Thesis no. MSEE-2016:37

Customer Churn Prediction Using Big Data Analytics

Naren Naga Pavan Prithvi Tanneedi

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering with Emphasis on Telecommunication Systems. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author(s):
Naren Naga Pavan Prithvi Tanneedi
E-mail: nata15@student.bth.se,
pavanprithvi27@yahoo.com

University advisor:
Prof. Dr.-Ing. Markus Fiedler
Dept. of Communication Systems

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

ABSTRACT

Customer churn is always a grievous issue for the Telecom industry as customers do not hesitate to leave if they don't find what they are looking for. They certainly want competitive pricing, value for money and above all, high quality service. Customer churning is directly related to customer satisfaction. It's a known fact that the cost of customer acquisition is far greater than cost of customer retention, that makes retention a crucial business prototype. There is no standard model which addresses the churning issues of global telecom service providers accurately. BigData analytics with Machine Learning were found to be an efficient way for identifying churn. This thesis aims to predict customer churn using Big Data analytics, namely a J48 decision tree on a Java based benchmark tool, WEKA. Three different datasets from various sources were considered; first includes Telecom operator's six month aggregate active and churned users' data usage volumes, second includes globally surveyed data and third dataset comprises of individual weekly data usage analysis of 22 android customers along with their average quality, annoyance and churn scores by accompanying theses. Statistical analyses and J48 Decision trees were drawn for three different datasets. From the statistics of normalized volumes, autocorrelations were small owing to reliable confidence intervals, but confidence intervals were overlapping and close by, therefore no much significance could be noticed, henceforth no strong trends could be observed. From decision tree analytics, decision trees with 52%, 70% and 95% accuracies were achieved for three different data sources respectively.

Data preprocessing, data normalization and feature selection have shown to be prominently influential. Monthly data volumes have not shown much decision power. Average Quality, Churn Risk and to some extent, Annoyance scores may point out a probable churner. Weekly data volumes with customer's recent history and necessary attributes like age, gender, tenure, bill, contract, data plan, etc., are pivotal for churn prediction.

Keywords: Big Data, churn prediction, decision tree, Quality of Experience.

ACNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my supervisor Markus Fiedler for his constant support, fortitude, understanding and encouragement throughout my thesis study. His expert guidance and comments helped me in exploring key topics, accomplishing various tasks and composing the report on time. His immense generosity, patience and colossal knowledge makes him the best mentor.

I would like to thank the course responsible, Prof. Kurt Tutschku for his timely updates despite his busy schedule. His encouragement and guidance throughout my master's education is commendable.

I am very thankful to my friendly theses partners, Hemanth and Mounika for their constant support and imperative advices throughout my academic exploration.

Finally, I would like to thank my parents for their unconditional love and support throughout my education. Without their encouragement and motivation, I may not be where I am today.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Networks
ARFF	Attribute-Relation File Format
CI	Confidence Intervals
CHAID	Chi-squared Automatic Interaction Detector
CSV	Comma Separated Values
FP	False Positive
GP	Genetic Programming
ID3	Iterative Dichrometer 3
KDD	Knowledge Discovery in Databases
MCC	Matthews Correlation Coefficient
PRC	Precision and Recall
QoE	Quality of Experience
QUEST	Quick Unbiased Efficient Statistical Tree
RA	Research Answer
ROC	Receiver Operating Characteristic
RQ	Research Question
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis
WiFi	Wireless Fidelity

1 INTRODUCTION

In competitive Telecom market, the customers want competitive pricing, value for money and high quality service. Today's customers won't hesitate to switch providers if they don't find what they are looking for. This phenomenon is called churning. Customer churning is directly related to customer satisfaction. Since the cost of winning a new customer is far greater than cost of retaining an existing one, mobile carriers have now shifted their focus from customer acquisition to customer retention [1].

After substantial research in the field of churn prediction over many years, BigData analytics with Machine Learning was found to be an efficient way for identifying churn. These achieve results more efficiently and receive insights that sets alarm bells ringing before any damage could happen, giving companies an opportunity to take precautionary measures. These techniques are usually applied to predict customer churn by building models and learning from historical data [2]. However, most of these techniques provide a result that customers might churn or not, but only few tell us why they churn.

Conducting experiments with end users' perspective, gathering their opinions on network, data normalization, preprocessing data sets [7], employing feature selection [6], eliminating class imbalance and missing values [5], replacing existing variables with derived variables [1] improves the accuracy of churn prediction which assists Telecom industries to retain their customers more efficiently.

Comparatively, a smaller study was done on user's perspective, taking into consideration their quality of experience. In fact, no study was done taking into consideration only user's data volumes. Estimation of Quality of Experience by finding relationships between QoE and traffic characteristics could help the service providers to continuously monitor the user satisfaction level, react timely and appropriately to rectify the performance problems and reduce the churn [3] [4].

1.1 MOTIVATION

The Telecom industry is humongous, vibrant and dynamic with extremely large base of customers, making customer acquisition and customer retention imperative concerns for its survival and good profitability. The new entrants focus on customer acquisition, while old and matured one emphasize to focus on customer retention. Globalization enables customers to choose the best available services, which encourages the customers not to stick with a single company, rather opt from a diverse range of products/services. Customer churning is directly related to customer satisfaction [1]. Since the cost of acquiring new customer is much higher than retaining old news, operators lay preeminent significance on various customer related methodologies and analytics to ensure customer retention.
There is no clear common consensus on the prediction technique to be used to identify churn. Significant research in the field of churn prediction is being carried out using various statistical and data mining techniques since a decade. BigData analytics with Machine Learning were found to be an efficient way for churn prediction. Several previous works [1] [7] [8] [14] focused on various data mining techniques for churn prediction based on call detail records. The work in [13] focused on service failures and disconnections recorded to identify churn. Study [5] focusses to detect early warnings of churn by assigning "Churn Score" for numerous customer transaction logs.
So far, customer churn has been majorly studied on network parameters. Barely, any study could be addressed regarding churn prediction with user's perspective taking their Quality of Experience into consideration. No study was done taking only data usage volumes into consideration. This thesis aims to predict customer churn using Big Data analytics, J48

decision tree on a Java based tool, WEKA; considering only users' data usage volumes from three different datasets. There is no standard model which addresses the churning issues of global telecom service providers accurately. This thesis predominantly focuses to identify churn using decision trees, one of the most popular data mining techniques. A decision tree is an eminent categorizer that use a flowchart-like process for categorizing instances. During the process of customer churn prediction, Telecom operators would often need to analyze the steps to figure out the probable cause and rationale instigating customers to churn. This could be only possible with decision trees as they are easy to interpret, visualize and analyze.

1.2 PROBLEM STATEMENT

In the competitive Telecom industry, public policies and standardization of mobile communication allow customers to easily switch over from one carrier to another, resulting in a strained fluidic market. Churn prediction, or the task of identifying customers who are likely to discontinue use of a service, is an important and lucrative concern of the Telecom industry. The aim of this thesis is to study and analyze customer churn prediction based on mobile data usage volumes with respect to QoE and users' perspective with the help of BigData analytics.

1.3 RESEARCH QUESTIONS

RQ1. What role does data normalization play in churn prediction?

RQ2. To which extent can data usage volumes be used for customer churn?

RQ3. Which kind of information is needed for a prominent churn prediction?

1.4 HYPOTHESIS

Through literature study, surveys and previous works, various discrepancies, challenges and difficulties are identified. After the data acquisition from the anonymous Telecom provider and an experimental survey by Mounika Reddy Chandiri [28], statistical and BigData analytics were carried out to draw different convictions on usage trends. From the individual data usage traffic analysis by Hemanth Kumar Ravuri [27], a certain trend of variation on the usage pattern is expected. The analysis is also expected to result in certain correlations between the varying data traffic, annoyance, churn risk and the quality of experience with respect to users and the BigData analytics indicating churn. From the study of the thesis work, a derivation of a general relation between users' satisfaction and users' traffic volume is expected to be reached.

1.5 METHODOLOGY

This thesis aims to study and analyze customer churn based on data usage volumes with respect to QoE and users' perspective using BigData Analytics. Three different datasets were analyzed statistically and with the help of J48 decision trees. Statistical analysis includes calculation and analysis of Mean, Standard deviation, Autocorrelations and Confidence intervals. Decision tree analysis includes data acquisition, data preparation that includes normalization, data preprocessing, data extraction and finally decision making.

1.6 MAIN CONTRIBUTION

Telecom operators need to be able to accurately predict churn in order to respond in time. The prime aspiration of this thesis is to predict customer churn from monthly and weekly mobile data usage volumes using BigData analytics. This thesis along with two partner theses is collaborated and united to form a lone crucial and dominant project. The main objectives include:

- Conduct survey with different sections of people regarding their data usage and numerous other questions [28].
- Analyze weekly QoE polls and volume measurements by Android-based tool compared [27].
- Study the importance of data preprocessing, data normalization and feature selection.
- Carefully analyze and assess six-month aggregate data usage volumes for active and churned users given by an anonymous Telecom provider.
- Carry out statistical and decision tree analyses for three datasets; one from Telecom provider and others from accompanying theses.
- Correlate and compare the results to know to which extent only data usage volumes could be used to predict churn.
- Finally, affirm the necessary information required for prominent churn prediction.

1.7 SPLIT OF WORK

The scope of churn prediction being vast, three thesis topics were designed on the same platform. All three of us share the results and data sources amongst ourselves. There might be few overlaps in analysis and results, but each individual has his/her own perspective in analyzing the data.

1. Survey on a global scale by Mounika Reddy Chandiri [28].

2. Android-based tool for weekly QoE polls and volume measurements by Hemanth Kumar Ravuri [27].

3. Churn prediction using BigData analysis by Naren Naga Pavan Prithvi Tanneedi.

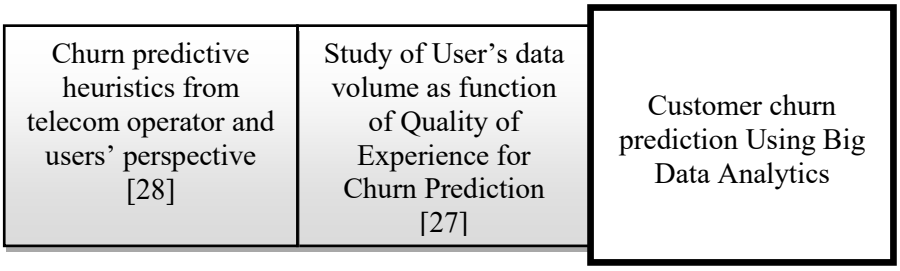


Figure 1. Split of work

1.8 THESIS OUTLINE

The outline of this thesis is briefly described in chapter 1. Introduction along with motivation, Problem statement, Research Questions, expected Hypothesis, main contribution, pithy Methodology and Split of Work are explained.

Chapter 2 presents background of related works. It includes recent ant prominent publications of literature review of various papers, journals and articles in the field of churn prediction, data preparation and data mining techniques.

Chapter 3 presents a detailed description of various analyses that have been adopted to identify customer churn. It explains about various tools and methods that have been enforced during the process of approach.

Chapter 4 gives a brief overview of results of statistical and BigData analytics (Decision trees) from three different data sources.

Chapter 5 explains a detailed interpretation of statistical and decision tree analyses mentioned in chapter 4.

Chapter 6 illustrates the conclusions drawn from various analysis accomplished in chapter 5.

Chapter 7 provides the references to related work.

2 RELATED WORK

Research in the area of customer churn is always a trending topic. The unbridled growth of databases in recent years brings data mining to the forefront of new business technologies., becoming our only hope for elucidating the patterns that underlie it [23]. Significant research in the field of churn prediction is being carried out using various statistical and data mining techniques since a decade. This chapter presents the recent and prominent publications on churn prediction in the recent years.

Gavril Toderean et al [8] presented an advanced data mining methodology that predicts customer churn in the pre-paid mobile telecommunications industry using call detail records dataset that consists of 3333 customers with 21 attributes each and a churn dependent variable with two classes Yes/No. Few attributes include the information about their corresponding inbound/outbound SMS count and voice mail. A principal component algorithm was applied to reduce the dimensionality of data and to eliminate the problem of multicollinearity. Three machine learning algorithms, namely neural networks, support vector machines and Bayesian networks were used to predict churn variable based independent variables. These models were evaluated using confusion matrix, gain measure and ROC curve. An overall accuracy of 99.10%, 99.55% and 99.70% were achieved for Bayesian networks, neural networks and support vector machines respectively.

Kiran Dahiya et al [9] proposed a new framework for churn prediction model, implemented it using WEKA data mining software. Each customer was classified as a potential churner or non-churner. The framework discussed was based on Knowledge Discovery Data process. Three different datasets, small, medium and large with varying attributes were considered. The efficiency and performance of decision tree and logistic regression techniques have been compared. Accuracy achieved with decision tree was much greater than logistic regression.

Utku Yabas et al [10] explains about subscriber churn analysis and prediction for mobile and wireless service providers. A real and complied dataset by Orange Telecom, 2009 was used. Main emphasis was laid on ensemble methods that encompass single methods to improve the solution to churn prediction problem. These results were compared with that of meta-classifiers, namely logistic regression, decision trees and random forests; and had encouraging values when considered for both ROC score and computing efficiency.

Saad Ahmed Qureshi et al [1] aims to present commonly used data mining techniques for churn prediction. The dataset used was obtained from Customer DNA website and contains traffic data of 1,06,000 customers and their usage behavior for three months. The class imbalance problem was solved by re-sampling. Regression analysis, Artificial Neural Networks, K-Means Clustering, Decision Trees including CHAID, Exhaustive CHAID, CART and QUEST were taken into consideration to identify churn. The results were compared based on the values of precision, recall and F-measure. Decision trees, especially Exhaustive CHAID were found to be the most accurate algorithm in identifying potential churners.

Muhammad Raza Khan et al [5], presented a unified analytic framework for detecting the early warnings of churn, and assigning a "Churn Score" to each customer that indicates the likelihood of a particular customer to churn within a predefined amount of time. The approach uses a brute force approach to feature engineering that generates a large number of overlapping features from customer transaction logs, then uses two related techniques to identify the features and metrics that are most predictive of customer churn. These features are then fed into a series of supervised learning algorithms that can accurately predict subscriber churn. For a dataset of roughly 1,00,000 subscribers from a South Asian mobile operator observed for 6 months, an approximate of 90 percent accuracy was achieved.

In order to solve the problem of big customer churn of about 5.23 million customers from China Telecom and China Netcom for fixed communication network operators, Yue He et al [11], proposed a prediction model based on RBF neural network. It then subdivides the customers by Analog Complexion Cluster to guide and help manage marketing and related work.

Genetic Programming (GP) based approach along with AdaBoost for modeling the challenging churn problem was proposed by Adnan Idris et al [7]. The GP's evolution process was exploited by integrating an AdaBoost style boosting to evolve multiple programs per class and final predictions are made on the basis of weighted sum of outputs of GP programs. This was tested on two standard datasets, one by Orange Telecom and the other by cell2cell. The accuracy achieved was 89% for cell2cell dataset and 63% for the other.

Xiaohang Zhang et al [12], investigated the effects of network attributes on the accuracy of churn prediction. Network attributes refer to the interaction among customers and the topologies of their social network, which is constructed by the customer calling behaviors. The predictions of traditional attribute-based models, network attribute-based models and combined attributes models are compared and found that incorporating network attributes into predicting models can greatly improve the prediction accuracy. The network attributes can be useful complements to the traditional attributes.

Michael J.Prez et al [13] proposed to identify customers with service failures and determine the propensity for a customer to disconnect based on the frequency of a recent service failure reported and success of repair. The dataset used in this study was from monthly statistical reports of a national multi-system operator in the telecommunications industry over a 10month period during January to October 2008. Two approaches were used in this study. The first looked at the service experience of customers with a service failure, from provider's "phone survey statistics" of current customers with a service failure. The second approach looked at the frequency of customers who had disconnected their services following a service failure within a 30 day (monthly) reporting period, using empirical data from the telecommunications provider's "billing system". The proceedings stated that the customers subscribed for the triple-play of voice, video and internet access were more likely to cancel all services after a service failure than other customers.

In paper [14] by L.Bin et al, call details of 6000 customers of Personal Handy phone System Service in China are observed for 180 days. After data pretreatment, data of 4799 customers was preserved. In order to build an effective and accurate model, three experimentations were considered to improve the ability of churn prediction. These include: changing sub-periods for training data sets, changing misclassification cost in churn model, changing sample method for training data sets. The results suggested that these churn models have excellent performance, quite effective and feasible only for limited information and skewed class distribution.

In paper [6] by A.Idris and A.Khan, a dataset of 40,000 instances provided by cell2cell Telecom Company was pre-processed to a balanced form. In the preprocessing stage, in order to provide discriminating features to the classifiers mRMR, Fisher's ratio and F-Score feature extraction methods were used. For each of these methods, a linear search is performed to select the features which provide maximum discriminating information to the classifiers and hence produce better performance. When a linear search is performed for all the methods with rotation forest, for mRMR the accuracy for predicting the churners was 76.2%, while it was 69.1% and 65.2% for Fisher's ratio and F-Score respectively. For Random Forest, the accuracy of churn prediction for mRMR, Fisher's Ratio and F-Score were 74.2%, 71.6% and 71.3% respectively.

3 METHODOLOGY

This chapter presents detailed description of various analyses that have been adopted to identify customer churn. Brief overviews of Weka tool and Decision tree are presented in sections 3.1 and 3.2 respectively. A concise critique of numerous modules that have been implemented during analyses are conferred in section 3.3.

3.1 WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is a free open source Java based data mining software issued under the GNU General Public License. It is a collection of various machine learning algorithms and classifiers determined for diverse data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [20].

WEKA has several graphical user interfaces that enable easy access to the underlying functionality. The main graphical user interface is the "Explorer". It has a panel-based interface, where different panels correspond to different data mining tasks [24]. The first panel is "Preprocess" panel, where data can be loaded from various data sources and revamped using WEKA's data pre-processing tools, called "filters". Supported file formats include ARFF, CSV, LibSVM, and C4.5.

The second panel in the Explorer, "Classify" gives access to WEKA's classification and regression algorithms. By default, this panel runs a cross-validation for a selected learning algorithm on the dataset after pre-processing. It also provides textual and graphical representation of various applicable models built from the full dataset. Moreover, it can visualize prediction errors in scatter plots that allows evaluation through different threshold curves. WEKA supports both supervised and unsupervised algorithms. These are accessible in the Explorer from third and fourth panels.

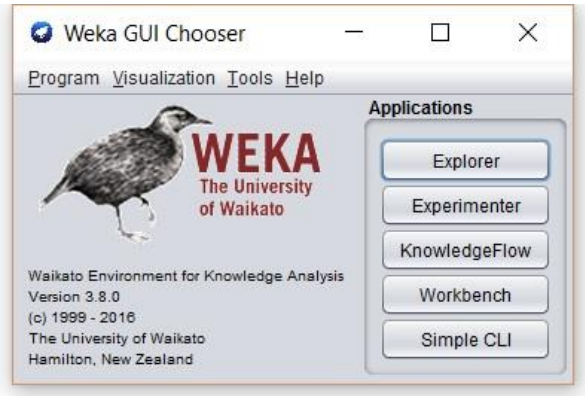


Figure 2. WEKA workbench

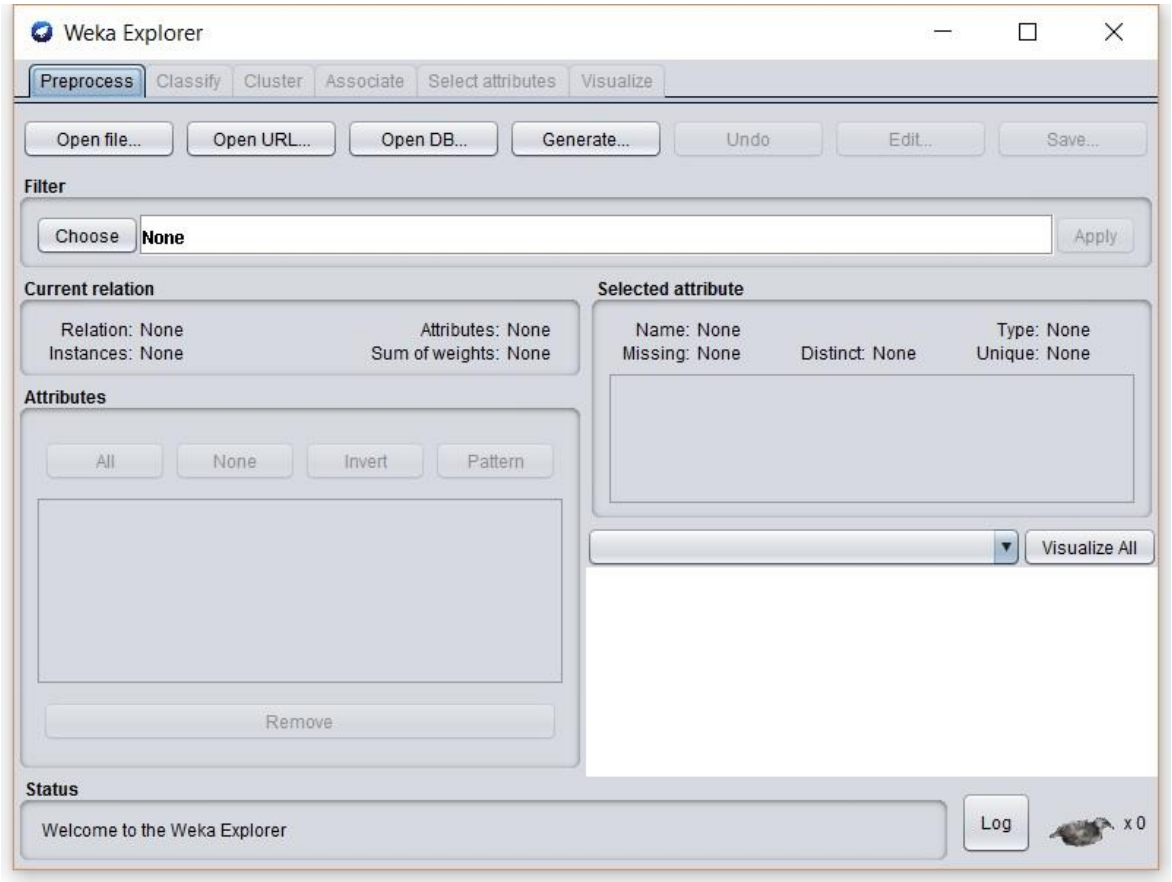


Figure 3. "Explorer" interface from WEKA GUI.

3.2 DECISION TREE

After a substantial research in the field of churn prediction over many years, Big Data analytics were found to be an efficient way for identifying customer churn. Big Data analytics achieve better results more efficiently and receive insights that sets alarm bells ringing before any damage could happen, giving companies an opportunity to take precautionary measures.

Decision trees are one of the predictive modeling approaches extensively used in data mining, where in a tree is used to explicitly represent decisions and decision making. These are the structured regression models. The goal of decision tree is to create a model that predicts the value of a target based on several input variables. Each node of a decision tree represents one of the traffic usage attributes of the customer. Leaves represent class labels and branches represent conjunctions of features that lead to class labels.

The decision tree used in this thesis is J48, which is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. J48 is specifically chosen for its pruning ability and exceptional handling of missing classes, which no other tree could perform. Algorithm C4.5, often referred to as a statistical classifier is an extension to ID3 algorithm which builds decision trees from a dataset using the concept of information entropy [25]. C4.5 algorithm has been quite successful in achieving the discrepancies of ID3 algorithm as it could handle both continuous and discontinuous attributes by creating thresholds, then branching the values based on these thresholds. This algorithm could propitiously handle the missing attributes and helps in pruning the trees after creation, where in the size of decision tree is reduced by removing branches/leaves that provide very little power to classify instances.

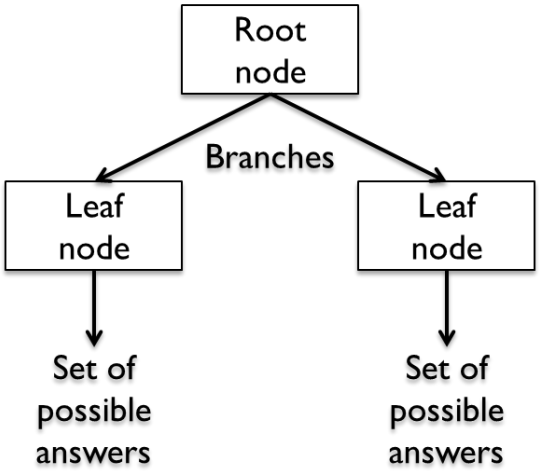


Figure 4. Decision tree

The trees are constructed in a top-down recursive divide-and-conquer manner. In a J48 decision tree, after the decisions are made, leaf nodes have two values. The first value demonstrates the total number of instances reaching the leaf and second value indicates the number of misclassified instances. In the case of missing values, fractional values are exhibited at the leaves. This tree by default uses 10-fold cross validation which states that 90% of the data is used for training and 10% for testing. As 90% is not too far from 100%, it gives affair estimate of the value. These cross validation folds can be switched at the "Test options" from Classify in WEKA. The accuracy achieved with a decision tree is far much higher than other data mining techniques which clearly states that decision tree is an efficient technique to predict churn [9].

3.2.1 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classifier on a particular dataset. It contains information about actual and predicted classifications done by a classification system. The entries in the confusion matrix have the following meaning in the context of our study [29].

TP (True Positive)- The number of instances when the predicted churner is truly a churner.

TN (True Negative)- The number of instances when the predicted non-churner is truly a non-churner.
FP (False Positive)- The number of instances when the predicted churner is non-churner in real.
FN (False Negative)- The number of instances when the predicted non-churner is a churner in real.

Confusion Matrix	Predicted YES	Predicted NO	
Actual YES	TP	FN	Total Real YES's
Actual NO	FP	TN	Total Real NO's
	Total YES predictions	Total NO predictions	

Figure 5. Confusion Matrix

3.3 APPROACH TOWARDS CHURN PREDICTION

The data sources for this thesis include:

a. Monthly data volumes for churned and active users from October 2015 to March 2016, provided by an anonymous Telecom provider.
b. Results of a global Telecom survey carried out by accompanying thesis, Chandiri Mounika Reddy [28].
c. Individual analysis results from Android based tool for weekly QoE polls and volume measurements, by Hemanth Kumar Ravuri [27].

A process known as Knowledge discovery in databases (KDD) has been implemented in our decision tree analyis. Knowledge discovery is defined as 'the non-trivial extraction of implicit, unknown, and potentially useful information from data' [26]. Data mining comes under this process KDD, where diverse patterns are extracted from different databases.

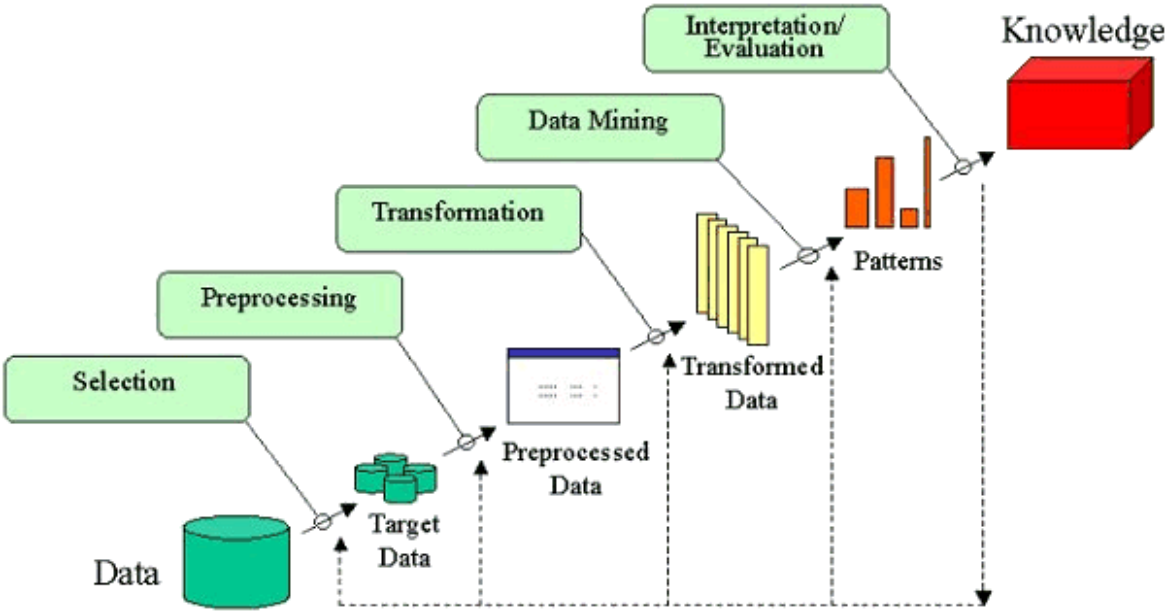


Figure 6. KDD approach in Data mining [22]

This thesis includes two types of analyses, namely Statistical Analysis and Decision tree analysis.

3.3.1 Statistical Analysis

Statistical Analysis is the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Telecom companies use statistics to optimize network resources, improve service and reduce customer churn by gaining greater insight into subscriber requirements [19]. Mean, Standard deviation, Standard error, Lag1 Autocorrelation, 95% Confidence Intervals have been calculated for various combinations.

a) **Mean:** Mean means the statistical average of a dataset. It usually depicts the central value of a set of numbers.
b) **Variance:** Variance is the average of squared differences from the Mean.

c) **Standard deviation:** The Standard Deviation is a measure of how spread out numbers are. In simple words, it's the square root of Variance.
d) **Standard error:** Standard error is defined as the standard deviation of sampling distribution (Mean). Mathematically, the division of standard deviation and square root of number of total instances of sampled data gives the Standard error.
e) **95% Confidence Intervals (CI):** Confidence intervals are a type of interval estimates that gives the most likely range of an unknown population. Confidence intervals consists of different ranges of values, 90%, 95% and 99%. In practice, confidence intervals are usually stated at 95% confidence level, 95 being not too far away from 100. Statistically, if there is a large overlap in confidence intervals, difference is not significant; whereas if the intervals do not overlap, there is a difference with 95% confidence value.
Mathematically, CI= $\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, where $\bar{X}$ means Mean, $z_{1-\frac{\alpha}{2}}$ is percentile of Normal distribution, $\frac{s}{\sqrt{n}}$ is estimation of variance of Mean, $z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ is half-size of confidence interval.
f) **Lag 1-Autocorrelation:** Autocorrelation is correlation of data with itself at different points in time. It often refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called "lagged correlation" or "serial correlation", which refers to the correlation between members of a series of numbers arranged in time [19].

3.3.2 Decision Tree Analysis

The framework of this analysis consists of five modules:

a) **Data Acquisition:** It is a process of gathering facts, details, statistics and necessary information from different sources. Acquiring data from the Telecom industry is the biggest task because of the fear of misusing it. The data sources have been cited at the peak of this section.
b) **Data Preparation:** Data preparation means to transform the acquired data into suitable forms for further analysis. The acquired datasets cannot be directly applied to the churn prediction models. In this regard, the aggregated data is provided with varied variables by looking at the usage behavior of the customers.
c) **Data Preprocessing:** It means to transform the raw data into human/machine understandable data. Data preprocessing is the most important phase in decision making as the raw data contains numerous ambiguities, duplicates, errors, missing values and redundant values. These values have no significance in predictive modeling. Therefore, all such posts are straightaway eliminated. Likewise, fields with too many null values needs to be discarded.
d) **Data Extraction:** Data extraction means to analyze and scrutinize the data to retrieve relevant information from different data sources in a specific pattern. Here, the attributes are identified for classifying process.
e) **Decision:** Here comes the final module of Decision tree Analysis, where the provider is set to make the final decision amongst various possibilities. The rule set will let the Telecom provider identify and classify in the different categories of churners and active customers by setting a particular threshold value.

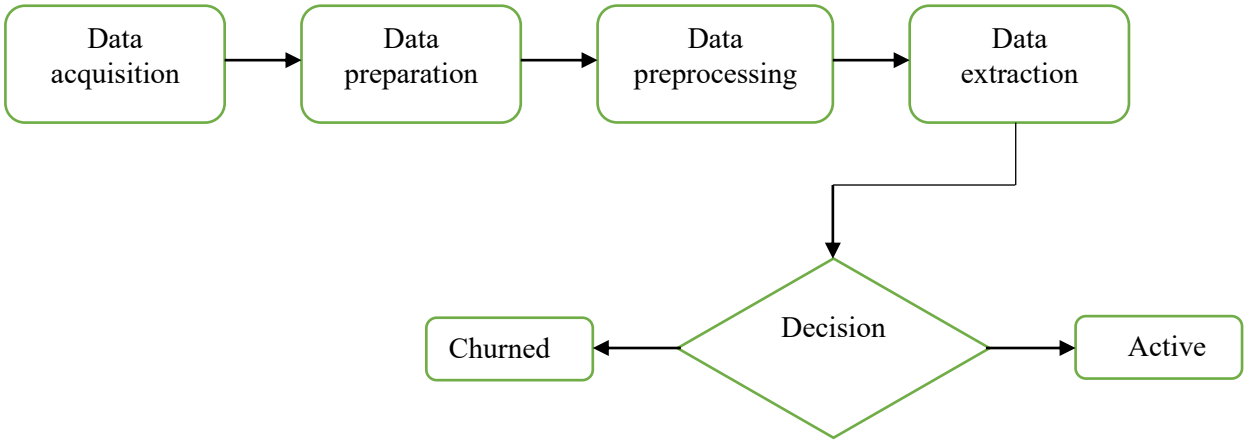


Figure 7. Framework of Decision Tree Analysis

4 RESULTS

This chapter gives a brief overview of the statistical and decision tree analytical results for three different data sources.

4.1 Anonymous Telecom provider

Reasons for Churn	Before	After
Poor service	65	45
Better price	148	128
Altered needs	312	169
No contact	4	2
Information	31	27
Coverage problem	38	24
Parallel Subscription	12	10
Consumer	8	8
Inter correction	3	3
Already has another subscription	31	28
Reference	4	3
Wrong info in buying opportunity	2	0
Fixed broadband	6	6
Not aware of order	2	0
Retains previous subscriptions	3	1
Changing orders	3	0

Table 1. Reasons for churn before and after data pre-processing

Table 1 construes the actual reasons given by the churners for an anonymous Telecom provider. Only 673 out of total 4106 have given the reasons for churn. The majority of the reasons cover Altered needs followed by Better price and Poor service. After the data pre-processing, only 454 out of total 2670 churners remained with reasons for churn. Majority were Altered needs, followed by better pricing options and poor service.

	March	February	January	December	November	October
Mean [MB]	2551.95	2877.20	3149.64	3030.09	3221.05	2997.06
Std-dev [MB]	9385.84	10776.39	11091.94	10072.81	10975.07	8698.05
95% CI half-size [MB]	287.09	329.63	339.28	308.10	335.70	266.05
Autocorrelation	15.18%	20.50%	17.36%	11.11%	11.13%	15.31%

Table 2. Statistical analysis for total churners

Table 2 depicts the statistical analytics for the total 4106 churners from the Telecom provider during October,2015 to March,2016. Mean, standard deviation, 95% confidence intervals and lag1 autocorrelations were calculated for the monthly data usage volumes.

	December	November	October
Mean [MB]	4261.52	4567.11	4141.24
Std-dev [MB]	11664.81	12835.09	10021.18
95% CI half-size [MB]	442.46	486.85	380.12
Autocorrelation	9.61%	9.56%	13.89%

Table 3. Churners without zeros and light users

Table 3 illustrates the statistical results for the churned customers after the data pre-processing, for a total of 2670 customers. Data pre-processing includes deduction of unwanted entries, zeros, duplicates and missing values. It eliminates the unnecessary redundant entries.

	December	November	October
Mean	0.969	1.023	1.004
Std-dev	0.508	0.444	0.519
95% CI half-size	0.019	0.016	0.019
Autocorrelation	5.09%	4.63%	5.00%

Table 4. Churners without zeros and light users (normalized)

Table 4 shows the statistical analytics for the churned customers after data pre-processing, preparation and normalization. The cleaned data is further normalized for better comparability of usage trends between months and properly organized for the WEKA tool to process the results.

	March	February	January	December	November	October
Mean [MB]	4212.89	3649.64	3767.99	3885.40	3554.85	3918.58
Std-dev [MB]	13338.81	11215.28	13188.09	12692.13	11199.46	13007.92
95% CI half-size [MB]	471.31	396.28	465.99	448.46	395.72	459.62
Autocorrelation	8.49%	3.91%	1.97%	2.12%	1.89%	2.52%

Table 5. Statistical analysis for total active customers

Table 5 depicts the statistical analytics of 3077 anonymous Telecom provider's active customers during October,2015 to March,2016.

	December	November	October
Mean [MB]	4823.36	4456.13	4899.77
Std-dev [MB]	13954.11	12380.08	14394.88
95% CI half-size [MB]	552.10	489.83	569.54
Autocorrelation	1.04%	0.59%	1.18%

Table 6. Active customers without zeros and light users

Table 6 illustrates the statistical results for active customers after the data pre-processing, for a total of 2454 active customers.

	December	November	October
Mean	0.984	0.977	1.039
Std-dev	0.521	0.449	0.518
95% CI half-size	0.020	0.017	0.020
Autocorrelation	0.26%	-0.92%	1.70%

Table 7. Active customers without zeros and light users (normalized)

Table 7 portrays the statistical analytics for active customers after data pre-processing, data preparation and normalization.

GROUP	PREDICTION OF	NORMALIZATION	CORRECTLY CLASSIFIED
Churners only	Reasons	No	31.53%
Churners only	Reasons	Yes	33.42%
Active & Churn users	Churn vs. Active	Yes	51.75%

Table 8. Decision Tree Analysis

Decision tree is a decision modeling tool that graphically displays the classification process of a given input for given output class labels [21] . The prime goal of decision tree learning is to achieve perfect classification with minimal number of decisions, although not always possible due to noise or inconsistencies in data. Table 8 gives a brief critique regarding the correct classification of decision trees carried out for the most relevant trees under various instances. The figure below depicts the decision tree for the best classification that could be achieved after numerous trials and normalizations.

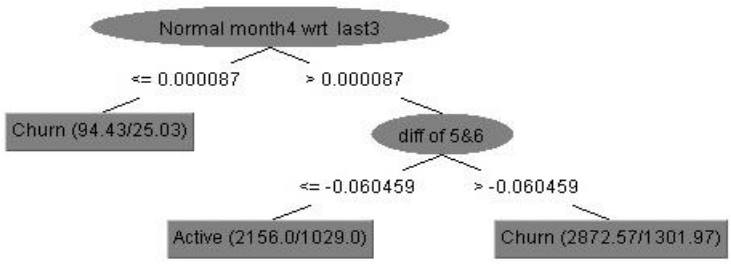


Figure 8. Visualization of decision tree for normalized Active and Churned users

The usage data from October 2015 to March 2016 from anonymous Telecom provider was received in order of month1, month2, month3, month4, month5 and month6, starting from March 2016 going backwards, where November 2015 represents month5 and October 2015 represents month6. Here, the main node is represented by 'Normal month 4 wrt last 3'. Normal here means the normalized value of month4 with respect to aggregate data usage average of last quarter of 2015, namely October, November and December. 'Diff of 5 & 6' means the difference of months 5 & 6, i.e., the data usage (in MB) difference of November and October. The two branches from the root node corresponds to two possible outcomes. If the value of normalized October is less than or equal to 0.0000087, the outcome is Churn. If it is greater than 0.000087, a second test is made, this time on difference between November and December (November-December). Eventually, if the difference is less than or equal to '-0.060459', outcome is Active and if difference is greater than '-0.060459', outcome is Churn.

4.2 Telecom Survey

This telecom survey was carried out by accompanying thesis partner, [28]. A total of 770 customers across the globe answered this survey. The survey includes the following questions:

a) Enjoying your stay at?
b) Age is just a number and your number is?
c) Present Telecom Provider?
d) Have you ever changed your Telecom Provider?
e) The Telecom Provider you previously used?
f) Looking to change the provider in the future?
g) What are the experiences that will make/made you change? (Please select all that applies)
h) The main problem with calls that will make/made you change? (Please select all that applies)?
i) Problems with Data plan that will make/made you change? (Please select all that applies)
j) Suggestions for Telecom Providers.

Upon receiving the feedback from customers across the globe, categorizing the answers of customers by data preprocessing and data preparation, a huge decision tree could be generated with about 69.7% accuracy. The tree was terribly huge and can be seen from the Appendix A3. The tree illustrates the current Telecom provider to which the user might tend to churn.

4.3 Individual data usage analysis

This analysis was carried out by accompanying thesis, [27]. Individual data usage of about 22 customers from three countries, India, USA and Sweden were analyzed for a period of about 8weeks with the help of an Android-based tool. Initially a survey was conducted, which includes the following questions:

a) Country?
b) Sex?
c) Age?
d) Present Telecom Provider?
e) How was your mobile's internet quality in the past week?
f) How was your wifi's quality in the past week?
g) How annoyed were you with your mobile's internet service in the past week?
h) How annoyed were you with your wifi's service in the past week?
i) Problems that you have faced frequently with your internet service in the past week? (Please select all that applies)
j) Would these experiences drive you to change your mobile service provider?
k) Would these experiences drive you to change your wifi service provider?
l) Data usage?

The number of customers being small, simple decision trees could be formed based on Annoyance, Quality and Churn risk. The figures below depict different decision trees that could be conceived.

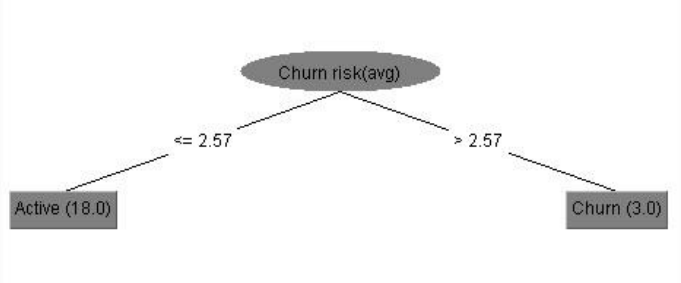


Figure 9. Decision tree with respect to churn risk

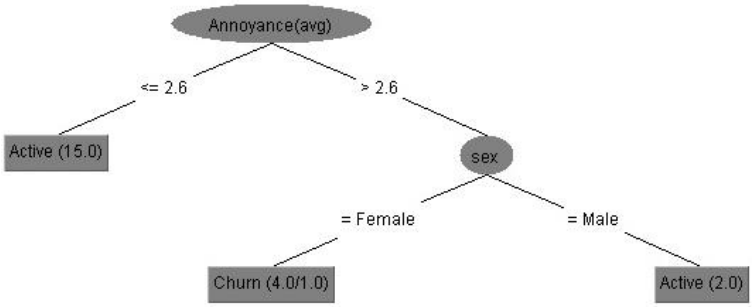


Figure 10. Decision tree with respect to annoyance

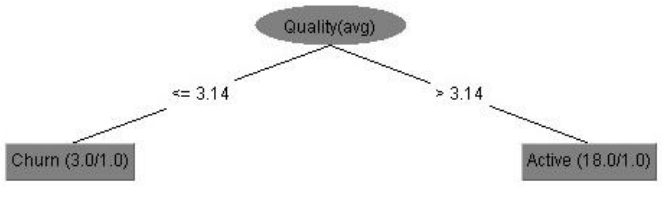


Figure 11. Decision tree with respect to quality

GROUP	PREDICTION OF	CORRECTLY CLASSIFIED
Churn Risk	Churn vs. Active	95.24%
Annoyance	Churn vs. Active	71.43%
Quality	Churn vs. Active	76.19%

Table 9. Decision tree analysis for Individual data usage

Table 9 gives a brief outline regarding the correct classification of decision trees carried out for Annoyance, Churn risk and Quality. A5, A6 and A7 in Appendix portrays the confusion matrices with respect to churn risk, annoyance and quality respectively.

5 ANALYSIS AND DISCUSSION

This chapter presents an expounded interpretation of statistical and decision tree analytics mentioned in the previous chapter. All the interpretations were carried out from three different data sources. These include:

a. Monthly data volumes for churned and active users from October 2015 to March 2016, provided by an anonymous Telecom provider.
b. Results of a global Telecom survey carried out by accompanying thesis [28].
c. Individual analysis results from Android based tool for weekly QoE polls and volume measurements [27].

5.1 Anonymous Telecom provider

Starting with the data from an anonymous Telecom provider, this data acquisition itself was a biggest task because of their fear of misusing it, concerning to customer's confidentiality and company's security issues. We were provided with monthly data usage volumes (in megabytes) of small and medium enterprises from October,2015 to March,2016. Initially only data about the churners were provided, which was later augmented with the active users.

Tables 2 and 5 from previous chapter denotes the statistical analytics for the total churners and active users respectively. This is the raw data without any data pre-processing. From Table2, it could be seen that the mean volumes for all the months were of similar order ranging from 2551 to 3221. Same is the case with standard deviations as well. Calculating the 95% confidence intervals, it could be observed that, there is either a significant overlap, or the confidence intervals are coming very close to each other. As discussed earlier, noteworthy differences could not be observed. The lag-1 autocorrelations were mostly positive and not too small either, which points at the risk that the real confidence intervals may be larger than the estimations presented here. Therefore, no vital statements could be made regarding these volumes. Same is the case with active users which could be observed from Table 5, with highly overlapping confidence intervals. But, they have rather small autocorrelations that tends to reliable confidence intervals. These results are therefore in the need for data-processing, hoping for better trends in confidence intervals and correlations.

A total of 4106 churned customer mobile data usage volumes were given, out of which 1231 were users with zeros volumes for all the six months. Such values add to redundancy and needs to be deducted. As per the norms of this Telecom provider, a user needs to decide about the churn three months prior the cancellation. In this regard, the last quarter of 2015, that includes October, November and December become the deciding factors for churn prediction. Therefore, eliminating posts related to customers with missing values, duplicates and zero volumes during two or three months in last quarter of 2015 avoid useless calculations. Upon data-preprocessing, there were a total of 2670 churned customers and 2454 active customers. The active users with zero volumes are the immediate probable churners that the provider needs to focus on.

From the total 4106 churners data provided initially, only 673 customers have given the actual reasons for churn, which is just 16% of the total churners. A total of 16 different reasons for churn could be found from Table1 in the previous chapter. Table 1 provides with the churned customers who were kind enough to cordially give a proper reason for churn. After data-processing removing zeros and missing values, only 454 remained with reasons, which is still just about 17% of the total 2670 churners. There were quite many duplicates in

these 454 churners with reasons. Excluding these duplicates, there remained only 404 churners with reasons.

The figure below is an illustration of churners with reasons during October 2015 to March 2016. It is quite clear that Altered needs were of major concern that made these enterprises churn to a fellow provider constituting a major share of 46% of the total reasons. It was followed by Better pricing opportunities from fellow Telecom providers constituting 22% and Poor service comprising 10%. It proves that the customers won't hesitate to switch providers if they don't find what they are looking for [1].

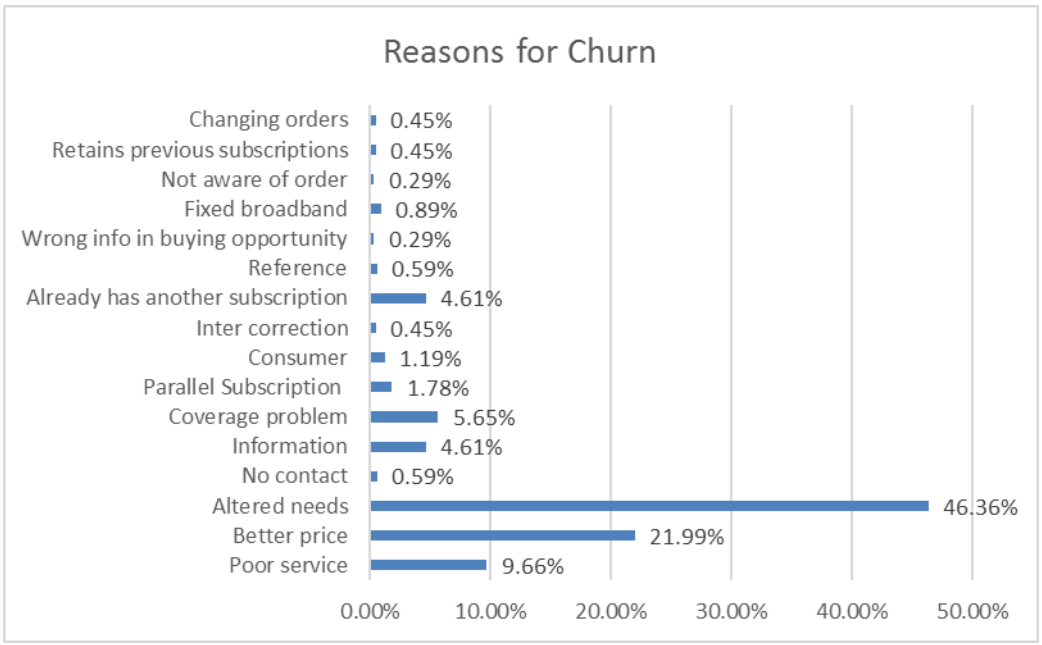


Figure 12. Reasons for churn

Tables 3 and 6 portrays the data of churned customers and active users after data-processing. As discussed earlier, only the last quarter of 2015 become the deciding factor of churn, preeminent emphasis was laid on December, November and October. From Table 3, it is evident that there is a decrease in autocorrelation from October to December, which is a rather positive sign that provides more satisfying confidence intervals, but the confidence intervals of December and October are completely overlapping each other, along with three quarter share overlap with November. Therefore, no strong affirmations could be given. The same is the case with active users in Table 6. Complete overlap in confidence intervals of October and December, with more than half overlap with November. Autocorrelations are quite small favoring better confidence intervals, but these correlations decreases from October to November, again increases from November to December. Such overlaps of confidence intervals in spite of small autocorrelations indicate unclear trends.

In the view of progressing the results interpretation, data normalization was proposed. Normalization refers to the creation of shifted and scaled versions of statistics, for better comparability of usage trends between months. It is just a simple technique of processing datasets such that the results are clear and univocal. Normalizations improve the data integrity so that the overall picture is not dominated by the heavy users. Heavy users here refer to the users with unusual high data usage trends during the last quarter of 2015. Numerous trials have been done for various combinations of datasets. Recollecting the decision factor period of last quarter of 2015, average monthly usage volume of every customer was divided with the average of October, November and December. Light usage users have been eliminated. Light users here refer to the users with minutest data usage volumes and with two or three zeros in aggregate monthly data usage volumes during

October, November and December. Tables 4 and 7 represent the normalized statistics of churners and active users. Mean data usage were of similar order in both the cases. In the case of churners, from confidence intervals, there is a fair overlap for October to December, but the values are very close-by, therefore no much significance could be noticed. Autocorrelations though being small adding to reliability in confidence intervals, show bizarre trends by going down and up. In the case of active users, autocorrelations decreased from October to December, strengthening the trust in confidence intervals. But there is a large overlap of confidence intervals amongst every month. In consideration of overlapping, proximate and bordering confidence intervals that misleads the robust meager autocorrelations, weighty proclamations regarding churn could not be made.

Numerous decision trees have been tried for different combinations of datasets, few with and without normalization, with and without data preprocessing, churners alone, churners with reasons alone, active and churners together including and excluding zeros, etc. Few are tabulated in Table 8. It was quite evident that combining active and churners together without any data preprocessing and data preparation (normalization), no decision tree could be formed. The reason could be due to randomness and redundant values. When the monthly usage volumes were normalized with average of last quarter of 2015, a decision tree with almost 52% accuracy could be achieved. Figure 8 from previous chapter represents this tree. The screenshot in Appendix A4 depicts the correctly and incorrectly classified instances along with TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area, PRC Area, MCC that have been calculated from the confusion matrix [29].

- **TP Rate**: Rate of true positives (instances correctly classified as a given class)
- **FP Rate**: Rate of false positives (instances falsely classified as a given class)
- **Precision**: Proportion of instances that are truly of a class divided by the total instances classified as that class
- **Recall**: Proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
- **F-Measure**: A combined measure for precision and recall calculated as (2 x Precision x Recall) / (Precision + Recall)
- **ROC Area**: ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC Area denotes area under the curve.
- **PRC Area**: Typically, Precision and Recall are inversely related, as Precision increases, Recall decreases and vice-versa. A balance between these two needs to be achieved, thus the precision-recall curves come in handy. Area under this curve is called PRC Area.
- **MCC**: **Matthews Correlation Coefficient** is a measure of the quality of binary classifications. It returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation [18].

The confusion matrix for this tree is shown below in table 10. Usually the diagonal elements are big and adjacent elements are small. Here, all the values are of same order, which clearly depicts what 52% means. Therefore, notable statements could not be made regarding the churn prediction.

Confusion Matrix	Predicted Active	Predicted Churn
Actual Active	1026	1428
Actual Churn	1044	1625

Table 10. Confusion Matrix for Telecom provider's data

5.2 Telecom Survey

A personal survey was conducted in various countries to cognize the reasons for churn from one Telecom provider to another. A total of 788 customers from about 15 countries participated in our survey. Data preprocessing including removal of unwanted data, deduction of wrong entries, zeros and error correction were made. Numerous details of a total of 770 customers with about 45 Telecom providers from various countries were tabulated. Reasons includes the problems with data and calls that made them churn. After grouping similar reasons together and allocating individual values, decision tree analysis was made. Allocation of alphabets for similarly grouped reasons is shown in the following table.

No.	Feedback	Alphabet
1	Bad customer care support	B
2	Disturbances in call/video streaming, Call drops or misconnections, Too long connecting or Loading time, Disturbances (Freezes/Buffering/Waiting time)	D
3	Bad signal strength/Coverage, No 3G/4G coverage, Poor network, Less coverage for 3G services	S
4	Just for change	J
5	Economic rates from other mobile operators, Not better than other providers	E
6	Network issues even when there is a signal, Insufficient speed	N
7	High call rates & poor offers, Prepaid was too pricey, Call rates too high, Choices of amount of data plans	H
8	Plans that keep changing/ Increase of offers, Internet packages too expensive, Expensive data plan	P
9	Change of country	C
10	They rip your money off & annoying advertisements	A
11	GPRS problem, Internet problem	G
12	Roaming is expensive	R
13	No problem, Pleased with calls, All okay	O

Table 11. Allocation of alphabets for similarly grouped reasons

Without clear distinction between the reasons for problems with call and data, we get an accuracy of 41.69%. With clear distinction between their classes, like Ddata for disturbance problems with respect to data and Dcall for disturbances with respect to calls, we get an accuracy of 69.74%. The tree illustrates the current Telecom provider to which the user might tend to churn. Owing to size and complexity of decision tree and confusion matrix, the resulting decision tree could not be shown here. Instead, the interested reader is referred to appendix A3.

5.3 Individual data usage Analysis

A personal survey was conducted to investigate the relationship between session volumes and QoE with respect to customers and analyze the possibility of churn. 22 Android users were asked to install a network analyzer and their individual week data was collected in addition with their quality, annoyance and churn risk scores. A total of 8week data was tabulated and correlations were made between Quality-Annoyance, Quality-Churn risk, Annoyance-Churn risk. At the end of eight weeks, surprisingly three customers turned out to churn and the remaining nineteen were still active.
After data-preprocessing, one customer with unusual and missing values has been eliminated. Based on the tabulated analysis sheet from Hemanth Kumar Ravuri's appendix [27], various decision tree analyses were made. The trees were represented in previous chapter. Table depicts the decision tree analysis for individual data usage.
With respect to average churn risk score, we get an accuracy of 95.23%. With churn risk score less than or equal to 2.57, eighteen users were active and for churn risk greater than 2.57, three turned to churn out. Confusion matrix for this tree is depicted in appendix A5.

With respect to average annoyance score, we get an accuracy of 71.43%. With annoyance less than or equal to 2.6, fifteen users were stringently active. With annoyance greater than 2.6, two males were active and four females were shown to churn, which means one user was incorrectly classified. This user needs to be observed for further weeks, as she might be a probable churner in near future. Confusion matrix for this tree is presented in appendix A6.

With respect to average quality score, we get an accuracy of 76.19%. With quality score greater than 3.14, eighteen users were active, unfortunately with one misclassification and with quality score less than or equal to 3.14, three users have churned with one misclassification. The confusion matrix is displayed in appendix A7. These misclassifications need to be keenly observed for further weeks to investigate the possibility of churn.

5.4 ANSWERS TO RESEARCH QUESTIONS

RA1. The data normalization is a part of data preparation. It is a simple process of organizing datasets such that the results are unambiguous and clear. In the case of churn prediction, normalization plays a dominant role. The acquired datasets from the anonymous Telecom provider could not be directly applied to the churn prediction models, here J48 decision tree. Data preprocessing along with normalization are extremely indispensable. Normalization, here refers to the creation of shifted and scaled versions of statistics, for better comparability of usage trends between months. Normalization boosts the faith in autocorrelation and confidence interval values. It also improves the data integrity so that the overall picture is not totally dominated by the heavy users, namely the users with unusual high data usage trends during the last quarter of 2015. From decision tree analysis, it was quite evident that combining active and churners together without any data preprocessing and data preparation (normalization), no decision tree could be formed. The reason could be due to randomness and redundant values. WEKA tool doesn't accept values with utter chaos. They need to be properly organized without any blanks between columns, etc. All such erroneous values are taken care of, by data preprocessing. A decision tree for preprocessed data just gave about 33% accuracy. When the monthly usage volumes were normalized with average of last quarter of 2015, a decision tree with almost 52% accuracy could be achieved. Figure 8 from previous chapters represents this tree. This explains the sole importance of data normalization that makes it so crucial for churn prediction.

RA2. We were provided with three data sources to carry out various analyses, out of which we get data usage volumes from two sources. First source comprises the monthly data usage volumes (in megabytes) of small and medium enterprises from October,2015 to March,2016 by an anonymous Telecom provider. The preprocessed monthly usage volumes were normalized with average of last quarter of 2015. Analyses including statistical and data mining analytics, decision tree (J48) were constructed. From the statistics of normalized volumes, confidence intervals were overlapping and close by, therefore no much significance could be noticed. Though autocorrelations were reasonably small, which is underpinning the credibility of the confidence intervals, no strong trends could be observed. From decision tree analytics, a decision tree with just 52% accuracy could be achieved. This is not a great success in the field of data mining.

The next data source comprises of weekly data volumes of 22 android users from three different countries along with quality, annoyance and churn risk scores for a period of eight weeks. Different relationships between Quality and Annoyance, Quality and Churn risk, as well as Annoyance and Churn risk were analyzed. Surprisingly three customers turned out to churn. Based on the decision tree analysis, 95.23%, 71.42%, 76.19% accuracies were achieved with respect to average churn risk, annoyance and quality scores respectively when dealt with churn prediction. Though the number of users being limited, these percentages are quite appreciable.

When comparing the results from these two sources, so far, the monthly volumes have not shown much decision power. Monthly data usage volume analyses were unfortunately not strong enough. The outcomes looked random. Where as in the case of weekly data, confirmed trends could be observed. Increase in annoyance and churn risk scores leads to decrease in quality and data volumes. Analysis of weekly data usage volumes with customer's recent history and essential attributes can contribute to exceptional results for predicting churn.

RA3. Drawing conclusions from the outcomes of sections 5.1, 5.2 and 5.3 along with the answers to first and second research questions, the following convictions were made. Data preprocessing, data normalization and feature selection have shown to be prominently influential. Average Quality, Churn Risk and to some extent, Annoyance scores may point out a probable churner. "Happy users surf more and churn less" [4]. Weekly data volumes with customer's recent history and necessary attributes like age, gender, tenure, bill, contract, data plan, etc., are pivotal for churn prediction

6 CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Customer churn is always a grievous issue for the Telecom industry as customers do not hesitate to leave if they don't find what they are looking for. Customer churning is directly related to customer satisfaction. There is no standard model which addresses the issues of global telecom service providers accurately. Keeping all such things into consideration, a research thesis on customer churn prediction based on mobile data usage volumes with respect to QoE and users' perspective was studied. Statistical and J48 Decision trees from BigData analytics were proposed for analysis.

We were provided with three sources of data. Firstly, the monthly data volumes for churned and active users by anonymous Telecom provider for a six-month period starting from October 2015 to March 2016. Initially only data about the churners was given, which was later augmented with the active users. The acquired datasets from the anonymous Telecom provider could not be directly applied to the churn prediction models, here J48 decision tree. Data preprocessing along with normalization are extremely indispensable for better comparability of usage trends between months. From the statistics of normalized volumes, confidence intervals were overlapping and close by, therefore no much significance could be noticed. Though autocorrelations were small owing to reliable confidence intervals, no strong trends could be observed. From decision tree analytics, a decision tree with just 52% accuracy could be achieved.

Secondly, the results of surveyed data by accompanying thesis [28]. Preprocessed details of a total of 770 customers with about 45 Telecom providers from various countries were tabulated. 339 customers have churned from one Telecom provider to another. 271 customers are in a plan to churn in near future. Considering the reasons from already churned users in order to predict the probable churners, similar reasons were grouped together and allocated an alphabet to carry out decision tree analysis. Without clear distinction between the reasons for problems with call and data, we get an accuracy of 41.69%. With clear distinction between their classes, like Ddata for disturbance problems with respect to data and Dcall for disturbances with respect to calls, we get an accuracy of 69.74%.

Thirdly, weekly data volumes of 22 android users from three different countries along with quality, annoyance and churn risk scores for a period of eight weeks were noted [27]. Different relationships between Quality-Annoyance, Quality-Churn risk, Annoyance-Churn risk were analyzed. Surprisingly three customers turned out to churn. Based on the decision tree analysis, 95.23%, 71.42%, 76.19% accuracies were achieved with respect to average churn risk, annoyance and quality scores respectively when dealt with churn prediction. Though the number of users being limited, these percentages are quite appreciable.

Confirmed trends observed through correlations: As Quality increases, Volume increases, accordingly Annoyance and Churn risk decreases.

Data preprocessing, data normalization and feature selection have shown to be prominently influential. Average Quality, Churn Risk and to some extent, Annoyance scores may point out a probable churner. The bigger the screen, higher the data consumption. Weekly data volumes with customer's recent history and necessary attributes like age, gender, tenure, bill, contract, data plan, etc., are pivotal for churn prediction.

6.2 FUTURE WORK

Inclusion of more data samples to third data source might increase confidence in quantitative results. Though the problem of customer churn was addressed by many researchers in numerous ways, still there is no standard model which addresses the issues of global telecom service providers accurately. There is lot of scope for development of such a model, which could take the above mentioned factors and many more into consideration.

7 REFERENCES

[1] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," in *2013 Eighth International Conference on Digital Information Management (ICDIM)*, 2013, pp. 131–136.

[2] W. Yu, D. N. Jutla, and S. C. Sivakumar, "A churn-strategy alignment model for managers in mobile telecom," in *Communication Networks and Services Research Conference, 2005. Proceedings of the 3rd Annual*, 2005, pp. 48–53.

[3] D. Collange, M. Hajji, J. Shaikh, M. Fiedler, and P. Arlos, "User impatience and network performance," in *2012 8th EURO-NGI Conference on Next Generation Internet (NGI)*, 2012, pp. 141–148.

[4] J. Shaikh, M. Fiedler, and D. Collange, "Quality of Experience from user and network perspectives," Annals of Telecommunications, 65(1—2):47—57, Jan./Feb. 2010. [Accessed: 12-Feb-2016].

[5] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, "Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty," in *2015 IEEE International Congress on Big Data (BigData Congress)*, 2015, pp. 677–680.

[6] A. Idris and A. Khan, "Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers," in *Multitopic Conference (INMIC), 2012 15th International*, 2012, pp. 23–27.

[7] A. Idris, A. Khan, and Y. S. Lee, "Genetic Programming and Adaboosting based churn prediction for Telecom," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 1328–1332.

[8] I. Brânduşoiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in the pre-paid mobile telecommunications industry," in *2016 International Conference on Communications (COMM)*, 2016, pp. 97–100.

[9] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015, pp. 1–6.

[10] U. Yabas and H. C. Cankaya, "Churn prediction in subscriber management for mobile and wireless communications services," in *2013 IEEE Globecom Workshops (GC Wkshps)*, 2013, pp. 991–995.

[11] Y. He, Z. He, and D. Zhang, "A Study on Prediction of Customer Churn in Fixed Communication Network Based on Data Mining," in *Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09*, 2009, vol. 1, pp. 92–94.

[12] X. Zhang, Z. Liu, X. Yang, W. Shi, and Q. Wang, "Predicting customer churn by integrating the effect of the customer contact network," in *2010 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI)*, 2010, pp. 392–397.

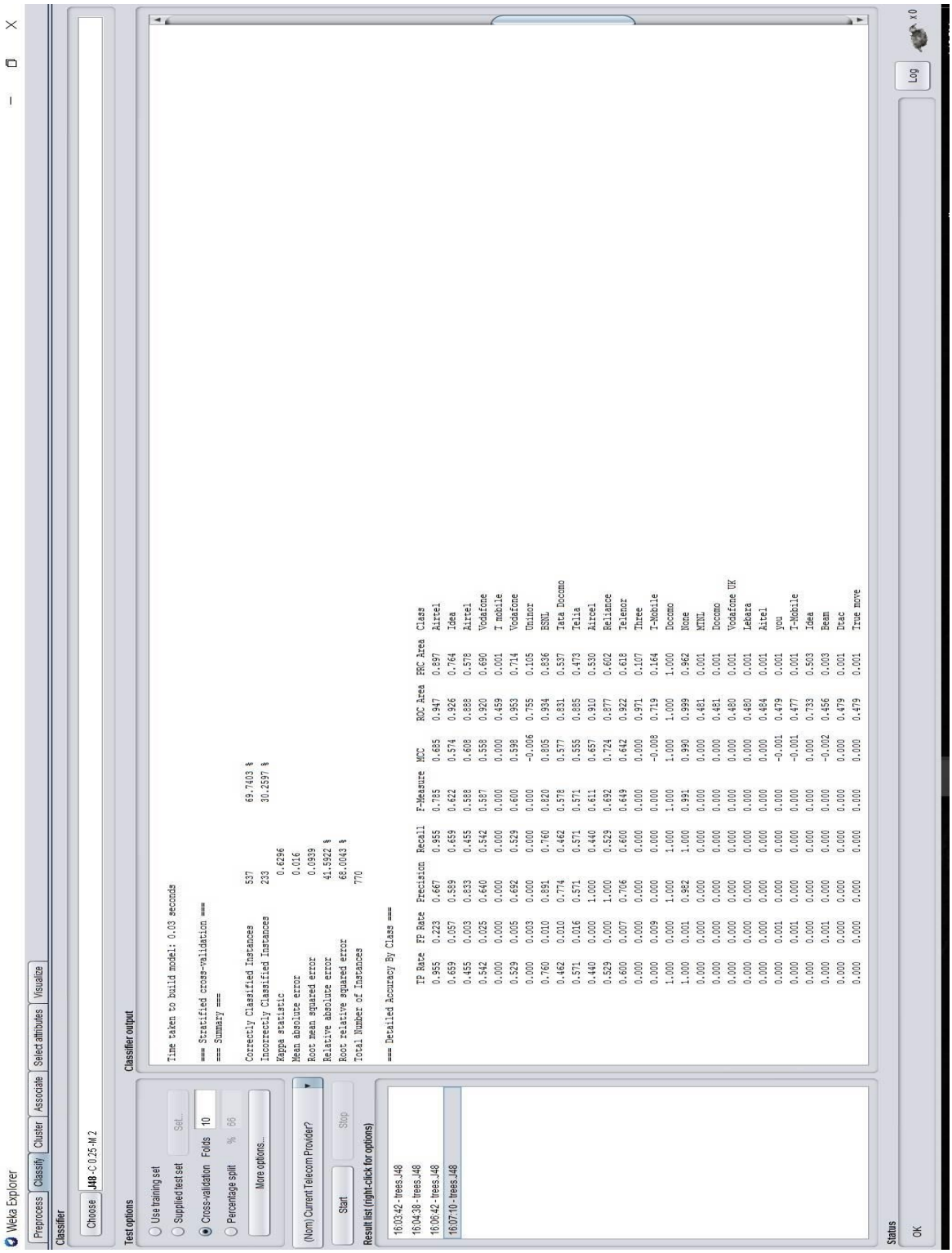
[13] M. J. Perez and W. T. Flannery, "A study of the relationships between service failures and customer churn in a telecommunications environment," in *PICMET '09 - 2009 Portland International Conference on Management of Engineering Technology*, 2009, pp. 3334–3342.

[14] L. Bin, S. Peiji, and L. Juan, "Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service," in *2007 International Conference on Service Systems and Service Management*, 2007, pp. 1–5.

[15] "AUTOCORRELATION" [Online]. Available: http://www.ltrr.arizona.edu/~dmeko/notes_3.pdf. [Accessed: 14-Sep-2016].
[16] W. M. C. Bandara, A. S. Perera, and D. Alahakoon, "Churn prediction methodologies in the telecommunications sector: A survey," in *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2013, pp. 172–176.
[17] A. Idris and A. Khan, "Ensemble Based Efficient Churn Prediction Model for Telecom," in *2014 12th International Conference on Frontiers of Information Technology (FIT)*, 2014, pp. 238–244.
[18] "Matthews correlation coefficient," *Wikipedia, the free encyclopedia*. 08-Sep-2016.
[19] "Statistical Analysis - What is it?" [Online]. Available: http://www.sas.com/en_us/insights/analytics/statistical-analysis.html. [Accessed: 14-Sep-2016].
[20] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 14-Sep-2016].
[21] Sam Drazin and Matt Montag, "Decision Tree Analysis using Weka." *Machine Learning-Project II, University of Miami*, pp.1-3
[22] "Overview of the KDD Process" [Online]. Available: http://www.ryerson.ca/~rmichon/mkt700/readings/KDD%20Process%20Overview.html. [Accessed: 14-Sep-2016].
[23] "Data Mining: Practical Machine Learning Tools and Techniques, Second Edition - Data Mining Practical Machine Learning Tools and Techniques - WEKA.pdf."
[24] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), pp.10-18.
[25] "C4.5 algorithm," *Wikipedia, the free encyclopedia*. 27-Jun-2016.
[26] Kumar, R. and Verma, R., 2008. KDD Techniques: A survey. *International Journal of Electronics and Computer Science Engineering, IJECSE*, *1*, pp.2042-2046.
[27] Hemanth Kumar Ravuri, "Study of user's data volume as function of Quality of Experience for churn prediction" [M.Sc.E.E. thesis 2016:32, BTH, submitted].
[28] Mounika Reddy Chandiri, "Churn predictive heuristics from Telecom operator and users' perspective" [M.Sc.E.E. thesis 2016:05, BTH, submitted].
[29] "Weka Data Analysis" [Online]. Available: http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaDataAnalysis.pdf. [Accessed: 29-Sep-2016].

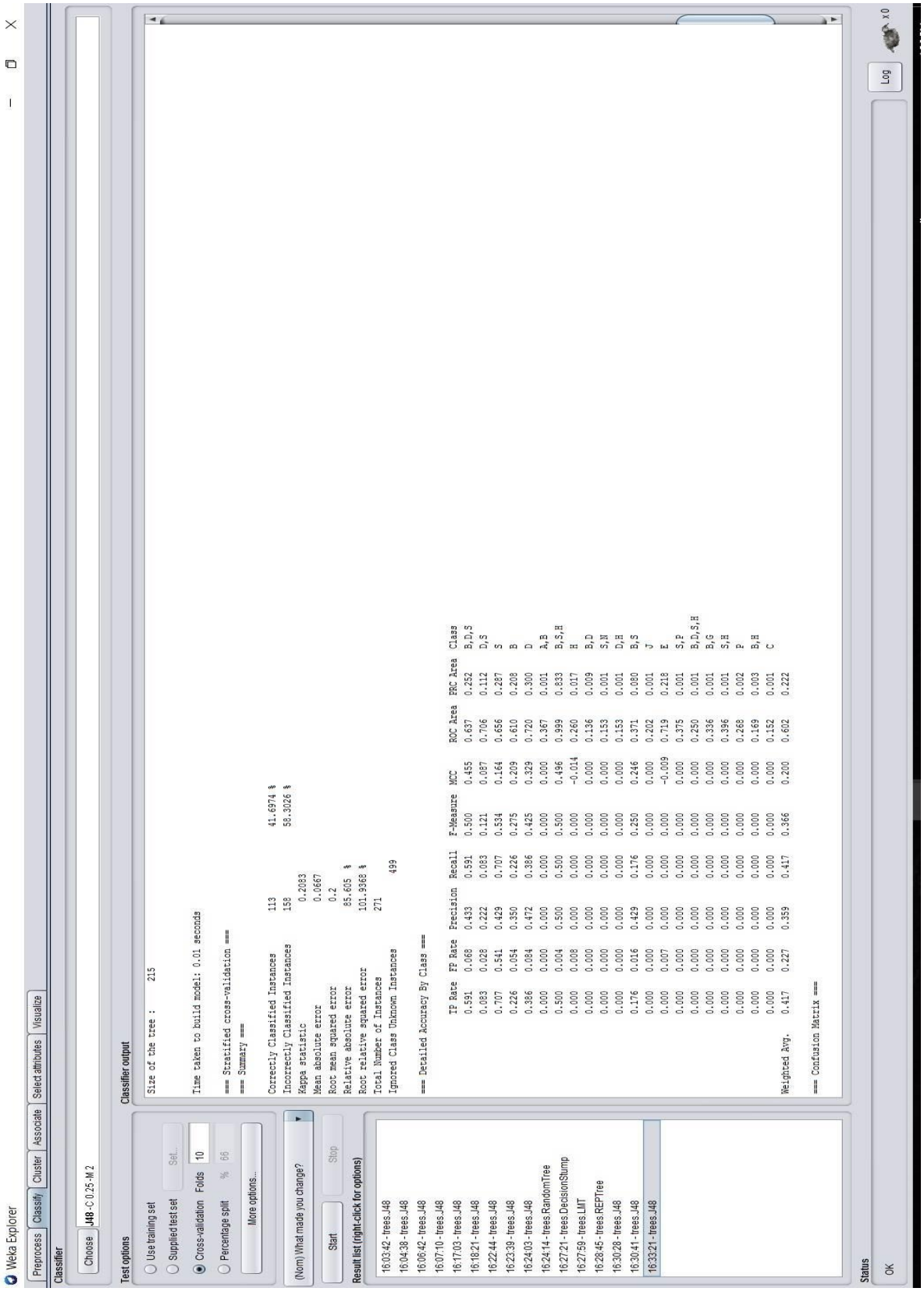
8 APPENDIX

A1. Decision tree analysis for surveyed data



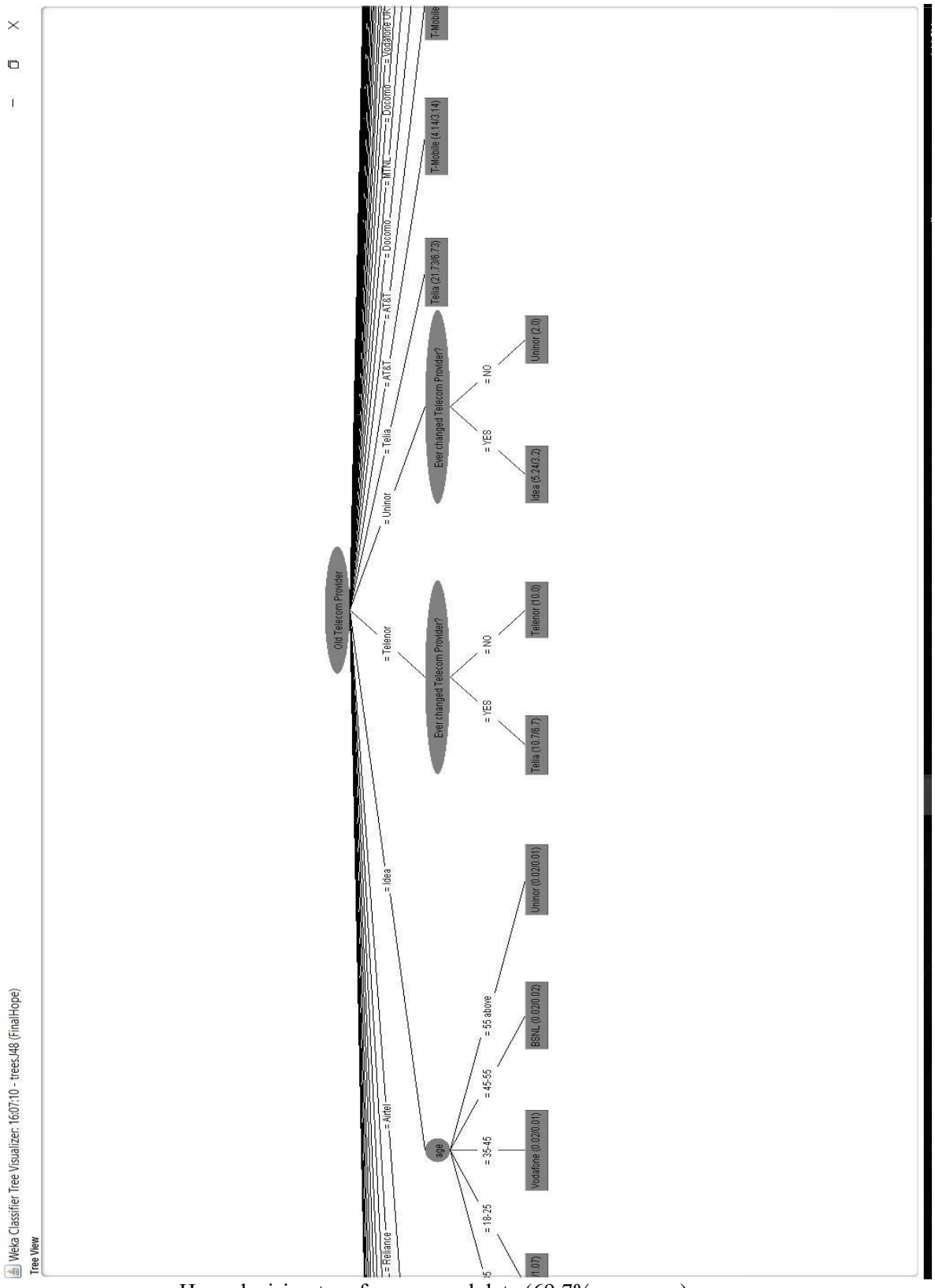
Decision tree with distinction between calls and data. (for sections 4.2 and 5.2)

A2. Decision tree analysis for surveyed data



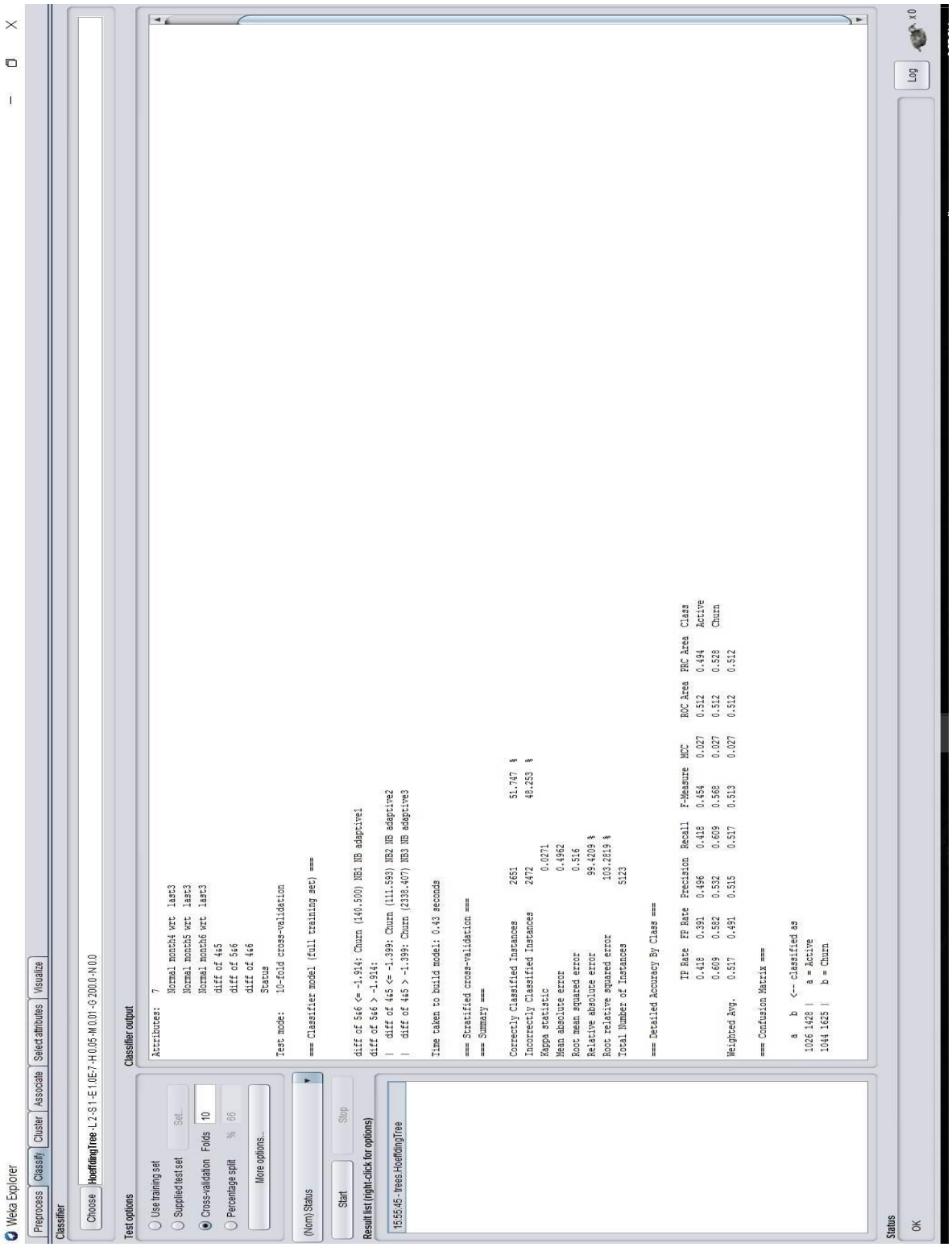
Decision tree without distinction between calls and data (for sections 4.2 and 5.2)

A3. Huge Decision tree for surveyed data



Huge decision tree for surveyed data (69.7% accuracy)

A4. Decision tree analysis for anonymous Telecom provider



TP rate, FP rate, Precision, Recall, etc have been calculated from the confusion matrix.
(for section 5.1)

A5. Decision tree analysis for individual data analysis (Churn risk)

Confusion Matrix	Predicted Churn	Predicted Active
Actual Churn	3	0
Actual Active	1	17

Confusion matrix with respect to churn risk

A6. Decision tree analysis for individual data analysis (Annoyance)

Confusion Matrix	Predicted Churn	Predicted Active
Actual Churn	1	2
Actual Active	4	14

Confusion matrix with respect to annoyance

A7. Decision tree analysis for individual data analysis (Quality)

Confusion Matrix	Predicted Churn	Predicted Active
Actual Churn	1	2
Actual Active	4	14

Confusion matrix with respect to quality