561 Final Project

TA Name: Karthikeyan K.

Group Members:
Elinor Cheng (yc493)
Haoran Liu (hl449)
Junghwa Jang (jj368)
Jason Liu (jl1184)
Qiyu Wu (qw112)
Yinting Zhong (yz790)

April 27, 2023

PR Statement

Duke University Students Launch Stock Price Prediction Tool to Promote Healthy Investment Culture

Durham, North Carolina – Duke University students have unveiled a new tool to support stock investment decisions. Combining their expertise in finance, statistics, and computer science, the interdisciplinary group of students aimed to tackle the perceived black-box nature and element of luck that often characterizes stock market investment culture. Their recently launched tool is expected to provide investors with not only a "go or not go" decision, but also a better understanding of the investment process, with the ultimate goal of cultivating a healthier investment culture that promotes safer and more successful investments.

The decision support tool was developed by integrating financial analysis with machine learning methodology. The students selected the most relevant macroeconomic and financial variables from various sources, collected stock market data from (Source), and processed it using feature engineering to capture various characteristics of stock market data. The financial performance of each firm was captured through the Wharton Research Data Services (WRDS), and then processed using the EM algorithm to be adjusted into the correct data format. They also collected macroeconomic variables and processed them with the time series algorithm, the Seasonal Trend Decomposition (STD), which adjusts the macroeconomic variables into the given data format for better prediction performance. Finally, the students collected social listening data and used the FinBERT model to calculate the daily sentiment scores in financial text. For the pilot study, they used 10 years of data (2010-2020) from the top 10 most influential tech companies during the last decade, such as Apple, Amazon, and Meta.

The pilot study was conducted using New York Times articles, but the students plan to expand the social listening data scope to include not only more news articles but also social media such as Twitter, Facebook, Instagram, and TikTok, as well as online

communities. Research has shown that online articles and social media activities can directly influence stock prices, as evidenced by Elon Musk's Dogecoin incident.

The students emphasized that their stock price decision support tool is not intended to be an oracle or a golden goose in the stock investment market. Rather, they hope it will provide people with a chance to better understand financial investments and cultivate a healthier investment culture. They want their platform to serve as a minimal guard against the black-box and reckless investment habits that can be costly, even life-altering, for some people. They aim to deliver not only simple "go or not go" decision-making advice, but also provide accessible education to general investors. Additionally, they plan to develop a platform to advise firms on financial decisions and marketing/PR decisions.

The students view their contribution as a bridge between financial analysis, social listening analysis, and machine learning techniques. They plan to expand their data collection scope to further enhance the performance of their algorithms.

Frequently Asked Questions

What value can your service provide?

Our service provides valuable insights into the business ecosystem by leveraging the integration of once scattered and heterogeneous large-scale data sets. We offer value to both general investors and firms by accurately predicting the stock market prices of a target company and providing explanations of why the investor's decision is encouraged or discouraged based on various financial, economic, and social listening perspectives. Additionally, our algorithm can diagnose the financial soundness and marketing/PR effectiveness of a given company, which can directly measure its stock market performance and help companies improve their overall business performance and growth.

What are unique or highly compelling aspects of your service?

Our service is compelling because we have adopted a groundbreaking approach to the integration of once scattered and heterogeneous data sets, which directly influence stock market performance. We incorporate social listening data, which is key to our service, and we plan to expand our scope to broader online social data, such as social media, Twitter, Facebook, Instagram, and TikTok, and online communities such as Reddit, etc. We are the first to attempt incorporating social listening data in investment prediction services, which can provide more meaningful insights to both general investors and firms. Additionally, our algorithm is not a black-box model, and it can diagnose the contributing predictors and how they contributed to the final result, which sets us apart from our competitors.

How does the minimal viable product/service (MVP) work?

Our service will be developed in two perspectives. First, for general investors, we will provide advice on their investment decisions and later investment portfolio. We will train our model using the (MODEL) model and provide advice on their choice of a target company, whether their investment decision is encouraged or discouraged.

Based on the individual companies' stock price prediction, we will integrate the information and provide advice on whether the portfolio choice is encouraged or discouraged. We will also provide a brief explanation of this decision to support investors in understanding the overall stock market environment.
Second, for corporate consumers, we will provide diagnostic tools for their financial soundness and marketing/PR performance evaluation. Our service utilizes a broad range of firms' financial performance and digital marketing indices through social listening data. Based on our current algorithm, we can develop a diagnostic algorithm to identify further strategies to improve current firm performances. For example, we can identify whether net profit margin is the deteriorating factor in the financial soundness of the given firm and provide an initial diagnosis to guide them in the right direction for following financial strategies. We can also provide advice on which social media platform is more influential for the firm's performance and which platform the firm should put more work into.

Are there any patentable components in your service, or in the near future?

We expect to further optimize our algorithms and develop our diagnostic algorithms, which we expect will be patentable in the future. Possessing a patent will provide a great leverage for our service in terms of marketing and broader adoption of our service.

Who are your target users of the technology?

We have two primary target groups. First, general investors, who we plan to provide services to in the form of a mobile application or web-based services. Second, corporate consumers, who we plan to provide with our services in the form of a web report. If requested, we can also provide ad-hoc and customized platform services for them.

How is the system trained to make stock price predictions?

To train our service, we will use data from four distinctive and heterogeneous datasets. First, we will use stock price data to produce 12 variables that capture the characteristics of stock price data. Our feature engineering algorithms will process this data. Second, we will use firms' financial statement data from the Wharton Research Data Service (WRDS), which provides firm-level financial variables for all U.S. companies across eight different characteristics. We will adjust this dataset with the EM algorithm to match the stock price data. Third, we will use macroeconomic indices, which will be processed using the Seasonal Trend Decomposition (STD) to adjust into the stock price data. Finally, we will collect and process social listening data using the FinBERT model to extract sentimental scores.
We will apply (MODEL) to predict the stock price and evaluate the importance of each variable to provide meaningful information for general investors and firm consumers.

Who are the existing or potential competitors in this area and how does your service differentiate itself from them?

While there are many platforms to predict stock prices, none have attempted to incorporate the breadth of datasets we are utilizing, particularly social listening data. Our service not only predicts the stock market data but also provides insights for future investment and for firms to advance their financial and marketing performance.

How will legal responsibilities be addressed if the system provides inaccurate suggestions?

To mitigate the liabilities of our erroneous suggestions, we will follow a similar approach to the prediction system for general investors by including a liability exemption clause in the terms and conditions and obtaining consent from users. For corporate users, we will follow the traditions of professional services, such as consulting firms, marketing research firms, or ad agencies.

Are there Privacy Concerns when you incorporate the various data sets?

Most of our data sources are publicly accessible, and for the social listening data, we only use publicly open sources. Even if we were accessible to private social listening data, it is not relevant to our model since stock market prices are not influenced by private social data.

What is your marketing plan?

We plan to release our beta version of the application for general users and collect consumer testimonies to use for future advertising. Based on our initial trial, we can implement influencer marketing by collaborating with famous investing advisor channels like Nick Black or Garrett Ashe. For corporate users, we can launch pilot projects with influential firms and international firms to build our client portfolio in the U.S. and overseas.

What resources will be needed to further improve the service?

To improve our service, we will require financial investment and business sway to expand and develop it further. First, we will need funding to hire more personnel to develop our mobile application and web report platform. We will also optimize our algorithm to make it more applicable in a broader context, such as the entire stock market and overseas markets. Additionally, we need to expand our social listening data collection, for which we will need to build an MOU with stronger social listening platforms like Brandwatch consumer intelligence services.

1 Data Collection

1.1 Data collection overview

To develop our stock trade decision support tool, we tested various machine learning approaches to identify the models with the best prediction performance. We also gathered data from publicly available sources, including stock market APIs, company financial reports, macroeconomic indices, and New York Times article APIs, which we used as training data. The data collection period spanned 10 years, from January 1, 2010, to December 31, 2019. We intentionally excluded data from 2020 due to the disruptive effects of the Covid-19 pandemic on the stock market. In this section, we provide detailed explanations of each data source's fields and a broad overview of the data.

1.2 Stock price data

To begin, we gathered data from the stock market API. Due to the immense amount of trade data spanning 10 years, we limited our selection to the top 10 tech industry stocks during this period.

Figure 1: The list of 10 Companies for this project

Abbreviation	Company
AAPL	Apple Inc.
AMZN	Amazon.com, Inc.
BRK-B	Berkshire Hathaway Inc Class B
GOOG	Alphabet Inc Class C (Google)
JNJ	Johnson & Johnson
META	Meta Platforms Inc (Facebook)
MSFT	Microsoft Corp
NVDA	NVIDIA Corporation
TSLA	Tesla Inc
V	Visa Inc

Our stock price data includes the following data fields, and we utilized the "Adjusted Close" price as the dependent variable (DV) because it accurately reflects the stock price on a given day:

Figure 2: The data fields and descriptions

Data Field	Description
Date	The day the stock was traded
Open	The opening stock price
Close	The closing stock price
High	The highest stock price of the day
Low	The lowest stock price of the day
Adjusted Close	The closing price after adjustments for all applicable splits and dividends to the Center for Research in Security Prices (CRSP) standards
Volume	The amount of the asset or security traded during the day
Dividend Amount	The amount of the dividend paid per share of Common Stock, multiplied by (x) the Purchase Amount divided by (y) the Liquidity Price
Split Coefficient	The number of stocks that were changed after the procedure

Moreover, we collected stock prices on a daily, weekly, and monthly basis, in line with the stock market's intrinsic characteristic of trading only on weekdays. Furthermore, as other financial and economic variables are generated on a monthly, quarterly, and even yearly basis, we aimed to determine the optimal data collection period through feature engineering and preliminary analysis.

1.3 Fundamental Indicators (Financial Ratios)

Financial ratios, as fundamental indicators, have been extensively studied to explain the performance of firms and the stock market (Beaver, 1968; Lewellen, 2004; Siew and Nordin, 2012). For instance, Beaver (1968) was among the first researchers to analyze the impact of financial ratios on stock market prices, while Lewellen (2004) developed more accurate prediction approaches using financial ratio data. Moreover, recent studies, such as Siew (2012), have advanced prediction techniques by utilizing machine learning.

To apply these findings to our prediction model, we collected financial ratio data from Wharton Research Data Services (WRDS), which provides over 70 pre-calculated firm-level financial ratios for all U.S. companies across 8 different categories, including valuation, liquidity, profitability, and financial soundness. The original accounting data is obtained from Compustat Quarterly and Annual file, which we had to adapt to daily, weekly, and monthly basis datasets through feature engineering. Additionally, we used CRSP and IBES databases to collect price-related data and earnings-related data. Out of the over 70 financial ratios available, we selected the 16 most relevant data fields, as shown in Figure.

Figure 3: The financial ratios

Data Fields	Descriptions
pcf	Stock price to cash flow ratio
PEG_trailing	Trailing P/E to growth ratio
dpr	Dividend payout ratio
npm	Net profit margin
gpm	Gross profit margin
roa	Return on assets
roe	Return on equity
capital_ratio	Capitalization ratio
de_ratio	Total debt to equity ratio
cash_ratio	Cash ratio
curr_ratio	Current ratio
inv_turn	Inventory turnover
pay_turn	Payables turnover
sale_nwc	Sales to working capital ratio
rd_sale	Research and development to sales ratio
accrual	Accruals to average assets ratio

1.4 Macroeconomic Variables

Research in stock market forecasting has demonstrated that macroeconomic variables have a significant impact on prediction performance (Mukherjee and Naka, 1995; Tinoco and Wilson, 2013). Including the dynamics of the macroeconomic

environment has been found to enhance the prediction of various economic events, such as economic distress and the behavior of stock markets. Therefore, we collected ten years' worth of macroeconomic variables shown in Figure 4.

Incorporating this macroeconomic data with our daily, weekly, and monthly stock market price data posed a challenge, as much of the macroeconomic data is generated annually or quarterly, which is a longer term compared to our stock price data. To address this, we employed various feature engineering methodologies, which we will explain in the feature engineering section.
The following macroeconomic variables were collected:

Figure 4: The macroeconomic variables and definition

Data	Description	Data Period
Real GDP	An annual data type that reflects the value of all goods and services produced by an economy in a given year, adjusted for inflation	Annual
Real GDP per capita	A quarterly data type that is calculated by dividing the GDP at constant prices by the population of a country or area	Quarter
Treasury Yield	Short-term (3-month) and long-term (30 years) government debt obligations, indicating how much investors can earn when they purchase them	Short-term (3 months) Long-term (30 years)
Federal Funds Rate	It refers to the target interest rate set by the Federal Open Market Committee (FOMC)	Annual
CPI	Monthly data that measures the monthly change in prices paid by U.S. consumers through the Consumer Price Index (CPI)	Month
Inflation	Annual data that measures the rate of increase in prices over a given period of time	Annual
Unemployment	The unemployment rate, which represents the percentage of unemployed people in the labor force	Annual
Nonfarm Payroll	Monthly data on the number of U.S. workers in the economy, excluding proprietors, private household employees, unpaid volunteers, farm employees, and the unincorporated self-employed	Monthly

1.5 New York Times API

In recent years, researchers have increasingly turned to social listening data such as Twitter and news articles to improve stock market predictions (Korivi et al., 2022; Li et al., 2020; Swathi et al., 2022). For instance, Korivi et al. (2022) and Swathi et al. (2022) use Twitter data to generate sentiment-based features for stock price forecasting, while Li et al. (2020) conduct a case study of Hong Kong stock market price prediction using sentiment analysis of news articles. These studies demonstrate the effectiveness of sentiment analysis in improving the accuracy of stock market predictions.

Building on this literature, we have collected social listening data from a free API that is relevant to stock price prediction. As one of the most reputable media companies, the New York Times is ranked second in the category of online news and publishers in the United States. To collect our data, we used the New York Times API, which is an open source. Specifically, we focused on three data fields, as readers tend to skim headlines and the lead paragraph rather than read the entire article. We collected 10 years' worth of data using R.

Figure 5: The NYT API data fields and descriptions

Data Field	Description
Publish Date	The date when the news article was published.
Main	The main headline of the news article.
Lead Paragraph	A brief summary of the news article in one paragraph.

2 Data Processing and Feature Engineering

2.1 Data Processing

2.2 Feature Engineering

Stock Price Feature Engineering

We performed feature engineering on the collected stock price data to capture various distribution characteristics, which we utilized in our prediction model. The target variables for our models were Difference and Direction, while the remaining ten variables served as predictors.

- Difference: The difference between the adjust close price of the day t and the adjust close price of the day t-1. We used two different versions: Direction1 and Direction2

$$Direction1 = AdjustClosePrice_t - AdjustClosePrice_{t-1}$$

$$Direction2 = AdjustClosePrice_{t+13} - AdjustClosePrice_{t-1}$$

- Direction: The dummy variable of **Difference**

$$f(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- Simple Moving Average (SMA): an arithmetic moving average calculated by adding recent prices and then dividing that figure by the number of time periods in the calculation average.

$$SMA = \frac{A_1 + A_2 + ... + A_n}{n}$$

, where $A_n = thepriceofanassetatperiodn$ and $n = thenumberoftotalperiods$

- Exponential Weighted Moving Average (EWM): An exponential moving average (EMA) is a type of moving average (MA) that places a greater weight and significance on the most recent data points.

$$EMA_t = (Value_t * \left(\frac{Smoothing}{1 + Days}\right) + EMA_{t-1} * \left(1 - \left(\frac{Smoothing}{1 + Days}\right)\right)$$

- The Relative Strength Index (RSI): The Relative Strength Index (RSI), developed by J. Welles Wilder, is a momentum oscillator that measures the speed and change of price movements. The RSI oscillates between zero and 100.

$$RSI_{stepone} = 100 - \left[\frac{100}{1 + \frac{Averagegain}{Averageloss}}\right]$$

- The Money Flow Index (MFI): The Money Flow Index (MFI) is a technical indicator that generates overbought or oversold signals using both prices and volume data.

$$MoneyFlowIndex = 100 - \frac{100}{1 + MoneyFlowRatio}$$

, where $MoneyFlowRatio = \frac{14PeriodPositiveMoneyFlow}{14PeriodNegativeMoneyFlow}$,
$RawMoneyFlow = TypicalPrice * Volume$, and $TypicalPrice = \frac{High+Low+Close}{3}$

- The Stop and Reverse (SAR) Index: The parabolic SAR (stop and reverse) indicator is used by technical traders to spot trends and reversals.

$$RPSAR = PriorPSAR + [PriorAF(PriorEP - PriorPSAR)]$$

$$FPSAR = PriorPSAR - [PriorAF(PriorPSAR - PriorEP)]$$

, where $RPSAR = RisingPSAR$, $FPSAR = FallingPSAR$,
$AF = AccelerationFactor$, $EP = ExtremePoint$

- The Advance/Decline (A/D) line: The advance/decline line (or A/D line) is a technical indicator that plots the difference between the number of advancing and declining stocks on a daily basis.

$$A/D = NetAdvances + \begin{cases} PA, & \text{if PA value exists} \\ 0, & \text{if no PA value} \end{cases}$$

, where $NetAdvances =$
$Difference between number of daily ascending and declining stocks$,
$PA = PreviousAdvances$, and $PreviousAdvances = Prior indicator reading$

- Money Flow Multiplier (MFM): The accumulation/distribution indicator (A/D) is a cumulative indicator that uses volume and price to assess whether a stock is being accumulated or distributed.

$$MFM = \frac{(Close - Low) - (High - Close)}{High - Low}$$

- The Volume-Weighted Average Price (VWAP): The volume-weighted average price (VWAP) is a technical analysis indicator used on intraday charts that resets at the start of every new trading session.

$$VWAP = CumulativeTypicalPrice * \frac{Volume}{CumulativeVolume}$$

- The Moving Average Convergence / Divergence (MACD): Moving average convergence/divergence (MACD, or MAC-D) is a trend-following momentum indicator that shows the relationship between two exponential moving averages (EMAs) of a security's price.

$$MACD = 12 - periodEMA - 26 - periodEMA$$

- The Stochastic Oscillator (STOCH): A stochastic oscillator is a momentum indicator comparing a particular closing price of a security to a range of its prices over a certain period of time.

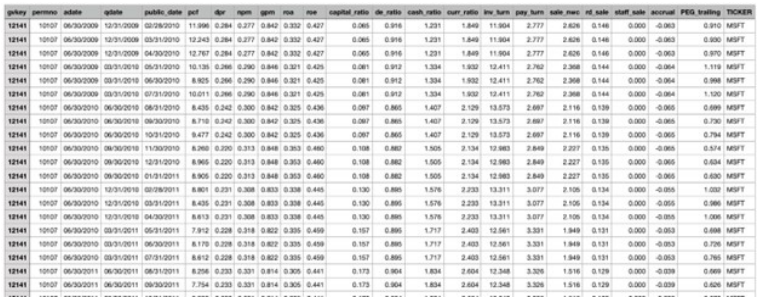
$$\%K = \left(\frac{C - L14}{H14 - L14}\right) * 100$$

, where $C = The most recent closing price$,
$L14 = The lowest price traded of the 14 previous trading sessions$,
$H14 = The highest price traded during the same 14 - day period$, and
$\%K = The current value of the stochastic indicator$

Fundamental Indicators

As previously mentioned, we encountered a challenge with the mismatch of data periods for our variables. While the stock price data are generated on a daily basis, other financial indices such as fundamental indicators and financial ratios are generated quarterly or even annually. Simply filling in missing values with quarterly or annual data would lead to misleading analysis results. To address this issue, we decided to employ feature engineering methodologies to impute the missing values in a more relevant way, leading to better analysis results.

Figure 6: The raw data for the fundamental indicators



gvkey	permno	adate	qdate	public_date	pcf	dpr	npm	gpm	roa	roe	capital_ratio	de_ratio	cash_ratio	curr_ratio	inv_turn	pay_turn	sale_nwc	rd_sale	staff_sale	accrual	PEG_trailing	TICKER
12141	10107	06/30/2009	12/31/2009	02/28/2010	11.996	0.284	0.277	0.842	0.332	0.427	0.065	0.916	1.231	1.849	11.904	2.777	2.626	0.146	0.000	-0.063	0.910	MSFT
12141	10107	06/30/2009	12/31/2009	03/31/2010	12.243	0.284	0.277	0.842	0.332	0.427	0.065	0.916	1.231	1.849	11.904	2.777	2.626	0.146	0.000	-0.063	0.930	MSFT
12141	10107	06/30/2009	12/31/2009	04/30/2010	12.767	0.284	0.277	0.842	0.332	0.427	0.065	0.916	1.231	1.849	11.904	2.777	2.626	0.146	0.000	-0.063	0.970	MSFT
12141	10107	06/30/2009	03/31/2010	05/31/2010	10.135	0.266	0.290	0.846	0.321	0.425	0.081	0.912	1.334	1.932	12.411	2.762	2.368	0.144	0.000	-0.064	1.119	MSFT
12141	10107	06/30/2009	03/31/2010	06/30/2010	8.925	0.266	0.290	0.846	0.321	0.425	0.081	0.912	1.334	1.932	12.411	2.762	2.368	0.144	0.000	-0.064	0.998	MSFT
12141	10107	06/30/2009	03/31/2010	07/31/2010	10.011	0.266	0.290	0.846	0.321	0.425	0.081	0.912	1.334	1.932	12.411	2.762	2.368	0.144	0.000	-0.064	1.120	MSFT
12141	10107	06/30/2010	06/30/2010	08/31/2010	8.435	0.242	0.300	0.842	0.325	0.436	0.087	0.865	1.407	2.129	13.573	2.697	2.116	0.139	0.000	-0.065	0.899	MSFT
12141	10107	06/30/2010	06/30/2010	09/30/2010	8.710	0.242	0.300	0.842	0.325	0.436	0.087	0.865	1.407	2.129	13.573	2.697	2.116	0.139	0.000	-0.065	0.730	MSFT
12141	10107	06/30/2010	06/30/2010	10/31/2010	8.477	0.242	0.300	0.842	0.325	0.436	0.087	0.865	1.407	2.129	13.573	2.697	2.116	0.139	0.000	-0.065	0.794	MSFT
12141	10107	06/30/2010	06/30/2010	11/30/2010	8.260	0.220	0.313	0.848	0.353	0.460	0.108	0.882	1.505	2.134	12.983	2.849	2.227	0.135	0.000	-0.065	0.574	MSFT
12141	10107	06/30/2010	09/30/2010	12/31/2010	8.965	0.220	0.313	0.848	0.353	0.460	0.108	0.882	1.505	2.134	12.983	2.849	2.227	0.135	0.000	-0.065	0.634	MSFT
12141	10107	06/30/2010	09/30/2010	01/31/2011	8.905	0.220	0.313	0.848	0.353	0.460	0.108	0.882	1.505	2.134	12.983	2.849	2.227	0.135	0.000	-0.065	0.630	MSFT
12141	10107	06/30/2010	12/31/2010	02/28/2011	8.801	0.231	0.308	0.833	0.338	0.445	0.130	0.895	1.576	2.233	13.311	3.077	2.105	0.134	0.000	-0.055	1.032	MSFT
12141	10107	06/30/2010	12/31/2010	03/31/2011	8.435	0.231	0.308	0.833	0.338	0.445	0.130	0.895	1.576	2.233	13.311	3.077	2.105	0.134	0.000	-0.055	0.986	MSFT
12141	10107	06/30/2010	12/31/2010	04/30/2011	8.613	0.231	0.308	0.833	0.338	0.445	0.130	0.895	1.576	2.233	13.311	3.077	2.105	0.134	0.000	-0.055	1.006	MSFT
12141	10107	06/30/2010	03/31/2011	05/31/2011	7.912	0.228	0.318	0.822	0.335	0.459	0.157	0.895	1.717	2.403	12.561	3.331	1.949	0.131	0.000	-0.053	0.698	MSFT
12141	10107	06/30/2010	03/31/2011	06/30/2011	8.170	0.228	0.318	0.822	0.335	0.459	0.157	0.895	1.717	2.403	12.561	3.331	1.949	0.131	0.000	-0.053	0.726	MSFT
12141	10107	06/30/2010	03/31/2011	07/31/2011	8.612	0.228	0.318	0.822	0.335	0.459	0.157	0.895	1.717	2.403	12.561	3.331	1.949	0.131	0.000	-0.053	0.765	MSFT
12141	10107	06/30/2011	06/30/2011	08/31/2011	8.256	0.233	0.331	0.814	0.305	0.441	0.173	0.904	1.834	2.604	12.348	3.326	1.516	0.129	0.000	-0.039	0.669	MSFT
12141	10107	06/30/2011	06/30/2011	09/30/2011	7.754	0.233	0.331	0.814	0.305	0.441	0.173	0.904	1.834	2.604	12.348	3.326	1.516	0.129	0.000	-0.039	0.626	MSFT

Initially, we tried forward filling, where earlier time values are used to fill the time gap. However, we found that the data quality decreased significantly, as shown by the high correlation matrix among all the fundamental indicators at different time frequencies.

To address this issue and find the most relevant way to handle it, we decided to employ the Expectation-Maximization (EM) algorithm, which assumes an underlying normal distribution. We learned about this algorithm during our lab sessions.

> The Expectation-Maximization (EM) algorithm is an iterative method to derive the maximum likelihood estimate (MLE) in presence of missing or hidden data. Let $X = (Y, Z)$ be the complete data, with log-likelihood $\log p_c(x;\theta) = l_c(x;\theta)$, where θ is the parameter of interest. It is common that if full data X is observed, the MLE is easy to derive, but if only Y is observed and Z is missing, the observed log-likelihood $l(y;\theta) = \log \int p_c(y,z;\theta)dz$ can be complex and hard to optimize.

We utilized the "impyute" package to impute the missing values between the time gap and checked the correlation matrix for improvement. As shown in the Figure, the correlation differences among predictors decreased significantly. We observed that a lower correlation led to better results, so we decided to apply the EM algorithm for feature engineering for the fundamental indicators.

Macroeconomic Variables Feature Engineering

Financial analysts often use macroeconomic data to help predict stock prices. However, macroeconomic data is typically generated on a monthly, quarterly, or even annual basis, which can create a mismatch with daily or weekly stock price predictions. To bridge this gap, we utilized the seasonal trend decomposition (STL) method, which breaks down the data into seasonal, trend, and residual components. We then applied Fourier transformation to fit the seasonal component and interpolated it with simulations. Finally, we combined the interpolated seasonal data with the backfilled trend, adding the residuals to fill any gaps. This process helped us to better predict daily and weekly stock prices using macroeconomic data despite the seasonality of the data.

$$y_i = s_i + t_i + r_i$$

, where $y_i = The value of the time series at point i$,
$s_i = The value of the seasonal component at point i$,
$t_i = The value of the trend component at point i$,
$r_i = The value of the remainder component at point i$

New York Times Data

To generate sentiment scores for the ten stocks we are interested in, we combined the headline and lead paragraph text of each article and filtered them using company-specific keywords. However, since the coverage varied widely across companies, we decided not to produce company-specific NLP features. Specifically, we found that the companies had 94, 85, 2, 98, 45, 191, 23, 7, 18, and 39 articles of coverage, respectively.

Figure 9: The list of keywords for each company

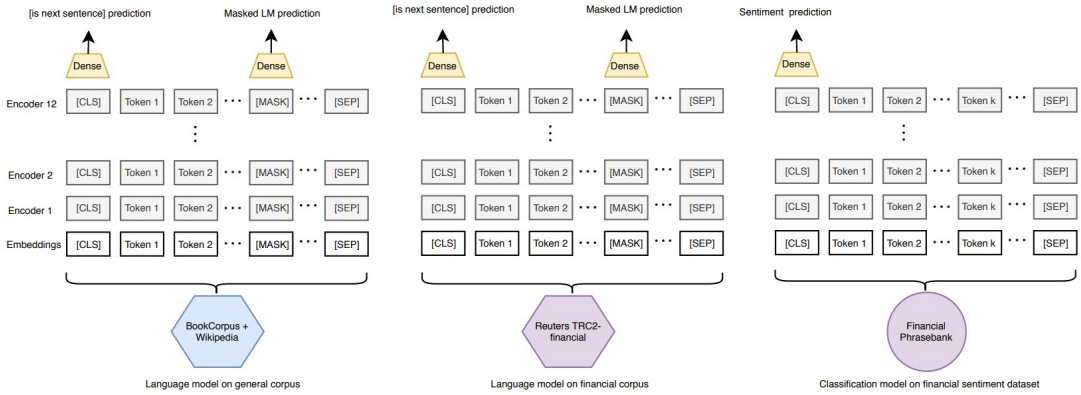
Stock	**Keywords**
AAPL	'AAPL', 'Apple', 'Tim Cook', 'iPhone'
AMZN	'AMZN', 'Amazon', 'Jeff Bezos'
BRK-B	'BRK-B', 'Berkshire Hathaway', 'Warren Buffett'
GOOG	'GOOG', 'Google', 'Sundar Pichai', 'Alphabet'
JNJ	'JNJ', 'Johnson & Johnson', 'harmac'
META	'META', 'Meta', 'Facebook', 'Mark Zuckerberg'
MSFT	'MSFT', 'Microsoft', 'Windows'
NVDA	'NVDA', 'NVIDIA', 'Jensen Huang', 'GPU'
TSLA	'TSLA', 'Tesla', 'Elon Musk'
V	'Visa', 'redit card'

To analyze the sentiment of the financial text, we applied the FinBERT model, which is a pre-trained NLP model designed for financial sentiment analysis. Financial

sentiment analysis is particularly challenging due to the specialized terminologies and lack of labeled data in the domain (Araci, 2019). FinBERT uses a combination of techniques, including LSTM, ELMo, ULMFit, Transformer, and BERT, to analyze the sentiment in financial text and calculate daily sentiment scores for each trading date.

The pre-training process is a crucial element of this model, as further pre-training on a target domain corpus has been shown to improve classification performance (Howard and Ruder, 2018). The pre-training process is depicted in Figure X.

Figure 10: Overview of pre-training, further pre-training and classification fine-tuning



To capture additional sentiment information, we included sentiment scores from the previous three days (lag1, lag2, and lag3) to represent daily background sentiments and sentiment scores from the previous five days (lag1, lag2, lag3, lag4, and lag5) to indicate weekly trading sentiments. The sentiment scores are available as separate dataframes in our GitHub repository.

To compute the sentiment scores for a day, we computed the sentiment logits for all articles from that day and averaged the logits to produce the three sentiment scores. We used logits instead of softmax probability to allow the model to convey the magnitude and confidence of the financial sentiments of a given day, which allows different days to have different weights.

Initially, we considered using monthly sentiment scores to inform our predictions. However, we ultimately decided against it due to concerns about overfitting. Furthermore, the noise and difficulty in consolidating article text data from a month would likely lead to biased and potentially meaningless results.

3 Model Fitting and Preliminary Results

3.1 Model Methodology

For our analysis, we attempted to fit both classification and regression models using the initial dependent variables Direction1 (t-1), and Direction2 (t+13 – t-1). This involved using two difference variables. Also for regression models, we used the stock price of (t) period as a target variable.

To prevent the use of future events as predictors and past events as target variables, we utilized Time Series Cross Validation instead of the general version which might shuffle the time sequence randomly. By using the Time Series Cross Validation, we ensured that we always kept the past information to predict future events. An example code for this is shown below:

- We trained data from future events (t+1, t+2, t+3) to predict the current event (t).

Figure 11: Time Series CV Code Example

```
from sklearn.model_selection import TimeSeriesSplit

X = np.array([[1, 2], [3, 4], [1, 2], [3, 4], [1, 2], [3, 4]])
y = np.array([1, 2, 3, 4, 5, 6])
tscv = TimeSeriesSplit(n_splits=3)
print(tscv)
TimeSeriesSplit(n_splits=3)
for train, test in tscv.split(X):
    print("%s %s" % (train, test))
[0 1 2] [3]
[0 1 2 3] [4]
[0 1 2 3 4] [5]
```

We also incorporated the forward selection process for feature selection in this model. This allowed the function to decide the optimal subset for each stock under each model. After this process, we investigated the feature selection statistics.

To select the best subset of features, we used the weighted F1-score as the scoring method. This approach considers the F1 score under each class, ensuring that the evaluation method adds penalties to the final classification result if the F1 score is not balanced, leading to decreased performance.

We divided the entire dataset into train and test datasets using the function shown below:

Figure 12: Train-Split Code Example

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,shuffle=False)
```

Next, we applied the time series cross-validation method to X_train and y_train to select the best feature subsets. We then evaluated the model using X_test and y_test to reduce the overfitting problem.

Finally, we applied three models for classification (SVM, Random Forest, and Logistic Regression) and another three models for regression (Lasso, Riedge, and Lasso PCA). For each model, we utilized the forward feature selection method and Time Series Cross Validation as mentioned earlier.

3.2 Logistic Regression Results

For the logistic regression analysis, we found that using Direction 2 as the dependent variable yields better overall results. However, the F1 score is generally around 0.4, with

monthly performance being better, reaching nearly 0.8. These findings are consistent with the results obtained from SVM analysis, which we will discuss later. The tables below show the results obtained using daily frequency data.

Figure 13: The result of Direction 1

f1-score	AAPL	AMZN	BRK-B	GOOG	JNJ	META	MSFT	NVDA	TSLA	V
0	0.03	0.65	0.67	0.01	0.67	0.66	0.67	0.02	0.65	0.64
1	0.71	0.01	0.01	0.65	0.05	0.05	0.03	0.66	0.19	0.02
accuracy	0.55	0.49	0.51	0.49	0.51	0.5	0.51	0.49	0.51	0.48
macro avg	0.37	0.33	0.34	0.33	0.36	0.35	0.35	0.34	0.42	0.33
weighted avg	0.4	0.32	0.34	0.32	0.36	0.36	0.35	0.33	0.42	0.32

Figure 14: The result of Direction 2

F1 Score	AAPL	AMZN	BRK-B	GOOG	JNJ	META	MSFT	NVDA	TSLA	V
0	**0**	**0.51**	**0**	**0.43**	**0**	**0.6**	**0.44**	**0**	**0**	**0**
1	**0.8**	**0.05**	**0.74**	**0.32**	**0.75**	**0.18**	**0.05**	**0.77**	**0.65**	**0.84**
accuracy	0.66	**0.35**	**0.59**	**0.38**	**0.59**	**0.46**	**0.3**	**0.62**	**0.48**	**0.72**
macro avg	0.4	**0.28**	**0.37**	**0.38**	**0.37**	**0.39**	**0.24**	**0.38**	**0.32**	**0.42**
weighted avg	0.53	**0.22**	**0.44**	**0.37**	**0.44**	**0.36**	**0.16**	**0.47**	**0.31**	**0.61**

3.3 Random Forest Results

The results of our Random Forest analysis showed very poor prediction performance overall. We tried both the normal Random Forest and Xgboost models, but neither provided us with any meaningful results. As a result, we concluded that Random Forest is not an appropriate model for our data.

Figure 15: The Random Froest results

	Direction 1		Direction 2	
F1 Score	RF	XGBOOST	RF	XGBOOST
0	0.56	0.56	0.27	0.25
1	0.38	0.41	0.76	0.74
accuracy	0.48	0.50	0.64	0.62
macro avg	0.47	0.49	0.52	0.50
weighted avg	0.46	0.48	0.60	0.59

3.4 SVM

SVM Results

Our SVM analysis indicated that Direction 2 produced better overall results. We experimented with several kernels, including linear, polynomial, RBF, and Sigmoid, and found that RBF and polynomial kernels performed the best. Therefore, we decided to use the RBF kernel for future tuning to simplify the model.

Figure 16: The SVM results

	Difference 1				Difference 2			
	Linear Kernel	Polynomial Kernel	RBF Kernel	Sigmoid Kernel	Linear Kernel	Polynomial Kernel	RBF Kernel	Sigmoid Kernel
0	0.03	0.7	0.36	0.46	0.46	0.02	0.21	0.38
1	0.71	0.5	0.73	0.45	0.12	0.81	0.84	0.44
accuracy	0.55	0.62	0.62	0.45	0.33	0.68	0.73	0.41
macro avg	0.37	0.6	0.54	0.45	0.29	0.42	0.52	0.41
weighted avg	0.4	0.59	0.56	0.45	0.23	0.55	0.64	0.42

SVM Frequency Analysis

Our analysis also revealed that classification performance improves with increasing frequency, from Daily to Monthly. We obtained the best prediction results with monthly frequency data.

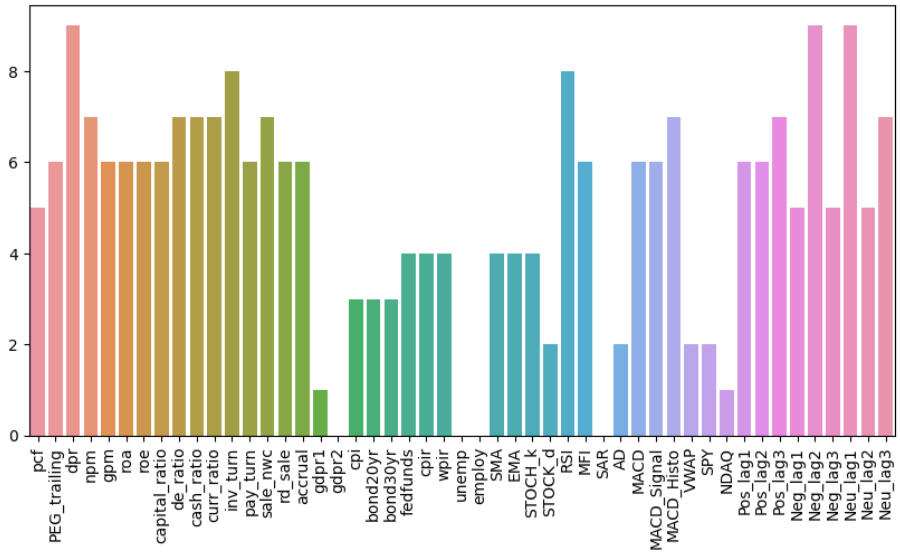
Figure 17: The SVM Frequency Analysis Results

	Daily	Weekly	Monthly
AAPL	0.799	0.781	0.829
AMZN	0.763	0.791	0.800
BRK-B	0.728	0.716	0.667
GOOG	0.689	0.545	0.802
JNJ	0.673	0.668	0.703
META	0.692	0.678	0.667
MSFT	0.717	0.687	0.857
NVDA	0.748	0.738	0.769
TSLA	0.493	0.625	0.509
V	0.828	0.844	0.829

SVM Feature Selection

Regarding feature selection, we found that fundamental indicators had the most significant impact on the model. These indicators supported the prediction tasks and improved the accuracy of our predictions.

Figure 18: The SVM Feature Selection



3.5 Lasso

In contrast to the classification models, we used regression models to predict continuous stock prices. One technique we employed for feature selection and regularization was Lasso regression, short for Least Absolute Shrinkage and Selection Operator. Given the large number of predictors, we suspected that using an optimal subset of them would lead to better prediction performance.

Our initial analysis showed that the best subset consisted of just 5 features. As depicted in the Lasso Regression Results graph, the model performed well for most stocks, except for Apple and Visa.

In terms of feature selection, we found that technical variables, such as the Stochastic Oscillator, contributed most to this model.

Figure 19: The Lasso Regression Results

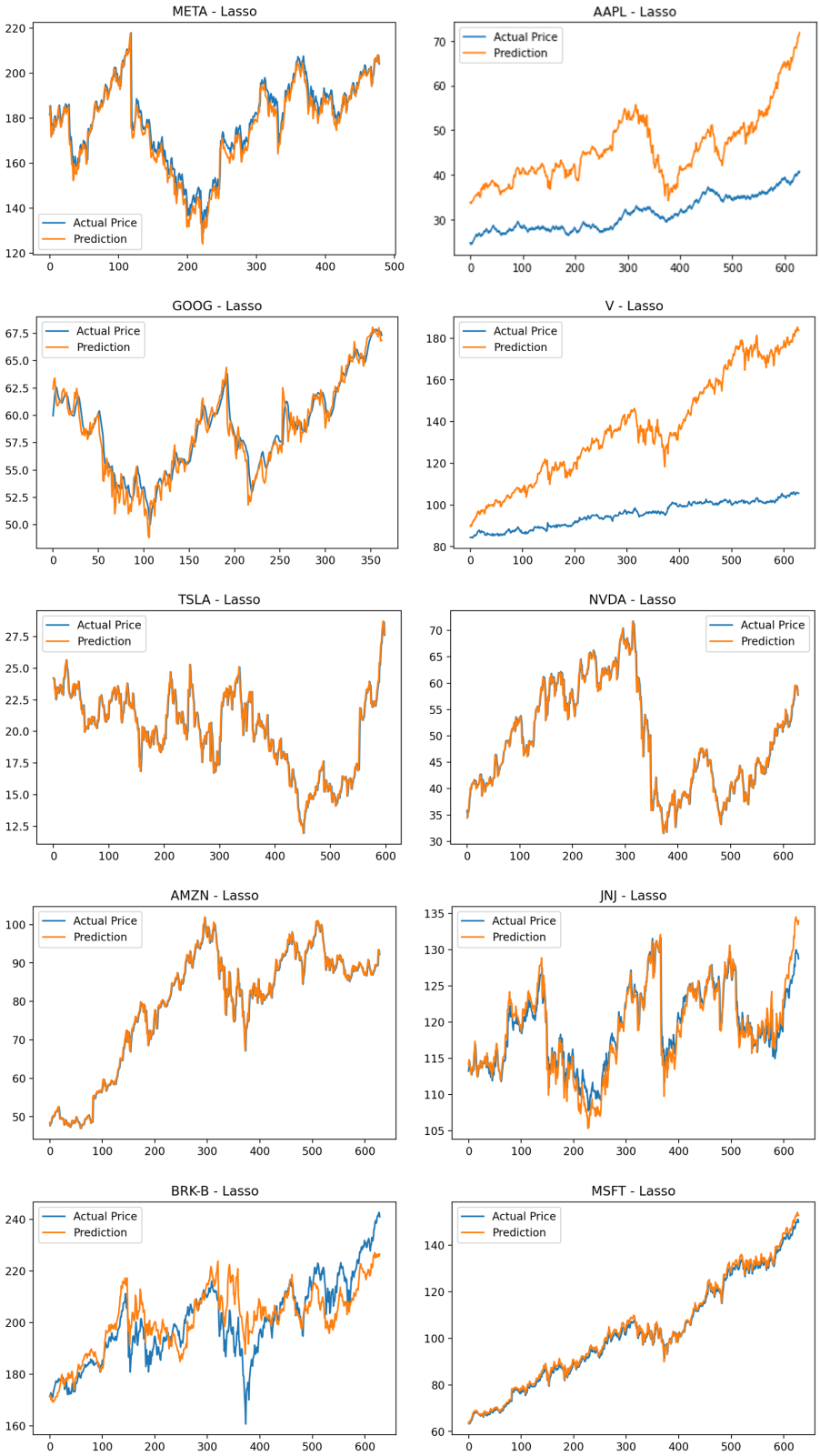
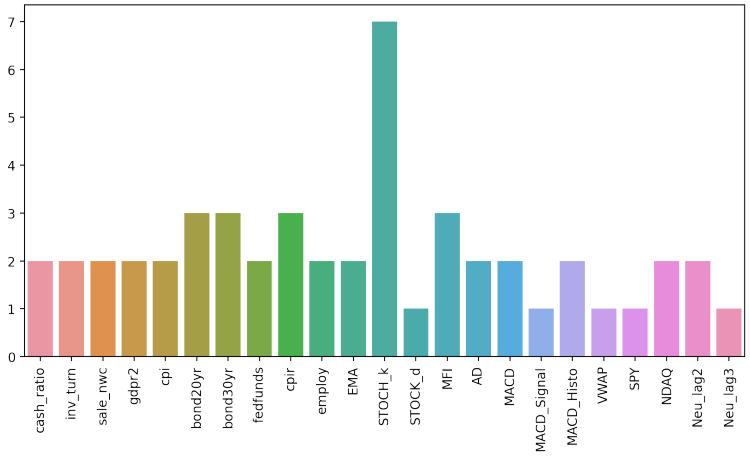


Figure 20: The Lasso Feature Selection



3.6 Ridge

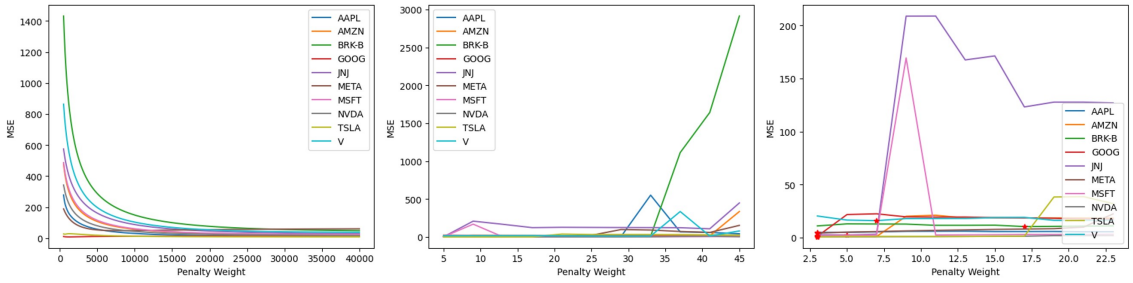
Variable Selection

Due to the high penalty from our initial analysis, we suspect that the dataset contains many irrelevant features. To improve the model's performance, we aim to use feature selection techniques to identify the most informative features.

Upon analyzing the MSE values, we observed that using a smaller number of features (less than 25) leads to lower MSE. Therefore, we plan to re-tune the feature selection step to obtain the best subset of features for the training dataset.

Overall, we found that seven features are optimal for most stocks, with the exception of BRK-B, which requires 17 features for optimal performance.

Figure 21: Ridge variable selection

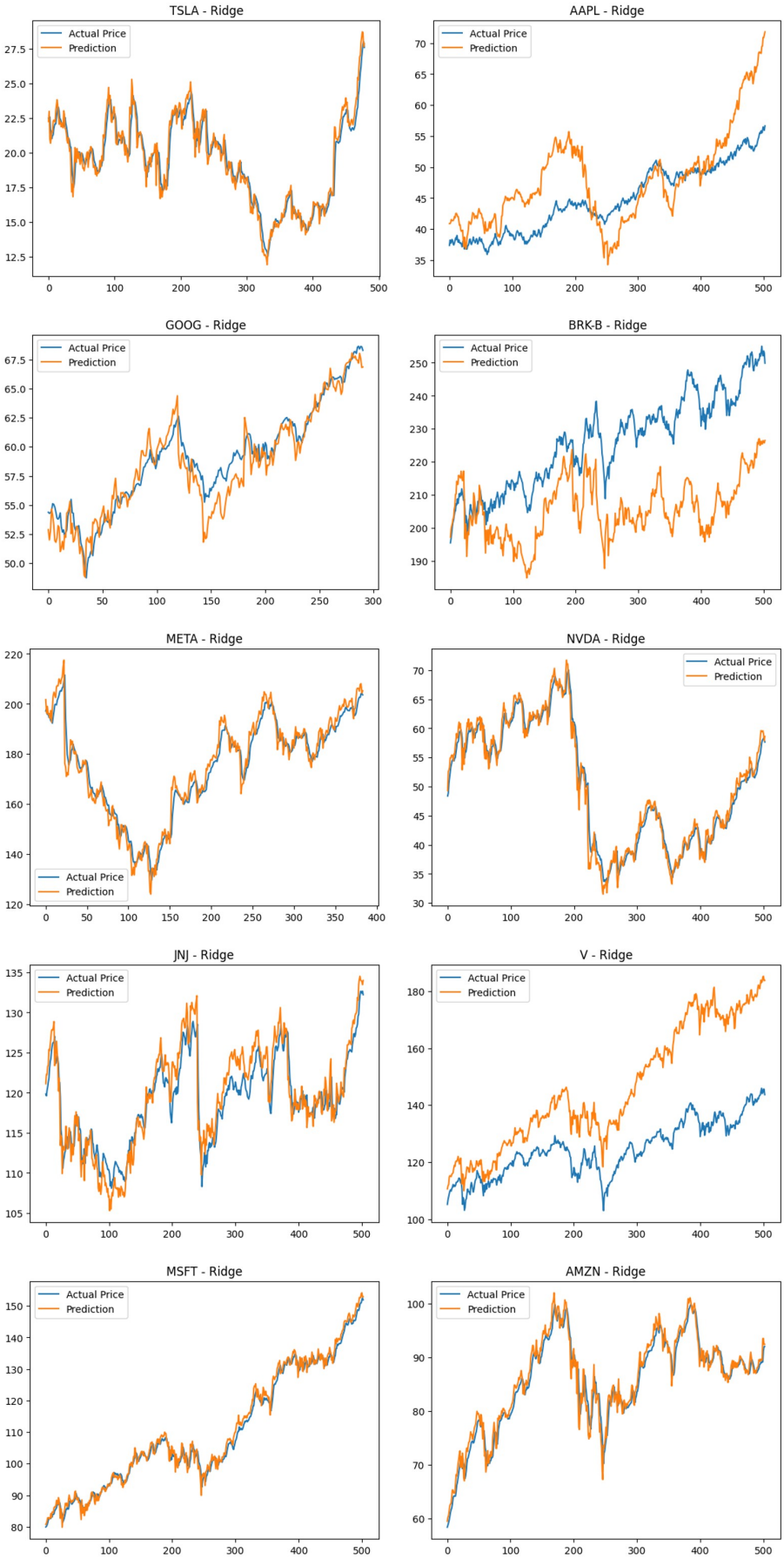


Ridge Result

Based on the analysis using the Ridge model, we have reached the following conclusions:

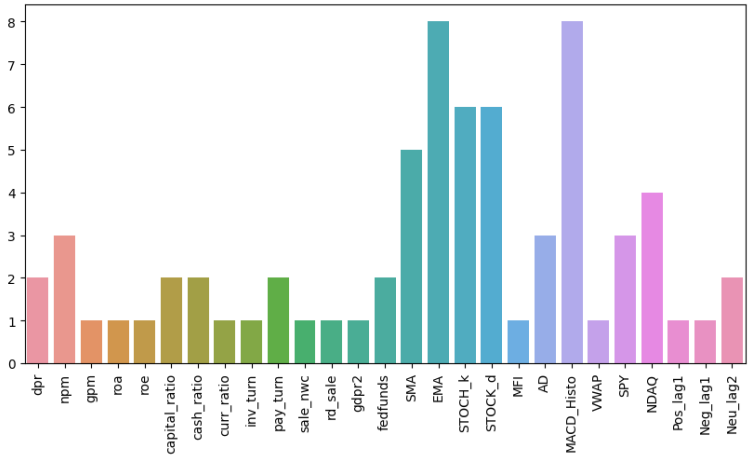
First, with the exception of Apple, Berkshire Hathaway, and Visa, the model performed very well in predicting the stock prices for all other items, as indicated by the very low MSE values.

Figure 22: Ridge Results Graph



Second, in terms of frequency, the daily data resulted in the best predictions, followed by weekly and monthly data. As the frequency increased, the prediction accuracy decreased, likely due to the inability of the model to capture the larger changes between data points.

Figure 23: Ridge Feature Selection



Third, the technical indicators were found to be the strongest predictors of stock prices. Almost all technical indicators showed a significant level of importance in the model. Additionally, several fundamental indicators such as dpr and npm, as well as macroeconomic indicators like the federal funds rate, were found to have a strong influence on the model.

3.7 Lasso with PCA

PCA Results

Lasso, an acronym for Least Absolute Shrinkage and Selection Operator, is a linear regression technique used for feature selection and regularization. Meanwhile, Principal Component Analysis (PCA) is a popular technique in machine learning and data science for dimensionality reduction. Since stock data is a time series, it may contain a lot of noise that can adversely affect model performance. Therefore, we use PCA to reduce the dimension of the data and Lasso for feature selection.

Figure 24: PCA Cumulative Explained Variance

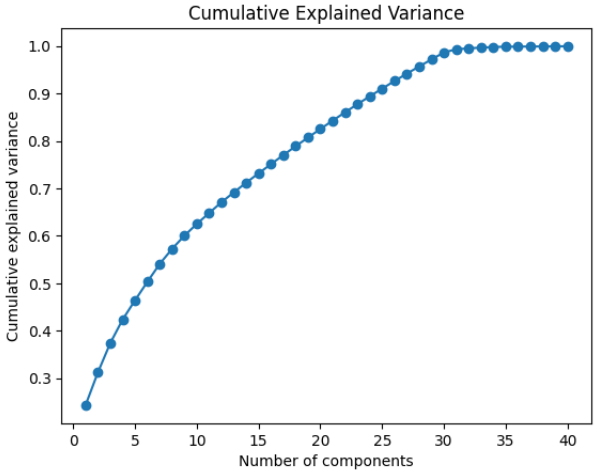


Figure 25: PCA Variables

Out[5]:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC31	PC32	PC33	PC34	PC
0	-4.676183	0.717328	-3.916821	-0.865202	0.423330	2.126407	2.989734	0.572820	0.663627	0.446939	...	-0.638703	0.319669	-0.314367	0.063588	-0.157
1	-4.793984	0.903469	-3.945507	-1.064112	-1.683572	1.632482	-2.088154	0.717673	0.613422	0.464931	...	-0.632267	0.343135	-0.313338	0.029663	-0.141
2	-5.020662	1.230514	-3.873630	-1.392689	1.928160	-1.131456	-2.776604	-0.295735	0.426009	-0.199288	...	-0.746628	0.303498	-0.309260	0.007566	-0.150
3	-5.036326	1.347989	-3.367817	-1.713100	-0.875235	-2.079474	0.093144	-0.026547	0.181758	-0.030956	...	-0.686063	0.157077	-0.305340	0.026791	-0.198
4	-4.306113	0.691473	-3.390819	-1.370803	-0.895087	-2.646459	1.216676	2.506918	0.343153	-0.312097	...	-0.572702	0.048366	-0.238183	0.002306	-0.211
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2511	6.949113	-4.233779	0.953912	-1.065472	1.639143	-0.873742	0.991530	1.205190	1.166008	0.212502	...	1.150519	-0.099034	-0.081145	0.415315	-0.322
2512	7.268151	-4.805499	0.901674	-0.768415	-0.341862	0.862394	-0.731504	0.779333	1.227972	2.598307	...	1.284844	-0.181124	-0.067615	0.372019	-0.309
2513	6.699623	-4.215744	1.135155	-1.275359	1.356284	0.400320	-0.853101	-0.796311	1.402859	2.157178	...	1.140486	0.033756	-0.091053	0.401800	-0.375
2514	7.205620	-4.765040	1.129899	-1.470075	-1.346731	0.928555	1.203885	0.978552	1.669729	1.751471	...	1.230614	-0.557988	0.004180	0.466260	-0.373
2515	8.602598	-6.005073	0.776634	-0.362563	0.768441	-0.499811	0.233304	2.604945	2.498059	8.489611	...	1.561908	-0.566440	0.032455	0.488093	-0.381

2516 rows × 40 columns

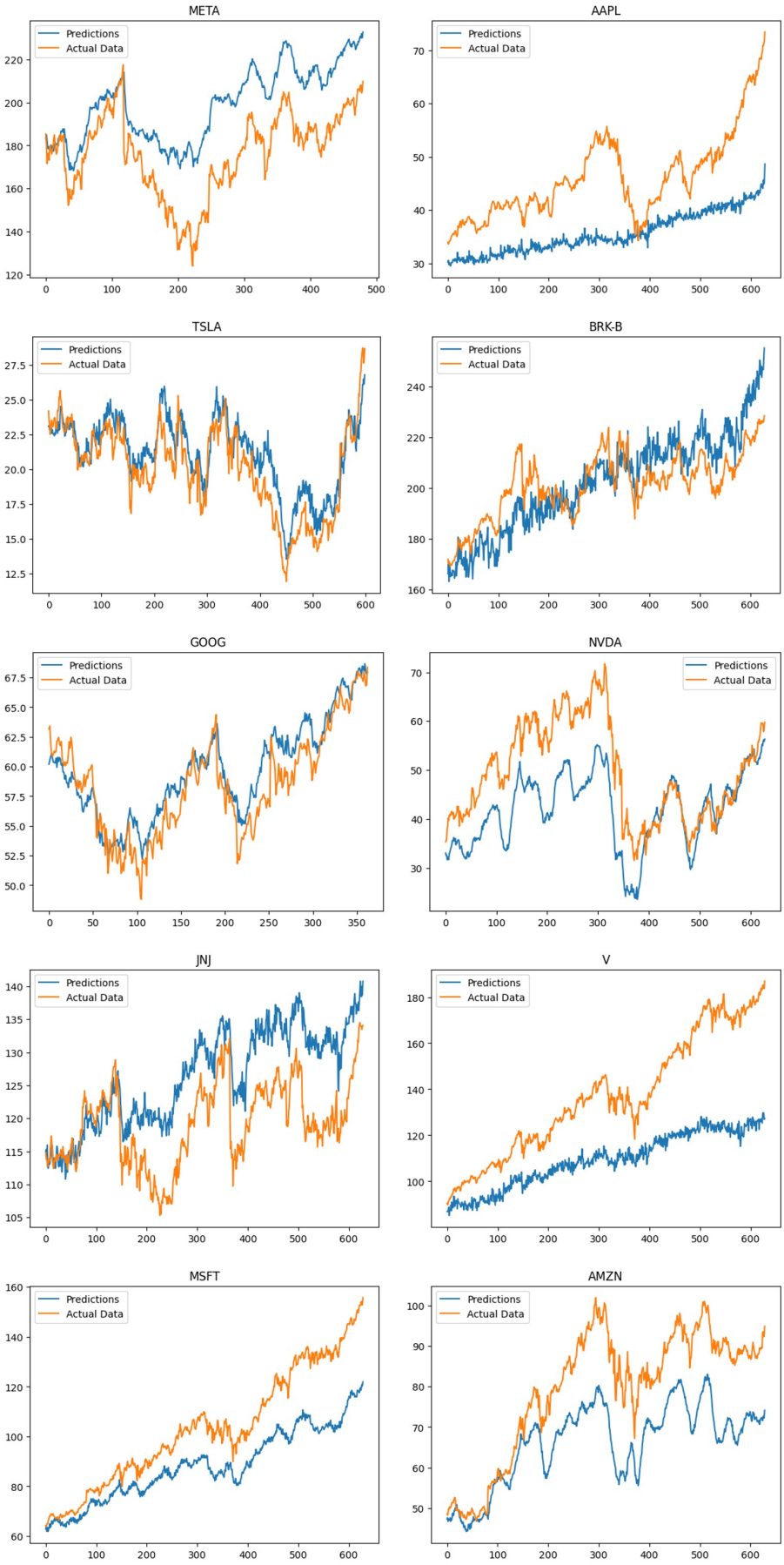
To remove as much noise as possible while maintaining interpretability, we use the cumulative explained variance (CEV) plot to determine the optimal number of components, which is 30. This number of components explains 99.9% of the data while eliminating most of the noise. Additionally, since some data is missing, we use KNN to impute the missing values before conducting PCA.

As the range of the training data is unknown, we use cross-validation to determine the best n_splits for each stock to achieve optimal performance.

Lasso PCA Results

Based on the analysis, the Lasso PCA model performed poorly in predicting stock prices compared to the previous Lasso and Ridge models. The MSE values for all stocks were relatively high. Therefore, we conclude that PCA was not the optimal dimensionality reduction approach for our data. Thus, for the most accurate predictions of stock prices, we recommend not considering the Lasso PCA model.

Figure 26: Lasso PCA Results



4 Limitation and Next Step

Based on our findings, we have determined that SVM with RBF is the best model for predicting simple direction, while Lasso performs the best for predicting exact stock prices. Both models are able to provide us with important features and insights, particularly Lasso which allows us to examine the associations between predictors and stock prices.

However, our analysis has several limitations. Firstly, due to subscription fees and costs, we were unable to include a broader scope of social listening data. In the future, we would like to expand our data scope to include broader news media and social media platforms such as Twitter, Facebook, Instagram, and TikTok, as well as specialized online trading communities.

Secondly, we acknowledge that the New York Times (NYT) may not be the most optimal source for financial information since it covers a wide range of topics. Despite this, we filtered articles based on specific financial keywords and used the FinBERT model to give more weight to financially-oriented articles. We encourage further research into alternative data sources that may provide more focused financial content.

Lastly, combining data sets with different generating periods proved to be a significant challenge. Although we used EM algorithms and Seasonal Trend Decomposition (STD) methodologies to address this issue, we observed that monthly frequency data showed the best performance, indicating that our imputation methods may not be successful for predicting daily frequency data. As such, we would like to explore more optimal methodologies to impute our data in the future.

Overall, our project aimed to find a way to predict stock market prices, which has been likened to the mythical "Fountain of Youth" by Eric. We faced numerous challenges, but we overcame them by employing a combination of statistical, computer science, and financial techniques. In the end, we were able to successfully merge disparate and diverse data sets with varying frequencies of data generation, and applied a range of models, including both classification and regression, to predict the future direction and exact prices of stocks. Our project underscores the importance of interdisciplinary collaboration and innovative problem-solving, and we hope our findings will contribute to future advancements in the field of stock market prediction.

Figure 7: The correlation matrices for the forward filled

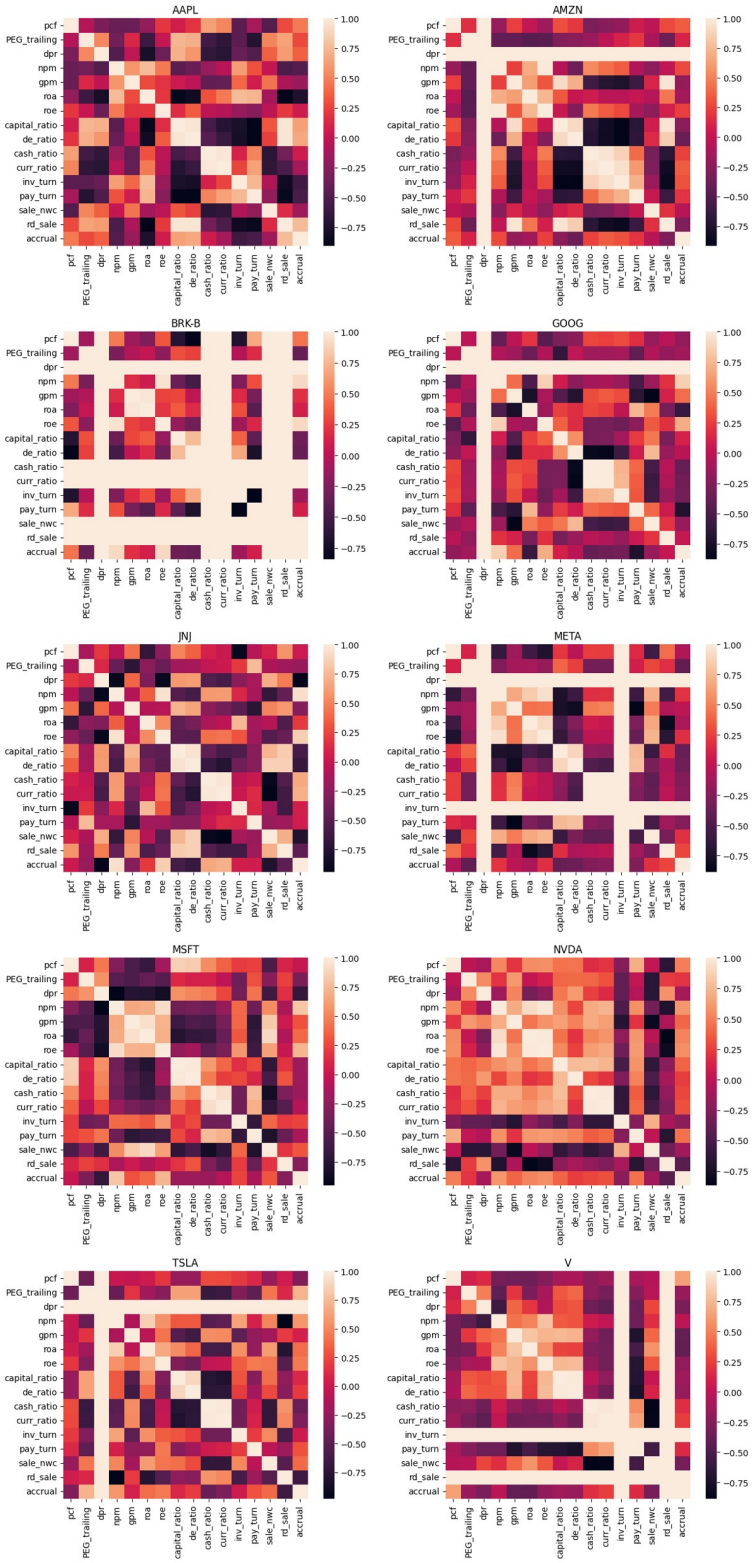
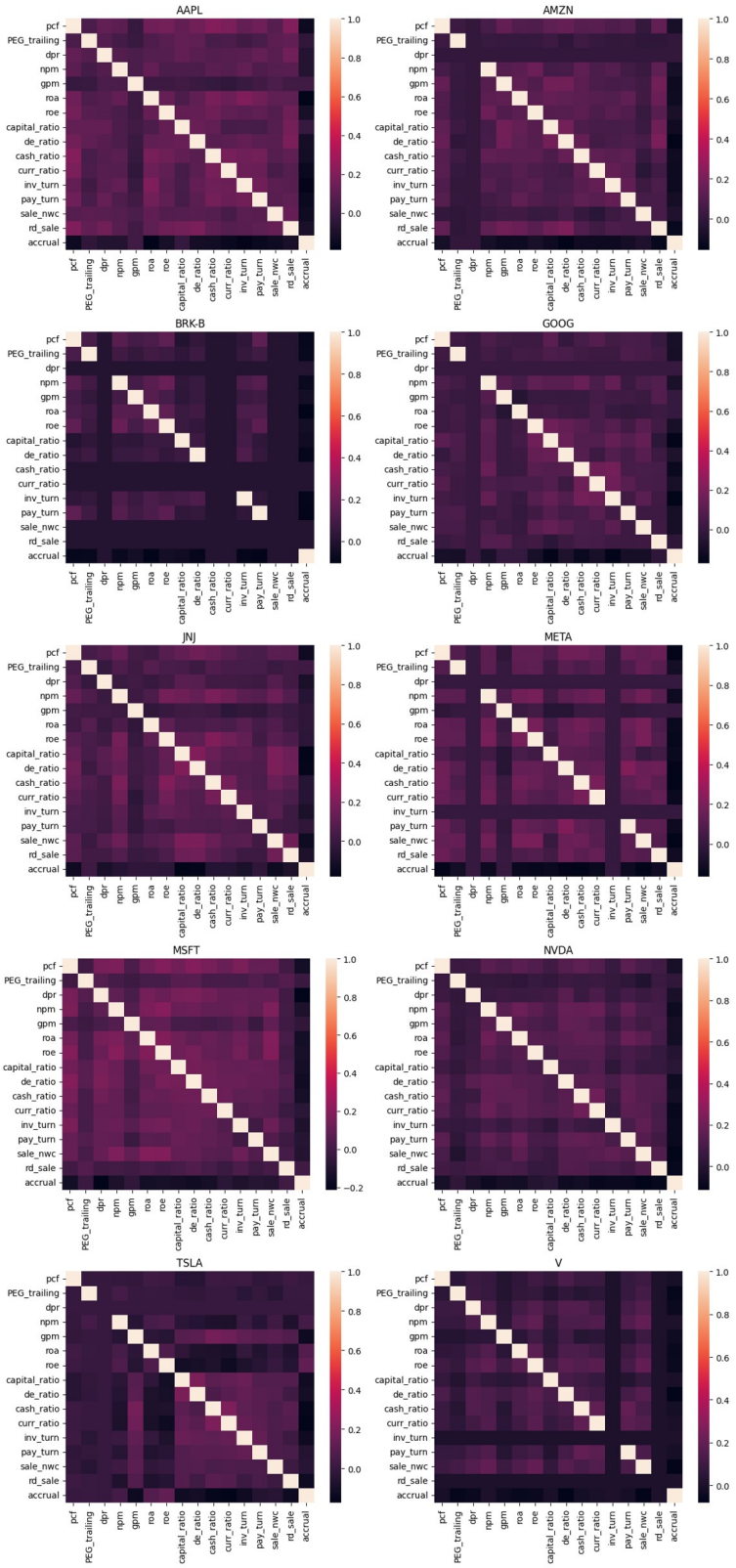


Figure 8: The correlation matrices for the EM algorithm



5 Reference

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Beaver, W. H. (1968). Market prices, financial ratios, and the prediction of failure. Journal of accounting research, 179-192.

Howard, J., Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Korivi, N., Naveen, K. S., Keerthi, G. C., Manikandan, V. M. (2022, January). A Novel Stock Price Prediction Scheme from Twitter Data by using Weighted Sentiment Analysis. In 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence) (pp. 623-628). IEEE.

Lewellen, J. (2004). Predicting returns with financial ratios. Journal of Financial Economics, 74(2), 209-235.

Li, X., Wu, P., Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing Management, 57(5), 102212.

Mukherjee, T. K., Naka, A. (1995). Dynamic relations between macroeconomic variables and the Japanese stock market: an application of a vector error correction model. Journal of financial Research, 18(2), 223-237.

Tinoco, M. H., Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. International review of financial analysis, 30, 394-419.

Siew, H. L., Nordin, M. J. (2012, September). Regression techniques for the prediction of stock price trend. In 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE) (pp. 1-5). IEEE.

Swathi, T., Kasiviswanath, N., Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. Applied Intelligence, 52(12), 13675-13688.