Turkish Online Journal of Qualitative Inquiry (TOJQI)

Volume 12, Issue 3, May 2021: 4666-4681

Research Article

CUSTOMER CHURN PREDICTION IN TELECOM INDUSTRY USING DWHBI APPROACHES AND R PROGRAMMING

K. Prakash Krishnan[1], Dr.A.Kumar Kombaiya[2]

ABSTRACT

In recent days, telecom industry plays a vital role in our daily life. During corona pandemic time entire worlddependson thetelecom services domain. But telecome industry has been facing many surivival problems in the globalmarketsince last 10 years due to heavy competition between competitors. To stand in this field, service providers have to understand the complete customer requirements and provide the efficient services to stop the customer movement from one network to another network. Customer churn is one of the most critical problem faced by the telecom industry. In this industry, it is more expensive to bring the new customers as compared to retain the existing customers. The objective of customer churn prediction is to find the subscribers that are ready to move from the currentservice provider in advance. Churn prediction allows the service providers to offer new benefits and campaign offers to retain the existing customer in the same network. Technically this term would be call it as 'Win back Situation' in telecom industry. The high volume of data generated by telecom industry , with the help of datawarehosuing and business intelligence implementationwould become the main asset for predicting the customer churn. To prevent the churn many models and methods are used by researchers.

This research paper is using data ware housing business intelligence method, Oracle SQL developer and R programming to predict the churn result.

DWHBI model used to get the historical and current data information based on the mapping transformation logics.

Oracle SQL tool has represented to get the consistent data set from various internal & external source files by using optimizational SQL queries.

The R Tool help us to process the large level dataset churn in form of graphics, chart and different unique visualizations. R is the most powerful statistical programming tool which can represent the dataset in any kind of digital format. It also have different packages to predict the result.

[1]Ph.DScholar,DepartmentofComputer science, ChikkannaGovtArtsCollege,Tirupur.
[2]AssistantProfessor, Department of Computer Science, ChikkannaGovtArtsCollege,Tirupur.

Index Terms → Churn, R Tool, Oracle SQL developer, Telecom,Optimization SQL query

1. INTRODUCTION

India is the world's 2nd largest telecommunication market. The telecom market can be divided into 3 categories.

1) Wire-less
2) Wire-line
3) Internet Services

The total number of telephone subscriber in India reached 1.18 billion as of 31 March 2020. As per TRAI the number of wireless subscribers is over 1.17 billion and the number of wireline subscribers are 20.58 Million.

As of 2019, India holds the highest data usage per smartphone average 9.8 GB per month. It is expected to double by 2024 as 18 GB.

Customer churn is a term, which is used to mention the customer movement from one service provider to another service provider. Literature discovers following type of customer churns.

A) Volunteer churn - Customer quit the contract and move to the next service provider
B) Non – Volunteer churn - Company quit the service to a customer
C) Rotational Churn - Customer can terminate the contract without any prior knowledge of both parties (Customer and Company). Sometime it call it as Silent churn.

1.1 REASONS FOR CUSTOMER CHURN:

The first two type [1.A] [1.B] of churn can be predicted easily with the help of customer usages and log histories but third type [1.C] of churn is quit difficult to predict because such type of customers can terminate the contract at any time in near future.Below are some reasons for the customer churn.

- Poor customer service
- Network coverage
- In sufficient network towers
- Hidden charges and Rates
- Unwanted Spam messages
- Weak customer relationship
- Ignoring customer complaints
- Over due SLA

The main objective of customer churn is used to predict the customers with high tendency to leave a company. In order to retain the existing customers , The telecom industry

needs to find out the reason for churn, which can be extracted based on customer usages and customer behavioralstatistics. Customer demographic attributes (age, sex, education, status, location) are used for subscriber churn calculation.It will be used to predict the prompt accuracy of newly developed model.

2. LITERATURE SURVEY

Bo Jin, Jia shi [1]proposed a new set of approach called PBCCP neural network algorithm which executes particle classification optimization (PCO) and particle fitness calculation (PFC). PCO classifies the categories according to their fitness but PFC calculates the fitness values of each forwarding training process of Back Propagation NN (Supervised Learning Algorithm). Here gradient descent is used to minimize the error function. The actual performance of BP on a specific problem is dependent on this input data. This experiment demand us to do multiple time until we reach the desired result.

$$Wx = Wx – a (\eth \text{ error} / \eth Wx)$$

Asif & Kadan[2]discovered the approach based on machine learning and big data platform. In order to measure the performance they used Area Under Curve (AUC) and Social Network Analysis (SNA). Here SNA has given the significant performance compared to AUC. This model was developed and tested through hadoop framework environment which handles large set of dataset. This model experimented with 4 Algorithms: Decision Tree, Random Forest, Gradient Boosted Machine (GBM), Extreme Gradient Boosting (XGBOOST). Here the data model needs to be updated each time. They concluded XGBOOST gives the best result with 93% accuracy.

Pamina J, Beschi Raja, Sathya Bama [3] divulged the method which based on machine learning process and the objective of this churn prediction is to perform binary classification with subscriber records and to figure out who are likely to cancel the contract in near future. Performance of the model evaluated by various measures like Accuracy, Recall, Precision, F-Score, ROC & AUC. This model experimented with KNN model, Decision Tree, Random Forest model, NN model. After experimental they found decision tree is the best model for real time business problem. Here they additionallyPCA model for preprocessing and reduce the dimension of the data which is important from the whole dataset.

Adnan Amin, Sajid Anwar [4] proposed and intelligent rule-based decision making method based on Rough set theory to remove the despensable attribute from the given dataset. RST approach applied on different rules generation algorithms (EA, GA, CA and LA). Here they discovered genetic algorithm dominated well than other rules generation algorithms in terms ofprecision, recall, the rate ofmisclassification, lift, coverage, accuracy, and F-measure. Below RST concepts are involved to reduce the data set

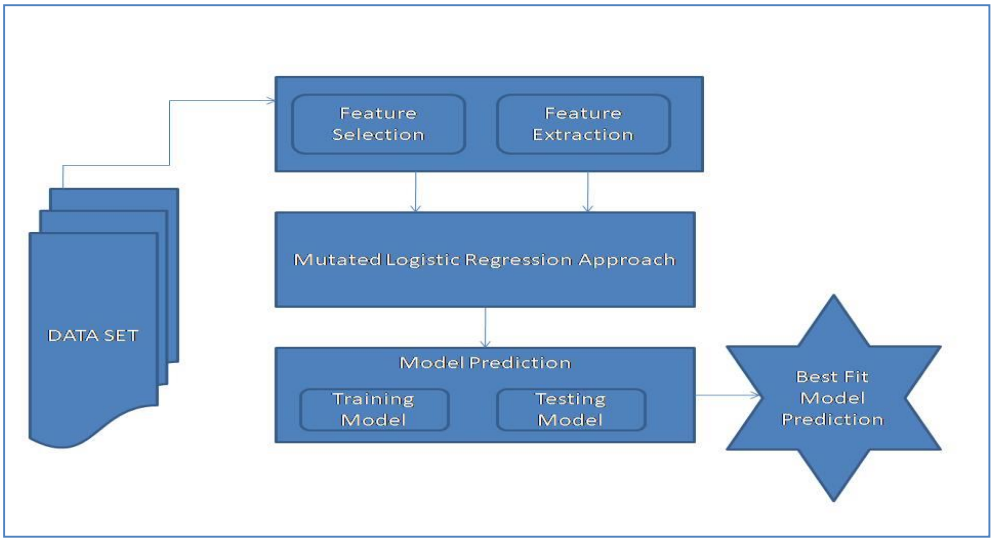
a. Lower Approximation
b. Upper Approximation
c. Boundary Region
d. Outside Region
e. Reduction core

Maurus Riedweg , Pavol Svaba and Gwendolin Wilke [6] discovered the semi auto predictive analytics and campaign management. They have chosen large data set from switzerland telecom company and they perfomed the calculation based on Naïve bayes prediction model and corresponding nomogram values and this prediction accuracy exceeded decision tree approach as well as the currently used bench mark.

V.Umayaparvathi , K.Iyakutti [7] have divulged with general churn prediction models as data collection, preparation, classification and prediction. They have analyzed with existing clustering and classification methods are used to distinguishing churners and several standard performance metrics are proposed for predicting and analyzing the performance model.

Tools / Technique Used	**Purpose**
DWHBI	1)To mainitan Current & Historical information by using various mapping transformation logics 2) Staging the data from various internal & external sources 3) We can use structured , semi structred & un-structured data 4) Data scrubbing , Aggreation , Data transimission
Oracle SQL developer	1) Convert the file data into Tabular format 2) Easy to create & handle the consistent data set from in consistent records 3) Prepare the selective column dataset from various tables. 4) It wont take much time to fix the errorness records in the table 5) We can effortlessly import/export columns in the dataset as per our requirement based on join conditions
R Programminng Tool	1) It can clean, analyze & graph our dataset 2) To find the best fit model for our problem 3) R programming language used for stastical computing 4) Better method to create the effective Dashboard. 5) Large scale , Robust-code , maintainable 6) Deployment & Reproduicbility 7) Running code without compiler

3. PROPOSED METHODOLOGY



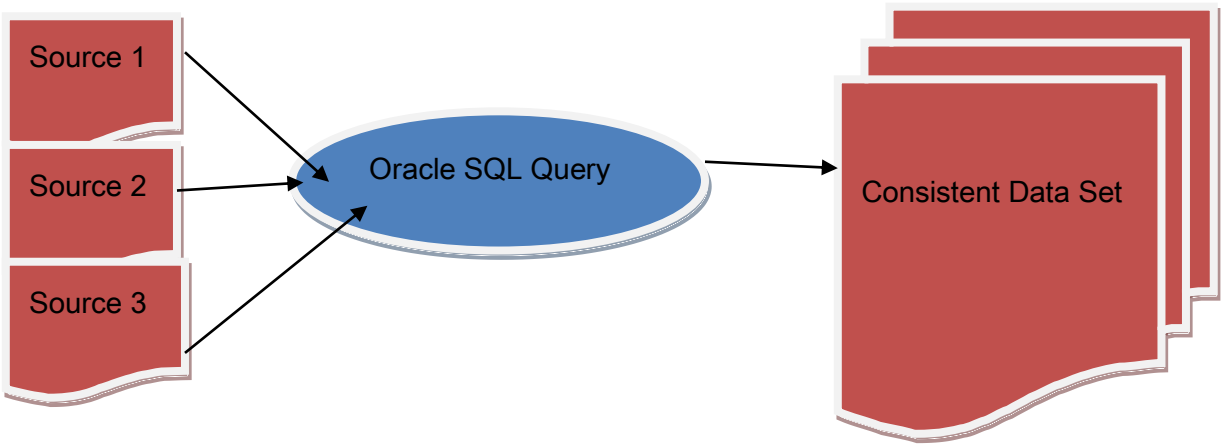
The above diagram represent the proposed MLR (Mutated Logistics Regression) , basically the digram can be splitted into 5 parts as below

Part 1 : Represent the Data
Part 2: Selection & Extraction of Particular atributes
Part 3: Applied the alogirthm approach on the data set
Part 4: Prepare the Test & Training model from the given data set
Part 5: Estimated the churn data.

4. CONSISTENT Vs INCONSISTENT

In Telecom industry normally deals with vast amount of data. Some time it demand us to do numerous operations on the data set to make it as consistent. We can do all the operations in R programming but it is not a time consuming process. It would required longer time than actual.

In order to save our time, we can use Oracle SQL developer tool to fix this issues in the table by using strucred query language. It will help us to get the consistent dataset which we can use directly in the R programming tool without using any modifications. Some time we need such preventive mechanism to complete our work in a quick turn around time.



4.1 POTENTIAL OPERATIONS

- To create single data set from Multiple tables → Using Join condition
- Predict & Replace Null values → Using NVL , NVL2 function
- Eliminate un wanted columns → Using Select operations
- Find and Fix the numerical issues → Using Mathematical functions
- Convert the factorial values → Using Case (or) decode operations
- Grouping the values → Using Group function
- Combine the one or more columns → Using concat opertions
- Fix the Date & time stamp issues → Using To_date() function

We can execute a single optimization query to fix all the aforemnetioned process. It will help us to get the proper consistent data set which is ready to use for the next step.

5. GROUND WORK – MUTATED LOGISTICS REGRESSION APPROACH

In this section, the proposed Mutated LR (Logistics Regression) method is utilized. Many conventional algorithm such as Decision Tree, Random Forest, Naïve bayes and classification tree has been proposed. These above mentioned algorithm is capable of predicting the churn rate. But they have several issues such as Naïve bayes which can converges quicker but it has higher error than logical regression. It would be good for small data sets but telecome industry is vast. Similary, in case of decision tree it built on entire dataset using all variables and it can drastically reduce the performance.

Like wise in Random forest , it may change considerably by a smlla change in data and RF algorithms compuations may got far more complex compared to other algorithms. It can easily occur overfitting and it need to choose the number of trees. Hence the above discussion can be achieved that concluded LR method and it do not produce the efficient result when we have huge datasets. It would produce the bad unbalance out come. Our proposed MLR (Mutated Logistics Regression) method performs much better in various parameter such as accuracy , roubustness and others. In this method we will be going to add some new variable which reduce the null deviance and residual deviance.It mostly used for binary output such as Yes (or) No, 0 (or) 1 , Right (or) Wrong. It can help in reducing problems of classification. Proposed method helps in focusing estimation of consumer churn rate in telecom services.

6. DATA SET EXTRACTED

The Attributes in our data are taken from Astro Telecom database. Here some of the prime table names are listed as below

Subscriber Table:

Subscriber	**DataType**
SubscriberNumber	varchar2(100)
FullName	varchar2(100)
Gencode	varchar2(10)
SeniorCitizen	int

AddOnServices Table:

Partner	varchar2(10)
Dependents	varchar2(10)
tenure	int
Period	varchar2(30)
AreaCode	int
MobileNumber	number
Churn	varchar2(10)
OnlineBackup	varchar2(10)

AddOnServices	**DataType**
SubscriberNumber	varchar2(100)
MultipleLines	varchar2(30)
NetService	varchar2(30)
OnlineSecurity	varchar2(30)
OnlineSupport	varchar2(30)
InterNationalPlan	varchar2(10)
InterNationalSMS	varchar2(10)
VoiceMail	varchar2(10)
IPTV	varchar2(10)
OTT	varchar2(10)
DeviceSecurity	varchar2(10)
PhoneService	varchar2(10)

Payment Process Table:

PaymentProcess	**Datatype**
SubscriberNumber	varchar2(100)
ModeOfPayment	varchar2(100)
ChargesPerMonth	decimal(38,1)
CustomerCare	varchar2(30)
GreenSource	varchar2(10)
TotalCharges	decimal(38,1)

SQL Query:

```
select
SU.SubscriberNumber, SU.Gencode, SU.SeniorCitizen,
SU.Partner, SU.Dependents,SU.tenure, AD.PhoneService,

case when AD.MultipleLines ='No phone service'
then 'No'
else AD.MultipleLines End as MultipleLines,

AD.NetService, AD.NetService,AD.OnlineSecurity, SU.OnlineBackup,
AD.OnlineSupport, SU.MobileNumber, AD.VoiceMail,
AD.InterNationalPlan, AD.InterNationalSMS, SU.AreaCode,

case when AD.DeviceSecurity = 'No internet service'
then 'No'
else AD.DeviceSecurity End as DeviceSecurity,

PP.CustomerCare,AD.IPTV,

case when AD.OTT = 'No internet service'
```

```
then 'No'
else AD.OTT End as OTT,

SU.Period, PP.GreenSource, PP.ModeOfPayment,
NVL(PP.ChargesPerMonth,0)ChargesPerMonth,
NVL(PP.TotalCharges,0)TotalCharges,SU.Churn

from Subscriber SU
Left join AddOnServices AD on SU.SubscriberNumber = AD.SubscriberNumber
Left join PaymentProcess PP on SU.SubscriberNumber = PP.SubscriberNumber;
```

Above Optimized query can used to fetch the consistent dataset and this query process will help us to get the file extracts as per our requirement. We can do any number of operations on the dataset and it all can achieved easily by wrtiting optimized query language.

Table – 1 : Dataset Description

```
'data.frame':   7050 obs. of  27 variables:
$ SubscriberNumber : Factor w/ 7050 levels "0002-ORFBO","0003-MKNFE",..: 5380 3964 2566 5541 6517 6557 1003 4772 5610 4536 ...
$ Gencode          : 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 ...
$ NetService       : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
$ OnlineSecurity   : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
$ OnlineBackup     : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
$ OnlineSupport    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
$ MobileNumber     : num  7.37e+09 7.37e+09 7.37e+09 7.37e+09 7.37e+09 ...
$ VoiceMail        : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
$ InterNationalPlan: Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
$ InterNationalSMS : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
$ AreaCode         : int  408 415 415 415 510 408 415 415 408 415 ...
$ DeviceSecurity   : Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
$ CustomerCare     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
$ IPTV             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
$ OTT              : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
$ Period           : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
$ GreenSource      : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
$ ModeOfPayment    : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
$ ChargesPerMonth  : 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 ...
```

Library & Package Used:

i) Correlationfunnel:

It can speed up the Exploratory Data Analysis (EDA) Process and it can be an effective time consuming process.

Using coorelation we can determine the good features prior to spending siginificant time developing machine learning models

It will give the insight of the business in faster manner

ii) dplyr:

It is the most powerful package to transform the summarized data into tabular format.

It used to perform common data manipulation operation such as re-ordering rows, selecting specific columns, adding new columns and summarizing data.

```
library(dplyr)
library(correlationfunnel)
data<-read.csv("D:\\Source Feed Files\\Astro Telecom Dataset.csv",header=T)
data %>% glimpse()
data_binarized_tbl <- data %>%
  select(-SubscriberNumber,-MobileNumber) %>%
  mutate(ChargesPerMonth=ifelse(is.na(ChargesPerMonth),200,ChargesPerMonth)) %>%
  mutate(TotalCharges =ifelse(is.na(TotalCharges),ChargesPerMonth, TotalCharges)) %>%
  binarize(n_bin = 5, thresh_infreq = 0.01, name_infreq = "-OTHER", one_hot = TRUE)
data_corr_tbl <- data_binarized_tbl %>%
  correlate(Churn__Yes)
data_corr_tbl %>%
  plot_correlation_funnel()
```

Explanation:

Below attributes are correlated with churn

- Month to Month Period
- Electronic Mode of payment
- Fiber Optic Internet service
- No device security
- No online security

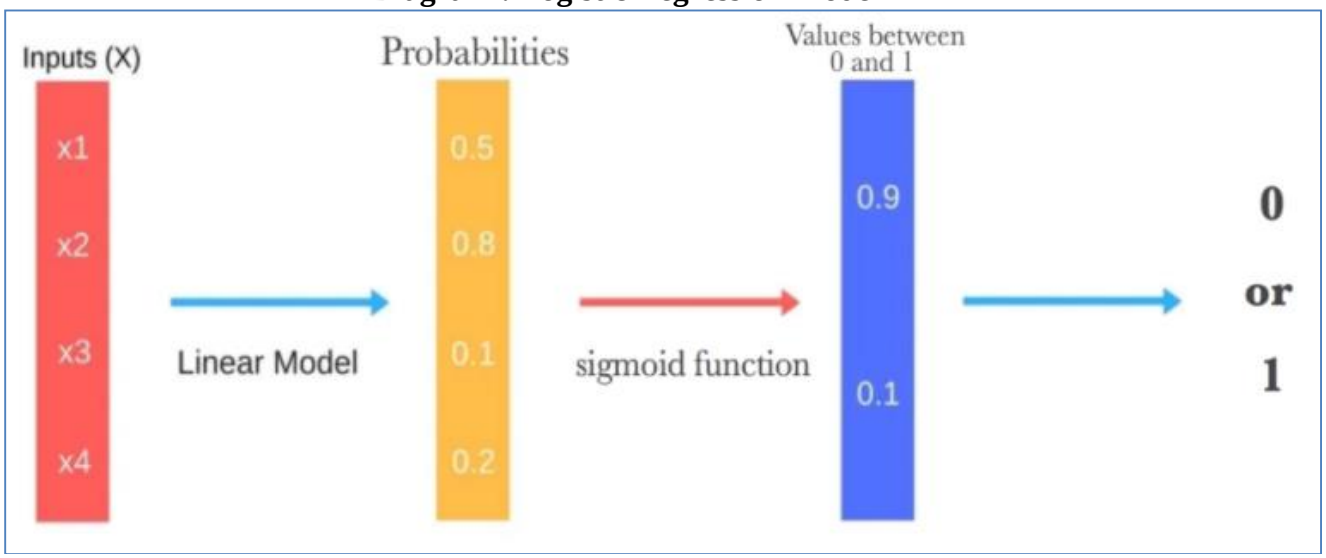
Below attributes are not correlated with churn

- Period with Two year contract type
- Tenure more than 5 years
- Payment Mode – Automatic Credit Card
- DSL Internet Service

Based on the above insights we need to promote the offers to customers to make them stay back in the same network.

Logistics Regression Method:

Diagram : Logistic Regression Model



- ✓ Logistic Regression comed under supervised learning Method
- ✓ It predicts the output of a categorical dependent variable
- ✓ Logistic Regrssion is a specific set of generalized linear models
- ✓ Types : Binary (0/1), Multi (a,b,c) , Ordinal (Low, Medium, High)
- ✓ The key representation of logistics regression are the coefficients

Steps in Logistics Regression:

1) Dataset Preparation
2) Bifurcate Test & Training dataset
3) Assumes no error in the output variable , consider removing outliers
4) Removing correlated variables from the dataset
5) Fitting the model to the training set
6) Predict the test outcome

7) Test accuracy of the result
8) Visualizing the output

iii) ggplot2 :

it is dedicatedly used for data visualization . it improve the quality of graphics. It also used to create complex plots from the dataset.

iv) GGally:

it is extension of ggplot2 and it used to reduce the complexity of combining geometric objects with transformed data.

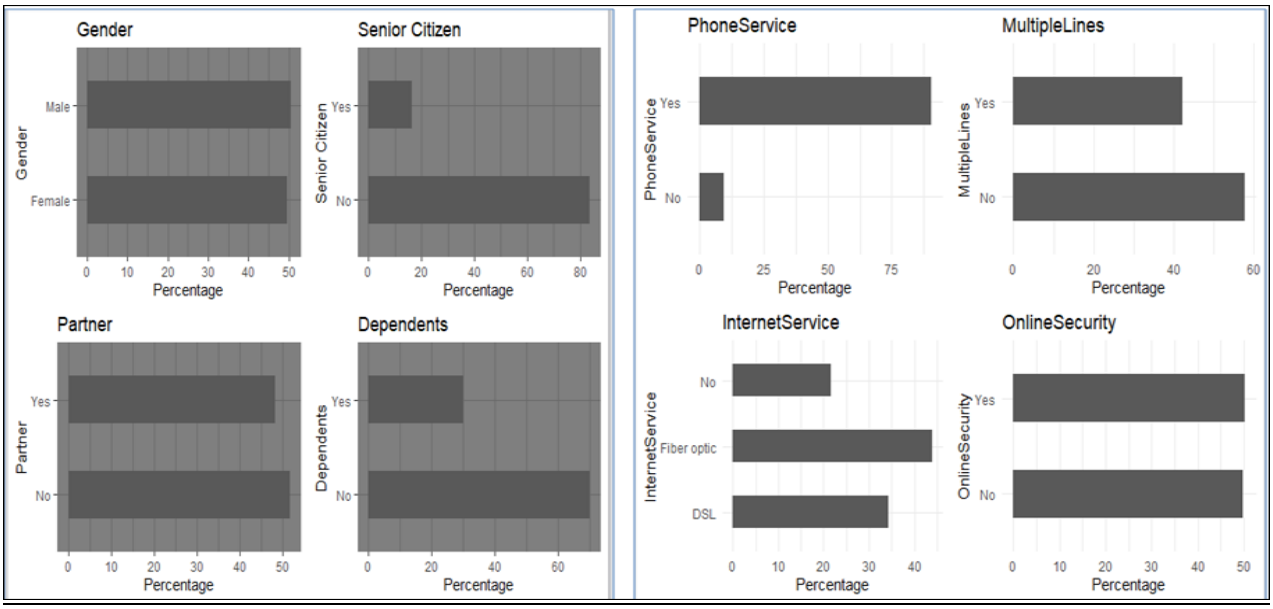
7) EXPERIMENTAL RESULTS & OBSERVATIONS:

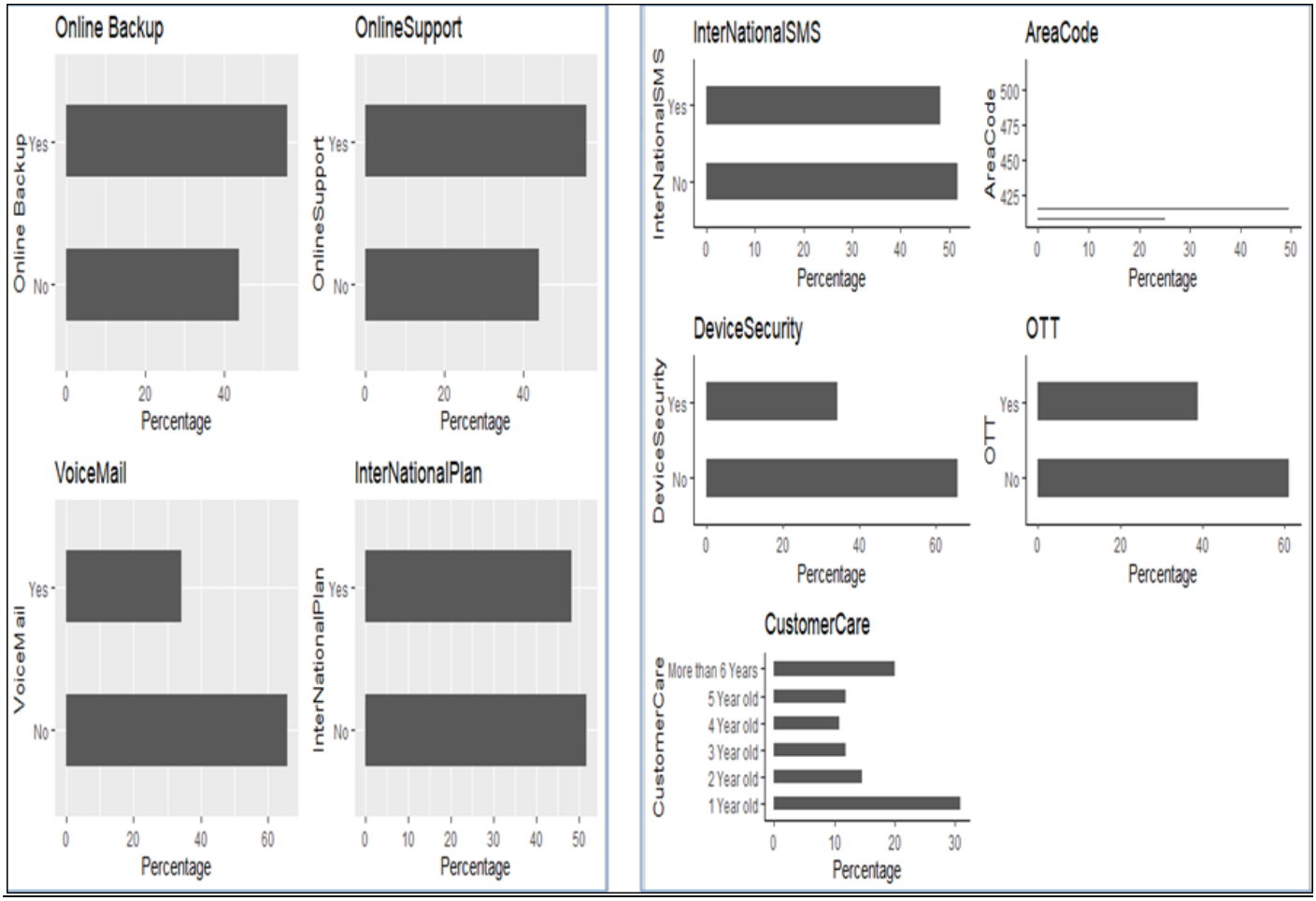
7.1 Reading Data Set churn from the file:

```
data<-read.csv("D:\\Source Feed Files\\Astro Telecom Dataset.csv",header=T)
```

7.2Plots for categorical values:

```
p1 <- ggplot(data1, aes(x=Gencode)) + ggtitle("Gender") + xlab("Gender") +
 geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
coord_flip() + theme_dark()
p2 <- ggplot(data1, aes(x=SeniorCitizen)) + ggtitle("Senior Citizen") + xlab("Senior Citizen") +
 geom_bar(aes(y = 100*(..count..)/sum(..count..)),width = 0.5) + ylab("Percentage") +
coord_flip() + theme_dark()
p3 <- ggplot(data1, aes(x=Partner)) + ggtitle("Partner") + xlab("Partner") +
 geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
coord_flip() + theme_dark()
p4 <- ggplot(data1, aes(x=Dependents)) + ggtitle("Dependents") + xlab("Dependents") +
 geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
coord_flip() + theme_dark()
grid.arrange(p1,p2,p3,p4,ncol=2)
```





Logisitics Regression Model:

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Formula – A symbolic description of the model to be fitted
Family – A description of the error distribution and link function to be used in the model
Data – A data frame containing variable of the model

LogModel <- glm(Churn ~ .,family=binomial(link="logit"),data=training)

```
Coefficients: (3 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.019294   0.635182  -0.030 0.975768
GencodeMale                      0.006324   0.075151   0.084 0.932939
SeniorCitizenYes                 0.158751   0.097154   1.634 0.102258
PartnerYes                      -0.018286   0.089356  -0.205 0.837851
DependentsYes                   -0.138211   0.103205  -1.339 0.180510
PhoneServiceYes                 -0.124091   0.501061  -0.248 0.804400
MultipleLinesYes                 0.386750   0.151622   2.551 0.010749 *
NetServiceFiber optic            1.170822   0.613329   1.909 0.056267 .
NetServiceNo                    -0.520022   0.982596  -0.529 0.596644
OnlineSecurityYes               -0.297956   0.154885  -1.924 0.054390 .
OnlineBackupYes                 -0.162753   0.148206  -1.098 0.272137
OnlineSupportYes                       NA         NA      NA       NA
VoiceMailYes                    -0.271926   1.357081  -0.200 0.841187
InterNationalPlanYes                   NA         NA      NA       NA
InterNationalSMSYes                    NA         NA      NA       NA
DeviceSecurityYes                0.308911   1.361901   0.227 0.820561
OTTYes                           0.369507   0.256761   1.439 0.150120
CustomerCareYes                 -0.214517   0.155373  -1.381 0.167384
IPTVYes                          0.333240   0.257023   1.297 0.194791
PeriodOne year                  -0.873888   0.125853  -6.944 3.82e-12 ***
PeriodTwo year                  -1.667027   0.207300  -8.042 8.87e-16 ***
GreenSourceYes                   0.313497   0.086691   3.616 0.000299 ***
ModeOfPaymentCredit card (automatic) -0.175871   0.132859  -1.324 0.185588
ModeOfPaymentElectronic check    0.300644   0.108799   2.763 0.005722 **
ModeOfPaymentMailed check        0.030045   0.131468   0.229 0.819234
ChargesPerMonth                 -0.010994   0.024132  -0.456 0.648695
tenure_group2 Year old          -0.999890   0.113085  -8.842  < 2e-16 ***
tenure_group3 Year old          -1.270344   0.133734  -9.499  < 2e-16 ***
tenure_group4 Year old          -1.144559   0.150813  -7.589 3.22e-14 ***
tenure_group5 Year old          -1.380611   0.164874  -8.374  < 2e-16 ***
tenure_groupMore than 6 Years   -1.782152   0.197407  -9.028  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6118.7  on 5287  degrees of freedom
Residual deviance: 4377.6  on 5260  degrees of freedom
AIC: 4433.6

Number of Fisher Scoring iterations: 6
```

Watched Items:
It wont be a fit model when **Null deviance < Residual deviance**
It wont be a fit model when the convergence requires many fisher’s scoring iterations
It wont be a fit model when the coefficients are larger size with significally standard errors

Residual → It is the deviation of an outcom from the predicted mean value.

Logit Function:

Logit Function is used as a Link function in a binomial distribution.

Logit(p(y=1/x))=log(p/1-p)=log(p)-log(1-p)

Logistic Equation:

Logistics Regression achive by extracting the log odds of the event ln[p/1-p]. Where P is the probability of event. So P always between 0 and 1. Mathematical equation as below

y=1/(1+e^-(a+b1x1+b2x2+b3x3+...))

y is the response variable , x is the predictor variable , a & b are coefficient

Formula Used:

Null Deviance = 2(LL(Saturated Model) - LL(Null Model)) on df = df_Sat - df_Null

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) df = df_Sat - df_Proposed

ANOVA MODEL

Chisq → It is a statistical model which used to find the significant correlation between two table

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

O – Observed Value ; e – Expected Value

Null Hypothesis(H0) → The Row and Column variables are independent

Alternative Hypothesis(H1) → The Row and Column variable are dependent

$$e = \frac{row.\,sum * col.\,sum}{grand.\,total}$$

```
anova(LogModel, test = "Chisq")
testing$Churn <- as.character(testing$Churn)
testing$Churn[testing$Churn=="No"] <- "0"
testing$Churn[testing$Churn=="Yes"] <- "1"

fitted.results <- predict(LogModel,newdata=testing,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != testing$Churn)
print(paste('Logistic Regression Accuracy',1-misClasificError))

table(testing$Churn, fitted.results > 0.5)
```

```
   FALSE TRUE
0   1180  115
1    211  256
```

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N}$$

```
> (1180+256) / ((1180+256) + (115+211))
[1] 0.814983
>
```

"Logistic Regression Accuracy 0.814982973893303"

CLASSIFICATION ACCURACY

Algorihtm	Classification Accuracy
Random Forest	78.18
Decision Tree	79.10
Logistics Regression	81.40

4. CONCLUSION

The proposed research has used datawarehousing technique , Optimizational SQL query and R package to predict the customer churn result from the company managerialtesting environment dataset Astro telecom.csv. It has evaluated, the number of churns using the classification technique Logistic Regression. The R tool has processed the large dataset churn in form of graphs which depicts the outcomes vividly and in a unique pattern visualization manner. The Churn Factor is used in many functions to depict the various areas or scenarios when the churn rate is high. Here we used optimizational sql query to get the consitent extract where we complete most of the preprocessing models such as data scrubbing, aggregarion, data integration ,data masking and duplicate elimination. In this research work we developed MLR algorithm which provides high classification accuracy , better AUC model. MLR simply outperforms the other existing methology of churn prediction including the standard RF model.

REFERENCES

[1] Yu, R., An, X., Jin, B. et al. Particle classification optimization-based BP network for telecommunication customer churn prediction. Neural Comput & Applic 29, 707–720 (2018)

[2] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using

machine learning in big data platform. J Big Data 6, 28 (2019)

[3] Pamina J., Beschi Raja J., Sam Peter S., Soundarya S., Sathya Bama S., Sruthi M.S. (2020) Inferring Machine Learning Based Parameter Estimation for Telecom Churn Prediction. In: Smys S., Tavares J., Balas V., Iliyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham

[4] Adnan Amin, Sajid Anwar(2016) Custome churn prediction in Telecom using Rough set approach Neuro computing 237(2017)242-254. 2016 Elsevier B.V. All rights reserved.

[5] K.Kim , J.Lee improved churn prediction in telecom industry by analyzing a large network , Expert syst. Appl 41 (15) (2014) 6575-6584

[6] Maurus Riedweg , Pavol Svaba and Gwendolin Wilke (2018)RevenueOptimization of Tele marketing campaigns for prepaid customers.18714.1PFES-ES.

[7] V.Umayaparvathi , K.Iyakutti (2016) A survey on customer churn prediction in Telecom industry : Datasets, mehtods and metrics ISSN: 2395 -0056

[8] A..M. AlMana and M. S. Aksoy, "An Overview of Inductive Learning Algorithms," Int. J. Comput. Appl., vol. 88, no. 4, pp. 20–28, 2014.

[9] M. S. Aksoy, H. Mathkour, and B. A. Alasoos, "Performance evaluation of rules-3 induction system

on data mining," Int. J. Innov. Comput. Inf. Control, vol. 6, no. 8, pp. 1–8, 2010.

[10] H. Harb, A. Makhoul and C. Abou Jaoude, "A Real-Time Massive Data Processing Technique for Densely Distributed Sensor Networks," in IEEE Access, vol. 6, pp. 56551-56561, 2018.

[11] Telecommunication Subscribers' Churn Prediction Model Using Machine Learning‖ , Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal

[12] Vijaya, J., Sivasankar, E. (2018). Improved Churn Prediction Based on Supervised and Unsupervised Hybrid Data Mining System. ICT4SD2016, pp. 485–499. Springer, Singapore

[13] Amoroso, D., and P. A. J. A. R. E. E. Ackaradejruangsri. "Exploring the adoption of mobile technologies in Thailand: Development of a research model." Journal of Business Management and Research 6.1 (2016): 19-28.

[14] MALISUWAN, SETTAPONG, and WASSANA KAEWPHANUEKRUNGSI. "Estimation of consumer surplus in mobile services: Case study on telecommunication market in Thailand." International Journal of Computer Networking, Wireless and Mobile Communications 5.6 (2015): 27-40.

[15] Opuni, FRANK FRIMPONG, and K. W. A. M. E. Adu-Gyamfi. "An analysis of the impact of emotional intelligence on service quality and customer satisfaction in the telecommunication sector in Ghana." International Journal of Sales & Marketing Management Research and Development 4.3 (2014): 11-26.

[16] Bernard, Mushirabwoba, et al. "Preliminary Study Of Frictional Power Losses In Spur Geared Transmissions." International Journal 7.4 (2017): 199-208.

[17] Zeb-Obipi, Isaac. "Frameworks of idustrial relations analysis: A re-visit of industrial relations theory." International Journal of Human Resources Management 7.1 (2007): 1-12.