Analysis and Predictions on the Causes of Customer’s Churn for Businesses

Manisha Reddy Gopu

15/06/2022

## Summary

Customer churn is known to cause a decrease in flow of customers to businesses. In most instances where businesses experience customer churn, a competitive advantage of the business is lost. In addition, the businesses will always fall in a survival dilemma in the case whee the growth of new potential customers cannot meet the needs of the business development. Therefore, focusing in the customer churn prediction, this paper looked into the Telecom industry based in USA. Through the use of the logistic regression algorithm, the project has established a customer churn prediction model. From this analysis using the logistic regression model, it was possible to identify potential churned customers and the strategies that will act as win-back of customers. Through the data mining algorithm implemented in this paper, it was possible to identify the causes and trends of customer customers. This help to answer the research question of whether there is a positive impact on the repeated purchase behavior for customers from previous consumption as well as whether the willingness of customers can be significantly affected negatively by prices and the customer churn rate can be increased by monthly consumption. From the results of this project, it is possible for the businesses to have a reference for identifying high risk churned customers in advance, continue to provide customers with value, enhancement of the viscosity and loyalty of customers, reduction of the maintenance cost of the customers, and maintaining the high value customers. The dataset used was retrieved from Kaggle.com website.

## Introduction

For any business enterprise, loyal customers are essential as they have a crucial role in promoting the core competitiveness and improving their performance. In term of cost of publicity, loyal customers help in reducing this costs and also help in attracting more new potential customers in a business enterprise. Loyal customers also help in increasing the opportunity for businesses to obtain basic profits as well as help businesses achieve premium income. Therefore, this project intends to solve the problem of predicting high-value customer churn based on various factors of Telecom industry. The project implements the logistic regression algorithm that helps to realize customer churn prediction. The content in this report is arranged as follows. The next section, section 2, is the literature review, that is summaries the key points from related literature review for this project. The next section includes the theory section that includes some of the hypotheses that were tested in this project. The next section include the data section that describes the source of the data and explanation of variables. The next section includes the methodology section, then the results section, and lastly the conclusion section. ## Literature Review

Customer churn includes the phenomenon where the customers are no more getting services or products from a business for a number of reasons. With the basis of the problem associated with customer churn, there has been previous research conducted in trying to identify the main cause of the customer churn as well as look for strategies that will help to win back potential customers.

According to Zhao et al. (2021), the main leading factors that cause customer churn for the Telecom industry are household income and monthly ISP consumption. In another study by Gabhane & Aslam Suriya (2022), the authors performed investigations on the influence of switching costs and customer satisfaction on customer churn. The authors concluded that with unchanged customer satisfaction, the higher the switching cost, the less likely for customer churn. In a study by Sato et al. (2010), the authors performed a comparison between the impact of decision tree algorithm and principal component analysis on customer churn prediction. From a comprehensive analysis of the existing studies, it has been identified that customer churn is an important aspect for customer relationship management that may lead to huge losses to business’s development and profits. It has also been identified that the research in the strategies needed to win-back loyal studies has been the main focus. However, with the advancement and increased large amounts of data there have been a limited research on customer churn in Telecom industry. This project has a main focus in answering the research question of whether length of contract and the tenure group have any impact on customer churn.

## Theory

H1: There is no customer churn on the repeated month-to-month purchase behaviour for customers.

## Data

The data for this project is retrieved from <https://www.kaggle.com/code/supratimhaldar/telco-customer-churn-exploratory-data-analysis/data>. This data involves:

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

The dataset contains 21 features and 7043 customers. The target variable in this dataset is the churn column. To have accurate results the dataset needs to be clean with no missing values. Therefore, the first step to check for any missing values. The Totalcharges column has 11 missing values.

sapply(customer\_churn, function(x) sum(is.na(x)))

## customerID gender SeniorCitizen Partner   
## 0 0 0 0   
## Dependents tenure PhoneService MultipleLines   
## 0 0 0 0   
## InternetService OnlineSecurity OnlineBackup DeviceProtection   
## 0 0 0 0   
## TechSupport StreamingTV StreamingMovies Contract   
## 0 0 0 0   
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges   
## 0 0 0 11   
## Churn   
## 0

The rows with missing values in the TotalCharges columns are removed.

customer\_churn <- customer\_churn[complete.cases(customer\_churn), ]

Next, the No internet service is changed to NO. This column includes six column, that is, “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies”.

recode\_cols <- c(10:15)  
for(i in 1:ncol(customer\_churn[,recode\_cols])) {  
 customer\_churn[,recode\_cols][,i] <- as.factor(mapvalues  
 (customer\_churn[,recode\_cols][,i], from =c("No internet service"),to=c("No")))  
}

Next, the No phone services for column MultipleLines is changed to No.

customer\_churn$MultipleLines <- as.factor(mapvalues(customer\_churn$MultipleLines,   
 from=c("No phone service"),  
 to=c("No")))

With tenure being between 1 month and 72 months, it was grouped to five groups, that is, “0–12 Month”, “12–24 Month”, “24–48 Months”, “48–60 Month”, “> 60 Month”.

Tenure\_Group <- function(tenure){  
 if (tenure >= 0 & tenure <= 12){  
 return('0-12 Month')  
 }else if(tenure > 12 & tenure <= 24){  
 return('12-24 Month')  
 }else if (tenure > 24 & tenure <= 48){  
 return('24-48 Month')  
 }else if (tenure > 48 & tenure <=60){  
 return('48-60 Month')  
 }else if (tenure > 60){  
 return('> 60 Month')  
 }  
}  
customer\_churn$Tenure\_Group <- sapply(customer\_churn$tenure,Tenure\_Group)  
customer\_churn$Tenure\_Group <- as.factor(customer\_churn$Tenure\_Group)

Next, the values for SeniorCitizen column are changed to No 0r Yes.

customer\_churn$SeniorCitizen <- as.factor(mapvalues(customer\_churn$SeniorCitizen,  
 from=c("0","1"),  
 to=c("No", "Yes")))

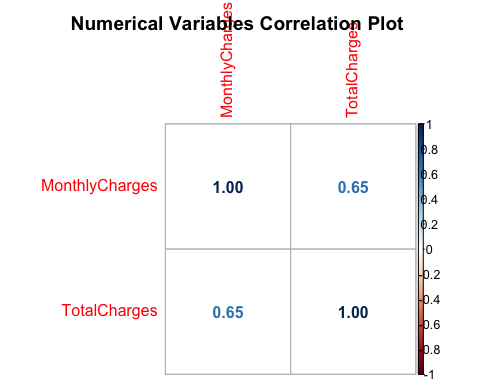
Since the customerID and tenure columns are not needed for the analysis in the project, we remove them.

customer\_churn$customerID <- NULL  
customer\_churn$tenure <- NULL

## Methodology

The dataset is cleaned and now ready for analysis. Therefore, the first thing is to identify the correlation of numerical variables.

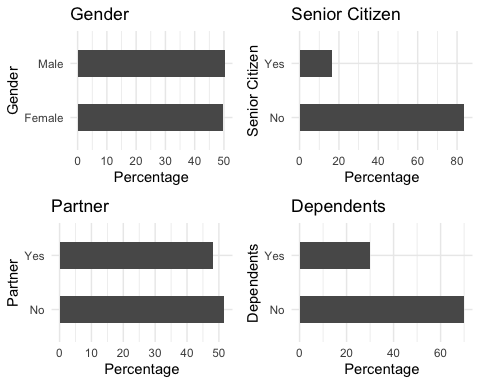
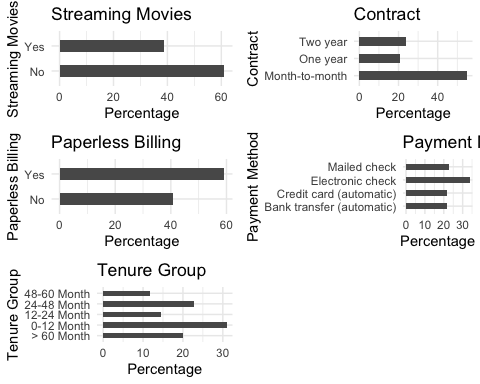
variables\_numer <- sapply(customer\_churn, is.numeric)  
matrix\_corre <- cor(customer\_churn[,variables\_numer])  
corrplot(matrix\_corre, main="\n\nNumerical Variables Correlation Plot ", method="number")



There is a correlation between the MonthlyCharges and TotalCharges. Therefore, we remove the TotalCharges column.

customer\_churn$TotalCharges <- NULL

Next, to illustrate distrubution of some of the categorical variables, boxplots were created as shown below.

Next, the dataset is split to training set and testing set.

split\_data<- createDataPartition(customer\_churn$Churn,p=0.7,list=FALSE)  
set.seed(2017)  
training<- customer\_churn[split\_data,]  
testing<- customer\_churn[-split\_data,]

Having data split, the next step is to fit the logistic model to identify most relevant features for predicting and deciciding if a customer will churn or not.

Model\_LR <- glm(Churn ~ .,family=binomial(link="logit"),data=training)  
print(summary(Model\_LR))

##   
## Call:  
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9801 -0.6631 -0.2871 0.6710 3.0762   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.719854 0.990371 -0.727 0.467316   
## genderMale -0.101087 0.077810 -1.299 0.193891   
## SeniorCitizenYes 0.187685 0.101175 1.855 0.063589 .   
## PartnerYes -0.086651 0.094111 -0.921 0.357193   
## DependentsYes -0.148431 0.108909 -1.363 0.172917   
## PhoneServiceYes 0.390944 0.778913 0.502 0.615731   
## MultipleLinesYes 0.485944 0.213230 2.279 0.022669 \*   
## InternetServiceFiber optic 1.963366 0.955457 2.055 0.039889 \*   
## InternetServiceNo -1.895331 0.967332 -1.959 0.050073 .   
## OnlineSecurityYes -0.219058 0.215084 -1.018 0.308451   
## OnlineBackupYes 0.081326 0.209487 0.388 0.697857   
## DeviceProtectionYes 0.210490 0.212285 0.992 0.321421   
## TechSupportYes -0.098246 0.216445 -0.454 0.649895   
## StreamingTVYes 0.738369 0.391199 1.887 0.059100 .   
## StreamingMoviesYes 0.683448 0.391969 1.744 0.081224 .   
## ContractOne year -0.772653 0.128162 -6.029 1.65e-09 \*\*\*  
## ContractTwo year -1.634580 0.213062 -7.672 1.70e-14 \*\*\*  
## PaperlessBillingYes 0.283740 0.089755 3.161 0.001571 \*\*   
## PaymentMethodCredit card (automatic) -0.143765 0.137636 -1.045 0.296240   
## PaymentMethodElectronic check 0.348114 0.111351 3.126 0.001770 \*\*   
## PaymentMethodMailed check -0.003897 0.134481 -0.029 0.976885   
## MonthlyCharges -0.043475 0.038052 -1.143 0.253239   
## Tenure\_Group0-12 Month 1.671525 0.204734 8.164 3.23e-16 \*\*\*  
## Tenure\_Group12-24 Month 0.780701 0.200848 3.887 0.000101 \*\*\*  
## Tenure\_Group24-48 Month 0.476086 0.182072 2.615 0.008927 \*\*   
## Tenure\_Group48-60 Month 0.082454 0.201020 0.410 0.681675   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5702.8 on 4923 degrees of freedom  
## Residual deviance: 4080.5 on 4898 degrees of freedom  
## AIC: 4132.5  
##   
## Number of Fisher Scoring iterations: 6

It is identified that Contract, Tenure\_Group and PaperlessBilling are the most relevant features.

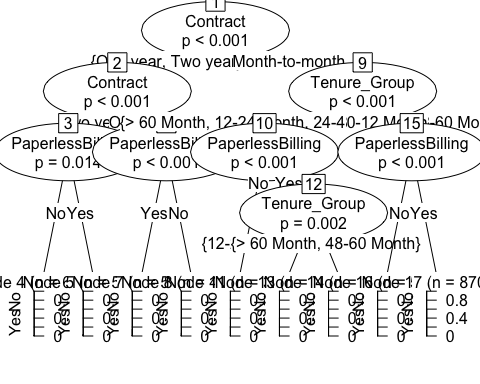
Therefore, this features are used in creating the decision tree model.

D\_Tr <- ctree(Churn~Contract+Tenure\_Group+PaperlessBilling, training)

## Results

Having created the two models, there were various results that have been identified. The results of the analysis illustrated that there are three features that needs some keen look. The Tenure\_Group, Contract and PaperlessBilling have been identified to be relevant features that may contribute to customer churn. Apart from that, the reults from the Decision Tree model illustrate that Contract is one of the most important variable that the Telco company can use to predict whether the customer will churn or not. Also, in the case where the customer has contract for one or two years, there is higher likelihood to churn. In addition, the customer is more likely to churn in the case where they have a month-to-month contract, using PaperlessBilling, and are in Tenure group of 0-12 months.

plot(D\_Tr)



The Decision Tree has an accuracy of 76.3 % in predicting customer churn.

pred\_tree <- predict(D\_Tr, testing)  
pred <- predict(D\_Tr, training)  
tab1 <- table(Predicted = pred, Actual = training$Churn)  
tab2 <- table(Predicted = pred\_tree, Actual = testing$Churn)  
print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))

## [1] "Decision Tree Accuracy 0.763757115749526"

## Conclusion

From the project, the contract, monthly charges, internet service, and tenure group are a major player in customer churn. The theory, there is no customer churn on the repeated month-to-month purchase behaviour for customers, is not true.This is because, from the analysis, there is a higher likelihood of customer churn for customers with month-to-month contract and have a tenure within 12 months.Gender and churn do not have any relationship.

## References

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A prediction model of customer churn considering customer value: an empirical research of telecom industry in China. Discrete Dynamics in Nature and Society, 2021.

Gabhane, M. D., & Aslam Suriya, D. S. (2022). Churn Prediction in Telecommunication Business using CNN and ANN. Journal of Positive School Psychology, 4672-4680.

Sato, T., Huang, B. Q., Huang, Y., Kechadi, M. T., & Buckley, B. (2010, November). Using PCA to predict customer churn in telecommunication dataset. In International Conference on Advanced Data Mining and Applications (pp. 326-335). Springer, Berlin, Heidelberg.