IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 11, November 2024| DOI: 10.15680/IJIRCCE.2024.1211060

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Approaches for House Price Prediction

Dr.A.Ram Kumar, A.Vamsi Kiran, B.Sravani, B.Padma Naga Durga, B.Shiva Krishna, B.Yashwanth

Professor, NSRIT, Vishakhapatnam, India

Student, Department of CSE (AIML), NSRIT, Vishakhapatnam, India

ABSTRACT: This paper seeks to explore which machine learning techniques would predict house prices because forecasting house prices has always been the most crucial challenge that real estate analysis has always been facing. The problem now stands on how one will use the algorithms of machine learning effectively for property valuation according to features like location, size, amenities, and so on. It compares the performances of different regression models: Linear Regression, Decision Trees, Random Forests, SVM, and Gradient Boosting Machines (GBM). The models were trained on a large sample of history property sales with features selected using feature importance analysis. From the above results, it is clear that the ensemble models such as Random Forest and Gradient Boosting are showing higher accuracy compared with the classical models like Linear Regression. The performance of the models is evaluated based on some metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2. It could be concluded that machine learning approaches, especially ensemble methods, could be of significant help in enhancing house price predictions due to its great value to investors, real estate analysts, and developers.

KEYWORDS: Machine Learning, House Price Prediction, Regression Models, Random Forest, Gradient Boosting, Feature Selection, Real Estate Analytics, Predictive Modeling.

I. INTRODUCTION

In informing buyer, seller, and investor decisions, house price prediction plays an important role in the real estate market. Traditionally, estimates of prices were based on expert judgment and very basic statistical models. They often failed to capture the vast complexities of factors that affect housing prices or sometimes failed to fully factor in location or provision of amenities and subsequent economic trends, all resulting from the extensive heterogeneity of houses. However, by virtue of machine learning, such sophisticated and accurate predictive models have been created. Machine learning has great many advantages over the traditional methods in that large, high-dimensional datasets can be processed and high-order, intricate, non-linear patterns between variables may be revealed. Techniques such as linear regression, decision trees, random forests, support vector machines, and neural networks have been successfully applied to the house price prediction problem with improved predictive accuracy and flexibility. The application of various ML techniques in predicting house prices and outlining strengths, limitations, and practical considerations will be discussed in this paper

II. METHODOLOGY

This research employed regression model to analyze Boston housing datasets in order to predict the prices of houses based on the features that are in the datasets. The fundamental step taken for the implementation include data collection, data exploration which was used to understand the datasets and identify features in the dataset; data pre-processing stage which was used to clean the dataset so as to make it suitable for model development. Afterwards the model was developed using the proposed random forest algorithm.

2.2. Data Pre-Processing

Data used for model training and testing must be analyzed properly so that the models learn to recognize patterns quickly. Numerical values were normalized, and categorical values were encoded separately. Once all data is explored and the most suitable features are prepared by using a heatmap, one has to pre-process those features. Normally, datasets have feature sets of different scales; it's why model performance would be lowered down. Scaling was done by

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the function Standard Scaler contained in the Scikit-learn module of Python. This function assumes data to be normally distributed and scales it to have a mean of 0 and a standard deviation of 1. Then a linear regression plot, called regplot was drawn to see the correlation between features and the variable to be forecasted, MEDV.

2.3. Model Development

The proposed model was built using the random forest algorithm. The random forest was implemented using the RandomForestClassifier available in Python Scikit-learn (sklearn) machine learning library. Random Forest is a popular supervised classification and regression machine learning technique. It employs the concept of ensemble learning to solve complex problems by incorporating several classifiers to improve the model's accuracy. Random Forest is a classifier that averages the outcomes of multiple decision trees applied to various subsets of a dataset to improve the dataset's predictive accuracy. Rather than relying on a single decision tree, the random forest uses the projections from each tree to determine the final performance based on the majority of votes. The algorithm for the random forest is

- Create an n-sample random bootstrap sample by selecting n samples at random from the training set.
- At each node, build a decision tree using the bootstrap sample: a. Select d functions at random without replacement. b. Split the node using the attribute that offers the optimal split according to the objective function (e.g., optimizing knowledge gain).
- Repeat steps 1-2 k times.
- Combine predictions from each tree using a majority vote to determine the class name.

The n-estimators parameter in the RandomForestClassifier is set to 500 to create 500 trees, enhancing accuracy and avoiding overfitting. Although more trees improve accuracy, they also increase training time. The bootstrap parameter is set to True to introduce variation into random forest subsets. Efficiency is improved by iterating the model several times and adding parameters during initialization.

III. RESULTS AND DISCUSSION

3.1. Results of the Data Exploration Process To understand the dataset better, data exploration was carried out. Fig. 1 show the distribution of the data in each of the features in the datasets. It shows the total count of the data, the mean, the standard deviation, the minimum value, 25%,50%,75% and the maximum value. From this, two data columns show interesting summaries. ZN (proportion of suburban property zoned for lots above 25,000 sq. ft.), with 0 representing the 25th and 50th percentiles. Second, with 0 for the 25th, 50th, and 75th percentiles, CHAS: Charles River dummy vector (1 if tract borders river; 0 otherwise). Since both variables are conditional + categorical, these summaries make sense.

```
print(data.describe())
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

	AGE	DIS	RAD	TAX	PTRATIO	B	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	
std	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	
min	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	

	LSTAT	MEDV
count	506.000000	506.000000
mean	12.653063	22.532806
std	7.141062	9.197104
min	1.730000	5.000000
25%	6.950000	17.025000
50%	11.360000	21.200000
75%	16.955000	25.000000
max	37.970000	50.000000

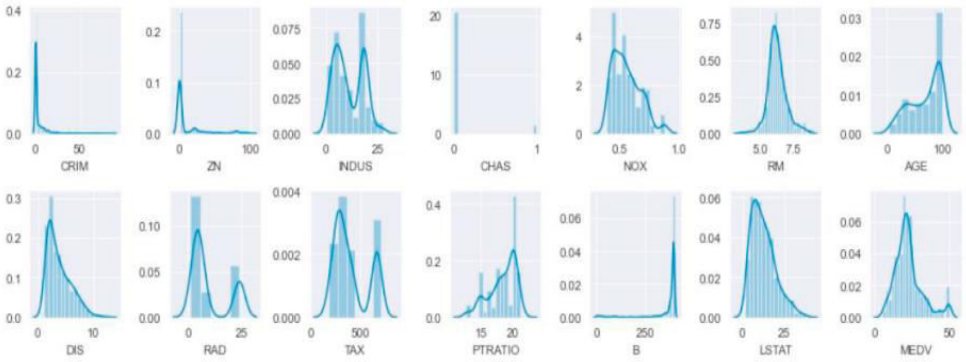
Fig. 1 Data Distribution

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The next exploration that was carried out is the generate the histogram of the data as shown in Fig. 2.



The histogram reveals that the distributions of columns CRIM, ZN, and B are heavily distorted. Except for CHAS, MEDV has a regular distribution, while other columns show normal or bimodal distributions. The final stage of data exploration involves the correlation matrix, which displays the coefficients of correlation between variables. To visualize these correlations, a seaborn heatmap is used. Heatmaps use colors to represent data values, guiding viewers to the most important areas. Seaborn heatmaps are visually appealing and effectively convey data messages, making them a popular choice for data analysts and scientists. The heatmap generated is shown in Fig. 3.

Fig. 3 Heatmap

3.2. Testing the Proposed Model

After training the model with the training dataset, the next phase of the study is to test the predictive prowess of the model. This was achieved by removing the actual prices from the dataset and simulating the model to predict the house prices. The predicted and actual house prices were then combined together and the difference were computed. These are captured in Table 1. The results obtained revealed that, though exact prices were not predicted in some cases, the difference between the predicted value and the actual value were in the range of ±5.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 1. Actual Vs Predicted House Prices

S/N	Actual Value	Predicted Value	Difference
1	27.967	26.700	1.267
2	14.755	13.400	1.355
3	21.137	20.600	0.537
4	40.790	43.100	-2.31
5	9.700	11.500	-1.8
6	25.928	29.400	-3.472
7	30.926	33.100	-2.174
8	32.652	33.200	-0.548
9	10.306	11.000	-0.694
10	14.755	13.400	1.355
11	21.137	20.600	0.537
12	31.737	35.100	-3.263
13	23.101	21.000	2.101
14	19.989	18.900	1.089
15	21.768	18.500	3.268
16	21.533	24.300	-2.767
17	19.067	14.100	4.967
18	22.944	24.800	-1.856
19	21.341	21.100	0.241
20	16.939	18.000	-1.061

3.3. Performance Evaluation of the Proposed Model

After training and testing the model, performance evaluation metrics were used to get the performance of the model. These are the Mean Absolute Error (MAE), R^2 or Coefficient of Determination and the Root Mean Square Error (RMSE).After getting the performance of the model a scatter plot was generated to show the linear regression between the actual value and the predicted value from the model.

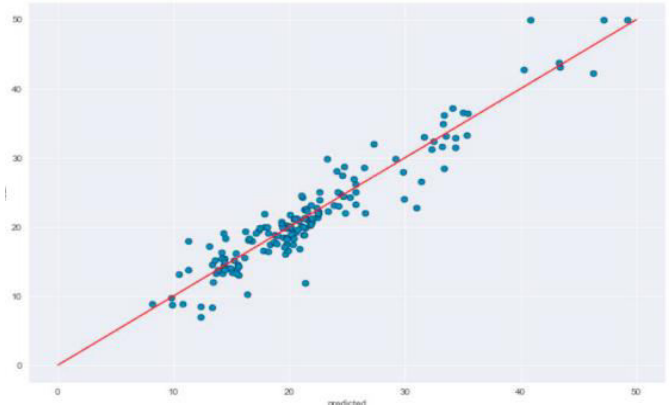


Fig. 4 Scatter Plot Real vs Predicted.

IV. CONCLUSION

Every year, house prices increase, making it important to have a way to predict future prices. Landowners, property valuers, and policymakers can use house price predictions to assess property value and determine fair sale prices. This also helps potential buyers decide when the right time to buy a home is. While the main factors that affect house prices are physical condition, style, and location, there are many other specific factors that can vary by region. Therefore, an effective prediction model must take these factors into account. This study highlights the effectiveness of the Random

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Forest machine learning technique in predicting house prices, using the Boston Housing dataset. A comparison of the predicted and actual prices, shown in Table 1, demonstrated that the model's predictions were within a ±5 price difference, suggesting that it can accurately predict house prices. Additionally, other machine learning models, particularly deep learning models, could also be explored to further improve house price predictions.

REFERENCES

[1] Garriga, C., Hedlund, A., Tang, Y., & Wang, P. (2020). Regional Science and Urban Economics Rural-urban migration and house prices in China. Regional Science and Urban Economics, March, 103613. https://doi.org/10.1016/j.regsciurbeco.2020.103613

[2] Wang, X., Li, K., & Wu, J. (2020). House price index based on online listing information : The case of China. Journal of Housing Economics, 50(May 2018), 101715. https://doi.org/10.1016/j.jhe.2020.101715

[3] Zhou, T., Clapp, J. M., & Lu-andrews, R. (2021). Is the behavior of sellers with expected gains and losses relevant to cycles in house prices ? Journal of Housing Economics, 52(May 2020), 101750. https://doi.org/10.1016/j.jhe.2021.101750

[4] Truong, Q., Nguyen, M., Dang, H., Mei, B., Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174(2019), 433– 442. https://doi.org/10.1016/j.procs.2020.06.111

[5] Lu, S., Li, Z., Qin, Z., Yang, X., Siow, R., & Goh, M. (2017). A Hybrid Regression Technique for House Prices Prediction. December. https://doi.org/10.1109/IEEM.2017.8289904

[6] Greenaway-mcgrevy, R., & Sorensen, K. (2021). A Time-Varying Hedonic Approach to quantifying the effects of loss aversion on house prices. Economic Modelling, 99(March), 105491. https://doi.org/10.1016/j.econmod.2021.03.010

[7] Malang, C. S., Java, E., & Febrita, R. E. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization. International Journal of Advanced Computer Science and Applications, 8(10), 323–326.

[8] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2020). Land Use Policy Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, July, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

[9] Filip F.G., Zamfirescu CB., Ciurea C. (2017) Collaboration and Decision-Making in Context. In: ComputerSupported Collaborative Decision-Making. Automation, Collaboration, & E-Services, vol 4. Springer, Cham. https://doi.org/10.1007/978-3-319-47221-8_1

[10]Kauko, T.; d'Amato, M. Introduction: Suitability Issues in Mass Appraisal Methodology. In *Mass Appraisal Methods*; Blackwell Publishing Ltd.: Oxford, UK, 2008; pp. 1–24. [Google Scholar] [CrossRef]

[11] Grover, R. Mass valuations. *J. Prop. Investig. Financ.* 2016, *34*, 191–204. [Google Scholar] [CrossRef]

[12] IAAO, International Association of Assessing Officers. *Standard on Mass Appraisal of Real Property (2017)*; International Association of Assessing Officers: Kansas City, MI, USA, 2019; p. 22. Availableonline: https://www.iaao.org/media/standards/StandardOnMassAppraisal.pdf (accessed on 22 August 2022).

[13] Wang, D.; Li, V.J. Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. *Sustainability* 2019, *11*, 7006. [Google Scholar] [CrossRef] [Green Version]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details