Stock Price Prediction Using Earnings

Austin Lew, Ethan Tien, Isabelle Do, Victor Yeh

University of California, Davis

Contents:

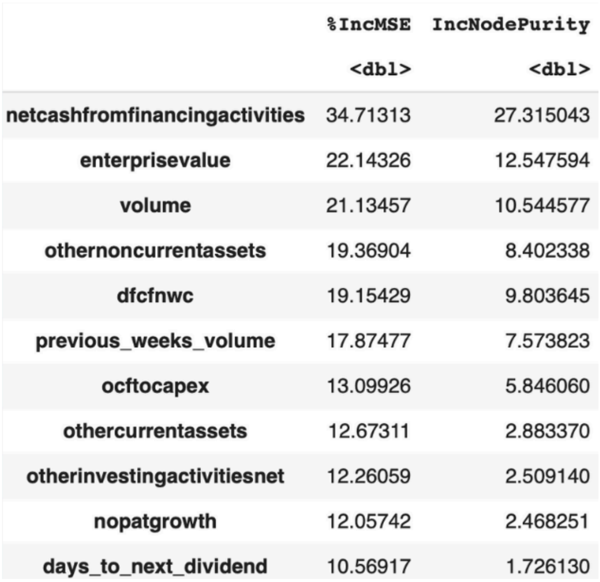
Introduction:

In this project, our objective is to predict stock market price changes the week following an earnings release using that corresponding week's financial reports and various technical indicators. The Dow Jones Index from the UC Irvine Repository comprised several different technical metrics of stocks like volume, opening price, and days to next dividend that provide information related to the characteristics of stocks that might be directly correlated to future returns. Each row represents data for a distinct week of trading activity with each week spanning Friday to the following Thursday. In addition to this dataset, we used data from a third party company, Intrinio, which we cleaned by looking up earnings release dates for each of the relevant companies during that time period. This gives us the ability to isolate the weeks if any abnormal behavior exists when earnings are released. Our motivation for looking at stocks was to see if our research would allow us to maximize earnings, since we were collectively interested in what variables may contribute to the stocks.

Data Exploration and Methodology:

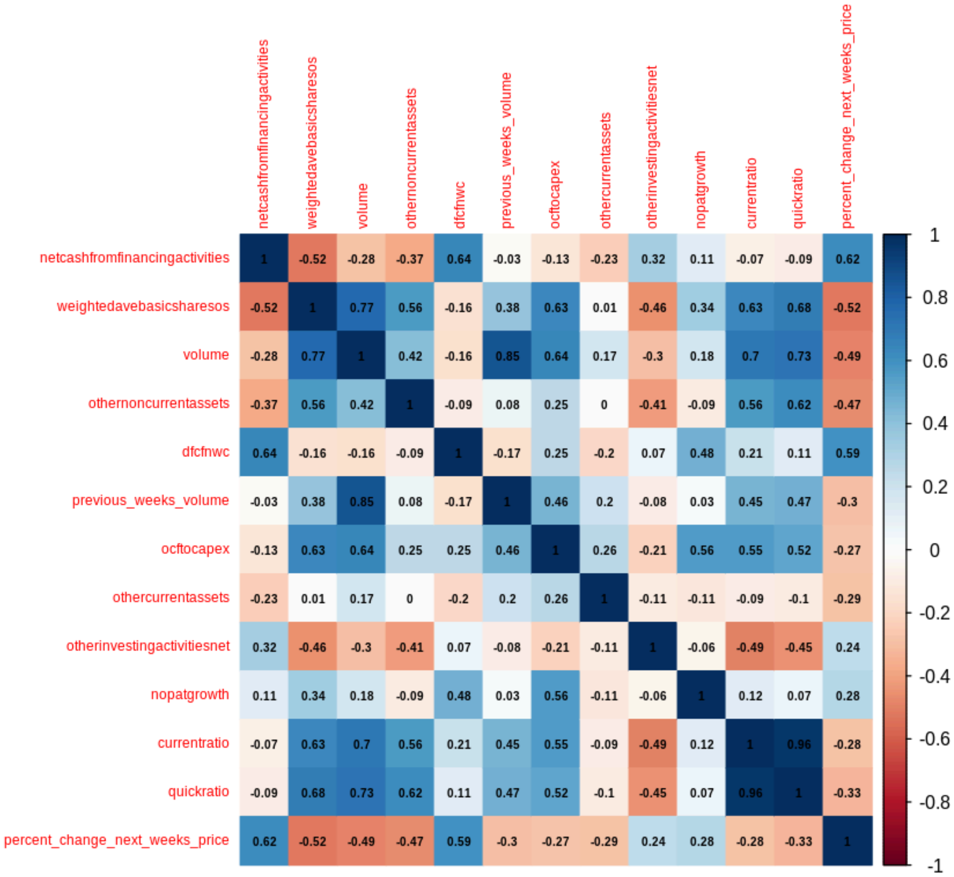
The dataset we've constructed for our analysis is designed where each of its 44 rows encapsulates a quarter's worth of financial data for an industrial company in the Dow Jones Index. Despite an initial glance suggesting a modest dataset size, it unfolds into a complex array with 275 predictors per row, which significantly adds to the depth of our information. We achieved this large array of predictors by first subsetting our UCI dataset into earnings weeks and non-earnings weeks and conducting a series of Welch's Two Sample t-tests comparing variables that allowed us to compare variables for these two groups of weeks and assess if there were any significant differences in behavior or performance. We chose to test for statistical significance in the variables “Volume”, “Percent Change in Volume over the Last Week”, and “Percent Change in the Next Week’s Price” using a significance level of 0.05. Comparing the volume of both of these samples, we got a p-value of 0.7841. Since this is extremely high, we fail to reject the null hypothesis that there is no statistically significant difference in volumes between earning weeks and non-earning weeks. Comparing the percent change in the next week’s price, we got a p-value of 0.8173. We also fail to reject the null hypothesis that there is no statistically significant difference in percent change in the next week’s price between earning weeks and non-earning weeks. Interestingly when comparing the percent change in volume over the last week, we got a p-value of 0.2787. This is also greater than our significance level so we fail to reject the null hypothesis that there is no statistically significant difference in this variable between the weeks. However, this p-value was smaller than the others by a lot, so it may be worth further investigation. We then merged the earnings release weeks with their corresponding earnings reports for that time period with data from Intrinio thus giving us the full 275 columns of data. This ensures we have an accurate depiction of stock movements with their corresponding earnings releases. This dataset then allows for a nuanced exploration of the dynamics between earnings reports and stock price fluctuations over time, focused on a constant cohort of 22 industrial companies reflective of the real-world Dow Jones Industrial Index.

Initially, we tried feature selection by applying forward and backward feature selection to understand what was driving changes in stock prices. This method had two problems: (1) It took a long time to run; and (2) The extracted subset of predictors were extremely prone to overfitting and overdimentionality. We then switched to using the Random Forest model for feature selection because this method allowed us to analyze and rank all the most easily classified features of the dataset at once this gave us better insight on the most critical subset of predictors.



	%IncMSE	IncNodePurity
	<dbl>	<dbl>
netcashfromfinancingactivities	34.71313	27.315043
enterprisevalue	22.14326	12.547594
volume	21.13457	10.544577
othernoncurrentassets	19.36904	8.402338
dfcfnwc	19.15429	9.803645
previous_weeks_volume	17.87477	7.573823
ocftocapex	13.09926	5.846060
othercurrentassets	12.67311	2.883370
otherinvestingactivitiesnet	12.26059	2.509140
nopatgrowth	12.05742	2.468251
days_to_next_dividend	10.56917	1.726130

Through the Random Forest method, we were able to find the top 11 most useful variables that contributed to earnings.



Utilizing the top 11 features that had the highest influence on MSE and highest Node Purity from our Random Forest Model, we created a heat map and observed the correlation between each individual feature with the percent change in a company's stock price, along with "Quick Ratio" and "Current Ratio" which were added based on domain knowledge.

After seeing the correlations we believed that these factors represented a reasonable sample of features to use in a Linear Regression model to see which feature combinations held the highest predictive power. Using the standard error and p-values with a significance level of 0.1 we identified the top 4 best linear predictors of a percent change in price were "Other Non-Current Assets", "Debt-free, Cash-free Net Working Capital (dfcfnwc)", "Net Operating Profit after Tax (NOPAT) growth", and "Operating Cash Flow to Capital Expenditures (ocftocapex)." These factors make sense as each factor examines a company's earnings and makes predictions based on a combination of available cash, profitability, and overall liquidity of a given stock. We ran a series of k-fold cross validations from 2-44 on our linear model to check for overfitting/over dimensionality. We found that, aside from 5 cases where the MSE jumps to outlandish numbers our model was fairly consistent in its predictions. When we plot the residuals of our linear regression model we can see the cause of the MSE jumps are results from a handful of outliers that have undue leverage over the other data points. These outliers might be cause for additional independent study. Seeing as the model used "Non-Current Assets" in its predictive model, we decided to try removing it in favor of metrics that measured "Current assets" using background accounting knowledge. The result was a model that used "Current Ratio" as a quadratic term and "Quick Ratio" as a linear term. This lowered our overall MSE of the quadratic model as well as fewer outlandish MSE values when testing across 2-44 k-folds. After reviewing the residual plots for each of the corresponding models, we feel that the models avoid the pitfalls of over-dimentionality and overfitting as well as being simple enough to understand conceptually. The factors we found to be the most significant predictors make sense under the lens of accounting as well. Most financial text regarding a company's earnings, like "The Intelligent Investor," put a heavy emphasis on examining a given company's ability to pay off debt (Quick/Current Ratio, Other Non-Current Assets), Profit Growth (NOPAT growth), and ability to reinvest using free cash flow (dfcfnwc, ocftocapex). The theory being that a well funded, profitable company will have low amounts of debt relative to their available cash as well as being able to use free cash flow to reinvest in itself.

Conclusions and Findings:

To conclude our study, we examined the potential of using earnings data to predict stock prices the week following an earnings release. Using the dataset from the UCI database the Dow Jones index supplemented by earnings from Intrinio, we feel we have found a significant correlation. With the features available to us, we achieved an MSE of 0.9 on our Random Forest at 13 k-folds, 1.7 on our linear regression at 9 k-folds and 1.62 on our quadratic regression at 9 k-folds. It is important to note that the correlation in our predictors do not necessarily imply causation. The cause for these specific features holding more predictive power than other features in our data set is still unknown to us. It is entirely possible we simply identified items that investment banks value the most driving the price up as a result. Nonetheless, our analysis not only highlights the viability of using earnings data for predictive purposes but also opens up new avenues for future research in financial market dynamics. Our project gives us a deeper understanding of how specific financial indicators can be utilized to forecast market trends, offering valuable perspectives for investors and analysts alike.

Disclaimers:

Our findings We can only assume our conclusion pertains to industrial stocks in the Dow Jones Index for Q1 and Q2 of 2011; beyond which would be speculation. For a more thorough study we could have used more time series data with earnings weeks. We can only assume our findings pertain to industrial stocks in the Dow Jones Index for the first 6 months of 2011. None of us are certified financial advisors, this is not financial advice.

Contact Information:

atlew@ucdavis.edu | vyeh@ucdavis.edu | ittdo@ucdavis.edu | eetien@ucdavis.edu

References

- Full Data set: https://drive.google.com/file/d/1PHD9M1csSliAS9nhuTBaPYLKTRYraULc/view
- Brown,Michael. (2014). Dow Jones Index. UCI Machine Learning Repository.https://doi.org/10.24432/C5788V.
- Graham, Benjamin. *The Intelligent Investor*.
- Google Colab: STA141C_final_proj.ipynb
    - Note: we used R for runtime so all our code is R