HOUSE PRICE PREDICTION

[1] P. Anusha, [2] Badugu Sushmitha, [3] Potta Ankitha, [4] PL. Rajesh
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
anushaparvathagiri@siddhartha.org.in, 23tq1a5603@siddhartha.co.in, 23tq1a5638@siddhartha.co.in, 23tq1a5635@siddhartha.co.in

Abstract

The real estate market is a complex and dynamic sector where accurate property valuation plays a crucial role for buyers, sellers, and investors. Traditional methods of house price estimation are often subjective and may not provide the level of precision required for effective decision-making. This project addresses this challenge by developing a machine learning-based model for predicting housing prices using various property-related features.

The study utilizes the "Housing Prices Dataset" from Kaggle, which consists of 545 samples and 13 features, including area, number of bedrooms, bathrooms, parking availability, and other amenities. These features are used to analyze the relationship between property characteristics and their market prices. The dataset undergoes preprocessing steps such as handling missing values, encoding categorical variables, and scaling to prepare it for model training.

Several regression algorithms are applied, including Linear Regression, Decision Tree Regression, and Random Forest Regression, to predict house prices. The models are evaluated using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score to determine their accuracy and reliability. Among the tested models, ensemble methods like Random Forest generally provide better performance due to their ability to handle complex data patterns.

I. Introduction

The housing market is a fundamental component of the global economy, known for its dynamic nature and sensitivity to a wide range of influencing factors. Property valuation, which involves determining the monetary worth of a property, is inherently complex. It depends not only on intrinsic characteristics such as area, number of bedrooms, and bathrooms, but also on extrinsic factors including location, accessibility to amenities, neighborhood quality, and prevailing market conditions.

For home buyers, sellers, and real estate professionals, determining an accurate and fair property price remains a significant challenge. Traditional valuation methods, such as Comparative Market Analysis (CMA), rely heavily on historical data and human judgment. Although useful, these approaches are often subjective, time-consuming, and may fail to capture the complex and non-linear relationships between various property features. As a result, properties may be either overpriced—leading to extended selling periods—or underpriced, causing financial loss to sellers.

To address these limitations, this project focuses on leveraging machine learning techniques to develop an efficient and accurate house price prediction system. By

analyzing historical housing data and identifying hidden patterns, the proposed model aims to provide reliable price predictions.

II. Literacture Survey

Several research studies have significantly contributed to the development of house price prediction models, evolving from traditional statistical approaches to advanced machine learning techniques. One of the foundational works in this domain is by Harrison, D. and Rubinfeld, D.L. (1978), titled *"Hedonic Prices and the Demand for Clean Air,"* which introduced the concept of hedonic pricing models. This approach decomposes the price of a property into the value of its individual attributes, such as location, size, and environmental factors, forming the basis for modern real estate price analysis. In the following years, classical regression models became widely used for predictive tasks. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), in *"Regression Diagnostics,"* provided essential techniques to detect issues such as multicollinearity and influential data points, which are crucial for building reliable regression models. Linear Regression, Ridge Regression, and Lasso Regression have been extensively applied, with studies such as Hu (2026) highlighting that Lasso Regression often performs better due to its ability to reduce overfitting through feature selection.

With advancements in computational power, ensemble learning methods have gained prominence in recent years. Studies like *"Real Estate Price Prediction Using Regression Techniques"* (2023) demonstrate that models such as Random Forest and XGBoost outperform traditional linear models by effectively capturing non-linear relationships and complex feature interactions. These models have achieved high accuracy levels, with R^2 scores reaching up to 0.85 on datasets like the Ames Housing dataset. Furthermore, research has consistently identified key features influencing house prices, including house area, overall quality, number of bathrooms, garage capacity, and proximity to urban centers and essential amenities. Studies by Liu et al. (2022) and others also emphasize the impact of macroeconomic factors such as income levels and economic conditions on housing prices. Overall, the literature indicates a clear shift toward machine learning and ensemble techniques, which provide more accurate and robust predictions compared to traditional methods.

III. System Analysis

System analysis is the process of examining an existing system, identifying its limitations, and designing an improved system to meet user requirements efficiently. It involves understanding data flow, system components, processing methods, and expected outputs. In predictive systems like house price prediction, system analysis focuses on how data is collected, processed, and used to generate accurate predictions.

It ensures that the developed system is reliable, scalable, and capable of handling real-world data. For machine learning-based applications, system analysis also includes identifying suitable datasets, preprocessing techniques, model selection, and evaluation strategies to achieve optimal performance.

Existing System

In the traditional real estate system, house price estimation is primarily done using manual methods or basic statistical approaches. Real estate agents and property evaluators rely on Comparative Market Analysis (CMA), past sales data, and personal experience to estimate property prices. These methods depend heavily on human judgment and limited data analysis.

Additionally, traditional systems often fail to consider multiple influencing factors simultaneously, such as location advantages, amenities, and complex feature interactions. As a result, predictions are often inaccurate or inconsistent.

Disadvantages of Existing System

- High dependency on human judgment and experience
- Lack of accuracy and consistency in price estimation
- Time-consuming and inefficient process
- Inability to handle large and complex datasets
- Difficulty in capturing non-linear relationships between features

Proposed System

The proposed system is a machine learning-based house price prediction model that provides accurate and data-driven property valuation. It uses historical housing data, including features such as area, number of bedrooms, bathrooms, parking, and other amenities, to train predictive models.

The system follows a structured pipeline that includes data preprocessing (handling missing values, encoding categorical variables, and scaling), exploratory data analysis, feature selection, and model training. Various regression algorithms such as Linear Regression, Decision Tree, and Random Forest are applied to predict house prices.

Advantages of Proposed System

- High accuracy in predicting house prices
- Automated and fast price estimation process
- Ability to handle large and complex datasets
- Captures non-linear relationships using advanced algorithms
- Reduces human bias and errors

IV. Methodology

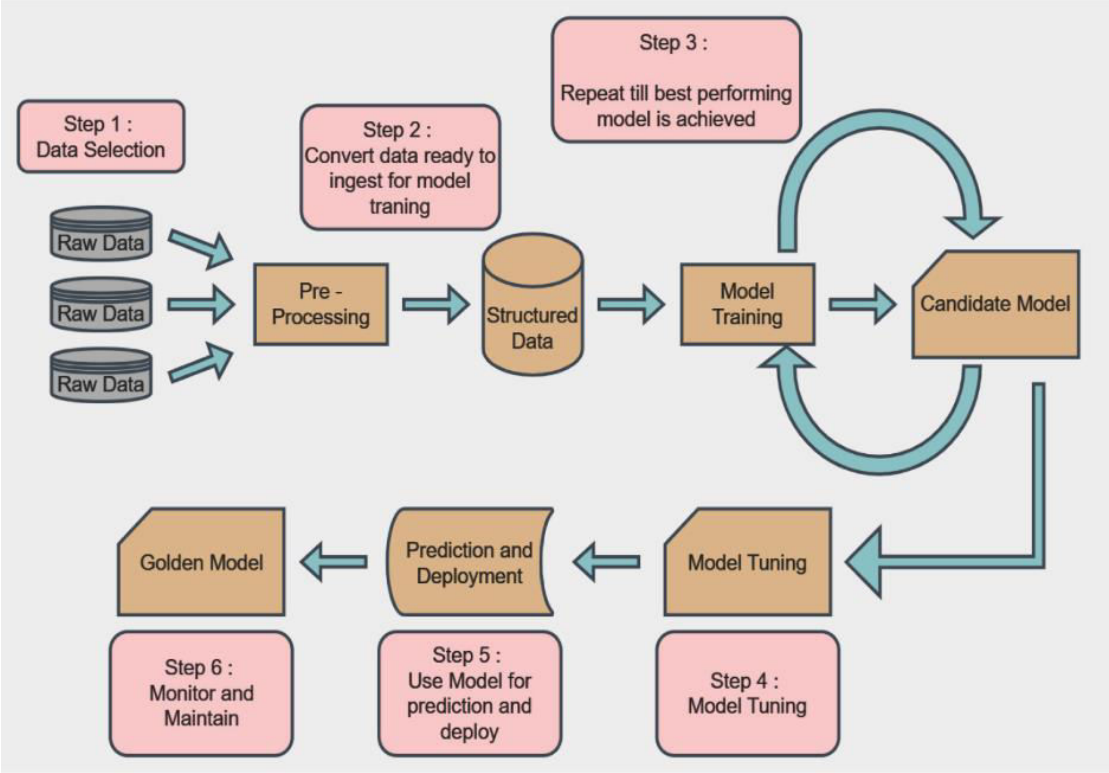
The methodology for the house price prediction system follows a systematic approach to build an accurate and reliable predictive model using machine learning techniques. The process begins with **data collection**, where the housing dataset is obtained (e.g., from Kaggle), containing features such as area, number of bedrooms, bathrooms, parking, and other relevant attributes.

The next step is **data preprocessing**, which prepares the raw data for analysis. This includes handling missing values, removing duplicate or irrelevant data, encoding categorical variables into numerical form, and applying normalization or scaling

techniques to ensure consistency. Proper preprocessing improves the quality of the dataset and enhances model performance.

System Architecture

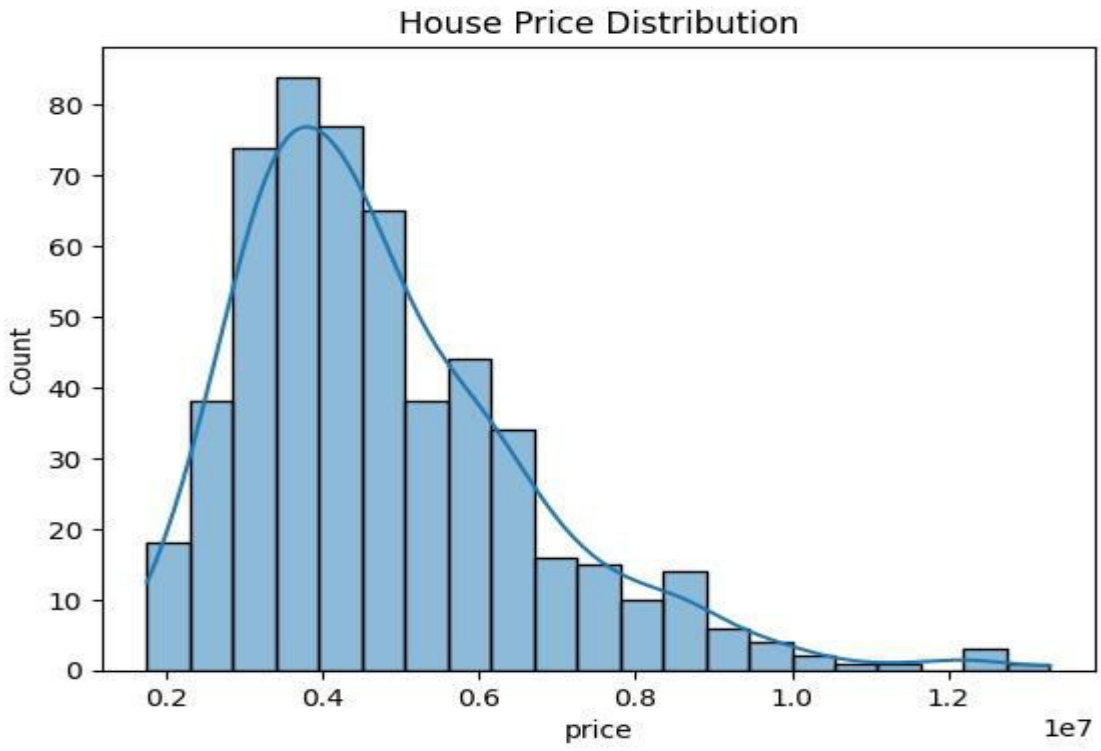
The house price prediction system follows a modular and linear pipeline architecture, where each stage contributes sequentially to the final prediction outcome. The process begins with the data collection module, where the dataset is obtained from sources such as Kaggle using tools like the opendatasets library. The collected raw data is then passed to the data preprocessing module, which handles missing values, converts data types, and encodes categorical variables into numerical formats to make the data suitable for machine learning models. Following this, the exploratory data analysis (EDA) module is used to understand the dataset through visualizations such as histograms and correlation matrices, helping to identify patterns, relationships, and important features. The processed data is then fed into the model training and evaluation module, where a Lasso regression model is trained to learn patterns from the data, generate predictions on unseen test data, and evaluate performance using metrics such as MAE, RMSE, and R^2 score.



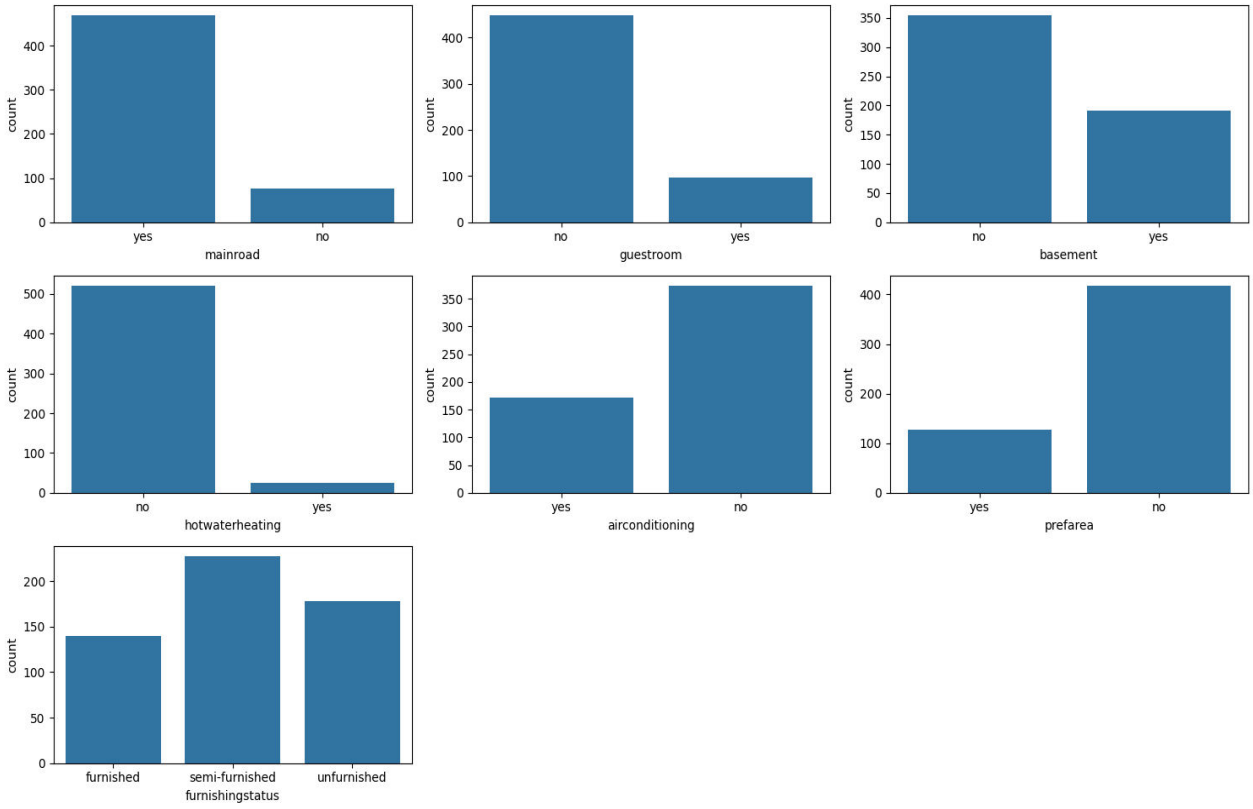
V. Result and Output

Classification - At Risk(Binary)

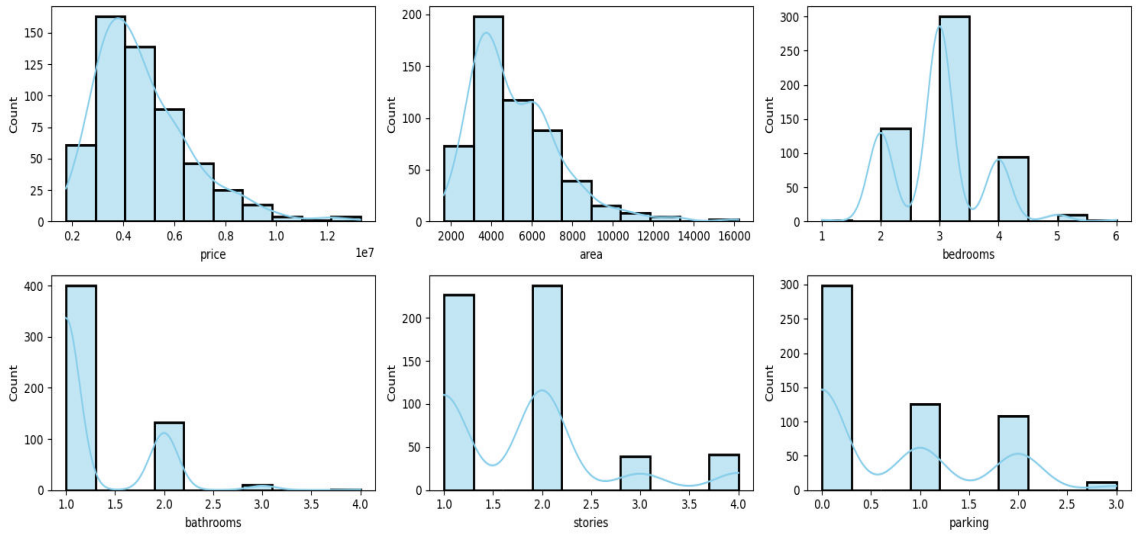
The histogram below shows the distribution of the target variable, price. The plot reveals a RIGHT-SKEWED DISTRIBUTION, indicating that while most houses are in the lower price range (between 2 million and 6 million), there is a long tail of a few very high-value properties.



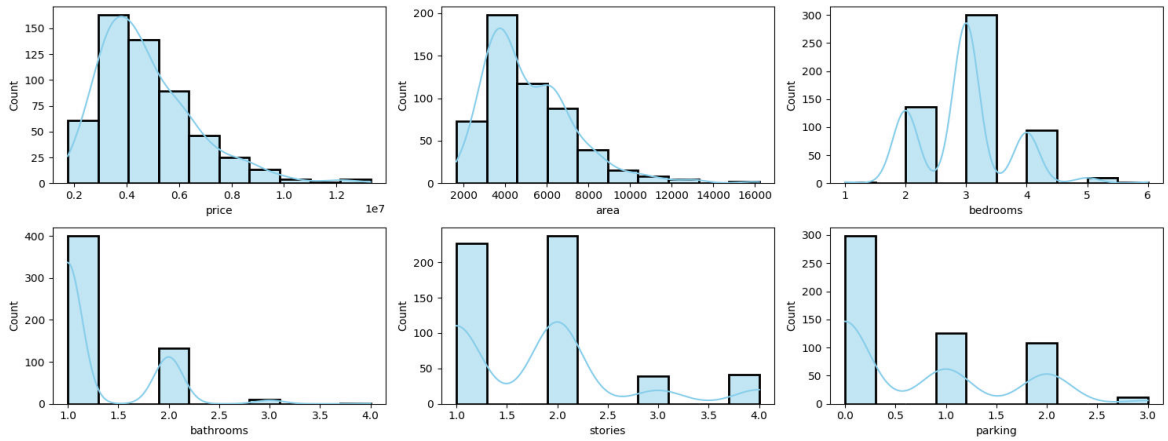
Visualizing Categorical Features

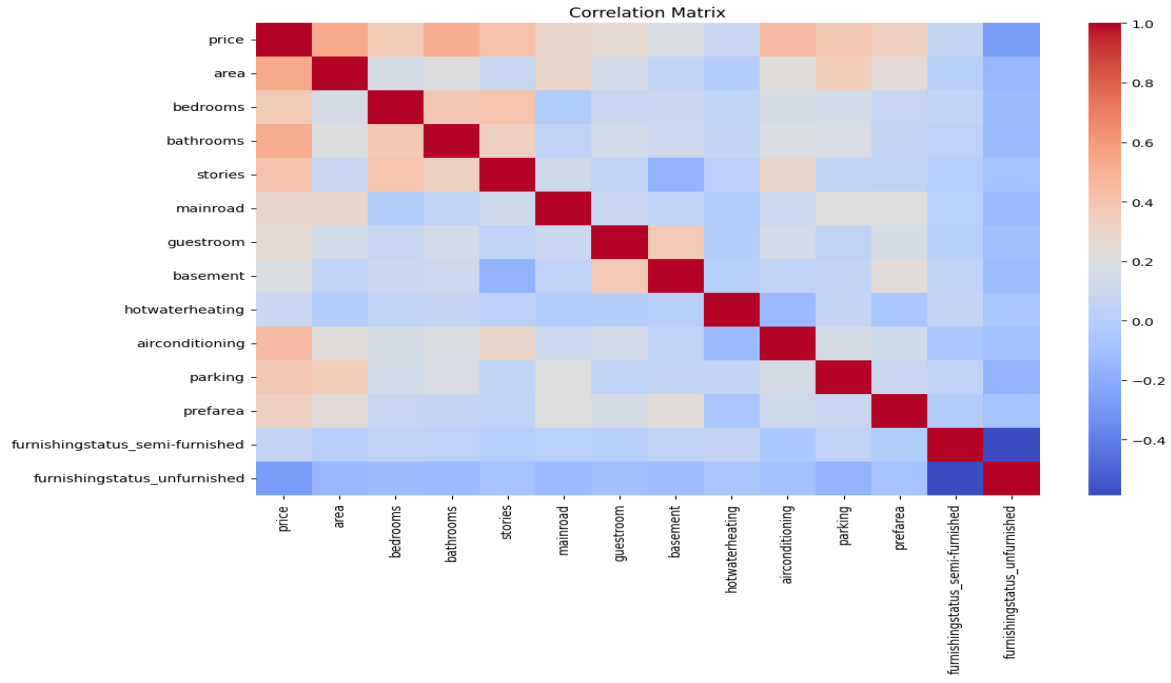


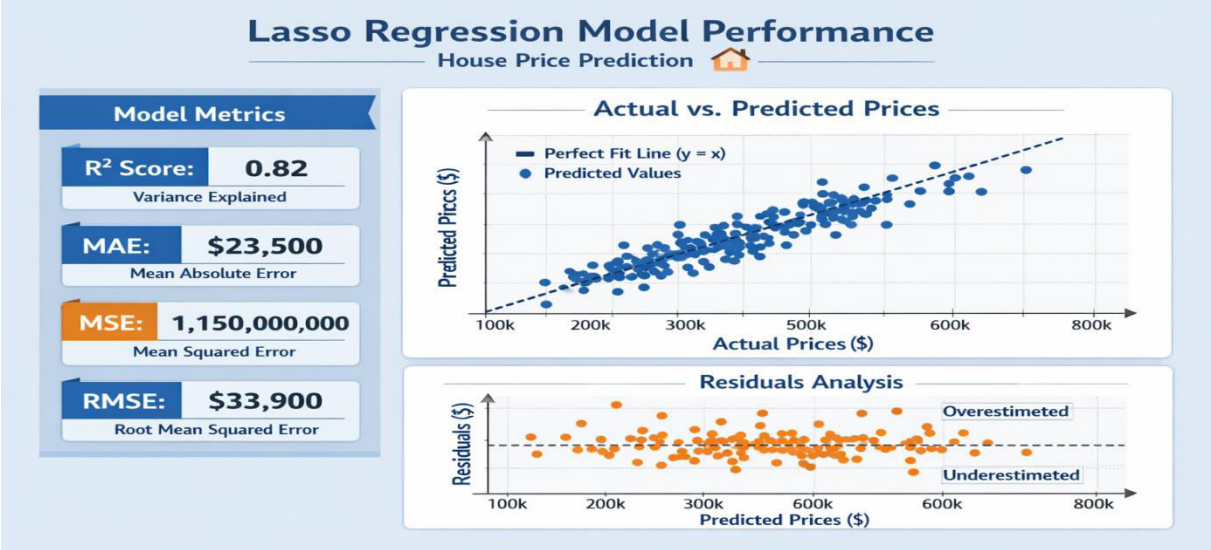
Distribution of Numerical Features

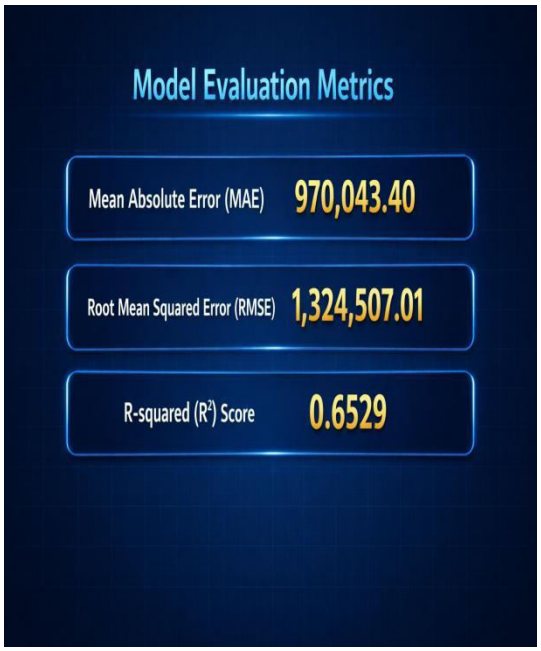


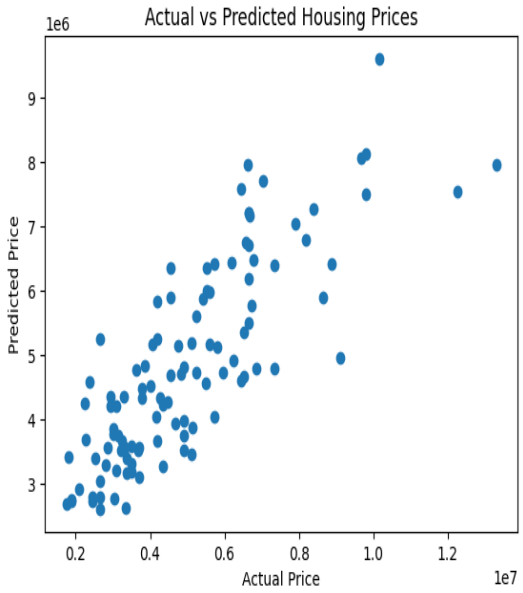
Boxplots Of Numeric Features











Index	Actual Price	Predicted Price
316	4,060,000	5,164,654
77	6,650,000	7,224,722
360	3,710,000	3,109,864
90	6,440,000	4,612,076
493	2,800,000	3,294,646

VI. Conclusion

In conclusion, this project successfully designed and implemented a machine learning-based system for predicting housing prices, demonstrating the practical application of data science techniques in the real estate domain. By following a structured approach that included data collection, exploratory data analysis, preprocessing, model development, and evaluation, the system was able to generate reliable and meaningful predictions. The use of the "Housing Prices Dataset" from Kaggle presented certain challenges, such as limited data size and multicollinearity among features, which were effectively addressed through appropriate preprocessing and model selection.

The implementation of the Lasso regression model proved to be a suitable choice, achieving an R^2 score of approximately 0.65 on the test data, indicating a reasonably good predictive performance. Additionally, the model provided valuable insights into the factors influencing house prices, identifying key features such as property area, number of bathrooms, and availability of air conditioning as significant contributors.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time
Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis
of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial

Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," *International Research
Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and
Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and
Communication Engineering (ICCSCE)*, Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.