US Home Prices Analysis by Siris Gupta

Dataset: Previous 20 years

Data Science Model: Random Forests

Model Performance:

* Mean Squared Error (MSE): 1.54
* Root Mean Squared Error (RMSE): 1.24
* Mean Absolute Error (MAE): 0.87
* R-squared (R^2) Score: 1.00

From the True Value vs Predicted Value chart, we can see that the trend displays a linear line. Which is a favourable outcome as it indicates that the model predictions are very close to the true values. This graph also shows that the model is consistently making predictions that align with the actual outcomes across the range of data points boasting a highly accurate and consistent result.

A graph with blue dots

Description automatically generated

The Following table shows:

* The features selected to predict the target variable (Home prices)
* Feature importance to understand which features have the most influence on the model's predictions.

|  |  |
| --- | --- |
| **Features** | **Importance** |
| Feature 1: Consumer Price Index for All Urban Consumers | 0.04 |
| Feature 2: Bank Prime Loan Rate | 0.0094 |
| Feature 3: Employment Cost Index | 0.2194 |
| Feature 4: Residential Real Estate Loans | 0.0352 |
| Feature 5: Total Construction Spending | 0.2323 |
| Feature 6: Total Housing Units | 0.1788 |
| Feature 7: Employment Rate: Aged 15 - 64 | 0.01 |
| Feature 8: Interest Rates: Long - Term Government Bond Yields: 10 - Year | 0.0074 |
| Feature 9: Share Prices: All Shares/Broad | 0.1223 |
| Feature 10: Federal Funds Effective Rate (interest rate) | 0.0099 |
| Feature 11: GDP | 0.1353 |

Significant Features are:

* Total Construction Spending (0.2323)
* Employment Cost Index (0.2194)
* GDP (0.1353)
* Share Prices: All Shares/Broad (0.1223)
* Total Housing Units (0.1788)

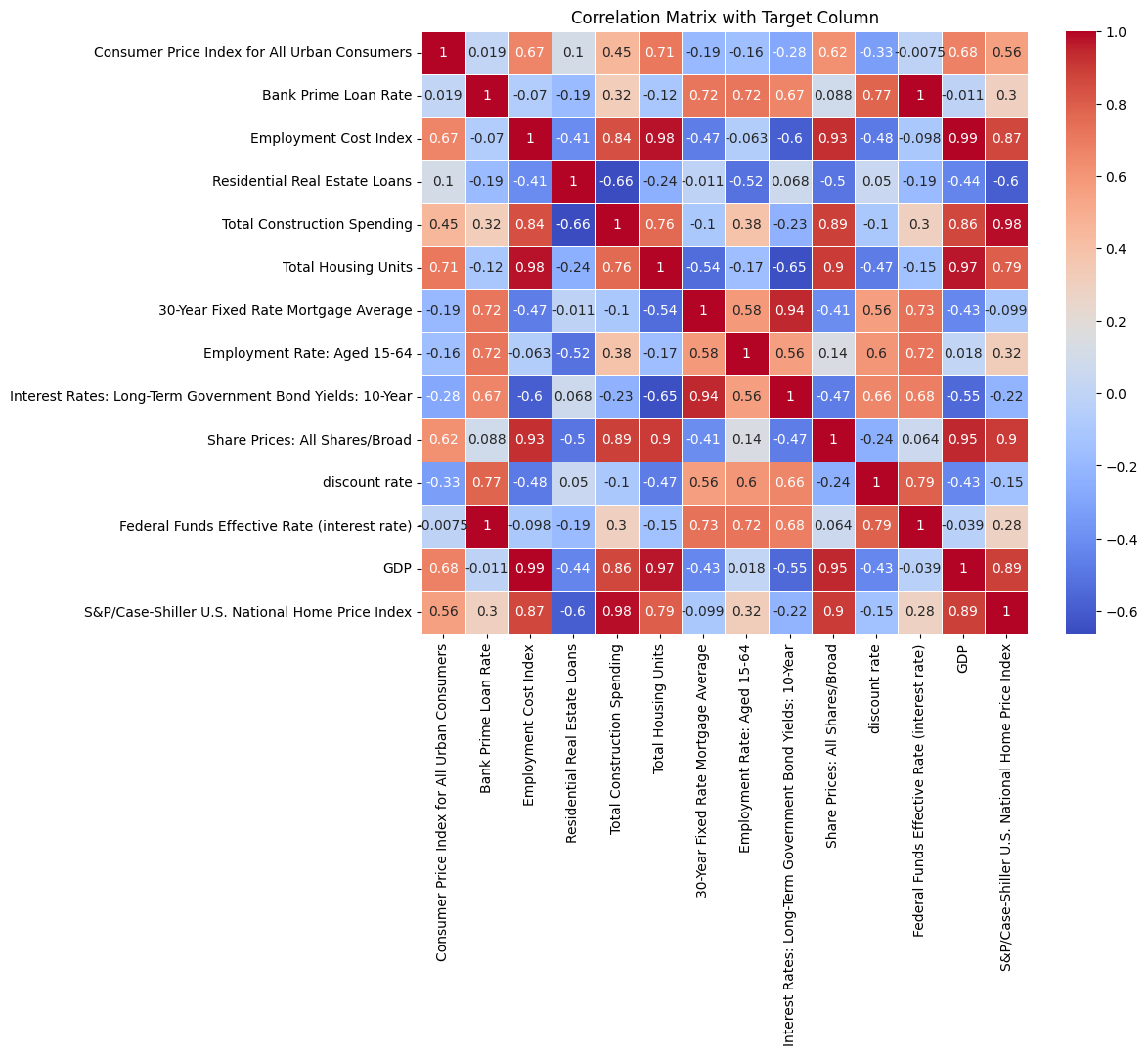
The factors that have the least impact on home prices are:

* Bank Prime Loan Rate
* Interest Rates: Long-Term Government Bond Yields: 10-Year
* Employment Rate: Aged 15-64

Correlation Matrix of all Features:

Note: Some features were removed while preprocessing due to very high

Multicollinearity with other features or due to very low correlation strength with the target variable.



Feature-wise Analysis:

Consumer Price Index (CPI): A higher CPI indicates that inflation is increasing, which can lead to higher home prices. This is because when inflation is high, consumers have less purchasing power, which can make it more difficult for them to afford homes.

Bank Prime Loan Rate (BPLR): A higher BPLR indicates that the cost of borrowing money is increasing, which can make it more expensive to buy a home. This is because when the BPLR is high, lenders charge higher interest rates on mortgages, which can make it more difficult for borrowers to qualify for a mortgage or afford a monthly payment.

Employment Cost Index (ECI): A higher ECI indicates that the cost of labor is increasing, which can lead to higher home prices. This is because when the ECI is high, builders have to pay more for labor, which can drive up the cost of new homes.

Residential Real Estate Loans (RRRL): An increase in RRRL indicates that more people are taking out loans to buy homes, which can lead to higher home prices. This is because when there is more demand for homes, prices tend to rise.

Total Construction Spending (TCS): An increase in TCS indicates that there is more investment in new construction, which can lead to higher home prices. This is because when more homes are being built, the supply of homes increases, which can put downward pressure on prices. However, if there is a shortage of homes, prices can rise.

Total Housing Units (THU): An increase in THU indicates that there is more housing available, which can help to keep prices in check. However, if there is a shortage of homes, prices can rise.

Employment Rate: Aged 15-64 (ER): A higher ER indicates that more people are employed, which can lead to higher home prices. This is because when more people are employed, they have more income, which can make it more affordable to buy a home.

Interest Rates: Long-Term Government Bond Yields: 10-Year (IR): A higher IR indicates that the cost of borrowing money is increasing, which can make it more expensive to buy a home. This is because when the IR is high, lenders charge higher interest rates on mortgages, which can make it more difficult for borrowers to qualify for a mortgage or afford a monthly payment.

Share Prices: All Shares/Broad (SP): Higher SP indicate that the stock market is doing well, which can lead to higher home prices. This is because when the stock market is doing well, investors have more confidence in the economy, which can lead them to invest in real estate.

Federal Funds Effective Rate (FFR): A higher FFR indicates that the cost of borrowing money is increasing, which can make it more expensive to buy a home. This is because when the FFR is high, banks charge higher interest rates on loans, which can make it more difficult for borrowers to qualify for a mortgage or afford a monthly payment.

Gross Domestic Product (GDP): A higher GDP indicates that the economy is growing, which can lead to higher home prices. This is because when the economy is growing, there is more demand for goods and services, including housing.

Overall, the data suggests that a number of factors can affect home prices, including inflation, interest rates, labour costs, demand for housing, and the overall health of the economy.

Process followed:

* Individual feature’s data from past 20 years was scraped from multiple sites and compiled into one master file.
* Some of the data was only collected quarterly, so I have created a python script convert the data from quarterly to monthly to create a standard interval of data points.
* Further preprocessing was done to remove null values and perform feature selection using correlation matrix.
* Dataset was finally split into training and testing data after which the dataset was normalized except for the target variable as we would like to retain the original scale of home prices.
* I have picked random forest for this regression type problem which was tuned using GridSearchCV and trained on the training dataset using the best parameters found from GridSearchCV.
* Finally some scikit learn metrics libraries were used to get the metrics and performance of the model which were displayed using and matplotlib.