Report on Predictive Modeling for Nashville Real Estate Investment

A Comparative Analysis of Regression Algorithms

Daniel Makala Mackaemba
Northeastern University
Febuary 4, 2026

Abstract

Identifying undervalued real estate properties is critical for maximizing investment returns in competitive housing markets. This study evaluates multiple predictive modeling approaches to estimate whether properties in Nashville sell above or below their assessed value. Using a cleaned and preprocessed dataset of recent housing transactions, four regression-based models—Linear Regression, Decision Tree, Random Forest, and Gradient Boosting—were developed and compared. Model performance was evaluated using RMSE, R^2, and MAE. Results indicate that Gradient Boosting provides the most accurate and stable predictions, making it the preferred model for identifying high-value investment opportunities.

1. Introduction

The Nashville real estate market has experienced rapid growth, creating both opportunity and risk for investors. While assessed property values provide a baseline, actual sale prices often diverge significantly due to location, structural characteristics, market dynamics, and buyer behavior. Accurately predicting these deviations allows investors to identify properties that are undervalued and avoid overpriced assets.

This project focuses on predicting the variable Sale Price Compared To Value, which captures the relative difference between sale price and assessed value. The objective is to build and compare multiple predictive models to determine which algorithm best supports data-driven investment screening. Four models are evaluated: Linear Regression, Decision Tree, Random Forest, and Gradient Boosting.

2. Data Description and Preparation

2.1 Dataset Overview

The dataset contains 56,644 residential property transactions from the Nashville area, including structural attributes, location indicators, and assessed values. The target variable, Sale Price Compared To Value, measures whether properties sell above or below assessment benchmarks.

2.2 Data Cleaning Strategy

To ensure modeling reliability, a structured data cleaning process was applied:

Duplicate Removal: Duplicate records were identified and removed to prevent biased learning.

Missing Value Treatment:

Numerical variables were imputed using the median, providing robustness against extreme values common in real estate data.

Categorical variables were imputed using the mode, preserving empirical distributions.

- **Outlier Assessment:** Outliers were identified using the Interquartile Range (IQR) method. Rather than removing them, all outliers were retained because extreme values often represent legitimate market conditions such as luxury or distressed properties.
- **Categorical Encoding:** One-hot encoding was applied to all categorical variables to ensure algorithm compatibility.
- **Feature Scaling:** Numerical features were standardized for Linear Regression to ensure coefficient comparability.
- **Train-Test Split**: The dataset was split into 70% training and 30% testing to evaluate generalization performance.

3. Modeling Approach

3.1 Linear Regression (Baseline Model)

Linear Regression was implemented as a baseline due to its interpretability. After feature scaling, the model achieved:

RMSE: 0.422

R^2: 0.028

The low R^2 reflects the complexity of housing markets and the influence of unobserved variables such as negotiation dynamics and property condition. However, coefficient analysis revealed meaningful drivers of price deviation, including location indicators, square footage, and structural attributes. Residual plots showed no strong systematic bias, supporting the model's validity as an interpretive baseline.

3.2 Decision Tree Regression

A Decision Tree model was trained with depth optimization using cross-validation. The optimal depth of seven produced:

RMSE: 0.424

R^2: 0.017

While performance was comparable to Linear Regression, Decision Trees exhibited mild overfitting and limited generalization. However, their hierarchical structure provided intuitive rule-based insights, identifying key split variables such as property size, location, and age.

3.3 Random Forest Regression

Random Forest was implemented to reduce overfitting through ensemble averaging. Despite testing multiple tree counts, the model produced:

RMSE: 0.466

R^2: –0.185

The negative R^2 indicates performance worse than predicting the mean. This result likely stems from high-dimensional feature space created by one-hot encoding and insufficient signal-to-noise ratio for ensemble aggregation. While predictive performance was poor, feature importance rankings remained informative for exploratory analysis.

3.4 Gradient Boosting Regression

Gradient Boosting achieved the strongest overall performance by sequentially correcting residual errors:

RMSE: 0.421

R^2: 0.034

MAE: 0.350

This model consistently outperformed alternatives across all metrics. Feature importance analysis highlighted structural and location-based attributes as the strongest predictors of sale price deviation. Residual diagnostics showed improved error stability, indicating better capture of non-linear interactions.

4. Model Comparison and Benchmarking

Model.	**RMSE**	**R^2**	**MAE**
Linear Regression	0.422	0.028	0.353
Decision Tree	0.424	0.017	0.358
Random Forest	0.466	-0.185	0.375
Gradient Boosting	0.421	0.034	0.350

Gradient Boosting outperformed all other models on every evaluation metric, demonstrating superior predictive accuracy and generalization.

5. Business Recommendation

Based on empirical performance and stability, Gradient Boosting is recommended for deployment as the company's investment screening model. Its advantages include:

- Highest predictive accuracy for identifying undervalued properties
- Robust feature importance rankings for focused due diligence
- Minimal overfitting and consistent test performance
- Ability to model complex non-linear relationships automatically

Predictions should be used to rank properties, not replace expert judgment. Properties predicted to sell well below assessed value should be prioritized for site visits and further inspection.

6. Limitations and Future Work

Despite improved performance, modest R^2 values indicate that many determinants of sale price deviations remain unobserved. Future enhancements should incorporate:

- Temporal pricing trends
- Neighborhood-level market indicators
- Property condition metrics
- GIS-based spatial features
- Economic and demographic variables

Stacked ensemble approaches may further improve accuracy.

7. Conclusion

This study demonstrates that machine learning models—particularly Gradient Boosting—can provide meaningful decision support in real estate investment contexts. While no model can fully capture market complexity, Gradient Boosting offers the most reliable framework for identifying high-value opportunities in the Nashville housing market and establishes a scalable methodology for future market expansion.