NON - TECHNICAL DATA REPORT

1. BUSINESS UNDERSTANDING

Business Overview

Syriatel (Arabic: سيريتل) is a leading telecommunications company in Syria, known for its rapid growth and extensive market presence. With a robust network of 63 Points of Service across the country, Syriatel handles over 25,000 customer queries daily through its Call Centers and operates 2,783 radio base stations. The company proudly serves over 6 million customers, holding a 55% share of the Syrian market. Their skilled team is committed to delivering high-quality services and solutions, solidifying Syriatel's position as one of the region's fastest-growing telecom operators.

Problem Statement

As new customers begin using a product, each contributes to the growth rate of that product. However, over time, some customers may discontinue their usage or cancel their subscriptions for various reasons. Churn refers to the rate at which customers cancel or choose not to renew their subscriptions, and a high churn rate can significantly impacts revenue.

Syriatel has observed an increase in customer churn and is concerned about the financial losses associated with customers who discontinue their services prematurely.

Objectives

To Determine the features that serve as early indicators of customer churn.

To Analyze and identify the underlying reasons why customers discontinue their service.

To Build a Predictive Model that is capable of accurately predicting when a customer is likely to discontinue their service.

Success Criteron

This analysis aims to:

Identify Key Features: Determine at least five key features that strongly correlate with customer churn, providing actionable insights for Syriatel to monitor and address customer dissatisfaction effectively.

Develop a Predictive Model: Build a classifier model that achieves: At least 90% accuracy in predicting customer churn. A minimum precision of 75%, ensuring the model minimizes false positives and provides reliable predictions.

Support Business Decision-Making: Enable Syriatel to use the identified features and model predictions to implement targeted retention strategies, reducing churn and mitigating revenue loss.

2. DATA UNDERSTANDING

The Churn in Telecom's dataset was sourced from Kaggle. The dataset has 3,333 rows and 21 columns. Each column represents a customer and the columns represent the customer details. The customer details are as follows:

1. the state the cusomer lives in,
2. account length- the number of days the customer has had an account,
3. the area code of where the customer lives,
4. the customer's phone number,
5. international plan- true if the customer has the international plan, otherwise false,
6. voice mail plan- true if the customer has a voice mail plan, otherwise false,
7. number vmail messages- the number of voicemails the customer has sent,
8. total day minutes- the total number of minutes the customer has been on call during the day,
9. total day calls- total number of calls the user has done during the day,
10. total day charge- total amount of money the customer was charged for calls during the day,
11. total eve minutes- the total number of minutes the customer has been on call in the evening,
12. total eve calls- the total number of calls the customer has been on in the evening,
13. total eve charge - total amount of money the customer was charged calls during the evening,
14. total night minutes - the total number of minutes the customer has been on call at night,
15. total night calls- total number of calls the user has done at night,
16. total night charge- total amount of money the customer was charged for calls at night,

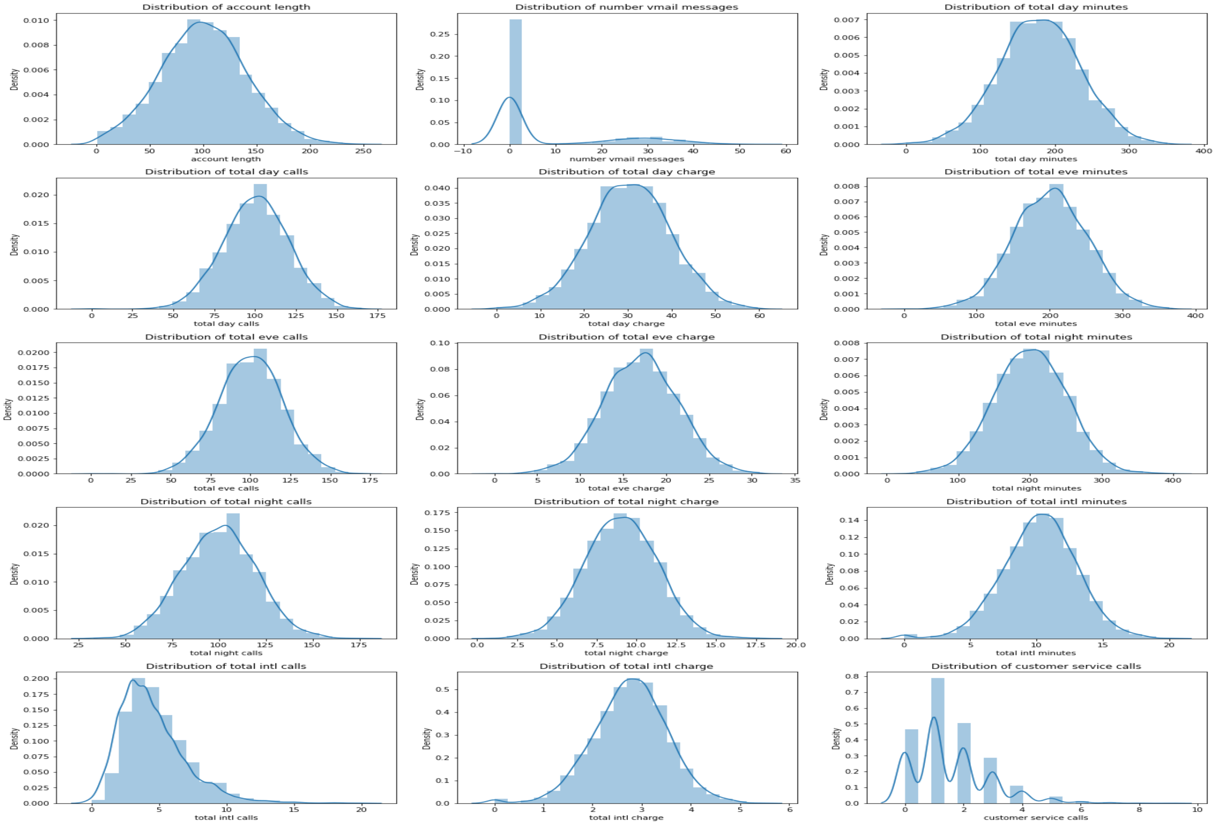
17. total intl minutes- the total number of minutes the customer has been on international calls,
18. total intl calls- total number of international calls the customer has done,
19. total intl charge- total amount of money the customer was charged for international calls,
20. customer service calls- number of calls the customer has made to customer service,
21. churn- true if the customer terminated their contract, otherwise false

3. DATA PREPARATION & ANALYSIS

Upon checking the data for duplicates, missing values and null values and there were none.

Univariate analysis

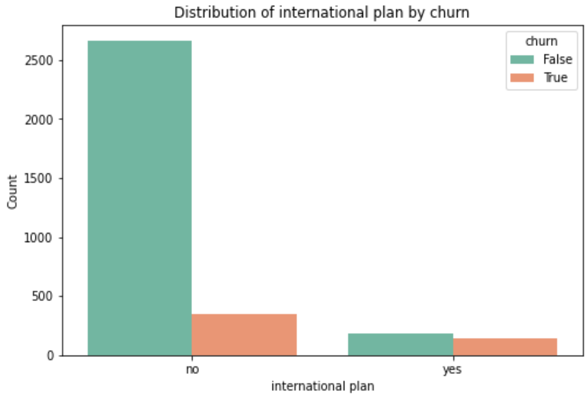
To better understand the dataset, we analyzed the distributions of our numerical columns using distribution plots.

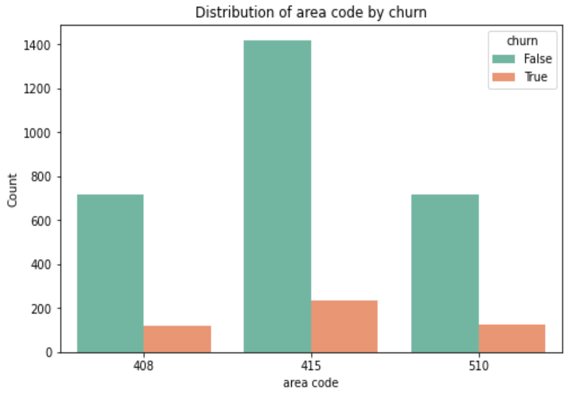


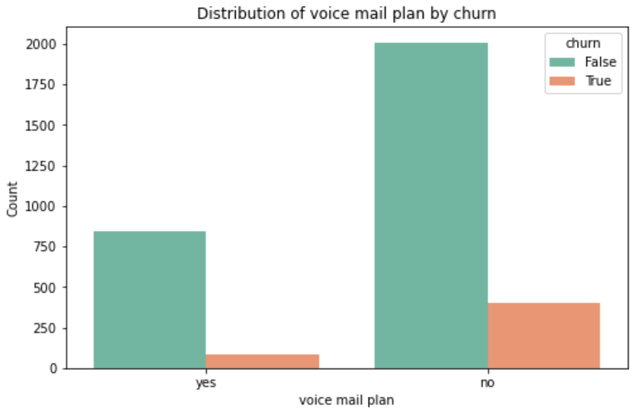
The analysis revealed that 12 out of the 15 numerical variables follow a normal distribution, which is advantageous for modeling purposes. The selected numerical columns include: account length, number of voicemail messages, total day minutes, total day calls, total day charge, total evening minutes, total evening calls, total evening charge, total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, and customer service calls.

Bivariate analysis

We visualized the relationship between our selected categorical variables: area code, international plan, and voice mail plan, and the target variable- churn using countplots.







The count plots reveal that customers with an international plan and those without a voice mail plan are more likely to churn, suggesting these features may influence customer retention. However, area code does not appear to have a significant impact on churn, as the distribution is relatively consistent across different codes. These insights highlight the importance of service plans in understanding customer churn.

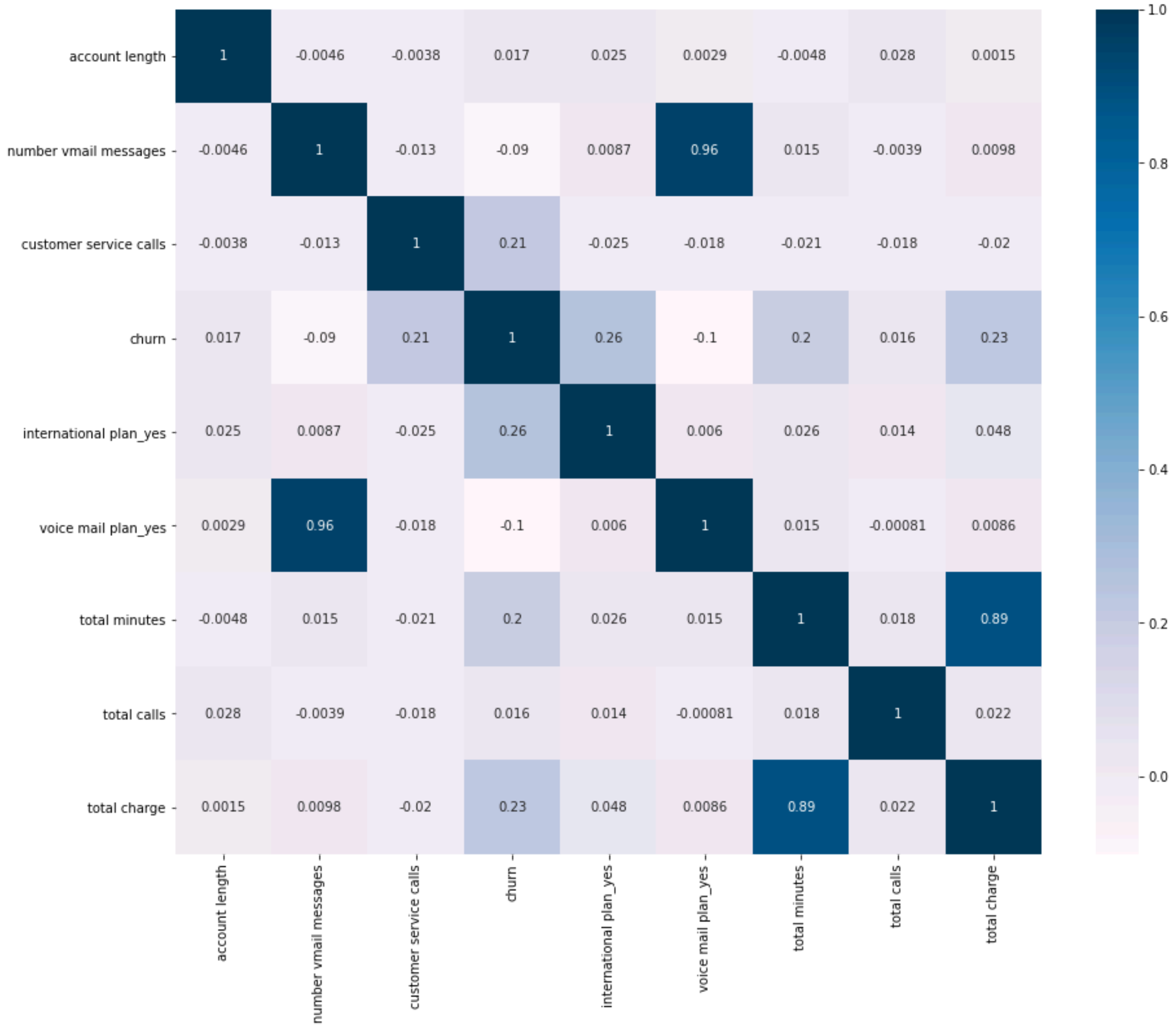
To further prepare the data for analysis, we;

1. Dropped/Removed irrelevant columns such as area code, phone number, and state.
2. Transformed the categorical columns into binary format using one-hot encoding, with adjustments made to avoid multicollinearity.

3. Merged the total minutes, calls, and charges into single combined columns for each category, reducing the dataset's size and improving its efficiency for analysis.

Multivariate analysis

We visualized the correlations among the remaining features to address any multicollinearity and optimize the dataset for modeling.



The results showed that most features exhibit little to no correlation, indicating they are largely independent. However, we observed strong correlations between voice mail plan_yes and number vmail messages, as well as between total charge and total minutes. To address this, we evaluated their correlations with churn and dropped the feature with the weaker correlation to ensure the dataset remains relevant and non-redundant.

4. MODELING & EVALUATION

In this analysis, we built and evaluated several machine learning models to predict customer churn for Syriatel.

Models used

We employed the following models in our analysis:

1. **Logistic Regression**: A simple model for binary classification tasks. It provides a strong baseline, allowing us to measure performance improvements with more complex models.
2. **Modified Logistic Regression (with SMOTE and MinMax Scaling)**: To improve upon the baseline Logistic Regression model, we applied the SMOTE technique to address class imbalance and used MinMax scaling to standardize the feature range. This modification aimed to enhance the model's performance, especially in terms of handling imbalanced data.
3. **Decision Tree Classifier**: A decision tree is a non-linear model that works well for problems with complex relationships between features. It is robust to outliers and doesn't require feature scaling, which made it a great option for this analysis.
4. **Random Forest Classifier**: This ensemble model is made up of multiple decision trees and is known for its high accuracy, robustness to overfitting, and ability to handle complex data.

The success of these models is evaluated based on the following criteria:

- **Accuracy**: A high accuracy score indicates that the model correctly classifies the majority of instances.
- **Precision**: This metric shows how many of the predicted churns were actual churns, minimizing false positives.
- **Recall**: Recall is important for identifying the actual churns, minimizing false negatives.
- **F1 Score**: A balanced measure of precision and recall.
- **AUC and ROC Curve**: The AUC score quantifies the overall ability of the model to distinguish between classes, while the ROC curve provides insights into performance across various thresholds.

Model Performance

	MODEL / CLASSIFIER

	LOGISTIC REGRESSION	MODIFIED LOGISTIC REGRESSION	DECISION TREE	RANDOM FOREST
ACCURACY SCORE	0.851	0.601	0.918	0.945
F1 SCORE	0.228	0.366	0.728	0.791
PRECISION	0.154	0.804	0.769	0.727
RECALL	0.44	0.237	0.692	0.867

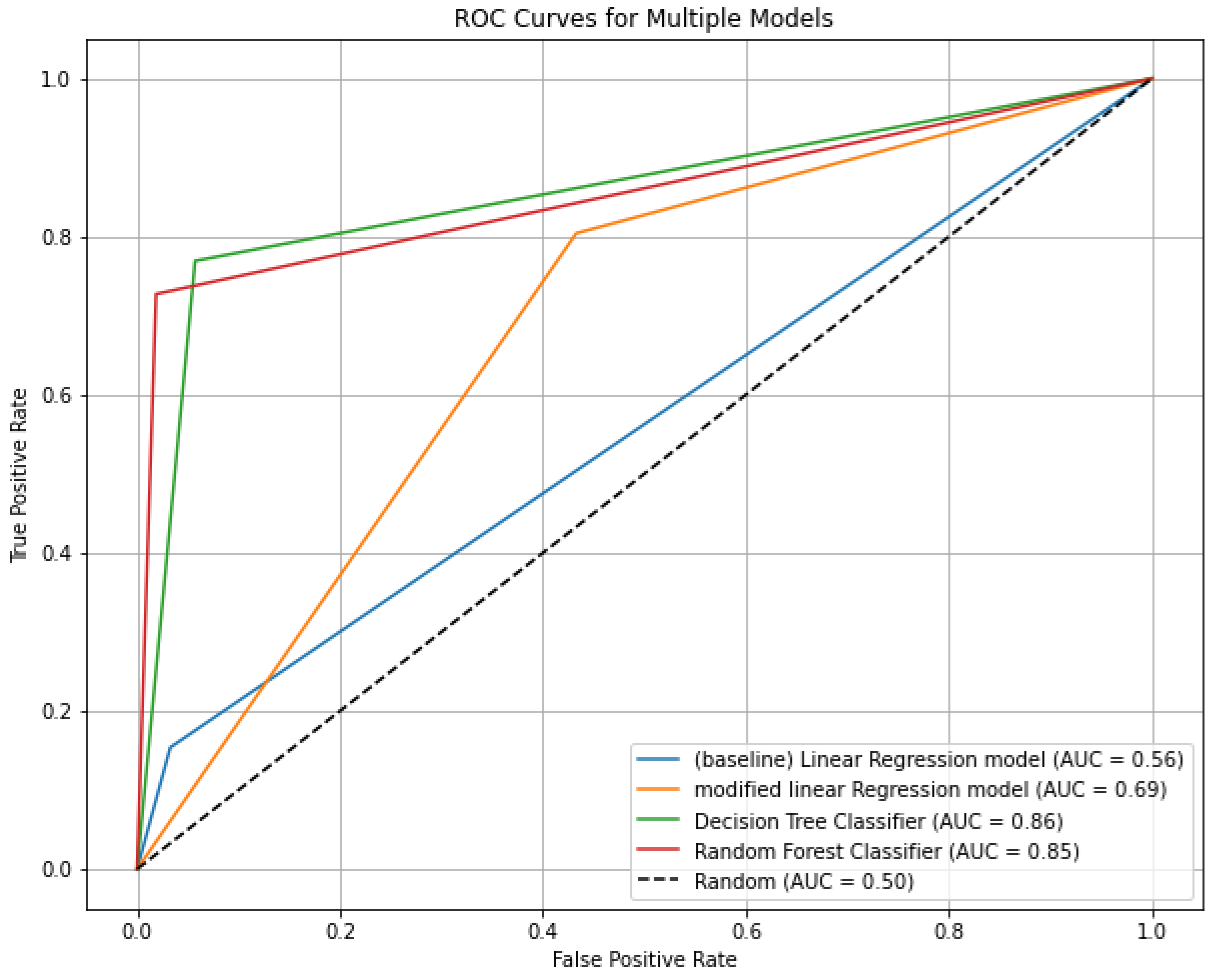
The Logistic Regression model provided us with a baseline performance. While it achieved a reasonable accuracy, its F1 score, precision, and recall were low, indicating that the model struggled with identifying churn customers effectively.

The modified Logistic Regression model, after applying SMOTE and MinMax scaling, performed better in terms of recall, successfully identifying churn instances. However, it showed a lower precision, meaning there were many false positives. The overall accuracy also dropped compared to the baseline Logistic Regression model.

The Decision Tree model performed better overall, showing improved precision and recall. It successfully identified churn cases, though it was slightly less accurate than other models.

The Random Forest model exhibited the highest accuracy and F1 score among all models. It had excellent precision, meaning it correctly identified churn customers, and was also good at minimizing false positives. It balanced recall and precision well, making it a very effective model.

In addition to the model summaries, we visualized the ROC curve and AUC to more precisely compare model performance, assessing the balance between true positive and false positive rates and providing a single value to quantify overall effectiveness as seen below:



Based on the results, the Random Forest Classifier outperformed all other models. It achieved the highest accuracy and precision, and balanced recall effectively. Therefore, we recommend using this model for predicting customer churn in order to maximize prediction accuracy and minimize false positives.

5. CONCLUSIONS

Through data analysis and model building, we identified key features such as 'account length', 'customer service calls', 'churn', 'international plan_yes', 'voice mail plan_yes', 'total calls', and 'total charge' that significantly predict customer churn. The Decision Tree Classifier performed well, while the Linear Regression model struggled despite SMOTE and MinMax scaling. The Random Forest Classifier excelled, achieving over 90% accuracy, demonstrating that ensemble methods like Random Forest, along with decision trees, are highly effective for churn prediction.

6. RECOMMENDATIONS AND NEXT STEPS

Enhance Feature Selection:
Further exploration of customer behavior and satisfaction related features could improve model accuracy in predicting churn and help identify key factors leading to service discontinuation.

Optimize Model Performance:
Fine-tuning the Random Forest and Decision Tree models, along with experimenting with alternative sampling techniques, can help improve prediction accuracy, especially in identifying at-risk customers.

Focus on Customer Retention Strategies:
Use the model's predictions to implement targeted retention efforts, such as personalized interventions or loyalty programs, and continuously monitor the model to ensure its effectiveness in predicting when customers are likely to discontinue their service.

7. REFERENCES

These are the references that used:

Worldfolio
Kaggle
Productplan.com