Javier Lozano

Dr. Steve Levkoff

ECON 494-01 – Intro to Business Analytics

November 15, 2020

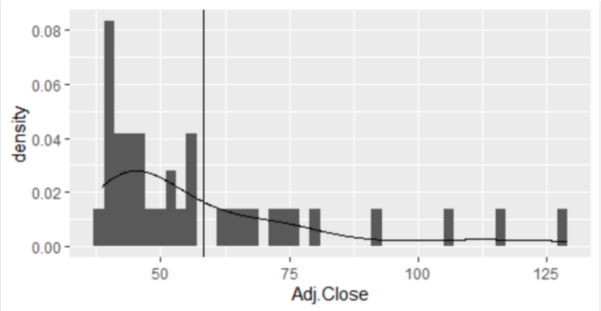
Project II: Predictive Analytics Project – Summary Report

I. Executive Summary

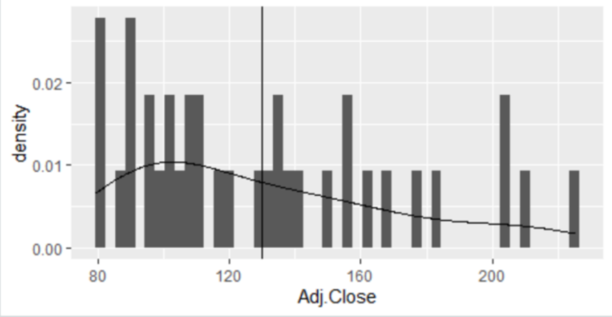
a. Previously on the first part of the project, I concluded that the companies' and index's returns distribution are much higher for when there was no pandemic, but the adjusted closing prices are higher for when there was pandemic. The illustrated graphs in the R code clearly demonstrated the impact of COVID-19 by showing how the adjusted closing prices rose and how the adjusted closing prices before were not as large as they are "today" (the month of October being the last time period observation). So, once I understood this, I did further research upon these findings that the code provided, and I found an article about one of the companies that was examined. The article said that Apple's dividend percentage has indeed been falling recently and that it is better to invest in companies with higher dividend yields. Since the data sets of the companies had high adjusted closing prices recently, simultaneously it makes sense that the S&P 500 has also high adjusted closing prices, especially since these tech companies are a part of the index fund and it is safe to assume that they all have been decreasing their dividend yields, stock splits, and new offerings. As for their roads to recovery, you can look at the data sets using the View() function (import data first as it indicates in this R Project II file), take a look at the "Adj.Close.Change" column (which was created in Excel), and see that this month's adjusted closing price has actually decreased, so hopefully this is a good signal foreshadowing the companies' higher share prices increasing to what they were before the pandemic.

b. Picking up from the last part of the project the final graphs were made to represent the TIDY data set of the three companies (Apple, Microsoft, and Facebook) and the S&P 500 Index. I created a histogram with the geom_vline() and geom_density() functions to illustrate my main point that I had made way in the beginning when I used regular histograms to illustrate the data. It shows that the adjusted closing prices increased drastically when the pandemic began. There is a skewness to the right given the greater number of observations that are pre-COVID rather than post-COVID. However, it is important to remember that pre-COVID adjusted closing prices were much lower that what they are now. This tells us that there the companies that we analyzed have been decreasing dividends, stock splits, or new stock offerings ultimately affecting the S&P 500 Index's adjusted closing price as well. The figures below show the TIDY data.

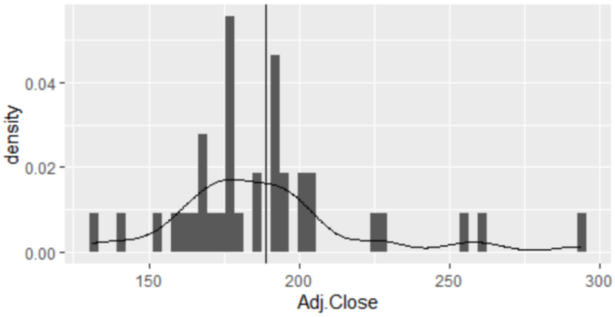
Apple:



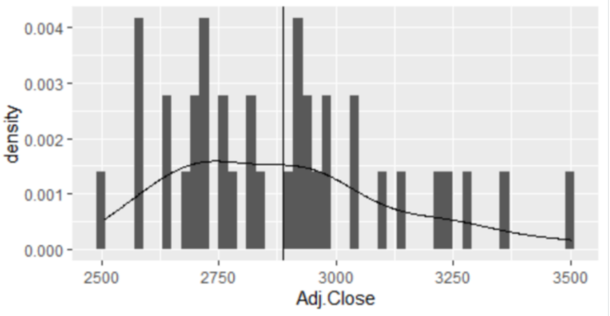
Microsoft:



Facebook:



S&P 500 Index:

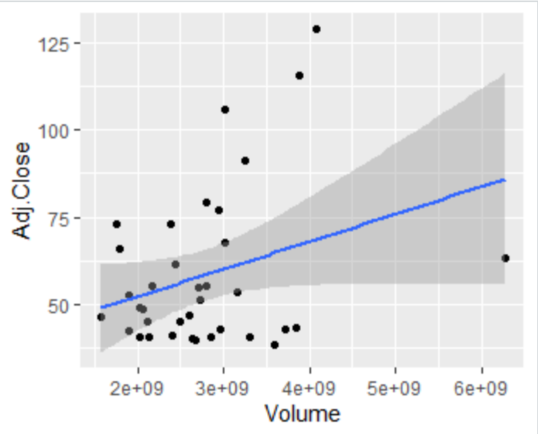


c. Our current variables of interest being the relationship of the adjusted closing prices, the volume of shares sold, and the current timeline of pre-pandemic and post-pandemic. Our focus now is to evaluate only Apple (being the most successful tech company out of the three that we evaluated) and how its returns can be predicted based on the increase of its volume, the periods of when the pandemic was in effect and not, the volume with the pandemic effects, and the volume plus its value squared. This means all four models are trying to predict Apple, to achieve this the variables in the right-hand side and transformations of variables will ultimately be used on Apple stock to predict its returns. The idea behind these strategies are to help us identify the significance of the variables that will be observed and come up with in and out of sample prediction to help us establish which model will be the optimal to utilize when analyzing our data. For this we will have to partition the data into training and testing categories after creating the models to create new models including our predictions that will be generated randomly and ultimately decide on which model would best fit our data.

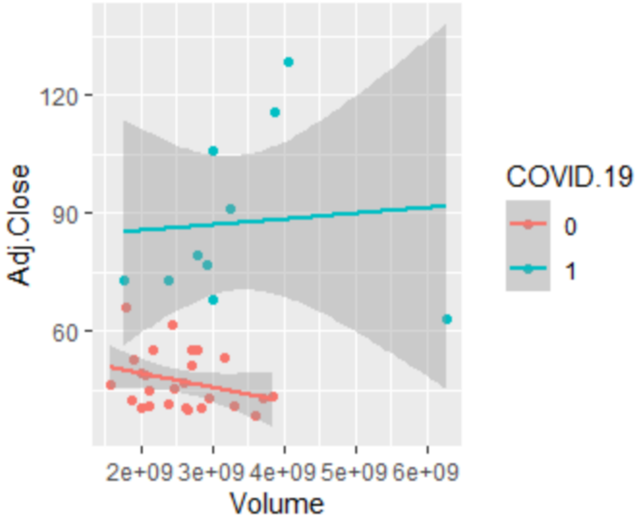
II. Linear Regression Proposals

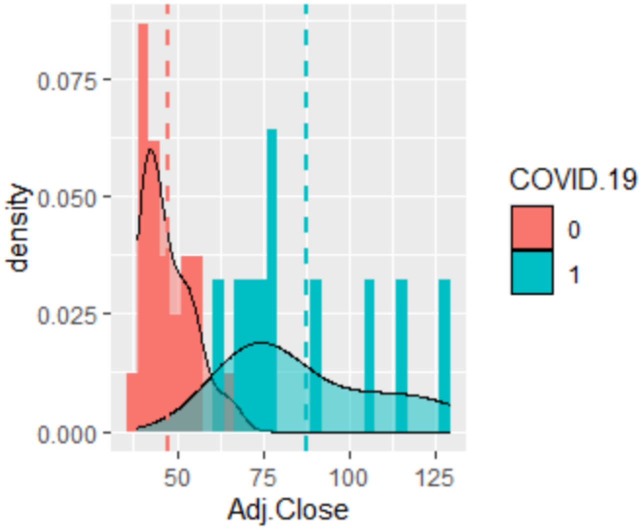
a. The linear regression models are four in total. One of them with one linear term, the second one with a different linear term, the third one with two linear terms, and the final model with a linear and a quadratic term. These were carefully selected to allow us to identify the best model fit for Apple's data.

b. The first model has the adjusted close for the past three years in the y-axis and the volume for each on the x-axis. The idea behind this structure of the model is to quickly analyze the significance of our two most important together variables. When the summary is viewed, the multiple R-squared and adjusted R-squared are both low at around 0.1 and 0.07, respectively. The p-value for the intercept shows a level of significance at 1% and the Volume has a low significance of 10%. This is not a crucial problem because the idea is to test these variables and their level of significance to later compare them with the other models' variables. I then decided to add a visualization of the model using the ggplot tools to perform that and the figure is available below.



c. The second model considers the presence of post and pre pandemic time periods. For this I transformed the variable from numeric to factor, that way 0 means no pandemic and 1 means pandemic in our data set. I again used ggplot tools to provide various visualizations of the graph above with COVID-19 presence indications. The new graphs taking COVID-19 into consideration is below.





d. Now that it is visually evident that COVID-19 did have an effect in our previous variables from our first model, it is safe to move on with our second model taking into account the pandemic’s effects on the company’s adjusted close prices first. In other words, it is our new and vital dummy variable for this analysis. The summary of the model tells us that both p-values of the intercept and our new variable (COVID.19) are significant at 0.1% and our multiple R-squared and adjusted R-squared increase to 67.1% and 66.1%, respectively. So far, our second model is the best fitting model to our overall data.

e. Our third model now includes the variables from the second model and adds the variable that we took out from the first model. This model would consider the adjusted closing prices, volume, and COVID-19 presence; with the first two variables as our main observations and COVID.19 being our dummy variable. In this model both the intercept and the COVID.19 presence have strong significance, but the Volume variable does not have any at all. This indicates that this variable can offset the model to an extent and can be proven by observing how the adjusted R-squared went from 66.1% from model 2 to 65.1% in model 3.

f. Lastly, in my fourth model I decided to take a closer look at the Volume because it's strange that it showed some but not much significance in the first model and no significance at all in the third. The fourth model considers volume and volume squared. The idea behind this model is to attempt to find some sort of meaning in the variable by squaring it, giving it more value. But again, not even the intercept nor the volume variables show significance.

III. Training Data Partition

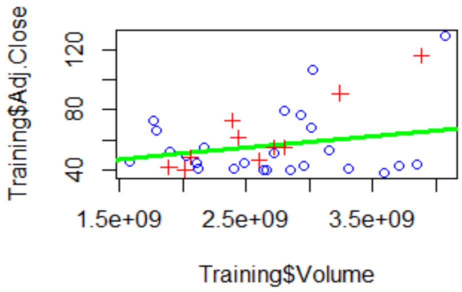
a. It is clear that our most optimal model according to the summary of the models above is the second model given that it shows the highest level of significance out of the four and this means that COVID-19 presence is crucial when analyzing the upward or downward movements in adjusted closing prices. The next step is to partition the data. The typical thing to do in data science is to use 70% of the data to train the model and that is why I set the "p" to 70% and applied the floor function to my observations to round down the number of observations in order for the models to make sense. The other 30% of the data will be used for benchmarking the models. The idea is to keep the identical training data for the four new models and have the testing data to figure out which of the four models performs the best. The observations will be sampled in a random fashion, hence the set.seed(1234) function and set the observations into their own categorized data sets in the global environment called "Training" and "Testing" containing 25 and 11, respectively.

b. M1 is created using the same variables as the previous model 1 that I generated (which are Apple's adjusted close and volume). Using the training data only, I generated in-sample predictions and viewed the data using the fitted values and the regular view function and both have the same numbers. Like the previous model, this model shows a low level of significance with only 10% in the intercept and a multiple R-squared of 5.3% and an adjusted R-squared of 1.2%. Throughout the models I decided to calculated the root mean squared error to measure the average deviation in the actual data relative to the prediction (the expected value conditional on the inputs) and have the same units of measurement as the output variable, y. Keep in mind that the RMSE is analogous to the standard deviation, meaning we measure it relative to its conditional mean – the expected / predicted value conditional on the independent variables in the model. The in-sample error on M1 comes out to equal 21.27. I did somewhat expect this outcome because like in model 1, volume seems to be generating error in our model.

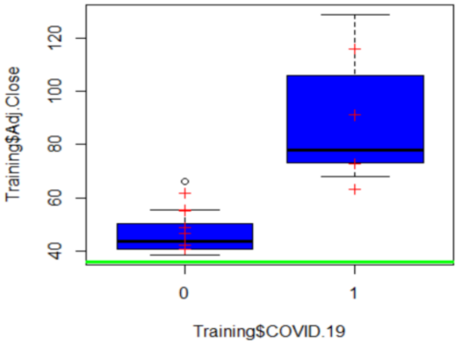
c. In M2 we also take 70% of model 2 to create an in-sample training analysis. The model's variables are now the adjusted closing prices and the COVID-19 presence. The model's summary shows great level of significance in the intercept and COVID-19 variable at 0.1% with a multiple R-squared of 69% and an adjusted R-squared of 67.6%. I generated measures of in-sample error using the training data and got 12.17 for the root mean squared. Less error than M1.
d. M3 considers all variables that we have observed so far (adjusted close, volume, and COVID-19 presence). This model shows less significance than the previous one where volume shows no significance, adjusted close lessened to 1% and COVID-19 remains at 0.01%. As per usual, I generated measures of in-sample error using the training data and got 12.07 for the root mean squared, a little bit lower that the previous error in M2.
e. Finally, M4 shows that once again that volume has little to no significance with the intercept being the only variable with some degree of significance at 10% and the volume and volume squared at no significance. The multiple R-squared in this model would result to 12.47% and the adjusted R-squared to 4.5%. I went ahead and I generated measures of in-sample error using the training data and got 20.45 for the root mean squared. As of right now, this has been the least optimal model by demonstrating extremely low significance compared to the rest when I originally thought that would changed if I squared the variable that has been giving us low results which is volume.
f. Overall, like I mentioned previously, I did expect the models that include only the adjusted closing prices and the volumes to perform poorly when compared to COVID-19 presence inclusion models given that COVID-19 is a determining factor and acts as a dummy variable in our data. Rendering our data more efficient logically speaking. The best model fit from the training data is Model 3 and it is evident because not only are our residuals normally distributed but also our model includes all the variables we would ideally want to take into account when answering the main question of how can we predict Apple's success with the data that is currently available.

IV. Testing Data Partition

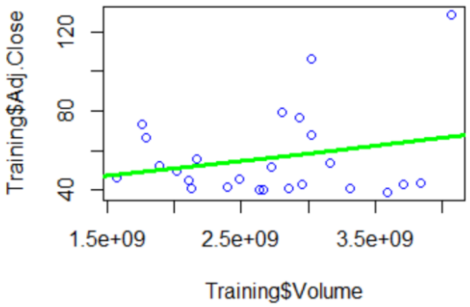
a. The parameters for M1 in the testing data are to be paired with the out-of-sample predictions. The root mean squared for the out-of-sample predictions from the testing data gives us 20.39. Generally, the out-of-sample error should be greater than the in-sample error and in this case, it is the other way around. The results of M1's regression is plotted and can be viewed in the figure below. The error in the blue dots being the training data (measures of in-sample error) and the red crosses are the testing data (measures of out-of-sample error).



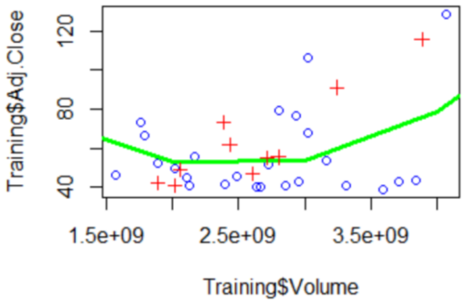
b. The relationships in the figure above will fit better if I replace volume with COVID-19 presence as a variable so that is what I did for M2. I generated measures of out-of-sample error using the testing data and got 13.77 for the root mean squared. This time a high value that the in-sample error, like things should typically be. The figure below shows the two time periods (being pre and post COVID-19) with its fitted values.



c. M3 considers all variables that we have observed so far (adjusted close, volume, and COVID-19 presence). I generated measures of out-of-sample error using the testing data and got 14.73 for the root mean squared. With a higher value in the out-of-sample error. The visualization can be viewed below.



d. Lastly, M4 shows that once again that volume has little to no significance with the intercept being the only variable with some degree of significance at 10% and the volume and volume squared at no significance. The multiple R-squared in this model would result to 12.47% and the adjusted R-squared to 4.5%. I created measures of out-of-sample error using the testing data and got 52.21 for the root mean squared. This is the highest value we have got so far from the out-of-sample error meaning this model would be the less optimal to utilize. The visualization of the model can be viewed below.



e. Overall, the best performing model in our testing data turned out to be M2 instead of M3 as in our training data. I believe that this result speak more accurately than the previous result given that the out-of-sample measurements generally create better predictions as opposed to the in-sample measurements because in-sample measurement primarily focus on fitting the data rather than testing it.

V. Optimal Model Proposal

a. In conclusion, the in-sample model comparison turned out to be M3>M2>M4>M1 while the out-of-sample comparison resulted in M2>M3>M1>M4. I believe that from our analysis of the data the model that would fit our data best is model 2 because the model represents only two variables, the adjusted closing prices and the presence of COVID-19, which currently according to my analysis are crucial variables to account for when predicting Apple's adjusted closing prices in reaction to the impact of the pandemic. When I began the project, I thought the best and most important variables would be the adjusted closing prices and the volume. Given that those are generally the most important things to consider when looking at our TIDY data from the first part of the project. However, if we want to take into account the impact that the pandemic had in regards to affecting Apple's adjusted closing price then it is evident that model 2 is the most optimal model to predict data within our data set. Especially now, speaking in real world terms, if we were to apply this model it would be beneficial for institutional and individual investors when it comes to analyzing the company's different adjusted closing prices over two separate periods of time, which are pre and post COVID-19. Since the future still remains uncertain of the economic impact that the virus will have for the next quarter or year, it is safe to say at least that this model clearly shows the seriousness of the pre and post pandemic considerations and how they can predict future stock value and company tendencies like decreasing dividend yields, stock splits, and new offerings for investors to consider.