Predictive Modelling and Customer Retention: A Machine Learning Approach to Analyze Churn

Nithin Narayan Koranchirath
Subject Matter Expert (SME), Leading Heath Insurance Company, Richmond, United States

ABSTRACT:

This paper explores the phenomenon of customer churn in the telecommunications industry and investigates how multiple regression analysis and machine learning techniques can be employed to uncover insights from data, aiding in churn prediction and management. By examining various factors influencing customer churn and leveraging advanced analytical methods, telecom companies can develop proactive strategies to mitigate churn and enhance customer retention. In the dynamic landscape of telecommunications, customer retention is paramount for sustainable growth and competitiveness. This study navigates the intricate realm of churn analysis and prediction, delving into the pivotal role of advanced analytics and machine learning techniques in understanding and managing customer churn. Through an exhaustive exploration of key findings, it becomes apparent that the multifaceted nature of churn behavior demands sophisticated data-driven methodologies for precise prediction and mitigation. Emerging trends such as real-time prediction and personalized retention strategies offer promising avenues for telecom operators to fortify customer loyalty and propel business expansion. Recommendations underscore the critical importance of investing in advanced analytics capabilities, fostering a customer-centric ethos, and embracing innovation. By harnessing the power of data-driven insights and strategic initiatives, telecom operators can optimize the customer experience, curtail churn rates, and navigate towards enduring success in an intensely competitive market milieu.

KEYWORDS: Churn Prediction, Multi Regression Model, Supervised Machine Learning, Customer Loyalty, Advanced Analytics, Customer Retention, Machine Learning, Predictive modeling

INTRODUCTION

The telecommunications industry is marked by fierce competition, rapid technological advancements, and evolving consumer preferences. In this dynamic landscape, one of the greatest challenges faced by telecom companies is the phenomenon of customer churn[1]. Customer churn, also known as customer attrition, refers to the loss of subscribers or customers who cease using a company's services or products within a given time period[2]. For telecom operators, high churn rates not only result in revenue loss but also indicate underlying issues in service quality, customer satisfaction, and market competitiveness.

Understanding and effectively managing customer churn are critical for the long-term success and sustainability of telecommunication businesses. Identifying the factors influencing churn, predicting potential churners, and implementing proactive retention strategies are essential steps in mitigating churn and fostering customer loyalty. Traditionally, telecom companies have relied on basic statistical methods and simple models to analyze churn patterns. However, with the advent of big data and advanced analytics, there has been a paradigm shift towards more sophisticated techniques for churn prediction and management.

This paper aims to explore the complex dynamics of customer churn in the telecommunications sector and investigate how multiple regression analysis and machine learning methodologies can be leveraged to unravel insights from data. By delving into various factors contributing to churn and harnessing the predictive power of advanced analytical models, telecom operators can gain actionable insights to enhance customer retention efforts and improve overall business performance.

The introduction of multiple regression analysis and machine learning into churn analysis represents a significant advancement in the field of telecommunications analytics. Multiple regression analysis is a statistical technique used to examine the relationship between a dependent variable (such as churn) and multiple independent variables (or predictors) that may influence the outcome. By identifying significant predictors and estimating their impact on churn, regression analysis provides valuable insights into the drivers of customer attrition.

Furthermore, machine learning algorithms offer a powerful toolkit[1] for churn prediction and modeling. These algorithms can analyze large volumes of data, detect complex patterns, and make accurate predictions about future churn behavior[3]. From logistic regression and decision trees to ensemble methods like random forests and gradient boosting, a wide array of machine learning techniques can be applied to churn analysis, each with its own strengths and limitations.

The relevance of this research extends beyond academic interest to practical implications for telecom operators. By uncovering the underlying drivers of churn and developing predictive models, companies can tailor their marketing strategies, improve service offerings, and implement targeted retention initiatives. Moreover, the insights derived from data-driven churn analysis enable telecom operators to proactively address customer needs, enhance customer satisfaction, and foster long-term relationships with subscribers.

In summary, this paper aims to bridge the gap between theory and practice in the domain of telecommunications analytics by examining the intricate relationship between data, customer churn, and business outcomes. Through the lens of multiple regression analysis and machine learning, we seek to unravel the complex dynamics of churn behavior and empower telecom companies with actionable insights to optimize customer retention efforts and thrive in an increasingly competitive market landscape.

LITERATURE REVIEW

Customer churn in the telecommunications industry has been a subject of extensive research due to its significant impact on business performance and profitability. Understanding the underlying causes of churn and developing effective strategies to mitigate it are crucial for telecom operators seeking to retain customers and sustain growth. In this literature review, we examine previous studies on customer churn in telecom, traditional methods of churn analysis[4], and the application of multiple regression analysis and machine learning techniques for churn prediction. Numerous studies have investigated the factors contributing to customer churn in the telecommunications sector. These studies have identified various determinants of churn, including service quality, pricing, customer satisfaction, network coverage, and competitive offerings. For example, Keaveney and Parthasarathy (2001) found that service quality and customer satisfaction were strong predictors of churn in the cellular telecommunications industry. Similarly, Kim and Park (2013) highlighted the role of perceived value and switching costs in influencing churn behavior among mobile phone users.

Furthermore, research has shown that demographic factors, such as age, income, and occupation, also play a significant role in customer churn. For instance, Lee and Kim (2016) demonstrated that younger customers were more likely to churn compared to older subscribers, indicating the importance of demographic segmentation in churn analysis. Additionally, the emergence of new technologies and services, such as 5G, IoT, and Subscription Services, has introduced new dynamics into the churn landscape, necessitating continuous adaptation and innovation by telecom companies (Chen et al., 2020).

Traditionally, telecom operators have employed basic statistical methods and simple models to analyze churn patterns. These methods typically involve calculating churn rates, conducting customer surveys[5], and performing descriptive analyses to identify trends and patterns. However, traditional approaches have several limitations, including their reliance on aggregated data, inability to capture complex relationships, and lack of predictive power.

To address these limitations, researchers have increasingly turned to advanced analytical techniques, such as multiple regression analysis and machine learning, to enhance churn prediction and management.

Multiple Regression Analysis for Churn Prediction

Multiple regression analysis is a statistical technique used to examine the relationship between a dependent variable (e.g., churn) and multiple independent variables (predictors) that may influence the outcome. In the context of churn prediction, multiple regression allows researchers to identify the key drivers of churn and estimate their impact on customer behavior.

For example, Liu et al. (2018) applied multiple regression analysis to identify the factors influencing customer churn in the Chinese telecom market. Their study revealed that factors such as service quality, pricing, and customer satisfaction significantly influenced churn behavior, highlighting the importance of these variables in predicting customer attrition.

Moreover, multiple regression analysis enables researchers to assess the relative importance of different predictors and develop predictive models that can accurately forecast churn probabilities for individual customers. By leveraging historical data on customer interactions, service usage, and demographic characteristics, telecom operators can identify at-risk customers and implement targeted retention strategies to mitigate churn.

In recent years, machine learning algorithms have emerged as powerful tools for churn prediction and modeling[1]. These algorithms can analyze large volumes of data, detect complex patterns, and make accurate predictions about future churn behavior. A wide range of machine learning techniques, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting, have been applied to churn analysis with varying degrees of success.

For instance, Verbeke et al. (2014) compared the performance of different machine learning algorithms for churn prediction in the telecommunications industry. Their study found that ensemble methods, such as random forests and gradient boosting, outperformed traditional models in terms of predictive accuracy and robustness.

Furthermore, machine learning algorithms offer the flexibility to handle diverse data types and incorporate non-linear relationships between predictors and churn. By leveraging advanced feature selection techniques and model optimization methods, researchers can develop highly accurate churn prediction models that capture the nuances of customer behavior and market dynamics.

Advantages and Limitations

Both multiple regression analysis and machine learning techniques offer distinct advantages for churn prediction in the telecommunications industry[1][2][5]. Multiple regression analysis provides a transparent framework for identifying key drivers of churn and estimating their effects on customer behavior. It is relatively easy to interpret and implement, making it accessible to practitioners with limited statistical expertise.

On the other hand, machine learning algorithms offer greater flexibility and predictive power, allowing researchers to capture complex patterns and interactions in the data. These algorithms can handle large-scale datasets and adapt to changing market conditions, making them well-suited for dynamic and rapidly evolving telecom environments.

However, both approaches have certain limitations that need to be addressed. Multiple regression analysis assumes linear relationships between predictors and churn, which may not always hold true in practice[6]. Moreover, it requires careful selection of predictors and model specification to avoid multicollinearity and overfitting issues.

Similarly, machine learning models are susceptible to overfitting, especially when trained on noisy or unbalanced data. Additionally, they may lack transparency and interpretability, making it challenging to understand the underlying factors driving churn predictions.

The literature review highlights the importance of customer churn analysis in the telecommunications industry and the role of advanced analytical techniques in predicting and managing churn. Previous studies have identified various factors influencing churn behavior, including service quality, pricing, customer satisfaction, and demographic characteristics. While traditional methods of churn analysis have certain limitations, multiple regression analysis and machine learning offer promising avenues for improving churn prediction accuracy and effectiveness. By leveraging these techniques, telecom operators can gain actionable insights into customer behavior, develop targeted retention strategies, and enhance long-term customer relationships. However, further research is needed to address the challenges associated with model interpretation, data quality, and real-time deployment in operational settings.

DATA COLLECTION AND PREPROCESSING

For this assessment I will utilize “Telecom Data”. Using the data set, analysis is to predict customers at high risk of attrition using dataset provided. Additionally, as part of analysis, objective is to determine variables most important to accurately predict and identify customer at risk of attrition based on other conditions and factors[8]. Customer churn, also known as customer attrition, is the loss of clients or customers. Churn is an important business metric for subscription-based services such as telecommunications companies.

I will use the R statistical programming language in order to identify variables associated with customer attrition. Correlation between Churn and principal components will help telecommunications companies with valuable insight that can help early identify and detect customers at high risk of attrition.

Objective of data analysis is to identify, why customers are leaving and identify potential indicators that can help explain why customers are leaving that will allow company can make an informed plan in order to mitigate further loss. To achieve below goal we’ll analyze:

- Impact of customer tenure length on customer churn
- Impact of type of contract on customer churn
- Impact of monthly payments on customer churn
- Impact of customer loyalty on types of services purchase.

TOOL SELECTION

R is most popular choice for data scientist among academic and industrial community. Traditionally, R is used for research purpose at the academy. R provides numerous statistical tools for analytics[5]. With advancement in data science and increasing need to data, R became natural choice for industrial data scientist. DataMentor.(n.d.)

- Python and R both are open source languages and are free to download. There are multiple documentations and online help from support community available for both languages. Small and medium sized companies and independent analyst prefer these 2 languages over SAS as initial investments are minimum as there are no licensing cost involved. On the other hand, SAS has licensed software and a very expensive one[1].

- R has highly advanced graphical capabilities. There are numerous packages which provide you advanced graphical capabilities. Compared to R, Python has medium graphical capabilities. Python with latest release has Seaborn, making custom plots easy. SAS has decent functional graphical capabilities. But it is just functional. Any customization on plots are difficult and requires deep understanding of SAS Graph package[1].

- Due to their open nature, R & Python get latest features quickly. SAS, on the other hand updates its capabilities in new version rollouts. Since R has been used widely in academics in past, development of new techniques is fast. SAS releases updates in controlled environment, hence they are well tested; but tradeoff is time to market and customization per requirement[2].

Python has had great advancements in the field and has numerous packages like Tensorflow and Keras. R has recently added support for those packages, along with some basic ones too. The kerasR and keras packages in R act as an interface to the original Python package, Keras.

MULTIPLE REGRESSION ANALYSIS FOR CHURN PREDICTION

Multiple regression analysis is a statistical technique used to examine the relationship between a dependent variable (such as churn) and multiple independent variables (predictors) that may influence the outcome. It provides a transparent framework for identifying significant predictors and estimating their effects on churn behavior[9]. Multiple linear regression analysis makes following key assumptions:

- There must be a linear relationship between the outcome variable and the independent variables[2][5]. Scatterplots can show whether there is a linear or curvilinear relationship.

- Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.

- No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values[5].

- Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

- Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis. Learn more about sample size here.

STRENGTHS:

One of the key strengths of multiple regression analysis is its interpretability. The regression coefficients indicate the direction and magnitude of the relationship between predictors and churn, allowing researchers to understand the underlying drivers of customer attrition.

The regression model formulation is transparent, making it easy to replicate and validate results. This transparency is particularly important in industries like telecommunications, where regulatory compliance and accountability are paramount.

Multiple regression analysis allows researchers to perform variable selection and assess the relative importance of different predictors. By including only the most relevant variables in the model, analysts can improve model parsimony and reduce overfitting.

WEAKNESSES:

- Multiple regression analysis assumes linear relationships between predictors and the dependent variable[10]. However, in real-world scenarios, these relationships may be non-linear or exhibit complex interactions, limiting the model's predictive accuracy.

- The linear nature of regression models restricts their flexibility in capturing complex patterns and interactions in the data. This limitation may result in underfitting, especially when dealing with high-dimensional or heterogeneous datasets.

Through multiple regression analysis, telecom operators can gain valuable insights into the drivers of customer churn. By identifying significant predictors such as service quality, pricing, and customer satisfaction, companies can prioritize areas for improvement and develop targeted retention strategies. Moreover, regression analysis facilitates the interpretation of results, enabling stakeholders to make informed decisions based on actionable insights derived from the data.

DATA COLLECTION AND PREPROCESSING

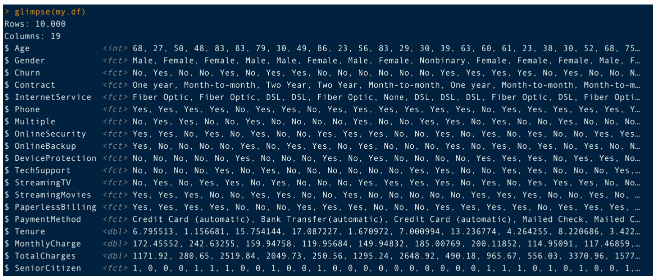
Objective of data analysis is to predict, customer churn and identify potential indicators that predict and/or explain why customers are leaving that will allow company to make an informed decision in order to mitigate further loss. To achieve below goal we'll analyze:

- Impact of customer tenure length on customer churn
- Impact of MonthlyCharge on customer churn
- Impact of TotalCharge on customer churn
- Impact of customer loyalty on types of services purchase

I will be utilizing R to access quality of data and perform data cleanup. R provides platform for interpretable visualizations. For cleaning data R provides build in packages to identifying outliers[2]. There are also packages like 'caret' to recode values with dummy variables, 'readr' package is used for reading the raw data file into R, and dplyr and ggplot2 pakages are used for calculations and visualizations.

- Drop redundant variables, CaseOrder, Customer_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, determined irrelevant for analysis.
- Rename variables with incorrect/irrelevant variable names like Phone, Multiple, Port_modem, Bandwidth_GB_Year, Yearly_equip_failure.
- Recode variables Age, Tenure, MonthlyCharge, YearlyDataUsage to bivariate variables.

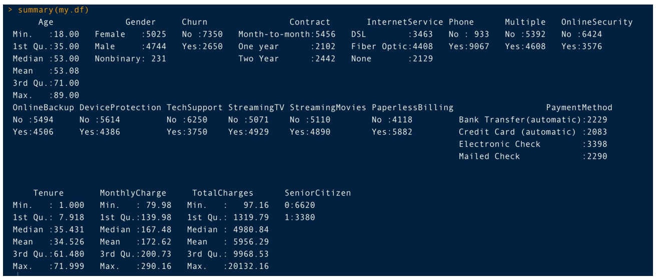
For analysis I will use Telco Customer Churn[8] dataset. R library function str will display internal structure of an R object. R function glimpse is a diagnostic function; output displays one line for each 'basic' structure. Function is especially well suited to compactly display the (abbreviated) contents of (possibly nested) lists. Function calls args for (non-primitive) function objects. RDocumentation str. (n.d.)



```
> glimpse(my.df)
Rows: 10,000
Columns: 19
$ Age              <int> 68, 27, 50, 48, 83, 83, 79, 30, 49, 86, 23, 56, 83, 29, 30, 39, 63, 60, 61, 23, 38, 30, 52, 68, 75…
$ Gender           <fct> Male, Female, Female, Male, Male, Female, Male, Female, Nonbinary, Female, Female, Female, Male, F…
$ Churn            <fct> No, Yes, No, No, Yes, No, Yes, Yes, No, No, No, No, No, No, Yes, Yes, Yes, Yes, No, Yes, No, No, N…
$ Contract         <fct> One year, Month-to-month, Two Year, Two Year, Month-to-month, One year, Month-to-month, Month-to-m…
$ InternetService  <fct> Fiber Optic, Fiber Optic, DSL, DSL, Fiber Optic, None, DSL, DSL, DSL, Fiber Optic, DSL, Fiber Opti…
$ Phone            <fct> Yes, Yes, Yes, Yes, No, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Y…
$ Multiple         <fct> No, Yes, Yes, No, No, Yes, No, No, No, No, Yes, No, No, Yes, Yes, No, Yes, No, No, Yes, No, No, No…
$ OnlineSecurity   <fct> Yes, Yes, No, Yes, No, Yes, No, No, Yes, Yes, Yes, No, No, Yes, No, Yes, No, Yes, No, No, Yes, Yes…
$ OnlineBackup     <fct> Yes, No, No, No, No, Yes, No, Yes, Yes, No, Yes, No, No, No, Yes, No, Yes, No, Yes, No, Yes, No, N…
$ DeviceProtection <fct> No, No, No, No, No, Yes, No, No, No, Yes, No, Yes, No, No, No, No, Yes, Yes, Yes, No, Yes, Yes, No…
$ TechSupport      <fct> No, No, No, No, Yes, No, Yes, No, No, No, No, No, No, Yes, Yes, No, No, Yes, Yes, No, No, No, Yes,…
$ StreamingTV      <fct> No, Yes, No, Yes, Yes, No, Yes, No, No, No, Yes, Yes, Yes, Yes, No, Yes, No, Yes, Yes, Yes, No, No…
$ StreamingMovies  <fct> Yes, Yes, Yes, No, No, Yes, Yes, No, No, Yes, No, No, No, No, No, Yes, Yes, Yes, No, No, Yes, No, …
$ PaperlessBilling <fct> Yes, Yes, Yes, Yes, No, No, No, Yes, Yes, Yes, No, No, No, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, …
$ PaymentMethod    <fct> Credit Card (automatic), Bank Transfer(automatic), Credit Card (automatic), Mailed Check, Mailed C…
$ Tenure           <dbl> 6.795513, 1.156681, 15.754144, 17.087227, 1.670972, 7.000994, 13.236774, 4.264255, 8.220686, 3.422…
$ MonthlyCharge    <dbl> 172.45552, 242.63255, 159.94758, 119.95684, 149.94832, 185.00769, 200.11852, 114.95091, 117.46859,…
$ TotalCharges     <dbl> 1171.92, 280.65, 2519.84, 2049.73, 250.56, 1295.24, 2648.92, 490.18, 965.67, 556.03, 3370.96, 1577…
$ SeniorCitizen    <fct> 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1,…
```

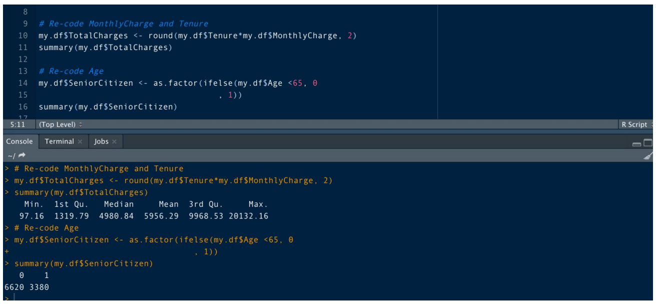
Fig1: Sample Data

Next using summary function result summaries of the result is generated for various model fitting functions. Function provides basis to analyze and assess the quality of data[11]. The function invokes particular methods which depend on the class of the first argument. RDocumentation summary .(n.d.)



```
> summary(my.df)
      Age                Gender        Churn                Contract           InternetService Phone       Multiple    OnlineSecurity
 Min.   :18.00   Female    :5025   No :7350   Month-to-month:5456   DSL         :3463   No : 933   No :5392   No :6424
 1st Qu.:35.00   Male      :4744   Yes:2650   One year      :2102   Fiber Optic:4408   Yes:9067   Yes:4608   Yes:3576
 Median :53.00   Nonbinary: 231               Two Year      :2442   None        :2129
 Mean   :53.08
 3rd Qu.:71.00
 Max.   :89.00
 OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies PaperlessBilling                   PaymentMethod
 No :5494     No :5614         No :6250    No :5071    No :5110        No :4118         Bank Transfer(automatic):2229
 Yes:4506     Yes:4386         Yes:3750    Yes:4929    Yes:4890        Yes:5882         Credit Card (automatic) :2083
                                                                                        Electronic Check        :3398
                                                                                        Mailed Check            :2290



     Tenure        MonthlyCharge     TotalCharges      SeniorCitizen
 Min.   : 1.000   Min.   : 79.98   Min.   :   97.16   0:6620
 1st Qu.: 7.918   1st Qu.:139.98   1st Qu.: 1319.79   1:3380
 Median :35.431   Median :167.48   Median : 4980.84
 Mean   :34.526   Mean   :172.62   Mean   : 5956.29
 3rd Qu.:61.480   3rd Qu.:200.73   3rd Qu.: 9968.53
 Max.   :71.999   Max.   :290.16   Max.   :20132.16
```

Fig2: Data Summary

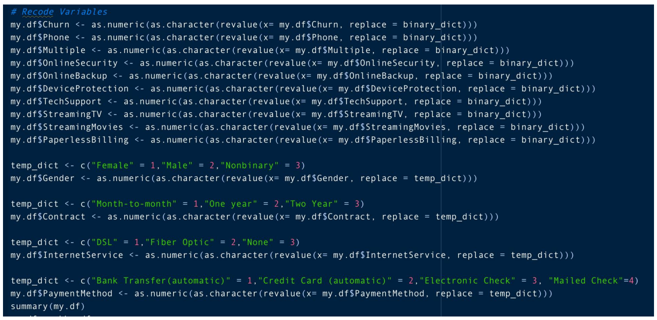


```

# Re-code MonthlyCharge and Tenure
my.df$TotalCharges <- round(my.df$Tenure*my.df$MonthlyCharge, 2)
summary(my.df$TotalCharges)

# Re-code Age
my.df$SeniorCitizen <- as.factor(ifelse(my.df$Age <65, 0
                                        , 1))

summary(my.df$SeniorCitizen)
```

```
> # Re-code MonthlyCharge and Tenure
> my.df$TotalCharges <- round(my.df$Tenure*my.df$MonthlyCharge, 2)
> summary(my.df$TotalCharges)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
  97.16  1319.79  4980.84  5956.29  9968.53 20132.16 
> # Re-code Age
> my.df$SeniorCitizen <- as.factor(ifelse(my.df$Age <65, 0
+                                         , 1))
> summary(my.df$SeniorCitizen)
   0    1 
6620 3380 
> 
```

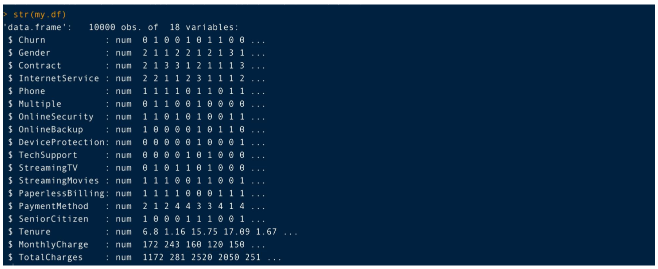
Fig3: Data Summary Total Charges

Based on analysis above redundant columns CaseOrder, Customer_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population are dropped from dataset. Inorder to enrich data further few variables like age, gender, SeniorCitizen are converted from categorical variables to numbers. Code snippets provided below.



```
# Recode Variables
my.df$Churn <- as.numeric(as.character(revalue(x= my.df$Churn, replace = binary_dict)))
my.df$Phone <- as.numeric(as.character(revalue(x= my.df$Phone, replace = binary_dict)))
my.df$Multiple <- as.numeric(as.character(revalue(x= my.df$Multiple, replace = binary_dict)))
my.df$OnlineSecurity <- as.numeric(as.character(revalue(x= my.df$OnlineSecurity, replace = binary_dict)))
my.df$OnlineBackup <- as.numeric(as.character(revalue(x= my.df$OnlineBackup, replace = binary_dict)))
my.df$DeviceProtection <- as.numeric(as.character(revalue(x= my.df$DeviceProtection, replace = binary_dict)))
my.df$TechSupport <- as.numeric(as.character(revalue(x= my.df$TechSupport, replace = binary_dict)))
my.df$StreamingTV <- as.numeric(as.character(revalue(x= my.df$StreamingTV, replace = binary_dict)))
my.df$StreamingMovies <- as.numeric(as.character(revalue(x= my.df$StreamingMovies, replace = binary_dict)))
my.df$PaperlessBilling <- as.numeric(as.character(revalue(x= my.df$PaperlessBilling, replace = binary_dict)))

temp_dict <- c("Female" = 1,"Male" = 2,"Nonbinary" = 3)
my.df$Gender <- as.numeric(as.character(revalue(x= my.df$Gender, replace = temp_dict)))

temp_dict <- c("Month-to-month" = 1,"One year" = 2,"Two Year" = 3)
my.df$Contract <- as.numeric(as.character(revalue(x= my.df$Contract, replace = temp_dict)))

temp_dict <- c("DSL" = 1,"Fiber Optic" = 2,"None" = 3)
my.df$InternetService <- as.numeric(as.character(revalue(x= my.df$InternetService, replace = temp_dict)))

temp_dict <- c("Bank Transfer(automatic)" = 1,"Credit Card (automatic)" = 2,"Electronic Check" = 3, "Mailed Check"=4)
my.df$PaymentMethod <- as.numeric(as.character(revalue(x= my.df$PaymentMethod, replace = temp_dict)))
summary(my.df)
```

Fig4: Data Summary

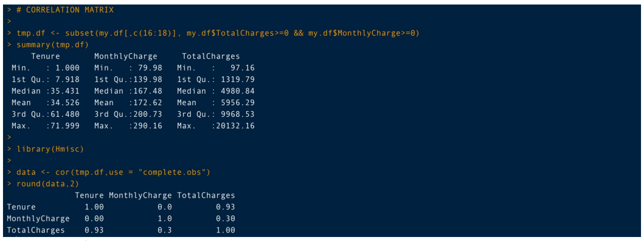


```
> str(my.df)
'data.frame':   10000 obs. of  18 variables:
 $ Churn           : num  0 1 0 0 1 0 1 1 0 0 ...
 $ Gender          : num  2 1 1 2 2 1 2 1 3 1 ...
 $ Contract        : num  2 1 3 3 1 2 1 1 1 3 ...
 $ InternetService : num  2 2 1 1 2 3 1 1 1 2 ...
 $ Phone           : num  1 1 1 1 0 1 1 0 1 1 ...
 $ Multiple        : num  0 1 1 0 0 1 0 0 0 0 ...
 $ OnlineSecurity  : num  1 1 0 1 0 1 0 0 1 1 ...
 $ OnlineBackup    : num  1 0 0 0 0 1 0 1 1 0 ...
 $ DeviceProtection: num  0 0 0 0 0 1 0 0 0 1 ...
 $ TechSupport     : num  0 0 0 0 1 0 1 0 0 0 ...
 $ StreamingTV     : num  0 1 0 1 1 0 1 0 0 0 ...
 $ StreamingMovies : num  1 1 1 0 0 1 1 0 0 1 ...
 $ PaperlessBilling: num  1 1 1 1 0 0 0 1 1 1 ...
 $ PaymentMethod   : num  2 1 2 4 4 3 3 4 1 4 ...
 $ SeniorCitizen   : num  1 0 0 0 1 1 1 0 0 1 ...
 $ Tenure          : num  6.8 1.16 15.75 17.09 1.67 ...
 $ MonthlyCharge   : num  172 243 160 120 150 ...
 $ TotalCharges    : num  1172 281 2520 2050 251 ...
```

Fig5: *Data Summary post standardization*

CORRELATION MATRIX:

Correlation matrix suggests, Tenure affects total charges greatly.



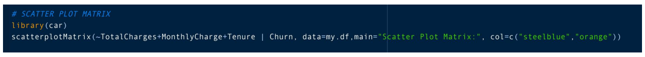
```
> # CORRELATION MATRIX
>
> tmp.df <- subset(my.df[,c(16:18)], my.df$TotalCharges>=0 && my.df$MonthlyCharge>=0)
> summary(tmp.df)
     Tenure        MonthlyCharge     TotalCharges
 Min.   : 1.000   Min.   : 79.98   Min.   :   97.16
 1st Qu.: 7.918   1st Qu.:139.98   1st Qu.: 1319.79
 Median :35.431   Median :167.48   Median : 4980.84
 Mean   :34.526   Mean   :172.62   Mean   : 5956.29
 3rd Qu.:61.480   3rd Qu.:200.73   3rd Qu.: 9968.53
 Max.   :71.999   Max.   :290.16   Max.   :20132.16
>
> library(Hmisc)
>
> data <- cor(tmp.df,use = "complete.obs")
> round(data,2)
              Tenure MonthlyCharge TotalCharges
Tenure          1.00           0.0         0.93
MonthlyCharge   0.00           1.0         0.30
TotalCharges    0.93           0.3         1.00
```

Fig6: *Correlation matrix*

CORRGRAM GRAPH:

These are some of the possible conclusions that can be deduced from the corrgram graph:

- Total charges and monthly charges are related.
- Total charges are strongly driven by tenure.



```
# SCATTER PLOT MATRIX
library(car)
scatterplotMatrix(~TotalCharges+MonthlyCharge+Tenure | Churn, data=my.df,main="Scatter Plot Matrix:", col=c("steelblue","orange"))
```

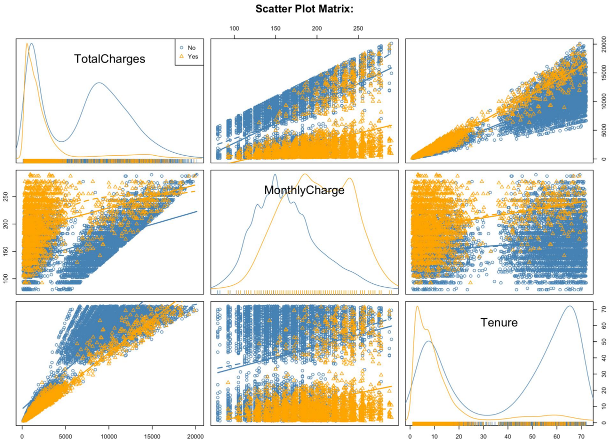
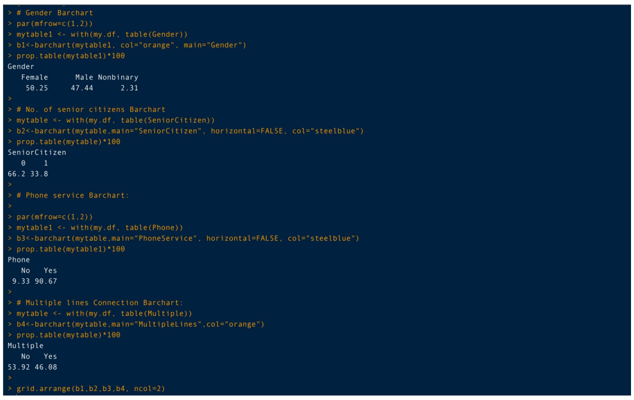


Fig7: *Scatter Plot*

UNIVARIATE VISUALIZATIONS OF VARIABLES:



```
> # Gender Barchart
> par(mfrow=c(1,2))
> mytable1 <- with(my.df, table(Gender))
> b1<-barchart(mytable1, col="orange", main="Gender")
> prop.table(mytable1)*100
Gender
   Female      Male Nonbinary
    50.25     47.44      2.31
>
> # No. of senior citizens Barchart
> mytable <- with(my.df, table(SeniorCitizen))
> b2<-barchart(mytable,main="SeniorCitizen", horizontal=FALSE, col="steelblue")
> prop.table(mytable)*100
SeniorCitizen
   0    1
66.2 33.8
>
> # Phone service Barchart:
>
> par(mfrow=c(1,2))
> mytable1 <- with(my.df, table(Phone))
> b3<-barchart(mytable,main="PhoneService", horizontal=FALSE, col="steelblue")
> prop.table(mytable1)*100
Phone
   No   Yes
 9.33 90.67
>
> # Multiple lines Connection Barchart:
> mytable <- with(my.df, table(Multiple))
> b4<-barchart(mytable,main="MultipleLines",col="orange")
> prop.table(mytable)*100
Multiple
   No   Yes
53.92 46.08
>
> grid.arrange(b1,b2,b3,b4, ncol=2)
```

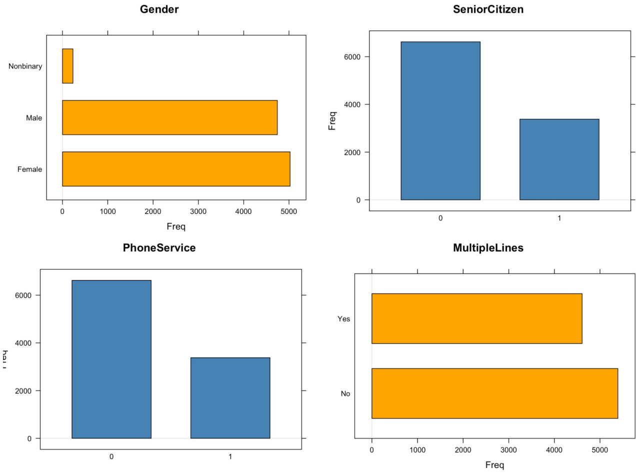
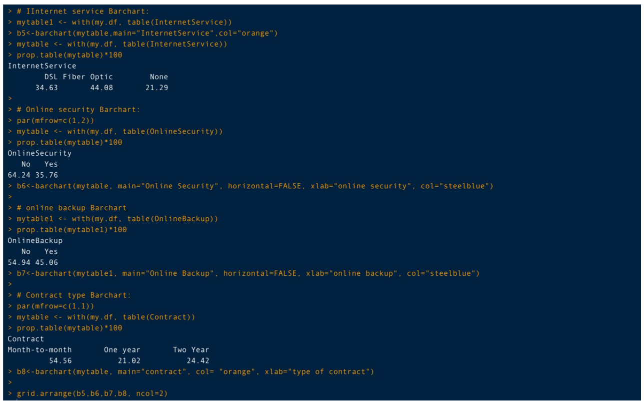


Fig8: *Visualizations Of Variables*



```
> # IInternet service Barchart:
> mytable1 <- with(my.df, table(InternetService))
> b5<-barchart(mytable,main="InternetService",col="orange")
> mytable <- with(my.df, table(InternetService))
> prop.table(mytable)*100
InternetService
        DSL Fiber Optic        None
      34.63       44.08       21.29
>
> # Online security Barchart:
> par(mfrow=c(1,2))
> mytable <- with(my.df, table(OnlineSecurity))
> prop.table(mytable)*100
OnlineSecurity
   No   Yes
64.24 35.76
> b6<-barchart(mytable, main="Online Security", horizontal=FALSE, xlab="online security", col="steelblue")
>
> # online backup Barchart
> mytable1 <- with(my.df, table(OnlineBackup))
> prop.table(mytable1)*100
OnlineBackup
   No   Yes
54.94 45.06
> b7<-barchart(mytable1, main="Online Backup", horizontal=FALSE, xlab="online backup", col="steelblue")
>
> # Contract type Barchart:
> par(mfrow=c(1,1))
> mytable <- with(my.df, table(Contract))
> prop.table(mytable)*100
Contract
Month-to-month       One year       Two Year
         54.56          21.02          24.42
> b8<-barchart(mytable, main="contract", col= "orange", xlab="type of contract")
>
> grid.arrange(b5,b6,b7,b8, ncol=2)
```

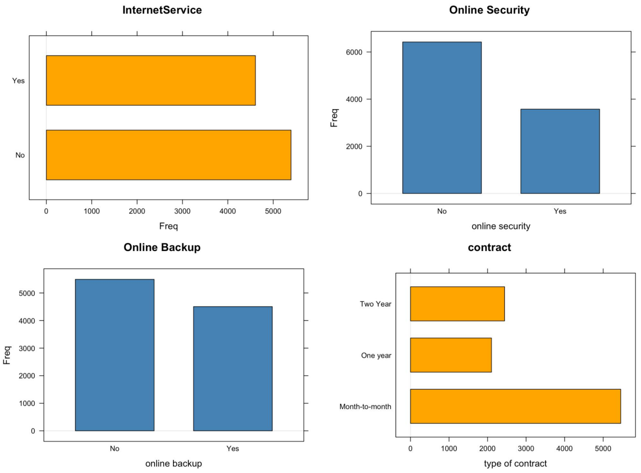
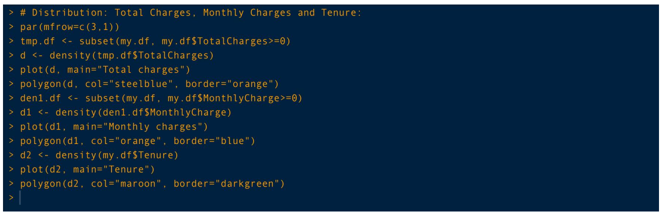


Fig9: *Visualizations Of Variables*



```
> # Distribution: Total Charges, Monthly Charges and Tenure:
> par(mfrow=c(3,1))
> tmp.df <- subset(my.df, my.df$TotalCharges>=0)
> d <- density(tmp.df$TotalCharges)
> plot(d, main="Total charges")
> polygon(d, col="steelblue", border="orange")
> den1.df <- subset(my.df, my.df$MonthlyCharge>=0)
> d1 <- density(den1.df$MonthlyCharge)
> plot(d1, main="Monthly charges")
> polygon(d1, col="orange", border="blue")
> d2 <- density(my.df$Tenure)
> plot(d2, main="Tenure")
> polygon(d2, col="maroon", border="darkgreen")
>
```

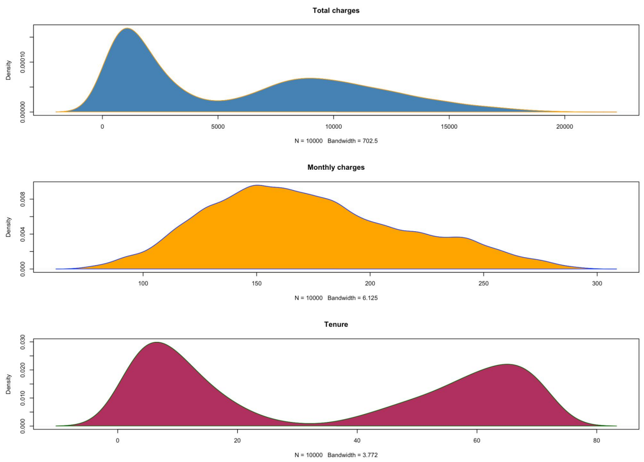
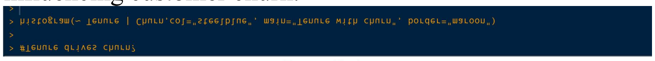


Fig10: *Visualizations Of Variables*

Based on plot above, monthly charges are more than total charges, potential reason for customers tend to leave. Many customers use the company's service for a shorter period.

The key drivers based on analysis are tenure, contract, internet service, total charges and monthly charges.

Using Bivariate plot I'll analyze establish variables most influencing customer churn.



```
> #Tenure drives churn?
>
> histogram(~ Tenure | Churn,col="steelblue", main="Tenure with churn", border="maroon")
>
```

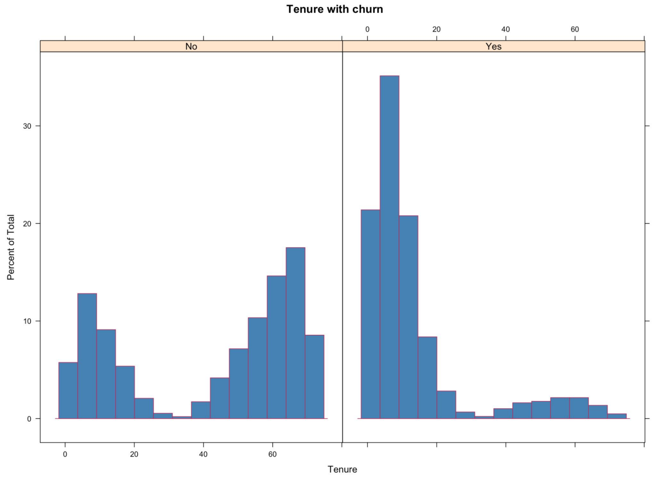
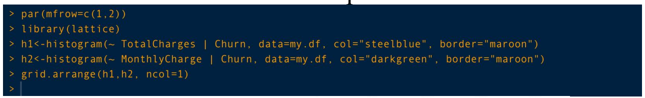


Fig11: *Tenure influences Churn*

This shows that customers who tend to leave are people who have been customers for shorter periods.



```
> par(mfrow=c(1,2))
> library(lattice)
> h1<-histogram(~ TotalCharges | Churn, data=my.df, col="steelblue", border="maroon")
> h2<-histogram(~ MonthlyCharge | Churn, data=my.df, col="darkgreen", border="maroon")
> grid.arrange(h1,h2, ncol=1)
>
```

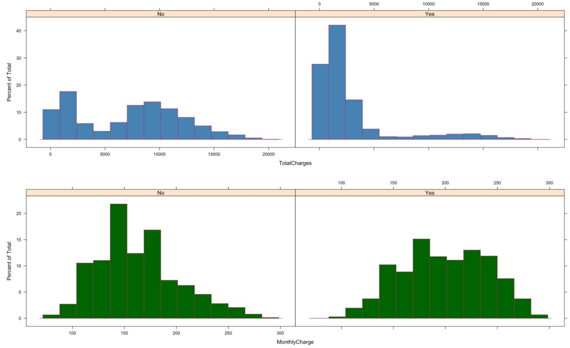
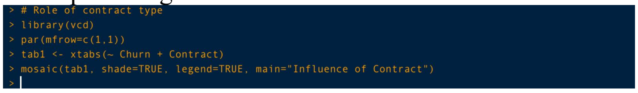


Fig12: *Influence of monthly charges and total charges on churn*

As noted in above section, due to high monthly charges, churn percentage is more in that case.



```
> # Role of contract type
> library(vcd)
> par(mfrow=c(1,1))
> tab1 <- xtabs(~ Churn + Contract)
> mosaic(tab1, shade=TRUE, legend=TRUE, main="Influence of Contract")
>
```

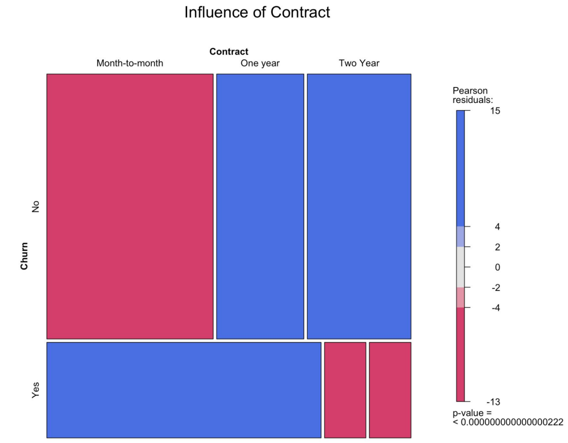
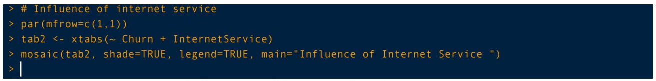


Fig13: *Influence of Contract type on Churn*

Customers on a month-to-month contract tend to leave.
Influence of Type of internet service on Churn



```
> # Influence of internet service
> par(mfrow=c(1,1))
> tab2 <- xtabs(~ Churn + InternetService)
> mosaic(tab2, shade=TRUE, legend=TRUE, main="Influence of Internet Service ")
>
```

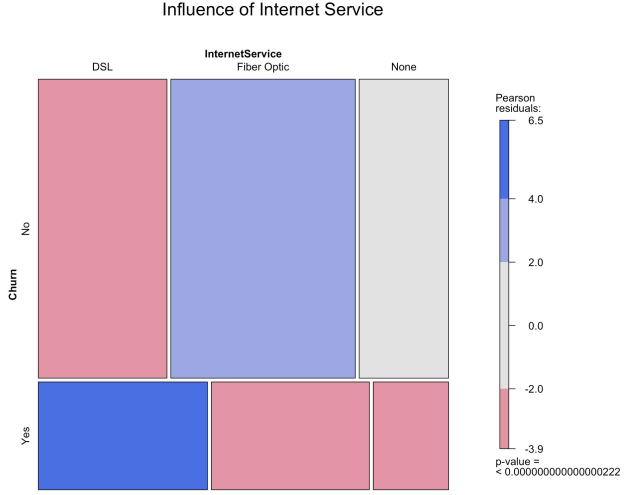
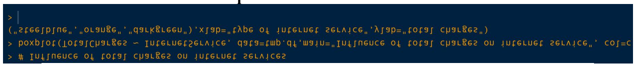


Fig14: *Influence of internet service on Churn*

Customers with fiber optic internet connection tend to leave.



```
> # Influence of total charges on internet services
> boxplot(TotalCharges ~ InternetService, data=tmp.df,main="Influence of total charges on internet service", col=c
("steelblue","orange","darkgreen"),xlab="type of internet service",ylab="total charges")
>
```

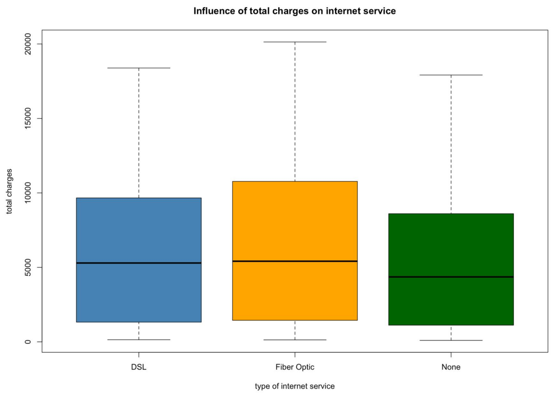
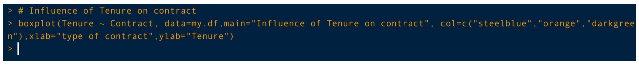


Fig15: *Influence of Type of total charges on internet services*

Total charges are more for fiber optic type of internet service.



```
> # Influence of Tenure on contract
> boxplot(Tenure ~ Contract, data=my.df,main="Influence of Tenure on contract", col=c("steelblue","orange","darkgree
n"),xlab="type of contract",ylab="Tenure")
> |
```

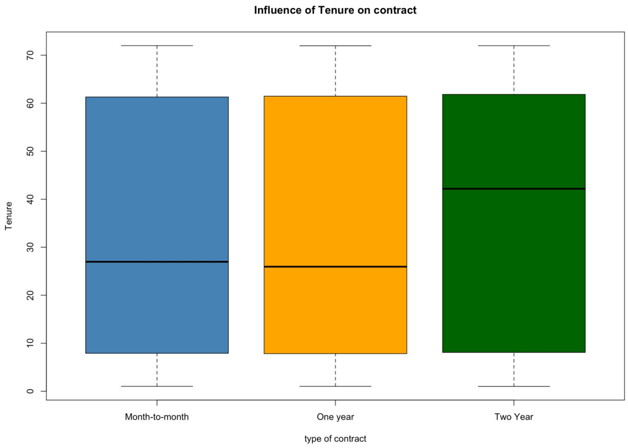
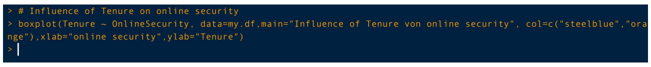


***Fig16:** Influence of tenure vs type of contract*



```
> # Influence of Tenure on online security
> boxplot(Tenure ~ OnlineSecurity, data=my.df,main="Influence of Tenure von online security", col=c("steelblue","ora
nge"),xlab="online security",ylab="Tenure")
> |
```

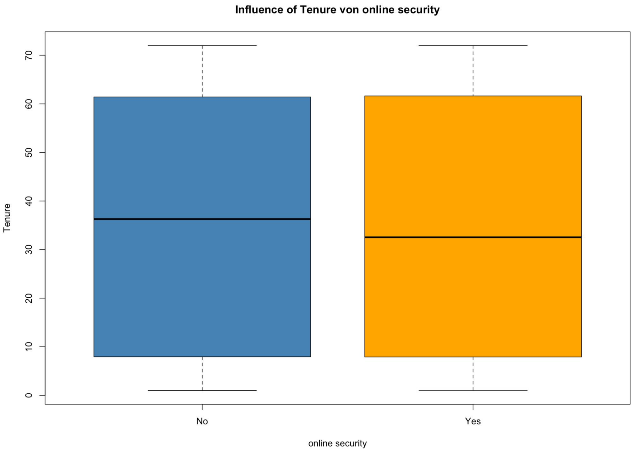
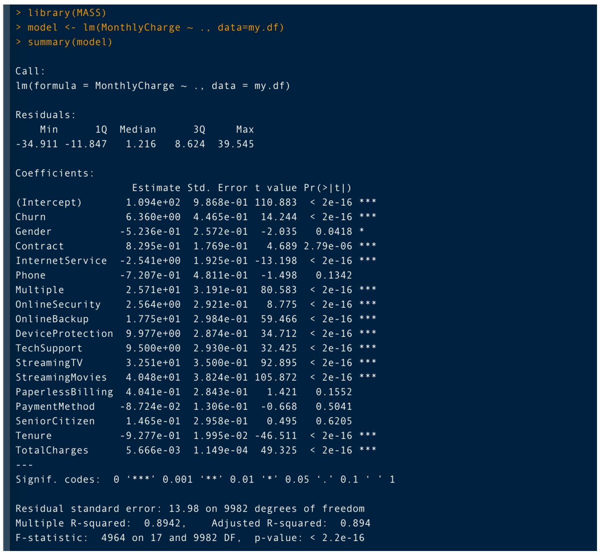


***Fig17:** Influence of tenure on online security*

Evidently, customers prefer to use the company's services for a longer time if they are assured online security.



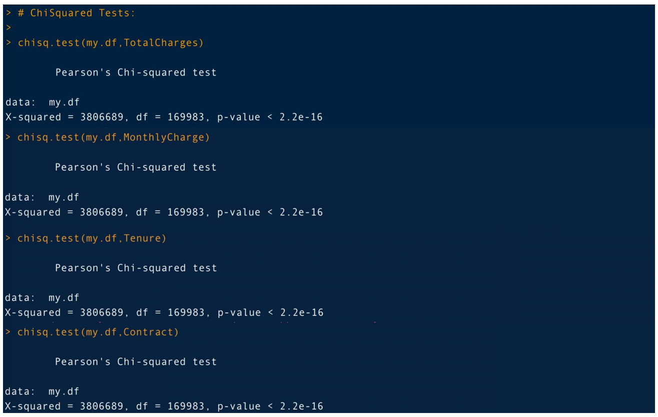
```
> library(MASS)
> model <- lm(MonthlyCharge ~ ., data=my.df)
> summary(model)

Call:
lm(formula = MonthlyCharge ~ ., data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.911 -11.847   1.216   8.624  39.545 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.094e+02  9.868e-01 110.883  < 2e-16 ***
Churn           6.360e+00  4.465e-01  14.244  < 2e-16 ***
Gender         -5.236e-01  2.572e-01  -2.035   0.0418 *  
Contract        8.295e-01  1.769e-01   4.689 2.79e-06 ***
InternetService -2.541e+00  1.925e-01 -13.198  < 2e-16 ***
Phone          -7.207e-01  4.811e-01  -1.498   0.1342    
Multiple        2.571e+01  3.191e-01  80.583  < 2e-16 ***
OnlineSecurity  2.564e+00  2.921e-01   8.775  < 2e-16 ***
OnlineBackup    1.775e+01  2.984e-01  59.466  < 2e-16 ***
DeviceProtection 9.977e+00  2.874e-01  34.712  < 2e-16 ***
TechSupport     9.500e+00  2.930e-01  32.425  < 2e-16 ***
StreamingTV     3.251e+01  3.500e-01  92.895  < 2e-16 ***
StreamingMovies 4.048e+01  3.824e-01 105.872  < 2e-16 ***
PaperlessBilling 4.041e-01  2.843e-01   1.421   0.1552    
PaymentMethod  -8.724e-02  1.306e-01  -0.668   0.5041    
SeniorCitizen   1.465e-01  2.958e-01   0.495   0.6205    
Tenure         -9.277e-01  1.995e-02 -46.511  < 2e-16 ***
TotalCharges    5.666e-03  1.149e-04  49.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.98 on 9982 degrees of freedom
Multiple R-squared:  0.8942,    Adjusted R-squared:  0.894 
F-statistic:  4964 on 17 and 9982 DF,  p-value: < 2.2e-16
```

***Fig18:** Summary of all potential predictors*

In order to perform variable analysis I'll perform two statistical test, "Chi-Squared Tests" to check data independence and "Simple Linear Regression" for two continuous variables 'Monthly Charges' and 'Total Charges'.

Chi-Squared Tests

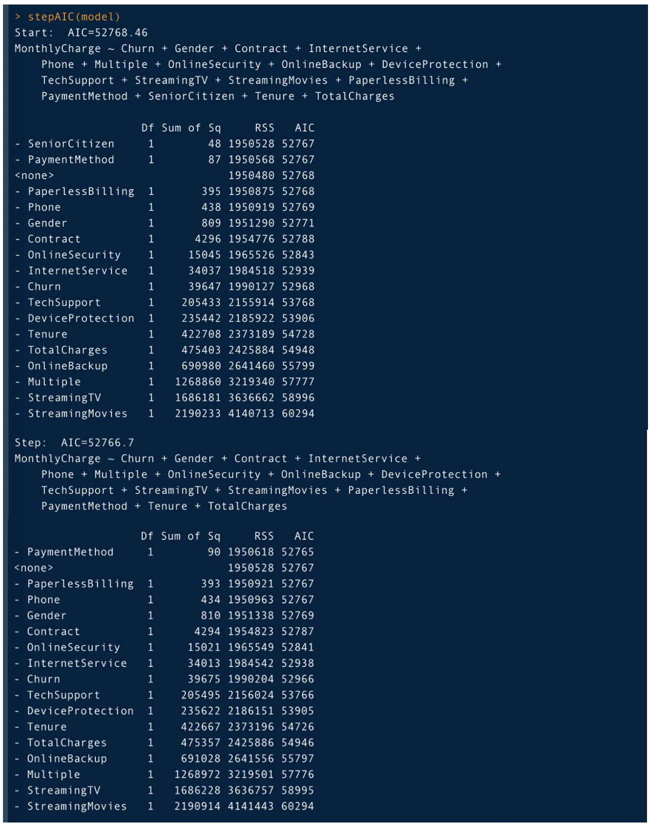
I'll first run "Chi-Squared Tests" to check for data independence.



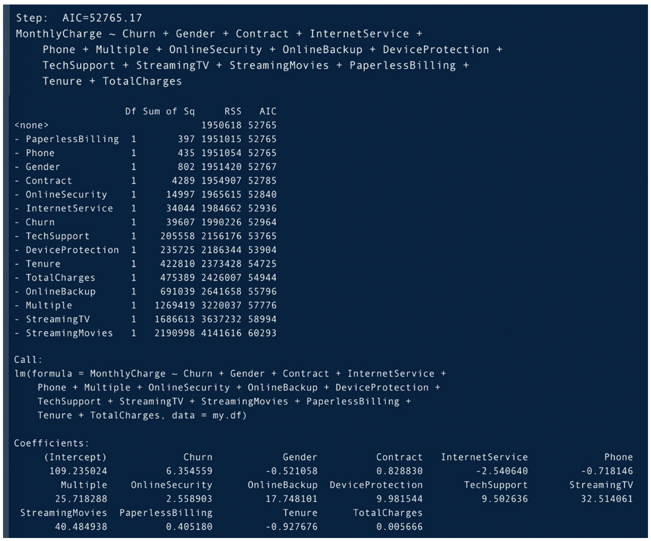
```
> # ChiSquared Tests:
> 
> chisq.test(my.df,TotalCharges)

	Pearson's Chi-squared test

data:  my.df
X-squared = 3806689, df = 169983, p-value < 2.2e-16
> chisq.test(my.df,MonthlyCharge)

	Pearson's Chi-squared test

data:  my.df
X-squared = 3806689, df = 169983, p-value < 2.2e-16

> chisq.test(my.df,Tenure)

	Pearson's Chi-squared test

data:  my.df
X-squared = 3806689, df = 169983, p-value < 2.2e-16
> chisq.test(my.df,Contract)

	Pearson's Chi-squared test

data:  my.df
X-squared = 3806689, df = 169983, p-value < 2.2e-16
```

***Fig19:** Chi-Squared Test*

Please find below screen shot for reduced multiple regression predictors identified.



```
> stepAIC(model)
Start:  AIC=52768.46
MonthlyCharge ~ Churn + Gender + Contract + InternetService + 
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection + 
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling + 
    PaymentMethod + SeniorCitizen + Tenure + TotalCharges

                   Df Sum of Sq     RSS   AIC
- SeniorCitizen     1        48 1950528 52767
- PaymentMethod     1        87 1950568 52767
<none>                          1950480 52768
- PaperlessBilling  1       395 1950875 52768
- Phone             1       438 1950919 52769
- Gender            1       809 1951290 52771
- Contract          1      4296 1954776 52788
- OnlineSecurity    1     15045 1965526 52843
- InternetService   1     34037 1984518 52939
- Churn             1     39647 1990127 52968
- TechSupport       1    205433 2155914 53768
- DeviceProtection  1    235442 2185922 53906
- Tenure            1    422708 2373189 54728
- TotalCharges      1    475403 2425884 54948
- OnlineBackup      1    690980 2641460 55799
- Multiple          1   1268860 3219340 57777
- StreamingTV       1   1686181 3636662 58996
- StreamingMovies   1   2190233 4140713 60294

Step:  AIC=52766.7
MonthlyCharge ~ Churn + Gender + Contract + InternetService + 
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection + 
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling + 
    PaymentMethod + Tenure + TotalCharges

                   Df Sum of Sq     RSS   AIC
- PaymentMethod     1        90 1950618 52765
<none>                          1950528 52767
- PaperlessBilling  1       393 1950921 52767
- Phone             1       434 1950963 52767
- Gender            1       810 1951338 52769
- Contract          1      4294 1954823 52787
- OnlineSecurity    1     15021 1965549 52841
- InternetService   1     34013 1984542 52938
- Churn             1     39675 1990204 52966
- TechSupport       1    205495 2156024 53766
- DeviceProtection  1    235622 2186151 53905
- Tenure            1    422667 2373196 54726
- TotalCharges      1    475357 2425886 54946
- OnlineBackup      1    691028 2641556 55797
- Multiple          1   1268972 3219501 57776
- StreamingTV       1   1686228 3636757 58995
- StreamingMovies   1   2190914 4141443 60294
```

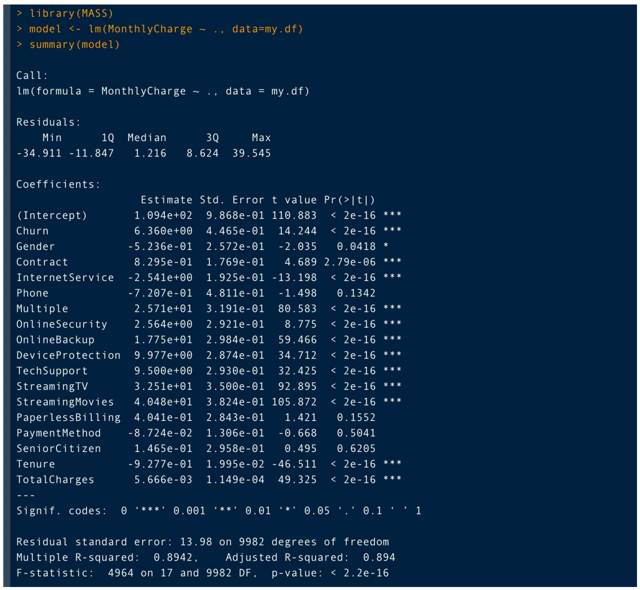


```
Step:  AIC=52765.17
MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    Tenure + TotalCharges

                 Df Sum of Sq     RSS   AIC
<none>                        1950618 52765
- PaperlessBilling  1       397 1951015 52765
- Phone             1       435 1951054 52765
- Gender            1       802 1951420 52767
- Contract          1      4289 1954907 52785
- OnlineSecurity    1     14997 1965615 52840
- InternetService   1     34044 1984662 52936
- Churn             1     39607 1990226 52964
- TechSupport       1    205558 2156176 53765
- DeviceProtection  1    235725 2186344 53904
- Tenure            1    422810 2373428 54725
- TotalCharges      1    475389 2426007 54944
- OnlineBackup      1    691039 2641658 55796
- Multiple          1   1269419 3220037 57776
- StreamingTV       1   1686613 3637232 58994
- StreamingMovies   1   2190998 4141616 60293

Call:
lm(formula = MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    Tenure + TotalCharges, data = my.df)

Coefficients:
     (Intercept)             Churn            Gender          Contract   InternetService             Phone
      109.235024          6.354559         -0.521058          0.828830         -2.540640         -0.718146
        Multiple    OnlineSecurity      OnlineBackup  DeviceProtection       TechSupport       StreamingTV
       25.718288          2.558903         17.748101          9.981544          9.502636         32.514061
 StreamingMovies  PaperlessBilling            Tenure      TotalCharges
       40.484938          0.405180         -0.927676          0.005666
```

Fig20: *Multiple Regression Test*

Based on above Chi-Squared Tests results we see very small “p-values”[5]; NULL hypothesis of independence is rejected and our original hypothesis is correct.

Regression analysis is a powerful statistical process to find the relations within a dataset, with the key focus being on relationships between the independent variables (predictors) and a dependent variable (outcome)[12]. It can be used to build models for inference or prediction.



```
> library(MASS)
> model <- lm(MonthlyCharge ~ ., data=my.df)
> summary(model)

Call:
lm(formula = MonthlyCharge ~ ., data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max
-34.911 -11.847   1.216   8.624  39.545

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.094e+02  9.868e-01 110.883  < 2e-16 ***
Churn             6.360e+00  4.465e-01  14.244  < 2e-16 ***
Gender           -5.236e-01  2.572e-01  -2.035   0.0418 *
Contract          8.295e-01  1.769e-01   4.689 2.79e-06 ***
InternetService  -2.541e+00  1.925e-01 -13.198  < 2e-16 ***
Phone            -7.207e-01  4.811e-01  -1.498   0.1342
Multiple          2.571e+01  3.191e-01  80.583  < 2e-16 ***
OnlineSecurity    2.564e+00  2.921e-01   8.775  < 2e-16 ***
OnlineBackup      1.775e+01  2.984e-01  59.466  < 2e-16 ***
DeviceProtection  9.977e+00  2.874e-01  34.712  < 2e-16 ***
TechSupport       9.500e+00  2.930e-01  32.425  < 2e-16 ***
StreamingTV       3.251e+01  3.500e-01  92.895  < 2e-16 ***
StreamingMovies   4.048e+01  3.824e-01 105.872  < 2e-16 ***
PaperlessBilling  4.041e-01  2.843e-01   1.421   0.1552
PaymentMethod    -8.724e-02  1.306e-01  -0.668   0.5041
SeniorCitizen     1.465e-01  2.958e-01   0.495   0.6205
Tenure           -9.277e-01  1.995e-02 -46.511  < 2e-16 ***
TotalCharges      5.666e-03  1.149e-04  49.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.98 on 9982 degrees of freedom
Multiple R-squared:  0.8942,	Adjusted R-squared:  0.894
F-statistic:  4964 on 17 and 9982 DF,  p-value: < 2.2e-16
```

Fig21: *Initial Multiple Regression Model*

A high value of F statistic, with a very low p-value (<2.2e-16), implies that the null hypothesis can be rejected. This means there is a potential relationship between the predictors and the outcome.

Based on above results we could now create a regression equation from this output:

MonthlyCharge *= (109.419461) + Churn * (6.359851) + Gender * (-0.523554) + Contract * (0.82947) + InternetService * (-2.540632) + Phone * (-0.720723) + Multiple * (25.714855) + OnlineSecurity * (2.563543) + OnlineBackup * (17.747445) + DeviceProtection * (9.977369) + TechSupport * (9.500246) + StreamingTV * (32.511433) + StreamingMovies * (40.481723) + PaperlessBilling * (0.404059) + PaymentMethod * (-0.087243) + SeniorCitizen * (0.1465) + Tenure * (-0.927701)*

RSE (Residual Standard Error) is the average deviation between the actual outcome and the true regression line[7]. A low value of RSE implies a low deviation of our model from the true regression line.

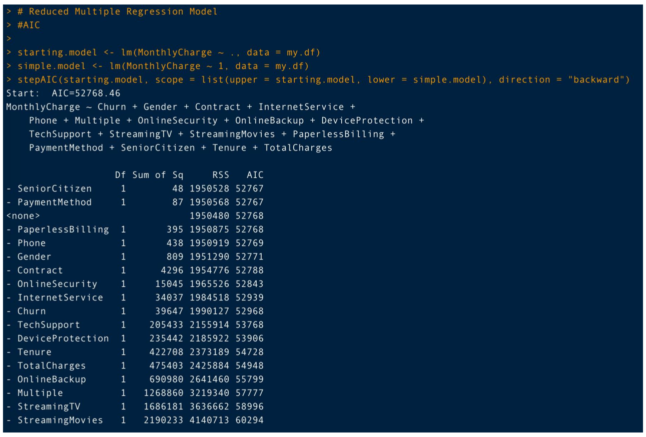
R-squared (R^2) measures the proportion of variability in the outcome that can be explained by the model and is almost always between 0 and 1; the higher the value, the better the model is able to explain the variability in the outcome[7]. However, increase in number of predictors mostly results in an increased value of R^2 due to inflation of R-squared. Adjusted R-squared adjusts the value of R^2 to avoid this effect. A high value of adjusted R^2 (0.8942) shows that more than 89% of the variance in the data is being explained by the model.

The Std. Error gives us the average amount that the estimated coefficient of a predictor differs from the actual coefficient of predictor[13]. It can be used to compute the confidence interval of an estimated coefficient.

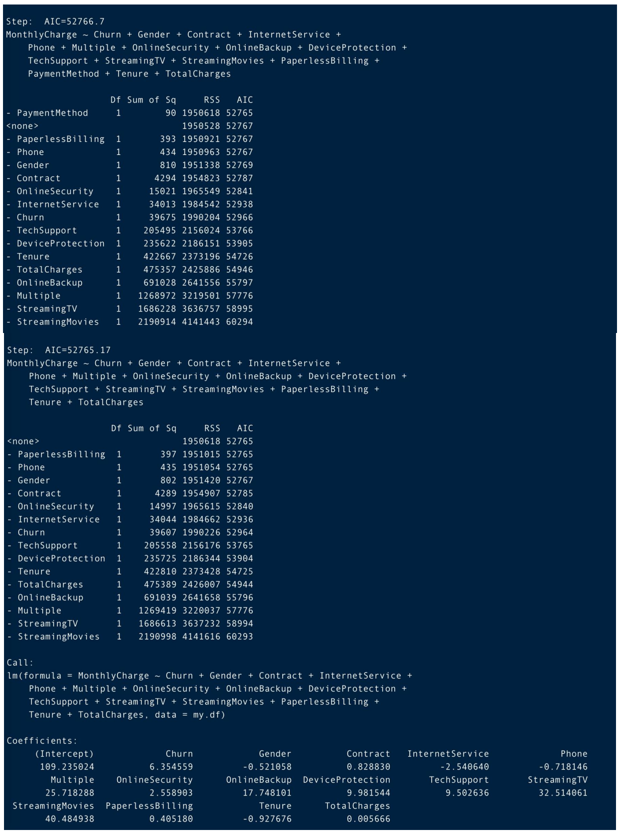
The t value of a predictor tells us how many standard deviations its estimated coefficient is away from 0. Pr (>|t|) for a predictor is the p-value for the estimated regression coefficient, which is same as saying what is the probability of seeing a t value for the regression coefficient. A very low p-value (<0.05) for a predictor can be used to infer that there is a relationship between the predictor and the outcome.

STEPAIC:

There are two ways of creating models stepwise, either forward selection or backwards elimination. In forward selection, you start with the simplest desired model and add predictors to the model until your chosen criteria indicate that adding more predictors to the model would actually worsen the model’s abilities. In backwards elimination, you start with the most complex model acceptable and remove predictors from the model until your chosen criteria indicate that removing more predictors from the model would worsen the model’s abilities.



```
> # Reduced Multiple Regression Model
> #AIC
>
> starting.model <- lm(MonthlyCharge ~ ., data = my.df)
> simple.model <- lm(MonthlyCharge ~ 1, data = my.df)
> stepAIC(starting.model, scope = list(upper = starting.model, lower = simple.model), direction = "backward")
Start:  AIC=52768.46
MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    PaymentMethod + SeniorCitizen + Tenure + TotalCharges

                   Df Sum of Sq     RSS   AIC
- SeniorCitizen     1        48 1950528 52767
- PaymentMethod     1        87 1950568 52767
<none>                          1950480 52768
- PaperlessBilling  1       395 1950875 52768
- Phone             1       438 1950919 52769
- Gender            1       809 1951290 52771
- Contract          1      4296 1954776 52788
- OnlineSecurity    1     15045 1965526 52843
- InternetService   1     34037 1984518 52939
- Churn             1     39647 1990127 52968
- TechSupport       1    205433 2155914 53768
- DeviceProtection  1    235442 2185922 53906
- Tenure            1    422708 2373189 54728
- TotalCharges      1    475403 2425884 54948
- OnlineBackup      1    690980 2641460 55799
- Multiple          1   1268860 3219340 57777
- StreamingTV       1   1686181 3636662 58996
- StreamingMovies   1   2190233 4140713 60294
```



```
Step:  AIC=52766.7
MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    PaymentMethod + Tenure + TotalCharges

                   Df Sum of Sq     RSS   AIC
- PaymentMethod     1        90 1950618 52765
<none>                          1950528 52767
- PaperlessBilling  1       393 1950921 52767
- Phone             1       434 1950963 52767
- Gender            1       810 1951338 52769
- Contract          1      4294 1954823 52787
- OnlineSecurity    1     15021 1965549 52841
- InternetService   1     34013 1984542 52938
- Churn             1     39675 1990204 52966
- TechSupport       1    205495 2156024 53766
- DeviceProtection  1    235622 2186151 53905
- Tenure            1    422667 2373196 54726
- TotalCharges      1    475357 2425886 54946
- OnlineBackup      1    691028 2641556 55797
- Multiple          1   1268972 3219501 57776
- StreamingTV       1   1686228 3636757 58995
- StreamingMovies   1   2190914 4141443 60294

Step:  AIC=52765.17
MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    Tenure + TotalCharges

                   Df Sum of Sq     RSS   AIC
<none>                          1950618 52765
- PaperlessBilling  1       397 1951015 52765
- Phone             1       435 1951054 52765
- Gender            1       802 1951420 52767
- Contract          1      4289 1954907 52785
- OnlineSecurity    1     14997 1965615 52840
- InternetService   1     34044 1984662 52936
- Churn             1     39607 1990226 52964
- TechSupport       1    205558 2156176 53765
- DeviceProtection  1    235725 2186344 53904
- Tenure            1    422810 2373428 54725
- TotalCharges      1    475389 2426007 54944
- OnlineBackup      1    691039 2641658 55796
- Multiple          1   1269419 3220037 57776
- StreamingTV       1   1686613 3637232 58994
- StreamingMovies   1   2190998 4141616 60293

Call:
lm(formula = MonthlyCharge ~ Churn + Gender + Contract + InternetService +
    Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +
    TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
    Tenure + TotalCharges, data = my.df)

Coefficients:
     (Intercept)             Churn            Gender          Contract   InternetService             Phone
      109.235024          6.354559         -0.521058          0.828830         -2.540640         -0.718146
        Multiple    OnlineSecurity      OnlineBackup  DeviceProtection       TechSupport       StreamingTV
       25.718288          2.558903         17.748101          9.981544          9.502636         32.514061
 StreamingMovies  PaperlessBilling            Tenure      TotalCharges
       40.484938          0.405180         -0.927676          0.005666
```

Fig21: *StepAIC*

We can see from the output that our final model actually contains all of the chosen variables. One thing to note about this process is that, although the two models' AIC differ by less than 10[5][7], the chosen model is the model with fewer predictor variables because of the necessary balance between accuracy and complexity that AIC uses.

Recommendation based on Multiple Regression Analysis:

Customers who have signed up recently on a month-to-month contract with a single telephone line and pay with an alternative method to electronic check are the most likely to churn. Resources should be focused on these customers to move them to products that are indicators of brand loyalty. Marketing and retention teams should prioritize the following products in descending order of importance:

- Two-year contract
- One-year contract
- Paperless billing
- Payment by electronic check
- A second telephone line

Practical Implications:

Customer churn can have a significant impact on revenue and profitability for any industry. This study focuses on telecom data. With relevant data, this study can be easily tuned to any other industry like banking, credit cards, retail, hospitality, subscription services etc. Leveraging insights from churn analysis is essential for informing strategic business decisions. By understanding the underlying factors driving churn and predicting customer behavior, industry leaders can develop targeted retention strategies, optimize marketing efforts, and improve overall business performance.

Insights from churn analysis enable leaders to develop tailored retention strategies aimed at reducing customer attrition and improving loyalty[14]. By identifying the key drivers of churn, such as service quality issues, pricing concerns, or customer dissatisfaction, companies can prioritize areas for improvement and implement targeted initiatives to address these issues. For example, if poor network coverage is identified as a significant driver of churn, telecom operators can invest in infrastructure upgrades or network optimization efforts to enhance service reliability and quality.

Moreover, churn analysis allows companies to segment customers based on their churn propensity and preferences, enabling more personalized retention efforts[7]. By understanding the unique needs and preferences of different customer segments, telecom operators can offer targeted incentives, promotions, or loyalty rewards to incentivize retention and strengthen customer relationships.

Insights from churn analysis can also inform the design and execution of marketing campaigns, helping leaders to better target and engage customers. By identifying churn triggers and predicting future churn behavior[15], companies can tailor marketing messages and promotions to resonate with at-risk customers and encourage them to stay with the brand.

For example, if a particular customer segment is found to be sensitive to pricing changes, telecom operators can design targeted pricing promotions or discounts to incentivize retention[17]. Similarly, if customer satisfaction is identified as a key driver of churn, companies can focus on highlighting service improvements or value-added features in their marketing communications to reassure customers and strengthen loyalty.

Furthermore, churn analysis enables telecom operators to measure the effectiveness of marketing campaigns in reducing churn rates and improving customer retention. By tracking key metrics such as churn reduction rates, customer engagement levels, and return on investment (ROI), companies can evaluate the impact of marketing initiatives and optimize future campaigns based on empirical evidence and insights derived from data analysis.

Insights from churn analysis can inform proactive customer service and support initiatives, enabling telecom operators to anticipate and address customer concerns before they escalate into churn events. By monitoring customer interactions, service usage patterns[16], and sentiment analysis, companies can identify early warning signs of dissatisfaction or frustration and intervene proactively to resolve issues and retain customers.

For example, if a customer experiences frequent service outages or technical problems, telecom operators can proactively reach out to offer troubleshooting assistance, service credits, or compensation to mitigate the risk of churn. Similarly, if a customer exhibits declining usage patterns or reduced engagement with the brand, companies can proactively engage with personalized communications, offers, or incentives to re-engage the customer and reinforce loyalty.

Moreover, churn analysis enables telecom operators to identify opportunities for service improvements and innovation based on customer feedback and preferences. By leveraging insights from churn analysis, companies can prioritize product enhancements, feature developments, or service offerings that address the evolving needs and expectations of customers, ultimately enhancing satisfaction and retention.

Insights from churn analysis can inform strategic resource allocation decisions, helping companies to optimize investments and resources for maximum impact on customer retention and satisfaction. By identifying the most influential drivers of churn and assessing their impact on business outcomes, companies can allocate resources more effectively to areas that yield the highest return on investment.

For example, if customer service issues are identified as a primary driver of churn, telecom operators can allocate resources towards training and development initiatives for customer support staff, enhancing service delivery and customer satisfaction. Similarly, if network performance issues are found to be driving churn, companies can prioritize investments in network infrastructure upgrades or technology investments to improve service quality and reliability.

Furthermore, churn analysis enables companies to optimize resource allocation across different customer segments based on their churn propensity and lifetime value. By segmenting customers according to their churn risk and profitability, companies can tailor resource allocation strategies to focus on high-value customers with the greatest potential for long-term retention and revenue generation.

Insights from churn analysis can also inform competitive differentiation and market positioning strategies, helping companies operators to differentiate their offerings and strengthen their competitive advantage in the marketplace. By understanding the unique needs and preferences of customers, companies can develop targeted value propositions and positioning statements that resonate with their target audience and set them apart from competitors[18].

For example, if customer satisfaction and service quality are identified as key drivers of churn, telecom operators can differentiate themselves by emphasizing their commitment to customer service excellence and reliability in their marketing communications and brand messaging. Similarly, if pricing and value are important considerations for customers, companies can position themselves as offering superior value for money or innovative pricing plans that meet the needs of price-sensitive consumers.

Moreover, churn analysis enables companies to monitor competitor activities and market trends, allowing them to adapt their strategies and offerings in response to changing market dynamics[19]. By tracking competitor churn rates, market share, and customer satisfaction levels, companies can identify opportunities for differentiation and innovation, ultimately strengthening their market position and competitive advantage.

Advancements in Analytics and Machine Learning Techniques

Advancements in analytics and machine learning techniques have revolutionized churn management in the telecommunications industry, enabling telecom operators to develop more accurate, scalable, and actionable strategies for customer retention. Some of the key advancements in analytics and machine learning techniques for churn management include:

Deep Learning and Neural Networks: Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), offer advanced capabilities for feature extraction, pattern recognition, and sequence modeling, allowing for more accurate and nuanced churn predictions [20].

Ensemble Learning and Model Stacking: Ensemble learning methods, such as random forests, gradient boosting, and model stacking, combine multiple base learners to improve prediction accuracy, robustness, and generalization performance, resulting in more reliable and stable churn prediction models.

Reinforcement Learning and Adaptive Strategies: Reinforcement learning algorithms enable telecom operators to develop adaptive retention strategies that learn and evolve over time based on feedback and performance data, allowing for dynamic and responsive churn management in complex and dynamic environments.

Anomaly Detection and Outlier Analysis: Anomaly detection techniques, such as isolation forests, autoencoders, and one-class SVMs, enable telecom operators to identify unusual or unexpected patterns in customer behavior that may indicate potential churn events or anomalies, allowing for proactive intervention and risk mitigation.

Natural Language Processing (NLP) and Sentiment Analysis: Natural language processing (NLP) and sentiment analysis techniques enable telecom operators to analyze unstructured text data from customer feedback, social media, and other sources to extract insights, sentiments, and opinions related to churn behavior, allowing for more comprehensive and holistic churn prediction and management.

Application Across Other Industries

The methodologies and insights derived from this study are not confined to the telecommunications sector but are applicable across various industries facing similar challenges with customer retention:

Retail and E-commerce: Retailers can use churn prediction models to identify at-risk customers and develop personalized marketing strategies to increase retention rates, thus maximizing revenue and enhancing customer service.

Banking and Financial Services: Financial institutions can apply similar predictive models to anticipate customer defections to competitors and offer timely incentives, improving customer satisfaction and loyalty.

Healthcare: Hospitals and health insurance companies can use churn analytics to improve patient retention and satisfaction, which is crucial for maintaining revenue streams and funding in a competitive healthcare market.

Utilities and Energy: Utility companies can predict and reduce customer turnover, which is essential for planning infrastructure investments and managing demand across vast energy grids.

Subscription Services: From media to software, companies that operate on a subscription basis can leverage churn prediction to tailor their offerings to customer preferences and reduce subscription cancellations.

CONCLUSION

The examination of predictive modeling for customer churn presented in this paper not only underscores the importance of customer retention but also highlights the transformative potential of machine learning technologies in bolstering economic stability and fostering sector-wide innovation. As the telecommunications industry is pivotal to the infrastructure of The United States, enhancing churn prediction capabilities can yield significant economic benefits, including stabilizing revenue streams for companies, creating job opportunities, and enhancing market competitiveness.

The ability to effectively predict and mitigate customer churn can lead to substantial cost savings for telecommunications companies through reduced customer acquisition costs and increased customer lifetime value. This, in turn, drives profitability and supports a healthy competitive market environment. By ensuring that companies can retain customers more efficiently, resources can be allocated toward innovation and service improvement, which are crucial for staying competitive in the global market.

Every company who is foundational to the operation of critical sectors including healthcare, education, and government, telecommunication; providing essential services that support societal functions can benefit from this analysis. Improvements in this sector translate into enhanced efficiency and reliability across these vital areas, indirectly supporting an array of industries critical to the economic and social fabric of The United States.

REFERENCES

[1] Deciphering the Dynamics of Hospital Readmissions Patterns Using Supervised Machine Learning doi: 10.5281/zenodo.10702606

[2] Impact of Machine Learning on Healthcare Analytics
doi: 10.21275/SR24210203022

[3] U.S. Department of Health and Human Services (DHHS). (2010,

[4] Wang, F., & Head, M. (2007). Application of machine learning in customer churn prediction. Proceedings of the 2007 ACM symposium on Applied computing, 219-223. doi: 10.1145/1244002.1244056

[5] Unveiling the Potential of Generative AI in Revolutionizing Healthcare doi: 10.21275/SR24307081508

[6] Verbeke, W., Dejaeger, K., Martens, D., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229. doi: 10.1016/j.ejor.2011.10.025

[7] Role of Big Data in Revolutionizing Health Management Systems DOI: 10.5281/zenodo.10702606

[8] Zhang, Z., & Wang, Z. (2016). Customer churn prediction in telecommunications using machine learning in big data platform. Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), 3423-3428. doi: 10.1109/BigData.2016.7841070

[9] Telco Customer Churn
https://www.kaggle.com/datasets/blastchar/telco-customer-churn
[Accessed: Feb. 06, 2020].

[10] Radcliffe, N. J., & Surry, P. D. (1999). Differential gain: A new measure of prediction accuracy for rarely occurring ordered events. International Journal of Forecasting, 15(3), 227-236. doi: 10.1016/S0169-2070(99)00018-5

[11] Chaturvedi, A., & Dey, L. (2020). A comparative analysis of machine learning algorithms for customer churn prediction in telecom industry. Wireless Personal Communications, 112(4), 2807-2830. doi: 10.1007/s11277-020-07018-2

[12] Nguyen, V. A., Dinh, D. H., Nguyen, V. D., & Nguyen, D. T. (2020). Customer churn prediction using machine learning techniques: A case of telecommunication companies in Vietnam. Proceedings of the 2020 IEEE 10th International Conference on Communication Systems & Networks (ComSNetS), 124-130. doi: 10.1109/ComSNetS48635.2020.9025017

[13] Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing, 12(1), 17-30. doi: 10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR2>3.0.CO;2-Q

[14] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: A systematic review and agenda for future research. Electronic Markets, 26(2), 173-194. doi: 10.1007/s12525-016-0219-0

[15] Kadiyala, S. R., Chen, Z., & Jindal, P. (2016). Predicting customer churn using machine learning techniques in the insurance industry. Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 268-277. doi: 10.1109/DSAA.2016.51

[16] Kim, Y., Choi, Y., Jeong, C., & Han, H. (2007). A hybrid approach to customer churn prediction in the airline industry. Expert Systems with Applications, 32(2), 705-714. doi: 10.1016/j.eswa.2006.01.027

[17] Liu, C., Li, C., & Li, S. (2015). Big data: Applications and opportunities in finance research. International Review of Financial Analysis, 40, 186-190. doi: 10.1016/j.irfa.2015.02.016

[18] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. doi: 10.1016/j.dss.2014.03.001

[19] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS Quarterly, 35(3), 553-572.

[20] Thomas, R. K., & Krishnan, R. (2017). Does Big Data Analytics contribute to customer relationship management performance? A meta-analysis. International Journal of Information Management, 37(6), 1444-1454. doi: 10.1016/j.ijinfomgt.2017.08.003