Modeling Short Term Insider Trades

STA302 - Final Assignment

Sahil Patel - 1006747905

December 17, 2021

Contents

Introduction

The purpose of this investigation is to attempt to generate a profitable trading strategy. A linear model will be created over short-term insider trades to model the percent change in a stock's price based on how many shares an insider traded, whether they bought or sold shares, how much the shares cost, how long since the insider traded, and the position of the insider. An insider is an individual with prominent status in a company who trades shares of that company. Insiders could trade with only private information, which is illegal according to the U.S. Securities and Exchange Commission, plausibly making their trades more profitable. According to previously conducted research, some insiders yield abnormally high returns over a year (Cziraki & Gider, 2019 and Gangopadhyay, Yook & Sarwar, 2009). Thus copying their trades would be profitable. Furthermore, whether an insider buys or sells is a good indication as to whether the stock might increase or decrease in price (Jeng, Metrick, & Zeckhauser, 2003).

The only issue with previous studies is they analyze the long term change. Thus the importance of this analysis is to investigate the percent change in these stocks in the short term. By analyzing the short-term movements of these stocks (<7 days), it might be possible to identify key exit points in trades to not lose money. Thus, it might be possible to make lots of short-term trades to yield more profits.

Methods

First, the data will be split 70/30 into a training and testing dataset. All of the following steps will be conducted on the training dataset first. Model assumptions are then validated so the results from variable selection can be trusted. Then variable selection is conducted and lastly, the model is validated.

Model Assumptions

Pairwise scatterplots of all the predictors are created to verify the conditional mean of each predictor is a linear function of another predictor. Next, the full model is fitted as seen below:

$$\widehat{\text{Price Change}}_i = \hat{\beta}_0 + (\hat{\beta}_1 \cdot \text{Stock Price}_i) + (\hat{\beta}_2 \cdot \text{Shares Exchanged}_i) + (\hat{\beta}_3 \cdot \text{Days Since Trade}_i) + (\hat{\beta}_4 \cdot \mathbb{I}_{i=\text{sell}})$$

A plot between the fitted values (using the full model) and the response values is created to verify that the conditional mean response is a single function of a linear combination of the given predictors. Failure for this graph to somewhat adhere to the identity function would make it reasonable to believe a linear model might not work for this problem. Next, a QQ Plot is generated to verify the normality of the model. Minimal deviation from the standard normal indicates the assumption is satisfied. Then leverage statistics of all the observations are calculated. If there are any leverage points, the standardized residual plots between each predictor and the response variables is created, else the normal residual plot is created. If a

fanning pattern is noticed, then the assumption of constant variance is not verified and a variance stabilizing transformation can be applied. If a systematic pattern is found in the residual plot then a Box-Cox transformation could be applied. If either of the transformations is applied, then the model will need to be refitted and model assumptions re-verified. Variance inflation factors are then calculated for all the predictors. If any predictor has a variance inflation factor over 5, it will be removed as a result of it being too related to another existing predictor. After all of these steps are taken and any important notes are mentioned, then the assumptions for the model are satisfied.

Variable Selection

First, we conduct an ANOVA F-test. If a p-value less than 0.05 is calculated, then we can proceed with variable selection knowing at least one predictor is linearly related to the response. A T-test is conducted on all the predictors and if there is a predictor with a p-value greater than 0.1, a reduced model is created without the predictor with the highest p-value. Then the adjusted coefficient of determination is calculated of the reduced and full model. If the adjusted coefficient of determination decreases, then the full model is kept. If not, we repeat this process on the reduced model. This is done to generate a simpler and more generalizable model while maintaining model accuracy. Lastly, a partial F-test between the full model and the reduced model is conducted. If the p-value is less than 0.05, then we have a final model. Model assumptions are once again verified and any violations of model assumptions are noted.

Model Validation

The test data is used to fit another model with the same parameters. Then differences in the predictors, coefficients of determination, and predictor coefficients are noted. Any major differences are noted as limitations in the final model as it means the model could've been overfitted. Furthermore, model assumptions are verified on the test model, as differences in model violations could indicate the model will not have a strong predictive performance.

Model Violations and Diagnostics

Leverage and outlier points are then calculated. If any are found, they are noted as they might've impacted the model by altering the coefficients. Then Cook's distance, DFFITS, and DFBETAS values are calculated for all the observations to determine any influential points that might've severely affected the model. This process is done for both the model generated from the training data and the testing data. A large number of influential, outlier, or leverage points in either of the datasets could be used to explain why a model might not have been validated. Lastly, any model violations noted during the model assumption verification phase are noted as limitations to the model. Given the goal of this analysis was to generate a predictive linear regression, some of the drawbacks of model violations are noted, but model generation proceeds.

Results

Data

Table 1: Numerical Summaries of Variables
Training Data Table

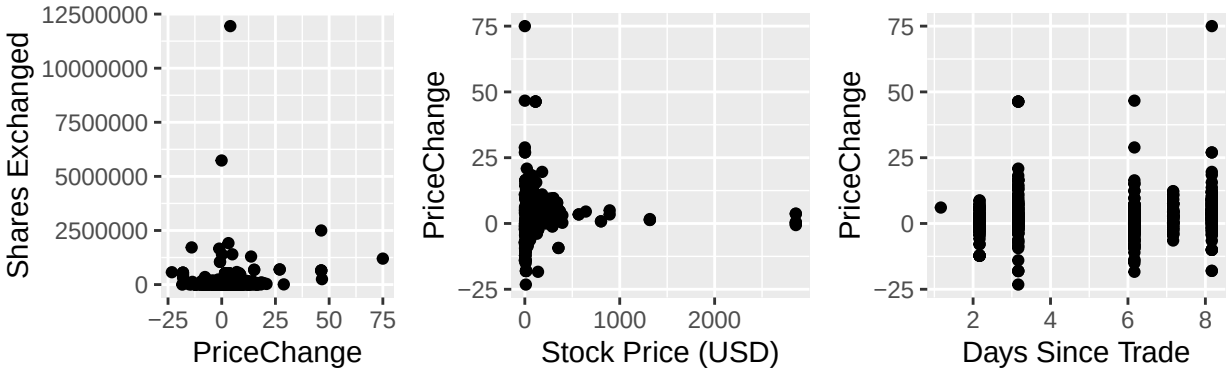
Predictor	Mean	Median	Variance
Stock Price	112.8055361	43.94	8.5117688×10^4
Shares Exchanged	1.2076603×10^5	1.5313×10^4	4.2756901×10^{11}
Asset Price Change	2.2127133	1.58	53.0245777
Time Since Trade	5.098833	6.1666667	4.7826692

Testing Data Table

Predictor	Mean	Median	Variance
Stock Price	133.2138265	53.46	1.0068133×10^5
Shares Exchanged	6.2689087×10^4	1.07475×10^4	4.2581034×10^{10}
Asset Price Change	2.0969388	1.68	64.4117752
Time Since Trade	5.1513605	6.1666667	4.8972004

Large differences in the testing and training data indicate that it might be hard to validate the model. Furthermore, large variances for the stock price, shares exchanged, and the percent change in the asset price could be indicative of a weak linear model.

Figure 2. Scatterplot of Numerical Predictors Against the The Stock Price Change

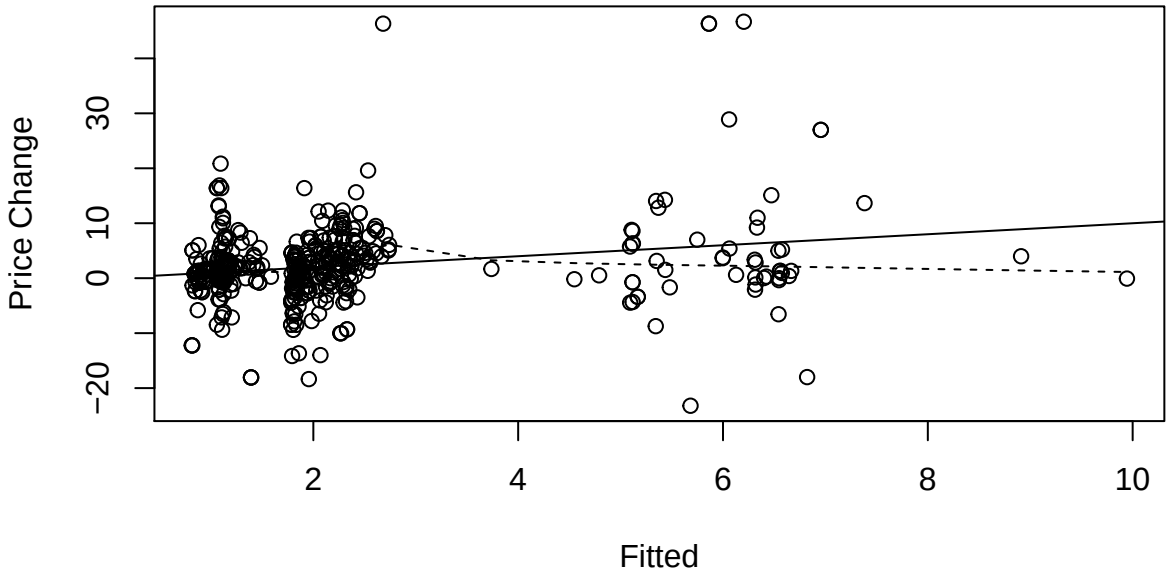


As seen in figure 2, the data appears to contain numerous outliers, and as a result, could also make model validation incredibly difficult.

Model Assumption Verification

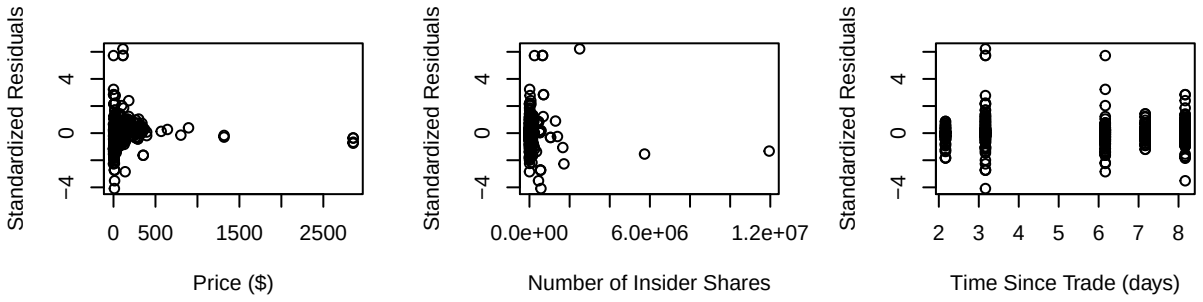
Large deviations in the fitted versus residual plot as seen below in figure 2 are indicative of how a linear model could be a very poor tool for analyzing the given predictors against the response.

Figure 2. Model Fitted Values versus Response Values



Next, numerous leverage points are calculated/observed so it is necessary to inspect standardized residuals to identify any patterns that might either require a Box-Cox transformation or a variance stabilizing transformation.

Figure 3. Standardized Residual Plots of the Numerical Predictors



As noticed in Figure 3, no discernible pattern is detected, thus neither a Box-Cox transformation nor a variance stabilizing transformation is applied. All other model assumptions hold noting the existence of leverage points.

Variable Reduction

First, the position of the insider is removed as no external research supports it as a relevant predictor. Furthermore, the large variety of positions in the data would only be supported by a few data points and thus no strong conclusions can be made about this factor. A p-value of 0.0002638 is obtained after an F-test on the full model, thus variable reduction can proceed. A T-test is conducted on each predictor, and it is observed that the price of the stock has the highest p-value of 0.263208. As a result, this predictor is removed and another model without this predictor is generated. However, a reduction in the adjusted coefficient of determination is noticed from 0.0377 to 0.03711. As a result, it is decided to keep this predictor in the model and end variable selection.

Model Validation and Diagnostics

After fitting another full model to the testing data, a difference in the coefficients and the significant predictors was noticed. As a result, outlier, influential, or leverage points within the training/testing data were calculated. It was determined that there are numerous outliers/leverage/influential points and as a result model validation is nearly impossible. As a result, it is noted that the final model generated by the training could not be validated using the testing data.

Model Specifics

Table 2. Model Coefficients

Coefficient	Value
β_0	4.5643112
β_1	0.0013071
β_2	5.9397193×10^{-7}
β_3	0.24155
β_4	-4.2800466

From Table 2 and the methodologies equation, we notice how an increase in the price of a stock by one dollar can increase the estimated percent change in the stock by about 1/10th of a cent. Furthermore, we notice that as an insider purchase one more share, the estimated percent change in the stock price only increases by a marginally small about of 5.9397193×10^{-7}. Next, we notice that after a day, the percent change of a stock is expected to increase by approximately 0.24155. Lastly, it was noticed that if an insider sells shares, rather than buying shares, the expected percent change in the price of the stock is -4.2800466.

Overall, this model has a low r-squared (0.0461423). As a result of the model diagnostics and this low r-squared value, its noted that this model has low predictive performance and is not representative of any pattern that might exist in the data.

Discussion

As described in table 2 and by the coefficient of determination, the final model predicting the percent price change of a stock traded by an insider is poorly described as a linear combination of the number of shares exchanged, the price of the stock, whether the insider bought or sold shares, and how many days its been since the insider traded. Given the very low coefficient of determination, the goal to generate an accurate model was not achieved and as a result, we fail to accurately predict the percent change in a stock's price traded by an insider trader.

Limitations

Given the data, the number of leverage points, outliers, and influential points, it is understandable why the model has such poor performance in predicting the percent change of the stock's price. As shown in figure 2, a linear model might not fully be appropriate, however, no obvious patterns were detected in the residual plot thus no transformation could be justified. Since the data was only collected over a single month, it is very hard to extrapolate the results from this data; another limitation. Overall, the large number of limitations seems to invalidate the use of this model.

Bibliography

Cziraki, Peter and Gider, Jasmin, The Dollar Profits to Insider Trading (March 31, 2021). Review of Finance, Forthcoming, TILEC Discussion Paper No. DP 2017-005, 14th Annual Mid-Atlantic Research Conference in Finance (MARC), Available at SSRN: https://ssrn.com/abstract=2887628 or http://dx.doi.org/10.2139/ssrn.2887628

Gangopadhyay, P., Yook, K. C., & Sarwar, G. (2009). Profitability of Insider Trades in Extremely Volatile Markets: Evidence from the Stock Market Crash and Recovery of 2000-2003. Quarterly Journal of Finance and Accounting, 48(2), 45–61. http://www.jstor.org/stable/40473485

Jeng, Leslie A. and Zeckhauser, Richard J. and Metrick, Andrew, Estimating the Returns to Insider Trading: A Performance-Evaluation Perspective. The Review of Economics and Statistics, pp. 453-471, May 2003, Available at SSRN: https://ssrn.com/abstract=146029 or http://dx.doi.org/10.2139/ssrn.146029

Ganti, A. (2021, September 3). What is insider trading? Investopedia. Retrieved October 18, 2021, from https://www.investopedia.com/terms/i/insidertrading.asp.

Hayes, A. (2021, August 22). Volatility. Investopedia. Retrieved October 18, 2021, from https://www.investopedia.com/terms/v/volatility.asp.

Tun, Z. T. (2021, August 12). Buy stock with insiders: How to track insider buying. Investopedia. Retrieved October 19, 2021, from https://www.investopedia.com/articles/investing/040915/buy-stock-insiders-how-track-insider-buying.asp.

U.S. Securities and Exchange Commission. (n.d.). Insider trading. Investor.gov. Retrieved October 18, 2021, from https://www.investor.gov/introduction-investing/investing-basics/glossary/insider-trading.