Objective: To analyze publicly available data on economic, demographic, and real estate indicators to build a predictive model that explains the impact of these factors on the S&P/Case-Shiller Home Price Index, a key indicator of the U.S. home prices over the last two decades.

Introduction:

Housing market has always been a strong case to study not only for its dependability and dynamics on various factors, but also for the fact that it impacts people's lives and livelihood in a big way. Price levels of homes in the United States can be studied well by following the S&P CoreLogic Case-Shiller Home Price Indices. Housing market is generally affected by multiple factors and if one wants to monitor and understand the dynamics of home prices, these indices offer reliable and valuable insights into the ever-changing housing market. A key component of this system is the S&P CoreLogic Case-Shiller U.S. National Home Price Index. Its comprehensive insight of the overall value of the single-family homes provides very valuable input to homebuyers nationwide. The way it works is by aggregating data from nine different regions of the nation. Furthermore, it updates data on a monthly basis. It also offers average price changes in specific geographic markets, covering 20 major metropolitan areas. People benefit from these indices by focusing on these city indices. Structurally, these areas are further grouped into two composites, one with 10 metro areas and the other that encompasses all 20. These indices have been built to measure percentage variations in housing market prices. However, these indices also maintain a constant level of quality. This is done by ensuring that variations due to factors such as house types, sizes, or physical characteristics are excluded from the calculations. This makes it a very much reliable index.

Supply Side Factors That Affect Home Prices

Important factors that affect the supply of houses available for sale are discussed below:

Months of Supply

Months of supply is one of the basic measures of supply. It tells how many months it would take for the current inventory of homes to sell. It quantifies the supply and demand in the housing market. Houses for sale is another measure of the same.

Data Source:

Months of Supply: https://fred.stlouisfed.org/graph/?g=zneA

Homes For Sale: https://www.census.gov/construction/nrs/historical_data/index.html

Unemployment

Unemployment is one of the key factors that affect both demand and supply in the real estate market. A high rate of unemployment means that people do not have the money to spend on houses. It may also mean that there is a dip in the investment side, and hence lower supply. Fear of unemployment where people are not really unemployed, but think they may lose their employment, can cause a deep impact in the real estate market.

Data Source: https://fred.stlouisfed.org/series/UNRATE

Mortgage Rate

Mortgage rates are one of the biggest factors that work as a pulse of the housing industry. It can simply predict how well the real estate market will or will not perform. It is such a factor that it actually plugs into both supply and demand side of the equation. It impacts the options of availability

of finance to buyers. It eases the inflow of finance for new constructions. Another aspect is the delinquency rate and the number of refinances for mortgages, which it affects. People are more likely to default on a higher mortgage rate. However, too low mortgage rate can create imbalance in the supply side as there could be demands outgrowing the supply.

Data Source: https://fred.stlouisfed.org/graph/?g=zneW

Federal Funds Rate

This comes into direct play when the Federal Funds Rate makes a direct dent into the housing market. Although mortgage rate and Federal Funds Rate are usually closely related, sometimes they may not be. There have been instances in the past when the Federal Government lowered the Fed Funds Rate, but banks did not reciprocate by lowering mortgage rates for multiple reasons. Sometimes the lowering happened, but did not complement each other by being proportionately lowered. Variations in the Federal Funds Rate influence the economy in many ways, not just the real estate market. Many of these factors again indirectly influence the real estate market. So the Fed Funds Rate impacts the housing market in a complex way.

Data Source: https://fred.stlouisfed.org/series/DFF#0

USA GDP

The GDP is a measure of the health of the economy, including the output of the overall economy. An economy that is doing well is usually indicated by healthy GDP and its year on year growth. This implies more investment and economic activity, and more buying. Housing sector also follows the same trend as the overall GDP.

Data Sources:

Monthly Index: https://fred.stlouisfed.org/graph/?g=znfe

Quarterly Real GDP (adjusted for inflation): https://fred.stlouisfed.org/series/GDPC1#0

Building Permits

By measuring the number of building permits allotted one can not only predict just the health of real estate industry, but also how free the real estate market is. It is an indicator of the extent of regulation/de-regulation of the market. It affects the supply through ease of putting a new property on the market.

Data Source: https://www.census.gov/construction/bps/

Housing Starts

This factor is a measure of the number of units of new housing projects started in a given period. This is either measured in terms of the number of units or in valuation of housing projects started in a given period.

Data Source: https://www.census.gov/construction/nrc/historical_data/index.html

Construction Spending

The amount of money spent (in millions of USD, seasonally adjusted) in the housing industry, is a measure of the activity in that industry. It is also an indicator of the supply that going to be available for future months.

Data Source: https://fred.stlouisfed.org/series/TTLCONS

Demand Side Factors That Affect Home Prices

Demand for housing is affected by many factors. Many of these factors also affect the supply side in the housing market. Below are a few factors that are prominent in influencing the demand for home buying.

Affordability: Wages & Disposable Personal Income

The "weakly earnings" are a measure of overall wages and earning of all employed persons.

The Disposable Personal Income is the other measure of how much of the earning is actually available to an individual for expenditure. This is an important measure as it takes into account taxes.

Data Sources:

Median usual weekly nominal earnings, Wage and salary workers 25 years and over: https://fred.stlouisfed.org/series/LEU0252887700Q#0

Real Disposable Personal Income: https://fred.stlouisfed.org/series/DSPIC96#0

Delinquency Rate on Mortgages

If one has to predict the number of foreclosures in real estate, one would look for the delinquency rate on housing mortgages. In the last economic recession, higher delinquency rate (higher than credit card delinquency rate) was a key indicator of the recession, the poorly performing industries, and the economic condition of the nation as a whole. It also indicates the feasibility of a homeowner to buy a house at a certain point of time. It also indicates the overall demand in the industry.

Data Source: https://fred.stlouisfed.org/series/DRSFRMACBS#0

Personal Savings

People can save their money after making all the necessary expenses. Therefore, the savings are indicator of one's own financial condition. The extent by which people are saving their money, matters in overall investments and capital availability. It is also a factor that decides the interest rate of loans (and not just the mortgage rate). It also signifies the comparison between people's inclination to spend their money and save it for future use. This therefore proves to be a significant indicator of the demand for home ownership as well.

Data Source:

Personal Saving: https://fred.stlouisfed.org/series/PMSAVE

Behavioral Changes and Changes in Preferences

Going for a home ownership is a significant and important decision in one's life. It is derived from a combination of factors including change in preferences and attitudes of people towards home buying. Change in cultural trends can be taken as a revealed metric of propensity for home buying.

There is another metric to track changes in preferences in home buying, and that is, personal consumption expenditure. However, there may not be a one to one correlation. For example, if expenditure is increasing, but not the homeownership, it would indicate a change in preferences towards home buying and ownership. It is quite possible that people prefer to rent a home than buying one. Hence, both of these parameters are used.

Date Sources:

Home Ownership Rate: http://bit.ly/homeownershiprate

Personal Consumption Expenditures: https://fred.stlouisfed.org/series/PCE

Data Cleaning

One of the challenges of dealing with this acquired data is that some of the metrics are reported as monthly, whereas many others are computed as quarterly.

For my dataset,

A) In one set, I have down-sampled the monthly data to quarterly, and after that changed it annually. So, in final computation, all data is annually. This data is target variable.

B) In another set, I have down-sampled the monthly data to quarterly, and after that changed it annually. So, in final computation, all data is annually. This data is feature.

C) In the next step merged both data sets as 'final_df' (Final DataFrame).

In my dataset, most of the data is official government data for key economic indicators, and hence, there are very few "missing values". The ones that are missing, are replaced with averaged values (average of one value before and after it). However, I have taken the training data only for the period in which all the variables have data available, namely, from 2003-2023. To down-sample monthly data, the value for the month is added as per the quarter (Jan+Feb+March, April+May+June, July+Aug+Sep, Oct+Nov+Dec). This data is mostly for migration data, and this seems to be a fair way to approximate the value, even if not very accurate.

Building The Model

I have used Linear Regression for building the model. I have selected the relevant features, including 'MSACSR' (monthly supply of new houses), 'UNRATE' (unemployment rate)'PERMIT' (number of new housing units authorized), 'QUSR628BIS' (Real Residence property price), 'TTLCONS' (total construction spending on residential projects, 'PSAVERT' (Personal saving rate), 'GDP' (Gross Domestic Product), 'HSN1F'(New One Family Houses Sold), 'HOUST1F' (New Privately-Owned Housing Units Started Single-Family Units), 'UMCSENT' (consumer sentiment index), 'INTDSRUSM193N' (interest rates or discount rates), and. The target variable aimed to predict is 'CSUSHPISA' (S&P Case-Shiller U.S. National Home Price Index).

After preparing the, split it into training and testing sets using an 70:30 ratio, where 70% of the data used for training the model, and 30% was for testing.

Next, we defined a dictionary of candidate models, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression, and Neural Network. These models represent different algorithms with varying complexities and learning capabilities.

To select the best performing model, I have used cross-validation with five folds. I used the mean squared error (MSE) as the evaluation metric, where lower values indicate better performance.

Based on the cross-validation results, identified Linear Regression as the best model with the lowest MSE and trained the Linear Regression model on the entire training data.

Model Evaluation:

To evaluate the performance of our model, we used two key metrics: mean squared error (MSE) and R-squared score. Based on evaluation, the Linear Regression model performed well.

In summary, model evaluation showed that the Linear Regression model performed well, with low prediction errors (MSE) and a high proportion of variance explained (R-squared score). The coefficients provided insights into the importance of each feature and their impact on the predicted home price index.