

PROFESSIONAL TRAINING REPORT

at

Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Engineering Degree in Information Technology

By:
SAKTHI GANESH

REG.NO:39120099



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

**SATHYABAMA INSTITUTE OF SCIENCE AND
TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**

NOVEMBER 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY



(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI

CHENNAI– 600119

www.sathyabama.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SAKTHI GANESH(Reg. No: 39120099)** who carried out the project entitled “**Telecom Churn Prediction Using Decision Tree Model**” under my supervision from June 2021 to November 2021.

Internal Guide

Dr. Sendru.srinivasulu,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **SAKTHI GANESH** hereby declare that the project report entitled **Telecom Churn Prediction Using Decision Tree Model** done by me under the guidance of **DR.Sendru.srinivasulu** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Information Technology

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean**, School of Computing, **Dr.R.Subhashini M.E., Ph.D, Head of the Department of Information Technology** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.Sendru.srinivasulu** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

CERTIFICATE OF COMPLETION

This certifies that

SAKTHI GANESH S of Sathyabama Institute of Science and Technology

Has successfully completed Training Program on

Data Science Using R

From 9th Sep 2021 to 30th Oct 2021



Dr. Dharani Senthilkumar
CEO - Vani Analytics
<http://www.vanianalytics.com>

ABSTRACT

In recent years, the telecom market has been very competitive. The cost of retaining existing telecom customers is lower than attracting new customers. It is necessary for a telecom company to understand customer churn through customer

relationship management. Therefore, CRM analyzers are required to predict which customers will churn. In this Project, a customer churn prediction model has been proposed that uses Decision Tree data science algorithm to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. By knowing the significant churn factors from customers data, CRM can improve productivity, recommend relevant promotions to the group of likely churn customers based on similar behavior patterns, and excessively improve marketing campaigns of the company.

TABLE OF CONTENTS		
CHAPTER	TITLE	PAGE NO
	ABSTRACT	06
1	INTRODUCTION	09
2	SYSTEM ANALYSIS	10

	2.1 EXISTING SYSTEM	10
	2.2 PROPOSED SYSTEM	11
3	SYSTEM SPECIFICATION	11
	3.1 HARDWARE SPECIFICATION	11
	3.2 SOFTWARE SPECIFICATION	12
4	SYSTEM IMPLEMENTATION	14
	4.1 SYSTEM ARCHITECTURE	14
	4.2 MODULES	15
5	SYSTEM DESIGN	19
	5.1 DATA FLOW DIAGRAM	19
	5.1.1 LEVEL 0	20
	5.1.2 LEVEL 1	21
	5.1.3 LEVEL 2	22
	5.1.4 LEVEL 3	22
6	CONCLUSION	23
	REFERENCES	24
	APPENDIX A: SNAP SHOTS	25

	APPENDIX B: SOURCE CODE	31
--	--------------------------------	----

Fig No	Figure Number	Page No.
1	4.1	14
2	5.1.1 Level 0	20
3	5.1.2 Level 1	21
4	5.1.3 Level 2	22
5	5.1.4 Level 3	22

CHAPTER 1

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.

Owing to fierce competition among telecom companies, customer churn is inevitable. Customer churn is the act of a customer ending a subscription to a service provider and choosing the services of another company. Companies must reduce customer churn because it weakens the company. A survey showed that the annual churn rate in the telecom industry ranges from 20% to 40%, and the cost of retaining existing customers is 5–10 times lower than the cost of obtaining new customers.

The cost of predicting churn customers is 16 times lower than that for obtaining new customers. Decreasing the churn rate by 5% increases the profit from 25% to 85%. This shows that customer-churn prediction is important for the telecom sector. Telecom companies consider customer relationship management (CRM) an important factor in retaining existing customers and preventing customer churn. To retain existing customers, CRM analyzers must predict which customers will churn and analyze the reasons for customer churn.

The goal of this project, is to provide solutions to customer churn using data science algorithm. For prediction, Decision Tree algorithms is used to predict the customer churn using factors, such as voice_mail_plan, number_vmail_messages, total_day_minutes, total_day_calls, total_day_charge, total_eve_minutes, total_eve_calls, total_eve_charge, total_night_minutes, total_night_calls, total_night_charge, total_intl_minutes, total_intl_calls, total_intl_charge, number_customer_service_calls etc. Once the at-risk customers are identified, the company must perform marketing campaigns for churn customers to maximize the churn-customer retention.

CHAPTER 2

SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

Existing concept deals with Qualitative observations and simple statistical analysis. The qualitative observations deals with the data that can be observed through human senses. They do not involve measurements or number. The simple statistical analysis includes mean, standard deviation, median, finding the size of data, variance etc., the results produced by these techniques are not precise. Using some programming languages and data mining tools, telecom customer churn dataset is classified. There is an increase in need to manage and understand data.

2.2 PROPOSED SYSTEM

Churn prediction is vital in the telecom sector as telecom operators have to retain their valuable customers. Telecom customer churn dataset is collected and analyzed to predict the customer churn. For prediction Data Science techniques of Decision Tree algorithm have been applied. Data Science is an interdisciplinary field that incorporates computer science, mathematics, statistics and domain knowledge. Decision Tree Algorithm is applied to the churn data, efficient prediction accuracy is achieved by delivering the result of churn and non-churn classification

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE SPECIFICATION

Processor	:	Processor Intel CORE i3 and above
Internet Connection	:	Existing telephone lines, Data card, Fiber net
RAM	:	4 GB

3.2 SOFTWARE SPECIFICATION

Operating System	:	Windows, Mac, Linux
Language	:	R Programming – R-4.1.1
GUI	:	RStudio

R is a Programming language and environment for Data Manipulation, statistical computing and graphics. Developed at Bell Laboratories by John Chambers and colleagues. Similar to the S language and available across all platforms -Linux, Mac, Windows. It provides variety of Graphical and Statistical techniques – Regression, Clustering, Classification, Association Rule Mining etc.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS. The source code for the R software environment is written primarily in C, FORTRAN, and R.

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.

RStudio is an Integrated Development Environment (IDE) for R. RStudio makes R easier to use. It includes a code editor, compiler or interpreter, debugger and visualization tools as well as tools for plotting, history, syntax-highlighting editor that supports direct code execution and workspace management. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion.

RStudio is available in open source and commercial editions available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application (Windows, macOS, and Linux) and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. RStudio is written in the C++ programming language. RStudio helps to keep R more organized and it adds more functionality to it.

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 SYSTEM ARCHITECTURE

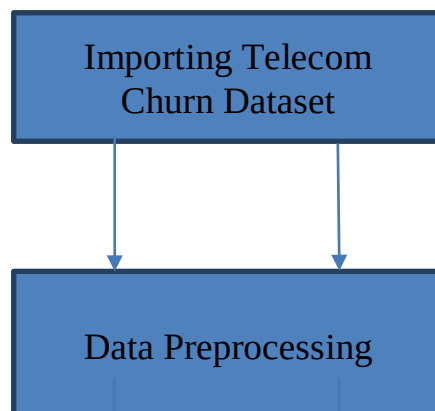


Fig.4.1 Architecture Diagram

The fig 4.1 shows the architectural diagram of this project defines the flow of data for predicting the customer churn. The first step is data importing, Data have to be loaded in to the R environment for analysis. The second step is, preprocessing the collected raw data into understandable format. Standardization and Normalization is the technique which is used to transform the various format of data into the common format and min-max technique is used for normalization of data values.

It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis. The missing values problem have to be solved by simple statistical techniques. Decision Tree algorithm is applied on the preprocessed data to create a predictive Decision Tree model. The algorithm is tested with some predictor variables, the Decision Tree Model will generate the output of whether to the customer will churn or not.

4.2 MODULES

There are four Modules

- Importing Telecom Churn Dataset
- Data Preprocessing
- Model Generation Using Decision Tree Algorithm
- Prediction using Decision Tree Model

4.2.1 IMPORTING TELECOM CHURN DATASET

Data is available in any file format like .txt, .csv, .xlsx, .spss etc. Data have to be loaded in to R environment for analysis. Once data have been extracted from the file it should be stored in a data frame. Packages necessary for classification algorithm – Decision Tree have to be installed into the R environment. For Decision Tree, party package have to be installed and loaded in to R environment.

4.2.2 DATA PREPROCESSING

Data preprocessing is the data mining technique that involves transforming raw data into understandable format. The raw data is highly susceptible to noise, missing values, and inconsistency.

Real world data is often incomplete, inconsistent, and is likely to contain errors. The missing Values problem have to be solved by simple statistical techniques. Data preprocessing is the proven method for resolving such issues. In order to improve the quality of the data consequently, the mining results of raw data is preprocessed so the efficiency process improved. It is not necessary to hold all the

attributes for doing the analysis, we can hold only the attributes which is affecting the analysis.

Data standardization is the process by which similar data is collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and large scale analytics. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Formula used for Normalization:

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

4.2.3 MODEL GENERATION USING DECISION TREE ALGORITHM

Classification is a technique where we categorize data into a given number of classes. Classification is to determine the extent to which a thing will or will not be a part of a category or type. Classification can be performed on structured or unstructured data. The main goal of a classification problem is to identify the category/class to which a new data will fall under.

Decision trees are powerful and popular tools for classification and prediction. J. Ross Quinlan originally developed ID3 at the University of Sydney. ID3 uses Entropy and Information Gain to construct a decision tree. Decision tree is a

classifier in the form of a tree structure. Root node represents entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting is a process of dividing a node into two or more sub-nodes. When a sub-node splits into further sub-nodes, then it is called decision node. Nodes do not split is called Leaf or Terminal node. When we remove sub-nodes of a decision node, this process is called pruning. It is the opposite process of splitting. A sub section of entire tree is called branch or sub-tree. A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node. The entropy is a measure of the uncertainty or impurity associated with a random variable. Values range from 0 – 1 to represent the entropy of information

$$Entropy(S) = \sum_{i=1}^c - p_i \log_2 p_i$$

Information gain is a measure of the effectiveness of an attribute in classifying the training data. Information gain is used as an attribute selection measure. Pick the attribute that has the highest Information Gain

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Algorithmic steps of Decision Tree algorithm

Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split.

The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain

Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

Outlook is having max gain, so tree have to start with outlook.

Step4a: A branch with entropy of 0 is a leaf node.

Step4b: A branch with entropy more than 0 needs splitting

The Previous steps have to be repeated till all the data has been classified.

Decision tree algorithm is faster than other traditional statistical techniques. It is very flexible, easy to implement and debug. It handles both continuous and discrete data. It is used for both binary and multiclass classification. Once the preprocessed data is applied to the Decision tree algorithm, Decision tree Model have been created, which is used to classify the customers for churn. Decision tree model output can be visualized by classification rules and binary classification tree.

4.2.4 PREDICTION USING DECISION TREE MODEL

By using test data, model can be evaluated. The algorithm is used for prediction by passing predictors variable, the Decision Tree Model will perform the churn or non-churn classification.

CHAPTER 5

SYSTEM DESIGN

5.1 DATA FLOW DIAGRAM

Data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. It differs from the flowchart as it shows the data flow instead of the control flow of the program. A data flow diagram can also be used for the visualization of data processing. The DFD is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

5.1.1 LEVEL 0

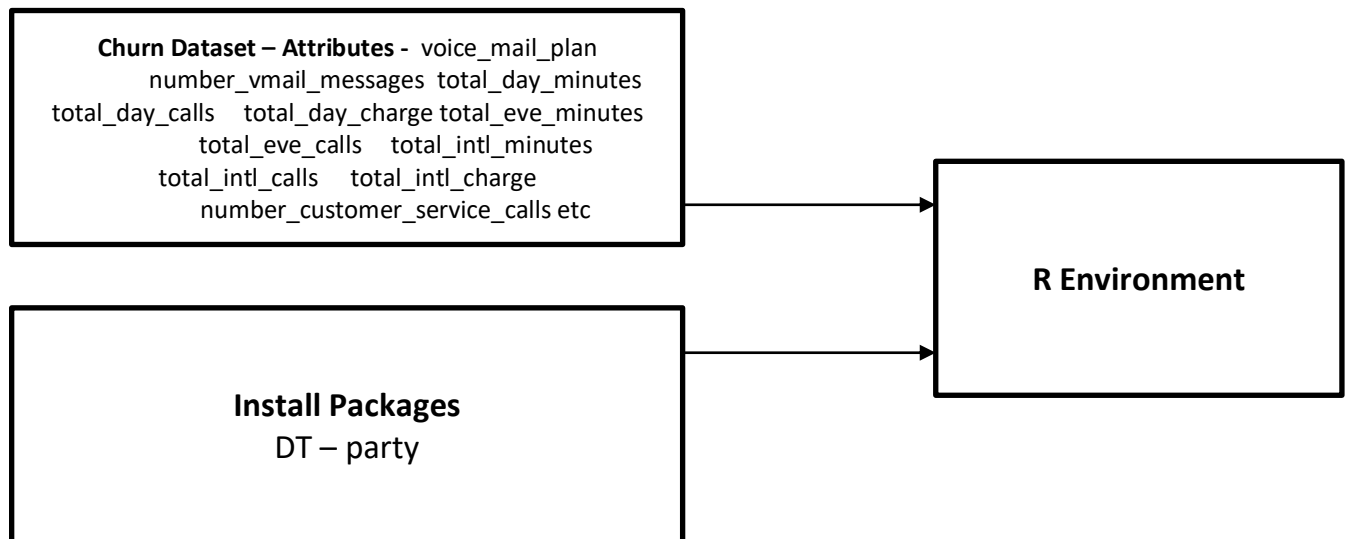


Fig 5.1.1 LEVEL 0

In the fig 5.1.1, Data is imported in to R environment for analysis. The input dataset is Churn dataset, is in csv format have to be loaded into the program to start the analysis. Package necessary for Decision Tree algorithm have to be installed and loaded in to the program. Decision Tree - party.

5.1.2 LEVEL 1

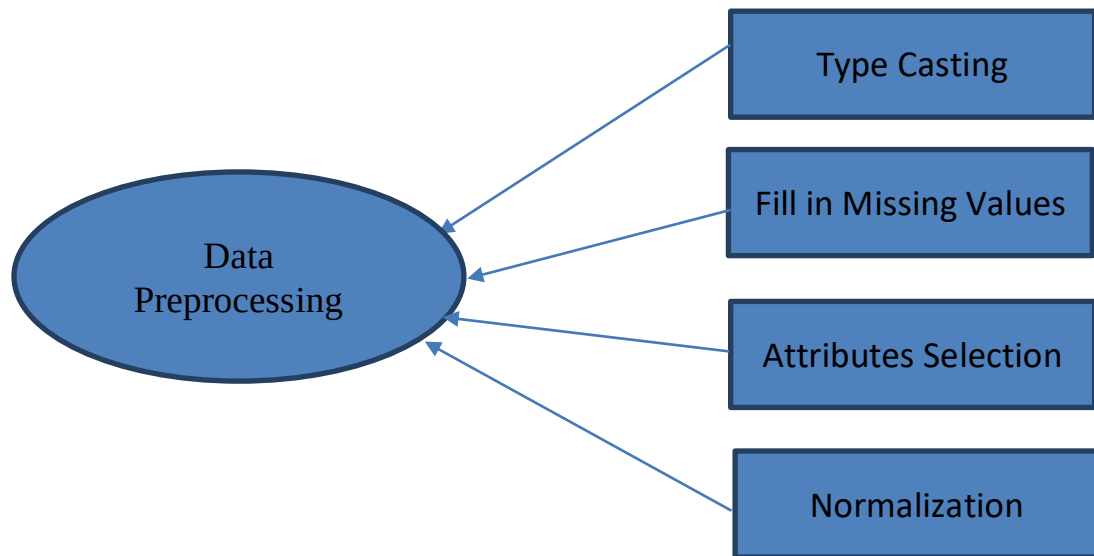


Fig 5.1.2 LEVEL 1

The fig 5.1.2 shows the data preprocessing, which includes attribute selection, standardization and normalization functions. In standardization, raw data is transformed into common, understandable format. In attribute selection, hold only the attributes which is affecting the analysis and it is not necessary to hold all the attributes for doing the analysis. In Normalization, mean of the attribute will be 0 and standard deviation will be 1.

5.1.3 LEVEL 2

In fig 5.1.3, preprocessed data is given as an input to Decision Tree algorithm and Decision Tree model has been created, which is used for classification and prediction.

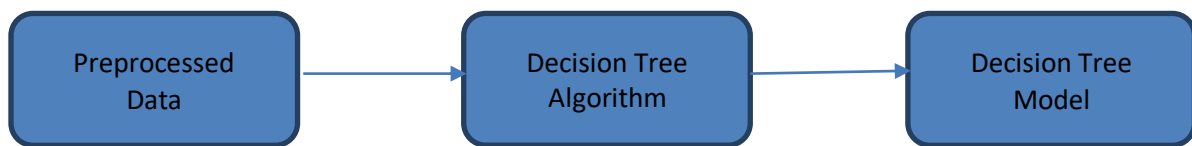


Fig 5.1.3 LEVEL 2

5.1.4 LEVEL 3

In fig 5.1.4, Predictors variables is given as an input and the Decision Tree Model will generate the output of whether to the customer will churn or not. Thus, prediction process is implemented successfully. Using this, model classification is also achieved by passing more number of customer churn datasets, it will classify the customer as two groups. One group is the set of customers who will not churn and the other group is the set of customers who will churn.

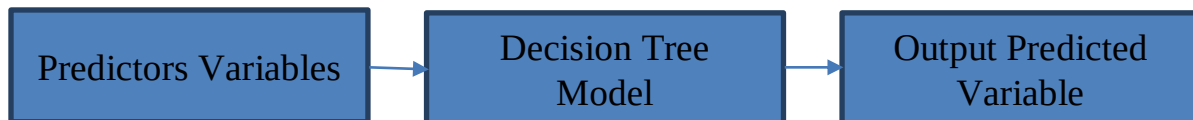


Fig 5.1.4 LEVEL 3

CHAPTER 6

CONCLUSION

In the present competitive market of telecom domain, churn prediction is a significant issue of the CRM to retain valuable customers by identifying a similar group of customers and providing competitive offers/services to the respective groups. Decision Tree algorithm is used to build the predictive model which helps to retain customers and solve the problems of CRM and decision maker of a company. This project provides the guidelines on customer retention for decision-makers of the telecom companies. The performance of this algorithm is much better when compared to other statistical techniques. This project can be further extended to explore the changing behavior patterns of churn customers by applying Deep Learning techniques for predictions and trend analysis.

REFERENCES

1. Laurie Butgereit, “Work Towards Using Micro-services to Build a Data Pipeline for Machine Learning Applications: A Case Study in Predicting Customer Churn", IEEE International Conference on Innovative Trends in Communication and Computer Engineering, 2020.
2. Ishpreet Kaur, “Customer Churn Analysis and Prediction in Banking Industry using Machine Learning", IEEE International Conference on Parallel, Distributed and Grid Computing, 2021.
3. Ahmet Tugrul Bayrak, Asmin Alev Aktas, “Personalized Customer Churn Analysis with Long Short-Term Memory", IEEE International Conference on Big Data and Smart Computing, 2021.
4. Jas Semrl; Alexandru Matei, “Churn prediction model for effective gym customer retention”, IEEE International Conference on Behavioural, Economic, Socio-cultural Computing, 2017.
5. Laurie Butgereit, “Big Data and Machine Learning for Forestalling Customer Churn Using Hybrid Software”, IEEE Conference on Information Communications Technology and Society, 2020

APPENDIX A

SNAP SHOTS

SAMPLE DATASETS

area_code	international	voice_mail	number_v	total_day	total_day	total_day	total_eve	total_eve	total_eve
area_code	no	yes	26	161.6	123	27.47	195.5	103	16.62
area_code	no	no	0	243.4	114	41.38	121.2	110	10.3
area_code	yes	no	0	299.4	71	50.9	61.9	88	5.26
area_code	yes	no	0	166.7	113	28.34	148.3	122	12.61
area_code	no	yes	24	218.2	88	37.09	348.5	108	29.62
area_code	yes	no	0	157	79	26.69	103.1	94	8.76
area_code	no	no	0	184.5	97	31.37	351.6	80	29.89
area_code	yes	yes	37	258.6	84	43.96	222	111	18.87
area_code	no	no	0	129.1	137	21.95	228.5	83	19.42
area_code	no	no	0	187.7	127	31.91	163.4	148	13.89
area_code	no	no	0	128.8	96	21.9	104.9	71	8.92
area_code	no	no	0	156.6	88	26.62	247.6	75	21.05
area_code	no	no	0	332.9	67	56.59	317.8	97	27.01
area_code	no	yes	27	196.4	139	33.39	280.9	90	23.88
area_code	no	yes	33	189.7	66	32.25	212.8	65	18.09
area_code	no	no	0	224.4	90	38.15	159.5	88	13.56
area_code	no	no	0	155.1	117	26.37	239.7	93	20.37
area_code	no	no	0	62.4	89	10.61	169.9	121	14.44
area_code	no	no	0	183	112	31.11	72.9	99	6.2
area_code	no	no	0	110.4	103	18.77	137.3	102	11.67
area_code	no	no	0	81.1	86	13.79	245.2	72	20.84
area_code	no	yes	39	213	115	36.21	191.1	112	16.24
area_code	no	no	0	134.3	73	22.83	155.5	100	13.22

DATA IMPORTING

The screenshot shows the RStudio interface with a script editor, console, and environment pane. The script editor contains the following code:

```
1 library(party)
2
3
4 churn_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\churn.csv", header = TRUE)
5 print(churn_data)
6 head(churn_data)
```

The console shows the output of the `head(churn_data)` command, displaying the first six rows of the data frame. The environment pane shows the `churn_data` object with 2000 observations and 18 variables.

	area_code	international_plan	voice_mail_plan	number_vmail_messages	total_day_minutes	total_day_calls
1	area_code_415	no	yes	26	161.6	123
2	area_code_415	no	no	0	243.4	114
3	area_code_408	yes	no	0	299.4	71
4	area_code_415	yes	no	0	166.7	113
5	area_code_510	no	yes	24	218.2	88
6	area_code_415	yes	no	0	157.0	79

	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes	total_night_calls
1	27.47	195.5	103	16.62	254.4	103
2	41.38	121.2	110	10.30	162.6	104
3	50.90	61.9	88	5.26	196.9	89
4	28.34	148.3	122	12.61	186.9	121
5	37.09	348.5	108	29.62	212.6	118
6	26.69	103.1	94	8.76	211.8	96

	total_night_charge	total_intl_minutes	total_intl_calls	total_intl_charge	number_customer_service_calls	churn
1	11.45	13.7	3	3.70	1	no
2	7.32	12.2	5	3.29	0	no
3	8.86	6.6	7	1.78	2	no
4	8.41	10.1	3	2.73	3	no
5	9.57	7.5	7	2.03	3	no
6	9.53	7.1	6	1.92	0	no

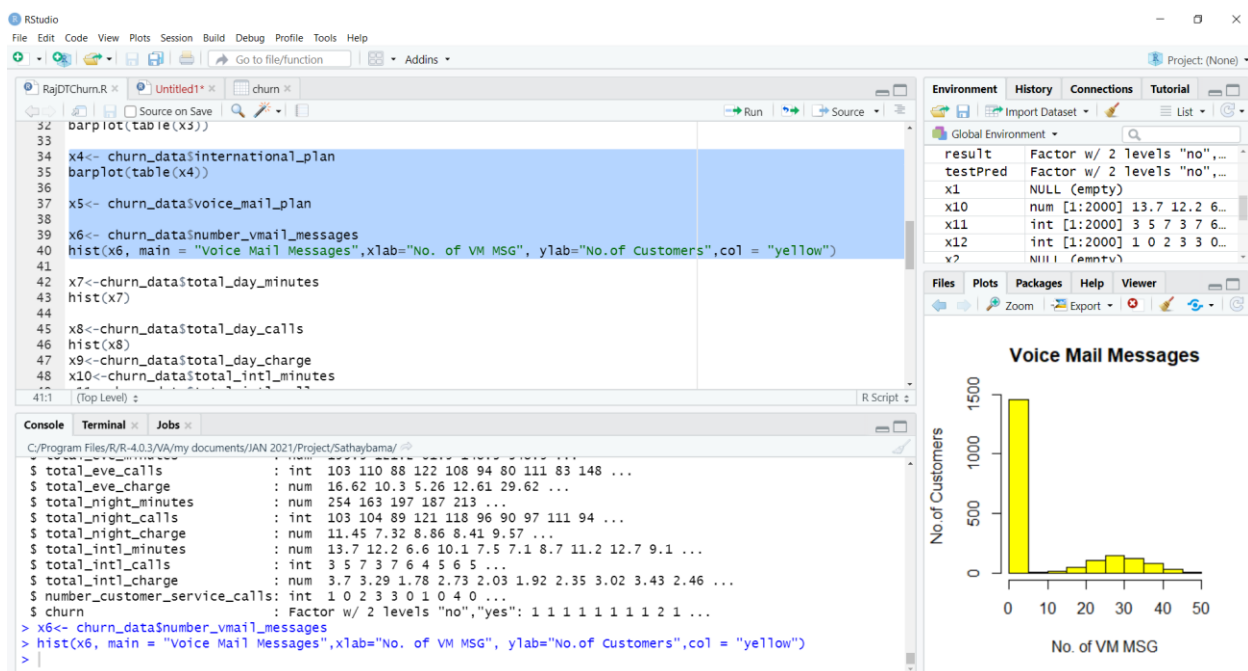
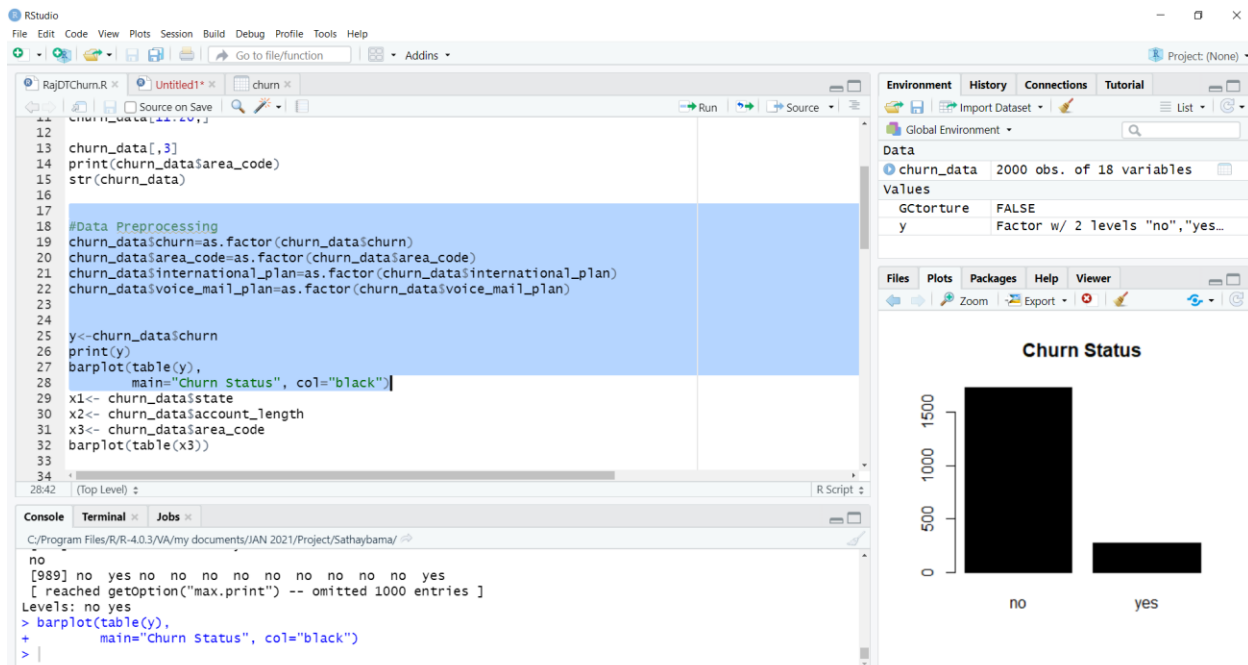
The screenshot shows the RStudio interface with a script editor, console, and environment pane. The script editor contains the following code:

```
1 head(churn_data)
2
3
4
5
6
7
8
9 length(churn_data)
10 churn_data[3,]
11 churn_data[11:20,]
12
13 churn_data[,3]
14 print(churn_data$area_code)
15 str(churn_data)
```

The console shows the output of the `str(churn_data)` command, displaying the structure of the data frame. The environment pane shows the `churn_data` object with 2000 observations and 18 variables.

```
$data.frame': 2000 obs. of 18 variables:
 $ area_code      : Factor w/ 3 levels "area_code_408",...: 2 2 1 2 3 2 1 2 2 2 ...
 $ international_plan : Factor w/ 2 levels "no","yes": 1 1 2 2 1 2 1 2 1 1 ...
 $ voice_mail_plan   : Factor w/ 2 levels "no","yes": 2 1 1 1 2 1 1 2 1 1 ...
 $ number_vmail_messages : int 26 0 0 0 24 0 0 37 0 0 ...
 $ total_day_minutes  : num 162 243 299 167 218 ...
 $ total_day_calls    : int 123 114 71 113 88 79 97 84 137 127 ...
 $ total_day_charge   : num 27.5 41.4 50.9 28.3 37.1 ...
 $ total_eve_minutes  : num 195.5 121.2 61.9 148.3 348.5 ...
 $ total_eve_calls    : int 103 110 88 122 108 94 80 111 83 148 ...
 $ total_eve_charge   : num 16.62 10.3 5.26 12.61 29.62 ...
 $ total_night_minutes : num 254 163 197 187 213 ...
 $ total_night_calls  : int 103 104 89 121 118 96 90 97 111 94 ...
 $ total_night_charge : num 11.45 7.32 8.86 8.41 9.57 ...
 $ total_intl_minutes : num 13.7 12.2 6.6 10.1 7.5 7.1 8.7 11.2 12.7 9.1 ...
 $ total_intl_calls   : int 3 5 7 3 7 6 4 5 6 5 ...
 $ total_intl_charge  : num 3.7 3.29 1.78 2.73 2.03 1.92 2.35 3.02 3.43 2.46 ...
 $ number_customer_service_calls : int 1 0 2 3 3 0 1 0 4 0 ...
 $ churn             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 2 1 ...
```

DATA PREPROCESSING



DECISION TREE ALGORITHM

Decision Tree Classification Model

```

65 testdata <- churn_data[,1:9]
66 print(trainData)
67 print(testData)
68
69 dtree_model <- ctree(churn ~., data=trainData)
70
71 print(dtree_model)
72 plot(dtree_model)
73
74 # Testing Classification Model
75
76 testPred <- predict(dtree_model, newdata = testData)
77
78:1 (Top Level)

```

Console Output:

```

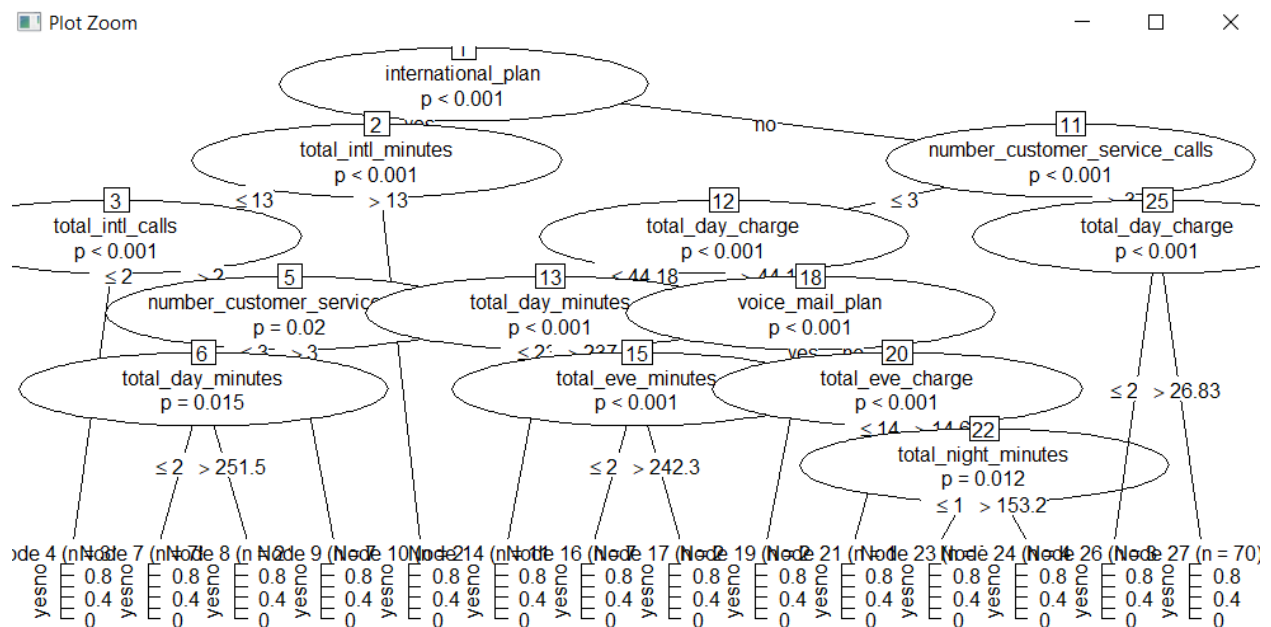
Number of observations: 1592
1) international_plan == {yes}; criterion = 1, statistic = 148.402
2) total_intl_minutes <= 13; criterion = 1, statistic = 18.416
3) total_intl_calls <= 2; criterion = 1, statistic = 20.346
4)* weights = 31
3) total_intl_calls > 2
5) number_customer_service_calls <= 3; criterion = 0.98, statistic = 10.514
6) total_day_minutes <= 251.5; criterion = 0.985, statistic = 11.063
7)* weights = 78
6) total_day_minutes > 251.5
8)* weights = 21
5) number_customer_service_calls > 3
9)* weights = 7
2) total_intl_minutes > 13
10)* weights = 29
1) international_plan == {no}
11) number_customer_service_calls <= 3; criterion = 1, statistic = 86.23

```

Environment:

Variable	Type	Levels	Values
x3	Factor w/ 3 levels	area_cod...	
x4	Factor w/ 2 levels	"no", "yes...	
x5	Factor w/ 2 levels	"no", "yes...	
x6	int	[1:2000]	26 0 0 24 0 0...
x7	num	[1:2000]	162 243 299 167...
x8	int	[1:2000]	123 114 71 113 ...
x9	num	[1:2000]	27.5 41.4 50.9 ...

DT Classification Rule Graph



DT model Testing Using Test Data

The screenshot shows the RStudio interface with a script file named 'RajDTChurn.R'. The script contains the following code:

```
71 print(Dtree_model)
72 plot(Dtree_model)
73
74 # Testing Classification Model
75
76 testPred <- predict(Dtree_model, newdata = testData)
77 print(testPred)
78
79 churn_pred<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\churnpred.csv", header = TRUE)
80 print(churn_pred)
81 str(churn_pred)
82 churn_pred$area_code=as.factor(churn_pred$area_code)
83
```

The console output shows the execution of the script, including the prediction of test data and the structure of the churn data.

```
C:/Program Files/R/R-4.0.3/NA/my documents/IAN 2021/Project/Sathayabama/
> testPred <- predict(Dtree_model, newdata = testData)
> print(testPred)
[1] no no no yes no no no no no no no no no no no yes no no no no no no
[27] no no no no no no no no no no no no no no no yes no no no no no no
[53] no no no no no no no no no no no no no no no yes no no no no no no
[79] no no no no no no no no no no no no no no no yes no no no no no no
[105] no no no no no no no no no no no no no no no yes no no no no no no
[131] no no yes no no no no no no no no no no no no no no no no no no no
[157] no no no yes yes no no no no no no no no no no no no no no no no no
[183] no no no no no no no yes no yes no no no no no no no no no no no no
[209] no no no no no no no no no no no no no no no yes no no no no no no
[235] no no no no no no no no no no no no no no no yes no no no no no no
[261] no no no yes no no no no no no no no no no no no no no no no yes no
[287] no no no no no yes no yes no no no yes no no no no no no no yes no
[313] no yes no no no no no no yes no no no no no no no yes yes no no no
[339] no no no no no no no no no no no no no no no yes no no no no no
[365] no no no yes no no no no yes yes no yes no no no no no no no yes no
[391] no no no no no no no no no no no no no no no no no no no no no
```

DT model- Generates Target- For given predictors

The screenshot shows the RStudio interface with a script file named 'RajDTChurn.R'. The script contains the following code:

```
78
79 churn_pred<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\churnpred.csv", header = TRUE)
80 print(churn_pred)
81 str(churn_pred)
82 churn_pred$area_code=as.factor(churn_pred$area_code)
83 churn_pred$international_plan=as.factor(churn_pred$international_plan)
84 churn_pred$voice_mail_plan=as.factor(churn_pred$voice_mail_plan)
85 str(churn_pred)
86
87 result <- predict(Dtree_model,churn_pred)
88 print(result)
89
90
```

The console output shows the execution of the script, including the prediction of targets for given predictors.

```
C:/Program Files/R/R-4.0.3/NA/my documents/IAN 2021/Project/Sathayabama/
> result <- predict(Dtree_model,churn_pred)
> print(result)
[1] no no no no yes no yes no no no yes no no no no yes no no no no yes no
[27] no no no no no no no no no no no no no no no no no no no no no no
Levels: no yes
```

APPENDIX B

SAMPLE CODING

```
#Importing Packages
```

```
library(party)
```

```
#Data Importing
```

```
churn_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\churn.csv", header  
= TRUE)
```

```
print(churn_data)
```

```
head(churn_data)
```

```
dim(churn_data)
```

```
length(churn_data)
```

```
churn_data[3,]
```

```
churn_data[11:20,]
```

```
churn_data[,3]
```

```
print(churn_data$area_code)
```

```
str(churn_data)
```

```
#Data Preprocessing
```

```
churn_data$churn=as.factor(churn_data$churn)
```

```
churn_data$area_code=as.factor(churn_data$area_code)
```

```
churn_data$international_plan=as.factor(churn_data$international_plan)
```

```
churn_data$voice_mail_plan=as.factor(churn_data$voice_mail_plan)
```

```
y<-churn_data$churn
```

```
print(y)
```

```
barplot(table(y),
```

```
      main="Churn Status", col="black")
```

```
x1<- churn_data$state
```

```
x2<- churn_data$account_length
```

```
x3<- churn_data$area_code
```

```
barplot(table(x3))
```

```
x4<- churn_data$international_plan
```

```
barplot(table(x4))
```

```
x5<- churn_data$voice_mail_plan
```

```
x6<- churn_data$number_vmail_messages
```

```
hist(x6, main = "Voice Mail Messages",xlab="No. of VM MSG", ylab="No.of  
Customers",col = "yellow")
```

```
x7<-churn_data$total_day_minutes
```

```
hist(x7)
```

```
x8<-churn_data$total_day_calls
```

```
hist(x8)
```

```
x9<-churn_data$total_day_charge
```

```
x10<-churn_data$total_intl_minutes
```

```
x11<-churn_data$total_intl_calls
x12<-churn_data$number_customer_service_calls
```

```
barplot(table(x12))
new_data <- cbind(x4,x5,x6,x7,x8,x9,x10,x11,y)
print(churn_data)
print(churn_data[,c(7,8,9)])
churn_data$churn=as.factor(churn_data$churn)
```

#Data Sampling

```
set.seed(1)
ind <- sample(2, nrow(churn_data), replace=TRUE, prob=c(0.8, 0.2))
trainData <- churn_data[ind==1,]
testData <- churn_data[ind==2,]
print(trainData)
print(testData)
```

#Decision Tree Model Generation

```
Dtree_model <- ctree(churn ~., data=trainData)
print(Dtree_model)
plot(Dtree_model)
```

Testing Classification Model

```
testPred <- predict(Dtree_model, newdata = testData)
print(testPred)
```

```
# Generates Target- For given predictors
churn_pred<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\churnpred.csv",
header = TRUE)
print(churn_pred)
str(churn_pred)
churn_pred$area_code=as.factor(churn_pred$area_code)
churn_pred$international_plan=as.factor(churn_pred$international_plan)
churn_pred$voice_mail_plan=as.factor(churn_pred$voice_mail_plan)

result <- predict(Dtree_model,churn_pred)
print(result)
```