SE 207 Final Project:
Predicting Stock Market Opening Prices using GRU

Austin Paxton

3/19/2022

Goal:
The objective of this project is to investigate the effectiveness of predicting stock market Opening Prices using Gated Recurrent Units (GRU). A GRU is a special type of RNN which has an update gate and reset gate. These gates allow the model to have a long term memory which RNN's suffer from. GRU was selected over a LSTM neural network because the LSTM has 3 gates which results in longer training times (see figure 1). The speed and long term memory capabilities made this type of neural network the ideal candidate for stock market prediction.

Additionally, another objective is to analyze the prediction accuracy on daily and weekly time periods. Traditionally, stock market traders look at longer time frames first before looking into shorter time frames, this is because the longer time periods have less noise and give a better sense of direction in the stock market. Based on this, one would assume that the weekly time period would yield a more accurate GRU result, since the weekly time period has less noise than the daily time period. However, neural networks require a lot of data to train and if sufficient data is not provided this can detrimentally affect the results. Therefore, from the perspective of a data scientist, the daily period should yield more accurate results because more data is provided. (Daily period has 7734 data points while weekly period has 1521 data points)

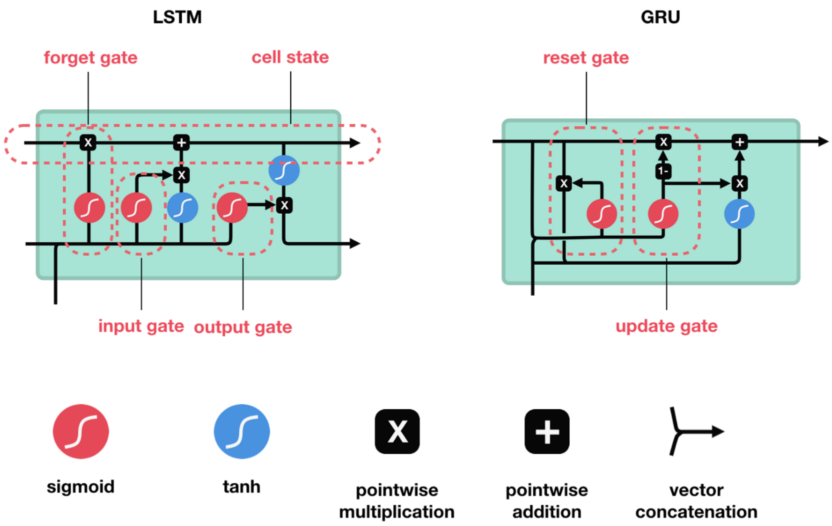

Figure 1: Comparison of LSTM and GRU Neural Networks

Stock Background:
The ETF $SPY was chosen to be analyzed in this report for two reasons. The first reason is that it has existed since 1993 so there is enough data to train, validate, and test the GRU. Additionally, the $SPY ETF is composed of 500 large cap and mid-cap stocks and is generally believed to mirror the overall stock market.

Data Collection and Usage:

The $SPY price history was downloaded in CSV format from Yahoo finance for daily and weekly time periods. For each time period, the Open, High, Low, Closing, adjusted Close, and volume are provided along with the date. However, for this project, only the open, high, low, and close prices remained after data-cleaning. Only the opening price will be analyzed with the GRU. Although the model was trained on open, high, low, and close prices, only the opening price will be analyzed. The weekly period has a total of 1521 data points (see figure 2), while the daily time period has a total of 7334 data points (see figure 3).

	Date	Open	High	Low	Close
0	1/25/93	43.968750	43.968750	43.750000	43.937500
1	2/1/93	43.968750	45.093750	43.968750	44.968750
2	2/8/93	44.968750	45.125000	44.531250	44.593750
3	2/15/93	44.468750	44.468750	42.812500	43.562500
4	2/22/93	43.687500	44.437500	43.468750	44.406250
...	...	...	...	...	...
1516	2/14/22	439.920013	448.059998	431.820007	434.230011
1517	2/21/22	431.890015	437.839996	410.640015	437.750000
1518	2/28/22	432.029999	441.109985	427.109985	432.170013
1519	3/7/22	431.549988	432.299988	415.119995	420.070007
1520	3/14/22	420.890015	441.070007	415.790009	441.070007

[1521 rows x 5 columns]

Figure 2: Sample of Cleaned Weekly Data

	Date	Open	High	Low	Close
0	1/29/93	43.968750	43.968750	43.750000	43.937500
1	2/1/93	43.968750	44.250000	43.968750	44.250000
2	2/2/93	44.218750	44.375000	44.125000	44.343750
3	2/3/93	44.406250	44.843750	44.375000	44.812500
4	2/4/93	44.968750	45.093750	44.468750	45.000000
...	...	...	...	...	...
7329	3/8/22	419.619995	427.209991	415.119995	416.250000
7330	3/9/22	425.140015	429.510010	422.820007	427.410004
7331	3/10/22	422.519989	426.429993	420.440002	425.480011
7332	3/11/22	428.119995	428.769989	419.529999	420.070007
7333	3/11/22	428.119995	428.769989	419.529999	420.070007

[7334 rows x 5 columns]

FIgure 3: Sample of Cleaned Daily Data

After cleaning the data, the opening price was plotted for both time periods. (See figure 4) However, because the data set is so large; the weekly and daily time periods are overlapped and indiscernible in figure 4. Therefore, it was plotted on a smaller time period (12/1/2020 to 7/1/2020) in order to see the difference between the two time periods. (see figure 5) Note that the daily Opening price data has more variance than the weekly Opening price data.

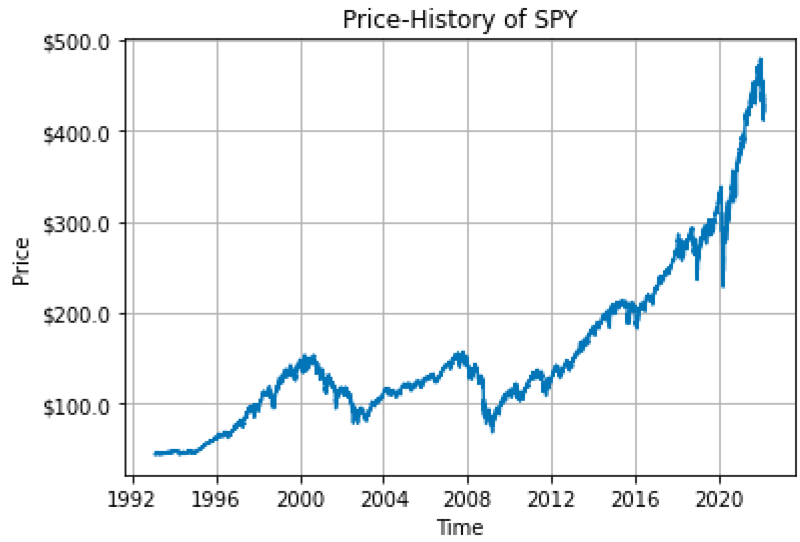

Figure 4: Price-History of SPY

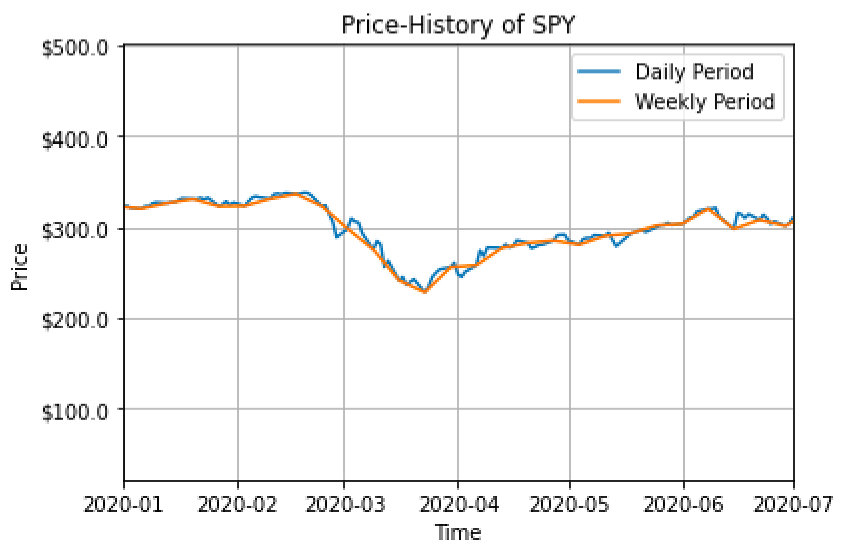

Figure 5: Price-History of SPY on Small Time Period

The data was normalized for both time periods as well in order to avoid exploding gradients during training. (see figure 6)

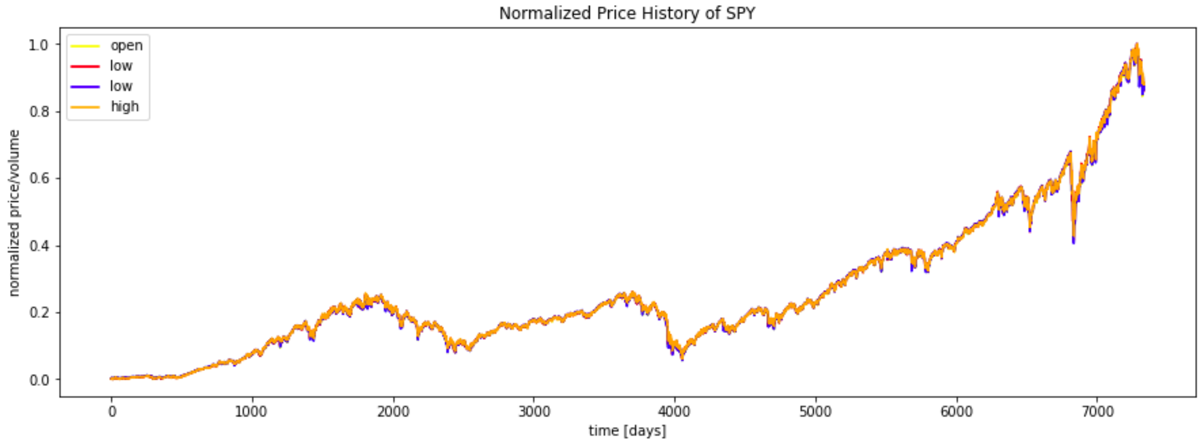

Figure 6: Normalized SPY data

For this project, 80% of the data will be utilized for training and the remaining 20% will be split evenly between validation and testing for 10% each. The data sets are both under 10,000 points and would not be considered big data. Therefore, larger validation and testing sizes were not practical.

Table 1: Sizes of Training, Testing, and Validation Data for Weekly and Daily Time Frame

	Weekly Time Frame	**Daily Time Frame**
Training Samples	1196	5847
Validation Samples	150	731
Test Samples	150	731

GRU Information:

In order to compare the prediction accuracy between the weekly and daily time frames, the same GRU parameters were applied to both models. Additionally, tensor flow was utilized to create the GRU.

Table 2: Parameters for GRU

	Weekly Time Frame	Daily Time Frame
Batch Size	30	30
Number of Epochs	100	100
Sequence Length	25	25
Number of Neurons	100	100

Learning Rate	0.001	0.001
Optimizer	Adam Optimization	Adam Optimization
Activation Function	Leaky Relu	Leaky Relu

Model Training:

The weekly time period was trained first because it has less data and is easier to debug errors without having to wait an extensive period of time to train. The training history of the of the weekly and daily time periods is shown below in table 3. From the training history of the weekly time period, a significant reduction in MSE train/val is observed after epoch 9.98. For the daily time period, a significant reduction in MSE train and MSE validation is observed after Epoch 5. After these Epochs, the MSE train and MSE validation do not continue to consistently decrease. Instead, they oscillate up and down before ending at the final MSE train and MSE validation (shown in table 4).

Table 3: Training History for Weekly and Daily Time Periods

Weekly Time Period	**Daily Time Period**
0.00 epochs: MSE train/valid = 0.026472/0.166495	0.00 epochs: MSE train/valid = 0.025734/0.161304
4.99 epochs: MSE train/valid = 0.000128/0.000289	5.00 epochs: MSE train/valid = 0.000022/0.000047
9.98 epochs: MSE train/valid = 0.000139/0.000412	9.99 epochs: MSE train/valid = 0.000010/0.000025
14.97 epochs: MSE train/valid = 0.000061/0.000149	14.99 epochs: MSE train/valid = 0.000008/0.000018
19.97 epochs: MSE train/valid = 0.000046/0.000114	19.99 epochs: MSE train/valid = 0.000009/0.000023
24.96 epochs: MSE train/valid = 0.000031/0.000075	24.99 epochs: MSE train/valid = 0.000009/0.000034
29.95 epochs: MSE train/valid = 0.000057/0.000123	29.98 epochs: MSE train/valid = 0.000008/0.000016
34.94 epochs: MSE train/valid = 0.000031/0.000133	34.98 epochs: MSE train/valid = 0.000008/0.000022
39.93 epochs: MSE train/valid = 0.000026/0.000091	39.98 epochs: MSE train/valid = 0.000007/0.000017
44.92 epochs: MSE train/valid = 0.000034/0.000173	44.98 epochs: MSE train/valid = 0.000007/0.000017
49.92 epochs: MSE train/valid = 0.000042/0.000221	49.97 epochs: MSE train/valid = 0.000011/0.000046
54.91 epochs: MSE train/valid = 0.000028/0.000081	54.97 epochs: MSE train/valid = 0.000009/0.000028
59.90 epochs: MSE train/valid = 0.000024/0.000080	59.97 epochs: MSE train/valid = 0.000007/0.000017
64.89 epochs: MSE train/valid = 0.000027/0.000124	64.97 epochs: MSE train/valid = 0.000011/0.000026
69.88 epochs: MSE train/valid = 0.000025/0.000140	69.96 epochs: MSE train/valid = 0.000008/0.000026
74.87 epochs: MSE train/valid = 0.000026/0.000081	74.96 epochs: MSE train/valid = 0.000007/0.000024
79.87 epochs: MSE train/valid = 0.000056/0.000314	79.96 epochs: MSE train/valid = 0.000007/0.000019
84.86 epochs: MSE train/valid = 0.000033/0.000167	84.96 epochs: MSE train/valid = 0.000007/0.000018
89.85 epochs: MSE train/valid = 0.000040/0.000100	89.95 epochs: MSE train/valid = 0.000008/0.000018
94.84 epochs: MSE train/valid = 0.000025/0.000102	94.95 epochs: MSE train/valid = 0.000006/0.000017
99.83 epochs: MSE train/valid = 0.000033/0.000167	99.95 epochs: MSE train/valid = 0.000007/0.000021

Table 4: Final MSE Test and MSE Validation for Weekly and Daily Time Frames

	Weekly Time Period	**Daily Time Period**
MSE Train	0.000033	0.000007
MSE Validation	0.000167	0.000021

Model Predictions:

The model predictions for opening price on both time frames were calculated and plotted below. Note that two plots are given for each time frame because the entire data set is so large it is difficult to see the discrepancies between predicted and actual opening price. The future prices come from the test data set. The plots corresponding to the weekly time frame are shown below in figure 7 and 8. Note that the GRU for weekly data consistently underpredicts the opening stock price. This is possibly due to the extensive uptrend that occurs in the test data. The GRU was trained on the red data in figure 7 and does not experience a significant uptrend as that occurs in the testing data. This is most likely why it is underpredicting the testing data. However, this underprediction does not occur for the daily time frame. From figure 10, it is apparent that the daily time frame future predictions are much more accurate than the weekly time frame (figure 8). The error prone areas occur at the peaks for the weekly time frame.

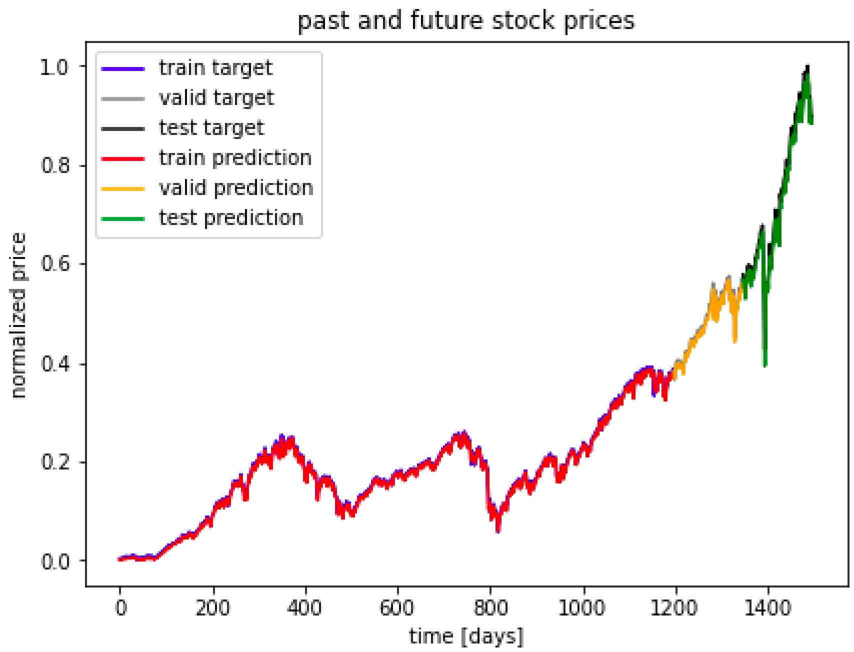

Figure 7: Past and Future Prices Predicted from GRU Trained on Weekly Data

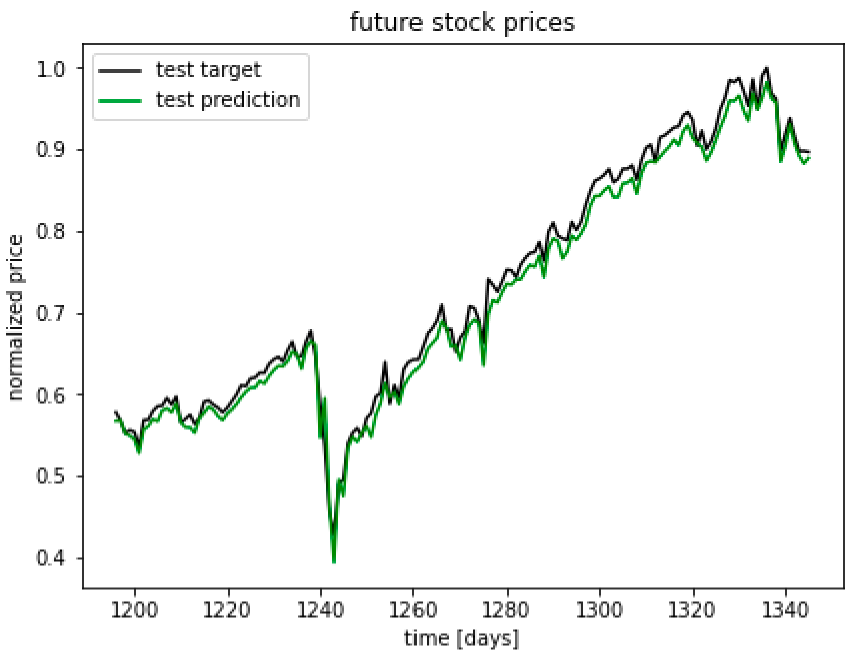

Figure 8: Future Stock Prices Predicted from GRU Trained on Weekly Data

The predictions for the daily data GRU are shown below in figures 9 and 10.

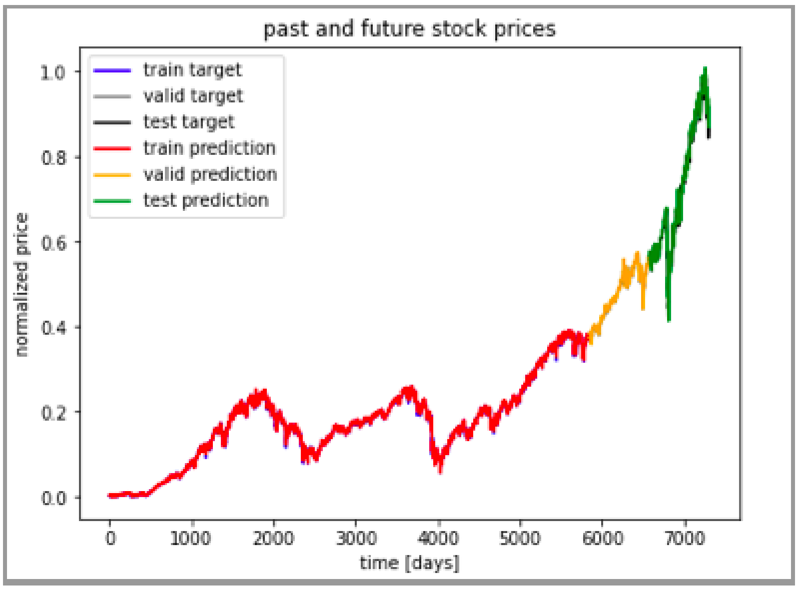

Figure 9: Past and Future Prices Predicted from GRU Trained on Daily Data

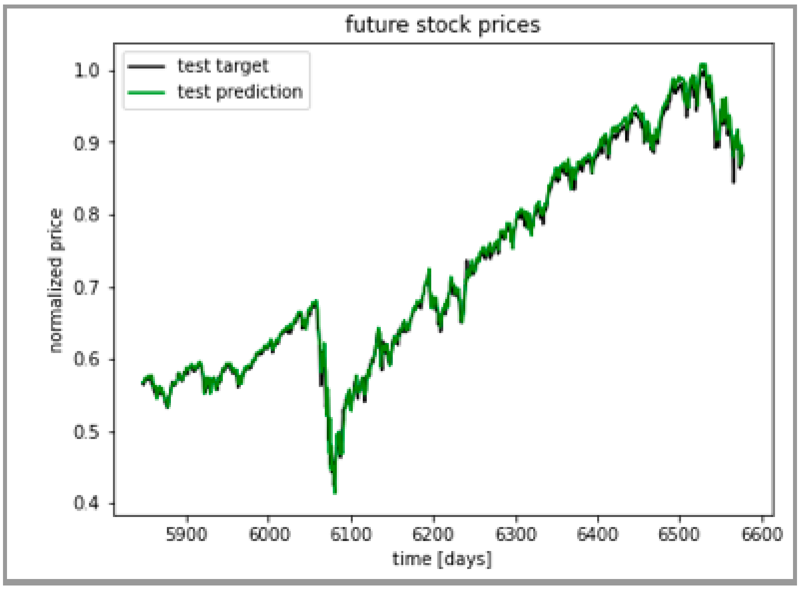

Figure 10: Future Stock Prices Predicted from GRU Trained on Daily Data

Conclusion:

It appears that the data scientist has triumphed over the stock trader. Although stock traders prefer to view longer time periods for more accurate results and noise reduction, this is not preferred for the data scientist training a GRU. The weekly time period's MSE train and MSE validation were both a magnitude larger than the MSE train and MSE validation for the daily time period. Additionally, when comparing the predictions on the test data (figures 8 and 10) it is apparent that the daily time frame has much more accurate results. The weekly time frame severely underpredicted the opening prices while the daily time frame matched up much more closely with the actual opening prices. The inaccuracy in the weekly time frame is likely due to its lack of extensive data, since the daily GRU was trained on approximately 4x more data points than the weekly GRU. For both models, the MSE validation was significantly higher than the MSE train value which implies that the model has an issue with over-fitting. To improve future results, a normal distribution of noise should be added to the stock market data to help alleviate the over-fitting issue.

Sources:

Raoulma, Raoul. “Ny Stock Price Prediction RNN LSTM GRU.” *Kaggle*, Kaggle, 8 Feb. 2018, https://www.kaggle.com/code/raoulma/ny-stock-price-prediction-rnn-lstm-gru/notebook.