Predicting Stock Price Movements Remain Unfeasible

Benjamin Minkin

May 16, 2024

1. Introduction

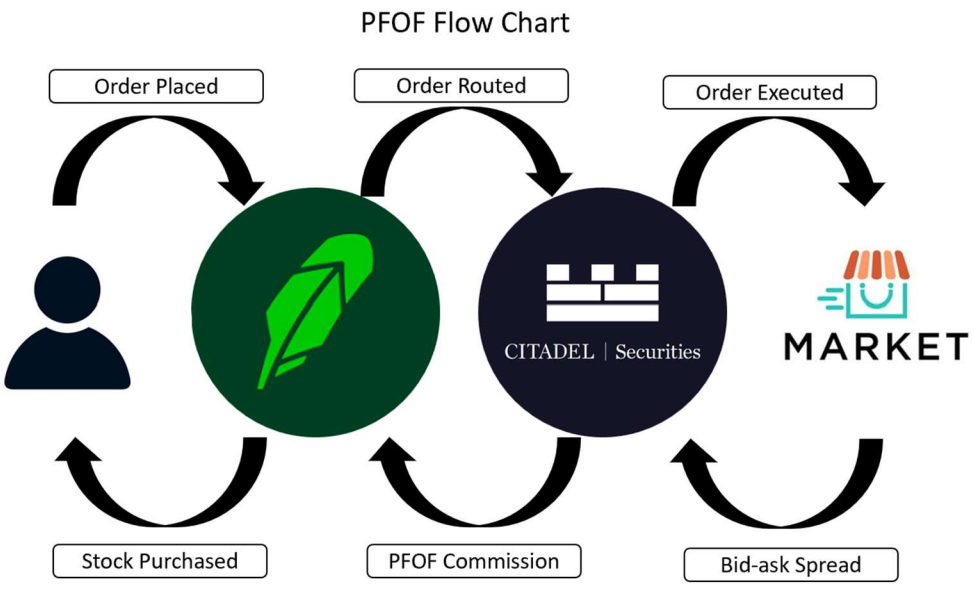
1.1 GameStop Mania

The month is January 2021. The first year of the global pandemic has concluded, and there is cautious optimism that the new year will be one of recovery. Lockdowns stemming from the pandemic have had devastating effects on retailers' valuations. Yet one company, forlorn for bankruptcy, sees its stock price begin climbing to unimaginable highs, defying all expectations. This event, later dubbed GameStop mania, caused billions of dollars to exchange hands. Over the course of this month, one investor, Keith Gill, saw his $53,000 initial investment skyrocket to a peak valuation of approximately $48 million, a 90,000% increase. [1] On the flip side, short-selling institutions saw exorbitant losses. One New York investment firm, Melvin Capital, lost an estimated $3 billion and required a bailout from hedge funds, including Citadel LLC. [2] This market frenzy reached its peak on January 28, when brokers including RobinHood, WeBull, and Interactive brokers removed their customers' ability to purchase shares of GameStop stock, preceding a steep crash in the price. Embroiled in the midst of this episode and the subsequent hearing in the House Financial Committee was a controversy regarding the practice called payment for order flow.

1.2 Payment for Order Flow

Payment for order flow (hereinafter PFOF) is the practice of brokers routing their customers' trades through market makers, such as Citadel LLC, in exchange for a fee. The earnings from PFOF were a driving force behind the aforementioned trading platforms offering

commission-free trading. In 2020 alone, the amount paid to brokers by market makers for PFOF is estimated to be around $2.6 billion. [3] These market makers execute the trade for the customer and are obligated to match the current national best bid and offer. One way market makers profit is by connecting buyers and sellers who place market orders. These orders carry instructions for the broker to execute the order immediately at the market price. This allows the market maker to profit on the bid-ask spread, as they can have one customer sell the stock at the bid and the other customer buy the stock at the ask. The market maker pockets this difference in what the buyer pays and what the seller gets paid. It is important to note that this is different from the market maker buying the security in order to push up the price to resell to a buyer, an illegal practice called frontrunning. In PFOF, the customer gets the stock at the market price commission-free, the market maker profits from executing the trade, and the broker gets a cut for routing the trade to the market maker.



1.3 The Untapped Potential of PFOF Order Flow Information

The primary focus of the hearing was to discuss incentives influencing market makers that may have been against their consumers' best interests. This includes the motive to maximize the bid-ask spread, the motive for frontrunning, and their soft power over brokers. The focus of this paper is not to speculate on whether the parties in question acted improperly, but rather to examine ways to extract valuable insights from this order flow data. Crucially, market makers have first access to this order flow information. This grants them a time-based advantage, as they can start digesting this data before anyone else. This information asymmetry, paired with fast processing algorithms, may allow this data to be used in high-frequency trading (hereinafter HFT) to profitably predict short-term market movements.

2. Overview

2.1 Scientific Modeling

The scientific method is a process created in the 17th century to learn about the world around us while limiting bias. This same process is used today in data science to gain insights from data. The first step of the scientific method is to observe a phenomenon. A phenomenon is any notable event, process, or feature of the observable world. This can be a result of a natural process, like the volume of snowfall on a particular mountain, or man-made processes, like the price to go skiing. After observing a phenomenon, a scientist can conjecture a test hypothesis regarding the causal drivers that underlie this phenomenon. This hypothesis informs the creation of a test model for experimentation. This test model is an approximation of reality that may be used to help understand how things truly operate. This is because a model, by definition, will never be a perfect match to reality, but it can be good enough to be used in inferential learning that has real-world applications. As George Box aptly put it, "All models are wrong, but some are useful." The experimentation conducted with the test model leads to gathering data, which is

used to infer the relative accuracy of the scientist's hypothesis. In hypothesis testing, this accuracy is compared to a baseline null model, where the goal is to reject the validity of the baseline for the test hypothesis. This is done by showing the outputs are unlikely to have resulted from the null hypothesis. We will go through each step of this process and how it is applied in our situation, starting with the observation.

3. Observation

3.1 The Phenomenon

In our case, the phenomenon we are measuring is stock price movement. Over a period of time, the purchase price of a stock fluctuates. This output (y) is measured in many ways, including total dollar change and percent change. For the purposes of this paper, we will define the output as either a net gain or loss over the time period. This means that our model will have an output space of $Y \in \{0,1\}$ where 0 is a decrease in price and 1 is a gain in price. A sideways movement where there is no price change will be set to 0 due to opportunity loss. By limiting the output to two possibilities, predicting stock price movements becomes a binary classification problem.

3.2 The Efficient Market Hypothesis

Since there has been a market where goods are bought and sold, speculators have attempted to predict future prices. If one believes that an item will increase in price, they could purchase the item now and resell it later. Nowadays, predicting future stock prices is a multi-trillion-dollar industry. Analysts from hedge funds, banks, and mutual funds are constantly creating models to identify profitable opportunities. Nevertheless, studies show that these professional predictors rarely outperform the baseline performance of the S&P 500 in the long

term.[4] One possible explanation for this underperformance is the efficient market hypothesis. The efficient market hypothesis is an economic theory that states that assets are priced according to all available information. To profit consistently, one must use information that isn't already integrated into the stock price. Consequently, as strategies exhibit their profitability by uncovering information, that information is acted on and becomes priced in. This negates the strategy's profitability in the long run. Put simply, a winning strategy over one time period may not be winning in the next. This dynamic evolution of the underlying generation process is called non-stationarity. This causes a serious dilemma when predicting future movements from past data. However, one way to consistently profit is to react to new information faster than everyone else. Already, HFT firms utilize tools that react to new information nearly instantaneously. [5] The data pipeline from PFOF provides a unique opportunity to react to news even quicker, almost as if it is still happening. By reacting to this news the fastest, strategies using PFOF data may be able to bypass the efficient market hypothesis. This would also suggest that the process can have stationarity and consistent profitability. The legal and ethical ramifications of this practice will not be explored. As far as this paper is concerned, as long as the market maker supplies the consumer with the shares at the best price, it has every right to use the information in its proprietary prediction models.

4. Hypothesis

4.1 Theoretical Test Hypothesis

Before predicting the direction of price movements, it is important to discuss the mechanism that sets the price of a stock. There is a fundamental economic theory called the law of supply and demand, which posits that prices are determined by the supply and demand for a

product. When the supply rises, the price decreases, and when the demand rises, the price increases. It is therefore natural to assume that changes in stock prices reflect changes in this supply-demand dynamic. If this assumption holds true, this relationship can be modeled using the indicator function $\{y = \mathbf{1}(z1 > z2)\}$, where z1 is the change in demand and z2 is the change in supply. If demand exceeds supply, then the output y will be 1 and the stock will go up in price. Conversely, if supply exceeds demand, the output y will be 0 and the stock will go down in price.

4.2 Practical Test Hypothesis

Unfortunately, this model is practically impossible to implement as supply and demand are not perfectly quantifiable. We must therefore introduce quantifiable proxies (x) that are similar to the true market mechanics (z). This would modify the previous model to the mathematical model $\{\hat{y} = \mathbf{1}(x_1 > x_2)\}$. Its interpretation is as follows: If the proxy for change in demand (x_1) exceeds the proxy for change in supply (x_2), the expected output ($\hat{y}$) is that the stock price will increase. This model, unlike the first, is a mathematical model. A mathematical model is one where the inputs are numerically quantifiable. The model takes these numerical inputs and uses them to deterministically generate an output. This makes the process replicable and transparent. It is, however, simply a mathematical function that does not logically understand the underlying data. For this reason, choosing the best proxies is of utmost importance. This is because the more they reflect the market mechanics, the more accurate the model will be. If proxies that have no relation to the causal drivers are chosen, the model will not have any predictive power. Because these proxies will never completely reflect supply and demand, there will always be some error induced, which will be discussed later on.

4.2 Null Hypothesis

A core assumption of this paper is that price movements are determined by supply and demand. Then, by creating a model to proxy these mechanisms, price movement can be predicted. This assumption may be incorrect on its face. A competing hypothesis about the mechanism driving stock price movements is the random walk hypothesis. The random walk hypothesis states that stocks' movements can be modeled by a random walk process. A natural extension of this hypothesis is that predicting movements is inherently impossible. This assumption of randomness underlies many Brownian financial models, including the Nobel Prize-winning Black-Scholes option pricing model. For the purpose of this paper, by assuming the drift constant is negligible, this hypothesis will be simplified to assume a stock will go up 50% of the time and down the other 50% within a short time span. Formulated next to the test hypothesis, this looks like: $\{H_{Null}: y \sim Bern(0.5)\}$ $\{H_{Test}: y = \mathbf{1}(z1>z2)\}$. The goal of the test model will be to outperform this baseline in a statistically significant way. This would make the model, at the very least, theoretically useful. Then, if the model can be practically implemented by HFT firms, it will be declared practically useful. The practical requirements will inform the selection of the test model.

5. Model selection

5.1 Model Requirements

Our model will also have to contend with being practically implementable. Even if the model can predict directional movements in the next second, if it takes longer than a second to compute, the opportunity to profit will pass. These technical constraints, as well as the model selection problem, will help determine the specifications of the test model.

5.2 The Model Selection Problem

The model selection problem describes a balancing act that all supervised learning models must endure. On one side, a model can focus too strongly on fitting the training data, where it will not be able to generalize the relationships between the inputs and the outputs. This is called overfitting and often happens when there is not enough training data, among other reasons. When overfit, a model captures the noise in the data rather than the signal. It is unable to separate normal fluctuations from causal relationships. Consequently, when the model is tested on new data outside of the training set, it will do poorly. On the other side of the spectrum, a model can make such broad generalizations between input and output that the true complicated relationships are not captured. This is called underfitting and happens when the number of high-quality proxies is too small. In this case, the model would benefit from more dimensionality to capture the complicated relationship between inputs and outputs.

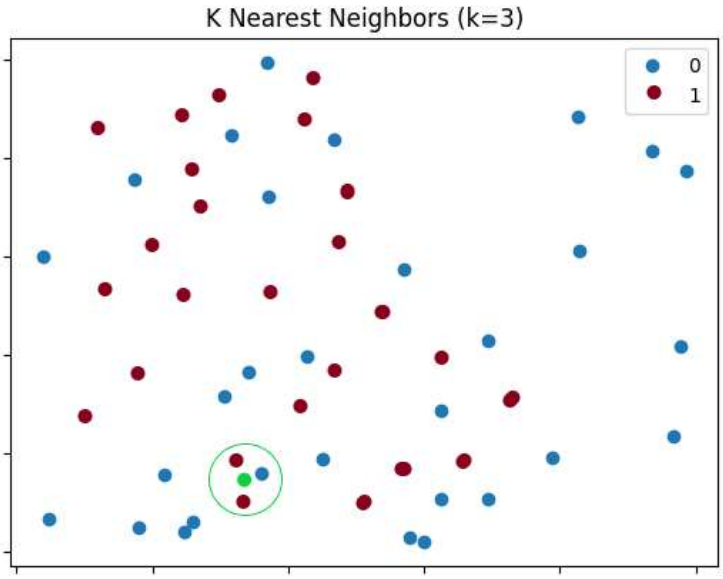
5.2 Model Candidates for Supervised Learning

Supervised machine learning is the process of training a predictive model using historical data. The first step is to gather data and label it for use. In this case, this will involve downloading stock price movements from sources like Yahoo Finance and modifying the outputs to be 1 for when the price change is positive and 0 for when the price change is negative. Next, the model will also require gathering the data for the proxies and linking that data with its respective output. For example, one record can be represented as a tuple (y = 1, x1 = 17,695, x2 = 42,951, x3 = 2.35). The tuples can be stored in a data frame with n rows corresponding to the number of entries in the training data. The next step is to define the candidate set (H). This encompasses the possible relationships between the inputs and the output. The more restrictive the candidate set, the less flexible the model is to fit the data, and the more error is induced. However, a candidate set that allows for complex relationships may capture unwanted noise in

the data. After the candidate set is defined, an algorithm (A) is chosen to systematically determine the best-fitting model in the candidate set. Like selecting the candidate set, the algorithm is decided on by the scientist and is a source of error. Luckily, due to the ubiquitous nature of binary classification problems, there are many models that can be used. These include the K-nearest neighbor model, the support vector machine (hereinafter SVM) model, and the logistic regression model.

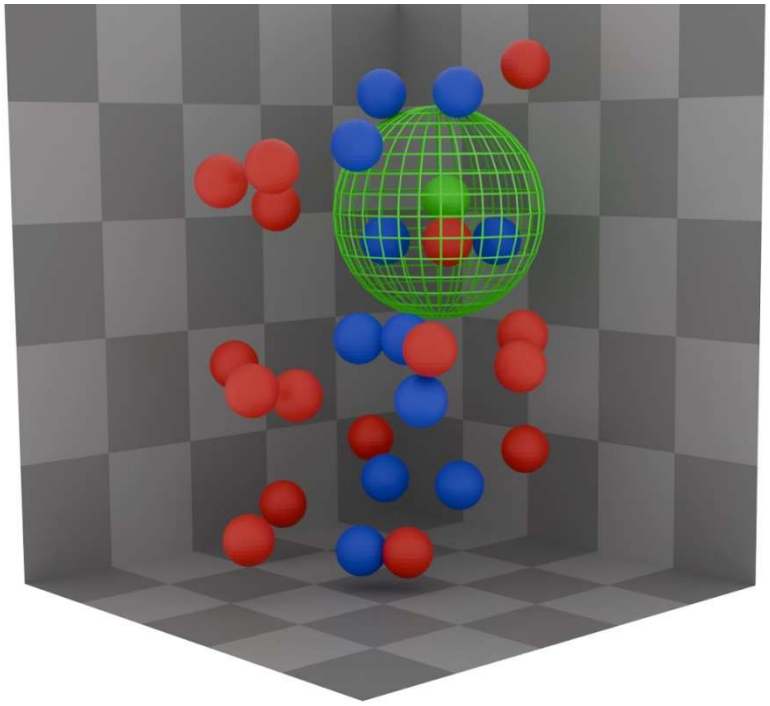
5.2 K-Nearest Neighbors Model

The K-nearest neighbor model finds the most similar historical instances to the situation at hand. It uses the most common outcome from these "neighbors" as the prediction for the current event. The most similar instances are determined by finding the points closest in distance. Using Euclidean distance in two dimensions, this process can be pictured with a circle.



In the figure, the model is making a prediction on the green data point. The predicted output will be zero as the majority of the k neighbors are 0. The parameter k=3 was chosen by the scientist

and may yield different results from k=4 or k=5. In three dimensions, the closest neighbors can be pictured using a sphere.

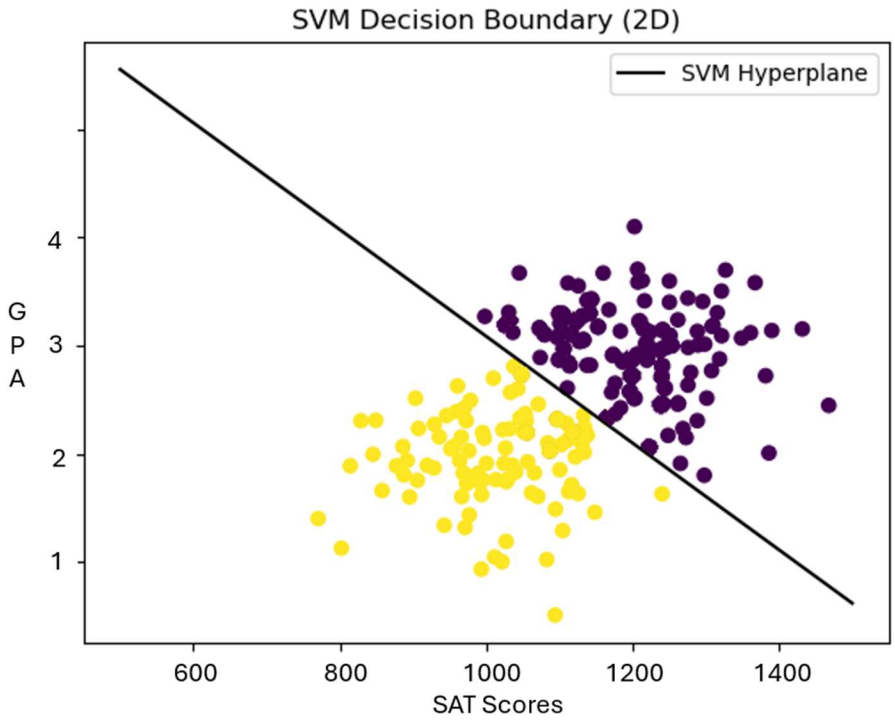


This model, albeit simple to understand, is not best suited for our application. This is because it must sort through the data to find the nearest neighbors, a time intensive process. In order to take full advantage of the market opportunity using HFT, speed is key.[6]

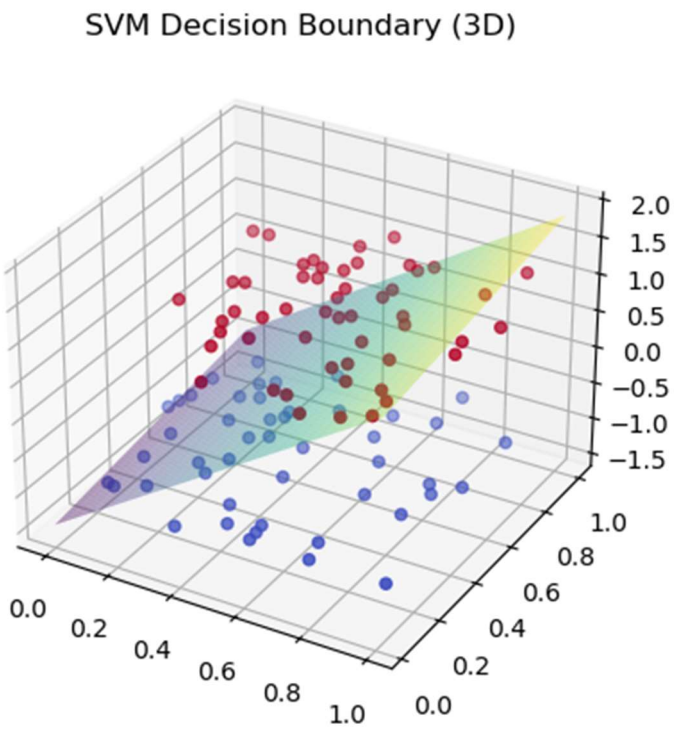
5.3 The SVM Model

The SVM model is used to classify p-dimensional data by separating past data using a (p-1)-dimensional hyperplane. This is most easily understood in an example using two-dimensional data. For example, assume college acceptance is two-dimensional and determined solely by GPA and SAT score. Each dimension of the data, also called a feature, can be plotted on an axis, forming a two-dimensional plane. If the data is linearly separable, a one-dimensional line can be drawn that will bisect the applications, forming a cluster of acceptances on one side and a cluster of rejections on the other. If a future application falls above the line, it is predicted

to be an acceptance, and vice versa. When many possible lines can be drawn, the SVM attempts to create the best-separating line, maximizing the margin between the closest points (the support vector) of each cluster.



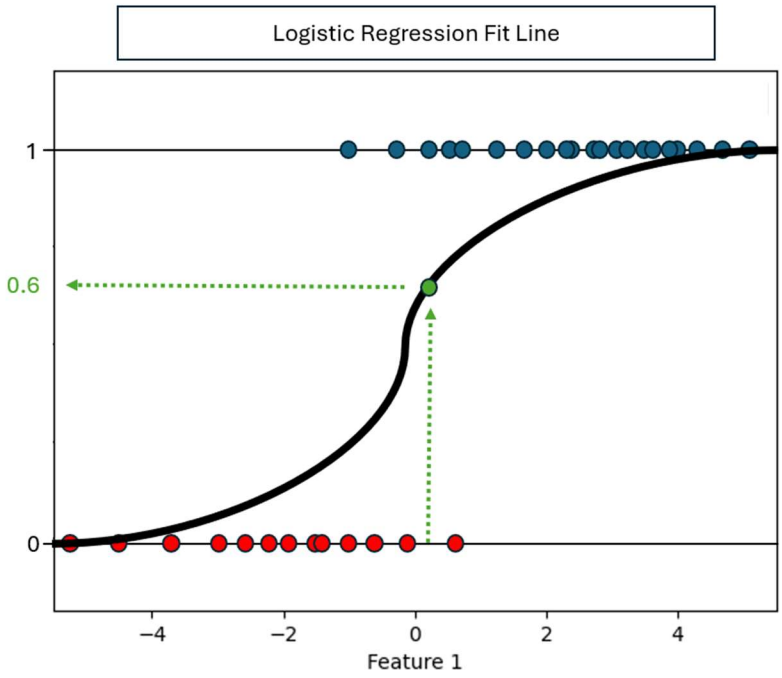
With three features, the separating hyperplane is a 2-dimensional plane. Here, the two classification regions are above and below the plane.



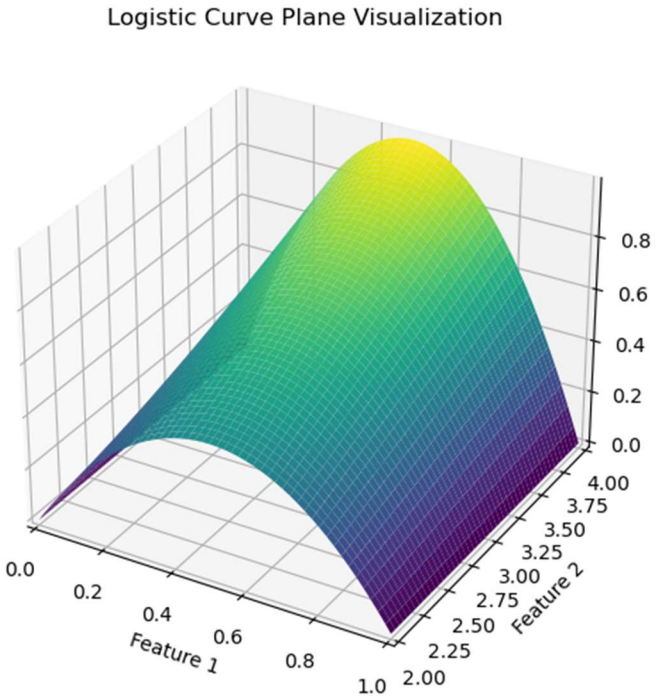
While this approach may work for our use case, this model has its downsides. Firstly, the output is simply a prediction of 1 or 0. There is no indication of how much conviction the model has in its prediction. Secondly, this approach does not show the importance of each proxy in the decision-making process. These are the main advantages of using the logistic regression model.

5.4 The Logistic Regression model

The logistic regression model is a function which forms a sigmoid through a data set of binary outputs. This is done by using the logistic continuous distribution function. Like any other continuous distribution function, the parameter space ranges from 0 to 1. This curve is then used to forecast the outcomes of new inputs. For example, if a new data point falls under where the curve is equal to 0.6, there is a predicted 60% chance of that input resulting in a 1.



Likewise, with two features, the logistic regression can be visualized in 3 dimensions with a 3-dimensional sigmoid logistic curve. Here, the height is the percentage prediction of an output of 1.



To simplify the resulting percentages for model validation, the output will be redefined as decision $\{\hat{y} = \mathbf{1}(\hat{y} > 0.5)\}$. This means that an output above 0.5 is considered a prediction of 1. However, the raw output may have some value in informing decision-making. For example, a more risk-averse trader may only buy when the original prediction is greater than 0.6 and sell when it is below 0.4.

5.4 Proxy selection

The following are possible proxies for supply and demand to be used to predict stock price movements for any individual stock using its order flow information. These proxies are loosely inspired by the square root law of market impact. [7]

Title	Symbol	Formula	Explanation
Order size	x_1	$\sum B_m - \sum S_m$	Difference between market buy orders and sell orders in the period preceding prediction.
Volume	x_2	$\frac{1}{\lambda_1}\sum_{i=1}^{\lambda_1} v_i$	Average volume over training period.
Volatility	x_3	$\frac{\sum_{i=1}^{\lambda 2} \frac{p_i - p_{i-1}}{p_{i-1}}}{\lambda_2}$	Average percent change between time periods.

Here is a possible scenario where these factors can be used to make predictions. If a stock sees a spike in volume and price relative to its usual volatility and the majority of orders are buys, this could indicate strong sentiment in the short term.

6. Experiment

6 Data Availability

Unfortunately, it is not possible to isolate trades made exclusively from PFOF sources. However, order flow data from the entire market is both recorded and easily attainable from the New York Stock Exchange website. This data is recorded on a tick-by-tick basis, encompassing every trade and bid-ask spread that occurred over that time period. Because the initiator of a trade cannot be isolated, x_1 will be the total volume for that time period. For example, x_2 and x_3 can be determined by a model trained on the past 5 days, while x_1 is determined by the current day's volume. The model will then predict the directional movement over the next day. Market volume will be used as a substitution of x1 on the assumption that PFOF order flow data is a simple random sample of all market activity.

6 Simple Random Sampling

When inferring about a population, it is best practice to learn from a randomly selected sample of the data. This is done by giving each data point in the population an equal opportunity to be selected. If the data is independently and identically generated, as it is in our null hypothesis, this would cause the characteristics of the random sample group to better match the characteristics of the population group. This assumption allows for better statistical inference, including the creation of Z-tests and confidence intervals, as will be discussed later.

6.1 Choosing Hyperparameters

For this model, there are two timeframes that need to be defined. They are as follows:

Title	Symbol	Explanation
Training period	λ_1	How much data the model is exposed to prior to prediction.
Time horizon	λ_2	How far into the future the model considers.

For use in HFT, λ_2 will be a fraction of a second, but λ_1 needs to be selected. This can be done through trial and error by testing a few different values and selecting the one that works best. For our practical experimentation, I will define λ_1 as one of the following {1, 5, 10, 15} days. λ_2 will be defined as an investment time horizon of 1 day.

7. Interpreting the Results

7 The Central Limit Theorem

The central limit theorem states that the means of realizations from any distribution will converge towards a normal distribution. In our case, even though the null hypothesis predicts outputs generated from a Bernoulli distribution, the mean of these outputs is expected to be normally distributed. This theorem underlies many core concepts of statistical inference, including the construction of confidence intervals.

7.1 Constructing a Confidence Interval

According to the null hypothesis, the resulting directional movements of stocks are n independent realizations from a Bernoulli distribution with p = 0.5. This means we can employ the central limit theorem to construct a 95% confidence interval for the number of predictions made. Then, because the mean of the data is assumed to be normally distributed, a Z-test can be employed. The one-tailed retainment region at an alpha of 5% will come out to $\{n * 0.5 \pm 1.645 * \sigma\}$ where $\{\sigma = \sqrt{(n * p * (1 - p))} = \sqrt{(0.25n)}\}$. For example, if we predict the movement of 100 stock prices, if 59 or more stocks are predicted correctly, then we can reject the null hypothesis.

7.2 Sources of Error

This model suffers from many sources of error. These include ignorance, misspecification, and estimation errors. The ignorance error is caused by substituting proxies for the real causal drivers. The ignorance error in this case is caused by estimating the partially quantifiable supply and demand. The inclusion of more relevant proxies could reduce ignorance error but may incur overfitting. The misspecification error is caused by the reduction of the candidate space. The larger the candidate space, the fewer misspecification errors are expected to occur. The estimation error is caused by the algorithm not being able to perfectly choose the best candidate available. This happens because we are unable to know the true characteristics of the generating process. Instead, we can only infer them from analyzing the training data using statistical inference. The model will also have its interpretation changed due to all of its assumptions.

7.3 Assumptions

There are many assumptions underlying this model. The biggest assumption is that PFOF orders are a simple random sample of the total market order flow. However, this assumption is most likely false. PFOF services retail traders, which are estimated to be around only 20% of the overall market. This group of traders, mockingly called “dumb money,” almost certainly selects stocks differently than institutional investors. Secondly, this model assumes that this information has not already been priced in. There are more than one market maker that executes trades from PFOF. The data from competing market makers may have already been used to purchase or sell stocks, thus moving the stock price and eliminating any opportunity to profit. The third major assumption is that past data has predictive power over future movements. It may be the case that the market is dynamic, even in the short term, and stock prices move for reasons that keep rapidly changing.

8 Conclusion

8.1 Model Usefulness

To be useful, the model must contend with three major hurdles. These are being able to reject the null hypothesis, being practically implementable, and being profitable despite the sources of error. I do believe that with the right proxies and hyperparameters, the null model can be rejected. As already outlined above in the model selection section, I believe this model to be practically implementable. What I am least convinced about is that this model will be profitable. I believe that the assumption that PFOF orders are simple random samples of market orders to be too heavy to overcome. I believe the difference between retail and institutional investors is too great. To be useful, this model should be utilized as one cog in a complex trading strategy. This is because this model may be able to predict trading momentum of retail traders. This prediction, paired with models that predict trading momentum of institutional investors, may reveal some hidden value in PFOF order flow data.

References

[1] K. Warren, “Who is Keith Gill? the man behind roaring kitty”, Investopedia. (n.d.). https://www.investopedia.com/keith-gill-roaring-kitty-8303143.

[2] Mohamed, Theron. “The Firms of Billionaire Investors Steve Cohen and Ken Griffin Pour $2.8 Billion into a GameStop Short-Seller That’s Lost 30% This Year.” *Markets Insider*, markets.businessinsider.com/news/stocks/steve-cohen-ken-griffin-invest-3-billion-gamestop-short-seller-2021-1-1030003305.

[3] “Capital Market Professional Perspective on Payment for Order Flow.” *Bloomberglaw.com*, 2023, www.bloomberglaw.com/external/document/X1RP679S000000/capital-markets-professional-perspective-payment-for-order-flow-.

[4] Neufeld, Dorothy. ”Success rate of actively managed funds” visual capitalist.com, 2023, https://advisor.visualcapitalist.com/success-rate-of-actively-managed-funds/.

[5] Beschwitz, Klein, Massa. “Media-driven High Frequency Trading”, conference.nber.org, 2013, https://conference.nber.org/confer/2013/MMf13/von_Beschwitz_Keim_Massa.pdf.

[6] Loveless, Stoikov, Waeber. “Online Algorithms in High Frequency Trading” acm.org, 2013 https://dl.acm.org/doi/fullHtml/10.1145/2523426.2534976#:~:text=HFT%20firms%20are%20able%20to,book%E2%80%94and%20react%20within%20microseconds.

[7] Almgren, Thum, Hauptmann, Li “Direct Estimation of Equity Market Impact”, Upenn.edu, 2005, https://www.cis.upenn.edu/~mkearns/finread/costestim.pdf.