Report 03

WILL THEY LEAVE? AN EXPLORATION OF CHURN USING PREDICTOR MODELS

Brendan D. Cubberly

December 9, 2022

Explorations in Data Science

Part 1: Establishing Baseline and Building a Model

After trying to examine if customers will leave the cell phone company, this is what I found. I started by observing the data of how many previous customers have left the company, and those who have stayed. Based on the training data I have, I was able to have a baseline model of around 2509 total customers, 2153 of which have stayed with the company. This gave me an accuracy of around 85.8%. I also created a CART model (shown below in Figure 1), which is a predictive model to help see how values can be influenced on others. As we can see in Figure 1, we have the data split at nodes by TRUE and FALSE. If the data is split by TRUE, then the customer left the company. If the data is split by FALSE, then the customer stayed with the company.

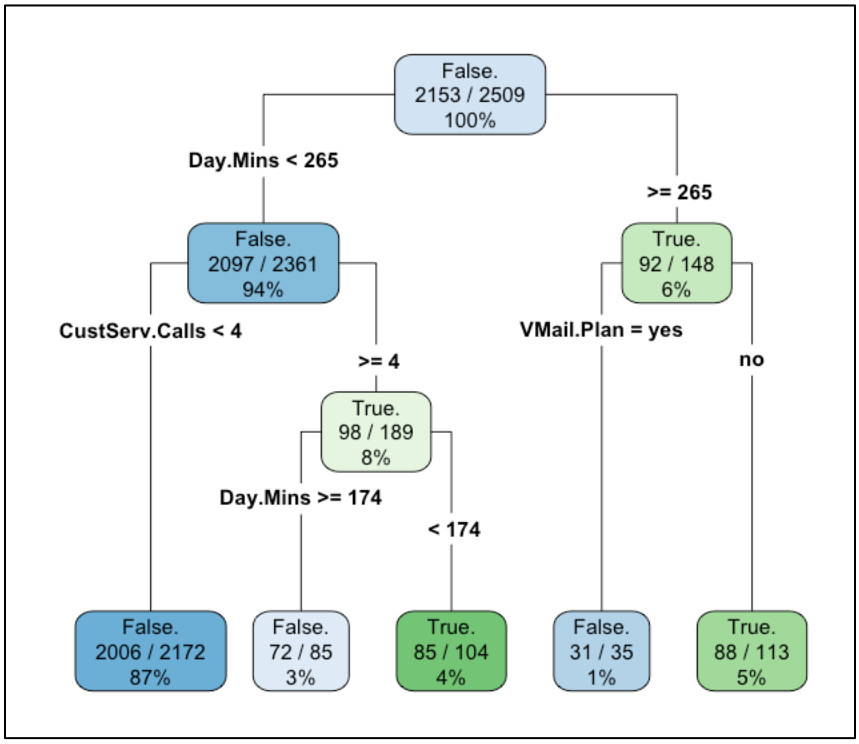

Figure 1: CART Model of the Churn Training Data Set

Another predictive model is the C5.0 model, which is shown as Figure 2 below. A C5.0 Model is not limited to binary splits, whereas a CART model is. As we can see in Figure 2, the model had nodes that are split based on the data, most notably the three largest segments. The largest is node is node 3, where it contained 2121 customers. Here, these customers stayed with the company if they spent less than or equal to 254.9 minutes on the phone and had 3 or less phone calls with customer service. This node had most customers stay with the company. The second largest node is node 5, with 104 customers. Here, the customers had also spent less than 255 minutes on the phone, but they had more than 3 phone calls with customer service. In addition,

this node contained customers who spent less than 173.6 minutes with customer service. Most of the customers in node 5 had left the company. The last largest node is node 6 with 80 total customers who mostly stayed with the company. Here, the customers spent less than 255 minutes on the phone, more than 3 calls with customer service, but had more than 173.5 minutes with them.

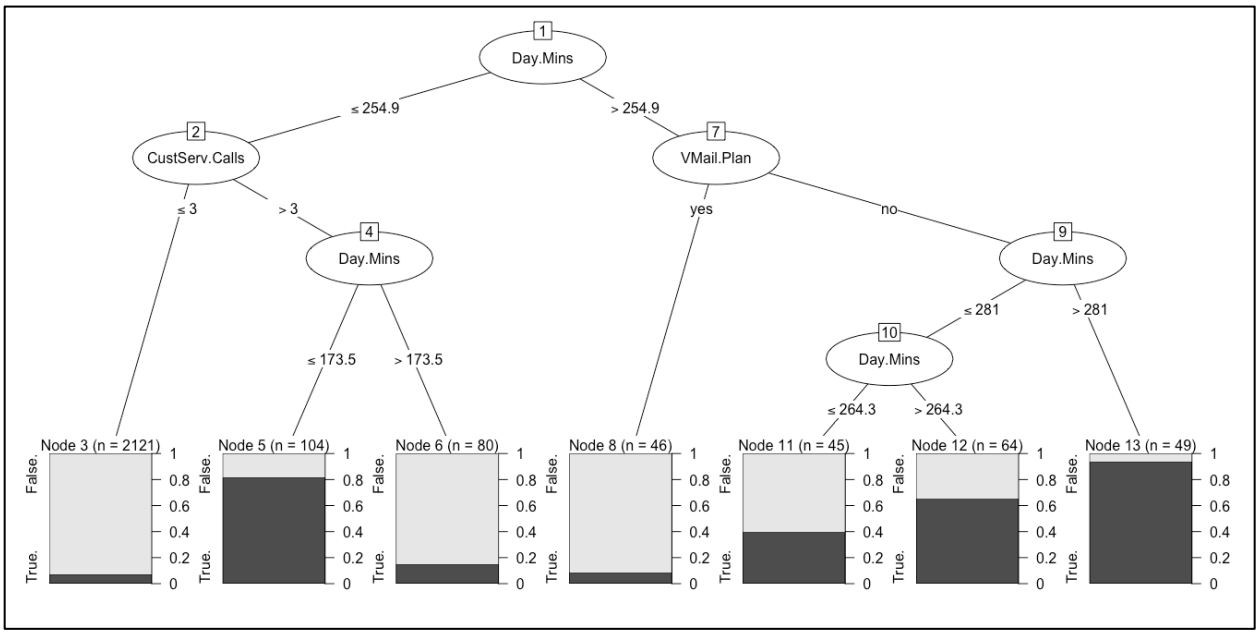

Figure 2: C5.0 Model based on the Churn Training Data Set

Part 2: Evaluate the Model

From Figure 3 below, we can see the difference between what the predicted values were, and the actual values of the data. Here, the values that are classified under the same predicted value and the same actual value mean they match. Essentially, the actual value of customers who stayed with the company is 2153, but the model predicted that 44 of them would leave. The value 2109 is classified as a True Negative, and the value 173 is classified as a True Positive. The numbers that didn't match are False Positives and False Negatives.

	Predicted	
Actual	*Churn False (Negative)*	*Churn True (Positive)*
Churn False (Negative)	2109	44
Churn True (Positive)	183	173

Figure 3: Contingency Table of Actual vs. Predicted Values

```
Evaluation on training data (2509 cases):

            Decision Tree
          ----------------
          Size      Errors

             7  227( 9.0%)   <<


           (a)   (b)    <-classified as
          ----  ----
          2109    44    (a): class False.
           183   173    (b): class True.
```

Figure 4: Diagram of Actual vs. Predicted Values Calculated in R

In a model, accuracy is the overall proportion of correct classifications. For this data set, accuracy is given by the following formula:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} = \frac{TN + TP}{n}$$

Figure 5: Formula for Accuracy in a Data Set

To calculate accuracy, we get the following:

$$\frac{2109 + 173}{2109 + 173 + 183 + 44} = \frac{2282}{2509} = 0.9095$$

To calculate sensitivity of the model, we can use the following formula:

$$Sensitivity = \frac{TP}{TP + FN}$$

Figure 6: Formula for Sensitivity

To calculate the sensitivity of the data set,

$$\frac{173}{173 + 183} = 0.486$$

To calculate the specificity of a model, we can use the formula:

$$Specificity = \frac{TN}{TN + FP}$$

Figure 7: Formula for Specificity

To calculate the specificity of the data set,

$$\frac{2109}{2109 + 44} = 0.9796$$

From the above calculations, we can see that there are some differences in the numbers. An example of this is the sensitivity of the data set is not ideal. For a good model, we would want the data sensitivity to be as close to 100% as possible. Since our sensitivity is only at 48.6%, it is not a good model to use. This is one weakness to the C5.0 model, as both the accuracy and specificity are high percentages, but sensitivity is not. However, this model is a good specific model with a percentage of 97.96% and is also an overall accurate model with an accuracy of 90.95%!

Part 3 (Optional): Considering the Costs of Predictions!

Considering the following statement:
If you send the intervention to a customer who was, in fact, going to leave, you have retained the $100 that you would have lost (a $100 profit) while spending the $10 cost for the intervention.
This is classified as True Positive.

Considering the following statement:
If you send the intervention to a customer who was not, in fact, going to leave, then you've spent $10 but gained the company nothing.
This is classified as False Positive.

Considering the following statement:
If you do not send the intervention to a customer who was not, in fact, going to leave, then you've spent nothing and gained nothing.
This is classified as False Negative.

Considering the following statement:
If you do not send the intervention to a customer who was, in fact, going to leave, then you've cost the company $100.
This is classified as True Negative.

Costing/Saving the Company

Actual	**Predicted**	
	Churn False (Negative)	*Churn True (Positive)*
Churn False (Negative)	$100	-$90
Churn True (Positive)	$100	-$90

Calculations from Question 4:

Actual	**Predicted**	
	Churn False (Negative)	*Churn True (Positive)*
Churn False (Negative)	2109 * 100 = 210900	44 * -90 = -3960
Churn True (Positive)	183 * -100 = -18300	173 * -90 = -15570

My C5.0 model costed the company $173,070. Since the negative numbers in the table are the savings, we can add up all the numbers together to get:

$$210900 - 18300 - 3960 - 15570 = 173070$$

Since the final result is a positive number, that is how much the C5.0 model costed the company.

Appendix

```
library(caret)
dtrain <- churn_train
dtrain$Int.l.Plan <- as.factor(dtrain$Int.l.Plan)
dtrain$VMail.Plan <- as.factor(dtrain$VMail.Plan)
dtrain$Churn. <- as.factor(dtrain$Churn.)

table(dtrain$Churn.)

set.seed(107)
inTrain <- createDataPartition(y=dtrain$Churn., p = .75, list = FALSE)
d.train <- dtrain[inTrain,]
d.test <- dtrain[-inTrain]

library(rpart)
cart01 <- rpart(Churn. ~ ., data = dtrain, method = "class")
cart01
install.packages("rpart.plot")
library(rpart.plot)
rpart.plot(cart01, type = 4, extra = 102)

install.packages("C50")
library(C50)
mod1 <- C5.0(Churn. ~ ., data = dtrain)
summary(mod1)
plot(mod1)
```