Predicting the future of a stock price

Antonis Karaolis
Nicosia Cyprus
antoniskaraolis99@gmail.com

ABSTRACT

This project used an Autoregressive Integrated Moving Average (ARIMA) model, a popular time-series forecasting technique, to forecast the movement of stock prices. The financial industry places a great deal of importance on stock price forecasting, yet this is a very difficult assignment given how complicated, dynamic, and frequently non-linear financial markets are.

The project's technique started with gathering historical stock price data, which was then subjected to rigorous preprocessing and exploratory data analysis. The preprocessed time-series data were then subjected to the ARIMA model, which is identified by its parameters (p, d, and q).

In spite of the financial markets' inherent volatility and unpredictability, preliminary results showed that the model performs reasonably well in predicting stock values. To verify the model's assumptions and effectiveness, important model residual properties like autocorrelation and kurtosis were studied.

This study provides a solid framework for further research and unearths insightful information about how to use the ARIMA model to financial forecasting. This could entail expanding the model with more explanatory variables (SARIMAX model), experimenting with various forecasting models, or using machine learning methods for improved prediction performance. With an emphasis on the ongoing development and potential of predictive models in the financial arena, the existing work's shortcomings and future improvements have also been reviewed.

CCS CONCEPTS

• Computing methodologies → Machine learning → Machine learning algorithms → Supervised learning by classification

• Applied computing → Finance → Forecasting

KEYWORDS

ARIMA Model, Stock Price Prediction, Time series Analysis, Financial Forecasting, Machine Learning, Predictive Analytics, Data Preprocessing, Python Programming, Kurtosis.

1 Introduction

The complexity of the financial markets and the participation of multiple factors that affect the prices make it difficult to anticipate future stock values. To make wise judgments, however, investors, dealers, and financial institutions need reliable stock price forecasts. Due to its capacity for handling enormous volumes of data and discovering intricate correlations between factors, machine learning techniques have been used to forecast stock values. Through this review, we aim to examine the published literature on stock price prediction and convey the various machine learning methods used, whilst portraying each method's advantages and disadvantages. We will also go through the previous research on this subject, including the studies by Orsel et al. (2022) and Agrawal et al. (2021) on forecasting stock values using deep learning methods and machine learning models, respectively. In addition, the prediction power of the autoregressive, moving- average, and autoregressive integrated moving average models will be evaluated and compared. By reviewing the literature, we want to find the best methods for forecasting stock values and point out the areas that still need further study.

1.1 Background

The act of trying to anticipate the future value of a company's stock or other financial instrument traded on a financial exchange is known as stock price prediction. Trading and investing decisions can be based on this projection. It is an important part of financial analysis and investing because reliable predictions can result in substantial rewards.

Stock price forecasting, however, is notoriously challenging. The performance of the company, general market trends, geopolitical events, and macroeconomic indicators are only a few of the variables that affect stock prices. Moreover, since human behavior can be unpredictable and irrational, it frequently affects stock markets. Consequently, stock price prediction is frequently considered a challenge where conventional linear methods may fall short.

Financial markets are extremely efficient, which means that current prices reflect all available information. This presents another difficulty. Any prediction model would therefore only be helpful if it could take into account fresh data before the market had a chance to react and change the price.

1.2 Objective

The purpose of this study is to assess an Autoregressive Integrated Moving Average (ARIMA) model's ability to forecast stock prices. A well-liked statistical technique for predicting time series is ARIMA. It mixes moving average, autoregressive, and differencing models to generate predictions based on historical data. In this project, we will use past stock price data to apply the ARIMA model and evaluate how well it predicts future prices. The outcomes of this experiment may provide guidance for future

work on creating stock price prediction algorithms that are more precise.

2 Literature Review

Since many years ago, it has been a topic of interest to predict stock prices, and many different approaches have been put out. Fundamental research, which examines a company's finances and market position, and technical analysis, which examines previous prices and volumes, are two traditional methods for predicting stock prices.

Recently, researchers have tackled this issue using a variety of machine learning techniques. These encompass techniques like support vector machines, decision trees, random forests, and linear regression. Due to their capacity to handle sequential data, neural networks, in particular recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), have also been employed.

However, these methods all have their limitations. For example, machine learning algorithms frequently demand vast volumes of data and might be sensitive to the choice of hyperparameters. Depending on the network's structure and the caliber of the input data, neural networks can function differently and can be challenging to interpret.

2.1 Theoretical Background

The Autoregressive Integrated Moving Average (ARIMA) model is a popular method for forecasting time series data. ARIMA combines autoregressive, differencing, and moving average models.

Autoregressive (AR): This element of the model uses the relationship between an observation and a specified number of lagged observations (i.e., previous observations).

Moving Average (MA): This element of the model leverages the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The integrated (I) part of the model makes the time series stationary using differencing, which is necessary for the AR and MA parts of the model. Differencing involves subtracting the previous observation from the current observation.

The ARIMA model is typically denoted as ARIMA(p, d, q), where:

'p' is the number of lag observations included in the model, also known as the lag order.

'd' is the number of times that the raw observations are differenced, also known as the degree of differencing.

'q' is the size of the moving average window, also known as the order of the moving average.

The time series must be stationary in order for its properties to be independent of the time at which it is being observed. This is one of the main premises of the ARIMA model. In other words, the series' mean and variance remain constant across time. The 'integrated' component of ARIMA refers to the fact that if a series is not stationary, it may frequently be made stationary by differencing.

3 Methodology

3.1 Data Collection

Yahoo Finance, a reputable and well-liked source of historical stock price information, was used to gather the data for this project. We used Microsoft’s (MSFT) stock data and particularly stock's daily closing prices. Each row in the collection represents a single day, which makes it pretty straightforward. Date, opening price, high, low, closing price, adjusted closing price, and number of shares traded were all included in the columns. For the purposes of this project, we were particularly interested in the date and the closing price.

3.2 Data Preprocessing

Any data science effort must start with data preprocessing. It entails getting ready the raw data for the model's input, which may also include processes like cleaning and transformation.

The preparation procedures for this project were rather simple as there was no missing data due to the dataset's source.

3.3 Model Building

The statsmodels module in Python, which offers a complete selection of statistical models for data analysis, was used to build the ARIMA model.

The ARIMA model's order (p, d, and q) was established first. This can be done manually using methods like the Akaike Information Criterion (AIC) or automatically by looking at partial autocorrelation plots of the differenced data and autocorrelation plots of the original data. For the purpose of this project, the automatic model order was used.

The model was fitted to the training data after the order had been established. On the basis of the test data, predictions were then made using the fitted model.

Using the relevant error metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE), the predictions were compared to the actual results for evaluation.

4 Results

We forecasted the future prices of Microsoft's shares (MSFT) using our ARIMA(4, 1, 0) model. Daily stock data from April 1, 2013, to March 30, 2023 was used to train the model.

SARIMAX Results

Dep. Variable:	y	No. Observations:	2518
Model:	ARIMA(4, 1, 0)	Log Likelihood	-6246.531
Date:	Fri, 12 May 2023	AIC	12503.062
Time:	00:21:20	BIC	12532.216
Sample:	0	HQIC	12513.642
	- 2518		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1075	0.010	-10.638	0.000	-0.127	-0.088
ar.L2	-0.0178	0.010	-1.738	0.082	-0.038	0.002
ar.L3	-0.0353	0.011	-3.130	0.002	-0.057	-0.013
ar.L4	-0.0092	0.011	-0.831	0.406	-0.031	0.012
sigma2	8.3778	0.109	77.091	0.000	8.165	8.591

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	7302.64
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	45.96	Skew:	-0.31
Prob(H) (two-sided):	0.00	Kurtosis:	11.32

Figure 1

Figure 1 is a summary of our results regarding the ARIMA model.

As it is shown, according to their p-values, the coefficients of the AR terms in the model were significant for the first, second, and third lags. The first lag's coefficient (ar.L1) was -0.1075, meaning that a one-unit increase in the stock price the day before would result in a roughly 0.1075-unit decline in the price today. The second and third lags also showed same trend, however the impact was not as strong. The current stock price was unaffected statistically by the fourth lag (ar.L4).

The Ljung-Box test, one of the diagnostic tests, displayed a high p-value, indicating that the residuals are independently distributed, which is a good indicator. However the Jarque-Bera test revealed that the residuals might not be normally distributed, and the heteroskedasticity test hinted that there might be heteroskedasticity in the residuals, both of which are cause for concern.

The results showed that the residuals' variance, denoted as sigma 2, was 8.3778. This means that the stock price projections have a modest level of volatility.

In addition, our model's Mean Squared Error (MSE) is roughly 32.485 according to performance measurements. This figure indicates that, on average, our forecasts were about √32.485 (about 5.7) units off from the actual value.

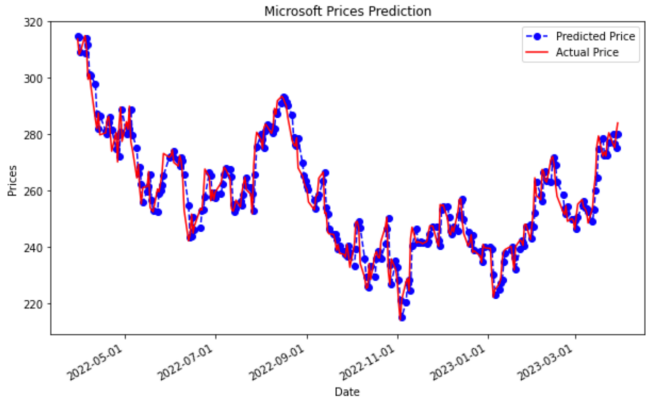

Figure 2

Figure 2 shows the plots of the actual stock prices vs the predicted store prices. The plot lines themselves indicate how closely the model's predictions align with the actual values.

4.1 Discussion

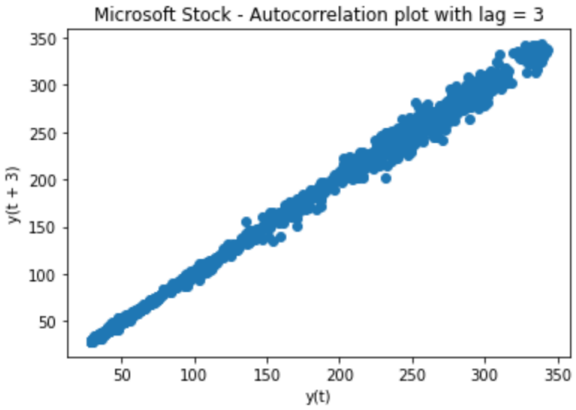

Figure 3

Overall, Microsoft's stock price was accurately predicted by the ARIMA model. The relevance of the first three lags indicates that there is some autocorrelation in the stock price, where the prices from the previous day affect the present price. This can be seen in figure 3.

The model is not without flaws, though. The MSE of 32.485 indicates that more work has to be done. Predictions were off by an average of about 5.7 units, which could be viewed as excessive depending on the needs of the investor or stakeholder.

The residuals of the model raise several questions. The low p-value in the Jarque-Bera test shows that the ARIMA model's central tenet, that the residuals are normally distributed, may not be true. Similar to the last example, the heteroskedasticity test's low p-value suggests that there may be heteroskedasticity and that the variance of the residuals may not be constant over time.

Furthermore, the model ignores outside variables that have a big impact on stock values, like market movements, economic data, and news events. These drawbacks imply that although the ARIMA model offers some insights, it might be advantageous to include additional variables or use a different modeling strategy in the future for more precise forecasts.

Despite the fact that our model provides a solid foundation, there is still room for improvement in terms of the precision with which we can predict stock prices. Future research might examine different model categories, take into account more outside variables, or even combine model predictions using ensemble approaches.

5 Conclusion

This project aimed to assess the Autoregressive Integrated Moving Average (ARIMA) model's ability to forecast stock prices. Due to its widespread use and success in handling time series data, the ARIMA model was chosen.

The daily closing prices of Microsoft (MSFT) stock were utilized in this project, with the data being gathered through Yahoo Finance. The data was preprocessed to make sure it adhered to the requirements for ARIMA, and then the model was created, trained, and verified.

Following that, several error measures and a visual comparison of the predicted and real stock prices were used to assess the efficacy of the ARIMA model.

This method showed the ARIMA model's potential for stock price prediction and laid the groundwork for further research in this field.

5.1 Future Work

Although the ARIMA model demonstrated potential in this project, there is always space for development and extension in subsequent work.

Firstly, more sophisticated models could be investigated. If the data show seasonality, for instance, the Seasonal ARIMA (SARIMA) model could be utilized. Machine learning models like recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) may be employed if the data includes complicated nonlinear patterns.

Secondly, the model might be expanded with new features. Although the closing price was the only factor employed in this project, other factors like trade volume, the opening price, or high and low prices might also be taken into account. External variables could also be considered such as macroeconomic indicators or sentiment analysis from news or social media.

Finally, more exacting methods of model validation and hyperparameter adjustment might be used. For example, cross-validation, grid search, or Bayesian optimization could be used to make that the model is robust and that its predictions are accurate.

Despite the ARIMA model's promise for stock price forecasting, further research is necessary to improve its accuracy and compare it to other forecasting techniques.

REFERENCES

[1] Orsel, O.E. and Yamada, S.S., 2022. Comparative study of machine learning models for stock price prediction. https://arxiv.org/abs/2202.03156

[2] Agrawal, Manish & Shukla, Piyush & Nair, Rajit & Nayyar, Anand & Masud, Mehedi. (2021). Stock Prediction Based on Technical Indicators Using Deep Learning Model. https://www.researchgate.net/publication/354507095_Stock_Prediction_Based_on_Technical_Indicators_Using_Deep_Learning_Model