INDENG 290: Problem Set 3

Loïc Jannin

Due: October 27, 2023 at 3 pm PST

1 Data Retrieval and Initial Observations

- Let's start by obtaining daily stock data for AMZN (Amazon) and NCLH (Norwegian Cruise Lines) from Yahoo Finance for the years 2016, 2017, 2018, 2019, and 2020 using Python.

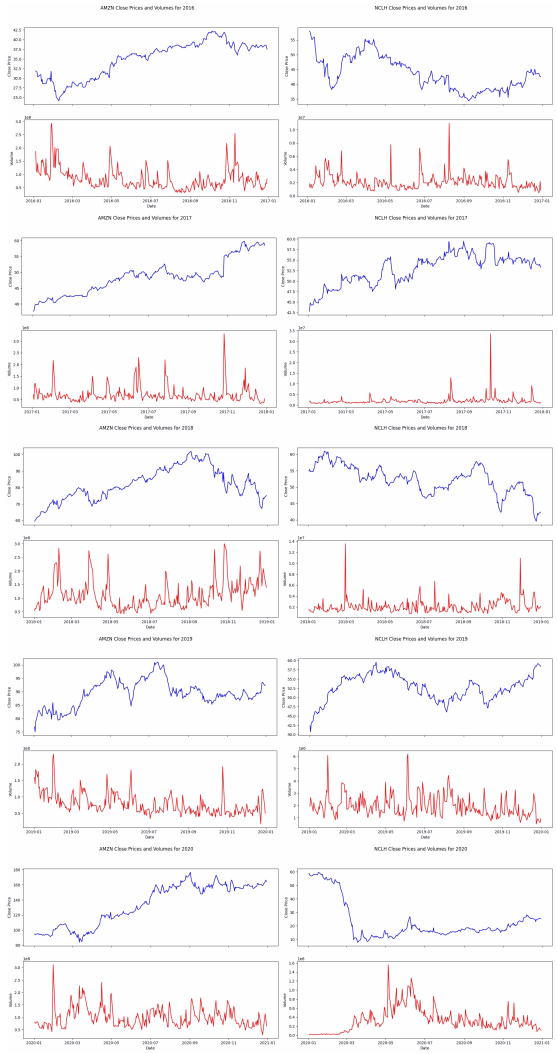

Figure 1: Stock Data for AMZN and NCLH

- Upon analyzing the data patterns, it becomes evident that the two stocks, AMZN (Amazon) and NCLH (Norwegian Cruise Lines), exhibited markedly different behaviors over the specified time frame.

  - AMZN displayed consistent growth throughout, maintaining an upward

trajectory in its stock price. In contrast, NCLH faced challenges, even before the COVID crisis, with its stock struggling to gain momentum during the same period.

- Despite the differing overall trends, it's worth noting that both stocks showed some similarities in terms of market volatility. For instance, there are noticeable resemblances in the patterns during October, November, and December of 2019, suggesting that specific market conditions or events impacted both stocks during those months.

- Unsurprisingly, when the COVID-19 crisis hit, NCLH, a cruise company, faced a tough time. After all, cruise companies rely on people coming together to enjoy vacations, which was the last thing on everyone's mind during the pandemic. In contrast, Amazon, the go-to online retail giant, thrived. It made perfect sense; as people refrained from going out to shop, they turned to Amazon for their needs. It was the convenient and safe alternative to in-person shopping, and the numbers showed it.

- To predict stock volumes and close prices effectively, a model should have the following characteristics:

    1. Be capable of reproducing known stylized facts to ensure the data appears realistic.
    2. Generate a wide range of scenarios without bias.
    3. Avoid overfitting the training data.
    4. Account for lag effects, considering that past close prices can influence the current price.
    5. Select relevant features or factors that influence stock prices and volumes.
    6. Eliminate outlier values.

- Analysis of 2020: In 2020, the COVID-19 crisis brought about a complete shift in the market, resulting in several distributional shifts. Amazon was relatively unaffected by the crisis, so we will focus more on the NCLH stock.

    1. First, a common and unusual bear market occurred. After a spike in volatility (evident in the significant drop in early 2020), stock markets saw a downward trend.
    2. During a crisis like this one, there was a shift in correlations. Notably, when previously almost all stocks were rising, some sectors were more impacted than others, leading to a correlation shift. Several other correlation shifts may have occurred but are not visible in these charts.
    3. For the NCLH stock, the stock distribution during the year 2020, after the initial drop, is fundamentally different from the year before, indicating a more stable distribution. The distribution shift in volumes is remarkable for the NCLH stock as this plot shows:

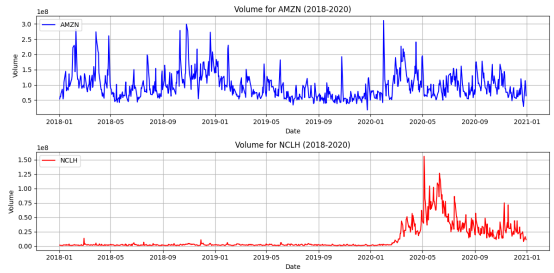

Figure 2: Volumes for NCLH and AMZN

- Model Enhancement: To enhance their models, one could consider developing models specifically designed to account for increased volatility. Volatility modeling techniques like GARCH (Generalized Autoregressive Conditional Heteroskedasticity) can be valuable in capturing changing volatility patterns. Another option is to use reinforcement learning (RL) algorithms, which passively interact with the environment to predict various scenarios, including crisis situations. Bayesian methods can also be useful, as they estimate the distribution of unobserved data given observed data.

2 Average and Median Forecasting method

Using the N-day sliding window method, we applied both the N-day average and N-day median methods to predict daily stock volumes.

2.1 Forecast for 10th, 31st and 61st day of the year.

N	Sliding Average	Sliding Median	Real Volume
10	1.3789e+08	1.2875e+08	1.2734e+08
30	9.5322e+07	8.7970e+07	9.9498e+07
60	9.1306e+07	8.3294e+07	6.2288e+07

Table 1: Stock Forecast Results for Amazon in 2019

N	Sliding Average	Sliding Median	Real Volume
10	6.9095e+07	6.9489e+07	5.7932e+07
30	8.2853e+07	7.0365e+07	5.2124e+07
60	1.1062e+08	9.6963e+07	1.2252e+08

Table 2: Stock Forecast Results for Amazon in 2020

2.2 Monthly mean square error.

Month	**MSE for Average**	**MSE for Median**
January	1.1299e+15	1.2351e+15
February	1.1526e+15	9.2122e+14
March	7.6900e+14	9.8292e+14
April	6.0308e+14	7.5847e+14
May	2.9683e+14	3.6475e+14
June	8.2447e+14	8.3978e+14
July	3.1315e+14	3.7018e+14
August	3.9135e+14	3.8211e+14
September	2.6635e+14	3.2195e+14
October	8.6798e+14	1.0035e+15
November	2.1504e+14	1.4599e+14
December	5.3531e+14	5.9899e+14
Average	6.2726e+14	7.0524e+14

Table 3: Mean Square Error (MSE) for Amazon in 2019 (N = 10)

In January, the error is observed to be higher compared to other months. This phenomenon can be attributed to the method used for calculating the projected price. The inclusion of December data introduces variations in the

projections due to the holidays during that period, which causes the algorithm to consider data from the previous year. As a result, the error for January tends to be larger.

2.3 Yearly mean square error.

Stock comparison: As expected, the accuracy of the Mean Square Error (MSE) for our linear model is influenced by the nature of the time series data. In the case of NCLH, our model yielded accurate results because the volumes are relatively stable. However, for stocks like AMZN, where the volume grows, or at least has some identified trends, the model's predictions based on the median or average are not as accurate. This model is efficient for stocks that do not exhibit substantial movement or have no strong trend. In such cases, the future price is likely to be closer to the average of past prices.

Distribution shift Impact: Moreover, in years marked by significant distributional shifts, such as 2020, the model's accuracy was notably reduced. High volatility and changes in market distribution make an average-based projection obsolete.

Role of N: Furthermore, the choice of the parameter N plays a crucial role in the model's accuracy. When N is large, calculating the mean or median of a substantial data set can result in values that deviate significantly from the most recent observations. Given that time series data often exhibit autocorrelation, where future values tend to resemble recent ones, a large N can lead to a less accurate model prediction.

Average			
Year	**N = 10**	**N = 30**	**N = 60**
2017	1.0650e+15	1.1443e+15	1.1701e+15
2018	1.7497e+15	2.1703e+15	2.4198e+15
2019	6.1375e+14	7.9447e+14	9.6645e+14
2020	1.0926e+15	1.3987e+15	1.6892e+15
Median			
Year	**N = 10**	**N = 30**	**N = 60**
2017	1.1692e+15	1.2027e+15	1.2139e+15
2018	1.8633e+15	2.3784e+15	2.8145e+15
2019	6.6042e+14	7.4984e+14	8.6189e+14
2020	1.1989e+15	1.4971e+15	1.8448e+15

Table 4: Average Mean Square Error (MSE) for Amazon in Different Years and N values

Average			
Year	**N = 10**	**N = 30**	**N = 60**
2017	4.5775e+12	5.1801e+12	5.3269e+12
2018	1.6242e+12	1.6930e+12	1.8396e+12
2019	7.3778e+11	7.6456e+11	8.1064e+11
2020	2.3037e+14	3.5914e+14	5.7053e+14
Median			
Year	**N = 10**	**N = 30**	**N = 60**
2017	5.1504e+12	5.4904e+12	5.6268e+12
2018	1.8219e+12	1.8498e+12	1.9678e+12
2019	8.0551e+11	7.9844e+11	8.1930e+11
2020	2.6708e+14	4.0649e+14	6.4210e+14

Table 5: Average Mean Square Error (MSE) for NCLH in Different Years and N values

2.4 Banking holidays

Since Yahoo Finance does not provide data for banking holidays (for example, no data on December 25th), it becomes challenging to directly compare the forecasted volume for these holidays with the actual price. However, we can assess our model's performance by examining the holidays during which the market is not closed. The following result for AMZN stock in 2019, for N=10. The prices show poor accuracy. When we compare the Mean Square Error (MSE) for these specific holiday dates to the monthly average MSE, we find that the MSE for these holidays is consistently higher than the monthly average.

Date	**MSE for Average**	**MSE for Median**
Columbus Day	2.874×10^{15}	2.618×10^{15}
Veteran's Day	2.247×10^{15}	2.233×10^{15}
Friday after Thanksgiving	3.268×10^{15}	3.169×10^{15}
December 31	5.238×10^{15}	4.894×10^{15}

Table 6: Mean Squared Error (MSE) for Selected Holidays, N=10

3 Linear Autoregressive Models

3.1 Forecast for 10th, 31st and 61st day of the year.

Using N-day sliding window, we find coefficients A, B, C in linear autoregressive models of lag 1 and lag 2, to predict daily stock prices for the N+1st day in 2019 and 2020 for N values of 10, 30, and 60. To predict the N+1st day, we use a N day slicing window.

N	lag 1 forecast	lag 2 forecast	Real Price
10	1.2904e+08	1.1667e+08	1.2734e+08
30	7.3801e+07	7.5668e+07	9.9498e+07
60	6.9352e+07	7.0035e+07	6.2288e+07

Table 7: Stock Forecast Results for Amazon in 2019

N	lag 1 forecast	lag 2 forecast	Real Price
10	6.7922e+07	8.0487e+07	5.7932e+07
30	7.5834e+07	7.7315e+07	5.2124e+07
60	1.0928e+08	7.6804e+07	1.2252e+08

Table 8: Stock Forecast Results for Amazon in 2020

It's a bit better than the first method.

3.2 Yearly mean square error.

General performance analysis: Surprisingly, the performances of this model are similar to the previous ones (Table 9 and 10 vs table 4and 5). The autoregressive model captures the underlying trend. But the randomness of the volumes makes it difficult to model.

Lag analysis: Lag 1 and lah 2 generates the same results, we can assume that the randomness of the function we try to approximate is too important to be modeled with a linear model, whatever the dimension of this model. I previously did all the work with stock prices and not volumes (I mixed...), and We were able to observe that Lag 1 is more effective for NCLH stock, whereas Lag 2 is more effective for AMZN stock. This can be explained by the fact that NCLH exhibits greater volatility. In the case of Amazon stock, its price tends to follow a more linear trend. Consequently, the next price is likely to be influenced by the two preceding prices, making Lag 2 more suitable. I know this is not the answer to the question and I don't have the data anymore but the idea seemed worth noticing.

Role of N: For both stocks, we notice that increasing the value of N leads to improved results. This aligns with the typical behavior of linear regressions,

where smaller samples can result in more scattered data, making the regression less accurate.

Distributional shift: Once again, we can observe that our model struggles to account for the market disruptions caused by the COVID-19 crisis in 2020. Linear models are not well-suited to modeling highly volatile and chaotic markets.

Lag 1			
Year	**N = 10**	**N = 30**	**N = 60**
2017	6.1200e+15	1.9903e+15	1.2035e+15
2018	2.1817e+15	1.4232e+15	1.3694e+15
2019	1.2901e+15	6.6736e+14	6.2881e+14
2020	3.1660e+15	1.1627e+15	1.1665e+15
Lag 2			
Year	**N = 10**	**N = 30**	**N = 60**
2017	1.1647e+16	2.1911e+15	1.2216e+15
2018	3.5829e+15	1.5768e+15	1.4147e+15
2019	2.1158e+15	7.1278e+14	6.5929e+14
2020	7.2396e+15	1.4502e+15	1.3012e+15

Table 9: Average Mean Square Error (MSE) for Amazon in Different Years and N values

Lag 1			
Year	**N = 10**	**N = 30**	**N = 60**
2017	9.5911e+12	6.2522e+12	5.5509e+12
2018	3.3527e+12	1.7710e+12	1.7713e+12
2019	9.1992e+11	6.1704e+11	6.0309e+11
2020	2.7982e+14	2.3027e+14	2.3852e+14
Lag 2			
Year	**N = 10**	**N = 30**	**N = 60**
2017	1.8438e+13	6.5282e+12	5.5976e+12
2018	8.1788e+12	2.2675e+12	1.8112e+12
2019	1.3703e+12	6.1347e+11	5.9761e+11
2020	3.6144e+14	2.5022e+14	2.4359e+14

Table 10: Average Mean Square Error (MSE) for NCLH in Different Years and N values

3.3 Banking holidays.

Once again, we are going to assess the quality of our model during holidays when the stock market is open. We will compare the forecasted values with the

actual values for Columbus Day, Veterans Day, the day after Thanksgiving, and December 31 for Amazon stock in 2019.

In this analysis, we observe that the Mean Squared Error (MSE) for holidays is slightly lower than that of the first method. However, when looking at the annual average error for the year 2019, we find that the holiday predictions exhibit bigger degree of error. This suggests that forecasting stock prices during holidays still pose a significant challenge for our model.

Date	**Lag 1 MSE**	**Lag 2 MSE**
Columbus Day	3.5237×10^{15}	3.6432×10^{15}
Veteran's Day	2.2809×10^{15}	2.0598×10^{15}
Friday after Thanksgiving	3.2421×10^{15}	3.1347×10^{15}
December 31	5.6384×10^{15}	1.2878×10^{15}

Table 11: Mean Squared Error (MSE) for Selected Holidays, N=10

4 Neural Networks for Volume Forecasting

In this section, we employ neural networks to enhance the daily volume forecast. We provide a detailed account of the neural network approach, including its architecture, training data, hyperparameters, training procedures, and training loss.

Neural Network Architecture

To forecast daily volume, we design a neural network model with the following architectural details:

- **Input Dimension:** N, where N is the size of the slicing window, eventually N = 10 or 5 seems to deliver the best performances.
- **Hidden Layer Dimension:** 30 (chosen for optimal performance). Other values such as 10, 30, 50 100 were tested. We use 5 hidden layers for better performances.
- **Output Dimension:** 1 (representing the forecasted price).
- **Activation Functions:** ReLU activation function, a classic choice for time series forecasting models.

All data were normalized to facilitate the convergence of out model.

Training Data

For volume forecasting, we used daily trading data spanning a 6 months period. The neural network was trained on data from mid 2018 to 2019. Training data are composed of a N day slicing window of volumes and a Target volume to replicate. The training process was computationally efficient, making it more practical to train the algorithm over the entire period.

Hyperparameters and Training Details

The neural network's training relied on various hyperparameters and specific configurations:

- **Learning Rate:** After testing different values, 0.00005 was selected as the learning rate, as it produced the best results. Surprisingly, lr = 0.05 always blocks the neural network into a unique prediction.
- **Loss function:** Three loss functions were tested for this task, but we decided to use both **Mean Squared Error (MSE)** and **L1 Loss** for their effectiveness. The third option was Huber loss function.
- **Optimizer:** Two optimizers were considered: Stochastic Gradient Descent and Adam. Ultimately, we chose **Adam** for its superior performance.

- **Batch Size:** 2. We experimented with lots of batch sizes , but the model performed suboptimally with this values.
- **Number of Epochs:** After numerous trials with various values, we settled on **100 epochs**.

Training Procedure

During training, the neural network processes is given a N sized window of closing price and the N+1th price as target price to forecast. We perform forward and backward passes, updating the model's parameters to minimize the loss function. The training procedure is executed over the specified number of epochs to ensure model convergence. No premature stop was given at this time, the choose of the number of epoch is doing this work.

Training Loss

The training loss is an essential indicator of model convergence and performance. We monitor the training loss over epochs, analyzing its behavior to detect potential overfitting or underfitting. The training loss curve provides insights into how well the neural network is learning the training data. Here is a plot of the training loss over the epochs:

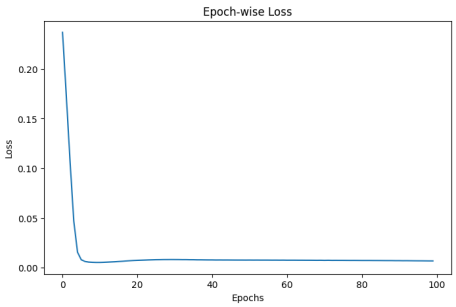

Figure 3: Training loss as a function of epoch

Model results

The results of the model have been satisfactory. We can compare the MSE for the two first month with the different models:

	Model 1	**Model 2**	**NN**
MSE	1.2258e+15	1.4372e+15	1.3563e+14

Table 12: Average MSE for AMZN in Different Years and N values

The model neural network model creates results 10 times more accurate than the two others. This show promising results for this method. To visualize, we can plots the time series of the volumes for our neural network generated data:

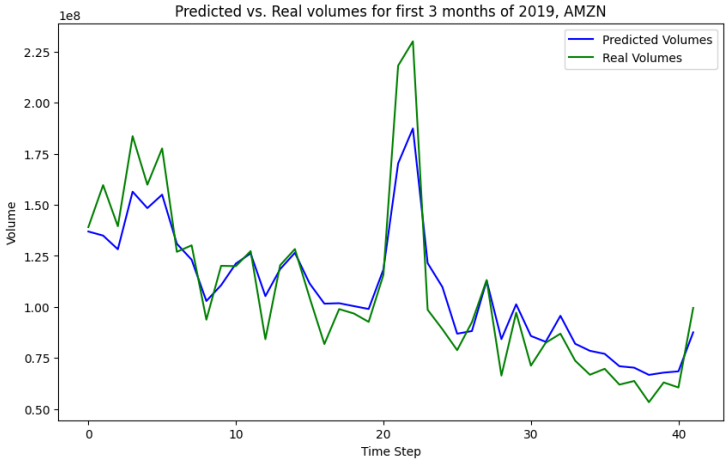

Figure 4: Training loss as a function of epoch

Conclusions

As we observed, the model did outperform the linear forecasting method. However, the loss is still consequent, several explanations for this performance gap are possible:

- The efficiency of my code might be suboptimal. This project was my first experience with coding neural networks, and it's possible that I missed some crucial details or optimal hyperparameter choices.
- We could use indicators or other features to help the algorithm predict the volume, such as the volatility.
- A conventional neural network model might not be well-suited for time series forecasting. Alternative architectures, such as Recurrent Neural Networks (RNNs) or their variants like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), could offer superior performance in capturing temporal dependencies and patterns in time series data.
- Overfitting can occur if the model becomes too complex relative to the amount of available data. Regularization techniques, such as dropout or weight decay, may help prevent overfitting and improve generalization.

5 Bonus

5.1 Introduction.

During the entire course of this analysis, holidays have proven to be particularly challenging to predict with our models. The distribution of trading activity on such days significantly deviates from the typical market behavior, and our existing models struggle to perform effectively. To enhance our forecasting accuracy for these specific days, I propose developing a dedicated neural network designed to address this issue.

5.2 Choice of data.

Train data: For the training dataset, we will follow a similar structure as before, using pairs composed of N days of sliced window volumes and the target volume for the N+1 day. However, in this case, the N+1 day will specifically correspond to one of four holidays: Columbus Day, Veteran's Day, the Friday after Thanksgiving, or December 31. To achieve this, our algorithm will be trained to predict one of these specific holidays over the past ten years. While the training data is somewhat limited, we may also consider incorporating data from other stocks to enhance the training dataset.

Test data: The test dataset is straightforward, consisting of the same holidays observed in the year 2019. The neural network will be provided with an N-day sliced window to forecast the volume for the N+1 day, which is one of the specified holidays.

5.3 Results.

Date	NN MSE
Columbus Day	4.7220e+13
Veteran's Day	1.8866e+14
Friday after Thanksgiving	5.0537e+10
December 31	2.1123e+15

Table 13: Mean Squared Error (MSE) for Selected Holidays, N=10

The results significantly outperform those achieved with alternative methods. To enhance the model's performance, one can employ Generative Adversarial Networks (GANs) to create additional input data. This synthetic data can then be seamlessly integrated into the training dataset, effectively expanding its size and potentially resulting in improved model accuracy and predictive capabilities.