Module 4 Project: Investing in Nashville

Sarvesh Thorve

College of Professional Studies, Northeastern University

Prof. Justin Grosz

**Introduction**

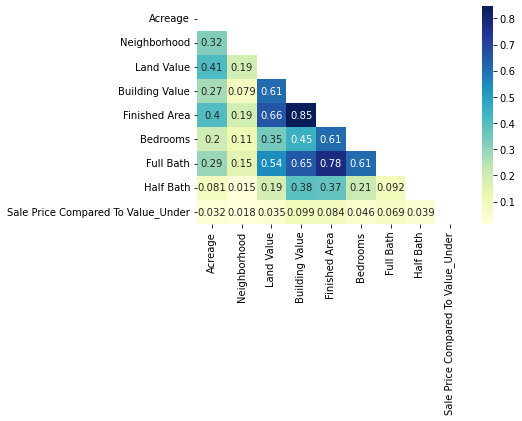
The Aim of this Project is to come up with actionable insights and recommendations to help the company to understand the key factors to classify a property as undervalued or overvalued to eventually come up with predictions to decide whether to invest in different kind of properties. The dataset has 22651 observations and a total of 26 variables. Predicting the results after developing and contrasting Logistic Regression, Decision Tree Models, Random Forest Models, and Gradient Boosting Classifiers to categorize the properties as undervalued or overpriced and discover what the crucial elements in locating the best deal may be. "Sale Price Compared To Value," the goal variable, indicates whether the property is undervalued or overvalued. There are many categorical variables, including "Land Use," "Sold As Vacant," "Multiple Parcels Involved in Sale," "City," "Tax District," "Foundation Type," "Exterior Wall," and "Grade." Acreage, Land Value, and Building Value are among the numerical variables.

# Data Cleaning

Chart

Description automatically generated with medium confidenceWe begin by eliminating the factors that are not important to our research. While checking the data types and null values to prevent contradictions in our results, we excluded variables such "Unnamed: 0," "Parcel ID," "Legal Reference," and "Property City" that are not relevant to our study. We decided to exclude some null values from the dataset because they made up less than 1% of the total data. There are several outliers in our data. Outliers need to be dealt with if they are the result of measurement mistake or if they have a big impact on the model's predictions. But, in our case it is not necessary to treat the outliers if they are actual observations that provide valuable information. For better analysis, certain subcategories in the categorical variables "Land Use" and "Foundation Type" were combined, such as "Piers" and "Typical" under the "Crawl" foundation type, and "Duplex" and "Quadplex" under another subcategory "Residential/Misc”. Our target variable “Sale Price Compared To Value” is unbalanced as there are more number of Overvalued properties than Undervalued. The Numerical variables are skewed and other numerical variables such as “building Value” and “Land Value” has several outliers. But, in this case we should not normalize the data because if the skewed data contains outliers, normalizing the data may exaggerate their impact on the results, thereby undermining the validity of the research. Additionally, normalizing the data may modify the original meaning of the values, making it harder to interpret the results. It could also lead to the loss of information, particularly if the data is significantly skewed.

## Analysis and Data Modeling

This report will complete further analysis by establishing models based on the preliminary conclusions drawn in the EDA after completing the preliminary analysis of the data set. In this section, the logistic regression model, Decision Tree model, Random Forest model, and Gradient boosting model will be used, and after the model is established, the results from these models will be compared to choose one for better insights. We begin by using a correlation plot to better understand our data. Some numerical variables such as ‘Finished Area” and “Building Value” have a high correlation with one another. As a result of high correlation between variables, multicollinearity can occur. The coefficients of the model may be difficult to interpret and may become unstable due to multicollinearity. As, these are the driving factors of Real Estate property value benchmarking, we should not handle the multicollinearity because it might tamper with our data and actual observations. The dataset had 1455 duplicate rows which were dropped. Upon checking for linear dependency in variables, we found out that there were no linearly dependent rows. One-hot encoding was done on variables 'Land Use’, ‘Sold As Vacant', 'Multiple Parcels Involved in Sale', 'City', 'Tax District', 'Foundation Type', 'Exterior Wall’, ‘Grade' for modeling purposes.

*Fig 1. Correlation Matrix*

### **Logistic Regression Model**

In this case, the logistic regression model has two advantages: To begin, the dependent variable " in the problem under consideration is a binary classification problem. Second, the model is extremely interpretable. By examining the weight of features, we can see how different features influence the outcome. We divide the data set into a training data set and a testing data set in a 3:2 split rather than a 4:1 split to avoid overfitting problems because there are very few observations of undervalued properties. Categorical variables are used to create dummy variables. Finally, a logistic regression model is developed for prediction. By interpreting the P value and coefficient of each variable in Appendix 2, we can draw some conclusions. Among the variables, "Land Value", "Building Value", "Full Bath”, “Land Use by single Family", "Property Sold as Vacant", "Multiple Parcels Involved in Sale", "City GOODLETTSVILLE" ,"City MADISON”, “City Joelton" ,"City NASHVILLE, “City OLD HICKORY" , all the Tax Districts, and “grade type E” houses have significant effects on the variable "Sale Price Compared To Value Under". Precision in classes 0 and 1 is 76% and 54%, respectively. It assesses how accurately the model predicts successful outcomes. A high precision indicates that the model predicts few false positives. As a result, our model correctly predicts undervalued properties only 54% of the time. The recall rates for classes 0 and 1 are 100% and 1%, respectively. It assesses the model's ability to detect every positive instance in the data. The accuracy is 76%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 86% for class 0 and 2% for class 1 which is undervalued properties. Many false positives are when many negative instances are mistakenly classified as positive. This means that our model considers the properties to be overvalued.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 6787 | 19 |
| F | 2187 | 22 |

#### Decision Tree Model

From the Decision Tree Model, Precision is 76% and 69% in classes 0 and 1, respectively. Our decision model correctly predicts undervalued properties 69% of the time based on significant features. Class 0 and 1 have 100% and 3% recall rates, respectively. The accuracy is 76%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 86% for class 0 and 5% for class 1, indicating that the properties are undervalued. As a result, our model outperforms the logistic regression model slightly. We examine the values assigned to each feature in the model to interpret the feature importance. These values represent the reduction in impurity achieved by splitting the data on that feature, such as Gini impurity. A higher value indicates that the feature is more important in predicting accurately. From the Decision Tree model in the appendix, “Sold as Vacant” is the most significant attribute followed by “Building Value” and “Bedrooms” and “Land Value”.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 6779 | 27 |
| F | 2150 | 59 |

**Random Forest Model**

From the Decision Tree Model, Precision is 76% and 91% in classes 0 and 1, respectively. Our decision model correctly predicts undervalued properties 91% of the time based on significant features. Class 0 and 1 have 100% and 2% recall rates, respectively. The accuracy is 76%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 86% for class 0 and 5% for class 1, indicating that the properties are undervalued. As a result, our model outperforms both the logistic regression and Decision Tree model slightly. From the Random Forest model, “Sold as Vacant” is the most significant attribute followed by “Building Value”, “Finished Area” and “Land Value” are significant key factors for determining the benchmarking of properties as overvalued or undervalued.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 6801 | 5 |
| F | 2157 | 52 |

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 6779 | 27 |
| F | 2150 | 59 |

**Gradient Boost Model**

From the Gradient Boost Model, Precision is 76% and 61% in classes 0 and 1, respectively. Our decision model correctly predicts undervalued properties 61% of the time based on significant features. Class 0 and 1 have 99% and 6% recall rates, respectively. The accuracy is 76%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 86% for class 0 and 11% for class 1, indicating that the properties are undervalued. As a result, our model outperforms all the models including logistic regression, Decision Tree model and Random Forest Model slightly because recall values are the highest for class 0 in this model. From the Gradient Boost Model, “Building Value” is the most significant attribute followed by “Land Value”, “Finished Area” and “Acreage” are significant key factors for determining the benchmarking of properties as overvalued or undervalued.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 6718 | 88 |
| F | 2071 | 138 |

Finally, we can compare the F1 score which is the weighted average of precision and recall and accuracy in the table below,

Table 1. Table for metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F-1 Score(0) - Overvalued** | **F-1 Score(1)- Undervalued** |
| Logistic Regression | 76% | 86% | 2% |
| Decision Tree | 76% | 86% | 5% |
| Random Forest | 76% | 86% | 5% |
| Gradient Boosting | 76% | 86% | 11% |

**Conclusion**

We were able to predict the key factors influencing the value of a property using the above study and models. A high F1 score indicates that the model has sufficient precision and recall. That is, the model makes a large proportion of correct positive predictions while making a small proportion of incorrect positive predictions. All the models in class 0 have a high F-1 score. As a result, the model can predict it correctly. Due to unbalanced data, F-1 scores for class "1" are extremely low. These factors are related because real estate prices are determined by a variety of key factors that are considered before pricing the property. Finally, we were able to deduce some significant key factors such as "Building Value," "Land Value," "Finished Area," and "Acreage" using our gradient boosting model, which is by far the most accurate model.

**Recommendation**

Using this Analysis, we can gain a thorough understanding of the key factors that influence the decision to label properties as overvalued or undervalued. Buying real estate involves considering a variety of factors to determine whether a property is a good investment. Furthermore, key stakeholders should prioritize metrics such as "Building Value," "Land Value," "Finished Area," and "Acreage" when considering an investment and categorize properties based on these factors.

**References:**

1. The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
2. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems" by Aurélien Géron

**Appendix**

1. Logistic Regression ModelTable

Description automatically generated

2. Decision Tree

Diagram

Description automatically generated