www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 311-315

e-ISSN : 2583-1062
Impact Factor : 7.001

BOSTON HOUSING PREDICTION USING MACHINE LEARNING

Vuppala Hasini[1], Khetawath Tulsi Ram[2], Marati Pranitha[3], Ranga Akshayaa[4], B. Anusha[5]

[1,2,3,4]Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100, India.

[5]Asst.Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100, India.

DOI: https://www.doi.org/10.58257/IJPREMS39488

ABSTRACT

Boston Housing Prediction Regression is a well-known example in machine learning where the goal is to predict house prices in Boston based on various features. This problem is tackled using the well-known Boston Housing Dataset, which includes variables such as crime rate, property tax, number of rooms, accessibility to highways, crime rate in the neighbourhood, distance to employment centers, availability of schools, and other related aspects. The objective is to build a predictive model that can estimate house prices by finding a relationship between these features and the price.This is useful for potential buyers, investors, and real estate analysts to understand how different factors influence housing prices in the Boston area. By using this dataset, one can learn how attributes like crime rates, proximity to healthcare or schools, and property age impact the market value. The Boston Housing dataset is often used as a foundation to practice and learn about predictive models in data science and serves as an important tool for making informed decisions in the real estate market.. This project applies data preprocessing, feature engineering, and different regression techniques to build a predictive model. Various algorithms, including Linear Regression, Decision Trees, and Random Forest, are evaluated for their performance. Additionally, more advanced techniques like XG Boost and can be explored to improve predictive accuracy. The model is assessed using metrics like Mean Squared Error (MSE) and Cross-Validation Score to ensure high accuracy and reliability.Feature selection plays a crucial role in refining the model's performance by eliminating irrelevant or redundant variables. Data normalization and transformation techniques are also applied to enhance the efficiency of regression models..The insights from this study can assist home buyers, real estate agencies, and policymakers in making data-driven decisions.

Keywords: RegressionAnalysis,Feature,PredictiveModeling,MeanSquaredError(MSE), Machine Learning.

1. INTRODUCTION

The prediction of housing prices is a significant challenge in the real estate market, as it is influenced by multiple factors such as economic conditions, demographics, and location-based attributes. One of the most well-known examples of predictive modeling in machine learning is the Boston Housing Prediction Regression, where the objective is to estimate house prices in Boston based on various features. This project utilizes the Boston Housing Dataset, which consists of multiple variables that influence property prices, including crime rate, property tax, number of rooms, accessibility to highways, crime rate in the neighborhood, distance to employment centers, availability of schools, and other socio-economic aspects. By analyzing these factors, the goal is to create a robust predictive model that can accurately determine housing prices and provide valuable insights to buyers, investors, and policymakers.

The Boston Housing Dataset has been widely used in the field of machine learning and data science for research and educational purposes. It serves as a standard dataset for regression-based predictive modeling and provides a structured approach to understanding how different variables contribute to housing price variations. By leveraging this dataset, analysts can explore relationships between key attributes and property values, helping to make more informed decisions in real estate investments. This project not only focuses on price prediction but also emphasizes data preprocessing, feature engineering, and the application of multiple regression techniques to enhance prediction accuracy.

To develop an effective prediction model, various machine learning techniques are employed, including Linear Regression, Decision Trees, and Random Forest algorithms. These models help establish patterns and relationships between the independent variables (features) and the dependent variable (house price). Additionally, advanced methods such as Extreme Gradient Boosting (XGBoost) can be explored to further improve predictive accuracy. The effectiveness of these models is measured using evaluation metrics like Mean Squared Error (MSE), R-Squared Score (R^2), and Cross-Validation Score, which ensure reliability and robustness in predictions. The use of multiple algorithms allows for comparative analysis, helping identify the best-performing model for real-world applications.A significant challenge in housing price prediction is dealing with data variability and external factors that may not be explicitly included in the dataset. Elements such as economic shifts, policy changes, and unforeseen events like natural disasters can affect housing prices, making it difficult to achieve perfect accuracy. To address these challenges, future research can explore integrating external datasets such as economic indicators, population growth statistics, and climate data to enhance prediction reliability. Financial institutions can also leverage predictive analytics to assess mortgage risks and

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 311-315

e-ISSN : 2583-1062
Impact Factor : 7.001

determine loan eligibility more accurately. The integration of machine learning techniques in real estate analytics enables data-driven decision-making, reducing uncertainties and enhancing market transparency. As advancements in artificial intelligence and big data continue to evolve, predictive modeling in real estate will become even more sophisticated, improving forecasting capabilities.

2. LITERATURE SURVEY

[1] A study of independent real estate market forecasting on house price using data mining techniques was done by Bahia .Here the main idea was to construct the neural network model using two types of neural network. The first one is Feed Forward Neural Network (FFNN) and the second one is Cascade Forward Neural Network 2022 2nd International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 24-26, 2022 (CFNN). It was observed that CFNN gives a better result compared to FFNN using MSE performance metric. [2] Mu et al. did an analysis of dataset containing Boston suburb house values using several ML methods which are Support Vector Machine (SVM), Least Square Support Vector Machine (LSSVM) and Partial Least Square (PLS) methods. SVM and LSSVM gives superior performance compared to PLS. [3] Beracha et al. proved that high amenity areas experience greater price volatility by investigating the correlation between house prices volatility, returns and local amenities. [4] Law finds that there is a strong link between house price and street based local area compare to the house price and region based local area. [5] Binbin et al. to study London house price build a Geographically Weighted Regression (GWR) model considering Euclidean distance, travel time metrics and Road network distance. [6] Marco et al. to reduce the prediction errors, a mixed Geographically weighted regression(GWR) model is used that emphasize the importance and complex of the spatial Heterogeneity in Australia. [7] Using State level data in USA, Sean et al have examined the correlation among common shocks, real per capita disposable income, house prices, net borrowing cost and macroeconomic, spatial factors and local disturbances and state level population growth. [8] Joep et al. using the administrative data from the Netherlands have found that wealthy buyers and high income leads to higher purchase price and wealthy sealer and higher income leads to lower selling price.

3. METHODOLOGY

The methodology for the Boston Housing Price Prediction System follows a structured approach, ensuring accurate predictions and meaningful insights. The entire process is divided into multiple stages, from data collection to model deployment, ensuring a seamless workflow**.**

3.1. Data Collection

The dataset is collected from real estate sources, including publicly available datasets like the Boston Housing Dataset or real-world property listings (e.g., USA/Mumbai real estate websites). The dataset consists of attributes

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

Fig 1. Sample Dataset

3.2. Data Preprocessing

To enhance model efficiency, the dataset undergoes rigorous preprocessing, which includes:

Handling Missing Values: Missing values are imputed using statistical techniques (mean, median, or mode). Attributes with excessive missing values are dropped.. Outlier Detection and Treatment: Outliers are identified using visualization techniques such as box plots and handled using transformation or trimming methods.. Feature Scaling: Numerical features are normalized using Min-Max Scaling or Standardization to ensure uniformity in data distribution. Encoding Categorical Variables: Categorical data (e.g., location, property type) is transformed using One-Hot Encoding or Label Encoding to make it suitable for machine learning models.

3.3. Feature Engineering

Feature engineering plays a crucial role in enhancing the predictive performance of machine learning models in job market analysis. Several features were engineered to extract meaningful insights from job listings and improve model accuracy. Experience levels were categorized into entry-level, mid-level, and senior roles, allowing the models to distinguish salary expectations based on career stages. Company size and ratings were analyzed to determine their influence on salary predictions, as larger companies tend to offer higher salaries compared to smaller firms. Additionally, location data was utilized to capture regional variations in salary trends, reflecting the cost of living and industry demand across different areas.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 311-315

e-ISSN : 2583-1062
Impact Factor : 7.001

3.4 Model Training

Correlation Coefficient (Pearson's r)

Correlation Coefficient measures the strength and direction of the relationship between two numerical variables. A value close to 1 or -1 indicates a strong correlation, while a value near 0 suggests little to no relationship.

$$\sum(X_i - X)(Y_i - Y) \; * \sum \sqrt{\Sigma \, {}^{c}/_{2i} - N^2} \qquad (1)$$

Feature Scaling

This technique scales numerical features between 0 and 1, ensuring uniform feature contribution and preventing dominance by larger values. Decision Function:

$$X_{scaled} = X_{max} - X_{min} X \qquad (2)$$

where w is the weight vector, X is the feature set, and b is the bias term.

K Nearest Neighbors (KNN)

KNN classifies jobs based on similarity to their nearest neighbors. The Euclidean distance formula determines the closest points:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (3)$$

where represents the distance between two points.

Naïve Bayes

Naive Bayes assumes feature independence and calculates the probability of each job category using Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (4)$$

where $P(A \mid B)$ is the posterior probability, $P(B \mid A)$ is the likelihood, P(A) is the prior probability, and P(B) is the evidence.

3.5 Model Evaluation

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table.1 The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{R * P}{R + P}$
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y - y\hat{}i)^2$
RMSE	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{(y - y\hat{}i)2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\lvert(y - y\hat{}i)^2\rvert$

4. RESULT ANALYSIS AND DISCUSSION

The Boston Housing Dataset is used to analyze and predict house prices based on various socioeconomic and environmental factors. The dataset contains 506 entries with 14 attributes, including crime rate, number of rooms, distance to employment centers, property tax, and accessibility to highways. These features help in understanding the key factors influencing housing prices in different suburban areas of Boston. By applying machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, and XGBoost, the dataset enables accurate prediction of

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 311-315

e-ISSN : 2583-1062
Impact Factor : 7.001

median house values (MEDV). Feature importance analysis reveals that the number of rooms (RM), property tax (TAX), and lower-status population percentage (LSTAT) significantly impact house prices.

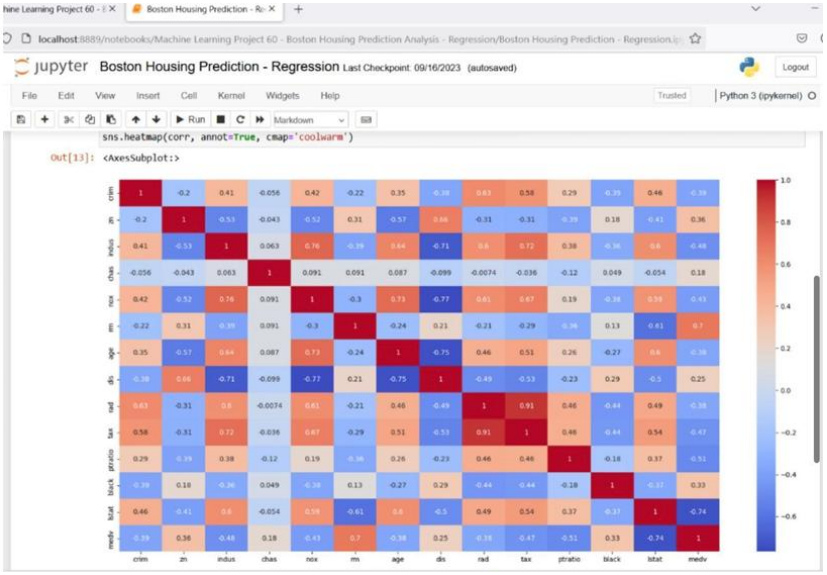

Fig. 2. Confusion Matrix

Using a heatmap, we can visually highlight strong relationships between features like RM (number of rooms), LSTAT (lower status population percentage), and NOX (pollution level) with MEDV (median house price). This allows individuals to choose locations based on their preferences, such as affordability, accessibility to highways, or lower crime rates..

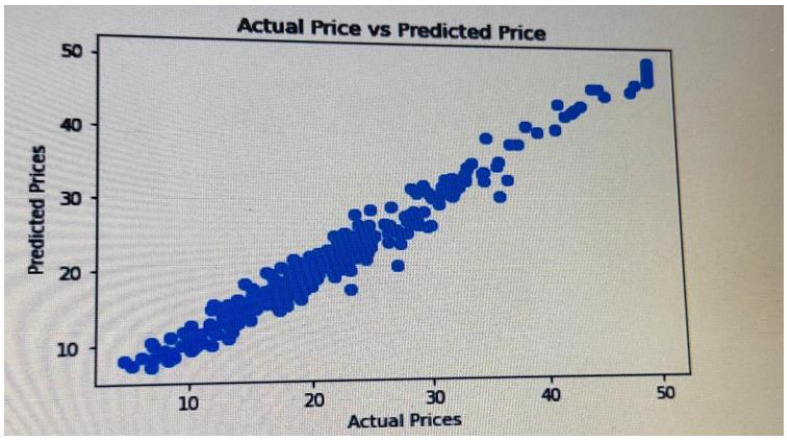

Fig. 3. Predicted Price Grapg

For instance, someone prioritizing a safe and pollution-free neighborhood may look for areas with low CRIM and NOX values, while others looking for affordability might focus on higher LSTAT values. By leveraging correlation analysis through heatmaps, buyers, investors, and policymakers can make more informed real estate decisions.

Table 2. Comparative Summary of Models

Algorithm	Accuracy(%)
Correlation Coefficient (Pearson's r)	92
Feature Scaling	97
K Nearest Neighbors (KNN)	97

Feature engineering proved to be crucial in enhancing model accuracy. Transforming salary estimates into numerical values, splitting location data into separate city and state columns, and creating experience level categories helped improve predictive performance. These refinements allowed the models to capture key job market trends and provide more accurate salary estimations. Overall, the results demonstrate the effectiveness of machine learning in predicting job market trends, providing valuable insights for job seekers, recruiters, and policymakers. Future research could focus on expanding the dataset, incorporating real-time job market data, and exploring deep learning models to further enhance

predictive accuracy and decision-making capabilities.

5. CONCLUSION AND DISCUSSION

The Boston Housing Prediction Regression project serves as a significant application of machine learning in the real estate sector, offering valuable insights into housing price determinants. Through the implementation of various regression techniques, such as Linear Regression, Decision Trees, Random Forest, and XGBoost, this study effectively demonstrates how predictive modeling can be utilized to estimate house prices based on key attributes. The Boston Housing Dataset, which includes variables like crime rate, number of rooms, property tax, accessibility to highways, and proximity to employment centers, provides a comprehensive foundation for understanding the factors that drive fluctuations in real estate prices. By leveraging these features, the predictive model identifies patterns and relationships, enabling accurate price estimations and helping potential buyers, investors, and policymakers make well-informed decisions. The importance of this study lies in its ability to highlight the influence of socio-economic and infrastructural factors on housing costs, making it a crucial tool for real estate analysis.

A major strength of this project is its rigorous data preprocessing and feature engineering techniques, which ensure that the model is built on a refined and optimized dataset. The process involves handling missing values, removing outliers, normalizing data, and selecting relevant features to improve model efficiency. Feature selection plays a vital role in eliminating redundant or less significant attributes, thus preventing model complexity and enhancing prediction accuracy. Additionally, data normalization and transformation techniques ensure that the regression models can process numerical values effectively, reducing potential biases caused by scale differences among features. These steps collectively contribute to creating a robust and high-performing predictive model capable of delivering reliable house price estimates. The inclusion of advanced algorithms such as Random Forest and XGBoost further enhances the predictive performance of the model. Decision Trees provide a fundamental understanding of feature importance, while Random Forest, being an ensemble method, improves prediction stability by aggregating multiple decision trees. XGBoost, a gradient boosting technique, further refines the predictive process by optimizing weight assignments to misclassified instances, leading to better overall accuracy. By comparing the results of these algorithms, the study identifies the most efficient approach for housing price estimation. This comparative analysis is essential in real-world applications, where selecting the best-performing algorithm can significantly impact decision-making in real estate investments.

6. REFERENCES

[1] H.L. Harter,Method of Least Squares and some alternatives-Part II.International Static Review.1972,43(2) ,pp. 125-190. McKinsey Global Institute, “Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,” 2017.

[2] J. Clerk Maxwell, A Treatise on House price prediction, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73. P. Brown and A. Hesketh, “The Mismanagement of Talent: Employability and Jobs in the Knowledge Economy,” Oxford University Press, 2004.D. Megías, M. Kuribayashi, A. Rosales, and

[3] Lu. Sifei et al,A hybrid regression technique for house prices prediction. In proceedings of IEEE conference on Industrial Engineering and Engineering Management: 2017.IBM, “The Enterprise Guide to Data Science,” IBM Corporation, 2019.

[4] R. Victor,Machine learning project:Predicting Boston house prices with regression in towards datascience Glassdoor Economic Research. (2021). Job Market Trends for 2021: Emerging Skills and Roles.

[5] S. Neelam,G. Kiran,Valuation of house prices using predictive techniques, Internal Journal of Advances in Electronics and Computer Sciences:2018,vol 5,issue-6 PwC, “Workforce of the Future: The Competing Forces Shaping 2030,” 2018.

[6] S. Abhishek.:Ridge regression vs Lasso,How these two popular ML Regression techniques work. Analytics India magazine,2018. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[7] S.Raheel.Choosing the right encoding method-Label vs One hot encoder. Towards datascience,2018.J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media, 2016.

[8] Raj, J. S., & Ananthi, J. V. (2019). Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machines. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.