Real Estate Sales Price Prediction Project

Predictive Modeling for Enhanced Real Estate Valuation

Presented By: Donn Bryan Julian
Date: 12/05/2024

Project Overview

Goal:

- Predict real estate sales prices using historical data.

Scope:

- Focused on properties from 2001 to 2022, using a variety of statistical and machine learning methods to develop an accurate prediction model.

Key Takeaway:

- The best model has an R-squared of 0.949, demonstrating strong predictive capabilities.

Source Data

- **Data Source:**

Acquired from https://catalog.data.gov/dataset/real-estate-sales-2001-2018 covering real estate sales between 2001-2022.

- **Details:**

Over 1 million records with features like assessed value, property type, town, and sale amount.

Exploratory Data Analysis (EDA) Overview

- **Objective:**

Understand data distribution, identify outliers, and recognize relationships.

- **Initial Insights:**

Data contains multiple types, non-uniform missing values, and varying distribution across categorical features.

Data Cleansing

Actions Taken:

- Removed towns with fewer than 500 entries to eliminate noise.
- Dropped non-numeric fields like 'Address' and 'Date Recorded' that didn't contribute to prediction.

Outcome:

- Focused on significant features to enhance model performance.

Statistical Tests Conducted

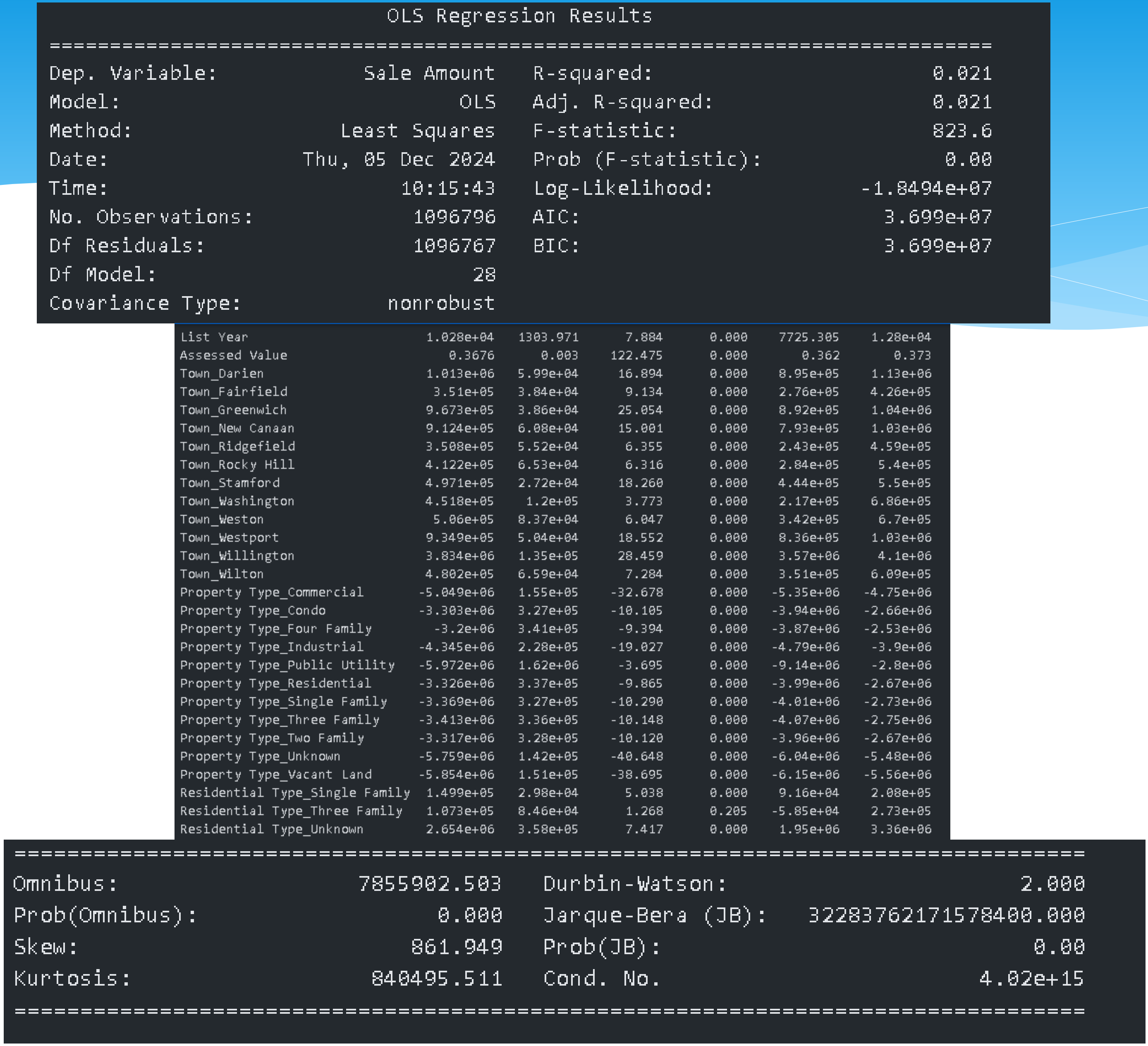
Multicollinearity (VIF Analysis):

Identified and removed highly correlated features to reduce redundancy.

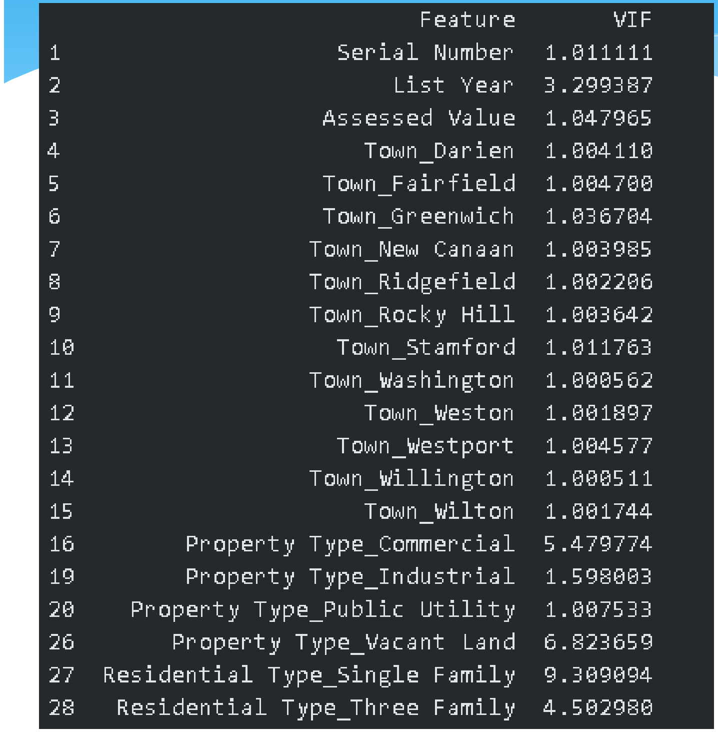
Autocorrelation: Performed Ljung-Box test to verify residual independence.

Heteroskedasticity: Breusch-Pagan test confirmed minimal variance issues.

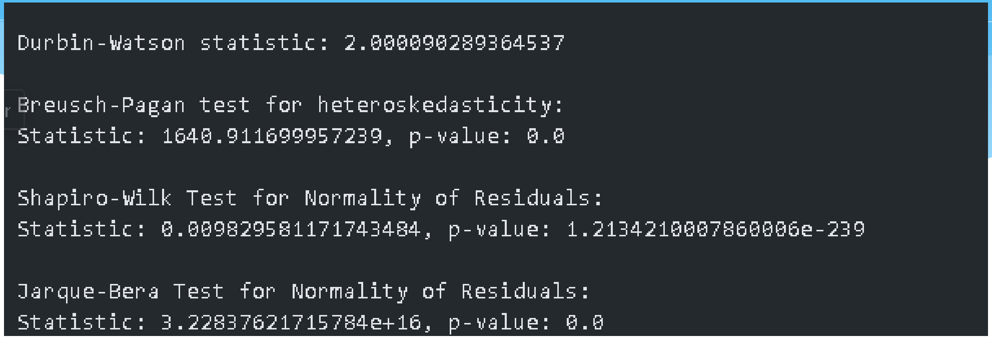
Normality: Applied Lilliefors and Shapiro-Wilk tests; deviations from normality informed the selection of ensemble models.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Sale Amount   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.021
Method:                 Least Squares   F-statistic:                     823.6
Date:                Thu, 05 Dec 2024   Prob (F-statistic):               0.00
Time:                        10:15:43   Log-Likelihood:            -1.8494e+07
No. Observations:             1096796   AIC:                         3.699e+07
Df Residuals:                 1096767   BIC:                         3.699e+07
Df Model:                          28
Covariance Type:            nonrobust
List Year                    1.028e+04   1303.971      7.884      0.000    7725.305    1.28e+04
Assessed Value                  0.3676      0.003    122.475      0.000       0.362       0.373
Town_Darien                  1.013e+06   5.99e+04     16.894      0.000    8.95e+05    1.13e+06
Town_Fairfield                3.51e+05   3.84e+04      9.134      0.000    2.76e+05    4.26e+05
Town_Greenwich               9.673e+05   3.86e+04     25.054      0.000    8.92e+05    1.04e+06
Town_New Canaan              9.124e+05   6.08e+04     15.001      0.000    7.93e+05    1.03e+06
Town_Ridgefield              3.508e+05   5.52e+04      6.355      0.000    2.43e+05    4.59e+05
Town_Rocky Hill              4.122e+05   6.53e+04      6.316      0.000    2.84e+05     5.4e+05
Town_Stamford                4.971e+05   2.72e+04     18.260      0.000    4.44e+05     5.5e+05
Town_Washington              4.518e+05    1.2e+05      3.773      0.000    2.17e+05    6.86e+05
Town_Weston                   5.06e+05   8.37e+04      6.047      0.000    3.42e+05     6.7e+05
Town_Westport                9.349e+05   5.04e+04     18.552      0.000    8.36e+05    1.03e+06
Town_Willington              3.834e+06   1.35e+05     28.459      0.000    3.57e+06     4.1e+06
Town_Wilton                  4.802e+05   6.59e+04      7.284      0.000    3.51e+05    6.09e+05
Property Type_Commercial    -5.049e+06   1.55e+05    -32.678      0.000   -5.35e+06   -4.75e+06
Property Type_Condo         -3.303e+06   3.27e+05    -10.105      0.000   -3.94e+06   -2.66e+06
Property Type_Four Family     -3.2e+06   3.41e+05     -9.394      0.000   -3.87e+06   -2.53e+06
Property Type_Industrial    -4.345e+06   2.28e+05    -19.027      0.000   -4.79e+06    -3.9e+06
Property Type_Public Utility -5.972e+06  1.62e+06     -3.695      0.000   -9.14e+06    -2.8e+06
Property Type_Residential   -3.326e+06   3.37e+05     -9.865      0.000   -3.99e+06   -2.67e+06
Property Type_Single Family -3.369e+06   3.27e+05    -10.290      0.000   -4.01e+06   -2.73e+06
Property Type_Three Family  -3.413e+06   3.36e+05    -10.148      0.000   -4.07e+06   -2.75e+06
Property Type_Two Family    -3.317e+06   3.28e+05    -10.120      0.000   -3.96e+06   -2.67e+06
Property Type_Unknown       -5.759e+06   1.42e+05    -40.648      0.000   -6.04e+06   -5.48e+06
Property Type_Vacant Land   -5.854e+06   1.51e+05    -38.695      0.000   -6.15e+06   -5.56e+06
Residential Type_Single Family 1.499e+05  2.98e+04     5.038      0.000    9.16e+04    2.08e+05
Residential Type_Three Family 1.073e+05   8.46e+04     1.268      0.205   -5.85e+04    2.73e+05
Residential Type_Unknown     2.654e+06   3.58e+05      7.417      0.000    1.95e+06    3.36e+06
==============================================================================
Omnibus:                  7855902.503   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   32283762171578400.000
Skew:                         861.949   Prob(JB):                         0.00
Kurtosis:                  840495.511   Cond. No.                     4.02e+15
==============================================================================
```



```
                               Feature       VIF
1                        Serial Number  1.011111
2                            List Year  3.299387
3                       Assessed Value  1.047965
4                          Town_Darien  1.004110
5                       Town_Fairfield  1.004700
6                       Town_Greenwich  1.036704
7                      Town_New Canaan  1.003985
8                      Town_Ridgefield  1.002206
9                      Town_Rocky Hill  1.003642
10                       Town_Stamford  1.011763
11                     Town_Washington  1.000562
12                         Town_Weston  1.001897
13                       Town_Westport  1.004577
14                     Town_Willington  1.000511
15                         Town_Wilton  1.001744
16            Property Type_Commercial  5.479774
19            Property Type_Industrial  1.598003
20        Property Type_Public Utility  1.007533
26           Property Type_Vacant Land  6.823659
27      Residential Type_Single Family  9.309094
28       Residential Type_Three Family  4.502980
```

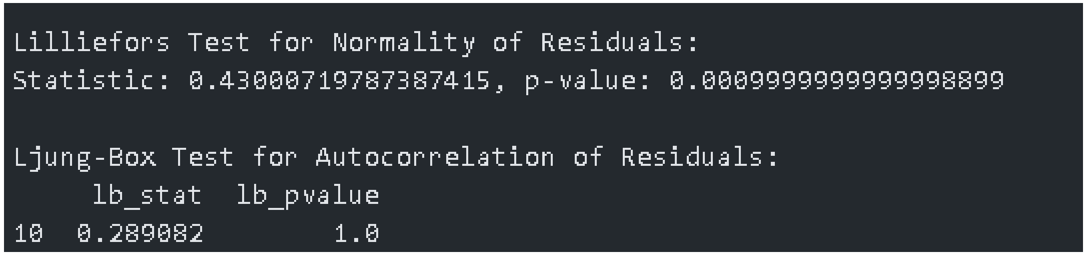


```
Durbin-Watson statistic: 2.000090289364537

Breusch-Pagan test for heteroskedasticity:
Statistic: 1640.911699957239, p-value: 0.0

Shapiro-Wilk Test for Normality of Residuals:
Statistic: 0.009829581171743484, p-value: 1.2134210007860006e-239

Jarque-Bera Test for Normality of Residuals:
Statistic: 3.22837621715784e+16, p-value: 0.0
```



```
Lilliefors Test for Normality of Residuals:
Statistic: 0.43000719787387415, p-value: 0.0009999999999998899

Ljung-Box Test for Autocorrelation of Residuals:
     lb_stat  lb_pvalue
10  0.289082        1.0
```

Feature Selection Techniques

- **Variance Thresholding:**

Removed features with low variance that provided little predictive information.

- **ANOVA F-test:**

Dropped statistically insignificant features.

- **Outcome:**

Reduced dimensionality to retain the most impactful predictors.

Data Preparation

One-Hot Encoding:

- Converted categorical features like 'Town' to dummy variables.

Sampling:

- Used 10% of the dataset to speed up the model training.

Final Dataset Size:

- Over 100,000 records used for modeling.

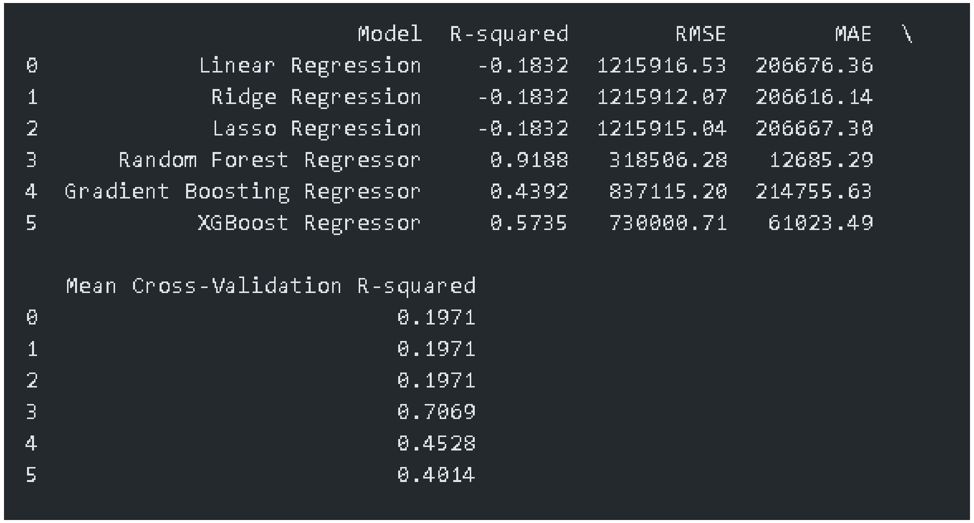
Model Testing and Selection

Models Tested:

- Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost.

Selection Criteria:

- R-squared, RMSE, and MAE for performance evaluation.



```
                         Model  R-squared        RMSE        MAE  \
0            Linear Regression    -0.1832  1215916.53  206676.36   
1             Ridge Regression    -0.1832  1215912.07  206616.14   
2             Lasso Regression    -0.1832  1215915.04  206667.30   
3      Random Forest Regressor     0.9188   318506.28   12685.29   
4  Gradient Boosting Regressor     0.4392   837115.20  214755.63   
5            XGBoost Regressor     0.5735   730000.71   61023.49   

   Mean Cross-Validation R-squared  
0                           0.1971  
1                           0.1971  
2                           0.1971  
3                           0.7069  
4                           0.4528  
5                           0.4014  
```

Model Performance Overview

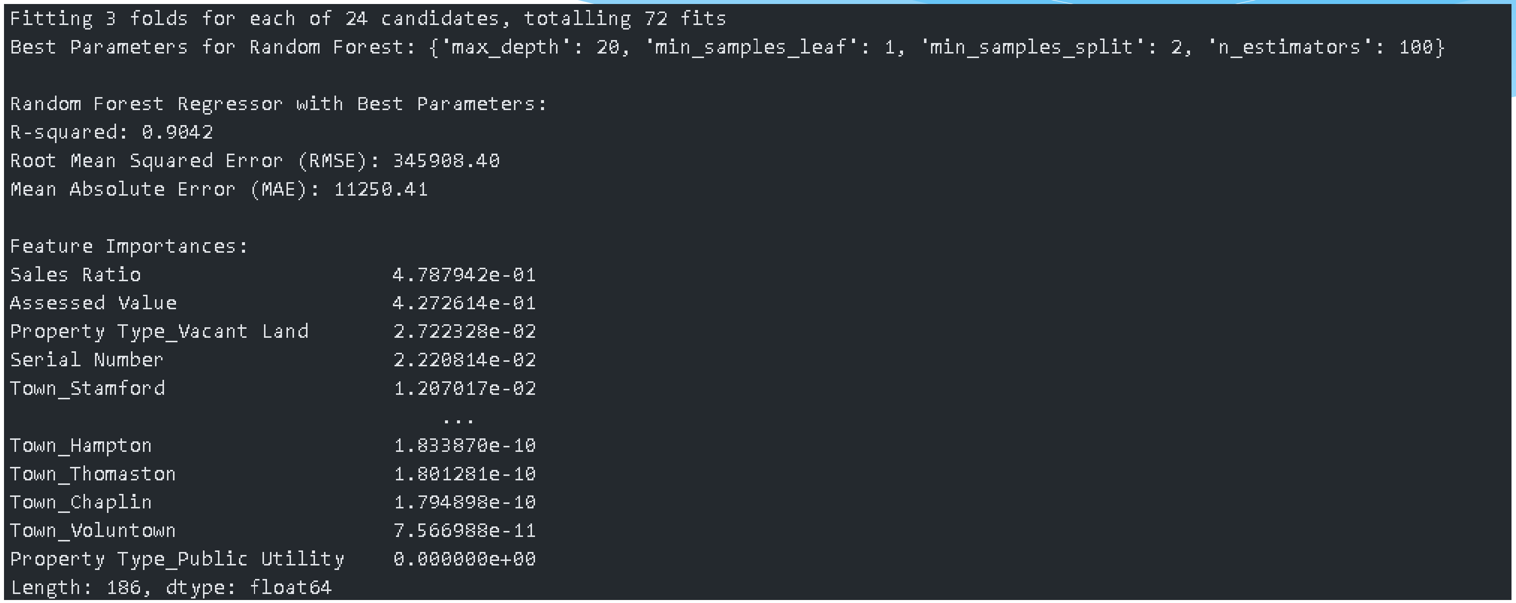
Random Forest Results:R-squared: 0.9042

Root Mean Squared Error (RMSE): 345,908

Mean Absolute Error (MAE): 11,250

Insight: Random Forest was the best-performing model, indicating strong predictive power.

Random Forest Model - Detailed Evaluation



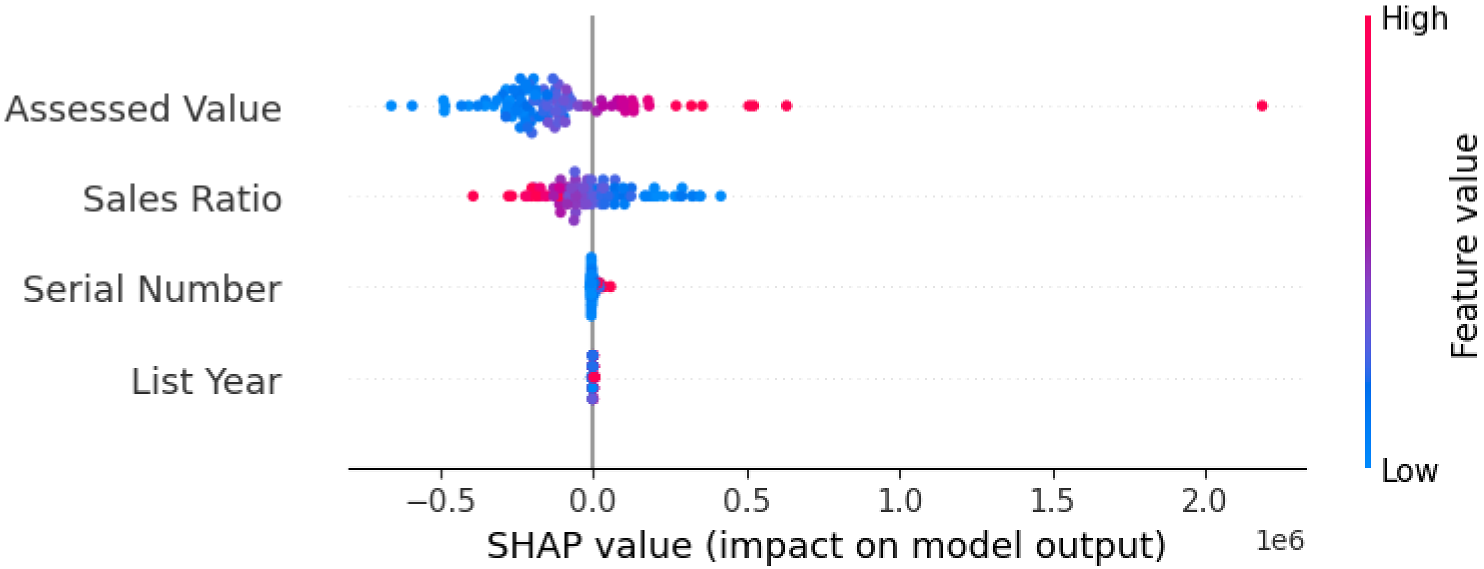
```
Fitting 3 folds for each of 24 candidates, totalling 72 fits
Best Parameters for Random Forest: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}

Random Forest Regressor with Best Parameters:
R-squared: 0.9042
Root Mean Squared Error (RMSE): 345908.40
Mean Absolute Error (MAE): 11250.41

Feature Importances:
Sales Ratio                     4.787942e-01
Assessed Value                  4.272614e-01
Property Type_Vacant Land       2.722328e-02
Serial Number                   2.220814e-02
Town_Stamford                   1.207017e-02
                                    ...
Town_Hampton                    1.833870e-10
Town_Thomaston                  1.801281e-10
Town_Chaplin                    1.794898e-10
Town_Voluntown                  7.566988e-11
Property Type_Public Utility    0.000000e+00
Length: 186, dtype: float64
```

- **Cross-Validation:** Average R-squared of 0.788 across 5-fold validation.
- **Residual Analysis:** Confirmed no visible pattern in residuals—indicative of a well-fitting model.

Understanding Predictions with SHAP



- **SHAP Analysis:**

Key features driving model predictions were Assessed Value and Sales Ratio.

- **Interpretation:**

SHAP values help visualize each feature’s contribution—red bars indicate higher values pushing the prediction up, while blue reduces it.

This provides transparency into model decision-making, aligning predictions with real-world expectations.

Deployment Preparation

Model Packaging:

- Saved the model as a .pkl file.

Deployment Plan:

- Created a REST API using Flask for real-time predictions.
- Future deployment on Hugging Face for broader accessibility.

Business Insights & Next Steps

Insights Gained:

- Key predictors like Assessed Value can be leveraged for targeted marketing and strategic pricing.

Next Steps:

- Deploy the model on Hugging Face.
- Monitor performance and plan for retraining.