**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 4

Nashville Housing Analysis

Jeff Hackmeister

6/15/2025

**INTRODUCTION**

For this study of the Nashville real estate market, we began with a dataset of recent transactions. The goal of the study is to develop a model for identifying properties that are either over or undervalued, leading to maximizing the return on investment in this market. The steps taken towards this goal will be to first conduct an exploratory analysis of the provided data, then perform any needed data cleaning and feature engineer steps deemed necessary, followed by building a series of predictive models and evaluating their performance against one another to determine the optimal model. The different models will include a linear regression, decision tree, random forest and gradient boost. In the end, a model will be developed and chosen to accurately predict if a property is over or undervalued and therefore determine how to invest in the market.

**DATA EXPLORATION AND MODEL PREPERATION**

To begin this project, all needed packages are loaded, these include pandas, numpy, matplotlib, statsmodels, and several tools from sklearn.

A computer code with black text

AI-generated content may be incorrect.

The dataset contains 22,651 entries for recently sold properties in the greater Nashville area. For basic data cleaning steps, any duplicative variables (such as Property City) are removed from the dataset. There were 108 null values for Half Bath, these are likely data entry errors that should be 0s, these were corrected. Then, the remaining 8 observations with null values are removed. To evaluate the distributions of each numerical value, a series of histograms was produced (see Figure 1 in Image Appendix). From these, several outlier values were identified. Rather than dropping these values, they were capped by determining the 1st and 99th percentile of each value and clipping any value outside of that range, this preserves the observation in the dataset without giving it outside influence on future models.

A computer code with black text

AI-generated content may be incorrect.

A screenshot of a computer screen

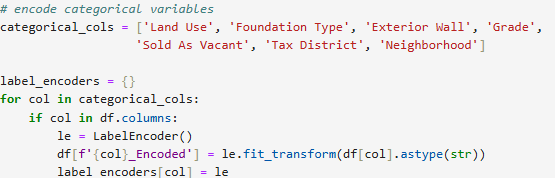
AI-generated content may be incorrect.

Additional feature engineering tasks were to combine land value and building value into a Total Assessed Value variable, calculating a price per square foot variable and total bathrooms (a sum of full and half bathroom counts).

A blue and orange pie chart

AI-generated content may be incorrect.With these steps completed, another series of visualizations were created comparing key variables to the binary Sale Price Compared to Value variable (Figure 2). A key observation here is that over 75% of the observations in the dataset were properties where the sales price was over the expected value. This indicates a so-called seller’s market, where buyers exceed the inventory of available properties and prices quickly rise due to the competition between buyers. While this rises prices across the board, it also represents opportunities to potentially find undervalued properties, make minor improvements and quickly “flip” or resell the property for a significant profit.

Next, the categorical values, such as Land Use and Foundation Type, were encoded as numeric values to be included in future analysis.

****

**CORRELATION ANALYSIS**

Before creating various models of study, a correlation matrix was created to explore relationships between variables and identify potential areas of multicollinearity (Figure 3).

A screenshot of a computer

AI-generated content may be incorrect.

From these correlation scores we can see that the encoded Tax District is the strongest predictor of a property being sold above assessed value (adding credence to the adage that the most important element of real estate is location, location, location!). Next was property age, a surprising result that will be watched in the modeling. At the other end of the report, properties sold as vacant were the strongest predictor of a property selling below assessed value. This is a potential area of interest for investment if retail buyers are turned off by vacant properties, they could be purchased below assed value, properly stagged and flipped. They also represent potential value buys to be converted into rental properties. It is important to note that all correlations are relatively weak, indicating that property values are complex and no single variable will be highly predictive.

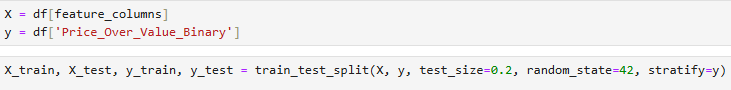
After exploring the correlation results, the feature variables were chosen for our models.

A white background with red text

AI-generated content may be incorrect.

**MODELING**

To begin the modeling stage of the project, the feature variables and the target variable were separated and then a training/test set were created with a standard 80/20 split. To keep the proportions of the data similar in both test and training sets, the stratify option was used. This ensures both splits contain roughly 75% of properties sold over assessed value as it was in the original dataset.



A screenshot of a computer code

AI-generated content may be incorrect.

The data was also scaled, meaning each numerical value is centered around a mean of 0 and standard deviation of 1. This ensures that variables with larger values do not have an outsized influence in the model over other variables that contain smaller values. These scaled values are critical to the linear regression model but will not be used for all techniques.

LINEAR REGRESSION

With the date property split and scaled, a linear regression was fit to the training data and used to make predictions on the test data and a series of diagnostic statistics were produced.

A close-up of a white background

AI-generated content may be incorrect.

A screenshot of a computer error message

AI-generated content may be incorrect.

Looking at the key features of the model, we again see Sold As Vacant at the top of the list, followed by Finished Area and Tax District.

A screenshot of a computer program

AI-generated content may be incorrect.

The predictive strength of this model is very low with an R² score of 0.0105. This indicates that the model can explain approximately 1.05% of the variances in the data. Given the dynamic and complex nature of a real estate market, it is unsurprising that a linear model would produce these types of results. The root mean squared error of 0.4295 also indicates a highly unreliable model. While this model is far from ideal, it serves as a helpful baseline for other modeling techniques to come. As George Box correctly noted “all models are wrong, but some are useful”.

DECISION TREE

Moving on to the decision tree model, the max depth is set at 10 with a minimum sample split of 20 for the same training data and the model is used to make predictions on the test data.

A screenshot of a computer program

AI-generated content may be incorrect.

The max depth of 10 is an attempt to balance the need to account for complexity in the data while avoiding overfitting the model to the training data. The minimum samples of 20 ensures at least 20 observations are required before making a split, thus limiting the influence of outlier values. Once again, diagnostic statistics are produced along with a list of the five most influential variables.

A screenshot of a computer error

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

The negative R² value immediately negates the viability of the model. While the regression model could explain less than 2% of the variances in the model, the decision tree actually performs worse than simply using the 75/20 split from the training dataset as the guide. While decision trees are more complex models, the added complexity does not add value in this use case. The changes in feature importance should also be noted, with Sold As Vacant dropping to fourth and Finished Area rising to the top.

RANDOM FOREST

With the disappointing results from the decision tree, we move on to a random forest approach. This ensemble method uses multiple decision trees (we will use 100) and averages the predictions from all for the final prediction. This method should be much more sensitive to the nuances of the real estate market in Nashville.

For the random forest model, once again the model is built from the training data and used to make predictions on the test data.

**A close-up of a computer screen

AI-generated content may be incorrect.**

And the same diagnostic statistics are produced.

**A screenshot of a computer error

AI-generated content may be incorrect.**

Much stronger results are produced from this model. The R² value of 0.0391 is modest but nearly a 4x improvement from the linear regression model. The RMSE value of 0.4229 also shows improvement from the linear regression and decision tree models. Looking into feature importance, the results are much closer to the single decision tree above.

**A screenshot of a computer program

AI-generated content may be incorrect.**

GRADIENT BOOST

The final model type will be gradient boost. While the random forest built 100 trees simultaneously, the gradient boost technique builds them one at a time, each tree learning from the errors from the previous tree.

A close-up of a computer code

AI-generated content may be incorrect.

For this model, another 100 trees will be created, the max depth will be lowered from the 10 used in the decision tree and random forest and the learning rate will be set to 0.1. The learning rate indicates that each tree essentially contributes 10% to the development of the model. This is a modestly conservative setting that ensures gradual improvement from one tree to the next.

Once the model is fit and predictions are made, the same diagnostic statistics are created.

A black text on a white background

AI-generated content may be incorrect.

The R² value for this model comes back at 0.0323 – close to but slightly lower than the 0.0391 achieved by the random forest approach. The RMSE of 0.04244 is also close but slightly behind (in this case a high value indicates worse performance) the random forest. Looking at the feature importance, there are a few notable differences. This is the first model to include Acreage in the top five and not to include Sold As Vacant. The gradient boost also produced the highest score for Price per SqFt of any model.

A screenshot of a computer screen

AI-generated content may be incorrect.

**MODEL INTERPRETATION**

After evaluating all models, we can compare the diagnostic statistics together.

A white paper with black numbers and lines

AI-generated content may be incorrect.

To further evaluate the performance of each model, a 5-fold cross validation was used. This tests each model’s consistency by running it against 5 different subsets of the training data.

A screenshot of a computer code

AI-generated content may be incorrect.

The if statements in the code ensure that the scaled data is used for linear regression while the unscaled data is again used for all other models. Using this cross-validation method further validates the predictability of the results from these models. To evaluate the results of the cross-validation, the mean R² values for each will be compared.

A white background with black text

AI-generated content may be incorrect.

Once again, the random forest model has the best results of the group with an R² of 0.0450 (+/- 0.0110) this is an improvement over the initial test R² of 0.039. The gradient boost model came in second with a mean R² of 0.0351 but also showed more unpredictability. Not only do these results show a clear performance gap between models but the initial performance of the random forest model was conservative with even better results in subsequent trials.

Additionally, the accuracy, precision, recall and F-1 scores of each model can be compared (also see Figure 4).

A screenshot of a computer screen

AI-generated content may be incorrect.

These results confirm the superiority of the random forest model in this study. The accuracy and precision scores are all slightly above the standard expectation that 75% of properties in Nashville will sell above the established values but in real estate investments, small improvements in prediction can lead to substation financial performance.

**CONCLUSION**

Based on the full analysis conducted, the random forest model is recommended for implementation for use in investment decsions in the Nashville market. Of all methods explored, this model provided the best results given the complex and nuanced nature of real estate investment analysis. Caution should be used in implementing the model in that is is highly efficient in identifying properties that are likely to be overvalued in the market but it more conservative in identifying under-valued properties, therefore the model should be used in conjunction with local experts who can provide additional instight into property values, outside of that available to the model in the data.

**REFERENCES**

[1] Box, G. E. P., & Norman Richard Draper. (1987). *Empirical model-building and response surface* (p. 424). Wiley.

‌

[2] GeeksforGeeks. (2024, February 22). *Random Forest Algorithm in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[4] OpenAI. (2024). ChatGPT (Claude 4) [Large language model]. https://claude.ai

‌ [5] Starmer, J. (n.d.). *Gradient Boost*[Review of *Gradient Boost*]. YouTube. https://www.youtube.com/watch?v=3CC4N4z3GJc

[6] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

**IMAGE APPENDIX**

Figure 1

A group of blue and white graphs

AI-generated content may be incorrect.

Figure 2

A group of graphs and diagrams

AI-generated content may be incorrect.

Figure 3

A colorful chart with numbers

AI-generated content may be incorrect.

Figure 4

A group of colorful graphs

AI-generated content may be incorrect.