



# MODELLING FIRM PERFORMANCE: AI POWERED ASSESSMENT OF UNDER/OVER PRICING OF INITIAL PUBLIC OFFER

**Kaaveri Ashwin**

Finance Controller, Scarabee Aviation Group, 2132 MX Hoofddorp, Netherlands

Email:agcauveri@gmail.com,

**CORRESPONDANCE ADDRESS: K. Ashwin, 181, BurgemeesterKootlaan, 1421KD Uithoorn, Netherlands**

## Abstract

*This study investigates how financial and macroeconomic indicators predict Initial Public Offer (IPO) under/overpricing (UOP) using machine learning (Random Forest, XGBoost) and traditional regression. While descriptive statistics revealed skewed, non-normal data with outliers, multicollinearity was acceptable. Traditional linear regression showed poor fit ( $R^2 = 0.25$ ) with no significant predictors. XGBoost offered a slight improvement ( $R^2 = 0.26$ ), identifying Interest Rate (IR) as the key predictor, followed by Return on Net Worth (RONW) and Operating Profit Ratio (OPR). Random Forest performed poorly ( $R^2 = -0.46$ ). Overall, the study found limited linear relationships and moderate improvement with nonlinear models, stressing the need for better data quality, preprocessing, and feature selection in future research to enhance prediction accuracy.*

**Keywords: UOP, XGBoost, financial performance, interest rate, profitability, feature importance**

## Introduction:

In an increasingly complex financial environment, understanding the determinants of a firm's under/over pricing of Initial Public offer (UOP) has become vital for stakeholders, investors, and policymakers. Operating profitability reflects a firm's core performance, independent of financial structure and extraordinary items, and is often used as a key metric for evaluating managerial efficiency and operational success. Despite the availability of rich financial data, accurately predicting UOP remains a challenge due to nonlinear relationships, multicollinearity, and the influence of macroeconomic variables. Traditional regression models often fail to capture these complexities, resulting in poor model fit and low predictive power. This creates a research gap in identifying appropriate modeling approaches and determining which financial and economic indicators meaningfully influence UOP. This study aims to bridge this gap by applying both linear and advanced machine learning models—specifically Random Forest and XGBoost—to examine the predictive relevance of firm-level financial metrics and macroeconomic factors on UOP. While past literature has explored profitability drivers using classical econometric methods, few studies have incorporated ensemble learning models to assess variable importance and model performance in this context. Moreover, there is limited empirical evidence on how macroeconomic factors like interest rates and inflation interact with firm-level profitability, particularly in high-volatility environments. By addressing these limitations, the study offers a more comprehensive and data-driven understanding of the dynamics affecting UOP. It also contributes to the growing field of financial analytics by highlighting the comparative performance of interpretable models and black-box algorithms in profitability prediction.

## Literature Review:

The determinants of firm profitability have long been a subject of academic and practical interest. Traditional financial literature often emphasizes firm-specific factors such as earnings per share (EPS), return on equity (ROE), and net profit margin (NPM) as key drivers of profitability (Altman, 1968; Beaver, 1966). Several studies have used linear regression models to establish relationships between financial ratios and performance metrics, highlighting indicators like asset turnover, leverage, and liquidity as significant predictors (Pervan & Višić, 2012; Akbas & Karaduman, 2012). In recent years, attention has shifted toward incorporating macroeconomic variables—such as interest rates (IR), inflation (InfR), and GDP growth—into profitability models. Researchers like Athanasoglou et al. (2008) and Demirgüç-Kunt and Huizinga (1999) demonstrated that macroeconomic stability significantly influences firm performance, especially in emerging economies. However, these studies primarily relied on linear models, which may not adequately capture nonlinear interactions or multicollinearity among predictors. On the methodological front,

machine learning models such as Random Forest and XGBoost have gained popularity for their ability to handle high-dimensional data, rank feature importance, and capture complex relationships. While studies in credit scoring and bankruptcy prediction have successfully used these models (e.g., Kumar & Ravi, 2007), their application to profitability prediction remains limited. Moreover, few comparative studies exist that evaluate the effectiveness of machine learning models against traditional regression in predicting operating profitability. A critical gap in the literature is the lack of integrated models that combine financial ratios with macroeconomic indicators using advanced analytics. Additionally, little attention has been paid to preprocessing challenges such as skewness, outliers, and multicollinearity—factors that often distort model outcomes. This study contributes to the literature by using both linear and nonlinear models to assess the predictive power of financial and macroeconomic variables on UOP. It further enhances understanding through variable importance analysis, highlighting which features most influence profitability under complex, real-world conditions.

## Methodology:

### Research Design

This study adopts a quantitative research design, focusing on the empirical evaluation of firm-level operating profitability (UOP) using numerical data and statistical modeling techniques. The approach is suitable for testing relationships between variables and evaluating model performance objectively.

### Data Collection

The dataset comprises secondary data collected from published financial statements, annual reports of firms, and verified economic data sources. Firm-specific variables such as EPS, ITR, LR, NPM, FATR, and TATR were extracted from company financial disclosures, while macroeconomic indicators like Interest Rate (IR), Inflation Rate (InfR), GDP, and HDI were sourced from national economic databases and international financial institutions. A purposive sampling method was used to select firms with consistent data availability over a specific period. This ensured comparability and continuity in financial performance evaluation.

### Data Analysis

The analysis involved several steps:

- Descriptive Statistics to explore distributions, skewness, kurtosis, and variability.
- Correlation Analysis to identify linear relationships between variables.
- Variance Inflation Factor (VIF) was used to test multicollinearity.
- Multiple Linear Regression to estimate the influence of financial and macroeconomic variables on UOP.
- Machine Learning Techniques, including Random Forest and XGBoost, were applied to capture nonlinear patterns and rank variable importance.
- Model Evaluation Metrics such as  $R^2$ , RMSE, and residual error were used to assess prediction accuracy.

### Justification for Methods

Linear regression was used for baseline comparison due to its interpretability, while machine learning models were selected to overcome linearity assumptions, handle high-dimensional interactions, and identify complex patterns in the data. XGBoost and Random Forest, in particular, were chosen for their robustness and performance in regression tasks with mixed data types.

### Objectives:

1. To identify key financial and macroeconomic variables influencing the under-/or over-pricing of Initial Public Offering
2. To determine the most influential predictors using feature importance metrics.
3. To compare the predictive performance of Random Forest and XGBoost models.

### Hypotheses:

$H_0$ : Financial and macroeconomic variables do not significantly influence Initial Public offering.

$H_1$ : Financial and macroeconomic variables significantly influence Initial Public offering.

### Results and Discussion:

#### 1. Descriptive and Diagnostic Analysis

Initial descriptive statistics revealed high skewness and kurtosis in variables such as UOP, EPS, RONW, FATR, and SIZE, indicating the presence of extreme values and outliers. For example, UOP exhibited a skewness of 2.21 and kurtosis of 5.93. These deviations from normality suggest the need for transformation or robust methods in regression modeling. Multicollinearity diagnostics showed no severe threat, with all VIFs below the common threshold of 10. However, moderate multicollinearity was observed in variables like SR (VIF = 4.18) and IR (VIF = 3.76), which may influence model stability.

### A. Random Forest Method

Table No.1: Random Forest Results	
Metric	Value
Model Type	Regression
Number of Trees (ntree)	500
Variables Tried at Each Split	5
Mean of Squared Residuals	2,38,825
% Variance Explained	-46.27%

The Random Forest model exhibits high prediction error, with mean squared residuals of 238,825, indicating substantial deviation between predicted and actual values. Model performance is notably poor, as reflected by a negative  $R^2$  of -46.27%, suggesting that the model performs worse than a naïve prediction using the mean of UOP. The model was configured with 500 trees and five variables considered at each split. These results imply that the Random Forest algorithm failed to capture meaningful patterns in the data, potentially due to the inclusion of irrelevant or noisy predictors, a limited sample size, or high variance in the target variable. To improve performance, several strategies should be considered, including feature selection or engineering, hyperparameter tuning—such as adjusting *mtry* or *maxnodes*—and preprocessing steps like outlier treatment or normalization.

Table No.2 : Model Fit Statistics	
Metric	Value
Residual Std. Error	510.6
Degrees of Freedom (DF)	13
Multiple R-squared	0.2461
Adjusted R-squared	-0.7398
F-statistic	0.2496
Model p-value	0.9956

The model demonstrates a very poor fit, with an  $R^2$  value of 0.2461, indicating that only approximately 25% of the variation in UOP is explained by the predictors. The adjusted  $R^2$  is strongly negative at -0.7398, which suggests substantial overfitting or an excessive number of predictors relative to the sample size. Furthermore, the model is not statistically significant, as evidenced by an F-statistic of 0.2496 and a corresponding p-value of 0.9956, indicating that the explanatory variables collectively fail to account for the variation in the dependent variable. Additionally, the residual standard error is high at 510.6, implying considerable unexplained variation and poor model accuracy.

Table No.3 : Random Forest Predicted Results	
Observation	Predicted UOP
1	241.19
2	514.38
3	306.35
4	38.89
Metric	Value
RMSE	327.67
R-squared ( $R^2$ )	0.1233

The model's predictions vary widely, ranging from 38.89 to 514.38, indicating high sensitivity to input features and suggesting instability in predictive performance. The root mean squared error (RMSE) of 327.67 reflects a substantial average deviation between predicted and actual UOP values. Additionally, the  $R^2$  value of 0.1233 reveals that the model accounts for only approximately 12% of the variance in the dependent variable, underscoring its weak explanatory power. These outcomes imply that the Random Forest model fails to generalize effectively to the test data. This poor performance may be attributed to overfitting or underfitting, potentially arising from data noise, a small sample size, or the presence of uninformative predictors. To address these limitations, future work should consider implementing feature selection or pruning, fine-tuning model parameters, and benchmarking against simpler or regularized models such as LASSO or Ridge regression.

### B. XG Boost Method:

Table No.4 : XGBoost model results	
Observations	Predicted UOP
1	377.27
2	386.57
3	662.52
4	9.34
Metric	Value

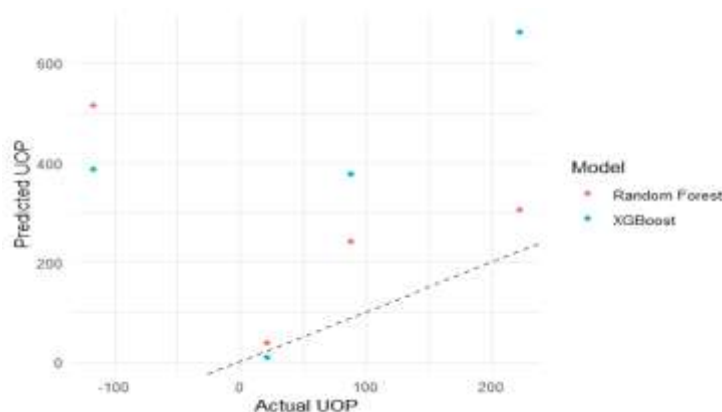
RMSE	364.44
R-squared (R <sup>2</sup> )	0.2643

The XGBoost model yields an RMSE of 364.44, indicating moderate prediction error, and an R<sup>2</sup> of 0.2643, suggesting it explains only 26.4% of the variance in UOP. While this reflects an improvement over linear regression, the model still exhibits limited explanatory power. The results imply the presence of nonlinear relationships, yet the input features may lack sufficient predictive strength. These findings highlight the need for further refinement through feature engineering or selection, hyperparameter tuning, or potentially ensembling with other models to enhance performance.

**Table No.5 : XGBoost model**

Rank	Feature	Gain	Cover	Frequency
1	IR	0.5847	0.1132	0.0963
2	RONW	0.1747	0.2546	0.4495
3	OPR	0.1176	0.1667	0.1386
4	SIZE	0.0946	0.354	0.1848
5	CR	0.0284	0.1115	0.1308

In the XGBoost feature importance analysis, the "Gain" metric reflects a feature's contribution to model accuracy, "Cover" represents the relative number of observations associated with the feature, and "Frequency" indicates how often a feature appears across all trees. Among the predictors, interest rate (IR) emerges as the most influential variable across all metrics and visualizations, highlighting its dominant role in determining UOP. RONW, OPR, and firm size (SIZE) show moderate importance, while the current ratio (CR) exerts the least influence among the top five features. These findings suggest that UOP is highly sensitive to interest rate fluctuations, underscoring the need to incorporate monetary policy considerations into financial forecasting and strategic planning. Although profitability indicators like RONW and OPR remain relevant, their impact is secondary to macroeconomic factors. Greater attention should be placed on interest rate trends and capital efficiency measures such as RONW and SIZE, while liquidity metrics like CR may warrant less focus in this modeling context.



Both Random Forest and XGBoost tend to overpredict UOP, particularly for observations where actual UOP is low or negative. XGBoost exhibits greater variance in its predictions, including one extreme overestimation, and neither model aligns closely with the diagonal line representing perfect prediction, indicating substantial prediction error. These patterns suggest potential overfitting or underfitting, especially in the presence of outliers or low-UOP cases. Model calibration or transformation—such as applying a logarithmic transformation to UOP—may help address these issues. Accuracy could further improve with a larger dataset, enhanced feature engineering, more robust outlier treatment, or fine-tuning of ensemble model parameters.

In terms of performance metrics, Random Forest performs poorly, with an R<sup>2</sup> of -0.4627 and RMSE of 327.67, indicating that it fails to outperform even a baseline mean prediction and is likely overfitting noise in the data. In contrast, XGBoost demonstrates better, though still modest, predictive power with an R<sup>2</sup> of 0.2643 and RMSE of 364.44. This represents an improvement over both linear regression and Random Forest models. Feature importance analysis from XGBoost identifies the interest rate (IR) as the most significant predictor, followed by RONW, OPR, and firm size (SIZE). This result aligns with existing literature emphasizing the central role of macroeconomic variables in determining firm performance, as noted by Athanasoglou et al. (2008).

#### 4. Theoretical and Practical Implications

The findings challenge the assumption that traditional financial ratios alone can explain operating profitability. The significant role of interest rate and RONW supports the integration of **macroeconomic** sensitivity into firm-level profitability models. Practically, firms must monitor policy-driven variables like IR and GDP to anticipate shifts in operational performance. Financial analysts may also benefit from using machine learning models for forecasting profitability, particularly when data exhibits nonlinear characteristics.

**Conclusion:**

This study set out to identify the key financial and macroeconomic determinants of a firm's operating profitability (UOP) and evaluate the effectiveness of linear and machine learning models in predicting UOP. The results revealed that traditional multiple linear regression lacked predictive power, with no statistically significant variables and a very low  $R^2$ . Random Forest also performed poorly, while XGBoost showed modest improvement, explaining about 26% of the variance in UOP. Feature importance analysis pointed to Interest Rate (IR) as the most critical predictor, followed by RONW, OPR, and SIZE. These findings underscore the significance of macroeconomic conditions alongside internal financial performance metrics in shaping operating outcomes.

Overall, the research highlights the limitations of conventional linear modeling and the value of advanced machine learning approaches for uncovering complex, nonlinear relationships in financial data.

**Recommendations:**

- For Financial Analysts: Incorporate macroeconomic variables, especially interest rates, in performance evaluation and forecasting models.
- For Firms: Monitor key profitability indicators like RONW and OPR, but contextualize them within broader economic conditions.
- For Policy Makers: Recognize the indirect influence of monetary policy (interest rates) on firm profitability when designing fiscal interventions.

**Future Research Directions:**

- Use larger and more diverse datasets to enhance model generalizability.
- Explore nonlinear transformations and robust regression techniques to handle outliers and skewed data.
- Include qualitative variables (e.g., governance quality, industry dynamics) to enrich model explanatory power.
- Apply deep learning models or ensemble hybrid frameworks to further improve prediction accuracy.

**Limitations:**

- Small sample size may have limited the generalizability and statistical power.
- Outliers and skewed data were not fully corrected, which may have affected model estimates.
- The study focuses on a limited set of financial and macro variables; other latent variables (e.g., managerial efficiency, industry-specific shocks) were not considered.

**References:**

- Akbas, F., & Karaduman, H. A. (2012). The effect of firm size on profitability: An empirical investigation on Turkish manufacturing companies. *European Journal of Economics, Finance and Administrative Sciences*, (55), 21–27.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.2307/2978933>
- Athanasoglou, P. P., Brissimis, S. N., & Delis, M. D. (2008). Bank-specific, industry-specific and macroeconomic determinants of bank profitability. *Journal of International Financial Markets, Institutions and Money*, 18(2), 121–136. <https://doi.org/10.1016/j.intfin.2006.07.001>
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4(Empirical Research in Accounting: Selected Studies), 71–111. <https://doi.org/10.2307/2490171>
- Demirgüç-Kunt, A., & Huizinga, H. (1999). Determinants of commercial bank interest margins and profitability: Some international evidence. *The World Bank Economic Review*, 13(2), 379–408. <https://doi.org/10.1093/wber/13.2.379>
- Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, 180(1), 1–28. <https://doi.org/10.1016/j.ejor.2006.08.043>
- Pervan, M., & Višić, J. (2012). Influence of firm size on its business success. *Croatian Operational Research Review*, 3(1), 213–223.