Northeastern University

ALY6020: Predictive Analytics

Assignment 4: Investing in Nashville

Abhigna Ramamurthy, 002982276

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Feb 5, 2023

Introduction

This report helps in understanding the general business due diligence that companies make to understand the existing market and make best value deals. The scenario involves a real estate company looking to make huge investment in the swiftly growing Nashville region. The report is aimed to understand what is driving the prices of houses in the Nashville region. The recent sales dataset that is provided has sales of houses along with some properties of houses that can be used to determine if the house was overpriced or underpriced during the sale. Evaluating the pricing of houses can help with investments because it affects the return on investment. If a house is overpriced, it may not yield a high enough return to make it a worthwhile investment. On the other hand, if a house is underpriced, it may provide a better return and thus be a more attractive investment opportunity. In addition, knowing the pricing trends in a particular area can help investors make informed decisions about when to buy or sell property.

Data cleaning

For this analysis, the data was presented in an excel format which was imported into Jupyter Notebook python where the analysis was conducted. Main libraries used include pandas, NumPy, and datetime for data cleaning and manipulation, seaborn, and plotly, for visualization, and statsmodels, and sklearn for data modelling. The initial data had 22,651 records and 26 attributes with 5 integer columns, 6 floating value columns, and rest are categorical or object data type columns.

The first step in cleaning this data was to delete unnecessary columns. ‘Unnamed: 0’ attribute was used to first check if we had any duplicate records as it serves are the primary identifier for the different records. As we had no duplicates ‘Unnamed: 0’ column was deleted as it did not hold any analytical value. Likewise, Parcel ID, and Legal Reference also were

eliminated due to the same reason. Property Address would be irrelevant to the analysis as it individually identifies each record too which makes it harder to classify or analyze that data. Suite / Condo is empty hence need to be eliminated. State value are all TN, hence can be deleted. Property City, City, and Neighborhood are excluded as Tax District has more aggregated data along with completeness.

The next stage of data cleaning is to check for missing values and there are 5 attributes with less than 1% missing values. All the attributes with missing values were checking for the distribution of data in the dataset to make an informed decision. With right skewed distributions, the attributes Finished Area, Bedrooms, Full Bath, and Half Bath missing values were filled with median value of the distribution. Foundation type, being a categorical attribute, the best way to fill the small number of missing values was to use mode of the column which was 'CRAWL' value. The final size of the dataset is 22,651 records (same as before as no records were deleted), and 18 attributes. The outliers were also checked using box plots. The outliers seen were just deviation from the normal values and nothing to eliminate at this stage.

Feature engineering was the next step in data cleaning where some attributes were transformed to hold more analytical contribution. Sale Date attribute holds the date of the last sale on the house which was converted to Sale Age value. This new attribute analyzes the market trend that might affect the next sale of the property as it would indicate the current value.

Similarly, Property Age was calculated using the Year Built attribute. Knowing the age of the property can be more useful in many ways as compared to just the year built.

Exploratory Data Analysis

Exploratory data analysis is a visual way to uncovering hidden patterns that can help the investors make better decisions knowing the pattern of the buyers. Zip code neighborhoods between 6200-6299 have the highest number of houses overpriced sale compared to value as shown in Figure 1, Appendix 1 histogram. More than 70% of the houses are residential with a medium quality grade of C, which indicates that a smaller number of high-quality properties. Majority of the houses are for single family use and are in urban services district areas for tax purposes with at least 51-75 years of built age as shown in bar chart of Figure 2, Appendix 1. The houses also had their last sale event 6.5-10.5 years ago which indicates the sale age ranges.

The bar graphs in Figure 3, Appendix 1 shows bar graphs of consumer preferences along with available properties for sale in the region with 3 bedrooms with 1 or 2 full bathrooms and no half bathrooms. Finally, the heatmap in Figure 4, Appendix 1 shows the correlation matrix for the dataset. It is a statistical tool that measures the linear relationship between the numerical attributes of the dataset. I'm most interested in identifying the relationships of attributes with 'Sale Price Compared to Value' which is the dependent variable that we classify on in the model analysis section. I can see sale age (which is how long it has been from the latest house sale) has the highest negative correlation with the attribute indicating latest the property sale, higher are the chances of overpriced sale. Property age has the highest positive correlation with the sale attribute indicating that newer properties are overpriced as compared to older properties.

Analysis

Data Pre-processing

Upon data cleaning and exploratory data analysis, I did several steps to get the data ready for modelling. The first order of business was to understand what the attribute for prediction was using models. The attribute 'Sale Price Compared to Value' is the dependent variable for the analysis. Hence this was converted to Boolean value as, 1 for overpriced and 0 for underpriced. Similarly attributes with just two responses and categorical were converted to the Boolean values. This step included attributes 'Sold As Vacant', and 'Multiple Parcels Involved in Sale' as 1 for yes and 0 for no. There were other attributes that could only help with exploratory data analysis and could not contribute to actual modeling of data. These attributes included 'Age Group', 'Tax District', 'Foundation Type', 'Exterior Wall', and 'Grade'.

'Land Use' attribute is an important one which identifies the type of the house, such as single family, residential combo, duplex, or quadplex. Hence this was converted to generate dummy variables using one hot encoding. The next check was for multicollinearity where correlation between the predictor variables is quantified using variance inflation factor and helps get rid of unnecessary attributes to improve the model performance. At this stage all values were justified hence no elimination were made. The dataset was further divided into training and testing data as an 80-20 split for model training and testing respectively. The training data was further scaled using standard scaler to ensure all the independent attributes are in the same range.

Data Modeling

Logistic Regression: The aim of this analysis is to understand the properties contributing to over pricing or underpricing of the sale of the house. Knowing this will help the company investors to make informed decisions on better investing. The logistic regression model output displays significant variables which can be determined by the P(>|t|), p-value, and depicts the positive and negative impact over dependent variable price through the coefficients. The Significance of the variable is determined if the P-Value is less than the significance level (0.05 A significance level of 0.05 indicates a 5% risk) then it terms that the model fits the data well. The results of the regression model are shown in Figure 1 of Appendix 2. The two most significant attributes obtained in this analysis are 'Sold as Vacant' and 'Sale Age' with coefficients of -7.91 and -0.58 respectively. The negative coefficient indicates the reciprocal relationship with sale price compared to value.

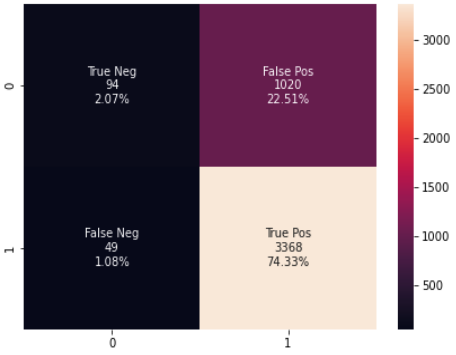

Figure 1: Confusion Matrix for Logistic Reg

The confusion matrix shown here represents the summary of performance for the model. As seen in the Confusion Matrix result, our Model has an accuracy of 76.4%. The model also yielded 76.8% precision which means the model makes a correct prediction 76.8% of the times and recall of the model is 98.6% which means if the response is 1 then the model can predict it 98.6% of the times. The F1-score is 0.69 which indicates good model performance. AUC, area under the curve for logistic regression is 0.54 which indicates the model has an average performance. The mean squared error value is 0.236 which tells how close a regression line is to a set of points. This can be used in comparison of the models.

Decision Tree: To get a better outlook on the price influencing attributes, decision tree is the second choice of model. Decision tree helps in dividing the feature space into half and helps obtaining a non-linear decision boundary. The hyperparameters used helped get a max depth of 4 and minimum samples per leaf of 5. The model also outputs visual of the model result in a tree plot as shown in Figure 2, Appendix 2. As seen in the figure, 'Sale Age' is the root node of the model which implies that it is the best predictor variable. The visualization also shows that the 'Sale Age' is class=0, which is under priced house sale. This significant variable is also in line with what logistic regression showed. The feature importance in Figure 3, Appendix 2 also shows Property Age as the second most significant predictor. Both of these values do contribute to under or over evaluation of the property.

The model achieved accuracy of 77.8% which is slightly better than that of logistic regression. The model also yielded 78% precision which means the model makes a correct prediction 78% of the times and recall of the model is 98.1% which means if the response is 1 then the model can predict it 98.1% of the times. The precision and recall values are very similar to that of logistic regression. The F1-score is 0.72, slightly better than logistic regression, which indicates good model performance. AUC, area under the curve for logistic regression is 0.57 which indicates the model has a good performance. The mean squared error value is 0.222 which is lower than logistic regression making decision tree a step-up form logistic regression. The confusion matric of the decision tree is shown in Figure 4, Appendix 2 which confirms the above listed results.

Random Forest: The Random Forest algorithm trains multiple decision trees using different samples of the training data, and different features are used as the root node in each tree. The final prediction is based on the collective predictions of all the decision trees in the forest. This ensemble model could provide better understanding of the decision tree itself as well as outputs a feature importance chart. The number of estimators is set to 5 for this analysis. The model feature importance is shown in Figure 5, Appendix 2. The top two predictors of the model are 'Building value' and 'Finished Area'. These are top predictors that are different from the ones in previous models.

Evaluating the performance of the model using confusion matrix shown in Figure 4, Appendix 2, accuracy of the mode is 77.8% which is exactly same as decision tree model. The model also yielded 78% precision which means the model makes a correct prediction 78% of the times and recall of the model is 98.2% which means if the response is 1 then the model can predict it 98.2% of the times. The precision and recall values are same as that of decision tree. The F1-score is 0.72, slightly better than logistic regression, which indicates good model performance. AUC, area under the curve for logistic regression is 0.62 which indicates the model has a good performance. The mean squared error value is 0.222 which is lower than logistic regression and same as decision tree. Hence, we can infer that random forest model has very similar results to decision tree and did not give the performance boost that was hoped for.

Gradient Boosting: Attributed as a self-correcting model, gradient boosting combines multiple low performing models by correcting the errors to produce a single strong model with improved performance. The process is repeated iteratively, where each new model focuses on the errors made by the previous models in the ensemble. With a 100 n_estimators, the model is restricted with maximum tree depth of 4 and learning rate of 0.1. The model also acts like a black box as it is not easy to extract feature importance from gradient boosting. However, its performance can be used to compare with other models, and it is also deemed as one of the most popular machine learning algorithms as it can achieve high accuracy on a wide range of problems.

The model achieved accuracy of 77.8%. The model also yielded 78% precision which means the model makes a correct prediction 78% of the times and recall of the model is 98.1% which means if the response is 1 then the model can predict it 98.1% of the times. The F1-score is 0.72, which indicates good model performance. AUC, area under the curve for logistic regression is 0.60 which indicates the model has a good performance. The mean squared error value is 0.222 which is lower than logistic regression but exactly same as decision tree model. The confusion matric of the decision tree is shown in Figure 7, Appendix 2 which confirms the above listed results.

Model comparison using benchmarking metrics

Model comparison is done using benchmarking metrics, which are quantitative measures to evaluate the performance of the models. The benchmarking metrics chosen for evaluation in this report are accuracy, precision, recall, mean squared error, F1 score, false positive rate, and speed of model execution. Accuracy helps understand the percentage pf correct predictions made by the model, however, on its own it is just one side of the coin. Precision, recall, and F1 score provide a better insight into the model performance. MSE is for comparison of models, along with AUC and speed of execution to evaluate the heaviness of model calculations.

Model	***Accuracy***	***Precision***	***Recall***	***MSE***	***F1 score***	***FPR***	***AUC***	***Speed***
Logistic Regression	76.4%	76.8%	98.6%	0.236	0.688	22.51%	0.54	0:00:00.083410
Decision Tree	77.8%	78%	98.1%	0.222	0.718	20.83%	0.57	0:08:14.752729
Random Forest	77.8%	78%	98.2%	0.222	0.717	20.86%	0.62	0:05:35.567447
Gradient Boosting	77.8%	78%	98.1%	0.222	0.718	20.83%	0.60	21:09:19.121860

Table 1: Benchmarking metrics used in the models

The above table shows different model metrics that are used to compare the models. Given that the main aim of the analysis is to understand what attributes affect the pricing of the houses in Nashville area, it is important to have a balance between all the different metrics to arrive at

the optimal model predictors. Due to slightly low performance logistic regression will be ruled out. However, other three tree-based models show comparably similar performances providing evidence that the data works well for tree-based classification algorithms. Among the three models, I selected decision tree model as the final model to be used as it has the perfect balance of speed, and performance. Thus, the attributes of importance from the decision tree model are sale age, property age, and land value.

Conclusion

From the above analysis we can conclude that higher performance is obtained from the decision tree model. The age attributes of house sale age and property age along with land value are the top significant features that drive the price group of the houses in Nashville. The results emphasizes that the original land value of the property coupled with property age, and how long ago the sale happened, the value of the property might be over or underpriced.

Based on the correlation analysis, latest the property sale, higher are the chances of overpriced house sale. This gives an important information about the market trend for the Nashville area. If a property was recently sold at a high price, it may indicate a good investment opportunity. On the other hand, if the last sale was a long time ago, the current market value of the property may have changed significantly. Knowing the last sale date of a property can help investors, buyers, and sellers make informed decisions and understand the current market conditions. On the other hand, knowing the actual age of the property itself can be very helpful in estimating the value of the property. Older the property, the property value can go down as maintenance and upkeep costs would be high. Keeping up with current building regulations can also put a heavy cost on the property during renovation. Properties that are newer or

recently renovated tend to have a higher resale value compared to older properties. The age of properties in a neighborhood can affect the overall value of the area and its desirability as a place to live or invest. Overall, properties depreciate over time, and the age of the property can affect its market value and investment potential.

Secondly, the value of the land on which a house is built can contribute to both under and over pricing of the house. Land value can be impacted by the location, particularly high-demand areas where amenities are good like good infrastructure can drive the value of the houses in the region high. Zoon restrictions can impact the real-estate value, along with fluctuating market conditions. The value of the land can also be affected by broader market conditions, such as supply and demand, interest rates, and economic growth. the value of the land can have a significant impact on the overall value and pricing of a house, and it is important for the investors to consider the value of the land when making real estate decisions. On a final note, this information can help the real-estate company focus on these details of the new properties before investing in them.

References

Bujokas, E. (2022, June 2). Feature Importance in Decision Trees. Medium. https://towardsdatascience.com/feature-importance-in-decision-trees-e9450120b445

Logistic Regression using Statsmodels - GeeksforGeeks. (2023). GeeksforGeeks. https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/

Singh, A. (2021, September 22). How to compare multiple machine learning models? Medium. https://medium.com/nerd-for-tech/how-to-compare-multiple-machine-learning-models-a679f9802e5d

sklearn.ensemble.GradientBoostingClassifier. (2023). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

sklearn.ensemble.RandomForestClassifier. (2023). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

sklearn.tree.DecisionTreeClassifier. (2023). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Appendix 1 – EDA

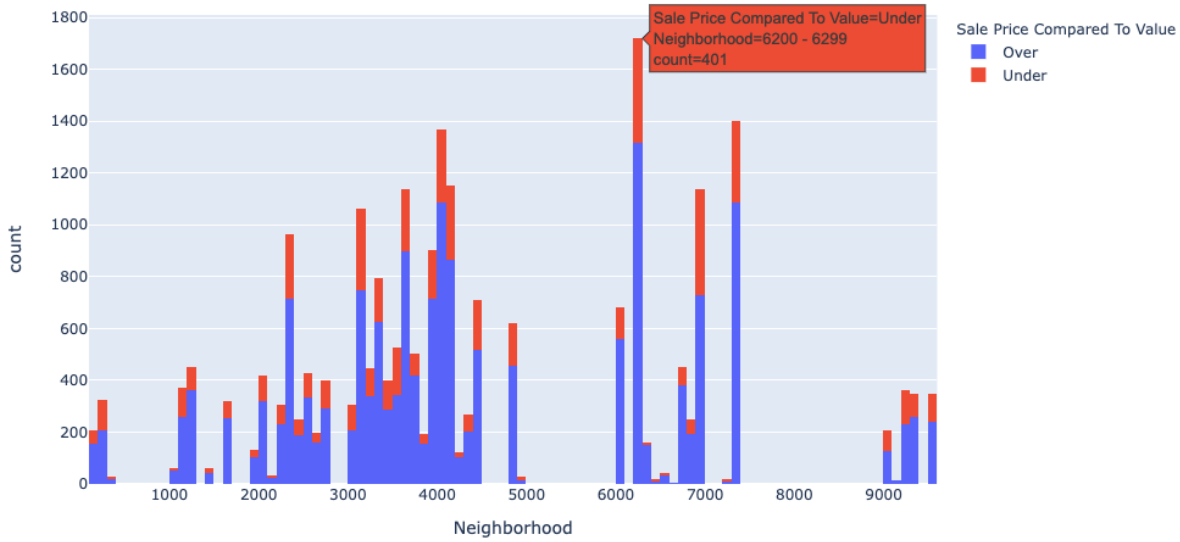

Figure 1: Histogram of neighborhood zip codes.

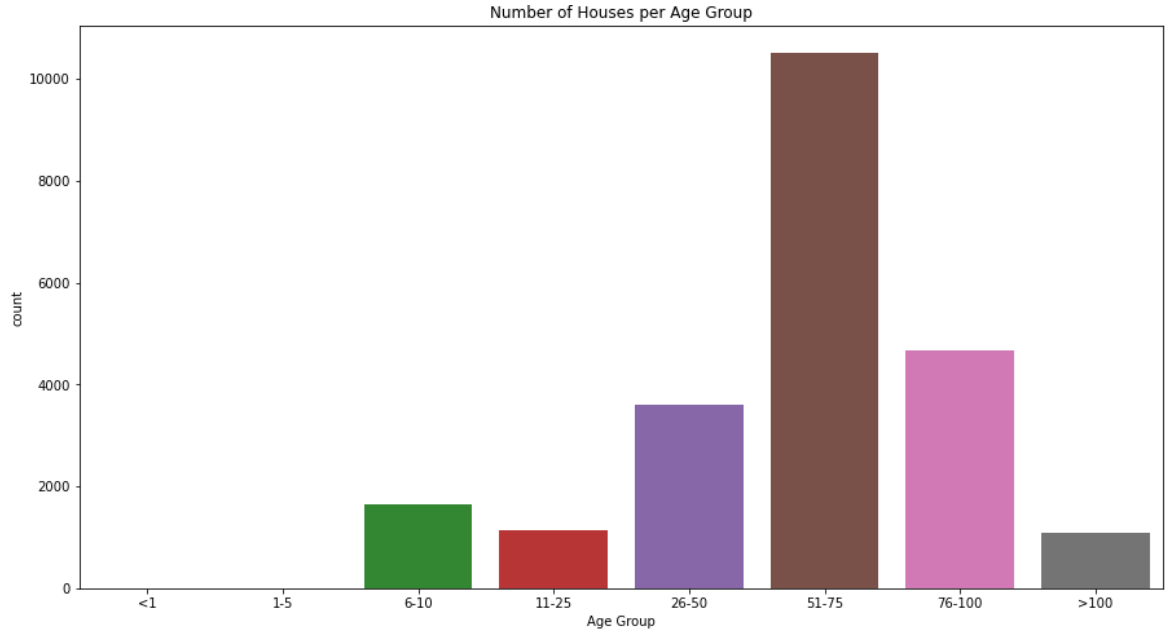

Figure 2: Number of houses per age group of properties.

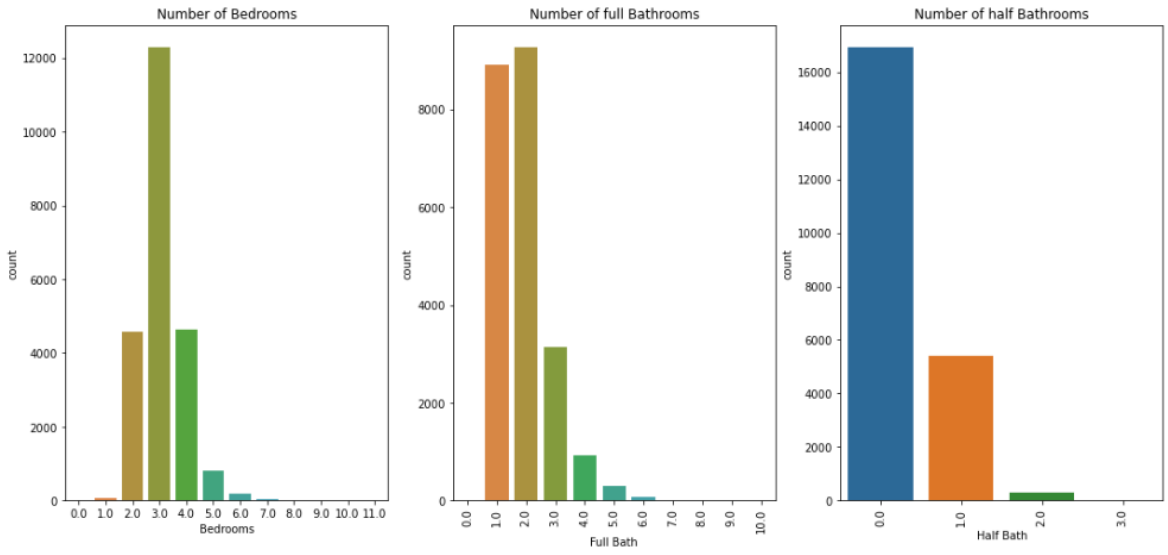

Figure 3: Count plots of bedrooms, full bath and half bath.

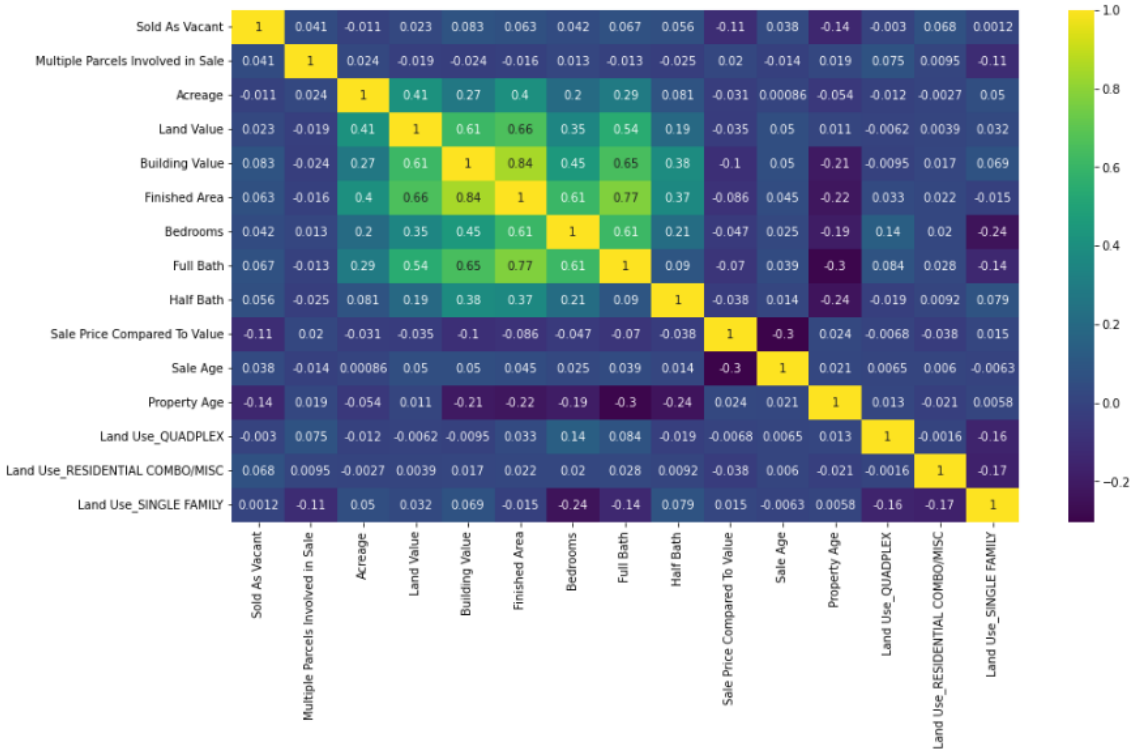

Figure 4: Correlation plot.

Appendix 2 – Model Results

```
Optimization terminated successfully.
         Current function value: 0.614289
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:     Sale Price Compared To Value   No. Observations:                18120
Model:                                 Logit   Df Residuals:                    18106
Method:                                 MLE   Df Model:                           13
Date:                      Sat, 04 Feb 2023   Pseudo R-squ.:                -0.09715
Time:                              01:09:36   Log-Likelihood:                -11131.
converged:                             True   LL-Null:                       -10145.
Covariance Type:                 nonrobust   LLR p-value:                     1.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Sold As Vacant                  -7.9065      0.223    -35.471      0.000      -8.343      -7.470
Multiple Parcels Involved in Sale   0.0505      0.017      2.897      0.004      0.016       0.085
Acreage                         -0.0536      0.018     -2.940      0.003      -0.089      -0.018
Land Value                       0.1531      0.024      6.358      0.000       0.106       0.200
Building Value                  -0.2159      0.036     -5.933      0.000      -0.287      -0.145
Finished Area                   -0.0555      0.042     -1.316      0.188      -0.138       0.027
Bedrooms                         0.0367      0.022      1.643      0.100      -0.007       0.080
Full Bath                       -0.0089      0.030     -0.296      0.767      -0.068       0.050
Half Bath                        0.0043      0.019      0.222      0.824      -0.034       0.042
Sale Age                        -0.5779      0.017    -34.878      0.000      -0.610      -0.545
Property Age                    -0.0197      0.018     -1.101      0.271      -0.055       0.015
Land Use_QUADPLEX               -0.0118      0.016     -0.744      0.457      -0.043       0.019
Land Use_RESIDENTIAL COMBO/MISC   -0.0549      0.018     -3.028      0.002      -0.090      -0.019
Land Use_SINGLE FAMILY           0.0391      0.018      2.231      0.026       0.005       0.073
================================================================================================
speed: 0:00:00.083410
```

Figure 1: Logistic Regression results

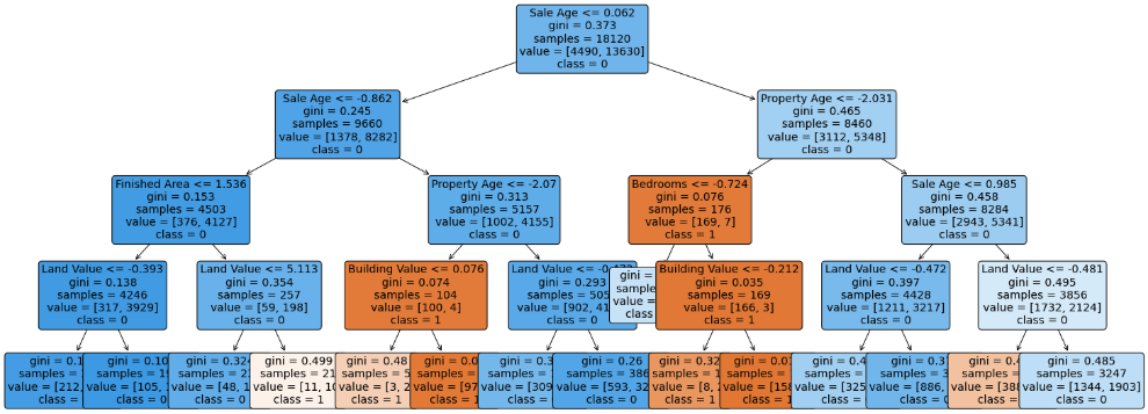

Figure 2: Decision Tree result

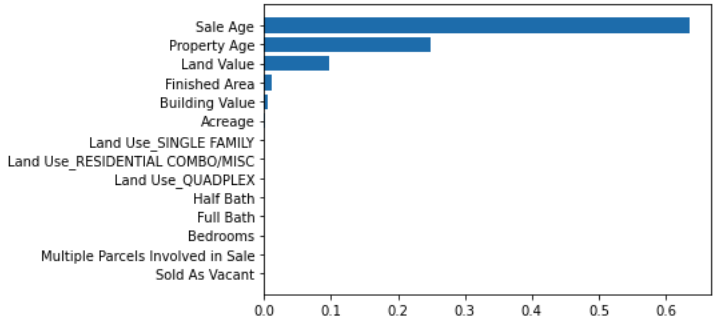

Figure 3: Feature importance of decision tree

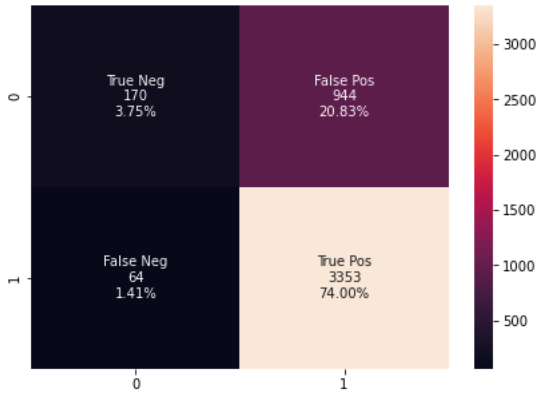

Figure 4: Confusion Matrix for decision tree

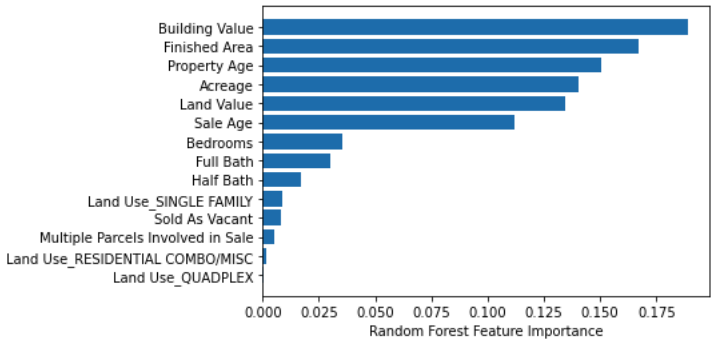

Figure 5: Random Forest Feature Importance bar graph

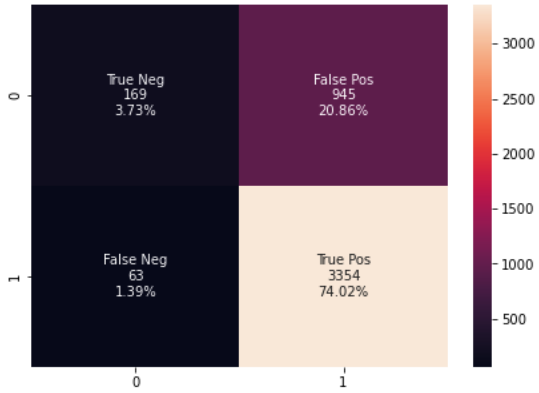

Figure 6: Confusion matrix for random forest.

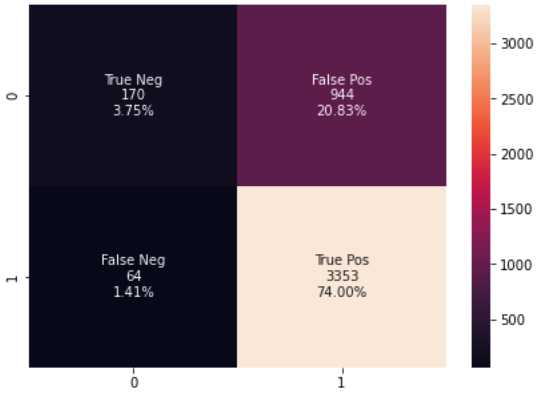

Figure 7: Confusion matrix for gradient boosting.