IFA
IRAN FINANCE ASSOCIATION

Iranian Journal of Finance

Print ISSN 2676-6337
Online ISSN 2676-6345

Smart-Beta Portfolio Optimization Using Machine Learning Techniques

Fatemeh Salehirad

MSc. in Financial Engineering and Risk Management, Department of Financial Engineering, College of Management, University of Tehran, Tehran, Iran. (Email: f.salehirad@ut.ac.ir)

Farid Tondnevis*

*Corresponding Author, Assistant Prof., Department of Financial Engineering, College of Management, University of Tehran, Tehran, Iran. (Email: farid.tondnevis@ut.ac.ir)

Iranian Journal of Finance, 2025, Vol. 9, No.4, pp. 91-116.
Publisher: Iran Finance Association
doi: https://doi.org/10.30699/ijf.2025.524928.1518
Article Type: Original Article

Received: March 01, 2025
Received in revised form: June 14, 2025
Accepted: October 28, 2025
Published online: December 01, 2025

Abstract

This study examines the integration of machine learning techniques with smart beta investment strategies to enhance portfolio performance. Traditional market indices often fail to meet investors' expectations, especially during volatile market periods, leading to a growing interest in alternative strategies such as smart beta methodologies. These strategies combine the cost and risk efficiency of passive investing with the performance advantages of active strategies by employing alternative weighting schemes based on financial

factors such as value, quality, and momentum. In this research, Return on Invested Capital (ROIC) is selected as a value-based factor due to its strong reflection of a company's operational efficiency and value creation driver. We employ three machine learning models—Support Vector Regression (SVR), Random Forest, and XGBoost—to forecast ROIC based on various financial ratios. Each model is fine-tuned using Bayesian optimization techniques to achieve the highest forecasting accuracy. The dataset includes financial data from 85 manufacturing companies listed on the Tehran Stock Exchange. Model performance is evaluated using R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), with the optimized Random Forest model achieving the best results based on higher R^2 and lower error values compared to the other models. The forecasted ROIC values are then used to construct a smart beta portfolio, which is compared to a traditional market-cap-weighted portfolio. The findings demonstrate that a machine learning-enhanced, ROIC-based smart beta strategy can significantly outperform traditional approaches, offering investors a more robust and data-driven method for portfolio construction and risk-adjusted return enhancement.

Keywords: ROIC, Machine learning, Bayesian optimization, Smart-beta strategy, Factor investing

JEL Classification: C5, C6, G1

Introduction

A range of investment approaches is used in the portfolio management process, including active and passive strategies. The passive approach, which is based on the efficiency of capital market information, considers specific criteria for the investment fund manager by limiting the manager's operational flexibility (Beasley et al., 2003)

Index tracking, as one of the passive investment management approaches, seeks to form a portfolio in such a way that replicates the performance of the index by investing in a group of shares that make up the index. On the other hand, active management allows fund managers a high degree of flexibility to attempt to "pick winners," or stocks whose values are expected to outperform other stocks over a period of time. (Beasley et al., 2003)

Several studies have compared active and passive approaches in investment. In general, the active approach imposes more costs and risks on investors. Finding and buying winning stocks and then selling others creates

high transaction and analysis costs, which also impose unsystematic risks on investors. On the other hand, passive approaches, which involve a buy-and-hold strategy, have lower transaction costs. Due to proper diversification, passive approaches only impose systematic risk on investors. (Beasley et al., 2003)

Investors always want to have strategies that offer higher returns and lower risk. Moreover, traditional market indices have shortcomings and cannot meet the needs of investors. So, responding to dramatic fluctuations in the stock market and the impact of inefficient markets, investors seek investment methods with higher returns than passive investment and also seek lower volatility than active investment (Chen et al., 2023). One approach that has revolutionized the financial investment methods is the Smart-Beta strategy, which has exceeded the traditional indices' performance.

Smart beta investing combines the benefits of passive with the advantages of active investing strategies. Smart beta defines a set of investment strategies that emphasize the use of alternative index construction rules compared to traditional market capitalization-based indices. Smart beta focuses on capturing investment factors or market inefficiencies in a rules-based and transparent way. The image below provides a good overview of investment strategies. (Malkiel, 2014)

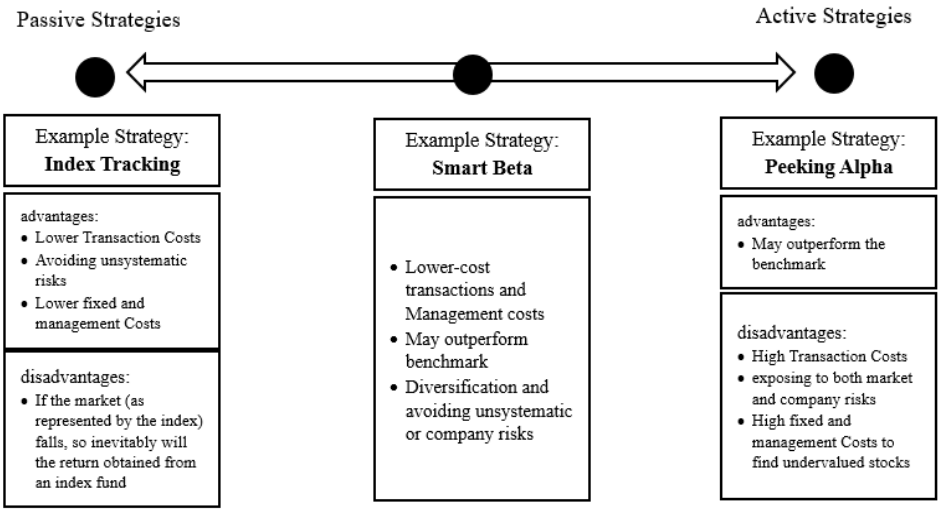

Figure 1. Classification of Investment Strategies

Index tracking, as one of the most frequent approaches used in passive strategies, seeks to replicate the performance of a market value-weighted index. Smart beta strategies, similar to passive approaches, aim to replicate the performance of an index but use alternative index weighting schemes. The smart beta strategy employs an active approach in determining how to weight the index it seeks to replicate. Weighting schemes used in smart beta strategies include volatility, liquidity, quality, value, size, and momentum.

In simpler terms, smart beta approaches do not replicate the performance of classic indices such as the S&P 500 or FTSE 100. Instead, they use index tracking approaches to reconstruct an index expected to perform better than market-based indices. Therefore, the smart beta strategy exploits the advantages of both passive and active approaches: lower transaction costs and unsystematic risk, while striving to outperform classic market indices.

Smart beta strategies have high transparency, low management costs, and better long-term performance, but are at risk of severe short-term declines due to a lack of risk control tools. Although there are some methods to use historical volatility for risk control, it is still difficult to adapt to the rapid switch of market styles. How to strengthen the risk control management of the portfolio while maintaining the original advantages of smart beta has become a new issue of concern in the industry. (Zhao et al., 2023)

To solve these problems and reach better risk management, researchers have turned to machine learning methods. Machine learning, with its ability to analyze vast amounts of data and detect complex patterns, offers suitable solutions for predicting the future and decreasing risks. By such advanced algorithms, machine learning techniques provide more robust risk management strategies compared to traditional methods.

One important application of machine learning in financial markets is the prediction of market trends and asset prices. Therefore, taking advantage of machine learning models in stock forecasting, many researchers incorporate these models into portfolio optimization models to improve portfolio performance (Chen et al., 2022). However, forecasting stock prices comes with significant uncertainties due to the multitude of factors that influence price movements, such as country-specific economic conditions, investors' sentiments towards a particular company, and political events (Basak et al., 2018). The complexity and interdependencies of these variables make accurate stock price forecasting challenging. To address these uncertainties, researchers

have focused on forecasting more stable and reliable financial metrics.

An auspicious approach is the forecasting of the return on invested capital (ROIC). Unlike stock prices, which are at risk of fluctuations and market noise, ROIC is dependent on a company's fundamental performance.

In this research, we apply machine learning models, which can be trained on historical financial data and a variety of relevant features, to forecast future ROIC values accurately. Consequently, using the forecasted ROIC as a basis for investment decisions can lead to more stable and forecastable returns compared to strategies that solely focus on stock price movements.

Moreover, integrating ROIC value predictions into smart-beta strategies can lead to better portfolio optimization. By selecting stocks based on their expected ROIC, investors can build portfolios that not only offer higher returns but also exhibit lower volatility. This approach provides a reliable foundation for investment decisions, thereby addressing the limitations of traditional stock price predictions. The rest of this paper first discusses the existing literature and the theoretical background. Then, it explains the data and methodology used in the study, followed by the presentation of the empirical results and their discussion. Finally, the paper concludes with key findings and suggestions for future research.

Literature Review

Machine learning techniques to forecast financial metrics

In the field of finance, many studies have combined machine learning techniques and financial ratios to forecast various financial metrics such as stock prices, credit risks, and company performance. Moreover, financial ratios have been used by many researchers for different purposes. Before reviewing research that used machine learning techniques, we try to review the application of financial ratios in the field of finance. The first person to use financial ratios in his research was (Beaver, 1966), who selected six financial ratios to distinguish between failed and non-failed corporations (Altman, 1968) and created a quantitative model, which was based on multiple financial ratios, to predict the bankruptcy of companies. Moreover, the primary goal of Ohlson's research was to predict bankruptcy, which outperformed previous studies using various financial ratios (Ohlson, 1980). Subsequent investigations have broadened the application of financial ratios, often integrating machine learning algorithms to enhance predictive accuracy.

Machine learning techniques have revolutionized the field of finance. These methods enable researchers to analyze a vast amount of data and forecast factors with high accuracy. In contrast, traditional methods, such as linear regression, cannot find the complex relationship among financial data. These studies consist of important financial domains such as credit scoring, financial crisis, and portfolio management. In the following, there is an overview of the application of machine learning methods in these important fields.

In the field of credit scoring, (Trivedi, 2020) used different machine learning methods, including support vector machine (SVM), decision tree, random forest, and Bayesian model to determine the credit score, and concluded that the random forest classifier has more accuracy compared to other classifiers. In the field of financial crisis, many researchers have focused on predicting bankruptcy by using financial ratios. For example, (Shetty et al., 2022) applied extreme gradient boosting (XGBoost), support vector machine (SVM), and a deep neural network, and just three financial ratios, including the return on assets, the current ratio, and the solvency ratio. They were able to reach a high accuracy in predicting bankruptcy. In the field of portfolio management, in the research of (Ma et al., 2020), two machine learning models and three deep learning models are applied to forecast stock return, and they showed that the random forest regressor outperformed other methods.

Smart Beta Strategy

Smart beta is a new way of investing that mixes the low cost of passive investing with some benefits of active investing. Instead of using traditional market value weighting, smart beta builds indices with other rules to catch special patterns in the market. These rules are clear and straightforward, and they help investors to get better returns in the long run.

Many studies have focused on Smart-Beta investment strategies to manage portfolios more effectively. These strategies employ different factors to achieve better portfolio performance. This literature review will provide an overview of various Smart Beta strategies, including Mean-Variance optimization, Minimum-Variance portfolio, Equal weighing, Risk Parity Strategy, and Factor Investing. In this research, our focus is on the Factor Investing strategy. In Table 1, we summarize the previous research that focused on various Smart-Beta investment strategies.

Table 1. Overview of Studies on Smart Beta Approaches

Study	Tittle	Selected smart beta strategy	Main findings
(Benartzi & Thaler, 2001)	Naive Diversification Strategies in Defined Contribution Saving Plans	Equal weighing(1/N)	The research concludes that while diversification can produce a reasonable portfolio, it does not assure sensible or coherent decision-making.
(Almahdi, 2015)	Smart Beta Portfolio Optimization	Mean-Variance optimization (MV)	The research showed that smart-beta, which is defined as a portfolio management strategy combining cap weight, economic scale, and minimum variance, replicates the benchmark index.
(Richard & Roncalli, 2015)	Smart Beta: Managing Diversification of Minimum Variance Portfolios	Minimum-Variance portfolio	The research concludes that smart beta strategies, especially risk-based ones, provide better diversification and risk management compared to traditional CW portfolios. However, their success depends on how much volatility is reduced.
(Arnott et al., 2005)	Fundamental indexation	Factor Investing	The fundamental index outperformed the S&P 500 by an average of 1.97% per year from 1962 to 2004. Regardless of interest regimes, business cycles, and bear or bull stock markets, the performance of this strategy is good.

Combination of machine learning methods and Smart-Beta strategies

Smart-Beta strategies can perform well over the long term, but they often face big drops in the short term. At first, these strategies focused on managing style exposure but did not pay enough attention to how to optimize weights. On the other hand, traditional portfolio management mainly looks at historical data and does not make the most of what future trends might tell us. These methods typically use the average historical return as the expected return, which can oversimplify stock market behavior and lead to incorrect predictions about short-term returns (Agrawal et al., 2021).

Recently, the combination of computational techniques and investment strategies has gained attention in the field of finance. In (Ta et al., 2018), linear regression and support vector regression were used for stock movement prediction, and these algorithms have demonstrated high accuracy and attractive returns compared to the S&P 500.

One important domain, which our research focused on, is the combination of machine learning methods and Smart-Beta strategies. The main aim of Smart-Beta strategies is to perform better than traditional market indexes. Machine learning techniques can forecast key factors such as ROIC to help the Smart-Beta strategy reach its primary goal.

The reason for combining Smart Beta strategies with machine learning methods is that Smart Beta needs an accurate forecasting of financial factors to outperform traditional indices. While it could be combined with other techniques, traditional statistical or econometric models are limited because they mainly rely on linear assumptions and historical averages. These models cannot fully capture the complex and nonlinear relationships in financial data. In contrast, machine learning methods can process large datasets, detect hidden patterns, and provide more accurate forecasts of financial metrics, such as ROIC. This predictive ability makes machine learning especially suitable for supporting Smart Beta strategies and improving their performance compared to other approaches.

This combination of machine learning techniques and Smart-Beta strategies is not only theoretical. Previous studies have demonstrated that such an approach can generate superior risk-adjusted returns compared to traditional methods. In this part of the literature review, we focus on research that integrates machine learning with Smart-Beta strategies. We review key studies, outline the methods applied, and highlight their unique contributions. A summary of these studies is provided in Table 2.

Table 2. Summary of Previous Research on Smart Beta Investment Strategies with Machine Learning

Study	Methodology	Forecasted Financial Metrics	Application in Smart-Beta Strategies	Unique Aspects
(Huang, 2011)	Support vector regression (SVR)	Stock prices	Equal weighing(1/N)	Genetic algorithms (GAs) are used to optimize the model parameters and input features.
(Huang et al., 2025)	Robust Linear Regression, Random Forest, and LSTM	Stock prices	Mean-Variance optimization(MV)	This paper presents an approach that integrates stock return prediction using machine learning algorithms with the mean-variance model to enhance performance
(Chen et al., 2022)	RF SVR LSTM	Stock returns	modified mean–variance (MMV) model	Applying the diversification level, which is measured by the Pearson correlation coefficient.
(Zhao et al., 2023)	natural gradient boosting	stock prices and their probability distribution	MV 1/N	Examining the quality of forecast uncertainty, which was neglected by previous research
(Zhu et al., 2023)	Extreme gradient boosting (XGBOOST)	Stock prices	Factor investing	Price indicators were used. Nonlinear integration was functional to improve the investment performance when combining characteristics.
Proposed Research	XGBoost, Random forest with Bayesian optimization	ROIC	Factor investing	Employs advanced ML techniques for ROIC prediction. Unique focus on robust portfolio construction

Research Methodology

This methodology comprises two sections: theoretical and numerical. The theoretical section introduces the Smart-Beta strategy and ROIC, forming the basis of the study. The numerical section applies machine learning techniques to forecast ROIC using financial ratios, followed by portfolio construction using a Smart-Beta strategy.

Theoretical section

This paper focuses on a profitability-based weighting approach to form an smart beta portfolio. In classical value-based weighting approaches, ratios such as P/E and P/B are used because empirical evidence has shown that stocks with lower P/E and P/B ratios usually generate higher returns for investors. This paper focuses on Return on Invested Capital (ROIC) as a profitability-based weighting measure. Return on Invested Capital (ROIC) is a financial metric that can help to assess whether a company is creating value with its investments.

ROIC (return on invested capital) represents a business's ability to generate operational income using invested capital. In other words, all aspects of working capital, asset management, and business profitability are considered in ROIC. On the other hand, the Expected Return of financial resources providers, which is directly affected by business risks, is represented by WACC (weighted average cost of capital). If a company's ROIC is greater than its WACC, the company's operations have provided a return greater than the expected return of investors. As a result, value creation occurs in this company (Tim Koller , Marc Goedhart, David Wessels, 2020).

The formula for calculating ROIC is as follows:

$$ROIC = \frac{NOPLAT}{IC} \quad (1)$$

Where NOPAT is net operating profit after tax, and IC is the amount of financial resources invested in corporate operations.

The above equation shows that if a company can create more operating profit with lower capital employed, it has obtained a higher return on its operations, leading to the creation of more value as a result of the company's operations.

DuPont decomposition can separate the ROIC into two parts:

$$ROIC = \frac{NOP(1-t)}{IC} = \frac{NOP(1-t)}{Sale} \times \frac{Sale}{IC} \quad (2)$$

Where Sale is the amount of operating income of a company, the above formula shows that ROIC is a function of operating profit margin ($\frac{NOP(1-t)}{Sale}$), and Invested Capital Turnover ($\frac{Sale}{IC}$).

In other words, if a company can create more operating income with less investment and simultaneously increase its operating profit margin by managing costs, it will generate higher returns.

Cost management strategies of companies are shown in the first part of ROIC because those strategies lead to an increase in the operating profit margin. Strategies based on more sales (marketing, product selection, pricing, etc.) and working capital management (receivable period, inventory turnover period, etc.) are shown in the second part of ROIC.

Numerical section:

The numerical section of this study aims to apply advanced machine learning techniques to forecast the ROIC of manufacturing companies that are listed on the Tehran Stock Exchange (TSE), and the application of this prediction in Smart-Beta strategy and index tracking. We used financial data that was gathered over the past five years, from 2019 to 2023. This section is divided into six main parts: Data Gathering, Data Preprocessing, Feature Engineering, Choosing the Right Machine Learning Model, Model Training and Evaluation, and Application in Smart Beta Strategy.

Data Gathering

In this study, we focused on 85 manufacturing companies listed on the Tehran Stock Exchange (TSE). Although more companies are listed on the TSE, only these firms had complete and consistent financial data for the entire five-year period from 2019 to 2023. To ensure consistency and comparability, we excluded financial institutions such as banks and insurance companies, as well as holding and investment companies, because their business models and financial structures are fundamentally different from those of manufacturing firms. Companies with incomplete or irregular financial statements were also

removed from the sample.

By limiting the analysis to manufacturing companies with complete and reliable data, we ensured that the results are robust and comparable across all firms. The primary data sources included the TSE database, company financial statements, and relevant economic reports. The financial metrics considered for each company included Return on Invested Capital (ROIC) as detailed in the theoretical section. Additionally, fifteen financial ratios were calculated using the financial statements of these companies. These ratios are summarized in Table 3.

Table 3. List of Financial Ratios Used for ROIC forecasting

Feature	Financial ratio	Description	Type
F1	Gross Profit Margin	Gross profit divided by revenue	Profitability
F2	Operating Profit Margin	Operating income divided by revenue	Profitability
F3	Pre-Tax Profit Margin	Pre-tax income divided by revenue	Profitability
F4	Net Profit Margin	Net income divided by revenue	Profitability
F5	Return on Assets (ROA)	Net income divided by total assets	Profitability
F6	Return on Equity (ROE)	Net income divided by shareholders' equity	Profitability
F7	Asset Turnover	Revenue divided by total assets	Efficiency
F8	Inventory Turnover	Cost of goods sold divided by average inventory	Efficiency
F9	Receivables Turnover	Revenue divided by average receivables	Efficiency
F10	Total Debt to Total Assets	Total debt divided by total assets	Solvency
F11	Debt Ratio	Total liabilities divided by total assets	Solvency
F12	Interest Coverage Ratio	EBIT divided by interest expenses	Solvency
F13	Current Ratio	Current assets divided by current liabilities	Liquidity
F14	Quick Ratio	(Current assets - Inventory) divided by current liabilities	Liquidity
F15	Cash Ratio	Cash and cash equivalents divided by current liabilities	Liquidity
F16	Target Feature	ROIC	-

These ratios align well to estimate ROIC because they cover the factors that contribute to both value creation (profitability and efficiency). This type of selection ensures that the model captures the fundamental financial factors that impact a company's performance.

Data Preprocessing

Data preprocessing is a crucial step in ensuring the quality and consistency of the dataset and impacts the performance of the machine learning models. The following steps were taken to preprocess the data:

Handling missing values

In this study, the Interest Coverage Ratio had missing values because some companies had zero interest expenses, resulting in a zero denominator and an undefined or infinite ratio. To address this, we replaced these missing values with three times the maximum observed value of the Interest Coverage Ratio. This approach ensures that the imputed values remain high enough to reflect the strong ability of these companies to cover their interest expenses, without skewing the data with actual infinity values. Other imputation methods, like mean or median substitution, would underestimate these ratios in this context.

Handling outlier detection

Outliers were detected using the IQR method. Then, detected outliers were scaled to the dataset to minimize their impact on the model by the RobustScaler method because traditional scaling methods, such as Min-Max scaling, can be heavily influenced by extreme values, which may distort the feature distributions. RobustScaler, on the other hand, uses the median and the interquartile range (IQR) to scale the data, making it more robust to outliers. This approach ensures that the impact of extreme values on the scaling process is minimized, leading to a more stable and reliable dataset for training the models. Then, the Power Transformer (Yeo-Johnson) method is used to stabilize variance and make the data more Gaussian-like. It is beneficial for handling skewed data.

Data Normalization and Standardization

To ensure that all features contribute equally to the model performance, they were standardized using StandardScaler from Scikit-learn. This method standardizes features by removing the mean and scaling to unit variance, which is essential for algorithms sensitive to feature scaling, such as Support Vector Machines.

Feature Engineering

In this study, feature engineering was performed to ensure the most informative features were selected for the prediction model. The feature engineering process comprised the following steps:

Correlation Analysis

To understand the relationships among the 15 selected features, we conducted a correlation analysis. This step was crucial for identifying multicollinearity, which can adversely affect model performance. Features with high correlation coefficients were carefully examined to determine if any could be removed to simplify the model and reduce redundancy.

Table 4. Correlation Results and Feature Selection

Category	Variables	Explanation
Highly correlated pairs	F2 – F3, F7 – F8, F12 – F13	These variables show high pairwise correlation, indicating potential redundancy.
Weak correlation with ROIC (F16)	F5, F9, F14	These features exhibit very low correlation with ROIC compared to others, suggesting limited forecasting power.

Handling Novel Methods

We chose not to employ other traditional feature selection methods, such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA), because the machine learning models used in this study, such as XGBoost and Random Forest, inherently manage feature importance and interactions. These models are capable of handling a large number of features and automatically identifying the most significant ones during the training process.

This approach allowed us to leverage the strengths of advanced machine learning techniques while ensuring that the input features were relevant and not overly redundant. By combining correlation analysis with the inherent feature-handling capabilities of modern algorithms, we aimed to optimize model performance and interpretability.

Choosing the Right Machine Learning Model

SVR

Support Vector Regression (SVR) is a powerful machine learning technique derived from Support Vector Machines (SVM), introduced by (Cortes & Vapnik, 1995). SVR is particularly well-suited for regression problems, as it focuses on finding a hyperplane that best fits the data points while minimizing the prediction error (Drucker et al., 1996). Unlike traditional regression methods, SVR aims to maintain a margin of tolerance, within which errors are acceptable. The algorithm selects support vectors, which are the data points that lie closest to the hyperplane, to determine the optimal regression line. SVR's effectiveness lies in its ability to handle high-dimensional data and its robustness to outliers, making it a popular choice for various applications, including financial modeling and forecasting (Sun & Yu, 2019)

Random forest

The Random Forest is a type of machine learning algorithm used for classification and regression, introduced by (Breiman, 2001). It gets its name because it is made up of many decision trees, which are built with some randomness. This means each tree in the forest is different from the others. Unlike regular decision trees that consider all variables when splitting, Random Forest trees use only a random subset of the input variables. This approach reduces the risk of overfitting and enhances the model's generalization capabilities. Additionally, Random Forest provides importance scores for each feature, enabling researchers to understand the contribution of different variables in the prediction process. Its robustness and versatility make it useful for complex datasets in financial applications.(Kamdem & Selambi, 2022)

XGBOOST

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm that extends the Gradient Boosted Decision Tree (GBDT) framework. Similar to Random Forest, XGBoost is composed of many decision trees; however, unlike Random Forest, the trees in XGBoost are built sequentially rather than independently. Each new tree is trained to correct the errors of the previous ones, thereby minimizing the loss function iteratively.

XGBoost introduces several innovations that significantly improve predictive performance. First, it applies both L1 and L2 regularization, which

prevents overfitting and enhances generalization. Second, it can efficiently handle missing data by automatically learning the best direction to substitute them. Third, its parallel and distributed computing framework makes it highly scalable to large datasets. Moreover, XGBoost incorporates shrinkage (learning rate) and column subsampling, further boosting its robustness.

Another key advantage of XGBoost is its ability to provide feature importance scores, which help researchers interpret the relative contribution of variables in prediction. Due to its high accuracy, computational efficiency, and flexibility, XGBoost has become one of the most widely used algorithms in various fields, including finance.

Recently, XGBoost has been applied to the forecasting of financial markets.(Chen et al., 2022)

In Table 5, we list the advantages and disadvantages of these machine learning methods.

Table 5. advantages and disadvantages of machine learning methods

Method	Advantages	Disadvantages
SVR	- This method works well with data that has many features. -It typically resists overfitting in high-dimensional data. -It adapts to nonlinear relationships with the kernel trick.	-Its performance heavily depends on choosing the correct kernel. -It requires significant computational power for large datasets. -This method is less transparent and more complex to interpret than tree-based approaches.
Random Forest	-This method has strong performance in both regression and classification problems. - It shows which features are most important for predictions. -It is suitable for larger datasets and can handle missing values.	-This method can be time-consuming to train if using many trees or deep tree structures. - It has higher computational demands than simpler algorithms.
XGBoost	- This method has high predictive accuracy, especially with optimal tuning. -It can handle missing data and reduce overfitting through regularization.	-The process of hyperparameter tuning can be complicated -It is usually less interpretable than simpler models.

Model Training and Evaluation

Data Splitting

To ensure a robust model assessment, the dataset was split into training and test sets with a 70/30 ratio, respectively. The training set, comprising 70% of the data, was used to train the machine learning models, while the remaining served as an independent test set to evaluate the model's generalization capability.

Hyperparameter Tuning

To optimize the model's performance, Bayesian optimization was employed for hyperparameter tuning. Root Mean Squared Error (RMSE) was used as the objective metric to guide the search, allowing for an efficient and systematic exploration of the hyperparameter space. Bayesian optimization enabled more targeted parameter adjustments to minimize RMSE, so it helps to identify the optimal model configuration for each algorithm.

Model Evaluation

It is crucial to understand if a machine learning method can perform well. Therefore, we need some evaluation metrics for gauging the performance of a machine learning technique. The following is a summary of the functional formula commonly employed for evaluating ML models, where Y_i is the actual value of the variable, F_i is the forecasted value from the model, Y_m is the mean of actual values, and n is the number of data points (Jian Huang, Junyi Chai, Stella Cho, 2020).

- R-Squared (R^2): R^2 measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, where a higher value indicates better explanatory power.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{i-}F_i)^2}{\sum_{i=1}^{n}(Y_{i-}Y_m)^2} \quad (3)$$

- Mean Squared Error (MSE): calculates the average of the squared differences between forecasted and actual values. It penalizes larger errors more significantly, making it sensitive to outliers. The metric is helpful in

evaluating the overall accuracy of the model and is widely used in regression tasks. However, since it is in squared units of the target variable, it may not be directly interpretable.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2 \quad (4)$$

- Root Mean Squared Error (RMSE): This metric is the square root of MSE, making it easier to interpret as it is expressed in the same units as the dependent variable. This metric provides an intuitive measure of the model's prediction error and is particularly valuable when comparing different models or interpreting the magnitude of the error in real-world terms.

$$RMSE = \sqrt{MSE} \quad (5)$$

- Mean Absolute Error (MAE): MAE measures the average absolute difference between predicted and actual values, treating all errors equally regardless of their direction. It is robust to outliers compared to MSE and RMSE. MAE is suitable when a straightforward, interpretable metric is needed to understand the model's prediction accuracy without overemphasizing significant errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i| \quad (6)$$

Application in Smart Beta Strategy

In this section, the ROIC was used to construct a fundamentally weighted portfolio under a Smart-Beta strategy. Stocks were selected based on their ROIC, with those showing the highest forecasted ROIC included in the portfolio. Unlike traditional market-capitalization weighting, this portfolio was weighted according to the ROIC of each stock, enabling the selection process to reflect fundamental valuation metrics rather than market size alone.

To manage outliers, an exponential weighting method was applied. This approach, which assigns larger weights to higher ROIC, mitigates the impact of

outliers more effectively than a standard min-max normalization method. In this algorithm, weights w_i for each stock i are calculated by applying the exponential function to the forecased ROIC R_i as follows:

$$w_i = \frac{e^{R_i}}{\sum e^{R_i}} \tag{5}$$

This exponential weighting formula emphasizes differences among stocks while controlling for the influence of extreme outliers, allowing the portfolio better to capture variations in the fundamental strength of each stock.

Conclusion

Factor Investing and Portfolio Performance

In this study, we explored the efficacy of factor investing by focusing on Return on Invested Capital (ROIC) as a primary factor. The main objective was to construct a Smart-Beta portfolio that could outperform a traditional market-capitalization-weighted index. To achieve this, we conducted a back-testing analysis over the past five years, using monthly data from 85 manufacturing companies listed on the Tehran Stock Exchange. This back testing aimed to assess whether the ROIC-based portfolio strategy could deliver consistent and superior performance, which could make it suitable for future application.

Our findings demonstrated that the ROIC-based Smart-Beta portfolio achieved higher returns compared to the traditional market-capitalization-weighted portfolio. This superior performance is rooted in the fundamental insight that companies with a higher ROIC are more likely to generate greater returns by effectively utilizing their capital to create value. Furthermore, the ROIC-based portfolio maintained a more favorable risk profile, reinforcing its potential as a viable investment strategy.

The empirical results, illustrated in Figure 3, support these conclusions. Figure 2 displays the individual monthly performance of the portfolios, highlighting the consistent outperformance of the Smart-Beta approach. Figure 3 presents the cumulative performance over the five-year period, which confirms the sustained advantage of the ROIC-based portfolio. These results suggest that the ROIC factor can be effectively used to construct portfolios that are not only more profitable but also better aligned with long-term investment goals.

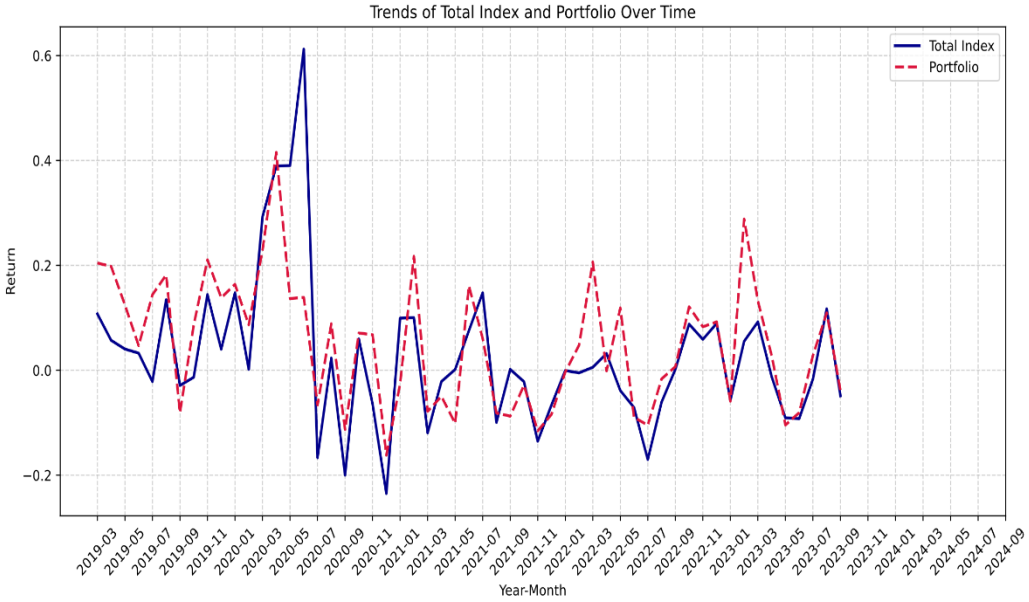

Figure 2. Trends of the total index and portfolio

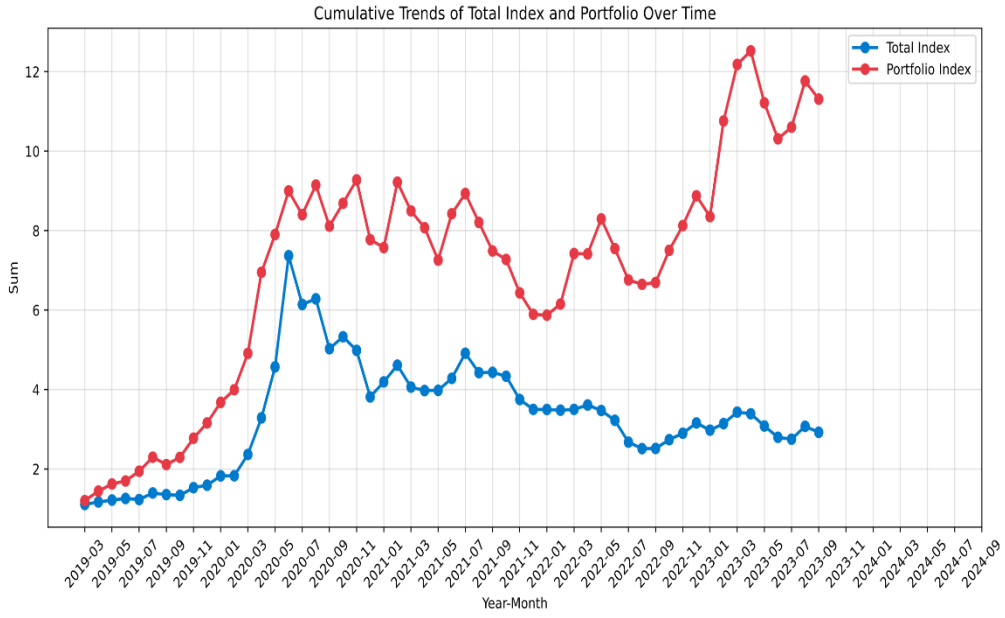

Figure 3. cumulative trends of the total index and portfolio

Based on the superior performance of the ROIC-based portfolio, we decided to use the predicted ROIC for the next year to construct a forward-looking smart beta portfolio.

Machine Learning Methods for ROIC forecasting

The performance comparison of the three models—SVR, Random Forest, and XGBoost—reveals that the Random Forest model outperforms the others across most evaluation metrics. It achieves the highest R^2 value of 0.41, indicating better explained variance. Additionally, it has the lowest error values, including MSE (0.28), $RMSE$(0.53), and MAE (0.44), demonstrating greater accuracy and reliability in predictions. While SVR and XGBoost also provide reasonable results, their performance metrics are less favorable, with XGBoost showing the lowest R^2(0.27) and relatively higher error rates compared to Random Forest.

Table 6. Performance of different models

Metric	SVR	Random Forest	XGBoost
R2	0.36	0.41	0.27
MSE	0.65	0.28	0.35
RMSE	0.80	0.53	0.59
MAE	0.64	0.44	0.48

Moreover, the predictive capabilities of these machine learning methods, combined with the results that were gained from our factor investing strategy, can be leveraged to create optimized portfolios for the future. By using the forecasted ROIC values and constructing portfolios based on the ROIC factor, investors can potentially enhance their returns.

Application of Predictions for Future Portfolio Construction

To further validate the effectiveness of our findings, we applied the ROIC predictions generated by our machine learning models to construct a portfolio for the following year. We then compared the performance of this forecasted ROIC-based portfolio to a traditional cap-weighted portfolio over the next year.

The analysis aimed to assess whether the forecasted ROIC-based portfolio could maintain its superior performance in a future scenario. The results showed that the ROIC-based portfolio continued to outperform the cap-weighted portfolio in terms of returns and risk-adjusted metrics. Figure 5 presents the trend of the portfolio performance over the next year period.

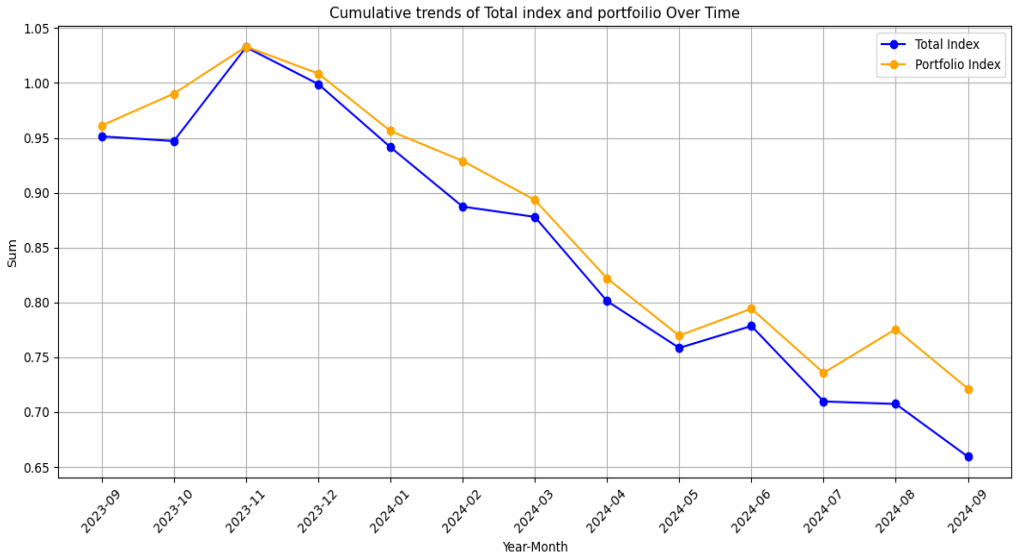

Figure 4. prediction of portfolio

The Information Ratio (IR) is a metric used to evaluate the performance of a portfolio relative to a benchmark by measuring the excess return generated per unit of risk. Equation 8 presents the formula of the Information Ratio, where R_p is the portfolio Return and R_I is the Benchmark return. The higher the IR, the better the portfolio performance compared to the index. The portfolio has been able to generate higher returns than the index with lower tracking error.

$$Information\ Ratio = \frac{\mathrm{E}(R_p - R_I)}{\mathrm{stdev}(R_p - R_I)} \tag{8}$$

Based on Table 7, the portfolio achieved an Information Ratio of approximately 0.25.

Table 7. Information ratio of portfolio

Market return $\boldsymbol{R_I}$	Portfolio return $\boldsymbol{R_p}$	$\boldsymbol{R_p - R_I}$
-4.87%	-3.88%	0.99%
-0.45%	3.03%	3.48%
9.03%	4.32%	-4.71%
-3.25%	-2.37%	0.88%
-5.76%	-5.19%	0.57%
-5.75%	-2.85%	2.91%

-1.05%	-3.84%	-2.78%
-8.74%	-8.00%	0.74%
-5.33%	-6.35%	-1.01%
2.63%	3.19%	0.55%
-8.83%	-7.36%	1.47%
-0.33%	5.42%	5.75%
-6.80%	-7.00%	-0.21%
Average:		0.66%
Standard Deviation:		2.64%
Information Ratio:		25.04%

This result indicates that the portfolio, which is based on the value of ROIC, generated some additional returns compared to the benchmark. While the portfolio shows the ability to have excess returns, the result shows future improvements in balancing return with risk.

These findings reinforce the potential of using machine learning predictions to guide factor investing strategies, which offer a method for constructing portfolios that have high performance.

Moreover, while the results indicate that the Random Forest model achieved a moderate level of predictive accuracy, there are opportunities for further improvement. The relatively high values of RMSE and MAE for Random Forest highlight the challenges in forecasting financial metrics such as ROIC, which are influenced by a wide range of factors.

Future research could focus on improving these models by incorporating additional features, using them with different machine learning algorithms, and further optimizing hyperparameters.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest concerning the research, authorship and, or publication of this article.

Funding

The authors received no financial support for the research, authorship and, or publication of this article.

References

Agrawal, M., Shukla, P. K., Nair, R., Nayyar, A., & Masud, M. (2021). Stock Prediction Based on Technical Indicators Using Deep Learning Model. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *70*(1), 287. https://doi.org/10.32604/cmc.2022.014637

Almahdi, S. (2015). Smart Beta Portfolio Optimization. *Journal of Mathematical Finance*, *5*(2), 202. https://doi.org/10.4236/jmf.2015.52019

Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance*, *23*(4), 589. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Arnott, R. D., Hsu, J., & Moore, P. (2005). Fundamental Indexation. *Financial Analysts Journal*, *61*(2), 83. https://doi.org/10.2469/faj.v61.n2.2718

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2018). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, *47*, 552. https://doi.org/10.1016/j.najef.2018.06.013

Beasley, J. E., Meade, N., & Chang, T.-J. (2003). An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, *148*(3), 621. https://doi.org/10.1016/s0377-2217(02)00425-3

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, *4*, 71. https://doi.org/10.2307/2490171

Benartzi, S., & Thaler, R. H. (2001). Naive Diversification Strategies in Defined Contribution Saving Plans. *American Economic Review*, *91*(1), 79. https://doi.org/10.1257/aer.91.1.79

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5. https://doi.org/10.1023/a:1010933404324

Chen, L., Pelger, M., & Zhu, J. (2023). Deep Learning in Asset Pricing. *Management Science*, *70*(2), 714. https://doi.org/10.1287/mnsc.2023.4695

Chen, W., Zhang, H., & Jia, L. (2022). A novel two-stage method for well-diversified portfolio construction based on stock return prediction using machine learning. *The North American Journal of Economics and Finance*, *63*, 101818. https://doi.org/10.1016/j.najef.2022.101818

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*,

20(3), 273. https://doi.org/10.1007/bf00994018

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support Vector Regression Machines. *Neural Information Processing Systems*, *9*, 155. https://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf

Huang, C. (2011). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, *12*(2), 807. https://doi.org/10.1016/j.asoc.2011.10.009

Huang, M., Dang, S., & Bhuiyan, M. A. (2025). Multi-objective portfolio optimization for stock return prediction using machine learning. *Expert Systems with Applications*, *298*, 129672. https://doi.org/10.1016/j.eswa.2025.129672

Kamdem, J. S., & Selambi, D. (2022). Cyber-Risk Forecasting using Machine Learning Models and Generalized Extreme Value Distributions. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.archives-ouvertes.fr/hal-03814979

Ma, Y., Han, R., & Wang, W. (2020). Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, *165*, 113973. https://doi.org/10.1016/j.eswa.2020.113973

Malkiel, B. G. (2014). Is Smart Beta Really Smart? *The Journal of Portfolio Management*, *40*(5), 127. https://doi.org/10.3905/jpm.2014.40.5.127

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395

Richard, J.-C., & Roncalli, T. (2015). Smart Beta: Managing Diversification of Minimum Variance Portfolios. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2595051

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy Prediction Using Machine Learning Techniques. *Journal of Risk and Financial Management*, *15*(1), 35. https://doi.org/10.3390/jrfm15010035

Sun, H., & Yu, B. (2019). Forecasting Financial Returns Volatility: A GARCH-SVR Model. *Computational Economics*, *55*(2), 451. https://doi.org/10.1007/s10614-019-09896-w

Ta, V.-D., Liu, C.-M., & Addis, D. (2018). *Prediction and Portfolio Optimization in Quantitative Trading Using Machine Learning*

Techniques. 98. https://doi.org/10.1145/3287921.3287963

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Zhao, C., Yang, S., Qin, C., Zhou, J., & Chen, L. (2023). A Novel Smart Beta Optimization Based on Probabilistic Forecast. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *75*(1), 477. https://doi.org/10.32604/cmc.2023.034933

Zhu, S., Zhou, C., Liu, H., & Ren, Y. (2023). Commodity factor investing via machine learning. *Pacific-Basin Finance Journal*, *83*, 102231. https://doi.org/10.1016/j.pacfin.2023.102231

Bibliographic information of this paper for citing:

Salehirad, Fatemeh & Tondnevis, Farid (2025). Smart-Beta Portfolio Optimization Using Machine Learning Techniques. *Iranian Journal of Finance*, 9(4), 91-116.