Churn Analysis

Albert Egron

Introduction

A telecommunications provider in United States with a customer churn rate of 14.5% would like to investigate characteristics of customer churn and improve retention rate via machine learning. The company has provided 3333 observations of customer data and would like to answer the following questions:

- Who are the customers that are most likely to churn?
- What are the common characteristics for customers who decided to churn?
- How can these customers be identified before they churn?
- What can be done to retain them?

With the dependent variable, Churn, as prediction target, experiments were conducted on 3 classification algorithms using data on nine independent variables provided by the organization. The preferred algorithm in this analysis, J48 Decision Tree, yields a supervised predictive modelling accuracy of 90.1%.

Post-predictive experiments were then performed on the Churn-only portion of the dataset (i.e. 483 of 3333 observations) using K-means clustering, which identified 3 characteristics relevant to customer churn:

- High Day Charges
- Higher than average customer service calls
- High International call charges

The following solutions were recommended to improve customer retention rate:

- Package for day-time calls/day-time and evening-time calls
- A retention plan with an expiry date in 12 months upon receipt of customer service call
- An enhanced Voice Mail Plan with more storage and options
- International call bundle: $3.5 for 20 minutes, 18 cents/min thereafter
- International calls and day calls bundle

R, Weka, and MS Excel were used in this analysis.

Data Preparation

1. Dataset Description

The dataset being analyzed comprises of 21 attributes, including a binary class attribute—namely Churn—indicating whether customer churn is observed or not. Of the 3333 records, 2850 were classified as ‘FALSE’ (i.e. customer attrition not observed) and 483 were classified as ‘TRUE’ (i.e. customer attrition observed). Therefore, the dataset is considered to have an imbalanced class distribution.

1.1. Attribute Type

	Attribute	Type
1	State	Character
2	Account Length	Integer
3	Area Code	Integer
4	Phone	Character
5	International Plan	Character
6	Voice Mail Plan	Character
7	No. of Voice Mail Messages	Integer
8	Total Day Minutes	Numeric
9	Total Day Calls	Integer
10	Total Day Charge	Numeric
11	Total Evening Minutes	Numeric
12	Total Evening Calls	Integer
13	Total Evening Charge	Numeric
14	Total Night Minutes	Numeric
15	Total Night Calls	Integer
16	Total Night Charge	Numeric
17	Total International Minutes	Numeric
18	Total International Calls	Integer
19	Total International Charge	Numeric
20	No. of Call Customer Service	Integer
21	Churn	Character(Class)

1.2. Attribute Summary

Churn		**International Plan**		**Voice Mail Plan**		**State**		**Phone**	
Length:	3333	Length:	3333	Length:	3333	Length:	3333	Length:	3333
Class:	Character	Class:	Character	Class:	Character	Class:	Character	Class:	Character
Mode:	Character	Mode:	Character	Mode:	Character	Mode:	Character	Mode:	Character

No. of Voice Mail Message		**No. of Call Customer Service**		**Account Length**		**Area Code**	
Min:	0.0	Min:	0.0	Min:	1.0	Min:	408.0
1st Qu:	0.0	1st Qu:	1.0	1st Qu:	74.0	1st Qu:	408.0
Median:	0.0	Median:	1.0	Median:	101.0	Median:	415.0
Mean:	8.1	Mean:	1.6	Mean:	101.1	Mean:	437.2
3rd Qu:	20.0	3rd Qu:	2.0	3rd Qu:	127.0	3rd Qu:	510.0
Max:	51.0	Max:	9.0	Max:	243.0	Max:	510.0

Total Day Minutes		**Total Day Calls**		**Total Day Charge**	
Min:	0.0	Min:	0.0	Min:	0.0
1st Qu:	143.7	1st Qu:	87.0	1st Qu:	24.4
Median:	179.4	Median:	101.0	Median:	30.5
Mean:	179.8	Mean:	100.4	Mean:	30.6
3rd Qu:	216.4	3rd Qu:	114.0	3rd Qu:	36.8
Max:	350.8	Max:	165.0	Max:	59.6

Total Evening Minutes		Total Evening Calls		Total Evening Charge	
Min:	0.0	Min:	0.0	Min:	0.0
1st Qu:	166.6	1st Qu:	87.0	1st Qu:	14.2
Median:	201.4	Median:	100.0	Median:	17.1
Mean:	201.0	Mean:	100.1	Mean:	17.1
3rd Qu:	235.3	3rd Qu:	114.0	3rd Qu:	20.0
Max:	363.7	Max:	170.0	Max:	30.9

Total Night Minutes		Total Night Calls		Total Night Charge	
Min:	23.2	Min:	33.0	Min:	1.0
1st Qu:	167.0	1st Qu:	87.0	1st Qu:	7.5
Median:	201.2	Median:	100.0	Median:	9.1
Mean:	200.9	Mean:	100.1	Mean:	9.0
3rd Qu:	235.3	3rd Qu:	113.0	3rd Qu:	10.6
Max:	395.0	Max:	175.0	Max:	17.8

Total International Minutes		Total International Calls		Total International Charge	
Min:	0.0	Min:	0.0	Min:	0.0
1st Qu:	8.5	1st Qu:	3.0	1st Qu:	2.3
Median:	10.3	Median:	4.0	Median:	2.8
Mean:	10.2	Mean:	4.5	Mean:	2.8
3rd Qu:	12.1	3rd Qu:	6.0	3rd Qu:	3.3
Max:	20.0	Max:	20.0	Max:	5.4

2. Exploratory Analysis

No missing data was identified in the dataset; hence no treatment of missing data was performed.

11 attributes were removed as a result of the exploratory analysis.

Outliers were present in most of the attributes, however, treatment of outliers was not applicable in this analysis.

Among the 21 attributes, 3 attributes, namely No. of Voice Mail Messages, Total International Calls and No. of Call Customer Service appeared to be positively skewed, yet treatment of skewed attributes was not deemed necessary in this analysis.

2.1. Attributes Selected for Analysis*

Account Length

International Plan

Voice Mail Plan

No. of Voice Mail Messages

Total Day Charge

Total Evening Charge

Total Night Charge

Total International Charge

No. of Call Customer Service

Churn

*See appendix for detailed attribute breakdowns.

2.2. Reasons for the removal of attributes

The following attributes were to be excluded from the analysis:

- State, Area Code;
- Phone;
- Total Day Minutes, Total Evening Minutes, Total Night Minutes, and Total International Minutes;
- Total Day Calls, Total Evening Calls, Total Night Calls, and Total International Calls

State, Area Code

Of the 3333 observations, 50 states were represented by merely three area codes. Upon further inspection, the area codes were not divided geographically to imply any regional patterns. Therefore, these attributes were removed from further analysis.

Phone

Phone numbers were removed due to its arbitrary nature which do not reflect statistical value.

Total Day Minutes, Total Evening Minutes, Total Night Minutes, and Total International Minutes

4 attributes pertaining to the count of user minutes were removed due to their perfect correlation with their respective charges.

	Total Day Minutes	Total Evening Minutes	Total Night Minutes	Total International Minutes
Total Day Charge	1.00	0.01	0.00	(0.01)
Total Evening Charge	0.01	1.00	(0.01)	(0.01)
Total Night Charge	0.00	(0.01)	1.00	(0.02)
Total International Charge	(0.01)	(0.01)	(0.02)	1.00

Total Day Calls, Total Evening Calls, Total Night Calls, and Total International Calls

Feature selection by Information Gain analysis on Weka indicated attributes pertaining to “calls” offer lowest information gain. Therefore, these 4 attributes were excluded from the analysis accordingly.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.077 +- 0.004     1   +- 0       6 Total Day Charge
 0.05  +- 0.002     2   +- 0      13 No of Calls Customer Service
 0.037 +- 0.002     3   +- 0       2 Inter Plan
 0.008 +- 0.001     4.7 +- 1       3 VoiceMail Plan
 0.008 +- 0.001     4.9 +- 0.54    4 No of Vmail Mesgs
 0.007 +- 0         6.3 +- 0.78   11 Total Int Calls
 0.007 +- 0.001     6.5 +- 1.43   12 Total Int Charge
 0.005 +- 0.002     7.8 +- 0.87    8 Total Evening Charge
 0     +- 0         9   +- 0       7 Total Evening Calls
 0     +- 0         9.8 +- 0.6     5 Total Day calls
 0     +- 0        11   +- 0       9 Total Night Calls
 0     +- 0        12   +- 0      10 Total Night Charge
 0     +- 0        13   +- 0       1 Account Length
```

2.3. Outliers

Of the 3333 observations, the Churn class, represented by 483 observations, accounts for 14% of the whole dataset. The table below indicates that the outliers have a significant representation in churned customers. In light of this, treatments of outliers such as removal or replacement by mean/median were not deemed applicable in this analysis.

Attribute	No. of Outlier	No. of Churner	% churners	% Churn Class in dataset
Account Length	18	5	28%	14%
Total Day Charge	25	12	48%	14%
Total Evening Charge	24	5	21%	14%
Total Night Charge	30	2	7%	14%
Total International Charge	49	6	12%	14%
No. of Call Customer Service	267	138	52%	14%

Predictive Modeling/Classification

1. Experiment

Following data preparation, iterative experiments were performed in the following classification algorithms on Weka:

1. J48 Decision Tree
2. Random Forest
3. Naïve Bayes

Cost Sensitive Classifier was implemented in each classification algorithms to balance the imbalanced data identified in the exploratory analysis. The Cost Sensitive Classifier offers more manual control than the Class

Balancer. It allows you to apply a penalty of your choice to the incorrectly classified instances to the minority class thus allowing more control, therefore creating a more balanced dataset and offering more potential accuracy.

For each classification algorithms, three strategies for dataset split were employed: simple training, 10-fold cross validation, and 66% split. Recall, Precision, and Accuracy for False Churners and True Churners were recorded respectively.

Of the 3 strategies, Random Forest generates the highest Accuracy, Recall and Precision. An accuracy of 100% on Training set generated by Random Forest suggests the possibility of over-fitting.

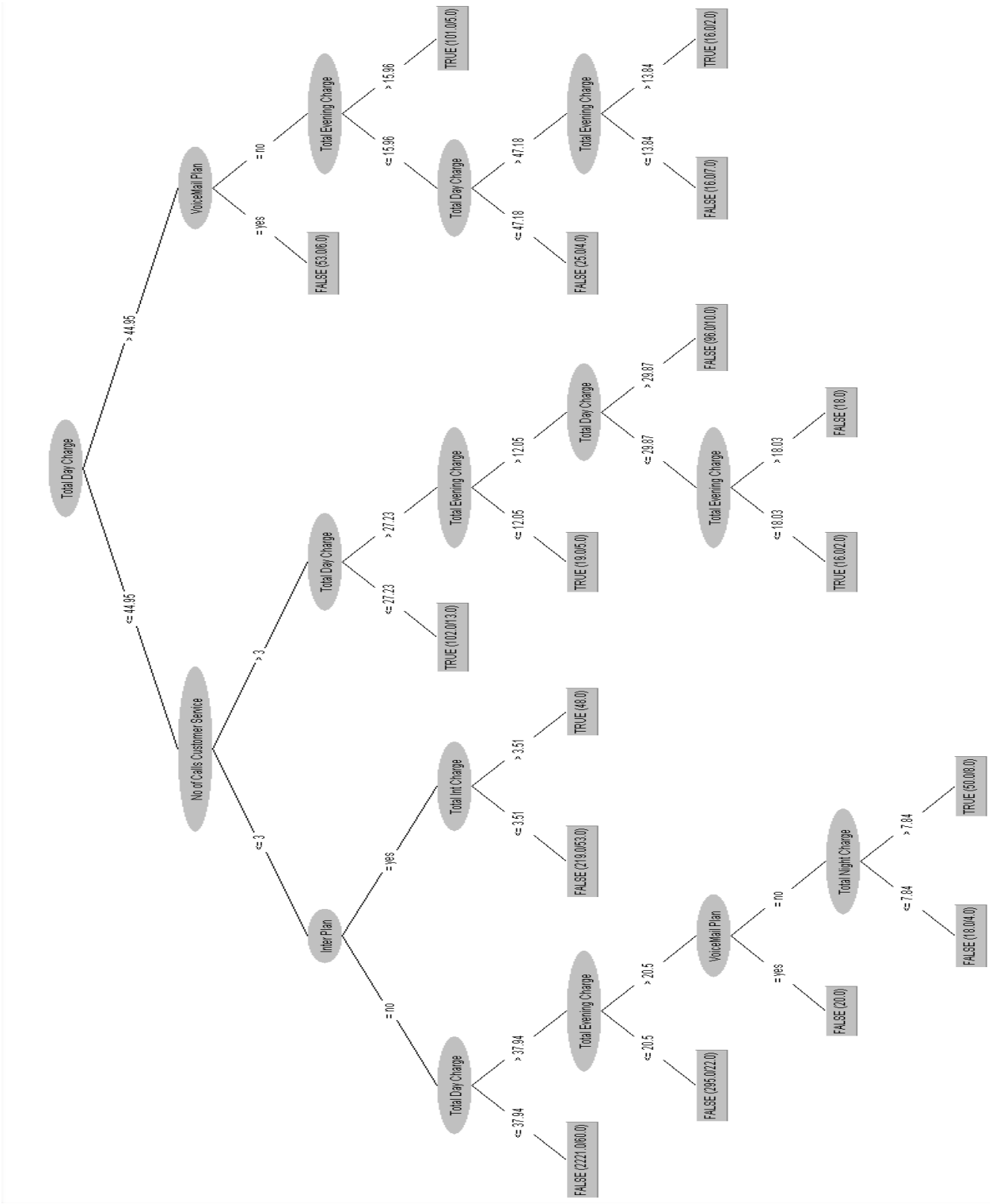
Based on the results, the best classifier is J48 Decision Tree with 10-fold cross validation for its ability to generate neither the highest nor the lowest accuracy among the 3 methods.

J48 Decision Tree	Recall		Precision		Accuracy
	F	T	F	T	
10-fold Cross	96.5	67.3	94.6	76.5	90.1
Training	94.9	96.3	99.3	76.2	91.6
66% split	96.5	62.8	93.9	75.2	89.3

Random Forest	Recall		Precision		Accuracy
	F	T	F	T	
10-fold Cross	97.9	70	95.1	85.1	93.9
Training	100	100	100	100	100
66% split	98.7	68.9	94.9	89.7	94.3

Naïve Bayes	Recall		Precision		Accuracy
	F	T	F	T	
10-fold Cross	93.7	49.3	91.6	56.9	87.2
Training	93.8	50.1	91.7	57.8	87.4
66% split	93.3	49.4	91.6	55.5	86.9

J48 Decision Tree



*Decision tree was modified with “minObj” in Weka set to 50 to fit on the report

Post-prediction Analysis

1.Experiment

Following Predictive Modelling, experiments were performed on K-means clustering on Weka using the Churn-only portion of the dataset (i.e. 483 observations) on 66% split. 3 clusters were identified as optimal number of clusters.

2.Results

The customers identified by the K-means algorithm are:

1. High Value Customers
2. International Callers
3. Deal Seekers

Attribute	Average	Deal Seekers	High Value Customers	International Callers
Account Length	101.80	105.22	89.47	112.76
# of Voicemail Messages	4.65	9.34	0.56	4.52
Total Day Charge	35.48	24.15	43.44	38.09
Total Evening Charge	17.91	15.05	19.84	18.68
Total Night Charge	9.24	8.77	10.13	8.69
Total International Charge	2.91	2.73	2.46	3.63
# Calls to Customer Service	2.29	4.13	1.31	1.48

Conclusions and Recommendations

1.Conclusions

A. K-means clustering: Three highest indicators of Churn

i. Total Daily charge ($43.4 *vs* mean of $35.4) with higher than average Total Evening Charge
ii. Customer service calls (4.1 calls *vs* mean of 2.9 calls) with higher than average Voicemail Messages
iii. Total International Charge ($3.6 vs mean of $2.9) with a higher than average Total Day Charge

B. K-means: Voice Mail Plan is a secondary determinant for customer retention.

C. Differential pricing of airtime as follows:

		cents/min
i.	Day	17.0
ii.	Evening	8.5
iii.	Night	4.5
iv.	International	27.0

2.Recommendations

A. High Value Customers (First Priority)
   a. Deploy a proactive retention approach
   b. Offer a bundle rate for day calls
   c. Offer a bundle rate for day and evening calls
   d. Goal is to maintain revenue from customers per month but add value by offering more minutes

B. International Callers (Second Priority)
   a. Deploy a proactive retention approach
   b. Offer an International Calling bundle
      i. $3.50 for 20 Minutes of International Calls
      ii. More minutes at same cost= More Value
      iii. Subsidized International Rate beyond 20 minutes
   c. Offer a Hybrid International+Day Calling bundle
      i. Based on higher than average Day Calls
      ii. Offering more value for high worth customer

C. Deal Seekers (Third Priority)
   a. Deploy a reactive retention approach
      i. They are likely to call asking for a deal before churning
   b. Offer a Loyalty Retention Package
      i. Offer expires after a time period
      ii. Buys company time and opportunity to turn Deal Seekers into High Value Customers
      iii. Offer Enhanced Voicemail with more storage and options

Appendix

Account Length

$stats
[1] 1 74 101 127 205

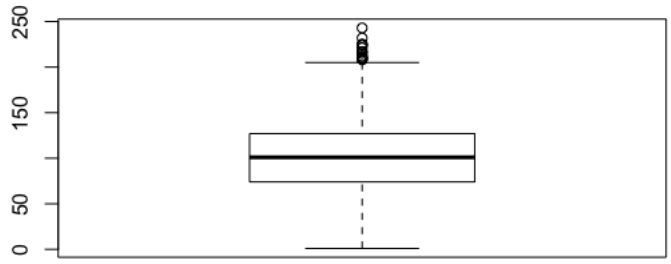
$n
[1] 3333

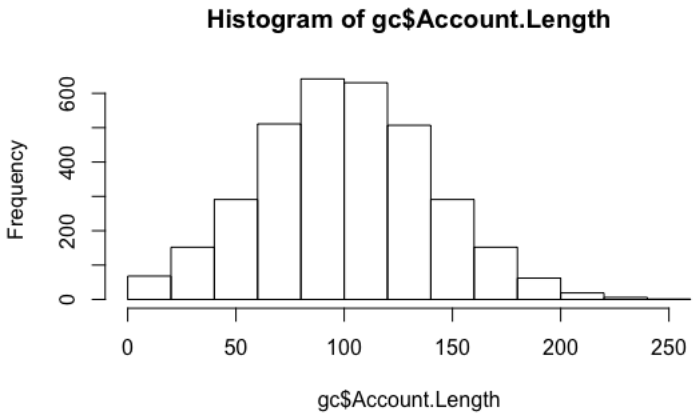
$conf
[1] 99.54951 102.45049

$out
 [1] 208 215 209 224 243 217 210 212 232 225 225 224 212 210 217 209 221 209

Variable looks approx. normally distributed

Total # of outliers = 18 (5 are churners 28%)





No. of Voice Mail messages

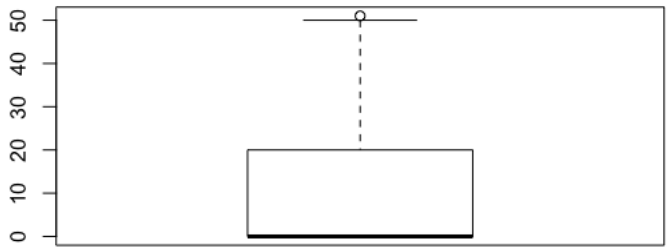
$stats
[1] 0 0 0 20 50

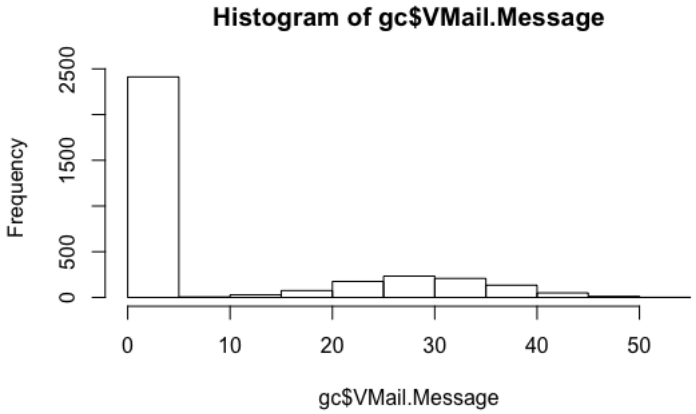
$n
[1] 3333

$conf
[1] -0.5473554 0.5473554

$out
[1] 51

Total number of customers with VM messages = 922 (28%) and out of those only 80 (9%) are churners
Total number of customers with 0 VM messages = 2411 (72%)





Total Day Minutes

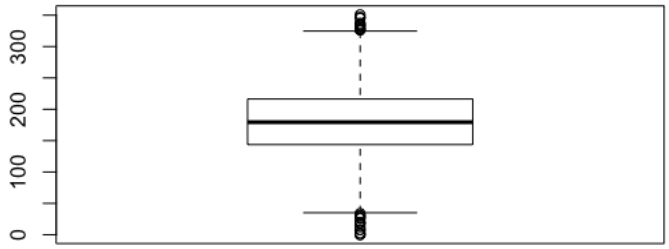
$stats
[1] 35.1 143.7 179.4 216.4 324.7

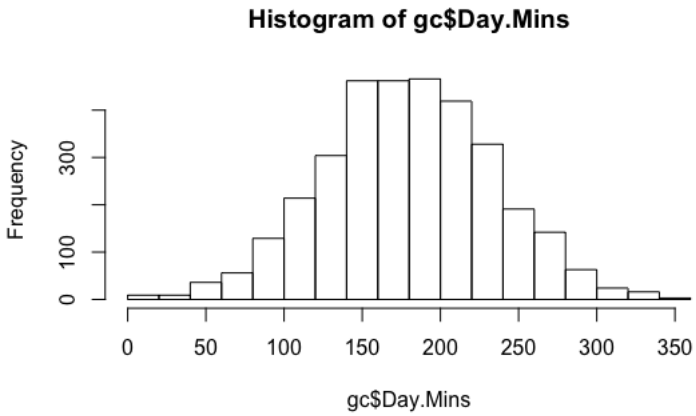
$n
[1] 3333

$conf
[1] 177.4104 181.3896

$out
 [1] 332.9 337.4 326.5 350.8 335.5 30.9 34.0 334.3 346.8 12.5 25.9   0.0   0.0 19.5
[15] 329.8   7.9 328.1 27.0 17.6 326.3 345.3   2.6   7.8 18.9 29.9

Data looks approx. normally distributed





Total Day Calls

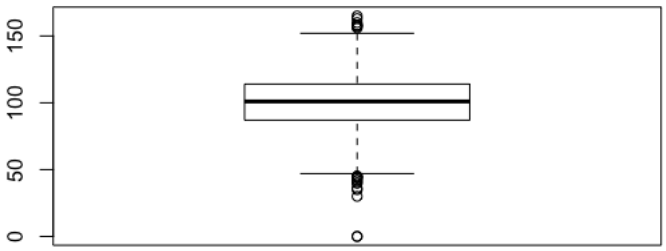
$stats
[1] 47 87 101 114 152

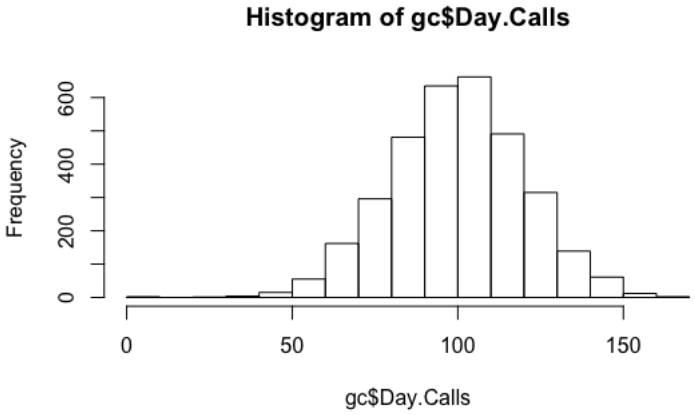
$n
[1] 3333

$conf
[1] 100.2611 101.7389

$out
 [1] 158 163 36 40 158 165 30 42   0 45   0 45 160 156 35 42 158 157 45 44 44
[22] 44 40

Data is approx. normal





Total Day Charge

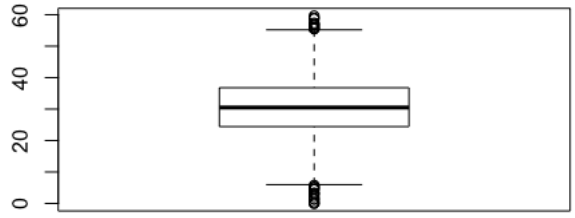
$stats
[1] 5.97 24.43 30.50 36.79 55.20

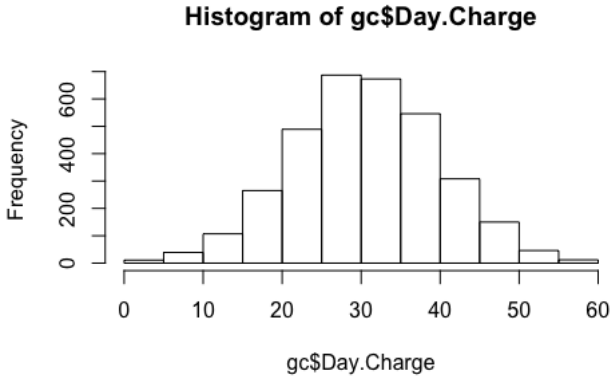
$n
[1] 3333

$conf
[1] 30.16173 30.83827

$out
 [1] 56.59 57.36 55.51 59.64 57.04 5.25 5.78 56.83 58.96 2.13 4.40 0.00 0.00 3.32 56.07
[16] 1.34 55.78 4.59 2.99 55.47 58.70 0.44 1.33 3.21 5.08

Total # of outliers = 25 (12 are churners 48%)





Total Evening Minutes

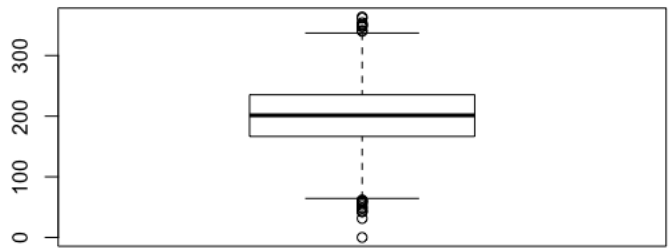
$stats
[1] 64.3 166.6 201.4 235.3 337.1

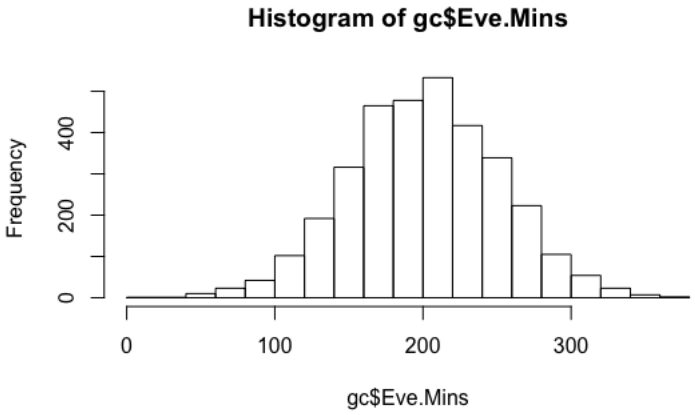
$n
[1] 3333

$conf
[1] 199.5198 203.2802

$out
 [1] 61.9 348.5 351.6 31.2 350.5 42.2 347.3 58.9 43.9 52.9 42.5 60.8 58.6 56.0
[15] 48.1 60.0 350.9 49.2 339.9 361.8 354.2 363.7   0.0 341.3

Data is approx. normal





Total Evening Calls

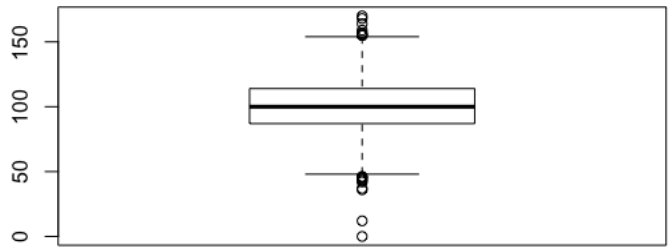
$stats
[1] 48 87 100 114 154

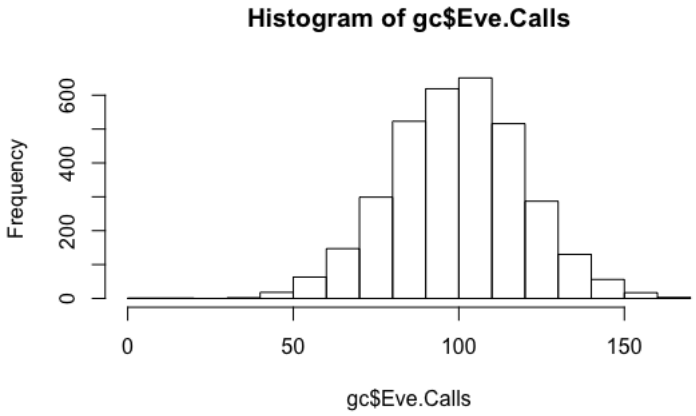
$n
[1] 3333

$conf
[1] 99.26107 100.73893

$out
 [1] 164 46 168 42 37 12 157 155 45 36 156 46 44 155 46 43   0 155 159 170

Data is approx. normal





Total Evening Charge

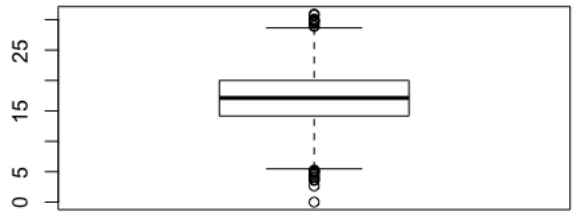
$stats
[1] 5.47 14.16 17.12 20.00 28.65

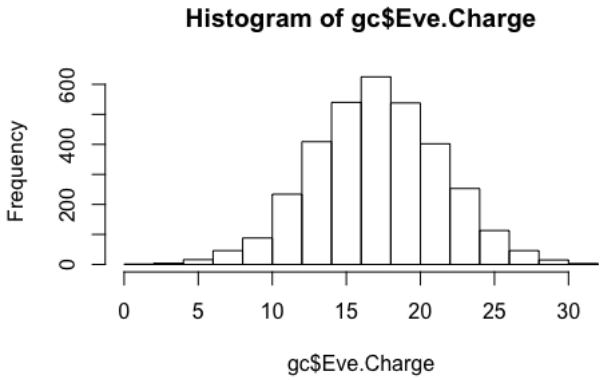
$n
[1] 3333

$conf
[1] 16.96017 17.27983

$out
 [1] 5.26 29.62 29.89 2.65 29.79 3.59 29.52 5.01 3.73 4.50 3.61 5.17 4.98 4.76 4.09
[16] 5.10 29.83 4.18 28.89 30.75 30.11 30.91 0.00 29.01

Total # of outliers = 24 (5 are churners 21%)





Total Night Minutes

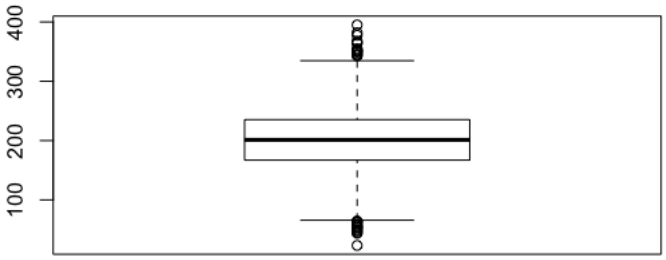
$stats
[1] 65.7 167.0 201.2 235.3 334.7

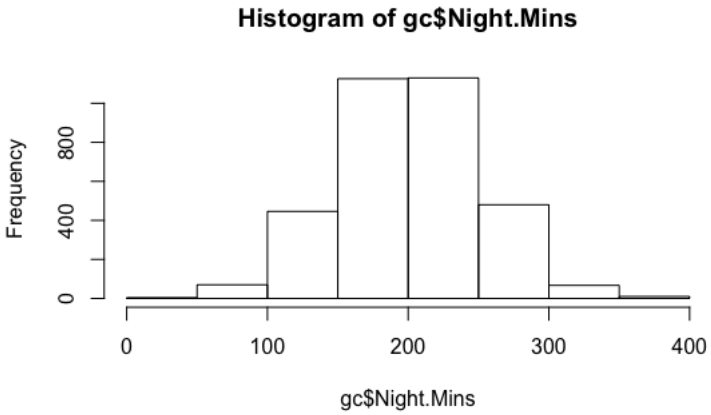
$n
[1] 3333

$conf
[1] 199.3308 203.0692

$out
 [1] 57.5 354.9 349.2 345.8 45.0 342.8 364.3 63.3 54.5 50.1 43.7 349.7 352.5 23.2
[15] 63.6 381.9 377.5 367.7 56.6 54.0 64.2 344.3 395.0 350.2 50.1 53.3 352.2 364.9
[29] 61.4 47.4

Data is approx. normal





Total Night Calls

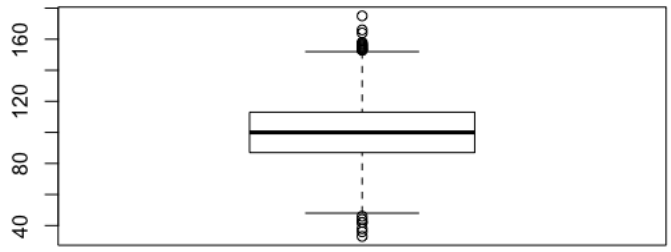
$stats
[1] 48 87 100 113 152

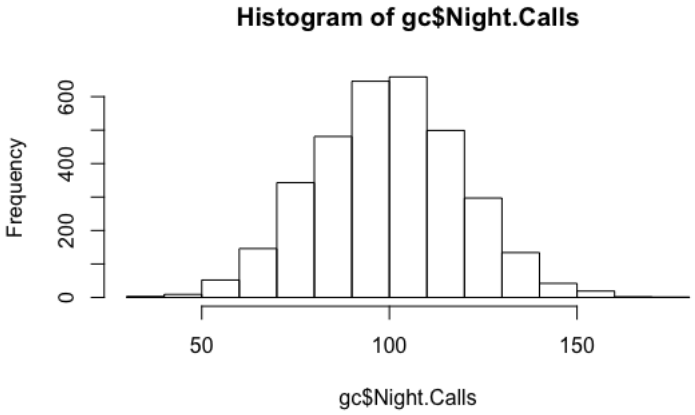
$n
[1] 3333

$conf
[1] 99.28844 100.71156

$out
 [1] 46 42 44 42 153 175 154 158 155 157 157 154 153 166 33 155 156 38 36 156 164
[22] 153

Data is approx. normal





Total Night Charge

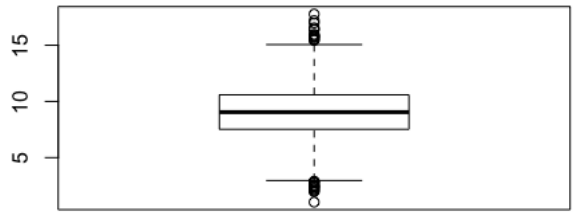
$stats
[1] 2.96 7.52 9.05 10.59 15.06

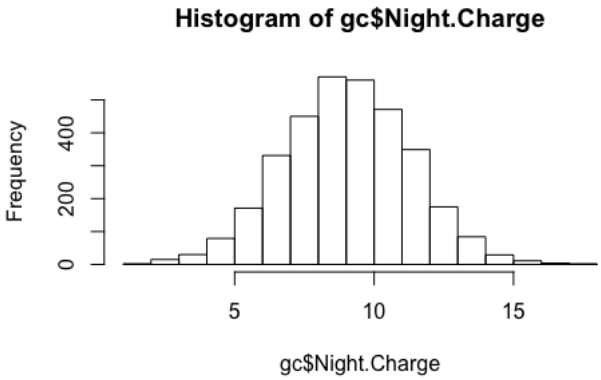
$n
[1] 3333

$conf
[1] 8.965981 9.134019

$out
 [1] 2.59 15.97 15.71 15.56 2.03 15.43 16.39 2.85 2.45 2.25 1.97 15.74 15.86 1.04 2.86
[16] 17.19 16.99 16.55 2.55 2.43 2.89 15.49 17.77 15.76 2.25 2.40 15.85 16.42 2.76 2.13

Total # of outliers = 30 (2 are churners 7%)





Total International Minutes

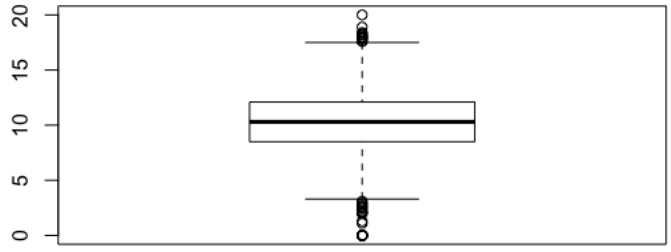
$stats
[1] 3.3 8.5 10.3 12.1 17.5

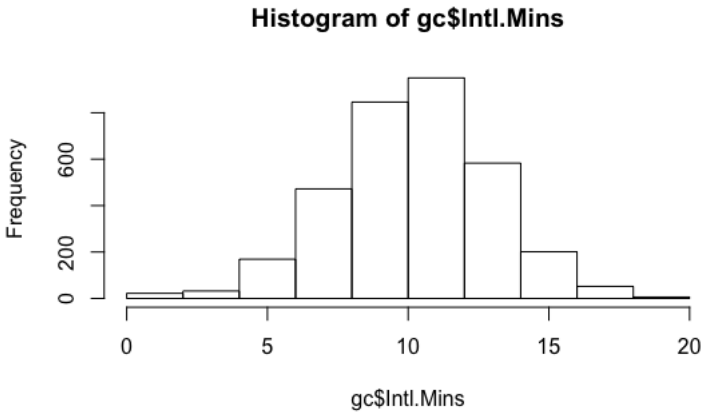
$n
[1] 3333

$conf
[1] 10.20148 10.39852

$out
 [1] 20.0 0.0 17.6 2.7 18.9 0.0 18.0 2.0 0.0 18.2 0.0 0.0 1.3 0.0 0.0 0.0
[17] 2.2 18.0 0.0 17.9 0.0 18.4 2.0 17.8 2.9 3.1 17.6 2.6 0.0 0.0 18.2 0.0
[33] 18.0 1.1 0.0 18.3 0.0 0.0 2.1 2.9 2.1 2.4 2.5 0.0 0.0 17.8

Data is approx. normal





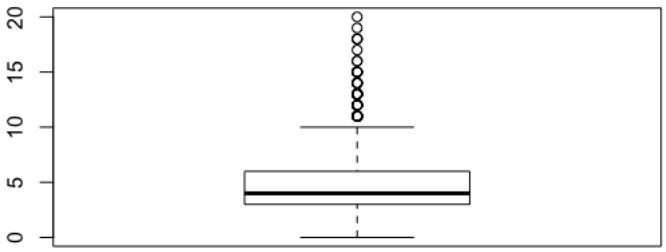
Total International Calls

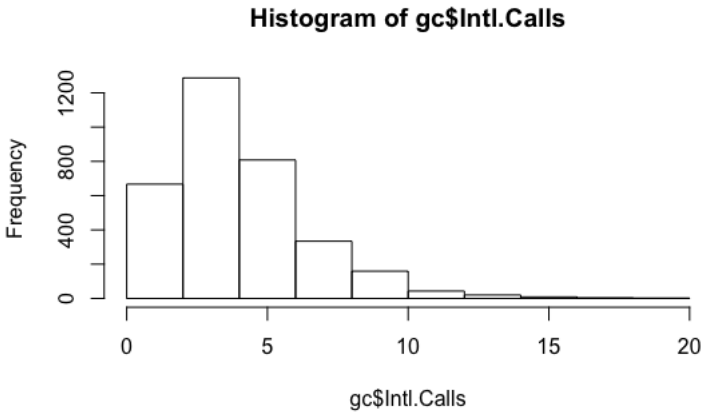
$stats
[1] 0 3 4 6 10

$n
[1] 3333

$conf
[1] 3.917897 4.082103

$out
 [1] 19 15 11 12 13 11 12 11 13 12 11 11 18 11 12 13 12 12 11 15 13 15 11 11 14 13 11 13
[29] 13 12 11 14 15 18 12 13 11 14 11 12 14 15 12 11 16 11 11 11 11 15 11 14 11 11 12 13
[57] 11 11 16 13 11 13 11 15 11 12 13 18 12 12 12 11 13 11 13 14 20 17





Total International Charge

$stats
[1] 0.89 2.30 2.78 3.27 4.67

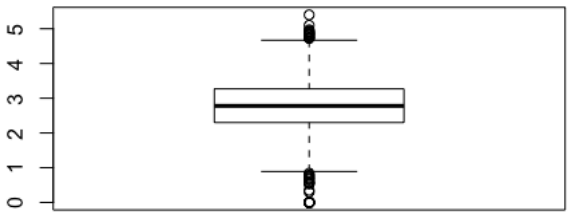
$n
[1] 3333

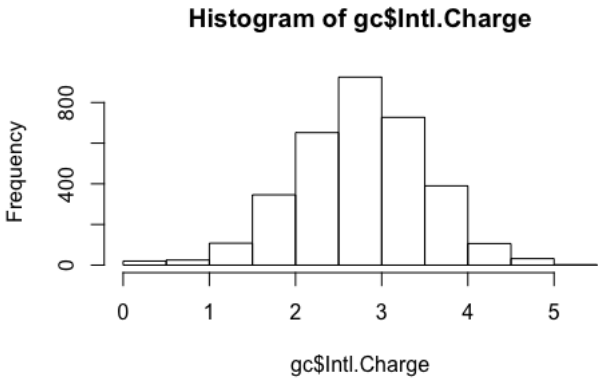
$conf
[1] 2.753453 2.806547

$out
 [1] 5.40 0.00 4.75 0.73 5.10 0.00 4.86 0.54 0.00 4.73 4.73 4.91 0.00 0.00 0.35 0.00 0.00 0.00
[19] 0.59 4.86 0.00 4.83 0.00 4.97 0.54 4.81 0.78 0.84 4.75 0.70 0.00 0.00 4.91 0.00 4.86 0.30

[37] 0.00 4.94 0.00 0.00 0.57 0.78 4.73 0.57 0.65 0.68 0.00 0.00 4.81

Total # of outliers = 49 (6 are churners 12%)





Number of Call Customer Service

$stats
[1] 0 1 1 2 3

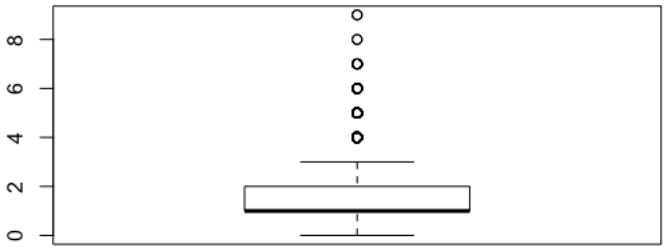
$n
[1] 3333

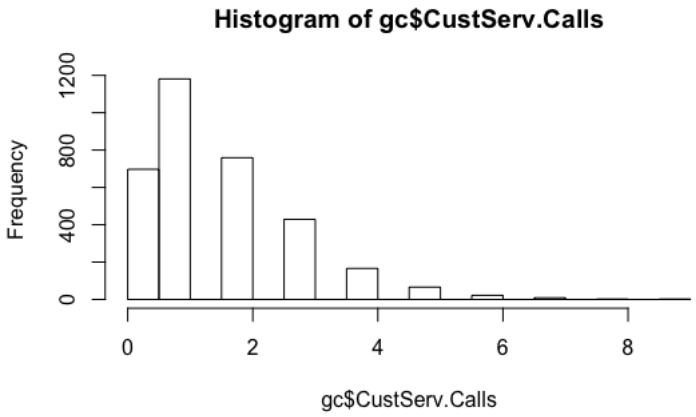
$conf
[1] 0.9726322 1.0273678

$out
[1] 4 4 4 5 5 5 4 4 4 4 4 4 4 4 4 4 5 5 4 5 4 4 5 4 4 4 4 4 5 4 4 7 4 4 4 4 4 5 4 4 4
[42] 4 4 5 4 7 4 9 5 4 4 5 4 4 5 5 4 6 4 6 5 5 5 6 5 4 4 5 4 4 7 4 6 5 4 4 4 6 4 4 5 4
[83] 4 4 4 4 4 5 5 6 5 4 4 4 5 4 4 4 4 5 5 4 4 4 4 6 4 5 4 6 4 4 4 4 4 4 4 4 4 6 4 4 4
[124] 4 8 4 4 5 4 4 4 6 5 5 7 4 4 5 4 4 5 4 4 5 7 4 4 5 7 4 4 4 4 8 6 4 4 5 5 5 4 4 5 4

[165] 4 4 4 4 4 4 4 4 4 5 6 4 5 4 4 5 5 4 6 4 4 4 9 6 4 5 5 4 6 4 4 5 4 4 4 5 5 6 4 5 4
[206] 4 4 4 5 4 4 4 5 4 5 6 4 4 5 4 4 4 5 4 4 4 4 4 5 7 6 5 6 7 5 5 4 6 4 4 4 4 5 6 7 4
[247] 4 4 5 5 5 4 4 4 5 6 5 5 4 4 4 4 4 4 4 4 5

Total # of outliers = 267 (138 are churners 52%)





International Plan, Voice Mail Plan and Churn

International Plan

no	yes
3010	323
90%	10%

Voice Mail Plan

no	yes
2411	922
72%	28%

Churn

False. True.

2850	483
86%	14%

Churn. = False = 2850

	gc.VMail.Plan		Total	% of 2850
gc.Int.l.Plan	no	yes		
no	1878	786	2664	93%
yes	130	56	186	**7%**
Total	2008	842		
% of 2850	70%	30%		

Churn. = True = 483

	gc.VMail.Plan		Total	% of 483
gc.Int.l.Plan	no	yes		
no	302	44	346	72%
yes	101	36	137	**28%**
Total	403	80		
% of 483	83%	17%		

J48 Decision Tree with 10 Fold Cross Validation (2X Cost Sensitive Matrix Balancer)

```
Cost Matrix
 0 1
 2 0


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3075               92.2592 %
Incorrectly Classified Instances       258                7.7408 %
Kappa statistic                          0.6713
Mean absolute error                      0.1195
Root mean squared error                  0.2584
Relative absolute error                 48.1743 %
Root relative squared error             73.3954 %
Total Number of Instances             3333

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965    0.327    0.946      0.965    0.955      0.673    0.844     0.944     FALSE
                 0.673    0.035    0.765      0.673    0.716      0.673    0.844     0.708     TRUE
Weighted Avg.    0.923    0.285    0.919      0.923    0.921      0.673    0.844     0.910

=== Confusion Matrix ===

    a    b   <-- classified as
 2750  100 |    a = FALSE
  158  325 |    b = TRUE
```

Random Forest 10 Fold Cross Validation (2X Cost Sensitive Matrix Balancer)

```
Cost Matrix
 0 1
 2 0


Time taken to build model: 0.88 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3129               93.8794 %
Incorrectly Classified Instances       204                6.1206 %
Kappa statistic                          0.7333
Mean absolute error                      0.1218
Root mean squared error                  0.2299
Relative absolute error                 49.1185 %
Root relative squared error             65.3104 %
Total Number of Instances             3333

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.300    0.951      0.979    0.965      0.738    0.894     0.965     FALSE
                 0.700    0.021    0.851      0.700    0.768      0.738    0.894     0.824     TRUE
Weighted Avg.    0.939    0.260    0.936      0.939    0.936      0.738    0.894     0.944

=== Confusion Matrix ===

    a    b   <-- classified as
 2791   59 |    a = FALSE
  145  338 |    b = TRUE
```

Naïve Bayes 10 Fold Cross Validation (2X Cost Sensitive Matrix Balancer)

```
Cost Matrix
 0 1
 2 0


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2908               87.2487 %
Incorrectly Classified Instances       425               12.7513 %
Kappa statistic                          0.455
Mean absolute error                      0.2296
Root mean squared error                  0.3153
Relative absolute error                 92.5885 %
Root relative squared error             89.5563 %
Total Number of Instances             3333

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.937    0.507    0.916      0.937    0.926      0.457    0.841     0.956     FALSE
                 0.493    0.063    0.569      0.493    0.528      0.457    0.841     0.518     TRUE
Weighted Avg.    0.872    0.443    0.866      0.872    0.869      0.457    0.841     0.893

=== Confusion Matrix ===

    a    b   <-- classified as
 2670  180 |    a = FALSE
  245  238 |    b = TRUE
```

K-Means Clustering

```
kMeans
======

Number of iterations: 30
Within cluster sum of squared errors: 199.68240256562984

Initial starting points (random):

Cluster 0: 144,no,no,0,47.35,20.46,5.4,3.13,1
Cluster 1: 42,no,no,0,51.66,19.74,6.62,1.57,1
Cluster 2: 55,no,no,0,48.57,19.63,10.38,4,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                     Full Data         0          1          2
                                (318.0)    (104.0)    (116.0)     (98.0)
=============================================================================
Account Length                101.7956   105.2212    89.4655   112.7551
Inter Plan                          no         no         no         no
VoiceMail Plan                      no         no         no         no
No of Vmail Mesgs              4.6509     9.3365     0.5603     4.5204
Total Day Charge              35.4812    24.1459    43.4406    38.0893
Total Evening Charge          17.9138    15.0479    19.8394    18.6759
Total Night Charge             9.2423     8.7746     10.129      8.689
Total Int Charge               2.9092     2.7338     2.4609     3.6259
No of Calls Customer Service   2.2862     4.1346     1.3103     1.4796



Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0       55 ( 33%)
1       65 ( 39%)
2       45 ( 27%)
```

R code

```
tc <- read.csv("/Users /Documents/Ryerson/CIND 119/project/provided weka data sets/churn.csv", header = T,
stringsAsFactors = F, na.strings = c("","NA"))

str(tc)
summary(tc)

boxplot(tc$)
boxplot.stats(tc$)

cor(tc$Day.Mins, tc$Day.Charge)
cor(tc$Eve.Mins, tc$Eve.Charge)
cor(tc$Night.Mins, tc$Night.Charge)
cor(tc$Intl.Mins, tc$Intl.Charge)

non_numberic <- data.frame(tc$Int.l.Plan, tc$VMail.Plan, tc$Churn.)
table_non_numneric <- table(non_numberic)
table_non_numneric

non_numberic_1 <- data.frame(tc$Int.l.Plan)
table_1 <- table(non_numberic_1)
table_1
```