US House Price Analysis

Gaurav Kanava
gauravkanava217@gmail.com

Introduction:

The objective of this analysis is to understand the key factors influencing US home prices over the past two decades. In pursuit of this, publicly available data sources have been harnessed to build a data science model that figures out the relationship between these factors and the S&P Case-Schiller Home Price Index, serving as a representative proxy for home prices on a national scale. This report delves into the methodology, data sources, findings, and conclusions derived from a comprehensive investigation of the factors that have impacted the US home prices from the past 20 years.

Methodology:

The methodology employed for this analysis encompasses three core phases: *data collection, data cleaning, and data analysis.*

Data Collection:
The initial phase involved brainstorming and identifying the potential factors that influence housing prices on a national scale. A keen focus was placed on ensuring that data for these features could be obtained from publicly accessible sources such as FRED (Federal Reserve Economic Data), Yahoo Finance, etc. Following this ideation, data for the identified features, including "Civilian Population," "Interest Rate," "Inflation," "Unemployment Percentage," "Median Family Income," "S&P 500," "GDP," and "Case-Schiller Home Price Index," was sourced from the respective platforms.

Data Cleaning:
The collected data was then subjected to rigorous cleaning and processing. This entailed addressing missing values, combining all files, data interpolation and converting date formats for consistency. The aim was to create a unified and coherent dataset that would facilitate comprehensive analysis.

Data Analysis:

The analysis phase was multi-faceted, encompassing various statistical and machine learning techniques to determine the impact of the identified features on the S&P Case-Schiller Home Price Index.

1. **Correlation Analysis:** The dataset was initially subjected to correlation analysis, where the Pearson correlation coefficient was computed for each feature in relation to housing prices. A high positive or negative correlation signified a strong linear relationship between the feature and housing prices.

2. **Linear Regression Modeling:** A multiple linear regression model was constructed, treating the features as independent variables and housing prices as the dependent variable. The coefficients of the model were interpreted to gauge the impact of each feature on housing prices. Positive coefficients indicated a positive impact, while negative coefficients suggested a negative impact.

3. **ANOVA Test:** The one-way ANOVA (Analysis of Variance) test was conducted to assess the significance of each continuous feature on housing prices. A low p-value from the ANOVA test indicated that the feature significantly influences housing prices.

4. **Random Forest Analysis:** A Random Forest model was implemented to evaluate feature importance scores. This approach quantified the contribution of each feature to the model's predictive performance, offering insights into their relative impact on housing prices.

These methods ranged from assessing linear relationships to understanding the significance and importance of features. The combined insights from these analyses contributed to a holistic interpretation of the influence of the identified factors on US home prices over the last two decades.

Data Collection and Dataset Preparation:

Data Collection: The foundation of this analysis was laid with the collection of data from various public sources. The following is a list of the data sources and the corresponding features that were sourced:

1. **Civilian Population (Thousands):** https://fred.stlouisfed.org/series/CNP16OV
2. **Interest Rate (Percentage):** https://fred.stlouisfed.org/series/MORTGAGE30US

3. **Inflation (Percentage):** https://fred.stlouisfed.org/series/T10YIE#0
4. **Unemployment (Percentage):** https://fred.stlouisfed.org/series/UNRATE
5. **Median Family Income (Current Dollars):** https://fred.stlouisfed.org/series/MEHOINUSA646N
6. **S&P 500 Index:** https://finance.yahoo.com/quote/%5EGSPC/history?period1=1010361600&period2=1699315200&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true
7. **GDP (Billions of Dollar):** https://fred.stlouisfed.org/series/GDP
8. **Case-Schiller Home Price Index (Housing Prices):** https://fred.stlouisfed.org/series/CSUSHPISA

Dataset Preparation: Creating the final dataset for analysis involved two critical steps:

1. **Draft Dataset Integration:** An initial draft dataset was assembled, primarily comprising features with readily available data, namely "Civilian Population," "Interest Rate," "Inflation," "Unemployment Percentage," and "Case-Schiller Home Price Index (Housing Prices)." Data integration was implemented using code to ensure uniformity in the date format.

   Code File: US Home Price Impact Dataset.ipynb

2. **Final Dataset Compilation:** Incorporating "GDP," "Median Income," and "S&P 500" data into the final dataset presented certain challenges. For "GDP" and "Median Income," where data was available only on a quarterly or annual basis, careful manual data compilation was carried out. Regarding the "S&P 500" data, due to its ambiguous format, a manual creation process was employed to ensure data integrity and consistency.

Post initial compilation, the final dataset was subjected to thorough preprocessing, addressing missing values, null entries, and ensuring data integrity. This was done in the other code file where it was finally Analyzed.

Analysis and Findings:

Code File: US Home Price Impact.ipynb

1. **Pearson Correlation:**

The Pearson correlation coefficients provide insights into the linear relationships between features and the housing price index (CSUSHPISA).

- *Notable correlations:*
- Strong positive correlation with "SP500" (0.931535), "GDP" (0.884637), and "Median Family Income" (0.892649).
- Strong negative correlation with "Unemployment Percentage" (-0.535753).

- Interpretation:
- Positive correlations suggest that as the values of these features increase, the housing prices tend to increase.
- Negative correlation with "Unemployment Percentage" indicates an inverse relationship, where higher unemployment percentages are associated with lower housing prices. It's as more people are unemployed means housing prices must fall down.

2. **Linear Regression:**

- Coefficients represent the estimated impact of each feature on the housing price.

- *Notable impacts:*
- Positive impact: "Interest Rate," "SP500," "GDP."
- Negative impact: "Population," "Inflation," "Unemployment Percentage."

- Interpretation:
- The coefficients represent the estimated change in the target variable (housing prices) for a one-unit change in each feature, assuming all other factors remain constant. Positive coefficients indicate a positive impact on housing prices, while negative coefficients indicate a negative impact.

3. **ANOVA Test:**

- ANOVA p-values assess the significance of each feature in relation to housing prices.

- *Notable results:*
- Significantly low p-values for "Unemployment Percentage" and "Median Family Income."
- High p-values for "Population," "Interest Rate," and "Inflation."

- *Interpretation:*
- Low p-values indicate that "Unemployment Percentage" and "Median Family Income" significantly impact housing prices.
- High p-values suggest that "Population," "Interest Rate," and "Inflation" may have less significant impact or require further investigation.

4. **Random Forest:**

- Feature importance scores quantify the contribution of each feature to the model's predictive performance.

- *Notable feature importance:*
- High importance: "GDP," "SP500," "Population."

- *Interpretation:*
- Features with higher importance contribute more to predicting housing prices according to the Random Forest model.

Overall Implications:

- The findings collectively suggest that economic indicators such as GDP, stock market performance (SP500), and demographic factors (population) play significant roles in influencing housing prices.
- The negative impact of "Unemployment Percentage" on housing prices is noteworthy, highlighting the sensitivity of the housing market to labor market conditions.
- While interest rates and inflation show some impact, their significance may be subject to further investigation.
- The combination of multiple analyses provides a robust understanding of the complex relationships between features and housing prices, enabling informed decision-making in the realm of real estate economics.

Limitations:

The analysis has a few limitations worth noting. The data varied in frequency—some features were available quarterly or annually, impacting the consistency of time-related comparisons. Missing values, especially in "Median Income," presented challenges and required Interpolation to add missing values. Additionally, the chosen factors were based on my understanding of their potential influence on housing prices. This subjective selection may overlook other relevant variables, and the assumptions made in the analysis might not fully capture the complexity of the relationships involved. In summary, these limitations suggest any ambiguity if it might come in the analysis.

Conclusion:

In examining US home prices over the past two decades, my analysis took a multi-faceted approach. I gathered data from public sources, tackling challenges like uneven timeframes and missing values. Through straightforward methods like correlation and regression, I uncovered some key influencers—economic factors like GDP, stock performance (SP500), and demographics, with unemployment percentages playing a negative role. While the findings shed light on housing trends, it's worth noting the limits, like data gaps and subjective feature choices.