

## An Efficient Classification Algorithm for Real Estate domain

Geetali Banerji\*, Kanak Saxena\*\*

*\*(IT Department, Institute of Information Technology & Management, New Delhi, India)*

*\*\* (Computer Applications Department, Samrat Ashok Technological Institute, Vidisha, India)*

**Abstract:** Classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier. In classification rule mining there is one and only one predetermined target. In this paper, we proposed an algorithm, which performs preprocessing and cleaning prior to traditional classification. Experimental results show that the classifier built this way is, in general, more accurate than that produced by the state-of-the-art classification system. In addition, this helps to solve a number of problems that exist in the current classification systems. This algorithm is especially fit to applications where it may assist domain experts in their decisions. We have taken Real estate field, which is a good example, where such applications may appear. In our dataset, there is a considerable amount of information associated with the customer (e.g. Income, personal, educational, demographic and professional details). We have applied statistical analysis and Association rules on real estate actual data and found that if we consider only customer demographic and professional details then relationship among various attributes are not that strong and it only predicts the customers who are in higher income group are willing to invest. After removing demographic details like caste, status, religion etc. we found that the relationship among remaining attributes become much stronger and it shows that customer in middle income group, house owners are also willing to invest. Further not only a person's professional details helps to identify a potential customer, in Indian context, his other responsibilities, expenditure heads also plays a crucial role. It is further proven that by taking into account the bank transaction details and buying interest of a customer, there is a drastic improvement in the results i.e. not only those customers, who are in high income group and potentially strong enough to invest but because of their other social responsibilities they are not able to do so and at the same time those who are following in middle or lower income group or having rented house with other investment are also likely to invest.

**Keywords:** Association Rules, Classification, DT (Decision Table), Statistical Measures, RepTree

### I. INTRODUCTION

Data mining is a process, which involves the application of specific algorithms for extracting patterns (test models) from data. New knowledge may be obtained in the process while eliminating one of the largest costs, viz., data collection [2]. Real Estate data, for example, often exists in vast quantities in an unstructured format. Our proposed classification algorithm which applies statistical analysis and association rules prior to traditional classification is proved better than the traditional ones because it cleans the data and removes the noise. Experimental results show that the classifier built this way is, in general, more accurate than that produced by the state-of-the-art classification system. The traditional classifiers are faster but in many cases accuracy is not so high. Moreover many of the rules found by proposed classification algorithm cannot be discovered by traditional classification algorithm. In earlier paper [9], we have already proved that in case of statistical analysis regression is used to determine the relationship among various attributes. We have tested various regression algorithms such as linear Regression, Least Median Square Isotonic and Pace Regression. Pace regression outperforms among various regression techniques because in this method it generates a pattern which assigns weights prior to processing and it is suitable with respect to Indian context because in India, caste, religion, social status plays a prominent roles In general context, multilayer perceptron is best among various machine learning algorithms because it applies weights on different hidden layer on the basis of different input weightage during processing which is suitable for other countries where the status, caste, religion does not have any role in purchasing behavior of a customer [11].

The rest of the paper is organized as follows. The concept of Classifiers is being discussed in section II. In section III, Statistical Analysis and Association Rules are discussed, the proposed algorithm has been introduced in Section IV, Section V discusses about Empirical Evaluation and Section VI concludes the paper.

### II. CLASSIFICATION

In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. It is a two step process, in the first step; a classifier is built describing a predetermined set of data classes or concept. This is the learning step, (or training phase), where a classification algorithm builds the classifier by learning from a training set made up of database tuples and their associated class labels. In the second step, the test model is used for classification [3, 4].

### **Preparing the data for classification**

The following preprocessing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification process [3].

**Data Cleaning:** This refers to reduce noise or handling missing values. This step help reduce confusion during learning.

**Relevance Analysis:** This refers to remove irrelevant or redundant attributes. Correlation analysis can be used to identify whether two given attributes are statistically related. For Example a strong correlation between attribute A1 and A2 would suggest that one of the two could be reduced for further analysis.

**Data Transformation and Reduction:** It refers to generalizing the data to higher-level concepts or normalizing the data. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1.0 to 1.0.

The data mining consists of various classification methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. In this paper, we are discussing Decision Table and RepTree.

### **Decision Table**

Decision Table algorithm classifier summarizes the dataset with a 'decision table' which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item [6, 7]. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a test model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table. We choose to explore decision tables because it is a simpler, less compute intensive algorithm. In this paper, we have used the Weka tools [8, 13, 14, 15], which determine which attribute or combination of attributes needs to be included for predicting the class best. The attribute space is searched greedily either top to bottom or bottom to top. A top-to-bottom search adds attributes at each stage; this is called *forward selection*. A bottom-to-top search starts with a full set of attributes and deletes attributes one at a time; this is *backward elimination*.

### **RepTree**

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

Classification rule mining and association rule mining are two important data mining techniques. They are similar except that classification involves prediction of one attribute, i.e., the class, while association rule discovery can predict any attribute in the data set [16, 17]. Classification rule mining aims to discover a small set of rules in the database to form an accurate classifier [18, 19]. Classification rule mining is indispensable to practical applications. Thus, great savings and conveniences to the user could result if the two mining techniques can somehow be integrated. [1]

## **III. STATISTICAL ANALYSIS AND ASSOCIATION RULES**

Statistical analysis involves use of observational data together with domain knowledge to develop a model to study and understand the trend of the data for prediction and decision. The data analysis is used to refine the model or possibly to select a different model, to determine appropriate values for terms in the model, and to use the model to make inferences concerning the process. Many alternative views of the data can be examined. Many different models can be explored. Massive amount of simulated data can be used to study the model/data possibilities. Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the “statistical significance” of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship. Regression analysis with a single explanatory variable is termed as *Simple Regression* and with multiple explanatory is termed as *Multiple Regression*. At the outset of any regression study, one formulates some hypothesis about the relationship between the variables of interest. Multiple regression allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables. [12]

In general, the association rule is an expression of the form  $X \Rightarrow Y$ , where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the *support* and *confidence* value.

**Support:** It is the probability of item or item sets in the given transactional data base:

$\text{Support}(X) = n(X)/n$  where  $n$  is the total number of transactions in the database and  $n(X)$  is the number of transactions that contains the item set  $X$ .

**Confidence:** It is conditional probability, for an association rule  $X \Rightarrow Y$  and defined as  $\text{Confidence}(X \Rightarrow Y) = \text{support}(X \text{ and } Y) / \text{support}(X)$ . All the traditional association rule mining algorithms were developed to find positive associations between items. Positive associations refer to associations between items existing in transactions. In addition to the positive associations, negative associations can provide valuable information. In practical there are many situations where negation of products plays a major role. For example, if I work full time in IITM then I do not work full time in any other institution. If I am a female then I am not a male.

Negative association rule is an implication of the form  $X \Rightarrow \neg Y$  where  $X$  and  $Y$  are item sets and  $X \cap Y = \emptyset$ .

Mining association rules can be broken down into the following two sub-problems:

1. Generating all item sets that have support greater than, or equal to, the user specified minimal support. That is, generating all large item sets.
2. Generating all the rules that have minimum confidence.

We can generate the association rule with more than one number of consequent items is generated by the following method:

- I. Find the rule in which number of consequents = 1.
- II. For the given rules  $p(x \rightarrow y)$  and  $p(x \rightarrow z)$ , the rule  $p(x \rightarrow yz)$  is generated by the intersection of both the association rules and get a new rule  $p(x \rightarrow yz) = p(xyz) / p(x)$ .

#### IV. PROPOSED ALGORITHM

In this algorithm, we have tried to remove noise by applying statistical and data mining methods on actual data and cleaned the data after that the classification is done on different methods and best method is identified. The steps are as follows: [5]

1. *Preprocessing the data based on statistical measures.*
  - i. *Identify relevant attributes by applying statistical techniques on actual data with Bank Transactions of customer and in general customer purchase Interest data.*
  - ii. *Applying Association rules to identify rules on actual data with Bank Transactions of customer and in general customer purchase Interest data.*
2. *Cleaning the data*
  - i. *Removing irrelevant data from the actual data set.*
  - ii. *Applying rules to remove noise from attributes.*
3. *Application of Classification algorithm*

*Classification methods are applied on cleaned dataset. In our case, we have chosen the following two methods:*

  - i. *Decision Table*
  - ii. *Rep Tree*
4. *Application of Classification algorithm on three different test modes.*
  - i. *Complete dataset as a training set*
  - ii. *Cross validation with 10 fold*
  - iii. *Splitting 75% as Training data set, 25% as Test data set*
5. *Determining the best combination.*

This algorithm makes the following contributions:

➤ It proposes a new way to build accurate classifiers.

Our Experimental results show that classifiers built this way are, in general, more accurate than those produced by the state-of-the-art classification system. (Table 1, Table 2, Graph 1 & Graph 2)

➤ It helps to solve the following important problems with the existing classification systems

- Most of the traditional data mining techniques failed because of the sheer size of the data. In our algorithm this problem is solved. In case of complete data set, the number of tuples got reduced from 5821 to 3421 attributes from 43 to 41 after preprocessing & cleaning. Where as in case of selected data set tuples are reduced from 5821 to 1440 and attributes from 26 to 24. (Table 1, Table 2, Graph 1 & Graph 2)

- Most of the algorithms assume the data to be noise-free. As a result, the most time-consuming part of solving problems becomes data preprocessing. Data formatting and experiment/result management are frequently just as time-consuming and frustrating. The concept of noisy data can be understood by the example of mining logs. A real life scenario can be if one wants to mine information from web logs. A user may have gone to a web site by mistake - incorrect URL or incorrect button press. In such a case, this information is useless if we are trying to deduce a sequence in which the user accessed the web pages. The logs may contain many such data items. These data items constitute data noise. A database may constitute upto 30-40% such Noisy data and pre-processing this data may take up more time than the actual algorithm execution time. This problem is handled at the time of preprocessing.

## V. EMPIRICAL EVALUATION

We have taken Real Estate data set containing 5821 tuples and 43 attributes. Firstly, we have applied statistical analysis to identify the attributes that are strongly correlated with each other, followed by applicability of Association rules to identify rules on actual data with customer Bank transactions and customer purchase Interest data. Those attributes were removed which doesn't have any relationship. In next step we applied classification algorithm on three test modes in two different environments. In first case, we have taken all the attributes (complete) and compared the patterns, results of both the methods. In second case, the same tests are performed, after discarding demographic details of customers (selected) and then compared the various results. It is found that if we are not considering demographic details of customers, we are getting very accurate results as well as the error rate is decreased. After considering the Bank Transaction details and Customer Purchase Interest, there is a drastic improvement in the results. The following paragraphs discuss the outcomes of various tests.

### *Comparative Analysis*

Table 1 shows the results of Decision table on Actual Real Estate complete dataset and selected data sets (after removing demographic details). Each of the tests is done on three test modes. The results are generated using WEKA 3-6-2, open source software for regression analysis and data mining.

The following are the findings, as per Table 1:

- As per traditional classification, in case of complete dataset, Training test mode takes minimum time, maximum correlation, minimum error rate. In general, if we consider complete data set then training test mode is the best method. It also depicts that the customers who are in the higher income group are willing to invest.
- As per our algorithm, in case of complete dataset, minimum time is taken by splitting test mode(75% training set and 25% test set)and maximum correlation, minimum error rates are generated by training test mode. In this case, customer in higher as well as higher middle income groups are willing to invest.
- As per traditional classification, in case of selected dataset (after removing demographic details), training test mode takes minimum time, maximum correlation, minimum error rates. Here it predicts that the customers who are in higher, middle income group, House owner are willing to invest.
- As per our algorithm, in case of selected dataset(after removing demographic details), minimum time is taken by Cross validation and splitting test mode, maximum correlation, minimum error rates are generated by training test mode. In this case, the customers from higher, middle as well as lower income group with other saving, rented house are willing to invest.

From the above results it is proved that there is a drastic improvement in results after preprocessing and cleaning, the factors which are very crucial in case of Indian context are having impact on the results. It also proves that training set gives the best results. Merit of best set is also improved after preprocessing and cleaning.

Table 2 shows the results of Reptree on Actual Real Estate dataset and selected data sets (after removing demographic details). Each of the tests is done on three test modes.

The following are the findings, as per Table 2:

- As per traditional classification , in case of complete dataset, minimum time is taken by splitting test mode(75% training set and 25% test set), maximum correlation, minimum error rate is taken by Training test mode i.e. in general, if we go for complete data set then training test mode is the best test mode.
- As per our algorithm, in case of complete dataset, minimum time is taken by splitting test mode(75% training set and 25% test set)and maximum correlation, minimum error rates are generated by training test mode.
- As per traditional classification , in case of selected dataset , all the test modes takes same time maximum correlation, minimum error rates are generated by training test mode.
- As per our algorithm, in case of selected dataset (after removing demographic details), minimum time is taken by Cross validation test mode, maximum correlation, minimum error rates are generated by training test mode.

After analyzing results of both classification methods, it is very much clear that after preprocessing and cleaning the results are more efficient and accurate. It is also proved that training test mode is best among other modes because it treats each tuple equally. Decision table method is better than Reptree because it is a simple, less compute intensive algorithm that involves the manipulation of counts associated with each of the table's entries. The attribute space is searched by a best-first search because this strategy is less likely to get stuck in a local maximum. The attribute selection is evaluated using a leave-one-out cross validation.

## VI. CONCLUSIONS

This paper discusses about the traditional classification algorithm, its limitations and suggests an efficient algorithm which does preprocessing using statistical techniques and association rules to remove the noise and clean the data. The effectiveness of this algorithm is that it not only considers the customer real estate data but also considers the customer's bank transactions as well as purchasing interest and gives more accurate results. The classification is done using Decision Table and RepTree. After analyzing results from both the classification methods, it is very much clear that our algorithm outperforms traditional classification algorithm. The results are more efficient, accurate and the error rates are reduced drastically. It is also proved that training test mode is best among other modes because it treats each tuple equally. Decision

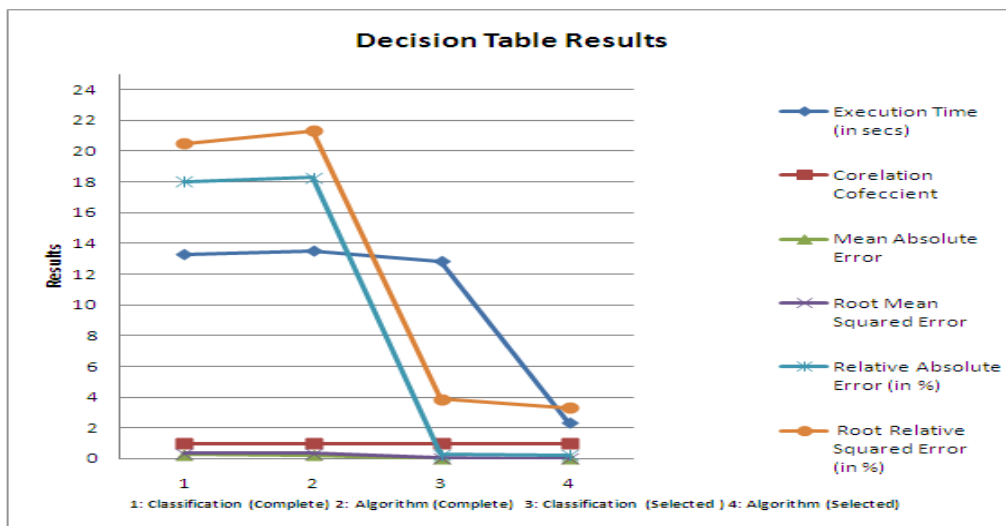
table method is better than Reptree because it is a simple, less compute intensive algorithm that involves the manipulation of counts associated with each of the table's entries. The attribute space is searched by a best-first search because this strategy is less likely to get stuck in a local maximum. The attribute selection is evaluated using a leave-one-out cross validation. It is proved that not only higher income group are willing to invest but also who in middle and higher middle group with other saving/investment with rented house or house owner are also willing to invest if, the bank transaction details and customer purchase interest are taken into account.

### References

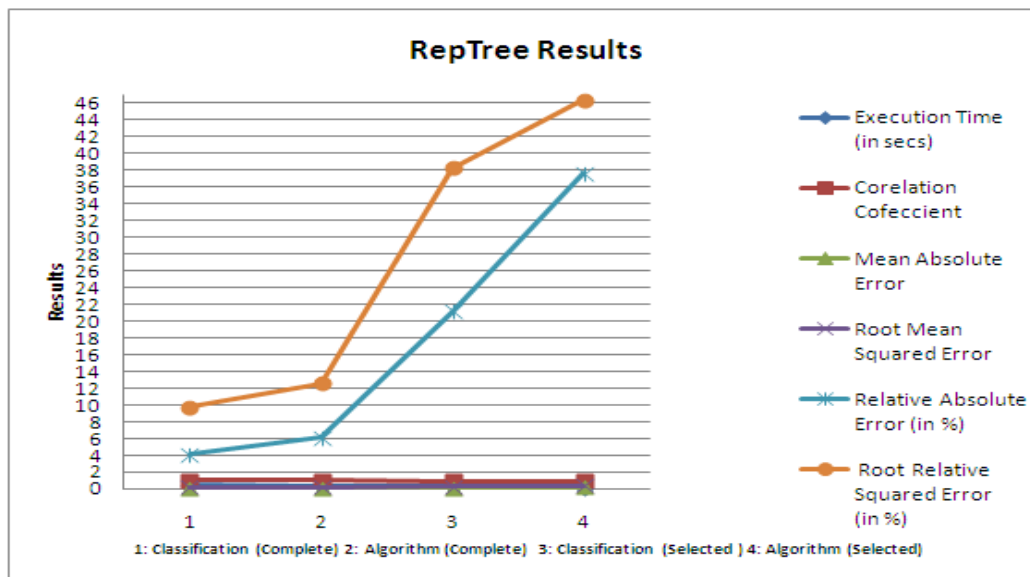
- [1] Bing Liu Wynne Hsu Yiming Ma, "Integrating Classification and Association Rule Mining", KDD-98 Proceedings. Copyright © 1998, AAAI ([www.aaai.org](http://www.aaai.org)).
- [2] Sunita Soni O.P.Vyas,"Using Associative Classifiers for Predictive Analysis in Health Care Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010
- [3] Han J. and Kamber M.,"Data Mining: Concepts and Techniques", 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001
- [4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine Volume 17 Number 3 (1996) (© AAAI)
- [5] Geetali Banerji, Kanak Saxena," An Algorithm for Rule based Classification", National Conference on Emerging Trends in Information Technology, March 2012, IITM New Delhi
- [6] Kathleen T. Durant Michael D. Smith , "Predicting UNIX commands using decision tables and decision trees", Research was supported in part by research grants from Compaq, IBM, Intel, Hewlett-Packard and Microsoft.
- [7] Ron Kohavi,"The Power of Decision table", 8th European Conference on Machine Learning, 174-189, 1995.
- [8] R. Bouckaert Remco, Eibe Frank et. al, "WEKA Manual for Version 3-6-2", January 11, 2010[8] R. Bouckaert Remco , Eibe Frank et. al, "WEKA Manual for Version 3-6-2",January 11, 2010
- [9] Geetali Banerji, Kanak Saxena, "Predictive Test model- A Boon for real estate" , International Journal for Wisdom Based Computing Volume(1) 2, April 2012
- [10] Geetali Banerji, Kanak Saxena," An Improved Apriori Based Algorithm with Single Scan of Database", National Conference on Converging Technologies Beyond 2020, April 2011, UIET, Kurukshetra University, India
- [11] Geetali Banerji, Kanak Saxena, "Analysis of Data Mining techniques on Real Estate", International Journal of Soft Computing and Engineering (IJSCE) ISSN:2231-2307 Volume-2, Issue-3, July 2012
- [12] <http://home.ubalt.edu/ntsbarsh/stat-data>
- [13] Witten I. frank, E. (2000),"Data Mining: *practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann", San Francisco.
- [14] CBA. Data mining tool. Downloading page. [http://www.comp.nus.edu.sg/~dm/p\\_download.html](http://www.comp.nus.edu.sg/~dm/p_download.html). Viewed on February 2010.
- [15] Weka. Data Mining software in Java. <http://www.cs.waikato.ac.nz/ml/weks>. Viewed on February 2010
- [16] Alaa AI Deen Mustafa Nofal and Sulieman Bani-Ahmad, "Classification Based on association Rule Mining Techniques: A general survey and Empirical Compative Evaluation".
- [17] B. Tunc and H. Dag, "Geneating Association Rules with Modified Apriori Algorithm", Proceedings of the 5<sup>th</sup> WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006
- [18] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and Regression Trees, Wadsforth International Group.
- [19] Quinlan, J. R. 1992, "C4.5: Program for Machine Learning", Morgan Kaufmann.

Decision Tree*	Traditional Algorithm on Complete Data Set			Our Algorithm on Complete Data Set			Traditional Algorithm on Selected Data Set			Our Algorithm on Selected Data Set		
	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii
Attributes	43			41			26			24		
No of Instances	5821			3421			5821			1440		
Rules	55			160			1652			444		
Subsets	278			300			264			200		
MBS	0.365			0.396			0.288			0.363		
Feature set:	INH,INH,IPC			INH,INH,INMH,IPC			INH,INH,INMH,HO,INM,IPC			INH,INH,INMH,HO,INM,RH,OI,IPC		
Test Modes	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii
Time (in secs)	13.3	21.25	22.97	13.52	13.59	8.31	12.84	13	13.33	2.34	2.16	2.16
CC	0.9788	0.977	0.976	0.977	0.9675	0.9703	0.999	0.928	0.907	0.999	0.852	0.847
MAE	0.2505	0.253	0.2649	0.2324	0.2747	0.2702	0.0019	0.0898	0.114	0.0012	0.1104	0.1202
RMSE	0.361	0.376	0.3841	0.345	0.409	0.4029	0.0306	0.296	0.336	0.024	0.383	0.399
RAE (in %)	18	18.17	18.78	18.29	21.62	20.77	0.284	13.59	17.13	0.248	23.624	24.347
RRSE (in %)	20.5	21.33	21.79	21.32	25.27	24.21	3.878	37.45	42.33	3.292	52.413	53.553

RepTree <sup>+</sup>	Traditional Algorithm			Our Algorithm on			Traditional Algorithm on			Our Algorithm on Selected		
Attributes	43			41			26			24		
No of Instances	5821			3421			5821			1440		
Rules	305			219			425			121		
Test Modes	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii
Time (in secs)	0.49	0.44	0.42	0.25	0.27	0.23	0.33	0.33	0.33	0.08	0.06	0.31
CC	0.9953	0.989	0.9873	0.9921	0.9846	0.9843	0.9239	0.8704	0.846	0.8862	0.7696	0.2535
MAE	0.057	0.097	0.1151	0.0779	0.1132	0.125	0.1399	0.1863	0.201	0.1754	0.255	0.2535
RMSE	0.171	0.257	0.28	0.203	0.283	0.295	0.3023	0.3904	0.427	0.3386	0.4695	0.4307
RAE (in %)	4.1	6.99	8.16	6.13	8.912	9.61	21.17	28.2	30.36	37.6	54.57	51.325
RRSE (in %)	9.73	14.56	15.88	12.57	17.51	17.71	38.27	49.41	53.85	46.34	64.22	57.852



Graph 1. Decision Table Results



Graph 2. Rep Tree Results

\*CC: Coefficient of Correlation, MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, RAE: Relative Absolute Error (%), RRSE: Root Relative Squared Error (%)

Real Estate

DATADICIONARY

NO.	Name	Description	NO.	Name	Description
1	CUSTYPE	Customer Subtype	21	SRS	Service
2	NOH	Number of houses	22	MGMT	Management
3	ASH	Avg size household	23	TL	Trained labor
4	AS	Avg age	24	UTL	Untrained labor
5	CMT	Customer main type	25	SCA	Social class A
6	NAT	Nationality	26	SCB1	Social class B1
7	CS	Caste	27	SCE2	Social class B2
8	SCS	Subcaste	28	SCC	Social class C
9	NR	No religion	29	SCD	1 Social class D
10	MRD	Married	30	RH	Rented house
11	LT	Living together	31	HO	Home owners
12	OR	Other relation	32	C1	1 car
13	SNG	Singles	33	C2	2 cars
14	HWTC	Household without children	34	NC	No car
15	HWTC	Household with children	35	PL	Policy Investment
16	HLE	High level education	36	OI	Other Investment
17	MLE	Medium level education	37	INL	Income < 20,000
18	LLL	Lower level education	38	INM	Income 20-55,000
19	HS	High status	39	INMH	Income 55-85,000
20	BUS	Business	40	INH	Income 85-125,000
41	INHH	Income >125,000	42	AI	Average income
43	IPC	Investment power class			

Bank Transaction Details

No.	Name	Description
1.	age	Age of the Customer
2.	sex	Sex of the Customer
3.	region	Region belongs
4.	income	Income Range
5.	save_act	Saving Account
6.	Curent_act	Current Account
7.	mortgage	Mortgage
8.	month_deposit	Monthly Deposit
9.	month_withdrawal	Monthly Withdrawal

Customer Purchase Interest

No.	Name	Description
1.	Location	Location of Flat/Plot
2.	Price	Total Price
3.	Bedrooms	No. of Bedrooms
4.	Bathrooms	No. of Bathrooms
5.	Size	Size of plot/Flat
6.	Price/SQ.Ft	Price per SQ. Feet