Final Project

Abdullah Habeh, Jason Steen, Suhani Patel

Kent State University

Business Analytics

Final Project

Contributions of team members:

Project member	Project Contribution
Jason Steen	Contribution to the data exploration and modeling in addition to the final report.
Suhani Patel	Contribution to the data exploration, final report and the final presentation.
Abdullah Habeh	Contribution to the data modeling, the final report the final presentation

Project Goal:

Given data for subscribers for which the churn outcome is known, explore and preprocess the data with the goal of developing a classification model to predict churn for current subscribers. The accuracy of the model can be evaluated according to the area under the receiver operating characteristics curve. Another potential evaluation metric is to compare the true positive rate to the false negative rate. The project is silent regarding which type of error, false positive or false negatives, is worse for the company. It is implied that both are bad, but it can also be inferred that false negatives are worse. This is based on the description that churn is costly for the company, from which it can be implied that providing incentives to customers not likely to churn is less costly. Under this framing the goal could be to maximize true positives while minimizing false negatives.

Overview of data and exploration analysis:

The data contains both numeric and categorical data. Two numeric attributes have negative values. Some of the attributes are missing values while others are not. We did not attempt to modify the state identification data, as each model iteration provided different results that spanned multiple regions of the country. In fact, no modifications were made to categorical data as there were no missing values to impute.

Several of the attributes containing usage data are missing the same number of values implying that the data may not be missing at random for whatever reason. This led to an important decision regarding whether to drop the records containing the NAs. One of the most important variables discovered after running the model, membership in the international plan, did not have any NAs, in addition to a few other variables. Because of the importance of this variable, we decided to keep the records with the NA values and impute the mean value for the attribute in their place. We tested using median versus mean values for imputation and actually began with median for most of the values at the beginning. Better results were achieved using mean imputation. For the attributes with negative values, we reassigned all negative values as NAs and imputed the mean value excluding the NAs from consideration. The NAs were then replaced with the calculated mean.

There were four attributes that were noticeably skewed; day minutes evening minutes, number of customer service calls and number of voicemail messages. Of these, logarithmic transformations were taken of day and evening minutes and number of voicemail messages. Number of voicemail messages

provided better results after being transformed, in addition to day and evening minutes. We decided to use number of calls to customer service as a multiplier for all applicable charges, therefore a logarithmic transformation would have weakened these generated attributes.

Details of modeling strategy:

Although linear regression might seem like a good choice for modeling considering the regression, however, since the target variable is categorical, we had to use a classification method (logistic regression modeling). It would have been even more convenient to use more classification methods such as Decision Tree, K- Nearest Neighbors Algorithm. However, since the Logistic Regression is a powerful classification method, we decided to only use it for this project.
The provided data set had a total of 3,333 customers of the total 483 were churning customers. For the modeling, to avoid over fitting the data to the model we split the data randomly to training data (trainingdata = 2444 observation) and a validation data (evaldata = 889 observations). The start was to explore the correlation among all variables in the dataset which revealed 5 variables with statistically significant correlation with the customer churn behavior (International_planyes/ voice_mail_planyes/ total_day_charge/ total_intl_calls/ number_customer_service_calls).

Initially, when attempting to improve the model, we focused on attribute removal for those attributes that provided little predictive power. While this is a worthwhile venture when scalability is concerned, for this exercise it is best to leave all original variables in place, regardless of their predictive power. Use of random selection for the training and testing data improved our area under receiver operating characteristics curve for our testing data reducing the possibility of an over fit. The two most important attributes according to the first model were membership in the international plan and number of customer service calls. Another observation from the initial model is that the area under receiver operating characteristics curve is inflated due to some records not being classified due to the presence of NAs in the model and the attributes to be classified. Additionally, false negatives outweigh true positives.

```
Call:
roc.default(response = trainingdata$churn, predictor = m1r)

Data: m1r in 1641 controls (trainingdata$churn no) < 278 cases (trainingdata$churn yes).
Area under the curve: 0.8412
         True
Predicted   no  yes
      no  1598  199
      yes   43   79
```

Because a significant indication of churn is number of customer service calls, we decided to multiply this attribute with another attribute indicating customer dissatisfaction, charges. This yielded very significant results for the product of customer service calls and day charge and still rather significant results for evening charge. Night and international charges, with p values between 2-3%, were less significant, but still provide desirable predictive ability. We generated three additional variables for ratio of each type of call to total of all type of calls. These new variables provided very significant results. We decided to try creating a similar ratio for each specific category of minutes to total minutes. This resulted in our logarithmic transformation of all types of minutes rather than just the two skewed types of minutes. This ratio had very little predictive power but did result in the idea to generate a new

variable. The new variable is average length of each type of call in logarithmic minutes! Note that the interpretation of this attribute is somewhat difficult, and its success could be attributed to over fitting.

When creating a new model with generated variables that exist in the original data, the significance of the original attribute increases if the new generated variable is significant, itself. This is can be thought of as amplifying the effect of the significant variable twice, once in the variable generation and another time by increasing the zstat for the original variable. We also noticed that some generated variables would result in the zstat for the original variable decreasing. This is a relationship which we are uncertain as to how it fully works. Maybe all the juice has been extracted from that attribute?

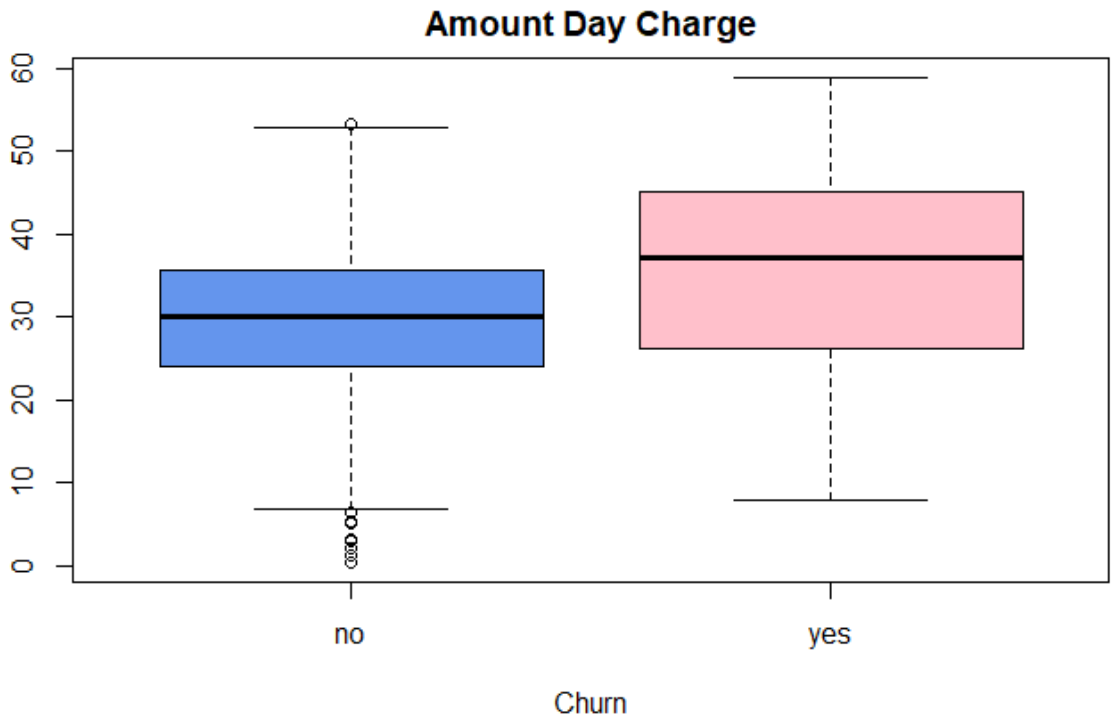
Estimation of model's performance:

Based on our testing data, our prediction for the area under the receiver operating characteristics curve is approximately 83.8% with a true positive rate of 37.5%. However, when this model was run on other group members' computers, the area under the receiver operating characteristics curve was greater than my result for the training data and lesser than my result for the testing data. We decided to present the prediction with the least spread between the two.

Insights and conclusions:

There is no doubt that reducing customer churn can have a large impact on the numbers of customer. In addition, the high cost to acquire new customers necessitates the need of telecom companies to reduce churn rate in order to decrease costs and boost revenue.
The data revealed:

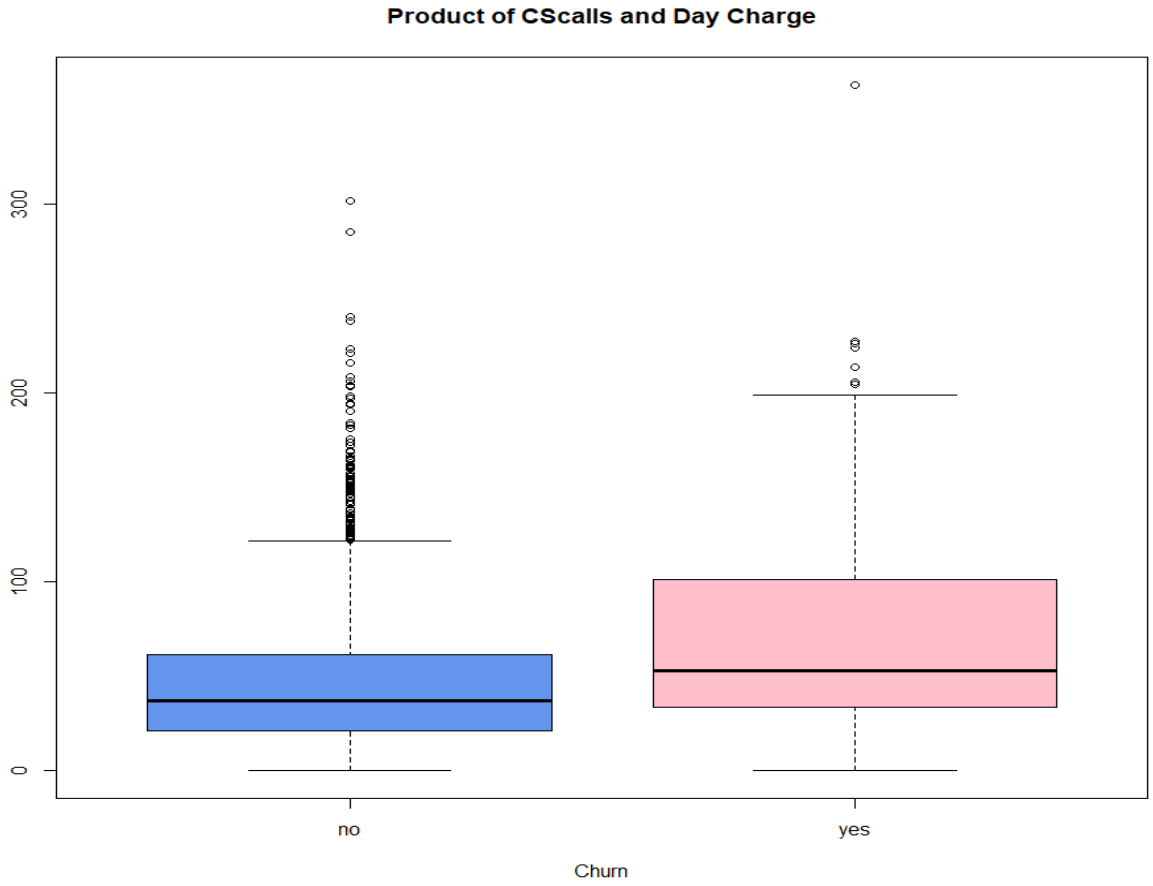
1- High churn rate among extreme day time users. It seems it is very likely that the more the customer pays the more likely they will look for cheaper providers.



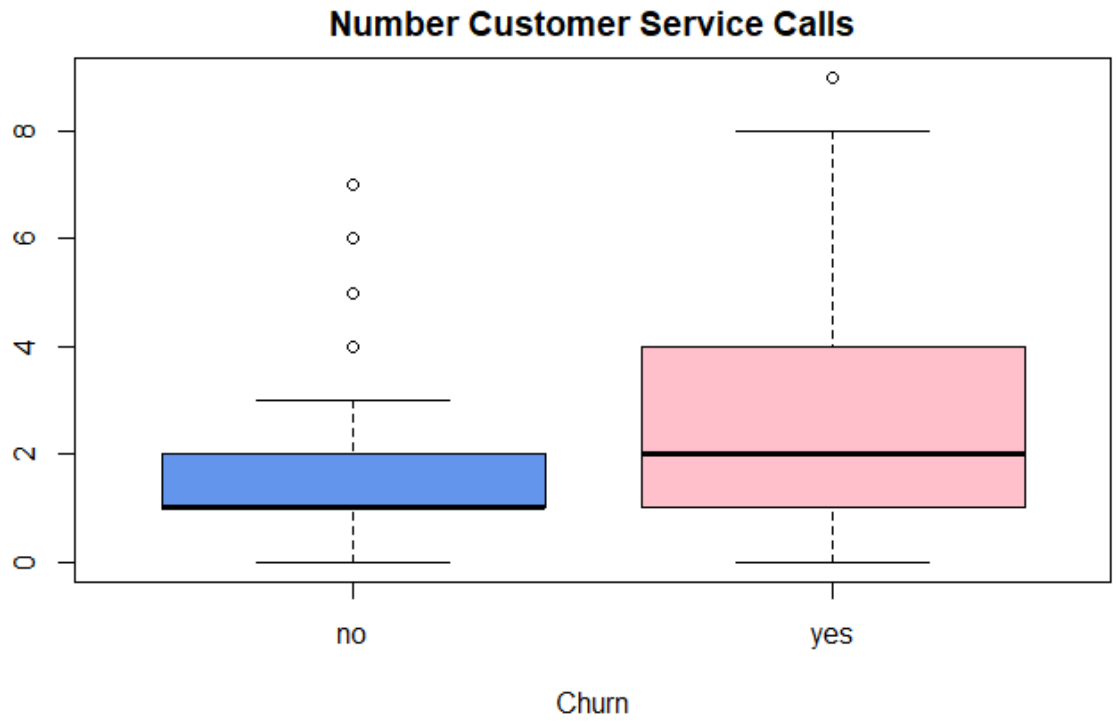
2- High churn rate among customers with international plans. Therefore, international plans will need to be restructured. The image below details the churn for the customers that have international plans. Customers that are enrolled in international plans are more likely to churn than those that are not enrolled. P(churn|IP-yes) = 43.3%

	Churn	
Int_Plan	no	yes
no	1957	254
yes	132	101

3- The product of two highly significant variables, customer service call and the day charges, creates a new variable that is also highly significant. Assumption is the higher the service center calls and the higher the day rate the higher customers will churn. The combination between both predictors creates a new highly significant variable. This indicates a correlation between the two variables and the churn outcome.



4- Customers with two or more service calls are more likely to churn. Telecom providers should improve response and customer service call centers to resolve customer issues in fewer than two calls. Increased number of customer service calls is an indicator of customer dissatisfaction.



5- Generation of a ratio of calls for a specific call type to the total calls for all call types, for all types of calls, indicates concentration of calls for each specific call type. Customers with a higher concentration of calls in a specific time period are more likely to churn. Interpretation in business might indicate that time specific promotions in either time of the day might increase customer satisfaction as the ratio for all 3 times of day yield very similar zstats and coefficients.

ratio_daytoallc	2.244e+02	9.139e+01	2.456	0.01405 *
ratio_evetoallc	2.385e+02	9.133e+01	2.612	0.00901 **
ratio_nighttoallc	2.398e+02	9.171e+01	2.614	0.00894 **

Note that no preference was indicated between the possible the types of error present and most prevalent in the classification results. Given the quantified cost of customer churn, it is implied that false negatives are more costly to ABC wireless than false positives. So while evaluating our model, we monitored area under the receiver operating characteristics curve, in addition to true positives and false negatives. One the model has been optimized, ABC wireless may wish to trade greater classifications of false positives for fewer classifications of false negatives. The end result would be that more customers likely to churn would be incentivized to stay with the company, while more customer unlikely to churn would receive those same incentives. It is up to management to decide if offering a greater amount of incentives to both customers likely to churn and those unlikely to churn is worth the potential reduction in churn that such an offering may provide. The screenshots for our evaluations of training and testing data follow.

```
Call:
roc.default(response = modtraining$churn, predictor = m2r)

Data: m2r in 2079 controls (modtraining$churn no) < 355 cases (modtraining$churn yes).
Area under the curve: 0.8771
          True
Predicted   no  yes
      no  2028  201
      yes   51  154


Call:
roc.default(response = modeval$churn, predictor = m2er)

Data: m2er in 754 controls (modeval$churn no) < 128 cases (modeval$churn yes).
Area under the curve: 0.8383
          True
Predicted  no yes
      no  727  80
      yes  27  48
```

Customers that have been identified with a higher churn rate should be targeted for an upgrade of monthly plan or other incentives to prevent churning. Especially offering incentives on International plans, improving customer service calls, and incentives to high day users. Using this model may reduce churn rate by up to 14%, leading telecom revenues to increase by millions of dollars.
In real world; annual churn rates in the wireless industry varies by company and ranges between 1.23% - 2.8% Verizon scored the lowest and Sprint scored the highest between the last quarter of 2013 and the first quarter of 2018.[1]

After looking into the results of the data modeling it is apparent how important monitoring the churn rate and how the number of variables actually helped point out the ones that have more correlation and can impact the churn rate. It is likely that the customers with a high probability of churn had a greater number of significant predictors than those with lower probabilities of churn. The churn rate would help telecom companies identify gaps in advance and plan policies and actions before they lose customers. Based on our data we can conclude to some ways telecom companies can use to reduce the churn rate:

- Increase efficiency of customer service.

  Resolving issues and complaints quickly should be a priority in telecom companies industry and providing a hassle- free service should reduce churn rate, increase customer experience and customer loyalty.
- Adding more attractive services especially focusing on valued international plans.

  This is one of the great techniques to reduce churn rate, as the more the customer feels satisfied with the amount of services for the money, the more loyal they become.
- More personalized plans for excessive users

  Customers are more attracted to companies that recognizes their usage and would increase loyallty if they received a “tailored” experience. That should give customers the feeling of a specialized services.

[1] http://www.dbmarketing.com/telecom/churnreduction.html

Lastly, keep in mind that customer behavior is very dynamic and can change by time, therefore, the model should be reevaluated based on new data periodically to measure and detect any sudden customer behaviors. Additionally, Churn rates are based on the correlation between the variables and the churn probability, which does not necessarily indicate causation, therefore any recommendations or corrective actions should also be tested after applying.