Predicting Stock Prices of Apple, Tesla, and Amazon using Linear Regression Models

Ivan Patel

December 26, 2020

1 Introduction

In this course project, I build three linear regression models to predict closing stock prices of Apple, Amazon, and Tesla. Although professional data scientists with a strong knowledge of financial markets can build a more complex model, I believe that a linear regression model does a surprisingly great job of predicting future stock prices.

I chose to analyze Apple, Amazon, and Tesla due to their stock's popularity. I want to emphasize that the model results are not meant to be an investment advise. Rather, I did this project because I was curious about how reliable is linear regression when making stock predictions.

2 Main objective of my analysis

The main objective of my analysis is to make predictions. It will be difficult to interpret a model's coefficients because the predictors and responses were normalized using MinMaxScaler() function from the Scikit-learn library.

I build three models for all three companies. For all models, the response variable is a given stock's next day closing price. The three models and their respective predictors are summarized in Table 1.

Table 1: Model's and predictors

Model Name	Predictors
Simple Linear Regression	Previous day's closing price
Multiple Linear Regression	Previous day's closing price + previous day's volume
Ridge Regression (alpha = 2)	Previous day's closing price + previous day's volume

Although the ridge regression's alpha value does seem high at first, this choice is intentional and will highlight where and why do the predictions of ridge regression fall short. In other words, an alpha of 2 will prove without a reasonable doubt why the multiple linear regression model, in this case of predicting stock prices, is the best choice.

3 Brief description of the data set and a summary of its attributes

Jason Strimpel's bulk stock data series download web system allows anyone to easily download daily stock prices into a csv file.[1] You can choose any companies you like by manually entering their tickers and selecting the variables.

I downloaded Apple, Tesla, and Amazon's closing price into a one csv file and those company's day volume into an another csv file. My data ranges from January 3, 2011 to December 24, 2020, and neither data sets have any missing observations. Figure 1 and Figure 2 summarize both data sets.

[1] http://finance.jasonstrimpel.com/bulk-stock-download/. This is a very useful tool because it is free. Please consider donating to the website to keep it alive

	AAPL	AMZN	TSLA
count	2513.000000	2513.000000	2513.000000
mean	34.375212	927.728305	62.764073
std	24.721960	817.809814	91.809640
min	9.714680	160.970001	4.366000
25%	17.348335	282.100006	23.436001
50%	26.253185	587.000000	45.344002
75%	42.640545	1601.859985	61.669998
max	133.948898	3531.449951	695.000000

Figure 1: Stock Closing Prices Summary

	AAPL	AMZN	TSLA
count	2.513000e+03	2.513000e+03	2.513000e+03
mean	2.550876e+08	4.293271e+06	3.164611e+07
std	2.007307e+08	2.380718e+06	2.856358e+07
min	2.420510e+07	8.813000e+05	1.198000e+06
25%	1.141584e+08	2.767900e+06	1.269600e+07
50%	1.841928e+08	3.688000e+06	2.489000e+07
75%	3.345692e+08	5.044600e+06	4.027350e+07
max	1.880998e+09	2.413420e+07	3.046940e+08

Figure 2: Stock Day Volume Summary

Notice that stock prices for Tesla and Amazon have the highest standard deviations. In the case of stock prices, standard deviations measure how widely prices are dispersed from the average price. In other words, a stock's standard deviation is that stock's measure of volatility.[2]

Daily trading volume is simply the total amount of shares traded for the day. Out of all the stocks in our data, Tesla has the largest average trading volume. When more money is moving a stock price, it means there is more demand for that stock. Thus a stock that's appreciating on high volume is more likely to be a sustainable move.[3]

4 Brief summary of data exploration and actions taken for data cleaning and feature engineering

4.1 Basic Approach

For a given company, I used the first nine years of closing prices and volume to train its three models. Therefore, the model predicts the closing prices of the year 2020 given the previous day's closing price and the previous day's volume.

I do not use cross validation cross validation because it shuffles the data and before training and testing the model on k folds k times. Instead, I use the closing prices of years 2011-2019 to train the models. Finally, there is no intuitive reason to include polynomial features because the linear regression models, as I show below, do an excellent job predicting closing prices in 2020. Although one can try a fitting a polynomial model between volume and closing prices, I would not recommend this approach because doing so will lead to a complex model at the expense of test accuracy.

4.2 Stock Price Visual

Say you could go back to 2011-01-03 with $10,000 and the stocks data analyzed in this project. You know you could make a lot of money in the next 10 years because you know the prices. But which stock should you buy?

Like many investors, you want to maximize your profit which is done buy buying the stock with the highest return. This is where the idea of normalization is useful. The normalized stock price graphs the price movement of an stock using 100 as the base value of one. For instance, if you bought a Tesla's share worth $1 on 01/03/2011, that share would've been worth more that $120 on 12/8/2020.

[2]https://www.investopedia.com/ask/answers/021015/what-best-measure-given-stocks-volatility.asp. Standard deviation is the most common way to measure market volatility

[3]https://www.investopedia.com/articles/investing/060315/stocks-trade-volume-important.asp

Figure 3 plots the normalized stock prices of all three companies. It shows that investing in Tesla would be the most lucrative investment rather than Apple or Amazon despite Tesla's volatility and the uncertainties investors had about the company.

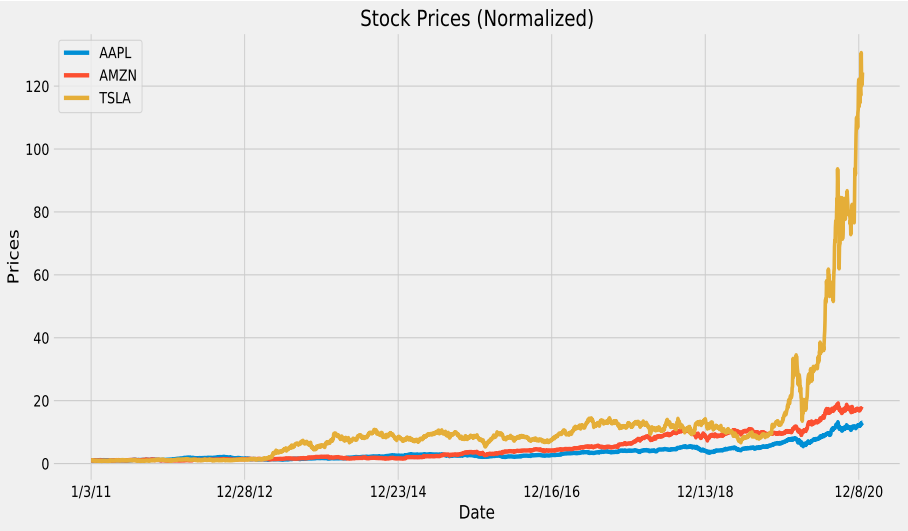


Figure 3: Stock Prices Normalized

5 Summary of three linear regression models

Table 2 below shows how accurate were a model's 2020 closing price predictions for a given company. Rounding validation scores up to three digits reveals that the performance of the the multiple linear regression and simple linear regression is identical. This reveals that the previous day's closing price is the biggest indicator of what is next day's closing price.

Table 2: Model's and R^2 on the Test Set

Model Name	Apple	Amazon	Tesla
Simple	0.995	0.993	0.995
Multiple	0.995	0.993	0.995
Ridge (alpha = 2)	0.822	0.935	0.552

Table 2 also shows that compared to the first two models, ridge regression with an alpha value of two severely under performs on the test set. This makes intuitive sense because the linear regression models are doing an excellent job of predicting closing prices, and restricting the coefficients is hurting the model's predictive ability. The figures in the Appendix compare the performance of multiple linear regression and ridge regression.

6 Recommendations

Although I don't show it here, multiple linear regression performs slightly better than the simple linear regression. Because of this slight edge, I would recommend that model. However, the difference in performance between the simple and multiple model is next to nothing. Thus, picking the simple model is also a very good choice.

7 Suggestions for next steps in analyzing this data

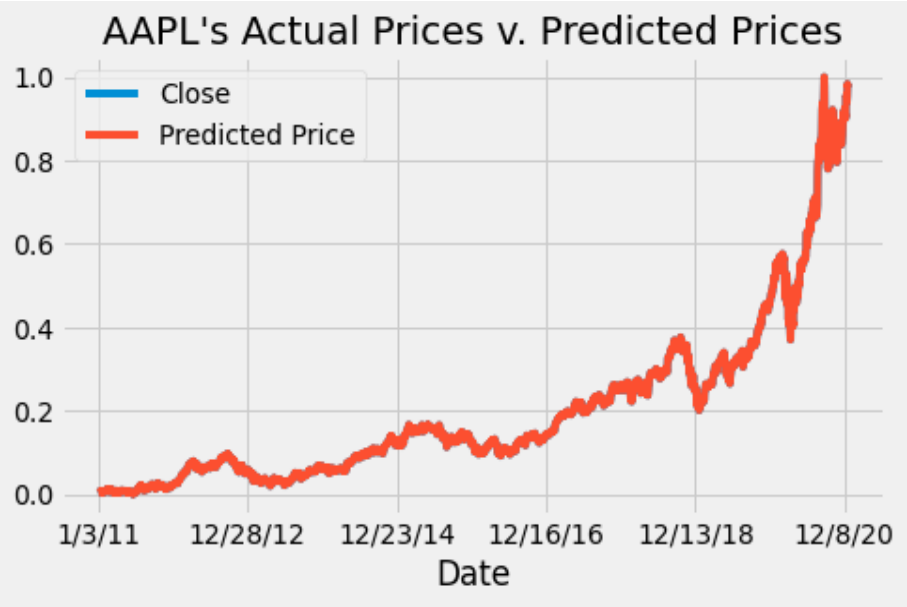
I would suggest using the same models to predict stock prices of different companies. Although I analyzed three companies, those companies are giants in the software industry. Training and testing the model on companies from

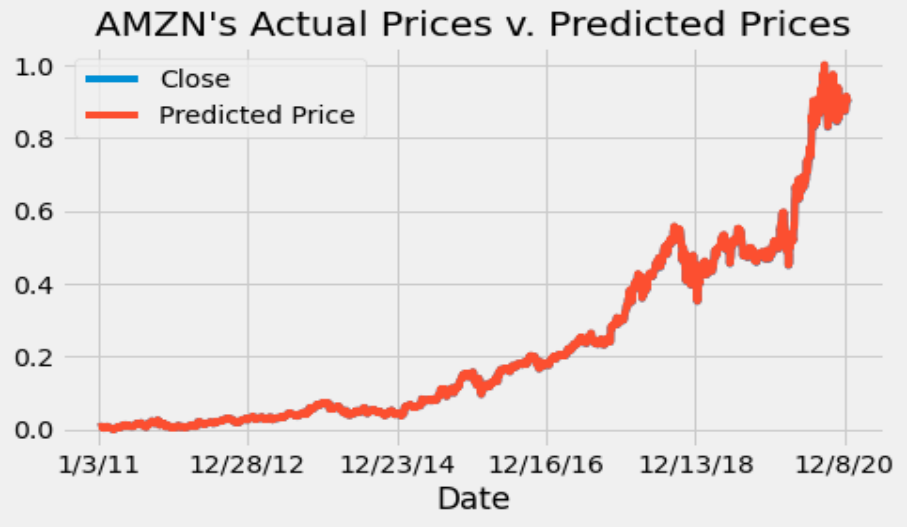
other industries such as hospitality, or entertainment could be worthwhile for people interested in investing in those sectors.
One can further analyze this data by using a deep learning model such as a neural network to predict future prices. If that model's predictions have a lower RMSE than the linear models, I would suggest using the neural network.

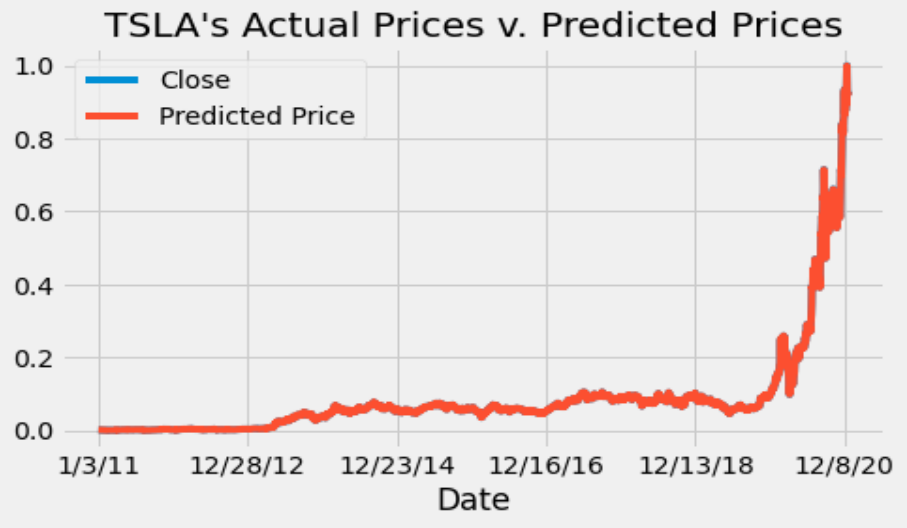
8 Appendix

8.1 Multiple Linear Regression Results

The reason why you cannot see the closing prices is because the predictions curve is on top of the closing price curve. This proves that model is doing an excellent job of predicting closing prices.

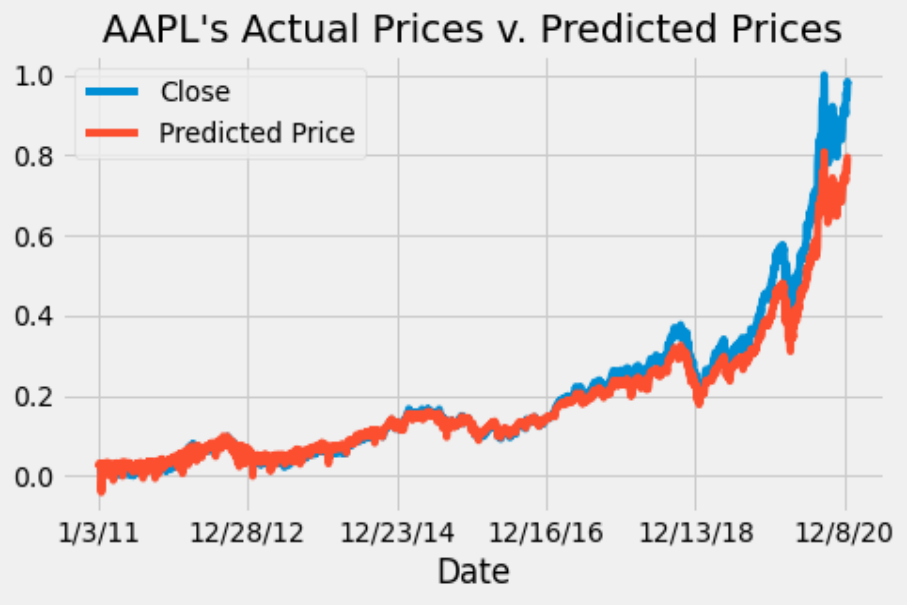


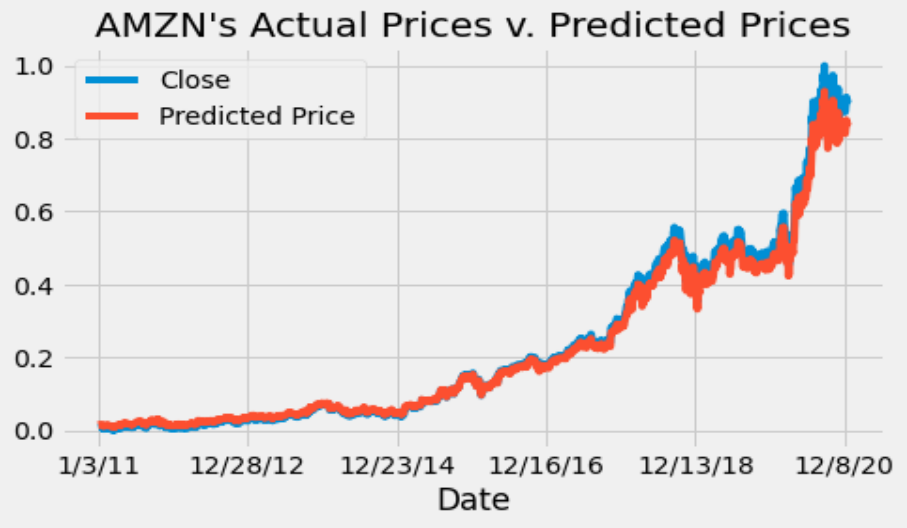


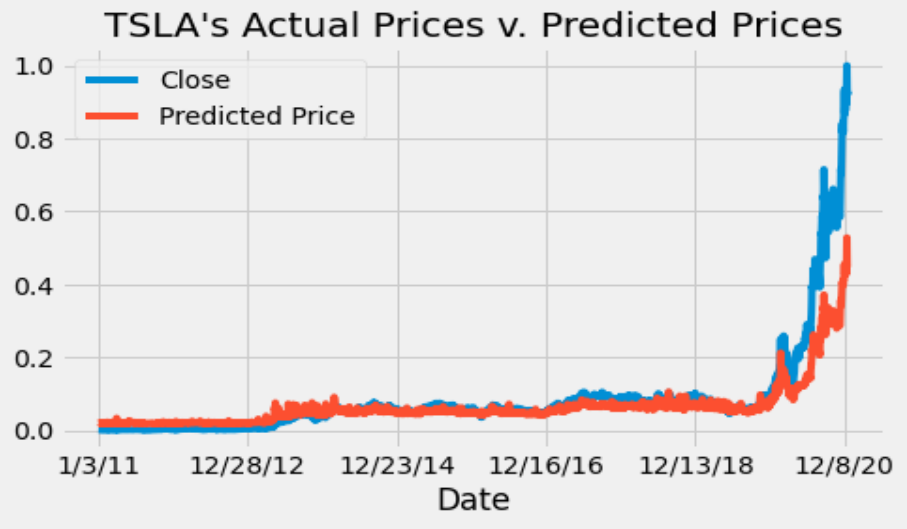


8.2 Ridge Regression Results

Notice that in figures for Tesla and Apple, 2020's predictions are much lower than the actual prices. This is a sign of under fiting.




AMZN's Actual Prices v. Predicted Prices
Close
Predicted Price
1.0
0.8
0.6
0.4
0.2
0.0
1/3/11
12/28/12
12/23/14
12/16/16
12/13/18
12/8/20
Date


TSLA's Actual Prices v. Predicted Prices
Close
Predicted Price
1.0
0.8
0.6
0.4
0.2
0.0
1/3/11
12/28/12
12/23/14
12/16/16
12/13/18
12/8/20
Date