

Project Customer Churn

Garima

15 December 2018

Contents

1 Introduction

1.1 Problem Statement	2
1.2 Data	2-4

2 Methodology

2.1 Missing Value Analysis	5
2.2 Data Visualisation	5-10
2.3 Outlier Analysis	11-14
2.4 Feature Selection	15-16
2.5 Modeling	17-19

3 Conclusion

3.1 Model Evaluation & Model Selection.	20
---	----

References.	21
---------------------	----

Chapter 1

Introduction

1.1 Problem Statement:

Customers play the most valuable position in any industry. However, loss of customer to the competitors, often termed as customer churn, can lead to huge loss for the company. The company will not only lose the business, but might also suffer from bad publicity. Also, it is always costlier to acquire new customers than retaining the older ones. Hence, we want to predict customer churn to retain company's business.

Objective:

To develop an algorithm to predict the churn score based on usage pattern.

1.2 Data

Given:

Data Sets –

- 1) [Test_data.csv](#)
- 2) [Train_data.csv](#)

Attribute Information:

The predictors provided are as follows:

- 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
- number of customer service calls

Target class:

Churn: if the customer has moved (1=yes; 0 = no)

Dimension of the data (including target class)

train data : 3334 x 21

test data : 1668 x 21

Give below are the first few observations of the data set :

Table 1.1 Customer Churn data set

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	...
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	...
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	...
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	...
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	...
5	AL	118	510	391-8027	yes	no	0	223.4	98	37.98	...
6	MA	121	510	355-9993	no	yes	24	218.2	88	37.09	...
7	MO	147	415	329-9001	yes	no	0	157.0	79	26.69	...
8	LA	117	408	335-4719	no	no	0	184.5	97	31.37	...
9	WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	...
10	IN	65	415	329-6603	no	no	0	129.1	137	21.95	...
11	RI	74	415	344-9403	no	no	0	187.7	127	31.91	...
...	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	number customer service calls	Churn	
...	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False.	
...	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False.	
...	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False.	
...	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.	
...	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.	
...	101	18.75	203.9	118	9.18	6.3	6	1.70	0	False.	
...	108	29.62	212.6	118	9.57	7.5	7	2.03	3	False.	
...	94	8.76	211.8	96	9.53	7.1	6	1.92	0	False.	
...	80	29.89	215.8	90	9.71	8.7	4	2.35	1	False.	
...	111	18.87	326.4	97	14.69	11.2	5	3.02	0	False.	
...	83	19.42	208.8	111	9.40	12.7	6	3.43	4	True.	

Chapter 2

Methodology

2.1 Missing value analysis:

Many times, our data set can have missing values due to various reasons like error in data extraction or data collection. These missing values can be treated in several ways like deleting them or imputing with mean, median, mode, KNN imputation or make use of other predictive models. However, after analysing we saw that, our data set is free from missing values.

Table 2.1 Missing value analysis for each feature

state	0
account length	0
area code	0
phone number	0
international plan	0
voice mail plan	0
number vmail messages	0
total day minutes	0
total day calls	0
total day charge	0
total eve minutes	0
total eve calls	0
total eve charge	0
total night minutes	0
total night calls	0
total night charge	0
total intl minutes	0
total intl calls	0
total intl charge	0
number customer service calls	0
Churn	0

2.2 Data visualisation :

Let us see the distribution of Churn percent amongst different feature set. First of all let us know how representative is the train data of customer churn as this data will be provided to the model.

Churn	0 (False)	1(True)
	2850	483

Churn percent: $483/(2850+483) = 14.49\%$

The churn percent present in our train data set is on the lower side. If required, we would oversample the minority class (here, when Churn == 1) to improve our model efficiency.

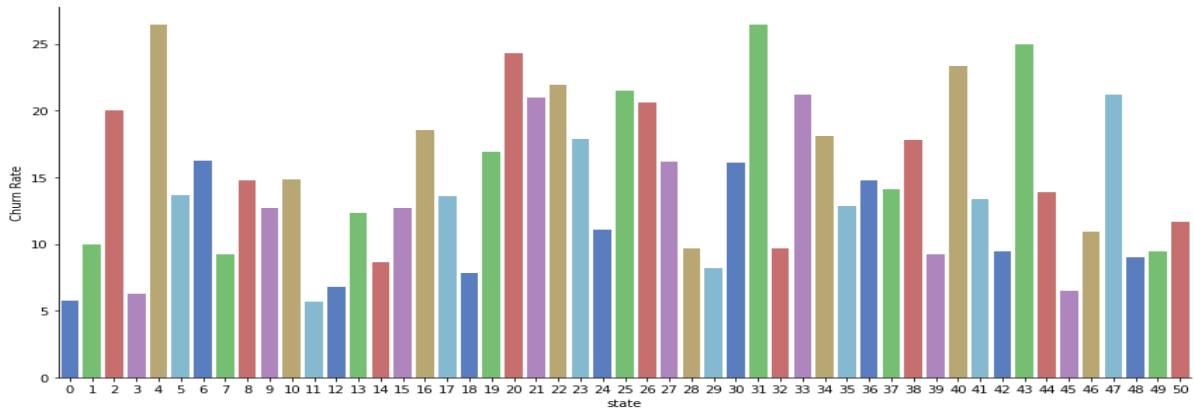
Now let us visualise Churn Rate amongst some categorical feature set in the given train data

State :

We can see certain states like 4 , 20 ,31, 41 , 43 ,40 have higher churn rate .

Table 2.2 Showing Churn Rate , State wise in the decreasing order of churn percent.

S.No.	State Code	Churn Rate (in %)	S.No.	State Code	Churn Rate (in %)	S.No.	State Code	Churn Rate (in %)	
1.	31	26.47	18.	6	16.22	35.	46	10.96	
2.	4	26.47	19.	27	16.18	36.	1	10.00	
3.	43	25.00	20.	30	16.07	37.	32	9.68	
4.	20	24.28	21.	10	14.81	38.	28	9.68	
5.	40	23.33	22.	8	14.75	39.	49	9.43	
6.	22	21.92	23.	36	14.75	40.	42	9.43	
7.	25	21.53	24.	37	14.10	41.	7	9.26	
8.	33	21.21	25.	44	13.89	42.	39	9.23	
9.	47	21.21	26.	5	13.64	43.	48	8.97	
10.	21	20.97	27.	17	13.56	44.	14	8.62	
11.	26	20.59	28.	41	13.33	45.	29	8.20	
12.	2	20.00	29.	35	12.82	46.	18	7.84	
13.	16	18.57	30.	9	12.70	47.	12	6.82	
14.	34	18.07	31.	15	12.68	48.	45	6.49	
15.	23	17.86	32.	13	12.33	49.	3	6.25	
16.	38	17.77	33.	50	11.69	50.	0	5.77	
17.	19	16.92	34.	24	11.11	51.	11	5.66	

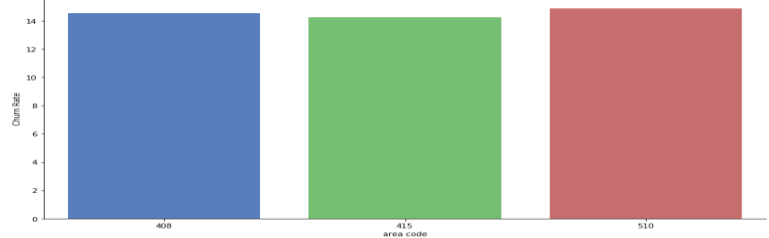


Area code :

Churn rate across different area code is almost the same i.e, around 14.50%.

Table 2.3 Area code wise churn percent in decreasing order

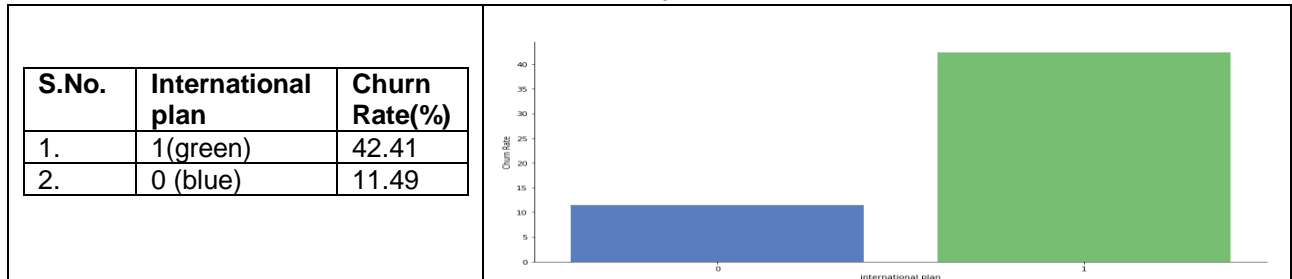
S.No.	Area Code	Churn Rate(%)
1.	510(red)	14.88
2.	408(blue)	14.56
3.	415(green)	14.25



International plan :

We can see that churn rate is higher amongst customer having international plan,i.e around 42.41%.Customers who do not have international plan have churn rate around 11.49%

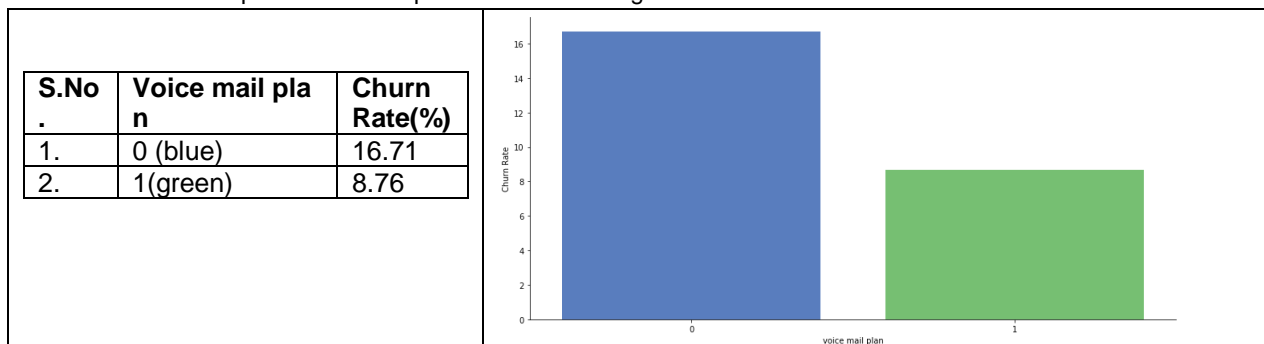
Table 2.4 International plan wise churn percent in decreasing order.



Voicemail plan :

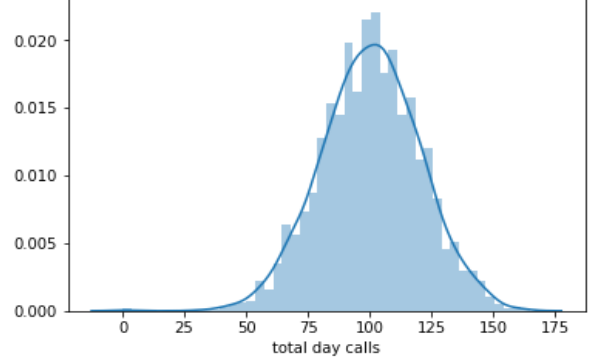
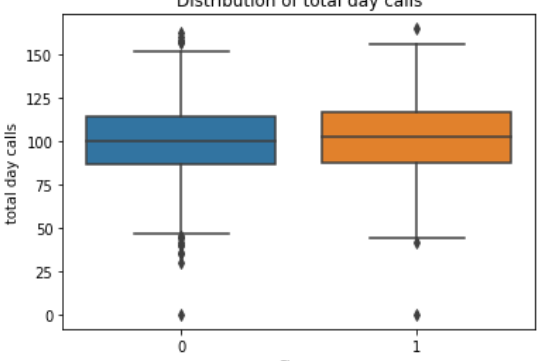
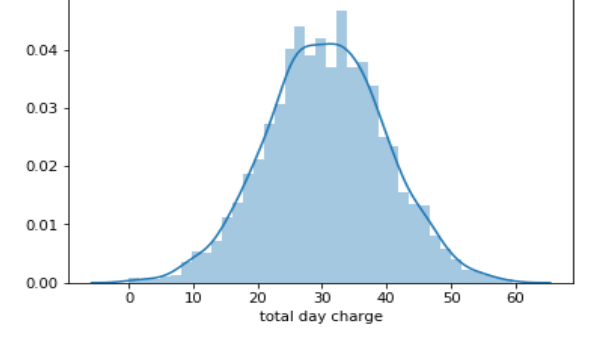
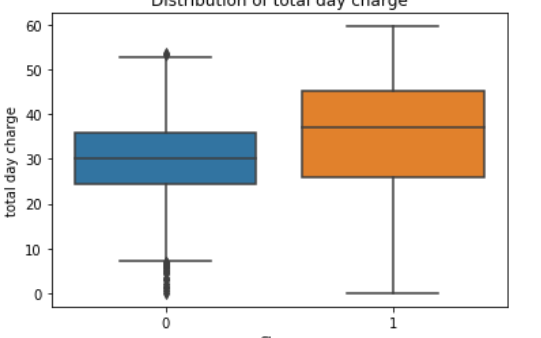
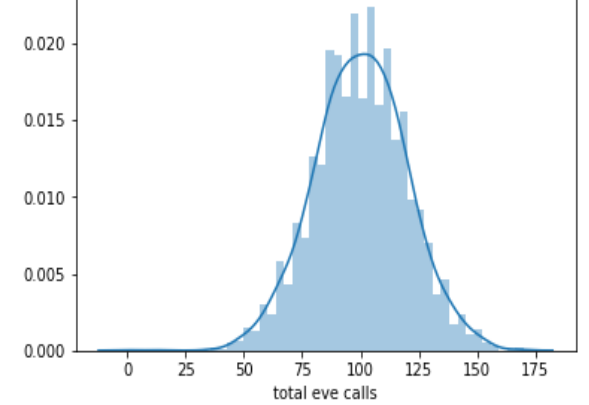
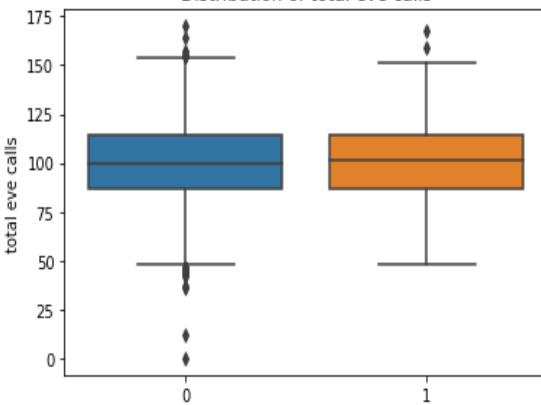
Customer not having voice mail plan has higher churn rate than those who have .Churn rate for customers without voice mail plan is 16.71% and customers with voice mail plan has churn rate of around 8.76%.

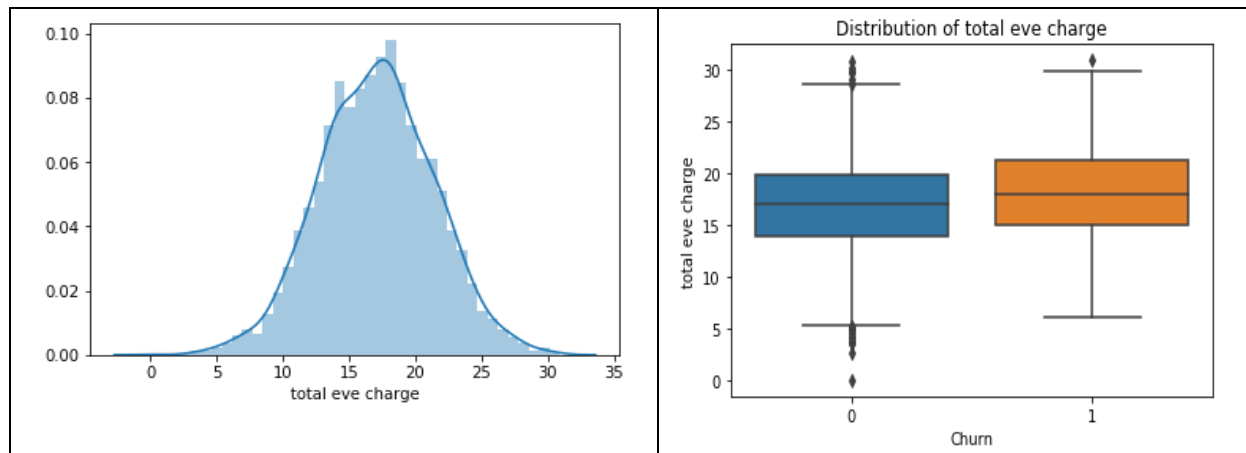
Table 2.5 Voicemail plan wise churn percent in decreasing order



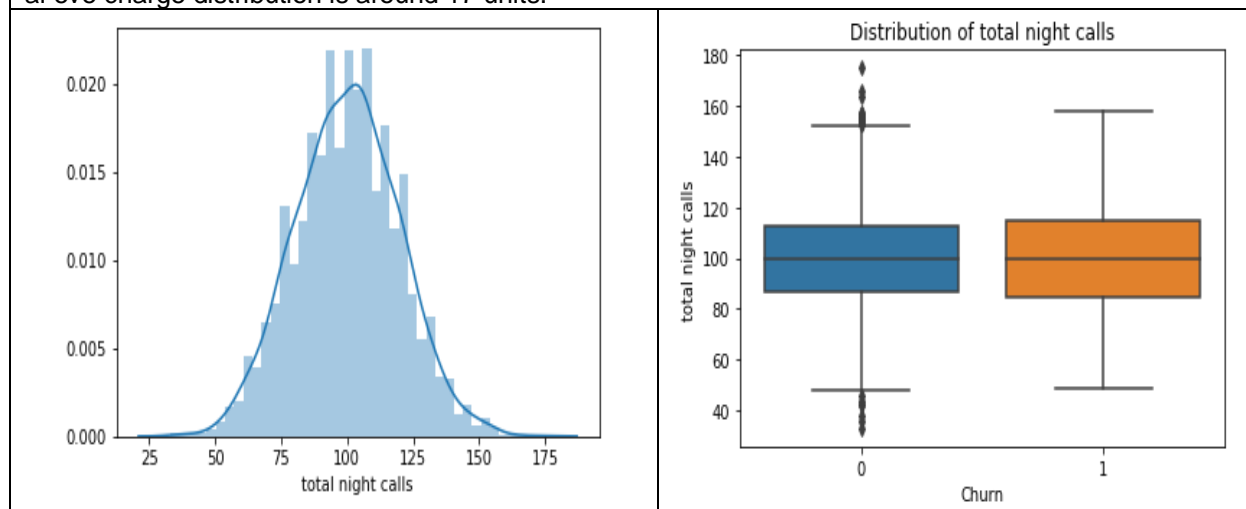
Visualisation of some numerical features:

Table 2.6 Histogram and boxplot of some numerical feature

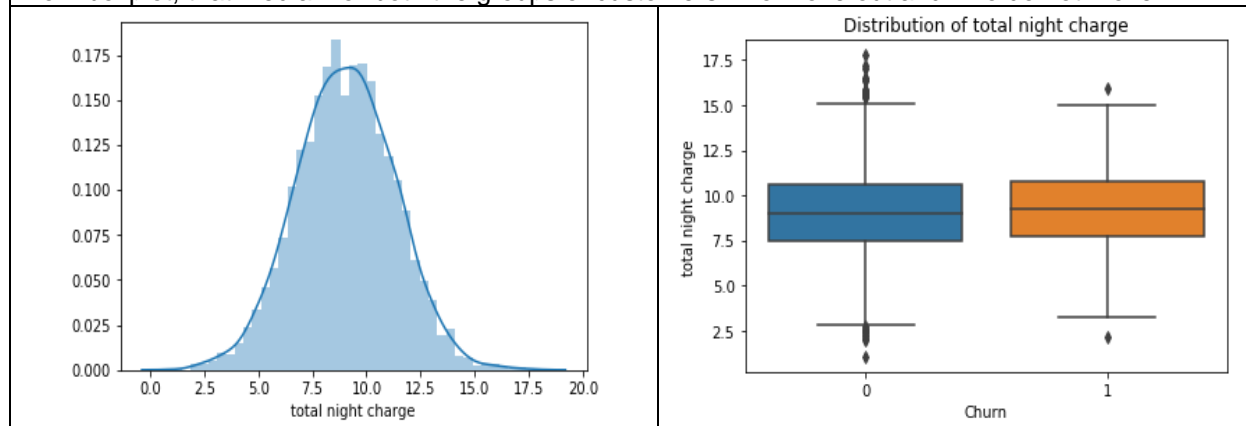
	
<p>Total day calls : We can see the above histogram of total day calls ,which is approximately distribute d around 100.Also the total day calls for customers who churn are slightly more than the customers who do not churn</p>	
	
<p>Total day charge : Total day charge is around 30 units.Customers who churn out bear more total day charge than customers who don't owing to the higher usage pattern demonstrate by them.</p>	
	
<p>Total eve calls : Total evening calls are distributed around 100,with customers who churn out have slightly higher median than those who do not churn</p>	



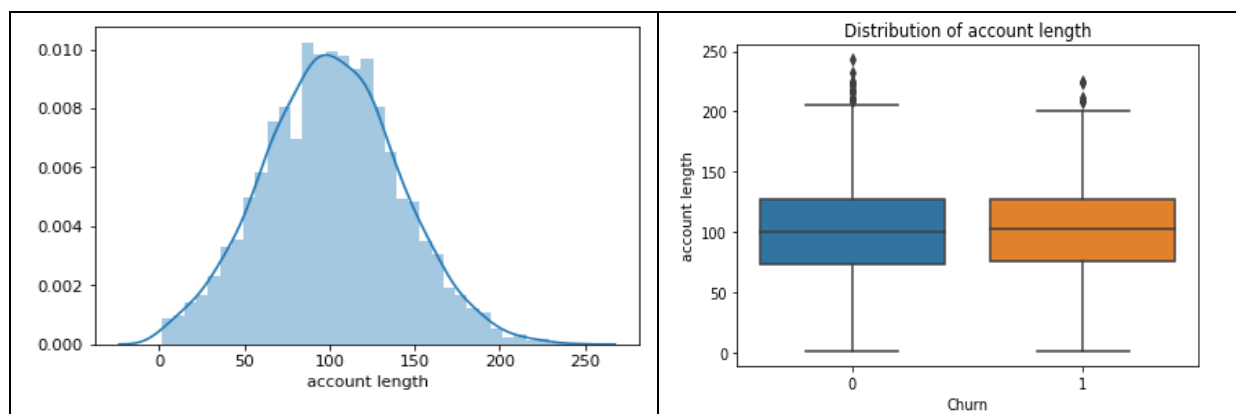
Total eve charge : Customers who move have higher total eve charge than those who do not .The total eve charge distribution is around 17 units.



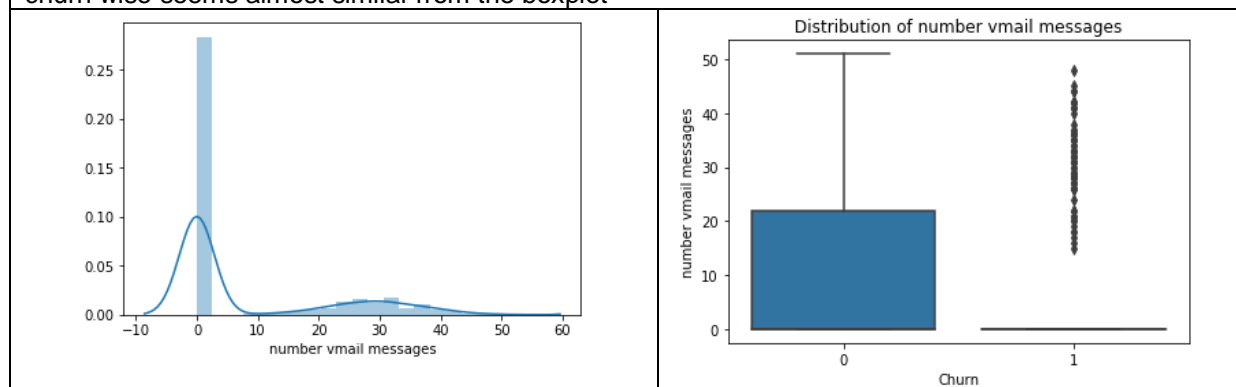
Total night calls : Total night calls are around 100 ,as we can see from the histogram and as well as from boxplot, that median for both the groups of customers who move out and who do not move.



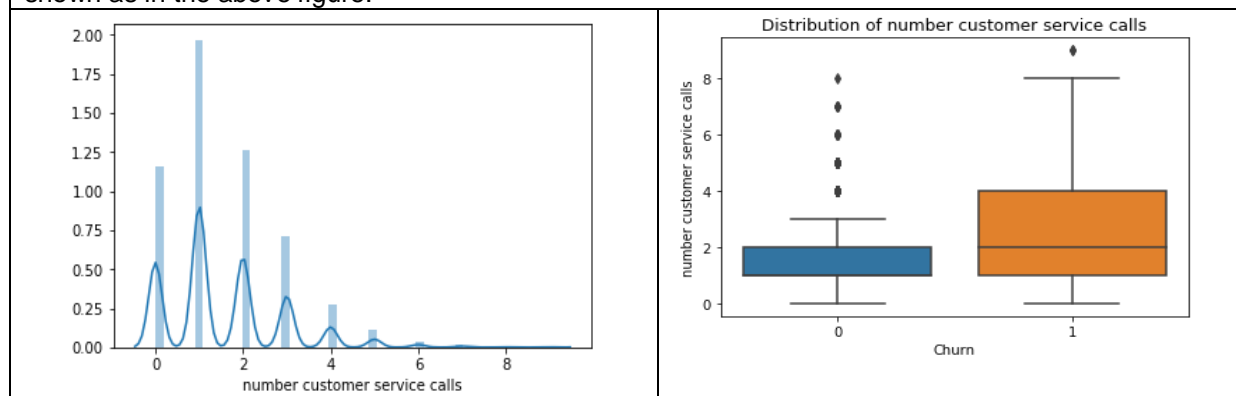
Total night charge : Total night charge is spread around 9 .The boxplot figure of total night charge, churn wise represents slightly higher charges for people who move out .



Account length : Median of the account length is around 100 units .The distribution of account length churn wise seems almost similar from the boxplot



Number vmail messages : Vmail messages shows a peak value at 0 , indicating that customers do not make vmail messages.However some variations are also seen towards right end of the histogram plot,although with lesser probability.The boxplot shows a pattern among customers who churn and who do not churn.People who churn out generally do not use vmail messages,except from some outliers shown as in the above figure.



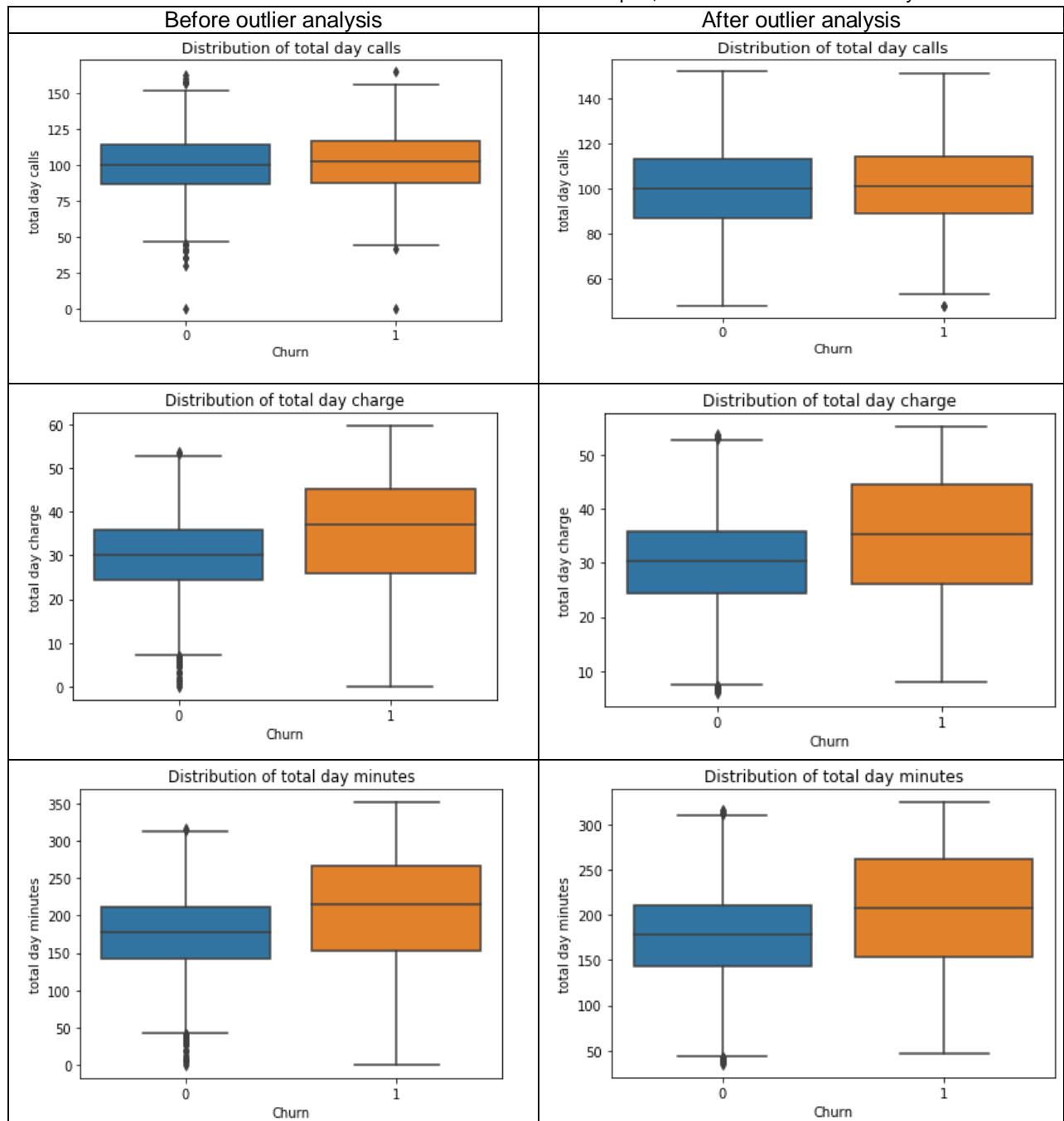
Number of customer service calls: Number of customer service calls for people who churn out are undoubtedly very high. This seems to be a very important aspect of the pattern among customers who move out.

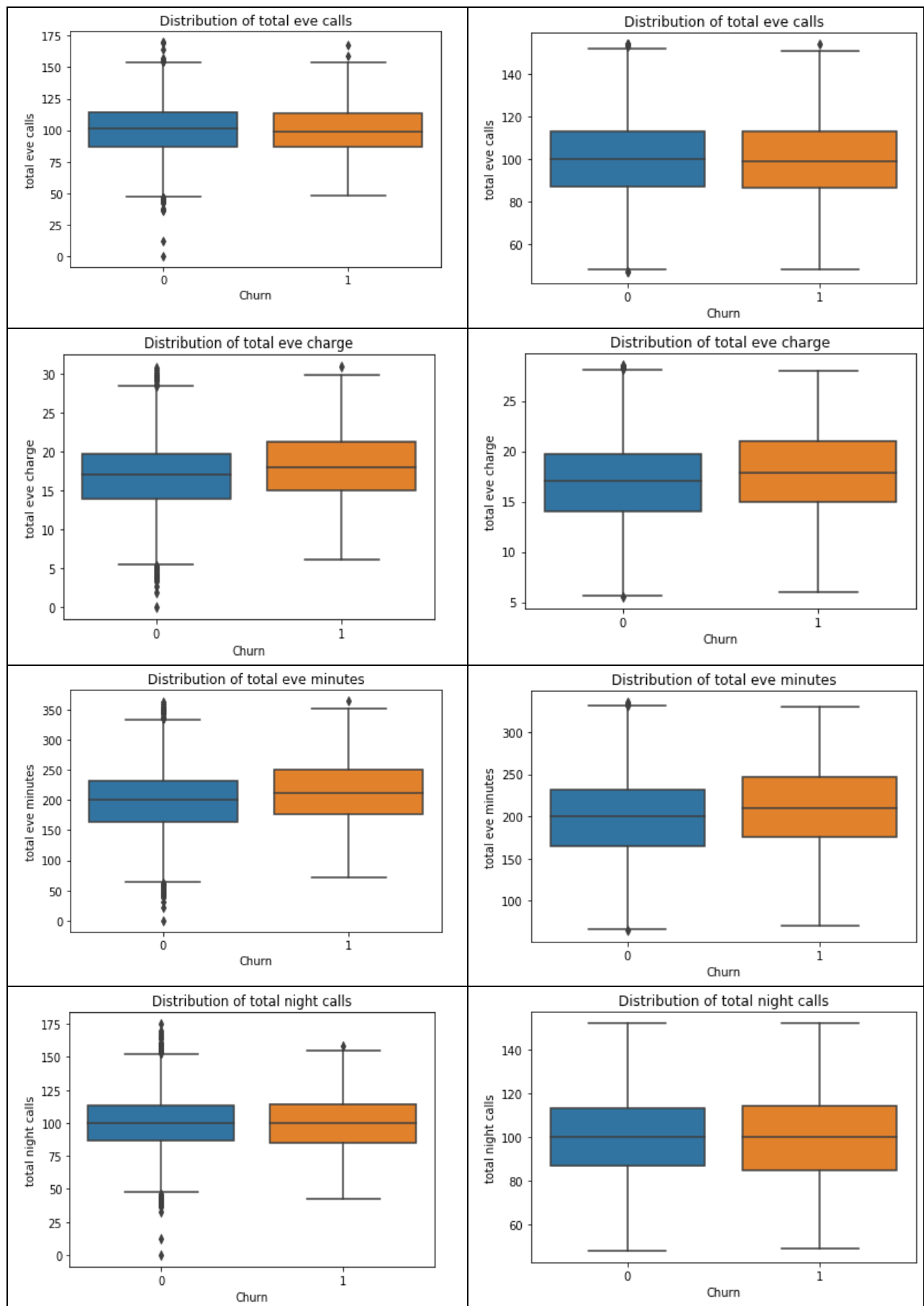
2.3 Outlier detection:

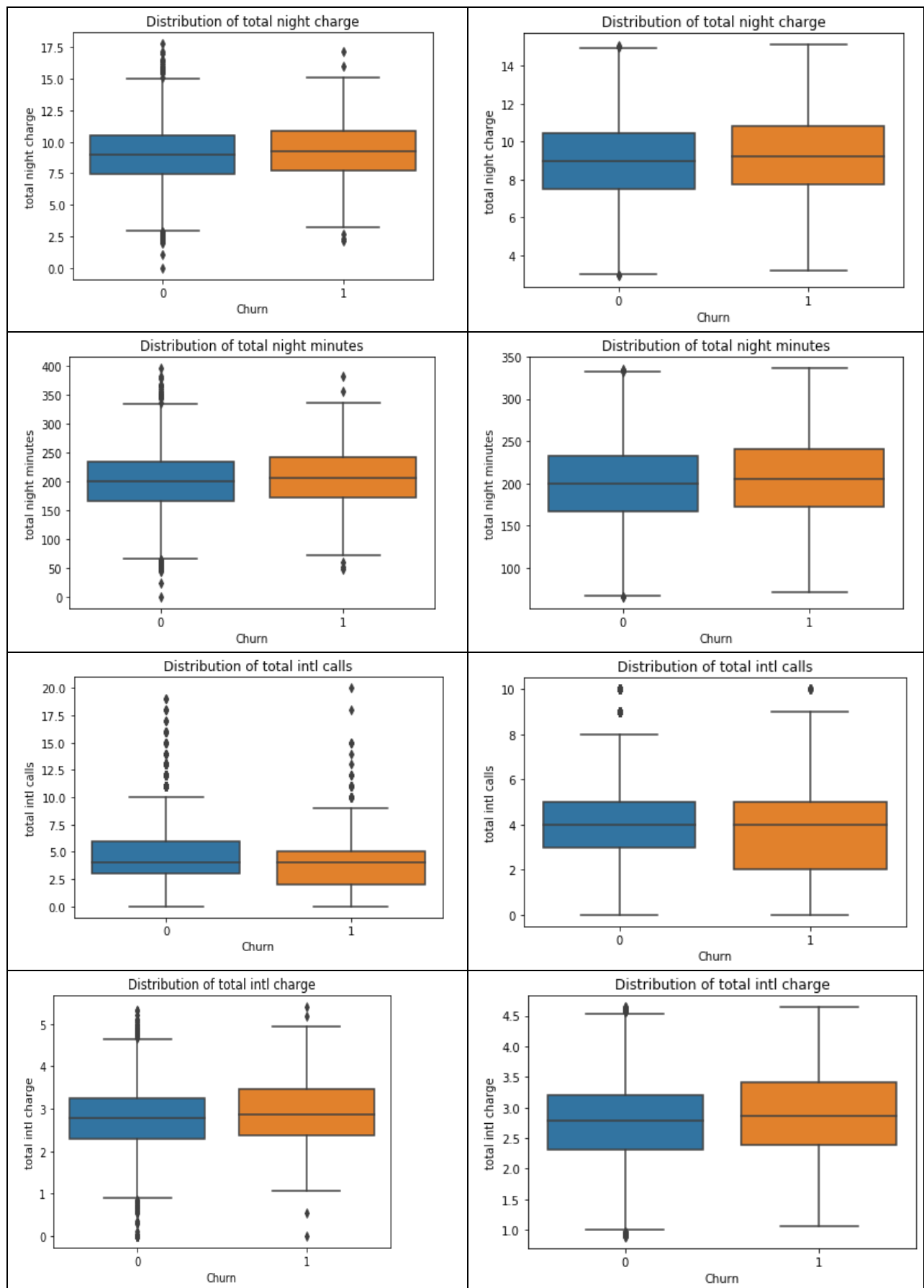
Before feeding the data to our model, we would like to analyse outlier in our data set. If not treated they can substantially effect the results of our model. Here, we do outlier analysis for numerical feature set with the help of boxplot method. Any data point that is less than $1.5 \times \text{IQR}$ (Inter Quartile range) times the 25th percentile or more than $1.5 \times \text{IQR}$ the 75th percentile, is to be treated as an outlier. We have already seen that our train data set contains only 14 % of the event rate that we want to predict, thus we should be careful enough to treat these outliers.

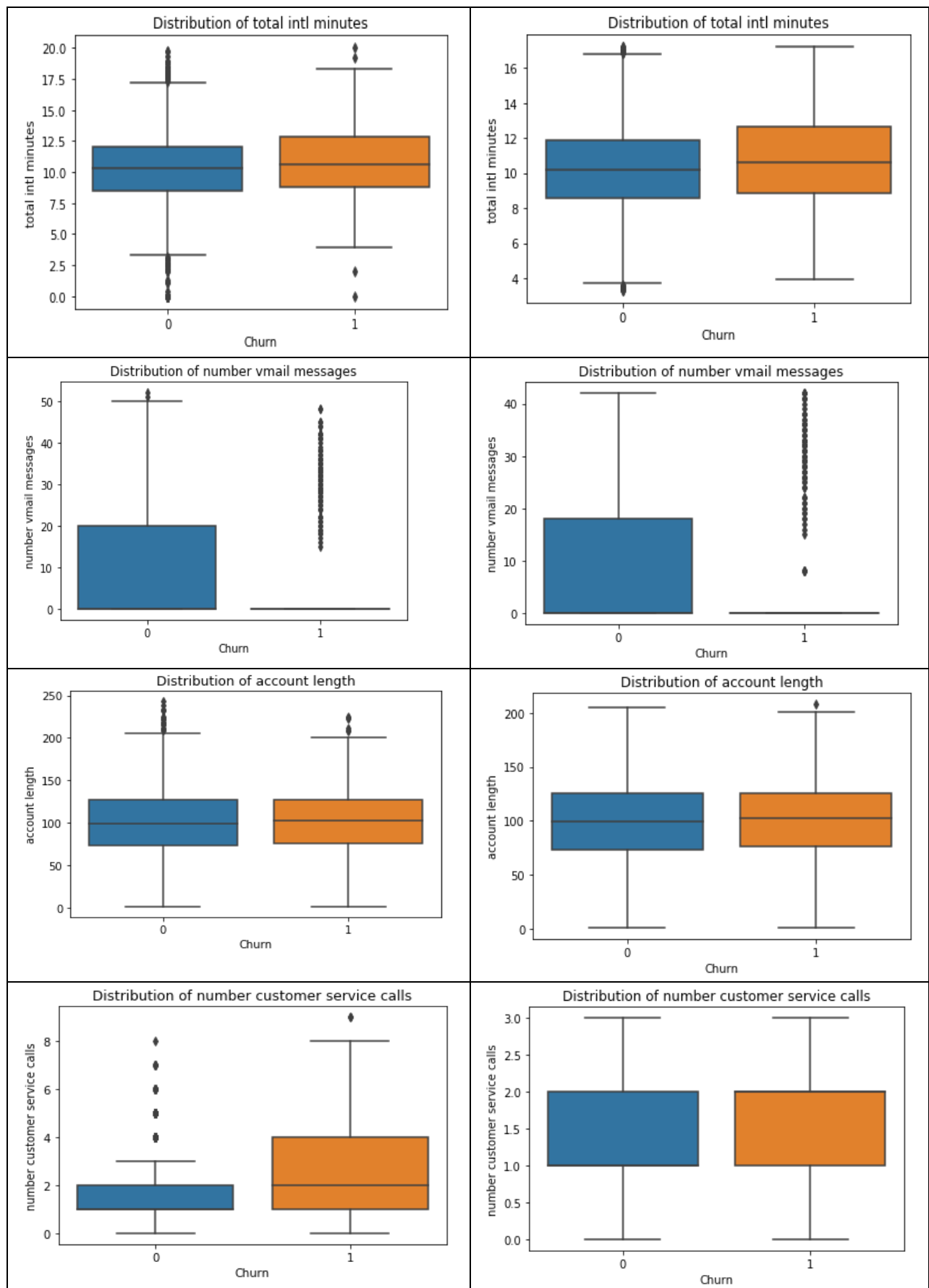
In our problem statement, the outliers are imputed with mean value.

Table 2.7 A brief view of our results can be seen from the below boxplot, before and after outlier analysis.









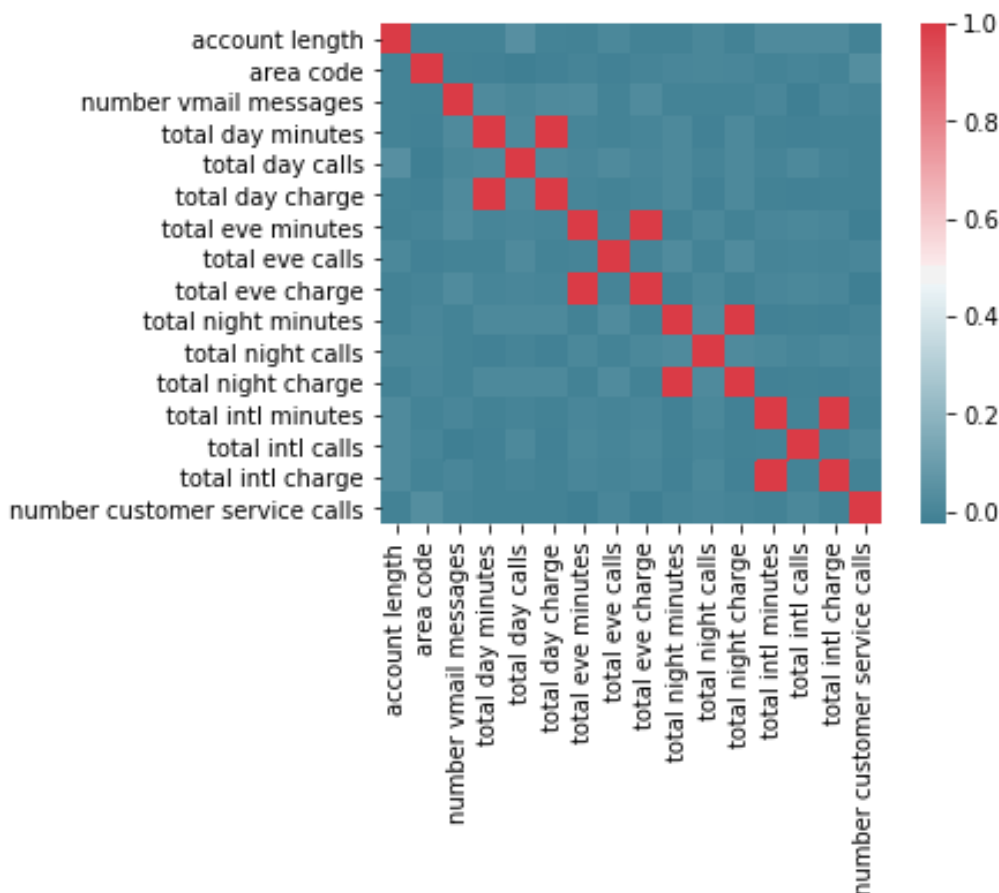
2.4 Feature Selection:

Numerical feature :

For numerical feature set, we perform correlation analysis for feature selection. Correlation tells us how two variables are linearly related to each other. If correlation between two or more variables are high i.e around more than ± 0.80 , it means that they are carrying same level of information, which suggests that one of the features can be safely dropped.

Correlation plot:

Table 2.8 Correlation plot of numerical features



From the above correlation plot, we observe that features "total day minutes", "total eve minutes", "total night minutes", "total intl minutes" are highly correlated with "total day charge", "total eve charge", "total night charge", "total intl charge" respectively. Thus, we will remove one of the features from the highly correlated feature pairs.

Categorical feature:

Feature reduction for categorical feature set with categorical target variable can be done using Chi Square test. This test is used to test the statistical significance of relationship between the variables.

$$\text{Chi square } (X^2) = \sum (O-E)^2/E$$

Where O = observed frequency

E = expected frequency under the null hypothesis

Probability values (p value) of different feature set are given below. Probability less than 0.05 indicates that the relationship between the variables is significant at 95% confidence interval. Thus, we reject those variables which have p value more than 0.05.

The p value observed from chi square test is given below:

state
7.85083622437e-05
area code
0.754658138533
phone number
0.493350889587
international plan
1.9443947475e-74
voice mail plan
7.16450178099e-15

Thus, as p value for area code and phone number is greater than 0.05, we remove these features from our feature set

After feature selection, we observe that we could finally reduce our feature dimension from 20 to 14. We are now ready to feed data to our model.

2.5 Model Development

In our problem statement, we have been provided with 2 files - train and test set .Let us now try different machine learning algorithms for classification. The model is built on train set and predictions are then made on test set

We start with the most basic algorithm for classification then move to the complex ones.

Logistic Regression:

Code :

Python :

```
model_LR = LogisticRegression()  
modelLR_fit = model_LR.fit(x_train,y_train)  
modelLR_predict = model_LR.predict(x_test)  
model_evaluation(y_test,modelLR_predict)
```

Churn (Actual\Predicted)	0	1
0	1426	194
1	17	30

Naïve Bayes :

Code :

Python :

```
model_NB = GaussianNB()  
modelNB_fit = model_NB.fit(x_train,y_train)  
modelNB_predict = model_NB.predict(x_test)  
model_evaluation(y_test, modelNB_predict)
```

Churn (Actual\Predicted)	0	1
0	1358	154
1	85	70

K Nearest Neighbour :

Code :

Python :

```
model_KNN = KNeighborsClassifier(n_neighbors=5)  
modelKNN_fit = model_KNN.fit(x_train,y_train)  
modelKNN_predict = model_KNN.predict(x_test)  
model_evaluation(y_test,modelKNN_predict)
```

Churn (Actual\Predicted)	0	1
0	1424	207
1	19	17

Decision Tree:

Code:

Python:

```
model_C50 = DecisionTreeClassifier()
modelC50_fit = model_C50.fit(x_train,y_train)
modelC50_predict = model_C50.predict(x_test)
model_evaluation(y_test,modelC50_predict)
```

Churn (Actual\Predicted)	0	1
0	1319	110
1	124	114

Random Forest:

Code :

Python :

```
model_RF = RandomForestClassifier()
modelRF_fit = model_RF.fit(x_train,y_train)
modelRF_predict = model_RF.predict(x_test)
model_evaluation(y_test,modelRF_predict)
```

Churn (Actual\Predicted)	0	1
0	1440	122
1	3	102

3. Conclusion

Model Evaluation and Model Selection:

In the given context of our problem statement we wanted to reduce customer churn rate. Hence, it is very important to reduce False Negative Rate, as we would not like to get misled if our model fails to predict actual customers who will churn out. Hence, here False Negative Rate is the most important parameter than accuracy for model evaluation on the basis of which we will select our model.

Machine Learning Model	Accuracy = $(TP+TN)/(TP+TN+FP+FN)$	FNR(False Negative Rate) = $FP/(FP+TP)$
	Python	Python
Logistic Regression	87.34%	36.17%
Naïve Bayes	85.66%	54.84%
K Nearest Neighbour	86.44%	52.77%
Decision Tree	85.96%	52.10%
Random Forest	92.50%	2.86%

We can see that Random Forest performs best in R as well as in Python.

References :

- i.) <https://edwisor.com/>
- ii.) <https://www.analyticsvidhya.com>
- iii.) <https://www.kdnuggets.com/>
- iv.) <https://towardsdatascience.com/>

Note: Figures and References are made from Python code outputs