Real Estate Value Prediction Using Linear Regression

Nehal N Ghosalkar
Department Of Computer Engineering
Sardar Patel Institute of Technology
Maharashtra, India
nehal.ghosalkar@spit.ac.in

Sudhir N Dhage
Department Of Computer Engineering
Sardar Patel Institute of Technology
Maharashtra, India
sudhir_dhage@spit.ac.in

***Abstract*—The real estate market is a standout amongst the most focused regarding pricing and keeps fluctuating. It is one of the prime fields to apply the ideas of machine learning on how to enhance and foresee the costs with high accuracy. There are three factors that influence the price of a house which includes physical conditions, concepts and location. The current framework includes estimating the price of houses without any expectations of market prices and cost increment. The objective of the paper is prediction of residential prices for the customers considering their financial plans and needs. By breaking down past market patterns and value ranges, and coming advancements future costs will be anticipated. This examination means to predict house prices in Mumbai city with Linear Regression. It will help clients to put resources into a bequest without moving toward a broker. The result from this research proved linear regression gives minimum prediction error which is 0.3713.**

***Index Terms*—Machine Learning, Linear Regression, MSE, RMSE**

I. Introduction

The study of real estate value is felt critical to help the choices in urban arranging. The land framework is a precarious stochastic process. Financial specialists choices depend on available patterns to procure most extreme returns. Designers are intrigued in knowing the future patterns for their basic leadership. To precisely gauge real estate costs what's more, future patterns, vast measure of information that impacts arrive cost is required for examination, demonstrating and determining.

According to the Census 2011, eight out of ten households in India own a house. But, this is mainly due because most people in rural areas have their own houses. Only 69 per cent of urban households own a home. The reason for this is soaring high prices of the property in urban area and nondeterministic nature of the house prices.[3]

In India, the property is sold as per the wish of seller. Thus, it is a biased procedure to buy a house in India since there is no standard way to list the selling price of the property. Very less work on real estate valuation is done in India. People in India believe on what is shown on the mass media. But mass media can manipulate the content as per their convenience and profits.[2]

As a financial specialist in the real estate business, one needs to comprehend the subtle elements of land showcase and what parameters influence its costs. As a typical man it isn't conceivable to contemplate the different market patterns and its impact on the property costs in detail. Henceforth, a device which comprehends these patterns and impact of different parameters on the property costs is required. There are different machine learning procedures that can be utilized to foresee the future esteems.

In any case, we require a model that can foresee the future property estimations with more noteworthy precision and least mistake. With a specific end goal to prepare the model, we require substantial measure of memorable dataset. Since less research work is done on forecast of land property in India, We might want to manufacture a framework. That can predict the property cost by taking into consideration the various parameters that affect the target value. And also to measure the prediction accuracy by taking into consideration different error metrics.

II. Background

The components that influence the land cost have to be considered and their effect on cost has additionally to be demonstrated. An examination of the past information uncovered that the costs demonstrate a non-direct trademark. It is construed that building up a basic direct numerical relationship for these time-arrangement information is found not reasonable for anticipating. Thus it wounds up basic to build up a non-direct model which fits the information trademark to dissect and estimate future patterns.

As the land segment is quick creating in Mumbai Metropolitan Area (MMA), the examination and figure of land costs utilizing numerical displaying and other logical procedures is a prompt critical requirement to take necessary decisions.

R. Manjula,[1] have come across with some calculation called as arrange plunge calculation which radically decreases the calculation workload,limiting the number of highlights while choosing the main essential ones. Organizations like "Zillow.com","magicbricks.com", consists of a vast dataset of houses whose costs they anticipate utilizing machine learning. One of the procedures they utilize is Linear and multivariate regression, profound learning to take in the idea of models from the past outcomes.

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

In the above paper author have characterized direct model information utilizing just a single element, multivariate model, utilizing a few highlights as its information and designed model utilizing the contribution as cubed or squared and henceforth ascertained the root mean square blunder (RMS esteem) for the proposed model.

Nissan Pow, anticipated both soliciting and sold costs from land properties in view of highlights, for example, topographical area, living territory, and number of rooms, and so forth.[8] Extra geological highlights, for example, the closest police headquarters and re station were removed from the Montr'eal Open Data Portal. They used techniques, like direct regression, Support Vector Regression (SVR), k-Nearest Neighbors (kNN), and Regression Tree/Random Forest Regression. Their result stated the soliciting cost with a mistake from 0.0985 utilizing an outfit of kNN and Random Forest calculations. Furthermore, where relevant, the nal cost sold was additionally anticipated with a mistake of 0.023 utilizing the Random Forest Regression. Their conclusion was on basis of the subtle elements of the expectation addresses, the examination of the land postings, and the testing and approval that comes about for the distinctive calculations in this paper.

Eduard Hromada [1] speaks about the distinctive strides from gathering the information from different promotions and land sites and sending out it into different classifications which is additionally broke down and confirmed. After the confirmation of information, the product device makes measurements plans which will examine relations among checked factors and depict the land showcase as indicated by clients A-Z request.

Machine learning algorithms are applied for the analysis of real data on the new housing market of Santiago, Chile.[10] Their goal is to look at the prescient execution of the Neural Network, Random Forest and Support Vector Machine approaches with conventional Ordinary Least Squares Regression. The database for our examination comprises of an example of 16,472 value records for new lodging units or private properties inside the region secured. The consequences of the examination demonstrate that Random Forest performed superior to alternate models in displaying lodging costs. All the more by and large, it can be assumed that machine learning procedures can give a valuable arrangement of apparatuses for securing information on housing prices.

Li Li and Kai-Hsuan Chuet has studied that real estate price variation has complicated the behaviors non-linearly and some uncertainty.[7] Author has used mathematical model free feature of neural network algorithm. They have used back propagation neural system (BPN) and outspread premise work neural system (RBF) two plans are utilized to set up the nonlinear model for genuine homes value variety expectation of Taipei, Taiwan in view of driving and concurrent monetary lists.The mean supreme value and root mean square blunder two lists of the value variety are chosen as the execution list. Thus based on the research author has concluded that the variation of house price trend is not that accurate.

House costs increments consistently, so there is a requirement for a framework to anticipate house costs later on.[6] House value expectation can enable the constructor to decide the offering cost of a house and it also helps the client to organize the correct time to buy a house. So This paper tends to foresee house costs in light of NJOP houses in Malang city with regression investigation and particle swarm advancement (PSO). PSO is utilized for choice of effect factors and relapse investigation is utilized to decide the ideal coefficient in forecast. The outcome from proposed paper gives the exploration demonstrated mix relapse and PSO is reasonable and the least error is IDR 14.186.

As the land prices keeps on changing consequently the study regarding its cost is necessary.[3] In this paper the atuhors comprehensively take a shot at two fundamental stages. The principal stage positions a gathering of client characterized areas to locate a perfect region and the second stage predicts the most appropriate zone as per their necessities and intrigue. It utilizes a traditional method called straight relapse and it gives an investigation of the outcomes acquired. It helps builds up the relationship quality between subordinate variable and other changing autonomous variable known as name characteristic and customary property individually. Regression shows consistent estimation of the reliant value i.e. name quality which is being utilized for the forecast.

Lastly, the authors here [9] have made a similar investigation of different Machine Learning calculations to be specific Linear Regression utilizing slope plunge, K closest neighbour relapse and Random backwoods relapse for forecast of land value patterns. The point of this paper is to look at the practicality of these machine learning calculations and select the most attainable one. To accomplish the point, parameters like Living Area, Number of rooms, Distance from air terminal/ expressway/ station/ significant,historic points, Proximity to healing facilities, Shopping choices, Number of theaters, land location(harbor/central/western) are utilized as the contribution to the model and land cost in the following quarters is the yield variable.

The quarterly information amid 2005-2016 is utilized as the informational collection to develop the model and the information has been gotten utilizing Web Scraping from sites like 99acres.com, Magicbricks.com, Google.com. The test comes about on the preparation informational index are utilized to think about the different calculations in view of blunder computation utilizing Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE).

As different research are done by authors using various Machine Learning Algorithms, it is seen than predicting real estate cost is a complex study. The study shows various results obtained from each of the papers but the missing factor is that it do not foresee future costs of the houses specified by the client. Because of this results, the hazard in interest is a condo or a zone increments extensively. To limit this mistake, clients tend to procure a broker which again expands the cost of the procedure. This prompts the alteration and improvement of the current framework.

III. System Design and Architecture

The Design is separated into three main stages: Initial, Middle,Last stage. The Initial stage is identified with Data accumulation and Analysis. The center stage comprise of different sub-stages like Feature Selection, Training the Linear Regression model and SVM display, Validating the model and Measure the mistake measurements. Last Phase includes the Visualization of the final products.

Phase I: Collection of Data

Collection of data is the process of measuring and gathering the information with the help of a software. There are many techniques and procedures to collect the data. We will be collecting the quantitative data which is structured and categorized. Real Estate house price indices reports and documents are put up on the internet by the Indian Government. These reports consist of the historical data of the past years which will be useful in our project. Before any kind of machine learning analysis, data collection is required. However validity of the dataset is a must otherwise there is no point in analysing the data. Hence we need to careful about the source of the data and check its validity beforehand.

Phase II: Data Cleaning and Loading

Data cleansing is the process of cleaning our data set. Their could be various garbage values present in the dataset. These garbage values can be removed by checking whether any missing values are present in the data or not. We also need to nd the validity of our dataset. Also the values need to be present in a given range. If a variable has many missing values we can drop those values. We also need to normalize the data before applying algorithm to it because every parameter has different units and the output will not be normalized. Hence, we need to normalize the dataset.

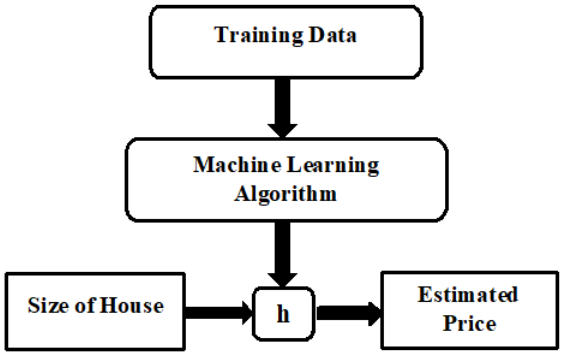

Fig. 1. Flow Model

Phase III: Feature Selection

The selection of feature is also known as variable selection. It is a process in which subset selection of parameters or variables from a large universal set of parameters is done. Feature selection techniques are used for four reasons:
To simplify the model
To shorten the training time of the model
To reduce the dimensionality.

Phase IV: Train LR model

Since the data is divided into two modules: Training set and Test set, we will be initially training the model. Target variable will be present in the training set. Thus linear regression tries to t the curve according to the given dataset with minimum error.

Phase V: Validation of model

Validation is the process of checking whether the applied algorithm ts the given dataset or not. Thus the accuracy of the model should be as high as possible. After applying the algorithm we can check how well our model ts the data. We can also apply two or more models to check the model or which ts our dataset the best. The model is viewed as an input-output transformation for these tests. The validation test compares the outputs from the system that is under consideration to the outputs that are obtained from the model provided that same input parameters are given to the model. The output values obtained from the model are recorded.

Fig. 2. Real Estate Prediction

IV. Methodology

Predicting the real estate values requires large number of factors such as locality, urban proximity, number of floors, shelf life, general rental units, number of bedrooms, bathrooms provided, parking space alloted, elevator, style of construction, total oor space, balcony space, condition of building, price per meter square of oor space. Thus there are various parameters which decide the price of a property which are co related to each other. Thus it becomes difcult to use numerous variables which are dependent. We will predict our target value using: Linear Regression Model.[5]

Linear Regression is extremely valuable device in prescient examination.

$E(Y / X) = f(X, B)$

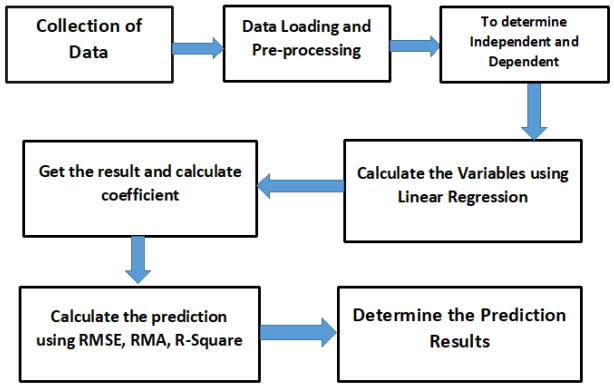

Fig. 3. Data Flow Model

Algorithm used: Linear Regression

The database of property rates contains properties like quarter, upper, normal and lower. The section upper comprises of the normal estimations of the houses that are high in costs, similarly normal and lower segment comprises of normal estimations of center range and low range house. Keeping in mind the end goal to utilize straight relapse the quarter trait is alloted on x-axis and the estimations of rates on y-axis. For every one of the quality direct relapse is performed once. The x-axis being autonomous is the decision accessible to the client to choose from a dropdown list.

Fig. 4. Price Deviation

In Linear Regression, we accept that there is a connection between autonomous variable vector and the needy target variable. By utilizing the free parameters, we can anticipate the objective variable. The autonomous information vector can be a vector of N parameters or properties. They are otherwise called regressors. It accept that the connection between subordinate variable and regressors is direct. The aggravation in anticipated esteem and the watched esteem is named as blunder.[6]

The subsequent stage is to distinguish best-fitting relationship (line) between the factors. The most widely recognized technique is the Residual Sum of Squares (RSS). This technique ascertains the distinction between watched information (real esteem) and its vertical separation from the proposed best-fitting line (anticipated esteem). It squares every distinction and includes every one of them.

The MSE (Mean Squared Error) is a quality measure for the estimator by partitioning RSS by add up to watched information focuses. It is dependably a non-negative number. Qualities more like zero speak to a littler blunder. The RMSE (Root Mean Squared Error) is the square base of the MSE. The RMSE is a measure of the normal deviation of the appraisals from the watched esteems. This is less demanding to watch contrast with MSE, which can be a vast number.

Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

Linear Regression will predict the exact numerical target value unlike other models which can only classify the output. Thus Linear Regression plays a strong role in predicting the price value of real estate property.

V. Conclusion

In the present real estate world, it has turned out to be difficult to store huge amount of information and concentrate them for one's own prerequisite. Likewise, the separated information ought to be helpful. The framework makes ideal utilization of the Linear Regression Algorithm. It makes use of such information in the most effective way. The direct relapse calculation satisfies customer by expanding the exactness of their decision and diminishing the danger of putting resources into a home. A ton's of highlights that could be added to make the framework all the more generally satisfactory.[8] One of the real future extensions is including home database of more urban areas which will give the client to investigate more domains and achieve an exact choice. More factors like subsidence that influence the house costs should be included. Top to bottom subtle elements of each property will be added to give plentiful points of interest of a coveted domain. This will help the framework to keep running on a bigger level.

References

[1] R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher, "Real estate value prediction using multivariate regression models," IOP Conference Series: Materials Science and Engineering, 2017.

[2] V.Sampathkumara, M.Helen Santhib and J.Vanjinathan, "Forecasting the land price using statistical and neural network software," 3rd International Conference on Recent Trends in Computing, 2015.

[3] Nihar Bhagat, Ankit Mohorkar and Shreyas Mane, "House Price Forecasting using Data Mining," International Journal of Computer Applications, 2016.
[4] Eduard Hromada, "Mapping of real estate prices using data mining techniques," Czech Technical University, Czech Republic, 2015.
[5] Pallav Ranka and Prof. Kripa Shanker, "Stock Market Prediction using Artificial Neural Networks," Indian Institute of Technology, Kanpur (208016), India.
[6] Adyan Nur Alfiyatin and Ruth Ema Febrita, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization," International Journal of Advanced Computer Science and Applications, 2017
[7] Li Li and Kai-Hsuan Chu, "Prediction of Real Estate Price Variation Based on Economic Parameters," Department of Financial Management, Business School, Nankai University, 2017.
[8] Nissan Pow, Emil Janulewicz and Liu Dave, "Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal," 2016.
[9] Dr. Swapna Borde, Aniket Rane, Gautam Shende and Sampath Shetty, "Real Estate Investment Advising Using Machine Learning," IRJET, 2017.
[10] "Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile," AMSE Conference Santiago, Chile, 2016.
[11] Mansurul Bhuiyan and Mohammad Al Hasan, "Waiting to be sold: Prediction of Time-Dependent house selling probability," IEEE International Conference on Data Science and Advanced Analytics, 2016.
[12] Wan Teng Lim, Lipo Wang,Yaoli Wang and Quing Chang, "Housing Price Prediction Using Neural Networks," IEEE 12th International Conference on Natural Computations, Fuzzy Systems and Knowledge Discovery, 2016.
[13] Muhammad A. Razi and Kuriakose Athappilly, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classication and regression tree(CART) models," Western Michigan University, 2005.
[14] Youness El Hamzaoui and Jose Alfredo Hernandez Perez, "Application of articial neural networks to predict the selling price in the real estate valuation process," Morelos, Mexico, 10th Mexican International Conference on Articial Intelligence, 2011.
[15] Ruben D. Jaen, "Data Mining: An empirical Application in Real Estate Valuation," Florida International University, 2002.