Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining using Decision Tree for Sales Force Optimization

Mr. Tanupriya Choudhury* *Assistant Professor CSE Dept, Lingaya's University(India)*
Mrs. Vasudha Vashisht *Assistant Professor CSE Dept, Lingaya's University(India)*
Prof.Vivek Kumar *Principal,DCTM-Palwal, Haryana,India*
Mr. Himanshu Srivastava *Undergraduate degree CSE G.B.T.U.(India)*

Abstract— In this paper, the application of Decision Tree to optimize the sales, by identification of most prospective leads, is studied. The probability of the conversion of the lead into an account or the customer is calculated on the basis of previous data of customers such as their demographic attributes, age, gender etc. that can help the end–user in deciding which leads to follow up.

Keywords — Data Warehouse, Data Mining, Customer Relationship Management, Decision tree.

I. INTRODUCTION

Every organisation implements a CRM system internally that may be computer based or manual. However, with the increase in the market competition and the advent of big players, the customer is spoilt with wide range of options to choose from. To tackle this problem, CRM systems are often implemented with the Decision support system. Such systems use Data Warehouse at the back end so that the bulk of the data can be analysed. Any CRM system is fed with huge amount of data and most of the implementation uses this raw data to build statistics about the customer preferences. However, this huge amount of data has implicit knowledge in it which can be dug out to give analysis of the customer preferences that can be used for planning future marketing strategies. Every company plans for a campaign to market their products and services. Massive amount of money is spent on these campaigns. Leads or the prospective customers, generated by these campaigns come from different background and can be categorised into different groups according to their spending power. Information about these leads is stored in the CRM database regularly. It could be useful for the sales department if the customer data is classified based on some attributes. The decision tree approach is most useful in such classification problems. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple of the database and thus, results in a classification of that tuple.

II. PROBLEM DEFINITION

The marketing campaigns target to a wide population and the generated leads differ largely in their attributes. Given such a huge amount of leads, the task of following up each lead is quite herculean. Randomly choosing the leads to follow up will not be efficient enough. A possible solution is to classify the leads on the basis of the probability of a lead to convert into an account. If such a probability can be found out, the sales team can filter the mammoth database and then can follow up the most prospective lead. The sales of the organisation would increase if the potential customers can be identified from the given leads. Given 100 leads for a particular campaign, if despite of contacting each lead one by one, only that lead having the positive prediction of conversion is contacted, the use of time and resource will be efficient.

III. CUSTOMER RELATIONSHIP MANAGEMENT

Customer Relationship Management is a business approach that integrates people, process and technology to enhance the relationship with the existing customers and to widen the scope for future customers. CRM is the need of the market where businesses have gone ubiquitous and the competition is cut-throat. Companies are spending billions of dollars for improving their product and their services. 25% of the product price is constituted from the marketing expenditure. With such high investment at stake, it is important that the marketing strategy is efficient and results in the growth of the company.

A. Leads

Leads are the prospects or tentative customers of the organisation. Anybody who is interested to take a product or services of a company can be considered as a lead. The main emphasis in any CRM system is given on its ability to convert the maximum leads as the customers of the organisation. For generating leads, the organisation plans and executes various campaigns ranging from mass e-mails (online) to seminars (offline) and conferences. The budget of these campaigns is the amount on which the return of investment of the campaign is calculated.

B. CRM Data Warehouse

Today, companies have made a global presence. The culture of multinationals has flourished during the decade. It is important that each organisation maintains a central data-warehouse which is integrated with all the local source systems so that data can be transferred to the data-warehouse at regular intervals. Any data-warehouse stores data in the form of

dimensions and facts .This model is also called as multidimensional modelling which largely differs from the traditional relational modelling in OLTP based systems. The need of the use of data warehouse arises from the fact that the centralisation of the data is needed for data mining.

1. Measures

At the centre of the dimensional model are the numeric measures that we are interested in understanding, such as enrolments or sales revenue. Related measures are collected into fact tables that contain columns for each of the numeric measures. Every time something measurable happens, such as a sales transaction, an inventory balance or when an event occurs, a new record is added to the fact table with these numeric values.

2. Dimensions

There are usually many different ways that people can look at these measures. For example, they could look at totals for a product category or show the totals for a particular set of stores. These different ways of looking at the information are called dimensions, where a dimension is a particular area of interest such as Product, Customer, or Time. Every dimension table has a number of columns with descriptive text, such as product category, colour, and size for a Product dimension. These descriptive columns are known as attributes.

IV. DECISION TREE

A. Principle

Decision Trees is the most popular data mining technique for classification problems. The principal idea of decision tree is to split your data recursively into sub-sets so that each subset contains more or less homogenous states of your target variable. At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When the recursive process is completed, a Decision Tree is formed.

Given a database $D=\{t_1, \ldots, t_n\}$ where $t_i=\{t_{i1},\ldots t_{ih}\}$ and the database schema contains the following attributes $\{A_1,A_2,\ldots,A_h\}$. Also given is a set of classes $C=\{C_1,\ldots,C_2\}$. A Decision Tree (DT) or classification tree is a tree associated with D that has the following properties.

1. Each internal node is labelled with an attribute A
2. Each arc is labelled with a predicate that can be applied to the attribute associated with the parent.
3. Each leaf is labelled with class C_j.

B. Feature Selection

Data mining algorithms are very sensitive to the number of attributes you include. Too many attributes requires extensive CPU and memory resources for processing. Also, not all the attributes are equally important in terms of the prediction accuracy. Feature selection is a process that selects a subset of attributes so that the processing time can be substantially reduced but with no or limited sacrifices on the model accuracy.

Statistical function such as entropy is used to calculate the impact of each input attribute related to the predictable attribute, and then select the most significant attributes for the model .Entropy is used to measure the amount of uncertainty or surprise or randomness in a set of data. Certainly when all data in a set belongs to a single class, there is no uncertainty. In this case, the entropy is zero. The objective of the decision tree classification algorithm is to iteratively partition the given data set into the subsets where all the finals subsets belong to a given same class.

C. Basic Concepts of tree growth

The basic idea of a decision tree is straightforward. The entire algorithm that builds the decision tree can be explained with a example from the CRM domain. Consider a huge database of leads stored in the data warehouse of an organisation. The table contains around 3000 rows. With information about their gender, age, marital status, yearly income, the predictable attribute here is the IsConverted, a binary column indicating if the customer is planning to buy the product or in the given scenario, planning to pursue a given course. The first step is to build the correlation count table. Each column in the correlation count table is an attribute/value pair of input attributes. Each row is a state value of predictable attribute. The cells in the table are the counts of correlations of input attribute values and predictable states. From the table, you can see that there are 400 Male students, 300 of them are associated with IsConverted = Yes, while 100 of them associated with IsConverted = No.

TABLE I

INITIAL CORRELATION TABLE

		Gender		**Age**		**Income**		**Marital Status**	
		Male	**Female**	**Below 25**	**Above 25**	**Low**	**High**	**Single**	**Married**
Is Converted	**Yes**	500	500	700	300	400	600	500	600
	No	1100	900	400	1600	700	1600	400	600

In our example, there are only two predictable states: Yes and No, $n = 2$.

Using the preceding formula, we can calculate the entropies of the splits on the four input attributes:

1) Split on Gender

Entropy (700, 400) + Entropy (300, 1600) = 0.946 +0.629 = 1.571

2) Split on Age
Entropy (300, 100) + Entropy (500, 1000) + Entropy (200, 900) = 0.814 + 0.918 + 0.684 = 2.416
3) Split on Yearly Income
Entropy (400,400) + Entropy (600, 1600) = 1.0 + 0.845 = 1.845
4)Split on Marital Status
Entropy (500, 1100) + Entropy (500, 900) =0.896 + 0.941 = 1.837

Based on these calculations, we find the subsets after the splits on Gender have the lowest entropy. Thus the most significant attribute to split at the root level is Gender. Once the data is split into two subsets, the algorithm repeats the same process on each leaf node to grow the tree.

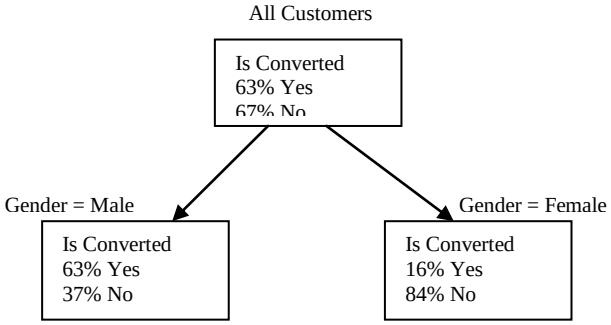

Fig 1. Decision Tree After First Split

V. DATA COLLECTION

The data used in training the mining model is collected from a CRM data warehouse. The CRM source system receives thousand of rows o leads data throughout the day. This data is integrated with a centralised data warehouse using integration packages that are scheduled to run at regular intervals. The data mining model is created by providing the input attributes and the predictable attributes. Once the mining model is created, the model is trained with the training data which is the data of previous customers. After the model is trained, it can be used for classification of any lead that enters into the system.

VI. RESULTS AND DISCUSSIONS

The Decision tree algorithm used here to classify the leads with respect to their probabilities can be efficiently implemented in any data warehouse. One such implementation is being done using Microsoft Business Intelligence Development Studio. The Decision tree data mining model can be created by providing the information such as input parameters, predictable parameters and the case column. When the decision tree data mining model was trained with a leads database having a sufficiently large number of rows, the data mining model could predict the probability of the leads which are not still converted. As can be seen in Table II , when the new leads are fed into the mining model ,the model calculates the predicted conversion of each of the lead. The second column only shows that the leads are still not converted while the third column shows the predicted conversion of those leads as 'Yes' or 'No'.

TABLE III
PREDICTED CONVERSION OF LEADS

CustomerKey	IsConverted	PredictedConversion
2	N	N
14	N	Y
17	N	Y
29	N	N

VII. CONCLUSIONS

The application of decision tree implemented in this paper can be effectively used in a CRM system of any domain. The analysis offered by this model can prove vital to any sales force department. By predicting the conversion of a lead into a customer, the sales department can effectively utilize its time and resources on the right person and hence the net output will be optimized.

REFERENCES

[1] John C. Hancock and Roger Toren, Practical Business Intelligence With SQL Server 2005, 2005
[2] Barton J. Goldenberg, CRM in Real Time: Empowering Customer Relationship, 2008
[3] Jamie MacLennan ,ZhaoHui Tang, Data Mining with SQL Server 2005, 2005
[4] David Tanwar, Monash University, Australia, Research and Trends in Data Mining Technologies and Applications, 2007
[5] Vincent Rainardi, Building a Data Warehouse for Customer Relationship Management With Examples in SQL Server, 2009
[6] Thomas C. Hammergren and Alan R. Simon, Data Warehousing For Dummies®, 2nd Edition, 2009
[7] Thomas H. Davenport and Jeanne G. Harris, The Architecture of Business Intelligence, March 2007.

