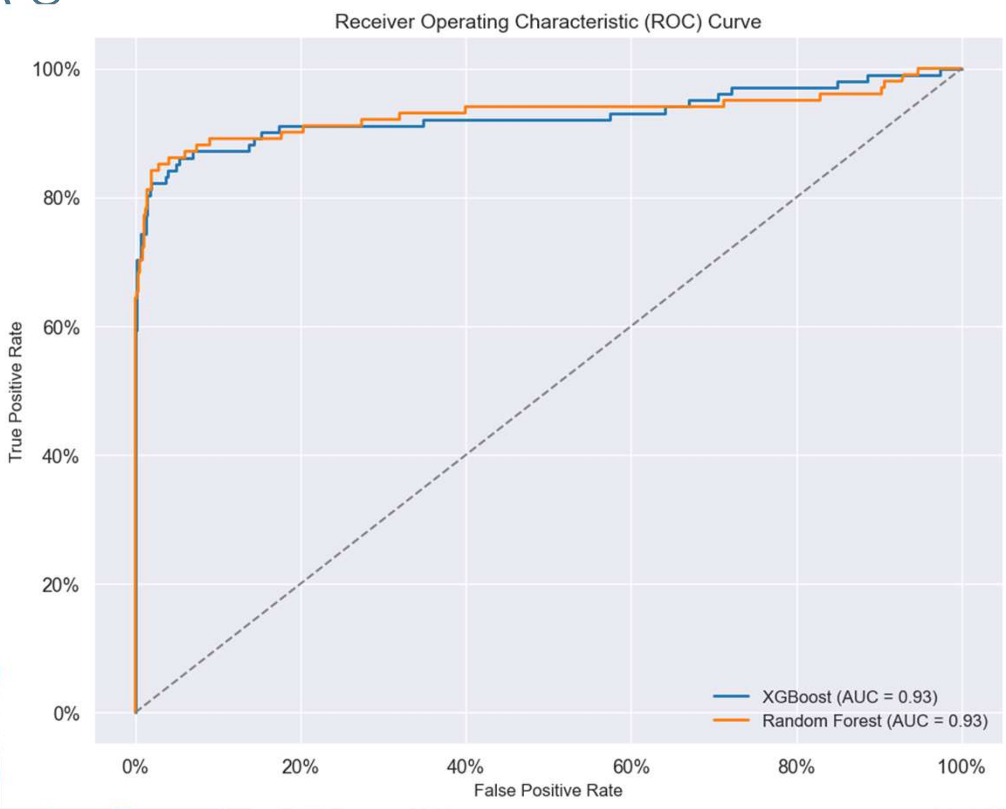
  The model exhibits high accuracy at 93.7%, with high precision, recall, and F1-score in the majority class ('False'). However, it tends to miss some 'True' samples, with a lower recall at 60%. The weighted averages show robust performance across both classes.

DATA PROCESSING & MODELING



This XGBoost model demonstrates high precision, indicating that when it predicts a customer will churn (True), it is correct 96% of the time. However, its recall for churn is lower at 77%

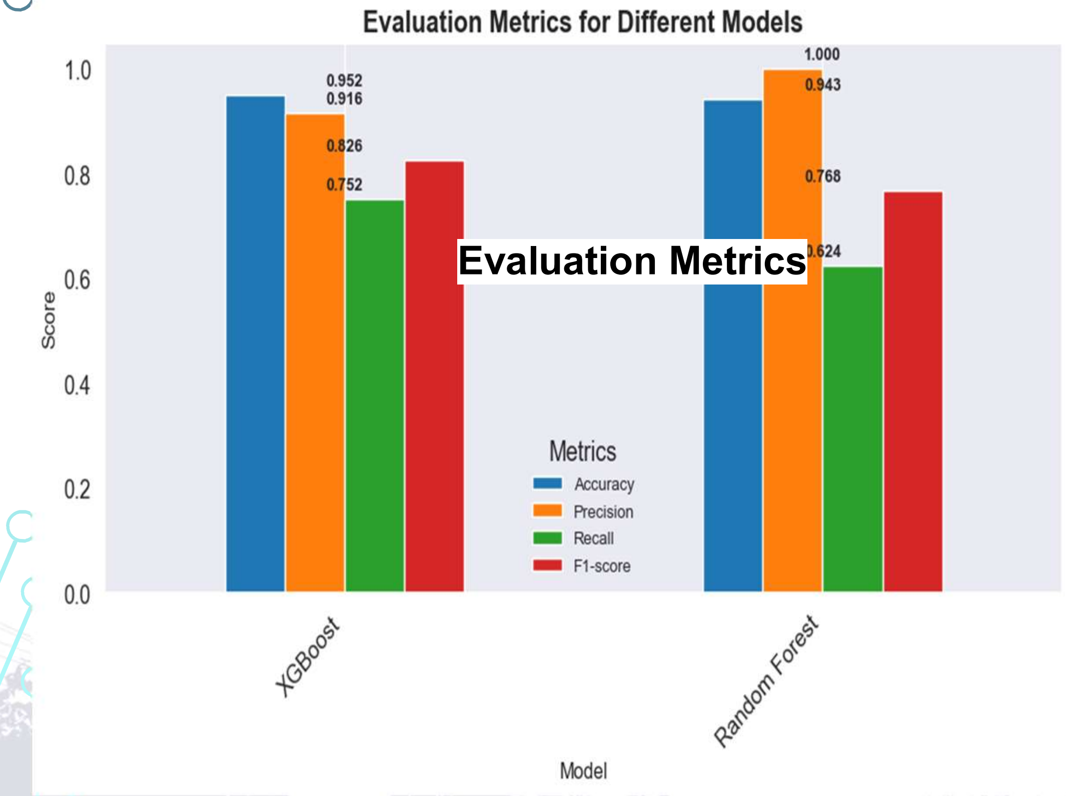
DATA PROCESSING & MODELING



Hyperparameter Tuning:
We are going to select the two best performing models and tune them

The ROC AUC score of 0.93 for both models demonstrates strong discriminatory power, reducing false positives and false negatives, indicating their ability to accurately capture data patterns and provide well-informed predictions across various threshold values, thereby enhancing their reliability in classification tasks.

DATA PROCESSING & MODELING



Evaluation Metrics

XGBoost vs Random Forest Evaluation

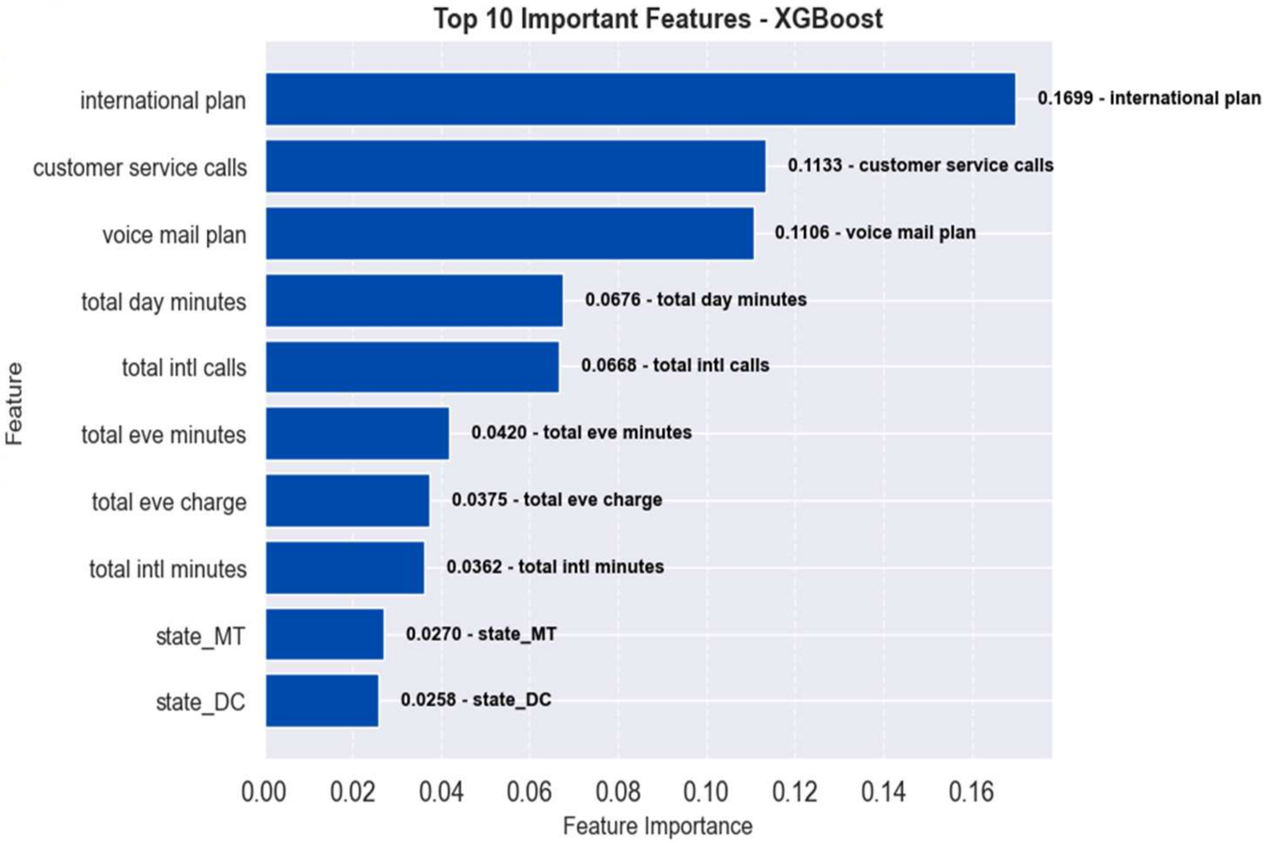
•XGBoost outperformed Random Forest in accuracy, recall, and F1-score.

•XGBoost achieved a higher accuracy (0.9610) than Random Forest (0.9415).

•Random Forest had a perfect precision (1.0000), indicating all positive predictions were correct.

•XGBoost's higher recall (0.7723) indicates its ability to identify more true positives.

DATA PROCESSING & MODELING



- Feature selection is crucial in enhancing model performance, reducing overfitting, and improving interpretability. Prioritizing informative features simplifies the model, reduces computational costs, and may improve predictive accuracy.

CONCLUSION & RECOMENDATION

The project developed machine learning models to predict customer churn for SyriaTel using classification algorithms like Logistic Regression, Random Forests, and XGBoost. Key predictors of churn were identified and accurate predictive models were developed.

The best course of action is focusing on customer retention strategies, enhancing service offerings, continuous model monitoring, establishing a customer feedback loop, and investing in data analytics and infrastructure.

Targeted customer retention strategies, especially for high churn risk customers, can reduce churn rates and improve customer satisfaction.

Enhancing service offerings, such as international plans and call durations, can also help. Continuous monitoring and retraining with updated data are recommended to ensure model effectiveness.