Logistic Regression and Classification Tree on Customer Churn in Telecommunication

Y. Long, T. Nguyen, and I. Tareque[1]

Abstract

The term *churn* refers to a customer unsubscribing to a service. Accurately identifying why customers churn is critical for telecommunication companies as such information enables them to improve specific services important for greater client retention. In our study, using customers' usage data, we perform logistic regression and classification tree analyses to develop two models that can predict with over 80% accuracy whether a customer will churn or not. We hypothesize that customers' international call lengths and subscription to international plans are important factors that companies should consider in order to predict their likelihood of churning. Our models affirm this hypothesis.

[1] Authors ordered by last name. This draft incorporates minor verbiage edits for greater clarity and is slightly different from the draft that was submitted for the competition.

Background and Significance

The term *churn* refers to a customer unsubscribing to a service. The ability to successfully identify which customers are likely to churn and why can play a critical role in boosting a telecommunication company's success, since firms can leverage such information to strategically improve specific products or give customers relevant promotions.

Published literature on this topic usually has large datasets available for predicting churn. For instance, Hung et. al (2006) created a model that can accurately predict churn using customer demographics, billing information, contract or service status, detailed call records, and service change log entries. Similarly, Jungxiang Lu performed survival analysis to develop a model consisting of 29 explanatory variables to predict churn in the telecom industry. The model had high predictive power, with the top two and top five deciles capturing about 60% and 90% of churners respectively.

However, given the real-world problem of limited customer information in small companies, we are interested in predicting customers' probability of churning based on their usage history only. Moreover, since in today's globalized world, demand for international communication is very high and there are many cost-effective alternatives to telecom services such as Skype and Viber, customers' demand for international communication and the quality of international plans likely significantly affects their decision to remain with a particular telecom company (Hughes, 2007). Therefore, we hypothesize that whether a customer subscribes to an international plan plays a significant role in determining if they will churn, and there is an interaction effect between a client's international plan subscription and international communication history.

Methods

Data Description

We use a public dataset provided by the CrowdAnalytix competition in 2012. The dataset provides a list of 22 variables regarding the phone usage pattern of 5,000 customers.[2] To avoid overfitting the model, we use two randomly divided subsets - *train* with 3,333 observations and *test* with 1,667. The predictive models are built based on the *train* dataset and tested on *test.*

Analytic Methods

1. Logistic Regression

As charge and minutes variables are perfectly correlated, we did not include the charge variables in our analysis. Next, as churn prediction is an understudied topic, we used stepwise variable selection technique along with drop-in deviance tests to screen all possible variables for association with the outcome variable at a significant level of 0.05 (Hosmer, Lemeshow, & Sturdivant, 2013). Since we are particularly interested in the *international plan* variable, we left it out of the stepwise procedure. After obtaining the list of all significant variables, we added all possible interaction terms and repeated the stepwise procedure to deduce Model 1. We then used a drop-in deviance test to compare this model to Model 2, created by adding the *international plan* variable and its interaction term (*international plan* with *total international minutes*) to Model 1.

2. Classification Tree

Classification and Regression Tree (CART) analysis uses binary recursive partitioning to find the best splitting points to build a decision tree that predicts if a customer will churn. We used the CRAN rpart package to construct the classification tree.[3] We first considered all the available variables (excluding charge variables as explained in the Logistic Regression section above) in

[2] See Appendix for detailed variable definitions.

[3] R code is included in the Appendix.

our model, and then pruned the tree to avoid overfitting. We chose a complexity parameter[4] of 0.05 because according to the plot of complexity, there is a large drop of x-val relative error from 0.08 to 0.05, but the change is small below 0.05.

Results

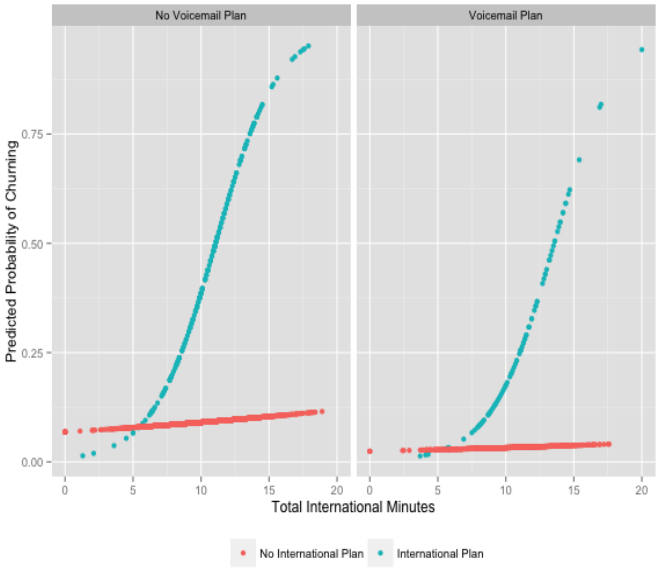

Figure 1. Predicted probability of churning based on Model 2 against *total international minutes*, *international plan*, and *voicemail plan* while controlling for other variables.

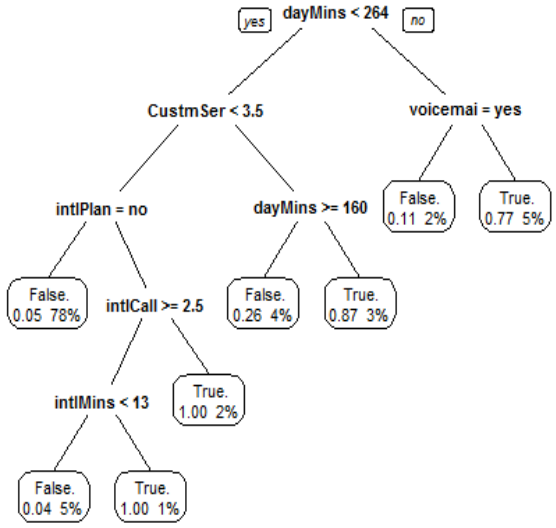

Figure 2. Classification Tree for customer churn in telecommunication. Each node is a criterion. Left branch meets the criterion, and right branch does not. "False" indicates that the individuals in that group are not likely to churn, whereas "True" indicates that they are likely to churn.

Logistic Regression

Model 1 has 8 variables: *total day minutes*, *total evening minutes*, *total night minutes*, *total international minutes*, *total international calls*, *number of customer service calls*, having a *voicemail plan* (dummy), *number of voicemail messages*, and two interaction terms: *total day minutes* with *total night minutes*, and *total day minutes* with *total evening minutes*.

The drop-in deviance test with Model 2 as the full model and Model 1 as the reduced model yielded a very high $G = 233.23$ and $p < 0.05$ (df = 2). Therefore, we can reject the null hypothesis and conclude that either *international plan* or the interaction term of *international plan* with *total international minutes* is an important indicator for predicting the probability of churning. Figure 1 indicates that customers who subscribe to an international plan have a higher predicted probability of churning compared to those who do not, especially when they use more international minutes.

[4] "The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue. We could also say that tree construction does not continue unless it would decrease the overall lack of fit by a factor of cp" (Williams, 2010). The smaller the cp, the larger the tree size and higher the potential of overfitting. The default cp is 0.01.

Classification Tree

Classification Tree analysis identified 6 significant variables at the complexity level of 0.05: *total day minutes*, *number of customer service calls*, *having an international plan* (dummy), *total international calls*, *total international minutes*, and having a *voicemail plan* (dummy). Therefore, *international plan* is identified as a significant factor by the Classification Tree analysis as well.

Discussion

Is International Plan an Important Variable in Predicting Customer Churn?

Both methods consistently show that whether a customer subscribes to an international plan is an important variable in predicting a customer's probability of churning. More interestingly, both Figure 1 and Figure 2 indicate that customers with international plans are even more likely to churn as their total international minutes increase. This finding suggests that this telecommunication service's international plan may not work well for the clients with a high demand for international calls.

However, the Classification Tree analysis indicates that having an international plan might not be the most crucial factor for churn prediction. For example, according to Figure 2, if a customer uses more than 264 daytime minutes in total and does not use a voicemail plan, they will likely churn regardless of whether they are subscribed to an international plan or not.

Furthermore, it is important to keep in mind that as our models are based on only one company's data; this conclusion cannot be generalized to the entire telecommunication industry. However, this low external validity should not be too concerning. Given the nature of the telecom industry, each service might have quite unique customer segments and accordingly, it is better to build separate predictive models by company. Another limitation is that we can only conclude that the international plan is an important variable among usage pattern variables. It is possible that this variable is relatively not as important when we include other types of variables like customer demographics.

Accuracy of Logistic Regression (LR) and Classification Tree Analysis

According to Table 1, both methods show relatively good and similar accuracy for both the *train* and *test* data. Therefore, we do not overfit our models.

Table 1. Percentage of concordant and discordant pairs in LR and Classification Tree

Method	Train		Test	
	Concordant	Discordant	Concordant	Discordant
Logistic Regression (Model 2)	83.90%	15.80%	84.90%	14.80%
Classification Tree	82.25%	17.75%	84.31%	15.69%

Recommendation and Future Approach

Both models predict that whether a client subscribes to an international plan or not, and their total international minutes used play a significant role in determining their likelihood of churning. This implies that the international plan's quality likely affects this particular telecom company's churn rate. We therefore recommend that this specific telecom company should divert some resources towards improving their international plans. Moreover, in order to better understand the behavior of clients calling internationally, companies overall should collect more behavioral and usage data that can help them make appropriate changes to the international plan for increasing client retention rates. Finally, as the classification tree analysis conveys a slightly different story than the logistic regression analysis, researchers should investigate the significant variables identified by our models in greater depth.

References

Lu, J. (n.d.). *Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS*. Retrieved from http://www2.sas.com/proceedings/sugi27/p114-27.pdf

Hughes, A. (2007). *Customer Churn Reduction and Retention for Telecoms: Models for All Marketers*. Racom Communications.

Hung, S., Yen, D., & Wang, H. (2006). *Applying data mining to telecom churn management. Elsevier, 31, 515–524-515–524.*

Chris, Clair, and Aditya. (2014). *Comparison between CART Analysis and Logistic Regression Models.* YouTube video from: https://www.youtube.com/watch?v=IcDP62bKfFc

Hosmer, D. W., Lemeshow, S., and Sturdivant, Rodney X. (2013). *Applied logistic regression*. New York: Wiley. Print.

James, Witten, Hastie, and Tibshirani. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer Texts in Statistics. New York. Print.

Kabacoff, Robert. (2014). *Quick-R: Tree-Based Models*. Retrieved from: http://www.statmethods.net/advstats/cart.html

Lewis. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis.* Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf

Morgan, Jake. (2014). *Classification and Regression Tree Analysis*. Retrieved from: https://www.bu.edu/sph/files/2014/05/MorganCART.pdf

Shashia. (2014). *Binary Logistic Regression on R: Concordance and Discordance.* Retrieved from: http://shashiasrblog.blogspot.com/2014/01/binary-logistic-regression-on-r.html

Williams, Graham. (2010). *Data Mining Desktop Survival Guide.* Retrieved from: http://datamining.togaware.com/survivor/Complexity_cp.html

Appendix

1. Variable Definition

Variable	Definition
state	The customer's registered state.
account length	The number of days since the customer registered their account.
area code	The customer's area code.
phone number	The customer's phone number.
international plan	Dummy variable for whether the customer subscribed to an international plan or not.
voicemail plan	Dummy variable for whether the customer subscribed to a voicemail plan or not.
number vmail messages	The number of voicemail messages since the customer registered his or her account.
total day minutes	Total number of daytime minutes used since the customer registered their account.
total day calls	Total number of daytime calls made since the customer registered their account.
total day charge	Total dollars charged by daytime minutes used since the customer registered their account.
total eve minutes	Total number of evening minutes used since the customer registered their account.
total eve calls	Total number of evening calls made since the customer registered their account.
total eve charge	Total dollars charged by evening minutes used since the customer registered their account.
total night minutes	Total number of nighttime minutes used since the customer registered their account.
total night calls	Total number of nighttime calls made since the customer registered their account.
total night charge	Total dollars charged by nighttime minutes used since the customer registered their account.
total intl minutes	Total number of international minutes used since the customer registered their account.

total intl calls	Total number of international calls since the customer registered their account.
total intl charge	Total dollars charged by international minutes used since the customer registered their account.
number customer service calls	Number of customer service calls since the customer registers his or her account.
Churn	Dummy variable identifying if the customer left the telecom company.

2. **Interpreting the classification tree**

The decision tree represents the hierarchy of the variables in predicting the dummy outcome variable (churn). For example, when total daytime minutes is less than 264, the model suggests we consider the number of customer service calls. “False” indicates that the individuals in that group are not likely to churn, whereas “True” indicates that they are likely to churn. The number on the bottom left shows the proportion of the data points that placed in the opposite group. For example, 5% of the sample do not have an international plan and churned, while the decision tree predicts that those who do not have an international plan are not likely to churn. The percentage on the bottom right side shows the percentage of all data in the given group.

3. **Code for classification tree**

```
#Read in the data:
churn <- read.csv(“churn.data.csv”)

#Load the rpart package and rpart.plot package
install.packages(“rpart”)
require(rpart)
install.packages("rpart.plot")
require(rpart.plot)

#Generate the default tree:
tree <- rpart(churn~total.intl.minutes+international.plan+voicemail.plan+nightMins+
        total.day.minutes+eveMin+total.intl.calls+number.customer.service.calls+acctLen+numV
        mail+dayCalls+eveCalls+nightCalls+total.intl.calls,data=churn,method="class")
prp(tree)        #plot the tree
plotcp(tree)     #plot the complexity parameter

#Generate the pruned tree:
prunedTree <- rpart(churn~total.intl.minutes+international.plan+voicemail.plan+nightMins+
        total.day.minutes+eveMin+total.intl.calls+number.customer.service.calls+acctLen+numV
        mail+dayCalls+eveCalls+nightCalls+total.intl.calls, data=churn, method="class",
        cp=0.05)
prp(prunedTree)        #plot the tree
```