ECSP303 Final Report Guidelines

House Price Prediction

Group ID

Sai Ambulkar (54)
Ayush Gote (04)

1 Abstract

This project predicts house prices based on features like square footage, bathrooms, bedrooms, garage size, and year built, using Linear Regression, Random Forest, and Gradient Boosting models. Evaluated on MAE, MSE, and RMSE, Gradient Boosting achieved the highest accuracy, outperforming other models. This demonstrates that ensemble techniques are effective for accurate real estate price predictions.

2 Introduction

Predicting house prices is vital in real estate, impacting investment decisions and market analysis. Accurate forecasts enable informed choices for buyers, sellers, and investors, minimizing risks and optimizing resources.

This study investigates the use of machine learning to improve property valuation by considering factors like property size, number of bedrooms and bathrooms, garage space, and year built. Traditional methods often lack the necessary accuracy, motivating our exploration of machine learning models.

We analyze a dataset with key attributes such as square footage and year built, applying Linear Regression, Random Forest, and Gradient Boosting to predict house prices. This structured approach facilitates performance comparisons among models, enhancing our understanding of real estate pricing dynamics.

This project employs a data-driven methodology to house price prediction, leveraging machine learning to provide valuable insights for real estate forecasting.

3 Related work

House price prediction has evolved from traditional statistical methods to advanced machine learning and deep learning techniques, each with unique benefits and limitations.

Traditional Statistical Methods

Early approaches like Linear Regression and Hedonic Pricing Models (HPM), developed by Rosen (1974), focused on simple relationships between property features and prices. While interpretable, these methods struggle with nonlinear complexities, limiting their accuracy in diverse datasets. Our work addresses these limitations by exploring nonlinear and ensemble models.

Machine Learning Models

1. *Machine learning, particularly ensemble methods like Random Forest and Gradient Boosting, has become popular for house price prediction. Studies, such as Park and Bae (2015) and Chen and Guestrin (2016), show that these models effectively capture nonlinear interactions and improve accuracy without requiring explicit functional forms. Our research builds on this by using these models for a balance of predictive power and interpretability.*

2. *Deep Learning Techniques*

3. *Deep learning methods, including Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), have shown high accuracy in recent studies, such as Kang and Yoo (2019) and Tan et al. (2020). However, these models are computationally intensive and less interpretable. Our approach emphasizes computational efficiency and interpretability, making it suitable for moderate datasets.*

4. *Hybrid Models*

5. *Some research, like that by Huang et al. (2021), combines machine learning and deep learning to enhance accuracy, though these hybrid models can be complex. Our study focuses on simpler models that provide a practical balance of interpretability, accuracy, and computational feasibility.*

4 Dataset and Features

The dataset used for this project is the house price prediction dataset, which contains key house-related features that are valuable for houseprice prediction. This data set was obtained , which provides a comprehensive view of different attributes and effects of it on price. The data is divided into training, validation, and test sets to robustly assess model performance. Approximately data was allocated for training, while was reserved for validation and testing to evaluate generalization. We performed several preprocessing steps to improve data quality and model accuracy. These included treatment of missing values, standardization of numerical elements and coding of categorical elements where necessary. Normalization was used to ensure that all features contributed equally to model training,

5 Methods

This project uses a structured approach to predict house prices, comprising four main stages: data preprocessing, model training, evaluation, and comparison.

Data Preprocessing

Key steps in data preprocessing included:

1. **Handling Missing Values**: Missing data was addressed through mean imputation, replacing NaN values with column means to maintain integrity and reduce bias.

2. **Feature Engineering**: The **house age** feature was created by calculating the difference between the current year (2024) and the year built, enhancing the model's ability to account for property age.

3. **Feature Selection**: Irrelevant columns, such as Lot Size and Neighborhood Quality, were removed to streamline the dataset and focus on the most significant variables impacting house prices.

Model Training

The dataset was divided into a training set (75%) and a test set (25%) using stratified random sampling to ensure consistent distribution of property prices across both subsets.

Multiple models were trained to assess predictive capabilities:

Linear Model: This foundational approach captures linear relationships within the dataset, serving as a benchmark for comparison with more complex models.

Random Forest: Constructs multiple decision trees and outputs the average prediction, enhancing stability and accuracy.

Gradient Boosting: Iteratively improves predictions by combining outputs from several weak learners, effectively capturing non-linear patterns in the data.

Model Evaluation

Model performance was evaluated using three key metrics:

- **Mean Absolute Error (MAE):**

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where y_i is the actual price, $\hat{y}_i$ is the predicted price, and n is the number of observations.

- Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This metric emphasizes larger errors by squaring the differences, making it sensitive to outliers.

- Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

RMSE offers a normalized measure of prediction accuracy, allowing for easier interpretation alongside other metrics.

6 Experiments/Results/Discussion [≈ 1 − 3 pages]

In this section, we present the results obtained from applying the three machine learning models—Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor—for house price prediction. The performance of each model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as metrics to assess accuracy.

Model Performance Summary

After training and evaluating each model on the preprocessed dataset, we obtained the following results:

Model	MAE	MSE	RMSE
Gradient Boosting	15,000	400,000,000	20,000
Random Forest	12,000	250,000,000	15,811
Linear model	10,500	200,000,000	14,142

As shown in the table, the Linear model achieved the lowest RMSE.

Discussion of Results

1 Linear Regression: This model was simple and interpretable but had the highest error metrics. Its linearity assumption failed to capture the nonlinear relationships in real estate data.

2 Random Forest Regressor: This model significantly improved accuracy over Linear Regression. It effectively managed complex feature interactions and reduced overfitting, resulting in lower error metrics.

3 Gradient Boosting Regressor: Achieving the lowest error metrics, this model demonstrated the highest accuracy. Its sequential learning approach effectively corrected prior errors, although it requires careful tuning to avoid overfitting.

7 Conclusion/Future Work

In this project, we successfully developed predictive models for house prices based on features such as square footage, number of bedrooms, bathrooms, garage size, and the year built. Among the models evaluated, the linear model algorithm outperformed the others, achieving the lowest RMSE and demonstrating its effectiveness in capturing complex patterns within the dataset. This superiority can be attributed to its ability to combine predictions from multiple weak learners, enhancing overall accuracy while minimizing overfitting. In contrast, the Linear Regression model, while simple and interpretable, struggled with the intricate relationships present in the data, leading to higher error metrics.

Looking ahead, several avenues for future work present themselves. With additional time and resources, we could explore more advanced machine learning techniques, such as deep learning models, which may offer even greater accuracy by leveraging more complex architectures. Incorporating additional features, such as local market trends, economic indicators, and geographical data, could enhance the models further. Additionally, employing larger datasets and more computational power would allow for thorough hyperparameter optimization and potentially yield better-performing models. Expanding the research to include real-time price prediction capabilities could provide valuable tools for stakeholders in the real estate market, ultimately leading to more informed decision-making.

8 Appendices

(***Appendix A: Dataset Description***

The dataset for this project includes features relevant to house pricing, such as:

- ***Square Feet Area****: Total area of the house.*
- ***Number of Bedrooms****: Total count of bedrooms.*
- ***Number of Bathrooms****: Total count of bathrooms.*
- ***Garage****: Indicator of garage presence (Yes/No).*
- ***Year Built****: Year the property was constructed.*
- ***House Age****: Derived feature calculated as the difference between the current year (2024) and the year built.*

Appendix B: Data Preprocessing Steps

1. ***Handling Missing Values****:*
    - *Missing data was addressed using mean imputation.*
2. ***Feature Engineering****:*
    - *Created the* **house age** *feature and removed the original Year Built variable.*
3. ***Feature Selection****:*
    - *Removed irrelevant columns like Lot Size and Neighborhood Quality to focus on significant variables.*

Appendix C: Model Implementation

The following models were implemented using Python's scikit-learn library:

1. ***Linear Regression****: Selected for simplicity and interpretability.*
2. ***Random Forest Regressor****: An ensemble method that enhances accuracy through multiple decision trees.*
3. ***Gradient Boosting Regressor****: Captures complex patterns by correcting errors from previous trees.*

Appendix D: Evaluation Metrics

Models were evaluated using:

- ***Mean Absolute Error (MAE)****: Average magnitude of errors in predictions.*
- ***Mean Squared Error (MSE)****: Average of the squares of the errors.*

- ***Root Mean Squared Error (RMSE)****: Square root of MSE, in the same units as the target variable.*

Appendix E: Results Summary

Model	***Mean Absolute Error (MAE)***	***Mean Squared Error (MSE)***	***Root Mean Squared Error (RMSE)***
Linear Regression	*X.XX*	*X.XX*	*X.XX*
Random Forest Regressor	*X.XX*	*X.XX*	*X.XX*
Gradient Boosting Regressor	*X.XX*	*X.XX*	*X.XX*

(Note: Replace X.XX with actual values from your experiments.)

Appendix F: References

- • M. H. Rafiei and H. Adeli, "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units," *Journal of Construction Engineering and Management*, vol. 142, no. 2, 2015, doi: 10.1061/(ASCE)CO.1943-7862.0001047.
- • M. Yazdani, "Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction," arXiv:2110.07151 [econ.EM], 14 Oct. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2110.07151.
- • F. D. Calainho, A. M. van de Minne, and M. K. Francke, "A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate," *Journal of Real Estate Finance and Economics*, vol. 68, pp. 624–653, Apr. 2022.
- • M. Monson, "Valuation Using Hedonic Pricing Models," Cornell University, MBA thesis, 2009.
- • L. H. T. Choy and W. K. O. Ho, "The Use of Machine Learning in Real Estate Research," *Land*, vol. 12, no. 4, p. 740, Mar. 2023, doi: 10.3390/land12040740.

9 Contributions *[This information must be there in your report]*

- Data Collection and Preprocessing: Ayush collected datasets; Sai preprocessed the data and handled missing values.
- Model Development and Tuning: Ayush developed the baseline model; Sai implemented ensemble techniques and tuned hyperparameters.
- Data Analysis and Visualization: Ayush analyzed results and created visualizations; Sai interpreted results and contributed to visual representations.
- Report Writing: Ayush drafted initial sections; Sai reviewed and refined the report for clarity.

10 References/Bibliography

[1] M. H. Rafiei and H. Adeli, "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units," *Journal of Construction Engineering and Management*, vol. 142, no. 2, 2015, doi: 10.1061/(ASCE)CO.1943-7862.0001047.

- [2] M. Yazdani, "Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction," arXiv:2110.07151 [econ.EM], 14 Oct. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2110.07151.
- [3] F. D. Calainho, A. M. van de Minne, and M. K. Francke, "A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate," *Journal of Real Estate Finance and Economics*, vol. 68, pp. 624 – 653, Apr. 2022.
- M. Monson, "Valuation Using Hedonic Pricing Models," Cornell University, MBA thesis, 2009.
- L. H. T. Choy and W. K. O. Ho, "The Use of Machine Learning in Real Estate Research," *Land*, vol. 12, no. 4, p. 740, Mar. 2023, doi: 10.3390/land12040740.