ISSN:0975-9646

Vishal Venkat Raman et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3081 - 3084

Identifying Customer Interest in Real Estate Using Data Mining Techniques

Vishal Venkat Raman, Swapnil Vijay, Sharmila Banu K

School of Computing Science and Engineering
VIT University, Vellore, Tamil Nadu 632014, INDIA

***Abstract*-- Real estate industry has become a highly competitive business with an enormous amount of unstructured documents and resources. The process of data mining in such a business provides an advantage to the developers by processing those data, predicting future trends and helping them to make favorable knowledge-driven decisions. In this paper, we focus on data mining techniques and its applications to develop a model which not only predicts the most suitable area for a customer based on his interest, but also identifies the most preferred and ideal location of real estate in any given area by ranking them.**

***Keywords*— data mining, infogainattributeeval, linear regression, ranking, rapidminer, weka.**

I. INTRODUCTION

Each and every company in today's real estate business is working productively to gain a competitive edge over other competitors. To achieve this, a reliable and a fast growing technology such as data mining is extensively used. With an enormous amount of data available in huge databases and data warehouses, it is important to develop a powerful model for analysis and extraction of such data and mining them for interesting knowledge [1].

Various data mining tools available in today's market helps organizations to make important and useful knowledge driven decisions. Hence, in this paper, we use data mining techniques in two different phases, where the first one deals with ranking of areas in each of the 5 zones of Delhi and identify an ideal location for real estate customers using the InfoGainAttributeEval function of WEKA tool [2]. The second phase is about predicting the most suitable area for any given customer based on his/her interest. We predict using a classical technique called linear regression and try to give an analysis of the results obtained.

II. LOCATION SURVEY & ITS FINDINGS

The first phase of this papers is all about differentiating the residential areas available in each of the zones of Delhi and ranking them. For this, we listed different attributes that affects the pricing of real estate in an area and then, identified the most influencing attribute for ranking. The city of Delhi, in total has 5 zones, namely, South Delhi, East Delhi, West Delhi, North Delhi and Central Delhi

A. The first step in this process is to identify all the residential locations in each of the zones. For this, we used a Geographic Information Systems (GIS) map of Delhi provided by the Municipal Corporation of Delhi (MCD) and Map My India. [3]

B. Using the GIS map, we were able to identify the different zones and wards of Delhi which can be considered as residential areas.

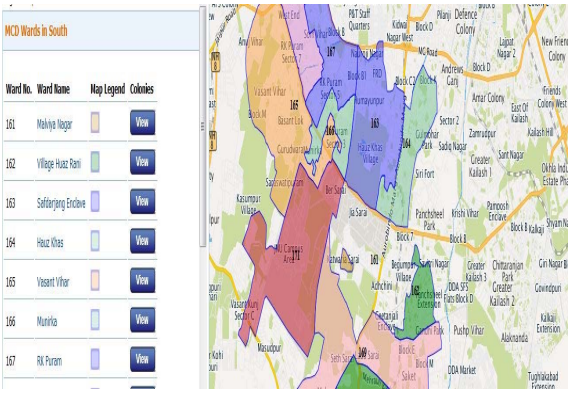
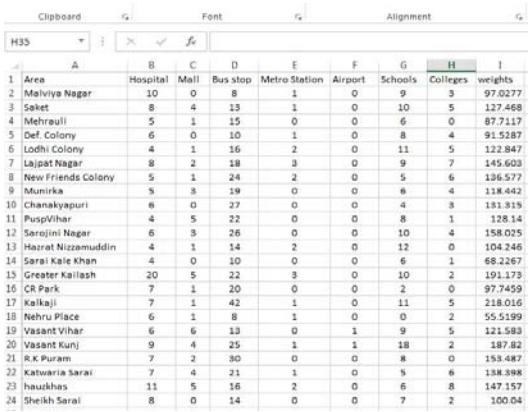


Fig 1. South Delhi zone showing all the residential areas

C. The identified residential locations of South Delhi are considered for finding the localities available in each one of them. The list of identified areas are: Malviya Nagar, Saket, Mehrauli, Defence Colony, Lodhi Colony, Lajpat Nagar, New Friends Colony, Munirka, Chanakyapuri, Pushp Vihar, Sarojini Nagar, Hazrat Nizamuddin, Sarai Kale Khan, Greater Kailash, CR Park, Kalkaji, Nehru Place, Vasant Vihar, Vasant Kunj, R.K. Puram, Katwaria Sarai, Hauz Khas and Sheikh Sarai.

D. The next step involves the finding of different facilities available in each of these areas. In our project, we have taken the following attributes that affect the pricing of real estate in an area: Hospitals, Malls & Markets, Metro Stations, Bus Stops, Airport, Schools and Colleges.



	A	B	C	D	E	F	G	H	I
1	Area	Hospital	Mall	Bus stop	Metro Station	Airport	Schools	Colleges	weights
2	Malviya Nagar	10	0	8	1	0	9	3	97.0277
3	Saket	8	4	13	1	0	10	5	127.468
4	Mehrauli	5	1	15	0	0	6	0	87.7117
5	Def. Colony	6	0	10	1	0	8	4	91.5287
6	Lodhi Colony	4	1	16	2	0	11	5	122.847
7	Lajpat Nagar	8	2	18	3	0	9	7	145.603
8	New Friends Colony	5	1	24	2	0	5	6	136.577
9	Munirka	5	3	19	0	0	6	4	118.442
10	Chanakyapuri	6	0	27	0	0	4	3	181.315
11	PuspVihar	4	5	22	0	0	8	1	128.14
12	Sarojini Nagar	6	3	26	0	0	10	4	158.025
13	Hazrat Nizzamuddin	4	1	14	2	0	12	0	104.246
14	Sarai Kale Khan	4	0	10	0	0	6	1	68.2267
15	Greater Kailash	20	5	22	3	0	10	2	191.173
16	CR Park	7	1	20	0	0	2	0	97.7459
17	Kalkaji	7	1	42	1	0	11	5	218.016
18	Nehru Place	6	1	8	1	0	0	2	55.5199
19	Vasant Vihar	6	6	13	0	1	9	5	121.583
20	Vasant Kunj	9	4	25	1	1	18	2	187.82
21	R.K Puram	7	2	30	0	0	8	0	153.487
22	Katwaria Sarai	7	4	21	1	0	5	6	138.398
23	hauzkhas	11	5	16	2	0	6	8	147.157
24	Sheikh Sarai	8	0	14	0	0	7	2	100.04

Fig. 2 Areas with attributes and its weights

E. The excel sheet containing all the areas and the number of total attributes as mentioned above is given as input to the WEKA tool and using the InfoGainAttributeEval function [4], the most influential attribute is identified. Using the math function obtained as given in the below screenshot, we calculate the weights of each area and rank them accordingly.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 Area):
        Information Gain Ranking Filter

Ranked attributes:
 3.3816   4 Bus stop
 3.2247   7 Schools
 3.0125   8 Colleges
 2.9097   2 Hospital
 2.577    3 Mall
 1.7901   5 Metro Station
 0        6 Airport

Selected attributes: 4,7,8,2,3,5,6 : 7
```

Fig. 3 Attribute Ranking using InfoGainAttributeEval in WEKA

Using the ranks obtained from the above mentioned steps, we find that Kalkaji is the ideal location for any real estate customer, followed by Greater Kailash, Vasant Kunj and so on. The same procedure can be applied in other zones of Delhi for finding their respective ideal locations.

III. DATA PREPARATION

A. Understanding of data

Huge amount of data is available in structured and unstructured format in various databases. [5] To select which data is useful and which is not is a major step towards data preparation. First, we identified the various amenities that directly or indirectly affect the pricing of a real estate. [6] We came up with a total of 23 attributes that were selected based on its relation to the real estate,

- Area in which it is locate
- Covered area
- Number of Bedrooms available
- Number of Bathrooms available
- Number of Balconies available
- Floor number
- Whether the real estate is furnished, unfurnished or semi-furnished
- Total Cost
- 24 hours water availability
- Security
- Piped Gas
- Power backup
- Reserved Parking
- Intercom facility
- Waste disposal provision
- Maintenance services
- Laundry service
- Gymnasium facilities
- Internet / WiFi services
- Park
- Vastu Compliant
- Lift services
- Club house

B. Calculating Cost of Real Estate

The standard rate / squeare feet value is taken from the Delhi Development Authority chart which assigns default rate for its 3 housing categories, Lower Income Group (LIG – 1 BHK), Middle Income Group (HIG – 2 BHK) and Higher Income Group (HIG / SFS – 3 BHK). [7]

Flats with lift	i) For upcoming schemes : - Rs.15,700 per sqm. (*Inclusive of one time cost component for operation and maintenance charges for lifts and fire fighting equipment and the like for a period of five years) ii) For existing Scheme - Rs.15,200 per sqm.
Flats without lift	- Rs.7,400 per sqm. For Janta/EWS/one Room Tenement - Rs.9,400 per sqm. For LIG/EHS Type –A - Rs.10,000 per sq.m. for LIG flats constructed on turnkey basis/Mega project - Rs.10,500 per sqm. For MIG/EHS Type-B - Rs.10,700 per sqm. For MIG flats constructed on Turnkey basis/mega project - Rs.11,000 per sqm. For HIG/SFS flats - Rs.700/- per sqm. Additional for underground common parking

Fig. 4 DDA standard rate chart

Based on the above rate per square feet, cost of real estate is determined by multiplying the area of the real estate with the rate per square feet of the corresponding category. Additionally, the cost also varies according to the amenities available in the real estate.

C. Organization of Data

Total of 23 attributes are listed along with area in which the real estate is located and the number of bedrooms available in it. We have collected about 500 data sets for this process. [8][9] After the processing stage, the data set is shown in the below figure.

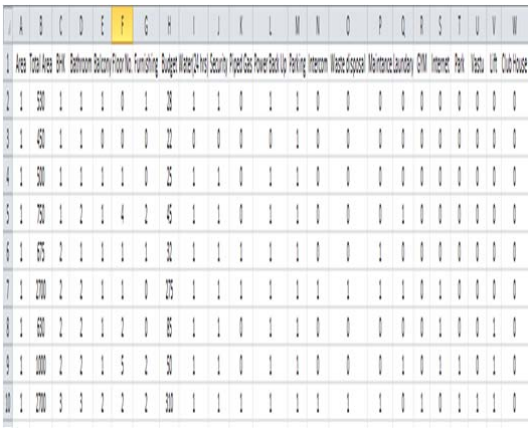


Fig. 5 Count of 23 attributes in each housing option

Numbers are assigned to the areas as shown in Fig. 7. The "*Furnishing*" attribute can take 3 values such as, 1 for fully furnished, 2 for semi – furnished and 0 for unfurnished. Further, the obtained attributes are converted in such a way that it is compatible with data mining operators available in RapidMiner. [10]

IV. ANALYSIS OF DATA MINING MODEL

Based on the dataset available, we are using 2 data mining modeling techniques namely, linear regression and rule induction for predicting the area for a real estate customer using Rapid Miner tool.

A. Linear Regression

Regression is a statistical measure used for prediction. It determines the relationship strength between dependent variable (known as label attribute) and other changing variables also called as independent variables (known as regular attribute). [6] Regression gives continuous value of the dependent variable that is used for prediction. Linear Regression operator in RapidMiner uses Akaike criterion for model selection. [10] The Akaike information criterion is a measure of the relative goodness of a fit of a statistical model. Mathematically, the linear regression formula is given by,

$$y = a + bx$$

where:

$$a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{n\sum (xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

The sample data set is loaded using the Retrieve operator. The Linear Regression operator is applied with default values of all parameters. Using the *ApplyModel* operator, the regression model generated by the *LinearRegression* operator is applied on the training data set which contains all the attributes except the "Area" attribute.

LinearRegression

```
  0.000 * Total Area
- 0.520 * BHK
+ 0.168 * Bathroom
- 0.305 * Balcony
+ 0.301 * Floor No.
+ 0.438 * Furnishing
- 0.000 * Budget
- 1.461 * Water(24 hrs)
+ 0.782 * Security
+ 1.573 * Piped Gas
- 3.204 * Power Back Up
+ 0.917 * Parking
- 2.208 * Laundary
- 3.146 * GYM
+ 1.723 * Internet
+ 2.241 * Park
+ 0.940 * Vastu
- 1.307 * Lift
+ 3.062 * Club House
+ 8.553
```

Fig. 6 Linear Regression function

The result obtained using the above obtained function is shown in the below figure

Row No.	prediction(Area)	Total Area	BHK	Bathroom	Balcony	Floor No.	Furnishing	Budget	Water(24 hrs)	Security	Piped Gas	Power Back Up	Parking	Intercom
1	7.478	1600	2	2	2	1	1	80	1	1	0	1	1	0

Fig. 7 Prediction result for a given customer

The result shows that, for a sample data set with *TotalArea*=1600, *BHK*=2, *Bathroom*=2, *Balcony*=2, *Floor No.*=1, *Furnishing*=1, *Budget*=80 with 24hrs water supply, security service, piped gas, power backup, reserved parking, intercom, etc. has predicted area value of 7.478, which also happens to be a continuous value which is expected because the *LinearRegression* operator always gives dependent variable values in continuous form. We can round of the predicted value to get the serial number of the area.

V. RESULTS AND ANALYSIS

The first phase deals in finding the most ideal location in a zone of any given city. In this paper, we have taken South Delhi zone of Delhi which contains 23 residential areas. To find the preferred location of South Delhi, we have used the GIS map of Delhi by which we were able to identify all the possible residential areas along with the local facilities like bus stops, metro stations, etc available in each area. Using WEKA, we could rank the attributes which influence the pricing of a real estate. The most influencing attribute as calculated by WEKA was bus stops, followed schools, colleges and so on.

The second phase dealt mostly with a survey process to predict a suitable area for any given real estate customer. First, we collected the standard rate/square ft. values of each type of housing as given by the Delhi Development Authority (2009 edition). The DDA provides housing in 3 categories, namely LIG (1 BHK), MIG (2 BHK) and HIG (3 BHK). Further, to calculate the actual cost of a real estate, the list amenities provided should be counted. It is the amenities provided by owners / builders and the area that finally affect the total cost of the apartment. For this, we have identified 23 possible amenities that are most commonly found in the apartments of Delhi, such as, piped gas, reserved parking, lift services, etc.

The survey contained questions for different categories of housing in all the areas and was stored in an excel sheet. To apply data mining technology in real estate, we have taken a random data with all the attributes, except for the "Area" attribute. To predict the area, we have used *Linear Regression* operator available in RapidMiner tool which gives us the predicted value of area as 7.478. Since the linear regression model gives a continuous value, the obtained result is converted to an integer, which is 7 in our case. This numerical value is nothing but the rank of the predicted area obtained in the first phase. The square error value of our linear regressions model is calculated to be 5.419 +/- 0.416.

VI. CONCLUSION

In today's highly competitive real estate business, it has become difficult to store such huge data and extract them for one's own requirement. Also, the extracted should be useful which is another complicated step, otherwise, there would be no use of using such data. Implementing the concepts of data mining applications in such a field has definitely proven to be useful, time saving and advantageous to both, the customers and the real estate dealers. Divided in two phases, we have applied data mining techniques such as linear regression for predicting a suitable area to the customer and identifying the most preferred residential area in Delhi. The same methods can be applied in finding the ideal location in rest of the zones of Delhi or in any given city.

VII. FUTURE WORK

Though, we were able to identify most of the residential areas in South Delhi, there may be some more places that have housing complexes or multi-storey apartments which are located in commercial areas. Such apartments were not included in this paper and can be counted in future to give a more accurate result. We have taken only those apartments that were constructed by the Delhi Development Authority (DDA) and were either sold directly to the customers / real estate dealers, or were given on lease. With more and more demand for housing in metropolitan cities, there is a definite increase in the number private builders that provide real estate with additional amenities to attract more customers. We have identified 23 amenities which can be found in any of the common housing plan that were provided by DDA. Hence, we have not considered apartments that were built by private builders and including those can provide more information about amenities such as DTH/Cable service, Private Garden/Terrace, etc. in the future.

There are several other models available that can be implemented for prediction. Data given as input to such model should be compatible with the tool used and the operators involved in the process. Also, more number of data sets can be used to increase the accuracy of the model. The main objective of using a different model should be to reduce the calculation time and carry out the whole process in ease.

REFERENCES

[1] Lu Ansheng, "Application Analysis of Clementine-based Data Mining Algorithm", 2011 International Conference on Intelligence Science and Information Engineering

[2] Swati Singh and Gaurav Dubey, "Finding Interest of People In Purchasing Real Estate By Using Data Mining Techniques", International Journal of Recent Technology and Engineering (IJRTE), Volume-X, Issue-X

[3] GIS Map of Delhi-NCR, http://alpha.mapmyindia.com/mcdApp/

[4] InfoGainAttributeEval – WEKA, "http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html"

[5] Aman Gupta and Gaurav Dubey, "Identifying Buying Preferences of Customers in Real Esstate Industry Using Data Mining Techniques", CPMR-IJT Vol. 2, No. 1, June 2012

[6] Zhou Xiaoming and Guan Wenjing, "Logistic Regression Analysis of Various Factors Affecting Customer Credit Risk", Financial Computer of Huanan 2003

[7] Delhi Development Authority Schemes, http://www.dda.org.in/housing/salient_features_conversion.htm

[8] www.magicbricks.com

[9] www.sulekha.com

[10] Fareed Akthar and Caroline Hahn, "RapidMiner 5: Operator Reference", Rapid-I GmbH 2012