House Price Prediction Using SARIMA AND VAR Model

Faaiz Anwar
dept. Data science (Amrita University)
Amrita University
Banglore, India
faaizanwar13@gmail.com

***Abstract*—These days, we see utilizations of Machine Learning (ML) and Computerized reasoning (AI) in practically every one of the areas yet for quite a while frame the land industry was very delayed in embracing information science and AI for issue addressing and improving their cycles.Present AI calculation helps us in upgrading security alarms, guaranteeing public wellbeing and improve clinical upgrades. Because of expansion in urbanization, there is an expansion sought after for leasing houses and buying houses. The costs of land and inherent houses consistently increment by time. Cities have consistently given freedoms to settle down individuals all over the country. The cost of houses in the city has expanded radically in the last 10 years. Here we have the quantity of houses bought from 2008 to 2019. We dissect the time arrangement information and using the SARIMA model we attempt to foresee the future extension.Informational index we have taken is a multivariate time arrangement information. So here we will likewise carry out a vector autoregressive(VAR) model as it helps in discovering the unique conduct of monetary and monetary time arrangement estimation.**

***Index Terms*—SARIMA, multi variant. Var,**

I. Introduction

Venture is a business action that the vast majority are keen on this globalization period. There are a few items that are frequently utilized for venture, for instance, gold, stocks and property. Specifically, property investment has increased significantly since 2011, both on demand and property selling.[1]In the United States house deals have developed by 34 percent somewhat recently and arrived at a record high of 5.51 million a year ago. In Australia, house deals have expanded by 36 percent since 2013.House price prediction [2] has consequently drawn in far reaching considerations on the grounds that the forecast results can help different land partners to settle on more educated choices.Traditionally, expectation of house cost is regularly controlled by proficient appraisers. However, an appraiser is probably going to be one-sided because of personal stake from the moneylender, contract representative, purchaser, or dealer. Accordingly, a mechanized forecast framework is useful to fill in as an autonomous outsider source that might be less one-sided.The Hedonic cost model proposed from the viewpoint of financial aspects is the most normal delegate, and has been concentrated widely in the writing of house value forecasts[3] . Be that as it may, it is essentially utilized for examining the connection between house cost furthermore, house highlights, where it ordinarily receives relapse strategies. As of late, with the broad use of Machine Learning in different fields, house value forecasting through more machine learning strategies, like ANN (Artificial Neural Network) , SVM (Support Vector Machine), AdaBoost (Adaptive Boosting) has additionally gotten increasingly more consideration[4]. Here we are using a time series model for our data set. House price prediction is a challenging problem.From one perspective, the compelling components of house cost are intricate and changed nonlinearly, making the customary models typically have huge forecast mistakes. Then again, the day by day information of the housing market is extremely immense and it is expanding quickly.To address these concerns, a house price prediction model based on Multivariate data is proposed first. The house price trend is predicted by using the ARIMA model. The multivariate data is handled by the VAR model, separately describing and predicting every attribute. To demonstrate the effectiveness and utility of the proposed approach, comparative experiments were conducted on the real housing data extracted from the Internet.

Identify applicable funding agency here. If none, delete this.

II. Design and implementation

A. Exploratory Data Analysis

Multivariate information consists of individual estimations that are gained as an element of multiple factors, for instance, energy estimated at numerous frequencies and as an element of temperature, or as an element of pH, or as a component of beginning fixations, etc, of the responding arrangements. A Multivariate Time Series has more than one time-subordinate variable. Every factor depends on its past qualities as well as has some reliance on different factors. We have property deals information for the 2007-2019 period for one explicit locale. The information contains costs for houses and units with 1,2,3,4,5 rooms. These are the cross-dependent factors [8] fig[1] fig[2].

It is evident that the 2 bedroom curve before 2009 is not an accurate representation of the actual median price.It is not possible for a 2 bedroom median price to be above that of 3 bedroom median price.This is due to low number of sales in that timeframe, which skews the calculated median price.

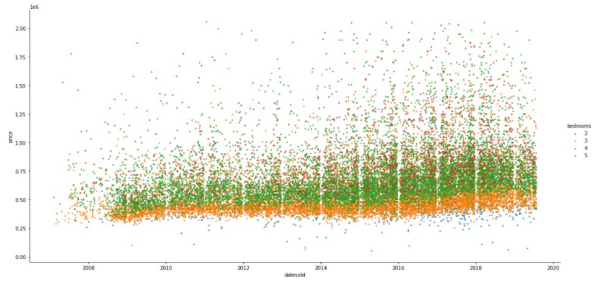

Fig. 1. Data Set.

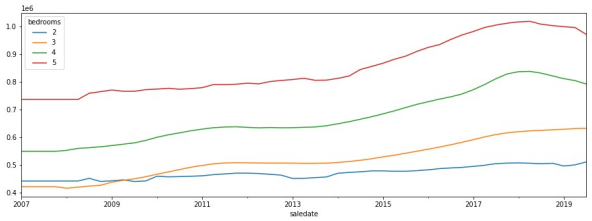

Fig. 2. G sampled median aggregator.

For the SARIMA model we found the box plot of the total mean including the price and the bedroom in the data set. And found Trend and seasonality in the time series data set.

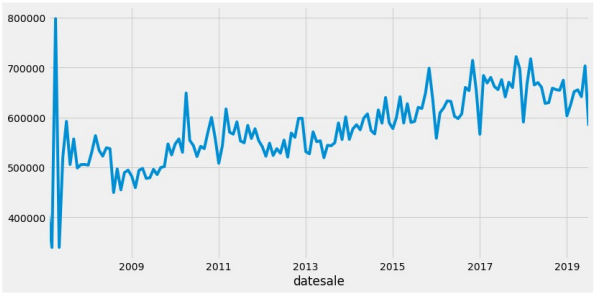

Fig. 3. Trend and Seasonality.

Multivariate information consists of individual estimations that are gained as an element of multiple factors, for instance, energy estimated at numerous frequencies and as an element of temperature, or as an element of pH, or as a component of beginning fixations, etc, of the responding arrangements.

B. Prediction of house price trend

sarima(P ,D,Q,s) is another option, shorthand documentation for indicating the multiplicative seasonal parts of models with ARMA aggravations. The reliant variable and any free factors are lag s seasonally differenced D times, and 1 through P seasonal lags of autoregressive terms and 1 through Q seasonal lags of moving-normal terms are remembered for the mode.

In the event that the arrangement is long and fixed and the hidden information producing measure doesn't have a long memory, evaluations will be comparable, regardless of whether assessed by unlimited most extreme probability (the default),

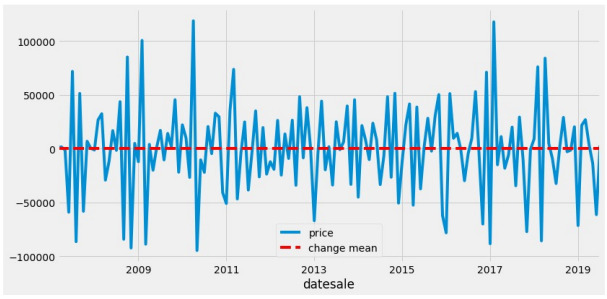

Fig. 4. Stationarity check with dickey-furley test.

$$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \theta_1 \epsilon_{t-1} + \theta_4 \epsilon_{t-4} + \epsilon_t$$

restrictive most extreme probability (condition), or greatest probability from a diffuse earlier (diffuse).

$$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \theta_1 \epsilon_{t-1} + \theta_4 \epsilon_{t-4} + \epsilon_t$$
$$(1 - \phi_1 L)\Delta y_t = c + (1 + \theta_1 L + \theta_4 L^4)\epsilon_t$$

By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the differenced series, you can tentatively identify the numbers of AR and/or MA terms that are needed [5][6].

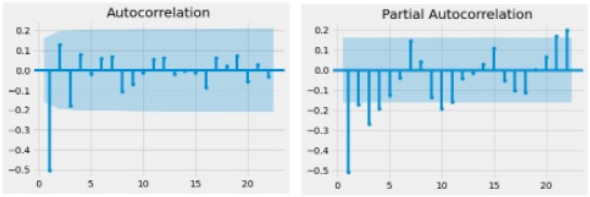

Fig. 5. ACF and PACF test.

C. Prediction of Individual House Price

The vector autoregression (VAR) model broadens the possibility of univariate autoregression to k time arrangement relapses, where the slacked upsides of all k arrangements show up as regressors [7]. Put in an unexpected way, in a VAR model we relapse a vector of time arrangement factors on slacked vectors of these factors. Concerning AR(p) models, the slack request is meant by p so the VAR(p) model of two factors Xt furthermore, Yt (k = 2) is given by the conditions [9][10].

III. EXPERIMENTAL RESULTS

The proposed approach was carried out on the python structure. The exploratory dataset was obtained from the Kaggle site. A sum of 29,581 examples were separated from the site, where 28,581 of them were utilized to be the preparation information and the rest 1,000 to be the test information.

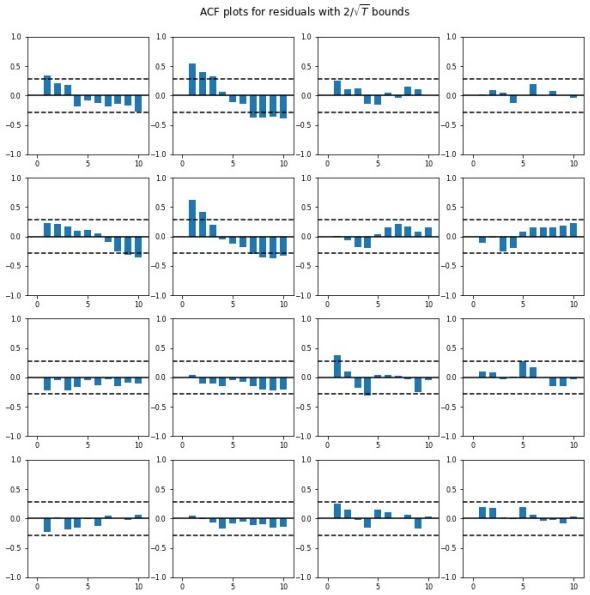

Fig. 6. ACF residual test in var model.

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \ldots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \ldots + \gamma_{1p}X_{t-p} + u_{1t},$$
$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \ldots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \ldots + \gamma_{2p}X_{t-p} + u_{2t}.$$

A. Prediction of House Price Trend

The sequential execution of the SARIMA model worked in the very same manner as its equal part with the exception of that the consecutive hereditary calculation executed utilizing a solitary center rather than being executed by multithreading.It very well may be tracked down that the anticipated house cost presents a light diminishing pattern. The future cost inside one or on the other hand two months anticipated by the SARIMA model is essentially reliable with the genuine information. Yet, with the increment of the expectation date, the anticipated future house cost would become conflicting with the genuine information on the grounds that the SARIMA model is a brief time frame forecast model.

Fig. 8 shows the quantile-quantile plot. Clearly the anticipated information and the genuine information harmonize with something very similar dispersion and the example information is an around normal distribution.

Fig. 9 shows the normalized lingering plot. It tends to be tracked down that the residuals are dispersed around the centerline arbitrarily. In this manner, the model is trustworthy as indicated by the factual hypothesis.

The forecast results if fig 11. The next 30 steps were forecasted on the data set. And it showed a decrease in the seasonality of the time series data.

The root mean square error for checking the prediction is 1.22.

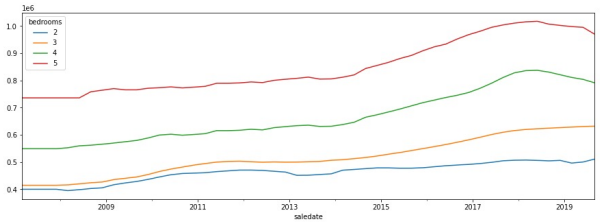

Fig. 7. Smoothning with var model.

SARIMAX Results

Dep. Variable:	price	No. Observations:	149
Model:	SARIMAX(1, 1, 4)	Log Likelihood	-1755.738
Date:	Sun, 30 May 2021	AIC	3523.475
Time:	10:15:04	BIC	3541.459
Sample:	03-31-2007 - 07-31-2019	HQIC	3530.782
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5387	0.542	-0.993	0.321	-1.602	0.524
ma.L1	-1.1606	0.551	-2.108	0.035	-2.240	-0.081
ma.L2	-0.1561	0.935	-0.167	0.867	-1.989	1.677
ma.L3	0.1709	0.461	0.370	0.711	-0.733	1.075
ma.L4	0.1515	0.139	1.094	0.274	-0.120	0.423
sigma2	1.389e+09	1.27e-09	1.09e+18	0.000	1.39e+09	1.39e+09

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	3.88
Prob(Q):	0.82	Prob(JB):	0.14
Heteroskedasticity (H):	0.98	Skew:	0.17
Prob(H) (two-sided):	0.95	Kurtosis:	3.72

Fig. 8. SARIMA result.

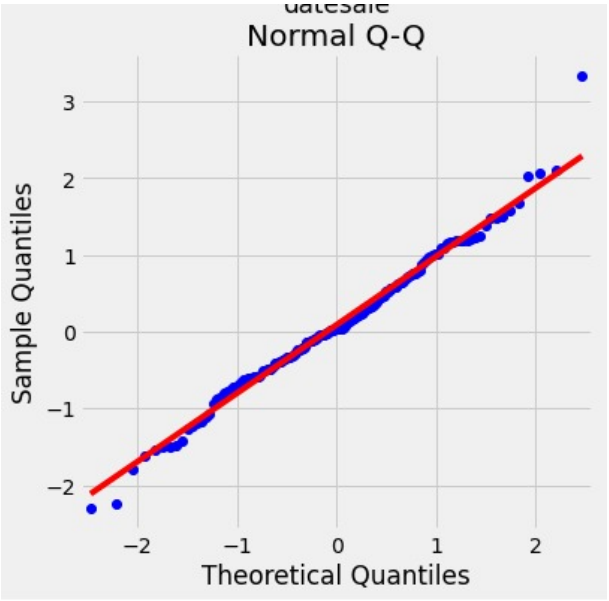

Fig. 9. The quantile-quantile plot.

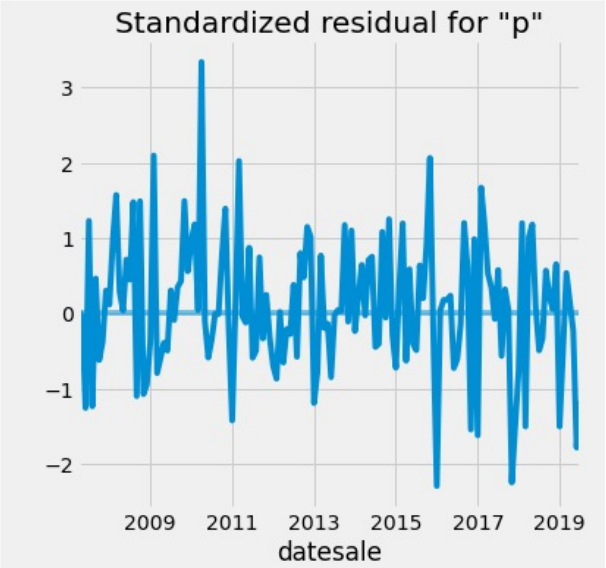

Fig. 10. Residual Plot.

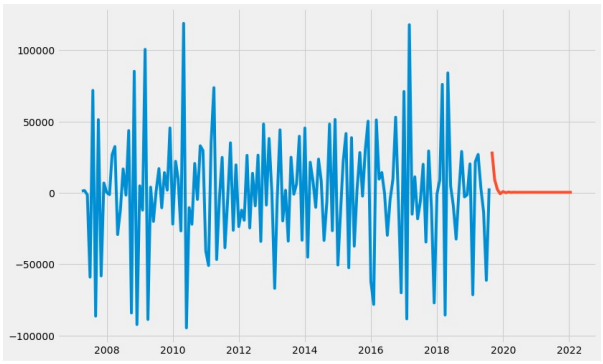

Fig. 11. SARIMA FORECASTING.

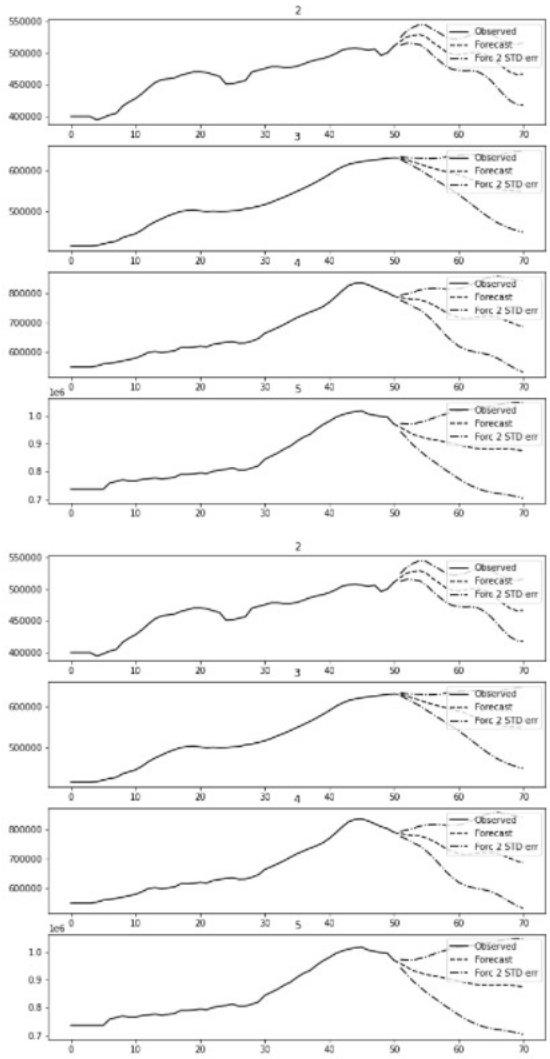

Fig. 12. VAR forecasting.

B. Prediction of Individual House Price

The summery result for forecasting from VAR model.

By utilizing the VAR model we had the option to foresee the costs of every individual houses with 1 2 3 4 and 5 rooms and we discovered the precision by checking the root mean square from the test informational collection to our real worth and the exactness for the root mean square were, 2 bedrooms 1.027963025803897 3 bedrooms 2.3523955523069557 4 bedrooms 8.347488048618313 5 bedrooms 7.229671356483708

CONCLUSION

In section 2 and 3. when applying SARIMA model in a multivariate data set. It can only find the prediction in the trend of the data set and we had a root mean square of 1.2 which is good to have a quantitative research and when applying aVAR model in our data set. We found that the VAR model works better on a multivariate time series data set where we had multiple variables in the house column. Model can predict better in two bedroom and three bedroom types of house.Rather in 4 bedrooms and 5 bedrooms houses the accuracy was low. Further in this research we can make our bedrooms an exogenous factor in hours SARIMA model which can be further implemented in the SARIMAX model.

IV. REFERENCES

[1] Qiu, H., Zhao, H., Xiang, H. et al. Forecasting the incidence of mumps in Chongqing based on a SARIMA model. BMC Public Health 21, 373 (2021). https://doi.org/10.1186/s12889-021-10383-x

[2] R. M. A. van der Schaar, —Analysis of Indonesian Property Market; Overview and Foreign Ownership,‖ Investment Indonesian. 2015.

[3] R. Schulz and A. Werwatz, "A state space model for berlin house prices: Estimation and economic interpretation," The Journal of Real Estate Finance and Economics, vol. 28, no. 1, pp. 37–57, 2004.

[4] World Health Organization. Global Status report on road safety. Geneva: WHO; 2015.

[5] K. M. Habibullah, A. Alam, S. Saha, A. Amin and A. K. Das, " A Driver-Centric Carpooling: Optimal Route-Finding Model using Heuristic Multi–Objective Search," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[6] 12th IEE International Conference on Road Transport Information and Control, 2004. RTIC 2004.

[7] Mutangi, K.,"Time Series Analysis of Road Traffic Accidents in Zimbabwe", InternationalJournal of Statistics and Applications, 5(4), pp.141-149, 2015.

[8] Data Set link- https://www.kaggle.com/htagholdings/property-sales

[9] P. Goertler, B. Mahardja, and T. Sommer, "Striped bass (Morone saxatilis) migration timing driven by estuary outflow and sea surface temperature in the San Francisco Bay-Delta, California," Scientific Reports, vol. 11, no. 1, pp. 1510.1–1510.11, 2021.

[10] Rabbani, M.B.A., Musarat, M.A., Alaloul, W.S. et al. A Comparison Between Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing (ES) Based on Time Series Model for Forecasting Road Accidents. Arab J Sci Eng (2021). https://doi.org/10.1007/s13369-021-05650-3

[11] Nokeri T.C. (2021) Forecasting Using ARIMA, SARIMA, and the Additive Model. In: Implementing Machine Learning for Finance. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-7110-0_2