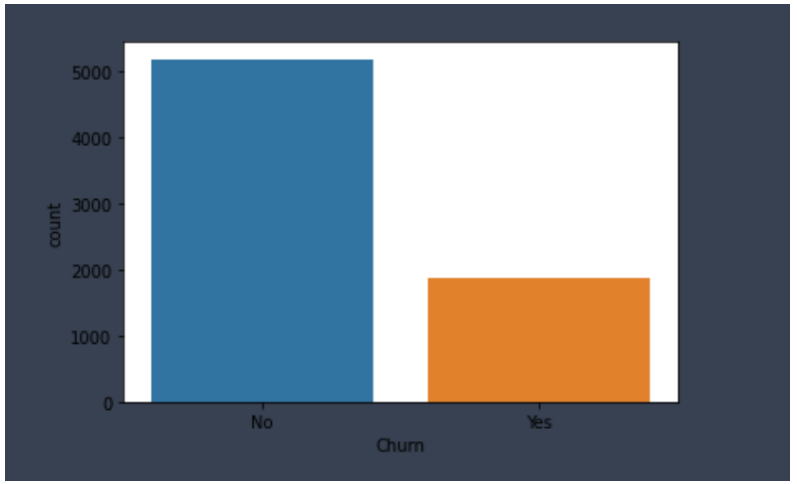
Customer Churn Analysis

(SyriaTel)



Author: Charles EGAMBI

Agenda

Project Overview

Data Modelling

Business Understanding

Conclusions

Data Analysis

Recommendations

Project Overview

SyriaTel is a telecommunications company in Syria.

They have been informed that some of their customers have started to churn and discontinue their service.

This analysis will determine what features will indicate if a customer will ("soon") discontinue their service.

Business Understanding

Customer Churning
How to prevent it

In the highly competitive telecom industry, customer churn represents a critical challenge that directly impacts profitability and market share.

The dataset under analysis offers essential insights into customer behaviour, helping to identify the key factors influencing churn.

By utilizing predictive analytics, telecom companies can proactively mitigate customer attrition, thereby optimizing retention strategies and improving overall business performance.

Problem Statement

The objective of this analysis is to develop a predictive model to anticipate customer churn in the telecom sector. By leveraging supervised classification techniques, we aim to identify key attributes and patterns that indicate potential churn among telecom customers.

Accurate churn prediction will enable telecom companies to implement targeted retention initiatives, such as personalized offers and proactive customer service interventions, ultimately reducing customer attrition and fostering long-term customer loyalty.

Data Understanding

Customer Churn indicates if a customer has terminated his or her contract with SyriaTel.

Predicting churn can help a telecom company focus its customer retention marketing efforts (such as providing special offers) on the subset of clients most likely to switch service providers.

Therefore, the “churn” column has been chosen as the target variable for this predictive analysis, which is a supervised classification problem.

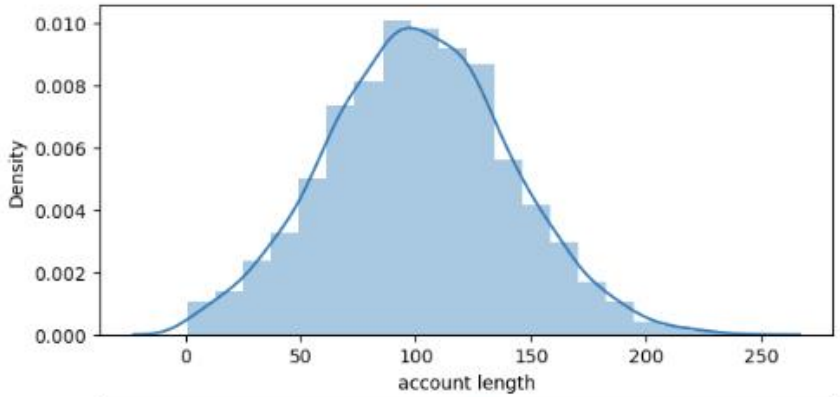
Target Variable - churn

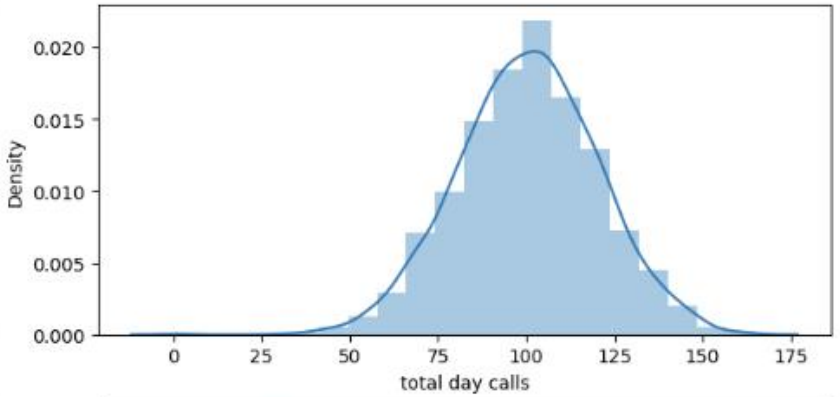
Unique identifier - phone number

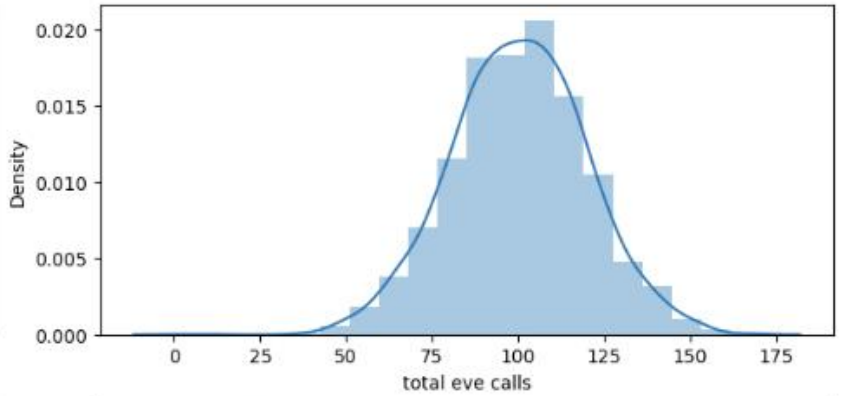
Exploratory Data Analysis - EDA

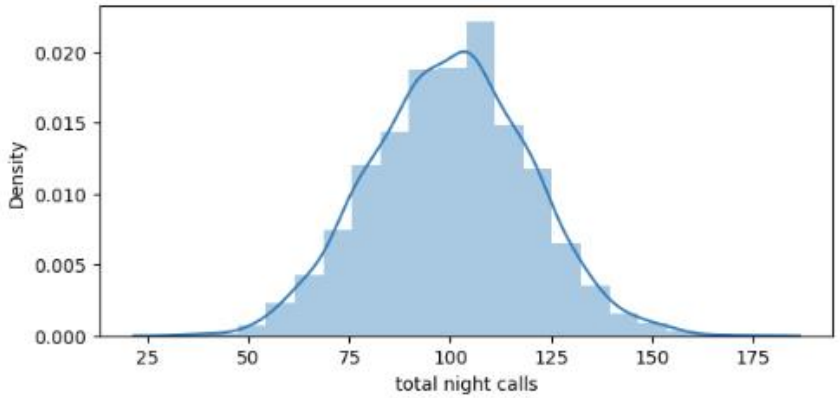
	account length	area code	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000

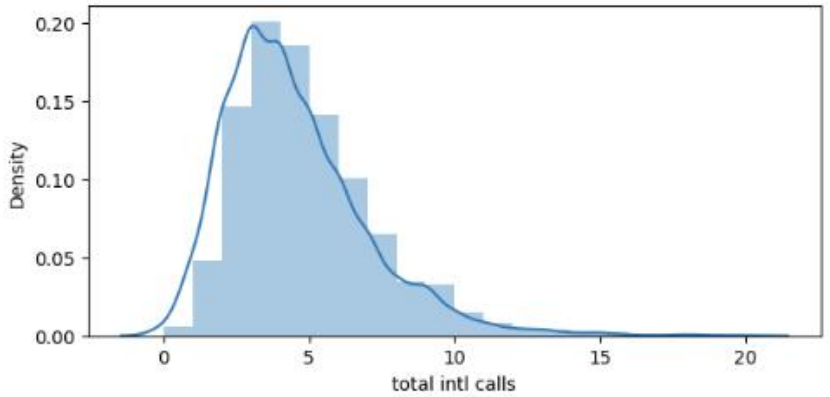
Univariate Analysis

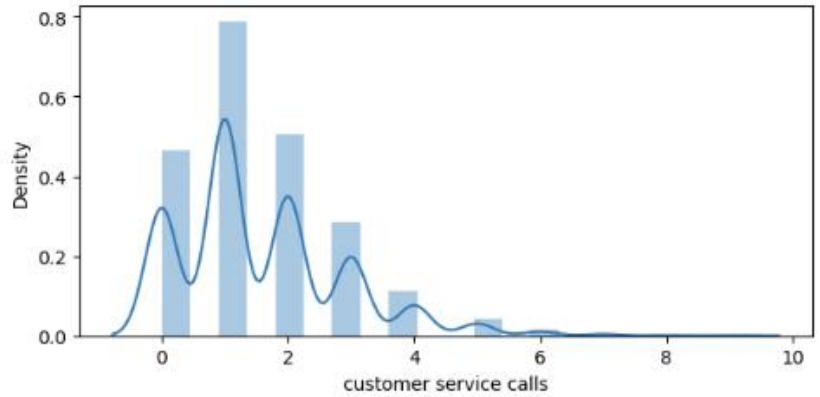










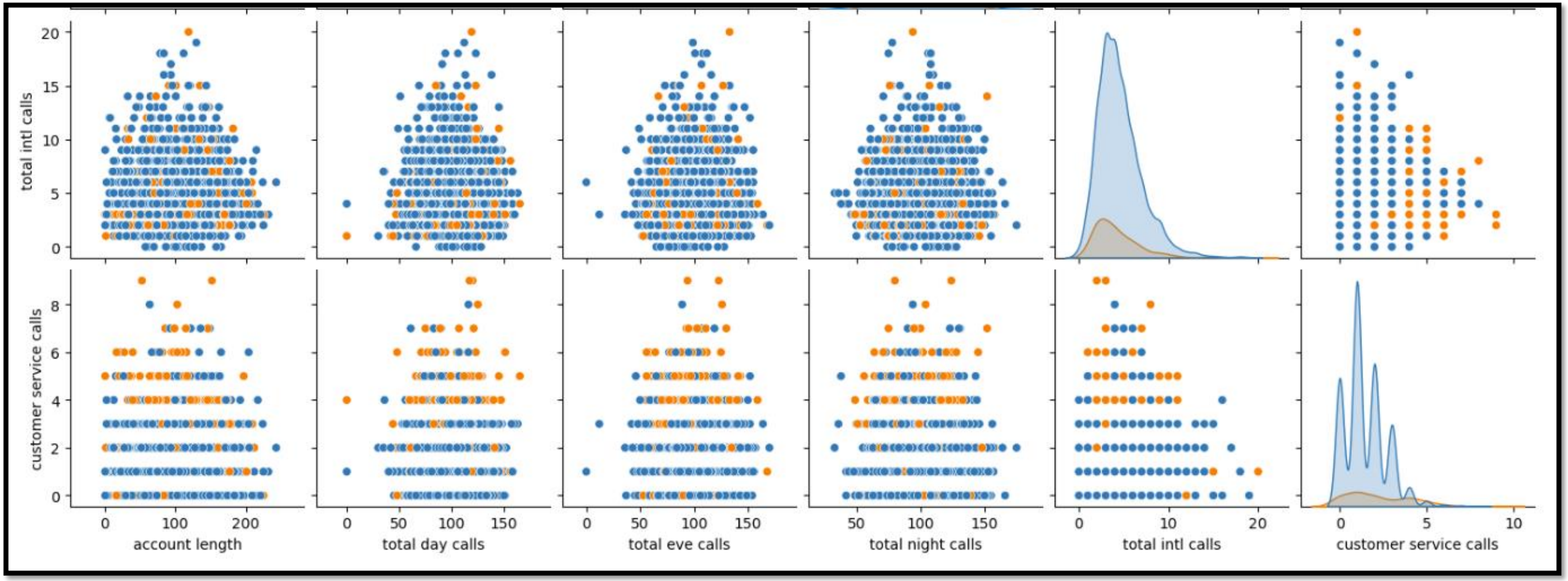


Findings

1. From the distribution plots, all the features apart from customer service calls, have a normal distribution. Although total international calls appears skewed to the right, it still maintains a normal distribution pattern.

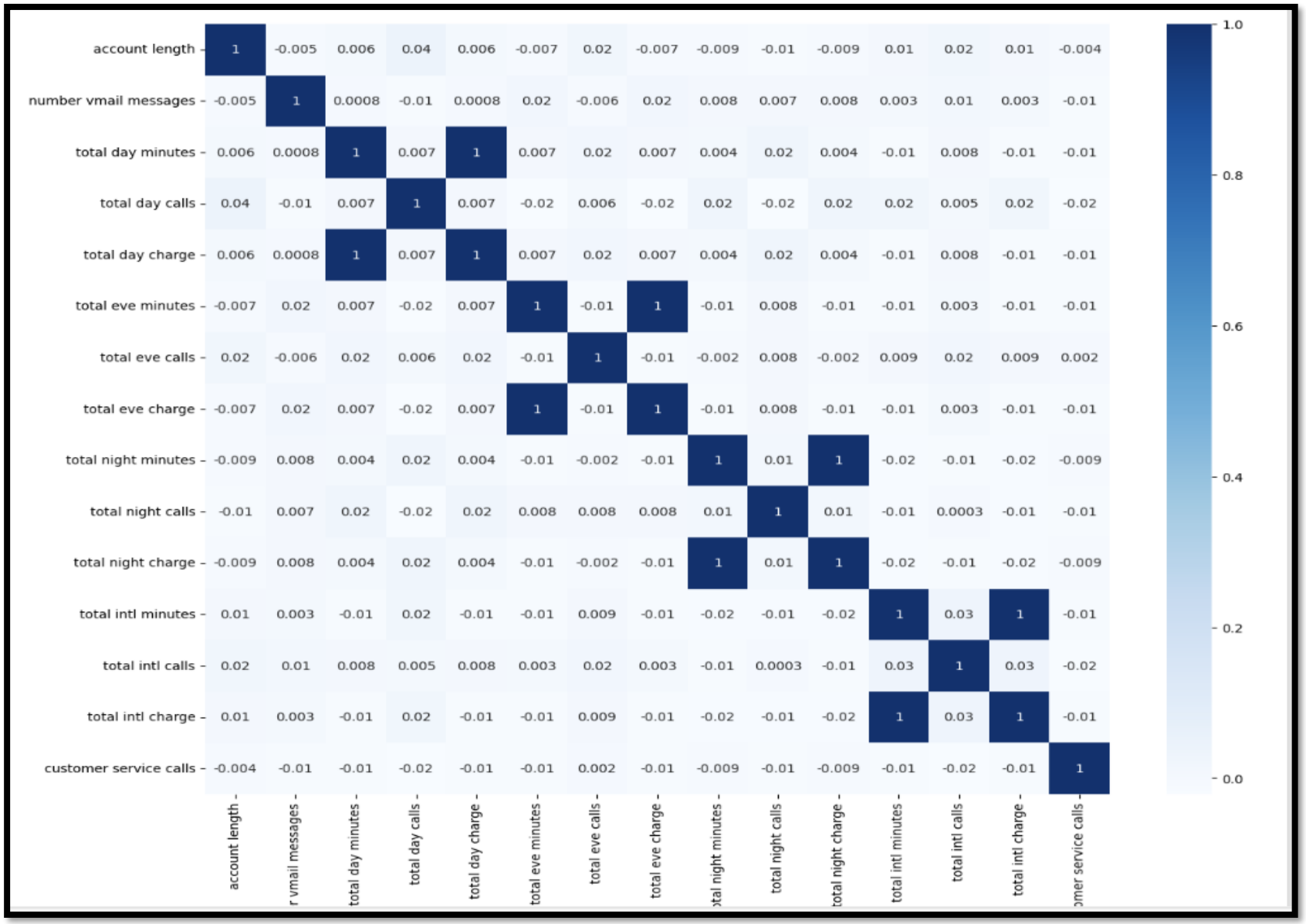
2. Customer service calls has multiple peaks, indicating several modes in the population.

Bivariate Analysis



There appears to be a clear relationship between customer service calls and true churn values. After 4 calls, customers are a lot more likely to discontinue their service.

Correlation



Findings

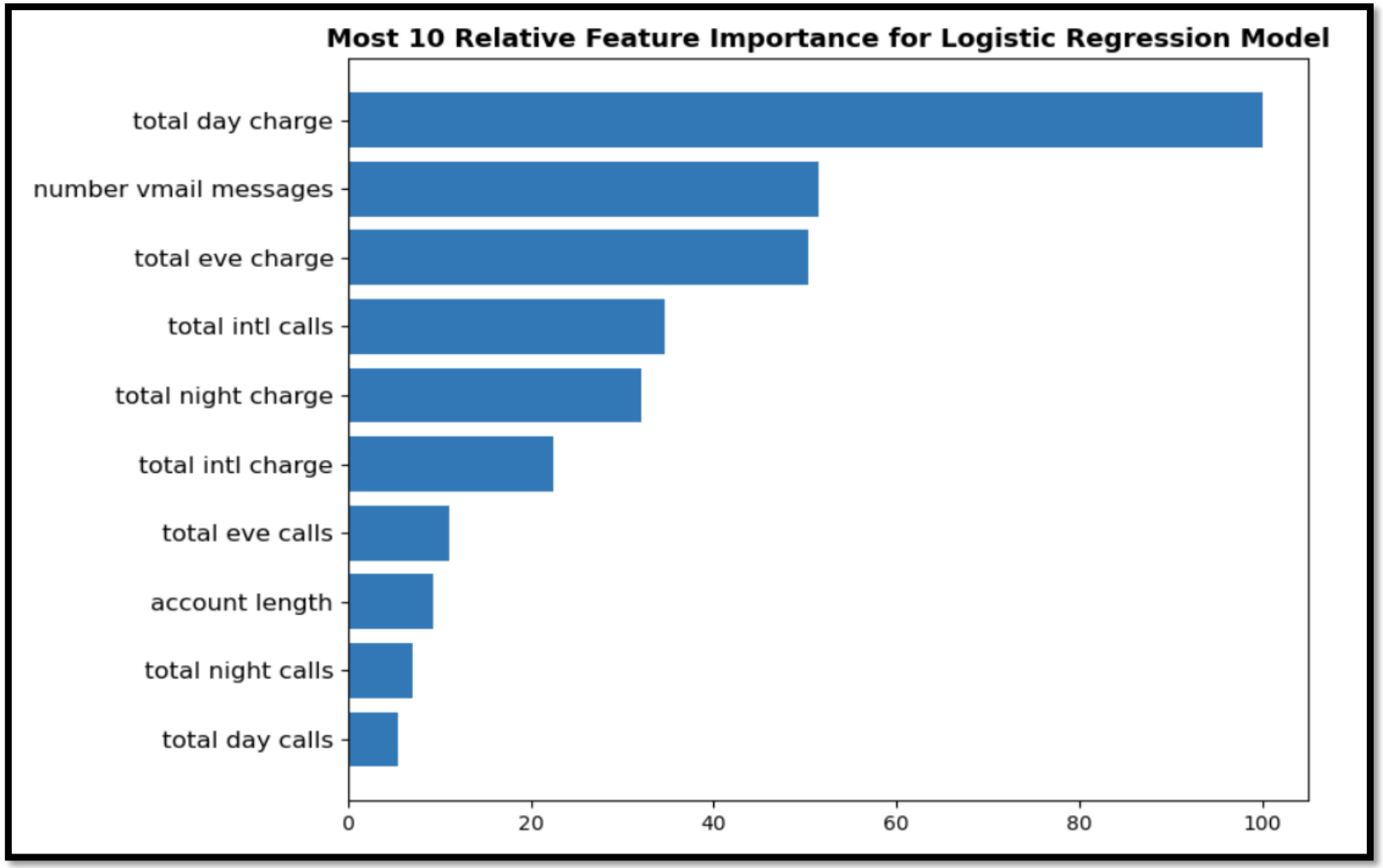
Features with a perfect positive correlation:

1. Total day charge and total day minutes.
2. Total eve charge and total eve minutes.
3. Total night charge and total night minutes.
4. Total int charge and total int minutes.

The perfect correlation of 1 indicates perfect multicollinearity.

Data Modelling

1. Logistic Regression Model



	precision	recall	f1-score	support
0	0.94	0.77	0.85	664
1	0.38	0.73	0.50	129
accuracy			0.77	793
macro avg	0.66	0.75	0.67	793
weighted avg	0.85	0.77	0.79	793

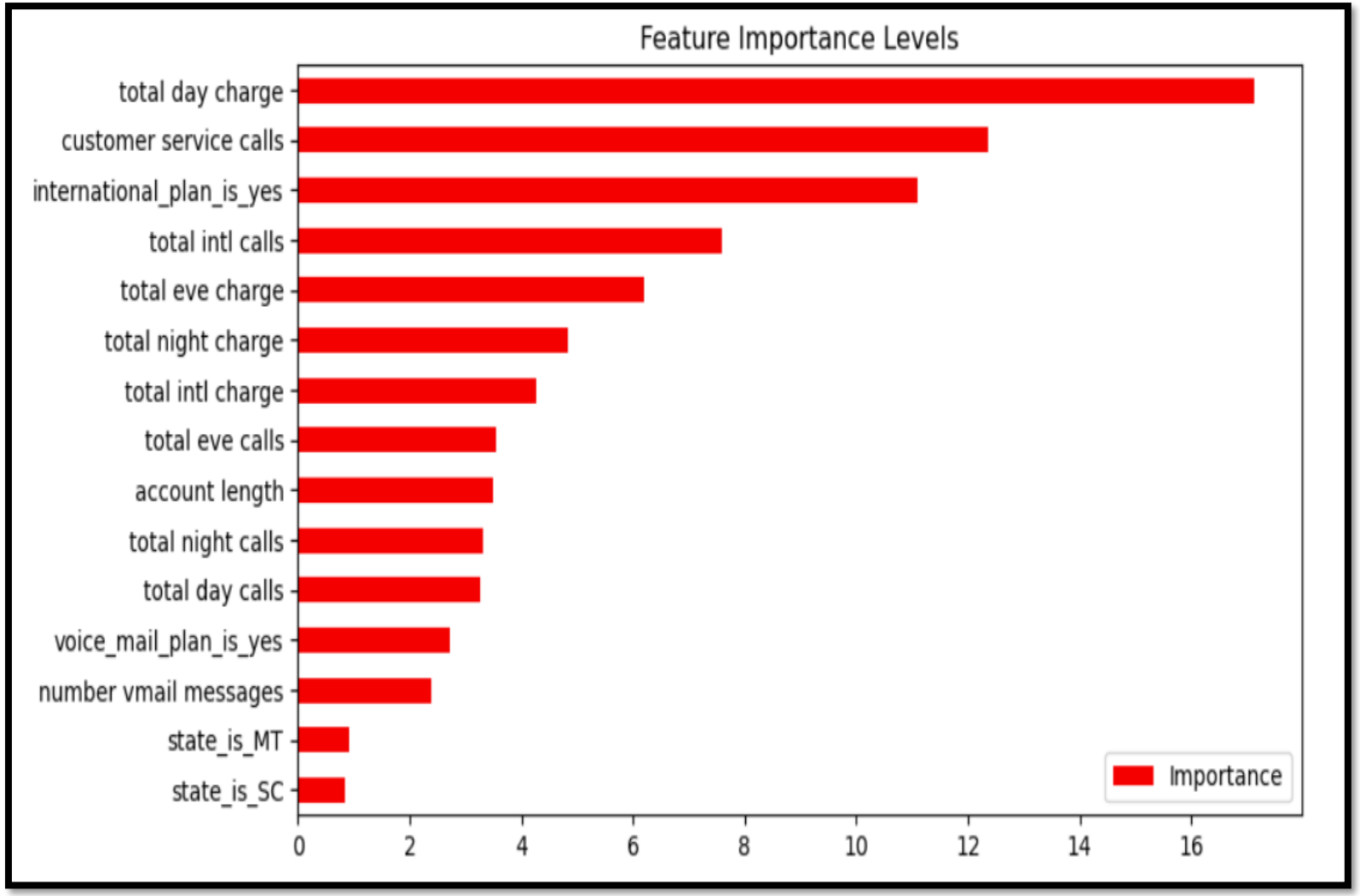
'total day charge', 'number of voicemail messages' and 'total evening charge' are the top three important features.

LOGISTIC REGRESSION MODEL RESULTS

Accuracy score for testing set:	0.76545
F1 score for testing set:	0.50267
Recall score for testing set:	0.72868
Precision score for testing set:	0.38367

Data Modelling

2. Random Forest Model



	precision	recall	f1-score	support
0	0.95	0.96	0.95	664
1	0.78	0.72	0.75	129
accuracy			0.92	793
macro avg	0.86	0.84	0.85	793
weighted avg	0.92	0.92	0.92	793

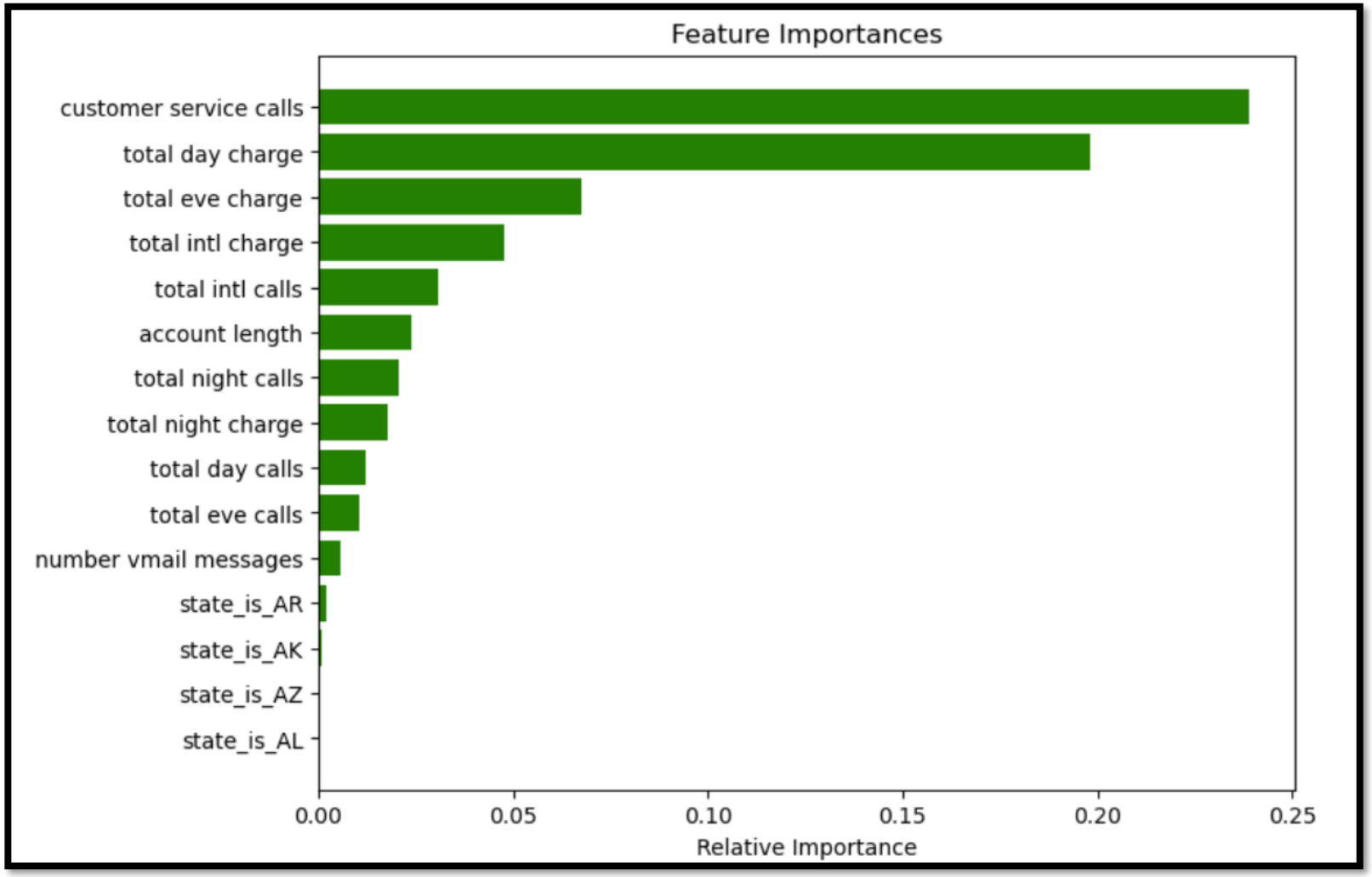
'total day charge', 'customer service calls' and 'international plan' features have the highest impact on the model.

RANDOM FOREST MODEL RESULTS

Accuracy score for testing set:	0.9218
F1 score for testing set:	0.7500
Recall score for testing set:	0.7209
Precision score for testing set:	0.7815

Data Modelling

3. Decision Tree Model



	precision	recall	f1-score	support
0	0.95	0.91	0.93	664
1	0.62	0.74	0.68	129
accuracy			0.89	793
macro avg	0.79	0.83	0.80	793
weighted avg	0.89	0.89	0.89	793

'customer service calls', 'total day charge' and 'total evening charge' are the three most important for the model..

DECISION TREE MODEL RESULTS

Accuracy score for testing set:	0.8852
F1 score for testing set:	0.6761
Recall score for testing set:	0.7364
Precision score for testing set:	0.6250

Regression Results and Conclusion

	Models	F1
1	DecisionTreeClassifier	90.111197
2	RandomForestClassifier	84.832549
0	LogisticRegression	81.077623

Looking at the results, we can see that Decision Tree Model performed well on our dataset compared to the Random Forest Model and Logistic Regression Model.

Recommendations

1. Based on the findings, it is recommended to focus on the Decision Tree Model for predicting customer churn in the telecom sector. This model has shown superior performance on the dataset.

2. While the Random Forest and the Logistic regression did not perform as well, further exploration into advanced feature engineering and threshold adjustments could potentially enhance their effectiveness.

Thank you