

# Analyzing Homeownership in the United States

## Data Analysis 2

### Group 14

Jonathan Goldstein, Tadeus Marchesi, Alyssa Blodgett, Rajesh Nandakumar

April 29, 2023

---

## ABSTRACT

Homeownership in the United States has long been regarded as a significant indicator of economic growth and overall prosperity. Over the past 40 years, the country has experienced significant fluctuations in homeownership rates, influenced by various factors such as the Federal Reserve's Funds Rate, GDP, Median Home Sale Prices and Unemployment Rate. This paper aims to analyze the trends in homeownership rates and their correlation with economic indicators over the past four decades. By analyzing different factors, this analysis seeks to provide insights into the interdependence of the housing market and the overall economy, and the implications of these trends on future economic growth and stability.

## 1. INTRODUCTION

Several studies have explored the relationship between homeownership in the United States and various economic indicators. For instance, research by Mankiw and Weil (1989) found that homeownership rates were positively correlated with economic growth and income levels. Furthermore, the study suggested that changes in interest rates and mortgage availability were important predictors of homeownership rates. Another study by Cutts and Zandi (2002) analyzed the impact of demographic and economic factors on homeownership rates. The authors found that age, income, and education levels were significant predictors of homeownership rates. Additionally, the study suggested that interest rates, housing prices, and credit availability were also important determinants of homeownership. Overall, these studies suggest that a combination of economic, demographic, and policy factors contribute to predicting homeownership rates in the United States. Understanding these factors can inform policymakers and stakeholders in developing effective strategies to promote sustainable homeownership and support economic growth.

The given datasets intuitively are expected to be correlated in some sense. It is expected for GDP and Homeownership Rate to have a positive correlation since GDP is an indicator of a country's wealth. Similarly, we can expect the median home price, Federal Funds Rate and Unemployment Rate data to have a negative correlation with the Homeownership Rate because lower home prices and interest rates are ideal when purchasing a home. Though we can expect those relationships intuitively, we cannot say statistically whether those intuitions are correct without proper analysis. Multivariate modelling methods such as Vector Autoregression seemed promising in the brainstorming phase to simultaneously assess all factor's effects on the Homeownership Rate. Univariate Modeling and Prediction could serve as another valuable option if multivariate methods prove not to be successful. Regardless, through multiple methods of analysis, it was expected that at least one of the four data sets analyzed would aid in predicting future home prices.

## 2. EXPLORATORY ANALYSIS

Crucial steps towards determining whether Homeownership and the other economic factors are related, are cleaning, transforming and normalizing data. In order to accomplish this, a visualization of all time series were produced to give a picture of what degree of data preparation was required. Figure 1 shows the behavior of all five time series with data normalized between 0 and 1. Each data set clearly indicates trend, seasonality or both.

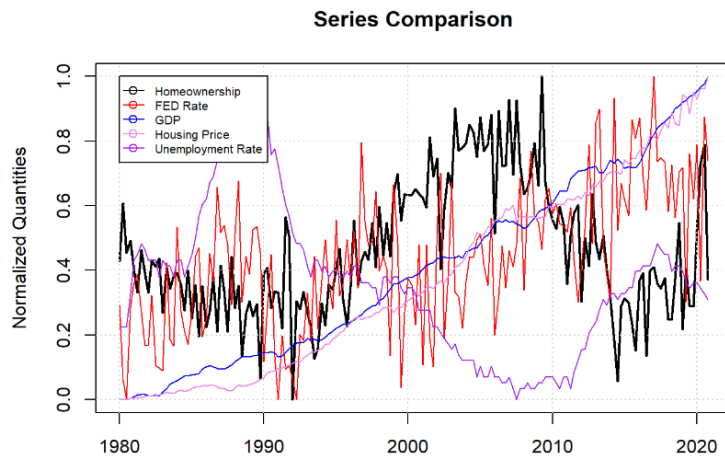


Figure 1 - Time Series Visualization

To alleviate the obvious violations of the three stationarity assumptions, differencing of was performed (Figure 2 below).



Figure 2 - First Order Differencing of Time Series Data

With the first order-differencing applied to all five time series, an Augmented Dickey-Fuller test was performed to ensure stationarity. Without stationarity, further modelling should not occur. Thankfully, the hypothesis test outputted p-values below the significance level of 0.05. Hence, the null hypothesis of non-stationarity can be rejected, and therefore, the analysis could proceed without further transformations.

Next, an assessment of correlation between the time series was performed to determine which series would be more valuable in the analysis. As presented in Figure 3, The addition of the correlation matrix does not increase confidence in the relationship between homeownership and the other factors to any degree. The highest correlation is between Housing Price and Homeownership, but still, that is only -0.13. This indicates an inverse relationship between the two, which does make sense intuitively. As house prices decrease, homeownership would increase and vice versa. In addition, correlations between Homeownership and the other variables are all below  $|0.1|$ , suggesting that that the relationship between variables is virtually non-existent.

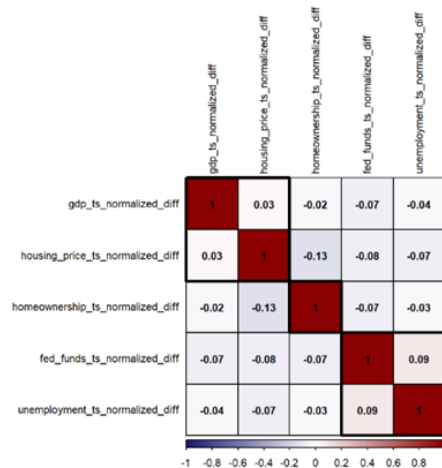


Figure 3 - Correlation Matrix

### 3. MULTIVARIATE ANALYSIS

The next step considered was a Vector Autoregression (VAR) model to perform a joint analysis on the four factors against Homeownership time. The first attempt was on an unrestricted VAR model with Federal Funds Rate, GDP, Median Home Price and Unemployment as predictors and Homeownership as the dependent variables. To determine predictive power, the regression coefficients were analyzed. A hypothesis test was then completed on each predictor to determine whether the specified predictor helps explain any variability in the model. In this particular test, the null hypothesis is that the regression coefficient is statistically 0. If this test outputs a p-value greater than 0.05, the null hypothesis is not rejected.

In the Unrestricted VAR Model (Figure 4, Upper) included up to 3 lags of all predictors, in which the summary did not provide any promising evidence. All predictors aside from Homeownership have large p-values, where lags 1, 2 and 3 explain the variability in the data. The Adjusted R-Squared value provides evidence that three lags of Homeownership explain 30% of the variability in the model. Therefore, another model is required to continue attempting to understand relationships within the datasets.

A Restricted VAR Model (Figure 4, Lower) was performed next, and the results were similar to that of the Unrestricted VAR Model. The only three variables selected when restricting the model were the three lags of Homeownership and the Adjusted R-Squared value deviated by less than a percent. Furthermore, the predictors in the model only explained 30% of the variability in the model. To further explore the data, Univariate Modelling should be completed.

	P Value	Adjusted R Squared
<b>Unrestricted VAR Model</b>		<b>0.299</b>
Homeownership Lag 1	1.02E-13	
Fed Funds Lag 1	0.911	
GDP Lag 1	0.14	
Housing Price Lag 1	0.501	
Unemployment Lag 1	0.211	
Homeownership Lag 2	5.86E-05	
Fed Funds Lag 2	0.264	
GDP Lag 2	0.388	
Housing Price Lag 2	0.768	
Unemployment Lag 2	0.394	
Homeownership Lag 3	0.035	
Fed Funds Lag 3	0.543	
GDP Lag 3	0.638	
Housing Price Lag 3	0.604	
Unemployment Lag 3	0.184	
<b>Restricted VAR Model</b>		<b>0.308</b>
Homeownership Lag 1	0.08035	
Homeownership Lag 2	0.0924	
Homeownership Lag 3	0.08077	

Figure 4 - VAR Model Output

#### 4. UNIVARIATE ANALYSIS

Of the univariate approaches available, the ARMA(p,q) - GARCH(m,n) model was chosen. This was mainly due to the fact that an ARMA model is relatively accurate in its predictions based on data from past lags, while concurrently adjusting to unpredictable shocks. In conjunction, A GARCH model is useful for predictions using heteroskedastic data. Order selection for both components commenced, testing ARMA orders up to 5 and GARCH orders up to 2. Combinations of all orders were collected and sorted by lowest BIC. The best model chosen from the selection process was the ARMA(5, 5) – GARCH(2, 1). This model will be used to attempt predicting 2021 Homeownership Rates.

Quarterly Homeownership Rates from March 1980 – December 2020 was used as the training set to predict Homeownership Rates for all of 2021 using the ARMA(5, 5) – GARCH(2, 1) model. Figure 5 compares the forecasted predicted values to the Homeownership Rate. This was completed to determine how accurate the prediction is by plotting bands displaying the forecasted values plus and minus the standard deviation of the forecasted values. The observation falls entirely within the prediction interval for 2021. The predicted observations are at most 2% higher than the forecast, which can be interpreted as relatively good model.

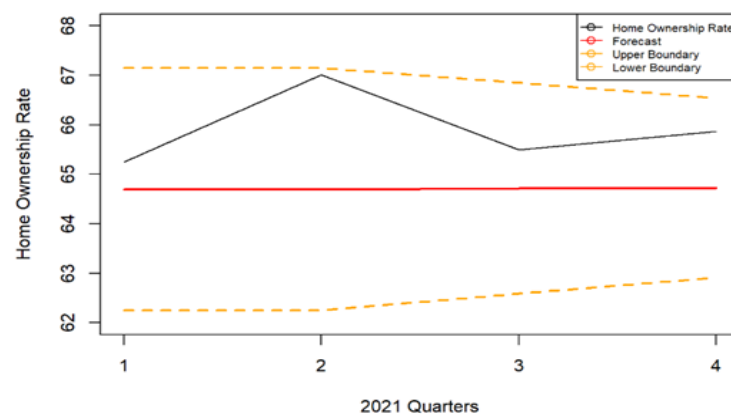


Figure 5 - ARMA (5, 5) - GARCH (2, 1) Prediction

The above prediction merely helped to prove that an ARMA-GARCH model may be useful in helping to predict Homeownership Rates in the future using past homeownership data. A different approach is required to progress the analysis, and therefore, Neural Networks were attempted.

#### 5. NEURAL NETWORKS

This session discusses two alternative approaches for forecasting Homeownership Rates based on Neural Networks<sup>1</sup>, which are Univariate (NAR)<sup>2</sup> and Multivariate (Nueral Fit)<sup>3</sup>

The first approach proposes a nonlinear autoregressive network (NAR) model as an alternative to univariate models such as ARMA and ARIMA. The defining equation of a NAR model is:

$$y(t) = f(y_{t-1}, y_{t-2}, \dots, y_{t-n_y})$$

Where the next value of the time series  $y(t)$  is regressed on previous values of the same series<sup>4</sup>.

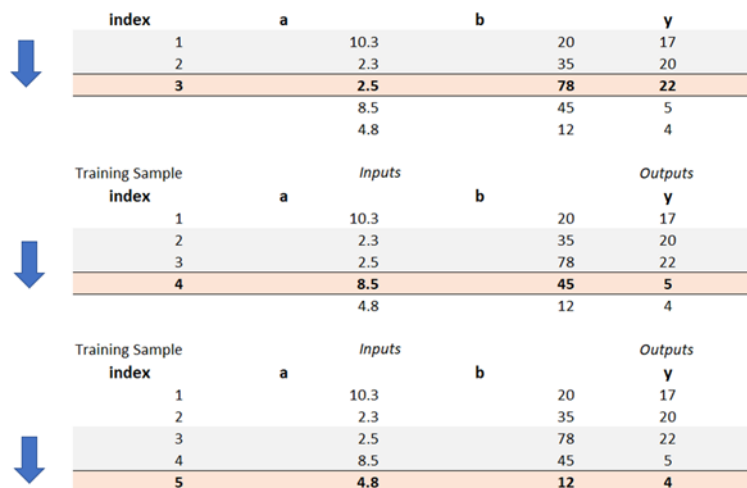
The second approach is based on a two-layer-feed-forward network (Neural FIT). These models can fit multi-dimensional problems; thus, they emerge as alternatives to multi-variate models (e.g., VAR). It's essential to distinguish between how these neural networks are typically used and how they are adapted in this project to work with time series.

In Figure 6, The Neural Net is presented to the input pairs and their respective outcome  $y$ , in order for its hyperparameters to be automatically adjusted based on the backpropagation algorithm<sup>5</sup>. Typically, in order to improve performance and avoid overfitting, the sets are presented to the Neural Net randomly. Such an approach is not appropriate for a time series once it does not capture how the output  $y(t)$  is influenced by  $y_{t-1}$ .

Training Sample index	a	b	y
3	10.3	20	17
	2.3	35	20
1	2.5	78	22
	8.5	45	5
2	4.8	12	4

Figure 1 - Typical Training Sampling

To overcome this, the training process follows the illustration in Figure 7. In this example, the first two entries predict the third. The time series window then slides one instant in time forward so the following two elements can be used to forecast the next.



index	a	b	y
1	10.3	20	17
2	2.3	35	20
3	2.5	78	22
	8.5	45	5
	4.8	12	4

Training Sample index	a	b	y
1	10.3	20	17
2	2.3	35	20
3	2.5	78	22
4	8.5	45	5
	4.8	12	4

Training Sample index	a	b	y
1	10.3	20	17
2	2.3	35	20
3	2.5	78	22
4	8.5	45	5
5	4.8	12	4

Figure 2 - Time Series Training Scheme

It is important to note that throughout the Neural Net creation process, the known data is divided into training, validation, and test sets, where the test set used to assess the quality of the Neural Network to be deployed used for 2021 predictions (Figure 8).

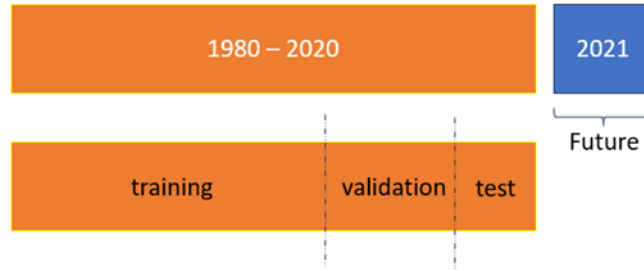


Figure 8 - Data Subdivision for Neural Net Training

## 5.1 UNIVARIATE APPROACH (NAR)

A NAR net with a time lag = 4 was trained using 70% of the data set, leaving 30% for validation and testing. Figure 9 displays the errors and the respective output-target regression equations. The best performance was obtained in the 2nd training epoch during the third pass of the algorithm.

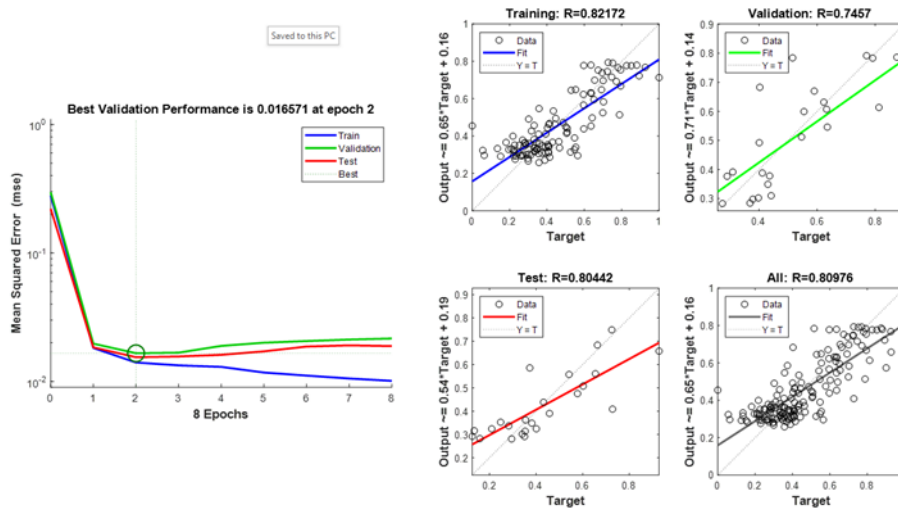


Figure 3 - NAR Training, Validation and Test

The residuals fit well (Figure 10), whereas the ACF plot (Figure 11) seems to indicate a random process. Figure 12 represents the forecast as per the trained net. This approach outperformed the ARMA-GARCH model previously discussed in this report, indicated by a lower MAPE (0.01) and Prediction Measure (1.33), than the other 13 models tested in this report.

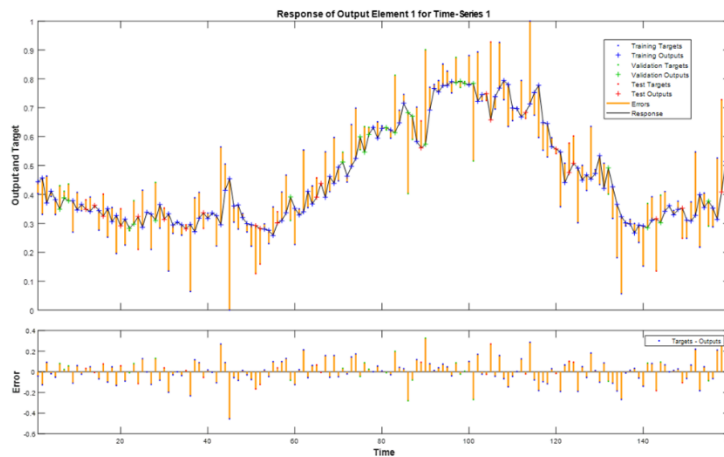


Figure 10 - NAR Residuals

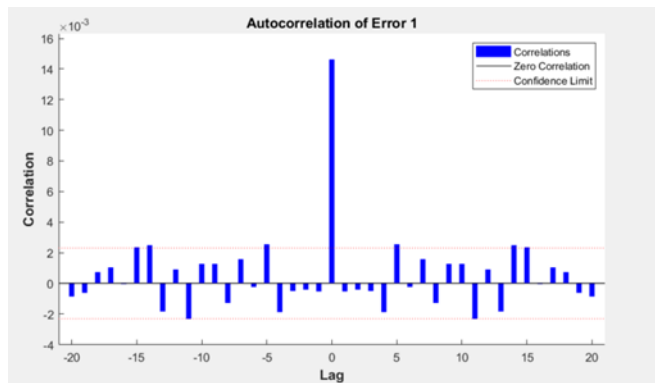


Figure 11 - ACF of NAR Residuals

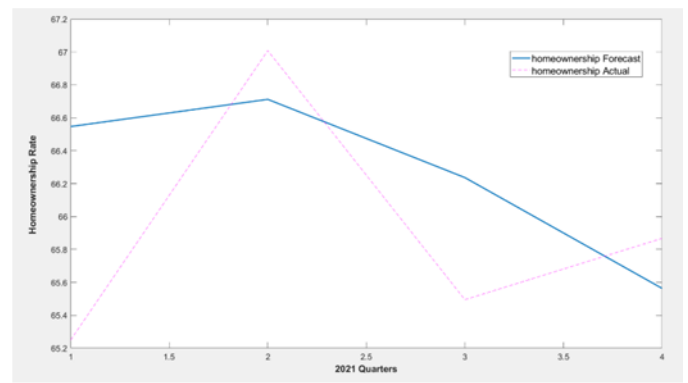


Figure 12 - NAR Forecast for 2021 vs. Actual data

## 5.2 MULTIVARIATE APPROACH (Neural FIT)

Part of the idea of using a multi-variate approach goes beyond the expectation of improving the prediction performance by considering exogenous factors. To question the variable selection previously discussed, a brute-force strategy was adopted: the variables were arranged in 12 models (Figure 13), and the performances were compared using PM and MAPE.

Model No.	Input Series			
model 1	Fed Funds	GDP	Housing Price	Unemployment
model 2	Fed Funds	GDP	Housing Price	
model 3	Fed Funds	GDP		
model 4	Fed Funds			
model 5		GDP	Housing Price	Unemployment
model 6		GDP	Housing Price	
model 7		GDP		
model 8			Housing Price	Unemployment
model 9			Housing Price	
model 10		GDP		Unemployment
model 11	Fed Funds			Unemployment
model 12	Fed Funds		Housing Price	

Figure 13 - Variable Combination Table

As evident in Figure 14, the Fed Funds and Unemployment Rates appeared as the best predictor input variables.

Model No.	PM	MAPE
model 1	0.8	0.0087
model 11	1.1751	0.0083
NAR	1.33	0.01
model 12	1.4054	0.0091
model 9	1.4729	0.0108
model 7	1.72	0.0092
ARMA-GARCH	4.16	0.018
model 3	9.25	0.0294
model 10	15.5462	0.0344
model 2	15.84	0.0343
model 6	39.95	0.0637
model 8	46.5	0.0676
model 4	65	0.0645
model 5	89.41	0.096

Figure 14 - Overall Results

The training, test, and validation sets for Model 1 are shown in Figure 15.1 — the error of 0.937 in the training set displayed very poor performances in the validation (negative correlation) and test data sets. This seems to indicate overfitting, and the predictions shown in Figure 15.2 are likely random results.

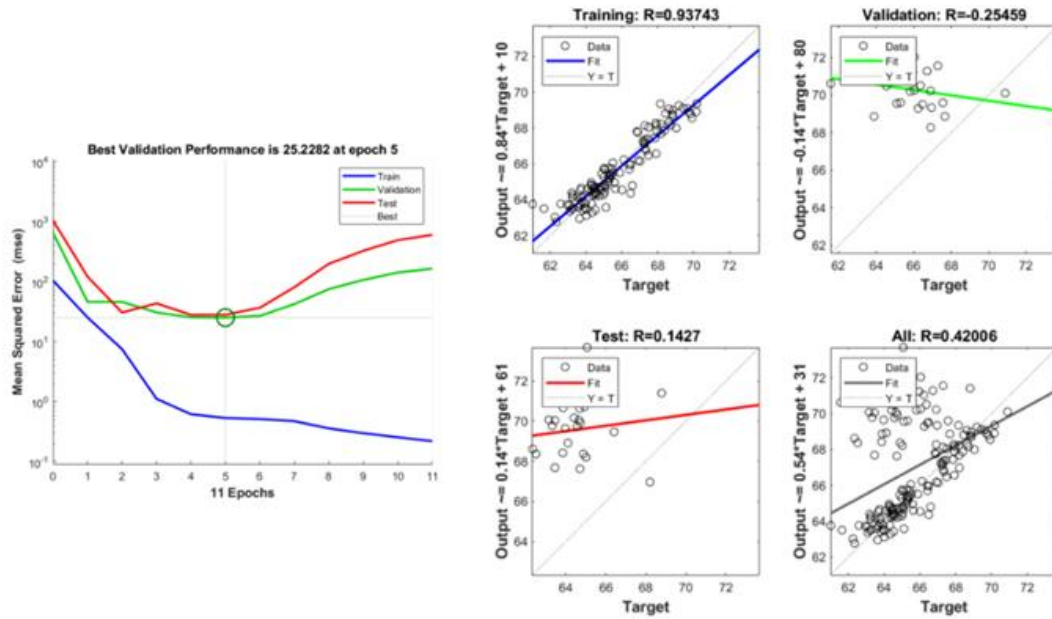


Figure 15.1- Training, Validation and Test of Model 1

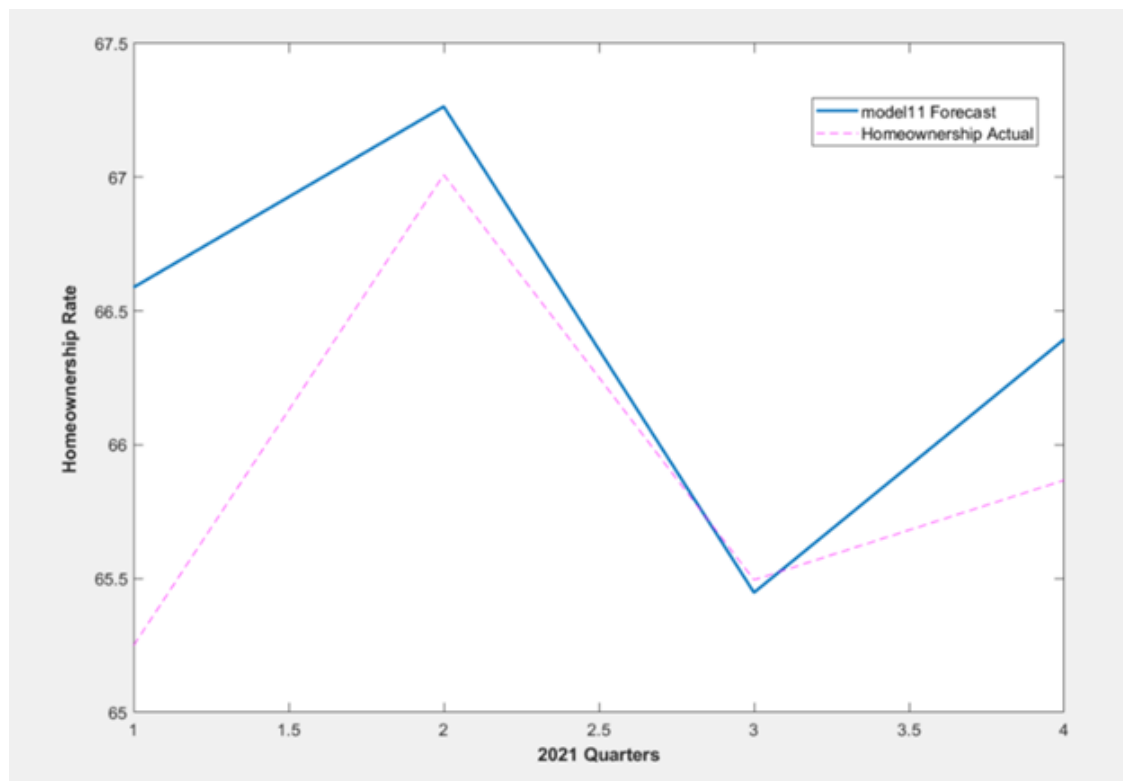


Figure 15.2 - Training, Validation and Test of Model 1

The training, test, and validation for Model 11 (Figure 16.1) — opposed to Model 1, Model 11 exhibits better validation results (error = 0.46) but with poor test performance. Despite that, the predictions shown in Figure 16.1 closely match the actual data.



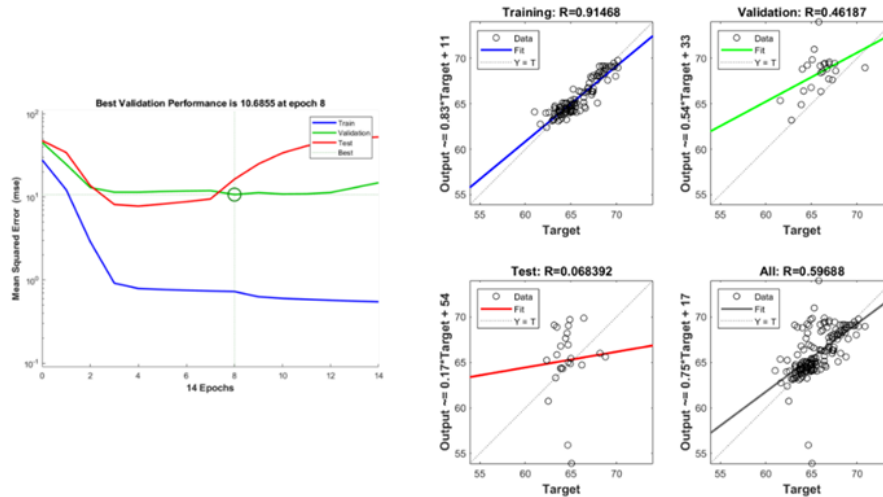


Figure 16.1 - Model 11 Training, Validation and Test

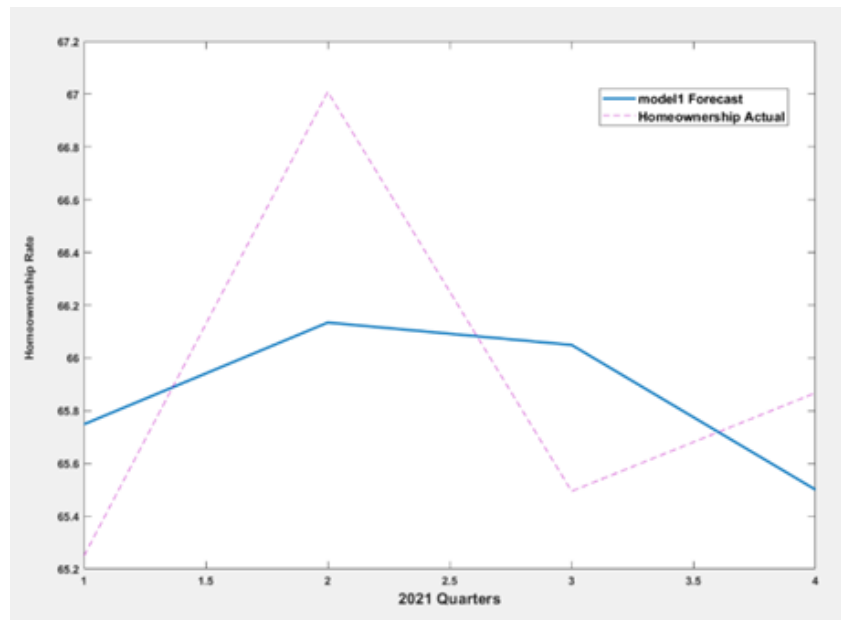


Figure 16.2 - Model 1 Forecast for 2021 vs. Actual Data

## 6. CONCLUSION

Although the VAR models proved GDP, Federal Funds Rate, Median Home Price and Unemployment Rate to be exogenous factors to Homeownership, the Neural Net approach tells a different story. In particular, the Federal Funds data Rate combined with Unemployment seems to produce strong predictions (Model 11 results). It is important to note that Model 1 outperformed all the others models tested via metrics such as PM and MAPE, however, Model 11 produced the best prediction. All predictions performed in this analysis, failed to capture shocks similar to Model 11. This could indicate potential pitfalls in PM and MAPE as prediction methods.

One important assumption made in the Neural Fit (Models 1 to 12) was that the future states of the input series were known, which, in a practical context, is never the case. Thus, it's fair to say that the NAR model was the best, given its performance and realism. Since the best performing model is a Univariate Neural Network approach, the analyses within this paper didn't show any statistical evidence towards an interdependence between Homeownership and the four economic factors tested. Homeownership data from prior time periods is very capable in forecasting future Homeownership Rates fairly accurately.

---

## 7. FUTURE ANALYSIS

This extensive analysis, though it led to using advanced and interesting techniques such as Neural Networks, did not give information as to what factors are directly related to homeownership. On the other hand, there were a few points discovered while modeling that may lead to interesting extensions to the analysis.

The Unrestricted VAR model showed that Housing Price was lagged by Homeownership and Unemployment. Instead of predicting Homeownership Rates, forecasting housing prices may provide more fruitful results.

The Neural Network approach was quite useful compared to other models, but potentially fine-tuning hyper neural net parameters would further improve results. To further improve this method, the ARMA-GARCH models could be combined to predict the Neural Fit inputs and establish a more realistic framework.

Overall, a good mix of univariate and multivariate time series modeling methods were applied in this analysis, and homeownership prediction quality was maximized given the tools at our disposal. Naturally, the research question the report aimed to answer brought about more questions that can be answered in the future given a lot of knowledge gained by this analysis.

## References

- [1] Greenwood, D. (1991). An overview of neural networks. *Behavioral science*, 36(1), 1-33.
- [2] Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, 14(3), 101039.
- [3] Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1), 44-55.
- [4] Pontes, F. J., Ferreira, J. R., Silva, M. B., Paiva, A. P., & Balestrassi, P. P. (2010). Artificial neural networks for machining processes surface roughness modeling. *The International Journal of Advanced Manufacturing Technology*, 49, 879-902.
- [5] Wythoff, B. J. (1993). Backpropagation neural networks: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 18(2), 115-155.
- [6] Mankiw, N. G., & Weil, D. N. (1989). The baby boom, the baby bust, and the housing market. *Regional Science and Urban Economics*, 19(2), 235-258.
- [7] Cutts, A. C., & Zandi, M. (2002). The rise and fall of subprime mortgages. *Journal of Housing Research*, 13(2), 139-158.
- [8] United States. Bureau of Labor Statistics. "(Seas) Unemployment Level". Bureau of Labor Statistics Data Tools. U.S. Dept. of Labor