Phase 3 Project

Name:Laura Kanda Mwichekha

Study Mode: Partime

Topic: SyriaTel Customer Churn Prediction

SyriaTel Customer Churn Prediction-Overview

Customer churn, the rate at which customers stop doing business with a company, is a critical metric for telecommunication companies. Churn can negatively impact revenue, customer lifetime value, and brand reputation. Therefore, predicting customer churn is essential for effective customer retention strategies.

The telecom provider SyriaTel wants to minimize revenue loss from client attrition. SyriaTel can enhance customer satisfaction and retain consumers by implementing focused retention methods aimed at identifying high-risk clients.

Project Objectives

Below are the objectives for the prediction model

1. Determine which customers are at danger of churning: Make an accurate guess as to which ones are most likely to cut off service.

2. Put focused retention methods into practice: Create specialized marketing strategies to speak to the unique requirements of at-risk clients and persuade them to stick with SyriaTel.

3. Boost client contentment Improve the whole customer experience to lower churn rate and increase steadfast loyalty.

4. Efficient resource allocation can be achieved by concentrating retention efforts on customers who are most likely to leave.

Project Overview

- Business Problem
- Data loading and understanding
- Data Preparation
- Modelling
- Evaluation
- Model of choice-Deciscion Trees
- Recommendation and future investigation
- Conclusion.

Business Problem

I have been tasked by SyriaTel to build a binary classification model to predict whether a SyriaTel customer will churn,i.e stop doing businesss in the near future.

The primary goal is to reduce financial losses due to customer churn

Data loading and Understanding

During the EDA, there are some findings that came up and needed to be addressed before getting into modeling. Below are the findings

Finding 1: Data type Conversion

The area code column was represented as an integer data type,the values were essentially labels or placeholders and not numerical values that can be used for mathematical calculations.It was therefore necessary to convert the data type of the area code column to a categorical datatype to accurately model .

Finding 2: High correlation - Multicollinearity

The heatmap analysis revealed several columns in the dataset to exhibit high levels of correlation with each other.This multicollinearity ,a condition where independed variables are highligh interrelated,can cause issues with our model,making it difficult to determine the individual impact of each target variable.It was therefore necessary to address multicollinearity

Finding 3: Outliers

I also observed several outliers with the data set illustrated by boxplots.This had the potential to impact the modelling process.The outliers in this case however are not abnormalities that needed to be eliminated. Rather, they represent a significant feature of the dataset that i had to take into consideration while modeling. Ensuring the robustness and correctness of the data requires an understanding of the nature and impact of these outliers.

Heat map Showing how the columns are correlated

Interpretation:

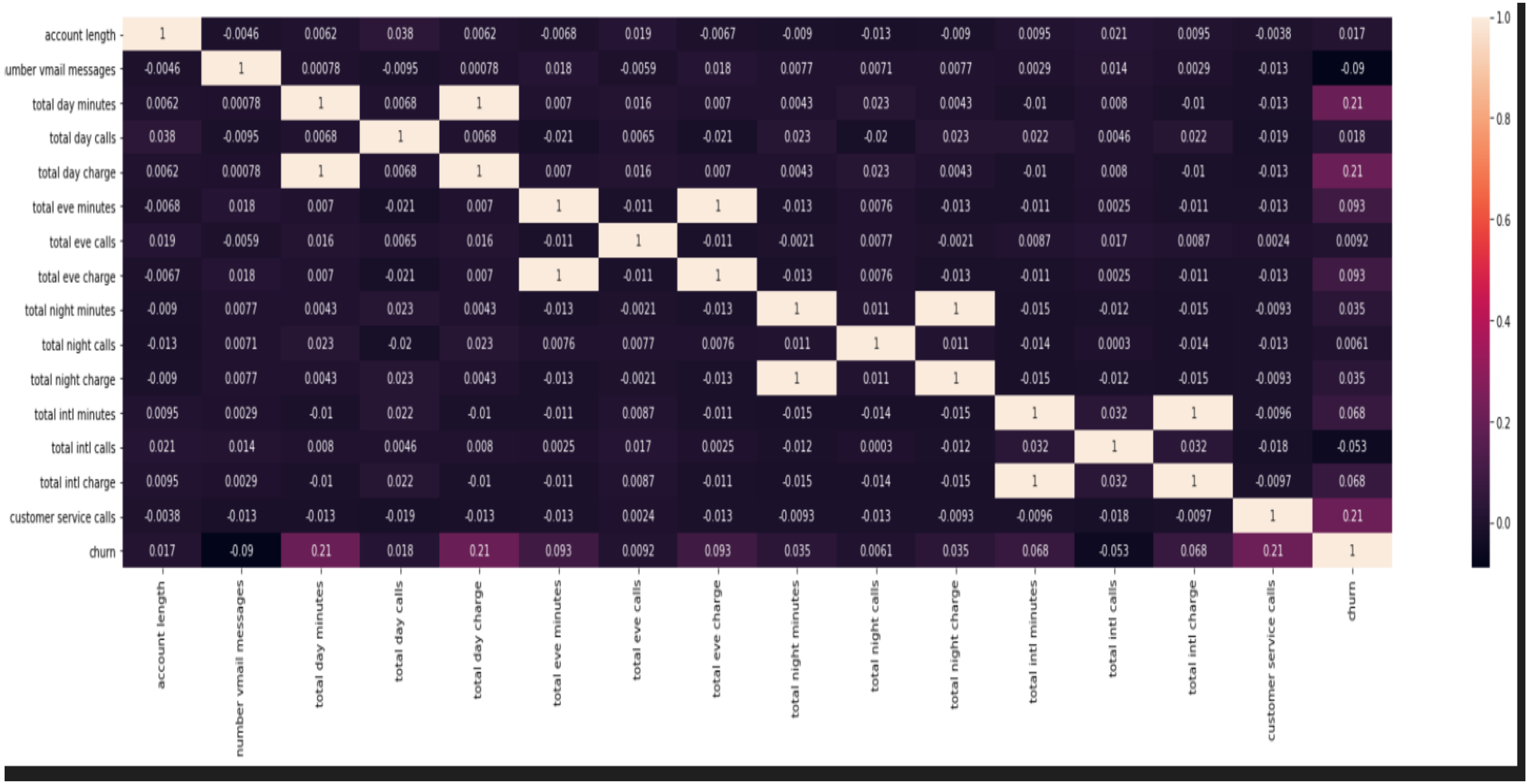
Colors in the heat map represent the correlation coefficient between features.

Red indicates a positive correlation, where higher values in one feature correspond with higher values in another feature.

Blue indicates a negative correlation, where higher values in one feature correspond with lower values in another feature.

White close to zero indicates no correlation between the features.

The intensity of the color represents the strength of the correlation. Darker colors represent stronger correlations (positive or negative)



	account length	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	churn
account length	1	-0.0046	0.0062	0.038	0.0062	-0.0068	0.019	-0.0067	-0.009	-0.013	-0.009	0.0095	0.021	0.0095	-0.0038	0.017
umber vmail messages	-0.0046	1	0.00078	-0.0095	0.00078	0.018	-0.0059	0.018	0.0077	0.0071	0.0077	0.0029	0.014	0.0029	-0.013	-0.09
total day minutes	0.0062	0.00078	1	0.0068	1	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
total day calls	0.038	-0.0095	0.0068	1	0.0068	-0.021	0.0065	-0.021	0.023	-0.02	0.023	0.022	0.0046	0.022	-0.019	0.018
total day charge	0.0062	0.00078	1	0.0068	1	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
total eve minutes	-0.0068	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
total eve calls	0.019	-0.0059	0.016	0.0065	0.016	-0.011	1	-0.011	-0.0021	0.0077	-0.0021	0.0087	0.017	0.0087	0.0024	0.0092
total eve charge	-0.0067	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
total night minutes	-0.009	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
total night calls	-0.013	0.0071	0.023	-0.02	0.023	0.0076	0.0077	0.0076	0.011	1	0.011	-0.014	0.0003	-0.014	-0.013	0.0061
total night charge	-0.009	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
total intl minutes	0.0095	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0096	0.068
total intl calls	0.021	0.014	0.008	0.0046	0.008	0.0025	0.017	0.0025	-0.012	0.0003	-0.012	0.032	1	0.032	-0.018	-0.053
total intl charge	0.0095	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0097	0.068
customer service calls	-0.0038	-0.013	-0.013	-0.019	-0.013	-0.013	0.0024	-0.013	-0.0093	-0.013	-0.0093	-0.0096	-0.018	-0.0097	1	0.21
churn	0.017	-0.09	0.21	0.018	0.21	0.093	0.0092	0.093	0.035	0.0061	0.035	0.068	-0.053	0.068	0.21	1

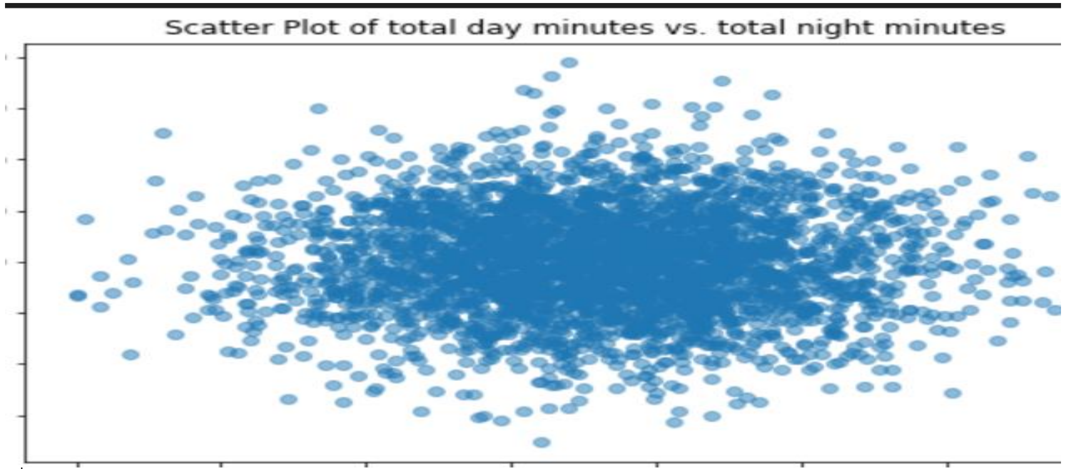
Scatter plot showing multicollinearity

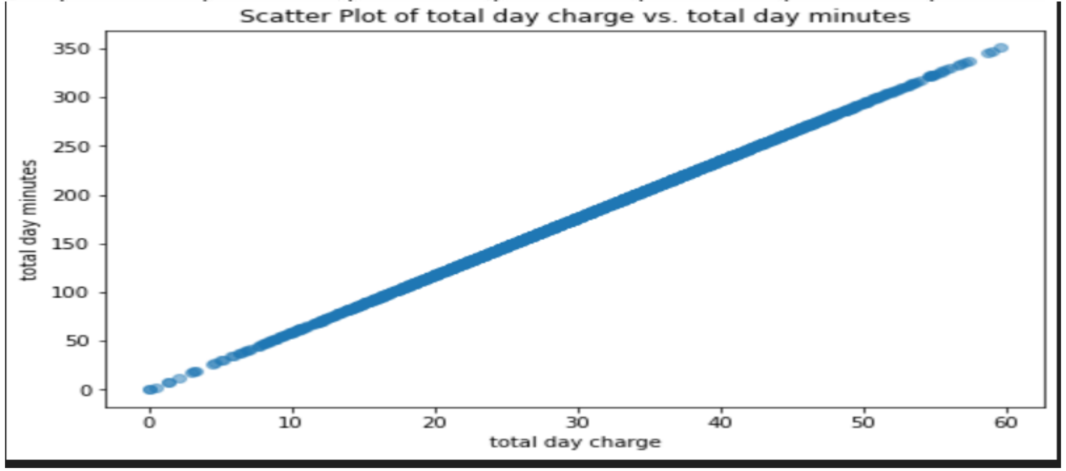
The scatter plots have revealed features that exhibit a perfect correlation with each other. Multicollinearity:

This perfect correlation indicates multicollinearity, a situation where independent variables are highly interrelated, often moving in perfect synchronization.

Modeling Challenges: Multicollinearity poses a significant challenge to the modeling process and can lead to less reliable statistical inferences.

In essence, the analysis has identified a strong linear relationship between certain features, suggesting that they may be redundant or providing similar information





Box plots to identify outliers and visualize the spread of data.

The boxplot visualizes the distribution of the data for a single feature.,in this case it’s the account length.

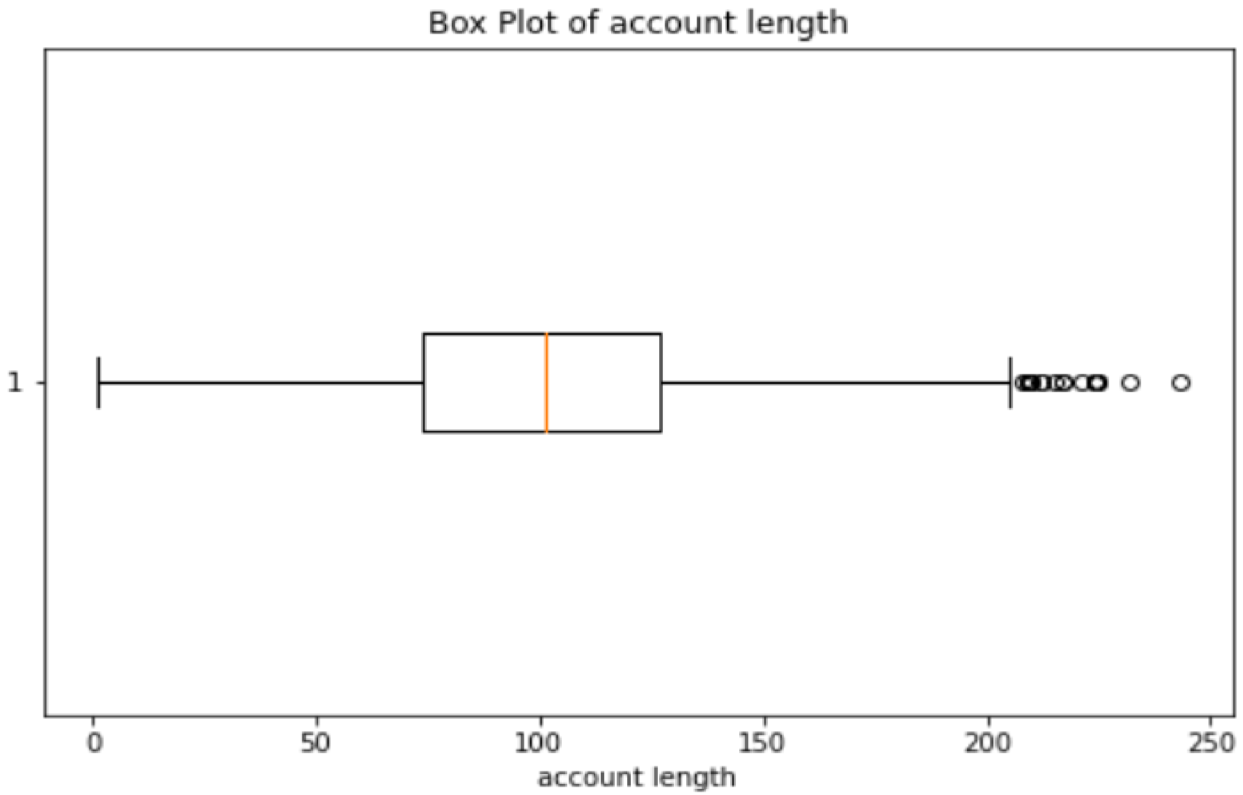
The box part of the plot represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3) of the data.

The line in the middle of the box represents the median (Q2) of the data.

The whiskers extend from the box towards the minimum and maximum values, up to 1.5 times the IQR.

Points beyond the whiskers are considered outliers.

By analyzing the boxplots, you can identify features that have outliers, skewness, or a large spread in their data



Data Preparation

I had to go through various steps to prepare the data for modelling.

Illustrated below are some of the steps I undertook.

1.Removing irrelevant columns

Given that the dataset does not contain any missing values and duplicates, the next step was to streamline the data by removing columns that were not essential for the analysis or modeling.

One of the columns I removed was phone number, which is a unique identifier for customers but is unlikely to be a predictive feature for the model

This process of column removal was crucial for several reasons such as

Dimensionality Reduction: By eliminating unnecessary columns, I was able to reduce the dimensionality of my data. This simplification led to faster model training and improved model interpretability.

Efficiency: Removing non-essential columns enhanced the efficiency of data processing and model training. It reduces computational demands, making our workflow more efficient.

Model Performance: Unnecessary columns can introduce noise into my analysis and modeling, potentially leading to less accurate results. Eliminating such columns can result in a cleaner and more focused dataset.

2.Feature Engineering

In our dataset, I identified that both the target variable and certain feature columns are categorical in nature. To effectively use this data in the modeling process, I decided to **encode** the data into numerical format

Data Preparation

Encoding:

Encoding categorical variables was a crucial step, as many machine learning algorithms require numerical inputs for model training. By performing appropriate encoding techniques, I was able to perform label encoding this enabled me to convert the categorical values into a numerical representation that the model could understand.

Handling imbalanced data:

From the dataset,the churn column had imbalanced data, I split the data roughly in half: 85.51% of the data fell into the 'False' or 0 category, and 14.49% of the data fall into the 'True' or 1 category. This suggested the imbalance in the dataset, with a larger percentage of non-churned clients.To address this issue, I employed the Synthetic Minority Over-sampling Technique, or SMOTE. Through the creation of duplicates,this increased the number of cases in the minority class.

Modelling

I developed a predictive model designed to anticipate whether a customer is on the verge of discontinuing their engagement with Syria. The primary objective was to curtail financial losses stemming from customers whohave a short-lived association with the entity.

Models Used were:

- Logistic Regression
- Decision Trees
- KNN Classifier Model

Evaluation

The logistic regression model showed a balanced performance with reasonably good accuracy, precision, recall, and F1 score. It captured positive cases effectively while maintaining precision. The ROC AUC score was also decent.

The decision tree model exhibited high accuracy, especially for the majority class. However, it had lower precision, recall, and F1 score for the minority class. This suggested that it may not perform as well on classifying the minority class.

The KNN classifier model achieved decent accuracy and performance for the majority class but struggled with the minority class just like the decision tree

Model of choice-Deciscion Trees

The decision tree model was evaluated with precision, recall, and F1 score for both Class 0 and Class 1.

From the accuracy results the model correctly predicted the class labels for the majority of instances in the test data. The precision metric is very important as it measures how accurate the model is at identifying the majority class, which is the customers who don't churn.

The downside of the model is that it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class, the same way it does the majority class.

But this is majorly attributed to the fact that our data is also highly imbalanced. But compared to the other two models, this one performs much better.

The model performs well in terms of precision and recall for both classes on the holdout test data, thus it can be deployed in a real-world scenario.

Recommendations and Future Investigations

- Customer Service Calls Investigation: Dig deeper to understand why some customers need to contact customer service frequently. This will help in finding ways to better assist them.

- High Churn States Analysis: Look into the states where many customers are leaving to identify any patterns or reasons for the high churn rates.

- Incentives for High Bill Customers: Find ways to encourage customers with high daily charges (over $55) to stay with Syria Tel. This might involve offering extra benefits and perks. Currently, all of these high-bill customers are leaving, which is a concern.

- Incentives for Customers who stay more than 6 months: Find ways to encourage customers to stay with the company even longer, e.g., giving them loyalty points, offers, etc., as this will help in creating a form of loyalty. CONCLUS

Conclusion

In conclusion, the decision tree model appears to be providing reasonably good results, especially for the majority class, which is typical for imbalanced datasets and is the best option out of the three models built above.

The decision tree model is the most suitable model for the given task, especially considering the imbalanced nature of the dataset. This means that the decision tree model is able to effectively predict the target variable for the majority class (i.e., the class with the most instances) while maintaining reasonable performance for the minority class.

This is a common scenario in real-world datasets where one class is significantly more prevalent than the other. In such cases, models that are specifically designed to handle imbalanced data, like the decision tree model, often outperform other models.

The End!!

Thank You

Name:Laura Kanda Mwichekha

Email Adddress: lmwichekha@gmail.com

Phone Number:0702873977

Any Questions?