International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

# IMPROVING PROFITABILITY IN CRM BASED ON DATA MINING TOOL

**Nisha Gupta[1], Munish Kumar[2], Rajiv Jain[3]**

*[1] Department of Computer Applications, SBSSTC Ferozepur (India),*
*Currently deputed at Department of Computer Applications GZSCCET Bathinda (India)*
*[2] Department of Computer Applications GZSCCET Bathinda (India)*
*[3]Department of Applied Sciences, MIMIT Malout (India)*

## ABSTRACT

*Data mining tools are used to discover hidden knowledge, unknown patterns and new rules from large data sets, which may be useful for a variety of decision-making activities. With the increasing economic globalization and improvements in information technology, huge amounts of financial data are being generated and stored. These can be subjected to data mining tools to discover hidden patterns and obtain predictions for trends in the future and the behavior of the financial markets. In this paper, study work on classification tree induction for large data sets, and analytical and statistical data mining on real data using statistical tools like SPSS are described. We have also discussed broader areas of application, like trading, customer profiling and customer care, where data mining tools can be used in banks and other financial institutions to enhance their business performance through Customer Relationship Management (CRM). Data mining on a real data of banks debit cardholders has been done to support the customer relationship management in banking sector.*

***Keywords: Analytical data mining, Classification trees, Data mining, Statistical data mining, SPSS.***

## I. INTRODUCTION TO DATA MINING

Data mining is the process of extracting useful and previously unknown information out of large complex data collections. Large amounts of data are collected routinely in business, government departments and research organizations. They are typically stored in large data warehouses or databases. For data mining tasks suitable data has to be extracted, cleaned and integrated with other sources. Further data analysis is required to find accurate, useful and understandable information. Data mining aims to discover hidden knowledge, unknown patterns, and new rules from large databases, that are potentially useful and ultimate understandable for making crucial decisions.

Database today can range in size into the terabytes-more than 1,000,000,000,000 bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how the meaningful conclusion about forest can be perfectly drawn? The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. Over the past several years

ISSN (O) 2454-9665
ISSN (P) 2454-0663

there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year. At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.

Data mining, the extraction of hidden predictive information from large databases [1], is powerful new technology with great potential to help companies focus on the most important information in their data warehouses and databases. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support system. Data mining tools can answer business question that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

## II. DATA MINING PROCESS

Most of the companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server on parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, " which clients are most likely to respond to my next promotional mailing, and why?"

The data mining concept [3] can be explained with the help of a figure 1. The first step in data mining tool is collect, clean and summarizes the historical data. In second step, build a model based on the application. After that learn and test the model with help of the historical data. Finally apply this model on new data.

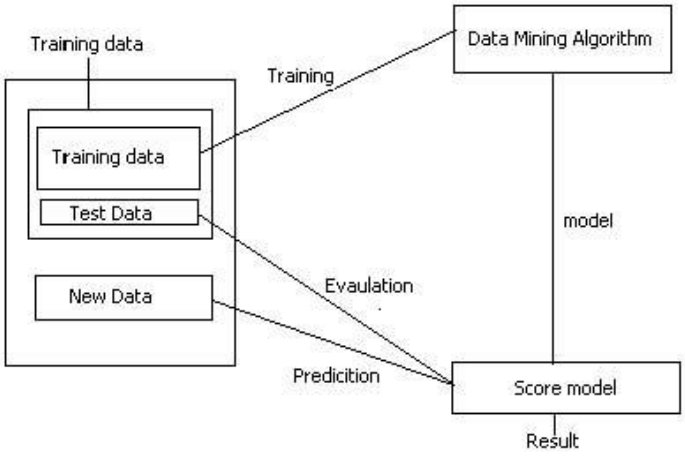


Figure 1: Data Mining Process

ISSN (O) 2454-9665
ISSN (P) 2454-0663

As shown in figure 2.1, historical data is used to train the data mining algorithms which is later evaluated on some part of the same data. Historical data is divided into two pats: training data and test data. Training data is used to train the model and test data is used to evaluate the model. Now this learned model is used to predict the new data.

**2.1 Detailed Data Mining Process**

A typical data mining process includes methodology for extracting and preparing data as well as making decision about actions to b e taken once data mining has place. When a particular application involves the analysis of large volume of data stored in several locations, data extraction and preparation become the most time consuming part of the discovery process. A detailed seven step data mining process is shown in figure2. A brief description of each step follows:

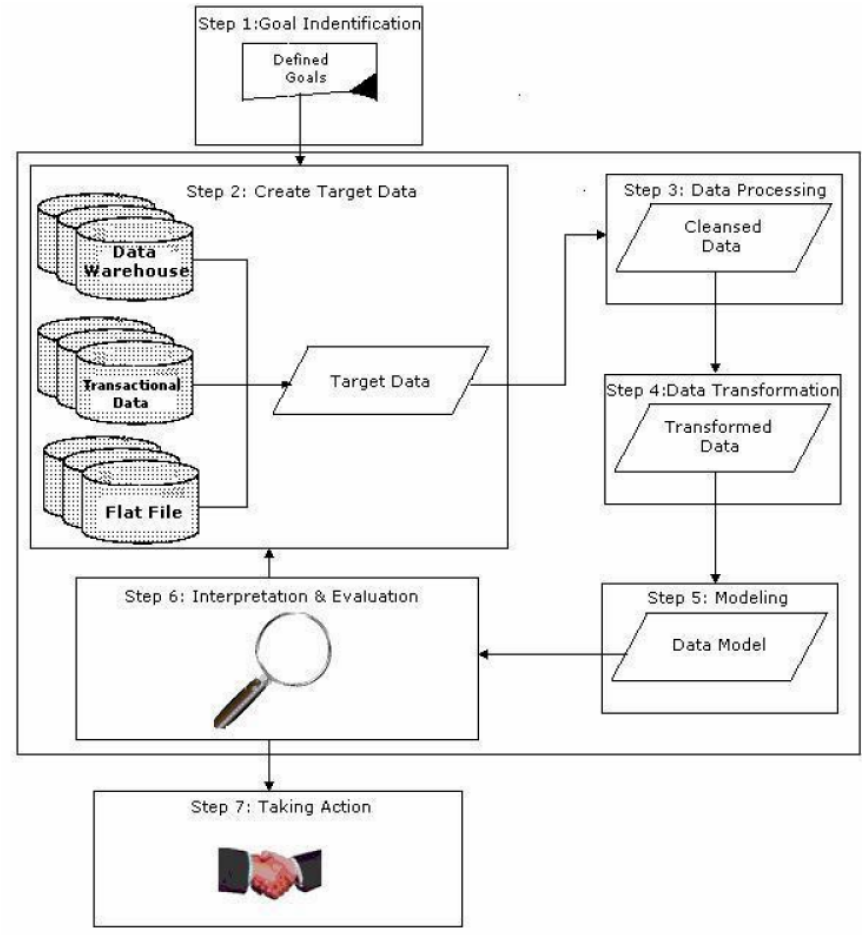


Figure2*:* Detailed Data Mining Process

*Step 1: Goal Identification:* The focus of this step is on understanding the domain being considered for knowledge discovery. We write a clear statement about what is to be accomplished.

*Step 2: Creating the target data set:* With the help of one or more human experts, we choose an initial set of data to be analyzed.

*Step 3: Data preprocessing:* We use available resources to deal with noisy data. We decide what to do about missing data values and how to account for time-sequence information.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

*Step 4: Data Transformation:* Attributes and instances are added and/ or eliminated from the target data. We decide on methods to normalize, convert, and smooth data.

*Step 5: Modeling:* A best model for representing the data is created by applying one or more data mining algorithms.

*Step 6: Interpretation and Evaluation:* We examine the output from step 5 to determine if what has been discovered is both useful and interesting. Decisions are made about whether to repeat previous steps using new attributes and / or instances.

*Step 7: Taking Action:* If the discovered knowledge is deemed useful, the knowledge is incorporated and applied directly to appropriate problems.

### 2.2 Data Mining Techniques

The data mining concepts involve learning from the training data. The learning algorithms can be very broadly classified as:

- Supervised
- Unsupervised

In unsupervised learning, there is no tutor which defines the class of priori. The system must itself find some way of clustering the objects into the classes, and find appropriate description for these classes.

In supervised learning, a tutor must help the system in construction of the model, by defining classes and providing positive and negative examples of objects belonging to these classes. Thus supervised learning is a function fitting algorithm, while unsupervised learning is cluster-finding algorithm [2].

Some of most frequently used data mining techniques are as below:

- Clustering
- Classification
- Link Analysis

## III. DATA MINING AND CUSTOMER RELATIONSHIP MANAGEMENT

Customer relationship management (CRM) helps companies improve the profitability of their interactions with customers, while at the same time; makes the interactions appear friendlier through individualization. To succeed with CRM, companies need to match products and campaigns to prospects and customers – in other words, to intelligently manage the customer life cycle. However, the sheer volume of customer information and increasingly complex interactions with customers, have propelled data mining to the forefront of making customer relationships profitable. Data mining is a process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that are used to understand what your customers want and predict what they will do. Data mining can help to select the right prospects on whom to focus, offer the right additional products to your existing customers and identify good customers who may be about to leave. This results in improved revenue because of a greatly improved ability to respond to each individual contact in the best way and reduced costs due to properly allocated resources. CRM applications that use data mining are called analytic CRM.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

### 3.1 Data Mining

The first and simplest analytical step in data mining is to describe the data. For example, data's statistical attributes (such as means and standard deviations), can be summarized and visually reviewed using charts and graphs and look at the distribution of field values in data. But data description alone cannot provide an action plan. Build a predictive model based on patterns determined from known results and then test that model on results outside the original sample. A good model should never be confused with reality (you know a road map isn't a perfect representation of the actual road), but it can be a useful guide to understanding your business.

Data mining can be used for both classification and regression problems. In classification problems prediction of what category something falls into – for example, whether or not a person is a good credit risk or which of several offers someone is most likely to accept. In regression problems, you're predicting a number, such as the probability that a person will respond to an offer.

In CRM, data mining is frequently used to assign a score to a particular customer or prospect indicating the likelihood that the individual behaves the way you want. For example, a score could measure the propensity to respond to a particular offer or to switch to a competitor's product. It is also frequently used to identify a set of characteristics (called a profile) that segments customers into groups with similar behaviors, such as buying a particular product. A special type of classification can recommend items based on similar interests held by groups of customers. This is called collaborative filtering.

### 3.2 Defining CRM

Customer relationship management in its broadest sense simply means managing all customer interactions. In practice, this requires using information about the customers and prospects to more effectively interact with your customers in all stages of relationship with them. We refer to these stages as the customer life cycle. The customer life cycle has three stages:

- Acquiring customers
- Increasing the value of customers
- Retaining good customers

Data mining can improve your portability in each of these stages when you integrate it with operational CRM systems or implement it as independent applications.

### 3.3 Applying Data Mining to CRM

In order to build good models for your CRM system, there are a number of steps to follow. The Two Crows data mining process model described below is similar to other process models such as the CRISP-DM model, differing mostly in the emphasis it places on the different steps.

Keep in mind that while the steps appear in a list, the data mining process is not linear – you will inevitably need to loop back to previous steps. For example, what you learn in the "explore data" step (step 3) may require you to add new data to the data mining database. The initial models you build may provide insights that lead you to create new variables.

The basic steps of data mining for effective CRM are:

1 Define business problem

ISSN (O) 2454-9665
ISSN (P) 2454-0663

2 Build marketing database

3 Explore data

4 Prepare data for modeling

5 Build model

6 Evaluate model

7 Deploy model and results

Customer relationship management is essential to compete effectively in today's marketplace. The more effectively we can use information about our customers to meet their needs, the more profitable you will be. Operational CRM needs analytical CRM with predictive data mining models at its core. The route to a successful business requires that we understand our customers and their requirements, and data mining is the essential guide.

## IV. PROBLEM STATEMENT

Classification trees are most widely used for classification mining. When data set grows larger the classification tree becomes thicker. Even after pruning which costs more time for the large data sets, trees are still too large size. With the large amount and wide diversities of the data in large data sets, many more data partitions are generated in recursive process of tree induction. Some leaf notes can achieve the purity even when no more attributes are available to branch on [4][5]. In this paper, following is the work pattern:

1. Classification tree induction

In this paper, work relating to classification tree induction for large data sets of debit cards holding of various bank customers has been presented. There is description of the construction of classification trees with majority of leaf nodes, which can handle large data sets. Due to majority leaf node tree size will remain concise.

2. Analytical Data Mining

Analytical data mining using data mining tools like SPSS has been performed. Real data of debit card holders of various banks has been collected through a simple questionnaire and has been worked out. Results and analysis of analytical data mining is shown in form trees, tables.

## V. ANALYTICAL DATA MINING OF DEBIT CARD HOLDERS FOR CRM USING SPSS

In this problem, there is use of Data Mining tool called SPSS [6] for analytical and statistical data mining. The general idea behind use of analytical and statistical data mining is to discover meaningful patterns, relationship, and information of strategic importance that lie hidden within the database. In this paper, the analytical data mining is done on the data of debit card holders for the improvement of CRM and debit card business.

There is mining of the real data of bank customers using SPSS 16.0 Windows Evaluation Version collected through the questionnaire. In analytical data mining, a model is created and evaluated with the help of real data. This model is used and evaluated with the help of collected data of debit card holders of banks of local area and this model is used to classify the data of customer's views and their withdrawal through debit cards. A surveywas conducted in the local area of Malout dist. Muktsar Punjab. A questionnaire was filled by the debit cardholders of the banks in local area of Malout. The data was collected for amount of withdrawal in a year by customer, category, segment of profession to he/she belongs, age, education, annual income, whether he/she

ISSN (O) 2454-9665
ISSN (P) 2454-0663

needs more facilities, need of more seminars & workshops, security of transaction through debit card, distance from ATM machine & his/her satisfaction from bank.

### 5.1 Model Summary Table

The model summary table provides broad information about the specifications used to build the model and the resulting model. Model summary table is shown in Table 1. The specifications section provides information on the settings used to generate the tree model, including the variables used in analysis. The results section displays information on the number of total and terminal nodes, depth of the tree (number of levels below the root node), and independent variables included in the final model. The variables for need of more facilities, need of seminars & workshops and satisfaction with the service did not make a significant contribution to the model, so they were automatically dropped from the final model.

Table 1 Model Summary Table

| | | | |
|---|---|---|---|
| Specifications | Growing Method | CHAID | |
| | | Dependent Variable | Annual Transaction |
| | | Independent Variables | Age Group, Education level, Category, Segment, Annual Income, Distance of ATM Machine, Need More Facilities, Debit Card security, Need of Seminars and Workshops, Satisfied with Service |
| | | Validation | None |
| | | Maximum Tree Depth | 3 |
| | | Minimum Cases in Parent Node | 50 |
| | | Minimum Cases in Child Node | 25 |
| Results | Independent Variables Included | Annual Income, Satisfied with Service, Category, Distance of ATM Machine, Segment, Need of Seminars and Workshops | |
| | | Number of Nodes | 18 |
| | | Number of Terminal Nodes | 11 |
| | | Depth | 3 |

### 5.2 Classification Tree Diagram

The tree diagram is a graphic representation of the tree model. Figure3 shows the tree diagram of our tree

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

model. This tree diagram shows a lot of information about classification. Using CHAID method, Annual Income is best predictor of Annual Transactions. For the income between 1.5 lacs to 3 lacs category, Distance of ATM machine is the next predictor of the Annual Transactions. For income between 3.0 to 5 lacs, satisfied with the service attribute is the next predictor for Annual Transactions. For the income less than 1.5 lacs, need of seminars & Workshops is the next predictor of annual Transactions. Debit card holders whose income is in between 1.5 lacs to 3 lacs and whose distance of ATM machine is in between 1-3 kms, the next predictor is the segment of profession. In this the debit card holders whose Annual Income is between 3 lacs to 5 lacs and who are satisfied with the services, the next predictor is category.

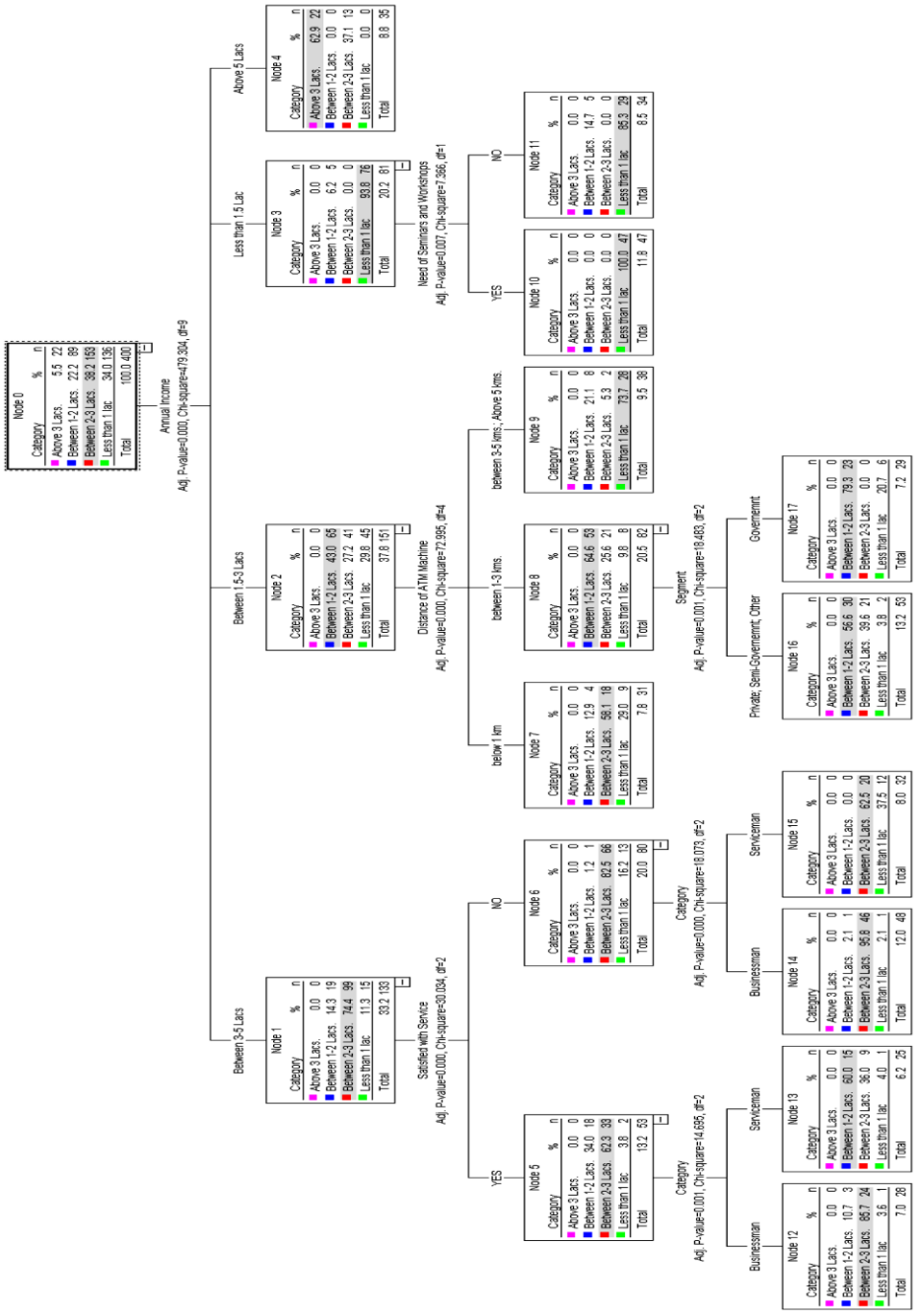

Figure 3: Classification Tree of Debit Card Customers

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

Table 2: Classification Tree Table

| Node | Above 3 Lacs. | | Between 1-2 Lacs | | Between 2-3 Lacs | | Less than 1 lac | | Total | | Predicted Category | Parent Node | Primary Independent Variable | | | | Split Values |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Percent | N | Percent | N | Percent | N | Percent | N | Percent | | | Variable | Sig.[a] | Chi-Square | df | |
| 0 | 22 | 5.5% | 89 | 22.2% | 153 | 38.2% | 136 | 34.0% | 400 | 100 % | Between 2-3 Lacs. | | | | | | |
| 1 | 0 | .0% | 19 | 14.3% | 99 | 74.4% | 15 | 11.3% | 133 | 33.2% | Between 2-3 Lacs. | 0 | Annual Income | .000 | 479.304 | 9 | Between 3-5 Lacs |
| 2 | 0 | .0% | 65 | 43.0% | 41 | 27.2% | 45 | 29.8% | 151 | 37.8% | Between 1-2 Lacs. | 0 | Annual Income | .000 | 479.304 | 9 | Between 1.5-3 Lacs |
| 3 | 0 | .0% | 5 | 6.2% | 0 | .0% | 76 | 93.8% | 81 | 20.2% | Less than 1 lac | 0 | Annual Income | .000 | 479.304 | 9 | Less than 1.5 Lac |
| 4 | 22 | 62.9% | 0 | .0% | 13 | 37.1% | 0 | .0% | 35 | 8.8% | Above 3 Lacs. | 0 | Annual Income | .000 | 479.304 | 9 | Above 5 Lacs |
| 5 | 0 | .0% | 18 | 34.0% | 33 | 62.3% | 2 | 3.8% | 53 | 13.2% | Between 2-3 Lacs. | 1 | Satisfied withService | .000 | 30.034 | 2 | YES |
| 6 | 0 | .0% | 1 | 1.2% | 66 | 82.5% | 13 | 16.2% | 80 | 20.0% | Between 2-3 Lacs. | 1 | Satisfied with Service | .000 | 30.034 | 2 | NO |
| 7 | 0 | .0% | 4 | 12.9% | 18 | 58.1% | 9 | 29.0% | 31 | 7.8% | Between 2-3 Lacs. | 2 | Distance of ATM | .000 | 72.995 | 4 | below 1 km |
| 8 | 0 | .0% | 53 | 64.6% | 21 | 25.6% | 8 | 9.8% | 82 | 20.5% | Between 1-2 Lacs. | 2 | Distance of ATM | .000 | 72.995 | 4 | between 1-3 kms |
| 9 | 0 | .0% | 8 | 21.1% | 2 | 5.3% | 28 | 73.7% | 38 | 9.5% | Less than 1 lac | 2 | Distance of ATM | .000 | 72.995 | 4 | between 3-5 kms; Above5 |
| 10 | 0 | .0% | 0 | .0% | 0 | .0% | 47 | 100% | 47 | 11.8% | Less than 1 lac | 3 | Need of Seminars | .007 | 7.366 | 1 | YES |
| 11 | 0 | .0% | 5 | 14.7% | 0 | .0% | 29 | 85.3% | 34 | 8.5% | Less than 1 lac | 3 | Need of Seminars | .007 | 7.366 | 1 | NO |
| 12 | 0 | .0% | 3 | 10.7% | 24 | 85.7% | 1 | 3.6% | 28 | 7.0% | Between 2-3 Lacs. | 5 | Category | .001 | 14.695 | 2 | Businessman |
| 13 | 0 | .0% | 15 | 60.0% | 9 | 36.0% | 1 | 4.0% | 25 | 6.2% | Between 1-2 Lacs. | 5 | Category | .001 | 14.695 | 2 | Serviceman |
| 14 | 0 | .0% | 1 | 2.1% | 46 | 95.8% | 1 | 2.1% | 48 | 12.0% | Between 2-3 Lacs. | 6 | Category | .000 | 18.073 | 2 | Businessman |
| 15 | 0 | .0% | 0 | .0% | 20 | 62.5% | 12 | 37.5% | 32 | 8.0% | Between 2-3 Lacs. | 6 | Category | .000 | 18.073 | 2 | Serviceman |
| 16 | 0 | .0% | 30 | 56.6% | 21 | 39.6% | 2 | 3.8% | 53 | 13.2% | Between 1-2 Lacs. | 8 | Segment | .001 | 18.483 | 2 | Private; Semi-Government; Government |
| 17 | 0 | .0% | 23 | 79.3% | 0 | .0% | 6 | 20.7% | 29 | 7.2% | Between 1-2 Lacs. | 8 | Segment | .001 | 18.483 | 2 | |

Growing Method: CHAID

Dependent Variable: Annual Transaction

a. Bonferroni adjusted

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

### 5.3 Tree Table

The tree table, as the name suggests, provides most of the essential tree diagram information in the form of table. Tree table is shown in Table 2. For each node, first the table displays the number and percentage of cases in each category of the dependent variable. Secondly, predicted category for the dependent variable. In this data set, the predicted category is the annual Income category with more than 38.2% of case in that node. Note that node 5– the Annual Income Above 5 lacs – is not the parent node of any node. Since it is the terminal node, it has no child node. Next in table column is the independent variable which is used to split the node and at last the split values of independent variable for that node.

### 5.4 Risk Estimate and Classification Table

The risk and classification tables provide a quick evaluation of how well the model works. The risk and classification table of our model is shown in table 3 and table 4 respectively. In our model, the risk estimate of 0.245 indicates that the category predicted by the model is wrong for 24.5 % of the cases. So the risk of misclassifying a customer is approximately 25%.

Table 3: Risk Table of the model

| Estimate | Std. Error |
|---|---|
| .245 | .022 |

Growing Method: CHAID

Dependent Variable: Annual Transaction

Table 4: Classification Table of Model

| Observed | Predicted | | | | |
|---|---|---|---|---|---|
| | Above 3 Lacs. | Between 1-2 Lacs. | Between 2-3 Lacs. | Less than 1 lac | Percent Correct |
| Above 3 Lacs. | 22 | 0 | 0 | 0 | 100.0% |
| Between 1-2 Lacs. | 0 | 68 | 8 | 13 | 76.4% |
| Between 2-3 Lacs. | 13 | 30 | 108 | 2 | 70.6% |
| Less than 1 lac | 0 | 9 | 23 | 104 | 76.5% |
| Overall Percentage | 8.8% | 26.8% | 34.8% | 29.8% | 75.5% |

The results in the classification table are consistent with the risk estimate. The table shows that our model classifies approximately 75.5 % of the customers correctly. The classification table does, however reveal one potential with this model: for those debit card holders with annual transaction between 2 to 3 lacs, it predicts customers having annual transaction between 2 to 3 lacs for only 70.6 % of them, which means that 29.4% of customers having annual transaction between 2 to 3 lacs are inaccurately classified with the other remaining customers.

## VI. CONCLUSIONS

In this paper, the work is based on classification tree induction for large data sets. And analytical data mining using SPSS 16.0 Windows Evaluation Version is performed. The work emphasizes on the usefulness of data

ISSN (O) 2454-9665
ISSN (P) 2454-0663

mining for customer relationship management in debit card business. For this data is collected, and mined using classification trees and following useful results are found.

As the annual income is the best predictor of Annual Transactions, only 5.5% of the total debit card holders annually transact more than 3 lacs. In more than 5 lacs Annual Income category, only 62.9% debit card holders annually transact more than 3 lacs but in annual income category 3 lacs to 5 lacs, this ratio suddenly decreases to 0%. As it can be interpreted from the classification tree that more than 60% of the debit card holders of this category (annual income 1-3 lacs) are not satisfied with the services of bank. And further out of these debit card holders 60% belongs to businessman category. So according to this classification tree, the bank should take the immediate correct actions to satisfy the businessman category having annual income between 3 to 5 lacs and who are not satisfied with the services of the bank. This will result to better the CRM with enhance their profitability in debit card business.

In the Annual Income less than 1.5 lacs category, only 6.2% of this category annually transact between 1 lac to 2 lacs. In this category, about 60% debit card holders feels the need of Seminars and Workshops to be conducted by the banks regarding the use of debit card. So banks should conduct the seminars and workshops to improve the customer-bank relations to maximize the profit from debit card business.

In the Annual Income between 1.5 to 3 lacs category the next predictor of Annual Transaction is Distance of ATM Machines. For the distance of ATM machine below 1 km, Annual Transactions between 2 to 3 lacs are 58% of its category but for the distance of ATM machine between 1 to 3 kms. category, Annual Transactions between 2 to 3 lacs are 25.6%. As the distance of ATM machine increases to 3 to 5 kms and above 5 kms category, the Annual Transactions between 2 to 3 lacs are only 5.3% of its category. So the banks should increase the number of ATM machines to improve the Annual Transactions and make the debit card business more profitable. Further, many results can be interpreted from the classification tree and statistical mining of the data.

**REFERENCES**

[1] L. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman, CA, 1996.

[2] M.D. Alder, “An Introduction to patterns recognition: Statistical, Neural Net and Synthetic Methods of getting robots to see and hear”, 1997. http://ciips.ee.uwa.edu.au/mike

[3] Jiawei Han, and Micheline Kamber. “Data Mining: Concepts and Techniques.” New York: Morgan Kaufmann Publishers, 2001.

[4] Surachai Wiwattanacharoenchai and Anongnart Srivihok. “Data Mining of Electronic Banking in Thailand: Usage Behavior Analysis by Using K-Means Algorithm.” Regional Conference on Digital GMS. 2003 Feb 26-28; 211-215.

[5] Rajanish Dass., “Data Mining in Banking and Finance: A Note for Bankers.”, Indian Institute of Management Ahmedabad.

[6] SPSS 16.0 Data Mining Tool, www.spss.com