Predicting Stocks of Tesla and NIO

Nexus Feng

1. Abstract

This project uses model building and statistics to predict the performance of electric vehicle stocks Tesla and NIO. The project utilizes data collection, statistical analysis, and model building to predict stock prices traded during 3/16/2020 - 9/16/2020. Using independent variables such as the High, Low, Volume of a stock (Common Stock predictors), the project first utilizes simple linear regression models to directly compare one independent variable with the individual dependent variable. Then, a multivariable linear regression model is used to predict the dependent variable: The Next Day Open price of a stock. By analyzing the coefficients, p-values, and R^2 value, this paper can determine the reliability of these models. Next, a logistic model will be used to predict if the stock will either rise or fall on any given next day by predicting the probability if the stock will go up or down. After both models are tested, a percentage error will be generated to determine which is more accurate. In conclusion, the multivariable linear regression model is the more accurate model with an average of 4% error while the logistic regression has an average of 41% error. The multivariable linear regression is chosen because of the displaying of actual values and the preciseness of the model.

Contents

2. Introduction

Have you ever wanted to make money you are lying in bed, watching TV, or playing games? Then you are looking for a passive income. Passive income is money you earn in a way that requires little to no daily effort to maintain. The most popular and effective method for people to earn a passive income is investing in the stock market.

Stocks are a type of security that give stockholders a share of ownership in a company. If the stock rises in price, stockholders earn capital appreciation. You must be wondering, what are the downsides to such an easy investment? Just like the stock can rise in price, it can drop in price as well. There is no guarantee that the company whose stocks you hold will grow in value, meaning you can lose the money you invest in stocks if you do not know which stock to buy. However, if you were able to predict the performance of stocks through some sort of mathematical model, your prediction will allow one to make more informed decisions.

The invention of electric vehicles has shaken the world. These cars offered all the benefits that gasoline-powered cars do not – no issues with gasoline, quiet, easy to drive, and no emission of pollutants. The electric vehicle has made great improvements in the last 20 years from the basic, mass-produced hybrid of 1997, the Toyota Prius, to the modern-day Tesla that can run 300 miles on a single charge.

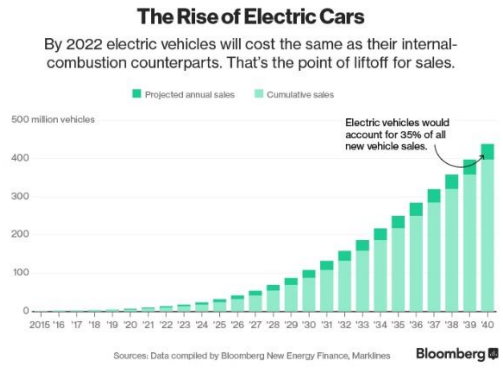


Figure 1 The Rise of Electric Cars (Bloomberg)

As seen in Figure 1, the production of electric cars will continue to rise in the future (Fig. 1). This multi-billion-dollar industry attracts investors and has created competition between automotive companies to earn the most revenue. The stock market reflects this future production increase.

Whenever a new idea is formed and demand rises, there will naturally be competition. It is no different in the electric vehicle industry. To compete with Tesla, electric automobile manufacturer NIO Inc. emerged. Although not as well-known as Tesla, NIO Inc. has emerged as a serious competitor in the industry. In September 2018, NIO Inc. filed for a $1.8 billion initial public offering on the New York Stock Exchange and later the Nasdaq.

This paper intends to compare the prediction of the stocks of Tesla and NIO on the Nasdaq through Simple Linear Regression, Multivariable Linear Regression, and Logistic

Regression to determine which regression model is the best and most reliable prediction method.

3. Research Background

The act of stock market prediction is the act of trying to determine the future value of a company stock (Wiki). The successful prediction of a stock's future price can yield opportunities for investment. Some people think that stocks are inherently unpredictable because of the efficient-market hypothesis, which suggests that stock prices reflect all current information and any price changes (Investopedia). However, others disagree as they believe there are mathematical, statistical, or technological ways to predict the market. There are mainly three types of analysis: fundamental analysis, technical analysis (charting), and technological methods such as machine learning.

The analysis this paper intends to use technical analysis to determine future stock price. This paper intends to use the previous day capitalization prices of a stock to analyze the stock open price of the next day. Some basic assumptions for this analysis include 1) stock price shows everything important about the company, 2) stock prices have trends, and 3) trends repeat itself.

This paper intends to use three types of regression models. First, this paper will use a simple linear regression model. A linear regression model is an approach to model the relationship between an independent variable and a dependent variable. The formula for a Simple Linear Regression model is shown below (Fig. 2).

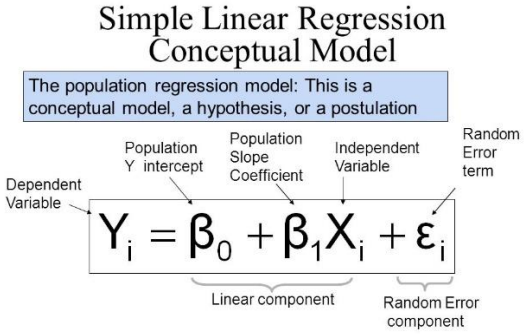


Figure 2 Simple Linear Regression Model (Luis Fok)

Using the independent variables (High, Low, Volume traded) and the dependent variable (Stock Open Price), I will create multiple Simple Linear Regression models for both Tesla and NIO to determine either a positive or negative relationship between individual independent variables and the dependent variable.

Second, I will combine the independent variables to use in a multivariable linear regression model. This model, instead of using one independent variable, uses all to determine

a relationship with the dependent variable. The formula for a Multivariable Linear Regression Model is shown below (Fig. 3).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i=n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

$\beta_1 \sim \beta_p$ = slope coefficients for each explanatory variable

ε = the model's error term (residual)

Figure 3 Multivariable Linear Regression Model (Kenton)

Combined with this formula, I will use this Multivariable Linear Regression Model to find the relationship of all independent variables and the dependent variable, the open price of a stock. Combined with the Multivariable Linear Regression, I will use the concept of the coefficient of determination "R^2" to find how close my data fits to the regression line. I will also use the correlation coefficient "r" to determine the correlation between the independent and dependent variables.

My third and last regression model will be the Logistic Regression model. A logistic regression model can be applied when the dependent variable is categorical, meaning it must be binary, either 1 or 0, up or down. To apply this to stocks, we must switch the dependent variable from being non-categorical to categorical. Therefore, the dependent variable will be if either the price of the open of the stock goes up (1) or goes down (0). The equation for the Logistic Regression Model is $P(Y=1) = e^{(\beta_0+\sum(\beta_i X_i)} / (1+e^{(\beta_0+\sum(\beta_i X_i))}$ (Wikipedia) where P is the probability for the dependent variable for case i, and x is a value on the independent variable for case i. If the Logistic Regression Model yields unideal results, this paper will utilize Logistic Regression Bagging to separate the data set into multiple groups and average out the dependent variable in the end.

4. Methodology

a. Data Collection

I collected data from Yahoo Financial for both the stocks Tesla and NIO from the time 9/17/19 – 3/16/20, a total of 128 data points for each independent variable – High, Low, and Volume. In total, there are 768 data points collected for both stocks. A part of the data table is shown below:

Open	High	Low	Volume
48.494	49.12	48.074	19327000
49	49.634	48.474	20851000
49.2	49.588	48.968	23979000
49.298	49.39	47.632	31765000
48	49.036	47.844	21701000
48.304	48.398	44.522	64457500
44.912	45.796	43.672	47135500
46.132	48.662	45.48	59422500
48.44	49.742	47.746	55582000
48.6	48.796	47.222	29399000

Table 1 Tesla Stock Data

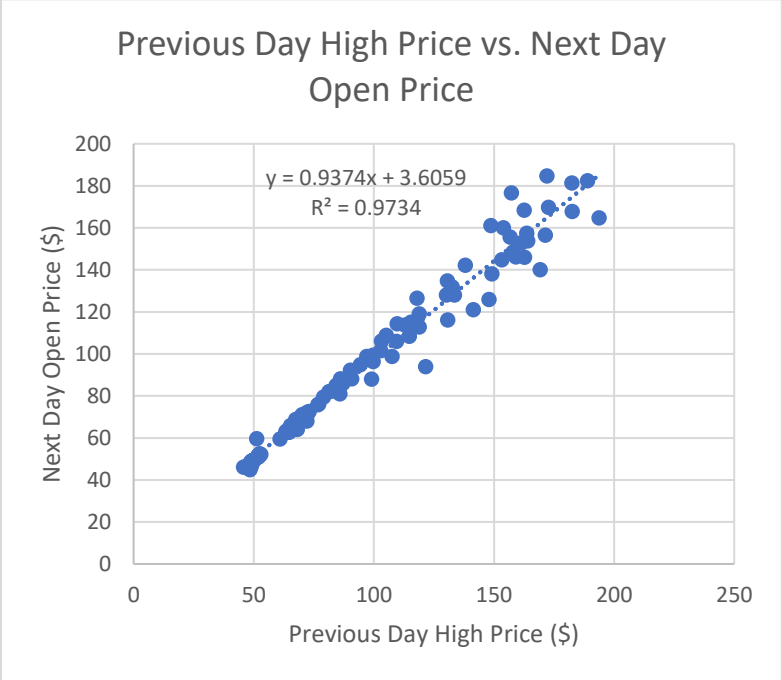
Open	High	Low	Volume
3.11	3.18	3	12905800
3.21	3.21	3.03	12519200
3.12	3.16	3.08	8733200
3.14	3.16	3.02	11999300
2.98	2.98	2.71	40356100
2.22	2.24	1.97	1.22E+08
2.15	2.15	2.02	38938700
2.05	2.06	1.9	50967800
1.94	2	1.71	59824600
1.72	1.73	1.53	58815600

Table 2 NIO Stock Data

b. Simple Linear Regression

1. Graph of Previous Day High Price vs. Next Day Open Price for Tesla

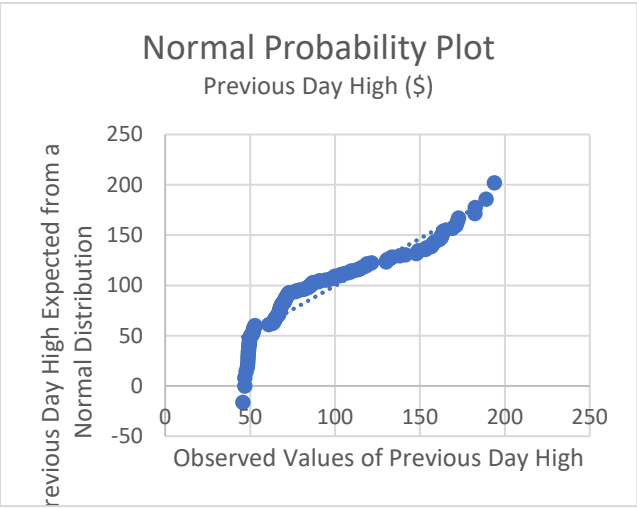
The graph with Previous Day High Price ($) in the X-axis, and the Next Day Open Price ($) in the Y-axis for Tesla is shown below.

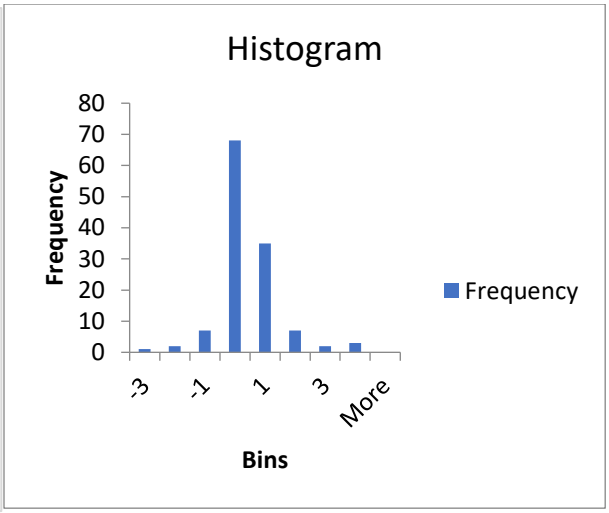


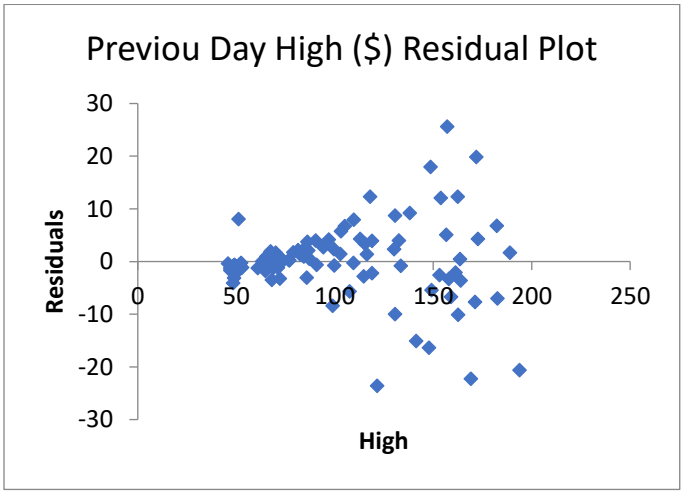
Based on the graph, there is a strong positive correlation between these two variables.

2. Check Conditions for Linear Regression

Before fitting this data to a linear model, there must be three tests done: residual histogram, residual plot, and residual probability plot. The histogram must display a normal distribution of the residual, indicating that a linear model is the best choice. The residual plot must display no clear patter, suggesting the linear model should be utilized. The residual needs to also follow the normal probability line. The three diagrams are displayed below:







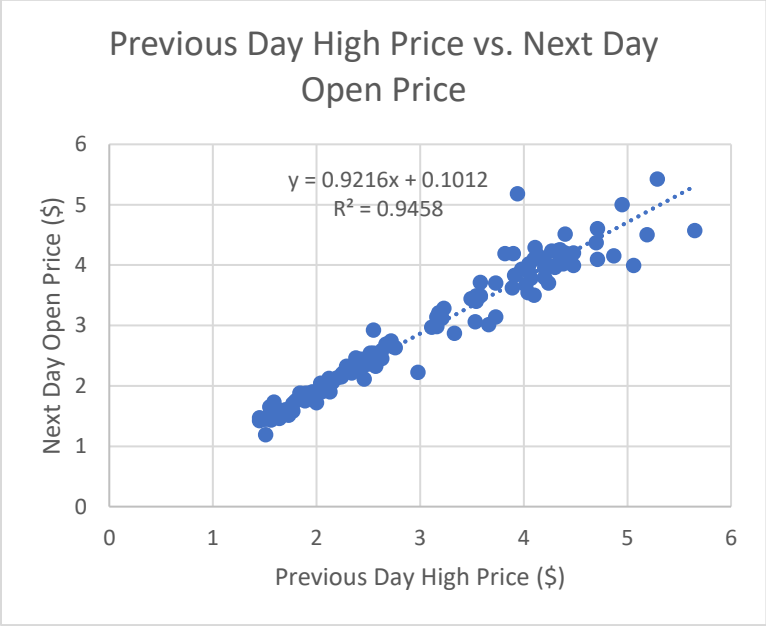
The Residual histogram is centered at 0, with a range from -3 to 3. The shape is normal (bell-shaped), meeting this condition.

The residual plot has no clear pattern, suggesting that the linear model will be the best fit of the data.

The residuals also follow the normal probability plot. Looking at these three tests, the linear model is the best fit for the data.

3. Graph of Previous Day High Price vs. Next Day Open Price for NIO

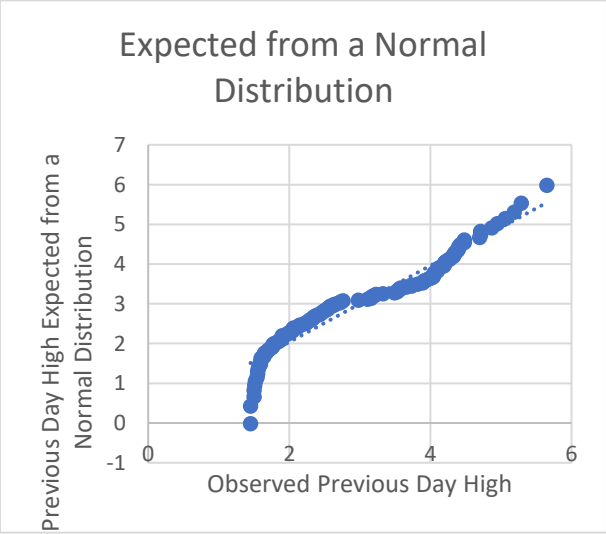
The graph with Previous Day High Price ($) in the X-axis, and the Next Day Open Price ($) in the Y-axis for NIO is shown below.

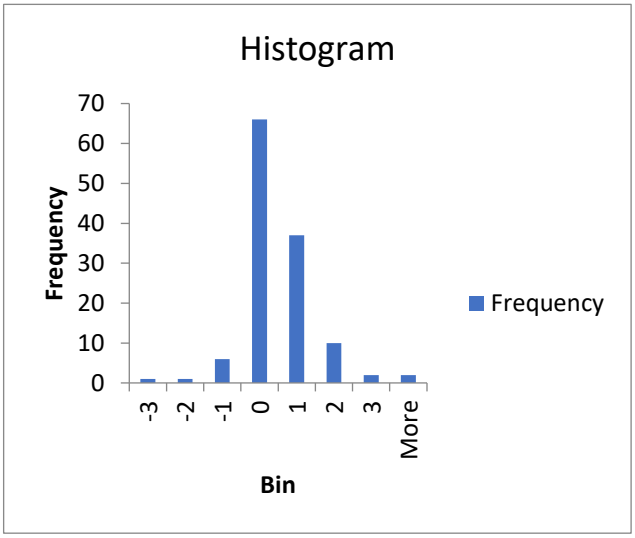


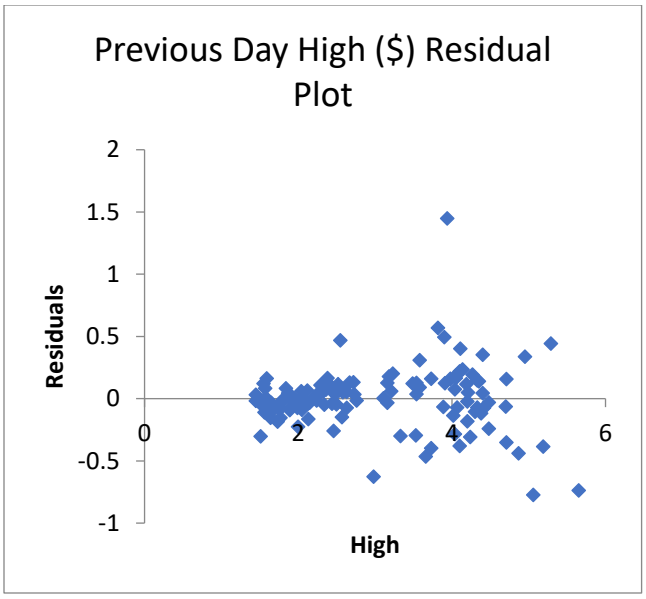
Based on the graph, there is a strong positive correlation between these two variables.

4. Check Conditions for Linear Regression

The same three diagrams for the graph displayed above is displayed below:







The Residual histogram is centered at 0, with a range from -3 to 3. The shape is normal (bell-shaped), meeting this condition.

The residual plot has no clear pattern, suggesting that the linear model will be the best fit of the data.

The residuals also follow the normal probability plot. Looking at these three tests, the linear model is the best fit for the data.

Additional simple linear regression graphs and their components for the rest of the independent variables can be found in the Appendix.

Section Conclusion

From the models above, we can conclude that the Previous Day Open Price and the Previous Day Low price are two very important factors when determining the Next Day Open Price. Volume traded, on the other hand, is generally not as important but still significant in predicting.

c. ANOVA table and Multivariable Linear Regression with Initial Independent Variables

Both multivariable linear regression models below incorporate the original independent variables (Previous Day High Price, Previous Day Low Price, and Previous Day Volume Traded).

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.98858197							
R Square	0.97729432							
Adjusted R Square	0.97673136							
Standard Error	5.96695596							
Observations	125							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	185430.7695	61810.2565	1736.02063	0			
Residual	121	4308.152174	35.6045634					
Total	124	189738.9217						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	0.22818033	1.538426924	0.14832055	0.8823367	-2.81754144	3.2739021	-2.81754144	3.2739021
High	0.26246844	0.153015224	1.71530934	0.08884822	-0.04046554	0.56540243	-0.04046554	0.56540243
Low	0.74000862	0.165285793	4.47714596	1.724E-05	0.412781808	1.06723543	0.412781808	1.06723543
Volume	1.41E-08	1.37E-08	1.02875897	0.30564502	-1.30E-08	4.11E-08	-1.30E-08	4.11E-08
Model = 0.228180331214474 + 0.262468441960457 (high) + 0.740008619030035 (low) + 1.40630918297912E-08 (Volume)								

Figure 10 Tesla Multivariable Linear Regression Model Using Data from 9/17/19-3/16/20

The R Square, also known as the correlation coefficient, explains the importance of the variables relative to the output. For the Tesla Model, the R Square came out as 0.97729432, meaning that 97% of the variation in the model is explained. This value illustrates a very strong correlation between the variables and the output. The small P-values indicates strong evidence against the null hypothesis, therefore meaning this variable rejects it. In my model, I am setting the significance of the p-value to be lower than 0.05 to reject the null hypothesis, as the confidence interval is set at 95%. The p-values of the High Price and the Volume Traded are greater than 0.05 at 0.08884822 and 0.30564502 respectively. The p-value of the Low Price is less than 0.05 at 1.724E-05. This demonstrates that in this model, the independent variable of Low Price rejects the null hypothesis while High Price and Volume Traded does not. The high price also demonstrates evidence as the p-value of this variable is not so far from 0.05 with a difference of 0.03884822.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.97567841							
R Square	0.95194836							
Adjusted R Square	0.950757							
Standard Error	0.23781826							
Observations	125							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	135.5754196	45.19181	799.0415	0			
Residual	121	6.843460443	0.056558					
Total	124	142.41888						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	0.05740998	0.061180447	0.938371	0.349923	-0.06371	0.178533	-0.06371	0.17853281
High	0.40284385	0.22516284	1.789122	0.076097	-0.04293	0.848613	-0.04293	0.848613065
Low	0.57402639	0.231114461	2.483732	0.01437	0.116474	1.031578	0.116474	1.031578409
Volume	5.82E-10	9.44E-10	0.617002	0.538393	-1.29E-09	2.45E-09	-1.29E-09	2.45E-09
Model = 0.0574099823555631+0.402843849585934 (High) + 0.574026389946539 (Low) + 5.82150555562649E-10 (Volume)								

Figure 11 NIO Multivariable Linear Regression Model Using Data from 9/17/19-3/16/20

For the NIO Model, the R Square came out as 0.95194836, meaning that 95% of the variation in the model is explained. This value illustrates a very strong correlation between the variables

and the output. The p-values of the High Price and the Volume Trade are greater than 0.05 at 0.076097 and 0.538393 respectively. This demonstrates that in this model, the independent variable of Low Price rejects the null hypothesis while High Price and Volume Traded does not. The high price also demonstrates evidence as the p-value of this variable is not so far from 0.05 with a difference of 0.026097.

d. Testing the Multivariable Linear Regression Model for Both Stocks

After completing the models for each stock, I plan to test the accuracy of prediction. From Yahoo Financial, I gathered the data from 3/17/20 – 9/16/20, a total of 768 data points for each stock. A part of the data table is shown below:

Next Day Open	High	Low	Volume
	98.974	88.434	20489500
88.001999	94.37	79.2	23994600
77.800003	80.972	70.102	23638100
74.940002	90.4	71.692	1.51E+08
87.639999	95.4	85.158	1.41E+08
86.720001	88.4	82.1	82272500
95.459999	102.738	94.8	1.14E+08
109.050003	111.4	102.222	1.06E+08
109.477997	112	102.45	86903500
101	105.16	98.806	71887000

Table 3 Tesla Stock Data (Used for Accuracy of Prediction)

Next Day Open	High	Low	Volume
	3.12	2.79	34719000
2.4	2.58	2.11	94431900
2.3	2.55	2.23	35499000
2.49	2.64	2.35	50528400
2.34	2.37	2.15	47435200
2.55	2.75	2.4	64750100
2.72	2.99	2.6	56588300
2.81	3.07	2.78	43579400
2.85	2.89	2.76	25132000
2.81	2.83	2.7	20115300

Table 4 NIO Data (Used for Accuracy of Prediction)

Using this data from 3/17/20-9/16/20, I plugged in the data into my individual models. The Average Percent Error for the prediction was 3% for the Tesla stock while 5% for the NIO stock. Below is a subset comparing the Predicted Next Day Open Price compared to the Actual Next Day Open Price with the percent error calculated:

Next Day Open	High	Low	Volume		Predicted Next Day Open	% Error
	98.974	88.434	20489500			
88.001999	94.37	79.2	23994600		91.9357981	4%
77.800003	80.972	70.102	23638100		83.94344666	8%
74.940002	90.4	71.692	150977500		73.68928178	2%
87.639999	95.4	85.158	141427500		79.13123711	10%
86.720001	88.4	82.1	82272500		90.2742299	4%
95.459999	102.738	94.8	114476000		85.34210299	11%
109.050003	111.4	102.222	106113500		98.95636866	9%
109.477997	112	102.45	86903500		106.6046102	3%
101	105.16	98.806	71887000		106.6606585	6%
102.052002	103.33	98.246	59990500		101.9576078	0%

Table 5 Tesla Predicted with % Error

Next Day Open	High	Low	Volume		Predicted Next Day Open	% Error
	3.12	2.79	34719000			
2.4	2.58	2.11	94431900		2.936028106	22%
2.3	2.55	2.23	35499000		2.36291638	3%
2.49	2.64	2.35	50528400		2.385406411	4%
2.34	2.37	2.15	47435200		2.499294898	7%
2.55	2.75	2.4	64750100		2.273921072	11%
2.72	2.99	2.6	56588300		2.580588211	5%
2.81	3.07	2.78	43579400		2.787324617	1%
2.85	2.89	2.76	25132000		2.915303737	2%
2.81	2.83	2.7	20115300		2.820572152	0%
2.83	2.88	2.68	30261400		2.759039463	3%

Table 5 NIO Predicted with % Error

As seen with the Average % Error between the prediction and the actual, we can conclude that using multivariable linear regression to predict the Next Day Open has a very high success rate. With the % Error ≤ 5%, the prediction model seems to be very successful.

e. Logistic Regression

Using a logistic regression requires the dependent variable to be binary. In order to categorize the Next Day Open Price, I split the values into two parts: if the Next Day Open Price goes up from the Previous Day Close Price, I gave it a 1, if not, 0. None of the values stayed the same. A subset of the Tesla table is shown below:

High	Low	Volume	Close	Next Day Open	Modified
49.119999	48.074001	19327000	48.958		
49.633999	48.473999	20851000	48.698	49	1
49.588001	48.967999	23979000	49.32	49.200001	1
49.389999	47.632	31765000	48.124	49.298	0
49.035999	47.844002	21701000	48.246	48	0
48.397999	44.521999	64457500	44.642	48.304001	1
45.796001	43.672001	47135500	45.74	44.911999	1
48.661999	45.48	59422500	48.512	46.132	1
49.742001	47.745998	55582000	48.426	48.439999	0
48.796001	47.222	29399000	48.174	48.599998	1

Table 6 Tesla Logistic Regression with Assigned Value

With the stock values gathered from Yahoo Financial for Tesla and NIO, the initial Logistic Regression Models is as follows:

$$Tesla\ Logistic\ Regression\ Model = \frac{e^{0.91783699-0.2113303\ (High)-0.0704468\ (Low)-1.01887155451383E-08\ (Volume)+0.28465218\ (Close)}}{1+e^{0.91783699-0.2113303\ (High)-0.0704468\ (Low)-1.01887155451383E-08\ (Volume)+0.28465218\ (Close)}}$$

$$NIO\ Logistic\ Regression\ Model = \frac{e^{0.33798528+1.57805078\ (High)-5.0960662\ (Low)-7.956E-09\ (Volume)+3.24999213\ (Close)}}{1+e^{0.33798528+1.57805078\ (High)-5.0960662\ (Low)-7.956E-09\ (Volume)+3.24999213\ (Close)}}$$

f. Testing the Logistic Regression Models

Below are subsets of the test results using the initial Logistic Regression equations:

High	Low	Volume	Close	Next Day Open	Translated Original	Predicted Next Day Open	Translated Predicted	True or False			
98.974	88.434	20489500	89.014								
94.37	79.2	23994600	86.04	88.001999	0	0.249928695	0	TRUE	# of True	49	38%
80.972	70.102	23638100	72.244	77.800003	0	0.411494479	Close Call	Close Call	# of False	43	34%
90.4	71.692	150977500	85.528	74.940002	1	0.308127005	0	FALSE	# of Close	36	28%
95.4	85.158	141427500	85.506	87.639999	1	0.394262508	0	FALSE			
88.4	82.1	82272500	86.858	86.720001	1	0.087569548	0	FALSE			
102.738	94.8	114476000	101	95.459999	1	0.583856341	Close Call	Close Call			
111.4	102.222	106113500	107.85	109.050003	1	0.527804977	Close Call	Close Call			
112	102.45	86903500	105.632	109.477997	1	0.448435268	Close Call	Close Call			
105.16	98.806	71887000	102.872	101	0	0.313138759	0	TRUE			
103.33	98.246	59990500	100.426	102.052002	0	0.57054589	Close Call	Close Call			

Table 7 Tesla Logistic Regression Model Test Result 1

High	Low	Volume	Close	Next Day Open	Translated Original	Predicted Next Day Open	Translated Predicted	True or False			
3.12	2.79	34719000	2.9								
2.58	2.11	94431900	2.43	2.4	0	0.547910511	Close Call	Close Call	# of True	54	42%
2.55	2.23	35499000	2.38	2.3	0	0.690581078	1	FALSE	# of False	49	38%
2.64	2.35	50528400	2.4	2.49	1	0.610719681	1	TRUE	# of Close	25	20%
2.37	2.15	47435200	2.37	2.34	0	0.481573581	Close Call	Close Call			
2.75	2.4	64750100	2.6	2.55	1	0.609809182	1	TRUE			
2.99	2.6	56588300	2.76	2.72	1	0.59432978	1	TRUE			
3.07	2.78	43579400	2.93	2.81	1	0.580870222	1	TRUE			
2.89	2.76	25132000	2.84	2.85	0	0.547678784	Close Call	Close Call			
2.83	2.7	20115300	2.71	2.81	0	0.465912155	Close Call	Close Call			
2.88	2.68	30261400	2.78	2.83	1	0.423586289	0	FALSE			

Table 8 NIO Logistic Regression Model Test Results 1

This paper first calculated the Predicted Next Day Open using the regression model equations in the previous section. Since the predicted column is a probability, this paper decided to translate it so that this column did not contain so many decimals. This paper deemed the prediction 1 if it was greater than 0.65, 0 if it was less than 0.45, and Close Call if it was in between. The True or False column determines if the predicted value matches with the original value. This paper gave the column a true if the predicted value matched the original value, a false if it did not, and a close call if it landed in the interval as shown above. After the first trial run, the prediction for the Tesla Model had a 38% true rate, 34% false rate, and a 28% close call rate, deeming this model highly unsuccessful. The prediction for the NIO model had a 42% true rate, 38% false rate, and a 20% close call rate, deeming this model slightly better than the Tesla model but just as bad.

g. First Improvement of the Logistic Regression Models

SUMMARY OUTPUT								
Regression Statistics								
Chi Square	14.1952198							
Residual Dev.	156.772395							
# of iterations	6							
Observations	125							
	Coefficients	Standard Error	P-value	Odd Ratio	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	0.91783699	0.549499666	0.09485725	2.503869	0.8528577	7.35100139	0.852857705	7.35100139
High	-0.2113303	0.091515532	0.02093091	0.809507	0.67658485	0.9685423	0.676584846	0.9685423
Low	-0.0704468	0.085655247	0.41082305	0.931977	0.78794418	1.1023391	0.787944185	1.1023391
Volume	-1.019E-08	5.65732E-09	0.07170627	1	0.99999998	1	0.999999979	1
Close	0.28465218	0.121398758	0.01903881	1.3293	1.04782333	1.6863887	1.047823326	1.6863887

Table 9 Tesla Model Summary Table

SUMMARY OUTPUT								
Regression Statistics								
Chi Square	5.49600375							
Residual Dev.	167.782791							
# of iterations	5							
Observations	125							
	Coefficients	Standard Error	P-value	Odd Ratio	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	0.33798528	0.524871346	0.5196148	1.40212	0.5012032	3.9224412	0.5012032	3.9224412
High	1.57805078	2.180585737	0.469261	4.845502	0.0674871	347.90181	0.0674871	347.90181
Low	-5.0960662	2.588577925	0.0489905	0.006121	3.832E-05	0.9777056	3.832E-05	0.9777056
Volume	-7.956E-09	8.54611E-09	0.3518524	1	1	1	1	1
Close	3.24999213	2.003799084	0.1048215	25.79014	0.5079451	1309.4548	0.5079451	1309.4548

Table 10 NIO Model Summary Table

Looking at the P-values of each table, the P-value for Volume on both models are higher than the others. Because of this, I decided to remove this independent variable altogether in order to improve the model performance. The model equations came out as follows:

$$Tesla\ Updated\ Logistic\ Regression\ Model = \frac{e^{0.604752922 - 0.212262501\ (High) - 0.008477088\ (Low) + 0.221929966\ (Close)}}{1 + e^{0.604752922 - 0.212262501\ (High) - 0.008477088\ (Low) + 0.221929966\ (Close)}}$$

$$NIO\ Updated\ Logistic\ Regression\ Model = \frac{e^{0.322229152 - 0.041487652\ (High) - 3.34797935\ (Low) + 3.119458256 (Close)}}{1 + e^{0.322229152 - 0.041487652\ (High) - 3.34797935\ (Low) + 3.119458256 (Close)}}$$

h. Second Testing of the Logistic Regression Models

With the new models, I once again tested it on the data. Below are subsets of the test results:

Predicted Next Day Open	Translated Predicted	True or False			
0.653880116	1	TRUE	# of True	74	59%
0.614423244	1	TRUE	# of False	39	31%
0.647840356	1	FALSE	# of Close	12	10%
0.598083365	1	FALSE			
0.621966641	1	TRUE			
0.465468262	Close Call	Close Call			
0.660347888	1	TRUE			
0.658443157	1	FALSE			
0.595999731	1	TRUE			
0.631379825	1	TRUE			
0.64997838	1	FALSE			

Table 11 Tesla Logistic Regression Model Test Result 2

Predicted Next Day Open	Translated Predicted	True or False			
0.516553354	Close Call	Close Call	# of True	35	28%
0.414014945	0	FALSE	# of False	24	19%
0.389282871	0	FALSE	# of Close	66	53%
0.392547516	0	TRUE			
0.403837278	0	TRUE			
0.599408687	1	FALSE			
0.466301622	Close Call	Close Call			
0.497435524	Close Call	Close Call			
0.493315675	Close Call	Close Call			
0.498600501	Close Call	Close Call			
0.572665057	1	FALSE			

Table 12 NIO Logistic Regression Model Test Results 3

Looking at the results, we can see a significant increase of 11% in the Tesla test results for TRUE but a 14% decrease in the NIO test results for TRUE. We can also see that the % of CLOSE decreased by 18% for the Tesla model while increasing by 33% for the NIO model. This makes me wonder what the results would look like if there were only the results of TRUE and FALSE, without Close Call. Below are the results after removing this condition:

Tesla Results

# of True	79	63%
# of False	46	37%

NIO Results

# of True	64	51%
# of False	61	49%

Without the presence of Close Call, we see an increase of 4% for the # of TRUE in the Tesla results and an increase of 23% for the # of TRUE in the NIO results. However, we also see an increase of 6% for the # of FALSE in the Tesla results and an increase of 30% for the # of FALSE in the NIO results. This proposes an interesting question, as most of the Close Call for the Tesla Results turned into TRUE, but most Close Call for the NIO results turned into FALSE. However, these results are still substandard as they have an average of 57% accuracy in predicting the Next Day Open Price.

i. Second Improvement of the Logistic Regression Models with Bagging

After further researching, I discovered Group Bagging which separates the data into different parts, calculating individual models for each part, and averaging out the results in order to find the average percentage of TRUE and FALSE. I implemented this into my original data, dividing the data into three groups using the highest value of the Next Day Open column. The groups came out as follow:

Tesla

	x=Next Day Open		
Group 1	0<x<61.5666657	Group 1 #	28
Group 2	61.566657<x<123.133331	Group 2 #	69
Group 3	123.133331<x<185	Group 3 #	28

NIO

	x=Next Day Open		
Group 1	0<x<1.80667	Group 1 #	28
Group 2	1.80667<x<3.61334	Group 2 #	56
Group 3	3.61334<x<rest	Group 3 #	41

Using this division, I followed the bagging by separating my testing data into these groups as well. However, because the stock market tends to rise over time, my testing data only fitted into the groups of 2 and 3 because of this growth. The testing data was split into these two groups:

Tesla Split		NIO Split	
# of Group 2	19	# of Group 2	39
# of Group 3	109	# of Group 3	89

After separating both data sets, I created models for each individual group using logistic regression. The models are as follows:

$$Tesla\ Group\ 1\ Model = \frac{e^{6.622227943+1.364371327(High)+0.50814401(Low)-2.0074471(Close)}}{1+e^{6.622227943+1.364371327(High)+0.50814401(Low)-2.0074471(Close)}}$$

$$Tesla\ Group\ 2\ Model = \frac{e^{0.674686451-0.537504453(High)+0.10288302(Low)+0.443151682(Close)}}{1+e^{0.674686451-0.537504453(High)+0.10288302(Low)+0.443151682(Close)}}$$

$$Tesla\ Group\ 3\ Model = \frac{e^{0.757500585-0.288111139(High)-0.059561333(Low)+0.347976164(Close)}}{1+e^{0.757500585-0.288111139(High)-0.059561333(Low)+0.347976164(Close)}}$$

$$NIO\ Group\ 1\ Model = \frac{e^{-5.3297455+9.43060204(High)+13.3107501(Low)-19.946954(Close)}}{1+e^{-5.3297455+9.43060204(High)+13.3107501(Low)-19.946954(Close)}}$$

$$NIO\ Group\ 2\ Model = \frac{e^{0.328647283+2.113132037(High)+2.328861916(Low)-4.42606686(Close)}}{1+e^{0.328647283+2.113132037(High)+2.328861916(Low)-4.42606686(Close)}}$$

$$NIO\ Group\ 3\ Model = \frac{e^{-6.5053505+4.71646868(High)+2.03410555(Low)-5.3324728(Close)}}{1+e^{-6.5053505+4.71646868(High)+2.03410555(Low)-5.3324728(Close)}}$$

After I came up with the models, I tested my variables with these models for individual groups. As explained above, only the models of groups 2 and 3 were utilized. After testing my variables, I came up with these tables:

Tesla Group 2			Tesla Group 3		
# of True	9	47%	# of True	43	39%
# of False	7	37%	# of False	52	48%
# of Close	3	16%	# of Close	14	13%

NIO Group 2			NIO Group 3		
# of True	17	44%	# of True	46	52%
# of False	17	44%	# of False	41	46%
# of Close	5	13%	# of Close	2	2%

For the Tesla Models, the average # of TRUE came out as 43%, the # of FALSE came out as 42.5%, and the # of Close Call came out as 14.5%. For the NIO Models, the average # of TRUE came out as 48%, the # of FALSE came out as 45%, and the # of Close Call came out as 7.5%.

j. Model Comparison

First, the simple linear regression gives a prediction of the exact value of the Next Day Open Price using one independent variable. The strength of this model is the fact that we can tell what variables have a direct correlation with the dependent variable by looking at the R-square value to determine correlation. We can use this to update the multivariable linear regression model in the next section. In my model, I found that the only independent variable with a very low correlation with the dependent variable is the "Volume".

Second, the multivariable linear regression gives a prediction of the exact value of the Next Day Open Price of the stock market using three main independent variables: Previous day High, Low, and Volume. The strength of this model is that it is extremely accurate and easy to understand. This model also presents an exact value.

Third, the logistic regression model provides prediction on whether the Next Day Open Price would either go up or down in value. This provides for a vague prediction, unlike that of a multivariable linear regression, as it only directs the trend instead of what the exact value is. I deemed all movies with a probability of more than 0.55 as "goes up", between 0.45 and 0.55 as "Cloe Call", and less than 0.45 being "goes down". Hence, this prediction only talks about what the probability for each possibility is most likely to happen.

Comparing the results of the last two models, compared to the % error of the multivariable linear regression models with 3% error for the Tesla Model and 5% error for the NIO model, we can see a difference of 41.5% for Tesla and 43.5% for NIO in terms of difference in error in the logistic model. This difference of greater than 40% clearly deems multivariable linear regression the better model.

5. Results/Conclusion

Using Simple Linear Regression, Multivariable Linear Regression, Logistic Regression, and followed with Data Elimination, we find the most important variables for each model and generated a total of 18 models to try to find which one gives the most accurate prediction. Finally, a multivariable linear regression model of *Y = 0.2281803 + 0.2624684 (high) + 0.7400086 (low) + 1.4063092E-08 (volume)* for Tesla and *Y = 0.0574100+0.4028438 (High) + 0.5740264 (Low) + 5.8215056E-10 (Volume)* for NIO is generated. This allowed for the results represented in % Error of 3% and 5% for Tesla and NIO respectively.

I followed this model by applying a logistic regression equation to find the model of whether the price would go up or down. Using excel, a logistic model of

$$Tesla\ Updated\ Logistic\ Regression\ Model = \frac{e^{0.6047529 - 0.2122625\ (High) - 0.0084771\ (Low) + 0.2219300\ (Close)}}{1 + e^{0.6047529 - 0.2122625\ (High) - 0.0084771\ (Low) + 0.2219300\ (Close)}}$$

$$NIO\ Updated\ Logistic\ Regression\ Model = \frac{e^{0.3222292 - 0.0414877\ (High) - 3.3479794\ (Low) + 3.1194583(Close)}}{1 + e^{0.3222292 - 0.0414877\ (High) - 3.3479794\ (Low) + 3.1194583(Close)}}$$

Is found to predict whether the Tesla and NIO stock will go up or down with a 37% and 49% Error for Tesla and NIO respectively.

In conclusion, the multivariable linear regression model has higher accuracy, with percentage errors an average of 4%. The logistic regression models performed worse with an average percent error of 43%. I chose the multivariable linear regression model over the logistic regression model as it is not only easy to understand with specific, everyday values but is greatly more accurate.

6. Future Work

I intend to further research into machine learning and on-the-fly algorithm building that is better suited for stocks which are very irrational. I also plan to expand my models onto different stocks as to see if this type of modeling only model growing markets such as that of the Electric Vehicle Market or if it also applies for other markets as well.

7. Lesson Learned

a. It is necessary to make sure you know what variables you are dealing with in the models you are using. In the future, I must make sure that my variables are further defined. Because I am predicting a single day value over a course of few months, I cannot utilize other stock variables such as P/E Ratio to incorporate in my data that may provide a very strong correlation in my model. I intend to adjust my sample size in the future so that I can utilize this variable as well.
b. Another thing I must account for is the market that I am analyzing. The Electric Vehicle Market is growing as people become more interested in saving the environment and making sure they are doing their part. This has led to dramatic increase in the Electric Vehicle sector as the stock market reflect a person's interest in that sector. If I intend to analyze other stocks in the future, I must make sure to utilize models that are more built for other stock patterns.
c. Throughout this process, I have also learned that independent research projects like this allows you to dive extremely deep into a subject that inspires and interests you. Through real life applications, you can develop this subject into a project you can take with you into college. This paper has also benefitted me as it improved my time management skills as independent research projects have no plan except for your own. You must choose time wisely on top of everything else in your academic career and plan accordingly.

8. Bibliography

"Here's How Electric Cars Will Cause the Next Oil Crisis." *Institute of Energy of South East Europe*, www.iene.eu/heres-how-electric-cars-will-cause-the-next-oil-crisis-p3240.html.

"Logistic Regression." *Wikipedia*, Wikimedia Foundation, 27 Oct. 2020, en.wikipedia.org/wiki/Logistic_regression.

"NIO Inc. (NIO) Stock Price, News, Quote & History." *Yahoo! Finance*, Yahoo!, 28 Oct. 2020, finance.yahoo.com/quote/NIO/.

"Stock Market Prediction." *Wikipedia*, Wikimedia Foundation, 7 Oct. 2020, en.wikipedia.org/wiki/Stock_market_prediction.

"Tesla, Inc. (TSLA) Stock Price, News, Quote & History." *Yahoo! Finance*, Yahoo!, 28 Oct. 2020, finance.yahoo.com/quote/TSLA/.

Kenton, Will. "How Multiple Linear Regression Works." *Investopedia*, Investopedia, 21 Sept. 2020, www.investopedia.com/terms/m/mlr.asp.

Nguyen, Joseph. "Regression Basics for Business Analysis." *Investopedia*, Investopedia, 14 Sept. 2020, www.investopedia.com/articles/financial-theory/09/regression-analysis-basics-business.asp.

Blokhin, Andriy. "What Is the Difference Between Linear and Multiple Regression?" *Investopedia*, Investopedia, 16 Sept. 2020, www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp.

Randall, Tom. "Here's How Electric Cars Will Cause the Next Oil Crisis." *Bloomberg.com*, Bloomberg, www.bloomberg.com/features/2016-ev-oil-crisis/.