Customer Churn Prediction

# Members

Ajay Herod, ajay.herod@ryerson.ca

Summary: A telecommunications company would like to study and analyze customer churn, based on user’s usage and experience. The predictive data modeling used to predict customer churn are classification from decision tree and naïve bayes. The results show that the classification from the decision tree model outperforms the naïve bayes model by a small margin. This is shown with a 92 percent accuracy from the decision tree model, while the naïve bayes model has an 89 percent accuracy. The decision tree also slightly outperforms in the other categories recall and precision. The post-predictive analysis tells us that customer churn can be accurately predicted based on the customer attributes; customer service calls, total day minutes, total evening minutes, total night minutes and voice mail plan. Based on the model customer churn is higher among customers that have a voice mail plan, more than three customer service calls, or above average minutes in either time of day. The model tells us that customer service provides a positive customer experience resulting in more customers churning. Also based on the data given I can also see that customers are linearly charged per minute. For the company I recommend two plans. First, increasing customer service calls. Second, that they create a phone bundle plan that includes voice mail and charges customers on a log scale per minute. This plan will create more customers likely to churn and provide a discount to them.

# Workload Distribution

|  |  |
| --- | --- |
| Member Name | List of Tasks Performed |
| Ajay Herod | Everything. |

# Data Preparation

The first step in data preparations is to understand it. Looking at the data there are 3333 customers and 21 variables.

Data Attributes:

State : character

Account.Length : integer

Area.Code : integer

Phone : character

Int.I.Plan : character

VMail.Plan : character

VMail.Message : integer

Day.Mins : number

Day.Calls : integer

Day.Charge : number

Eve.Mins : number

Eve.Calls : integer

Eve.Charge : number

Night.Mins : number

Night.Calls : integer

Night.Charge : number

Intl.Mins : number

Intl.Calls : integer

Intl.Charge : number

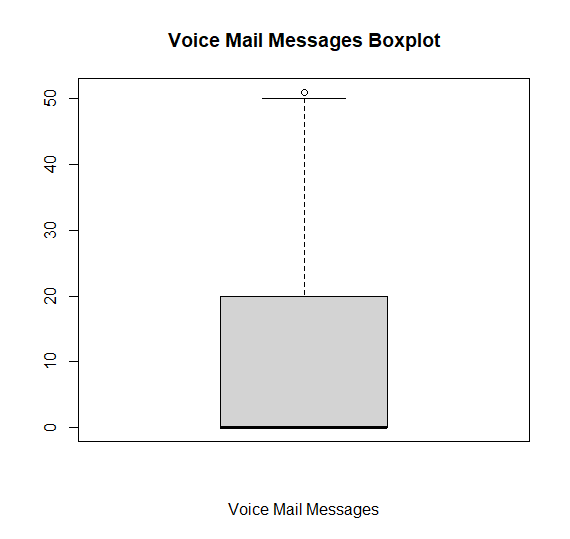
CustServ.Calls : integer

Churn : character

Second, I checked for missing data, which was complete. Next, I began with data cleaning. Based on the data attributes, we can eliminate Account Length, State and Phone number. These attributes are unique and are not characteristics of a customer in the data modeling for customer churn. Fourth, I looked at data preparation. To create the models, I converted the churn attribute into a binary variable that our models will be able to recognize. Last, I viewed each attributes min, max and mean using their boxplot. Then checked for any correlation between the attributes that may give us any information.

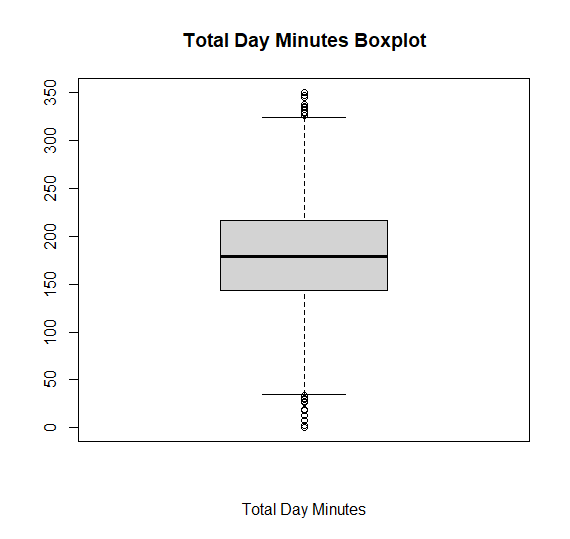
The voice mail messages attribute shows:

* Min ≃ 0
* Max ≃ 50
* Mean ≃ 0



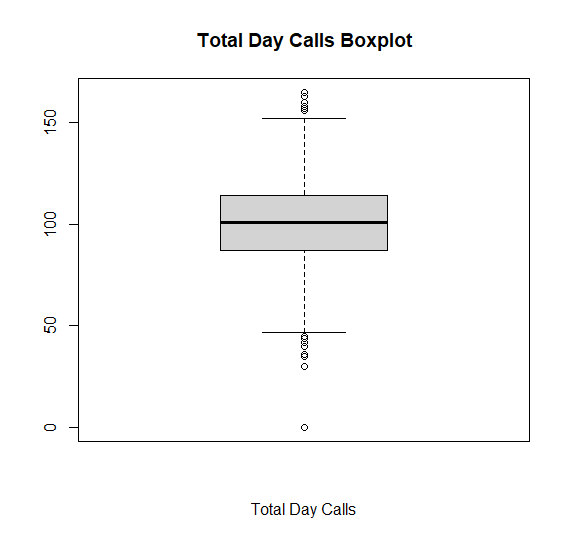
The total day minutes attribute shows:

* Min ≃ 0
* Max ≃ 350
* Mean ≃ 175



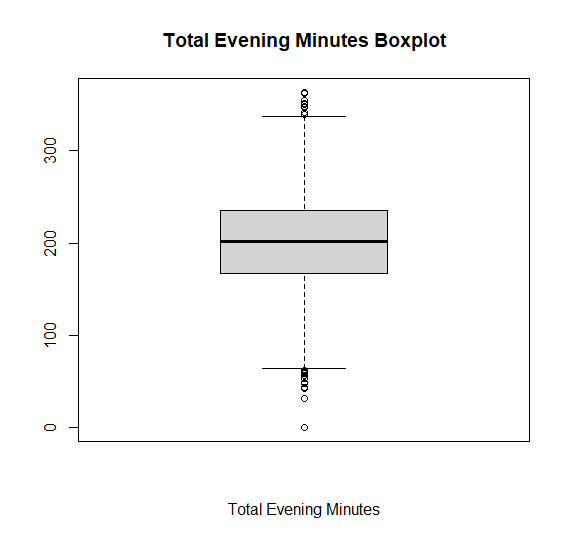
The total day calls attribute shows:

* Min ≃ 0
* Max ≃ 160
* Mean ≃ 100



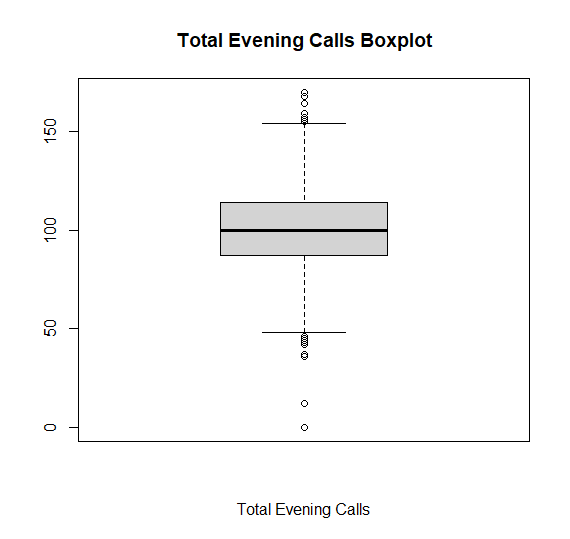
The total evening minutes attribute shows:

* Min ≃ 0
* Max ≃ 350
* Mean ≃ 200



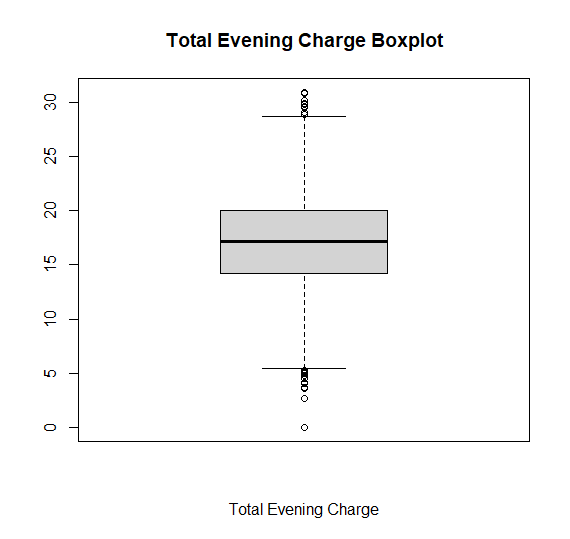
The total evening calls attribute shows:

* Min ≃ 0
* Max ≃ 175
* Mean ≃ 100



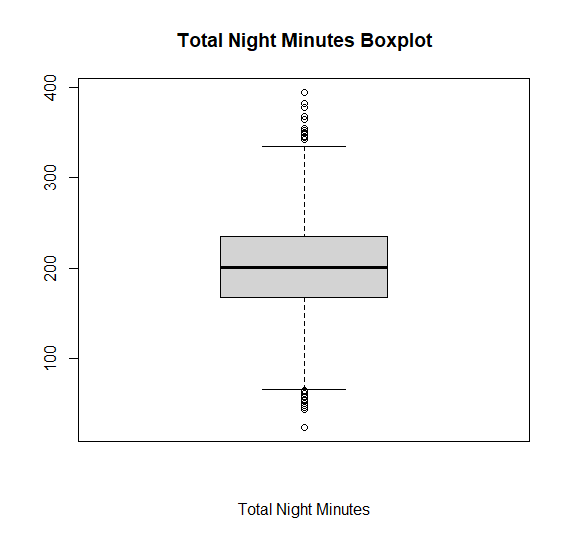
The total evening charge attribute shows:

* Min ≃ 0
* Max ≃ 30
* Mean ≃ 17



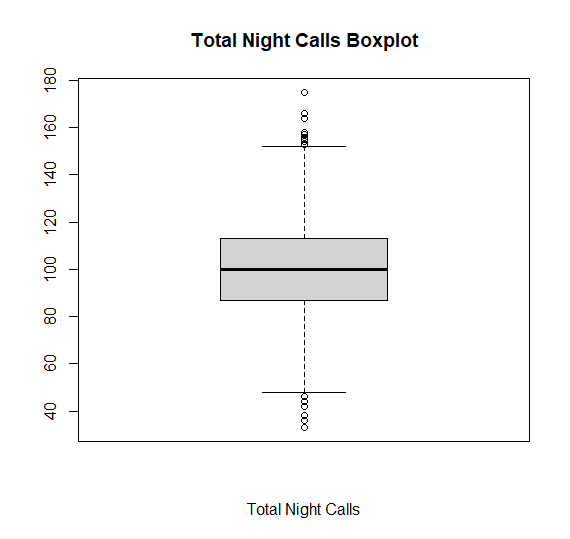
The total night minutes attribute shows:

* Min ≃ 0
* Max ≃ 400
* Mean ≃ 200



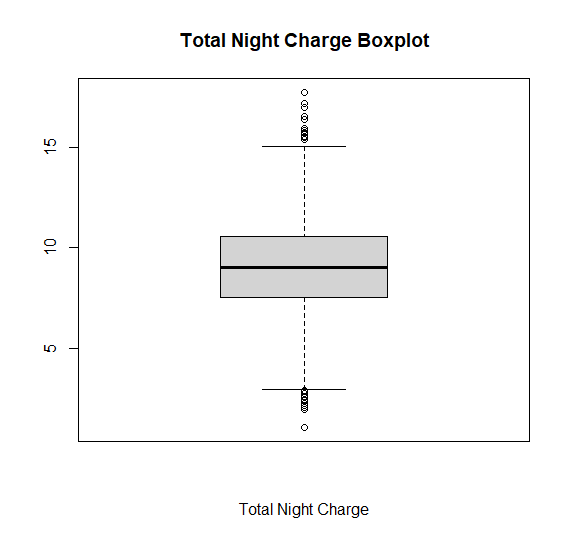
The total night calls attribute shows:

* Min ≃ 30
* Max ≃ 170
* Mean ≃ 100



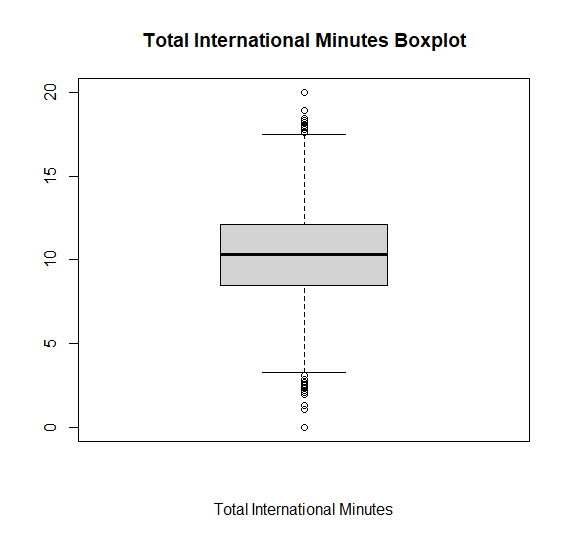
The total night charge attribute shows:

* Min ≃ 0
* Max ≃ 18
* Mean ≃ 10



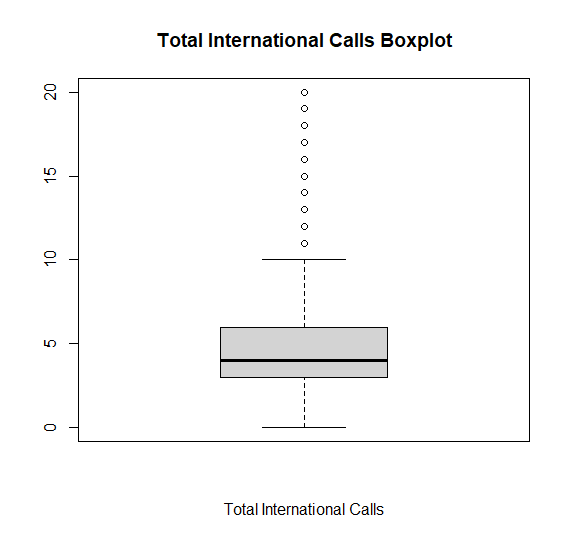
The total international minutes attribute shows:

* Min ≃ 0
* Max ≃ 20
* Mean ≃ 10



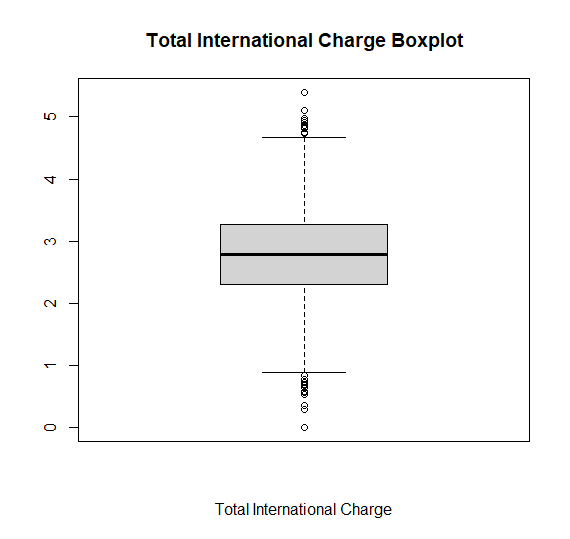
The total international calls attribute shows:

* Min ≃ 0
* Max ≃ 20
* Mean ≃ 4



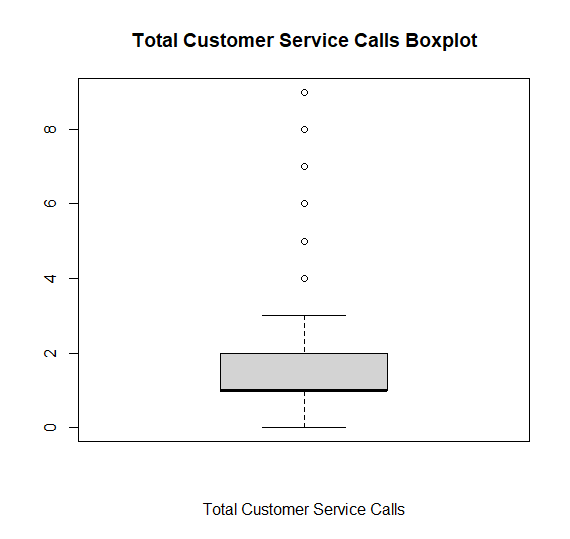
The total international charge attribute shows:

* Min ≃ 0
* Max ≃ 5
* Mean ≃ 3

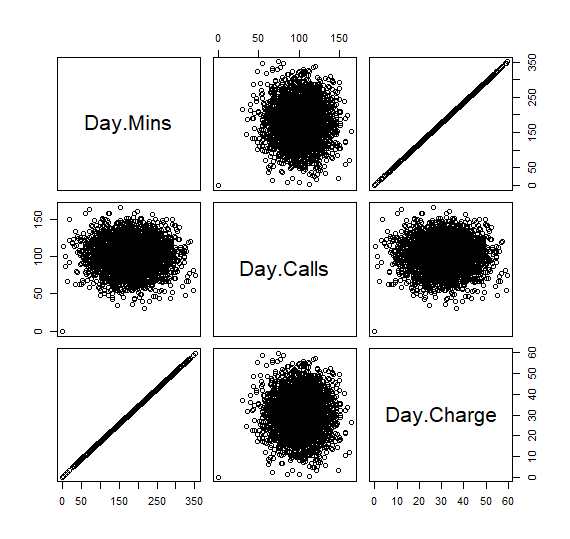
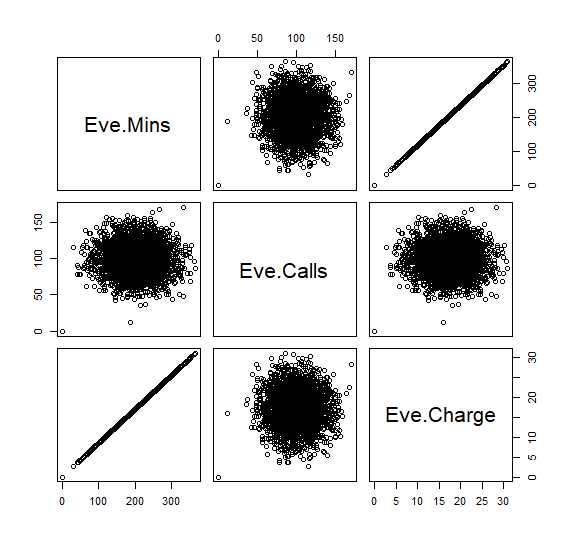


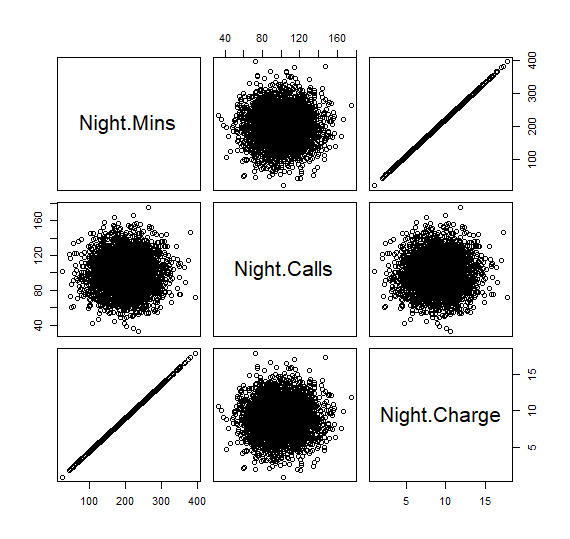
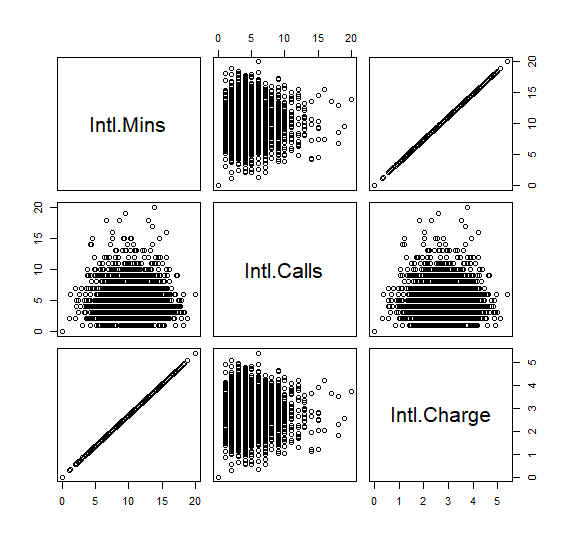
The total customer service calls attribute shows:

* Min ≃ 0
* Max ≃ 9
* Mean ≃ 1



Correlations

As we can see each set of day, evening, night and international share some correlation with their respected total calls, minutes, and charge. Total calls show a cluster with minutes and charge for each set, indicating no correlation. While total minutes and charge show a strong linear relationship in each set. This indicates that the company linearly charges per minute for each set.

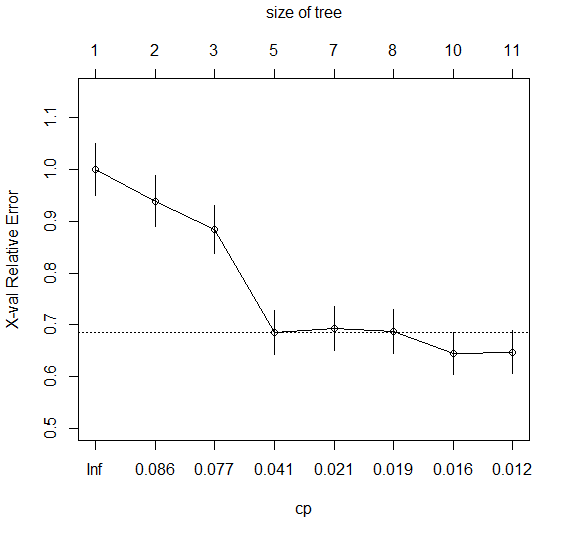
# Predictive Modeling/Classification

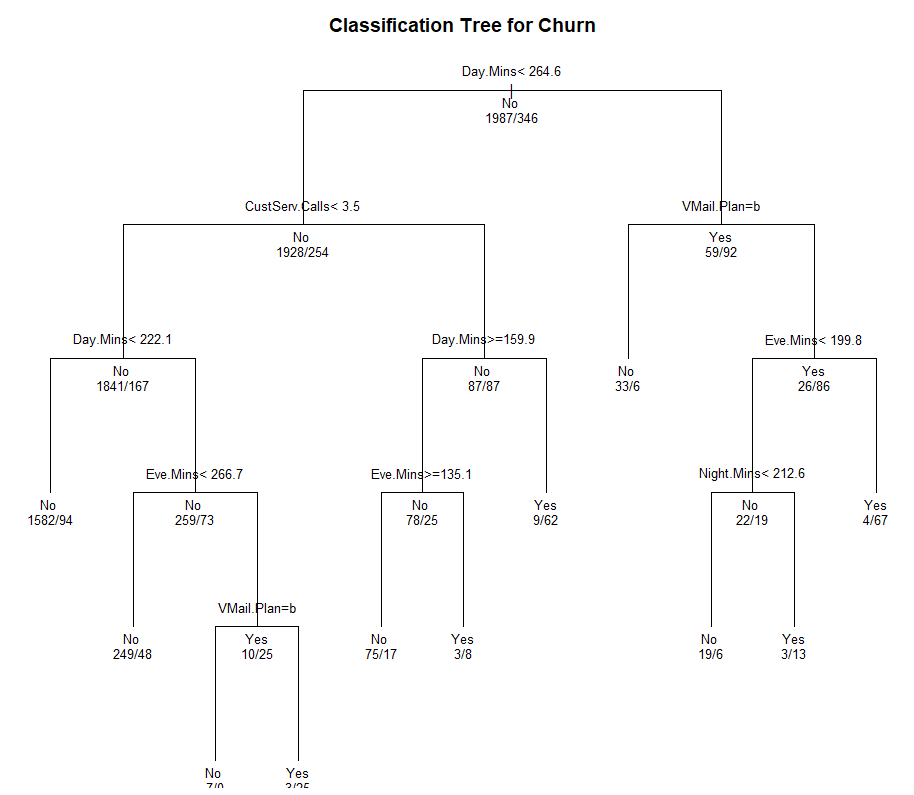
Data Split

To create the models first the data needs to be split into training and test sets. The training set will help teach the models how to define a customer’s churn based on their characteristics. The test set will test the model’s predictions and give a confusion matrix based on the model’s performance. To split the data into its sets we will use simple random sampling with a 70/30 split. The reason why we make sure the split is random is to ensure the there is no bias on which data is in either set.

Decision Tree

The variables used in the model’s construction are customer service calls, total day minutes, total evening minutes, total night minutes and voice mail plan. The root node error is 346/2333 = 0.14831.





Confusion Matrix for Decision Tree:



Accuracy = TP + TN / Total

= 923 / 1000

= 0.92

Recall = TP / (TP + FN)

= 68 / 137

= 0.49

Precision = TP / (TP + FP)

= 68 / 76

= 0.89

Naïve Bayes:

Naive Bayes Classifier for Discrete Predictors

A-priori probabilities:

False True

0.8516931 0.1483069

Conditional probabilities:

Area.Code

False 437.3548 42.51378

True 437.2486 42.31951

VMail.Message

False 8.965778 14.16525

True 5.676301 12.38369

Day.Mins

False 175.9873 49.81555

True 206.8072 68.19458

Day.Calls

False 100.3065 19.64901

True 102.8497 20.98420

Day.Charge

False 29.91834 8.468565

True 35.15763 11.593108

Eve.Mins

False 199.5548 50.51046

True 214.2173 51.03701

Eve.Calls

False 100.0584 20.00143

True 100.7717 19.47774

Eve.Charge

False 16.96242 4.293442

True 18.20864 4.337909

Night.Mins

False 200.0414 51.14855

True 203.8486 47.62280

Night.Calls

False 99.65123 19.69226

True 99.56358 20.26928

Night.Charge

False 9.001928 2.301756

True 9.173237 2.143112

Intl.Mins

False 10.13392 2.833359

True 10.64220 2.929314

Intl.Calls

False 4.542023 2.469679

True 4.158960 2.615593

Intl.Charge

False 2.736638 0.7649528

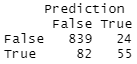
True 2.873728 0.7908848

CustServ.Calls

False 1.441872 1.187811

True 2.202312 1.853972

Confusion Matrix for Naive Bayes



Accuracy = TP + TN / Total

= 894 / 1000

= 0.89

Recall = TP / (TP + FN)

= 55 / 137

= 0.40

Precision = TP / (TP + FP)

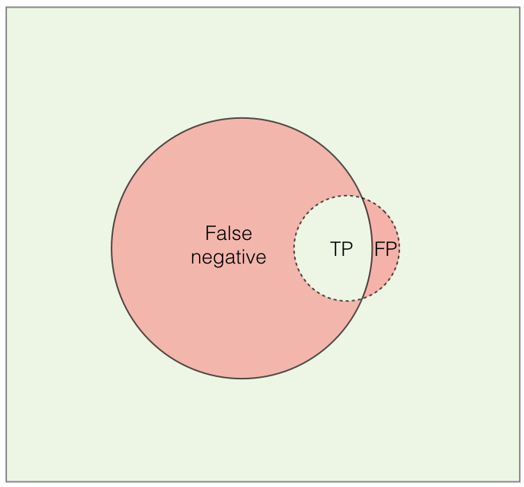
= 55 / 79

= 0.69

# Conclusions and Recommendations

From analyzing the data during the data preparations, we learn a lot about the data and how they are related. The main take away is the correlations of the attributes. We learned that the telecommunications company charges linearly based on minutes.

From the predictive modeling we learn that the decision tree uses five attributes to predict customer churn while the naïve bayes uses all given attributes. Both models show a strong accuracy with a low recall and a high precision. This means both classifiers wrongly predict that customers will not churn when they do. This also means that when the classifiers are predicting customers churn they are typically correct.



Based on the respected confusion matrix we can see that the decision tree model outperforms the naïve bayes model. A greater accuracy of 3%, recall 9% and precision 20% indicates that the decision tree model predicts customer churn overall better. These percentages differences may not seem great but affect the number of customers churns given a greater population.

The decision tree model tells us that customer churn can be accurately predicted based on the customer attributes customer service calls, total day minutes, total evening minutes, total night minutes and voice mail plan. Based on the model customer churn is higher among customers that have a voice mail plan, more than three customer service calls, or above average minutes in either time of day. Also based on the model, we can determine that customer service provides a positive customer experience resulting in more customers churning. For the company I recommend two plans. One, to help retain more customers I recommend that the company finds ways to increase the customer service calls for those that are lower than average in this attribute. Second, that they create a phone bundle plan that includes voice mail and charges customers on a log scale per minute. Customers with less minutes will be charged more while customers with more minutes will be charged less compared to the current linear charge per minute. Charging on a log scale per minute will entice customers to have a greater total minute and result in more possible customers churning. This plan will create more customers that fit under the classifications of churning. It will provide a discount to customers more likely to churn in the future and help the company retain more customers.