Final Project Report

MIS-64036 - Business Analytics

**ABC Wireless Inc. - Customer Churn Prediction Project**

Group 3 – Rohit Koyyala, Riba Khan, Manaswini Purumandla and Tejasvini Mavuleti

# **Abstract**

A company's bottom line can suffer significantly if it fails enough clients. Another reason why improving client retention and reducing churn is crucial is because finding new customers is often more expensive than keeping existing ones. Customer churn is a huge worry for large firms and one of the most difficult challenges. Because lost customers mean lost money, customer churn is an important measure to track. Companies that lose consumers face the loss of revenue from those clients and the high expense of acquiring new customers. Because it directly influences profits, companies are striving to develop various approaches to predict potential client churn, notably in the telecom industry. As a result, determining the causes that cause customer churn is crucial to making the necessary efforts to reduce churn. The project's main contribution is developing a churn prediction model that assists ABC Wireless Inc. in identifying customers who are more likely to churn and maintaining those customers to lower the churn rate. Offering them special promotions or incentives is part of this retention strategy.

**Table of Contents**

[**Abstract** 1](#_Toc103198927)

[**Introduction** 3](#_Toc103198928)

[**Data cleaning** 3](#_Toc103198929)

[**Outputs and observations** 4](#_Toc103198930)

[**Removing missing variables** 9](#_Toc103198931)

[**Correlation Plots** 9](#_Toc103198932)

[**Model Selection** 9](#_Toc103198933)

[**Logistic Regression** 10](#_Toc103198934)

[**Confusion Matrix of Logistic Regression Model** 12](#_Toc103198935)

[**Building Decision Tree model** 12](#_Toc103198936)

[**Confusion Matrix of Decision Tree** 13](#_Toc103198937)

[**Model Performance** 13](#_Toc103198938)

[**Building the Final Model to predict the churn using Test data and Decision Tree Algorithm** 13](#_Toc103198939)

[**Prediction of the Test data** 13](#_Toc103198940)

[**Plot for summary of the Test data** 14](#_Toc103198941)

[**Conclusion** 14](#_Toc103198942)

[**Recommendations** 16](#_Toc103198943)

[**References** 17](#_Toc103198944)

## **Introduction**

In developed countries, telecommunications have become one of the most critical industries. The degree of competition has increased due to technological developments and an increase in the number of operators. Companies rely on various techniques to succeed in this competitive industry. A couple of solutions have been proposed to enhance revenue attract new customers, upsell existing customers, and extend customer retention periods. When these tactics are compared, the third strategy is the most profitable, demonstrating that retaining an existing customer cost far less than obtaining a new one and is much easier than the upselling plan. Companies must reduce their costs to implement the third strategy. Machine learning technology is extremely effective at predicting this circumstance. The primary goal of this project is to examine the data and use predictive modeling to determine whether an existing client of ABC Wireless will switch service providers. Customers’ churn is a considerable concern in service sectors with highly competitive services. On the other hand, predicting the customers who are likely to leave the company will represent a potentially large additional revenue source if it is done in the early phase.

To better comprehend the nature of the data, data analysts utilize data visualization and statistical tools to convey dataset characterizations, such as size, amount, and accuracy. The churn data shows the different types of services used by the customers of the ABC Wireless Telecom company. The following were the attributes – state, account\_length, area\_code, international\_plan , voice\_mail\_plan, number\_vmail\_messages , total\_day\_minutes, total\_day\_calls, total\_day\_charge, total\_eve\_minutes, total\_eve\_calls, total\_eve\_charge, total\_night\_minutes, total\_night\_calls, total\_night\_charge, total\_intl\_minutes, total\_intl\_calls, total\_intl\_charge, number\_customer\_service\_calls, and the churn result.

## **Data cleaning**

You can clean data by detecting flaws or corruptions, repairing or eliminating them, or manually processing data to avoid repeating the same mistakes. Although software tools can help with most parts of data cleansing, some tasks must be completed manually. While cleaning the data the descriptive statistics helped us decide on results. Let’s look at our statistics.

A screenshot of a computer

Description automatically generated with medium confidence

## **Outputs and observations**

Graphs and charts condense large amounts of information into easy-to-understand formats that clearly and effectively communicate essential points. For example, calculating churn and implementing best practices to minimize it are crucial to improving the health of your business. This metric will also help you analyze past business performance and accurately forecast future revenue. So far, we’ve discussed churn in terms of how many customers or how much income you are losing. And we compared how much you are gaining. However, that is an oversimplification. You can use more in-depth churn metrics to get a complete picture of your company’s health. Here is what we got when we applied data visualization.

Diagram, shape, arrow

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

We see that there are 2850 customers who have not churned, and 483 have canceled their services and switched to another provider. Compared to other states, the turnover rate in Texas, New Jersey, Maryland, and Michigan is about average. According to the above boxplot distribution of churn data by total day charge, customers with a 35 or 40 are more likely to cancel their subscriptions. According to the above box plot distribution, customers who call customer care more than twice are more likely to churn.

## **Removing missing variables**

Some variables in the dataset have missing values or n/a. The variations in accuracy and prediction results based on the method we used in handling the missing values. First we delete the rows that have n/a. Then, impute the n/a using the median of the columns and plug the n/a attributes.

## **Correlation Plots**

When churn equals Yes, we’ll look at the correlation of variables, we discovered a high negative association between the number of customer support calls and total day, total evening, and total international costs among those who churned from the plots above.

We can deduce several potentially interesting truths from the correlation. When churn = yes, the higher the costs, the more calls to customer support were made.

Chart, scatter chart

Description automatically generated

## **Model Selection**

To find the most accurate model for predicting which customers would churn and which will not.

A predictive model based on regression and Decision Tree Model was used to highlight the effect of numerous factors and their relevance in forecasting the result of the target variable.

Regression can be done in two ways:

1. Linear Regression
2. Logistic Regression

Because the data’s target variable is categorical, a logistic regression model is the best option. When predicting a binomial property, it’s tempting to use linear regression as a model, however the performance likelihood might be negative or more than 1, making it ineffective. A likelihood or chance of odds between 0 and 1, as determined by logistic regression, is the optimum outcome for this model. We also picked Logistic Regression and Decision Models as appropriate after examining the dataset because categorization is our primary goal. In that vein, we’ll test both models on our dataset and choose the best one to predict the test dataset as the final model.

Determining the predictive ability of Logistic regression and Decision trees models :

To prevent overfitting the model, divide the dataset into two sections: training and validation. Building a logistic regression model, predicting the outcomes on the validation data set, and validating the model’s performance with a confusion matrix Building a Decision Tree Model and Predicting the Results on the Validation Data Set and Validating the Model’s Performance with the Confusion Matrix Compare the Model Performance and Chose the best model.

## **Logistic Regression**

Logistic Regression - Logistic regression is a type of regression that uses a combination of continuous and discrete factors to predict discrete or categorical variables. To put it another way, the Y or goal variables must always be categorical variables, but the X variables might be categorical or continuous variables.

Graphical user interface, text, application, email

Description automatically generated  
  
Text

Description automatically generated

Chart, line chart

Description automatically generated

**Confusion Matrix of Logistic Regression Model**

The Accuracy of the logistic Regression Model is 86.49 %

Sensitivity : 97.54%

Specificity : 20.83%

**Building Decision Tree model**

#Decision Tree Model:

The main goal of the decision tree model is to Classify or predict an outcome based on a set of predictors.  
Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated with low confidence

**Confusion Matrix of Decision Tree**

The Accuracy of the Decision Tree Model is 92.04 %

Sensitivity : 97.02%

Specificity : 62.50%

**Model Performance**

Because of its excellent accuracy, we picked the Decision Tree Model to forecast the churn of the test data after testing the Model’s performance.

**Building the Final Model to predict the churn using Test data and Decision Tree Algorithm**

Diagram

Description automatically generated

Area under the curve: 0.8931

**Prediction of the Test data**

## predict\_churn  
## no :1445   
## yes: 155

**Plot for summary of the Test data**

Chart, bar chart

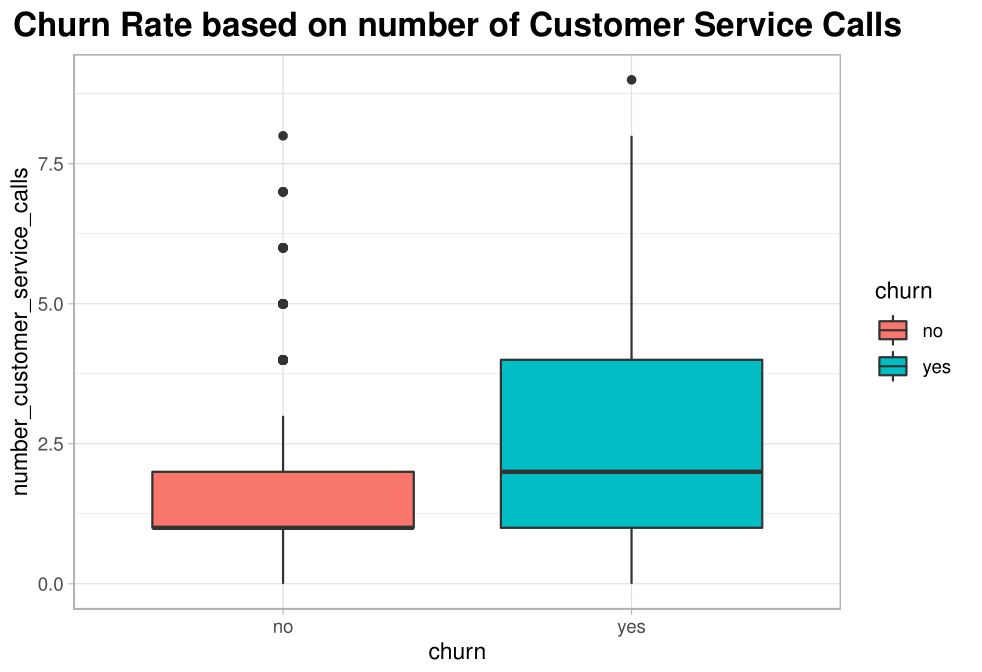
Description automatically generated

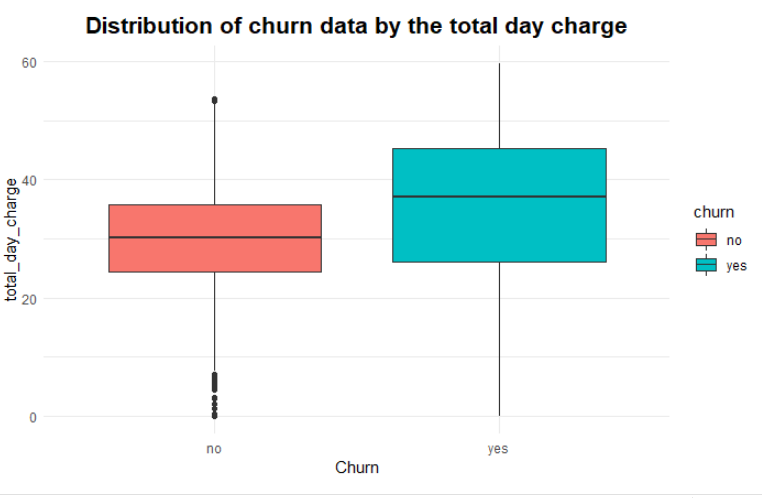
predict\_churn

no :1444 yes: 156

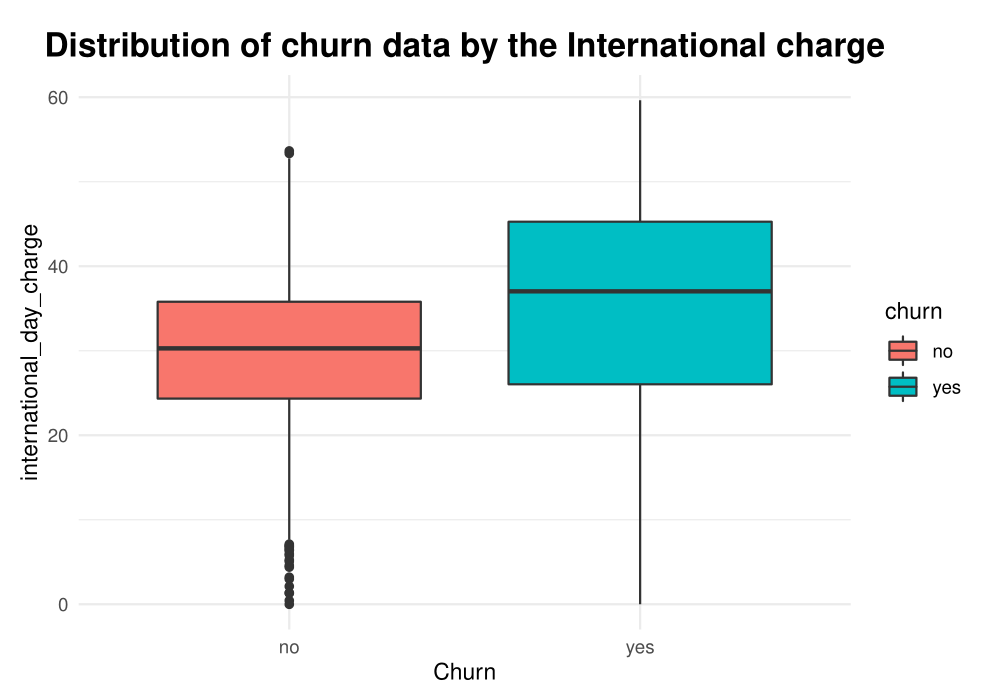
# **Conclusion**

Effect on the churn

 The customers who called Customer Service more than 2 to 4 times are high likely to cancel the service, as those customers who switched the company were not satisfied.



The customers who paid over 35-day charges are likely to move away.



The customers who paid more than 30 international day charges are likely to cancel the service.

From the data exploration, consumers who phoned Customer Account more than 2 to 4 times are more likely to cancel the service, implying that customers who moved companies were dissatisfied with the service. customers who paid higher above 35-day rates are more likely to discontinue the service, we may deduce. International day costs, like day charges, have an impact on the turnover rate. Customers that spent above 30 international day costs are more likely to terminate the service, according to the above box plot distribution.

### **Recommendations**

It would be preferable to lower the daytime and international daytime charges or keep these two groups competitive. Improve and deliver outstanding customer service. With a high-grade network, you can maintain or improve overall quality. It is good to maintain customer retention by rewarding loyal consumers. ABC Wireless Inc. should get customers' input and should be sought regularly.

## **References**

J. (2021, July 22). Churn Rate: Definition, Advantage and Disadvantage, How to Reduce It. Sales Glossary - LeadMine. <https://www.leadmine.net/glossary/churn-rate/>

The Value of Model Accuracy | DataRobot Blog. (2022, March 24). DataRobot AI Cloud. [https://www.datarobot.com/blog/the-value-of-model-accuracy/#:%7E:text=Model%20accuracy%20is%20defined%20as,certainly%20not%20the%20only%20way](https://www.datarobot.com/blog/the-value-of-model-accuracy/%23:%7E:text=Model%20accuracy%20is%20defined%20as,certainly%20not%20the%20only%20way)

What is Decision Tree Analysis? Definition and FAQs | OmniSci. (n.d.). Www.heavy.ai. <https://www.heavy.ai/technical-glossary/decision-tree-analysis‌>

Wikimedia Foundation. (2022, March 29). Model selection. Wikipedia. [https://en.wikipedia.org/wiki/Model\_selection](https://en.wikipedia.org/wiki/Model_selection%20)