

Telecom CUSTOMERs CHURN ANALYSIS

BUSINESS ANALYTICS FINAL PROJECT REPORT



April 30, 2023

Group – 8

Peddireddigari Jyothsna - Nandini Raveendran Nair Subhadra -Srujana Kasturi -

Krishna Kumar Tavva

Contents

[Project Abstract 2](#_Toc133785972)

[Team Contributions 2](#_Toc133785973)

[Introduction 2](#_Toc133785974)

[Project Goal 2](#_Toc133785975)

[Assumptions 2](#_Toc133785976)

[Overview of Data 2](#_Toc133785977)

[Data 2](#_Toc133785978)

[Data Preparation 3](#_Toc133785979)

[Exploratory Data Analysis 4](#_Toc133785980)

[Model Building Strategy 6](#_Toc133785981)

[Data Partition 6](#_Toc133785982)

[User Defined Functions 6](#_Toc133785983)

[Optimal Threshold Selection 6](#_Toc133785984)

[Metric Calculations 6](#_Toc133785985)

[Model Building 6](#_Toc133785986)

[Logistic Regression 6](#_Toc133785987)

[K-nearest neighbors 8](#_Toc133785988)

[Decision Tree 9](#_Toc133785989)

[Model Evaluation 11](#_Toc133785990)

[Results 12](#_Toc133785991)

[Insights and Conclusion 12](#_Toc133785992)

[Insights 12](#_Toc133785993)

[Conclusion 12](#_Toc133785994)

[Limitations 12](#_Toc133785995)

[Appendix 13](#_Toc133785996)

[User Defined Functions 13](#_Toc133785997)

# Project Abstract

Customer turnover is a major concern and one of the most difficult challenges for multinational organizations. Companies are exploring a variety of ways to forecasting potential client churn because it has such a direct impact on their revenues, particularly in the telecoms industry. As a result, understanding the factors that contribute to client attrition is critical for reducing churn. The fundamental contribution of this project is developing a churn prediction model that supports ABC Wireless Inc. in identifying customers who are more likely to churn to lower churn rates. This retention strategy includes offering them special promotions or incentives.

## Team Contributions

### Contributors

Peddireddigari Jyothsna, Nandini Raveendran Nair Subhadra, Srujana Kasturi, and Krishna Kumar Tavva.

We have divided all the work between everyone in the team equally. Everyone contributed ideas to implement in R and snippets required for the initial data exploration and manipulation. Then Jyothsna worked on Model development and Model selection. The team reviewed the resulted and finalized the best-performing model of all the models built in the project. Krishna and Nandini worked on the presentation and on the audio for the presentation. Followed by Srujana created a report document to complete all the required artifacts of the project. After completion of all the required files for the submissions, each of us exchanged the work completed by the other to review and make the necessary changes.

# Introduction

Customer churn is a common problem for businesses across various domains. It refers to customers who stop buying products or services from a company. The reasons could be varied, such as poor customer service, high costs, a better alternative, etc. Identifying and predicting customer churn is essential for businesses as it helps them retain their customers and increase their revenue.

# Project Goal

The Project goal is to identify the probable customers that would discontinue subscriptions using the best-performed predictive model of all the models built during the project. ABC Wireless Inc. would be able to target the right customers that are exploring the new service plans offered by other services and implement marketing strategies such as targeted offers or discounts based on their usages. The data used in this project is from ABC Wirelss Inc., which offers phone and internet services. The dataset contains customer information with churning information to identify the active customers and inactive customers.

The project was carried out using the R programming language and Rstudio(IDE).

## Assumptions

The project was completed with the following assumptions.

* Customer acquisition cost is significantly higher than retaining customers with possible targeted offers.
* Dataset represents an overall population of ABC telecom customers.
* As the dataset used for the analysis is entirely related to the ABC telecom customers. The project ignored the competitor’s pricing and plans for similar services.

# Overview of Data

## Data

This information pertains to the services used by ABC Wireless Telecom users. This dataset includes 20 attributes, 3333 Customers. Following is the snapshot of the list of variables in the dataset and their data type. 15 numeric and 5 categorical attributes are there in the dataset.

Table

Description automatically generated with medium confidence

Data Preparation

State, actual\_length and area\_code variables are removed from the dataset. Actual\_length variable is removed from the dataset after verifying that there is no significance impact in predicting the customer churn.

A picture containing calendar

Description automatically generated

Following snapshot of r code is used to converting international plan, voice mail plan and churn variables to binary for making the variables suitable for building predictive models for the project.

Graphical user interface, text, application

Description automatically generated

Following snapshot of r code determined that the NA values are present in some of the attributes.

Graphical user interface, text, application

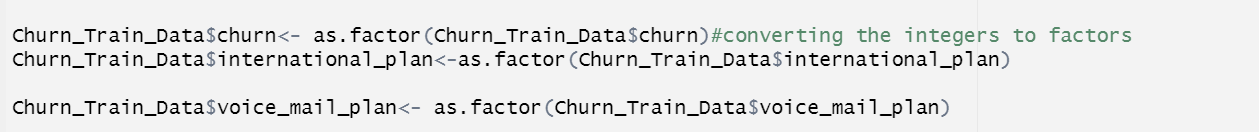
Description automatically generated

For all the numeric variables in the modified dataset, NA values are imputed with median value of the variable. Median is selected for the imputation in place of mean as the median would be helpful even if the data distribution is skewed. Following is the snapshot of the imputation.

Text

Description automatically generated

As treating the churn column as integer would cause problems, churn column is converted into factor for building the model. Additionally, international\_plan and voice\_mail\_plan variables are also converted to factors for using the attributes as a predictor for the model inputs.



## Exploratory Data Analysis

Summary of all the columns present in the modified data set after data imputation, conversions and attribute exclusions.



Text

Description automatically generated

Snapshot of r code for the histograms on all numerical variables and scatter plots between all the numeric variables in the dataset.

Graphical user interface, text, application, email

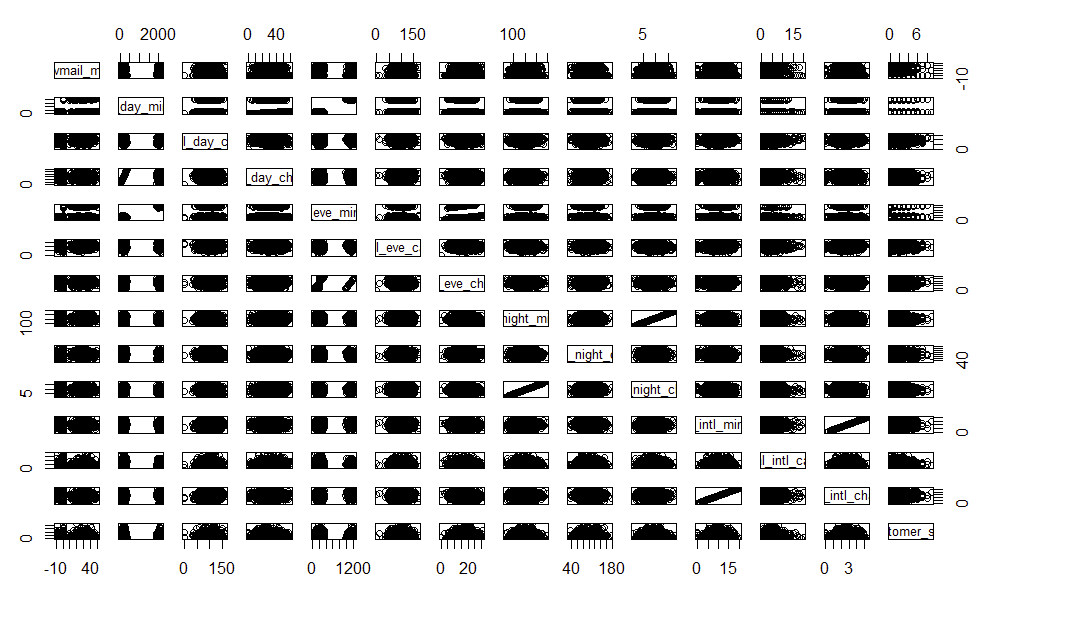
Description automatically generated

The distribution of each variable was explored using histograms.

Chart

Description automatically generated

The relationship between pairs of variables was explored using scatter plots.



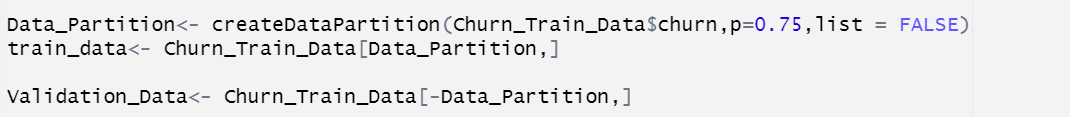
From the scatter plot between all the numeric variables, total night minutes and total night charges are correlated. Additionally, international minute and internal charges are highly correlated.

# Model Building Strategy

After data manipulation and data exploration, data was divided into train and validation. Train data set is used for building the predictive model and validation data set to validate the accuracy of the selected model.

## Data Partition

The data was partitioned into a training set of 75% and a validation set of 25% using the following r code.



## User Defined Functions

### Optimal Threshold Selection

FindThreshold returns the sensitivity, specificity and accuracy of the predictions by changing the threshold increments from 0.02 to 1. This is helpful in identifying the best cut off values for the probabilities by considering all three metrics sensitivity, specificity and accuracy. This is helpful as team focused on the improvement of sensitivity which would be the key for the project.

### Metric Calculations

MetricCalculation is to calculate all the metrics True Negatives, True Positives, False Positives, False Negatives, Sensitivity, Specificity and Accuracy for the table object in the r with predicted and actual values count.

## Model Building

Logistic regression, K-nearest neighbors (Knn) and decision tree models were built in the project.

### Logistic Regression

Following the r code for building the logistic regression and the output of the model.

Text

Description automatically generated

Text

Description automatically generated

Following is the snapshot of r code to predict the churn of the validation data set using the logistic regression model built.

Graphical user interface, text, application, email

Description automatically generated

Next step is to find the best threshold value to categorize for customers churning Yes or No.

Graphical user interface, table

Description automatically generated

From the above results, 0.14 threshold value selected after reviewing the metrics sensitivity, specificity and accuracy. Even though 0.16 threshold has higher specificity, sensitivity is important metric as the customer acquisition time is costly for the company than retaining the customers.

Following r code is used to calculate the AUC value of the model after applying the threshold.

Graphical user interface, text, application

Description automatically generated

Confusion matrix for the actual and validated data is viewed using the following r code.

Graphical user interface, text, application

Description automatically generated

Snapshot of all evaluation metrics of the logistic regression model when applied on the validation dataset.

Graphical user interface, text, application

Description automatically generated

### K-nearest neighbors

Following the r code for building the K-nearest neighbors (Knn) model with k value as 60 observations and the output of the model.

Graphical user interface, text, application, email

Description automatically generated

Next step is to find the best threshold value to categorize for customers churning Yes or No.

Table

Description automatically generated

From the above results, 0.14 threshold value selected after reviewing the metrics sensitivity, specificity and accuracy.

Graphical user interface, text, application

Description automatically generated

Confusion matrix for the actual and validated data is viewed using the following r code.

Graphical user interface, application

Description automatically generated

Snapshot of all evaluation metrics of the K-nearest neighbors model when applied on the validation dataset.

Graphical user interface, text, application

Description automatically generated

### Decision Tree

Following the r code for building the decision tree model with best complexity factor(cp) and the output of the model.

Graphical user interface, text, application, email

Description automatically generated

A picture containing text

Description automatically generated

Next step is to find the best threshold value to categorize for customers churning Yes or No.

Graphical user interface, text, application

Description automatically generated

From the above results, 0.14 threshold value selected after reviewing the metrics sensitivity, specificity and accuracy.

Graphical user interface, text, application, email

Description automatically generated

Confusion matrix for the actual and validated data is viewed using the following r code.

Graphical user interface, text, application

Description automatically generated

AUC value for the decision tree is 0.875 which is better than logistic regression.

Graphical user interface, text, application, email

Description automatically generated

Following decision tree plot could help in identifying the significant variables. total\_day\_charge and number\_customer\_service\_calls are important variables in the decision tree model.

Diagram

Description automatically generated

Snapshot of all evaluation metrics of the K-nearest neighbors model when applied on the validation dataset.

Graphical user interface, text, application

Description automatically generated

# Model Evaluation

The models were evaluated using the validation data. Sensitivity, specificity, and accuracy were used to evaluate the performance of the models. Following is the snapshot of the r code to compare metrics of all the models built in the project.

Graphical user interface

Description automatically generated with medium confidence

After comparing the sensitivity, specificity, and accuracy metrics of the three models built in the project, the Decision tree and Logistic regression have the best sensitivity. Even though decision tree and logistic regression have similar sensitivity, the specificity and accuracy of logistic regression are lower, which could cost the telecom company significantly higher costs as the target customers for reducing churn will increase if specificity is low. So, the **Decision tree** has the best sensitivity, specificity, and accuracy, resulting in lower marketing costs to retain ABC telecom customers that are more likely to churn.

## Results

Decision Tree is the best model of three models built as part of the project with 89% accuracy, 93% specificity and 71.6% sensitivity.

# Insights and Conclusion

## Insights

1. The predictive model was built over the “training data with 2501 observations out of which 2018 observations had the churn value as “no” and 483 observations had the churn value as “yes”. This can be a pure example of an “**Imbalanced Dataset”.**
2. It can be understood that model works well for most of the data that has a Churn value of “no”. This can be called out by the TNR/Specificity = 92.84%, since the trained model didn’t get to capture the trends much for the Churn value of “yes” the Sensitivity is lagging a bit short with 71.67%.
3. Imbalanced data can be a possible reason for less TPR and pruning the model isn't that effective as the dataset is highly imbalanced.
4. Total Day Charge and Number of Customer Service Calls are significant based on the decision tree.

## Conclusion

Overall, the two most important variables responsible for predicting the customers being churned at ABC Wireless Inc. are “Total Day Charge”, and “Number of Customer Service Calls”.

The reason for this could be that customers who are concerned with the price may find it expensive, meaning that the telecom company may be charging high rates to those customers who frequently use their services. This group of customers might find the services they use as expensive and might be shifted to another service provider where the charges are less.

Limitations

The analysis has certain limitations, such as the limited scope of the data used. The dataset used in this project is specific to a telecommunications company and may not be generalizable to other domains. Furthermore, other variables that may influence customer churn, such as customer satisfaction and loyalty, were not included in the dataset.

# Appendix

Following snapshot from the .Rmd file details the list of the libraries used to complete the customer churn analysis project.

Table

Description automatically generated

Below snapshot code is used for loading the churn\_train.csv file(to build the model) and customers\_to\_predict.Data file(to predict the customer churn “Yes” or “No” based on the final model selected after evaluating each of the model metrics).

Graphical user interface, text

Description automatically generated with medium confidence

## User Defined Functions

As part of the project to reduce the repetitive r code, two functions are developed. One of the functions is FindThreshold which returns the sensitivity, specificity and accuracy of the predictions by changing the threshold increments from 0.02 to 1. This is helpful in identifying the best cut off values for the probabilities by considering all three metrics sensitivity, specificity and accuracy. This is helpful as team focused on the improvement of sensitivity which would be the key for the project.

Following is the code for the FindThreshold.

Graphical user interface, text, application, email

Description automatically generated

Another function MetricCalculation is to calculate all the metrics True Negatives, True Positives, False Positives, False Negatives, Sensitivity, Specificity and Accuracy for the table object in the r with predicted and actual values count.

Text

Description automatically generated