ISSN: 2321-8169
4705 - 4713

# Efficient Extraction of Actionable Data using Decision Tree

KK Tangod
Department of Information Science & Engineering
Gogte Institute of Technology
Belagavi, Karnataka, India
*kirankt@git.edu*

GH Kulkarni
Department of Electrical & Electronics
Jain Engineering College
Belagavi, Karnataka, India
*ghkulkarni1@rediffmail.com*

***Abstract*—** Extracting Actionable Knowledge from decision tree is aimed towards providing a novel algorithm which can predict churn in telecommunication industry which helps in maximizing the expected net profit. The approach we take, integrates a mathematical model which predicts the profitable customers using the transaction data and a novel data mining algorithm which would predict whether a profitable customer is going to churn and taking necessary actions directly on top of data mining results in post processing step where, there by suggesting the special promotional offers and discounts only to profitable customers to maximize the expected net profit.

***Keywords***-*churn, pre-processing, Classification, Entropy, Cumulative-Net-Present-Value(cnpv)*

*****

## I. INTRODUCTION

The digital revolution has given rise to a host of technologies that are transforming marketing practices. Powerful databases and electronic data networks are allowing companies to collect concise information about customers and their buying patterns more effectively and efficiently than ever before. The Internet, in particular, has increased the ability of firms to track the behavior of individual consumers as they visit numerous Web pages.

In order to reap the benefits of detailed customer knowledge, firms need to systematically estimate the profitability associated with its use. The ultimate goal is to develop highly committed customers who not only make repeat purchases and generate continual revenue streams, but also require minimal maintenance along the way. It is entirely possible that while some customers do not bring profits with their initial purchases, the margins from their future expected transactions paint a different picture. As a result, firms need to track initial acquisition costs and compare them to the profits to be generated over the customer's expected relationship with the company. The above activities allow marketers to decide which customers to go after, how to change the promotional mix as a function of past and recent transactions and, if necessary, which specific customers to discontinue serving. Indeed, many practitioners and scholars are expressing the view that marketing is rapidly becoming "the science and art of finding, retaining, and growing profitable customers".

For many firms in a very competitive market and business environment, customer churn management becomes one of the most critical success factors mainly due to higher acquisition costs for new customers. In particular, with exceptionally high annual churn rates (20–40%), the firms in the mobile telecommunications industry try to develop predictive models that accurately identify which customers are most likely to terminate the current relationship. Therefore, the optimal selection of customer targets (e.g., most probable churners) and marketing campaign size where the profit of marketing campaign is maximized has been considered a critical success factor in the Customer Relationship Management (CRM) program.

Many marketing managers still select their targets either intuitively or based on the long-standing methods such as cross tabulation or RFM (Regency, Frequency, Monetary), a segmentation based database marketing method. For example, in RFM analysis, marketing managers first divide customers into a total 125 groups (5 groups for each of three past purchasing pattern dimension) in such a way that customers within the same group are similar in terms of past purchasing patterns such as how recently they have made a purchase (Regency), how often they have made purchases (Frequency), and how much they have spent (Monetary). Then, marketers narrow down their focus on customer groups who have recently made a purchase, have made multiple purchases, and have spent more money in their customer lifetime. This RFM analysis and its variants have been successfully used for more than 30 years because they are simple and easy to apply. However, they are very limited in the sense that they cannot be applied to new customers in target markets because no RFM information is available for new customers. In addition, they do not provide any insights about new customers and markets, and hence cannot be used to expand customer bases because marketers shift their focus on only customers with high scores of RFM variables, which, in return, will boost the values of RFM scores for targeted customers while lowering down those of customers with low RFM scores. Finally, RFM analysis may not be applicable to certain products. For example, customers who recently purchased an expensive car are not likely to buy another car very soon. In contrast, customers who purchased a car several years ago are more likely to purchase another car.

Fortunately, the rapid development of new data mining and business intelligence models over two decades in computer science, cognitive science, and information system area makes it possible for marketing managers to disseminate micromarketing messages targeted toward a specific group of households that are most likely to open to the customized incentives. When a churn management program is operated successfully, marketing managers can further identify most likely churners in advance, and then develop and present the right retention offers customized across different customer segments. Both marketing and data mining researchers have presented various database marketing approaches for successful CRM programs. Several studies presented models to target households by using the knowledge extracted from the customer‘s purchase history. In other studies the profitability condition of a campaign was explicitly formulated as a function of the model performance along with campaign cost and revenue factors such as mailing costs and marginal revenue per identified positive record. In another study, a business intelligence model was presented to identify as many customers as possible who will respond to a specific solicitation campaign letter by studying the effects of variable selection and class distribution on the performance both of a primitive classifier system and of a relatively more sophisticated classifier system.

In particular, several studies aim to provide accurate models for churn prediction in the telecommunication industry based on various statistical and data mining models such as neural network models, support vector machines, or clustering algorithms along with customer demographics, billing, and call detail records. However, most of these models have not focused on identifying critical success factors but on building an accurate churn prediction model only, which may only serve as the first step of developing churn marketing programs.

## II. Related Work

The first and foremost step in determining customer profitability is to have clear sense of characteristics of customer activities or transactions, which provides the firms with information on the buying patterns of both new and existing customers. The lifetime value is calculated based on the cumulative net present value. Based on which the classification of customers as profitable and non profitable is done [4].

Predicting customer life time value involves modeling the churn hazard as a function of tenure and other customer attributes. The lifetime values are computed using future time intervals with the hazard model. Hazard models are used to estimate the shape of the hazard function (the time effect) and how the shape is affected by the co-variants. Hazard models in prediction can be used to score customers. The inputs are the current customer tenure and the current values of the other co-variants. Output of this is the churn hazard at that time and probability of churn.

The definition of customer lifetime value is suitable for businesses with contractual products such as banking, insurance, and telecom. The aim of this work [5] is to determine the customer who intends to churn, and to create specific campaigns to them by using a customer data of a major telecommunication firm. To determine the reasons for the churning of the customer logistic regression and decision trees analysis, neural networks, clustering techniques are employed.

Telecom service providers will want to meet the needs of customers and retain customers. This dissertation use data exploration technology to build a predictive model, find out the possible churners and provide personalized service. Data mining is used in churn analysis to predict whether a particular customer will churn, when churn is expected to happen and reasons for churn as suggested in [3].

In any firm, the business goal has to be measurable, reachable, realistic and well-timed. Next is to translate the business problems into data mining problems, where churn data modeling is done. So churn predictive modeling refers to the task of building a model for the target variable (churner or non-churner) as a function of the explanatory variables as intimated in [6, 8, and 13]. That is Customer data in the data warehouse will be extracted and explored to identify data items of interest. Next designing of programming scripts to extract and transform data into the desired form after the variables of interest will be selected. Next Churn prediction and value modeling is done.

In a Churn analysis applications, the first activity is to access to the customer data. Then, factors are classified to decide which factor or factors affect customer churn decision. After determining which customers are likely to churn, different and specific marketing and retention strategies can be applied to the target customers, in a defined time period. Based on historical data these methods try to find patterns which can point out possible churners. There are four data set variables: behavior of the customer, perception of the customer, demographic details and other environment variables. Several categories of prediction models in statistics and data mining that can be applied to churn prediction problems. Researches show that Decision Tree is well suited to this type of problem [9, 11].

Exploring data analysis, then extracting of the data, creating predictive model using decision tree, test models, verify its effectiveness and stability this is done using decision tree classification algorithm C5.0 is used. There are two steps in classification of decision tree; first step, use the training set to establish a decision tree and second step, use the decision tree to make classification to input record. Create a classification C5.0 algorithm to minimize expected cost of classification error rates [2, 7].

Demographic features have the lowest affect on the churn prediction. But it was not easy to get to the conclusion about the billing and usage features. It seems that these two types are equally important for predicting churn [1].

Once the probable churn customer list is predicted the next step is to explore the reduction of churning through the use of data mining for marketing plan development. This is done by identifying combinations of voice mail and/or international calling plan packages that could effectively be offered to the various customer groups depending on the usage [10]. With information pattern gained from mining of call data, the firm expects to identify which of the plan-customer group pairs have the lowest churn rates. Such pattern identification enables the firm to optimize its marketing strategy and promote the special offers for churn reduction. Company should make voice-mail a standard part of all calling plans. The company needs to target improvements to customer service as a way to reduce churning.

## III. System Design and Implementation

### A. *Designing of system:*

The first step is the Knowledge Discovery process which would start with collecting the relevant data and end up with the valuable knowledge, which in this case is the probable customer churn list who must be retained by the company to remain profitable. Once the data is collected from different sources it has to be pre-processed and transformed into required format, after the data has been transformed it will be the input to the data mining model which would provide the valuable results. The two models applied to the pre-processed data are explained below.

The first and foremost step in determining customer profitability is to have a clear sense of characteristics of customer activities or transactions, which provide the firms with information on the buying patterns of both new and existing customers. Based on these customer activities cumulative net present value is calculated. Depending on which the customers are classified as profitable and non profitable.

Telecom service providers should meet the needs of customers and also aim at retaining them. This proposed system uses the data exploration technology to build a predictive model which will help in find out the possible churners and also personalized services can be provided. Data mining technique is employed in churn analysis to predict whether a particular customer will churn or not.

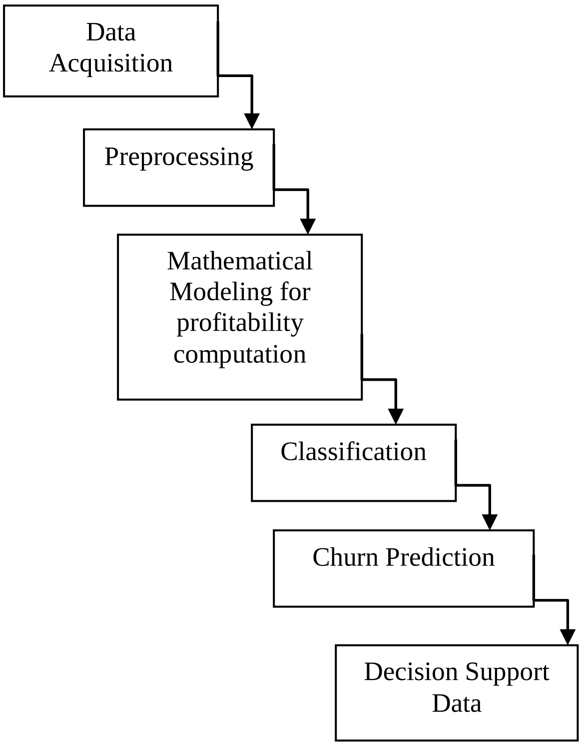


Fig. 1 : System Architecture

Fig: 1 shows the overall system architecture and the building blocks are explained as follows:

*Data Acquisition:* The data is collected from various sources with respect to telecommunication domain like customers and their respective transactions and usage data.

*Pre-Processing:* The data will be usually having large set of attributes. Some of the attributes the value may be missing, these missing values are treated in the Pre-Processing step.

*Mathematical Model for Profit Computation:* One way to think about measuring a company's promise for long-term economic growth is to measure its customers by computing customer profitability which can be computed mathematically using customer life time value method.

*Classification(using Decision Tree):* Once the customer profitability is computed the next step is to classify the customer into various classes or groups using decision trees.

*Churn Prediction:* Churn prediction helps in knowing which customer may churn prior to losing him. Because retaining the existing customer is much easier and cost effective than acquiring a new customer.

*Decision Support data:* Churn prediction helps in further marketing strategy. Customers who are predicted to get churned can be suggested with various offers and discounts as a part of marketing strategy and can be attracted.

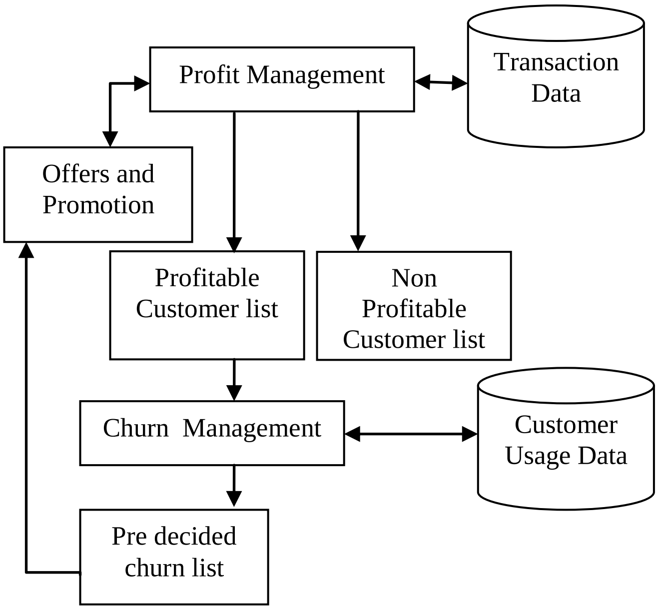


Fig. 2 : Profit management system

Block diagram shown in the Fig. 2 depicts various processes involved in the system. In Profit Management process the transaction data is accessed this data is further processed by the system to classify customers into Profitable or Non-Profitable customer based on the Mathematical model, Only Profitable customers are considered for the next process by filtering out Non-Profitable customers. Profitable customer usage details(Total-day -min, total-night-min, Total-day-cost so on 15 more attribute details are given) are given as the input for the Churn Management process where the customers are classified as Churner or Non Churners based on the decision tree algorithm(C4.5 algorithm). Once Churner details are obtained the system suggests offers and Promotions to Churners who are profitable because acquiring new customer is costlier then retaining the old customer.

### B. *Implementation:*

Instance of Customer Information:

| Cust_id | Name | Address | Status |
|---|---|---|---|
| 1 | Atul | Jharkhand | Active |
| 2 | Amit | Rajastan | Active |
| 3 | Aniruddh | Maharastra | Active |
| 4 | Ram | Karnataka | Active |
| 5 | Priti | Assam | Active |
| 6 | Sameer | Karnataka | Churn |

Table 1: Customer Information

Instance of Plan table:

| Plan_id | Cost | Margin |
|---|---|---|
| 1 | 75 | 7 |
| 2 | 100 | 10 |
| 3 | 50 | 4 |
| 4 | 150 | 15 |

Table 2 : Instance of Plan

Instance of Transaction Table

| cust_id | Transaction | plan_id | date_of_tran |
|---|---|---|---|
| 1 | 1 | 4 | 20-02-2013 |
| 1 | 2 | 2 | 01-03-2013 |
| 2 | 3 | 3 | 27-02-2013 |
| 1 | 4 | 1 | 01-01-2014 |
| 2 | 5 | 3 | 14-02-2014 |
| 3 | 6 | 2 | 29-01-2014 |
| 4 | 7 | 2 | 17-03-2014 |
| 5 | 8 | 4 | 11-02-2014 |
| 6 | 9 | 1 | 27-02-2013 |
| 7 | 10 | 3 | 10-02-2014 |
| 8 | 11 | 4 | 24-02-2014 |
| 9 | 12 | 3 | 29-01-2014 |
| 10 | 13 | 2 | 17-03-2014 |
| 7 | 14 | 1 | 27-02-2014 |
| 8 | 15 | 4 | 01-01-2014 |

Table 3: Instance of Transaction

Following features are derived from input attributes:
Purchase- Frequency=No. of times transaction done per 3 months
Average Margin= total margin ÷ No of plans subscribed
Retention Rate(R)=(total num of customers-new customers) ÷ total customers previous period
Projection of Retention rate for period t=(R) (t-1);
Present Value of Customer = [{(Average Margin*Purchase-Frequency)- Marketing_Cost}*R] ÷ (1+discount) t }]
Considering t=3
Customer_id=1
Avg_Margin=(7+10+15) ÷ 3 =10.67
Purchase_Frequency=3
Marketing_cost=3rs
Retention_rate=7 ÷ 10=0.7
Retention_rate_for_3_months=(0.7)3-1=0.49
Acquisition_cost=Marketing_cost/Response_rate =15
Present_value_of_customer=[{(10.67*3)-3}*0.49] ÷ (1.03)3=13
Net_present_value= Present_value_of_customer-Acquisition_cost=13-15=-2
For t=6
Retention_rate=(0.7)6-1=0.168
Present_value_of_customer=8.14
Cumulative_net_present_value=∑ Net_present_value=8.14-2=6.14
Since the Cumulative_Net_Present_Value is positive the customer is considered to be Profitable.

Example 2:

Considering t=3
Customer_id=3
Avg_Margin=10
Purchase_Frequency=1
Marketing_cost=3rs
Retention_rate=7 ÷ 10=0.7
Retention_rate_for_3_months=(0.7)3-1=0.49
Acquisition_cost=Marketing_cost/Response_rate=15
Present_value_of_customer= [{(10*1-3)*0.49}÷1.03)3] = 3.138
Net_present_value= Present_value_of_customer-Acquisition_cost=3.138-15 = -11.862
For t=6
Retention_rate=(0.7)6-1 = 0.168
Present_value_of_customer = 1.96
Cumulative_net_present_value = ∑Net_present_value = -11.862+1.96 = -9.902
Since the Cumulative_Net_Present_Value is negative the customer is considered to be Non-Profitable.
The following code in our system will fetch the data from the table and predict the profitable customers.
“rs1=st1.executeQuery("select * from trans where cust_id='"+str1.get(i)+"'"); “

The above statement fetches the transaction details about that particular customer id, the plan id’s are fetched from the result set and for each plan id the amount and margin details are fetched from the following result set obtained by the statement below.

“rs1=st1.executeQuery("select * from trans where cust_id='"+str1.get(i)+"'");”

Then the Cumulative-Net-Present value is calculated by following code.

```
“avg_margin/=freq;
 avg_amount/=freq;
 profit_per_cust=(avg_margin*freq-3)*0.49f;
 pv=profit_per_cust/(1.03f*1.03f*1.03f);
 npv=pv-6;
 profit_per_cust2=(avg_margin*(freq*2)-6)*0.168f;

pv2=profit_per_cust2/(1.03f*1.03f*1.03f*1.03f*1.03f*1.03f);
 npv2=pv2;
 cnpv=npv+npv2;”
```

Depending on this cnpv value the customer is classified as Profitable or Non-Profitable.

*1. Data Pre-Processing*

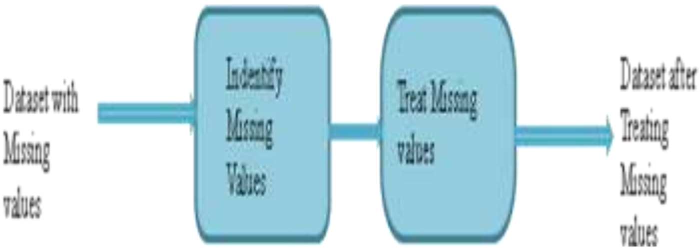


Fig. 3: Data Pre-Processing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost activity before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Missing data can occur because of non-response: no information is provided for several items or no information is provided for a whole unit. Some items are more sensitive for non-response than others, for example items about private subjects such as income. Sometimes missing values are caused by the researchers themselves, if data collection was not done properly or if mistakes were made with the data entry.

Our system contains a module to treat these missing values which identifies the missing values in the given dataset and replaces the null value with a particular format value. System treats the nominal and numerical value differently, numerical values are replaced by the mean value of that particular column and nominal values are replaced by the mod value of that particular column.

*2. Classifying Profitable and Non-Profitable customers*

Instance of Plan table:

| Plan_id | Cost | Margin |
|---|---|---|
| 1 | 75 | 7 |
| 2 | 100 | 10 |
| 3 | 50 | 4 |
| 4 | 150 | 15 |

Table 4: Instance of Plan

The proposed module in classifying Profitable and Non-Profitable customers intends to find the customer profitability based on transaction is to have a clear sense of relevant features of customer action. To this end, analysis of historical data can be very powerful in providing firms with information on the buying patterns of both new and existing customers.

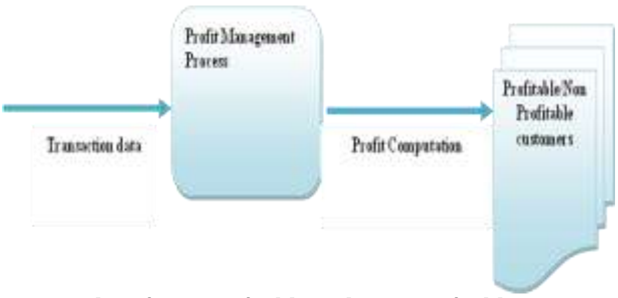


Fig. 4: Classifying Profitable and Non-profitable customers

Taking into account the responsiveness of particular customers to certain offerings, the cost of making these offerings, and the probability a customer is expected to keep generating revenue for the firm enables the calculation of customer profitability.

In this module, we have assumed values for following attributes: Marketing Cost, discount rate, response rate (E.g.: Marketing _Cost=3rs/3months, discount rate=3%, response rate=20%) The discount rate grants the seller to control for the fact money received in the future is not equal to the value of money received today. Money can be invested today and interest on it can be earned to adjust for discounting, awaited profits in totality for each customer should be deflated for each period by as factor of Dt

= ( 1+ i )t, Where 'i' is the discount rate, and 't' the number of periods we wait to receive the money.

Input for the customer profitability can be various transaction details. The following attributes related transaction that are used for computation are: Customer_id, status, Plan_id, transaction_id, date_of_transaction, cost and margin.

Instance of Customer Information:

| Cust_id | Name | Addr | Status |
|---|---|---|---|
| 1 | Atul | Jharkhand | Active |
| 2 | Amit | Rajastan | Active |
| 3 | Aniruddh | Maharastra | Active |
| 4 | Ram | Karnataka | Active |
| 5 | Priti | Assam | Active |
| 6 | Sameer | Karnataka | Churn |

Table 5: Customer Information

Instance of Transaction Table

| cust_id | Transaction | plan_id | date_of_tran |
|---|---|---|---|
| 1 | 1 | 4 | 20-02-2013 |
| 1 | 2 | 2 | 01-03-2013 |
| 2 | 3 | 3 | 27-02-2012 |
| 1 | 4 | 1 | 01-01-2013 |
| 2 | 5 | 3 | 14-02-2013 |
| 3 | 6 | 2 | 29-01-2013 |
| 4 | 7 | 2 | 17-03-2013 |
| 5 | 8 | 4 | 11-02-2013 |
| 6 | 9 | 1 | 27-02-2012 |
| 7 | 10 | 3 | 10-02-2013 |
| 8 | 11 | 4 | 24-02-2013 |
| 9 | 12 | 3 | 29-01-2013 |
| 10 | 13 | 2 | 17-03-2013 |
| 7 | 14 | 1 | 27-02-2012 |
| 8 | 15 | 4 | 01-01-2013 |

Table 6: Instance of Transaction

Following features are derived from input attributes:

Purchase- Frequency=No. of times transaction done per 3 months

Average Margin= total margin ÷ No of plans subscribed

Retention Rate(R)=(total num of customers-new customers) ÷ total customers previous period

Projection of Retention rate for period t=(R) (t-1);

Present Value of Customer = [{(Average Margin*Purchase-Frequency)-
Marketing_Cost}*R] ÷ (1+discount) t }]

Refer Table 1, Table 2, Table 3 for following computation:

Example 1:
Considering t=3
Customer_id=1
Avg_Margin=(7+10+15) ÷ 3 =10.67
Purchase_Frequency=3
Marketing_cost=3rs
Retention_rate=7 ÷ 10=0.7
Retention_rate_for_3_months=(0.7)3-1=0.49
Acquisition_cost=Marketing_cost/Response_rate=15
Present_value_of_customer=[{(10.67*3)-3}*0.49] ÷ (1.03)3=13
Net_present_value = Present_value_of_customer-Acquisition_cost=13-15=-2
For t=6
Retention_rate=(0.7)6-1=0.168
Present_value_of_customer=8.14
Cumulative_net_present_value=∑ Net_present_value=8.14-2=6.14

Since the Cumulative_Net_Present_Value is positive the customer is considered to be Profitable.

Example 2:
Let n be number of customers
Considering t=3
Customer_id=3
Avg_Margin=10
Purchase_Frequency=1
Marketing_cost=3rs
Retention_rate=7 ÷ 10=0.7
Retention_rate_for_3_months=(0.7)3-1=0.49
Acquisition_cost=Marketing_cost/Response_rate=15
Present_value_of_customer= [{(10*1-3)*0.49} ÷ (1.03)3]=3.138
Net_present_value = Present_value_of_customer-Acquisition_cost=3.138-15= -11.862
For t=6
Retention_rate=(0.7)6-1=0.168
Present_value_of_customer=1.96
Cumulative_net_present_value=∑Net_present_value=-11.862+1.96= -9.902

Since the Cumulative_Net_Ptresent_Value is negative the customer is considered to be Non-Profitable.

The following code in our system will fetch the data from the table and predict the profitable customers.

“rs1=st1.executeQuery("select * from trans where cust_id='"+str1.get(i)+"'"); “

The above statement fetches the transaction details about that particular customer id, the plan id’s are fetched from the result set and for each plan id the amount and margin details are fetched from the following result set obtained by the statement below.

“rs1=st1.executeQuery("select * from trans where cust_id='"+str1.get(i)+"'");”

Then the Cumulative-Net-Present value is calculated by following code.

```
“avg_margin/=freq;
 avg_amount/=freq;
 profit_per_cust=(avg_margin*freq-3)*0.49f;
 pv=profit_per_cust/(1.03f*1.03f*1.03f);
 npv=pv-6;
 profit_per_cust2=(avg_margin*(freq*2)-6)*0.168f;
 pv2=profit_per_cust2/(n*1.03f)
 npv2=pv2;
cnpv=npv+npv2;”
```

Depending on this cnpv value the customer is classified as Profitable or Non-Profitable.

### 3. *Building Decision Tree learning*

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision tree generates the decision rules, the rules extracted from decision trees can be applied to specific domain.

Decision tree programs construct a decision tree T from a set of training cases. Decision tree learning algorithm functions recursively. First, an attribute must be selected as the root node. In order to create the most efficient (i.e, smallest) tree, the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the

ISSN: 2321-8169
4705 - 4713

same classification. The best split is the one that provides what is termed the entropy.

Entropy is frequently used in machine learning and data mining algorithms for evaluating splits in decision trees. C4.5 algorithm is employed in this system for decision tree construction.

*C4.5 Algorithm*

1. Check for base cases
2. For each attribute *a*
   Find the lowest entropy from splitting on *a*
3. Let *a_best* be the attribute with the lowest gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*

The input to this module will be the training dataset from which decision tree will be built by the above algorithm and rules are extracted from this tree which will be used in the next module to predict the probable churn customers. This particular module implemented in this system would be general and will be the learning model and can be applied to any of the realistic pre-processed dataset and is provided with the various options to control the tree growth by allowing the user to change the depth and minimum leafs and training sets to be used while splitting.

The following are the important classes used in the Building Decision Tree

Decision Tree: This class represents a decision tree. A decision tree is a recursive data structure. It has root node, some data associated with the root node and a list of child, each child is a decision tree itself. If the child list is empty then the root is a leaf, otherwise it is an inner node. It is useful to associate a decision to the inner node not just to the leaves.

Pair: This class represents a generic pair class. For the sake of efficiency data encapsulation is broken and the inner variables are declared to be public. This class is a C++ struct rather than a real class.

Tree Expander: Expands a given node in a JTree.
Parameters for the constructor of this class are
tree: The JTree to expand.
model: The TreeModel for tree.
node: The node within tree to expand.
row: The displayed row in tree that represents node.
depth: The depth to which the tree should be expanded. Zero will just expand node, a negative value will fully expand the tree, and a positive value will recursively expand the tree to that depth relative to node.
Tree Frame: is a class responsible for the last frame. It contains the tire, confusion matrix, score functions, and three buttons.

The following are the methods important to build a decision tree
buildClassDistribution(): It associates and distribute the training points to particular class .
findBestSplit(): It distributes the dataset according to the attribute values and verifies the entropy for each attribute and returns the condition with less entropy
divideTrainingPoints(): It divides the training points according to the condition by expanding nodes using the class tree expander.
JTree class is used to create and expand the tree.

4. *Classifying Churn-able and non churn-able Customers*

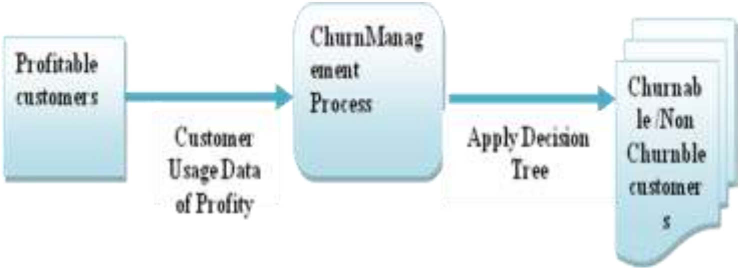


Fig. 5: Classifying Churnable and Non Churnable Customers

The proposed module intends to predict whether the customer is going to churn or not, based on the call log detail (usage details) the decision tree is implemented. The decision tree technique is able to classify specific entities into specific classes based on feature of entities. To divide up a large collection of records into successively smaller sets of records we apply decision rules. And among other models of prediction, Decision Tree is a good choice since it provides the rules that Business users can understand and provides a better accuracy. Churn prediction is important because Competition and customer churn are perhaps the biggest challenges in the telecommunications industry. It is necessary to know which customer may churn prior to losing their client, not after they're already gone.
Our system compares the usage of each customer to the rules extracted from the decision tree and predicts whether a customer is going to churn.

5. *Input Attribute for churn prediction:*

phone number: discrete.

international plan: discrete.

voice mail plan: discrete.

number vmail messages: continuous.

total day minutes: continuous.

total day calls: continuous.

total day charge: continuous.

total eve minutes: continuous.

total eve calls: continuous.

total eve charge: continuous.

total night minutes: continuous.

total night calls: continuous.

total night charge: continuous.

total intl minutes: continuous.

total intl calls: continuous.

ISSN: 2321-8169
4705 - 4713

total intl charge: continuous.

number customer service calls: continuous.

*6. Derived Feature:*

A target variable CHURN, which holds the binary value TRUE or FALSE that, indicates the customer who is likely to churn. Churn prediction module offers suggestions and promotions Once the decision tree is applied on to the customer data, the predicted churn list can be obtained. Because the predicted probable churn customers will be profitable, we can suggest various plans and offers to them.

This module suggests various plans and offers to the customers in specific to their requirement using the usage details; with this the customers could get the messages in which they are interested in. Here customers are put into particular class or group like Local callers, National callers, International callers etc, and suggestions and offer promotions are done on these classes. The offer is sent to the customers through SMS using the STMP protocol through the gateway ipipi.com to which the system sends the username, password, customer number and promotional offer details to be sent

## IV. RESULTS

**1. Data size versus Accuracy**

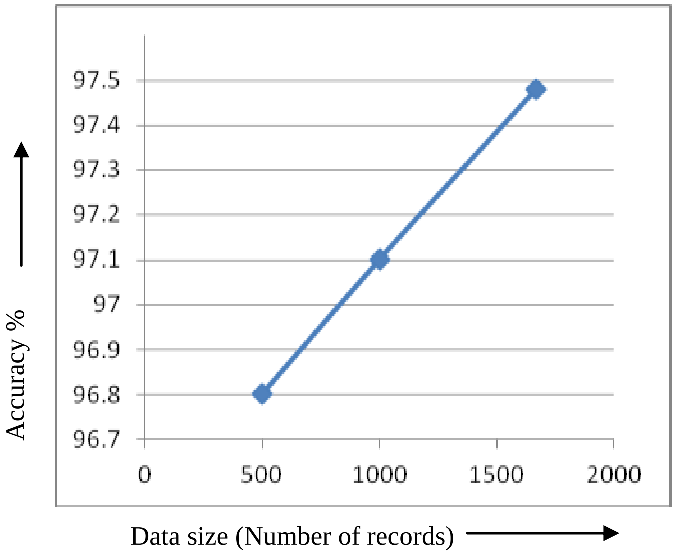


Fig. 6: Data versus Accuracy graph.

System is built using different size of training set which are directly responsible for the accuracy of the underlying system. Hence it can be see that the accuracy is linearly related to the size and quality of the dataset used.

**2. Data size versus Error rate**

The decision tree is built with different data set size and it was observed that the error rate was decreasing as the training size increases as the large dataset contains large variety of customer details which would provide the sufficient data to extract different rules and predict the customer class.

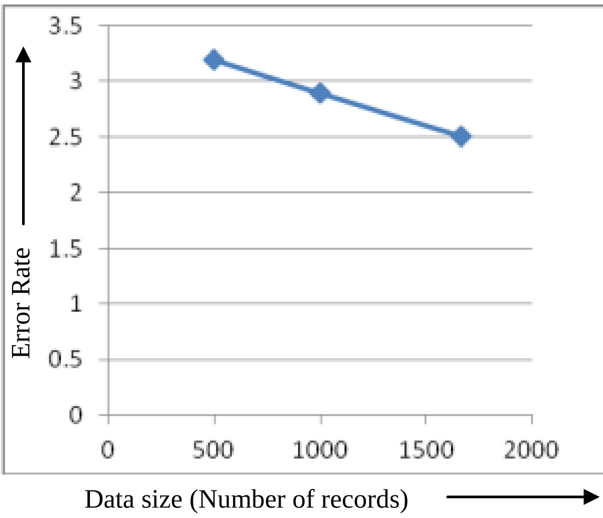


Fig. 7: Data size versus Error rate

## V. CONCLUSION

In this work the proposed system is designed to discover knowledge from the historical data and to be specific it discovers knowledge from the data obtained from telecommunication domain, using the customer details with the help of decision trees and provides a valuable knowledge about the pattern present in the customer details to predict whether the particular customer is probable churner or not and in addition it provides the support to suggest some special offers those customers so that they can retain them in a cost effective manner by predicting whether the customer is profitable to the company and is he worth to that offer so that the company can concentrate and spend more money in retaining these customers rather than wasting money on non-profitable customers. It also show that the decision tree built with larger and quality dataset has the high accuracy than the decision tree built with smaller training set. The system also directs the suggested offers only to the customers who are interested in that particular category determined by their usage.

## REFERENCES

[1] Applying Data Mining to Customer Churn Prediction in an Internet Service Provider, Afaq Alam Khan BGSB University J&K, India, Sanjay Jamwal BGSB

[2] Data Mining in Churn Analysis Model for Telecommunication Industry, Khalida binti Oseman, Sunarti binti Mohd Shukor, Norazrina Abu Haris, Faizi bin Abu Bakar, Operation Support Systems Division TM Research & Development Sdn. Bhd., Vol. 1 No. 19-27, 2010

[3] The Research on Applying Data Mining to Telecom Churn Management Shan Jin,Yun Meng1, Chunfen Fan, Feng Peng, Qingzhang Chen, College of Information, Ningbo Dahongying University, Ningbo, China, College of Computer, Zhejiang University of Technology, China

[4] Predicting Customer Value, Will Potts, Data Miners, Inc., Paper 073-30

[5] Customer churn analysis in telecommunication sector, Umman Tuğba Şimşek Gürsoy,Department of Quantitative Methods, School of Business Administration Istanbul University, Istanbul, Turkey.

[6] Knowledge Discovery on Customer Churn Prediction, LI-SHANG YANG, CHAOCHANG CHIU, Department of Business Administration, St .John's University, Tamsui, Taiwan, ROC.

ISSN: 2321-8169
4705 - 4713

[7] Providing a method for determining the index of customer churn in industry, Keyvan Vahidy Rodpysh, Department of e-commerce, Nooretouba University, Tehran, Iran, Vol.3, No.1, January 2013.

[8] [Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services, Goran Kraljevi´c, Sven Gotovac.

[9] Churn Prediction, Vladislav Lazarov, Technische Universität München, Marius Capota, Technische Universität München.

[10] Marketing Plans for Reducing Customer Churning, Bapuran, Manoj Deng, Meixia.

[11] Improving Customer Churn Models as one of Customer Relationship Management Business Solutions for the Telecommunication Industry, Ecaterina Oana Slăvescu PhD candidate, The Bucharest Academy of Economic Studies, Iulian Panait PhD candidate, The Bucharest Academy of Economic Studies, Volume XII, Issue 1 /2012.

[12] Churn Prediction in Telecommunication Using Data Mining Technology, Rahul J. Jadhav, Bharati Vidyapeeth Deemed University,Pune, Usharani T. Pawar, Shivaji University Kolhapur.

[13] Applications of Data Mining Techniques in Telecom Churn Prediction, V. Umayaparvathi, DDE, Madurai Kamaraj University Madurai, Tamilnadu, India, K. Iyakutti, Madurai Kamaraj University Madurai, Tamilnadu, India, Volume 42– No.20, March 2012.

[14] Model to Predict the Behavior of Customers Churn at the Industry , Keyvan Vahidy Rodpysh Department of e-commerce, Nooretouba University, Tehran, Iran, Volume 49– No.15, July 2012.