Predicting Price Reaction using Machine Learning

scriado

Hypothesis

Nowadays, it often seems like one of the least predictable elements of the financial world is the stock market. Earnings and profits often fluctuate and vary seemingly at random, but always seem to have underlying logic when the state of a company or organization is looked at. We sought to investigate if the 7-day price reaction of a given stock could be predicted, and if so, what ML technique would be best suited to doing so. We thus tested that there existed a relationship between financial variables related at an earnings call and the price reaction, that some of the variables were irrelevant and could thus be excluded, and that our model making use of several feed forward layers would be in general, more efficient than a simple quadratic regression model.

Data

This project used data from the LSEG Refinitiv database, a student accessible database that hosts the records of several top companies from their quarterly earnings calls. Along with being easily accessible, the data columns were fairly standardized across the entire dataset, with all the units and data columns being kept consistent and included several fairly acknowledged financial factors as well as the resulting 7-day price reaction. As some companies sometimes did not report some data columns, such as the SmartEstimate, we experimented with both zeroing out or eliminating those rows altogether and found that zeroing provided overall better accuracy. While this is not necessarily accurate to the true representation of these values, for practicalities sake we found it better to handle the data accordingly.

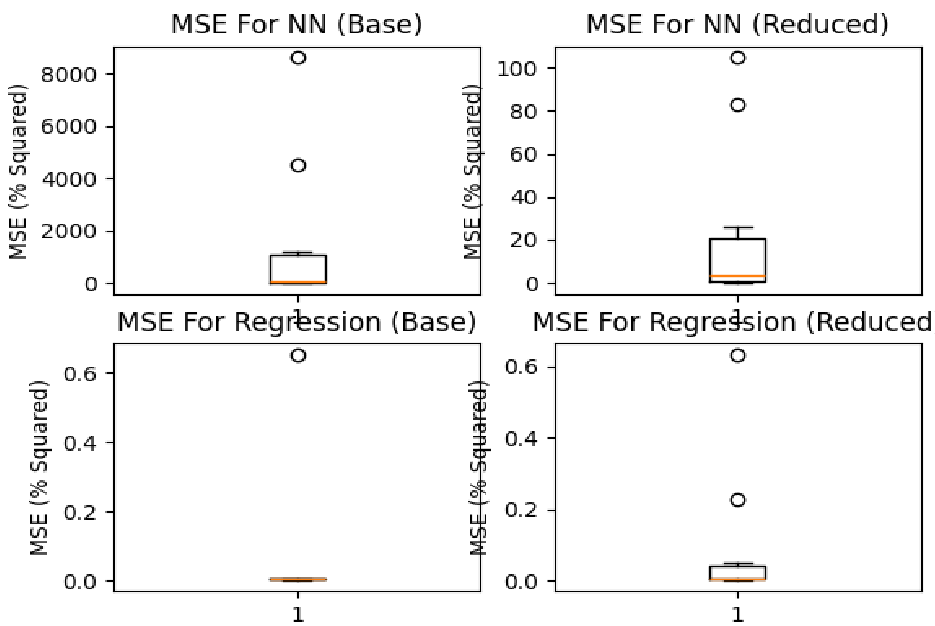
Model Setups:

Both a 2-layer feed forward neural network and a quadratic regression model were trained on a randomly selected 80% of the dataset at any given time and were evaluated on the remaining 20% for a given run. These models are only expected to work on this data, and therefore will likely not generalize well outside of this specific dataset. No data was thrown away, however due to some of the input datasets not containing specific values, some empty data was zero'd in some cases for consistencies sake. We only had around 1100 datums, so randomized validation selection rather than cross-validation was a necessity. The neural network used MSE as a loss function, and accuracy was also measured using MSE.

Findings

Claim #1: There exists a relationship between financial variables reported at an earnings call and the 7-day price reaction.

Support for Claim #1: We experimented with both a DL and a quadratic regression approach, and found that overall, we could consistently get extremely low mean squared errors with a quadratic regression model. While the DL models fluctuated in accuracy wildly, we found that on average a degree 1 quadratic model trained on an 80/20 split almost always got an MSE < .003, which strongly suggests that there is a highly predictable relationship between the earnings call factors and the 7-day price reaction. In addition, given that the more complex DL model and the higher degree quadratic models reported both higher and more inconsistent MSEs on average, this relationship is likely relatively simple.



Claim #2: There is little statistically significant evidence that some reported factors are irrelevant to the prediction of 7-day price reaction, and models trained without these factors will perform at the same or an improved level as a larger, more complex model.

Support for Claim #2: We trained both a multi feed-forward layer neural net and our quadratic regression approach with both a selected subset of our data that eliminated what we believed were related factors as well as the entire dataset. We then performed a two-sample t test comparing the MSEs of these models upon shuffling the data and re-training with the same 80/20 split used above. When comparing the two DL approach samples, we found that we got p-values of .1142 for the DL approach's MSEs and p-values of .78 for the regression model. This implies that from a statistical aspect, we can't say that any models performed better/worse with or without the added factors. Another reason this may be difficult to say is that the DL approach had serious variability issues, which may have impacted this analysis.

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

Claim #3: A simpler model is more capable and more consistent at predicting the 7-day price reaction based off earnings-call data then a more complex DL approach.

Support for Claim #3: While we cannot completely reject the null hypothesis based on t-testing, when looking at the MSE samples due to variance issues in the DL approach, there is a clear difference in accuracy over the 80/20 training/testing split across all tests. Therefore, we can confidently claim that a basic quadratic model is currently a better fit for 7-day price reaction prediction with our limited dataset than a more complex DL model. Further, based on testing of complexity for the quadratic model, a linear one works best then a higher degree model, showing that the simplest model works best in this case.