ThomasOgada /
SyriaTel_Churn_Project

Code Issues Pull requests Actions Projects Wiki Security Insights

main SyriaTel_Churn_Project / Go to file t Add file ...

ThomasOgada Create README.md 16 minutes ago

Name	Name	Last commit date
README.md	Create README.md	16 minutes ago

README.md

SyriaTel Churn Project.

Project Overview.

This project is about SyriaTel, a telecommunication company based in Syria in the Middle has a telecommunications sector experiencing rapid growth in mobile and internet penetration. The company plays a vital role in connecting people and businesses. However, increasing competition and evolving customer preferences pose challenges for customer retention. Understanding and addressing the drivers of churn are crucial for SyriaTel to sustain business success and enhance customer satisfaction. Customer churn is a phenomenon where customers cease doing business with a company, is a critical concern for telecommunications companies like SyriaTel. Retaining customers is essential for maintaining revenue and growth in this competitive industry. Identifying factors contributing to churn, such as service dissatisfaction or competitive offers, SyriaTel can take targeted actions to mitigate churn and improve customer retention.

Problem Statement

SyriaTel, a telecommunications company, faces the challenge of customer churn, where customers discontinue their services. This attrition impacts revenue and profitability. The business seeks to proactively identify customers at risk of churning and implement effective retention strategies to mitigate revenue loss and maintain customer loyalty.

Specifically, the project aims to address the following questions:

1.What are the primary factors driving customer churn for SyriaTel?

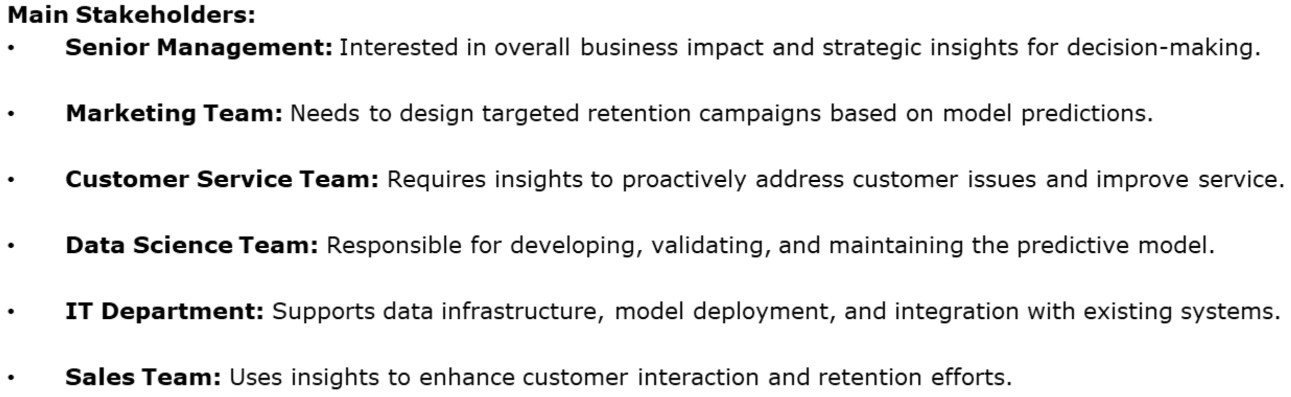
2.Which machine learning modelling technique to apply in accurately predicting Churn so as to take proactive measures?

3.What actionable insights can SyriaTel derive from the predictive model to improve customer retention efforts?

4.What strategies can SyriaTel put in place to reduce churn rate?.

Main Stakeholders

Main Stakeholders: Senior Management: Interested in overall business impact and strategic insights for decision-making. Marketing Team: Needs to design targeted retention campaigns based on model predictions. Customer Service Team: Requires insights to proactively address customer issues and improve service. Data Science Team: Responsible for developing, validating, and maintaining the predictive model. IT Department: Supports data infrastructure, model deployment, and integration with existing systems. Sales Team: Uses insights to enhance customer interaction and retention efforts.



Main Stakeholders:

- **Senior Management:** Interested in overall business impact and strategic insights for decision-making.
- **Marketing Team:** Needs to design targeted retention campaigns based on model predictions.
- **Customer Service Team:** Requires insights to proactively address customer issues and improve service.
- **Data Science Team:** Responsible for developing, validating, and maintaining the predictive model.
- **IT Department:** Supports data infrastructure, model deployment, and integration with existing systems.
- **Sales Team:** Uses insights to enhance customer interaction and retention efforts.

Methodology

Data Collection:

Gather and preprocess customer data, including numerical, categorical, and string columns. For purposes of this project, the 'SyriaTel_df.csv' dataset was used. The dataset had 3,333 rows and 21 columns The columns provided had numerical, categorical and string data types.

Data Preparation.

Outlier identification and handling. One hot and Cording Categorical Columns. After Data preparation, a dataframe of 3,333 rows and 67 columns were adopted for further analysis. Main columns colums considered for the analysis included Numerical Colums :'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’ Categorical Columns: “State”, “International Plan, “Voice Mail Plan”,”Churn”

Data Analysis:

Perform exploratory data analysis (EDA) to identify key churn indicators. Assesing descriptive statistics of the dataset. visualizing outputs in Barcharts, Histograms, scatterplots and Heatmap to understand the distribution and correlation of various features. Computing the normality and spread of the numerical variables. Inferential Statistics Hypothesis testing using ANOVA(Analysis of Variance)

Feature Engineering:

Encode categorical variables including State', 'International Plan', 'Voice Mail Plan', 'Churn'. normalizing all the features using the StandardScaler MinMax.

Model Building:

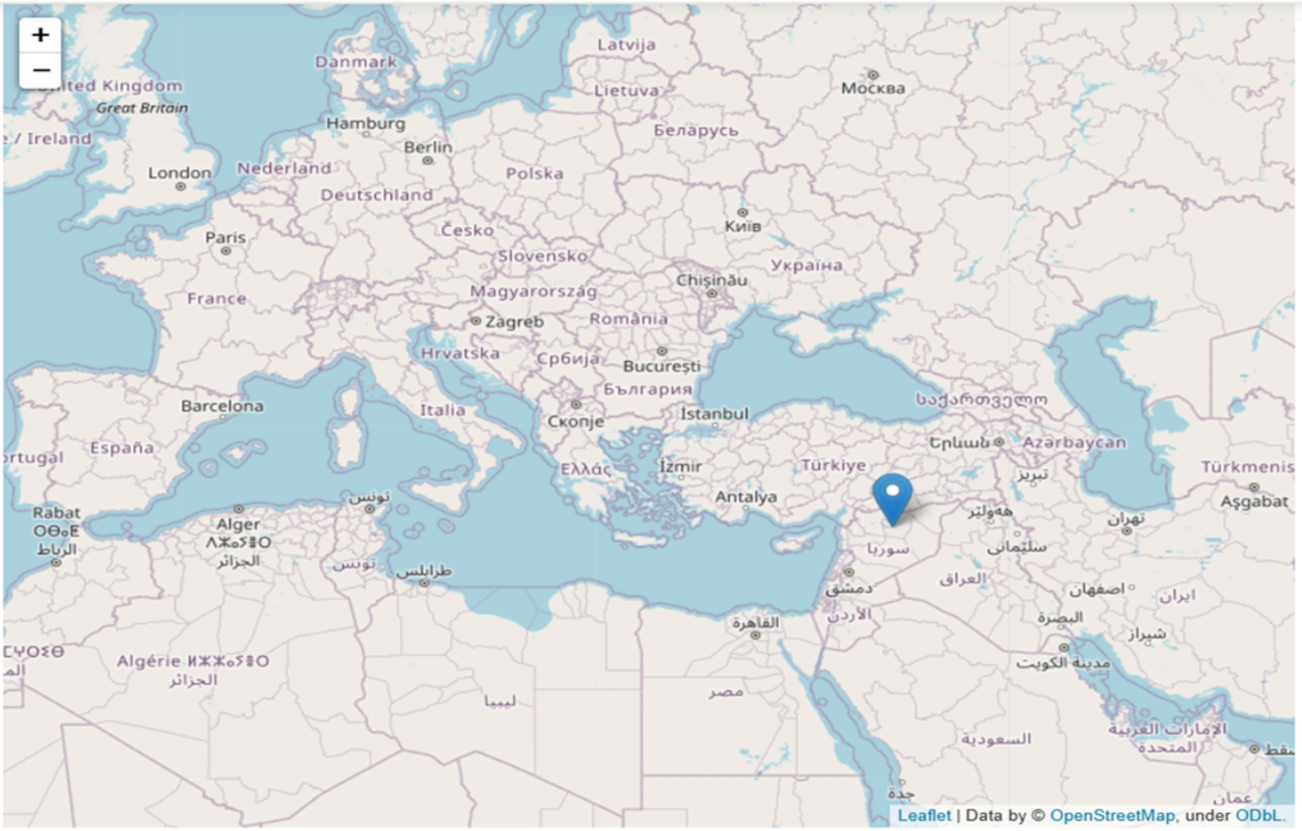
Feature Engineering One Hot Encoding(Dealing with categorical Data) Normalization/Standadization Split- Train Test Model Evaluation Use machine learning models such as Logistic Regression, Decision Trees, KNN, and XGBoost.

Model Evaluation:

Validate models using k-fold cross-validation and select the best-performing model. Deployment: Implement the model to predict churn and support retention strategies.

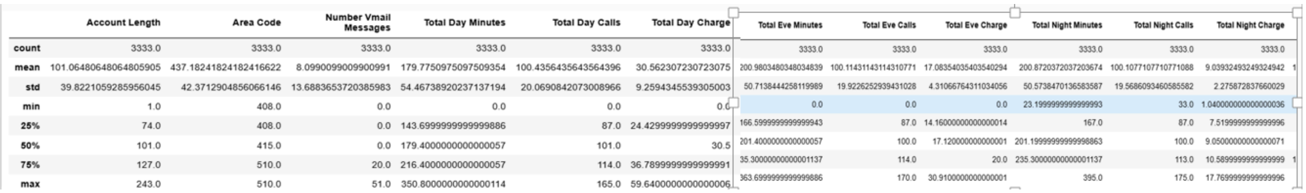
Results.

Geographical Location of the Syria.



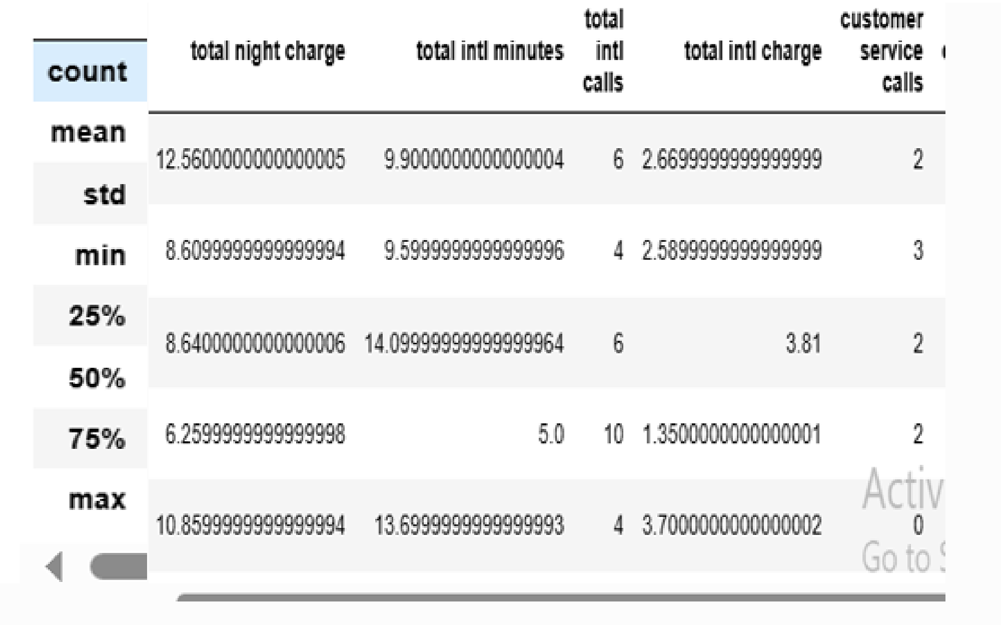
Leaflet | Data by © OpenStreetMap, under ODbL.

Descriptive Statistics



	Account Length	Area Code	Number Vmail Messages	Total Day Minutes	Total Day Calls	Total Day Charge
count	3333.0	3333.0	3333.0	3333.0	3333.0	3333.0
mean	101.06480648064805905	437.18241824182416622	8.0990099009900991	179.7750975097509354	100.4356435643564396	30.562307230723075
std	39.8221059285956045	42.3712904856066146	13.6883653720385983	54.46738920237137194	20.0690842073008966	9.2594345539305003
min	1.0	408.0	0.0	0.0	0.0	0.0
25%	74.0	408.0	0.0	143.6999999999999886	87.0	24.4299999999999997
50%	101.0	415.0	0.0	179.4000000000000057	101.0	30.5
75%	127.0	510.0	20.0	216.4000000000000057	114.0	36.7899999999999991
max	243.0	510.0	51.0	350.8000000000000114	165.0	59.6400000000000006

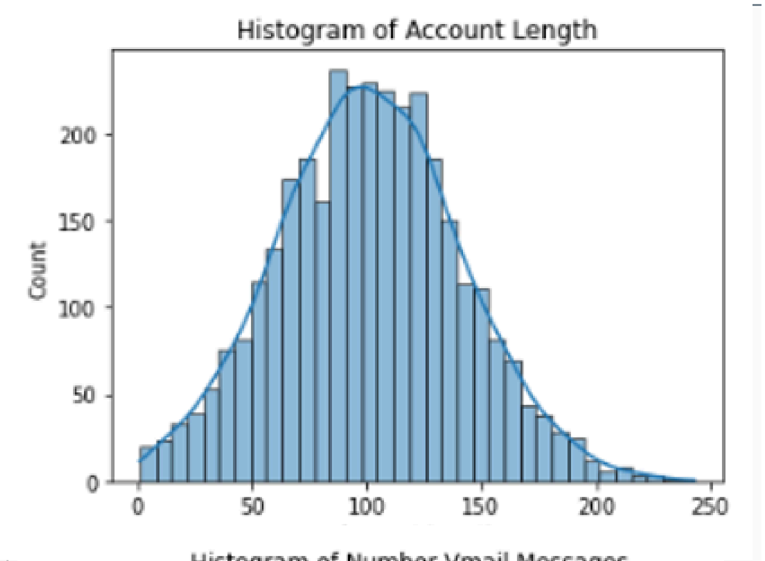
	Total Eve Minutes	Total Eve Calls	Total Eve Charge	Total Night Minutes	Total Night Calls	Total Night Charge
count	3333.0	3333.0	3333.0	3333.0	3333.0	3333.0
mean	200.9803480348034839	100.11431143114310771	17.08354035403540294	200.8720372037203674	100.10771077107710 88	9.03932493249324942
std	50.7138444258119989	19.9226252939431028	4.31066764311034056	50.5738470136583587	19.5686093460585582	2.275872837660029
min	0.0	0.0	0.0	23.1999999999999993	33.0	1.0400000000000000036
25%	166.5999999999999943	87.0	14.1600000000000014	167.0	87.0	7.5199999999999996
50%	201.4000000000000057	100.0	17.120000000000001	201.1999999999998863	100.0	9.0500000000000071
75%	35.3000000000001137	114.0	20.0	235.3000000000001137	113.0	10.5899999999999999
max	363.6999999999999886	170.0	30.910000000000001	395.0	175.0	17.7699999999999996

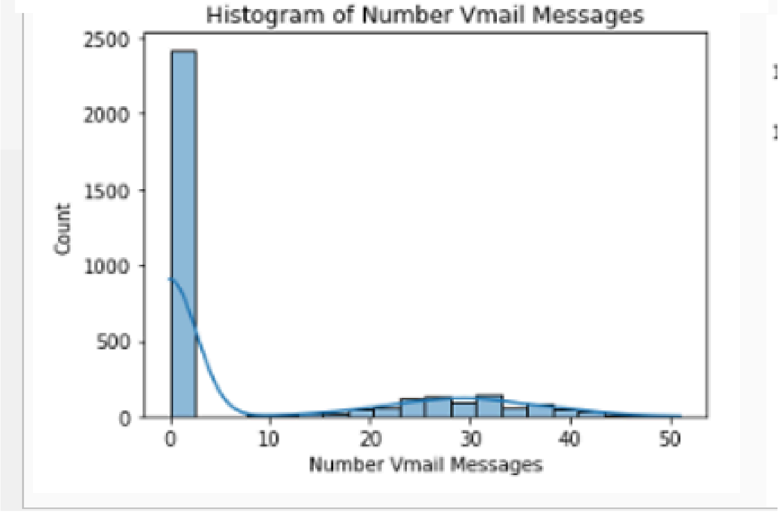


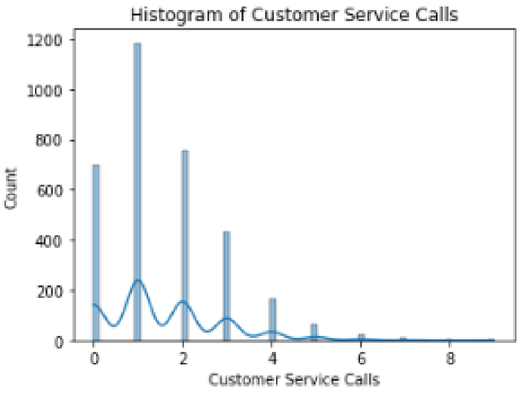
count	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls
mean std	12.5600000000000005	9.9000000000000004	6	2.6699999999999999	2
min	8.6099999999999994	9.5999999999999996	4	2.5899999999999999	3
25% 50%	8.6400000000000006	14.0999999999999964	6	3.81	2
75%	6.2599999999999998	5.0	10	1.3500000000000001	2
max	10.8599999999999994	13.6999999999999993	4	3.7000000000000002	0

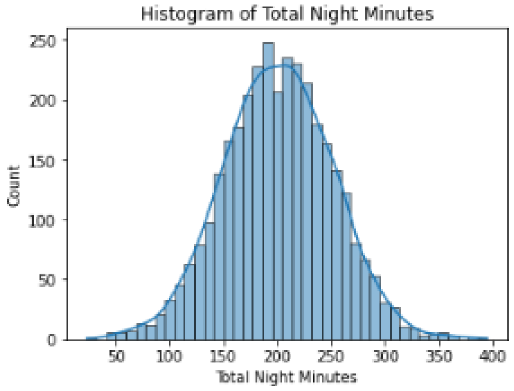
Total number of customer is 3,333. Mean account length 101.1. Max account length 243. Mean Total Day Calls is approximately 100 calls. Max Total Night Calls is 175. Std for Total Day Change is 9.3. Max Customer Service Calls is 9.

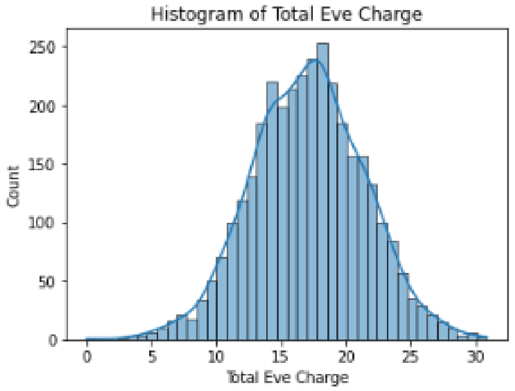
Univariant Analysis



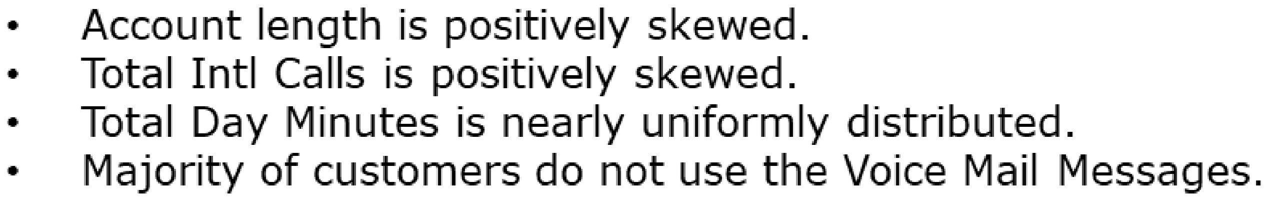








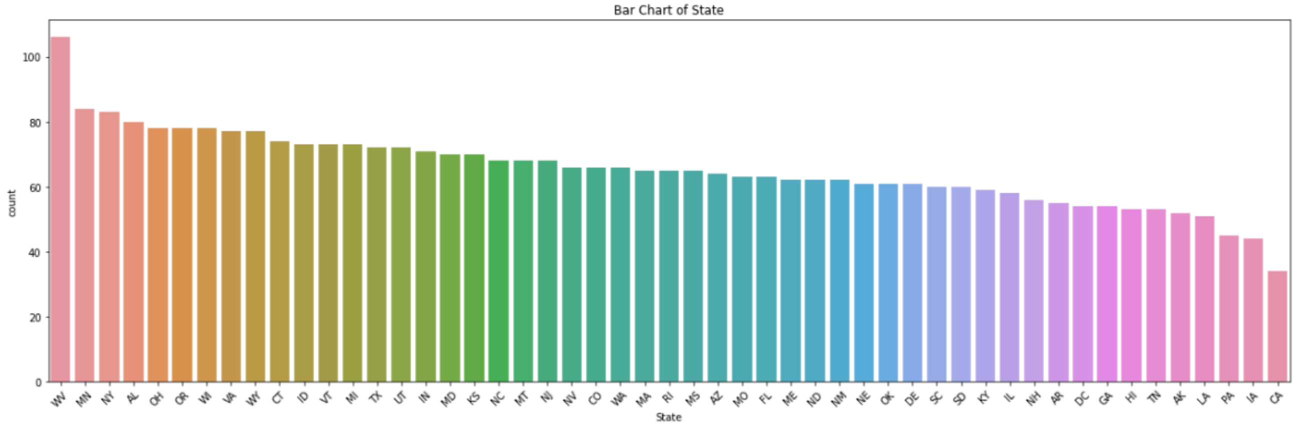
Account length is positively skewed. Total Intl Calls is positively skewed. Total Day Minutes is nearly uniformly distributed. Majority of customers do not use the Voice Mail Messages.

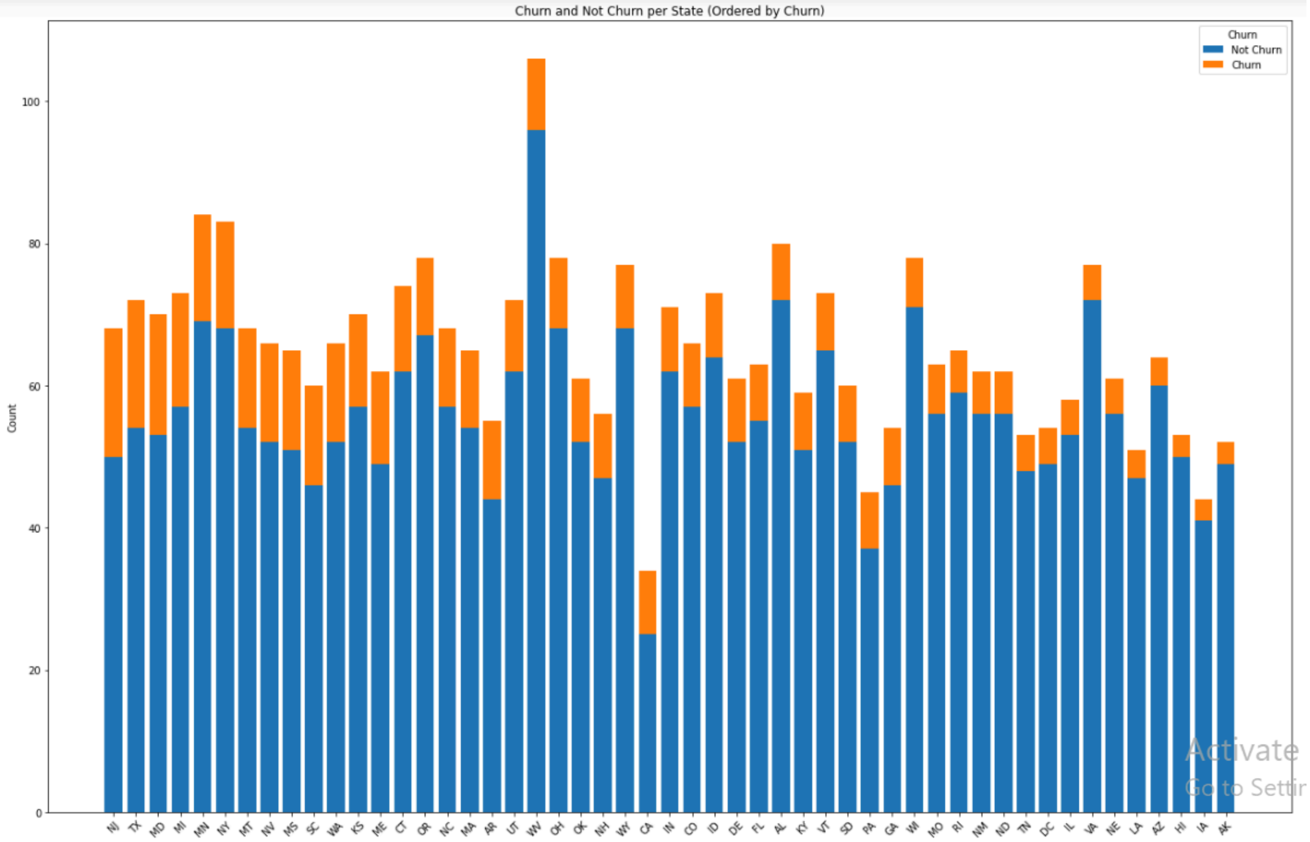


- Account length is positively skewed.
- Total Intl Calls is positively skewed.
- Total Day Minutes is nearly uniformly distributed.
- Majority of customers do not use the Voice Mail Messages.

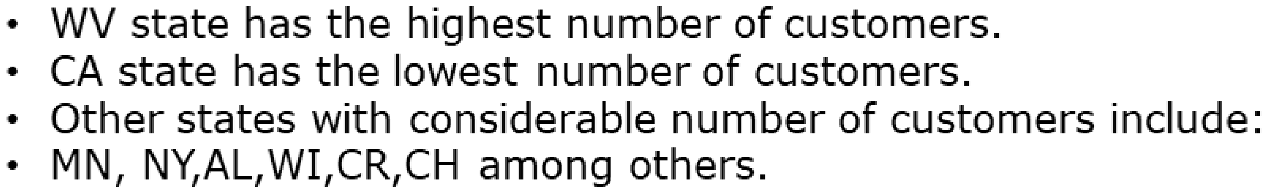
Distribution of Churn per State

Below plots shows how the churn customer are distriuted per State.





WV state has the highest number of customers. CA state has the lowest number of customers. Other states with considerable number of customers include: MN, NY,AL,WI,CR,CH among others.

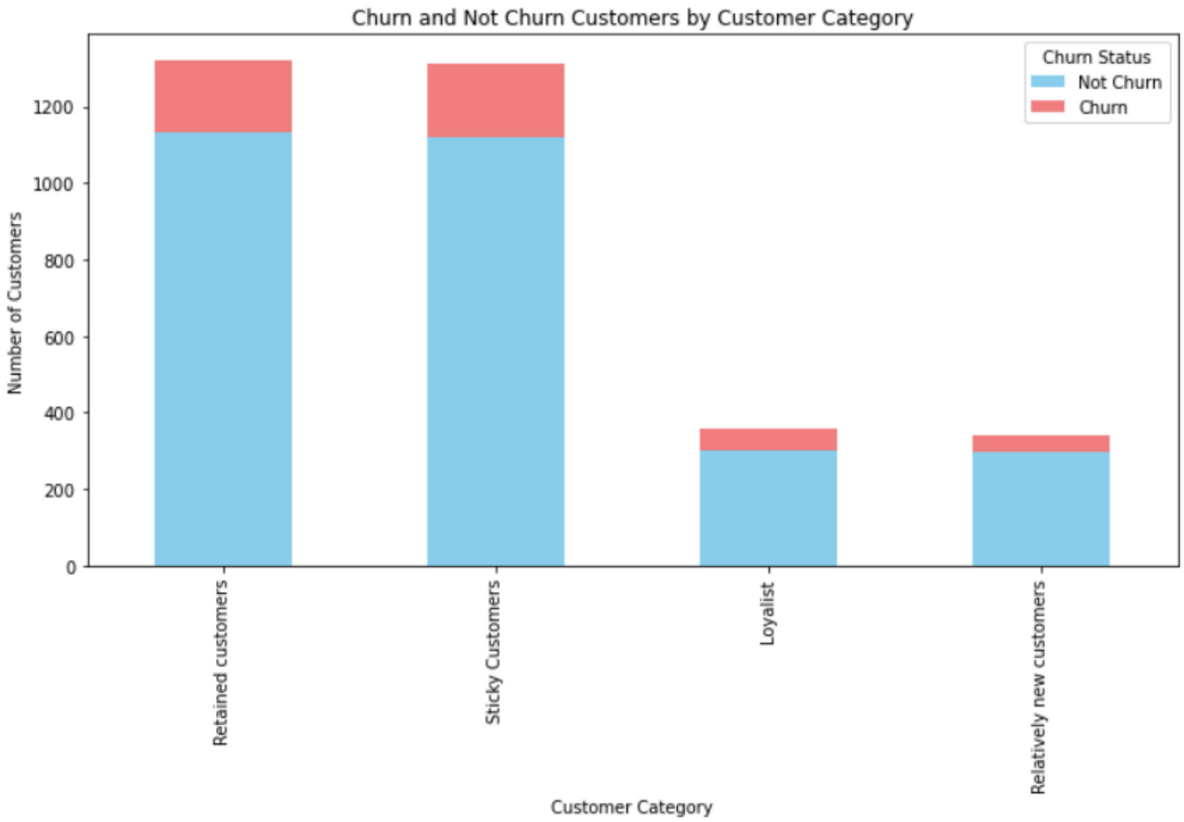


- WV state has the highest number of customers.
- CA state has the lowest number of customers.
- Other states with considerable number of customers include:
- MN, NY,AL,WI,CR,CH among others.

Bivariant Analysis

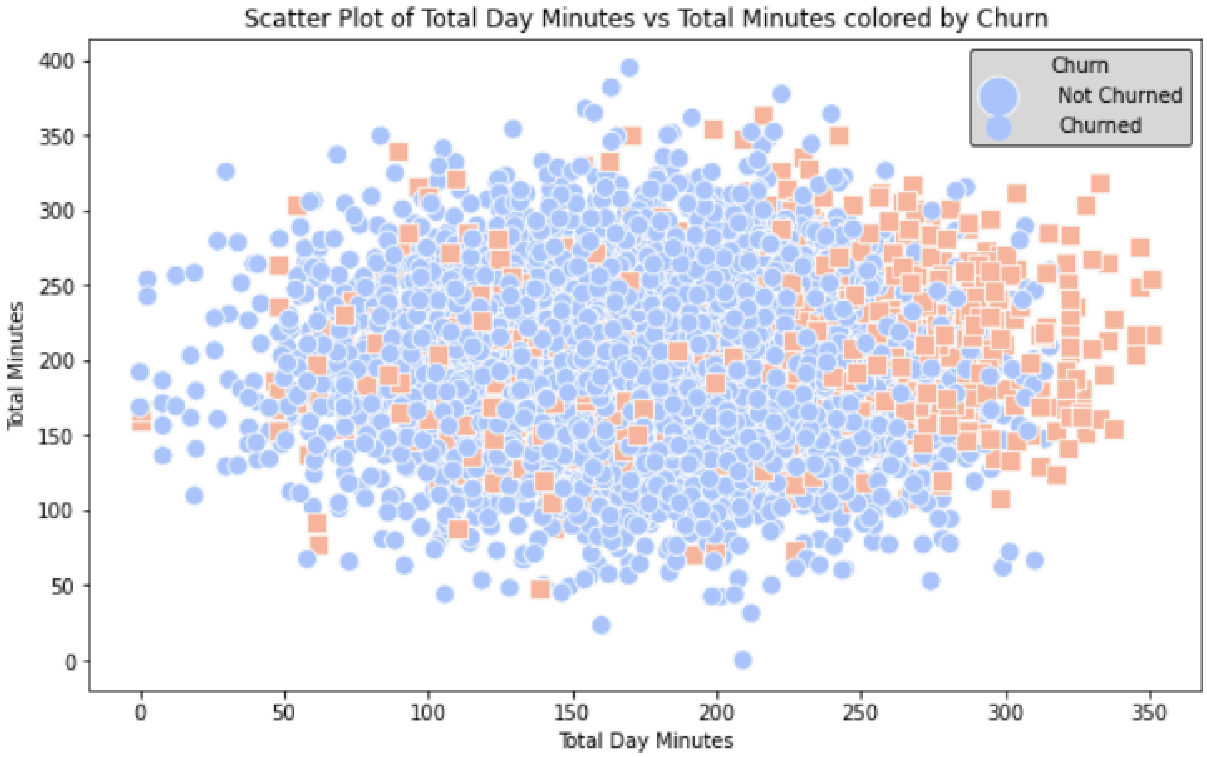
Distribution per Customer Category

Customers were categorised into Relatively New, Retained, Sticky and Loyalists. Customers with with account length <= 50 "Relatively New" Where account lenght <= 100 "Retained Customers" Where Account length <= 150 "Sticky Customers" Else "Loyalists". The distiribution was as below:



Results indicated that: • Retained and Sticky customer categories have majority of the churn. • Loyalists and relatively new customers have least least churn. • Majority of the total customers are under the Retained and Sticky customer categories.

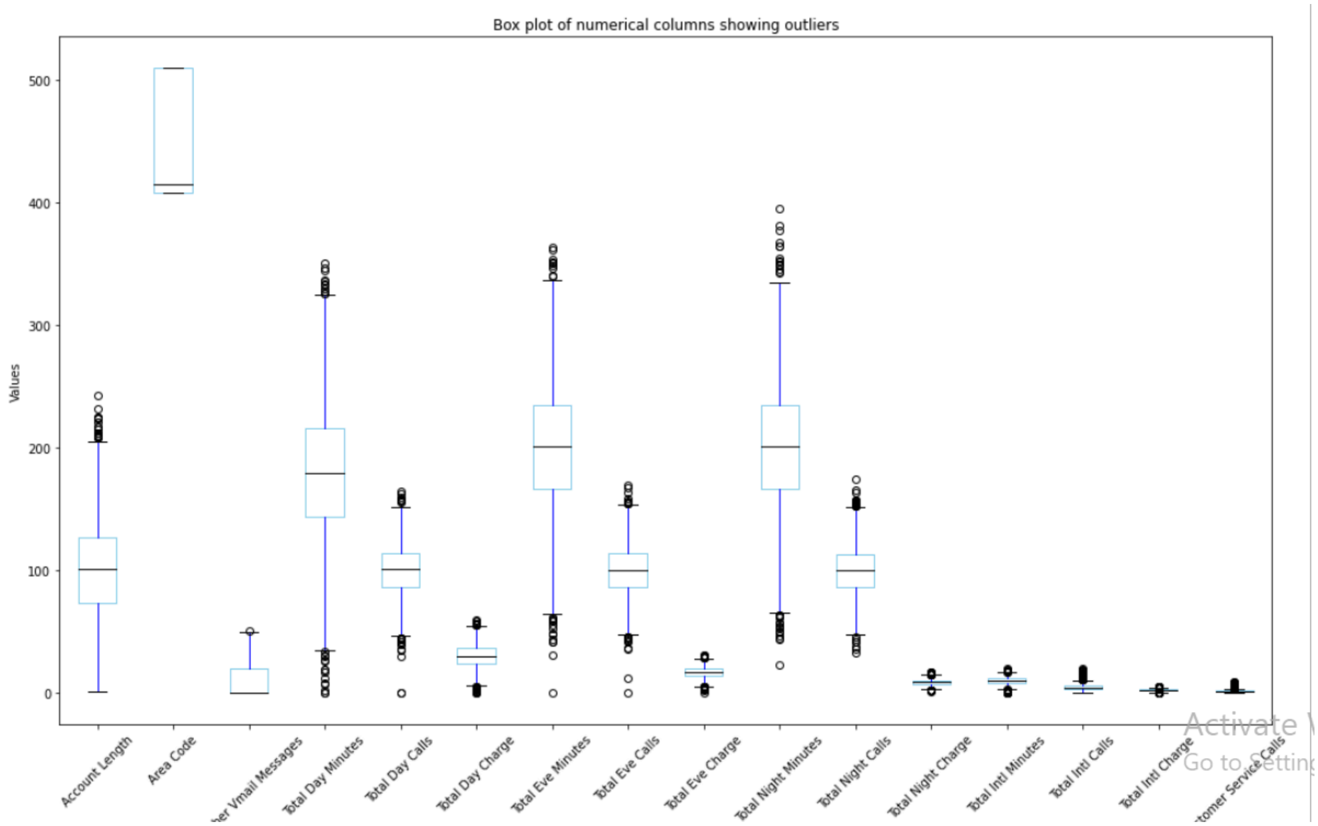
Multivariant Analysis



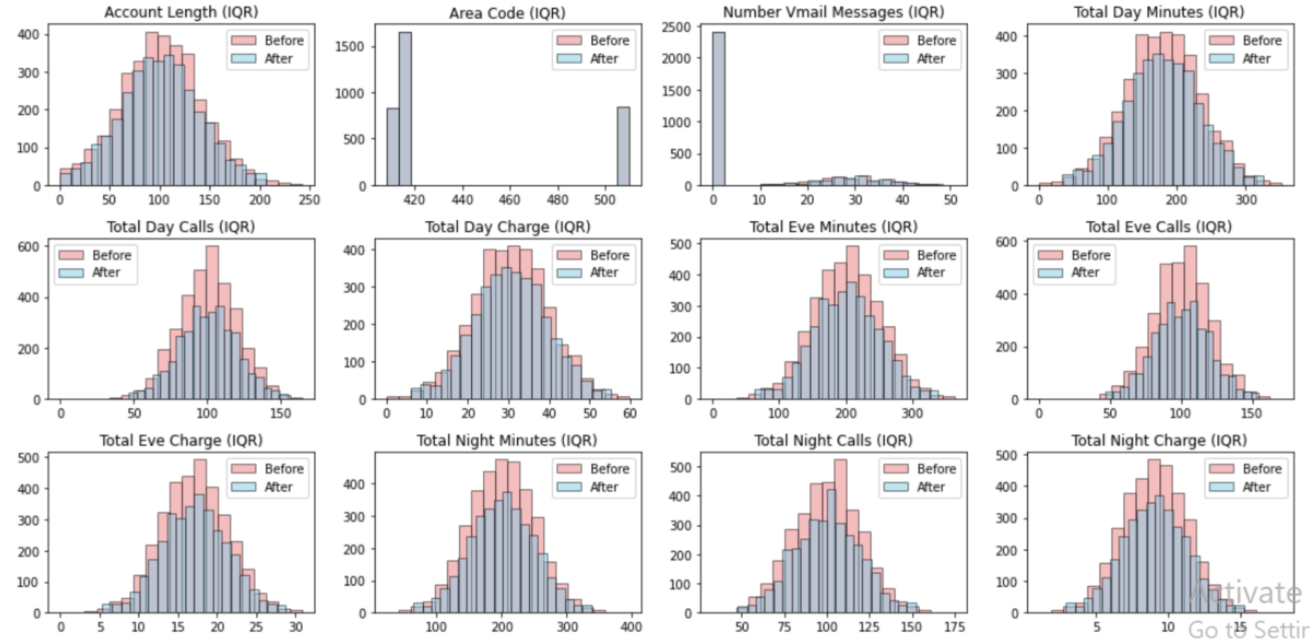
The scatter plot shows the distirituibion of churn and not churn using two numerical columns such as Total Minutes,Total Day Minutes.

Data Preparation

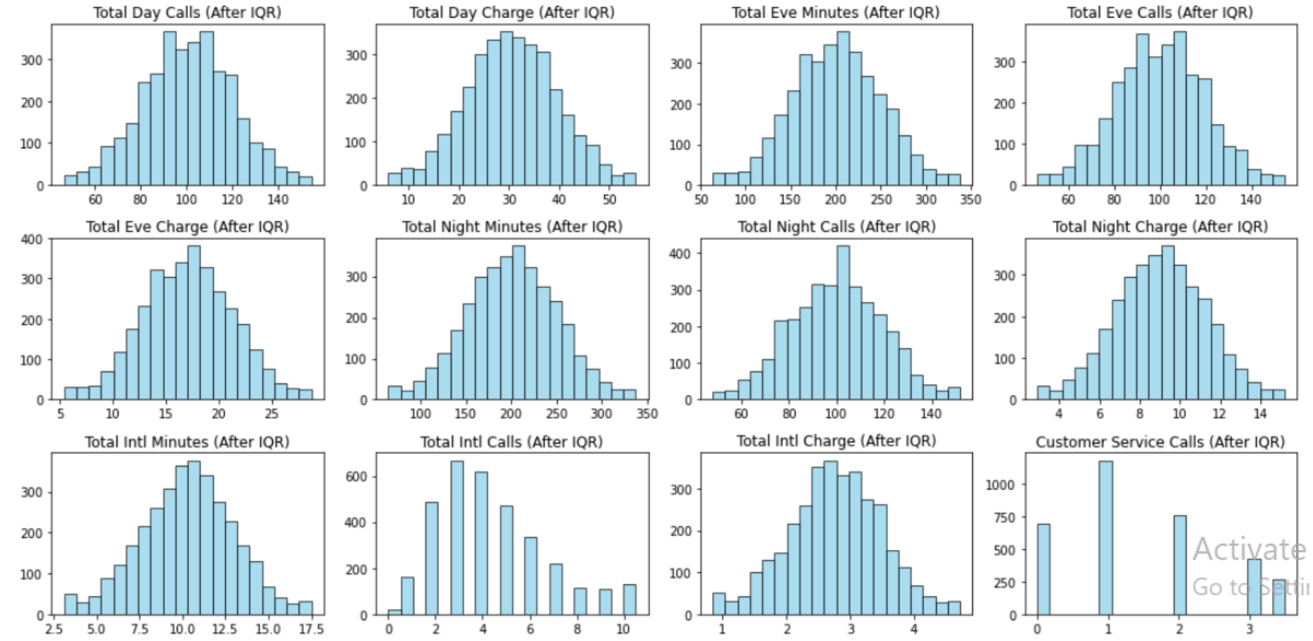
Idenfying Outliers



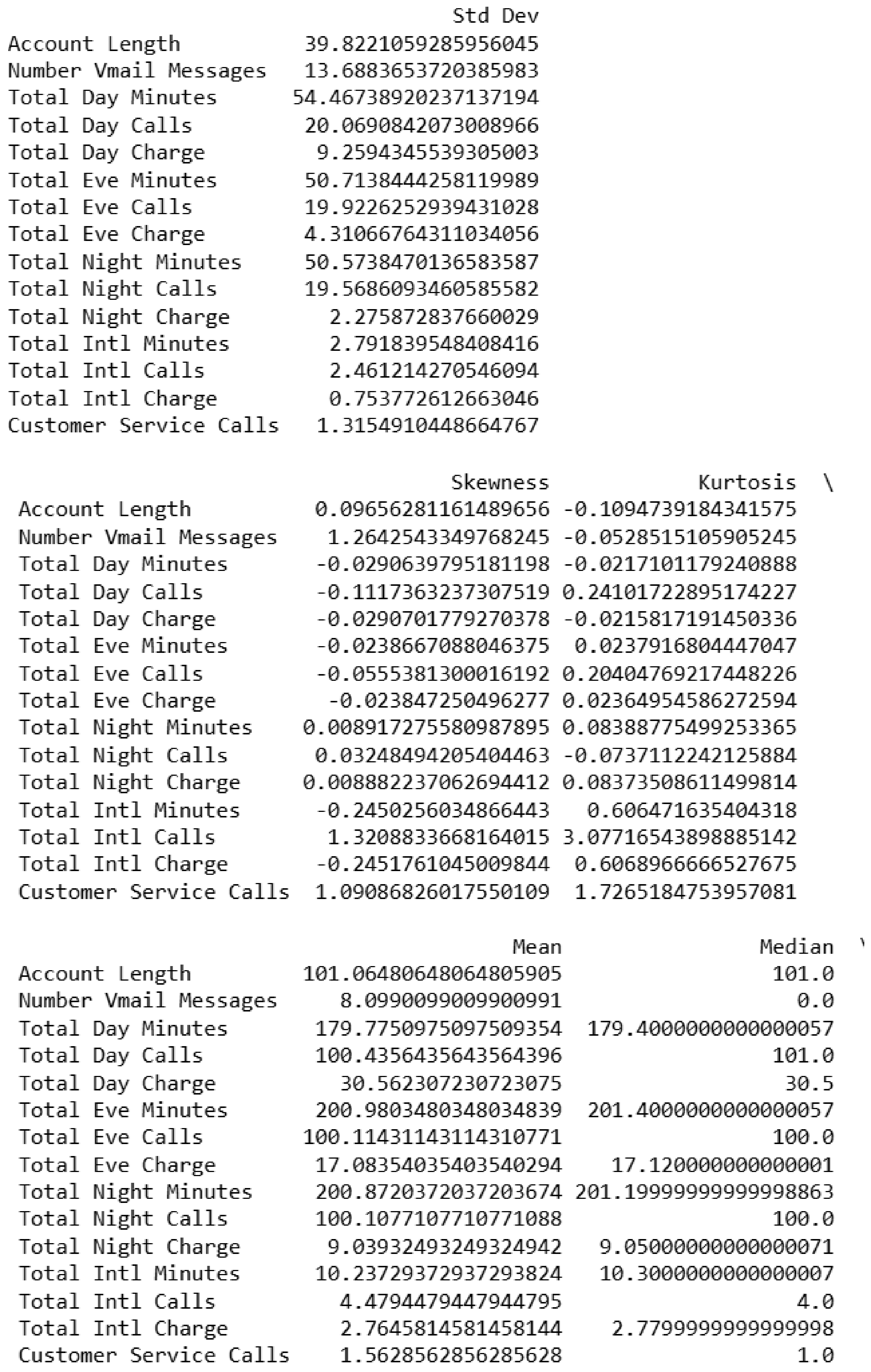
Indicating numerical columns with outliers. Distributions before and after handling outliers



Distributions of the numerical columns after removing outliers.



Normality and Spread

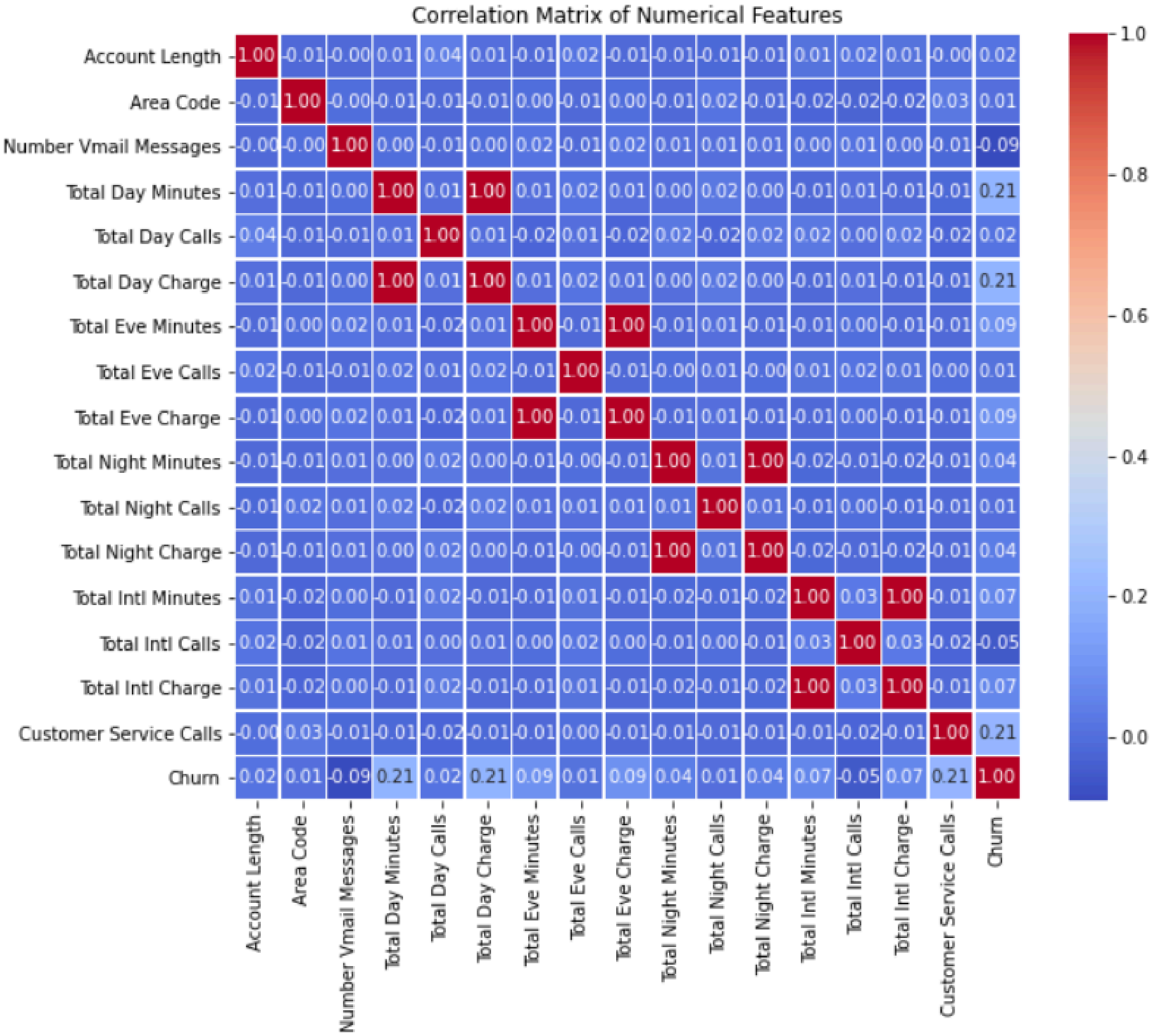


```
                                  Std Dev
Account Length          39.8221059285956045
Number Vmail Messages   13.6883653720385983
Total Day Minutes      54.46738920237137194
Total Day Calls         20.0690842073008966
Total Day Charge         9.2594345539305003
Total Eve Minutes       50.7138444258119989
Total Eve Calls         19.9226252939431028
Total Eve Charge         4.3106676431103405 6
Total Night Minutes     50.5738470136583587
Total Night Calls       19.5686093460585582
Total Night Charge        2.275872837660029
Total Intl Minutes        2.791839548408416
Total Intl Calls          2.461214270546094
Total Intl Charge         0.753772612663046
Customer Service Calls   1.3154910448664767

                                  Skewness             Kurtosis  \
 Account Length          0.09656281161489656 -0.1094739184341575
 Number Vmail Messages    1.2642543349768245 -0.052851510590524 5
 Total Day Minutes       -0.0290639795181198 -0.0217101179240888
 Total Day Calls         -0.1117363237307519 0.24101722895174227
 Total Day Charge        -0.0290701779270378 -0.0215817191450336
 Total Eve Minutes       -0.0238667088046375  0.0237916804447047
 Total Eve Calls         -0.0555381300016192 0.20404769217448226
 Total Eve Charge         -0.023847250496277 0.02364954586272594
 Total Night Minutes    0.008917275580987895 0.08388775499253365
 Total Night Calls       0.03248494205404463 -0.0737112242125884
 Total Night Charge     0.008882237062694412 0.08373508611499814
 Total Intl Minutes      -0.2450256034866443   0.606471635404318
 Total Intl Calls         1.3208833668164015 3.07716543898885142
 Total Intl Charge       -0.2451761045009844  0.6068966666527675
 Customer Service Calls  1.09086826017550109  1.7265184753957081

                                     Mean                 Median  \
 Account Length         101.06480648064805905                  101.0
 Number Vmail Messages    8.0990099009900991                    0.0
 Total Day Minutes       179.7750975097509354  179.4000000000000057
 Total Day Calls         100.4356435643564396                  101.0
 Total Day Charge          30.562307230723075                   30.5
 Total Eve Minutes       200.9803480348034839  201.4000000000000057
 Total Eve Calls         100.1143114311431077 1                100.0
 Total Eve Charge         17.08354035403540294    17.120000000000001
 Total Night Minutes     200.8720372037203674 201.1999999999998863
 Total Night Calls       100.1077107710771088                  100.0
 Total Night Charge        9.03932493249324942   9.0500000000000071
 Total Intl Minutes      10.23729372937293824   10.3000000000000007
 Total Intl Calls          4.4794479447944795                    4.0
 Total Intl Charge         2.7645814581458144    2.7799999999999998
 Customer Service Calls    1.5628562856285628                    1.0
```

Account length is positively skewed. Total Intl calls is negatively skewed Total Day Minutes has the highest std.

Total Intl Charge has the lowest std. Avg Total Day Charge is 30.5 Avg Total Eve Calls is 100.1 Avg Intl minutes 10.2

Correlation Matrix

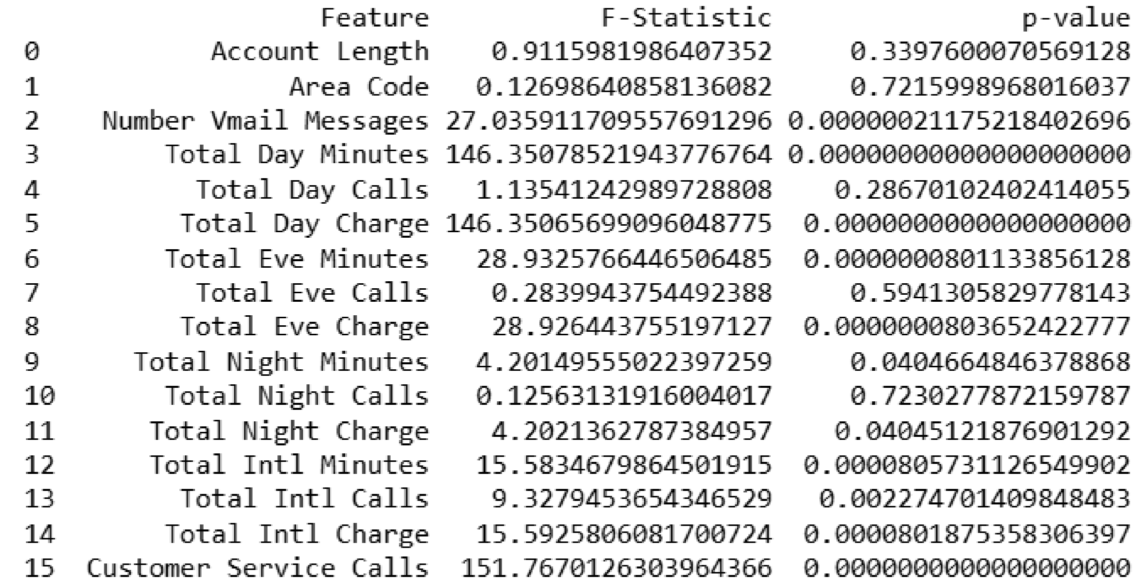


Results indicated that features that are highly correlated included: Total Day Charge and Total Day Minutes. • Total Eve Charge and Total Eve Minutes. • Total Night Charge and Total Nights Minutes Total Intl charge and Total Minutes. • This offers insights on opportunities for better packages and loyalty programs.

Hypothesis Testing

Null Hypothesis (H0): There is no significant influence of the various factors to churn rate in SyriaTel.

Alternate Hypothesis (H1): There is a significant influence of the various factors to churn rate in SyriaTel. Results were as follows



```
                  Feature           F-Statistic                p-value
0          Account Length    0.9115981986407352     0.3397600070569128
1               Area Code   0.12698640858136082     0.7215998968016037
2   Number Vmail Messages 27.035911709557691296 0.00000021175218402696
3       Total Day Minutes 146.35078521943776764 0.00000000000000000000
4         Total Day Calls   1.13541242989728808    0.28670102402414055
5        Total Day Charge 146.35065699096048775  0.0000000000000000000
6       Total Eve Minutes   28.9325766446506485  0.0000000801133856128
7         Total Eve Calls    0.2839943754492388     0.5941305829778143
8        Total Eve Charge    28.926443755197127  0.0000000803652422777
9     Total Night Minutes   4.20149555022397259     0.0404664846378868
10      Total Night Calls   0.12563131916004017     0.7230277872159787
11     Total Night Charge    4.2021362787384957    0.04045121876901292
12     Total Intl Minutes   15.5834679864501915  0.0000805731126549902
13       Total Intl Calls    9.3279453654346529   0.002274701409848483
14      Total Intl Charge   15.5925806081700724  0.0000801875358306397
15 Customer Service Calls  151.7670126303964366  0.0000000000000000000
```

Conclusion Features such as 'Account Length', 'Area Code', 'Total Day Calls', 'Total Eve Calls', and 'Total Night Calls' have their p-values are greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis (H0) for these features. This suggests that there is no significant influence of these factors on the churn rate in SyriaTel.

Remaining features including 'Number Vmail Messages', 'Total Day Minutes', 'Total Day Charge', 'Total Eve Minutes', 'Total Eve Charge', 'Total Night Minutes', 'Total Night Charge', 'Total Intl Minutes', 'Total Intl Calls', 'Total Intl Charge', and 'Customer Service Calls', the p-values are extremely low (close to 0). Therefore, we reject the null hypothesis (H0) for these features. This indicates that there is a significant influence of these factors on the churn rate in SyriaTel.

In conclusion, there was evidence to suggest that most numerical features have a significant influence on the churn rate in SyriaTel, except for 'Account Length', 'Area Code', 'Total Day Calls', 'Total Eve Calls', and 'Total Night Calls'

Modeling

Models performed included Logistic Regression, Decision Tree, KNN and XGBoost models have been built to answer the research questions.

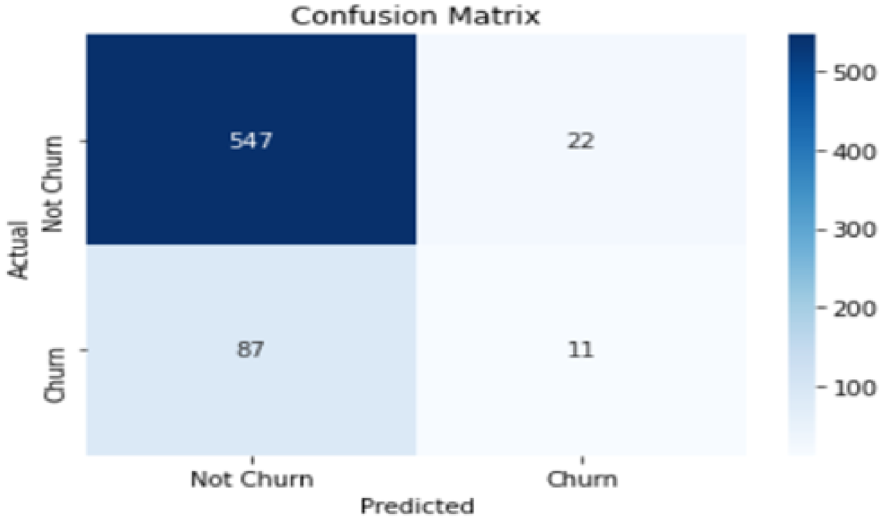
Data Split criteria

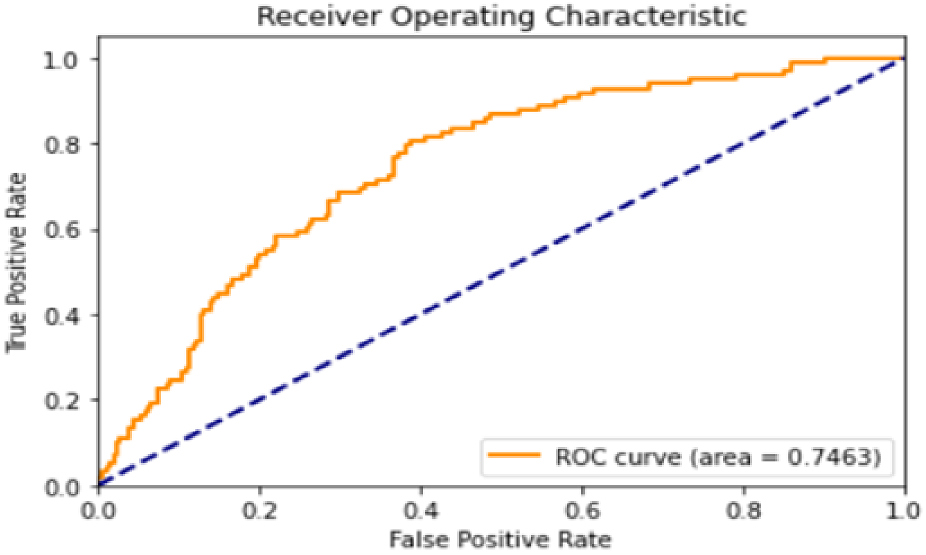
Split the data into training and testing sets

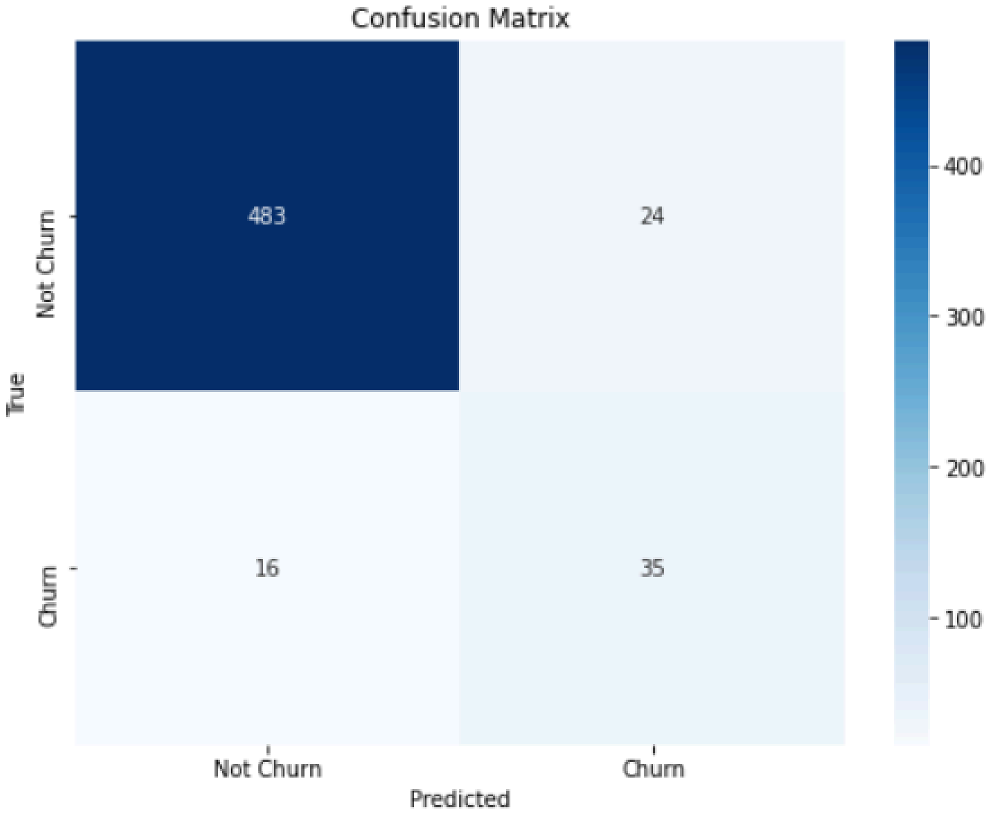
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=1, stratify=y)

Baseline Logistic Regression Model

From the above split, logististic regression model was performed. Results:





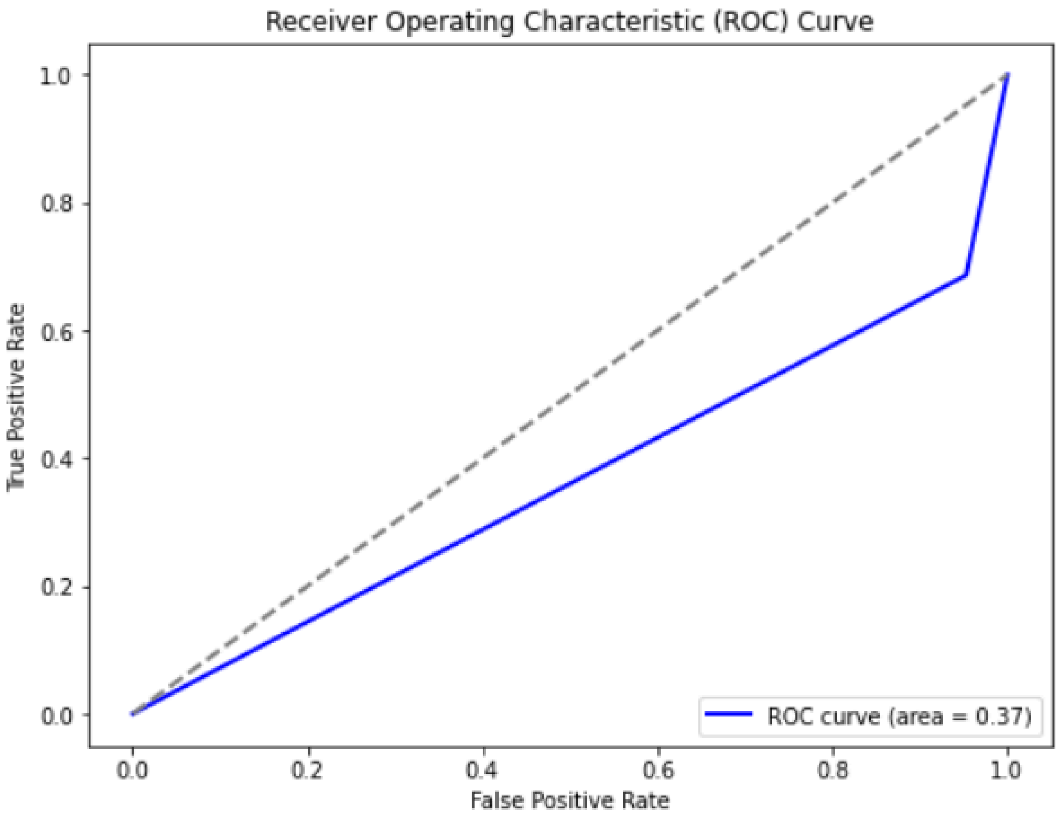


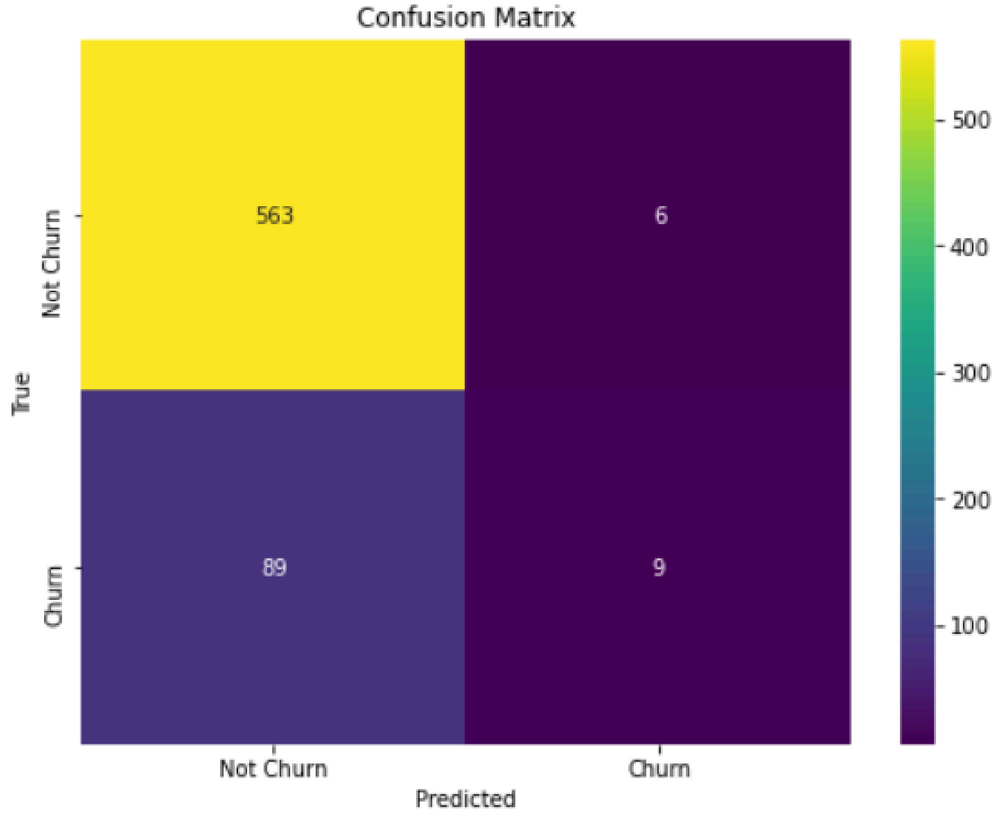
```
Confusion Matrix:
[[547  22]
 [ 87  11]]
Accuracy: 0.8366
Precision: 0.3333
Recall: 0.1122
F1-Score: 0.1679
ROC-AUC: 0.7463
```

• The model has high accuracy but struggles with precision and recall for the churn class. • Suggesting that while it correctly predicts the majority of 'no churn' cases, • It fails to adequately identify 'churn' cases. • Therefore the need to consider other model techniques

Decision Tree Model

Given the above performance, a decision tree model was conducted that gave below results.



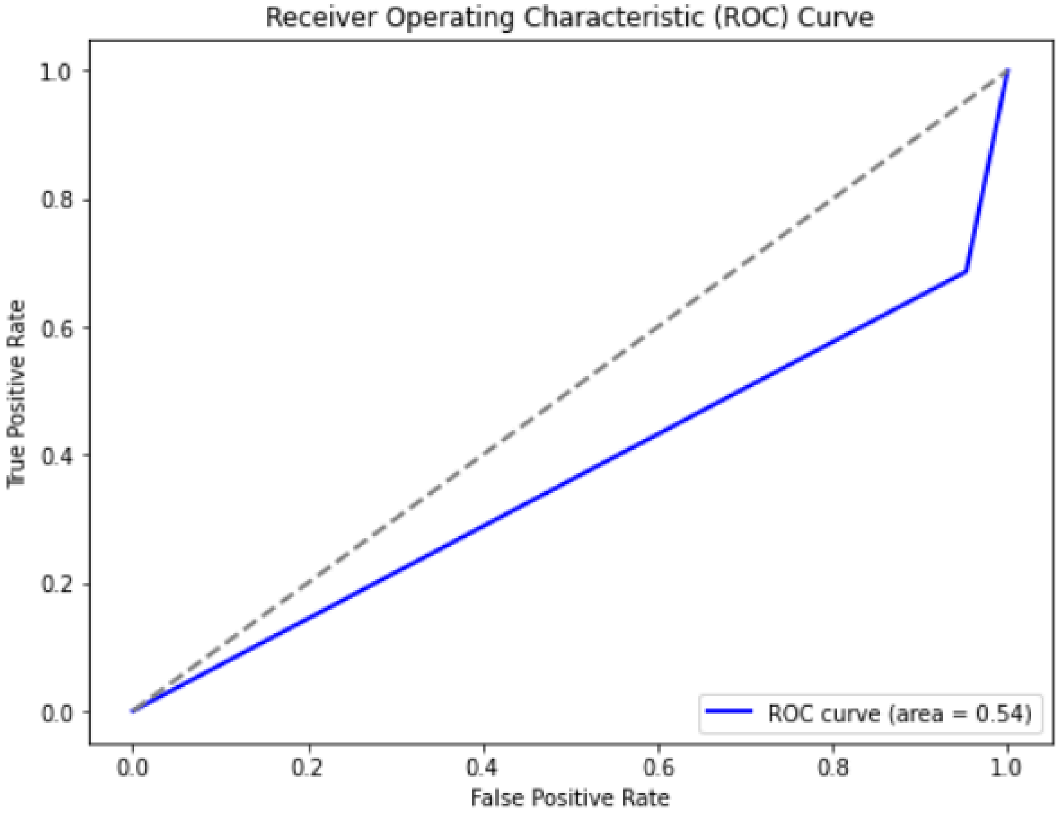


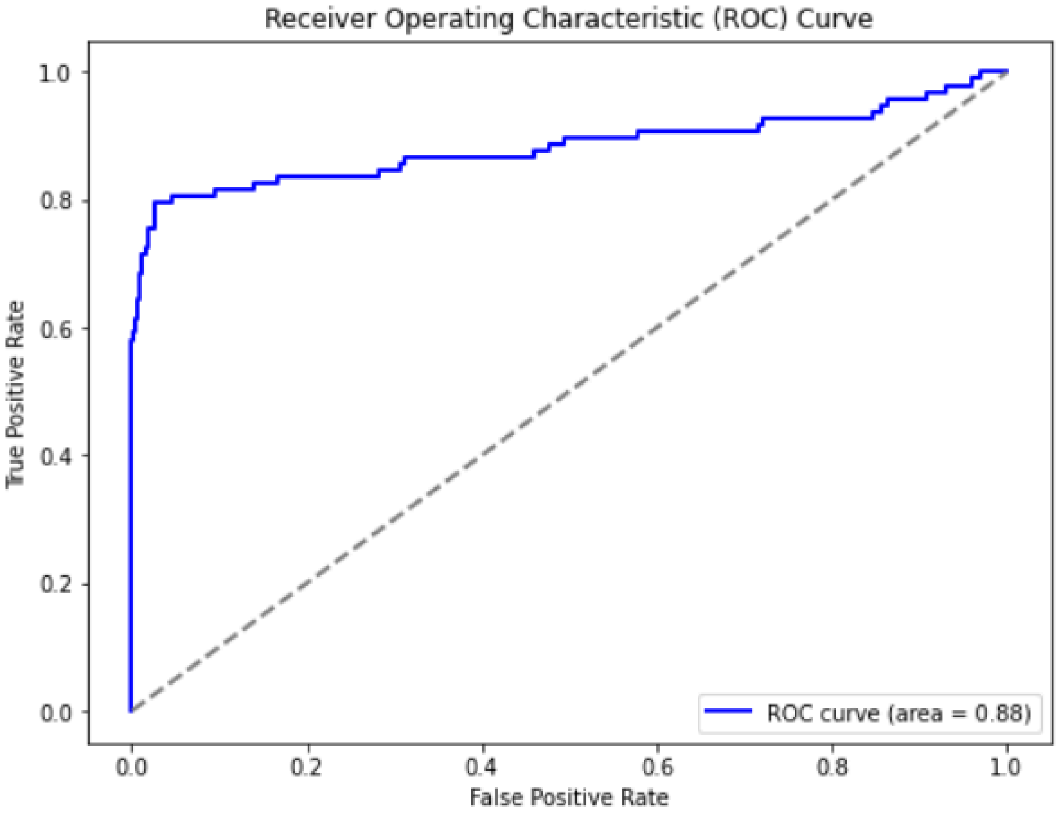
```
Confusion Matrix:
[[483  24]
 [ 16  35]]
Accuracy: 0.9283
Precision: 0.5932
Recall: 0.6863
F1-Score: 0.6364
ROC-AUC: 0.3668
```

Observations While the Decision tree model has improved in terms of accuracy, precion,Recall and F1-Score, It has a lower ROC-AUC meaning that it may predict alot of false positives which may in this case mean predicting alot of Churn which may not be the true case. Based on the above comparison, we proceed to perform other models.

KNN model

With use of the appriapriate libraries, split data a KNN model was perform. Results



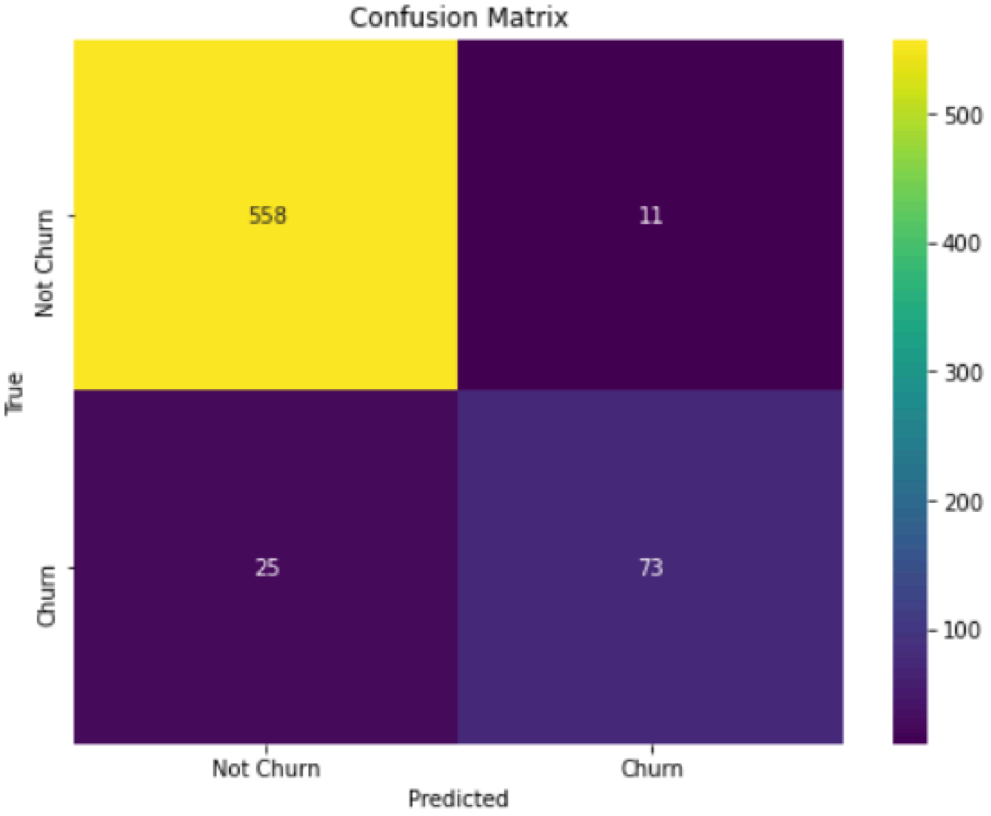


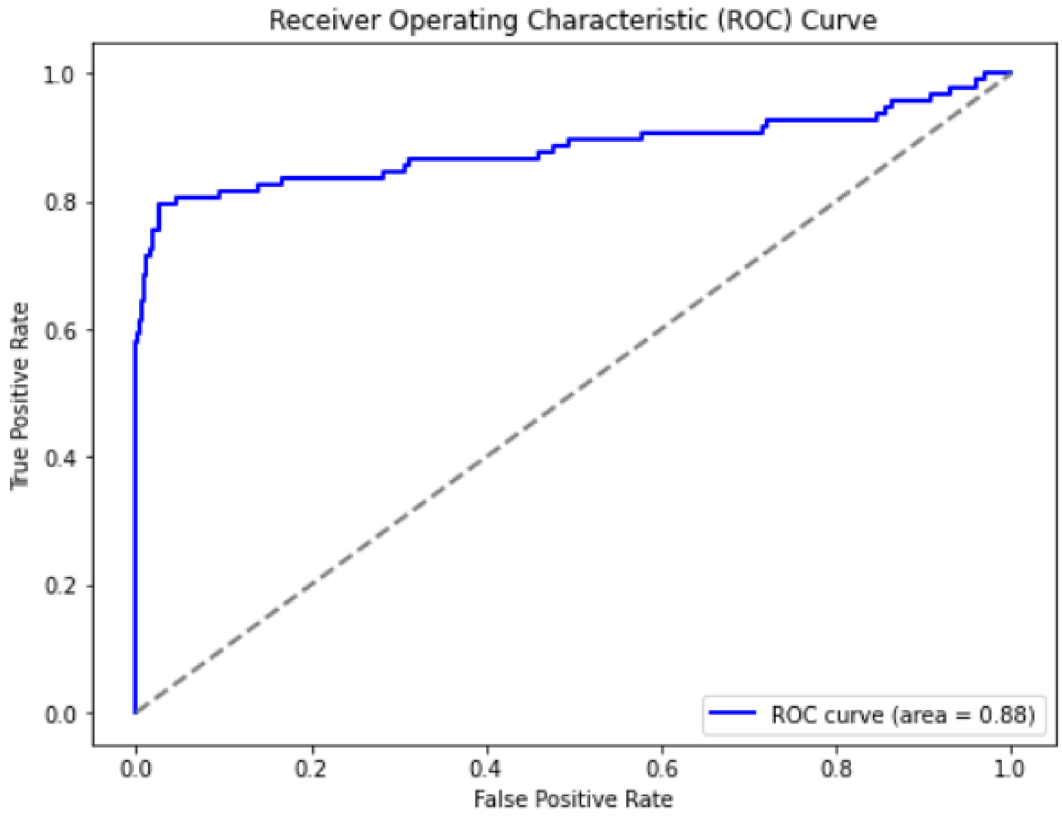
```
KNN Model Performance:
Accuracy: 0.8575712143928036
Precision: 0.6
Recall: 0.09183673469387756
F1-Score: 0.1592920353982301
ROC-AUC Score: 0.5406459596140741
```

KNN Model Performance: Accuracy: 0.8575712143928036 Precision: 0.6 Recall: 0.09183673469387756 F1-Score: 0.1592920353982301 ROC-AUC Score: 0.5406459596140741 • The accuracy is reduced while it has a higher ROC-AUC Score compared with the previous model. As this was not convincing, another model model was performed.

XGBoost Model

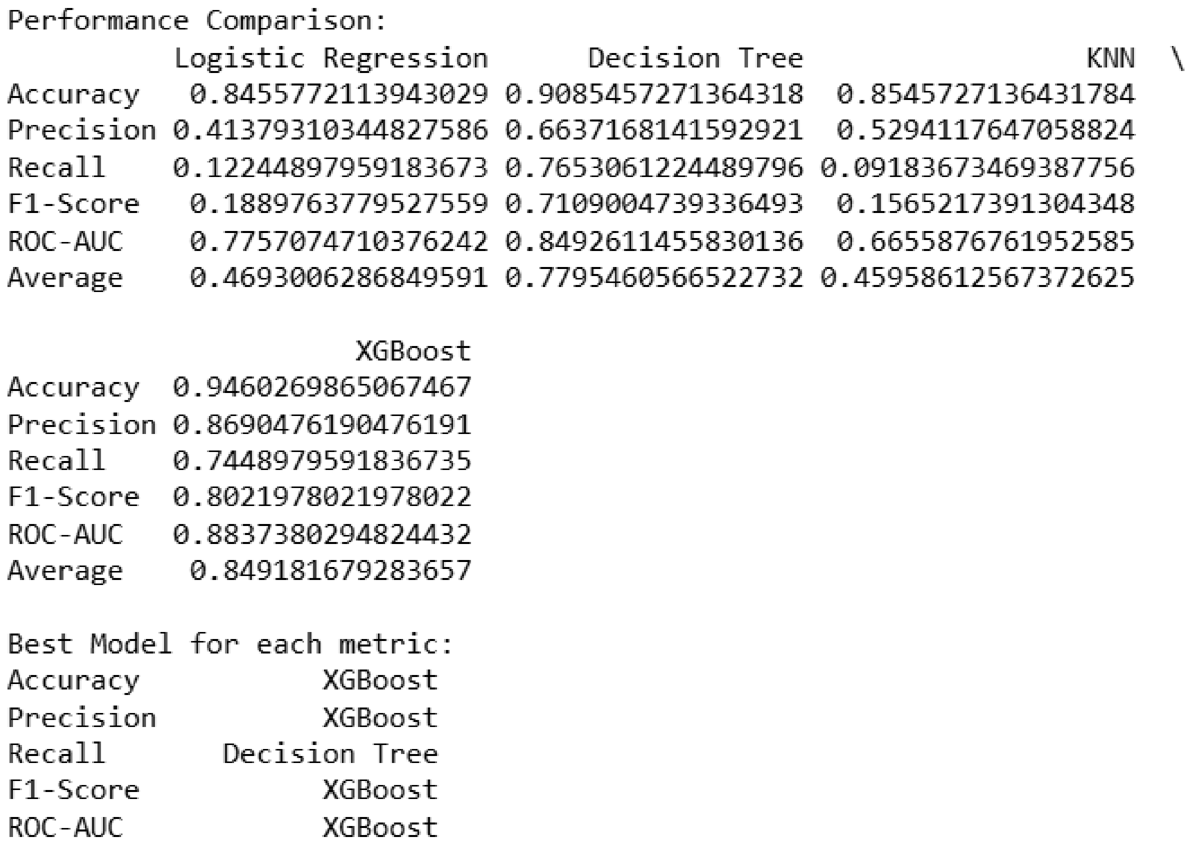
Relevant libraries were loaded, XGBoost Model was performed gave the following results.



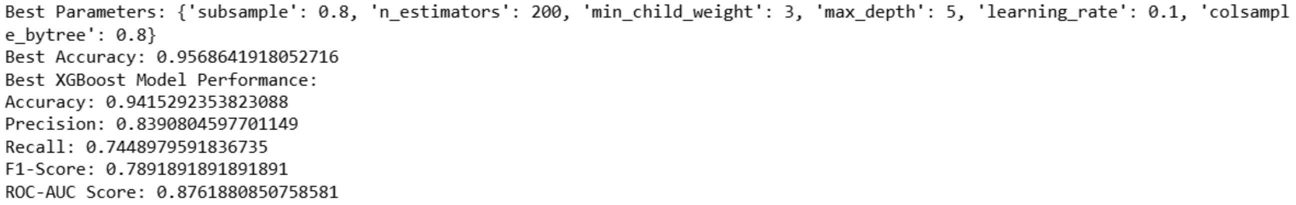


```
XGBoost Model Performance:
Accuracy: 0.9460269865067467
Precision: 0.8690476190476191
Recall: 0.7448979591836735
F1-Score: 0.8021978021978022
ROC-AUC Score: 0.8837380294824432
Confusion Matrix:
[[558  11]
 [ 25  73]]
```

Model Performance Comparison



```
Performance Comparison:
          Logistic Regression      Decision Tree                KNN  \
Accuracy   0.8455772113943029 0.9085457271364318  0.8545727136431784
Precision 0.41379310344827586 0.6637168141592921  0.5294117647058824
Recall    0.12244897959183673 0.7653061224489796 0.09183673469387756
F1-Score   0.1889763779527559 0.7109004739336493  0.1565217391304348
ROC-AUC    0.7757074710376242 0.8492611455830136  0.6655876761952585
Average    0.4693006286849591 0.7795460566522732 0.45958612567372625

                     XGBoost
Accuracy  0.9460269865067467
Precision 0.8690476190476191
Recall    0.7448979591836735
F1-Score  0.8021978021978022
ROC-AUC   0.8837380294824432
Average    0.849181679283657

Best Model for each metric:
Accuracy            XGBoost
Precision           XGBoost
Recall        Decision Tree
F1-Score            XGBoost
ROC-AUC             XGBoost
```



```
Best Parameters: {'subsample': 0.8, 'n_estimators': 200, 'min_child_weight': 3, 'max_depth': 5, 'learning_rate': 0.1, 'colsampl
e_bytree': 0.8}
Best Accuracy: 0.9568641918052716
Best XGBoost Model Performance:
Accuracy: 0.9415292353823088
Precision: 0.8390804597701149
Recall: 0.7448979591836735
F1-Score: 0.7891891891891891
ROC-AUC Score: 0.8761880850758581
```

XGBoost Model Performance: Accuracy: 0.9460269865067467 Precision: 0.8690476190476191 Recall: 0.7448979591836735 F1-Score: 0.8021978021978022 ROC-AUC Score: 0.8837380294824432

Based on the above score, a for loop function was implemented to identfiy the best performing model. The results were as follows. Uploading image.png…

XGBoost Model was identified as the best performing model.

Cross Validation of the Best performing Model.

A cross validation was done with a k = 5. The results were as below. Logistic Regression: Average Cross-validation Score = 0.8668 Decision Tree: Average Cross-validation Score = 0.9096 KNN: Average Cross-validation Score = 0.8571 XGBoost: Average Cross-validation Score = 0.9542 Indicating that XGBoost had the highest Average Cross-validation Score = 0.9542.

Hyperparameter tuning on the XGBoost Model

In order to improved the performance of the model, hyperparameter tuning was conducted on the model. The results indicated that: Uploading image.png…

Variable of Importance

Identify important variables that SyriaTl can keep monitoring, variable of importance indicated the following variables were identified as Variables of Importance in descending order include. • Total Day Minutes Total Eve Minutes • Total Night Minutes • Total Night Calls • Account Length • Total Eve Calls • Customer Service Calls • Total Intl Calls • Intl Pla • Number of Vmail Messeges • This provides insights of the variables to consider in developing churn mitigation strategies