Analysis Deliverables

- Hypothesis testing component: the group must formulate a minimum of three distinct research hypotheses and test them using a statistical testing method.

1. There exists a statistically significant causal relationship between price reaction and certain earnings related parameters, that can be modeled using statistical models to a significant degree.
2. There does not exist a strong causal relationship between price reaction and certain earnings call-related parameters (i.e., “Predicted Surprise”, “Median”: median analyst expectations), conditional on the existence of other parameters (i.e., “Mean”), based on a deep learning network, and removing these will have no effect or improve the model accuracy.
3. The deep learning model will have a statistically significantly lower mean squared error loss than a quadratic model on the same variable choices for the relationship between price reaction and certain earnings call-related parameters.

- Machine learning component: you should use at least two of the Machine Learning techniques shown in class. You can use either a supervised or unsupervised learning method.
    1. We are using a basic three layer deep learning model and a quadratic regression model for our two machine learning components. Both models are supervised, and have had parameter optimization done on layers to attempt to find the best possible model layer size in the DL model’s case and polynomial degree count in the quadratic regression model’s case. Both models had 80/20 training/test cross-validation done to assess the model for a given training run, and the dataset was shuffled and resegmented every time we did an independent training run to evaluate the mean squared error.

- Why did you use this statistical test or ML algorithm? Which other tests did you consider or evaluate? What metric(s) did you use to measure success or failure, and why did you use it? What challenges did you face evaluating the model? Did you have to clean or restructure your data?
    1. We went with a DL approach for data prediction under the hypothesis that a quadratic or linear model might not be complex enough or might require too much understanding on our end of how the data interacts with the other factors. Since we were not actually sure of how to connect these correlations, we believe these factors have the strongest predictive power over price reaction: "Surprise

%", "Mean", "Standard Deviation", "Standardized Unexpected Earnings (SUE)", "Number of Estimates", "YoY Growth %", "Mean % Chg (7d Post Rpt)".

a. Since each hypothesis we were testing had two sample sets of data that we wanted to evaluate for a statistically significant difference, we made use of a two sample t-test test to compare both groups. We were using mean squared error, as is typical for regression task evaluation to evaluate the accuracy of the ML/Quadratic regression models. Due to variance on the model end and having limited data, we performed several training runs to be able to confidently get a range of model accuracies and evaluate which model performed "best" at a given architecture.

- What is your interpretation of the results? Do you accept or deny the hypothesis, or are you satisfied with your prediction accuracy? For prediction projects, we expect you to argue why you got the accuracy/success metric you have. Intuitively, how do you react to the results? Are you confident in the results?

| Deep Learning Model Mean Squared Errors on Test Set: | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Selective Data | 26.027 | 3.004 | 104.835 | 0.104 | 82.890 | 0.255 | 3.629 | 3.613 | 1.861 | 0.102 |
| Raw Dataset | 8612.483 | 699.841 | 7.304 | 87.995 | 30.443 | 1178.999 | 4507.689 | 22.892 | 51.736 | 5.986 |

| Quadratic Regression Model Mean Squared Errors on Test Set (degree 1): | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Selective Data | 0.003 | 0.652 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.004 | 0.003 | 0.003 |
| Raw Dataset | 0.050 | 0.004 | 0.003 | 0.003 | 0.009 | 0.004 | 0.005 | 0.228 | 0.632 | 0.003 |

1. Given the extremely low mean squared error for the test set on the quadratic regression model, we can reject the null hypothesis that there would be no way to determine a relationship between 7 day price reaction and the financial factors we chose. The extremely low MSE, although there is no way to perform a stats test against the best possibility, suggests that there is a high degree of model accuracy between the quadratic regression model and the true data graph.
2. When performing the t test to compare both groups data makeups, despite there being a fairly strong MSE difference for both models, the unpaired t test results did not suggest a statistically significant difference. When comparing the two MSE samples for the DL model group, we got a p-value of .1142, for the DL

model, which fails to allow us to reject the null hypothesis. When comparing the two regression models, we got a p-value of .78. This means that statistically speaking, we cannot reject the null hypothesis that there exists a relationship between all input variables and the output predictive accuracy.

3. When performing the t test to compose both groups, we performed two separate sub t-tests. One to compare the smaller factors dataset, and one to compare the larger factors dataset. The smaller factor dataset had a p-value of .08 for the two sample t tests, and the larger factor dataset had a p-value of .11 for the two sample t tests. Both results do not allow us to reject the null hypothesis due to not falling within the 95% confidence interval requirement, and therefore we cannot conclusively state that our hypothesis is true.
4. Frankly, I was surprised that despite the gigantic difference in MSE, there was no clear statistically significant difference between the regression and DL models, and that from just looking at the raw MSEs, the quadratic regression model seemed to be more accurate. However, I would theorize that since after experimenting, we found that degree 1 was the most accurate (making this in effect a linear regression model), it simply means that the modeling function is not as complex as we would have expected. Further, the underperformance of the DL model seems to be due to lack of data. There was far more variance in the DL model's MSE, which would imply either overfitting or simply not enough data. Given we were only working with around 1000 data points, I would theorize the second is true.

Provide comments and an interpretation of the results you obtained:

- Did you find the results corresponded with your initial belief in the data? If yes/no, why do you think this was the case?
    - Hypothesis 1: Honestly, given that the models were relatively simple, we were surprised by the low degree of error we found with our regression model overall. This implies that the modeling function for this data is far simpler than expected, given that not only did a basic quadratic model work, but a quadratic model of degree 1 was shown to be most effective.
    - Hypothesis 2: Given group knowledge about how regression models process data, it does not surprise us that there was no statistically significant difference on how the models handle extraneous variables. These variable's influence in theory should be limited to nothing overtime in the weights matrix / coefficients, and therefore it makes sense to us that we would not be able to reject the null hypothesis.
    - Hypothesis 3: The high accuracy of the linear regression model was very surprising to us, especially compared to the underperformance of the DL model. We had not realized how simple our data was, and we are now curious whether

    adding more data from companies in the S&P 500 would still optimally use such a simple model.

- Do you believe the tools for analysis that you chose were appropriate? If yes/no, why or what method could have been used?
    - I believe that our deep learning component may be slightly too complex for the data we have currently. However, it is not necessarily inappropriate. The advantage to having a deep learning model is that our project is very scalable. The more data we add, the more complex relationships between our parameters we will be able to find.
    - As for the quadratic regression aspect, given the extremely low MSE, we believe it was an adequate tool for an analysis.
- Was the data adequate for your analysis? If not, what aspects of the data were problematic and how could you have remedied that?
    - Our data is extremely simple, and seems to be best modeled by a linear regression. However, we also theorize that we had nowhere near enough data to properly train a DL model to the correct degree of accuracy, which may have impacted our results in this area. We might find it beneficial to expand our dataset to more S&P 500 companies, which would maybe result in more complex functions.