Logistic Regression and Classification Tree on Customer Churn in Telecommunication

Abstract
Knowing what makes a customer unsubscribe from a service (called churning) is very important for telecom companies as such information enables them to improve important services that can enable them to retain more customers. In our study, we perform logistic regression and classification tree analyses to develop two models that can predict whether a customer will churn or not using only customer usage data. We hypothesize that international communication history and subscription to international plans are important indicators that can help companies predict churn rate. Both of our models affirm this hypothesis.

Background and Significance

The term *churn* refers to a customer unsubscribing to a service. Successfully identifying which customers are likely to churn and why they churn plays a key role in the success of a telecommunication company since it can better provided products or give those customers relevant promotions.

Published literature on this topic usually has large data sets available for predicting churn. For instance, Hung et. al (2006) created a model that can accurately predict churn using customer demographics, billing information, contract/service status, call detail 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 has 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 available to small companies, we are interested in predicting customers' probability of churning based on their usage history only. Moreover, in today's globalized world, demand for international communication is very high and there are many cost-effective alternatives to telecom services like Skype and Viber so customers' demand for international communication and the quality of international plans should significantly affect customers' decision to remain with a particular telecom company (Hughes, 2007). Therefore, we hypothesize that whether a client subscribes to an international plan plays a key role in determining if he or she churns. There might also be an interaction effect between subscription and the customer's international communication history.

Methods

Data Description

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

Analytic Methods

We used 2 methods – Logistic Regression and Classification Tree – to build models predicting the probability of churning in order to see whether both methods give the same conclusion on the impact of the subscription to international plan.

1. Logistic Regression

There are few strongly supported theories on churn prediction so we used stepwise variable selection technique to screen all possible variables for association with the outcome at the significant level of 0.05 (Hosmer, Lemeshow, & Sturdivant, 2013). Model 1 was created by doing stepwise selection on all variables (except international plan dummy variable) and their interaction terms. Model 2 was obtained by adding the international plan dummy variable and the interaction term – international plan x total international minutes to Model 1. We then used drop-in deviance test to compare the reduced model (Model 1) to the full model (Model 2).

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 churns based on input variables (Loh, 2011). We use the rpart package (Therneau, Atkinson, & Ripley, 2015) in

[1] The variables include: State, Area code, Account length, Total day/ evening/ night/ international minutes, Total day/ evening/ night/ international calls, Total day/ evening/ night/ international charges, Voicemail plan (dummy), International plan (dummy), Number of voicemail messages, Number of customer service calls, Churn? (dummy). See Appendix for detailed definition.

Rstudio to construct the classification tree[2]. First, we entered the same variables in our model and then pruned the tree to avoid overfitting. We chose a complexity parameter[3] of 0.05 since there seems to be a big drop of x-val relative error from 0.08 to 0.05, but not very much change below 0.05 according to the plot of complexity.

Results

Logistic Regression

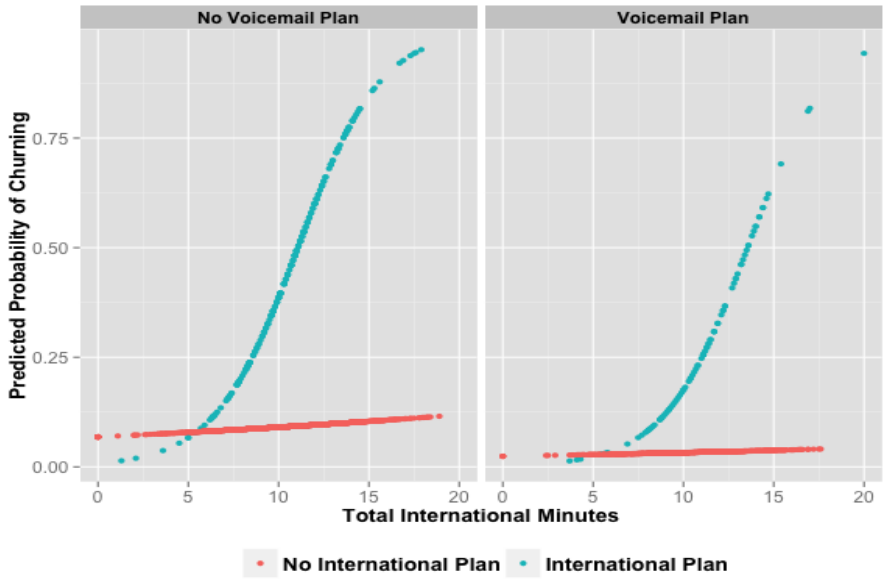

Figure 1. Predicted Probability of Churning based on Model 2 with response to Total international minutes, International Plan, and Voicemail Plan (controlling for other variables)

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 voicemail plan (dummy), number of voicemail messages, and 2 interaction terms: total day minutes with total night minutes and with total evening minutes.

The drop-in deviance test 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 International plan x Total international minutes is an important variable in predicting the probability of churning. Figure 1 indicates that customers who subscribe to international plan have higher predicted probability of churning than those who do not, especially when they use more international minutes.

Classification Tree

The Classification Tree identified 6 significant variables at the complexity level of 0.05: total day minutes, number of customer service calls, having international plan (dummy), total international calls, total international minutes, and having voicemail plan (dummy). Therefore, international plan is considered important by the classification tree as well.

[2] R code is included in the Appendix.

[3] "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 potential of overfitting. The default of cp is 0.01.

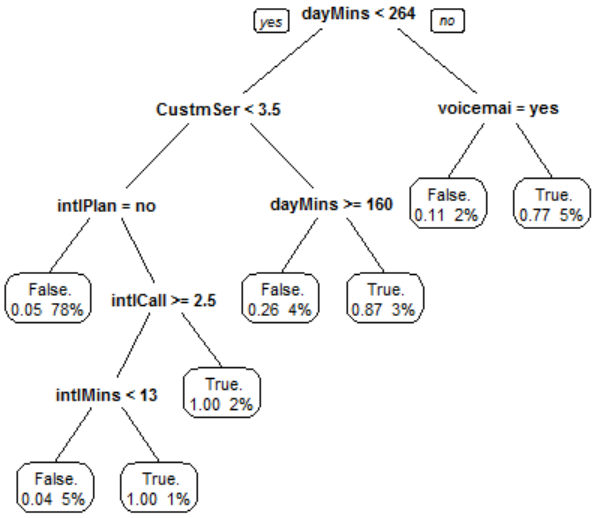

Figure 2. Classification Tree for Customer Churn in Telecommunication. Each node is a criterion. Each left branch represents meeting the criterion and failing to churn.
See Appendix for details.

Accuracy of Logistic Regression (LR) and Classification Tree

According to Table 1, both methods show relatively good and similar accuracy for both train and test data. Therefore, overfitting is not a problem in these two methods.

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

Method	Train		Test	
	Concordant	Discordant	Concordant	Discordant
LR (Model 2)	83.90%	15.80%	84.90%	14.80%
Classfication Tree	82.25%	17.75%	84.31%	15.69%

Discussion

Both methods consistently show that whether a customer subscribes to an international plan is an important variable in predicting the probability of a customer's churning. More interestingly, both Figure 1 and Figure 2 indicate that customers with international plan are even more likely to churn than those without the plan when their total international minutes increase. This finding might suggest that this telecommunication service's international plan did not work well for people with high demand for international calls.

However, the Classification Tree indicates that international plan might not be the most crucial factor in churn prediction. For example, according to Figure 2, if a customer uses more than 264 day minutes in total and does not use the voicemail plan, he or she will likely churn no matter whether he or she subscribes to international plan.

Furthermore, we should keep in mind that as this model is 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. The nature of this industry leads to the fact that each service might have quite unique customer segments and accordingly, it is better to build separate predictive models for each company.

From both models, it seems that whether a client subscribes to the international plan or not, and their total international minutes used play a huge role in determining whether they will churn. This might imply that the quality of the international plan could affect a telecom company's churn rate. So we recommend that this telecom company should divert some resources towards improving their international plans. Moreover, in order to better understand the behavior of clients calling internationally, companies should collect more behavioral and usage data that can help them make appropriate changes to the international plan for increasing client retention rates. Besides, as the classification tree tells a slightly different story as discussed before, researchers should investigate the variables identified by the analysis in greater depth.

References

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

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.* Retrieved from http://www.rmcet.com/lib/Resources/Archieve Report/E-Journals/Experimental Thermal and Fluid Science/Vol.31 No03 Oct 2006/Applying data mining to telecom churn management.pdf

Loh, Wei-Yin. (2011). *Classification and regression trees.* Retrieved from http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

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

Therneau, T., Atkinson, B., & Ripley, B. (2015). *Package 'rpart'.* Retrieved from https://cran.r-project.org/web/packages/rpart/rpart.pdf

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 registers his or her account
area code*	The customer's area code
phone number*	The customer's phone number
international plan (intlPlan)	The dummy variable returns yes if the customer subscribes to an international plan and no if not
voice mail plan (voicemail)	The dummy variable returns yes if the customer subscribes to a voice mail plan and no if not
number vmail messages	Number of voice mail messages since the customer registers his or her account
total day minutes (dayMins)	Total number of minutes used in the day since the customer registers his or her account
total day calls	Total number of calls used in the day since the customer registers his or her account
total day charge*	Total dollars charged by day minutes used since the customer registers his or her account
total evening minutes	Total number of minutes used in the evening since the customer registers his or her account
total evening calls	Total number of calls used in the evening since the customer registers his or her account
total evening charge*	Total dollars charged by evening minutes used since the customer registers his or her account
total night minutes	Total number of minutes used in the night since the customer registers his or her account
total night calls	Total number of calls used in the night since the customer registers his or her account
total night charge*	Total dollars charged by night minutes used since the customer registers his or her account
total international minutes (intlMins)	Total number of minutes used in the day since the customer registers his or her account
total international calls	Total number of international calls since the customer

(intlCall)	registers his or her account
total international charge*	Total dollars charged by international minutes used since the customer registers his or her account
number of customer service calls (CustmSer)	Number of customer service calls since the customer registers his or her account
Churn?	The dummy variable returns True if the customer switches the service when data are collected and False if not

** Variables were not included in the selection method due to inappropriateness and perfect correlation.*

2. **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
```

3. **How to interpret the classification tree?**

The decision tree represents the hierarchy of the variables in predicting the categorical outcome (churn or not). For example, when total minutes in the day is less than 264, we go to the left of the tree and look at the number of customer service calls. "False." indicates that the overall conclusion for a specific group is that the objects in the group are not likely to churn, and "True." indicates that they are likely to churn. The number on the left side of the box suggests the proportion of the data points that are placed in the wrong group. For example, there are 5% of data points that do not have international plan and churned when the tree shows that the data points that do not have international plan are not going to churn. The percentage on the right side of the box shows the percentage of the data points in this group out of the total number of data points.