Knowledge Based Churn Prediction In Telecom Networks

By

Name	Roll No.	Registration No:
Rashida Jahan	*11700116058*	*161170110055*
Nitish Prasad Kushwaha	*11700116067*	*161170110046*
Shailza Kumari	*11700116047*	*161170110066*
Anisha Annu	*11700116102*	*161170110011*

UNDER THE GUIDANCE OF
Asst. Prof. Somenath Nag Choudhury
Dept. of Computer Science & Engineering
RCC Institute of Information Technology

PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY

श्रमम् बिना न किमपि साध्यम्

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY
[Affiliated to West Bengal University of Technology]
CANAL SOUTH ROAD, BELIAGHATA, KOLKATA-700015

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY

श्रमम् बिना न किमपि साध्यम्

TO WHOM IT MAY CONCERN

I hereby recommend that the Project entitled **“Knowledge Based Customer Churn Prediction For Telecom Networks”** prepared under my supervision by Nitish Prasad Kushwaha(11700116067), Shailza Kumari (11700116047), Rashida Jahan (11700116058) and Anisha Annu(11700116102) of B.Tech (8th Semester), may be accepted in partial fulfillment for the degree of **Bachelor of Technology in Computer Science & Engineering** under West Bengal University of Technology (WBUT).

.

Somenath Nag Choudhury
……………………………………………
Project Supervisor
Department of Computer Science and Engineering
RCC Institute of Information Technology

Countersigned:

………………………………………
Head
Department of Computer Sc. & Engg,
RCC Institute of Information Technology
Kolkata – 700015.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY

श्रमम् बिना न किमपि साध्यम्

CERTIFICATE OF APPROVAL

The foregoing Project is hereby accepted as a credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project only for the purpose for which it is submitted.

FINAL EXAMINATION FOR
EVALUATION OF PROJECT

1. ______________________________

2. ______________________________

3. ______________________________

(Signature of Examiners)

ACKNOWLEDGEMENT

We acknowledge our overwhelming gratitude & immense respects to our revered guide, Mr. Somenath Nag Choudhury (Asst. Prof., RCC Institute of Information Technology) under whose scholarly guideline, constant encouragement & untiring patience; we have proud privilege to accomplish this entire project work. We feel enriched with the knowledge & sense of responsible approach we inherited from our guide & shall remain a treasure in our life.

Nitish Prasad
Roll no: 11700116067

Shailza Kumari
Roll no: 11700116047

Anisha Annu
Roll no: 11700116102

Rashida Jahan
Roll no: 11700116058

ABSTRACT

Customer churn is a critical and challenging problem affecting business and industry, in particular, the rapidly growing, highly competitive telecommunication sector. It is of substantial interest to both academic researchers and industrial practitioners, interested in forecasting the behavior of customers in order to differentiate the churn from non-churn customers. The primary motivation is the dire need of businesses to retain existing customers, coupled with the high cost associated with acquiring new ones. A review of the field has revealed a lack of efficient, rule-based Customer Churn Prediction (CCP) approaches in the telecommunication sector .The proposed approach effectively performs classification of churn from non-churn customers, along with prediction of those customers who will churn or may possibly churn in the near future. Further, comparative results demonstrate that our proposed approach offers a globally optimal solution for CCP in the telecom sector, when benchmarked against several state-of-the-art methods. Churn prediction helps business to gain a better understanding of future expected revenue , allows to target individual in an attempt to prevent them from discontinuing their subscription with the company and helps to understand what preventative steps are necessary to ensure lost revenue is minimized. Finally, we show how attribute-level analysis can pave the way for developing a successful customer retention policy that could form an indispensable part of strategic decision making and planning process in the telecom sector.

CONTENTS

LIST OF ABBREVIATIONS

1. Gender=gen
2. SeniorCitizen=sc
3. Partner=part
4. Dependent=dep
5. Tenure=ten
6. PhoneService=ps
7. MultipleLines=ml
8. InternetService=is
9. OnlineSecurity=os
10. DeviceProtection=dp
11. TechSupport=ts
12. StreamingTV=st
13. Contract=con
14. PaperlessBilling=pb
15. Payment=pm
16. Monthly Charges=mc
17. Total Charges=tc

Chapter-01

Introduction

In the present world, a huge volume of data is being generated by telecom companies at an exceedingly fast rate. There is a range of telecom service providers competing in the market to increase their client share. Customers have multiple options in the form of better and less expensive services. The ultimate goal of telecom companies is to maximize their profit and stay alive in a competitive market place. A customer churn happens when a vast percentage of clients are not satisfied with the services of any telecom company. It results in service migration of customers who start switching to other service providers. There are many reasons for churning. Unlike postpaid customers, prepaid customers are not bound to a service provider and may churn at any time. Churning also impacts the overall reputation of a company which results in its brand loss. A loyal customer, who generates high revenue for the company, gets rarely affected by the competitor companies. Telecom companies consider policy shift when the number of customers drops below a certain level which may result in a huge loss of revenue .Churn prediction is vital in the telecom sector as telecom operators have to retain their valuable customers and enhance their Customer Relationship Management (CRM) administration . The most challenging job for CRM is to retain existing customers . Due to recent advancements in the field of big data, there exist many data mining and machine learning solutions which can be used to analyze such data. These techniques analyze the data and identify reasons behind customer churning. CRM can employ these techniques to maximize their profit [2]. Furthermore, it may be used to design retention strategies to reduce the ratio of customers that are going to churn. The CRM can achieve the customer retention objective of a company by identifying accurate customer needs by using data mining techniques. Data mining involves the process of identifying the behavior of churn customers from the patterns extracted from the data. Data mining is known by many different names such as business intelligence, predictive modeling, knowledge discovery, and predictive analytics. Our contribution to this study is to propose a churn prediction model.

Chapter-02

Related Work

Sl no	Author Name	Paper Name	Contribution	Challenge
1	Nabgha Hashmi, Naveed Anwer Butt and Dr.Muddesar Iqbal	Customer Churn Prediction in Telecommunication : A Decade Review and Classification	Research make a contribution in the field of customer churn predictive modeling in telecommunication. Thus the paper draws a sketch line for the researchers for reviewing and accumulation of the trends about data mining applications in the field of telecommunication.	The classification techniques are good for analyzing qualitative and continuous data and afterwards interpreting results but these techniques do not guarantee the appropriable accuracy of prediction model for large enough, highly dimensional, non linear or time series datasets.
2	IRFAN ULLAH , BASIT RAZA , AHMAD KAMRAN MALIK , MUHAMMAD IMRAN , SAIF UL ISLAM , AND SUNG WON KIM	A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector	A customer churn model is provided for data analytics and validated through standard evaluation metrics. The obtained results show that our proposed churn model performed better by using machine learning techniques	Investigate eager leaning and lazy learning approaches for better churn prediction. The study can be further extended to explore the changing behavior patterns of churn customers by applying Artificial Intelligence techniques for predictions and trend analysis

Tb 2.1

Work Flow

We have taken the data of the customers which is purified via KDD Process(Knowledge discovery in databases).
Then afer collecting the data we have calculate the churn rate during to all the attributes of the customer.
After that we have made decision tree to specify the important factors. we have made the decision tree upto 6 levels.
Then we have used classfier to get the equation to find the churn rate of the customer , which will be randomly input.

SOFTWARE & HARDWARE REQUIREMENTS

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.
R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.
One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.
R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,

- a suite of operators for calculations on arrays, in particular matrices,

- a large, coherent, integrated collection of intermediate tools for data analysis,

- graphical facilities for data analysis and display either on-screen or on hardcopy, and

- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Our Proposal and Approach

Decision Tree:

A **decision tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

The decision tree classifiers organized a series of test questions and conditions in a tree structure. The following figure shows a example decision tree for prediction. In the decision tree, the root and internal nodes contain attribute test conditions to separate recordes that have different characteristics. All the terminal node is assigned a class lable with probability of churn rate.

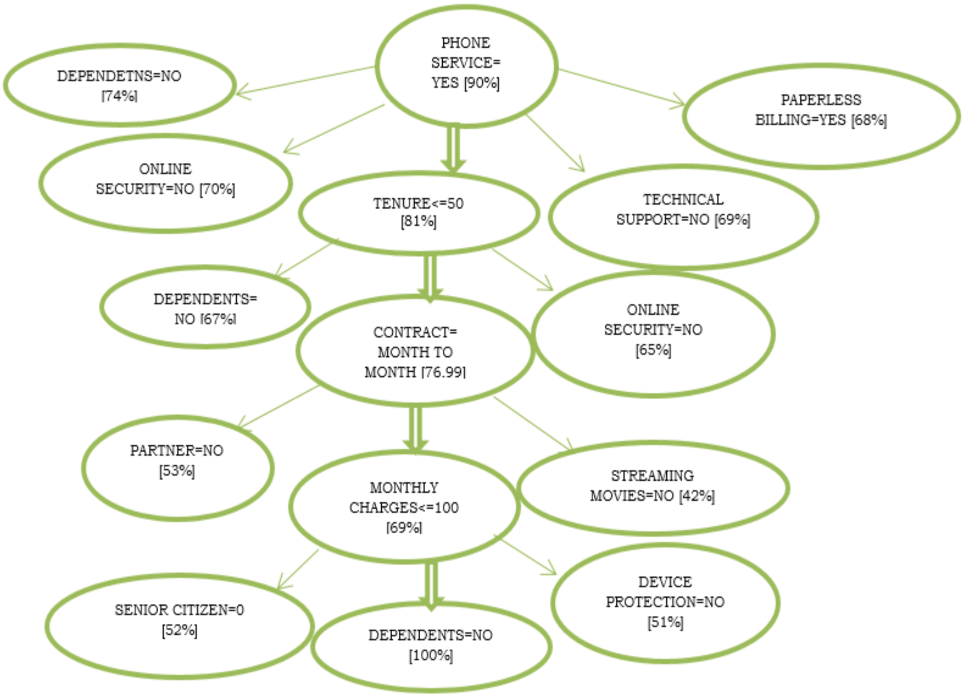

Fig 2.1

Once the decision tree has been constructed, classifying a test record is straightforward. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. It then lead us either to another internal node, for which a new test condition is applied, or to a leaf node. When we reach the leaf node, the class lable associated with the leaf node is then assigned to the record, As shown in the follwoing figure , it traces the path in the decision tree to predict the class label of the test record, and the path terminates at a leaf node.

LEVEL-1: ROOT OF THE DECISION TREE

Features	**Churn Contribution in %**
Female	50.24
Youth	74.53
Independent	82.55
Multiple Lines	45.47
Fibre Optic	69.39
No Online Security	78.17
No Backup	65.97
No Device Protection	64.79
No Technical Support	77.36
No Streaming TV	50.4
No Streaming Movie	50.4
Paperless Billing	74.9
Electronic Transfer	57.3
Phone Service	90

T 2.2

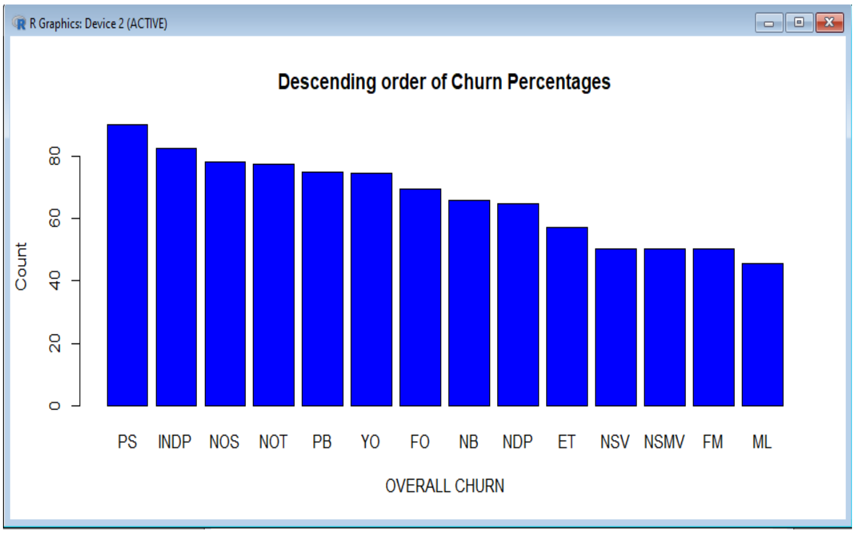

Fig 2.2

LEVEL-2: NEXT ROOT OF THE DECISION TREE

PHONE SERVICE IS “YES” AND CAUSING MAXIMUM CHURN.
NEXT ROOT SELECTION :

Features	**Churn Contribution in %**
Female	45.96
Youth	67.79
Independent	74.69
Multiple Lines	45.47
Fibre Optic	69.39
No Online Security	70.49
No Backup	59.60
No Device Protection	58.31
No Technical Support	69.76
No Streaming TV	44.72
No Streaming Movie	44.94
Paperless Billing	68.59
Electronic Transfer	52.40
Contract	80.36
No Partner	58.37
Tenure	81.59
Monthly Charges	77.36

Tb 2.3

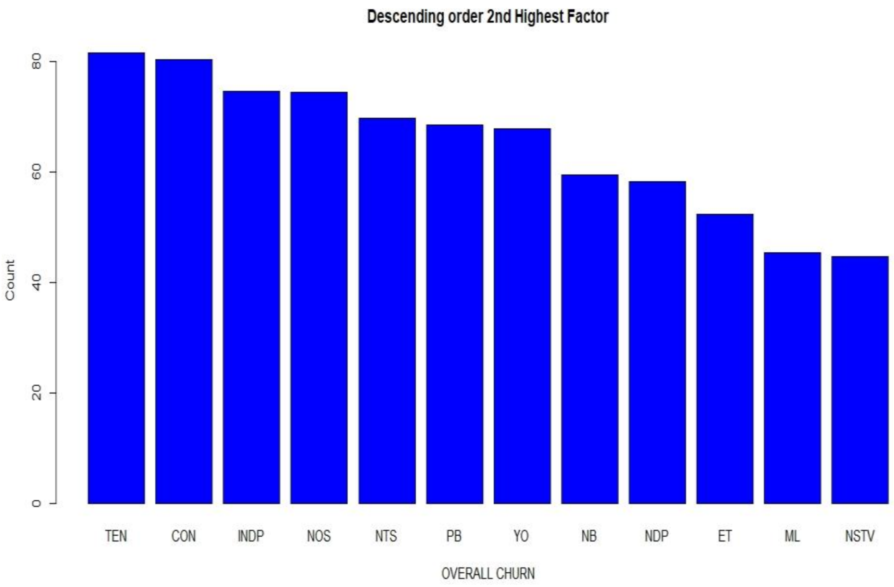

Fig 2.3

LEVEL-3: NEXT ROOT OF THE DECISION TREE

PHONE SERVICE IS “YES” AND TENURE <=50 CAUSING MAXIMUM CHURN.

NEXT ROOT SELECTION:

Features	Churn Contribution in %
Female	42.16
Youth	61.04
Independent	67.46
Multiple Lines	37.66
No Online Security	65
No Backup	56.66
No Device Protection	55.21
No Technical Support	63.99
No Streaming TV	43.07
No Streaming Movie	43.23
Paperless Billing	61.36
Electronic Transfer	47.83
Contract	76.99
No Partner	55.37
Monthly Charges	73.19

Tb 2.4

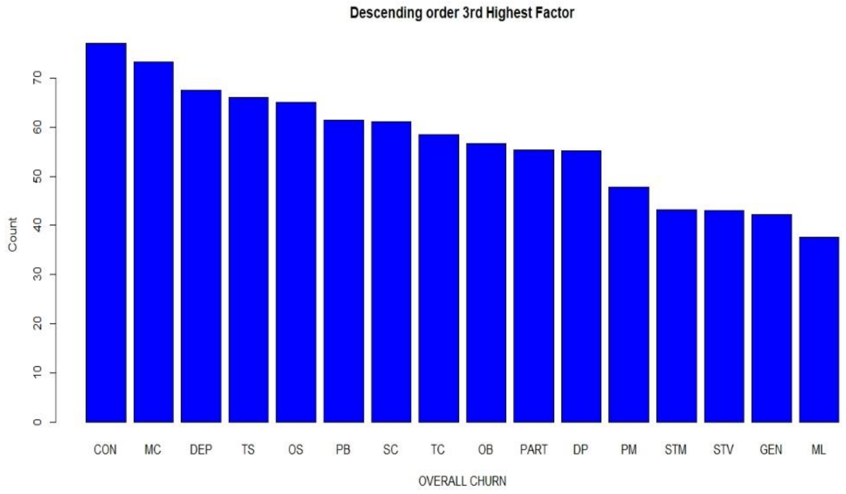

Fig 2.4

LEVEL-4: NEXT ROOT OF THE DECISION TREE

PHONE SERVICE IS “YES” AND TENURE <=50 AND CONTRACT M2M CAUSING MAXIMUM CHURN.

NEXT ROOT SELECTION:

Features	Churn Contribution in %
Female	39.07
Youth	56.98
Independent	64.84
Multiple Lines	35.58
No Online Security	62.38
No Backup	54.36
No Device Protection	53.39
No Technical Support	62.38
No Streaming TV	41.94
No Streaming Movie	42.16
Paperless Billing	58.21
Electronic Transfer	46.33
No Partner	53.18
Monthly Charges	69.87

Tb 2.5

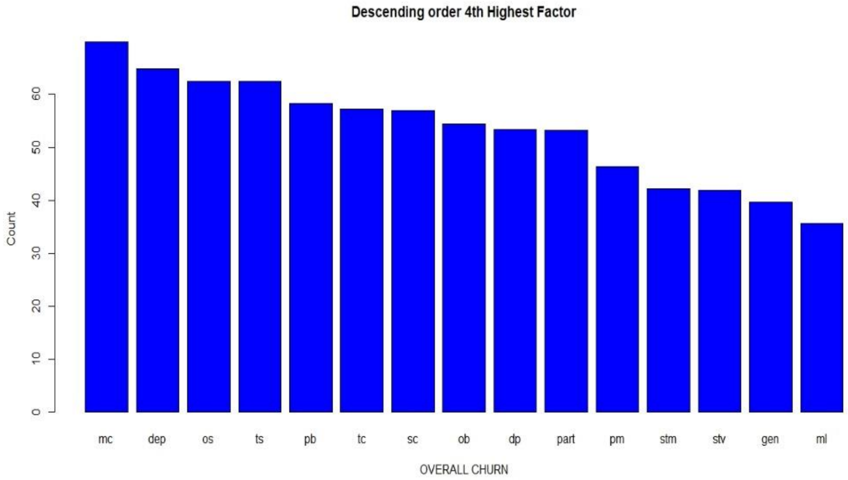

Fig 2.5

LEVEL-5: NEXT ROOT OF THE DECISION TREE

PHONE SERVICE IS “YES”, TENURE <=50, ,MONTHLY CHARGES AND CONTRACT M2M CAUSING MAXIMUM CHURN.

NEXT ROOT SELECTION:

Features	Churn Contribution in %
Female	36.43
Youth	52.38
Independent	100
Multiple Lines	29.10
No Online Security	56.92
No Backup	51.25
No Device Protection	51.63
No Technical Support	57.24
No Streaming TV	41.94
No Streaming Movie	42.05
Paperless Billing	52.05
Electronic Transfer	41.51
No Partner	49.33

Tb 2.5

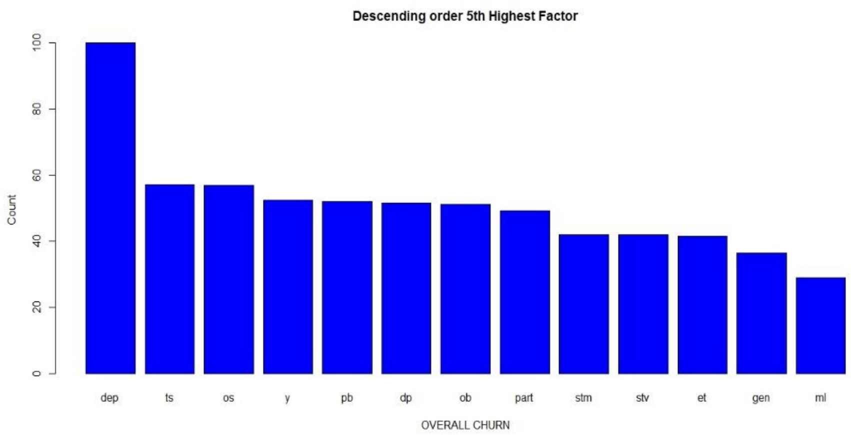

Fig 2.6

Implementation of Classifier:

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Now, before moving to the formula for Naive Bayes, it is important to know about Bayes’ theorem.

Bayes’ Theorem

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:

where A and B are events and P(B) ? 0.

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

where, y is class variable and X is a dependent feature vector (of size *n*) where:

P(y|X) = (P(X|y)P(y))/P(X)
Where X=x1,x2….xn

Chapter-03

Observed Output

After getting the most valuable factors using decision tree, we have

6TH level: 100% churning ->Independents+PHONESERVICE+Contract+tenure+MonthlyCharges

And we have also generated the formula using naïve base classifier to predict the churn rate when some random input will be given.

Churn rate = ax1+bx2+cx3+dx4+ex5
Where a,b,c,d,e are the positive constants and x1,x2 …x5 are the attributes of the Customer.

INPUT:

```
gen = readline(prompt="Input Gender: (M/F):")
sc =  readline(prompt="Input SeniorCitizen or Not: (0/1):")
part= as.integer(readline(prompt="Input Have Partner or Not: (0/1):"))
dep= as.integer(readline(prompt="Input Dependent or Not: (0/1):"))
ten= as.integer(readline(prompt="Input Tenure:"))
ps= readline(prompt="Input Having PhoneService or Not: (y/n):")
ml=readline(prompt="Input Having Multiple Lines or Not: (y/n):")
is= readline(prompt="Input Type of Internet Service : ")
os=readline(prompt="Input Having Online Security or Not: (y/n):")
ob=readline(prompt="Input Having OnlineBackup or Not: (y/n):")
dp= readline(prompt="Input Having DeviceProtection or Not: (y/n):")
ts=readline(prompt="Input Having TechSupport or Not: (y/n):")
stv=readline(prompt="Input Having StreamingTV or Not: (y/n):")
con=readline(prompt="Input Contract Type:")
pb=readline(prompt="Input Having PaperlessBilling or Not: (y/n):")
pm=con=readline(prompt="Input Payment Method:")
mc= as.integer(readline(prompt="Input Monthly Charges:"))
tc= as.integer(readline(prompt="Input Total Charges:"))
print("NEW DATA READ::")
print(paste("Gender=",gen,"SeniorCitizen=",sc,"Partner=",part,"Dependent=",dep,"Tenure=",ten,
"PhoneService=",ps,"MultipleLines=",ml,"InternetService=",is,"OnlineSecurity",os,"DeviceProtection=",dp,"TechSupport=",ts,"StreamingTV=",st,"Contract=",con,"PaperlessBilling=",pb,"Payment=",pm,"Monthly Charges=",mc,"Total Charges=",tc))
```

1.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N   Dependent=N Tenure=70 PhoneService=Y
Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y
TechSupport=Y StreamingTV=Y Contract=Y2Y PaperlessBilling=Y Payment=BankTransfer
Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 83%

2.

NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N   Dependent=N Tenure=40 PhoneService=Y
Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y
TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y
Payment=BankTransfer Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 72%

3.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N   Dependent=N Tenure=40 PhoneService=Y
Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y
TechSupport=Y StreamingTV=Y Contract=Y2Y PaperlessBilling=Y Payment=BankTransfer
Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 55%

4.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N   Dependent=N Tenure=40 PhoneService=N
Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y
TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y
Payment=BankTransfer Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 34%

5.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N   Dependent=N Tenure=40 PhoneService=Y
Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y
TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y
Payment=BankTransfer Monthly Charges=120 Total Charges=500

Conclusion: Probability of Churn is: 60%

6.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=N Tenure=40 PhoneService=Y Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y Payment=BankTransfer Monthly Charges=120 Total Charges=500
Conclusion: Probability of Churn is: 60%

7.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=Y Tenure=40 PhoneService=N Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y Payment=BankTransfer Monthly Charges=120 Total Charges=500
Conclusion: Probability of Churn is: 57%

8.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=N Tenure=40 PhoneService=N Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=Y2Y PaperlessBilling=Y Payment=BankTransfer Monthly Charges=120 Total Charges=500
Conclusion: Probability of Churn is: 45%

9.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=Y Tenure=70 PhoneService=Y Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y Payment=BankTransfer Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 95%
10.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=N Tenure=70 PhoneService=Y Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y Payment=BankTransfer Monthly Charges=120 Total Charges=500
Conclusion: Probability of Churn is: 88%

11.
NEW DATA READ::
Gender=F SeniorCitizen= 0 Partner=N Dependent=N Tenure=70 PhoneService=Y Multiple Lines= N InternetService=Y OnlineSecurity=Y DeviceProtection=Y TechSupport=Y StreamingTV=Y Contract=M2M PaperlessBilling=Y Payment=BankTransfer Monthly Charges=60 Total Charges=500
Conclusion: Probability of Churn is: 100%

Performance Analysis

Since we have probability based classifier we are getting the accurate results upto some extent , since we have used the 5 most effective attributes .

But it may be possible that some customer attributes will not fall in this category , then it may give some unexpected results.

References

- http://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf
- https://pdfs.semanticscholar.org/b666/3d699e3f93a096209a2c29e9c6c85049c56a.pdf
- http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8706988
- https://www.r-project.org/about.html
- https://www.researchgate.net/publication/311484488_Customer_Churn_Prediction_in_Telecommunication_Sector_using_Rough_Set_Approach