EasyChair Preprint

№ 10283

House Price Prediction System Using Machine Learning and Data Science

Satyam Sobhraj, S.P.S Chauhan and Anas Ghani Ur Rahman

May 29, 2023

House Price Prediction System Using Machine Learning and Data Science

Satyam Sobhraj
B. Tech Department CSE
Galgotias University
Gautam Budh Nagar, India
Satyam_sobhraj.scsebtech@galgotiasuniversity.edu.in

S.P.S. Chauhan, Professor
School of Computing Science and Engineering
Galgotias University
Greater Noida
sps.chauhan@galgotiasuniversity.edu.in

Anas Ghani Ur Rahman
B. Tech Department CSE
Galgotias University
Gautam Budh Nagar, India
anas_ghani.scsebtech@galgotiasuniversity.edu.in

***Abstract* – Real estate is the area with the least transparency in our economy. Housing costs fluctuate every day and are sometimes artificially exaggerated. Using real factors to forecast real estate values is the main goal of our research project. In this case, we strive to base our reviews on all the important aspects that go into pricing. We use several different regression algorithms in this strategy. Rather than solely relying on one technique to determine our results, we instead use the weighted averages of several different techniques that produce the most accurate results. The results showed that this strategy provides the lowest error and maximum accuracy compared to using separate methods. We also recommend using Google Maps to get accurate real-world reviews by using up-to-date local data. Homes in Bengaluru most upscale and cheap neighborhoods have very different prices, as is obvious.**

Keywords: **House Price Prediction System, Bengaluru Dataset, SVR, Hedonic Price Model, Forest Regression.**

I. INTRODUCTION

A person's basic need, a home, varies in price depending on the neighborhood and features such as parking spaces. One of the biggest and most important decisions a family will ever make is buying a home as families invest all of their money and eventually cover it with loans. Machine learning modeling is the process of using data that computers have learned to generate new data.

Also, consumers are financially able to make a large investment, and a plentiful supply of housing in the country is a sign that the construction industry is doing well.

These criteria are used to measure our performance:

- Develop a reliable pricing prediction model.
- Verify the model's ability to make accurate predictions.
- Determine the key home price characteristics that support the model's prediction ability.

II. LITERATURE REVIEW

The idea is described in this article. It aims to use cutting-edge technologies to save time and money for real estate sellers and buyers. VR4RE, one of Blue Mind Software's innovative projects, is currently in an advanced stage. This study also illustrates the history of internal technology initiatives to provide suitable 3D and VR presentation tools for real estate. The virtual world.

To develop a system for predicting real estate prices based on certain input characteristics, we propose using machine learning and artificial intelligence techniques. By using a few input characteristics and predicting an accurate and reasonable price, this algorithm is used by classified websites to directly estimate the value of new properties that are being offered and avoid inaccurate home valuations. The proof-of-concept (POC) study described here can be viewed as an evaluation.

III. TOOLKIT

Jupyter Notebook is a component of the distribution of Python and other programs known as Anaconda. You can use the Windows start menu, command line, or Anaconda Navigator to open a Jupyter Notebook with Anaconda. Anaconda Server channels can be used to share and store Jupyter Notebook files. A web program called a Jupyter Notebook can store and run Python or R code, text, graphics, and other types of media in rich text files.

Software Used:

- Anaconda
- Jupyter Notebook
- Html CSS JavaScript
- Python Flask Server
- Visual Studio Code

Library Used:

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit Learn
- XG BOOST

The source of the dataset is the Bengaluru House Dataset which is accessible on Kaggle. It consists of a target variable and several house prediction variables. After downloading our dataset into pandas, perform data cleansing using a variety of methods.

area_type	availability	location	size	society	total_sqft	bath	balcony	price
Super built	19-Dec	Electronic	2 BHK	Coomee	1056	2	1	39.07
Plot Area	Ready To I	Chikka Tir	4 Bedroom	Theanmp	2600	5	3	120
Built-up A	Ready To I	Uttarahalli	3 BHK		1440	2	3	62
Super built	Ready To I	Lingadheer	3 BHK	Soiewre	1521	3	1	95
Super built	Ready To I	Kothanur	2 BHK		1200	2	1	51
Super built	Ready To I	Whitefield	2 BHK	DuenaTa	1170	2	1	38
Super built	18-May	Old Airpor	4 BHK	Jaades	2732	4		204
Super built	Ready To I	Rajaji Nag	4 BHK	Brway G	3300	4		600
Super built	Ready To I	Marathahi	3 BHK		1310	3	1	63.25
Plot Area	Ready To I	Gandhi Ba	6 Bedroom		1020	6		370
Super built	18-Feb	Whitefield	3 BHK		1800	2	2	70
Plot Area	Ready To I	Whitefield	4 Bedroom	Prrry M	2785	5	3	295
Super built	Ready To I	7th Phase	2 BHK	Shncyes	1000	2	1	38
Built-up A	Ready To I	Gottigere	2 BHK		1100	2	2	40
Plot Area	Ready To I	Sarjapur	3 Bedroom	Skityer	2250	3	2	148
Super built	Ready To I	Mysore Rc	2 BHK	PrntaEn	1175	2	2	73.5
Super built	Ready To I	Bisuvanah	3 BHK	Prityel	1180	3	2	48
Super built	Ready To I	Raja Rajes	3 BHK	GrrvaGr	1540	3	3	60
Super built	Ready To I	Ramakrish	3 BHK	PeBayle	2770	4	2	290
Super built	Ready To I	Manayata	2 BHK		1100	2	2	48
Built-up A	Ready To I	Kengeri	1 BHK		600	1	1	15
Super built	19-Dec	Binny Pete	3 BHK	She 2rk	1755	3	1	122
Plot Area	Ready To I	Thanisandi	4 Bedroom	Soitya	2800	5	2	380
Super built	Ready To I	Bellandur	3 BHK		1767	3	1	103
Super built	18-Nov	Thanisanc	1 RK	Bhe 2ko	510	1	0	25.25
Super built	18-May	Mangamm	3 BHK		1250	3	2	56

FIG. 1. *BENGALURU_HOUSE_DATASETS*

First, visit Kaggle.com, then click the download dataset button to download the dataset as a CSV file. Therefore, this dataset has the following properties, all of which are referred to as independent variables: type of liberty, location, size, square foot, etc. Price is a dependent variable that I am trying to make predictions on.

Brief of Working –
First, install the Anaconda distribution, which includes the data science technology stack from Jupyter Notebook, Pandas, Python, Escalon, and others. You need to import some important libraries into Jupyter Notebook. You should be reading the CSV file I just had opened in a panda's data frame once the libraries have been imported.

```
df1 = pd.read_csv("bengaluru_house_prices.csv")
df1.head()
```

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

FIG. 2. *READY PANDA SETS TO READ CSV FILE*

The number of rows and columns in my data is shown when I run Shape. The data set I have is decent enough at 13 000 rows. Then, for each of these region-type categories, I create a report of the data sample's data set. To do this, first, organize your data frame by region type, and then aggregate the count.

Outlier Detection and Removal: Outliers are data points that either contain errors in the data or, in some cases, contain no data at all but only show the most extreme variation. It makes sense to eliminate them because they are valid.

We use a variety of methods to identify and eliminate outliers, including the standard deviation between them. For example, if you have a two-bedroom apartment, you can use a simple or domain-specific one as an example. Its total size cannot exceed 500 square meters.

How many square meters are bedrooms usually? I can tell you it's about 300 square meters. So, if there is ever a situation where your house is saying 600 square meters but has eight beds in total. Therefore, you can confidently eliminate these as they appear to be clear data errors, anomalies, or outliers. These data points are all eliminated.

We have around 13000 in our data frame so you need to remove it as this is what I will be doing. You must do df6, and that's what I want to argue against here. If you want to exclude specific outliers from all these rows, filter them based on your criteria and the df6 dot shape. Your rows now have a total of 1,200,502, so this eliminates outliers.

For my cross-validation shuffle, I'm making a split shuffle. Shuffle split will randomly distribute my samples so that they are distributed equally throughout the folds. It does not only focus on one region. I consistently achieve a score of more than 80% while using cross-validation.

Other regression methods include lasso regression, decision tree regression, and others. Regression algorithms come in many different varieties. We employ a technique called grid search CV for that. Escalon has a decent API that allows you to run your model with various regressors and parameters. So, I can also use grid search CV to find my best model. I'll provide the inputs x and y, and this should tell me which algorithm is the best.

Once everything is ready, we will be our UI application. Our user interface will be a simple HTML CSS JavaScript website. Your UI organizational structure and your application elements will be contained in HTML. Cascaded Stylesheets, or CSS, contain the stylesheets that determine an application's color scheme and user interface. JavaScript contains dynamic code that uses HTTP requests to our backend to collect the data. As a result, JavaScript acts more like a server communicator, delivering the data to you while executing dynamic code.

IV. PROBLEM STATEMENT

Python is a programming language commonly utilized in computational science and mathematics; A set of software for mathematics, science, and engineering is called SciPy. A modelling and data analysis library is called Pandas. A robust interactive shell, Python makes it simple to record and edit work sessions, offers visualization, and supports parallel computing. The Software Carpentry Course runs boot camps and offers free educational resources to teach the fundamentals

of scientific computing.

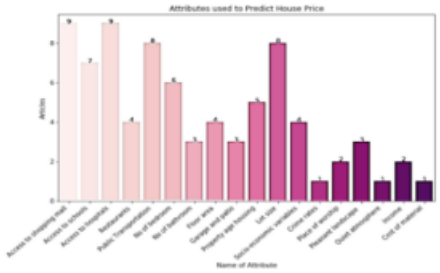

FIG. 3. ASPECTS USED TO PROJECT HOUSE VALUE

V. RESOURCE ALGORITHM

Data Exploration: The first phase of data analysis, known as data exploration, typically involves summarizing key aspects of a data set, such as B. its size, accuracy, the early patterns it uncovers, and other elements. Python can also be used to run statistical programs, which data analysts typically do. However, Python is a more complex statistics program. When analyzing information collected from multiple sources and stored in data warehouses, an organization must be aware of the number of instances in a dataset, the variables it contains, the set of missing values, and the general hypotheses the data is likely to support.

Data Visualization: In information visualization, data and information are displayed graphically. Information visualization devices give an available w Devices for information visualization make it possible to identify and find informational patterns, outliers, and designs. ay to watch and get patterns, exceptions, and designs in information utilizing visual components such as charts, charts, and maps. To analyze gigantic sums of information and make data-driven choices within the world of Huge Information, information visualization apparatuses and innovations are pivotal.

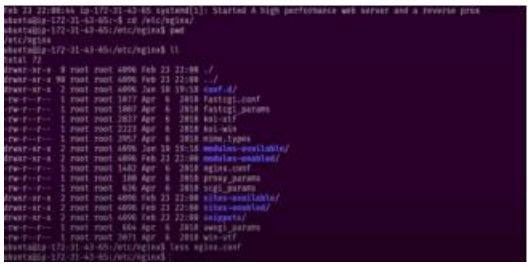

FIG. 4. WEB SERVER WITH HIGH PERFORMANCE

***Data Selection*:** Data selection is the process of deciding what data source, type, and equipment for data collection will be most effective. The data is selected after the actual data collection process. This concept distinguishes between active/interactive data selection (monitoring activities/events using collected data or performing secondary data analysis) and selective data reporting. (Selective removal of data not supporting research premise). How pertinent data is chosen for a research study can jeopardize data integrity. Finding the appropriate data kinds, sources, and tools is the main objective of data selection in order to provide researchers the best chance of successfully responding to research questions. Visualization tools make it possible to observe and understand patterns, exceptions, and designs in data.

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000

FIG. 5. AVERAGE PRICE PER SQ/FT

Data Transformation: The log transformation can be used to reduce the skewness of highly skew distributions. Assumptions of inferential statistics can thereby be satisfied, and the patterns in the data can thereby become clearer. The top panel makes it difficult to discern a pattern, in contrast to the bottom panel, which makes the strong relationship clear. Comparing geometric means essentially compares the means of data that have been logarithmically transformed. This happens because, as shown below, the geometric mean of logarithmically transformed numbers is equal to the antilogarithm of the arithmetic mean.

Data Preprocessing: This transformation takes place before the data is sent to the algorithm. When used, it converts dirty data into clean information. With this information mining technique, illogical material is converted into a logical framework. Add disorganized information in a logical order. The result of the data pre-processing is the final data set, which is used for preparation and as a basis for tests.

Data Collection: Learning about different variables is a thorough process that feeds into data collection. It promotes research, generates many theories, and evaluates results. To facilitate social interactions and estimate data on specific characteristics within the existing framework, data collection will be carried out. This enables the discussion of relevant questions and the evaluation of the results.

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

FIG. 6. ROOM AVAILABILITY AND RATES

Data Cleaning: To maximize the value of the data, errors are detected and eliminated during data cleaning. Data cleaning is carried out using data processing technologies. In this way, erroneous data records can be identified and changed in a data record, a table, or a database. It replaces the missing data with encrypted information. Before publication, the information is

checked for completeness and correctness.

Machine Learning Model: The housing demand valuation paradigm can be divided into two categories: the traditional technique and the elevated valuation method according to. Advanced scoring techniques include the hedonic pricing tool, artificial neural network (ANN), and spatial analysis framework. Multiple regression techniques and a stepwise regression procedure are part of the conventional scoring scheme. Considering the available models, choosing the best model to predict house price is crucial. One of the most common models used in this research area is regression analysis. Another common technique for forecasting house prices is Support Vector Regression (SVR).

Random Forest Regression: A random forest is an ensemble methodology that can handle both regression and classification problems by using many decision trees and a technique called bootstrap aggregation, also known as bagging. The main concept here is to combine multiple decision trees to achieve the result, rather than relying on individual decision trees. The process of training each decision tree in the random forest technique with a different data sample, where sampling is done with replacement.

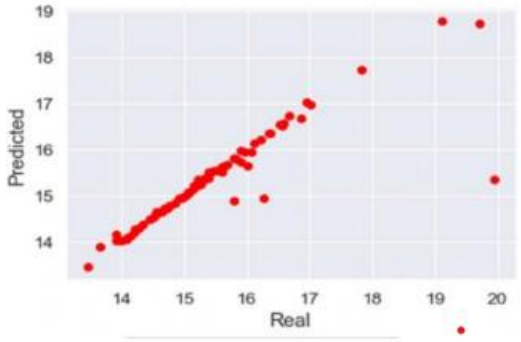

FIG. 4. *REAL VS. PREDICTED*

Artificial Neural Network: These Network was consistently chosen any time a nonlinear characteristic present. Since housing costs are This model should be used as a spatial factor in house price estimation study, which is likewise non-linear. However, this system's performance is severely limited. ANN can mimic complicated nonlinear interactions because house price estimates contain several nonlinear variables. It replaces the missing data with encrypted information. Before publication, the information is checked for completeness and correctness.

Support Vector Regression (SVR): Support vector regression, a powerful subset of supervised learning, is a three-layer neural network that uses SVM as its basis for prediction. A portion of the training data is used to create the model. Supporting vector regression has several advantages, including the ability to deal with nonlinear results, the fact that It has the capacity to get over small-sample learning issues, but only provides one feasible optimal solution. This model can produce market forecasts for a range of markets, including the real estate market, which overcomes the challenges of non-linear regression and small-sample learning.

Multiple Linear Regression: Regression analysis is a method of determining how variables are related to each other. How closely the variables are correlated can be determined using the regression equation or the correlation coefficient. Several regression models can be used to identify the features that are most important in explaining the dependent variable. Multiple regression analysis is used to collect the data from independent and dependent variables, which also allows for some price estimates. The power of the multiple regression model may be demonstrated while analyzing the significance of the relationship between the dependent and independent variables. Multiple regression modeling was used to explain changes in an independent variable using a dependent variable.

Hedonic Price Model: Compared to the typical good consumption, the home market is a bit different. Because it shows resilience, adaptability, and spatial fixation, the housing market is said to be independent. The hedonic method is therefore preferred for an accurate estimation of market disparity.

VI. EXPERIMENT

Python will be our chosen programming language, Madrid Pandas Data Cleaning For data visualization, switch to SK Learn Python Flask for a backend server, HTML, CSS, and JavaScript for our website, and learn for model creation.

2.66GHz, Core-to-Dual, Core-to-Quad, and i3 processors, 3GB, and 4GB RAM, and these specs; Visual Studio, Machine Learning, a 24-port network switch, and a Python Flask server.

RESULTS

Using linear regression, the best performance for this dataset was demonstrated and deployable. Because they are so far behind, it is not recommended to use Random Forest Regressor or XG Boost Regressor for further deployment.

The expected vs. real price chart below illustrates the accuracy of the forecasts:

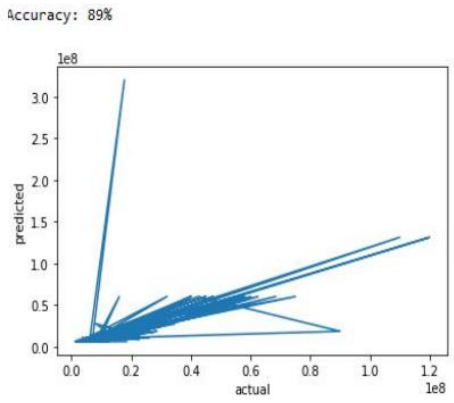

FIG. 7. *GRAPH OF PREDICTED VS. REALIZED PRICE BASED ON THE DATASETS*

VII. CONCLUSION

In this paper, we have reviewed, discussed, and examined current research on the key variables affecting house prices, as well as data mining techniques. Homes in more enviable regions, including those close to a mall or other amenities, typically cost more than those in less enviable rural settings. Investors, home buyers, and builders could select an affordable home price as well as the actual price of a property with the help of the accurate prediction model. The subject of this study was the characteristics that previous researchers used to predict a house price with different prediction algorithms. The results of the survey show that SVR, ANN, and XG Boost can predict home values.

A predictive model is created using the machine learning decision tree method to project potential sale values for any property. The dataset was expanded to include additional variables, such as air quality and crime rates, to improve the precision of the pricing predictions. Our approach distinguishes out from the competition since these characteristics are mostly missing from the datasets of these other forecasting systems. Given that they influence people's purchasing decisions, you should consider these aspects when forecasting house prices. The UI and the learned model are brought together using the Flask Framework. The method has an 89% accuracy rate in forecasting house prices.

REFERENCES

[1]. A S. Temür, M. Akgün, and G. Temür, “Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models,” J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, Doi: 10.3846/jbem.2019.10190.

[2]. A. Ebekozien, A. R. Abdul-Aziz, and M. Jaafar, “Housing finance inaccessibility for low-income earners in Malaysia: Factors and solutions,” Habitat Int., vol. 87, no. April, pp. 27–35, 2019, doi: 10.1016/j.habitatint.2019.03.009.

[3]. A. Jafari and R. Akhavian, “Driving forces for the US residential housing price: a predictive analysis,” Built Environ. Proj. Asset Manag., vol. 9, no. 4, pp. 515–529, 2019, Doi: 10.1108/BEPAM-07-2018- 0100.

[4]. Choong Wei Cheng, “Statistical Analysis of Housing Prices in Petaling,” University Tunku Abdul Rahman, 2018.

[5]. Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." 2011 National Conference on Innovations in Emerging Technology. IEEE, 2011.

[6]. Manjula, R., et al. "Real estate value prediction using multivariate regression models." Materials Science and Engineering Conference Series. Vol. 263. No. 4. 2017.

[7]. A. Varma et al., “House Price Prediction Using Machine Learning and Neural Networks,” 2018 Second International Conference on Inventive Communication and Computational Technologies, pp. 1936–1939, 1936.

[8]. Arietta, Sean M., et al. "City forensics: Using visual elements to predict non-visual city attributes." IEEE Transactions on Visualization and computer graphics 20.12 (2014): 2624-2633.

[9]. Yu, H., and J. Wu. "Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5." (2016).

[10]. Sinha, Anurag & Rammish, Md. (2021). HOUSE COST ESTIMATION OF BANGALORE REGION USING FEATURE SELECTION ALGORITHM OF MACHINE LEARNING.

[11]. Lee, Min-feng & Chen, Guey-shya & Lin, Shao-pin & Wang, Wei-Jie. (2022). A Data Mining Study on House Price in Central Regions of Taiwan Using Education Categorical Data, Environmental Indicators, and House Features Data. Sustainability. 14. 6433. 10.3390/su14116433.

[12]. Sampath Kumar, V., Sanithi, M. H., & Vanjinathan, J. (2015).