**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Churn Prediction in Telecommunication Industry

Submitted by

Shwetha S. Patil 1MS12CS106

Komal K. Saraiya 1MS12CS046

Shewtha S.N. 1MS12CS108

Aniq Ayub Khan 1MS11CS013

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), Dec **2015**

# Abstract

Telecommunication market is expanding day by day. Companies are facing a severe loss of revenue due to increasing competition hence the loss of customers. They are trying to find the reasons of losing customers by measuring customer loyalty to regain the lost customers. The customers leaving the current company and moving to another telecom company are called churn.

The research paper is using data mining technique and R package to predict the results of churn customers on the benchmark Churn dataset available from (http://www.dataminingconsultant.com/data/churn.txt). The R tool has represented the large dataset churn in form of graphs which depicts the outcomes in various unique pattern visualizations. The Churn Factor is used in many functions to depict the various areas or scenarios where churners can be distinguished. The paper is considering churn factor in account to depict various patterns for churners. R is a powerful statistical programming tool which can represent the dataset graphically with respect to different parameters and it also uses different packages available.

Churns can be reduced by analyzing the past history of the potential customers systematically. In the past few years, the fast emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tool for visualization, computational statistics and data science.

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

1 **INTRODUCTION**

* 1. General Introduction……………….
  2. Statement of the Problem…………..
  3. Objectives of the project……………
  4. Current Scope………………………
  5. Future Scope……………………….

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**
3. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirements

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.4.6 Performance

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram and Classes (represent Inheritance, Aggregation and Association)
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK**
2. **REFERENCES**

**1. INTRODUCTION**

1.1 General Introduction

Numerous telecom companies are present all over the world. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up in the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning. Therefore, it has become important for the operators to earn back the amount they invested along with at least the minimum profit within a very short period of time.

1.2 Statement of Problem

Businesses in the consumer market and in all enterprise sectors have to deal with churn. Sometimes churn is excessive and influences policy decisions. The traditional solution is to predict high-propensity churners and address their needs via a concierge service, marketing campaigns, or by applying special dispensations. These approaches can vary from industry to industry and even from a particular consumer cluster to another within one industry (for example, telecommunications).

The common factor is that businesses need to minimize these special customer retention efforts. Thus, a natural methodology would be to score every customer with the probability of churn and address the top N ones. The top customers might be the most profitable ones; for example, in more sophisticated scenarios, a profit function is employed during the selection of candidates for special dispensation. However, these considerations are only a part of the holistic strategy for dealing with churn.

1.3 Objectives of the Project

Churn is problem in Telecom Industry and which is much more critical with Mobile Number Portability. Telecom Companies want to identify churn and prevent Churn by offer discounts to customer so that they don’t switch. So we use classified machine learning techniques in our data analytics to identify Churners early and Churn score to prevent future churn .

1.4 Current Scope

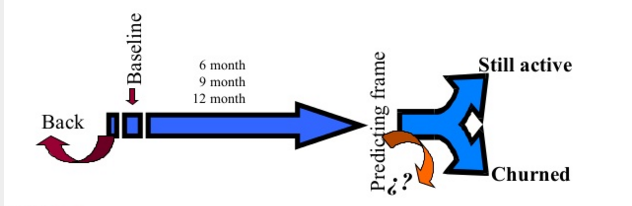
Currently the result will help the company to understand the churn. Based on this prediction company will try to prevent the probable churn from happening by providing benefiting offers and discounts to customers.

1.5 Future Scope

After giving discounts and offers, the company will analyze the actual churn and based on that they will know the effectiveness of churn score and can help them from suffering loss of customer and finance.

**2. PROJECT ORGAINSATION**

2.1 Software Process Model



2.2 Roles and Responsibilities

Each member took their responsibilities and did their work religiously. We had equally divided the roles and we could successfully complete the work in due time.

Aniq generated datasets using “mockaroo” random data generator. The Presentation was also done by him.

Komal and Shwetha S.N contributed in understanding and visualizing datasets in SAP Lumira. Moreover they were responsible for documentation.

Shwetha S. Patil understood the R Tool, and the R language, and the implementation of logistic regression in R and visualization in R.

After understanding and implementing the algorithm and getting the results. All the team members wrote the paper.

**3. LITERATURE SURVEY**

Churn customer is one who leaves the existing company and become a customer of another competitor company. The management that was assumed to determine the customer turnover is called as Churn management. Customer movement from one provider to another in telecommunication industry is called customer churn and the operator‘s process to retain profitable customers counted as churn management.

There have been many papers written on churn and many experiments done. Logistic regression and decision tree J48 have been used the most in the experiments. These are the most efficient techniques to generate the data.

**4. SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

Prediction of churn score to help the tecommucation company which is our client. This can be done by providing the churn score by the product created by us. By this result they will provide

4.2 External Interface Requirements

1. **User interface**

In this project, user interface is in the form of visualization of data colletcted in statistical form, eg. bar graph.

1. **Hardware interface**

Since the algorithm and data sets to be used do not have any designated hardware it does not require any specific hardware interfaces. The raw data collected is pre-processed using R.

1. **Software interface**

The software communicates with datasets acquired and an algorithm is run on them to predict the churn score.

1. **Communication Interface**

The communication between the different parts of the system is important since they depend on each other. However, in what way the communication is achieved is not important for the system and is therefore handled by the underlying operating systems.

4.3 Functional requirements

a) Collect data from the client and form data set

b) Run the algorithm on the obtained data sets

c) Predict the churn score using linear regression

d) Submit the result to the client

4.4 Software System Attributes

1. **Reliability**

TAG: SystemReliability

GIST: The reliability of the system.

SCALE: The reliability that the algorithm derives the right churn score.

METER: Churn score obtained from 100000 datasets during testing.

MUST: More than 98% of the datasets.

PLAN: More than 99% of the datasets.

WISH: 100% of the datasets.

1. **Availability**

TAG: SystemAvailability

GIST: The availability of the system when it is used.

SCALE: The average system availability .

METER: Churn score obtained from 100000 datasets during testing.

MUST: More than 98% of the time.

PLAN: More than 99% of the time.

WISH: 100% of the time.

1. **Security**

TAG: SystemSeurity

GIST: The Security of the churn score obtained

SCALE: Only authenticated clients can see the churn score.

METER : Attempts to get the password to access churn score.

MUST: More than 98% of the time.

PLAN: More than 99% of the time.

WISH: 100% of the time.

1. **Portability**

TAG: SystemPotability

GIST: The algorithm for churn prediction can be run on different platforms.

SCALE: Number of platforms it is compatible with.

RAT: The adaptable platform for the application to run on.

1. **Maintainability**

TAG: Application extendibility

GIST: The application should be easy to extend. The code should be written in a way that it favors implementation of new functions.

RAT: In order for future functions to be implemented easily to the application.

1. **Performance**

TAG: SystemPerformance

GIST: The application should perform well under all conditions

RAT: No data should be lost in the process and

4.5 Performance requirements

* + Database transactions per second
  + Accuracy of the churn score
  + Response time

4.6 Database requirements

We require SQL Server Name with the following :

* Database
  + Name
  + Size
  + Anticipated growth rate or capacity planning
  + Database configurations
  + Storage configurations
* Data
  + Data elements
    - Tables
    - Columns
    - Data types
    - Accept null
    - Defaults
  + Data access
    - Determine how the data will be accessed for index selection
  + Source
    - Uploads or downloads from an existing system
    - Data entry by customers.

4.7 Design Constraints

* Complexity of Logistic algorithm
* All coding will be done in R.
* Logistic regression requires huge amount of data to achieve stable and meaningful output.

**5. DESIGN**

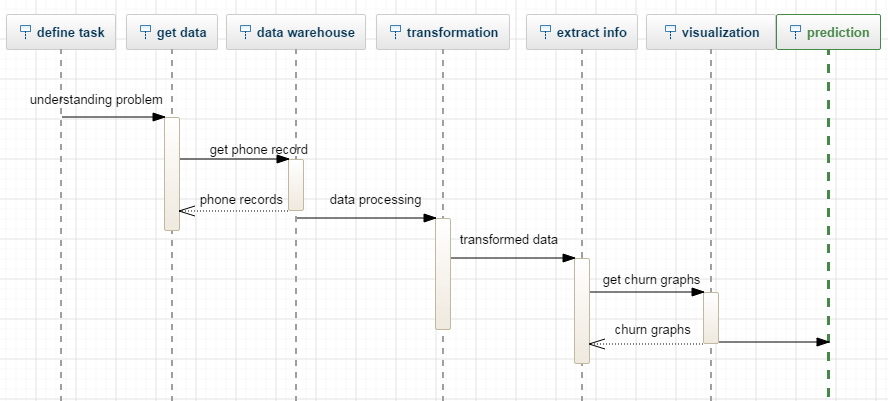
Introduction

First, understand the problem task and collect the data from data warehouse. After acquiring the data preprocess it and reduce variables and rows and reduce the overall data to the limited amount which we require. Then the data is transformed into a format suitable for input. Try the different models and use the one most suitable to the project and identify the significant variables. Interpret the output of models and score new data and test that data on hold out samples.

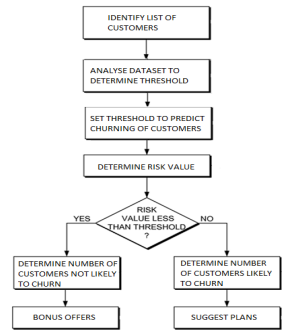
Graphical User Interface

R statistical tool, RStudio and SAP Lumira are the graphical user interfaces used in the project.

Sequence Diagram



Data flow diagram



Metric calculation

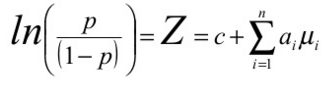
After Model is evaluated based on the performance on the Hold Out Sample, the model is scored in real time to identify churners on real data.

P= Probability of getting the condition of “churn” .

P=0 customer will remain active.

P=1 customer will churn.

LOGISTIC REGRESSION :-



**6. IMPLEMENTATION**

Tools Introduction

**R** is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

Technology Introduction

**SAP Lumira** is a self-service, data visualization application for business users created by Microsoft. It creates visualization like pie chat, bar graph etc based on the attributes selected by the user from the datasets provided.

Overall view of the project in terms of implementation

Churns can be reduced by analyzing the past history of the potential customers systematically. In the past few years, the fast emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tool for visualization, computational statistics and data science

Using logistic regression, classification tree, decision tree, confusion matrix we predict the churn score for each customer and visualize the data using RStudio and SAP Lumira.

Explanation of Algorithm and how it is been implemented

**Algorithm** :

**Step 1**. A flow-chart-like tree structure. Internal node denotes a test on an attribute. Outcome of the test is represented by Branch. Class labels are represented by Leaf nodes.

**Step 2**. Decision tree generation comprised of two phases.

Tree construction: At start, root contains all the training examples. Tree pruning: Branches that reflect noise and outliers are identified and removed.

**Step 3**. Decision tree is used to classify an Unknown sample. Attribute values of the sample are tested against the decision tree.

**Step 4**. When all samples for a given node belong to the same class, or there are no remaining attributes for further partitioning then the partitioning is stopped.

**Implementation:**

* Extracting Classification Rules from Trees

1. IF-THEN rules are used.

2. From root to leaf one rule is created for each path.

3. Each attribute-value pair along a path forms a conjunction.

4. The leaf node holds the class prediction.

5. Rules are easier for humans to understand.

* Using rpart package

rpart (formula, data, method)

f<-rpart(Churn.~CustServ.Calls+Eve.Charge+Intl.Charge+Night.Charge+Day.Charge, method="class", data=churn)

Package rpart is used in plotting the graphs. The functions within rpart that are used are as follows:

* Using plotcp function

Applied on the set of possible cost- complexity pruning of a tree from a nested set. A cross- validation is already performed by rpart on the geometric means of the Interval values of cp where pruning is optimal. The mean and standard deviation of errors in cross- validated prediction against each of the geometric means is stored in cptable in ‗f‘ are plotted by this function. A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal line.

* Using Plot function

plot(Churn. ~., data = churn, type = "c")

lines(Churn.~ Day.Charge,type="l")

In plot function, x and y axis are mentioned along with the data source and the type of graph that is, curve, line etc.

Information about the implementation of Modules

4.1 Reading Data Set Churn from CSV file

churn<-read.csv("C:\\Users\\Documents\\R\\win-library\\3.1\\RWeka\\R\\churn.csv", header=T)

4.2. Names of all the attributes

> names(churn)

[1] "State" "Account.Length" "Area.Code" "Phone"

[5] "Int.l.Plan" "VMail.Plan" "VMail.Message" "Day.Mins"

[9] "Day.Calls" "Day.Charge" "Eve.Mins" "Eve.Calls"

[13] "Eve.Charge" "Night.Mins" "Night.Calls" "Night.Charge"

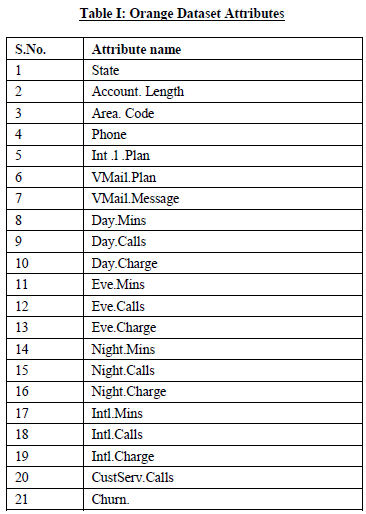
[17] "Intl.Mins" "Intl.Calls" "Intl.Charge" "CustServ.Calls"

[21] "Churn."

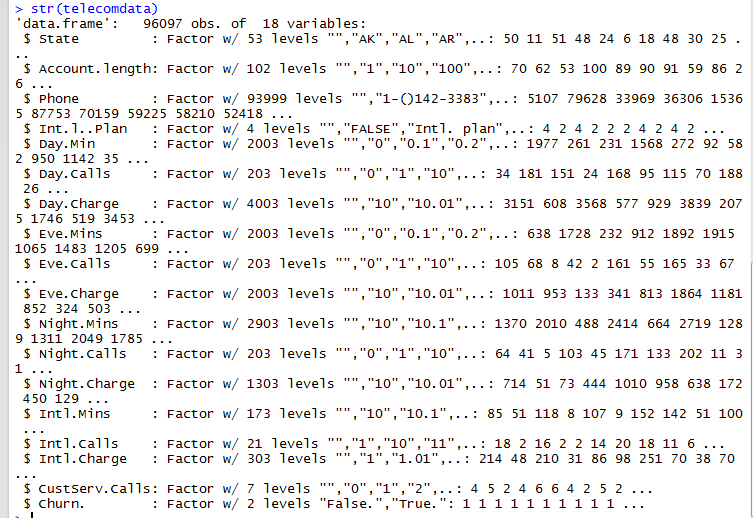
**7. TESTING**

7.1 RESULTS AND SNAPSHOTS

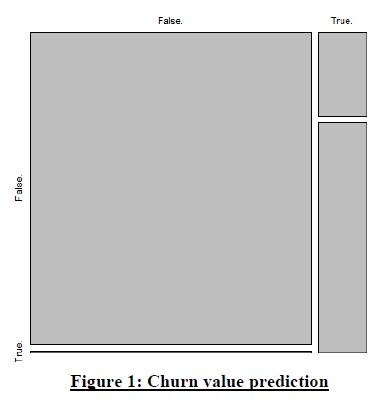
These are the attributes used in the datasets

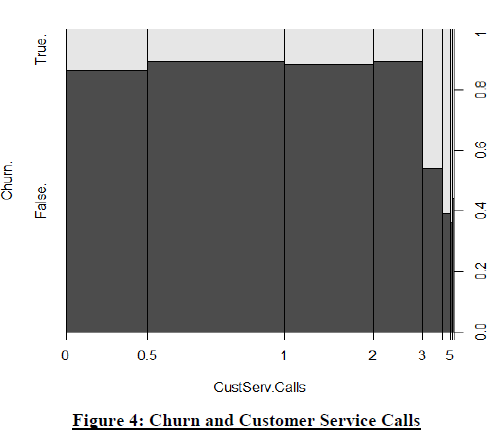


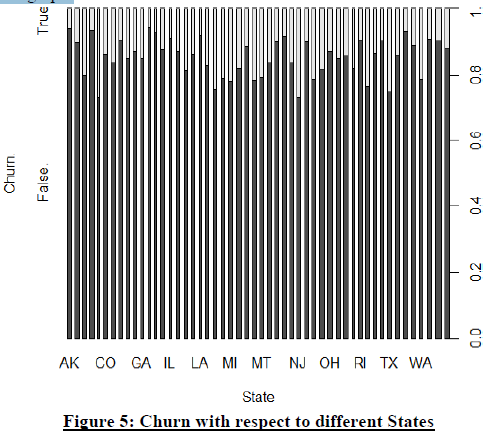
These are the attributes in RStudio

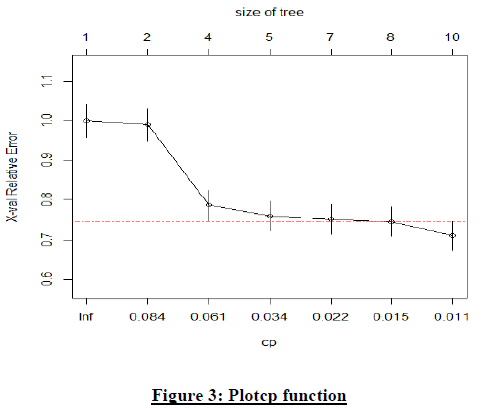


Following are the results obtained in RStudio

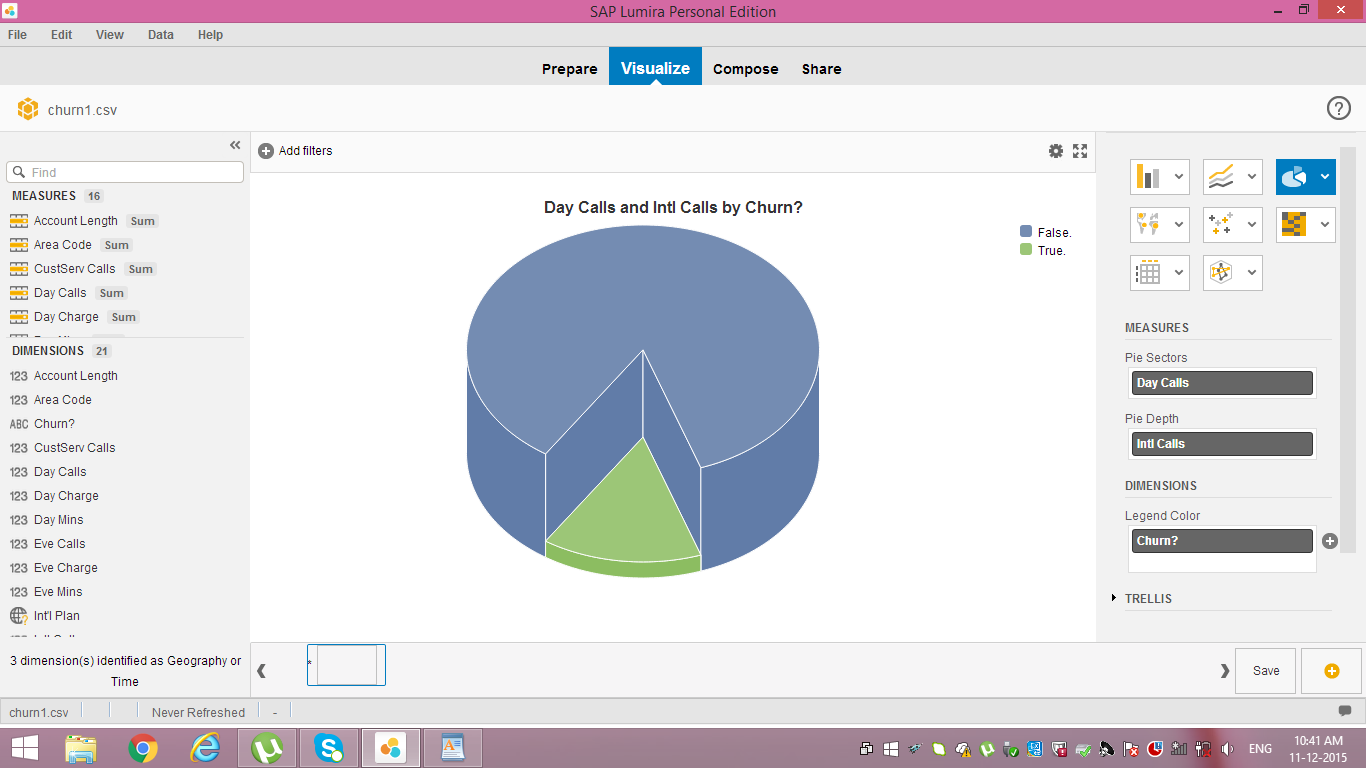


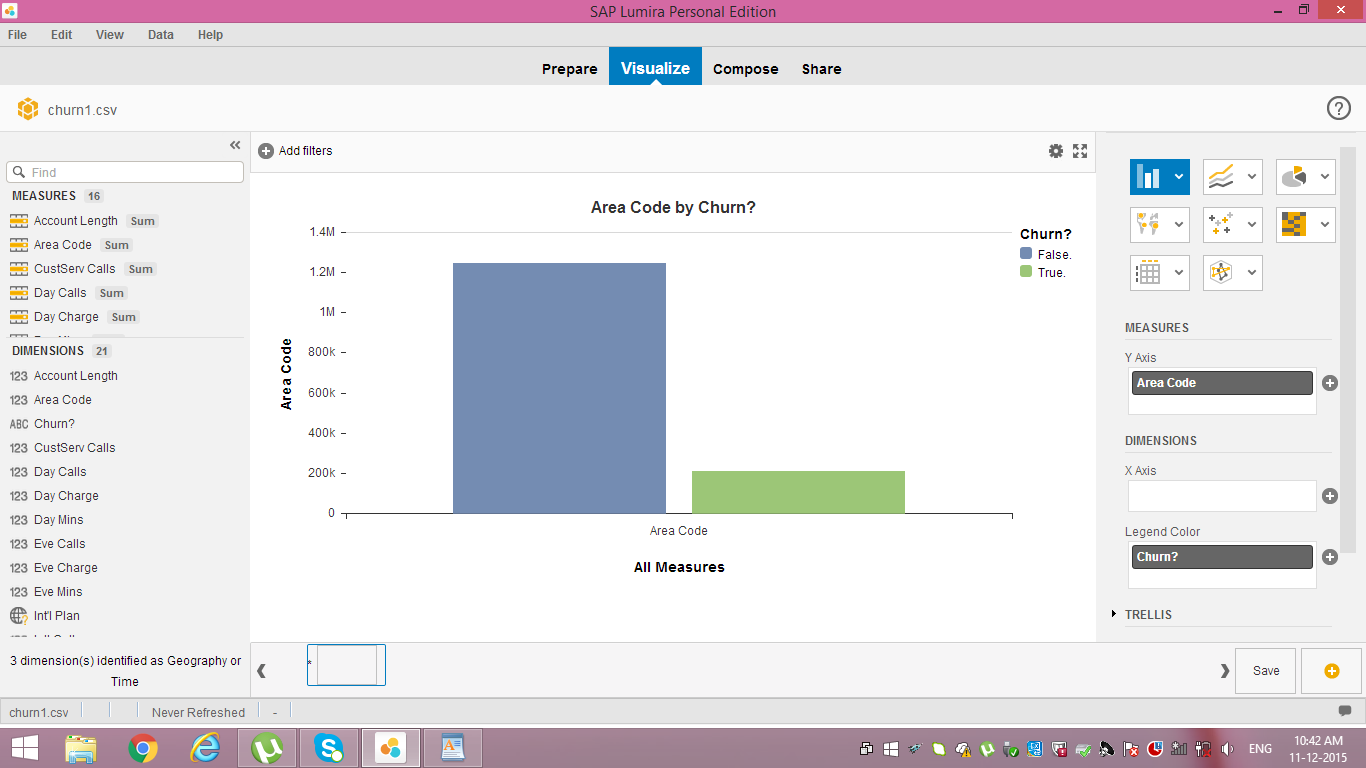


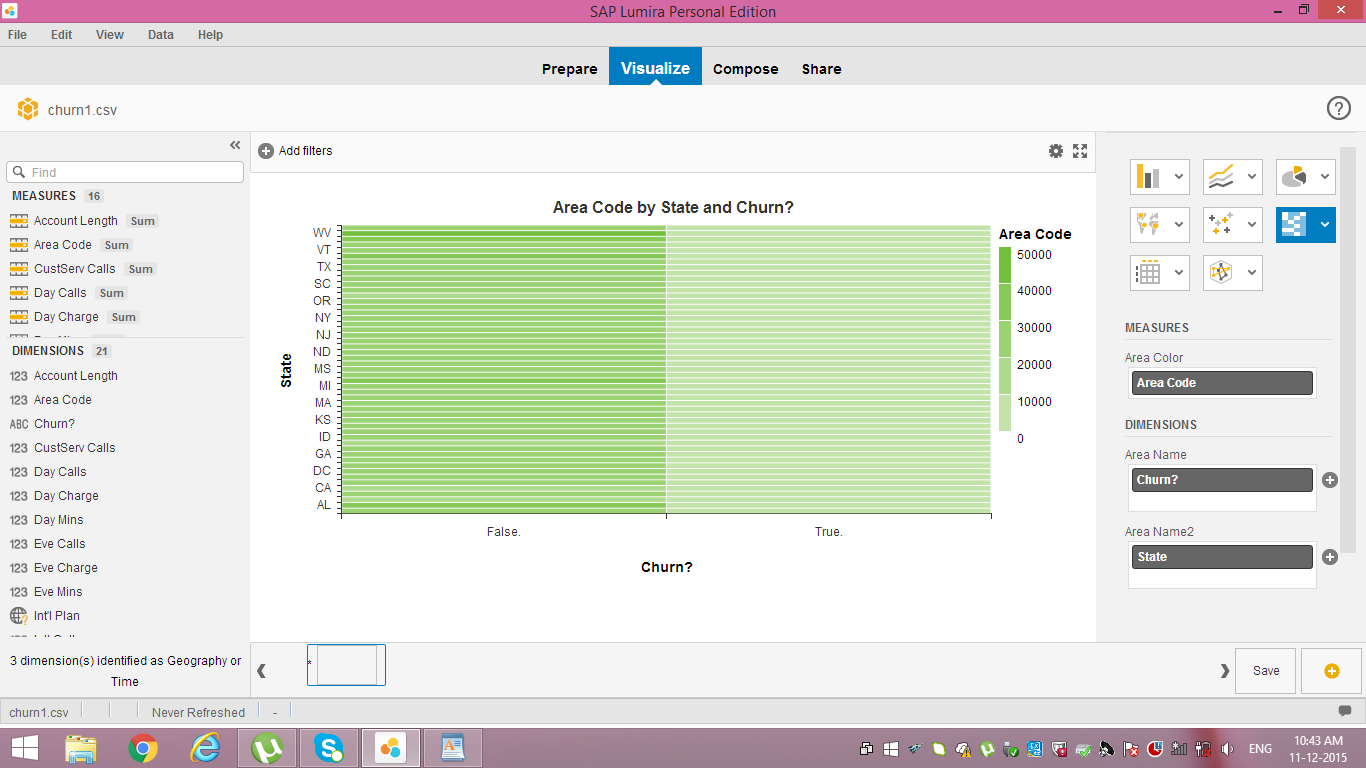




Using SAP Lumira following are the results obtained







**8. CONCLUSION & SCOPE FOR FUTURE WORK**

The proposed research has used data mining technique and R package to predict the results of churn customers on the benchmark Churn dataset available at http://www.sgi.com/tech/mlc/db/ and mockaroo.com. It has evaluated, the number of churns using logostic regression and decision tree. The R tool has represented 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. The study predicts that there is a huge deviation in graph of churners when customer service calls are measured. The graphs are made taking churn factors as the deciding parameters. Graphs represent the different ways of observing the number of churners from the dataset. Once the root area is recognized the steps can be taken by Telecom Company to improve their services and retain their old customers from churning. SAP Lumira has been used just for visualization fro better understanding.

**9. REFERENCES**

[1] ―Applying Data Mining Techniques in Telecom Churn Prediction‖ , International Journal of Advanced Research in Computer Science and Software Engineering, N.Kamalraj Dr.A.Malathi , Department of Computer Technology PG and Research Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science Govt. Arts College.

[2] Anshul Goyal and Rajni Mehta, Assistant Professor,CSE. ―Performance Comparison of Naïve Bayes and J48 Classification Algorithms.‖

[3] R Data: http://cran.r-project.org/

[4] Vladislav Lazarov, Technische Universität München and Marius Capota, Technische Universität München, ―Churn Prediction‖,

[5] Prof S. Chandrasekhar, ―Predicting The Churn In Telecom Industry‖‘.

[6] Dr. Rajni Jain, ―Introduction to Data Mining Techniques‖.

[7] Marcin Owczarczuk, ―Churn models for prepaid customers in the cellular telecommunication industry using large data marts‖, Expert Systems with Applications, 37 (2010) 4710–4712.

[8] Gary M. Weiss, ―Data Mining in the Telecommunications Industry‖, Data Mining and Knowledge Discovery Handbook, 2005 - Springer

[9] ―Data Mining And Business Analytics with R‖, Johannes Ledolter, Dept. of Management Studies, University of IOWA.