SYRIATEL CUSTOMER CHURN ANALYSIS

Abstract

Customer churn has become a critical concern in the competitive telecommunications industry, where retaining existing customers is significantly more cost-effective than acquiring new ones. This report focuses on analyzing customer churn for SyriaTel, a telecommunications company in Syria, with the aim of understanding customer behavior and predicting churn accurately.

Contents

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 soaring 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 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 be incorporated into their customer retention management, such as whether to improve the service experience, design initiative-taking 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 generate 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:

- **Accuracy metric:** Measures the total number of correctly identified instances. An accuracy of between 75% and 85% is desired.
- **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 is sourced from Kaggle and contains records of SyriaTel customers. The dataset contains 3,333 rows with each row representing a customer record.

In addition, the dataset has twenty-one columns which broken down as follows:

i. **Customer usage**: The following columns provide further insight to the customer phone usage based on time of day:
    a. Usage during the day - total day minutes, total day calls, total day charge
    b. Usage during the evening - total eve minutes, total eve calls, total eve charge
    c. Usage at night - total night minutes, total night calls, total night charge
    d. No of voicemail messages

ii. **Plan subscription:** These columns give us a view of the plans that each of the customer has:
    a. International Plan
    b. Voice mail Plan

iii. **Unique customer details**. The columns falling under this section are:
    a. State
    b. Account length
    c. Area code
    d. Phone Number

iv. **International phone usage/customer service**: The columns under this category include:
    a. Total intl calls
    b. Total intl minutes
    c. Total intl charge
    d. Customer Service Calls
v. Likelihood of churn: Column Churn which takes on two values:
    a. True – Customer has churned.
    b. False – Customer is yet to churn.

DATA PREPARATION AND ANALYSIS

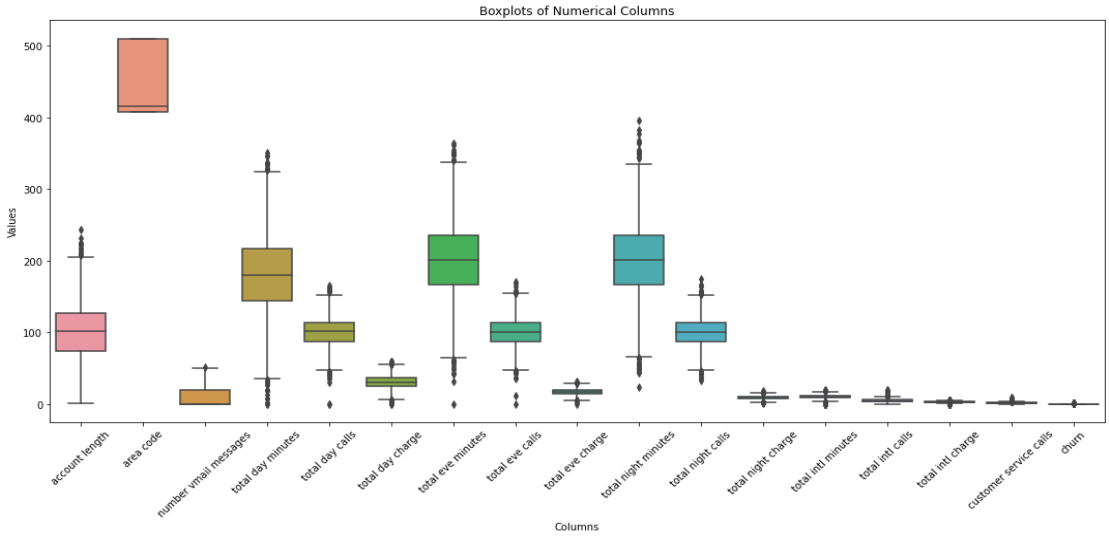
Data Preparation

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 outliers. This has been done step by step as explained below:

1. **Null / Missing values**: The dataset does not contain any missing data in any of the twenty-one 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 were completed. 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.



Majority of the columns do have a few outliers as expected given that the calling patterns may differ from customer to customer and some far removed from the median. However, this pattern is similar amongst the unique features we are looking at. For instance, the number of calls has a similar pattern of outliers regardless of whether it is 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

The aim of exploratory data analysis is to get further insights on SyriaTel customer behaviour as well as check which columns will be suitable to function 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 distinctive 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.

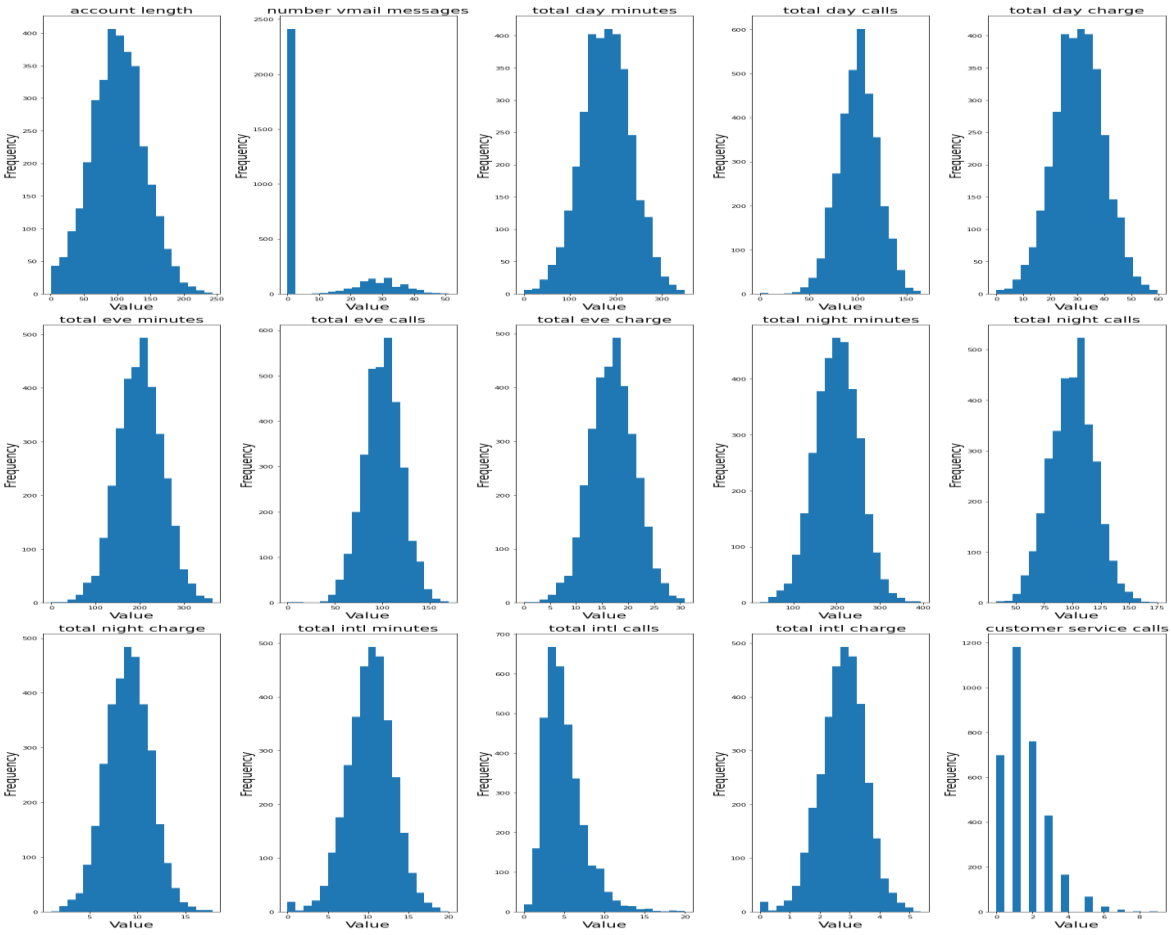
a. Numerical Columns

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 is the output:



From the diagram above, most of the numerical columns appear to be normally 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 approximately 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 700 customers having had no contact with customer care.

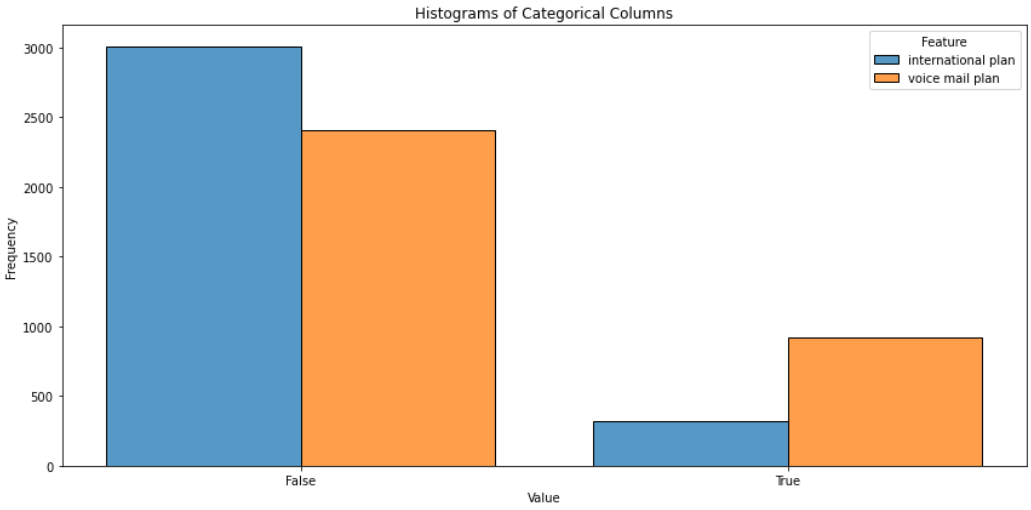
Finally, some of the columns such as "*area code*" and "*account length*" do not seem to be giving a lot of information with regards to customer behaviour. As such, these will be dropped, and all the other columns can be used as features to our predictive model.

b. Categorical columns

The categorical columns looked at are split as follows:

i. **Plan details**

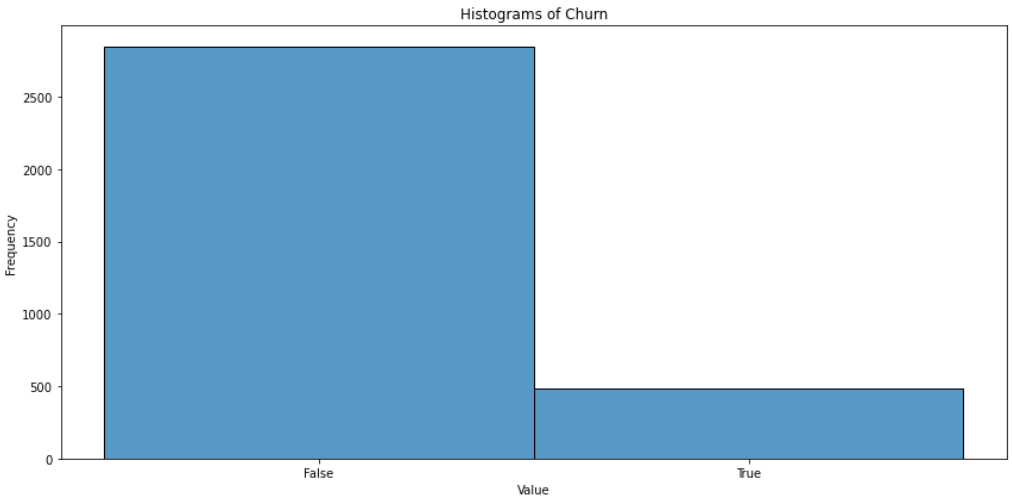
The columns falling under plan details are “international plan” and “voice mail plan”. These are represented by the histograms in the diagram below.



It can be observed that roughly 3,000 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.

ii. **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.



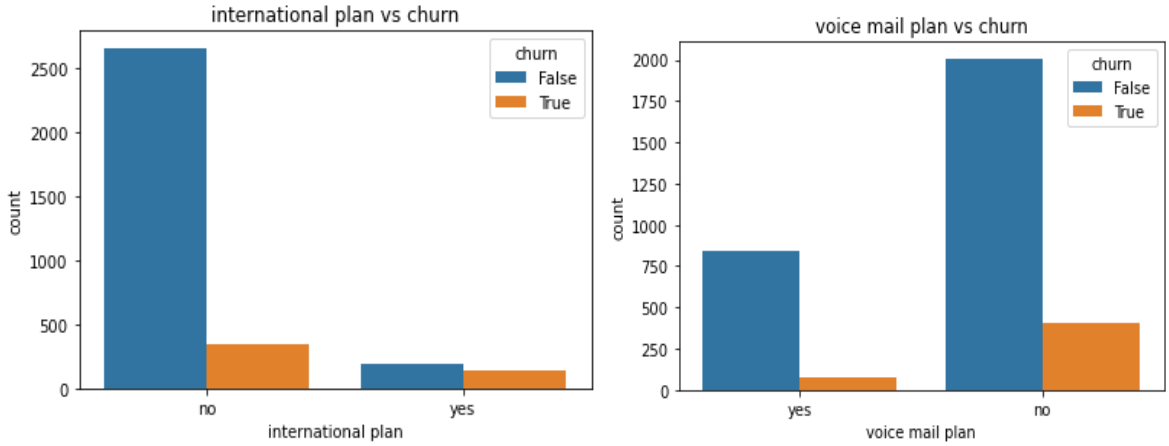
It can be observed that a lot of the customers have not churned. Roughly 2,800 customers are still with SyriaTel while roughly 500 customers have left. Therefore, we need to understand if any of the features looked at before may have caused this churn leading us to the next section which will delve more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

2. *Bivariate Analysis*

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

a. 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.

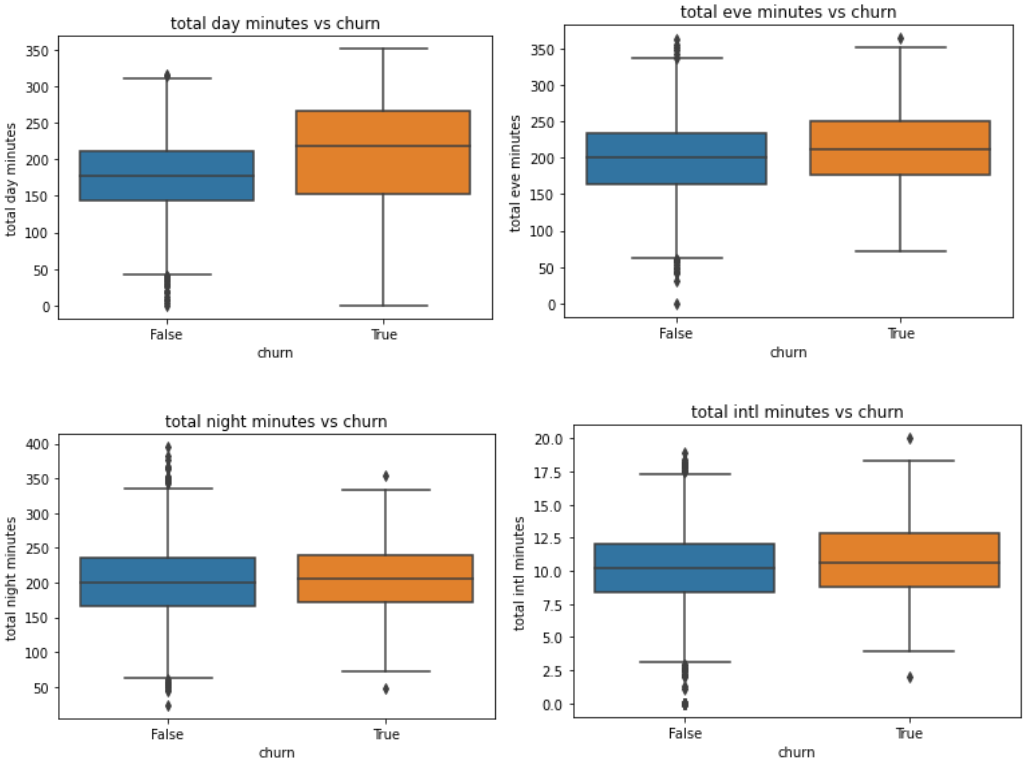


Similar pattern is observed with customers on the voicemail plan. However, the proportion of the customers who churned compared to those who are still active was much more pronounced on the voicemail plan. This proportion was quite close on the international plan. Therefore, there was high churn for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

b. Relationship between Churn and the Numeric Features

i. 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.

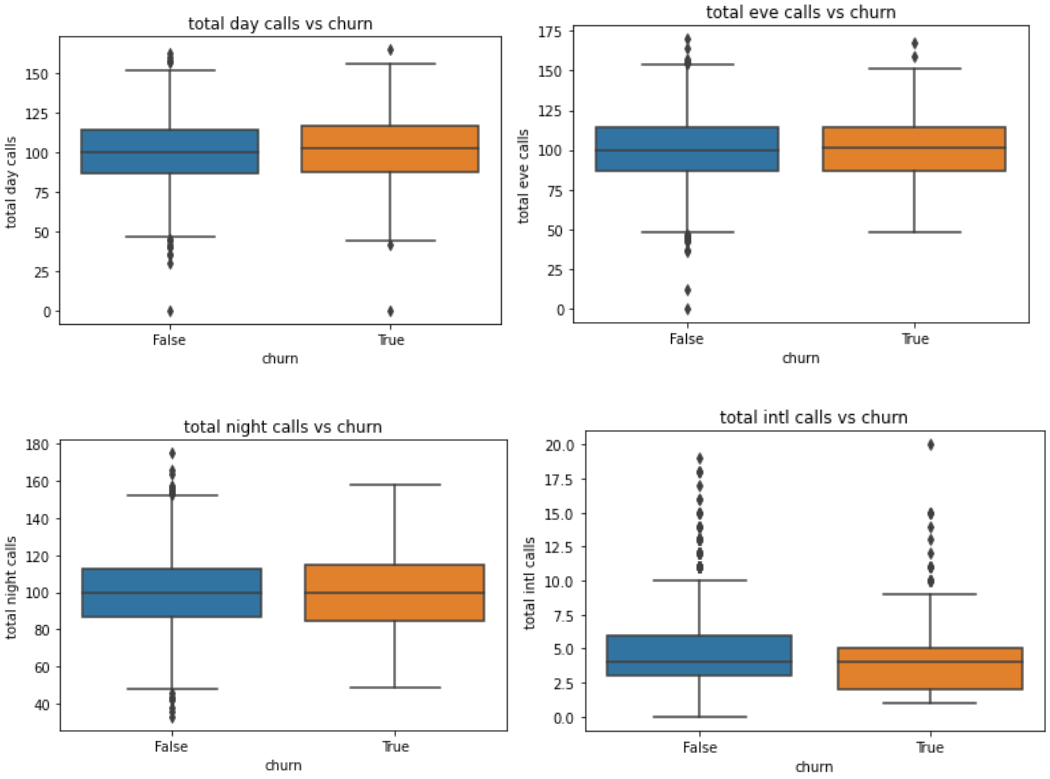


This is quite evident during the day period where customers who churned spent a median period of around 200 minutes on the phone compared to 180 minutes for active customers. As for the evening and night, the difference is quite negligible with both sets of customers (those who churned and those who did not) having a similar median amount spent on the phone. The variation observed is also quite similar amongst both sets of customers. A stark 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.

ii. Relationship between churn and number of calls

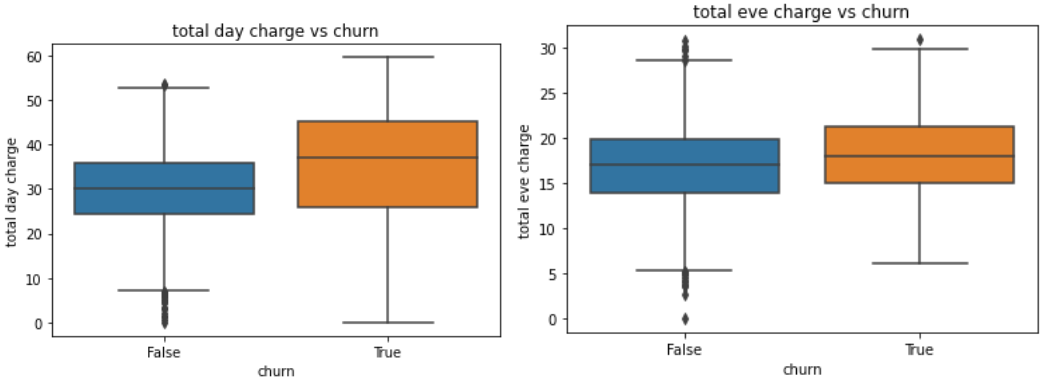
Unlike the relationship between churn and total minutes, there is no noticeable difference in the number of calls made by churned and active customers. Both groups make a similar number of calls during the day, evening, and night, with a median of roughly 100 calls. The variation in the number of calls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

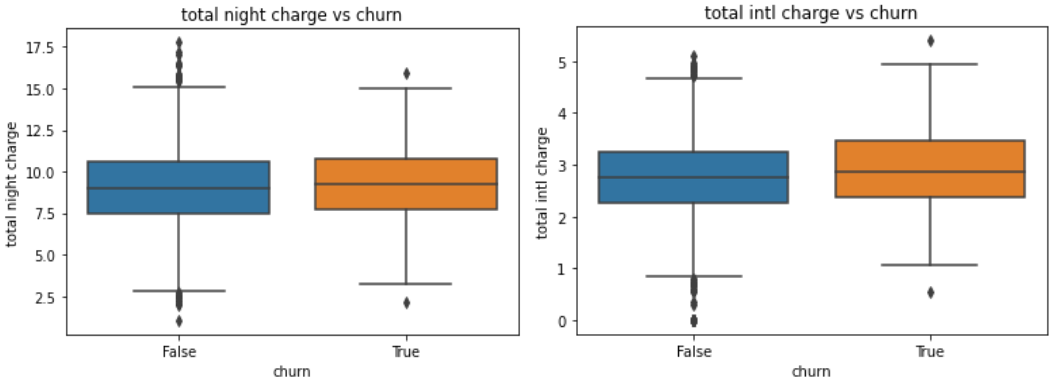
overall (a median of about 5 calls). Interestingly, churned customers tend to make fewer international calls, while active customers make slightly more.



iii. Relationship between churn and phone charges

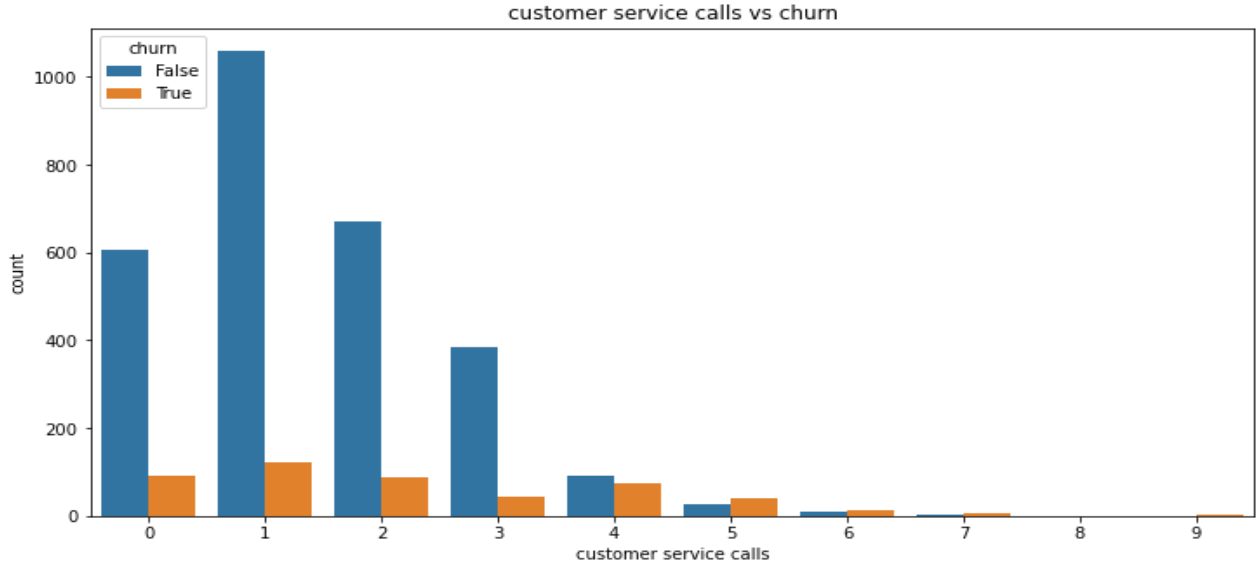
The boxplots show that total charges are higher for customers who churn compared to active customers. Charges are highest for calls made during the day and decrease as the day progresses into the evening and night. During the day, churned customers pay a median of around $35, while non-churned customers 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.





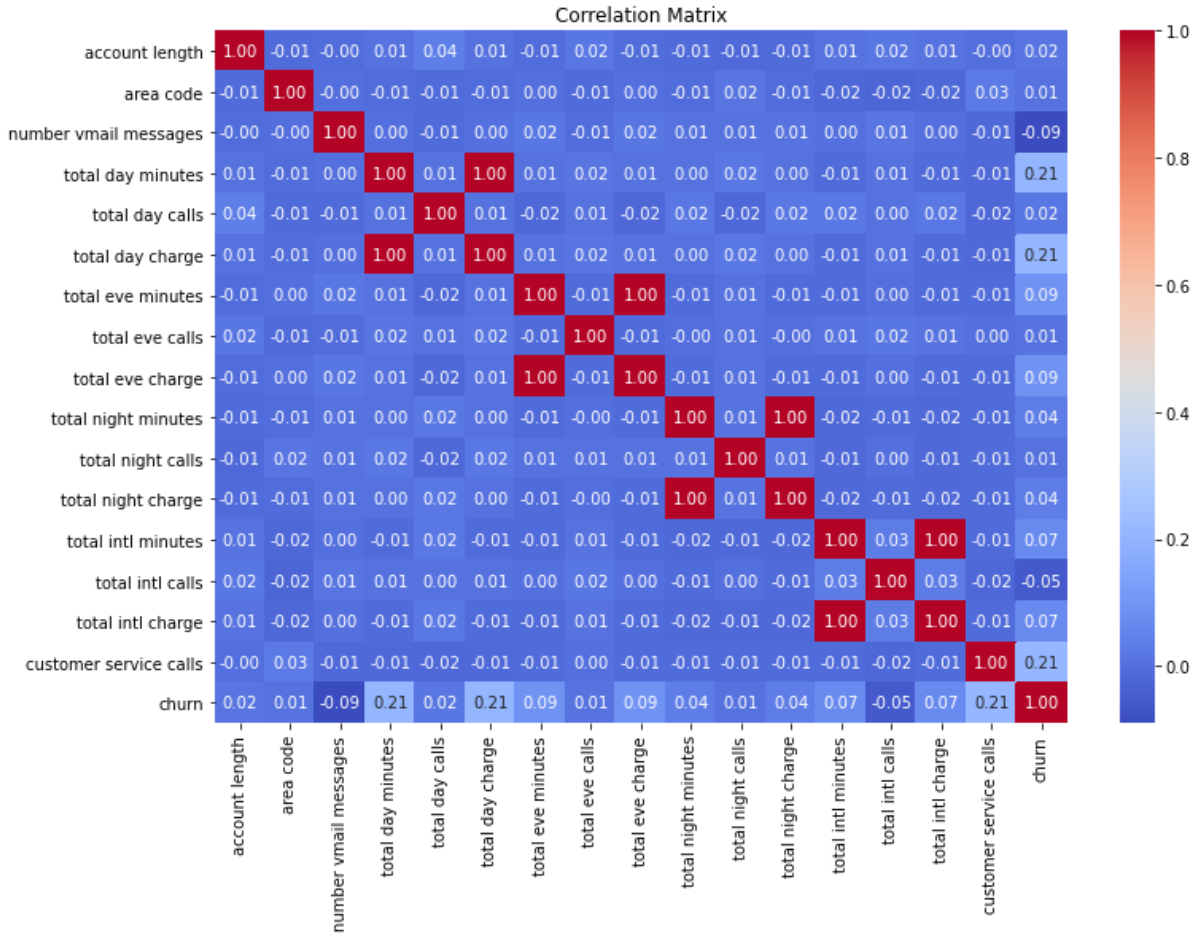
iv. **Relationship between churn and customer service**

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 is possible that unresolved issues or poor customer service contributed to their decision to leave.



3. *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.

From the heatmap, strong correlations, indicated by red, particularly among variables like total minutes, total charges, and their respective times of day (e.g., day, evening, and night) can be observed. In contrast, many variable pairs show weak correlations, with values close to 0. The variables most positively correlated with churn are total day minutes, total day charges, total international charges, and customer service calls. Conversely, churn shows a negative correlation with the number of voicemail messages, suggesting that offering voicemail services may help retain or attract customers. High day and international charges, combined with customer service interactions, appear to be significant factors contributing to customer churn at SyriaTel.

MODELLING

Data Preprocessing

Following the exploratory data analysis, next steps is to process the data based on the findings from the exploratory data analysis in readiness for modelling following the steps listed below:

i. *Drop irrelevant columns*

The following columns will be dropped as they will not be used going forward:

- **state** and **area code**: Contains the location of the customer, which would be useful in an inferential study but not a predictive study like this.
- **account length** and **phone number**: Contain the customer's unique identifier which does not necessarily provide us with information that would be useful in predicting churn.
- **total day minutes**, **total eve minutes**, **total night minutes** and **total intl minutes**: From the correlation matrix, these are correlated with *total day charge*, *total eve charge*, *total night charge* and *total intl charge,* respectively. To prevent multicollinearity of features we need to remove one of the columns for a more robust model.

ii. Label-encode the categorical variables

The following binary categorical variables are mapped to numbers (0 and 1):

- international plan: The values Yes/No are encoded to 1/0
- voice mail plan: The values Yes/No are encoded to 1/0
- churn: The values True/False are encoded to 1/0

iii. Identify target and features

Identify which columns represent features and which column represents the target. Recall that in this instance, we are trying to predict customer churn so that will be the target. All other remaining columns will be the features.

iv. Check for imbalance in the target variable

The target variable *Churn* has a class imbalance issue as majority of the customers are still with SyriaTel. Thus, if we had a model that always picked customers who did not churn (majority class) then we would expect an accuracy score of around 86%.

```
------------Distribution of Churn----------
0    2850
1     483
Name: churn, dtype: int64


0    0.855086
1    0.144914
Name: churn, dtype: float64
```

v. Perform a Train Test split

Split the data into training and test data sets. The model will be built using the training data set and its predictive power evaluated on the test set.

Model Building and Evaluation.

Given this is a classification problem, we will be using Logistic regression and Decision Tree algorithms to build the predictive model. The approach will be to start with a baseline model and refine it to get the model with the best performance. Metrics for evaluation used will be precision, recall, accuracy, F1 score and AUC score.

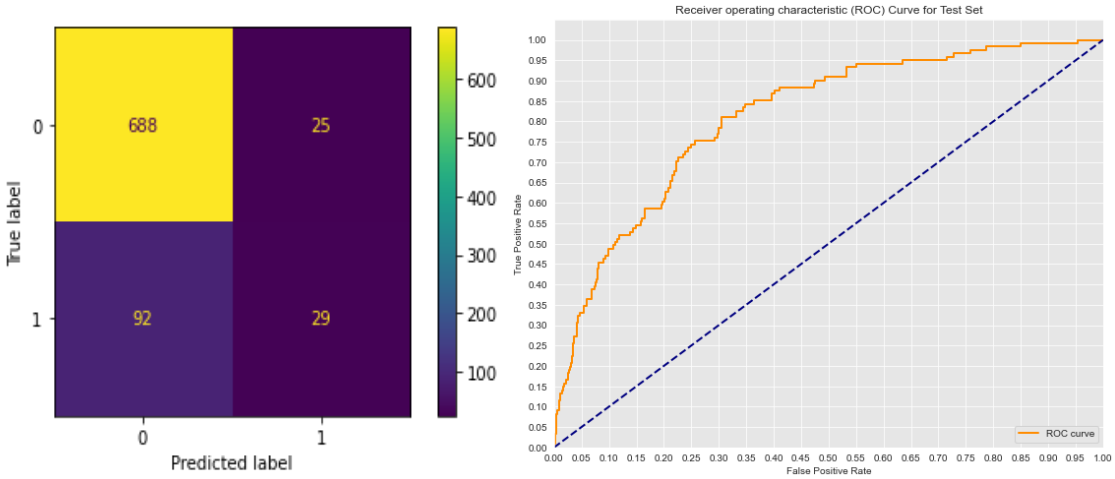
Logistic Regression:

Baseline Model: The baseline model for this analysis is a logistic regression model with the default model parameters and a random state = 42 implemented through the Scikit learn library. The random state is specified to ensure we get the same results when we run the model more than once.

Evaluation

The baseline model has an AUC of 0.811. In addition, the recall is around 24% and the precision is at 53%. Accuracy is at 86 % which is like a model that predicts the majority class all the time (customers who did not churn). Looking at the confusion matrix we can see that the number of false negatives is quite high at 92 (i.e. Customers who have

churned but are classified as if they did not churn). Thus, it does seem the model is penalizing the minority class due to the class imbalance. We will now proceed with building additional logistic regression models by tweaking the hyperparameters to rectify this imbalance.



Model with balanced weights

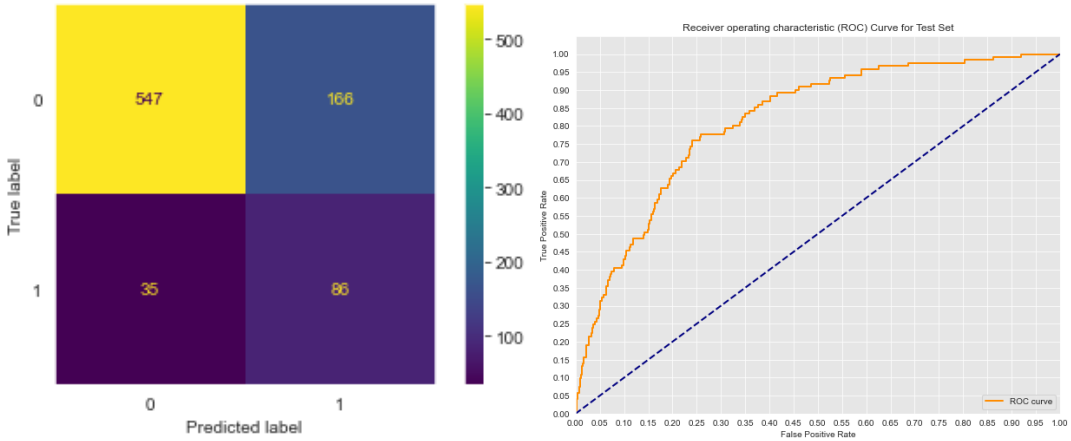
Due to the class imbalance in our target variable *churn*, the model is biased towards predicting the majority class (customers who did not churn) leading to inferior performance on the minority class, which happens to be our class of interest i.e. customers who churned.

Thus, setting class_weight="balanced" adjusts the weights assigned to each class in the loss function inversely proportional to their frequency in the training data. As a result, the minority class receives a higher weight, increasing its influence on the model during training while the majority class receives a lower weight, reducing its dominance in the model's decisions. Applying resampling technique like SMOTE would lead to overfitting in this case.

Evaluation

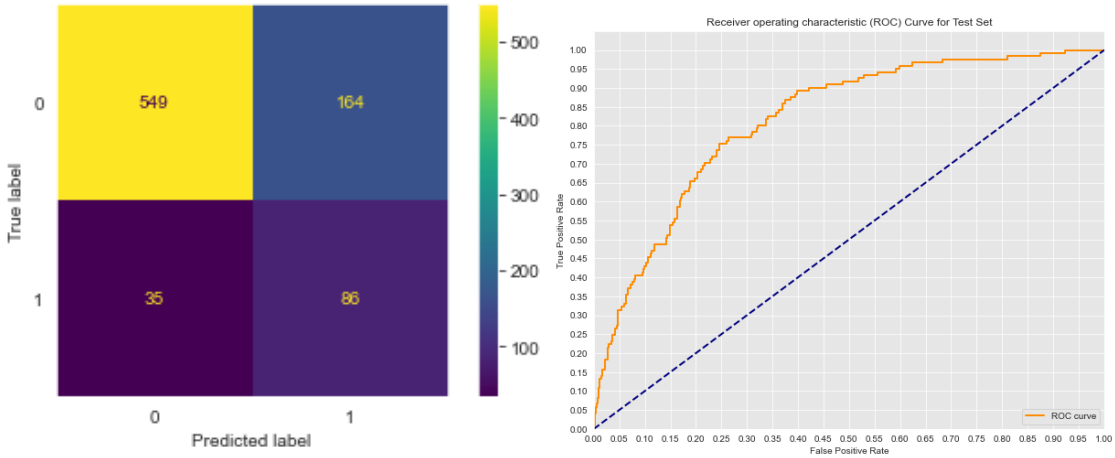
The AUC of this model is slightly higher than the baseline model AUC of 0.811 indicating better model performance. In addition, the test recall jumped from 23% to 77%. F1 score has also gone up from 0.33 to 0.46 hence model seems to be capturing customers who churned much better than the baseline model. This is evident in the confusion matrix, where the number of false negatives has reduced from 92 customers to 35 customers.

Comparing our test to training metrics, we seem to be getting slightly better metrics on the training data hence indicating that we could be overfitting.



Model with more regularization

To reduce overfitting that we observed in the previous model, regularization has been increased to reduce the overfitting by reducing the value "C" in the logistic regression from the default 1 to 0.1



Evaluation

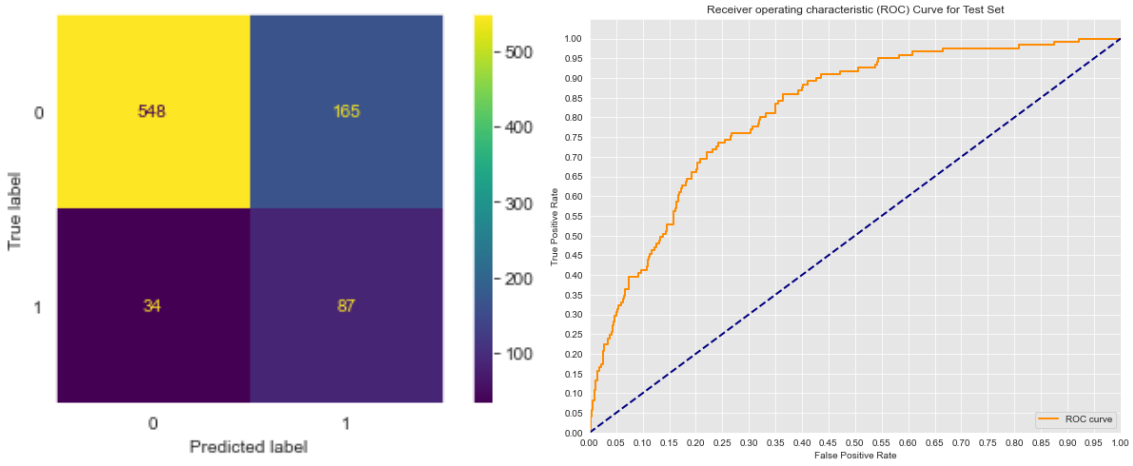
After increasing regularization, we see no major difference in the AUC of the model. This is evident in the confusion matrix above as the numbers are like the previous model. The metrics are slightly better compared to the model with balanced weights, i.e. precision, recall, accuracy and F1 score have slightly increased. Even though it is not a substantial change, we will proceed with this low regularization.

Model with alternative solver

Finally, a model using the liblinear solver was generated given the results were not getting better with the default solver which uses the Ridge (L2) penalty. This was used in combination with the Lasso (L1) penalty while increasing the max iterations to 10,000 to allow the gradient descent algorithm to take more steps in finding an optimal solution.

Evaluation

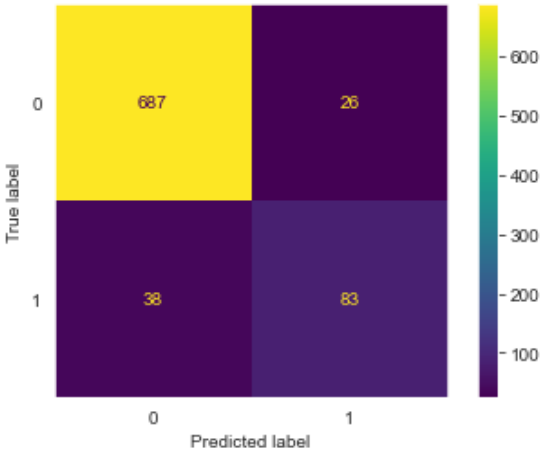
This model metrics are slightly better than all the three logistic models we have looked at thus far. AUC is quite like the model with less regularization while recall increased by a percentage point to 72% pushing the F1 score up to 0.47. Looking at the confusion matrix, the model captures less false negatives (customers who have churned being captured as not churned). However, it still classifies a substantial number of customers who have not churned as having churned (false positives) with the number being equal to 165 customers. As such, next step is to use the decision tree algorithm and evaluate if our results improve.



Decision Tree model:

Following on from the logistic regression model evaluation, a decision tree model is generated using the Scikit learn library and fit it to the training data. This will follow similar steps of beginning with a vanilla decision tree model and then prune it to check if we can get better results based on our test results.

Evaluation



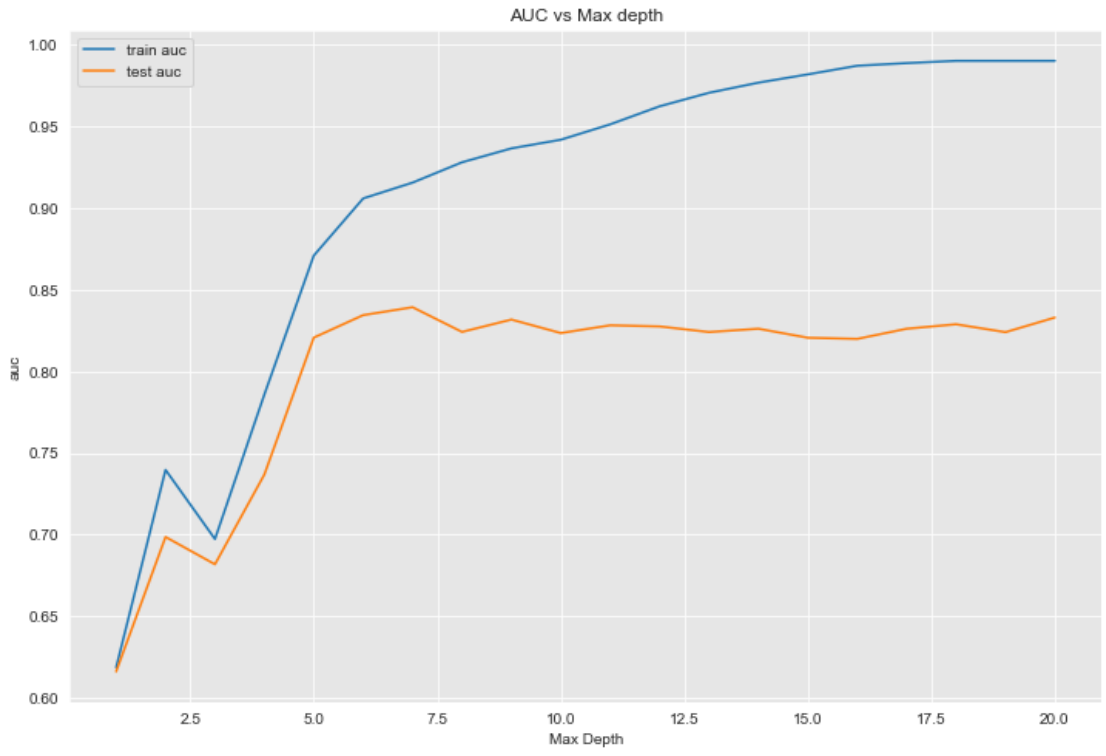
From the evaluation metrics, the decision tree has a higher accuracy and precision compared to the logistic models looked at thus far. This is quite evident in the confusion matrix plot above as the number of false positives is quite low at 26 customers. However, in terms of recall, it is much lower at 68% hence still misclassifying customers who have churned worser than the logistic regression. AUC is higher hence showing the model is performing much better than all models looked at thus far.

Pruned Decision Tree model:

This is an optimized version of the vanilla decision tree model where the following parameters will be tweaked to improve the model's performance:

a. *Max depth*

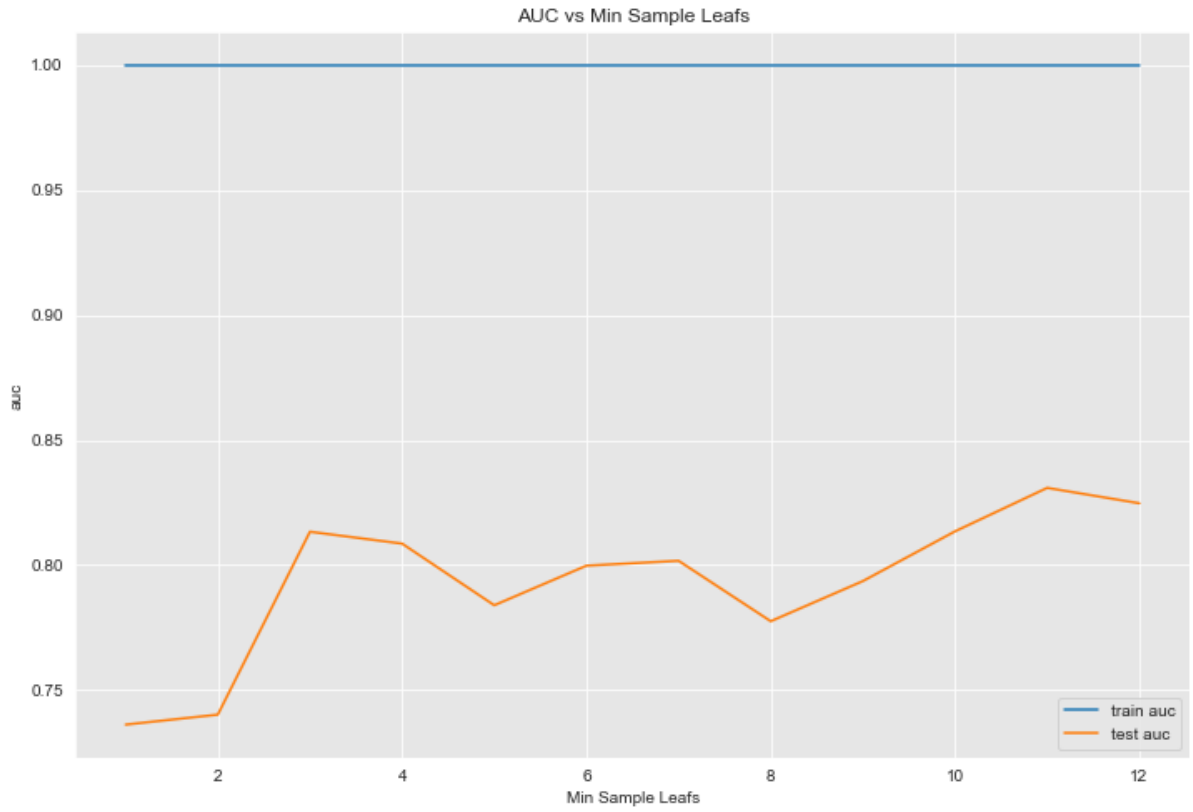
To check for the maximum depth, we will iterate over *max depth* values ranging between 1 and 20 and train the decision tree for each depth value. Following this, we will calculate the training and test AUC for each run then plot a graph to show underfitting/overfitting as well as the optimal value



From the graph above, the training error decreases with increasing tree depth which shows signs of overfitting. Test error increases after depth=7. Thus, there is nothing more to learn from deeper trees (some fluctuations, but not stable). Hence the optimal value for max depth is 7.

b. *Max Features*

To check for the maximum features, we will iterate over *max features* values ranging between 1 and 12 and train the decision tree for each depth value. Following this, we will calculate the training and test AUC for each run then plot a graph to show underfitting/overfitting as well as the optimal value



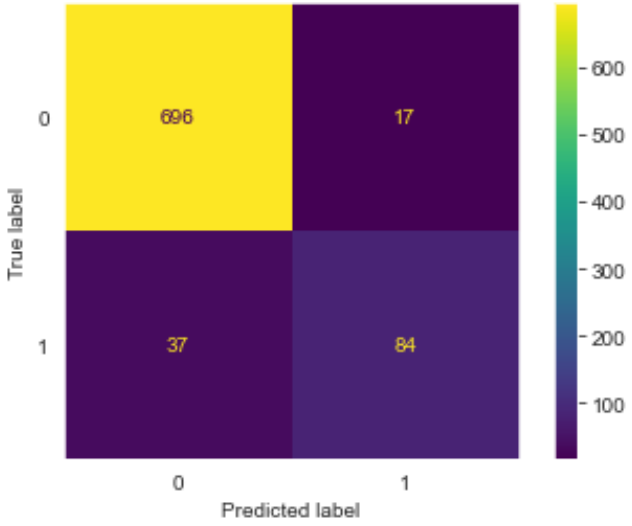
From the graph above, we see no effect on the training dataset when varying the maximum features as the training AUC remains flat. On the other hand, there are fluctuations observed in the test AUC there is a bit of fluctuations observed with the optimal value being 11.

Fitting the pruned tree model

With these two updated values, *max depth*=7, *max features*=11, the decision tree is pruned accordingly, training data fit to it and test data ran to evaluate its performance.

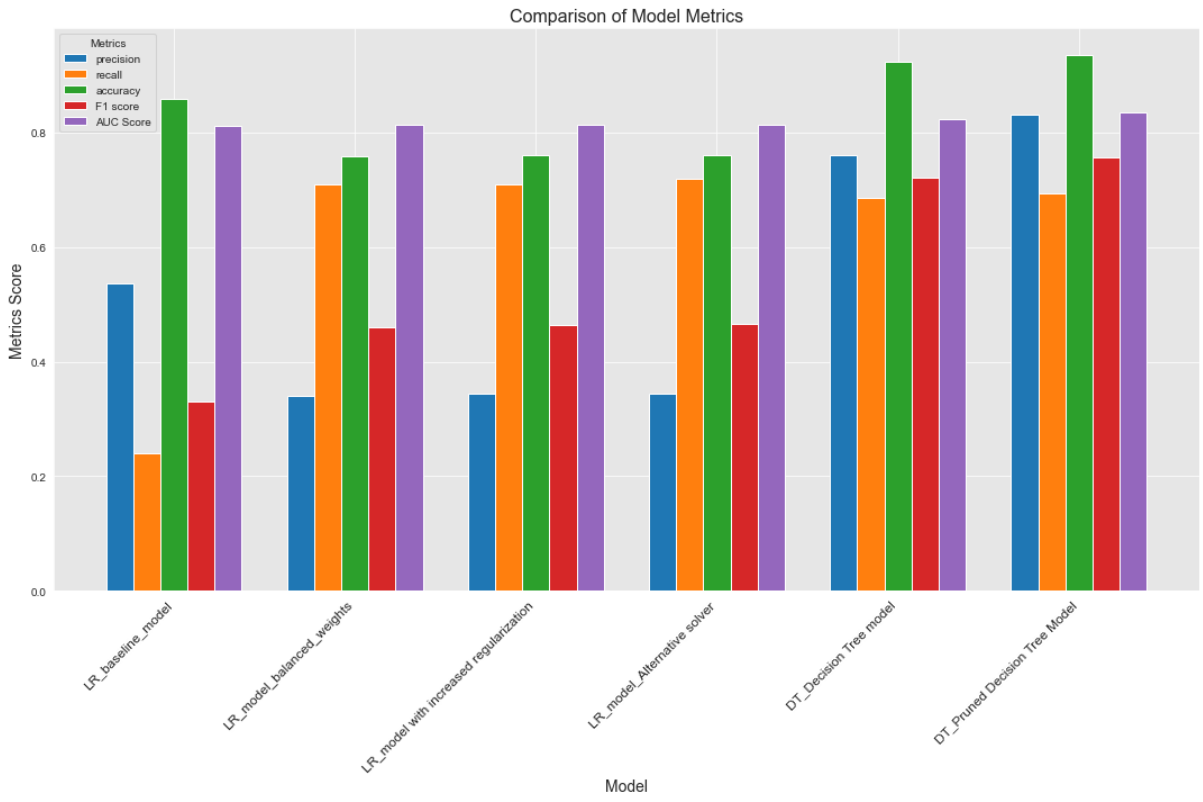
Evaluation:

Using the pruned decision tree gave better metrics compared to the vanilla decision tree model. AUC score was much higher as well as the precision, accuracy, F1 score and recall. Taking all metrics into consideration, this pruned decision tree model seems to be performing the better than the vanilla decision tree model and this can be confirmed by the confusion matrix below.



FINAL MODEL EVALUATION:

All the model evaluation findings have been compiled and plotted to get a better view of how the models compare against each other as well as against our success metrics.



From the bar graph above, decision tree models perform better than the logistic regression models. Additional check is how the models performed against our metrics of success:

- Accuracy of between 75% and 85% is desired: All the models evaluated were able to meet this threshold
- Precision of between 50% and 60% is desired: Only the baseline logistic model and the decision tree models met this threshold. The baseline logistic regression model was within this range while both decision tree models (unpruned and pruned) were able to surpass it.
- Recall of between 60% and 70% is desired. All the models were able to meet this criterion except the baseline logistic regression model. Logistic models containing tuned hyper parameters performed the best i.e. model with alternative solver, model with balanced weights and model with less regularization

- F1 score of between 0.55 and 0.65 is highly desirable: Only the decision tree models were able to meet this criterion. Due to the precision scores being low, all logistic regression models had a F1 score less than 0.5
- Area under the curve (AUC): A higher result indicates a more accurate model performance. The pruned decision tree had the highest AUC. However, the differences observed across all models evaluated was quite minimal.

Taking all this account, the pruned decision tree model meets all the metrics of success. It strikes a balance across precision and recall. Thus, it will be able to give a balanced view of customers who will churn as well as those who will not churn. Also, the accuracy score and the AUC is the highest making it the model that has the most accurate predictive performance.

RECOMMENDATION

Below are the business recommendations for SyriaTel based on the analysis performed:

- SyriaTel should go for the pruned decision tree model when predicting whether a customer will churn. It will be able to give the most accurate predictive view of customers who will churn. This model provides a balanced view across all the evaluation metrics, and it is easy to interpret the model results to stakeholders in the company.
- Key factors making customers leave SyriaTel are high international calling charges, day charges and poor customer service. Customers who left made the highest number of day and international calls, but their charges were still quite high. In addition, customers who made more calls to customer service were more likely to leave as opposed to those who did not.
- SyriaTel can reduce customer churn by reducing their day and international calling charges as they seem to be highly uncompetitive. They could produce discount schemes to reward customers who call more often. It should also improve on their customer service through training of its customer care agents. Finally, it should improve its service in general across the board to ensure calls by customers to customer care are reduced to a minimum.

CONCLUSION

This analysis looked at SyriaTel customer data to determine a predictive customer churn model. In addition, customer's patterns have been studied to determine the reason behind customers leaving and ways in which this can be mitigated. The pruned decision tree model has been found to be the best predictive model for the data analysed as it gives the most accurate and balanced view of when a customer will churn. Moreover, customers in SyriaTel churned because of high calling charges as well as poor customer service. It is imperative for the company to offer deals in the form of discounts to customers who call much more often and improve their customer service by training their customer service agents.

NEXT STEPS

Deployment of the model to end users is the next step. The model will be exported into a format suitable for integration through embedding it into a software application (such as web or mobile application) where end users can input their data and receive predictions.

Other sophisticated models need to be considered such as Random Forest, XG Boost to get better predictive performance. In addition, a much bigger dataset should be sought to increase the training and predictive power of the models.