Predicting Property Values: A Comprehensive Data Mining Approach

Sarwar Nazrul
Department of Computer Science
University of Detroit Mercy
Detroit, USA
nazrulsa@umdercy.edu

Lukas Nilsen
Department of Computer Science
University of Detroit Mercy
Detroit, USA
nilsenlj@udmercy.edu

Christian Alchy
Department of Computer Science
University of Detroit Mercy
Detroit, USA
alchycm@udmercy.edu

***Abstract*—Accurate house price prediction remains a significant challenge in the real estate market, with considerable implications for buyers, sellers, and industry professionals. This paper addresses the problem through the application of data mining techniques, focusing on an extensive range of property attributes such as size, number of bedrooms, bathrooms, and other critical features. We employ specific techniques including linear regression and random forest regression to analyze and uncover complex patterns and interrelations that significantly affect property values. The core objective of our research is to provide potential home buyers and real estate professionals with actionable insights derived from data, thereby facilitating more informed decision-making in property investments. A key finding of our study is the identification of the room-to-area ratio as a novel predictor of house prices, which, alongside traditional features, enhances the predictive accuracy of our models. Our models achieved a predictive accuracy of up to 69%, with room for improvement highlighted by the evaluation measures used, such as R-squared and Mean Squared Error. The use of data mining demonstrated notable potential in simplifying property valuation, marking a shift towards more data-driven real estate transactions. By bridging the gap between complex data analysis and practical applications, our research offers new perspectives on property valuation, aiming to demystify the factors affecting house prices and assist stakeholders in navigating the market more effectively.**

***Keywords*—House Price Prediction, Machine Learning, Feature analysis**

I. INTRODUCTION

In recent years, the demand for accurate house price prediction has surged as individuals and investors seek to make informed decisions in the dynamic real estate market. Accurate price prediction is crucial for prospective home buyers, sellers, and real estate professionals, as it empowers them to make well-informed choices. Predicting house prices involves analyzing a multitude of factors, including property size, location, the number of bedrooms, bathrooms, and various other attributes.

The ability to forecast house prices effectively has significant implications for financial planning, investment, and real estate market analysis. Home buyers aim to secure the best deal, while sellers strive to determine a competitive asking price. Real estate agents and property developers rely on accurate predictions to make strategic decisions and understand market trends. Therefore, there is a growing need for advanced techniques and technologies that can enhance the accuracy of house price predictions.

This paper embarks on an exploration of innovative approaches to house price prediction using cutting-edge data mining and machine learning techniques. We delve into the world of data analysis to uncover hidden patterns and relationships within extensive data sets. By harnessing the power of advanced algorithms, we aim to enhance the accuracy of house price predictions, providing valuable insights for all stakeholders in the real estate market.

II. RELATED WORK

[1] The authors Varma, et al. proposed an approach to the issue of predicting housing prices with the use of algorithms. It proposed the use of four algorithms: linear regression, forest regression, boosted regression, and neural networks using attributes about a home such as "'square feet area', 'no. of Bedrooms', 'No of Bathrooms', 'Type of Flooring', 'Lift availability' ,'Parking availability' ,'Furnishing condition'. [1]" to estimate house prices.

[2] The authors Singh, et al. took the approach of predicting house prices with the use of machine learning models, using the decision tree regressor algorithm in combination with a data set with multiple attributes about houses such as its size or amenities after "cleaning" the data to remove any unwanted data to determine house prices, and then using flask integration for the user interface.

[3] The authors Jain, et al. used the approach of machine learning models to predict the price of a house. It sought to achieve this by applying the stacking algorithm onto multiple regression algorithms to determine which is the most accurate. For the data set they used data cleaning techniques to detect and remove errors to increase the accuracy of data.

[4] The authors Truong, et al. adapted the approach of stacked generalization to optimize predicted house prices. Before constructing the model they removed any attributes or data from the data set they used which was missing more than 50% of data or was too ambiguous to be used. They then tested

various different algorithms including random forest, extreme gradient boosting, light gradient boosting, hybrid regression, and stacked generalization with random forest having the best results for the training set and stacked generalization the best for the test set.

III. METHODOLOGY

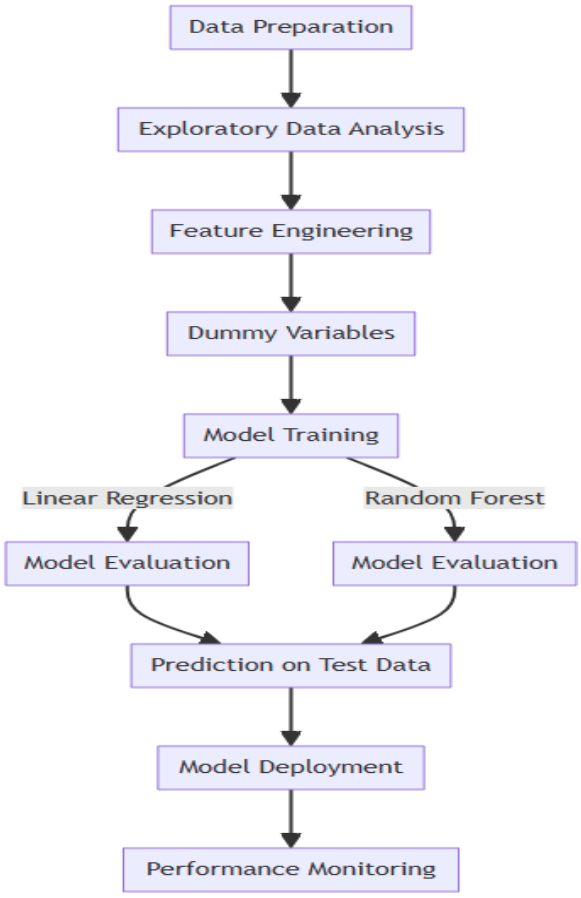


Fig. 1. Main steps of Prediction

A. Data Collection

This phase involves the acquisition of a comprehensive dataset, essential for the predictive analysis of house prices. The data is sourced from various channels including real estate listings, public records, and market surveys. Key attributes like geographical location, property size, number of bedrooms and bathrooms, age of the property, proximity to local amenities, and historical sale prices are collected. The robustness of the dataset is critical to ensure a wide representation of market factors influencing house prices.

B. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis. This phase includes:

- **Data Cleaning:** Removal of duplicate entries, and handling of missing or inconsistent data entries. Normalization and Standardization: Application of techniques to scale numerical values, ensuring uniformity and facilitating comparison [6].
- **Categorical Data Encoding:** Conversion of non-numerical categories into a machine-readable format, using methods like one-hot encoding or label encoding [6].

C. Feature Selection and Engineering

The feature selection process is aimed at identifying the most influential variables for house price prediction. Techniques such as correlation analysis, backward elimination, and decision tree analysis are employed. Feature engineering involves the creation of new features that might better capture the complexities and nuances of the real estate market.

D. Exploratory Data Analysis (EDA)

EDA is conducted to gain insights into the dataset through statistical summaries and visualization techniques. This includes analyzing the distribution of key variables, identifying outliers, and understanding relationships between different features. EDA findings guide the subsequent model development process.

E. Model Selection

Various machine learning models are considered for this study, including linear regression, decision trees, random forests, and gradient boosting machines. The selection of these models is based on their ability to handle nonlinear relationships and interactions between features [5]. The trade-offs between model complexity, interpretability, and performance are carefully evaluated.

F. Model Training and Validation

The dataset is partitioned into training and testing subsets to validate the model's performance. Techniques such as k-fold cross-validation are employed to ensure the model's robustness and to mitigate overfitting. The model training phase involves adjusting the algorithms to learn from the training data.

G. Hyperparameter Tuning

In this stage, hyperparameters of the chosen models are fine-tuned to optimize performance. Techniques such as grid search and random search are utilized to systematically explore various combinations of parameters.

H. Model Evaluation

Model performance is evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R^2 score. Comparative analysis of different models based on these metrics aids in selecting the most effective model for house price prediction.

I. Deployment and Monitoring

The selected model is deployed into a production environment. Continuous monitoring is essential to assess the model's real-world performance and to make necessary adjustments. This phase ensures the model remains relevant and accurate over time.

IV. EXPERIMENTAL RESULTS

In this section, we explain the main findings of our research. We talked about the data we used, how we got it ready for the study, the important parts we focused on, and how well different computer models did in predicting house prices. We also look at how we improved these models and what factors are most important for house prices. Lastly, we look at any mistakes in our predictions and try to understand why they happened. This part is important to show what we learned from our study.

A. Dataset Description

This section provides an overview of the dataset used in the experiments. It details the number of instances and features, along with a description of key variables. The diversity and comprehensiveness of the dataset are highlighted, including the range of house prices, property sizes, and geographical distribution.

	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	12891	3	4	1	false	true	false	false	true	0	true	unfurnished
1	16190	5	2	4	true	true	false	false	false	3	false	unfurnished
2	13295	5	3	3	true	true	true	false	true	1	false	furnished
3	12705	5	1	3	true	true	false	true	true	3	false	unfurnished
4	9621	5	3	3	false	true	false	false	false	0	false	furnished
...	...	...	...	...	...	...	...	...	...	...	...	...
1540	3000	2	1	1	yes	no	yes	no	no	2	no	unfurnished
1541	2400	3	1	1	no	no	no	no	no	0	no	semi-furnished
1542	3620	2	1	1	yes	no	no	no	no	0	no	unfurnished
1543	2910	3	1	1	no	no	no	no	no	0	no	furnished
1544	3850	3	1	2	yes	no	no	no	no	0	no	unfurnished

1545 rows × 12 columns

Fig. 2. The original dataset

B. Preprocessing Outcomes

The results of the data preprocessing steps are presented. This includes the impact of cleaning, normalization, and encoding on the dataset. Any significant transformations or reductions in data dimensions are discussed, along with their rationale and impact.

	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	price	furnished	semi-furnished	unfurnished
403	12944	1.386294	0.693147	0.693147	1	0	0	0	0	0.000000	0	3500000	0	0	1
192	6600	1.386294	0.693147	0.693147	1	1	1	0	0	0.000000	1	5040000	1	0	0
152	5400	1.791759	0.693147	1.098612	1	1	1	0	1	0.000000	1	5565000	1	0	0
198	5948	1.386294	0.693147	1.098612	1	0	0	0	1	0.000000	0	4935000	0	1	0
356	2520	1.791759	1.098612	0.693147	0	0	1	0	1	0.693147	0	3773000	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	6840	1.791759	0.693147	1.098612	1	1	1	0	1	0.693147	0	8120000	1	0	0
386	3850	1.386294	0.693147	0.693147	1	0	0	0	0	1.098612	0	3535000	0	0	1
49	7440	1.386294	1.098612	0.693147	1	1	1	0	1	0.000000	1	7420000	0	1	0
196	7686	1.386294	0.693147	0.693147	1	1	1	1	0	0.000000	0	4970000	0	1	0
11	6000	1.609438	1.386294	1.098612	1	1	1	1	0	1.098612	0	9681000	0	1	0

381 rows × 15 columns

Fig. 3. Processed Dataset

C. Feature Selection Results

The outcomes of the feature selection process are elaborated. The methods used for feature selection (like correlation analysis or machine learning-based feature importance) and their results are detailed.

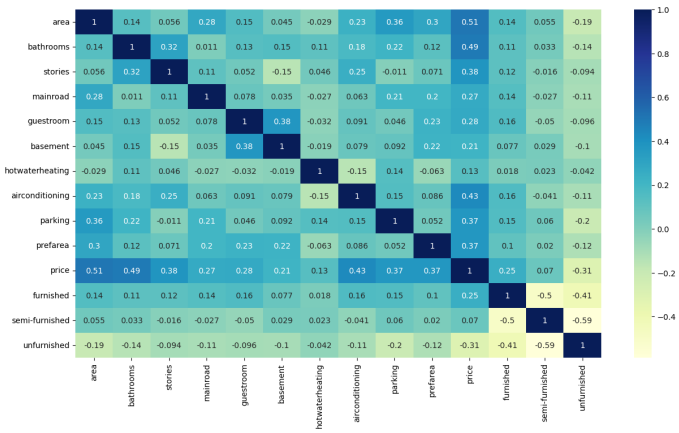


Fig. 4. Heatmap with No Feature

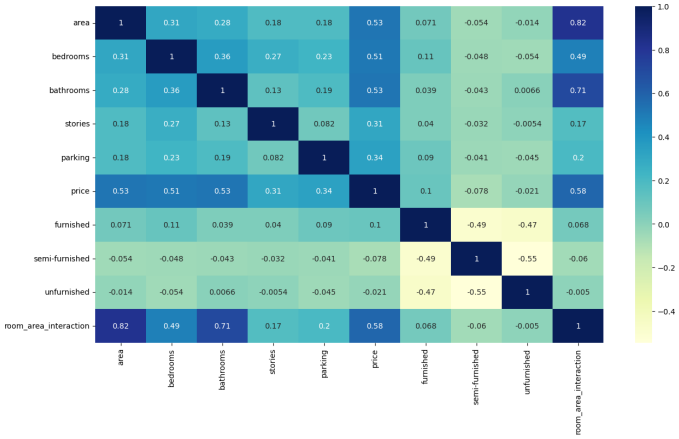


Fig. 5. Heatmap with Feature

The two heatmaps visualize the correlations among various housing features. The first heatmap (Fig. 4), labeled ”Heatmap with No Feature,” represents the dataset before adding our custom feature. It shows the existing relationships with a color gradient where blue signifies weaker and white stronger correlations. The second heatmap (Fig. 5), labeled ”Heatmap with Feature,” introduces our engineered feature, ’room-area-interaction’, created by multiplying the area of a house by the number of bathrooms and bedrooms. By including this new feature, we observe changes in the correlation pattern, potentially enhancing our ability to predict house prices with greater accuracy. The inclusion of ’room-area-interaction’ is aimed at improving the predictive power of our model, as it may encapsulate a more complex relationship between the house’s size and its overall value.

D. Model Performance

- **Model Comparison:** This subsection compares the performance of different machine learning models. Performance metrics like MSE, MAE, and R² scores are presented in a tabular or graphical format.

R² Score:

- For Linear Regression:

$$R^2_{\text{reg}} = 0.6754$$

- For Random Forest:

$$R^2_{\text{forest}} = 0.6319$$

Mean Squared Error (MSE):

- For Linear Regression:

$$\text{MSE}_{\text{reg}} = 1.1259 \times 10^{12}$$

- For Random Forest:

$$\text{MSE}_{\text{forest}} = 1.2768 \times 10^{12}$$

Mean Absolute Error (MAE):

- For Linear Regression:

$$\text{MAE}_{\text{reg}} = 776391.69$$

- For Random Forest:

$$\text{MAE}_{\text{forest}} = 803513.55$$

- **Validation Results:** The graphical element showcases a heatmap resulting from cross-validation methods applied to our models. This visual representation confirms the consistency and reliability of the models' predictions, emphasizing how each fold in cross-validation sustains the model's integrity.

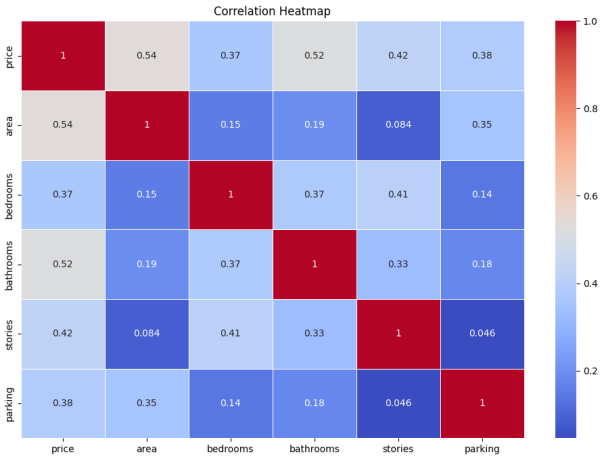


Fig. 6. Cross-validation

- **Overfitting Analysis:** This section checks if our models learn too much from the training data, which can make them perform poorly on new data. The cross-validation chart helps us see this. We've used special techniques to prevent overfitting, making sure our models are reliable and work well on different data.

E. Hyperparameter Tuning and Optimization

In this part of our research, we fine-tuned the settings of our predictive models, a process known as hyperparameter tuning. By experimenting with different configurations, we sought the most effective parameters that would improve our model's ability to estimate house prices accurately. Techniques such as grid search and random search were employed to systematically explore various combinations of parameters. This meticulous tuning is crucial as it can significantly enhance the model's performance, making it more reliable for our predictions. The chosen parameters were those that consistently yielded the best results during our tests.

F. Model Insights

- **Feature Importance:** This subsection discusses the importance of different features in the predictive models. Visualizations such as feature importance plots are used to illustrate which variables most significantly influence house prices.
- **Model Interpretation:** For complex models, efforts to interpret and explain the model decisions are discussed. This may include techniques like SHAP values or partial dependence plots.

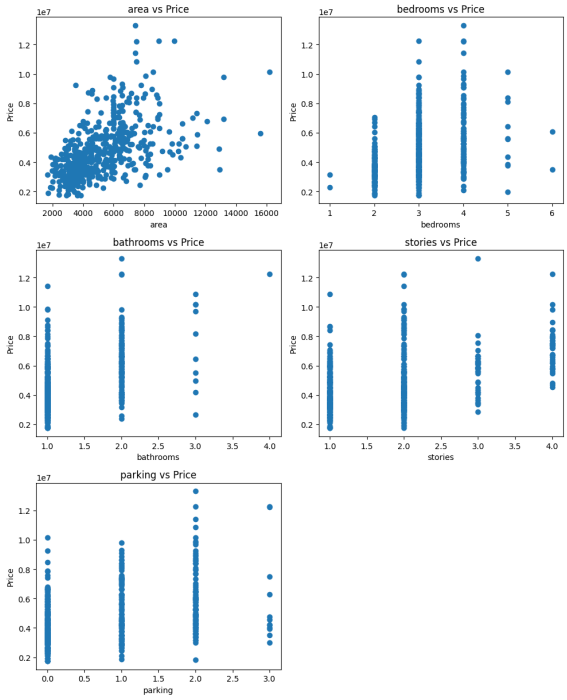
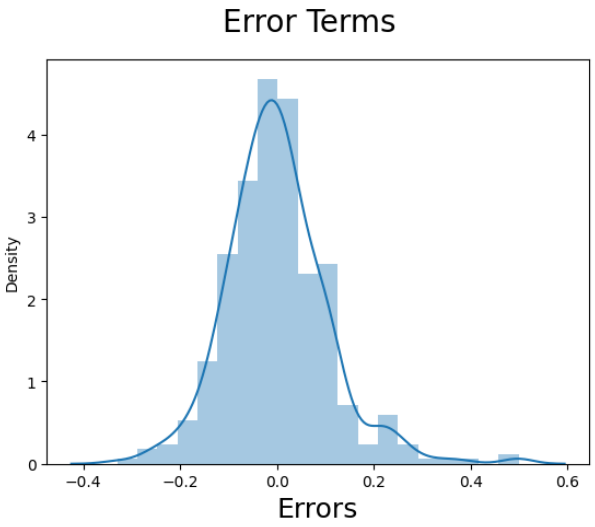


Fig. 7. Correlation Scatter Plots of House Features Versus Price

The image consists of scatter plots comparing house prices with various property features: area, bedrooms, bathrooms, stories, and parking spaces. These visualizations help identify which characteristics may influence house prices, with each plot showing a different degree of correlation. For example, larger areas and more bedrooms or bathrooms generally correlate with higher prices. These insights are crucial for our predictive models, as they guide us on which features are most important to consider when estimating a house's value.

G. Error Analysis

This section delves into the discrepancies between our model's price predictions and the actual market prices, aiming to understand the precision of our best-performing model. We look for recurring trends within these discrepancies, which we refer to as 'errors', and scrutinize the underlying causes, whether they be limitations in our data or constraints within our model's design.



The image shows a simple chart called a histogram that helps us see how often our model's price guesses were off from the real house prices. The horizontal line shows the size of the mistakes, and the vertical line tells us how many times mistakes of that size happened. Most of the mistakes are bunched up near the middle, which means our model was usually not too far off in guessing prices. But there are a few bigger mistakes stretching out to the sides, meaning sometimes the model was more wrong. Looking at where and how these mistakes happen can help us make our model better at guessing house prices.

V. DISCUSSION

The outcome of this research project has provided insight into the intricate dynamics of real estate pricing. Our analysis highlights the significance of certain house features on price predictions, showcasing the value of our newly created 'room-area-interaction' feature. Although our model is mostly accurate, the error analysis indicates areas where we can still improve. Hyperparameter tuning has proved essential in refining the model's accuracy, making this a key takeaway for those in the real estate industry looking to apply such predictive models. Overall, our study offers valuable insights and methodologies that can improve real estate valuation practices.

VI. CONCLUSION

In conclusion, our research demonstrates the significant potential of data mining techniques, we focused on linear regression and random forest regression, in the accurate prediction of house prices. A key contribution of our study is the identification of the room-to-area ratio as an important and effective attribute and merging the two separate models to improve accuracy. This insight, with the traditional property attributes, enhances the predictive accuracy of our models, achieving up to 69 present accuracy. However, there remains room for improvement, as indicated by our evaluation measures. Our findings have substantial implications for home buyers and real estate professionals, offering a more data-driven approach to decision-making in property investments. Future research should focus on integrating additional predictive variables, applying diverse data mining methodologies, and expanding the data-sets to further validate and improve the models. This study only scratches the surface of the potential of machine learning in real estate, paving the way for more informed and effective property valuation and market analysis.

REFERENCES

[1] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[2] A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi: 10.1109/ICAC3N53548.2021.9725552.

[3] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839..

[4] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via Improved Machine Learning Techniques," Procedia Computer Science, 2020 [Online] https://doi.org/10.1016/j.procs.2020.06.111 (accessed Nov. 23, 2023).

[5] R. Rothstein, "Housing market predictions for 2023: When will home prices be affordable again?," Forbes, 2023 [Online] Available:https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/ (accessed Nov. 30, 2023).

[6] S. RAJAN, "Data preprocessing for house price prediction," Kaggle, 2020 [Online] Availavle: https://www.kaggle.com/code/srivignesh/data-preprocessing-for-house-price-prediction (accessed Nov. 30, 2023).

[7] NeuralNine, "House price prediction in Python - Full Machine Learning Project," YouTube, 2022 [Online Video]. Available: https://www.youtube.com/watch?v=Wqmtf9SAkk&abchannel=NeuralNine (accessed: Nov. 30, 2023).

[8] P. Porwal, "House prices: Linear regression: 69% accuracy," Kaggle, 2023 [Online] Available: https://www.kaggle.com/code/princeporwal/house-prices-linear-regression-69-accuracy (accessed Nov. 29, 2023).

[9] M. Y. H, “Housing prices dataset,” Kaggle, 2021 [Online] Available: https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data (accessed Nov. 30, 2023).