ABC Wireless Inc Churn Analysis

MIS-64036-001-Business Analytics Final Project

Dr.Razavi Rouzbeh

Project Goal

Telecom companies generally spend more in acquiring new customers in comparison with retaining customers. ABC Wireless Inc. has hired our team to help them with their **Targeted** approach of identifying potential customers who might churn in the upcoming days. It is to be understood that misidentifying customers would result in redundant expenses for the company as the same customers would have stayed with the service provider anyway. Our goal is to deliver a model using logistic regression that identifies maximum potential churn customers (true positives) and has minimum misidentifications (false positives).

Overview of data, including data exploration analysis

A model is only ever as good as the data that drives it. For that reason, it is essential to spend time investigating the properties of your data and whether any manipulation may be in order prior to beginning the modeling process. For our data exploration and clean-up we followed the steps outlined below.

1) We reviewed the data for any null values. If there are null values in the data, you must decide how they will be treated: will they be removed, imputed, or something else? A quick run of the code below confirmed that there were no null values in either the training dataset or in the customers to predict:

colMeans(is.na(churnTrain))

colMeans(is.na(Customers_To_Predict))

2) We ran *summary(churnTrain)* to provide a high level view of all of the data within the data frame. From this it became clear that we had a few categorical variables (state, international plan, voicemail plan, area code, churn) and that the remaining variables were continuous. The quartiles, mean, and median for the numerical variables gave the impression that most were normally distributed (in most cases, mean and median were roughly the same, and quartiles appeared evenly distributed around those values). However, the variables below appeared to be skewed:

```
> summary(churnTrain$number_vmail_messages)
   Min. 1st Qu.  Median          Mean 3rd Qu.  Max.
  0.000   0.000   0.000   8.099  20.000  51.000
> summary(churnTrain$total_intl_calls)
   Min. 1st Qu.  Median          Mean 3rd Qu.  Max.
  0.000   3.000   4.000   4.479   6.000  20.000
```

3) While logistic regression does not require that variables follow a normal distribution, we still sought to normalize our skewed variables as much as possible, just to see if it made any amount of difference in our model. Using the *skewness* function, we were able to test whether or not we were able to reduce the skew of the variables above. We also used this function to confirm that the other variables showed little to no skew. The skew of the total international calls was able to be drastically reduced (1.32 to -0.19), while the skew of the voicemail messages was only slightly improved (1.26 to 0.99). In both cases, log function proved the most effective, with both variables requiring the addition of a constant to prevent log(0).

4) Variables that are highly correlated result in redundancies when modelling. To check for correlations between variables, we used the corrplot function.

Using this, it was clear that there was a correlation near 1 for the minutes and charges associated with day, evening, night and international calls. Intuitively, this makes sense – all customers are being charged the same rates and the minutes and charges are merely transformations of each other, based on the plan rate. We pulled the exact correlation between these variables to confirm they are close to or equal 1, using the cor function. In all cases, correlation was .999 or 1. We confirmed that the use of either the minutes *or* the charges would be equivalent, by running one model with minutes and the other with charges; the z-scores for each variable and the overall AUC was the same for both models.

5) Additional variables that may be of interest in a model were created based on the results of our rough model in step 4) and intuition:

a. Normalized international calls
    i. churnTrainClean$total_intl_calls_nm = log(churnTrain$total_intl_calls+1)
b. Normalized voicemail calls
    i. churnTrainClean$number_vmail_messages_nm = log(churnTrain$number_vmail_messages+.000001)
c. Sum of all charges. Intuition suggests that if your bill is larger, you may be more likely to churn. These charges were summed, as we thought the total charges may be a more powerful predictor than the charges separated by time of day. If the model with the sum of charges proved to be as good as modeling each variable separately, the use of the summed charges would also simplify our model.
    i. churnTrainClean$total_charge = churnTrainClean$total_day_charge +

churnTrainClean$total_eve_charge +
churnTrainClean$total_night_charge +
churnTrainClean$total_intl_charge

d. Sum of all domestic charges. We wanted to be open to the possibility that international charges by themselves may be significant on their own, and in a different way, then domestic.

i. churnTrainClean$total_charge_no_intl = churnTrainClean$total_day_charge + churnTrainClean$total_eve_charge + churnTrainClean$total_night_charge

e. Rather than using all of the states in the model, we created an indicator to let us know if a customer was from a state of statistical significance. The z-scores for each of these states were positive, so in all cases, customers from these states were more likely to churn. If these states proved to be a good predictor when combined, it would be a simpler model than having all the states included

```
churnTrainClean$state_ind <- ifelse(churnTrainClean$state == "CA" |
                churnTrainClean$state == "ME" |
                churnTrainClean$state == "MI" |
                churnTrainClean$state == "MS" |
                churnTrainClean$state == "MT" |
                churnTrainClean$state == "NJ" |
                churnTrainClean$state == "NV" |
                churnTrainClean$state == "SC" |
                churnTrainClean$state == "TX" |
                churnTrainClean$state == "WA", "Y","N")
```

f. While the number of day, evening and night calls were not statistically significant by themselves, we wanted to check if the total number of calls combined were:

i. churnTrainClean$total_calls_no_intl = churnTrainClean$total_day_calls + churnTrainClean$total_eve_calls + churnTrainClean$total_night_calls

ii. churnTrainClean$total_calls = churnTrainClean$total_day_calls + churnTrainClean$total_eve_calls + churnTrainClean$total_night_calls + churnTrainClean$total_intl_calls

6) Finally, we calculated some overall summary data to get a feel for what we may expect to see in our model.

The overall percent of churn within the churnTrain population was calculated and compared to the churn rate of those with a voicemail plan, and those with an international plan. Based on the data below, we anticipated our final model to confirm that those with an international plan were far more likely to churn than the regular population, and those with a voicemail plan were less likely to churn.

churn	Overall_churn_percent	Intl_plan_churn_percent	Vmail_plan_churn_percent
1 no	85.5	57.6	91.3
2 yes	14.5	42.4	8.68

For numerical variables, averages were calculated and compared among those who churned and those who did not. To minimize the number of variables in the output, we utilized the total domestic charges instead of day, evening, and night individually. In the table below, d = domestic, I = international, and cs = customer service.

Based on the summary data below, variables we might expect to influence whether customers churn include: number of voicemails, total domestic charges, and total customer service calls. It seemed possible international charges and the number of international calls could prove significant and seemed unlikely that account length or a number of domestic calls would be significant.

churn	avg_vm	Avg_length	Avg_d_charge	Avg_i_charge	Avg_d_calls	Avg_i_calls	Avg_cs_calls
1 no	8.60	101	55.7	2.74	300	4.53	1.45
2 yes	5.12	103	62.5	2.89	302	4.16	2.23

At this point, we felt prepared to begin iterations of modeling using our variables and had a rough idea of some of the output we may expect to see based on the summary data.

Details of your modeling strategy (i.e. what technique and why)

The team's modeling strategy is to include all the variables in the generalized logistical model function in r and analyze the output to determine the performance of the model. The team considered logistic regression as the foundation of the model.

Reasons for choosing Logistic regression:

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, logistic regression is a predictive analysis
- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- Linear regression assumes a linear relationship between each variable. However, if the customer is paying less for a particular service compared to a competitor, the chances of the customer leaving the service provider will vary exponentially.
- Linear regression assumes that the probability increases proportionately as the independent variable increases.
- However, when it comes to predicting a discrete variable — for example, whether a customer will stay with a service provider or not, or whether it will rain or not —Logistic regression will give us better prediction in various aspects.

In order to analyze the output of the generalized logistic model, our team used the *summary()* to review the output of the model. Through doing this, we were able to examine the significance of each of the variables utilized within the model. The team found 6 significant variables when creating the model. The significant variables will be analyzed and discussed within the "Insights and Conclusions" section of this paper.

Estimation of the model's performance

There are so many ways one can access the performance of the logistic regression model. we have considered the **ROC curve and AUC**.

- Receiver Operating Characteristic (ROC) is basically, ROC curve is the plot of true positive rate (TPR)(sensitivity) against the false positive rate (FPR)(1-Specificity) using different cutoff points.
- In the ROC plot, you want your points on the curve to get closer to the northwest (0,1) for your model to be more accurate. The closer your points are to the diagonal line, the less accurate your model is. AUC is the area under the curve of ROC.

The ROC values can be found using the command

```
roc(data_Test$churn, Predict_test_final)
```

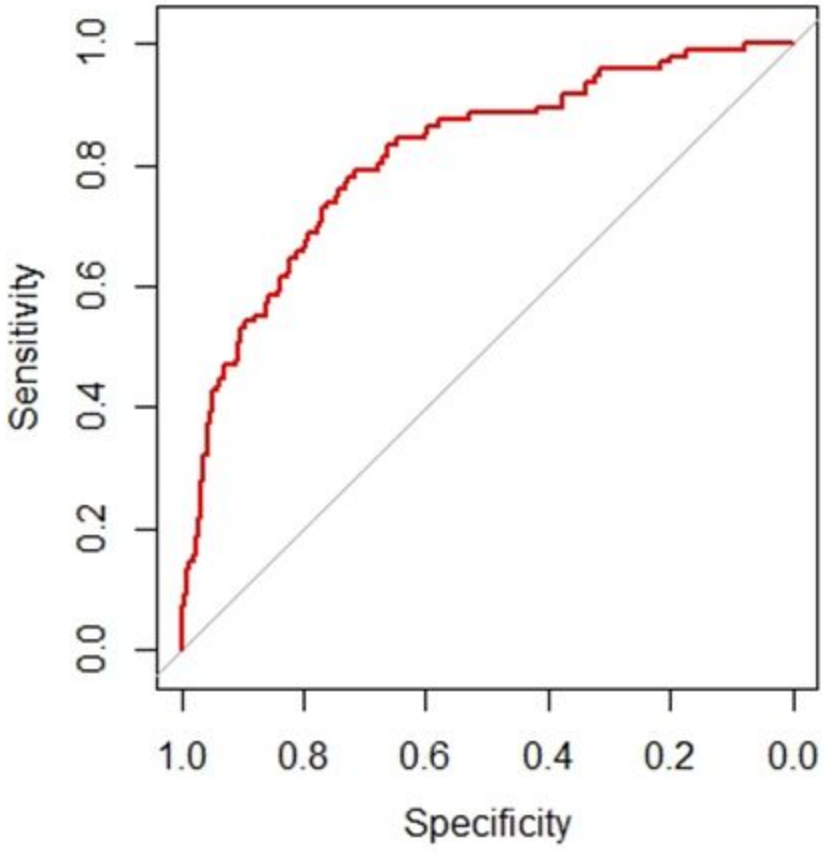
```
Call:
roc.default(response = data_Test$churn, predictor = Predict_test_final)

Data: Predict_test_final in 570 controls (data_Test$churn no) < 96 cases (data_Test$churn yes).
Area under the curve: 0.852
```

As seen from the image, the area under the curve, AUC = 85.2%

The same can be plotted in a graph using the command

```
plot(roc(data_Test$churn, Predict_test_final), col="red")
```



Insights and conclusions

With an ROC value of 85.2 % we can safely infer that the proposed model is pretty good and reliable. On analyzing the model using the ***summary()*** function we were able to determine some key factors that dominantly influence the customers to churn.

```
> summary(Model_ABC_Wireless)

Call:
glm(formula = churn ~ state_ind + international_plan + voice_mail_plan +
    number_vmail_messages + total_intl_charge + number_customer_service_calls +
    total_charge_no_intl + total_intl_calls_nm, family = "binomial",
    data = data_Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9602  -0.5104  -0.3324  -0.1872   3.0811

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -7.480526   0.540154 -13.849  < 2e-16 ***
state_indY                    0.881275   0.140904   6.254 3.99e-10 ***
international_planyes         1.938102   0.163601  11.847  < 2e-16 ***
voice_mail_planyes           -2.269772   0.642366  -3.533  0.00041 ***
number_vmail_messages         0.045079   0.019871   2.269  0.02330 *
total_intl_charge             0.245631   0.084554   2.905  0.00367 **
number_customer_service_calls 0.518854   0.044296  11.713  < 2e-16 ***
total_charge_no_intl          0.079150   0.006484  12.206  < 2e-16 ***
total_intl_calls_nm          -0.587705   0.144803  -4.059 4.94e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2209.0  on 2666  degrees of freedom
Residual deviance: 1718.6  on 2658  degrees of freedom
AIC: 1736.6

Number of Fisher Scoring iterations: 6
```

The key parameters in their order of significance are as follows:

Churn Influencers	
Attributes	**Z-values**
Total Charge (domestic)	**12.206**
International_Plan (Yes)	**11.847**

No. of Customer Service Calls	**11.713**
States ("CA","ME,"MI","MS","MT","NJ","NV","SC", "TX","WA")	**6.254**
Total International Calls	**-4.059**
Voicemail Plan (Yes)	**-3.533**
Total International Charge	**2.905**
No. of Voicemail Messages	**2.269**

Total Charge (Domestic - Excluding International)

The parameter that we found has the most influence in churning a customer is the charges the customers are getting charged. This parameter is obtained by aggregating the day, evening and night charges together. From analyzing the data available we were able to infer that there were no significant differences among the rates offered to customers overall either during the day or evening or night. So the total calls duration has a direct proportion with the charges i.e., higher the consumption higher the charges. This factor seems to have the most influence on customers.

International Plan_Yes

The second most significant parameter in predicting customer churn is the categorical variable ***international_plan.*** We recognized that customers who have their International Plan activated are more likely to churn in comparison with those who have not. The rationale behind this could be multiple factors: the customers could be foreign and they are in the country temporarily. For instance, students, tourists, international businessmen, etc. Another reason could be that the customer could be traveling internationally on a regular basis and there could be a compromise

on the service quality due to factors that are not entirely under the control of the telecommunication company.

Number of Customer Service Calls

The parameter that falls next in line would be the '**number_customer_service_calls**'. It is quite intuitive as calls to customer service are usually made more often than not due to having issues with the telecommunication service. Customers who are likely to have issues with their network service are probably more likely to churn.

States

From the descriptive analysis, we were able to identify certain states that had a significant influence on the customer churn in comparison with others. The following are those significant states: Montana, California, South Carolina, Texas, New Jersey, Washington, Michigan, Mississippi, Maine, Nevada. Some of the primary reasons that contribute to these variations could be varying network coverage, presence of regional competitors, state tariff rates, etc.

Total Number of International Calls

We infer from the model that the parameter '**total_intl_calls**' has a negative correlation effect with the churn. It is to be understood that these customers have their international plan activated and the fewer the number of international calls they make the higher are the chances that they will churn. This could be attributed to absence from the country, poor service offered, non-affordable rates for international calls, etc.

Voicemail Plan

The variable 'voice_mail_planyes' seems to have a negative correlation with the churn. i.e. The more the customers has Voicemail plan activated the less are the chances that they will churn. Voicemails are activated for most customers these days and a failure to activate a voicemail plan indicates a lack of usage and familiarity with the system.

Total International Charge

The significance of 'Total International Charge' goes inline with the total international calls made. High charges for international calls prompts the user to make fewer international calls in spite of opting for an international plan.

Number of Voicemails

The significance of '**number_vmail_messages'** implies that the customer might not be an active user or there could be the case of poor network reception which probably directs the calls to a Voicemail box.

Conclusion

The proposed model serves the purpose of identifying potential customers who might churn. Using this model the team was able to figure out the significant factors influencing the customer to churn. Based on the analysis, the team has come up with the following suggestions that might help the organization neutralize or lower the churn rate.

- The total domestic charge is identified as a key factor and from the data, we see that there were no significant differences among the rates offered to customers overall either during the day or evening or night. As the company follows the '**Targeted Approach**' this situation could be handled by proactively identifying the customers with high domestic charges and extending customized offers to those customers.
- The model has identified the following states to have a more churn rate than the rest. Montana, California, South Carolina, Texas, New Jersey, Washington, Michigan, Mississippi, Maine, Nevada. It would be advisable to allocate more resources to look into the services offered in these states to assess the discrepancy.
- Another key factor is the number of calls made to customer service. Apart from identifying the issues that prompt the customers to contact customer service, focus on providing improved customer service by addressing and resolving the issues might be an ideal approach.