Discover Artificial Intelligence

Research

A machine learning approach to predict the S&P 500 absolute percent change

F. S. Rodriguez[1] · P. Norouzzadeh[1] · Z. Anwar[2] · E. Snir[3] · B. Rahmani[1]

Received: 5 July 2023 / Accepted: 2 January 2024
Published online: 25 January 2024

Abstract
Models of the stock market often focus on predicting the direction of the stock market. Instead of following this approach, we created a model to predict the daily absolute percent change of the S&P 500. An accurate model of this metric would greatly increase profitability of option trading strategies such as straddles and iron condors. In this project, novel features were created based on historical data and fed to machine learning algorithms such as Decision Trees, Rule Based Classifiers, K-mean Clusters, and Kernels. Based on our findings, Decision Trees and Kernels showed an accuracy of 80% when predicting absolute percent change, while Rule Based Classifiers had an accuracy of 88%.

1 Introduction

Recently, there is growing interest in using machine learning and data analysis methods to predict stock prices. The main objective is often predicting the direction of stock trading.

The S&P 500 index follows the weighted performance of the 500 most valuable companies in the US. It is often used by economists to measure the health of the US stock market and the overall economy [1]. The massive adoption of this index has resulted in an increased interest in retail and institutional investment leading to exponential growth in trading of Exchange-Traded Funds (ETFs) and options on the index. Modeling the performance of the index is valuable.

Multiple factors like earnings season, wars, political news, and traders' psychology influence stock prices [2]. There are two main approaches to predict stock performance: fundamental and technical analysis. Fundamental analysis focuses on how a company is performing and the success of the relevant industry and market. Information such as earnings per share, dividend yield, price-to-earnings ratio, and fair market value of the stock are important metrics in fundamental analysis [3].

To predict possible price movements, some traders use technical analysis to find trends or patterns. This discipline employs past data such as price movement and volume to identify trading opportunities [4]. Although past performance does not correlate highly with future returns, people's emotions have autocorrelation. Technical analysts create compound features using historical data. These technical indicators may find recurring patterns in stock price movement [5].

Numerous machine learning papers propose models that could accomplish this feat. Macchiarulo has tried both technical analysis and machine learning methods to predict the movement of stock prices [6]. Other models look at the relationship between social sentiment and the stock price to forecast future movements [7].

✉ B. Rahmani, Bahareh.Rahmani@slu.edu | [1]Saint Louis University, Saint Louis, USA. [2]North Dakota State University, Fargo, USA. [3]Washington University in St. Louis, Saint Louis, USA.

Brunhuemer et al. evaluated a machine learning model which focuses on trading straddles of buying put and call options at the same strike price. They focus on volatility and other technical analysis data as inputs into the machine learning model [8]. They show that predicting volatility is as useful as predicting a specific market direction.

Machine learning techniques such as support vector regression, decision trees, k-nearest neighbor, random forest and multilayer perceptron have been used to predict the stock market. Techniques such as MLP and LSTM have showed promising results [9].

Fu [10] applied hybrid machine learning methods to increase the accuracy of S&P index prediction [10]. Dr. Wei applied a support vector machine method to predict S&P indices [11]. Wei applied different machine learning methods with boosting approach to increase the accuracy of prediction.

In this paper, a combination of technical and statistical analysis and machine learning methods are used to create models that can improve the likelihood of accurately forecasting the absolute movement of the S&P 500 index. The predictions focus on large versus small changes in the index. An absolute change of less than 1% is coded as a small change (Group 0), while a change above 1% is coded as a large change (Group 1). The reasoning for this is that S&P 500 options have a high likelihood of yielding over 100% daily return when the index moves more than 1% in either direction.

2 Data description

Historical data for the S&P 500 from January 2000 until August 2022 was taken from Yahoo Finance [12]. The selected features include opening price, highest price in the day, lowest price in the day, closing price, and volume. These factors were constructed and preprocessed for analysis with machine learning methods. Two main groups of features were created: Features based on percentages, and features based on days.

2.1 Features based on percentages

Percentage features are datapoints that come from changes among attributes of the original data, transformed into percentages. The features that belong to this group are:

- Daily Percent Change—the percentage difference between opening and closing price.
- Overnight Percent Change—percent change between the previous day's close, and current day's open
- Percentage Volatility—percent change between the daily low and high price
- 5-day and 3-day Percent Moving Average—moving average of Daily Percent Change and Absolute Percent Change for 3 or 5 days
- Absolute Percent Change (Abs Mov)– absolute value of the Daily Percent Change

These features were then passed through binarization and discretization to facilitate implementation in the algorithms.

2.2 Features based on days

These features investigate the characteristics of the trading day to create data points. The features that belong to this group are:

- Day of the Week
- Consecutive Red Days—the number of days that have closed below the open price.
- Consecutive Green Days—the number of days that have closed above the open price.
- Discretization Consecutive days—merging of the Red Days and Green Days

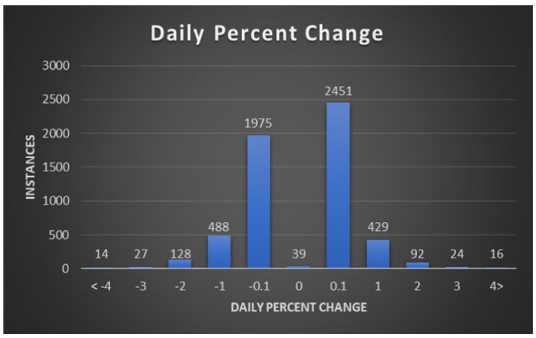


Fig. 1 Daily percent change with number of instances

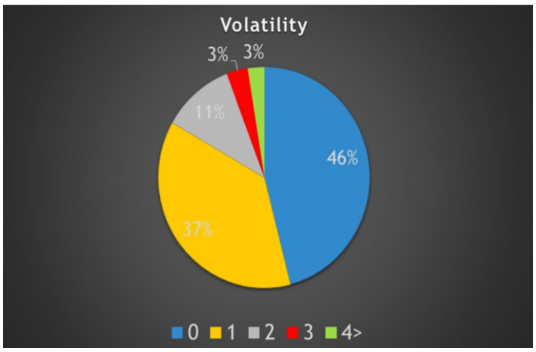


Fig. 2 Intraday volatility, based on percent change

3 Technical and statistical analysis

Before feeding the data into machine learning models, data analysis was performed to investigate any correlations between the features. The likelihood of the closing price of the S&P 500 with an Absolute Percent Change above 1%, is 21.43%. The probabilities of closing price with a percent change between − 1 and + 1% are 78.5%. (Fig. 1).

On average, volatility is 1.34%, and there is a 54% chance that intraday volatility is above 1%. (Fig. 2).

Scatter plots including the correlations between Consecutive Days and Absolute Percent Change shows higher streaks than Absolute Percent Change. Similarly Consecutive Days ranging from − 4 to + 2 show the biggest probabilities of producing a movement above 3%. (Fig. 3).

4 Machine learning models and results

4.1 Decision tree

The Decision Tree is one of the most applied machine learning methods due to their easy implementation and robustness. These trees separate continuous or discrete datasets based on their Gini and Entropy [13].

In our project, the model extracted from the entire dataset for 22 years was 80% accurate. But starting from 2009—after the crisis of 2008—the accuracy of decision tree rose to 87%. The low Gini (~ 0.26) shows the high accuracy of decision tree (Figure 4).

4.2 Rule based classifier

After multiple iterations of decision trees, rule-based classifiers were created to predict the absolute percent change of sessions below 1% [Group 0] or above that threshold [Group 1].

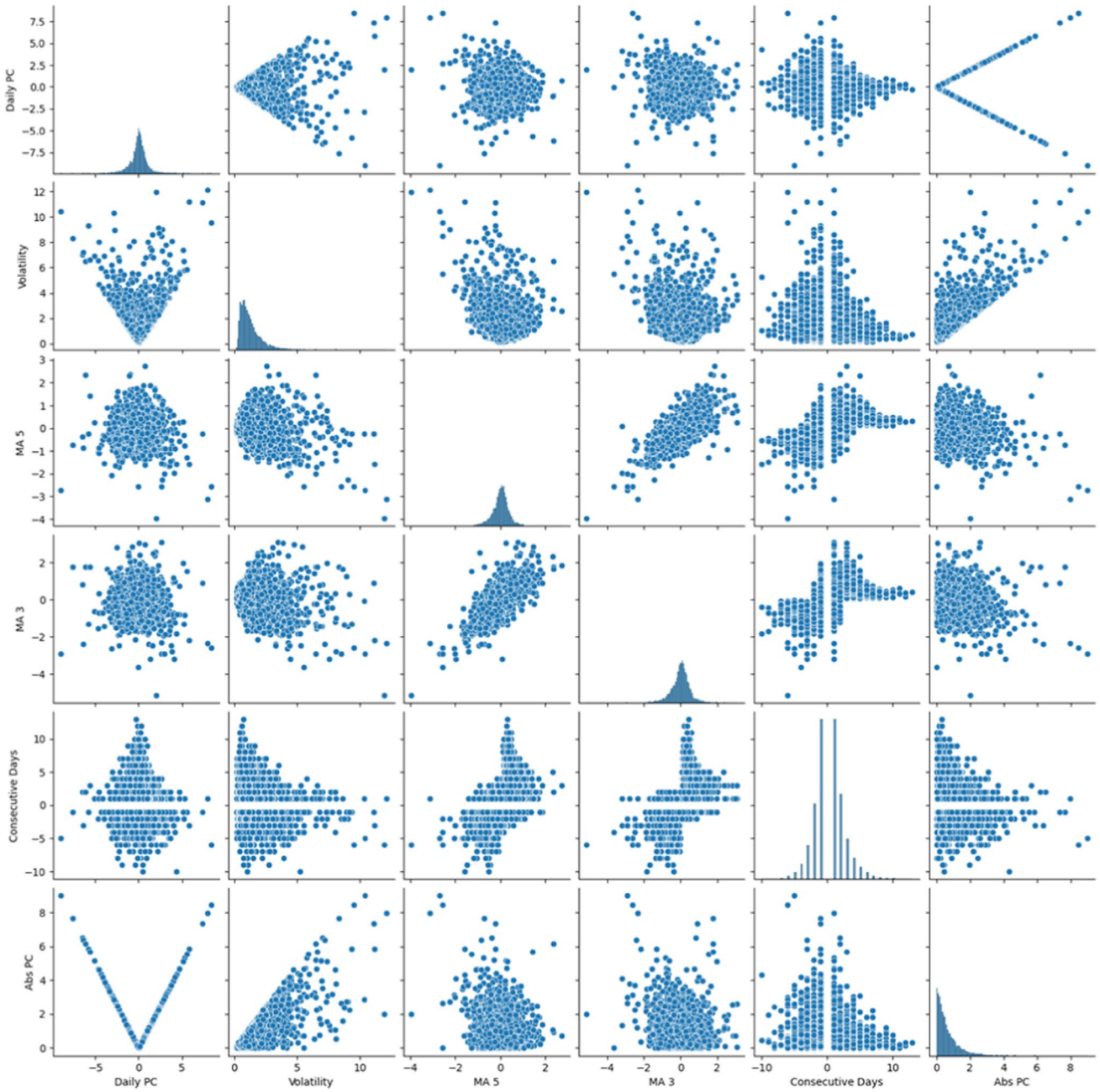


Fig. 3 Scatter plot

Rule 1: used the 3-day moving average for yesterday, today, and tomorrow to predict a low percent change, of less than 1%. The results showed an accuracy of 88.14% with a coverage of 38.62% (Fig. 5). Note that the chance of a day ending the session with a percent change below 1% is 78.5% as seen previously (Fig. 1). This rule generates a 10% improvement.

Rule 2: the 5-day absolute moving average performed better than Rule 1, exhibiting an accuracy of 91.09% when predicting low volume, and a coverage of 42.09% (Fig. 6). This is a 13% improvement over the naïve estimate.

Rule 3: was created to predict absolute price changes above the 1% threshold. Since Abs MA5 performed well with rule 2, it was used again. Rule 3 had a prediction percentage of 50.62% and a coverage of 14.12% (Fig. 7).

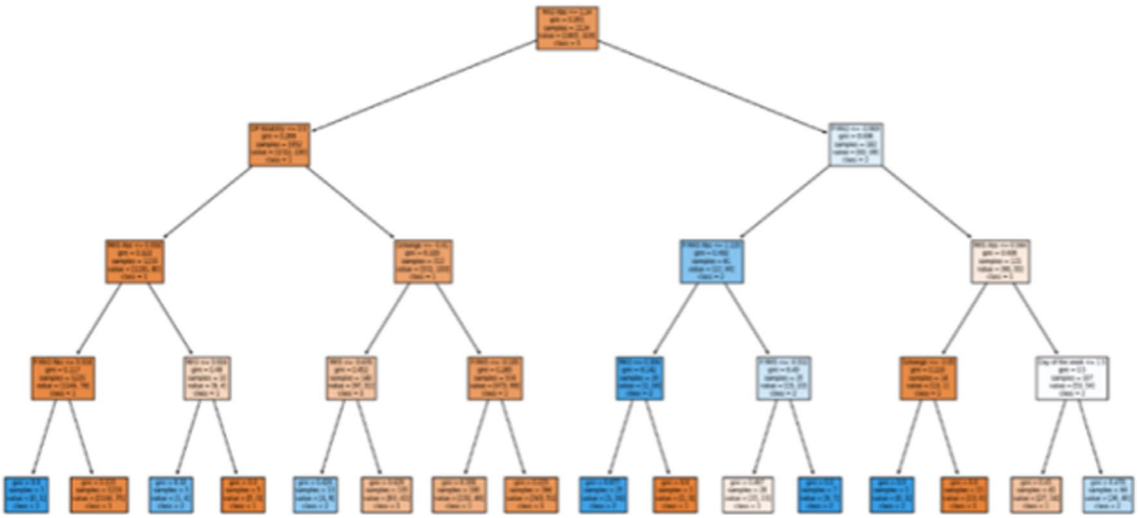


Fig. 4 Decision tree

Fig. 5 Rule 1 statistical data. R1: (– 0.3 ≥ Previous MA3 ≤ 0.3) ^ (– 0.3 ≥ Current MA3 ≤ 0.3)—> Abs Mov = 0

Statistical Data	
Total	5678
Population	2193
When True	1933
Coverage	38.62
Accuracy	88.14

Fig. 6 Rule 2 statistical data. R2: (Previous Abs MA5 ≤ 0.516) ^ (Current Abs MA5 ≤ 0.516)—> Abs Mov = 0

Statistical Data	
Total	5677
Population	2279
When True	2076
Coverage	40.14
Accuracy	91.09

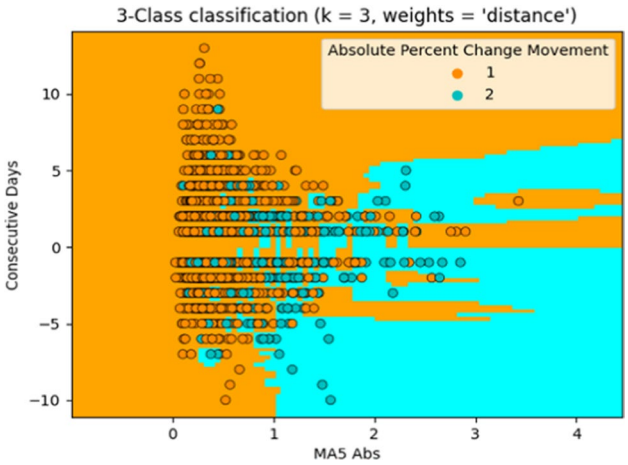
Fig. 7 Rule 3 statistical data. R3: (Previous Abs MA5 ≥ 1) ^ (Current Abs MA5 ≥ 1)—> Abs Mov = 1

Statistical Data	
Total	5677
Population	802
When True	406
Coverage	14.13
Accuracy	50.62

4.3 K-mean classifier

Based on the previous analysis, the Consecutive Days and MA5 Absolute features were fed to a K-mean Classifier algorithm. The 3-mean classifier had an accuracy of 85%. It successfully classified 88% of the Absolute Percent Change Movement of group 0 and 51% of group 1 (Fig. 8).

Fig. 8: 3-mean classifier



5 Conclusion

Looking at the results, all three machine learning methods were able to improve the base probabilities of a low or high Absolute Percent Change Movement. The Rule Based Classifier had the highest accuracy of 91.09% to predict a low percent change in prices, while the K-mean Classifier had the best prediction of a high percent change with 51% accuracy.

Moving average and boosting machine learning methods performed well to predict SP 500 stock prices. In this project a decision tree with 87% accuracy was created. Three rules with the highest accuracy of 91% were derived from decision tree. The features selected from the rule with the highest accuracy (rule 2) were used to classify data with a 3-mean classifier with an accuracy of 88%

Technical and machine learning analysis made the prediction of the S&P 500 index possible with high accuracy. There is a plan to apply Random Forest and Convolutional Neural Network (CNN) to increase the accuracy of predictions. Besides that, we can apply the introduced methods: Decision Tree, Rule Based Classifier and K-Mean to predict prices of other stock indices.

Author contributions FS, analyzed data, developed the codes, created the graphs, wrote the 1st draft. PN, reviewed the paper, made the comments, confirmed the results. ZA, reviewed the paper, rephrased some sentences, made the comments. ES, advisor for the business analytics part, reviewed the paper. BR, defined and lead the project, confirmed results, reviewed the paper.

Funding There is no funding or grant related to this project.

Data availability Historical data for the S&P 500 from January 2000 until August 2022 was taken from Yahoo Finance. It is open source and available for public [10].

Declarations

Competing interests All authors declare that they have no conflicts of interest.

References

1. S&P Global. https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview.

2. Gidofalvi G, Elkan C. Using news articles to predict stock price movements. San Diego: University of California; 2001.
3. Bohl L, Frederick R. How to pick stocks using fundamental and technical analysis. Westlake: Schwab Brokerage; 2022.
4. Hayes A, Battle A, Jackson A. Technical Analysis: What it is and how to use it in investing Investopedia. Accessed 14 Mar 2022.
5. Jiao Y, Jakubowicz J. Predicting stock movement direction with machine learning: an extensive study on S&P 500 stocks. IEEE Int Conf Big Data. 2017. https://doi.org/10.1109/BigData.2017.8258518.
6. Macchiarulo A. Predicting and beating the stock market with machine learning and technical analysis. J Int Bank Commer. 2018;23(1):1–22.
7. Porshnev A, Redkin I, Shevchenko A. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. IEEE Int Conf Data Min Workshops. 2013. https://doi.org/10.1109/ICDMW.2013.111.
8. Brunhuemer A, Larcher L, Seidl P, Desmettre S, Kofler J, Larcher G. Supervised machine learning classification for short straddles on the S&P500. Ithaca: Cornell University; 2022.
9. Medarhri I, Hosni M, Nouisser N, Chakroun F, Najib K. Predicting stock market price movement using machine learning techniques. Piscataway: IEEE; 2022.
10. Fu Z. Machine Learning Models' Combination for Higher Accuracy of S&P 500 Index Prediction. Piscataway: IEEE; 2020.
11. Wei X. Predicting the price of SP500 index based on machine learning Methods. Quant Fin. 2023. https://doi.org/10.2991/978-94-6463-098-5_59.
12. SPY History. Yahoo Finance. Available from https://finance.yahoo.com/quote/SPY/history/
13. Bahzad C, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021;2(01):20–8.