SMU Data Science Review

Volume 7 | Number 1 Article 7

Following the Crowd: Beginners Investors Guide to the Options Market

Jeremy Dawkins
Southern Methodist University, jeremydawkins80@gmail.com

Alexy Morris
Southern Methodist University, alexym@mail.smu.edu

Jacob Gipson
Southern Methodist University, gipsonj@mail.smu.edu

Masoud Valizadeh
Northern Illinois University, mvalizadeh@niu.edu

Following the Crowd: Beginners Investors Guide to the Options Market

Alexy Morris, Jacob Gipson[2], Jeremy Dawkin[3], Masoud Valizadeh[4]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{jeremyd, alexym, gipsonj}@smu.edu
{mvalizadeh@niu.edu}

Abstract. While the options market may be intimidating for a beginner, having the right tools can help improve the outcome of their investments. This project aims to develop a tool that uses time-series analysis and forecasting to model the future demand of S&P 500 and AAPL options contracts. The open interest of these contracts will be analyzed using various models such as AR, ARIMA, Neural Networks, and VAR, along with the put-call ratio. The goal is not to make buy or sell recommendations, but alert the user when money is flowing into a security or index. Of all the models, the use of the ARMA model provides the best results for predicting the open interest in contracts for these specific symbols.

1 Introduction

The world of financial opportunity has made way for individuals to take part and make their fortune through investments, real estate, stock markets, options markets, and many other ventures. Many individuals claim that participation in these ventures requires wealth and success. However, this is not entirely true. Having the market know-how can turn a small sum of money into a sizable profit with big rewards. This holds true because everyday investors' awareness of their money helps them process any potential risks and their possible yield results. When individuals can invest their hard-earned money, diversifying their portfolio, it is safe to assume that they know what they are doing. Why? Because of their experience in the financial market, consistent investors can expand their portfolios. Future investors who want to invest in the market are often left out, missing earnings that could help them.

One investment article state that "55% of Americans aren't investing…" (Anderson, 2019, p. 1) because of investing ignorance, impatience with seeing a profit, and fear. These factors deter individuals from placing their money somewhere risky while waiting for it to grow. Most of the time, the easiest way to enter the world of investing is either through the stock market, equity securities like Google and Amazon, or indexes such as S&P 500 or DOW Jones. A minuscule amount of money has the potential for gain within a certain period. When it comes to stock markets, one must know what they are doing, who to invest in, when to invest, and how to keep up with the current conditions of the American economy. There are even strategies used

A warning and disclaimer to anyone expecting guaranteed rewards: this study does not guarantee that someone will become wealthy from investing, but it will offer a financial education tool to teach them how to invest their money.

by investment firms or artificial intelligence advisors where they control or advise individuals on which securities to invest in. An issue for someone with extra money set aside in their savings account is that investing in the normal stock market takes too long. Thus, many Americans who work normal jobs question why they should invest in something that would take too long to see rewards.

Another issue preventing individuals from investing is a lack of education. Financial Literacy is a plague for Americans only because there are few readily available informational resources. American schools do not traditionally focus their curriculum on finance. (Camberato, 2022, p. 1) Topics such as saving money, calculating and filing taxes, or investing are left out of early education. When young Americans attend college, many stay unaware of financial basics unless they are business majors or decide to take extracurricular courses in business or finance. Unfortunately, the gap between participation in this financially driven world and comprehension of financial basics is growing. For example, when an individual gets their first job and starts their career, one of the first onboarding tasks requires setting up individual retirement accounts such as the Roth IRAs or 401ks. Many employees complete these tasks with little understanding of personal investing; leading them to teach themselves via the internet, friends, or other sources. One might mention that the internet is accessible to all. However, it is difficult to scour through the time-consuming and often complex subjects. The use of brokerage or investment firms serves their purpose of helping individuals invest in the market. Still, with the many vendors to choose from, the selection process can be overwhelming for some. Another fear contributing to the lack of knowledge on investing is the mindset of losing money due to a bear market or economic downturn. For example, during the 2007-08 financial crisis, the American market system took a downturn due to the housing bubble's burst and mortgage crises. Simultaneously this resulted in people losing their jobs, banks being in financial ruin, and even companies going bankrupt. This led many to depend on individual savings or other income streams because the market tanked, and Americans lost their money. Potential investors fear this repeating and happening to them, but the fear lives mainly with the uninformed. Their lack of skills in this specific area hinders them from making any money.

Despite these concerns from the public, some still consider venturing into investments to expand their income streams. This research focuses on targeting people interested in investing in the market but lack the market knowledge to effectively participate. It also is focused on people who do not have time to consider these strategies deeply. It is not openly discussed that a great beginner market to tap into for beginner investors is the options market. The options market can be risky but can supply multiple advantages over the typical stock market, leading to less initially invested money and potentially greater gains. Bearing this in mind, this research aims to provide the fundamentals of the options market and what can be done with an option contract. An investor would not need to know the intricate details of the options market besides the basic knowledge provided in this research paper. The options market supplies many opportunities in which one can make money quickly compared to the stock market in a short amount of time. Allowing one to predict

whether security will go up or down based on the market and a set date, one could see profit no matter how big or small.

To further discuss the detailed part of this research paper, one must know what a choice is and what calls and puts are. "Options are contracts that give the bearer the right - but not the obligation to either buy or sell an amount of some underlying asset at a predetermined price at or before the contract expires." (Downey, 2022, p. 1) Many brokerage firms and investment banks offer these to individual investors in a manner like the stock market. Many retail investors would say that people should spread out their portfolio with some investments in the stock market and some in the options market to help with offsetting losses that one can gain, this is called hedging. One can hedge with options to reduce the amount of risk they take at a reasonable cost to the investor. The option contracts allow the investor to predict future price events of securities. They supply two key types of options contracts. One is the *call option* which "gives the holder the right, but not the obligation, to buy the underlying security at the strike price on or before expiration. A call option will become more valuable as the underlying security rises in price" (Downey, 2022, p.1). An example of this is allowing one to bet on if the price of Apple will go up soon before the set expiration date of the contract. The other important contract choice is the *put option* which is "Opposite to call options. A put gives the holder the right, but not the obligation, to instead sell the underlying stock at the strike price on or before expiration" (Downey, 2022, p.1). Continuing with the previous example, the put gains value as the underlying price falls for Apple. Investors and corporations make billions of dollars annually just from predicting price moves throughout the year on different securities.

This research will serve as an opportunity to bridge the gap between novice retail investors and experts in investing in the options market. Many scholars who supply insights on the stock market or options market only present information to people who are already knowledgeable on these topics. This allows them to keep expert investors well-informed and to create strategies they could use to beat or outperform the market. No techniques will be provided for predicting the price of a stock or evaluating options but instead, a toolkit that could be used to help novice investors with when to step into the market. The toolkit supplies an opportunity and freedom for any investor who wants to continue learning about the options market. This study will use data science and machine learning to discuss investment opportunities for corporations and experts. Examining open interest for these specific securities over time will create a time series discussion about timing and opportunities novices can invest in.

This research intends to inform that a non-expert investor can predict open interest in any security at any time. Can a time series model forecast open interest for any type of security using exploratory data analysis and research? By supplying an opportunity for evidence to support this hypothesis, one can introduce novice retail investors into a new world of diversifying one's portfolios or streams of income. This research will be valuable to the field by extending its findings to leading experts, investment firms, or brokerages to pour resources into advertising for more

consumers. This could also present a toolkit to be used as applications, or apps, developed by firms who wish to expand accessibility to novice investors and inform them on how the options industry works. Doing this makes way for entry into the world of finance without significant knowledge of options but increased confidence in one's ability to invest money.

2 Literature Review

2.1 Financial Literacy

Financial Literacy is a global problem that can seem far-off and intangible. For this paper, Financial Literacy can be defined as the ability to use financial knowledge to obtain a specific money-related goal. For example, opening a high-yield savings account to eventually pay for a predictable high-cost expense, including a house, car, further education, and more. There are many ways to quantify financial Literacy, with one of the most prevalent being the Big 3, a series of questions created by Lusardi and Mitchell (2014) in their research on financial Literacy in a global setting:

> 1. Suppose you had $100 in a savings account and the interest rate was 2% per year. After five years, how much do you think you would have in the account if you left the money to grow?
>
> A. More than $102
> B. Exactly $102
> C. Less than $102
> D. Do not know
> E. Refuse to answer
>
> 2. Imagine that the interest rate on your savings account was 1% per year, and inflation was 2 % per year. After one year, how much would you be able to buy with the money in this account?
>
> A. More than today
> B. Exactly the same
> C. Less than today
> D. Do not know
> E. Refuse to answer
>
> 3. Please tell me whether this statement is true or false. 'Buying a single company's stock usually provides a safer return than a stock mutual fund.'
>
> A. True
> B. False
> C. Do not know
> D. Refuse to answer. (p. 499)

These questions serve as a benchmark to determine an individual's understanding of interest rates, inflation, and stocks. The answer choices for each question allow us to easily categorize an individual into truly literate, moderately literate, and others for those who withdrew from every question.

One of the biggest examples of the need for Financial Literacy is how retirement plans have changed in recent years. "In many countries, employer-sponsored defined benefit (DB) pension plans are swiftly giving way to private defined contribution (DC) plans, shifting the responsibility for retirement saving and investing from employers to employees" (Lusardi, 2019, p. 1). In addition, the commonality of targeted at younger citizens has increased tremendously; student loans and short-term loan services have become much more easily accessible. Although short-term loan services can be avoided, student loans have become a standard for those attending further education. Unfortunately, this massive shift has not been met with an equal increase in financial Literacy. This shows that there is ample space and a definite need to create more tools and resources to further the average person's understanding of and comfort with all forms of finances, including but certainly not limited to the stock market.

Thus, Financial Literacy experiences wide-scale boosts, and tools that help these endeavors along must be vital. In this day and age, "Financial literacy can be seen as an investment in human capital" (Amagir et al. l, 2017, p. 57), and such investments lead to advancements in the financial field.

2.2 Individual Investors

The stock market is a slightly more tangible representation of a country's economic status, which means that one could argue that outside extreme circumstances a given company's stocks will remain constant quarterly. However, this assessment is routinely proven false by the simple fact that human involvement is present in every aspect of the marketplace. Investors can affect the market in a variety of ways; one of the most common is done by what are called *noise traders.* These are individuals taking part in the stock market without company backing or elevated levels of experience. These traders are sensitive to outside influences, like the weather, news, races, etc. Thus, it is possible to check these outside influences and gather them up into a *fear index* that can predict when these noise traders are likely to sell off riskier assets and the following, temporary, stock lowering. That said while the effect on the stock market is always resolved around 20 days (about 3 weeks) after there is little data showing that these same investors are coming back with the same consistency (Kostopoulos et al, 2020). Looking at more experienced investors, usually with company support, there is a higher degree of familiarity with the market that leads to increased returns. Specifically, investors will typically see a higher and higher profit margin as they grow in the industry. It is said that investors will see around a 20% increase in profits from early portfolios to more recent ones (Nicolosi et al, 2009, p. 5). From this standpoint a machine learning model leveraging the knowledge of these experienced investors would show similar profits and on average they did. Of course, the limitations here are the need for a steady stream of data sourced from these individuals (Pagliaro et al, 2021, p. 3). To get around this one can instead forget the investor insights and create a model purely using market data both old and new. By focusing on a single index, for instance the S&P 500, one can easily compare a variety of models and thus decide the *best* strategy for predicting the stock market.

The *best* is a supervised time series model (Soni et al, 2022, p.4). This paper will continue from the ideas and strategy used here to create a more ideal solution. In this paper, the focus will be on using the knowledge of novice and experienced investors in more easily understood time series models to give everyday noise traders a solid ground into more intricate stock strategies.

2.2 Using Machine Learning Techniques to Generate Returns in the Stock Market

From historical data to 8-k filings, never has there been a time where more information has been available to the average retail investor. With increasingly widely available and real-time financial data, various machine-learning techniques have been used to improve profits in the stock market. Sakarwala created a predictive model on whether a stock moved up, down, or sideways based on the content of a corresponding 8-k filing. Twenty thousand filings of S&P 500 companies were pulled from the SEC (Securities and Exchange Commission) EDGAR website and were combined with corresponding historical price data. The text was preprocessed, and The Stanford NLP Wikipedia 2014 + Gigaword 100 dimensions were chosen for pre-trained word embeddings under the assumption that it would carry information for specialized, industry-specific words found in the texts since it was trained from the Wikipedia corpus (Sakarwala et al., 2019, p.11). They then compared various deep learning techniques (Multi-Layer Perceptron, Convolutional Neural Networks, and Recurrent Neural Networks) in their ability to predict the stock's price action based on the contents of the 8-k. They found that RNN (Recurrent Neural Networks) and RNN-CNN performed the best. Now specifically about the options market, many studies have been completed about using machine learning to perfect an options trading strategy. Žabčić-Matić's paper looked at the use of neural networks to predict whether the price of a stock will move, within the next trading day, by more or less than the cost of its associated straddle option spread, with the expiry date set to be the next trading day (Žabčić-Matić, 2019, p. 1). The stocks that were studied were the following companies: Apple, IBM, Microsoft, Intel, Cisco, Qualcomm, Walmart, General Electric, Goldman Sachs, and American Express. Many features were used along with obvious ones like open, close, high, and low prices and daily volumes like on-balance volume, accumulation/distribution line, and MACD. The results were alarmingly good. The models did produce consistent profits and only very rarely incurred a loss (Žabčić-Matić, 2019, p. 1). This shows that machine learning can lead to profitable outcomes for the user, even in a volatile environment like the options market.

Zeng explored using machine learning to predict the best timing of exercise options. Specifically, they used a Q-learning algorithm to consider the early exercise opportunity of American possibilities as a potential profit-generating source (The Q-learning algorithm has two stages, an optimization, and an evaluation stage. In the optimization stage, an iterative progressive hedging algorithm is used to find all options' weights and exercise time at each time, where the Q-values are approximated by regression (Zeng, 2017, p. 50). In the evaluation stage, real-time trading is used to refine when to exercise the options for each simulated sample path (trajectory) given

the weights from the optimization stage (Zeng, 2017, p. 54). In one of the evaluation algorithms, they changed the Least Square Monte Carlo algorithm to evaluate their performance in discovering the exercise time. Their model was successful because the algorithms outperformed the LSMC about the utility gap from best and certainty equivalent of return. This further proves that machine learning is beneficial in creating profit in the options market.

Not only in the United States market that is being predicted but similar machine learning algorithms have been used to predict across the world. Using machine learning, Zhang and company studied market trends and prices in the Chinese stock market. (Zhang et al., 2018, p. 54). They analyzed the Shanghai Stock Exchange (SSE) Index stocks to decide to create a model to predict profitable patterns. Techniques such as support vector machines, random forest, Naïve Bayes classifier, and neural networks are compared to decide which model predicted future data using the scoring metric of accuracy. From their conclusion, the neural network outperformed all the other models in considering patterns in the index compared to other deep learning techniques.

2.3 ARIMA/ML Forecasting Modeling

Investment managers are always looking forward to the options market. That way, they are prepared and can perfect a strategy to make money regardless of the current market. In the aftermath of the 2008 financial crisis data, the Volatility Index (VIX) outlook was trending downwards. An S&P 500 index options database involved 43,014 calls and 38,784 put premiums for 81,798 options traded on the CME (Rostan et al., 2020). The calls and puts were collected over 60 months from January 2009 through December 2013. The strategy was to examine how an autoregressive integrated moving average (ARIMA) model can be used to forecast calls and put options with a comparison of the typical generalized autoregressive conditional heteroskedasticity (GARCH) model. This paper only focused on European call-and-put options contracts. After careful evaluation and exploratory analysis of the contracts, the ARIMA model that proved to have impressive results was the ARIMA forecasting model. When looking at the data, the researchers broke the data into two pieces of calls and put and used ARIMA modeling techniques on them. From the discovery, they could generate forecasts of the S&P 500 Composite Index obtained from the ARIMA (1,1,1) model. When this is done, an options strategy is used, allowing the investor to decide which direction they would like to go depending on their call or selected choice. The ARIMA forecasting results in a being a much better-produced result compared to the forecasting using the GARCH (1,1) model. The authors assessed that there is more profit with call than put options when the economy is in a bear market. They were able to see the intrinsic value of puts decrease (Rostan et al., 2020, p. 4). Their research used the ARIMA model for forecasting calls and put options to assess if dollar profit can be made based on undervalued or overvalued performances. This strategy can be used today and improve on further research in a bear market and discuss the feature importance of sentimental analysis. Alotaibi discovered that an ARIMA (0,1,1) contributed to forecasting prediction returns for the stock market (Alotaibi, 2022, p. 7). They examined the Willis Towers Watson

company spanning from 2010 to 2016 and compared multiple linear regression and ARIMA. With the S&P 500 Index being a popular security to examine, Sun studied this index with an ARIMA and ETS model close price (Sun, 2020, p. 10). An ARIMA (2,1,3) is the best model fitted using the year 2005 to 2014, which succeeded in having a success of short-term predictions. There is the belief that even this current model can be improved using deep learning techniques or hybrid models when modeling the S&P 500 Index close prices.

Hedge fund managers examine VIX Futures due to the forecasting of predicting market volatility around a certain time. This allows the managers to set up their potential strategies to attack the market by investing. Hosker looks at the VIX Futures forecasting using machine learning, comparing three existing financial models to six machine/deep learning supervised regression methods. These six models generated compared to the financial models will be analyzed and determined to present the highest results showing lower mean absolute error (MAE), higher explained variance, and higher correlation (Hosker, 2019, p. 3). To decide the best model, an accuracy matrix was generated for every model displaying better overall accuracy.

2.4 Time Series Analytics and Deep Reinforcement Learning in Portfolio Management

With so many people and corporations investing in the stock market, they hold various securities in their portfolios that require management. Some would consider stocks, options, ETFs, and bonds serious investments. For someone to see substantial profit with their portfolio, managing these accounts requires a good knowledge of markets and periodic adjustment of assets in the account (Chavin et al., 2021, p. 2). With public market exchanges, individuals who want to invest in the market and have limited assets are not afforded the same investment guidance level as those with wealth managers supplying them with advice. People use many portfolios and investment advisor tools to try to close the gap in smarter investing. With these advisor tools being automated to an extent, one just needs to put their money in based on personal preference and sit back and watch the artificial intelligence (AI) advisor suggest securities for you automatically. Chavin's paper's sole focus is to use the success of the AI/ML stock portfolio trade management research to date (Chavin et al., 2021, p. 2). The portfolio management advisor implements various data science techniques. This paper focuses on reinforcement learning (RL) algorithms (Q-Learning & Policy Optimization) which use time series analysis. Investment firms are always trying to improve their algorithms, allowing the advisor to make wiser decisions so that the retail investor sees adequate results. Focusing on the RL algorithm in advising allows the algorithm to learn and adjust its behavior while interacting with a live environment. Many factors affect an economic influence that the advisors consider, which is modeled using the Markov Decision Process (MDP). This process uses a mathematical framework that helps with the decision-making of economic influences of policy, war, supply, demand, etc. The data prepared in this paper utilizes the DOW Jones Index using investment options DOW 30-day multi-stock trading, DOW electronic trading fund single stock trading, and lastly, the DOW 30 portfolio rebalancing trading. An analytical approach is taken here, with models

generated using the RL algorithm and presented with results that improve the portfolio managers' advisor strategies. The RL approach generated multiple models that can be averaged out on the dataset, which is considered the ensemble method. They discovered that MDP and RL proved reputable evidence in the stock evidence in the stock trading decision models in selecting which proper investment options to select. The ensemble method produced the highest Sharpe Ratio and cumulative portfolio final balance in their findings. This paper also contains a thorough use of time series that contributed to a part of the selection environment and MDP. With the advisor, this paper could use AI/ML techniques that can supply future aid to retail investors who do not have time to manage their own investments or are not knowledgeable compared to the wealthy and corporations.

This paper will use previous research knowledge to focus on future investors. A time series and multi-layer perceptron model will be used to analyze SPX and AAPL open interest and forecast to predict future interest. Allowing someone to see how open interest is moving throughout the year can give investors insight into what is happening. The aim is to inform and stress the importance of allowing people who want to invest in options an opportunity to do so despite their feelings that there is a higher barrier to entry.

3 Methods

3.1 Background of Models

This research will use several diverse models to describe the open interest data for SPY and AAPL: Autoregressive (AR), Moving Average (MA), ARMA, and ARIMA. Stationary time series tend to be described by AR, MA, and ARMA models, while nonstationary time series tend to be described by ARIMA models. It is safe to predict that the type of model that best describes open interest data would be an ARIMA model, as it is known that open interest is not stationary as the mean depends on the window of time being observed.

The general form of an ARIMA (p, d, q) model is:

$$(1-B)^d\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)X_t=\left(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q\right)a_t \quad (1)$$

$$\text{Were } BX_t=X_{t-1} \text{ and } Ba_t=a_{t-1} \quad (2)$$

Where fp are the autoregressive constants, qq are the moving average constants, Xt is the value of the time series at time t, and at is the white noise component at time t. The values of p and q indicate the number of autoregressive and moving average constants are present in the model.

3.2 Data

The data will be attained from The Options Clearing Corporation's website (theocc.com), which is a clearing house that is the world's largest equity derivatives clearing organization. The OCC provides data about the market integrity for options, futures, and securities lending transactions.

A hypothesis testing will be conducted modeling open interest contract data using the index SPY and the equity security Apple (AAPL). The hypothesis testing will determine if the model is stationary or not using time series examining data provided from the beginning of the year January 4, 2022, which is the first trading day of the year through August 31st, 2022. A model will be created using a time series that will prove the ability to forecast open interest for this upcoming fall of 2022 and winter of 2022. This forecasting will contribute to the ability of individual investors to where money is being placed in the market during certain trading quarters that can be beneficial to the investor.

An evaluation of different ARMA and ARIMA models that will be tested on the data provided and provide detailed analysis with the equations provided. Creating a time series model will be conducted using a five traditional step method, with step one doing data preparation or preprocessing. A full examination of time series decomposition, modeling, forecasting and the last is model evaluation.

3.3 Methods

The dataset that was used was open interest and put/call ratio data dating from 01/26/2021 to 09/22/2022 for both SPY and APPL tickers. Data until 07/13/2022 was used to train our models, while the rest of the data was used to evaluate performance. The cutoff was selected to evaluate predictions 50 trading days out. The primary objective of the project was to model open interest, so the initial approach was to analyze both tickers in a traditional univariate fashion before including the put-call ratio as a potential regressor. This was done to establish the performance of the univariate models as a baseline for comparison with more advanced models later.

Upon initial review of the realization plot of the SPY open interest data, wandering behavior was observed. The lack of constant mean, independent of the section of the time series, suggests that the series is not stationary. On the other hand, the AAPL open interest data appeared to be stationary, with a constant mean and constant variance that were independent of the time period, and the autocorrelation plot showed signs of dampening. Although seasonality was suspected from the autocorrelation plot, it was not confirmed through an overfitting test. For the sake of completeness, both stationary and non-stationary models were explored as a single realization is not always indicative of a variable’s stationarity.

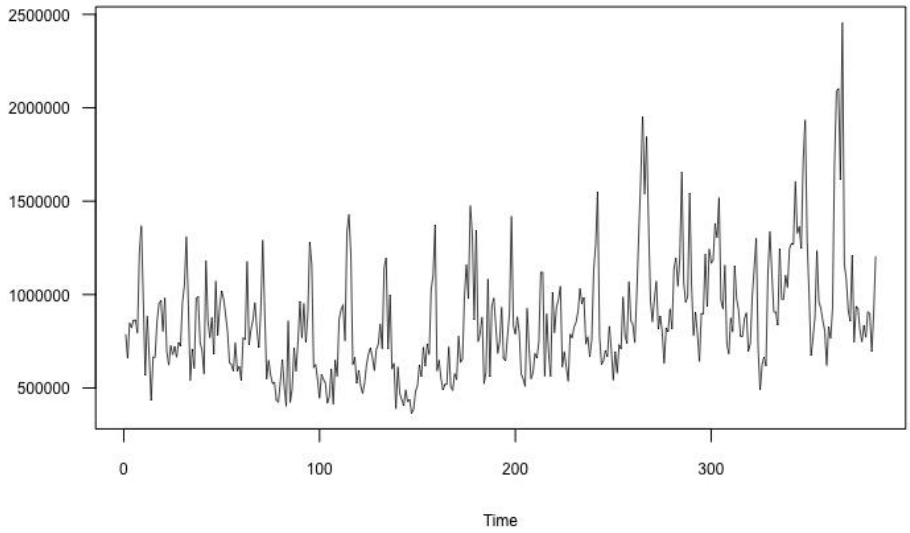

Figure 1: Realization of SPY open interest data from 01/26/2021 to 07/13/2022

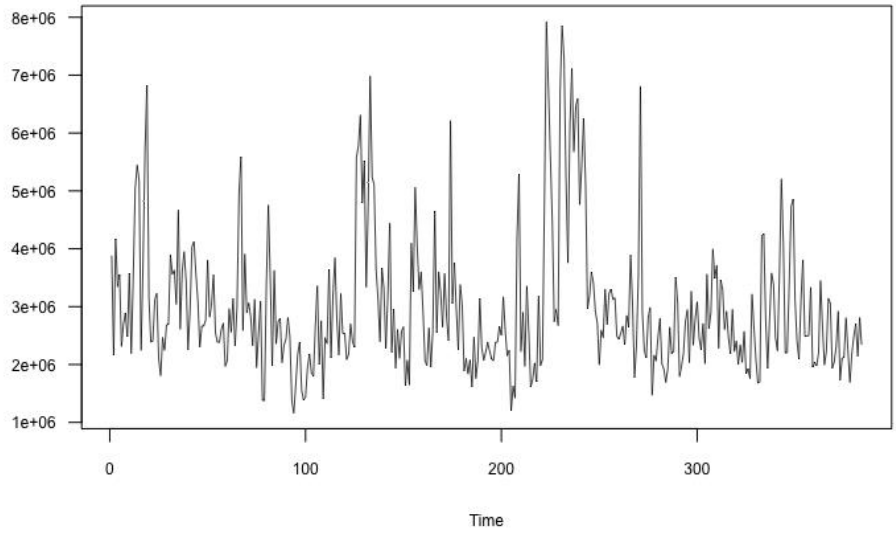

Figure 2: Realization of AAPL open interest data from 01/26/2021 to 07/13/2022

First non-stationary models were explored. The wandering behavior was removed from the data by first differencing. The aic.wge() function from the tswge package was used to estimate the number of Autoregressive (AR) and Moving Average (MA) components in the model. Both AIC and BIC were used as selection criteria to

achieve the best fit while also placing value on model simplicity. For the SPY open interest data, the same 3 models were selected using both criteria: ARIMA (1,1,1), ARIMA (3,1,2), and ARIMA (4,1,1). The same strategy was applied to the AAPL open interest data, and the following models were selected: ARIMA (5,1,6), ARIMA (9,1,5), and ARIMA (1,1,1). The AR and MA components were estimated and forecasted. Model performance was evaluated using 5, 10, and 50-day Mean Squared Error (MSE) values, along with rolling window root mean squared errors (RWRMSE) of the same window lengths.

Open interest data was also modeled using stationary models. A similar strategy was taken except tTe aic.wge() function was used on non-differenced data. For SPY, the AIC and BIC selection criteria favored an AR (2,0) and ARIMA (1,1), with ARMA (1,2) as the only other model appearing on both lists of candidates. For AAPL, the selected models were ARMA (6,8), ARMA (4,7), and AR (5). Forecasts using these models and the same performance metrics were generated and can be seen in the results.

Next, neural network models were fitted and compared to the traditional univariate time series models. Multiple iterations were created for both SPY and AAPL data by specifying lags, differencing order, and different numbers of hidden nodes, as well as different selection strategies. The impact of daily put-call ratio on prediction accuracy was also explored as a possible regressor.

Lastly, the introduction of the Vector Autoregressive (VAR) model is used to investigate the relationship between the put-call ratio and open interest. The model is intended to see how the two variables relate to each other or influence each other in some way. The put-call ratio and open interest were combined to form a vector, which was then modeled to fit an AR model. After determining the selected AR model, the parameters were then used in a VAR model to determine the overall equation, which produced a VAR (1). The two features, open interest, and put-call ratio were given a weighted value by adding them together. From analyzing these two features, the model produced a VAR (1) model, which was then used to forecast using the same horizon as the previous models.

4 Results

SPY Open Interest

Model	5 Day MSE*	10 day MSE*	50 day MSE*	5 day RWRMSE **	10 day RWRMSE **	50 day RWRMSE **	Ljung-Box test p-values	
							K = 24	K = 48

ARIMA (1,1,1)	0.844	1.31	1.04	4.21	4.94	5.42	0.0678	0.0678
ARIMA (4,1,1)	0.987	1.39	1.06	4.33	5.11	5.48	0.274	0.229
ARIMA (3,1,2)	0.844	1.26	1.03	3.70	4.44	4.97	0.071	0.0643
ARMA (1,1)	0.673	0.604	1.24	2.49	2.72	3.01	0.11	0.065
ARMA (1,2)	0.682	0.604	1.24	2.50	2.72	3.02	0.095	0.063
AR (2)	0.678	0.603	1.24	2.50	2.72	3.02	0.12	0.077
Univariate NN	0.631	0.596	1.21	2.49	2.71	3.02	N/A	N/A
Multivariate NN	0.636	0.624	1.17	N/A	N/A	N/A	N/A	N/A

* X 10^{11}
**X 10^{5}

Figure 7: Summary of performance metrics for candidate models for SPY Open Interest

All non-neural network models failed to reject at the 95% confidence level when conducting the Ljung-box test on the residuals, indicating that the models do not show a lack of fit. Only the ARIMA (1, 1, 1), ARIMA (4, 1, 1), and ARMA (1, 2) failed to reject at a 90% confidence level using two different maximum lags (K). The stationary models outperformed the non-stationary candidates significantly in all measures except for the 50-day MSE. In the end, the ARMA (1, 1, 1) model was selected as the top performing univariate model, as it scored the best or second best in all 6 metrics and passed the Ljung-box test at the 95% confidence level. The model can be seen below:

$$(1 - 0.76B)(X_t - 860{,}410) = (1 - 0.17B)a_t \text{ where } a_t = 5.2 \times 10^{10} \quad (3)$$

The top performing neural network (NN) model was made up of 5 hidden nodes and one input, a univariate lag of 1. It can be seen below:

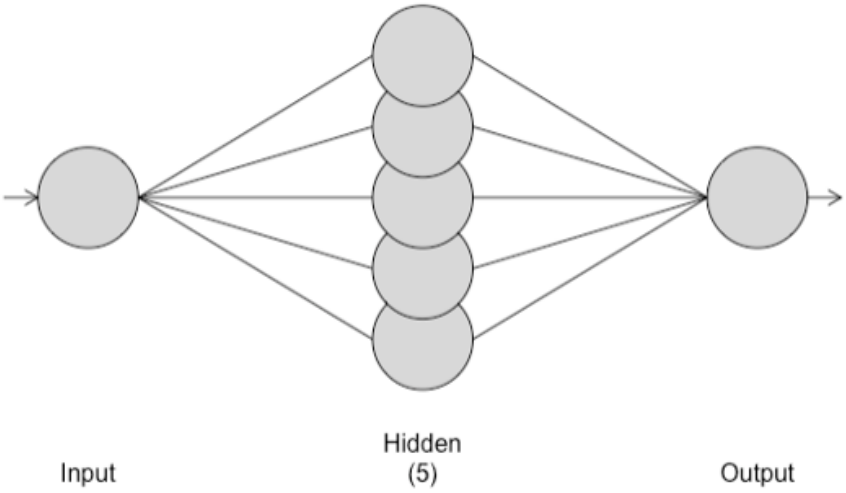

Figure 3: $SPX Multi Layered Perceptron model diagram

It performed similarly or slightly better than the previous AR (2) model in all metrics. Another NN model was trained using the put-call ratio data, but it did not improve performance and, in fact, the 10-day MSE worsened.

AAPL Open Interest

Model	5 day MSE *	10 day MSE *	50 day MSE *	5 day RWRMSE **	10 day RWRMSE** *	50 day RWRMSE **	Ljung-Box test p-values	
							K = 24	K = 48
ARIMA (5,1,6)	3.06	4.77	7.22	1.32	1.57	2.06	0.038	0.35
ARIMA (9,1,5)	2.46	4.80	6.46	1.19	1.39	1.92	0.38	0.72
ARIMA (1,1,1)	4.65	4.79	8.15	1.23	1.37	1.65	0.0032	0.0027
ARMA (6,8)	3.06	7.84	6.49	0.95	1.01	1.14	0.46	0.79
ARMA (4,7)	2.56	7.34	6.21	0.95	1.01	1.15	0.77	0.89
AR (5)	3.52	7.38	6.51	0.89	0.96	1.09	0.70	0.67
Univariate NN	3.27	6.28	6.44	0.87	0.94	1.08	N/A	N/A
Multivariate NN	3.63	4.61	7.30	N/A	N/A	N/A	N/A	N/A

* X 10^{11}
** X 10^{6}

Figure 4: Summary of performance metrics for candidate models for AAPL Open Interest

The Ljung-Box failed to reject all the models at a 95% confidence level except for the ARIMA (1,1,1) model, which was eliminated from consideration. The non-stationary ARIMA models all outperformed the other models in the 5 and 10 day MSE scores, but underperformed in the other 4 metrics. The ARMA (4, 7) model had a very high 10 day MSE score, but the low 5 day MSE score in conjunction with the lowest RWRMSE scores gave us the most confidence in selecting it as the best performing non-neural network model. The equation for the model is shown below:

$$(1 - 0.93B + 0.83B^2 - 0.23B^3 - 0.37B^4)(X_t - 2{,}992{,}078) = (1 - 0.39B + 0.64B^2 + 0.08B^3 - 0.21B^4 - 0.03B^5 - 0.04B^6 + 0.10B^7)a_t \text{where } a_t = 8.1 \times 10^{11} \quad (4)$$

The univariate neural network model that had the best performance had 5 hidden nodes and 2 inputs, univariate lags at 1 and 4. Its performance was an improvement over the AR (5) model in all four metrics. A diagram of the model can be seen below:

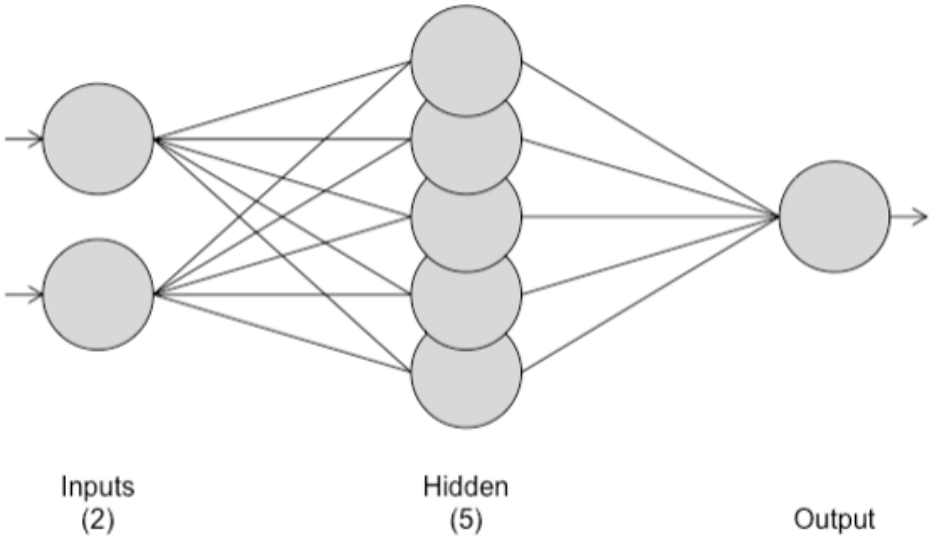

Figure 5: $AAPL Multi Layered Perceptron model

Unlike the SPY neural network model, incorporating put-call ratio as a regressor significantly improved the 10-day MSE, but also increased the 5 and 50 day MSE.

A Vector Auto-Regressive (VAR) model was created for SPY and AAPL put-call ratio. This is a widely used measurement to understand how the equity or index is moving and to predict if there will soon be a bull or bear market. A training and test set is created to train the model and test its forecasting prediction. With the VAR model, the interrelationships between open interest and put-call ratio are considered to

use all variables. The correlation between variables is leveraged to improve fitting and forecasting.

There is a relationship between open interest and put and call options that influences investors to buy or sell certain contracts. A cross-correlation between the put-call ratio and open interest for SPY is shown:

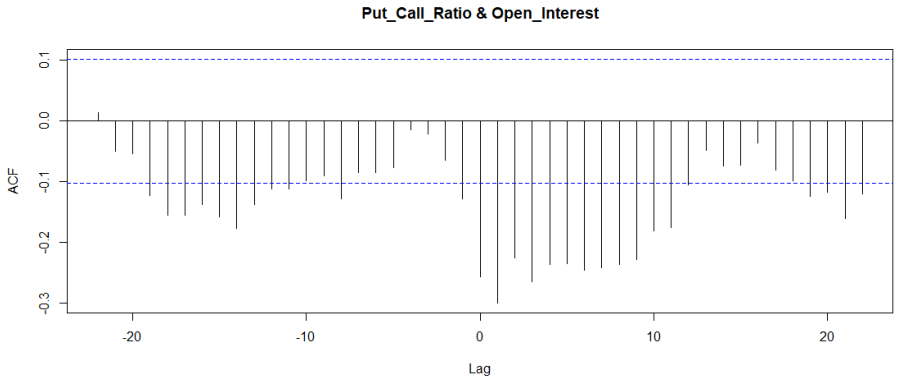

Figure 6: Autocorrelation Function of PCR and Open Interest for SPY

As shown in the graph, the PCR and the open interest of SPY produced the greatest autocorrelation at lag 1, with lag 3 closed behind. The VAR select process, considering a maximum lag is 3, was used. The VAR selects used the AIC, which picked lag 3, and the BIC picked lag 1. To reduce the number of parameters, the BIC selection was used, and the parameter estimation is created by considering the lag at 1. A summary of the VAR model produced the following equation:

$$PCR = OI.L1 + PCR.L1 + Const \quad (5)$$

The forecasting of the last 50 dates of options trading is then used to display how well the model is performing on the test set. The forecast of the VAR (1) model is shown, displaying its performance of future dates.

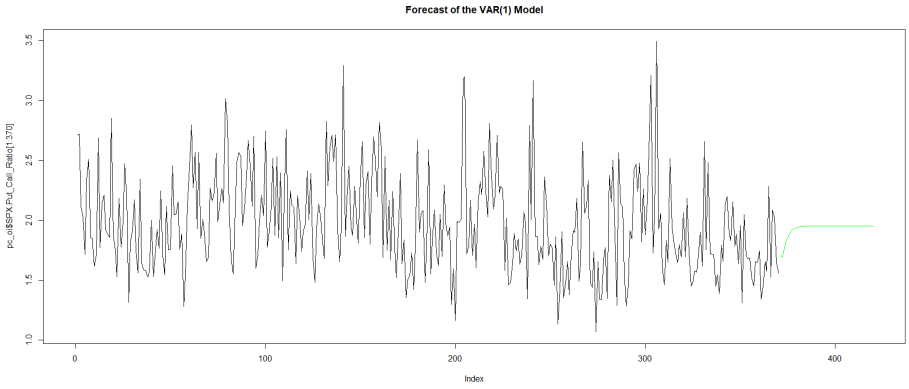

Figure 7: Forecast of the VAR Model for SPY

Then RMSE for the model came out to be around 0.33, which is small and suggests that the model did a decent job at predicting the put-call ratio.

A VAR model was then examined with similar settings but with the AAPL equity. The cross-correlation of the PCR and the open interest was analyzed in this equity. From the chart below, there appear to be significant differences around lag – 2 and around lag 16.

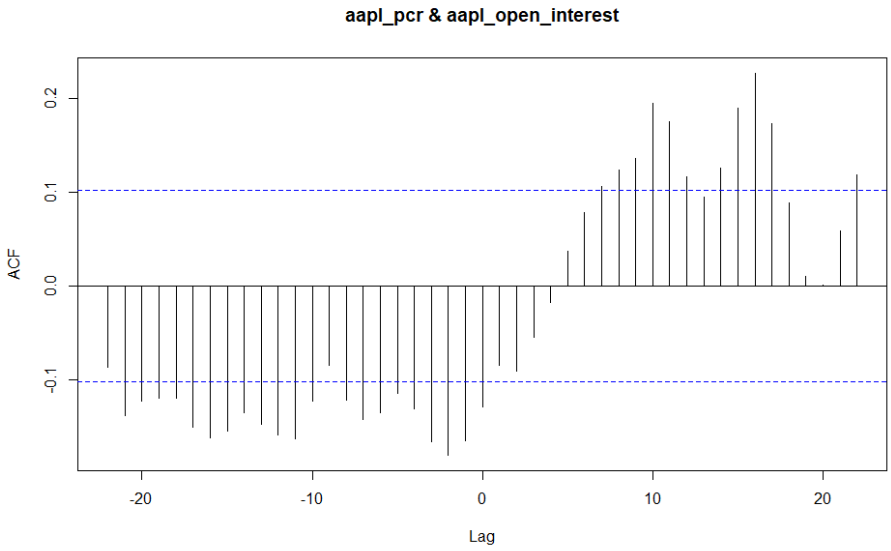

Figure 8: Autocorrelation Function of PCR and Open Interest of AAPL

The maximum lag to capture any of the trading days is lag 20. Considering the increase at lag 16, the AIC selects a lag of 5, and the BIC selects a lag of 1. To minimize the parameters of the model, the BIC selection of lag 1 was used. The equation for the VAR model with lag 1 is shown below:

$$PCR = OI.L1 + PCR.L1 + Const \quad \textbf{(6)}$$

The RMSE from the prediction forecast of the last 50 dates was compared to what the VAR model produced, resulting in an RMSE of 0.38. The testing set of the last 50 dates was close to what was predicted.

A prediction of the last 50 dates is used and plotted with the training data, as shown below:

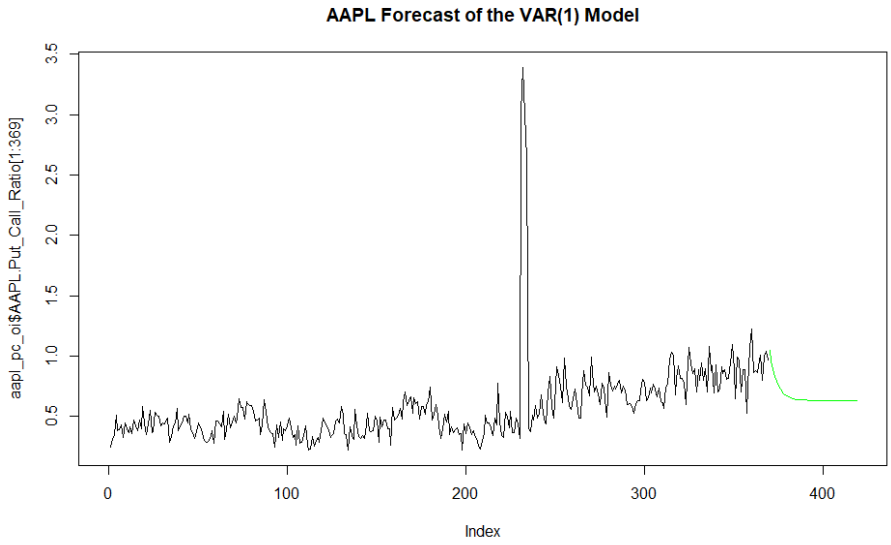

Figure 9: Forecast of the VAR of AAPL

5 Discussion

5.1 Discussion

The results discovered in this report show that the ability to forecast open interest is immense. It could provide insight into how the market is trending day-to-day and

the possibility of predicting what comes next. Since the options market constantly moves up and down based on trends or business, and many outside influences, our top performing models for each ticker displayed accurate results in predicting the forecast. The 10 and 50-day forecasts for each ticker using their corresponding top performing, non-neural network model can be seen below:

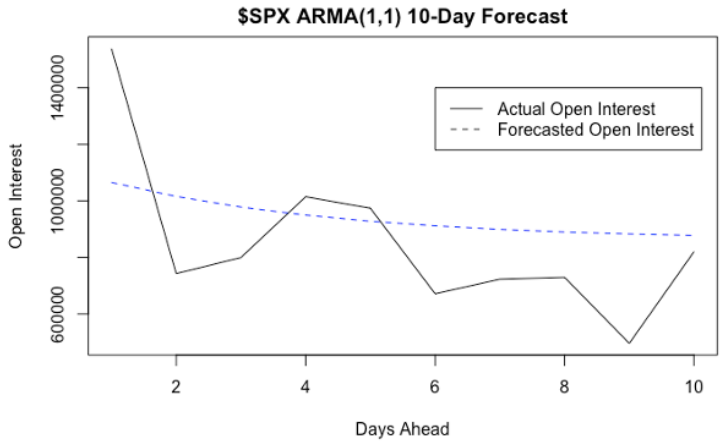

Figure 10: 10-day forecast for $SPX Open Interest using ARMA (1, 1) model

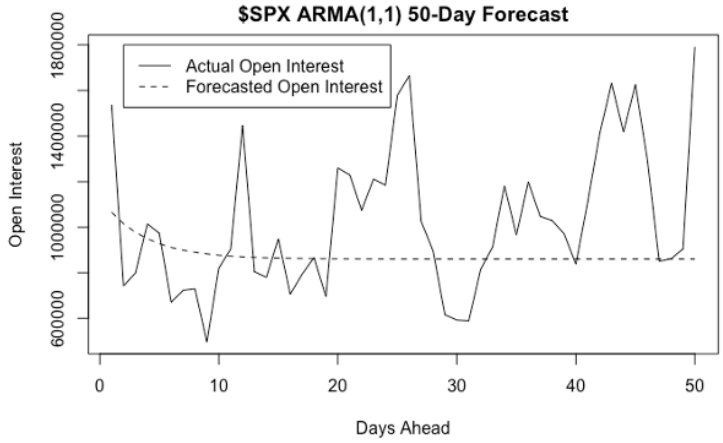

Figure 11: 50-day forecast for $SPX Open Interest using ARMA (1, 1) model

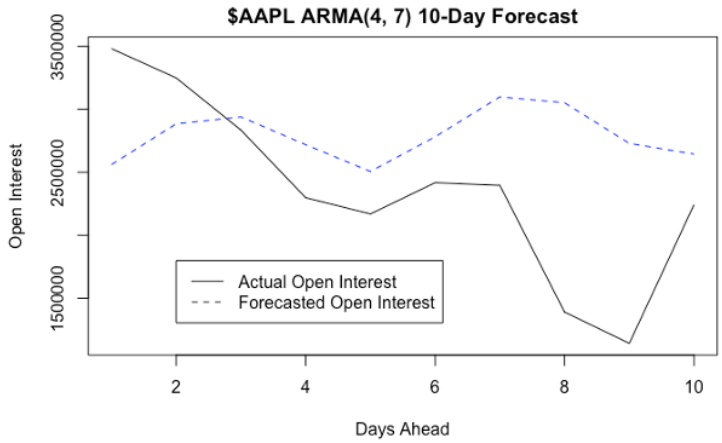

Figure 12: 10-day forecast for $AAPL Open Interest using ARMA (4, 7) model

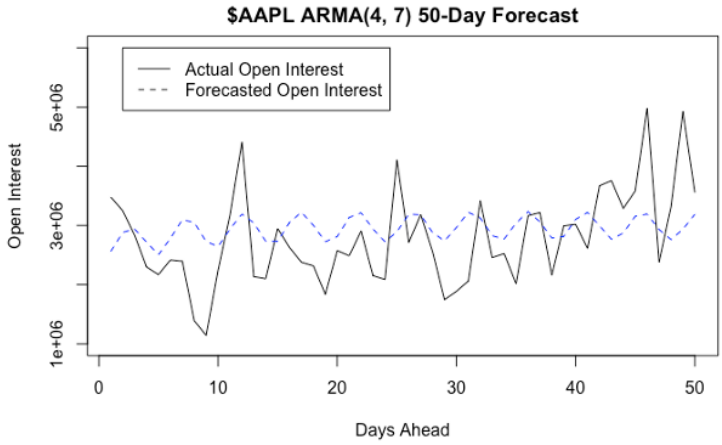

Figure 12: 50-day forecast for $AAPL Open Interest using ARMA (4,7) model

We can see in our foreasts that our short term forecasts are more valuable than our long term (50 day) forecasts. The short term (10 day) forecasts mimic the behavior of the open interest data, while the 50-day forecast just converges to the mean open interest value. This makes conceptual sense as what we know today can make better predictions for tomorrow, but 50 days out would not be helpful in this case due to the nature of the markets and the lack of knowledge about what may happen in the short term. The uncertainty of these predictions is compounded the longer the time horizon is for the predictions. The short-term look into the future is usually the best, allowing one to be a proactive member of the market, rather than just having a contract expire.

The SPY and AAPL symbols move differently from each other since the SPY represents the S&P 500, which combines multiple companies from different

industries, whereas AAPL is just one technology company. Apple is a very volatile company, so its open interest contracts are abundant and constantly changing daily. The SPY contract open interest contracts tend to vary based on any market impact, which could change open interest drastically.

A Vector Autoregressive model is examined to see if the put-call ratio somehow correlates to open interest. PCR does have and had in determining how options are moving from day to day, but the VAR model did not show accurate results when it came to predicting open interest. From this aspect, PCR is useful as a tool alongside open interest. With PCR, one could see if more puts or calls are being used for certain symbols, they are interested in investing in. The use of open interest can be used as a tool and a guide for new investors as well as people who want to invest more who try their hand at options trading. With this tool, a possible investment company that has mobile applications or trading platforms could provide a tool like this to be of service to attract newcomers.

5.2 Ethical Discussion

This research is believed to not contain any violations of ethical standards imposed on the paper by the affiliated institutions. The data used to create the previously discussed models was obtained through legal channels and does not contain any information that can be regarded as outside of the public domain. For full transparency, the specific data used here was acquired from a specific company's database, but the exact same information can be gained from a multitude of publicly accessible sources.

That said, the models and any results-based discussion presented in this research paper are not intended to be taken as actionable recommendations. This is strictly intended as an educational tool, and the authors, nor any affected institutions, can be held responsible for any effects of implementing this work.

Lastly, because of the nature of this research paper, it does not violate any ethical standards set by the financial industry. Specifically, the paper is not susceptible to common financial ethical dilemmas such as insider trading, conflicts of stakeholder interest, or improper investment management.

6 Conclusion

Overall, the ARMA model provided outstanding results on the S&P 500 (SPY) and Apple's (AAPL) symbols. With contracts, there are expiration dates attached to them, so depending on the investor's intention, it can be monumental. This provides a different aesthetic not only to them but allows one to potentially make money at a bit of a faster pace than traditional stocks. This also introduces investors to help alleviate risks in their portfolios and not keep all their money in one venture. The models used

here are just the beginning, and other models could prove to be a bit more accurate. One could continue this work to possibly look at Long Short-Term Memory or Convolutional Networks as well. One could also advance this research by taking into consideration sentiment analysis as well. The market does not just thrive on numbers but many outside influences, so other factors could be added to the data to emphasize these features.

The market, in general, is very difficult to navigate and predict due to these factors, so the challenge itself of looking at non-stationary models comes into play. Other tools could be used in hand with open interest and the put-call ratio as well, which could benefit in predicting. The ability to educate people in financial literacy and provide an avenue for them to learn about other investing options is important. As technology continues to improve, many things will begin to become more streamlined, which could indeed help people catch up and be active players.

References

1. Alotaibi, R. (2022). Arima model for stock market prediction. *2022 8th International Conference on Computer Technology Applications*. https://doi.org/10.1145/3543712.3543723
2. Amagir, A., Groot, W., Maassen van den Brink, H., & Wilschut, A. (2017). A review of financial-literacy education programs for children and adolescents. Citizenship, Social and Economics Education, 17(1), 56–80. https://doi.org/10.1177/2047173417719555
3. Anderson, J. (2019, October 14). *This is why 55% of Americans aren't investing*. GOBankingRates. Retrieved October 31, 2022, from https://www.gobankingrates.com/investing/strategy/this-is-why-55-of-americans-arent-investing/
4. Camberato, J. (2022, October 12). *Council post: Should schools teach financial literacy classes?* Forbes. Retrieved December 1, 2022, from https://www.forbes.com/sites/forbesfinancecouncil/2022/10/11/should-schools-teach-financial-literacy-classes/?sh=7be75a8d4633
5. Chavan, S., Kumar, P., & Gianelle, T. (n.d.). *Intelligent Investment Portfolio Management using time-series analytics and deep reinforcement learning*. SMU Scholar. Retrieved September 18, 2022, from https://scholar.smu.edu/datasciencereview/vol5/iss2/7/
6. Downey, L. (2022, August 23). *Essential Options Trading Guide*. Investopedia. Retrieved October 31, 2022, from https://www.investopedia.com/options-basics-tutorial
7. Faulkner, A. (2021). Financial Literacy around the world: What we can learn from the national strategies and contexts of the top ten most financially literate nations. The Reference Librarian, 63(1-2), 1–28. https://doi.org/10.1080/02763877.2021.2009955
8. Hosker, J., Djurdjevic, S., Nguyen, H., & Slater, R. (n.d.). *Improving VIX futures forecasts using machine learning methods*. SMU Scholar. Retrieved September 18, 2022, from https://scholar.smu.edu/datasciencereview/vol1/iss4/6/

9. Kostopoulos, D., Meyer, S., & Uhr, C. (2020). Google search volume and individual investor trading. Journal of Financial Markets, 49, 100544. https://doi.org/10.1016/j.finmar.2020.100544
10. Lusardi, A. (2019). Financial Literacy and the need for financial education: Evidence and implications. *Swiss Journal of Economics and Statistics*, *155*(1). https://doi.org/10.1186/s41937-019-0027-5
11. Lusardi, A., & Mitchell, O. S. (2014). The economic importance of Financial Literacy: Theory and Evidence. Journal of Economic Literature, 52(1), 5–44. https://doi.org/10.1257/jel.52.1.5
12. Nicolosi, G., Peng, L., & Zhu, N. (2009). Do individual investors learn from their trading experience? Journal of Financial Markets, 12(2), 317–336. https://doi.org/10.1016/j.finmar.2008.07.001
13. Pagliaro, C., Mehta, D., Shiao, H.-T., Wang, S., & Xiong, L. (2021). Investor behavior modeling by analyzing financial advisor notes. Proceedings of the Second ACM International Conference on AI in Finance. https://doi.org/10.1145/3490354.3494388
14. Rostan, P., Rostan, A., & Nurunnabi, M. (2020). Options trading strategy based on ARIMA forecasting. *PSU Research Review, 4*(2), 111-127. 10.1108/PRR-07-2019-0023
15. Sakarwala, M. A., & Tanaydin, A. (n.d.). *Use advances in Data Science and computing power to invest in stock market*. SMU Scholar. Retrieved September 18, 2022, from https://scholar.smu.edu/datasciencereview/vol2/iss1/17/
16. Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: A systematic review. *Journal of Physics: Conference Series*, *2161*(1), 012065. https://doi.org/10.1088/1742-6596/2161/1/012065
17. Sun, Z. (2020). Comparison of trend forecast using Arima and ETS models for S&P500 Close Price. *2020 The 4th International Conference on E-Business and Internet*. https://doi.org/10.1145/3436209.3436894
18. The Options Clearing Corporation. (n.d.). *OCC - the Foundation for secure markets - theocc.com*. Retrieved September 19, 2022, from https://www.theocc.com/.
19. Zabcic-Matic, T. F. (2019). *One Step Back, Two Steps Forward - a Machine Learning-Powered Options Trading Strategy* (thesis). Retrieved September 18, 2022, from https://dash.harvard.edu/handle/1/37364632.
20. Zeng, Y. (2017). *Machine Learning in Option Markets* Available from Dissertation Abstracts International http://www.pqdtcn.com/thesisDetails/A3ABA5AA9895B956AF8603F237954168
21. Zhang, C., Ji, Z., Zhang, J., Wang, Y., Zhao, X., & Yang, Y. (2018). Predicting Chinese stock market price trend using machine learning approach. *Proceedings of the 2nd International Conference on Computer Science and Application Engineering - CSAE '18*. https://doi.org/10.1145/3207677.3277966

Appendix

Use if needed for additional information