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

University of Toronto

Yuanhan Peng

August 19, 2022

1. Introduction

In contrast to Bitcoin and other cryptocurrencies, Dogecoin emerged in 2013 as an unconventional entrant to the digital currency space. Conceived as a humorous response to the rapidly expanding cryptocurrency market by creators Billy Markus and Jackson Palmer, Dogecoin features the Shiba Inu dog as its mascot, positioning it as the inaugural "meme" cryptocurrency. This unique branding has garnered substantial attention from investors. The global pandemic further catalyzed Dogecoin's rise to prominence, marking it as one of the most valuable cryptocurrencies in recent times. According to Yahoo Finance (2022), Dogecoin's closing price experienced a notable increase, moving from $0.0056 on January 1, 2021, to $0.071 on August 17, 2022.

This research endeavors to forecast the closing price of Dogecoin utilizing a blend of time series and machine learning methodologies, including the Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX), K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM). To assess the efficacy of these models, three metrics will be employed: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). By evaluating each model's performance through in-sample forecasting, this study aims to identify the most accurate model and its optimal parameters for prediction. Subsequently, the chosen model will be applied to out-of-sample forecasting to verify its real-world applicability, maintaining consistency in parameter estimation.

2. Data

1) Dogecoin

The dataset on Dogecoin's daily transactional activities in USD, sourced from Yahoo Finance for the year 2022, encompasses six attributes: date, opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume. For the purpose of this study, only the closing price is utilized as the predictive variable, with the date

attribute facilitating the implementation of time series analyses. This dataset is bifurcated into two distinct phases. The initial phase, spanning from January 1, 2021, to July 20, 2022, serves for in-sample forecasting to estimate parameters and select the appropriate model, incorporating a total of 566 data points. As depicted in Figure 1, Dogecoin's closing price commenced at a modest level in early 2021 but witnessed a notable surge by mid-May of the same year. Despite subsequent price volatility, a general declining trend was observed, culminating in a value of approximately $0.10 by July 2022.

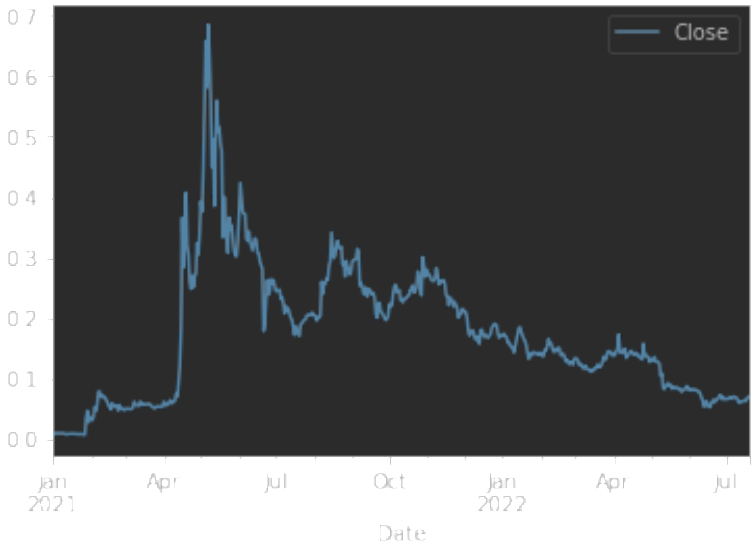

Figure 1

The subsequent segment of the dataset covers the period from July 21 to July 31, 2022, comprising the closing prices for 11 days, designated for out-of-sample forecasting. Following the identification of the most efficacious model and the establishment of its parameters, this model will be employed to predict the closing prices for these 11 days. The predicted values will then be juxtaposed with the actual data to evaluate the model's predictive accuracy. As illustrated in Figure 2, during this period, the closing price of Dogecoin exhibited minor fluctuations, ranging between $0.07 and $0.062.

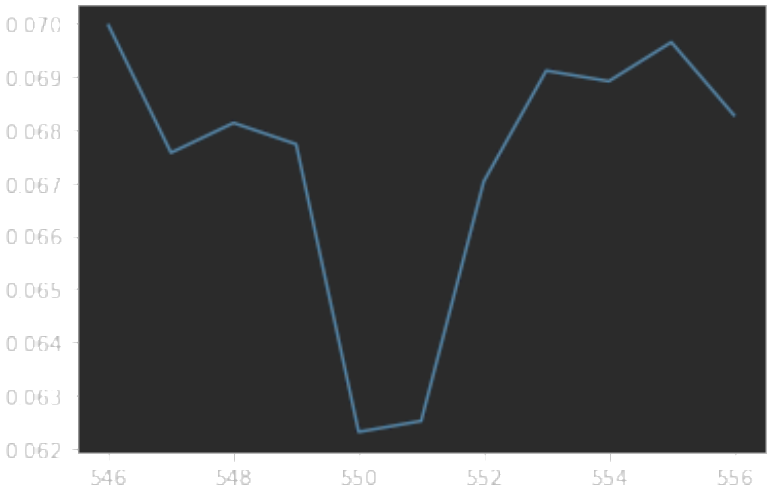

Figure 2

2) ARIMA Model

A fundamental prerequisite for deploying the Auto-Regressive Integrated Moving Average (ARIMA) model is the stationarity of the time series data, implying consistent variance and mean over time. A rolling statistics analysis, as depicted in Figure 3, reveals that the mean and variance of Dogecoin's closing price from January 1, 2021, to July 20, 2022, exhibit variability, thereby suggesting the data's non-stationarity. Further affirmation of this characteristic comes from the Augmented Dickey-Fuller (ADF) test, a commonly employed method for assessing data stationarity. The ADF test applied to the original closing price data yields a p-value of approximately 0.08, surpassing the threshold of 0.05. This outcome indicates a failure to reject the null hypothesis, confirming the non-stationary nature of the data.

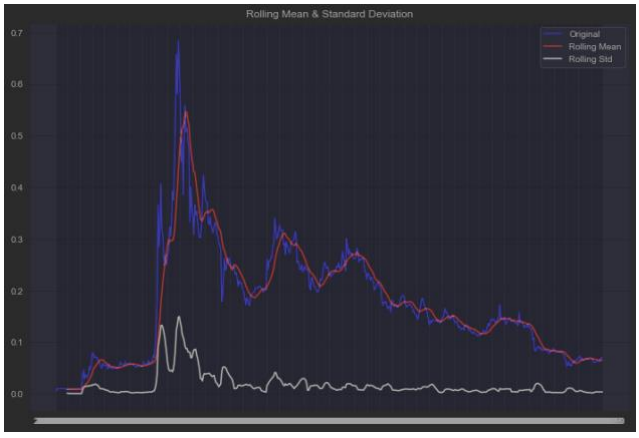

Figure 3

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.

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

Figure 4

The implementation of first differencing on the dataset reduced the total number of variables to 565. This preprocessing step aimed to achieve data stationarity was validated through a subsequent Augmented Dickey-Fuller (ADF) test on the modified dataset, which returned a p-value significantly below the 0.05 threshold. This result leads to the rejection of the null hypothesis, confirming the dataset's transition to stationarity. The establishment of data stationarity is further corroborated by the auto-correlation function (ACF) and partial auto-correlation function (PACF) plots, as illustrated in Figures 5 and 6. The absence of a slow decaying trend in these plots reaffirms the dataset's stationary status. The final phase involves dividing the dataset into two distinct segments: a training set comprising 535 data points and a testing set consisting of 30 data points, thereby setting the stage for subsequent model training and evaluation processes.

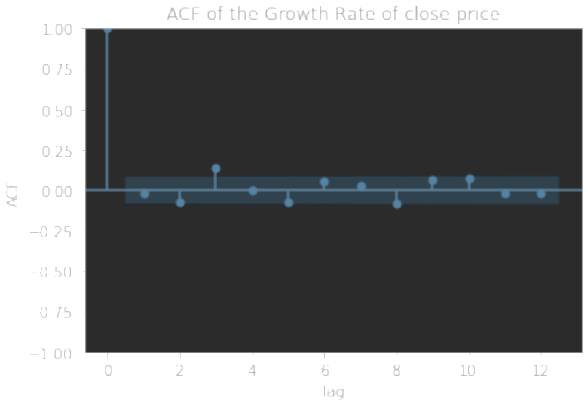

Figure 5

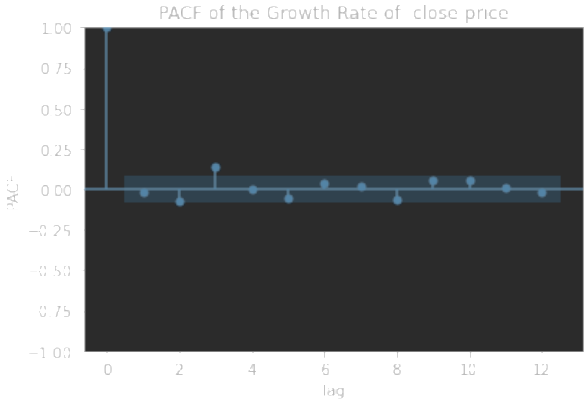

Figure 6

3) SARIMAX Machine Learning Model

Similar to the ARIMA model, the Seasonal Auto-Regressive Integrated Moving Average with Exogenous factors (SARIMAX) model also utilizes the dataset spanning from January 1, 2021, to July 20, 2022, necessitating that the data be stationary. By applying a logarithmic transformation to the original closing price data, stationarity is further pursued. The Augmented Dickey-Fuller (ADF) test conducted on the logarithmically transformed data yields a p-value of approximately 0.04, falling below the 0.05 significance level. This outcome signifies the rejection of the null hypothesis, thereby confirming that the logarithmic transformation renders the close price data stationary.

Subsequent analysis involves decomposing the transformed dataset to examine its underlying components, as depicted in Figure 7. This decomposition visualizes the observed series, trend line, seasonal patterns, and residual fluctuations. The seasonal component of the analysis reveals discernible seasonality within the data, characterized by weekly recurring patterns. Despite the presence of seasonality, its magnitude is relatively minimal, ranging from 0.0 to 0.0025. This suggests that while seasonality is a feature of the dataset, its influence on enhancing predictive accuracy is likely marginal due to its limited scale.

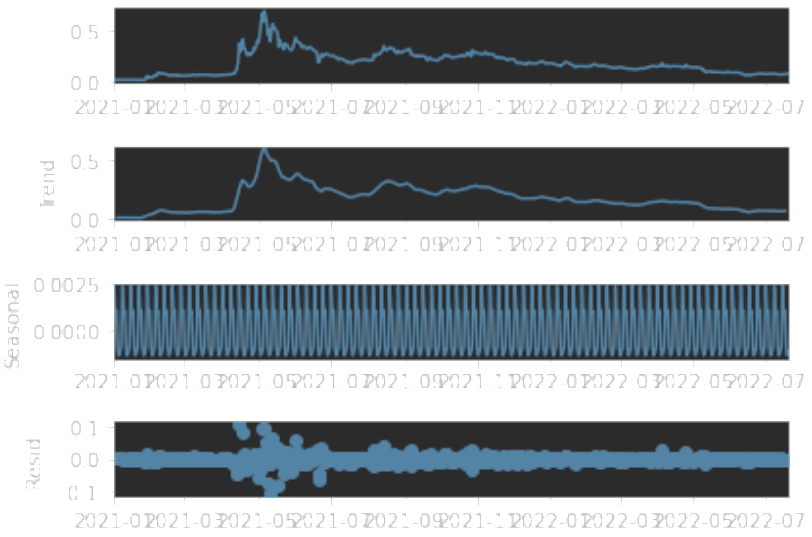

Figure 7

Incorporating exogenous factors into the SARIMAX model holds potential for enhancing the precision of forecasts. Data for six exogenous variables, sourced from Yahoo Finance and Wind, can be categorized into three distinct groups based on their relevance and impact.

The first category encompasses macroeconomic indicators and pandemic-related statistics, including the US Dollar Index ("UDI"), the closing price of Brent Oil ("Brent"), the daily count of new COVID-19 cases ("New_Confirmed_Cases"), and the cumulative number of confirmed COVID-19 cases in the USA to date ("YTD"). These factors are presumed to influence Dogecoin's price indirectly through their impact on global financial stability and investor sentiment.

The second category comprises a single variable, the MSCI All Country World Index ("MSCI_ACWI"), which is widely recognized as a benchmark for global equity markets. This index reflects overall stock market performance and investor confidence, potentially correlating with cryptocurrency market movements.

The final variable considered is the closing price of Bitcoin ("BTC_Close"), arguably the most influential cryptocurrency. Given Bitcoin's dominance in the cryptocurrency market, its price movements are often indicative of broader market trends, which could similarly affect Dogecoin.

To address any gaps in the dataset, missing values among these exogenous variables have been imputed with the nearest preceding or following non-missing value, ensuring a complete dataset for analysis. This method of imputation helps maintain the integrity of the time series data, allowing for a more accurate assessment of how these diverse factors may contribute to forecasting Dogecoin's price dynamics.

The outcomes from an exploratory SARIMAX model reveal that none of the included predictors significantly enhances forecast accuracy. This is evidenced by the observation that all predictors display a p-value greater than the established significance level of 0.05, as illustrated in Figure 8. Given this lack of statistical significance, the logical next step involves excluding these non-contributory predictors from the model

and conducting a refit to potentially improve its forecasting performance. Additionally, in anticipation of employing the ARIMA model, the dataset should be methodically divided into two distinct portions: a training set encompassing 536 data points and a testing set consisting of 30 data points. This division lays the groundwork for a systematic approach to model training and evaluation, ensuring a rigorous assessment of the model's predictive capabilities.

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
Close	0.9726	0.009	114.252	0.000	0.956	0.989
UDI	-0.0034	0.006	-0.552	0.581	-0.016	0.009
BTC_Close	3.214e-07	6.85e-07	0.469	0.639	-1.02e-06	1.66e-06
MSCI_ACWI	0.0002	0.000	0.671	0.502	-0.000	0.001
YTD	3.304e-10	1.82e-09	0.182	0.856	-3.23e-09	3.89e-09
New_Confirmed_Cases	-3.165e-08	7.97e-08	-0.397	0.691	-1.88e-07	1.24e-07
Brent	6.708e-05	0.002	0.038	0.970	-0.003	0.004
ar.L1	-0.8694	0.007	-117.919	0.000	-0.884	-0.855
ma.L1	0.9295	0.007	129.018	0.000	0.915	0.944
ar.S.L7	-0.0002	6.69e-05	-2.656	0.008	-0.000	-4.66e-05
ma.S.L7	-0.9702	0.001	-1904.418	0.000	-0.971	-0.969
sigma2	0.0120	0.000	47.390	0.000	0.012	0.013

Ljung-Box (L1) (Q):	2.16	**Jarque-Bera (JB):**	39166.40
Prob(Q):	0.14	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.10	**Skew:**	3.43
Prob(H) (two-sided):	0.00	**Kurtosis:**	44.96

Figure 8

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

The Random Forest, KNN, and SVM models, like their predecessors, utilize the dataset spanning from January 1, 2021, to July 20, 2022. A distinctive approach for these models involves the generation of lag variables, which capture the delayed effects of a predictor on the variable of interest, as highlighted by Hyndman and Athanasopoulos in 2016. The determination of an optimal number of lag variables is model-specific and subject to empirical testing to identify the configuration that yields the most accurate predictions.

Subsequent to the creation of lag variables, the dataset is partitioned into training and testing sets, comprising 536 and 30 data points, respectively. An additional preparatory step for the KNN and SVM models involves feature scaling. This process standardizes the range of the data features, ensuring that no single feature dominates the model due to its scale, thereby enhancing model performance and the reliability of predictions.

Following the fitting of these models and the generation of forecast values, an inverse transformation of the predicted data is necessary. This final step converts the scaled or transformed forecast values back to their original scale, enabling a direct comparison with actual historical data and an assessment of the model's forecasting accuracy.

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 analyzes univariate time series models and forecasts 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 basic ARIMA methodological steps used in this project (figure 9):

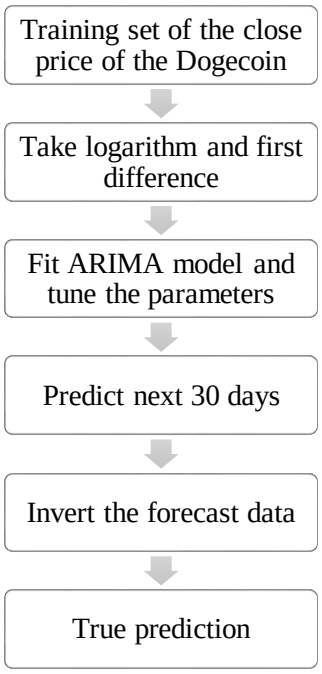

Figure 9

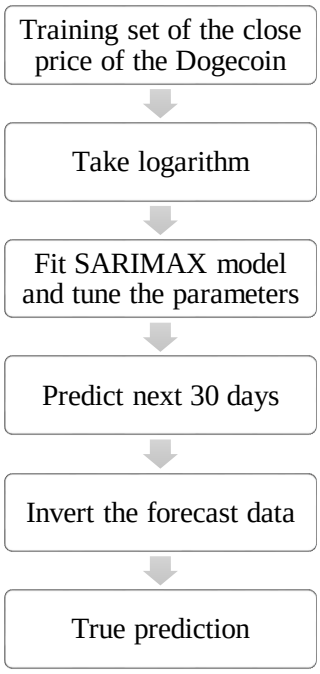

Figure 10

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 neighbors will be selected, whose weighted average would be the prediction value (Kohli, Godwin, & Urolagin, 2020). As a result, the number of neighbors is a significant parameter for the performance of this model. The graph (figure 11) below shows the process of using KNN regression to forecast.

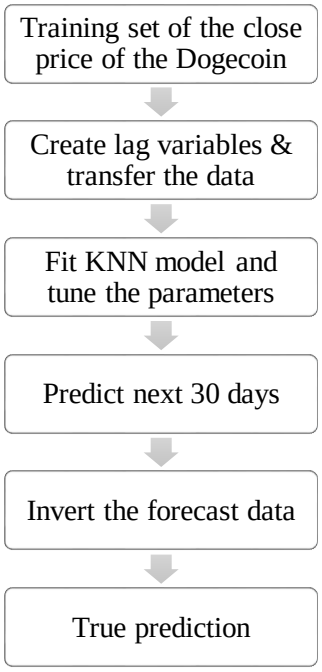

Figure 11

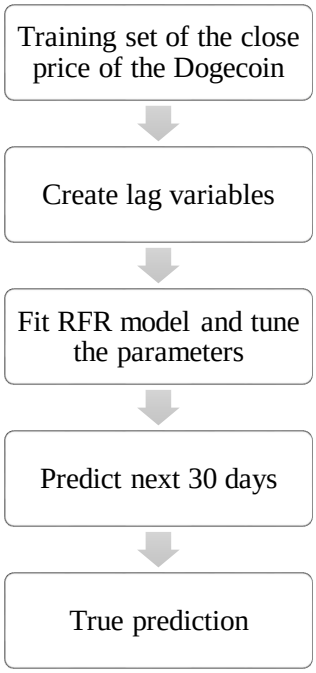

Figure 12

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 of applying SVM:

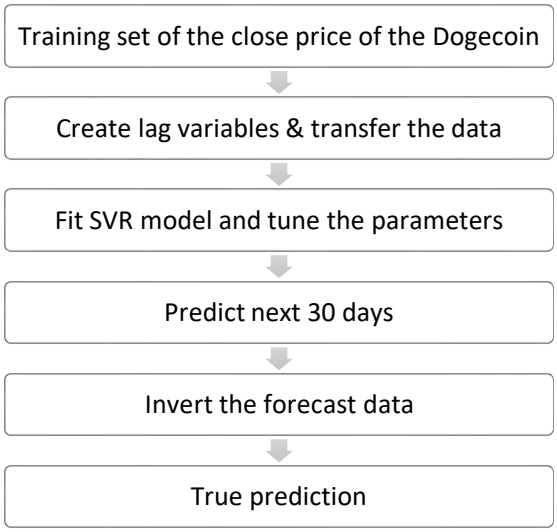

Figure 13

4. Result

The performance of each model is evaluated based on the following three factors commonly used in similar research: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Coefficient of determination (R^2). In general, the model with smaller MSE, smaller MAPE but higher R^2 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.

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

c) R^2 indicates the goodness of the fit of the model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Here are the performance graphs for each model using 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.

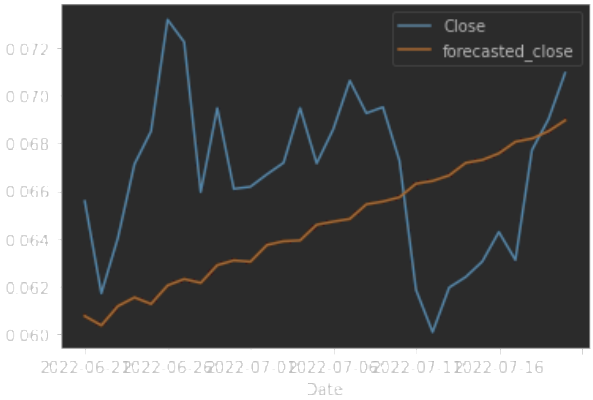

Figure 14

2) Performance of SARIMAX Model

The best SARIMAX model with smallest AIC has parameters (1, 0, 3)×(2, 1, 3, 7) and the none of the external variable can be used to improve accuracy.

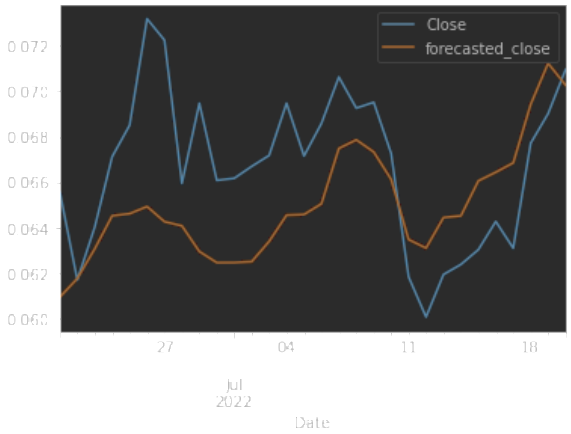

Figure 15

3) Performance of KNN Model

When the number of lag variables is 4, the number of neighbors 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:

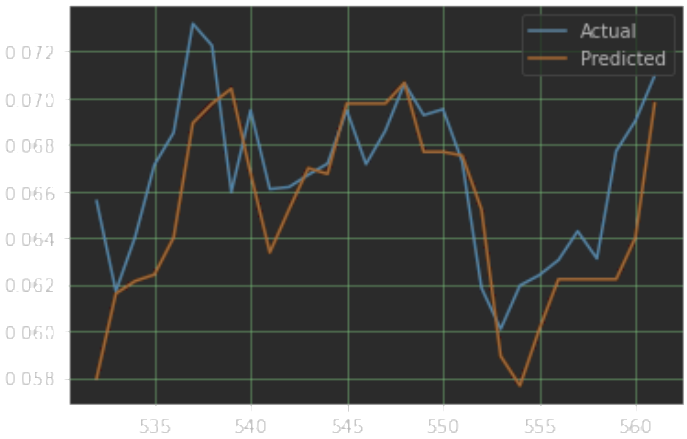

Figure 16

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 variable is 3.

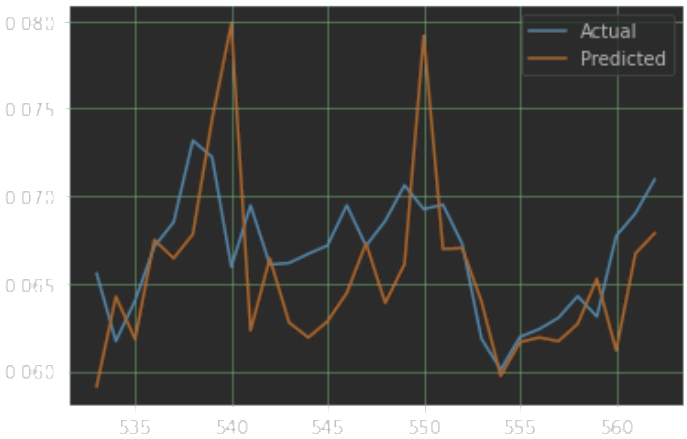

Figure 17

5) Performance of SVM Model

When the kernel is RBF and the number of lag variable 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.

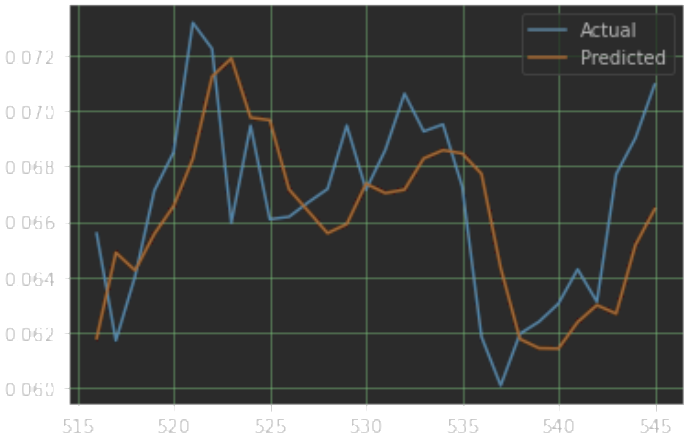

Figure 18

Overall, the KNN and SVM models outperform the others with the highest accuracies of 96.43% and 96.76%, respectively, attributed to their lower MAPE values and positive R^2 values, indicating a better fit to the data compared to the other models. The SVM model, in particular, demonstrates the lowest MSE and a favorable R^2 value, marking it as the most effective model in this analysis for predicting Dogecoin's closing price. Conversely, the SARIMAX model shows the least favorable outcomes with the highest MSE, highest MAPE, and a significantly negative R^2 value, translating to an accuracy of just 87.03%. This suggests a less accurate model fit and predictive capability. Below is the summary table of the three measurements for each model:

Model	**MSE**	**MAPE**	**R^2**	**Accuracy (100%-MAPE)**
ARIMA	2.390034920 1360483e-05	5.582765484 399139	-1.2068778120554655	94.42%
SARIMAX	8.590652286 213458e-05	12.69349205 1808628	-6.932319215004968	87.03%
KNN	9.215537760 544227e-06	3.570366462 9135456	0.1490682567622608	96.43%

Random Forest	2.126341517 4878986e-05	4.923552645 485006	-0.9451844070036048	94.96%
SVM	7.516501284 206293e-06	3.240963507 905344	0.2055625365282191	96.76%

Figure 19

In summary, while KNN and SVM models show promising results in forecasting accuracy, SARIMAX, ARIMA, and Random Forest models exhibit challenges that could be addressed with further refinement and parameter tuning. Negative R^2 values, although mathematically anomalous, serve as a stark indicator of a regression model's suboptimal performance. In practical terms, such occurrences suggest that the assumed regression line is even less effective in explaining the variance in the data than simply using the mean value as a predictor. This underscores the critical importance of R^2 as a diagnostic tool, providing insights into the validity and utility of the regression model in question.

In our study, the out-of-sample forecasting method employing the SVM model yielded highly accurate predictions for Dogecoin's closing price from July 21 to July 31, 2022, achieving an impressive accuracy rate of approximately 96.8%. Additionally, the model exhibited a remarkably low Mean Squared Error (MSE) of 7.227917674045728e-06 and a Mean Absolute Percentage Error (MAPE) of 3.1987196860232965. These metrics affirm the SVM model's efficacy in generating precise forecasts, underscoring its practical utility in financial forecasting applications. The graphical representation in Figure 20 further illustrates the alignment between the actual and predicted data, validating the reliability of the SVM model's forecasts.

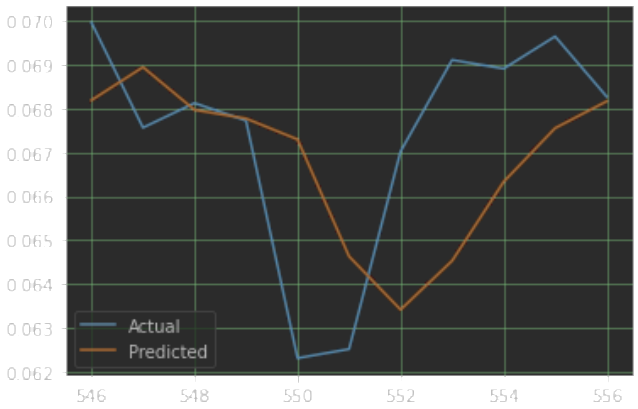

Figure 20

5. Conclusion

Our study focused on analyzing the daily close price of Dogecoin through a rigorous process of data cleaning and manipulation tailored to meet the requirements of each forecasting model. Employing five distinct models—ARIMA, SARIMAX, KNN, Random Forest, and SVM—we sought to forecast Dogecoin's close price accurately. In the in-sample forecasting phase, the SVM regression model emerged as the top performer, boasting an accuracy of 96.76%, followed closely by the KNN model with an accuracy of 96.43%. The remaining three models exhibited comparable accuracies, hovering around 94%.

Moving forward, we conducted out-of-sample forecasting using the SVM model, the best-performing model from the in-sample phase, to evaluate its predictive accuracy. Impressively, the SVM model achieved an accuracy rate of approximately 96.8%, accompanied by a remarkably low Mean Squared Error (MSE) of 7.227917674045728e-06 and a Mean Absolute Percentage Error (MAPE) of 3.1987196860232965. Despite these favorable outcomes, it's noteworthy that the SVM model exhibited a small positive R^2 value in the out-of-sample phase, indicating room for improvement in the model's parameter estimation process.

In summary, our findings underscore the efficacy of the SVM model in forecasting Dogecoin's close price, particularly in the out-of-sample context. However, our study is not without limitations. While we meticulously cleaned and manipulated the data to suit each forecasting model's requirements, the exclusion of certain external factors, such as market

sentiment and regulatory changes, may have limited the comprehensiveness of our analysis. Additionally, our focus on solely analyzing the daily close price of Dogecoin overlooks other relevant factors that could influence price movements, such as trading volume or social media trends. These omissions could potentially impact the robustness and applicability of our forecasting models. Therefore, future research should aim to incorporate a broader range of variables and consider additional external factors to enhance the accuracy and reliability of cryptocurrency price forecasts.

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