Building a Predictive Model to Identify Customer Churn for ABC Wireless Inc.

Final Assignment

Business Analytics

Dr. Rouzbeh Razavi

Kent State University

Date 16-Dec-2020

**Table of Contents**

|  |  |
| --- | --- |
|  |  |
| Executive Summary……………………………………………………………………………………………………. | 3 |
| Project Goal ……………..……………………………………………………………………………………………….. | 4 |
| Process Flowchart …………………………………………………………………………………………………….. | 5 |
| Overview ……………………………………………………………………………………………………................. | 6 |
| Modeling Strategy……………………………………………………………………………………………………… | 9 |
| Model Performance ……………………….…………………………………………………………………………. | 12 |
| Insights ……………………………………………………………………………………………………………………… | 15 |
| Conclusion ………………………………………………………………………………………………………………… | 16 |

**Executive Summary**

ABC Wireless Inc. has hired our team to prevent customer churn. We decided to create a model that could predict the customers who would churn soon. Using the results from our model, ABC Wireless inc. would be utilizing appropriate marketing efforts to ensure that those customers stay with the company. To create our model, we used historical data of 1000 customers that included categorical and continuous data.

The data was thoroughly analyzed by our team members, who were able to impute missing data through a machine learning algorithm and notice both positive and negative correlation between different variables. Our team created multiple models to accurately predict customer who would churn soon. We started with 15% accuracy in prediction. With our final model, we were able to predict with 71% accuracy. Out of the 1000 customers, it was predicted that 93 customers would churn.

Our model allowed us to identify the key attributes which influence the customers to churn. Those findings are included in the report.

**Project Goal**

Telecom companies spend a lot of resources to acquire new customers. The cost of retaining a customer is less than the cost of acquiring a new customer. ABC Wireless has hired our team to identify customers, who are likely to churn soon, and assist the company with their targeted approach to keep those customers. We shall be creating a model for prediction of customer churn. The model would accurately identify maximum true positives and minimum false positives. Misclassifications (false positives) would result in spending resources on customers who would continue to stay with the company. Our goal is to deliver a model that would allow ABC Wireless to retain existing customers and maximize their revenue.

**Process Flowchart**

This High-level Process flowchart depicts sequential steps taken to predict the customers of ABC Wireless who are likely to churn:

**Overview of Data**

**Initial Analysis:**

Summary of each variable of the dataset was viewed to get a feel for the data’s overall quality.

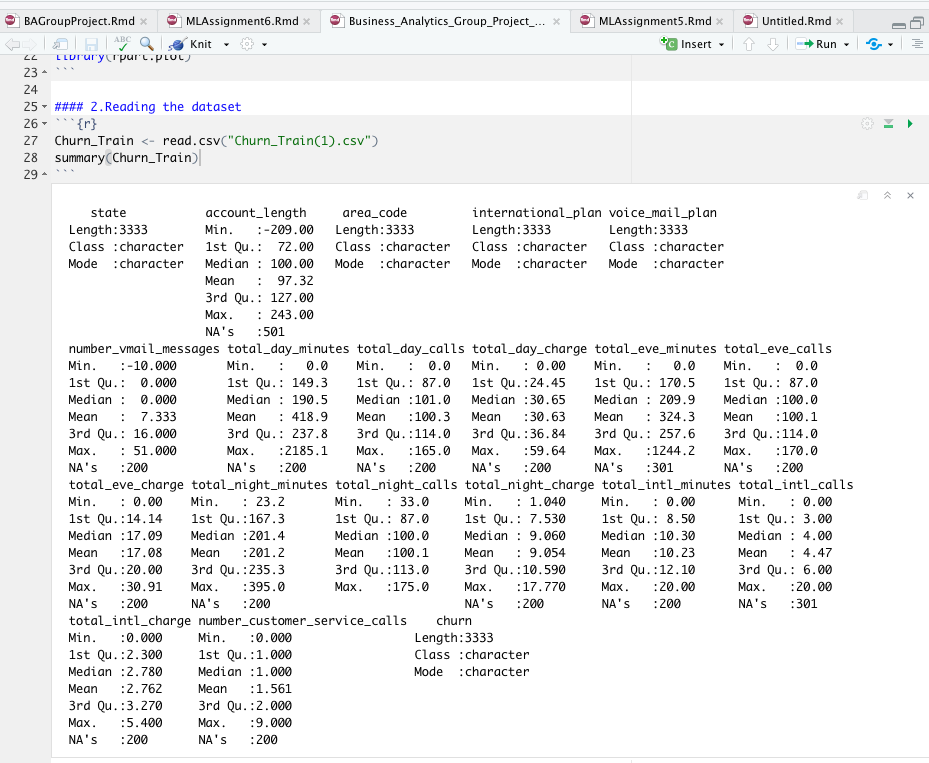


Figure 1 Summary of data

From this summary, we evaluated that 2 variables ‘account\_length’ and ‘number\_vmail\_messages’ contained negative values and 14 other variables had at least 200 rows of missing data (NA’s). To look further into the missing data, we ran ‘md.patterns’ using the VIM package in R. It generated the graphic representation to show the layout of the ‘NA’ data. According to this graphic, all 14 variables shared 200 common rows of missing data and 3 of those 14 variables have an additional 101 rows each of missing data.

Chart, bar chart, histogram

Description automatically generated

Figure 2 Missing data

**Missing Value Imputation:**

MICE package in R was used to impute the missing values in the data set. MICE does this by determining the best method suitable for continuous incomplete variables. In this case, PMM or Predictive Mean Matching. PMM finds a set of observed values with the closest predicted mean as the missing one and imputes the missing values by a random draw from that data set. MICE uses all available variables in its’ imputation model including categorical values.

**Factor Correlation Analysis and Interpretation:**

After imputing the missing values, a correlation plot was created to view any correlation between the variables in the dataset.

A picture containing text, businesscard

Description automatically generated

Figure 3 Correlation Plot

**Modeling Strategy**

**Selection of Prediction Algorithm:**

Once the data was ready for analysis, we evaluated the following algorithms for performance and ranking suitability:

* **Decision Tree**: is a flowchart-like structure where each internal node represents a test on a feature.
* **Logistic Regression**: is a supervised learning classification algorithm used to predict the **probability** of the target variable.

Since prediction of customer churn is a classification problem, we have considered Logistic Regression and Decision Tree to create our model. These models are very popular algorithms and have strong predictive performance and good comprehensibility. The model with better performance would be chosen for prediction.

Here are the steps taken by the team to measure each model’s performance:

**Evaluation of Model’s Performance:**

Once the confusion matrix is generated, we evaluate the accuracy, precision, and recall for both models:

* **Accuracy** is the proportion of identifications that are actually correct.
* **Precision** is proportion of positive identifications that are actually positive.
* **Recall** is the proportion of actual positives that were identified correctly.

Our effort was to reduce false positive and false negatives. False positives lead to unnecessary use of company’s resources on customers who would continue to stay with the company. False negatives lead to loss of customers who could have been retained through active marketing.

**Model Analysis & Performance**

**Part 1: Predictive Models**

**Model 1**

The first model that we used on our dataset was logistic regression. The reason we choose to begin with this model was due to the dataset having both continuous and categorical variables. Logistic regression handles both of those variable types very well and will classify the customers who are likely to churn (see **Figure 1**).

**Model 2**

We wanted to improve the predictive power of our first model. We used a search program to discover any variable combinations that could be used to increase the model’s customer churn prediction accuracy (see **Figure 1**).

**Decision Tree Predictive Model**

The second logistic model improved our prediction accuracy by about 30% which was a significant increase. However, we still wanted to improve the accuracy as 42% is not much better than a person randomly predicting if a customer would churn. Decision Trees also handle both categorical and continuous variables. These models are also easy to understand and interpret. The graphic below shows the customer churn prediction accuracy (see **Figure 1**).

**Figure 1**

Customer Churn

Prediction Accuracy

**15%**

**Logistic Model 1**

**42%**

**Logistic Model 2**

**71%**

**Decision Tree Model**

**Part 2: Model Prediction on Test Dataset**

The decision tree model outperformed our other two logistic models by about 60% and 30%, respectively. Based upon this, we selected it as the best model for predicting customers who were more likely to churn. Using the decision tree model on the Customer\_To\_Predict dataset, it predicted that **93 of the 1000 customers** would likely churn (see **Figure 2** below).

**Insights**

Customer churn (also known as customer attrition) refers to when a customer ceases his or her relationship with a company. It is a common problem for telecom companies that costs them hundreds of dollars in brand promotion and advertisement expenses and results in loss of future revenue as well. One of the important tasks is to identify customers who are likely to churn and provide suitable intervention to encourage them to stay.

The Decision Tree model allowed out team to identify the key attributes which influence the customers to churn. Key attributes causing customer churns are:

* Having International plans
* Having voice mail plans
* Amount of daytime call charges.
* Number of customer service calls
* Total international calls and charges

Insights on this data should help the Company management to take appropriate action forward to prevent customers moving out to the peer companies.

Below are some recommendations to the company management to prevent customer churn: ​

* **Enhancing the customer service**: Customer service calls are indicators of the dissatisfaction. Follow up calls and proper resolution is the key to the customer satisfaction. ​
* **International calls:** International calling plays a vital role in losing customers. Better rate for international calls and promotional offers would help in preventing churn of customers who use the international calling service
* **Call Volumes**: Discounts can be offered to customers with high calling volumes during day and evening after a certain threshold to encourage them to stay and keep using the service.

**Conclusion**

Our team has done a detailed analysis for ABC Wireless Inc. using historical data to create multiple predictive models. We were able to impute missing (NA) values using PMM or Predictive Mean Matching. We analyzed the results to create the best possible model to predict churn.

Two models were created:

* Logistic Regression
* Decision Tree

The first logistic regression model was able to predict with 15% accuracy. To optimize the model, we used a search program to discover any variable combinations that could be used to increase the model’s prediction accuracy. We were able to take the accuracy to 42%. We believed that 42% accuracy was not good enough. To increase the accuracy further, we created another model based on the decision tree algorithm. The Decision Tree model performed better than the logistic regression model. The Decision Tree model predicted 71% of the data correctly on the validation data set.