PORTFOLIO PROJECT

PROJECT REPORT
ON
CHURN REDUCTION

ABHIMANYU BHATIA

01 February 2019

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. This problem statement is targeted at enabling churn reduction using analytics concepts.

The objective of this Case is to predict customer behaviour. We have been provided with a public dataset that has customer usage pattern and if the customer has moved or not.
We need to develop an algorithm to predict the churn score based on usage pattern.

1.2. Exploratory Data Analysis

The objective is to apply the machine learning algorithms and build models on this dataset in order to predict the customer behaviour.
Given below is a sample of the data set that shared to us:

Table 1: Churn Reduction sample data (Columns: 1-7)

state	account length	area code	phone number	international plan	voice mail plan	number vmail messages
KS	128	415	382-4657	No	yes	25
OH	107	415	371-7191	No	yes	26
NJ	137	415	358-1921	No	no	0
OH	84	408	375-9999	Yes	no	0

Table 2: Churn Reduction sample data (Columns: 8-14)

total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes
265.1	110	45.07	197.4	99	16.78	244.7
161.6	123	27.47	195.5	103	16.62	254.4
243.4	114	41.38	121.2	110	10.3	162.6
299.4	71	50.9	61.9	88	5.26	196.9

Table 3: Churn Reduction sample data (Columns: 15-21)

total night calls	total night charge	total intl minutes	total intl calls	total intl charge	number customer service calls	Churn
91	11.01	10	3	2.7	1	False.
103	11.45	13.7	3	3.7	1	False.
104	7.32	12.2	5	3.29	0	False.
89	8.86	6.6	7	1.78	2	False.

As you can see in the Table 4 we have the following 20 predictor variables, using which we have to predict the churn:

Table 4: Predictor Variables of our Dataset

S. No.	PREDICTOR
1.	state
2.	account length
3.	area code
4.	phone number
5.	international plan
6.	voice mail plan
7.	number vmail messages
8.	total day minutes
9.	total day calls
10.	total day charge
11.	total eve minutes
12.	total eve calls
13.	total eve charge
14.	total night minutes
15.	total night calls
16.	total night charge
17.	total intl minutes
18.	total intl calls
19.	total intl charge
20.	number customer service calls

By observing the dataset we have created four new variables:

Total minutes = Total day minutes+ Total eve minutes+ Total night minutes+ Total intl minutes

Total calls = Total day calls+ Total eve calls+ Total night calls+ Total intl calls

Total charge = Total day charge+ Total eve charge+ Total night charge+ Total intl charge

Total average minutes/call = Total minutes/ Total charge

Using the describe function in python we have obtained a brief description of our dataset

Table 5: Brief Description of the Dataset

	count	mean	std	min	25%	50%	75%	max
account length	5000.0	100.258600	39.694560	1.0	73.000	100.00	127.00	243.00
area code	5000.0	436.911400	42.209182	408.0	408.000	415.00	415.00	510.00
number vmail messages	5000.0	7.755200	13.546393	0.0	0.000	0.00	17.00	52.00
total day minutes	5000.0	180.288900	53.894699	0.0	143.700	180.10	216.20	351.50
total day calls	5000.0	100.029400	19.831197	0.0	87.000	100.00	113.00	165.00
total day charge	5000.0	30.649668	9.162069	0.0	24.430	30.62	36.75	59.76
total eve minutes	5000.0	200.636560	50.551309	0.0	166.375	201.00	234.10	363.70
total eve calls	5000.0	100.191000	19.826496	0.0	87.000	100.00	114.00	170.00
total eve charge	5000.0	17.054322	4.296843	0.0	14.140	17.09	19.90	30.91
total night minutes	5000.0	200.391620	50.527789	0.0	166.900	200.40	234.70	395.00
total night calls	5000.0	99.919200	19.958686	0.0	87.000	100.00	113.00	175.00
total night charge	5000.0	9.017732	2.273763	0.0	7.510	9.02	10.56	17.77
total intl minutes	5000.0	10.261780	2.761396	0.0	8.500	10.30	12.00	20.00
total intl calls	5000.0	4.435200	2.456788	0.0	3.000	4.00	6.00	20.00
total intl charge	5000.0	2.771196	0.745514	0.0	2.300	2.78	3.24	5.40
number customer service calls	5000.0	1.570400	1.306363	0.0	1.000	1.00	2.00	9.00

We can infer that out of the 21 variables in our dataset 15 are continuous and 6 are categorical variables.

- **CONTINUOUS VARIABLES:** account length, 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, phone number

- **CATEGORICAL VARIABLES:** state, area code , international plan, voice mail plan, number customer service calls ,Churn

2. DATA PRE-PROCESSING

2.1. Assigning levels to Categorical Variables

- We have converted the following categorical variables into factor variables:
  - international plan
  - voice mail plan
  - Churn
  - state

2.2. Missing Value Analysis

- Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.
- If the missing data is less than 30% of the total data its value can be imputed using various statistical techniques.
- Table 6 shows the missing values in our dataset.

Table 6: Missing Values in Churn Reduction dataset

Feature Variables	No. of missing values	Percentage of Missing Values (%)
State	0	0.00
account length	0	0.00
area code	0	0.00
phone number	0	0.00
international plan	0	0.00
voice mail plan	0	0.00
number vmail messages	0	0.00
total day minutes	0	0.00
total day calls	0	0.00
total day charge	0	0.00
total eve minutes	0	0.00
total eve calls	0	0.00
total eve charge	0	0.00
total night minutes	0	0.00
total night calls	0	0.00
total night charge	0	0.00
total intl minutes	0	0.00
total intl calls	0	0.00
total intl charge	0	0.00
number customer service calls	0	0.00

- From the table 6 it can be clearly seen that there are no missing values.
- Thus, there is no requirement for imputation of the missing values.

2.3. Outlier Analysis

- An outlier is an observation point that is distant from other observations.
- An outlier may be due to variability in the measurement or it may indicate experimental error.
- An outlier can cause serious problems in statistical analyses
- Box plots are non-parametric i.e. they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution.

- Thus, Box Plot technique for detecting Outliers can be used for any type of data distribution.
- To perform Outlier Analysis box plots for all the continuous variables are plotted and shown below.

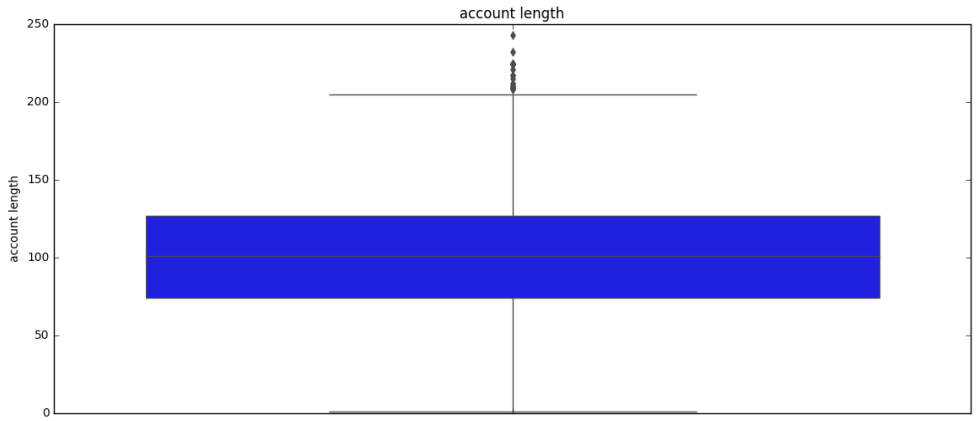

Fig.1: Box Plot of variable “Account length”

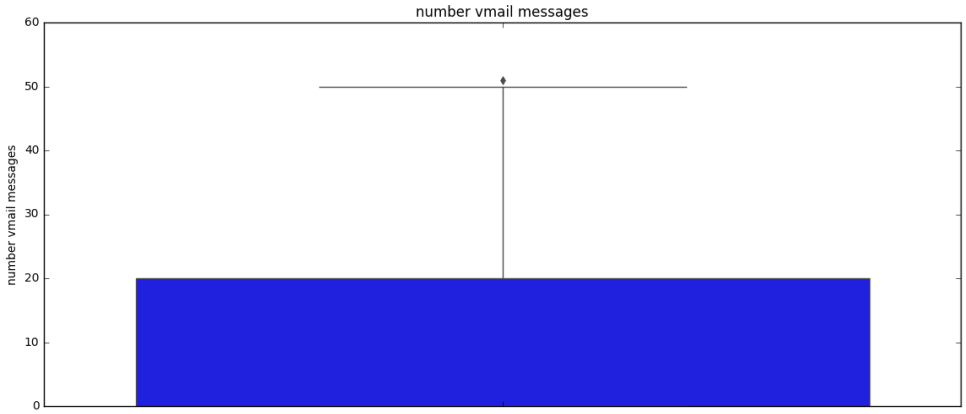

Fig.2: Box Plot of variable “Number vmail messages”

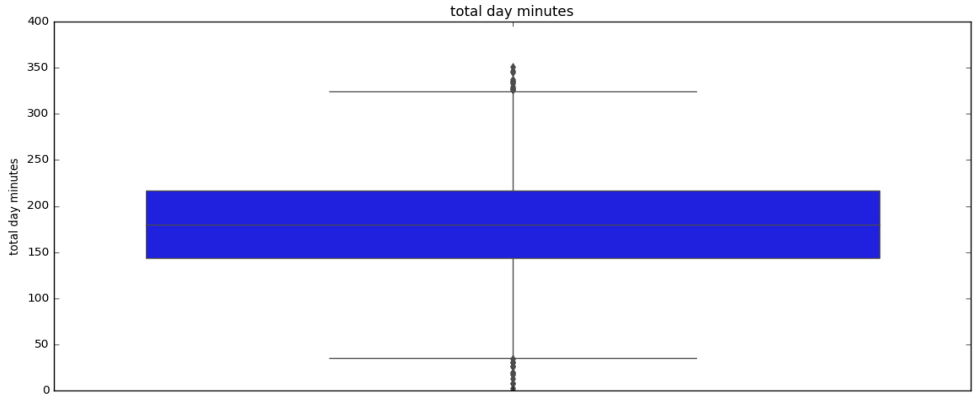

Fig.3: Box Plot of variable “Total day minutes”

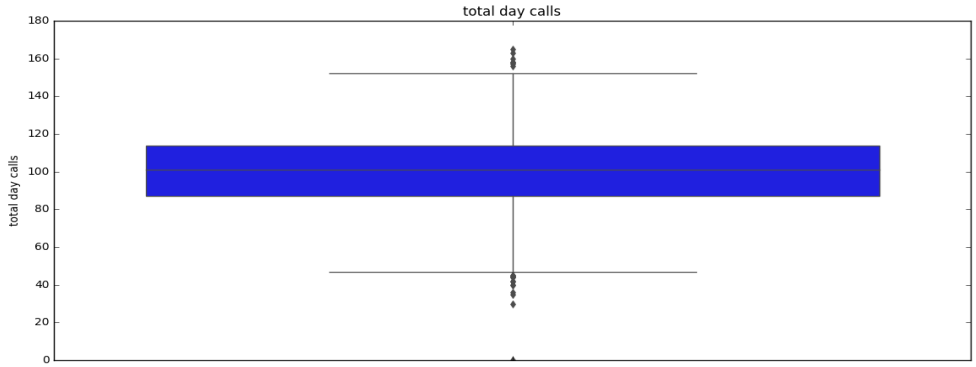

Fig.4: Box Plot of variable “Total day calls”

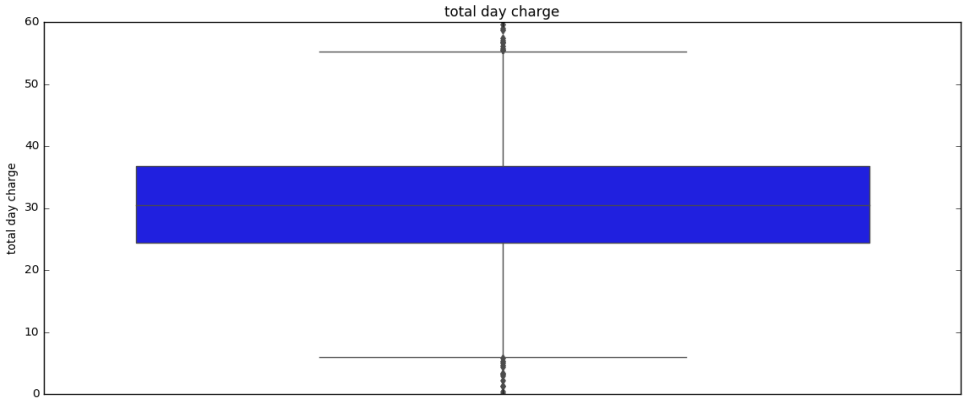

Fig.5: Box Plot of variable “Total day charge”

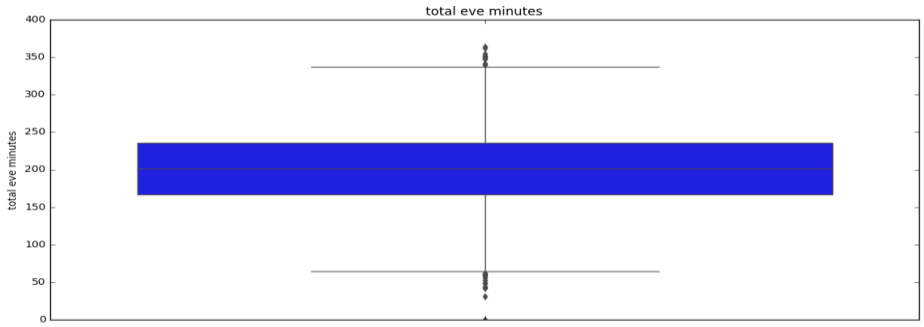

Fig.6: Box Plot of variable “Total eve minutes”

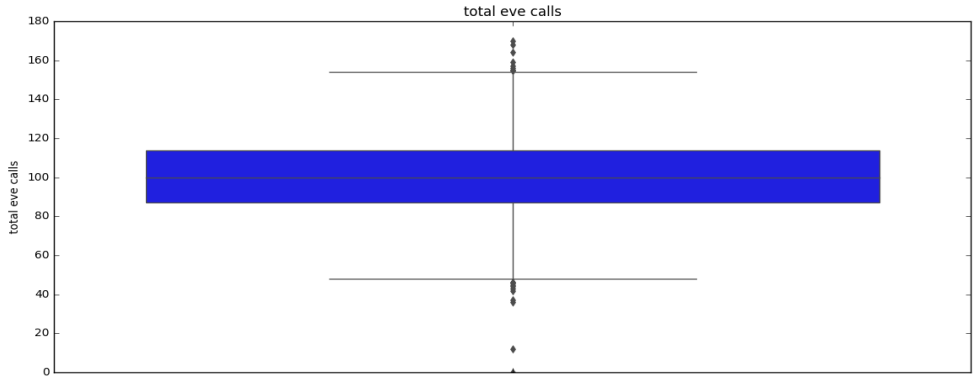

Fig.7: Box Plot of variable “Total eve calls”

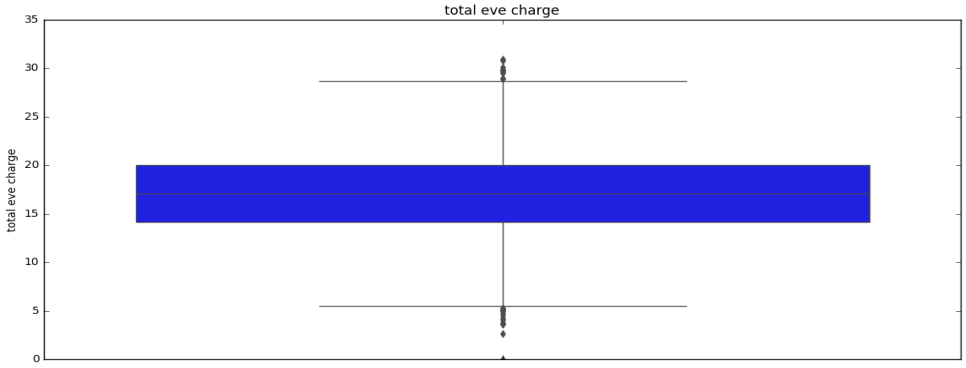

Fig.8: Box Plot of variable “Total eve charge”

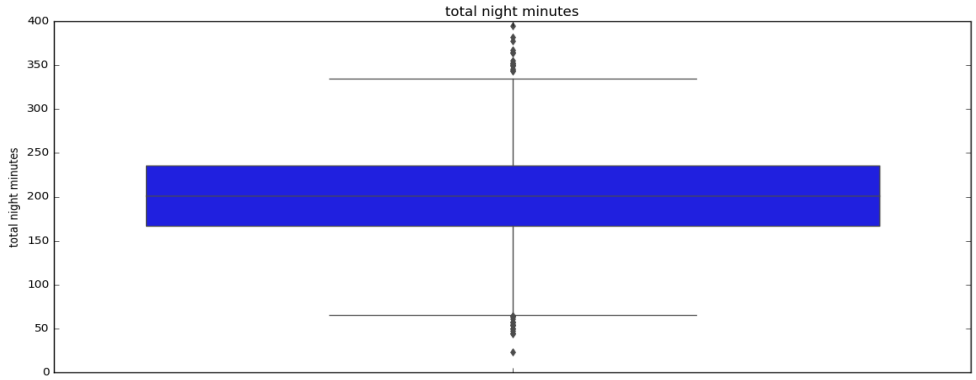

Fig.9: Box Plot of variable “Total night minutes”

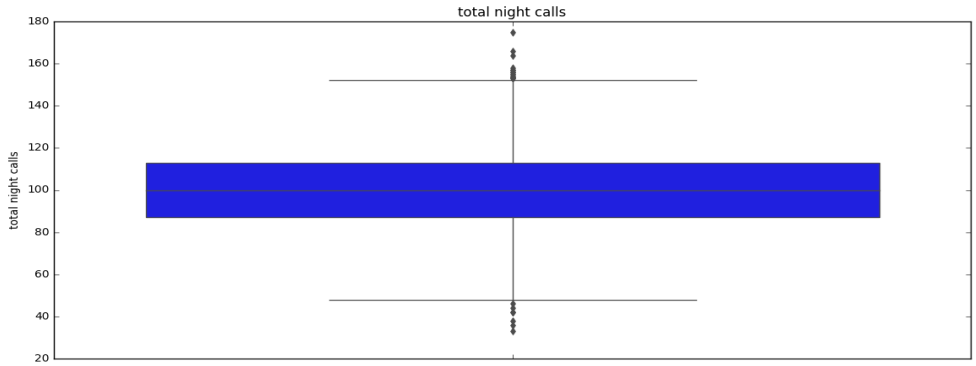

Fig.10: Box Plot of variable “Total night calls”

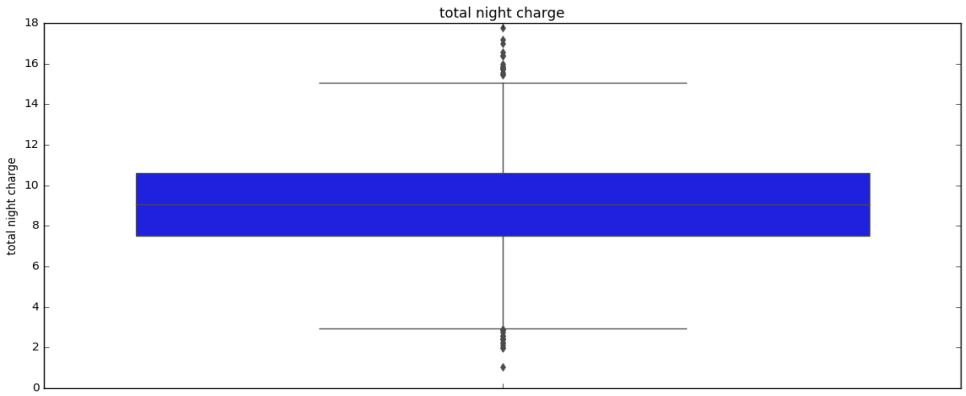

Fig.11: Box Plot of variable “Total night charge”

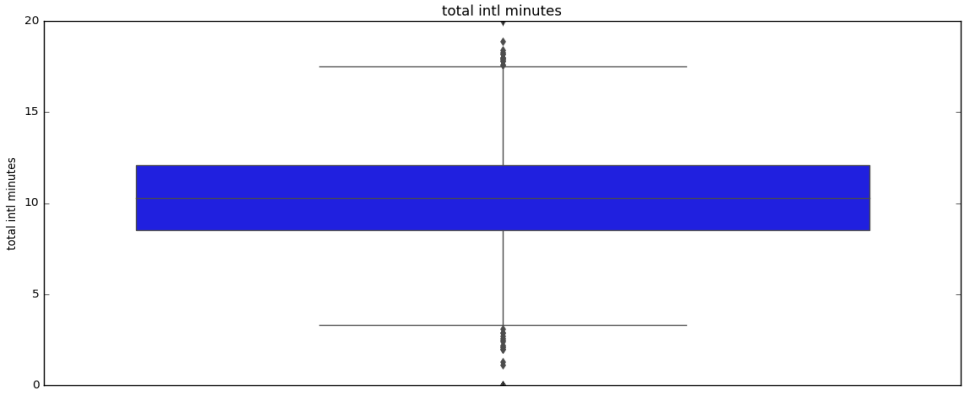

Fig.12: Box Plot of variable “Total intl minutes”

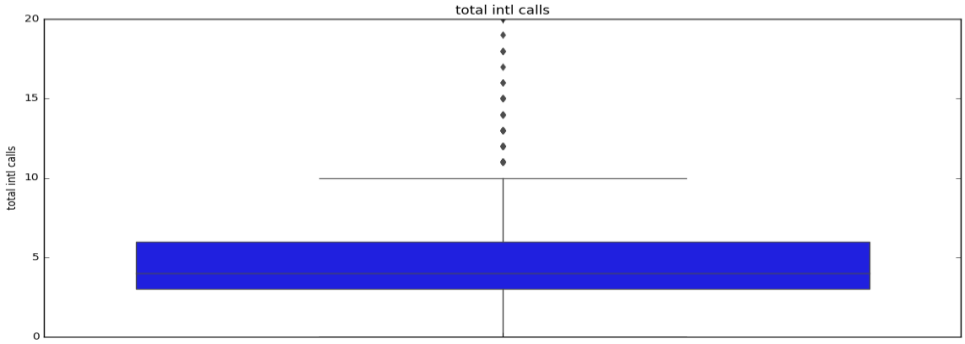

Fig.13: Box Plot of variable “Total intl calls”

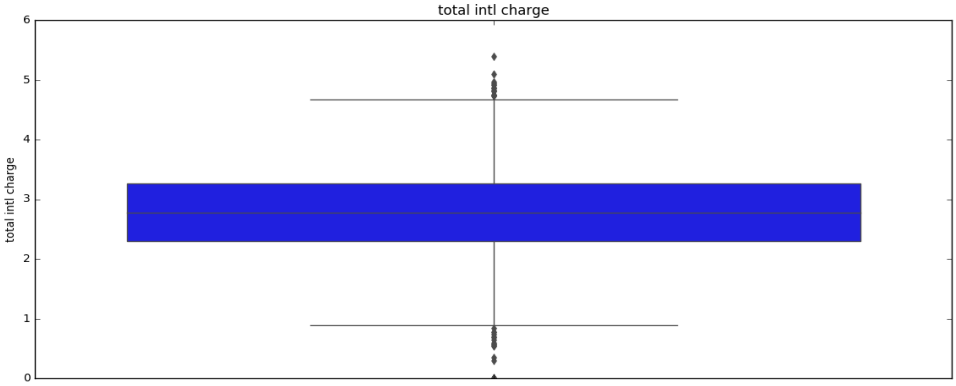

Fig.14: Box Plot of variable “Total intl charge”

2.4. Outliers Imputation

- From the boxplots we were able to detect outliers
- For imputation three methods were tested: Mean, Median & KNN.
- KNN imputation was found to be the most accurate.
- KNN imputation is used for imputation of Outliers.

2.5. Feature Selection

- In large datasets the number of predictor variables are sometimes very high.
- When these variables have high correlation with each other the problem of multi-collinearty arises.
- By using techniques of Correlation plots and chi-square tests we can check the variables that contribute towards multi-collinearity.
- Thus, some of these variables can be dropped.
- It then becomes easier to analyse the dataset.

2.4.1 Creation of new variables

We have created four new variables by combining the existing predictor variables.

Total minutes = Total day minutes+ Total eve minutes+ Total night minutes+ Total intl minutes

Total calls = Total day calls+ Total eve calls+ Total night calls+ Total intl calls

Total charge = Total day charge+ Total eve charge+ Total night charge+ Total intl charge

Total average minutes/call = Total minutes/ Total charge

Further we converted these newly created variables into factor type for easy analysis.

Table 7: Creation of new variables in the dataset

Total Charge	Total Calls	Total Minutes	Category Assigned	Level Assigned
25-50	150-250	200-400	Low	0
50-75	250-350	400-600	Medium	1
More than 75	More than 350	More than 600	High	2

2.4.2 Correlation

The correlation plot of all the continuous variables is shown below:

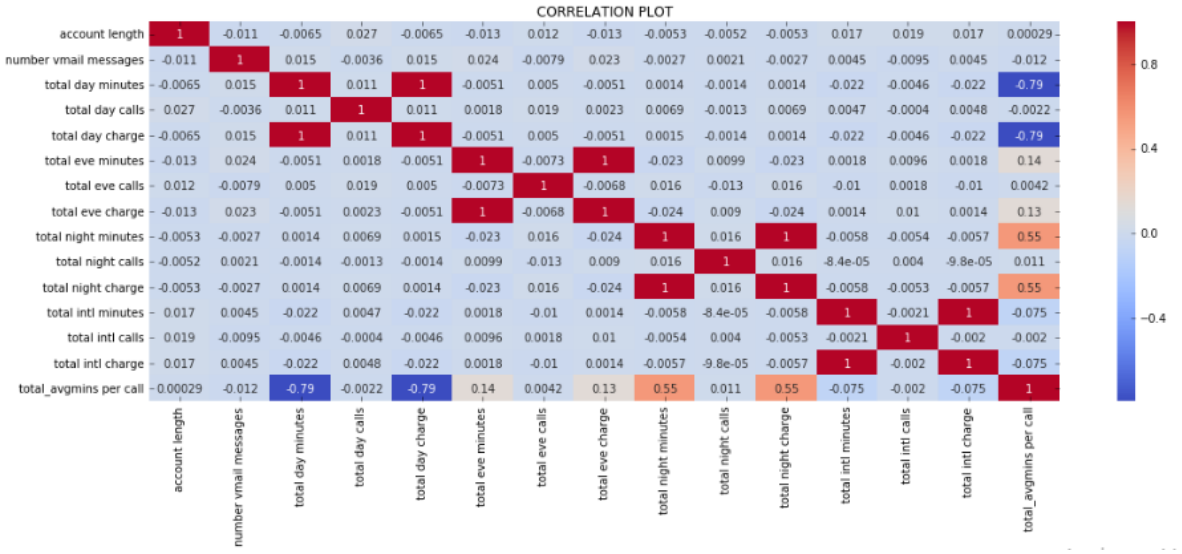

Fig.15: Correlation Plot

From the above plot following observations can be made:

- Total day minutes, Total day charge and total_avgmin per call are highly correlated (factor of -0.8).
- Also since we have created four new variables, we can delete the following variables:
  - total day minutes, total eve minutes, total night minutes, total intl minutes, total day calls, total eve calls, total night calls, total intl calls, total day charge, total eve charge, total night charge, total intl charge, phone number.

2.4.3 CHI-SQUARE TEST

We have used the chi-square test for categorical variables.
The P value of state is 7.850836224371827e-05
The P value of international plan is 1.9443947474998577e-74
The P value of voice mail plan is 7.164501780988496e-15
The P value of number customer service calls is 4.186291993492475e-101
The P value of Total minutes is 1.2163282287655295e-11
The P value of Total calls is 0.7470342892016619
The P value of Total charge is 8.713882039076516e-212

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

- From the above chi-square test we observed that the p values for some of the categorical variables is more than 5%
- Thus, we can drop those categorical variables from our model.
- Following categorical variables have been dropped: Total calls.

2.6. Feature Scaling

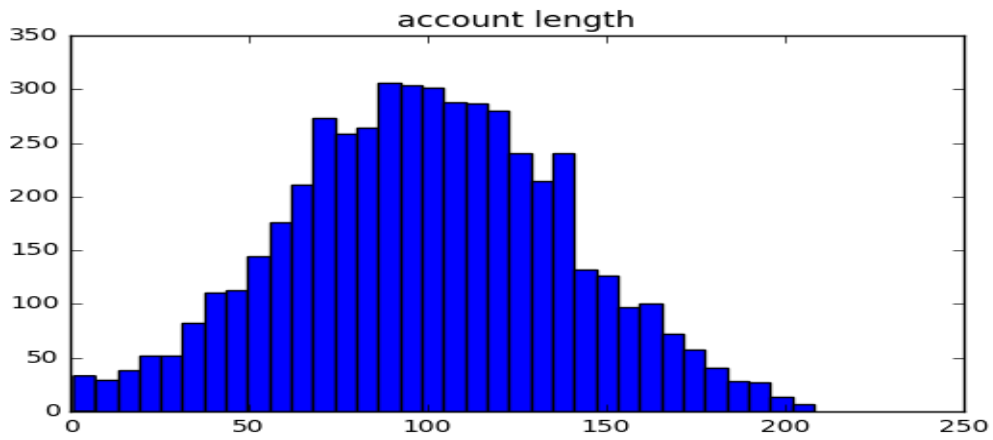

Fig. 16: Histogram of “Account length”

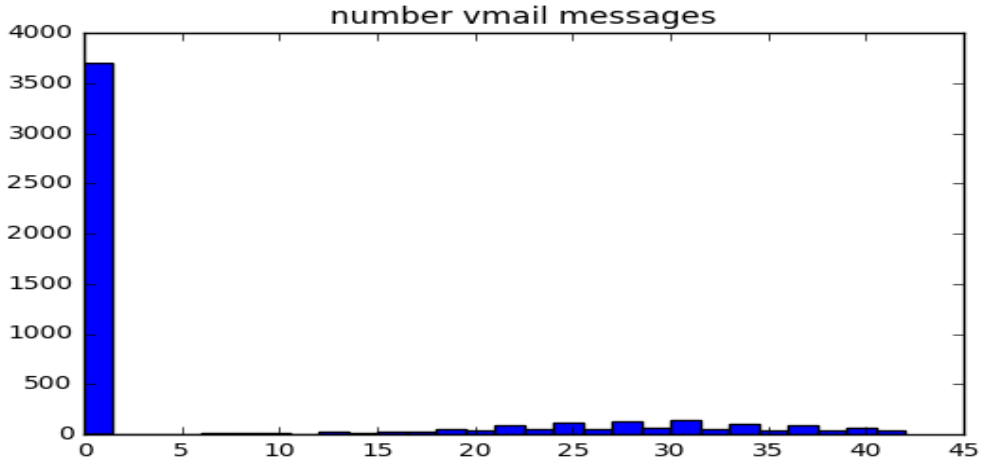

Fig. 17: Histogram of “Number vmail messages”

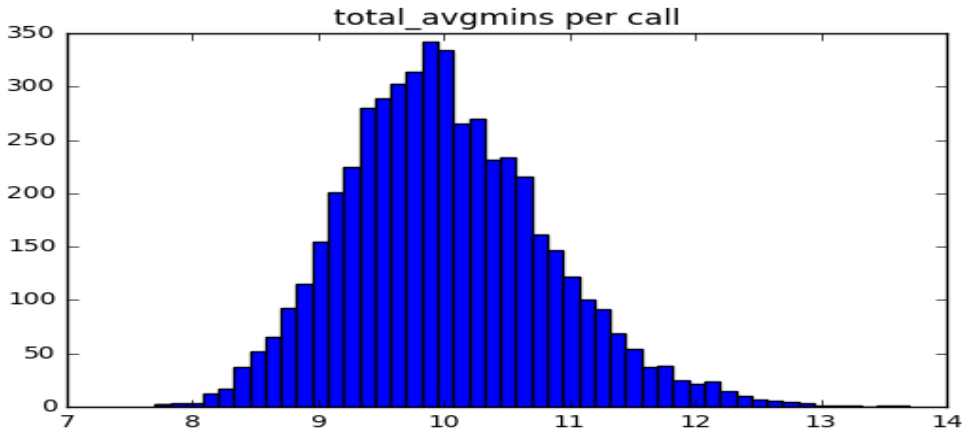

Fig. 18: Histogram of “Total avg. minutes per call”

- Normalization & Standardization are the two techniques used for Feature Scaling.
- These techniques are used for reducing unwanted variation either within or between variables.
- Standardization is used when data is uniformly distributed.
- Since the data is non-uniformly distributed as can be seen from the Histograms above, we have used Normalization to bring all of the variables into proportion with one another.

$$Value_{new} = \frac{Value - minValue}{maxValue - minValue}$$

3. DATA VISUALIZATION

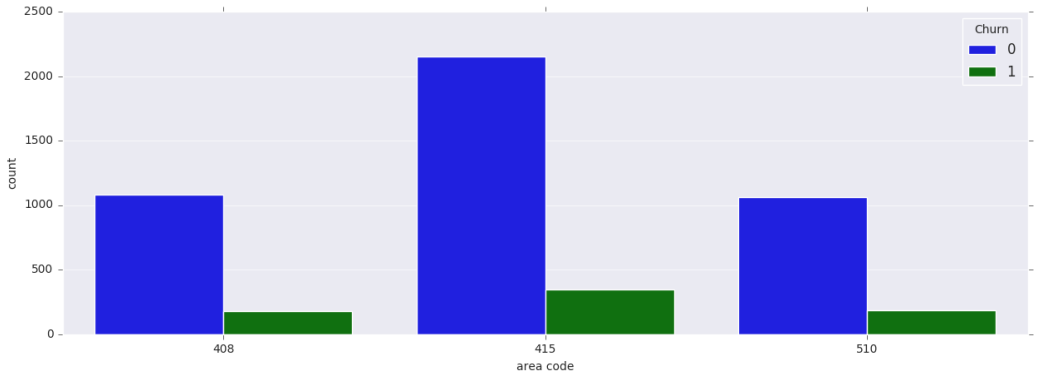

Fig.19: Churn based on area code

Fig.20: Churn based on number of customer service calls

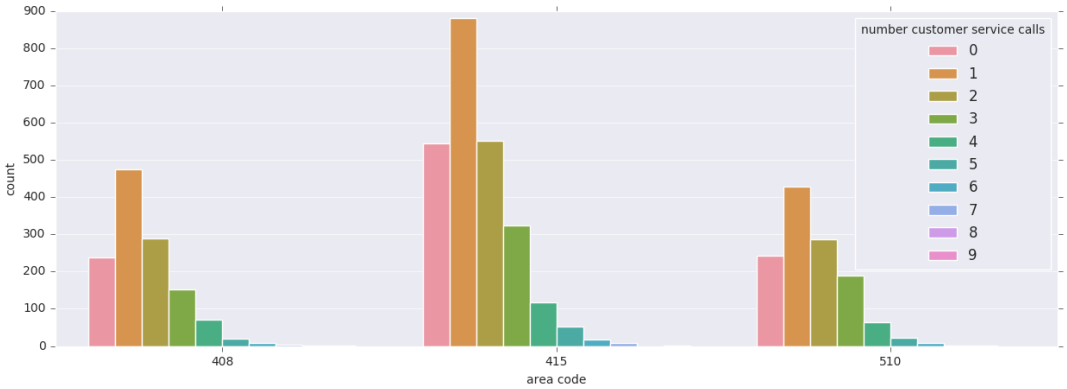

Fig.21: Customer service calls based on area code

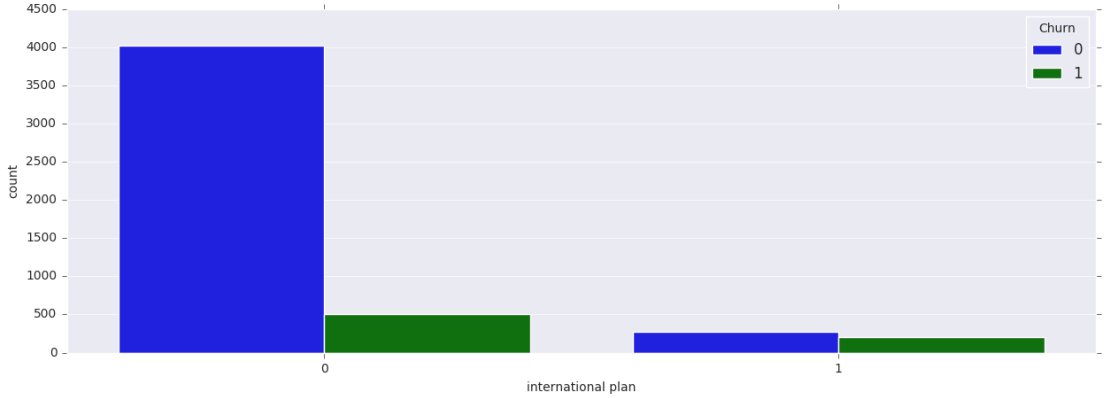

Fig.22: Churn based on International Plan

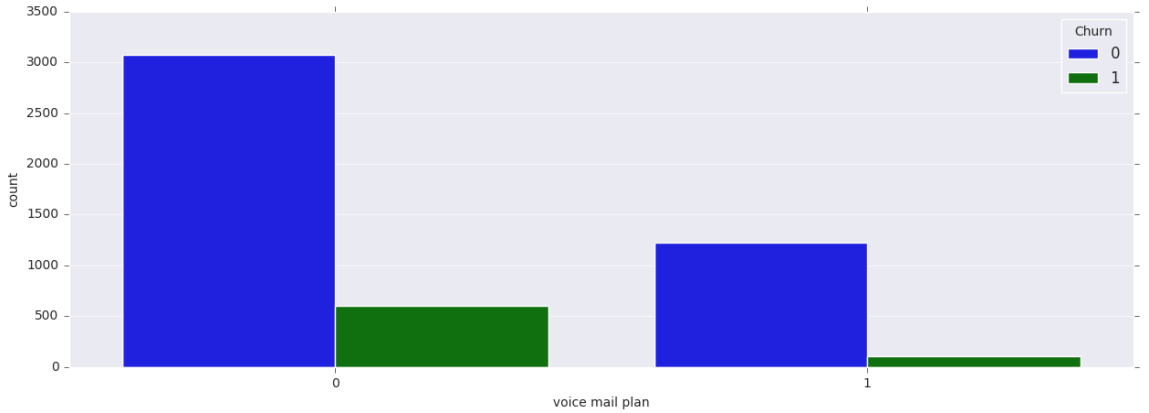

Fig.23: Churn based on Voice mail Plan

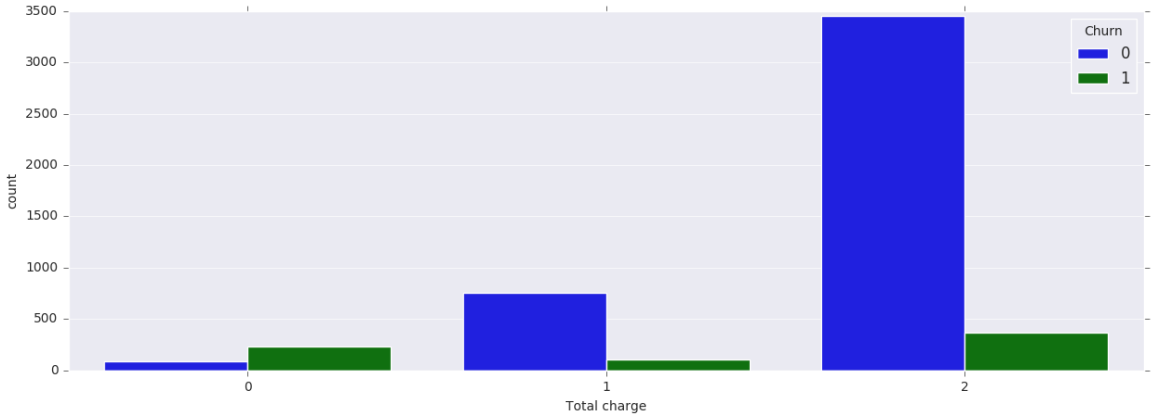

Fig.24: Churn based on Total Charge

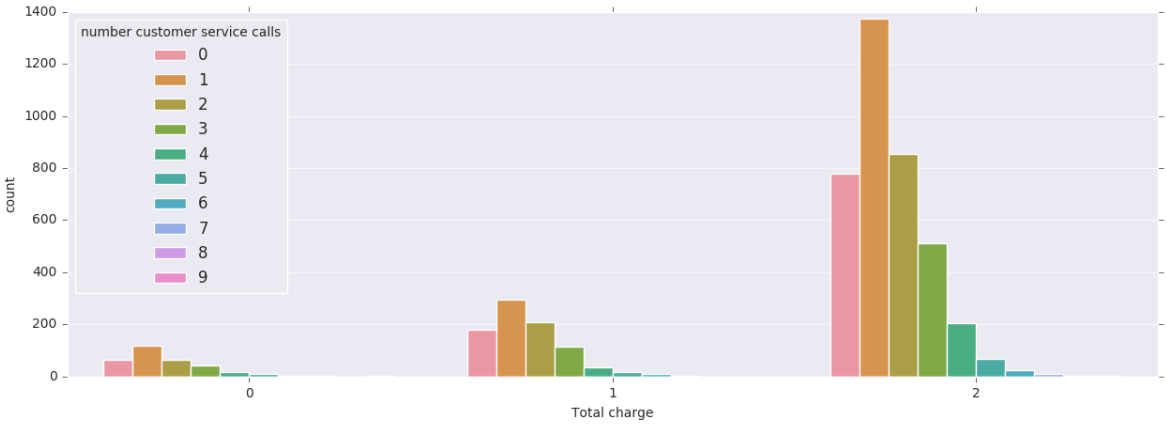

Fig.25: Customer service calls based on Total Charge

4. INFERENCES AND RECOMMENDATIONS

Based on the above plots some inferences and recommendations have been discussed:

4.1. Inferences

- More than 50% of Customers which have made 4 or more customer service calls are going to churn out.
- Approximately half of the customers having an international plan are likely to churn out.
- Customers who are paying having the lowest charge i.e. 25-50 seem highly unsatisfied. Most of them are likely to churn out.
- Since Total charge and total minutes are correlated, it can be inferred that customers who are paying less are not getting proportionate talk-time or minutes from the company in comparison to the customers who are paying more.
- Area code 415 is having the highest churn as well as the highest number of customer service calls. It seems likely that network services in this area code are not good.

4.2. Recommendations

- Customers having high number of customer calls (3 or above) need to be attended on a priority basis.
- The service provider must provide better services and features in their lower range packs so that the customer paying less also get adequate benefits and are satisfied.
- The company should also improve their international plan by either slashing down the prices or providing more benefits at the same price.
- There seems to be a high network issue in the area code 415. The service provider should provide special attention to this area code on priority basis.

5. MACHINE LEARNING MODELS

To predict the churn of customers we have trained the data using the following supervised machine learning algorithms:

- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- NAIVE Bays

After treating the missing values and outliers the data was split into train data (80% of total data) and test data (20% of total data).
All the five algorithms were applied on the training data and then the values of the target variable of test dataset were predicted.

After the prediction of churn, Confusion Matrix for all the models was compared.

- Random Forest was the most accurate with **92%** accuracy.
- Naive Bays, Decision Tree and Random Forest had the lowest False Negative rate of approximately **7.6%.**
- Random Forest achieved the lowest False Positive rate of **11.25%**.
- All the models had an accuracy of more than **85%.**

5.1. Logistic Regression

The error metrics for logistic regression were obtained as follows:

- CONFUSION MATRIX:
  [838 21]
  [92 49]
- ACCURACY: 88.7%
- FALSE NEGATIVE RATE: 8.82%
- FALSE POSITIVE RATE: 30.0%

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.2. Decision Tree

The error metrics for decision tree were obtained as follows:

- CONFUSION MATRIX:
  [801 58]
  [66 75]
- ACCURACY: 87.6%
- FALSE NEGATIVE RATE: 7.61%
- FALSE POSITIVE RATE: 43.61%

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.3. Random Forest

The number of estimators were taken as 300.
The error metrics for random forest were obtained as follows:

- CONFUSION MATRIX:
  [850 9]
  [70 71]
- ACCURACY: 92.1%
- FALSE NEGATIVE RATE: 7.61%
- FALSE POSITIVE RATE: 11.25%

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.4. KNN

The number of nearest neighbours were calculated using elbow method and the optimum k value was found to be 3.

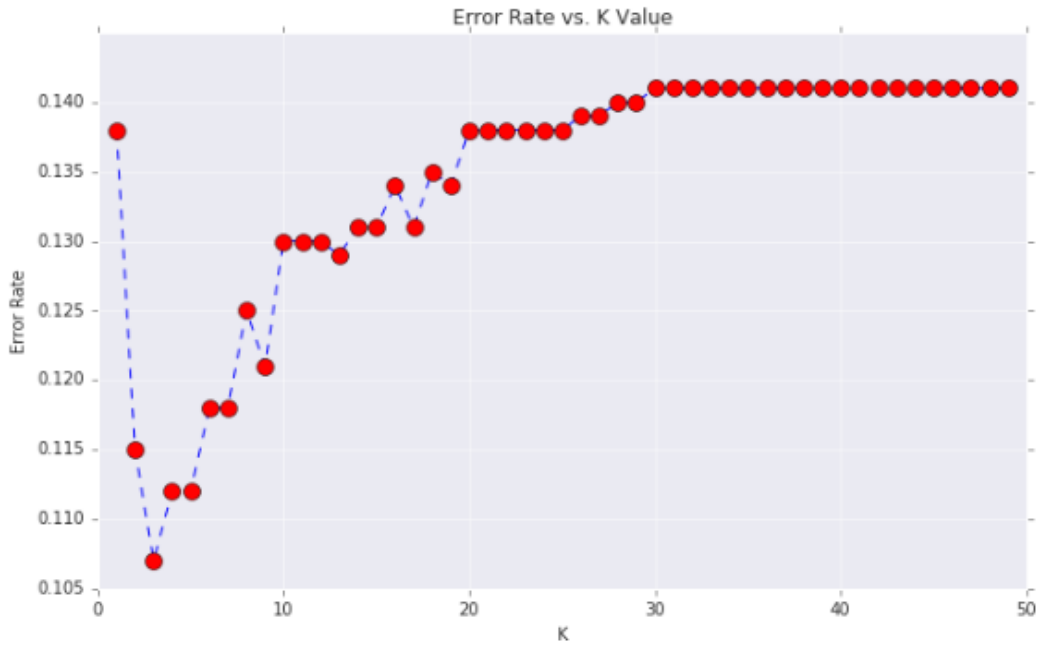

Fig 26: Elbow method for obtaining k-value

The error metrics for KNN were obtained as follows:

- CONFUSION MATRIX:
  [845 14]
  [93 48]

- ACCURACY: 89.3%
- FALSE NEGATIVE RATE: 9.91%
- FALSE POSITIVE RATE: 22.58%

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.5. Naive Bays

The error metrics for NAIVE Bays were obtained as follows:

- CONFUSION MATRIX:
  [804 55]
  [68 73]
- ACCURACY: 87.7%
- FALSE NEGATIVE RATE: 7.79%
- FALSE POSITIVE RATE: 42.97%

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

6. CONCLUSIONS

- Comparing the error metrics of all the machine learning models above, Random Forest was found to be the most optimum machine learning model.
- Random Forest achieved the lowest False Negative rate and False Positive rate of 7.61% and 11.25% respectively. It was also the most accurate with an accuracy of 92.1%.

Thus, for predicting the churn in our dataset we can choose Random Forest with the number of estimators being 300.