Seeing into the Future: Using Time Series Forecasting to Predict Housing Trends

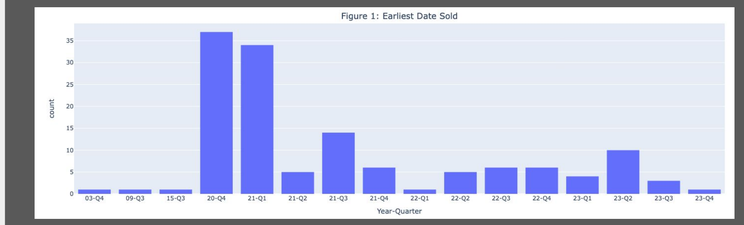
María José del Granado (Mary Jo) ‘24 Data Science Major Capstone
Advisor: Eni Mustafaraj

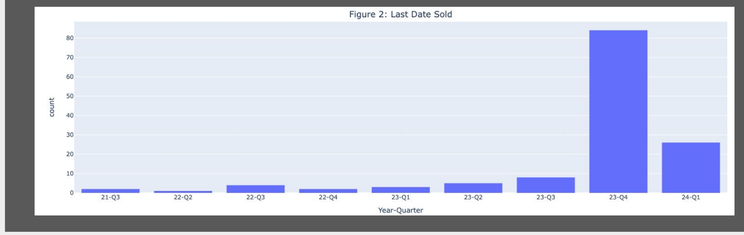
Research Question

To what extent can ARIMA and SARIMA models accurately forecast average prices for different cities in the Boston Metropolitan Area?

Data Collection

In order to perform this study, I scraped data on houses in the recently sold section off of Zillow. I focused on houses within the Boston Metropolitan Area—meaning homes that were within Middlesex, Essex, Norfolk or Suffolk county. The full dataset has 72,461 rows and 4325 columns, although for my analysis I will only work with three variables: date sold, price and city. Once I calculated the average price per city, my dataset was reduced to 1235 rows and 3 columns. However, the starting periods and final periods of my data weren’t the same. So, I picked both starting and final periods that maximized the number of rows within my dataset: 2020-Q4 to 2023-Q4 (See Figures 1-2). In the end, the data I used for my analysis consisted of 510 rows and 3 columns. In terms of splitting my data for training and testing, I labeled data up until period 2023-Q4 as training, and the last four periods as testing data.





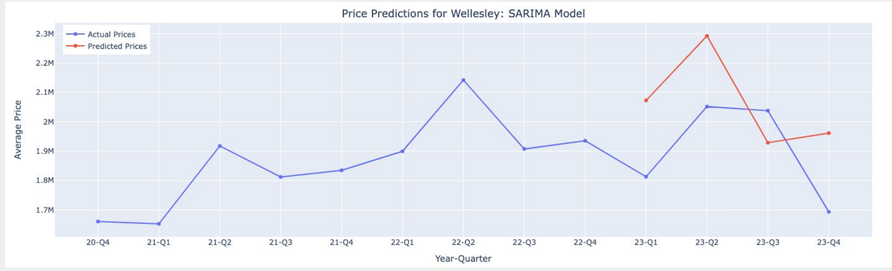


Table 1. RMSE by Model

Model	RMSE Minimum	RMSE Average	RMSE Maximum
ARIMA	20,528 (Quincy)	173,809	767,497 (Dover)
SARIMA	56,967 (Lowell)	329,986	1,204,566 (Westwood)

Model & Analysis

I fit two different autoregressive models: ARIMA and SARIMA. Both of these models predict future values based on past values, the main difference being that SARIMA takes seasonality patterns into account. I used both models to predict four different average prices.

Model 1: ARIMA (p,d q) = ARIMA(2,1,4)
Model 2: SARIMA (p, d, q)(P,D,Q)m = SARIMA(2,1,1)(1,3,2)4
where:
p = number of autoregressive terms
d = degree of differencing
q = order of moving average
P: Seasonal autoregressive order.
D: Seasonal difference order.
Q: Seasonal moving average order.
m: The number of time steps for a single seasonal period.

ToDo: hyperparameter tuning, remedial measure, dataset

Discussion and Conclusion

My results suggest that predicting the average sale price of a home, merely on historical data isn’t sufficient to get accurate predictions. After all, there several different factors that play a role in determining the home sale prices, that I don’t use. However, at the same time my models are limited by the fact that the data that I collected doesn’t go that far back in time. It could very well be that better performance could be achieved, if I had more data.