Analyzing the Impact of Macroeconomic Factors on Real Estate: A Study of CPI, Interest Rates, and Home Inventory

Project Progress Report (Phase 3)

Team 68
Nihar Reddy Moramganti, GTID: 903848828
Yuchen Hu, GTID: 903844369
Michael Cardenas, GTID: 903835761
Linh Nguyen Dang , GTID: 903757931

Georgia Institute of Technology

MGT6203: Data Analytics For Business

Frederic Bien

July 09, 2023

Background Information

Being a first-time home buyer and completing that purchase is one of the best moments in a person's life. For many people, especially recent graduates and newly married couples, looking into making their first ever housing purchase, they often struggle deciding when is the right time to buy as the market might be too expensive at that given moment. There are many who work for years to save up enough money to buy their first home, but ultimately, the market dictates when most have an opportunity to buy. In addition, during the COVID-19 pandemic, housing prices went up exponentially, and many people who wanted to buy a house had to defer, not knowing how long unfavorable market conditions would persist.

Problem Statement

We want to find which variables have the most influence on housing prices so we can better predict housing prices. One of the reasons we want to look into this is because there are many drivers that affect housing market prices. With housing prices being at their highest during the COVID-19 pandemic, we want to explore how these drivers influenced housing prices during that time, and also how these factors played pre-pandemic as well. Based on housing data from 2016-2023, we want to build a model that would be the best fit for determining the most influential variables that will predict future housing prices. These factors will also play a role in investment opportunities for people to invest in property based on the housing market trends.

Business Justification

The most important aspect of what the purpose of this project intends to do for first-time home buyers is being able for them to buy homes at the right time. This will make sure that they are well prepared and knowledgeable about housing market trends and be smart on when to buy a house. When families buy a home at the right time, they will be able to buy it for cheaper when demand is less. Another aspect of this is property appreciation. Based on the prices of other homes on the market, a home buyer can buy that property, work to renovate a certain aspect of that home, and then sell it at a higher price than what it was initially sold for. This will make home buyers more encouraged to invest in these properties and help generate more income for themselves. These two points mainly will not only help home buyers save money, but also gain wealth based on the property appreciation.

Data Exploration

1. 30-Year Fixed Rate Mortgage Average in the United States.
2. Adjusted Consumer Price Index (CPI) for All Urban Consumers.
3. Housing Inventory: Active Listing Count in the United States
4. Annual Estimates of the Resident Population by Counties in the United States

Initial Hypothesis

1. Housing prices have a negative correlation with interest rates. The lower interest rate will lead to lower monthly payments as well as better borrowing power. This increased affordability will increase the housing popularity and drive up prices. However, the impact of interest rates on housing prices may have a lagged effect taking several months to materialize.

2. Housing prices have a positive correlation with CPI. Higher inflation, as indicated by CPI, can increase the costs of labor and materials involved in construction and maintenance. This increased cost will translate into higher housing prices.
3. Housing prices have a negative correlation with active listings. When there is a limited number of housing supply compared to the number of potential buyers, it will create a competitive market. The increased competition can drive up the demand and have a subsequent impact on housing price growth.
4. Housing prices have a positive correlation with population growth. Change in population, especially with population growth, can influence housing demand, and potentially lead to higher prices to limited housing supply.

Data Clean up in R

1. Standardizing the Time Series: To align the dataset, we recognized that datasets were reported at different levels of time granularity. We decided to convert them to monthly which ensures we have better comparability across the variables.
2. Selecting Appropriate Time Frame: While we aimed to use a longer date period for our model to capture diverse trends, we faced a limitation with the active house listing dataset, which only provided data from July 2016 onwards. We made the decision to restrict our model's time frame between July 2016 and May 2023. This timeframe still captures pre and post-pandemic periods and has sufficient data to perform accurate prediction while acknowledging the limitation of datasets.
3. County Selection based on Population: Considering the vast number of countries and incomplete data for smaller counties. We experimented with different population thresholds to identify the most representative counties. Ultimately, we chose 16 counties from 8 different states based on a population over 2 Millions as of 2022.

By implementing those steps, we aligned the datasets, established a relevant timeframe and selected representative counties for our analysis. These efforts improved the data quality and allowed us to perform our prediction model effectively.

Approach/Methodology

Our approach starts with an exploratory analysis of the various data sources we were able to pull together. The aim is to gain insights into the data and identify potential features that may influence housing prices. Pulling together data from the vast amount of sources found online will require a significant amount of cleaning and pre-processing. We expect to have to normalize various data sets, as well as one-hot encode categorical variables such as states, counties, and property types. Some features that we anticipate having a significant impact on housing prices are macroeconomic factors, demographic indicators, and historic pricing. Once the variables have been selected, we will split the data into training, validation, and testing sets. From there we will fit multiple models using various methodologies such as linear regression, decision trees, and random forest regressions. To optimize our model, we will use established procedures such as Bayesian optimization, grid search or randomized search. Since each approach to optimization carries its own benefits and drawbacks, we will decide on how to proceed once we have completed the variable selection phase. To evaluate different models, we will use AUC and R-squared to compare model performance. We will also use cross-validation to ensure that our model generalizes well to unseen data. The model with the best overall performance will then be used in our final analysis.

We expect our analysis to produce results in line with expert opinion. Additionally, the final model will have a reasonably low mean absolute error while exhibiting a relatively high R-squared value which should indicate a strong ability to accurately predict housing prices. Given such a model, we will be able to predict housing prices across various regions in order to identify potential investment opportunities.

Our analysis directly impacts various stakeholders in the real estate industry such as individuals, investors and even government bodies. The goal of the analysis is to provide these stakeholders with the knowledge to make better informed, strategic decisions regarding the real estate market. This will ultimately lead to benefits such as increased profitability, risk mitigation and improved operational efficiency.

Once we are able to cleanse and wrangle the data, we will be able to test the code using multivariate linear regression analysis, and create scatter plots in R or Tableau. The closer the r^2 value is to 1, then that represents a positive correlation, and vice versa. We want our data visualizations to tell a story in order for the audience to understand which variables played a profound role in housing prices. Once we are able to get the r^2 values, we will be able to support or disprove the hypothesis based on which variables we thought would have a positive or negative correlation on the housing prices.

Analysis Performance

To begin, the data was reduced to include data points between July 2016 to May 2023, and counties with greater than 2 million in population. This approach was selected to narrow our scope and compare metropolitan regions. Secondly, we randomly sampled the data into training, validation, and testing subsets: 60%, 20%, 20%, respectively.

The first selection of models built and compared were a combination of simple and multivariate linear regressions. Using a single or a combination of the independent variables in the model, the models were first compared on fit and significance using adjusted R-square and p-values. The independent variables were mortgage interest rates, CPI, active listings, and population, the value (price) was the dependent variable, and county was an indicator variable for regional segmentation.

Comparing the simple linear regression models, a key takeaway is that each model has high fit rates, greater than 0.85 R-square, with CPI having the highest at 0.96. The population only model had a p-value of 0.19, therefore it was excluded from additional consideration of model selection. One multivariate linear regression model was built, which includes all four dependent variables, the model showed evidence of heteroscedasticity, and so, to adjust, log transformation of the value (price) was applied. The result was that each independent variable, sans county indicator variable, is statistically significant ($p=<0.05$), including the population. The change of population from statistical insignificance to statistically significant was unexpected shows that population has collinearity with at least one of the other variables. Lastly, mortgage interest rates are seen as the least statistically significant variable in this model (p-value = 0.0065).

Using the top 3 models by adjusted R-square and statistical significance, the validation dataset was used to evaluate the models' performance via calculation of the root mean squared error (RMSE). Between those models, the model that included all 4 variables (mortgage interest rates, CPI, active listings, population) performed with the lowest RMSE:

1. CPI only: 40765.98
2. Active listings only: 53172.70
3. All variables: 36906.55

Finally, the selected model was run with the test dataset. From that, the RMSE returned 35672.95, which is similar to the validation output. Between the model's R-square and the RMSE, this is a good result at this current stage in the project.

Problems we encountered

While data gathering and performing feature selection, we started with too many features. The housing industry has so many metrics to choose from, it took the team a good amount of time to determine which were appropriate to complete our objective. We were able to decide on which features by narrowing our scope around the dependent dataset and macro-economic factors as outlined in the problem statement.

A current problem that we are dealing with is determining further which variables to include. When we ran a correlation matrix, mortgage interest rates and CPI are strongly correlated and, so, the current discussion is if we should remove one to simplify.

A second potential problem we have is the different magnitudes of the variables. We have so far used the raw data numbers as is, however, there may be improvements in the model if we normalize the scale of the data across the board.

Next Steps

Based on our findings thus far, besides choosing the correct model, we need to determine how to best present our data. Since our complete data set only contains 6 years of monthly data points, we need to take into account that the model may not accurately be representative of market cycles which occur over a longer time scale. Additionally, the population dataset stops at December 2022 while the other data continues into May 2023. Next steps would include determining if we should impute the population data using an ARIMA time series model or filter the rest of the data through December 2022. This then leads to the next issue of choosing the correct model.

Initially, we had planned to use a linear regression model to predict price. However, as shown above, larger predicted values lead to larger variance which suggests heteroscedasticity and nonlinear data. This is one of the inherent limitations of using a linear model, as it operates under the assumption of linearity and having data points that are independent of each other. To remedy this, we used a log transformation on the dependent variable. This means we will have to alter our interpretation of the model as the relationship between the variables is no longer linear. Also, since the data consists of time series, we will explore the use of AutoRegressive Integrated Moving Average (ARIMA) models, Gradient Boosting Machines, and Random Forests. We will then choose the best model based on both its predictive performance and interpretability.

One other aspect that we will have to address is the issue of multicollinearity in the underlying data. As mentioned previously, the US 30 year average mortgage rate is highly correlated with CPI. Whereas, active listings is negatively correlated with CPI. In our preliminary

discussions, we have entertained the idea of removing CPI as a predictor. The reason being that CPI already takes into account housing data as part of the underlying figure and since it is already correlated with mortgage rates it is seemingly extraneous data.

To summarize, we will have to focus on the issues of confined data, imputation, model selection, and multicollinearity. As we get closer to our final model, it is important that we shift focus to these areas as they will ultimately define our success in being able to predict housing prices using macroeconomic indicators.

Survey 3 Sources

When delving into the body of research related to our topic on the Housing Market, we came across numerous publications attempting to solve the problem of forecasting housing prices. Through the wealth of knowledge already available we were able to refine our focus and steer clear of the common pitfalls associated with the type of analysis we are pursuing. In our investigation we came across several research articles outlining not only what to avoid during our data discovery phase, but also which models might prove best for what we are trying to achieve.

The first source we examined, "Forecasting EREIT Returns", looked at the role of financial assets, direct real estate, and the Fama and French factors in explaining Equity Real Estate Investment Trust (EREIT) returns. The results showed that neural networks with multiple factors yielded the best predictions. The second source, "Forecasting Real Estate Prices", addressed the difficulty in predicting changes in real estate prices. It discussed the challenges inherent in constructing a reliable model for the real estate market such as heterogeneity, high transaction costs, and illiquidity of the underlying asset. They also noted that REITs show the highest predictability when looking at yearly horizons. The third source, "Housing Prices and Inflation", analyzed the relationship between housing prices and the Consumer Price Index (CPI). It was shown that the relationship often exhibits a lag and is not a direct 1:1 correlation.

Overall, these studies contribute to the body of research as it relates to the real estate market by examining different variables, models, and forecasting methods. Additionally, they highlight the challenges of constructing reliable models due to limited time-series data, long time horizons, and illiquidity of the overall market.

Works Cited

Serrano, Camilo, and Martin Hoesli. "Forecasting Ereit Returns." *Taylor & Francis Online*, 18 June 2020, www.tandfonline.com/doi/abs/10.1080/10835547.2007.12089784.

Ghysels, Eric, et al. "Forecasting Real Estate Prices." *Handbook of Economic Forecasting,* Elsevier BV, 2013, pp. 509–80. https://doi.org/10.1016/b978-0-444-53683-9.00009-8.

Bernstein, Jared, et al. "Housing Prices and Inflation." *The White House*, Sept. 2021, www.whitehouse.gov/cea/written-materials/2021/09/09/housing-prices-and-inflation.