**DATA MINING IN TELECOMMUNICATIONS INDUSTRY TO**

**STUDY CONSUMER BEHAVIOUR**

PROJECT REPORT

Submitted by

**ADITHYASHANKAR AJITH (11400002)**

**ASHOK M. D (11400005)**

**SANJAY J (11400015)**

*To*

*The University of Kerala*

*In partial fulfilment the requirements for the award of the Degree*

*Of*

*Bachelor of Technology in Mechanical Engineering*



**Department Of Mechanical Engineering**

College of Engineering, Thiruvananthapuram – 16

March, 2014

**ACKNOWLEDGEMENT**

We express our whole hearted gratitude to our guide Sri.Mahesh S, Asst. Professor, College of Engineering, Trivandrum for the expert guidance and advice in doing this project.

We express our sincere thanks to Prof. G L Aswini Kumar, Head of the Department, Department of Mechanical Engineering, College of Engineering, Trivandrum, for his kind cooperation during the course of this work.

We would also wish to record our gratefulness to all our friends and classmates for their help and support in carrying out this work successfully.

Last but not the least; we would like to express our heartfelt thanks to *God Almighty*.

**DEPARTMENT OF MECHANICAL ENGINEERING**

**COLLEGE OF ENGINEERING**

**THIRUVANANTHAPURAM – 16**

**CERTIFICATE**



This is to certify that the report entitled “**DATA MINING IN TELECOMMUNICATIONS INDUSTRY TO STUDY CONSUMER BEHAVIOUR**”, submitted by “**ADITHYASHANKAR AJITH, ASHOK M D** and **SANJAY J**” to the University of Kerala in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Industrial Engineering is a bonafide record of work carried out by them under our guidance and supervision.

**Sri. Mahesh S**

**Assistant Professor**

**Dept.of Mechanical Engineering,**

**College of Engineering**

**Trivandrum**

**Prof. G L Aswini Kumar**

**Head of the Department**

**Dept. of Mechanical Engineering**

**College of Engineering Trivandrum**

**ABSTRACT**

Having a high teledensity of over 90, Kerala is proving to be a fierce ground for competition among service providers. Customer satisfaction is given a high priority by the service providers and hence it is imperative that they look into the factors that will cause customer retention and churning. ‘**Data mining in telecommunications industry to study consumer behaviour**’ is a scientific approach to understanding the consumer behaviour in the telecommunications industry in the context of the urban population of Trivandrum district, Kerala, India. The data was collected through an extensive survey done among a sizeable sample population. A questionnaire was created for the purpose and both online and offline survey methods were employed in the collection of data. Data analysis is done on the obtained data to study the behavioural patterns of the consumers currently existing in the telecommunications industry in Trivandrum city, Kerala and it also helps predict the same by means of decision trees.

**CONTENTS**

CHAPTER 1

1.1 INTRODUCTION

1.2 OBJECTIVE

1.3 PROBLEM SCENARIO

CHAPTER 2

2.1 LITERATURE REVIEW

2.1.1 DATA MINING

2.1.2 R

2.1.3 IBM SPSS

2.1.4 BEHAVIOURAL ANALYSIS OF THE DIFFERENT SEGMENTS TO PROPOSE

MARKETING RECOMMENDATIONS FOR MOBILE OPERATORS.

2.1.5 CUSTOMER BEHAVIOUR

CHAPTER 3

3.1 DATA COLLECTION

3.1.1 QUESTIONNAIRE

3.1.2 SAMPLE SIZE DETERMINATION

3.1.3 COLLECTION OF DATA

3.2 DATA ANALYSIS

3.2.1 INTRODUCTION

3.2.2 DESCRIPTIVE STATISTICS

3.2.3 ANALYSIS USING R

CHAPTER 4

PREDICTIVE MODELING USING DECISION TREE

CHAPTER 5

CONCLUSION

CHAPTER 6

BIBLIOGRAPHY

**LIST OF FIGURES AND TABLES**

**FIGURES**

1. Taxanomy of data mining
2. Data mining process
3. Occupation vs service providers
4. Age vs Purpose
5. Age vs Parameter
6. Calltime vs Age
7. Onlinetime vs Occupation
8. Clustering
9. Clustering Vector
10. Result of associate rule mining of cluster 1
11. Result of associate rule mining of cluster 2
12. Result of associate rule mining of cluster 3
13. Result of associate rule mining of cluster 4
14. Result of associate rule mining of cluster 5
15. Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and para in Rhs

1. Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and CT in Rhs

1. Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and OT in Rhs

1. Result of associate rule mining when gender,occup,salary are considered in

Lhs and purpose in Rhs

19. Decision Tree

**TABLES**

1.Average Rating of service providers

**CHAPTER-1**

**1.1 INTRODUCTION**

Telecom is one of the fastest-growing industries in India. India's telecommunication network is the second largest in the world based on the total number of telephone users (both fixed and mobile phone) according to a study conducted by EconomicTimes on 16 August 2012. According to the Internet And Mobile Association of India (IAMAI), the Internet user base in the country stood at 190 million at the end of June, 2013.The mobile phone industry in India would contribute US$ 400 billion in terms of gross domestic product (GDP) of the country in 2014.

Data traffic powered by third generation (3G) services grew at 146 per cent in India during 2013, higher than the global average that saw usage double, according to an MBit Index study by Nokia Siemens Networks (NSN). Indian telecom industry has grown from a tele-density of 3.58% in March 2001 to 74% in June 2013. This great leap in both number of consumers as well as revenues from telecom services has not only provided sufficient contribution in Indian GDP growth but also provided much needed employment to India youth.

Giving a definite boost to its image as the State that stays “connected”, Kerala has achieved the highest teledensity rate among the states in India with an overall teledensity of 96.74%. Because of such high market potential the competition between service providers in the state is high. The main players in the telecom industry in Kerala are BSNL, Idea, Airtel, Vodafone, Reliance, Tata Docomo etc.

Customer satisfaction is a term frequently used in  marketing, it is a measure of how products and services supplied by a company meet or surpass customer expectation. Customer satisfaction is defined as “the number of customers, or percentage of total customers, whose reported experience with a firm, its products, or its services exceeds specified satisfaction goals. In a competitive marketplace where businesses compete for customers, customer satisfaction is seen as a key differentiator and increasingly has become a key element of business strategy. Therefore, it is essential for businesses to effectively manage customer satisfaction. To be able do this, firms need reliable and representative measures of satisfaction.

Sometimes companies are misguided by the notion that customers depend on them.  The truth of the matter is that we very much so depend on them. The level of satisfaction a customer has with a company has profound effects. The level of customer’s satisfaction has a positive effect on profitability:

* A totally satisfied customer contributes 2.6 times as much revenue t oa company as a somewhat satisfied customer.
* A totally satisfied customer contributes 17times as much revenue as a somewhat dissatisfied customer.
* A totally satisfied customer contributes 17times as much revenue as a somewhat dissatisfied customer.
* A totally dissatisfied customer decreases revenue at a rate equal to 18 times what a totally satisfied customer contributes to a company.

When a person is satisfied with a company or service they are likely to share their experience with other people to the order of perhaps five or six people.  However, dissatisfied customers are likely to tell another ten people of their unfortunate experience.

**1.2 OBJECTIVE**

The project aims at studying the present market scenario of Telecom Service in Trivandrum city. The major players in the market are (i) BSNL (ii) Airtel (iii) Idea (iv) Vodafone (v) other service providers viz. Reliance, Tata DoCoMo.

The main objectives of the study are:-

(i) To study the customer satisfaction towards mobile service providers.

(ii) To predict the customer satisfaction based on certain key variables.

(iii) To study and understand the customer behaviour.

(iv) To understand the performance of different brands in the market on various parameters like Network Accessibility, Tariffs,Billing/Recharge Services, Package offers and Customer Satisfaction.

**1.3 PROBLEM SCENARIO**

The introduction of the wireless telecommunication has changed the trend of the telecommunication industry in Trivandrum city to become highly competitive. Customers have the freedom to choose and switch from one package to another package offered by the service provider, or from one service provider to another service provider. Thus, customer churn has become a major concern that is now harassing network providers in Trivandrum which has a large population. Through research, it is found that the cost of acquiring new customers is much higher than retaining the existing ones (Ali Tamaddoni Jahromi, 2009; Huang Ming, Niu Wenying, Liang Xu, 2009). Hence it is important for service providers to ensure that the customers are satisfied with their products. “Data mining in telecommunications industry to study consumer behaviour” aims at assessing the satisfaction of customers based on certain key parameters and also helps to predict the same.

**CHAPTER 2**

**2.1LITERATURE REVIEW**

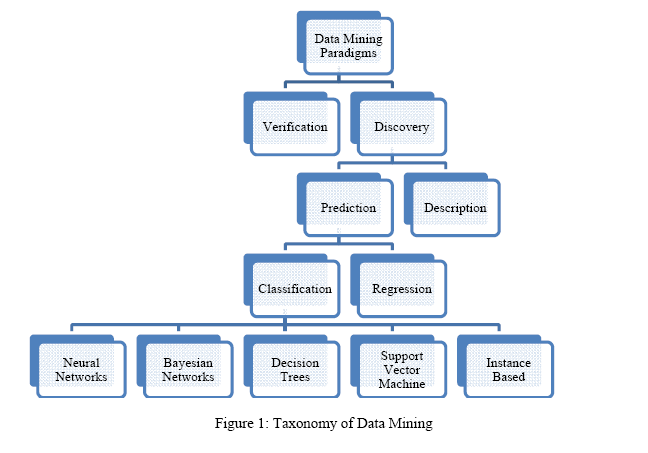
**2.1.1 DATA MINING**

High growth of data in databases created a need for technologies which can extract and uncover the hidden information in large amount of data which can be useful in decision making in business organizations. Data mining is a technology that could solve this problem with approach combining machine learning, statistics and database management that are used for finding useful and valid patterns in data. Data mining methodology is used for analyzing large amounts of data for secondary analysis by extracting information from operating application systems. It can be used for finding and identifying relationships among data (Wu et al., 2012). Data mining is one of the most popular processes of analysing huge amount of data to find relationships among them with the goal to gain knowledge and make decisions (Ngai et al., 2009).

The main goal of data mining is to provide insight into disorganized information in order to enhance business knowledge and future business activities for the user in business organizations. Data mining techniques can be categorized into two groups: descriptive (verification) and predictive (Lejeune, 2001). Descriptive data mining techniques are used for better understanding of the data, while predictive data mining techniques are used for forecasting and

devising.

It usually involves four classes of tasks which are the classification, clustering, regression and association rule learning [Gary Cokins and Ken King]. The taxonomy of data mining methods can be summarized as Figure 1 (MO Zan, ZHOA Shan, LI Li and LIU Ai-Jun, 2007).



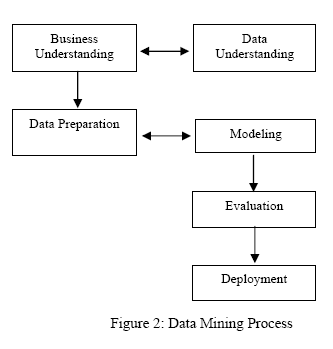
Description (Verification) method deals with evaluation of a hypothesis proposed by an external source (i.e. an expert etc.). This method includes the traditional statistical methods such as the goodness-of-fit test, t-test of means, and analysis of variance. These methods are less associated with data mining than their discovery oriented counterparts because most data mining problems are concerned with selecting a hypothesis (out of a set of hypotheses) rather than testing a known one.

On the other hand, discovery methods are methods that identify patterns in the data automatically. The discovery method branch consists of prediction methods and description methods. Description oriented data mining methods focus on understanding the way the underlying data operates while the prediction-oriented methods aim to build a behavioral model that can get newly and unseen samples and is able to predict values of one or more variables related to the sample. The classification methods arrange data into predefined group and can be divided into five (5) branches; neural networks, Naïve Bayesian networks, decision trees, support vector machine and instance based methods.

2.1.1.1 THEARCHITECHTURE/PROCESS IN DATA MINING

Data mining methodology consists of ten (10) processes; translate a business problem into a data mining problem, select appropriate data, get to know the data, create a model set, fix the problems with the data, transform data, build models, assess models, deploy models and assess results. However, these steps can be further summarized into five main stages:

The initial exploration incorporates with business and data understanding, data preparation, model building, evaluation and deployment. These are the entire steps taken to generate predictions. Figure 2 shows the flow chart of standard processes for data mining.



i. Business and Data Understanding

Here, the objectives and goals are defined by understanding the problems to obtain clearer

objectives [Huang Ming, Niu Wenying and Liang Xu, 2009]. Next, the type or criteria of data

which is available to us is determined in order to solve our objectives need. Data selected

should represent enough quantity of data in a given period of time. Then, the hidden trends in

data are discovered by finding the relation of the data selected [Shyam V.Nath and Dr. Ravi S.

Behara, 2005].

ii. Data Preparation

Data preparation is the most time-consuming phase in data analysis or data mining processes. In data preparation phase,data is collected, integrated and cleaned. Integration of data may require extracting from multiple sources. Once, the data is in a tabular format, it should be fully characterized. Data needs to be cleaned by resolving any ambiguities, errors, and removing redundant and problematic data. Data which clearly do not contribute in the analysis are removed. Finally, the table should be divided, where appropriate, into subsets in order to optimize the performance of the database, simplify the analysis and allow specific queries to be

performed easily.

iii. Modeling Phase

In this phase, we are required to develop a model for future prediction and select appropriate modeling technique to suite with the main objectives and evaluate them in the next phase. Once the model has been selected, different parameters are obtained to make improvement on the

results.

iv. Evaluation Phase

This phase is the vital stage in CRISP-DM (Cross Industry Standard Process- Data Mining) model. Evaluation on the response time, confidence level, cost, error rate and the usefulness ofthe model in achieving the objectives and goals previously defined is done.

v. Deployment

This stage deploys the results through the objectives and goals previously defined at the initial

process.

2.1.1.2 DESCRIPTIVE STATISTICS

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of information, or the quantitative description itself. Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example in a paper reporting on a study involving human subjects, there typically appears a table giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, and the proportion of subjects with related comorbidities.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables, kurtosis and skewness.

Descriptive statistics provides simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation.

The use of descriptive and summary statistics has an extensive history and, indeed, the simple tabulation of populations and of economic data was the first way the topic of statistics appeared. More recently, a collection of summarisation techniques has been formulated under the heading of exploratory data analysis: an example of such a technique is the box plot.

In the business world, descriptive statistics provides a useful summary of many types of data. For example, investors and brokers may use a historical account of return behavior by performing empirical and analytical analyses on their investments in order to make better investing decisions in the future.

2.1.1.3 PREDICTIVE MODELLING

Predictive modelling leverages statistics to predict outcomes. Predictive modelling is a process used in predictive analytics to create a statistical model of future behaviour. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. In many cases the model is chosen on the basis of decision theory to try to guess the probability of an outcome given a set amount of input data. Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set. Depending on definitional boundaries, predictive modelling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts.

In this project we have explicitly asked the respondents about the satisfaction with their service providers and then based on the assumption that the dissatisfied customers do certainly change and the satisfied customers continue to be associated with their service provider, we create a predictive model to predict the customer satisfaction. The decision tree algorithm is used to create the predictive model.

DECISION TREE

Decision tree learning is one of the predictive modelling approaches used in statistics ,data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data.

Data comes in records of the form:

(\textbf{x},Y) = (x_1, x_2, x_3, ..., x_k, Y)

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **x** is composed of the input variables, x1, x2, x3 etc., that are used for that task.

Now we describe some of the practical hurdles involved and how trees are sometimes used to overcome them.

Variable selection

Data often arrive at the analyst's door with lots of variables. The baggage sometimes includes a dictionary that makes uninteresting reading. Yet the analyst must find something interesting in the data. Most of the variables are redundant or irrelevant and just get in the way. A preliminary task is to determine which variables are likely to be predictive.

A common practice is to exclude input (independent) variables with little correlation to the target (dependent) variable.

An alternative practice is to use inputs that appear in splitting rules of trees. Trees notice relationships from the interaction of inputs. The analyst would typically use the selected variables as the inputs in a model such as logistic regression. In practice trees often provide far fewer variables than seem appropriate for a regression. The sensible solution is to include some variables from another technique, such as correlation. No single selection technique is capable of fully prophesizing which variables will be effective in another modeling tool.

Variable importance

The analyst may want the variable selection technique to provide a measure of importance for each variable, instead of just listing them. This would provide more information for modifying the selected list to accommodate other considerations of the data analysis. Intuitively, the variables used in a tree have different levels of importance.

A formulation of variable importance that captures this intuition is as follows: (1) Measure the importance of the model for predicting an individual. Specifically, let the importance for an individual equal the absolute value of the difference between the predicted value (or profit) of the individual with and without the model. (2) Divide the individual importance among the variables used to predict the individual, and then (3) average the variable importance over all individuals.

Interaction detection

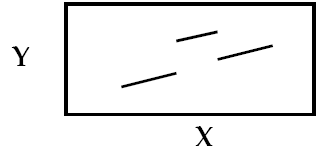
Having selected inputs to a regression, the analyst will typically consider possible interaction effects.

Consider a data set containing one target Y and one input X that are related through the equation

Y = a X + b W

where a and b are constants, and W is an indicator: it takes values 0 and 1, it is omitted from the data set, but it is 1 if and only if X is in some interval. The analyst must infer the influence of W on Y using only X.

The diagram illustrates the situation in which W equals one when X is near the middle of its range. The tree creation algorithm searches for a split on X that best separates data with small values of Y from large ones. If b is large enough, values of Y with W = 1 will be larger than values of Y with W = 0, and the algorithm will make the first two splits on X so as to isolate data with W = 1. A tree with three leaves appears in which one leaf represents the desired indicator variable.



When b is smaller so that data with extreme values of Y have W = 0, the data with W = 1 have less influence on the algorithm. A split that best separates extreme values of Y is generally different from a split that best detects interactions in linear relationships. A tree creation algorithm specially designed for discovering interactions is needed. Otherwise, even if the W = 1 data is eventually separated, the tree will have many superfluous leaves.

Stratified modeling

The analyst preparing for a regression model faces another hidden pitfall when the data represent two populations. A different relationship between an input and the target may exist in the different populations. In symbols, if Y = a + b X + e and Y = a + c X + e express the relationship of Y to X in the first and second population respectively, and b is very different from c, then one regression alone would fit the combined data poorly.

The problem for the analyst is to recognize the need to perform two regressions instead of one. For this purpose, some analysts first create a small decision tree from the data, and then run a separate regression in each leaf. This is called stratified regression.

Missing value imputation

The analyst may have to contend with missing values among the inputs. Decision trees that split on one input at a time are more tolerant to missing data than models such as regression that combine several inputs. When combining several inputs, an observation missing any input must be discarded. For the simplest of tree algorithms, only observations that need to be excluded are those missing the input currently being considered to split on. They can be included when considering splitting on a different input.

Trees may be the best modeling tool for this purpose because of their tolerance to missing data, their acceptance of different data types, and their robustness to assumptions about the input distributions.

Model interpretation

Trees are sometimes used to help understand the results of other models.

SOFTWARE PACKAGES

Commonly used software packages include R, Rapidminer, IBM SPSS Modeler, SAS Enterprise miner, WEKA etc.

**2.1.2 R**

R is a [programming language](https://en.wikipedia.org/wiki/Programming_language) and software environment for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics. The R language is widely used among [statisticians](https://en.wikipedia.org/wiki/Statistician) and [data miners](https://en.wikipedia.org/wiki/Data_mining) for developing [statistical software](https://en.wikipedia.org/wiki/Statistical_software) and data analysis. Polls, [surveys of data miners](https://en.wikipedia.org/wiki/Rexer%27s_Annual_Data_Miner_Survey), and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

R is an implementation of the [S programming language](https://en.wikipedia.org/wiki/S_(programming_language)) combined with [lexical scoping](https://en.wikipedia.org/wiki/Lexical_scoping) semantics inspired by [Scheme](https://en.wikipedia.org/wiki/Scheme_(programming_language)). [S](https://en.wikipedia.org/wiki/S_(programming_language)) was created by [John Chambers](https://en.wikipedia.org/wiki/John_Chambers_(programmer)) while at [Bell Labs](https://en.wikipedia.org/wiki/Bell_Laboratories). There are some important differences, but much of the code written for S runs unaltered.

R is an integrated suite of software facilities for data manipulation, calculation and graphical

display. It has 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 directly at the computer or on hardcopy.

R and its libraries implement a wide variety of statistical and [graphical](https://en.wikipedia.org/wiki/Graphical) techniques, including [linear](https://en.wikipedia.org/wiki/Linear) and [nonlinear](https://en.wikipedia.org/wiki/Nonlinear) modeling, classical statistical tests, [time-series analysis](https://en.wikipedia.org/wiki/Time-series_analysis), classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, [C](https://en.wikipedia.org/wiki/C_(programming_language)), [C++](https://en.wikipedia.org/wiki/C%2B%2B), and [Fortran](https://en.wikipedia.org/wiki/Fortran" \o "Fortran)code can be linked and called at run time.

R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger [object-oriented Programming](https://en.wikipedia.org/wiki/Object-oriented_programming) facilities than most statistical computing languages. Extending R is also eased by its [lexical scoping](https://en.wikipedia.org/wiki/Lexical_scoping) rules. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages.[[20]](https://en.wikipedia.org/wiki/R_(programming_language)#cite_note-20)

**2.1.3 IBM SPSS**

SPSS is a comprehensive and flexible statistical analysis and data management solution. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses.  SPSS is a widely used program for [statistical analysis](http://en.wikipedia.org/wiki/Statistics) in [social science](http://en.wikipedia.org/wiki/Social_science). It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners,[[3]](http://en.wikipedia.org/wiki/SPSS" \l "cite_note-3) and others.

SPSS Statistics Base provides essential statistical analysis tools for every step of the analytical process.

* A comprehensive range of statistical procedures for conducting accurate analysis.
* Built-in techniques to prepare data for analysis quickly and easily.
* Sophisticated reporting functionality for highly effective chart creation.
* Powerful visualization capabilities that clearly show the significance of your findings.
* Support for all types of data including very large data sets.

 In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation are features of the base software.

Statistics included in the base software:

* [Descriptive statistics](http://en.wikipedia.org/wiki/Descriptive_statistics): [Cross tabulation](http://en.wikipedia.org/wiki/Cross_tabulation), [Frequencies](http://en.wikipedia.org/wiki/Statistical_frequency), Descriptives, Explore, Descriptive Ratio Statistics
* Prediction for numerical outcomes: [Linear regression](http://en.wikipedia.org/wiki/Linear_regression)
* Prediction for identifying groups: [Factor analysis](http://en.wikipedia.org/wiki/Factor_analysis), [cluster analysis](http://en.wikipedia.org/wiki/Cluster_analysis) (two-step, [K-means](http://en.wikipedia.org/wiki/K-means_clustering), [hierarchical](http://en.wikipedia.org/wiki/Hierarchical_clustering)), [Discriminant](http://en.wikipedia.org/wiki/Discriminant_analysis_(in_marketing)" \o "Discriminant analysis (in marketing))

The many features of SPSS Statistics are accessible via [pull-down menus](http://en.wikipedia.org/wiki/Pull-down_menus) or can be programmed with a proprietary [4GL](http://en.wikipedia.org/wiki/4GL)*command syntax language*. Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax and are not accessible through the menu structure. The pull-down menu interface also generates command syntax: this can be displayed in the output.

SPSS Statistics places constraints on internal file structure, [data types](http://en.wikipedia.org/wiki/Data_types), [data processing](http://en.wikipedia.org/wiki/Data_processing), and matching files, which together considerably simplify programming. SPSS datasets have a two-dimensional table structure, where the rows typically represent cases (such as individuals or households) and the columns represent measurements (such as age, sex, or household income). Only two data types are defined: numeric and [text](http://en.wikipedia.org/wiki/String_(computer_science)) (or "string"). All data processing occurs sequentially case-by-case through the file. Files can be matched one-to-one and one-to-many, but not many-to-many.

The [graphical user interface](http://en.wikipedia.org/wiki/Graphical_user_interface) has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS Statistics window. The 'Data View' shows a [spreadsheet](http://en.wikipedia.org/wiki/Spreadsheet) view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. The 'Variable View' displays the metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as [statistical surveys](http://en.wikipedia.org/wiki/Statistical_survey) are more often created in [data entry](http://en.wikipedia.org/w/index.php?title=Data_entry_program&action=edit&redlink=1) software, or entered during [computer-assisted personal interviewing](http://en.wikipedia.org/wiki/Computer-assisted_personal_interviewing), by scanning and using [optical character recognition](http://en.wikipedia.org/wiki/Optical_character_recognition) and [optical mark recognition](http://en.wikipedia.org/wiki/Optical_mark_recognition) software, or by direct capture from [online questionnaires](http://en.wikipedia.org/wiki/Online_questionnaires). These datasets are then read into SPSS.

SPSS Statistics can read and write data from [ASCII](http://en.wikipedia.org/wiki/ASCII) text files (including hierarchical files), other statistics packages, [spreadsheets](http://en.wikipedia.org/wiki/Spreadsheets) and [databases](http://en.wikipedia.org/wiki/Databases). SPSS Statistics can read and write to external [relational database tables](http://en.wikipedia.org/wiki/Table_(database)) via [ODBC](http://en.wikipedia.org/wiki/ODBC) and [SQL](http://en.wikipedia.org/wiki/SQL).

Statistical output is to a [proprietary file format](http://en.wikipedia.org/wiki/Proprietary_format) (\*.spv file, supporting [pivot tables](http://en.wikipedia.org/wiki/Pivot_table)) for which, in addition to the in-package viewer, a stand-alone reader can be downloaded. The proprietary output can be exported to text or [Microsoft Word](http://en.wikipedia.org/wiki/Microsoft_Word), PDF, Excel, and other formats. Alternatively, output can be captured as data (using the OMS command), as text, tab-delimited text, [PDF](http://en.wikipedia.org/wiki/PDF), [XLS](http://en.wikipedia.org/wiki/Microsoft_Excel#File_formats), [HTML](http://en.wikipedia.org/wiki/HTML), [XML](http://en.wikipedia.org/wiki/XML), SPSS dataset or a variety of graphic image formats ([JPEG](http://en.wikipedia.org/wiki/JPEG), [PNG](http://en.wikipedia.org/wiki/Portable_Network_Graphics), [BMP](http://en.wikipedia.org/wiki/Windows_and_OS/2_bitmap) and [EMF](http://en.wikipedia.org/wiki/Windows_Metafile)).

**2.4 BEHAVIOURAL ANALYSIS OF THE DIFFERENT SEGMENTS TO PROPOSE MARKETING RECOMMENDATIONS FOR MOBILE OPERATORS**

(Journal Referred: Exploring consumer adoption of new services by analyzing the behavior of 3G subscribers: An empirical case study By *Li-Chen Cheng and Li-Min Sun*).

Diverse mobile applications and services undoubtedly influence the daily life of mobile users around the world significantly. Furthermore, the maturing of mobile technologies effectuates e-commerce over mobile platforms. Though the number of 3G subscribers is raising rapidly, average mobile revenue per user (ARPU) is dropping. Mobile operators must manage nonvoice

ARPU more effectively to overcome this business challenge (Kuo and Yen 2009, Deng et al. 2010). To fulfill this requirement, the telecommunications industry is constantly developing new

and innovative 3G value-added services to meet the various needs of consumers. Although a variety of mobile value-added services have been released, whether consumers will purchase these services remains unknown (Teng et al. 2009). Understanding customer-purchasing behavior has become a key issue.

3G operators employing objective data analysis to identify heavy users of new products and investigate the behavioral differences between heavy and non-heavy users has become a critical exercise (Koivumaki et al. 2006, Schierza et al. 2010). These results will support the development of different consumption behavior models, providing differentiated products and services regarding marketing and maintenance of customer relationships, which will significantly enhance future marketing strategies (Wu and Wang 2005).

Data mining generally entails technologies discovering previously unknown information and summarizing relevant information from a vast number of databases to assist business decisions

(Cabana et al. 2008). Appropriate customer relationship management (CRM) strategies can be adopted with the assistance of data mining technologies, which can manage the data required to enhance understanding of customers (Ngai 2005).

Two main contributions of the study: from a conceptual viewpoint, this study proposes a comprehensive CRM strategy framework that contains a customer segmentation process and a

behavior analysis process.

The most critical customer strategy should comprise examining the existing and potential customer base and identifying which forms of segmentation is most appropriate. Upon conducting customer segmentation, enterprises could focus on the character of different user groups to propose more efficient operations when acquiring and retaining the relationship to increase overall business profits (Payne and Frow 2005, Thomas and Sullivan 2005). The objective of these strategic operations is to build long-term relationships for increasing customer satisfaction, strengthening customer loyalty, and raising profitability (Swift 2000, Mithas et al. 2005).

Association rule is the best-known technique for customer purchase analysis. Liu applied association rules to extract knowledge from customer purchase history for the development of one-to-one marketing (Liu and Shih 2005). The method has been widely used in various areas, such as for mining user access patterns on websites, using POS information to extract interorganizational retailing knowledge, recommending products to users, and cross sailing (Berson et al. 2000, Lin et al. 2003, Sohn and Kim 2003, Liu and Shih 2005, Chen et al. 2009).

Customer segmentation and RFM models Customer segmentation divides all customers into an appropriate number of clusters to effect customized strategies for meeting different customer needs. Early approaches for segmentation include the use of demographic, geographic, situation (Gehrt and Shim 2003), lifestyle and behavioral characteristics of consumers (Plummer 1974, Assael and Roscoe 1976, Punj and Stewart 1983, Hoek et al. 1996, Schijns and Schroder 1996, Gehrt and Shim2003, Bose and Chen 2010). However, Gupta et al. (2006) suggests that the past purchases of consumers are superior predictors for determining future purchase behavior compared to demographics.

Association rules

Association rule mining is another major data mining technique mostly used to discover multiple independent elements that frequently co-occur, and seek to extract rules in association with

the co-occurring elements in a given dataset (Agrawal et al. 1993). A well-known application of association rules is market basket analysis, which contains a customer’s purchasing transactions,

that is, the itemsets that are purchased by a customer in a single transaction. The number of customer transactions is typically quite high, and the number of items in frequent itemsets can increase exponentially. Association rules can be used to examine as many frequent itemsets as possible, answering queries such as what products tend to be purchased together. Association rules can reduce a large amount of information to a small and more comprehensive set of statistically supported statements (Han and Kamber 2007). A typical association rule assumes the form A ? B, where A and B are in an item set that contains only a single atomic condition. A rule also can be represented as an antecedent (left-hand side) and a consequent (right-hand side). The intersection between the antecedent and the consequent is empty. The support of an association

rule is the percentage of records containing items A and B together. The confidence of a rule is the percentage of records containing item A that also contains item B. The confidence represents the strength of the rule, and the support indicates the frequency of the patterns occurring in the rule. Rules with high confidence and robust support are referred to as strong rules.

The proposed framework completely maps the customer strategy process, customer segmentation versus segment granularity, rules extraction versus customer characteristics. Thus, our results can be beneficial in the value creation and multichannel integration process. The aim of this study is to produce a strategy for developing marketing and product recommendations based on real data generated by customers, in this case mobile operators. These providers have to continually meet new customer demands and be able to dynamically shift their business strategies in order to gain a competitive advantage among intense rivalries. The analysis of the consumer behaviors of valuable customer segments is a starting point to accomplish this goal. We construct a model for clustering customers based on some special attributes and then analyze customer behaviors by finding association rules from each segment.

Clusters will be analyzed based on their use of 3G services and association rules. Rules generated from each segment will be inspected for diverse levels of support and confidence. Implicit rules reveal what kinds of services are usually used together. The association rules that emerged from the experimental dataset will be summarized in the next section. There may be thousands of rules related to hundreds of products output by the data mining tool; such rules become complicated and are not easy to interpret. The only thing we can do is to reduce the number of rules without altering their essence.

The whole process of behavior comparison can be divided into two steps, including intra-cluster analysis that can help to understand specific cluster characteristics, and inter-cluster analysis

for discussing the differences for one-on-one behavioral comparison.

Intra-cluster analysis.

At the beginning, we discover the general rules within each cluster that have the same support value. Then, the dependency network can be determined to illustrate the relationships among five product categories. Among these general rules, we filter out some interesting rules for marketing purposes. Changed rules show the implicit patterns that target products are consequents under different support values. We want to discover significant rules that change as the support values decrease, such as the change from A?B to C?B. These patterns can describe various services associated with new services. Otherwise, the trend for adopting services appears while adjusting the support values.

Inter-cluster analysis

In this study, customers are divided into three groups, based on high, medium and low values. The goal of the analysis is to discover meaningful market knowledge. Inter-cluster analysis may be used to generate special patterns for unexpected rules and inter-cluster changing rules. The unexpected rules are filtering the changing rules among clusters with the same support value for

designing suitable marketing packages. Finally, we observe changes in consumption behaviors among clusters under different support values. The changes in adopting specific services among

the three clusters may be found.

**2.4 CUSTOMER BEHAVIOUR**

(Journal Referred: A Study of Customer satisfaction on Telecom Service Providers By Dinesh Kumar Pandiya , Dr Brajesh Kumar and Mazahidul Haque Choudhury ,ASSAM UNIVERSITY, SILCHAR, Department of Commerce)

Customer satisfaction is a term frequently used in  marketing, it is a measure of how products and services supplied by a company meet or surpass customer expectation. Customer satisfaction is defined as “the number of customers, or percentage of total customers, whose reported experience with a firm, its products, or its services exceeds specified satisfaction goals. In a competitive marketplace where businesses compete for customers, customer satisfaction is seen as a key differentiator and increasingly has become a key element of business strategy. Therefore, it is essential for businesses to effectively manage customer satisfaction. To be able do this, firms need reliable and representative measures of satisfaction.

Business always starts and closes with customers and hence the customers must be treated as the King of the market. All the business enhancements, profit, status, image etc. of the organization depends on customers. Hence it is important for all the organizations to meet all the customers’ expectations and identify that they are satisfied customer. Customer satisfaction is the measure of how the needs and responses are collaborated and delivered to excel customer expectation. It can only be attained if the customer has an overall good relationship with the supplier. In today’s competitive business marketplace, customer satisfaction is an important performance exponent and basic differentiator of business strategies. Hence, the more is customer satisfaction; more is the business and the bonding with customer.

It is necessarily required for an organization to interact and communicate with customers on a regular basis to increase customer satisfaction. In these interactions and communications it is required to learn and determine all individual customer needs and respond accordingly. Even if the products are identical in competing markets, satisfaction provides high retention rates.

Higher the satisfaction level, higher is the sentimental attachment of customers with the specific brand of product and also with the supplier. This helps in making a strong and healthy customer-supplier bonding. This bonding forces the customer to be tied up with that particular supplier and chances of defection very less. Hence customer satisfaction is very important panorama that every supplier should focus on to establish a renounced position in the global market and enhance business and profit.

Importance of Customer Satisfaction

Sometimes companies are misguided by the notion that customers depend on them.  The truth of the matter is that we very much so depend on them. The level of satisfaction a customer has with a company has profound effects. The level of customer’s satisfaction has a positive effect on profitability:

• A totally satisfied customer contributes 2.6times as much revenue to a company as a somewhat satisfied customer.

• A totally satisfied customer contributes 17times as much revenue as a somewhat dissatisfied customer.

• A totally satisfied customer contributes 17times as much revenue as a somewhat dissatisfied customer.

• A totally dissatisfied customer decreases revenue at a rate equal to 18 times what a totally satisfied customer contributes to a company.

Method to Measure Customer Satisfaction

Companies use the following methods to measure  customer satisfaction.

1) Complaints and suggestion system: Companies obtaining complaints through their  customer service centers, and further suggestions were given by customers to satisfy their desires.

2) Customer satisfaction surveys: Responsive companies obtain a direct measure of customer satisfaction  by periodic surveys. They send questionnaires to random sample of their customers to find out how they feel about various aspects of the company’s performance and also  solicit views  on their competitor’s performance. It is useful to measure the customer’s willingness to recommend the company and brand to other persons.

3) Lost Customer Analysis: Companies should contact customers who have stopped buying or who have switched to another supplier to learn why this happened.

4) Consumer Behavior V/s Consumption Behavior:

**CHAPTER -3**

**3.1 DATA COLLECTION**

The objective of most marketing research projects is to obtain information about the characteristics or parameters of a population. Sample characteristics, called statistics, are then used to make inferences about the population parameters.

For the purpose of the study, during the initial stages we met with a few qualified personnel from all the major service providers in the city. Based on their valuable inputs a questionnaire was created in such a way that all the major factors required to assess the customer behaviour were included.

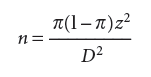
3.1.1 QUESTIONNAIRE

The questionnaire consisted of 19 questions of which 2 were dichotomous, 5 were scalar and 12 were multiple choice questions in nature. It dealt with questions ranging from the age, gender and occupation of the subject to rating of different parameters of selection of service providers like network accessibility, call rates etc.

3.1.2 SAMPLE SIZE DETERMINATION

Several qualitative factors should also be taken into consideration when determining the sample size . These include the importance of the decision, the nature of the research, the number of variables, the nature of the analysis, sample sizes used in similar studies, incidence rates (the occurrence of behaviour or characteristics in a population), completion rates and resource constraints. The statistically determined sample size is the net or final sample size: the sample remaining after eliminating potential respondents who do not qualify or who do not complete the interview.Depending on incidence and completion rates, the size of the initial sample may have to be much larger.

Since the sample statistic in this case, i.e the number of persons having smartphone access cannot be determined in exact magnitude, a proportion based on the market penetration of smartphones in Trivandrum (source: Telecom Regulatory Authority of India) was used for calculating the initial sample size by the following equation:



Where Π = population proportion

Z= z value at a particular % confidence interval)

D = desired precision

The the finite

population correction (fpc) is now applied since the population proportion in greater than 10%. The required sample size is then calculated from the formula:

C:\Users\user\Desktop\ashokproject\pics\awd.PNG

Where N denotes the total population,

denotes sample size with fpc

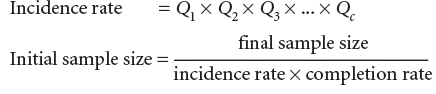
n denotes saple size without fpc.

The sample size is then adjusted statistically based on the inclusion of the incidence rate and the completion rate.

Incidence rate refers to the rate of occurrence or the percentage of persons eligible to participate in the study. Incidence rate determines how many contacts need to be screened for a given sample size requirement.

Similarly, the determination of sample size must take into account anticipated refusals by people who qualify. The completion rate denotes the percentage of qualified

respondents who complete the interview.



3.1.3 COLLECTION OF DATA

Data was collected from a sizeable sample whose size was found to be 554. The sample was chosen by convenience. For the purpose of data collection both online and offline survey methods were employed. According to the census report of the population of Trivandrum city, it could be understood that approximately 70% of the population aged between 20-60 years belong to the group of 20-40 years. This age category of 20-40 years also witnesses the highest percentage of internet users (approx. 55%). Hence a larger proportion of the survey was done online.

360 respondents were obtained through online survey which was conducted through social media. 194 respondents were approached by offline survey in places that are generally considered to have a heterogeneous mixture of people.

3.1.4 RELIABILITY OF THE QUESTIONNAIRE

It is necessary to check whether the survey instrument is reliable. In this case the survey instrument is a questionnaire. To assess the reliability, Cronbach’s α statistic was calculated. The reliability coefficient or the Cronbach’s Alpha value lies between 0 and 1.

|  |  |  |
| --- | --- | --- |
| Cronbach’s alpha | Cronbach’s alpha with standardization of items | Number of items |
| 0.713 | 0.714 | 11 |

The value 0 indicates an instrument with full of error. While the value 1 indicates an instrument with total absence of error. An alpha value greater than 0.65 is considered as acceptable reliability. In our scenario, the Cronbach’s Alpha value is equal to 0.713, which is greater than 0.65. Therefore the questionnaire is reliable.

**3.2 DATA ANALYSIS**

3.2.1 INTRODUCTION

Having a large amount of data does not become useful unless you derive useful information from it. Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data Analysis is thus defined as procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

3.2.2 DESCRIPTIVE STATISTICS

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. It is the discipline of quantitatively describing the main features of a collection of information, or the quantitative description itself. Descriptive statistics provides simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative , i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation.

In this project, IBM SPSS Statistics software was used to describe, summarize and identify patterns from the data collected.

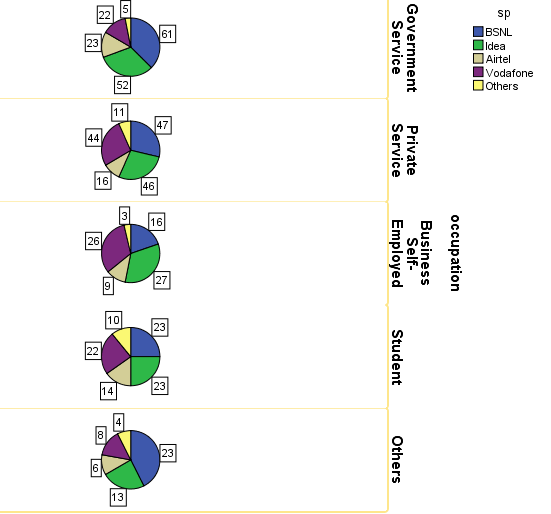


Figure 3: Occupation vs service providers

The above figure indicates which all service providers are preferred by the people working in different sectors. The above statistics indicate that BSNL dominates the industry with it being the most favoured service provider in most of the cases. After BSNL, Idea is the next most preferred service provider and it even surpasses BSNL in case of Business/Self-employed sector.

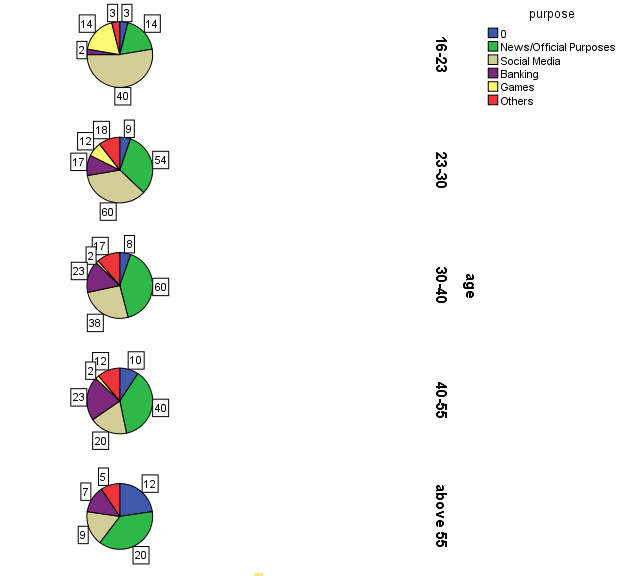


Figure 4: Age vs Purpose

The above pie-charts show for what all purposes people of different age groups use the mobile internet. People in the age group 16-23 use the internet most for social media. People in the age group 23-30 use mobile internet for Social media as well closely followed by news/official purposes. In the age group 30-40, People mostly prefer news/official purposes. In the age group 40-55, again people prefer to use mobile internet for news/official purposes mostly and the same trend is shown in the above 55 age group.

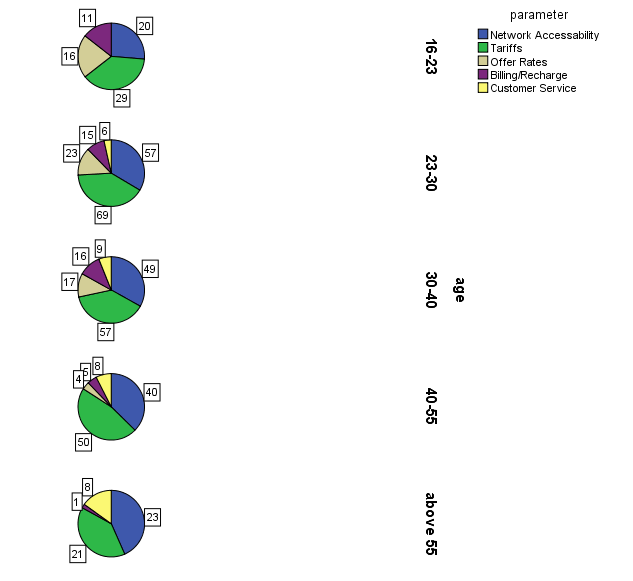


Figure 5: Age vs Parameter

The above series of pie charts denote the parameter which people give most importance to while selecting service provider. From the above figures we can see see that tariff is the most important parameter for all the age groups except those above 55.

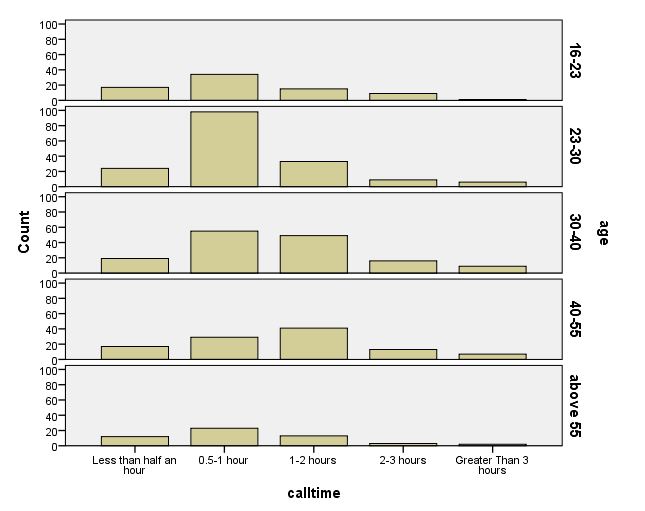


Figure 6: Calltime vs Age

From the above figure, it is clear that people tend to call between half an hour to one hour daily.

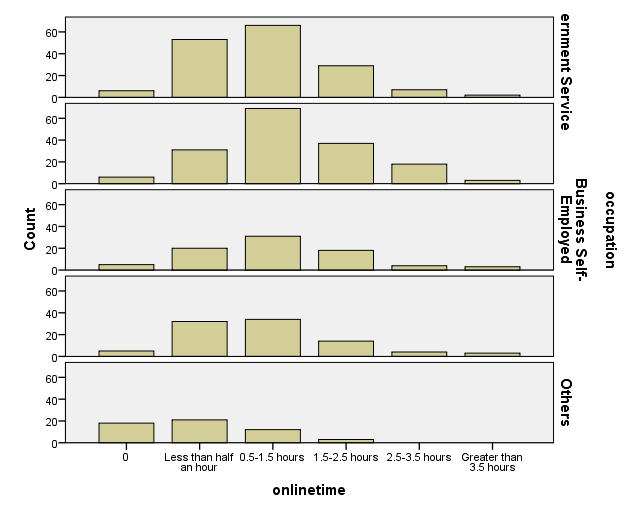


Figure 7: Onlinetime vs Occupation

The above figure shows that people working in almost all the sectors tend to spend half an hour to 1.5 hours online daily

Table 1: Average Rating of service providers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Service Providers | RNA | RBR | RT | RO | RCS |
| BSNL | 3.58 | 3.34 | 2.85 | 3.01 | 3.42 |
| Idea | 3.49 | 3.22 | 2.84 | 2.98 | 3.42 |
| Airtel | 3.44 | 2.79 | 3.03 | 3.25 | 3.44 |
| Vodafone | 3.16 | 3.13 | 3.35 | 3.49 | 3.56 |
| Others | 2.97 | 3.52 | 2.61 | 2.70 | 2.48 |

The table given above depicts the overall customer satisfaction levels of the major telecom service providers in Trivandrum city. From the table it can be seen that of all the service providers, BSNL has a high rating of 3.58 on 5 for the criterion of network accessibility followed by Idea (3.49).Vodafone has a rating of 3.16 and enjoys the lowest rating among the service providers. For the criterion of Billing/Easy of recharge also BSNL has the higher rating(3.34) among the four providers and Airtel has the lower rating(2.79).For the other three criterions, vodafone has the higher rating.

3.2.3 ANALYSIS USING R

R is an integrated suite of software facilities for data manipulation calculation and graphical display. It has 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 directly on computer or on hard copy.

In this project, R is used to implement discovery based data mining techniques. The main techniques used regarding discovery based data mining are Cluster analysis and Associate Rule Mining.

1.Cluster Analysis

The clustering technique is one of the core tools that is used by the data miner. Clustering gives us the opportunity to group observations in a generally unguided fashion according to how similar they are.This is done on the basis of a measure of the distance between observations.Clustering allows the data miner to break data into more meaningful groups and then contrast the different clusters against each other. Clusters can also be useful in grouping observations to help make the smaller datasets easier to manage. The aim of clustering is often to identify groups of observations that are close together but as a group are quite separate from other groups.

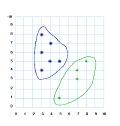


Figure 8: Clustering

K-means Clustering

The K-means clustering was used to implement the Cluster analysis. The K-means clustering is a simple and computationally efficient technique used popularly for Cluster analysis.A model built using the k-means algorithm represents the clusters as a collection of k means. The observations in the dataset are associated with their closest \mean" and thus are partitioned into k clusters. The mean of a particular numeric variable for a collection of observations is the average value of that variable over those observations. The means for the collection of observations that form one of the k clusters in any particular clustering are then the collection of mean values for each of the input variables over the observations within the clustering.

The R code for the k-means clustering is shown below



The first line of code implements the K-means algorithm on the data file projectdata.csv. The number 5 denotes the number of clusters into which the entire number of observations are divided. The second line is the code to get the output of the clustering process.

The results of the k-means clustering is shown in below figure

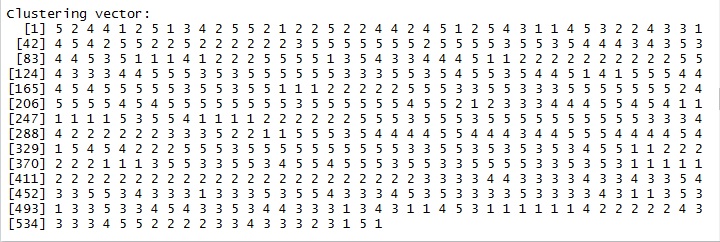


Figure 9: Clustering Vector

This clustering vector shows in which all clusters the observations are placed. The first observation is placed in cluster 5, the second observation is placed in cluster 2, third observation in cluster 4 and so on. The sizes of each cluster are 58,99,125,83 and 189 for cluster 1, cluster 2, cluster 3, cluster 4, cluster 5 respectively.

2.Association Rule Mining

Association rules are rules presenting association or correlation between item sets. Associationanalysis identifies relationships or correlations between observations and/or between variables in our datasets. These relationships are then expressed as a collection of so-called association rules. The approach has been particularly successful in mining very large transactional databases, like shopping baskets and on-line customer purchases.  
Association analysis is one of the core techniques of data mining.

A representation of association rules is required to identify relationships between items within transactions. Suppose each transaction is thought of as a basket of items (which we might represent as *{A, B, C, D, E, F}*). The aim is to identify collections of items that appear together in multiple baskets (e.g., perhaps the items *{A, C, F}* appear together in quite a few shopping baskets). From these so called *itemsets* (i.e., sets of items) we identify rules like *A, F ⇒ C* that tell us that when *A* and *F* appear in a transaction (e.g., a shopping basket) then typically so does *.*

An association rule is in the form of *A ⇒ B*, where *A* and *B* are two disjoint item sets, referred to respectively as the lhs (left-hand side) and rhs (right-hand side) of the rule. The three most widely-used measures for selecting interesting rules are *support*, *confidence* and *lift*. *Support* is the percentage of cases in the data that contains both *A* and *B*, *confidence* is the percentage of cases containing *A* that also contain *B*, and *lift* is the ratio of confidence to the percentage of cases containing *B*.

The formulae to calculate them are:

support(*A ⇒ B*) = *P*(*A ∪ B*)   
confidence(*A ⇒ B*) = *P*(*B|A*)   
=*P*(*A ∪ B*)  
*P*(*A*)  
lift(*A ⇒ B*) = confidence(*A ⇒ B*)  
 *P*(*B*)  
=*P*(*A ∪ B*)  
*P*(*A*)*P*(*B*)

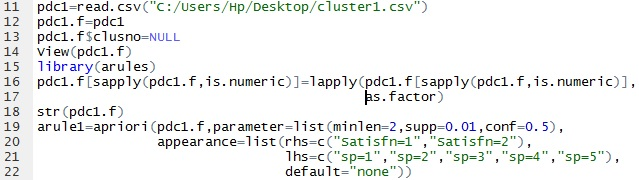
where *P*(*A*) is the percentage (or probability) of cases containing *A*.  
In addition to support, confidence and lift, there are many other interestingness measures, such as chi-square, conviction, gini and leverage.

Definition

The problem of association rule mining is defined as: Let I=\{i_1, i_2,\ldots,i_n\} be a set of n binary attributes called *items*. Let D = \{t_1, t_2, \ldots, t_m\} be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form X \Rightarrow Y where X, Y \subseteq I and X \cap Y = \emptyset. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

APRIORI algorithm was used to implement association rule mining. It is a level-wise, breadth-first algorithm which counts transactions to find frequent itemsets and then derive association rules from them. An implementation of it is function apriori() in package  
*arules.* Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

In this project, association rule mining was done to find relationships between variables. First, the intra cluster analysis was done on all the clusters to find the satisfaction of different service providers among the different people who took part in the survey process. The code for implementing the association rule in the first cluster is shown below.



The result is

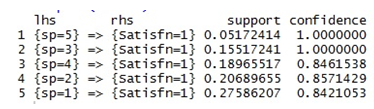


Figure 10: Result of associate rule mining of cluster 1

The result shows that there are only 5 association rules in this case which have support value greater than 0.01 and confidence value greater than 0.5. The confidence value shows the strength of the association. The value 1 shows that the association is very strong. The support value shows the frequency. In this cluster, service provider 3(Airtel) and service provider 5(others) are the service providers which have most satisfaction among people.

Similarly the results of the other clusters are:

Cluster 2:

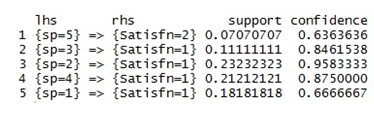


Figure 11: Result of associate rule mining of cluster 2

In cluster 2, there are only 5 rules have which have confidence level greater than 0.5. Service provider 2(Idea) is the service provider with the highest satisfaction level with confidence value of 0.9583333 followed by service provider 4(Vodafone). The first rule indicates that majority of people using service provider 1(others eg: Reliance, Docomo) are not satisfied since its confidence is greater than 0.5 for Satisfn=2( Not satisfied).

Cluster 3:

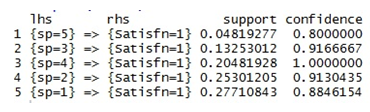


Figure 12: : Result of associate rule mining of cluster 3

In cluster 3, there are 5 rules which satisfy the minimum support and confidence value conditions. Service provider 4(Vodafone) is the service provider with the highest satisfaction level followed by service provider 3(Airtel) and service provider 2(Idea). Service provider with lowest satisfaction level is again service provider 5(others).

Cluster 4:

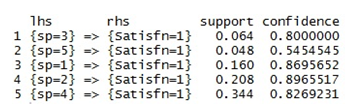


Figure 13 : Result of associate rule mining of cluster 4

In cluster 4, the service provider with highest satisfaction level is service provider 2(Idea) closely followed by service provider 1(BSNL). The service provider with lowest satisfaction level is again service provider 5(others).

Cluster 5:

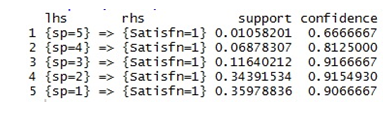


Figure 14: Result of associate rule mining of cluster 5

In cluster 5, the service provider with highest satisfaction level is service provider 3(Airtel) closely followed by service provider 2(Idea). Service provider with lowest satisfaction level is again service provider 5(others).

From this, we can observe that service provider 2 i.e Idea is the service provider with the highest satisfaction level followed by Vodafone and Airtel. The service provider with the least level of satisfaction is service provider 5 i.e other service providers such as Reliance, Docomo etc.

Next we look at different characteristics of customer behaviour with the target variables(rhs) being parameter, call time, online time and purpose depending on the age, gender, occupation and salary. This is to identify patterns and relationships between these variables.

The minimum confidence is set as 0.75 and it is assumed that all rules occurring above this confidence level is significant.

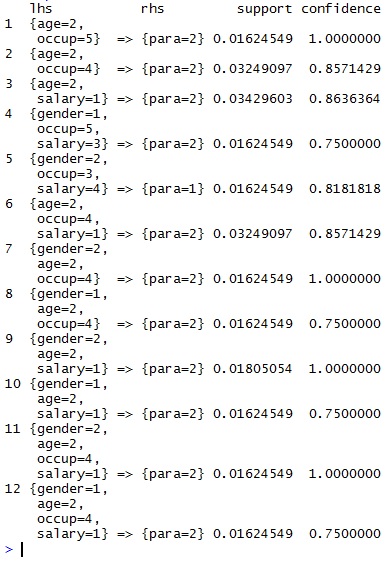


Figure 15: Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and para in Rhs

In the above result of association rule mining, we get all the rules that have confidence value greater than 0.75. For example, in case 11, a person who has gender=2(female), age=2(23-30), occupation=4(student) and salary=1(nil) looks for the parameter=2(). Similarly, such results are predicted above.

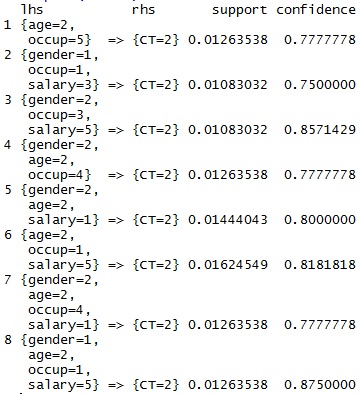


Figure 16 : Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and CT in Rhs

In the above result, 8 rules are obtained which are significant. Call time=2(0.5 hours to 1.5 hours) is the only rhs result obtained with a confidence value greater than 0.75. Hence that is the most significant range of call times which the service providers have to consider.

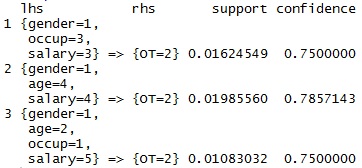


Figure 17:Result of associate rule mining when gender,age,occup,salary are considered in

Lhs and OT in Rhs

In the above result, only 3 rules are obtained. Online time=2() is the only rhs result obtained with a confidence value greater than 0.75. Example, a person with gender=1(male), age=4(40-50) and salary=4(3-5) spends time online=2(1.5 hours to 1.5 hours).

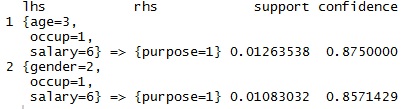


Figure 18 : Result of associate rule mining when gender,occup,salary are considered in

Lhs and purpose in Rhs

With purpose as the rhs variable, only 2 results are obtained.

**CHAPTER-5**

**5.1 PREDICTIVE MODELING USING DECISION TREE**

A decision tree is a [decisionsupport](http://en.wikipedia.org/wiki/Decision_support_system" \t "_blank" \o "Decision support system) tool that uses a tree-like [graph](http://en.wikipedia.org/wiki/Diagram) or [model](http://en.wikipedia.org/wiki/Causal_model) of decisions and their possible consequences, including [chance](http://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](http://en.wikipedia.org/wiki/Utility). It is one way to display an [algorithm](http://en.wikipedia.org/wiki/Algorithm).

A decision tree is a [flowchart](http://en.wikipedia.org/wiki/Flowchart)-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules.

In [decision analysis](http://en.wikipedia.org/wiki/Decision_analysis) a decision tree and the closely related [influence diagram](http://en.wikipedia.org/wiki/Influence_diagram) are used as a visual and analytical decision support tool, where the [expected values](http://en.wikipedia.org/wiki/Expected_value) (or [expected utility](http://en.wikipedia.org/wiki/Expected_utility)) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:

1.   Decision nodes - commonly represented by squares

2.   Chance nodes - represented by circles

3.   End nodes - represented by triangles

Decision trees are commonly used in [operations research](http://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](http://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](http://en.wikipedia.org/wiki/Goal). If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a [probability](http://en.wikipedia.org/wiki/Probability) model as a best choice model or online selection model [algorithm](http://en.wikipedia.org/wiki/Algorithm). Another use of decision trees is as a descriptive means for calculating [conditional probabilities](http://en.wikipedia.org/wiki/Conditional_probability).

Decision trees, [influence diagrams](http://en.wikipedia.org/wiki/Influence_diagrams), [utility functions](http://en.wikipedia.org/wiki/Utility_function), and other [decision analysis](http://en.wikipedia.org/wiki/Decision_analysis) tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of [operations research](http://en.wikipedia.org/wiki/Operations_research) or [management science](http://en.wikipedia.org/wiki/Management_science) methods.

ALGORITHM USED: ID3 ALGORITHM

ID3 is a simple decision learning algorithm developed by J. Ross Quinlan (1986). ID3 constructs decision tree by employing a top-down, greedy search through the given sets of training data to test each attribute at every node. It uses statistical property call information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training examples according to their target classification.

Entropy It is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on c different values, then the entropy S relative to this c-wise classification is defined

WEW

where is the proportion/probability of S belonging to class i. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. A key point to note here is that the more uniform is the probability distribution, the greater is its entropy.

Information gain

It measures the expected reduction in entropy by partitioning the examples according to this attribute. The information gain, Gain(S, A) of an attribute A, relative to the collection of examples S, is defined as

Capture

where Values A( ) is the set of all possible values for attribute A, and v S is the subset of S for which the attribute A has value v. We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root.

Algorithm

Let set *E* is a collection contains of samples, category can take *m* different values correspond to different  
category *Ci , i = 1, 2, 3, ..., m*. If the selected test property is *A* property, with *A* has *v* different values;  
*a*1 , *a*2 ,..., *av* , *A* will divided the set *E* into v sub-sets; *E*1 , *E*2 ,..., *Ev* . Set *Eij* is the sample size for thesubset  
*E j* belongs to *Ci* . When *A* is used to divide the current sample collection, the required  
information entropy is calculated as follows;

Calculate the expected A;

A

Then compute the information entropy for set *E*;

A

For a given subset *E j* , the information entropy is;

QRQ3

111

*pij* is the probability of the samples of subset *E j* belong to category *Ci* .

Get the information gain for using property A; *Gain*( *A*=) *I*(*E*1, *E*2 ,..., *Em* −) *E*( *A*)

The calculation of the information entropy for each property, selection of the biggest gain property as a test property for a given set *E* and generation of the branches and node can be done.

DECISION TREE

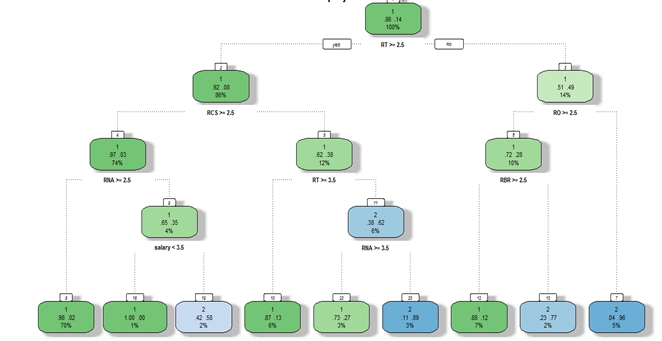


Figure 19: Decision Tree

In this case, the variable with the largest amount of information is 'Rating of Tariffs' and hence it is used as the classification property.

From the total population, there is a probability of 0.86 that people who have given a rating greater than 2.5 are likely to be satisfied. Among this satisfied population, there is a probability of 0.92 that customers who have given a rating of more than 2.5 on customer service are likely to be satisfied. and among this population there is a probability of 0.97 that customers who have given, as their rating of the network accessibility score of 2.5 or more are likely to be satisfied.

In a similiar manner, after the analysis of all the branches and all the nodes we arrive at the following result. There is a 0.98 probability that 70% of the population that have given a rating of more than 2.5 for network accessibility, and a rating of more than 2.5 for customer service and a rating of more than 2.5 for tariffs are likely to be satisfied.

**CHAPTER-5**

**CONCLUSION**

The consumer behaviour in the telecom industry was studied. The various results regarding this study are:

* By the associate rule mining done on each cluster, we have obtained the result that the customer satisfaction is most for the service provider Idea followed by Airtel and Vodafone.
* A predictive model was built using decision tree algorithm to predict the customer satisfaction and it was predicted that 70% of the population who have given a rating of more than 2.5 for network accessibility, customer service and tariffs are likely to be satisfied with the service provider. Similar predictions were made based on different combinations of variables however, the customer satisfaction in all these cases were less than 10%.
* The different trends that explain the customer behaviour were found out using descriptive statistics.
* The average ratings of various parameters like network accessibility, tariffs, billing/recharge services, package offers, and customer satisfaction were found out against the different service providers were found out.

**CHAPTER-6**

**BIBLIOGRAPHY**

Data Mining in Churn Analysis Modelfor Telecommunication Industry, *Khalida binti Oseman1 Sunarti binti Mohd Shukor1Norazrina Abu Haris1Faizin bin Abu Bakar11Operation Support Systems Division TM Research & Development Sdn. Bhd.*

Exploring consumer adoption of new services by analyzing the behaviorof 3G subscribers: An empirical case study; *Li-Chen Cheng , Li-Min SunDepartment of Computer Science and Information Management, Soochow University, Taipei 100, Taiwan, ROC*

Customer experience management: An exploratory study on theparameters affecting customer experience for cellular mobileservices of a telecom company ; *Sujata Joshi*

*Symbiosis Institute of Telecom Management, Symbiosis International University, Lavale, Pune, India*

A Study of Customer satisfaction on Telecom Service Providers; *Dinesh Kumar Pandiya Dr Brajesh Kumar, f ASSAM UNIVERSITY, SILCHAR, Department of Commerce*

Telecom Sector in India: Past, Present and Future;*Dr. Papori Baruah Professor, Department of Business Administration, Tezpur University, India & Rashmi Baruah Asst. Professor, Dept. of Business Administration, University of Science & Technology, Meghaloya, India*

An Implementation of ID3 --- Decision Tree Learning Algorithm;*Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia*

Highlights on Telecom Subscription Data as on 30th November, 2013: *(Press Release No. 04/2014), TELECOM REGULATORY AUTHORITY OF INDIA New Delhi, 29th January, 014*

MARKETING RESEARCH: An applied approach (Updated second edition) *by Naresh S Malhotra, David F. Birks*

Data Mining using Rattle and R, The Art of Excavating data for Knowledge Discovery*; Graham Williams*