

**Investing in the Digital Age: Integrating Alternative Data Sources**

*Joel Esparza, Adam Benko, Subham Shaurav, Brian Kotin, Nilesh Choudhary*

*Gies College of Business, University of Illinois Urbana-Champaign*

*FIN580 Quantimental Investments*

*Professor Tony Zhang*

*5/5/2024*

***Abstract***

In the rapidly evolving landscape of artificial intelligence (AI) technology, companies spanning various industries are at the forefront of pioneering novel approaches to data analysis. This trend is particularly pronounced within the financial sector, where firms are increasingly integrating traditional financial metrics with alternative data sources to gain deeper insights into market dynamics. Alternative data encompasses a wide array of non-traditional datasets, ranging from web-scraped news articles and microblogs to satellite imagery, credit card transactions, and social media data. This enables firms to analyze traditional financial metrics with a whole new perspective. Amid this diverse landscape, our research is centered on User Generated Content (UGC) originating from Virtual Investing Communities.

By leveraging this alternative data alongside conventional financial indicators, our investigation seeks to unlock valuable insights into the potential impact of market sentiment on company performance. The proliferation of UGC within Virtual Investing Communities presents a unique opportunity to capture real-time opinions, sentiments, and discussions surrounding various stocks and financial instruments. Integrating this rich source of data with traditional metrics enables us to paint a comprehensive picture of market behavior and sentiment, empowering decision-makers to make more informed and timely investment decisions. Through our analysis, we aim to investigate the synergistic relationship between alternative data and conventional financial metrics, shedding light on new avenues for enhancing investment strategies and driving financial performance in an increasingly data-driven world.

Our research consists of two different trading strategies, one for short-term, high frequency trading and another for long-term trading. The short-term narrows in on the sentiment of news articles and their influence on the near future returns of stocks. We explore this relationship through different angles, as we split our analysis and trading strategies to datasets of companies in the Russell 2000, the S&P 500, the Nasdaq 100, and the Dow Jones 30. In addition to this, we further analyze each of these datasets at the daily, weekly, and monthly level. The long-term strategy attempts to quantify a key component in a company’s success, culture. This was done through the utilization of four different factor variables: innovation aggressiveness, leadership retainment, risk appetite, and culture emphasis. We explore the impact these factors have on companies’ annual returns for the ten largest companies in the market. While much work was devoted to this research, our investigation of the relationship between alternative data and traditional financial metrics only offers a glimpse of the true power that the combination of these two have when it comes to making smarter investing decisions.

***Short-Term, High-Frequency Strategy***

Traditional financial metrics are essential for investing decisions; however, while holding immense value, they do not fully grasp all necessary details when it comes to choosing those companies to invest in. A key component that is not captured by these traditional metrics is market sentiment. Market sentiment can be described as the overall attitude or outlook of investors and traders towards a particular asset, market, or financial instrument. Each day, users across a wide variety of virtual investing communities share their thoughts and advice on the stock market. For retail investors, this shared information is crucial for investing decisions, as it offers them key insights from “experts” of their respective news sources. This fluctuates market sentiment for each company, as these retail investors will undervalue or overvalue the company’s value of their shares daily. We aim to capitalize on this noise through the utilization of artificial intelligence, big data for UGC, and statistical applications.

**Data Collection & Preparation**

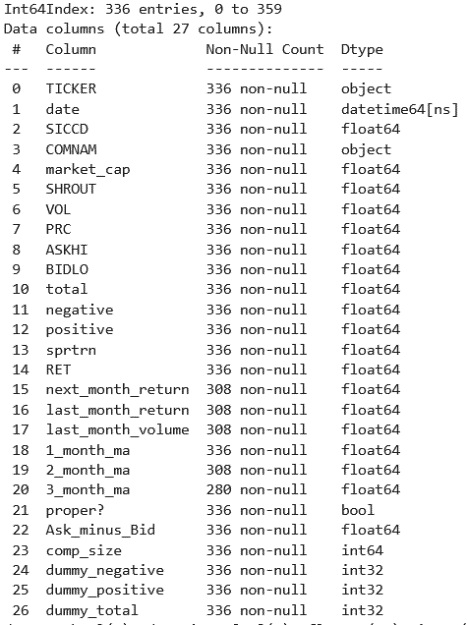
The UGC that we decided to focus on throughout this short-term strategy were news articles written by a wide variety of different authors that have been collected by Capital IQ. This has all been compiled into a dataset and was accessible to our team through the Wharton Research Data Services. In this dataset, each observation is a specific headline that was declared by Capital IQ as a “Key Development”. It contains columns such as date, the company the headline was about, the text of the headline, and other characteristics regarding the observation. Due to the large size of this data, we were restricted to only a year’s worth of data (computational limitation), which was 1,942,227 observations. We then further filtered for only observations that correspond to companies included in the Russell 2000, resulting in 230,934 observations. These were all observations for the year 2023.

While the text of the headlines was included in this data, we were unable to identify the sentiment that was reflected in these headlines. To do this, we incorporated the OpenAI python package to conduct a sentiment analysis on each headline that was included in the dataset. By doing so, artificial intelligence was able to classify each headline as either positive, negative, or neutral. We then appended these results to our dataset and proceeded further. From the Compustat data services, we were able to attain a dataset of the daily stock returns and other relevant metrics for the Russell 2000 listed companies in the year 2023. When joining the data together, we ended up with ~490,000 observations, each indicating a unique date and ticker combination for the year 2023. All data from the daily stock dataset maintained the same, while we performed the following aggregation on our headlines data: sum the number of negative, positive and neutral headlines grouped by date and ticker. This resulted in three new columns that identifies the total count of positive, neutral, and negatively written articles about a company on a given day.

Furthermore, once we completed the dataset of daily metrics for the Russell 2000 companies, we then expanded our horizons and created numerous subsets of this data. A dataset was created to only include companies in the S&P500, as well as for the Nasdaq 100 and the Dow Jones 30. The number of observations for these subsets were ~125,000, ~25,000, and ~7000, respectively. After successfully creating subsets for daily stock data of each of these indexes, we then created an additional two subsets for each daily dataset we had. One of these subsets was an aggregation of weekly data and the other was an aggregation of monthly data. To do this, we grouped by week/month and ticker and performed the following aggregations: average market capitalization, average volume, average shares outstanding, average price, maximum ask high price, minimum bid low price, sum of total news articles, sum of total positive news articles, sum of total negative news articles, cumulative returns of the S&P500 during this time period, and the cumulative returns of the specified ticker during this time period. After performing these groupings and aggregations, we ended up with our Russell, S&P500, Nasdaq, and Dow Jones datasets having the following totals of weekly observations: ~104,000, ~26,000, ~5,200, and ~1,560 observations respectively. The totals of monthly observations were: ~24,000, ~6,000, ~1,200, and ~360 observations, respectively.

Additional columns were then added to each of these subsets. These columns include historical and future values, such as previous period’s returns and volume and the return of the next period. Moving averages were then calculated. For daily, these moving averages were calculated for the previous 10 days, 20 days, and 30 days. For weekly, these moving averages were calculated for 2 weeks, 3 weeks, and 4 weeks. And for monthly, these moving averages were calculated for 2 months and 3 months. A dummy variable names proper was then created that indicated whether these moving averages were in proper form (i.e 10 day moving average > 20 day moving average > 30 day moving average) to indicate a stocks momentum. Three other dummy variables were created to indicate whether the company received at least one negative article written about it, at least one positive article, and at least one article written about it. The last two columns added were a categorical variable to classify the company’s size with respect to the other companies included in the index and a column that contains the difference between the ask high and bid low price for the respective period. Numeric variables in this dataset were then normalized using a z-score method to satisfy the normal distribution assumption for regression analysis. After doing so, these datasets were finalized for analysis. The table below displays the dataset structure for the Dow Jones monthly stock data; however, all columns contain the same information, just with a different number of observations.

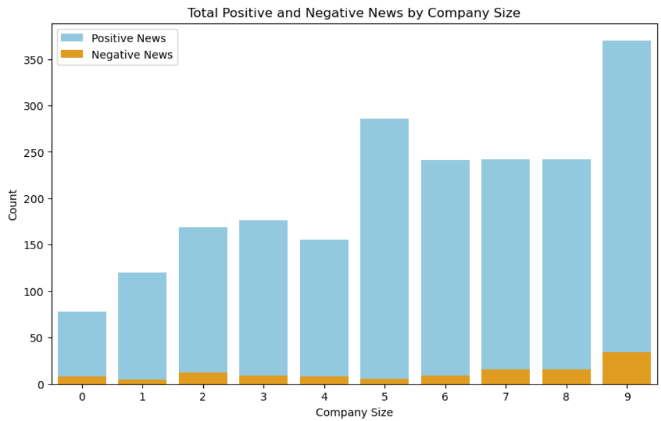
**Table 1: Short-Term Strategy Dataset**



**Data Exploration**

When exploring each of these datasets, we were able to identify many similarities amongst these datasets. A common theme was that the larger companies attract a higher volume of news articles in every dataset. This would only make sense as these larger companies are much more publicly known and therefore attract the attention of retail investors to a much larger scale, thus leading to authors to lean towards covering news centered around larger companies. Looking at the figures below, we can better see this relationship. Figure 1.1 has the size of the company ranked 0-9, with 9 being the 200 companies in the Russell 2000 with largest market capitalization and 0 being the 200 companies in the index with the smallest market capitalization. We can identify that most of these news articles are being written for larger companies. Figure 1.2 displays this same exact relationship but narrowing in on companies included in the Nasdaq. We see that most of these articles are about companies with higher market capitalization, especially those in the top 10.

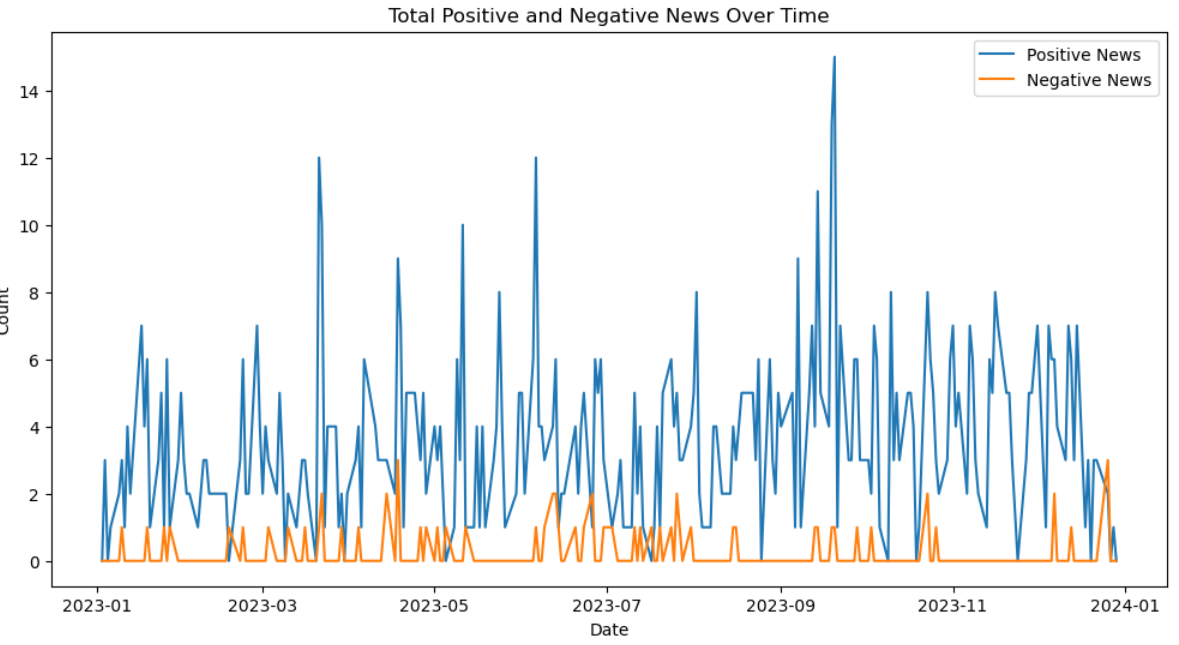
**Figure 1.1: Russell Distribution of Daily Articles**   **Figure 1.2: Nasdaq Distribution of Daily Articles**

Figure 1.1: Russell 2000 Distribution of Daily New Articles by Company Size


Another common theme identified in our analysis were the periods in the calendar year that experienced the most news articles. When analyzing the time trends these numbers experience, we identified spikes near March, June, and October in most of our datasets. This would make sense, as it is in these months when the quarterly earnings reports are typically released to the public. With investors being given the performance updates of their respective companies, there is much noise during these periods. Figure 2.1 shows this trend over time. We can identify the three spikes during these periods to display this relationship.

When it comes to the industries that are attracting the greatest news volume, the five industries extracting the majority news articles were SIC Code 283: Drugs, SIC Code 737: Computer Programming, Data Processing, and Related Services, SIC Code 602: Commercial Banks, SIC Code 367: Electronic Components and Accessories, SIC Code 481: Telephone Communications, except Radiotelephone. In addition to this, we were able to identify that those industries that experience the highest returns in their respective periods, are also the industries that contain a higher number of total articles per company as well as a higher number of total positive articles per company in their respective periods

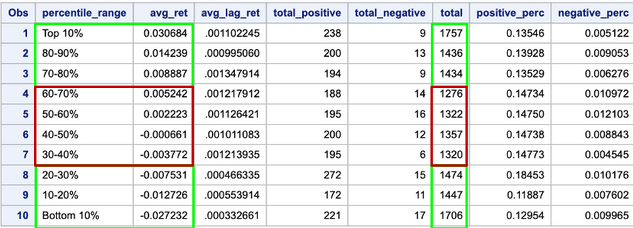
**Figure 2.1: Dow Jones News Daily Volume Trends in 2023**



After gaining a better understanding of our data, we then began to take a deeper dive into the primary focus: returns. Our approach was to better understand this relationship from the perspective of multiple angles. The first was through a grouping of performances based off the respective periods. For each of the datasets, we aggregated a new column that placed each observation into deciles, depending on the ranking of their period’s return. For all datasets, the following pattern appeared: More attention is drawn to those companies that have received relatively high return values or relatively low return values. his would make sense, as retail investors are interested in hearing news regarding those companies that performed well or those that performed poorly. This relationship can be displayed in Table 2, showing us the results for the Nasdaq100 daily returns. We see the value in the total column is higher for the percentile ranges at the high and low ends, indicating more noise is drawn to more extreme values.

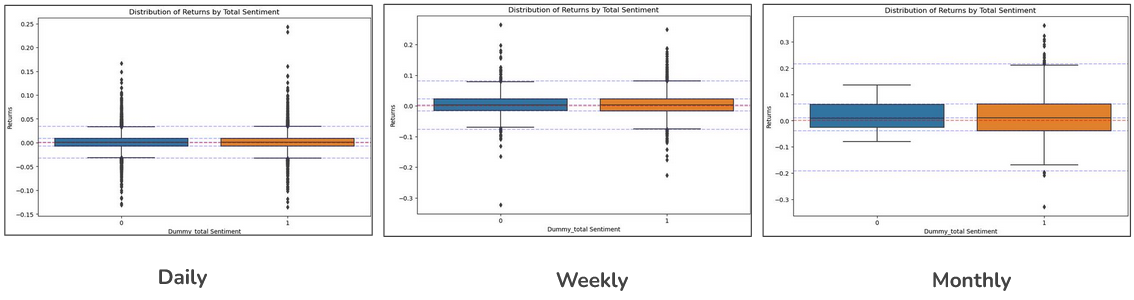
Furthermore, when grouping our observations by our dummy variables: “dummy\_total”, “dummy\_negative”, and “dummy\_positve”, we were able to identify another insightful piece of information. When focusing in on our Ask\_minus\_Bid column (value that represents the difference between the max price in the period and the min price in the period), we conducted statistical tests to determine whether there are differences between the mean and volatilities between those companies with the dummy indicator compared to those without the dummy indicator. Performing these statistical tests, while not significant at the 5% level for several of the datasets, we were able to identify the general relationship that those companies with the dummy indicator experienced greater differences in their ask high and bid low prices, and experienced higher variance amongst these differences. This tells us that the perceived value of these financial instruments tends to vary significantly amongst buyers and sellers in the market when news articles are being written about them. This exemplifies the claim that this form of USG influences retailer opinions about companies to either undervalue or overvalue the stocks' true price.

**Table 2: Nasdaq 100 Daily Returns Deciles News Distribution**



Although this information offers us valuable insights, it only informs us on the relationship of USG and stock returns regarding the same periods. We became more interested in analyzing the relationship that these news articles have with future returns for companies. More specifically, we wished to determine to what extent does this alternative data contributes to the predicting power of future period returns. To do this, we utilized the three dummy variables once again, but this time focusing on future returns instead of current. While all datasets display different values and insights, there was once again a common theme extracted from each analysis. When comparing the next period returns at the daily level, we see the distribution of returns is very similar between those with the dummy\_total indicator (at least 1 news article in the given day) and those without the dummy\_total indicator. As we shift our attention to weekly cumulative returns, we were able to see the distribution for the dummy\_total indicator (for weekly, switched on if at least 2 articles were written in the given week) contains a slightly wider distribution of next week cumulative returns. And lastly, as shift our attention to monthly cumulative returns, we see this relationship becomes much more apparent, as the distribution for the monthly dummy\_total indicator (10+ articles in given month) is significantly wider than those without the indicator. This tells us that those companies that receive x amount of news articles in a specified period, not only tend to have more extreme returns on the day of, but also experience more extreme returns in the period immediately after as well. This information will prove to be of great value in our analysis; however, it does not give us any sort of prediction power in the directional movement of stocks. To determine that information, we constructed several statistical models.

**Figure 3.1: Nasdaq Distribution of Next Period’s Returns by News Dummy Indicator**



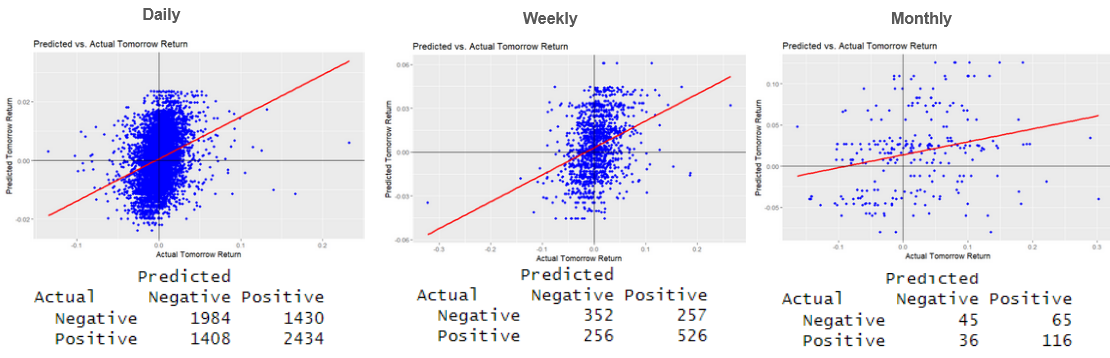
**Predicting Directional Movement for Returns**

Once we have gathered these insights, we then began to believe that if we were able to improve the amount of prediction power we had for the directional movement of returns, we believe that we would be able to capitalize on these extreme returns that are generated from those companies that received a higher volume of news articles. For us to achieve this, we applied three different models to offer us greater prediction power: Lasso Regression, Random Forest, and Logistic Regression. Lasso and the random forest were models that we chose due to the way they conduct variable selection on the variables to filter out irrelevant factors. These models were constructed to predict future returns. Our logistic regression was constructed in a way to predict the directional movement of the next period’s returns. These models were constructed uniquely for all 12 of the datasets. Although each model is unique, the process of creating these models was similar in every aspect. In addition to this, all four indexes experienced similar patterns when transitioning from daily, to weekly and monthly. All three models for each Russell 2000, S&P 500, Nasdaq, and Dow Jones witnessed more prediction power as we transition from daily to weekly to monthly. To elaborate further on these models, we will describe the process using the Nasdaq datasets; however, this is the same process and has similar results to that of the other indexes.

Before these three models were built, we first identified columns that are highly correlated. We proceeded by removing these variables from our models. For the Lasso Regression, each of the daily, weekly, and monthly datasets were partitioned into a training and testing set, with 70% of observations in the training set and 30% in the testing set. After doing so, 5-fold cross validation was applied to this model to capture the optimal lambda to use in our model. Once we attained the best model, we used this model to predict the future returns for all datasets and appended these values into a column named “lasso\_pred”. The model does a terrible job at predicting the actual return value, as this model has an adjusted R-squared of 0.015; however, we are focused on the directional movement and not the actual return. When analyzing confusion matrices of the Nasdaq, we were able to see greater prediction power in the monthly models, and then weekly, and lastly the daily. The daily model was able to accurately predict positive returns in the testing set 53.66% of the time and 50.17% of the time for negative returns. Both numbers increase for the weekly model, with 56.38% and 54.7% respectively. And once again, as we move to our monthly model, we once again see these numbers increase, with 65.56% and 68.62% accuracy.

Moving on to the random forest, we were able to yield much higher accuracy percentages. The data was trimmed to remove any outliers located in the top and bottom 2.5% of their respective distributions. After doing so, the data was once again partitioned into 70% of observations in a training set and 30% into a testing set. Due to computational limitations, we performed manual cross validation by testing the model with different combinations of parameters until we identified the best one. We found a random forest with a max depth of 30 and a cp of 0.0001 results in the best model performance. The random forest model that was constructed for the Nasdaq daily returns resulted in 63.35% accuracy in accurately predicting positive returns and 58.11% in predicting negative returns for the testing set. For the weekly model, we see the accuracy of positive returns increase to 67.26% while the accuracy of negative returns decreases to 57.80%. For the monthly model, we see this pattern continue as the accuracy of positive returns increases to 76.82% while the accuracy of negative returns decreases to 40.91%. While the random forest model does a tremendous job in accurately predicting the directional movement of future returns, it also had an extremely low adjusted R-squared. The results of the random forest are displayed in Figure 4.1.

**Figure 4.1: Nasdaq Random Forest Prediction Analysis**



The final model that was implemented was a logistic regression to predict the directional movement of the future period’s returns. This model is designed to predict binary outcomes, so a new column was added to our dataset that had a value of “Positive” or “Negative” to indicate the actual directional movement of the return for the following period. Once this column was created, the data was once again partitioned into 70% of observations in a training set and the remaining 30% to the testing set. Cross validation was then applied to this model using 10-folds. After cross-validation was performed, we identified those significant variables. Noticeably common variables across majority of all 12 logistic regression models included returns of the present period, company size, “proper” dummy variable regarding moving averages, and volume. For the monthly returns, we found the returns of the S&P500 to be highly significant with a positive correlation. For the Nasdaq, the daily logistic model was able to accurately predict positive future returns 53.36% of the time and 51.28% of the time for negative future returns. Both values increase as we move towards our weekly model, 57.28% and 53.09% respectively. As we shift further to our monthly returns, we once again see these values increase, to 70.36% and 66.08%. Now that we have created three models for all twelve datasets, we appended the predictions of these models to their respective datasets.

**Implementation of Trading Strategies**

Now that we have a thorough understanding of the relationship of UGC, in the form of news articles, and the returns of stocks in the market, we combine these insights with the prediction models that were constructed to implement potential trading strategies for the year 2023. During this year, the S&P500 yielded 21% annual returns. We will use this number as the benchmark to compare our trading strategies. For each of these strategies, the initial investment amount of $100,000 is systematically allocated among all companies meeting the prescribed conditions for taking a position (either 1 for a long position in the next period, 0 to do nothing in with this stock in the next period, or -1 for a short position during the next period) during the specified period. This allocation ensures an equitable distribution of funds across eligible companies, enabling a balanced approach to investment strategy execution. For each date within the designated period, the code identifies the companies with positions of 1 or -1 and divides the total investment amount equally among them. This means that each company with a position of 1 or -1 receives an identical portion of the initial investment, facilitating a uniform implementation of the trading strategy.

Following the allocation of funds, the code computes the returns for each company based on their respective positions and subsequent market performance. For companies with a position of 1 (representing a long position), the code calculates the return by multiplying the next week's return for that company by the investment per company. Conversely, for companies with a position of -1 (indicating a short position), the code calculates the return using the next week's return but with a negative sign to reflect the short position. By multiplying these returns by the investment per company and summing them across all companies with positions of 1 or -1, the code derives the total return for the strategy on each date. This calculated sum of daily returns will then be used as the initial investment at the beginning of the next period, and the process restarts. This is iterated for all periods in the dataset in sequential order. This dynamic approach to investment allocation and return calculation enables the strategy to adapt to market conditions and capitalize on opportunities while mitigating risks associated with short-term fluctuations.

When implementing these models, there were common trends for each of the twelve datasets. In all datasets, we find that utilizing a combination of our predictive models tends to lead to an increase in the amount cumulative returns that is generated by each strategy. In addition to this, we find that adding a criterion for total news articles or the amount of total positive sentiment articles also tends to increase the total cumulative returns of the strategy. For these reasons, we find that nearly all of our strategies tend to outperform the S&P500 benchmark of 21% annual returns. In addition to this, we find that our strategies that reinvest daily outperform those similar strategies that reinvest weekly, and both of those tend to outperform those strategies that reinvest monthly. Something we discovered was that including the random forest predictions in our criteria results in a strategy that yields remarkable returns. For this reason, much caution was brought to these implementations, as we believe that the random forest is exposed to bias. Both the lasso and the logistic provide reasonable returns.

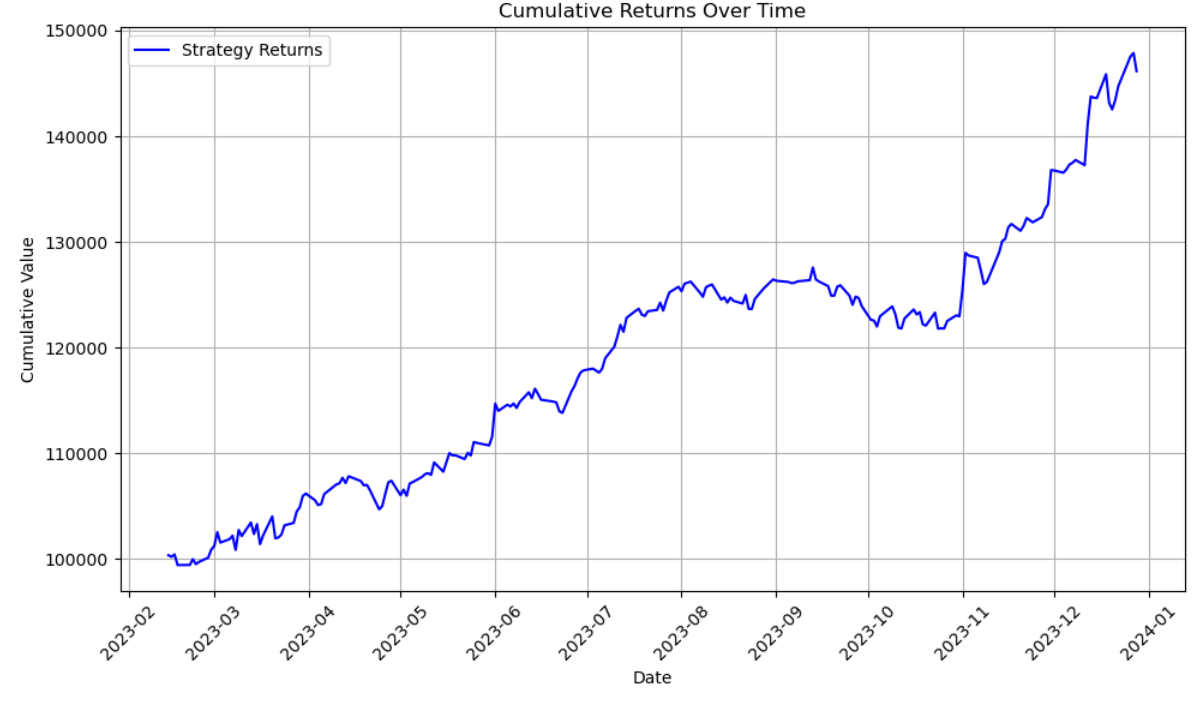
For Dow Jones, the lasso regression and news dummy on their own was able to achieve 24% when reinvesting daily, 24% when reinvesting weekly, and 17% when reinvesting monthly. The trading strategy with only the logistic regression model and news dummy was able to achieve 35% when reinvesting daily, 17% for weekly, and 25% when reinvesting monthly. For the random forest, this strategy achieves 168% cumulative returns when reinvesting daily, 65% returns when reinvesting weekly, and 21% returns when reinvesting monthly. When we combine the lasso, random forest, and news dummy, our strategies yield 150% for daily, 69% for weekly, and 26% for monthly. For our strategy using lasso, logistic, and our news dummy, we find our strategy yielding 32% for reinvesting daily, 28% for reinvesting weekly, and 28% when reinvesting monthly. When implementing the logistic regression, random forest, and news dummy, our strategy yields 210% cumulative returns when reinvesting daily, 73% when reinvesting weekly, and 37% when reinvesting monthly. Finally, when utilizing all three models as well as our news dummy variable, we find our strategy yielding 158% returns when reinvesting daily, 71% returns when reinvesting weekly, and 34% returns when reinvesting monthly.

For the Nasdaq, the lasso regression and news dummy on their own were able to achieve 36% when reinvesting daily, 25% when reinvesting weekly, and 29% when reinvesting monthly. The trading strategy with only the logistic regression model and news dummy was able to achieve 31% when reinvesting daily, 21% for weekly, and 24% when reinvesting monthly. For the random forest, this strategy achieves 260% cumulative returns when reinvesting daily, 72% returns when reinvesting weekly, and 29% returns when reinvesting monthly. When we combine the lasso, random forest, and news dummy, our strategies yield 390% for reinvesting daily, 81% for reinvesting weekly, and 38% for reinvesting monthly. For our strategy using lasso, logistic, and our news dummy, we find our strategy yielding 40% for reinvesting daily, 29% for reinvesting weekly, and 28% when reinvesting monthly. When implementing the logistic regression, random forest, and news dummy, our strategy yields 390% cumulative returns when reinvesting daily, 80% when reinvesting weekly, and 39% when reinvesting monthly. Finally, when utilizing all three models as well as our news dummy variable, we find our strategy yielding 380% returns when reinvesting daily, 80% returns when reinvesting weekly, and 41% returns when reinvesting monthly.

For the S&P 500 Dataset, the lasso regression and news dummy on their own were able to achieve 24% when reinvesting daily, 22% when reinvesting weekly, and 28% when reinvesting monthly. The trading strategy with only the logistic regression model and news dummy was able to achieve 35% when reinvesting daily, 25% for weekly, and 27% when reinvesting monthly. For the random forest, this strategy achieves 166% cumulative returns when reinvesting daily, 128% returns when reinvesting weekly, and 48% returns when reinvesting monthly. When we combine the lasso, random forest, and news dummy, our strategies yield 150% for reinvesting daily, 136% for reinvesting weekly, and 60% for reinvesting monthly. For our strategy using lasso, logistic, and our news dummy, we find our strategy yielding 32% for reinvesting daily, 26% for reinvesting weekly, and 35% when reinvesting monthly. When implementing the logistic regression, random forest, and news dummy, our strategy yields 210% cumulative returns when reinvesting daily, 138% when reinvesting weekly, and 60% when reinvesting monthly. Finally, when utilizing all three models as well as our news dummy variable, we find our strategy yielding 157% returns when reinvesting daily, 140% returns when reinvesting weekly, and 63% returns when reinvesting monthly.

For the Russell 2000 Dataset, the lasso regression and news dummy on their own were able to achieve 44% when reinvesting daily, 28% when reinvesting weekly, and 31% when reinvesting monthly. The trading strategy with only the logistic regression model and news dummy, was able to achieve 41% when reinvesting daily, 24% for weekly, and 30% when reinvesting monthly. For the random forest, this strategy achieves 480% cumulative returns when reinvesting daily, 158% returns when reinvesting weekly, and 65% returns when reinvesting monthly. When we combine the lasso, random forest, and news dummy, our strategies yield 485% for reinvesting daily, 180% for reinvesting weekly, and 83% for reinvesting monthly. For our strategy using lasso, logistic, and our news dummy, we find our strategy yielding 46% for reinvesting daily, 31% for reinvesting weekly, and 34% when reinvesting monthly. When implementing the logistic regression, random forest, and news dummy, our strategy yields 520% cumulative returns when reinvesting daily, 175% when reinvesting weekly, and 82% when reinvesting monthly. Finally, when utilizing all three models as well as our news dummy variable, we find our strategy yielding 480% returns when reinvesting daily, 176% returns when reinvesting weekly, and 84% returns when reinvesting monthly. Figure 5.1 displays the daily trading strategy that was implemented for the Russell 2000 daily dataset. The position of stocks is 1 (buy at market open tomorrow) if the lasso and logistic regression predict positive returns tomorrow & if the company had a news article written about it. The position is –1 (short at market open tomorrow) if both the lasso and logistic regressions predict negative returns tomorrow & if the company had a news article written about it. If this criterion does not hold true for an observation, then the position is 0 which indicates a neutral position.

**Figure 5.1: Russell 2000 Logistic, Lasso, and News Dummy Daily Trading Strategy**



**Conclusion**

As mentioned in the beginning of this section, there is much concern regarding the random forest. We believe that constructing random forest with a high complexity, as cp = 0.0001, resulted in a highly biased model. The only way to know for sure would be to continue our research and used this model for data outside of the year 2023. Something that can be done in our research's future would be to implement this model for current metrics (2024) to see how the strategy would perform if it were to be implemented now. In addition to this, some other noteworthy constraints that were present throughout the research were computational, financial, and time limitations. In terms of the computational limitations, the dataset of the news articles that was compiled by Capital IQ contained 20+ million observations since the year 2000. Due to this dataset's size, we were unable to use it, resulting in us being limited to only 2023. If we were to have computational power, we would have been able to better understand this true relationship and have more reliable statistical models. In terms of financial limitations, iterating the Open AI function in python for every observation was costly. If more money were invested, we would have been able to include more observations in our research. In addition to this, access to other data of UGC is available online but costs money. If we could include more UGC factors, we would be able to better understand true market sentiment surrounding each company in the market. And lastly, time constraints. There are infinitely many ways to analyze our datasets and gather a better understanding of the relationship between UGC and stock returns. There are also infinitely many ways that we could position our stocks in our trading strategies. If time was not a constraint, we could have dug deeper and explored including more factors in our criteria that were used to position stocks in our trading strategies.

Regardless of these limitations, we were able to find a great deal of insights throughout our research. A key point that was found was that the attention of news articles is drawn to companies that are experiencing extreme returns relative to the specified period. Another key finding was that the distribution of future returns for those companies that experience a relatively high volume of news articles is larger than those that do not receive such a volume of articles. And the biggest takeaway was that although the random forest was exposed to much bias, we believe that the combination of traditional financial metrics as well as alternative data enables us to make smarter investing decisions. When only looking at our strategies that include lasso and/or logistic regressions, we were able to reasonably outperform our annual benchmark of 21% consistently. Since these trading strategies had relatively simple criteria in our investing positions, we believe that further research can be done to achieve higher returns. As for now, we conclude that the power of integrating alternative data with traditional financial metrics will be crucial in the financial industry for years to come, as we enter this new age of artificial intelligence.

***Long-Term Strategy***

At the end of the day, what drives a company’s successes are the people that are devoting hours upon hours of time towards ensuring their respective companies strive. Steve Jobs made the following remarks about the employees of a company, “Innovation has nothing to do with how many R&D dollars you have. It is about the people you have, how you are led, and how much you get it.” This statement exemplifies the importance of hiring the right people and establishing a shared set of values, goals, attitudes, and practices. This can be referred to as the company’s culture. Maintaining a sustainable and strong company culture is one of the company’s best interests as this often leads to great long-term success. If we were able to quantify a company's culture, we can potentially capitalize on this information and yield greater returns in our trading strategies. There are many ways to quantify such a broad term; for this reason, we have created four factors to represent various aspects of the company’s culture. These aspects include innovation aggressiveness, leadership retainment, risk appetite, and cultural emphasis. This analysis was performed on the companies with the 10 largest market capitalizations from 2014-2023.

**Data Collection & Factor Creation**

We took a comprehensive approach to data collection, sourcing information from multiple high-quality databases to capture a wide spectrum of metrics that we felt would quantify culture. We utilized Capital IQ to collect key developments and headlines that highlight significant events and innovations affecting the top 10 firms by market capitalization. Yahoo Finance provides daily stock market metrics, crucial for analyzing financial performance over time. COMPUSTAT was used for its detailed executive compensation data, offering insights into the tenure of top executives. Additionally, 10-K filings were analyzed to gauge the emphasis companies place on their corporate culture, providing a textual analysis base for understanding qualitative aspects of corporate governance.

Let us look at data collection and processing methodology of each of the factors used:

1. **Innovation Aggressiveness:** This factor assesses the extent to which a company encourages and embraces innovation. It involves evaluating the organization's approach to introducing new ideas, products, or processes, and its willingness to take risks and invest in research and development. Companies with high innovation aggressiveness may prioritize experimentation, foster a culture of creativity, and actively seek out opportunities for disruptive innovation in their industry. This factor was created using data from Capital IQ.  
     
   CapitalIQ Key Development dataset provides a quantitative measure of the company's innovation activity and its commitment to driving change and progress within the industry.  
     
   Steps :

* Find the count of articles in the Capital IQ Key development database for each year.
* This data is saved as “ART\_COUNT” column.

1. **Leadership Retainment:** This factor focuses on the organization's ability to attract and retain talented leaders. It considers the strategies and practices in place to identify, develop, and retain effective leaders within the company. Factors such as leadership development programs, succession planning, and employee satisfaction levels may be assessed to gauge leadership retainment. Companies that excel in this area often have strong leadership pipelines, low turnover rates among key executives, and a supportive and inclusive leadership culture. We used COMPUSTAT Executive Compensation Dataset to find the average executive tenure of current executives. This feature is stored as “TENURE” in our dataset.

Steps:

* Download the start date and end date for all executives each year for a company.
* Use the start date and end date to calculate TENURE for each executive for each year.
* Calculate the average of all executive TENURE for each year for each company.
* This feature is stored as TENURE\_MEAN.

1. **Risk Appetite:** This refers to the level of risk that an organization is willing to accept in pursuit of its objectives. It involves evaluating the organization's tolerance for uncertainty and its willingness to take calculated risks to achieve strategic goals. Factors such as the organization's risk management processes, decision-making frameworks, and past risk-taking behavior may be considered when assessing risk appetite. Companies with a high-risk appetite may pursue ambitious growth strategies, invest in emerging markets or technologies, and embrace innovation despite potential uncertainties. We have used Income Stability Ratio as a proxy for this factor. The income stability ratio provides insights into the financial stability of the company and its ability to withstand economic fluctuations. A higher income stability ratio may indicate a lower risk appetite, as the company prioritizes stability and consistent performance over riskier ventures.

Steps:

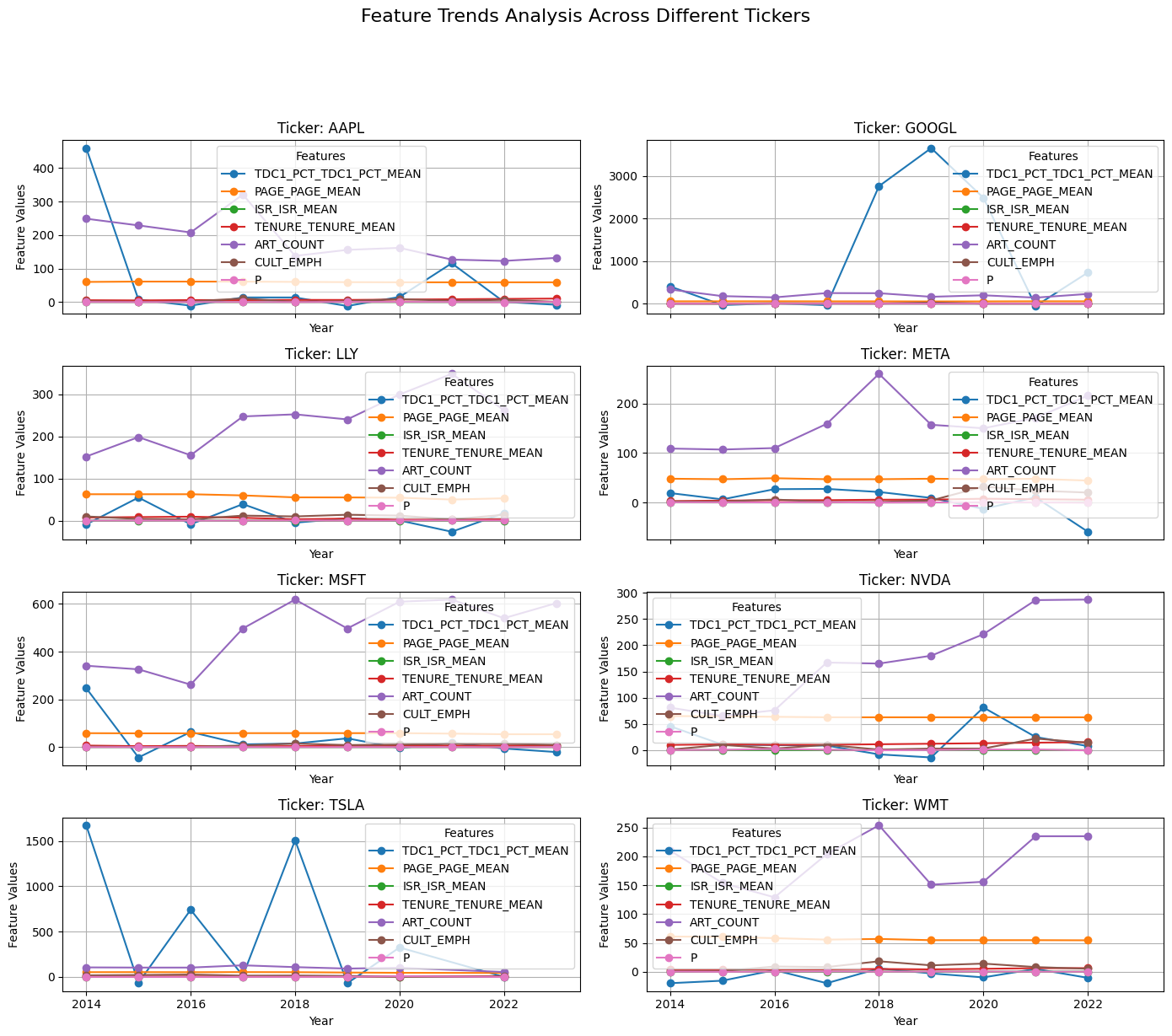
* ISR = (Salary / Total Compensation) \* 100
* Calculate this number for all the executives for every year for every company
* Take the mean per year per company to generate “ISR\_MEAN”
* Apart from this we also store mean total direct compensation for each company for each year as TDC1\_MEAN to complement the ISR values.

1. **Culture Emphasis:** This factor examines the importance placed on organizational culture and values within the company. It involves assessing the extent to which the organization's culture aligns with its stated values and strategic objectives. Factors such as leadership behavior, communication practices, and employee engagement initiatives may be evaluated to understand the emphasis placed on culture. Companies with a strong culture emphasis often prioritize values such as transparency, integrity, and collaboration, fostering a sense of purpose and belonging among employees and driving performance and innovation.  
     
   Steps:  
   - Collection of 10-K filings for all 10 companies for the past 10 years (2014-2023): The 10-K filings provided detailed insights into the company's operations, financial performance, and strategic priorities.  
   - Scraping of the next 10,000 words from the 4th page of each 10-K filing: This text data was then processed using an OpenAI function. We counted the number of sentences which talk about culture in the extracted text. Despite encountering mildly inconsistent results, this process allowed for textual analysis to investigate the emphasis placed on organizational culture and values within the company.

- This feature is saved as “CULT\_EMPH” for each year for each company.

**Data Exploration:**

Visualizing variation of features across time for each of the companies.

**Figure 6.1: Time Trends of Four Long-Term Factors**  
 

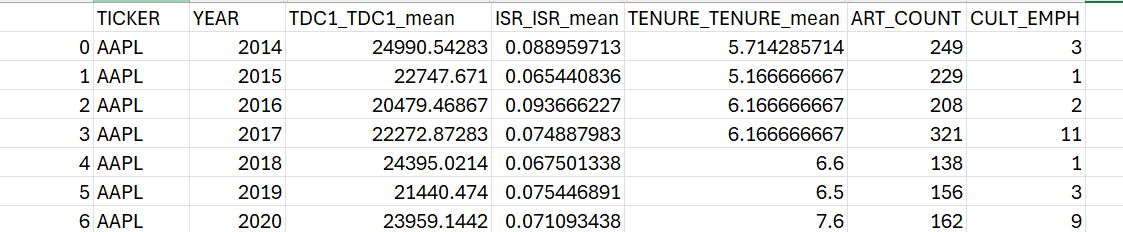
Most companies show variability in the metrics over the years, with no single trend dominating across all features and companies. This suggests that company performances and strategies are influenced by many factors that change over time. AAPL shows a more stable trend across features with slight year-to-year fluctuations, indicating consistent management practices and possibly a steady market environment. GOOGL and TSLA exhibit significant spikes in certain years for metrics like TDC1\_PCT\_MEAN and PAGE\_PAGE\_MEAN, which might indicate specific years of high activity or strategic shifts. MSFT and WMT display moderate fluctuations, which could suggest adaptive strategies in response to market conditions or internal changes.

The Capital IQ Key Development dataset offers insights into companies' strategic focuses, revealing differing levels of emphasis across dimensions. ART\_COUNT, reflecting Innovation Aggressiveness, varies significantly among companies, indicating diverse commitments to innovation and public visibility. Conversely, TENURE\_TENURE\_MEAN, depicting Leadership Retainment, shows less variation, suggesting stable leadership structures or effective retention strategies. CULT\_EMPH, representing Culture Emphasis, exhibits wide variance, reflecting diverse corporate values or strategic priorities outlined in annual reports. These metrics provide nuanced perspectives, aiding stakeholders in understanding companies' strategic orientations amidst the corporate landscape's complexities.

**Signal Generation and Portfolio Construction:**

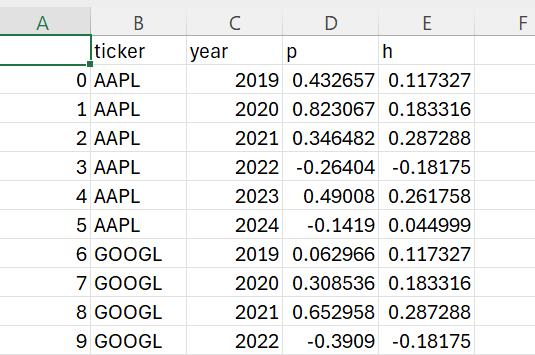
There are 2 datasets required for signal generation. The first data set contains features (as discussed in the previous section) for each year in our observation period, i.e. 2014 to 2023. The second data set contains annual returns for all the training years. This is the target variable on which we will train the model. Let us review the relevant columns of both data sets.

Dataset 1: Features



Column Description:  
1. TICKER: Ticker of the company   
2. YEAR: Year of annual return  
3. TDC1\_TDC1\_MEAN:   
4. ISR\_ISR\_MEAN  
5. TENURE\_TENURE\_MEAN  
6. ART\_COUNT  
7. CULT\_EMPH

Dataset 2: Annual Returns



Column Description:

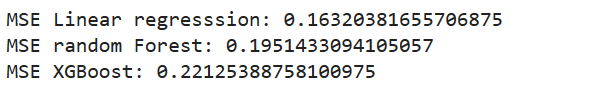
1. ticker: Ticker of the company
2. year: Year of analysis
3. p: Return of the stock
4. h: Return of the hedge (SPY US for the same period)

Modelling: We have run tests using 3 machine learning models of varying complexity. The 3 models that we have tested are: Lasso Regression, Random Forest and XG Boost. Let us briefly review the pros and cons of the three models. Note that this list is not exhaustive and is just indicative of the way the model may handle the data.

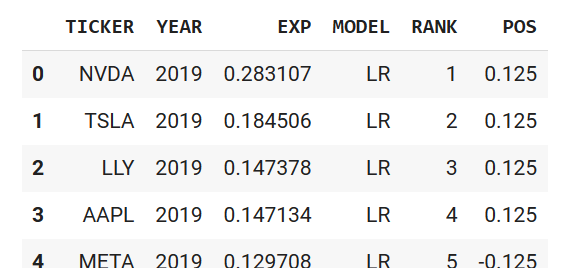
**Table 3: Pros & Cons of Statistical Models**

|  |  |  |
| --- | --- | --- |
| **Model** | **Pros** | **Cons** |
| Lasso Regression | Easy feature selection | Instability in correlated features |
|  | Regularization | Simple Model |
| Random Forest | High Accuracy | Lack in interpretability |
|  | Handles non-linearity | Computationally expensive |
| XG Boost | High Performance | Lack in interpretability |
|  |  | Computationally expensive |

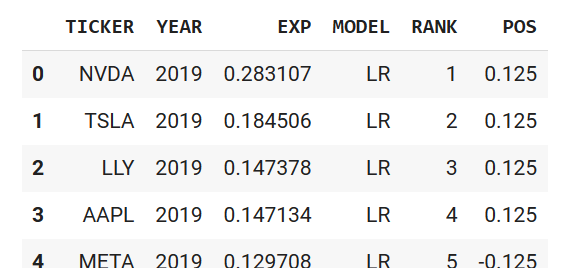
We fit the model using the features (dataset 1) and the annual returns (dataset 2). The result of model fitting i.e. Mean Squared Error (MSE) is mentioned below. The MSE for the Lasso (Linear) Regression came out to be the least and the MSE for the XG Boost came out to be the highest. This implies that Lasso (Linear) has the best fit on the training data.



After the models are fitted, we use the features from 2019 to 2023 to generate predicted annual return values for all the tickers individually. Then for each year, all the securities are ranked from highest to lowest in the order of predicted return. This is represented in the example output below.



This rank is our indicative buying and selling signals for each security. We use this ranking to generate a portfolio as described in the portfolio construction process next. We have used a heuristic portfolio construction process which is a long-short market and size neutral construction strategy. The idea behind the construction process was to generate a portfolio which is risk averse to market factors and is relatively simple for explainability.   
  
 All securities in our portfolio belong to the US market so if we have a 4 long and 4 short position every year, at the market level, the portfolio will be neutral by construction. Similarly, all the securities which we have used for this paper are Large-Cap securities and hence by construction the portfolio is Size factor neutral as well. We add the position column in the signal dataset indicative of the weight of the securities at the start of the year. Note the extra column has been added in the dataset below indicative of positions.



We ran the above-mentioned signal generation and portfolio construction process for every year and generated suggested positions for every year. Then we generated cumulative Profit & Loss for the entire period (2019-2023) based on the position and we have compared the numbers to the Profit & Loss of the S&P 500 which is the benchmark for the US market. We repeated the experiment with training the model once with just p (stock return) and once with p-h (stock–market) returns, instead of just the stock return. We will summarize our findings in the next section of the paper.

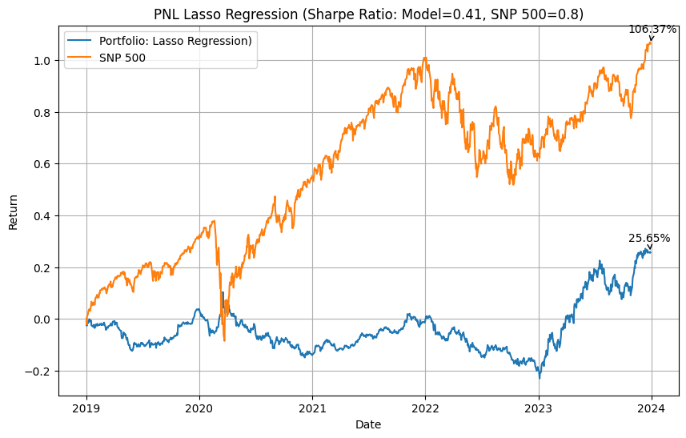
**Results:**

Version 1: Model trained on return of securities

1. Lasso regression

The portfolio constructed exhibits a cumulative return of around 25.65%, notably lower than the S&P 500's return of approximately 106.37% over the same period. Additionally, the Sharpe Ratio for the Lasso Regression model is 0.41, indicating a less favorable risk-adjusted return compared to the S&P 500's ratio of 0.8. The portfolio did avoid the short dip seen by the S&P 500 in 2020, possibly because of the long-short nature of the portfolio.

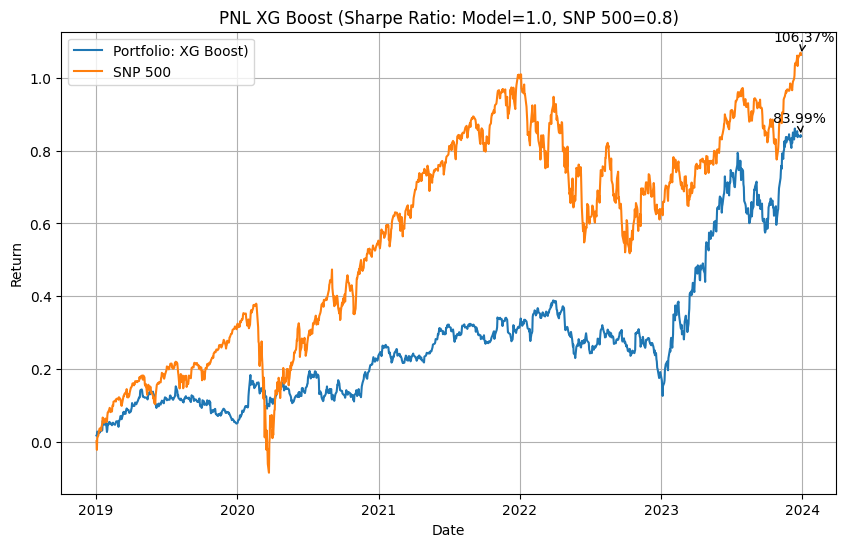
**Figure 7.1**



1. XG Boost

The XG Boost portfolio concludes with a return of approximately 83.99%, comparable to the S&P 500's overall return of 106.37%. However, the Sharpe Ratio for the XG Boost portfolio stands at 1.0, surpassing the S&P 500's ratio of 0.8, indicating a superior risk-adjusted return. Despite the XG Boost portfolio yielding lower absolute returns than the S&P 500, its risk-adjusted performance is better. Furthermore, the performance improves significantly when compared to the lasso regression-based model.

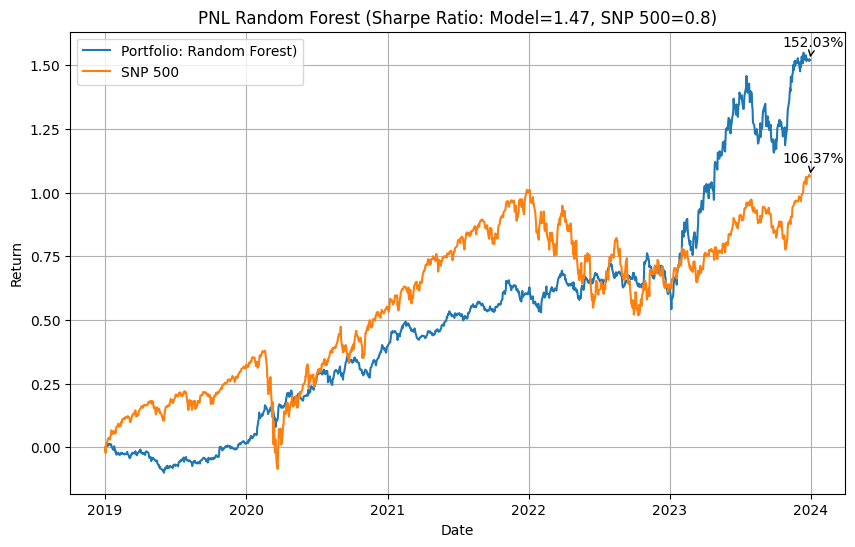
**Figure 7.2**



1. Random Forest

The portfolio, utilizing the Random Forest model, showed remarkable performance, boasting a return of approximately 152.03%, significantly surpassing the S&P 500's return of about 106.37%. Furthermore, the Sharpe Ratio for the Random Forest model stands at an impressive 1.47, indicating superior risk-adjusted performance compared to the S&P 500. This suggests that not only did the portfolio generate higher absolute returns, but it also managed risk more effectively, making it an attractive investment option.

**Figure 7.3**

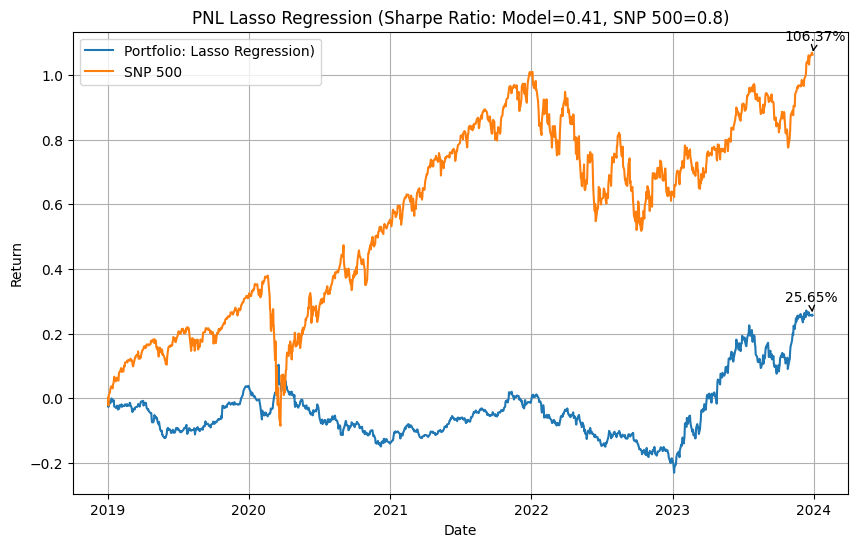
.

Version 2: Model trained on return of securities minus the return of market

1. Lasso regression

The portfolio exhibits a cumulative return of around 25.65%, noticeably lower than the S&P 500, which boasts a return of about 106.37% during the same period. Additionally, the Sharpe Ratio for the Lasso Regression model stands at 0.41, falling short of the S&P 500's ratio of 0.8, suggesting a less favorable risk-adjusted return. The cumulative return is same as the one noticed for the model trained on just the security return.

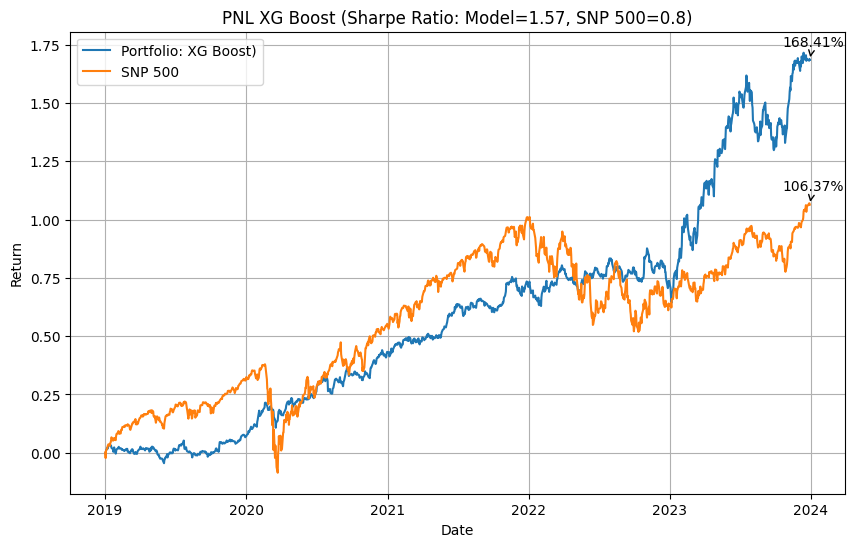
**Figure 8.1**



1. XG Boost

The XG Boost portfolio concludes with an impressive return of approximately 168.41%, outperforming the S&P 500, which achieves an overall return of 106.37%. The Sharpe ratio of the model is also significantly better than the S&P 500. Another point to notice is the jump in the performance when using the securities minus the market return when training the model.

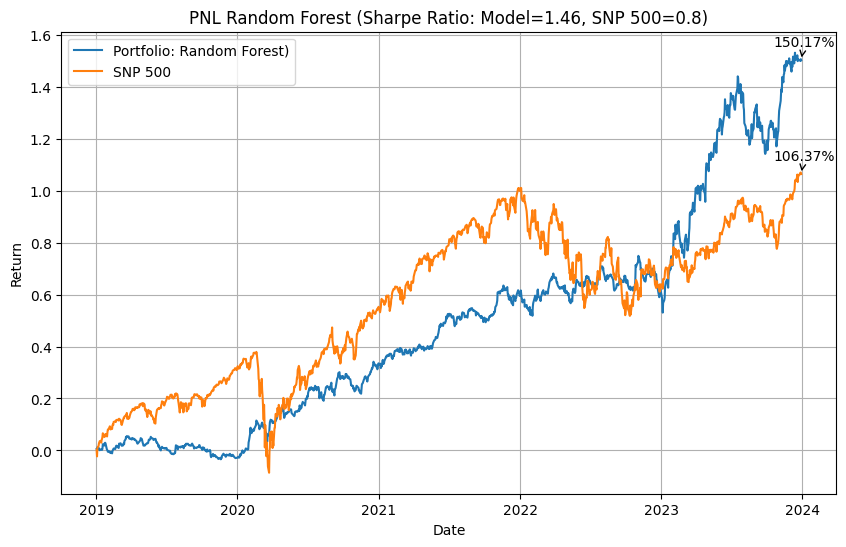
**Figure 8.2**



1. Random Forest

The portfolio exhibited exceptional performance, surpassing the S&P 500 with a return of approximately 152.03%, compared to the S&P 500's return of about 106.37%. Moreover, the Random Forest model utilized in the portfolio demonstrated a higher Sharpe Ratio of 1.47, indicating superior risk-adjusted performance relative to the S&P 500. This suggests that not only did the portfolio generate significantly higher absolute returns than the market index, but it also managed risk more effectively.

**Figure 8.3**



**Conclusion:**

The study highlights the crucial role of selecting appropriate features like features that encapsulate financial stability, operational efficiency, leadership, and public perception in modeling of trading strategies and are vital for crafting predictive models that accurately forecast stock returns. Random Forest and XG Boost models proved to be the most effective in this setting, suggesting that complexity in model design can indeed translate to better predictive performance and financial outcomes.

The strategy has notable benefits and limitations, along with potential areas for future enhancement. Benefits include a relatively lower risk profile and minimal trading costs due to annual rebalancing, making it an attractive option for conservative investors seeking steady returns. However, the limitations are significant, with the strategy not substantially outperforming the benchmark, reliance on limited data sources, and unclear price drivers influencing stock performance. To address these issues, future enhancements could involve expanding the universe of stocks and incorporating more diverse data sources to enrich the model inputs. Additionally, adopting dynamic portfolio construction methods, such as neural networks, could refine decision-making processes. Implementing profit and loss attribution would further help in understanding and isolating the price drivers, enhancing the strategy’s overall effectiveness and responsiveness to market dynamics.

This study underscores that more research needs to be done to uncover relations between factors such as financial stability, operational efficiency, leadership, company culture and the stock returns over larger horizons. The mentioned factors can be a good source of long-term investments.

***References***

*Algorithmic Trading and the Market for Liquidity*, faculty.haas.berkeley.edu/hender/ATMonitor.pdf. Accessed 6 May 2024.

*News versus Sentiment: Predicting Stock Returns From ...*, www.federalreserve.gov/econresdata/feds/2016/files/2016048pap.pdf. Accessed 6 May 2024.

Deveikyte, Justina, et al. “A Sentiment Analysis Approach to the Prediction of Market Volatility.” *Frontiers*, Frontiers, 29 July 2022, www.frontiersin.org/articles/10.3389/frai.2022.836809/full.