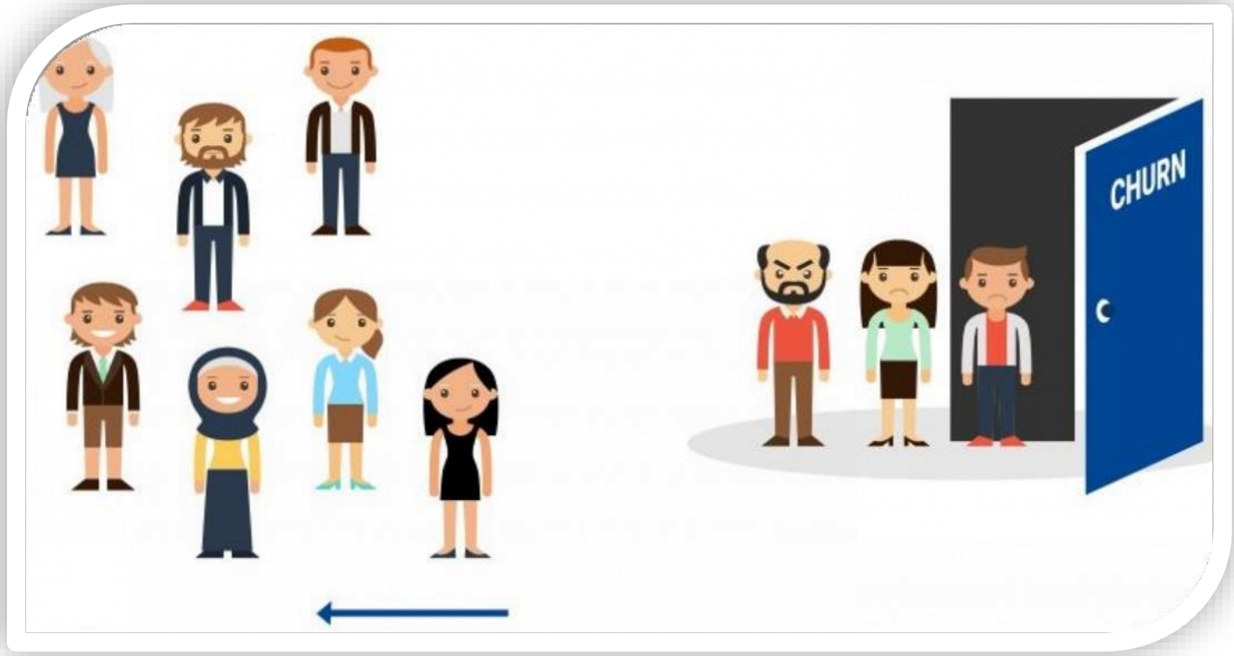
TELECOMMUNICATION CUSTOMER CHURN ANALYSIS



OBJECTIVE: Predict the reasons of losing customers by measuring customer loyalty to regain the lost customers.

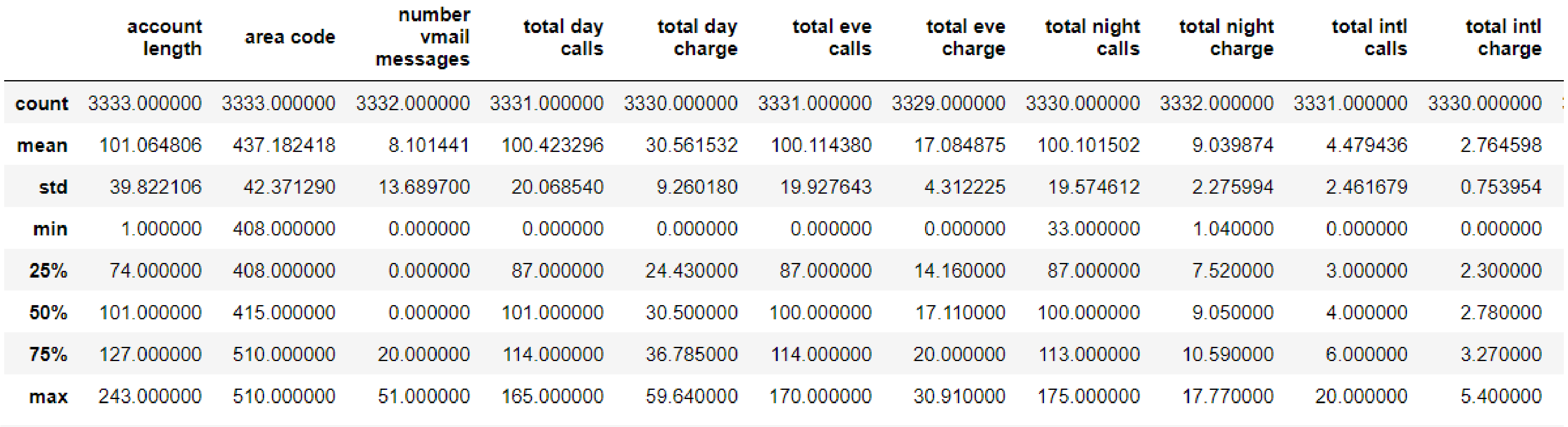
VARIABLES: Our dataset consists of more than 3000 customers. We describe below the target and predictor variables of our Case Study. There are 16 variables in all of which the Independent Variable are:

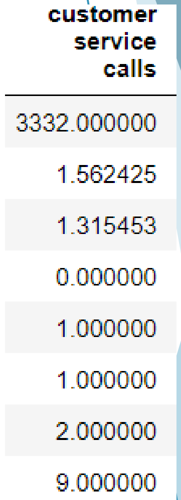
- State: The 51 state in which the customer resides, indicated by a two-letter abbreviation
- Account Length: Number of months the customer has stayed with the company
- Area Code: The three digit area code of the corresponding customer’s phone number
- Phone Number : The remaining seven- digit phone number
- International Plan : International plan activated (yes , no)
- Voice Mail Plan : Voice Mail plan activated (yes , no)
- number vmail messages : No. of voice mail messages
- Total Day Calls : The total number of calls placed during the day
- Total Day Charge : The billed cost of daytime calls
- Total Eve Calls : The total number of calls placed during the evening
- Total Eve Charge : The billed cost of evening time calls
- Total Night Calls : The total number of calls placed during the night
- Total Night Charges : The billed cost of nighttime calls
- Total Intl Calls : The total number of international calls
- Total Intl Charge : The billed cost for international calls
- Customer Service Calls : Number of calls to customer service made

DATA OVERVIEW AND DATA MANIPULATION

Summary Statistics

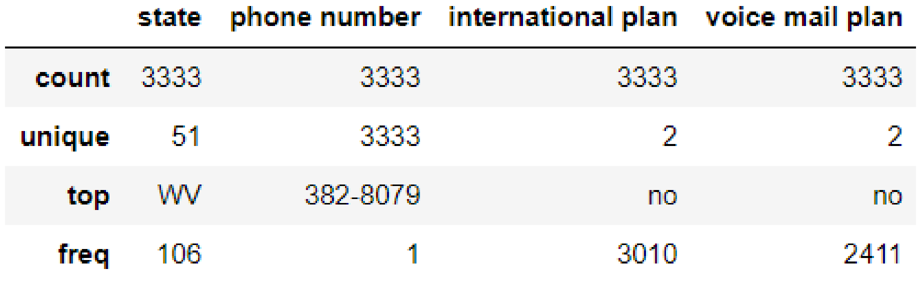
> For Numerical Variables





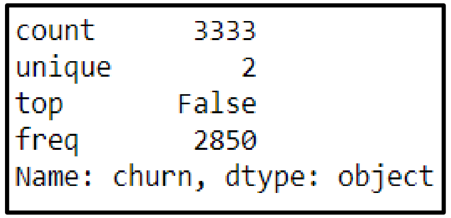
	account length	area code	number vmail messages	total day calls	total day charge	total eve calls	total eve charge	total night calls	total night charge	total intl calls	total intl charge	customer service calls
count	3333.000000	3333.000000	3332.000000	3331.000000	3330.000000	3331.000000	3329.000000	3330.000000	3332.000000	3331.000000	3330.000000	3332.000000
mean	101.064806	437.182418	8.101441	100.423296	30.561532	100.114380	17.084875	100.101502	9.039874	4.479436	2.764598	1.562425
std	39.822106	42.371290	13.689700	20.068540	9.260180	19.927643	4.312225	19.574612	2.275994	2.461679	0.753954	1.315453
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	33.000000	1.040000	0.000000	0.000000	0.000000
25%	74.000000	408.000000	0.000000	87.000000	24.430000	87.000000	14.160000	87.000000	7.520000	3.000000	2.300000	1.000000
50%	101.000000	415.000000	0.000000	101.000000	30.500000	100.000000	17.110000	100.000000	9.050000	4.000000	2.780000	1.000000
75%	127.000000	510.000000	20.000000	114.000000	36.785000	114.000000	20.000000	113.000000	10.590000	6.000000	3.270000	2.000000
max	243.000000	510.000000	51.000000	165.000000	59.640000	170.000000	30.910000	175.000000	17.770000	20.000000	5.400000	9.000000

> For Categorical variables



	state	phone number	international plan	voice mail plan
count	3333	3333	3333	3333
unique	51	3333	2	2
top	WV	382-8079	no	no
freq	106	1	3010	2411

churn



```
count       3333
unique         2
top        False
freq        2850
Name: churn, dtype: object
```

Data Manipulation

The variables, ‘Area Code’, ‘International Plan’, ‘Voice Mail Plan’ and ‘Account length_group’ are dummy variables. The outcome variable of our study, a binary variable, is ‘Churn’ (Whether the customer left the service).

First, area code was converted into object type.

Second, converted churn into integer type and replaced the following values:

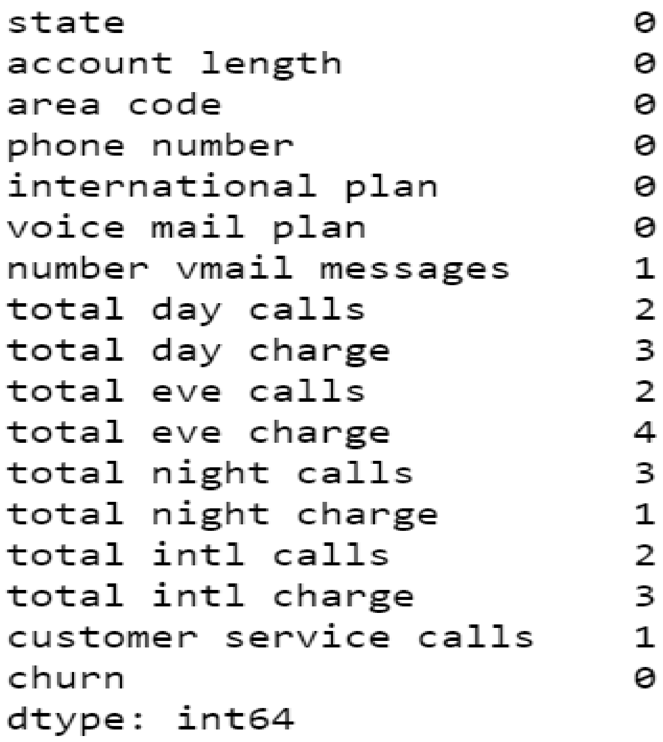
‘True’ with 1 and ‘False’ with 0 for the column ‘Churn’.

Third, conversion of ‘account length’ to categorical column ‘account length_group’ in which the data has been grouped into 6 sub-categories:

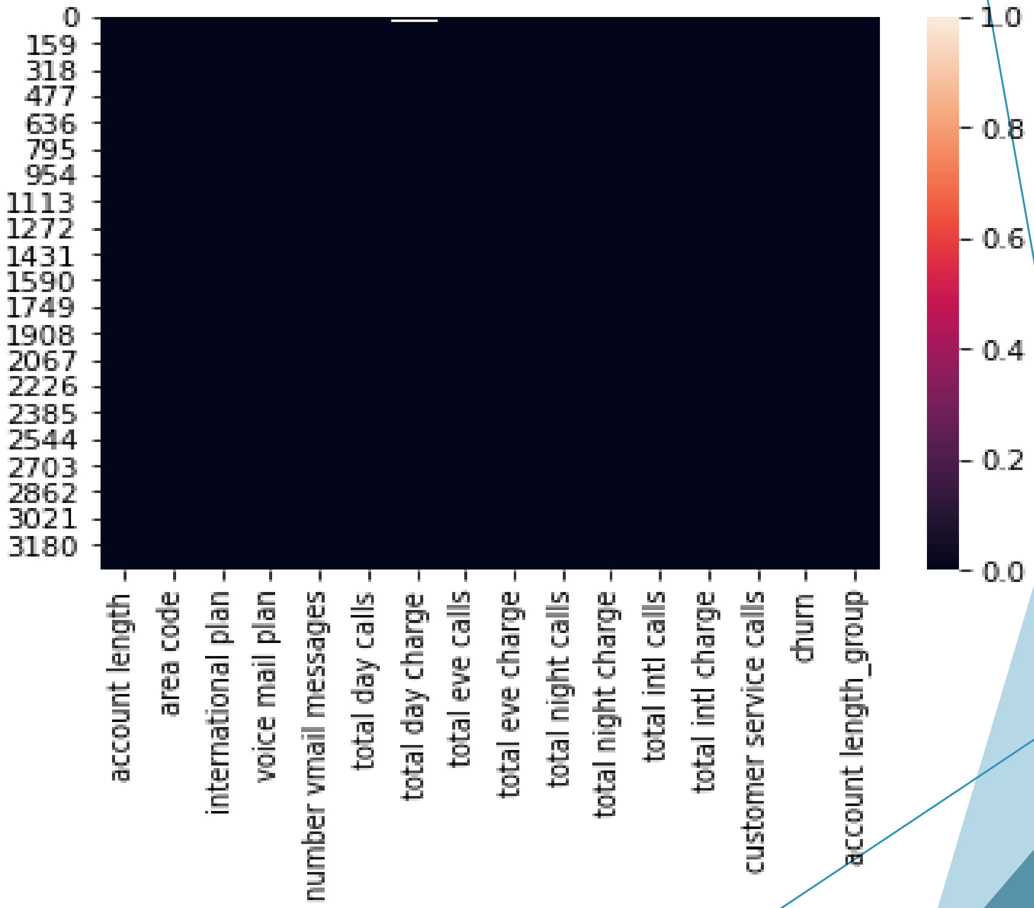
Tenure_0-12 for the tenure period ranging from 0 to 12 months and likewise, Tenure_13-60 for the tenure period ranging from 13 to 60 months,
Tenure_61-120 for the tenure period ranging from 61 to 120 months ,
Tenure_121-180 for the tenure period ranging from 121 to 180 months ,
Tenure_181-240 for the tenure period ranging from 181 to 240 months and Tenure_gt_240 for more than 240 months.

Missing values imputation

We observed that there were few missing values in the data. So we imputed their values by applying ‘b-fill’.



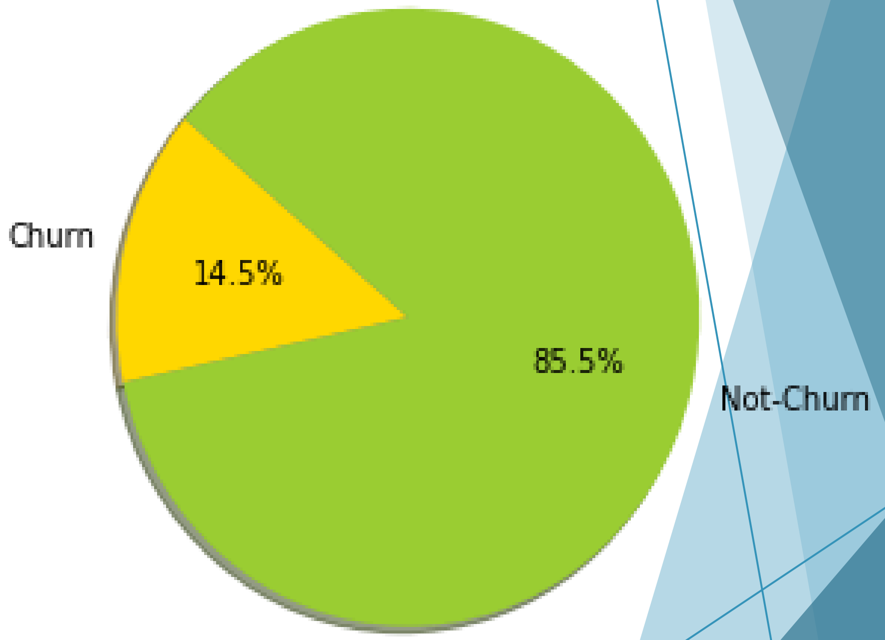
```
state                     0
account length            0
area code                 0
phone number              0
international plan        0
voice mail plan           0
number vmail messages     1
total day calls           2
total day charge          3
total eve calls           2
total eve charge          4
total night calls         3
total night charge        1
total intl calls          2
total intl charge         3
customer service calls    1
churn                     0
dtype: int64
```



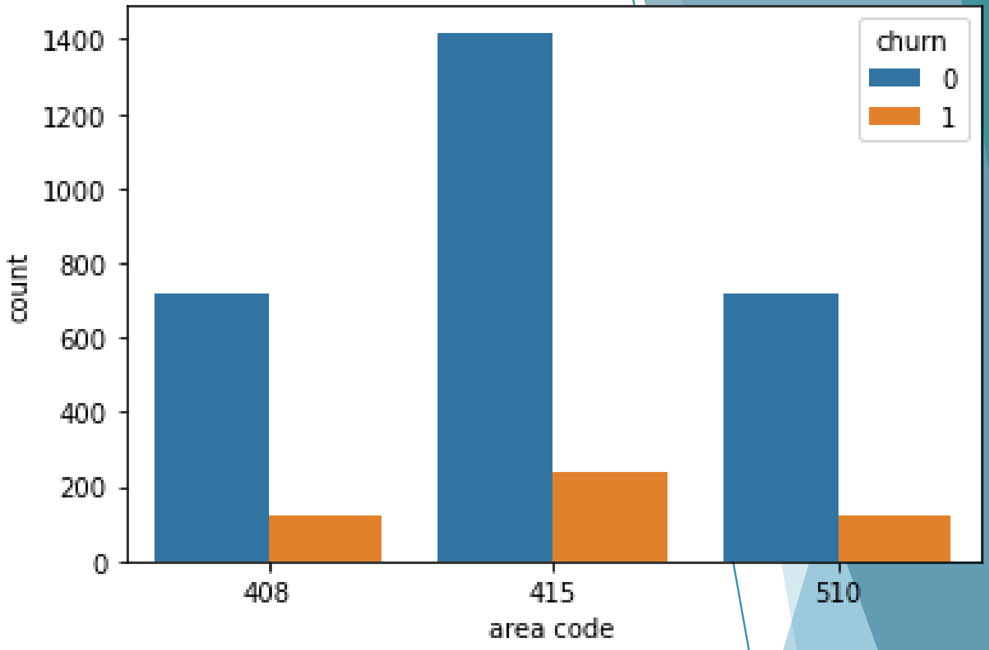
EXPLORATORY DATA ANALYSIS AND INTERPRETATIONS

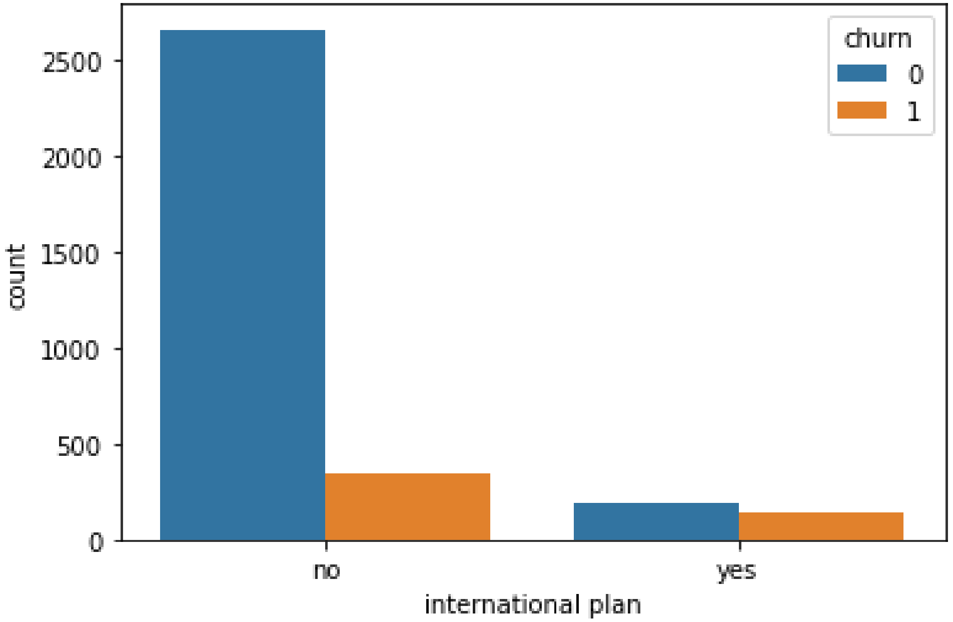
Univariate Analysis

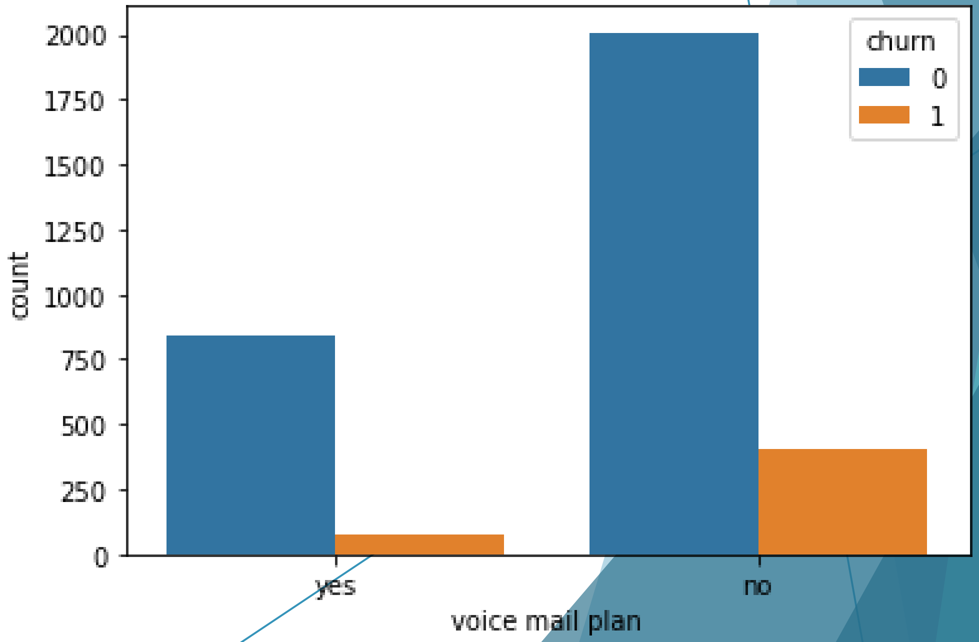
- Customer churn is a metric that measures the number of subscribers who leave. A low churn rate is ideal. Companies that experience a high churn rate are under more pressure to generate revenue from other areas or gain new customers.
- From the pie chart, we observe that the churn rate for the company is 14.5%. We infer that customers are gradually migrating to competition.



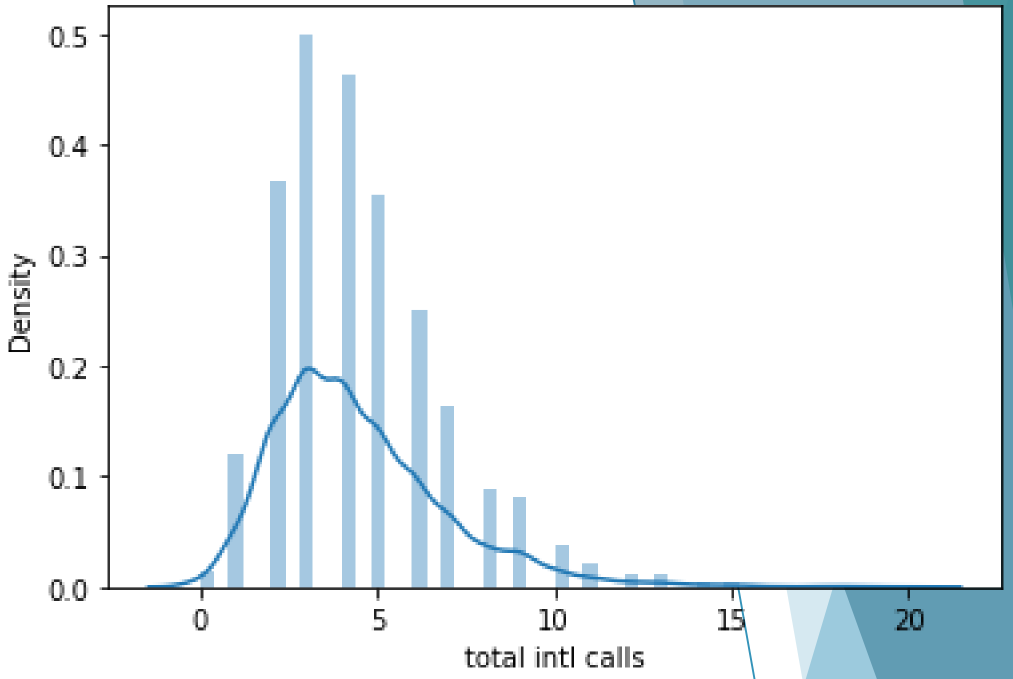
- When International Plan is enabled, the churn rate is much higher; the usage of the international plan by the customer is a strong feature. We do not observe the same effect with Voice mail plan.
- Also, Area code 415 has the maximum churned and non-churned customers.

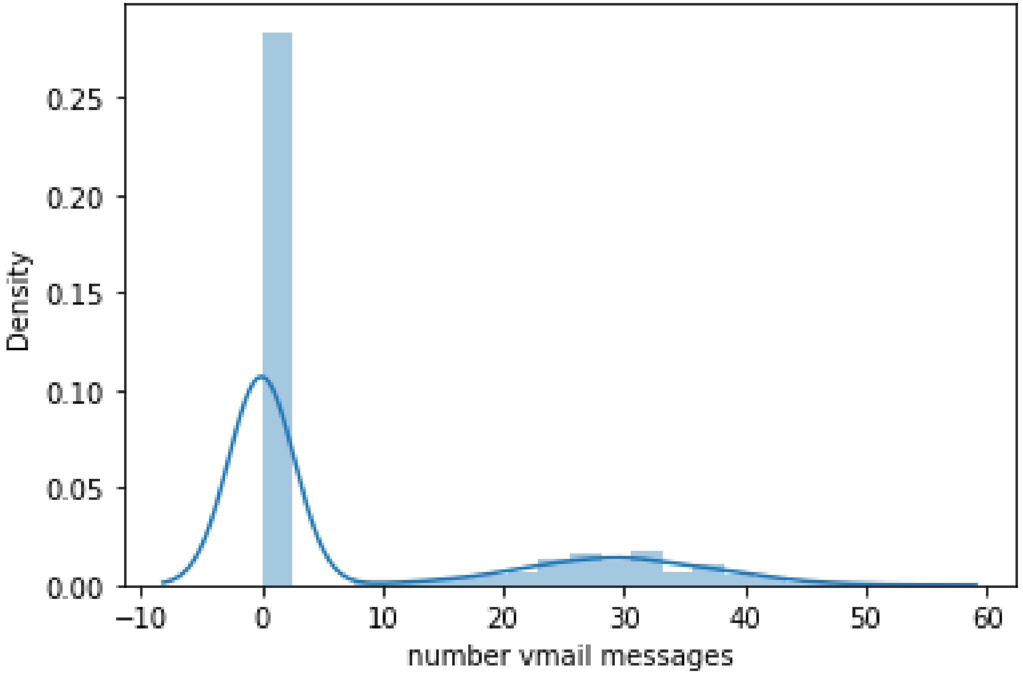


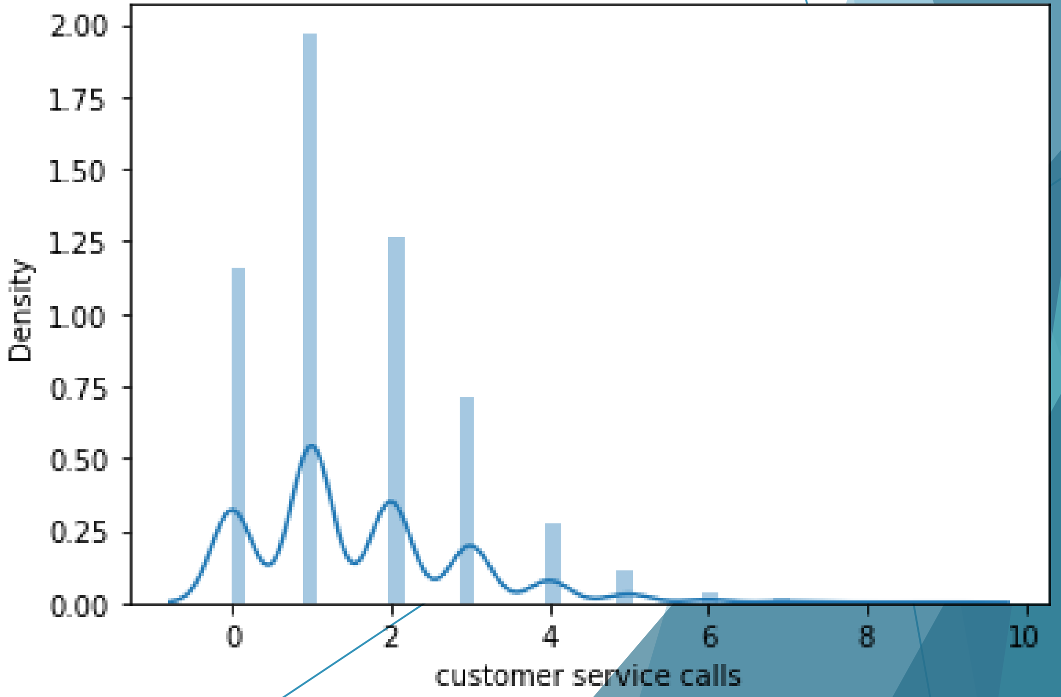




- After plotting the Density plot and Pair plot, we came to the conclusion that all the variables are normally distributed, except ‘Total international calls’, ‘Number vmail messages’ and ‘Customer service calls. They all are **right–skewed.**

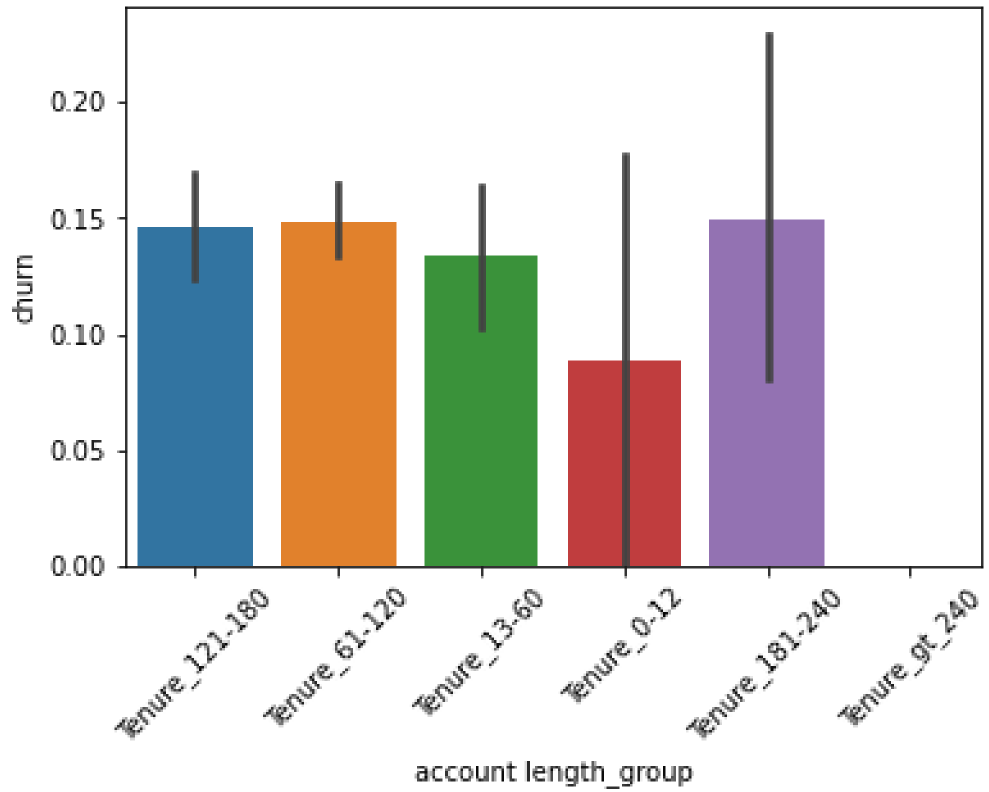


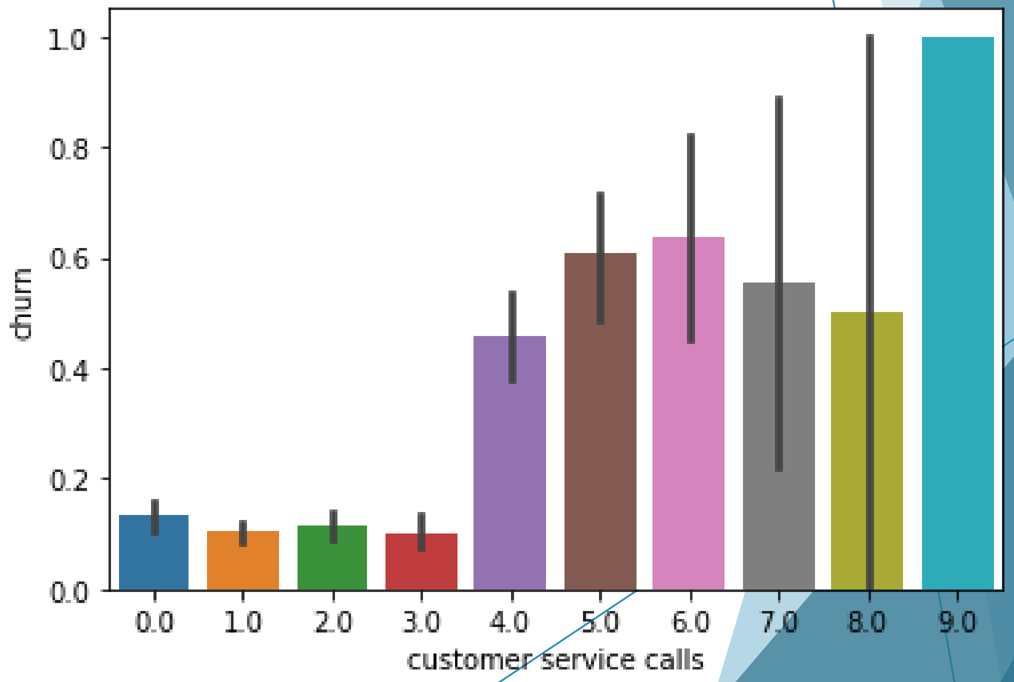




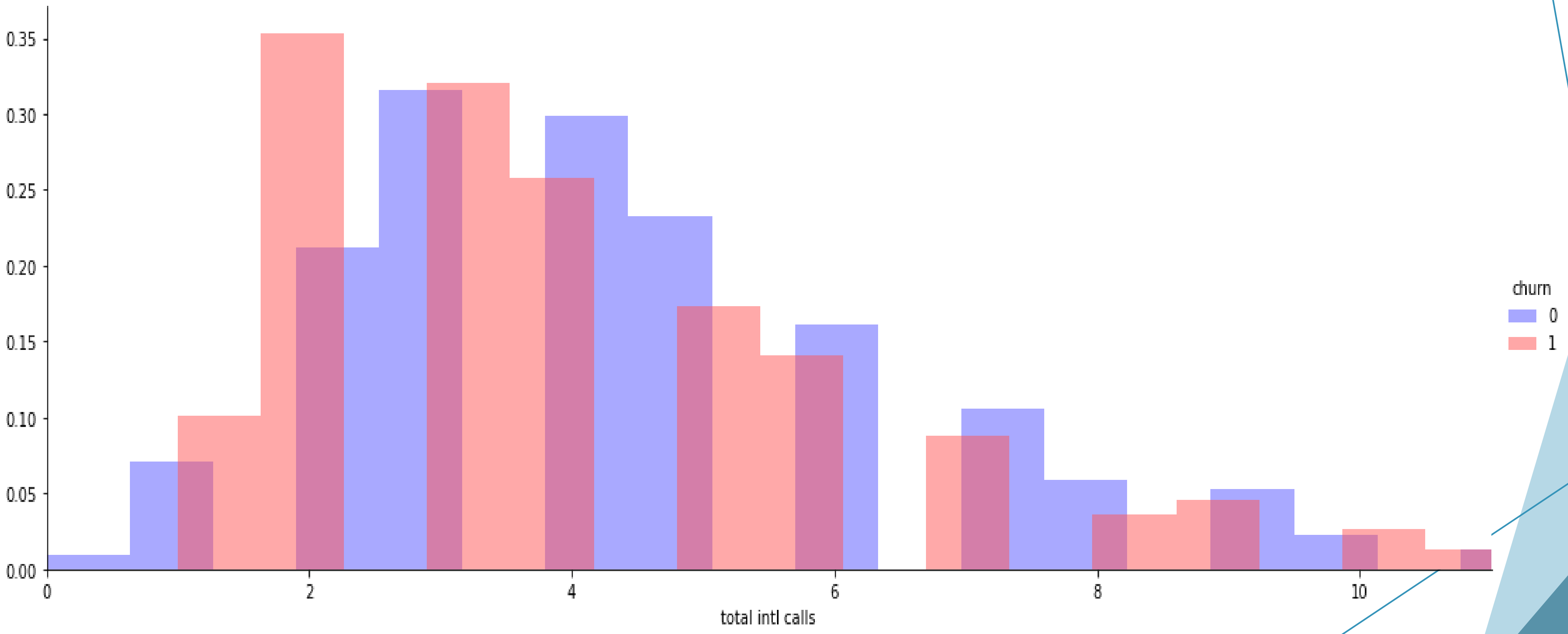
Bivariate Analysis

- Comparing account length_group and churn, we observe that churn is lowest for members with tenure 0–12, and increases as the tenure period increase.
- Churn rate increases after 4 or more calls to the customer services.



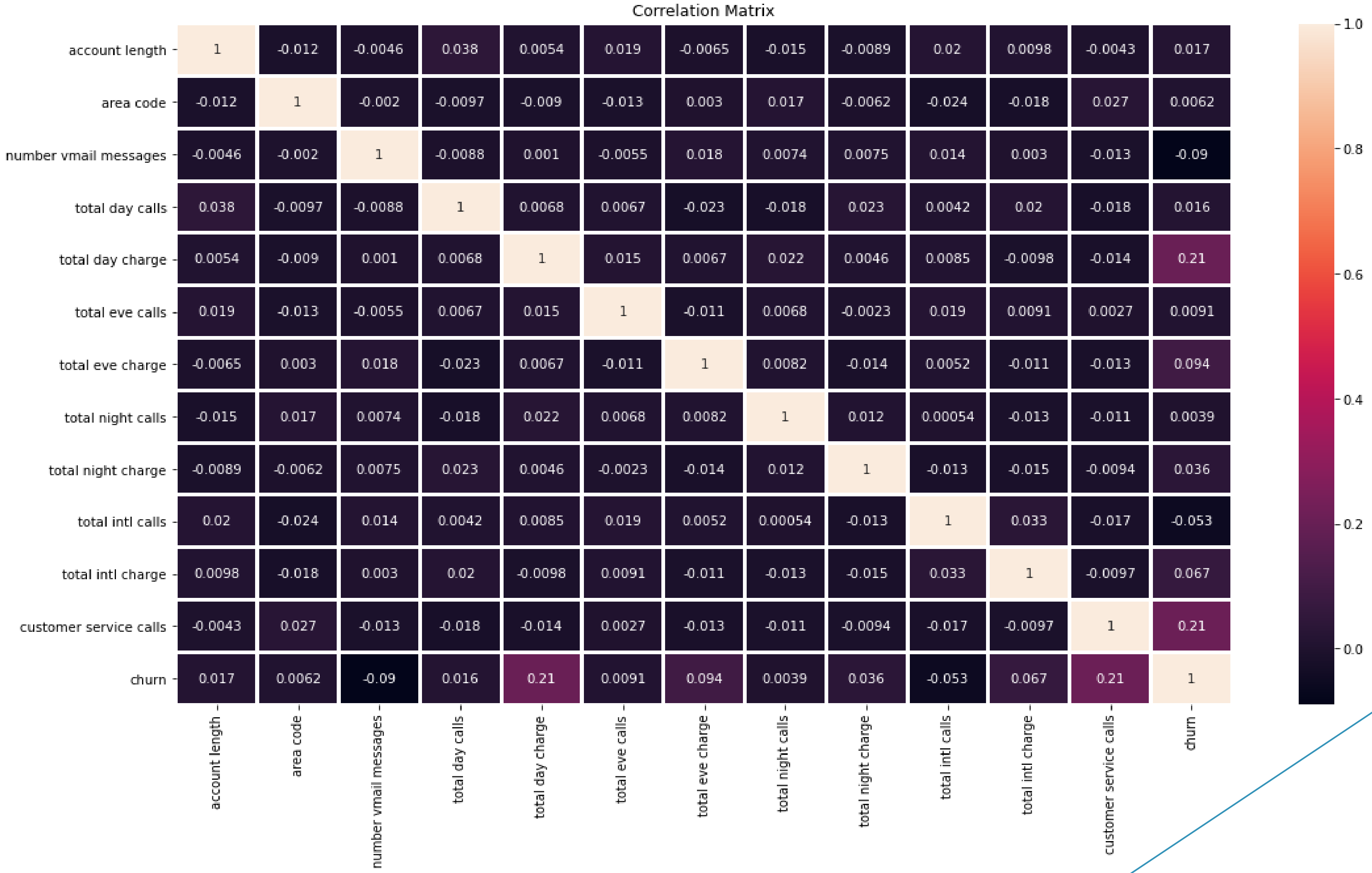


The total international calls data with churn shows that the churn rate was high when the total international calls were less and it gradually reduced with more and more international calls.



Correlation Matrix

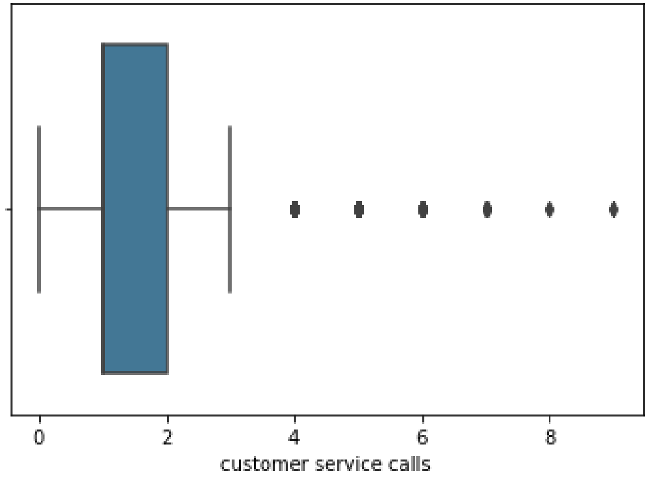
The heatmap below shows the correlation between the variables. We observe that there is very less correlation among the variables. The maximum correlation value is 0.21 that is between 'total day charge' and 'churn' and 'customer service calls' and 'churn'.

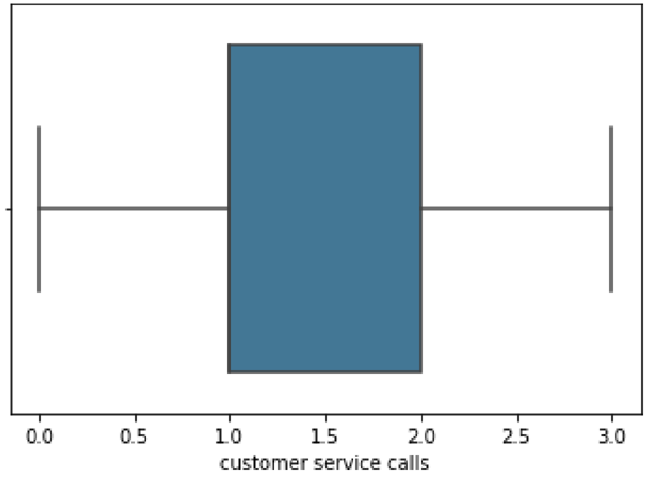


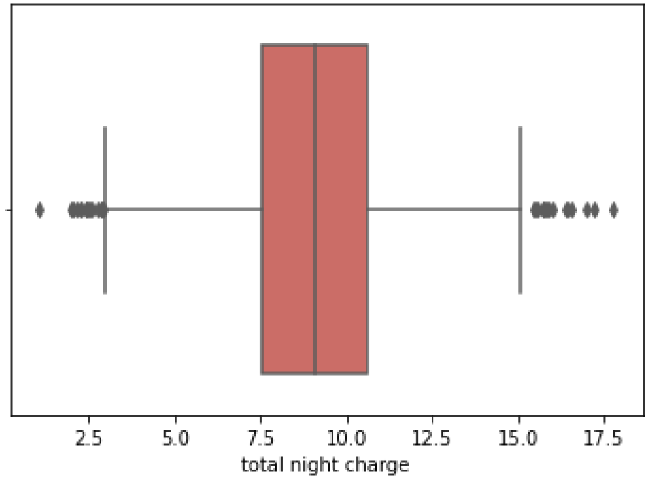
	account length	area code	number vmail messages	total day calls	total day charge	total eve calls	total eve charge	total night calls	total night charge	total intl calls	total intl charge	customer service calls	churn
account length	1	-0.012	-0.0046	0.038	0.0054	0.019	-0.0065	-0.015	-0.0089	0.02	0.0098	-0.0043	0.017
area code	-0.012	1	-0.002	-0.0097	-0.009	-0.013	0.003	0.017	-0.0062	-0.024	-0.018	0.027	0.0062
number vmail messages	-0.0046	-0.002	1	-0.0088	0.001	-0.0055	0.018	0.0074	0.0075	0.014	0.003	-0.013	-0.09
total day calls	0.038	-0.0097	-0.0088	1	0.0068	0.0067	-0.023	-0.018	0.023	0.0042	0.02	-0.018	0.016
total day charge	0.0054	-0.009	0.001	0.0068	1	0.015	0.0067	0.022	0.0046	0.0085	-0.0098	-0.014	0.21
total eve calls	0.019	-0.013	-0.0055	0.0067	0.015	1	-0.011	0.0068	-0.0023	0.019	0.0091	0.0027	0.0091
total eve charge	-0.0065	0.003	0.018	-0.023	0.0067	-0.011	1	0.0082	-0.014	0.0052	-0.011	-0.013	0.094
total night calls	-0.015	0.017	0.0074	-0.018	0.022	0.0068	0.0082	1	0.012	0.00054	-0.013	-0.011	0.0039
total night charge	-0.0089	-0.0062	0.0075	0.023	0.0046	-0.0023	-0.014	0.012	1	-0.013	-0.015	-0.0094	0.036
total intl calls	0.02	-0.024	0.014	0.0042	0.0085	0.019	0.0052	0.00054	-0.013	1	0.033	-0.017	-0.053
total intl charge	0.0098	-0.018	0.003	0.02	-0.0098	0.0091	-0.011	-0.013	-0.015	0.033	1	-0.0097	0.067
customer service calls	-0.0043	0.027	-0.013	-0.018	-0.014	0.0027	-0.013	-0.011	-0.0094	-0.017	-0.0097	1	0.21
churn	0.017	0.0062	-0.09	0.016	0.21	0.0091	0.094	0.0039	0.036	-0.053	0.067	0.21	1

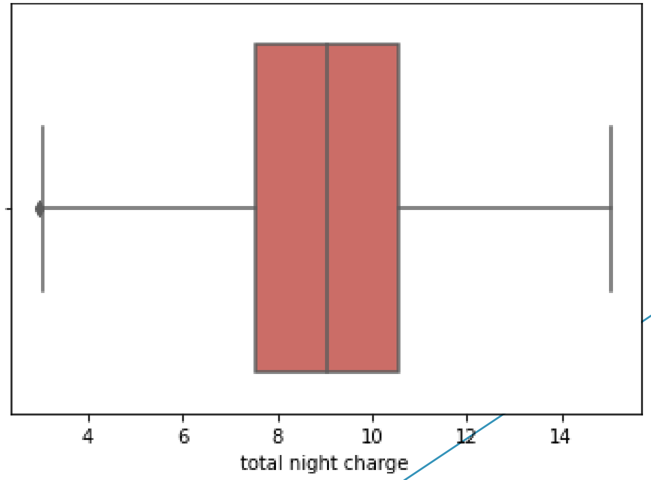
Outliers detection and treatment

There were some outliers present in all the variables. To treat those outliers we replaced them with the median value of those variables. For example:









Supervised Classification Algorithms

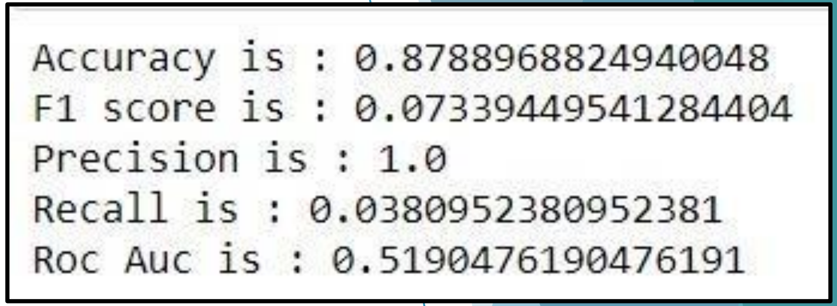
LOGISTIC REGRESSION

Logistic regression is a fundamental classification technique. We run logistic regression instead of linear regression as our target variable, "Churn" is categorical variable. Logistic regression is fast and relatively uncomplicated, and it's convenient to interpret

After running the model, following are the results:

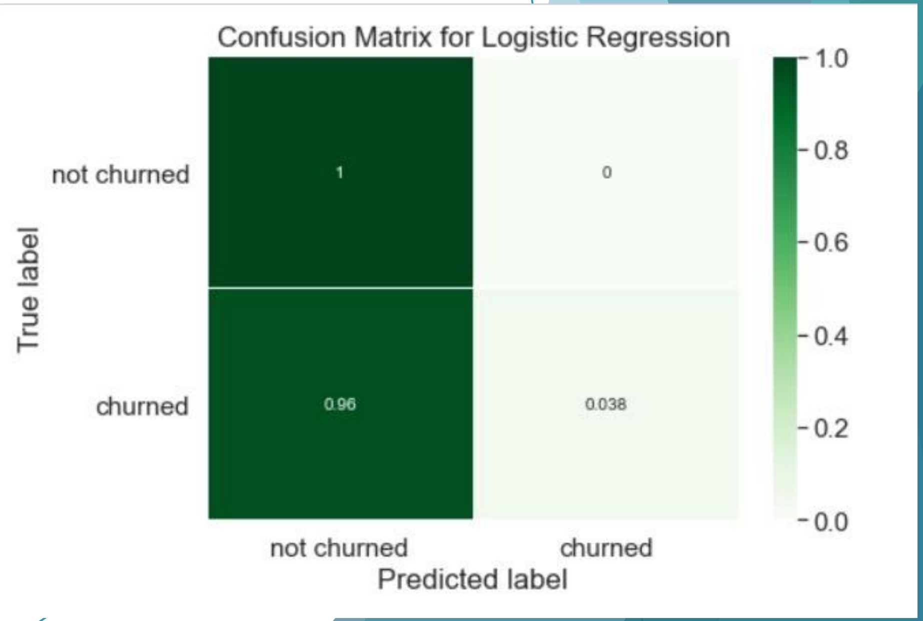
1. Area under Curve is a perfect performance metric for ROC curve, also referred as Accuracy Index. We got AUC as 51.90 % (Higher the AUC, better the prediction power of the model)

2. Accuracy of the model is 87.88%



```
Accuracy is : 0.8788968824940048
F1 score is : 0.07339449541284404
Precision is : 1.0
Recall is : 0.0380952380952381
Roc Auc is : 0.5190476190476191
```

Classification Report for Logistic Regression



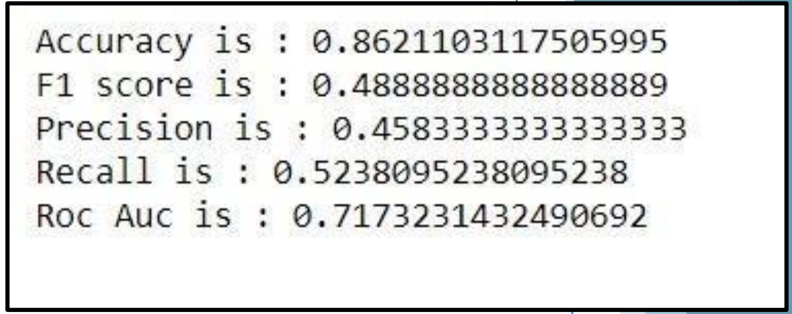
DECISION TREE CLASSIFIER

Decision trees are a popular and powerful tool used for classification and prediction purposes. It works for both continuous and categorical input and output variables.

Results:

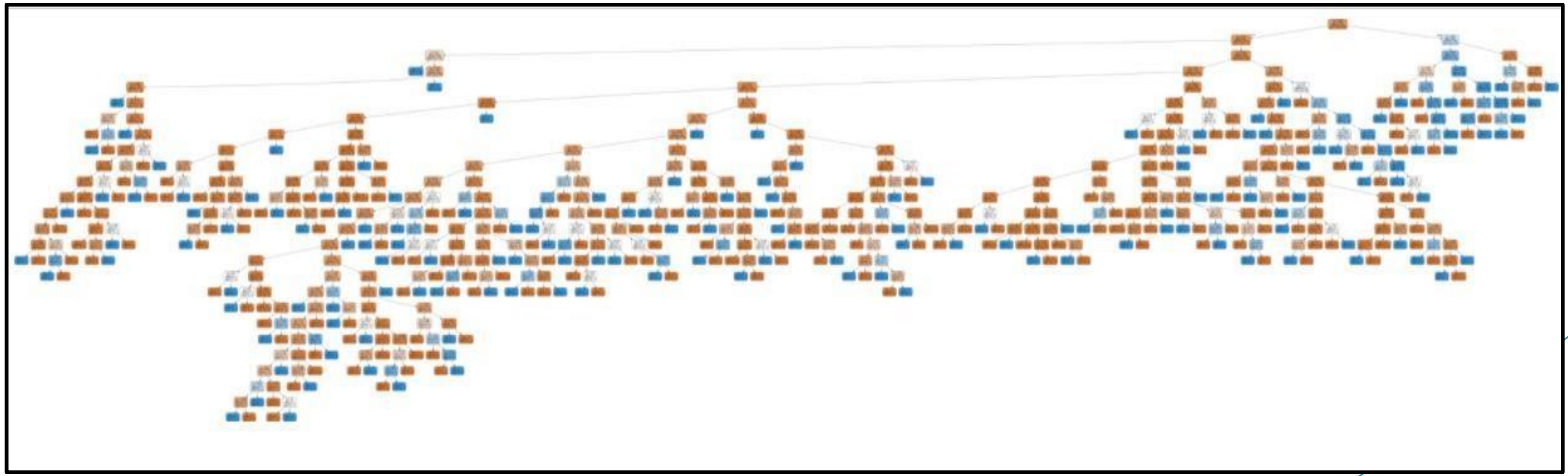
1. Accuracy of the model is 86.21%

2. Area under Curve is 71.73%



```
Accuracy is : 0.8621103117505995
F1 score is : 0.4888888888888889
Precision is : 0.4583333333333333
Recall is : 0.5238095238095238
Roc Auc is : 0.7173231432490692
```

Classification report for Decision Trees



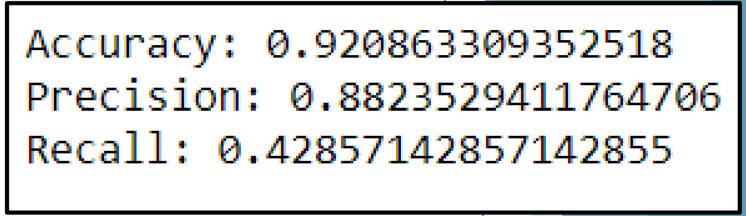
RANDOM FOREST CLASSIFIER

A random forest takes random samples, forms many decision trees, and then takes the average of those decisions to form a more refined model. It emphasizes on feature selection and does not assume that the model has a linear relationship — like regression models do.

Results:

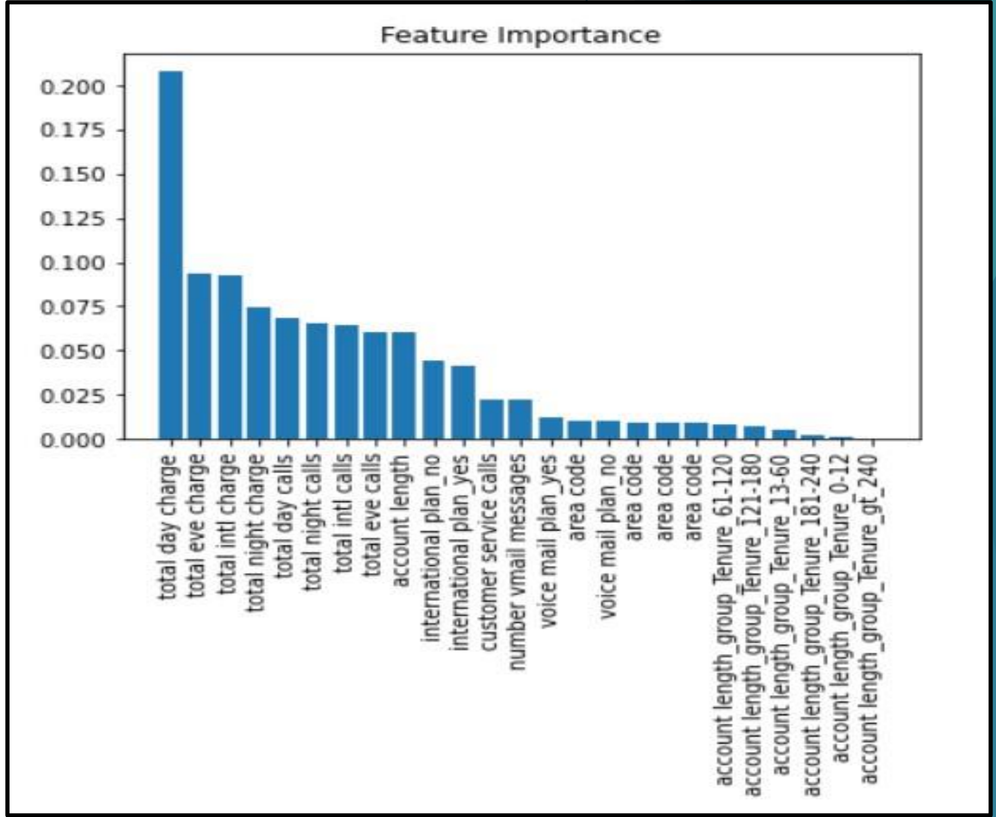
1. Accuracy of the model is 92.08%
2. Area under Curve is 73.59%.

According to Feature Importance plot, total day charge is the highly significant variable followed by Total eve charge, total international charge, and tenure is the least significant variable in determining the customer churn rate.



```
Accuracy: 0.920863309352518
Precision: 0.8823529411764706
Recall: 0.4285714285714285S
```

Accuracy, model precision, recall of Random Forest



Hyperparameter Tuning techniques

Hyperparameter tuning refers to the shaping of the model architecture from the available space. This, in simple words, is nothing but searching for the right hyperparameter to find high precision and accuracy.

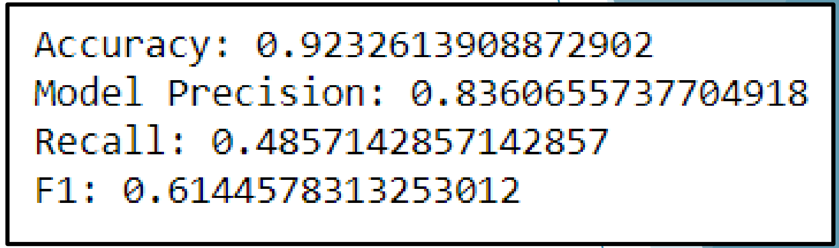
1. Grid Search

Grid search is a technique which tends to find the right set of hyperparameters for the particular model.

2. Random Search

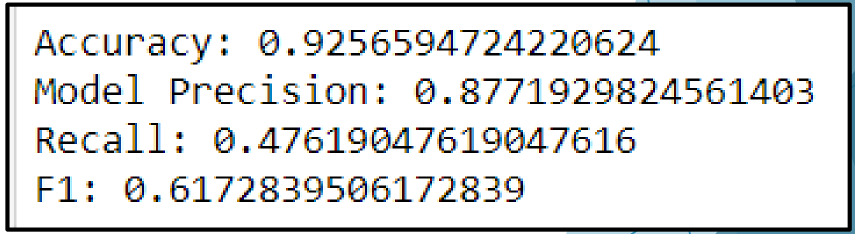
Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model.

Comparing both, we can see that Random search yields relatively better results due to high precision and accuracy.



```
Accuracy: 0.9232613908872902
Model Precision: 0.8360655737704918
Recall: 0.4857142857142857
F1: 0.6144578313253012
```

Accuracy, model precision, recall and F1 of Grid search



```
Accuracy: 0.9256594724220624
Model Precision: 0.8771929824561403
Recall: 0.47619047619047616
F1: 0.6172839506172839
```

Accuracy, model precision, recall and F1 of Random search

MODEL SELECTION

Models	Accuracy	F1 score
Logistic Regression	0.878896	0.073394
Decision Tree	0.862110	0.48888
Random Forest	0.92565	0.61728

Results:

Accuracy- Random forest produced the best accuracy rate at 92.52.

F1 score- Out of all the models, Random Forest is clearly winning with 61.72.

Since, the key motive is to regain the customers and increase profitability to further expand their business, F1 score is important.

Also, due to uneven churn data, F1 score holds more weightage in model selection and is far more useful than accuracy.

Therefore, **Random forest** is probably the best choice.

STRATEGIES TO REDUCE CHURN RATE

- **Increasing Customer tenure:** A telecom company can ensure that customers remain loyal for a longer tenure by increasing the switching costs for them. This can be done by providing bundled services like providing handset with a locked SIM, Broadband internet along with OTT services. They can also incentivize their customers by offering them annual loyalty points that can be redeemed for a discount on OTT streaming services. A telecom company can engage its customers in a contract of two years or more as longer term contracts significantly reduce the churning rate.

- **Economizing daily charges:** A telecom company's customers who churn are most likely to have total day charges greater than $45. We observed that 21% of the churn rate is due to total day charges. The probable reasons are low quality streaming services, poor network quality and connectivity. It should introduce family plans in order to alleviate the high churn rate.

- **Customer Service Issues:** Regardless of industry, customer service has always been an internal driver that affects your customer satisfaction and this is highly significant in the telecom industry where communication services are a necessity. In our model, 21% customer churn is because of customer service.

From issues reaching the call centers to the actual resolution of the customer's problems, keeping your customer service issues to a minimum is key.

This can be done by firstly tracking KPIs that include: Turn-around time for issue resolution, Average reply time, The time to first response, Ticket backlog and First contact resolution rate.

All these are long-tail KPIs that affect churn due to customer service issues and should be constantly monitored by your team.