

Research Article

Predicting the Return of an Equity Investment Fund Using the Random Forest Algorithm: A Case Study of a Fund Managed by Charisma Asset Management

Running title: Random Forest Prediction of Equity Fund Returns

Hossein Eslami Mofid Abadi^{1*} • Amirali Nematzadeh²

¹ Department of Financial and Banking Management, Central Tehran Branch (CT.C.), Islamic Azad University, Tehran, Iran

² Financial Engineering Program, Central Tehran Branch (CT.C.), Islamic Azad University, Tehran, Iran

* Corresponding author: Hossein Eslami Mofid Abadi, hossein.eslamimofidabadi@iau.ac.ir, ORCID: 0000-0003-4067-6622

ARTICLE INFORMATION

Manuscript ID

[IJFTGRS-XXXX]

Article type

Research Article

Received

7 July 2026

Revised

23 July 2026

Accepted

1 September 2026

Published

1 October 2026

DOI

<https://doi.org/10.22034/ijftgrs.2026.252602>

KEYWORDS

Equity investment fund; Fund return; Iranian capital market; Machine learning; Random Forest; Time-series cross-validation

JEL CLASSIFICATION

C45; C53; G11; G17; G23

ABSTRACT

PURPOSE: This study evaluates the ability of the Random Forest algorithm to predict the daily return of an equity investment fund managed by Charisma Asset Management, compares it with linear regression and a single decision tree as reference models, and identifies the most influential predictors of fund return.

DESIGN/METHODOLOGY/APPROACH: Daily data for the Karis (Sepehr Charisma) equity fund from Farvardin 1399 to Esfand 1403 (2020–2025) were cleaned by removing non-trading days and verified data-recording errors, resulting in 1,019 actual trading days. Fund-return, market-return, and asset-composition data were matched by date, and return-based features were constructed using only lagged information, leaving 989 observations for model estimation. Fund-specific and market variables were used to predict the total daily return. The models were trained using a chronological split, tuned with five-fold time-series cross-validation, and evaluated using RMSE, MAE, R^2 , and directional accuracy.

FINDINGS: On the test set, linear regression achieved the highest predictive performance ($R^2 = 0.721$), whereas Random Forest produced comparable results ($R^2 = 0.680$) and outperformed the single decision tree. In five-fold time-series cross-validation, Random Forest achieved the highest mean R^2 and the lowest mean RMSE. Market return and short-term momentum variables were the dominant predictors. Excluding the contemporaneous market return reduced the R^2 of both models to approximately 0.28. Because no formal forecast-comparison tests were conducted, these comparisons are descriptive.

RESEARCH IMPLICATIONS: For asset managers and investors, daily fund returns are driven mainly by systematic market movements, so the information set matters more than algorithmic complexity. The single-fund design limits generalizability, the asset-allocation variables varied little over time, which may have reduced their predictive contribution, and macroeconomic predictors were not available at a daily frequency.

ORIGINALITY/VALUE: This study provides a case study of daily return prediction for a specific Iranian equity fund using Random Forest, with particular emphasis on performance variation across time-series validation folds and on the distinction between contemporaneous explanation and genuine forecasting.

RESEARCH HIGHLIGHTS

- Market return was the strongest predictor of daily equity fund returns.
- Linear regression outperformed Random Forest on the held-out test set.
- Random Forest achieved the highest mean R^2 in time-series cross-validation.



Authors retain the copyright and full publishing rights.

Published by Avidan Pooya Publication. This article is an open access article licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0)

1. Introduction

Predicting the return of financial assets has long been one of the most fundamental and persistently debated problems in finance. The efficient market hypothesis argues that prices reflect all available information, so that returns follow a random walk and the systematic prediction of excess returns is impossible (Fama, 1970); yet a growing body of evidence shows that advanced computational methods can extract part of the return dynamics that linear models fail to capture (Gu *et al.*, 2020). Within this context, investment funds serve as key financial intermediaries that channel retail savings into the capital market and give investors who lack the expertise or time for direct trading access to professional asset management (Bargos & Claro Romão, 2025). The rapid growth in the number and size of these funds in recent years has made them one of the main vehicles of indirect investment in the Iranian economy and has heightened the importance of studying their return behaviour.

Among the various fund types, equity funds allocate the bulk of their assets to listed shares and therefore combine high return potential with greater volatility and risk—a feature that makes predicting the return of an equity fund a critical problem for investors, asset managers, and regulators (Bian & Lin, 2025). The return of an equity fund is the outcome of many factors, ranging from fund-specific characteristics such as size, asset composition, and portfolio turnover to market variables such as the return and volatility of the overall index (Bargos & Claro Romão, 2025). This multifactor nature and the complex interactions among these variables make modelling fund returns difficult and call for flexible approaches (Gu *et al.*, 2020).

Predicting the return of an equity fund in the Iranian capital market is challenging because fund performance is influenced by complex interactions between fund-specific characteristics and market conditions. The Iranian capital market also operates within a volatile macroeconomic environment characterized by inflation, exchange-rate fluctuations, oil-price shocks, and economic sanctions. These macroeconomic conditions provide the broader context for the study; however, they are not directly incorporated into the empirical models because consistent observations at the required daily frequency were not available. Accordingly, the present analysis focuses on fund-specific and market-based variables that are observable at a daily frequency. The potentially nonlinear relationships and interactions among these variables motivate the use of flexible machine-learning methods alongside conventional benchmark models.

Over the past two decades, machine-learning algorithms have emerged as a powerful alternative to traditional methods in finance, because they can extract complex, nonlinear patterns from data without imposing restrictive assumptions about the data distribution or functional form (Gu *et al.*, 2020). Their theoretical foundation lies in statistical learning theory, which provides a framework for generalizing models from training data to unseen data (Vapnik, 1998). Among these algorithms, the Random Forest introduced by Breiman (2001) is notable for combining an ensemble of decision trees; it offers robustness against overfitting, the ability to model nonlinear relationships and interactions, and a built-in measure of variable importance. These properties make Random Forest well suited to volatile markets such as Iran's, where interpretability and the identification of the drivers of return are as important as predictive accuracy itself.

Despite the rapid expansion of machine-learning applications in Iranian finance, domestic research has concentrated mainly on predicting financial crises, bankruptcy, fraud, and risk (Naderi *et al.*, 2025; Taheri Haftasiabi *et al.*, 2023). International studies indicate that machine-learning methods can capture complex and nonlinear patterns in financial return prediction more effectively than conventional linear approaches (Gu *et al.*, 2020), and previous studies have reported promising applications of Random Forest and other machine-learning methods in predicting stock and fund returns (Bargos & Claro Romão, 2025; Bian & Lin, 2025; Fischer & Krauss, 2018). Evidence from the mutual fund literature also shows that fund-specific characteristics, including fund momentum and flows, may help explain differences in fund performance (Kaniel *et al.*, 2023). In contrast, the Iranian machine-learning literature reviewed in this study has primarily examined financial crises, bank bankruptcy, banking risk, fraud detection, and related problems (Hashemi Golsefidi *et al.*, 2021; Naderi *et al.*, 2025; Soleymani *et al.*, 2023; Taheri Haftasiabi *et al.*, 2023). Comparatively limited attention has therefore been devoted to modelling the daily return of an individual Iranian equity investment fund by jointly considering fund-specific and market-based variables.

Accordingly, the main objective of this study is to evaluate the predictive performance of the Random Forest algorithm in modelling the daily total return of the Karis equity fund and to compare it with two benchmark models:

linear regression and a single decision tree. The study further examines variation in model performance across time-series validation folds and evaluates the relative importance of fund-specific and market-based predictors. Rather than presenting the model as a definitive investment decision-making tool, the study aims to provide evidence on the predictive capability, stability, and interpretability of Random Forest under the available information set. The remainder of the paper is organized as follows. Section 2 reviews the theoretical background and literature and states the research questions, Section 3 describes the data and methodology, Section 4 reports the results, Section 5 discusses the findings, and Section 6 concludes.

2. Theoretical Background and Literature Review

2.1. Conceptual and theoretical foundations

Predicting the return of securities has always been a central concern of financial scholars and investors, and the foundational theories of finance offer different frameworks for understanding this phenomenon. In his classic review of efficient capital markets, Fama (1970) argued that prices reflect all available information, so that changes in returns follow a random walk and the systematic prediction of excess returns is impossible. In contrast, Markowitz's modern portfolio theory showed that a rational investor can, through diversification, select an optimal combination of return and risk, thereby laying the foundation for the quantitative analysis of return and risk (Fabozzi *et al.*, 2008; Markowitz, 1952). Sharpe (1964) subsequently introduced the capital asset pricing model, establishing a linear relationship between expected return and systematic risk (beta), and Fama and French (1993) extended this into multifactor models by adding size and value factors to explain the cross-section of stock returns. These frameworks shaped the predominantly linear paradigm for explaining returns—a paradigm that, despite its explanatory power, faces serious limitations in capturing nonlinear relationships and complex interactions among variables.

In the specific domain of investment funds, the central question is whether fund performance is persistent and predictable or merely the result of chance. By introducing the alpha measure, Jensen (1968) provided a framework for assessing risk-adjusted fund performance and showed that, on average, fund managers are unable to earn persistent excess returns. Carhart (1997), by adding a momentum factor to the three-factor model and examining the persistence of fund performance, found that most of the persistence in performance stems from common risk factors and expenses rather than from managers' persistent skill. While classical asset-pricing theories provide the conceptual foundation for understanding the determinants of fund performance, they generally rely on linear relationships between explanatory variables and expected returns. In contrast, machine-learning methods are designed to capture nonlinear relationships and complex interactions without imposing a predetermined functional form. Consequently, machine-learning approaches should be viewed as complementary to, rather than replacements for, classical financial theories, particularly when the underlying data-generating process is expected to exhibit nonlinear behaviour (Kaniel *et al.*, 2023).

2.2. Machine learning and the Random Forest algorithm

As the limitations of linear models in dealing with the dynamic, nonlinear, and nonparametric nature of financial markets became apparent, the literature gradually turned toward machine-learning methods. The theoretical foundation of these methods lies in statistical learning theory, which formalizes the concepts of generalizability, the bias–variance trade-off, and the control of model complexity (Vapnik, 1998). On this basis, a family of supervised algorithms was developed and brought together in a unified framework in the reference text on statistical learning (Hastie *et al.*, 2009). Among these methods, ensemble approaches—which aggregate the outputs of several base models—gained a special place because of their variance reduction and improved generalizability (Zhou, 2025). Comprehensive reviews of forecasting have likewise documented the growing role of these methods across a wide range of economic and financial forecasting problems (Petropoulos *et al.*, 2022).

The Random Forest algorithm, introduced by Breiman (2001), is a prominent example of tree-based ensemble methods. By building a large number of decision trees on bootstrap samples of the data and selecting a random subset of variables at each node, it reduces the correlation among trees and limits overfitting; the final prediction is then obtained by averaging (in regression problems) or majority voting (in classification problems). Three features make this algorithm suitable for financial data: its ability to capture nonlinear relationships and interactions without an

explicit specification of the functional form; its robustness to outliers and noise, which are common in volatile markets; and its variable-importance measure, which aids model interpretability and the identification of the key drivers of return (Breiman, 2001; Gu *et al.*, 2020). This interpretability makes Random Forest an attractive choice for supervisory and decision-support applications in asset management.

2.3. Empirical evidence and research gap

At the international level, growing evidence has accumulated on the usefulness of machine learning in return prediction. In a large-scale study of a broad cross-section of U.S. stocks, Gu *et al.* (2020) showed that machine-learning models—especially tree-based methods and neural networks—substantially outperform linear methods in asset pricing, with the strongest predictive signals attached to momentum, liquidity, and volatility characteristics. In the fund domain, Kaniel *et al.* (2023), using neural networks, showed that fund characteristics can persistently distinguish high- from low-performing funds and that fund momentum and flows are the most important predictors of risk-adjusted performance, whereas the characteristics of the stocks held in the portfolio have little predictive power. Bian and Lin (2025), applying Random Forest to predict the quarterly returns and volatilities of NASDAQ 100 stocks, confirmed the usefulness of the algorithm for portfolio optimization. In a study of Brazilian equity funds, tree-based models such as Random Forest and extremely randomized trees achieved accuracy above 90% in classifying future fund returns and trained substantially faster than gradient boosting (Bargos & Claro Romão, 2025). Fischer and Krauss (2018), by contrast, found that long short-term memory (LSTM) networks outperformed memory-free methods, including Random Forest, in predicting the directional movements of S&P 500 constituents, although Random Forest outperformed the remaining benchmark methods.

In the domestic literature, the application of machine learning in finance and banking has expanded rapidly, although its focus has been mainly on predicting crises, bankruptcy, and risk. Taheri Haftasiabi *et al.* (2023), using the slope of the yield curve and a bank credit index, predicted Iran's financial crisis with a machine-learning approach, and in a subsequent study proposed a hybrid model combining artificial intelligence and spatial econometrics to predict bank bankruptcy, in which risk contagion among neighbouring banks played an important role (Haftasiabi *et al.*, 2025). Soleymani *et al.* (2023) used structural equation modelling to anticipate financial crises in banks. Naderi *et al.* (2025) found that Random Forest and support vector machines performed better in predicting operational risk in the banking industry, and Ehsani and Hosseini (2025) showed that tree-based classifiers combined with feature optimization achieve high accuracy in predicting customer churn. Hashemi Golsefidi *et al.* (2021) applied machine-learning methods to detect falsification in the financial statements of listed companies, and Asgarian *et al.* (2021) modelled the likelihood of systemic banking crises in developing countries.

Overall, previous studies in the Iranian capital market have primarily focused on bankruptcy prediction, financial distress, banking crises, operational risk, and fraudulent financial reporting rather than on equity investment funds. The existing literature also suggests that the predictive performance of machine-learning algorithms depends not only on the selected modelling technique but also on the characteristics of the available data and the validation framework employed (Bian & Lin, 2025; Fischer & Krauss, 2018; Gu *et al.*, 2020). While international studies have demonstrated the potential advantages of machine-learning methods in large and heterogeneous datasets, limited evidence is available for equity investment funds in the Iranian capital market using rigorous time-series validation procedures (Bergmeir *et al.*, 2018; Cerqueira *et al.*, 2020). Accordingly, this study addresses the following research questions:

RQ1. *How accurately does Random Forest predict the daily total return of the Karis equity fund compared with linear regression and a single decision tree?*

RQ2. *How stable is the predictive performance of the competing models across time-series validation folds?*

RQ3. *What is the relative importance of market-based and fund-specific predictors, and how much predictive power remains when only information available before the prediction date is used?*

3. Data and Methodology

3.1. Research design and analytical framework

This applied, descriptive–correlational study was designed to predict the daily return of an equity investment fund managed by Charisma Asset Management using the Random Forest algorithm within a time-series framework. A machine-learning approach, and ensemble algorithms such as Random Forest in particular, was adopted because of its ability to model complex nonlinear relationships and interactions among financial variables and to achieve satisfactory generalizability in the presence of a large number of daily observations (Breiman, 2001; Gu *et al.*, 2020). The study period covers Farvardin 1399 to Esfand 1403 (March 2020 to March 2025), with data collected at a daily frequency. The conceptual framework of the research is presented in Figure 1.

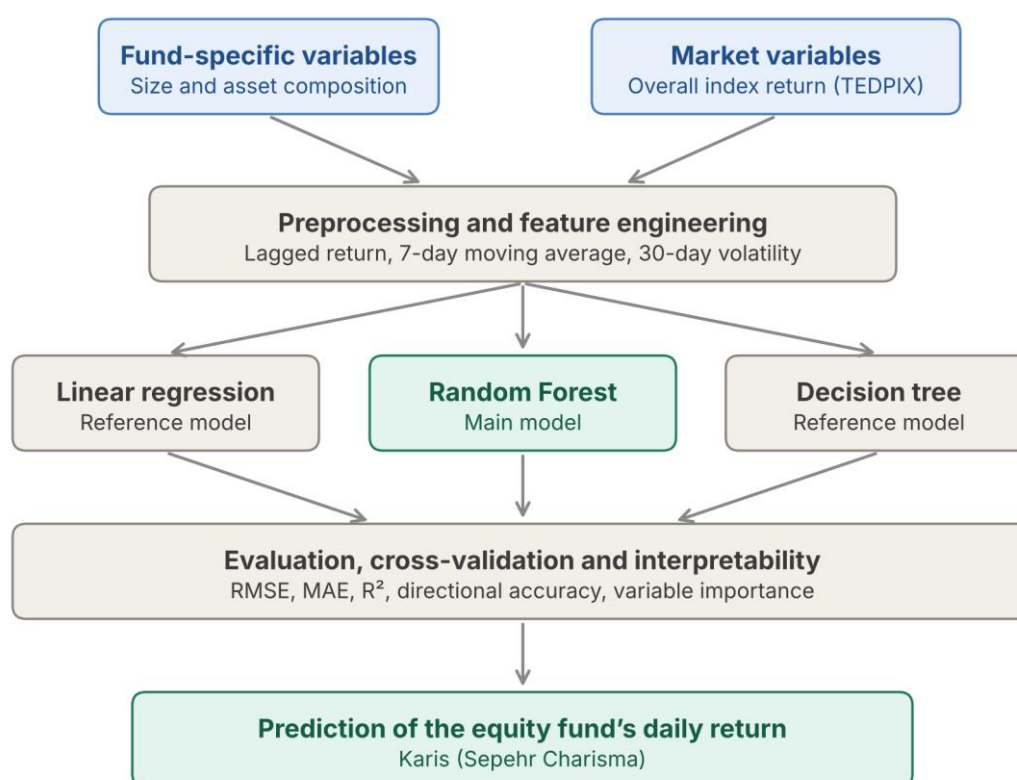


Figure 1. Conceptual framework of the research

Notes: TEDPIX = Tehran Stock Exchange overall index. Return-based features use only information available before the prediction date.

3.2. Population, sample, data sources, and study period

The statistical population comprises all equity investment funds active in the Iranian capital market. The sample was selected purposively and consists of one equity fund managed by Charisma Asset Management. The inclusion criteria were as follows: (a) the fund is of the equity type; (b) it was continuously active throughout the study period (Farvardin 1399 to Esfand 1403); and (c) complete daily net asset value (NAV) data and asset-composition reports were available. The selected fund is Karis (Sepehr Charisma), an equity exchange-traded fund established on 15/07/1392 (7 October 2013). Given the daily frequency of the data, the sample size equals the number of trading days over the study period, providing sufficient observations for model training and testing (Table 1).

Data on the fund's NAV and daily return were obtained from the official website of Charisma Asset Management, and data on the Tehran Stock Exchange indices were collected from the Tehran Securities Exchange Technology Management Company (TSETMC). Data quality was assessed in terms of accuracy, consistency, completeness, and timeliness to minimize measurement error (Géron, 2022; Han *et al.*, 2012; Kuhn & Johnson, 2019).

Table 1. Profile of the studied fund

Attribute	Value
Fund name	Karis (Sepehr Charisma)
Fund type	Equity ETF
Fund manager	Charisma Asset Management
Inception date	15/07/1392 (7 October 2013)
Study period	Farvardin 1399 – Esfand 1403 (March 2020 – March 2025)
Data frequency	Daily
Observations after cleaning	1,019 trading days
Observations used for model estimation	989 trading days

Notes: ETF = exchange-traded fund. Dates follow the Iranian (Solar Hijri) calendar.

3.3. Variables, constructs, and measurement

The main dependent variable is the total daily return of the fund. Daily fund returns were calculated from the official total-return NAV reported by Charisma Asset Management, as in Equation (1), where NAV_t is the net asset value per unit at the end of trading day t . Because the published NAV already incorporates all cash distributions, no additional dividend adjustment was required. Observations corresponding to non-trading days were excluded, and returns were calculated only for actual trading days.

$$R_t = \frac{NAV_t - NAV_{t-1}}{NAV_{t-1}} \times 100 \quad (1)$$

The main objective is to estimate the conditional expectation of the current day's return given a set of predictors, as expressed in Equation (2), where X_t is the vector of explanatory variables and $f(\cdot)$ is the mapping function estimated by the learning algorithm:

$$R_t = f(X_t) + \varepsilon_t \quad (2)$$

The explanatory variables were selected on the basis of established financial theories and prior empirical evidence on investment fund performance. The market return (overall market index return) was included because asset-pricing theories identify systematic market movements as a primary determinant of expected returns (Fama & French, 1993; Sharpe, 1964). The lagged fund return, the 7-day moving average, and the 30-day historical volatility were included to capture short-term return persistence and recent market dynamics documented in the mutual fund literature (Carhart, 1997). Fund-specific variables, including fund size, the equity ratio, and the cash and deposit ratio, were selected because previous studies suggest that fund characteristics and portfolio composition influence fund performance, risk exposure, and investment outcomes (Berk & Green, 2004; Kaniel *et al.*, 2023).

In addition to the one-day lagged return, two historical return-based features were constructed. To ensure that only information available before the prediction date was used, the features were calculated after the removal of non-trading days and were shifted by one trading day. The 7-day moving average (MA7) was calculated as the arithmetic mean of the previous seven daily fund returns, as shown in Equation (3).

$$MA_{7,t} = \frac{1}{7} \sum_{i=1}^7 R_{t-i} \quad (3)$$

The 30-day historical volatility (VOL30) was calculated as the sample standard deviation of the previous thirty daily fund returns, as shown in Equation (4), where $\bar{R}_{t-30:t-1}$ is the mean fund return over the preceding 30 trading days.

$$VOL_{30,t} = \sqrt{\frac{1}{29} \sum_{i=1}^{30} (R_{t-i} - \bar{R}_{t-30:t-1})^2} \quad (4)$$

Neither feature contains the contemporaneous return R_t . Because a complete 30-trading-day historical window was required, the first 30 trading-day observations were excluded from model estimation. Table 2 summarizes the operational definitions of the variables.

Table 2. Operational definition and measurement of the research variables

Category	Symbol	Variable	Definition / measurement	Source
Dependent variable	R_t	Total daily fund return	Daily percentage change in the fund's official total-return NAV, as defined in Equation (1)	Charisma Asset Management
Fund-specific	$SIZE_t$	Fund size	Natural logarithm of the fund's total net asset value on trading day t	Charisma Asset Management
Fund-specific	EQ_t	Equity ratio	Market value of equity holdings divided by total fund assets on trading day t	Asset-composition reports
Fund-specific	DEP_t	Cash and deposit ratio	Value of cash and bank deposits divided by total fund assets on trading day t	Asset-composition reports
Return-based	R_{t-1}	Lagged fund return	Fund return on the immediately preceding trading day	Computed from fund-return data
Return-based	$MA_{7,t}$	7-day moving average	Arithmetic mean of fund returns over the preceding seven trading days ($t - 7$ to $t - 1$)	Computed from fund-return data
Return-based	$VOL_{30,t}$	30-day historical volatility	Sample standard deviation of fund returns over the preceding 30 trading days ($t - 30$ to $t - 1$)	Computed from fund-return data
Market	RM_t	Overall market index return	Daily percentage return of TEDPIX on trading day t	TSETMC

Notes: NAV = net asset value; TSETMC = Tehran Securities Exchange Technology Management Company; TEDPIX = Tehran Stock Exchange overall index.

The asset-composition dataset contained one observation for each calendar date in the study period and was matched with the fund-return dataset by date. Following the removal of non-trading days, the corresponding asset-composition observations were retained. No forward-fill, interpolation, or other temporal imputation procedure was applied.

3.4. Model specification and estimation procedure

The main model is the Random Forest algorithm, evaluated alongside two reference models—multiple linear regression and a single decision tree—for performance comparison. Random Forest uses bootstrap sampling to create different training subsets and averages the predictions of all trees, as in Equation (5), where T_b is the b -th tree and B is the total number of trees. The relative importance of each variable is computed from the mean decrease in impurity.

$$\hat{R}_t^{RF} = \frac{1}{B} \sum_{b=1}^B T_b(X_t) \quad (5)$$

As a reference, the multiple linear regression model is defined as in Equation (6), where β_0 is the intercept, β_j the coefficient of the j -th variable, p the number of predictors, and ε_t the error term.

$$R_t = \beta_0 + \sum_{j=1}^p \beta_j X_{j,t} + \varepsilon_t \quad (6)$$

The data-analysis procedure comprised six main steps. First, the raw dataset, which initially contained 1,827 calendar-day observations, was cleaned before feature construction. Observations corresponding to non-trading days were removed because they did not represent actual market sessions. In addition, three observations containing clearly erroneous market-return values were excluded after examination of the original records. No winsorization, trimming,

or general statistical outlier rule was applied to valid return observations, and no temporal imputation was used. After cleaning, 1,019 actual trading-day observations remained. Second, feature engineering included the construction of the one-day lagged return, the 7-day moving average, and the 30-day historical volatility, using only information available before the prediction date. Third, a chronological split was used to divide the observations into training and test sets, and time-series cross-validation was conducted exclusively within the training sample to prevent future information from entering model estimation. Fourth, the model hyperparameters were selected within the training sample; the final settings adopted for the Random Forest model are presented in Table 3.

Table 3. Final hyperparameter settings of the Random Forest model

Hyperparameter	Description	Final value
n_estimators	Number of trees in the forest	200
max_depth	Maximum depth of each decision tree	5
min_samples_split	Minimum number of samples required to split an internal node	10
min_samples_leaf	Minimum number of samples required at a terminal node	1
max_features	Number of predictor variables considered at each split	All features

Notes: Hyperparameters were selected using five-fold time-series cross-validation within the training sample. Parameter names follow the scikit-learn implementation.

Fifth, the models were trained and evaluated using the root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and directional accuracy (DA), as defined in Equations (7)–(10):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^n (R_k - \hat{R}_k)^2} \quad (7)$$

$$\text{MAE} = \frac{1}{n} \sum_{k=1}^n |R_k - \hat{R}_k| \quad (8)$$

$$R^2 = 1 - \frac{\sum_{k=1}^n (R_k - \bar{R})^2}{\sum_{k=1}^n (R_k - \bar{R})^2} \quad (9)$$

$$\text{DA} = \frac{1}{n} \sum_{k=1}^n \mathbb{I}(\text{sign}(R_k) = \text{sign}(\hat{R}_k)) \quad (10)$$

where R_k and $\hat{R}_k$ denote the actual and predicted returns, $\bar{R}$ the mean actual return, $\mathbb{I}(\cdot)$ the indicator function, and n the number of test observations. Sixth, variable-importance analysis was conducted using the mean decrease in impurity to interpret the model and identify the most influential predictors of return.

3.5. Validity, robustness, and reproducibility

All data preprocessing, feature construction, model training, hyperparameter tuning, and performance evaluation were implemented in Python using the pandas and scikit-learn libraries. Two procedures were used to assess the robustness of the results. First, five-fold time-series cross-validation within the training sample was used to evaluate the stability of model performance across consecutive validation folds (Bergmeir *et al.*, 2018; Cerqueira *et al.*, 2020). Second, a genuine forecasting specification that excludes the contemporaneous market return was estimated to distinguish contemporaneous explanation from prediction based solely on information available before the prediction date.

4. Results

4.1. Descriptive statistics and preliminary analysis

Following the data-cleaning procedure described in Section 3.4, the cleaned dataset contained 1,019 actual trading-day observations. After the 30-day historical volatility variable was constructed and observations without a complete historical window were removed, 989 observations remained for model estimation. Table 4 reports the descriptive statistics of the variables used in the empirical analysis.

Table 4. Descriptive statistics of the research variables

Variable	Mean	Median	Std. dev.	Minimum	Maximum
Fund daily return (%)	0.179	0.150	1.531	−5.52	5.43
Overall index return (%)	0.178	0.070	1.462	−5.15	11.64
Lagged return (%)	0.109	0.000	1.373	−5.52	5.43
7-day moving average (%)	0.115	0.030	0.620	−2.03	2.50
30-day volatility (%)	1.130	1.054	0.435	0.36	2.48
Equity ratio (%)	92.45	93.00	5.03	71.30	99.69
Cash and deposit ratio (%)	2.69	2.13	2.72	0.00	18.88
Fund size (log net assets)	16.39	16.33	0.35	14.65	17.23

Notes: Std. dev. = standard deviation. Returns, ratios, and volatility are expressed in percent; fund size is the natural logarithm of total net assets.

The daily fund return averaged 0.179%, whereas the average daily return of the overall market index was 0.178%. The equity ratio averaged 92.45%, indicating that the fund maintained a predominantly equity-based portfolio throughout the sample period. The descriptive statistics also show sufficient variation in the explanatory variables to support the subsequent predictive modelling.

4.2. Main findings

As reported in Table 5, linear regression achieved the best test-set performance, with an R^2 of 0.721, an RMSE of 0.578, an MAE of 0.442, and a directional accuracy of 84.85%. Random Forest produced an R^2 of 0.680, an RMSE of 0.620, an MAE of 0.476, and a directional accuracy of 80.30%. Both models performed better than the single decision tree. Because no formal pairwise forecast-comparison test was conducted, these differences are interpreted descriptively and should not be regarded as evidence of statistically significant superiority.

Table 5. Performance comparison of the models on the test set

Model	RMSE	MAE	R^2	Directional accuracy
Linear regression (reference)	0.578	0.442	0.721	84.85%
Single decision tree (reference)	0.743	0.576	0.539	74.24%
Random Forest (main)	0.620	0.476	0.680	80.30%

Notes: RMSE = root mean squared error; MAE = mean absolute error; R^2 = coefficient of determination. RMSE and MAE are expressed in percentage points of daily return.

According to Table 6, the overall market index return was by far the dominant predictor, with a relative importance of 0.888. It was followed by the lagged fund return and the 7-day moving average, with relative importance values of 0.053 and 0.024, respectively. The remaining variables, including historical volatility, fund size, and the asset-composition ratios, made comparatively limited contributions to the Random Forest predictions.

Table 6. Variable importance in return prediction (Random Forest)

Variable	Relative importance	Rank
Overall index return	0.888	1
Lagged return (previous day)	0.053	2
7-day moving average of return	0.024	3
30-day volatility	0.013	4
Fund size	0.011	5
Cash and deposit ratio	0.008	6
Equity ratio	0.004	7

Notes: Relative importance is the mean decrease in impurity, normalized to sum to one.

4.3. Robustness, sensitivity, and additional analyses

As reported in Table 7, Random Forest achieved the highest mean cross-validation R^2 (0.628) and the lowest mean RMSE (0.812). The decision tree produced the lowest standard deviation of R^2 across the validation folds, whereas linear regression showed the greatest variation in performance. However, no formal variance-equality test was conducted; therefore, these differences are interpreted descriptively and should not be regarded as statistical evidence of superior model stability.

Table 7. Five-fold time-series cross-validation results

Model	Mean R^2	Std. dev. of R^2	Mean RMSE
Random Forest	0.628	0.091	0.812
Decision tree	0.587	0.065	0.849
Linear regression	0.466	0.329	0.903

Notes: Results are computed within the training sample using five consecutive validation folds.

As shown in Table 8, excluding the contemporaneous market return reduced the R^2 of both models to approximately 0.28. Directional accuracy was 63.2% for Random Forest and 65.7% for linear regression. Because no one-sample proportion test against the 50% benchmark was conducted, these results are reported descriptively and should not be interpreted as statistically significant evidence of predictive performance above random chance.

Table 8. Performance in the genuine forecasting setting (excluding the contemporaneous market return)

Model	R^2	RMSE	Directional accuracy
Random Forest	0.284	0.919	63.2%
Linear regression	0.291	0.914	65.7%

Notes: The predictor set excludes the contemporaneous overall index return and contains only information available before the prediction date.

5. Discussion

5.1. Interpretation relative to theory and prior evidence

The findings indicate that the daily return of the examined equity fund was driven primarily by broad market movements rather than by slowly changing fund-specific characteristics. The dominant contribution of the overall market return is economically plausible because the fund maintained a consistently high equity allocation throughout the study period, causing systematic movements in the Tehran Stock Exchange to be transmitted directly to the fund's total-return NAV. This interpretation is consistent with classical asset-pricing theory, which identifies systematic market risk as the principal determinant of diversified equity portfolio returns (Fama & French, 1993; Sharpe, 1964). It also agrees with empirical evidence showing that market-wide information explains a substantial proportion of short-horizon return variation in machine-learning applications to financial markets (Gu *et al.*, 2020).

The comparison of the competing models provides additional insight into the characteristics of the underlying data. Although Random Forest achieved the highest average coefficient of determination in time-series cross-validation,

linear regression produced the strongest performance on the final holdout test set. This outcome suggests that the relationship between the selected explanatory variables and fund returns remained largely linear during the evaluation period, limiting the additional benefit of a more flexible nonlinear algorithm. Previous research has shown that machine-learning methods generally provide their greatest advantages when financial relationships contain substantial nonlinearities, complex interactions, and high-dimensional predictor spaces (Gu *et al.*, 2020; Hastie *et al.*, 2009). In the present study, the dominant influence of the contemporaneous market return appears to have reduced the opportunity for nonlinear methods to generate meaningful improvements in predictive accuracy. Because no formal forecast-comparison procedures, such as the Diebold–Mariano test, or variance-equality tests were conducted, the observed differences among the competing models should be interpreted descriptively.

The variable-importance analysis further highlights the relative contribution of different sources of information. Besides the market return, short-term momentum indicators, including the lagged return and the 7-day moving average, were the most influential predictors of daily fund performance. This finding is consistent with the literature documenting short-term return persistence and momentum effects in financial markets (Jegadeesh & Titman, 1993; Moskowitz *et al.*, 2012). Similarly, Kaniel *et al.* (2023) reported that momentum-related fund characteristics contribute meaningfully to explaining differences in mutual fund performance. By contrast, the comparatively low importance of the equity ratio, the cash and deposit ratio, and fund size should not be interpreted as evidence that portfolio allocation is economically unimportant. Rather, these variables exhibited relatively limited variation, because the fund remained predominantly invested in equities throughout most of the sample period, and broad asset-allocation ratios are likely to influence medium-term performance more strongly than daily return fluctuations. Unlike factor-model studies such as Carhart (1997), which evaluate systematic exposures to market, size, value, and momentum factors, the present study relied on aggregate portfolio characteristics together with short-term return variables; the relatively low importance of the asset-allocation variables therefore reflects differences in variable definition, sampling frequency, and modelling objectives rather than inconsistency with the broader literature.

An additional contribution of the study is the distinction between contemporaneous explanatory performance and genuine forecasting based exclusively on previously available information. When the contemporaneous market return was excluded from the predictor set, the predictive performance of both Random Forest and linear regression declined substantially. This finding suggests that a considerable proportion of the explanatory power observed in the main specification depended on information available on the same trading day rather than *ex ante*. Such a result is broadly consistent with the efficient market hypothesis, which holds that future returns are difficult to predict using publicly available historical information alone (Fama, 1970). Although both models achieved directional accuracy exceeding 60% in the genuine forecasting specification, no one-sample proportion test against the 50% benchmark was conducted; consequently, these results should not be regarded as statistical evidence that the models predicted return direction better than chance.

5.2. Theoretical and methodological contributions

More broadly, the findings indicate that forecasting performance depends not only on algorithmic sophistication but also on the information content of the available predictors. Increasing model complexity alone cannot compensate for an information set that omits economically relevant variables. In the Iranian capital market, macroeconomic conditions, exchange-rate fluctuations, policy interventions, inflation expectations, and sanctions frequently influence investor expectations and market dynamics. Consequently, improvements in forecasting performance may depend more on expanding the available information set than on adopting increasingly complex machine-learning algorithms, a conclusion consistent with recent developments in empirical asset pricing and forecasting research (Gu *et al.*, 2020; Petropoulos *et al.*, 2022). Methodologically, the study illustrates the importance of chronological splitting, time-series cross-validation, and an explicit separation between contemporaneous and *ex ante* predictors when machine-learning models are applied to daily fund returns.

5.3. Managerial, regulatory, investment, and policy implications

From a practical perspective, Random Forest provides an interpretable modelling framework through its variable-importance analysis, allowing researchers and asset managers to examine the relative contribution of market-wide and fund-specific variables. However, the present findings should not be interpreted as evidence that the models constitute economically profitable investment or trading strategies. The genuine forecasting experiment demonstrated

substantially lower predictive performance once contemporaneous information was excluded, and the study did not evaluate transaction costs, bid–ask spreads, portfolio turnover, execution constraints, or risk-adjusted investment performance. For regulators and investors, the dominance of the market factor implies that the daily return risk of a highly invested equity fund is largely systematic and cannot be substantially reduced through fund-level allocation decisions over short horizons. Accordingly, the results should be interpreted as evidence on statistical predictive performance under the available information set rather than as support for implementable investment decisions.

6. Conclusion

This study evaluated the predictive performance of the Random Forest algorithm for modelling the daily return of an Iranian equity investment fund and compared it with linear regression and a single decision tree. Random Forest produced predictive performance comparable to the benchmark models but did not outperform linear regression on the final test set. The dominant role of the overall market return indicates that systematic market movements were the primary driver of daily fund performance, whereas fund-specific allocation variables contributed comparatively little at the daily frequency examined. Furthermore, the substantial decline in predictive performance after the contemporaneous market return was excluded demonstrates the importance of distinguishing contemporaneous explanatory performance from genuine forecasting based solely on previously available information. Overall, the study contributes empirical evidence on the application of machine-learning methods to equity fund return prediction in the Iranian capital market and shows that the usefulness of increasingly sophisticated algorithms depends fundamentally on the quality, timing, and economic relevance of the available information. The findings should be interpreted as descriptive rather than as evidence of statistical superiority or economic profitability.

6.1. Limitations and future research

Several limitations should be acknowledged. First, the analysis was based on a single equity investment fund, limiting the generalizability of the findings to other funds, management styles, and market conditions. Second, the principal empirical model included the same-day market return, so part of the reported explanatory performance reflects contemporaneous association rather than *ex ante* prediction. Third, the empirical models did not incorporate macroeconomic or policy-related variables, such as exchange-rate movements, commodity prices, interest-rate conditions, liquidity indicators, inflation expectations, and policy interventions, which were unavailable at a consistent daily frequency. Fourth, the sample period included exceptional events, including the 2020 market boom and subsequent decline, the COVID-19 pandemic, and continuing economic sanctions, which may have affected return dynamics. Fifth, the study evaluated statistical forecasting performance only; economic profitability, transaction costs, and risk-adjusted investment performance were not assessed. Finally, model comparisons were based on descriptive performance measures; formal procedures such as the Diebold–Mariano test, variance-equality tests, and one-sample proportion tests for directional accuracy were beyond the scope of the study.

Future research may extend the analysis in several directions. First, examining multiple equity funds within a panel-data framework would allow broader generalization and comparisons across investment strategies and management styles. Second, future studies may incorporate additional predictors, including daily exchange-rate movements, commodity prices, liquidity indicators, market sentiment, monetary-policy announcements, and other macroeconomic variables aligned with their actual information-release dates to avoid look-ahead bias. Third, gradient-boosting algorithms such as XGBoost may be compared with Random Forest under identical time-series validation procedures, whereas sequence-based deep-learning models such as LSTM should be evaluated only after substantially expanding the dataset, because reliable estimation of highly parameterized neural networks requires considerably larger samples. Fourth, explainable artificial intelligence techniques such as SHAP may be employed to investigate whether the importance of market-wide and fund-specific predictors changes across bullish, bearish, and highly volatile regimes. Finally, statistical forecasting evaluation should be complemented with formal forecast-comparison tests and assessments of the economic value of predictive models under realistic trading conditions.

Declarations and Transparency Statements

Author contributions (CRediT)

Amirali Nematzadeh: Conceptualization, Methodology, Data curation, Formal analysis, Writing—original draft. Hossein Eslami Mofid Abadi: Supervision, Validation, Writing—review and editing. Both authors have read and approved the final version of the manuscript.

Funding

This research received no external funding.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval

Not applicable. The study uses publicly available fund and market data and does not involve human participants or animals.

Informed consent

Not applicable, because the study did not involve human participants or personally identifiable data.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Fund NAV data are published by Charisma Asset Management, and market index data are available from TSETMC.

Code availability

The code used for data preprocessing, model estimation, and performance evaluation is available from the corresponding author upon reasonable request.

Use of generative AI and AI-assisted technologies

The authors used an AI-based assistant for language editing and for assistance in drafting and structuring the manuscript. An AI-assisted tool was also used during editorial preparation for copy-editing, formatting to the journal template, and checking the bibliographic details of the references. All research design decisions, the data analysis and its validation, the interpretation of results, and the final conclusions were independently reviewed, verified, and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

Acknowledgments

Not applicable.

References

- Asgarian, M. R., Daei Karimzadeh, S., & Sharifi Renani, H. (2021). Predicting the likelihood of systematic banking crises in selected developing countries (An approach from the multinomial logit model). *Quarterly Journal of Applied Theories of Economics*, 7(4), 117–138. [In Persian] https://ecoj.tabrizu.ac.ir/article_12364.html
- Bargos, F. F., & Claro Romão, E. (2025). Enhanced forecasting of equity fund returns using machine learning. *Mathematical and Computational Applications*, 30(1), Article 9. <https://doi.org/10.3390/mca30010009>
- Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>
- Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6), 1269–1295. <https://doi.org/10.1086/424739>
- Bian, J., & Lin, J. (2025). Optimizing investment strategies: A random forest approach to stock return prediction and portfolio management. In *Proceedings of the International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)* (pp. 674–681). Atlantis Press. https://doi.org/10.2991/978-94-6463-652-9_70
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57–82. <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>
- Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11), 1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>
- Ehsani, F., & Hosseini, M. (2025). Customer churn analysis using feature optimization methods and tree-based classifiers. *Journal of Services Marketing*, 39(1), 20–35. <https://doi.org/10.1108/JSM-04-2024-0156>

- Fabozzi, F. J., Markowitz, H. M., & Gupta, F. (2008). Portfolio selection. In F. J. Fabozzi (Ed.), *Handbook of finance* (Vol. 2, pp. 3–13). John Wiley & Sons. <https://doi.org/10.1002/9780470404324.hof002001>
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.2307/2325486>
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. <https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/>
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
- Haftasiabi, R. T., Naderi, A., & Piri, P. (2025). Hybrid model for predicting bankruptcy in Iranian banks: Combining artificial intelligence with spatial econometrics. *Quarterly Journal of Quantitative Economics*, 21(3), 1–29. [In Persian] <https://doi.org/10.22055/jqe.2025.48811.2684>
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. <https://doi.org/10.1016/C2009-0-61819-5>
- Hashemi Golsefid, A., Lashgari, Z., & Hajiha, Z. (2021). The application of machine learning model for detection of falsification of accounting. *Journal of Management Accounting and Auditing Knowledge*, 10(37), 271–283. [In Persian] https://www.jmaak.ir/article_17193_en.html
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. <https://doi.org/10.1111/j.1540-6261.1993.tb04702.x>
- Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2), 389–416. <https://doi.org/10.1111/j.1540-6261.1968.tb00815.x>
- Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1), 94–138. <https://doi.org/10.1016/j.jfineco.2023.07.004>
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press. <https://doi.org/10.1201/9781315108230>
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91. <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>
- Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, 104(2), 228–250. <https://doi.org/10.1016/j.jfineco.2011.11.003>
- Naderi, H., Rastegar Sorkhe, M. A., Ostadi, B., & Kargari, M. (2025). Predicting the likelihood of operational risk occurrence in the banking industry using machine learning algorithms. *Journal of Asset Management and Financing*, 13(4), 77–96. <https://doi.org/10.22108/amf.2025.143030.1929>
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., ... Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, 38(3), 705–871. <https://doi.org/10.1016/j.ijforecast.2021.11.001>
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442. <https://doi.org/10.1111/j.1540-6261.1964.tb02865.x>
- Soleymani, M., ArabMazar Yazdi, M., Shekarkhah, J., & Safarzadeh, M. H. (2023). Anticipation of financial crisis in banks by structural equations model. *Financial Management Strategy*, 11(2), 1–28. https://jfm.alzahra.ac.ir/article_7051.html
- Taheri Haftasiabi, R., Piri, P., Naderi, A., & Esmaily, N. (2023). Forecasting the financial crisis in Iran according to the slope of the yield curve and the bank credit index: A machine learning approach. *Iranian Journal of Economic Studies*, 12(1), 217–248. <https://doi.org/10.22099/ijes.2024.51275.1973>
- Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons. <https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034>
- Zhou, Z.-H. (2025). *Ensemble methods: Foundations and algorithms* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781003587774>