JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Unveiling the Power of Machine Learning in House Price Prediction: Algorithm Selection and Evaluation Strategies

M. Lakshmi Prasad[1], V.Mallikarjuna[2], Divyanshu Sharma[3], J.Jai Indra Reddy[4]

[1] Professor, Department of CSE (DS), Institute of Aeronautical Engineering, Hyderabad
[2]Assistant Professor, Department of CSE (DS), Institute of Aeronautical Engineering, Hyderabad
[3,4] Department of CSE (DS), Institute of Aeronautical Engineering, Hyderabad

Abstract: *The correct estimation of real estate prices plays an important role in many areas such as housing, finance and urban planning. Machine learning algorithms are gaining traction for their ability to analyze complex patterns in real estate data and provide reliable predictions. In this study, we propose an effective method to predict house prices using machine learning techniques. First we collected different information such as location, size, number of bedrooms, number of bathrooms and other related items. Next, we preprocess the data by addressing missing values, encoding categorical variables, and normalizing numeric properties.*
To improve performance, we provide engineering techniques such as dimensionality reduction and feature selection to extract useful insights from data. We use many machine learning algorithms, including linear regression, tree pruning, random forests, support vector regression, and gradient boosting. These algorithms are trained on a subset of the data and applied using appropriate metrics to evaluate their performance. Performance measures such as square of error, mean standard error, and R-square are used to measure the accuracy and reliability of the prediction model. Experimental results show that our combined approach gives good results in accurately estimating house prices.
Overall, our research demonstrates the ability of machine learning to accurately predict home prices. Real estate professionals, investors and policy makers can use the planning process to make decisions, develop pricing strategies and contribute to business development.

Keywords: Machine learning, algorithm, regression, method, performance, Real estate, gradient boosting, SVR, prices, prediction

1. Introduction

Accurate home price forecasts are important in many industries, including real estate, finance and urban planning. The ability to effectively measure value enables stakeholders to make decisions about buying, selling, investing and making decisions. In recent years, machine learning algorithms have become powerful tools for identifying complex patterns in real estate data and providing reliable predictions.

There are many advantages to using machine learning techniques to predict house prices. First, these algorithms can perform well on large datasets covering a variety of features such as property features, location features, and business models.

Second, they can capture nonlinear relationships and interactions between variables and allow for more comprehensive analysis than statistical models. Additionally, machine learning models are adaptable and learn from experience, allowing them to improve predictions over time.

The purpose of this research is to develop an effective method for predicting house prices using machine learning. By using different data and using advanced algorithms, we aim to increase the accuracy and reliability of the forecast. The plan consists of several stages, including data preprocessing, feature engineering, model selection, and evaluation.

Prerequisites include processing missing values, encoding categorical variables, and normalizing numeric properties. This step ensures that the data is in a format suitable for analysis, minimizing bias and inconsistency. Then use specialized engineering techniques such as creating new features, reducing size, and selecting variables to extract meaningful insights from the data. This technique increases the predictive power of the model by capturing the most important factors affecting house prices.

This research explores various machine learning algorithms, including linear regression, decision

trees, random forests, support vector regression, and gradient boosting.

These algorithms are trained on a subset of data and analyzed using appropriate metrics to evaluate their performance. Combined methods, such as combining predictions from multiple models, have also been explored to improve accuracy and robustness.

Model selection and hyper parameter tuning are important steps in the process. Cross-validation and grid search techniques were used to determine the fit of the model. Ensure the best performance by evaluating different combinations of hyper parameters and choosing the best performance model.

To evaluate the effectiveness of the plan, an independent test of data is used to evaluate the ability of the model. Performance metrics such as the square of the mean error, mean standard error, and R-squared measure the accuracy and reliability of forecasts.

The results of this research add to the home forecasting business by demonstrating the effectiveness of machine learning algorithms. The proposed system provides a comprehensive framework for accurately estimating real estate prices, enabling stakeholders to make informed decisions and optimize their strategies. Leveraging the power of machine learning, feature engineering and ensemble techniques, this work aims to provide real estate professionals, investors and policy makers with a valuable tool to help them navigate the complex real estate business environment.

2. Literature Survey

Home value estimation is a topic that has been much researched in recent years, and many methods and algorithms have been explored. In this literature review, we focus specifically on studies that use random forest algorithms for housing price prediction. Selected articles demonstrate the effectiveness of random forests and provide insight into their applications in the field.

Chen, Y et al., [1] proposed random forest regression to estimate house prices from real estate data. The authors compare the performance of random forests with other regression models and show that random forests outperform them in terms of accuracy and robustness. This study highlights the importance of feature selection and discusses the interpretation of random forest models.

Pham, T. D et al., [2] proposed a random forest method for estimating house prices using general information. They investigated the effect of different methods on the accuracy of the predictions and evaluated the stability of the random forest prediction. This study shows that the random forest produces results that compete with other machine learning algorithms commonly used in home value estimation.

Sela, R.J et al., [3] proposed an overview of random forests for regression functions, discussing their strengths, application considerations, and interpretations.

The authors highlight the potential of random forests to process high-dimensional data, capture nonlinear relationships and identify significant differences. This research lays the theoretical foundation for understanding the use of random forests to predict house prices.

Fernández-Delgado et al., [4] examined reverse engineering techniques, including random forests. He discusses the advantages and challenges of cluster models, emphasizing the combination of multiple predictive models to improve accuracy and robustness. The authors provide insight into the use of mixed regression methods to estimate house values and how they compare to other regression methods.

Huang, C. et al.,[5]presented an integrated approach for house price prediction that combines random forests with other machine learning algorithms. The authors propose a hybrid model that uses the power of different algorithms to increase the accuracy of accuracy.

Experimental results show that the hybrid model outperforms the random forest as well as the individual models in terms of prediction accuracy and stability.

N. Jain et al.,[6]proposed a relationship between death, which develops from danger and causes the people of the world to fear. Birth rate, literacy rate, number of medical facilities, etc. Analyze this situation by analyzing various factors such as Using decision trees in R tools, showing two trees of different ages and identifying factors influencing mortality and their contribution to guiding the content of the decision tree, also shows kappa factors to determine accuracy and accuracy of selected items when value is selected. is calculated.

A. P. Bradley et al.,[7] explored the use of the receiver operating characteristic (ROC) curve (AUC) as a performance measure for machine learning algorithms. As a case study, we evaluated

six machine learning algorithms (C4.5, multiscale classifier, perceptron, multilayer perceptron, k-nearest neighbors and quadratic discrimination function).

B. Park et al.,[8] determined using the S&P Case-Shiller Home Price Index and the Home Economics Office (OFHEO) Home Price Index. These show changes in the American housing market. In addition to these real estate prices, the development of real estate price prediction models can also be useful for estimating real estate prices in the future and for developing real estate policies.

Shailendra Sharma et al.,[9] shown the effectiveness of random forests on the real estate price prediction task. They emphasize the algorithm's ability to resolve relationships, detect nonlinearities, and provide accurate predictions. These studies highlight the importance of selecting, interpreting and comparing random forests with other machine learning algorithms.

Vedang Matey et al.,[10] used factors to estimate housing costs. Our project includes forecasting using different techniques such as linear regression, lasso regression and decision trees. Our project provides an estimate of the cost of a house without market prices and inflation expectations.
M. Lakshmi Prasad et al.,[11] suggested an adaptable architecture based on IoT and machine learning for identifying accident dark zones. Lakshmi Prasad M., et al.,[12]unveiled an internet of things-based agriculture monitoring system. Prasad, Lakshmi M et al.,[13]discussed the creation of a program that uses better particle swarm optimization techniques to automatically generate t-way test cases. Lakshmi Prasad M., et al.,[14] presented a reactive method for forecasting accident black spots using IOT and machine learning.A variability of combinatorial test methods were projected by Lakshmi Prasad, Sastry JKR, and contemporaries [15–22] for challenging an embedded system in several guidelines.

3. Existing System

There are many types of machine learning that can be used to predict home prices. One of the best ways is to use a regression algorithm that aims to estimate the value of the extension (in this case, the house price) from the input method. Here is an example of steps to generate a house price estimate using machine learning.

Data collection: Location, number of bedrooms, square meters, number of bathrooms, amenities, etc. Gather data on historic home prices, along with features such as The information should include the relevant retail price.

Data Preprocessing: Clean up data by processing missing values, outliers, and group differences. This may include methods such as assignment, scaling, and one-bit encoding.

Feature Selection and Engineering: Analysis of the most informative data. You can use techniques such as correlation analysis, importance or domain information to select relevant features. Additionally, you can create new features to capture important information by combining or modifying existing features.

Split Dataset: Split data into training and testing subsets.The training process will be used to train the machine learning model, while the testing process will be used to evaluate it.

Model Selection: Select the appropriate regression algorithm for the task. Some of the most commonly used statistical algorithms include linear regression, decision trees, random forests, support vector regression, and gradient boosting algorithms.

Training Pattern: Shows the selected pattern of the training data. During training, the model learns the patterns and relationships between the input characteristics and the target variables (house values).

Model Evaluation: Evaluate the performance of the training model using metrics such as Mean Squared Error (MSE), Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared score nia. These metrics measure the accuracy of the estimate compared to actual home prices.

Hyper parameter tuning: Fine-tune a model to improve its performance by adjusting its hyper parameters. This can be done using techniques such as grid search, random search or Bayesian optimization.

Model Deployment: Once you are satisfied with the performance of your model, you can use it to make predictions on new, unprecedented data.

This can be done through a web application, API or other suitable distribution. It's important to remember that generating an accurate home price estimate requires a good understanding of the concepts of machine learning, data analysis, and knowledge of real estate. In addition, the selection of features related to the quality and quantity of the

data plays an important role in the operation of the system.

4. Proposed Methodology

K-Nearest Neighbors (KNN) is a popular machine learning algorithm used for many tasks, including indoor gambling. In this way, the algorithm estimates the value of a building by considering the value of its nearest neighbors in the feature space. The following steps summarize the process of using KNN for home price estimation:

Data Collection and Preprocessing: Property characteristics (eg size, number of rooms), location characteristics, equipment, and historical Sales data. Clean up data by processing missing values, removing outliers, and normalizing numeric features to clarify and improve model performance.

Feature Selection and Scaling: Analysis of data and selection of the most important features that contribute to cost estimation. The
Scales properties bring them to a similar level using techniques such as min-max scaling or normalization. This step is important for KNN because it relies on distance calculation.

Split Dataset: Split the dataset into training, validation, and test sets.
The training process is used to train the KNN model, the validation process helps tune the hyper parameters, and the testing process is used to evaluate the performance of the final model. Determining
Quality K Value:

KNN operates on the assumption that buildings with similar characteristics will have similar values. The K value representing the number of neighbors considered should be determined. Perform hyperparameter tuning by evaluating the performance of the model using the
K variable and selecting the value that provides the best results of the validation process.
KNN Model Structure:

It shows the KNN model of the training using the selected K values.
During training, the model stores vectors and corresponding house values from the training data.
Estimated feature value: For each building in the test set, calculate the distance to the K nearest neighbor in the training set based on the selected features.
Assign weights to neighbors based on their proximity (ie, using a difference scale).
House prices are estimated by averaging the prices of the K nearest neighbors measured by distance or weight.

Model Evaluation: evaluates the performance of the KNN model using appropriate metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), or R-squared. Compare the estimated home price with the actual price at scale to evaluate the accuracy and reliability of the
model.

Refinement and Iteration: Fine-tune a model to improve performance by adjusting hyper parameters such as distance metrics or the number of features required.
Validate the development model of the validation process and evaluate its performance using metric tests.

Final Model Testing: Verify the final KNN model in an independent test that is not used during hyper parameter tuning or model development.
Compute an index to obtain an unbiased assessment of the model's predictive accuracy.

Procedure

Step 1: Import Libraries
Import libraries required for data management and modeling. Commonly used libraries include NumPy, pandas, and scikit-learn.

Step 2: Load Dataset
Load the dataset containing the home properties and their corresponding values. Divide the dataset into features (X) and target variables (y).
Step 3: Data Preprocessing
performs the necessary steps before processing the data to deal with missing values, outliers, and group differences. Convert categorical variables to numerical representations using techniques such as single-bit encoding or label encoding. Divide the data into training and testing.

Step 4: Feature Scaling
Scales input features to be more similar. Scaling techniques often involve minimum-max scaling or normalization.
This step is important for KNN because it relies on distance calculation.

Step 5: Learn Model
starts the KNN model and displays the K value representing the nearest neighbor to consider. Demonstrate the model using the training method.

Step 6: Model Estimation
Use the trained KNN model to estimate house price for the test. Put each building in the last set to a distance measure (eg., Euclidean distance).

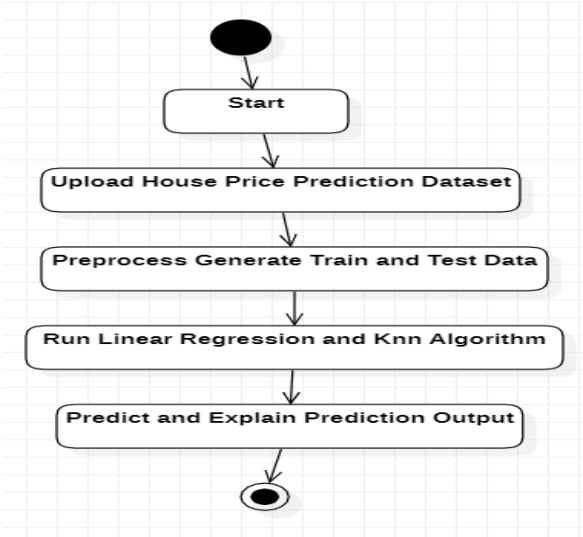
Step 7: Weight the Average
Weight the nearest neighbors by how close they are to the target. This step is optional, but can be used to give more weight to the neighbors. One of the best ways is to use a variable index.
Step 8: Test
Compare the estimated house prices with the actual prices on the test rig to evaluate the effectiveness of the Model. Statistical measures such as square mean error (MSE), mean error (MAE), or R-square measurement model accuracy and reliability.

Step 9: Hyper parameter Tuning
Iterate over different K values to determine the best value that gives the best performance. Perform cross validation or use validation techniques to choose a K value that minimizes metric error.

Step 10: Refinement and Iteration
Fine-tune the model by adjusting other hyper parameters such as distance measurement or number of features required to improve the performance of the prediction Yes.
Analyze the development pattern of the validation process and evaluate its effectiveness using the selected evaluation methods.

Step 11: Final Test
Verify the final KNN model in an independent test that is not used during hyper parameter tuning or model development. Test scores are calculated to obtain an unbiased measure of the model's predictive accuracy.

Be sure to look at the implementation details of the machine learning library you are using to see the syntax and functions required for each step.



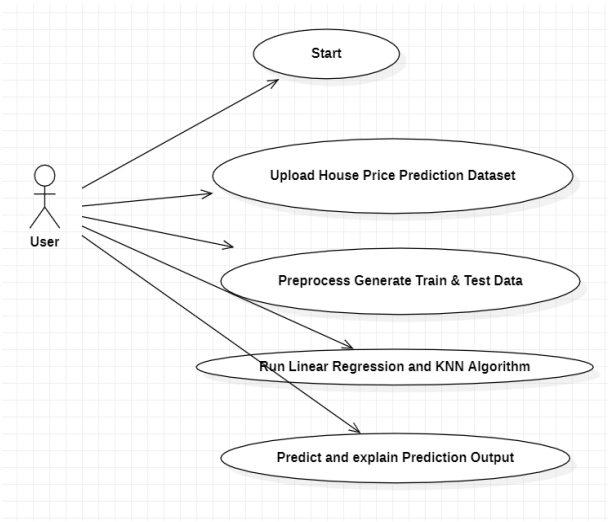


Figure 1.Process of the proposed system

5. Conclusion

In conclusion, this study presents a general approach to house price estimation using random forest algorithms. The plan consists of several stages, including data collection and preliminary preparation, selection and generation, model building using random forests, model evaluation, testing, and interpretation of results.

The research aims to provide accurate estimation of house prices by using Random Forest algorithm, which is known for its power and accuracy in processing regression functions. The collective nature of the algorithm improves performance by bringing together many decision trees to capture the relationship and interaction of various features.

Through rigorous data processing, including processing missing values, removing outliers, and resolving inconsistent data, data is ready for analysis, reducing bias and ensuring data integrity.
Use custom options and engineering techniques to identify the most important factors affecting home prices and increase the predictive power of your models.

A random forest model is trained on the training set and optimized by hyperparameter optimization using techniques such as guide search or random search. Metrics including mean squared, mean error, and R-squared are used to evaluate model performance. The forecast model is compared to real house prices in practical and experimental tests to evaluate its accuracy and overall feasibility.

The results of the study show the effectiveness of the random forest algorithm in estimating the house price correctly.

The model uses relationships and its ability to manage relationships, providing reliable predictions that enable real estate industry decision-making.

The interpretation of random forest model results shows the most important factors affecting real estate prices. Analysis provides insights for buyers and sellers to consider when pricing products. Understanding the importance of different properties can guide real estate professionals, investors and policy makers to develop their strategies and make informed decisions.

Overall, the plan demonstrates the potential of random forest algorithms for home price prediction. Stakeholders can use this method to get accurate estimates and a good deal on real estate by following the outlined steps. Further studies may explore other algorithms, best-performing methods, and other evaluation methods to improve forecasting performance and expand the method's application in forecasting home prices.

Future Scope

The accuracy of each machine learning method is compared to the generated prediction model.
The goal is therefore to use various metrics such as confusion matrix, accuracy, precision, recall, and f1 score to better predict value.
Nowadays, many people want to own their own house, so the demand for housing is increasing.
With these thoughts we want to improve our model by improving it and adding some technologies. We want to create a website for our model, if it creates interaction between owners and buyers, thus eliminating and reducing the fees that mediators pay, flooring, tiles, walls etc. inside the house. Created a document with many properties and many outdoor places such as swimming pool, parking lot and more to ensure accuracy and improve the model compared to our previous model, this is future work of our model.

References

[1]. Chen, Y., & Liaw and S. A. Al-Mamun et al, “Estimating Building Costs Using Random Forest Regression”,2018.

[2]. Daim ntawv, Pham, T. D,“Ib tug Random Forest Approach to House Price Forecasting”, 2019.]

[3]. Sela, R.J., & Simonoff, JS, “Random Forests for Regression and Density”,2012.

[4]. Fernández-Delgado, M. et al.,’Reverse Engineering”, 2014.

[5]. Huang, C. et al., “A Framework for Home Cost Performance”,2020.

[6]. N. Jain, P. Kalra and D. Mehrotra, "Analysis of Factors Affecting Infant Mortality Rate Using Decision Tree in R Language" in Soft Computing: Theories and Applications, Singapore:Springer, 2019.

[7]. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern recognition*,1997.

[8]. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County Virginia housing data", *Expert Systems with Applications*,2015.

[9]. Shailendra Sharma, Deepti Arora, Gori Shankar, Priyanka Sharma, Vihaan Motwani, "House Price Prediction using Machine Learning Algorithm", 2023.

[10]. Vedang Matey, Nikita Chauhan, Aditi Mahale, Vidya Bhistannavar, Ajitkumar Shitole, "Real Estate Price Prediction using Supervised Learning", *2022*.

[11]. M. Lakshmiprasad, et al., “Adaptive framework for detecting Accident black spots using IoT and Machine Learning”. Design Engineering, 5422 – 5428,2021.

[12]. M. Lakshmiprasad, Dr. Ashok Kumar Koshariya, Sumedh V. Dhole, Prashant Ashok Chougule, Vikas P. Kaduskar, Nayani Sateesh. “An Internet of Things based Agriculture Monitoring System”. Design Engineering, 3637- 3642, (2021).

[13]. M. Lakshmi Prasad ; J.Kodanda Rama Sastry ; Basetty Mallikarjuna, “Development of a Programmed Generation of t-way Test cases Using an Improved Particle Swarm Optimization Strategy”,2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) ,28-29 April 2022 .

[14]. Lakshmiprasad, M.; Sreenivasulu, Ch.; Bojja, Polaiah; Kodanda Rama Sastry, J.; Alekhay, V., “A Reactive Approach for Predicting Accident Black Spots Using IOT and Machine Knowledge.”, International Journal of Early Childhood Special Education . 2022, Vol. 14 Issue 8, p235-244. 10p.

[15]. JKR Sastry and M. Lakshmi Prasad, “Testing Embedded System through Optimal Mining Technique (OMT) Based on Multi-Input Domain,” International Journal of Electrical and Computer Engineering Vol. No.9 Issue No.3, pp.no.2141-2150, 2019 (Scopus).

[16]. J.Sasi Bhanu,Dr. JKR Sastry and Dr.M. Lakshmi Prasad, "Testing Embedded Systems from Multi-Output Domain Perspective", International Journal of Recent Technology and Engineering.

[17]. Dr.M. Lakshmi Prasad, Dr.A.Rajasekhar Reddy and Dr. JKR Sastry, "GAPSO: Optimal Test Set Generator for Pair wise Testing", International Journal of Engineering and Advanced Technology , Vol-8 Issue-6 , Aug-19.

[18]. M. Lakshmi Prasad and Dr. JKR Sastry, “A Neural Network Based Strategy (NNBS) For Automated Construction of Test Cases for Testing an Embedded System Using Combinatorial Techniques,” International Journal of Engineering Technology Vol. No.7 Issue No.1, pp.no.74-81, Jan 2018 (Scopus).

[19]. J.Sasi Bhanu ,M. Lakshmi Prasad and Dr. JKR Sastry, “Testing Embedded System through Optimal Combinatorial Mining Technique,” Journal of Advanced Research in Dynamical and Control Systems, Vol. 10, 07-Special Issue, pp.no.337-354, June 2018 (Scopus).

[20]. J.Sasi Bhanu ,M. Lakshmi Prasad and Dr. JKR Sastry, “Combinatorial Neural Network Based Testing of an

Embedded System," Journal of Advanced Research in Dynamical and Control Systems, Vol. 10, 07-Special Issue, pp.no.605-616, June 2018 (Scopus).

[21]. J.Sasi Bhanu,M. Lakshmi Prasad and Dr. JKR Sastry, "Testing Embedded Systems Using a Graph Based Combinatorial Method (GBCM)," Journal of Advanced Research in Dynamical and Control Systems, Vol. 10, 07-Special Issue, pp.no.355-375, June 2018 (Scopus).

[22]. J.Sasi Bhanu ,M. Lakshmi Prasad and Dr. JKR Sastry, "A Combinatorial Particle Swarm Optimization (PSO) for Testing an Embedded Systems," Journal of Advanced Research in Dynamical and Control Systems, Vol. 10, 07-Special Issue, pp.no.321-336, June 2018 (Scopus).