CS 4641-A

Project Final Report

Submitted to:

Dr. Madhi Roozbahani

Date: December 7, 2021

Submitted by Team 13:

Ethan Benville, Carson Earnest, Alex Reyna, Jim Sherman, Conor Walsh

Introduction

The S&P 500 stock index consists of many of the top global companies in terms of market cap across many sectors. To assess a company's financial performance, shareholders and investors utilize a wide variety of metrics to predict how market factors (ex. changes in supply chain demand, industry competitor sales growth) or business developments (ex. new product launch, company merger) may impact a company's stock price on a given day. As these prices can fluctuate greatly on an inter-day and intra-day basis, machine learning algorithms can help create models which predict daily stock price changes and give insight into how historical trends may influence future stock price volatility.

Problem Definition

The main problem we are seeking to solve with this project is finding a more accurate method of predicting future stock prices using machine learning algorithms. Originally, we aimed to do this by feeding the algorithm historical information about a company's fundamentals, meaning its core business and valuation. These would be metrics such as revenue, margins, the stock's beta, and the stock's price-to-earnings ratio. However, due to the difficulty of finding historical fundamental information on a day-by-day basis, combined with the fact that some of these metrics remain static for long periods of time (revenue and earnings are reported quarterly), we decided to change our approach to price-action trading. This is a process by which traders analyze the stock price exclusively with the intention of predicting future returns. New features may be created based on the price, such as a moving average, but the analysis is still based solely on the stock's price at its core. We anticipate that the algorithm will likely be more accurate in predicting shorter-term yield compared to long-term yield, as a company's long-term performance is typically based on business fundamentals and earnings, which our algorithm does not take into account.

Data Collection

In order to pull the necessary data for testing and training, we used the yahoo finance API. We pulled two primary data sets: the first was daily price data from September 9th, 2010 (oldest data we could pull) until December 31st, 2020, and current financial metrics about the companies in our testing pool. For the midterm report, we broke up the training data into training and testing data. However, for the final report, we will pull daily pricing data from January 1st, 2021 until the present for the testing data.

Once the data was collected, we performed two methods of preprocessing. The first was to create 10, 50, and 200-day moving averages as well as 5, 20 and 260 day forward returns for the prices. It is important to note that the 5 day period is equivalent to a week of trading days, 20 days is equivalent to a month of trading days, and 260 is the actual number of trading days in a year. The moving average was calculated by pulling the average price over the specified period. The forward return was calculated with the following formula: (price of day i+[4] - price of day i)/(price of day i) for 5 day forward return, (price of day i+[19] - price of day i)/(price of day i) for 10 day forward return, and (price of day i+[259] - price of day i)/(price of day i) for the 260 day forward return, where the range of the indexing variable i started at 1.

Methods

For this part of the project, we implemented ridge regression in python and divided our data into training and validation/testing data. Sklearn was used to split the data, create a ridge regression model, and evaluate error measurements. Data was extracted from our cleaned csv file for the ETF VOO to use in our examples, however, any of the tickers can be used. All moving averages including 10 dma, 50 dma, and 200 dma were then extracted from the resulting data frame. Next, the data was split into 80 percent training and 20 percent testing, Next, ridge regression is applied, and the number of days to predict in the future are specified (in our case, 30 days), and matplotlib is used to create graphs, including forecasts, ridge regression prediction values, and the actual data. The features used to implement the ridge regression model were the different moving averages. Furthermore, a lasso regression model was implemented using the same features to predict stock price. Also, long short-term (LSTM) neural nets was implemented. The neural nets used slightly different data. Each neural net only used one of the three given moving averages as data. However, each datapoint was the bundle of 60 moving price averages used to predict a day in time. If the ith day was being predicted, then the previous i-60 moving averages were used to predict the price of the ith day. Error measurements (RMSE and R squared) are calculated to evaluate the effectiveness of the models. R-squared values were not calculated for the neural net.

Ridge regression and lasso regression were used to predict 5 day, 20 day, and 260 day forward returns. Each model used the stock price and the three different moving averages from *x* days ago to predict return values, where x is either 4, 19, or 259. These models were also used to predict future return values outside of our dataset.

Results / Discussion

Examples of the moving averages plotted with their respective stock prices are shown in Figure 1. Notice that the 10 day moving average is the closest to the actual trend in the data. This makes sense, as the 10 day moving average is more representative of a current day's stock price compared to a 50 or 200 day moving average.

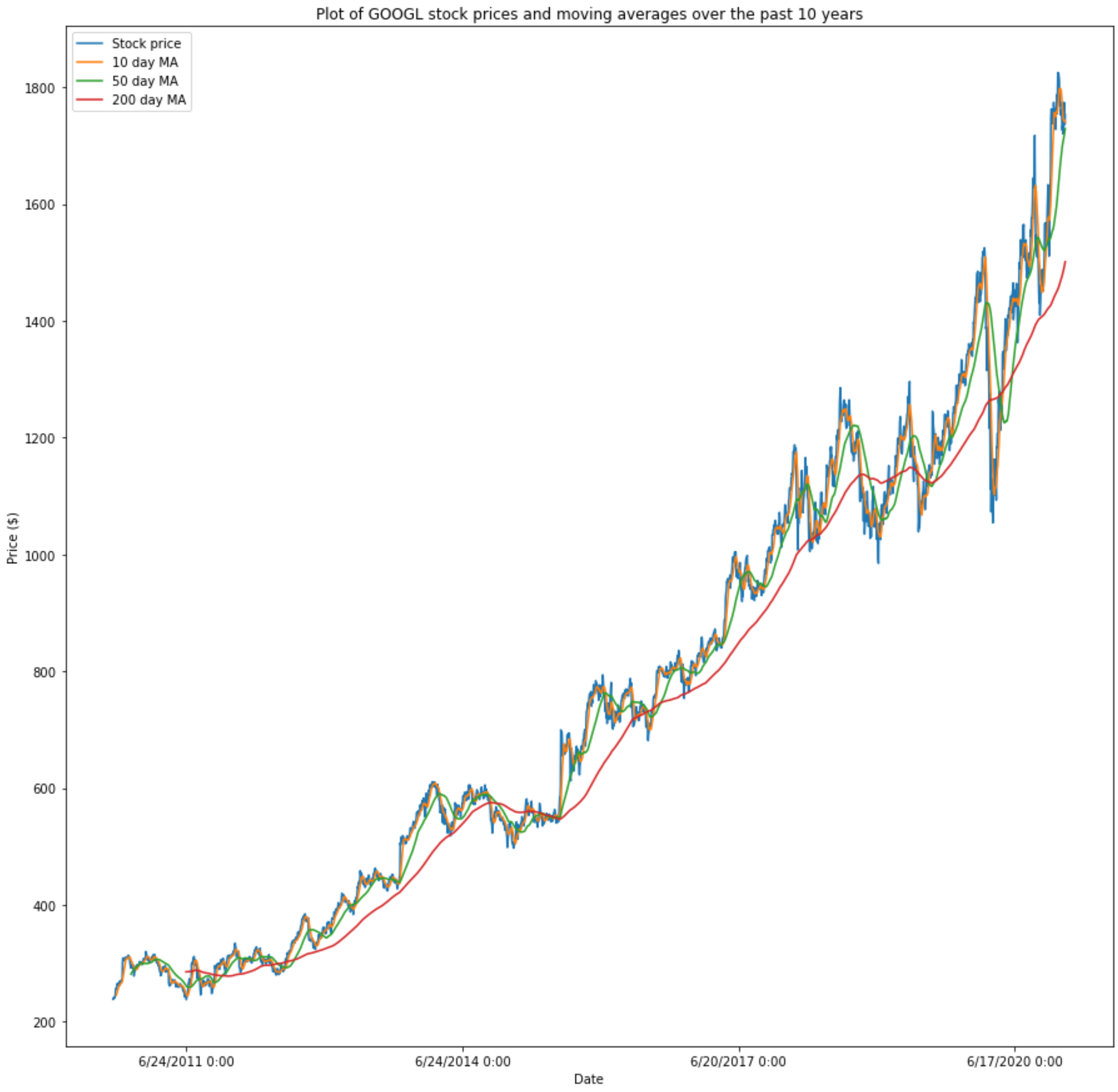


Figure 1: Plot of stock price by day along with moving averages

Supervised Learning

The supervised learning methods used for this project were ridge regression, lasso regression, and a Long Short-Term Memory (LSTM) neural network. For each method, a train test split was performed with a test size of 20% of the entire data. We used R-squared and RMSE to evaluate our ridge regression and lasso regression models. The neural net models only calculated RMSE values. The results suggest that predictions longer into the future tend to be less accurate.

Each section analyzes the results computed from ridge regression models, lasso regression models, and neural net models respectively. The final section, labeled Model Comparison, compares the results from each of the different models.

Ridge Regression

Examples of the ridge regression model using the moving averages as features are shown below in Figure 2. Each ridge regression graph is accompanied by another graph (shown in Figure 3) that displays the difference between the predicted stock price (red dots) and the actual stock price (blue line). Each graph has the respective RMSE and R-squared value generated from applying ridge regression to the data. The RMSE value expresses how well the predicted stock prices created from the training data match the actual stock prices. The R-squared value expresses how well the test data of the different moving averages predicts the actual test data of the stock prices.

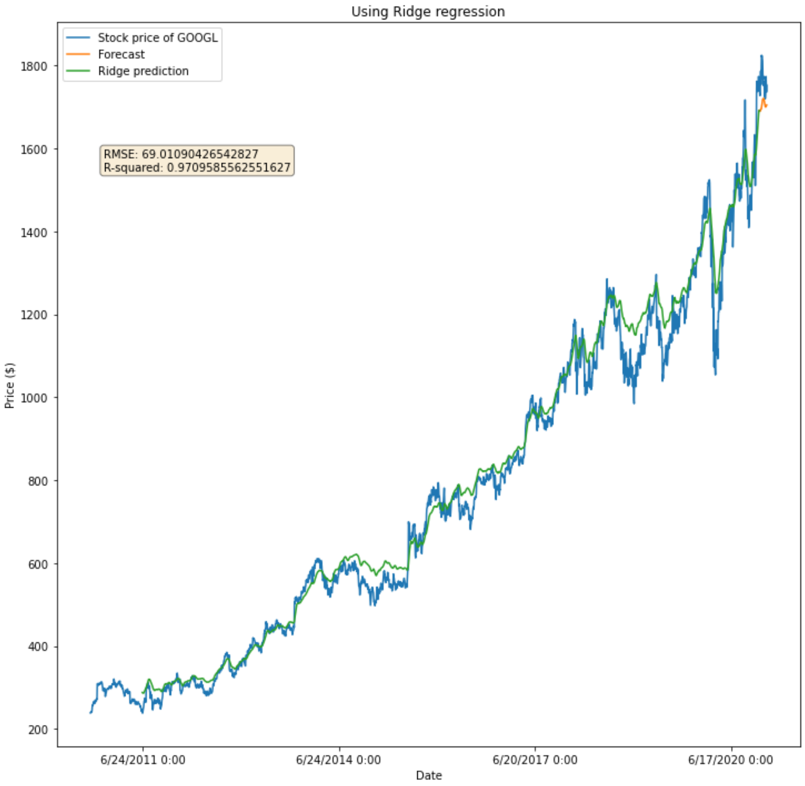


Figure 2: Fitted ridge regression stock price values

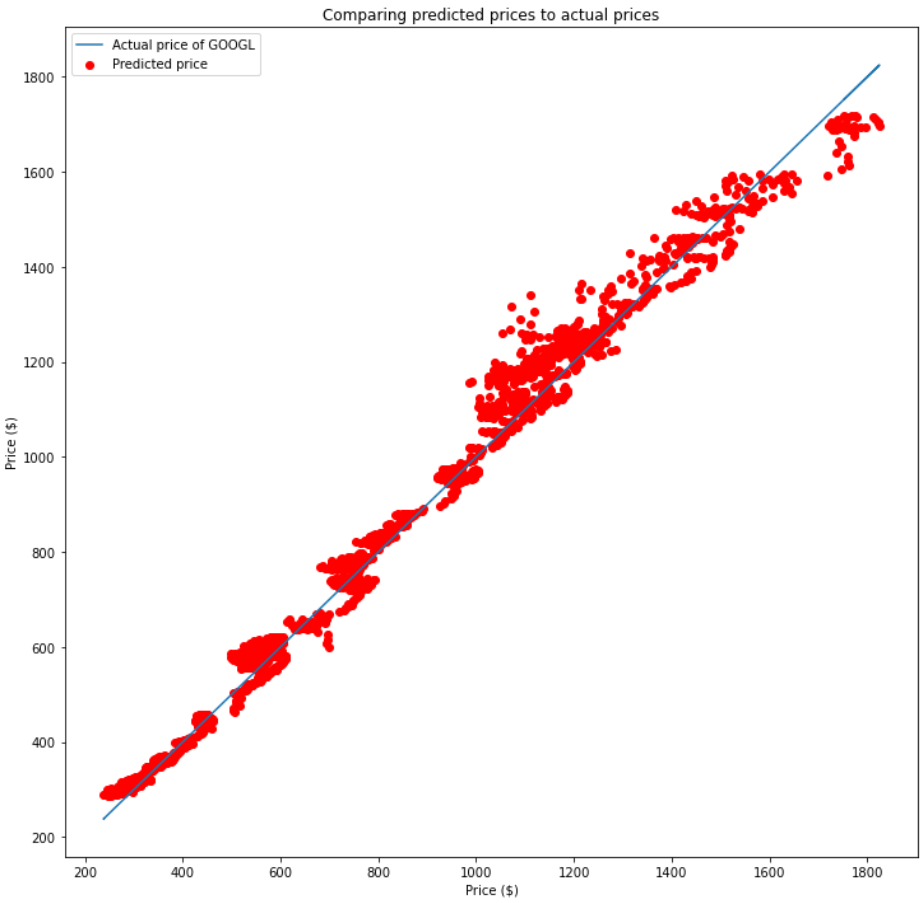


Figure 3: Plot of ridge regression stock price predictions vs actual stock prices

Notice the RMSE values are not very close to 0. However, compared to the 2500+ data points we used and the fact that the data ranges from $200 to $1,800, this RMSE value is relatively low. We can tell this value is relatively low by looking at how far off the red dots are from the blue line. For each of the examples above, the predicted prices (red dots) are relatively close to the actual price (blue line). As for the R-squared value, for both graphs, this value is very close to 1, which means the test data predicts the price of the stock pretty accurately. However, both the RMSE and R-squared values may increase or decrease depending on how much data we use to train and how much we use to test. Below in figures 4, 5 and 6 are examples of one stock's regression model using testing data of 25 days, 50 days, and 100 days.

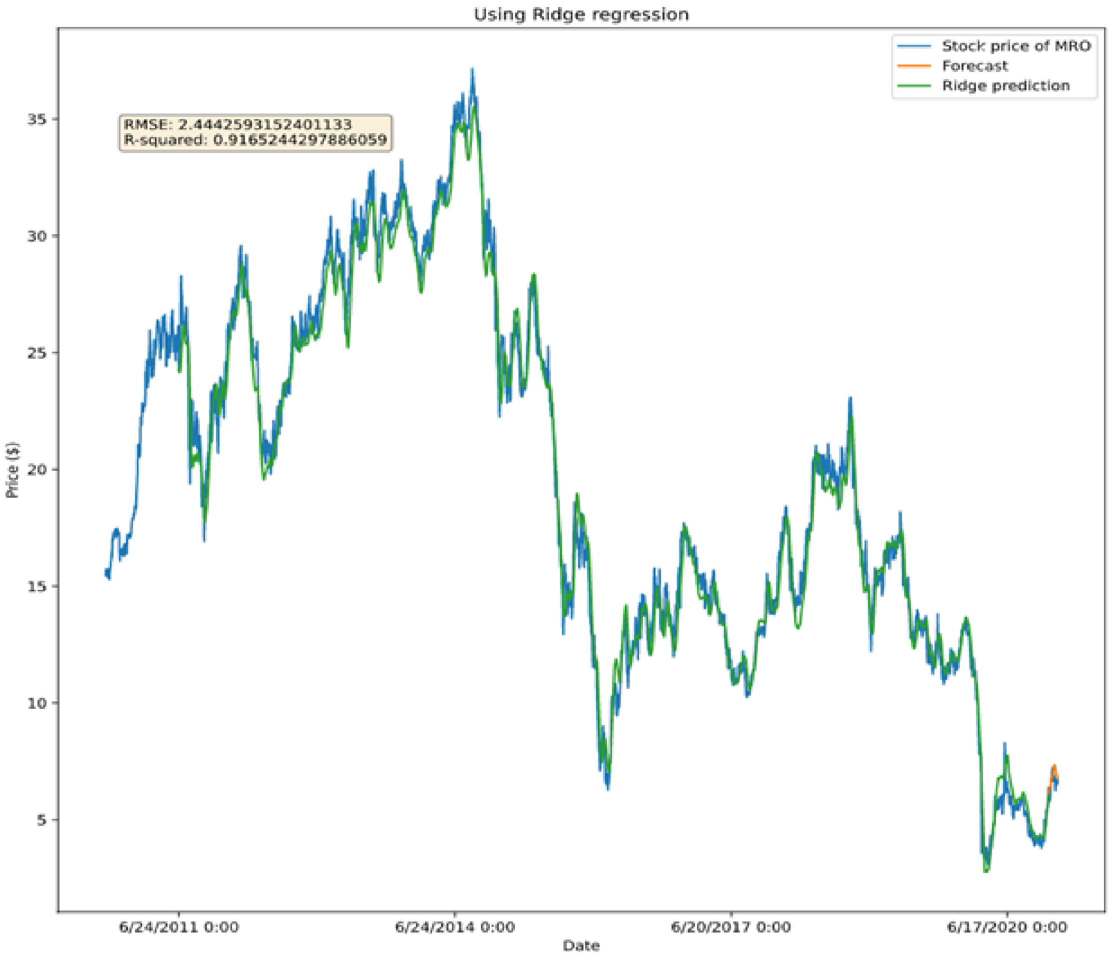


Figure 4: 25 day forecast using ridge regression stock price predictions

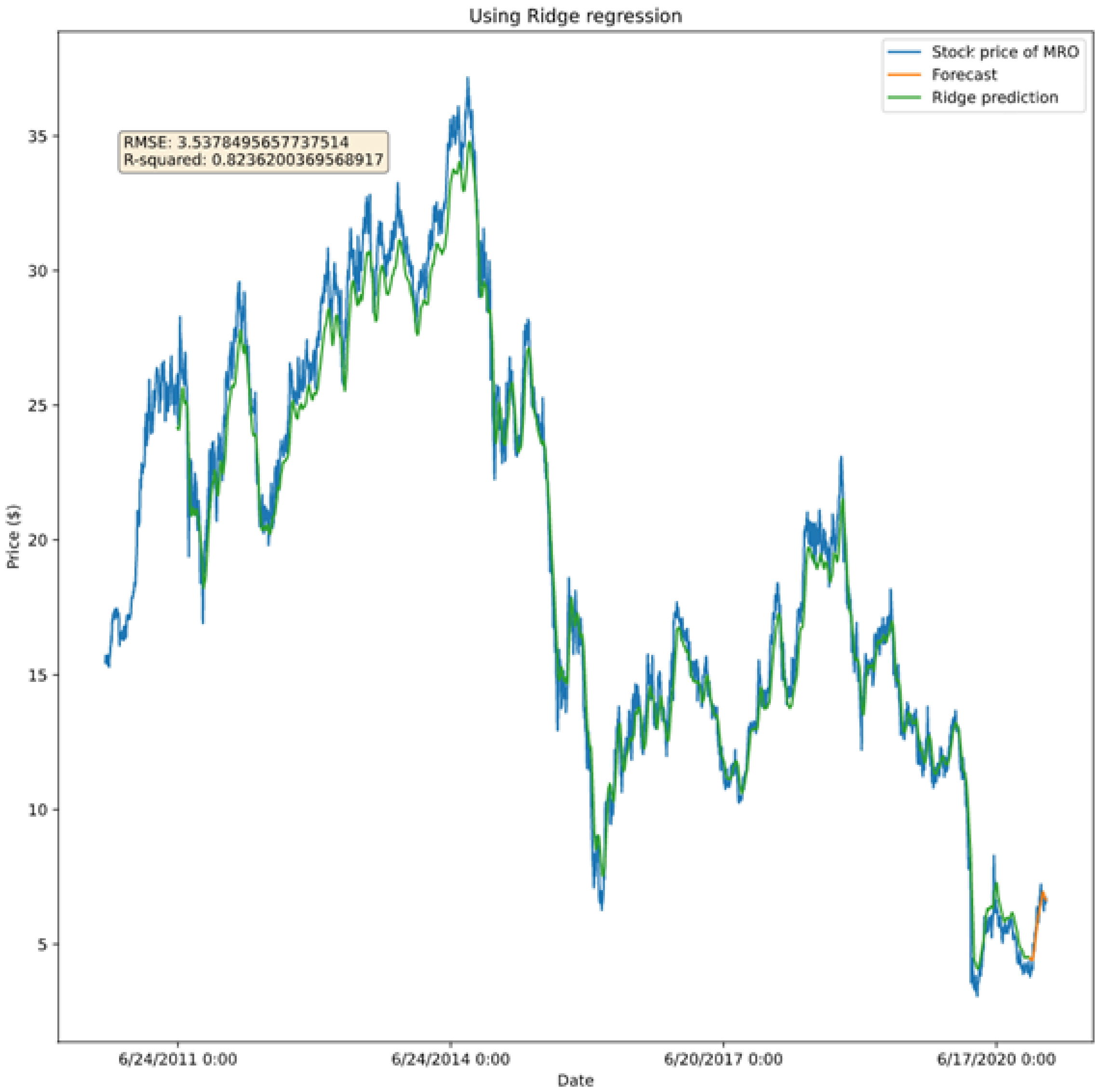


Figure 5: 50 day forecast using ridge regression stock price predictions

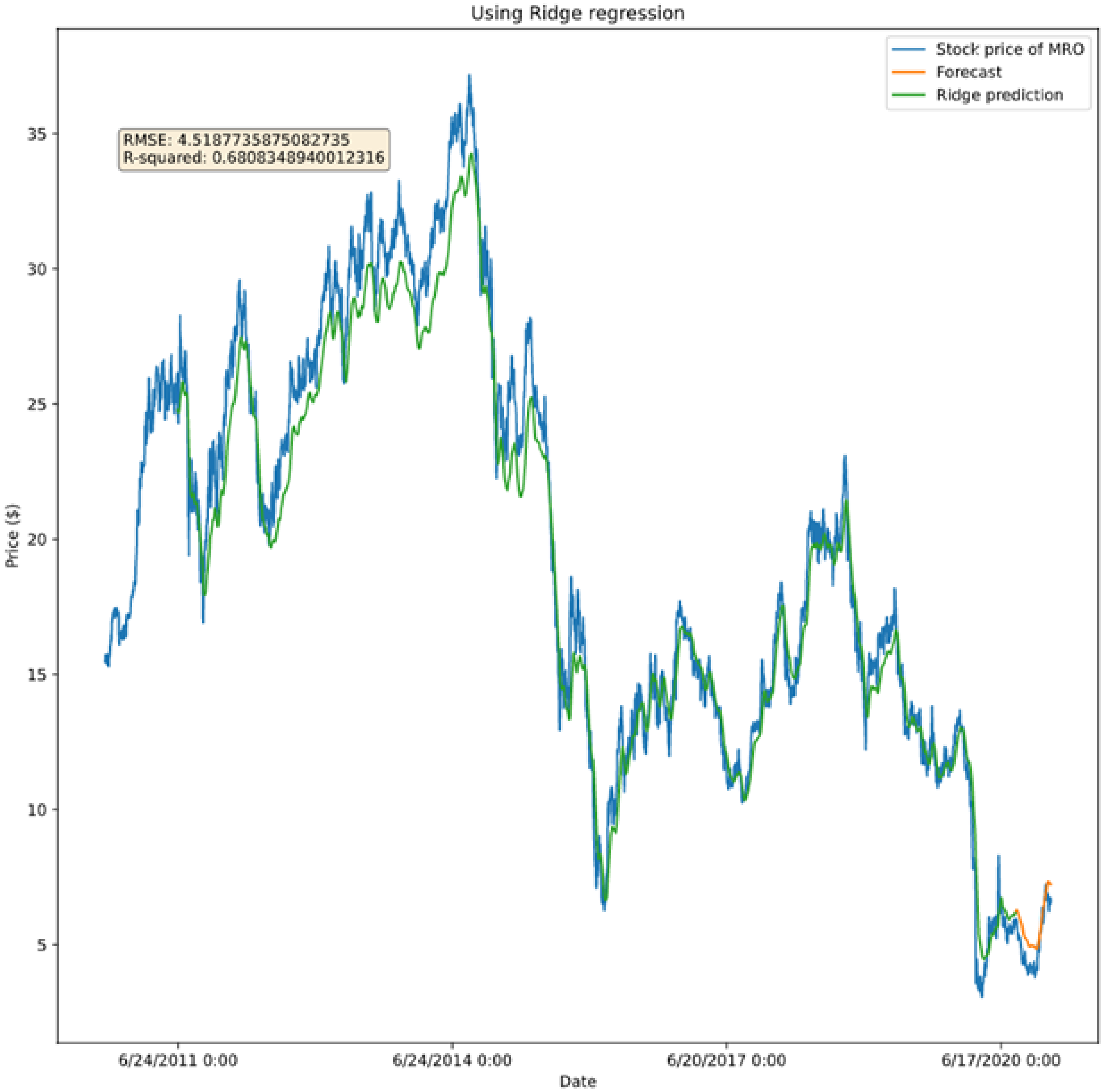


Figure 6: 100 day forecast using ridge regression stock price predictions

Notice that as we increase the number of days we are predicting, which decreases the size of our training, the RMSE values increase and the R-squared values decrease. These trends in RMSE and R-squared combined mean that the model gets less accurate the further you try to predict in the future. Therefore, ridge regression is best applied to stock price determination if you are only determining a short number of days ahead. This follows what we predicted earlier that ridge regression is more accurate for short term prediction compared to longer term prediction.

Furthermore, 5 day (one week), 20 day (one month), and 260 day (one year) moving return values were predicted using a separate ridge regression model. Figures 7-9 display the models predicting 5 day, 20 day, and 260 moving return values of Google's stock compared to the actual moving averages. The features used to create this model were stock price as well as the moving averages. For example, the 5 day return prediction used the starting stock price from 5 days before, as well as the moving averages recorded 5 days before.

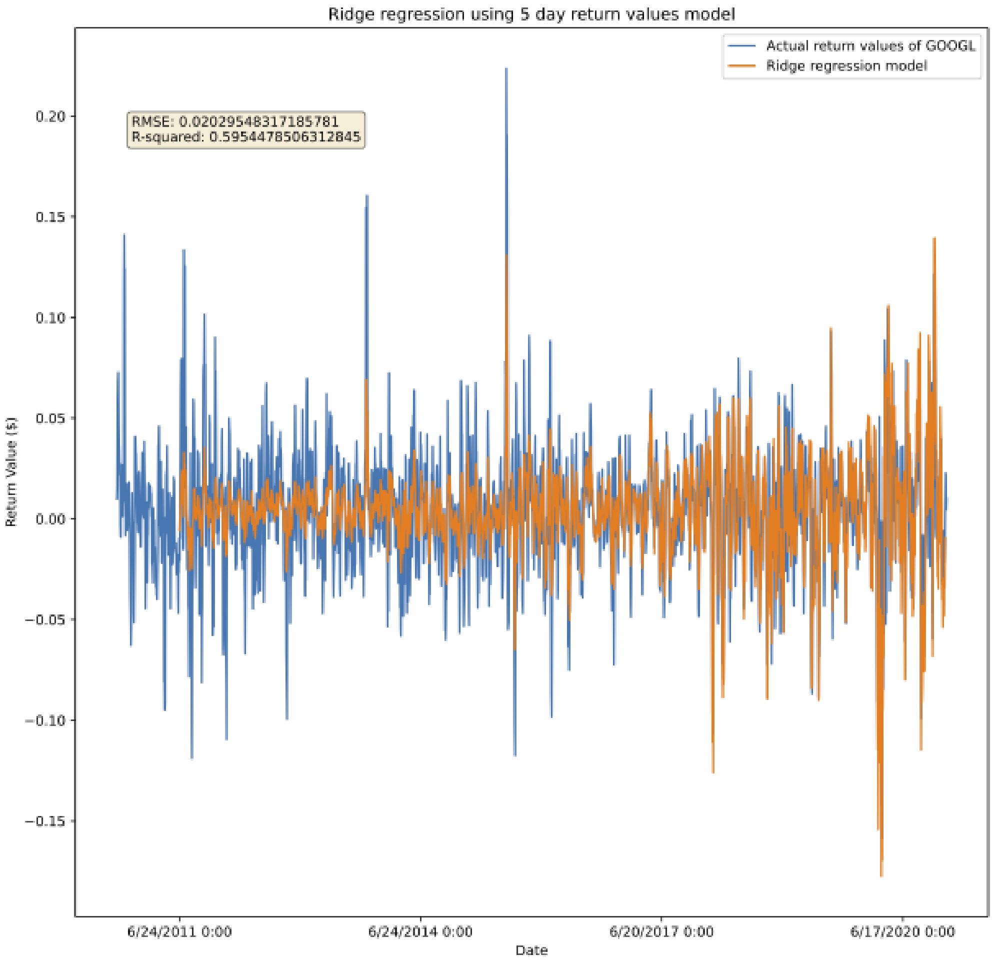


Figure 7: 5 day moving average predictions of GOOGL using ridge regression vs. actual 5 day return values

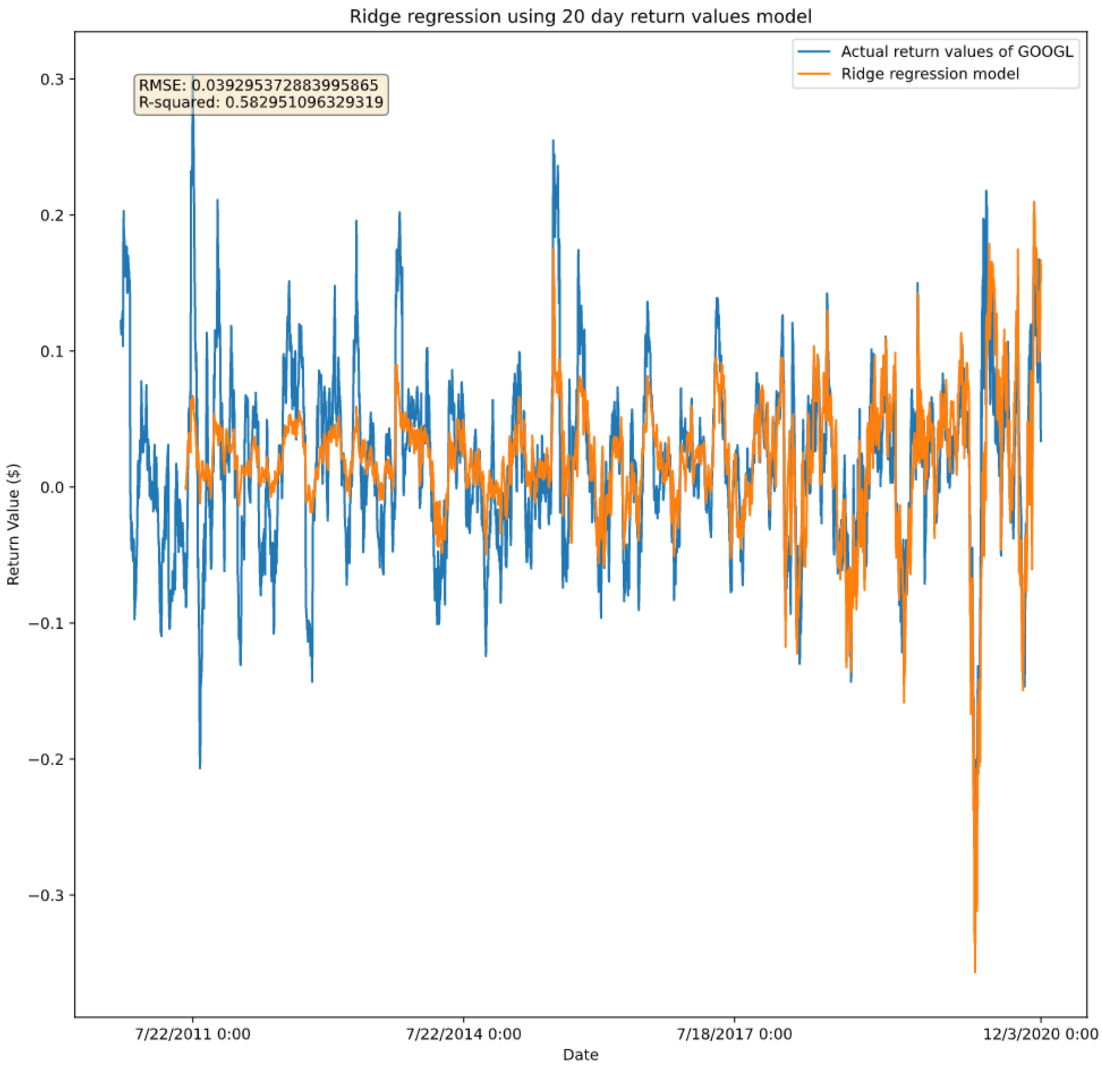


Figure 8: 20 day moving average predictions of GOOGL using ridge regression vs. actual 20 day return values

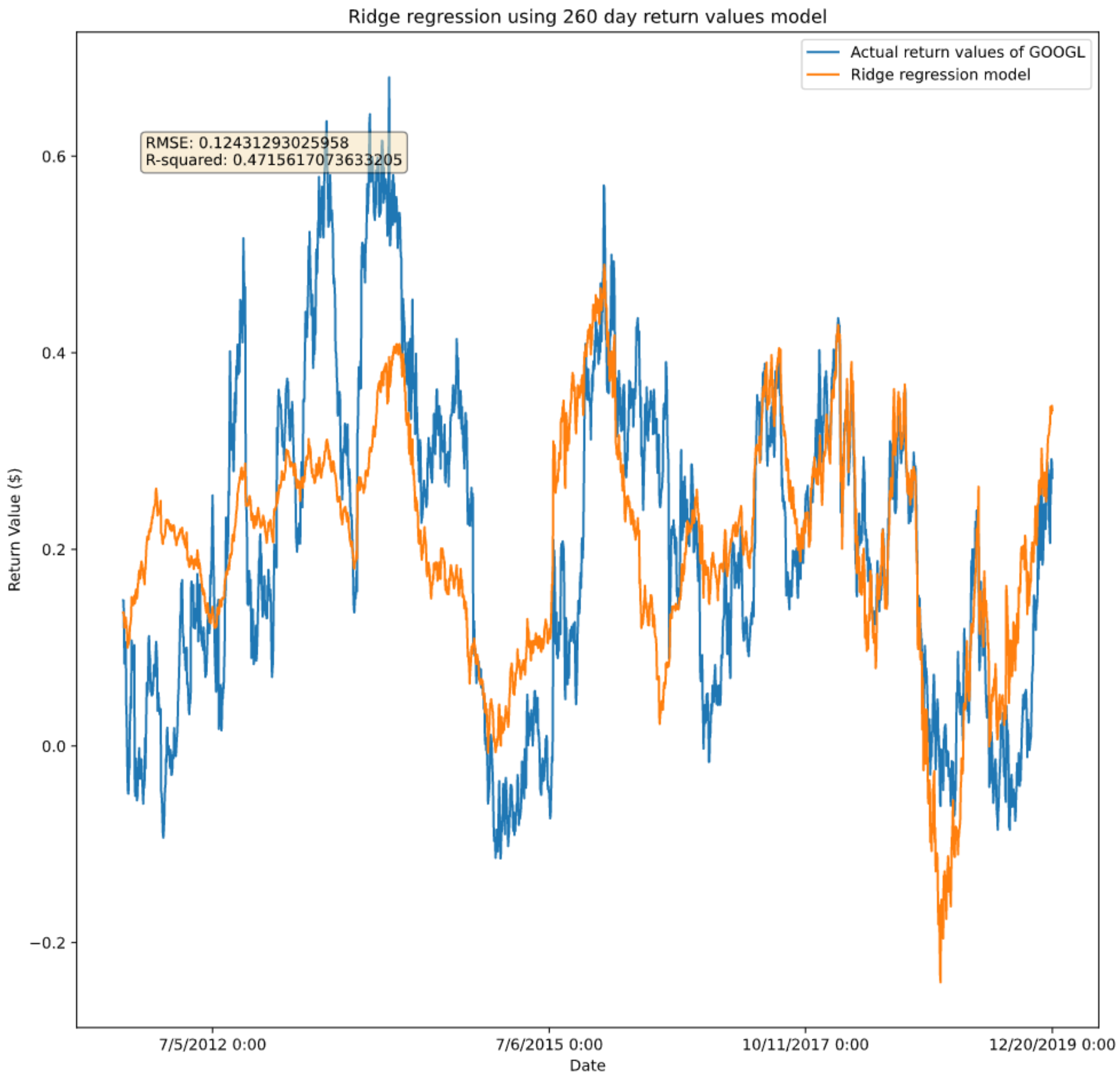


Figure 9: 260 day moving average predictions of GOOGL using ridge regression vs. actual 260 day return values

Tables 1 and 2 compare the r-squared and the RMSE values of the ridge regression models using 5 day, 20 day, and 260 day moving return values of GOOGL and MRO (not displayed above).

Moving day return value	**RMSE value**	**R-squared value**
5 day	0.0202955	0.59544785
20 day	0.03929537	0.5829511

260 day	0.12431293	0.4715617

Table 1: RMSE and R-squared values of different moving return predictions of GOOGL stock using ridge regression

Moving day return value	**RMSE value**	**R-squared value**
5 day	0.03985127	0.5938015
20 day	0.09171964	0.6481888
260 day	0.253353404	0.49152697

Table 2: RMSE and R-squared values of different moving return predictions of MRO stock using ridge regression

Notice how predicting one year ahead of time is much less accurate than trying to predict one week or one month ahead of time. This is shown through the RMSE values being much higher as well as the r-squared values being much lower in the 260 day return predictions compared to the 5 day and 20 day return value predictions. This further follows our prediction that ridge regression is more accurate for short term prediction compared to longer term prediction.

These models can be used to predict future return values outside of our dataset. Since we are predicting data outside of our dataset, there are no RMSE or r-squared values. Figures 10-12 display future return value predictions of Google's stock.

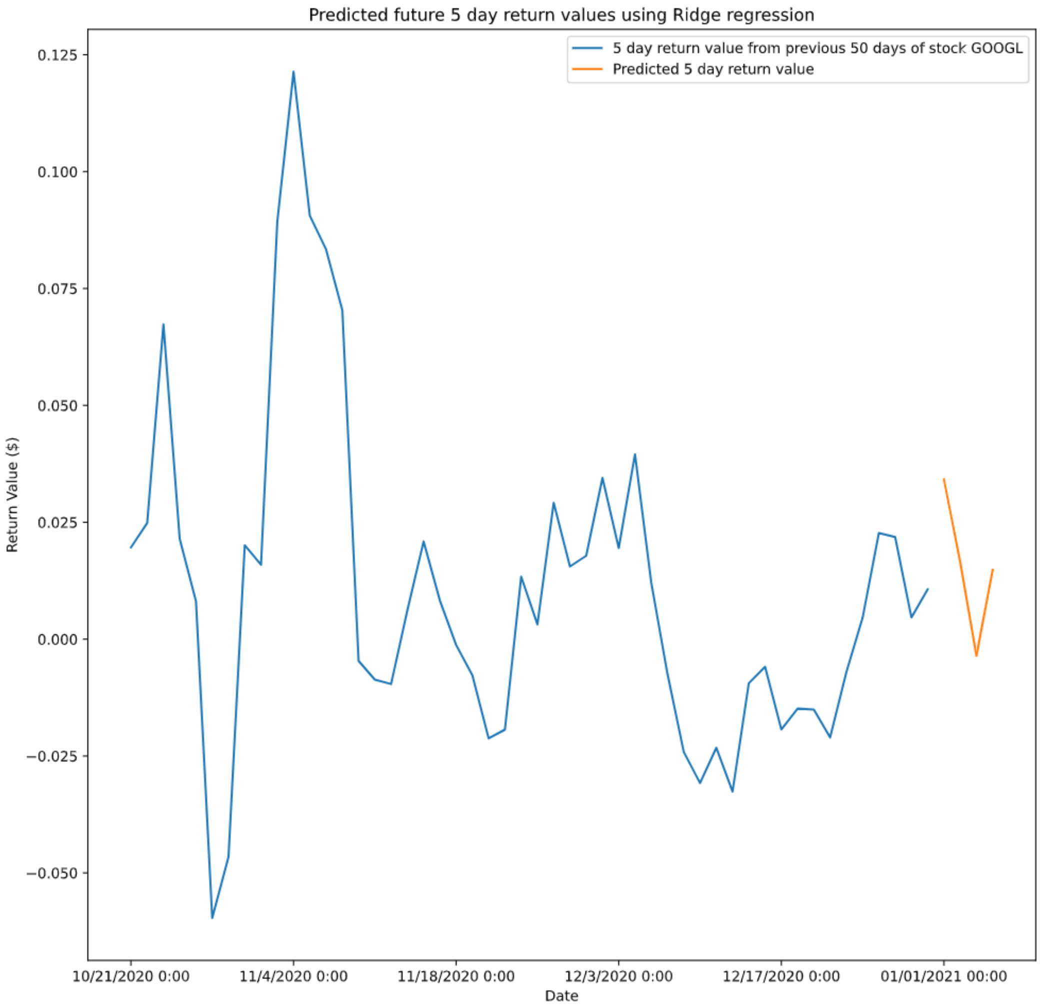


Figure 10: Predicted future 5 day return values of GOOGL using ridge regression

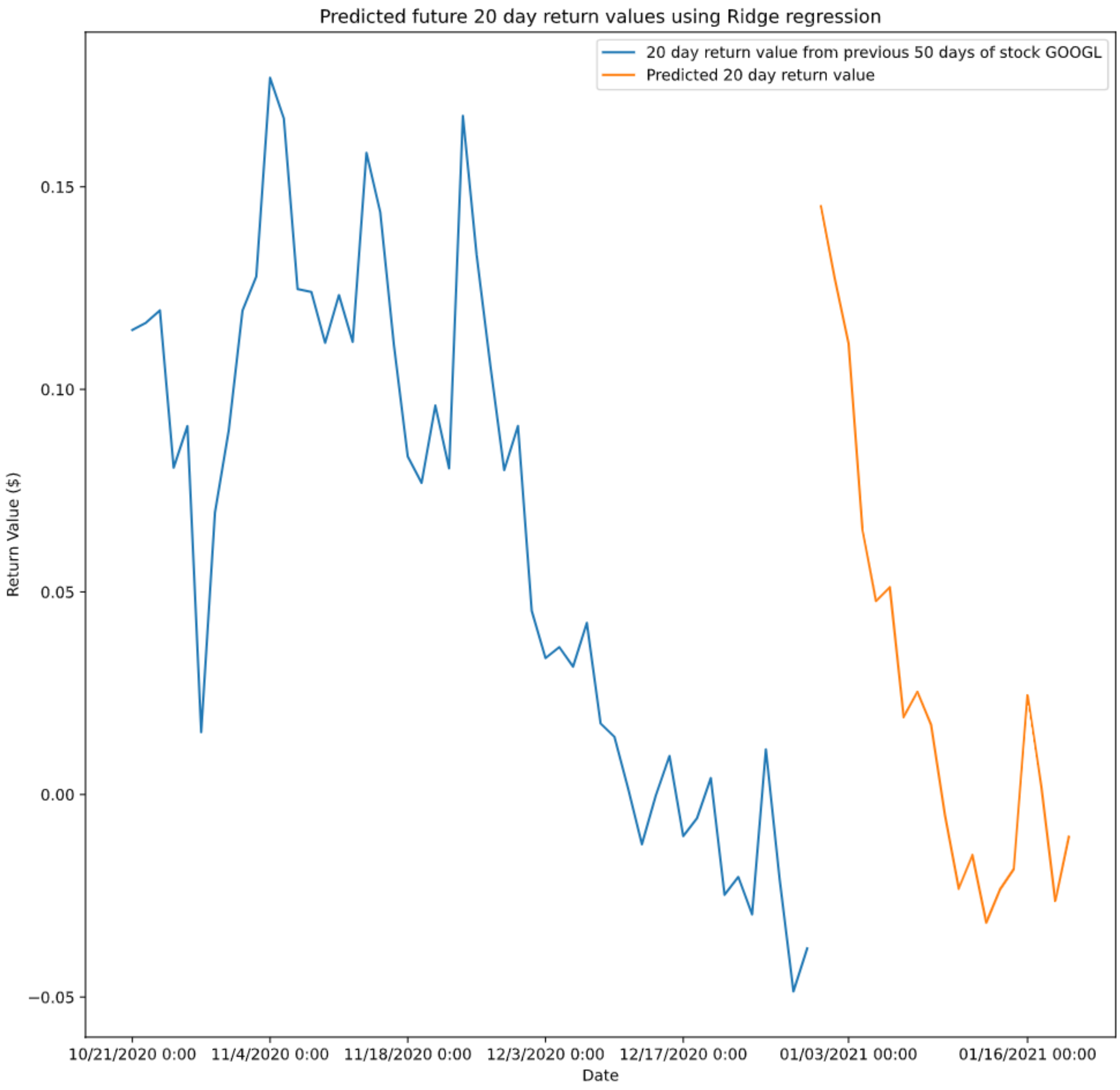


Figure 11: Predicted future 20 day return values of GOOGL using ridge regression

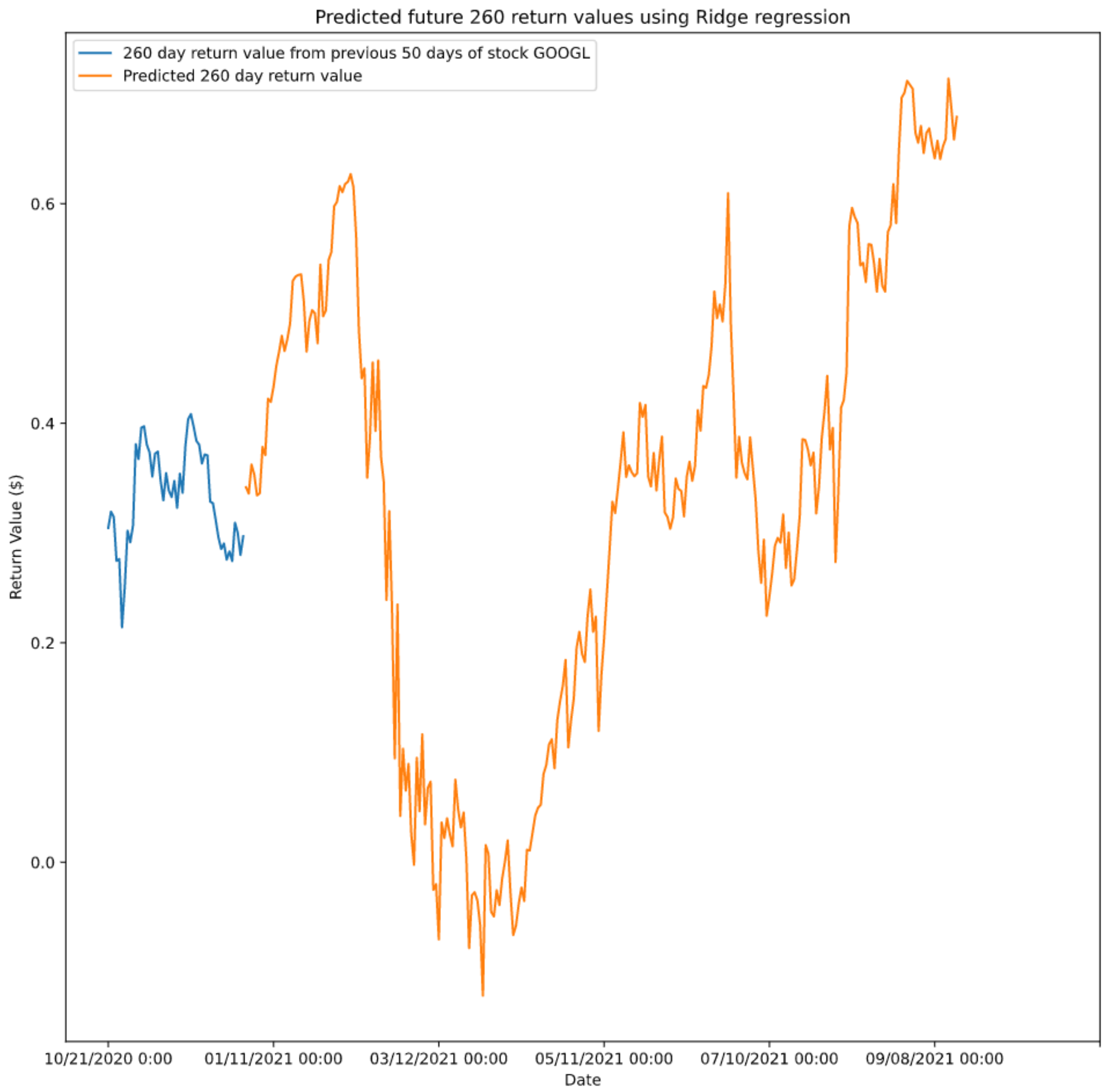


Figure 12: Predicted future 5 day return values of GOOGL using ridge regression

Lasso Regression

Using the same moving averages used to create the ridge regression model, a lasso regression model was implemented to create a model of stock prices over time. The model of

MRO stock price over time is shown below in Figure 13.

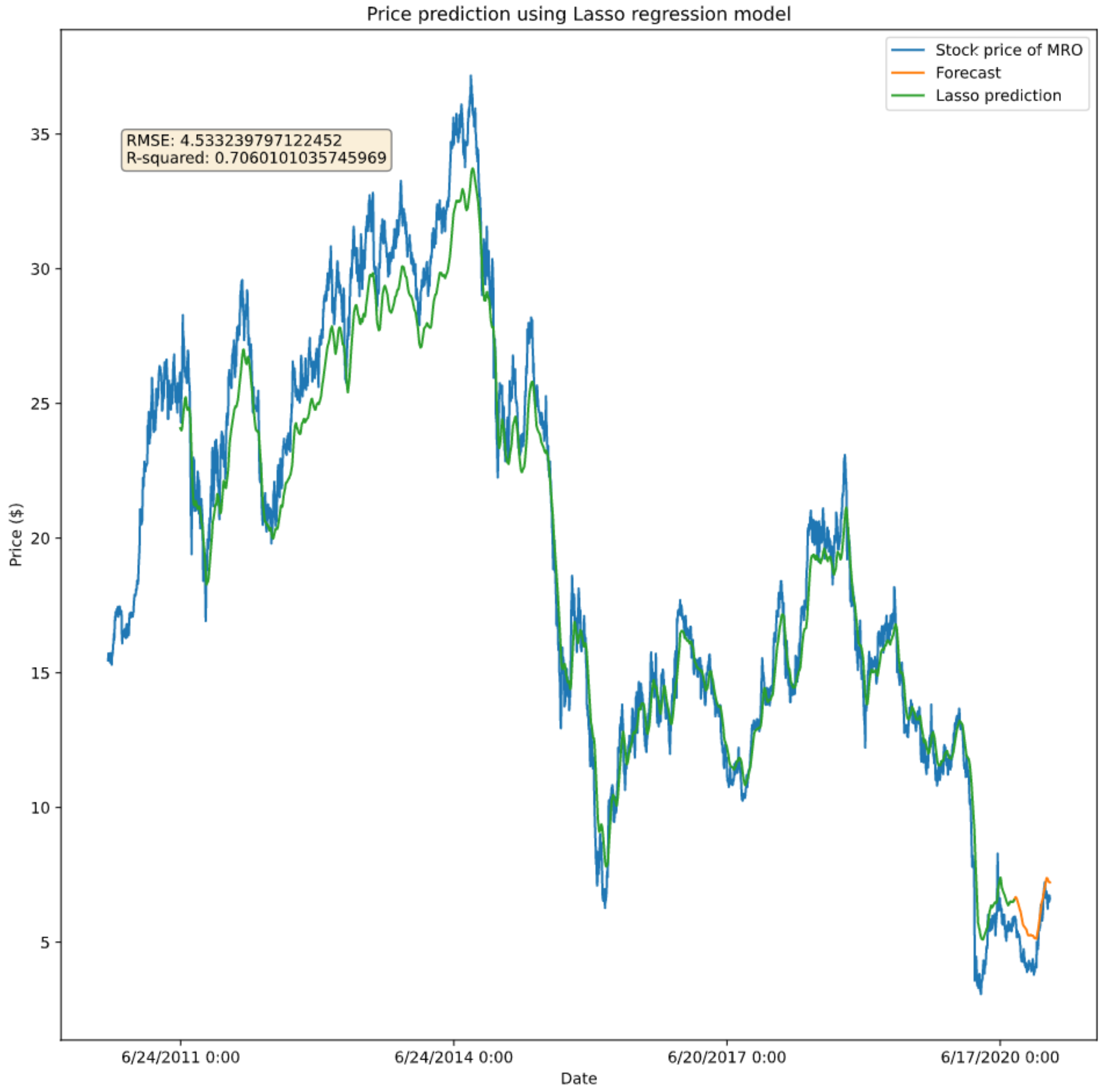


Figure 13: 100 day forecast using lasso regression stock price predictions

Furthermore, 5 day (one week), 20 day (one month), and 260 day (one year) moving return values were predicted using a separate lasso regression model. Figures 14-16 display the models predicting 5 day, 20 day, and 260 moving return values of Google's stock compared to the actual moving averages. The same features were used in this model as in the ridge regression model.

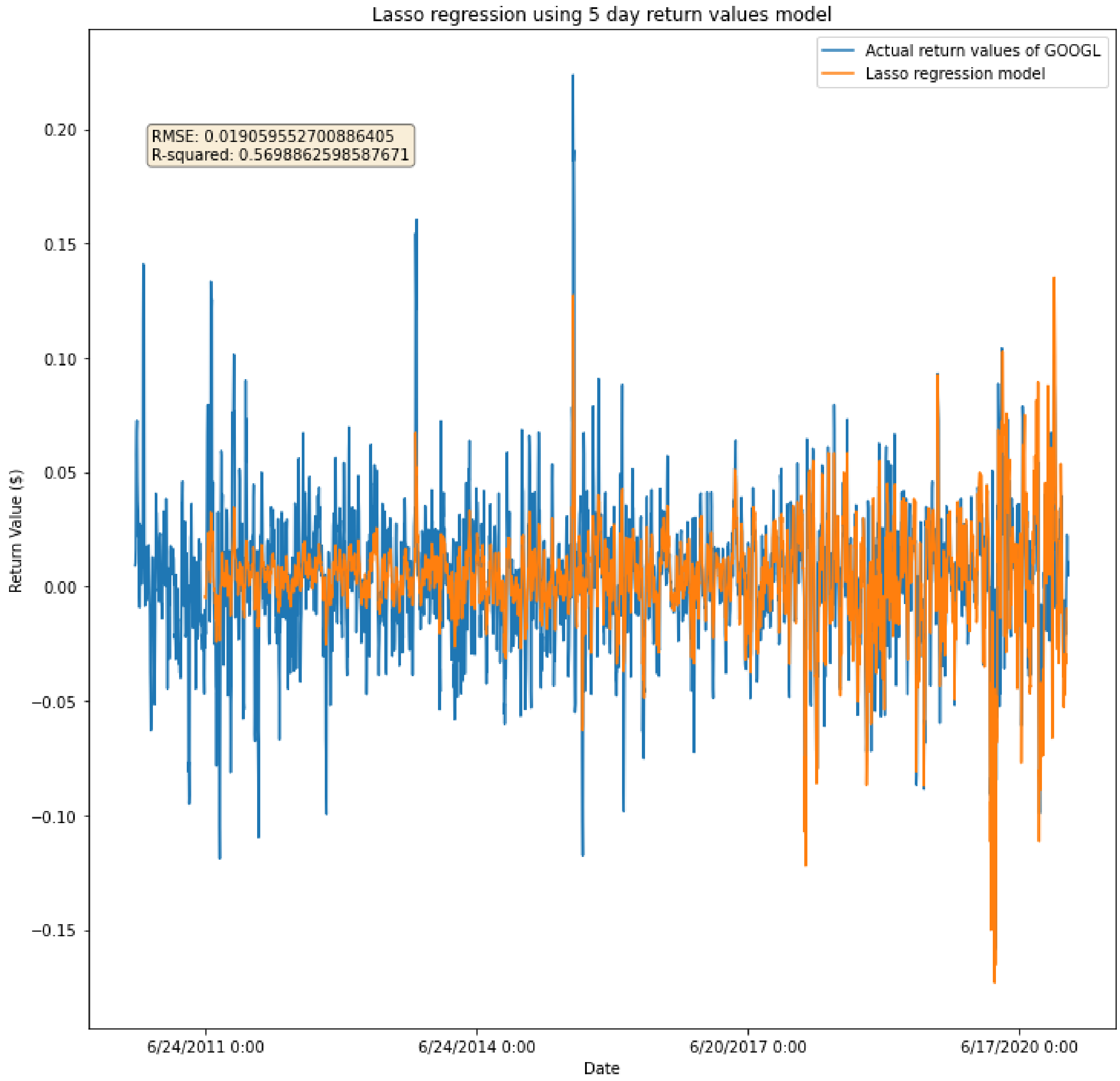


Figure 14: 5 day moving average predictions of GOOGL using lasso regression vs. actual 5 day return values

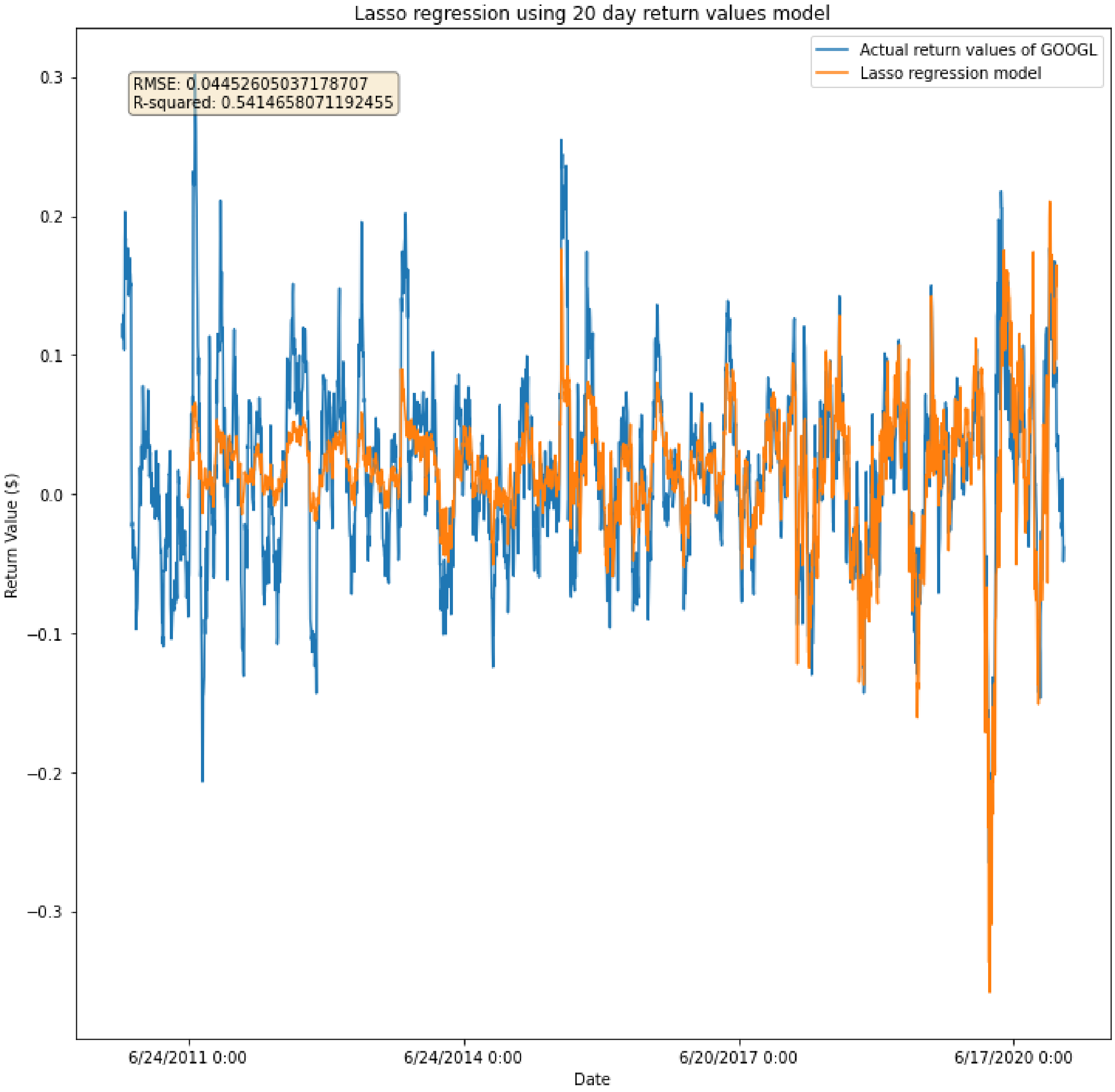


Figure 15: 20 day moving average predictions of GOOGL using lasso regression vs. actual 20 day return values

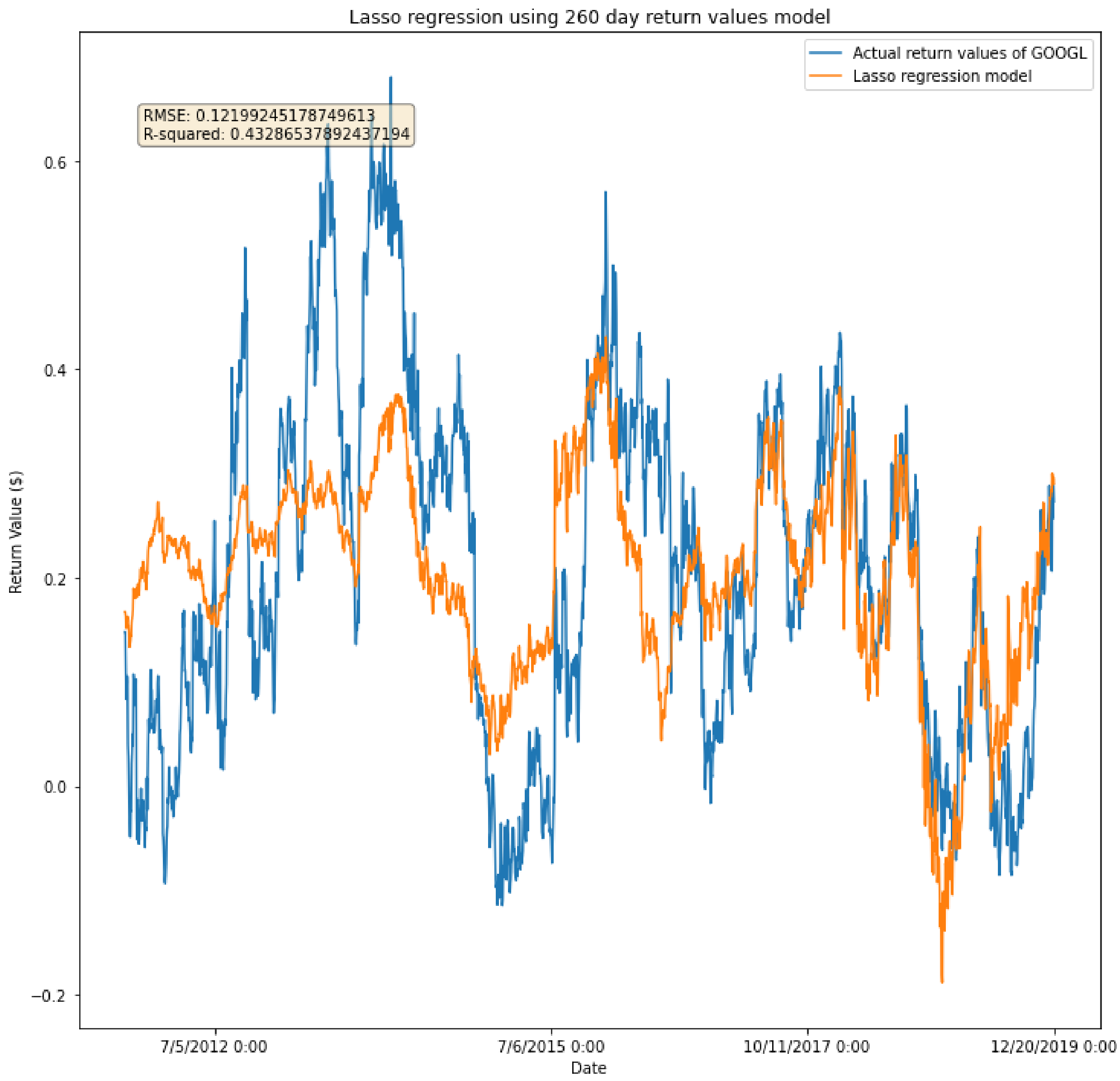


Figure 16: 260 day moving average predictions of GOOGL using lasso regression vs. actual 260 day return values

Tables 3 and 4 compare the r-squared and the RMSE values of the lasso regression models using 5 day, 20 day, and 260 day moving return values of GOOGL and MRO (not displayed above).

Moving day return value	**RMSE value**	**R-squared value**
5 day	0.01905955	0.56988626

20 day	0.044526050	0.541465807
260 day	0.121992451	0.43286538

Table 3: RMSE and R-squared values of different moving return predictions of GOOGL stock using lasso regression

Moving day return value	**RMSE value**	**R-squared value**
5 day	0.030613120	0.617875485
20 day	0.09060504	0.5740656
260 day	0.256229780	0.55424029

Table 4: RMSE and R-squared values of different moving return predictions of MRO stock using lasso regression

Just like with ridge regression, notice how predicting one year ahead of time is much less accurate than trying to predict one week or one month ahead of time. This is shown through the RMSE values being much higher as well as the r-squared values being much lower in the 260 day return predictions compared to the 5 day and 20 day return value predictions. This also follows our prediction that lasso regression is more accurate for short term prediction compared to longer term prediction.

Once again, these models can be used to predict future return values outside of our dataset. Figures 17-19 display future return value predictions of Google's stock.

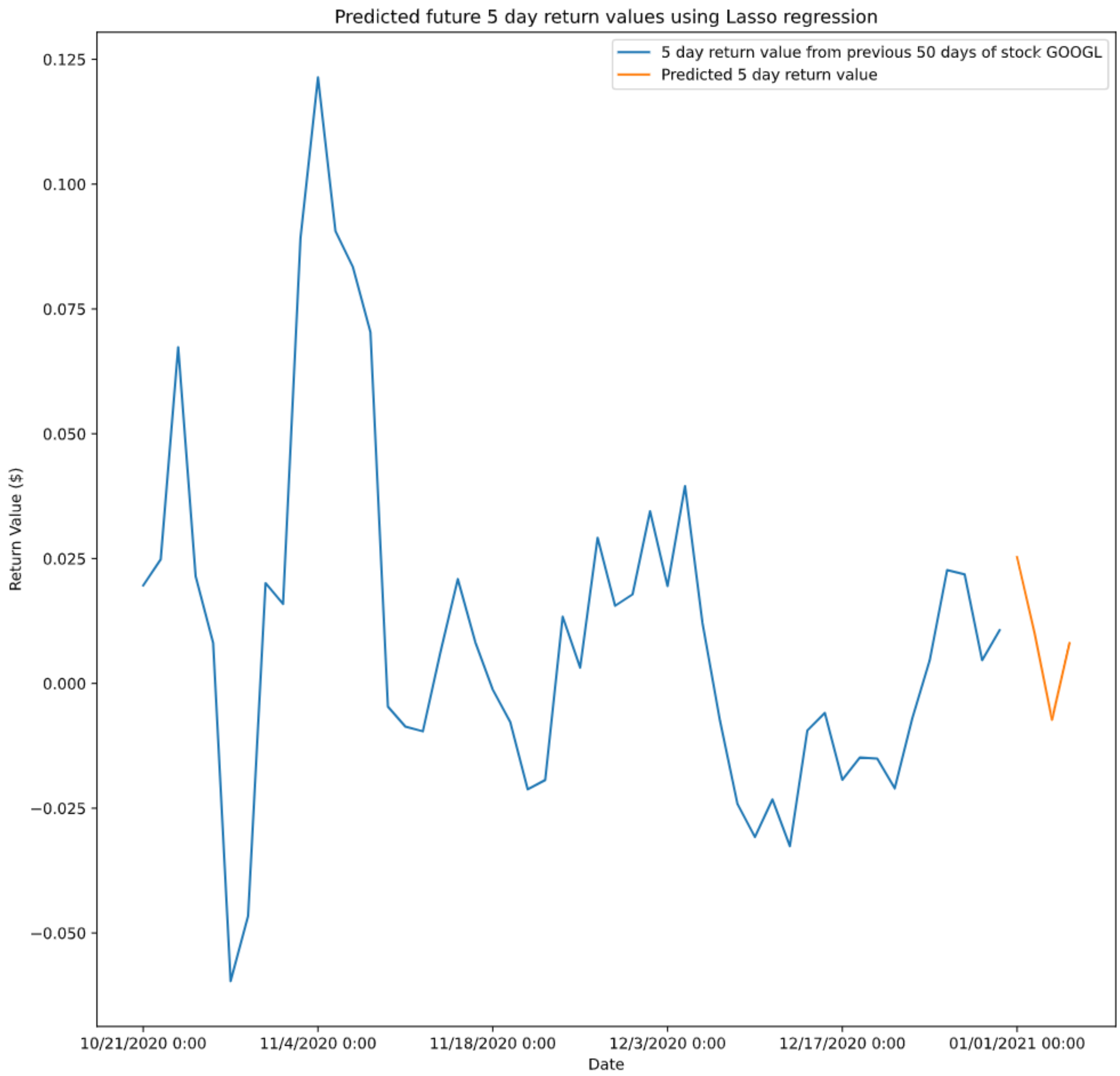


Figure 17: Predicted future 5 day return values of GOOGL using lasso regression

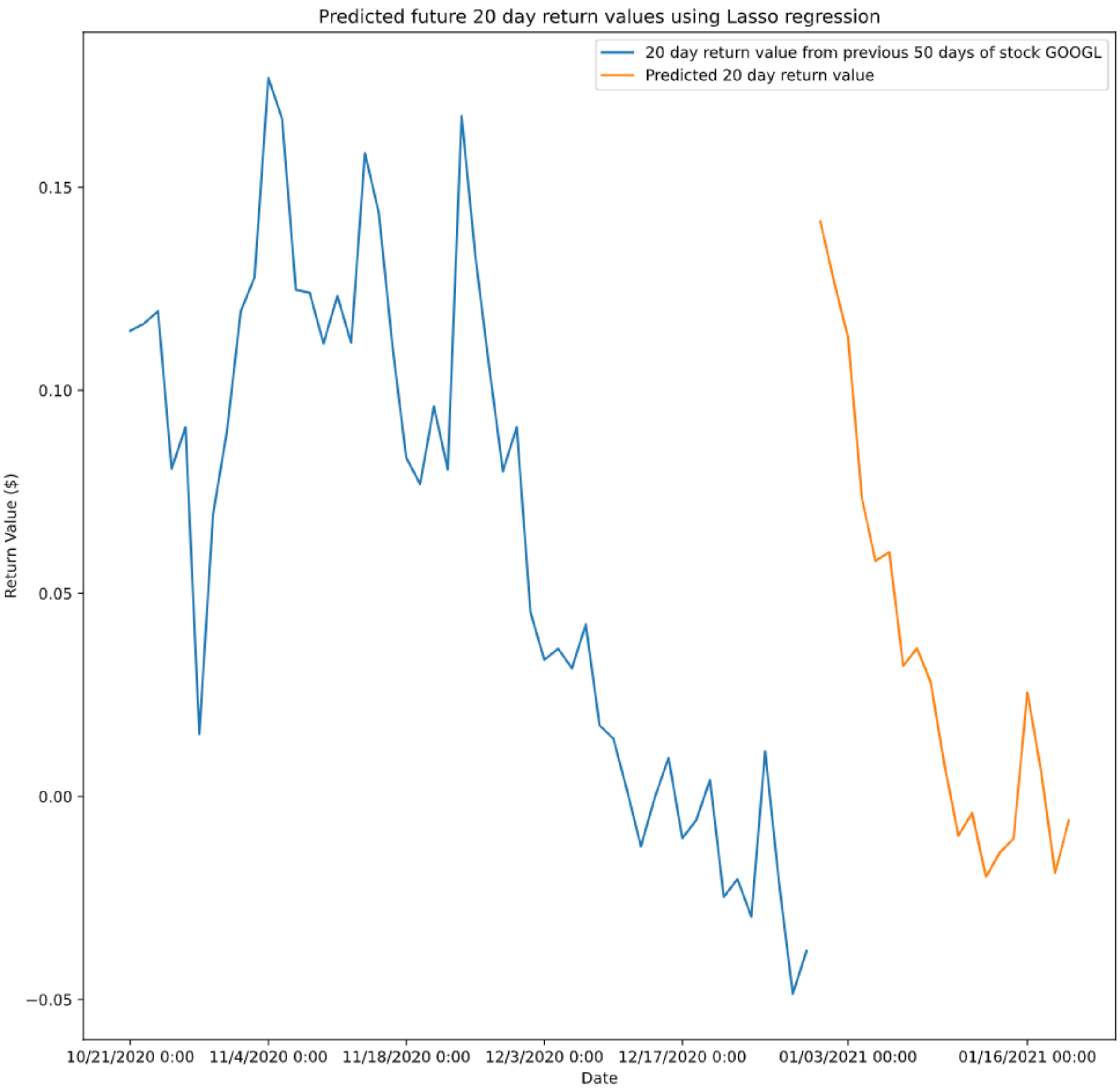


Figure 18: Predicted future 20 day return values of GOOGL using lasso regression

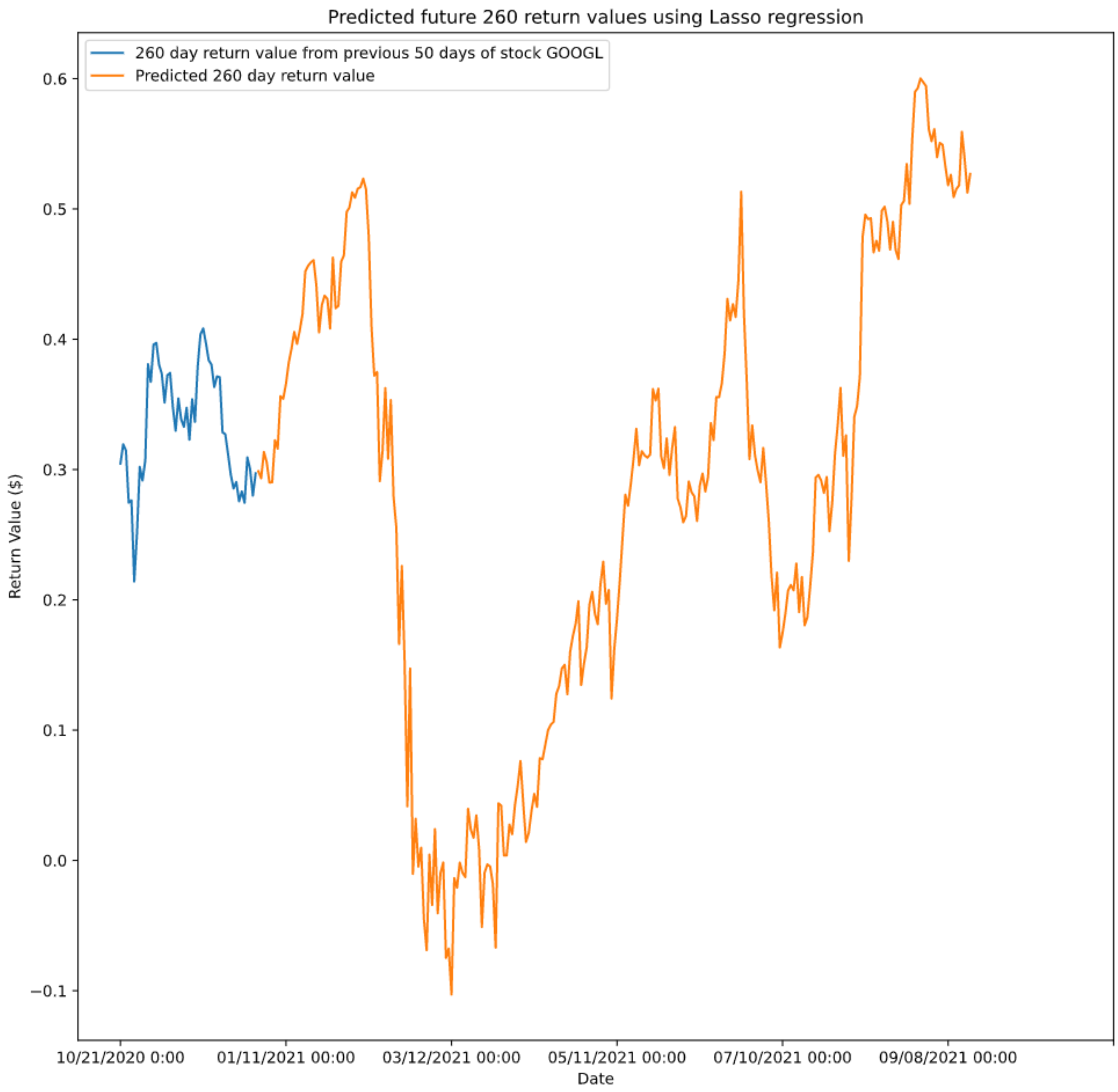
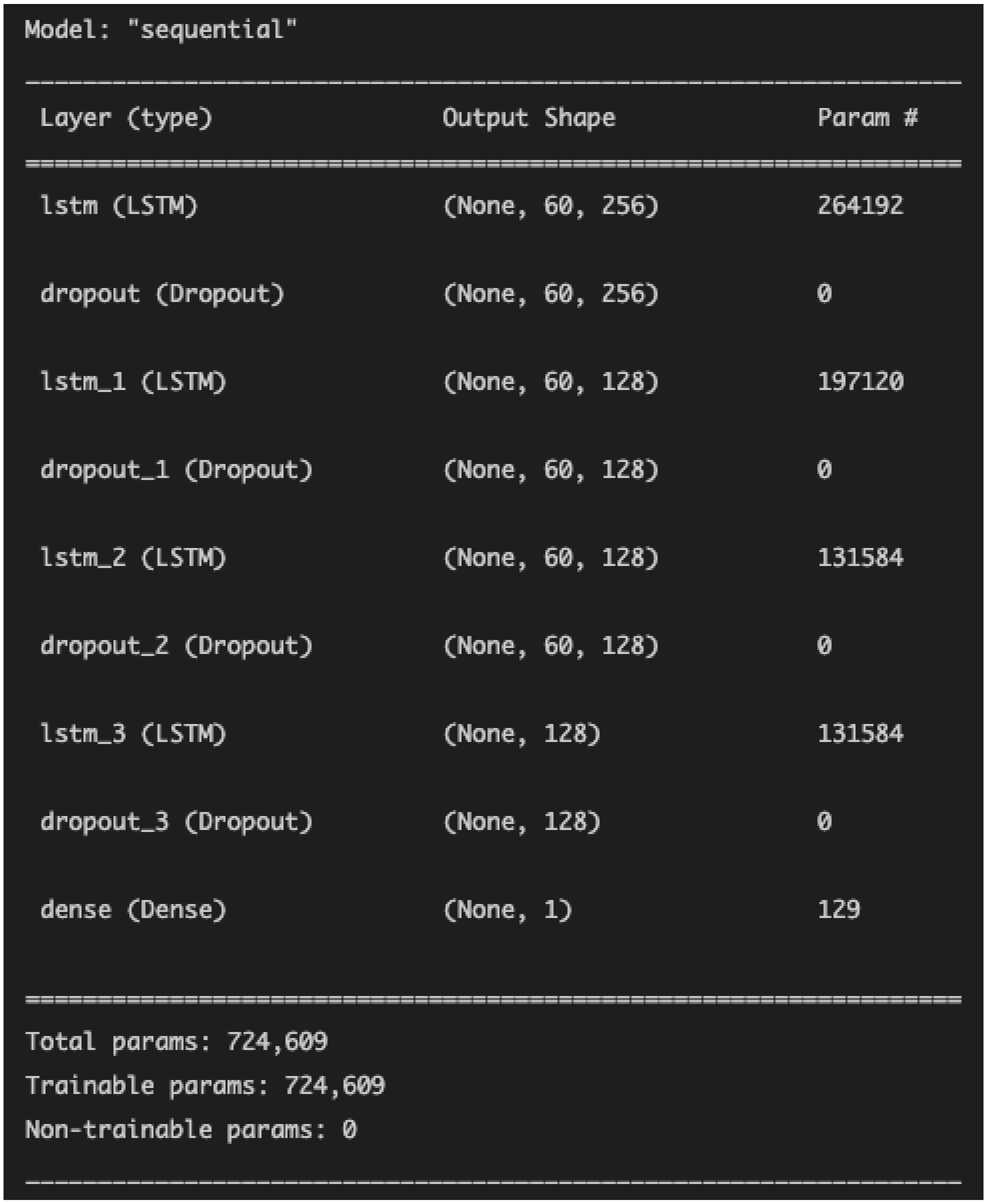


Figure 19: Predicted future 260 day return values of GOOGL using lasso regression

LSTM Neural Net

A long short-term memory (LSTM) neural net was created to model stock prices over time. Each data point was a time stamp of the moving averages from the previous 60 days. The averages from the previous 60 days were used to predict the current day. For example, the neural net model using 5 day moving averages predicted the stock's price on the ith day by looking at the 5 day moving averages from the previous 60 days. The size of each time step (60) can be seen in the summary of the neural net displayed below in Figure 20.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 60, 256)           264192

 dropout (Dropout)           (None, 60, 256)           0

 lstm_1 (LSTM)               (None, 60, 128)           197120

 dropout_1 (Dropout)         (None, 60, 128)           0

 lstm_2 (LSTM)               (None, 60, 128)           131584

 dropout_2 (Dropout)         (None, 60, 128)           0

 lstm_3 (LSTM)               (None, 128)               131584

 dropout_3 (Dropout)         (None, 128)               0

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 724,609
Trainable params: 724,609
Non-trainable params: 0
_________________________________________________________________
```

Figure 20: Summary of the LSTM neural network

Increasing the number of neurons led to more accurate results. However, continuing to increase the number of neurons at each layer greatly increased how long it took to train the data. Testing with different numbers of neurons, it was decided that the first layer should have 256 neurons, while the others have 128. These values led to accurate results while also having a reasonable runtime. The neural net models created to predict Google's stock price using the 60 day timestamps of 10 day moving averages (Figure 21), 50 day moving averages (Figure 22), and 200 day moving averages (Figure 23) are displayed below.

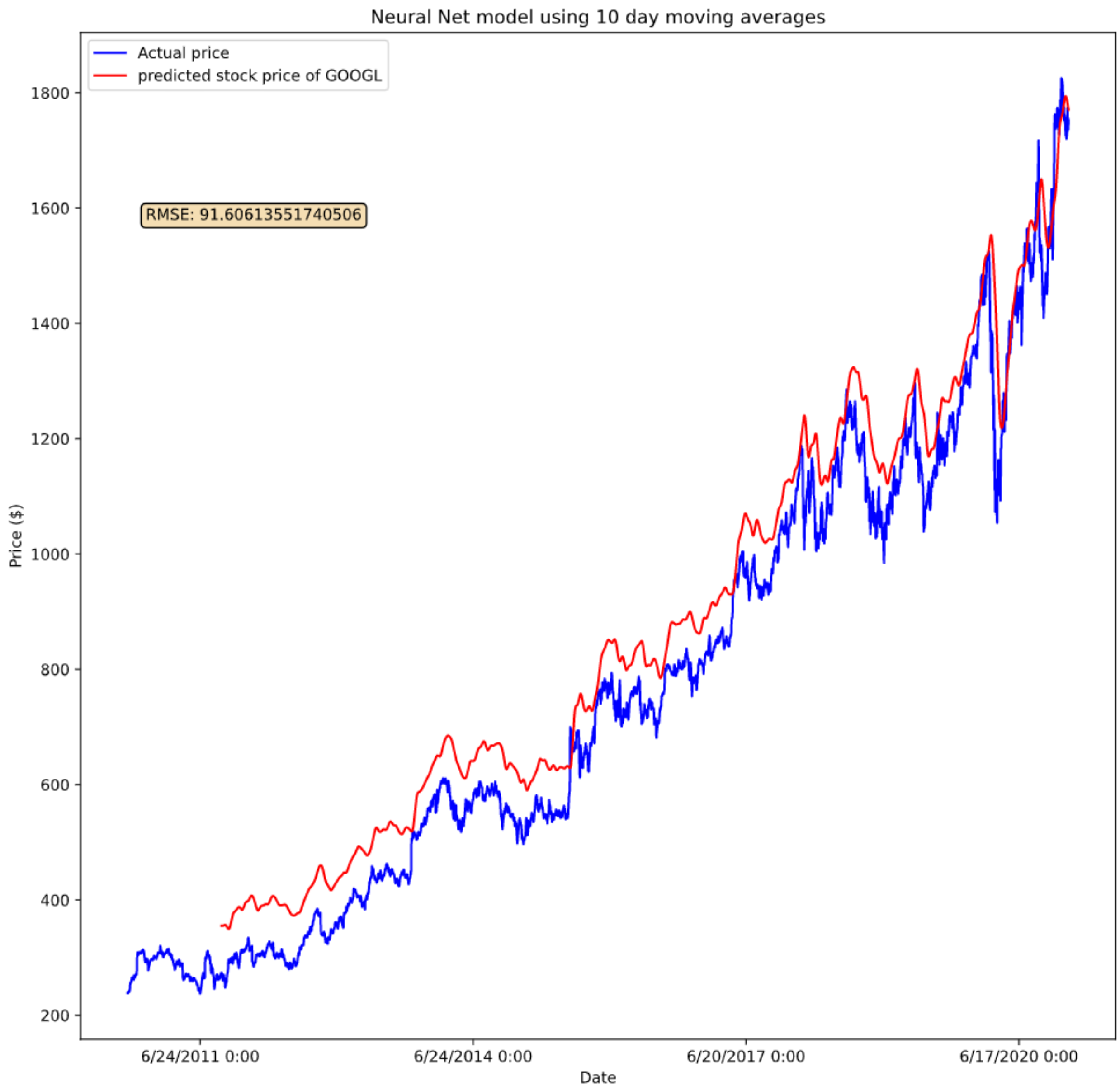


Figure 21: Stock price prediction neural net model using 10 day moving averages.

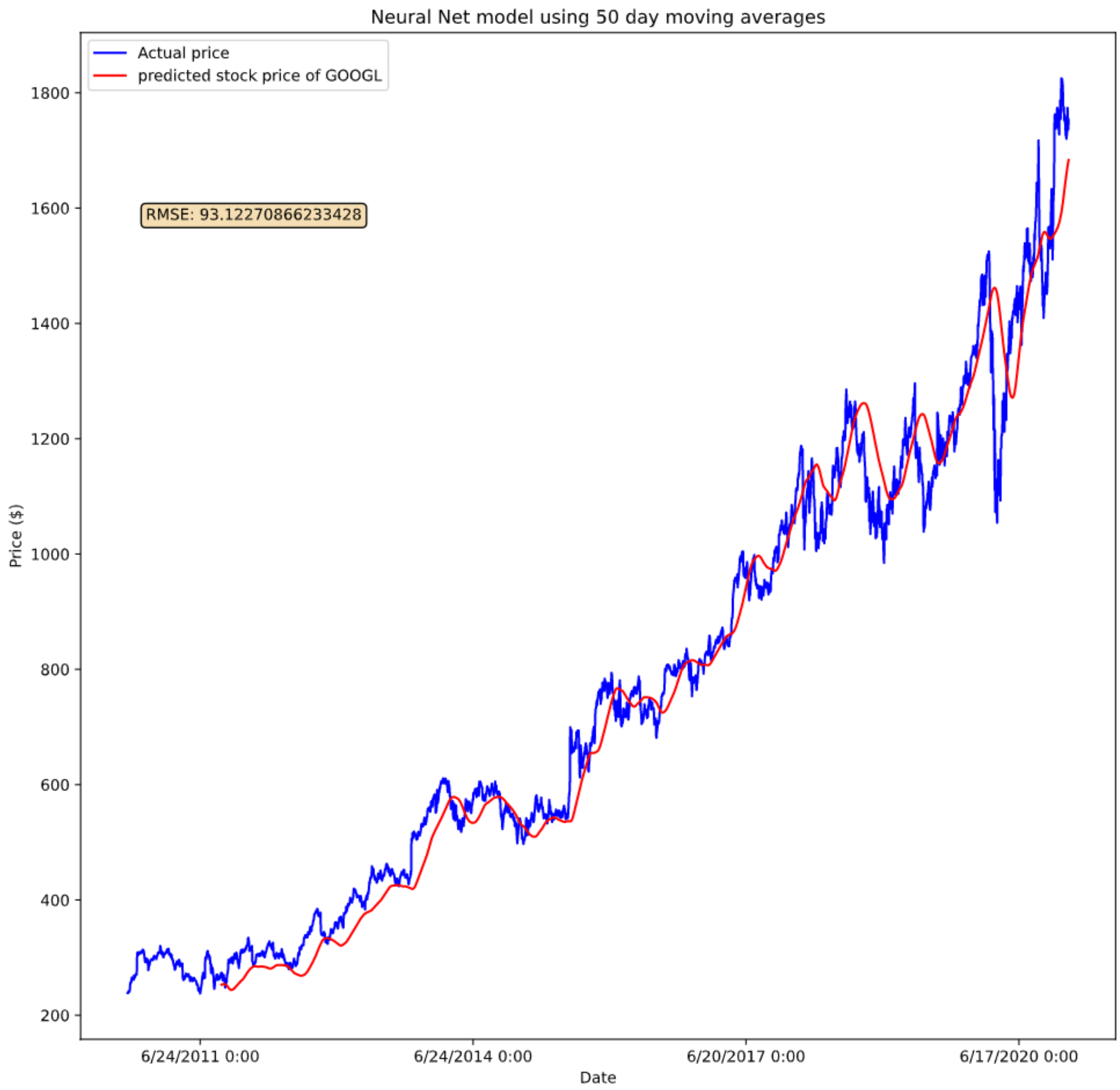


Figure 22: Stock price prediction neural net model using 50 day moving averages.

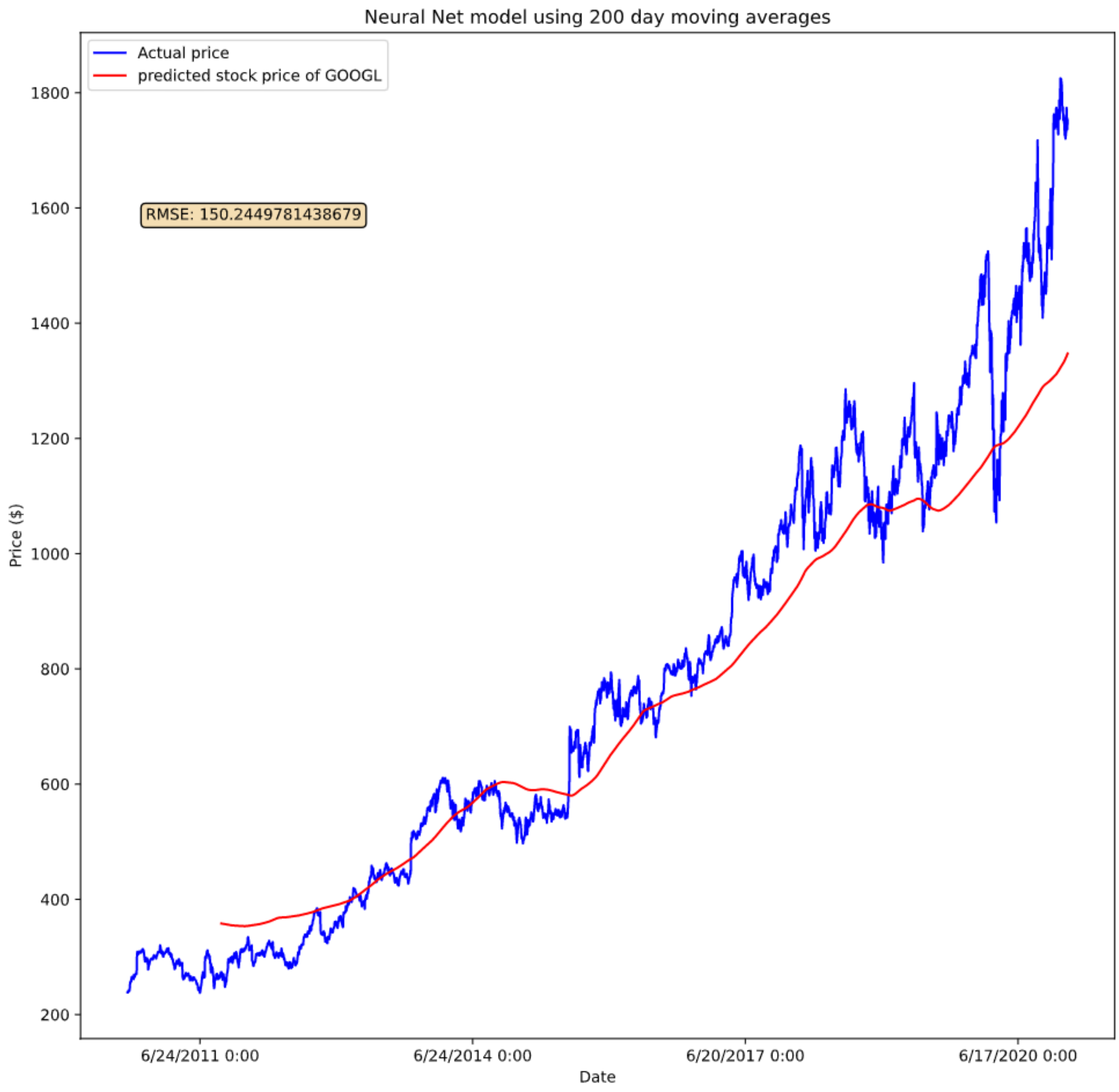


Figure 23: Stock price prediction neural net model using 200 day moving averages.

Tables 5 and 6 below compare the RMSE values of the neural net models using 10 day, 50 day, and 200 day moving average values of GOOGL and MRO (not displayed above).

Moving average value (the feature used)	**RMSE value**
10 day	91.60613555
50 day	93.1227086

200 day	150.24497814

Table 5: RMSE values of LSTM neural net models using different moving averages as features to predict GOOGL stock price

Moving day return value (the feature used)	RMSE value
10 day	0.827341418
50 day	0.26157376
200 day	1.96069617

Table 6: RMSE values of LSTM neural net models using different moving averages as features to predict MRO stock price

Clearly, the models using the 200 day moving price averages performed the worst. The models using the 10 day and 50 day moving price averages performed about the same.

Comparing models

Tables 7 and 8 below depicts the RMSE values and R-squared values (if applicable) of the different models described above that were used to model stock prices of GOOGL and MRO, respectively.

Model	**RMSE value**	**R-squared value**
Ridge Regression	69.010904217	0.970958556
Lasso Regression	82.2146226	0.951645504
LSTM Neural Net using 10 day moving price averages	91.60613555	NA
LSTM Neural Net using 50 day moving price averages	93.1227086	NA
LSTM Neural Net using 200 day moving price averages	150.24497814	NA

Table 7: RMSE and R-squared values of different models used to predict GOOGL stock price

Model	**RMSE value**	**R-squared value**
Ridge Regression	4.518773588	0.68083489

Lasso Regression	4.5332398	0.7060101
LSTM Neural Net using 10 day moving price averages	0.827341418	NA
LSTM Neural Net using 50 day moving price averages	0.261573764	NA
LSTM Neural Net using 200 day moving price averages	1.96069617	NA

Table 8: RMSE and R-squared values of different models used to predict MRO stock price

The tables suggest that ridge regression performed best when predicting the prices of GOOGL, and the LSTM neural net using 50 day moving price averages performed best on MRO. Looking at the actual stock prices of GOOGL (shown in Figure 2) and MRO (shown in Figure 4), GOOGL is much more linear and has a steady increase in price compared to MRO, which jumps up and down. The linearity of the GOOGL stock price is indicative of why ridge regression, a linear model, best predicted GOOGL stock price. Furthermore, the LSTM model using 50 day moving price averages best predicted the varying stock prices of MRO because the model looked at data in 60 day chunks, rather than looking at each day by itself, to model the prices of MRO stock.

Furthermore, we were able to use ridge regression and lasso regression to predict return values. Tables 9 and 10 compare the r-squared and the RMSE values of the ridge regression models and the lasso regression models using 5 day, 20 day, and 260 day moving return values of GOOGL and MRO (not displayed above).

Model/what is being predicted	**RMSE value**	**R-squared value**
ridge/5 day return	0.0202955	0.59544785
lasso/5 day return	0.01905955	0.56988626
ridge/20 day return	0.03929537	0.5829511
lasso/20 day return	0.044526050	0.541465807
ridge/260 day return	0.12431293	0.4715617
lasso/260 day return	0.121992451	0.43286538

Table 9: RMSE and R-squared values of different moving return predictions of GOOGL stock using ridge regression and lasso regression

Model/what is being predicted	**RMSE value**	**R-squared value**
ridge/5 day return	0.03985127	0.5938015
lasso/5 day return	0.030613120	0.617875485
ridge/20 day return	0.09171964	0.6481888
lasso/20 day return	0.09060504	0.5740656
ridge/260 day return	0.253353404	0.49152697
lasso/260 day return	0.256229780	0.55424029

Table 10: RMSE and R-squared values of different moving return predictions of MRO stock using ridge regression and lasso regression

For both models, the further into the future the model tried to predict, the worse it performed. Both models performed about the same, as the RMSE values do not differ by more than 0.01 when the models are predicting the same thing.

To conclude, the further a model tries to predict stock return values into the future, the worse it performs. Furthermore, ridge regression is most useful when modeling stock prices that have followed a somewhat linear path. A neural net performs best on modeling stock prices that fluctuate heavily.