Data References and Explanation

Prepared by:Adarsh Kumar

Date: 17-10-23

In this assignment, we needed to use the S & P Case-Schiller Home Price Index as a proxy for home prices and identify key factors that influence US home prices nationally. Then,we needed to build a data science model that explains how these factors impacted home prices over the last 20 years.

1.Introduction: The model aims to analyze the key factors that influence U.S. home prices at the national level. The analysis spans the period from 1st April 2003 to 1st April 2023 because of the availability of the data till that month for all key features and explores various economic and housing market indicators.

2.Data:Output:

House Price Index (Index Jan 2000=100): Since we had to use S&P Case-Schiller Home Price Index as a proxy for home prices. I downloaded the monthly data of last 20 years from 2003-04-01 to 2023-04-01(considering the data availability of other features) as a CSV file and renamed it:'HPI(Apr2003-Apr2023).csv'*(Uploaded the file in the data folder with same name)* and took it directly as House Price Index, which is our output in this model. Link of the website(As mentioned in the question)

Input:

1. **Population (In thousands):** Population growth reflects the demand and supply of housing in a given market. Population growth can also affect the income, preferences, and needs of consumers for housing. So, I considered this as a factor. I downloaded the monthly population data of last 20 years from 2003-04-01 to 2023-04-01 as a CSV file and renamed it:'Population(Apr2003-Apr2023).csv'*(Uploaded the file in the data folder with same name)*, from this website.
2. **GDP (In Billions of Dollars):** Gross Domestic Product is a measure of the total value of goods and services produced in a country. It reflects the economic activity and growth of a country. A higher GDP indicates a stronger economy, which may increase the income and purchasing power of consumers, and thus increase the demand and price of housing.That's why I considered it a factor. I downloaded a CSV file from this website and named it 'GDP(Apr2003-Apr2023).csv'*(Uploaded the same in the data folder)* which contained quarterly data and converted these quarterly data into monthly data with the same value for each month in respective quarters.

3. **Unemployment Rate (In Percent):** The Unemployment Rate represents the percentage of people actively seeking employment but currently jobless, serving as an economic health indicator. A high Unemployment Rate often lowers consumer confidence, reducing demand for housing and impacting prices. That's why I considered it as a factor. I downloaded the monthly Unemployment Rate data of last 20 years from 2003-04-01 to 2023-04-01 as a CSV file from this website and renamed it:'UnemploymentRate(Apr2003-Apr2023).csv' *(Uploaded the same in the data folder)*.
4. **Mortgage Rate (In Percent):** Mortgage rate is the interest rate charged by lenders to borrowers for home loans. It affects the affordability and availability of credit for home buyers. A lower mortgage rate reduces the cost of borrowing, which may stimulate the demand and price of housing. That's why I considered it as a factor. I downloaded a CSV file from this website and named it 'Mortgagerate(Apr2003-Apr2023).csv'*(Uploaded the same in the data folder)* which contained weekly data for each month. So, I aggregated the weekly data into monthly data with the help of mean.
5. **Consumer Price Index(CPI-U all items 1982-1984=100, seasonally adjusted):** One of the most common measures of inflation is the Consumer Price Index (CPI), which is produced by the Bureau of Labor Statistics (BLS).The CPI shows changes in the prices paid by urban consumers for a representative basket of goods and services, such as food, energy, clothing, **housing**, etc. I've used CPI as a factor and not CPPI(Commercial Property Price Index is a measure of the changes in prices of commercial real estate properties) because CPPI is not as widely available and updated as CPI, which is produced by the Bureau of Labor Statistics (BLS) every month. I downloaded the data of CPI as a CSV file from this website and named it:'CompleteCPI.csv'*(Uploaded the same in the data folder)*, which contained monthly data as Period with their years separately. Then, I created a dictionary:'month_dict' to map month names to their numerical representation, and obtained monthly data from 2003-04-01 to 2023-04-01, after converting the year and month into this format.
6. **Housing Credit Availability Index(In Percent):** Housing Credit Availability Index is a measure of the percentage of home purchase loans that are likely to default. It reflects the risk appetite and lending standards of lenders. A higher HCAI indicates that lenders are willing to offer more credit to borrowers, which may increase the supply and price of housing.That's why I considered it as a factor.I downloaded the HCAI data from this website as an excel file and named it:'HCAI_Chart.xlsx'*(Uploaded the same in the data folder)*,which now contains only Total Risk(%) *(sum of Borrower Risk & Product Risk)* as 'Housing Credit Availability Index' of **Whole market** with Quarter and Year separately. Earlier this excel file had 4 workbooks consisting of Whole market, GSE(Government Sponsored Enterprise),Govt., and PP(Portfolio and Private), but I removed other workbooks and considered only Whole market data because it's a combination of all the

other categories,and represents the entire market of home purchase loans. This category reflects the overall credit availability and riskiness of the mortgage market. Then, I removed unwanted columns in this excel workbook and left it with only 'Year','Quarter' and renamed 'Total Risk' column as **'Housing Credit Availability Index'**. Then, I created a dictionary:'quarter_to_months' to map quarter numbers to their corresponding months, and obtained monthly HCAI data, with each month in 'YYYY-MM-01' format.

3.Model Building: I chose a Random Forest Regression model for our analysis because of its ability to capture complex relationships and handle multicollinearity. This model makes predictions by aggregating predictions from multiple decision trees, and it's capable of capturing complex interactions and non-linear relationships between features and the target variable. And, we had seen that almost all of our features were highly correlated from the correlation table. And the random forest model is robust to multicollinearity (high correlation between input variables), making it less sensitive to strong correlations among the features. We split the data into a training set and a test set (80% training, 20% testing). And the above features were used for the model. We train the created random forest model on training data and utilize it to make predictions on the test data.

4.Model Evaluation: The Random Forest Regression model was trained and evaluated using the following metrics:

- R-squared (R^2) to measure the proportion of the variance in the House Price Index explained by our model.
- Mean Absolute Percentage Error (MAPE) to assess the accuracy of our predictions in percentage terms.

5.Result: Our Random Forest Regression model achieved the following performance metrics:

- R-squared (R^2): 0.9974, which is high R-squared value and suggests that the model explains a substantial portion of the variance in the House Price Index.
- Mean Absolute Percentage Error (MAPE): 0.9293%, which is low MAPE and indicates that our model's predictions are relatively accurate in the context of the target variable.