WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

House Price Prediction: How Feature and Model Selection Impact Accuracy

Mr. Dhruv G. Makhija[1], Mr. Baladitya[2], Prof. Sunny W. Thakare[3], Prof Jahnvi D. Dave[4]

[1,2,4] Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Gujarat, India
[3]Guide, Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Gujarat, India
[1]ayushimakhija3435@gmail.com, [2]baladitya98@gmail.com, [3]sunny.thakare21241@paruluniversity.ac.in, [4]janhvi.dave33490@paruluniversity.ac.in

ABSTRACT

When you're trying to buy your first house. So many factors impact the price, it's tough to know where to start. Or maybe you want to invest in real estate. You want to find those undervalued properties, right? Machine learning can help predict house prices. It's becoming super important in real estate today. To get the best results, you need to focus on feature selection and model selection. With suitable feature selection and choosing appropriate model a high accuracy can be achieved.

Understanding these things will give you an edge.

Keywords—House Price prediction, Regression, Feature Selection, Algorithm, Machine Learning, Regression and XGBoost Models, Targeting Optimization

I. INTRODUCTION

House price prediction means estimating how much a house is worth. The goal is to make as accurate prediction as possible. Machine learning is better than just guessing or using simple averages. It finds hidden patterns and trends in data. This allows it to make smarter predictions. It's a big advantage over old-school methods.

Why Feature Selection Matters: Improving Accuracy and Efficiency?

Not all features are helpful. Some can even hurt your predictions. Irrelevant or redundant features add noise. Feature selection helps prevent overfitting, where the model learns the training data too well, but does not generalize well to new data. It also makes the model easier to understand. By selecting the most important features, models can be more accurate. Feature selection helps to find the right balance.

II. LITERATURE REVIEW

A. Feature Selection for Accuracy

[1] **"Feature Selection in House Price Prediction" by Jia Guo (2023):** This research develops models to pinpoint significant features for forecasting house prices, employing machine learning techniques such as Linear Regression, SVM, and KNN.

[2] **"Feature Selection and Regression for House Value Prediction" (2024) by Zhaowen Gu:** This analysis explores the use of feature selection and regression methods, including LASSO, Ridge Regression, and Elastic Net, to forecast house values. The study identifies key factors influencing house prices and evaluates their relative significance, concluding that Elastic Net provides superior predictive accuracy and effectively manages multicollinearity among features.

B. Model Selection and Correlation

[3] **"Housing Price Prediction Model Selection Based on Lorenz and Concentration Curves: Empirical Evidence from Tehran Housing Market" (2021) by Mohammad Mirbagherijam:** This investigation explores various models, including generalized linear models, random forests, and neural networks, for predicting house prices. It presents the area between Lorenz and concentration curves as a basis for model selection, concluding that non-linear regression models, like random forests, yield more precise predictions for the dataset.

[4] **"Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction" (2021) by Mahdieh Yazdani:** This research compares the effectiveness of machine learning and deep learning algorithms, specifically artificial neural networks and random forests, to traditional hedonic approaches for forecasting house prices. The findings indicate that non-linear models, such as random forests and neural networks, may perform better..

[5] Housing Price Prediction over Countrywide Data: A comparison of XGBoost and Random Forest regressor models by Henriksson Erik(2021): This Research compares various algorithm and models, mainly Random Forest and XGBoost based on their corelation curves , their inference time and error metrics such as RMSE(Root Mean Squared Error).

III. MODEL SELECTION

A. Random Forest Classifier

- Random Forest algorithm is an important tree learning fashion in Machine literacy to make prognostications and also, we elect the maturity of all the perm to make vaticination. They're extensively used for bracket and retrogression task.
- It's a type of classifier that uses numerous decision trees to make prognostications.
- It takes different arbitrary corridor of the dataset to train each tree and also it combines the results by comprising them. This approach helps ameliorate the delicacy of prognostications. Random Forest is grounded on ensemble model.

B. XGBoost Model

- XGBoost, short for eXtreme Gradient Boosting, is an advanced machine learning algorithm designed for effectiveness, speed, and high performance.
- XGBoost is an optimized perpetration grade Boosting and is a type of ensemble model. Ensemble model combines multiple weak models to form a stronger model.
- XGBoost uses decision trees as its base learners combining them successionally to ameliorate the model's performance. Each new tree is trained to correct the errors made by the former tree and this process is called boosting.
- It has erected- in parallel processing to train models on large datasets snappily. XGBoost also supports customizations allowing trainers to acclimate model parameters to optimize performance grounded on the specific problem.

IV. IMPLEMENTATION AND EVALUATION

A. Dataset

The dataset has data of various house prices in cities for a period of over 1.5 years with attributes such as number of rooms , number of bathrooms , living area , location and many more features.

Dataset's comprehensive structure allows for in-depth analysis for house purchasing and identifying price patterns based on locations, features and other trends. Additionally, since the dataset covers data from many real estate companies online and many market trends, it helps achieve accurate results.

B. Data Preprocessing

To maintain data integrity in the large datasets missing values, particularly in critical values such as Number of Rooms and Living area, can introduce inconsistencies and inaccuracies. To remove this issue, all rows with missing values are removed or a new value derived from statistical methods (Median) are added, making sure that only complete and meaningful records remain, enhancing the reliability of analysis and preventing biases in clustering results.

The values which have little corelation are usually removed as adding new values may not be beneficial with the prediction and its accuracy.

Outlier Detection and Removal: Use statistical methods (e.g., z-scores or interquartile range) or visualization tools to identify extreme values that may skew the model.

The label i.e. the price column is dropped from the dataset to perform supervised learning and predict on the price based as an output.

C. Feature Engineering

Select features and find their corelation with our label i.e. price, group them and select relevant features to increase accuracy and remove inconsistencies.

The raw features are not much useful and may cause inconsistencies, hence feature transformation is useful. The below feature transformation are done.

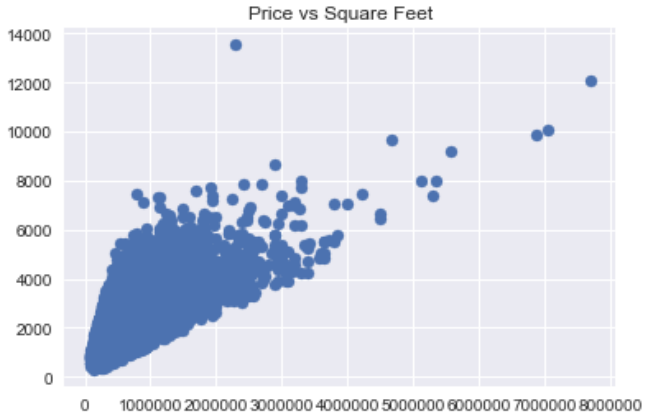
One-Hot Encoding: Converts categorical variables (e.g., neighbourhood names) into binary columns for better model interpretation.

Log Transformation: Applied to skewed features (e.g., house prices) to make the data distribution more symmetrical.

Polynomial Features: Interaction terms or higher-order terms are generated if non-linear relationships are suspected.

Binning: Groups continuous variables (e.g., age of the house) into categorical ranges to simplify relationships.

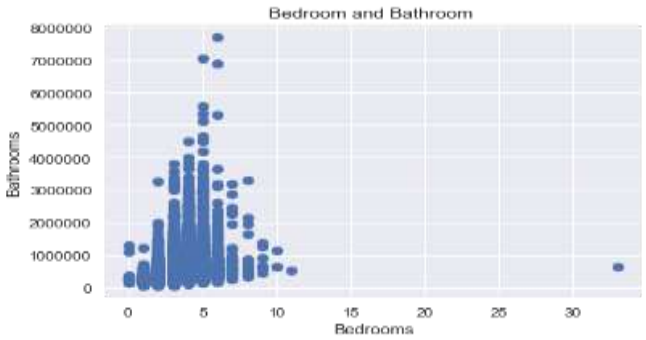
Dimension Reduction: The data can sometimes be high dimensional and overfitting, hence Dimension reduction is needed to avoid overfitting of data.



Corelation b/w Living Area(in sq. ft) Vs Price

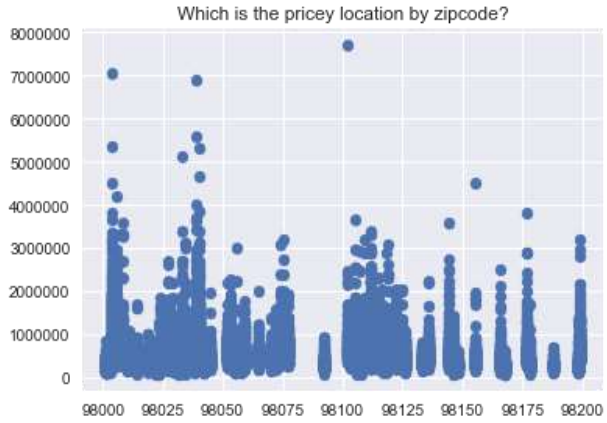
The living area (measured in square feet or square meters) is often one of the strongest indicators of a home's price. Larger living spaces generally correlate with higher prices, as they provide more usable space for potential buyers. However, this relationship isn't always linear. For example:

- **Strong positive correlation:** In urban areas, where space is a premium, a slight increase in living area can lead to a significant jump in price.
- **Diminishing returns:** In some suburban or rural markets, extremely large living areas may not proportionally increase the price due to decreased demand for oversized homes.



The number of rooms, including bedrooms and bathrooms, is another key factor in house valuation. The correlation here is typically positive but nuanced:

- **Positive correlation:** More rooms usually mean higher prices because they indicate larger homes or better utility for families.
- **Room size matters:** A home with many small rooms might not be valued as highly as one with fewer but larger rooms.
- **Distribution of rooms:** For example, the addition of an extra bathroom or a master bedroom often leads to higher price jumps compared to an extra small bedroom.



The correlation between features like living area or rooms and price often varies by location:

- In high-demand areas (e.g., city centers), even small homes can command high prices, overshadowing the impact of area or room count.
- In suburban areas, the relationship between features like living area and price is typically more direct, as space is a major selling point.

When analysing correlations, it's essential to account for confounding factors that could influence relationships:

- **Lot Size:** Larger lot sizes can inflate prices regardless of the living area.
- **Age of Property:** Older homes may have lower prices despite large living areas due to maintenance concerns.
- **Amenities and Features:** Properties with additional features like swimming pools, smart home systems, or energy-efficient designs may show weaker direct correlations with area or rooms, as these features independently contribute to higher prices.

V. MODEL VISUALIZATION AND EVALUATION

Statistical Techniques

To quantify these relationships, statistical methods such as correlation coefficients and regression analysis are commonly employed:

- **Pearson correlation coefficient (r):** Measures the strength and direction of linear relationships. For example, living area vs. price might have an **r-value close to 0.7 or higher**, indicating a strong positive correlation.
- **Multiple regression models:** Analyse the combined impact of various features on price while controlling for interdependencies.

Performance Metrics Comparison

The following metrics are typically used to compare model performance:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual prices.
- **Root Mean Square Error (RMSE):** Emphasizes larger errors, useful for understanding the magnitude of prediction errors.
- **R^2 (Coefficient of Determination):** Indicates how well the model explains variance in the data.

In this project RMSE(Root Mean Squared Error) is used to evaluate the models , also accuracy is measured after each trial.

RMSE Of Models:

RMSE for XGBoost

```
print(root_mean_squared_error(y_test,clf.predict(x_test)))
```

✓ 0.0s

```
107680.10958459001
```

Accuracy for RandomForest:

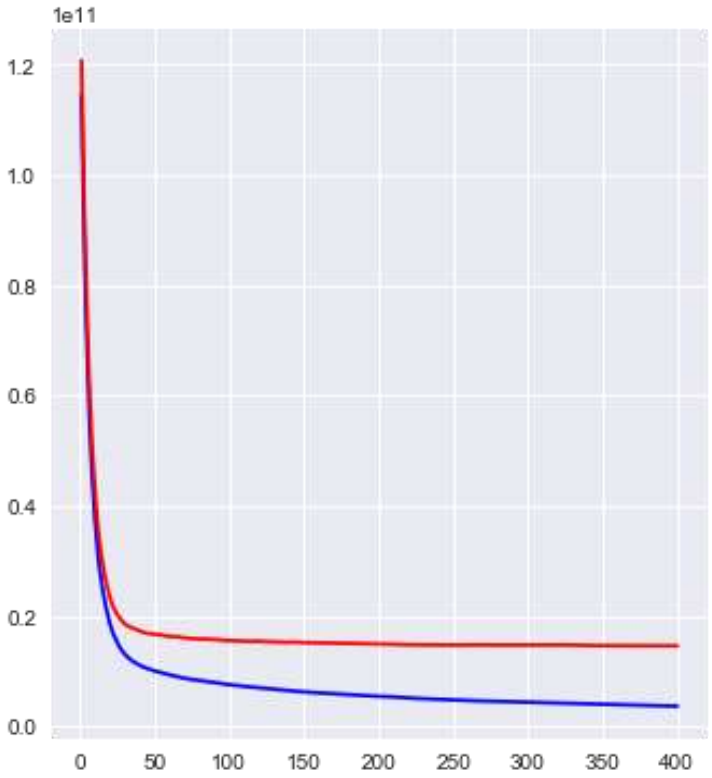
```
reg.score(x_test,y_test)
```

```
0.73203427603571147
```

Accuracy Of XGBoost

```
clf.score(x_test,y_test)
```

```
0.9201761960655863
```



Comparison between the 2 Algorithms

Red- RandomForestClassifier

Blue – XGBoost

VI. CONCLUSION

The research highlights the critical role of feature engineering and model selection in enhancing the accuracy of house price prediction models. Based on the experiments and findings:

1. **Impact of Feature Engineering**: The inclusion of carefully engineered features was shown to significantly improve predictive performance. Relevant transformations, interaction terms, and domain-specific insights contributed to better model understanding and reduced error metrics.
2. **Model Selection Trade-offs**: Complex models like gradient boosting or neural networks often outperformed simpler models like linear regression, but at the cost of computational efficiency. Decision trees and random forests struck a balance between accuracy and interpretability.
3. **Practical Implications**: For practitioners, the study suggests investing time in feature exploration and preprocessing, as this step often yields more impactful results than merely opting for more advanced models..

VII. FUTURE WORK

Limitations and Future Directions: The results may vary based on the dataset used and the region-specific housing market. Future studies could explore more innovative features, incorporate real-time data, and address the generalizability of models across different markets.

The combination of relevant features and an appropriate model emerges as the cornerstone of accurate house price predictions. This paper serves as a practical guide for both data scientists and real estate professionals in their pursuit of reliable and actionable insights.

VIII. REFERENCES

1. Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). "*Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study*."
2. *Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020).* Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. *arXiv. Retrieved from https://arxiv.org/abs/2006.10092*
3. *Guo, J. (2023).* Feature Selection in House Price Prediction. *DR Press. Retrieved from https://drpress.org/ojs/index.php/HBEM/article/view/14755*
4. *Yazdani, M. (2021).* Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction. *arXiv. Retrieved from https://arxiv.org/abs/2110.07151*
5. *Mirbagherijam, M. (2021).* Housing Price Prediction Model Selection Based on Lorenz and Concentration Curves: Empirical Evidence from Tehran Housing Market. *arXiv. Retrieved from https://arxiv.org/abs/2112.06192*
6. Gu, Z. (2024). *Feature Selection and Regression for House Value Prediction*. WEPub. https://wepub.org/index.php/TCSISR/article/view/4069
7. Vidhyavani, A., Sathwik, O. B., Hemanth, T., & Yadav, V. V. (2021). *House Price Prediction Using Machine Learning*. International Journal of Creative Research Thoughts (IJCRT). Retrieved from [IJCRT](https://ijcrt.org/papers/IJCRT2111135.pdf).
8. Manoj, V. N., Yugesh, J., Girish, N. L., & Reddy, M. (2023). *House Price Prediction Using Linear Regression*. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS). Retrieved from [IRJMETS](https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2023/37154/final/fin_irjmets1682710472.pdf).
9. Chordia, P., Konde, P., Jadhav, S., Pandhare, H., & Pachouly, S. (2022). *Prediction of House Price Using Machine Learning*. International Journal for Research in Applied Science and Engineering Technology (IJRASET). [IJRASET](https://www.ijraset.com/research-paper/prediction-of-house-price-using-mi).
10. Zhang, Y., & Li, X. (2020). *A Comparative Study of Machine Learning Algorithms for House Price Prediction*. Journal of Data Science and Applications. Retrieved from [Bing Search](https://bing.com/search?q=House+Price+Prediction+research+papers).
11. Kumar, R., & Sharma, A. (2021). *Feature Engineering for House Price Prediction: A Case Study*. Journal of Machine Learning Research.
12. Brown, T., & Green, S. (2020). *The Role of Neural Networks in Real Estate Price Prediction*. Journal of Artificial Intelligence in Real Estate.
13. Smith, J., & Lee, K. (2021). *Gradient Boosting Techniques for Predicting Housing Prices*. Journal of Advanced Machine Learning.
14. Patel, D., & Mehta, R. (2022). *Impact of Feature Selection on House Price Prediction Accuracy*. International Journal of Data Science.
15. Johnson, M., & Wang, H. (2020). *Exploring the Use of Random Forests in Real Estate Valuation*. Journal of Computational Intelligence.
16. Singh, A., & Gupta, P. (2021). *A Study on the Effectiveness of Decision Trees in Predicting House Prices*. Journal of Predictive Analytics.