Module 4 Project

Investing In Nashville

Ujwal Kumar

Northeastern University

ALY6020: Predictive Analytics

Prof. Justin Grosz

12/04/2022

Content

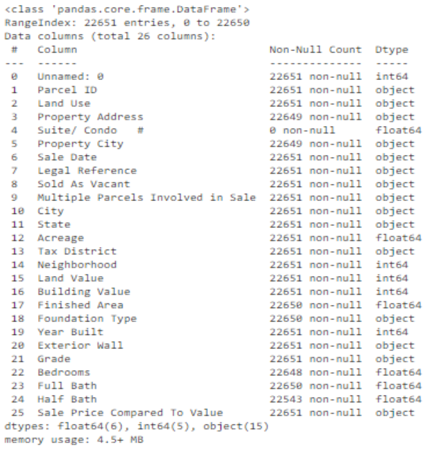
Introduction

Buying a property is a complicated undertaking since it involves several factors. After assessing each of the circumstances, such as the number of beds, baths, land areas, and so on. When these expectations are met, the most important factor is the price. The worth of a property is determined by its qualities, after which it is determined if it is overpriced or underpriced.

To start with, we will clean the dataset in this project. That is, we will check for null values and take appropriate action on the columns with null values, and then we will check for values in different columns to gain knowledge of the dataset. Following that, we will enter the exploratory data analysis phase, where we will plot several sorts of graphs and explore their relationships. Data preparation will take place during the feature engineering stage, and we will alter the data so that we may go on to the modeling step. We will utilize four models in the modeling stage: Logistic Regression, Decision Trees, Random Forest Classifier, and Gradient Boosting Classifier. Finally, we will compare the results and select the best-performing model from the group.

Data Cleaning

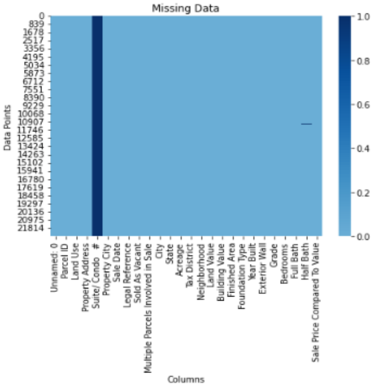
The dataset has around 22650 entries. It has 26 characteristics of various datatypes such as int, float, and object. After determining the number of features and records in the dataset, we examined the types of values included in the features to have a better idea of the sorts of values present in the columns.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22651 entries, 0 to 22650
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Unnamed: 0                       22651 non-null  int64  
 1   Parcel ID                        22651 non-null  object 
 2   Land Use                         22651 non-null  object 
 3   Property Address                 22649 non-null  object 
 4   Suite/ Condo   #                 0 non-null      float64
 5   Property City                    22649 non-null  object 
 6   Sale Date                        22651 non-null  object 
 7   Legal Reference                  22651 non-null  object 
 8   Sold As Vacant                   22651 non-null  object 
 9   Multiple Parcels Involved in Sale  22651 non-null  object 
 10  City                             22651 non-null  object 
 11  State                            22651 non-null  object 
 12  Acreage                          22651 non-null  float64
 13  Tax District                     22651 non-null  object 
 14  Neighborhood                     22651 non-null  int64  
 15  Land Value                       22651 non-null  int64  
 16  Building Value                   22651 non-null  int64  
 17  Finished Area                    22650 non-null  float64
 18  Foundation Type                  22650 non-null  object 
 19  Year Built                       22651 non-null  int64  
 20  Exterior Wall                    22651 non-null  object 
 21  Grade                            22651 non-null  object 
 22  Bedrooms                         22648 non-null  float64
 23  Full Bath                        22650 non-null  float64
 24  Half Bath                        22543 non-null  float64
 25  Sale Price Compared To Value     22651 non-null  object 
dtypes: float64(6), int64(5), object(15)
memory usage: 4.5+ MB
```

(Figure 1: Data Types of Columns)

The next critical step was to look for a null value. There were multiple columns in the data set that contained null values, but one column, "Suite/Condo," had 100% of null values, which signifies that the field was null and had no values. We chose to remove the column because it had 100% null data. Other columns, such as Half Bath, Full Bath, Finished Area, and so on, contained null values, but they accounted for less than 1% of the overall amount. We opted to exclude the records from the data set because the number of null values was quite low and we had a dataset with more than 22.6K entries. We deleted around 120 entries in total, leaving us with approximately 22.5K records to work with.



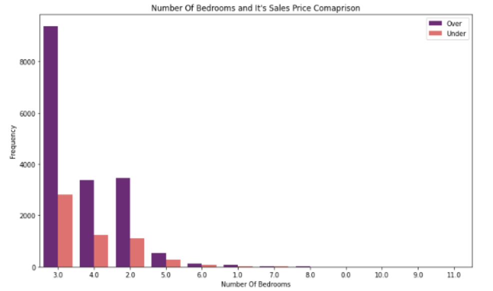
(Figure 2: Null Values Visualization)

We converted one column, "Sale Date," which was in the format YYYY-MM-DD, after dealing with null values. We separated the year, month, and day from each row of the Sale Date feature and placed them in new columns called Year, Month, and Day. This step was performed so that we could more effectively visualize the dates aspect, and the separated data will be useful throughout the model development stage.

Exploratory Data Analysis

To learn more about how Sale Price Compared To Value differs from other variables in the dataset. We will examine different figures to determine trends and which groups have influences on Sale Prices. Kindly refer to **Appendix 1** for EDA graphs and plots.

Sale price and their relationship with the number of beds feature is quite interesting; houses with 2, 3, and 4 beds are quite common as we had more records of them compared to another number of beds, and a common trend was observed that in each of the categories of beds, the sale price was overvalued compared to undervalued, whether it was in 2 beds, 3 beds, or 4 beds, etc.



(Figure 3: Bed Count and Corresponding Sales Price Category)

The relationship between Full Bath and Sales Price compared to Value Featured followed a similar pattern, with properties with 1, 2, or 3 Full Baths having a higher risk of being overpriced than houses with fewer Full Baths. Furthermore, if a property is unoccupied, there are greater odds of acquiring it at a lower price than if it is not vacant, and if a house is not vacant, there are more chances that it will be overvalued.

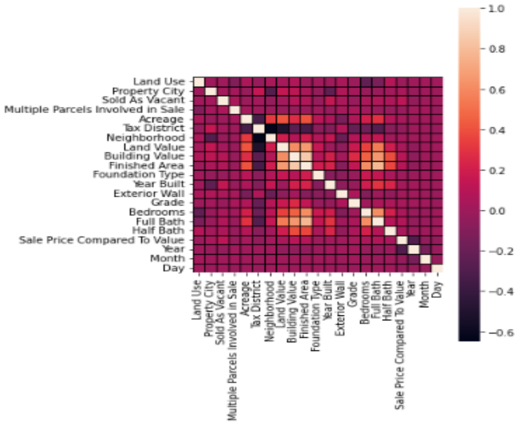
The trend in other features such as Grade, Exterior Wall Type, and Foundation Type, the majority of the values present in these features constitute price. Most of them are overpriced as compared to underpriced.

When looking at the months of house sales, March through May sees a consistent increase, with June seeing the largest amount of sales. Following then, the quantity of sales gradually decreases until October. However, there is a tiny increase in sales in December, leaving us with three months: January, February, and November. These three months have a higher possibility of receiving a good bargain than the other months.

Feature Engineering

After the completion of the data cleaning and EDA stage, we have adequate information about the attributes of the data sets. We will preprocess our data at this point so that it may be utilized for modeling. To begin, we addressed the outliers in continuous variables such as Finished Area, Land Value, Building Value, and Acreage. Outliers were handled using Quantile-based flooring and capping. The outlier is capped at a specified value above the 90th percentile or leveled at a factor less than the 10th percentile in this procedure. The graphs have been added to **Appendix 2**.

Following the successful treatment of outliers, the next step was to concentrate on categorical variables. As a result, we employed the Label Encoder approach to encoding all of the data set's category columns. Furthermore, we eliminated variables that were no longer relevant such as state, city, Legal Reference, etc in modeling and thus removed them. We finished these steps so that we could go on to the modeling stage.



(Figure 4: Correlation Matrix)

The above correlation matrix shows that our targeted variable, "Sale Price Compared To Value," is not substantially connected with any aspect. It is, nevertheless, connected to factors such as Furnished Area, Building Value, and Sold As Vacant. Now, in the modeling step, we'll evaluate which factors were significant in forecasting whether the sale price was under or overpriced.

Modeling

This report will complete further analysis by constructing models based on the preliminary findings obtained in the EDA and after finishing the preliminary examination of the data set. In this part, we will use several models such as Logistic Regression, Decision Trees Classifier, Random Forest Classifier, and Gradient Boosting Classifier. To analyze the results visually, all the graphs, classification reports, and confusion matrices have been added to **Appendix 3**.

After the feature engineering process was completed, the data was divided into 75:25 train and test sets. We fitted the training data to prediction models and then evaluated the models with test data. Following that, we recorded the model's accuracy, precision, and recall in order to evaluate the models.

Logistic Regression

Logistic Regression was the first model we evaluated. The model's output was extremely intriguing. This model's accuracy was about 75.3%, with precision and recall values of 0.53 and 0.01 respectively. This model's f1-score was 0.65. The model fared better in terms of precision, but not as well in terms of recall. Following that, we utilized p values and the SHAP approach to find the characteristics that were influencing the decisions. Building value, land value, finished area, and acreage were all crucial factors in prediction.

Decision Trees Classifier

The Decision Trees Classifier was the second model on which we evaluated our data. This model's accuracy was around 70.6%. The model scored 0.41 and 0.44 on assessment parameters such as accuracy and recall. This model's f1-score was 0.71. In terms of recall and f1-score, it outperformed Logistic Regression. However, the accuracy and precision value is slightly lower when compared to logistic regression. Next up, we used the feature importance method to determine important features for decision trees. Building Value, Year Built, Finished Area, and Year(Sale Year) were the factors influencing Decision Tree Classifier outcomes.

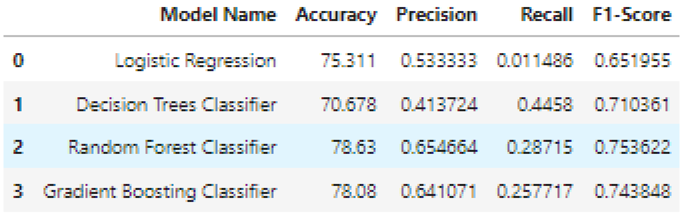
Random Forest Classifier

Random Forest Classifier is the third model in the case study. We utilized this technique to forecast whether the housing price will be over or under the price. The accuracy of this model was around 78.6%. Furthermore, the accuracy and recall values were 0.65 and 0.28, respectively. It also received an f1-score of 0.75. In terms of accuracy, precision, and f1-score, the Random Forest Model outperformed the prior two models. The recall value is lower than that of the Decision Trees Classifier but higher than that of Logistic Regression. The Random Forest

Classifier findings were influenced by the following factors: Building Value, Year Built, Finished Area, and Year (Sale Year).

Gradient Boosting Classifier

The Gradient Boosting Classifier is the final model we used in our report. The model provided us with 78% accuracy. Other measures, such as precision, recall, and f1-score, have values of 0.64, 0.25, and 0.74, respectively. This model outperformed Logistic Regression and Decision Trees Classifier in terms of accuracy, precision, and f1-score. However, it falls short of the Random Forest model since the evaluation metrics of this model are lower than those of the Random Forest Classifier. Next up, we used the feature importance method to determine important features for Gradient Boosting Classifier. Year (Sale Year), Year Built, Land Use, and Building Value are all important factors in Gradient Boosting Classifier prediction.



	Model Name	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	75.311	0.533333	0.011486	0.651955
1	Decision Trees Classifier	70.678	0.413724	0.4458	0.710361
2	Random Forest Classifier	78.63	0.654664	0.28715	0.753622
3	Gradient Boosting Classifier	78.08	0.641071	0.257717	0.743848

(Figure 5: Model Performances)

After reviewing the tabular data above, which includes model performance and measures such as Precision, Recall, F1 Score, and Accuracy. The call for the best model was close but Random Forest Classifier is the one. The Random Forest Classifier Model distinguishes out from the crowd since it not only has higher accuracy but also performs better on precision and F1 score. Furthermore, the Random Forest Classifier has a lower recall value than the Decision Trees Classifier but a higher recall value than the other two models we used, Logistic Regression and Gradient Boosting Classifier. Overall, the Random Forest Classifier is the model that the Real Estate business should use to estimate the value of residences as under or overpriced.

One of the most important components of post-prediction analysis is establishing which features had a substantial influence on prediction. To establish feature relevance, we employed SHAP and p values for logistic regression and feature significance parameters for the rest of the classifiers, and the key criteria that affected model predictions were Year (Sale Year), Year Built, Finished Area, and Building Value. These were the four most important criteria that affected prediction. These characteristics were also crucial in the EDA section. The real estate corporation have to prioritize these attributes before purchasing any real estate properties since, in our case study thus far, these features have driven the outcomes and can be highly useful in the future. Year built and building value can provide insight into the property and its likely value classification, whether it is under or overpriced. The real estate firm can make a judgment based on that.

Conclusion

After completing the case study, the following approach can help the Real Estate firm to invest in Nashville:

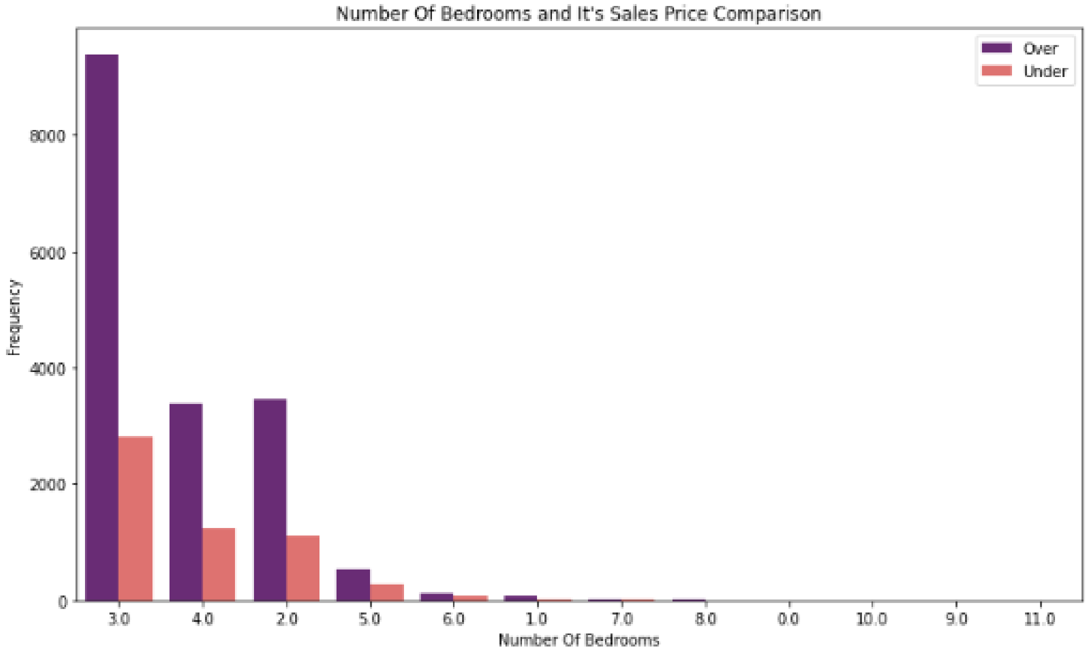
- The company should target areas with lower Finished Area values since the possibilities of obtaining an underpriced contract are better. The investment firm must hunt for a home with a low completed area since this will correspond to overpricing. They should first determine how much money they need to invest and how many finished square feet may be accommodated in order to receive a fair bargain. After considering these factors, they must determine if it is overpriced or underpriced. With this method, the corporation has a better chance of obtaining a lower-priced deal.
- Aside from the typical restrictions, time is an essential component of real estate investing. Prices are also affected by the passage of time. The corporation should aim to invest in the months of February or November, as our analysis shows that there are better possibilities of acquiring a deal at a lower price. Furthermore, the company should target unoccupied homes. As a result, whether it is a vacant property and the closing date is in November or February. There is a good probability of receiving a good bargain.

References

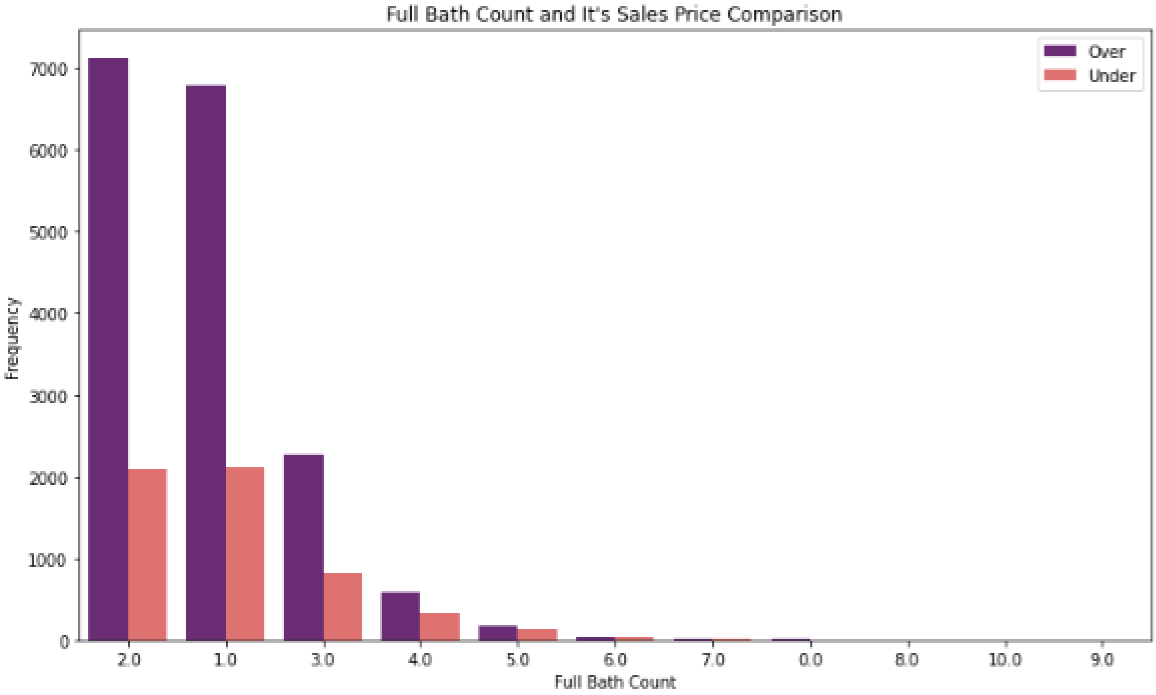
- Brownlee, J. (2020, August 14). *A gentle introduction to the gradient boosting algorithm for machine learning*. MachineLearningMastery.com. Retrieved December 4, 2022, from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
- *Random forest classifier: A complete guide to how it works in Machine Learning*. Built In. (n.d.). Retrieved December 4, 2022, from https://builtin.com/data-science/random-forest-algorithm

Appendix 1

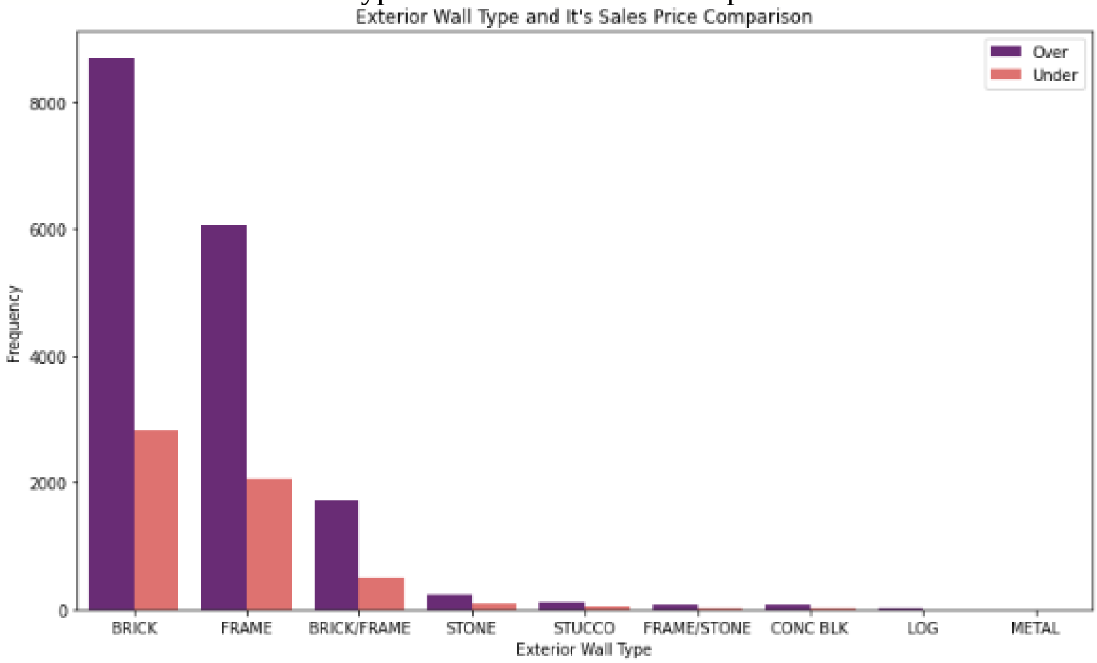
Count Plot of Number Of Bedrooms and It's Sales Price Comparison



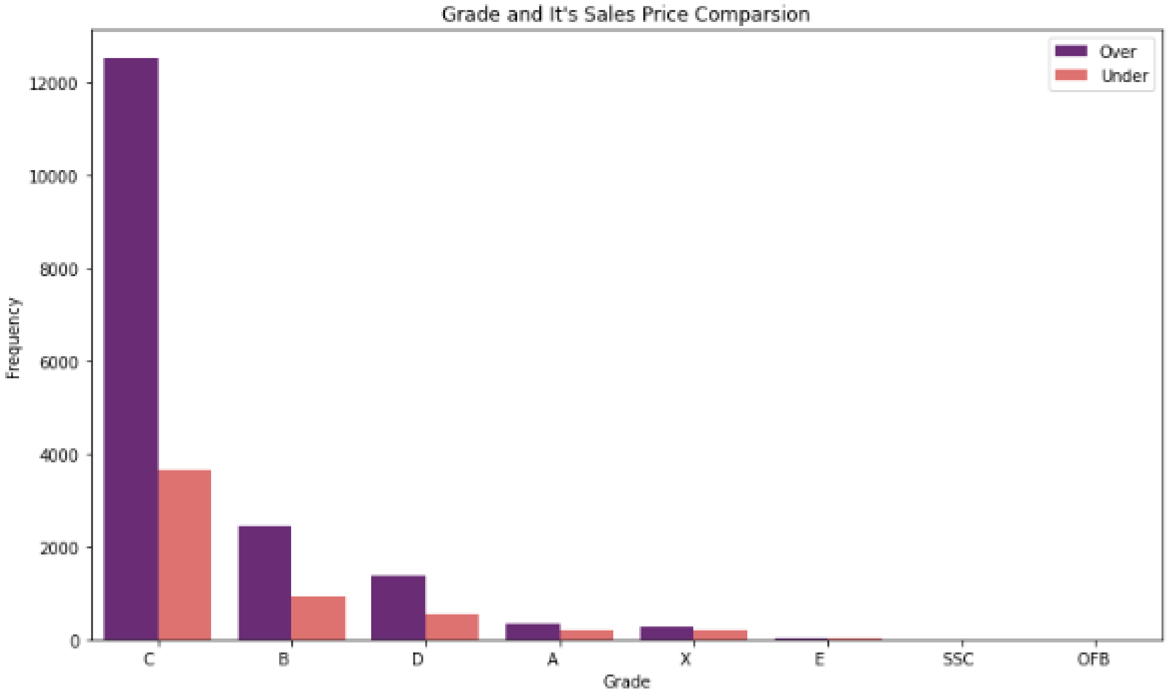
Count Plot of Full Bath Count Feature and It's Sales Price Comparison



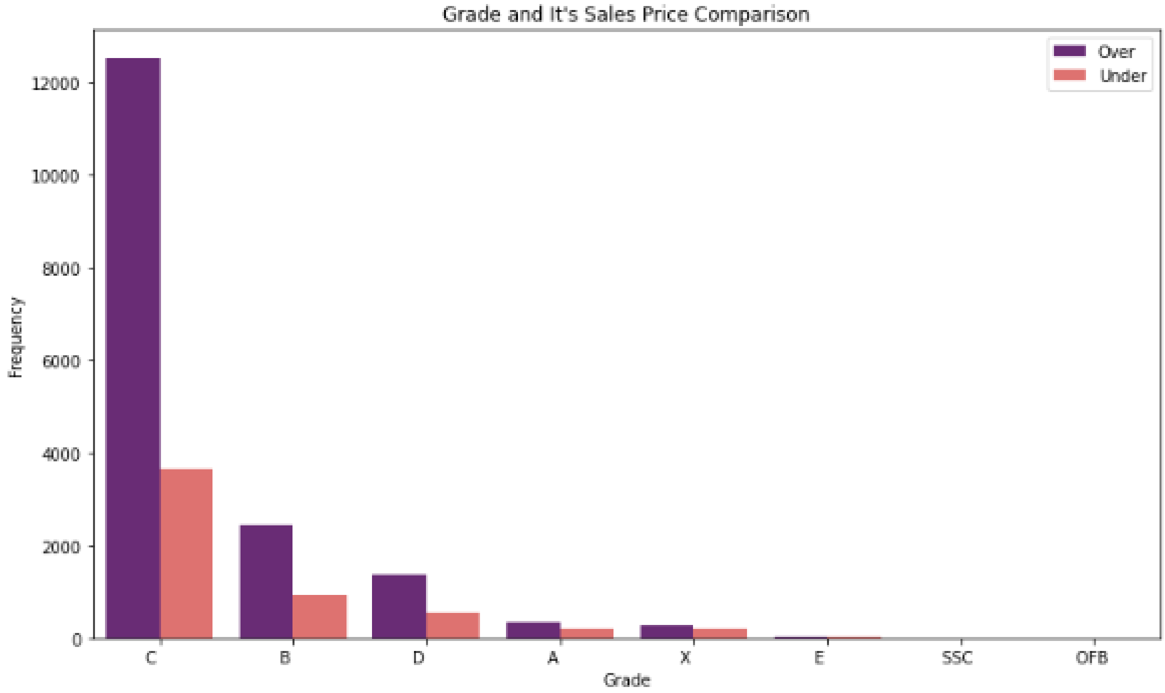
Count Plot of Exterior Wall Type and It's Sales Price Comparison



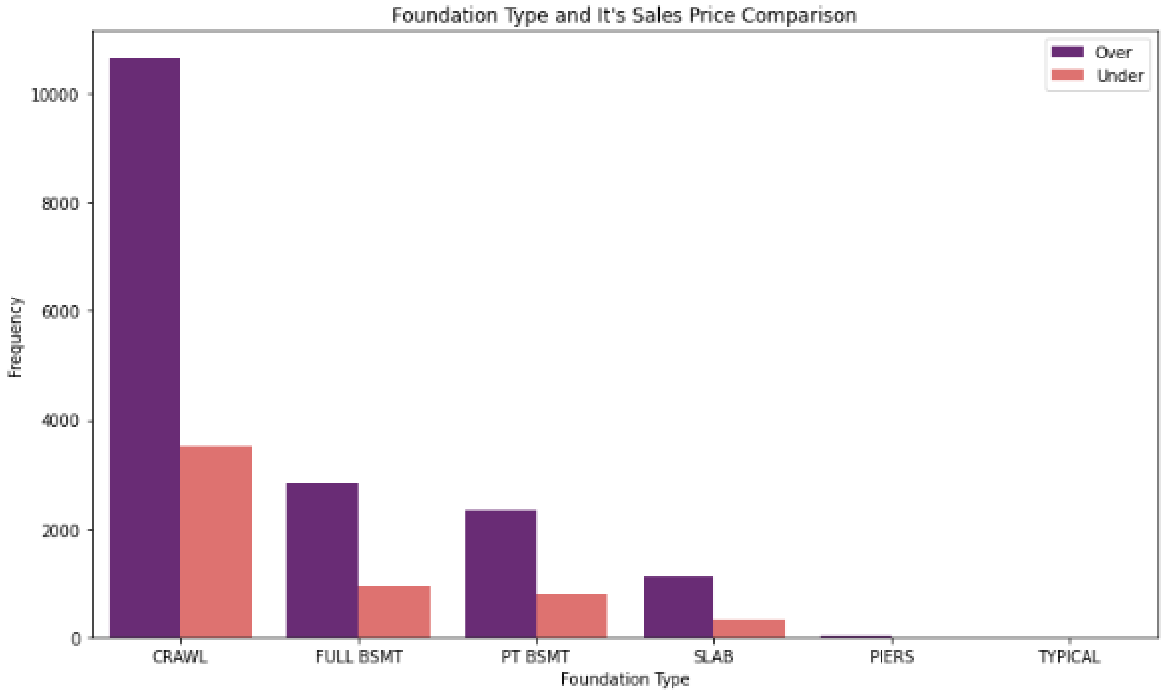
Count Plot of Grade and It's Sales Price Comparison



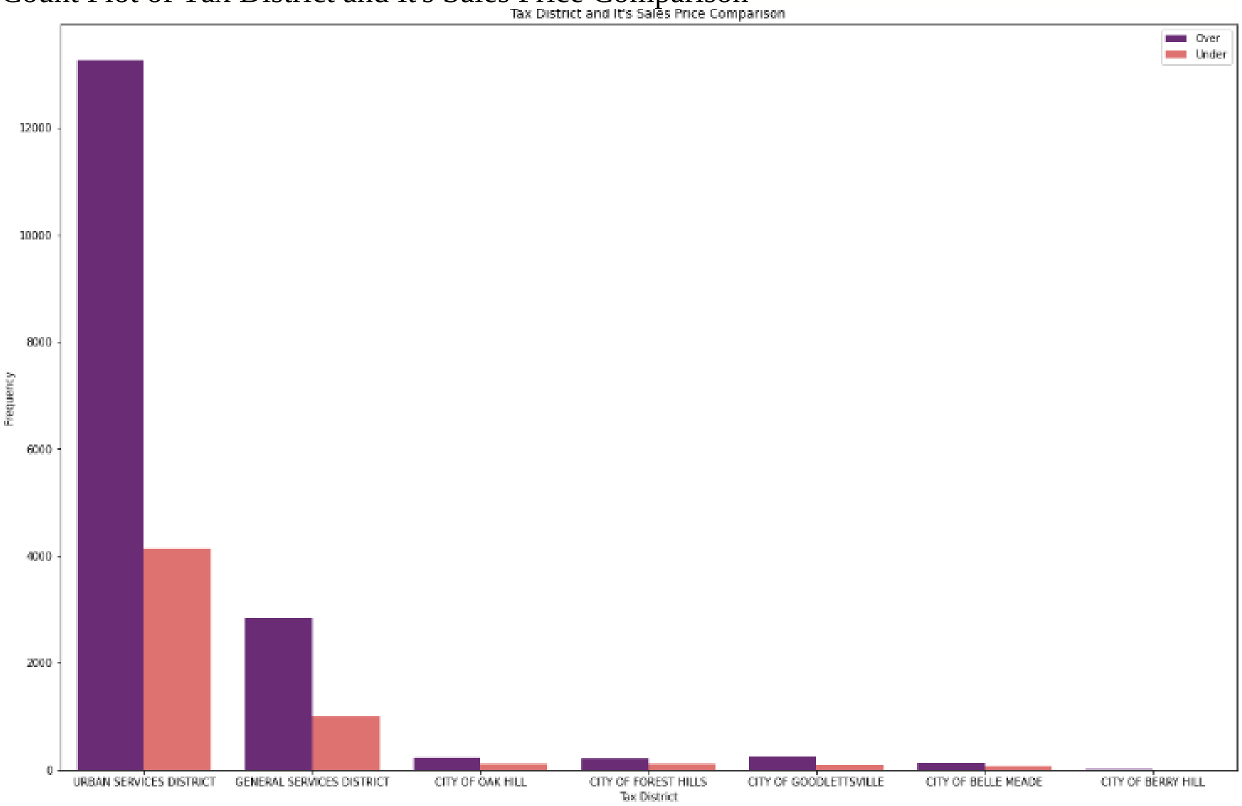
Count Plot of Grade and It's Sales Price Comparison



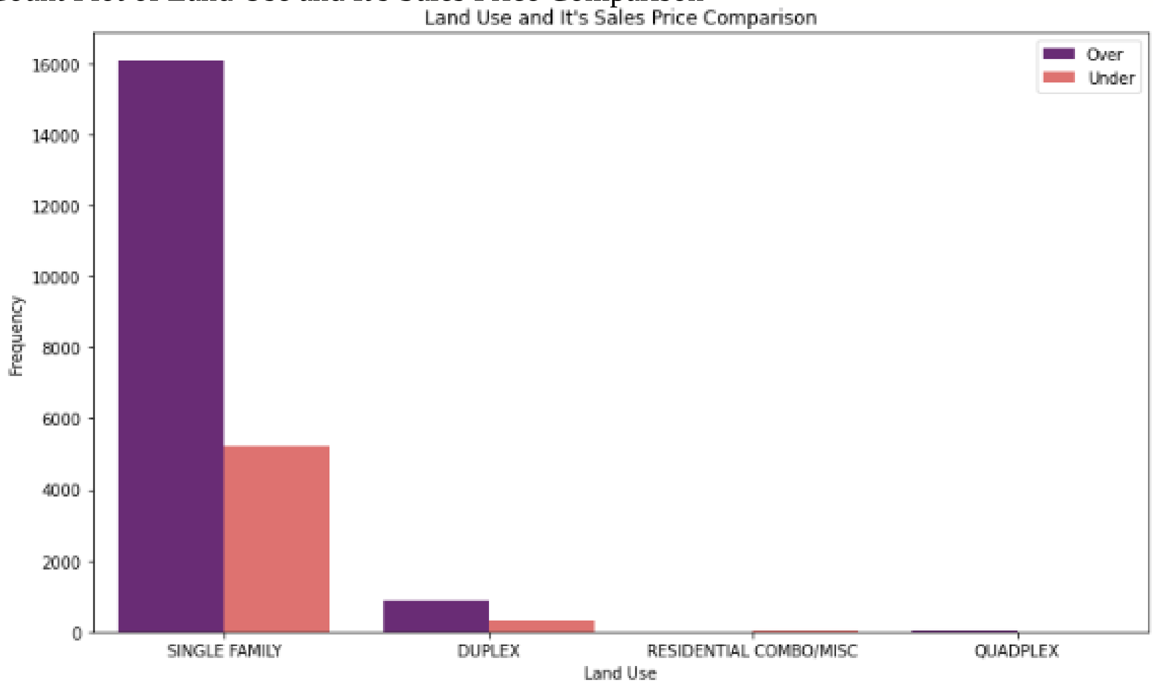
Count Plot of Foundation Type and It's Sales Price Comparison



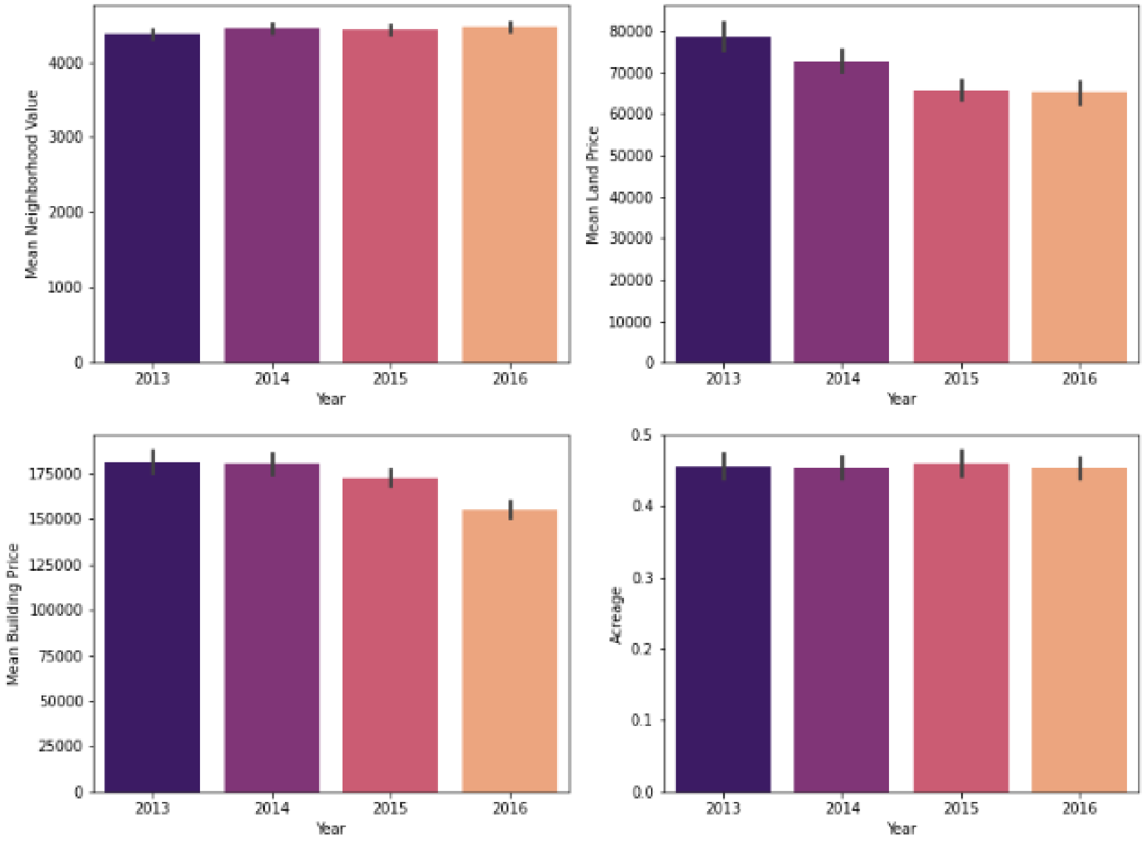
Count Plot of Tax District and It's Sales Price Comparison



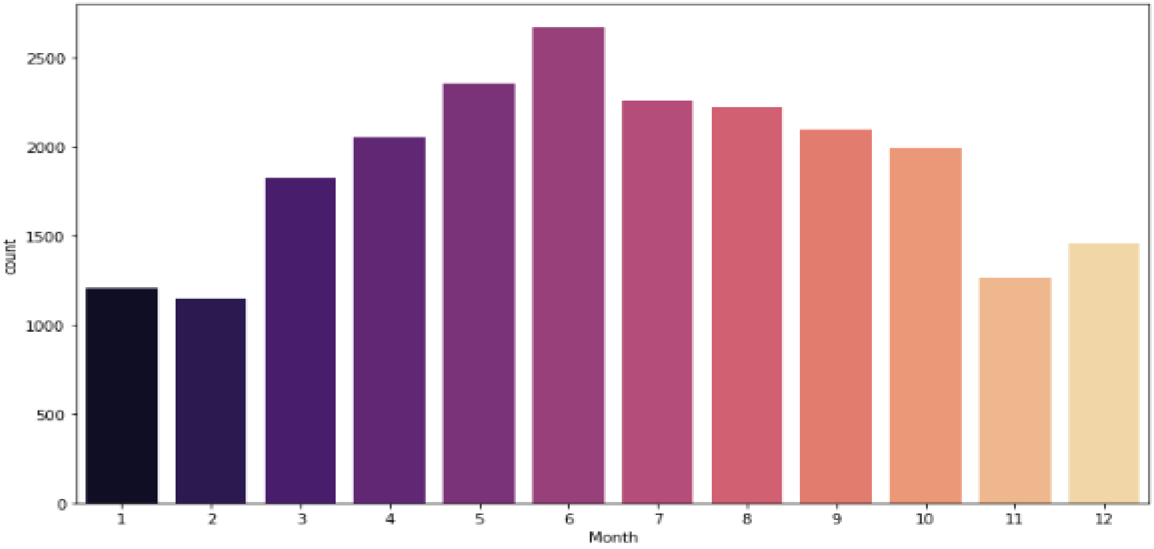
Count Plot of Land Use and It's Sales Price Comparison



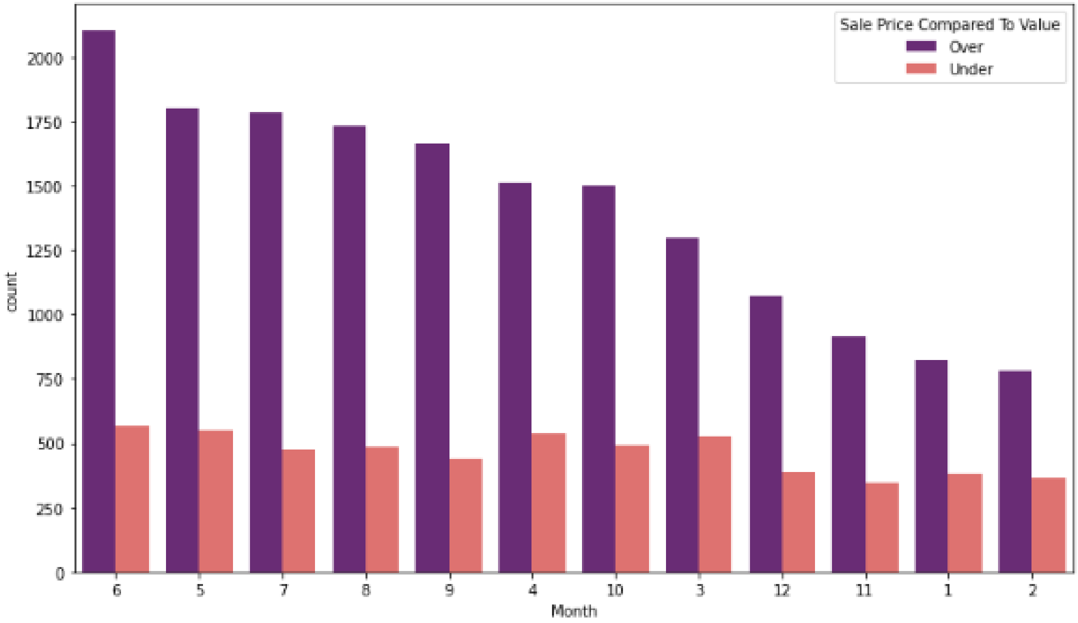
Mean Value of Continuous Features Across Sale Years



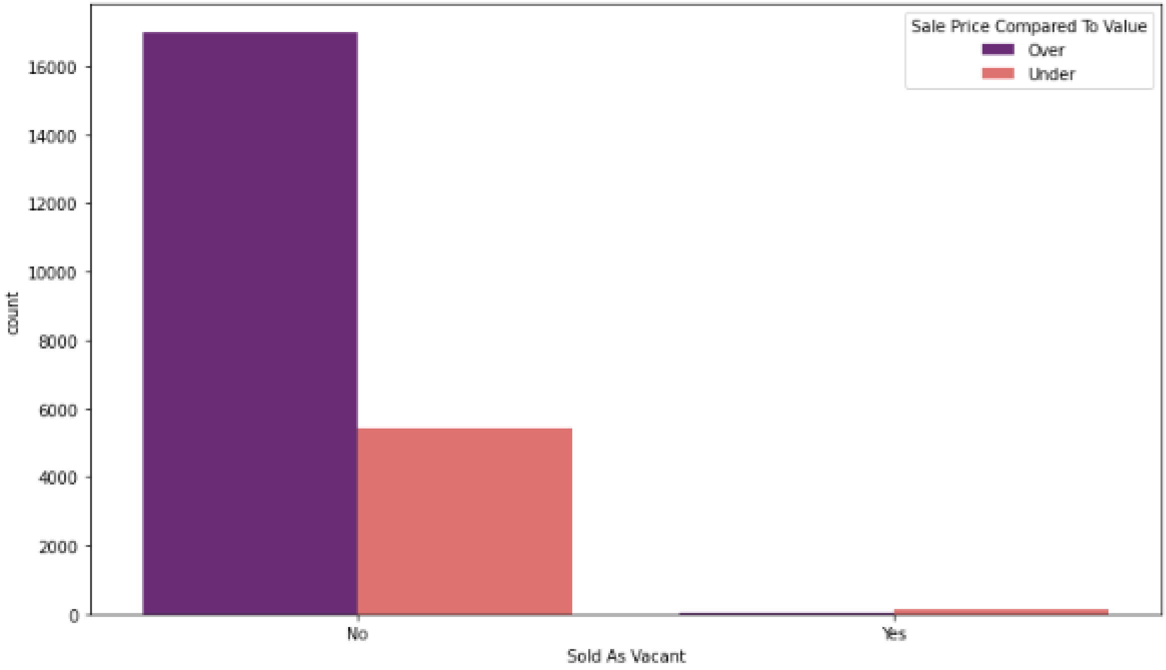
Count Plot of Month Feature



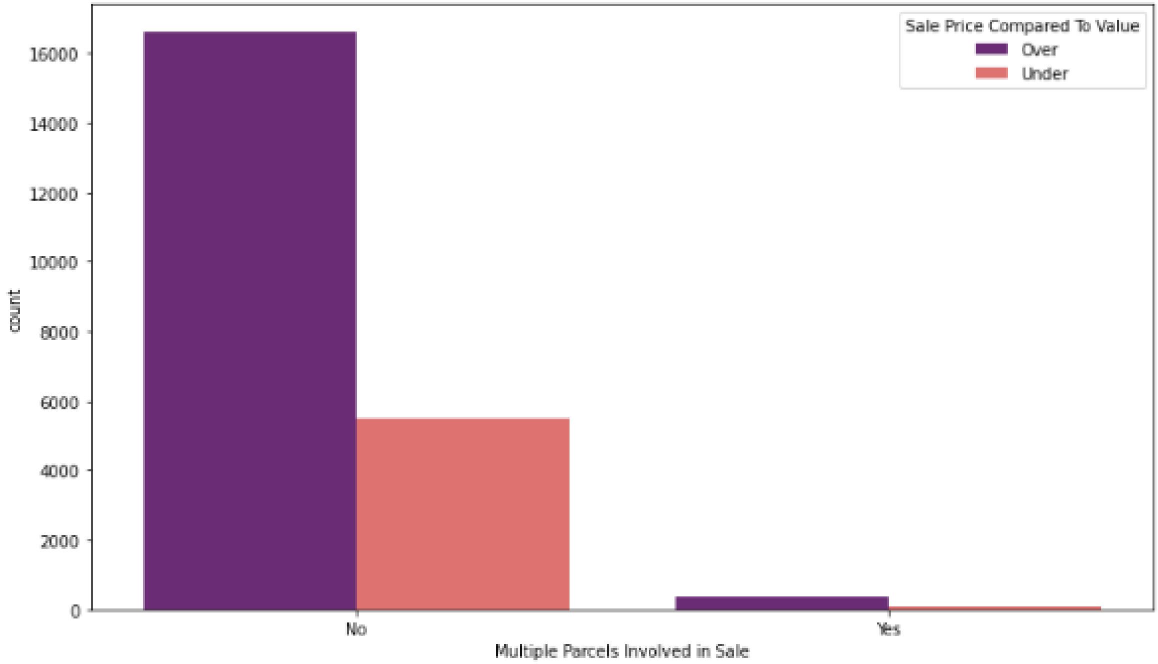
Count Plot of Month Feature and It’s Relation with Sale Price



Count Plot of Sold As Vacant Feature and It’s Relation with Sale Price

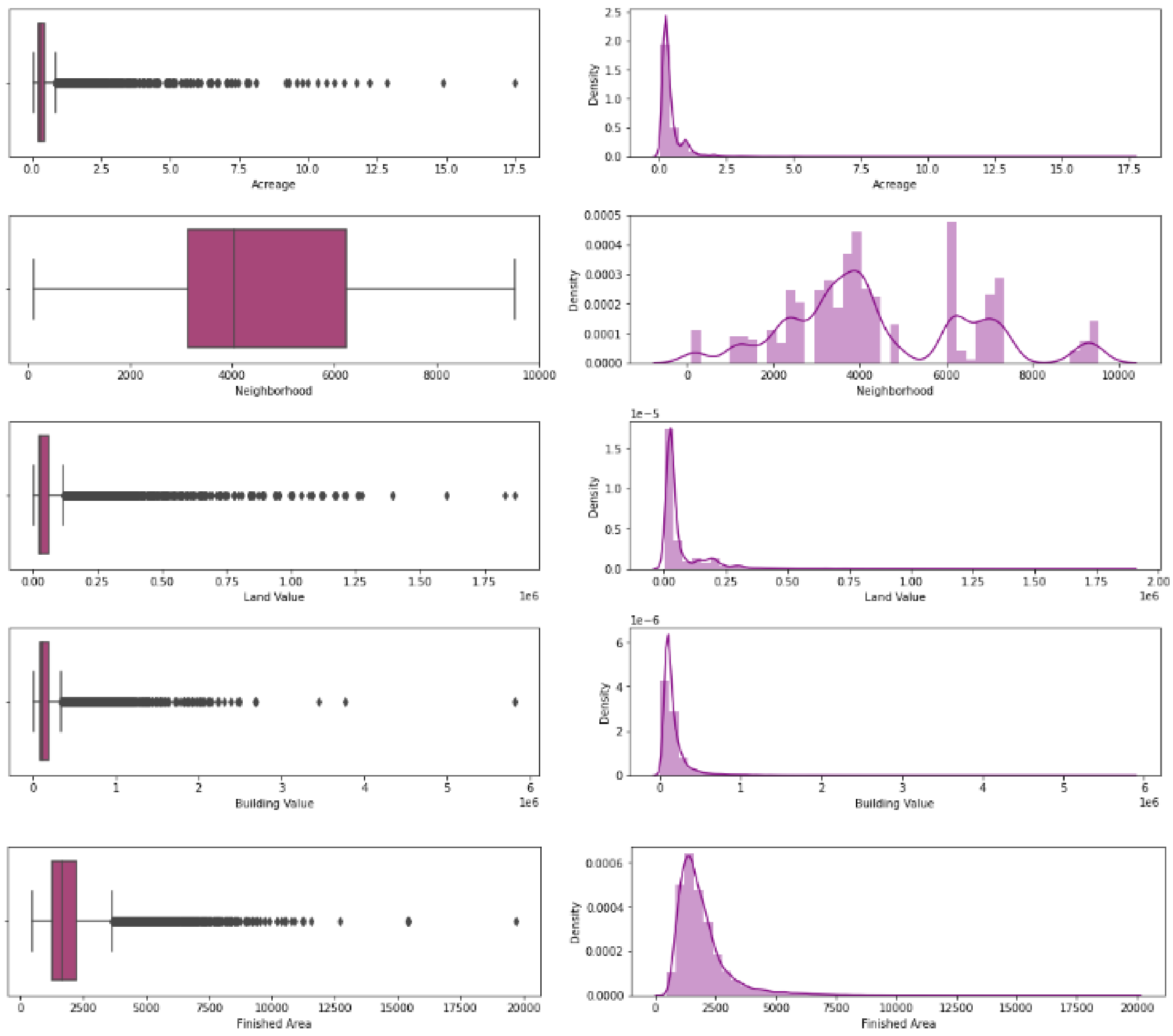


Count Plot of Multiple Parcels Involved in Sale and It’s Relation with Sale Price

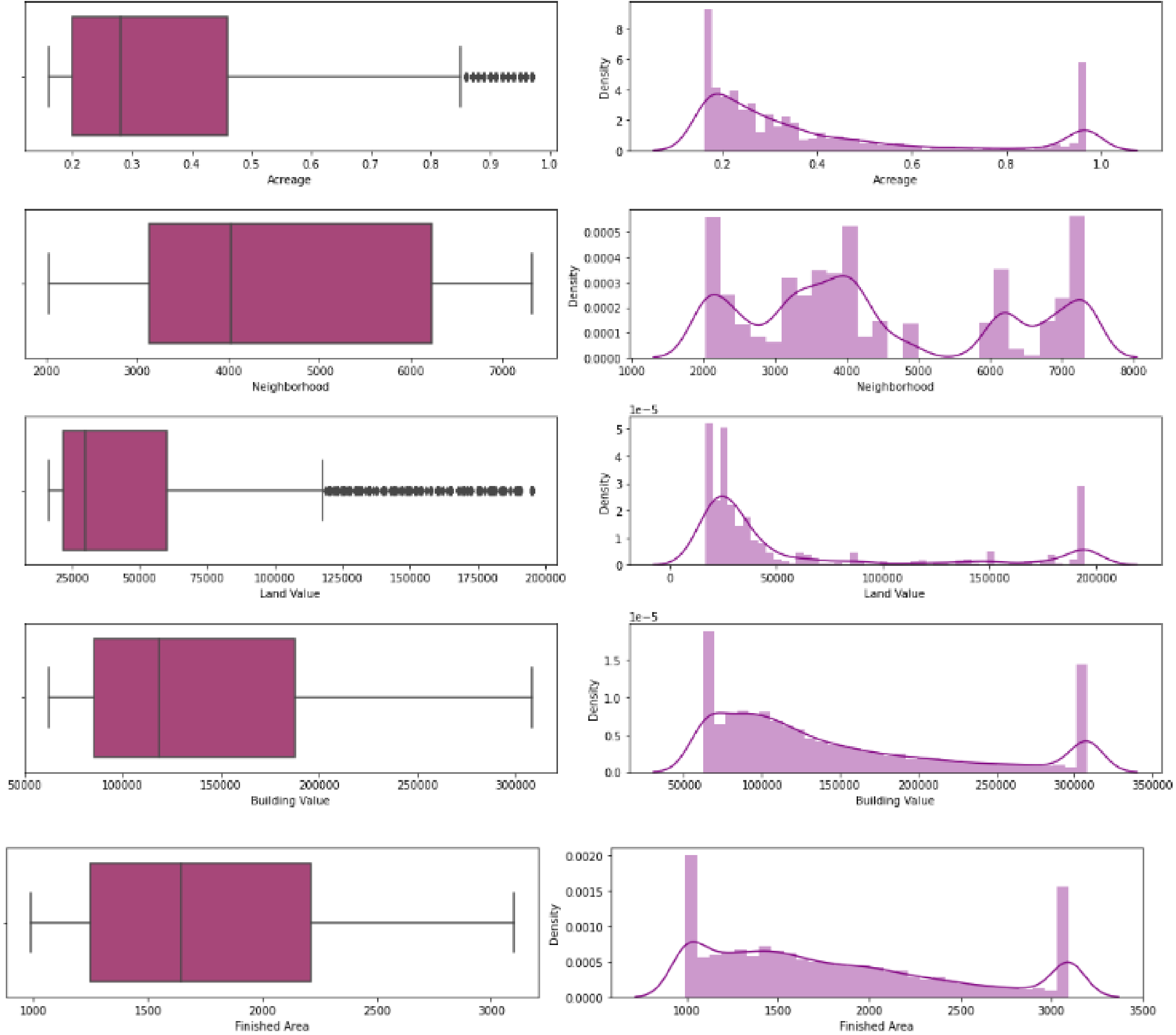


Appendix 2

Distribution and Box Plot of Continuous Variables Before Treating Outliers

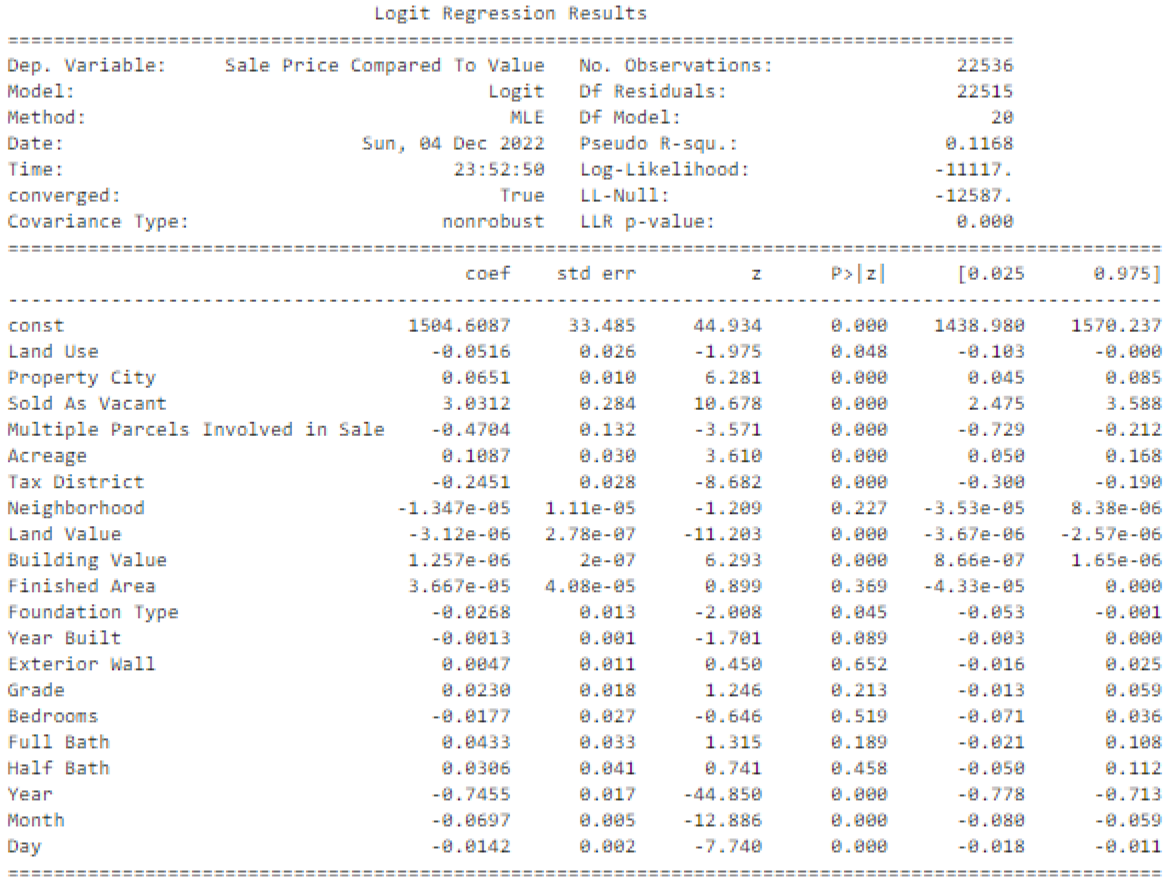


Distribution and Box Plot of Continuous Variables After Treating Outliers



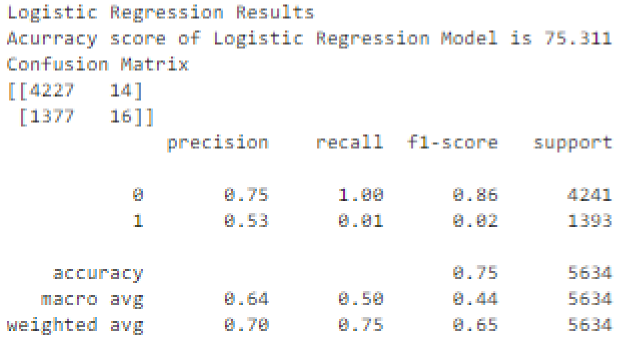
Appendix 3

Logistic Regression Results



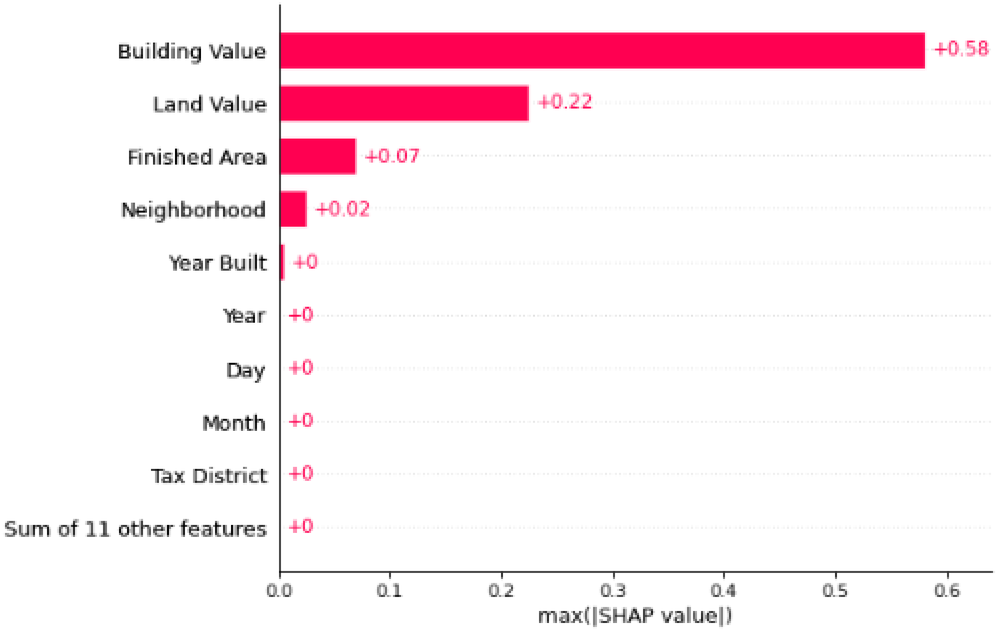
```
                           Logit Regression Results
==============================================================================
Dep. Variable:     Sale Price Compared To Value   No. Observations:                22536
Model:                                    Logit   Df Residuals:                    22515
Method:                                     MLE   Df Model:                           20
Date:                          Sun, 04 Dec 2022   Pseudo R-squ.:                  0.1168
Time:                                  23:52:50   Log-Likelihood:                -11117.
converged:                                 True   LL-Null:                       -12587.
Covariance Type:                      nonrobust   LLR p-value:                     0.000
=====================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                          1504.6087     33.485     44.934      0.000    1438.980    1570.237
Land Use                         -0.0516      0.026     -1.975      0.048      -0.103      -0.000
Property City                     0.0651      0.010      6.281      0.000       0.045       0.085
Sold As Vacant                    3.0312      0.284     10.678      0.000       2.475       3.588
Multiple Parcels Involved in Sale -0.4704     0.132     -3.571      0.000      -0.729      -0.212
Acreage                           0.1087      0.030      3.610      0.000       0.050       0.168
Tax District                     -0.2451      0.028     -8.682      0.000      -0.300      -0.190
Neighborhood                  -1.347e-05   1.11e-05     -1.209      0.227   -3.53e-05    8.38e-06
Land Value                     -3.12e-06   2.78e-07    -11.203      0.000   -3.67e-06   -2.57e-06
Building Value                 1.257e-06      2e-07      6.293      0.000    8.66e-07    1.65e-06
Finished Area                  3.667e-05   4.08e-05      0.899      0.369   -4.33e-05       0.000
Foundation Type                  -0.0268      0.013     -2.008      0.045      -0.053      -0.001
Year Built                       -0.0013      0.001     -1.701      0.089      -0.003       0.000
Exterior Wall                     0.0047      0.011      0.450      0.652      -0.016       0.025
Grade                             0.0230      0.018      1.246      0.213      -0.013       0.059
Bedrooms                         -0.0177      0.027     -0.646      0.519      -0.071       0.036
Full Bath                         0.0433      0.033      1.315      0.189      -0.021       0.108
Half Bath                         0.0306      0.041      0.741      0.458      -0.050       0.112
Year                             -0.7455      0.017    -44.850      0.000      -0.778      -0.713
Month                            -0.0697      0.005    -12.886      0.000      -0.080      -0.059
Day                              -0.0142      0.002     -7.740      0.000      -0.018      -0.011
=====================================================================================================
```

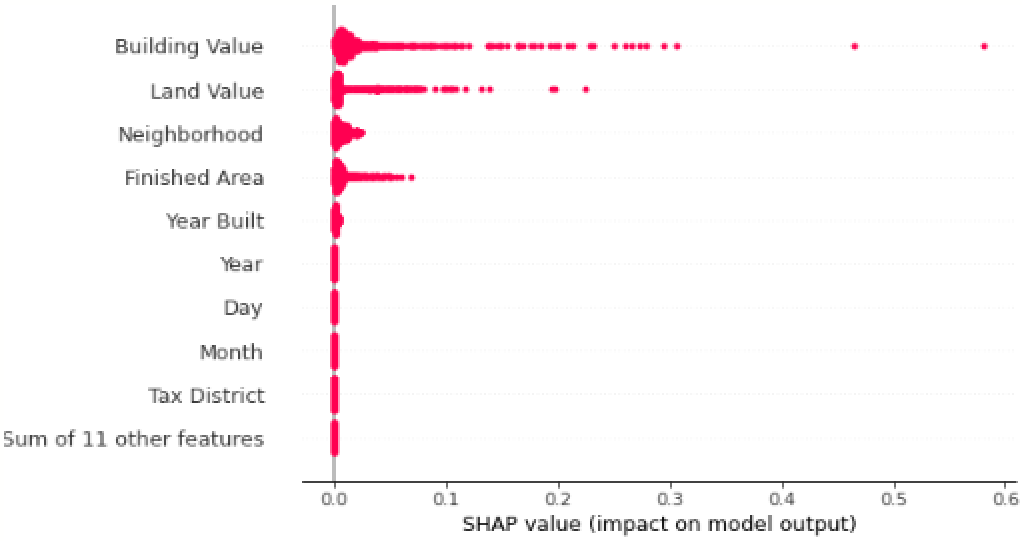
Classification Report, Confusion Matrix and Accuracy of Logistic Regression



```
Logistic Regression Results
Acurracy score of Logistic Regression Model is 75.311
Confusion Matrix
[[4227   14]
 [1377   16]]
              precision    recall  f1-score   support

           0       0.75      1.00      0.86      4241
           1       0.53      0.01      0.02      1393

    accuracy                           0.75      5634
   macro avg       0.64      0.50      0.44      5634
weighted avg       0.70      0.75      0.65      5634
```

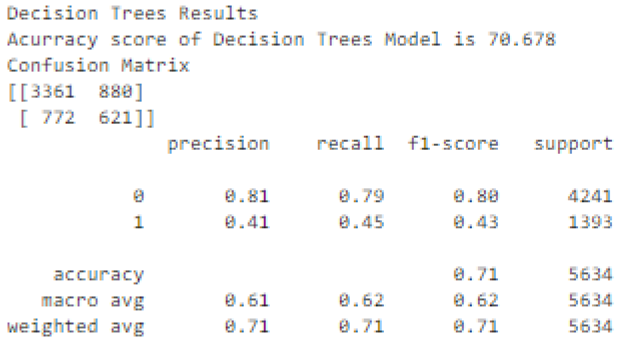
Using SHAP, visualizing feature importance





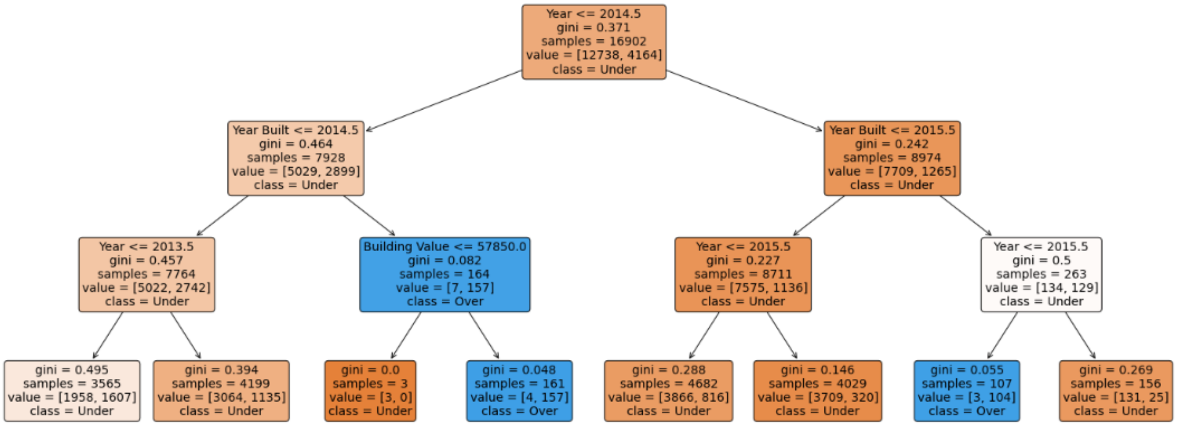
Decision Trees Classifier

Classification Report, Confusion Matrix, and Accuracy of Decision Trees Classifier

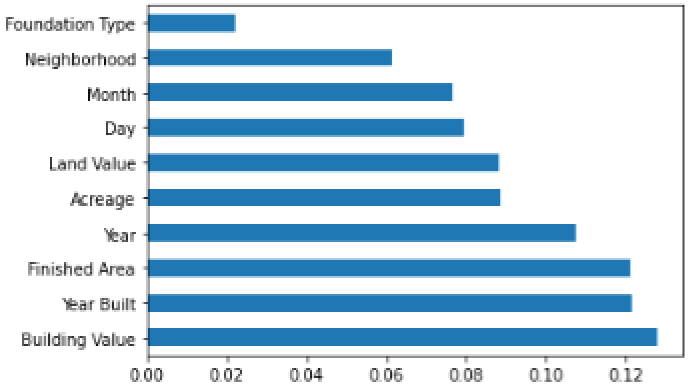


```
Decision Trees Results
Acurracy score of Decision Trees Model is 70.678
Confusion Matrix
[[3361  880]
 [ 772  621]]
              precision    recall  f1-score   support

           0       0.81      0.79      0.80      4241
           1       0.41      0.45      0.43      1393

    accuracy                           0.71      5634
   macro avg       0.61      0.62      0.62      5634
weighted avg       0.71      0.71      0.71      5634
```

Decision Tree Visual Representation

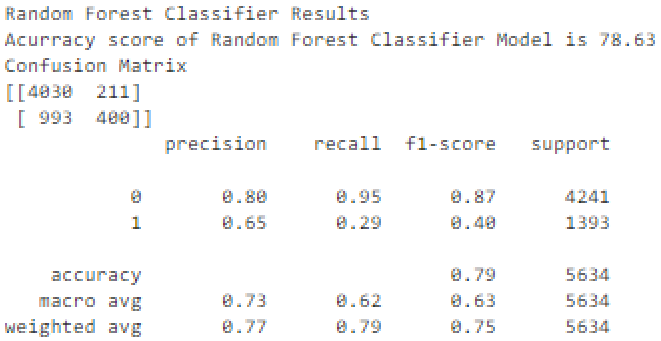


Feature Importance



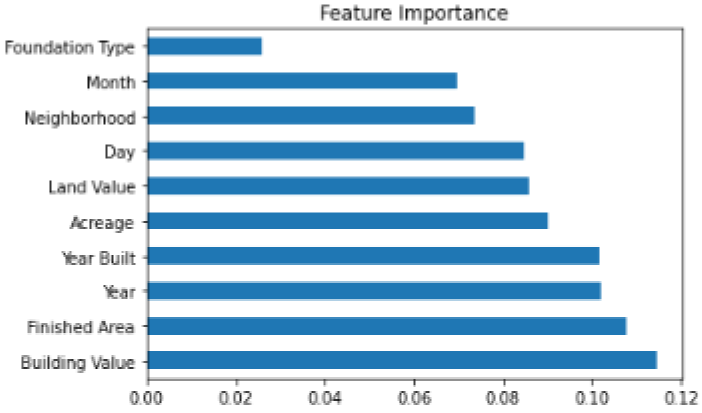
Random Forest Classifier

Classification Report, Confusion Matrix, and Accuracy of Random Forest Classifier



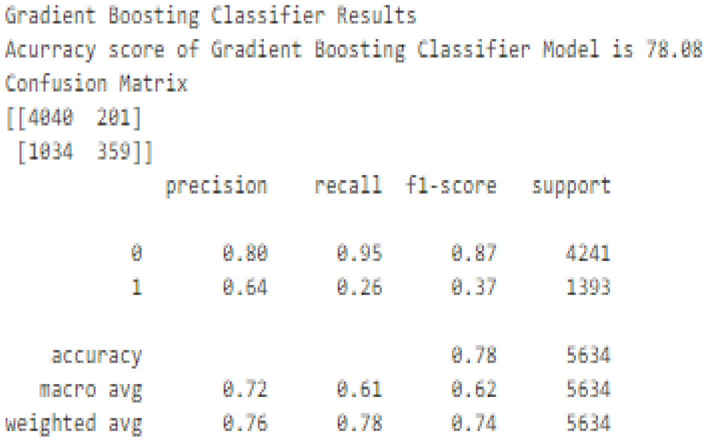
```
Random Forest Classifier Results
Acurracy score of Random Forest Classifier Model is 78.63
Confusion Matrix
[[4030  211]
 [ 993  400]]
              precision    recall  f1-score   support

           0       0.80      0.95      0.87      4241
           1       0.65      0.29      0.40      1393

    accuracy                           0.79      5634
   macro avg       0.73      0.62      0.63      5634
weighted avg       0.77      0.79      0.75      5634
```

Feature Importance



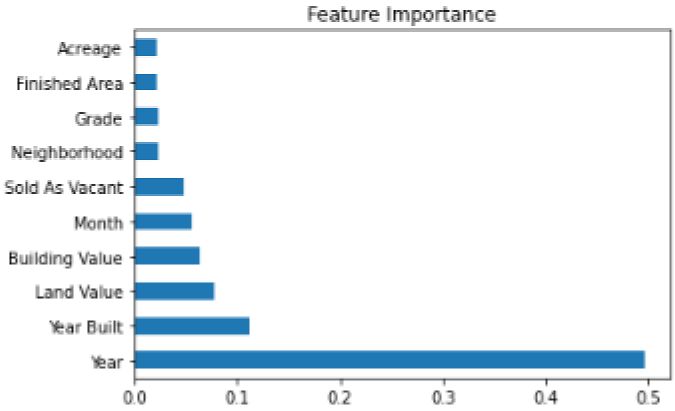
Gradient Boosting Classifier

Classification Report, Confusion Matrix, and Accuracy of Gradient Boosting Classifier



```
Gradient Boosting Classifier Results
Acurracy score of Gradient Boosting Classifier Model is 78.08
Confusion Matrix
[[4040  201]
 [1034  359]]
              precision    recall  f1-score   support

           0       0.80      0.95      0.87      4241
           1       0.64      0.26      0.37      1393

    accuracy                           0.78      5634
   macro avg       0.72      0.61      0.62      5634
weighted avg       0.76      0.78      0.74      5634
```

Feature Importance


Feature Importance
Acreage
Finished Area
Grade
Neighborhood
Sold As Vacant
Month
Building Value
Land Value
Year Built
Year