IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

Real Estate Price Prediction System

Rasika Bakre[1], Tanvi Kate[2], Prof. Harshwardhan Kharpate[3]
Students, Department of Computer Engineering[1,2]
Guide, Department of Computer Engineering[3]
Cummins College of Engineering for Women, Nagpur, Maharashtra, India

Abstract: *Due to the large increase in the land prices every year, the sale price of real estate, rented property and freeholds also increase consequently. People often find themselves paying unrealistic prices to the brokers and agents or even for the land, for a home of their preferences. So determining the fair price of the house based on various factors such as the location, the locality, the size of the house, etc becomes very crucial. To eliminate the overpriced rented property and heavy brokerage, we have designed a ML model using various regression techniques like decision tree, random forest and xgboost, so that it can predict genuine prices of real estate and can be used as a reference.*

Keywords: Real Estate, House Price, Machine Learning, Regression algorithm, decision tree.

I. INTRODUCTION

1.1. Background

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.

Machine Learning (ML) is a vital aspect of present-day business and research. It progressively improves the performance of computer systems by using algorithms and neural network models. Machine Learning algorithms automatically build a mathematical model using sample data also referred to as training data which form decisions without being specifically programmed to make those decisions.The primary aim of this paper is to use these Machine Learning Techniques and curate them into ML models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need.

1.2. Machine Learning

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. It allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on a predetermined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.
Several Machine Learning algorithms are used to solve problems in the real world today.

1.3. Regression Analysis

The term regression is used to indicate the estimation or prediction of the average value of one variable for a specified value of another variable. Regression Analysis is a statistical tool used to estimate the relationship between a dependent variable (y) and an independent variable (x).
The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

- Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

Three major uses for regression analysis are

1. Determining the strength of predictors,
2. Forecasting an effect, and
3. Trend forecasting.

II. Literature Survey

2.1. Indian Real Estate Industry

By 2040, the real estate market will grow to Rs. 65,000 crore (US$ 9.30 billion) from Rs. 12,000 crore (US$ 1.72 billion) in 2019. Real estate sector in India is expected to reach US$ 1 trillion in market size by 2030, up from US$ 200 billion in 2021 and contribute 13% to the country's GDP by 2025. Demand for residential properties has surged due to increased urbanization and raised household income. Driven by increased transparency and returns, there is a growth in private investment in this sector. The residential sector is expected to continue to demonstrate robust growth assisted by the rising penetration of housing finance and favorable tax incentives. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multi store and high-rise buildings. Investments in Real Estate Industry have grown significantly over the years and we have noticed a non-uniform pattern in terms of land pricing.

2.2. Need for House Price Prediction

The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. While our nation continues its growth trend and the construction industry lags behind demand, prices will continue to rise while interest rates bump upward. Securing investment property now with thorough due diligence and watching investment pay off over the next few years. Real property is not only a man's basic need, but it also represents a person's wealth and prestige today. Because their property values do not decline rapidly, investment in real estate generally seems to be profitable. Changes in the price of real estate can affect various investors in households, bankers, policy makers and many others. Investment in the real estate sector appears to be an attractive investment choice. Predicting the value of immovable property is therefore an important economic index.

III. Proposed System

Proposed system is a real time application used in real time. It is a browser-based application which is meant for real estate business. Proposed system is a generic application which can be accessed from different locations. System overcomes all the different drawbacks that we have in the existing system and come up with the solution. System major objective is to predict the price of a house. Proposed system uses machine learning techniques to predict price. We use supervised learning techniques for prediction. From online training data-sets downloaded, data collected from the sources such as Kaggle and data world websites. We used many parameters for house price prediction.

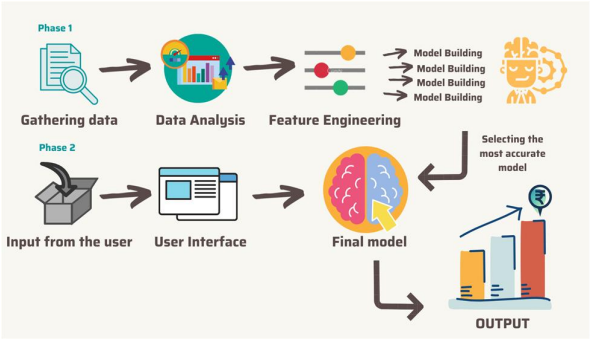

Figure 1: Architecture of Real Estate Price Prediction System

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

IV. METHODOLOGY

4.1. Data Collection

The first step for any kind of machine learning analysis is gathering the data which is valid. We need to pay attention to the source from where we take the data. For the purpose of this paper, we've relied on a database named "Real Estates prices in Metropolitan cities in India" from Kaggle. This dataset was available in CSV format and had approximately 1700 records over 40 independent parameters for each metropolitan city of India - Mumbai, Bangalore, Delhi, Chennai, Hyderabad, Kolkata. So there were in total around 10000*40 data values for training the models.

4.2. Data Analysis

A. Missing Null Values

Many of the variables had-1 values that had to be addressed. Based on what made the most sense, those values were filled out accordingly.

1. Imputation Technique: Values were filled with the median to prevent data from being skewed.
2. Dropping Technique: If the data for similar records was abundantly available, that complete record was dropped out of the data.

After eliminating all the -1 values and NaN records (null values), the data set was once again checked for null values and there were found none.

B. Removal of Outliers

An outlier is an extremely high or extremely low-value value in the data. Outliers are the observation points that are distant from other observations. For instance, if there is a house in the dataset with an area of 50 sq. feet for a price of 50,00,000 Rs. Such houses may exist on the market for various reasons, but they are not statistically meaningful. We want to make a price estimate based on the market average, and so we didn't take such outliers into consideration. As well as, most regression methods explicitly require outliers be removed from the dataset as they may significantly affect the results.

4.3. Feature Engineering-

A. Cardinality of parameters/variables

Variables within a dataset can be related for lots of reasons. For example:

- One variable could cause or depend on the values of another variable.
- One variable could be lightly associated with another variable.
- Two variables could depend on a third unknown variable.

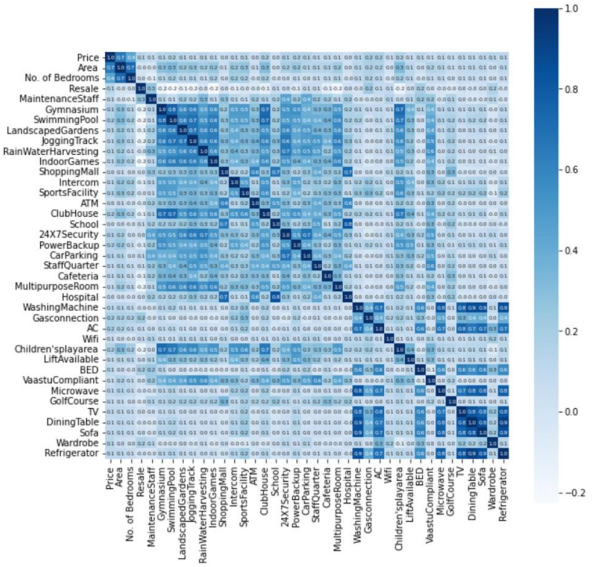

Figure 2: Heatmap for correlation

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

It can be useful in data analysis and modeling to better understand the relationships between variables. The statistical relationship between two variables is referred to as their correlation. Now if the correlation among two variables are strong, then remove one of the variables. In this project, no such parameter was removed due to strong correlation as each one of them had its individual effect on the price. The correlation was studied by plotting a heatmap.

B. Removal of Categorical Variables

There are many categorical variables as numeric variables. For eg. Name of the city and the Location. Categorical variables are strings which pose a threat to models. It is better to create dummy variables which are numeric constant for categorical variables which will help the models to operate on categorical variables. So the parameters city and location were changed into dummy variables - City no. and Location no.

4.4. Splitting Training and Testing Data

It is important to remember that if we use the same dataset for training and testing, the model may overfit. This means it will show excellent accuracy on the given dataset but completely fail for a new one. The reason is that when the dataset is split into train and test sets, there will not be enough data in the training dataset for the model to learn an effective mapping of inputs to outputs. There will also not be enough data in the test set to effectively evaluate the model performance. And so, we split the dataset into two parts and then use one for learning and other for testing. This way we simulate new data for our learning model and if there is an overfit, we can spot it. We split the dataset into a proportion of 80/20 which means 80% of the dataset will be used for learning and the remaining 20% for testing. So from around 10,000 data records, 8000 were used for training and 2000 for testing.

4.5. About Error and Accuracy

We need a defined method to check and decide which model will predict the prices most accurately. For this project, the method of least squares is used to find the best-fitting line for the observed data. R squared error is basically a squared difference between actual and predicted values. R-Squared is the ratio of the sum of squares regression (SSR) and the sum of squares total (SST). Sum of Squares Regression (SSR) represents the total variation of all the predicted values found on the regression line or plane from the mean value of all the values of response variables. The sum of squares total (SST) represents the total variation of actual values from the mean value of all the values of response variables. So, lesser this error, greater will be the model accuracy.

4.5. Model Building

A. Multivariate Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.
The visualized predictions showed an Error of 0.54 and so an accuracy of 46%.

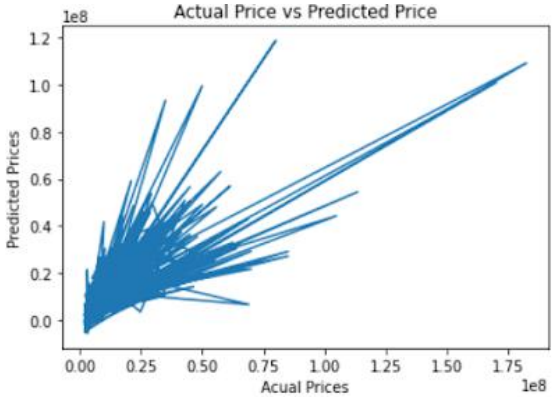

Fig 3: Visualization of Linear Regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

B. Decision Tree Regression

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. For instance in Fig 6, it shows data clustered at bottom left. The visualized predictions showed an Error of 0.19 and so an accuracy of 81%.

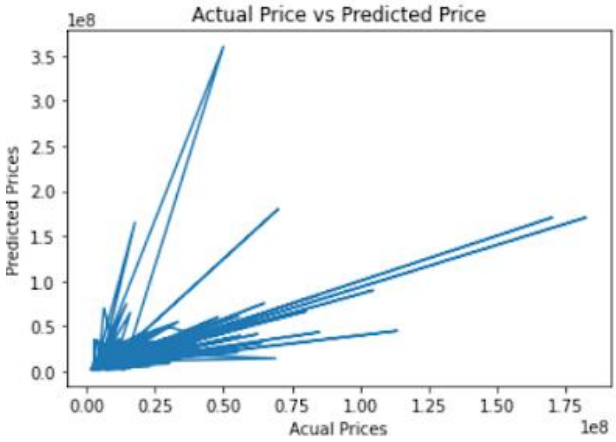

Fig 4: Visualization of Decision Tree Regression

C. XGBoost Regression

Gradient boosting refers to a class of ensemble machine learning algorithms. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. So the name is "gradient boosting".
The visualized predictions showed an Error of 0.55 and so an accuracy of 45%.

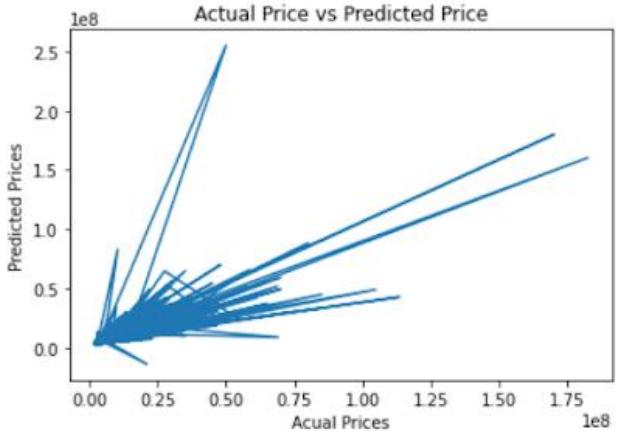

Fig 5: Visualization of XGBoost Regression

D. Random Forest Regression

Random forest is a Supervised Machine Learning Algorithm that is used widely *in* Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random forest Algorithm is that it can handle the data set containing continuous and categorical variables in the case of regression. The visualized predictions showed an Error of 0.56 and so an accuracy of 44%.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

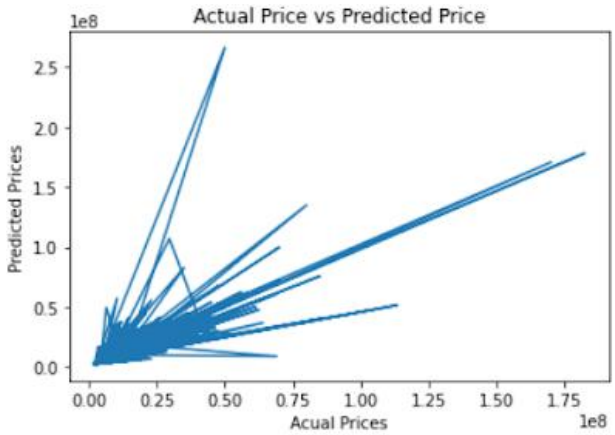

Fig 6: Visualization of Random Forest Regression

4.6. Choosing the Best Regressor

There is more than one way to do regression analysis. What we are looking for is the best prediction accuracy given our data. Following are the retained values for the corresponding accuracies of regressors. So the best regressor is the Decision Tree for this data set with an accuracy of 81%.

4.7. Creating an User Interface

For the final part, we created a Django API using the ATOM text editor for the user to be able to interact with the finalized model. We first created the views which would render the templates developed in raw HTML/CSS with a request-response mechanism. For the views to be visible, they were connected with corresponding urls. Every parameter which puts weight on the price prediction was inserted into the model. Now when the server runs, the user needs to enter values for every parameter and then the value for a similar kind of house is predicted by the model in the backend and displayed.

V. RESULT

As it can be seen through the descriptions of models, the decision tree regressor provided the best results with an accuracy of 81%, so we decided to use this model for production of user interface. Price prediction in production pretty much works the same as in our test code except there is no need to calculate r squared error and switch models anymore. At this point of accuracy, we can offer fair price predictions. We can compare the actual price of a house with our predicted price and observe the deviation.

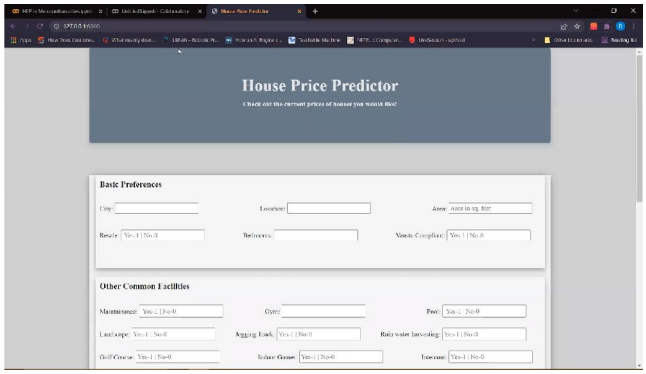

Figure 7: Demo data entry

We have defined several models with various features and various model complexities. There is a need to use a mix of these models. A linear model gives a high bias (under fit) whereas a high model complexity-based model gives a high variance (overfit). This is what ensemble learning is. However, redundancies in dataset as well as model training leads to the poor outcomes of ensemble learning. That is why the outcomes of ensemble learning algorithms like XGBoost and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

Random Forest were poor. It showcases that for the kaggle dataset that we have used, a single model is more efficient at predicting the prices properly. Here is how the results look from a user perspective for a dummy data case.

Figure 8: Demo Result

VI. CONCLUSION

This is a business-oriented application useful for real estate business. The system predicts the price of a house based on some parameters. System uses data science algorithms for price prediction. System is a real time application useful for real estate business. System predicts house prices exactly, so that buyers or sellers will not get lost. The outcome of this study can be used in annual revision of guideline value of land which may add more revenue to the State Government while land transaction is made. This study will support the policy makers to relook the movement of the identified factors to have control on rise in the land price and stabilize it. Since there is a greater need for good long-term data analysis about land price, general land market behavior and spatial development, the results produced in this research may be of great use for Government and non-Government agencies which are involved in land administration.

VII. FUTURE SCOPE

In the future, we can integrate this model with various real estate websites to expand it to a higher level. We can also use this to present a comparative study of the systems' predicted price and the price from real estate websites such as Housing.com for the same user input.. Right now, the dataset only includes data of Metropolitan cities of India like Mumbai, Bangalore, Chennai, Delhi, Kolkata and Hyderabad. Expanding it to other cities and states of India is the future goal. To make the system even more informative and user-friendly, we will be including GoogleMaps. This will show the neighborhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors increases the valuation of real estate property.

REFERENCES

[1]. House Price Prediction Forecasting And Recommendation System Using Machine Learning by Ashutosh Sharma1, Pranav Sonawale2, Deeksha Ghonasgi3, Shreya Patankar4

[2]. House Price Forecasting Using Machine Learning *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020*

[3]. Byeonghwa Park, Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction, Volume 42, Pages 2928-2934

[4]. Real Estate Price Prediction with Regression and Classification, CS 229 Autumn 2016

[5]. Real Estate Price Prediction Smith Dabreo, Shaleel Rodriguez, Valiant Rodrigues, Parshvi Shah Student, Fr. Conceicao Rodrigues College of Engineering, Mumbai. Assistant Professor, Fr. Conceicao Rodrigues College of Engineering, Mumbai.

[6]. Real Estate Price Prediction Using Machine Learning , Aswin Sivam Ravikumar, School of Computing, National College of Ireland

[7]. [Literature Review on Real Estate Value Prediction Using Machine Learning, Akshay Babu, Dr. Sanjana S. Chandran

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

[8]. Kaggle.com
[9]. ResearchGate.net
[10]. ieee.org
[11]. GoogleScholar
[12]. Youtube
[13]. irjet.net
[14]. Coursera

Abbreviations

- ML: Machine Learning
- AI: Artificial Intelligence
- XGBoost: Extreme Gradient Boosting