SYRIATEL CUSTOMER CHURN PREDICTION: A SUPERVISED LEARNING APPROACH.

PROJECT BY: Kelvin Kipyegon Rotich.

INTRODUCTION.

Welcome to this project on predicting customer churn at SyriaTel, a telecommunications company based in the United States. In this project, we will be focused on supervised learning algorithms and choose the best one which will predict whether a customer will churn or not. This is based on a comprehensive dataset which contains several factors which may or may not influence a customer to stop using SyriaTel as its telecommunications service provider.

The main purpose of this project is to create a predictive model that accurately classifies whether a customer will churn or not. This project aims to provide insights to SyriaTel stakeholders and in turn they will make informed decisions which will mitigate customer churn effects in the company revenue.

In the following sections we will delve into the project's methodology, key findings and the models' performances in predicting customer churn.

BUSINESS UNDERSTANDING.

Telecommunications industry background in the USA.

The telecommunication industry in the United States has evolved significantly over the years, playing a pivotal role in shaping the nation's communication landscape. Ever since the first telegraph was created in the mid-19th century, the industry has been witness to a continuous series of technological advancements.

The advent of the telephone services in the late 1800s marked a transformative era. The mid-20th century saw the rise of microwave and satellite technologies which facilitated long-distance communication. The divestiture of AT&T in 1984 led to increased competition, paving the way for creation of new telecommunication companies in the United States.

The late 20th century and early 21st century saw the proliferation of mobile communications, with the emergence of wireless networks and the widespread adoption of smartphones. This period also saw the expansion of broadband internet services, enabling high-speed data transformations.

Regulatory changes, like the Telecommunications Act of 1996, aimed to foster competition and innovation by breaking down monopolistic structures. As a result, numerous players like SyriaTel entered the market, offering diverse services ranging from traditional landline telephone to broadband internet, cable television and mobile services.

Today, the U.S. Telecommunications industry continues to be dynamic, with ongoing advancements in 5G technology, fiber-optic networks, and the convergence of services. Major companies in the industry have played central roles, contributing to the nation's connectivity and driving innovation in communication technologies.

Challenges facing the telecommunication industry in the United States.

1. Cybersecurity issues: The telecommunications industry in the USA has an interconnected nature. This, coupled with the rise of IoT devices, exposes it to network vulnerabilities and data breaches.
2. Spectrum congestion: The increasing demand for wireless communication causes a strain to the available frequencies which leads to slower data speeds and degraded network performance.
3. Digital divide: The disparities in broadband access, particularly in rural areas, limit residents’ opportunities for education, employment and essential services.
4. Customer churn: Some, or maybe all, of the above mentioned challenges may lead to customers that stop using services of telecommunication companies. The fact that this is an unpredictable outcome is a big issue to the companies and it may cause huge losses.

Solutions to curb these challenges.

1. Robust cybersecurity measures which include encryption protocols, proactive threat detection mechanisms and regular system audits should be put in place and if they are present they should be improved.
2. Collaborative spectrum management and allocation by the telecommunication firms should be done. They should also employ technologies like dynamic spectrum sharing for more efficient use.
3. There should be significant investments in expanding broadband infrastructure to underserved areas which ensures equitable access for all citizens.
4. Telecommunication companies should listen to their customer needs and improve on their services by creating promotions or tariffs that can bring more customers on board and guarantee customer satisfaction.

It can be seen that even though the American telecommunication landscape is more advanced than most of the countries globally, there are issues that plague them like their global counterparts. However, they have an advantage since they have the means that may try to solve their issues faster.

Project Problem Statement.

SyriaTel is grappling with the issue of customer churn. Despite offering a range of services, the company is experiencing a significant increase in customer attrition, leading to a decline in overall revenue and customer satisfaction and all because of customer churn. The company seeks to proactively predict customer churn, allowing for targeted retention strategies and ultimately reducing customer attrition rates to enhance overall long-term business sustainability in the dynamic telecommunications industry.

Project Objectives.

1. To investigate each feature and check for patterns. This will help in identifying features to be used in creating the models. It will also help in identifying distributions of numerical features and count the unique values in each feature for categorical features.
2. To investigate the relationship between the feature variables and the target variable. This will try to identify the patterns that may lead to customer churn. This will also help in filtering some of the features to be used in modelling.
3. To check the relationship between numerical features. This will determine the models to be used for this project.
4. To create a precise model that will be used to predict customer churn depending on a range of features.

DATA UNDERSTANDING.

The data used for the project was obtained from the ***customer_churn.csv*** file that was downloaded from kaggle.com. The dataset contains information about some SyriaTel customers, their information and whether the customer churned or not. There are 21 columns in the dataset with 3333 entries.

Additional Column information.

- ***state***(object): The state where the customer comes from.
- ***account length***(int): The number of months the customer has stayed with SyriaTel.
- ***area code***(int): The telephone area code the customer lives in.
- ***phone number***(object): The customer's phone number.
- ***international plan***(object): Whether the customer has an international plan or not.
- ***voice mail plan***(object): Whether the customer has a voicemail plan or not.
- ***number vmail messages(***int): The number of voicemail messages the customer has.
- ***total day minutes***(float): The number of minutes the customer spends on calls during the day.
- ***total day calls***(int): The number of calls the customer makes during the day.
- ***total day charge***(float): The amount the customer is charged from calls during the day.
- ***total eve minutes***(float): The number of minutes the customer spends on calls in the evening.
- ***total eve calls***(int): The number of calls the customer makes in the evening.
- ***total eve charge***(float): The amount the customer is charged from calls in the evening.
- ***total night minutes***(float): The number of minutes the customer spends on calls at night.
- ***total night calls***(int): The number of calls the customer makes at night.
- ***total night charge***(float): The amount the customer is charged from calls at night.
- ***total intl minutes***(float): The number of minutes the customer spends on international calls.
- ***total intl calls***(int): The number of international calls the customer makes.
- ***total intl charge***(float): The amount the customer is charged from international calls.
- ***customer service calls***(int): The number of customer service calls the customer has ever made.
- ***churn***(bool): Whether the customer has churned or not.

DATA PREPARATION.

For this stage, three steps were taken. The first one involved checking for missing, duplicated and placeholder values in the dataset. However, none of these values were found.

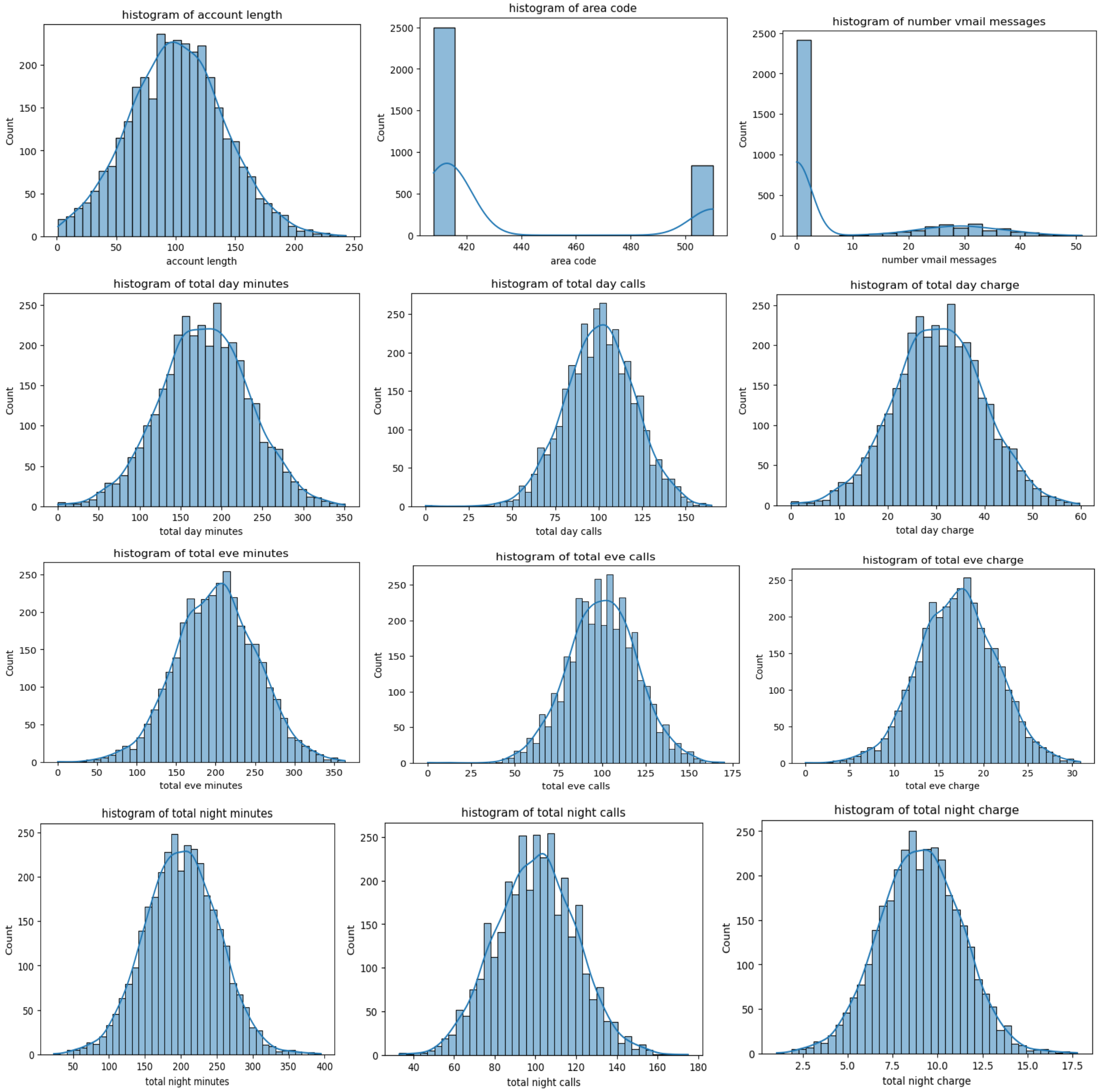
The next step was to find outliers in the dataset and deal with them appropriately. Even though outliers were present, the best option was to leave them in the dataset. This was because they were genuine events that took place and they could affect customer churn.

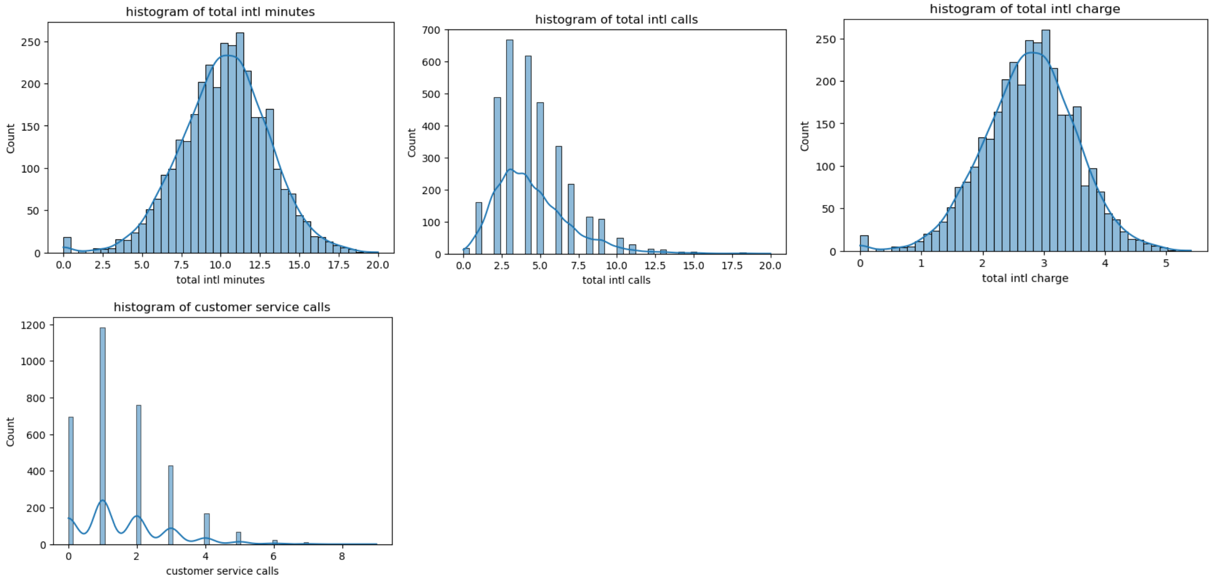
The final step was to look for columns that were not the right datatype and change them. All the columns seemed to have the right datatype.

EXPLORATORY DATA ANALYSIS.

Univariate Analysis.

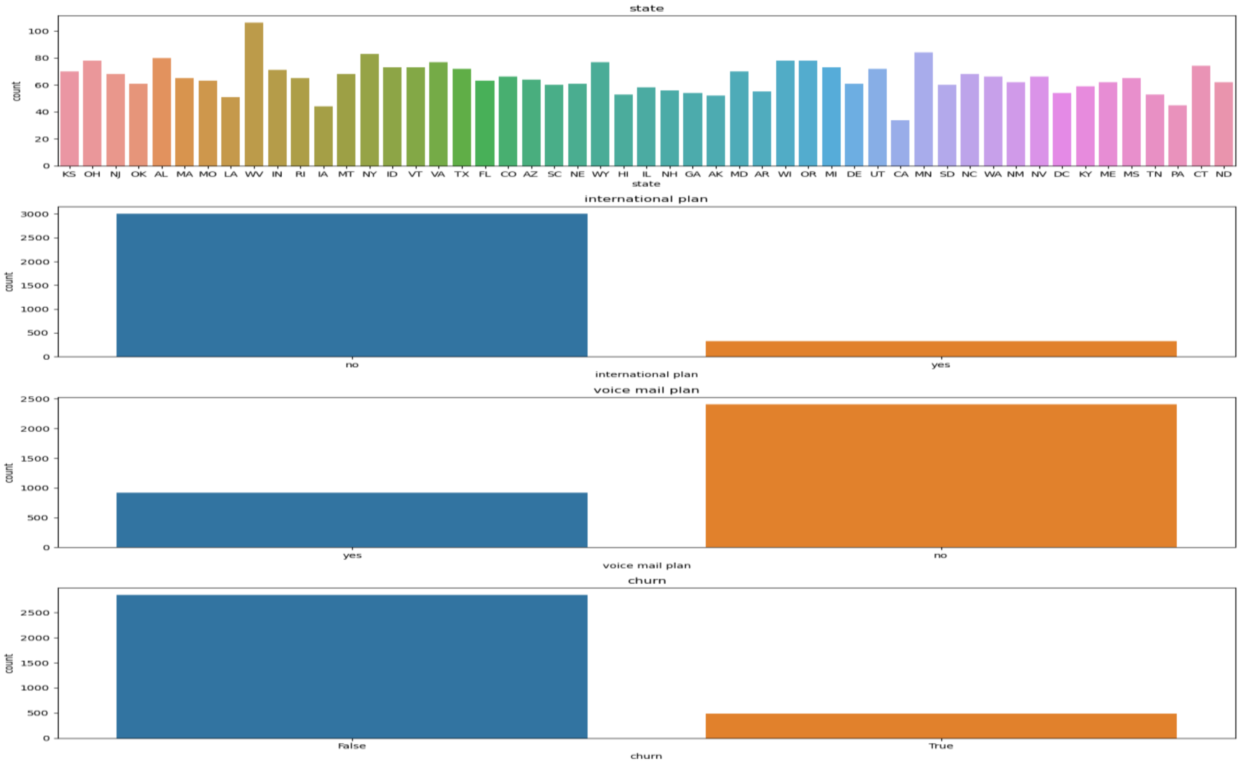
The analysis of numerical columns was the first step taken in the univariate analysis stage. Here, histogram plots were created to show the distributions of these features.





From the histograms, it is evident that most of the features seem to be normally distributed. Those that don't follow the same distribution underwent normalization in the modelling process.

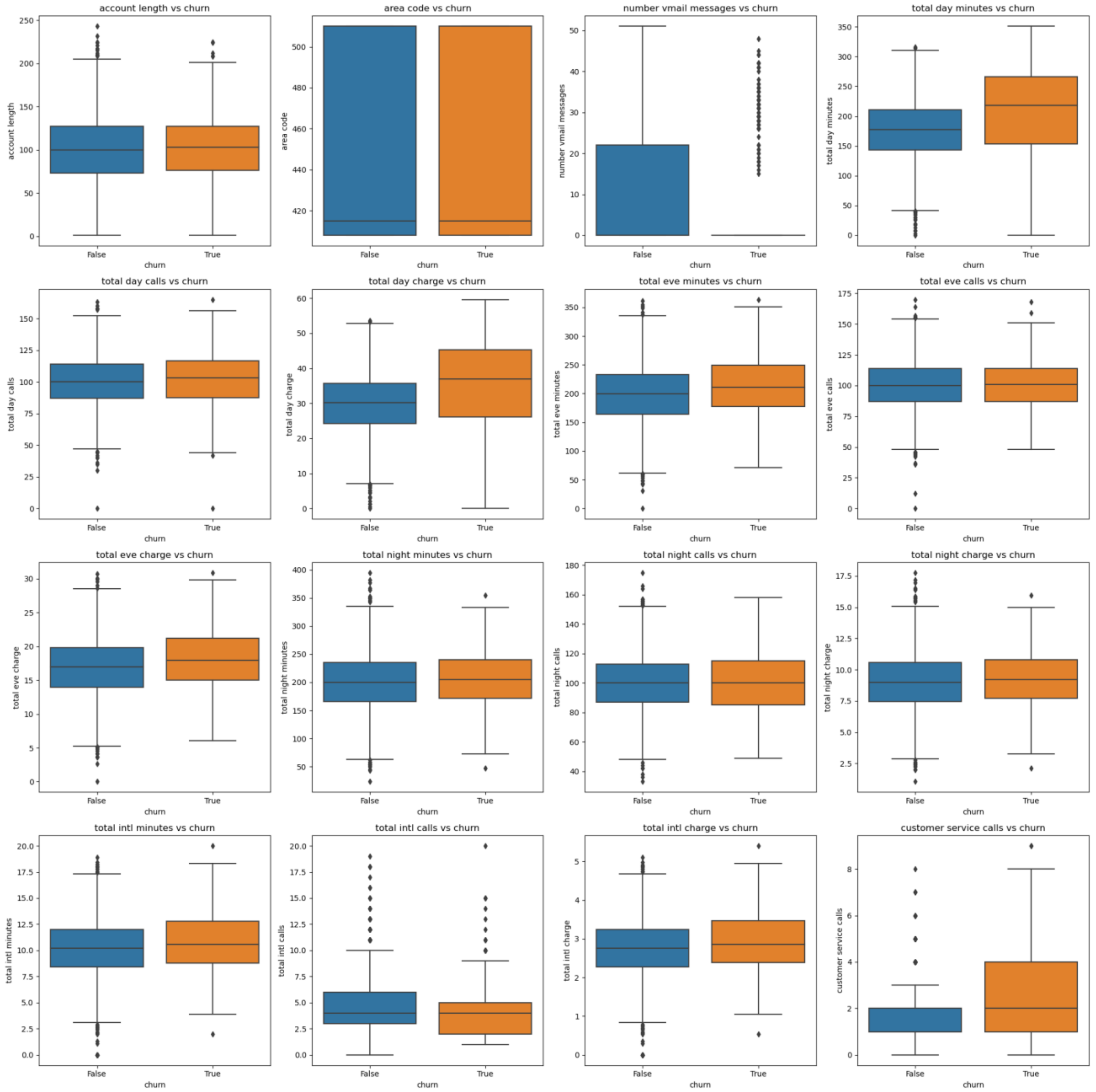
The next step was to analyze the categorical columns. A count on the number of unique values was done and it was noticeable that the ***phone number*** column had the number of unique values same as the number of rows in the dataset which meant that it would not bring any effect on the models we created and hence we opted to drop it in the data preprocessing stage. Count plots were then created to show how the other remaining columns were distributed.



Based on the visual, it is noticeable that all of the columns save for ***states*** have two classes. However, we opted not to drop the ***states*** column since it could have a hidden pattern on customer churn.

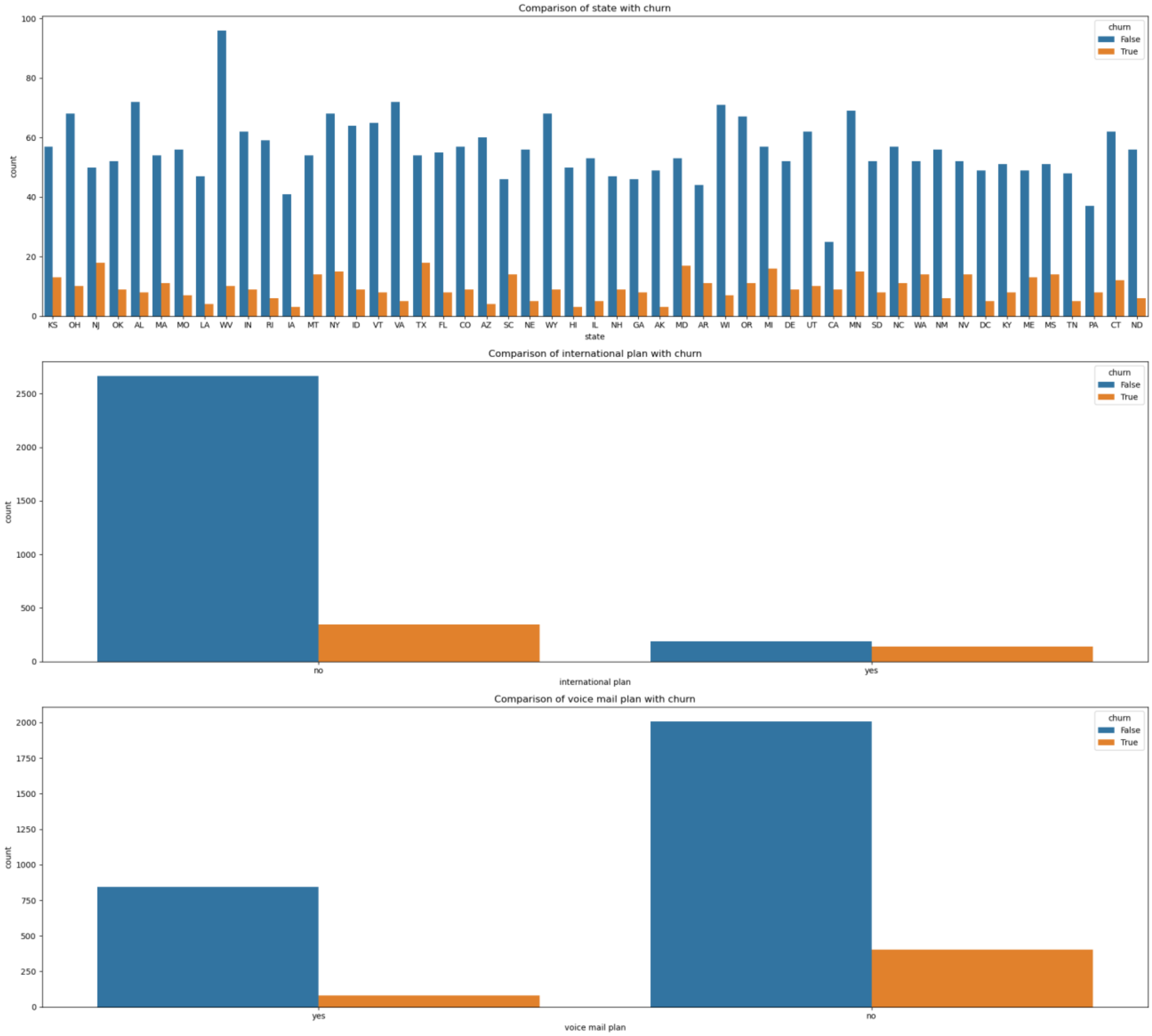
Bivariate Analysis.

Here, we were comparing the features to the ***churn*** column, which is our target variable. We began by comparing it with the numerical columns by plotting boxplots.



Based on the visual, we noticed that there was no visible pattern shown by the ***area code*** column. That means it won’t be used in the modelling process. We could also see that most of the customers who churned made a lot of customer service calls to the company and spent a lot of minutes making calls during the day which made them incur charges. With the other features they seem to have somewhat similar characteristics.

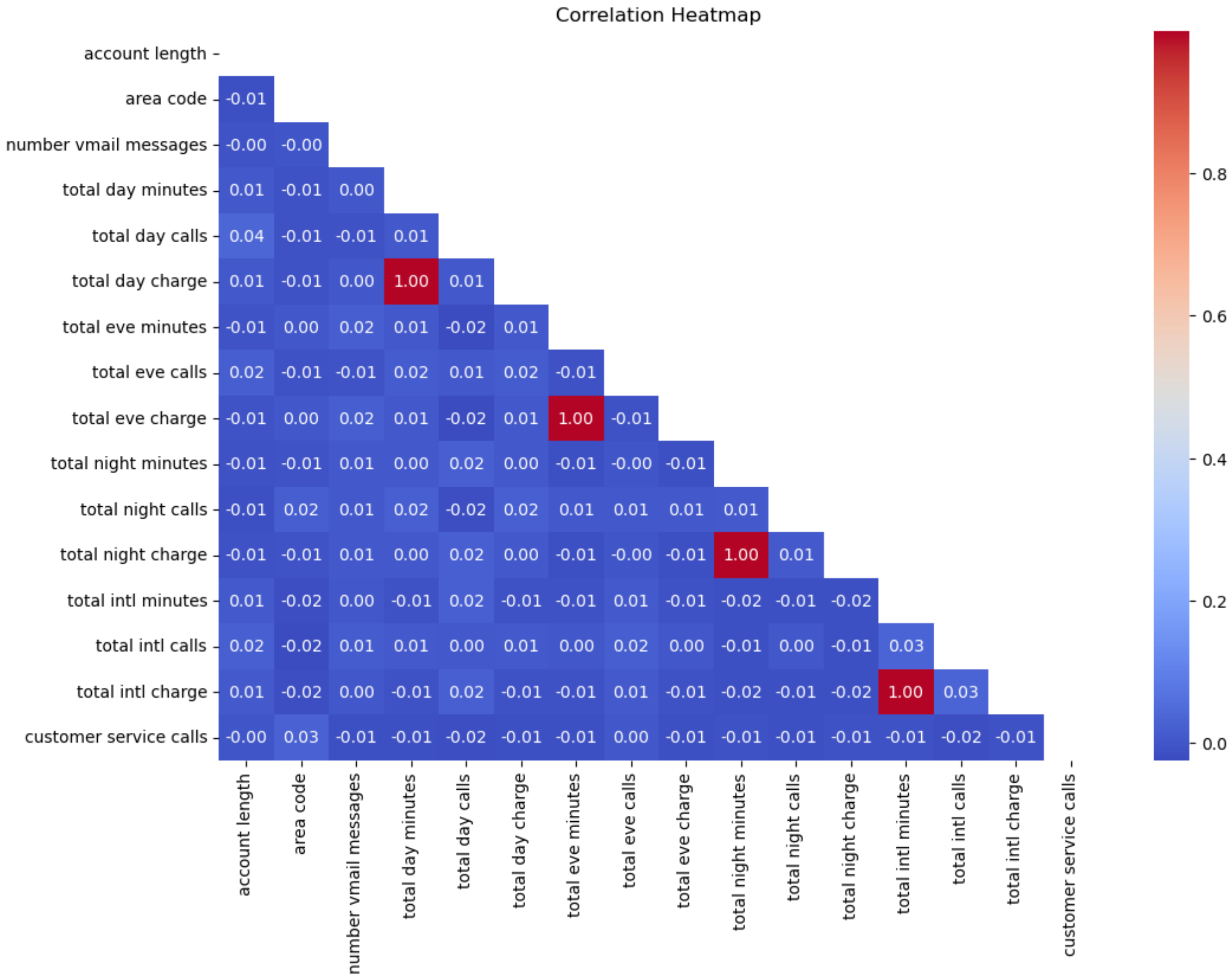
We then focused on comparing the target column with the categorical columns. We did this by creating bar plots.



Since it was hard to tell the pattern between those who’ve churned and those who hadn’t in the ***states*** column, dropping it was needed. Other than that, it was noted that most of the customers who had churned had subscribed for neither an international nor a voicemail plan.

Multivariate Analysis.

This was the final step taken in the Exploratory Data Analysis Stage. Here, a heat map was plotted to show the correlation between the numerical columns in the dataset.



Based on the plot, it is seen that most of them have either weak or no correlation with each other. However, total minutes spent had a perfectly positive correlation with the total charges incurred irrespective of the time. Even the minutes spent on international calls had the same characteristic when compared to the total international charges incurred. This meant that the two were dependent of each other and it also meant that the models that need features to be independent of each other such as logistic regression and any Naive Bayes algorithm weren’t used in the modelling stage.

MODELLING.

Data Preprocessing.

Before we began creating the models, we went through data preprocessing. Here, we dropped the ***state, area code*** and ***phone number*** columns. We also conducted feature engineering on the ***international plan*** and ***voice mail plan*** columns where we replaced the *yes* and *no* classes to *1* and *0* respectively. We also split the dataset into training and testing datasets which would help us train our models and evaluate our models using the metrics achieved. Since the target variable had class imbalance, where one class was way more represented than the other one, we had to undertake the SMOTE technique on the training dataset to curb this issue. We then split the dataset to training and validation datasets. These datasets were used in creating the models and the test dataset from the previous split was used in model evaluation.

Model Selection.

The datasets created from the second split were used in this stage. Here, five algorithms were used in the creating our models. For each algorithm, 3 models were created and the model with the best metrics were chosen. Below is a table showing the best metrics for each algorithm.

Algorithm	**Precision Score**	**Recall Score**	**Accuracy Score**	**F1 Score**	**Precision-Recall AUC**	**ROC AUC**
Decision Trees	0.83	0.90	0.86	0.87	0.89	0.86
K-Nearest Neighbors	0.89	0.96	0.92	0.92	0.94	0.92
Quadratic Discriminant Analysis	0.81	0.83	0.82	0.82	0.86	0.82
Random Forests	0.94	0.94	0.94	0.94	0.96	0.94
XGBoost	0.93	0.95	0.94	0.94	0.95	0.94

These models were used in the final model evaluation.

Model Evaluation.

We used the testing datasets from the first split to determine the model we would use for our prediction. The metric we used to determine the best model was the recall score. This is because the impact of our model predicting a customer isn’t going to churn and then they end up doing the opposite will be very bad to SyriaTel since they would incur losses from this issue. These were the results gotten from the evaluation.

Algorithm	**Final Evaluation Recall Score**
Decision Trees	0.77
K- Nearest Neighbors	0.60
Quadratic Discriminant Analysis	0.72
Random Forests	0.77
XGBoost	0.81

Based from this evaluation, it can be seen that the XGBoost model was the best for predicting customer churn at SyriaTel.

Limitations.

1. The classes were imbalanced. in that the *False* class was way more than the *True* class. This is why we had to use the SMOTE technique to mitigate the issue.
2. There were outliers present in the dataset. The issue was that they were genuine events and they could not be dropped from the dataset.

CONCLUSION.

In conclusion, the implementation of this customer churn prediction model stands as a pivotal initiative for SyriaTel. By harnessing advanced analytics and machine learning, we have fortified our ability to anticipate and mitigate customer churn, a crucial factor in sustaining and growing SyriaTel's customer base. This model not only give insights into potential customer churn indicators but also empowers the company to proactively engage with at-risk customers, offering personalized incentives and services to enhance their satisfaction. With this data-driven project, the company positions itself not just to retain customers but to cultivate lasting relationships and drive the long-term success of the business in the dynamic landscape of the telecommunications industry.

RECOMMENDATIONS.

These are some of the recommendations offered:

- SyriaTel should try and listen to the issues raised in the customer service calls and make improvements on them.
- The company should create more incentives to enhance customer satisfaction and attract new customers.
- The company should reach out to its customer base and do surveys in order to get insights on their customer needs.
- They should create more promotions and rewards to increase customer participation with their services.

NEXT STEPS.

Some of the next steps that can be taken after this analysis are:

- Model deployment.
- Customer surveys by the company.
- Customer churn comparison project with other telecommunication companies.