**SYRIATEL CUSTOMER CHURN ANALYSIS:**

OVERVIEW AND DATA UNDERSTANDING

**Business Overview:**

Customer churn analysis has recently become increasingly important in the ever evolving and competitive telecommunication industry. Customer churn analysis involves the study of customer behaviour to identify patterns and factors that lead to customers leaving their providers. As the cost of getting a new customer is five to twenty-five times more than keeping an existing customer, telecommunication as well as mobile operators see the need to pay more attention to retaining existing customers to increase their revenues.

There are myriads of reasons why a customer might leave such as high prices, poor network coverage or customer service. However, one of the most common reasons cited is customers simply getting a better deal elsewhere, especially in markets where there is a lot of competition. Therefore, understanding these churn drivers, even though it’s not straight forward, is critical for not just knowing why customers leave but identifying the warning signs of customers about to terminate contracts or switch providers.

Thus, accurate prediction of customer’s behaviours, using machine learning solutions assists companies in identifying necessary actions to be incorporated into their customer retention management, such as whether to improve the service experience, design proactive campaigns to boost adoption, or re-engage at-risk customers.

**Problem Statement:**

SyriaTel, a telecommunications company in Syria, would like to predict whether a customer will (“soon”) stop doing business with them(“churn”). As such, it would like to get an understanding of the customer’ s behaviour and accurately pre-empt whether the customer will stop using their services

**Objectives:**

Objectives for this analysis are as set out below:

1. To come up with a predictive model that shows whether a customer will churn
2. Identify the key factors affecting customer churn amongst SyriaTel customers
3. Identify what aspects of SyriaTel services need more prioritization to prevent customer churn.

**Metrics of success:**

The following measures, based on previous studies done on customer churn analysis, are evaluated on the predictive models to ensure we have the best performing model:

* The accuracy metric: Measures the total number of correctly identified instances. An accuracy of between 75% and 85% is desired.
* The precision metric: Measures how the predictive model is observing the actual number of positives against the predicted positives. A precision of between 50% 6and 70% is desired.
* Recall metric: Measures the predictive model's ability to correctly identify churners. A recall of between 60% and 70% is desired.
* F1-score: Measures how accurate the predictive model’s performance is. A F! score of between 0.55 and 0.65 is highly desirable
* Area under the curve (AUC): A higher result indicates a more accurate model performance.

**DATA UNDERSTANDING**

The SyriaTel dataset used in this analysis has been sourced from Kaggle and contains records of SyriaTel customers. The dataset is made up of 3,333 rows with each row representing a customer record.

In addition, the dataset is made up of 21 columns which can be segmented as follows:

1. Customer usage: The following columns provide further insight to the customer phone usage based on time of day:
   1. Usage during the day - total day minutes, total day calls, total day charge
   2. Usage during the evening - total eve minutes, total eve calls, total eve charge
   3. Usage at night - total night minutes, total night calls, total night charge
2. Plan subscription: These columns give us a view of the plans that each of the customer has:
   1. International Plan
   2. Voice mail Plan which can be linked to No of Voicemail messages column
3. Unique customer details. The columns falling under this section are:
   1. State
   2. Account length
   3. Area code
   4. Phone Number
4. International phone usage/customer service: The columns under this category include:
   1. Total intl calls
   2. Total intl minutes
   3. Total intl charge
   4. Customer Service Calls
5. Likelihood of churn: This is captured in the column Churn which takes on two values:
   1. True – Customer has churned
   2. False – Customer is yet to churn

**DATA PREPARATION AND ANALYSIS**

**Data Preparation**

From the previous section, we have a view of the columns and rows in our dataset. The next step is to check for missing values, duplicate values as well as check for outliers. This has been done step b step as explained below:

1. **Null / Missing values**: The dataset does not contain any missing data in any of the 21 columns or 3333 rows. As such there is no need to impute values and the values in the dataset will be used as they are.
2. **Duplicate values**: Similarly, the dataset does not contain any duplicate values. No additional modification is required on this end.
3. **Outliers**: To perform this check, box plots on the numerical variables was done. The dataset is a mix of both categorical(text) and numerical columns. Thus, as a first step, we split our columns into categorical(text) and numerical categories and plot the numerical box plots to observe if there are any outliers as shown in the diagram below.

A graph of different colored squares

Description automatically generated

From the diagram above, majority of the columns do have a few outliers as expected given that the calling patterns may differ from customer to customer and some maybe far removed from the median. However, this pattern is similar amongst the different features we are looking at. For instance, the number of calls has a similar pattern of outliers regardless of whether it's day, evening or night period. This occurrence is replicated in the phone charges as well as the number of minutes. As such, these outliers will be kept for our analysis to give us a better understanding of the customer patterns.

**Exploratory Data Analysis**

Following on from the data preparation step above, we now have a complete dataset that can be used for the data analysis. The aim is to get further insights on SyriaTel customer behaviour as well as check which columns will be suitable to act as features and target variables while building the predictive model. This exploratory data analysis will be divided into three sections:

1. Univariate analysis – Examining the different features independently
2. Bivariate Analysis - Examining the relationship between two features
3. Multivariate Analysis – Examining the relationship between more than two features

**1. Univariate Analysis**

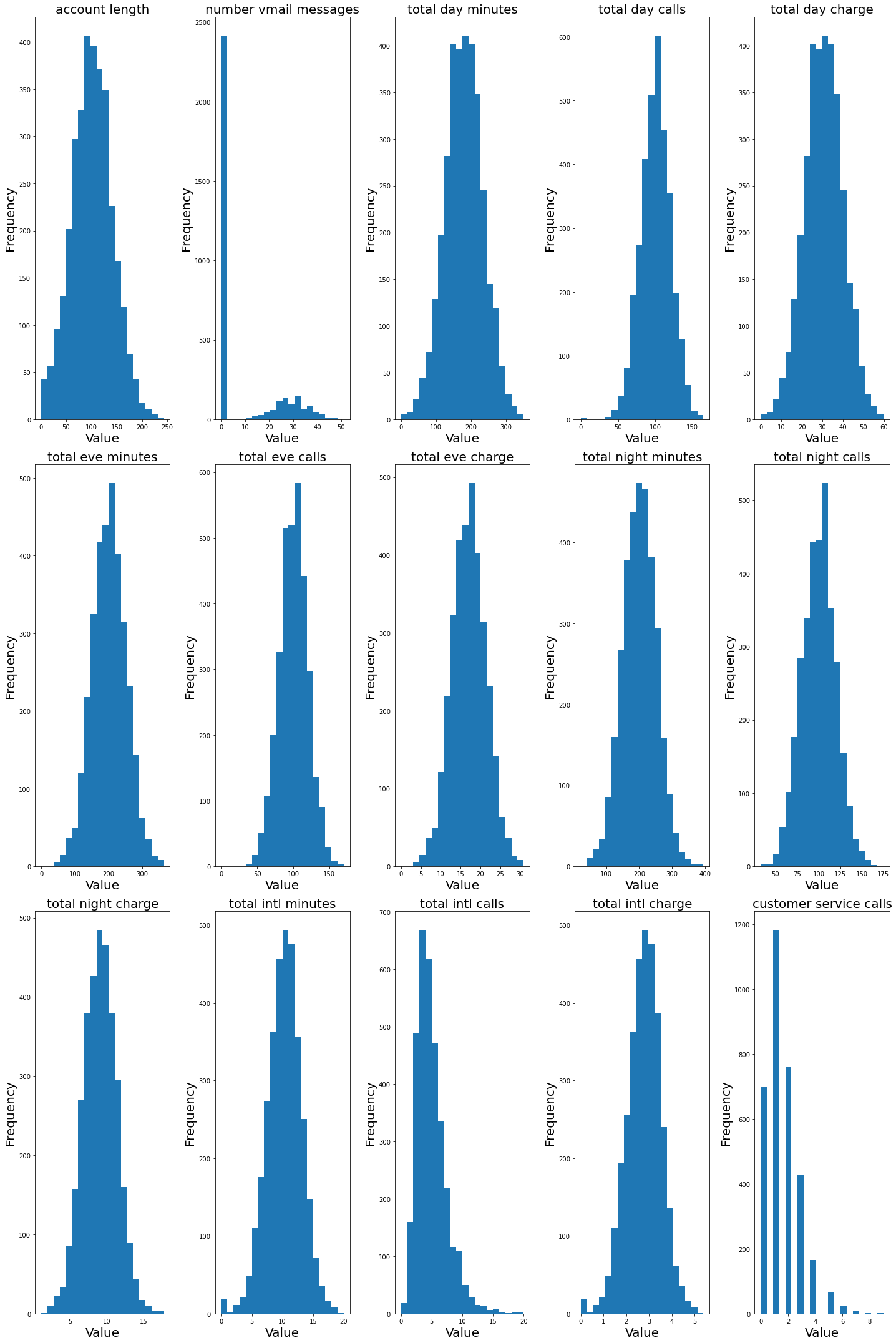
Given the dataset is made up of categorical and numerical columns, these were looked at separately.

1. **Numerical Columns**

The numerical columns are listed as follows:

* 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

To get a comprehensive look at these numerical columns, several histograms were plotted to get an understanding of their distribution and here’s the output:



From the diagram above, most of the numerical columns appear to be normally or closely distributed. Some columns such as “total intl call” appear to be positively skewed while “total international charge” is negatively skewed. Also, we can deduce that majority of the SyriaTel customers do not use the voice mail messaging services with roughly 2,400 never using this service. In addition, the customers rarely contact customer service as can be evidenced with approximately 1200 customers just making one call while roughly 700 customers having had no contact with customer care.

Finally, some of the columns such as “area code” and “account length” do not seem to be giving a lot of information with regards to customer behaviour. As such, these will be dropped, and all the other columns can be used as features to our predictive model.

1. **Categorical columns**

The categorical columns looked at are split as follows:

1. **Plan details**

The columns falling under plan details are “international plan” and “voice mail plan”. These are represented by the histograms in the diagram below.

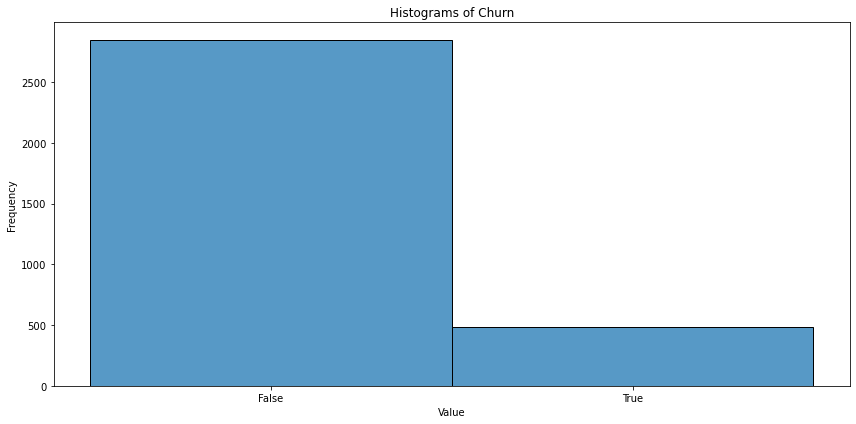
A blue and orange rectangles

Description automatically generated

It can be observed that roughly 3000 SyriaTel customers do not have an international plan, while roughly 2,400 customers do not have a voicemail plan. On the other hand, it does seem more customers are enrolled on the voicemail plan compared to the international plan. It could be that both these plana are not as attractive in terms of their offering and hence most customers tend to keep away from them or majority of the customers see no need for them hence the low uptake.

1. **Churn**

This is the categorical column, that shows how many of the customers have churned. This will also be the target variable when it comes to building our predictive model. A similar histogram to the one above was done and can be viewed below.



From the histogram above, we do observe that a lot of the customers have not churned. Roughly 2,800 customers are still with SyriaTel while roughly 500 customers have left. Therefore, we need to understand if any of the features looked at before may have caused this churn leading us to the next section which will delve more into this.

1. **Bivariate Analysis**

In this section we will be exploring the relationship between our target variable churn and all the other data features.

1. **Relationship between Churn and the Categorical Features**

The two main categorical features that will be investigated are international plan and voice mail plan. Using two separate count plots, as shown below, we can observe that the number of customers who churned were much lower regardless of whether they were on the international plan or not.

A screenshot of a graph

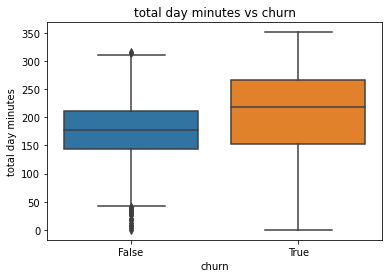
Description automatically generatedA screenshot of a graph

Description automatically generated

Similar pattern is observed with customers on the voicemail plan. However, the proportion of the customers who churned compared to those who are still active was much more pronounced on the voicemail plan. This proportion was quite close on the international plan. Therefore, there was high churn for customers who were on the international plan compared to those on the voice mail plan.

1. **Relationship between Churn and the Numeric Features**
   1. **Relationship between churn and total minutes**

Customers who churned had a higher median number of minutes spent on the phone. Below are boxplots showing the relationship between total minutes spent on the phone and churn.

A screenshot of a computer screen

Description automatically generated

A blue and orange rectangular shapes

Description automatically generatedA blue and orange rectangular shapes

Description automatically generated

This is quite evident during the day period where customers who churned spent a median period of around 200 minutes on the phone compared to 180 minutes for active customers. As for the evening and night, the difference is quite negligible with both sets of customers (those who churned and those who didn't) having a similar median amount spent on the phone. The variation observed is also quite similar amongst both sets of customers. A huge difference is observed in the amount spent on total international minutes as customers spent median duration of 10 minutes, which is quite removed from the day to day calling periods.

* 1. Relationship between churn and number of calls

Unlike the relationship between churn and total minutes, there is no noticeable difference in the number of calls made by churned and active customers. Both groups make a similar number of calls during the day, evening, and night, with a median of roughly 100 calls. The variation in the number of calls throughout the day is also similar for both groups. However, international calls show a different pattern, with fewer calls overall (a median of about 5 calls). Interestingly, churned customers tend to make fewer international calls, while active customers make slightly more.

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A blue and orange rectangular shapes

Description automatically generatedA blue and orange rectangular objects

Description automatically generated

* 1. Relationship between churn and phone charges

The boxplots show that total charges are higher for customers who churn compared to active customers. Charges are highest for calls made during the day and decrease as the day progresses into the evening and night. During the day, churned customers pay a median of around $35, while non-churned customers pay about $30, a significant difference that may contribute to customer dissatisfaction. Additionally, churned customers spend slightly more on international call charges compared to those who remain. Since churned customers also spend the most time on international calls, offering them lower rates could be an effective strategy for retention.A blue and orange rectangular shapes

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A blue and orange rectangular shapes

Description automatically generatedA blue and orange rectangular shapes

Description automatically generated

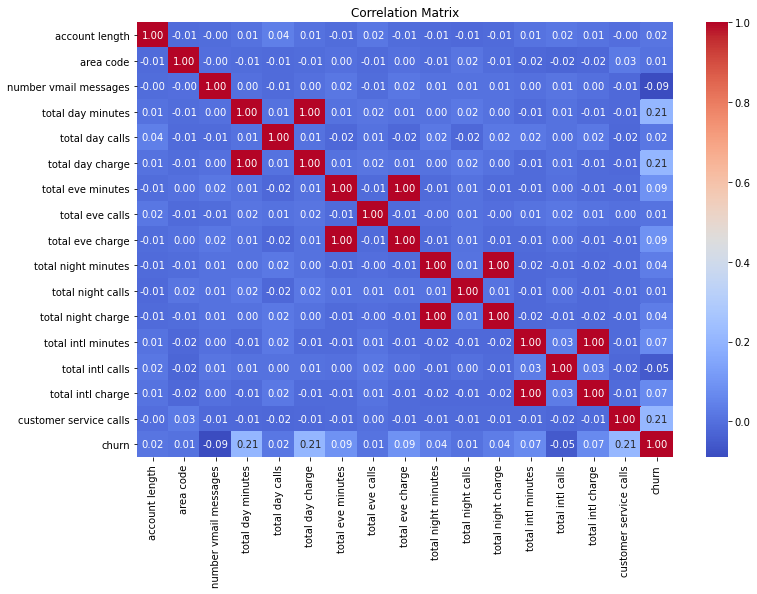
* 1. **Relationship between churn and customer service**

The count plot below reveals that customers who did not churn made the fewest customer care calls, with most making only one call. The number of calls decreases as the frequency of customer service interactions increases. In contrast, customers who churned made significantly more calls, with a fluctuating pattern as the number of customer service calls rises. This suggests that churned customers likely faced persistent issues, prompting frequent calls to customer care. It's possible that unresolved issues or poor customer service contributed to their decision to leave.A graph of blue and orange squares

Description automatically generated

Multivariate analysis:

This involves analysing all the numerical variables and examining the relationships between them. To aid in this process, a heatmap has been created as a visual representation of the correlation matrix shown below.



The heat map visualizes the relationships between different numerical variables in your dataset, ranging from -1 to 1 and can be explained as follows:

1. **Diagonal Elements**: All diagonal elements are 1.00, as a variable is perfectly correlated with itself.
2. **Positive Correlations**: Values closer to 1 indicate a strong positive relationship between variables (e.g., as one variable increases, the other also increases). These are represented by the red squares.
3. **Negative Correlations**: Values closer to -1 indicate a strong negative relationship (e.g., as one variable increases, the other decreases). These are represented by blue squares.
4. **Weak or No Correlation**: Values close to 0 (near white or light blue/red) indicate little to no linear relationship between variables.

Therefore, we can observe that strong correlations where red dominates, especially along total minutes, total charges, and corresponding times of day (e.g., day, evening, and night variables). On the other hand, weak correlations are visible in a lot of the pairs, with many correlations close to 0.