Northeastern University
College of Professional Studies

ALY 6020 – PAREDICTIVE ANALYTICS

Prof. Justin Grosz

Module 4 – Investing in Nashville

03/19/2023

Pooja Kairamkonda

INTRODUCTION

Investing in Nashville housing can be a lucrative opportunity for those interested in real estate. Nashville is a growing city with a strong economy and a thriving cultural scene, which has led to an increase in demand for housing. This demand has driven up property values and rental rates, making the city an attractive destination for real estate investors. When considering investing in Nashville housing, it is important to research and consider the factors such as exact location, tax, amenities, property type, land value, building value, finished area & foundation type. The project aims to analyze & examine these factors in order to decide whether or not investing in Nashville would be wise.

In this report, the more categorized Nashville housing dataset features were used to investigate the variables that could significantly affect the price of real estate. The results were analyzed to identify the trends & patterns of individual attributes including Land value, Building value, Finished area, Foundation type, Number of bedrooms, exterior wall material and many more relative features how they can be used to decide investing in Nashville property would be good or not. The goal of this project is to understand & predict the sale price compared to value is over or under. A logistic regression, decision tree, random forest & gradient boost model was used to accurately predict the outcome and determine the key factors in finding the best deal . The model results were compared and to provide recommendations to the real estate company to focus on the right features. By analyzing the dataset, the company can identify which factors contribute to identify the key factors to workout the best property purchasing deal.

DATA CLEANING

The dataset ‘Nashville_housing_data.csv’ includes 22651 records and 26 variables. At a first glance, the dataset was in a good shape as the datatype of each variable was accurate. With the help of null value check, for the variables having null values were less than 0.5% of the entire dataset, the records were dropped. Also, the column ‘Suite/Condo #’ was dropped since the entire column empty. Additionally, the columns, 'Parcel ID', 'City', 'Unnamed: 0' were dropped as Parcel ID & Unnamed: 0 doesn’t have any significant contribution to the analysis, & they do not provide any meaningful information about variables being studied. Moreover to avoid bias these columns were dropped. The column ‘City’ was dropped since it was duplicate of column ‘Property City’. After performing these steps, the final dataset contained 22536 rows records and 22 columns.The final dataset was used for exploratory data analysis in the next part.

EXPLORATORY DATA ANALYSIS

In this section, exploration of variables that may help understand the factors that could help identify the best property value were analyzed using a variety of graphs & plots that could possibly help the real estate company to make the decision of investing in Nashville properties through the target variable ‘Sale Price Compared to Value’.

A subset of numeric columns that can effectively have an impact on the deal, was used to observe the outliers in the dataset as shown in Appendix 1. Additionally, histograms were plotted to understand the data distribution. Here we have observed that the variables, Acreage, Land value, Building value, Finished Area, Bedrooms, Full Bath were found to be right skewed.

In order to understand how variables of the dataset are significantly correlated a correlation plot was generated. Here, we have used one hot encoding to convert the data into numerical by creating the dummy variables for each of the unique column value. Numerical information should be provided so that the results can be properly interpreted. One hot encoding also has the benefit of avoiding the issue of introducing hierarchy or order among categories. While performing one hot encoding we dropped the original variables for modelling.

Upon observing the correlation matrix, we can clearly see that few of the attributes that that have high collinearity were building value, finished area, full bath, Property City_GOODLETTSVILLE. Since these variables are highly correlated with each other, we performed VIF (Verification Infation Factor) check to identify & remove them in order to avoid model issues like unstable and unreliable coefficient estimates and ensure the model's accuracy and reliability. VIF measure is useful to check the degree of multicollinearity of predictor variables. We observed, the variables TAX DISTRICT_GENERAL SERVICES DISTRICT, TAX DISTRICT_URBAN SERVICES DISTRICT & Grade_C were found to VIF>10 hence we dropped them. Also, the variables Foundation Type_TYPICAL, Grade_OFB, Grade SSC with VIF value as infinite were dropped to avoid issues with accuracy and model reliability. Dropping these columns would improve the accuracy and stability of the model by reducing multicollinearity and allowing the remaining predictor variables to more accurately represent their unique contributions to the outcome variable. After performing the initial analysis with EDA we dived deeper into further analysis.

ANALYSIS

In this section, we dive deeper into our analysis. Here our target variable is ‘Sale Price Compared to Value’ which is valued as Over/Under.

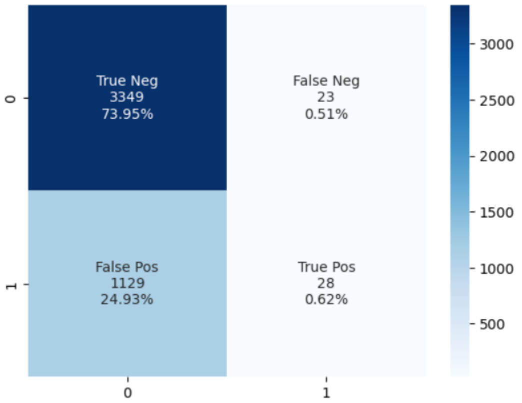
In the section, we are focusing on building 4 logistic models with all the independent variables and one dependent variable ‘Sale Price Compared to Value’ to make predictions about the best deal of the property price and compare the model results with the previous once. Among the 4 models, we will be providing model recommendations to the company that results more accurate outcome and can make investment in a property a wise decision.

The dataset was split into 80%-20% train-test split. Initially, a logistic regression model was fit on the dataset and the Logit summary of regression analysis was obtained to observe & interpret the results as shown in Appendix 1. Upon observing the summary results of regression analysis, it was clear that, variables such as Land value, Building Value, Sale Month, Property City_Madison, Property City_Nashville, Property City_old hickory, sold as vacant_yes, Grade_D & Grade_E are significant predictors of the outcome that will provide us with results if the property sale value is over/under priced. A considerable number of variables that are not significant (variables with pvalue>0.05) are listed in the summary results as shown in Appendix 1.

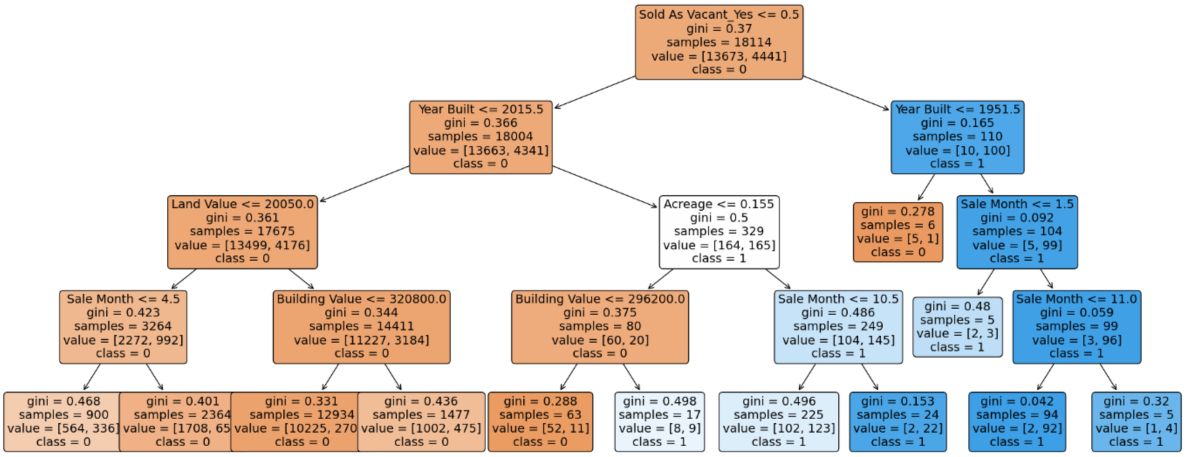
We move ahead to build our prediction models logistic, decision tree, random forest, and gradient boost compare the results to decide which one would yield better results.

Throughout our analysis, we are comparing our models based on Accuracy benchmark.

- **Logistic regression**: In order to interpret the results of logistic regression model, we generated a confusion matrix as shown. The x-axis indicates actual values & y-axis indicates predicted values. From the figure we can observe that, overall 74.56% predictions made by the model were accurate. On the other hand, a loss factor of 1129 (24.93%), since it shows misclassification of property values true that were actually false. This component needs to be optimized by the company because it can lead to misunderstandings and ultimately inappropriate investments that will eventually cause loss.



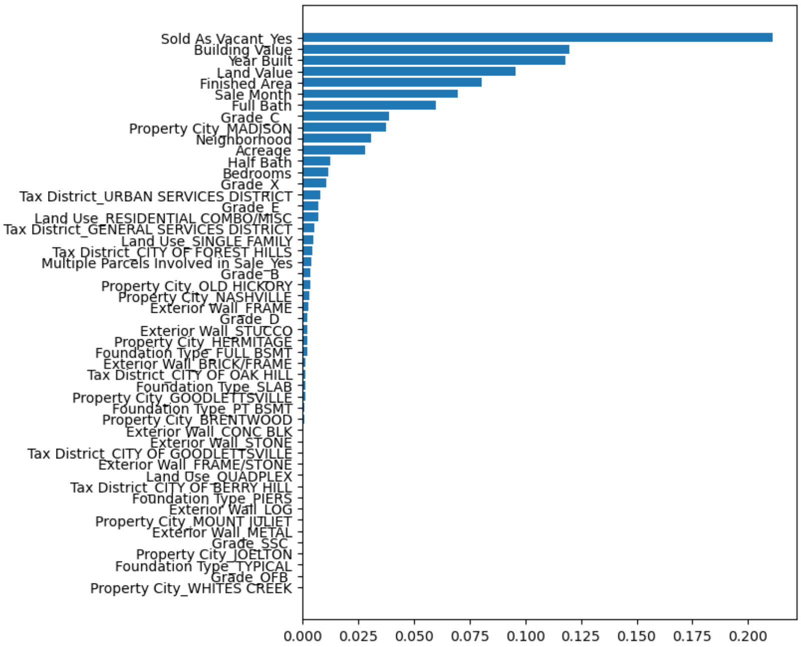
- **Decision Tree Model** : A Decision tree model was built, in order to understand, compare, interpret and check if the model accuracy can be improved to better predict the outcome. Decision trees are better at handling categorical & numerical variables of a dataset and thus we used encoded dataset except the target variable which is originally a categorical variable. The train-test split was kept as 80-20. A decision model with max_depth = 4 was fit to the dataset, where 4 indicates the tree level. This model is more interpretable as compared to logistic regression model as it builds a flowchart that displays how the model is making predictions, what variables were used at each level, number of records, and the gini index whose low value indicates less likelihood of misidentification. The decision tree is as follows:



Here the leaf nodes represent the prediction outcome. In this case, the variables Sale month, Building value, Land value and acreage plays an important role in identifying the over/under priced properties. The model is capable of achieving 75.31% of accurate results, which is slightly higher than the linear regression model. The lower gini index of variable ‘Sale month’ can mostly contribute to the prediction of target variable. Furthermore, feature importance technique was used to identify the variables that can significantly impact the predictor which is displayed in the decision tree feature importance plot displayed as shown in Appendix 1.

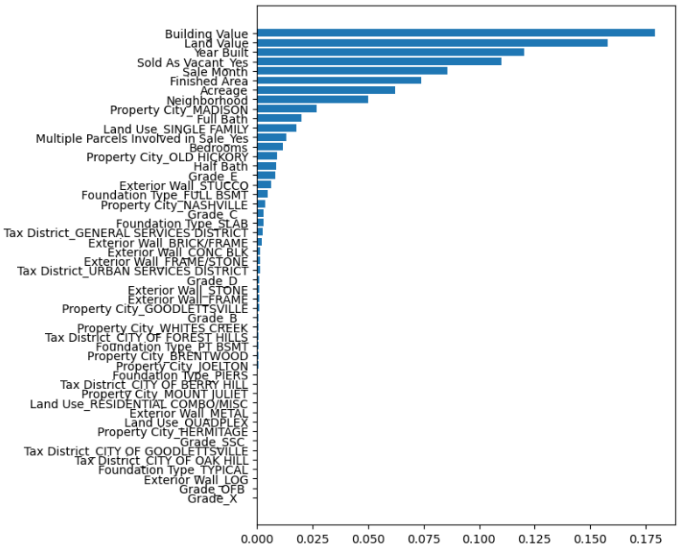
- **Random Forest Model :**

Next, a random forest model was built as they help achieve higher accuracy than decision tree model due to the ensemble of multiple decision trees that makes it less susceptible to overfitting caused by outliers or noise in the data. However, in this case, the model accuracy is 74.89% which is slightly less than Random forest model. This is because the feature importance selection that used more number of variables than decision tree. These variables can be observed from the Random Forest feature importance plot as shown below.



- **Gradient Boost Model :**

This is another model that used ensembled decision trees to improve the model prediction. The ensembled decision trees are built sequentially, where each subsequent tree learns from the errors of the previous tree. The model gradually improves its predictions by iteratively minimizing the loss function. A model accuracy of 75.51% was obtained by this model which is highest among all the 4 models. With the help of feature importance, the model highlights the below variables that potentially help predicting the over/under priced properties.



In order to understand the accurate & inaccurate predictions of all the models a detailed classification report was generated as below:

Model	Logistic		Decision Tree		Random Forest		Gradient Boost	
Accuracy	74.56%		75.31%		74.89%		75.51%	
	0	1	0	1	0	1	0	1
Precision	0.75	0.55	0.75	0.72	0.75	0.86	0.76	0.67
Recall	0.99	0.02	0.99	0.05	1.00	0.02	0.99	0.08
F-1 Score	0.85	0.05	0.86	0.10	0.86	0.04	0.86	0.14

The table help us analyze the figures as below:

- Precision: from all the classes we have predicted as positive, how many are actually positive.
- Recall: from all the positive classes, how many we predicted correctly.
- F-measure: It is used to compare scenarios where occurrence of low precision and high recall and vice versa is observed.

In next part, we will drilling down the variables that are important to business and provide recommendations.

CONCLUSION & RECOMMENDATIONS

Based on the analysis results & feature importance technique of different models help us to focus on the variable that are important to improve and generate better prediction results and few of them are: Land value, building value, sold as vacant_yes, Acreage, finished area.

Additionally, from the accuracy results it is evident that, the gradient boost model will predict the over/under priced property values with 75.51% accuracy. Along with this model we provide additional recommendations to the company to focus on the variables that could impact the outcome and help them to make the decision of investing in Nashville's properties.

The company should think about making an investment in more precise valuation techniques or technologies in order to increase the accuracy of the land value and building value factors. This can entail utilizing more sophisticated equipment or engaging a qualified appraiser to determine property values.

The company should concentrate on locating and focusing on properties that are likely to be sold as unoccupied in order to make a wise investment as it will have an impact on the decision of investing. This could entail researching market trends, finding homes that are regularly listed as unoccupied, and creating marketing plans that are especially aimed at these kinds of homes.

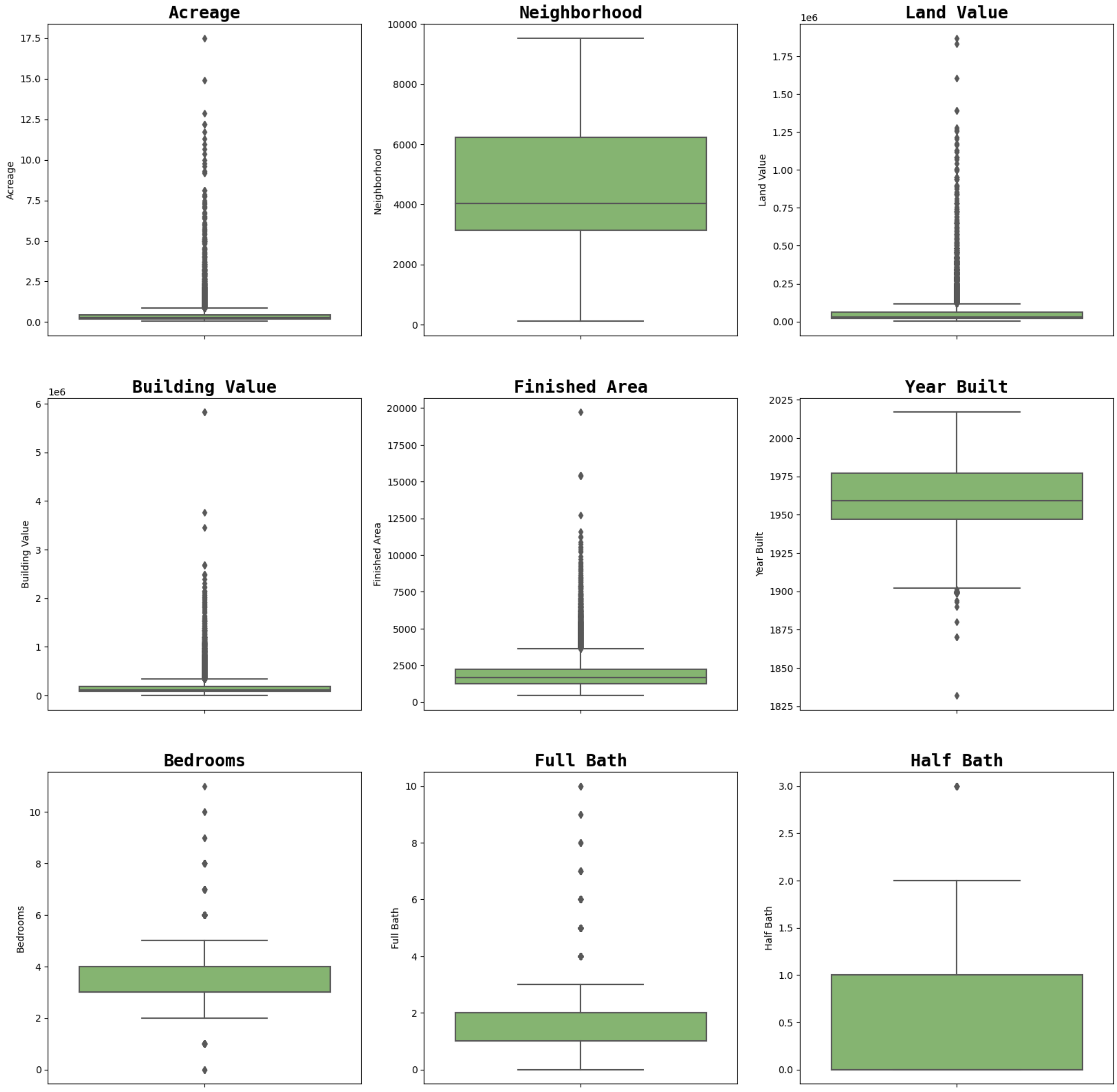
he acreage of a property is an important factor to consider when making a property investment decision. The size of the property can determine its potential use, such as whether it can be used for commercial or agricultural purposes, and may provide more opportunities for development or income-generating activities. The acreage can also affect the property's value, with larger properties generally commanding higher prices. Additionally, the potential returns on a property investment can be influenced by the acreage, as a larger property may provide more rental income or appreciate in value more quickly. For some investors, the size of the property is important for lifestyle reasons, such as having space for outdoor activities or hobbies, and providing more privacy and seclusion. Ultimately, the acreage of a property is just one of many factors that should be considered when making a property investment decision, alongside location, condition, and potential for growth. Finally, with our analysis we can recommend using gradient boost model as it has the highest more accuracy than other 3 regression models that will help the real estate company to generate more accurate predictions and ultimately make better investment decisions.

REFERENCES

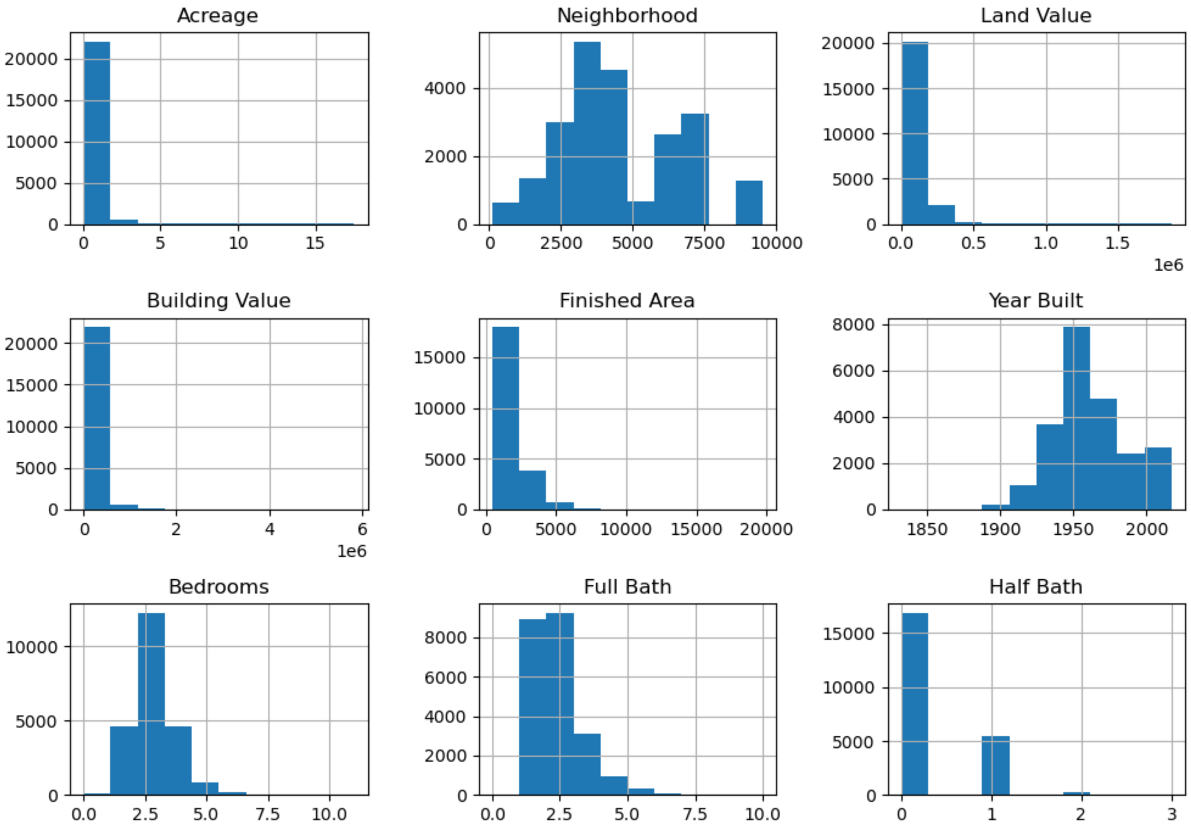
1. Sarang Narkhede (May 9, 2018). Understanding Confusion Matrix. Retrieved from - https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
2. Dennis T. (Jul 25, 2019). Confusion Matrix Visualization. Retrieved from - https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea
3. Valentin Richer. (Mar 2, 2019). Understanding Decision Trees (once and for all!). Retrieved from - https://towardsdatascience.com/understanding-decision-trees-once-and-for-all-2d891b1be579
4. Mehul Gupta. (Jul 17, 2019). Ensemble Models in Machine Learning. Retrieved from -https://medium.com/data-science-in-your-pocket/ensemble-models-in-machine-learning-d429c988e866

APPENDIX 1

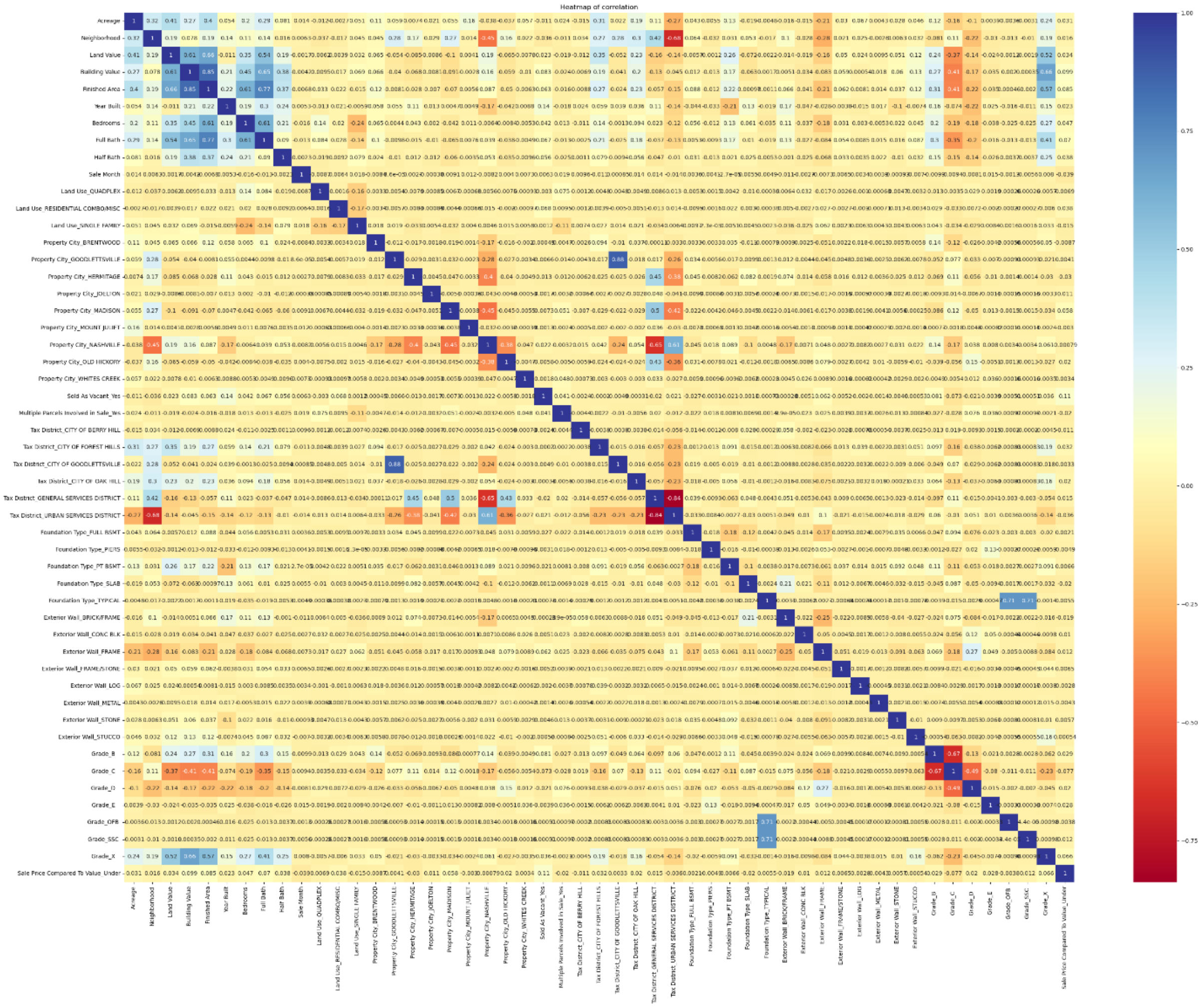
1. Detecting outliers of numerical variables



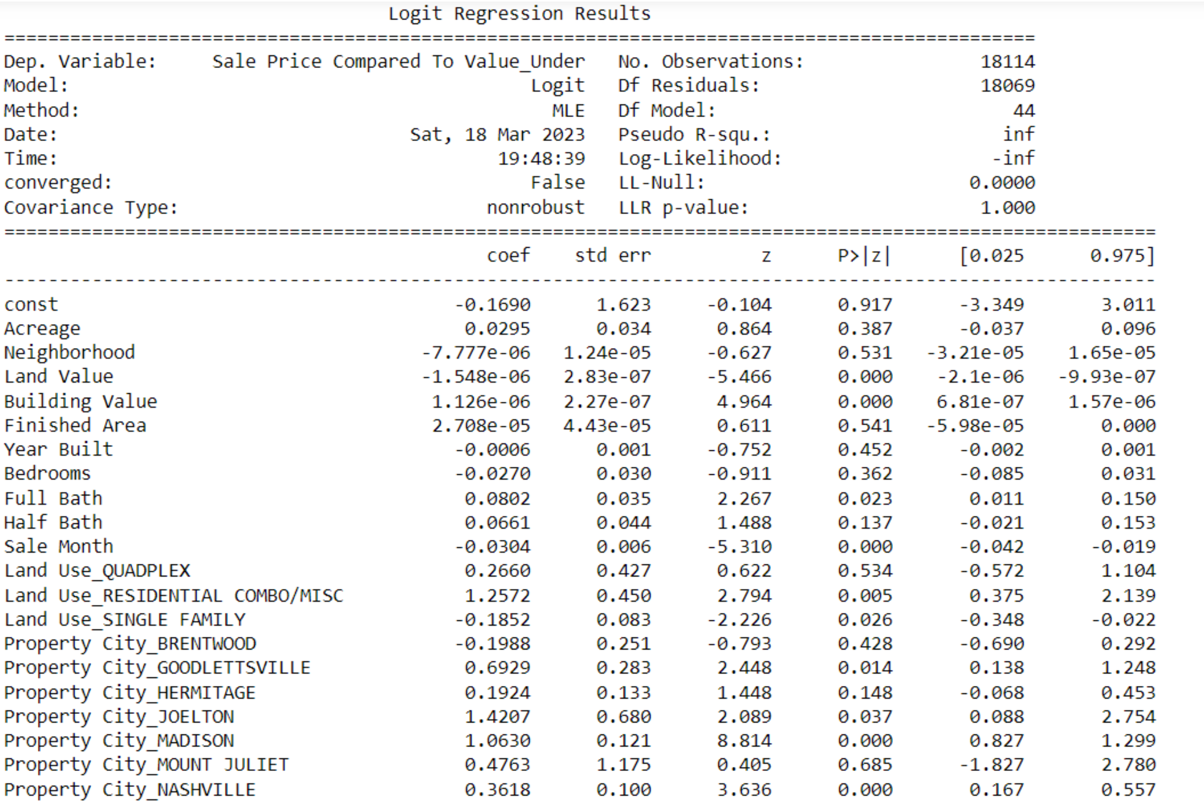
2. Data distribution of numerical variables



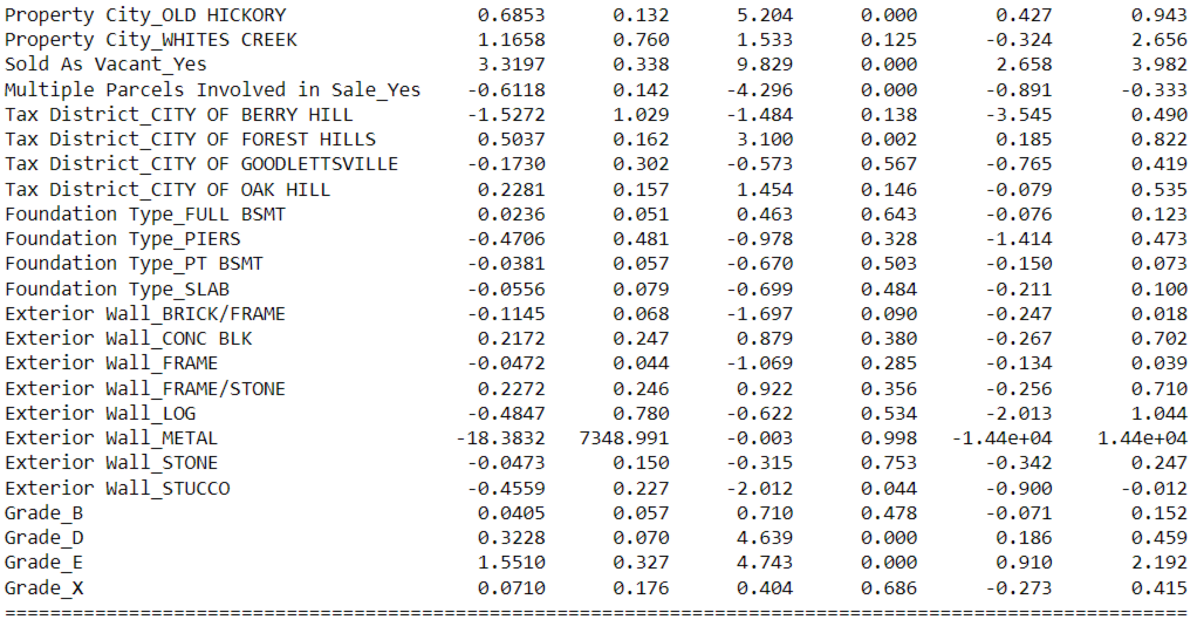
3. Heatmap of Correlation



4. Logit regression results

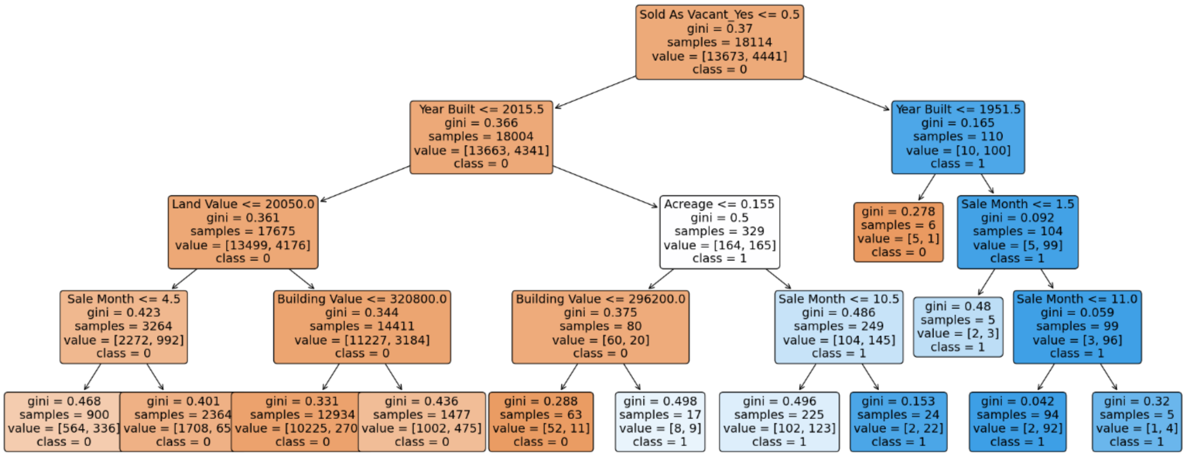


```
                           Logit Regression Results
==============================================================================
Dep. Variable:     Sale Price Compared To Value_Under   No. Observations:           18114
Model:                                   Logit   Df Residuals:               18069
Method:                                    MLE   Df Model:                      44
Date:                         Sat, 18 Mar 2023   Pseudo R-squ.:                inf
Time:                                 19:48:39   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                           -0.1690      1.623     -0.104      0.917      -3.349       3.011
Acreage                          0.0295      0.034      0.864      0.387      -0.037       0.096
Neighborhood                 -7.777e-06   1.24e-05     -0.627      0.531   -3.21e-05    1.65e-05
Land Value                   -1.548e-06   2.83e-07     -5.466      0.000    -2.1e-06   -9.93e-07
Building Value                1.126e-06   2.27e-07      4.964      0.000    6.81e-07    1.57e-06
Finished Area                 2.708e-05   4.43e-05      0.611      0.541   -5.98e-05       0.000
Year Built                      -0.0006      0.001     -0.752      0.452      -0.002       0.001
Bedrooms                        -0.0270      0.030     -0.911      0.362      -0.085       0.031
Full Bath                        0.0802      0.035      2.267      0.023       0.011       0.150
Half Bath                        0.0661      0.044      1.488      0.137      -0.021       0.153
Sale Month                      -0.0304      0.006     -5.310      0.000      -0.042      -0.019
Land Use_QUADPLEX                0.2660      0.427      0.622      0.534      -0.572       1.104
Land Use_RESIDENTIAL COMBO/MISC  1.2572      0.450      2.794      0.005       0.375       2.139
Land Use_SINGLE FAMILY          -0.1852      0.083     -2.226      0.026      -0.348      -0.022
Property City_BRENTWOOD         -0.1988      0.251     -0.793      0.428      -0.690       0.292
Property City_GOODLETTSVILLE     0.6929      0.283      2.448      0.014       0.138       1.248
Property City_HERMITAGE          0.1924      0.133      1.448      0.148      -0.068       0.453
Property City_JOELTON            1.4207      0.680      2.089      0.037       0.088       2.754
Property City_MADISON            1.0630      0.121      8.814      0.000       0.827       1.299
Property City_MOUNT JULIET       0.4763      1.175      0.405      0.685      -1.827       2.780
Property City_NASHVILLE          0.3618      0.100      3.636      0.000       0.167       0.557
```

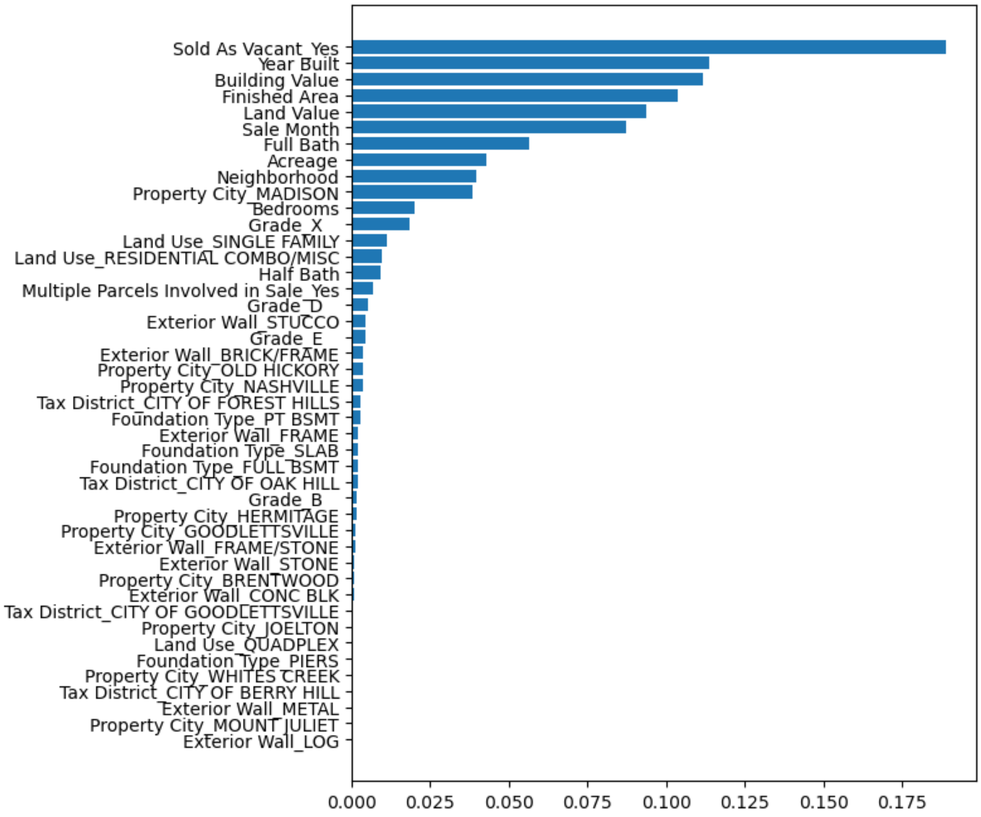


```
Property City_OLD HICKORY              0.6853      0.132      5.204      0.000       0.427       0.943
Property City_WHITES CREEK             1.1658      0.760      1.533      0.125      -0.324       2.656
Sold As Vacant_Yes                     3.3197      0.338      9.829      0.000       2.658       3.982
Multiple Parcels Involved in Sale_Yes -0.6118      0.142     -4.296      0.000      -0.891      -0.333
Tax District_CITY OF BERRY HILL       -1.5272      1.029     -1.484      0.138      -3.545       0.490
Tax District_CITY OF FOREST HILLS      0.5037      0.162      3.100      0.002       0.185       0.822
Tax District_CITY OF GOODLETTSVILLE   -0.1730      0.302     -0.573      0.567      -0.765       0.419
Tax District_CITY OF OAK HILL          0.2281      0.157      1.454      0.146      -0.079       0.535
Foundation Type_FULL BSMT              0.0236      0.051      0.463      0.643      -0.076       0.123
Foundation Type_PIERS                 -0.4706      0.481     -0.978      0.328      -1.414       0.473
Foundation Type_PT BSMT               -0.0381      0.057     -0.670      0.503      -0.150       0.073
Foundation Type_SLAB                  -0.0556      0.079     -0.699      0.484      -0.211       0.100
Exterior Wall_BRICK/FRAME             -0.1145      0.068     -1.697      0.090      -0.247       0.018
Exterior Wall_CONC BLK                 0.2172      0.247      0.879      0.380      -0.267       0.702
Exterior Wall_FRAME                   -0.0472      0.044     -1.069      0.285      -0.134       0.039
Exterior Wall_FRAME/STONE              0.2272      0.246      0.922      0.356      -0.256       0.710
Exterior Wall_LOG                     -0.4847      0.780     -0.622      0.534      -2.013       1.044
Exterior Wall_METAL                  -18.3832   7348.991     -0.003      0.998   -1.44e+04    1.44e+04
Exterior Wall_STONE                   -0.0473      0.150     -0.315      0.753      -0.342       0.247
Exterior Wall_STUCCO                  -0.4559      0.227     -2.012      0.044      -0.900      -0.012
Grade_B                                0.0405      0.057      0.710      0.478      -0.071       0.152
Grade_D                                0.3228      0.070      4.639      0.000       0.186       0.459
Grade_E                                1.5510      0.327      4.743      0.000       0.910       2.192
Grade_X                                0.0710      0.176      0.404      0.686      -0.273       0.415
=========================================================================================
```

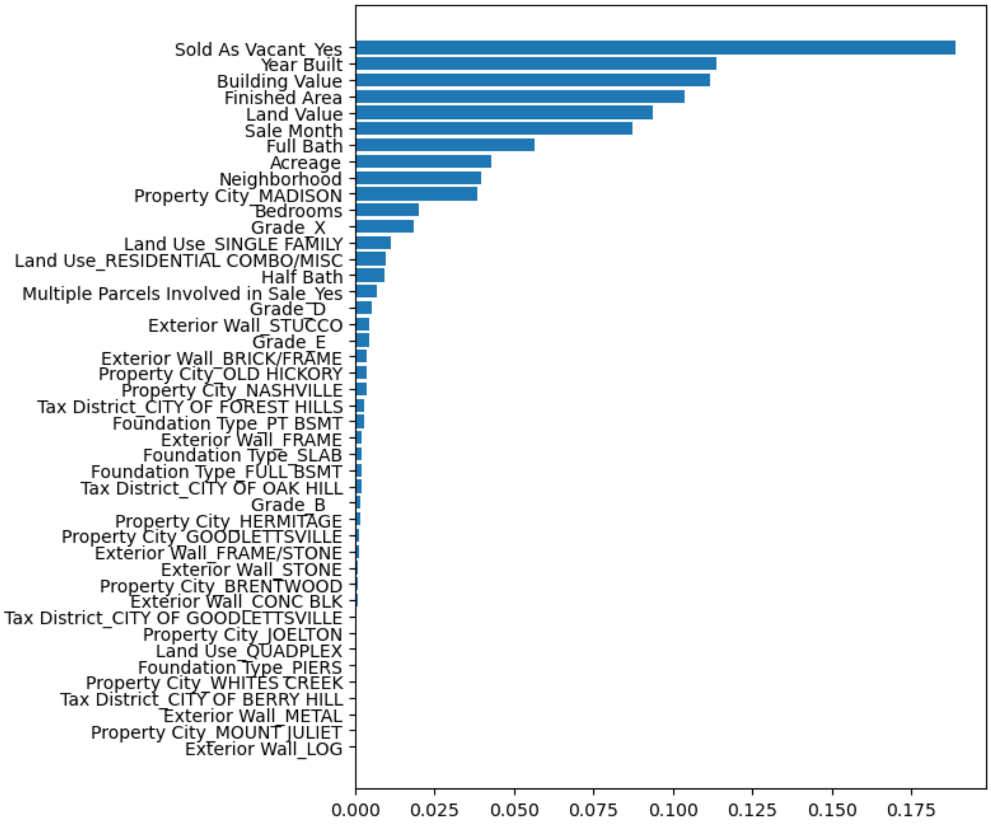
5. Decision Tree



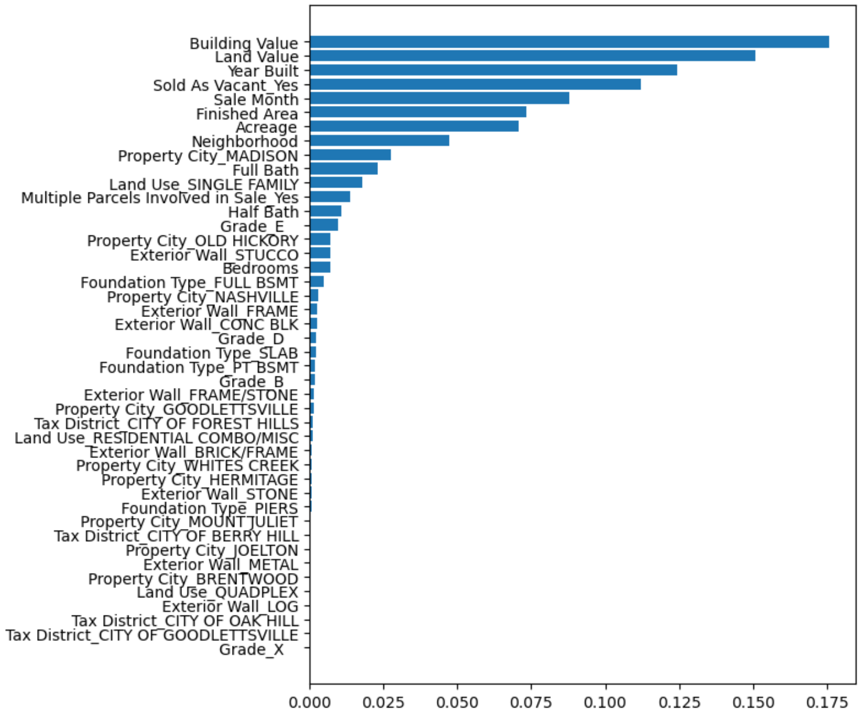
6. Decision Tree – Feature importance



7. Random Forest – Feature Importance



8. Gradient boosting – Feature Importance


Building Value
Land Value
Year Built
Sold As Vacant_Yes
Sale Month
Finished Area
Acreage
Neighborhood
Property City_MADISON
Full Bath
Land Use_SINGLE FAMILY
Multiple Parcels Involved in Sale_Yes
Half Bath
Grade_E
Property City_OLD HICKORY
Exterior Wall_STUCCO
Bedrooms
Foundation Type_FULL BSMT
Property City_NASHVILLE
Exterior Wall_FRAME
Exterior Wall_CONC BLK
Grade_D
Foundation Type_SLAB
Foundation Type_PT BSMT
Grade_B
Exterior Wall_FRAME/STONE
Property City_GOODLETTSVILLE
Tax District_CITY OF FOREST HILLS
Land Use_RESIDENTIAL COMBO/MISC
Exterior Wall_BRICK/FRAME
Property City_WHITES CREEK
Property City_HERMITAGE
Exterior Wall_STONE
Foundation Type_PIERS
Property City_MOUNT JULIET
Tax District_CITY OF BERRY HILL
Property City_JOELTON
Exterior Wall_METAL
Property City_BRENTWOOD
Land Use_QUADPLEX
Exterior Wall_LOG
Tax District_CITY OF OAK HILL
Tax District_CITY OF GOODLETTSVILLE
Grade_X