2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) | 979-8-3503-6472-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/AECE62803.2024.10911214

House Price Prediction Using Machine Learning

Saurabh Pratap Singh Rathore
Dept. of Management ICAPSR
Delhi, India
rathoresaurabhsingh@gmail.com

Mohammed Akram Khan
Dept. of
Madhav University
Sirohi, India
akramkhan@madhavuniversity.edu.in

Saurabh Kumar
Dept. of Management Mangalayatan
University Aligarh, India
saurabh.kumar@mangalayatan.edu.in

Arvind Hans
Dept. of Management Usha Martin
Univrsity Ranchi, India
arvind@umu.ac.in

Pallavi Gangwar
Dept. of Education
Lingayas Vidyapeeth
Faridabad, India
rsdhanusha@gmail.com

Chinmay Jain
Dept. of CSE
Chandigarh University
Perambalu, India
chinmayjain4804@gmail.com

Abstract—This study presents a machine learning algorithm aimed at predicting home values in the housing market. By utilizing a comprehensive dataset that includes details such as location, size, number of bedrooms, and bathrooms for each house, the model goes through several preprocessing steps. These steps include handling missing values, encoding categorical variables, and managing outliers. Feature selection techniques are applied to identify the most significant attributes. To ensure robust model evaluation, the dataset is divided into training and testing subsets. A comparative analysis of various machine learning algorithms is conducted, including Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, SVR, Decision Tree, Gradient Boosting Regressor, Extra Trees Regressor, K Neighbors Regressor, Dense Deep Learning, Regularization, Deep Learning, Random Forest Regressor, XGBoost Regressor, and Polynomial Regression (Degree=2). Evaluation metrics such as mean squared error, mean absolute error, and R-squared are used to assess model performance. Ensemble techniques, particularly Random Forests and Gradient Boosting, consistently demonstrate strong predictive capabilities, outperforming simpler models like Linear Regression with lower prediction errors and higher R- squared values. Neural networks, known for their ability to learn intricate patterns, also show success, especially in handling complex datasets. The gradient boosting model is further optimized through hyperparameter tuning, highlighting its capability to capture complex correlations within the data and generate accurate housing price predictions.

Keywords—Machine learning model, encoding, missing values, feature selection, random forests, gradient boosting, linear regression, neural networks, R-squared value, hyperparameter tuning, house price predictions.

I. INTRODUCTION

Considering its complexity and volatility, projecting home values in the real estate market is challenging. Recent developments in machine learning methods, however, provide encouraging answers to this problem. In order to produce accurate and insightful price estimations, this work intends to propose a house price prediction model that makes use of machine learning methods. The model intends to aid real estate professionals, homeowners, and investors make wise decisions in the dynamic real estate market by utilizing data and computational tools. The primary aim of this research endeavor is to devise a dependable and accurate prognostication technique for approximating residential property worth by leveraging a variety of property attributes.

Traditional techniques of price assessment sometimes ignore the complex interrelationships that affect property prices in favor of oversimplified measures and historical patterns. Large datasets may be analyzed using machine learning techniques, which can also identify subtle patterns and nonlinear correlations that affect pricing. An exhaustive dataset encompassing a plethora of factors, ranging from geographical coordinates, dimensions, to the count of bedrooms and bathrooms, along- side other relevant details, is utilized to achieve this objective. These characteristics add up to a property's distinctive value proposition as a whole and have a big influence on market value. The predictive model is developed over the course of numerous stages.

Initially the data is prepared to correct flaws such missing values and outliers and converting categorical variables into an appropriate format for analysis. Then, the most important properties are determined using feature selection techniques, and they are included to the prediction model. In the model evaluation step, the effectiveness of various machine learning methods is contrasted. These algorithms combine both conventional methods like linear regression and more sophisticated ones like gradient boosting, random forests, decision trees, and support vector 2 machines. Utilizing well-established criteria like mean squared error, mean absolute error, and Rsquared, the efficacy of each method is evaluated. The gradient boosting algorithm has greater prediction powers, according to preliminary studies.

The dataset's complex associations may be captured by this technique, which also produces the lowest prediction errors and greatest R-squared scores. The performance of the gradient boosting model is further optimized via hyper parameter adjustment. In conclusion, this work proposes a machine learning-based house price prediction model that tackles the issues brought on by the dynamic real estate market. The model demonstrates accuracy and dependability in projecting home values by taking into account a wide range of property features and utilizing cutting-edge computational techniques. The model may be used by real estate experts seeking precise estimates, by homeowners determining the value of their homes, and by investors making calculated judgments. This model offers chances for future improvement through data enrichment and the incorporation of cutting-edge approaches as the real estate sector continues to change.

II. LITERATURE REVIEW

Mora-Garcia et al. (2022) explore machine learning for housing price prediction during COVID-19, addressing how pandemic-related factors impact real estate prices [1]. Their study emphasizes ML's predictive power in uncertain market conditions.

979-8-3503-6472-9/24/$31.00 ©2024 IEEE

Agarwal and Vij (2024) investigate AI's role in Indian education, highlighting the challenges in adoption and potential for enhancing learning outcomes [2]. The paper outlines the educational opportunities AI offers in resource limited settings.

Prakash (2024) examines how big data analytics can improve organizational performance, focusing on data-driven decision-making [3]. The study emphasizes big data's potential to boost efficiency and strategic insight.

Ahtesham et al. (2020) analyze machine learning techniques for predicting house prices in Karachi, considering local market factors [4]. Their research demonstrates ML's applicability to real estate forecasting in emerging markets.

Saraswat, Saxena, and Vashist (2024) explore machine learning for electronic health record management in blockchain-cloud frameworks, enhancing EHR security and accessibility [5].

House Price Prediction using Machine Learning (2019) examines various machine learning models for predicting housing prices, focusing on model accuracy and applicability [6]. The study underscores the effectiveness of ML in real estate forecasting.

Prakash (2024) highlights blockchain's role in supply chain management, discussing how it improves transparency and operational efficiency [7]. The paper illustrates blockchain's transformative potential in logistics and data security.

V. Sharma (2022) studies data scaling methods crucial for machine learning, evaluating techniques that enhance model accuracy and stability [8]. The work highlights the importance of preprocessing in achieving reliable ML results.

Sharma, Harsora, and Ogunleye (2024) propose an optimized XGBoost algorithm for house price prediction, emphasizing enhanced accuracy and computational efficiency [9]. This study presents XGBoost as a competitive choice for real estate analytics.

Chaurasia and Haq (2023) explore machine learning models tailored for housing price prediction, with a focus on model selection and training effectiveness [10]. Their research sup- ports ML's adaptability in real estate valuation.

Kaushik et al. (2024) present PneumoAI, an advanced ML model for pneumonia detection, showcasing increased diagnostic accuracy [11]. This study emphasizes AI's growing role in medical imaging and diagnostics.

Li (2024) uses machine learning to predict house prices, focusing on model accuracy and scalability in diverse housing markets [12]. The study showcases the benefits of ML for precise property valuation.

Singh, Rastogi, and Rajpoot (2021) explore machine learning techniques for predicting house prices, emphasizing the comparative performance of algorithms in improving price accuracy and reliability [13].

Kaushik et al. (2024) present an AI model for skin cancer classification, achieving dermatologist-level accuracy [14]. This study demonstrates AI's impact in enhancing diagnostic precision in dermatology.

Annamoradnejad & Annamoradnejad (2022) discuss machine learning applications in housing price prediction, review ing various ML techniques and their practical implications for data-driven real estate markets [15].

Sharma et al. (2023) propose a machine learning algorithm for house price prediction, focusing on optimizing model selection and training [16]. Their study highlights ML's potential in dynamic housing markets.

Mubarak et al. (2022) introduce a map-based recommendation and price prediction model, enhancing real estate decision making with spatial data integration [17].

Sagala & Cendriawan (2022) apply linear regression to house price prediction, providing a simple yet effective model for estimating property values in developing markets [18].

Kuraishi (2023) examines big data and AI's role in Lithuanian supply chains, focusing on analytics-driven operational improvements and strategic insights in logistics [18].

Kaushik et al. (2024) discuss security and privacy in intelligent transportation systems, proposing trusted models to safeguard user data in smart transit networks [20].

This collection highlights the range of AI and ML applcations in fields from real estate to healthcare and supply chains. Let me know if further detail is needed for any specific reference.

III. METHODOLOGY

A. Dataset

The dataset was downloaded and imported from Kaggle and further it has two more datasets namely 'train.csv' and 'test.csv'.

B. Handling Outliers

The outliers for the train dataset were visualized using box and whisker plots.

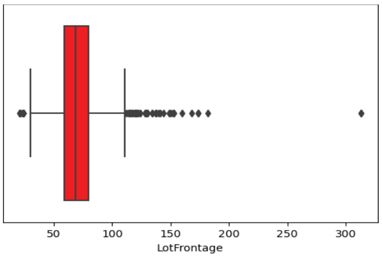

Fig. 1. LotFrontage column from training dataset plotted for outlier detection and visualization

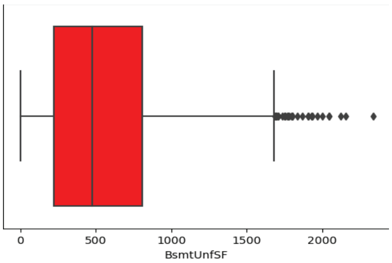

Fig. 2. BsmtUnfSF column from training dataset plotted for outlier detection and visualization

Dimensions for original train dataset were (1460, 81). After removing the outliers, the dimensions are (1101, 81).

C. Data Preprocessing

In this the train and test datasets were concatenated with (2560, 81) as the final dimensions. After concatenation the missing values were examined and handled. Total of 9 columns with missing values were filled with 2's and 0's. Replacing and changing the dataset with datasets having label encoding, dummy variable columns, principal component analysis, columns having more than 0.05 correlation with SalePrice column and feature selection as df le, df dum, df pca, df cor and df fs respectively. The df pca dum, df cor dum and df fs dum were created based on the df dum dataset as it was more accurate as compared to the df le dataset due to a greater number of columns.

D. Model Implementation

For the prediction of house prices total 14 machine learning models are implemented. The baseline architecture for the model implementation follows the concept of Transfer Learning. In order to find the evaluation metrics for various regression models that were implemented, we defined the evaluate() function. The metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared were further defined in the evaluate() function for each model. Then, using transfer learning the pre- defined models were trained on the dataset that was created via combining the train and test datasets. After training the models the evaluate() function was called to calculate the regression metrics for each model and the metrics for df le, df dum, df pca dum, df cor dum and df fs dum datasets were also evaluated for all the models.

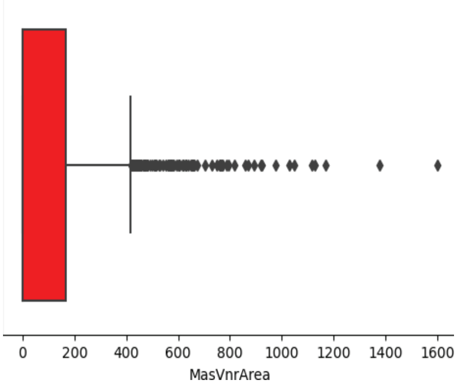

Fig. 3. MasVnrArea column from training dataset plotted for outlier detection and visualization

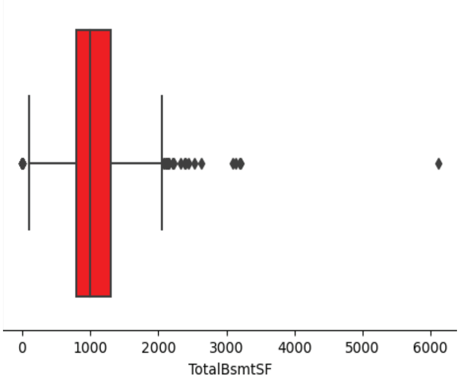

Fig. 4. TotalBsmtSF column from training dataset plotted for outlier detection and visualization

The following models were implemented

1) Linear Regression

2) Ridge Regression

3) Lasso Regression

4) Polynomial Regression

5) ElasticNet

6) Support Vector Regressor

7) Decision Tree

8) Random Forest

9) Gradient Boosting Regressor

10) Extra Trees Regressor

11) XGBoost (Extreme Gradient Boosting Regressor)

12) KNeighbors Regressor

13) Regularization + Deep Learning

14) Just Deep Sense Learning

According to the feature importance tally, the most important and decisive features for some models are as follows:

• Decision Tree –

df le (GarageArea)

df dum (TotalBsmtSF)

df fs (OverallQual)

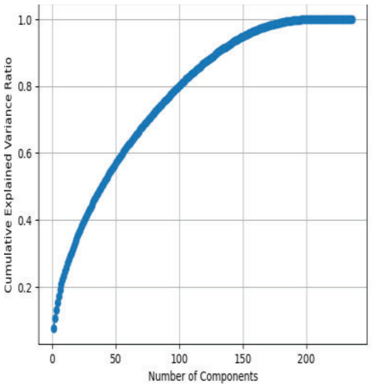

Fig. 5. Cumulative Explained Variance and Number of Components plotted for df pca dataset

df cor (n bathrooms)

• Random Forest –

df le (area with basement)

df dum (area with basement)

df fs (area with basement)

df cor (area with basement)

• Gradient Boosting Regressor –

df le (area with basement)

df dum (BsmtQual)

df fs (OverallQual)

df cor (n bathrooms)

• Extra Trees Regressor –

df le (GarageFinish)

df dum (SaleType New) df fs (Foundation PConc) df cor (GarageCars)

IV. Result

In this research paper we focused on housing price forecast- ing, and examined the effectiveness of 14 different machine learning models. These models encompass a spectrum of methodologies, spanning from conventional approaches like linear regression and decision trees to sophisticated techniques such as support vector regressors, gradient boosting, and deep learning. Using a diverse dataset with characteristics such as property size, location, amenities and historical price trends, the study found that regression models like Linear, Ridge, Lasso, Polynomial regression and ElasticNet along with Gradient Boosting Regressor, XGBoost and SVR, consistently deliver superior predictive performance. The research emphasizes the importance of function development and data preprocessing, while providing the real estate industry with valuable insight into effective housing price forecasting methods.

V. Conclusion

In general, our study comprehensively examines the application of 14 different machine learning models for housing price forecasting. Through careful evaluation, it is clear that regression models like Linear, Ridge, Lasso, Polynomial re- gression and ElasticNet along with Gradient Boosting Regressor, XGBoost and SVR, consistently outperform other models, highlighting their effectiveness in capturing complex patterns in the housing market. We highlight the critical role of feature engineering and data preprocessing in improving the predictive accuracy of all models. Our findings provide valuable insight into the real estate industry and demonstrate the potential for more accurate and reliable house price forecasting. As the need for accurate property valuations valuations continues to grow, the use of these advanced technologies can provide stakeholders with better decision-making tools in this dynamic and changing market.

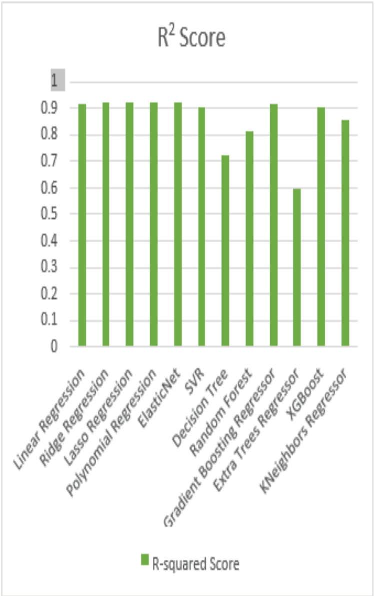

Fig. 6. R-Squared Scores compared for all Models

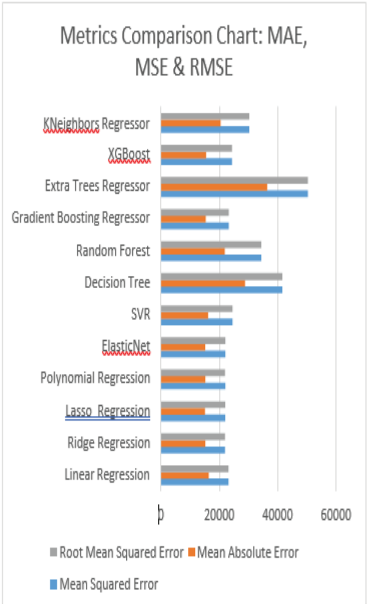

Fig. 7. MAE, MSE and RMSE values compared for all Models

References

[1] R.T. Mora-Garcia, M.F. Cespedes-Lopez, and V. R. Perez-Sanchez, “Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times,” Land, vol. 11, no. 11. MDPI AG, p. 2100, Nov.21, 2022. doi: 10.3390/land11112100.

[2] P. Agarwal and A. Vij, “Assessing the Challenges and Opportunities of Artificial Intelligence in Indian Education,” International Journal for Global Academic & Scientific Research, vol. 3, no. 1. International Consortium of Academic Professionals for Scientific Research, pp.36–44, Apr. 04, 2024. doi: 10.55938/ijgasr.v3i1.71.

[3] D. Prakash, “Data-Driven Management: The Impact of Big Data Analytics on Organizational Performance,” International Journal for Global Academic & Scientific Research, vol. 3, no. 2. International Consortium of Academic Professionals for Scientific Research, pp. 12–23, Jul. 02,2024. doi: 10.55938/ijgasr.v3i2.74.S

[4] M. Ahtesham, N. Z. Bawany and K. Fatima, ”House Price Pre- diction using Machine Learning Algorithm - The Case of Karachi City, Pakistan,” 2020 21st International Arab Conference on In- formation Technology (ACIT), Giza, Egypt, 2020, pp. 1-5, doi:10.1109/ACIT50332.2020.9300074.

[5] B. K. Saraswat, A. Saxena, and P. C. Vashist, “Machine learning for effective EHR management in blockchain-cloud integration,” Journal of Autonomous Intelligence, vol. 7, no. 4. Frontier Scientific Publishing Pte Ltd, Feb. 02, 2024. doi: 10.32629/jai.v7i4.1274.

[6] “House Price Prediction using Machine Learning,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 9. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 717–722, Jul. 30, 2019. doi: 10.35940/iji- tee.i7849.078919.

[7] A. Prakash, “Blockchain Technology for Supply Chain Management: Enhancing Transparency and Efficiency,” International Journal for Global Academic & Scientific Research, vol. 3, no. 2. International Consortium of Academic Professionals for Scientific Research, pp.01–11, Jul. 02, 2024. doi: 10.55938/ijgasr.v3i2.73.

[8] V. Sharma, “A Study on Data Scaling Methods for Machine Learning,” International Journal for Global Academic & Scientific Research, vol. 1, no. 1. International Consortium of Academic Professionals for Scientific Research, Feb. 23, 2022. doi: 10.55938/ijgasr.v1i1.4.

[9] H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm: XGBoost,” Analytics, vol. 3, no. 1. MDPI AG, pp. 30–45, Jan. 02, 2024. doi: 10.3390/analytics3010003.

[10] A. Chaurasia and I. U. Haq, ”Housing Price Prediction Model Using Ma- chine Learning,” 2023 International Conference on Sustainable Emerg- ing Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 497-500, doi: 10.1109/ICSEIET58677.2023.10303359.

[11] P. Kaushik, M. Arora, Y. Sharma, M. Poonia, P. Kumawat, and R. S.

[12] Charak, “PneumoAI: Redefining accuracy in pneumonia detection using advanced machine learning,” in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2024.

[13] C. Li, “House price prediction using machine learning,” Applied and Computational Engineering, vol. 53, no. 1. EWA Publishing, pp.225–237, Mar. 28, 2024. doi: 10.54254/2755-2721/53/20241426.

[14] A. P. Singh, K. Rastogi and S. Rajpoot, ”House Price Pre- diction Using Machine Learning,” 2021 3rd International Confer- ence on Advances in Computing, Communication Control and Net- working (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi:10.1109/ICAC3N53548.2021.9725552.

[15] P. Kaushik, Y. Chopra, A. Kajla, M. Poonia, A. Khan and D.Yadav, ”AI-Powered Dermatology: Achieving Dermatologist-Grade Skin Cancer Classification,” 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2024, pp. 1-6, doi:10.1109/IATMSI60426.2024.10502664.

[16] R. Annamoradnejad and I. Annamoradnejad, “Machine Learning for Housing Price Prediction,” Encyclopedia of Data Science and Machine Learning. IGI Global, pp. 2728–2739, Oct. 14, 2022. doi: 10.4018/978-1-7998-9220-5.ch163.

[17] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, ”House Price Prediction using Machine Learning Algorithm,” 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986, doi: 10.1109/ICCMC56507.2023.10084197.

[18] M. Mubarak et al., “A Map-Based Recommendation System and House Price Prediction Model for Real Estate,” ISPRS International Journal of Geo-Information, vol. 11, no. 3. MDPI AG, p. 178, Mar. 07, 2022. doi:10.3390/ijgi11030178.

[19] N. T. M. Sagala and L. H. Cendriawan, ”House Price Prediction Using Linier Regression,” 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia,2022, pp. 1-5, doi: 10.1109/ICCED56140.2022.10010684.

[20] Z. Mohammad Kuraishi, “Role of Analytics in Supply Chain Management Industry in Lithuania: Big Data Analytics & AI,” International Journal for Global Academic & Scientific Research, vol. 2, no. 4. International Consortium of Academic Professionals for Scientific Research, pp. 44–53, Dec. 31, 2023. doi: 10.55938/ijgasr.v2i4.65.

[21] P. Kaushik, S. P. S. Rathore, L. Sachdeva, M. Poonia, D. Singh, and L. Bir, “Intelligent transportation systems trusted user’s security and privacy,” in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2024.