Project Title: Using Machine Learning Techniques to Integrate Technical and Fundamental Analysis to Identify Trading Opportunities

Student Names:
Longxiang Dai, 70961656, longxiad@uci.edu
Yongheng Zhang, 42664320, yonghenz@uci.edu

Github: https://github.com/zhangyongheng78/Stats170B

1. Introduction and Problem Statement

For actively managed stock portfolios, there are two primary approaches to investing and trading, fundamental analysis, and technical analysis. Fundamental analysis is about evaluating a company's financial statements, the market it belongs to, and the overall economy to determine the intrinsic value of a share. Technical analysis, instead, evaluates investments purely on the market activity surrounding them. It seeks to identify patterns and trends by studying historical price and volume.

We believe that machine learning techniques can be implemented in trading and investments to better predict markets and execute trades at optimal times. Hence, in this project we experiment with various machine learning algorithms and measure their performance on trading stocks. We train models to learn from financial reports, macroeconomic indicators, and historical market data to forecast future stock movements and give suggestions on buying or selling the stocks. The approach we take to address this problem is 4-fold:

1) Build predictive models to conduct fundamental analysis and make long-term predictions.
2) Build predictive models to perform technical analysis and make short-term predictions.
3) Apply clustering methods to identify stocks that have common financial conditions and trends, and then make predictions on the basis of each cluster.
4) Collect tweets and perform sentiment analysis to generate a sentiment indicator for each stock in the S&P 500 index. Explore how it foretells the ups and downs of stocks.

2. Related Work

Since the official Twitter API does not allow users to get tweets older than a week, we refer to a free online source that bypasses the limitation of the official Twitter API (https://github.com/Jefferson-Henrique/GetOldTweets-python). We also refer to the SimFin API documentation (https://simfin.readthedocs.io/en/stable/) in order to load relevant datasets and generate financial signals. Since our research project relates to time-series and we are still new to this area, we also refer to a paper that discusses time-series clustering to cluster different stocks (https://www.sciencedirect.com/science/article/abs/pii/S0306437915000733). Inspired by the paper, we conduct a comparative analysis of machine learning methods for trading stocks (https://academic.oup.com/rfs/article/33/5/2223/5758276).

3. Data Sets

We use three datasets from Simfin (https://simfin.com/), Federal Reserve Economic Data (FRED) (https://fred.stlouisfed.org/), and Twitter (https://twitter.com/).

Simfin is an online platform for fundamental financial data from the US market and Germany market. Our focus is on the data from the US market, and those datasets can be accessed via free API or paid API. There are 13 datasets for the US market and we use the paid API to access the datasets, which contain more information than the datasets generated by the free API. We ignore 6 datasets that are related to banks and insurance companies and use the rest 7 datasets, which can be categorized into three parts:

1. Basic information such as Companies, Markets, and Sector/Industry;
2. Daily stock price data such as Share Prices;
3. Quarterly and annual financial statements such as Balance Sheet, Cash Flow, and Income Statement.

As of May 21st, the zipped file size for the Companies **(Figure 1)** dataset is 38.2 kB and has 2121 data rows. This dataset contains general information like companies' names and tickers for all companies in the US market. The zipped file size for Markets **(Figure 2)** dataset is 220.0 B and has 5 data rows. This dataset contains the markets and the currency of the markets. Each companies' currency of the share prices is the same as the currency of the market, which the company is in. The zipped file size for Sector/Industry **(Figure 3)** is 1.2 kB and has 71 data rows. This dataset includes sector and industry, which corresponds to the IndustryId of a company.

As of May 21st, the zipped file size for the Share Prices **(Figure 4)** dataset is 95.8 MB and has 5,715,030 data rows. This dataset contains all of the daily share prices. Users can also access the most recent data, which has less zipped file size (53.5 kB) and data rows (2121 rows).

The balance sheet indicates the financial situation of a company on a particular date. The financial situation of a company could be reflected as the company's assets, liabilities, and shareholder's equity. The Balance Sheet dataset could be accessed annually, quarterly, or trailing twelve months. As of May 21st, the zipped file size for the annual Balance Sheet dataset **(Appendix Figure 5)** is 2.6 MB and has 16,707 data rows; the zipped file size for the quarterly Balance Sheet dataset is 4.1 MB and has 53,654 data rows; thee zipped file size for trailing twelve months' Balance Sheet dataset is 4.0 MB and has 51,053 data rows. There are 27 features in this dataset, including Total Assets, Total Liabilities, and Total Equity, etc.,

The cash flow means how much money moves in and moves out of a company. The Cash Flow dataset could be accessed annually, quarterly, or trailing twelve months. As of May 21st, the zipped file size for the annual Cash Flow dataset **(Figure 6)** is 1.5 MB and has 16,707 data rows; the zipped file size for the quarterly Cash Flow dataset is 4.3 MB and has 53,655 data rows; the zipped file size for trailing twelve months' Cash Flow dataset is 4.4 MB and has

51,054 data rows.There are 25 features in this dataset, including Net Change in Cash, Net Cash from Investing Activities, and Cash from Equity, etc.,

The income statement shows the profitability of a company. The Income Statement dataset could be accessed annually, quarterly, or trailing twelve months. As of May 21st, the zipped file size for the annual Income Statement dataset **(Figure 7)** is 1.6 MB and has 16,707 data rows; the zipped file size for the quarterly Income Statement dataset is 4.4 MB and has 53,655 data rows; the zipped file size for trailing twelve months' Income Statement dataset is 4.5 MB and has 51,054 data rows. There are 25 features in this dataset, including Revenue, Net Income, and Gross Profit, etc.,

The Federal Reserve Economic Data is a database provided and maintained by the research division of the Federal Reserve Bank of St. Louis. The database offers macroeconomic indicators as macroeconomic data. We use GDP (Gross Domestic Product), Unemployment Rate **(Figure 8)**, CPI (Consumer Price Index), IPI (Industrial Production Index), Permits, Sentiments in our clustering dataset. We use Treasure Rate in our training dataset.

The Twitter datasets **(Figure 9)** are collected from Twitter. We have 505 Twitter datasets total, and each of them represents the tweets for a stock. We use the keyword search method and collected tweets from 2007/01/01 to 2020/03/22. Those tweets are highly related to stocks in the S&P 500 since we included "$" before a ticker in our keyword search, which is an indicator of a stock.

4. Overall Technical Approach

Section 1: Data Management and Preprocessing

- Data Collection
    - Frequency conversion: we resample the data to the same frequency and fill the value by the last valid observation
    - Merging datasets: we join financial data and macroeconomic indicators from Simfin and FRED by the date the data is released. In merging financial datasets, the keys are "Report Date" and "Ticker" on both datasets. In merging the financial dataset and the macroeconomic indicators, the joint key of Simfin datasets is called "Report Y-M" which takes the year and month of the column "Report Date", and the joint key of FRED datasets is "Date". Also, in joining the sentiment indicator to the stock data, the joint keys of sentiment dataset are "twitter" and "time", and the joint keys of stock data are "Ticker" and "Report Y-M".
    - Date offset: we apply 3-month offset to quarterly data and 1-month offset to monthly data, including twitter sentiment indicator, to avoid information leak when the model is making trading decisions.
    - Tweets Collection: We improve the code of GetOldTweets to scrape the tweets of stocks from Twitter.
- Data Cleaning

    - We drop infinite and negative infinite values on generated features.
    - For each stock, we first fill any missing value by the last valid observation. Then we fill it by the sector mean in the same period. If none of these methods work, we remove the whole row.
    - We use the robust scaler to scale the data that removes the median and scales the data according to the interquartile range.
    - We apply one-hot-encoding to convert the categorical variable 'Sector' into dummy variables.
    - Remove emoji and redundant characters (such as stock ticker) from collected tweets before we perform the sentiment analysis.
- Feature Engineering
    - We generated 83 features **(Figure 10)** from financial reports and macroeconomic indicators for stock clustering and quarterly over quarter (QOQ) stock movements prediction, which can be categorized into five types of features:
        - Financial signals (net profit margin, debt ratio, return on assets, etc.)
        - Growth signals (sales growth, earnings growth, etc.)
        - Macroeconomic indicators (unemployment rate, inflation, etc.)
        - Stock information (sector)
        - Sentiment indicator (monthly average sentiment score of each stock)
    - We generated 41 features **(Figure 11)** from daily price and volume for the week over week (WOW), and 1 year stock movements prediction, which can be categories into six types of features:
        - Valuation signals (P/E, P/Sales Ratio, etc.)
        - Volume signals (volume market-cap, volume turnover, etc.)
        - Price signals (moving average, exponential moving average, etc.)
        - Trade signals (MACD buy signal, etc.)
        - Relative strength indicators signals (value signal, crossover signal)
        - US treasury rates (from 1-month to 30-year)
        - US treasury yield spreads

<u>Section 2: Data Analysis and Machine Learning</u>

- Part 1: Train models to conduct fundamental analysis and make long-term predictions
    - Target variable:
        - stock movement in a quarter (up or down)
    - Inputs: 83 quarterly features, 12k observations
- Part 2: Train models to conduct technical analysis and make short-term predictions
    - Target variable:
        - stock movement in a week (up or down)
        - stock movement in one years (up or down)
    - Inputs: 41 daily features, 861k observations
- Machine Learning Modeling:
    - Performance metric:
        - precision = true positive / (true positive + false positive)

        - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. In building portfolios, we rely on the model to pick stocks that we believe will perform well in the future. Because we only care about the stocks that hold, we calculate the precision, instead of accuracy, to measure the performance of our portfolio.
    - Validation methods:
        - Train-test split (The training set includes stocks before and includes 2017-12-31, and the test set includes stocks after and includes 2018-01-01).
        - Use cross-validation (cv=5) on the train set for hyperparameter tuning
    - Classification Models:
        - Logistic Regression
        - Support Vector Machine (SVM)
        - K-Nearest Neighbor (KNN) Classifier
        - Random Forest
        - Gradient Boosting
- Backtesting:
    - Measure the performance on the test set
    - Performance metrics:
        - Precision
        - Annualized Return
        - Annualized Volatility
        - Annualized Sharpe Ratio = (Return - Risk-free Rate) / Volatility

5. Software

(a) Code we wrote:

- Link datasets
- Data cleaning and feature engineering
- Generate price signals and sentiment indicators
- Implement clustering and predictive models
- Perform trading strategies suggested by the model
- Calculate return and sharpe ratios

(b) Publicly available code/libraries:

- Simfin API
- Federal Reserve Economic Data (FRED) API
- GetOldTweets Code
- Pandas for data wrangling
- TextBlob for sentiment analysis
- Scikit-learn for selecting training models
- Seaborn and matplotlib for visualization
- Tslearn for time-series clustering

6. Experiments and Evaluation

First, we train models to conduct fundamental analysis and predict quarter over quarter (QOQ) movements of stocks. We conduct a comparative analysis of different machine learning methods. For comparison, we define passive investing as the baseline approach. If one seeks to invest in stock markets but has no knowledge about the finance, it's likely he or she will be suggested to diversify the portfolio by holding all stocks in a stock market index such as S&P 500, Dow Jones Industrial Average, etc. As the baseline approach, we simply hold all trading companies in S&P 500 in equal weights. In this way, the overall precision of QOQ prediction is 0.6293, which means the prices of 62.03% holding stocks go up in the following quarter. As shown in **Table 1**, we experimented with different types of models and we found that random forest with 25 maximum depth performs best. A random forest is simply a collection of decision trees whose results are aggregated into one final result. Relative to other models, it is less likely to overfit. Running with a 5-fold cross validation method, the model outperforms the passive investing strategy and achieves an average of 0.7301 precision with 0.0083 standard deviation on validation sets.

Model	Train Precision (5-fold cross validation)	Validation Precision (5-fold cross validation)
Baseline (passive investing)	0.6293	
RobustScaler + Logistic Regression	0.6981 (0.0020)	0.6928 (0.0043)
RobustScaler + SVM	0.7889 (0.0021)	0.6622 (0.0053)
KNN (n=6)	0.8002 (0.0036)	0.7134 (0.0093)
Random Forest (max_depth = 25)	0.9873 (0.0031)	0.7301 (0.0083)
Gradient Boosting (max_depth = 10)	0.8202 (0.0110)	0.6940 (0.0063)

Table1. Model Performance (QOQ Prediction)

As random forest has shown its advantage in prediction, we take a deeper look at its performance on the test set. In quantitative finance, backtesting refers to the general method for seeking how well a strategy or model would have done on historical data. Here, we adopt three additional metrics for measuring the portfolio performance including annualized return, annualized volatility, and sharpe ratio. In addition to the binary category (up or down), the classification models output the probability as well. The default threshold is 0.5 which means to buy a stock and hold for the quarter when the model shows over 50% confidence in the stock in our case. Taking advantage of this, we can set different threshold values for accepting a decision to maximize the portfolio performance. Here, we train two separate models, one with the sentiment indicator generated from tweets and one without. In **Figure 12**, we find that as the threshold increases, precision and annualized return go up. Even though higher thresholds

cause higher volatility, we are benefited more from increased annualized returns. As a result, the adjusted-return (a.k.a. sharpe ratio) is boosted. Looking at the two lines on the graph, we find that adding the sentiment indicator to the model improves the performance.

In **Figure 13**, we can see how the portfolio value changes over time under different threshold values. The result shows that higher thresholds doesn't imply higher cumulative returns. It's because when the threshold is close to one, it's likely that the model would suggest pulling money out of the market if there is much uncertainty. To sum up, the higher the threshold value is, the more conservative the model is. Taking a high threshold is a desirable action to conservative investors but not to aggressive investors who prefer to take risk for excess returns.

The 1-year stock movement takes the stock's adjusted closing price at time t and time t+1year, and then calculates the relevant change. Therefore, this model could tell stockholders the stock's movement after 1 years of buying the stock. By using passive investing, we consider buying all the target stocks in the S&P 500. 77.09% of the stocks in the training set and 61.68% of the stocks in the test set are moving up after 1 year of buying the stocks. After selecting and tuning the models, we used the RandomForestClassifier at max_depth = 20 as our model. We perform backtest as illustrated in **Figure 14**. As what we expect, the precision, annualized return, and annualized sharpe ratio increases as the threshold increases. The annualized volatility and number of holding stocks decreases as the threshold increases.

We also perform the time-series clustering method. We first construct a table which contains the target stocks in the S&P 500 and their adjusted closing prices on recorded dates. Since the time-series clustering method requires the range of the time periods for different stocks should be the same, we drop the stocks that do not have enough records and end up with 400 stocks for clustering. We use elbow method to determine the optimal number of clusters for K-means, and perform Euclidean distance and Dynamic Time Warping (DTW) distance to cluster the stocks. After generating the clusters, we use the dataset that was used for predicting 1-year stock movement, and evaluate the subsets of it based on the cluster numbers. **Figure 15** and **Figure 16** show the backtest performance on Euclidean distance and Dynamic Time Warping distance. From those two figures, we can see that the cluster generated by dynamic time warping distance has better precision, annualized volatility, and annualized sharpe ratio. Besides, the model is more conservative in the cluster as the number of holding stocks decreases to 318 in the end.

Second, we train models to conduct technical analysis and predict week over week (WOW) movements of stocks. Similarly, we train a random forest model with 25 max depth and perform backtesting on the test set. **Figure 17** shows how the performance metrics get improved as the threshold increases. From **Figure 18**, we see that the portfolio cumulative return is maximized when the threshold equals to 0.75.

7. Notebook Description

Our Notebook contains the following parts. The first part includes helpful functions for generating quarterly and daily features from financial reports and historical stock data. The

second part is loading and merging the datasets, including the datasets from Simfin, FRED, collected tweets, and financial features calculated from financial reports. The third part addresses time-series clustering on S&P 500 stocks. The fourth part focuses on the combination of quarterly financial signals and macroeconomic data from FRED. We performed quarterly over quarter (QOQ) stock movements prediction and backtesting based on the dataset. The last part focuses on the combination of daily financial signals and macroeconomic data from FRED. We perform week over week (WOW), and 1 year stock movements prediction and backtesting.

8. Members Participation

The approach we take to address this problem is 4-fold: First part is to build predictive models to conduct fundamental analysis and make long-term predictions. Longxiang focuses on the quarter over quarter (QOQ) stock movements prediction and backtesting, and Yongheng focuses on 1 year stock movements prediction and backtesting. Second part is to build predictive models to perform technical analysis and make short-term predictions. Longxiang focuses on the week over week (WOW) stock movements prediction and backtesting. Yongheng participates in the third step, which repeats step 2 to perform technical analysis of each cluster of stocks after applying clustering methods to identify stocks that have common financial conditions. Longxiang participates in the fourth step, which is to collect tweets for each stock and perform sentiment analysis to generate a sentiment indicator, and explore how it foretells the ups and downs of stocks.

Tasks	**Sub-tasks**	**Participants**
1. Build predictive models to conduct fundamental analysis and make long-term predictions.	Quarter over Quarter (QOQ) stock movements prediction and backtesting	Longxiang Dai
	1 year stock movements prediction and backtesting	Yongheng Zhang
2. Build predictive models to perform technical analysis and make short-term predictions	week over week (WOW) stock movements prediction and backtesting	Longxiang Dai
3. After applying clustering methods to identify stocks that have common financial conditions, repeat step 2 to perform technical analysis on each cluster of stocks.	1 year stock movements prediction and backtesting	Yongheng Zhang
4. Collect tweets for each stock and perform sentiment analysis to generate	Algorithm and implementation	Longxiang Dai

a sentiment indicator. Explore how it foretells the ups and downs of stocks.		

9. Discussion and Conclusion

- **What we have learned; the strengths and limitations of such methods and algorithms:**
    - For the clustering perspective, we have learned how to do clustering in a time-series dataset. We got to know the representation and clustering algorithm aspects of the time-series clustering. Particularly, we have learned two different but major distance measuring methods for time-series clustering. Those two distance measuring methods are called Euclidean distance and Dynamic Time Warping distance. For different time series trends, the Euclidean distance is one to one measure. Specifically speaking, Euclidean distance focuses on measuring different time-series at the same time point. However, for two time-series as an example, a closer distance at each time point does not mean the overall trend is the same. The Dynamic Time Warping distance focuses on measuring different time-series at different time points. The time-series' trends at different time periods may have the same shape and Dynamic Time Warping distance can catch that. However, this method takes a long time since it takes more factors into consideration than the Euclidean distance.
    - For the classification part, we have learned that hyperparameter tuning is time-consuming but important in the modeling process. Hyperparameters directly control the behaviour of the training algorithm and have a significant impact on the performance of the model being trained.
    - We learned that the financial data are inherently noisy, so it's usually difficult to identify any pattern or trend in the data. However, inspired by the methods of fundamental analysis and technical analysis, we generate a bunch of features from financial reports as inputs to train machine learning models. The result shows that with these features, the model outperforms the passive investing strategy that most investors would take.
    - We also found that public sentiment can be helpful in the real world trading and investing. The information derived from tweets could give some useful information on buying and selling stocks.

- **What ended up being harder than expected:**
    - From the online tutorial and sources about time-series clustering, the authors usually focus on the change of one variable over time. For example, they only focus on the trends of the stocks' closing prices for different stocks and calculate the distance based on the trends. We initially want to combine more than one variables and then form combined trends for different stocks. However, this

thought requires us to calculate distance between trends of different stocks at one feature and combine with trends of different stocks at the other features before applying the clustering algorithm. This requires us to modify the clustering library, which is hard to implement.
    - We tried to use the sentiment indicator generated from tweets to make WOW predictions. However, as not every stock in the S&P 500 index has enough tweets each week, the daily sentiment score could not reflect the true public sentiment towards the company. So, we decided to calculate the monthly sentiment for QOQ prediction.

- **What other lessons we learned, expected or unexpected:**
    - We also use the other clustering method that is not related to time-series clustering. We use our dataset which is generated by the combination of quarterly financial signals and macroeconomic data from FRED. We tried to cluster the dataset based on different stocks at different dates. However, based on the visualization of the two methods that determine the optimal number of clusters, Elbow Method and Silhouette Method, we found out that there is no optimal number of clusters that can be determined. We also tried Principle Component Analysis and took components that explain around 80% of the variance. However, we still cannot find the optimal number of clusters based on the Elbow Method and Silhouette Method. This result is not what we expected but is reasonable, since there are tons of data rows and they might not have clear gatherings and separations.

- **Possible Future Directions:**
    - In the future, one of the ideas that might lead to a better clustering result is to put multiple features together in a time-series dataset. By combining features and conducting overall trends, the clustering result will be more accurate since there are more features put into consideration. For stock price prediction, we can also apply regressions to the dataset so not only we can see the signs of movement of stocks, but also can see the percentage change of the stocks over time. As financial data are inherently noisy, we can build deep neural networks that better understand the non-linear relationships between features, to improve the prediction accuracy.

Appendix

Ticker	SimFinId	Company Name	IndustryId
A	45846	AGILENT TECHNOLOGIES INC	106001.0
AA	367153	Alcoa Corp	110004.0
AAC	939324	AAC Holdings, Inc.	NaN
AAL	68568	American Airlines Group Inc.	100006.0
AAME	450021	ATLANTIC AMERICAN CORP	104004.0
...	...	...	...
ZUMZ	45730	Zumiez Inc	103002.0
ZVO	901866	Zovio Inc	102006.0
ZYNE	901704	Zynerba Pharmaceuticals, Inc.	106002.0
ZYXI	171401	ZYNEX INC	106004.0
low	186050	LOWES COMPANIES INC	103002.0

2113 rows × 3 columns

Figure 1. The Companies Dataset

MarketId	Market Name	Currency
ca	Canada	CAD
de	Germany	EUR
it	Italy	EUR
sg	Singapore	SGD
us	United States	USD

Figure 2. The Markets Dataset

IndustryId	Sector	Industry
100001	Industrials	Industrial Products
100002	Industrials	Business Services
100003	Industrials	Engineering & Construction
100004	Industrials	Waste Management
100005	Industrials	Industrial Distribution
...	...	...
110004	Basic Materials	Metals & Mining
110005	Basic Materials	Building Materials
110006	Basic Materials	Coal
110007	Basic Materials	Steel
111001	Other	Diversified Holdings

71 rows × 2 columns

Figure 3. The Sector/Industry Dataset

Ticker	Date	SimFinId	Open	Low	High	Close	Adj. Close	Dividend	Volume
A	2007-01-03	45846	34.99	34.0500	35.48	34.30	22.85	NaN	2574600
	2007-01-04	45846	34.30	33.4600	34.60	34.41	22.92	NaN	2073700
	2007-01-05	45846	34.30	34.0000	34.40	34.09	22.71	NaN	2676600
	2007-01-08	45846	33.98	33.6800	34.08	33.97	22.63	NaN	1557200
	2007-01-09	45846	34.08	33.6300	34.32	34.01	22.65	NaN	1386200
...	...	...	...	...	...	...	...	...	...
low	2020-05-06	186050	110.29	108.9461	111.53	109.62	109.62	NaN	4434512
	2020-05-07	186050	111.15	111.0700	113.13	111.92	111.92	NaN	3206726
	2020-05-08	186050	113.77	112.1000	114.63	114.23	114.23	NaN	3951316
	2020-05-11	186050	113.10	112.5733	114.75	113.39	113.39	NaN	3862234
	2020-05-12	186050	114.82	111.1800	115.05	111.24	111.24	NaN	3682485

5683975 rows × 8 columns

Figure 4. The Share Prices Dataset

Ticker	Report Date	SimFinId	Currency	Fiscal Year	Fiscal Period	Publish Date	Shares (Basic)	Shares (Diluted)	Cash, Cash Equivalents & Short Term Investments	Accounts & Notes Receivable	Inventories	...	Short Term Debt	Tot
A	2008-10-31	45846	USD	2008	Q4	2009-10-05	3.570000e+08	3.590000e+08	1.429000e+09	770000000.0	6.460000e+08	...	NaN	1.3
	2009-10-31	45846	USD	2009	Q4	2009-12-21	3.430000e+08	3.430000e+08	2.493000e+09	595000000.0	5.520000e+08	...	1.000000e+06	1.1
	2010-10-31	45846	USD	2010	Q4	2010-12-20	3.440000e+08	3.560000e+08	2.649000e+09	869000000.0	7.160000e+08	...	1.501000e+09	3.0
	2011-10-31	45846	USD	2011	Q4	2011-12-16	3.460000e+08	3.530000e+08	3.527000e+09	860000000.0	8.980000e+08	...	2.530000e+08	1.8
	2012-10-31	45846	USD	2012	Q4	2012-12-20	3.480000e+08	3.530000e+08	2.351000e+09	923000000.0	1.014000e+09	...	2.500000e+08	1.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
low	2014-02-28	186050	USD	2013	Q4	2014-03-31	1.034000e+09	1.037000e+09	5.760000e+08	NaN	9.127000e+09	...	4.350000e+08	8.8
	2015-02-28	186050	USD	2014	Q4	2015-03-31	9.640000e+08	9.670000e+08	5.910000e+08	NaN	8.911000e+09	...	5.520000e+08	9.3
	2016-02-29	186050	USD	2015	Q4	2016-03-29	9.090000e+08	9.100000e+08	7.120000e+08	NaN	9.458000e+09	...	1.104000e+09	1.0
	2017-02-28	186050	USD	2016	Q4	2017-04-04	8.670000e+08	8.660000e+08	6.580000e+08	NaN	1.045800e+10	...	1.305000e+09	1.1
	2018-02-28	186050	USD	2017	Q4	2018-04-02	8.270000e+08	8.280000e+08	6.900000e+08	NaN	1.139300e+10	...	1.431000e+09	1.2

16707 rows × 27 columns

Figure 5. The Annual Balance Sheet

Ticker	Report Date	SimFinId	Currency	Fiscal Year	Fiscal Period	Publish Date	Shares (Basic)	Shares (Diluted)	Net Income/Starting Line	Depreciation & Amortization	Non-Cash Items	...	Net Cash from Operating Activities	
A	2008-10-31	45846	USD	2008	FY	2009-10-05	3.630000e+08	3.710000e+08	6.930000e+08	2.010000e+08	41000000.0	...	756000000	-1
	2009-10-31	45846	USD	2009	FY	2009-12-21	3.460000e+08	3.460000e+08	-3.100000e+07	1.620000e+08	215000000.0	...	408000000	-1
	2010-10-31	45846	USD	2010	FY	2010-12-20	3.470000e+08	3.530000e+08	6.840000e+08	2.020000e+08	-116000000.0	...	718000000	-1
	2011-10-31	45846	USD	2011	FY	2011-12-16	3.470000e+08	3.550000e+08	1.012000e+09	2.530000e+08	219000000.0	...	1260000000	-1
	2012-10-31	45846	USD	2012	FY	2012-12-20	3.480000e+08	3.530000e+08	1.153000e+09	3.010000e+08	-50000000.0	...	1228000000	-1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
low	2014-02-28	186050	USD	2013	FY	2014-03-31	1.059000e+09	1.061000e+09	2.286000e+09	1.562000e+09	54000000.0	...	4111000000	-8
	2015-02-28	186050	USD	2014	FY	2015-03-31	9.880000e+08	9.900000e+08	2.698000e+09	1.586000e+09	77000000.0	...	4929000000	-8
	2016-02-29	186050	USD	2015	FY	2016-03-29	9.270000e+08	9.290000e+08	2.546000e+09	1.587000e+09	673000000.0	...	4784000000	-1
	2017-02-28	186050	USD	2016	FY	2017-04-04	8.800000e+08	8.810000e+08	3.093000e+09	1.590000e+09	563000000.0	...	5617000000	-1
	2018-02-28	186050	USD	2017	FY	2018-04-02	8.390000e+08	8.400000e+08	3.447000e+09	1.540000e+09	574000000.0	...	5065000000	-1

16707 rows × 25 columns

Figure 6. The Annual Cash Flow Dataset

Ticker	Report Date	SimFinId	Currency	Fiscal Year	Fiscal Period	Publish Date	Shares (Basic)	Shares (Diluted)	Revenue	Cost of Revenue	Gross Profit	...	Non-Operating Income (Loss)	
A	2008-10-31	45846	USD	2008	FY	2009-10-05	3.630000e+08	3.710000e+08	5.774000e+09	-2.578000e+09	3.196000e+09	...	20000000.0	
	2009-10-31	45846	USD	2009	FY	2009-12-21	3.460000e+08	3.460000e+08	4.481000e+09	-2.189000e+09	2.292000e+09	...	-40000000.0	
	2010-10-31	45846	USD	2010	FY	2010-12-20	3.470000e+08	3.530000e+08	5.444000e+09	-2.514000e+09	2.930000e+09	...	-6000000.0	
	2011-10-31	45846	USD	2011	FY	2011-12-16	3.470000e+08	3.550000e+08	6.615000e+09	-3.086000e+09	3.529000e+09	...	-39000000.0	
	2012-10-31	45846	USD	2012	FY	2012-12-20	3.480000e+08	3.530000e+08	6.858000e+09	-3.254000e+09	3.604000e+09	...	-76000000.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
low	2014-02-28	186050	USD	2013	FY	2014-03-31	1.059000e+09	1.061000e+09	5.341700e+10	-3.494100e+10	1.847600e+10	...	-476000000.0	-4
	2015-02-28	186050	USD	2014	FY	2015-03-31	9.880000e+08	9.900000e+08	5.622300e+10	-3.666500e+10	1.955800e+10	...	-516000000.0	-5
	2016-02-29	186050	USD	2015	FY	2016-03-29	9.270000e+08	9.290000e+08	5.907400e+10	-3.850400e+10	2.057000e+10	...	-552000000.0	-5
	2017-02-28	186050	USD	2016	FY	2017-04-04	8.800000e+08	8.810000e+08	6.501700e+10	-4.255300e+10	2.246400e+10	...	-645000000.0	-6
	2018-02-28	186050	USD	2017	FY	2018-04-02	8.390000e+08	8.400000e+08	6.861900e+10	-4.521000e+10	2.340900e+10	...	-633000000.0	-6

16707 rows × 25 columns

Figure 7. The Annual Income Statement Dataset

Date	Unemployment Rate	Unemployment Rate QOQ	Unemployment Rate MOM	Unemployment Rate YOY
2007-01	4.6	0.045455	0.045455	-0.021277
2007-02	4.5	0.000000	-0.021739	-0.062500
2007-03	4.4	0.000000	-0.022222	-0.063830
2007-04	4.5	-0.021739	0.022727	-0.042553
2007-05	4.4	-0.022222	-0.022222	-0.043478
...	...	...	...	...
2019-09	3.5	-0.054054	-0.054054	-0.054054
2019-10	3.6	-0.027027	0.028571	-0.052632
2019-11	3.5	-0.054054	-0.027778	-0.054054
2019-12	3.5	0.000000	0.000000	-0.102564
2020-01	3.6	0.000000	0.028571	-0.100000

157 rows x 4 columns

Figure 8. The Unemployment Rate Dataset

1240864186357305345	2020-03-20 04:54:10	3016	25m share XT in $MMM there @ 40c	
1240853441456594945	2020-03-20 04:11:29	4101	Q1 YoY sales up 40% with only the first 2 weeks being a catalyst to the surge. Expect a further surge in sales for Q2 as bans on number of food items can buy in the supermarkets. $MMM @MarleySpoon in fact does not appear to have these issues in getting you the food you need.	
1240844876821880833	2020-03-20 03:37:27	86	The US dollar hits resistance, USD needs a great sell off @jimcramer $hnc $teva $mmm $ge $dia $uco $gld $iwm $fxi $xle $xlb $xlf $xlk $xli $xlu $xlv $xly $xlp $ipm $gs $bac $xom $cvx $hd $wmt $dis $mmm $mrk $aapl $intc $msft $amzn $pypl $nflx $twtr $tpr $tgt $fcx $fslr $bb	
1240843237033742336	2020-03-20 03:30:56	207	I'm with you! I will get hurt if/when this occurs. At the same time, I'm glad that there are numerous Aristocrats that are more than just "money machines" and trying to make a positive difference in this chaotic era... ie: $WMT, $JNJ, $CLX, $MMM, $WBA, $ABBV, $TGT, $CL, $PG, $KMB https://t.co/0qz4mwCeT6	
1240829875948150784	2020-03-20 02:37:50	7110	I'm with you! I will get hurt if/when this occurs. At the same time, I'm glad that there are numerous Aristocrats that are more than just "money machines" and trying to make a positive difference in this chaotic era... ie: $WMT, $JNJ, $CLX, $MMM, $WBA, $ABBV, $TGT, $CL, $PG, $KMB https://t.co/0qz4mwCeT6	
1240824044989054978	2020-03-20 02:14:40	77	Quarantine survival kit... 👌😇 very peaceful... just my interest me profession and hobbies. @YETICoolers @beatsbydre @Bose @BenchmadeKnives @ROLEX @3M #n95 #8210plus @lagunitasbeer @Corkcicle @CcelLofficial @Apple $AAPL $YETI $MMM $AMZN $SPY $V $FB $MSFT $INTC $AMD $QQQ $VIX https://t.co/2eAqMNGdHN	
1240818413443977217	2020-03-20 01:52:17	56	Quarantine survival kit... 👌😇 very peaceful... just my interest me profession and hobbies. @YETICoolers @beatsbydre @Bose @BenchmadeKnives @ROLEX @3M #n95 #8210plus @lagunitasbeer @Corkcicle @CcelLofficial @Apple $AAPL $YETI $MMM $AMZN $SPY $V $FB $MSFT $INTC $AMD $QQQ $VIX https://t.co/2eAqMNGdHN	
1240811368246173698	2020-03-20 01:24:18	5469	I really believe our best companies like @3M can make the changes in supply chain needed to relieve burdens hospitals feel right now in protective equipment. Talked to several hospital executives today who praised $MMM as a great partner.	
1240810129215213568	2020-03-20 01:19:22	16655	I really believe our best companies like @3M can make the changes in supply chain needed to relieve burdens hospitals feel right now in protective equipment. Talked to several hospital executives today who praised $MMM as a great partner.	
1240807659827167232	2020-03-20 01:09:33	1050	If anyone want to know what $MMM will be doing over the next few days, have a look at the $APRN chart https://t.co/XhEhyRBKSg	
1240803771933806594	2020-03-20 00:54:06	954	$MMM great sales update for pre made meals and delivered these stocks are bagging on the Nasdaq	
1240799646919692288	2020-03-20 00:37:43	954	$mmm home food delivery. Should bunt hard. https://t.co/Cv10pWlqs9	
1240793395468963840	2020-03-20 00:12:52	345	$MMM great sales update for pre made meals and delivered these stocks are bagging on the Nasdaq	
1240793139608084481	2020-03-20 00:11:51	345	$MMM 43.5 29% Ann: COVID-19 - Increased Sales and Deliveries	
1240793088555008000	2020-03-20 00:11:39	3016	$MMM great sales update for pre made meals and delivered these stocks are bagging on the Nasdaq	
1240791691373625344	2020-03-20 00:06:06	63	And there's the $MMM release - really solid numbers for what's generally their strongest quarter. Next quarter could be massive.	
1240790881910702080	2020-03-20 00:02:53	345	$MMM Increased Sales and Deliveries due to #COVID19 https://t.co/wOXvWrGePD	
1240780807311490048	2020-03-19 23:23:13	1055	$HON $MMM $LVMUY $GE $F: Ahead of potential shortages of supplies and equipment needed for hospitals to treat Covid…: https://t.co/9YdTTkYmN0	
1240776478582849537	2020-03-19 23:05:39	1327	Check which story for #DowJones has the highest score $V $MMM $MCD https://t.co/8zP8MKc2dK https://t.co/GeaUeayjIF	
1240769023434002436	2020-03-19 22:36:02	1451	Short sale volume (not short interest) for $APT at 2020-03-18 is 47%. https://t.co/PybtaIQOVr $MMM 37% $ALGN 47% $CAH 33% $MLSS 40%	
1240762902656753664	2020-03-19 22:11:42	6405	The Spanish Flu — Coo Coo Ca Choo — Stock Market (And Sentiment Results)… #coronavirus $JPM $MMM $PFE $UTX $WFC $CSCO $DIA $SPY $ABBV $SPX $DJI https://t.co/NVFugeiY8b	
1240759419060830210	2020-03-19 21:57:52	968	$MMM German customs seized protective masks and gear at Jüchen 3M plant as it was suspected to be illegally exported to US and Switzerland - Rheinische Post https://t.co/UqAINsqPZD	
1240757918879940614	2020-03-19 21:51:54	6405	The Spanish Flu — Coo Coo Ca Choo — Stock Market (And Sentiment Results)… $JPM $MMM $PFE $UTX $WFC $CSCO $DIA $SPY $ABBV $SPX $DJI https://t.co/NVFugeiY8b	
1240746625670275074	2020-03-19 21:07:02	3493	The Spanish Flu — Coo Coo Ca Choo — Stock Market (And Sentiment Results). $JPM $MMM $PFE $UTX $WFC $CSCO $DIA $SPY $ABBV $SPX $DJI https://t.co/TGZ0bESqMm	
1240743165788917760	2020-03-19 20:53:17	187	$HE $SWAY $HD $MMM! $I $AM $MR $AFT $TWO	
1240740658429734912	2020-03-19 20:43:19	336	On the watch list for tomorrow... Longs in $CRWD $NTES $DOCU $ROKU $NVDA $TWLO and $MMM. 📈 Staying short in $VXX $WMT and $DE. 📉 Watching $ZM $FB $BABA $ADBE and $FDX at the open. 👀 #StocksToTrade #StocksToWatch #StockMarket #OptionsTrading #SwingTrading #Friyay	
1240719044824772612	2020-03-19 19:17:26	4101	No worries $MMM got this! Production ramp up & help on way https://t.co/lyGRemYyey	
1240710158264799232	2020-03-19 18:42:07	2425	$MMM 19-Jun-20 ATM Implied Vol Climbs +5.9%. Straddle Implies a Move of ±26.9% https://t.co/W5himzXyob	
1240709647033675776	2020-03-19 18:40:05	2425	$MMM Option Order Flow Sentiment is 89.6% Bullish. https://t.co/MANLlNBCKq	
1240708884983144450	2020-03-19 18:37:04	377	No worries $MMM got this! Production ramp up & help on way https://t.co/lyGRemYyey	
1240704359127638016	2020-03-19 18:19:05	4101	$fdx $mmm $csco $estc $roku guess i ll wait for $mmm https://t.co/SHewem9tRk	
1240698020124962824	2020-03-19 17:53:53	463	$fdx $mmm $csco $estc $roku guess i ll wait for $mmm https://t.co/SHewem9tRk	
1240682121674588160	2020-03-19 16:50:43	1344	$MMM out small gain... jm done for now.. nothing gonna rip that I see.. will sit untill nice shreds	
1240679805110104065	2020-03-19 16:41:30	1344	$MMM pop... slow one.. with so much chatter about them having more masks to supply... one would think would go more.. but i have low expectiations.. people want out of the market for now not in... and no good news today.. except we borrows some masks from some consutions dudes	
1240679187427594240	2020-03-19 16:39:03	1344	long small position $MMM if can go over $136	
1240677353245687808	2020-03-19 16:31:46	39	My fav picks: $MMM 3M - 3.94 $DIS Disney - 1.84 🤔 $CAT Caterpillar - 2.37 🤔 $PEP Pepsi - 3.88 $JNJ J&J - 4.01 $HD Home Depot - 5.13 $NKE Nike - 6.33 $SQ Square - 5.54 $KDP Keurig Dr.P - 1.00 😇	
1240677045044199426	2020-03-19 16:30:32	172	Are any market participants paying attention to this disaster press conference? Pence just said $MMM makes 35mill masks per month BUT only 5mill are QUALIFIED for Hospital use MEANWHILE 10 minutes ago, it was said that construction quality masks were being given to hospitals!!!	
1240676001144885254	2020-03-19 16:26:23	1186	$MMM $WHR not moving w/this little rally today, despite #3M increasing masks production and me doing 2 loads of disinfecting laundry a day in my @whirlpoolusa washer and dryer! I like both here near their 52 week lows for a tiny bite... @CabotDividends @jimcramer @RickDucat_TDAN https://t.co/rQT8HnzxyX	
1240673917053947904	2020-03-19 16:18:07	0	$MMM makes the M95 masks the President is discussing.	
1240673503638102019	2020-03-19 16:16:28	630	Are any market participants paying attention to this disaster press conference? Pence just said $MMM makes 35mill masks per month BUT only 5mill are QUALIFIED for Hospital use MEANWHILE 10 minutes ago, it was said that construction quality masks were being given to hospitals!!!	
1240673337296240640	2020-03-19 16:15:48	28263	Pence Says 3M Increasing Face Mask Production to 420M/Yr $MMM https://t.co/dsZBJP6Wzt @benzinga	
1240663103660085254	2020-03-19 15:35:08	2440	Peak profit for the last 6 expired option alerts for $MMM 67.92 % | 80.47 % | 428.44 % | 497.35 % | 452.76 % | -31.43 % | https://t.co/y4cQtVXdaa https://t.co/u2QAJdnAJ4	
1240661078763339776	2020-03-19 15:27:06	2440	$MMM - Last six months, 33 option alerts peaked above 100% after they were triggered by the algo https://t.co/y4cQtVXdaa https://t.co/nazV1HyWbM	
1240660062567350274	2020-03-19 15:23:03	2440	$MMM - View historical options performance for MMM https://t.co/y4cQtVXdaa https://t.co/nc3zrtt24G	
1240653030451818497	2020-03-19 14:55:07	853	Top 10 Dividend Aristocrats to Buy In This Crazy Market $ABBV Abbvie, $AFL Aflac, $CAT Caterpillar, $CB Chubb, $EMR Emerson, $GD General Dynamics, $GWW Grainger, $MMM 3M, $PPG PPG Industries, $SWK Stanley Back & Decker https://t.co/gfjubq5FZr	

Figure 9. Sample Twitter Dataset

	(Dividends + Share Buyback) / FCF	Asset Turnover	CapEx / (Depr + Amor)	Current Ratio	Debt Ratio	Dividends / FCF	Dummy_Dividends	Gross Profit Margin	Interest Coverage	Inventory Turnover	...	x0_Consumer Cyclical	x0_Consumer Defensive
0	0.168927	0.711759	0.590551	3.201005	0.247890	-0.000000	0	0.532001	10.890411	7.216882	...	0.0	0.0
1	0.185950	0.738375	0.603113	3.470432	0.247687	-0.000000	0	0.532570	13.927536	7.205128	...	0.0	0.0
2	0.177064	0.730374	0.671937	3.031573	0.241250	-0.000000	0	0.533485	14.875000	7.366370	...	0.0	0.0
3	0.045863	0.739752	0.706349	3.351906	0.239697	-0.000000	0	0.532907	14.881579	7.183565	...	0.0	0.0
4	0.162579	0.721024	0.658635	3.275204	0.231276	0.031789	1	0.531163	14.207317	7.166843	...	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
12555	0.624005	0.540503	0.915584	3.596893	0.598682	0.161141	1	0.671931	9.131068	4.187635	...	0.0	0.0
12556	0.616107	0.543416	0.867052	4.047135	0.592116	0.174497	1	0.664863	8.595349	4.345334	...	0.0	0.0
12557	0.576439	0.550337	0.866142	4.171163	0.586747	0.183984	1	0.669203	8.437500	4.306268	...	0.0	0.0
12558	0.624564	0.545600	0.900990	4.389734	0.571948	0.205862	1	0.673821	8.646018	4.273801	...	0.0	0.0
12559	0.693215	0.542226	1.065534	2.629014	0.558424	0.231563	1	0.681789	9.049327	4.439716	...	0.0	0.0

12560 rows × 65 columns

Figure 10. Overview of Quarterly Features

	x0_Basic Materials	x0_Business Services	x0_Consumer Cyclical	x0_Consumer Defensive	x0_Energy	x0_Financial Services	x0_Healthcare	x0_Industrials	x0_Other	x0_Real Estate	...	Treasury 1-Year Rate	Treasury 3-Year Rate
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.10	0.37
1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.09	0.33
2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.09	0.35
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.09	0.33
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.10	0.33
...	...	...	...	...	...	...	...	...	...	...	...	...	...
861568	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.15	0.19
861569	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.15	0.21
861570	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.16	0.24
861571	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.16	0.22
861572	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.15	0.20

861573 rows × 83 columns

Figure 11. Overview of Daily Features

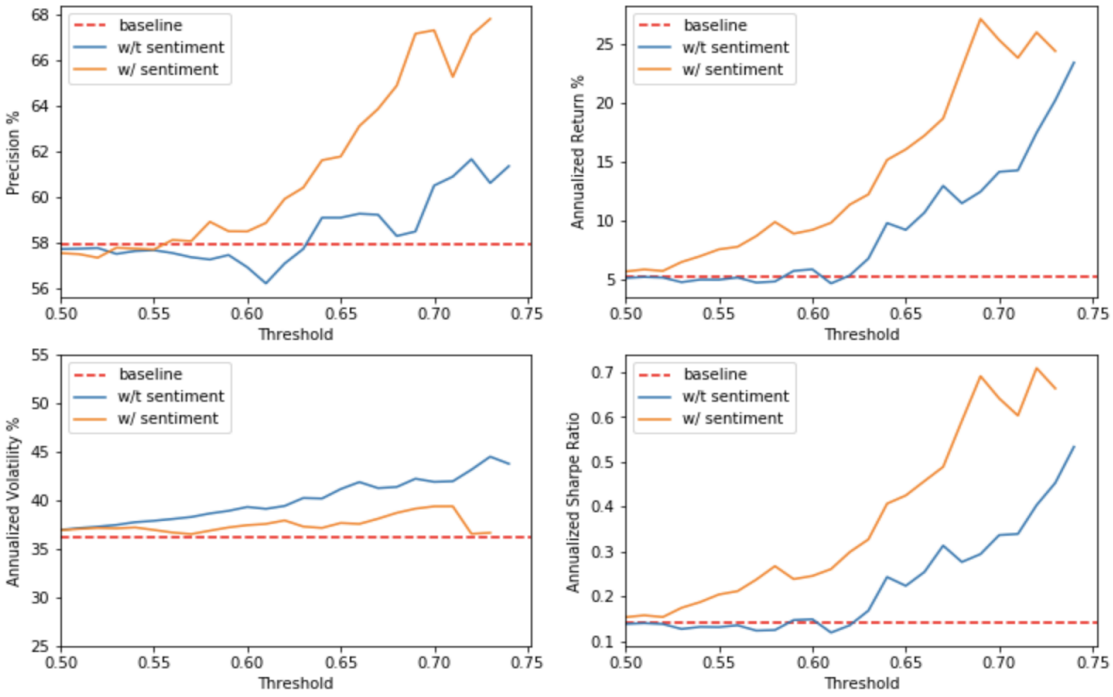

Figure 12. Backtesting Performance (QOQ Prediction)

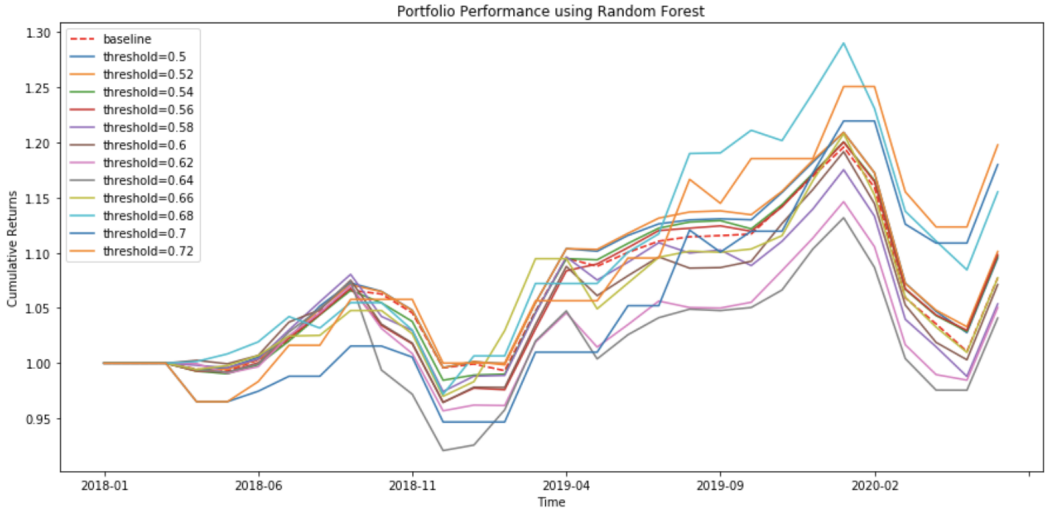

Figure 13. Portfolio Cumulative Return (QOQ Prediction)

	Threshold	Precision	Annualized Return	Annualized Volatility	Annualized Sharpe Ratio	Number of Holding Stocks
0	0.5	76.773825	15.150748	25.801156	0.470938	40252
1	0.6	80.856640	17.946314	24.725981	0.604478	25565
2	0.7	84.772130	20.134722	23.543019	0.727805	13692
3	0.8	87.312659	21.552814	23.405389	0.792673	10609
4	0.9	91.116751	22.752109	22.474815	0.878855	3940
5	1.0	92.866941	22.462708	20.942059	0.929360	729

Figure 14. Backtesting Performance (1 Year Prediction)

	Threshold	Precision	Annualized Return	Annualized Volatility	Annualized Sharpe Ratio	Number of Holding Stocks
0	0.5	82.553603	19.270455	22.401278	0.726318	34093
1	0.6	83.978655	20.050691	21.848517	0.780405	32232
2	0.7	86.312839	21.322157	21.183320	0.864933	28545
3	0.8	89.582168	22.868614	19.955834	0.995629	21444
4	0.9	93.737039	24.108630	17.313443	1.219205	9644
5	1.0	0.000000	NaN	NaN	NaN	0

Figure 15. Backtesting Performance on time-series clustering using Euclidean distance (1 Year Prediction)

	Threshold	Precision	Annualized Return	Annualized Volatility	Annualized Sharpe Ratio	Number of Holding Stocks
0	0.5	86.085536	19.802654	18.988376	0.884892	33670
1	0.6	87.037523	20.443244	18.968825	0.919574	30488
2	0.7	89.042934	21.500412	18.444775	1.003016	24596
3	0.8	91.192373	22.194520	17.687298	1.085215	17201
4	0.9	94.034091	23.281525	16.396079	1.236974	7040
5	1.0	94.654088	23.786604	15.734929	1.321048	318

Figure 16. Backtesting Performance on time-series clustering using DTW distance (1 Year Prediction)

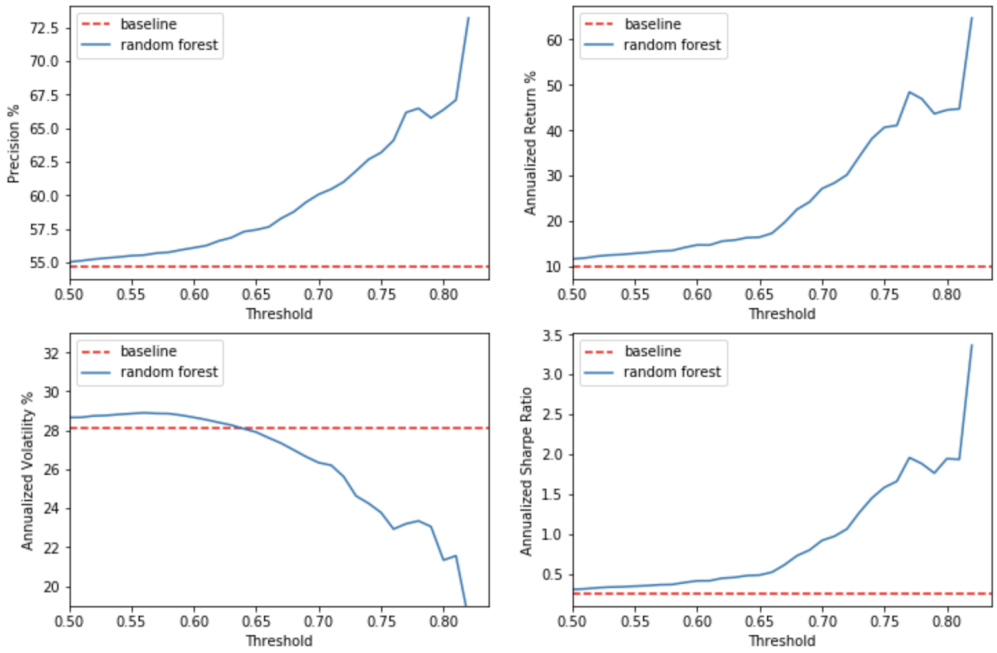

Figure 17. Backtesting Performance (WOW Prediction)

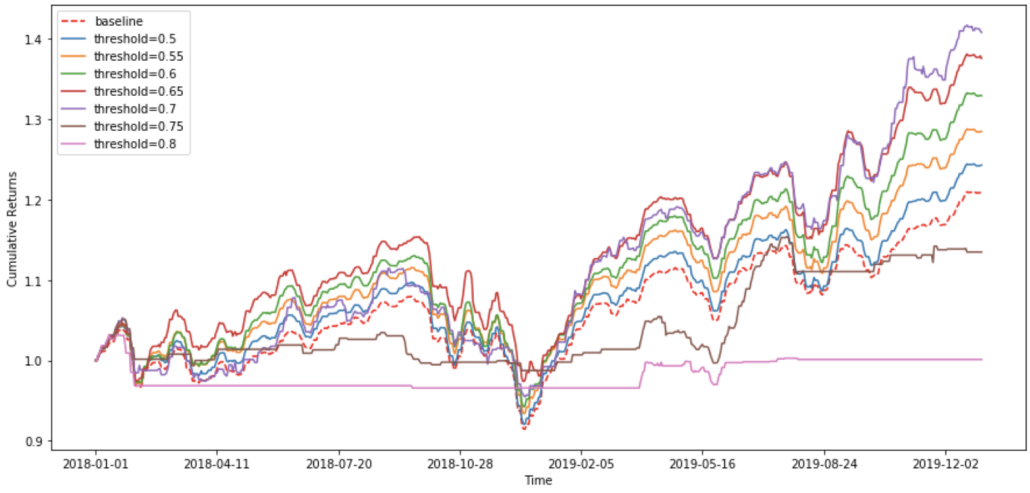

Figure 18. Portfolio Cumulative Return (WOW Prediction)