Sean McLean

ALY 6020

Module 4 Project

Investing In Nashville

Introduction

The focus of the assignment is to look at what factors contribute to the best values for a real estate company that has decided to make an investment in the growing Nashville, Tennessee area. Using a dataset from previous real estate sales, this can be accomplished using this week's module subject which centers on decision trees, gradient boosting, and random forest modeling. The dataset contains 22,652 cells with variables that provide the features of each home and some aspects of the previous sales of the homes. The target variable being evaluated is 'Sale Price Compared To Value' which will help determine which homes are being overvalued and undervalued in the Nashville market.

Part 1

In the data cleansing phase, the dataset was prepared for modeling by addressing missing values and removing irrelevant attributes. The target variable, "Sale Price Compared To Value," was converted from categorical to numeric form, mapping 'under' to zero and 'over' to one. The rows with missing target values were removed to ensure that only complete data was used. For numeric features, missing values were imputed using each column's median to maintain the data's distribution and reduce bias. Some of the irrelevant columns, such as 'property address' and 'sale date,' were alleviated to streamline the dataset and focus on relevant features for modeling.

The preprocessing steps included separate handling of the numeric and categorical features in the dataset. Numeric features were scaled to standardize their range, while categorical features were transformed using one-hot encoding to make them suitable for machine learning

models. This thorough preparation ensured that the data fed into the models was clean, consistent, and well-organized, laying a solid groundwork for building accurate predictive models (Anaconda, 2024).

Part 2

The linear regression model was designed to predict housing prices and determine key factors influencing these prices. The performance was assessed using mean squared error (MSE) and R-squared metrics, the coefficient of determination. The model achieved an MSE of 0.181 and an R-squared value of 0.038, indicating that it accounted for a little of the variance in housing prices. The low R-squared value suggested that the model had limited predictive power and was not effective in capturing the underlying patterns in the data (Anaconda, 2024).

The method served as a useful baseline for comparing with the other models being used in the module this week. The performance of the linear regression model also highlighted the limitations of using simple linear relationships for accurate property value predictions. These results pointed to the need for other modeling techniques to better understand and predict the factors affecting housing prices. A visual presentation of these models was executed to display and verify these relationships between the models (See Appendix A).

Part 3

The decision tree regressor was used to assess whether a more adaptable model could enhance prediction accuracy. The model, however, yielded a higher mean squared error (MSE) of 0.350 and a negative R-squared value of -0.858. These outcomes indicated subpar performance, with the negative R-squared value suggesting that the decision tree model performed worse than a

simple mean-based prediction. The elevated MSE further emphasized the model's inadequacy in capturing the variance of the target variable.

The performance of the decision tree highlighted its difficulties in handling the dataset's complexity. Unlike linear regression, the decision tree did not achieve improved predictive accuracy, indicating potential overfitting or an inability to discern critical patterns in the data. This pointed to the necessity of exploring some of the other methods that could achieve better results (Anaconda, 2024).

Part 4

The random forest regressor was implemented to improve model performance through ensemble learning. This model after being executed showed better results compared to the decision tree, with a mean squared error (MSE) of 0.204. Despite this improvement, it still produced a negative R-squared value of -0.084, indicating that it did not effectively capture the variability in housing prices. While the random forest model outperformed the decision tree, it emphasized the need for different techniques to garner more accurate predictions.

The random forest results demonstrated that although ensemble methods can enhance performance, they still struggled with explaining the variance in the dependent variable. This provided more awareness for further model refinement or the exploration of alternative approaches to achieve more dependable predictions (Anaconda, 2024).

Part 5

The gradient boosting regressor was evaluated next with the aim of achieving even better predictive accuracy. The model performed the best among the ones tested with an MSE of 0.176

and an R-squared value of 0.066. Although the R-squared value was still relatively low, the gradient boosting model showed an improvement over the previous models. The results also suggested that it was more capable of capturing the underlying patterns in the data compared to other models. Despite its better performance, the modest improvement highlighted the potential for further model refinement or feature engineering to enhance prediction accuracy (Anaconda, 2024).

Part 6

The model comparison involved assessing and visualizing the performance of linear regression, decision tree, random forest, and gradient boosting models. Performance metrics, including mean squared error (MSE) and R-squared values, were plotted to highlight each model's strengths and weaknesses. Among the models, gradient boosting exhibited the best performance, though all models faced challenges in accurately predicting housing values.

Feature importance visualizations revealed that key features such as "Building Value," "Finished Area," "Acreage," and "Year Built" were consistently important across the decision tree, random forest, and gradient boosting models (See Appendix B & Appendix C). The comparison and visualizations provided insights into the relative effectiveness of each model, with gradient boosting being the preferred choice due to its superior performance. This is chosen despite the need for further enhancements to fully address the prediction challenges (See Appendix D).

The highest rated variables indicate that these features in homes are most valuable when looking at previous sales and can be used for future predictors of home purchases. An ideal home for buyers pertains to price, how much land on the property, the age of the home, and how

developed the area is around them. These qualities can be subjective as some people prefer new homes over old ones, more land than less, and being in a more developed neighborhood over having fewer homes around them. The value of the home though is usually universal in that most buyers care about the price of the home they are buying or attempting to purchase.

Conclusion

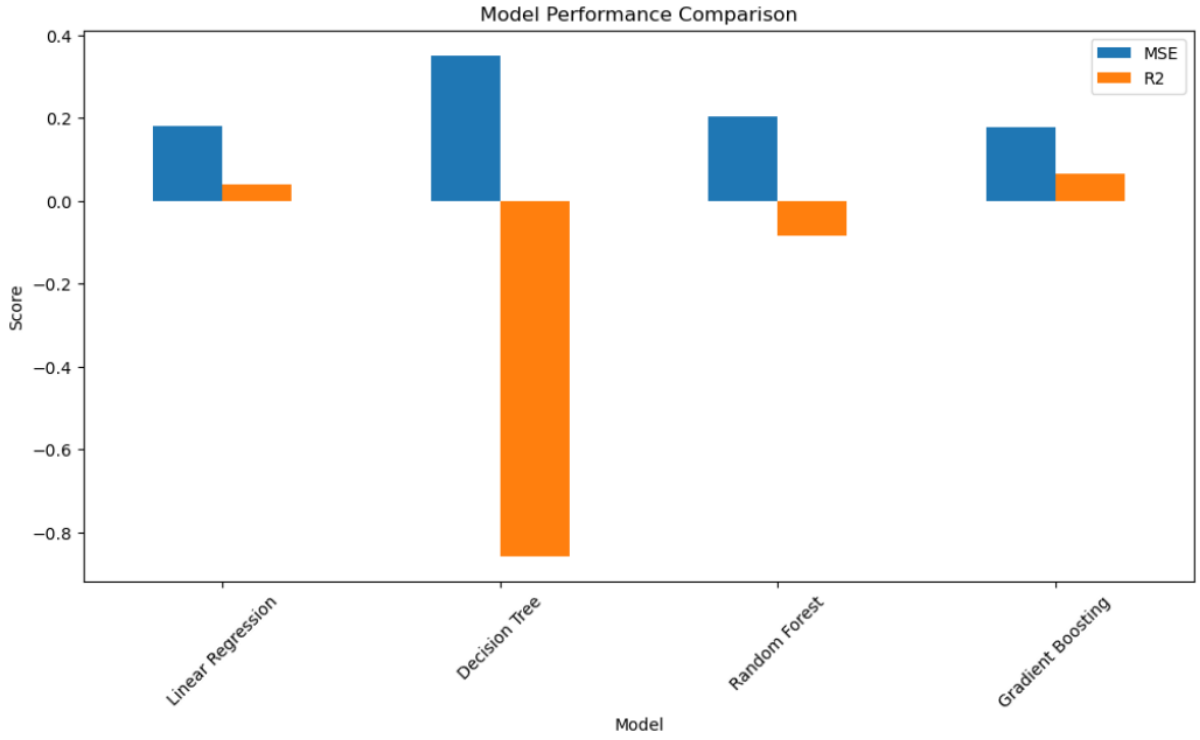
The models that were implemented provided a glimpse into how decision tree models can be an effective tool in highlighting real estate data for future housing sales. From the provided dataset the gradient boosting model was the strongest model from its mean squared error and coefficient of determination values. The strongest variables identified when looking at the dependent variable are the most predictive of whether a property is undervalued or overvalued. These features are recommended for looking at future insights into home buyer behaviors as well as future real estate sales in the Nashville area.

References

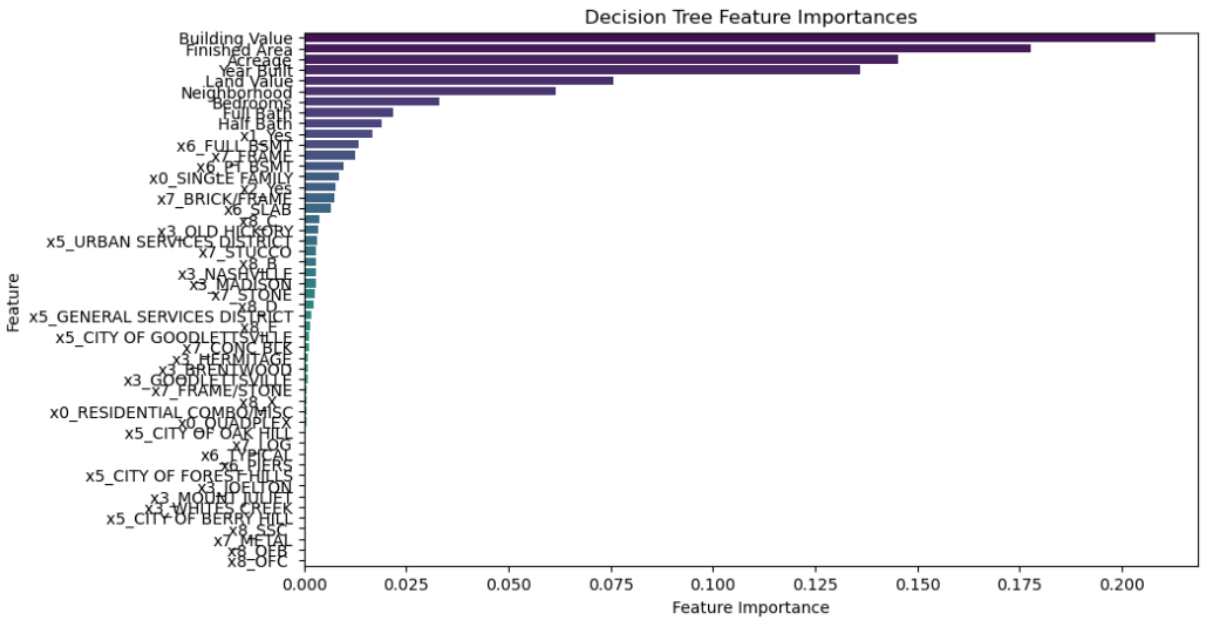
Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved August 4th, 2024, from https://www.anaconda.com

Appendix

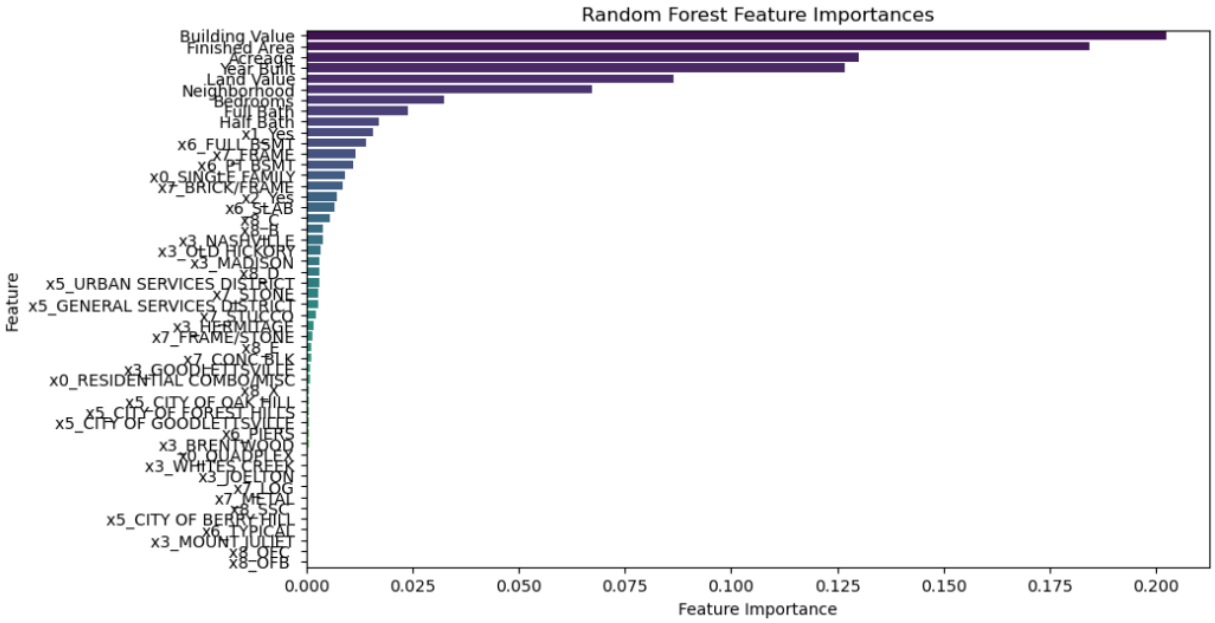
Appendix A



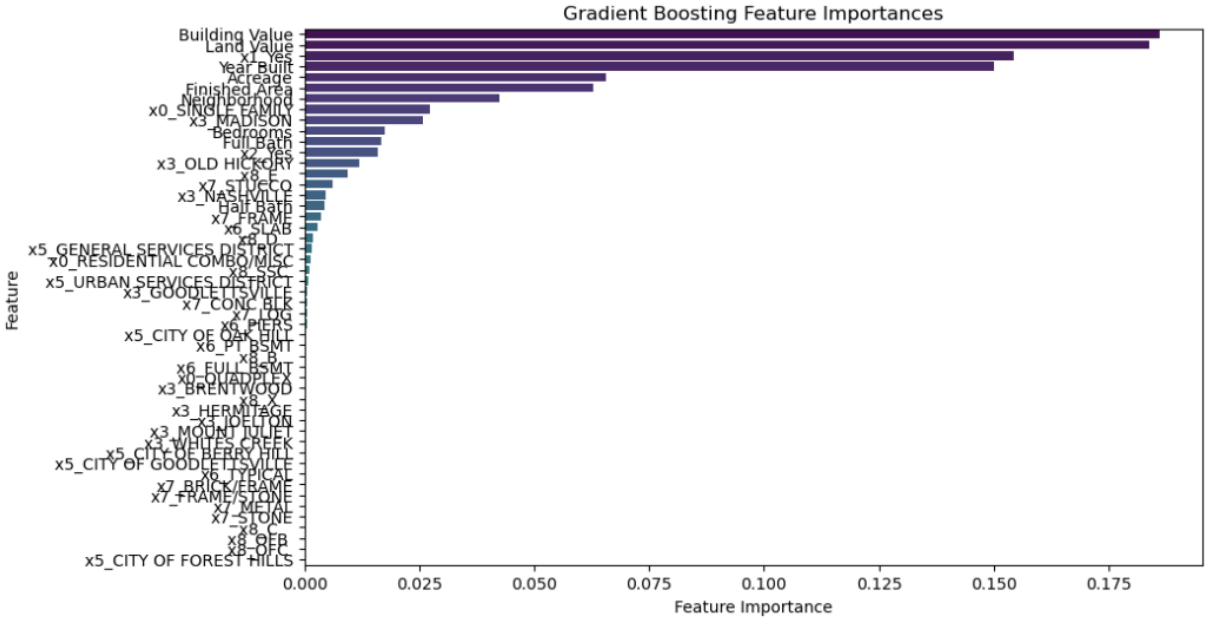
Appendix B



Appendix C


Random Forest Feature Importances
Building Value
Finished Area
Acreage
Year Built
Land Value
Neighborhood
Bedrooms
Full Bath
Half Bath
x1_Yes
x6_FULL BSMT
x7_FRAME
x6_PT BSMT
x0_SINGLE FAMILY
x7_BRICK/FRAME
x2_Yes
x6_SLAB
x8_C
x8_B
x3_NASHVILLE
x3_OLD HICKORY
x3_MADISON
x8_D
x5_URBAN SERVICES DISTRICT
x7_STONE
x5_GENERAL SERVICES DISTRICT
x7_STUCCO
x3_HERMITAGE
x7_FRAME/STONE
x8_E
x7_CONC BLK
x3_GOODLETTSVILLE
x0_RESIDENTIAL COMBO/MISC
x8_A
x5_CITY OF OAK HILL
x5_CITY OF FOREST HILLS
x5_CITY OF GOODLETTSVILLE
x6_PIERS
x3_BRENTWOOD
x0_QUADPLEX
x3_WHITES CREEK
x3_JOELTON
x7_LOG
x7_METAL
x8_SFC
x5_CITY OF BERRY HILL
x6_TYPICAL
x3_MOUNT JULIET
x8_OFC
x8_OFB
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Feature Importance

Appendix D


Gradient Boosting Feature Importances
Building Value
Land Value
x1_Yes
Year Built
Acreage
Finished Area
Neighborhood
x0_SINGLE FAMILY
x3_MADISON
Bedrooms
Full Bath
x2_Yes
x3_OLD HICKORY
x8_E
x7_STUCCO
x3_NASHVILLE
Half Bath
x7_FRAME
x6_SLAB
x8_D
x5_GENERAL SERVICES DISTRICT
x0_RESIDENTIAL COMBO/MISC
x8_SFC
x5_URBAN SERVICES DISTRICT
x3_GOODLETTSVILLE
x7_CONC BLK
x7_LOG
x6_PIERS
x5_CITY OF OAK HILL
x6_PT BSMT
x8_B
x6_FULL BSMT
x0_QUADPLEX
x3_BRENTWOOD
x8_A
x3_HERMITAGE
x3_JOELTON
x3_MOUNT JULIET
x3_WHITES CREEK
x5_CITY OF BERRY HILL
x5_CITY OF GOODLETTSVILLE
x6_TYPICAL
x7_BRICK/FRAME
x7_FRAME/STONE
x7_METAL
x7_STONE
x8_OFC
x8_OFB
x8_OFD
x5_CITY OF FOREST HILLS
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Importance