Using Multilinear Regressions on Alternative Data to Predict Movements in NASDAQ Index

Anay Contractor, Thor Mead, and Arya Bhansali

December 2022

1 Introduction

The financial markets are highly unpredictable systems subject to millions of variables and factors; a change in one variable can unhinge the entire system. To develop the thesis for our project, we decided to analyze the causal relationship we want to investigate from an economic perspective to provide a basis for the technical work we had to do.

There are various normative methods and data points for predicting market movements such as balance sheet analysis, back testing, and technical analysis. However, we wanted to build a model using different data and techniques, namely alternative data. We define alternative data as non-traditional data that is not within investors' traditional data sources (financial statements, SEC filings, press releases, historical price data). Data such as credit card spending, geo-location, and web traffic are some examples of alternative data sources [1].

We want to harness the power of alternative data - not only harness, but provide economic significance for our machine learning model - to generate meaningful insights about the stock market.

To gather data, we were able to use the Federal Reserve Bank of St. Louis website. We will use the following variables from there [2].

- Large Bank Consumer Credit Card Balances: Total Balances
  - The total credit card balances show us consumer consumption patterns as we can see how much they are spending in a given time period. We've specified the data for large banks, as most credit cards are issued from larger banks such as Bank of America, Citibank, Wells Fargo, and more. Using this data, we can examine more substantial consumer spending patterns over time. This data is collected by the FRED on a quarterly basis even though we get credit card bills on a monthly basis as an individual consumer. This is important to note because there's a difference between when consumers get sent their credit card balance and pay it each month compared to when the FRED gets access to that data quarterly.
- Commercial Bank Interest Rate on Credit Card Plans, All Accounts
  - Interest rates are important as they provide an indication for how spending will pan out. When interest rates are higher, this would encourage more saving and less spending within the economy. There would also be less borrowing within the economy. When interest rates are lower, there will be an opposite effect.
- Consumer Price Index for All Urban Consumers: Information Technology, Hardware and Services in U.S. City Average
  - The consumer price index measures the change in price over time for a specific set of goods/services. You get this measurement by taking the current market basket of goods and divide by the initial

market basket of goods and get a percentage. This is geared towards specifically for information technology and hardware, determining how the market for technology services is doing and if the prices are increasing or decreasing.

- Consumer Price Index for All Urban Wage Earners and Clerical Workers: Information Technology, Hardware and Services in U.S. City Average
    - This is very similar to the previous measurement of CPI for urban consumers. Since majority of the country are wage earners, this measurement affects the standard of living amongst people in the economy. When this metric is increasing, that means that there's potential inflation on the rise. Depending on the other factors, this could mean that the standard of living has increased in the economy as well as people are more confident of the overall trend of the economy. In this specific case, it is measuring people and their spending on technology services.
- NASDAQ Composite Index
    - This is one of the big three stock market indexes in the United States. This includes big-name companies such as Apple, Tesla, etc. The NASDAQ Composite is important to track as that is an important indicator of how the stock market is performing. Although not fully correlated to the economy, the health of the stock market can help determine consumer confidence in the future of our economy by looking at how companies such as Apple, Tesla, etc. are performing in the stock market.
- Net Percentage of Domestic Banks Tightening Standards for Credit Card Loans - Quarterly

  - When there is a tightening of loans, there could be a credit crunch due to the fact that there are not as many opportunities for people to get a line of credit from the bank. When there is a tightening of loans, it would lead to decreased spending due to the scarcity of credit available and vice versa if there is not as much restrictions by the Board of Governors. This can also affect how consumers obtain loans since there would be more regulations on them, making it harder if more banks are getting tighter on their loaning policies.

- Personal Savings

  - The rate of personal savings is important because it can determine how healthy the economy is overall. During periods of recession, we can expect this rating to increase since people are unable to spend as much and decrease during booming periods since there are more opportunities for people to spend. This rating can also determine consumer confidence in the economy as it is a percentage of disposable income, which is the amount left over after deducting taxes and other mandatory charges.

- Quarterly Financial Report: U.S. Corporations: Computer and Electronic Products: Net Sales, Receipts, and Operating Revenues

  - This measures quarterly reports for companies with computer and electronic products. This tells us how corporations are operating within this industry and whether they are making money or not. If corporations are increasing in sales and revenue, this would mean that they are able to make more profits, which means in turn that the demand for technology has increased. What this means is that consumers are more confident in the economy and corporations have a better incentive to produce more goods/services for people. This also means that corporations are able to find better ways to generate profit as well.

- Velocity of M1 Money Stock
  - This measures amount of times a dollar is used to buy goods/services as a unit of time. This rating is important as it determines how much demand and supply there is overall in the economy (indicating short-term consumption). During good economic times, you can expect this to increase because people are more able to buy products, so demand would increase. The suppliers would be able to produce more products as well since they will cost more during good economic times. During bad economic times, you can expect this variable to decrease as less people are able to buy products, decreasing demand along with less incentive from suppliers to produce goods since their products won't cost as much.

2 Data Cleaning

Before any testing and analysis was done, we first had to clean up our data. When looking through all of the CSV files, we saw that the data had different time blocks. We saw that since majority of the FRED data is quarterly along with the fact that most traditional financial reports are quarterly releases, we decided that standardizing our data to quarterly would make the most sense. Standardizing our data would eliminate unnecessary confusion for the model and helps us evaluate as necessary for the models we are using. We also noticed that not all of the data were collected during the same time frames. To accommodate for this, we decided to only collect data in which there was commonality amongst all the files for the dates so that we don't have a lot of null or zero value entries for data that the FRED didn't have on record for the specific variables that we are measuring in this case. What this resulted in was only 38 rows of data consisting of 9 predictors and 1 response variable for the model. We also had to add the previous quarter NASDAQ as another predictor since using the previous quarter helps predict what the next quarter would look like.

Although this made it easier for us to use for model selection and further analysis, having a smaller dataset only looking at quarterly averages could hurt the accuracy of each of the variables in the model. If there was data reported in more frequency, the quarterly average could hurt the accuracy of this. For example, having the NASDAQ averaged quarterly can hurt how accurate it is due to the volatility that each quarter could have within the stock market, in which we cannot capture in this scenario. If the data was reported less frequently, it could hurt the accuracy of these variables due to the fact that we don't have enough data to average out each of the variables quarterly. Since the data on more frequent reporting would not have existed, we would never know what the accuracy for that data could've been, which in turn could affect the results of our model.

When cleaning the data, we realized that there were a few variables which had a much higher order of magnitude than the rest. These were specifically the NASDAQ, Large Bank Consumer Credit Card Balances, and Quarterly Financial Reports. To solve this issue, we used a simple scaling method. The equation for this was $\frac{x_{ij}-\mu_i}{\sigma_i}$ for which $i = 1, .., 10$, the number of predictors and $j = 1, ..., 38$, the number of rows. This was done for every single data point within the three variables mentioned above. This process was done because when there are large differences in magnitude in the different variables we are using, it can lead to an ill posed data matrix. What happens is that this can generate larger errors when we are using the data to model, as we will see more in depth for the modeling process. Overall, normalizing the NASDAQ, Large Bank Consumer Credit Card Balances, and Quarterly Financial Reports had a drastic effect when modeling the data, as the accuracy different when running the different models as mentioned later.

3 Multicollinearity

Collinearity is a pertinent issue of any variable selection process, as the goal of producing any statistically significant regression model is selecting as many independent variables as possible that maximize the correlation between the predictors and the response. Highlighting the importance of this was a prerequisite for our model selection work, as we needed to understand the nature of the variables in order to move into the process of choosing an appropriate model.

To assess this, we decided to implement the Variance Inflation Factor (VIF) method. In essence, the VIF factor is a measure of how much the variance of a regression coefficient is inflated due to any multicollinearity in the model [3]. To elucidate, the variance between regressors is an important measure of independence, as more variance indicates less collinearity. However, in certain cases, the variance can be inflated (as the method is named), which can point towards issues in multicollinearity. The method works by regressing each predictor against every other predictor in the model, then plugging in the subsequent R^2 values into the VIF formula:

$$VIF = \frac{1}{1-R_i^2}$$

With this method, we were able to systematically diagnose multicollinearity issues and better understand the nuances of our data. With this in mind, we were able to move into the model selection process.

4 Model Selection

Making a prediction of this nature was clearly a problem to be solved with some sort of regression model. While we were cleaning our data we discovered that several of our explanatory variables had concerningly high collinearity, which could cause issues with a standard Linear Regression model. This suggested that it would be a good idea to use either a Ridge Regression or Lasso Regression model, as both techniques were suggested as ways of minimizing the impact of collinearity. We were unsure of which model would be more appropriate, so we decided to make both and compare their results.

We created these two models in Python, with the assistance of the sklearn package. Using cross validation, we were able to determine the optimal α values for each model. Some of the values in this report are different from those in the presentation, as we had not properly implemented cross validation at the time of the report. We found the ridge model had an α value of 0.248, while the Lasso model had an α value of 4.055.

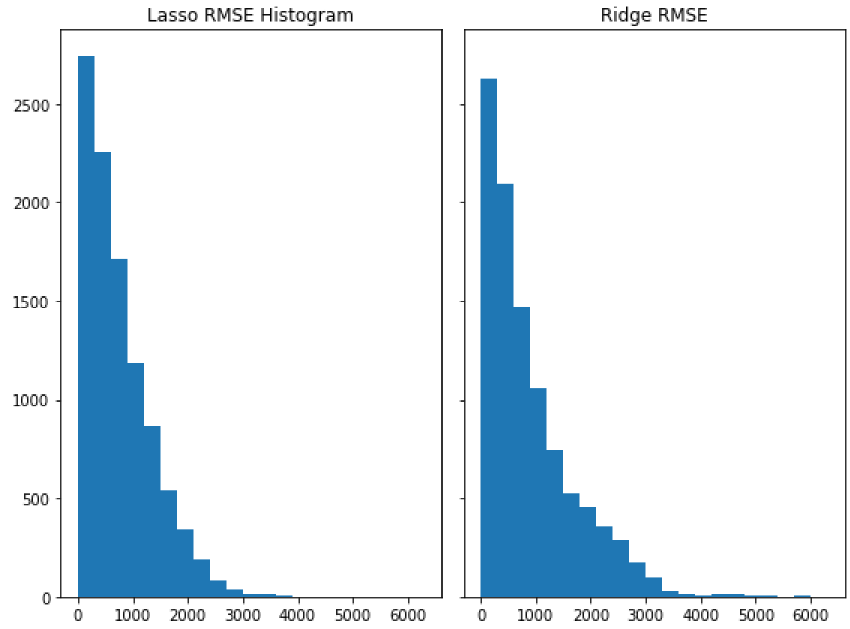
After creating the two models, we examined the weights they placed on each variable to see the differences in how the models were behaving. We noticed that the Lasso model removed Consumer Price Index for All Urban Wage Earners and Clerical Workers: Information Technology, Hardware and Services in U.S. City Average, but kept all of the other variables to some degree. The most impactful variables were NASDAQX and Consumer Price Index for All Urban Consumers: Information Technology, Hardware and Services in U.S. City Average. The coefficients of the Ridge Regression were much more balanced, all of their magnitudes were less than 400. This tells us that the Ridge model accounted for more of the variables, potentially being more impacted by covariance than the Lasso, which had a more dramatic feature selection. Below is a table of all the coefficients for reference.

Variable	Lasso Coefficient	Ridge Coefficient
Consumer Price Index for All Urban Consumers: Information Technology, Hardware and Services in U.S. City Average	-816.349	-389.269
Personal Savings	102.292	128.939
Consumer Price Index for All Urban Wage Earners and Clerical Workers: Information Technology, Hardware and Services in U.S. City Average	0.0	-388.736
Large Bank Consumer Credit Card Balances: Total Balances	-50.606	97.453
Commercial Bank Interest Rate on Credit Card Plans, All Accounts	-50.471	353.610
Net Percentage of Domestic Banks Tightening Standards for Credit Card Loans - Quarterly	-9.513	-39.714
Quarterly Financial Report: U.S. Corporations: Computer and Electronic Products: Net Sales, Receipts, and Operating Revenues	-253.505	-242.754
Velocity of M1 Money Stock	107.934	-22.663
NASDAQX	3339.007	-188.356

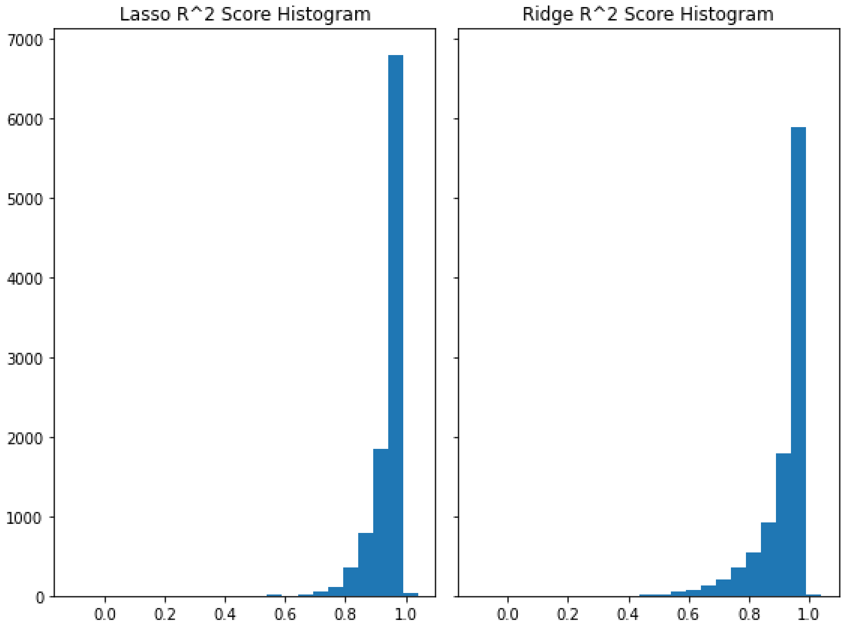
Upon analyzing the results of these two models, we found that the errors in predicting the test values ranged from values as low as 5 to as high as 2,700. Additionally, neither model seemed to be noticeably more accurate, sometimes the Ridge Regression produced a smaller error, and other times the Lasso Regression did. One obvious conclusion was that the models were much better at predicting the next quarter's average NASDAQ value when the value was closer around 4,000 to 8,000. The errors of over 1,000 points were often caused by trying to predict outlier prices that were greater than 10,000, which occurred during the pandemic. While these results were interesting, our testing data was made up of only 12 data points, which made it hard to assess larger trends.

The best solution we could come up with for this issue was to retrain both models with randomly partitioned training data sets 10,000 times. Each time we retrained our models, we took note of the RMSE, R^2 coefficient, average error, and the average positive and negative error. This helped us greatly to minimize the chances that our assessment of a model would be based on a test/train split that was overly influenced by outlier data points from the pandemic. Ideally we would have had a larger data set to work with that spanned back many more years to avoid this problem entirely, but that was unfortunately not possible.

First we assessed the average RMSE of Ridge and Lasso regression models. We found that over the 10,000 samplings, the average of the Lasso model's RMSE was 753.014, roughly 150 points lower than the Ridge model's RMSE, 901.509. Generating a histogram of each models' RMSE showed that both models' modes occurred around 0. The Lasso averages tapered off before 4,000, while those of the Ridge nearly passed 6,000.

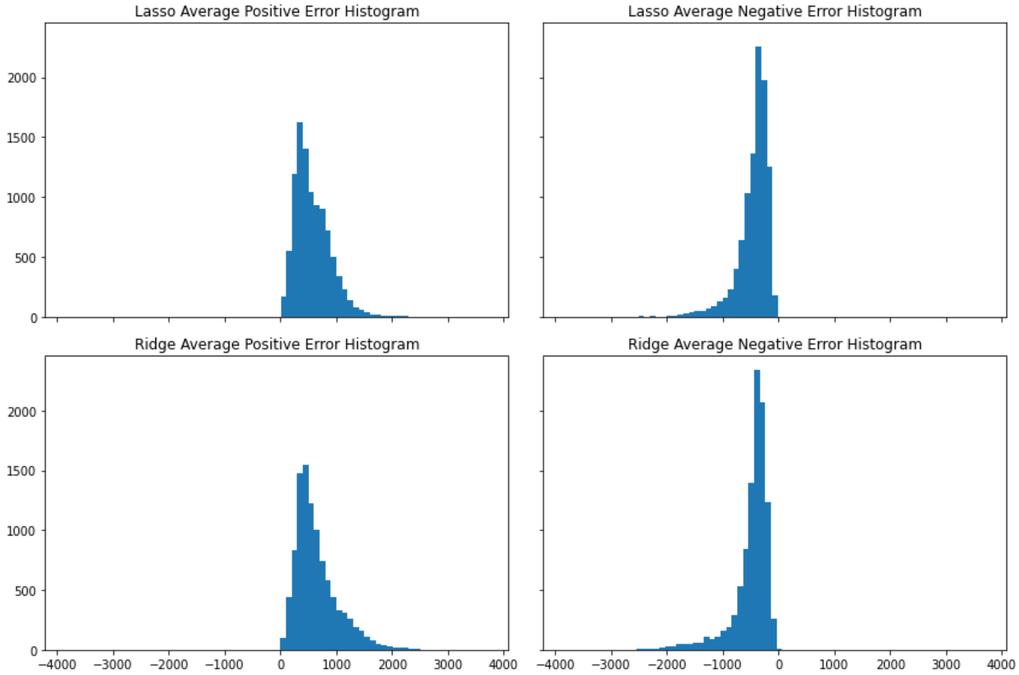


Next we assessed the R^2 scores of both models. We found that both models' modes were around 1.0, which is the best possible R^2 score. Following the trend in the RMSE values, we found that both models had tails heading towards 0, however, once again the tail of the Lasso model was shorter than that of the Ridge model. The tail of the Ridge model passed 0.5, while the Lasso model's did not.

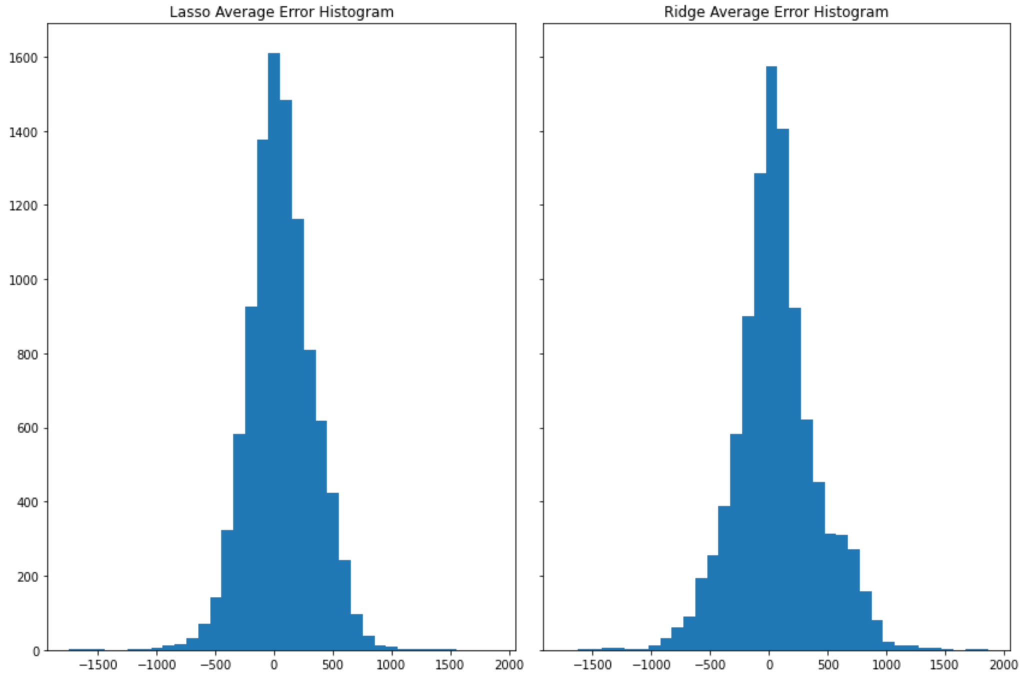


These two assessments suggested that the Lasso model was typically predicting with a higher accuracy than the Ridge model. However, these values don't tell us much about the nuances of the models' respective predictions. We still had questions, mainly whether one or both of the models tended to over or under predict. To answer this question, we continued our sampling and collected the error of each model 10,000 times.

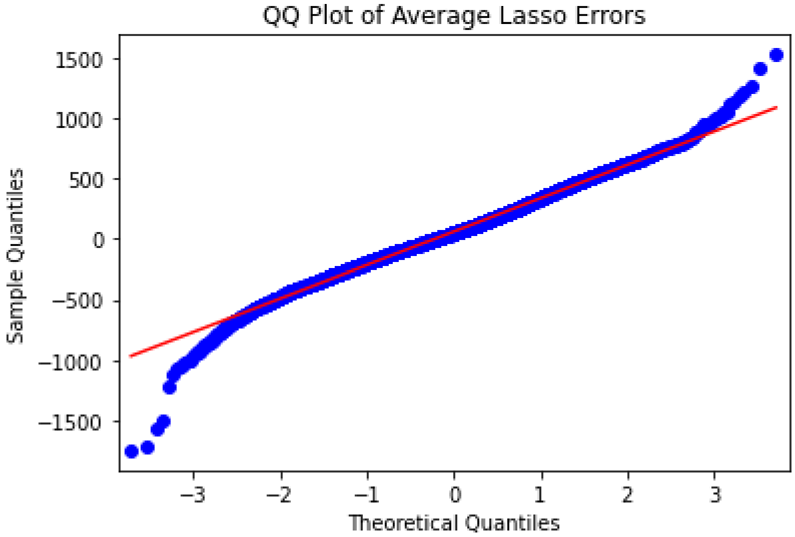
First we separated the average negative and positive errors. In this case, a positive average error corresponds to a model's average under-prediction, while a negative average error corresponds to a model's average over-prediction. A model that made more dramatic over or under predictions would be problematic, so we wanted to ensure that both of these values was as close to 0.0 as possible. When we plotted these values, we found that both models followed a very similar trend. Both model's average over and under estimates were clustered around 0.0, with tails extending away from 0.0. We can see that both models' tended to make over-predictions that were closer to 0.0, while they tended to under-predicted with a larger magnitude. In all cases average signed errors rarely exceeded 2,000.



Once we had confirmed that our models weren't making more over or under predictions, we assessed the average total error. Each average error was gathered over 12 quarters of test data. We found that the Lasso model had an mean average error of 59.473, while that of the Ridge model was 67.561. Overall when you consider the scale of the numbers we were predicting, a difference of roughly 8 points of error is quite small. We also calculated the standard deviations of both models' average error. The Lasso model's average error had a standard deviation of 275.633, while the Ridge model's average error had a standard deviation of 344.410. Again, these numbers are fairly close to one another. However, in combination with the previous results, this assessment convinced us that the Lasso model was superior to the Ridge model for our needs.

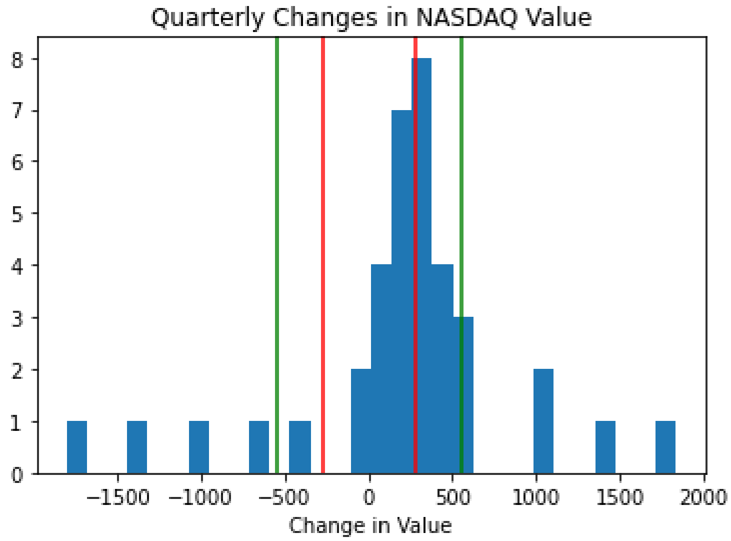


We noticed that the average errors of the Lasso model seemed to follow a normal distribution. To confirm this hunch, we created a QQ-plot, which confirmed that this was the case.



The fact that the errors followed a normal distribution allowed us to make the claim that roughly 68% if the Lasso models we generated were predicting with a long term average accuracy of roughly 275 points from the true value. Extending this further, we could claim that 95% of the generated models were predicting within 550 points of the true value on average. For these claims to be relevant, we then needed to find how the NASDAQ tended to behave quarter to quarter. If the NASDAQ tended to increase by less than 200 points each quarter, a model that predicted with over 200 points of error would be fairly useless. Inversely, if the NASDAQ tended to increase by over 600 points per quarter, our model would be quite useful.

We found that 62.162% of the observed quarters saw shifts in the NASDAQ's value of more than 275 points. About 24.324% of the observed quarters saw shifts in the NASDAQ's value of more than 550 points. In the histogram below, the values outside of the red lines represent quarters that saw values change by more than 275 points, and any quarters outside of the green lines saw changes of more than 550 points.



Thus we can say that in the long term, about 68% of the Lasso models we generated would have averaged favourable predictions. If you were to invest only when one of models within one standard deviation predicted a NASDAQ value of 275 points more than the current NASDAQ value, it would be a winning strategy. Of course there would be no way of knowing whether the model you were using was one of the 68% that this holds true for. When taking into account the relatively small data set we had access to, we feel that these results are promising. It could be worthwhile to expand the data set both in terms of the number of predictor variables and time frame to try to further increase the model's accuracy. Ultimately a model that can predict the NASDAQ score with an average accuracy of ±275 points would certainly be of use.

5 Conclusion

The model we have built is an exercise in showing our understanding of the concepts learned in this class and our desire to build a novel model rather than an applicable tool for the financial markets. The reason for this is because the financial markets are highly volatile and nondeterministic with millions of variables and possibilities that need to be considered and computed for any given model. Our model simply captures the behavior of a very small set of variables.

However, this is not to say that the model is not useful. The general framework, implementing a machine learning model to use alternative data to predict the NASDAQ index, is very applicable. The increase in the availability of data and rapid progression in technology has made these type of projects very useful.

In the future, increasing the number of variables, increasing back testing, and consistently refactoring the model to account for risk and tail events would ensure we have a more tenable model to eventually use for our own investment decisions, or potentially other investors' decisions.

References

[1] Alternative Data. *What is Alternative Data?* URL: https://alternativedata.org/alternative-data/.

[2] Federal Reserve Bank of St. Louis. *FRED Economic Data.* URL: https://fred.stlouisfed.org/.

[3] Statistics How To. *Variance Inflation Factor.* URL: https://www.statisticshowto.com/variance-inflation-factor/.