College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 4 Project - Investing in Nashville

Submitted on:

Feb 8, 2024

Submitted to:	**Submitted by:**
Prof: BEHZAD AHMADI	NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The Housing Data dataset, is a collection of housing transactions in Nashville. This dataset is valuable for real estate market analysis, predictive modeling, and urban planning. It serves as a rich source for understanding housing trends, pricing strategies, and the impact of location on property values.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set:

- **Unnamed: 0**: An index or unique identifier for each record.
- **Parcel ID**: A unique identifier for a parcel of land in the real estate registry.
- **Land Use**: The designated use of the property, such as "SINGLE FAMILY."
- **Property Address**: The street address of the property.
- **Suite/Condo #**: If applicable, the suite or condo number of the property.
- **Property City**: The city in which the property is located, here it's "NASHVILLE."
- **Sale Date**: The date on which the property was sold.
- **Legal Reference**: A reference number for the legal document of the sale.
- **Sold As Vacant**: Indicates whether the property was vacant at the time of sale ("Yes" or "No").
- **Multiple Parcels Involved in Sale**: Indicates if the sale involved multiple parcels ("Yes" or "No").
- **Owner Name**: The name of the property's owner.
- **Address**: Likely a duplication or variation of property address details.
- **Acreage**: The size of the property in acres.
- **Tax District**: The tax district in which the property is located.
- **Land Value**: The assessed value of the land.
- **Building Value**: The assessed value of the building on the property.
- **Total Value**: The total assessed value of the property (land + building).
- **Finished Area**: The total finished area of the property in square feet.
- **Foundation Type**: The type of foundation the building has (e.g., slab, crawl space).
- **Year Built**: The year in which the building was constructed.
- **Exterior Wall**: The material of the building's exterior wall (e.g., brick, frame).
- **Grade**: An assessment of the property's overall quality or grade.
- **Bedrooms**: The number of bedrooms in the property.
- **Full Bath**: The number of full bathrooms in the property.
- **Half Bath**: The number of half bathrooms in the property.
- **Sale Price Compared To Value**: A categorical variable indicating whether the sale price was over, under, or at the assessed value ("Over", "Under").

The dataset consists of 22,651 rows and 26 columns. The dataset includes a mix of data types, with 15 object (string) columns, 6 float columns, and 5 int columns.

DATA CLEANING

- **Handling missing values**
  'Suite/ Condo #' column has 22,651 missing entries, indicating the information is either consistently unavailable or not applicable. Minor missing data in 'Property Address', 'Property City', and a few other variables suggest targeted cleaning or imputation could easily remedy these gaps.

  The missing values in the 'Half Bath', 'Full Bath', 'Bedrooms', and 'Finished Area' columns were substituted with their respective column medians.

- **Dropping irrelevant columns**
  Columns 'Unnamed: 0', 'Suite/ Condo #', 'Legal Reference', 'Parcel ID', 'Property Address', and 'State' were removed because they do not contribute to the analysis or the insights.

- **Removing duplicate rows**
  Duplicate rows in the dataset can lead to inconsistencies and affect the accuracy of data analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.
  144 rows were identified as duplicate rows and they were eliminated for enhancing data quality.

- **Simplifying Categorical Data**
  The categorical columns were mapped to binary representations (0 or 1) in three columns ('Sold As Vacant', 'Multiple Parcels Involved in Sale', and 'Sale Price Compared To Value').

- **Combining columns**
  The 'Total Bath' and 'Total Value' columns were created by combining data from the from existing columns ('Half Bath' + 'Full Bath') and ('Land Value' + 'Building Value') respectively.

- **Changing datatype**
  Column ‘Sale Date’ was converted to datetime format.

EDA

Observations from the EDA:

Figure 1:

Key observations from the histograms:

- Acreage varies significantly with a minimum of 0.04 acres to a maximum of 17.5 acres, indicating a wide range of property sizes.
- Year Built ranges from 1832 to 2017, showcasing the diversity in property ages.
- Finished Area and Bedrooms suggest a variety of property sizes and capacities, from smaller homes to larger estates.

Figure 2 &3 :

- The correlation analysis indicates that there is a strong positive correlation between 'Finished Area' and 'Total Value', highlighting a significant relationship where larger areas are typically associated with higher property values.
- Variables such as 'Sold As Vacant,' 'Finished Area,' and 'Total Value' are positively correlated to 'Sale Price Compared To Value'.

LOGISTIC REGRESSION

We started with encoding the categorical values present in the data. Encoding is a crucial data preprocessing step in machine learning and statistical analysis, particularly when working with categorical variables. Without encoding, models may not consider categorical data, leading to errors. This process also ensures consistency in representing categories and plays a vital role in improving model performance by allowing models to learn from categorical data.

Then, the dataset is split into two parts: one containing the features and the other with the target variable, excluding specific columns ('Sale Price Compared To Value', 'Sale Date', 'City', 'Property City') from the features.
Subsequently, the dataset is further divided into training and testing sets using a 80-20 split ratio.

A **logistic regression model (Figure 4)** is constructed. The model's accuracy is calculated, and a confusion matrix and classification report are generated to assess its performance in detail.

The **Logit** function (Figure 5) is then applied to the data, and a summary of the logistic regression analysis is produced, providing coefficients, standard errors, z-values, p-values, and confidence intervals for each predictor. This summary helps to understand the significance of each predictor and the model's overall fit to the data.

Accuracy : The model attained an accuracy rate of 75.6%, indicating it correctly classified about 75.6% of the instances.

Confusion Matrix:

- True Positives (TP): Only 6 instances of accurately identified cases where housing prices are inaccurately represented.
- True Negatives (TN): High number of instances correctly identified as accurately represented prices (3396).
- False Positives (FP): High number of instances incorrectly classified as inaccurately represented prices (1098).
- False Negatives (FN): Only 1 instance incorrectly classified as accurately represented price.

Features such as 'Sold As Vacant' and 'Multiple Parcels Involved in Sale' appear to be significant drivers of housing prices based on their coefficients.

DECISION TREE

Accuracy: The accuracy of the current model is 76% (Figure 6).

Confusion Matrix:

- True Positives (TP): The model correctly identifies 32 instances of inaccurately represented prices.
- False Positives (FP): There are 1072 false positives, indicating instances where the model incorrectly predicts inaccurately represented prices.
- True Negatives (TN): The model correctly identifies 3388 instances of accurately represented prices.
- False Negatives (FN): Only 9 instances are incorrectly classified as accurately represented prices.

Both models have a similar accuracy, with the decision tree model achieving a slightly higher accuracy. The decision tree model identifies more true positives and fewer false negatives compared to the logistic regression model. Therefore, the decision tree model appears to be more effective in this scenario.

This decision tree model would be used to predict whether a data sample belongs to class 0 or class 1 based on features such as 'Total Bath', 'Foundation Type', 'Land Use', and 'Bedrooms'.

The decision tree (Figure 7) provides a clear and interpretable model, showing the decision rules derived from the training data. This is useful for understanding feature importance and the decision-making process in classification tasks. The first split is made based on 'Land Use', dividing the dataset into two major branches.

RANDOM FOREST

Accuracy: The accuracy of the Random Forest model is 72.3% (Figure 8).

Confusion Matrix:

- True Positives (TP): The model correctly identifies 164 instances of inaccurately represented prices.
- False Positives (FP): There are 307 false positives, indicating instances where the model incorrectly predicts inaccurately represented prices.
- True Negatives (TN): The model correctly identifies 3090 instances of accurately represented prices.
- False Negatives (FN): There are 940 false negatives, indicating instances where the model incorrectly classifies accurately represented prices.

The random forest model has a lower accuracy and significantly higher F1 score compared to the logistic regression and decision tree models. The random forest model has a higher number of false negatives and false positives compared to the logistic regression and decision tree models. It also has a lower number of true positives and true negatives.

GRADIENT BOOST

Accuracy: The Gradient Boost model achieves an accuracy of 76.1%, indicating that it correctly predicts housing prices in approximately 76.1% of cases (Figure 9).

Confusion matrix:

- True Positives (TP): The model correctly predicts 49 instances as inaccurately represented prices.
- False Positives (FP): There are 21 instances where the model incorrectly predicts inaccurately represented prices.
- True Negatives (TN): The model correctly predicts 3376 instances as accurately represented prices.
- False Negatives (FN): There are 1055 instances where the model incorrectly predicts accurately represented prices.

The Gradient Boost model outperforms all the previous models but has a lower F1 score compared to the random forest model.

CONCLUSION

From this comparison (Figure 10 & 11), we can observe that Gradient Boost has the highest accuracy among the models, followed closely by Decision Tree. However, Random Forest achieves the highest F1 score, indicating better balance between precision and recall. Gradient Boost also shows a relatively higher F1 score compared to Logistic Regression and Decision Tree, indicating better performance in capturing both precision and recall.

The Logistic Regression model is simple and interpretable, making it easier to understand the relationships between features and housing prices.

REFERENCES

Saini, A. (2024, January 10). *Gradient Boosting Algorithm: A complete guide for beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

APPENDIX

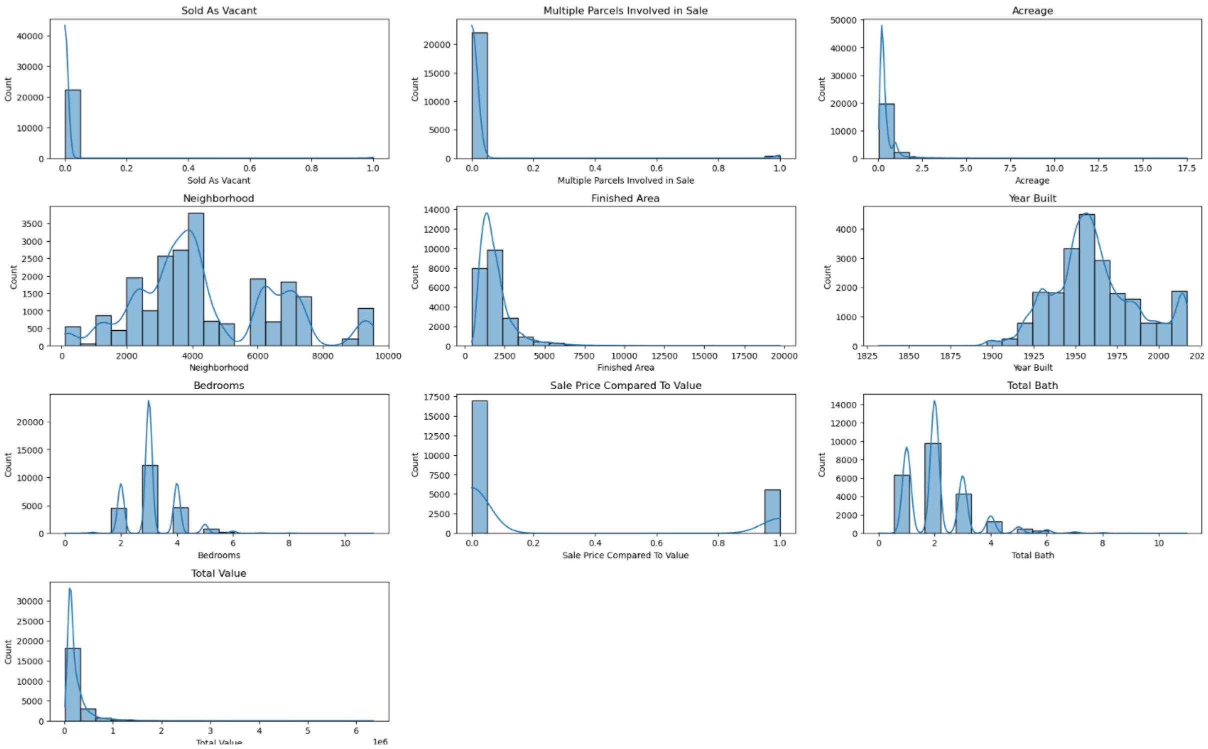
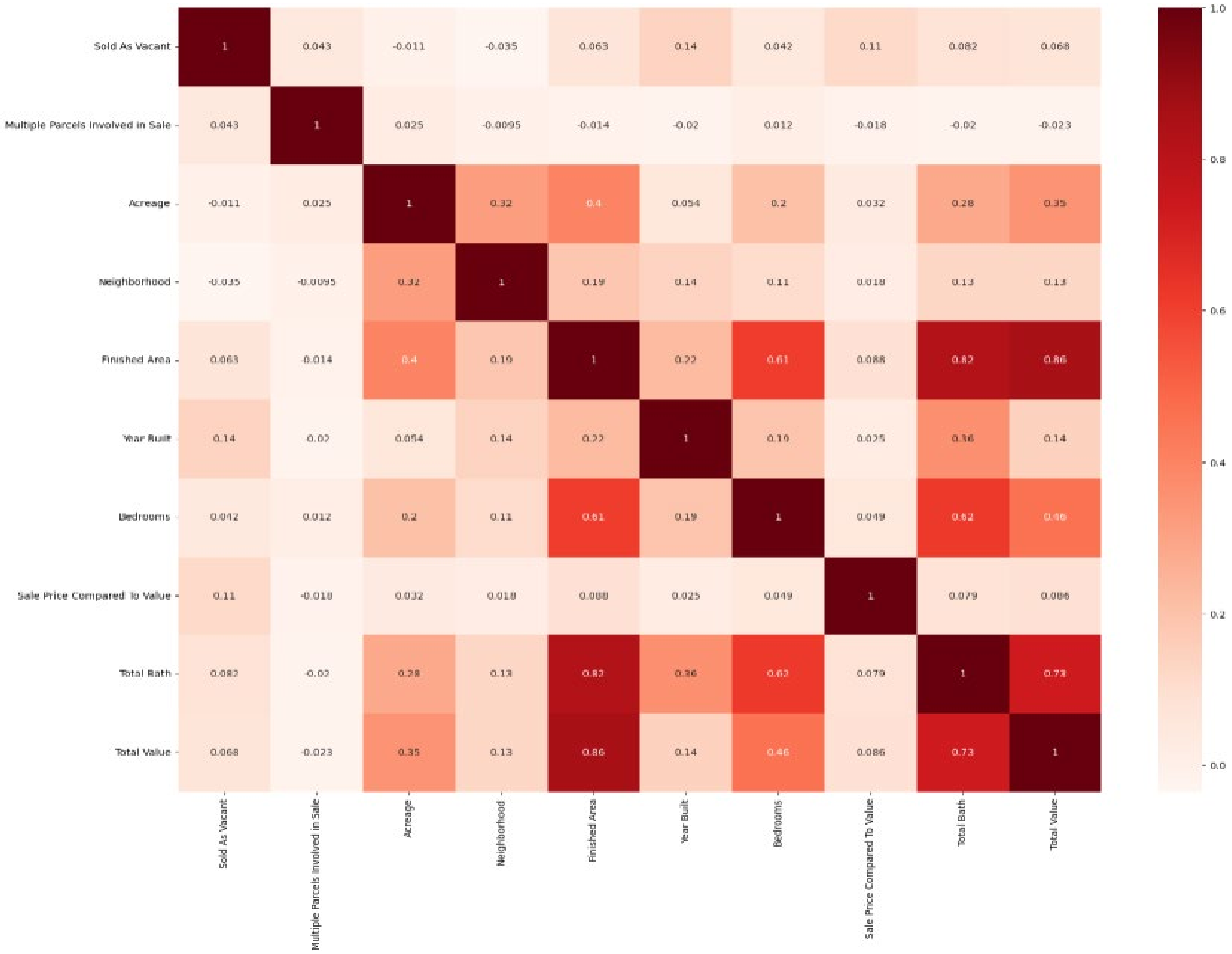
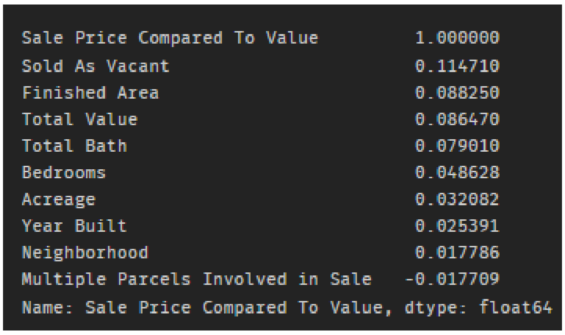


Figure 1 – Histograms for understanding distribution



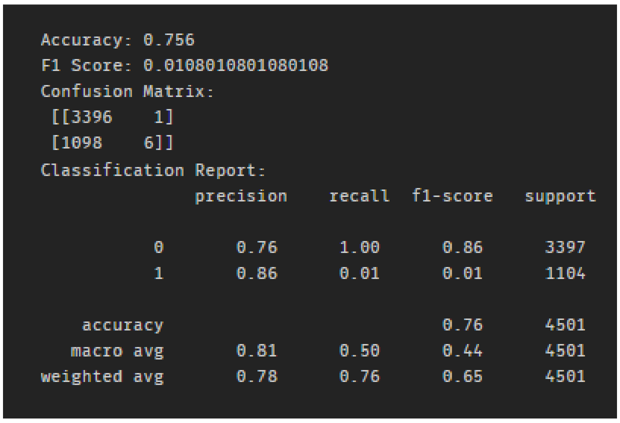
	Sold As Vacant	Multiple Parcels Involved in Sale	Acreage	Neighborhood	Finished Area	Year Built	Bedrooms	Sale Price Compared To Value	Total Bath	Total Value
Sold As Vacant	1	0.043	-0.011	-0.035	0.063	0.14	0.042	0.11	0.082	0.068
Multiple Parcels Involved in Sale	0.043	1	0.025	-0.0095	-0.014	-0.02	0.012	-0.018	-0.02	-0.023
Acreage	-0.011	0.025	1	0.32	0.4	0.054	0.2	0.032	0.28	0.35
Neighborhood	-0.035	-0.0095	0.32	1	0.19	0.14	0.11	0.018	0.13	0.13
Finished Area	0.063	-0.014	0.4	0.19	1	0.22	0.61	0.088	0.82	0.86
Year Built	0.14	0.02	0.054	0.14	0.22	1	0.19	0.025	0.36	0.14
Bedrooms	0.042	0.012	0.2	0.11	0.61	0.19	1	0.049	0.62	0.46
Sale Price Compared To Value	0.11	-0.018	0.032	0.018	0.088	0.025	0.049	1	0.079	0.086
Total Bath	0.082	-0.02	0.28	0.13	0.82	0.36	0.62	0.079	1	0.73
Total Value	0.068	-0.023	0.35	0.13	0.86	0.14	0.46	0.086	0.73	1

Figure 2 – Correlation matrix



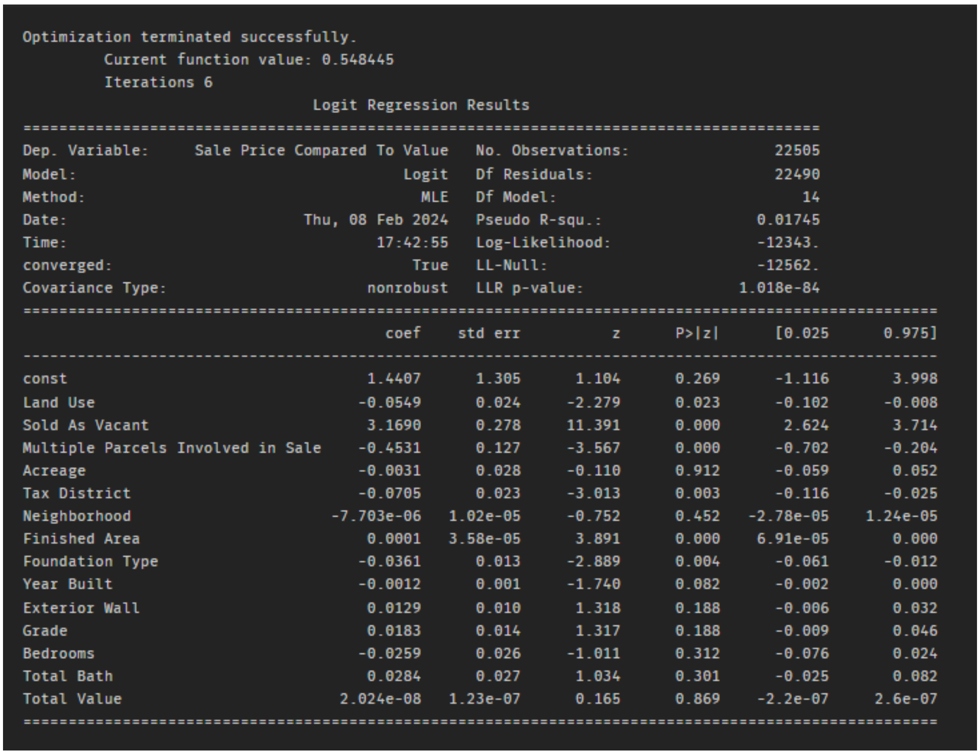
```
Sale Price Compared To Value        1.000000
Sold As Vacant                      0.114710
Finished Area                       0.088250
Total Value                         0.086470
Total Bath                          0.079010
Bedrooms                            0.048628
Acreage                             0.032082
Year Built                          0.025391
Neighborhood                        0.017786
Multiple Parcels Involved in Sale  -0.017709
Name: Sale Price Compared To Value, dtype: float64
```

Figure 3 – Correlation of other features with target variable



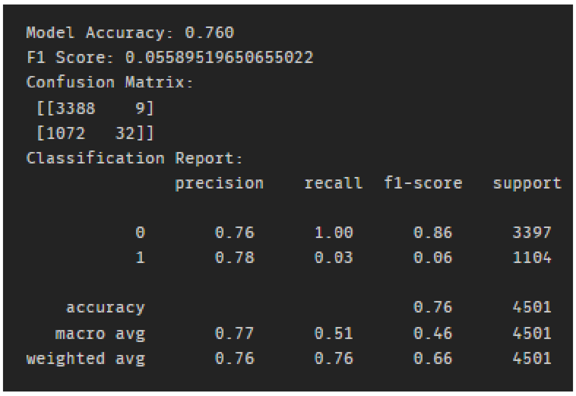
```
Accuracy: 0.756
F1 Score: 0.0108010801080108
Confusion Matrix:
 [[3396    1]
 [1098    6]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      1.00      0.86      3397
           1       0.86      0.01      0.01      1104

    accuracy                           0.76      4501
   macro avg       0.81      0.50      0.44      4501
weighted avg       0.78      0.76      0.65      4501
```

Figure 4 – Logistic Regression Model



```
Optimization terminated successfully.
         Current function value: 0.548445
         Iterations 6
                                Logit Regression Results
========================================================================================
Dep. Variable:     Sale Price Compared To Value   No. Observations:                22505
Model:                                    Logit   Df Residuals:                    22490
Method:                                     MLE   Df Model:                           14
Date:                          Thu, 08 Feb 2024   Pseudo R-squ.:                 0.01745
Time:                                  17:42:55   Log-Likelihood:                -12343.
converged:                                 True   LL-Null:                       -12562.
Covariance Type:                      nonrobust   LLR p-value:                 1.018e-84
=====================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                 1.4407      1.305      1.104      0.269      -1.116       3.998
Land Use                             -0.0549      0.024     -2.279      0.023      -0.102      -0.008
Sold As Vacant                        3.1690      0.278     11.391      0.000       2.624       3.714
Multiple Parcels Involved in Sale    -0.4531      0.127     -3.567      0.000      -0.702      -0.204
Acreage                              -0.0031      0.028     -0.110      0.912      -0.059       0.052
Tax District                         -0.0705      0.023     -3.013      0.003      -0.116      -0.025
Neighborhood                      -7.703e-06   1.02e-05     -0.752      0.452   -2.78e-05    1.24e-05
Finished Area                         0.0001   3.58e-05      3.891      0.000    6.91e-05       0.000
Foundation Type                      -0.0361      0.013     -2.889      0.004      -0.061      -0.012
Year Built                           -0.0012      0.001     -1.740      0.082      -0.002       0.000
Exterior Wall                         0.0129      0.010      1.318      0.188      -0.006       0.032
Grade                                 0.0183      0.014      1.317      0.188      -0.009       0.046
Bedrooms                             -0.0259      0.026     -1.011      0.312      -0.076       0.024
Total Bath                            0.0284      0.027      1.034      0.301      -0.025       0.082
Total Value                        2.024e-08   1.23e-07      0.165      0.869    -2.2e-07     2.6e-07
=====================================================================================================
```

Figure 5 – Logistic model result summary (statsmodels)



```
Model Accuracy: 0.760
F1 Score: 0.05589519650655022
Confusion Matrix:
 [[3388    9]
 [1072   32]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      1.00      0.86      3397
           1       0.78      0.03      0.06      1104

    accuracy                           0.76      4501
   macro avg       0.77      0.51      0.46      4501
weighted avg       0.76      0.76      0.66      4501
```

Figure 6 – Decision tree Model

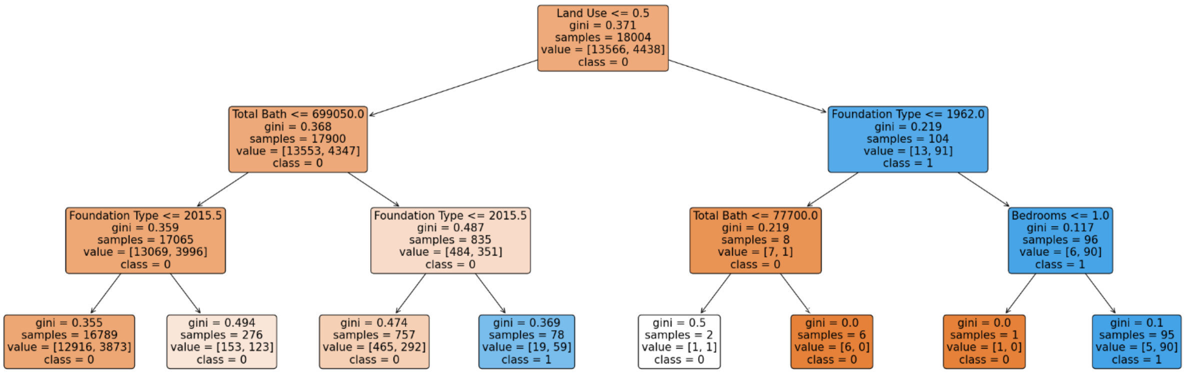
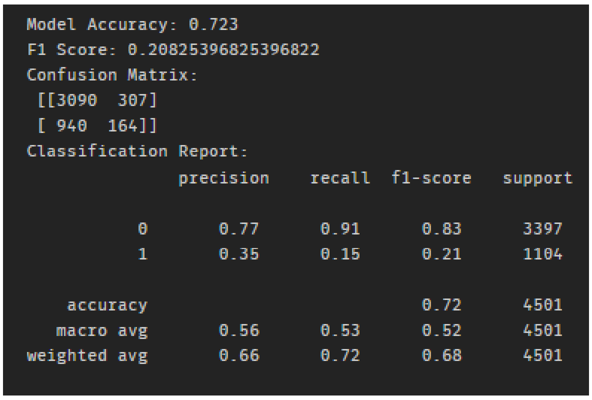
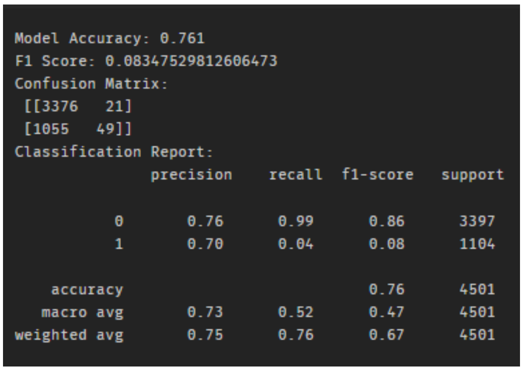

Land Use <= 0.5 gini = 0.371 samples = 18004 value = [13566, 4438] class = 0
Total Bath <= 699050.0 gini = 0.368 samples = 17900 value = [13553, 4347] class = 0
Foundation Type <= 1962.0 gini = 0.219 samples = 104 value = [13, 91] class = 1
Foundation Type <= 2015.5 gini = 0.359 samples = 17065 value = [13069, 3996] class = 0
Foundation Type <= 2015.5 gini = 0.487 samples = 835 value = [484, 351] class = 0
Total Bath <= 77700.0 gini = 0.219 samples = 8 value = [7, 1] class = 0
Bedrooms <= 1.0 gini = 0.117 samples = 96 value = [6, 90] class = 1
gini = 0.355 samples = 16789 value = [12916, 3873] class = 0
gini = 0.494 samples = 276 value = [153, 123] class = 0
gini = 0.474 samples = 757 value = [465, 292] class = 0
gini = 0.369 samples = 78 value = [19, 59] class = 1
gini = 0.5 samples = 2 value = [1, 1] class = 0
gini = 0.0 samples = 6 value = [6, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.1 samples = 95 value = [5, 90] class = 1

Figure 7 – Decision tree



```
Model Accuracy: 0.723
F1 Score: 0.20825396825396822
Confusion Matrix:
 [[3090  307]
 [ 940  164]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.91      0.83      3397
           1       0.35      0.15      0.21      1104

    accuracy                           0.72      4501
   macro avg       0.56      0.53      0.52      4501
weighted avg       0.66      0.72      0.68      4501
```

Figure 8 – Random forest Model



```
Model Accuracy: 0.761
F1 Score: 0.08347529812606473
Confusion Matrix:
 [[3376   21]
 [1055   49]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.99      0.86      3397
           1       0.70      0.04      0.08      1104

    accuracy                           0.76      4501
   macro avg       0.73      0.52      0.47      4501
weighted avg       0.75      0.76      0.67      4501
```

Figure 9 – Gradient boost Model

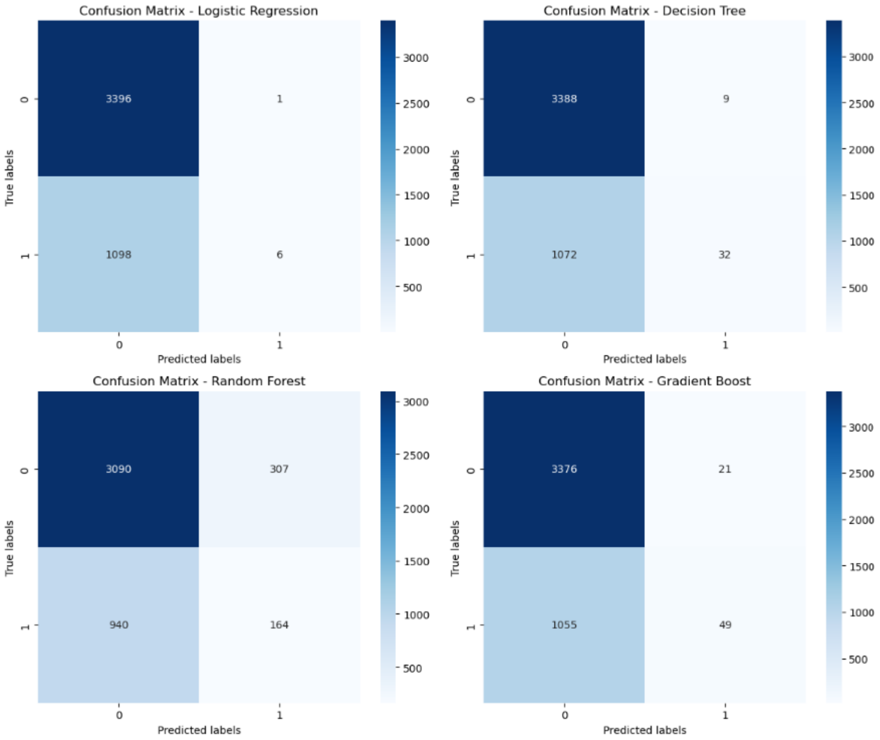


Figure 10 – Confusion matrix for all models

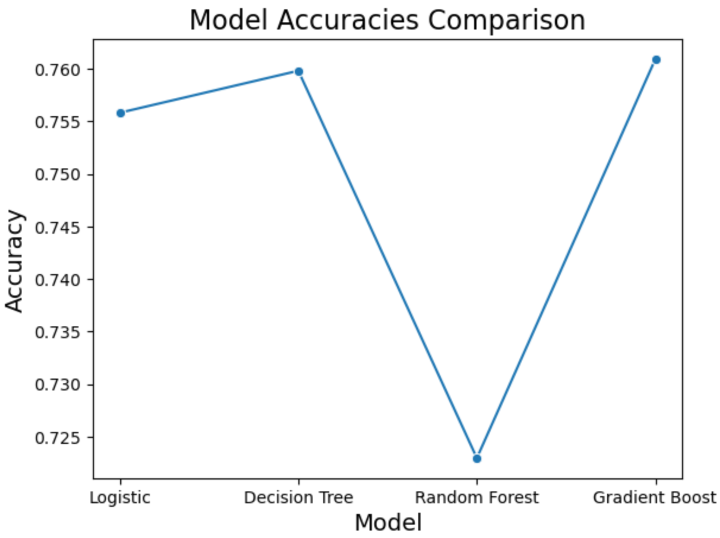


Figure 11 – Graph comparing accuracies of models