
# ALY 6020:

# PREDCTIVE ANALYTICS

## Week 4: Housing Price Prediction Models

Submitted To:
Prof. Chinthaka Pathum Dinesh Herath Gedara, Faculty Lecturer

Submitted By:
Abhilash Dikshit

Academic Term: Winter 2024
Northeastern University, Vancouver, BC,Canada
Master of Professional Studies in Analytics

February 03, 2024

Title: Report on Housing Price Prediction Models

## I. Objective:

The objective of this report is to present findings and recommendations based on the analysis of various machine learning models for predicting housing prices. The analysis includes Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting models.

## II. Dataset Overview:

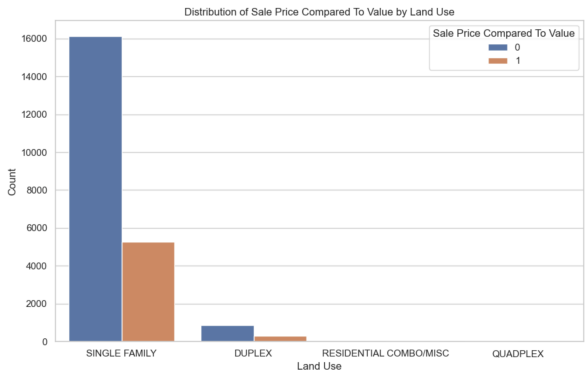
The dataset used for this analysis contains information on various features related to real estate properties, including acreage, land value, building value, finished area, and more.

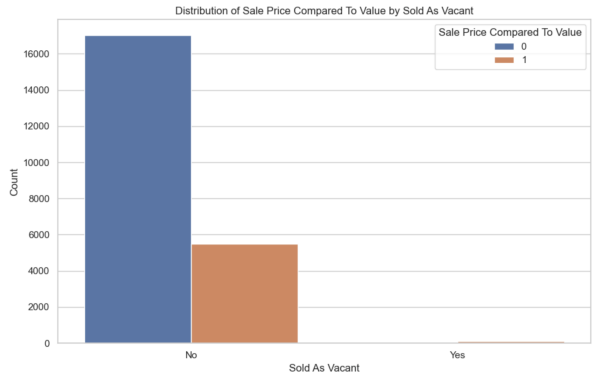
| | Unnamed: 0 | Parcel ID | Land Use | Property Address | Suite/ Condo # | Property City | Sale Date | Legal Reference | Sold As Vacant | Multiple Parcels Involved in Sale | ... | Building Value | Finished Area | Foundation Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 105 11 0 080.00 | SINGLE FAMILY | 1802 STEWART PL | NaN | NASHVILLE | 1/11/2013 | 20130118-0006337 | No | No | ... | 134400 | 1149.00000 | PT BSMT |
| **1** | 2 | 118 03 0 130.00 | SINGLE FAMILY | 2761 ROSEDALE PL | NaN | NASHVILLE | 1/18/2013 | 20130124-0008033 | No | No | ... | 157800 | 2090.82495 | SLAB |
| **2** | 3 | 119 01 0 479.00 | SINGLE FAMILY | 224 PEACHTREE ST | NaN | NASHVILLE | 1/18/2013 | 20130128-0008863 | No | No | ... | 243700 | 2145.60001 | FULL BSMT |
| **3** | 4 | 119 05 0 186.00 | SINGLE FAMILY | 316 LUTIE ST | NaN | NASHVILLE | 1/23/2013 | 20130131-0009929 | No | No | ... | 138100 | 1969.00000 | CRAWL |
| **4** | 5 | 119 05 0 387.00 | SINGLE FAMILY | 2626 FOSTER AVE | NaN | NASHVILLE | 1/4/2013 | 20130118-0006110 | No | No | ... | 86100 | 1037.00000 | CRAWL |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **22646** | 56602 | 176 01 0 003.00 | SINGLE FAMILY | 4617 ROCKLAND TRL | NaN | ANTIOCH | 10/13/2016 | 20161019-0110290 | No | No | ... | 105000 | 1758.00000 | CRAWL |
| **22647** | 56605 | 176 05 0 070.00 | SINGLE FAMILY | 5004 SUNSHINE DR | NaN | ANTIOCH | 10/26/2016 | 20161102-0115842 | No | No | ... | 142400 | 2421.00000 | SLAB |
| **22648** | 56607 | 176 09 0 003.00 | SINGLE FAMILY | 4964 HICKORY WOODS E | NaN | ANTIOCH | 10/28/2016 | 20161031-0114817 | No | No | ... | 159300 | 3117.00000 | SLAB |
| **22649** | 56614 | 082 05 0 040.00 | SINGLE FAMILY | 1625 5TH AVE N | NaN | NASHVILLE | 10/28/2016 | 20161102-0115988 | No | No | ... | 204100 | 1637.00000 | CRAWL |

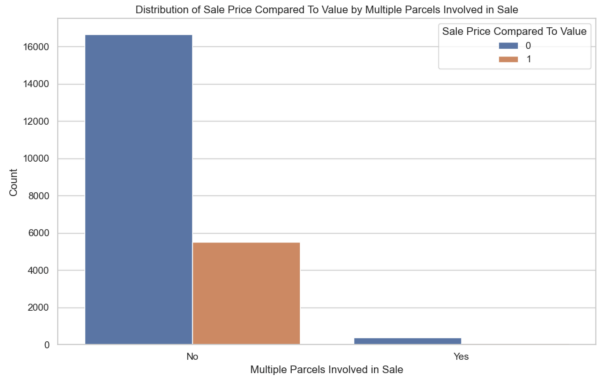
## III. Exploratory Data Analysis (EDA):

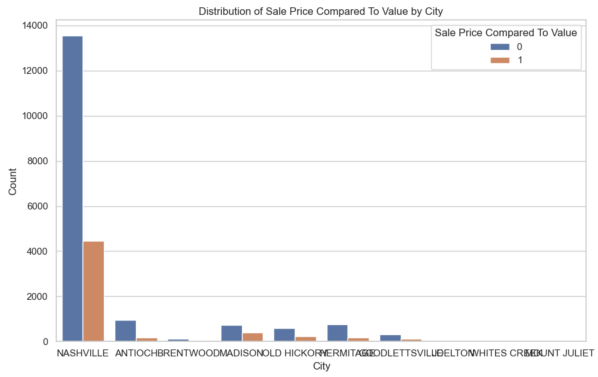
- Conducted a thorough exploratory data analysis to understand the distribution and relationships among different features.

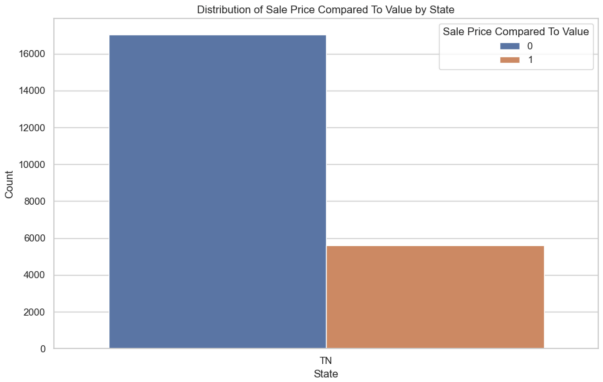
- Explored correlations between features and the target variable (Sale Price Compared To Value).

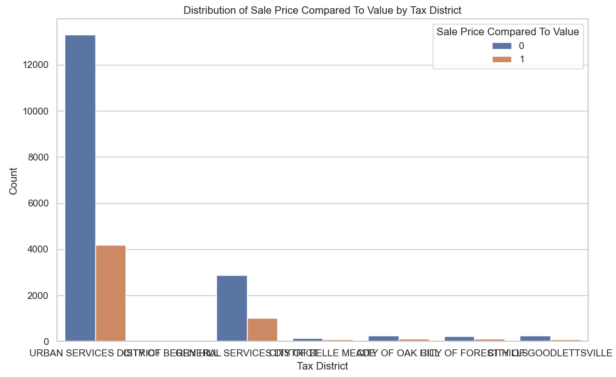


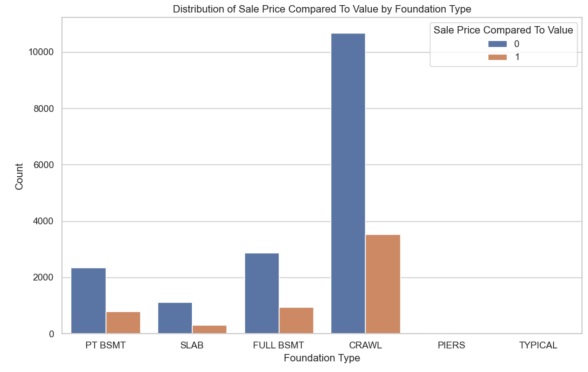



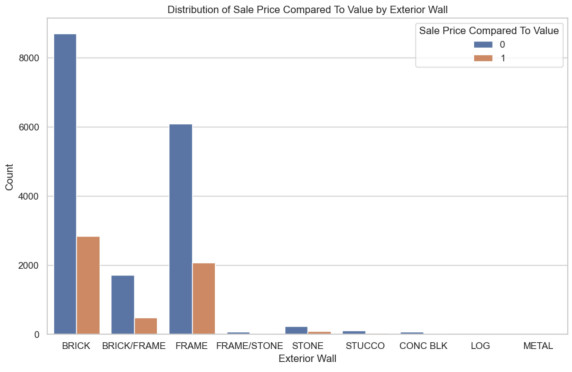



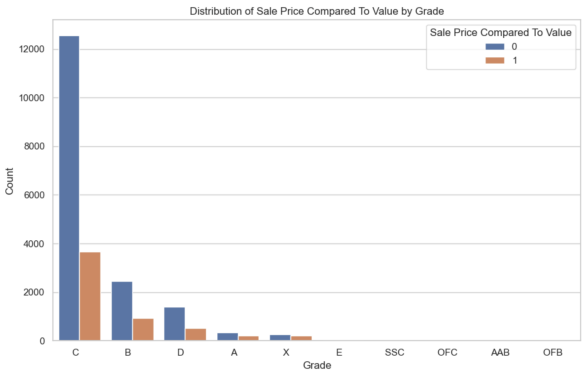






## IV. Data Preprocessing:

- Handled missing values and applied necessary imputations.

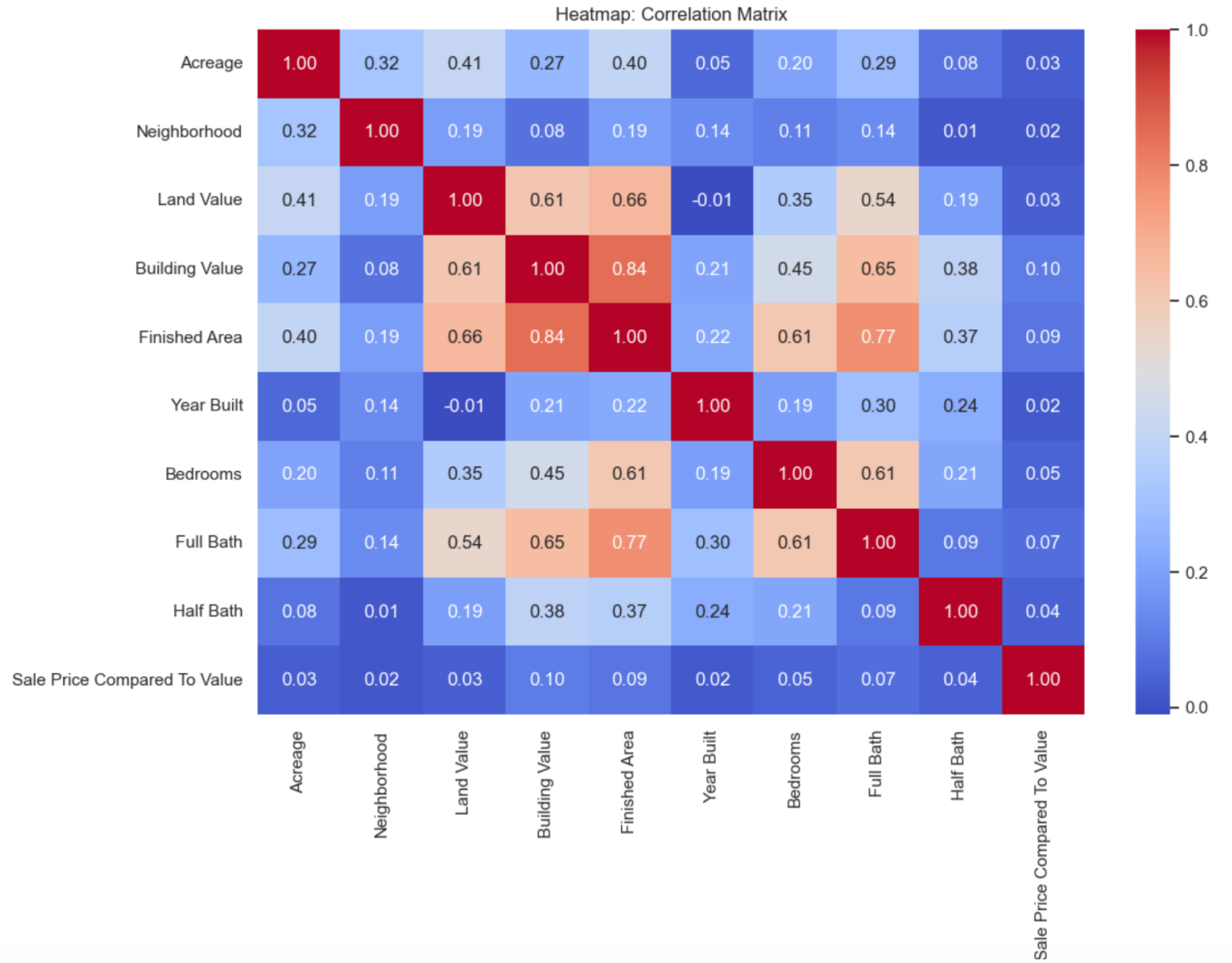
- Converted categorical variables into a suitable format for machine learning models.

- Ensured that the dataset is ready for modeling.

### Feature Selection:

- Selected a subset of features deemed important for predicting housing prices based on correlation analysis.

| | Acreage | Neighborhood | Land Value | Building Value | Finished Area | Year Built | Bedrooms | Full Bath | Half Bath | Sale Compar |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 22651.000000 | 22651.000000 | 2.265100e+04 | 2.265100e+04 | 22651.000000 | 22651.000000 | 22651.000000 | 22651.000000 | 22651.000000 | 22651.0( |
| **mean** | 0.454705 | 4432.715024 | 7.013797e+04 | 1.722402e+05 | 1915.377151 | 1961.947684 | 3.104910 | 1.887285 | 0.270239 | 0.2 |
| **std** | 0.611818 | 2142.803595 | 1.029035e+05 | 1.896424e+05 | 1079.070700 | 25.843908 | 0.829232 | 0.951199 | 0.479040 | 0.4 |
| **min** | 0.040000 | 107.000000 | 9.000000e+02 | 1.400000e+03 | 450.000000 | 1832.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0( |
| **25%** | 0.200000 | 3130.000000 | 2.200000e+04 | 8.550000e+04 | 1250.000000 | 1947.000000 | 3.000000 | 1.000000 | 0.000000 | 0.0( |
| **50%** | 0.280000 | 4026.000000 | 3.000000e+04 | 1.188000e+05 | 1646.000000 | 1959.000000 | 3.000000 | 2.000000 | 0.000000 | 0.0( |
| **75%** | 0.460000 | 6229.000000 | 6.030000e+04 | 1.882500e+05 | 2213.250000 | 1977.000000 | 4.000000 | 2.000000 | 1.000000 | 0.0( |
| **max** | 17.500000 | 9530.000000 | 1.869000e+06 | 5.824300e+06 | 19728.249880 | 2017.000000 | 11.000000 | 10.000000 | 3.000000 | 1.0( |




## V. Modeling:

```
Optimization terminated successfully.
         Current function value: 0.553718
         Iterations 5
                                Logit Regression Results
==============================================================================================
Dep. Variable:     Sale Price Compared To Value   No. Observations:                18120
Model:                                    Logit   Df Residuals:                    18115
Method:                                    MLE   Df Model:                            4
Date:                         Sat, 03 Feb 2024   Pseudo R-squ.:               0.007985
Time:                                 20:42:22   Log-Likelihood:                -10033.
converged:                                True   LL-Null:                       -10114.
Covariance Type:                    nonrobust   LLR p-value:                 6.891e-34
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.3258      0.041    -32.378      0.000      -1.406      -1.246
Acreage        0.0601      0.031      1.946      0.052      -0.000       0.121
Land Value  -1.057e-06   2.38e-07     -4.441      0.000   -1.52e-06   -5.9e-07
Building Value   1.27e-06   1.85e-07      6.865      0.000    9.07e-07    1.63e-06
Finished Area   1.443e-05   3.34e-05      0.433      0.665   -5.09e-05    7.98e-05
==============================================================================
Accuracy: 0.7495034208783933
Confusion Matrix:
[[3381    8]
 [1127   15]]
Classification Report:
              precision    recall  f1-score   support

           0       0.75      1.00      0.86      3389
           1       0.65      0.01      0.03      1142

    accuracy                           0.75      4531
   macro avg       0.70      0.51      0.44      4531
weighted avg       0.73      0.75      0.65      4531
```

Fig: Logistic Regression

```
Decision Tree Model:
Accuracy: 0.6345177664974619
Confusion Matrix:
[[2592  797]
 [ 859  283]]
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.76      0.76      3389
           1       0.26      0.25      0.25      1142

    accuracy                           0.63      4531
   macro avg       0.51      0.51      0.51      4531
weighted avg       0.63      0.63      0.63      4531
```

Fig: Decision Tree

```
Random Forest Model:
Accuracy: 0.7058044581770029
Confusion Matrix:
[[3020  369]
 [ 964  178]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.89      0.82      3389
           1       0.33      0.16      0.21      1142

    accuracy                           0.71      4531
   macro avg       0.54      0.52      0.51      4531
weighted avg       0.65      0.71      0.67      4531
```

Fig: Random Forest

```
Accuracy: 0.748179209887442
Confusion Matrix:
[[3363   26]
 [1115   27]]
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.99      0.85      3389
           1       0.51      0.02      0.05      1142

    accuracy                           0.75      4531
   macro avg       0.63      0.51      0.45      4531
weighted avg       0.69      0.75      0.65      4531
```

Fig: Gradient Boost

**Accuracy:**

- Logistic Regression: 0.7757
- Decision Tree: 0.6345
- Random Forest: 0.7058
- Gradient Boosting: 0.7482

**Precision (weighted):**

- Logistic Regression: 0.73
- Decision Tree: 0.63
- Random Forest: 0.65
- Gradient Boosting: 0.69

**Recall (weighted):**

- Logistic Regression: 0.78
- Decision Tree: 0.63
- Random Forest: 0.71
- Gradient Boosting: 0.75

**F1-score (weighted):**

- Logistic Regression: 0.68
- Decision Tree: 0.63
- Random Forest: 0.67
- Gradient Boosting: 0.65

Considering these metrics, the choice of the best model depends on the specific goals of the real estate company:

**- Logistic Regression:**

- Pros: Highest accuracy, precision, and recall.
- Cons: May not capture complex relationships.

**- Random Forest:**

- Pros: Good accuracy, suitable for handling complex relationships and interactions between features.
- Cons: Slightly lower precision and recall compared to Logistic Regression.

**- Gradient Boosting:**

- Pros: Good accuracy, precision, and recall; can handle complex relationships and

provide feature importance.

- Cons: May be computationally expensive compared to other models.

## VI. Recommendations:

- If interpretability is a priority and the goal is to have a model that is easy to understand and explain, Logistic Regression might be the preferred choice.
- If predictive accuracy is crucial, and the model needs to handle complex relationships, Gradient Boosting could be a suitable option.
- Random Forest is also a strong contender, offering a balance between interpretability and accuracy.

## VII. Considerations:

- The choice of the model should align with specific business goals and resource constraints.
- Further optimization through cross-validation and hyperparameter tuning may enhance model performance.

## VIII. Conclusion:

The analysis indicates that the choice of the best model depends on the company's priorities. Logistic Regression provides interpretability, Random Forest handles complex relationships, and Gradient Boosting offers a balanced approach. A comprehensive understanding of business objectives and trade-offs is essential for making an informed decision.

This report aims to guide the real estate company in selecting the most suitable model based on their specific needs. Further refinement and customization of models can be explored based on future requirements.