House Index Forecasting

Maria Shiferaw

May 2024

Abstract

This paper explores forecasting the U.S. House Price Index (HPI) utilizing advanced statistical and machine learning methods, including ARIMA and LSTM neural networks. By examining historical data spanning from 1991 to 2024, we demonstrate our approach's efficacy in predicting future trends and provide insights into the dynamics influencing the real estate market. This study integrates traditional time series decomposition and modern predictive modeling, aiming to offer a robust tool for economic analysis and investment decision-making.

Introduction

The U.S. House Price Index (HPI) serves as a critical economic indicator, reflecting the health of the real estate market and influencing both macroeconomic policies and individual investment decisions. Accurate predictions of HPI movements are essential for stakeholders, including policymakers, investors, and financial institutions. This paper details the application of both ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory) models to forecast HPI, providing a comprehensive view of methodological strengths and limitations inherent to time series forecasting.

Data Cleaning

The data cleaning process involved several critical steps to prepare the dataset for robust statistical analysis and predictive modeling. The raw data was initially loaded from a CSV file containing historical monthly U.S. House Price Index (HPI) values. Key steps in the data cleaning process included:

Conversion of Data Types: The dataset contained fields formatted as strings due to special characters and annotations which were converted into numeric format to enable mathematical operations essential for time series analysis.

Handling Missing Values: An examination of the dataset revealed missing values that could potentially skew the analysis. These were systematically removed to ensure the integrity and accuracy of the predictive models.

Data Normalization: Given the scale differences inherent in economic indices over extended periods, the data was normalized using standard scaling techniques. This step is crucial for neural network performance, as it ensures that the model does not become biased towards higher values and improves the convergence speed during training.

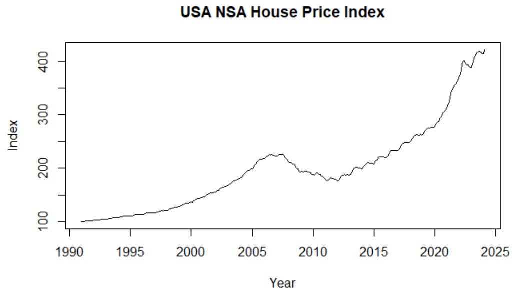

Figure 1: House Price Index Graph

Methodology

The methodology section is divided into several sub-sections detailing the approach for decomposing and analyzing the time series data:
Time Series Decomposition: Using the decompose() function, the time series data was broken down into its constituent components: trend, seasonality, and residuals. This step was critical for understanding the underlying patterns and ensuring that the model considerations align with the data's characteristics.

Stationarity Testing: To apply ARIMA models effectively, the time series must be stationary. The Augmented Dickey-Fuller (ADF) test was conducted to test the null hypothesis that the time series is non-stationary. This step determines the need for differencing to stabilize the mean of the time series.

ACF and PACF Analysis: The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were generated to identify the order of the ARIMA model. These plots indicate the lag after which the autocorrelations are close to zero, suggesting the terms to be included in the ARIMA model.

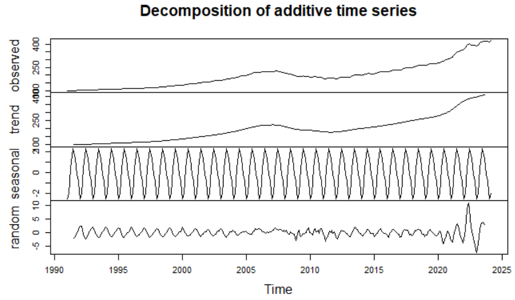

Figure 2: Decomposition Graph

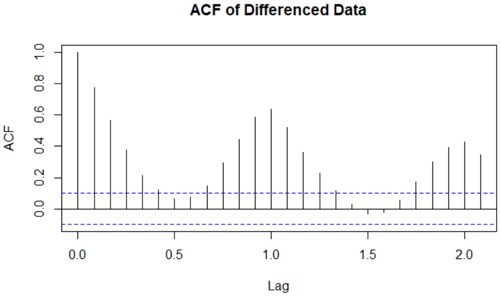

Figure 3: ACF Graph of HPI

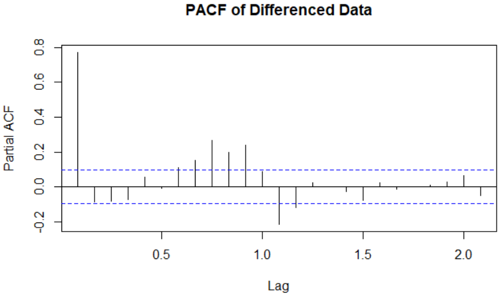

Figure 4: PACF Graph of HPI

Algorithms

This section delves into the specifics of the algorithms used for forecasting the HPI:
ARIMA Model: Based on the insights gained from ACF and PACF analysis, an ARIMA model was configured. The auto.arima() function from the forecast library was utilized, which automates the model selection process by iteratively exploring different combinations of parameters and selecting the best fit based on AIC (Akaike Information Criterion). The model's adequacy was then assessed by examining the residuals, ensuring that they resemble white noise, indicating a good fit.

LSTM Network: The Long Short-Term Memory (LSTM) network, a type of recurrent neural network, was chosen for its ability to capture long-term dependencies in time series data. The network was structured with an LSTM layer followed by a dense output layer to predict future values. The data was split into training and testing sets, with the LSTM model trained on the normalized and sequenced data. The training process involved adjusting the number of epochs and batch size to optimize the learning process and prevent overfitting.

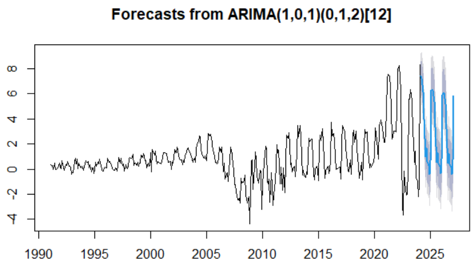

Figure 5: 3 years forecasting using ARIMA Model

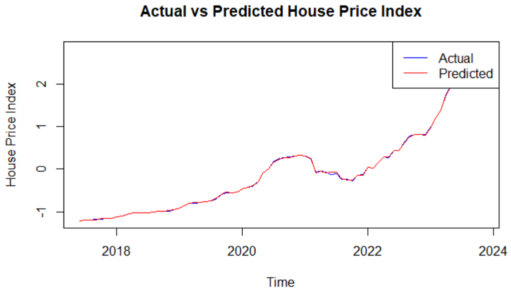

Figure 6: LSTM Model Prediction

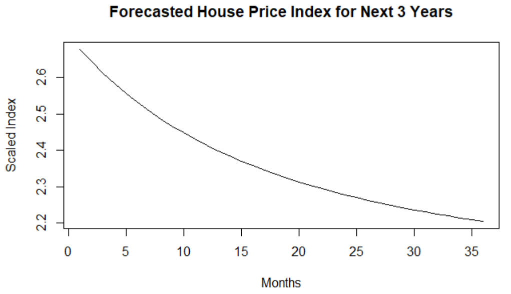

Figure 7: 3 years forecasting using LSTM Model

Conclusion

This study underscores the potential of integrating traditional statistical methods with advanced machine learning techniques to enhance the accuracy of economic forecasting. Future research could explore hybrid models that combine the strengths of ARIMA and LSTM, potentially leading to superior predictive performance in complex economic time series data.