2021

Feature Investigation for Stock Returns Prediction Using XGBoost and Deep Learning Sentiment Classification

Seungho (Samuel) Lee
Claremont McKenna College

Claremont McKenna College

Feature Investigation for Stock Returns Prediction Using XGBoost and Deep Learning Sentiment Classification

Submitted to
Professor Nishant Dass
and
Professor Mike Izbicki

by
Seungho (Samuel) Lee

for
Senior Thesis
Spring 2021
May 3, 2021

This page is intentionally left blank.

Acknowledgements

I would first like to express my gratitude to my thesis advisors, Professor Nishant Dass and Professor Mike Izbicki. Without their guidance and technical support, it would have been hard for this thesis to come to a fruition.

I would also like to thank Dr. Jeho Park and Cindy Cheng from Murty Sunak Quantitative and Computing Lab for providing a much-needed guidance and access to its NVIDIA DGX Station, which greatly reduced amount of time required to conduct this thesis project.

Lastly, I would like to thank my family for providing me emotional support throughout the pandemic and Vladyslav Ivanov, a close friend who helped me immensely throughout my academic journey.

Abstract

This paper attempts to quantify predictive power of social media sentiment and financial data in stock prediction by utilizing a comprehensive set of stock-related fundamental and technical variables and social media sentiments. For conducting sentiment analysis, this study employs a pretrained finBERT model that provides three different sentiment classifications and respective softmax scores. Hence, the significance of these variables is evaluated with XGBoost regression and Shapley Additive exPlanations (SHAP) frameworks. Through investigating feature importance, this study finds that statistical properties of sentiment variables provide a stronger predictive power than a weighted sentiment score and that it is possible to quantify the impact features make on so-called "black box" models.

Keywords: Feature Importance, Machine Learning, Sentiment Analysis, Shapley Value, Stock Prediction, Twitter Sentiment

Table of Contents

I. Introduction

Predicting stock returns has been one of the contentious topics in modern Financial Economics. From his empirical work "Efficient Capital Markets," Eugene Fama found that there is extensive evidence supporting the Efficient Market Hypothesis, suggesting that prices of securities "fully reflect" relevant information at any given time: Therefore, neither technical nor fundamental analysis is effective in seeking abnormal returns consistently under an efficient market condition (Malkiel, 2003).

Such notion has been challenged a lot more in recent years as a quick adoption of the internet and advancement in computational resources led to various ways to capture unrealized information. Such information has been captured in various ways from utilizing image recognition on satellite images to implementing a natural language processing (NLP) algorithm to capture public sentiment on social media. While a web scraping practice on publicly available media had been a contentious topic from a legal standpoint, it has been ruled from the *HiQ Labs, Inc. v. LinkedIn Corporation* case that scraping information from a public website does not violate the Computer Fraud and Abuse Act (Lee, 2019). As utilizing information captured from the public domain is becoming more accepted and adopted, it is imperative to measure the significance of using such data from an Economics perspective.

While there has been extensive research in utilizing advanced machine learning algorithms to predict stock returns, there has been a lack of research that attempts to define the magnitude of contributions features make on stock predictions. Therefore, this paper utilizes nonparametric models to quantify and rank important features utilized in stock predictions. Hence, this thesis contributes to the field of Quantitative Economics and Finance by:

1. Utilizing comprehensive financial data with sentiment scores to derive key features for stock returns prediction.
2. Incorporating game theory framework to quantify feature contribution to the predictive power of a “black box” machine learning model, such as an Extreme Gradient Boosting (XGBoost) algorithm.
3. Filling gaps in understanding impacts of social media and information role on stock price movements.

II. Literature Review

While the Efficient Market Hypothesis (EMH) has been influential throughout modern Financial Economics, it has been challenged numerous times. Grossman and Stiglitz (1980) argue that abnormal returns are present if there is a cost related to obtaining such information and that those returns will disappear once the costs are properly compensated. Hence, they further contend that the perfect information is impossible as there is no economic incentive for investors to search for information if it is fully reflected in the market, rendering financial markets obsolete.

As a response, Fama (1991) concedes to their arguments in his later paper that the strict EMH only works under assumptions of no "information and trading costs" and points that a more economically sensible definition suggests that the information is reflected in the price to the point where profit that one earns from having the information is not greater than the marginal costs associated with obtaining it. In other words, if the information is too costly to obtain, no agents in the market are willing to uncover such information, which allows markets to follow Random Walk.

While the above hypothesis had been highly regarded, there have been numerous studies in recent years that challenged a notion of a "strong form" of efficiency in stock markets. Abu-Mostafa and Atiya (1996) have argued that the existence of numerous price trends and "undiscounted serial correlations among fundamental events and economic figures" are present in their findings on foreign exchange markets. Hence, Lo, Mamaysky, and Wang (2000) utilize a technical pattern recognition approach to conclude that the analysis can be beneficial in seeking excess returns.

There have been two main factions that attempted to “beat the market”: fundamental and technical analyses. Fundamental analysis involves assessing the intrinsic value of a firm via macroeconomic indicators (e.g., GDP, CPI), industry analysis, and equity-specific analysis (Hu et al., 2015). In terms of company-specific information, there are numerous methods employed, including the P/E method, the Gordon Growth Model, and financial ratios (Shah et al., 2019). Furthermore, the above methods are benchmarked across time or similar firms to identify financial health and “provide the foundation for financial forecasting” (Schill, 2016). On the other hand, technical analysis involves looking into stock data and deriving useful indicators (e.g., price momentum) that capture patterns in stock price movements (Nesbitt and Barrass, 2004).

A rise of technological adoptions and computational resources has also expanded available information sets, leading to various investment approaches to acquire abnormal returns. Due to a recent surge in the internet usage and computational resources, there has been a constant expansion of available media for market information, including social media (e.g., Facebook, Twitter, Reddit) and real-time satellite images of store-level parking lots, which have been shown to increase information asymmetry between informed and uninformed investors (Katona et al., 2018).

Especially, there have been numerous discussions in recent years that attempt to explain the role of Twitter data in stock price movements. Taking a different approach from traditional stock prediction models based on the EMH assumption, Mittal and Goel (2011) delve into a perspective of Behavioral Economics and establish a premise that there is a direct correlation between public and market sentiments. Utilizing over 476 million publicly available tweets from June 2009 to December 2009, the authors categorize the tweets into “Calm, Happy, Alert, and Kind” via compiling and mapping a Profile of Mood States (POIMS) and computing daily scores with a simple counting method. Through the 5-fold sequential cross-validation method, they found that their Self Organizing Fuzzy Neural Network (SOFNN) algorithm performed well with Calm and Happiness states, achieving 75.56% accuracy in stock directions.

Yang et al. (2015) utilize network analysis to define a financial community with Twitter users that share similar interests with the financial market to prove that Twitter sentiment provides a predictive power in predicting the market movement. They extract a list of commonly used languages from critical user nodes identified in the above and evaluate each stock ticker’s weighted sentiment score, average sentiment value, and centrality measures (i.e., Out-degree, Betweenness, Closeness). From their analysis, they find that the sentiment regression with the betweenness group exhibits statistical significance across all logarithmic returns and that selecting the top 200 users in the betweenness group is most effective.

On the other hand, Gu and Kurov (2020) consider the Fama-MacBeth and the Carhart Four-factor models to quantify the informational role of social media and define its applicability in investment strategies. Using Bloomberg’s machine learning-based firm-specific Twitter sentiment scores, they find that the sentiments can predict abnormal returns and that they capture fundamental information reflected on one-day delayed stock prices. Looking at “value-relevant” events (i.e., analyst recommendation, analyst target price changes, quarterly earnings), they also find that the sentiments provide new fundamental information about firms.

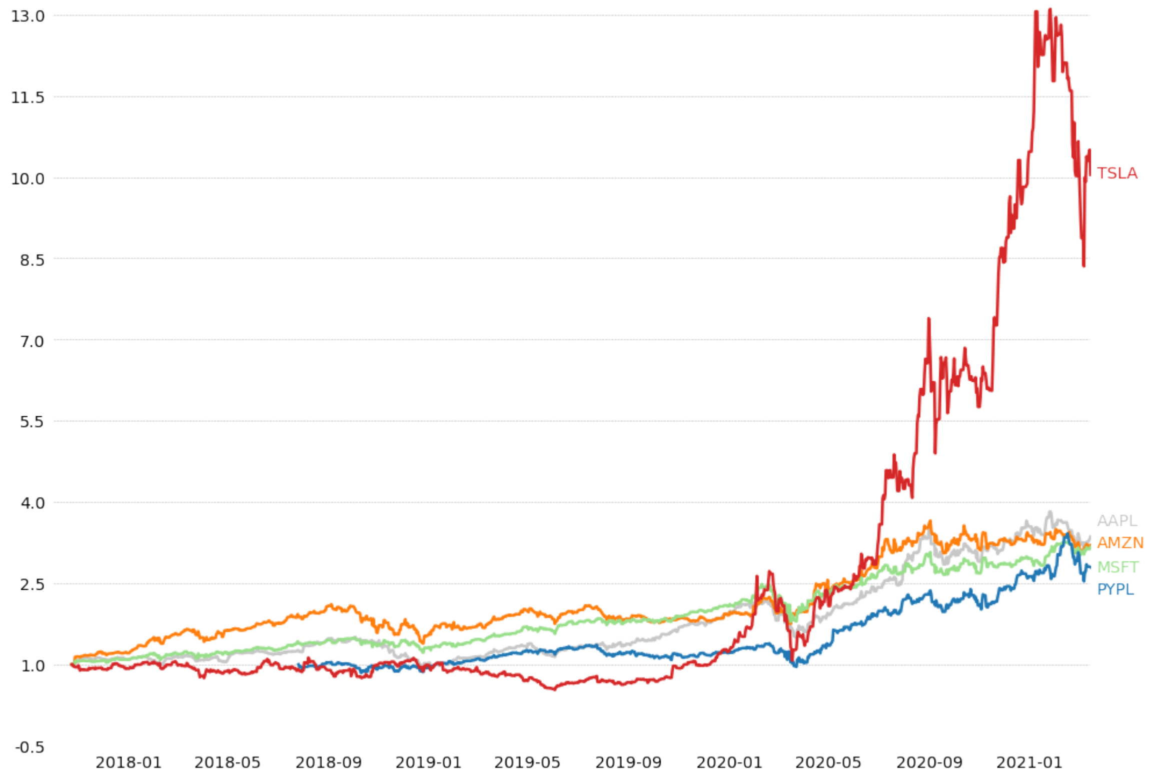
Lastly, Li et al. (2020) propose an approach that incorporates technical indicators from stock prices and sentiments of financial news articles in “building a two-layer LSTM network to learn the sequential information.” Utilizing more than five years of Hong Kong Stock Exchange data and four sentiment dictionaries, they found that the LSTM model with both technical indicators and news sentiment results in higher accuracy.

III. Data and Methodologies

1. Dataset

This thesis covers a total of five tickers with dates ranging from October 23, 2017 to March 16, 2021: Apple (Ticker: $AAPL), Amazon (Ticker: $AMZN), Microsoft (Ticker: $MSFT), PayPal (Ticker: $PYPL), and Tesla (Ticker: $TSLA). While longer time-series data on historical prices were available, a timeframe of the data is restricted largely due to two factors: Twitter Dataset and API Limitation. As the final Twitter dataset only contains tweets generated on and after October 23, 2017, the scope of this project has been restricted to the above timeline.

Figure 1: *Normalized Returns of Tickers (2017 – 2021)*



Data utilized in this study can be divided into five major segments: Historical Prices, Technical Analysis, Fundamental Analysis, Market Proxies, and Sentiment Analysis. Historical prices (daily adjusted time-series) and technical analyses were collected via Alpha Vantage. Backed by well-known exchanges and institutions, such as London Stock Exchange, Y Combinator, and Harvard Business School, Alpha Vantage provides time-series data on stock prices, company-specific fundamental data, forex, and an array of technical indicators. Daily adjusted time-series data from Alpha Vantage consist of *trading date*, *open*, *high*, *low*, *close*, *adjusted close*, *volume*, and *dividend amount* over 20 years of data. While there are various technical indicators available, the following variables were selected for this study (Alpha Vantage, 2021):

1) Simple Moving Average (*SMA*): It refers to an average price of a given stock over a certain period *t*. It is defined as

$$SMA_t = \frac{P_1 + P_2 + P_3 + \cdots + P_t}{t} \tag{1}$$

where *t* refers to time interval, and *P* refers to stock price at *t*. Hence, SMA is calculated with daily closed prices and on 20 days period.

2) *Exponential Moving Average* (*EMA*): It is a moving average metric that reduces the lag in price movement by providing more weights to recent prices. By default, a period is set as 20.
3) *Moving Average Convergence/Divergence* (*MACD*): Invented by Gerald Appel, it refers to the difference in the EMA values between 12- and 26-day and is generally utilized as a "trend or momentum indicator" (Schlossberg, 2021). Hence, the API provides the MACD histograms (spreads between MACD and 9-day EMA), MACD values, and the MACD Signal Lines on closed prices.
4) *Stochastic Oscillator* (*STOCH*): It provides the proximity of the current close price to the high-low range over a given period. The API retrieves SlowK and

SlowD metrics. Hence, the SlowK period is calculated with a 5-day period and SlowD with a 3-day period.

5) *Relative Strength Index* (*RSI*): As a momentum indicator, it gauges “the magnitude of recent price changes” to signal overbought or oversold situations (Blystone, 2021). By default, it is calculated over a 20-day horizon on closed prices.

Company-specific items from Alpha Vantage API’s Fundamental Analysis are utilized to derive the following fundamental ratios and measures:

1) Profitability Ratios: *Profit Margin*, *Asset Turnover*, *Return on Assets (ROA)*, *Return on Equity (ROE)*, *Return on Invested Capital (ROIC)*, *Return on Research Capital (RORC)*, and *Working Capital Turnover* are selected for measuring business profitability.
2) Operating Efficiency Ratios: *EBITDA Margin*, Gross *Margin*, *Operating Margin*, *Inventory Turnover*, *Property, Plant, and Equipment (PPE) Turnover*, *Receivables Turnover*, and *Payables Turnover* are selected.
3) Liquidity Ratios: *Current Ratios*, *Quick Ratios*, and *Operating Cash Flow Ratio* are chosen.
4) Leverage: *Debt Ratio*, *Debt-to-Equity Ratio*, and *Financial Leverage Ratio* are selected for analyzing the financial leverage of the firm.
5) Valuation: *P/E Ratio*, *Market Capitalization*, *Enterprise Values* are selected to gauge the firm’s value.

Formulas utilized to calculate the above variables can be found in Figure A1.

For Market Proxies, data were collected from various institutions and news sources, such as Nasdaq, U.S. Department of Treasury, Wall Street Journal (FactSet), and Yahoo Finance. Obtained from Yahoo Finance, daily quotes of Nasdaq Composite (IXIC) and CBOE Volatility Index (VIX) consist of *Open*, *High*, *Low*, *Close*, *Adjusted Close* prices, and *Volume*. As S&P 500 and Dow Jones Industrial Average (DJI) were not publicly available on Yahoo Finance, they were obtained from the Nasdaq website and FactSet database via Wall Street Journal, respectively. Lastly, Daily Treasury Yield Curve Rates were obtained from the U.S. Treasury Department's Resource Center page, which spans from 1-Month Treasury Bill to 30-Year Treasury Bond.

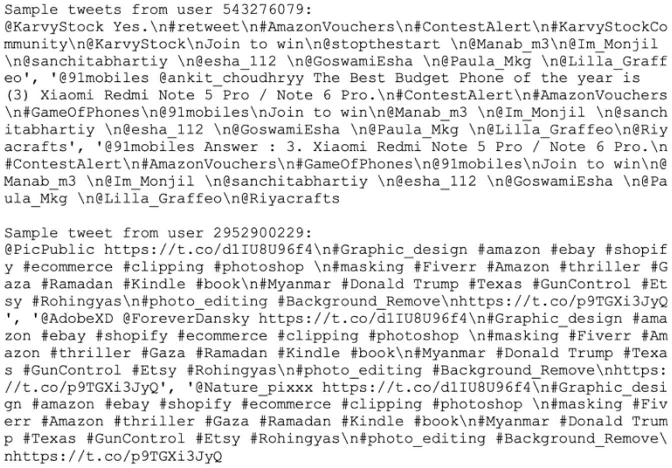
Finally, stock-specific daily sentiment scores were computed with a deep learning NLP model on the Twitter dataset. Originally collected and utilized by Izbicki, Papalexakis, and Tsotras (2019), the original data have all tweets with geolocation information spanning from October 17, 2017 to March 11, 2021 collected via the Twitter API. Since the datasets consist of 2.2 Terabytes of JSON files in over 100 languages, the dataset was filtered with four types of parameters: *Cashtags*, *Twitter Profiles*, *Hashtags*, and *Website URL*. As capturing general economic climate takes extensive filtering conditions, only stock relevant *Cashtags* and *Hashtags* were utilized. Furthermore, to ensure that the filter also captures any information related to the firms, Twitter profiles and relevant website URLs on user profiles were utilized. Hence, the following parameters were used in the data crawling process:

Table 1: *Filter Parameters for Twitter Data Import*

Filter Type	Parameters
Cashtags	$AAPL, $AMZN, $MSFT, $PYPL, $TSLA, $aapl, $amzn, $msft, $pypl, $tsla
Twitter Profiles	@Apple, @tim_cook, @Amazon, @AmazonNews, @JeffBezos, @Microsoft, @satyanadella, @PayPal, @AskPayPal, @Dan_Schulman, @Tesla, @elonmusk
Hashtags	#Apple, #AAPL, #aaple, #aapl, #Amazon, #AMZN, #amazon, #amzn, #Microsoft, #MSFT, #msft, #microsoft, #PayPal, #PYPL, #paypal, #pypl, #Tesla, #TSLA, #tesla, #tsla
Website URL	http://Apple.com, http://amazon.com, http://news.microsoft.com, http://tesla.com

As a result, a total of 732,489 tweets from October 22, 2017 to March 16, 2021 were parsed and imported to a PostgreSQL docker container. Hence, relevant tables were exported as comma-separated values (CSV) files, which were imported to a NVIDIA DGX Station maintained by the Quantitative and Computing Lab at Claremont McKenna College (QCL).

Figure 2: *Selected Samples of Removed Tweets*



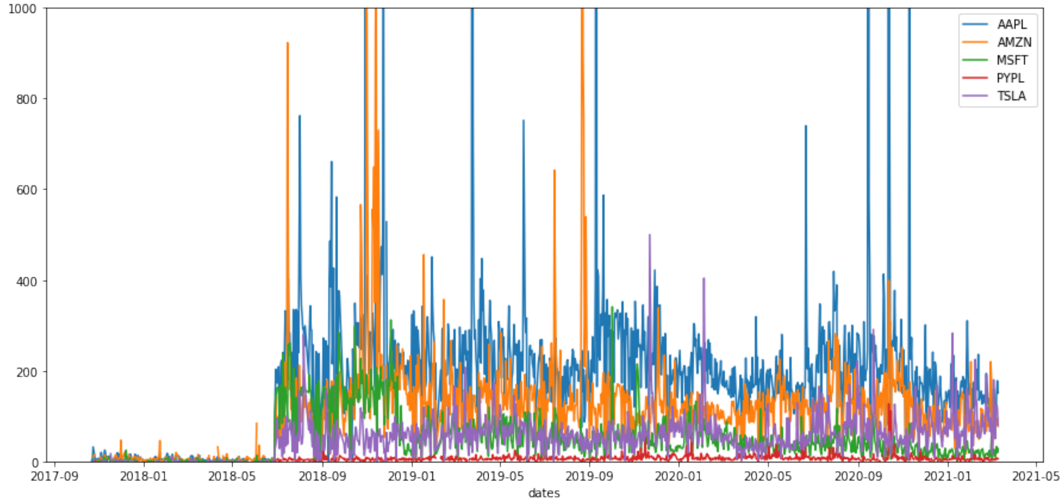
```
Sample tweets from user 543276079:
@KarvyStock Yes.\n#retweet\n#AmazonVouchers\n#ContestAlert\n#KarvyStockCo
mmunity\n@KarvyStock\nJoin to win\n@stopthestart \n@Manab_m3\n@Im_Monjil
\n@sanchitabhartiy \n@esha_112 \n@GoswamiEsha \n@Paula_Mkg \n@Lilla_Graff
eo', '@91mobiles @ankit_choudhryy The Best Budget Phone of the year is
(3) Xiaomi Redmi Note 5 Pro / Note 6 Pro.\n#ContestAlert\n#AmazonVouchers
\n#GameOfPhones\n@91mobiles\nJoin to win\n@Manab_m3 \n@Im_Monjil \n@sanch
itabhartiy \n@esha_112 \n@GoswamiEsha \n@Paula_Mkg \n@Lilla_Graffeo\n@Riy
acrafts', '@91mobiles Answer : 3. Xiaomi Redmi Note 5 Pro / Note 6 Pro.\n
#ContestAlert\n#AmazonVouchers\n#GameOfPhones\n@91mobiles\nJoin to win\n@
Manab_m3 \n@Im_Monjil \n@sanchitabhartiy \n@esha_112 \n@GoswamiEsha \n@Pa
ula_Mkg \n@Lilla_Graffeo\n@Riyacrafts

Sample tweet from user 2952900229:
@PicPublic https://t.co/d1IU8U96f4\n#Graphic_design #amazon #ebay #shopif
y #ecommerce #clipping #photoshop \n#masking #Fiverr #Amazon #thriller #G
aza #Ramadan #Kindle #book\n#Myanmar #Donald Trump #Texas #GunControl #Et
sy #Rohingyas\n#photo_editing #Background_Remove\nhttps://t.co/p9TGXi3JyQ
', '@AdobeXD @ForeverDansky https://t.co/d1IU8U96f4\n#Graphic_design #ama
zon #ebay #shopify #ecommerce #clipping #photoshop \n#masking #Fiverr #Am
azon #thriller #Gaza #Ramadan #Kindle #book\n#Myanmar #Donald Trump #Texa
s #GunControl #Etsy #Rohingyas\n#photo_editing #Background_Remove\nhttps:
//t.co/p9TGXi3JyQ', '@Nature_pixxx https://t.co/d1IU8U96f4\n#Graphic_desi
gn #amazon #ebay #shopify #ecommerce #clipping #photoshop \n#masking #Fiv
err #Amazon #thriller #Gaza #Ramadan #Kindle #book\n#Myanmar #Donald Trum
p #Texas #GunControl #Etsy #Rohingyas\n#photo_editing #Background_Remove\
nhttps://t.co/p9TGXi3JyQ
```

To prepare for the sentiment analysis, a further filtering process was needed as not all tweets were written in English nor contained relevant information. To efficiently filter out irrelevant tweets, I proceeded with grouping all tweets by Twitter User ID and ranked them based on the number of tweets after filtering out non-English messages. As can be seen from Figure 2, there were numerous tweets from Twitter users that are either spams or irrelevant to investment. Furthermore, a manual lookup of sample tweets from the

Twitter user list revealed that relevant user nodes often have relatively higher counts of *in_reply_to_status_id*, *in_reply_to_user_id*, and *quoted_status_id* instances. As a result, 43 Twitter user nodes were identified and removed from the Twitter dataset, resulting in a reduction of total tweet counts to 443,669.

Figure 3: *Counts of Total Tweet Counts by Tickers*



From Figure 3, it is observable that there are significantly larger sets of tweets for Apple, Amazon, and Tesla stock tickers. While sentiment variables extracted for the three tickers could be meaningful in the analysis, those for other tickers could not have a meaningful impact on the feature investigation process. Therefore, results for the three stocks can be compared with others to investigate whether a robust set of sentiment variables provide a meaningful impact on stock returns prediction. It is also imperative to note that there is a significant increase in tweet frequencies from June 2018. A sudden increase in the counts can be attributable to a change in a crawling program deployed by Izbicki et al. (2019) on June 28, 2018.

Table 2: *Summary of Sentiment Counts by Tickers*

Ticker	Negative Sentiment	Neutral Sentiment	Positive Sentiment	Total
Apple (Ticker: $AAPL)	5,634	201,037	3,082	209,753
Amazon (Ticker: $AMZN)	5,050	112,657	2,687	120,394
Microsoft (Ticker: $MSFT)	962	41,746	1,725	44,433
PayPal (Ticker: $PYPL)	206	6,483	88	6,777
Tesla (Ticker: $TSLA)	4,085	55,871	2,356	62,312

As modeling and training a deep-learning sentiment model requires an extensive computational resource, the FinBERT model, a pre-trained BERT language classification model, was utilized to compute sentiment scores of the tweets. Trained on a subset of Reuters' TRC2 corpus and Financial PhraseBank, The FinBERT outperformed the "state of the art" financial sentiment model by 15% in accuracy on FiQA sentiment dataset (Araci, 2019). To expedite the classification process, Hugging Face's pipeline API was leveraged to load the tweets and extract sentiment labels (Negative, Neutral, Positive) and softmax outputs on each tweet. As illustrated in Table 2, most tweets were categorized as neutral, with Apple, Amazon, and Tesla having the highest total tweet counts. Hence, it is also observed that there are generally more negative sentiments than positive sentiments except for Microsoft. To address a disproportionate number of neutral sentiments to other classifications, weighted sentiment scores were calculated on daily basis. Specifically, a weight or a proportion of each sentiment to a total tweet count is multiplied by a mean of relevant softmax values, which is then aggregated to derive a weighted sentiment score of given stock and time:

$$\begin{aligned}\text{Weighted Sentiment Score}_{it} &= \frac{w_{\mathrm{Positive}_{it}}\bar{\mathrm{s}}_{\mathrm{Positive}_{it}} + 0 \cdot w_{\mathrm{Neutral}_{it}}\bar{\mathrm{s}}_{\mathrm{Neutral}_{it}} - w_{\mathrm{Negative}_{it}}\bar{\mathrm{s}}_{\mathrm{Negative}_{it}}}{w_{\mathrm{Positive}_{it}} + w_{\mathrm{Neutral}_{it}} + w_{\mathrm{Negative}_{it}}} \\ &= \frac{w_{\mathrm{Positive}_{it}}\bar{\mathrm{s}}_{\mathrm{Positive}_{it}} - w_{\mathrm{Negative}_{it}}\bar{\mathrm{s}}_{\mathrm{Negative}_{it}}}{w_{Negative_{it}} + w_{Neutral_{it}} + w_{Positive_{it}}},\end{aligned} \tag{2}$$

where i refers to a stock, t refers to time, w refers to a weight of a relevant sentiment classification, and s refers to relevant sentiment softmax output.

Since the sentiment score was weighted and aggregated on daily basis by each stock, relevant descriptive statistics, such as means and standard deviations of daily sentiment labels and softmax values, were also included. Hence, a list of variables selected for the final analysis can be found in Table 3.

Table 3: *List of Variables in Final Dataset*

	Source / Method	Relevant Variables
Historical Prices	Alpha Vantage	Date, Open, High, Low, Close, Adjusted Close, Volume, Dividend Amount
Technical Analysis	Alpha Vantage	Simple Moving Average, Exponential Moving Average, Moving Average Convergence/Divergence (MACD-Histogram, MACD. MACD Signal Line), Stochastic Oscillator (Slow %K, Slow %D), Relative Strength Index (RSI)
Fundamental Analysis	Alpha Vantage	Shares Outstanding, Total Debt, Cash and Cash Equivalents, Reported EPS
	Calculated	ROA, ROE, ROIC, RORC, Profit Margin, Asset Turnover, Financial Leverage Ratio, Working Capital Turnover Ratio, EBITDA Margin, Gross Margin, Operating Margin, Inventory Turnover, PPE Turnover, Receivables Turnover, Payables Turnover, Current Ratio, Quick Ratio, Operating Cash Flow Ratio, Debt Ratio, Debt/Equity Ratio, P/E Ratio, Market Capitalization, Enterprise Value
Market Proxies	Nasdaq	S&P 500 (SPX): Date, Close/Last, Volume, Open, High, Low
	U.S. Department of Treasury	Daily Treasury Yields: Date, 1 Month, 2 Month, 3 Month, 6 Month, 1 Year, 2 Year, 3 Year, 5 Year, 7 Year, 10 Year, 20 Year, 30 Year
	Wall Street Journal (FactSet)	Dow Jones Industrial Average (DJI): Date, Open, High, Low, Close
	Yahoo Finance	Nasdaq Composite (IXIC): Date, Open, High, Low, Close, Adj Close, Volume CBOE Volatility Index (VIX): Date, Open, High, Low, Close, Adj Close, Volume
Sentiment Analysis	FinBERT	Tweet Counts (Negative, Neutral, Positive), Tweet Sentiment Scores – Mean (Negative, Neutral, Positive), Tweet Sentiment Scores – Standard Deviation (Negative, Neutral, Positive), Tweet Sentiment Scores – Minimum (Negative, Neutral, Positive), Tweet Sentiment Scores – Maximum (Negative, Neutral, Positive), Total Tweet Counts, Weighted Daily Sentiment Score

2. Experiment Setup

As the actual analysis involves predicting future logarithmic returns, all observations are lagged by a single day. While a traditional parametric model is helpful in examining a given hypothesis, it is not suitable for comparing across different arrays of variables as parameters must be specified. Therefore, I utilized the XGBoost regression model to analyze which variables are utilized to splitting decisions and have a meaningful impact on overall prediction.

Hence, the analysis can be broken into two large components: Baseline model for feature selection and optimized models via the Randomized Search Cross-Validation technique. For the baseline model, I utilize default XGBoost settings to conduct a feature selection process and decrease the number of parameters on the dataset. This process is crucial as it allows us to focus on parameters that have a significant impact on predicting stock returns. Also, 90% of the dataset is assigned for training and the rest for testing purposes as we are looking at a limited number of observations.

After the baseline models are assessed, I utilize two cross-validation methods, Time-Series Split Cross-Validation and Blocked Time-Series Cross-Validation, to ensure that our models do not overfit on the training dataset. Since we do not want the model to train in non-sequential order, it is imperative to adopt cross-validation strategies that account for such restrictions on time-series datasets. Time-Series Split Cross-Validation technique structures a training set on each *k* iteration before the validation sets, which guarantees that future observations are not utilized for predicting the past. While the method addresses the initial issue, it can also lead to another problem of data leakage in which the models might "observe and memorize" future patterns: Blocked Time-Series Cross-Validation addresses the issue by adding margins "between the training and validation folds" and between each iteration (Shrivastava, 2020). As the latter issue is not as critical as the former, we will utilize and consider the outcomes of the two methods. Furthermore, the following parameters were selected for Randomized Search Cross-Validation parameters:

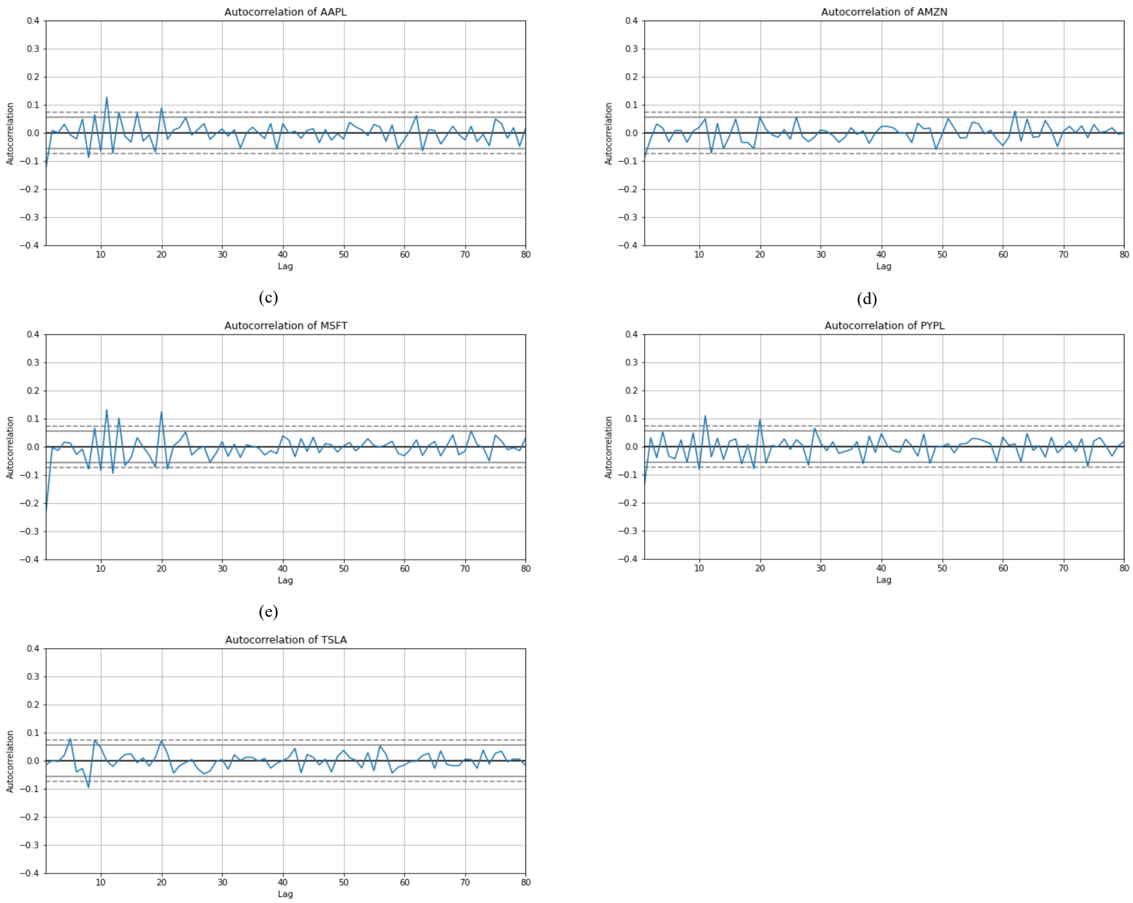
- Number of iteration: 5
- Number of Parameters Sampled: 200

- Learning Objective: Regression with squared loss
- Number of Estimators: Random integer within [150, 800]
- Learning Rate: Uniform continuous random variable within [0.01, 0.08]
- Subsample: Randomly Selected within [0.3, 0.5, 0.7, 0.9, 1.0]
- Max Depth: Randomly Selected within [6, 7, 8, 10, 12, 14, 16]
- Percent of Features Used for Tree: Uniform continuous random variable within [0.45, 0.90]
- Minimum Child Weight: Randomly Selected within [4, 8, 12, 14, 16]
- Evaluation Metric: Mean Squared Error

Lastly, I selected the XGBoost's F-score and SHAP values retrieved from the Shapley Additive exPlanations (SHAP) package as two metrics for quantifying the impact of the features. F-score is calculated by counting how many times a feature was used to make splitting decisions, thereby highlighting the importance of the feature in the fitting process of the model. On the other hand, the SHAP package utilizes a game theory framework called "Shapley Values" to make "black box" machine learning models more interpretable. Hence, the values utilize the "Shapley interaction index" to provide local interaction effects thereby capturing the impact of a certain feature among all variables (Lundberg et al., 2020). Hence, utilizing both metrics will allow us to identify which features are important in predicting stock returns.

IV. Empirical Results

Figure 4: *Autocorrelation Plots of Log Return Variables*



(a)

(b)

(c)

(d)

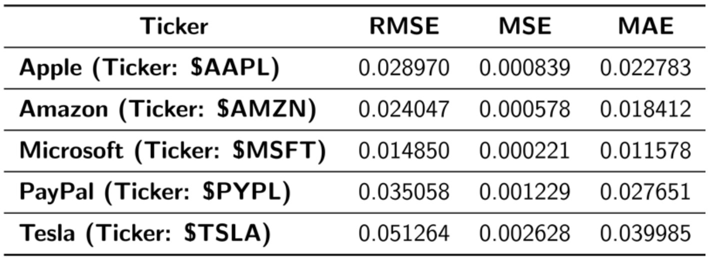
(e)

Unlike recurrent neural network models that are often implemented on time-series data (e.g., LSTM), the XGBoost regression model employs the tree method, which requires two additional considerations: lagged variables and time-related indicator variables. As Gu and Kurov (2020) emphasize in their research, it is important to control for return momentum as return correlation and "contemporaneous correlation of returns and sentiment" could lead to a lead-lag relationship. To identify a possibility of a lead-lag effect, I first looked at the autocorrelation of logarithmic returns of five stocks of interest. As can be seen from Figure 4, autocorrelation values jump 95% (solid line) and 99%

(dotted line) confidence bands for Apple, Microsoft, and PayPal stocks, especially up to approximately 20-day lags; while lesser in extents, such jumps can be seen from other stocks as well. While generating lagged variables for all available variables could be helpful, it is not suitable for this situation as it would result in a superfluous number of variables given limited time-series observations. Therefore, the lagged values from day t-1 to t-5 were only applied for adjusted closing prices of the stocks. To account for a possible seasonality in the stock prices, 6 different date-related indicator variables were added: day of the week, quarter, month, year, and two binary indicators for start and end of the month.

1. Baseline Results

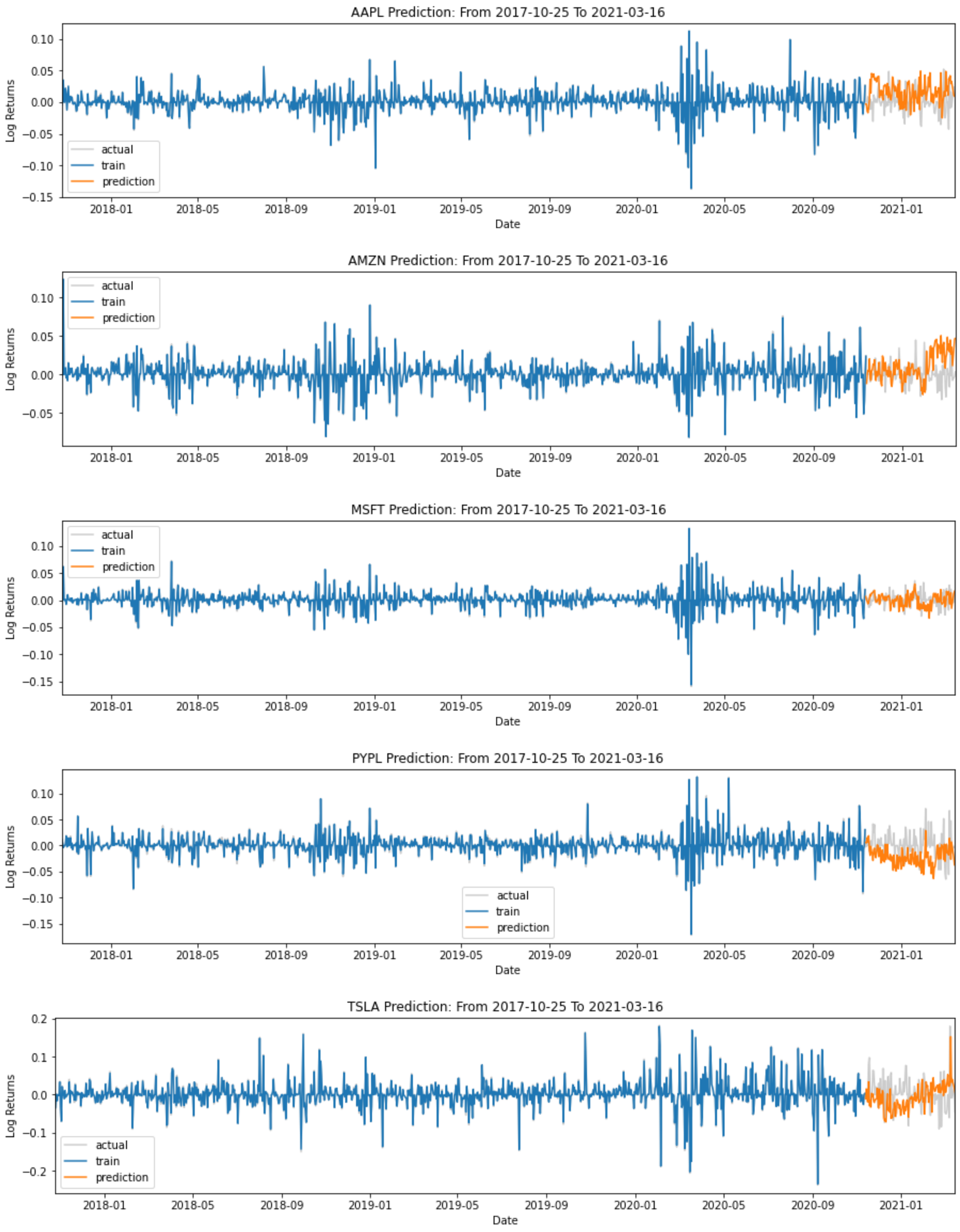
Table 4: *Metrics of Baseline XGBoost Regression Results*



Ticker	RMSE	MSE	MAE
Apple (Ticker: $AAPL)	0.028970	0.000839	0.022783
Amazon (Ticker: $AMZN)	0.024047	0.000578	0.018412
Microsoft (Ticker: $MSFT)	0.014850	0.000221	0.011578
PayPal (Ticker: $PYPL)	0.035058	0.001229	0.027651
Tesla (Ticker: $TSLA)	0.051264	0.002628	0.039985

Even though there was no hyperparameter tuning set in the baseline XGBoost regressions, the models fitted the test sets surprisingly well. From Table 4, we can see that the Microsoft model yielded the lowest Root Mean Squared Error (RMSE) while Tesla yielded the highest; such performance disparities were expected as normalized returns of Tesla were most volatile among the tickers. Furthermore, we can observe from Figure 5 that the PayPal and Tesla models fail to correctly capture accurate trends in stock returns on the test dataset. Therefore, it is imperative to account for an issue of overfitting and properly implement cross-validation and hyperparameter tuning techniques.

Figure 5: *Log Return Prediction Results of Baseline XGBoost Regressions*


AAPL Prediction: From 2017-10-25 To 2021-03-16

AMZN Prediction: From 2017-10-25 To 2021-03-16

MSFT Prediction: From 2017-10-25 To 2021-03-16

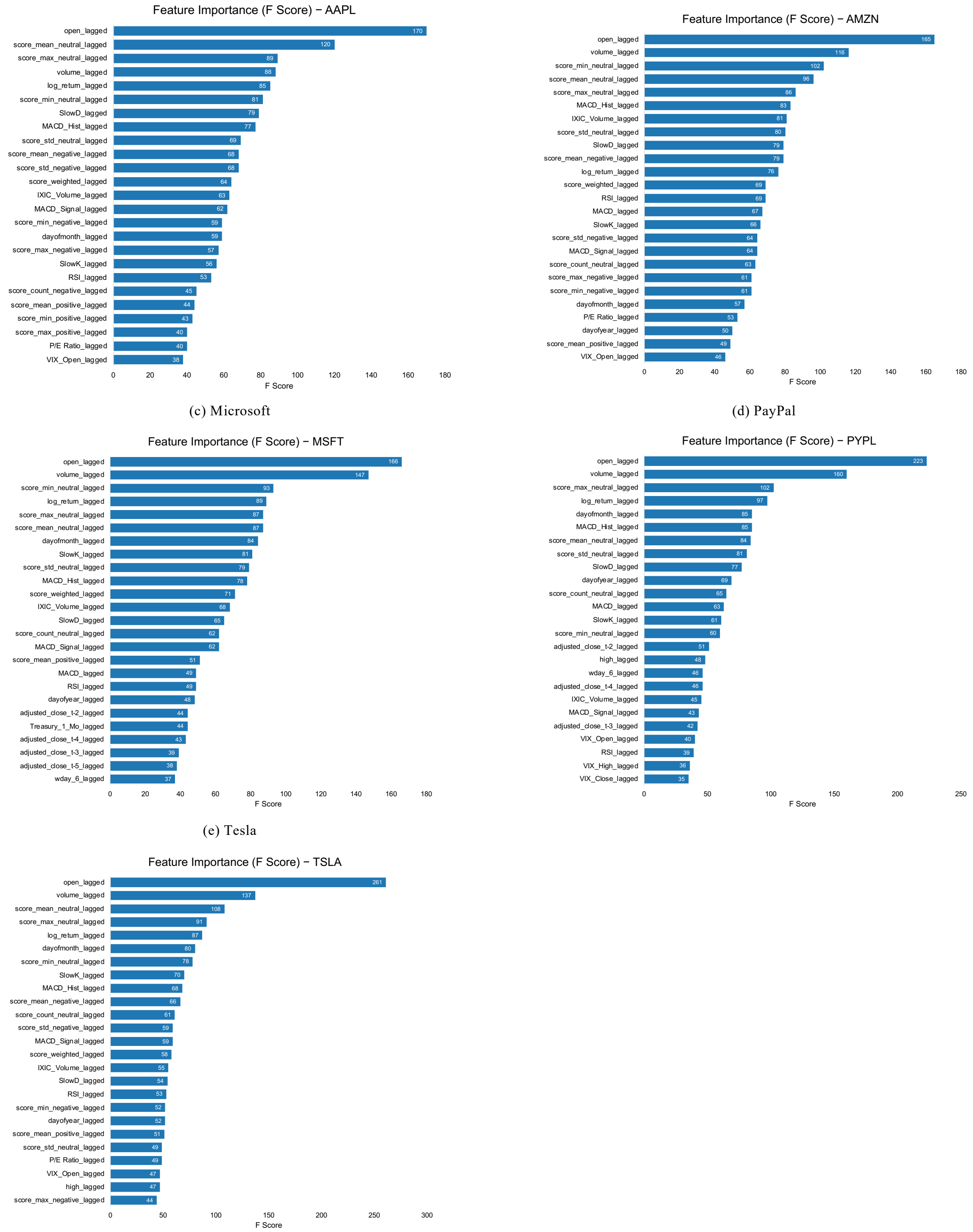
PYPL Prediction: From 2017-10-25 To 2021-03-16

TSLA Prediction: From 2017-10-25 To 2021-03-16

Comparing F-score outputs of the regressions, we can see that numerous types of variables are considered in splitting decisions. From Figure 6, it can be observed that previous days' historical price variables (e.g., *open*, *volume*, *log return*) are utilized the most across all stocks by the models to split the nodes. While there are still instances in which adjusted closing prices from t-3 to t-5 were used (e.g., Microsoft, PayPal), their scores rank relatively lower, suggesting that lag-lead relationships are not as pronounced as I initially believed. Interestingly, we can also see that neutral sentiment mean scores and their standard deviations are often employed, second to previous historical price variables mentioned above. It is important to note that technical indicators (i.e., *SlowD*, *MACD-Hist*) were also greatly contributed to splitting decisions made by the models. Lastly, we can also see that there are some seasonal aspects of the stock returns captured by the models as some date dummy variables, such as day of the months, rank relatively high in numerous stocks.

On the other hand, SHAP values provide some additional insights that are not well observed from F-scores. First, we can observe that low logarithmic returns in previous days generally have the largest positive impact on the next day returns prediction. For instance, it can be noted from Figure 7 that a low previous day return of Apple stock resulted in a SHAP value over 0.06, meaning that the variable provided an increase of over 0.06 in the predicted logarithmic return value for that specific observation. Furthermore, it seems that large daily high values of the VIX index and 1-Year Treasury Yields in previous days result in considerable decreases in predicted returns: such notion is validated in the actual stock market in which bearish market

Figure 6: *Feature Importance Plots of Baseline XGBoost Regressions*


(a) Apple

(b) Amazon

(c) Microsoft

(d) PayPal

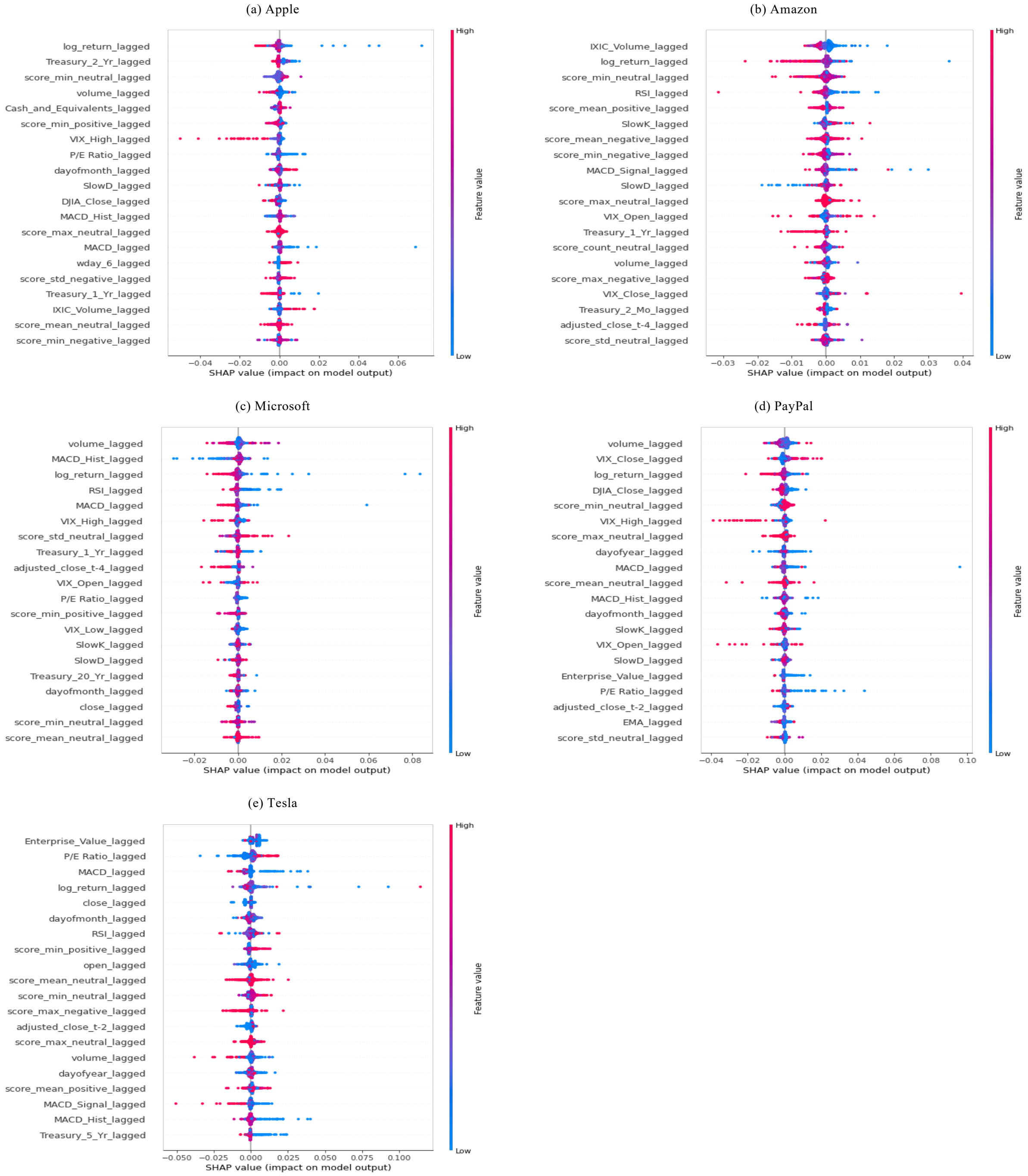
(e) Tesla

that *P/E Ratio* and *Enterprise Value* are utilized frequently. It is interesting to observe that the two variables both have a negative impact on next-day returns for PayPal while

P/E Ratio have a positive influence on Tesla: the differences can be attributable to a distinct market perception on Tesla as it recently turned a profit and is still considered in its growth stage.

While the above analysis exhibits crucial trends across all stocks, we can still observe some signs that the above models could suffer from the bias-variance tradeoff. While the cross-validation process alone can immensely solve the issue, it is still crucial to reduce parameters to reduce the chance of overfitting. Therefore, I extracted variables that have higher F-scores and SHAP values than the mean values of the two for all stocks. Hence, these variables were used for the cross-validated models.

Figure 7: *SHAP Plot of Baseline XGBoost Regressions*


(a) Apple

(b) Amazon

(c) Microsoft

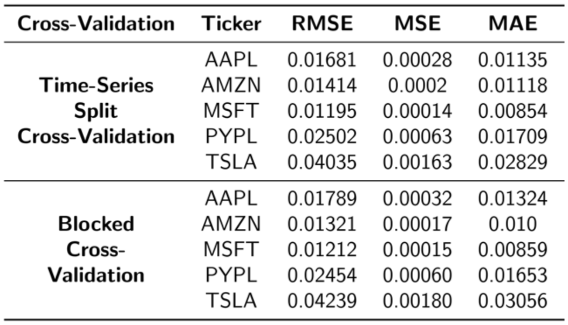
(d) PayPal

(e) Tesla

Note. Features are ordered in descending order by the cumulative SHAP values (overall impact)

2. Randomized Search Cross-Validation Results

Table 5: *Metrics of Cross-Validation Results*

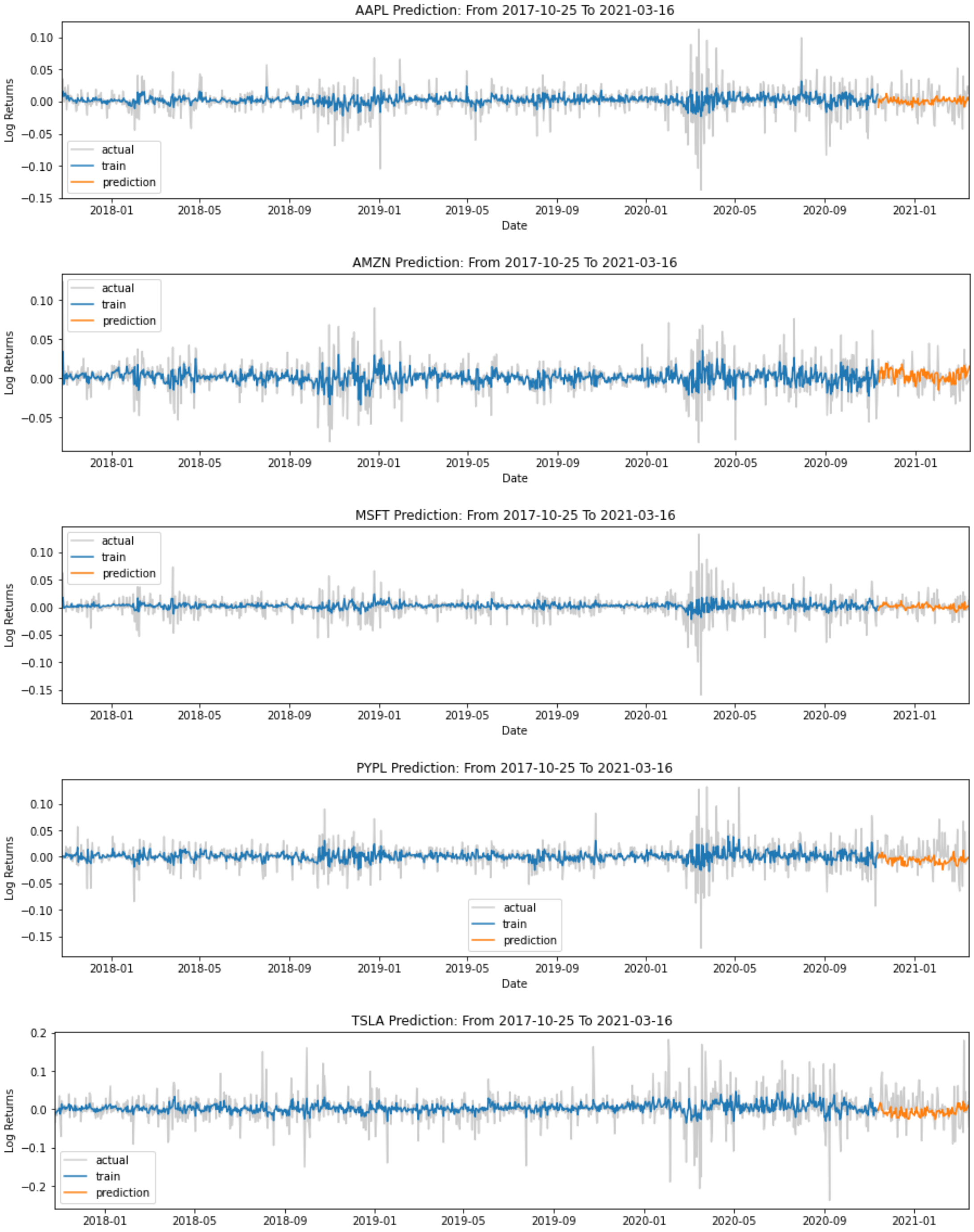


Cross-Validation	Ticker	RMSE	MSE	MAE
Time-Series Split Cross-Validation	AAPL	0.01681	0.00028	0.01135
	AMZN	0.01414	0.0002	0.01118
	MSFT	0.01195	0.00014	0.00854
	PYPL	0.02502	0.00063	0.01709
	TSLA	0.04035	0.00163	0.02829
Blocked Cross-Validation	AAPL	0.01789	0.00032	0.01324
	AMZN	0.01321	0.00017	0.010
	MSFT	0.01212	0.00015	0.00859
	PYPL	0.02454	0.00060	0.01653
	TSLA	0.04239	0.00180	0.03056

Through the Randomized Search Cross-Validation process, we can see that RMSE scores reduced greatly throughout all stocks. From Table 5, we can see that there have been mixed results, some having lower RMSE for the Time-Series Split Cross-Validation process. As there is only a slight difference between the two methods, I proceeded to select a model with the lowest RMSE value for each stock: the Time-Series Split Cross-Validation was selected for Apple, Microsoft, and Tesla and the Blocked Cross-Validation for Amazon and PayPal.

Even though it seems from Figure 8 that the XGBoost model performed best for Amazon, we can see that one fitted on Microsoft has the lowest RMSE. While implementation of a cross-validation strategy led to underfitting in all cases, most predictions seem to capture directions of logarithmic returns relatively well. While Randomized Search Cross-Validation was chosen due to the limited time of this study, it would be essential to consider a more robust set of hyperparameters and Grid Search Cross-Validation method to attain higher accuracies in the future.

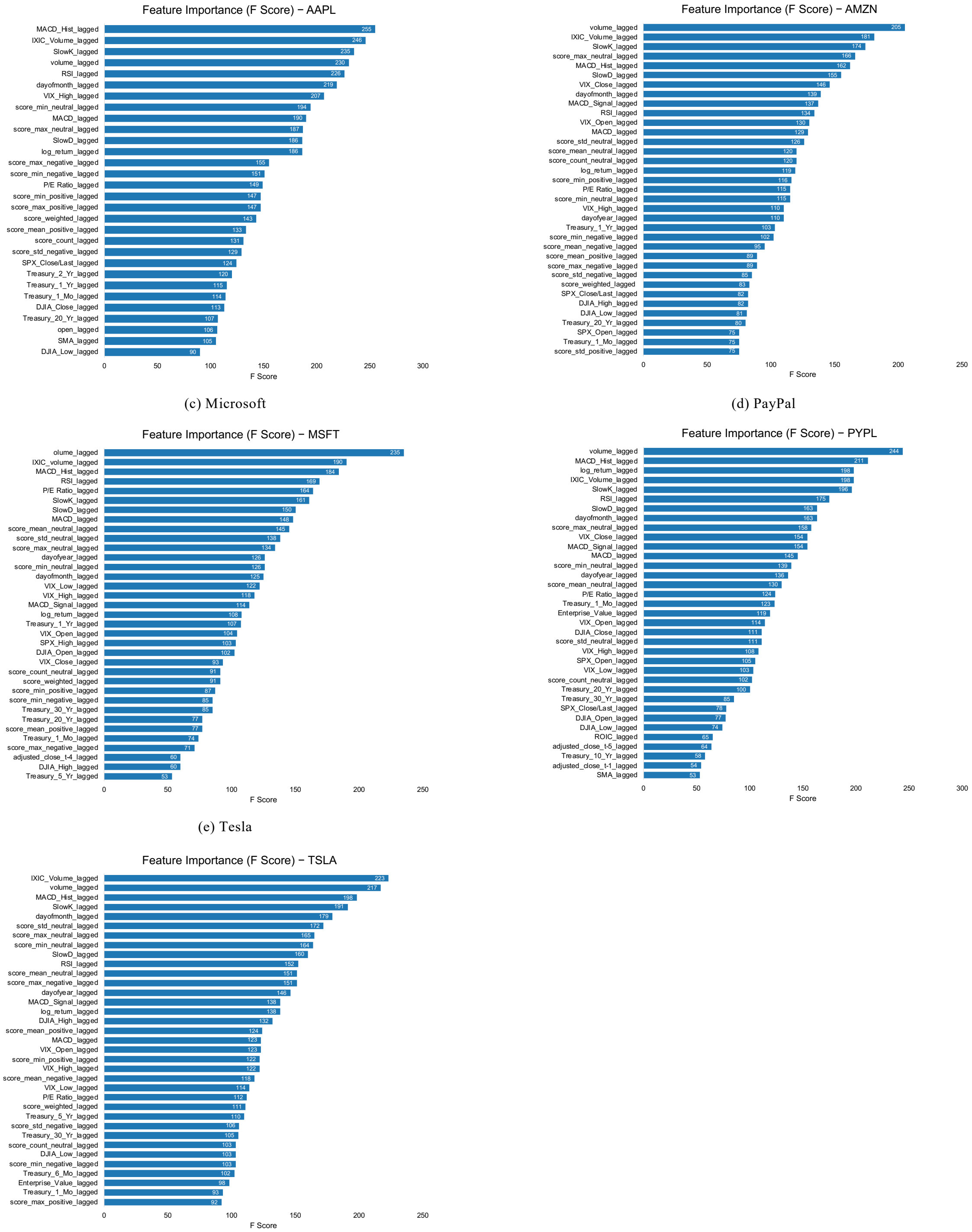
Figure 8: *Log Return Prediction Results of Selected Cross-Validated XGBoost Regressions*



As the number of features available for making splitting decisions decreased, it can be observed from Figure 9 that F-scores have significantly increased for the top features. While the trends that I have highlighted in the baseline models are well reflected for the cross-validated regressions, it is crucial to delve into SHAP values to further investigate whether any additional insights can be gained from the cross-validation results.

From Figure 10, we can observe that the effects of sentiment variables are much more pronounced for Apple, Amazon, Microsoft, and Tesla. For instance, high *score_max_neutral* values have significant positive and negative impact equally in predicting Amazon returns. However, high *score_count_neutral* values seem to have a negative impact on logarithmic return prediction, suggesting that there could be a high degree of ambivalence in public sentiment. Interestingly, *score_min_positive* is shown to have a positive impact on logarithmic returns, another notion that the finBERT variables well reflect the public sentiment of Amazon. Similar to the baseline models, the cross-validated models have strong inverse relationships for high previous days' *log returns*, *volume*, and *P/E Ratio* (except for Tesla).

Figure 9: *Feature Importance Plot of Selected Cross-Validated XGBoost Regressions*



(a) Apple

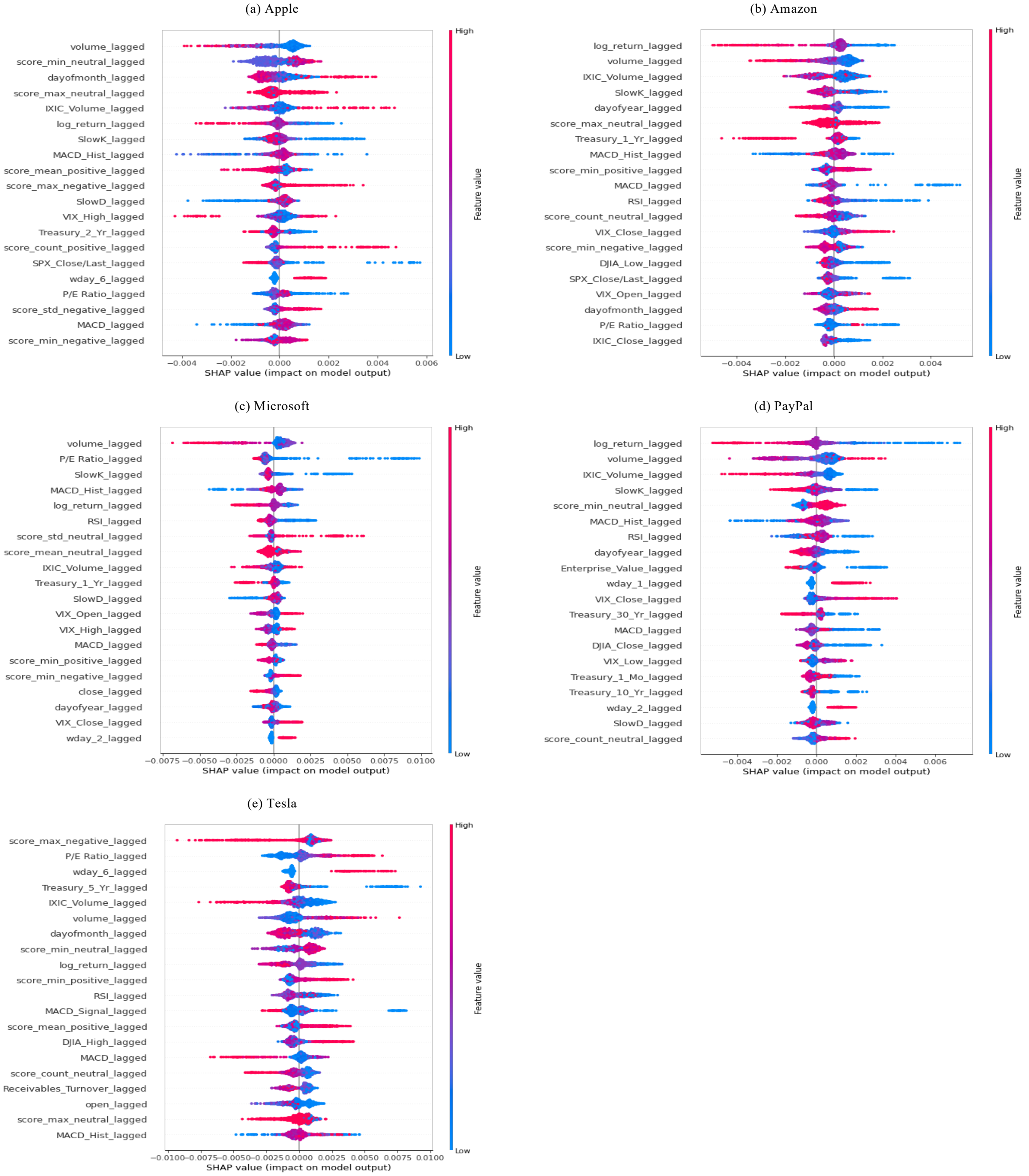
(b) Amazon

(c) Microsoft

(d) PayPal

(e) Tesla

Figure 10: *SHAP Plot of Selected Cross-Validated XGBoost Regressions*


(a) Apple

(b) Amazon

(c) Microsoft

(d) PayPal

(e) Tesla

V. Conclusion and Future Direction

As a goal of this paper is to identify how different variables impact stock prediction given stock-related sentiment, it clearly has shown that sentiments have predictive power and that both fundamental and technical analyses provide a significant impact on stock returns. Notably, all three sentiment classifications well reflect their presumed impact on stock returns, and some fundamental variables (e.g., *P/E Ratio*) have firm-specific effects in which their influence on the return predictions needs to be evaluated based on the qualitative aspects of the assets.

There are several implications from this research project. First, we have seen that it is more meaningful to assess statistical properties of sentiment variables than a weighted average measure. Second, it is possible to quantify the magnitude of impact features have on “black box” models, allowing quantitative researchers to incorporate qualitative metrics in gauging a firm’s value. Lastly, this study strengthens the notion that both fundamental and technical analyses are critical in identifying asset pricing.

Despite some success in measuring the impact of certain features on predicting stock returns, several issues need to be addressed. First, it is essential to have a larger dataset to ensure that we have enough observations for the training and test sets. Second, as the Twitter dataset utilized in this study is solely composed of geotagged tweets, it would be crucial to acquire non-geotagged tweets to guarantee that they represent the tweet population. Lastly, utilizing recurrent neural network techniques (e.g., LSTM) could be helpful as they can store memories of past observations, which is crucial in working with time-series data.

VI. References

Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to financial forecasting. *Applied Intelligence*, *6*(3), 205–213. https://doi.org/10.1007/BF00126626

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. http://arxiv.org/abs/1908.10063

Blystone, D. (2021). *Overbought or Oversold? Use the Relative Strength Index to Find Out*. Investopedia.

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance*, *46*(5), 1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x

Grossman, S. J., & Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, *70*(3), 393–408.

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, *121*, 105969. https://doi.org/10.1016/j.jbankfin.2020.105969

Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., & Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing*, *36*, 534–551. https://doi.org/10.1016/j.asoc.2015.07.008

Izbicki, M., Papalexakis, V., & Tsotras, V. (2019). Geolocating Tweets in any Language at any Location. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 89–98. https://doi.org/10.1145/3357384.3357926

Katona, Z., Painter, M., Patatoukas, P. N., & Zeng, J. (2018). On the Capital Market

Consequences of Alternative Data: Evidence from Outer Space. *9th Miami Behavioral Finance Conference*. https://doi.org/10.2139/ssrn.3222741

Lee, T. B. (2019). Web scraping doesn't violate anti-hacking law, appeals court rules. *Ars Technica*.

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. https://doi.org/10.1016/j.ipm.2020.102212

Lo, A., Mamaysky, H., & Wang, J. (2000). *Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation*. https://doi.org/10.3386/w7613

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, *17*(1), 59–82. https://doi.org/10.1257/089533003321164958

Mittal, A., & Goel, A. (2011). *Stock Prediction Using Twitter Sentiment Analysis*. Stanford University.

Nesbitt, K. V., & Barrass, S. (2004). Finding Trading Patterns in Stock Market Data. *IEEE Computer Graphics and Applications*, *24*(5), 45–55. https://doi.org/10.1109/MCG.2004.28

Schill, M. J. (2016). *Business Performance Evaluation : Approaches for Thoughtful Forecasting*.

Schlossberg, B. (2021). *Trading the MACD divergence*. Investopedia.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies*, *7*(2), 26. https://doi.org/10.3390/ijfs7020026

Shrivastava, S. (2020). *Cross Validation in Time Series*. Medium.

Alpha Vantage. (2021). *Alpha Vantage*. https://www.alphavantage.co/

Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, *15*(10), 1637–1656. https://doi.org/10.1080/14697688.2015.1071078

VII. Appendix

Figure A1: *Financial Ratios and Fundamental Variables Formulas*

1. Return on Equity (ROE) $= \frac{\text{Net Income}}{\text{Shareholders' Equity}}$
2. Return on Invested Capital (ROIC) $= \frac{\text{Net Operating Profit After Tax}}{\text{Total Debt + Shareholders' Equity}}$
3. Return on Research Capital (RORC) $= \frac{\text{Gross Profit}}{\text{R\&D Expense}_{t-1}}$
4. Profit Margin $= \frac{\text{Net Income}}{\text{Total Revenue}}$
5. Asset Turnover $= \frac{\text{Total Revenue}}{\text{Total Asset}}$
6. Financial Leverage $= \frac{\text{Total Asset}}{\text{Shareholders' Equity}}$
7. Working Capital Turnover Ratio $= \frac{\text{Total Revenue}}{\text{Average Working Capital}}$
8. EBITDA Margin $= \frac{\text{EBITDA}}{\text{Total Revenue}}$
9. Gross Margin $= \frac{\text{Total Revenue}-\text{Cost of Goods Sold}}{\text{Total Revenue}}$
10. Operating Margin $= \frac{\text{Operating Income}}{\text{Total Revenue}}$
11. Inventory Turnover $= \frac{\text{Cost of Goods Sold}}{\text{Average Inventory}}$
12. PPE Turnover $= \frac{\text{Total Revenue}}{\text{Average PPE}}$
13. Receivables Turnover $= \frac{\text{Total Revenue}}{\text{Average Accounts Receivable}}$
14. Payables Turnover $= \frac{\text{Cost of Goods Sold}}{\text{Average Accounts Payable}}$
15. Current Ratio $= \frac{\text{Current Assets}}{\text{Current Liabilities}}$
16. Quick Ratio $= \frac{\text{Cash and Cash Equivalents}+\text{Accounts Receivable}}{\text{Current Liabilities}}$
17. Opreating Cash Flow Ratio $= \frac{\text{Cash Flow From Operations}}{\text{Current Liabilities}}$
18. Debt Ratio $= \frac{\text{Total Debt}}{\text{Total Asset}}$
19. Debt/Equity Ratio $= \frac{\text{Total Debt}}{\text{Shareholders' Equity}}$
20. P/E Ratio $= \frac{\text{Closing Price}}{\text{Reported EPS}}$
21. Market Capitalization $= \frac{\text{Closing Price}}{\text{Shares Outstanding}}$
22. Enterprise Value $= \text{Market Capitalization}+\text{Total Debt}-\text{Cash and Cash Equivalents}$