STM JOURNALS

Current Trends in
Information Technology

https://journals.stmjournals.com/ctit/

ISSN: 2249-4707 (Online)
ISSN: 2348-7895 (Print)
Volume 14, Issue 1, 2024
January–April
DOI (Journal): 10.37591/CTIT

Review CTIT

Machine Learning Based House Price Forecasting

Santushti Betgeri[1], Dimple Thakar[2,*], Monisha Mohan[3]

Abstract

This research endeavours to craft a predictive model leveraging machine learning to estimate the market value of houses in Delhi. By integrating Python and its powerful libraries, pandas for data processing, Plot for interactive visualizations, scikit-learn for implementing machine learning algorithms, XGBoost for boosting the model's prediction accuracy, and to evaluate the model's performance cross-validation techniques are used. An interactive user interface is created using a Flask web application to enter characteristics of a house and according to that application will forecast the price of house. This project *sets out to equip users with a dynamic tool for determining house prices based on essential property attributes. The initiative underscores the potential of machine learning technologies in transforming the real estate sector by enabling more precise property valuation, enhancing market analysis, and bolstering investment and risk assessment strategies. Through the application of sophisticated data analysis and predictive modelling techniques, the project aims to provide valuable insights for real estate professionals, investors, and analysts, facilitating informed decision-making and fostering profitable investment opportunities.*

Keywords: Statistical models, XGBoost model, statistical models, data analytics, risk assessment

INTRODUCTION

One of the most vital and dynamic industries in every economy is real estate. House price prediction is a challenging endeavour with broad ramifications for buyers, sellers, investors, and policymakers. The blending of data science and machine learning methodologies has opened the door for creative approaches to this problem. In this project, the XGBoost algorithm and the Flask web framework are used to create a house price prediction model for properties in Delhi, India [1]. Data has become a useful resource in the modern era of information. The availability of extensive statistics pertaining to various housing qualities has created possibilities for the development of predictive models that can support making well-informed judgments. Our research makes use of a dataset obtained from Kaggle that includes elements such as the number of bedrooms, baths, living space, and lot size, existence of a shoreline, views, and numerous other features that are essential in deciding property pricing [2]. We preprocess and filter the data using libraries like pandas and scikit-learn to provide a solid foundation for our predictive model. Due to its capacity to recognize complex relationships within the data, manage missing values, and reduce over fitting, the XGBoost algorithm stands out as a top pick for regression problems. To build a predictive model for assessing property values using the provided features, our project employs XGBoost. We emphasize the significance of evaluating the model and utilize metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the R-squared score to gauge the accuracy and performance of the model. We expand our efforts beyond model development to provide a user-friendly interface that makes the predictive capability of our model accessible to a wider audience. We created an interactive online application using the Flask web framework that

***Author for Correspondence**
Dimple Thakar
E-mail: dimple.thakar@marwadieducation.edu.in

[1-3]Assistant Professor, Department of Computer Applications, Marwadi University, Rajkot, Gujarat, India

Received Date: March 01, 2024
Accepted Date: March 12, 2024
Published Date: April 03, 2024

Citation: Santushti Betgeri, Dimple Thakar, Monisha Mohan. Machine Learning Based House Price Forecasting. Current Trends in Information Technology. 2024; 14(1): 5–11p.

allows users to enter a property's characteristics and receive an estimated house price in return [3]. This fusion of web development with machine learning technologies is an example of how different fields are coming together to solve problems in the current world.

Dataset

The dataset used for this project includes relevant characteristics that affect home prices, such as the number of bedrooms, bathrooms, living space, lot size, number of floors, presence of a waterfront, views, house condition, grade, different area measurements, year of construction and renovation, postal code, and closeness to airports and schools [4].

The dataset is split into training and testing sets using scikit-learn's train-test split function after pre-processing and feature engineering. The XGBoost model was chosen because it can efficiently handle missing values and grasp complicated relationships in the data. The model is assessed on training data through the application of metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared score [5].

BLOCK DIAGRAM

Block Diagram for Model Development and Evaluation

The given Figure 1 represents the flow diagram of block diagram for model development and evaluation.

LITERATURE SURVEY

House price forecasting has long been a topic of interest in both academia and business. Researchers and practitioners have investigated several ways to create precise models for house price estimate since the introduction of machine learning and the availability of massive datasets [6]. The reach and utility of such models have also been significantly enhanced by the incorporation of online technology. An overview of pertinent research in the fields of housing price prediction, machine learning, and web application development is given in this literature assessment.

Due to the real estate sector's explosive growth, it is urgently necessary for all parties involved in the decision-making process to evaluate and forecast property prices using mathematical models and scientific approaches. In this regard, the use of machine learning algorithms for predicting property prices has attracted a lot of attention recently [7]. This review of the literature looks at numerous

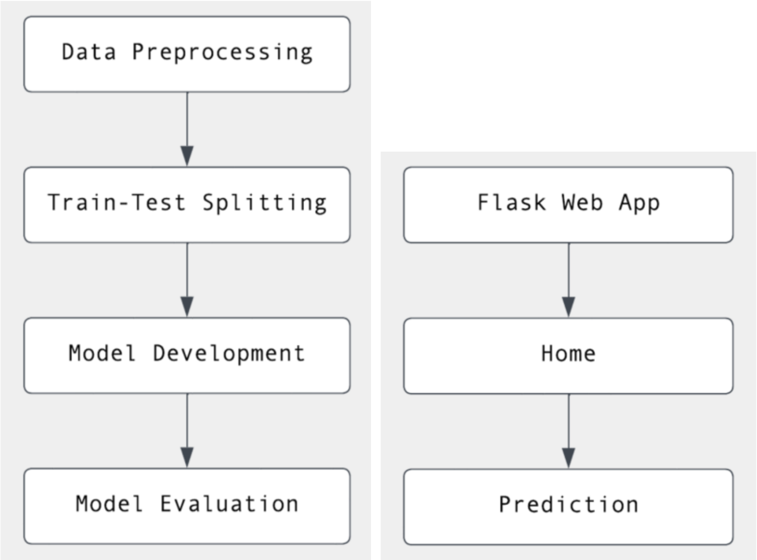


Figure 1. Model development and evaluation and flask web app.

researches that have predicted property prices in various locations across the world using regression analysis and machine learning methods like linear regression, Random Forest, and CNN. The assessment also emphasises how crucial it is to consider a variety of elements when creating property price prediction models, including topographical features, localities, and physical conditions. The results of these studies can help players in the real estate sector make wise judgments [8, 9].

In the research paper titled "House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis" by Chen, the author elucidates the prediction of property prices in Zhaoqing City spanning from 2010 to 2018 [10]. This is achieved through the application of multiple linear regression and Pearson coefficient correlation analysis. In order to determine the goodness of fit R2, the study identified variables that have a substantial correlation with home prices. A multiple linear regression model was applied to examine these variables. This method was employed to assess the disparity between the predicted and actual house prices for the years 2019 and 2020, resulting in |D|. The observation of the prediction effect involved combining the difference |D| between projected and actual home prices and the goodness-of-fit R2 value. Scholars have employed this strategy of employing linear regression models to forecast house prices in various locations throughout the world [10].

Ghosalkar and Dhage, in their study titled "Real estate value prediction using linear regression", investigated the impact of three crucial variables: physical conditions, philosophical beliefs, and geographical locations, on housing values in Mumbai, India. To forecast housing prices in the chosen area, the researchers choose to apply linear regression. It is noteworthy that the researchers' analysis did not take market pricing or cost growth into account. Ghosalkar and Dhage's model concentrates entirely on the effects of the three identified components, giving insights into their relative relevance in influencing Mumbai property values [11].

The paper titled "Housing Price Prediction Using Supervised Learning" by Mahale *et al.* details the viability of predicting the buying and selling prices of real estate properties. This prediction incorporates factors such as location, living space, number of rooms, and other relevant variables. Additionally, Mahale *et al.* highlighted the inclusion of geographical aspects, such as the proximity to the nearest police and fire stations. Mahale *et al*. combined Random Forest and CNN techniques to make this forecast [12].

In the study titled "House Price Forecasting Using Data Mining" by Bhagat *et al.*, the linear regression algorithm was utilized to predict property values and identify the influencing variables. They examined current data to create precise projections [13].

In the article "Real estate value prediction using multivariate regression models", Manjula *et al.* contended that several factors affect home values. A diverse range of characteristics can be employed, sourced from various outlets, to construct a precise prediction model. This is an important study on feature extraction forecasting house values using visual cues. This involved clustering residences with comparable specifications and costs [14].

The study, "Predicting land prices through statistical and neural network software", by Sampathkumar *et al.* illustrates the utilization of historical trends to predict upcoming land values. The determination of growth or decline rates is derived from these historical patterns. Additionally, the analysis may incorporate economic factors to establish a more accurate correlation. The study also referred to a poll done by 99acres.com [15].

The article "Deciding Between Label Encoding and One-Hot Encoding Methods" explains that one-hot encoding involves transforming a column with categorical data into multiple columns after initially applying label encoding to the column. Depending on which column the value is in, the values in these columns are then transformed to 1s or 0s. This essay appeared in the journal Towards Data Science.

METHODOLOGY USED

The XGBoost technique is used to build a reliable house price prediction model for this project, and the Flask web framework is used to build a user-friendly online application. The procedure can be divided into a number of crucial components, including data preprocessing, model construction, assessment, and web application creation [16].

Data Preprocessing

- Pandas is used to load the dataset, which contains attributes such as the number of bedrooms, bathrooms, living areas, lot areas, and more.
- To deal with missing values, outliers, and discrepancies, data cleaning is done.
- To make categorical variables acceptable for the XGBoost model, they are encoded using methods such as one-hot encoding or label encoding.
- To maintain a consistent scale for all features, one can employ feature scaling or normalization techniques.

Train-Test Split

- Utilizing scikit-learn's train-test split function, the pre-processed data set is divided into training and testing sets. A standard split ratio, 70–30, is employed.

Model Development (XGBoost)

- Due to its robustness in handling complex relationships and its effectiveness in handling missing data, the XGBoost algorithm was chosen.
- The training dataset is used to train the XGBoost model, with features serving as the input and housing prices serving as the target variable.
- To improve model performance, hyper parameter tuning can be done using strategies like grid search or random search.

Model Evaluation

- Different metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the R-squared score are employed for evaluating the performance of the trained XGBoost model.
- To evaluate the model's performance in terms of generalization and stability, cross-validation techniques may be used.

Flask Web Application Development

- An interactive user interface is created using a Flask web application.
- Routes in the programme deal with user input and model predictions.
- Users can enter property features into input fields on an HTML page that is rendered by the home route.
- The trained XGBoost model is used to handle user inputs before returning the estimated house price via the prediction route.

Error Handling and User Experience

- To handle possible exceptions during user input and prediction, error handling mechanisms are put in place.
- Intuitive design, useful messages, and responsive layouts improve user experience.

Deployment and Integration

- Users can access the Flask application through a web browser because it has been deployed on a web server.
- Users can get precise property price forecasts thanks to integration with the XGBoost model.

RESULT

The created house price prediction model, which is based on the XGBoost algorithm and incorporated into a Flask web application, performs well when projecting Delhi real estate prices. Through a variety of measures and user interactions, the model's precision and usability are assessed.

Model Evaluation Metrics

A thorough set of assessment metrics is obtained to rate the XGBoost model's predicting skills after it has been trained and tested. These metrics shed light on the model's capacity to represent associations between attributes and home values as shown in Figures 2 and 3.

- *The estimated mean square error (MSE)*: which represents the average squared difference between projected and actual housing values, is 22,216,041,242.109.
- *Root mean squared error (RMSE)*: The RMSE value, which represents the average size of prediction errors, is 149,050.465.
- *The computed mean absolute error (MAE)*: which represents the typical absolute difference between anticipated and actual prices, is 77,136.059.
- *Mean Absolute Percentage Error* (*MAPE):* The average percentage difference between expected and actual costs is 14.6%, according to the MAPE.
- *Score for R-squared (R2)*: With an R2 value of 0.836, the model appears to explain 83.6% of the variation in home prices.

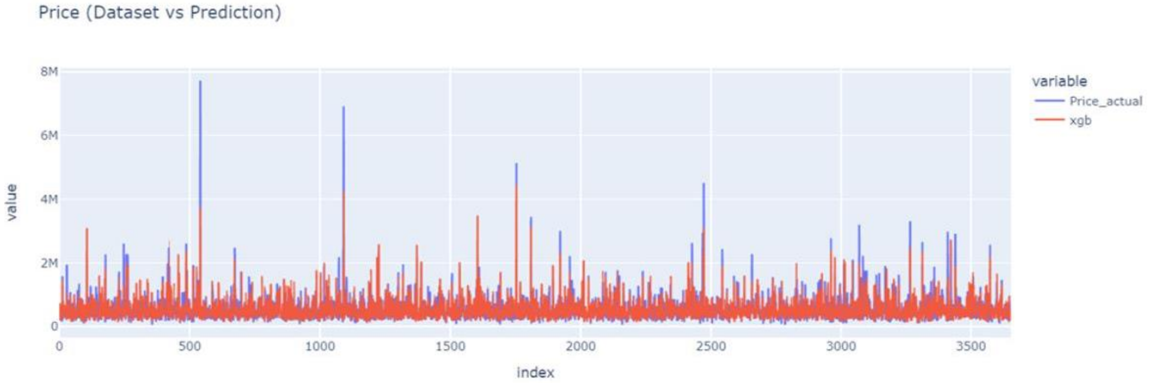
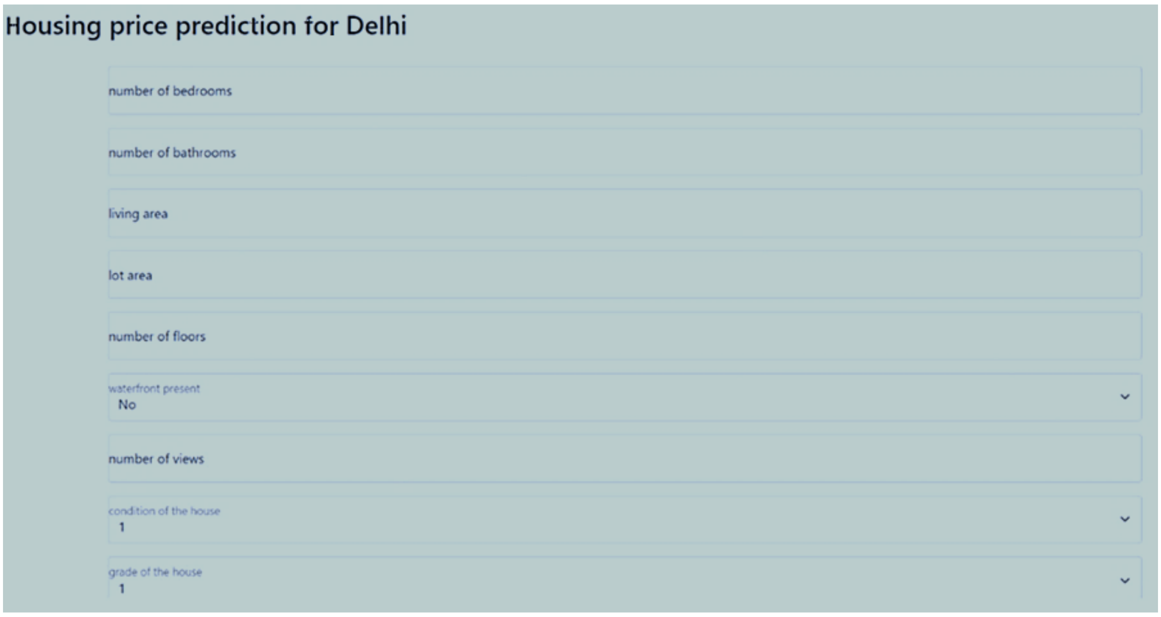


Figure 2. Actual vs. predicted.



Housing price prediction for Delhi

number of bedrooms

number of bathrooms

living area

lot area

number of floors

waterfront present
No

number of views

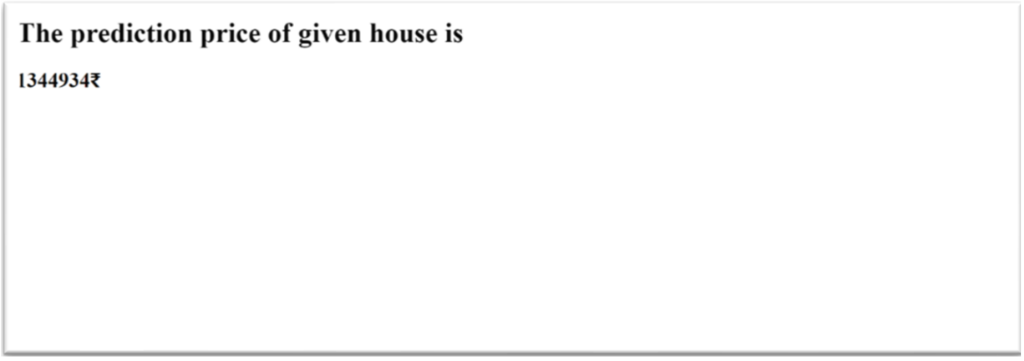
condition of the house
1

grade of the house
1

Figure 3. Home page.

Web Application and User Interaction

The Flask web application that interacts with the XGBoost model improves accessibility and user experience. Through the interactive interface, users can enter different aspects of a property, and the application uses calculations from the model to generate real-time projections of house prices as shown in Figure 4.



The prediction price of given house is

1344934₹

Figure 4. Prediction page.

CONCLUSION

The combination of machine learning and web development has produced a potent and creative way to forecast home prices while opening up this technology to a larger audience. By utilising the XGBoost algorithm and the Flask web framework, this project attempted the goal of creating a forecast model for house values in Delhi. The accomplishment of this project serves as a testament to the value of interdisciplinary approaches in tackling challenging real-world problems. The raw data was of high quality and dependability by carefully preparing it. The XGBoost algorithm, which is renowned for its ability to grasp complex associations, was used to produce a predictive model that showed excellent accuracy. The model's accuracy in estimating property values was proven by evaluation measures such Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error, and R-squared score. An intuitive Flask web application was used to integrate the prediction model, creating a smooth and engaging experience. The user interface made it simple for people to enter details about a property and instantly receive projections, promoting well-informed real estate decision-making. A smooth and engaging experience was developed by incorporating the predictive model within an approachable Flask web application. Decision-making in the real estate industry was facilitated by the user interface, which made it simple for people to input property features and instantly obtain projections. In conclusion, our research demonstrated admirably the ability of contemporary technologies to address difficult problems. The project created a useful application that enables users to precisely estimate home prices by fusing machine learning methodologies with web development frameworks. Beyond its direct use, the project also demonstrates the transformative power of multidisciplinary methods, establishing a precedent for the fusion of data science and software development in the solution of practical issues.

REFERENCES

1. Afsal M. (2023). House Price dataset of India. [Online]. Kaggle. Available from: https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india/data
2. GfG. (2021). XGBoost. [Online]. Geeks for Geeks. Available from: https://www.geeksforgeeks.org/xgboost/
3. Python. (2024). Welcome to Python.org [Online]. Python. Available from: https://www.python.org/
4. Scikit-learn. (2024). machine learning in Python — scikit-learn 1.4.1 documentation. [Online]. Scikit-learn. Available from: https://scikit-learn.org/stable/
5. Pandas. (2024). Python Data Analysis Library. Pydata. Available from: https://pandas.pydata.org/
6. Palletsprojects. (2024). Welcome to Flask: Flask Documentation (3.0.x). [Online]. Available from: https://flask.palletsprojects.com/en/3.0.x/

7. Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC). 2018; 6–10. doi:10.1145/3195106.3195133.
8. Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE). 2018; 35–42. doi:10.1109/icmlde.2018.00017.
9. Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). 2017; 319–323. doi:10.1109/ieem.2017.8289904.
10. Ningyan Chen (Business School, University of Aberdeen, Aberdeen, UK). House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis. Wirel Commun Mob Comput. 2022; 2022: 9590704.
11. Ghosalkar, Dhage. Real estate value prediction using linear regression. In 2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India. 2018; 1–5.
12. Aditi Mahale, *et al*. House price prediction using supervised learning. 3rd International Conference on Advances in Engineering, Technology & Business Management (ICAETBM-2022). 2022.
13. Nihar Bhagat, Ankit Mohorkar, Shreyas Mane. House Price Forecasting using Data Mining. Int J Comput Appl. 2016; 152(2): 23–26.
14. Manjula R, *et al*. Real estate value prediction using multivariate regression models. IOP Conf Ser: Mater Sci Eng. 2017; 263(4): 042098.
15. Sampathkumar V, *et al*. Forecasting the land price using statistical and neural network software. Procedia Computer Science. 2015 Dec; 57:112-121
16. Raheel S. (2018 Nov 9). Choosing the right encoding method Label vs One hot encoder. Towards data science. [Online]. https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b