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.

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.

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.

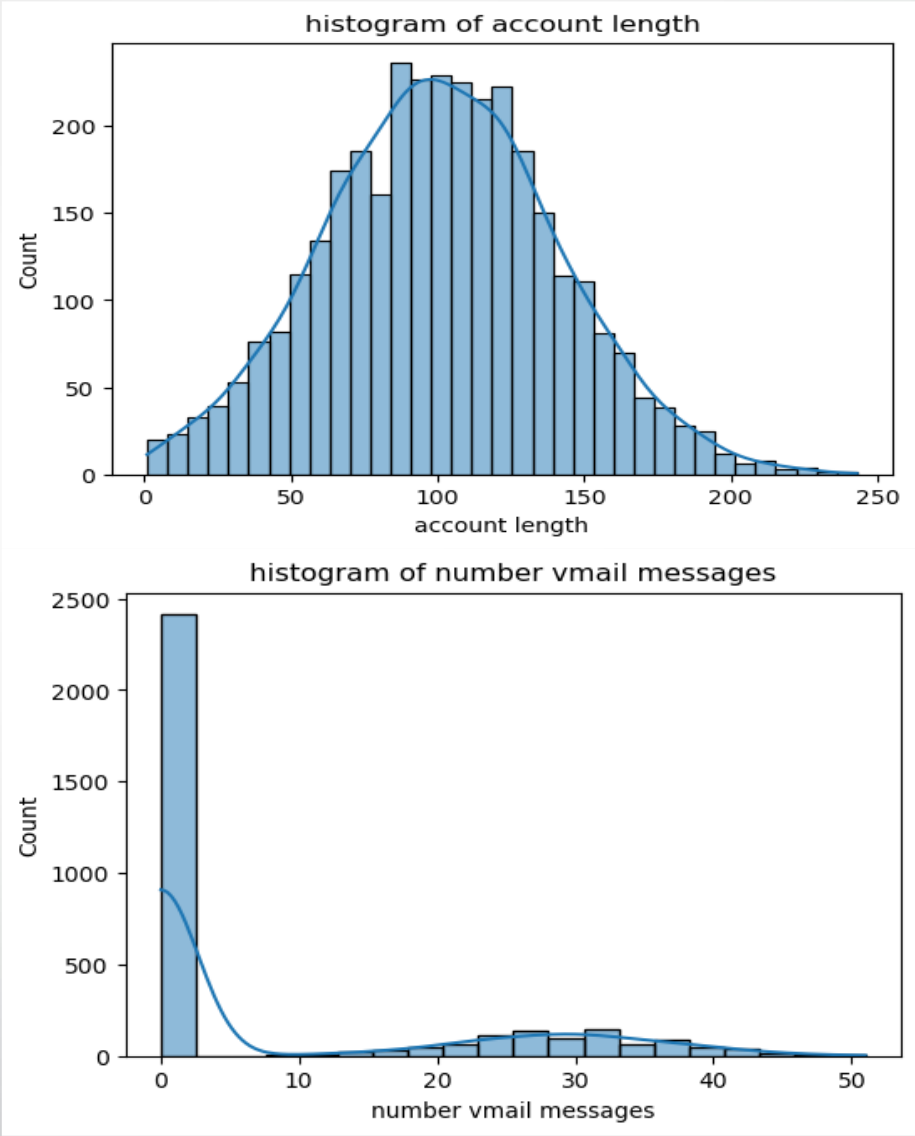
DATA PREPARATION

- The first one involved checking for missing, duplicated and placeholder values in the dataset. However, none of these values were found .
- The other step was to look for columns that were not the right datatype and change them. All the columns seemed to have the right datatype.
- The other 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.

EXPLORATORY DATA ANALYSIS

Most of the features in the dataset, for instance the first plot, seem to be normally distributed.

Those that don't follow the same distribution , as the second plot shown, underwent normalization in the modelling process.



- We began the next analysis with 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.
- This 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.

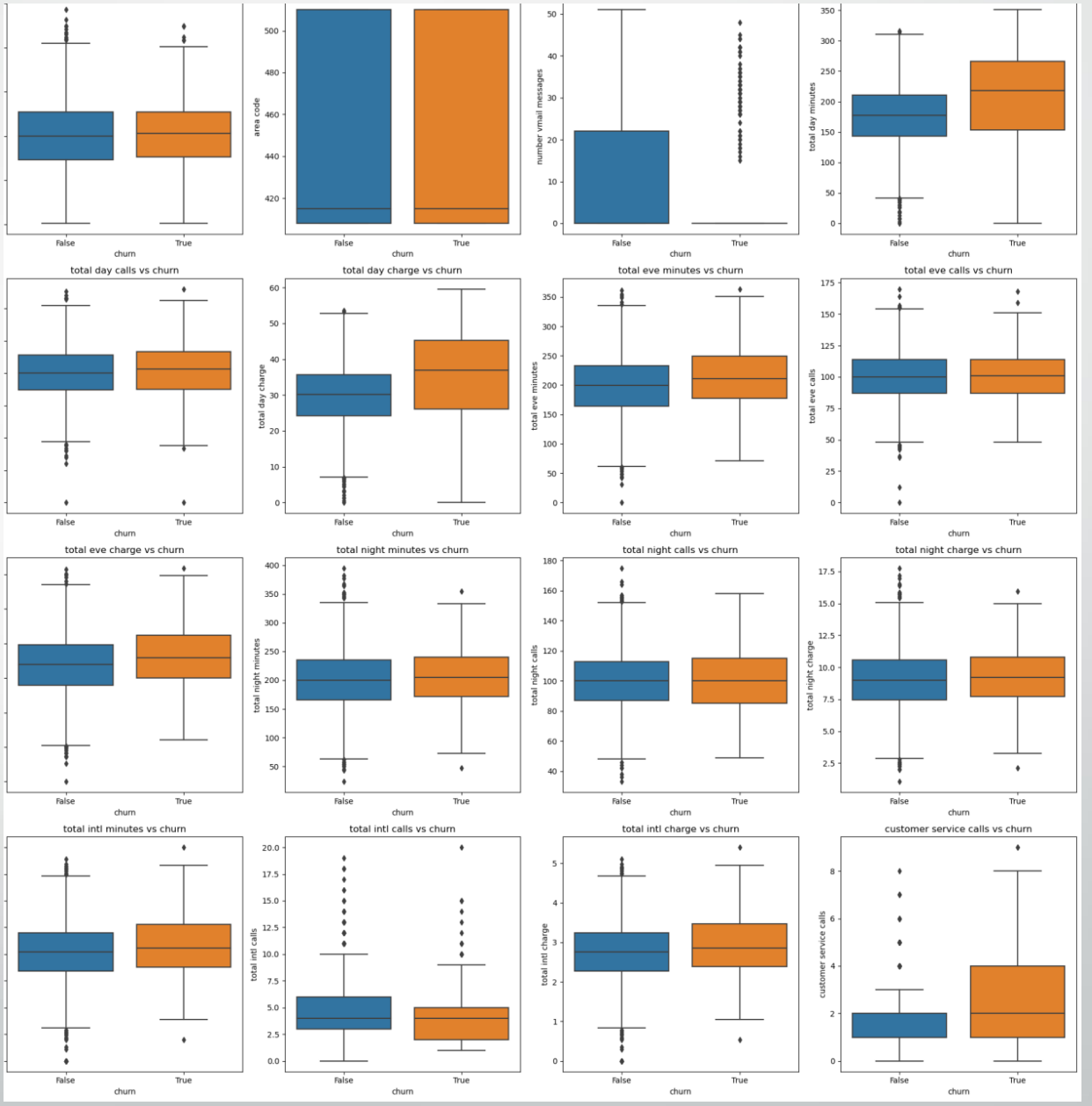
We then compared the numerical features to the ***churn*** column, which is our target variable by plotting boxplots.

Based on the visual, we noticed that there was no visible pattern shown by the ***area code*** column.

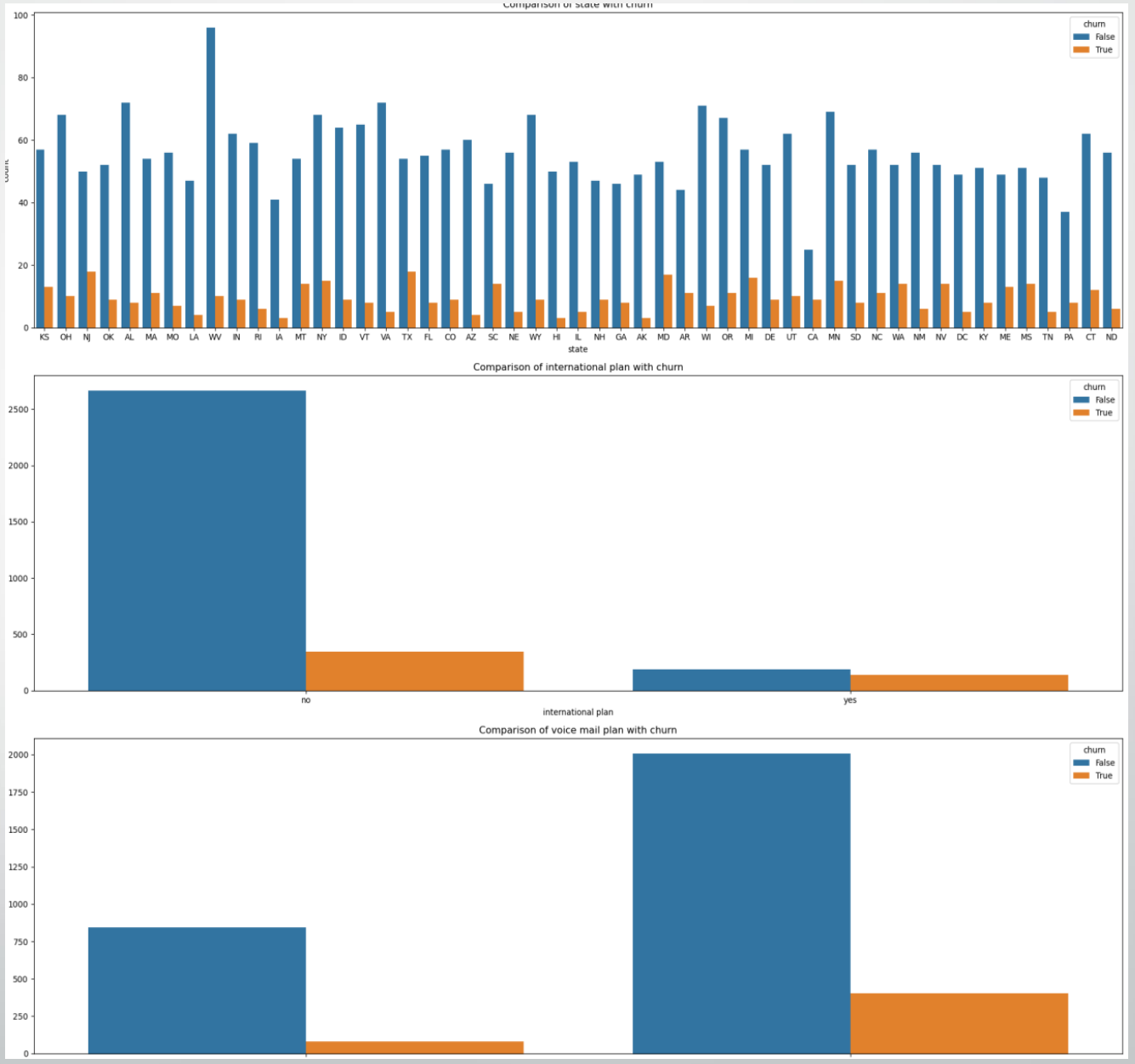
That means it won't be used in the modelling process.

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.

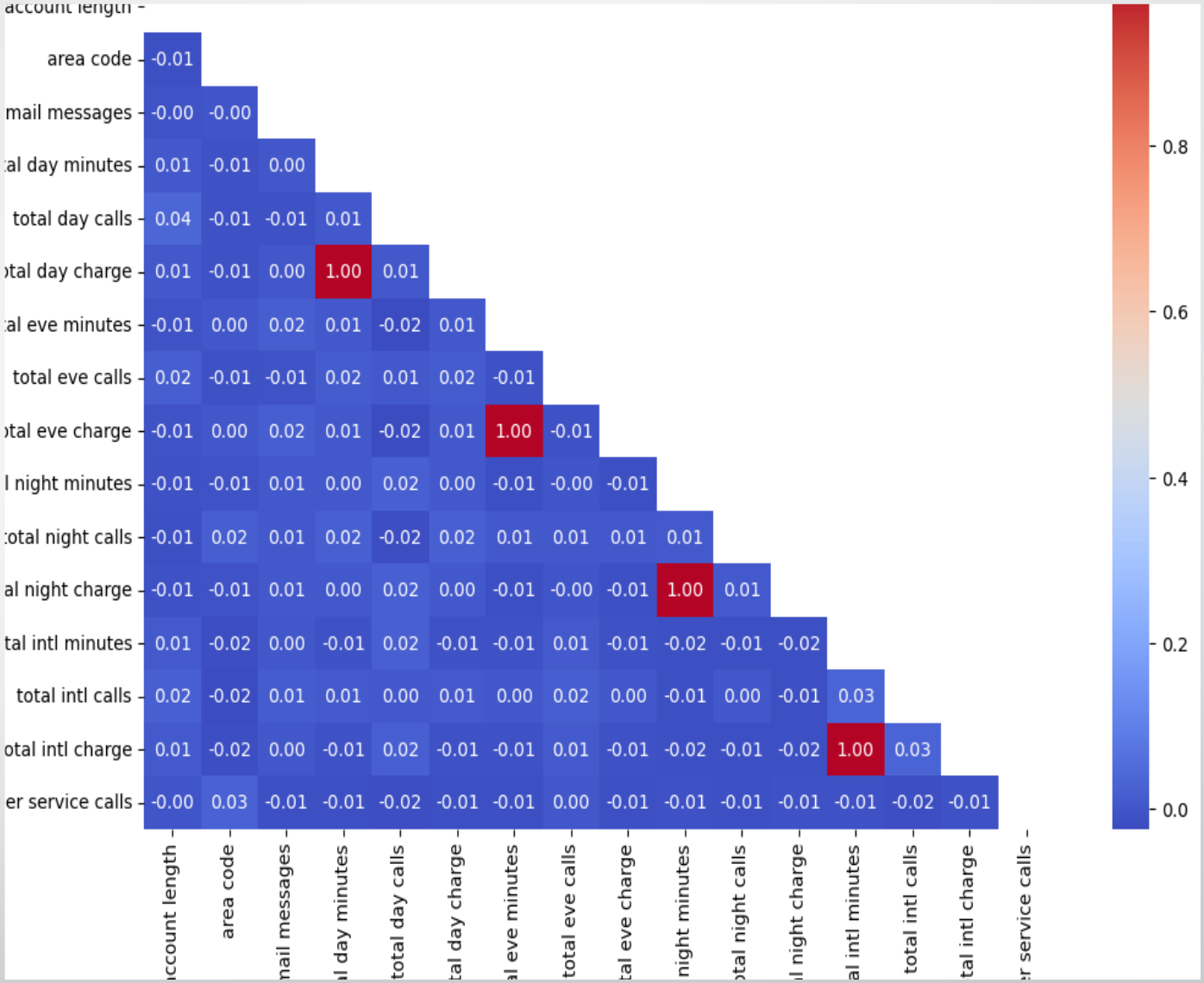
With the other features they seem to have somewhat similar characteristics.



- We then compared 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.



- 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.



MODEL 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.
- The data was also normalized to make the features follow a normal distribution.

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.
- The next slide contains a table showing the best metrics for each algorithm.
- These models were used in the final model evaluation.

Algorithm	Precision Score	Recall Score	Accuracy Score	F1Score	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

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.
- The next slide contains the results gotten from the evaluation.
- Based from this evaluation, it can be seen that the XGBoost model was the best for predicting customer churn at SyriaTel.

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

LIMITATIONS

- 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.
- 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

- 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 STEP

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

THANK YOU!

QUESTIONS?