House Price Forecasting using Data Mining

Nihar Bhagat
Department of Computer Engg.
Mumbai University,
RAIT, Nerul, India.

Ankit Mohokar
Department of Computer Engg.
Mumbai University,
RAIT, Nerul, India.

Shreyash Mane
Department of Computer Engg.
Mumbai University,
RAIT, Nerul, India.

ABSTRACT
People looking to buy a new home tend to be more conservative with their budgets and market strategies. The existing system involves calculation of house prices without the necessary prediction about future market trends and price increase. The goal of the paper is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted. The functioning of this paper involves a website which accepts customer's specifications and then combines the application of multiple linear regression algorithm of data mining. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

Keywords
Data mining, house price forecasting, prediction, linear regression, real estate.

1. INTRODUCTION
This paper brings together the latest research on prediction markets to further their utilization by economic forecasters. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This paper efficiently analyses previous market trends and price ranges, to predict future prices. This topic brings together the latest research on prediction markets to further their utilization by economic forecasters. It provides a description of prediction markets, and also the current markets which are useful in understanding the market which helps in making useful predictions. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This paper uses linear regression algorithm to predict prices by analyzing current house prices, thereby forecasting the future prices according to the user's requirements.

2. RELATED WORK
2.1 Identifying Customer Interest in Real Estate Using Data Mining Techniques
With a large amount of unstructured resources and documents, the Real estate industry has become a highly competitive business. The data mining process in such an industry provides an advantage to the developers by processing those data, forecasting future trends and thus assisting them to make favorable knowledge-driven decisions. In this paper, the main focus is on data mining method and its approach to develop a model which not only predicts the most suitable area for a customer according to his\her interests, and it also recognizes the most preferred location of real estate in any given area by ranking them.

This is used to predict a favorable location by ranking method. It analyses a set of locations selected by the customer. It broadly works on two basic phases. The first phase ranks a group of customer defined locations to find an ideal area and the second phase predicts the most suitable area according to their requirements and interest. It uses a classical technique called linear regression and tries to give an analysis of the results obtained. It helps establishes the relationship strength between dependent variable and other changing independent variable known as label attribute and regular attribute respectively. Regression displays continuous value of the dependent variable i.e. label attribute that is used for prediction. Linear Regression operator in RapidMiner [1].

2.1.1. Drawbacks:
It doesn't predict future prices of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process. This leads to the modification and development of the existing system.

3. PROPOSED SYSTEM
Nowadays, e-education and e-learning is highly influenced. Everything is shifting from manual to automated systems. The objective of this project is to predict the house prices so as to minimize the problems faced by the customer. The present method is that the customer approaches a real estate agent to manage his/her investments and suggest suitable estates for his investments. But this method is risky as the agent might predict wrong estates and thus leading to loss of the customer's investments.

The manual method which is currently used in the market is out dated and has high risk. So as to overcome this fault, there is a need for an updated and automated system. Data mining algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also the new system will be cost and time efficient. This will have simple operations. The proposed system works on Linear Regression Algorithm.

3.1. Algorithm used: Linear Regression
The database of property rates contains attributes like quarter, upper, average and lower, where each year from 2009 is divided into 4 quarters (q1: January-March, q2: April-June, q3: July- September, q4: October-December). The column *upper* consists of the average values of the houses that are high in prices, likewise *average* and *lower* column consists of average values of middle range and low range house [2]. In order to use linear regression the quarter attribute is assigned on x-axis and the values of rates on y-axis. For each of the attribute linear regression is performed once. The x-axis being independent is the choice available to the user to select from a dropdown list. The algorithm works as follows:

1. Read n //total number of points

2. Read x, y //x and y co-ordinates of points

3. Initialize diffx[n], diffy[n]
4. Initialize diffxy, diffx2 to 0
5. for i = 1 to n do
   calculate the mean of x : xm mean of y : ym
   diffx[i] = x[i] – xm //find the difference values between each x and mean of x
   diffy[i] = y[i] – ym //find the difference values between each y and mean of y
   diffx2 = Σ(diffx[i])2 //calculate the summation of all the difference values of x
   diffxy = Σ((diffx[i]) * (diffy[i])) //compute the product Of diff values of x and y
   end for
6. m = diffxy / diffx2 //the slope value is obtained by this Formula
7. c = ym – (m * xm) //the intercept value is obtained with this Formula
8. Equation complete: y = (m * x) + c
9. Stop.

By substituting the value of x in the obtained equation the respective y value can be found.

The figure below shows the real estate rate fluctuations in Navi Mumbai, India since 2009. The series divided as higher average and lower price ranges respectively. The equations denote their linear regression functions developed in accordance with the algorithm explained above. [3]

The following table consists of values of rates from January 2009 to December 2015. Each year divided in 4 quarters and represented as quarter no. in the table.

Table 1. Estate rates/ft^2

Quarter No.	upper	average	Lower
1	3187	2890	2677
2	3230	2975	2805
3	3315	3060	2890
4	3442	3230	3017
5	3612	3315	3060
6	3740	3442	3187
7	4080	3748	3485
8	4335	3995	3697
9	4675	4420	4080
10	4887	4505	4165
11	5100	4717	4420
12	5312	4930	4590
13	5270	4930	4590
14	5525	5100	4717
15	5625	5270	4887
16	5822	5482	5100
17	5992	5652	5227
18	6120	5652	5312
19	6120	5737	5397
20	6120	5780	5397
21	6417	5950	5525
22	6545	6120	5695
23	6715	6290	5822
24	7182	6630	6035
25	7267	6715	6120
26	7650	7097	6545
27	7862	7225	6672
28	7990	7310	6800

The graph below shows the linear regression performed on the available dataset. The dataset is divided into 3 categories namely; Upper, representing the houses by renowned builder associations and full of amenities followed by subsequent less categories as Average and Lower. The graph also contains the three equations of the respective categories which can be used to determine the prices of real estates in future. For example consider a customer needs to calculate the value of his real estate in January 2018. Then a difference in the time period is calculated quarterly to assign a value in the form of x-axis to January 2018. Now since according to the available dataset January 2009 is equal to 1 on x-axis, January 2018 will be equal to 37. Now by substituting this value of x-axis into whichever equation that represents his house category, consider higher range. Therefore the value of y, i.e. rate per sq. ft. generated is $y = 177.13*(37) + 2900.7$ .This value is approximately equal to 9455. Therefore the prediction can be made that the rate of that particular categorical house will increase to 9455 in January 2018. This information will provide help to customers looking to buy a new house to determine whether or not their investment will yield fruitfulness in the near future. Also investors looking to sell already owned properties will have a general idea about how much profit they will make by selling a property at what time. The entire system will work exactly similar with average as well as lower categories. The database will be added with new values as time progresses in order to consider the most recent developments in the real estate market.

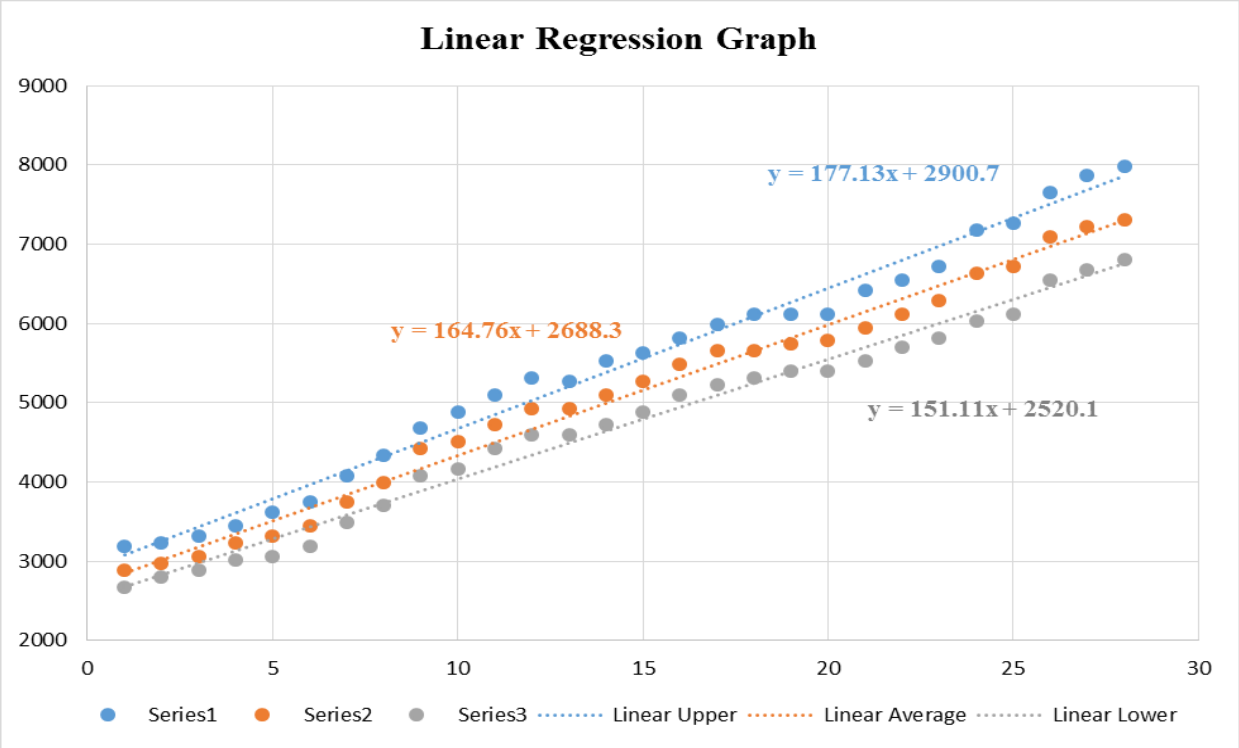

Figure 1: Price deviations from January 2009 in quarterly increments

3.2. Working of the System

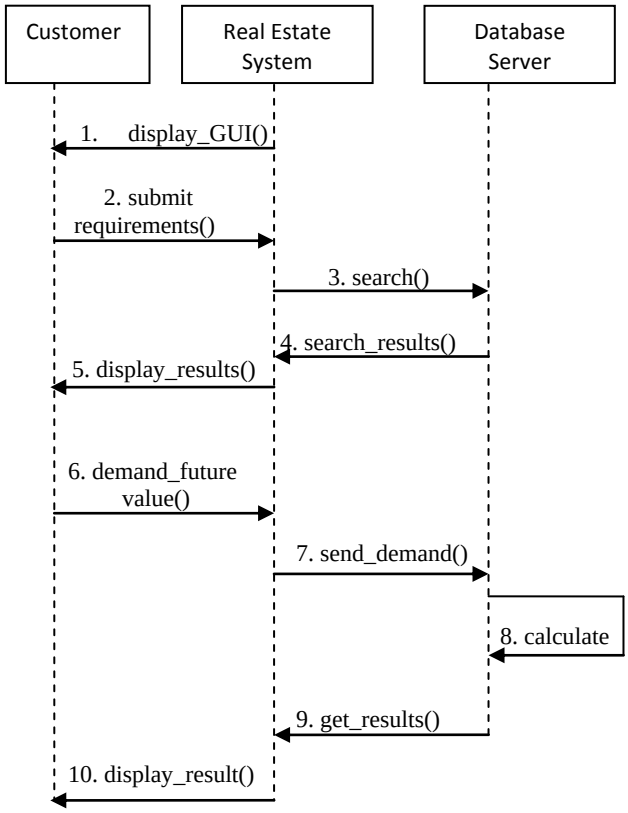

Figure 2: Sequence Diagram

The Sequence diagram above explains the working of the system. The proposed system is supposed to be a website with 3 objects namely: Customer, the Web Interface and the Database Server. The database server also includes the computational mechanism described in the algorithm. When the customer first enters into the website they are displayed with a GUI where they can enter inputs such as the type of house, the area in which it is located etc. A data index searching then provides with outputs consisting of matching properties. Now, if the customer wants to check the house price in future they can enter the date from the future. The system will identify the date and categorize it in the quarters. The algorithm then will compute the value of rate and provide the results back to the customer.

4. CONCLUSION

In today's real estate world, it has become tough to store such huge data and extract them for one's own requirement. Also, the extracted data should be useful. The system makes optimal use of the Linear Regression Algorithm. The system makes use of such data in the most efficient way. The linear regression algorithm helps to fulfill customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate. A lot's of features that could be added to make the system more widely acceptable. One of the major future scopes is adding estate database of more cities which will provide the user to explore more estates and reach an accurate decision. More factors like recession that affect the house prices shall be added. In-depth details of every property will be added to provide ample details of a desired estate. This will help the system to run on a larger level.

5. ACKNOWLEDGEMENT

It is a matter of great honour to bring out the paper on "**House Price Forecasting**". The project received excellent guidance of project guide **Mrs. Snehal Mumbaikar** and project co-ordinator **Mrs. Aditi Chhabria**. The project received their whole hearted assistance, inspiration, encouragement and valuable guidance in all phases. Also with the help of **Dr.Ramesh Vasappanavara**, Principal, Ramrao Adik Institute of Technology, **Dr. Leena Ragha** Head of the Department the project received great support. Lastly

thanking our group members and our class friends who directly or indirectly gave the ideas and help for the project. Also thanking to the non-teaching staff for providing the various facilities for the project.

6. REFERENCES

[1] Vishal Raman, May 2014. Identifying Customer Interest inReal Estate Using Data Mining.

[2] http://www.99acres.com/property-rates-and-price-trendsin-mumbai

[3] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introduction to Linear Regression Analysis

[4] Gongzhu Hu, Jinping Wang, and Wenying FengMultivariate Regression Modeling for Home ValueEstimates with Evaluation using Maximum Information Coefficient

[5] Iain Pardoe, 2008, Modeling Home Prices Using Realtor Data

[6] Aaron Ng, 2015, Machine Learning for a London Housing Price Prediction Mobile Application