AMAZON STOCK PRICE ANALYSIS AND PREDICTION USING MACHINE LEARNING

April 2023

Objective and Significance

- The scope of this project is to build several deep learning algorithms based on RNN techniques that can predict future values of an indicator using Time-Series Forecasting methods in order to achieve the highest possible accuracy
- Stock market prediction is the act of trying to determine the future value of company stock or other financial instruments traded on an exchange. The successful prediction of a stock's future price could yield significant profit.

Data Source and Description

Data source: Amazon Stock Data | Kaggle

Open = Price from the first transaction of a trading day

High = Maximum price in a trading day

Low = Minimum price in a trading day

Close = Price from the last transaction of a trading day

Adj Close = Closing price adjusted to reflect the value after accounting for any corporate actions

Volume = Number of units traded in a day

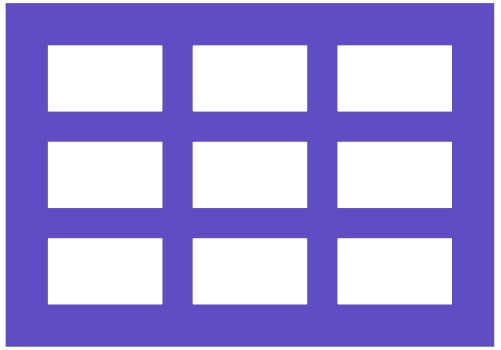
Wrangling

- The purpose of data wrangling and cleaning are:
  - To ensure that all features are of the correct data type.
  - To ensure that missing values are properly dealt with (imputation).
  - To prepare the data set for exploratory data analysis and statistical analysis

Column Name	Data Type
Date	datetime64
Open	float64
High	float64
Low	float64
Close	float64
Adj Close	float64
Volume	int64

Incorrect Values & Imputation

- No missing values in the dataset.
- Most of the features were the correct data type.
- Only change need was the date feature to a datetime64 object and set it to our index. This was done so we can work with it as a time series.



Seasonality

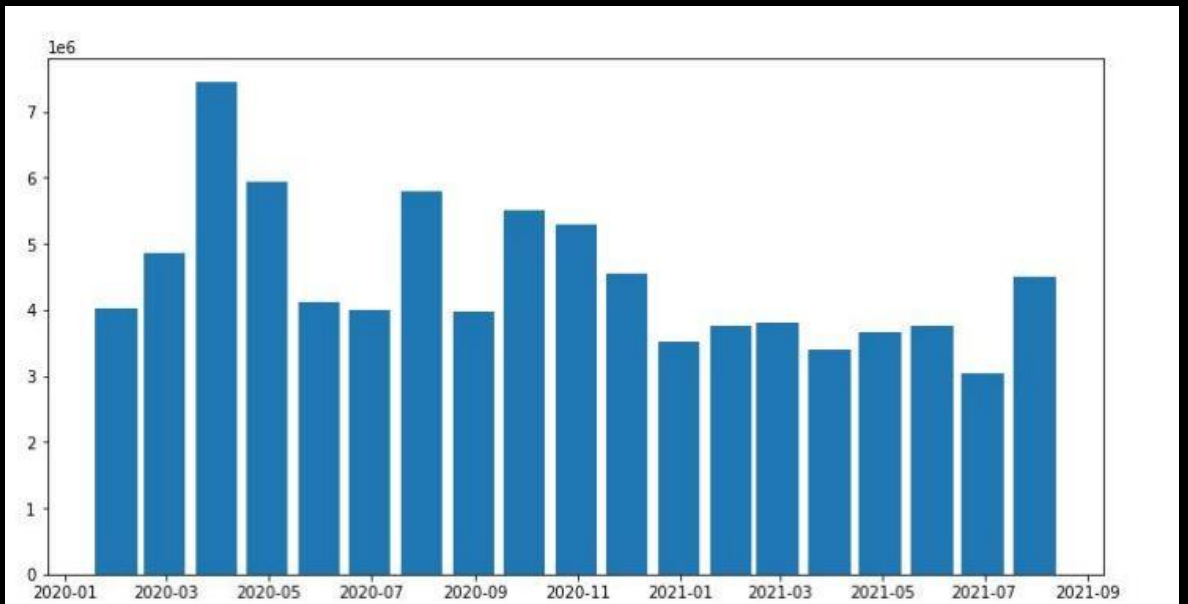
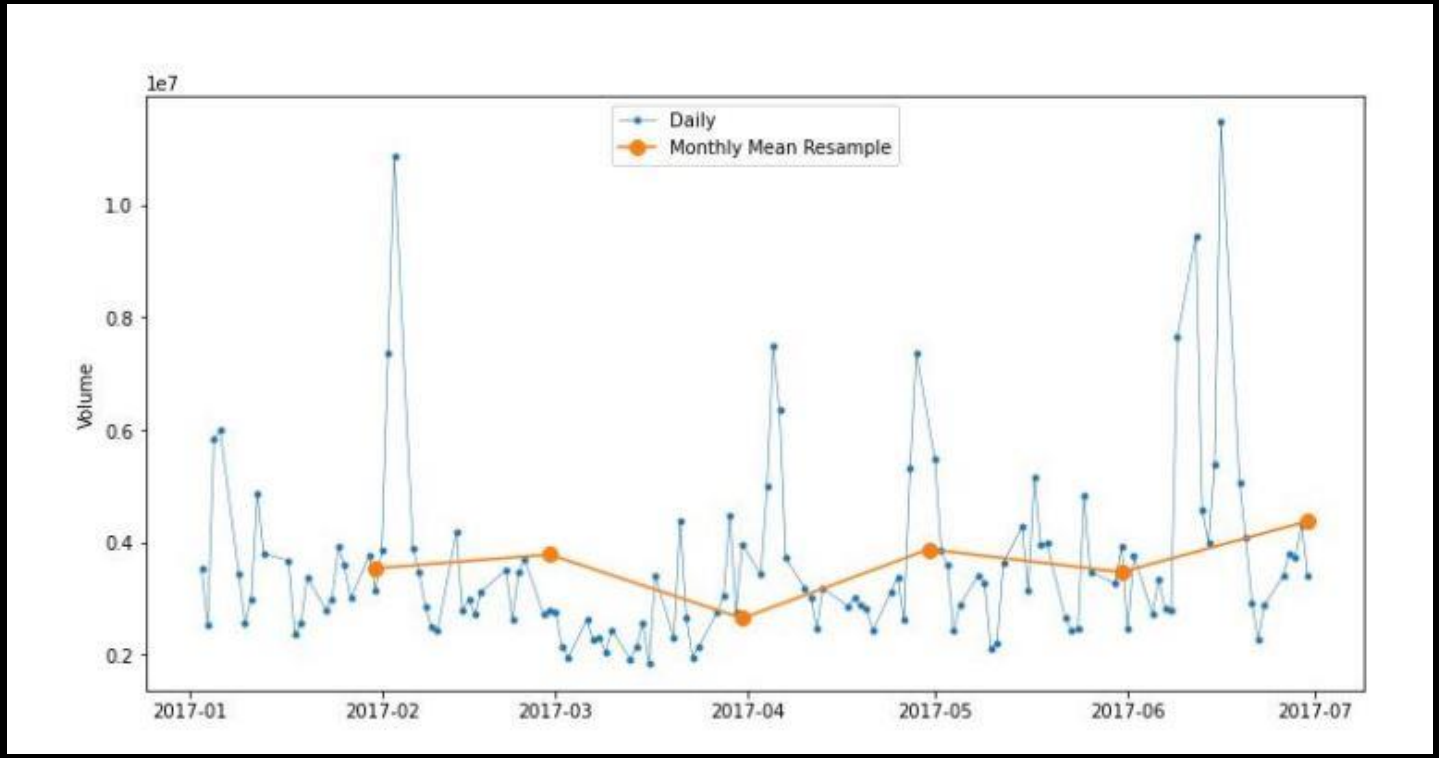
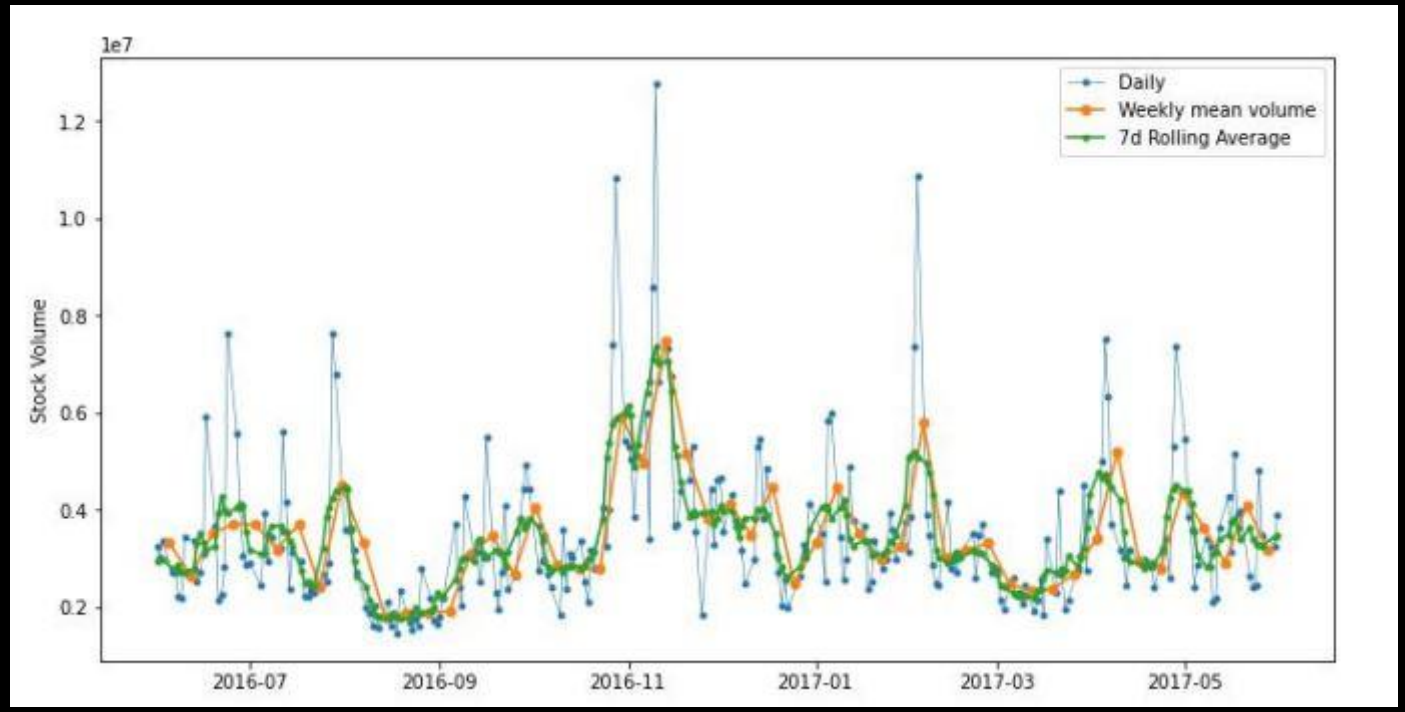


Figure 3: Bar Plot of Monthly Data in 2020

Resampling and Rolling



Rolling


1e7
1.2
1.0
0.8
0.6
0.4
0.2
Stock Volume
Daily
Weekly mean volume
7d Rolling Average
2016-07
2016-09
2016-11
2017-01
2017-03
2017-05

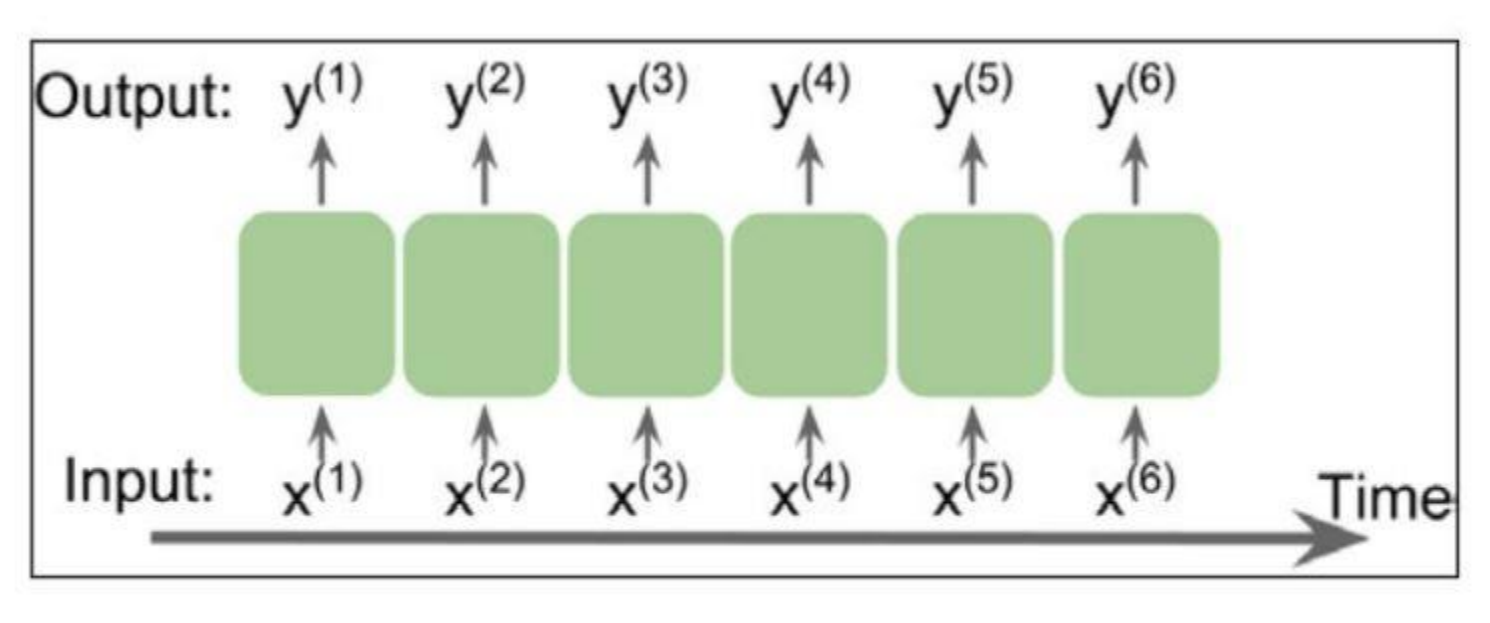
Data Modeling

- We elected to use Recurrent Neural Networks (RNNs) for modeling sequential data and a specific subset of sequential data—time-series data.

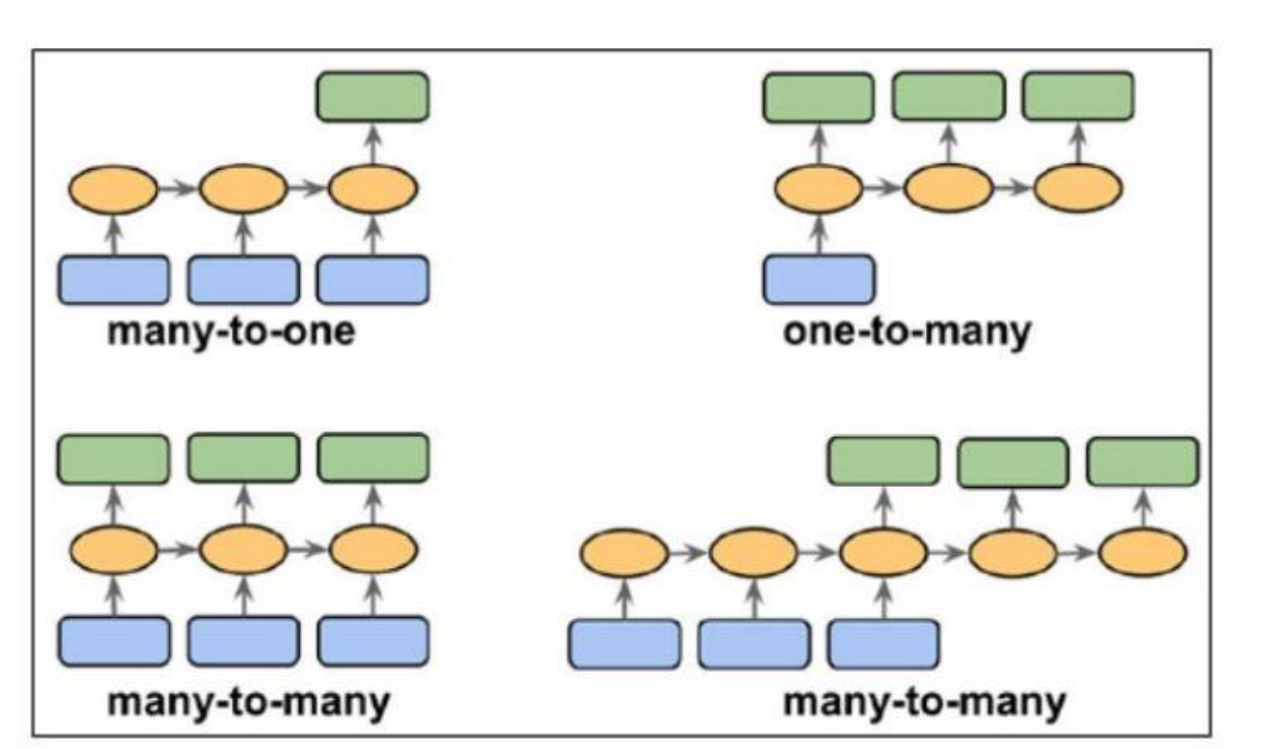
RNNs and Sequential Data

- What makes sequences unique, is that elements in a sequence appear in a certain order and are not independent of each other.
- typical machine learning algorithms for supervised learning assume that the input data is IID.
- This assumption is not valid anymore when we deal with sequences-by definition, order matters

Representing Sequences



Different Categories of Sequential Modeling



Challenges and Solutions

- Backpropagation through time will cause the vanishing gradient problem
- In practice, there are two solutions to this problem:
  - Truncated backpropagation through time (TBPTT)
  - Long short-term memory (LSTM)

  We elected to go with LSTM since it has been more successful in modeling long-range sequences by overcoming the vanishing gradient problem

Performance Evaluation

MODEL	MSE	RSME	MAE	R^2
RNN	54104.77	156.86	162.61	0.93

Conclusion and Recommendations

While the MSE, RSME and MAE are relatively high, we have very good R2 score. With a score of approximately 92.566% , we can interpret this to mean that 92.566% of the variance in explained by our RNN.

As data sets become more and more readily available, it may prove to be beneficial to conduct this analysis again in the future on different time series data sets and see how well the model performs.

It may also be helpful to build different models with different parameters to determine which one performs best on a given data set.

CONCLUSION

Thank you for listening. Any questions?