SYRIATEL CUSTOMER CHURN PROJECT REPORT *by Sarah Joshua*

OVERVIEW AND BUSINESS UNDERSTANDING.

Syriatel is a major telecommunications company in Syria. It offers a range of services, including mobile and fixed-line telephony, internet access, and data services. The company has a significant market share in Syria and plays a vital role in connecting people and businesses across the country.

However, due to the ongoing conflict in Syria, Syriatel's operations have been significantly impacted. The company has faced challenges such as infrastructure damage, power outages, and security threats. Despite these difficulties, Syriatel has continued to provide essential communication services to the Syrian people.

Customer churn is a significant concern for telecommunications companies. It refers to the loss of customers over a specific period. This can be a costly problem, as acquiring new customers is often more expensive than retaining existing ones.

Apart from the ongoing conflicts, customer churn in Syriatel may be caused by other various factors such as:

- Poor service quality
- High pricing
- Lack of Innovation
- Poor customer service
- Competitive pressure

PROBLEM STATEMENT

SyriaTel is a telecommunications company in Syria that aims to utilize customer data to uncover patterns and predict the probability of customer churn, enabling the company to take steps to retain high value customers and minimize value losses.

OBJECTIVES

Main Objective

- Find a machine learning model for correct classification of Churn and non-churn customers.

Secondary objectives

1. To analyze customer behavior patterns and identify key factors influencing churn decisions.

2. To identiwhich of the key factors that affect churn need to be given more attention or priority in order to reduce customer churn as soon as Possible. up.
3. To evaluate the effectiveness of retention strategies and measure the impact on customer churn rate

Metrics of success

Based on previous studies and research, the following are the measures that evaluate the success of models.

- Accuracy:80% total number of True positives (correctly identified instances)
- Precision:50% measures how predictive the model is in regard s to the number of true positives against false positives
- REcall:75% The ability of the model to identify churners correctly
- F1-Score: between 0.55 and o.65 measures the accuracy of the predictive model's performance
- Area Under the Curve (AUC): The higher the AUC the more accurate the performance

DATA UNDERSTANDING

The csv file used for this project was downloaded from Kaggle, which is a fantastic platform for data scientists and machine learning enthusiasts. It offers a wealth of datasets, competitions, and a vibrant community to learn and collaborate.

“bigml.csv” is a dataset that contains 3333 rows and 21 columns’

Here is an explanation of the columns that we have in this dataset.

1. state: The state where the customer resides.

2. account length: The number of months the customer has been with the telecommunications company.

3. area code: The area code of the customer's phone number.

4. phone number: The customer's phone number.

5. international plan: Whether the customer has an international plan.

6. voice mail plan: Whether the customer has a voicemail plan.

7. number vmail messages: The number of voicemail messages the customer has.

8. total day minutes: The total number of minutes1 used during the day.

9. total day calls: The total number of calls made during the day.

10. total day charge: The total charge for day time usage.

11. total eve minutes: The total number of minutes used during the evening.

12. total eve calls: The total number of calls made during the evening.

13. total eve charge: The total charge for evening time usage.

14. total night minutes: The total number of minutes used during the night.

15. total night calls: The total number of calls made during the night.

16. total night charge: The total charge for night time usage.

17. total intl minutes: The total number of minutes used for international calls.

18. total intl calls: The total number of international calls made.

19. total intl charge: The total charge for international calls.

20. customer service calls: The number of calls made to customer service.

21. churn: Whether the customer has churned (left the company) or not.

DATA PREPARATION AND ANALYSIS

Data Preparation(Data cleaning)

In data preparation and analysis, here is a list of what we are required to do.

1. **check for missing values.**
   After checking for missing values, we realized that there were no columns with missing values. Therefore, we don't have to worry about handling any.
2. **check for duplicate values.**
   After checking for duplicate values, we realized that there are no rows with duplicate values. Therefore, we also don't have to worry about handling any duplicate values
3. **check for null values**.
   After checking for null values, we realize that there are no null values in our data
4. **check for outliers**
   I checked for outliers and realized most of the columns had outliers but with very small percentages. Since the small percentages could not make very significant changes, I decided to leave them as they were.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an important step that helps understand the data, identify patterns, detect anomalies, and uncover relationships between variables before applying any machine learning models.

EDA can be done in 3 important steps.

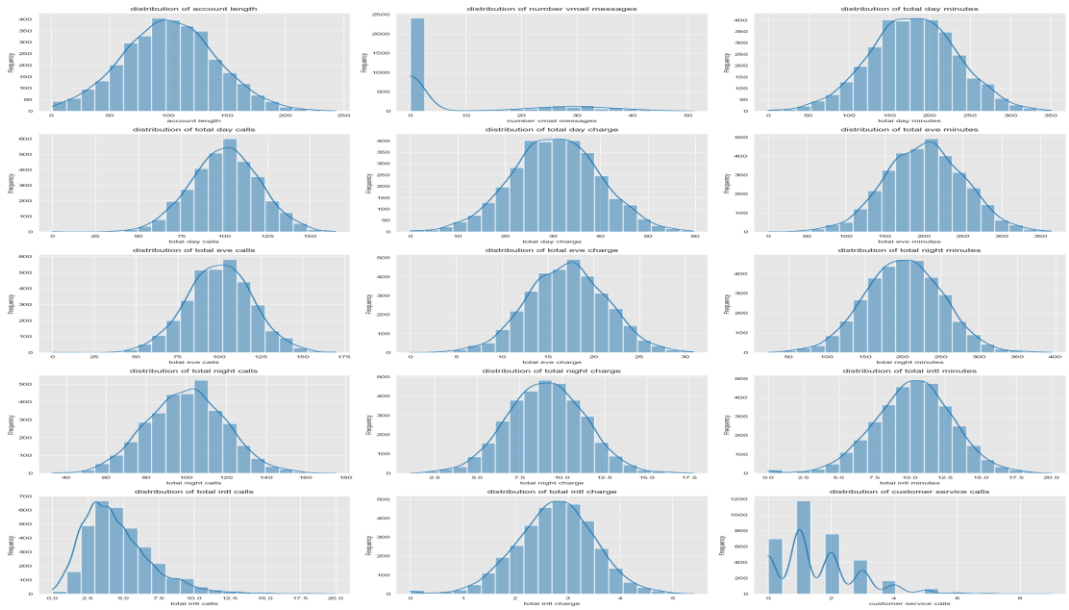
1. **Exploring Univariate Analysis**.

This refers to analyzing one variable at a time to understand it's distribution or It's Key Characteristics.

Below is a list of analysis that I did on Univariate analysis.

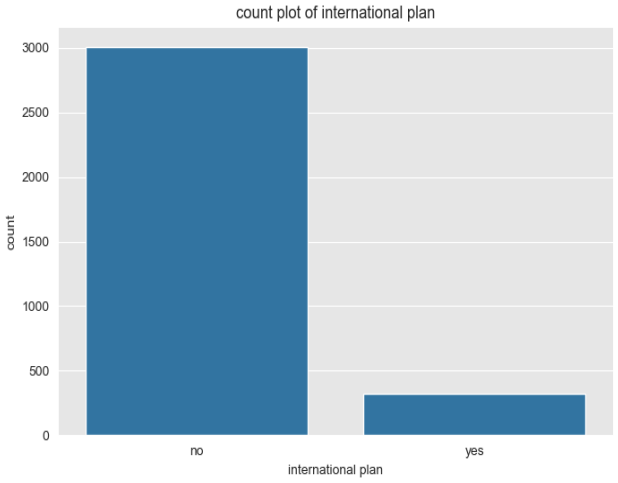
a). Plotting the Distribution of features

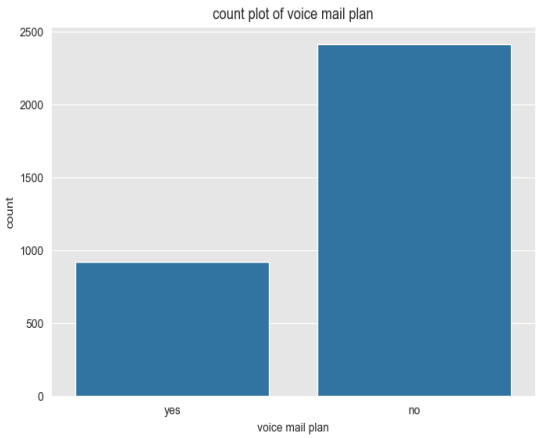
The histograms were plotted to get an understanding of their distributions.



From the output above, it is evident that most of the numerical variables are normally distributed. However total intl calls looks to be slightly positively skewed. number vmail messages, on the other hand, is extremely positively skewed as majority of the customers rarely have voicemail messages. Also, we can deduce that majority of the SyriaTel customers do not use the voice mail messaging services and rarely contact customer service as roughly 1200 only contacted the customer care once.

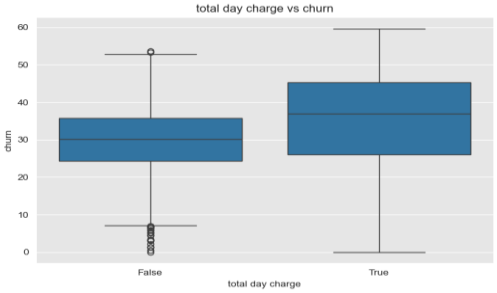
I also plotted the following graphs in Univariate analysis

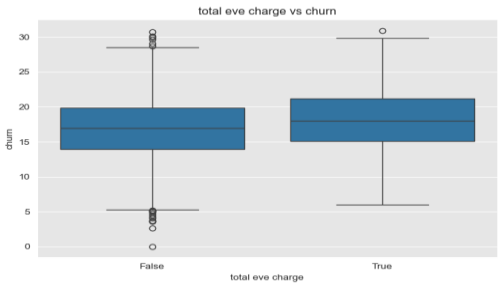




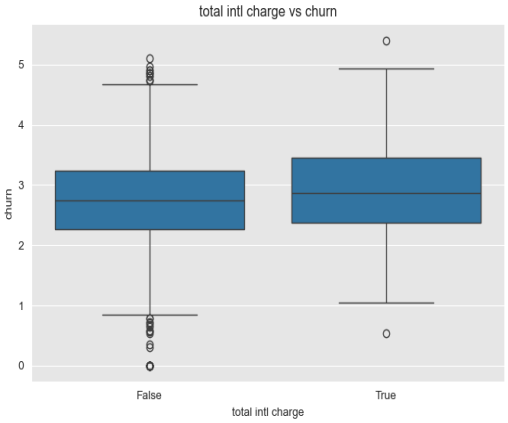
2, **Exploring Bivariate analysis**

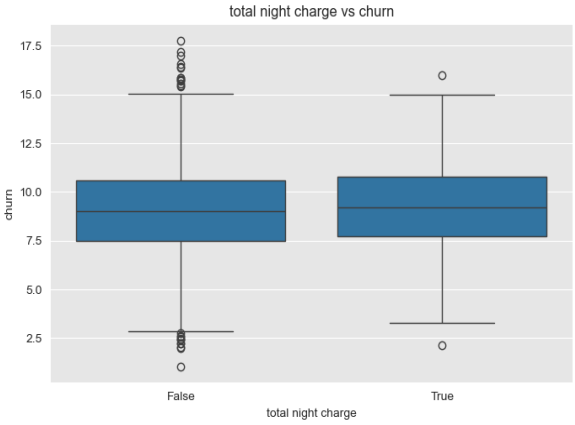
Analysing the relationship between two variables to explore potential correlations or dependencies



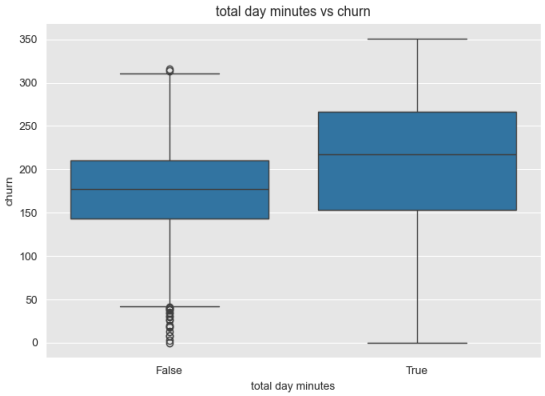


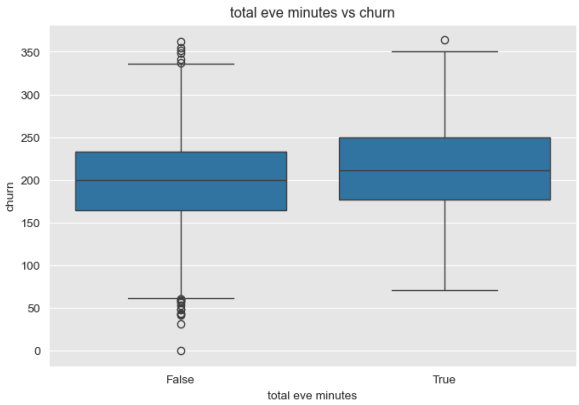
.

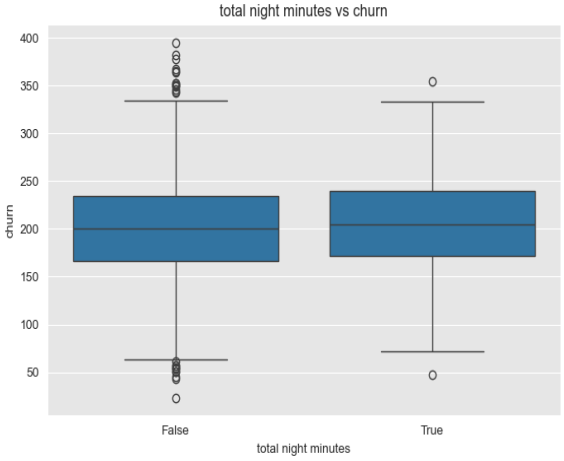




We observe that the total charges are higher for customers who churn compared to the active customers. The charge is the highest for calls made during the day and it reduces as the day progresses to the evening and night. During the day, customers who churned paid a median of around 35 dollars while those who did not churn paid roughly 30 dollars. This is quite a steep difference and it could explain why some of the customers are leaving.

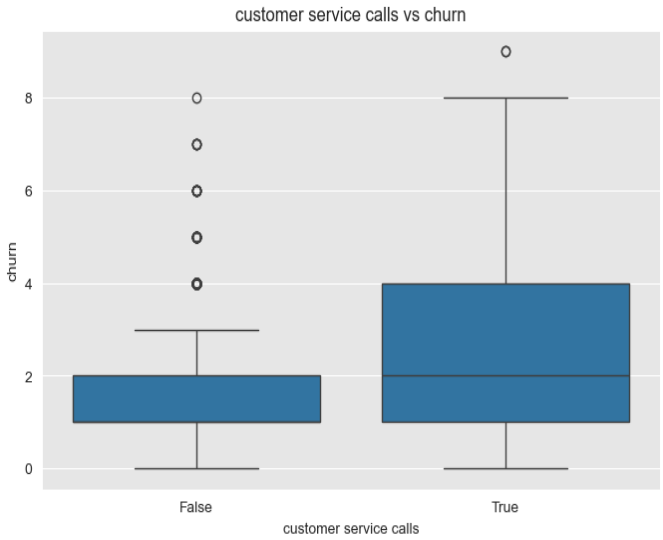






As can be observed by the box plots above, customers who churned had a higher median number of minutes spent on the phone. This is quite evident during the day period where customers who churned spent a median period of around 200 minutes on the phone compared to 180 minutes for the active customers. 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

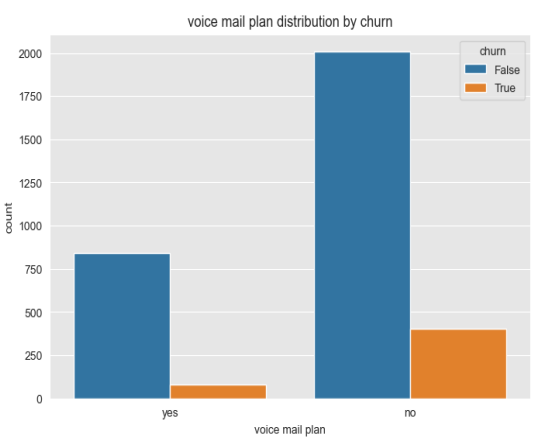
median time spent on the phone. The variation observed is also quite similar amongst the both sets of customers. A huge difference is observed in the amount spent on total international minutes as customers spent a median duration of 10 minutes, which is quite removed from the day to day calling periods.

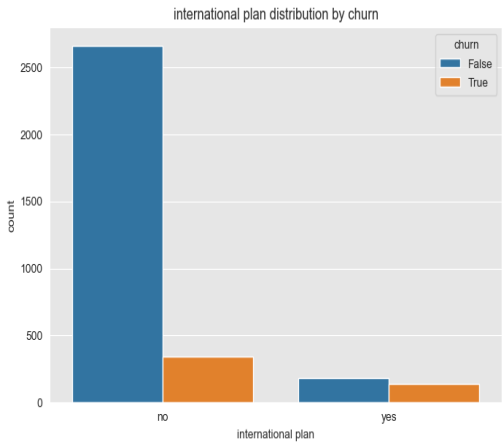


.

The box of the churners is wider than that of the non-churners. This means that their usage patterns vary significantly.

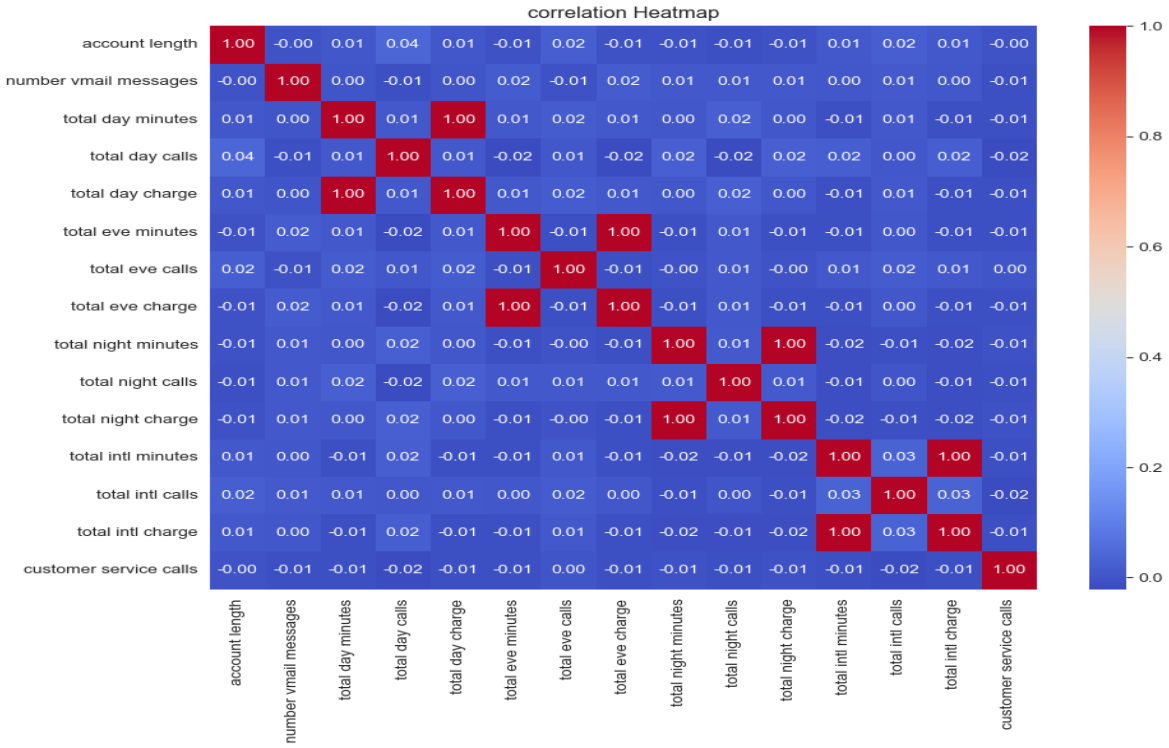
The e median of churners is higher compared to non-churners. This shows that customers whmake more customer service callses a morere likely to churn





- For those that use international plan, there are more non-churners than churners.
- For those that don't use international plan, there are also more non-churners compared to churnes
- There are more people that don't use voicemail plan than those who use voicemail plan
- For those that use voicemail plan, there are more non-churners than churners.
- For those that don't usevoicemaill plan, there are also more non-churners compared to churnes

3. Exploring Multivariate analysis.



The above plot is a plot showin correlation between the numeric columns. its shows that there is multicoleniarity wethin the dataset. The multicollinearity is there because total day charge isequal to total day minutes, total night charge is equal to total night minutes, total international charge is equal to total international minutes.

We can therefore deduce that the presence of a service allows customers to have vmail messages helps in retaining/attracting customers. However, the total day charges, international charges are quite prohibitive. These charges coupled with customer service could be the causes of customers leaving SyriaTel.

MODELING

Data Preprocessing

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. It involves preparing raw data to ensure it is clean, consistent, and suitable for analysis or modeling. High-quality data preprocessing can significantly enhance model accuracy and reliability.

Key Steps in Data Preprocessing

1. Handling Multicolinearity

From the previous subtopic, we saw that the heat map showed multicolleniarity within our dataset.

I went ahead and dropped the minutes columns because they were more or less the same or giving the same output as the charges columns.

2. Data Cleaning

I'm also going to drop state and account length columns because I'll not be using them

3. Encoding Categorical Data

Next, i'm going to do label encoding on the "international plan", "voice mail plan" and "churn" columns.

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

- international plan
- voice mail plan
- churn

I used mapping to label encode these columns

4. Feature Scaling

I do have a clear imbalance in my target variable churn given majority of the customers are still with Syria Tel. Therefore 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%. This clas imbalance issue will be looked at as part of building the model.

I did normalization to balance the target variable “churn”

5. Splitting the Dataset

I then went ahead and identified the x and y variables to be used for modelling, and I split the dataset into train and test sets.

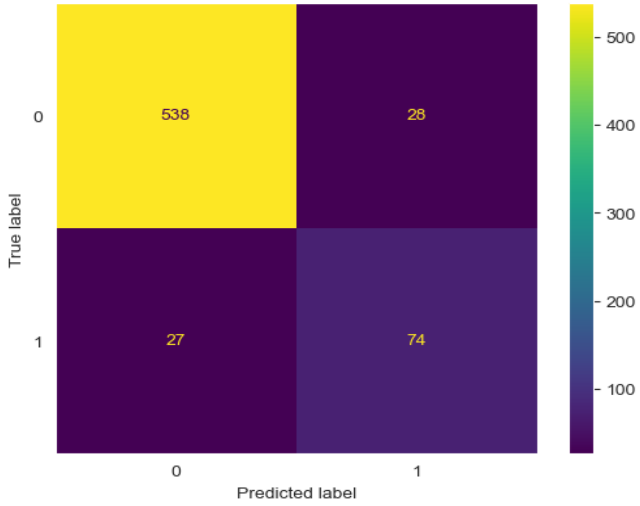
Decision Trees

A decision tree model is evaluated using the Scikit learn library and fit into the training data.

BUILDING A BASELINE MODEL

Our baseline model will be a logistic regression using the inbuilt model parameters

EVALUATION



Baseline Model Metrics:

Accuracy: 0.9175412293853074

Precision: 0.7254901960784313

recall: 0.7326732673267327

f1_score: 0.729064039408867

From the evaluation metrics, the decision tree has a high accuracy and precision. This is quite evident in the confusion matrix plot above as the number of false positives is quite low at 27 customers. In terms of recall, it's 73% , the accuracy is at 91%, the precision is at 72% and the AUC is 84.1.

I pruned the model by looking at the following:

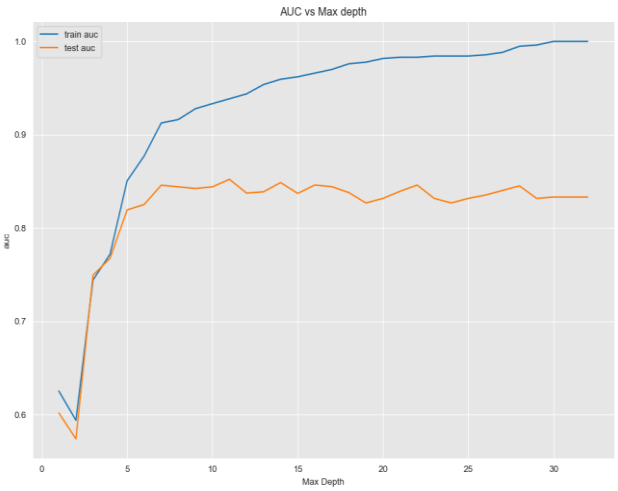
- Max_depth
- min_samples_leaf
- min_sample_splits
- maximum_features

max depth.

This refers to the maximum depth of the tree. It limits how deep the tree can grow which can help in preventing overfitting

To check for the maximum depth, I iterated over max_depth values ranging between 1 and 20 and trained 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

Evaluation



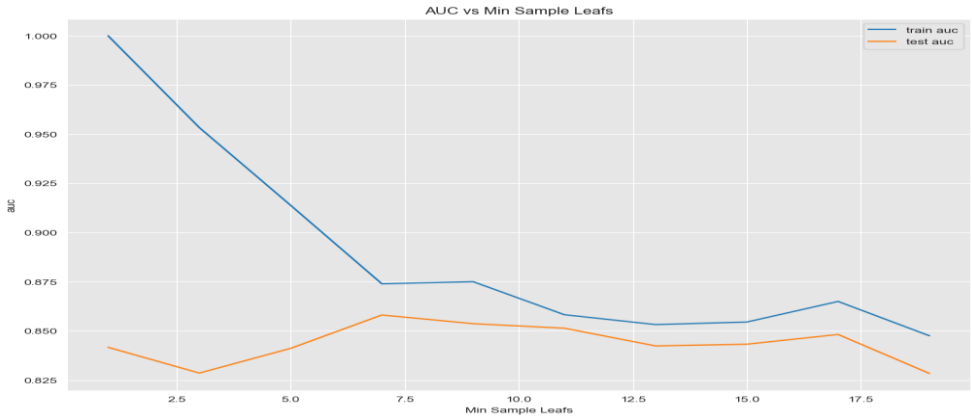
From the graph above, the training error decreases with increasing tree depth which definitely shows signs of overfitting. Test error increases after depth=7. Thus, there is nothing more to learn from deeper trees (some fluctuations, but not stable) Hence the optimal value for max_depth is 7

Minimum_samples_leaf

Refers to the minimum samples required to be at a leaf node. This ensures that leaf nodes have enough data points to make statistically significant decisions

To check for the minimun samples leaf, I iterated over minimum samples leaf values ranging between 1 and 20 and trained the decision tree for each depth value. Following this, we will calculate the training and test AUC for each run then plot a graph .

Evaluation



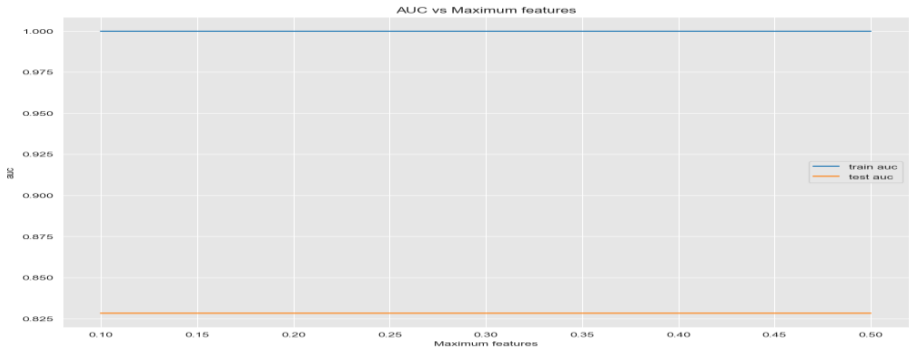
The region where the gap between train and test AUC is minimal, and the test AUC is near its peak (around `min_samples_leaf` = 7.5–10), is likely the best balance between bias and variance.

Maximum_features

Refers to the maximum number of features to consider when looking for the best split. This is meant to control the subset of features to evaluate at each split, which can help to reduce overfitting.

To check for the maximum features, I iterated over max_features values ranging between 1 and 12 and trained 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

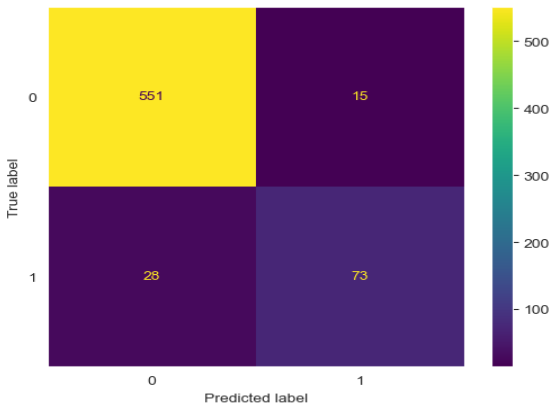
Evaluation



From the graph above, we see no effect on the training dataset as well as the testing dataset as the features of the training and testing dataset remains flat. This means that we cannot obtain any optimal values from the above graph.

Pruned model

Evaluation



Pruned model Metrics:

Accuracy: 0.9355322338830585

Precision: 0.8295454545454546

recall: 0.7227722772277227

f1_score: 0.7724867724867724

From the evaluation metrics, the decision tree has a high accuracy score and precision This is quite evident in the confusion matrix plot above as the number of false positives is quite low at 28 customers. In terms of recall, it's 72.27% , the accuracy is at 93.55%, the precision is at 82.95% and the AUC is 84.81%

[145]:

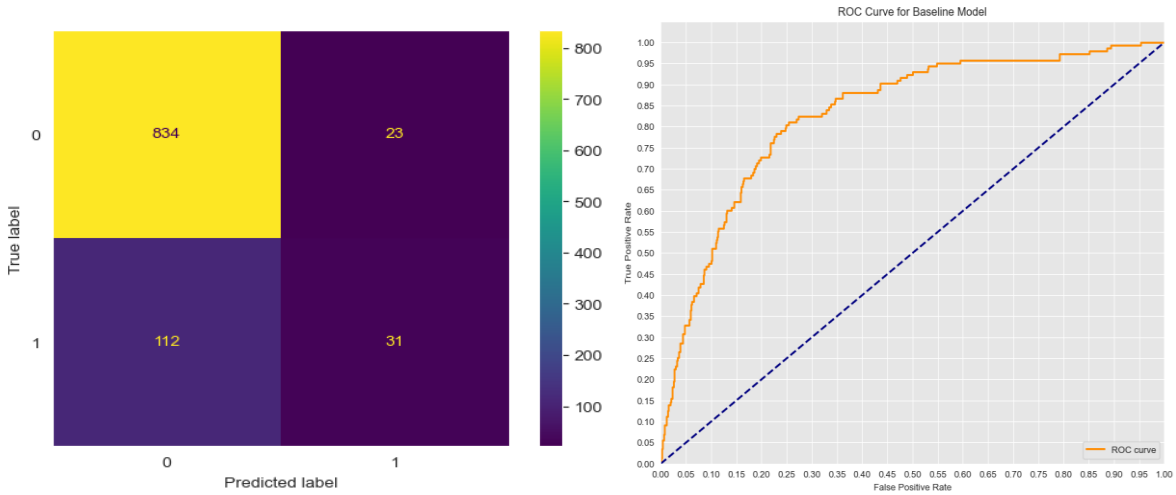
Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary and multiclass classification problems. Just like the decision trees, I also started bu building the baseline model.

Baseline Model

Our baseline model will be a logistic regression using the inbuilt model parameters

Evaluation



Baseline Model Metrics:

Accuracy: 0.865

Precision: 0.5740740740740741

recall: 0.21678321678321677

f1_score: 0.3147208121827411

Based on the evaluation metrics above, the baseline model has an AUC of 0.8305. In addition, the recall is around 22% and the precision is at 57%. Accuracy is at 87 % which is similar to 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 are quite high at 112 (i.e.customers who have churned but are classified as if they did not churn). Thus, it does seem the model is penalizing the

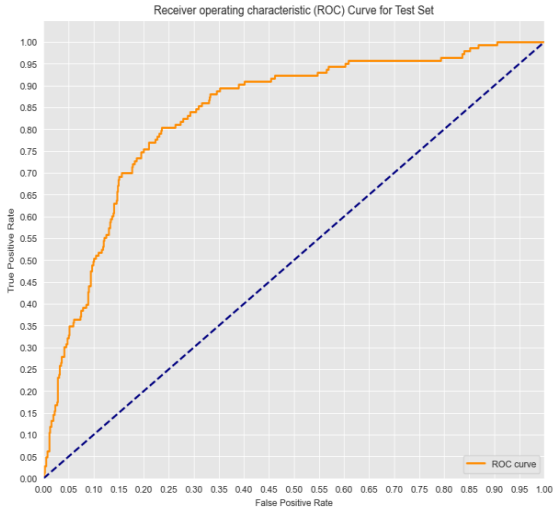
minority class due to the class imbalance. We will now proceed with building additional logistic regression models by tweaking the hyperparameters to rectify this imbalance.

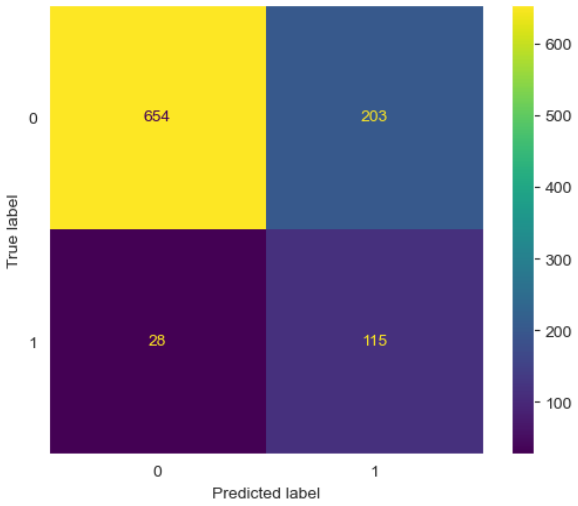
With balanced class

Due to the class imbalance we have noticed in our target variable churn, the model is biased towards predicting the majority class (customers who did not churn).This has lead to poor 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. Therefore, 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.

Evaluation





Oversampled Model Metrics:

Accuracy: 0.769

Precision: 0.36163522012578614

recall: 0.8041958041958042

f1_score: 0.49891540130151846

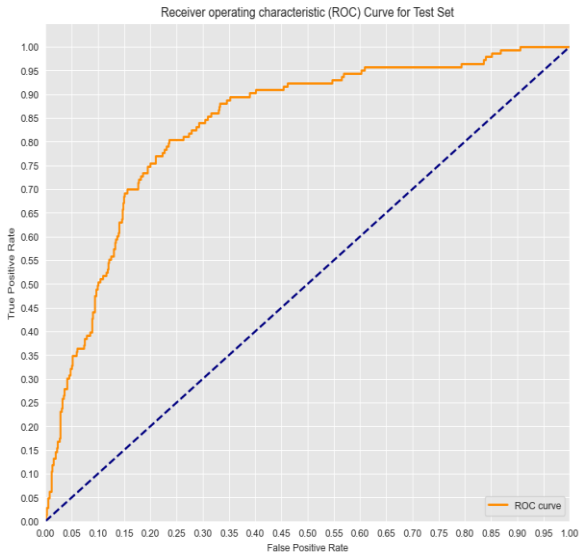
The AUC of this model is slightly higher than the baseline model AUC of 0.8347 indicating better model performance. In addition, the test recall has jumped from 22% to 80%. F1 score has also gone up from 0.315 to 0.499 hence model seems to be capturing customers who churned much better than the baseline model. This is evident in the confusion matrix, where the number of false negatives has reduced from 112 customers to 28 customers.

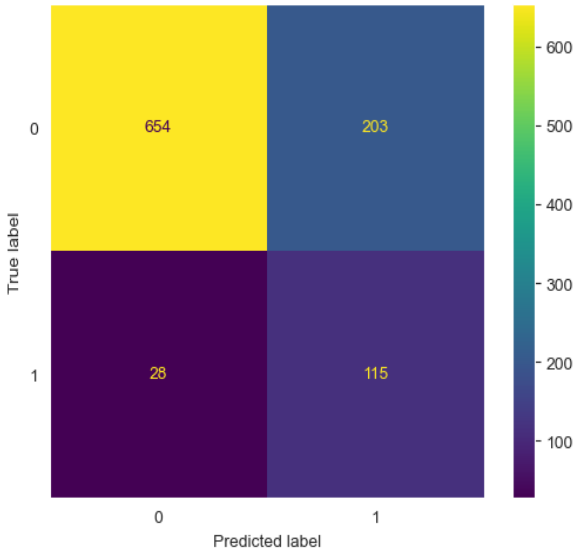
Comparing our test to training metrics, we seem to be getting slightly better metrics on the training data hence indicating that we could be overfitting.

With increased Regularization

One of the hyperparameters C which is a measure of regularization, has been set to 1 in our baseline model. We will increase regularization to reduce the overfitting by reducing C to a small number (e12). along with random_state=42. We will call this model model_increased_regularization

Evaluation





Regularized Model:

Accuracy: 0.769

Precision: 0.36163522012578614

recall: 0.8041958041958042

f1_score: 0.49891540130151846

After increasing regularization, we see no difference in the AUC of the model. The metrics are the same as the model_with_weights, i.e. precision, recall, accuracy and F1 score have not increased or reduced.

EVALUATION

model	DT baseline	DT pruned	LR baseline	LR balanced	LR reg
Precision	0.725	0.829	0.574	0.361	0.361
Recall	0.732	0.722	0.216	0.804	0.804
Accuracy	0.917	0.935	0.865	0.769	0.769
F1 score	0.729	0.772	0.315	0.499	0.499
AUC score	0.842	0.848	0.830	0.834	0.834

If we evaluate our models using our researched metrics of success:

- Accuracy:75% - 85% : Most of the models lied between this range, while others, especially from the deision tree models, surpassed the range.
- Precision:65% - 75% : The logistic regression models did not perform well with regards to our prediction metrics. Both models from the decision trees performed well
- REcall:70% - 80% : for recall, only the liner regression baseline model did not reach our predicted metrics, The other models lied between the range of our predicted metrics.
- F1-Score: between 0.55 and o.75 : The Decision Tree models performed well with regards to our predicted metrics of succes, but the Logistic Regression models did not.
- Area Under the Curve (AUC): All models increase their AUC but the pruned model from decision trees had the highest "AUC".

RECOMMENDATIONS

1. I would recommend Decision Trees as the best Machine Learning model to use when trying to predict customer churn for SyriaTel Company.
2. The key factor influencing customer churn in SyriaTel Company is Total charges for calls, which greatly impacts the total minutes taken for calls I would recommend that day Call charges and international call charges to be reduced in order to reduce customer churn in SyriaTel.
3. I would also recommend strategies like rewarding minuites and tokens that can be redeemed to encourage or "entice" their customers to use Syriatel and inrease the number of calls.

CONCLUSIONS

This analysis on Syriatel was better achieved using the Decision Tree model. It helped us conclude that Syriatel company should consider reducing call charges and include incentives in order to reduce customer churn and attract more customers.

NEXT STEPS

The model can now be deployed to the end-users, and it can also be use to predict future similar cases in future.

The project only used two modelling types to achieve it's goals. Other models should be considered just incase they are able to produce better results.