Andrew Johns, Tina Xu, Sonal Kanabar, Choe samba

Customer Churn rate analysis

For ABC Wireless

***Problem Statement***

ABC Wireless Inc. has hired us to help them with the customers’ churn issue. In this project, we will be working as a part of a team to use historical data from ACB Wireless Inc. to build a model that can predict/identify their customers who are likely to churn.

***Project Goal***

Our goal is to help predict and identify customers for ABC Wireless Inc that are likely to churn. And help predict future actions that can reduce the churn rate. We were able to do so through the following actions.

* **Discovering the data:** We dove into the dataset to identify the meaning of each variable and eliminate any noise.
* **Data Preparation:** Checked the quality of the dataset and cleaned it to prepare it for modeling planning.
* **Model Planning:** Incorporated logistic regression and multiple variables to figure out the model that we plan to use. GLM function to classify whether the variables would be churn or not.
* **Modeling Building and Predicting:** Tested the data of the Modeling Planning; to help identify if customers would churn or not based on our model.

***Overview of data***

To identify in advance customers who are likely to churn we have utilized the Churn\_Train.csv Dataset which contains descriptors that can be divided into three different categories:

* Graphical data: state, area code, etc.
* Account information: number of complaints, number of voicemails, length of account, etc.
* Services used: International, voicemail, etc.

We will also have a column with two labels for churn: Yes and No, which is our target to predict

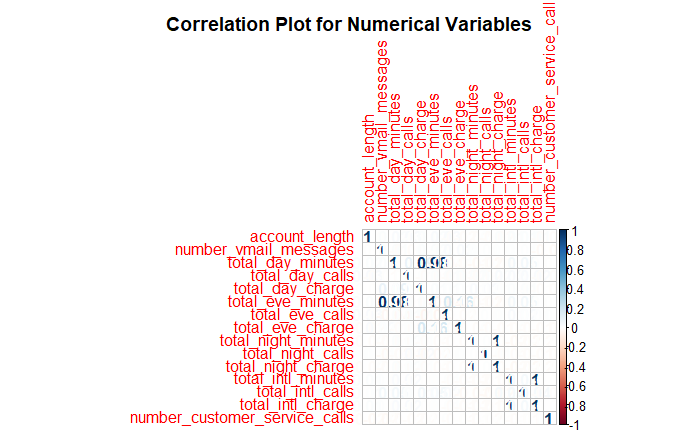
Below are the variables that we will be using to measure the historical data from ACB Wireless Inc. Each variable includes an explanation and we have highlighted the 7 variables that we hypothesize will have to the greatest impact on the likeliness of the customer to churn.

1. **State (categorical):** The state the customer resides in
2. **Account\_length:** How long the customer has been with the company
3. **Area\_code:** Area code the customer resides in
4. **International\_plan (yes/no):** Does the customer have an international plan?
5. **Voice\_mail\_plan (yes/no):** Does the customer have a voicemail plan?
6. **Number\_vmail\_messages:** # of voicemails the customer has left for customer service.
7. **Total\_day\_minutes:** Total # of minutes spent daily
8. **Total\_day\_calls:** Total # of calls
9. **Total\_day\_charge:** Daily amount charged to customer
10. **Total\_eve\_minutes:** Total # of minutes spent during evening hours
11. **Total\_eve\_calls:** Total # of calls during evening hours
12. **Total\_eve\_charge:** Amount charged to customer during evening hours
13. **Total\_night\_minutes:** Total # of minutes spent at night
14. **Total\_night\_calls:** Total # of calls a night
15. **Total\_night\_charge:** Amount charged to customer during night calls
16. **Total\_intl\_minutes:** Total # of minutes spent during international calls
17. **Total\_intl\_calls:** Total # of international calls
18. **Total\_intl\_charge:** Amount charged to customer during international calls
19. **Number\_customer\_service\_calls:** Numbers of calls made to customer service requesting assistance

***Data Exploration Analysis***

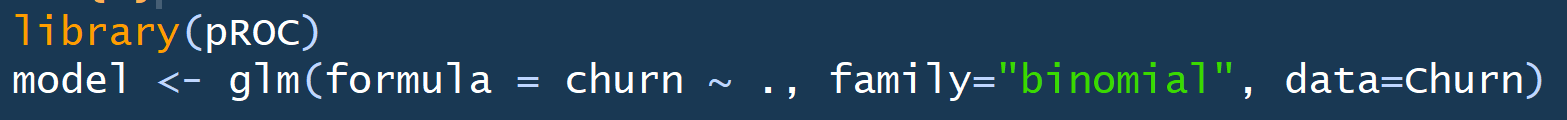
We found that the majority noise came from the variable Number\_vmail\_messages, identifying 200 lines that were almost exclusively filled with N/A values. We decided to remove these values. After removing the values, we saw there were still N/A values within account\_length, total\_eve\_minutes, and total\_intl\_calls. We decided to fill those N/A values with the average of each column to eliminate that noise as well. We also converted the values found in account\_length and number\_vmail\_ message to their absolute value, therefore eliminating any negative numbers. This decision was made on the assumption that the negative values were entered incorrectly.

We also used the Correlation Plot below to detect which variables were correlated and not needed when identifying variables that are significant to Churn. For exmaple it is not neccessaary to have both total\_eve\_minutes and total\_day\_minutes.



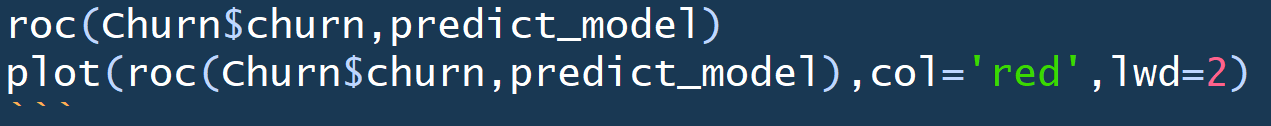
***Details of Modeling Strategy (i.e. what technique and why)***

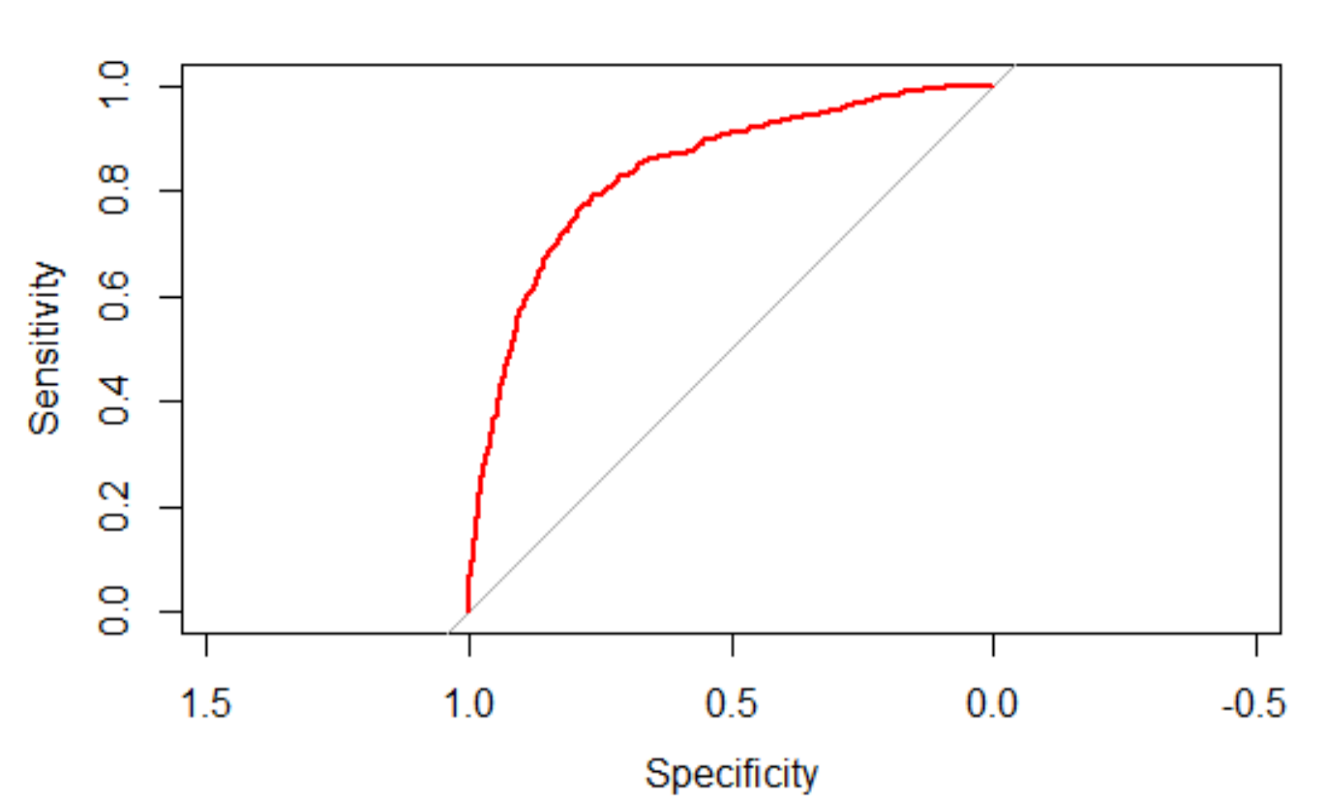
We decided to use the Generalized Linear Modeling strategy. We picked this model as we are classifying a binary variable and deciding whether it will be a “yes” or “no.” In R, this was done through the glm( ) function.

****

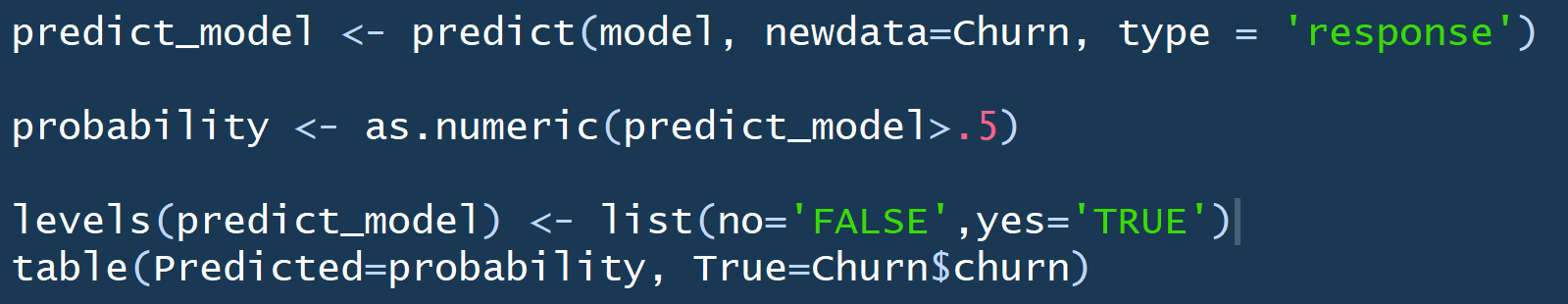
***Estimation of model’s performance***

We reviewed the model’s performance by using the RoC curve and maximized the Area under the curve. We believe that after cleaning the data and using all the available variables that we achieved a very high amount of AUC at 0.8377. Our code also includes a graphical representation of the RoC.





With this, we built a table to show the false-positives (type 1 errors) and false negatives (type 2 errors) using a threshold of the probability.



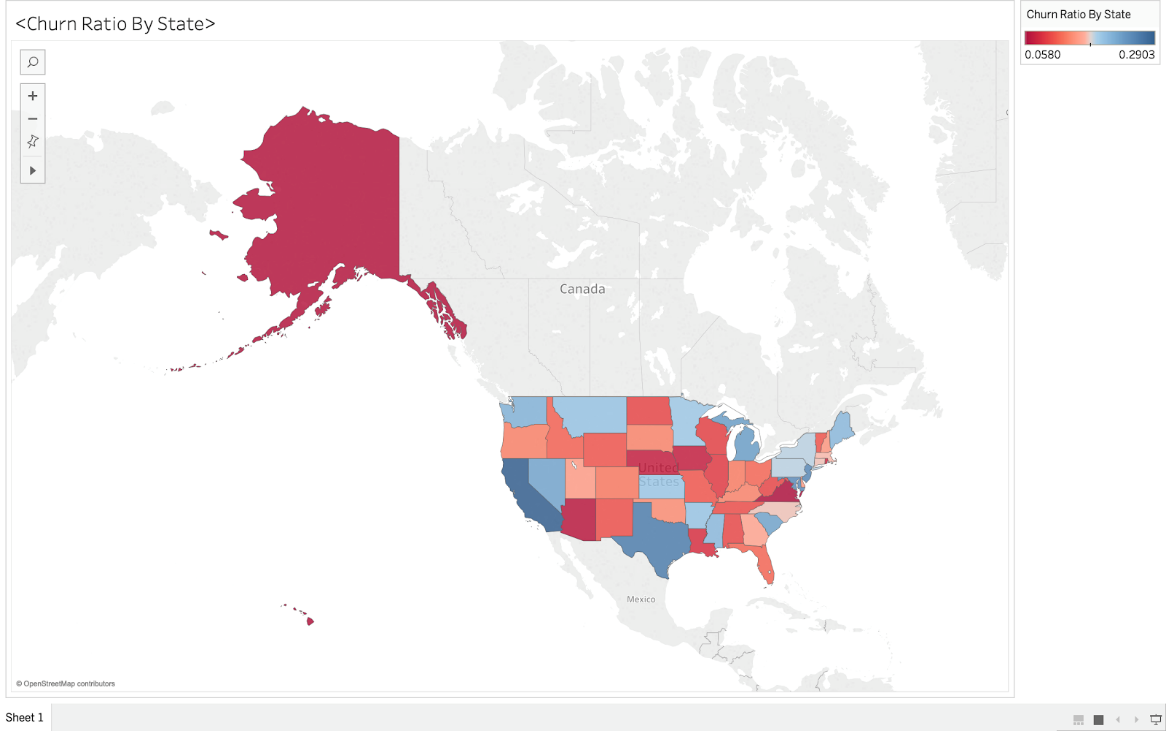
***Insights and conclusions***

After creating our model and reviewing the performance, we were able to make valid connections within the data. The variables that proved to have significant impact on the churn ratio were

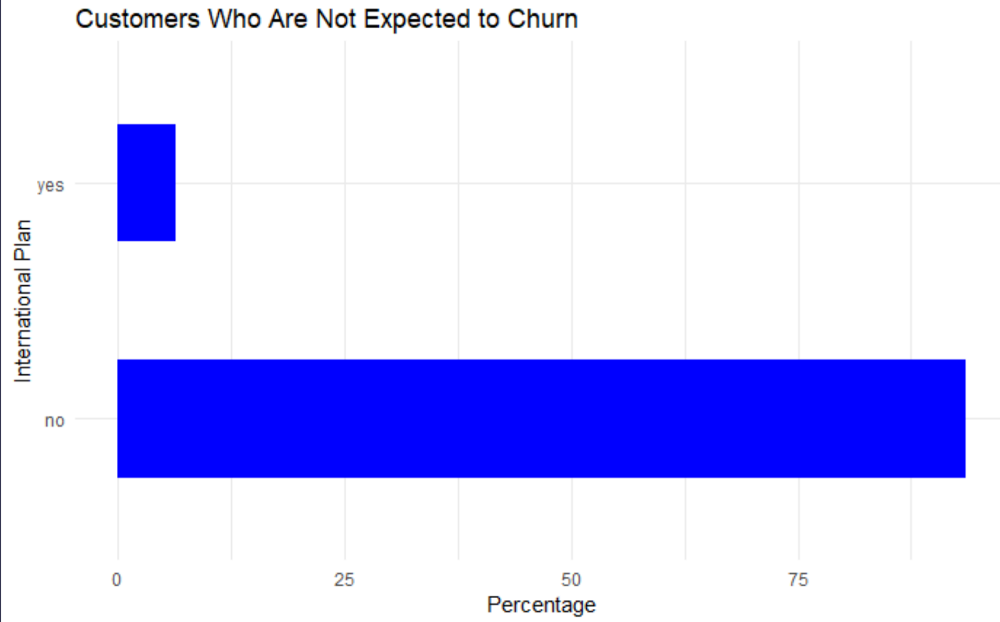
1. **State (categorical):** The state the customer resides in
2. **Account\_length:** How long the customer has been with the company
3. **International\_plan (yes/no):** Does the customer have an international plan?
4. **Voice\_mail\_plan (yes/no):** Does the customer have a voicemail plan?
5. **Total\_day\_charge:** Daily amount charged to customer
6. **Total\_intl\_calls:** Total # of international calls
7. **Number\_customer\_service\_calls:** Numbers of calls made to customer service requesting assistance

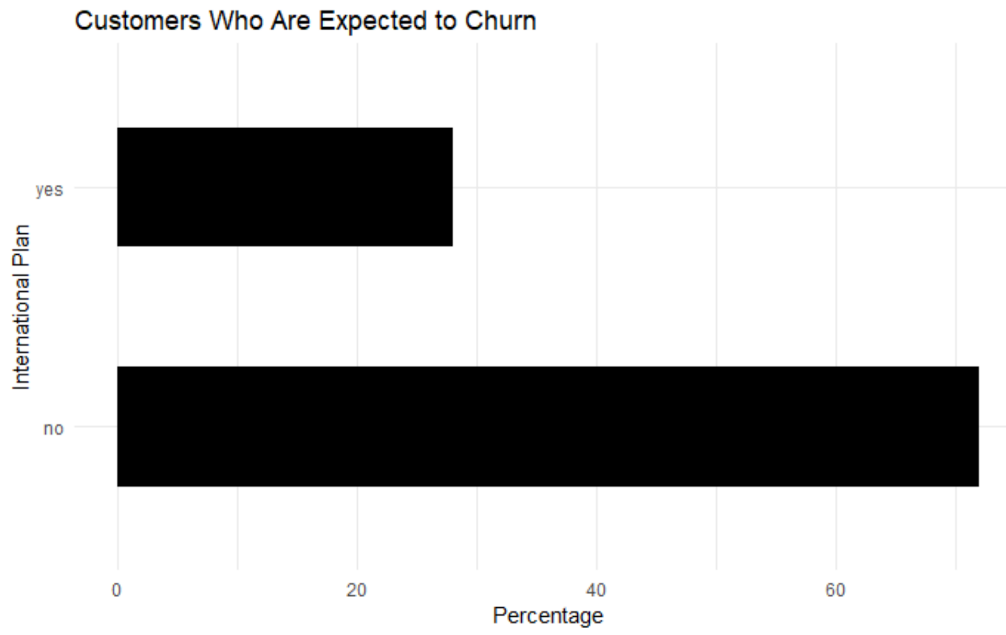
Most of the variables we believed to have significant predictive power proved to be correct. We also discovered other significant variables that weren’t initially expected to provide a connection to the probability of a customer churning.

In the graphic below we took a deeper look at the churn ratio for each state. The red equates to states below the average churn ratio and the blue equates to states above the average churn ratio.



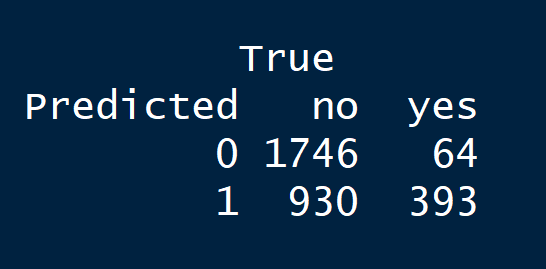
The binary column of international\_plan proved to be a significant variable and below are graphs that support this. Just under 30% of individuals that are expected to churn have international plans compared to under 10% for individuals that are expected to change.

****

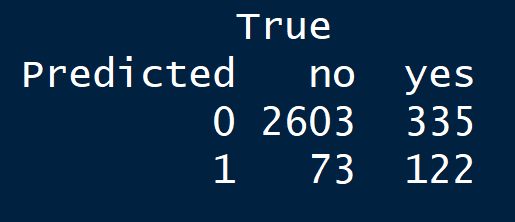


When we built a table to show type 1 and type 2 errors, we decided to show the extreme values for the threshold of probability and a few values in between. This will provide a range of results and allow the executives the flexibility to choose the best option for their situation, while still providing our recommendation. Below is the breakdown for the 10%, 50%, and 90%. We believe that a threshold of 55% is the best as it has a good balance of minimizing both the type 1 and 2 errors at roughly 13% of the data while providing a high ratio of correct positive predictions to false positive predictions at nearly 2:1 (Both areas where the company would spend money to prevent churning).

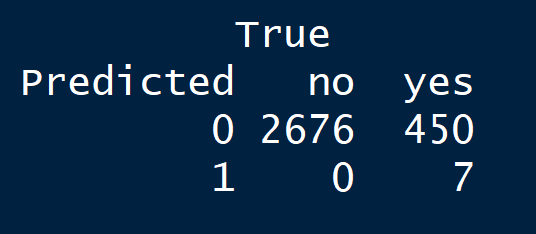
**10% Threshold:**



**50% Threshold:**

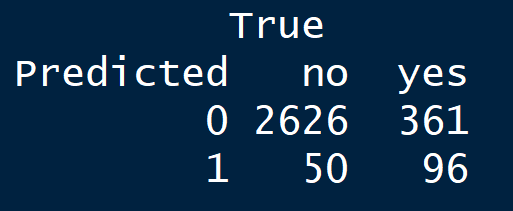
****

**90% Threshold:**

****

**Recommended**

**55% Threshold:**

****