Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 806–813

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management

(ITQM 2020 & 2021)

House Price Prediction using Random Forest Machine Learning Technique

Abigail Bola Adetunji[a], Oluwatobi Noah Akande[*b], Funmilola Alaba Ajala[a], Ololade Oyewo[a], Yetunde Faith Akande[c], Gbenle Oluwadara[b]

[a]*Department of Computer Science, Faculty of Computing and Informatics, Ladoke Akintola University of Technology, Nigeria*
[b]*Computer Science Department, College of Pure and Applied Sciences, Landmark University, Nigeria*
[c]*Accounting Department, College of Business Sciences, Landmark University, Nigeria*

Abstract

Predicting a price variance rather than a specific value is more realistic and attractive in many real-world applications. Price prediction can be thought of as a classification issue in this situation. However, the House Price Index (HPI) is a common tool for estimating the inconsistencies of house prices. Since housing prices are closely correlated with other factors such as location, city, and population, predicting individual housing prices needs information other than HPI. The HPI is a repeat-sale index that tracks average price shifts in repeat transactions or refinancing of the same assets. Therefore, HPI is ineffective at predicting the price of a single house because it is a rough predictor based on all transactions. This study explores the use of Random Forest machine learning technique for house price prediction. UCI Machine learning repository Boston housing dataset with 506 entries and 14 features were used to evaluate the performance of the proposed prediction model. A comparison of the predicted and actual prices predicted revealed that the model had an acceptable predicted value when compared to the actual values with an error margin of ±5.

Keywords Sales forecasting; House Price Prediction; Machine Learning; Random Forest Algorithm

1. Introduction

Housing is one of the integral components that can be used to measure how successful the economy of a nation is. As the economy increases, people tend to migrate from the urban to rural areas which results to an increase in

* Corresponding author.
E-mail address: akande.noah@lmu.edu.ng

1877-0509
10.1016/j.procs.2022.01.100

the population of urban dwellers. As the population of urban dwellers increases, the demand for accommodation increases. As the demand increases, the price of house also increases. In addition to these, the infrastructural developments in an area can result in a sudden rise in the price of houses in a particular area. For instance, once the challenges of unmotorable road and unstable electricity a residential area become resolved, house owners tend to increase the prices of house in that particular area. In several nations, such as the United States Federal Housing Finance Agency HPI, the United Kingdom National Statistics HPI, the United Kingdom Land Registry's HPI, the United Kingdom Halifax HPI, the United Kingdom Rightmove HPI, and Singapore's URA HPI, the House Price Index (HPI) is often used to calculate price increases in residential housing [1,2,3,4]. However, research has shown that the use of HPI is not enough in this 21st century [3,4,5]. Generally, house prices are influenced by a number of variables. Authors in [6] identified these factors to be physical condition, concept and location. Physical conditions that can be observed by physical perception include the size of the property, the number of rooms, the size of the kitchen and garage, the availability of the yard, the area of land and structures, and the age of the property. Physical characteristics of a house, such as the size of the structure, the year it was built, the number of bedrooms and bathrooms, and other facts that may define the house's interior features, may affect the price of a house [7]. Although concepts refer to various marketing tactics employed by developers to attract potential investors. This includes how close the property is to hospitals, markets, educational institutions, airports, major roads etc. The location of a property has a significant impact on its price. This is because the current land price is determined by the area.

Therefore, understanding house price patterns and determining factors is not only a thing of interest to tenants alone; it is also an issue of interest to home owners, analysts and policy makers in the real estate industry as well as urban and regional planning authorities [8]. A computer-based prediction system can help them to make informed decision about if a property should be acquired and the best time to acquire the property [9,10,11,12] Residential real estate is the primary store of equity for the middle class that serves as leverage for new businesses. However, rising house prices can boost demand by increasing homeowners' income, but they can also promote debt-financed consumption and weaken financial resilience. Price forecast strategies can be divided into two categories. The first category of strategies was intended to forecast market trends in a time-series format, such as stock and oil price forecasting. The second category of approaches focuses on estimating the price of particular goods based on their characteristics, such as the cost of a house or an airline ticket. The second form of price prediction task is the focus of this article. The time-series strategy involves looking for a relationship between present and previous rates. The second method involve the use of hedonic pricing and linear regression. The second approach that involves the use of Random forest algorithm was adopted into his study. Over the years, machine learning techniques have been greatly explored for price prediction. The results obtained have shown the predictive prowess of machine learning algorithm. Machine learning creates algorithms and builds models from data, then applies them to new data to make predictions. The key distinction between a model and a conventional algorithm is that instead of simply executing a sequence of instructions, a model is constructed from input data. Unsupervised learning uses unlabelled data, while supervised learning uses data with results labelled. Regression, inference, neural networks, and deep learning are some of the most popular machine learning algorithms. However, this work explores the use of random forest machine learning technique for house price prediction. UCI Machine learning repository Boston housing dataset was used to evaluate the performance of the proposed model while Mean Absolute Error (MAE), R^2 or Coefficient of Determination and the Root Mean Square Error (RMSE) were used as performance evaluation metrics.

2. 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.

2.1. Data Collection and Exploration

In the development of the model, the UCI Machine learning repository Boston housing dataset was used. The dataset was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusett dataset. Before constructing a regression model, exploratory data analysis is needed. Researchers may uncover the data's underlying trends in this manner, which aids in the selection of suitable machine learning approaches. Therefore, data exploration was carried out to understand the features present in the dataset and their purpose. The features present in the dataset are: CRIM which is the per capita crime rate by town, ZN which is the proportion of residential land zoned for lots over 25,000sq.ft, INDUS which is the proportion of non-retail business acres per town, CHAS which is the Charles River dummy variable (1 if tract bounds river, 0 otherwise), NOX which is nitric oxides concentration (parts per 10 million), RM is the average number of rooms per dwelling, AGE signifies proportion of owner-occupied units built prior to 1940, DIS is the weighted distances to five Boston employment centers, RAD is the index of accessibility to radial highways, TAX is the full-value property-tax rate per $10,000, PTRATIO is the pupil-teacher ratio by town, B 1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town, LSTAT is the percentage of lower status of the population and MEDV is the median value of owner-occupied homes in $1000's. Since the model uses a supervised learning method, the dataset must be divided into the training dataset and testing dataset. For the training dataset, 70% of the dataset was used to train the model while the remaining 30% was used for testing.

2.2. Data Pre-Processing

The data acquired for model training and testing should be analyzed appropriately before creating models so that the models can learn the patterns more quickly. Numerical values were normalized, while categorical values were encoded one-at-a-time. After the exploration of the data and selecting the most suitable feature with the use of the heatmap, the next stage is the pre-processing of the data of the selected features that will be used. Typically, the datasets acquired for the training and testing task have several features. It is highly probable that the values of various features are on a different scale which may lower the performance of the model, therefore, scaling was carried out to ensure that the features are on a relatively similar scale. The Standard Scaler function available in Phyton Skitlearn module was for this task. The Standard Scaler assumes that your data is naturally distributed within each function and scales it so that it is now clustered about 0 with a standard deviation of 1. The feature's mean and standard deviation are measured and then the feature is scaled based on:

$$\frac{\mathrm{xi} - \mathrm{mean(x)}}{\mathrm{stdev(x)}}$$

After scaling the features, a linear regression plot (regplot) was drawn to see the correlation between the features and MEDV. This is to understand the dataset better since MEDV is the variable that will be forecasted.

2.3. Model Development

The proposed model was built using the random forest algorithm. The random forest was implemented using the RandomForestClassifier available in Phyton 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

i. Create an n-sample random bootstrap sample (by substitution, select n samples at random from the training set).
ii. At each node, build a decision tree using the bootstrap sample:
    a. Select d functions at random without replacing them.
    b. Divide the node using the attribute that offers the optimal split according to the objective function, such as optimizing knowledge gain in this case.
iii. Repeat steps 1-2 k times more.
iv. By combining the predictions from each tree, a majority vote is used to give the class name.

Furthermore, the n_estimators parameter in the RandomForestClassifier helps us to choose how many trees to create which we set at 500. The greater the number of trees in the forest, the more accurate it is, and the issue of overfitting is avoided. Although increasing the number of trees in the random forest enhances accuracy, it also increases the model's average training time. The bootstrap parameter, which we set to True, is also included in the class. Only a limited set of features will be used to introduce variation into random forest subsets, however. We improved the efficiency of the RandomForestClassifier by iterating the model several times and adding a few parameters when we initialized it.

3. 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. The first premise is that these columns will be useless in a regression task like forecasting MEDV (Median value of owner-occupied homes).

```
In [98]: 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

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

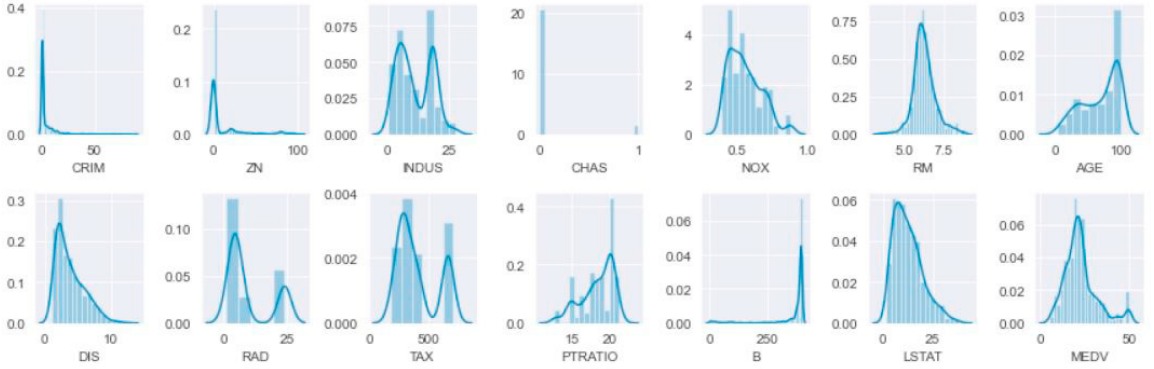

Fig. 2 Histogram Diagram

The histogram further reveals that the distributions of columns CRIM, ZN, and B are heavily distorted. Also, with the exception of CHAS, MEDV appears to have a regular distribution (the predictions) and the other columns appear to have a normal or bimodal distribution of data (which is a discrete variable). The final stage for the data exploration is the correlation matrix. A correlation matrix is a table that displays the coefficients of correlation between the table's random variables (Xi) and other table's values (Xj). This reveals the pairs with the highest correlation. For us to be able to get the correlation between the columns, seaborn heatmap was used. A heatmap is a graphical representation of data that uses colors to represent data values. That is, it makes use of color to convey a message to the reader. When there is a large amount of data, the use of heatmap is a great way to guide the viewer to the most important areas. Seaborn heatmaps are visually pleasing and appear to deliver straightforward data messages almost instantly. Therefore, data analysts and data scientists alike use this approach for correlation matrix visualization. The heatmap generated is shown in Fig. 3.

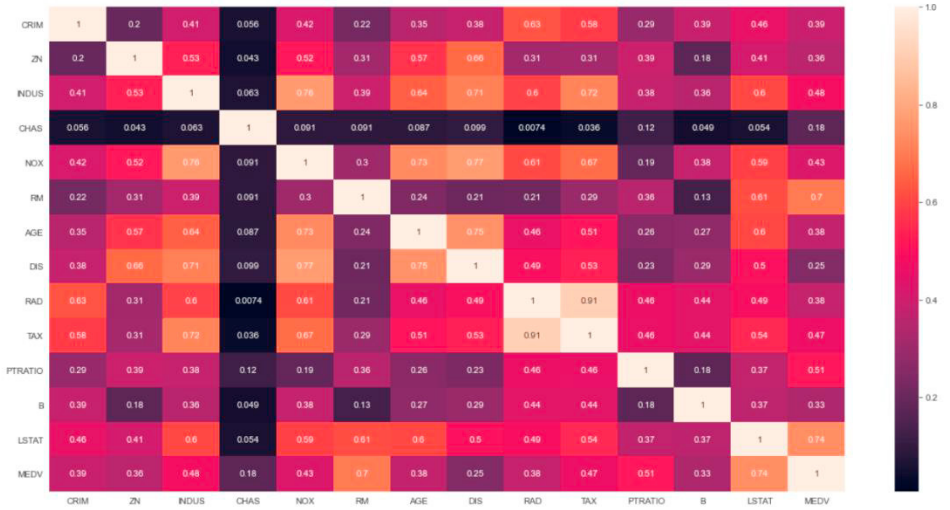

Fig. 3 Heatmap

From Fig. 4, it was observed that TAX and RAD are strongly correlated attributes from the correlation matrix. With respect to MEDV, the correlation scores of other columns were above 0.5. this is a good sign that they can be used as predictors. Therefore, we are going to be using this feature to predict the MEDV.

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.

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). The performance of the model based on the metrics is provided in Fig. 4.

```
.............Evaluation metrics..............
RSquared:  0.9001431198457122
MAE 1.9001315789473687
MSE:  6.702676631578947
RMSE:  2.588952805977534
```

Fig. 4 Evaluation of the Random Forest Model

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. This is shown in Fig.6. An appreciable performance of the predicted values over the actual values could be as a result of the K-fold cross-validation and Coupling effect of multiple regressions. The k-fold cross-validation approach is a good way to find a good bias-variance trade-off. This approach is used by Stacking Regression to determine the generalization efficiency of each variable model. Various regression methods can complement one another. The second stacking level will learn and correctly forecast house prices based on the first stacking level's pre-estimated prices.

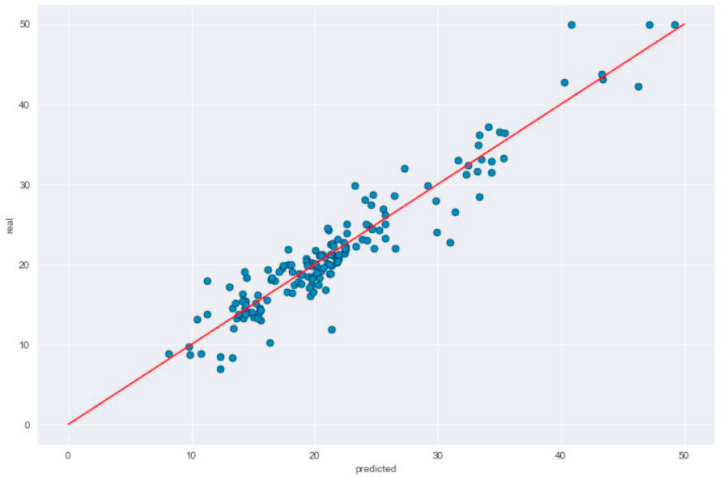

Fig. 5 Scatter Plot Real vs Predicted.

Conclusion

Every year, house prices rise, necessitating the creation of a mechanism to forecast future house prices. Land owners, estate valuers, and policymakers may use house price prediction to calculate the valuation of a home and the acceptable sale price. This will assist potential buyers in determining the right time to purchase a home. While physical conditions, styles, and location are the three main factors that influence a house's price, the individual variables that influence a house's price vary. Therefore, a perfect prediction model must accommodate the specific variables that influences the price of a house in the region being considered. This study has further affirmed the prowess of random forest machine learning technique in predicting the prices of a house based on variables made available in Boston housing dataset. A comparison of the predicted and actual prices sown in Table 1 revealed that the model achieved a prediction difference of ±5. This showed that the model can be used to predict house prices. Several other machine learning models especially deep learning models can also be explored for house price prediction.

Acknowledgements

Authors appreciate Landmark University Centre for Research, Innovation and Development for sponsoring the publication of this article.

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] 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.

[7] 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

[8] 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

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

[10] Aderonke Anthonia Kayode, Noah Oluwatobi Akande, Adekanmi Adeyinka Adegun, Marion Olubunmi Adebiyi (2019), "An automated mammogram classification system using modified support vector machine", Medical Devices: Evidence and Research, 12, 275—284.

[11] Kayode Anthonia Aderonke, Akande Noah Oluwatobi, Saheed O Jabaru, Oladele O Tinuke (2020), "An Empirical Investigation of the Prevalence of Osteoarthritis in South West Nigeria: A Population-Based Study", International Journal of Online and Biomedical Engineering (iJOE), 16(1), 100-114.

[12] Oluwatobi Noah Akande, Oluwakemi Christiana Abikoye, Aderonke Anthonia Kayode, and Yema Lamari (2020), "Implementation of a Framework for Healthy and Diabetic Retinopathy Retinal Image Recognition", Scientifica, Volume 2020, Article ID 4972527, pp. 1-14.