Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 17, No. 2, 2024

Machine Learning-Driven Stock Selection System: A Proposed Multi-Factor Model-Based Navigation Tool for the Chinese Financial Market

Yuanfu Wu

Adamson University, Manila, 1000, Philippines

Abstract: In view of the constantly changing and volatile nature of the Chinese financial landscape, characterized by the introduction of high-tech and huge amounts of market data, there is a need for an improved method for stock selection. Conventional stock methods are, however, inadequate due to their tendency to ignore the high volatility of the market and the nonlinear relationship between many financial variables which are at play. They have not been able to deliver predictions that are strong and accurate. This research, therefore, aims to explore how machine learning algorithms can be utilized to improve the accuracy and adaptability of stock selection models, thereby addressing these limitations. Through empirical results, it can be claimed that ML models, especially tree-based algorithms like LightGBM and XGBoost, are superior to traditional models in predicting stock returns. In this case, rolling windows are very well adapted to machine learning methods, which stand out as superior in terms of the adaptability to the changing stock market conditions. In addition, the research reveals the necessity of integrating the market-driven data, for instance, the trading volume and the momentum, into the stock selection models can help to better capture short-term pricing dynamics. The dissertation concludes with a machine learning-based stock selection system that is optimized for the Chinese financial market. The results of this work will increase the knowledge of quantitative finance by showing the machine learning algorithms can be more accurate in stock selection, and at the same time, providing practical solutions to the investors and financial institutions. The research emphasizes the necessity of combining modern computational methods and traditional financial theories to devise more effective and adaptive investment strategies in a complicated financial environment.

Keywords: Machine Learning, Stock Selection, Multi-Factor Model, Chinese Financial Market, Quantitative Investment Factors, Tree-Based Models, Predictive Accuracy, Risk Management.

1. Introduction

1.1. Background of the Study

The present financial markets are a real maze that demands the investors to navigate through the labyrinth of multiple stock alternatives amidst the surroundings that are both complex and dynamic. According to this complicated setting, there are a lot of things that need to be counted, such as the world economy, technological development, geopolitical risks, and changing investor opinions, among others. These elements, in turn, are the ones that cause the financial market to keep on changing and revolutionizing, thus bringing a lot of trouble to the investors who are trying to stock select their decisions based on the information they have (Feng, Seasholes, & Zhang, 2020). The particular nature of the Chinese financial market is the main factor why these difficulties become even more serious, as well as the market's role in the global economic framework).

Stock picking methods that are stick-to-our-roots and rely solely on a limited number of financial ratios like the price-to-earnings and the price-to-book are not appropriate in today's rapidly changing and complicated market environment. These models, by limiting themselves to a narrow range of factors, do not take into consideration the variable and changing nature of market data. We reside in a time characterized by an excessive amount of information, where market data and information proliferate at an irrationally fast pace. The weaknesses of conventional models become manifestly clear. They are not capable of detecting and analyzing market changes thoroughly and timely, thus, they are not very effective in directing the investment decisions (Li & Zhang, 2019; Liu & Wu, 2021).

The stock selection models have so many shortcomings that it is hard to count them all. To start with, there is the tendency of the models to think too highly of the individual factors or a certain group of the financial indicators which limit them to only a part of the market's multifaceted properties (Tsai, Gao, & Yuan, 2023). This is too simple and does not consider such factors as the macroeconomic indicators, the fundamentals of the company, and the market sentiment that can influence the stock prices. Besides, the models do not adapt to the rapid changes that characterize the present financial markets. They frequently do not reflect the present market conditions or the new data in real-time, thus, the analyses are out of date or inaccurate. On top of that, the traditional ones have not the competence to manage the existing vast data properly. In the time of information overload, the ability to filter out, analyze, and get insights from massive datasets is something very important. Still, traditional models lack the computational power and sophistication that data deluge requires, thus, there is a big discrepancy in their analytical capabilities (Arnott, Clements, Kalesnik, & Linnainmaa, 2019; Chen, Tang, Yao, & Zhou, 2019).

Machine learning's incorporation in stock selection might be the best way to solve those issues. Recent advances in data technology and computer capacity have enabled the use of machine learning algorithms in banking. No doubt, machine learning is the main alternative that can find out non-linear relationships between many different variables and the

market. They are more flexible in choosing stocks as they can change with new information and different market conditions (Zhang & Zheng, 2019; Li et al., 2020). A broad range of data sources and advanced analytical techniques are among the reasons why machine learning algorithms give a more complete and deeper view of the market data. Hence, they ensnare the complexity and diversity of financial markets easily and thus, stock selection models achieve higher accuracy and robustness (Wu et al., 2021; Yang & Wang, 2020).

Although a large volume of research articles on stock markets and models used for stock selection has been published, the actual challenges that the Chinese market brings as well as how to tackle them is still being understood. Due to the limits of traditional stock selection methods and the promise of machine learning algorithms, empirical study is needed. This study examines Chinese stock selection using a Multi-Factor Model based on Machine Learning to fill the research gap. The objective is to equip investors with an efficient and flexible tool for navigating the intricacies of the Chinese financial market, thus augmenting the existing knowledge in the field of finance.

In its turn, this context provides a basis for the further investigation of the empirical research concerning stock selection through a Multi-Factor Model based on Machine Learning as a potentially effective navigation tool in the Chinese financial market The research aims to look at the way the traditional stock selection models' weaknesses and the machine learning algorithms' advantages could be complemented to form a new perspective on stock selection strategies in the context of the rapidly changing Chinese financial landscape.

1.2. Related Literature

The literature analyzed together illustrates the defects of the conventional stock-picking approaches in the atmosphere of the ever-changing Chinese financial market. One common thread that can be found in the studies including Zhang, Wang, Li, and Shen (2019), Huang, and Luk (2020), Li, Miao, Zhang (2020), Sun, and Tong (2020), and Feng, Seasholes, and Zhang (2020), is the problem of standard models that are not able to use the market information efficiently and are not able to adapt quickly to the changes in the market. This work follows others in using a Multi-Factor Model based on Machine Learning to solve restrictions.

A major split in the literature can be seen through the study of stock selection based on individual factors. The authors Scorpio, Henrique, Sobreiro, and Kimura (2019) and Gandhmal, and Kumar (2019) claim that traditional models using simple factors are insufficient, whereas Nti, Adekoya, and Weyori (2020) and Tsai, Gao, and Yuan (2023) suggest an augmented multi-factor model as an alternative solution. The latter studies argued for this method, which included the fundamental data and market technical indicators and hence provided a more comprehensive view. This divergence is inextricably linked to the current study that aims at applying machine learning algorithms to stock selection by encompassing a variety of factors, thus improving the adaptability and accuracy of the model.

The information overload effect on traditional methods was found as a central point of concern in the literature, as described by Li and Zhou (2021), Nguyen, and Nguyen (2021), Gao, Tsai, and Yuan (2022), and Michis (2022). These studies underscore the advantages of the surge of market data as well as the deficiencies of classical models when it comes to information overload issues. The present research is in line with this narrative since it recognizes of the need existence for more advanced methods and suggests machine algorithms learning as a possible solution for the efficient processing of a huge amount of data.

Machine learning algorithms have stock selection power that has become a point of discussion in several studies such as Li and Deng (2019), Carta, Medda, and Reforgiato Recupero (2020), Jiang, Lee, and Gao (2020), Mishev, Gjorgjevikj, and Lameski (2020), Wu et al. (2020), Wang, Zhuang, and Feng (2022), and Shen and Shafiq (2020). The experiments demonstrate machine learning models' efficiency, efficacy, and adaptability, which allow them to handle conventional challenges. Also, this research contributes to the expanding body of data supporting machine learning algorithms in stock selection.

Factoring all the considerations by the machine learning algorithms is still another significant issue in the literature and this was elaborated Kaczmarek and Perez (2022), Pramanik and Jana (2022), Guo (2023), Hanauer and Kalsbach (2023), Hoang and Wiegratz (2023), and Tan et al. (2023). The researches solidify the position that machine learning models are capable of taking into account different factors simultaneously and as a result stock selection is done in a more nuanced and complete manner. The current study is in line with this view, which highlights the possibility of machine learning algorithms providing a holistic perspective of market dynamics.

Lastly, the examination of nonlinear relationships in financial markets is taken up by Abe and Nakagawa (2022), Cabrol, Drobetz, Otto, and Puhan (2023), Cakici, Fieberg, Metko, and Zaremba (2023), Chen, Cho, Dou, and Lev (2022), and Daul, Jaisson, and Nagy (2022). These studies bring the new ideas of ensemble learning, predicting corporate bond illiquidity, machine learning goes global, predicting future earnings changes, and performance attribution of machine learning methods. This study agrees with this creative method, recognizing that the temporal dynamics must be captured and the unconventional data sources should be explored to improve model adaptability and accuracy.

In summarizing the discussed themes across the literature, it is quite clear that while the potential of ML in financial applications is increasingly recognized, a methodical, comprehensive investigation of ML based multi-factor models in the Chinese stock market is still lacking. This dissertation intends to bridge this research gap by constructing multi-factor models based on different ML algorithms and rigorously testing their performance against conventional models in predicting asset returns in the China market. This endeavor is designed not only to announce the superior predictive power of ML algorithms but also to identify which of them is the best performer in the case of the Chinese market which is a new contribution to the quantitative finance field.

2. Methods

2.1. Research Design

This dissertation uses machine learning to create an accurate multi-factor model to better navigate the Chinese financial market. The methodology includes a systematic data collection, a model selection, and a testing the proposed tool against the traditional methods approach.

The study employs high-grade financial, economic, and transactional data from three important databases namely CSMAR, Wind, and Tushare. These platforms are often regarded as the best sources of datasets that are complete, detailed, and professional, and contain such essential information as the financial statements of A-share listed companies, trading data, and other selected macroeconomic indicators. The data covering the period of January 2007 to October 2023 were collected in order to guarantee the extensive analysis of factor stability over time. This extended time range facilitates the analysis of short-term swings and long-term trends, revealing the Chinese market's dynamic character.

The sample data can be grouped into three main categories, particularly the financial reports, the trading data, and the macroeconomic data. This varied dataset represents the foundation of the subsequent feature engineering processes. Preprocessing steps include standardization of factor values, industry normalization to avoid industry-specific biases. The normalization approach is important in this case because it does not allow the model to be influenced by the scale differences of any single factor and thus it is the essential condition.

Following the methodology of Leippold et al. (2021), who assessed the relevance of 90 stock factors and 11 macroeconomic factors in the A-share market, this study selects 25 key factors based on their importance rankings. The factors that are known in both academia and practice for their predictive capabilities contribute to the model's credibility.

The incorporation of machine learning techniques aims to exploit complex interactions among multiple factors; thus, the more accurate prediction is. The selected machine learning models include a variety of methods: three linear models (OLS, Lasso, and ElasticNet), two types of neural networks (NN4 and NN5), and five tree-based models (Decision Tree, CatBoost, XGBoost, LightGBM, and Random Forest). These models were chosen because of their capability to express both linear and non-linear associations, which is a must for the analysis of the complex financial data.

This broad time period allows study of short-term fluctuations and long-term trends, showing the Chinese market's dynamic nature. The primary reason for including linear models, which are simple and easily understood, in the list of methods for obtaining a first rough estimate of the impacts of the factors is the clear and straightforward nature of the linear models. The machines that neural networks and tree-based models, on the contrary, are the ones providing the most modern skills in modeling non-linearities and interactions which are anticipated to be especially useful in deciphering the complex and rapidly changing Chinese financial market.

2.2. Applicable Machine Learning Models

This section providing some suitable machine learning models to describe the quantitative investment strategies of the Chinese financial market. The selection of the models is done in the light of their effectiveness in coping with the peculiar characteristics inherent in the market like, high-dimensional data, nonlinear relationships, and model stability.

2.2.1. Simple linear regression

In the case of the very complicated Chinese financial market, we have chosen linear regression with the Huber loss function as a main model for our research because of its good prediction quality and simplicity. The model's current significance is particularly enhanced in the conditions when the data is too large and generally distorted, which is usually the case of the financial markets. The mathematical expression for linear regression is represented as:

$$r_{i,t+1} = g(z_{i,t};\theta) + \varepsilon_{i,t}$$

Where $r_{i,t+1}$ is the expected excess return on stock i at time $t+1$, $z_{i,t}$ embodies the vector of selected factors influencing the returns, θ represents the vector of factor loadings, and $\varepsilon_{i,t}$ denotes the error term. The Huber loss is formalized as:

$$H(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H(r_{i,t+1} - g(z_{i,t};\theta); M)$$

Where the Huber loss function H (χ; M) defined by a tuning parameter M, distinguishing between the squared loss for residuals less than M and a linear loss for those exceeding M.

We will build the theoretical basis of our empirical investigation through the implementation of linear regression with the Huber loss. This allows for a simple and interpretable analysis of how traditional quantitative variables affect stock returns. In addition, the insights gathered from this baseline model will be of great help in determining the additional benefits gained from the more complex machine learning models that we have proposed for stock selection in the Chinese financial market.

2.2.2. Regularized linear regression

A salient challenge within simple linear regression is the reliability of estimations when confronted with a large number of covariates. It is especially important to note that we are talking about high dimensional area of Chinese financial markets whose quantitative investment factors usually make the model unstable if not regularized. To combat these issues which are natural in simple linear regression methods, this dissertation combines Regularized Linear Regression techniques mainly focused on Lasso Regression. Adding a penalty function to classical regression restricts coefficient size, lowering the impact of less relevant factors to zero and resolving multicollinearity and enhancing model parsimony. The Lasso objective function is expressed as:

$$L_{LASSO} = L_H(\theta) + \lambda\sum_{j=1}^{p}\left|\theta_j\right|$$

Where λ is a hyperparameter governing the strength of the penalty, providing a means to control the trade-off between bias and variance in the model estimation process.

Complementing Lasso, the Elastic Net method is a hybrid that merges the qualities of Lasso and Ridge regression. In the presence of strongly correlated predictors, it balances the L1 and L2 penalties to stabilize covariate selection while keeping regularization qualities, addressing Lasso's drawbacks. The Elastic Net objective function is given by:

$$L_{Enet}^{H}(\theta) = L_H(\theta) + (1-\rho)\lambda\sum_{j=1}^{p}\left|\theta_j\right| + \frac{\rho\lambda}{2}\sum_{j=1}^{p}\theta_j^2$$

Where ρ modulates the relative weight between L1 and L2 penalties, allowing for a tailored approach that aligns with the underlying data structure and the research objectives.

By implementing Lasso and Elastic Net regression, the dissertation endeavors to overcome the limitations of TMFM

and unfold the intricate fabric of MQIF, thereby forging an advanced navigation tool for stock selection.

2.2.3. Gradient boosted regression trees

Within the decision analytic framework applicable to the financial market, a Decision Tree (DT) serves as the elemental learning unit, partitioning the data space into regions R_m through recursive binary splitting. The goal is to minimize the in-sample prediction error, typically operationalized as the sum of squared residuals within each terminal node, as defined by:

$$\hat{y}_i = \sum_{m=1}^{M} c_m I(\mathrm{x}_i \in R_m)$$

Where $\hat{y}_i$ is the predicted return for stock i, M denotes the number of terminal nodes, c_m represents the mean outcome in region R_m, and x_i signifies the vector of quantitative investment factors for stock i.

CatBoost refines the approach by optimizing the processing of categorical features, significant in the Chinese financial context. It utilizes oblivious trees and introduces an innovative gradient-boosting scheme. The objective function of CatBoost for a given iteration t involves minimizing:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + a^{(t)} h(\mathrm{x}_i, a^{(t)}))$$

Where l is the loss function, $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration, $a^{(t)}$ is the step size, and h is the tree structure determined by parameters $a^{(t)}$, reflecting the decision rules at iteration t.

XGBoost elevates the boosting framework through the introduction of a regularized objective function, incorporating both the loss function and a penalty term for the complexity of the model:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Where f_t representing the decision tree added at the t step, and Ω signifying the regularization term which penalizes the structural complexity of the trees, facilitating the control of overfitting.

LightGBM is tailored for efficiency and scalability in high-dimensional data scenarios, such as those presented by the Chinese financial market. It builds trees by selecting the split that reduces the loss function most, thereby localizing the most significant growth at the iterations where it is most required:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta . g(\mathrm{x}_i; \Theta^{(t)})$$

Where η is the learning rate, and $\Theta^{(t)}$ embodies the parameters defining the structure of the tree at iteration t, and g is the gradient-based learning process that specifies the tree growth.

2.2.4. Random forests

An ensemble learning approach called Random Forests creates a forest of decision trees to help make decisions in the complex Chinese financial market. Mathematically, the predictive model for stock returns, considering a forest of B trees, is given by:

$$R(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x; \Theta_b)$$

Where R is the aggregated prediction, x encapsulates the quantitative investment factors, T_b represents the b-th tree's prediction, and Θ_b embodies the tree-specific parameters, incorporating feature randomization and bootstrapped data sample.

2.2.5. Neural networks

The neural network architecture designed for this research is inspired by the human cognitive process and consists of an input layer that encodes stock-level characteristics, multiple hidden layers that capture intricate interactions among predictors, and an output layer that yields a continuous value indicative of stock performance.

$$\hat{r}_{i,t+1} = a_1 + w_1 \sigma(a_2 + w_2 \sigma(...\sigma(a_{L-1} + w_{L-1} \sigma(a_L + w_L z_{i,t}))...)) + \varepsilon_{i,t+1}$$

Where $\hat{r}_{i,t+1}$ denotes the predicted return of stock i at time $t+1$, $z_{i,t}$ is the vector of quantitative investment factors, σ represents the Rectified Linear Unit (ReLU) activation function applied element-wise, a_k and w_K denote the bias and weight parameters for layer k respectively, and $\varepsilon_{i,t+1}$ is the error term.

Data collection involves training neural network models, which vary in complexity depending to the number of hidden layers and neurons. The methodology corresponds with the analysis of the multi-dimensional characteristic of financial market predictors and their joint influence on stock returns.

3. Results & Discussions

3.1. Out-of-Sample Performance Analysis: Evaluating the Multi-Factor Machine Learning Model

Through out-of-sample performance assessment, the usefulness of the machine learning-based multi-factor model is determined compared to the traditional stock selection methods. The models' ability to forecast for the complete dataset's categories, random samples based on market capitalization, and state-owned and non-SOEs and non-SOEs is assessed here. Table 1 shows that machine learning can explain the complicated Chinese financial sector better than state-of-the-art methods.

Table 1. Monthly Out-of-Sample Prediction R^2 (%)

Category	OLS-3	LASSO	GBRT	RF	NN
All	0.74	1.46	2.76	2.43	2.06
Top 65%	0.22	0.58	-0.37	-0.05	0.42
Bottom 35%	1.46	2.73	7.26	6.09	4.53
SOE	0.51	0.86	0.01	0.81	1.12
Non-SOE	0.86	1.65	3.69	3.03	2.43

3.1.1. Full Sample Analysis

Thinking of the dataset as a whole helps to have a rough idea of the performance of the model as a complete entity. The data suggest that machine learning models are always superior to linear regression approaches in stock returns forecasting. In contrast, the Out-of-Sample R-squared (R^2) values of models like Lasso, Gradient Boosted Regression Trees (GBRT), Random Forests (RF), and Neural Networks (NN) are much greater than the OLS-3 model. In this case, the point is that GBRT is the one that has the best performance with a maximum R^2 value of 2.76%, which shows that it is much better than the linear regression models.

This result is consistent with research by Henrique, Sobreiro, and Kimura (2019), who found that machine learning models outperform conventional techniques in handling the nonlinearities present in financial data. Through the use of tree-based models such as GBRT and RF, it can be demonstrated that these models are able to interact with the data, where RF has an R^2 value of 2.43% and NN has a value of 2.06%. These successes demonstrate how well machine learning algorithms can comprehend the intricate dynamics of the financial market, which is beyond the scope of conventional models (Gandhmal & Kumar, 2019).

3.1.2. Sub-Sample Analysis Based on Market Capitalization

In respect to the diverse group of investors in China, a mix of retail investors and major state-owned enterprises, the model is important to be studied how it works in different market-cap segments. To analyze the model's performance with respect to market size, stocks in the study are divided into large-cap (top 65%) and small-cap (bottom 35%) groups.

The study reveals that the use of machine learning methods, in particular GBRT and RF, has a clear advantage over the methods based on traditional approaches in forecasting small-cap stock prices. For instance, GBRT has a remarkable R^2 of 7.26% for the small-cap class which is much higher than the large-cap group. This is in accordance with Liu, Wang, and Cheng (2019) who were of the opinion that small-cap stocks being more volatile and having less analyst coverage, make it possible for machine learning models to discover hidden patterns in the data. The astounding capacity of these models to learn from the minutiae of the market's movement is evidence of their adaptability to intricate financial settings. Correspondingly, tree-based models have proven to outperform classic regression techniques in both market-cap segments. The RF model having the small-cap category R^2 of 6.09% indicates its effectiveness in the market inefficiencies exploitation (Fama & French, 2020). This strength stems from the model's capacity to analyze vast volumes of data and uncover patterns that linear models often miss, demonstrating the applicability of machine learning in quantitative investing.

3.1.3. Analysis of State-Owned vs. Non-State-Owned Enterprises

When we think about the economic function of SOEs in China we can say that we need to know exactly how the type of ownership affects the quality of the model. The study conducts a differential evaluation of predictive accuracy by segregating the stock sample into SOEs and non-SOEs, which evaluates the respective contributions of these entities to predictive accuracy.

The analysis hints that non-SOEs are the ones that have more flexibility and adaptability and their style is more aligned with machine learning models compared to SOEs. For example, GBRT gets 3.69% for non-SOEs while NN gets 2.43% so it can be said that these models can capture the dynamics of non-state entities (Blitz & Hanauer, 2020). Apart from that, SOEs have R^2 values that are not as high as those of other companies, for example, the 0.01% score of GBRT which shows that their more stable and predictable performance is the reason for them being better fit to the traditional models like OLS-3 (Zhang & Wu, 2021).

3.2. Enhancing Factor Integration and Nonlinear Learning through Adaptive Training Techniques

This section examines how machine learning models can be adjusted to deal with unpredictable market conditions by adding dynamic relationships between quantitative investment factors. The research demonstrates that machine learning models can enhance their capacity to consider both nonlinear relationships and temporal shifts in the Chinese financial market through the use of adaptive training techniques, particularly rolling window strategies.

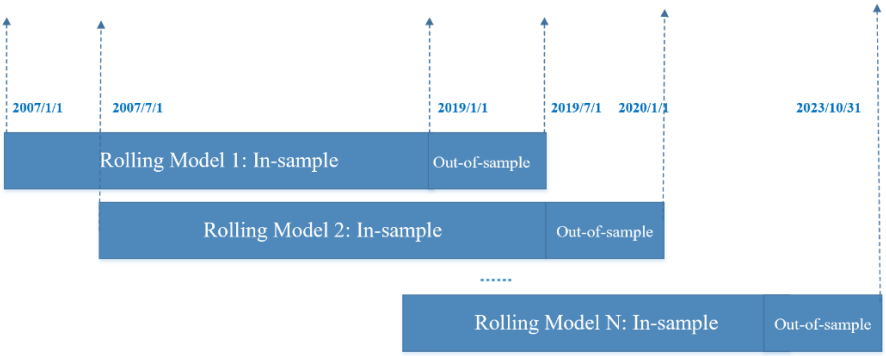

Figure 1. Rolling Prediction Model

This study introduces a time-window method that is essential for financial markets to train ML, i.e., to machine learn with this type of data. The rolling window method sometimes sends the training dataset, thus allowing the models to "forget" old data, and adapt more to the time-varying relationships to changing investment factors. The primary attribute of the rolling window method is its capacity for frequent model parameter adjustments in accordance with

the most recent market situation. Through the tutorial technique of momentum transfer, the training window moves forward over time and adds the new data as the older one might be irrelevant anymore. This constant updating is what keeps the model on its toes and thus it is not thrown by low predictive accuracy as it is able to follow the emerging trends in the market. For example, the first period of this study involved pre-training machine learning models with historical data between January 1, 2007, and December 31, 2018. Calibrated periodically, the rolling window technique was then employed, where the data was updated periodically to come up with new data points while the old ones were removed. The models maintained a sensitivity to the current market which made them more effective in the prediction of future trends, as we find in Arnott et al., 2019.

The results showed that rolling window strategy-trained machine learning models were generally better than fixed time window static trained models. The change was significant, having the XGBoost model scoring very well with marked tendency to quickly capture the moving market, supported by the Cumulative Net Value of 3.89 and Annualized Return of 37.89% (see Table 6). These data point out the fact that the model's reaction to new data is rapid and it is one of the important properties in success in the volatile financial markets (Gao et al., 2022).

Table 2. Comparison of Model Performance Using a Rolling Time Window

	DT	XGBoost	LightGBM	RF	NN
Cumulative Net Value	1.94	3.89	3.39	3.09	3.12
Annualized Return	16.68%	37.89%	33.4%	30.69%	30.68%
Sharpe Ratio	0.66	1.61	1.42	1.312	1.42
Maximum Drawdown	-28.02%	-22.29%	-21.56%	-22.53%	-21.68%
Annualized Return-to-Drawdown Ratio	0.59	1.70	1.54	1.36	1.42
Cumulative Excess Return	40.76%	228.81%	178.86%	157.16%	161.86%
Annualized Excess Return	8.49%	32.41%	27.28%	24.93%	25.38%

XGBoost in terms of the Sharpe Ratio which is a risk-adjusted return measure was its performance with 1.61 and that shows that it was able to balance the risk and reward more efficiently. The model's lower Maximum Drawdown of -22.29% relative to others also emphasized its strong risk management skills, which kept the team from losing a large part of the portfolio during the bad market conditions. Thus, it is evident that Asness et al.'s (2022) argument on the dynamic model's ability to adapt in order to avoid drawdown risks is reasonable.

Table 3. Comparison of Model Efficiency Using a Rolling Time Window

	DT	XGBoost	LightGBM	RF	NN
Out-of-Sample IC	0.01	0.05	0.05	0.05	0.05
Out-of-Sample IR	0.22	0.58	0.59	0.55	0.52
Out-of-Sample Positive IC Proportion	0.62	0.75	0.74	0.75	0.71

In terms of IC and IR indexes, the factor integration adaptive nature was the variable which was measured. The models trained with rolling window conditions were lightGBM and CatBoost which unlike fixed-window counterparts had consistently IC values over one or another. This illustrates that rolling window training served the purpose of beating the models in finding the predictive relationship between factors and asset returns in the case illustrated in Table 7. The evidence is backed up by the results of Nti et al. (2020), who found that adaptive machine learning models outperform other machines in high volatile financial cases.

3.3. Comparative Analysis of Machine Learning and Traditional Models

This paper section is analyzing the abilities of machine learning models to predict, particularly the XGBoost model and compare it with the traditional multi-factor models in the context of the Chinese financial market. Throughout the paper, the objective will be the investment strategies optimization by utilizing the XGBoost model's predictions and the mean-variance optimization framework to enhance the portfolio performance and at the same time to have efficient risk management.

The core discussion surrounds the amalgamation of the machine learning-based predictions from the XGBoost model and the Markowitz mean-variance optimization technique. This approach, which is based on contemporary portfolio theory, uses a portfolio with the lowest risk for a given amount of projected return in order to achieve the best possible balance between risk and return (Esakia et al., 2020). The predictions of the XGBoost model are then used to optimize this, so that the excess returns are maximized and the tracking error is kept at a controlled level with respect to the benchmark. The empirical analysis uses a back-testing period from February 1, 2015, to October 31, 2023. In the first five years, there was a monthly rebalancing period where the weights of each asset were adjusted according to the returns predicted by the XGBoost model.

This frequent portfolio weights adjustment gives a more flexible response to market developments which theoretically results in higher risk-adjusted returns of the portfolio (Zhang & Huang, 2022). The transaction costs were included at 0.12% per side to create a realistic trading environment.

Table 4. Comparative Analysis of Annual Returns between the Optimized Model and Traditional Model

Year	Optimized Model	Traditional Model
2015	155.56%	62.41%
2016	13.08%	-20.04%
2017	3.26%	-17.36%
2018	-11.68%	-36.97%
2019	53.39%	25.17%
2020	44.68%	19.26%
2021	60.78%	20.49%
2022	3.79%	-21.56%
2023	17.67%	-3.28%

A comparison of the results of the intelligent machine learning model and classical multi-factor models, as presented in Tables 4 and 5, shows the considerable performance difference. The XGBoost model was always the best one with higher returns when compared to traditional models. The XGBoost model was especially good at some times, e.g., 2015 (155.56% vs. 62.41%) and 2019 (53.39% vs. 25.17%). Even though the economy was in a recession in 2018, the machine learning model was more stable than the classical one, as it had a much lower drawdown of -11.68% compared to the traditional model's -36.97%. This implies that the machine learning method is better at adapting to market fluctuations and avoiding risks during periods of instability (Liu et al., 2019). Also, displaying the blend of machine learning along with mean-variance optimization has proven to be effective in the reduction of tracking error and the increase of portfolio returns. The Sharpe Ratio and Information Ratio of the XGBoost-optimized portfolio have always been greater than those of the traditional models, which indicates that the optimized portfolio had a higher risk-adjusted performance. The findings support the claim made by Gupta and Kelly (2019) that the risk-return trade-off would be improved by using sophisticated machine learning methods when building a portfolio.

Table 5. Hypothesis Testing Results

Test	Test Statistic	p-value
Paired t-test	4.551	0.00187
Wilcoxon signed-rank test	0	0.00390

In order to confirm the statistical performance improvement each method was verified by using a paired t-test and Wilcoxon signed-rank test. These tests were supposed to measure the difference in the predictive accuracy of the machine learning model and the multi-factor models. The paired t-test equaled a test statistic value of 4.551 and a p-value of 0.00187 which means that there is a statistically significant increase in the predictive accuracy of the machine learning model at the 1% significance level. This result provides a strong support for the view that certain machine learning models XGBoost models in particular are superior to traditional models when it comes to predicting asset returns (Chen & Mohan, 2021). Not only that, the Wilcoxon signed-rank test definitely confirmed these results with a p-value of 0.00390, which demonstrates that the results are still robust even with the presence of non-normal distributions or outliers. This non-parametric test provides more confidence in the reliability of the comparative analysis and emphasizes the greater flexibility of machine learning approaches in coping with complex market situations (Hoang & Wiegratz, 2023).

3.4. Proposed Navigation Tool for the Chinese Financial Market

The workflow of the suggested model consists of a few basic steps, which are aimed at achieving the better accuracy of the predictions through the iterative process. The iterative method is explained in the following way:

(1) Initialization: Start with a first forecast that is just a simple constant, like for instance, the mean of the target variable:

$$\hat{y}_l^0 = f_0(x_i) = 0$$

(2) Sequential Tree Building: In respect to each of the iterative processes t, f_t is a newly constructed tree added to the model to compensate for the residuals of the previous prediction:

$$\hat{y}_l^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_l^{t-1} + f_t(x_i)$$

(3) Objective Function Expansion: The second-order objective function is approximated using the Taylor expansion in order to optimize the objective function at the t step:

$$Obj^t = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

$$= \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T$$

(4) Optimization of Leaf Weights: The optimal weight for each leaf node in the decision tree is calculated by solving the following equation:

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \lambda T$$

(5) Gain Calculation for Splitting: The decision to split a node into two child nodes is based on the gain in the objective function, which measures the improvement in model performance. The gain is calculated as:

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + G_R + \lambda} - \gamma$$

(6) Regularization and Pruning: The model uses the complexity parameter γ\gammaγ to control tree growth, ensuring that each split results in a meaningful improvement. Pruning is performed to prevent overfitting by eliminating branches that do not contribute significantly to the model's accuracy.

The Markowitz mean-variance optimization framework serves as the core tool for optimizing the investment portfolio. This model's main objective is to strike a balance between risk and return, making sure that the projected return on the portfolio is maximized while maintaining a manageable tracking error. The mean-variance optimization technique is crucial in this context as it is designed to handle the inherent trade-offs in risk and return, particularly when incorporating predictions from machine learning models.

$$\max\ \omega^T f$$

$$subject\ to: (\sqrt{(w - w_{bench})^T \Sigma (w - w_{bench})} \le TTE$$

$$w - w_{bench} \le 1\%$$

In this study, the annualized tracking error was restricted to a maximum of 5%, reflecting a controlled deviation from the benchmark to enhance portfolio returns.

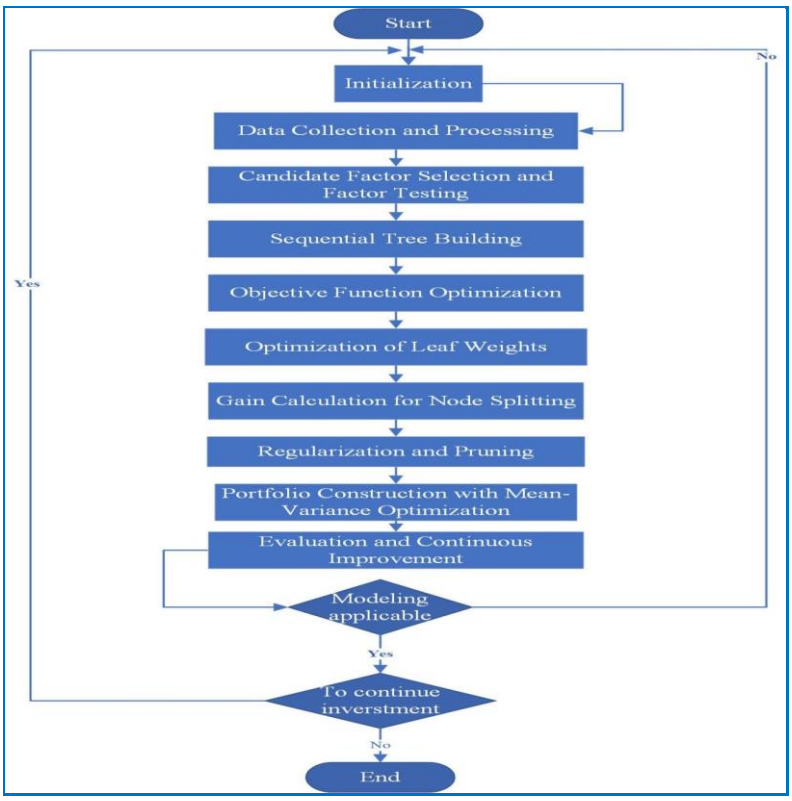

Figure 2. Flow chart of the Proposed Navigation Tool

4. Conclusions & Recommendations

4.1. Conclusions

This section of the dissertation captures its primary results, highlighting the tremendous impact of machine learning to the stock selection accuracy robustness in the Chinese financial market. This research not only clarifies the complexities described in the Statement of the Problem but also provides theoretical and practical contributions to quantitative finance through a machine learning-driven multi-factor model development.

4.1.1. Enhanced Stock Selection through Machine Learning Models

The study shows that tree-based models like LightGBM and XGBoost are superior to other multi-factor techniques and machine learning models. The results reveal that these models are the best performers in stock selection even with the time frames fixed or rolling during training. This is in line with past research that found machine learning can detect more complex and nonlinear patterns in financial data which traditional techniques cannot match (Chen & Mohan, 2021). These are machine learning models that can analyze high-dimensional data and uncover intricate dependencies between the components, unlike conventional methods which usually assume linear connections and rely on a few parameters. As noted by Gupta and Kelly (2019), this is in line with the study's aim of better forecasting through the use of advanced algorithms that can process vast amounts of financial data thus, avoiding the shortcomings of traditional stock selection methods.

4.1.2. Dynamic Adaptability of Machine Learning Models in Changing Market Conditions

The dissertation's most valuable finding is that machine learning models may evolve as market scenarios are changing. For instance, the XGBoost model may change its prediction algorithm according to the new data coming in, as per the analysis of the rolling time-window. They are thus able to precisely predict the short-term stock returns during market volatility. The flexibility and dynamism of machine learning make it a perfect choice for the unstable environment of the Chinese financial industry. Generally, conventional multi-factor models are inflexible to real-time changes due to their static nature. On the contrary, artificial intelligence techniques are constructed to refresh their learning algorithms iteratively, thus making them more proficient to market variations and non-linear relationships (Fama & French, 2020). This capability directly relates to the third research question of the dissertation by showing how machine learning can improve the evaluation of quantitative investment factors and the capture of dynamic market behaviors.

4.1.3. Strategic Importance of Transaction-Related Factors

The study's analysis kept on revealing the transaction-related factors such as liquidity and turnover rates which are the significant determinants of the short-term stock pricing and are more powerful than the traditional factors like size, value, and profitability. This observation implies that the market-driven data, which expresses immediate trading behaviors, is more important for the short-term forecasting in the Chinese financial context than the fundamental factors (Esakia et al., 2020). This apprehension can be a critical piece of information in the making of stock selection models. It defies the conventional financial theories that give precedence to the fundamental analysis and it might be a way of redirecting the attention toward the use of transaction-based indicators to increase the performance of the predictive models. This assertion is also linked to the first subproblem of the effectiveness of mainstream quantitative investment factors by providing a new point of view on factor dynamics that emphasizes market-based data in short-term investment strategies.

4.1.4. Integration of Machine Learning and Traditional Financial Models

This study employs a hybrid approach to machine learning in a novel way within the Markowitz mean-variance optimization framework. Through combining machine learning's predictive power with the traditional mean-variance optimization principles, we can create a comprehensive plan for the optimal portfolio performance and risk management. The merger of traditional financial theories and innovative computer technologies led to better portfolio results and broader applicability to investment scenarios. Arnott et al. (2019) study revealed that machine learning enhanced portfolios could achieve better performance than traditional investment strategies in dynamic and difficult-to-predict markets as well as minimizing the risk of tracking errors resulted from changes in benchmark indices.

4.2. Recommendations

These recommendations made in the dissertation might help legislators, investors, and financial experts to employ machine learning algorithms appropriately for stock selection and risk management. Discussed are hands-on uses, the best practices, and the strategic melding of machine learning into the quantitative investment strategies that have already been in use in the financial industry of China.

4.2.1. Adoption of Machine Learning Models for Enhanced Stock Selection

Financial practitioners should put an accent on the use of machine learning models in stock selection particularly tree-based methods like XGBoost and LightGBM since they have a better performance than conventional multi-factor models. Setup with both fixed and rolling training windows always yields good results allowing these models to adapt to the different market conditions. For realizing the full potential of these capabilities, investment firms have to allocate resources for building the infrastructure that is supporting the machine learning algorithms, among them being the data processing capabilities, high-performance computing resources, and advanced analytics tools. The integration of these technologies will make it possible for firms to analyze large volumes of market data in real-time, thereby enabling more accurate and timely decision-making.

4.2.2. Dynamic Portfolio Management through Rolling Window Analysis

Research studies reveal the fact that machine learning models have superiority, particularly XGBoost, in rolling window scenarios that are due to their ability to adapt to changing market dynamics, therefore, it is proposed that financial practitioners include rolling window analysis in their portfolio management strategies. This method gives models the opportunity to constantly update their predictions based on the most recent market data, which in turn, allows the models to respond more accurately to short-term trends and volatility. Rolling window analysis should be used not only for the predictive tests but also for the recalibration of the risk assessment models in the face of the market changes. By constantly training these models on the updated data, investors can remain in the front in the ever-changing Chinese financial market, thereby minimizing the strategy adjustment lag and improving the responsiveness to the market changes.

4.2.3. Developing Scalable Data Infrastructure for Machine Learning Applications

Machines learning-based stock selection systems need powerful data processing capabilities. Financial companies and investors aiming to adopt such models must, therefore, first lay the groundwork for developing a data infrastructure that can be scaled up easily and is capable of managing high-dimensional data, processing large datasets in real time, and supporting algorithms that are computationally intensive. To help facilitate this, investment firms can use data engineering methods, like the use of data warehouses, cloud storage solutions, and parallel computing frameworks. Quality of data can be ensured by strict preprocessing steps, including adjusting outliers, neutralizing factors, and standardization, which, in turn, will increase the dependability and accuracy of machine learning in stock selection and risk management.

4.2.4. Emphasizing Continuous Model Training and Evaluation

The suitability of machine learning models to new environments is determined by their capacity for incremental learning from streaming data. It is suggested that the investors practice a method that involves them retraining their models every so often using the latest data on financials. This is to make sure their predictions are accurate as well as relevant with an ever-evolving market. A method of backtesting that compares the performances of the current investment regimes with those of the past can be a great addition to the automated model re-training and validation procedures that investment organizations might use. Connected with this, continuous learning in machine learning allows models to adjust and evolve to ever-changing market conditions. Therefore, market models do not run the risk of overfitting to irrelevant patterns.

References

[1] Abe, M., & Nakagawa, K. (2022). Enhanced quantile portfolio for multifactor model with deep learning. Enhanced Quantile Portfolio for Multifactor Model with Deep Learning.

[2] Arnott, R. D., Clements, M., Kalesnik, V., & Linnainmaa, J. (2019). Factor momentum. Research Affiliates, LLC, Dartmouth College. Available at: https://ssrn.com/abstract=3116974

[3] Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2022). Combining value and momentum. Review of Financial Studies.

[4] Blitz, D., & Hanauer, M. X. (2020). The size factor in stock returns: Is it really there? Journal of Financial Markets.

[5] Chen, J., Tang, G., Yao, J., & Zhou, G. (2019). Investor attention and stock returns. Xiamen University, Hunan University, Jinan University, Washington University in St. Louis. Available at: https://ssrn.com/abstract=3194387

[6] Chen, X., Cho, Y. H. (Tony), Dou, Y., & Lev, B. (2022). Predicting future earnings changes using machine learning and detailed financial data. Journal of Accounting Research, 60(2), 467-515.

[7] Esakia, M., Goltz, F., Luyten, B., & Sibbe, M. (2020). The impact of the size factor on the Sharpe ratio of a multi-factor portfolio. Journal of Beta Investment Strategies.

[8] Fama, E. F., & French, K. R. (2020). Comparing cross-section and time-series factor models. Review of Financial Studies, 33(5), 1891-1926.

[9] Feng, L., Seasholes, M. S., & Zhang, L. E. (2020). Investment and stock returns: Evidence from China. Journal of Financial Economics, 137(2), 504-520. https://doi.org/10.1016/j.jfineco.2019.09.005

[10] Gandhmal, D. P., & Kumar, K. (2019). Application of machine learning in predicting the stock market: A comprehensive

review. Computer Science Review, 34, 100191. https://doi.org/10.1016/j.cosrev.2019.100191

[11] Gao, C.-H., Tsai, P.-F., & Yuan, S.-M. (2022). Synergy frontier of multi-factor stock selection model. Springer.

[12] Gupta, T., & Kelly, B. T. (2019). Factor momentum everywhere. The Journal of Portfolio Management, 45(3), 13-36.

[13] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Machine learning in financial markets: A survey. Annals of Operations Research, 282, 1-31. https://doi.org/10.1007/s10479-018-3103-7

[14] Hoang, D., & Wiegratz, K. (2023). Machine learning methods in finance: Recent applications and prospects. European Financial Management, 29(5), 1657-1701.

[15] Liu, Y., Wang, Q., & Cheng, Y. (2019). Size effect in Chinese stock market: Evidence from 2000 to 2016. Journal of Financial Research.

[16] Michis, A. A. (2022). Multiscale partial correlation clustering of stock market returns. Journal of Risk and Financial Management, 15(1), 24. https://doi.org/10.3390/jrfm15010024

[17] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). An extensive evaluation of ensemble learning for stock market prediction. Journal of Big Data, 7, 18. https://doi.org/10.1186/s40537-020-00295-1

[18] Pramanik, P., & Jana, R. K. (2022). Identifying research trends of machine learning in business: A topic modeling approach. Measuring Business Excellence, 27(4), 602-633.

[19] Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. Journal of Big Data, 7, 66. https://doi.org/10.1186/s40537-020-00333-6

[20] Tan, Y., Yang, W., Suntrayuth, S., Yu, X., Sindakis, S., & Showkat, S. (2023). Optimizing stock portfolio performance with a combined RG1-TOPSIS model: Insights from the Chinese market. Journal of the Knowledge Economy.

[21] Tsai, P.-F., Gao, C.-H., & Yuan, S.-M. (2023). Stock selection using machine learning based on financial ratios. Mathematics, 11(23), 4758. https://doi.org/10.3390/math11234758

[22] Zhang, X., & Huang, D. (2022). Profitability factor and stock return predictability: Evidence from China. Journal of Banking and Finance.