**ABC Wireless Inc.**

**Business Analytics - MIS-64036   
Kent State University**

**Group Members**

**Devesh Petwal:** Completed the presentation and assisted in visualizations for the overview of data section. **Elham Zare:** Completed the overview of the data section and introduction.

**Nicholas Golina:** Completed the results, and insights and conclusions section and contributed to the modeling.

**Satyasri Pavani Harika Penjerla:** Completed the modeling report section and assisted with the modeling design.

**Contents**

Introduction…………………………………………..3

Overview of data…………………………………….4

Model Planning……………………………………...6

Model Building……………………………………….10

Estimation of Model’s Performance……………….12

Results……………………………………………….13

Insights and Conclusions…………………………..15

**Introduction**

In the modern days of business competition, most of the telecom companies suffer from losing their customers to competition (churn). Churn is one of the main concerns of these companies because it is more costly to acquire new customers rather than keeping the existing one and it lowers profitability. In addition, Churn rate influences the lifetime value of the customer significantly because it affects the length of service and the future revenue of the company. The higher the churn rate is the shorter is the average customer lifetime. Telecom companies spend hundreds of dollars to acquire a new customer by advertising and offering special offers. However, when they lose the customer, the company loses the future revenue from that customer as well as the resources spent to acquire that customer.

There are several [reasons for customer churn](https://blog.hubspot.com/service/reasons-for-customer-churn) that are personal and unique to each customer, but they usually fall under a few common categories, for instance, price, product/market fit, user experience and customer experience. The path to successfully retain customers includes many factors and activities like onboarding, activation events and consistent feedback. Without defining the goals for declining churn, customer retention initiatives are not going to move to the top of an organization’s priority list. The lack of involvement to a churn goal is one of the main reasons of industries having problems in customer retention. If companies set a churn goal, all members and teams must be activated to respond. However, often teams are not aware of how and what systems should be involved in gathering and surfacing the correct data for a retention initiative. Moreover, sometimes the teams don't work together or within the same system. For churn to be reduced, it often requires coordination and cooperation across departments.

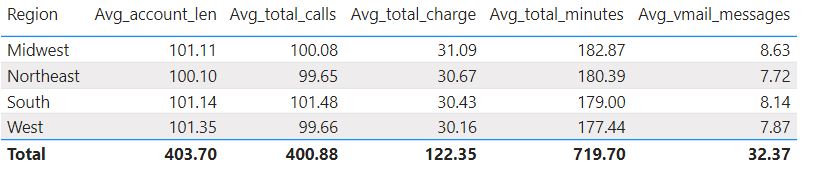
Telecom companies use mostly two approaches to address churn. First approach is Untargeted approach that relies on superior product and mass advertising to increase brand loyalty and hence retain customers. The second is targeted approach where the companies attempt to identify customers who are likely to churn, and provide suitable intervention, special programs or incentives to motivate them to stay. This approach is, however, risky and can cause huge financial and timing burden if the company doesn’t recognize the real churner.

In this project, our main analysis and goal are on enabling churn reduction in ABC Wireless Inc. and identifying the customers who are likely to churn. Churn prediction is one of the most common and popular Big Data use cases in business. It comprises detecting customers who are likely to cancel a subscription to a service and move to competitors. Being able to predict churn based on customer data has proven extremely valuable to big telecom companies.

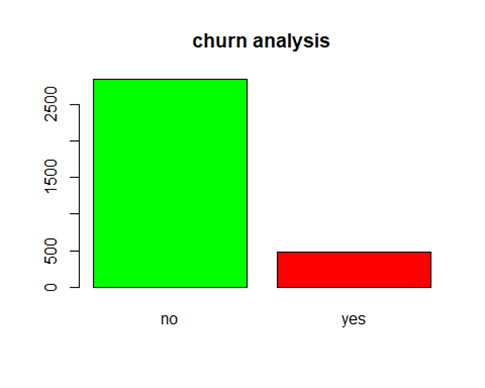
**Overview of data**

For our dataset, there were several variables, mostly numerical, to consider. As part of our analysis, we explored variables one by one and after identifying the data type and category of the variables, we created a dummy variable for several, categorical variables represented in the data. This included particular voicemail plans, the region the customer fell in and those quantitative variables that include like total day, minutes calls, charges, regions etc. There were several trends in the data that were important to consider.

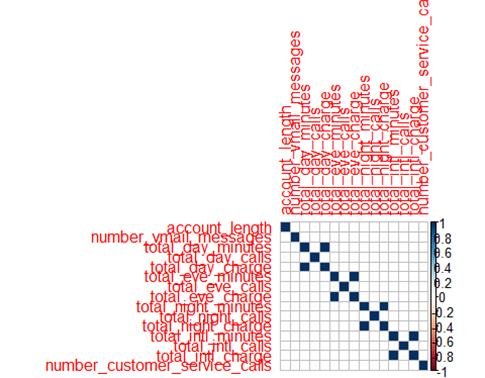
For the regions of US states the variation among these metrics like total calls, account length, charge, minutes, and vmail messages is very small. This is surprising since there usually is geographic variation when approaching any problem.



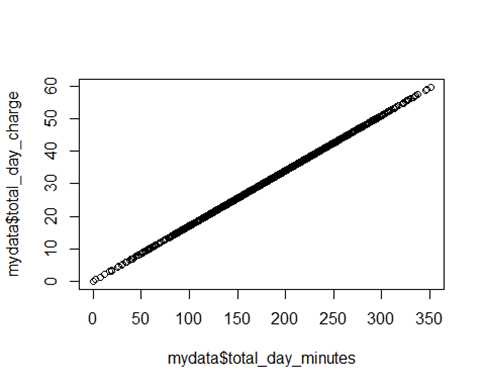
The below graph shows the overall churn in dataset:



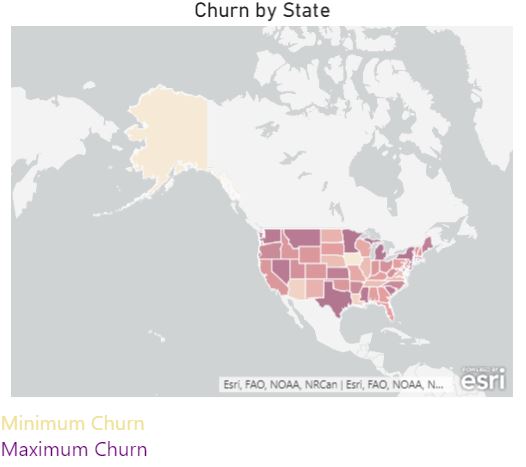
We found a correlation between the variables. The following graph displays the correlation between our data that are distributed between positive to negative one.



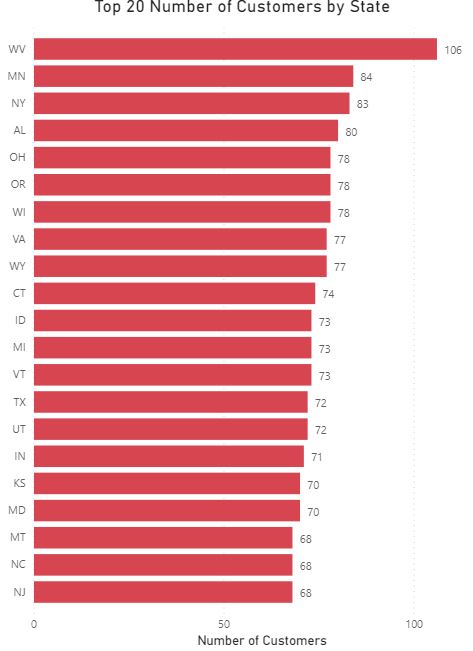
In the following graphics, we displayed the linear relationship between our variables, which served as the basis for dropping some variables from the model due to multicollinearity:



Here the churn rate by state can be examined. This shows a lot of geographic heterogeneity, which could be a factor in churn predictions.



The top 20 states by number of customers can be as well showing a lot of variation in this regard. This could matter greatly in terms of the over representation of some states in the model, which is why it is important to look at these states by region.



**Model Planning**

Model Planning is mainly of two selections

a.Variable Selection

b.Model Selection

Detailed explanation of the above two selections are described as below

**a.Variable Selection**

Feature creation is a process to creating new variables based on existing variables. For instance, we have a date, month/day/year, as an input variable in a data set. We can generate new variables like day, month, year, week, weekdays that may have better connection with target variable. State variable is rearranged as four regions like

Midwest

Northeast

South

West

Eliminated Total\_day\_charge,total\_night\_charge,total\_eve\_charge,total\_intl\_charge due to collinearity.Area\_code is also eliminated because it is redundant variable.

**b.Model Selection**

Regression analysis is a form of predictive modeling technique which gives the relationship between dependent variable and independent variables. Regression analysis also indicates the strength of impact of multiple independent variables on dependent variable. It is used for building predictive models by evaluating most useful set of variables and eliminating variables which are redundant. There are various kinds of regression technique mainly depends on three aspects – type of dependent variable, number of independent variables, shape of regression line. Most common types of regression analysis are

* · Linear Regression
* · Logistic Regression

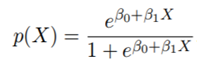
As said earlier, one of the important aspects in regression analysis is a type of dependent variable, in this project is churn which is binary categorical variable. Binary categorical variable has only two possible outcomes like 1 (yes) and 0 (no). In other words, if we are looking for mapping function (f) that can take a given vector of features (x) and predict a dependent variable (y) where ‘y’ is a categorical variable then is called classification. So, it is important for estimating the probability that x belongs to each of the categories of the dependent variable. In this scenario logistic regression is generally used which uses the logistic function because values strictly bounded ranges from 0 to 1.

For classification, we prefer probabilities between 0 and 1.

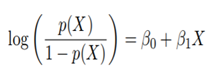
Let’s consider single variable model where

P(X)=(Y=1|X)

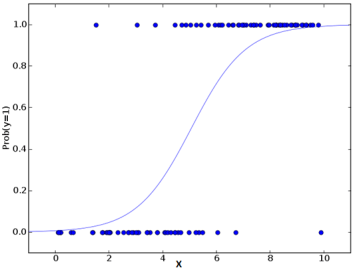
Logistic regression uses the form



We use log () function “odds” and wrapped in the logarithm is called log odds or logit transformation of p(X) which is monotone transformation. Odds can vary from 0 to infinity.



The typical logistic regression model plot is shown below:

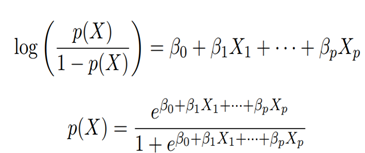


Next comes, Maximum Likelihood is probabilistic framework for optimizing models and to estimate the parameters. It is the objective function for the model.



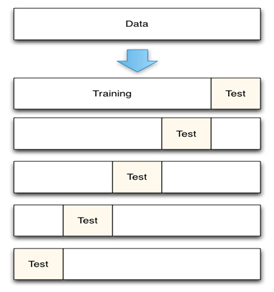
The above equation gives the probability of the observed zeros and ones in the data. We must pick 0 and to maximize the likelihood of the observed data points. The likelihood ‘L’ is high for P value closer to 1 when outcome is 1 and is closer to 0 when outcome is 0. Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm() function to implement this model. The coefficient, β, in the logistic regression can be interpreted as follows. For every unit increase in X, the logarithm of the odds ratio of the Y increases by β.

In this project, we have several independent variables and one dependent variable. So, the model will be in the form of



Key assumption in Logistic regression is that the dependent variable should be discrete or categorical variable. Also, one of the flexibilities using logistic regression is that independent variable does not have to be in normal distribution.

Before building the model in R, we use K-fold cross validation, is a resampling procedure. Single parameter in this approach is K that refers to number of groups that given data sample is to be split into. We use k=5 which splits the data into five groups. It considers one group as test and remaining group as training dataset and fit a model on training set and evaluate it on test set. Retain the evaluation score and discard the model. Finally, summarize the sample model evaluation scores. Below is a pictorial representation of cross validation.

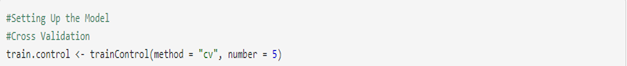


**Cross-Validation with K=5 folds**

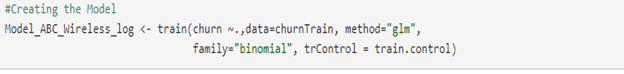
**Model Building**

Implementing model using R programming language using R-Studio environment.

After data exploratory analysis we did setting up the model using cross-validation with k=5 using method “cv”

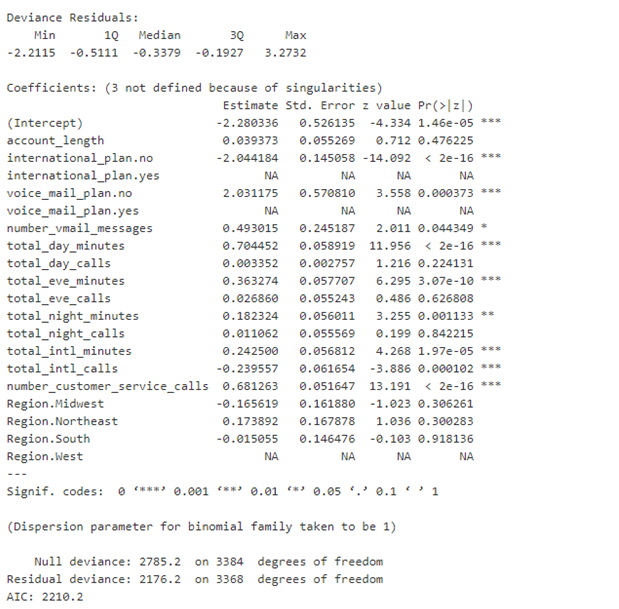


Creating the model using “glm” method for logistic regression in R.



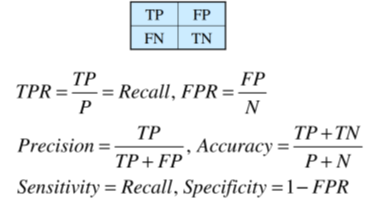
Summary of the model is shown below:



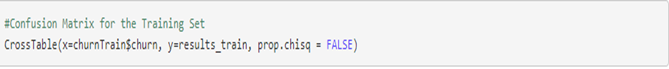


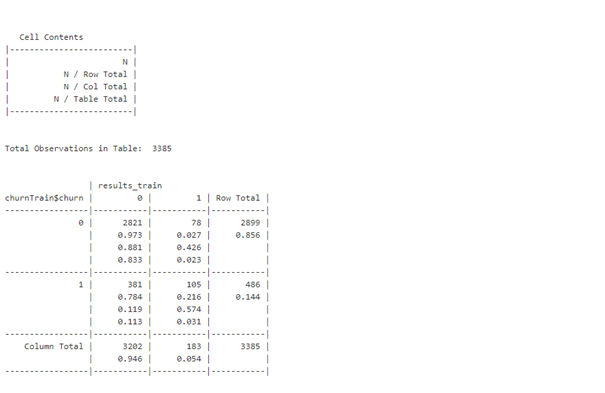
**Estimation of Model’s Performance**

1. Confusion Matrix: It is a tabular representation of actual and predicted values. This is important to find the accuracy of the model and avoid overfitting. The various parameters that can be calculated by confusion matrix are as follows



Implementing in R



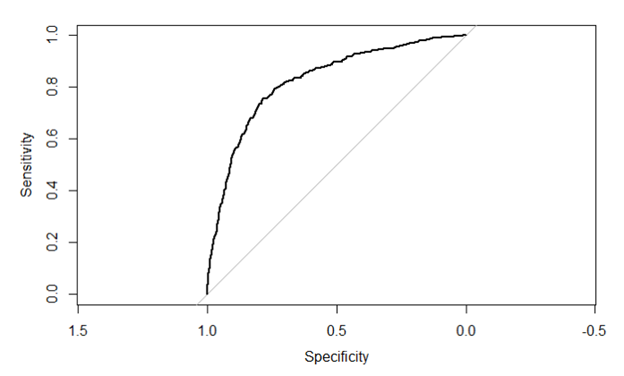


Accuracy for this model is 86.44%.

2. Receiver Operating Characteristics (ROC) curve: Summarizes the model performance by evaluating the tradeoff between sensitivity and specificity. The area under the curve referred as performance metric for ROC curve which is accuracy.

Implementing in R as

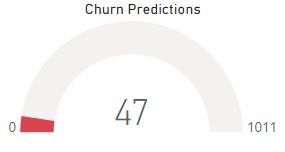
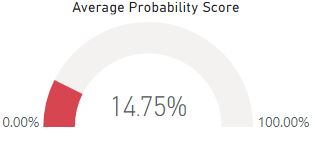




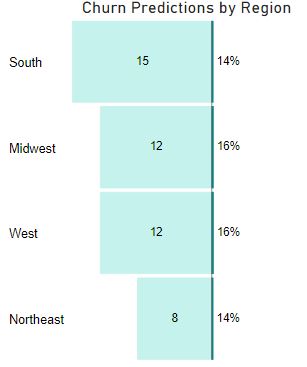
The ROC in this model with AUC is 82.42%. Similar results were observed for the testing set as well.

**Results**

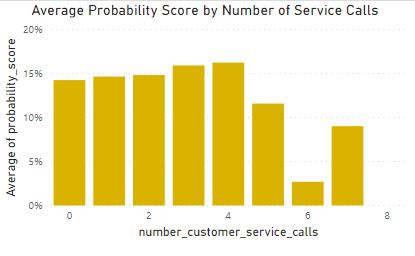
After making predictions on the dataset to determine churn risk, the results can be broken down in a number of different ways. The average probability of churn for the entire dataset was 14.84% and the number of churn predictions was 49 of the people sampled.



Additionally, when looking at the churn risk and probability of churn by census region there are many commonalities. Overall, the average probabilities are in the range of 14% to 16%. The Southern region had the highest predicted churn, whereas the northeast region had the lowest churn predictions.



Another finding was that the average probability increased gradually when customers had to make multiple customer service calls up to 4. This could be evidence that customers do not like to utilize the customer service route multiple times due to the time it takes to complete and resolve the issues.



**Insights and Conclusions**

In conclusion, customer churn is a risk for a small amount of people. However, it is important to analyze the trends within the data so that the reasons for customer churn can be elucidated. It is clear from the results that regional differences do not seem to explain the presence of high probability scores. Efficiency improvements can be made to reduce the risk of churn through the customer service route based on usability testing from other competitors. Another recommendation is to pay close attention to the users that are most active with their phones. These people should be monitored and kept in the loop of any new plans and improvements to the program the customers are a part of. This is necessary to ensure the trust of the end user and potentially prevent customer churn from occurring.