iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53921

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Prediction and Portfolio Optimization in Quantitative Trading Using Machine Learning Techniques

Ganesh Patil[1], Adarsh Salunke[2], Dhiraj Chaudhary[3], Aniruddha Pawar[4], Prof. Kirti Walke[5]
[1, 2, 3]*Undergrad Student, Dept. of Information Technologies SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra*
[4]*Undergrad Student, Dept. of Computer Engineering Sinhgad College of Engineering, Vadgaon, Pune, Maharashtra*
[5]*Asst. Professor, Dept. Of Information Tech, SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra*

Abstract: ***Quantitative trading is an automated procedure in which trading techniques and judgments are performed using mathematical models. Quantitative trading involves a vast spectrum of computational methods, such as statistics, physics, or machine learning to diagnose, forecast, and benefit big data in finance for acquisition. This work analyses the body components of a quantitative trading technique. Machine learning presents many consequential benefits over conventional algorithmic trading. Machine learning executes numerous trading techniques consistently and acclimates to real-time demands. To illustrate how machine learning techniques can meet quantitative trading, linear regression and asset vector regression pinnacles are utilized to predict stock tendency. In accumulation, numerous optimization strategies are used to optimize the recovery and manage hazards in trading. One typical attribute of both forecast measures is they virtually executed short-term predictions with high precision and repayment. However, in the short-term forecast, the linear regression instance outmatches corresponded to the support vector regression model. Predictability is extensively enhanced by adding specialized needles to the dataset, preferably by accommodating cost and importance. Despite the gap between prediction modelling and authentic trading, the suggested trading technique accomplished a higher retrieval than the S&P 500 ETF-SPY.***
Keywords: Trading technique, Stock forecast, Portfolio optimization, Machine learning, Quantitative trading, Portfolio optimization.

I. INTRODUCTION

Quantitative trading, known as algorithmic trading, is the technique of purchasing and dealing inventories/ acquisitions established by executing trading techniques in a disciplined and systematic way. These trading techniques are conceived via relentless analysis and mathematical analyses. When quantitative trading techniques were first familiarized, they were intensely beneficial and swiftly acquired market share. Heightened systematic trading (HFT) accounted for an elevated allocation of day-to-day trading magnitude with millions of shares. But as a competitor has risen, returns hold diminished unhurriedly. As the tribulation of seeking returns in trading additions, machine learning is considered the most powerful tool to acquire a competitive benefit and facilitate asset return. Machine learning presents several significant advantages over quantitative trading by furnishing a combination of influential computer algorithms that allow one to move from encountering relationships/designs established on documented data to determination and acclimating to market movements in a periodic style. The algorithms understand to use of predictor variables to predict the prey variable. An assortment of trading techniques uses machine learning for trading analysis, including linear regression, neural networks, reinforcement vector machines, and deep understanding. Among these techniques, regression is a familiar and most naturally used machine learning technique in finance, such as deserving forecast, distinction licking forecast, and stock market prediction. In this paper, we use linear regression and support vector regression to achieve stock tendency prediction. Based on the predicted results, multiple philosophical portfolio distributions and optimization strategies are utilized to optimize portfolio retrieval.

II. QUANTITATIVE TRADING SYSTEM

Quantitative trading can be described as the periodic undertaking of trading techniques that mortal beings assemble via specific analysis. In this context, occasional is described as a disciplined, methodological, and computerized technique. Individuals accomplish the research and determine which courses will be used to perform on the paradises of stocks for trading techniques. Those people after quantitative trading techniques are called quants or quant traders.

The concept of quantitative trading is organized to influence statistical computation, computer algorithms, and computational resources for continuous trading systems, which aspire to underestimate risk and maximize recovery established on the chronological implementation of the encoding techniques experimented against chronological economic data. Most acquisition techniques are designed and enforced by mortal beings in which decisions are driven by psychology and sentiment.

In discrepancy, quantitative trading is designed to eradicate arbitrariness by assembling disciplined trading techniques experimented with by a computational representative. A conclusion navigated by emotion, indiscipline, passion, greed, and fear will be taken out of the quantitative investment process. Moreover, investment banks use quantitative trading, a complex mechanism to derive business investment decisions from insightful data such as Goldman Sachs, Morgan Stanley, etc. Quantitative trading involves using complex mathematics to buy and sell orders for derivatives, equities, foreign exchange rates, and commodities at a very high speed. Quantitative trading techniques are acquisition techniques established on quantitative computation of financial markets and forecast of prospective enactment. The strategy and associated predictions depend on the time scale of the investment, as exemplified by the following classes of quantitative strategies: fundamental analysis when a stock is trading under intrinsic value. The long-term strategies motivated quant has a quarterly timescale periodic macro established on macroeconomics estimation or market circumstances and tendencies to determine acquisition prospects. Periodic macro techniques are model-based and accomplished by software with determinate mortal involvement. The above systematic macro has a monthly timescale. Intersection or relative value trades and other statistical arbitrage (StarArb) techniques refer to trading in similar assets predicted to congregate in significance. These illustrate StatArb techniques, with time hierarchies meandering from minutes to months. High-frequency trading (HFT) is known as purchasing/vending an enormous number of shares in a brief period.

The time scale of HFT in milliseconds and the holding period of the traded shares is usually less than one second. A typical quantitative trading system has three modules: an alpha model, a risk model, and a transaction cost model, which provide a portfolio structure bar, which in turn interacts with the implementation bar, as shown in Figure 1. The alpha bar is contained to predict the future of the pawns the quant wants to believe in trading to yield retrievals. For example, in a trend-following technique in the futures markets. On the other hand, the alpha bar is conceived to predict the tendency of whatever future demands the quant has decided to comprise in a strategy. Hazard bars, by disparity, are designed to limit the amount of vulnerability the quant has to those aspects dubious about generating returns but could drive casualties. For example, the trend follower could choose to restrict his directional exposure to a given asset class, such as commodities, because of concerns that too many forecasts he observes could streak up in the exact approach, conducting to surplus hazard; the hazard bar would contain the levels for these commodity orientation boundaries. The transaction expense bar is used to help resolve the expense of whatever trades are required to relocate from the current portfolio to the new portfolio that is sensual to the portfolio structure bar. Almost any trading transaction costs capital, whether the retailer desires to profit enormously or inconsequentially from the trade. The alpha, hazard, and transaction cost bars forage into a portfolio structure representative, which counterbalances the trade-offs demonstrated by the pursuit of profits, the limiting of risk, and the costs associated with both, thereby specifying the most profitable portfolio to hold.

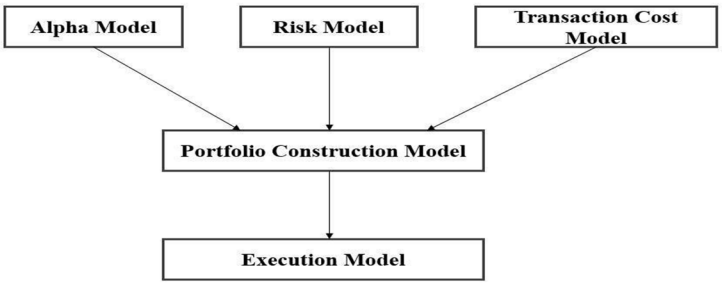


Figure 1: Quantitative trading modules.

Quantitative trading strategy workflow consists of 6 stages: data collection, data pre-processing, trade analysis, portfolio construction, back-testing, and execution as shown in **Figure 2.**

1) Stage 1: Data Collection

Multiple data sources can be collected through finance data API such as Morning Star, Quandl, Google, Yahoo Finance API or provided by security companies. Stock data are collected under various types of formats. It could be under fundamental, technical, macroeconomics, or even sentiment data in text form with variety of time-scales.

2) Stage 2: Data pre-processing

Collected data could be time series data, non-stationary data, unstructured data under the text or missing data. Therefore, it requires a heavy task for data normalization, scaling, and transformation in order to aggregate all the data sources. Prediction and Optimization Portfolio in Quantitative Trading Using Machine Learning Techniques

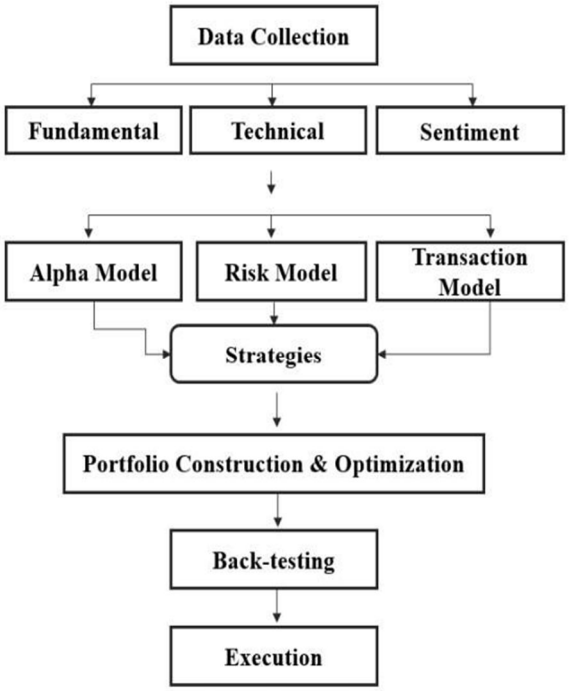


Figure 2: Typical quantitative trading workflow.

3) Stage 3: Trade Analysis

The prediction model is the core of the trade analysis stage, in which alpha, risk, and transaction models are implemented to develop effective trading strategies. A theory-driven or data-driven approach can conduct an alpha model. The theory-driven method accesses the cleaned data from fundamental, technical, macroeconomics, and sentiment data as the input for modelling. Based on the collected, the alpha model aims to predict or forecast stock price movement and then generate some passive trading strategies, which can be used for back-testing before execution. The data-driven approach ingests the data sources in real time and interacts with historical data. A data-driven approach mainly focuses on trading patterns as indicators and signal generation. They scanned the signals/ indicators in real-time that can generate real-time strategies for a trading system with widespread trading. All quantitative trading systems seek to archive high alpha as profit. However, from time to time, risk exposures will not generate profits as expected, but the risk can impact the return of trading strategy daily. The risk model enables quants to define, measure and control risks. There are two naturally obtained paths for estimating the number of hazards in the marketplace. The first measures risk by computing the standard deviation of various instruments' returns over time, known as volatility. The more volatility, the more risk is said to be present in the market. The second way to measure risk is to measure the level of similarity in the behaviour of the various instruments within a given investment portfolio. If all the appliances in a portfolio are flawlessly associated, then as one bet ranges, so go all the other gambles. The transaction cost model is designed to record all trading transaction costs. The transaction cost model has three major components: commissions and fees, slippage, and market impact. Transaction costs are significant to investors because they are one of the critical determinants of portfolio returns.

4) Stage 4: Portfolio Construction and Optimization

Portfolio construction and optimization: Stock investors are interested in combining multiple stocks into a single investment portfolio instead of investing only in a specific store to neglect the volatility of the investment portfolio. This can be done naively by having equal weights for all the stocks or automatically adjusted weights to maximize the portfolio return, which is called portfolio optimization. Portfolio construction starts with three basic questions: reduce the risk, maximize the return, and capital allocation.

5) Stage 5: Back-testing

Back-testing: A critical difference between a traditional investment management process and a quantitative investment process is the possibility of back-testing a quantitative investment strategy to see how it would have performed in the past. Therefore, before implementation, all the quantitative trading strategies are thoroughly back-tested. Back-testing is a simulation of a trading strategy used to evaluate the performance of the proposed method. While back-testing does not allow one to predict how a process will perform in future conditions, its primary benefit lies in understanding the vulnerabilities of a system through a simulated encounter with real world situations of the past Back-testing: A critical difference between a traditional investment management process and a quantitative investment process is the possibility of back-testing a quantitative investment strategy to see how it would have performed in the past. Therefore, before implementation, all the quantitative trading strategies are thoroughly back-tested. Back-testing is a simulation of a trading strategy used to evaluate the performance of the proposed approach. While back-testing does not allow one to predict how a system will perform in future conditions; its primary benefit lies in understanding the vulnerabilities of design through a simulated encounter with real-world situations of the past.

6) Stage 6: Execution

Quants build alpha models, risk models, and transaction cost models. These modules are fed into a portfolio construction model, which determines a target portfolio. But having a target portfolio on a piece of paper or computer screen is considerably different from owning that portfolio. The final stage of the quantitative trading system is to enforce the portfolio judgments created by the portfolio structure model. Although the employment era can be semi-automated or exhaustively automated, the implementation instrument can be manual or fully automated via high-performance application programming interfaces (APIs). For low-frequent trading techniques, manual and semi-manual procedures are standard. For highly systematic trading techniques, it is essential to assemble a comprehensively computerized execution instrument, which will continually be tightly associated with the trade generator. The critical reflections when developing an implementation technique are the interface to the brokerage, underestimation of trade costs, and variation of the arrangement of the live procedure from back-tested performance.

III. MACHINE LEARNING TECHNIQUES

Artificial Intelligence (AI) focuses on intelligent agents that can perceive their environment and take actions to solve tasks that replicate human cognitive functions. To acquire knowledge, AI systems need to learn from raw data. A rational agent should not only perceive but also know as much as possible from what it perceives. Learning is the ability to generalize from data to act optimally with new data. Machine Learning (ML), a subfield of AI, studies the perception, learning, and action tasks as algorithms that learn from data. ML has evolved from multiple fields due to real-world industrial demands for effective methods to handle extensive and high dimensional data. ML has various applications in industries such as natural language processing, computer vision, smart manufacturing, supply chain, and finance. The core part of ML is a computer algorithm. As ML has developed, multiple advanced computational algorithms have been designed to statistically estimate complex functions that are difficult to express in closed form. ML computational models can handle large-scale complex data as big data in a distributed environment and reduce computation time through real-time or stream processing. ML can be defined as a process of learning in which a computer program improves its performance at tasks, as measured by a performance metric, through experience. There are two types of ML tasks: perception tasks, where the ML algorithm learns from the dataset to perform a specific predefined action, and action tasks, where the final goal is to find a decisive action based on what was known from perception tasks. The performance metric, such as the error rate in a binary classification task, is specific to the study. The error rate is the ratio of incorrectly classified examples to the total number of models. It can be viewed as an estimate of the expected 0-1 loss, with a loss of 1 for each misclassified example and 0 for correctly classified samples.

$$Error\ rate = \frac{N_{incorrectclassified}}{N_{total}} \quad (1)$$

$$Accuracy = \frac{N_{correctclassified}}{N_{total}} = 1 - Error\ rate \quad (2)$$

However, such a performance measure is inconvenient in practice because it might change this continuously when the parameters of the models are changed continuously. In other words, such a performance measure metric would be a non-differentiable function of its parameters, so no gradient-based optimization could be applied in this setting. Therefore, a smooth and differentiable alternative to the error rate as a performance measure is considered instead of the probability or low probability of the observed data under the model's assumptions.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

This leads to a differentiable objective function for tuning all the parameters to the data, which can be efficiently done using gradient-based optimization techniques. In regression, one possible choice for performance measure is a mean square error (MSE) as represented in (3). These are the indeed observed outputs and are model estimates for these outputs. The sum runs over all observations so that the MSE is 0 only when all data points are matched precisely, without any errors

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)^2 \qquad (3)$$

The performance measure improves with experience as a result of learning. Here the learning from experience is the ability to generalize.

Learning from experience can roughly fall into three categories: supervised, unsupervised, and reinforcement learning.

In supervised learning, each training data consists of a pair of input objects and the desired output value or target. The main objective is to produce a function that will map input values to the output value so that when unknown new data is fed, the procedure enables us to make a reasonable prediction about the target value.

This type of learning is supervised learning because training data sets are predefined and can be considered an advisor overseeing the learning process. Unsupervised learning is the second one, where the expected outcome is not defined. Unsupervised learning refers to a broad array of machine learning algorithms that tempt hypotheses from information clusters consisting of infusion data without tagged comebacks.

This implies that the algorithm is not presented with the correct output for a sample input but is forced to learn the right way to produce a result in an unsupervised manner. Unsupervised learning primarily forms a significant part of learning for the human brain and hence is an important segment of machine learning. Reinforcement learning involves techniques that try to retro-feed the model to improve performance. To accomplish this, the model needs to interpret signals, decide on an action and then compare the outcome against a predefined reward system. Reinforcement learning tries to understand what needs to be done to maximize the rewards.

A. Machine Learning Taxonomy

Supervised and unsupervised learning are about perception tasks because an algorithm should perform one particular action in both of them.

For example, classify an image, spam email, and so on. On the additional indicator, underpinning understanding is all rough activity assignments. The initial appointment, commonly encountered in machine learning, is a regression to learn a real-valued function, a map of a dimensional space f: $R^n \rightarrow R$, be given the training set of input-output pairs (X_i, Y_i), where Y are an actual number and a vector.

The regression model can be used for sale demand forecasts and similar tasks. Another prevalent supervised learning task is classification, with the objective nearly the same as regression except that in the category are discrete numbers rather than continuous variables as in regression.

For Prediction and Optimization Portfolio in Quantitative Trading Using Machine Learning Techniques classification, typical industrial applications are spam detection, image recognition, or document classification. In unsupervised learning, clustering is one of the most popular tasks known as segmentation in business.

A dataset divides into clusters, or partitions, in which a collection is a set of a homogenous group of data points. In terms of functions, it looks the same as the classification, except that no class labels are given to the algorithm. Clustering strategies are constantly utilized for patron segmentation or irregularity detection. Another category of unsupervised understanding algorithms is dubbed expression understanding.

This term retains many additional approaches whose main idea is nevertheless always the same. Especially this task is to map initial and dimensional data where N can be a large number and to a lower dimensional space of dimension k where $k \leq n$. These assignments are also directed as dimensional decline or future schooling assignments. Among other things, drawing learning methods are used for text recognition and machine translation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

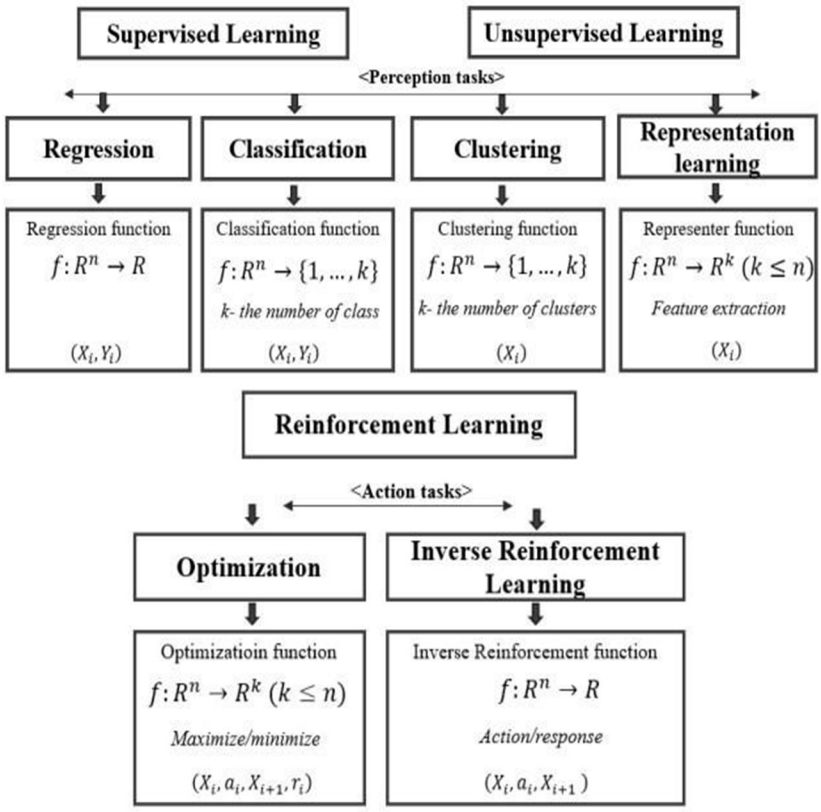


Figure 3: Landscape of machine learning.

The last category of the machine learning algorithm is reinforcement learning. A function that is optimized in enforcement lowering is called the policy function. This function describes what an algorithm should do by giving the current state of the dataset to maximize its total reward over some time. In this case, the training data should be in the form of tuples that contain the current state action taken next state 5) in reward from taking this section. In addition to direct reinforcement learning, there is also a varying interest in the alternative formulation known as inverse reinforcement learning (IRL). In this method, everything is the same as for direct reinforcement learning, but there is no information on rewards received by the agent upon taking actions. Reinforcement learning methods are widely used in robotics and computational advertising. The landscape of ML is shown in Figure 3.

B. Model Selection

Model selection is an integral part of machine learning. It does not mean different choices from the model type, like selecting Knearest neighbour, Naïve Bayes, Trees, SVM, etc. Model selection is choosing different complexity levels, called complexity selection. Here, the complexity can be expressed by the flexibility to fit or explain data. A complex model with more features tends to have a low bias because these models are more flexible and have more capacity to adjust data. However, the flip side is that adding more features would generally increase variance. In contrast, a simple model with fewer features tends to have high bias and low variance. Therefore, building a suitable model requires the right model complexity level matching the data complexity to optimize the trade-off. In some cases, the right level of model complexity and architecture could be established beforehand, guided by some data characteristics. For example, classification is commonly used for lower dimensional data, and neural networks are often used for high dimensional data. However, no single machine learning classification algorithm can be universally better than the others in all domains [15]. The advanced computational resources allow for building a complicated representative with subordinate tendencies and increased friction. One possible way to reduce the conflict is to normalize or bind somehow the value of model parameters so that the model output would vary less with a variance of input data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

IV. 4 MACHINE LEARNING FOR STOCK PREDICTION AND PORTFOLIO OPTIMIZATION

A. Machine Learning for Stock Prediction

A machine learning algorithm allows the system to search for trading patterns within complex data sets such as fundamental, technical, and even sentiment data. Therefore, machine learning enables quants to build up multiple trading strategies. On the other hand, event driven modelling improves prediction confidence to maximize profit and minimize risk. In a quantitative trading system, the vast universe of stock data puts a heavy task for feature selection or indicator selection, whether based on fundamental, technical, or sentiment analysis. Some popular types of machine learning algorithms for training, such as decision trees, random forest/ gradient boost, neural network (LSTM), and evolutionary algorithms, have been used to improve this task. The wrapper is a way of leveraging a machine-learning algorithm to select and evaluate indicator subsets. Wrapper techniques investigate the relationships to find the best collection of indicators instead of looking at hands individually. Chorus understanding is the path to unite considerable uncorrelated classifiers to develop an unmarried or more powerful movement. The advantage of using ensemble learning over a wrapper of random forest is that we can use a lot of different classifiers to find patterns and information from data. In addition, combining pattern recognition and association rule learning is a compelling way to leverage machine learning algorithms while still being able to interpret the output and trading strategies. Since the stock prediction model is considered the core component of the alpha model in quantitative trading. The prediction model investigates the input data, such as technical, fundamental, macroeconomic, or sentiment data. The outcome predicted values could be the input of the portfolio construction. To demonstrate the simple quantitative trading system, linear regression and support vector regression in Scikit-learn are used to simulate how quantitative trading performs. Three optimization models are used to optimize the system performance with the target of maximizing return and minimizing risk.

In this work, X is the set of N trading stocks in the large dataset, denoted as X = $\{X^T_1, X^T_2, \ldots, X^T_N\}$, where is the set of stock indicators or features which can be fundamental, technical, or sentiment indicators, the higher value of T will increase the dataset dimensions. Sample data can be collected by different K -factor (hourly, daily, weekly, monthly) rates.

In this work, regression models have been used to predict the tendency of commodities for a considerable duration of time $[t_0, t_k]$. The dataset is separated into exercise and test sets, and then fit the training set is to the regression models. As a result, the set prediction accuracy P for X is calculated by (3), denoted as P= $\mathbf{\{P_1, P_2, .., P_N\}}$. P bar is the prediction accuracy on average for each prediction model, which will be the key to evaluate both prediction models. Then cumulative return for each stock is calculated for the predicted period $[t_0:t_k]$, denoted as R =$\{r_1, r_2, .., r_N\}$. To construct the portfolio, M(M<N) stock with an accuracy rate greater than average predicted accuracy P bar And highest cumulative return in the forecast set is selected to construct a portfolio

B. Portfolio Construction and Optimization

1) Portfolio Allocation

An investment portfolio is just a set of various securities or stock allocations. Key statistics of a portfolio: daily return, cumulative return, average daily return, standard daily return. Daily or cumulative return may indicate as one part of the investment reward and evaluate the performance of a portfolio. However, the high return may come with high volatility or risk in investment. Sharpe proportion is a criterion for estimating risk-adjusted retrieval, and this ratio has become the industry standard for such calculations. The following relation defines the Sharpe ratio:

$$SR = \frac{R_p - R_f}{\sigma_p} \qquad (4)$$

Where Ais, the expected portfolio return, is a risk-free return. BA is portfolio standard deviation. Risk-free return is the return received by putting money in an investment such as a bank savings account, LIBOR, or treasury bond that are essentially risk-free. Because the risk-free is a really low value, we can assume σ_p as 0. The Annualized Sharpe Ratio (ASR) can be obtained by multiplying against a Kfactor based on the sampling rate. The ASR is calculated as follows:

$$ASR = \sqrt{k}\frac{R_p - R_f}{\sigma_p} \qquad (5)$$

2) Portfolio Optimization

Instead of allocating equal weights for all stocks in the portfolio, we can use optimization techniques to maximize some performance measures, which is called portfolio optimization. Monte Carlo Simulation (MCS) works by randomly assigning a weight to each security in the portfolio, then calculates its mean daily return and standard deviation of daily return. Sharpe ratio is calculated and selected through thousands of randomly selected allocations. Another approach for portfolio optimization uses a mathematical calculation called Efficient Frontier, as known as Mean-Variance Optimization.

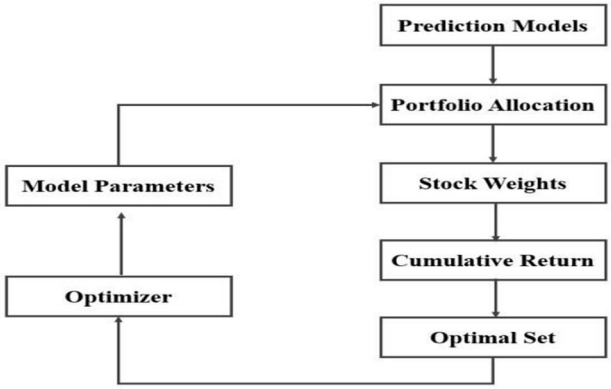


Figure 4: Optimization method.

Assume S is the set of stocks in the constructed portfolio, denoted as S = {X1, X2, …, XM }. The recovery of the portfolio is estimated as follows:

$$R = \sum_{i=1}^{M} w_i X_i \qquad (6)$$

Where W = {w1, w2, .., wM } is the set of weights as the allocation ratio for the portfolio. The optimization portfolio's objective is to find the optimal set Win order to maximize portfolio return in (6). A general portfolio optimization method is shown in Figure 4. Based on the stock prediction models, the prediction results are input for the portfolio allocation model to generate initial stock weights then the cumulative return of the portfolio is calculated to obtain an optimal set as a fitness measure of a portfolio. To archive optimal results as long as finding the best model parameters of allocation through the optimizer. Forecast and Optimization Portfolio in

Quantitative Trading Using Machine Learning Techniques.

V. EXPERIMENT AND RESULTS

A. Data Preparation

In this work, we take 500 stocks data from S&P500 as our predicting data, which are represented by 500 identical stock tickers. The 10-year daily historical data are collected by Quandy API from Jan 1st, 2008, till Mach 27th, 2018, with 2576 total trading days.

B. 5.2 Experimental Design

In this simulation, two regression models, linear regression (LR) and support vector regression (SVR), are used to predict stock movement for different periods, as shown in Table 1.

Table 1: Prediction time periods

Period	Start	End	Days
1	2018-02-20	2018-03-27	26
2	2018-01-11	2018-03-27	52
3	2017-04-12	2018-03-27	78
4	2017-10-25	2018-03-27	104
5	2017-09-20	2018-03-27	129

Based on the predicted accuracy for each trading period, four tickers with high predicted accuracy and the highest predicted return are selected to construct a monthly portfolio. The test size ratio used in both regression models is 1:4 (25%).

In order to evaluate the impact of technical indicators on the predicted result, moving average indicators for 9, 20, 50, and 150 days are used. Basic statistical values: High-Low Percentage Changed (HL_PCT) and Open–Close Percentage Changed (OC_PCT) are also added to the main stock's features. Therefore, there are two different datasets used to evaluate the performance of both regression models. One is 500 stocks' data with the main features. Another is the main features and the added technical indicators.

First, we calculate the prediction accuracy on average on each dataset, then select the model based on the higher prediction accuracy P . We construct a portfolio by selecting four stocks with the highest predicted return as well as pried cited accuracy higher than the average $\bar{P}$.

In order to maximize the return and minimize the volatility of the constructed portfolio, we use three different portfolio optimizers: Equal-weights (EQ) portfolio does not require any optimization technique since all the stocks in the portfolio are equally weighted. Monte Carlo Simulation (MCS) is used to find the optimal weights through thousands of randomly selected allocations. Mean Variance Optimization (MVO) is used to find an adaptive weights portfolio that adapts the stock weights using the prediction models. The mapping is a mathematical function whose parameters are found through optimization, as shown in Figure 4.

C. Experiment Results

1) Model Selection

As shown in Figure 5, the prediction accuracy on average decrease gradually by the prediction periods for both regression prediction models. In each period, the predicted accuracy is slightly higher with the support of technical indicators. Therefore, adding more features, such as technical indicators, could improve the predicted accuracy. In short-term trading, the LR model obtains a higher accuracy than SVR one. However, SVR predicted accuracy tends to be higher in long-term prediction.

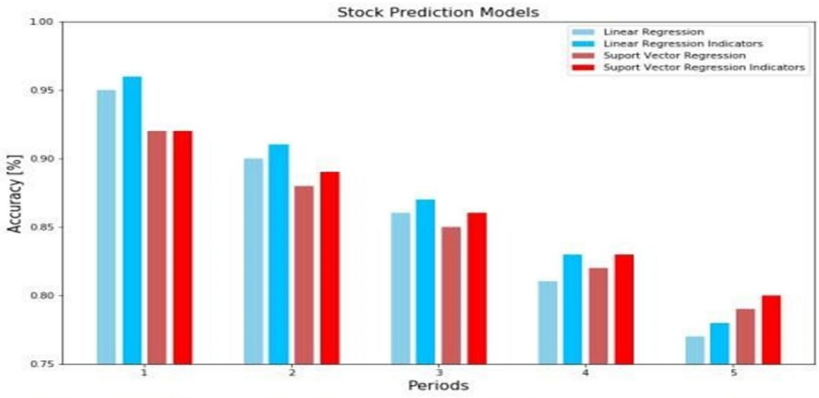


Figure 5: The prediction accuracy for regression models.

Table 2. Portfolio performance measurement results

Period	Optimizer	Prediction		Actual		S&P500 SPY	
		SR	R(%)	SR	R(%)	SR	R(%)
1	EQ	0.25	21.9	0.31	18.7	(1.77)	(3.6)
	MCS	0.55	22.4	0.32	13.7		
	MVO	0.55	22.4	0.33	13.7		
2	EQ	0.24	45.9	0.33	45.5	(1.14)	(5.2)
	MCS	0.89	47	0.23	38.5		
	MVO	0.96	47.3	0.25	38.5		
3	EQ	0.20	76.9	0.33	66.6	0	(0.4)
	MCS	0.38	58.7	0.31	57.0		
	MVO	0.39	59	0.31	57.4		
4	EQ	0.10	62	0.12	13.7	0.54	3
	MCS	0.12	78.7	0.10	13.9		
	MVO	0.13	79.7	0.11	13.7		
5	EQ	0.06	31.2	0.06	9.0	0.76	5.2
	MCS	0.08	61	0.07	9.8		
	MVO	0.08	61.6	0.07	9.6		

Portfolio Evaluation

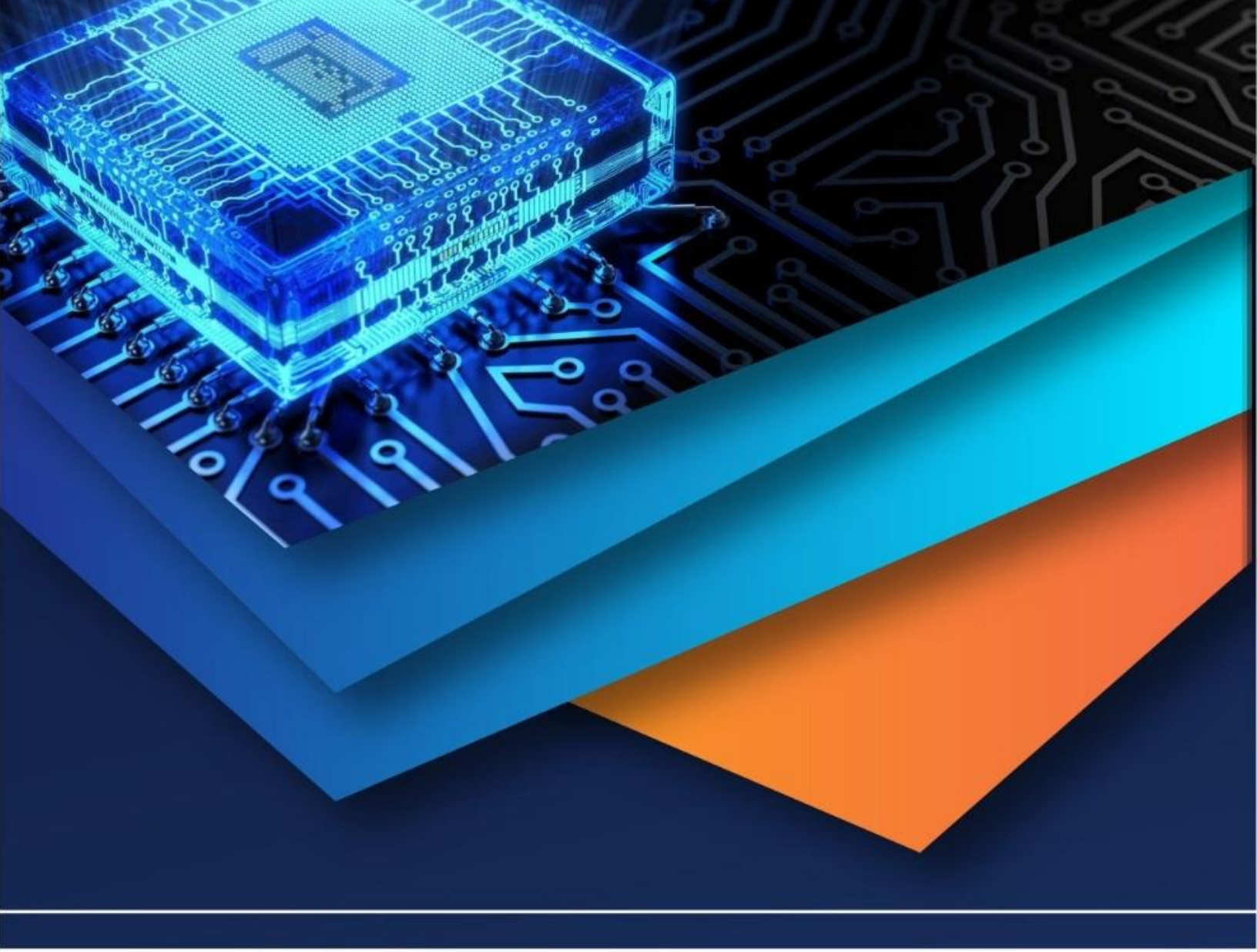
Because LR obtains the higher predicted accuracy on average in the predicted periods 1,2 and 3, then portfolios 1,2 and 3 are constructed based on the LR prediction results. Portfolios 4 and 5 are constructed based on the SVR prediction results. As presented in Table 2, the short-term portfolio investment obtains a higher Sharpe ratio (SR) and cumulative return. However, the predicted SR tends to be lower and less attractive in long-term investment. That can be caused by the low predicted accuracy leading to a higher risk or more volatility. The predicted optimal SR and cumulative return are nearly the same in both optimizers (MCS, MVO) and improve the results from the EQ optimizer. Despite the fact that the performance measures in actual are lower than predicted, all the portfolios obtain a profitable return and high SR. Especially portfolio 3 with more than 60 percent cumulative return, while the benchmark S&P 500 ETF-SPY lost 0.4 percent.

VI. CONCLUSIONS AND DISCUSSIONS

This paper presents the fundamentals of the quantitative trading system in terms of system architecture, benefits, and trading workflows. The taxonomy of machine learning techniques and applications in finance are also described in this work. Machine learning plays an important role and becomes the most powerful tool to build up trading strategies by improving prediction accuracy. In the experiment, both linear regression and support vector regression models are used to predict the stock price. As a result, both regression prediction models are shown effectively in prediction with high accuracy on average. The linear regression model performs better than support vector regression in short-term prediction. However, the support vector regression model tends to perform better than linear regression in long-term prediction. Using indications should improve prediction accuracy. To evaluate the confidence of the prediction results, five different portfolios are constructed. Despite the fact that the Sharpe ratios and returns are not exactly as predicted, the proposed strategy seems to work pretty well by achieving attractive returns and high annualized Sharpe ratios in shortterm trading. All the portfolios exist accomplished additional actually corresponded to S&P 500 EFT-SPY. Back-testing and diversification are considered the target for further research. There are several challenges to building effective quantitative trading strategies through machine learning. First, market data is non-stationary, while most machine learning techniques assume the datagenerating processing is stationary. Second, market data exhibit a high noise-to-signal ratio. The forecast measures can execute agreeably on the chronological information cluster. However, the stock market is always fluctuated by many factors, such as market psychology, macroeconomics, and even political issues. Therefore, high performance on the historical dataset does not guarantee to earn a desirable profit in practice. Third, back-testing is not only the tool to evaluate the discovered strategy but also helps to avoid false positives. Finally, developing flexible, efficient trading strategies is critically important for quantitative trading. It is the most challenging task in the quantitative trading system since the diversity of data sources and formats and the different characteristics of data. That will make the prediction getting more complex. In summary, the advanced computational resources and machine learning enable to design of an efficient quantitative trading system.

REFERENCES

[1] S.Y. Wang, Y. Nakamori, W. Huang(2005). Forecasting supply market tendency approach with approval vector apparatus, 32(10),
[2] F.C. Park, C. Han, E. Chong(2017). In-depth understanding grids for stock demand research and forecast: Procedure, data representations, and case studies, 83(),
[3] R.K. Narang. 2009. Inside the Black Box: The Straightforward Truth regarding Quantitative Trading. New Jersey, Wiley Finance Press Chapter 1.
[4] Quantitative Trading: How to Create Your Algorithmic Trading Industry (1st ed.). E. Chan. 2008, Wiley Press, Chapter 3.
[5] V. Lalchand, M. Galas, P. Treleaven(2013). Algorithmic Trading Review, 56(11),
[6] P. Norvig, S. Russell. 2009. Artificial Intelligence: A Modern Approach (3rd ed.). Pearson Press. Chapter 1.
[7] A.Y. Ng, A. Saxena, J. Michels(2005). High-speed obstruction release utilizing monocular concept and underpinning wisdom. In cruising with the 22nd global discussion on Machine learning. German,
[8] P. Barham, M. Abadi el.(2016). TensorFlow: A system for large-scale machine learning. In proceeding to the 12th international conference on operating systems design and Implementation. USA, pp.
[9] C. Fermüller, Y. AloimonosD. Summers-Stay, P. Wiriyathammabhum (2016). Computer Concept and Natural Language Processing: Contemporary Tendencies in Multimedia and Robotics, 49(4)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY