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

Project Final Report

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 20, 2023

Table of Contents

Overview of Project

Background

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.

Research

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. There are substantial factors that have caused real estate returns: GDP, inflation, short-term interest rates, dividend

yields, and price-earning ratio. Other factors include stocks and bonds as well. These factors are instrumental when impacting supply and demand, which also impact asset prices. However, there is no best predictor variable when identifying the increase in housing prices and thus, more of a combination of those predictor variables.

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 along with how monetary policy is used in order to influence real estate prices. They also noted that REITs show the highest predictability when looking at yearly horizons. There are certain predictors among valuation ratios such as rent-price and income-price, which is able to demonstrate the real estate returns, which are a return on an investment of real estate property. Other variables like construction costs and regulatory restrictions also play a huge role in determining proper real estate returns.

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. It highlighted throughout the pandemic other variables which have caused a huge increase in housing prices. The long term increases were due to restrictions on local zones and affordable housing, but the overall increase was exacerbated by other factors. Some of these factors showed that in May 2021, the price of lumber increased by 114% over one year and the prices of iron and steel increased by 73% the past one year. However, those prices have declined recently and now we see that housing prices have come down a bit. Additionally, the source dives deeper into the relationship between housing prices and CPI, where it shows that rent prices show homes that are both on the market and for homes that are already occupied, so there is less volatility. However, there can still be issues when it comes to housing prices due to supply constraints or shortages in the housing market.

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.

Overview of Data

Data Exploration

Housing prices are influenced by multiple factors, and to understand their impact, we have carefully selected four independent and one dependent datasets from reliable sources:

30-Year Fixed Rate Mortgage Average in the United States (Federal Reserve Economic Data): This dataset provides weekly average interest rates for 30-year fixed-rate mortgages in the United States since 1971. Analyzing this data allows us to assess the relationship between interest rates and housing prices over time.

Adjusted Consumer Price Index (CPI) for All Urban Consumers (U.S. BUREAU OF LABOR STATISTICS): The CPI is a measure of the average change over time in the prices paid by urban consumers. The dataset provides monthly CPI data for the United States since early 1973. Analyzing this information allows us to understand the inflation experienced by the urban consumers in their living expenses.

Housing Inventory: Active Listing Count in the United States (Federal Reserve Economic Data): This dataset provides information on active single-family, condo and townhouse listing count in the United States since July 2016 on monthly level. This housing

supply data will have us better measure its effect on housing prices.

Annual Estimates of the Resident Population by Counties in the United States (Census): This dataset provides annual population estimates by Counties and States between 2010 and 2022. By analyzing this, we can gain insights on how demographics and populations can impact housing prices.

Zillow Home Value Index - US Metro (Zillow Research Data): This dataset provides typical value for Single-family, condo/Co-op homes in the 35th to 65th percentile range. This will be our dependent data to be used in modeling.

Initial Hypothesis

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.

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.

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.

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

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.

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.

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 Million 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.

Overview of Modeling

Methodology/Approach

We expected our analysis to produce results in line with expert opinion. Additionally, we projected that the final model would 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 would then be able to predict housing prices across various regions in order to identify potential investment opportunities.

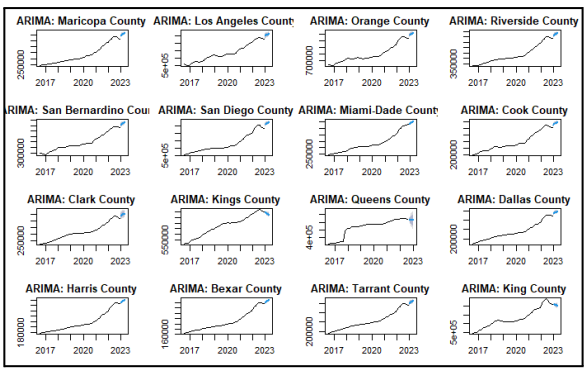
Our approach started with an exploratory analysis of the various data sources we were able to pull together. We were able to gain insights into the data and identify potential features that would influence housing prices. Pulling together data from the vast amount of sources found online required a significant amount of cleaning and pre-processing. We normalized various data sets, as well as one-hot encoded categorical variables such as counties.

Once we selected our variables, we split the data into training, validation, and testing sets. From there we fit multiple models using various methodologies such as autoregressive integrated moving average (ARIMA), linear regression, random forest regressions. We chose to pursue an ARIMA model due to its ability to capture trends and seasonality while also being best suited for handling time series data. Multiple linear regression was also included in our analysis as it is able to provide a proper benchmark from which to compare the other two models. Linear regression also has the added benefit of interpretability of the resulting coefficients. Additionally, we ran multiple random forest regressions which were able to capture more complex, non-linear relationships between our independent and dependent variables. To optimize our models, we used various R packages that autotune the hyperparameters, specifically for ARIMA and random forests. To evaluate the different models, we primarily used the root mean square error. We used the RMSE and adjusted R squared along with several visualizations to determine the model with the best overall performance. This multi-model approach allowed us to take the best characteristics from each to further provide a more comprehensive view of the real estate market and the factors that influence it.

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.

ARIMA

Considering that we were using time series data, it was fitting to start our analysis with an AutoRegressive Integrated Moving Average (ARIMA) model to forecast prices in the future. Due to the nature of using an ARIMA, we decided to segment the data by county for which we would create independent models. Each model utilizes independent variables including the 30 year average mortgage rate, Consumer Price Index, number of active listings, and the respective counties population with the dependent variable being the average monthly price.



Our data set originally spanned from July 2016 to December 2022 due to missing 2023 data for population. As a result, we forecasted the population numbers for January to May of 2023 using a basic ARIMA model for each county. These results were then appended to the working data set for use in the forecasting model. First we grouped the data by county and used the auto.arima

function in R to create our models for 2016 through 2022. Then we forecasted the prices in the first 5 months of 2023 using the forecast function. This allowed us to then create a dataframe of the Root Mean Square Error (RMSE) of the forecast models for each respective county. As is shown in the table below, the ARIMA models performed rather well as they are best suited for handling time-series data while accounting for trends and seasonality. Additionally, the ARIMA models allow us to include external variables which are crucial to forecasting price as it is inherently influenced by factors such as mortgage rates, CPI, population, and number of active listings.

ARIMA	County	RMSE
1	Bexar County	11098.501
2	Clark County	43131.538
3	Cook County	10363.172
4	Dallas County	21283.578
5	Harris County	12709.112
6	King County	4875.126
7	Kings County	20088.195
8	Los Angeles County	77080.477
9	Maricopa County	53606.172
10	Miami-Dade County	11728.835
11	Orange County	85898.414
12	Queens County	5824.351
13	Riverside County	49861.279
14	San Bernardino County	45027.62
15	San Diego County	75654.827
16	Tarrant County	23099.808

For the first approach at modeling housing prices, we were satisfied with the results for certain counties such as King and Queens County which had relatively low RMSE's. Through the comparison of predicted and actual values we were able to measure the performance across counties as well as across models as will be demonstrated below. Overall, our findings suggest that ARIMA models are useful for forecasting housing prices when segmented by county and will prove useful for informing strategic decisions related to the housing market. Additionally, the results provide a foundation for further research using similar and more refined methods.

Linear Regression

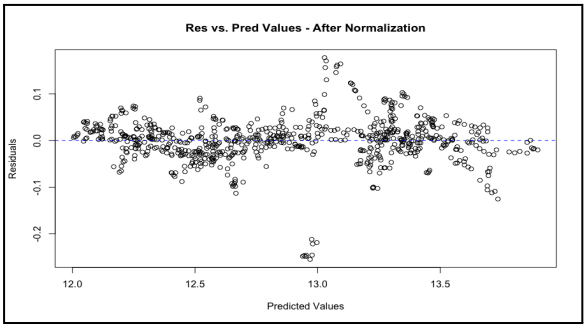
Linear Regression Model is the second approach we used to forecast housing prices. We have created a combination of simple and multivariate linear regressions. Using a single or a combination of the independent variables in the model, the models were first compared on fit and significance using adjusted 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.

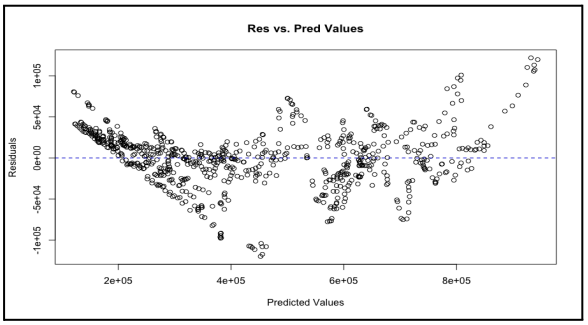
Comparing the simple linear regression models, a key takeaway is that each model has high fit rates, greater than 0.85 R-square, with CPI having the highest at 0.967. The population only model had a p-value of 0.19, therefore it was excluded from additional consideration of model selection. One multivariate 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).

To evaluate the performance of the top three models based on adjusted R-square and statistical significance, we utilized the validation dataset to calculate the root mean

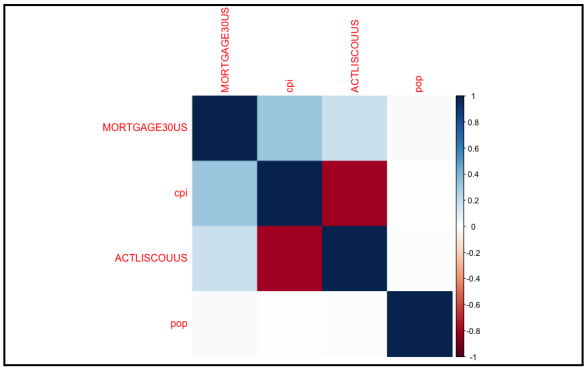
squared error (RMSE). The RMSE for the CPI only validation set was 35,578.5, for the Active Listing only validation set was 55,894.08, and for the model with all four variables, it was 34,678.08, indicating the lowest RMSE.

Based on the RMSE comparison, we divided to proceed with the model including all variables and ran it again with the test dataset, it was showing an RMSE of 32,516.82. This result demonstrates a good output of selected models for forecasting housing prices.

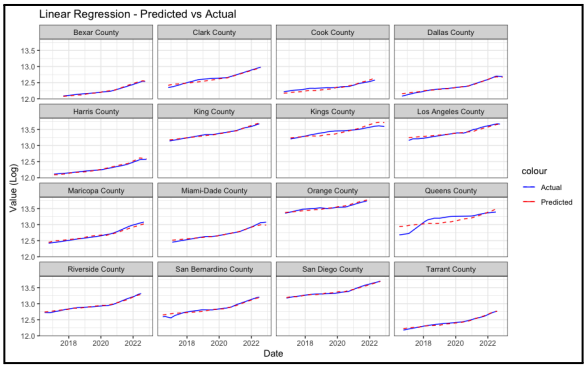




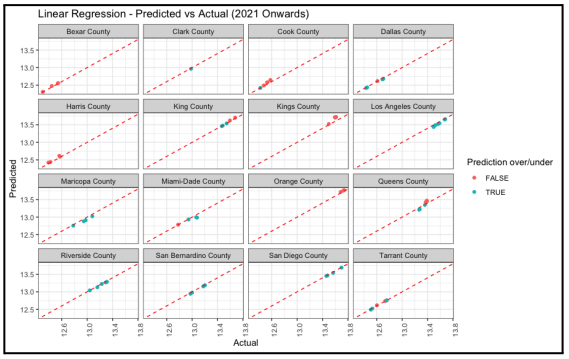
The next step we were going to assess the model's performance by comparing the residuals vs. predicted value plot. The initial plot suggested the model is heteroscedasticity and non-linear. To address those issues, we applied a Log-Linear transformation to the dependent variable. This transformation helps to normalize the residuals and improve the model's performance. As a result, the updated plot shows a much more satisfactory fit of the model to the data.



Additionally, we conducted a correlation matrix analysis to explore the relationship between different variables in the dataset. The results indicate a high positive correlation between mortgage rate and CPI. On the other hand, there is a negative correlation between active listing and CPI, indicating an inverse relationship between these variables.



Finally, using the selected Linear Regression model, we created a plot to compare the predicted and actual housing values by counties, providing valuable insights into the overall trend.



	lm_pred_df_filtered$county	count	mean_value	median_value	min_value	max_value	pred_over	pred_under
1	Bexar County	4	−0.0169369344	−0.0135118631	−0.0320698703	−0.008654141	0	4
2	Clark County	1	0.0165803531	0.0165803531	0.0165803531	0.016580353	1	0
3	Cook County	5	−0.0348011742	−0.0459102155	−0.0664382929	0.001710936	1	4
4	Dallas County	6	0.0149052333	0.0177168566	−0.0065201976	0.023389185	5	1
5	Harris County	4	−0.0309127899	−0.0260903295	−0.0535581301	−0.017912370	0	4
6	King County	5	−0.0085371015	0.0040870672	−0.0293147149	0.005899541	3	2
7	Kings County	5	−0.0943353794	−0.1089692775	−0.1206796087	−0.025454011	0	5
8	Los Angeles County	6	0.0363844546	0.0376398928	0.0167065309	0.060682351	6	0
9	Maricopa County	4	0.0542432294	0.0621390565	0.0245064637	0.068188341	4	0
10	Miami-Dade County	4	0.0410431165	0.0388969390	−0.0059458714	0.092324459	3	1
11	Orange County	3	−0.0288799142	−0.0285740187	−0.0310711517	−0.026994572	0	3
12	Queens County	7	−0.0181682518	−0.0502872234	−0.0806420557	0.060169206	3	4
13	Riverside County	5	0.0172315932	0.0238100772	0.0008690043	0.031266482	5	0
14	San Bernardino County	5	0.0193594397	0.0183033286	0.0139081617	0.026417574	5	0
15	San Diego County	4	0.0132804053	0.0090440235	0.0081768145	0.026856760	4	0
16	Tarrant County	6	−0.0007653894	−0.0006228031	−0.0154468743	0.013083935	3	3

To identify potentially undervalued counties, our analysis will focus on the most recent data from 2021 onwards. The results revealed that several counties consistently showed predicted values higher than the actual value. Among them Maricopa County, Miami-Dade County, and Los Angeles County stood out with the highest mean difference, suggesting a promising investing opportunity.

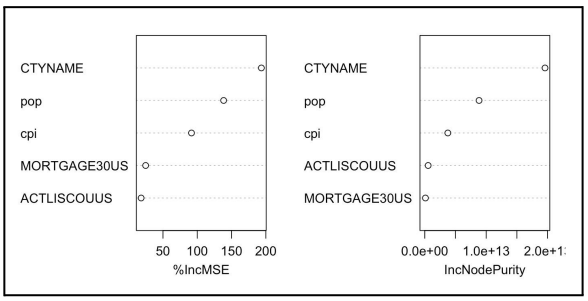
Random Forest

The last approach taken was using random forest regression. The parameters chosen include 500 trees ('ntree' = 500) and 5 random variable choices at each decision ('mtry' = 5). These were selected using the best common industry practices. 500 trees allows for a good balance between performance and accuracy and 5 'mtry' is roughly the square root of the max number of features which allows enough decisions for the model to make while allowing room for randomness without overfitting.

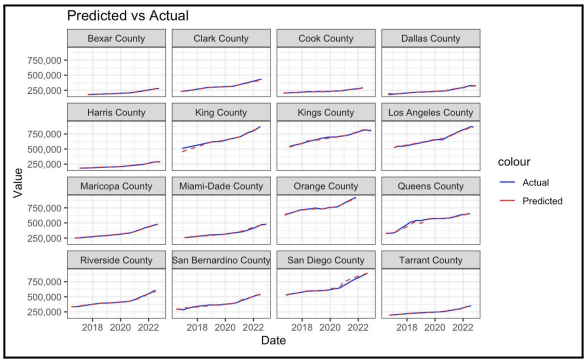
```
Call:
 randomForest(formula = value ~ MORTGAGE30US + cpi + ACTLISCOUUS +
pop + CTYNAME, data = train_set, importance = TRUE, mtry = 5)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 122505534
                    % Var explained: 99.72
```

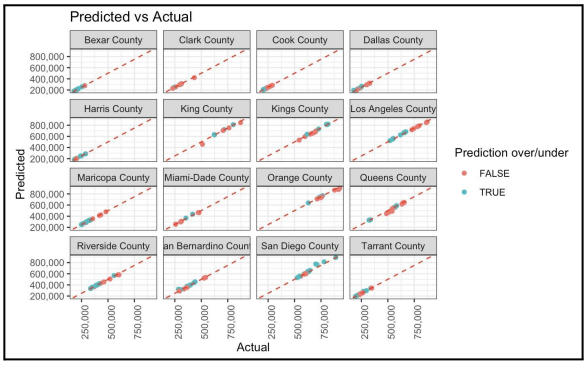
The random forest model performed extremely well. The variability explained is at 99.72% for the training set. To test if the model is overfitted, the same test set that was applied to the linear regression was used here in the random forest to calculate the RMSE; the output equaling 11,629.

Between the features, it shows here that counties, population, and cpi are the main drivers of the model, in that order. In this model, the mortgage interest rates and the active listings have much lower impact as seen under both % increase in mean square error and the increase in node purity.



As expected due to the % variance explained, when plotting the actual observation points and the prediction observations, the plots are aligned closely. However, when comparing the actuals vs predicted, it shows that there is a different amount of variance by county.





To compare counties, a difference was taken between the prediction value and the actual value (prediction minus actual). A positive output means that the home's value is predicted higher than the actual value and is currently undervalued. The opposite is true that a negative output means the prediction is lower than the actual and that the house is overvalued, according to the model. Median value was used to lower risk of outliers. San Diego County has the highest median value at $4,718, then Los Angeles at $3,377, and then King County at $2,600, while Queens County has the lowest median value at -$4,913, meaning the median home value is predicted to be lower than the actual value by $4,913.

	rf_all_pred_df$county	count	mean_value	median_value	min_value	max_value	pred_over	pred_under
1	San Diego County	21	11300.00112	4717.5755	-5037.5024	74214.242	15	6
2	Los Angeles County	14	826.15827	3376.8984	-17596.4439	17600.329	8	6
3	King County	9	-6701.05757	2599.8144	-53264.0468	9047.057	5	4
4	Harris County	13	1986.14593	1645.2641	-904.2441	7448.109	11	2
5	San Bernardino County	18	2667.65818	1056.9242	-11340.1312	38775.555	9	9
6	Bexar County	13	240.89515	870.3391	-4645.4132	4786.791	7	6
7	Riverside County	16	-1648.44039	830.4989	-22543.6052	10038.985	9	7
8	Tarrant County	16	578.05143	201.4488	-6012.4427	15209.153	8	8
9	Maricopa County	20	-10.67972	129.6056	-7892.4490	6483.969	11	9
10	Clark County	10	-950.95655	-407.1736	-13567.0278	4751.856	4	6
11	Dallas County	16	951.60182	-1537.3895	-8805.2512	17333.838	7	9
12	Cook County	24	-1613.10781	-2175.9182	-5802.0162	2854.686	7	17
13	Kings County	13	-1509.01903	-3428.7669	-17680.4431	19623.162	6	7
14	Miami-Dade County	11	-1367.65964	-3494.7717	-14088.6830	18474.213	2	9
15	Orange County	13	-7611.95586	-4017.4992	-37794.4762	8213.372	5	8
16	Queens County	21	-9989.51417	-4913.3425	-50356.4052	6570.387	7	14

RF	County	RMSE
1	Bexar County	2878.254
2	Clark County	5399.084
3	Cook County	2464.15
4	Dallas County	4603.31
5	Harris County	2484.828
6	King County	9874.904
7	Kings County	7113.504
8	Los Angeles County	9669.148
9	Maricopa County	5903.896
10	Miami-Dade County	5201.609
11	Orange County	9377.261
12	Queens County	9770.774
13	Riverside County	7510.469
14	San Bernardino County	8184.319
15	San Diego County	10483.422
16	Tarrant County	3940.105

Discussion

Based on our analysis, it appears that the random forest model outperforms the ARIMA and multiple linear regression models when it comes to predicting real estate prices using macroeconomic and demographic factors. This was evident from the lower average RMSE of 6,553.69 as opposed to the ARIMA RMSE of 34,458.19. Additionally, the random forest regression was able to account for over 99% of the variance in the model. While this shows high performance in terms of making accurate predictions, it may also be an indicator of overfitting. The results from the random forest regression may be attributed to its ability to capture complex non-linear relationships like those found in the real estate market. Furthermore, the random forest model was able to provide valuable insights on a county-by-county basis. The individual county models were able to identify potential opportunities for investment by comparing predicted and actual prices. The counties with overestimated prices could signal that prices are currently undervalued.

Although the random forest model performed the best, the ARIMA and multiple linear regression models were able to provide valuable insight as well. The ARIMA model, for instance, was able to provide highly accurate predictions for certain counties which may lead to areas of further exploration and analysis. On the other hand, the linear regression model provided the interpretive framework for being able to quantify the impact of each independent variable.

In conclusion, our approach has yielded favorable results and valuable insights into the dynamics of the real estate market. We have clearly shown that we are able to make accurate predictions in the short-term at the county level using macroeconomic indicators such as mortgage rate, cpi and active listings in conjunction with demographic data such as population. The predictive power of the random forest along with the strengths of the ARIMA and linear regression models will assuredly be able to provide various stakeholders with the tools

and knowledge to make better informed, strategic decisions regarding the real estate market. Further research is recommended to explore the utilization of more advanced models with additional variables, adjusting the time frame to a wider view, and expanding the geographic scope of our analysis. This will ultimately allow for better predictive capabilities and more valuable insights for practical applications.

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.