MULTIDISCIPLINARY SCIENCE JOURNAL

CONFERENCE PAPER
Published Online: August 18, 2023
https://doi.org/10.31893/multiscience.2023ss0206

Exploring the Intersection of Social Science and Humanities: Current Research and Challenges

Predicting future housing prices: a machine learning approach

Pushpendra Kumar Verma[a] ✉ | **Satyendra Arya**[b] | **Chetana Asbe**[c]

[a]IIMT University, Meerut, Uttar Pradesh, India, Associate Professor, School of Computer Science and Applications.
[b]Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Associate Professor, Management and Technology.
[c]ATLAS SkillTech University, Mumbai, India, Associate Professor, Management & Entrepreneurship.

Abstract In our ecosystem, the real estate sector is the least transparent. Daily changes in housing prices as well as sometimes exaggerated prices rather than valuations are a part of life in the housing market. Our research project's major focus is on predicting future housing prices using actual machine learning. Here, we want to concentrate our judgments on each fundamental factor that goes into calculating the price. Machine learning has played a significant role in picture identification, spam restructuring, normal speech command, product suggestion, and medical diagnosis in recent years. The current machine learning method aids us in improving security warnings, maintaining public safety, and improving medicinal advancements. Machine learning technology also improves customer service and makes automobiles safer. The current research discusses the prediction of future housing prices provided by a machine learning system. The data was preprocessed after it was collected. In this procedure, we employ the Forest Neural Gradient Boosting Algorithm (FNGBA). We evaluate and compare several prediction techniques for the selection of prediction methods. Our findings demonstrate the necessity for a successful approach to the problem and the capability of our method to provide predictions that can be compared to existing models of housing price prediction. When compared to employing independent methods, the findings showed that this technique delivers the least mistake and the maximum accuracy.

Keywords: housing price, machine learning, FNGBA

1. Introduction

One of the most important decisions a person makes in their life is purchasing a house. Numerous factors, such as a house's location, attributes, and supply and demand in the real estate market, may influence its price. Another crucial component of the national economy is the housing sector (Yap and Ng 2018). As a result, anticipating house prices benefits not only purchasers but also real estate brokers and economists. Housing market forecasting studies look at home prices, growth trends, and their links with various factors (Wu and Brynjolfsson 2015). Over the last several years, the development of machine learning methods and the explosion of data or big data accessible have opened the way for real estate research (Zhou et al 2019). There is a wide range of research that uses statistical learning approaches to analyze the housing market. Machine learning is heavily used in this approach. The machine learning prediction process goes through several phases. Before analysis, data are first collected and processed (Grieco et al 2016).

The data may need to be cleaned up, formatted properly, and given the proper features. The machine-learning model is subsequently trained using the prepared data (Hussain et al 2019). To acquire the perfect combination of features and algorithms for producing accurate predictions, the model is updated. Giving reliable datasets is a need of machine learning, and predictions made thereafter are dependent on them. A machine learning model may be trained to predict housing prices using historical data on sales of homes and information on the factors that influence house prices (Hu et al 2019). A model that can predict housing prices based on input characteristics may be trained using this data. For predicting house prices, a variety of machine learning techniques may be utilized, such as decision trees, random forests, and linear regression (Rodriguez-Galiano et al 2015).

The purpose of the essay was to assist both buyers and sellers in accurately estimating a home's selling price as well as to assist readers in determining when to start building a home (Madhuri et al 2019). (Wang et al 2019) presented a deep learning-based model for predicting future home prices, and we show how it can be deployed using the TensorFlow library. The model is trained using the Adam optimizer, and the Relu function is employed as the activation function. (Adetunji et al 2022) investigated the use of the Random Forest machine learning approach for predicting home prices. 506 items and 14 characteristics from the Boston housing dataset from the UCI Machine Learning Repository were used to assess the effectiveness of the suggested prediction model.

(Lu et al 2017) suggested a hybrid Lasso and Gradient boosting regression model to forecast the price of each home. Recently, the suggested strategy was implemented as the primary kernel for the Kaggle Challenge "House Prices: Advanced Regression Techniques." (Yu et al 2018) examined the secondary market for real estate in Beijing from three perspectives: a review and analysis of the market, a prediction of future housing prices, and a comparison of the results. In the end, the best model program was developed, which has significant implications for assessing and forecasting home prices. (Bency et al 2017) offered a methodology for employing Convolutional Neural Networks (CNNs) to model geospatial data (in this case, home price data) to automatically learn the spatial correlations between variables. (Wang et al 2019) compared the properties of a Random Forest (RF)-based home price predictor with those of a traditional hedonic pricing model to better understand its strengths and weaknesses. They analyzed condo sales in Gangnam, one of South Korea's most affluent neighborhoods, from 2006 through 2017. (Lisi et al 2019) provided commentary on the application of hedonic pricing models to real estate valuation. Some housing markets may benefit greatly from the concept. The Forest Neural Gradient Boosting approach (FNGBM), which provides a more accurate result than any other approach, is introduced in this research for predicting future housing prices.

2. Materials and Methods

From the major real estate agency in the Christchurch region, Harcourt, we randomly choose 200 house details to analyze. In May 2003, we obtained the dataset from Harcourt's website. Due to the concentration of businesses, restaurants, and stores in the downtown area, the percentage of homes is quite low. Only 15 house records are gathered from the central city, 25 from North Christchurch, and 40 from the other four designated areas. In all, 200 separate observations are used here.

2.1. Z-score normalization

Z-score normalization uses the data's mean and standard deviation as its starting points. In cases when the lowest and maximum values of the data are unknown, this technique is of great use. This is the formula that is used:

Y_{new}= $\frac{y-\mu}{\sigma}$ =$\frac{y-Mean(Y)}{stdDev(Y)}$

Y_{new}=The adjusted value obtained after scaling the data

Y = outdated value

μ = Statistics mean

σ = Estimated Standard Deviation

2.2. Forest Neural Gradient Boosting Algorithm

The Gradient Boosting algorithm iteratively combines weak learners who are just slightly superior to random—into effective learners. A regression approach similar to boosting is called gradient boosting. The objective of finding an approximation is done using gradient boosting. E (v), of the operator$E^*(v)$, which translates examples to their final values y, by examples to their final values of a specified loss function, $L\big(z, E(v)\big)$, provided a training dataset $C = \{v_j, z_j\}_1^M$, Gradient boosting produces a weighted average of variables that additively approaches the value of $E^*(v)$.

$$E_m(v) = E_{m-1}(v) + \rho_n g_n(n) \qquad (1)$$

Where $E_m$is the n[th] variable weight, ρ_n (v). Those are the ensemble's models, such as decision trees. The closest estimate is built up gradually. At first, an approximate constant representation of $E^*(v)$ is found as

$$E_0(v) = arg \min_{\alpha} \sum_{j=1}^{M} K\big(z_j, \alpha\big). \qquad (2)$$

Future models are anticipated to reduce

$$\big(\rho_n g_n(n)\big) = arg \min_{\alpha} \sum_{j=1}^{M} K\left(z_j, E_{n-1}\big(v_j\big) + \rho g\big(v_j\big)\right) \qquad (3)$$

In contrast to explicitly addressing the optimization issue, every $\rho_n$might seem to rapacious move in the gradient reduction efficiency for E^*. Every simulation (ρ_n) is then qualified on a different dataset $C = \{v_j, z_{nj}\}_{j\ =\ 1}^{M}$,, with pseudo-residuals (q_{nj}) determined by

$$q_{nj} = \left[\frac{\partial K\big(z_j, E(v)\big)}{\partial E(v)}\right]_{E(v)=E_{n-1}(v)} \qquad (4)$$

After that, we solve an optimization issue involving a line search to get the value of n. Over-fitting may occur in this approach

if the iterative procedure is not adequately regularized. If the model properly hm fts the pseudo-residuals, the procedure will stop after just one iteration if the loss function is a quadratic loss, for example. To regulate the cumulative effect of gradient boosting, several regularization hyper-parameters are taken into account. Regularizing gradient boosting is as simple as applying shrinkage to each gradient decent step $\rho_n(v) = p_{n-1}(v) + x\,\rho_n g_n(v)$ with x= (0, 1.0). Typically, is given a value of 0.1. Regularization may be improved by putting constraints on the learned models' complexity. For decision trees, we may restrict the lowest value of occurrences needed to split a node or the depth of the tree. While random forest's default settings for those hyper-parameters don't restrict the expressive capacity of the trees, the default values in gradient boosting do (for example the depth is often restricted to 3 5). Last but not least, random subsampling without replacement belongs to a class of hyper-parameters that is incorporated in several gradient-boosting variants and may further enhance the generalization of the ensemble. Gradient boosting's last traits to be put to the test are:

- Shrinkage, also known as learning rate.
- The tree's greatest depth: the same significance as that of the trees produced by random forests.
- The fraction of the whole sample is taken as a subsample. This is often done without replacement.
- Similar to a random forest, the maximum amount of features are used to determine the optimal split (max_features).
- A minimal number of samples are needed to divide an inner node, like in a random forest (min_samples_ split).

In forest regression, the bagging of trees technique is used. Making the different trees seem less like they belong together is the goal. Then, to reduce their variation, we average the trees. Several decision trees might be built using this strategy.
The random forest method of training uses bootstrap aggregation (sometimes called bagging) to train tree-based models. It selects a random subset of the training data $V = v1, \dots, vm$ and repeatedly fits trees to this subset $Z = z1, \dots, zm$ until convergence is achieved, a process called "bagging." $a = 1, \dots, A$:

1. Using V and Z as training data randomly choose n training samples (V_a, Z_a).
2. Put V and Z through a $ea\ on V_a, Z_a$ classification/regression tree's training process.

After training, you may make predictions about future samples b' by averaging the outputs of many regression trees that have been fed the same data:

$$\hat{e} = \frac{1}{A}\sum_{a=1}^{A} ea\,(v') \qquad (5)$$

The standard deviation of the predictions generated by each regression tree on a' provides an additional measure of the prediction's uncertainty.

$$\sigma = \frac{\sqrt{\sum_{a=1}^{A}(ea(v')-\hat{e})^2}}{A-1} \qquad (6)$$

In addition, the output from each of these methods is used as training data for the neural network. To provide a precise outcome, we use a neural network to boost regression. Neural networks are effective because they compare and compute all of the predictions to provide the most accurate outcome. Figure 1 illustrate how the system operates. The consumer, the database, and the website all constitute separate but equal objects. The algorithmic computational mechanisms are also included. The user interface prompts the consumer to input the desired location, desired area, and other factors related to the home purchase. The user enters their desired price range and desired location, and the system shows houses that fall inside that range.

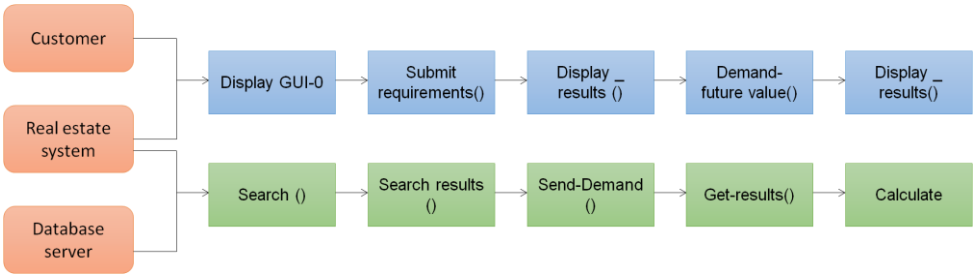

Figure 1 Diagram of the sequence.

3. Result

In this paper, we use the FNGBA Technique to predict the future price of housing. We have examined several of the previously established approaches, such as CNN, SVR, and LSTM, and contrasted them with our technique FNGBA.R2, RMSE, accuracy, and prediction rate are the metrics that are being used.

3.1. Accuracy

Accuracy is a statistical metric that is used to reflect the degree to which the anticipated value and the actual value are similar to one another. The higher denotes a better performance. Figure 2 and Table 1 show comparisons between the recommended strategy and more conventional methods. The FNGBA technique that was suggested had a higher degree of accuracy.

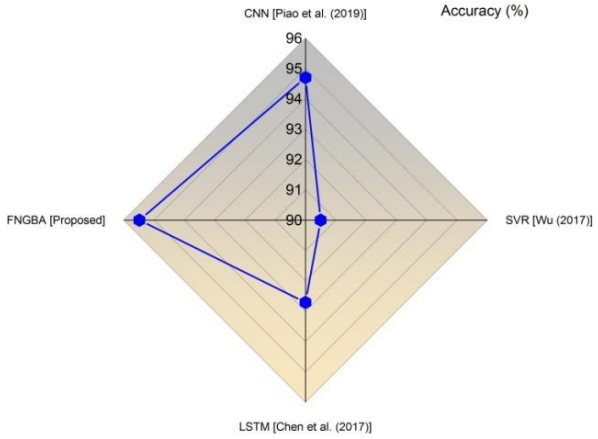

Figure 2 Accuracy comparisons between the suggested and current approaches.

Table 1 Comparison of Accuracy.

Methods	Accuracy (%)
CNN [Piao et al (2019)]	94.7
SVR [Wu (2017)]	90.5
LSTM [Chen et al (2017)]	92.71
FNGBA [Proposed]	95.48

$$\text{Accuracy} = \frac{[observed_t - predicted_t]}{observed_t} \times 100\% \quad (7)$$

3.2. The R^2

The R^2 or coefficient of determination is a statistical metric used to assess the extent that a model fits the data. It is a statistical measure used to assess the reliability of the regression line. Figure 3 and Table 2 show comparisons between the recommended techniques with more conventional methods, respectively. The FNGBA strategy that was suggested had a higher degree of R2 compared to the earlier research.

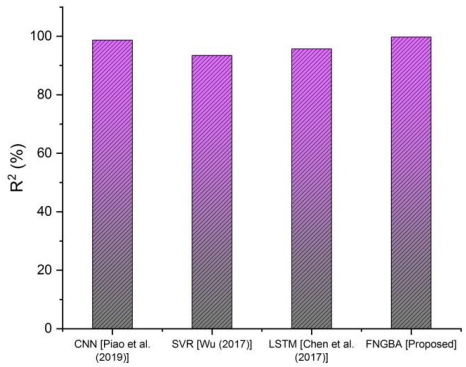

Figure 3 The R^2comparisons between the suggested and current approaches.

Table 2 Comparison of R^2.

Methods	R2 (%)
CNN [Piao et al (2019)]	98.68
SVR [Wu (2017)]	93.45
LSTM [Chen et al (2017)]	95.72
FNGBA [Proposed]	99.72

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - \hat{P}_i)^2}{\sum_{i=1}^{n}(P_i - \overline{P})^2} \quad (8)$$

3.3. RMSE

An estimator's RMSE for a population parameter is its MSE multiplied by its square root. The MSE is calculated by taking the square root of the anticipated value of the discrepancy between the estimate and the parameter. In Figure 4 and Table 3, respectively, comparisons between the recommended technique and more conventional methods are shown. In comparison to the earlier research, the FNGBA technique has a lowered error rate.

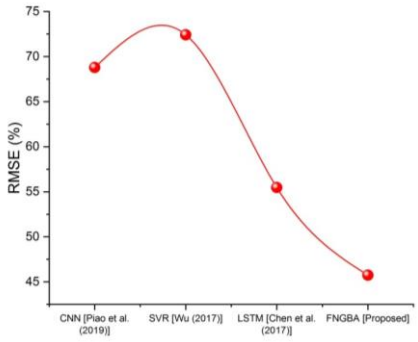

Figure 4 Comparison of RMSE of the existing and proposed methodologies.

Table 3 Comparison of RMSE.

Methods	RMSE (%)
CNN [Piao et al (2019)]	68.78
SVR [Wu (2017)]	72.42
LSTM [Chen et al (2017)]	55.48
FNGBA [Proposed]	45.75

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - \hat{P}_i\right)^2} \quad (9)$$

3.4. Precision

Precision is a statistic used in binary classification issues to evaluate the percentage of genuine among the most optimistic predictions generated by the model. In other words, precision measures the model's accuracy in making positive predictions. Comparisons of the suggested method with traditional approaches are presented in Figure 5 and Table 4, respectively. the FNGBA technique that was suggested had a higher precision rate.

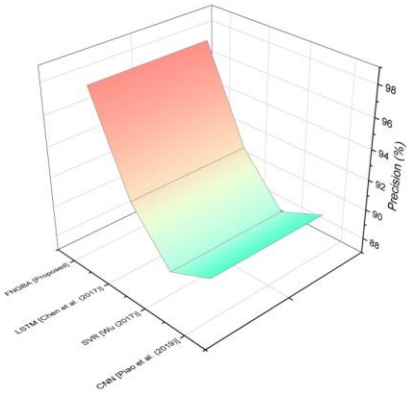

Figure 5 Comparison of Precision of the existing and proposed methodologies.

Table 4 Comparison of Precision.

Methods	Precision (%)
CNN [Piao et al (2019)]	89.78
SVR [Wu (2017)]	88.54
LSTM [Chen et al (2017)]	91.71
FNGBA [Proposed]	97.74

$$Precision = \frac{True\ Positives}{True\ Positives + False\ positives}$$

4. Discussion

The process of predicting house prices is intricate and diverse, requiring a thorough grasp of the housing market and the variables that influence it. It is possible to create exact projections about future housing prices and to decide on buying, sale, or investment in real estate by using the forest neural gradient boosting algorithm.

5. Conclusions

The goal of this approach is to create an exact prediction of future housing prices. Forest regression is used effectively by the system. The addition of neural networks has significantly enhanced the algorithm's efficiency. Consumers will be impressed since the method eliminates the possibility of purchasing the incorrect property and produces reliable results. The system's essential functioning may be maintained while new features for the customer's benefit are introduced. Larger cities may be added to the database in a future update, allowing our customers to see additional houses, get more precise information, and make an informed choice. The system's precision may be enhanced. If the system is made larger and has more processing capacity, it will be possible to include many more citations. In addition, using Augmented Reality, we may combine several UI/UX approaches for a more dynamic display of the outcomes. In addition, a learning system may be developed that takes into account user input and past activity to provide personalized outcomes.

Ethical considerations

Not applicable.

Declaration of interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

Adetunji AB, Akande ON, Ajala FA, Oyewo O, Akande YF, Oluwadara G (2022) House price prediction using random forest machine learning technique. Procedia Computer Science 199:806-813.

Bency AJ, Rallapalli S, Ganti RK, Srivatsa M, Manjunath BS (2017) Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In 2017 IEEE winter conference on Applications of computer vision (WACV), pp 320-329. IEEE.

Chen X, Wei L, Xu J (2017) House price prediction using LSTM. arXiv preprint arXiv:1709.08432.

Grieco G, Grinblat GL, Uzal L, Rawat S, Feist J, Mounier L (2016) Toward large-scale vulnerability discovery using machine learning. In Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp 85-96.

Hu L, He S, Han Z, Xiao H, Su S, Weng M, Cai Z (2019) Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land use policy, 82, pp.657-673.

Hussain M, Zhu W, Zhang W, Abidi SMR, Ali S (2019) Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review 52:381-407.

Lisi G (2019) Property valuation: the hedonic pricing model–location and housing submarkets. Journal of Property Investment & Finance.

Lu S, Li Z, Qin Z, Yang X, Goh, RSM (2017) A hybrid regression technique for house prices prediction. In 2017 IEEE international conference on industrial engineering and engineering management (IEEM), pp 319-323. IEEE.

Madhuri CR, Anuradha G, Pujitha MV (2019) House price prediction using regression techniques: A comparative study. In 2019 International Conference on smart structures and Systems (ICSSS), pp 1-5. IEEE.

Piao Y, Chen A, Shang Z (2019) Housing price prediction based on CNN. In 2019 9th international conference on information science and Technology (ICIST), pp 491-495. IEEE.

Rodriguez-Galiano, V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas MJOGR (2015) Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees, and support vector machines. Ore Geology Reviews 71:804-818.

Wang F, Zou Y, Zhang H, Shi H (2019) House price prediction approach based on deep learning and ARIMA model. In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), pp 303-307. IEEE.

Wu JY (2017) Housing price prediction using support vector regression.

Wu L, Brynjolfsson E (2015) The future of prediction: How Google searches foreshadow housing prices and sales. In Economic analysis of the digital economy, pp 89-118). University of Chicago Press.

Yap JBH, Ng XH (2018) Housing affordability in Malaysia: perception, price range, influencing factors and policies. International Journal of Housing Markets and Analysis.

Yu L, Jiao C, Xin H, Wang Y, Wang K (2018) Prediction on housing price based on deep learning. International Journal of Computer and Information Engineering 12:90-99.

Zhou T, Song Z, Sundmacher K (2019) Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 5:1017-1026.