Predicting Amazon Stock Prices Using Headlines

Reegan Fahy

09/12/2024

Introduction

According to an article in CNBC, investment gurus Warren Buffet and Mark Cuban spend hours every day reading. Buffet is quoted to "spend five or six hours a day reading", apparently favoring *The Wall Street Journal, The Financial Times, The New York Times, The USA Today, The Omaha World-Herald* and *American Banker.* Similarly, Cuban claims three hours a day for reading. Both men credit their voracious reading habits as a key to their business success. "I feel like if I put in enough time consuming all the information available, particularly with the net making it so readily available, I can get an advantage in any technology business," says Cuban (Montag, "Warren Buffett and Mark Cuban Agree This One Habit").

Unfortunately, most people don't have the time, financial resources or inclination to dedicate hours a day to reading. Indeed, a 2014 study published by the American Press Institute estimates that 6 in 10 Americans don't read more than the headlines (API Team, "How Americans Get Their News"). I am, admittedly, guilty of this exact practice; only choosing to fully read certain articles as they appeal to me. At the same time, who doesn't want to be as fiscally prophetic as Warren Buffet?

The model I set out to build is aimed at assisting headline-readers seeking insight into market movements based on headlines alone. Specifically, the user inputs one or more headlines as well as the day's Amazon stock price. The then model assesses a VADER sentiment score of a headline and uses that score to determine price changes in Amazon stock for the following day.

Methodology

I began by selecting two datasets from Kaggle from which to train and test my model. The first was a collection of stock prices on the US Market over a five year period and the second, 24 years of New York Times article headlines. Each dataset was inspected then cleaned for my purposes.

To judge the sentiment of headlines, I applied Sentiment Analysis (SIA) and TF-IDF Vectorization to obtain a VADER (Valence Aware Dictionary and Sentiment Reasoner) score. I selected this combination because SIA is especially useful for measuring tone in text and TF-IDF uses word frequency to identify significant words. The accuracy of my VADER sentiment score was analysed using Linear regression, and Random Forest and Gradient Boosting deep learning (see *Table 1*). The Gradient Boosting regression provided the best model results due to its ensemble nature. On average, the Mean Squared Error (MSE) between the predictions and the actual sentiments was quite small, suggesting that the model is effective in minimizing prediction errors. In addition, 59% of the variability in sentiment scores is accounted for by the model. Given the complexity of language, this is a reasonably satisfactory result.

My two datasets were then merged based on a common timeframe (03 Feb 2019 to 28 Dec 2023). For each day, there was one stock price available, but numerous headlines. To

account for this, I averaged each headline's VADER sentiment into one, daily sentiment score, then aggregated the data based on date. My expectation was that headline sentiment may not affect stock price right away, but might take a number of hours or days to "hit" the market. I therefore added a series of calculated columns, including lagged sentiment scores from 1 day, 2 days and 7 days, as well as some additional stock price insights such as change from previous day. The final cleaned and merged dataset results in 1,081 entries (see *Table 2*).

Now merged, I used a correlation matrix to assess which stock price was most highly correlated to VADER sentiment. It may come as a surprise to no one that the stock prices had little to no relationship to the headline sentiment score. Nor did the relationship improve or decline when comparing the price to sentiment scores from the same day, from 1 day ago, 2 days ago or 7 days ago. Interestingly, the sentiment scores themselves were also poorly correlated, showing that headline sentiment was mostly unrelated to that of neighboring days (see *Image 1*). At this stage, I considered reworking the basis of my project, but ultimately persevered; based on the correlation matrix, I selected Amazon stock to use in my model because it had the closest positive correlation to headline sentiment (lagged 1 day and lagged 2 days) (see *Image 2* and *Table 3).*

Conscious of the low correlation, I tested a wide range of models to find the method for predicting Amazon Price from VADER sentiment (see *Table 4*). The best result was a Ridge Regression Model which was surprising, since a Ridge Regression is a linear model and I would have expected stocks to be more non-linear. Although the best fit available from those tested, the Ridge Regression is still a poor fit. The MSE of 890.12 translates to a prediction deviation of 29.83 units compared to the actual result. The average price of Amazon Stock Price is $128, meaning the model deviates from the actual price of the stock on average by 23%. The R2 score of the Ridge Regression is also quite low, implying that the model can only explain around 8% of the variance.

Financial Implications

Taken for its accuracy, the model provides no substantial financial implications. Possibilities for a different approach include:

- Include different news outlets such as a financial media company
- Further narrow the news desks from which articles were sourced
- Include scoring for sub titles and/or abstracts in assessing relationship to stock prices
- Use a different language model, perhaps one that is focused on financial dialogue specifically
- Use a broader spread of stock prices
- Remove outliers in data to tighten results
- On the user interface, require a full day's worth of headlines, rather than select few

With inexhaustible time and resources, the ideal model would include, but not be limited to, news headlines. After all, the markets are influenced by an incredible array of factors, not just the newscycle. For all its flaws, the model is interesting in showing just how *un*related headlines are (when taken in isolation) to the movement of the markets. It's unfortunate for me and my fellow headline-skimmers; In the age when we have an endless spread of news outlets to choose from, perhaps the only aim of headlines is to capture your attention, rather than provide time-efficient insights.

Word count: 1,059

Image 1: Correlation Matrix Heatmap - Stock Prices versus Sentiment Scores

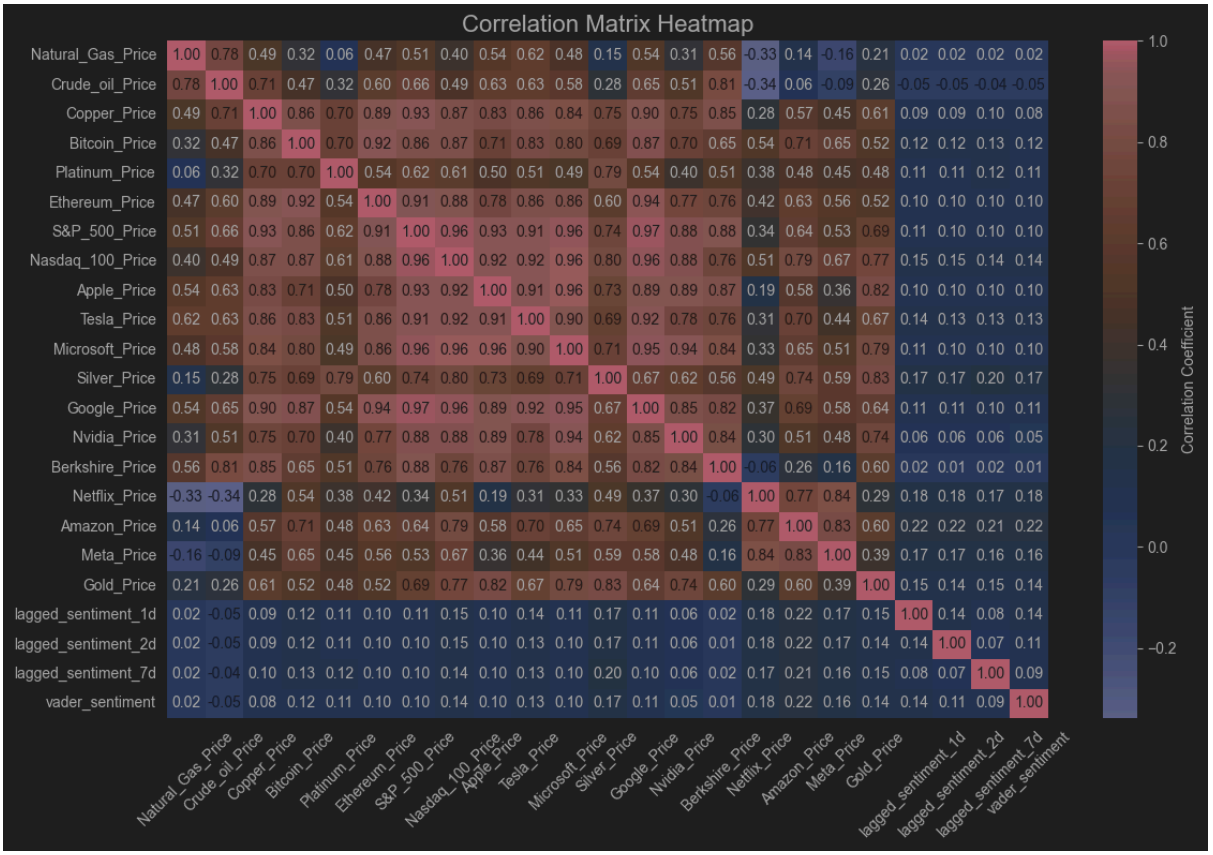

Image 2: Timeline of Amazon Prices versus Daily VADER Sentiment scores (scaled)

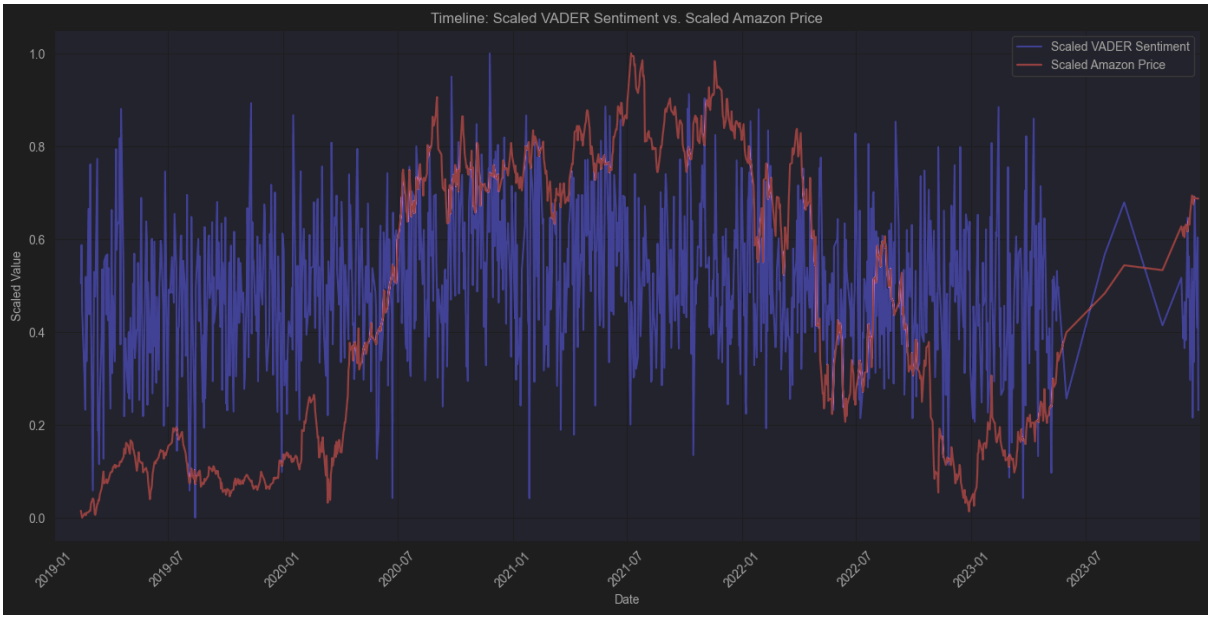

Table 1: Modeling VADER Sentiment Scoring

Model	**Model Type**	**MSE**	**R2**
Linear	Regression	0.08	0.55
Random Forest	Machine Learning	0.10	0.46
Gradient Boosting	Machine Learning	0.07	0.59

Table 2: Merged Dataframe

Entries	1,081
Column Size	81 Columns
Data Types	datetime64[ns](1), float64(60), int64(20)
Date Range	2019-02-13 to 2023-12-28
Columns	'Date' 'Vader_sentiment' 'Natural_Gas_Price' 'Crude_oil_Price' 'Copper_Price' 'Bitcoin_Price' 'Platinum_Price' 'Ethereum_Price' 'S&P_500_Price' 'Nasdaq_100_Price' 'Apple_Price' 'Tesla_Price' 'Microsoft_Price' 'Silver_Price' 'Google_Price' 'Nvidia_Price' 'Berkshire_Price' 'Netflix_Price' 'Amazon_Price' 'Meta_Price' 'Gold_Price' 'Natural_Gas_Price_Difference' 'Natural_Gas_Price_Percent_Change' 'Natural_Gas_Price_Change_Indicator' 'Crude_oil_Price_Difference' 'Crude_oil_Price_Percent_Change' 'Crude_oil_Price_Change_Indicator' 'Copper_Price_Difference' 'Copper_Price_Percent_Change' 'Copper_Price_Change_Indicator', 'Bitcoin_Price_Difference'

	'Bitcoin_Price_Percent_Change' 'Bitcoin_Price_Change_Indicator' 'Platinum_Price_Difference' 'Platinum_Price_Percent_Change' 'Platinum_Price_Change_Indicator' 'Ethereum_Price_Difference' 'Ethereum_Price_Percent_Change' 'Ethereum_Price_Change_Indicator' 'S&P_500_Price_Difference' 'S&P_500_Price_Percent_Change' 'S&P_500_Price_Change_Indicator' 'Nasdaq_100_Price_Difference' 'Nasdaq_100_Price_Percent_Change' 'Nasdaq_100_Price_Change_Indicator' 'Apple_Price_Difference' 'Apple_Price_Percent_Change' 'Apple_Price_Change_Indicator' 'Tesla_Price_Difference' 'Tesla_Price_Percent_Change' 'Tesla_Price_Change_Indicator' 'Microsoft_Price_Difference' 'Microsoft_Price_Percent_Change' 'Microsoft_Price_Change_Indicator' 'Silver_Price_Difference' 'Silver_Price_Percent_Change' 'Silver_Price_Change_Indicator' 'Google_Price_Difference' 'Google_Price_Percent_Change' 'Google_Price_Change_Indicator' 'Nvidia_Price_Difference' 'Nvidia_Price_Percent_Change' 'Nvidia_Price_Change_Indicator' 'Berkshire_Price_Difference' 'Berkshire_Price_Percent_Change' 'Berkshire_Price_Change_Indicator' 'Netflix_Price_Difference' 'Netflix_Price_Percent_Change' 'Netflix_Price_Change_Indicator' 'Amazon_Price_Difference' 'Amazon_Price_Percent_Change' 'Amazon_Price_Change_Indicator' 'Meta_Price_Difference' 'Meta_Price_Percent_Change' 'Meta_Price_Change_Indicator' 'Gold_Price_Difference' 'Gold_Price_Percent_Change' 'Gold_Price_Change_Indicator' 'Lagged_sentiment_1d' 'Lagged_sentiment_2d' 'lagged_sentiment_7d'

Table 3: Amazon Stock Price Details

Minimum Price	$80.40
Maximum Price	$186.57

Mean Price	$127.99

Table 4: Modeling Relationship between VADER sentiment and Amazon Price

Model	**Model Type**	**MSE**	**R2**
Linear	Regression	887.65	0.08
Random Forest	Machine Learning	1182.36	-0.22
Gradient Boosting	Machine Learning	2081.43	-1.15
Ridge	Regression	890.12	0.08
Lasso	Regression	890.74	0.08
ElasticNet	Regression	957.09	0.01
Support Vector	Machine Learning	928.10	0.04
K-Nearest Neighbors	Machine Learning	1087.87	-0.12
Multi-Layer Perceptron	Deep Learning	11,114	

Sources

API Team. “How Americans Get Their News.” *American Press Institute*, American Press Institute, 1 Dec. 2023, americanpressinstitute.org/how-americans-get-news/.

Montag, Ali. “Warren Buffett and Mark Cuban Agree This One Habit Is Key to Success-and Anyone Can Do It.” *CNBC*, CNBC, 15 Nov. 2017, www.cnbc.com/2017/11/15/warren-buffett-and-mark-cuban-agree-reading-is-key-to-success.html.