Sentiment Analysis to Predict Dow 30 Stock Direction

Kelvin Encarnacao, Matt Hodges, Michael Lee, Andrew McQueen

Quantitative Finance and Deep Learning

FIN:9160:0800

Data Description

Our project utilizes two datasets: 1) daily price, volume, and return information for the 30 stocks listed on the Dow Jones Industrial Average index (DJIA) for all market trading days in 2020, sourced from the Compustat North America database located on the Wharton Research Data Services (WRDS) subscription platform; and, 2) sentiment scores derived from Thomson Reuters news articles published on the given trading days for each of the 30 stocks in the Dow 30, scraped from the Thomson Reuters website (www.thomsonreuters.com). The first dataset contains the following six features:

PERMNO – unique permanent identification numbers assigned by Center for Research in Security Prices (CRSP) to all companies listed in CRSP dataset; datatype: int64.
date – series of integers identifying the date in the YYYYMMDD format; datatype: int64.
TICKER – string of capitalized letters of the ticker symbols assigned to each company by their given stock markets (i.e., NYSE or NASDAQ) and listed on the DJIA; datatype: object.
PRC – floating point number series giving the raw prices to the hundred thousandth decimal as they were reported at the end of the time of trading (daily); datatype: float64.
VOL – series of integers totaling raw number of shares of a stock traded on given day, not adjusted for splits nor containing over-allotment; datatype: int64.
RET – floating point number series giving the change in the total value of an investment in the security per dollar of initial investment; datatype: float64.

The second dataset also has six features:

timestamp – numerical date with year, month, and day each separated by a hyphen, then followed by hours, minutes, and seconds; datatype: datetime64.
headline – string containing each article's title; datatype: object.
body – string containing each article's body; datatype: object.
subject – string containing the ticker symbol of each stock mentioned in the article; datatype: object.

TextBlobTitle – sentiment score assigned using the TextBlob library on the *headline* feature values for each article; datatype: float64.
TextBlobBody – sentiment score assigned using the TextBlob library on the *body* feature values for each article; datatype: float64.

These two datasets require preprocessing for our analysis, which is discussed below, but before we continue with the data preparation we outline our project's goal.

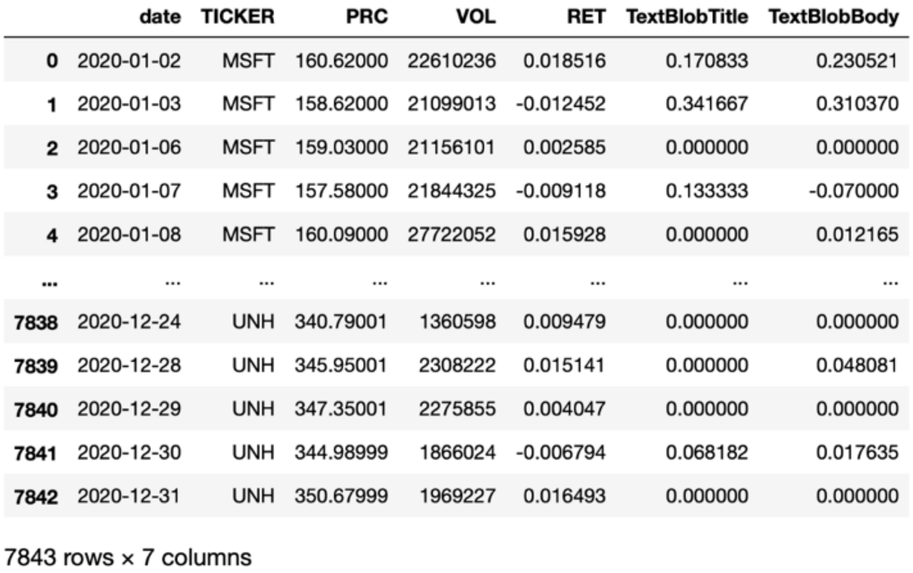
Project Goal

The success of publications like The Wall Street Journal, Financial Times, and Bloomberg—to name just a few—as trusted sources for investor information is undeniable. Each day millions of investors seek the latest news that can help them make the right stock picks that will provide the best returns. The question is, however, what is the relationship the news itself has to the performance of the public companies on which the publications report. Our project investigates that relationship by sourcing sentiment scores from Thomson Reuters news articles on the companies in the DJIA—which indexes the top thirty publicly traded US companies listed on either the New York Stock Exchange (NYSE) or the NASDAQ—and by analyzing the performance of those companies in relation to those sentiment scores.

Data Preparation

Each of our datasets required a little preprocessing before we could begin modeling and analyzing stock performance in relation to news sentiment. First, the *date* and *timestamp* features from each dataset needed to be made compatible by reformatting them and setting the time zone for the news dataset to the eastern US time zone to match the stock information from the NYSE and the NASDAQ. Next, again in the news dataset, the *subject* feature contained ticker symbol information for any of the Dow 30 companies that are mentioned in the given articles, but it was in a non-malleable format. We had to apply literal_eval from the *Abstract Syntax Tree* Python library to breakdown the syntax of the feature, then we exploded the feature in order to separate the values of a given news article that mentioned multiple companies, allowing us to provide unique values for each company. Finally with this feature, we created the *TICKER* feature for the news dataset by transforming the exploded *subject* values by removing any superfluous string characters around the letters of the ticker symbols. For instance, the original *subject* values were transformed from this ['R:MSFT.O'] to simply this MSFT, which matches the format of the stock performance dataset.

The final two steps of preprocessing were to group and merge. First, we grouped all the values in the news dataset by their *TICKER* values and their date, based on single day values (i.e., any news that came out on a given day on a particular company), then we needed to average the sentiment scores to get a single sentiment score for the news of the day for each company. Finally, we merged the two datasets—news sentiment and stock performance—by both date and ticker symbol, removed any null values, and then dropped the *timestamp* and *PERMNO* features because they were made redundant by the merged *date* and *TICKER* features. This gave us the dataset with which we were able to begin our modeling and analysis:



	date	TICKER	PRC	VOL	RET	TextBlobTitle	TextBlobBody
0	2020-01-02	MSFT	160.62000	22610236	0.018516	0.170833	0.230521
1	2020-01-03	MSFT	158.62000	21099013	-0.012452	0.341667	0.310370
2	2020-01-06	MSFT	159.03000	21156101	0.002585	0.000000	0.000000
3	2020-01-07	MSFT	157.58000	21844325	-0.009118	0.133333	-0.070000
4	2020-01-08	MSFT	160.09000	27722052	0.015928	0.000000	0.012165
...	...	...	...	...	...	...	...
7838	2020-12-24	UNH	340.79001	1360598	0.009479	0.000000	0.000000
7839	2020-12-28	UNH	345.95001	2308222	0.015141	0.000000	0.048081
7840	2020-12-29	UNH	347.35001	2275855	0.004047	0.000000	0.000000
7841	2020-12-30	UNH	344.98999	1866024	-0.006794	0.068182	0.017635
7842	2020-12-31	UNH	350.67999	1969227	0.016493	0.000000	0.000000

7843 rows × 7 columns

During the modeling and analysis process, we found one final necessary preprocessing step. We noticed that Citigroup (C) and Honeywell International (HON) had no sentiment values for the entire dataset, which indicated that Thomson Reuters did not publish any news articles during the year on these two companies in the Dow 30.

Modeling

Once the Dow30 stocks were merged with their news for the year 2020, we decided our target variable for the prediction, which is stock return. The variables we wanted to use were stock price, stock volume, and news sentiment scores. Given the amount of lagging for our dataset, we can simulate the effect of variables over time before moving on to perform the classification Tree-based static modeling, such as Random Forest, AdaBoost, and XGboost were used to predict the stock returns with and without sentiment scores. Modeling with and without the sentiment scores would allow us to determine how our models' accuracies changed with news sentiment. Static modeling simply means picking a date to separate the training and testing periods, we pick 2020-10-20 as our separator for the modeling. It's roughly 80% of time prior to 2020-10-20 and 20% after 2020-10-20. After creating a list of parameters for Random Forest, AdaBoost, and XGboost, we cross validated and found the optimal performances of stock return prediction for each algorithm we chose with or without news sentiment scores in their best parameter settings. For hyper tuning we tested several different max_depth variables and n_estimators for XGBoost and Random Forest and tested several n_estimators values for Ada Boost. It appears max model performance 0.53 on the accuracy very difficult to go pass that mark somehow no matter how we changed the parameter settings due to the limitations of our dataset. Moreover, we created several other models to conduct different comparisons. These included testing our prediction ability before and after the pandemic began, testing differences between industries included in the Dow, and differences between stocks with highly correlated returns and sentiment scores. However, the results of these models were often inconsistent with our hypotheses, or it was dependent on the individual models for each of the stocks. Some of these models are further discussed in the analysis section.

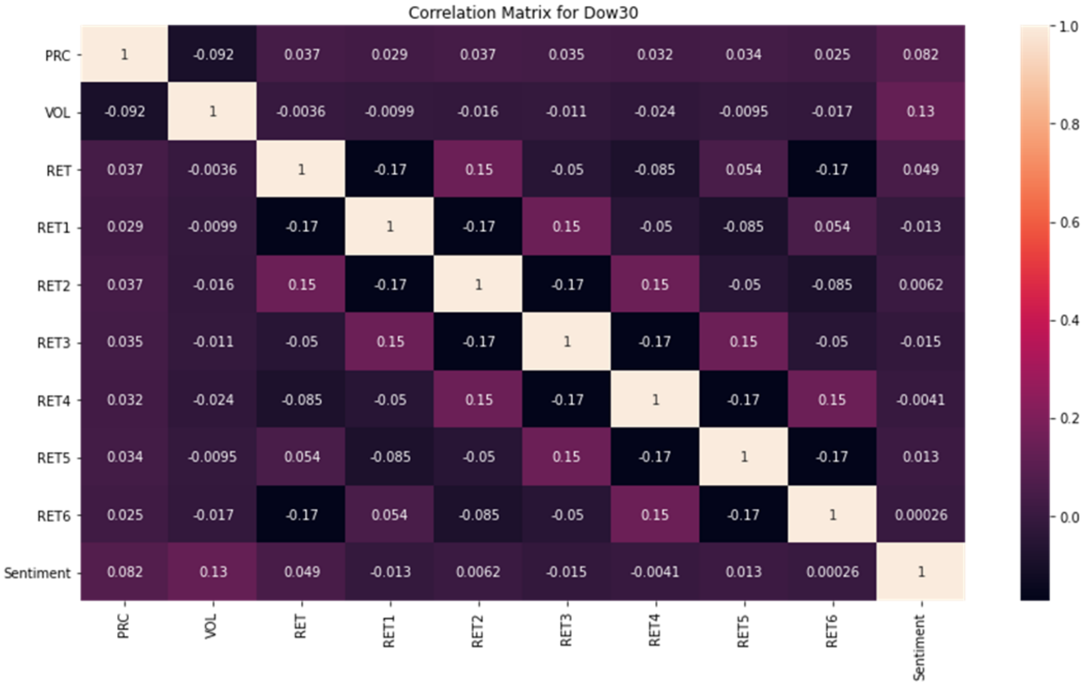
Data Limitations

Our main data limitations came from the amount of data and time from which the data was collected. 2020 is notoriously known as the year the COVID-19 pandemic began. This created a highly volatile market, which affected our ability to model because of the market's inconsistent nature. Market uncertainty. Another limitation was having only a year's worth of news data. This limited our dataset to 253 rows per stock. We expect that having more data would allow us to create better models and improve their accuracies.

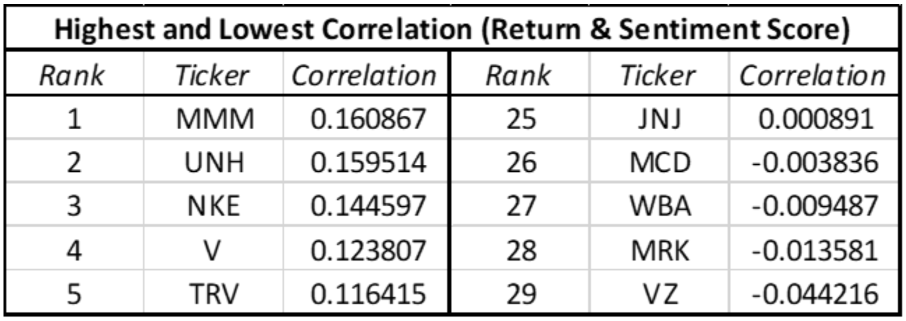
Analysis

Correlation of Returns and Sentiment Scores

The correlation between stocks' returns and sentiment scores can help determine whether there is a relationship between the two. We find that there is only a very small correlation (0.049) between the returns and sentiment scores of *all* stocks. The relationship between stocks' sentiment scores and their lagged returns are unpredictable and are highly *uncorrelated*. The weakness of this relationship may suggest market efficiency, as information contained in news articles is very quickly reflected in the price of a given stock. If market efficiency did not exist, we would be able to use sentiment scores and company announcements to easily predict the future price of stock. However, this is not the case.

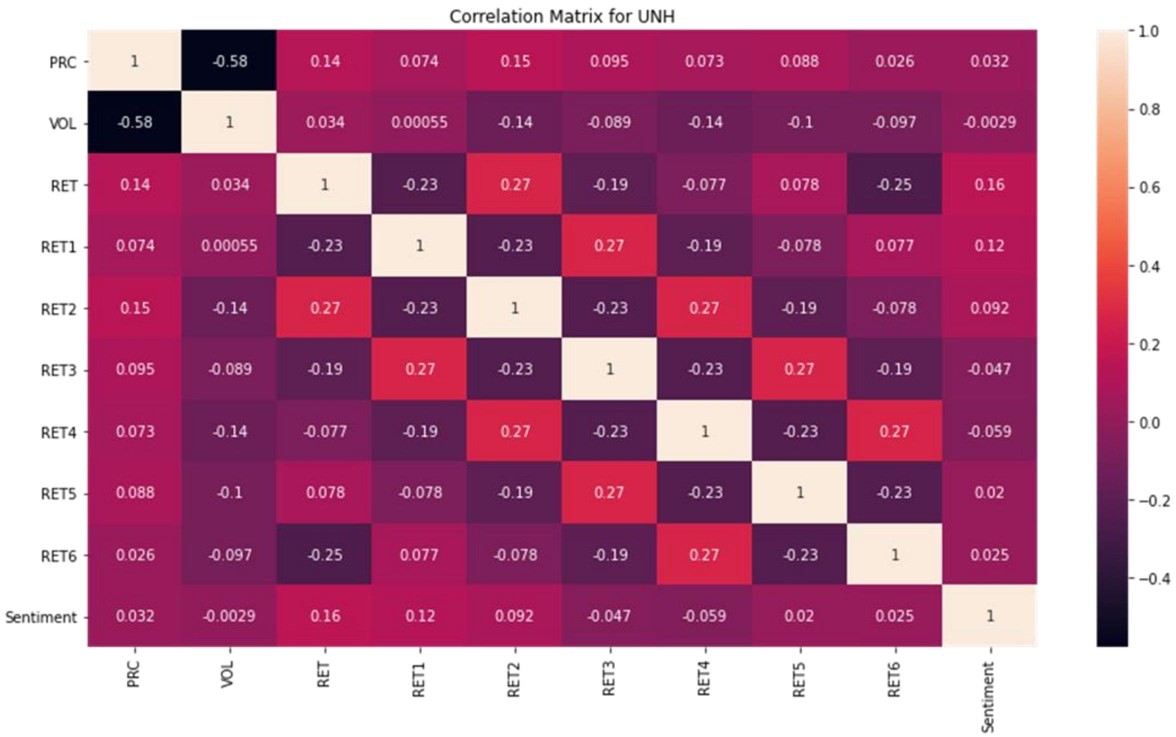


To determine the effect of this factor, we chose stocks with the strongest relationships between return and sentiment score. Initially, we tried grouping stocks by their respective industries, though we found that stocks within industries were unrelated in terms of this relationship. Some stocks' returns and sentiment scores even have a negative relationship, suggesting the inconsistency of using sentiment scores to predict the direction of multiple stocks.

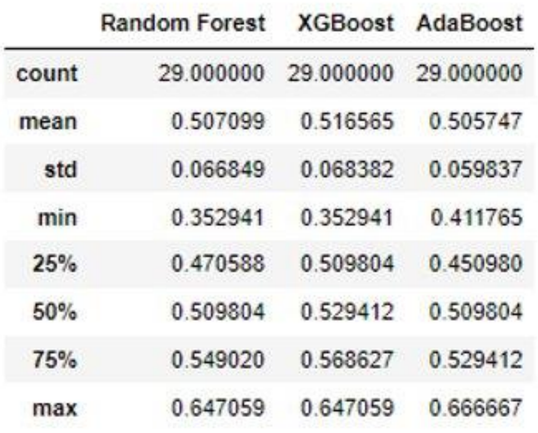


Highest and Lowest Correlation (Return & Sentiment Score)					
Rank	*Ticker*	*Correlation*	*Rank*	*Ticker*	*Correlation*
1	MMM	0.160867	25	JNJ	0.000891
2	UNH	0.159514	26	MCD	-0.003836
3	NKE	0.144597	27	WBA	-0.009487
4	V	0.123807	28	MRK	-0.013581
5	TRV	0.116415	29	VZ	-0.044216

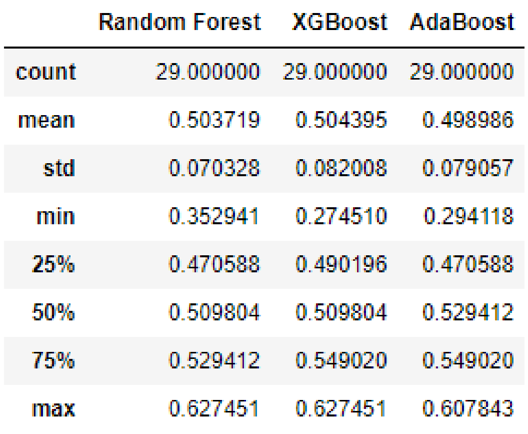
For one of the top stocks, UnitedHealth Group Inc, we see that the correlation between its stock returns and sentiment scores is around 0.16, and that the correlation of the first few return lags with sentiment scores is more consistent than the entire Dow. We wanted to test our models' ability to predict the direction of stocks' returns with high correlations, then compare them to the entire dataset.



After doing this, we decided to create and run our models. We had 2 different iterations of results, one with fixed parameters for everything and one using cross validation with multiple different parameters in order to use the best parameters for each model for each stock. The first set of results are as follows with the sentiment variables being used on the right and not being used on the left.



	Random Forest	XGBoost	AdaBoost
count	29.000000	29.000000	29.000000
mean	0.507099	0.516565	0.505747
std	0.066849	0.068382	0.059837
min	0.352941	0.352941	0.411765
25%	0.470588	0.509804	0.450980
50%	0.509804	0.529412	0.509804
75%	0.549020	0.568627	0.529412
max	0.647059	0.647059	0.666667

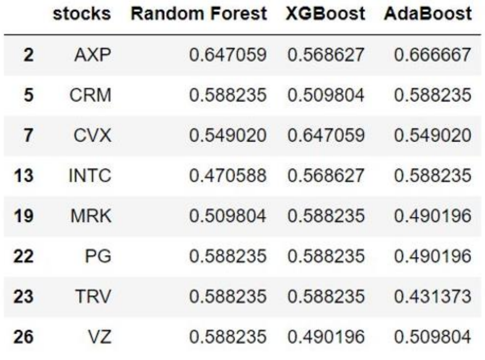


	Random Forest	XGBoost	AdaBoost
count	29.000000	29.000000	29.000000
mean	0.503719	0.504395	0.498986
std	0.070328	0.082008	0.079057
min	0.352941	0.274510	0.294118
25%	0.470588	0.490196	0.470588
50%	0.509804	0.509804	0.529412
75%	0.529412	0.549020	0.549020
max	0.627451	0.627451	0.607843

Looking at this the results are a bit inconclusive, on average without any cross validated hyper tuning the sentiment scores perform slightly better but it is so close so it’s unreliable to have any takeaways from this other than that we cannot observe a noticeable difference in accuracy using the sentiment values in the model.

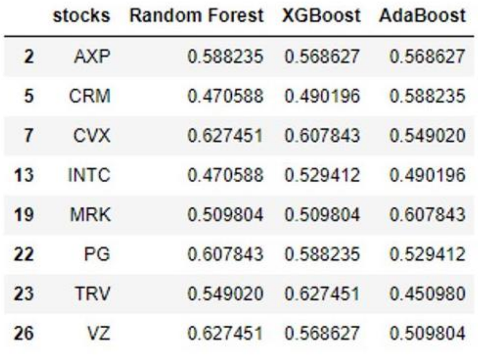
Next before getting into some hyper tuning for better results, we wanted to look at which stocks performed the best in the sentiment dataset for accuracy. The results are as follows:

Sentiment



	stocks	Random Forest	XGBoost	AdaBoost
2	AXP	0.647059	0.568627	0.666667
5	CRM	0.588235	0.509804	0.588235
7	CVX	0.549020	0.647059	0.549020
13	INTC	0.470588	0.568627	0.588235
19	MRK	0.509804	0.588235	0.490196
22	PG	0.588235	0.588235	0.490196
23	TRV	0.588235	0.588235	0.431373
26	VZ	0.588235	0.490196	0.509804

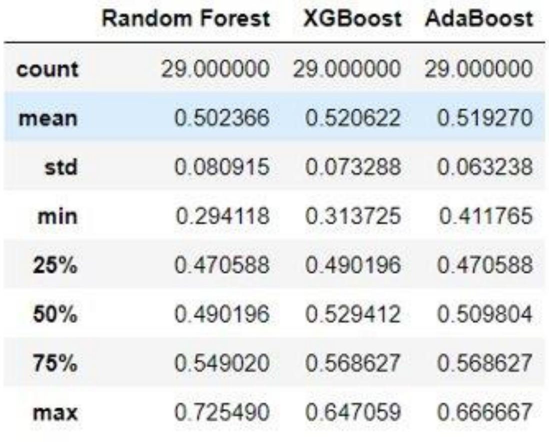
No Sentiment



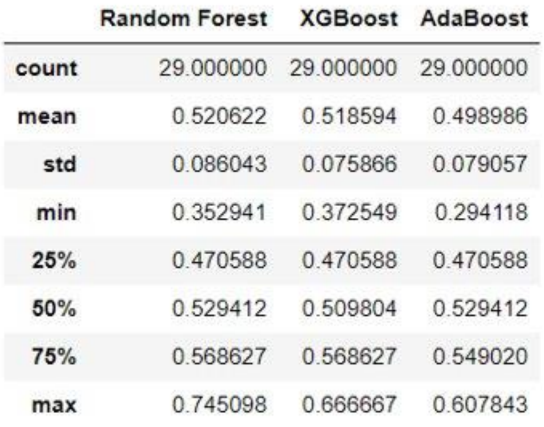
	stocks	Random Forest	XGBoost	AdaBoost
2	AXP	0.588235	0.568627	0.568627
5	CRM	0.470588	0.490196	0.588235
7	CVX	0.627451	0.607843	0.549020
13	INTC	0.470588	0.529412	0.490196
19	MRK	0.509804	0.509804	0.607843
22	PG	0.607843	0.588235	0.529412
23	TRV	0.549020	0.627451	0.450980
26	VZ	0.627451	0.568627	0.509804

Looking at this there are some stocks that perform significantly better than their counterparts in the non-sentiment dataset, however there are some that also perform worse, so we were unable to come to any conclusions at this point about why this might work for select stocks and not for others.

Next, we looked at hyptertuning our data with cross validation to see how these parameters might influence the sentiment feature's effectiveness to predict the direction of the stock. The results are as follows again with sentiment scores being on the left and non-sentiment on the right:



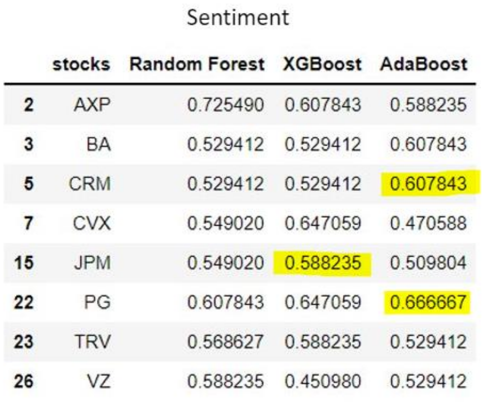
	Random Forest	XGBoost	AdaBoost
count	29.000000	29.000000	29.000000
mean	0.502366	0.520622	0.519270
std	0.080915	0.073288	0.063238
min	0.294118	0.313725	0.411765
25%	0.470588	0.490196	0.470588
50%	0.490196	0.529412	0.509804
75%	0.549020	0.568627	0.568627
max	0.725490	0.647059	0.666667



	Random Forest	XGBoost	AdaBoost
count	29.000000	29.000000	29.000000
mean	0.520622	0.518594	0.498986
std	0.086043	0.075866	0.079057
min	0.352941	0.372549	0.294118
25%	0.470588	0.470588	0.470588
50%	0.529412	0.509804	0.529412
75%	0.568627	0.568627	0.549020
max	0.745098	0.666667	0.607843

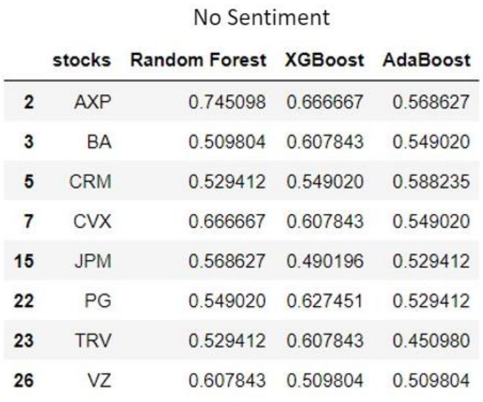
These results similar to our fixed parameter results earlier. Our top performing model on average was from the sentiment dataset and was XGBoost which slightly outperformed the others by a couple percentage points. The max accuracy scores, however, still belonged to the non-sentiment dataset for which Random Forest outperformed the sentiment data Random Forest models on average. Adaptive boosting models also performed almost an entire 2 percent better on average for the sentiment data and we saw improvement here from our previous fixed results.

We similarly printed the top stocks based on accuracy from this dataset:



Sentiment

	stocks	Random Forest	XGBoost	AdaBoost
2	AXP	0.725490	0.607843	0.588235
3	BA	0.529412	0.529412	0.607843
5	CRM	0.529412	0.529412	0.607843
7	CVX	0.549020	0.647059	0.470588
15	JPM	0.549020	0.588235	0.509804
22	PG	0.607843	0.647059	0.666667
23	TRV	0.568627	0.588235	0.529412
26	VZ	0.588235	0.450980	0.529412

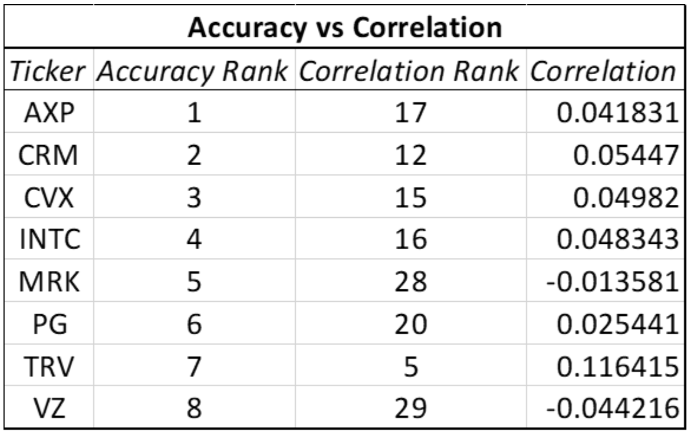


No Sentiment

	stocks	Random Forest	XGBoost	AdaBoost
2	AXP	0.745098	0.666667	0.568627
3	BA	0.509804	0.607843	0.549020
5	CRM	0.529412	0.549020	0.588235
7	CVX	0.666667	0.607843	0.549020
15	JPM	0.568627	0.490196	0.529412
22	PG	0.549020	0.627451	0.529412
23	TRV	0.529412	0.607843	0.450980
26	VZ	0.607843	0.509804	0.509804

Highlighted here are the stocks that performed much better using the sentiment features, the most notable of which is PG which was almost an entire 5 percent better on max accuracy in the sentiment data compared to the non-sentiment data. Looking at the rest of the top performers, however, it is hard to come to a definitive conclusion because across the board there are stocks that perform much better and others that seem to perform worse with the inclusion of the sentiment data.

The next thing we looked at was what kind of relationship our accuracy results could have when looking at correlations between return and sentiment for each individual stock. We calculated the correlations for the previously mentioned stocks with the top accuracies in our hyper tuned model results.

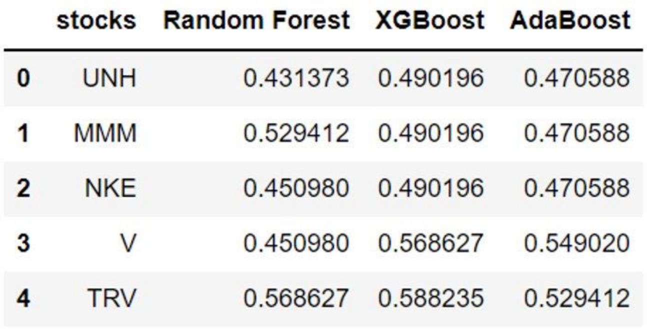


Accuracy vs Correlation

Ticker	*Accuracy Rank*	*Correlation Rank*	*Correlation*
AXP	1	17	0.041831
CRM	2	12	0.05447
CVX	3	15	0.04982
INTC	4	16	0.048343
MRK	5	28	-0.013581
PG	6	20	0.025441
TRV	7	5	0.116415
VZ	8	29	-0.044216

Looking at this there seems to be no direct correlation between our top accuracy scores and a high correlation value between sentiment and return. TRV is an exception which had a high accuracy and a high correlation, but the rest had average and below average correlation scores so our hypothesis could not be supported. Many of the stocks with the highest stocks had lower than average (0.0545) correlations, and the most inversely correlated stocks were also some of the most accurately predicted.

Furthermore, we looked specifically at the stocks who had the highest correlation scores to see what their accuracies looked like.



	stocks	Random Forest	XGBoost	AdaBoost
0	UNH	0.431373	0.490196	0.470588
1	MMM	0.529412	0.490196	0.470588
2	NKE	0.450980	0.490196	0.470588
3	V	0.450980	0.568627	0.549020
4	TRV	0.568627	0.588235	0.529412

These stocks performed differently across the board as stocks like TRV and V had an above average accuracy score but others like UNH which we looked specifically at earlier, NKE, and MMM all performed poorly or below average. Again, this showed us that we needed more information and could not support any claim that sentiment on average for every stock will predict the direction of return with high accuracy.

Conclusion

Sentiment scores act as one piece of information and it’s not reliable to predict for the entire market price return movements. For a select number stocks the sentiment data was useful and increased our predictive accuracy by a significant amount. Looking at the bigger picture there was not enough information to come to a consensus about these sentiment scores being a good predictor on average for any Dow30 stock or any stock in general based on our analysis. We think that this information could be very useful, however there are many things to consider such as the context behind these sentiment scores and news reports. There is much more work to be done to completely prove or disprove any hypothesis about sentiment scores and we think using more data from a longer timeline and perhaps from more than just Dow30 stocks could give us much different results. We faced some limitations as previously mentioned and market uncertainty certainly influenced our results, however, we cannot say for certain whether this method of prediction will achieve more accurate results on average for the Dow30 as a whole. It would be more useful to try more specific sentiment features that are a bit more intuitive and to try different models with many other parameters for further testing.