Babeș-Bolyai University

HOUSING PRICE PREDICTION USING DECISION TREES, GRADIENT BOOSTING, AND K-NEAREST NEIGHBORS

Author: Turcu Ciprian-Stelian

INTRODUCTION

Purpose of the Study:

- Accurate housing price prediction is vital for real estate planning and investment decisions.
- Advanced data analysis reveals patterns in property attributes influencing prices.

Methods Used:

- Decision Trees.
- Gradient Boosting.
- K-Nearest Neighbors.

Key Questions:

- Which model provides the highest predictive accuracy?
- What are the strengths and limitations of these methods?

DATASET

Dataset :

Housing Price Prediction.

Dataset Description:

- 545 entries with 13 numeric and categorical features.
- Examples of features: price, area, bedrooms, furnishing status.

Preprocessing:

- Standardization, through scaling of features
- Encoding for categorical variables.
- Handling missing data.

THEORETICAL PART

1

Decision Tree Regression:

- Splits data into subsets using feature-based decisions.
- Pros: Interpretability, robustness to outliers.
- Cons: Prone to under/overfitting.

2

Gradient Boosting Regression:

- Combines weak learners to improve predictive accuracy.
- Pros: High accuracy, handles complex patterns.
- Cons: Computationally intensive.

3

K-Nearest Neighbors (KNN):

- Predicts based on the mean of nearest neighbors.
- Pros: Simplicity, non-parametric.
- Cons: Sensitive to scaling, hyperparameter tuning.

EVALUATION

Metrics Used:

- Mean Squared Error (MSE).
- R-squared (R^2) Score.

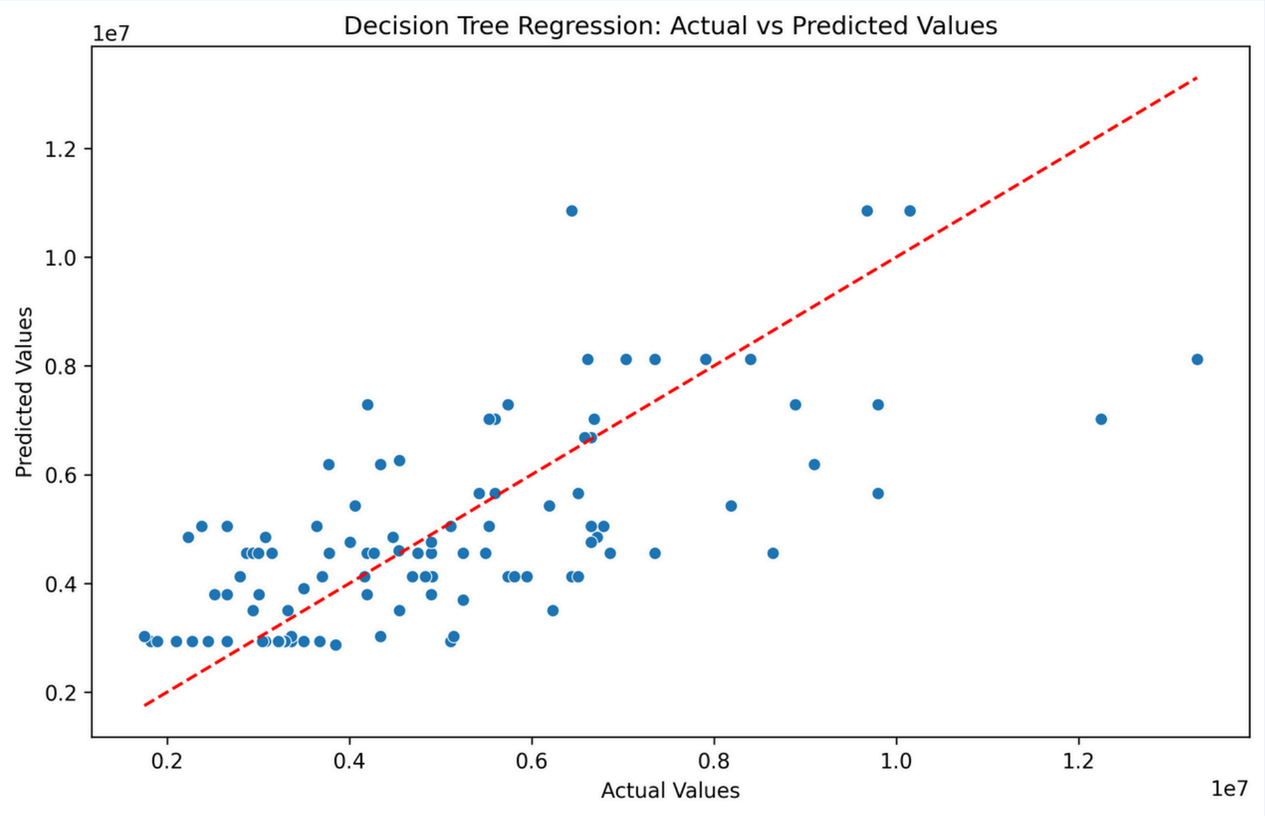
Training and Testing:

- Dataset split into 80% training and 20% testing.

ANALYSIS RESULTS

Decision Tree

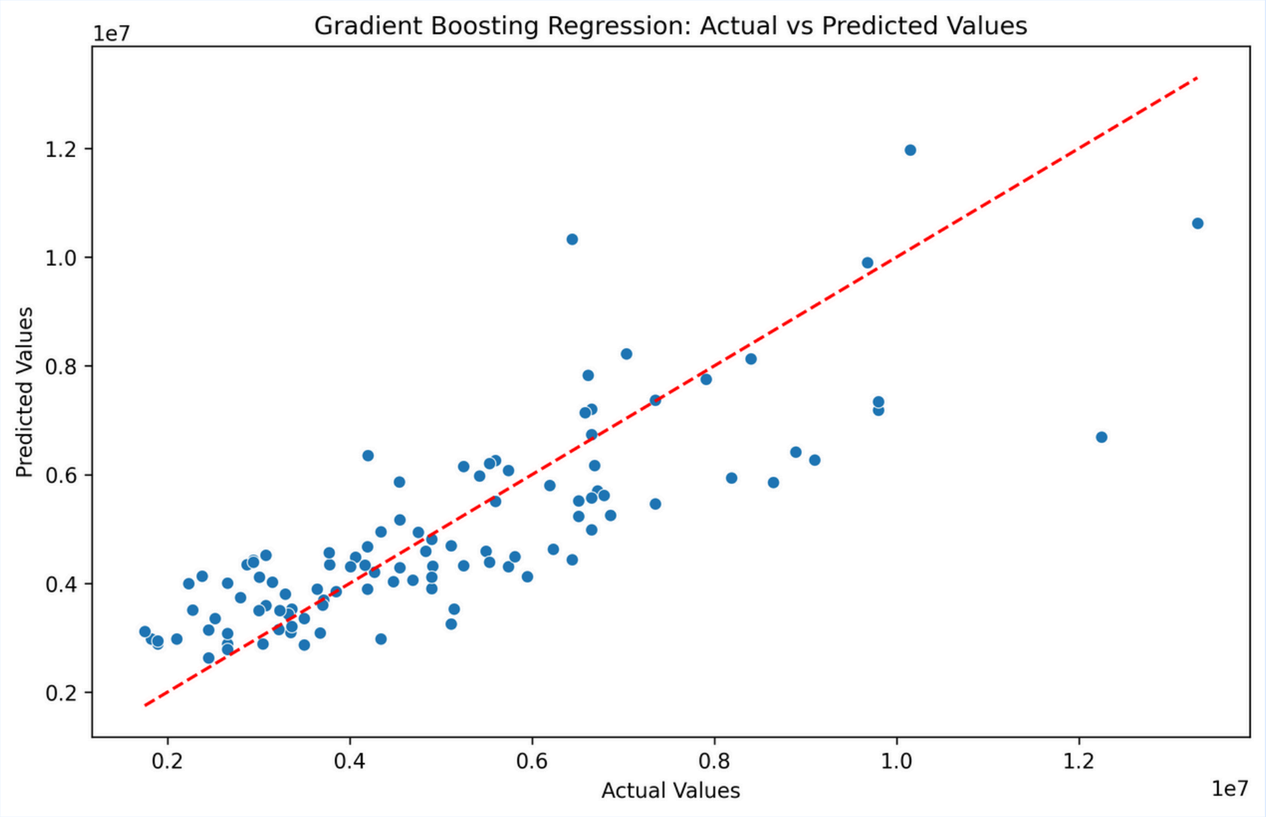
- R^2 = 0.465.
- MSE = 2.70 trillion
- Key Insight: Limited variability explained, prone to underfitting/overfitting.



ANALYSIS RESULTS

Gradient Boosting

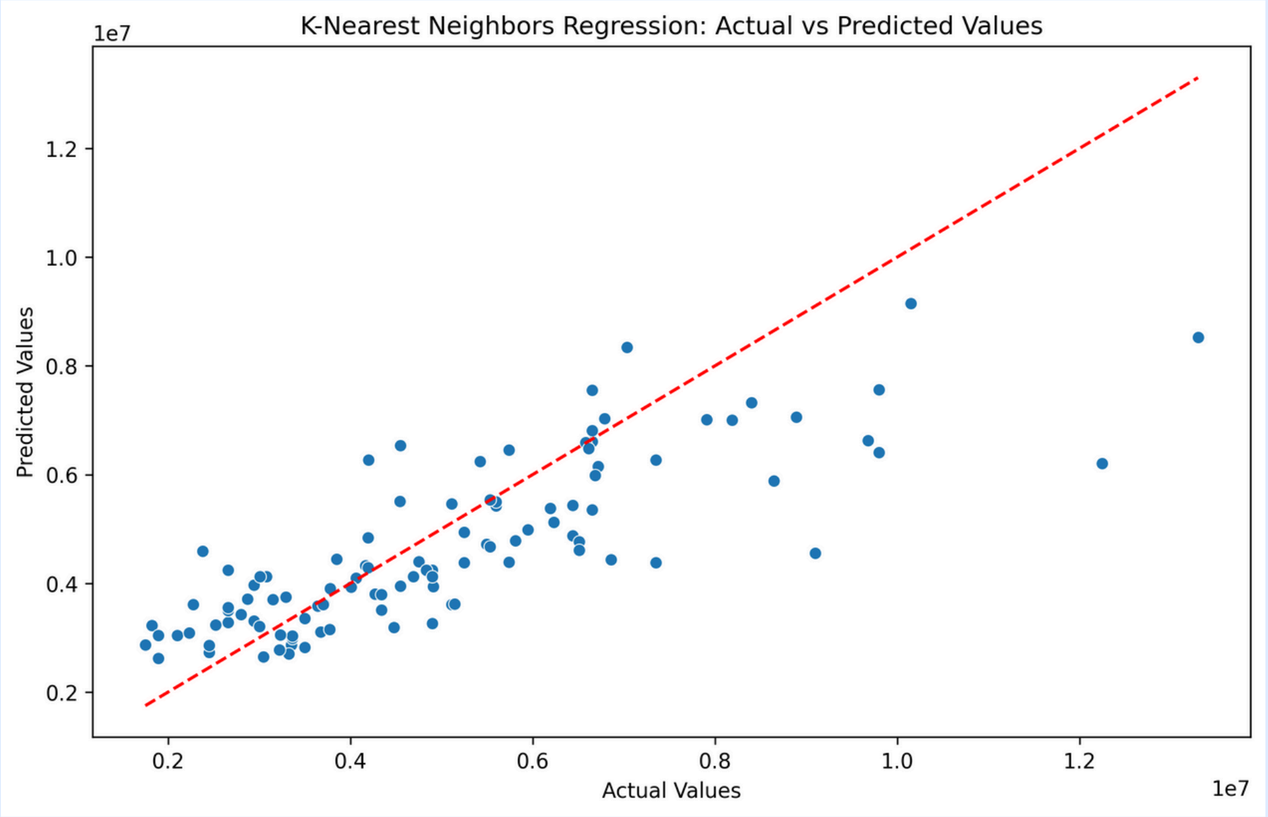
- R^2 = 0.665.
- MSE = 1.69 trillion
- Key Insight: Best performance, captured complex relationships.



ANALYSIS RESULTS

K-Nearest Neighbors (KNN)

- R^2 = 0.613.
- MSE = 1.96 trillion.
- Key Insight: Moderate performance, influenced by parameter tuning.



COMPARATIVE ANALYSIS

Model Performance:

- Gradient Boosting > KNN > Decision Trees.

Strengths and Weaknesses:

- Gradient Boosting excels in accuracy but is computationally expensive.
- KNN is versatile but sensitive to scaling.
- Decision Trees are interpretable but prone to underfitting/overfitting.

CONCLUSIONS & FUTURE WORK

Key Takeaways:

- Gradient Boosting is the most effective model for housing price prediction.
- Careful preprocessing and hyperparameter tuning are critical.
- Advanced data analysis uncovers key trends and relationships in real estate markets.

Improving Performance:

- Feature engineering to capture non-linear relationships.
- Experimentation with advanced models (e.g., XGBoost, LightGBM).
- Addressing multicollinearity among features.
- Expanding the dataset for better generalization.

Future Directions:

- Explore hybrid models and ensemble techniques.
- Leverage larger, more diverse datasets.

Babeș-Bolyai University

THANK YOU

Turcu Ciprian-Stelian | 2025