NON-TECHNICAL PRESENTATION FOR A CLASSIFICATION TO PREDICT WHETHER A CUSTOMER WILL STOP DOING BUSINESS WITH A TELECOMMUNICATIONS.

(a)OVERVIEW & DATA UNDERSTANDING

(i)Background Information

Telecommunication companies face several challenges in trying to retain their customers due to the industry being competitive with several Telecommunication companies. Customer churn is the act of customers leaving a service provider. This directly impacts revenue and growth. Understanding the reasons behind churn and predicting which customers are likely to leave allows companies to take proactive measures to retain them and also reduces the losses made by the company.

(ii)Problem Statement

Customer churn has a substantial effect on the profitability of telecommunications organizations. Retaining the customers they already have is way cheaper than getting new customers. This project seeks to predict churn using machine learning, from which actionable insights can be drawn for retention strategies and in reducing how much money is lost because of customers who don't stick around very long.

(iii)Objectives

1. To develop machine learning models to be able to predict customer churn with high accuracy.
2. To identify the key factors driving customer churn in the company.
3. To provide actionable recommendations based on data-driven insights to enhance customer retention efforts.

(iv)Success Criteria

1. Having models that have accurate results.(Accuracy)
2. Ensure that the model have the ability to distinguish between classes comprehensively.(ROC AUC)

(v)Data Understanding

This analysis uses data sourced from Kaggle named Churn in Telecom's dataset, a CSV file that provides comprehensive information on total day calls, total eve calls, customer service calls, total night calls, churn, total intl calls among other key performance factors to help the company make

informed decisions on whether a customer will ("soon") stop doing business with This project aims to analyze SyriaTel, a telecommunications company data to be able to tell if a customer will churn or not.This will help to identify at-risk customers and improve retention strategies.
The company is interested in reducing how much money is lost because of customers who don't stick around very long. By looking at various factors like total day calls, total eve calls, customer service calls, churn etc I will guide the company on how to reduce the churn rate so as to reduce loss.

The dataset has a total of 3333 rows and 21 columns.

The columns are: state, account length, area code, phone number, international plan, voice mail plan, 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 and churn

(b)DATA PREPARATION & ANALYSIS

(i)Checking the data

-I performed data cleaning so as to be able to use the data for analysis and modeling:

1. Checking for missing values:Here I got that there was no missing values in either of the columns in this dataset.
2. Checking for duplicate Values: No duplicates found were found in this dataset.
3. Checking for Outliers**:** Some outliers were identified in usage variables like total call minutes.

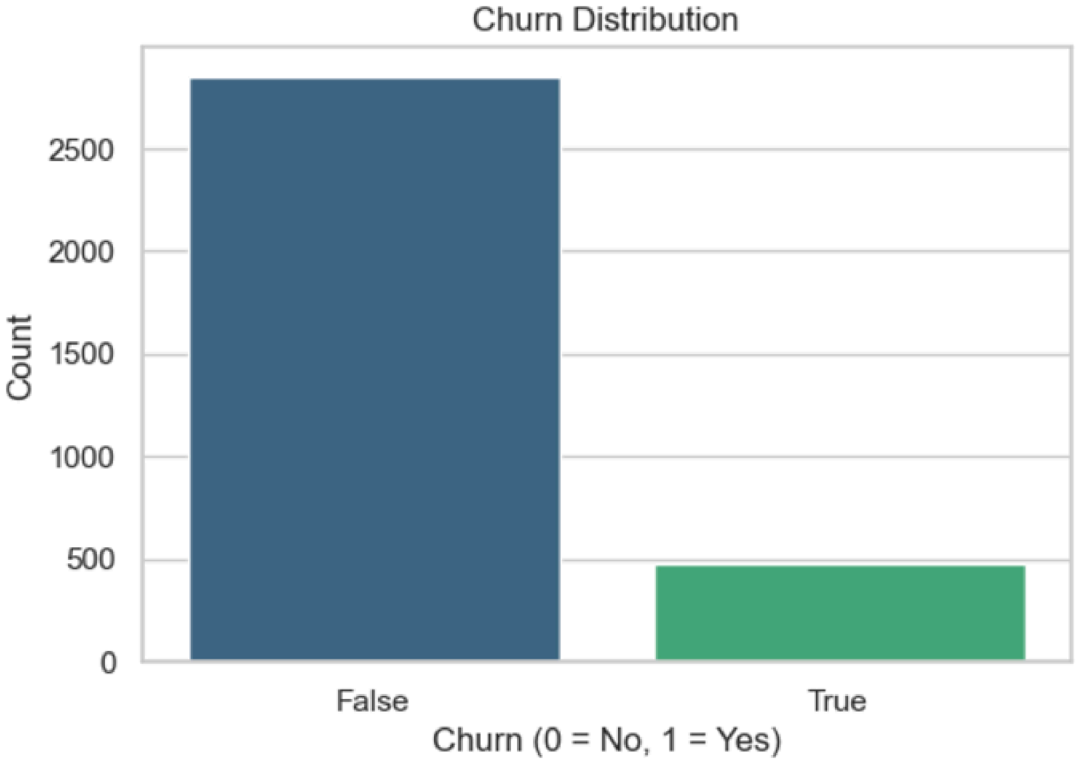
(ii)Data Cleaning

1. Missing values:This did not require any cleaning as there were no missing values in the dataset.
2. Duplicate Values:This did not require any cleaning as there were no duplicates in the dataset.
3. Outliers:Retained, as it would signify unusual customer behavior relevant to churn.
4. Dropped the phone number column as I did not need it since it is unique for every customer so it would not be helpful in modeling.
5. Categorical Encoding: Convert categorical variables like 'State' and 'International Plan' into numerical values using one-hot encoding.

(c)DATA ANALYSIS

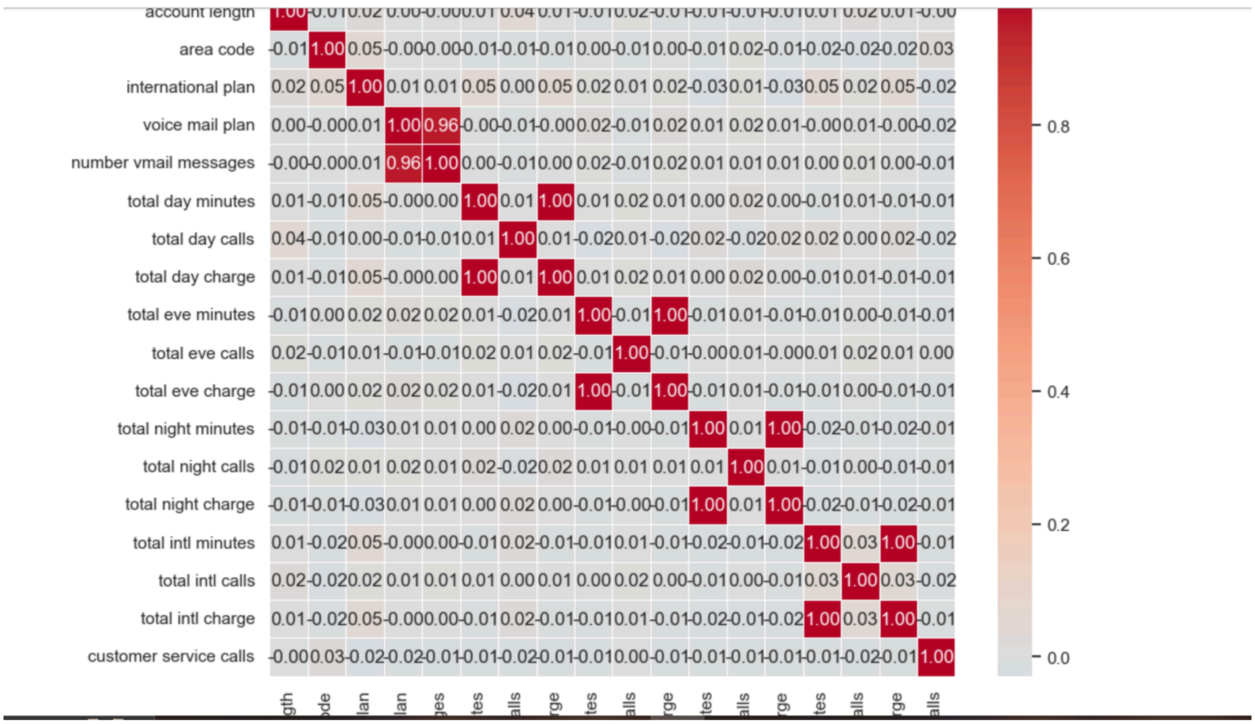
-These are the analysis I did:

1. Barchart to check the count of churn:
   -This is to check the distribution of churn rates for the telecommunication company.



2. Correlation heatmaps

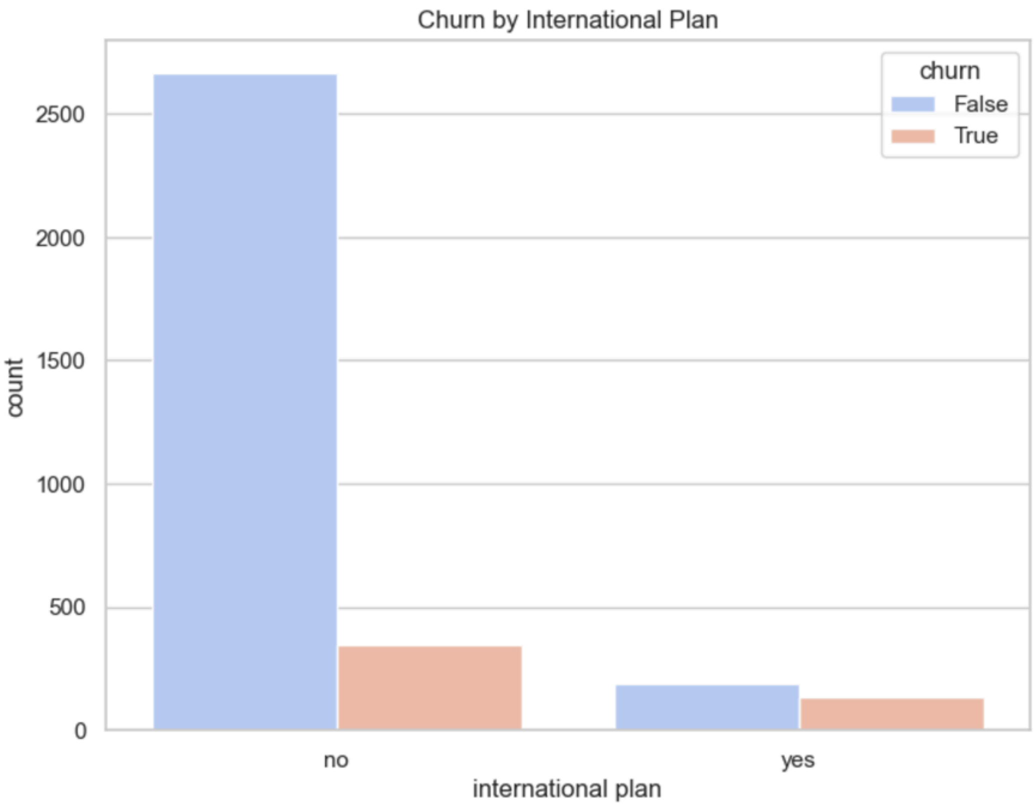
   This is to show the correlation and relation between the different features



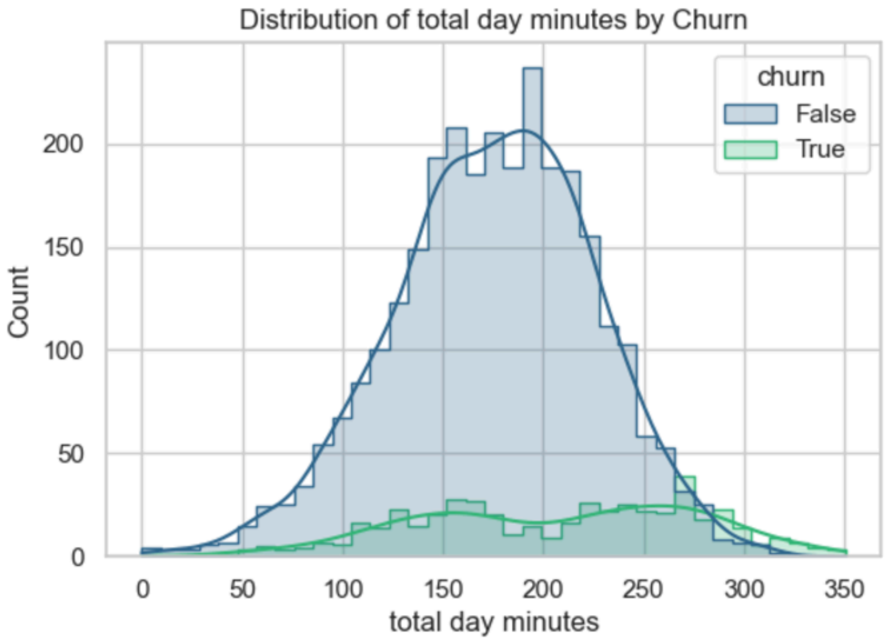
	gth	ode	lan	lan	ges	tes	alls	rge	tes	alls	rge	tes	alls	rge	tes	alls	rge	alls
account length	1.00	-0.01	0.02	0.00	-0.00	0.01	0.04	0.01	-0.01	0.02	-0.01	-0.01	-0.01	-0.01	0.01	0.02	0.01	-0.00
area code	-0.01	1.00	0.05	-0.00	-0.00	-0.01	-0.01	-0.01	0.00	-0.01	0.00	-0.01	0.02	-0.01	-0.02	-0.02	-0.02	0.03
international plan	0.02	0.05	1.00	0.01	0.01	0.05	0.00	0.05	0.02	0.01	0.02	-0.03	0.01	-0.03	0.05	0.02	0.05	-0.02
voice mail plan	0.00	-0.00	0.01	1.00	0.96	-0.00	-0.01	-0.00	0.02	-0.01	0.02	0.01	0.02	0.01	-0.00	0.01	-0.00	-0.02
number vmail messages	-0.00	-0.00	0.01	0.96	1.00	0.00	-0.01	0.00	0.02	-0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.00	-0.01
total day minutes	0.01	-0.01	0.05	-0.00	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01
total day calls	0.04	-0.01	0.00	-0.01	-0.01	0.01	1.00	0.01	-0.02	0.01	-0.02	0.02	-0.02	0.02	0.02	0.00	0.02	-0.02
total day charge	0.01	-0.01	0.05	-0.00	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01
total eve minutes	-0.01	0.00	0.02	0.02	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01
total eve calls	0.02	-0.01	0.01	-0.01	-0.01	0.02	0.01	0.02	-0.01	1.00	-0.01	-0.00	0.01	-0.00	0.01	0.02	0.01	0.00
total eve charge	-0.01	0.00	0.02	0.02	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01
total night minutes	-0.01	-0.01	-0.03	0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01
total night calls	-0.01	0.02	0.01	0.02	0.01	0.02	-0.02	0.02	0.01	0.01	0.01	0.01	1.00	0.01	-0.01	0.00	-0.01	-0.01
total night charge	-0.01	-0.01	-0.03	0.01	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01
total intl minutes	0.01	-0.02	0.05	-0.00	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01
total intl calls	0.02	-0.02	0.02	0.01	0.01	0.01	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	-0.01	0.03	1.00	0.03	-0.02
total intl charge	0.01	-0.02	0.05	-0.00	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01
customer service calls	-0.00	0.03	-0.02	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	1.00

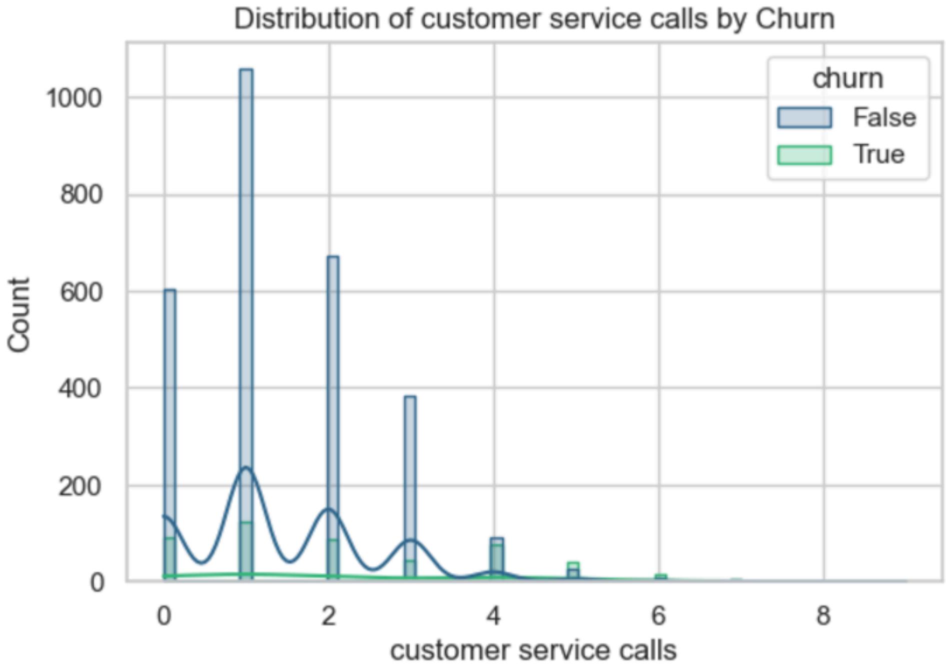
3. Barchart to check the count of churn and international plan

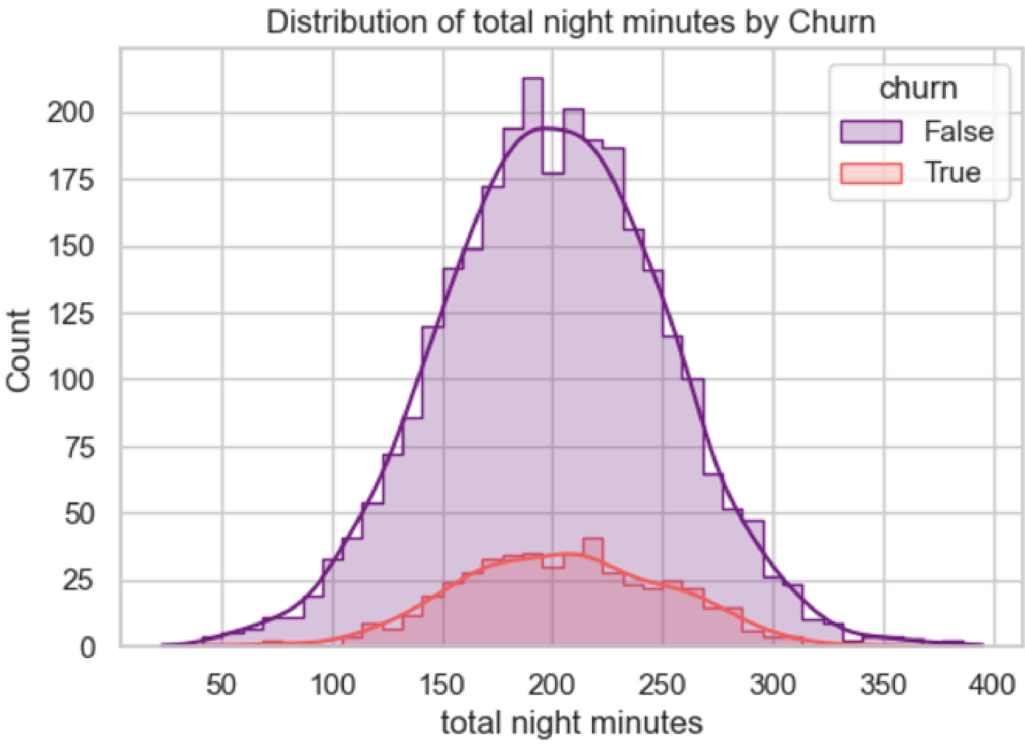
Correlation between international plan and churn: Customers with an international plan churned more frequently.

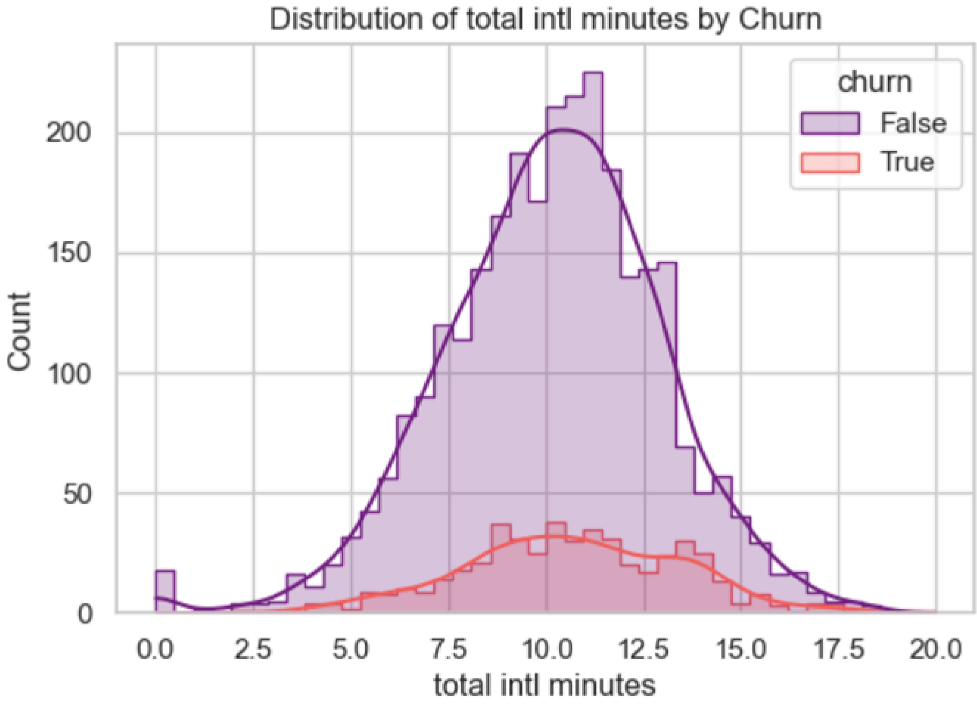


4.Plotting distributions of the: 'total day minutes', 'customer service calls', 'total night minutes', 'total intl minutes'

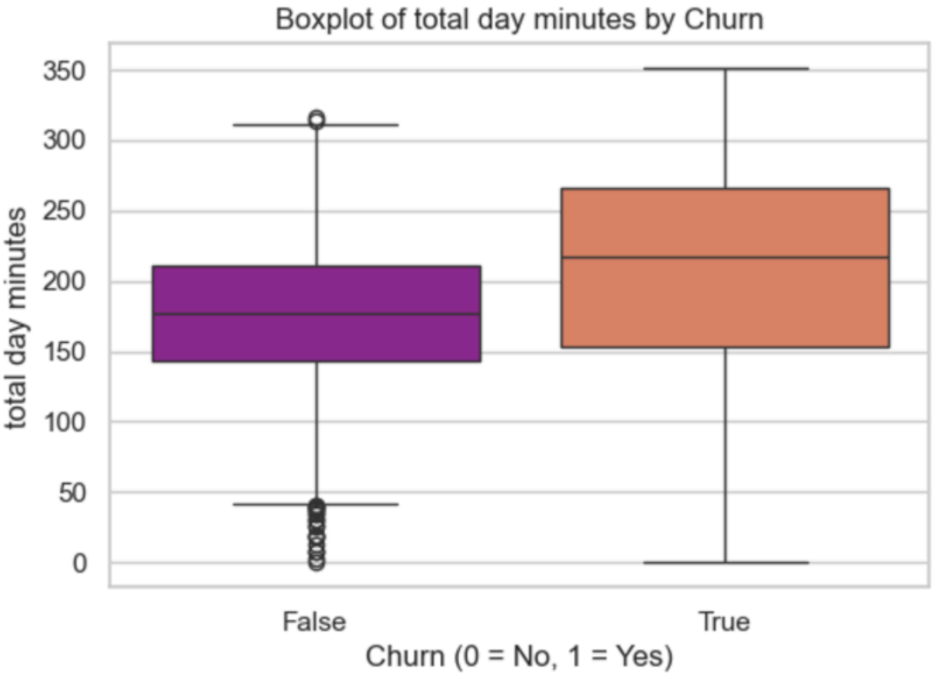

Distribution of total day minutes by Churn
churn
False
True
Count
0
50
100
150
200
0
50
100
150
200
250
300
350
total day minutes

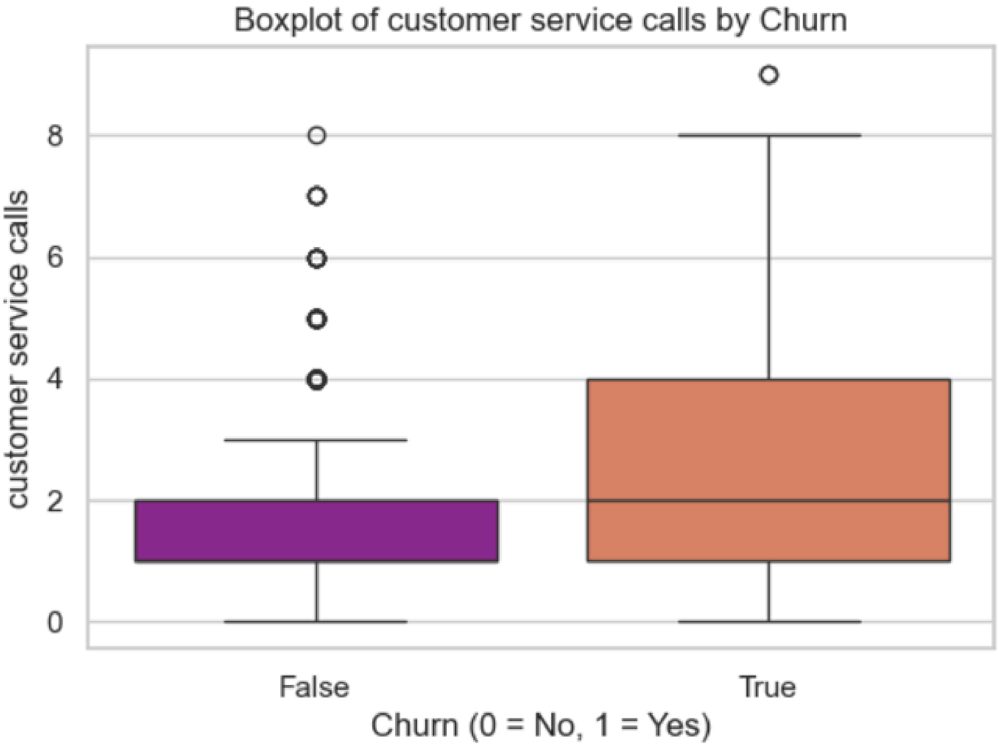

Distribution of customer service calls by Churn
churn
False
True
Count
0
200
400
600
800
1000
0
2
4
6
8
customer service calls

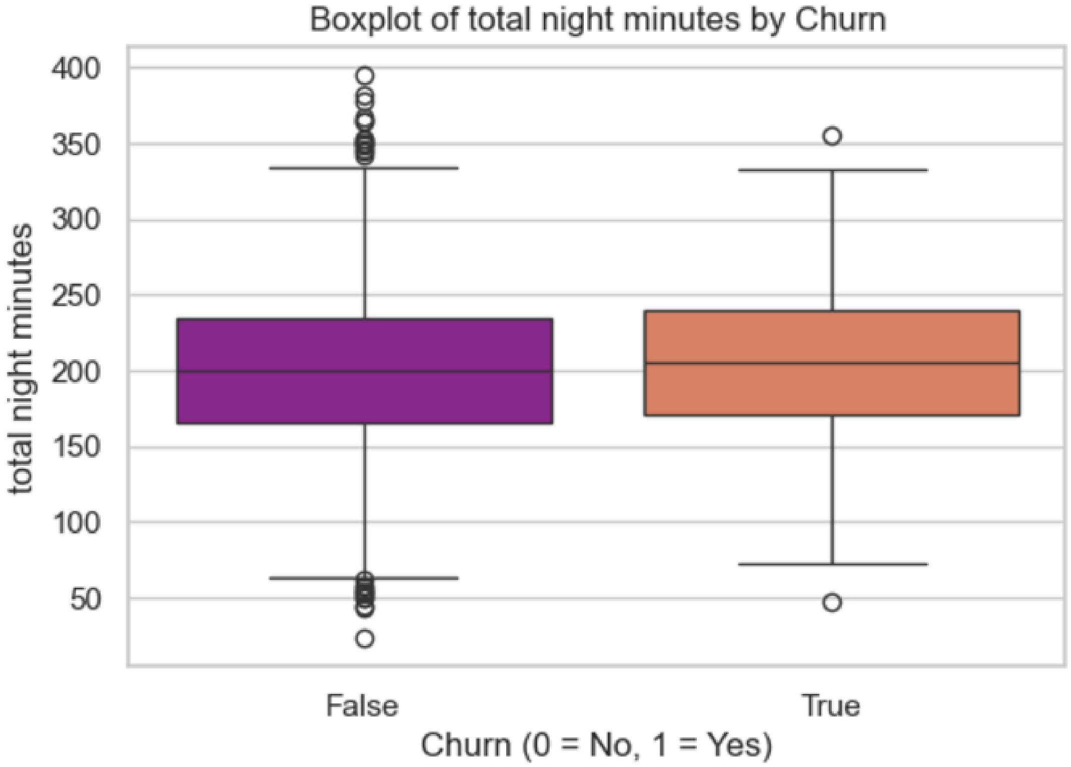


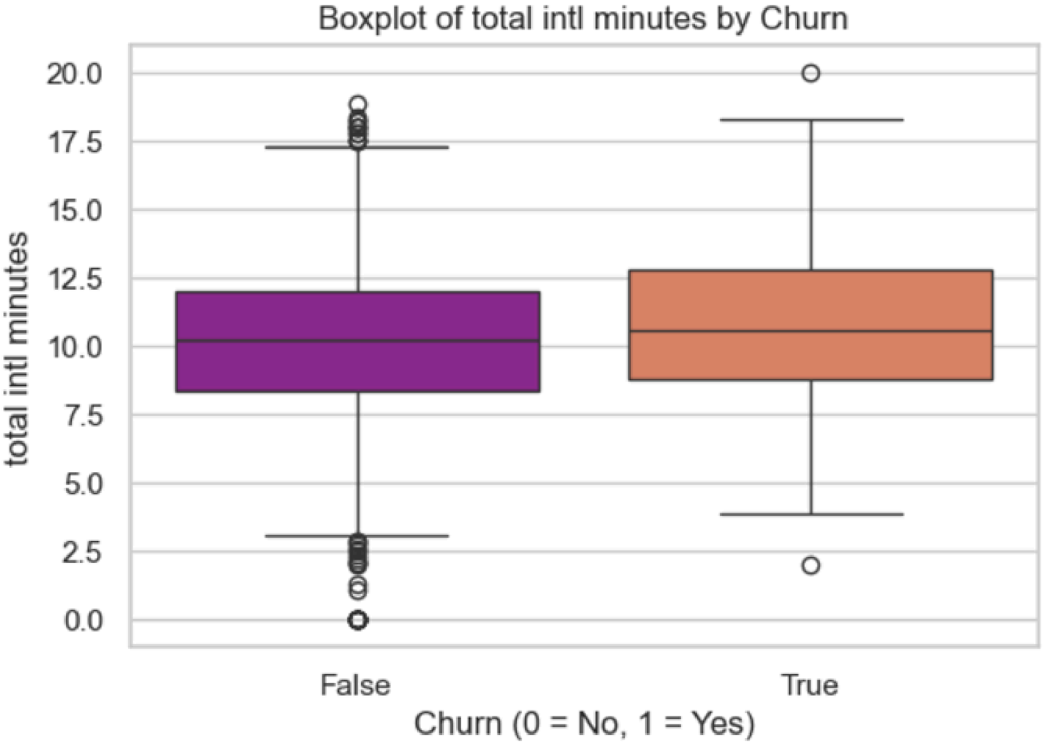


5. Plotting Boxplots for the:'total day minutes', 'customer service calls', 'total night minutes', 'total intl minutes'


Boxplot of total day minutes by Churn
350
300
250
200
150
100
50
0
total day minutes
False
True
Churn (0 = No, 1 = Yes)


Boxplot of customer service calls by Churn
8
6
4
2
0
customer service calls
False
True
Churn (0 = No, 1 = Yes)


Boxplot of total night minutes by Churn
total night minutes
400
350
300
250
200
150
100
50
False
True
Churn (0 = No, 1 = Yes)


Boxplot of total intl minutes by Churn
total intl minutes
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
False
True
Churn (0 = No, 1 = Yes)

(d). MODELING

The Models I chose to use are:

1. Baseline Model: Logistic Regression.
2. 2nd Model: Decision Trees with hyperparameter tuning
3. 3rd Model: Random Forest with hyperparameter tuning .

The reasons I used this models is because:

1. Logistic Regression is a simple model and interpretable.Also because it is a good starting point.
2. Decision Trees handles non-linear relationships well and robust to overfitting and adding hyperparameter tuning made it more accurate.
3. Random Forest with hyperparameter tuning is optimized for better accuracy, precision, recall and ROC-AUC.

Metrics of Success for my models was:

1.Accuracy: Overall prediction correctness of the model.

2.Precision & Recall: Key for identifying churners accurately the True Positive Rate and the False Positive Rate.

3.ROC- AUC: Measures the model's ability to distinguish between classes.In this case the classes was customers who have churned and those who have not.

(e)EVALUATION

(i)Each model and its performance

1.Logistic Regression

-Accuracy: 0.7691154422788605
-ROC-AUC: 0.8270475457439738
-Precision: True:0.37 False:0.95
-Recall: True:0.77 False:0.77

2.Decision trees with hyperparameter tuning
-Accuracy: 0.9175412293853074

-ROC-AUC:0.8889637196935241
-Precision: True:0.83 False:0.93
-Recall: True:0.57 False:0.98

3.Random Forest with hyperparameter tuning
-Accuracy:0.9265367316341829
-ROC-AUC: 0.9118007207081132
-Precision: True:0.79 False:0.95
-Recall: True:0.70 False:0.97

(ii)The model that performed best and why

The model that performed best is Random Forest with hyperparameter tuning and this is because of its balance of precision, recall, overall accuracy and ROC-AUC . Which shows the model is highly effective at distinguishing between churners and non-churners across all classification thresholds, which is critical for customer retention strategies.

(f)CONCLUSIONS

1.Random Forest models outperforms both Logistic Regression and Decision Trees in capturing complex patterns in the data and better modeling of the data.
2.The high accuracy and recall values of the Random Forest with hyperparameter tuning model demonstrates its reliability in predicting customer churn.
3.That Customers with international plans and high total charges are at higher risk of churn.

(g)RECOMMENDATIONS

1.I would recommend the company to focus on customers with high total charges and international plans.
2.I would recommend the company to use the tuned Random Forest model to be able to flag high-risk customers for retention efforts.
3.Based on churn insights, consider revisiting pricing structures or enhancing international plan features to better align with customer needs.
4.Introduce loyalty programs targeting high-usage customers to improve retention rates.

5.Regularly retrain the model using updated customer data to ensure its predictions remain accurate as customer behaviors evolve.