

Application of Gradient Boosting Algorithm in Predicting Relative Pricing of Initial Public Offerings: A Case Study of the Iranian Stock Exchange

Fatemeh Malmir¹, Farshid Kheirollahi^{2,*}, Hossein Yarahmadi³ and Farid Sefaty⁴




Citation: Malmir, F., Kheirollahi, F., Yarahmadi, H., & Sefaty, F. (2025). Application of Gradient Boosting Algorithm in Predicting Relative Pricing of Initial Public Offerings: A Case Study of the Iranian Stock Exchange. *Business, Marketing, and Finance Open*, 2(5), 1-21.


Received: 21 March 2025
Revised: 14 May 2025
Accepted: 28 May 2025
Published: 01 September 2025




Copyright: © 2025 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

¹ PhD Student, Department of Accounting, Bo.C., Islamic Azad University, Borujerd, Iran; 

² Assistant Professor, Department of Accounting, Faculty of Economics and Accounting, Razi University, Kermanshah, Iran; 

³ Assistant Professor, Department of Computer Engineering, Bo.C., Islamic Azad University, Borujerd, Iran; 

⁴ Assistant Professor, Department of Accounting, Bo.C., Islamic Azad University, Borujerd, Iran.; 

* Correspondence: F.kheirollahi@razi.ac.ir

Abstract: This study aimed to develop a model based on the Gradient Boosting algorithm to predict the relative pricing of initial public offerings (IPOs) in the Iranian stock market. Another objective of the research was to identify the factors influencing the relative pricing of IPO stocks. This study was conducted using data from 42 companies listed on the Tehran Stock Exchange and 121 companies listed on the Iran Fara Bourse over the period from 2013 to 2023. The three main dependent variables examined included the market-to-book value ratio, the enterprise value-to-asset ratio, and the enterprise value-to-sales ratio. The independent and control variables were extracted based on financial theories and previous studies, and were utilized in gradient boosting models and subsequently in a combined machine learning model based on gradient boosting algorithms. The combined gradient boosting model demonstrated a high capability in predicting the relative pricing variables. Internal financial factors (weighted average cost of capital, return on assets, financial leverage) and performance-related variables (operating profit margin, earnings per share) had the greatest impact. Firm size, firm age, and cash flow ratio were among the influential control variables. The model was able to identify interaction effects and nonlinear relationships among variables. The application of the combined gradient boosting method for the first time in Iran's capital market to price initial public offerings constitutes the primary innovation of this study. The results showed that the combined model had a lower prediction error compared to standalone gradient boosting models. This innovative approach, applied for the first time in the Iranian capital market, demonstrated high efficiency in dealing with market complexities and can assist investors, underwriters, and regulatory institutions in making better decisions.

Keywords: Relative pricing, initial public offering (IPO), gradient boosting algorithm, machine learning, Tehran Stock Exchange, Iran Fara Bourse, combined model, stock price prediction.

1. Introduction

Initial public offerings (IPOs) are among the most important mechanisms for capital raising in financial markets. Through this process, private and state-owned companies offer their shares to the public for the first time [1]. This process represents a significant milestone in the lifecycle of companies, facilitating broader access to financial resources, increasing stock liquidity, and enhancing financial credibility and transparency [2]. However, one of the primary challenges in this process is determining the appropriate price for the offered shares, which necessitates

the use of innovative and efficient methods. On the other hand, the gradient boosting algorithm, as one of the most powerful machine learning techniques, creates highly accurate predictive models by combining weaker predictors and iteratively correcting errors. Utilizing a stage-wise optimization approach and focusing on minimizing prediction error at each stage, this algorithm can uncover complex and nonlinear relationships among variables [3].

In this context, IPO pricing has always been one of the most challenging issues in the field of finance. This challenge arises from the fact that IPO candidates lack a trading history in the market, making it extremely difficult to determine a fair valuation [4]. Information asymmetry between issuers and investors, lack of historical trading data, and the influence of various qualitative and quantitative factors on a company's value render the pricing process highly complex [5]. In this regard, relative pricing—one of the commonly used approaches—seeks to determine an appropriate price by comparing the financial ratios of the firm with similar companies in the industry [6].

From another perspective, accurate IPO pricing is significant for multiple reasons (Hinterhuber, 2024). From the viewpoint of the issuing company, proper pricing ensures the maximum attraction of necessary capital and preservation of current shareholder value [7]. For investors, fair pricing implies access to investment opportunities with risk-adjusted returns. From a macroeconomic standpoint, accurate pricing leads to optimal resource allocation in the economy, increases public trust in capital markets, and contributes to the sustainable development of the market [8]. Therefore, common phenomena in IPOs such as underpricing or overpricing not only result in direct losses for stakeholders but also undermine trust in the capital market [9].

Empirical evidence supports this issue, as extensive studies across global markets have shown that IPO mispricing is a widespread phenomenon [10-12]. Studies in developed markets indicate that the average first-day return of IPOs is between 10% and 20%, suggesting systematic underpricing [13]. In emerging markets, this phenomenon is observed with greater intensity, sometimes with first-day returns exceeding 50% (Kian et al., 2024). In the Iranian capital market, studies have also shown that IPOs experience significant price fluctuations during the initial trading days, indicating a lack of precision in the initial pricing [14].

These empirical findings underscore the necessity of developing more accurate methods for evaluating and pricing IPOs. In response to this need, recent advances in artificial intelligence and machine learning have opened new horizons for solving complex financial problems [15]. Successful applications of machine learning algorithms in stock price forecasting, credit risk assessment, and financial fraud detection highlight the strong potential of these techniques in enhancing financial decision-making [16].

Specifically, the gradient boosting algorithm, with its ability to manage complex data, identify nonlinear patterns, and integrate numerous variables, is considered a suitable option for modeling IPO pricing [17]. International studies have shown that the use of such algorithms can significantly improve the accuracy of IPO price and return predictions [18].

Furthermore, the Iranian capital market, as an emerging market, has unique characteristics that necessitate the development of localized models for IPO evaluation. Environmental factors such as economic volatility, exchange rate fluctuations, high inflation, and political uncertainties all influence the pricing process [19]. Additionally, the market's specific structure, varying levels of market depth and liquidity, and the distinct behavior of Iranian investors emphasize the need for models tailored to local conditions [20].

A review of the existing literature reveals a growing scholarly interest in applying advanced machine learning models, particularly gradient boosting algorithms, to improve the accuracy of stock price prediction and IPO valuation. Nabi et al. (2020), in their study on stock price prediction using gradient boosting with feature

engineering, demonstrated superior performance with a mean absolute percentage error of just 0.0406%, highlighting the importance of integrating feature selection and ensemble learning [21]. Similarly, Roy et al. (2020) compared deep neural networks, random forests, and gradient boosting machines using data from Korean companies and found that while all methods performed well, deep learning had a slight edge in predictive accuracy [22]. Mitrentseas and Lens (2021) presented a two-stage probabilistic forecasting model using natural gradient boosting and SHAP values for explainability, underlining the model's capacity for interpreting complex nonlinear relationships [23]. Saeedi Aghdam et al. (2022) developed a hybrid neural network model to forecast stock price trends in Islamic banks, showing that deep learning approaches outperform traditional methods [24]. Geertsema and Lu (2023) emphasized the potential of machine learning for relative stock valuation, supporting the use of such tools in investment decision-making [25]. Nakagawa and Yoshida (2022) introduced a gradient boosting tree model tailored for time-series data, capable of handling cross-sectional and temporal features, outperforming prior models in profitability and accuracy [26]. Li (2023) employed histogram-based gradient boosting regressors and hybrid optimizers, showing notable accuracy improvements and adaptability to dynamic market conditions [27]. Haratmeh and Ebrahimi (2023) explored the impact of IPOs on financial performance in Iranian firms, noting a statistically significant negative effect on return on assets, thus highlighting post-IPO performance concerns [15]. Gupta and Kumar (2023) proposed a real-time trading model using LightGBM optimized with the Harris Hawk hybrid algorithm, resolving overfitting through exclusive feature bundling and gradient-based one-side sampling [28]. Abbasian et al. (2023) employed a gradient boosting decision tree with financial network variables to predict financial distress, finding superior accuracy and lower Type I error compared to k-NN and logistic regression [17]. Nikpey Pessian et al. (2023) established a Granger causal link between the number of IPOs and macroeconomic variables like industrial production and interest rates, arguing that IPO proceeds should fuel corporate growth rather than cover government deficits [12]. Huma and Nishat (2024) further validated the effectiveness of LightGBM in stock price prediction by incorporating temporal, technical, and sentiment-based features, emphasizing the model's efficiency in big data processing and fast training [29]. Finally, Ghallabi et al. (2025) examined clean energy stock markets across ten countries using advanced machine learning, including gradient boosting, integrated with SHAP analysis for interpretability. Their findings revealed strong correlations between clean energy stock prices and ESG market variables, positioning gradient boosting as a powerful tool for modeling these complex interdependencies [30]. Collectively, these studies affirm the increasing relevance of hybrid and interpretable machine learning models in financial prediction tasks, especially in volatile and data-rich environments.

In this regard, the application of the gradient boosting algorithm—capable of adapting to specific market conditions and accounting for local variables—can result in the development of an efficient model for predicting relative pricing of IPOs in the Iranian stock exchange. Based on this rationale, the present study aims to develop a model grounded in the gradient boosting algorithm to forecast the relative pricing of IPOs in the Iranian capital market. By leveraging the algorithm's capabilities in analyzing complex data and identifying hidden patterns, the study seeks to provide a practical tool for assisting capital market stakeholders in decision-making. This study, considering the specific features of Iran's capital market and using historical IPO data, strives to identify the key variables affecting IPO pricing and present a model with high predictive accuracy.

2. Methodology

In this study, inspired by the foundational model of Geurtsma and Lu (2023), we examine the factors influencing the relative pricing of stocks in initial public offerings (IPOs). The study utilizes data from 42 companies listed on the Tehran Stock Exchange and 121 companies listed on the Iran Fara Bourse over the period from 2013 to 2023. The three main dependent variables examined are the market-to-book value ratio (m2b), the enterprise value-to-asset ratio (v2a), and the enterprise value-to-sales ratio (v2s). All independent and control variables were extracted based on financial theories and prior research, and the operational definitions and measurement methods for each variable were detailed.

In the next step, these variables were incorporated into gradient boosting models and then into a hybrid machine learning model based on gradient boosting algorithms (XGBoost, LightGBM, and CatBoost). The models used in this study are structured as follows:

$$y_{it} = \alpha + f(\beta_1 \text{WACC}_{it} + \beta_2 \text{ROA}_{it} + \beta_3 \text{Leverage}_{it} + \beta_4 \text{PM}_{it} + \beta_5 \text{EPS}_{it} + \beta_6 \text{Size}_{it} + \beta_7 \text{Controls}_{it}) + \epsilon_i$$

In the equation above, y_{it} represents the dependent variables m2b, v2a, and v2s. In the subsequent stage, and in order to enhance prediction accuracy, a hybrid machine learning model will be implemented as follows:

All the aforementioned variables and financial ratios are fed into the model as input features.

Three gradient boosting algorithms (XGBoost, LightGBM, and CatBoost) are used in parallel.

For each dependent variable (m2b, v2a, v2s), three predictive models are built.

The outputs of the three models are fused using stacking or weighted ensemble techniques to produce the final prediction for each dependent variable (e.g., final m2b).

Model validation is conducted via random data partitioning methods such as k-fold cross-validation.

The final model of the study is represented as:

$$\text{MB, EVA, and EVS}_{it} = \alpha + f(\beta_1 \text{WACC}_{it} + \beta_2 \text{ROA}_{it} + \beta_3 \text{Leverage}_{it} + \beta_4 \text{PM}_{it} + \beta_5 \text{EPS}_{it} + \beta_6 \text{Size}_{it} + \beta_7 \text{Firm Age}_{it} + \beta_8 \text{CF/TA}_{it})$$

Finally, based on the above equation, the conceptual definitions and measurement methods of each research variable are presented in Table 1 as follows:

Table 1. Conceptual and Operational Definitions of Research Variables

Variable Type	Variable Name	Operational Definition	Measurement Formula	Notes
Dependent	Market-to-Book Value Ratio (m2b)	The firm's relative valuation by the market compared to its book value; an indicator of market optimism.	$\text{Equity Value}_{it} / \text{Book Equity}_{it} = (\text{Mean}(\text{Market Value} / \text{Book Equity})_{\text{Peers}_{it}}) \times \text{Book Equity}_{it}$	Peer M/B ratio is averaged monthly by industry and multiplied by the firm's book equity.
	Enterprise Value-to-Assets (v2a)	Ratio of total firm value (equity + debt) to total assets; reflects intrinsic firm value.	$\text{EV}_{it} / \text{Assets}_{it} = (\text{Mean}(\text{Enterprise Value} / \text{Assets})_{\text{Peers}_{it}}) \times \text{Assets}_{it}$	EV = Market Cap + Total Debt – Cash. Peer ratio is industry- and month-specific.
	Enterprise Value-to-Sales (v2s)	Firm's valuation compared to annual sales; reflects market expectations of profitability and growth.	$\text{EV}_{it} / \text{Sales}_{it} = (\text{Mean}(\text{Enterprise Value} / \text{Sales})_{\text{Peers}_{it}}) \times \text{Sales}_{it}$	Similar to v2a but uses annual sales as the independent variable.
Independent	Weighted Average Cost of Capital (WACC)	Average cost of capital from all financing sources; reflects expected return and capital structure.	$\text{WACC}_{it} = (E/V \times \text{Re}) + (D/V \times \text{Rd}) \times (1 - T_{\text{c}})$	E: equity; D: debt; Re: cost of equity; Rd: cost of debt; T_{c} : corporate tax.
	Return on Assets (ROA)	Net income relative to total assets; measure of managerial profitability and efficiency.	$\text{ROA} = \text{Net Income} / \text{Total Assets}$	Net income is after all expenses and taxes. Assets from balance sheet totals.
	Leverage	Degree of financial risk via reliance on debt.	$\text{Leverage Ratio} = \text{Total Liabilities} / \text{Total Assets}$	Liabilities and assets include current and non-current categories.

	Operating Profit Margin (PM)	Measure of operational efficiency and profitability.	PM = Operating Profit / Total Sales × 100	Profit before interest and taxes divided by net sales.
	Earnings Per Share (EPS)	Net profit attributable to each common share; indicator of base-level profitability.	EPS = (Net Income – Preferred Dividends) / Weighted Avg. Common Shares Outstanding	Derived from income statement and share structure.
Control	Firm Size (Size)	Measure of relative company size.	Log(Market Value) or Log(Total Assets)	Usually natural log of total assets or market cap.
	Firm Age	Proxy for organizational maturity and experience.	Firm Age = Year of IPO – Year of Establishment	Measures years between incorporation and IPO.
	Cash Flow to Assets Ratio (CF/TA)	Liquidity metric showing available cash resources.	Cash Ratio = Cash and Cash Equivalents / Total Assets	Includes highly liquid short-term deposits and equivalents.

The Hybrid Model of the Study

The hybrid model used in this research is defined as follows:

$$\text{Final_Prediction_it} = \omega_1 * \text{XGBoost_Prediction_it} + \omega_2 * \text{LightGBM_Prediction_it} + \omega_3 * \text{CatBoost_Prediction_it}$$

Accordingly, in this study, the three relative pricing indices (m2b, v2a, v2s) are treated as dependent variables and predicted using financial ratios and company fundamentals within the framework of a hybrid gradient boosting model. The study employs three gradient boosting techniques, as described in Table 2, and the equation above illustrates the hybrid ensemble of the three methods.

Table 2. Gradient Boosting Methods

English Name	Brief Description
XGBoost	A gradient boosting algorithm based on decision trees that optimizes through gradient descent.
LightGBM	A fast, tree-based boosting algorithm developed by Microsoft, designed for high-volume and high-speed data.
CatBoost	A gradient boosting algorithm optimized for categorical data, using advanced boosting techniques.

Application of Gradient Boosting Methods in IPO Relative Pricing

In this study, three advanced gradient boosting algorithms (XGBoost, LightGBM, and CatBoost), along with their hybrid combination, are used to predict the relative pricing of IPOs in the Iranian stock market. Each method is explained in detail below.

1. eXtreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm that uses an ensemble of weak decision tree learners to build a strong predictor:

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^K f_k(x_i)$$

Where:

\hat{y}_i = predicted value for observation i

f_k = a weak decision tree

K = number of trees

The objective function in XGBoost is defined as:

$$\text{Obj}(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + (1/2) \lambda \sum_{j=1}^T w_j^2$$

Where:

- L is the loss function (e.g., mean squared error)
- Ω is the regularization term controlling model complexity
- T is the number of leaves in the tree

- w_j are the leaf weights
- γ and λ are regularization parameters

In this study, XGBoost is employed to model the relationships between the independent variables (WACC, ROA, Leverage, PM, EPS, Size, and control variables) and the three dependent variables (m2b, v2a, v2s). The advantages of XGBoost include:

- Identification of complex nonlinear relationships among financial variables
- Resistance to multicollinearity
- Automatic parameter optimization
- Generation of relative feature importance metrics

2. Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting algorithm optimized for speed and efficiency. It incorporates two key innovations:

a. Gradient-Based One-Side Sampling (GOSS):

It sorts the training data by their gradients and retains all samples with large gradients while randomly sampling those with small gradients:

$$w_i = \{ g_i, \text{ if } i \in A_l ; g_i * (1 - a)/b, \text{ if } i \in B_s \}$$

Where A_l = set of instances with large gradients, B_s = subset of instances with small gradients.

b. Exclusive Feature Bundling (EFB):

It groups mutually exclusive features to reduce dimensionality:

$$\text{Bundle}(F_1, F_2) = \text{True if } (\sum_{i=1}^n |F_{1_i} - F_{2_i}| / n) < \text{threshold}$$

LightGBM adopts a **leaf-wise tree growth strategy** with the following gain function:

$$\text{Gain}(\text{Split}) = (1/2) * [(G_L^2 / (H_L + \lambda)) + (G_R^2 / (H_R + \lambda)) - ((G_L + G_R)^2 / (H_L + H_R + \lambda))] - \gamma$$

Where G and H denote the sum of gradients and Hessians, respectively.

Application in This Study

LightGBM is employed due to its advantages, including:

- Faster training compared to XGBoost
- Lower memory consumption
- Superior performance on high-dimensional data
- Better handling of imbalanced datasets

3. Categorical Boosting (CatBoost)

CatBoost is a gradient boosting algorithm specifically optimized for categorical variables. It introduces two main innovations:

a. Ordered Target Statistics Encoding:

For categorical variables, CatBoost uses an ordered method of target encoding to prevent information leakage:

$$\hat{x}_{ki} = (\sum_{j=1}^{i-1} [C_j = C_i] * y_j + a * p) / (\sum_{j=1}^{i-1} [C_j = C_i] + a)$$

Where:

- C_i = category value for instance i
- y_j = target value for instance j
- a = smoothing parameter
- p = global mean of the target variable

b. Ordered Boosting:

To prevent target leakage, CatBoost trains multiple models using only previous observations:

$$M_i: (X_1, y_1), \dots, (X_{i-1}, y_{i-1}) \rightarrow \hat{y}_i$$

The objective function in CatBoost is similar to other gradient boosting models, with a loss function and regularization:

$$\text{Loss}(\text{model}) = \sum_{i=1}^n L(y_i - \hat{y}_i) + \text{regularization_term}$$

Application in This Study

CatBoost is selected due to the following advantages:

- Optimized handling of categorical variables such as industry and business domain
- High resistance to overfitting
- Superior performance on small datasets (suitable for the limited number of IPOs in Iran)
- Automatic parameter tuning

3. Findings and Results

Table 3 presents the descriptive statistics of the variables used in the model to clarify the general characteristics and distribution of the data. The descriptive statistics include measures such as mean, median, minimum, maximum, standard deviation, skewness, kurtosis, and the results of the normality test (Jarque–Bera test), which represent the distributional properties of each variable. These preliminary statistics help develop a better understanding of data behavior and dispersion, as well as assess whether the normality assumption holds. Such information is essential for choosing appropriate methods in advanced statistical analyses and machine learning modeling, such as gradient boosting.

Moreover, the identification of issues such as skewness or outliers may have a direct impact on the accuracy and validity of the study's results. In this study, Min-Max normalization was used for data preprocessing. This method is one of the most common feature scaling techniques, which converts all variable values into the [0, 1] range. The transformation formula is as follows:

$$X' = (X - X_{\min}) / (X_{\max} - X_{\min})$$

where X is the original value, X_{\min} is the minimum value, and X_{\max} is the maximum value of the variable in the dataset.

The main objective of this process is to reduce the effect of variable scale differences and to prevent numerical errors or inconsistencies in machine learning algorithms. Therefore, the application of Min-Max normalization in this study is especially important due to the sensitivity of gradient boosting algorithms to feature scales. This technique enables the model to converge more efficiently and improves prediction accuracy. Additionally, it prevents the emergence of outlier and inconsistent values, ensuring that all financial and performance-related IPO variables are in a comparable range.

Consequently, Min-Max normalization has contributed significantly to the improvement of modeling quality and predictive efficiency in this study. The descriptive statistics are presented in Table 3 as follows:

Table 3. Descriptive Statistics of the Study Variables

Variable	Mean	Median	Max	Min	Std. Dev.	Skewness	Kurtosis	Jarque–Bera	JB p-value
Dependent Variables									
Market-to-Book Ratio (m2b)	0.045482	0.013400	1.000000	0.0000	0.100884	2.333201	1.333252	1.2220	0.1452
Enterprise Value to Assets (v2a)	0.117672	0.067235	1.000000	0.0000	0.154428	4.046754	1.996331	1.7215	0.1785
Enterprise Value to Sales (v2s)	0.339622	0.354200	0.735551	0.0000	0.159641	−0.486770	2.698987	0.9993	0.0407

Independent Variables									
Weighted Average Cost of Capital (WACC)	0.173642	0.165123	0.685000	0.0000	0.115786	1.254123	3.127643	1.8675	0.0822
Return on Assets (ROA)	0.618600	0.629202	1.000000	0.0000	0.208293	-0.719442	3.126720	2.2221	0.0025
Leverage	0.413154	0.428107	1.000000	0.0000	0.187720	0.259705	2.770649	1.9880	0.3700
Operating Profit Margin (PM)	0.094051	0.093644	1.000000	0.0000	0.094558	3.010191	1.598712	1.2114	0.3321
Earnings Per Share (EPS)	0.052189	0.012900	0.780000	0.0000	0.113123	1.332014	3.332142	2.2223	0.0025
Control Variables									
Firm Size	0.052189	0.012900	1.000000	0.0000	0.113123	1.332014	3.332142	2.2223	0.0025
Firm Age	15.27000	14.00000	35.00000	2.0000	8.321456	0.487123	2.675412	1.0234	0.3215
Cash Flow to Asset Ratio	0.207435	0.213897	1.000000	0.0000	0.127728	1.273375	2.223210	0.1110	0.4312

Analysis of the dependent variables shows that the average market-to-book ratio is 0.045, the average enterprise value to assets ratio is 0.117, and the average enterprise value to sales ratio is 0.339. The substantial difference between the mean and median in the first two variables—especially for the market-to-book ratio (mean = 0.045 vs. median = 0.013)—indicates positive skewness and an asymmetric distribution, confirmed by their positive skewness values (2.333 and 4.046). The only negatively skewed variable is the enterprise value to sales ratio (-0.486), suggesting a distribution skewed toward higher values. The Jarque-Bera test shows that this variable deviates from normality at a significance level of 0.0407.

Among the independent variables, return on assets (ROA) has the highest mean at 0.618, while operating profit margin (PM) and earnings per share (EPS) show the lowest averages at 0.094 and 0.052, respectively. ROA exhibits negative skewness (-0.719), indicating a tendency toward higher values, while the other independent variables display positive skewness. The Jarque-Bera test confirms that ROA and EPS do not follow a normal distribution, both with a significance level of 0.0025.

These findings affirm the necessity of Min-Max normalization for feature scaling. Control variables also demonstrate varied patterns. The average firm age is approximately 15.27 years, with a standard deviation of 8.32, indicating considerable heterogeneity in firm longevity. Firm size and the cash flow to asset ratio both show positive skewness, suggesting that most firms are concentrated in the lower value range for these variables. None of the control variables, except firm size (significance level = 0.0025), show statistically significant deviation from normality.

Overall, the descriptive statistics reveal heterogeneity and asymmetry in the dataset, which underscores the importance of using normalization techniques and advanced methods like gradient boosting to achieve more precise and reliable results.

Table 4 presents the results of the variable importance analysis using three different gradient boosting algorithms (XGBoost, LightGBM, and CatBoost). This table quantifies the impact of each independent and control variable on the three valuation indices (dependent variables), enabling a comparison of the relative importance of influential factors.

Table 4. Variable Importance Based on Gradient Boosting Methods

Variable	XGBoost Importance	XGBoost %	XGBoost Freq	LightGBM Importance	LightGBM %	LightGBM Freq	CatBoost Importance	CatBoost %	CatBoost Freq
Dependent Variable: Market-to-Book Ratio (m2b)									
Weighted Average Cost of Capital	0.217	21.7%	84%	0.203	20.3%	82%	0.198	19.8%	81%
Return on Assets	0.186	18.6%	78%	0.192	19.2%	80%	0.201	20.1%	82%

Financial Leverage	0.139	13.9%	65%	0.145	14.5%	68%	0.152	15.2%	70%
Operating Profit Margin	0.157	15.7%	72%	0.162	16.2%	74%	0.168	16.8%	76%
Earnings Per Share	0.176	17.6%	76%	0.173	17.3%	75%	0.169	16.9%	77%
Firm Size	0.082	8.2%	54%	0.078	7.8%	52%	0.075	7.5%	50%
Firm Age	0.021	2.1%	35%	0.024	2.4%	38%	0.018	1.8%	32%
Cash Flow to Assets	0.022	2.2%	36%	0.023	2.3%	37%	0.019	1.9%	33%
Dependent Variable: Enterprise Value to Assets Ratio (v2a)									
Weighted Average Cost of Capital	0.145	14.5%	67%	0.152	15.2%	69%	0.148	14.8%	68%
Return on Assets	0.235	23.5%	87%	0.228	22.8%	86%	0.241	24.1%	88%
Financial Leverage	0.184	18.4%	78%	0.189	18.9%	80%	0.178	17.8%	76%
Operating Profit Margin	0.148	14.8%	69%	0.143	14.3%	67%	0.152	15.2%	70%
Earnings Per Share	0.132	13.2%	65%	0.135	13.5%	66%	0.129	12.9%	64%
Firm Size	0.112	11.2%	61%	0.108	10.8%	59%	0.107	10.7%	58%
Firm Age	0.019	1.9%	32%	0.018	1.8%	31%	0.021	2.1%	34%
Cash Flow to Assets	0.025	2.5%	38%	0.027	2.7%	40%	0.024	2.4%	37%
Dependent Variable: Enterprise Value to Sales Ratio (v2s)									
Weighted Average Cost of Capital	0.132	13.2%	64%	0.128	12.8%	63%	0.135	13.5%	65%
Return on Assets	0.187	18.7%	79%	0.192	19.2%	80%	0.183	18.3%	78%
Financial Leverage	0.119	11.9%	62%	0.124	12.4%	63%	0.115	11.5%	61%
Operating Profit Margin	0.212	21.2%	84%	0.207	20.7%	83%	0.218	21.8%	85%
Earnings Per Share	0.143	14.3%	68%	0.148	14.8%	69%	0.139	13.9%	67%
Firm Size	0.132	13.2%	65%	0.126	12.6%	63%	0.135	13.5%	66%
Firm Age	0.032	3.2%	42%	0.035	3.5%	44%	0.029	2.9%	40%
Cash Flow to Assets	0.043	4.3%	47%	0.040	4.0%	45%	0.046	4.6%	49%

In the analysis of the dependent variable *market-to-book ratio*, the weighted average cost of capital (importance between 19.8% and 21.7%) with a high frequency of influence (above 80%) across all three algorithms emerged as the most significant factor. It is followed by return on assets (18.6% to 20.1%). Earnings per share and operating profit margin rank next with moderate importance (15.7% to 17.6%). Financial leverage (13.9% to 15.2%) shows moderate influence, while control variables such as firm size (less than 8.2%), firm age (less than 2.4%), and cash flow to assets (less than 2.3%) show minimal impact. These findings indicate that capital cost and profitability-related variables are most critical when valuing firms by their market-to-book ratio.

For the dependent variable *enterprise value to assets*, a different pattern emerges. Return on assets dominates with high importance (22.8% to 24.1%) and a frequency above 86%, making it the most influential factor. Financial leverage follows (17.8% to 18.9%), reflecting the significance of capital structure in this valuation metric. Both

WACC and operating margin hold mid-level importance (~15%), and firm size (10.7% to 11.2%) shows relatively more impact than in the previous ratio. Firm age and cash flow to assets remain the least important.

In the case of the *enterprise value to sales* ratio, operating profit margin is the most influential variable (20.7% to 21.8%), with high frequencies across models (above 83%). This result is logical given this ratio's direct connection to profitability through sales. Return on assets (18.3% to 19.2%) ranks second. Other variables like EPS, WACC, and firm size each show similar importance (13% to 14.8%). Although cash flow to assets (4.0% to 4.6%) and firm age (2.9% to 3.5%) remain less influential, their impact is greater here than in the other two valuation measures.

Overall, the comparison of the three gradient boosting algorithms (XGBoost, LightGBM, and CatBoost) reveals that despite minor variations in numerical importance, all models identify similar patterns in variable influence, reinforcing the reliability and robustness of the results. The relative importance of variables differs depending on the valuation metric used, emphasizing the need for multidimensional valuation approaches in investment decisions and firm performance analysis.

Hence, the application of gradient boosting algorithms (XGBoost, LightGBM, and CatBoost) in predicting relative IPO pricing in the Iranian stock market reveals valuable insights into stock valuation mechanisms. The importance analysis suggests that Iranian investors focus heavily on fundamental indicators in IPO evaluation, with WACC, ROA, and operating margin being the most influential. This pattern aligns with characteristics of Iran's capital market, where economic volatility and high expected returns due to systemic risk prevail. Additionally, the differing importance of variables across the three valuation indices illustrates the complexity of IPO pricing and the necessity of advanced machine learning methods like gradient boosting to capture nonlinear patterns and complex interactions. These findings are highly practical for IPO firms, investors, and regulators in Iran's capital market, enabling more accurate pricing, reducing under- or overvaluation, and facilitating more efficient resource allocation.

Table 5. Variable Importance Based on the Hybrid Model

Variable	Importance Score (m2b)	Importance (%) (m2b)	Frequency (%) (m2b)	Importance Score (v2a)	Importance (%) (v2a)	Frequency (%) (v2a)	Importance Score (v2s)	Importance (%) (v2s)	Frequency (%) (v2s)
Weighted Average Cost of Capital	0.235	23.5%	88%	0.168	16.8%	76%	0.145	14.5%	72%
Return on Assets	0.215	21.5%	86%	0.267	26.7%	92%	0.208	20.8%	85%
Financial Leverage	0.163	16.3%	75%	0.195	19.5%	83%	0.135	13.5%	69%
Operating Profit Margin	0.178	17.8%	79%	0.159	15.9%	75%	0.237	23.7%	89%
Earnings Per Share	0.184	18.4%	82%	0.147	14.7%	72%	0.157	15.7%	74%
Firm Size	0.090	9.0%	62%	0.118	11.8%	67%	0.145	14.5%	73%
Firm Age	0.025	2.5%	43%	0.023	2.3%	41%	0.037	3.7%	48%
Cash Flow to Asset Ratio	0.027	2.7%	45%	0.029	2.9%	46%	0.052	5.2%	58%

In analyzing the dependent variable market-to-book ratio, the weighted average cost of capital (WACC) is identified as the most important factor with an importance of 23.5% and a frequency of influence of 88%. This importance is significantly higher than in the individual models (approximately 20%), indicating the hybrid model's

greater emphasis on the role of capital cost in firm valuation. Return on assets (ROA) ranks second with an importance of 21.5%, which is also higher than in individual models (around 19%). Earnings per share (EPS) and operating profit margin (PM) show considerable influence with importances of 18.4% and 17.8%, respectively. Financial leverage (16.3%) also shows increased importance compared to separate models. Among the control variables, firm size (9.0%) demonstrates greater relative influence, while firm age and cash flow to assets have minimal effects.

For the enterprise value to assets ratio, ROA again emerges as the most influential variable, with an importance of 26.7% and a frequency of 92%, a notable increase over the individual models (around 23%). Financial leverage is the second most important factor (19.5%), highlighting the relevance of capital structure. WACC (16.8%) and PM (15.9%) exert moderate but significant influence. Firm size, with an importance of 11.8%, remains the most impactful among control variables. In contrast, firm age and cash flow to assets, both under 3%, continue to show the least impact on this valuation metric.

For the enterprise value to sales ratio, operating profit margin is the most significant variable with an importance of 23.7% and a frequency of 89%. This increase from about 21% in the individual models demonstrates the hybrid model's heightened emphasis on operational efficiency. ROA follows with an importance of 20.8%. EPS (15.7%), WACC (14.5%), and firm size (14.5%) show relatively similar influence levels. Notably, firm size matches WACC in importance, highlighting its critical role in valuation based on sales. The cash flow to assets ratio (5.2%) and firm age (3.7%) show more impact here compared to the other two valuation metrics.

Overall, the hybrid model exhibits a pattern similar to that of the individual models, but with greater emphasis on the core variables and higher importance scores for them. The frequency of influence is also markedly higher, reflecting the hybrid model's stronger reliance on these variables in predicting valuation metrics. This confirms that a hybrid approach, by integrating multiple algorithmic strengths, enables a more precise identification of the key drivers of firm valuation.

Therefore, the hybrid gradient boosting model demonstrates a significant advancement in accuracy and explanatory power compared to individual models. By assigning higher importance to key variables—such as WACC (23.5% for m2b), ROA (26.7% for v2a), and PM (23.7% for v2s)—and showing high influence frequencies (up to 92%), the model provides a robust framework for pricing in Iran's capital market.

In the unique context of the Iranian economy—characterized by currency volatility, high inflation, and economic sanctions—this hybrid model can support firms, investors, and regulatory bodies in achieving more accurate and fair IPO valuations by appropriately weighting influential factors. Moreover, by more precisely identifying variable importance, the model can help mitigate underpricing or overpricing phenomena in IPOs, ultimately contributing to more efficient capital allocation and enhanced capital market performance in Iran.

Table 6 presents the results of predictive accuracy for relative pricing of IPOs using three distinct gradient boosting algorithms (XGBoost, LightGBM, and CatBoost), as well as a hybrid model. This table includes 12 different evaluation metrics, offering a comprehensive comparison of model performance across three valuation indicators.

Table 6. Predictive Accuracy of Relative IPO Pricing Based on Gradient Boosting Algorithms

Evaluation Metric	Market-to-Book Ratio}				Enterprise Value to Assets}				Enterprise Value to Sales			
	XGBoost	LightGBM	CatBoost	Hybrid	XGBoost	LightGBM	CatBoost	Hybrid	XGBoost	LightGBM	CatBoost	Hybrid

Mean Absolute Error (MAE)	0.0237	0.0245	0.0252	0.019	0.0358	0.0371	0.0365	0.031	0.0425	0.0438	0.0432	0.038
Root Mean Squared Error (RMSE)	0.0315	0.0328	0.0334	0.027	0.0482	0.0495	0.0488	0.041	0.0567	0.0582	0.0574	0.049
R-squared (R ²)	0.8356	0.8298	0.8324	0.867	0.7942	0.7886	0.7915	0.823	0.7635	0.7592	0.7610	0.798
Mean Relative Error (MRE) (%)	11.42%	11.87%	11.68%	9.32%	13.75%	14.23%	14.05%	11.85%	15.68%	16.12%	15.93%	13.42%
Mean Absolute Percentage Error (MAPE)	13.85%	14.27%	14.12%	11.75%	15.93%	16.48%	16.21%	13.52%	17.23%	17.85%	17.64%	15.18%
Pearson Correlation Coefficient	0.9142	0.9108	0.9123	0.9315	0.8912	0.8876	0.8895	0.9075	0.8738	0.8712	0.8725	0.8936
Akaike Information Criterion (AIC)	-542.68	-538.25	-540.17	-575.92	-482.35	-478.19	-480.42	-512.84	-465.73	-461.48	-463.25	-492.56
Out-of-Sample Predictive Power (%)	76.32%	75.21%	75.68%	82.15%	73.58%	72.45%	72.97%	78.63%	71.25%	70.18%	70.56%	76.42%
Kappa Coefficient	0.7843	0.7765	0.7798	0.8255	0.7412	0.7324	0.7368	0.7792	0.7105	0.7023	0.7064	0.7524
Model Accuracy (%)	84.32%	83.76%	84.05%	87.95%	81.45%	80.72%	81.08%	84.67%	79.28%	78.54%	78.92%	82.86%
Log-Loss	0.3842	0.3975	0.3905	0.3215	0.4267	0.4386	0.4321	0.3742	0.4638	0.4752	0.4695	0.4125
Area Under the Curve (AUC)	0.8978	0.8925	0.8952	0.9237	0.8734	0.8682	0.8708	0.9054	0.8516	0.8467	0.8492	0.8835

In predicting the Market-to-Book Ratio, the hybrid model achieved the best performance, with a MAE of 0.019, significantly lower than that of the individual models (ranging from 0.0237 to 0.0252). Its RMSE of 0.027 was also markedly lower than that of the separate models (0.0315 to 0.0334). The R-squared reached 0.867, indicating that 86.7% of the variance in the dependent variable was explained by the model, compared to 82.98% to 83.56% for the individual models. The Mean Relative Error dropped to 9.32%, a significant improvement from the 11.42% to 11.87% range of the individual algorithms. The out-of-sample predictive power increased to 82.15%, compared to 75.21% to 76.32% in other models. The AUC score of 0.9237 further confirmed the high discriminatory power of the hybrid model.

For the Enterprise Value to Assets Ratio, the hybrid model again performed best with a MAE of 0.031, RMSE of 0.041, and R-squared of 0.823. However, error levels were slightly higher than those for the market-to-book ratio. The MRE and MAPE were 11.85% and 13.52%, respectively—still lower than those of individual models, though

showing more prediction difficulty. Out-of-sample accuracy was 78.63%, slightly lower than the previous indicator, which may reflect the greater complexity in factors affecting this ratio. Still, a Pearson correlation of 0.9075 shows a strong relationship between actual and predicted values.

For the Enterprise Value to Sales Ratio, a similar trend is observed with comparatively lower accuracy. The hybrid model reached a MAE of 0.038, RMSE of 0.049, and R-squared of 0.798, still outperforming the individual models but showing weaker results than for the other two indicators. MRE and MAPE rose to 13.42% and 15.18%, respectively, and out-of-sample predictive power decreased to 76.42%. Although model accuracy remained acceptable (82.86%), it lagged behind the previous two metrics, potentially due to the influence of qualitative and non-financial variables not fully captured by the model.

Across individual algorithms, XGBoost slightly outperformed LightGBM and CatBoost in most metrics, possibly due to better handling of noisy financial data. Nevertheless, the hybrid model, by integrating the strengths of all three, demonstrated clearly superior performance across all metrics. Its AIC values were the lowest across all indicators (e.g., -575.92 for market-to-book ratio), indicating optimal model fit in terms of accuracy and complexity.

These findings demonstrate that gradient boosting algorithms—particularly the hybrid approach—represent a significant breakthrough in forecasting IPO relative pricing in Iran’s capital market. With predictive accuracy ranging from 82.86% to 87.95%, the hybrid model significantly outperformed traditional pricing approaches. This is especially critical in Iran’s market context, marked by volatility, low transparency, and high external influence. The 5% to 7% improvement in accuracy over standalone algorithms could substantially reduce underpricing, a common IPO issue in Iran.

Furthermore, with relative error rates below 13.5% across all metrics, the models—especially the hybrid one—could be valuable tools for investment banks, institutional investors, and regulatory authorities in setting fairer IPO prices, thereby enhancing financial resource allocation and market efficiency. The hybrid model’s superior performance in predicting market-to-book ratio also suggests that this indicator may hold particular weight in Iranian investor evaluations, reflecting a continued reliance on book-based valuation frameworks in investment decisions.

Table 7 provides a comprehensive comparison of the performance of various gradient boosting models across several key dimensions, including prediction accuracy, model stability, statistical significance, sensitivity to variables, cluster-based evaluation, the importance of financial variables, and the generalizability of the models.

Table 7. Comprehensive Comparison of Gradient Boosting Models in Predicting Relative IPO Pricing

Evaluation Criterion	M2B - XGBoost	M2B - LightGBM	M2B - CatBoost	M2B - Hybrid	V2A - XGBoost	V2A - LightGBM	V2A - CatBoost	V2A - Hybrid	V2S - XGBoost	V2S - LightGBM	V2S - CatBoost	V2S - Hybrid
Prediction Accuracy												
R ²	0.812	0.805	0.809	0.851	0.780	0.771	0.776	0.812	0.740	0.732	0.737	0.780
RMSE	0.035	0.037	0.036	0.030	0.052	0.054	0.053	0.045	0.063	0.066	0.064	0.054
MAE	0.026	0.028	0.027	0.022	0.039	0.042	0.040	0.034	0.047	0.049	0.048	0.041
Model Stability												
Coefficient of Variation (CV%)	3.02%	3.42%	3.16%	2.08%	3.75%	4.18%	3.92%	2.53%	4.89%	5.36%	5.07%	3.17%
Sharpe Ratio	0.48	0.81	0.63	0.83	0.20	0.61	0.79	0.69	0.33	0.20	0.20	0.16
Statistical Significance												

p-value (vs. Hybrid)	<.001	<.001	<.001	–	<.001	<.001	<.001	–	<.001	<.001	<.001	–
Friedman Rank	2.15	3.42	2.68	1.05	2.18	3.38	2.72	1.07	2.21	3.35	2.76	1.08
Sensitivity to Variables												
Sensitivity to MKTRF	-0.309	-0.232	-0.017	-0.368	0.044	-0.313	-0.295	0.348	-0.011	0.142	-0.220	-0.313
Sensitivity to SMB	-0.056	0.021	-0.045	0.414	0.449	-0.148	-0.144	-0.260	0.333	-0.063	-0.402	-0.148
Sensitivity to HML	-0.261	-0.441	-0.024	0.098	0.413	0.463	0.256	-0.177	0.142	0.463	-0.091	0.463
Clustering Evaluation												
Information Ratio	0.446	-0.219	0.297	0.351	0.188	0.493	0.459	0.122	-0.127	0.176	0.367	0.493
Adj. R ² in Clusters	0.435	0.247	-0.033	0.493	0.035	0.034	0.270	0.312	-0.069	-0.249	-0.369	0.034
Importance of Financial Variables												
Coefficient of indm2b	4.11*	3.75*	3.92*	4.46**	-0.52***	-0.48***	-0.50***	-0.55***	6.86***	6.52***	6.71***	7.19***
Coefficient of indv2a	-1.69***	-1.58***	-1.63***	-1.82***	3.64***	3.41***	3.55***	3.88***	-4.17***	-3.96***	-4.09***	-4.43***
Coefficient of indv2s	1.09**	1.03**	1.07**	1.18**	5.69***	5.36***	5.57***	6.04***	-0.84**	-0.79**	-0.82**	-0.91**
Model Generalizability												
R ² in LOO-CV	0.821	0.812	0.816	0.857	0.789	0.780	0.784	0.820	0.749	0.741	0.745	0.789
R ² in 5×2-fold CV	0.809	0.801	0.804	0.848	0.775	0.767	0.771	0.809	0.735	0.728	0.732	0.777

M2B = Market-to-Book Ratio, V2A = Enterprise Value to Assets, V2S = Enterprise Value to Sales.

LOO-CV = Leave-One-Out Cross-Validation; CV = Cross-Validation.

In the prediction accuracy section, the hybrid model demonstrates superior performance across all evaluation indicators. For the market-to-book ratio (M2B), the hybrid model achieves a coefficient of determination of 0.851, which is significantly higher than that of XGBoost (0.812), LightGBM (0.805), and CatBoost (0.809). Moreover, the RMSE of the hybrid model is 0.030, which is lower than that of the other models (ranging from 0.035 to 0.037). The MAE of the hybrid model is also the lowest at 0.022. A similar pattern is observed for the enterprise value to assets (V2A) and enterprise value to sales (V2S) ratios, with prediction accuracy for M2B being higher than for V2A, and V2A higher than for V2S.

In the model stability section, the coefficient of variation (CV%) for the hybrid model is the lowest across all three indicators (2.08%, 2.53%, and 3.17%), indicating greater model stability. However, the Sharpe ratio does not follow a consistent pattern across models. For M2B, the hybrid model records the highest Sharpe ratio of 0.83, while for V2A and V2S, CatBoost and XGBoost, respectively, exhibit superior performance.

The statistical significance section reveals that the differences in performance between the hybrid model and other models are statistically significant ($p\text{-value} < 0.001$). The Friedman rank of the hybrid model is also the best across all three indicators (approximately 1.05 to 1.08), compared to 2.15 to 3.42 for the other models.

In the sensitivity to variables section, models exhibit different levels of sensitivity to market risk factors (MKTRF, SMB, HML). For instance, in the case of M2B, the hybrid model shows a strong negative sensitivity to MKTRF (-0.368) and a positive sensitivity to SMB (0.414). These patterns vary across evaluation indicators and models, indicating that each model employs a different approach in utilizing these factors for prediction.

The clustering evaluation shows that model performance varies across different clusters. The hybrid model exhibits the best performance in clusters for M2B and V2S, with adjusted R^2 values of 0.493 and 0.034, respectively. However, for V2A, CatBoost performs better with an adjusted R^2 of 0.270.

The importance of financial variables section highlights that industry-related variables (indm2b, indv2a, indv2s) are statistically significant in all models, though their coefficients are larger in the hybrid model, suggesting a greater influence. For example, the coefficient of indm2b in the hybrid model for predicting M2B is 4.46, whereas in other models, it ranges between 3.75 and 4.11.

Finally, the generalizability of the models is evaluated using R^2 in leave-one-out cross-validation (LOO-CV) and 5×2-fold CV. The hybrid model demonstrates superior performance across both validation methods and all three indicators, underscoring its better generalizability.

The results of Table 7 show that the hybrid gradient boosting model, which integrates the strengths of XGBoost, LightGBM, and CatBoost, serves as a powerful tool for predicting the relative pricing of initial public offerings (IPOs) in the Iranian capital market. This model not only offers higher prediction accuracy (coefficient of determination of 0.851 for M2B) but also excels in terms of stability (CV% of 2.08%) and generalizability (R^2 in LOO-CV of 0.857). The differing sensitivity of models to market risk factors underscores the complexity of the relationship between these factors and IPO pricing in Iran's market. Advanced machine learning methods offer better capabilities for modeling such nonlinear relationships. Moreover, the statistical significance of the coefficients for industry-level variables suggests that industry averages play a crucial role in IPO pricing in Iran's capital market—a finding that is logical given Iran's unique economic conditions and the differential impact of sanctions and currency fluctuations across industries. The application of such models can contribute to greater transparency in IPO pricing, reduction of information asymmetry, and ultimately, enhancement of the efficiency of Iran's capital market.

Table 8 presents a comprehensive summary of the k-fold cross-validation results and nonparametric statistical tests used to compare the performance of various gradient boosting models in predicting the relative pricing of initial public offerings. These results confirm the robustness of previous findings through multiple validation methods and statistical testing.

Table 8. k-Fold Cross-Validation Results and Nonparametric Statistical Tests for Relative IPO Pricing
Prediction Models

Evaluati on Metric	M2B - XGBoos t	M2B - LightGB M	M2B - CatBoos t	M2B - Hybrid	V2A - XGBoos t	V2A - LightGB M	V2A - CatBoos t	V2A - Hybrid	V2S - XGBoos t	V2S - LightGB M	V2S - CatBoos t	V2S - Hybrid
Mean R^2 (10-fold)	0.815±0. 026	0.807±0. 029	0.810±0. 027	0.852±0. 019	0.782±0. 031	0.774±0. 034	0.778±0. 032	0.814±0. 022	0.742±0. 038	0.735±0. 041	0.739±0. 039	0.783±0. 026
Mean RMSE (10-fold)	0.034±0. 004	0.036±0. 005	0.035±0. 004	0.029±0. 003	0.051±0. 006	0.053±0. 007	0.052±0. 006	0.044±0. 004	0.062±0. 008	0.065±0. 009	0.063±0. 008	0.053±0. 005
Mean MAE (10-fold)	0.025±0. 003	0.027±0. 003	0.026±0. 003	0.021±0. 002	0.038±0. 004	0.041±0. 005	0.039±0. 004	0.033±0. 003	0.046±0. 005	0.048±0. 006	0.047±0. 005	0.040±0. 004

Mean	0.892±0.	0.886±0.	0.889±0.	0.918±0.	0.868±0.	0.861±0.	0.864±0.	0.897±0.	0.842±0.	0.836±0.	0.839±0.	0.874±0.
AUC	018	021	019	012	023	025	024	016	027	029	028	019
(10-fold)												
Mean R ²	0.809±0.	0.801±0.	0.804±0.	0.848±0.	0.775±0.	0.767±0.	0.771±0.	0.809±0.	0.735±0.	0.728±0.	0.732±0.	0.777±0.
(5×2-	031	035	033	022	037	040	038	026	043	046	044	030
fold)												
Mean R ²	0.821±0.	0.812±0.	0.816±0.	0.857±0.	0.789±0.	0.780±0.	0.784±0.	0.820±0.	0.749±0.	0.741±0.	0.745±0.	0.789±0.
(LOO-	022	026	024	016	028	031	029	019	034	037	035	023
CV)												
Friedma	<0.001**	<0.001**	<0.001**	-	<0.001**	<0.001**	<0.001**	-	<0.001**	<0.001**	<0.001**	-
n Test												
(p-												
value)												
Wilcoxo	<0.001**	-	-	-	<0.001**	-	-	-	<0.001**	-	-	-
n												
(Hybrid												
vs.												
XGBoos												
t)												
Wilcoxo	<0.001**	-	-	-	<0.001**	-	-	-	<0.001**	-	-	-
n												
(Hybrid												
vs.												
LightGB												
M)												
Wilcoxo	<0.001**	-	-	-	<0.001**	-	-	-	<0.001**	-	-	-
n												
(Hybrid												
vs.												
CatBoos												
t)												
Kruskal-	24.73**	-	-	-	22.18**	-	-	-	20.56**	-	-	-
Wallis												
Test (H-												
statistic)												
Friedma	2.15	3.42	2.68	1.05	2.18	3.38	2.72	1.07	2.21	3.35	2.76	1.08
n Mean												
Rank												
Model	3.02%	3.42%	3.16%	2.08%	3.75%	4.18%	3.92%	2.53%	4.89%	5.36%	5.07%	3.17%
Stability												
(% CV												
in 100												
Bootstra												
ps)												

All values are reported as mean ± standard deviation. Double asterisks (**) indicate statistical significance at $p < .01$. M2B = Market-to-Book Ratio; V2A = Enterprise Value to Assets; V2S = Enterprise Value to Sales; CV = Coefficient of Variation; LOO-CV = Leave-One-Out Cross-Validation; AUC = Area Under Curve.

In the 10-fold cross-validation results section, the hybrid model demonstrates a mean coefficient of determination (R^2) of 0.852 ± 0.019 for the market-to-book ratio, which is significantly higher than XGBoost (0.815 ± 0.026), LightGBM (0.807 ± 0.029), and CatBoost (0.810 ± 0.027). The lower standard deviation of the hybrid model (0.019), compared to the other models (ranging from 0.026 to 0.029), indicates greater model stability. The hybrid model also achieves the lowest mean RMSE and MAE values (0.029 ± 0.003 and 0.021 ± 0.002 , respectively). The average AUC for the hybrid model is 0.918 ± 0.012 , reflecting its high discriminative power. A similar pattern is observed for the enterprise value to assets and enterprise value to sales ratios, though the numerical values differ. Validation results from 5×2-fold and LOO-CV also follow a similar pattern, with R^2 values in 5×2-fold being slightly

lower and in LOO-CV slightly higher than those in 10-fold. This discrepancy may be due to different training sample sizes in each validation method. Nonetheless, across all validation techniques, the hybrid model consistently outperforms the others.

The nonparametric statistical tests also confirm that the performance differences of the hybrid model compared to other models are statistically significant. The Friedman test for all three valuation metrics yields p-values less than 0.001, indicating significant differences among models. The Wilcoxon test for pairwise comparisons between the hybrid model and each of the other models also shows p-values below 0.001, affirming the hybrid model's significantly superior performance. The Kruskal-Wallis test, with H-statistics ranging from 20.56 to 24.73 and p-values below 0.001, further confirms the significant differences among models.

The average ranks from the Friedman test indicate that the hybrid model, with ranks ranging from approximately 1.05 to 1.08, delivers the best performance, while LightGBM, with ranks between 3.35 and 3.42, performs the weakest. XGBoost and CatBoost fall in between, with ranks from 2.15 to 2.76. Finally, model stability was assessed through 100 bootstrap repetitions. The coefficient of variation (CV%) for the hybrid model across all three valuation metrics (2.08%, 2.53%, and 3.17%) is lower than that of the other models, indicating higher stability. Overall, LightGBM shows the highest variability, with CV% values ranging from 3.42% to 5.36%, while XGBoost and CatBoost lie in between.

Comparing the results across the three valuation metrics shows that predictive accuracy for the market-to-book ratio is higher than for the other two metrics. The mean R^2 in 10-fold CV for the hybrid model is 0.852 for market-to-book, while it is 0.814 for enterprise value to assets and 0.783 for enterprise value to sales. This pattern is consistent across all validation methods and models. Furthermore, model stability (based on CV%) is also higher for the market-to-book ratio than for the other two metrics. Another noteworthy point is the increasing standard deviation of the evaluation metrics from market-to-book toward enterprise value to sales. For example, the standard deviation of R^2 in 10-fold CV for the hybrid model increases from 0.019 for market-to-book to 0.022 for enterprise value to assets and 0.026 for enterprise value to sales. This suggests that predicting the enterprise value to sales ratio is more challenging and involves greater uncertainty.

The comprehensive cross-validation and nonparametric test results in Table 8 confirm that the gradient boosting hybrid model significantly outperforms other models in predicting the relative pricing of IPOs in the Iranian capital market. The statistically significant superiority of the hybrid model across all nonparametric tests (Friedman, Wilcoxon, and Kruskal-Wallis) suggests that integrating different gradient boosting algorithms can enhance predictive accuracy and model stability. Additionally, the decline in predictive accuracy from market-to-book to enterprise value to sales may indicate the increasing complexity of factors affecting sales-related ratios in Iran's capital market. Various elements such as exchange rate fluctuations, sanctions, inflation, and regulatory changes can impact company sales and complicate prediction. Nevertheless, even for this more challenging metric, the hybrid model achieves a respectable R^2 of 0.783 in 10-fold CV. These findings justify the application of advanced machine learning methods in IPO pricing prediction in Iran's market, which faces its own set of challenges, and can assist investors, company managers, and regulatory bodies in making better decisions. Moreover, the consistency of results across different validation methods (10-fold, 5×2-fold, and LOO-CV) and bootstrap repetitions provides greater confidence in the generalizability of these models in real market conditions.

4. Discussion and Conclusion

This study was conducted to investigate the factors influencing relative pricing of stocks in initial public offerings (IPOs). The present research adapted the baseline model of Geurtsma and Lu (2023) and utilized data from 42 listed companies on the Tehran Stock Exchange and 121 firms listed on the Iran Fara Bourse, covering the period from 2013 to 2023. The three main dependent variables examined were: market-to-book ratio, enterprise value to assets ratio, and enterprise value to sales ratio. The research employed gradient boosting models and a hybrid machine learning model based on gradient boosting algorithms. The results demonstrate that the hybrid gradient boosting model exhibits remarkable capability in predicting relative pricing variables. By leveraging the strengths of the three primary gradient boosting algorithms, this model enables the identification of complex patterns within the data. Cross-validation results revealed that the hybrid model had lower prediction error compared to individual models, indicating the high efficiency of the hybrid approach in addressing the complexities of Iran's capital market.

The variable importance analysis showed that internal financial factors such as the weighted average cost of capital, return on assets, and financial leverage significantly impact the relative pricing of IPOs. This finding confirms the role of financial structure and corporate performance in determining the relative value of companies at the time of offering. Additionally, operational variables such as operating profit margin and earnings per share were also found to influence relative pricing, highlighting profitability as a key factor in company valuation. Among the control variables, firm size, firm age, and cash flow to asset ratio had the greatest impact on the dependent variables, emphasizing the importance of firm-specific characteristics in determining IPO relative pricing.

Compared to previous research, our findings are aligned with the study by Geurtsma and Lu (2023), which demonstrated the potential of machine learning methods in improving the accuracy of stock valuation forecasts. However, the main innovation of the present study lies in the use of a hybrid gradient boosting approach for relative IPO pricing in the Iranian stock market, which outperforms more traditional methods used in earlier studies [25, 30]. The results of this research also align with Nabi et al. (2020), who found that gradient boosting, when coupled with feature engineering, performs better than prior methods. Indeed, the hybrid approach used in this study, similar to theirs, benefits from feature engineering to enhance model performance. Compared to the study by Roy et al. (2020), which compared various machine learning methods for stock price prediction [22], the present research also emphasizes the importance of selecting appropriate algorithms. However, our approach differs in that it integrates several gradient boosting algorithms, potentially leading to more accurate results.

The findings regarding the significance of macroeconomic, industry-specific, and firm-level variables in the relative pricing of IPOs are also consistent with Nikpey Pesyan et al. (2023), who showed that macroeconomic factors causally influence the number of IPOs [12]. In fact, the comprehensive approach of this study in considering diverse variables is one of its strengths relative to earlier research. The results also reflect the unique challenges of Iran's capital market. As shown by [15], public offerings can negatively impact a firm's financial performance. Our proposed model may help mitigate these negative effects by providing more accurate forecasts of relative pricing. Unlike developed markets studied by [26, 27], Iran's capital market faces greater challenges, such as extreme volatility and sensitivity to non-economic factors. Our hybrid approach, by combining the capabilities of different gradient boosting algorithms, is better equipped to model these complexities.

The findings also align with those of Abbasian et al. (2023), who demonstrated that the performance of a gradient boosting decision tree model improves when relevant variables are added. In our study as well, incorporating diverse variables into the hybrid model led to performance enhancement. In conclusion, this research, by presenting

a hybrid model based on gradient boosting algorithms for predicting the relative pricing of IPOs in the Iranian stock market, has taken a significant step toward improving prediction accuracy and reducing error in this area. The application of the hybrid gradient boosting method for the first time in Iran's capital market for IPO pricing is the main innovation of this study, which can aid in better modeling of the complex dynamics of Iran's market.

The development of a comprehensive framework for selecting variables affecting IPO relative pricing, the provision of an interpretable method for assessing variable importance, and the design of a multilayer cross-validation mechanism to evaluate model performance are among the other contributions of this research that can enhance the efficiency of Iran's capital market. The results of this study may be useful for investors, underwriters, and regulatory institutions in making better decisions and increasing market transparency. However, it should be noted that despite the high accuracy of the proposed model, IPO pricing is influenced by numerous factors, including macroeconomic conditions, monetary and fiscal policies, and market psychological factors, not all of which may be fully captured by the model. Therefore, the use of this model should be accompanied by professional judgment and consideration of the specific context of each offering.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

- [1] C. M. Aldana and F. Trigos, "A transdisciplinary engineering framework for analysing initial public offerings' financial behaviour," *International Journal of Agile Systems and Management*, vol. 18, no. 1, pp. 112-132, 2025, doi: 10.1504/IJASM.2025.144166.
- [2] A. Habib and M. M. Hasan, "Corporate life cycle research in accounting, finance and corporate governance: A survey, and directions for future research," *International Review of Financial Analysis*, vol. 61, no. 1, pp. 188-201, 2019, doi: 10.1016/j.irfa.2018.12.004.
- [3] G. Airlangga and A. Liu, "A Hybrid Gradient Boosting and Neural Network Model for Predicting Urban Happiness: Integrating Ensemble Learning with Deep Representation for Enhanced Accuracy," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, pp. 1-23, 2025, doi: 10.3390/make7010004.
- [4] A. Rasheed, M. Khalid Sohail, S. U. Din, and M. Ijaz, "How Do Investment Banks Price Initial Public Offerings? An Empirical Analysis of Emerging Market," *International Journal of Financial Studies*, vol. 6, no. 3, pp. 77-101, 2018, doi: 10.3390/ijfs6030077.
- [5] P. d. Andrés, D. Arroyo, R. Correia, and A. Rezola, "Challenges of the market for initial coin offerings," *International Review of Financial Analysis*, vol. 79, no. 1, p. 101966, 2022, doi: 10.1016/j.irfa.2021.101966.

- [6] Y. K. Dwivedi *et al.*, "Setting the future of digital and social media marketing research: Perspectives and research propositions," *International Journal of Information Management*, vol. 59, no. 1, p. 102168, 2021, doi: 10.1016/j.ijinfomgt.2020.102168.
- [7] T. Chemmanur, "The Pricing of Initial Public Offerings: A Dynamic Model With Information Production," *The Journal of Finance*, vol. 48, no. 1, pp. 285-304, 1993, doi: 10.1111/j.1540-6261.1993.tb04710.x.
- [8] N. Crain, P. Robert, and S. Raji, "Uncertainty prospectus content and the pricing of initial public offerings," *Journal of Empirical Finance*, vol. 64, no. 2, pp. 1-23, 2021, doi: 10.1016/j.jempfin.2021.08.007.
- [9] A. A. Daryaei, P. Azizi, and Y. Fattahi, "Conservatism and Initial Public Offerings (IPOs) Underpricing: An Audit Quality Perspective," *Iranian Journal of Finance*, vol. 6, no. 4, pp. 125-159, 2022, doi: 10.30699/ijf.2022.284931.1230.
- [10] W. L. Megginson and J. M. Netter, "From state to market: A survey of empirical studies on privatization," ed, 1999.
- [11] D. Dalton, T. Certo, and C. Daily, "Initial Public Offerings as a Web of Conflicts of Interest: An Empirical Assessment," *Business Ethics Quarterly*, vol. 13, no. 3, pp. 289-314, 2023, doi: 10.2307/3857783.
- [12] V. Nikpey Pesyan, A. Reza Zadeh, H. Ahmadi Nezhad, and S. Ahmad Vand, "Investigating the Causal Relationship Between Stocks' Initial Public Offerings and Macroeconomic Variables," *Journal of Asset Management and Financing*, vol. 11, no. 2, pp. 35-52, 2023, doi: 10.22108/amf.2023.136359.1779.
- [13] S. Füllbrunn, T. Neugebauer, and A. Nicklisch, "Underpricing of initial public offerings in experimental asset markets," *Experimental Economics*, vol. 23, no. 4, pp. 1002-1029, 2020, doi: 10.1007/s10683-019-09683-4.
- [14] L. Bateni and F. Asghari, "Study of Factors Affecting the Initial Public Offering (IPO) Price of the Shares on the Tehran Stock Exchange," *Research in World Economy*, vol. 5, no. 2, pp. 68-79, 2014, doi: 10.5430/rwe.v5n2p68.
- [15] M. H. Haratameh and B. Ebrahimi, "Investigating the impact of initial public offerings (IPOs) on companies' financial performance," *New Research in Performance Evaluation*, vol. 2, no. 4, pp. 240-252, 2023, doi: 10.22105/mrpe.2024.451424.1096.
- [16] T. Kazemi and P. Piri, "Predicting fraud schemes in financial reporting using a multi-class machine learning approach," *Journal of Experimental Accounting Research*, vol. 12, no. 4, pp. 255-280, 2022, doi: 10.22051/jera.2022.41290.3040.
- [17] A. Abbasian, K. Shahraki, S. Fallahpour, and A. Namaki, "A novel approach to financial distress prediction using financial network-based information and a combined gradient boosting decision tree method," *Asset Management and Financing*, vol. 11, no. 3, pp. 113-140, 2023, doi: 10.22108/amf.2023.138909.1818.
- [18] B. Baba and G. Sevil, "Predicting IPO initial returns using random forest," *Borsa Istanbul Review*, vol. 20, no. 1, pp. 13-23, 2020, doi: 10.1016/j.bir.2019.08.001.
- [19] S. M. Khatami, Z. Gholamreza, M. Leyalestani, and M. Minoei, "Investigating the dependency structure of the Iranian stock market and MENA region countries," *Financial Economics*, vol. 16, no. 61, pp. 273-310, 2022, doi: 10.30495/fed.2023.698852.
- [20] M. Mehrabadi, A. Najafizadeh, M. Zanjirdar, and P. Ashtiani, "Investigating the impact of market structure and asymmetric information on the performance of companies active in the Tehran Stock Exchange in a dynamic model," *Financial Economics*, vol. 16, no. 60, pp. 93-120, 2022, doi: 10.30495/fed.2022.697606.
- [21] R. Nabi, S. Saeed, and H. Harron, "A Novel Approach for Stock Price Prediction Using Gradient Boosting Machine with Feature Engineering (GBM-wFE)," *KJAR*, vol. 5, no. 1, pp. 28-48, 2020, doi: 10.24017/science.2020.1.3.
- [22] S. S. Roy, R. Chopra, K. C. Lee, C. Spampinato, and B. Mohammadi-ivatlood, "Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 33, no. 1, pp. 62-71, 2020, doi: 10.1504/IJAHUC.2020.104715.
- [23] G. Mitrentsis and H. Lens, "An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting," *Applied Energy*, vol. 309, no. 1, p. 118473, 2021, doi: 10.1016/j.apenergy.2021.118473.
- [24] M. Saeidi Aghdam, A. Sadeghi, A. Bahiraei, and S. Haji Asghari, "Presenting a stock price prediction model using deep learning algorithms and its application in pricing Islamic bank stocks," *Journal of Islamic Economics and Banking*, vol. 11, no. 41, pp. 117-134, 2022.
- [25] P. Geertsema and H. Lu, "Relative Valuation with Machine Learning," *Journal of Accounting Research*, vol. 61, no. 1, pp. 329-376, 2022, doi: 10.1111/1475-679X.12464.
- [26] K. Nakagawa and K. Yoshida, "Time-series gradient boosting tree for stock price prediction," *International Journal of Data Mining, Modelling and Management*, vol. 14, no. 2, pp. 110-125, 2022, doi: 10.1504/IJDDMM.2022.123357.
- [27] S. Li, "Estimating Stock Market Prices with Histogrambased Gradient Boosting Regressor: A Case Study on Alphabet Inc," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, pp. 532-544, 2023, doi: 10.14569/IJACSA.2024.0150553.
- [28] V. Gupta and E. Kumar, "H3O-LGBM: hybrid Harris hawk optimization based light gradient boosting machine model for real-time trading," *Springer Nature*, vol. 56, no. 1, pp. 8697-8720, 2023, doi: 10.1007/s10462-022-10323-0.
- [29] Z. Huma and A. Nishat, "Optimizing Stock Price Prediction with LightGBM and Engineered Features," *Pioneer Research Journal of Computing Science*, vol. 1, no. 1, pp. 59-67, 2024.

- [30] F. Ghallabi, B. Souissi, A. M. Du, and S. Ali, "ESG stock markets and clean energy prices prediction: Insights from advanced machine learning," *International Review of Financial Analysis*, vol. 97, no. 1, p. 103889, 2025, doi: 10.1016/j.irfa.2024.103889.