Data Mining and Data Analytics for Analysing Customer Churn Rate

Sharan Kumar Paratala Rajagopal

Abstract — This research paper describes how to determine the various factors impacting the customers churn rate in telecom industry. And what factors impact customer to move from one telecom source to another. Using data analytics and data mining will analyze the factors for churn rate.

Index Terms — Data analytics; Data mining; R; Churn rate.

I. INTRODUCTION

One of the most critical factors in customer relationship management that directly impacts a company’s long-term profitability is customer attrition. When a company can better predict if a customer is likely to cut ties, it can take a more targeted approach to mitigate customer turnover.

Telecommunications Company is concerned about the number of customers leaving their landline business for cable competitors. The company needs to know which customers are leaving and attempt to mitigate continued customer loss. This paper describes the methods to analyze customer data to identify why customers are leaving and potential indicators to explain why those customers are leaving so the company can make an informed plan to mitigate further loss.

II. ANALYSIS APPROACH

A. Tool Selection

Firstly, we need to choose the best appropriate tools to perform the analysis on the data set. For this paper we are choose R programming to analyze the churn rate of the customer due to the fact that R provides various inbuilt packages and also features making statistical analysis of large data sets simple. R has the integrated development environment available in R studio and is accessible from number of scripting language and hence makes it easier to view and analyze the code. With R almost every tool a data scientist might need to manipulate and evaluate structured data is included. What isn’t in the base package has often been built and shared by other programmers and is freely available to download. Hence, based on the advantages of R programming it’s been chosen for the analysis of the customer data.

B. Descriptive Method

Multiple correspondence analysis will be used to analyze the variables which contribute most to the customer churn rate. To check if there is any collinearity and eliminate such data from analysis. This data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. Binning to create categorical variables if there any continuous.

C. Prediction Method

Logistic regression will be used as the target and dependent variables are categorical values and logistic regression holds good for the analysis.

III. DATA EXPLORATION AND PREPARATION

Before beginning the analysis, the target variable has to be determined in the data and indicate if the specific type of data the target variable is using. The target variable in this research is the “Churn” variable. The data which is used by the target variable will be either “Yes” or “No” and it is categorical.

“Fig.1” shows the rest of the variables used in the analysis process.

![Fig. 1. Data set for analyzing customer churn factors.](image)

“Fig. 2” provides the details of independent.

![Fig. 2. Independent variable for analyzing customer churn factors.](image)
A. Data preparation

- Data preparation is a very important step in any data analysis. As whole data set might not contribute entirely due to various factors such as outliers etc.
- In this analysis the data preparation is done by reducing the number of independent variables which are continuous and might not contribute much.
- Reduce number of categorical variable values. Make them into factors which will help in analysis.
- MonthlyCharges and TotalCharges are of similar values and both of them might not be required and one of the column can be dropped.
- Reduce the binary value to “No” instead of “No Internet Service” wherever possible.
- Bin the “Tenure” column for the analysis instead of continuous qualitative values.

B. Statistical Identity

“Fig.3” provides the details of the variables used pre and post cleaning of data preparation and divided into dependent and independent variables.

| Name                  | Type (pre-cleaning) | Type (post-cleaning) | Dependent/Independent | Notes                                                                 |
|-----------------------|---------------------|----------------------|-----------------------|----------------------------------------------------------------------|
| customer ID           | Random(Categorical) | Dropped              | Dropped               | Dropped due to lack of preditive characteristics.                   |
| Gender                | Binary              | Dropped              | Dropped               | Dropped from analysis due to lack of variance.                     |
| SeniorCitizen         | Binary              | Binary               | Independent           |                                                                      |
| Partner               | Binary              | Binary               | Independent           |                                                                      |
| Dependents            | Binary              | Binary               | Independent           |                                                                      |
| PhoneService          | Binary              | Binary               | Independent           |                                                                      |
| MultipleLines         | Categorical         | Binary               | Independent           | Reduced to a binary value by the recoding of the “No phone service” to “No” |
| InternetService       | Categorical         | Categorical          | Independent           | Three levels                                                         |
| OnlineSecurity        | Categorical         | Binary               | Independent           |                                                                      |
| OnlineBackup          | Categorical         | Binary               | Independent           |                                                                      |
| DeviceProtection      | Categorical         | Binary               | Independent           | Reduced to a binary value by the recoding of the “No internet service” to “No” |
| TechSupport           | Categorical         | Binary               | Independent           |                                                                      |
| StreamingTV           | Categorical         | Binary               | Independent           |                                                                      |
| StreamingMovies       | Categorical         | Binary               | Independent           |                                                                      |
| Contract              | Categorical         | Categorical          | Independent           | Three levels: Month-to-Month One-year Two-year                     |
| PaperlessBilling      | Binary              | Binary               | Independent           |                                                                      |
| PaymentMethod         | Categorical         | Categorical          | Independent           | Four levels                                                         |
| Churn                 | Binary              | Binary               | Dependent             |                                                                      |
| Tenure                | Continuous          | Categorical          | Independent           | Converted to Categorical variable by binning the values by year. Six levels. |
| MonthlyCharges        | Continuous          | Categorical          | Independent           | Converted to Categorical variable by binning the values by $50 increments |
| TotalCharges          | Continuous          | Dropped              | Dropped               | Dropped to high correlation to tenure.                              |

Fig. 3. Statistical identity and variable preparation

C. Raw data

The Raw data has been cleaned and the same is mentioned in “Fig.4”. Various possible values for each variable identified is described to give better understanding of the data we are looking at.
Import the raw data into the R and verify that variables are as expected with the values which was mapped to. Let’s run commands to verify if there are any missing data. “Fig.5” provides there is 11 missing data for the “Total Charges”.

“Fig.6” shows after removing the missing data.

The raw data set chosen has 21 columns and 7032 rows of data. Other cleanup activities performed on this data set includes:
- “No Internet service” is converted to “No”.
- “No Phone Service” is converted to “No”.
- Group the “Tenure” into bins based on the months. Minimum value is 1 and maximum value is 72. So, group into 5 tenure groups.
- Change the values in the column “SeniorCitizen” 0 – “No” and 1 – “Yes”.
- Removed the column “CustomerID”, “Gender” and “TotalCharges” as we don’t need for analysis. These columns don’t predict the churn rate and it’s of least importance in analysis. The churn rate is not dependent on the gender to predict if the customer is staying or leaving the telecom company.
- “Fig.7” shows that “Monthly Charges” and “Total Charges” are correlated. So, one of them will be removed from the model. “Total Charges” will be removed.

“Fig.8” shows bar plots for categorical variables.
Fig. 8. Bar plots for categorical variables.

“Fig. 9” provides the multiple correspondence analysis plot for different variables used in the analysis.

Fig. 9. MCA plot.
“Fig. 10” provides the information based on MCA analysis, the customer with month-month contract is near to the target variable “Churn” and it shows that customers more likely to churn when they are on month-to-month contract. Overall customer turn rate is 26.6%.

> round(prop.table(table(trainingChurn)) * 100, 1)

|     |   No | Yes |
|-----|-----|-----|
| 73.4 | 26.6 |

Fig. 10. Customer churn rate analysis

D. Logistic Regression analysis

Logistic regression is used for binary classification. Since, “Churn” variable is categorical and contains “Yes” or “No” values. Logistic regression will be appropriate. “Fig. 11” provides the logistic regression model values.

> LogModel <- glm(as.factor(churn) ~ ., family = binomial(link = "logit"), data = train)

> summary(LogModel)

Df    LogLikelihood     Deviance Residual Dev
  1     0.8475676   0.60608 -0.38282 0.66297 3.1169

> coefficients(LogModel)

(‘(‘ not defined because of singularities)

´

Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.818374 0.525843 -5.362 8.25e-08 ***
genreMale 0.003598 0.070014 0.046 0.96321
SenorCitizens 0.271267 0.100448 2.704 0.00685 **
PartnerYes -0.026553 0.061173 -0.438 0.66118
Dependents -0.127052 0.108834 -1.167 0.24305
PhoneService -1.211218 0.399897 -3.029 0.00244 **
MultipleLines -0.032431 0.197759 -0.165 0.86856
Internettierfiber optic -0.001940 0.413192 -0.009 0.99306
OnlineSecurityNo internet service -0.315662 0.18292 -4.090 0.00004 ***
OnlineBackupNo -0.373409 0.125959 -2.980 0.00328 ***
DeviceProtectionNo -0.299927 0.127908 -2.302 0.02085 *
TechSupportNo -0.529041 0.137737 -3.901 0.00009 ***
StreamingNo internet service -0.052472 0.19872 -0.264 0.79371
StreamingTVNo internet service -0.052472 0.19872 -0.264 0.79371
StreamingMoviesNo internet service -0.052472 0.19872 -0.264 0.79371
StreamingSuspensionNA NA NA NA NA
ContactType NA NA NA NA NA
Contract Year -0.890261 0.123207 -7.384 1.56e-13 ***
Contract Year NA NA NA NA NA
PaperlessBilling 0.279933 0.084662 3.319 0.00118 *
PaymentMethodCredit card (automatic) -0.236644 0.132124 -1.816 0.06956
PaymentMethodDebit card 0.167106 0.112423 1.489 0.13988
PaymentMethodOther 0.111653 0.134046 -0.838 0.40291
MonthlyCharges 0.038043 0.017030 2.234 0.02558 *
Tenure_12-month 1.167074 0.201981 5.820 2.16e-08 ***
tenure Group2-24 0.871311 0.200552 4.345 1.39e-05 ***
tenure Group24-48 0.420480 0.185588 2.278 0.02271 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5702.6 on 4923 degrees of freedom
Residual deviance: 5495.6 on 4899 degrees of freedom

AIC: 5480.6

Number of Fisher scoring iterations: 6

Fig. 11. Logistic regression model

The top three most relevant features include:
- Contact
- Tenure_group
- PaperlessBilling

“Fig. 12” provides the assessment of the logistic regression model with the accuracy value.

> testingChurn <- as.character(testingChurn)
> testingChurn[testingChurn == "No"] <- 0
> fitted.results <- predict(LogModel,newdata = testingChurn,type = "response")

In predict(linkobject, newdata, se.fit, scale = 1, type = ifelse(type == prediction from a rank-deficient fit may be misleading
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misClassfier.errror <- mean(fitted.results != testingChurn)
> print(paste("Logistic Regression Accuracy",1- misClassfier.errror))

[1] "Logistic Regression Accuracy 0.907096093468669"

Fig. 12. Logistic regression model accuracy

“Fig. 13” provides the fit for the linear regression.

> glm.fit <- glm(as.factor(churn) ~ ., family = binomial(link = "logit"), data = training)
> print(summary(glm.fit))

Deviance Residuals:

Min      1Q   Median      3Q     Max
-2.0455 -0.6060 -0.3828 0.6629 3.1169

Coefficients: (‘(‘ not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.818374 0.525843 5.362 8.25e-08 ***
genreMale 0.003598 0.070014 0.046 0.96321
SenorCitizens 0.271267 0.100448 2.704 0.00685 **
PartnerYes -0.026553 0.061173 -0.438 0.66118
Dependents -0.127052 0.108834 -1.167 0.24305
PhoneService -1.211218 0.399897 -3.029 0.00244 **
MultipleLines -0.032431 0.197759 -0.165 0.86856
Internettierfiber optic -0.001940 0.413192 -0.009 0.99306
OnlineSecurityNo internet service -0.315662 0.18292 -4.090 0.00004 ***
OnlineBackupNo -0.373409 0.125959 -2.980 0.00328 ***
DeviceProtectionNo -0.299927 0.127908 -2.302 0.02085 *
TechSupportNo -0.529041 0.137737 -3.901 0.00009 ***
StreamingNo internet service -0.052472 0.19872 -0.264 0.79371
StreamingTVNo internet service -0.052472 0.19872 -0.264 0.79371
StreamingMoviesNo internet service -0.052472 0.19872 -0.264 0.79371

Fig. 13. Fit for Logistic regression model

“Fig. 14” provides the details of confusion matrix for logistic regression.

> confusionMatrix(LogModel, testingChurn)

Confusion Matrix for Logistic Regression

|      | testingChurn |
|------|--------------|
| churn |          |
| 0     | 2468         |
| 1     | 233          |

Fig. 14. Confusion matrix for Logistic regression model

“Fig. 15” provides the details of odds ratio for logistic regression.
IV. LOGISTIC REGRESSION CONCLUSION

Logistic regression works better when there is single decision is to be made. Decision tree can be used when there is more than one decision to be made. Decision tree can be scaled up to more complex and more liable to over fit. Logistic regression is useful when there is categorical variable to analyze. Logistic regression is simple and has low variance and less prone to over fitting the values.

Fig. 15. Odds ratio for Logistic regression

Fig. 16. Tree plot for Logistic regression

Based on the “Fig.16”, “Contract” is the most important variable to predict customer churn or not. Customer contract even if its one year or two year contract and even without paperless billing, customer is less likely to churn. If the customer is on month-to-month contract and tenure group of 0-12 months and paperless billing is more likely to churn.
REFERENCES

[1] Selva Prabhakaran, 2016-2017, Logistic Regression. Available [Online] http://r-statistics.co/Logistic-Regression-With-R.html.

Sharan Kumar Paratala Rajagopal is a Senior Manager with Capgemini America, Inc. having 14+ years of design, development and architecture experience. He is specialized in Java/J2EE, Integration methodologies, Guidewire Product, Data Analytics, AI and Cloud technologies. He has vast domain experience in Public Services, Hospitality and Property & Casualty Insurance. Also contributed multiple technical articles to major Dev communities.