**Data Report on** **SyriaTel Telecommunication Company**

**GROUP 6**

**Members:**

1. Elvis Wanjohi
2. Stephen Gachingu
3. Elsie Almadi
4. Naledi Ambuku
5. Mark Muriithi

**Moringa school**

**Data Science**

**September 2025**

**TABLE OF CONTENT**

**Business Understanding………………………………………………………1**

**Data Understanding…………………………………………………………….2**

**Data Preparation.………………………………………………………..........3**

**Data Analysis...……………………………………………………………………4**

**Modeling…………………………………………………………………………….5**

**Recommendations ……………………………………………………………...6**

**Conclusion……………………………………………………………………………7**

**Business Understanding**

**Business Overview**

SyriaTel is a telecommunication company that specializes in the provision of data and voice services. Like its competitors, the company’s overarching goal is profit maximization. However, one of the major challenges it faces is customer churn, which occurs when subscribers cancel their services and switch to competitors. This project seeks to address that challenge through data-driven methods.

**Problem Statement**

SyriaTel is looking to increase their revenue by reducing the number of customers that are canceling their services. At present, the company does not have a reliable system to predict which customers are most likely to leave. Without such a system, it is difficult to intervene in time to retain customers, which ultimately leads to reduced profits and loss of market share.

**Business Objective**

The main business objective is to reduce customer churn by predicting which customers are at risk of leaving. Achieving this will allow SyriaTel to take timely action and improve customer retention.

**Project Objectives**

**Main Objective**

The main objective of this project is to develop a machine learning classifier that can accurately predict whether a SyriaTel customer is likely to churn.

**Specific Objectives**

These objectives include:

* To analyze customer attributes and usage patterns in order to identify the key drivers of churn.
* To develop, train, and evaluate predictive models that classify customers as churners or non-churners.
* To evaluate the classifier model performance using metrics such as F1-score, precision and recall.
* To generate actionable business insights and recommendations, based on the model outputs, that can support strategies to reduce customer churn.

**Business Success Criteria**

The success of this project will be assessed in three ways:

* It should generate actionable insights that SyriaTel can use to reduce churn rates in the future.
* The predictive model should achieve acceptable levels of performance, with high accuracy and a strong ability to correctly identify customers who are likely to leave.
* The results should be presented in a way that is clear and interpretable, so that they can be easily understood and applied by business managers and decision-makers**.**

Achieving this will clearly inform the shareholders at SyriaTel on which services have a large number of churns and take decisive action in improving the service.

**Requirements and Assumptions**

Requirements: The dataset contains the required information to complete the analysis.

Assumptions: This dataset acquired from Kaggle has the correct and accurate information on the number of churns in Syriatel telecommunication company.

**Data Understanding**

**Overview**

The dataset used in this project is the Churn in Telecoms dataset obtained from Kaggle. It contains customer account and usage information for a telecommunications company.

This dataset is in csv format and it contains 3,333 rows and 21 columns.

**Exploring the data**

The columns in this dataset include:

|  |  |
| --- | --- |
| Feature | Description of Feature |
| **State** | This contains the abbreviation of each statewhere every user resides in. |
| **Account length** | This refers to the amount of time a credit account has been open. |
| **Area Code:** | This is three-digit number that indicates the geographic region of where the number is located |
| **Phone Number:** | This is the user’s unique subscriber number. |
| **International Plan** | This an extended service from a mobile carrier that allows a customer to use their phone to call, text and data in another country. |
| **Voice mail Plan** | This is a service that allows callers to leave recorded audio messages. |
| **Total day minutes**. | This refers to the cumulative amount of time a customer spends interacting with a service within a given day |
| **Total day calls** | This refers to the total number of phone calls a customer made during the day. |
| **Total day charge** | This refers to the sum of all charges incurred by a customer within a single day. |
| **Total evening minutes** | This refers to the cumulative amount of time a customer spends interacting with a service in the evening. |
| **Total evening calls** | This refers to the total number of phone calls a customer made in the evening. |
| **Total evening charge** | This refers to the sum of all charges incurred by a customer in the evening. |
| **Total night minutes** | This refers to the cumulative amount of time a customer spends interacting with a service in the night. |
| **Total night calls** | This refers to the total number of phone calls a customer made during the night. |
| **Total night charge** | This refers to the sum of all charges incurred by a customer during the night. |
| **Total international minutes** | This refers to the cumulative amount of time a customer talks with someone from a different country. |
| **Total international calls** | This refers to the total number of phone calls a customer made to someone in a different country. |
| **Total international charge** | This refers to the sum of all charges incurred by a customer for all calls made internationally. |
| **Customer service calls** | This is an assistance and support provided by a company to its customers both before or after they buy or use its products or services |
| **Churn** | This refers to customers stopping their relationship with a company. |

The target variable is churn. This is a binary variable that indicates whether a customer has churned (True) or not (False). Since the target is categorical, it will be encoded during data preparation to allow machine learning models to process it.

**Data Preparation**

In this section the following steps were taken to prepare the data:

1. **Importing the required libraries**

The necessary libraries required for this project include: Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn.

1. **Loading the data**

The dataset was obtained from Kaggle and was compiled into a zipped CSV file, which was saved locally and loaded into our project for further analysis, modeling and evaluation.

1. **Data Cleaning**

In this step, we preview the dataset to understand its structure, identify missing values, detect duplicates, and check data types.

Our dataset contained no missing values and no duplicates.

1. **Column Removal**

The phone number column was dropped because it acts as an identifier and does not affect whether a customer churns**.**

1. **Feature Engineering**

In this section we added calculated columns which include:

* Total individual charges - This column compares the charges among different people in a day.
* Total individual minutes- This column shows us the customer interaction.
* Total individual calls - This column helps us understand the customer behaviors.
* Charges per call- This column is useful for assessing call value and identifying whether certain customers or call types are more profitable.

The Total Individual charges and Total individual minutes column are used to get the average cost per minute which is 0.100445.

The final dataset contained 3333 rows and 24 columns.

**Exploratory Data Analysis**

Exploratory Data Analysis entails examining and summarizing the main characteristics of a dataset using both statistical and visual techniques.

In this section EDA helps examine the states with the largest customer loss, the states with the largest number of charges, the percentage of customers who churn, and the overall customer interactions with the services provided by SyriaTel telecommunication company. By using visualizations EDA allows us to detect underlying patterns, trends and correlations within the data. EDA will help us identify whether the customers with more interaction churn.

**Univariate analysis**

We used univariate analysis to show the percentage of customers who churn and who don’t. The visualization showed about 80% of customers do not churn and about 15% who churn.

**Bivariate analysis**

We showed the relationship between the target variable and state, where the state with the largest number of churns was New jersey. We also showed the relationship between total individual minutes and the churn status This helped us identify a pattern where the customers who use more minutes tend to not churn and the customers who use less minutes are more likely to churn.

We were able to uncover the trends and patterns using a density plot, bar graph and scatter plot.

**Multivariate analysis**

This section involves analyzing data with multiple variables simultaneously. We showed the relationship between state and all the charges which include total day charge, total evening charge, total night charge and total international charge. This allows us to capture patterns and dependencies that would be missed with univariate analysis.

**Modeling**

In this section we fitted a Logistic regression model and Decision Tree model which are classifier models. We made use of the Scikit-learn library for modeling our features and target variable.

The goal for this section was to evaluate the classifier model performance using metrics such as, Accuracy, F1-score, precision, recall and AUC.

* Accuracy: is how many predicted instances were correct overall.
* Precision: of those we said "will churn", how many really churned.
* Recall: of those who really churned, how many did we find.
* F1: This is the balance between precision and recall.
* AUC: is how well the model ranks customers from low to high risk of churning.

Preparing data for modeling:

1. **One hot Encoding:** In this part we converted the categorical variables to numeric datatypes. The categorical columns included: International plan, Area code, Voice mail plan, and Churn.
2. **Performing a train test split:** We split our dataset with 70% for the training set and 30% for the test set. We used a random state of 42.
3. **Standardizing-** This involves scaling the features to a similar range typically a mean of 0 and a standard deviation of 1 since the features have vastly different ranges.
4. **Fixing Class imbalance-** We used SMOTE (synthetic minority oversampling technique) to create balance between the majority and minority class. There was a high number of customers who did not churn (about 80%) compared to customers who churned (about 15%). Fixing class imbalance ensures the classifier models don’t ignore the minority class.

In this section we also plotted a ROC curve and a Confusion matrix to illustrate the model performance.

**Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1\_score** | **AUC** |
| Logistic Regression | 0.755 | 0.337 | 0.710 | 0.457 | 0.808 |
| Decision Tree | 0.872 | 0.547 | 0.683 | 0.607 | 0.793 |

We used a confusion matrix to further illustrate both models’ performance and the decision tree classifier model was better at finding churners since it had a lower false positive rate than the logistic regression model.

Using the ROC curve the logistic regression model was slightly more reliable and was more stable across thresholds.

In this part we also tried to find the most important features that would lead to churn in both the decision tree and logistic regression model. These features influenced our insights and business recommendations.

**Recommendations**

* **Improve Customer support quality-** since Customer service calls are key drivers of churn, a high customer service call volume is strongly tied to churn since it means the customer might be having a problem with the service or is dissatisfied.
* **Reassess International Plan Pricing & Policies-** the feature importance results showed that customers with international plan are more likely to churn. Syriatel can review pricing and offer bundled discounts.
* **Retain High-Usage Customers-** Heavy day-minute users are at lower risk of churning since it indicates there’s high customer interaction with the product, but we've to keep the customers satisfied by creating targeted retention programs (loyalty points, discounted packages, or “VIP customer” care) to reduce churn on loyal customers.

**Conclusion**

The model shows that customer dissatisfaction and costly international plans are the main churn triggers. Focusing on customer experience, fair pricing, and rewarding loyal heavy users can significantly cut churn and boost retention.