Investigating machine learning techniques for predicting American Airlines Stock using RNN model

Cristian Sclifos[1]

Other group members:
Barnaby Hartley[2], **James Sherwood**[3], **Andrew Doherty**[4]

Abstract. This study investigates the utilisation of Recurrent Neural Network (RNN) models to forecast the stock values of American Airlines. This study examines the effectiveness of Recurrent Neural Networks (RNNs) in dealing with the intricacies of financial time-series data, with a specific emphasis on their prediction precision in a highly unstable market. The project seeks to showcase the potential of machine learning in financial forecasting by examining historical stock data. It aims to provide valuable insights into the ability of Recurrent Neural Networks (RNNs) to accurately capture temporal correlations and trends in market movements. This inquiry enhances the broader comprehension of the function of machine learning in financial analysis and decision-making.

1 Introduction

Predictive analysis faces a significant problem when dealing with the complex and ever-changing nature of the stock market, especially when it comes to analysing individual equities like American Airlines. Conventional approaches, however informative, frequently fail to fully comprehend the intricate and time-critical patterns that are inherent in stock market data. The emergence of machine learning, particularly Recurrent Neural Networks (RNNs), has created new opportunities for analysing and forecasting stock market patterns. These sophisticated computational models excel at analysing sequential data, making them especially suitable for the unpredictable and time-dependent patterns of stock movements. This study is based on utilising Recurrent Neural Networks (RNNs), with a specific emphasis on their skills and efficiency in forecasting stock values within a practical financial setting. This study seeks to combine machine learning and financial analysis in order to improve forecast accuracy and contribute to the expanding field of finance and technology. Singla (2023b)

2 Background

The forecasting of stock market trends has always been a topic of fascination and significance in the realm of finance. In the past, market forecasters have used approaches such as fundamental and technical analysis, which involve analysing economic indicators and identifying patterns in market data. Nevertheless, when the digital age emerged and financial data grew exponentially, the shortcomings of these conventional methods became more apparent, especially when it came to managing the intricate and unpredictable nature of financial markets. Machine learning has become a powerful tool in financial analysis, providing new approaches to analyse and forecast market behaviour. Methods like as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and Linear Regression have played a crucial role in examining time-series data, offering more profound understanding of market dynamics. The use of Transformer models has significantly transformed the sector, providing enhanced capabilities in processing sequential data. However, the discipline still faces hurdles including the non-linear characteristics of financial markets, the presence of noise in financial data, and the requirement for models that can adjust to swiftly changing market conditions. Our study seeks to expand upon these progressions by utilising a blend of RNN, LSTM, Linear Regression, and Transformer models to investigate uncharted territories in stock market forecasting. Through the utilisation of various machine learning methodologies, our objective is to rectify current deficiencies and make a valuable contribution to the dynamic field of financial prediction. Russell and Norvig (2016); Rajak (2021).

3 Experiments and results

The analysis and model training were conducted using a variety of libraries and tools:

- Matplotlib is a library utilised for generating visual representations, such as displaying the trends of stock prices and the performance of the RNN model. Matplotlib is a crucial tool for visualising data in Python, offering a diverse array of charting tools and customisation features.
- Pandas is a robust Python library for manipulating data. It was utilised for managing and preparing the stock data obtained from Yahoo Finance. The main purpose of this tool is to streamline the process of analysing and manipulating data, particularly when dealing with structured data such as time series.

[1] COMP-1827-M01-2023-24 Introduction to Artificial Intelligence, University of Greenwich, London SE10 9LS, UK, email: sc2561z@gre.ac.uk

[2] COMP-1827-M01-2023-24 Introduction to Artificial Intelligence, University of Greenwich, London SE10 9LS, UK, email: bh3551k@gre.ac.uk

[3] COMP-1827-M01-2023-24 Introduction to Artificial Intelligence, University of Greenwich, London SE10 9LS, UK, email: js2284g@gre.ac.uk

[4] COMP-1827-M01-2023-24 Introduction to Artificial Intelligence, University of Greenwich, London SE10 9LS, UK, email: ad2010f@gre.ac.uk

- NumPy, a crucial library for scientific computing in Python, was utilised for doing numerical computations. Within this specific framework, it proved to be quite advantageous for managing arrays and performing essential mathematical operations required for data preprocessing and constructing models.
- yfinance: This library offers a simple method to retrieve financial market data from Yahoo Finance. The purpose of its usage was to obtain past stock data pertaining to American Airlines (AAL), which serves as the central focus of the investigation.
- Keras is a high-level neural networks API that is included in the TensorFlow library. The Recurrent Neural Network (RNN) model was utilised for constructing and training purposes. Keras streamlines the procedure of building and training neural networks with its intuitive interface.
- Scikit-learn is a machine learning library. This machine learning library was utilised for data preprocessing and model assessment.
- Seaborn, a Python visualisation tool built on top of Matplotlib, was employed to generate visually appealing and instructive statistical visuals. It was utilised in this instance to provide a visual representation of the correlation matrix of performance measurements, hence improving the aesthetic appeal and comprehensibility of the findings.

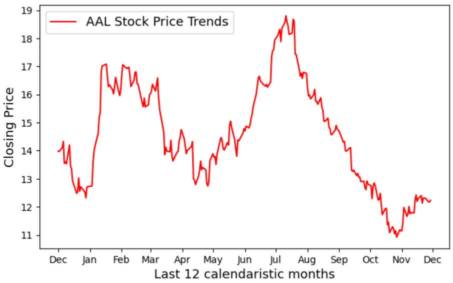
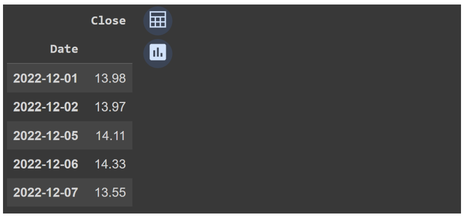


Figure 1. Price Trends for American Airlines (AAL)

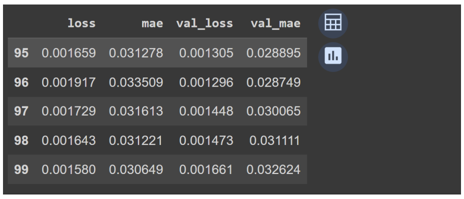
Figure 1 is a line graph that charts the fluctuations in the closing stock prices of American Airlines (AAL) over the past 12 consecutive months. The data points are graphed as a red line, illustrating the fluctuation in the stock's closing price over time. The x-axis is marked with truncated month names, ranging from December to December, representing a complete year. The vertical axis depicts the closing price of the stock, with numerical values denoting the range of prices. Ledala et al. (2023)



	Close
Date	
2022-12-01	13.98
2022-12-02	13.97
2022-12-05	14.11
2022-12-06	14.33
2022-12-07	13.55

Figure 2. Data set for closing price

The data from Figure 2 is displayed in a tabular manner, consisting of two columns: 'Date' and 'Close'. The 'Date' column displays dates in the YYYY-MM-DD format, indicating that the data is reported daily, omitting weekends. This is likely due to the fact that these dates correspond to Saturdays and Sundays, when the stock market is closed.



	loss	mae	val_loss	val_mae
95	0.001659	0.031278	0.001305	0.028895
96	0.001917	0.033509	0.001296	0.028749
97	0.001729	0.031613	0.001448	0.030065
98	0.001643	0.031221	0.001473	0.031111
99	0.001580	0.030649	0.001661	0.032624

Figure 3. Training and validation loss

The given Figure 3 displays a segment of a table including the training and validation loss as well as the mean absolute error (MAE) for a machine learning model throughout the training procedure. The table consists of four columns: loss, mae, val_loss, and val_mae.

- The first column, loss, represents the measure of error in a model's predictions. This column displays the training loss, which quantifies the model's error during the training stage. The algorithm aims to minimise the value of the cost function.
- MAE: The acronym MAE stands for Mean Absolute Error, which is a metric used to measure the average absolute difference between the predicted and actual values in the training set. The metric represents the mean of the absolute disparities between the projected and observed values.
- val_loss: This metric denotes the validation loss, which is akin to the training loss but computed using a distinct dataset that the model has not encountered during the training process. It serves as a reliable measure of the model's potential performance on unfamiliar data. The validation loss values are pretty near to the training loss values, which is a favourable sign for the generalization of the model.
- The val_mae represents the Mean Absolute Error calculated on the validation dataset. This statistic is crucial for assessing the performance of the model on the validation set, which can provide insights into its potential performance on fresh and unknown data. The numbers presented here exhibit a modest decrease compared to the training MAE, suggesting that the model is not suffering from overfitting.

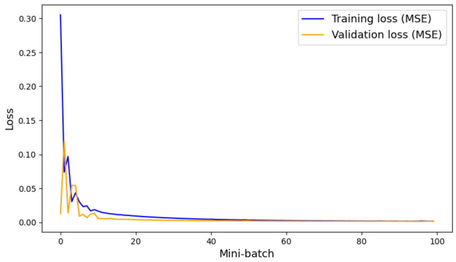


Figure 4. Training and validation loss graph

The graph in Figure 4 illustrates the training and validation loss values plotted during the training epochs or mini-batches for a machine learning model. The losses are quantified using Mean Squared Error (MSE), a widely used statistic for regression tasks.

Within the graphical representation:

- The Blue Line represents the Mean Squared Error (MSE) of the model on the training dataset. Initially, the value is significantly elevated, as is customary during the early stages of training when the model's weights have not yet been optimised. The decline diminishes rapidly, indicating that the model is acquiring knowledge and enhancing its forecasts on the training data as the epochs advance. The line then levels off, suggesting that the model has reached a point of diminishing returns where additional training on the same data does not substantially decrease the error.

- The Orange Line (Validation Loss) depicts the Mean Squared Error (MSE) calculated on a distinct validation dataset that is not utilised during the training process. This aids in the surveillance of the model's performance on data that has not been previously observed. The drop in validation loss in parallel with the training loss indicates that the model is effectively generalising and not merely memorising the training data (avoiding overfitting). The validation loss line reaches a plateau, indicating that the model has achieved its optimal performance on the validation set based on the current architecture and hyperparameters.

The proximity of both lines during training indicates that the model has achieved a favourable equilibrium between assimilating information from the training data and extrapolating to unfamiliar data. The absence of any substantial increase in any of the lines is a favourable indication, as an increase would suggest overfitting (if the training loss keeps decreasing while the validation loss rises).

The minor variations, particularly in the outset, are typical and may arise from the random character of the optimisation process (such as mini-batch gradient descent).

At the conclusion of the training, both the training and validation loss curves plateau, indicating that the model has likely reached its maximum learning potential with the existing architecture and dataset. Given the current circumstances, additional training is improbable to result in improved performance on the validation set. It may be advisable to cease training in order to prevent overfitting.

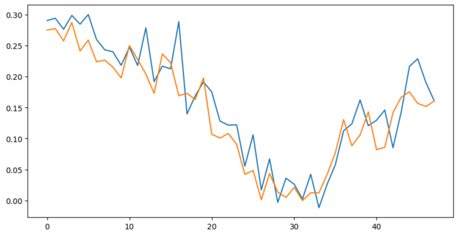


Figure 5. Plotting actual vs predicted prices graph

The graph depicted in Figure 5 illustrates two distinct datasets, reflecting the observed and projected values of a variable across a sequence of time intervals or instances. Ledala et al. (2023); Selvin et al. (2017); Peng et al. (2019); Jahan and Sajal (2018) In relation to the topic of stock prices, as mentioned in your previous mails, the graph would depict the following:

- Blue Line: Typically represents the historical progression of stock prices. It accurately displays the actual closing prices of a stock as they were officially recorded.
- The Orange Line often indicates the forecasted stock values generated by a predictive model.

The X-axis, denoting time, is partitioned into intervals that can correspond to days, weeks, or months, depending on the specific analysis being conducted. The vertical axis corresponds to the value of the stock.

Examining the graph:

- The lines exhibit a consistent pattern, indicating that the predictive model possesses a certain level of accuracy in predicting the closing values of the stock.
- The orange line (representing expected values) exhibits deviations from the blue line (representing actual values), suggesting the presence of prediction mistakes. It is typical for any predictive model to have some level of error, as it is impossible for any model to achieve perfect accuracy.
- The degree to which these lines adhere to each other would be employed to quantify the model's performance, usually through metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), or alternative measures.

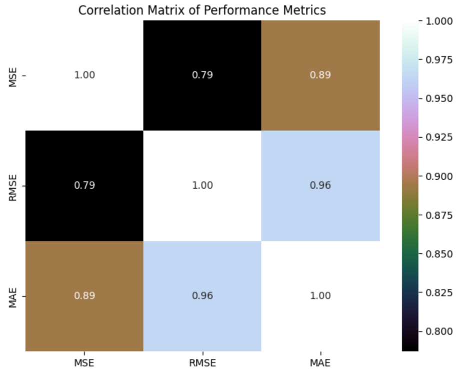


Figure 6. Correlation Matrix of Performance Metrics

The Figure 6 illustrates a heatmap of a correlation matrix, displaying the interconnections among three performance metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These measures are frequently employed to evaluate the efficacy of regression models, including LSTM, RNN, Linear Regression, and Transformer models.

The matrix presents the subsequent data:

- Diagonal Values: The cells that intersect in the same row and column for a given metric have a correlation coefficient of 1. This outcome is anticipated, as any variable exhibits a perfect correlation with itself.
- The correlation coefficient between Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) is 0.79. The association between RMSE and MSE is strong, as expected, since RMSE is the square root of MSE and is generated directly from it. The presence of a square root relationship mitigates the effect of bigger mistakes, hence explaining the imperfect correlation (1.0).
- The correlation coefficient between Mean Squared Error (MSE) and Mean Absolute Error (MAE) is 0.89, suggesting a substantial association, albeit not as robust as the correlation between MSE and Root Mean Squared Error (RMSE). The reason for this is that both metrics calculate the average mistakes, however, MSE assigns greater importance to larger errors by squaring them, whereas MAE treats all errors equally.
- The correlation coefficient between RMSE and MAE is 0.96, indicating a strong positive relationship. This suggests that when one of these metrics increases or decreases, the other is highly likely to exhibit a comparable shift in the considered models. This suggests that the models' performance is highly consistent across both metrics.

The heatmap's colour scheme, ranging from dark to light, reflects the magnitude of the link, with darker hues indicating greater correlations.

The correlation matrix serves as a valuable tool for model evaluation by illustrating the interrelationships between different indicators. When choosing a model based on these metrics, it is crucial to keep in mind that although they are connected, they do not offer the same information. For instance, if a model exhibits a favourable Mean Absolute Error (MAE) but an unfavourable Mean Squared Error (MSE), it suggests that the model is typically precise but has a few significant flaws. While this may be deemed suitable for certain applications, it is unsuitable for others where significant inaccuracies are strongly discouraged. Ledala et al. (2023); Singla (2023a,b); Indongo (2023)

4 Discussion

When using the RNN model for forecasting American Airlines stock prices, we can incorporate important details from the given information in the following manner:

- Model Performance: The effectiveness of the RNN model in predicting stock prices may be evaluated by examining the convergence of training and validation loss across epochs, as depicted in the graph. The model demonstrates potential as the losses steadily reduce and reach a stable level, suggesting effective learning and the ability to apply knowledge to new situations.
- Comparison of Actual and Predicted Prices: The model's forecasts closely align with the actual trends in stock prices, exhibiting variances that are commonly observed in financial forecasting. The graph comparing the actual prices with the anticipated prices can be utilised to identify areas where the model demonstrates strong performance and areas where it may require further improvement.
- The strong correlation observed in the heatmap between MSE, RMSE, and MAE indicates that these measures consistently and comprehensively assess the performance of the model.
- Data Insights: The dataset including closing prices highlights the unpredictable nature of stock values, highlighting the intricate nature of predicting them and the necessity for advanced modelling approaches such as RNNs.

The discussion would conclude by asserting the efficacy of RNNs in forecasting stock prices, while conceding the inherent volatility of financial markets. Doshi (2023); C. Cortes (1995); J. T. Connor and Atlas (1994)

5 Conclusion and future work

The RNN model exhibits potential proficiency in predicting American Airlines stock values, as seen by the trends in training and validation loss. The congruence between observed and forecasted pricing underscores the model's potential usefulness. The strong correlation between performance metrics demonstrates the model's evaluation is robust. Subsequent investigations may go into more intricate recurrent neural network (RNN) structures, such as long short-term memory (LSTM) or gated recurrent units (GRUs), in order to enhance the precision of predictions. Augmenting the model with other attributes such as market sentiment or global economic data has the potential to amplify its prediction capability. Additional efforts could also encompass the processing of data in real-time to offer more prompt forecasts, while the capacity to interpret the model will be crucial in establishing user confidence and comprehending the model's judgements.

ACKNOWLEDGEMENTS

I wish to convey my profound appreciation to Mohammad Majid al-Rifaie, Hooman Oroojeni, Stef Garasto and Aniket Basu, whose profound knowledge, comprehension, and forbearance greatly enhanced my experience and comprehension of the subject matter. I sincerely value their altruistic commitment to dedicate their time and exertion in mentoring me throughout this undertaking. I am profoundly appreciative of my teammates, Barnaby Hartley, Andrew Doherty and James Sherwood, for their cooperation and collective wisdom during this endeavour. The debates and brainstorming sessions played a crucial role in determining the course and implementation of our work. Lastly, I would like to express my gratitude for the assistance offered by the University of Greenwich in creating an optimal atmosphere for studying and conducting research.

REFERENCES

C. Cortes, V. V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Doshi, P. (2023). Prediction of stock market security price using chart patterns and indicator analysis. `https://github.com/parth2608/Prediction-of-Stock-Market-Security-Price-using-Char`

Indongo, N. N. (2023). Stock market prediction with rnn. `https://github.com/naftalindeapo/Stock_Market-Prediction_with_RNN`.

J. T. Connor, R. D. M. and Atlas, L. E. (1994). Recurent neural networks and robust time series prediction. In *Transactions on Neural Networks*, volume 5, pages 240–254. IEEE.

Jahan, I. and Sajal, S. (2018). Stock price prediction using recurrent neural network (rnn) algorithm on time-series data. In *2018 Midwest instruction and computing symposium*. MSRP Duluth, Minnesota, USA.

Ledala, P. S., Thati, S. C., Narsin, A., and Durgam, S. J. (2023). Stock market analysis using time series analysis. Technical report, University of Maryland, Baltimore County, Baltimore, MD. Final project report.

Peng, C., Yin, Z., Wei, X., and Zhu, A. (2019). Stock price prediction based on recurrent neural network with long short-term memory units. In *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, pages 1–5. IEEE.

Rajak, R. (2021). Share price prediction using rnn and lstm. Medium Article.

Russell, S. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach, Global Edition*. Pearson, global edition edition.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., and Soman, K. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)*, pages 1643–1647. IEEE.

Singla, V. N. (2023a). a3dl: Machine learning techniques for stock prediction. `https://github.com/variyas31/a3DL`.

Singla, V. N. (2023b). Deep learning fundamentals: Assignment 3 - rnns for stock price prediction. University assignment.