**Which customers are churning and why?**

Prepared by Jennifer Tatum

On March 28th, 2017

To determine which customers are churning and why, I have created a decision tree model for analysis. My analysis shows that there are several attributes that are shared by churning customers, some of which we may be able to influence.

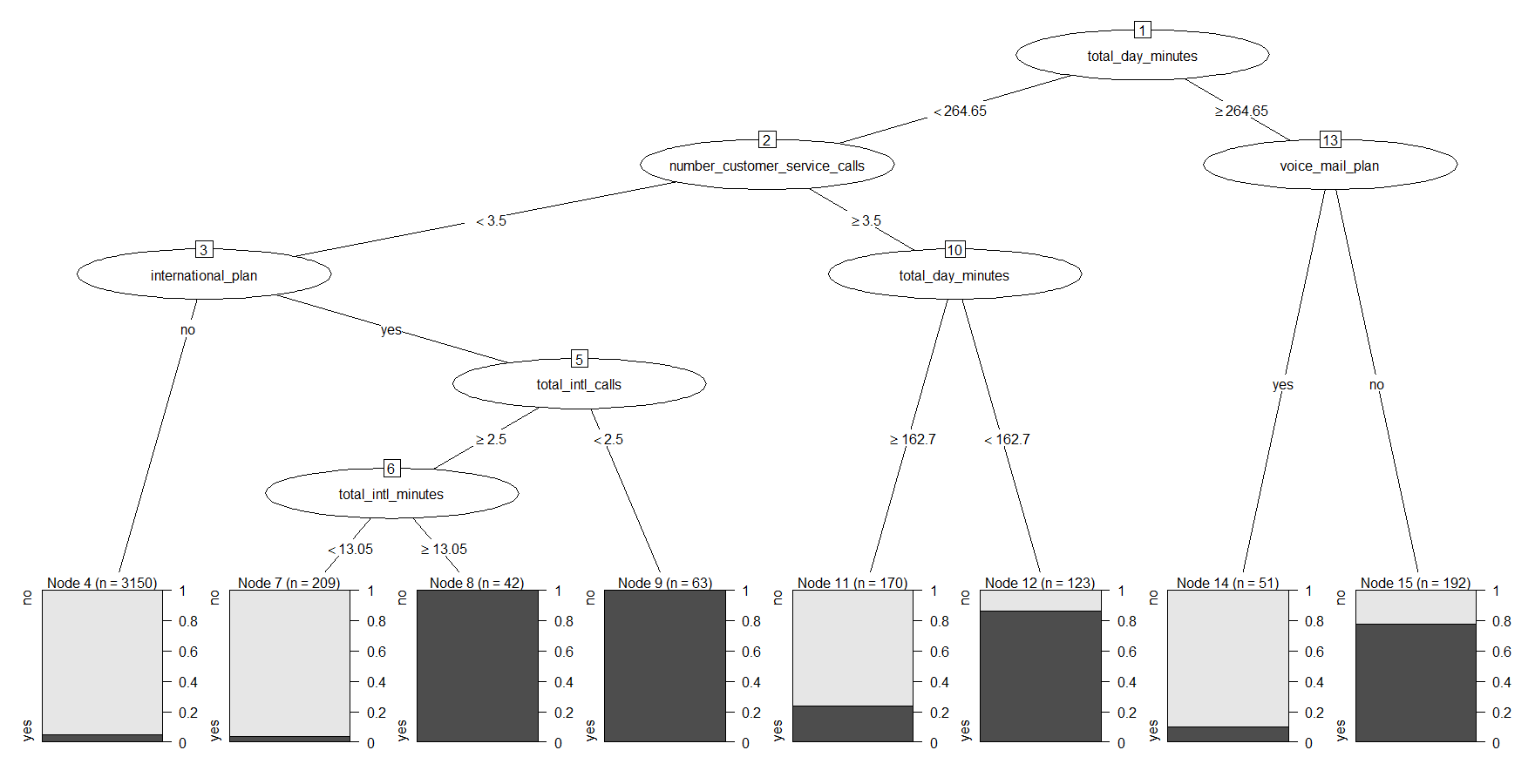
**Key Takeaways**

* Excessive talkers (more that 265 minutes per day) prefer to have a voicemail plan.
* Customers with lower day minutes are less tolerable of services call.
* International plan customers are picky about their plans.

**Analysis Summary**

To determine which of our customers are churning and why, I considered a data set with 5,000 observations of 20 different variables, one of which being whether the customer churned or not. The other variables include total minutes, calls and charges by day, evening, night and international. Geographic data is included with state and area code. It also includes how long someone has been our customer, how may customer service calls they have made, and whether or not they have an international or voicemail plan (with number of voicemail messages).

I have randomly divided this data into a training set with 4,000 observations and a validation set with 1,000 observations. The training set is used to develop the decision tree model and the validation set is used to measure how well the model performs. Decision trees models can be built with a small or large number of variables. If all 20 variables are used, it becomes cumbersome to evaluate the results and the splits at the bottom of the tree are amongst very small numbers of customers making those distinctions less significant. If we “prune” the tree to a manageable number of variables, it is easy to see the results below.

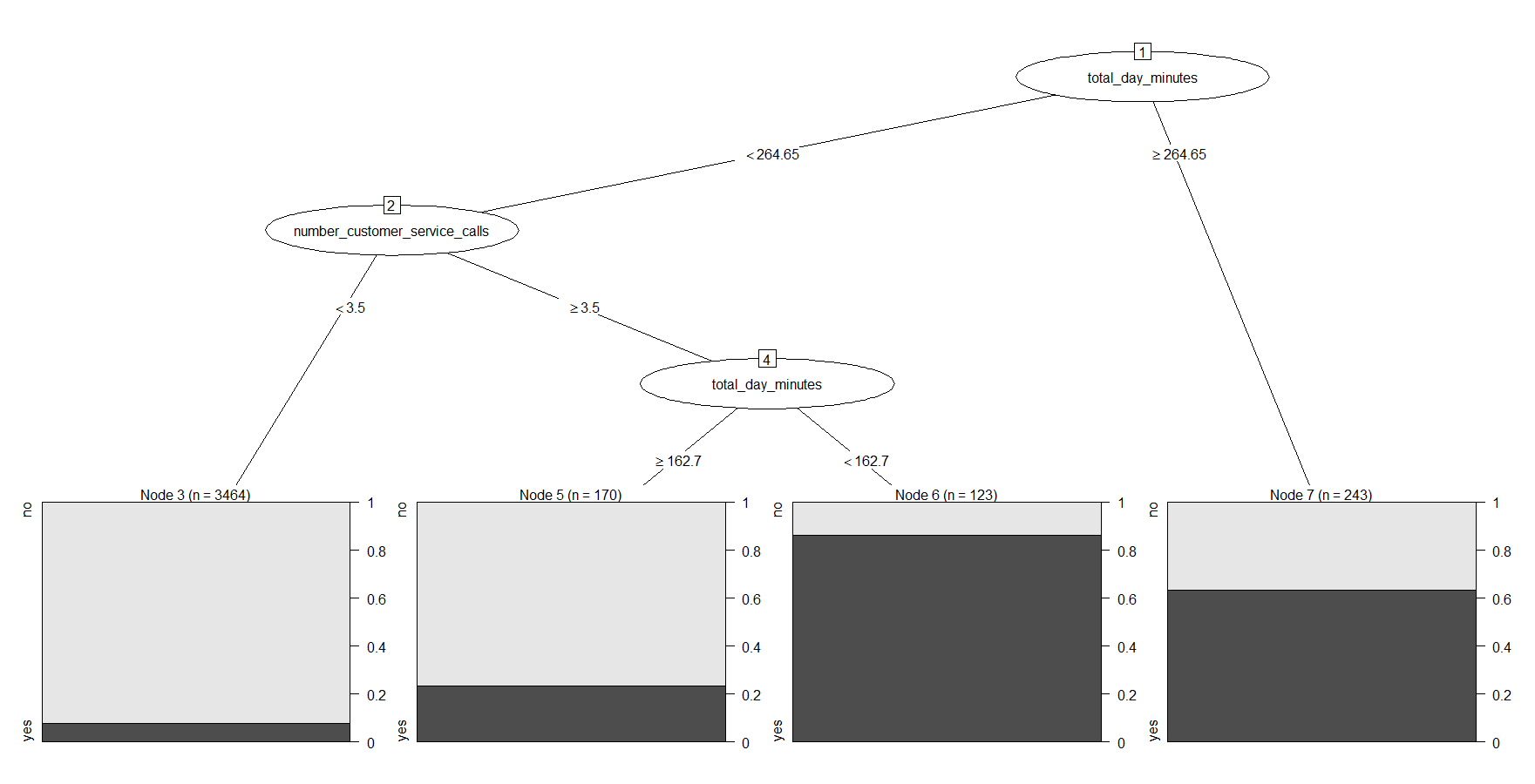


Visualizing this pruned decision tree model, we are able to see some interesting facts about our customers. Most of our current customers (3150) are regular talkers (less than 265 minutes per day), have less than 3.5 customer service calls, and do not have an international plan.

We can also see that the most significant factors in explaining customer churn are number of day time minutes, customer service calls, and international plans. The largest population of churned customers is excessive talkers without a voicemail plan (150 customers). Another 100 customers left who were regular talkers with more than 3.5 customer service calls and less than 163 minutes a day. And all of the customers who were regular talkers, with less than 3.5 customer service calls, an international plan and either made less than 2.5 international calls (63 customers) or more than 2.5 international calls with more than 13.5 international minutes total (42 customers) churned.

To test how well this decision tree model is at predicting churn, I compared the predicted values against the actual values from the validation data set of 1,000 observations. There are several performance indicators to consider. The accuracy rate (how often the model predictions the correct outcome, churn or not churn) is 93.6%. This means that 93.6% of the time the model correctly predicted whether a customer in the validation data churned or not. The sensitivity, how well the model can predict a churned customer, is 65%. The specificity, how well the model predicts a not churn and the actual value is not churn, is very high at 98%. And finally, precision, how well the model predicts a churn and the actual value is churn, is 81%.

Overall these performance indicators are reasonable. However, if reducing the number of variables made the decision trees easier to understand, I wanted to look at an even smaller, simpler model to see if it will perform as well.



Here the customers at risk of churning are simply the customers that talk excessively or regular talkers with more than 3.5 customer service calls and less than 163 day time minutes. Although this model is much easier to understand, the performance measures are not as good as the first model.

All of the previously described indicators have dropped. The accuracy rate drops from 93.3% to 88.2%, sensitivity from 65% to 43%, specificity from 98% to 95%, and precision from 81% to 58%.

**Commentary**

Although the second model does very simply provide some insight into our customer churn, I recommend using the first model for analysis. This more granular approach proves to be more accurate across all performance indicators.

Given the insights from this model, there are several things Earth Connect can consider to reduce churn. First, plans with more than 265 minutes could come standard with a voicemail plan. Second, add minutes to plans for customers who talk less than 163 minutes a day and have more than 3 customer service calls. This may increase customer tolerance of such calls. And lastly, revisit the international plan make up to determine if we can offer different plans that better suit our customer’s needs.

**Addendum Part B**

A KNN model is the simplest model to determine whether a customer will churn or not. It is an exploration model that will help identify customers who will churn but not the factors that make a customer churn as with the decision tree. It tries to use similarities between customers to determine if a new customer will churn.

I have taken the same data set of 5,000 observations and divided it into a training set of 4,000 observations and a validation set of 1,000 observations as before. I created a KNN model using the training set and then validated the model with the validation set.

My first rendition of the model used an industry standard for an input of “k” as starting point. Because finding the actual churned customer is our focus, I tweaked the model to be more accurate and sensitive. I used what I learned from the decision tree model and started by reducing the base argument “k”. This is similar to the notion of not using as many variables and narrowing my focus. It is important to note that this decrease makes the model less bias, but also more variant. These two factors are balanced to find the best fitting model.

The best fitting KNN model has an accuracy rate of 90% which is not as accurate compared to the 93% accuracy rate of the decision tree model. Specificity is relatively the same but there are significant differences in sensitivity and precision. Again, sensitivity is how well the model is at predicting that someone will churn. The KNN model is drastically lower with a sensitivity of 26% compared to 65% of the decision tree. Precision is how well the model predicts someone will churn and they do actually churn and the KNN model is slightly better with an 86% rate compared to 81%.

Overall the KKN model performs similar to the decision tree. However, if our main focus is to fit customers who churn, the decision tree is better based on the sensitivity rate.

**Addendum Part C**

To look at this same information with a “tried and true regression model”, I conducted a logistical regression. Again I divided the data into a training and validation set with the same ratio of 4,000 to 1,000 observations respectively.

Because the first decision tree model identified the most significant variables in our data set, I was able to start this analysis with the following variables: day time minutes, customer service calls, international calls, and international minutes. From this we can see that customers who are excessive talkers (more than 265 minutes a day) are 23 times more likely to churn than customers who are not excessive talker. If customers have more than 3.5 customer service calls, they are 17 times more like to churn than customers who have less than 3.5 customer service calls. And customers with international plans are 9 times more likely to churn than customers who don’t have an international plan.

These figures are valid when you look at them independently. Meaning all of the other variables have to be constant to say that a customer is 23 times more likely to churn if they are excessive talkers versus not. This is true when we look at probabilities as well. Holding everything else constant, a customer who is an excessive talker has a 96% probability of churning. And customers who have more than 3.5 customer service calls or have international plans have probabilities of 94% and 90% of churning respectively. These probabilities are significant and the decision tree’s output shows where we can focus our attention within these broad variables.

Using the validation set to determine how well the logistical regression is performing, I predicted a churn if the regression model’s probability for churn for a customer was over 50%. The model’s accuracy rate was 90%, specificity 97%, sensitivity 42% and precision 69%. In generally the model performed better than the KNN, but not as well as the decision tree when focusing on accuracy and sensitivity.