Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/74/20241522

Analysis of the Application of XGBoost in Exchange-Traded Funds

Haoran Peng[1,a,*]

[1]*Metropolitan college, Boston University, Boston, MA, USA, 02215*
a. haoranpeng66@gmail.com
**corresponding author*

Abstract: In the quickly expanding landscape of contemporary financial markets, the paramount significance lies in comprehending and effectively employing novel technologies. Trading, a practice that bears resemblances to the use of sticks, has become a prominent option for structuring investment portfolios owing to its inherent diversification properties. This concept has the potential to enhance individuals' comprehension and decision-making abilities inside the intricate and densely populated realm of contemporary finance. This research aims to explore the correlation between Exchange-Traded Funds (ETFs) and the utilization of sophisticated machine learning methodologies, with a particular focus on XGBoost (eXtreme Gradient Boosting). The study additionally offers an extensive overview of the significance of artificial intelligence (AI) in the field of finance. It employs the concept of Xgboost to address the challenge of handling the substantial volume of datasets in the stock market. This approach is based on the utilization of decision trees, which are a robust machine learning algorithm. The objective is to examine and investigate the profound impact of AI in the realm of finance.

Keywords: Technical Analysis, Machine Learning, Quantitative Trading, Portfolio Management

1. Introduction

In contemporary finance, there has been a notable increase in complexity and competition, resulting in a wider array of investment choices available to individuals for their portfolio allocation. The portfolio serves as a fundamental component of investment management, enabling the diversification of investment risk and facilitating the potential for individuals to generate more returns while minimizing risk. Individuals have the ability to enhance their investment portfolio by directly purchasing bonds or stocks. However, Exchange-Traded Funds (ETFs) have emerged as a prominent method for portfolio diversification. An ETF is a form of investment fund that functions as an exchange-traded product, meaning it can be bought and sold on stock markets. ETFs possess inherent diversification as a fund type and are notably exchanged on exchanges. Individuals have the ability to engage in the buying and selling of ETFs at their discretion, similar to the manner in which stocks are traded. Given that ETFs are sold on exchanges, they are supported by governmental sponsorship, thereby establishing a level of safety in their trading activities. In recent years, there has been significant progress in the field of Machine Learning and Deep Learning. Consequently, the utilization of Machine Learning (ML) techniques for the prediction of stock prices has emerged as a

DOI: 10.54254/2754-1169/74/20241522

crucial trading algorithm in the financial market. Prominent hedge funds, such as Bridgewater Associates and Millennium Management, have made substantial investments in machine learning algorithmic trading, resulting in significant financial gains amounting to billions of dollars. Among the various machine learning algorithms, XGBoost (eXtreme Gradient Boosting) emerges as a prominent and influential technique within the domain of stock trading. This research aims to explore the correlation between ETFs and the utilization of sophisticated machine learning methodologies, with a particular focus on XGBoost (eXtreme Gradient Boosting). The study additionally presents an extensive overview of the significance of AI in the field of finance. It employs the concept of Xgboost to address the challenge of handling the vast amount of datasets in the stock market. This approach is based on the utilization of decision trees, which are known for their robust machine learning capabilities. The work aims to explore and analyze the transformative potential of AI in the financial industry. In contrast, the author predominantly employs XGBoost as an exemplary instance within the domain of finance; yet, its functionalities transcend this specific context. If XGBoost demonstrates its ability to efficiently handle and analyze the voluminous, intricate, and dynamic datasets commonly encountered in stock markets, it undeniably possesses the capacity to foster progress in other domains.

ETFs have emerged as significant components within contemporary investment portfolios, exerting a substantial impact on the financial market. The positive trajectory of ETFs has the potential to induce a corresponding upward movement in other indexes, such as stocks and bonds. Therefore, it is vital to comprehend and forecast ETFs. This study use the Xgboost algorithm, a sophisticated Machine Learning technique, in conjunction with technical analysis indicators to examine and forecast ETFs. The indicators will be classified into four distinct aspects: Trend Indicators, Momentum Indicators, Volatility Indicators, and Volume Indicators. It will then assess the predictive capabilities of each element in relation to ETF prices. Upon accurately forecasting the trajectory of ETFs, people will get a comprehensive comprehension of the ultimate market conditions.

2. Review of AI and its development in Finance

2.1. The development of AI

In the contemporary era of artificial intelligence, AI has emerged as a pivotal instrument across various domains, notably finance. AI has proven to be highly effective in the field of facial recognition, making significant contributions to criminal detection. Additionally, AI technology has been successfully employed in the medical domain, particularly in the area of medication. For instance, AI algorithms can analyze X-ray images to determine the presence of illness or disease in patients [1]. The process of human face or X-ray photo detection traditionally requires a significant amount of time and effort. However, with the advent of AI, the ability to rapidly scan and analyze thousands of photos or faces in a single day has emerged, resulting in a substantial increase in efficiency. Furthermore, artificial intelligence algorithms can also be employed for the purpose of data analysis. An increasing number of young individuals are engaging in the study and use of algorithms across several domains. AI systems are capable of efficiently solving complex computational problems within a short period of time. Additionally, AI systems possess the ability to identify and establish correlations across data sets that may elude human comprehension. The data can be inputted into artificial intelligence systems, which can subsequently provide the desired outcome.

2.2. AI in Finance

In the realm of financial markets, data serves as the fundamental basis for investment decisions. In the context of fundamental analysis, the acquisition of data such as earnings per share (EPS) or the Sharpe Ratio is essential. Conversely, in the realm of technical analysis, the gathering of data pertaining to the opening, closing, highest, and lowest prices is necessary for constructing a bar chart.

DOI: 10.54254/2754-1169/74/20241522

Given its robust data analysis capabilities, AI has emerged as a significant player in the financial sector. According to Stephen Muggleton's scholarly article titled "Alan Turing and the Development of Artificial Intelligence," scientists such as Turing made predictions regarding the future development of AI and its potential to exert dominion or exert influence across various domains.In this study, we aim to investigate the effects of various environmental factors on plant growth. Machine Learning (ML) and Deep Learning (DL) algorithms have significantly enhanced the investment returns for both investors and fund managers. Intelligent systems possess the capability to effectively handle extensive datasets, doing tasks at speeds and depths that beyond human capacity for manual completion. According to the introduction by Pramila P. Shinde, the concept of AI emerged in the 1950s with the aim of developing robots that had human-like or even superior intelligence. The utilization of machine learning (ML) and DL techniques has demonstrated that AI is highly proficient in processing vast amounts of data sets [2]. The financial market data contains a significant amount of noise and irrational data. AI systems have the ability to analyze and comprehend this data, enabling them to distinguish between the helpful and potentially deceptive portions. Utilizing machine learning approaches, AI has the capability to reveal intricate and concealed patterns inside data, so enabling enhanced and prompt investment decision-making. Financial institutions have developed Robo Advisors, automated systems that assist individuals in constructing investment strategies. These Robo Advisors typically entail lower fees when compared to their human counterparts. Prominent hedge funds, such Bridgewater Associates and Millennium Management, have made substantial investments in ML algorithmic trading, resulting in significant financial gains up to billions of dollars. Additionally, an increasing number of hedge funds are publicly announcing their utilization of AI to assist in portfolio construction. AI algorithms have emerged as indispensable tools in the field of investment.

2.3. Exchange-Traded Funds

In recent times, ETFs have gained increasing significance within the realm of financial markets. ETFs are a type of investment instrument that combines characteristics of both mutual funds and equities. The ETF was established in 1989 with the primary objective of providing a more efficient and cost-effective investment choice. ETFs are investment vehicles that, as their name implies, function similarly to funds. These ETFs are designed to allocate investments among multiple stocks, thereby mitigating the potential risks associated with a concentrated portfolio. Additionally, ETFs has the capability to be exchanged on the exchange, similar to equities. Therefore, ETFs possess the dual benefits of diversification and liquidity. ETFs enable investors to diversify their risk by allowing them to participate in a portfolio of securities rather than a single security. Additionally, ETFs offer liquidity, facilitating the buying and selling of shares as desired by investors. ETFs encompass a variety of asset classes, such as stock ETFs, bond ETFs, and sector ETFs, among others. Providing a variety of options for investors to select from in terms of geographical areas. Due to their inherent simplicity and wide range of investment options, ETFs have gained significant popularity among both individual and institutional investors. During the second quarter of 2023, the average daily trading volumes for United States stocks and United States ETFs amounted to $495.4 billion and $141.6 billion, respectively [3]. US ETFs constituted more than 28% of the overall composite volume in the secondary market within the United States.

Nevertheless, ETFs are not exempt from encountering certain dangers, such as market risk, for instance. The suitability of applying AI approaches to handle the challenges posed by ETFs stems from their inherent characteristics, such as the abundance of datasets and the ability to identify intricate patterns. These attributes align with the specialized domain of AI, which excels in handling extensive data and deciphering complex patterns. The utilization of ETFs enables AI to execute financial market activities, signifying a substantial transformation in the approach of market players

DOI: 10.54254/2754-1169/74/20241522

towards asset allocation. According to the findings of Min-Yuh Day and Jian-Ting Lin, various deep learning modules have been employed in the ETF, resulting in an annual return of up to 12% [3]. The subsequent portions of this article will explore the utilization of artificial intelligence, particularly machine learning models such as XGBoost, to address the issue in the ETF domain. Additionally, the study will discuss the implementation of innovative and efficient financial methodologies to execute the program.

3. Decision tree and Xgboost

3.1. Decision tree

Decision trees have played a prominent role in the historical progression of algorithm development. Decision trees are a type of non-parametric supervised learning technique that is commonly used for both classification and regression applications. The process involves partitioning a given dataset into smaller subgroups, concurrently constructing an associated decision tree in an incremental manner. The ultimate outcome will manifest as a tree structure, wherein the decision branch and leaf nodes symbolize the rules and potential responses. Due to their distinctive structure, decision trees are highly appropriate for analyzing non-linear datasets. Fortunately, the nature of financial data is non-linear, which contributes to the inherent challenge of accurately predicting stock values.Therefore, the decision tree method is deemed appropriate for application in the financial market. However, decision trees possess a drawback in that they are prone to overfitting. In principle, it is possible to assign each output in the training dataset to a single leaf, so achieving a perfect fit of the training set. However, this approach would inevitably result in overfitting. Therefore, it is necessary to do tree pruning by modifying the parameter, and it is imperative to experiment with various hyperparameters in order to achieve optimal outcomes. According to J. Myles' paper titled "An Introduction to Decision Tree Modeling," the issue of overfitting arises when decision trees are not properly pruned or when decision rules are not adequately generalized. However, despite this concern, decision trees can still be effectively employed in large databases, playing a crucial role in discrimination and predictive modeling. The user's text does not provide any information to rewrite in an academic manner. Despite the persisting challenges associated with decision trees, they remain a crucial component in the field of machine learning.

To address such a challenge, the Random Forests algorithm was developed. Random Forests generate an ensemble of decision trees, collectively referred to as a "forest." The fundamental concept involves the creation of multiple independent trees, each tasked with performing either the most popular classification or averaging the regression. The proposed methodology relies on prediction accuracy and incorporates measures to mitigate overfitting by aggregating the outcomes obtained from individual trees. The abundance of trees inside the "forest" will enable the capture of intricate structures present in the dataset.

3.2. XGBoost

Xgboost, an abbreviation for eXtreme Gradient Boosting, builds upon the progression from decision tree to random forest, elevating this concept to a higher level. XGBoost is a meticulously engineered distributed gradient boosting library that has been specifically intended to exhibit exceptional efficiency, adaptability, and portability. Within the scope of Gradient Boosting, the technique leverages the process of minimizing errors in order to successfully execute and achieve machine learning algorithms. In the work presented by Tianqi Chen and Carlos Guestrin, they introduce XGBoost, a machine learning framework designed for scalable tree boosting. The system is accessible in the form of an open source package. The influence of the system has been widely acknowledged in many machine learning and data mining challenges [4]. XGBoost has the capability to effectively

DOI: 10.54254/2754-1169/74/20241522

address the issue of missing values as well as handle sparse data. Additionally, this technology has the capability to operate in both parallel and distributed computing environments, resulting in enhanced speed and scalability compared to numerous other similar approaches. In their study titled "XGBoost: A Scalable Tree Boosting System," Tianqi Chen and Carlos Guestrin discuss the incorporation of a regularized model in XGBoost to mitigate the issue of overfitting. The approach presented in this study bears resemblance to prior research on regularized greedy forest, albeit with a simplified target and algorithmic framework to facilitate parallelization [5]. This also elucidates the advantage of Xgboost and addresses the drawback of decision trees pertaining to their susceptibility to overfitting.

XGBoost offers numerous advantages, one of which is its ability to address the prevalent issue of missing data in the field of data science. In accordance with the scholarly work authored by Kyung Keun Yun, the study proposes the utilization of a hybrid GA-XGBoost algorithm in predicting the direction of stock prices. This approach involves a three-stage feature engineering procedure. The XGBoost model demonstrates higher accuracy compared to other classic machine models and exhibits faster computational speed [6]. The algorithm's capacity to execute regularized boosting not only helps its predicting accuracy but also improves its ability to generalize. The superiority of this model over other machine learning models can be attributed to its exceptional performance in classification and regression tasks. The adaptability and performance of this tool have contributed to its widespread popularity among data scientists across various domains.

4. Methodology

4.1. Data preparation

There are two primary forms of trading analysis methods, namely fundamental analysis and technical analysis. Fundamental analysis use the accounting methodology to assess the intrinsic worth of a company, thereby informing investment decision-making. Technical analysis, on the other hand, employs past price and volume data of a company to conduct analysis and subsequently formulate investment judgments. The fundamental tenets and underlying presumptions of technical analysis revolve around the notion that price movements exhibit repeating patterns and that the market incorporates all available information. The application of technical analysis extends to openly traded securities across the worldwide market, encompassing a diverse array of financial instruments including equities, bonds, commodities, and ETFs. Nevertheless, the application of technical analysis is more efficient in liquid markets due to the requirement of ample data for analysis. In contemporary times, the proliferation of big data has resulted in a substantial increase in the availability of financial market data. Consequently, the significance of technical analysis has been increasingly pronounced inside the financial market. Due to the capability of AI algorithms to identify concealed patterns within extensive datasets, the application of AI in investment has become prevalent, particularly in the field of technical analysis.

4.2. Dataset Introduction

The ETF data is obtained from Yahoo Finance, a widely utilized website among several financial investors. The dataset comprises a collection of 18 ETFs that are among the most actively traded in the United States. These specific ETFs have been chosen due to the requirement of ample data for AI systems to generate accurate forecasts. Adequate transaction records and data pertaining to the most heavily traded ETFs can be procured for the purpose of training our model. Presented is a visual representation of the SPY ETF, which is among the exchange-traded funds included in our dataset.

DOI: 10.54254/2754-1169/74/20241522

SPY dataset

	Date	Open	High	Low	Close	Adj Close	Volume
0	2010-01-08	113.889999	114.620003	113.660004	114.570000	88.440407	126402800
1	2010-01-11	115.080002	115.129997	114.239998	114.730003	88.563919	106375700
2	2010-01-12	113.970001	114.209999	113.220001	113.660004	87.737946	163333500
3	2010-01-13	113.949997	114.940002	113.370003	114.620003	88.478996	161822000
4	2010-01-14	114.489998	115.139999	114.419998	114.930000	88.718285	115718800
...	...	...	...	...	...	...	...
3477	2023-11-01	419.200012	423.500000	418.649994	422.660004	422.660004	98068100
3478	2023-11-02	426.579987	430.920013	426.559998	430.760010	430.760010	94938900
3479	2023-11-03	433.140015	436.290009	433.010010	434.690002	434.690002	100110800
3480	2023-11-06	435.470001	436.149994	433.679993	435.690002	435.690002	67831700
3481	2023-11-07	435.690002	437.589996	434.510010	436.929993	436.929993	64211000

Figure 1: SPY Dataset (source:yahoo finance)

As depicted in Figure 1, the dataset comprises daily data including several attributes such as open price, high price, low price, closing price, and volume for each ETF. The aforementioned data serve as the foundational elements for doing technical analysis. However, it is important to compute additional technical indicators in order to enhance the efficacy of the model. Technical indicators play a crucial role in extracting supplementary insights from fundamental daily data. These indicators possess the capability to gauge various patterns in price movements, market mood, and money flows, hence enabling the anticipation of future price fluctuations.

In this paper, we seperate the indictors mainly into four groups: Trend Indicators, Momentum Indicators, Volatility Indicators, Volume Indicators. Trend Indicators assist traders in visualizing asset price momentumes and expected price swings. Here MA (Moving Average), CCI (Commodity Channel Index), RSI (Relative Strength Index), OBV (On-Balance Volume) are selected as the trend Indicators.

Momentum indicators can help traders determine the strength or weakness of a stock's price. DMA(Day Moving Average), DMI (Directinal Movement Index), DPO(Detrended Price Oscillator), MACD (Moving Average Convergence Divergence) are selected as the momentum indicators. Volatility indicators can help traders understand how much and how quickly the prices of assets, like stocks or currencies, are changing. Here, ADX (Average Directional Index), ATR (Average True Range), CR (Chart), BBIBOLL (Bollinger Bands) are selected. Lastly, Volume indicators can indicate the perception of investors about a specific stock by measuring the number of traders that are interested in buying or selling ti at a given point. Here MAR (Moving Average Ribbons), VROC (Volume of Rate of Change), VRSI (Volume Relative Strength Index), MAD (Moving Average Difference) are selected.

4.3. Model Building

Currently, a total of 16 indicators have been obtained from four distinct groups, which will serve as the input for our analysis. The proposed methodology involves training distinct groups individually and subsequently assessing the predictive performance of each group. The market's level of influence can be observed by analyzing several groupings. Subsequently, the 16 indicators will be consolidated into a cohesive unit for the purpose of assessing their overall performance. Regarding the output, our approach entails constructing the XGBoost model as a binary classifier. In order to determine the % change of the daily closing price, it is necessary to categorize the resulting values into two distinct groups. Specifically, a value of 1 is assigned to non-negative percentage changes, while a value of 0

DOI: 10.54254/2754-1169/74/20241522

is assigned to negative percentage changes. The selection of XGBoostclassifier over XGBoostregressor is based on the belief that classifiers are superior models. Classifiers are robotic entities, and in the context of the stock market, our primary concern lies in determining if the price is ascending or not, rather than the specific magnitude of its increase. In the event that there is a significant likelihood of a price increase tomorrow, we will proceed with the purchase of this stock. In this study, we categorize a value of 1 to represent a non-negative percentage change. The group labeled as 1 holds the most significance, and our objective is to maximize the accuracy of predicting values within this group. The confusion matrix is a crucial tool for assessing predictions and is more effective than the mean squared error.

The information encompasses the entirety of the time period from 2010 to 2023, with around 3300 trade days. The initial 3000 trading days will be utilized as the training data for our XGBoost model, while the subsequent 300 trading days will serve as the testing data to evaluate the predictive performance. This study adopts the assumption that an initial investment of $100 is made for each ETF, and thereafter computes the monetary value based on the projected outcome. If the model's prediction is 1, it signifies that we will proceed with purchasing shares in this ETF, thus resulting in a modification in the monetary value. If the model predicts a value of 0, it would imply that we should refrain from both purchasing and selling this ETF. Consequently, there would be no alteration in the quantity of money.

5. Results and analysis

Initially, we will exclusively employ SPY data as the training dataset and assess the predictive capabilities of this model.

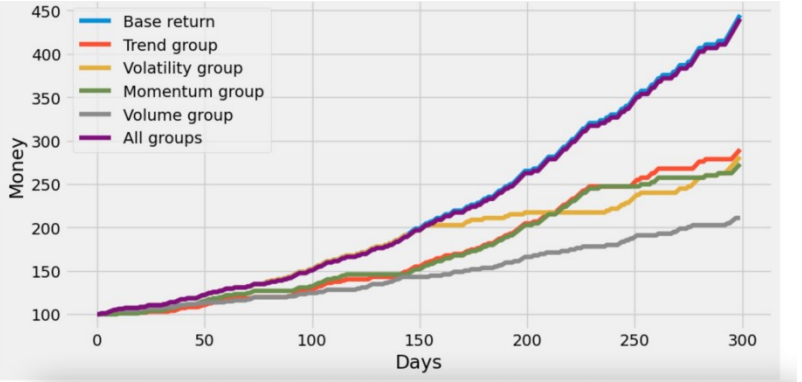

Figure 2: Return from different groups for SPY data

Based on the aforementioned output, the XGBoost investment commenced with an initial capital of $100. Over the course of around 300 trading days, the investment had a minimum threefold increase, with the highest recorded outcome reaching approximately $400. It is evident that the overall line (represented by the purple line), which incorporates all 16 indicators, has the highest level of return in comparison to the other groups. In order to optimize investment decision-making, it is imperative to incorporate the perspectives and insights of all four groups. Utilizing all categories of indicators enables a more comprehensive assessment of the market. One of the groups analyzed, namely the volume group, exhibits the lowest level of return. This observation suggests that while volume may have the capacity to influence market dynamics, it is not the sole determinant of market behavior. In order to enhance investment decision-making, it is imperative for volume indicators to be utilized in conjunction with other indicators. Nevertheless, our model exhibits inferior performance compared to the baseline strategy of purchasing and holding for a period of 300 trading

DOI: 10.54254/2754-1169/74/20241522

days. The reason for this phenomenon is the tremendous growth of ETFs in recent years, which has led to a situation where individuals can simply invest in ETFs and generate financial returns.

Following the utilization of a single ETF, we proceed to construct our portfolio by incorporating all 18 available ETFs, so evaluating the efficacy of our methodology. The methodology entails training each ETF individually, with a starting investment of $100 allocated to each ETF. The statement implies that an initial investment of 1800 units will be made, and the outcomes produced by our model will be observed.

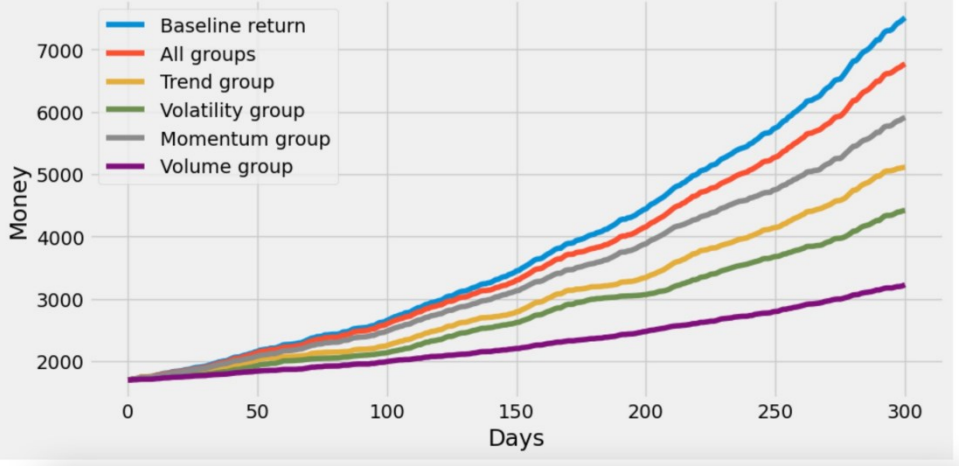

Figure 3: Return for ETF portfolio from different groups

Figure 3 illustrates the portfolio's return. However, the portfolio that incorporates all 16 indicators has superior performance. It commenced with an initial investment of $1800 and yielded a final value over $6000 over a span of 300 trading days. The proposition posits that the inclusion of all types of indicators is necessary for the development of a sound investment strategy. The trend group demonstrates a return that is second only to the 16-indicator group in terms of its level of profitability. The rationale behind this observation is that ETFs have exhibited a notable rising trajectory in recent years, resulting in a relatively favorable return for the trend group. The volume group exhibits the lowest rate of return compared to the other groups, thereby reinforcing the notion that relying just on volume indicators is not advisable.

6. Conclusion

Based on the findings from both experiments conducted, it is evident that the baseline exhibits a superior return compared to the other groups. This outcome can potentially be attributed to the robust upward trajectory observed in the ETFs. Despite the fact that our study generates a surplus of returns, there are still areas in which we can make improvements. The dataset has a noticeable bias, specifically a prominent upward trend, which may lead one to believe that a straightforward purchase would yield favorable outcomes. In the future, there will be a greater use of intricate datasets to evaluate the performance of our model in more intricate scenarios. In addition, our analysis encompasses the utilization of 16 established indicators. However, it is evident that these 16 indicators alone are insufficient to comprehensively elucidate the whole of the market dynamics. In this study, we intend to incorporate a greater number of indicators for user input. Additionally, we aim to develop novel indicators that offer enhanced explanatory power within the market context. Finally, we will attempt to enhance the XGBoost algorithm in order to get a higher rate of return compared to the existing implementation of XGBoost.

DOI: 10.54254/2754-1169/74/20241522

In this prospective scenario, a comprehensive exploration of various indicators and the systematic adjustment of hyperparameters will be undertaken to optimize the XGBoost model for the purpose of constructing portfolios. In addition, we plan to expand our investment pool by incorporating a diverse range of financial instruments into our portfolio, including bonds and equities, in addition to our existing holdings in ETFs. The financial industry is experiencing growth and advancement, prompting ongoing efforts to explore various methodologies for elucidating market dynamics and achieving superior investment returns.

References

[1] *Muggleton, S. (2014). Alan Turing and the development of Artificial Intelligence. AI Communications, 27(3), 3-10. DOI: 10.3233/AIC-130579.*

[2] *Shinde, P.P., Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6).*

[3] *Min-Yuh Day and Jian-Ting Lin. (2019). "Artificial Intelligence for ETF Market Prediction and Portfolio Optimization." In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1026-1033.*

[4] *Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics, 18(6), 275-285. DOI: 10.1002/cem.873*

[5] *Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).*

[6] *Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. Expert Systems With Applications, 186, 115716. https://doi.org/10.1016/j.eswa.2021.115716*