NJCU
NEW JERSEY CITY UNIVERSITY

Predicting Housing Prices with Machine Learning

Instructor
Dr. Michael Hudson
Student
Ngan Huynh
Date: Sep 1st, 2025

Abstract. This study focused on a comparative analysis between two models, Linear Regression and Decision Tree with pruning, for the problem of house price prediction. Using the dataset of factors influencing house prices, I applied data processing techniques and model training. The performance of the two models was evaluated based on error metrics as MSE, MAE, and the coefficient of R2. The results show that both models are capable of predicting house prices, with linear regression showing slightly better performance compared to the Decision Tree. The analysis highlights why these differences occur, discusses possible improvements, and considers the definition of "best" performance.

Keywords: housing prices, linear regression, decision tree, machine learning, model evaluation

1. Introduction

In the real estate sector, accurate house price prediction is a complex yet valuable challenge. It can assist buyers and sellers in making informed decisions, as well as help real estate professionals analyze market trends. This study focuses on applying two machine learning models to solve this problem: Linear Regression and Decision Tree. The objective is to compare the performance of these two models to determine which is more suitable for the selected dataset.

This study has developed a model capable of predicting housing prices using Machine Learning. The model delivers predictions with about 61% accuracy. The model relies on historical data such as area, bedrooms, bathrooms, stories, parking, and more. I researched and tested several Machine Learning algorithms to select the one that maximizes prediction accuracy.

1.1 Housing Market Analytics and Price Prediction

The dataset in this study contains approximately 545 residential properties, each described by a mix of numeric attributes (e.g., area, bedrooms, bathrooms, stories, parking) and categorical amenities (e.g., main road, air conditioning, furnishing status).

Table 1: Important KPIs for housing price Sourcing

KPI	Definition
price	Market price of the house.
area	Total the area of the house in square feet.
bedrooms	Number of bedrooms.
bathrooms	Number of bathrooms.
stories	Number of stories in the house.
mainroad	The house has access to a main road.
guestroom	The house includes a guest room.
basement	The house has a basement.
hotwaterheating	The house has hot water heating.
airconditioning	The house has air conditioning.
parking	Number of car parking spaces.
prefarea	The house is in a preferred residential area.
furnishingstatus	Furnishing level of the house: unfurnished, semi-furnished, furnished.

1.2 Housing Attributes and Market Dynamics

While prior work emphasizes the role of location in housing prices, this dataset focuses on structural features and amenities. These represent the tangible aspects of a property that directly affect perceived quality and cost. For example, a larger area and more bathrooms typically increase price, while air conditioning or furnishing status adds premium value.

Although the absence of geographic information is a limitation, the dataset remains a valuable case study for comparing algorithmic approaches to price modeling.

2. Literature Review

The dataset is coming from Kaggle. The dataset is publicly available on Kaggle from 4 years ago. I first read my dataset into a pandas data frame called df, and then I use the head() function to show the first five records from my dataset.

The dataset contains 13 variables, and I use ‘price’ as a target variable.

These feature variables will help me predict whether price increases or decreases: area, bedrooms, bathrooms, stories, mainroad, guestroom, basement, hotwaterheating, airconditioning, parking, prefarea, furnishingstatus.

This study contributes to the literature by directly comparing Linear Regression and a pruned Decision Tree Regressor on a Kaggle housing dataset. By evaluating model performance using multiple error metrics and visual diagnostics, the study provides insights into the practical balance between interpretability and predictive capability, offering guidance for analysts and lenders in real estate decision-making.

3. Business Goals and Problem Definition

3.1 Business Goals

The primary goal of this study is to build predictive models for housing prices that balance accuracy and interpretability. Stakeholders in the housing market as buyers, sellers. This project requires reliable forecasts that not only deliver precise price estimates but also provide insight into the drivers of those predictions.

The analysis is designed to achieve three goals:

The first thing establish a baseline model using Linear Regression to calculate the contribution of property attributes as area, bedrooms, bathrooms, and stories to housing prices.
The second thing implement a more flexible predictive model using a Decision Tree with pruning to capture non-linear relationships and interactions between features while controlling for overfitting.
The final thing compare and interpret results to assess trade-offs between model interpretability and predictive power, thereby providing actionable insights for practical decision-making.

3.2 Modeling Environment

The study leverages Python to conduct the analysis. Data preprocessing and exploratory analysis were performed using Pandas, NumPy, Matplotlib, and predictive modeling employed scikit-learn implementations of Linear Regression and Decision Tree. The dataset consists of about 545 records and 13 features.
The dataset was split into 80% for training and 20% testing. Calculation metrics consist of Mean Squared Error, Mean Absolute Error, and R2 Score, complemented by residual analysis and visualization of actual versus predicted prices.

4. Solution Methodology

4.1 Data models

The dataset used in this study was obtained from Kaggle’s Housing Prices dataset. The dataset includes an entity is a house, and attributes are price (target variable), area, bedrooms, bathrooms, stories, and several categorical features as mainroad, guestroom, basement, airconditioning, parking, and furnishingstatus.

4.2 Machine Learning Models

I apply the training/test method with two models as Linear Regression and Decision Tree with pruning. After running the models, I will choose and pick the best model based on the accuracy score.

5. Solution Results

5.1 Performance Metrics

The predictive price house performance of both models was evaluated on training and test data

Table 2: Training/Testing performance (R2, RMSE, MAE).

Model	Train R2	Test R2	Test MSE	Test MAE
Linear Regression	0.699 (69%)	0.588 (58%)	1,211,349	849,809
Decision Tree	0.732 (73%)	0.333 (30%)	4,870,169	1,085,615

The results indicate that Linear Regression achieved better generalization, with a test R2 of 0.588 (58%), compared to 0.333 (33%) for the Decision Tree with pruning. Although the Decision Tree obtained higher training performance (R2 = 0.733 (73%)), the sharp drop on the test set shows clear signs of overfitting. In contrast, Linear Regression maintained more consistent performance across training and testing.

Figure 1. Actual vs.Predicted — Linear Regression

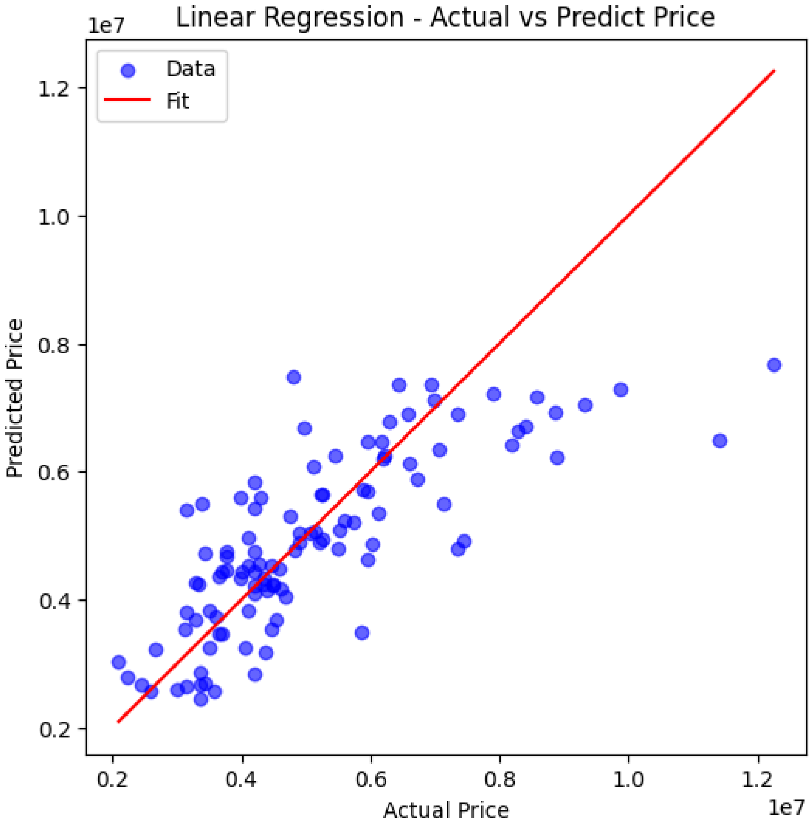
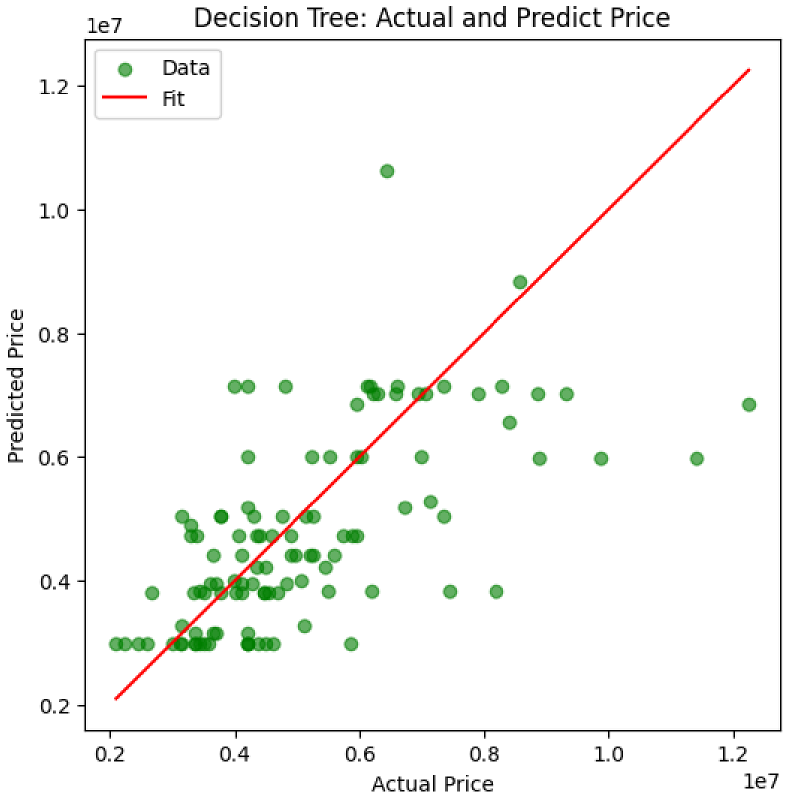


Figure 2. Actual vs. Predicted — Decision Tree



5.2 Learning curves

Learning curves were generated to evaluate model performance as the size of the training data increased. These plots show both the training R2 and the cross-validation R2 as a function of training size. I use tools because they help identify whether a model is limited by bias or by variance.

When the training and validation curves converge at a relatively low R2 value, the model suffers from high bias, which indicates underfitting. Conversely, when there is a large and persistent gap between the training and validation curves, the model exhibits high variance, which indicates overfitting.

Figure 3. Learning Curves - Linear Regression

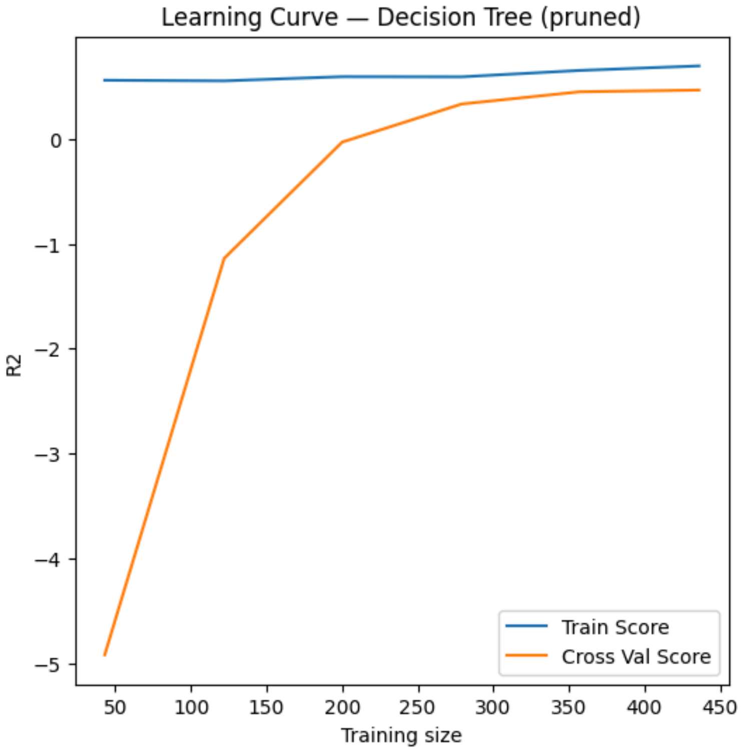
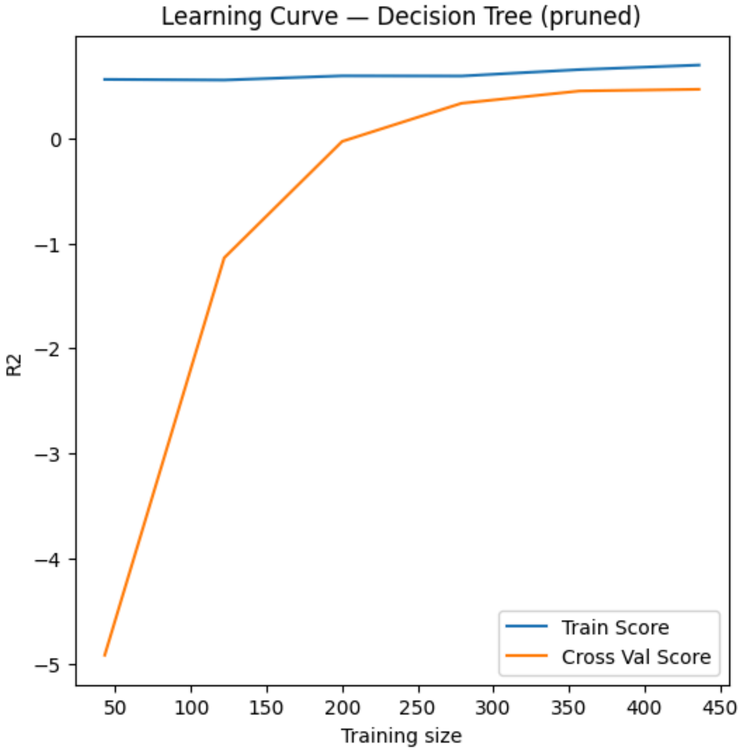


Figure 4. Learning Curves - Decision Tree



5.3 Timing

Both models trained and predicted extremely quickly, with execution times measured in milliseconds. Linear Regression required approximately 0.0022 seconds to fit and 0.00008 seconds to make predictions. The Decision Tree required 0.00215 seconds for fitting and 0.00117 seconds for prediction.

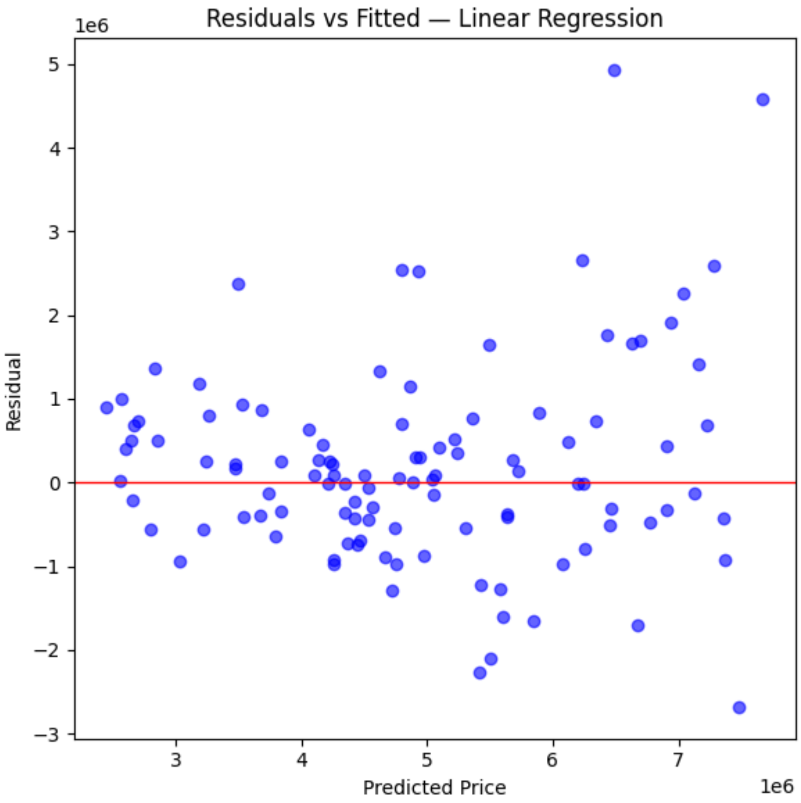
Although both times are negligible in the real world, Linear Regression was slightly faster at prediction, while the Decision Tree required additional traversal through its branches.

5.4 Residual

Residual analysis provides insight into the quality of model predictions.

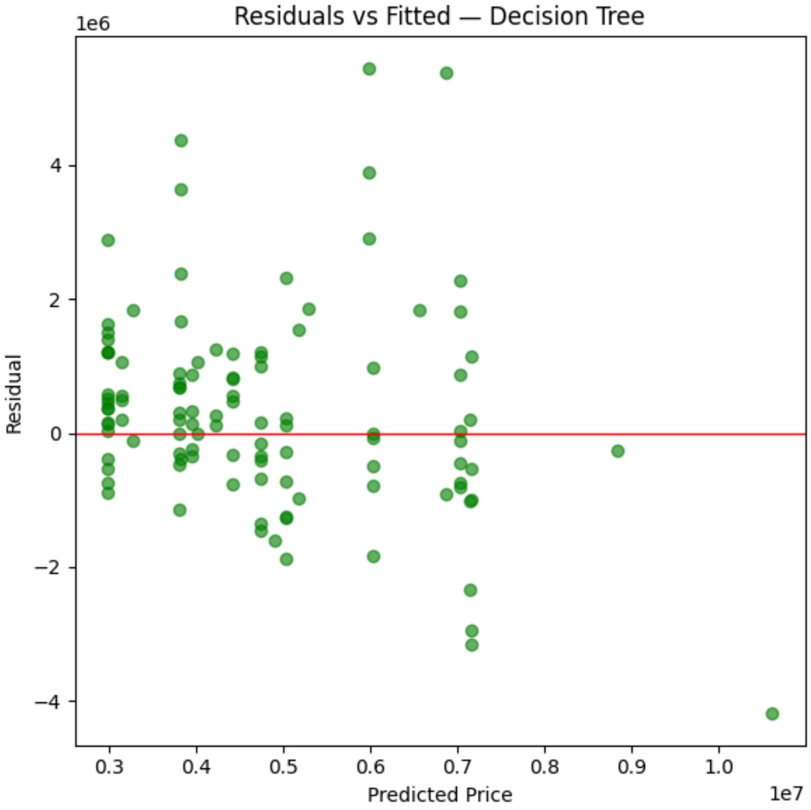
For Linear Regression, a random scatter like this is a sign of a good model. It shows that residuals tended to widen at higher predicted prices, which suggests that the variance in house prices increases with value.

Figure 5: Residuals vs Fitted — Linear Regression



For the Decision Tree, this is a sign of a poor model fit and indicates that the tree was overfit to the training data, suffering from high variance. The model's errors are systematic, not random, which compromises its ability to generalize beyond the training data.

Figure 6: Residuals vs Fitted — Decision Tree



6. Conclusion and future works

This study compared the performance of Linear Regression and a Decision Tree Regressor with pruning for predicting housing prices. The results demonstrated that Linear Regression generalized more effectively to unseen data, achieving a test R2 of 0.588(58%), while the Decision Tree overfitted the training set and achieved a lower test R2 of 0.333(33%). Learning curves confirmed that Linear Regression was bias-limited but stable, whereas the Decision Tree remained variance-driven even after pruning. Residual analysis provided further insights, with Linear Regression showing heteroscedasticity and the Decision Tree producing step-like residuals.

In this study, we limited our scope to comparing Linear Regression and a pruned Decision Tree on a relatively small housing dataset. However, the same methodology could be applied to larger and more diverse datasets, as long as sufficient and reliable data are available. The factors we identified as impactful for predicting housing prices, such as property size, location attributes, and amenities, remain relevant across different regions and markets. Based on the current analysis, several directions can improve result predictive performance. Future work could include applying a transformation to the target variable as using a log transformation, which I think will help reduce heteroscedasticity and improve the fit of Linear Regression.

Acknowledgments
This study was developed as part of an applied analytics study on predictive modeling for housing prices. Thank the Kaggle community for providing open access to the Housing Prices dataset, which served as the foundation for this research. I also acknowledge the developers and contributors of open-source tools such as Python, pandas, scikit-learn, and matplotlib, which enabled efficient data processing, visualization, and modeling.

References

Varun T, (2024). House Price Prediction with Machine Learning.

M Yasser H (2021). Housing prices dataset. In
https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Indhumathy, C. (2020) Building Linear Regression in Python. In
https://towardsdatascience.com/linear-regression-in-python-a4cfbab72c17/

Shaw, T. (2023). Decision Trees: Introduction & Intuition.
https://towardsdatascience.com/decision-trees-introduction-intuition-dac9592f4b7f/

Appendix A. Code Snippets

Key functions for metrics calculation (RMSE, MAE, R²).
Example of the code used for training Linear Regression and Decision Tree.
A short snippet for generating learning curves.

Appendix B. Additional Figures and Tables

Full-size versions of your Actual vs. Predicted plots (Figures 1 and 2).
Learning curves (Figures 3 and 4)
Residual plots (Figures 5 and 6)
Important KPIs for housing price Sourcing (Table 1)
Training/Testing performance (Table 2)