US House Price Index Analysis

Introduction and problem statement

The U.S. House Price Index (HPI) is a metric used to track and measure changes in the prices of residential properties in the United States over time. It provides insight into the trends and movements in the housing market, specifically the prices of homes. The HPI is usually calculated relative to a specific baseline year. The prices in the baseline year are set to 100, and changes in subsequent years are measured relative to that base year (We take year 2000 as base year in this study). HPI values are often reported on a regular basis, such as monthly, quarterly, or annually, to provide timely information about housing market trends (We consider monthly frequency in this study). In this study we consider the S&P Case-Schiller Home Price Index as a proxy for home proxy. The S&P Case-Shiller Home Price Index (often referred to simply as the Case-Shiller Index) is a widely followed and respected measure of residential real estate prices in the United States. It is named after its creators, economists Karl E. Case and Robert J. Shiller. The index is designed to provide a comprehensive and accurate view of changes in the value of residential real estate over time. We are trying to predict what different socio-economic factors affect this Index.

Approach

We consider different Demand and Supply factors that could affect The HPI. We first identify such factors, then collect the monthly data for these factors from year 2000 from open source web, then use python and its data manipulation libraries to accumulate this data and create a final dataset. Once this dataset is created we clean and preprocess this dataset and perform in-depth analysis to identify which factors affect HPI. We train a regression model based on Machine Learning algorithms to obtain relative importance of features to asses their power to predict and determine HPI.

Data

The Demand and Supply data features regarding House Price Index are collected from web. The main source for most of the data is Federal Reserve Bank of St. Louis(FRED) Economic Data Repository. The link to this source is here : <https://fred.stlouisfed.org/>. Different features considered are as follows:

**Demand Features:**

**Interest Rates, Discount Rate For US**: Interest rates, in a general sense, refer to the cost of borrowing money or the return on investment for lending or saving money. The discount rate, often referred to as the "Federal Reserve discount rate" or the "discount rate at the Fed," is the interest rate at which eligible depository institutions (mostly banks) can borrow money from their regional Federal Reserve Bank. When Interest rates are low it becomes more affordable for people to borrow money for buying homes. Hence this seems to be an important factor to consider.

**30-Year Fixed Rate Mortgage Average in the United States:** The 30-Year Fixed-Rate Mortgage Average in the United States refers to the average interest rate on 30-year fixed-rate mortgages as reported by various financial institutions and agencies. It is a key indicator of the prevailing mortgage interest rates in the U.S. housing market. Homebuyers and homeowners closely monitor these rates because they have a significant impact on the affordability of buying or refinancing a home.

**GDP Per Capita:** Gross Domestic Product (GDP) per capita is a commonly used economic indicator that measures the average economic output, or income, per person in a given country or region. It is calculated by dividing the total GDP of a country or region by its population. The result approximates the average income or economic well-being of the population. It makes sense that if GDP Per Capita is higher then HPI would be higher.

**Consumer Sentiment:** Consumer sentiment, in the context of the housing price index in the United States, refers to the attitudes, perceptions, and expectations of consumers regarding the housing market and home prices. It reflects how confident or optimistic consumers are about the state of the housing market, their willingness to buy or invest in real estate, and their expectations regarding future home price movements. Consumer sentiment plays a crucial role in influencing housing market dynamics and can impact supply and demand, as well as home prices.

**Inflation:** When there is overall inflation in the economy, it often leads to higher consumer prices and increased purchasing power for individuals. This can result in increased demand for housing, as people may want to invest in real assets like homes to protect their wealth from eroding due to inflation. As demand for housing rises, prices tend to go up, contributing to an increase in the housing price index.

**Population:** Population seems like a very common socio-economic factor that would affect HPI

**Employed Population**: Employed population would generally tend to affect HPI in some way.

**Supply Features:**

**Avg. House sales price:** The average house sales price can have a direct impact on the housing price index in the United States. The average house sales price is one of the key components used to calculate HPI. So, it makes sense to include this feature in the analysis.

**Monthly Supply of New Houses in the United States:** It seems fair to assume that greater the monthly supply of houses lesser the HPI. This could be thought in terms of supply-demand dynamics.

**New Housing Units Authorized:** This feature too could also affect HPI from the perspective of supply-demand dynamics.

**New Housing Units Completed:** Same principle of supply-demand dynamics applies here too

**producer price index:** Producer price index considered here is index that measures cost of construction materials. It makes sense that if this index is high then HPI might also be high.

So, these are the different features considered in this study to analyze their effect on HPI.

Data Preprocessing, EDA & Modelling:

Once we collect the data from web, we merge all the demand and supply features to create one final dataset. Most of the features have monthly frequency but some features have quarterly frequency. We interpolate quarterly frequency features using time-based interpolation technique since we are basically dealing with time series data here. Then we conduct in-depth EDA to understand the nature of each feature and how each feature affects our target variable. We do necessary feature transforms and engineer new features based on our analysis. To reduce multi-collinearity, we drop features that are linearly dependent on each other.

Before going ahead with modelling, we split the data into train and test sets. We do time-based splitting i.e. last or most recent 20 percent of the data is reserved for testing purpose. To start with we apply linear regression model to predict HPI. Before training the linear regression model, we standardize the dataset. We get acceptably good performance with linear regression. We try more complex models but they do not outperform linear regression. To understand which features, affect HPI mostly, we take a look at coefficients or weights assigned by optimization algorithm in linear regression to each feature. We get observations as below.

Observations:

The important features that affect HPI in positive manner are as follows (In order of their importance): GDP Per Capita, New houses completed in a given month, avg sales price of houses, New monthly houses introduced each month. It means higher the value of these variables, higher the House Price Index.

The important features that affect HPI in negative manner are as follows (In order of their importance): Avg Mortgage rate, Consumer Sentiment Index, Employed Population i.e. higher the value of these variables lower the house price index.

Graph below shows feature importance:

