California Housing

Freddie Mac

EXECUTIVE SUMMARY

Freddie Mac

In order to better serve its customers, it’s essential for FreddieMac to have a pulse on the housing market and be able to predict the rise and fall of housing prices.

The problem arises when predictive models struggle to capture anomalies, most recently seen during the COVID-19 pandemic. With a fall in employment, GDP, and other trustworthy factors, models predicted that housing prices should have fallen. They didn’t.

This called into question the robustness of the predictive models in use.

We believe that there is a need to examine:

What other factors can be used to increase model stability?

Can modern machine learning models improve predictive accuracy?

DATA HIGHLIGHTS

SOURCES

State of California Department of Finance	Labor Force & Job Numbers Inflation & Consumer Price Index
Federal Housing Finance Agency	Monthly Average Loan Amounts & Interest Rates for Single-Family Homes
Freddie Mac’s Mortgage Interest Rates	30-Year Fixed Mortgage Interest Rates 15-Year Fixed Mortgage Interest Rates
Federal Reserve Economic Data	New Private Housing Units Authorized by Building Permits for California
Bureau of Economic Analysis	Annual Gross Domestic Product by State
Google Trends	Housing Interest Index

DATA HIGHLIGHTS

DESCRIPTION AND CLEANING

FEATURES
15 TOTAL INDICATOR COLUMNS

DATA POINTS
216

YEARS ANALYZED
TRAIN: 2005 TO 2019
TEST: 2020 TO 2022

DATA CLEANING

NULL VALUES
Removed rows with consistently null values

DATA FILTERING
Cleaned excess data columns, disregarded columns with yearly data, filtered out California data

TIME RANGE SLICING
Sliced data to 2005-2022 time periods

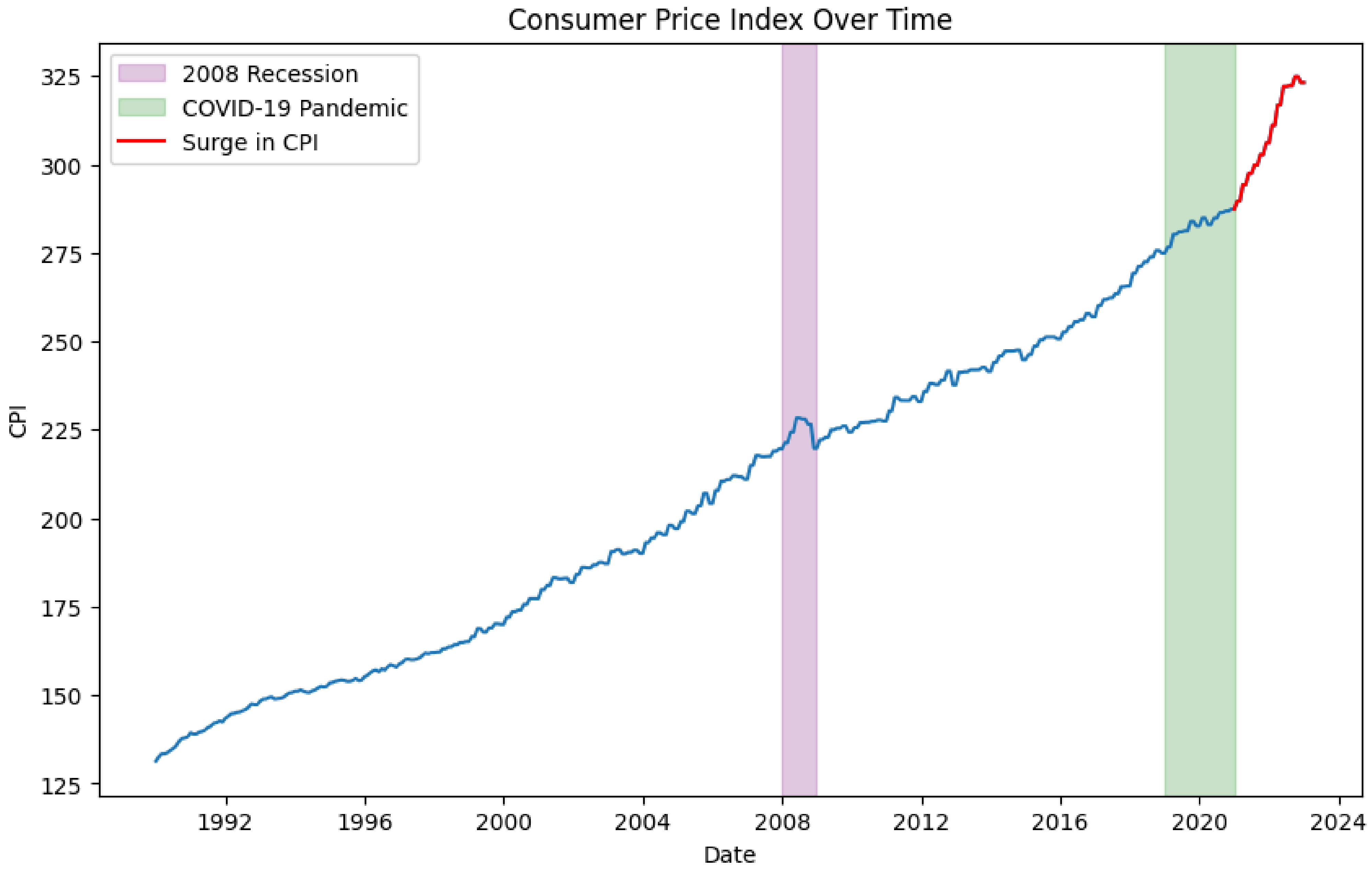
DATA PREPROCESSING

AGGREGATION
Converted data with weekly values to monthly values

IMPUTATION
Filled empty values with imputed estimates using percieved trend values

DATA TYPECASTING
Converted float & string values to Integers

EXPLORATORY DATA ANALYSIS

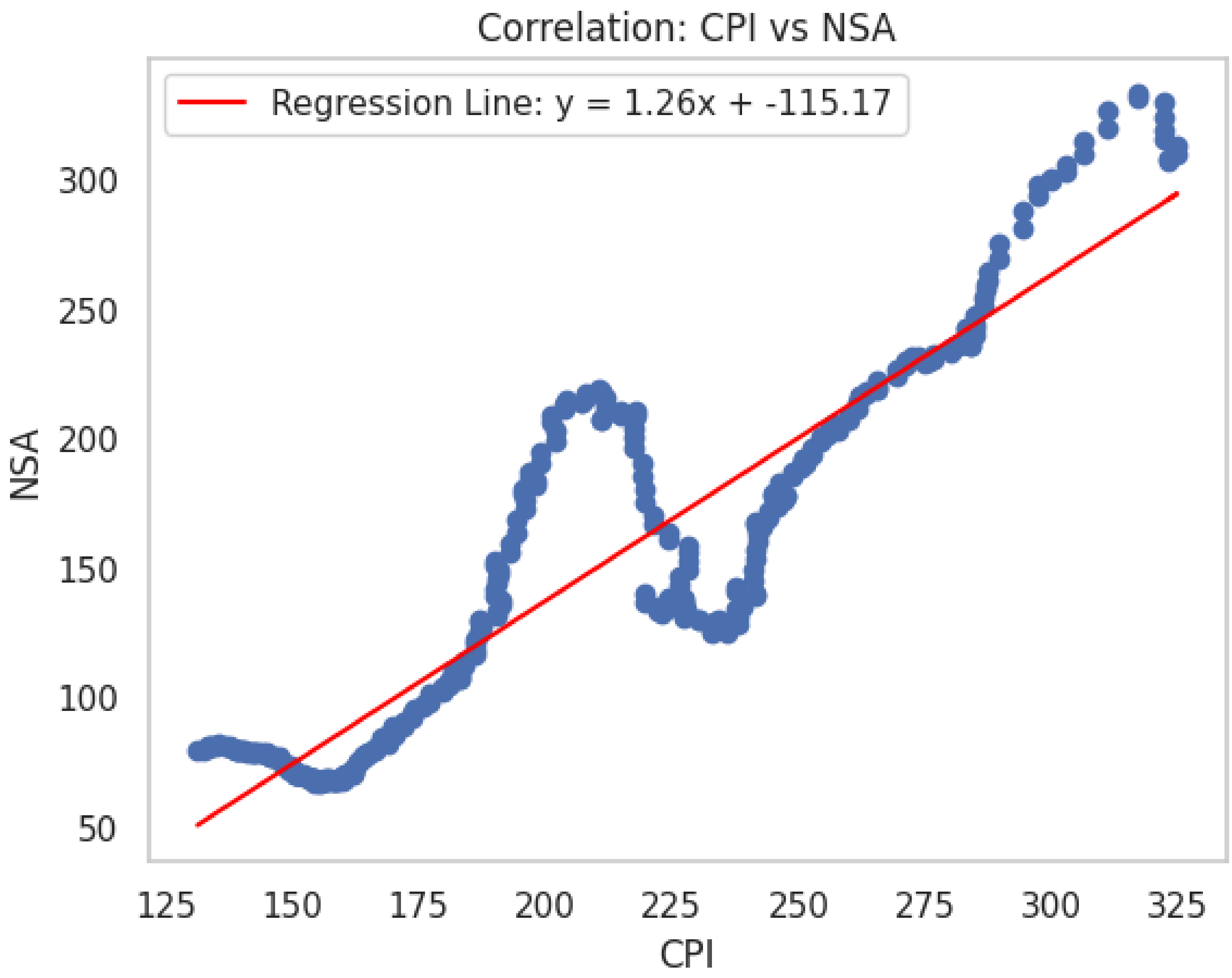


- **2008-2009 Downturn (Falling CPI)**: Reflects deflationary pressures amid the global financial crisis.

- **2019-2021 Stability (Constant CPI)**: Signifies an economically steady period with moderate inflation, providing a baseline for consistent housing market trends.

- **Post-2021 Inflationary Surge (Rising CPI)**: Indicates sudden CPI rise, highlighting inflationary pressures.

DOES CPI HAVE AN IMPACT ON NSA HPI?



NSA=1.26×CPI−115.17

- This indicates a positive correlation between CPI and NSA housing price index.
- CPI being the most important feature to predict the Housing Price Index.

ARIMA MODEL

- Excels in forecasting time-dependent data, capturing historical patterns for future predictions.

- Adjustable parameters make ARIMA versatile, fitting various datasets by tweaking autoregressive, differencing, and moving average orders.

- Used Random Forest as it excels in handling various types of features, dealing with multicollinearity, and providing insightful feature importance scores, making it a robust choice for ARIMA feature selection.

- ARIMA's interpretability and diagnostic capabilities make it a reliable choice for forecasting housing price indices over time.

- ARIMA models can be extended to SARIMAX (Seasonal ARIMA) to incorporate seasonality in the data.

MODEL COMPONENTS

ARIMA

Orders for Arima

- p: Autoregressive (AR) order - The number of lag observations included in the model.
- d: Integrated (I) order - The number of times that the raw observations are differenced (made stationary).
- q: Moving Average (MA) order - The size of the moving average window.

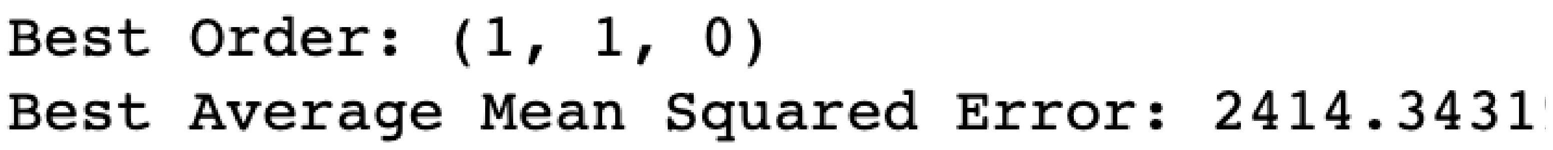
Orders for Sarimax

Has same orders as Arima with an additional order for seasonality.
s: The number of time steps in a seasonal period.

Order Selection

- Model performance varies drastically with the selection of orders.

- Orders were selected by training both the model multiple times using different values of orders and the best performing order values were finalized.

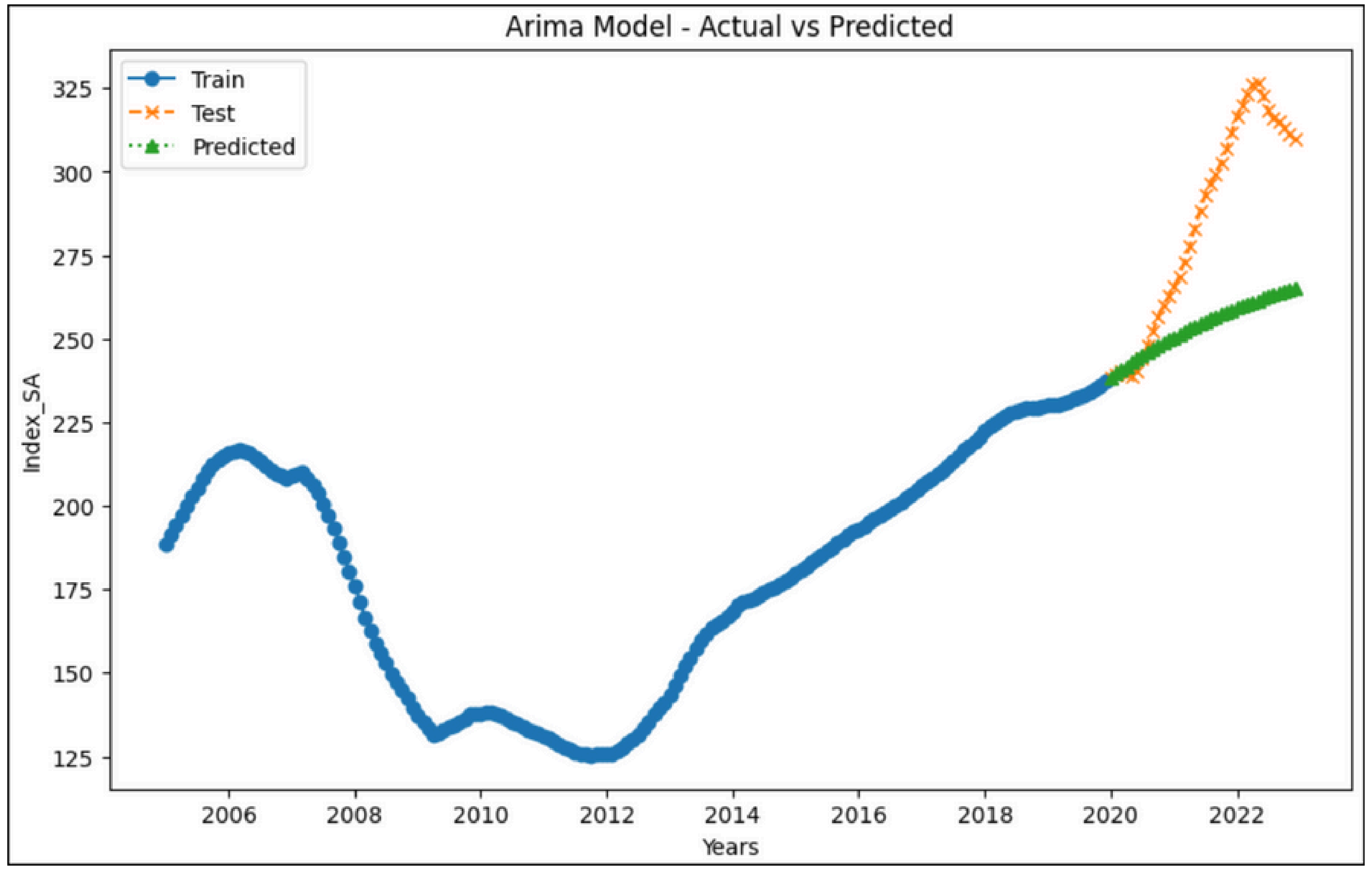


```
Best Order: (1, 1, 0)
Best Average Mean Squared Error: 2414.3431
```

PREDICTIONS

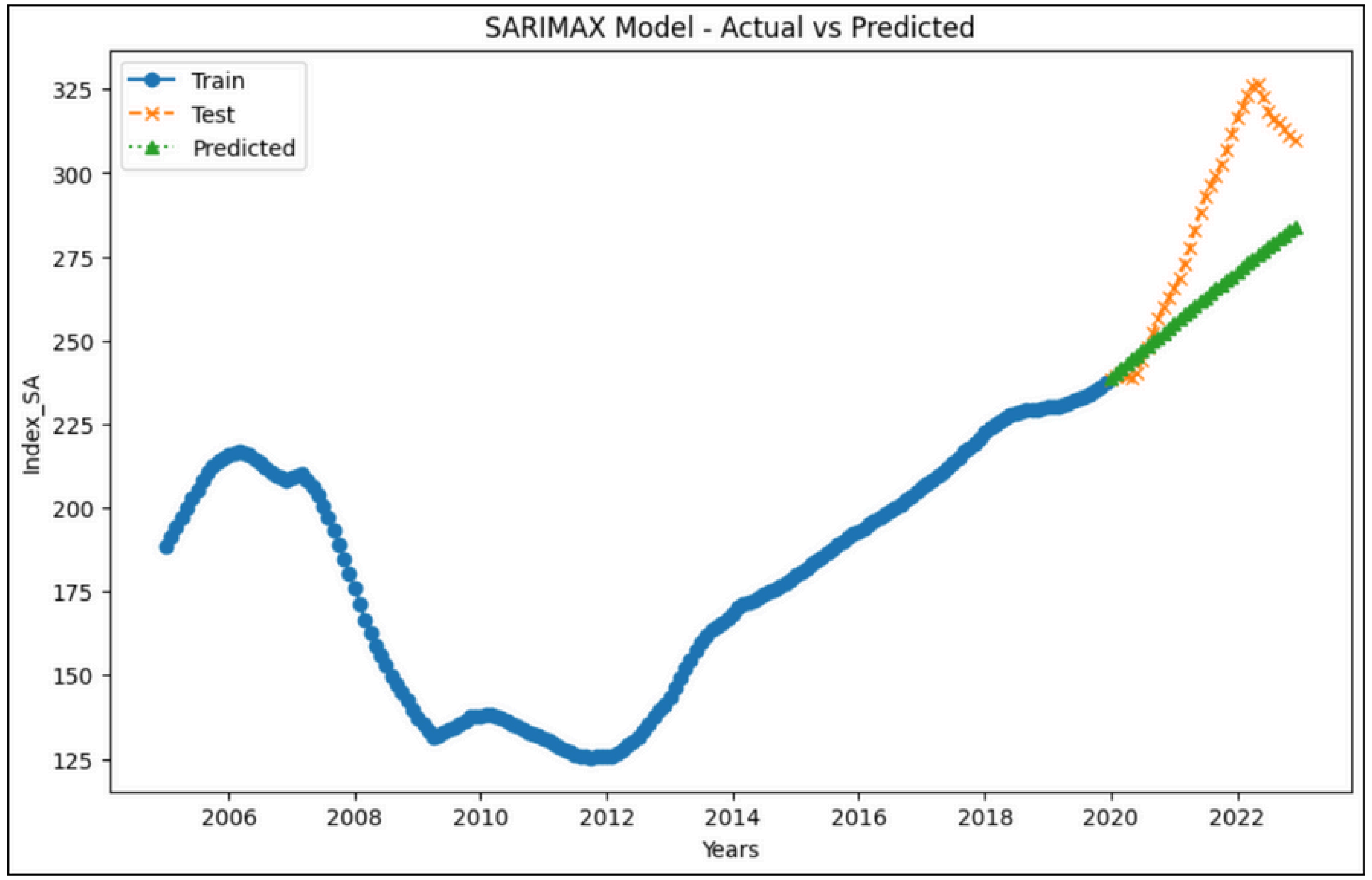
ARIMA VS. SARIMAX

ARIMA



RMSE = 49.13
MSE = 2414.34

SARIMAX



RMSE = 29.27
MSE = 857.12

Features Selected: CPI, Employment, Unemployment, Newly built homes - effective interest rates

SARIMAX performed better than ARIMA

RECOMMENDATIONS

- *Explore Alternative Modern Models*

  After implementation of machine learning and Neural Network models, there is evidence that these models are worth looking into with tuning and domain knowledge.

- *Looking Back is the Way Forward*

  Models were seen to benefit from the introduction of lag in selective factors. With iterative testing, further improvement may be achieved.

- *Contextual Sentiment Model*

  Allocate resources to train NLP models on housing market texts for more accurate sentiment classifications.

CHALLENGES

- *Data Availability*

  To predict trends accurately, features like customer profiles, socio-economic factors like salary, census data, and social media trend analysis data, and MSA level data can be more insightful.

- *Data Accuracy & Consistency*

  Multiple sources create a consistency issue in the data related to the data time frame, values, and formats.

- *Bias in Sentiment Analysis Models*

  Selection of news sources significantly affect classification of text, hence building a credible corpus can be difficult.