International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 5 Issue 7, July 2024

House Price Forecasting using Machine Learning Algorithms

Shagun Tiwari[1] and Priyanka Bhatele[2]
[1]Sagar Institute of Science Technology and Research, Bhopal, India, 462044
[2]Sagar Institute of Science Technology and Research, Bhopal, India, 462044
[1]shagun.tiwarii@gmail.com, [2]priyankabhatele@sistec.ac.in

Abstract: People who want to buy a new house are more conservative with their budget and market strategies. In the current system, the calculation of home prices without the necessary forecasting about future market trends and value addition is included. The goal of the paper is to predict efficient house pricing for real estate customers regarding their budgets and priorities. By analyzing the trends and price limits of the previous market, and future developments, future prices will also be estimated. The work of this letter includes a website that accepts client specifications and then combines the application of several linear regression algorithm of data mining. This app will help customers to invest in an asset without an agent. It also reduces the risk involved in the transaction.

Keywords: *Data mining, machine learning, house price forecasting, prediction, linear regression.*

1. INTRODUCTION

This paper brings together the latest research on predicted markets to further their use by economic forecaster. Thus, real estate customers need to predict efficient house pricing in relation to their budgets and priorities. This paper efficiently analyzes previous market trends and price limits to predict future prices. This topic brings together the latest research on predicted markets to further their use by financial forecasts. This prediction provides details of markets, and current markets are also useful in understanding the market which helps in making useful predictions. Thus, real estate customers need to predict efficient house pricing in relation to their budgets and priorities. This paper uses linear regression algorithm to predict prices by presently analyzing the home prices, thereby estimating the future prices according to the user's requirements.

2. RELATED WORK

Large number of unstructured resources and documents, the real estate industry has become a highly competitive business. In such industries, the data mining process helps the developers by processing those data, anticipating future trends and thus making them friendly knowledge-driven decisions. In this paper, the main focus is to develop a model on the data mining method and its approach, which not only predicts the most suitable area according to its interest to a customer, and it is the most preferred location of real estate in any Recognizes that he was given the field by ranking [1]. It is used to predict the location favourable by the ranking method. It analyzes a set of selected locations by the customer. It works roughly on two basic steps. The first stage ranks a group of customer defined places to find an ideal area, and the second stage predicts the most appropriate area according to their needs and interests. It uses a classical technique called linear regression and tries to give an analysis of the results obtained [2]. It helps in establishing the relationship between dependable variables and other changing independent variables, which are known as label attributes and regular characteristics respectively. Regression represents the constant value of the dependent variable i.e. label attribute that is used for prediction of linear regression operator in Rapid Miner [4]. Applied Machine Learning Project 4 Prediction of real estate property prices analyzed the real estate property prices in Montreal. The information on the real estate listings was extracted from Centris.ca and duProprio.com. They predicted both asking and sold prices of real estate properties based on features such as geographical location, living area, and number of rooms, etc. Additional geographical features such as the nearest police station and fire station were extracted from the Montreal Open Data Portal; the final price sold was also predicted with an error of 0.023 using the Random Forest Regression [5]. A (FFBP) network model and (CFBP) network model are one of these tasks used in Data Mining Model by Using ANN to compare results of them.

Ideal Sabri Hashim Bahia concludes that which one of the two networks appears to be a better indicator of the output data to target data network structure than maximizing predict. Paper aims to demonstrate the importance and possible value of housing predictive power which provides independent real estate market forecasts on home prices by using data mining tasks.[11]Yu, Jiafu Wu. has predicted house prices given explanatory variables that cover many aspects of residential houses. As continuous house prices, they are predicted with various regression techniques including Lasso, Ridge, SVM regression, and Random Forest regression; as individual price ranges, they are predicted with classification methods including Naive Bayes, logistic regression, SVM classification, and Random Forest classification.[1]Da-Ying Li proposes support vector regression (SVR) to forecast real estate prices in China. Aim of the paper is to examine the feasibility of SVR in real estate price prediction. To achieve the aim, five indicators are selected as the input variables and real estate price is used as output variable of the SVR. The quarterly data during 1998-2008 are employed as the data set to construct the SVR model. With the scenarios, real estate prices in future are forecasted and analyzed. The forecasting performance of SVR model was also compared with BPNN model. [10] In Prediction of Real Estate Price Variation Based on Economic Parameters L. Li, K.-H. Chu, has used macroeconomic parameters on real estate price variation are investigated before establishing the price fluctuation prediction model. Here, back propagation neural network (BPNN) and radial basis function neural network (RBF) two schemes are employed to establish the nonlinear model for real estate's price variation prediction of Taipei, Taiwan based on leading and simultaneous economic indices. The public related data of Taipei, Taiwan real estate variation during 2005-2015 are adopted for analysis and prediction comparison [3].

3. RELATED TERMS

Supervised Learning

Supervised machine learning is a machine learning algorithm that uses a labelled dataset for prediction. Labelled Data means an output is associated with every input. In other words, supervised learning builds a model using this dataset that can make a prediction of the new or unseen dataset. This unseen or new dataset is called training dataset and helps in validating the model. Supervised algorithms can be classification algorithms like Support Vector Machines, Naive Bayes, Decision trees or regression algorithms like linear regression, neural networks, and decision trees. Since decision trees can be used for both classification and regression, they are one of the best-known supervised algorithms.

Ensemble Learning

It is a machine learning paradigm in which multiple weak learners are trained and combined to form a strong learner. It combines a diverse set of individual learners together to improvise on the predictive power and stability of the model. A number of learners (base learners) form an ensemble. The ensemble has the stronger ability for generalization than that of individual base learners. Ensemble learning is a better choice because it has the ability to boost weak learners which are slightly better than a random guess to strong learners which can make very accurate predictions. "Base learners" are also called "weak learners". On training by a base learning algorithm like a decision tree, neural network or others, base learners are generated. Ensemble methods generate base learner in two ways:

1). Homogeneous base learners, in which same base learning algorithm is applied to generate individual learners.
2). Heterogeneous base learners, in which different base learning algorithms are applied to generate individual learners.[13]

Random Forest

A combination of many tree predictors (here decision tree) such that each tree is generated independently from another tree. As the number of trees in the forest increases the generalization error for the forest converges. Factors on which generalization error depends are i) strength of the individual trees and ii) correlation among the trees in the forest.[12] Definition : A random forest is a classifier consisting of a collection of tree-structured classifiers {h(, k),k=1…..} where the{ k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x .

Feature Extraction

Sometimes, a dataset may include some features that may not be as important as other features. Such features do not help much in classifying data or sometimes may lead to low accuracy. It is important to select the best features so that a better and efficient model is created. Using same

features, again and again, can cause overfitting and underfitting of data [15-18]. Techniques such as k-fold cross-validation can be used to solve overfitting. This will help in constructing a better model. Prediction of the house price is dependent on some of the major features of the house and few of them are:

1). Location.
2). Parking.
3). Amenities.
4). Stamp Duty Rate.
5). No of rooms.
6). Nearby Places
7). Facing towards.

Data is collected and stored in NoSQL / SQL format. That data is divided into two parts

1). Training data
2). Testing data.

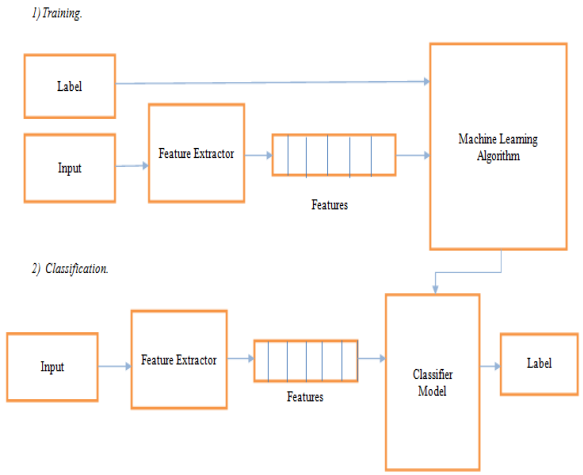

Fig. 1: ML Model

4. PROBLEM IDENTIFICATION

It does not estimate the future prices of houses mentioned by the customer. Because of this, exposure to investment in an apartment or an area increases significantly. To reduce this error, customers employ an agent, which increases the cost of the process again. This leads to modification and development of the existing system.

5. SYSTEM DESIGN

The overall system design consists of following modules:

(a) Data Collection.
(b) Preprocessing
(c) Data Classification.
(d) Data regression.
(e) Prediction of Output.

Data is collected from various data sources, that data can be structured or unstructured format. For structured data SQL techniques are used to extract data and for unstructured data NOSQL techniques are used to extract data. Classification and then regression algorithms can be applied for price prediction on data.

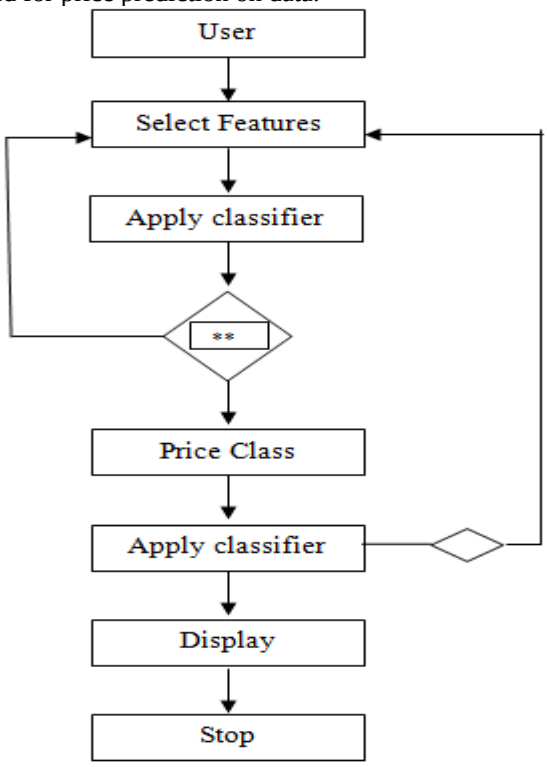

Fig. 2: System Flow Diagram

6. PROPOSED SYSTEM

In Recent Days, e-learning and e-learning are highly influenced. Everything is moving from manual to automated systems. The purpose of this project is to predict home prices so that the problems faced by the customer can be minimized. The current method is that the customer contacts a real estate agent to manage his investments and suggest suitable assets for his investment. But this method is risky because the agent can guess the wrong estates and thus can harm the customer's investment.

Manual method which is currently used in the market is dated and there is high risk in it. Therefore, to overcome this flaw, an update and automatic system is required. Data mining algorithms can be used to help investors invest in appropriate assets, according to their specific requirements. Also new system cost and time will be efficient. It will be simple operation. The proposed system works on linear regression algorithms. 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.

7. 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 of features 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.

REFERENCES

[1] Vishal Raman, May 2014. Identifying Customer Interest in Real 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 FengMulti-variate Regression Modeling for Home Value Estimates 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

[7] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: an overview," IEEE Transactions onIndustrialElectronics, 2014.

[8] C.L. Huang,M.C.Chen,and C.J.Wang, "Credit scoring with a data mining approach based on support vector machines," Expert Systems with Applications, vol. 33, no. 4, pp. 847–856, 2007.

[9] S. Ding, S. Yin,K. Peng, H.Hao, and B.Shen,"A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill," IEEE transaction on Industrial Informatics,vol.9,no.4,pp.2239–2247,2013.

[10] A. Abraham, Artificial Neural Networks, Hand book of Measuring System Design, 2005.

[11] Iteral Sabri Hashim Bahia, 2013. A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study, International Journal of Intelligence Science, pp. 162-169.

[12] Leo Breiman, 2001. Random Forests, Statistics Department University of California Berkeley, CA 94720.

[13] Zhou ZH. ,2015. Ensemble Learning. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA.

[14] World Population Review (2017). Retrieved from http://worldpopulationreview.com/worldcities/punepopulation/

[15] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[16] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[17] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[18] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.