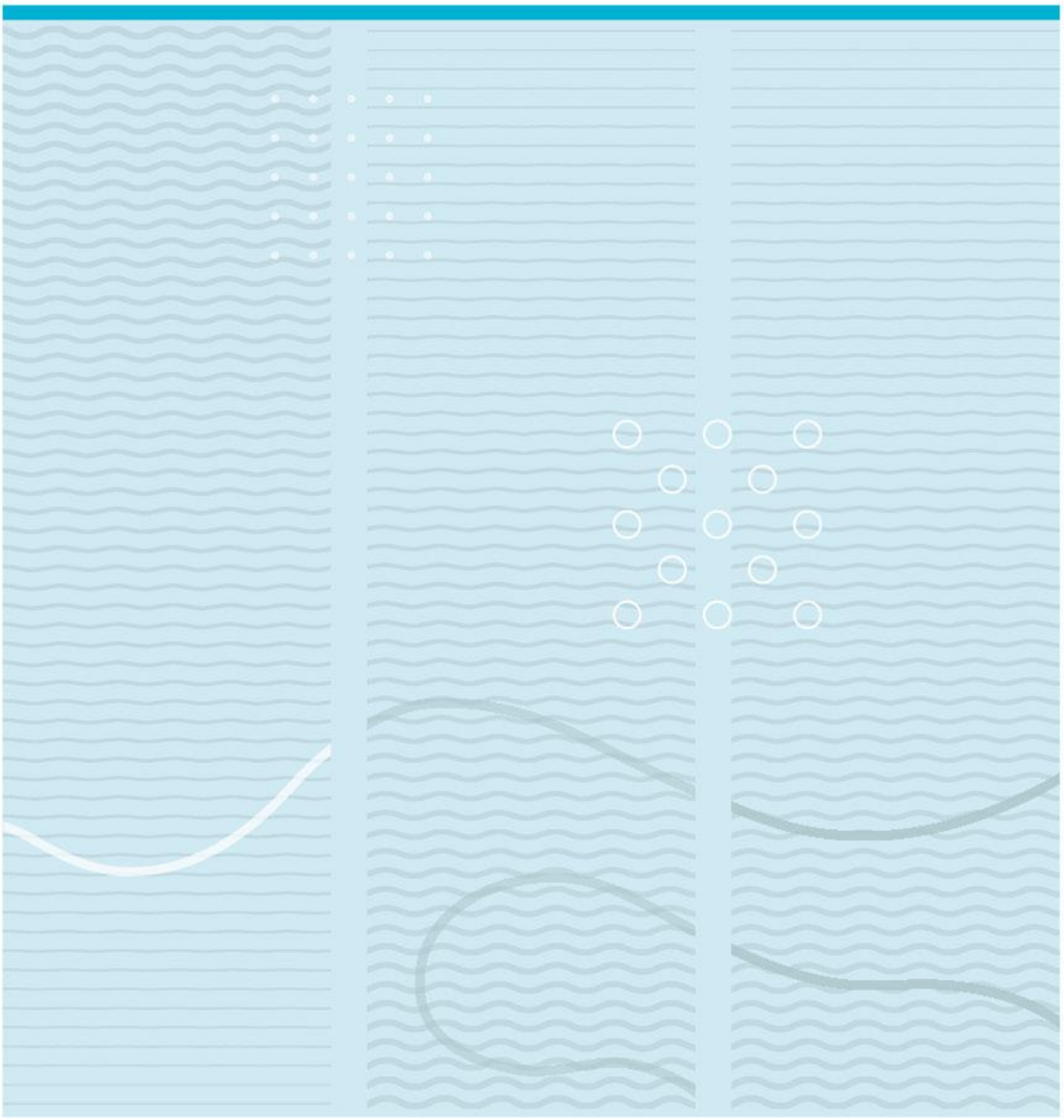
University of South-Eastern Norway
Faculty of Science, Technology, Natural Sciences and Maritime Sciences
–

CS4020: Data Science
Study programme: Master of Science in Computer Science
Spring/Autumn 2022

246108
250808
250822

House price prediction by time series analysis



University of South-Eastern Norway
Faculty of Science, Technology, Natural Sciences and Maritime Sciences
PO Box 235
NO-3603 Kongsberg, Norway

http://www.usn.no

This project is worth xx study points

Summary/Abstract

In this project, we worked to produce a time series model which can forecast the price of a house property. The reason for choosing this project was that we find this project fruitful and learning worthy. Fruitful in the sense that, in this project we will be using classic timeseries forecasting to build a prediction model. And such kind of prediction task is widely and frequently used in industry for business forecasting, understanding past behavior, plan future and evaluate current accomplishment.

We imported dataset from Kaggle that contains house sales price ranging from 2007-2019. We followed Scikit-learn approach as well as deep learning to perform the project. Specifically, we used univariate and multivariate time series analysis to produce models. For univariate analysis, we chose autoARIMA model and for multivariate, we chose VAR and LSTM. We made two assumptions initially to proceed forward. Then, we generated the models and compared the test results obtained from different models. Finally, we found that univariate time series analysis using autoARIMA was suitable for our dataset.

Contents

1 Introduction

Forecasting is an essential task in business nowadays because it helps companies to know their economical trend and formulate their plan and policies accordingly. This type of valuable forecasting information is mostly obtained by time series analysis. The analysis is widely used for predicting values in future in various business domains, it can be sales, house price or weather(Bontempi, Ben Taieb, & Borgne, 2012).

House price is an important aspect in society as it affects all stake holders, government regulating real estate market, real estate developers taking investment decision and consumers buying houses for fundamental shelter needs. The range of house price is great concern for both buyers and sellers, and thus prediction of house price is helpful for them to make decisions. The house price is a time series data. It can be used to generate a standard forecasting system that can serve as an independent third-party source and lessen bias problem with evaluators(Wang, Zou, Zhang, & Shi, 2019).

1.1 Problem statement

In our project “House price prediction by time series analysis”, we worked on a dataset from Kaggle which contains sales data of house from 2008-2019. Here, we built a machine learning model which can predict the price of house for future quarters. In order to build the model, we applied univariate as well as multivariate time series analysis. We use (Auto Regression Integration Moving Average) autoARIMA model for single time series, (Vector Auto Regression) VAR and (Long Short Term Memory) LSTM model to perform multiple series analysis. Our problem statement is to build different time series model for our dataset and select the best one among them.

1.2 Project objective

The objective of our project is to make a standard time series model that can forecast future quarter of house price.

1.3 Assumptions and limitations

Two assumptions were made to accomplish the project. First, we assumed that the prediction of a house price depends mainly upon its past price (lags) only. We considered only one dependent variable here. To model this assumption, we performed univariate time series analysis using autoARIMA.

Secondly, we assumed that there are also other factors which affects the forecasting of a house price, apart from its own past prices. For example, location, floor area, age, last renovated date,etc. have significant impact upon prediction. But our dataset was limited and lacked such influencing parameters. So, we targeted number of bedrooms parameter as our multiple variable and assumed that prediction of a house having certain number of bedrooms is also affected by the price of another house having other number of bedrooms. For instance, prediction of 2-bedroom house price is also affected by past price of 3,4,5-bedroom house in addition to its own previous lags.

1.4 Project management

We three people work equally to complete the project. We distributed our work as mentioned below:

Report writing

- Introduction – 246108
- Theoretical background – 250808
- Literature review – 250822
- Research methodology – 250822,250808
- Results – 246108
- Discussion and conclusion – 246108,250808,250822

Project

- Univariate analysis – 250822,250808
- Multivariate analysis – 246108

2 Theoretical background and Literature review

2.1 Theoretical background

Time series data often arise when monitoring industrial processes or tracking corporate business metrics. The essential difference between modelling data via time series methods and using the process monitoring methods.

According to tableau, Time series analysis is a specific way of analysing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

Time series forecasting requires *extra* preprocessing steps.

On top of the normality assumptions, most ML algorithms expect a *static relationship* between the input features and the output.

A static relationship requires inputs and outputs with constant parameters such as mean, median, and variance. In other words, algorithms perform best when the inputs and outputs are stationary.

This is not the case in time series forecasting. Distributions that change over time can have unique properties such as seasonality and trend. These, in turn, cause the mean and variance of the series to fluctuate, making it hard to model their behavior (Kaggle).

2.2 Literature review

Predicting the future is one of the complex and challenging tasks in applied science. Building effective predictors from historical data requires computational and statistical methods to infer dependencies between past and short-term future values from observations, as well as appropriate strategies for dealing with longer horizons (Darmawan & Sriyanti, 2017). The forecasting domain has been influenced by linear statistical methods such as ARIMA models since the 1960s. However, in the late 1970s and early 1980s it became increasingly clear that linear models were not suitable for many practical applications (Litsiou, 2021). In the same period, several useful nonlinear time series models were proposed such as the bilinear model (Poskitt & Tremayne, 1986). In comparison to linear time series analysis and forecasting, the development of non-linear time series analysis and forecasting is still in its infancy(Litsiou, 2021). However, Machine learning models have gained attention in the forecasting community during the last two decades and have established themselves as genuine competitors to traditional statistical models (Ahmed, Atiya, Gayar, & El-Shishiny, 2010). Nowadays, Machine learning techniques plays very important role in timeseries forecasting. Here in this project, we use both univariate and multivariate analysis. The simplest type of statistical analysis is univariate analysis. It can be inferential or descriptive, just like other types of statistics. The key fact is that only one variable is involved. Univariate analysis can yield misleading results (Palit & Popovic, 2006). But we used automatic univariate ARIMA here, Automatic univariate ARIMA modelling has been shown to produce one-step-ahead forecasts as accurate as those produced by competent modelers (Darmawan & Sriyanti, 2017).

3 Research methodology

We used Kaggle dataset which contains the record of house sales from 2007 to 2019. We analyzed this data and considered just house property type for our project and converted number of bedrooms into columns to fit into our model as shown in table 1.

Then we worked on this dataset to perform single variable and multiple variable time series analysis.

Table 1 – Converted number of bedrooms to columns

bedrooms	2	3	4	5
saledate				
2007-03-31	441854.0	421291.0	548969.0	735904.0
2007-06-30	441854.0	421291.0	548969.0	735904.0
2007-09-30	441854.0	421291.0	548969.0	735904.0
2007-12-31	441854.0	421291.0	548969.0	735904.0
2008-03-31	441854.0	416031.0	552484.0	735904.0

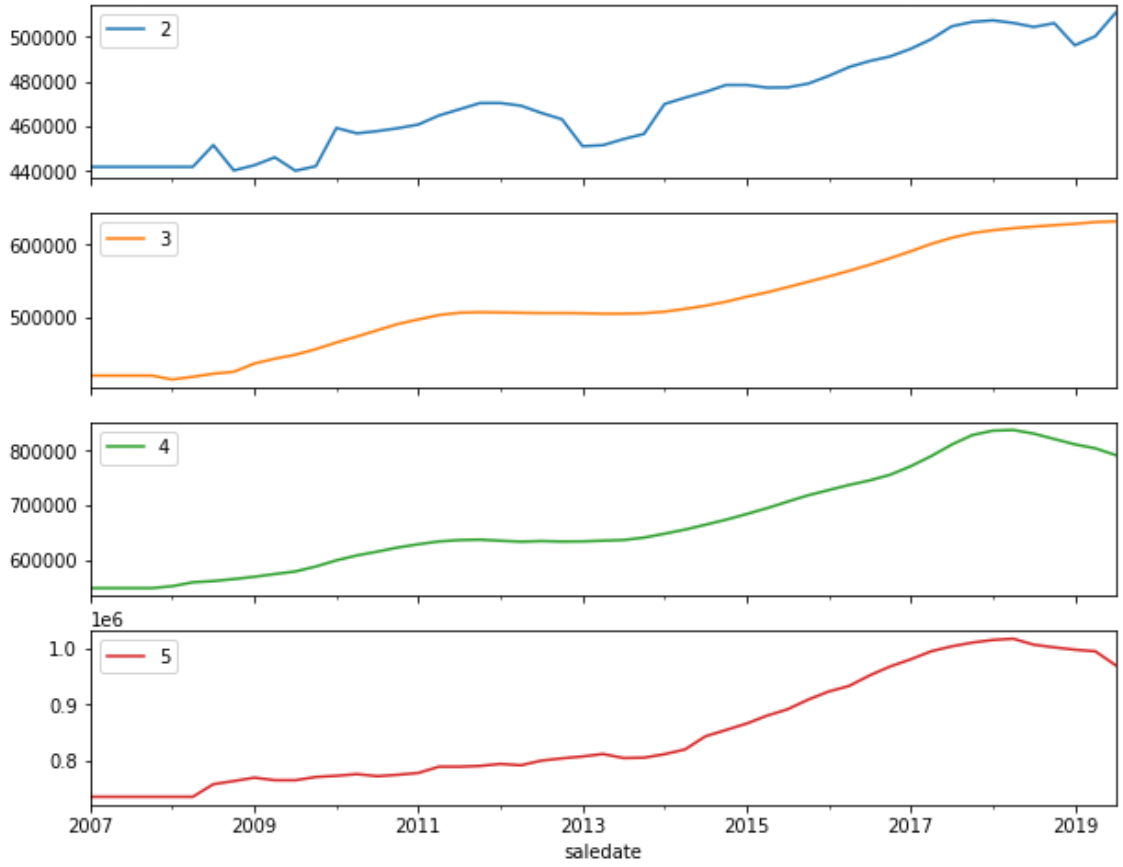


Figure 1- original dataset

4 Results

Table 2 – Test metrics of univariate time series analysis by autoarima model

No. of bedrooms	MAPE	RMSE	Mean (Test data)
2	1.73	10048	503906
3	0.57	1089	627464
4	5.93	56343	815906
5	3.36	40098	998492

Similarly, in table 2, multiple variable model predicted 2 and 3 bedroom prices efficiently than 4 and 5 bedrooms. But when we compare between two models, we can say that LSTM model outperforms VAR model as both MAPE and RMSE scores are lesser.

Table 3 – Test metrics of multivariate time series analysis by VAR and LSTM model

No. of bedrooms	MAPE (VAR)	MAPE (LSTM)	RMSE (VAR)	RMSE (LSTM)
2	1.02	0.86	6632	6043
3	2.35	1.55	15921	10199
4	8.34	4.92	78893	46148
5	7.22	4.87	83350	52059

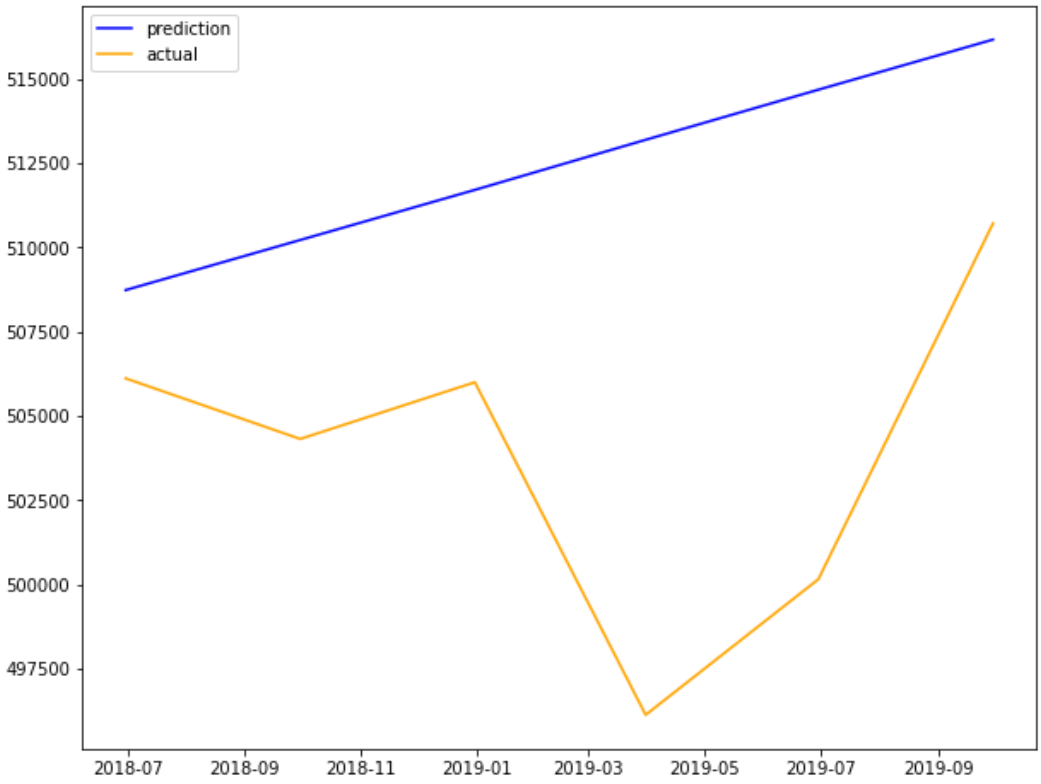


Figure 2- plotting actual and prediction for 2-bedroom using autoarima

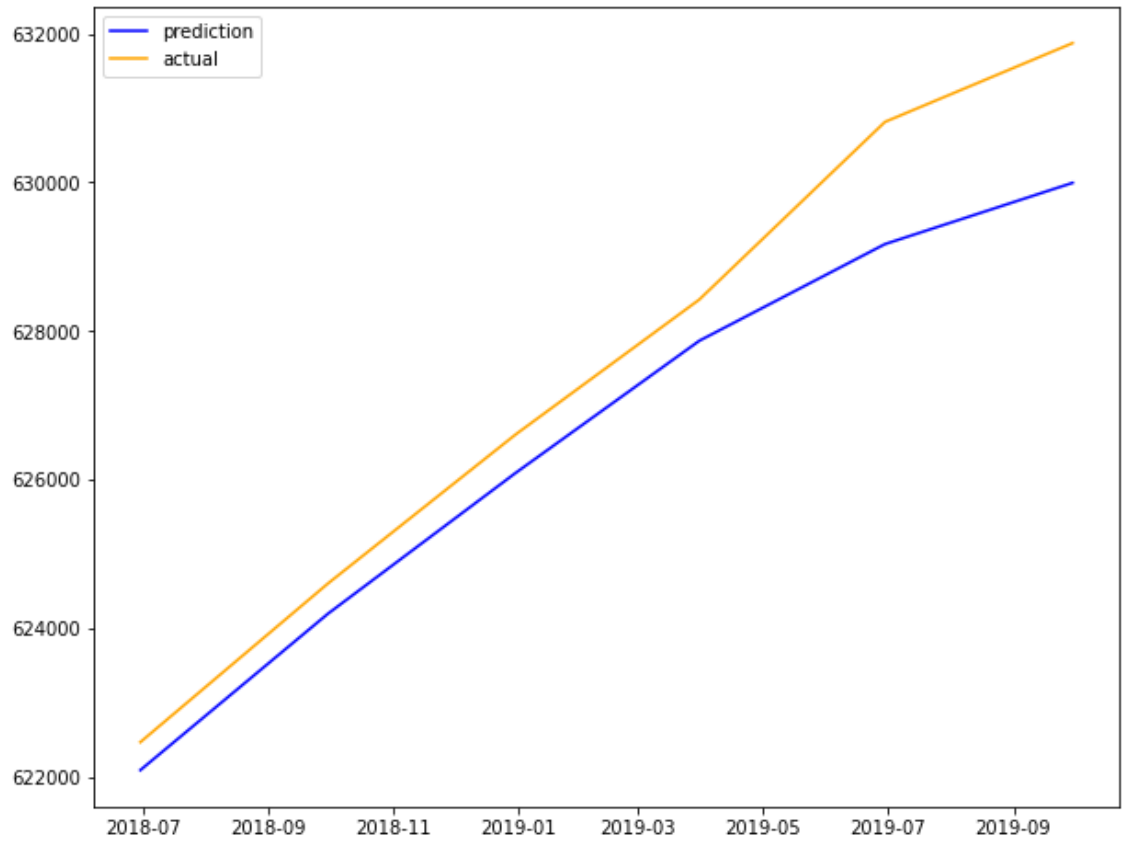


Figure 2- plotting actual and prediction for 3-bedroom using autoarima

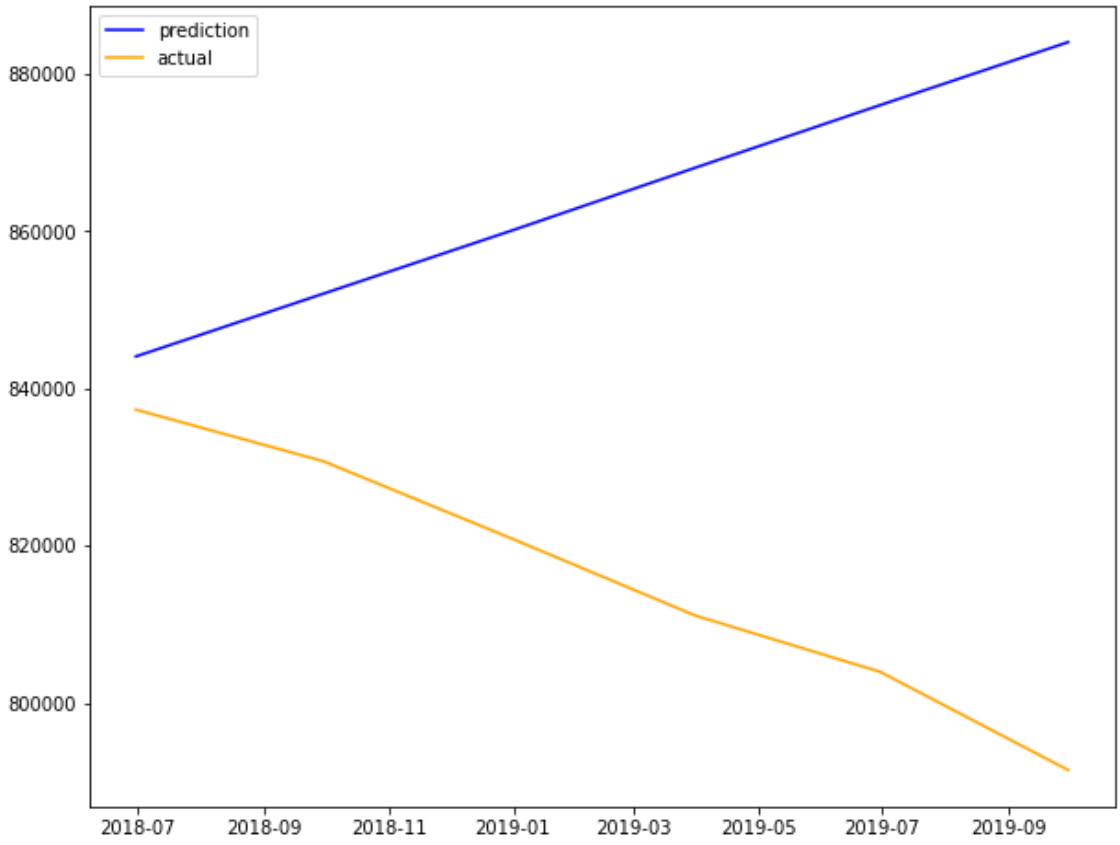


Figure 4- plotting actual and prediction for 4-bedroom using autoarima

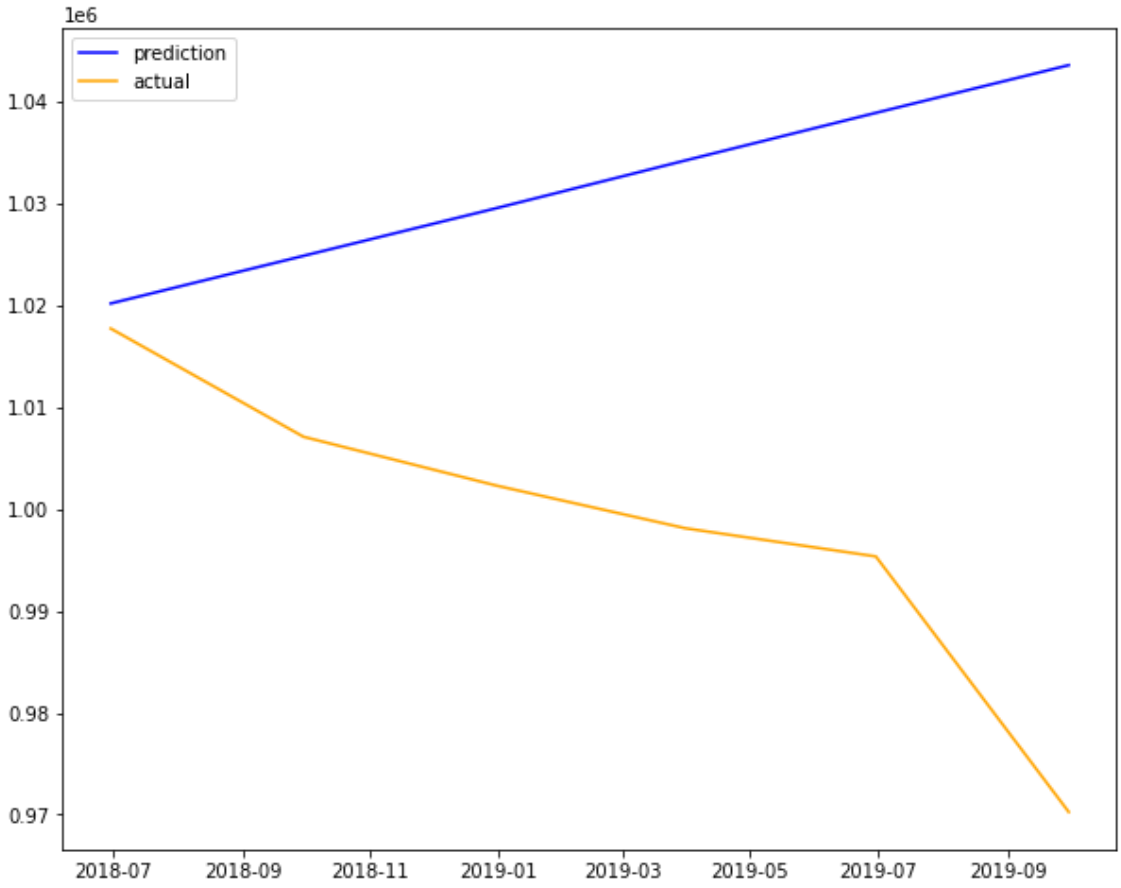


Figure 5- plotting actual and prediction for 5-bedroom using autoarima

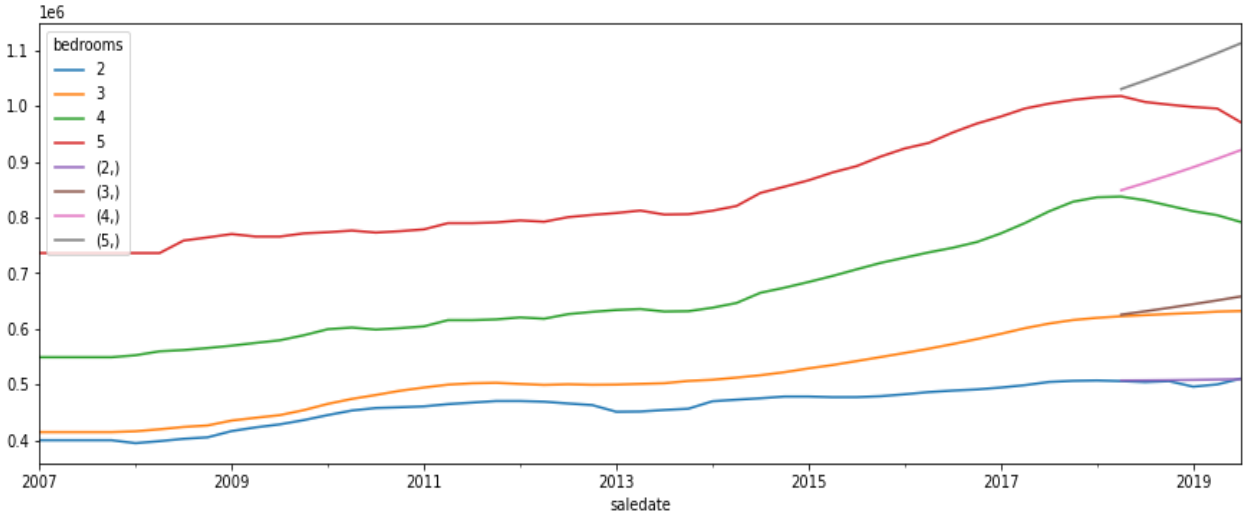


Figure 6- plotting actual and prediction for all-bedroom using VAR

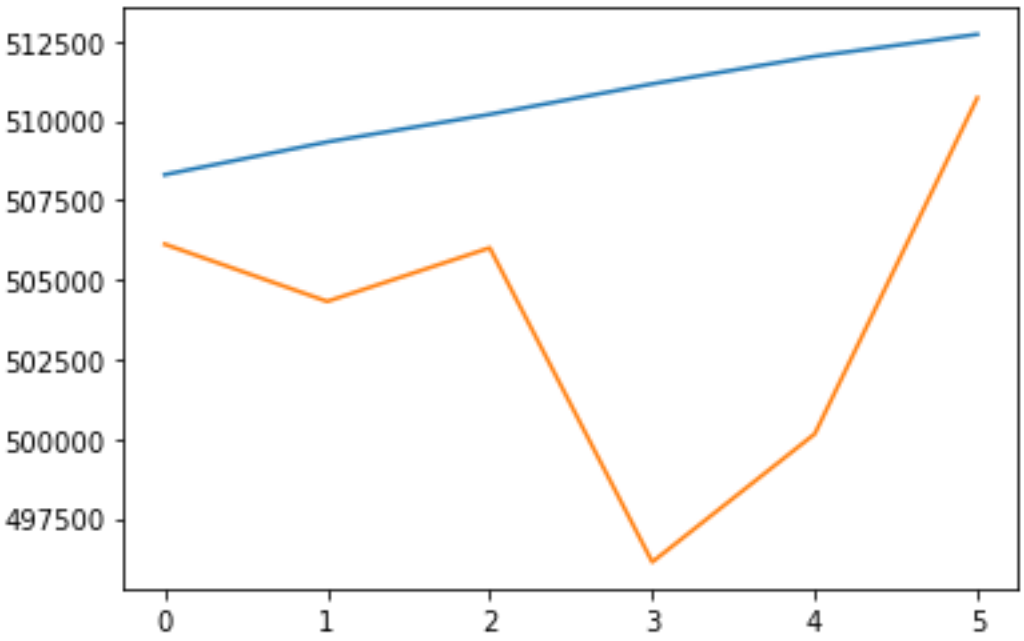


Figure 7- plotting actual and prediction for 2-bedroom using LSTM

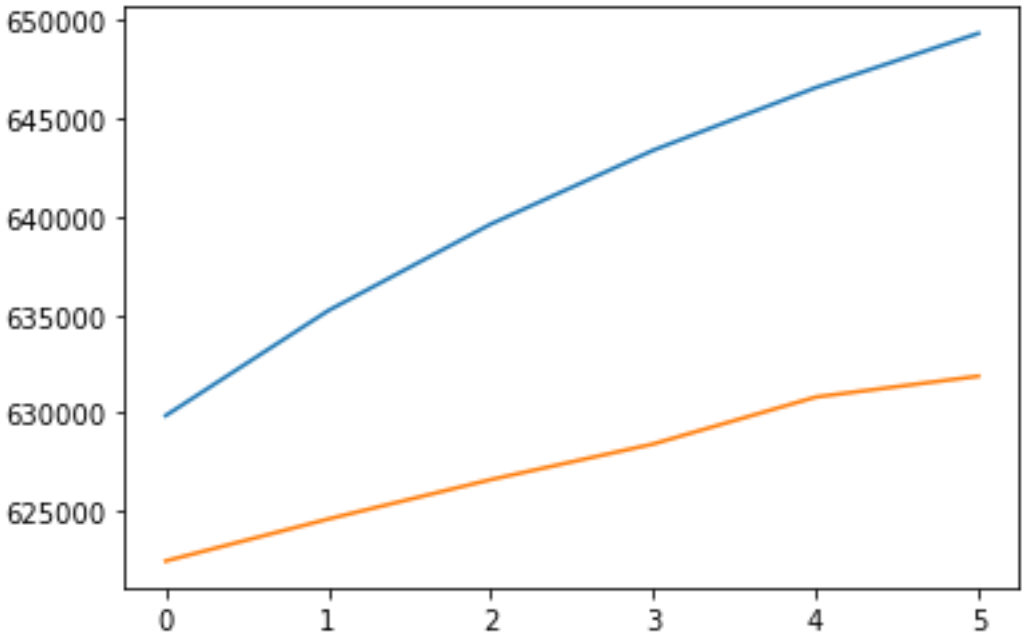


Figure 8- plotting actual and prediction for 3-bedroom using VAR

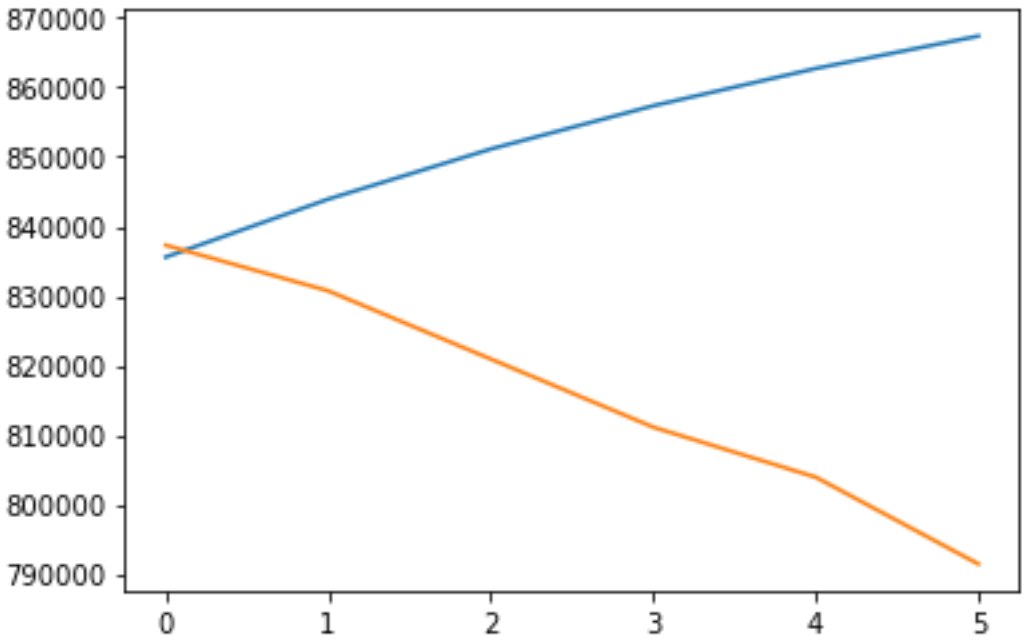


Figure 9- plotting actual and prediction for 4-bedroom using LSTM

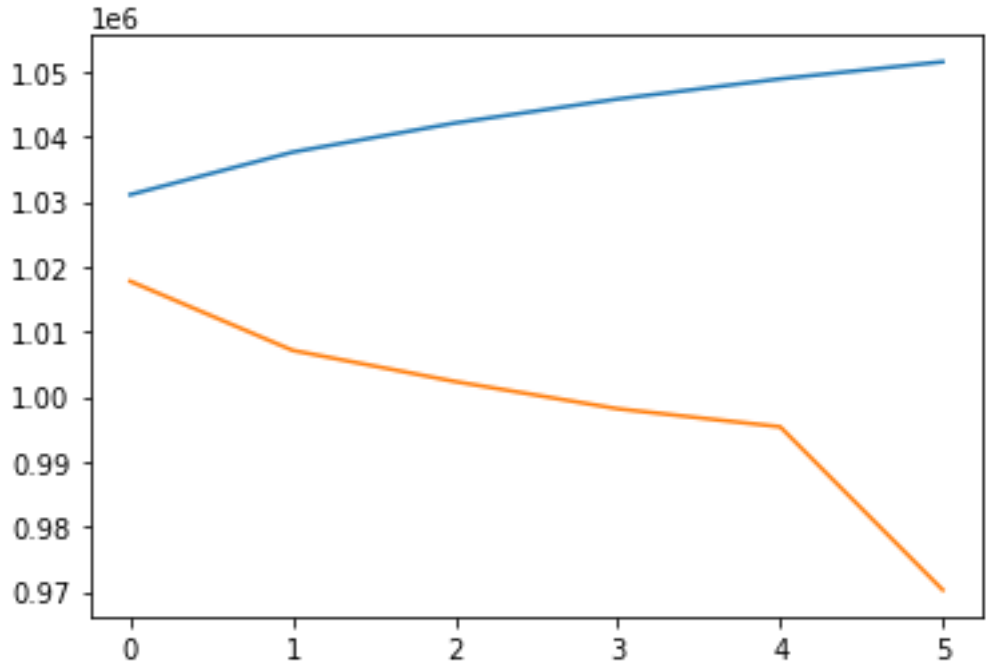


Figure 10- plotting actual and prediction for 5-bedroom using LSTM

5 Discussion and conclusions

5.1 Discussion

Table 1 shows the mean absolute percentage error (MAPE) and root mean square error (RMSE) obtained from univariate time series analysis by using autoarima model for each number of bedroom houses. It is noted from the table that the model predicted quite well with 2- and 3-bedrooms house price because the MAPE values are pretty less, 1.73 and 0.57 percent. It means that on an average our points forecast is off by just 1.73 and 057 percent. This can be also verified by analysing RMSE score of 2 and 3 bedroom as well which are very less compared to its average price. But the model's efficiency reduced in case of 4- and 5-bedroom houses since both MAPE and RMSE scores are increasing.

5.2 Conclusions

We were able to develop a forecasting model for house price. The analysis of result indicated that the model was able to predict house prices efficiently for 2-and 3-bedrooms, however the prediction efficiency decreased for 4-and 5-bedrooms house in both time series analysis, univariate and multivariate.

Since both univariate and multivariate analysis generated models with nearly same prediction result, we concluded that that there was negligible effect of one dependent variable to another variable in our dataset. So, we concluded that our first assumption stating prediction of a house depends only upon its past price, was verified. Therefore, we recommended univariate time series analysis using autoarima model for predicting house price with such type of datasets.

List of figures

References/bibliography

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews, 29*(5-6), 594-621.

Bontempi, G., Ben Taieb, S., & Borgne, Y.-A. L. (2012). *Machine learning strategies for time series forecasting.* Paper presented at the European business intelligence summer school.

Darmawan, S., & Sriyanti, A. (2017). Document details. *Advanced Science Letters, 23*(7), 6524-6526.

Litsiou, K. (2021). *Forecasting the success of megaprojects with Judgmental methods*: University of Salford (United Kingdom).

Palit, A. K., & Popovic, D. (2006). *Computational intelligence in time series forecasting: theory and engineering applications*: Springer Science & Business Media.

Poskitt, D. S., & Tremayne, A. (1986). The selection and use of linear and bilinear time series models. *International Journal of Forecasting, 2*(1), 101-114.

Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019). *House price prediction approach based on deep learning and ARIMA model.* Paper presented at the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT).