United States Real Estate House Price Forecast

For Single Families condo/coops and all homes with 1, 2, 3, 4 and 5+ bedrooms

Agenda

1 Project Overview

2 Dataset and Preprocessing

3 EDA Key Insights

4 Model comparison and interpretation

5 Product Demo

6 Takeaways and Conclusions

Project Overview

About the project

Problem Statement

The housing market is highly volatile, making it challenging for homeowners and real estate agents to accurately predict house prices for large purchase/ investments

Proposed Solution

Developed a ML model that uses historical housing data to forecast future house prices.

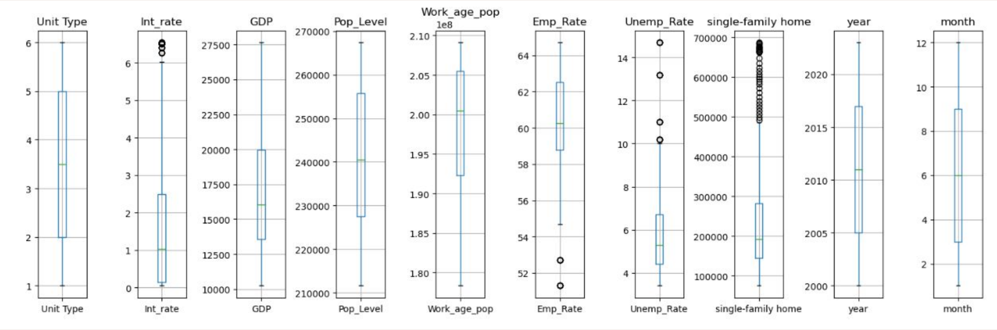
Potential Impact

Revolutionize the real estate industry by providing more accurate and reliable house price predictions. This can help homeowners make informed decisions about selling/buying their properties and assist real estate agents in setting competitive listing prices.

Dataset and Preprocessing

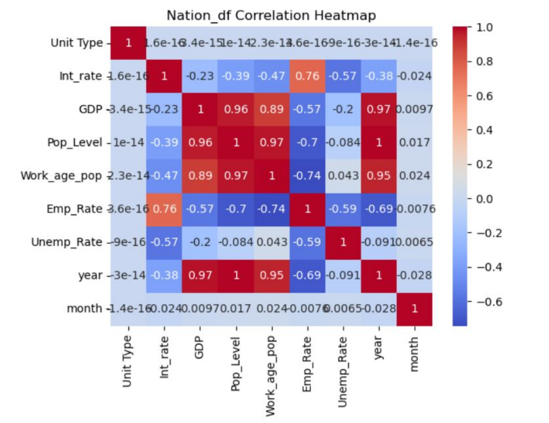
Raw Data	Dataset	Preprocessing	Outcome	Ready for EDA
	1. 523 columns for US cities housing average cost 2000-2023 2. Structured data 3. Other variables: Interest rate, Pop level, Working age population, Employment rate, unemployment rate, GDP	1. Stacked all house pricing in one column parallel to its location 2. Dealt with null values (31 location kept) 3. Separated date in Year and Month 4. Encoded Unit type	1. Two data frames: Nation and Cities 2. Nation_df: Only US data 3. Cities_df: Cities data and encoded locations	

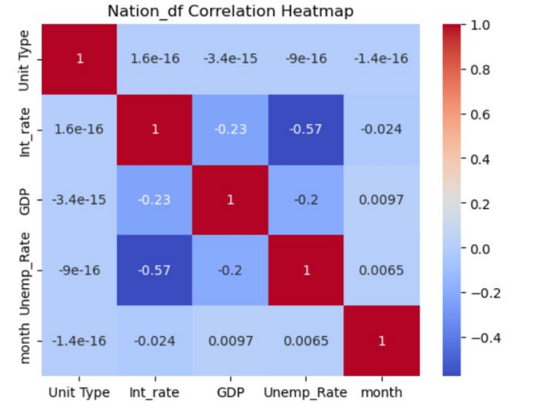
EDA- Nation Key Insights



Nation Box Plot Takeaway:

Most parameters did not have outliers and have skewed distributions. Most outliers present in home prices

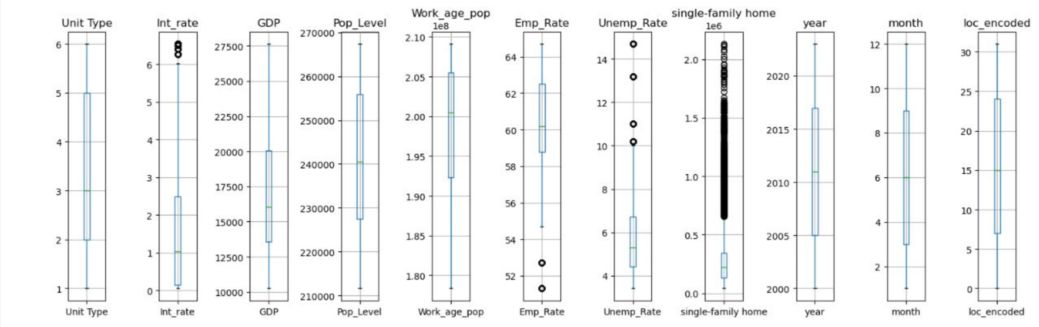




Correlation Takeaway:

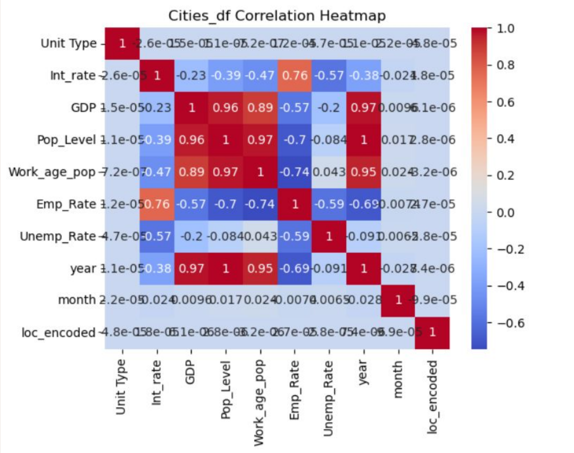
4 variables were removed due to high correlation to avoid multicollinearity

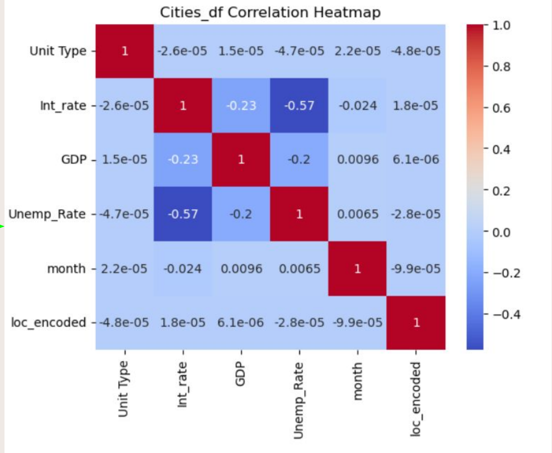
EDA- Cities Key Insights



Nation Box Plot Takeaway:

Most parameters did not have outliers and have skewed distributions. Most outliers present in home prices. The outliers are more noticeable in cities than nation





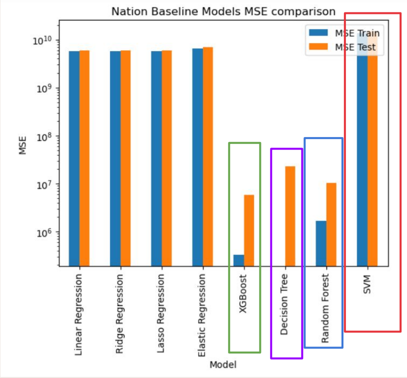
Correlation Takeaway:

Similar as with nations, 4 variables were removed due to high correlation to avoid multicollinearity

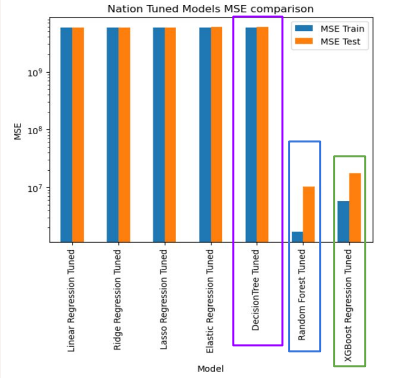
Model comparison and interpretation: MSE

Nation

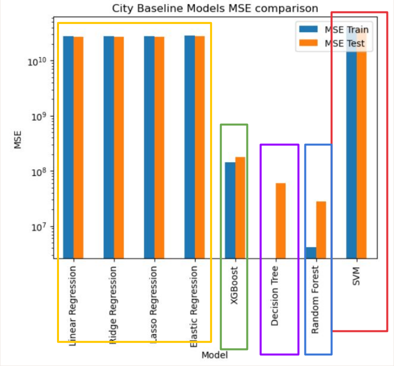
Baseline Model

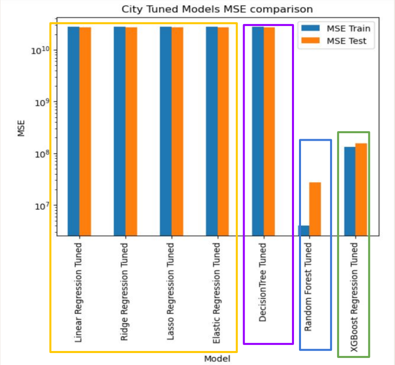


Tuned Model



Cities





Observations:

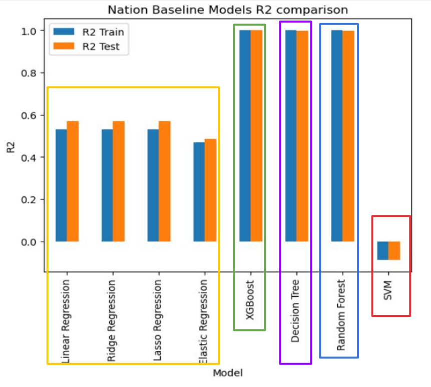
1. SVM worst MSE (largest value)
2. Liner, Ridge, Lasso and Elastic Regression large MSE even after tuning
3. Decision tree overfitting in baseline. Large MSE after tuning.
4. Random forest overfits after tuning in nations, not significant change in cities
5. XGBoost performs the best in terms of not overfitting and low MSE

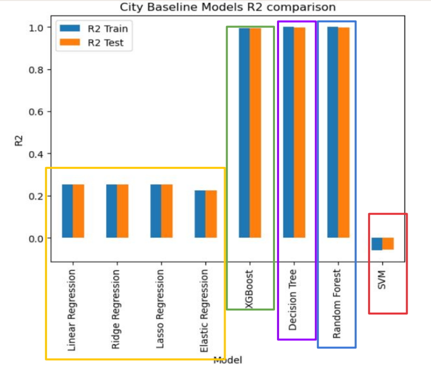
Model comparison and interpretation: R2

Nation

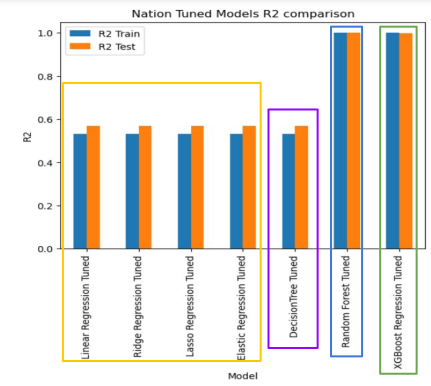
Cities

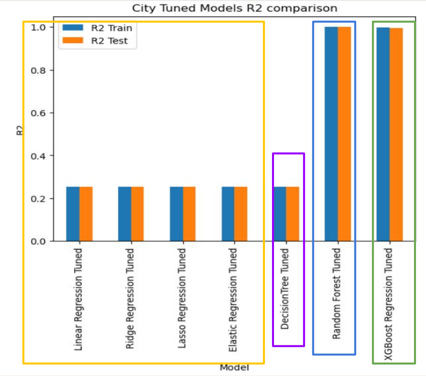
Baseline Model


Nation Baseline Models R2 comparison


City Baseline Models R2 comparison

Tuned Model


Nation Tuned Models R2 comparison


City Tuned Models R2 comparison

Observations:

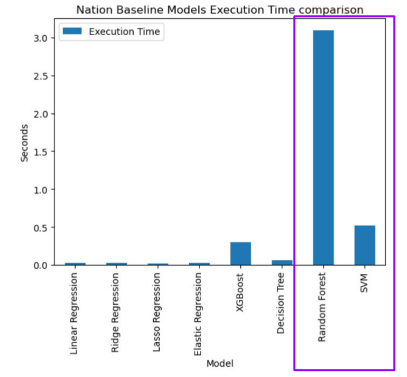
1. SVM worst R2 (smallest value)
2. Liner, Ridge, Lasso and Elastic Regression ok train to test R2 ratio but low R2 specially in cities
3. Decision tree overfitting in baseline. R2 close to 1 before tuning.
4. Random forest and XGBoost performs the best in terms of R2, very close to one even after tuning

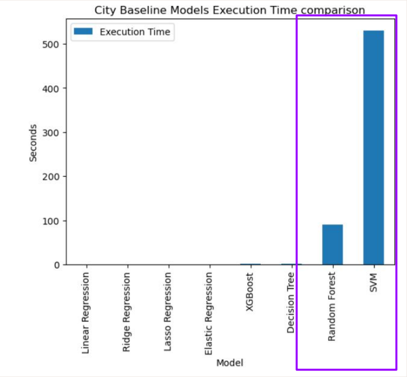
Model comparison and interpretation: Exec.Time

Nation

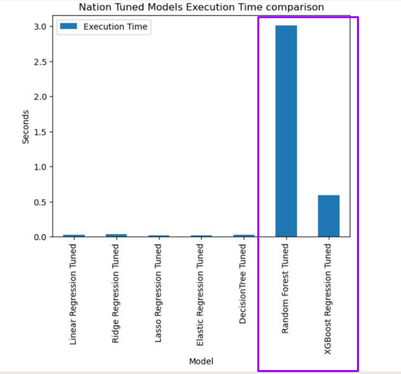
Cities

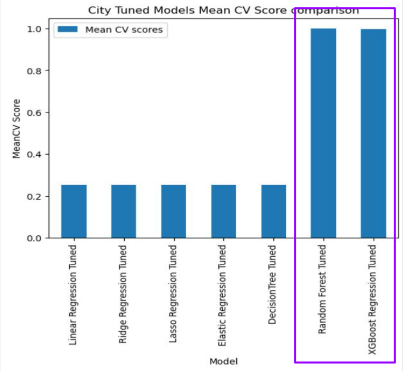
Baseline Model





Tuned Model





Observations:

1. The baseline models that take the longest are SVM and Random Forest. SVM was removed for tuning.
2. The tuned models that took the longest were Random Forest and XGBoost

Model comparison and interpretation: Takeaway

1. SVM was the worst performing model for this dataset. It was removed before tuning.

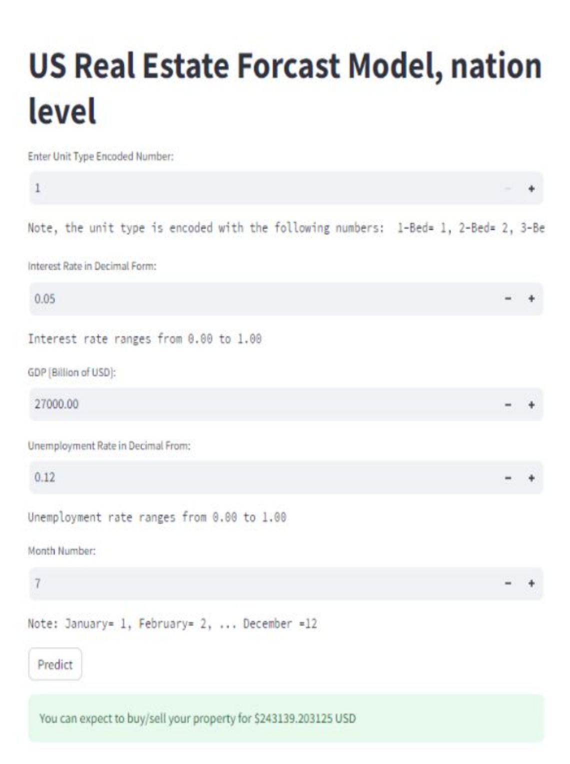
2. Linear, Ridge, Lasso and Elastic Regression performed ok for nation dataset, but poorly for cities. These models were very similar but not the best to tackle this problem. Similarly, decision Tree overfits before tuning. After tuning, MSE gets larger and R2 decreases.

3. Random Forest tends to overfit in MSE and has a high R2. This model is one that tends to take the longest execution time.

4. Finally, XGBoost is the best model for this project, even if it has a larger MSE than Random Forest, it does not over fits the data as random forest and it takes a bit less in execution time than random forest.

Product Demo

Note: The model was deployed using streamlit.



US Real Estate Forcast Model, nation level

Enter Unit Type Encoded Number:

1

Note, the unit type is encoded with the following numbers: 1-Bed= 1, 2-Bed= 2, 3-Be

Interest Rate in Decimal Form:

0.05

Interest rate ranges from 0.00 to 1.00

GDP [Billion of USD]:

27000.00

Unemployment Rate in Decimal From:

0.12

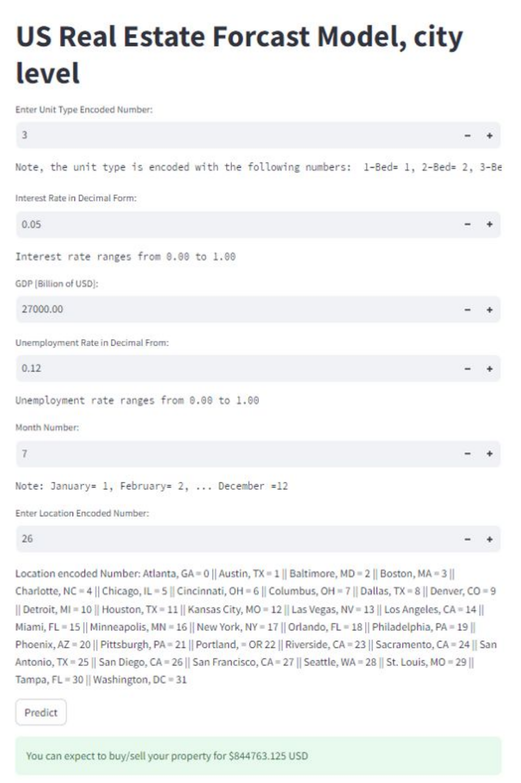
Unemployment rate ranges from 0.00 to 1.00

Month Number:

7

Note: January= 1, February= 2, ... December =12

You can expect to buy/sell your property for $243139.203125 USD



US Real Estate Forcast Model, city level

Enter Unit Type Encoded Number:

3

Note, the unit type is encoded with the following numbers: 1-Bed= 1, 2-Bed= 2, 3-Be

Interest Rate in Decimal Form:

0.05

Interest rate ranges from 0.00 to 1.00

GDP [Billion of USD]:

27000.00

Unemployment Rate in Decimal From:

0.12

Unemployment rate ranges from 0.00 to 1.00

Month Number:

7

Note: January= 1, February= 2, ... December =12

Enter Location Encoded Number:

26

Location encoded Number: Atlanta, GA = 0 || Austin, TX = 1 || Baltimore, MD = 2 || Boston, MA = 3 || Charlotte, NC = 4 || Chicago, IL = 5 || Cincinnati, OH = 6 || Columbus, OH = 7 || Dallas, TX = 8 || Denver, CO = 9 || Detroit, MI = 10 || Houston, TX = 11 || Kansas City, MO = 12 || Las Vegas, NV = 13 || Los Angeles, CA = 14 || Miami, FL = 15 || Minneapolis, MN = 16 || New York, NY = 17 || Orlando, FL = 18 || Philadelphia, PA = 19 || Phoenix, AZ = 20 || Pittsburgh, PA = 21 || Portland, = OR 22 || Riverside, CA = 23 || Sacramento, CA = 24 || San Antonio, TX = 25 || San Diego, CA = 26 || San Francisco, CA = 27 || Seattle, WA = 28 || St. Louis, MO = 29 || Tampa, FL = 30 || Washington, DC = 31

You can expect to buy/sell your property for $844763.125 USD

Takeaways and conclusion

1. The best model for the dataset collected was the tuned version of XGBoost. DF and RF tended to overfit the data and all linear mode did not have a good accurate measure

2. In order to improve the models, a more intensive and accurate/detailed data collection (usually given by premium or paid features) would be needed.

ATTENTION
YOUR
FOR
YOU
THANK

References

- https://www.zillow.com/research/data/
- https://fred.stlouisfed.org/tags/series?t=interest+rate%3Bmonthly%3Busa
- https://fred.stlouisfed.org/series/FEDFUNDS
- https://fred.stlouisfed.org/series/GDP
- https://fred.stlouisfed.org/series/CNP16OV
- https://fred.stlouisfed.org/series/LFWA64TTUSM647S
- https://fred.stlouisfed.org/series/EMRATIO
- https://fred.stlouisfed.org/series/UNRATE
- https://fred.stlouisfed.org/series/IHLIDXUS

Appendix 1- Nation Baseline Metrics

```
Linear Regression:
Pipeline execution time: 0.0259 seconds
Training MSE: 5799149884.892, Testing MSE:
5895238362.264
Training R^2: 0.531, Testing R^2: 0.569
Cross-validation Scores: [0.51391756 0.54872451
0.52211211 0.48879691 0.54264246]
Mean CV Score: 0.5232387095723793
------------------------------

Ridge Regression:
Pipeline execution time: 0.0293 seconds
Training MSE: 5799154668.605, Testing MSE:
5896224343.369
Training R^2: 0.531, Testing R^2: 0.568
Cross-validation Scores: [0.51393599 0.5487772
0.52214833 0.48891177 0.54250851]
Mean CV Score: 0.5232563608819628
------------------------------

Lasso Regression:
Pipeline execution time: 0.0160 seconds
Training MSE: 5799149889.575, Testing MSE:
5895268820.667
Training R^2: 0.531, Testing R^2: 0.569
Cross-validation Scores: [0.51391957 0.54872785
0.52211383 0.4888012  0.54264063]
Mean CV Score: 0.5232406138041081
------------------------------
```

```
Elastic Regression:
Pipeline execution time: 0.0257 seconds
Training MSE: 6558046792.195, Testing MSE:
7037165366.568
Training R^2: 0.470, Testing R^2: 0.485
Cross-validation Scores: [0.45624614 0.4897092
0.46590668 0.4649078  0.45276892]
Mean CV Score: 0.4659077462261256
------------------------------

XGBoost:
Pipeline execution time: 0.3012 seconds
Training MSE: 324979.403, Testing MSE: 5763229.121
Training R^2: 1.000, Testing R^2: 1.000
Cross-validation Scores: [0.99917359 0.99825071
0.99868394 0.99912043 0.99928007]
Mean CV Score: 0.998901749110788
------------------------------

Decision Tree:
Pipeline execution time: 0.0625 seconds
Training MSE: 0.000, Testing MSE: 22952039.291
Training R^2: 1.000, Testing R^2: 0.998
Cross-validation Scores: [0.99892911 0.99363193
0.99831137 0.99834542 0.99829013]
Mean CV Score: 0.9975015925802619
------------------------------
```

```
Random Forest:
Pipeline execution time: 3.0939 seconds
Training MSE: 1689648.019, Testing MSE: 10341116.988
Training R^2: 1.000, Testing R^2: 0.999
Cross-validation Scores: [0.99888472 0.99767833
0.99857687 0.99877712 0.99903424]
Mean CV Score: 0.9985902575697285
------------------------------

SVM:
Pipeline execution time: 0.5189 seconds
Training MSE: 13491434381.228, Testing MSE:
14874172763.144
Training R^2: -0.090, Testing R^2: -0.089
Cross-validation Scores: [-0.1311069  -0.04600841
-0.07302047 -0.0945792  -0.11483911]
Mean CV Score: -0.09191081934336895
------------------------------
```

Appendix 2- Nation Tuned Metrics

```
Linear Regression Tuned:
Pipeline execution time: 0.0300 seconds
Training MSE: 5799149884.892, Testing MSE:
5895238362.264
Training R^2: 0.531, Testing R^2: 0.569
Cross-validation Scores: [0.51391756 0.54872451
0.52211211 0.48879691 0.54264246]
Mean CV Score: 0.5232387095723793
------------------------------

Ridge Regression Tuned:
Pipeline execution time: 0.0328 seconds
Training MSE: 5799149983.094, Testing MSE:
5895298354.036
Training R^2: 0.531, Testing R^2: 0.569
Cross-validation Scores: [0.51392667 0.54873027
0.52211212 0.48880022 0.54263167]
Mean CV Score: 0.523240189587842
------------------------------

Lasso Regression Tuned:
Pipeline execution time: 0.0156 seconds
Training MSE: 5799149884.892, Testing MSE:
5895238392.870
Training R^2: 0.531, Testing R^2: 0.569
Cross-validation Scores: [0.51391756 0.54872452
0.52211211 0.48879691 0.54264246]
Mean CV Score: 0.5232387114797682
------------------------------
```

```
Elastic Regression Tuned:
Pipeline execution time: 0.0160 seconds
Training MSE: 5800008998.307, Testing MSE:
5909252528.401
Training R^2: 0.531, Testing R^2: 0.568
Cross-validation Scores: [0.51404664 0.54920399
0.52242642 0.48998104 0.5411575 ]
Mean CV Score: 0.5233631187732677
------------------------------

DecisionTree Tuned:
Pipeline execution time: 0.0313 seconds
Training MSE: 5800008998.307, Testing MSE:
5909252528.401
Training R^2: 0.531, Testing R^2: 0.568
Cross-validation Scores: [0.51404664 0.54920399
0.52242642 0.48998104 0.5411575 ]
Mean CV Score: 0.5233631187732677
------------------------------

Random Forest Tuned:
Pipeline execution time: 3.0072 seconds
Training MSE: 1689648.019, Testing MSE: 10341116.988
Training R^2: 1.000, Testing R^2: 0.999
Cross-validation Scores: [0.99888472 0.99767833
0.99857687 0.99877712 0.99903424]
Mean CV Score: 0.9985902575697285
------------------------------
```

```
XGBoost Regression Tuned:
Pipeline execution time: 0.5956 seconds
Training MSE: 5790942.339, Testing MSE: 17569412.846
Training R^2: 1.000, Testing R^2: 0.999
Cross-validation Scores: [0.99860613 0.99842979
0.9982562  0.99786821 0.99884665]
Mean CV Score: 0.9984013954132044
------------------------------
```

Appendix 3- City Baseline Metrics

```
Linear Regression:
Pipeline execution time: 0.0679 seconds
Training MSE: 27600232652.191, Testing MSE:
26936165905.391
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535485 0.24387534
0.25601862 0.25257623 0.25961198]
Mean CV Score: 0.2534874047675915
------------------------------

Ridge Regression:
Pipeline execution time: 0.0509 seconds
Training MSE: 27600232658.604, Testing MSE:
26936159331.547
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535534 0.24387586
0.2560182  0.25257617 0.25961154]
Mean CV Score: 0.25348742278188197
------------------------------

Lasso Regression:
Pipeline execution time: 0.0630 seconds
Training MSE: 27600232657.883, Testing MSE:
26936157147.189
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535503 0.24387591
0.25601863 0.25257641 0.2596118 ]
Mean CV Score: 0.25348755805819007
------------------------------
```

```
Elastic Regression:
Pipeline execution time: 0.0469 seconds
Training MSE: 28659422719.116, Testing MSE:
27896118574.410
Training R^2: 0.225, Testing R^2: 0.226
Cross-validation Scores: [0.22899289 0.22040419
0.22482168 0.22361812 0.22721197]
Mean CV Score: 0.2250097703606701
------------------------------

XGBoost:
Pipeline execution time: 2.1101 seconds
Training MSE: 145133144.509, Testing MSE:
176258286.485
Training R^2: 0.996, Testing R^2: 0.995
Cross-validation Scores: [0.99468931 0.99454418
0.9948928  0.99539891 0.99431643]
Mean CV Score: 0.9947683255455176
------------------------------

Decision Tree:
Pipeline execution time: 1.2250 seconds
Training MSE: 0.000, Testing MSE: 60500522.370
Training R^2: 1.000, Testing R^2: 0.998
Cross-validation Scores: [0.99782933 0.99758498
0.99794755 0.99824839 0.9977956 ]
Mean CV Score: 0.9978811702214216
------------------------------
```

```
Random Forest:
Pipeline execution time: 89.9446 seconds
Training MSE: 4097624.484, Testing MSE: 27876503.882
Training R^2: 1.000, Testing R^2: 0.999
Cross-validation Scores: [0.9988693  0.99878297
0.9988538  0.99894333 0.99891512]
Mean CV Score: 0.9988729043231597
------------------------------

SVM:
Pipeline execution time: 529.5205 seconds
Training MSE: 39221549155.927, Testing MSE:
38052733310.548
Training R^2: -0.060, Testing R^2: -0.056
Cross-validation Scores: [-0.06791118 -0.05694311
-0.06599032 -0.05498297 -0.06238686]
Mean CV Score: -0.06164289053802059
------------------------------
```

Appendix 4- City Tuned Metrics

```
Linear Regression Tuned:
Pipeline execution time: 0.0612 seconds
Training MSE: 27600232652.191, Testing MSE:
26936165905.391
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535485 0.24387534
0.25601862 0.25257623 0.25961198]
Mean CV Score: 0.2534874047675915
------------------------------

Ridge Regression Tuned:
Pipeline execution time: 0.3577 seconds
Training MSE: 27600232658.336, Testing MSE:
26936161702.965
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535542 0.24387604
0.25601805 0.25257565 0.25961182]
Mean CV Score: 0.2534873960658194
------------------------------

Lasso Regression Tuned:
Pipeline execution time: 0.0650 seconds
Training MSE: 27600232652.191, Testing MSE:
26936165896.625
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25535485 0.24387534
0.25601862 0.25257623 0.25961198]
Mean CV Score: 0.2534874049221946
------------------------------
```

```
Elastic Regression Tuned:
Pipeline execution time: 0.0629 seconds
Training MSE: 27601410078.271, Testing MSE:
26934491855.380
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25548646 0.24402188
0.25584642 0.25252572 0.25942869]
Mean CV Score: 0.25346183133210026
------------------------------

DecisionTree Tuned:
Pipeline execution time: 0.0725 seconds
Training MSE: 27601410078.271, Testing MSE:
26934491855.380
Training R^2: 0.254, Testing R^2: 0.252
Cross-validation Scores: [0.25548646 0.24402188
0.25584642 0.25252572 0.25942869]
Mean CV Score: 0.25346183133210026
------------------------------

Random Forest Tuned:
Pipeline execution time: 181.1773 seconds
Training MSE: 3954073.044, Testing MSE: 27601047.349
Training R^2: 1.000, Testing R^2: 0.999
Cross-validation Scores: [0.99887924 0.99875977
0.99887324 0.99895081 0.99893368]
Mean CV Score: 0.9988793472945297
------------------------------
```

```
XGBoost Regression Tuned:
Pipeline execution time: 4.5844 seconds
Training MSE: 134064071.188, Testing MSE:
154527785.549
Training R^2: 0.996, Testing R^2: 0.996
Cross-validation Scores: [0.99498713 0.99556252
0.99522476 0.9959914  0.99514503]
Mean CV Score: 0.9953821692299722
------------------------------
```