Forecasting the Price of Dogecoin Using Time Series Model and Machine Learning Techniques

University of Toronto

Yuanhan Peng

August 19, 2022

Contents

1. Introduction

Different from Bitcoin or other cryptocurrencies, Dogecoin was created as a “joke” by Billy Markus and Jackson Palmer in 2013. Since the logo of Dogecoin is a “Shiba Inu” dog, Dogecoin is considered the first “meme” cryptocurrency, attracting many investors. Due to the pandemic, Dogecoin has become one of the most valuable cryptocurrencies. By Yahoo Finance (2022), its close price grew significantly from 0.0056 on January 1, 2021, to 0.071 on August 17, 2022.

This project aims to use the time series model (ARIMA) and some machine learning models, including SARIMAX, KNN, Random Forest, and SVM, to forecast Dogecoin's close price. Three measurements are determined to evaluate the performance of each model, which are Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Coefficient of determination (). By comparing the performance of each model with the in-sample forecasting method, the best model and corresponding parameters for prediction would be selected. Then, this study would apply the out-of-sample forecasting technique to testify to the accuracy in practical, using the best performance model, where its parameter estimation remains the same.

2. Data

1) Dogecoin

The daily frequency data of Dogecoin in USD, downloaded from Yahoo Finance (Yahoo Finance, 2022), includes six features: date, open price, highest price, lowest price, close price, adjusted close price, and volume. Only the close price data is chosen as the predictor for this project, and the date feature is selected to perform time series models. Generally, the data of Dogecoin will be divided into two sections. The first section, from January 1, 2021, to July 20, 2022, is used for in-sample forecasting for parameter estimation and model selection. There are 566 variables in total. As figure 1 shows, the close price of Dogecoin was relatively low at the beginning of 2021, but then significantly increased in the middle of May 2021. After the boom, although the price of Dogecoin

fluctuated over time, it maintained a downward trend and reached about 0.1 USD in July 2022.

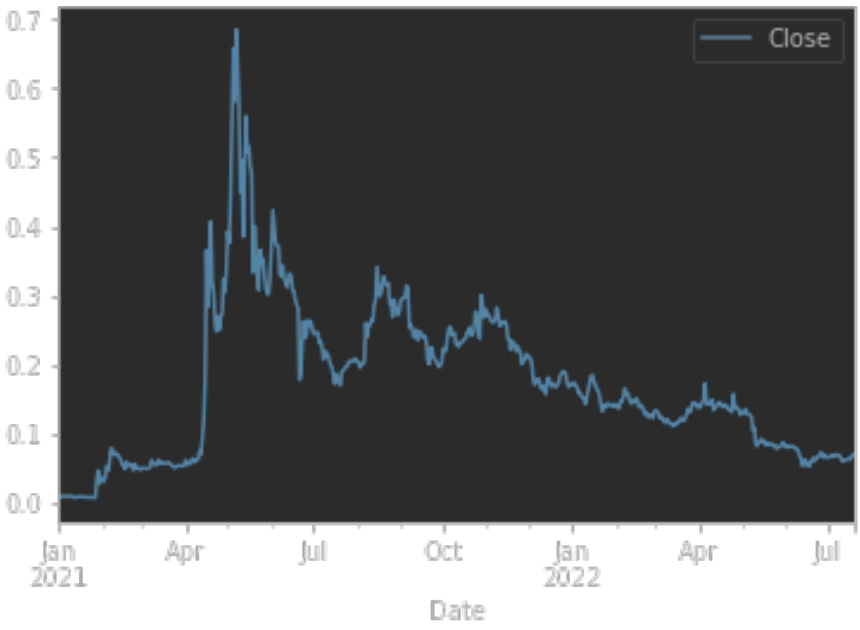

Figure 1

The second section of the data, from July 21 to July 31, 2022, includes the close price of 11 days. This is prepared for the out-of-sample prediction. After the model with the best performance is selected and the corresponding parameters are set, the best model would be applied to forecast the close price of these 11 days and compared with the actual data to test its accuracy. From figure 2, the close price of Dogecoin slightly violated between 0.07 to 0.062 USD.

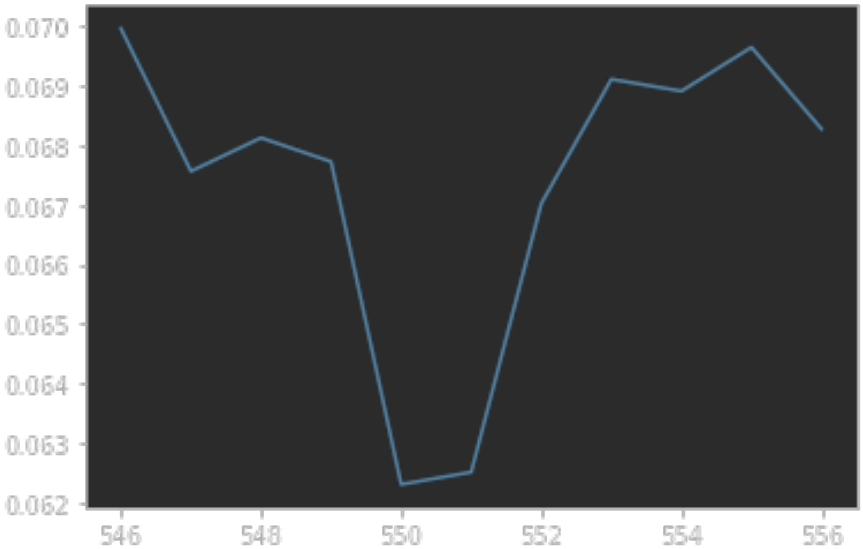

Figure 2

2) ARIMA Model

The precondition of applying the Auto-Regressive Integrated Moving Average model is that the time series data must be stationary, which means that the variance and mean of the data are constant over time. By performing a rolling statistics test, figure 3 showed that the mean and variance of Dogecoin's close price from January 1, 2021, to July 20, 2022, are not constant, which indicates the original data is not stationary. Moreover, the Argument Dicky-Fuller test (ADF), a typical method used to check the stationary of data, could confirm that the original data is not stationary. The result of the ADF test on original close price data has a p-value around 0.08, which is greater than 0.05. In other words, the null hypothesis that the data is non-stationary is not rejected.

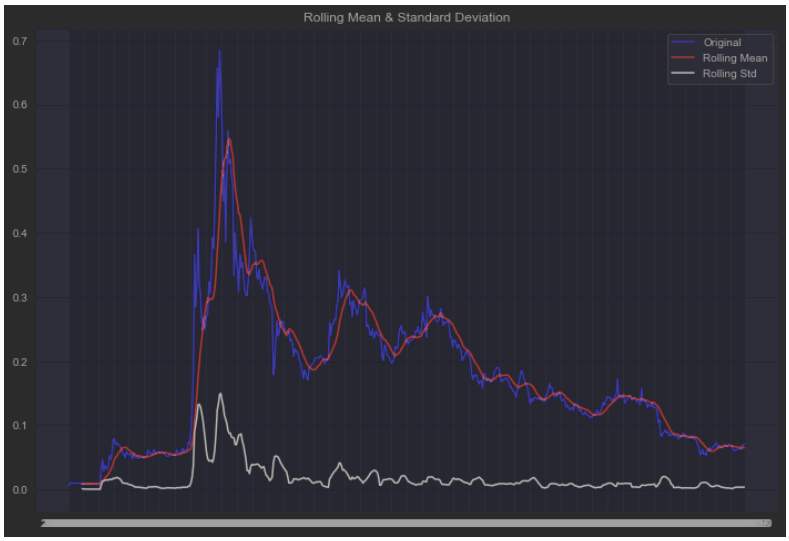

Figure 3

Thus, to make the time series data stationary, it is necessary to take a logarithm of the original close price data and then take the first difference of it, which is described in Figure 4. The number of variables became 565 because of the first differencing process. Then, by applying the ADF test, the p-value of the manipulated data is significantly less than 0.05, which indicates that the null hypothesis is rejected. As a result, the data is now stationary. The auto-correlation function (ACF) and partial auto-correlation function (PACF) are also plotted as Figures 5 and Figure 6 show. Since none of them shows a slow decaying trend, the data is further confirmed to be stationary. The last step is to

split the data into two sets: the training set and the testing set, which include 535 and 30 variables respectively.

$$close\ price\ \rightarrow ln(close\ price) \rightarrow$$
$$diff = \ln(close\ price)\ [i+1] - \ln(close\ price)\ [i]$$
$$where\ i\ \in [0, 565]\ and\ i\ \in \mathbb{Z}$$

Figure 4

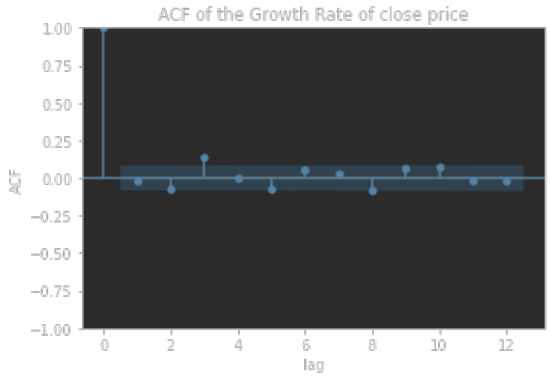

Figure 5

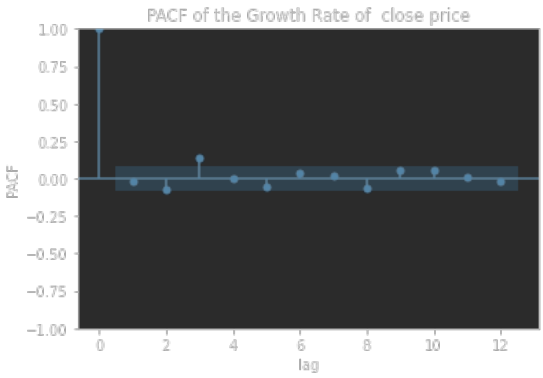

Figure 6

3) SARIMAX Machine Learning Model

Similar to the ARIMA time series model, Seasonal Auto-Regressive Integrated Moving Average with exogenous factors (SARIMAX) model uses data from January 1, 2021, to July 20, 2022, and requires stationary data as well. By taking a logarithm of the original close price data, the ADF test returned a p-value around 0.04, which is less than 0.05, which indicates the null hypothesis is rejected, and the logarithm of the close price is stationary. Next, Figure 7 is the result of decomposing the data, it shows the observed series, trend line, seasonal pattern, and residual plot. The seasonal plot indicates that the time series data display seasonality, with a pattern recurrence occurring weekly. Although the data still has seasonality, the seasonal scale is small because it is from 0.0 to 0.0025; thus, it won't significantly contribute to the accuracy improvement.

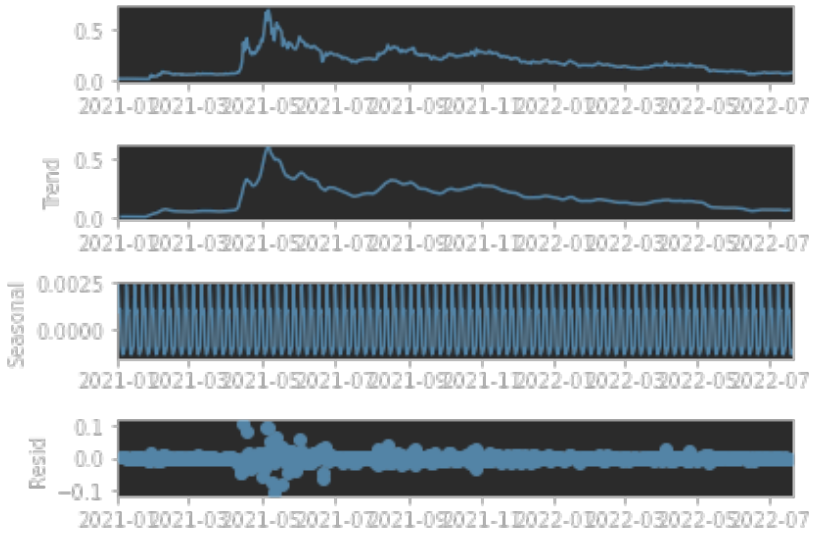

Figure 7

Furthermore, exogenous factors can be added to this model to improve forecasting accuracy. There are six exogenous variables data downloaded from Yahoo Finance and Wind, which can be classified into three aspects. The first group is macro factors, such as the US Dollar Index (named “UDI”), the Close Price of Brent Oil (named “Brent”), the number of new confirmed cases (named “New_Confirmed_Cases”), and the total number of confirmed cases of COVID-19 in the USA (named “YTD”). The second aspect includes MSCI ACWI (named “MSCI_ACWI”), an index usually used in the stock market. The last variable is the close price of Bitcoin (named "BTC_Close"), one of the most famous cryptocurrencies. All the missing values in exogenous variables data are filled in with the nearest value ahead or below them.

By running a random SARIMAX model, the summary result of the model indicates that only the close price of Bitcoin contributes to the improvement of prediction accuracy. Since, as figure 8 shows, only the p-value of the close price of Bitcoin is less than the significant level, which is 0.05, this coefficient is statistically significant. Last but not least, in preparation for the ARIMA model, the data should be split into two sets: training (536 variables) and testing sets (30 variables).

Covariance Type:opg					
	coef	std err	z	P>|z| [0.025	0.975]
UDI	-0.0089	0.022	-0.408	0.683-0.052	0.034
BTC_Close	2.365e-05	2.05e-06	11.549	0.000 1.96e-05	2.77e-05
MSCI_AIWC	0.0011	0.001	0.760	0.447-0.002	0.004
YTD	-5.771e-09	4.33e-08	-0.133	0.894-9.07e-08	7.91e-08
New_Confirmed_Cases	1.273e-08	2.03e-07	0.063	0.950-3.86e-07	4.11e-07
Brent	0.0002	0.004	0.052	0.959-0.007	0.008

Figure 8

4) Other Machine Learning Models (Random Forest, KNN, SVM)

Similarly, all the rest machine learning models, including Random Forest, KNN, and SVM, use the data from January 1, 2021, to July 20, 2022. However, different from the previous two models, the lag variables, a variable related to the past value of the panel data, are created for each model to allow lagged effects of the predictor (Hyndman & Athanasopoulos, 2016). However, the appropriate number of lag variables may vary among different models, and it would be tested for each model. Then, the data will be split into training and testing datasets, which have 536 and 30 variables respectively. For KNN and SVM models, there is one more step that is feature scaling, which can help normalize the data into a particular range. After the models fit and return forecasting values, the returned data needed to transform inversely to receive the actual prediction.

3. Methodology

1) Auto-Regressive Integrated Moving Average Model (ARIMA)

The ARIMA model is a typical and widely used time series model, introduced by Box and Jenkins. This model can analyze univariate time series models and forecast time series data. According to SAS Institute, there are three stages in ARIMA modeling: identify, estimate, and forecast. In the first identification stage, it is necessary to check and make sure that the time series data is stationary. Estimation and tuning of parameters will be in the second stage, and then the model will fit. The last stage is to predict future values and access the forecasting ability by comparing them with the actual data (SAS Institute Inc., 2000, p.

Chapter 7 The ARIMA Procedure). There are three parameters in the ARIMA model: p, d, and q, which are the number of autoregressive terms, the number of non-seasonal differences needed for stationary, and the number of lagged forecast errors in the forecast equation (Nau, 2020). Generally, the equation of the ARIMA (p, d, q)model can be written as:

$$\widehat{Y_t} = \mu + \emptyset_1 Y_{t-1} + \cdots + \emptyset_p Y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

Here are the basic ARIMA methodological steps used in this project (figure 9):

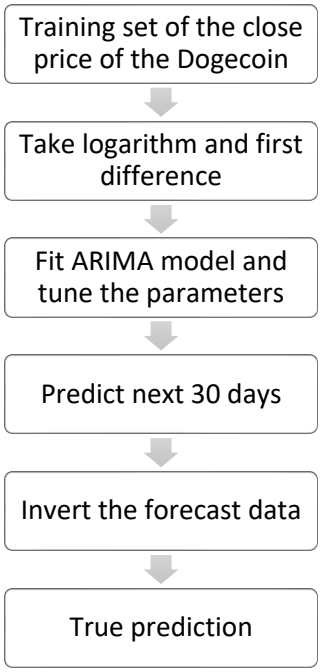

Figure 9

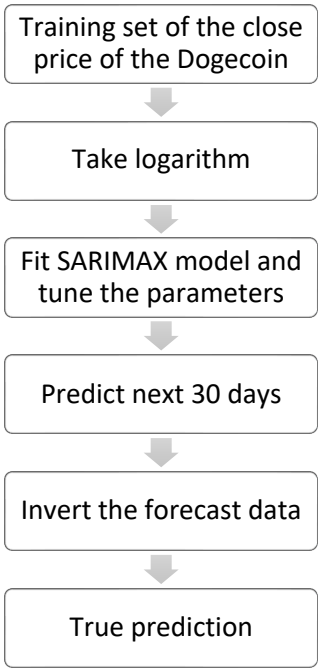

Figure 10

2) Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors Model (SARIMAX)

The SARIMAX is a machine learning model used for time series analysis, which likes an improved version of the ARIMA model since it could capture the effectiveness of seasonality and external variables. Thus, different from the simple ARIMA model, SARIMAX has two more parameters: $(P, D, Q)_s$ and (X), representing the seasonal order and external variable respectively. The process of utilizing the SARIMAX model is shown in figure 10.

In order to find the appropriate parameters, the grid search method would be applied to select the model with the smallest Akaike Information Criterion (AIC) usually used to measure how the model fits the data in statistics.

3) K-Nearest Neighbors Model (KNN)

KNN is a supervised machine learning algorithm, used for classification and regression. By analyzing the similarity between features, the KNN model can predict the values for new points. The distance between a new data point and every point from the training set will be calculated, which may use distance functions like Euclidean, Manhattan, and Minkowski. Next, based on the distance, the closest k neighbours will be selected, whose weighted average would be the prediction value (Kohli, Godwin, & Urolagin, 2020). As a result, the number of neighbours is a significant parameter for the performance of this model. The graph (figure 11) below shows the process of using KNN regression to forecast.

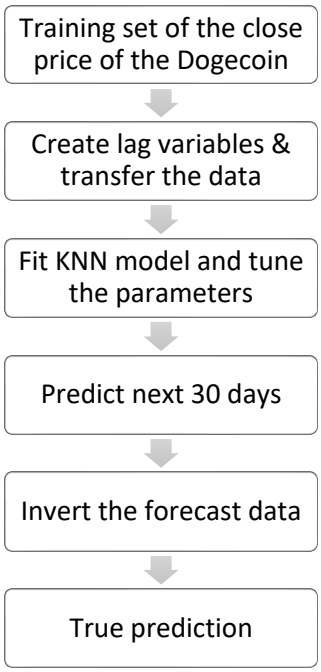

Figure 11

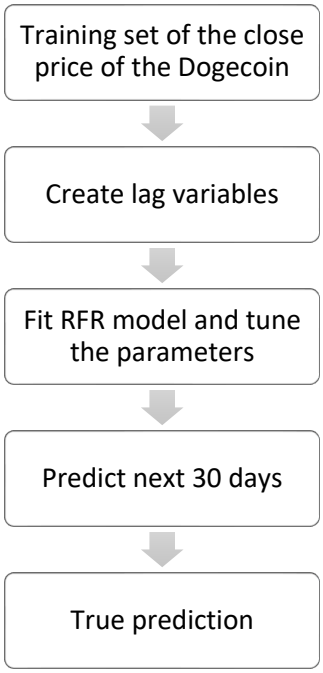

Figure 12

4) Random Forest Model

Random Forest is a supervised machine learning method, always used for classification and regression. For each decision tree, the value of each leaf usually represents the mean of the observations in a specific area, and each tree would return a prediction value. Since Random Forest is an ensemble of decision trees, the average of all prediction values from each tree would be the final prediction for the forest; therefore, the Random Forest Regressor always has higher accuracy than the decision tree. In this case, the number of decision trees may affect the performance of the random forest. Similarly, as the process of using KNN regressor, the flowchart of applying Random Forest Regressor (RFR) is shown in figure 12.

5) Support Vector Machines Model (SVM)

Like other machine learning models used in this project, the SVM is a supervised machine learning model used for classification and regression. Different from the KNN model, the SVM separates classes by deciding on a hyperplane, instead of calculating the distance between two points. Therefore, how to find an appropriate hyperplane would strongly influence the accuracy of this model. In other words, the choice of kernel plays an important role in the performance of this model. There are many types of kernels, such as linear, non-linear, and radial basis functions (RBF). Here is the flowchart for applying SVM:

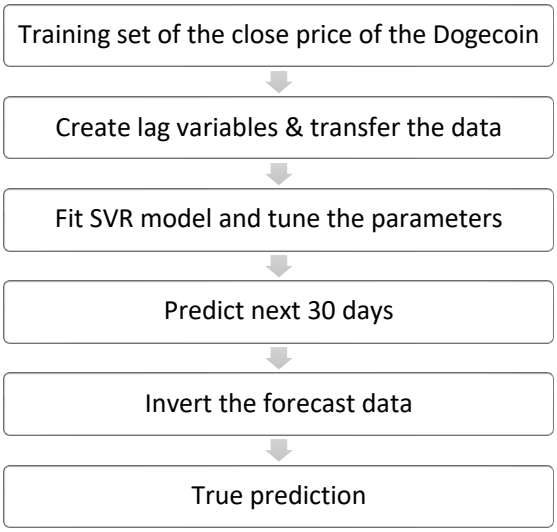

Figure 13

4. Result

The performance of each model is evaluated based on the following three factors commonly used in similar research: Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). In general, the model with a smaller MSE, and smaller is more desired and has better performance.

a) MSE represents the average squared difference between the estimated values and the actual values.

$$MSE = \frac{1}{n}\sum(y - \hat{y})^2$$

b) MAPE is always used to measure the forecasting accuracy in statistics.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|$$

Here are the performance graphs for each model using the in-sample prediction technique:

1) Performance of ARIMA Model

After selecting the model with the smallest MSE, the best ARIMA model has an ARMA order (5, 0, 4). Figure 14 shows the actual Dogecoin close price, and the predicted time forecast for the Dogecoin close price from June 21 to July 20, 2022.

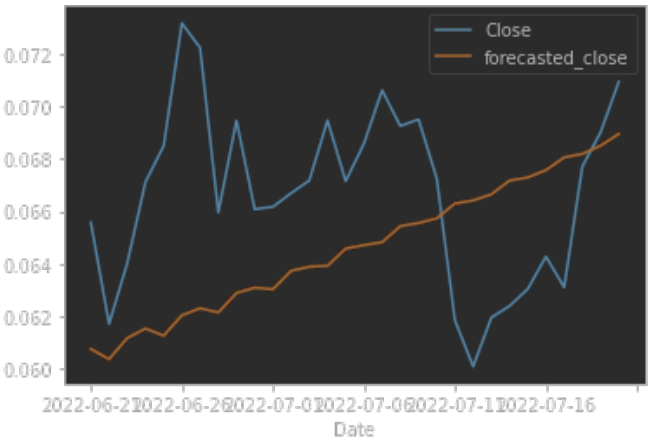

Figure 14

2) Performance of SARIMAX Model

The best SARIMAX model with the smallest AIC has parameters (1, 0, 3)×(2, 1, 3, 7) and the external variable used to improve accuracy is the Bitcoin close price.

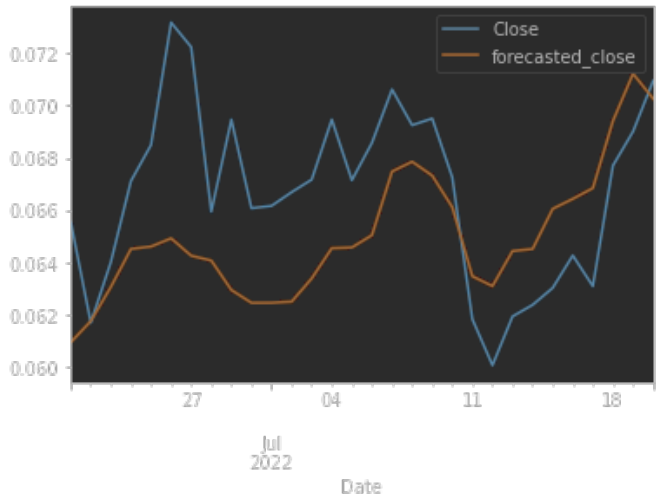

Figure 15

3) Performance of KNN Model

When the number of lag variables is 4, the number of neighbours is 7, and all the other parameters in the KNN model remain as the default values, this model has the best performance. Here is the plot shows the actual values and predicted values:

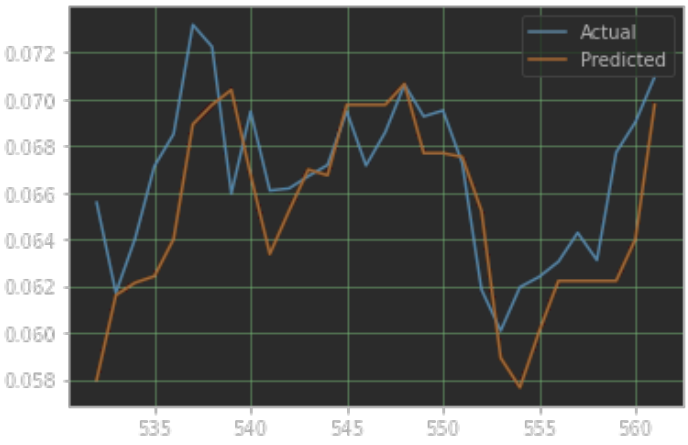

Figure 16

4) Performance of Random Forest Model

By comparing the performance of the Random Forest model with the different number of estimators, such as 100, 500, and 1000, the model fitted best with the data when the number of estimators is 1000, and the number of lag variables is 3.

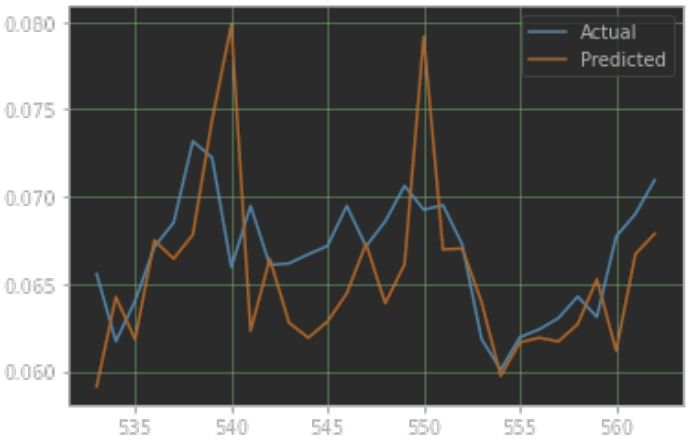

Figure 17

5) Performance of SVM Model

When the kernel is RBF and the number of lag variables is 20, the SVM model performed best, compared with its performance when the kernel is linear and non-linear. By the overlap between two lines, which represents the actual data and predicted data, figure 18 implies that the performance of SVM is better than the rest models in this project.

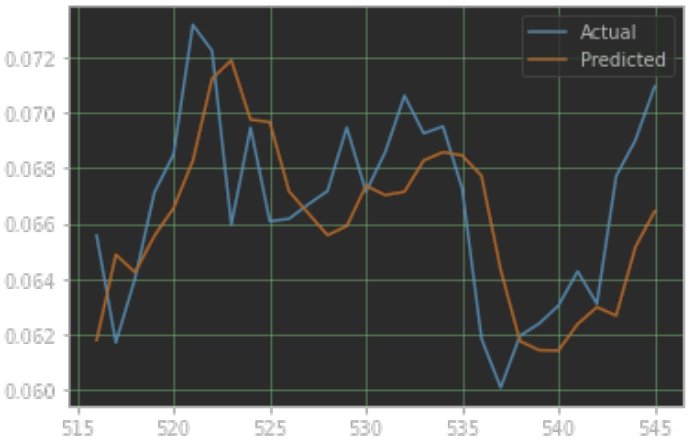

Figure 18

In all, the SVM model has the smallest MSE and MAPE and the highest accuracy, which indicates that SVM is the best model for predicting the close price of Dogecoin. Here is the summary table of the three measurements for each model:

Model	MSE	MAPE	Accuracy (100%-MAPE)
ARIMA	2.3900349201360483e-05	5.582765484399139	94.42%
SARIMAX	1.3312504129531544e-05	5.40052321881958	94.60%
KNN	9.215537760544227e-06	3.5703664629135456	96.43%
Random Forest	2.1263415174878986e-05	4.923552645485006	94.96%
SVM	7.516501284206293e-06	3.240963507905344	96.76%

Figure 19

In the end, by using the out-of-sample method with the SVM model, the accuracy of forecasting the close price of Dogecoin from July 21 to July 31, 2022, approached 96.8%. Its MSE is 7.227917674045728e-06, and its MAPE is 3.1987196860232965. Figure 20 shows the graphs for both actual and predicted data.

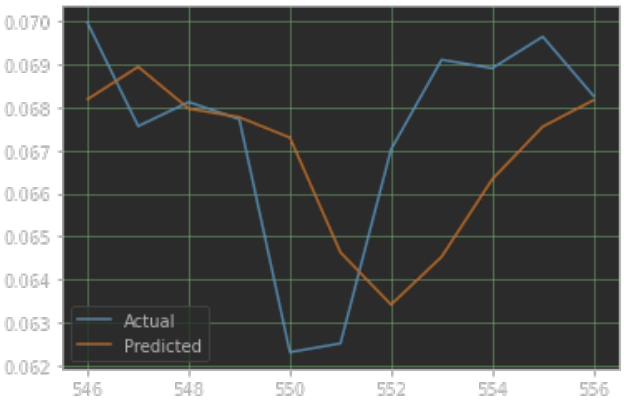

Figure 20

5. Conclusion

Our time series dataset contains the daily close price of Dogecoin. After cleaning and manipulating the data according to each model's preconditions, we applied five models: ARIMA, SARIMAX, KNN, Random Forest, and SVM, to forecast the close price of Dogecoin. For the in-sample forecasting stage, SVM regression performed the best with the highest accuracy, which was about 96.76%, while

the second-best model is KNN, whose accuracy approaches 96.43%. The rest three models had similar accuracies, around 94%. The second stage of this study is to perform out-of-sample forecasting, using the best model SVM with the same parameter to testify to its accuracy, which approached 96.8%. Moreover, its MSE was 7.227917674045728e-06, and its MAPE was 3.1987196860232965. Generally, the SVM model performed well in forecasting Dogecoin's close price. However, since the R^2 of the SVM model in the out-of-sample stage was negative, it indicates that this study and the overall performance of the SVM model in forecasting could be further improved with more appropriate parameter estimation.

Reference

Hyndman, R. J., & Athanasopoulos, G. (2016). *Forecasting: Principles and Practice 2nd.* Springer.

Kohli, S., Godwin, G. T., & Urolagin, S. (2020, July 26). Sales Prediction Using Linear and KNN Regression. *Advances in Machine Learning and Computational Intelligence*, pp. 321–329. https://link.springer.com/chapter/10.1007/978-981-15-5243-4_29#citeas

Iqbal, M., Iqbal, M. S., Jaskani, F. H., Iqbal, K., & Hassan, A. (2021, July 07). Time-Series Prediction of Cryptocurrency Market using Machine Learning Techniques. *EAI Endorsed Transactions on Creative Technologies, 8*(28). doi: 10.4108/eai.7-7-2021.170286

Nau, R. (2020, August 18). *Statistical forecasting: notes on regression and time series analysis.* Retrieved from https://people.duke.edu/~rnau/411home.htm

SAS Institute Inc. (2000, February). SAS OnlineDoc®, Version 8. https://dms.umontreal.ca/~duchesne/chap7.pdf

Yahoo Finance. (2022). *Yahoo Finance*. Retrieved from Yahoo Finance: https://finance.yahoo.com/quote/DOGE-USD/history?period1=1609459200&period2=1660694400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true