**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning for Business |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Daniela Mariano Barreto |
| **Student Number:** | 2023278 |
| **Assessment Due Date:** | 24/04/2024 |
| **Date of Submission:** | 24/04/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of Contents

[**Introduction** 3](#_Toc164872275)

[Motivation 3](#_Toc164872276)

[Problem domain and objectives 3](#_Toc164872277)

[Data description 3](#_Toc164872278)

[Word count 4](#_Toc164872279)

[**Clustering Algorithms** 4](#_Toc164872280)

[DBSCAN 5](#_Toc164872281)

[OPTICS 6](#_Toc164872282)

[Conclusion 8](#_Toc164872283)

[**Time Series Analysis** 8](#_Toc164872284)

[ARIMA 8](#_Toc164872285)

[Forecasting 12](#_Toc164872286)

[Conclusion 13](#_Toc164872287)

[**References** 14](#_Toc164872288)

# **Introduction**

A global trend in pedagogical approach, known as Massive Open Online Courses (MOOCs), serves as an alternative and supplement to traditional models of learning by utilizing online platforms (Sharma, 2013, p.19). It provides high-quality educational content that students worldwide can access more quickly and easily (Thuy et al., 2023, p.1-2).

Since 2012, numerous platforms have emerged for online education, among which Coursera stands out as the most popular due to its variety of courses and strong partnerships with prestigious higher educational institutions like MIT, Harvard, and Stanford (Sharma, 2013, p.19; Zotova et al., 2021, p.167).

## Motivation

During the pandemic period, MOOCs experienced a significant increase (Serravallo, 2020, p.1). This market is estimated to reach USD 22.8 billion this year, with an expectation of USD 119 billion by 2029 (www.mordorintelligence.com, n.d.), making it a lucrative investment opportunity. According to this source, the Coursera platform experienced a 640% increase during the pandemic compared to the previous period.

## Problem domain and objectives

With the modernization of the educational system with MOOCs, platforms such as Coursera developed and sought to serve the most diverse areas of knowledge, signalling promising investment prospects.

In this context, the project aims to compare two clustering algorithms (DBSCAN and OPTICS) to evaluate Coursera's courses based on student ratings and the approximate duration required to complete them. This evaluation could lead to recommendations regarding course duration, which might be useful for course creators to consider when developing courses. Additionally, the analysis of the stock market for Coursera using the ARIMA algorithm will be covered to forecast feature values.

## Data description

The dataset used is from the Kaggle repository (Elvin, 2024, p.1) and contains records of various courses available on the Coursera platform. The time series data is from the stock market of Coursera and was gathered from Yahoo Finance, from 2021-03-31 to 2024-03-26.

Time series data: [Coursera, Inc. (COUR) Stock Price, News, Quote & History - Yahoo Finance](https://finance.yahoo.com/quote/COUR/)

A close-up of a document

Description automatically generated

**Table 1:** Data dictionary based on the repository information.

## Word count

|  |  |
| --- | --- |
| Introduction: | 287 |
| Clustering algorithms: | 346 |
| Time series analysis: | 574 |
| Total: | **1207** |

# **Clustering Algorithms**

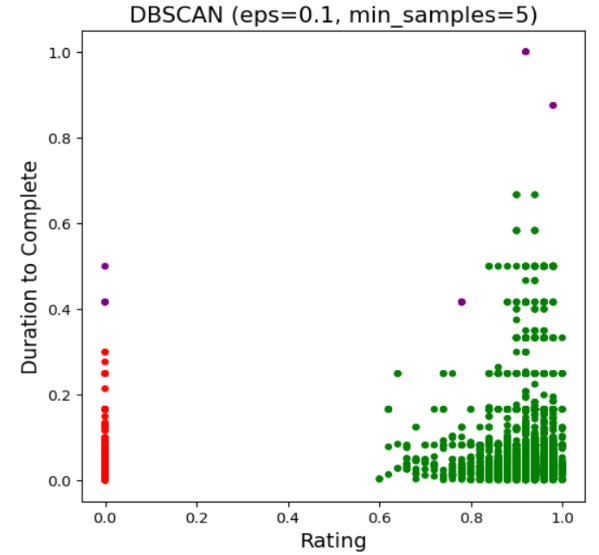
Unsupervised learning algorithms predict without a target label, relying solely on features (Müller and Guido, 2017, p.131). Clustering algorithms group objects based on similarity criteria, useful in business for recommendations and segmentation (Navlani, Fandango and Idris, 2021, p.325).

The Silhouette score and Davies-Bouldin index (DBI) assess clustering quality (Navlani, Fandango, and Idris, 2021, p.350). DBI considers compactness and separation, with lower scores indicating better clusters. Silhouette score measures cluster separation, with higher values indicating better results (Navlani, Fandango, and Idris, 2021, p.351).

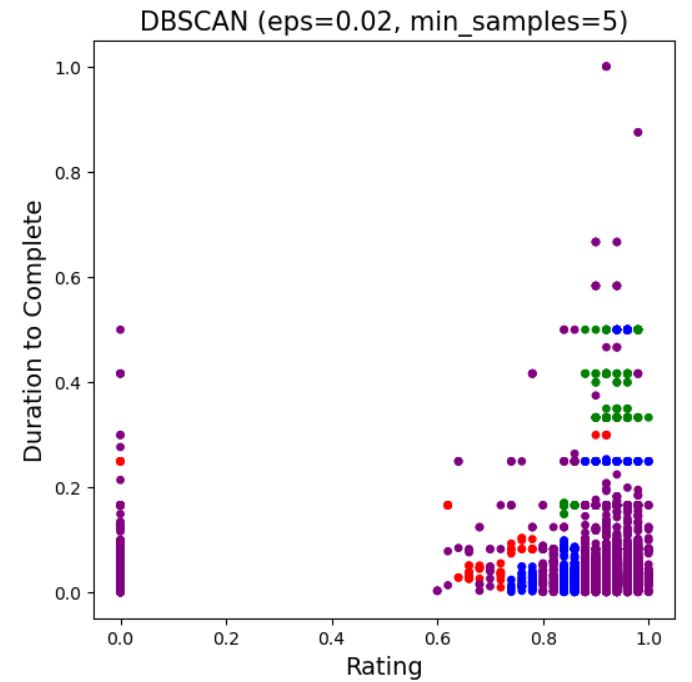
The clustering was performed using the Density Based Spatial Clustering of Application with Noise (DBSCAN) and the Ordering Points To Identify the Clustering Structure (OPTICS) algorithms.

## DBSCAN

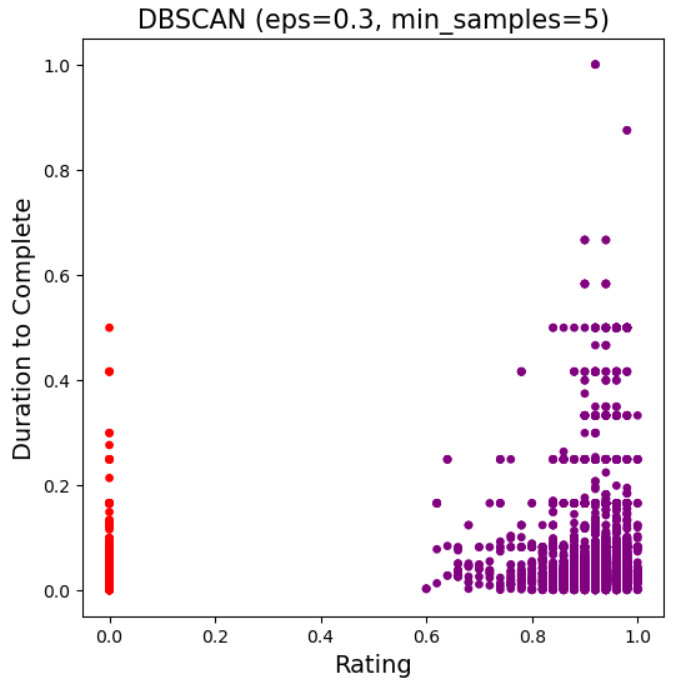
The model is possibly suitable for this dataset because it can handle relatively large datasets without pre-setting the number of clusters, capture complex and dense shapes, and identify points that are not part of any cluster (noise) (Müller and Guido, 2017, p.187), even though determining parameters might be challenging. Figure 2 presents the clustering plot with different epsilon values tested.



b



a



c

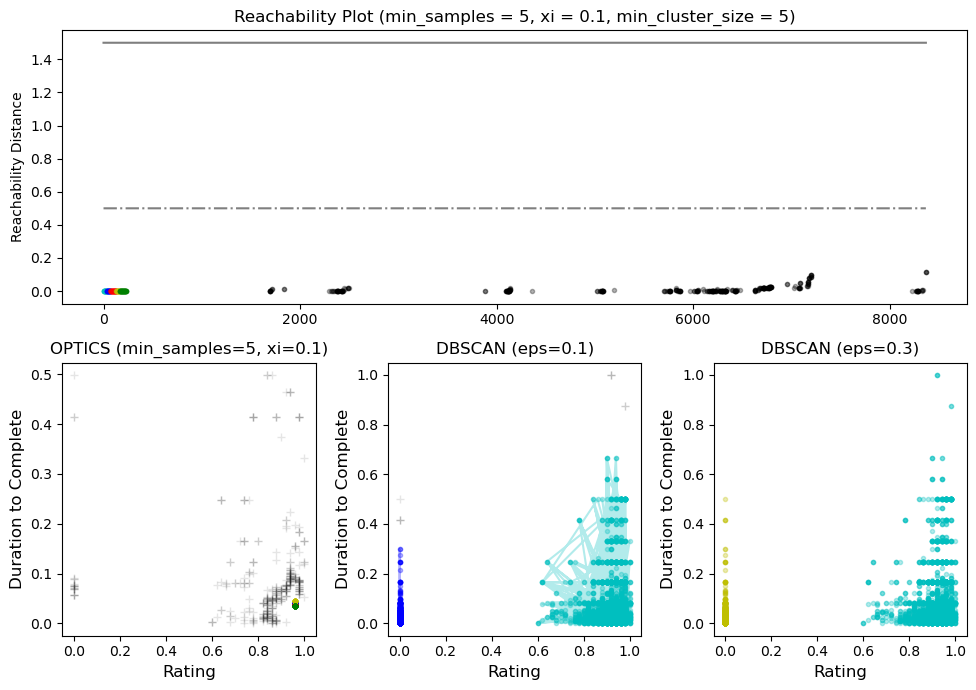
**Figure 2:** DBSCAN clustering results with different epsilon values. a) Silhouette score equal 0.2790 and Davies-Bouldin index (DBI) equal 1.3923. b) Silhouette score equal 0.8810 and BDI equal 0.7302. c) Silhouette score equal 0.9061 and DBI equal 0.0978.

In figure 2b we can see an improvement in separation with 2 clusters and noise (purple dot) along with better metrics and noise separation, suggesting high-quality clustering results in which the clusters are well-defined and distinct, with low intra-cluster variance and high inter-cluster variance.

## OPTICS

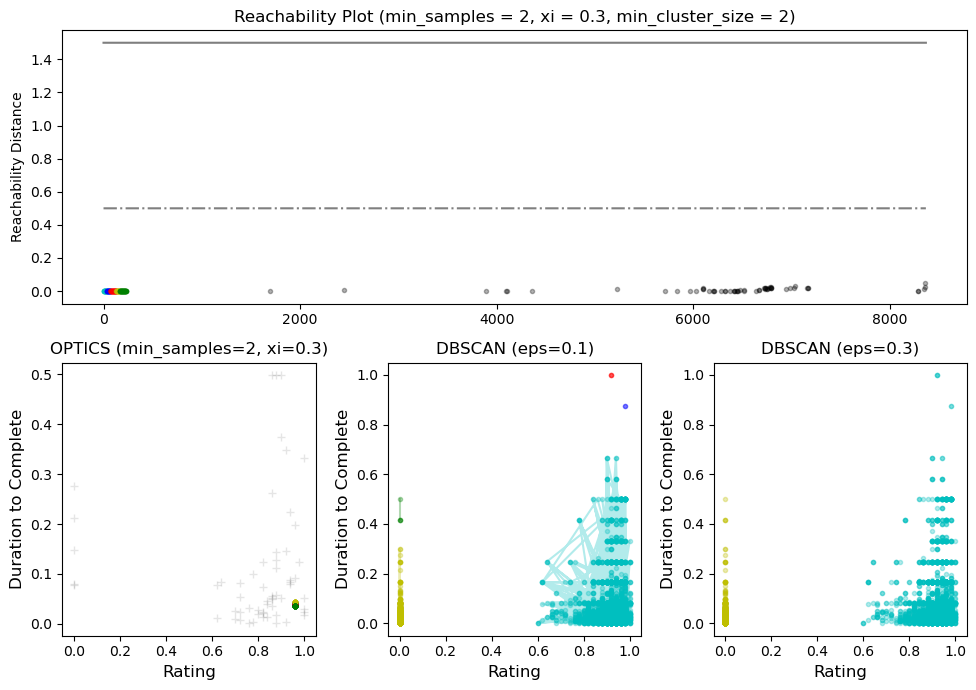
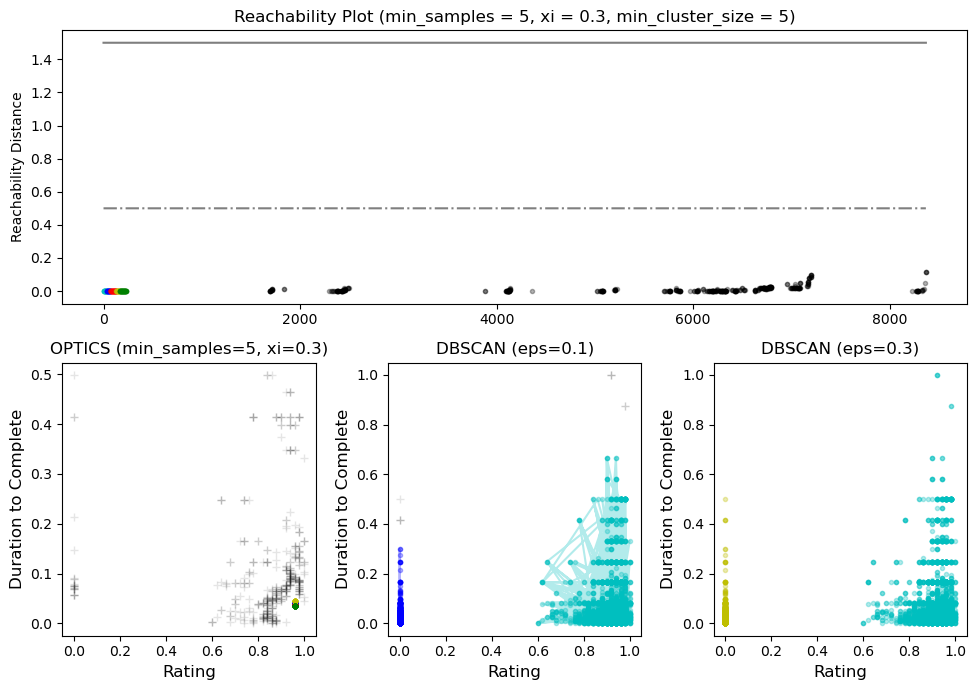
This algorithm tends to be beneficial to this project since it can provide insights into the distribution and correlation within the dataset by creating an ordering of the data based on its density, storing the core distance and a reachability distance for each object, and helping to identify clusters of varying densities in the data (Ankerst et al., 1999, pp. 49-60).

After employing equivalent values from DBSCAN in OPTICS, I found good performances (figures 3-5). Notably, Figure 5 exhibited excellent performance solely with a Silhouette score of 0.98, indicating high-quality clustering with good separation. However, the Davies-Bouldin Index (DBI) yielded unsatisfactory results.



**Figure 3:** OPTICS clustering results. Silhouette score = 0.90 and DBI = 1.39.

**Figure 4:** OPTICS clustering results. Silhouette score = 0.89 and DBI = 1.29.



**Figure 5:** OPTICS clustering results. Silhouette score = 0.98 and DBI = 1.42.

## Conclusion

We observe the reachability distance among the points in DBSCAN (eps=0.1), supporting its results (figure 5). Thus, DBSCAN appears to outperform OPTICS, as indicated by the Silhouette score and DBI. However, OPTICS effectively illustrates connections among the points. Overall, both algorithms performed well, suggesting that courses with shorter completion times are highly rated by students, emphasizing the need for course development to prioritize shorter completion times.

# **Time Series Analysis**

## ARIMA

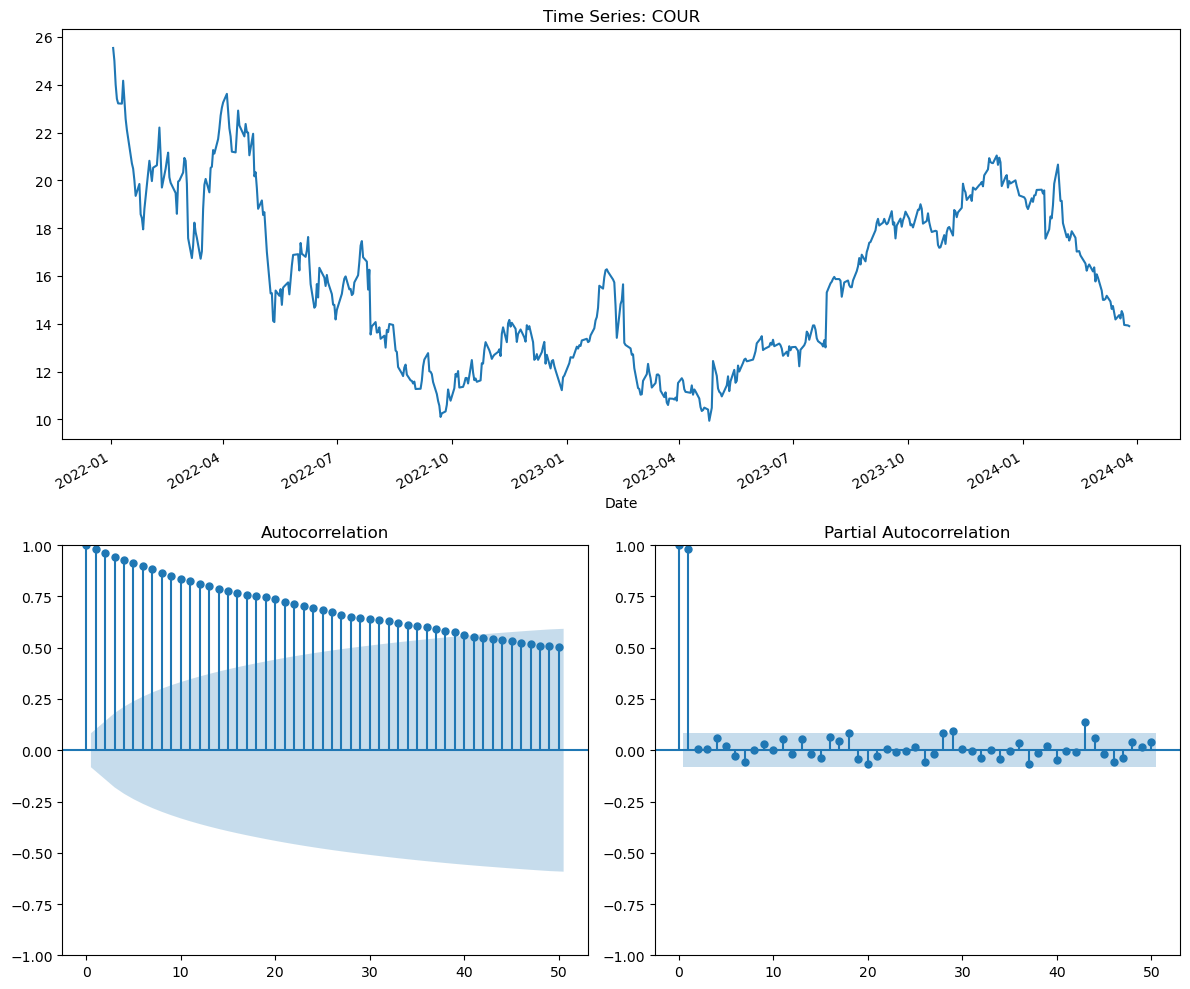
Traditional regression models often fail to capture time series dynamics. However, ARIMA, a widely used model for time series forecasting, incorporates correlations over time with autoregressive (AR), integrating nonstationary components (I) and autoregressive moving average (MA) models to address this problem and enhance predictions (Shumway and Stoffer, 2017, p.75; www.statsmodels.org, n.d.).

* Initial exploration:

The initial exploration of Coursera time series revealed a reduction trend coinciding with the pandemic's end around January 2022. Subsequently, it stabilized with minor fluctuations, necessitating the removal of pandemic period records as anomalies for precise forecasting.

After excluding this period (Figure 6a), the time series showed high fluctuations with two peaks and no apparent seasonality according to the ACF plot, in addition the slow decay in ACF indicated non-stationarity, confirmed by the Dickey-Fuller test (Figure 6b-c). To mitigate bias, first-order differencing was applied, removing correlation and collinearity with past data and rendering the time series stationary (RAVAL, 2024) (Figure 7).

**Figure 6:** a) Time series for Coursera stock market (without pandemic period). b) Autocorrelation Function (ACF) plot. c) Partial Autocorrelation Function (PACF) plot.

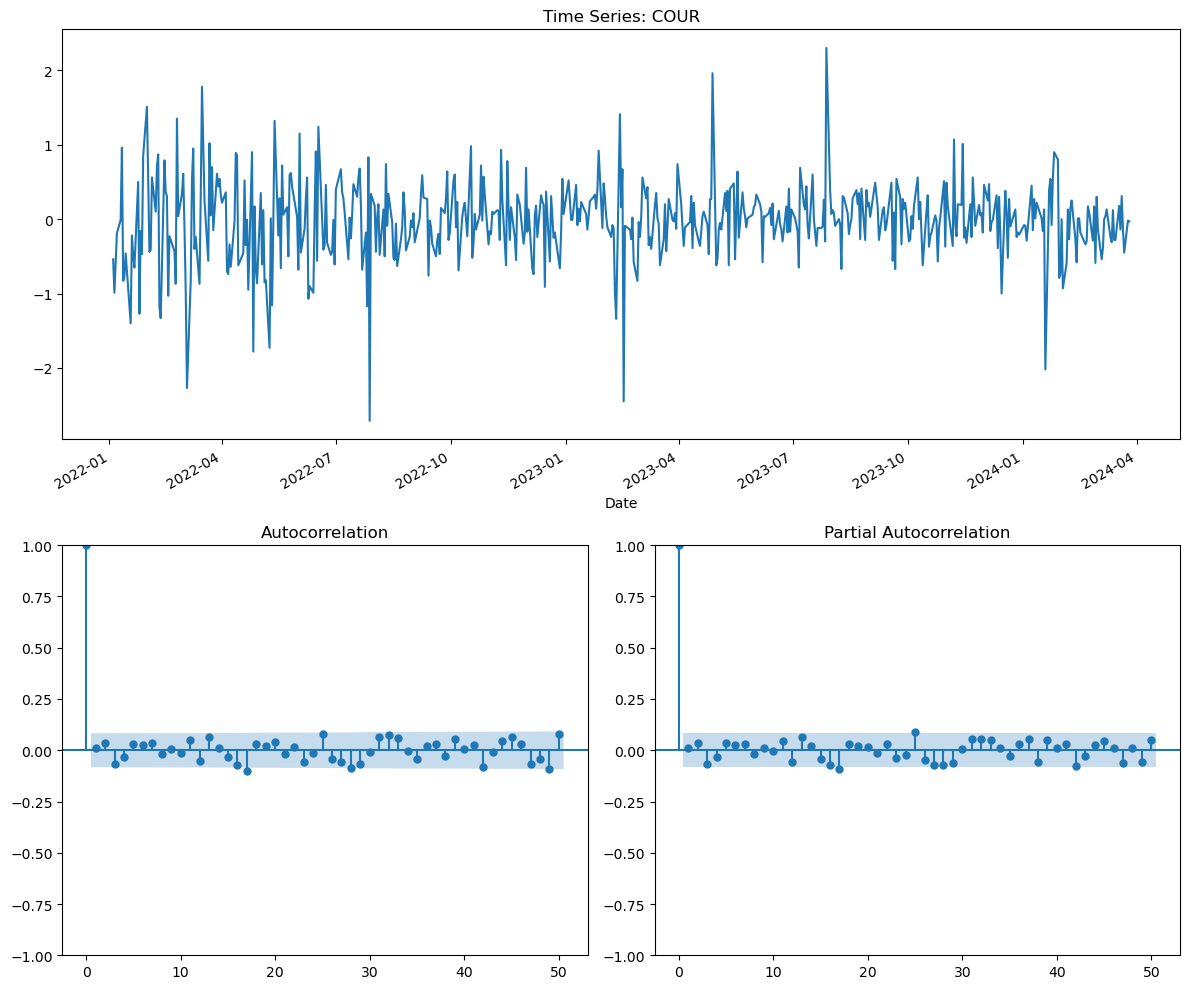


a

b

c

**Figure 7:** a) Stationary time series for Coursera stock market (without pandemic period). b) Autocorrelation Function (ACF) plot. c) Partial Autocorrelation Function (PACF) plot.



a

b

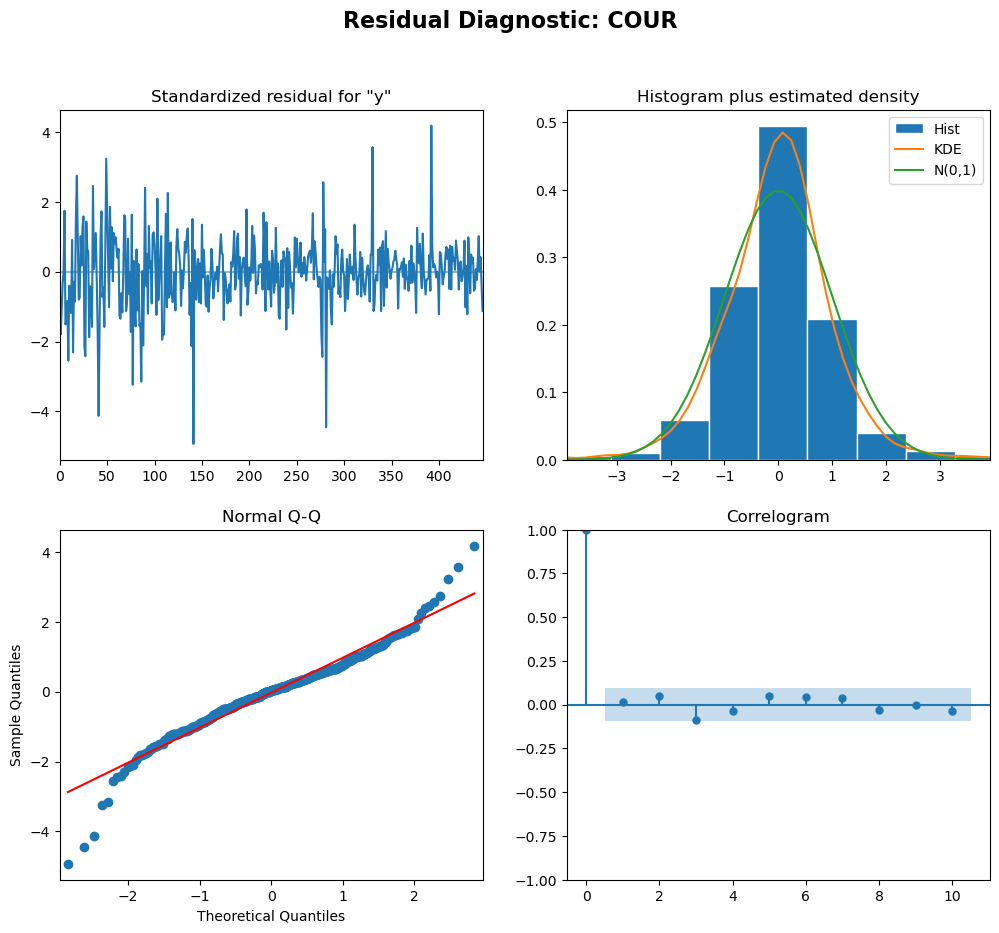
c

* ARIMA results and validation:

Considering the optimum values for ARIMA (0, 1, 0), I used them to train and test the model. The residual diagnostic is below, and the residual errors vary uniformly around a mean of zero (Figure 8b). The density histogram (Figure 8c) suggests a normal distribution, with residual errors appearing close to a mean of zero and exhibiting uniform variance, while the Q-Q plot indicates a slight left skew in the distribution, suggesting reasonable adherence. In Figure 8d, the correlogram shows that the residual errors are not autocorrelated, indicating white noise.

Root Mean Square Error (RMSE) and R2 score were used to verify how well the model in predict future values based on the patterns learned from the training set (Figure 9). The metrics resulted in an excellent model performance in which it can explain 95.5% of variance in the data, which is quite high. RMSE is close to zero (0.41), indicating a good predictability of the model since the variation between actual and predicted values is close.

**Figure 8:** a) Standardized residual. b) Histogram estimated density. c) Normal Q-Q. d) Correlogram.

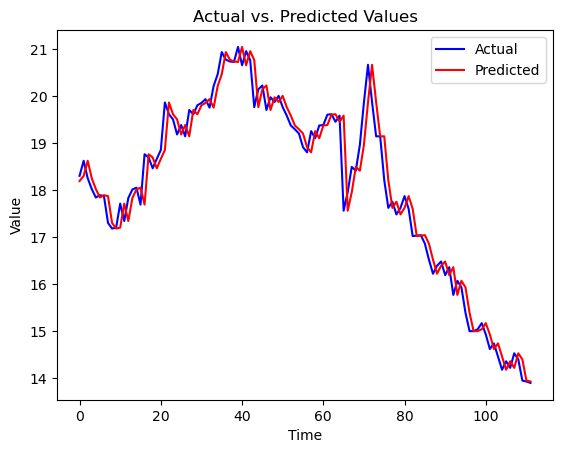


a

b

c

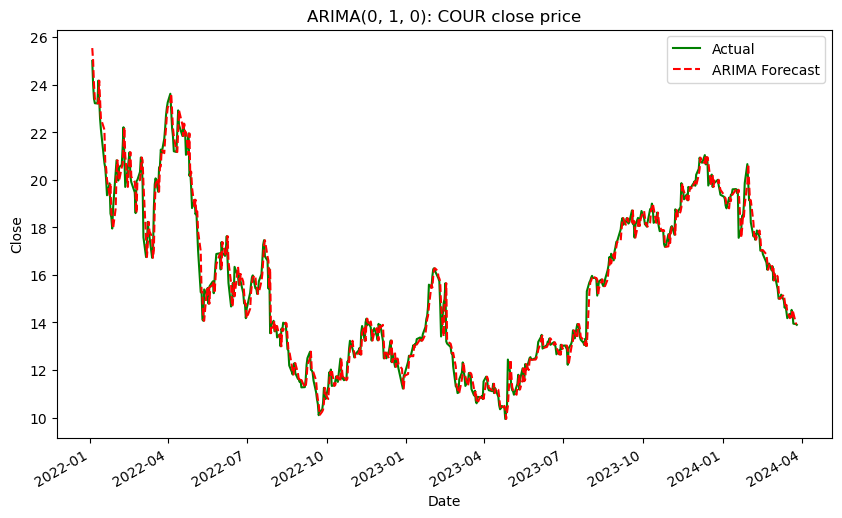
d



**Figure 9:** Chart of actual and predicted values for ARIMA (0,1,0).

## Forecasting

After using the entire 'Close' feature in the ARIMA (0, 1, 0), inclusive incorporating the residuals to obtain complete predicted values (figure 10), I observed an increase in the AIC. However, the sigma2 coefficient was reduced, indicating a decrease in the noise term when using the entire dataset. The metrics also resulted in good model performance, with the model explaining 88% of the variance, which is considered high. Nonetheless, the RMSE (1.20) was unsatisfactory, as the stock price varied by $1.20, which can be considered a significant fluctuation.

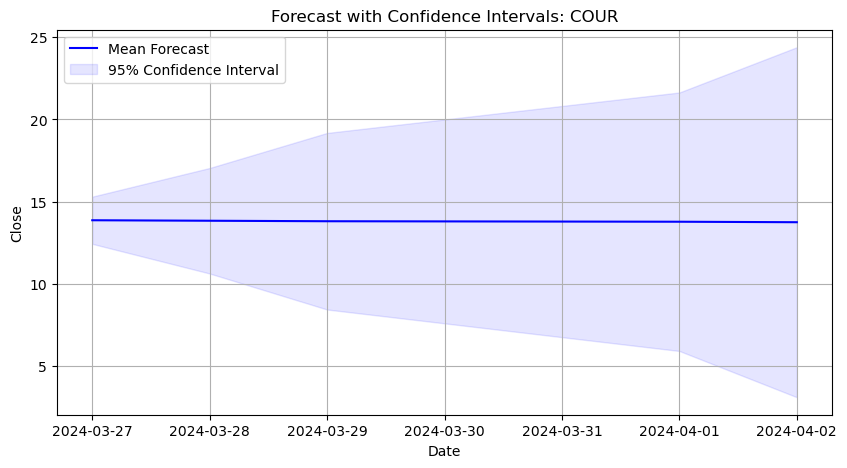


**Figure 10:** Chart of actual (‘Close’ feature) and predicted values for ARIMA (0,1,0).

* Predicting future values:

Using all available data in 'Close' feature in ARIMA (0,1,0) all the predicted values were identical. But, employing second-order differencing (0,2,0) improved alignment with the original data, possibly eliminating remaining trends and seasonality, resulting in predictions more compatible with the original data.

Predictions for five values in the feature are shown with a 95% confidence level (figure 11) with metrics resulting in good model performance, with the model explaining 88% of the variance. Moreover, RMSE showed improvement compared to previous forecasting, resulting in a more satisfactory outcome as the stock price varied by only 0.40 cents, making it considerably more suitable for this kind of business.



**Figure 11:** Forecasting the 'Close' price for five days with ARIMA (0,2,0) with 95% confidence interval.

## Conclusion

In general, the ARIMA (0,1,0) model can capture a significant portion of the variability in the data, 88%. However, incorporating residuals led to a notable increase in the RMSE (1.20). Conversely, the ARIMA (0,2,0) potentially reduced trends and seasonality within the data, resulting in predictions more aligned with the original dataset, thus decreasing the RMSE (0.40) while maintaining a high variance explainability.

However, despite the model's good performance metrics, it has limitations in forecasting future values. The increasing variability of the data, as indicated by the standard error over time, suggests that the underlying patterns in the data are becoming more complex as time progresses. Therefore, further investigation is justified to address this issue and uphold the model's forecasting accuracy over time.

# **References**

Ankerst, M., Breunig, M.M., Kriegel, H.-P. and Sander, J. (1999). *OPTICS: ordering points to identify the clustering structure*. ACM SIGMOD Record, 28(2), pp.49–60. doi:https://doi.org/10.1145/304181.304187.

Elvin, R. (2024). Coursera Dataset. [online] Available at: https://www.kaggle.com/datasets/elvinrustam/coursera-dataset. [Accessed 12 Feb. 2024].

Müller, A. C. and Guido, S. (2017). Introduction to machine learning with Python: a guide for data scientists. 1st ed. United States of America. O’reilly Media.

Navlani, A., Fandango, A. and Idris, I. (2021). *Python Data Analysis*. Packt Publishing Ltd.

Raval, P. (2024). How to Build ARIMA Model in Python for time series forecasting? [online] ProjectPro. Available at: https://www.projectpro.io/article/how-to-build-arima-model-in-python/544. [Accessed 16 Apr. 2024]

Serravallo, J. (2020). Connecting with students online : strategies for remote teaching & learning. Portsmouth, New Hampshire: Heinemann.

Sharma, R.C. (2013). How to Offer a Massive Open Online Course (page 19-20). *EduComm Asia*. [online] Available at: https://www.academia.edu/5505040/How\_to\_Offer\_a\_Massive\_Open\_Online\_Course\_page\_19\_20\_. [Accessed 16 Mar. 2024].

Shumway, R.H. and Stoffer, D.S. (2017). ARIMA Models. In: *Time Series Analysis and Its Applications*. Springer, Cham, pp.75–163.

Thuy, T., Thanh Tu Tran, Gia Khuong An and Nguyen, P. (2023). Students’ Perception Towards Learning Massive Open Online Courses on Coursera Platform: Benefits and Barriers. *International Journal of Emerging Technologies in Learning (ijet)*, 18(14), pp.4–23. doi:https://doi.org/10.3991/ijet.v18i14.39903.

www.mordorintelligence.com. (n.d.). Massive Open Online Course (MOOC) Market | Growth, Trends, and Forecasts (2020 - 2025). [online] Available at: <https://www.mordorintelligence.com/industry-reports/massive-open-online-course-mooc-market>. [Accessed 17 Mar. 2024].

Zotova, M., Likhouzova, T., Shegai, L. and Korobeynikova, E. (2021). *The Use of MOOCS in Online Engineering Education*. International Journal of Engineering Pedagogy (iJEP), [online] 11(3), p.157. doi:https://doi.org/10.3991/ijep.v11i3.20411. [Accessed 16 Mar. 2024].