Text

Description automatically generated with low confidence

# ALY6020, 80406: Predictive Analytics

SPRING 2023

Module 4 Project

Investing in Nashville

Instructor’s name: Justin Grosz  
Date of submission: 17th June 2023

Prepared by: Rahul Agarwal

**1.1 INTRODUCTION:**

This assignment report focuses on the analysis of a housing dataset to assist a real estate company in making informed investment decisions in the thriving Nashville area. The company has provided a dataset containing recent sales information and seeks to build a predictive model that accurately identifies the best value deals. The key objective is to analyze the variable "Sale Price Compared To Value" to determine overvalued and undervalued properties. By developing an accurate model, the company aims to uncover the essential factors that contribute to finding the most favorable deals.

The data cleansing was performed to ensure the highest data quality. A logistic regression model, decision tree model, Random Forest model, and a Gradient Boost model were built and used to predict housing prices and identify the driving factors behind them. Then the results of each model were compared evaluating their performance in comparison to the other models. The report also emphasizes the importance of benchmarking metrics to compare and contrast the four models. Based on the findings, recommendations will be provided to the real estate company. These recommendations include evidence-backed justifications for selecting a particular model and identifying two key features of homes that the company should prioritize. The objective of this research is to provide real-world evidence of how these features impact the value of properties, supporting the decision-making process for successful real estate investments.

**1.2 ANALYSIS:**

**About Dataset and Data Cleaning:**

The dataset provided for the assignment consists of recent sales data in the growing Nashville area. The real estate company wants to build a model to identify the best value deals. The key variable of interest is "Sale Price Compared To Value," which helps determine if properties are over or under-valued. The dataset includes 26 variables such as Parcel ID, Land Use, Property Address, Suite/Condo #, Property City, Sale Date, Legal Reference, Sold As Vacant, Multiple Parcels Involved in Sale, City, State, Acreage, Tax District, Neighborhood, Land Value, Building Value, Finished Area, Foundation Type, Year Built, Exterior Wall, Grade, Bedrooms, Full Bath, and Half Bath. The dataset contains 22,651 rows and the target variable for analysis is "Sale Price Compared To Value."

We begin by removing elements that are irrelevant to our investigation. We omitted variables such as "Unnamed: 0," "Parcel ID," "Legal Reference," and "Property City" that were not relevant to our study when evaluating the data types and null values to avoid conflicts in our results. We opted to exclude certain null values from the dataset because they accounted for less than 1% of the total data. Our data contains some outliers. Outliers must be addressed if they are the consequence of a measurement error or have a significant influence on the model's predictions. However, if the outliers are genuine observations that give useful information, we do not need to address them. Certain subcategories in the categorical variables "Land Use" and "Foundation Type" were merged for better analysis, such as "Piers" and "Typical" under the "Crawl" foundation type, and "Duplex" and "Quadplex" under the "Residential/Misc" subcategory. Our goal variable "Sale Price Compared To Value" is uneven since there are more Overvalued than Undervalued properties. The numerical variables are skewed, and there are multiple outliers in other numerical variables such as "building Value" and "Land Value." However, we should not normalize the data in this scenario because, if the skewed data contains outliers, normalizing the data may exaggerate their influence on the conclusions, weakening the research's validity. Furthermore, normalizing the data may alter the original meaning of the results, making interpretation more difficult. It may also result in the loss of information, especially if the data is severely skewed.

**Data Analysis:**

Following the completion of the preliminary analysis of the data set, in this report, we will perform additional analysis by constructing models based on the preliminary results obtained in the EDA. In this part, the logistic regression model, Decision Tree model, Random Forest model, and Gradient boosting model will be employed, and the results from these models will be examined to see which one provides the best insights. To better comprehend our data, we begin with a correlation plot. Some numerical variables have a high connection, such as "Finished Area" and "Building Value." Multicollinearity may emerge as a result of a significant correlation between variables. Due to multicollinearity, the model's coefficients may be difficult to interpret and may become unstable. Because these are the main variables of Real Estate property value evaluation, we should avoid dealing with multicollinearity because it may tamper with our data and true observations. The dataset contains 1455 duplicate rows, which were removed. We discovered that there were no linearly dependent rows when we checked for linear dependency in variables. For modeling reasons, variables 'Land Use,' 'Sold As Vacant,' 'Multiple Parcels Involved in Sale,' 'City,' 'Tax District,' 'Foundation Type,' 'Exterior Wall,' and 'Grade' were encoded once.

**Logistic Regression Model:**

The logistic regression model offers two advantages in this case: To begin, the problem at hand is a binary classification problem with a dependent variable. Second, the model is really easy to understand. We can understand how different factors impact the outcome by studying the weight of features. Because there are extremely few observations of undervalued qualities, we split the data into a training data set and a testing data set in a 60/40% split rather than an 80/20% split to prevent overfitting concerns. Dummy variables are created using categorical variables. Finally, a logistic regression model for prediction is created.

We may take some inferences by evaluating the P value and coefficient of each variable in the Appendix. Among the variables are "Building Value", "Land Value", "Full Bath", "Land Use by Single Family", "Multiple Parcels Involved in Sale", "Property Sold as Vacant", "City GOODLETTSVILLE", "City MADISON", "City Joelton", "City OLD HICKORY", "City NASHVILLE", all Tax Districts, and "grade type E" houses. Precision is 76% and 54% in classes 0 and 1, respectively. It measures how well the model predicts good outcomes. With high accuracy, the model predicts minimal false positives. As a result, just 54% of the time our approach properly anticipates undervalued properties. Classes 0 and 1 have 100% and 1% recall rates, respectively. It evaluates the model's ability to recognize every positive case in the data. The accuracy is 76%, and the F-1 score (weighted average of precision), recall, and accuracy are 86% for class 0 and 2% for class 1, indicating that the attributes are undervalued. Many false positives occur when a large number of negative cases are incorrectly labeled as positive. This suggests that our model believes the properties are overpriced.

**Decision Tree Model:**

Precision is 76% and 69% in classes 0 and 1, respectively, according to the Decision Tree Model. Based on major features, our decision algorithm properly identifies undervalued assets 69% of the time. Classes 0 and 1 have recall rates of 100% and 3%, respectively. The precision is 76%, while the F-1 score is 86% for class 0 and 5% for class 1, suggesting that the attributes are undervalued. As a consequence, our approach marginally outperforms the logistic regression model. To evaluate the feature significance, we check the values assigned to each feature in the model. These numbers show the impurity decrease obtained by segmenting the data on that attribute, such as Gini impurity. A greater score suggests that the trait is more relevant in properly forecasting. According to the Decision Tree model in the appendix, "Sold as Vacant" is the most important feature, followed by "Building Value", "Land Value", and "Bedrooms".

**Random Forest Model:**

Precision is 76% and 91% in classes 0 and 1, respectively, according to the Random Forest Model. Based on major features, our decision algorithm properly forecasts undervalued assets 91% of the time. Classes 0 and 1 have recall rates of 100% and 2%, respectively. The precision is 76%, while the F-1 score is 86% for class 0 and 5% for class 1, suggesting that the attributes are undervalued. As a consequence, our model marginally outperforms both the logistic regression and the Decision Tree models. According to the Random Forest model, "Sold as Vacant" is the most significant feature, followed by "Building Value", "Finished Area", and "Land Value" as crucial critical criteria for defining the benchmarking of properties as undervalued or overvalued.

**Gradient Boost Model:**

Precision in classes 0 and 1 is 76% and 61%, respectively, according to the Gradient Boost Model. Based on major features, our decision algorithm properly forecasts undervalued properties 61% of the time. Classes 0 and 1 had recall rates of 99% and 6%, respectively. The precision is 76%, while the F-1 score is 86% and 11% for classes 0 and 1 respectively, suggesting that the attributes are undervalued. As a consequence, our model marginally surpasses all other models, including logistic regression, Decision Tree Model, and Random Forest Model, since the values of recall for class 0 are the greatest in our model. According to the Gradient Boost Model, "Building Value" is the most important feature, followed by "Land Value", "Finished Area", and "Acreage" as crucial critical criteria for deciding whether a property is overpriced or undervalued.

**Model Comparison:**

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

Table 1: Model Metrics Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| **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% |

**1.3 CONCLUSION AND RECOMMENDATION:**

Using the previously discussed research and models, we were able to anticipate the critical elements determining the value of a property. A high F1 score suggests that the model has enough accuracy and recall. That is, the model produces a big proportion of right positive predictions while producing a small number of wrong positive predictions. All of the models in class 0 possess a high F-1 score. showing the model can accurately forecast it. Due to imbalanced data, F-1 ratings for class "1" are immensely low. These aspects are connected because the costs of real estate are controlled by several fundamental criteria evaluated before pricing the property. Finally, we were able to derive certain important essential elements such as "Building Value", and "Land Value" using our gradient boosting approach, which is by far the most accurate.

We may acquire a complete grasp of the important aspects that impact the choice to designate properties as overpriced or undervalued using this Analysis. To assess if a home is a suitable investment, buyers must examine a number of variables. Furthermore, while contemplating an investment, important stakeholders should prioritize criteria such as "Building Value", and "Land Value" and categorize assets based on these variables.

**1.4 REFERENCES:**

1. *Hastie, Tibshirani, & Friedman. (August 26, 2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction (2nd ed.). Springer. Retrieved June 14, 2023, from* [*https://doi.org/10.1007/978-0-387-84858-7*](https://doi.org/10.1007/978-0-387-84858-7)
2. *Adrian R. (May 31, 2021). Hyperparameter tuning for Deep Learning with scikit-learn, Keras, and TensorFlow. Pyimagesearch. Retrieved June 14, 2023, from* [*https://pyimagesearch.com/2021/05/31/hyperparameter-tuning-for-deep-learning-with-scikit-learn-keras-and-tensorflow/*](https://pyimagesearch.com/2021/05/31/hyperparameter-tuning-for-deep-learning-with-scikit-learn-keras-and-tensorflow/)

**1.5 APPENDIX:**

* Figure 1: Count Plot for Target Variable “Sales Price Compared to Value”

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

* Figure 2: Histograms for other Variables:

A screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidence

|  |  |
| --- | --- |
| * Figure 3: Boxplot for Land Value:   A picture containing screenshot, text  Description automatically generated | * Figure 4: Boxplot for Building Value:   A picture containing screenshot, text, rectangle, diagram  Description automatically generated |

* Figure 5: Correlation Plot:

A screenshot of a graph

Description automatically generated with low confidence

* Logistic Regression Model Results:

Logit Regression Results

==============================================================================================

Dep. Variable: Sale\_Price\_Compared\_To\_Value\_Under No. Observations: 13521

Model: Logit Df Residuals: 13476

Method: MLE Df Model: 44

Date: Sat, 17 Jun 2023 Pseudo R-squ.: inf

Time: 19:42:35 Log-Likelihood: -inf

converged: False LL-Null: 0.0000

Covariance Type: nonrobust LLR p-value: 1.000

==========================================================================================================

coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------------------------

const 0.2346 0.413 0.569 0.570 -0.574 1.043

Acreage 0.0719 0.038 1.889 0.059 -0.003 0.146

Neighborhood -2.84e-05 1.56e-05 -1.818 0.069 -5.9e-05 2.21e-06

Land\_Value -2.493e-06 3.89e-07 -6.406 0.000 -3.26e-06 -1.73e-06

Building\_Value 1.138e-06 2.63e-07 4.335 0.000 6.24e-07 1.65e-06

Finished\_Area 1.688e-06 5.07e-05 0.033 0.973 -9.77e-05 0.000

Bedrooms -0.0033 0.034 -0.099 0.921 -0.069 0.063

Full\_Bath 0.0322 0.040 0.799 0.424 -0.047 0.111

Half\_Bath 0.0056 0.050 0.112 0.911 -0.093 0.104

Land\_Use\_SINGLE FAMILY -0.2226 0.093 -2.390 0.017 -0.405 -0.040

Sold\_As\_Vacant\_Yes 2.8807 0.331 8.714 0.000 2.233 3.529

Multiple\_Parcels\_Involved\_in\_Sale\_Yes -0.5759 0.156 -3.687 0.000 -0.882 -0.270

City\_BRENTWOOD -0.5844 0.323 -1.808 0.071 -1.218 0.049

City\_GOODLETTSVILLE 0.6335 0.341 1.859 0.063 -0.034 1.301

City\_HERMITAGE 0.3243 0.190 1.709 0.087 -0.048 0.696

City\_JOELTON 1.7963 0.726 2.476 0.013 0.374 3.218

City\_MADISON 1.2422 0.176 7.044 0.000 0.897 1.588

City\_MOUNT JULIET -0.2360 1.135 -0.208 0.835 -2.460 1.988

City\_NASHVILLE 0.3370 0.112 3.008 0.003 0.117 0.557

City\_OLD HICKORY 0.8870 0.188 4.725 0.000 0.519 1.255

City\_WHITES CREEK 1.4179 0.703 2.018 0.044 0.041 2.795

Tax\_District\_CITY OF BERRY HILL -2.1120 1.084 -1.948 0.051 -4.237 0.013

Tax\_District\_CITY OF FOREST HILLS -0.2463 0.285 -0.865 0.387 -0.804 0.312

Tax\_District\_CITY OF GOODLETTSVILLE -0.9725 0.452 -2.152 0.031 -1.858 -0.087

Tax\_District\_CITY OF OAK HILL -0.7938 0.295 -2.689 0.007 -1.372 -0.215

Tax\_District\_GENERAL SERVICES DISTRICT -1.2257 0.309 -3.961 0.000 -1.832 -0.619

Tax\_District\_URBAN SERVICES DISTRICT -1.0607 0.294 -3.613 0.000 -1.636 -0.485

Foundation\_Type\_FULL BSMT 0.1089 0.058 1.871 0.061 -0.005 0.223

Foundation\_Type\_PT BSMT 0.0624 0.063 0.989 0.323 -0.061 0.186

Foundation\_Type\_SLAB 0.0175 0.089 0.196 0.845 -0.158 0.193

Exterior\_Wall\_BRICK/FRAME -0.0609 0.078 -0.786 0.432 -0.213 0.091

Exterior\_Wall\_CONC BLK 0.1527 0.298 0.513 0.608 -0.431 0.737

Exterior\_Wall\_FRAME -0.0370 0.051 -0.730 0.465 -0.136 0.062

Exterior\_Wall\_FRAME/STONE 0.1824 0.293 0.623 0.533 -0.391 0.756

Exterior\_Wall\_LOG -44.7976 3.11e+09 -1.44e-08 1.000 -6.09e+09 6.09e+09

Exterior\_Wall\_METAL -0.2905 1.120 -0.259 0.795 -2.486 1.905

Exterior\_Wall\_STONE 0.0066 0.168 0.039 0.969 -0.322 0.336

Exterior\_Wall\_STUCCO -0.4840 0.264 -1.830 0.067 -1.002 0.034

Grade\_B -0.3919 0.143 -2.740 0.006 -0.672 -0.112

Grade\_C -0.5351 0.157 -3.409 0.001 -0.843 -0.227

Grade\_D -0.2347 0.179 -1.309 0.191 -0.586 0.117

Grade\_E 0.6959 0.377 1.847 0.065 -0.043 1.434

Grade\_OFB -19.8742 2.75e+04 -0.001 0.999 -5.4e+04 5.39e+04

Grade\_SSC 24.9039 1.84e+05 0.000 1.000 -3.6e+05 3.6e+05

Grade\_X -0.1785 0.207 -0.861 0.389 -0.585 0.228

==========================================================================================================

precision recall f1-score support

0 0.76 1.00 0.86 6844

1 0.51 0.01 0.02 2171

accuracy 0.76 9015

macro avg 0.64 0.50 0.44 9015

weighted avg 0.70 0.76 0.66 9015

* Decision Tree Model Results:

Confusion Matrix:

[[6827 17]

[2116 55]]

Classification Report:

precision recall f1-score support

0 0.76 1.00 0.86 6844

1 0.76 0.03 0.05 2171

accuracy 0.76 9015

macro avg 0.76 0.51 0.46 9015

weighted avg 0.76 0.76 0.67 9015

Top Feature Importance Variables:

Feature Importance

0 Sold\_As\_Vacant\_Yes 0.373056

1 Building\_Value 0.276894

2 Land\_Value 0.270195

3 Neighborhood 0.070855

4 Finished\_Area 0.009000

5 Acreage 0.000000

6 Bedrooms 0.000000

* Random Forest Model Results:

Confusion Matrix:

[[6796 4]

[2177 38]]

Classification Report:

precision recall f1-score support

0 0.76 1.00 0.86 6800

1 0.90 0.02 0.03 2215

accuracy 0.76 9015

macro avg 0.83 0.51 0.45 9015

weighted avg 0.79 0.76 0.66 9015

Top Feature Importance Variables:

Feature Importance

0 Sold\_As\_Vacant\_Yes 0.233952

1 Building\_Value 0.162378

2 Finished\_Area 0.112583

3 Land\_Value 0.108903

4 Full\_Bath 0.076929

5 Acreage 0.048156

6 Neighborhood 0.044184

* Gradient Boosting Model Results:

Confusion Matrix:

[[6695 105]

[2067 148]]

Classification Report:

precision recall f1-score support

0 0.76 0.98 0.86 6800

1 0.58 0.07 0.12 2215

accuracy 0.76 9015

macro avg 0.67 0.53 0.49 9015

weighted avg 0.72 0.76 0.68 9015

Top Feature Importance Variables:

Feature Importance

0 Building\_Value 2.603118e-01

1 Land\_Value 1.985981e-01

2 Sold\_As\_Vacant\_Yes 1.107184e-01

3 Finished\_Area 1.088758e-01

4 Acreage 8.960166e-02

5 Neighborhood 7.726702e-02

6 Full\_Bath 2.414811e-02

* Two files have been attached along with this assignment report for reference viz. dataset file named **“Nashville\_housing\_data.csv”** and Python Jupyter Notebook code file named **“ALY6020\_M4Project\_Agarwal.ipynb”**.