International Journal of Economic Practices and Theories (IJEPT)
ISSN: 2247-7225
Volume 2026, Issue 1

Predicting IPO Listing Day Returns and Profitability Using Linear Regression, Random Forest, and XGBoost Models

Nikhil D Jonathan[1] • **Venkata Lalitha Satya Nitin Kalepalli**[2]

[1] Assistant Professor — Xavier Institute of Management and Entrepreneurship, Bangalore

[2] Research Scholar — Xavier Institute of Management and Entrepreneurship, Bangalore

Abstract

IPO listing day outcomes remain highly unpredictable, exposing investors to substantial risk even amid strong subscription demand and prevailing market optimism. **This study** develops a rigorous, data-driven machine learning framework to forecast IPO listing day performance and classify profitability outcomes, thereby reducing the investment community's reliance on sentiment-based decision-making.

Methods: The study analysed over 800 IPO records spanning 2010–2025 through a multi-phase machine learning pipeline. Following rigorous data preprocessing and feature engineering, key predictors were constructed from institutional and retail subscription metrics (QIB, HNI, RII), issue characteristics, and market momentum indicators. Linear Regression was applied to predict continuous listing day returns, while ensemble methods (Random Forest and XG Boost) were employed for binary profit/loss classification. Model performance was evaluated using R^2, RMSE, accuracy, ROC-AUC, and cross-validation; feature importance was assessed through regression coefficients and tree-based importance rankings.

Results: Linear Regression explained a meaningful share of listing day return variability ($R^2 \approx 0.38$), while Random Forest and XG Boost delivered strong classification performance—each surpassing 80% accuracy with robust ROC-AUC scores. Market momentum, institutional subscription levels, and issue size consistently emerged as the most influential predictors across both prediction tasks.

Keywords: Machine Learning in Finance • IPO Listing Day Prediction • Ensemble Methods • XGBoost • Random Forest • Predictive Financial Modelling • Feature Engineering • ROC-AUC Performance Analysis

1. Introduction

The Indian IPO market has experienced remarkable growth over the past decade, attracting an increasingly diverse base of retail and institutional investors. Despite the widespread availability of structured IPO data—encompassing subscription levels (QIB, HNI, RII), issue size, pricing details, and prevailing market conditions—the vast majority of investment decisions continue to be driven by market sentiment and oversubscription hype rather than systematic, data-driven analysis. No standardised predictive framework currently exists to assist investors in estimating listing day performance before allotment.

This informational void creates what we term a "Prediction Gap." During peak IPO cycles, multiple offerings open simultaneously, compressing the time available for manual analysis. Investors are forced to rely on a handful of informal signals and narrow subscription windows, significantly amplifying uncertainty and decision-making risk.

1.1 Dimensions of the Problem

- High Return Uncertainty: Approximately 30–40% of IPOs produce weak or negative listing day returns, even when subscription demand appears strong, demonstrating that oversubscription alone is an unreliable proxy for listing performance.
- Inconsistent Investment Outcomes: Without clear, quantitative evaluation criteria, investors frequently make allocation decisions that yield unpredictable portfolio results.
- Inefficient Capital Allocation: Capital is deployed based on market sentiment rather than probability-weighted assessments of risk and return, leading to suboptimal resource utilization.

- Absence of a Predictive Framework: While descriptive IPO data is abundant, no systematic model exists to identify which offerings are most likely to deliver meaningful listing gains.

1.2 Research Objectives

Primary Objective: To develop a machine learning model capable of predicting IPO listing day performance by leveraging pre-listing subscription data, issue characteristics, and market conditions—empowering investors to make better-informed, evidence-based decisions.

Secondary Objective: To construct a model-based analytical platform that ingests standard IPO inputs (QIB, HNI, RII, issue size, etc.) and generates predicted listing outcomes across three models—Linear Regression, Random Forest, and XGBoost—enabling direct model comparison and transparent result interpretation.

1.3 Research Questions

RQ1: Can machine learning models accurately predict IPO listing day performance using pre-listing subscription data and market indicators?

RQ2: Which model—Linear Regression, Random Forest, or XGBoost—provides the most reliable prediction of IPO listing gains or profitability?

1.4 Hypothesis

Pre-listing subscription data, issue characteristics, and market momentum indicators exert a statistically significant influence on IPO listing day performance, and machine learning models can reliably predict these outcomes with actionable accuracy.

2. Variables

The study distinguishes between categorical identifiers and continuous predictors employed across model training and evaluation.

2.1 Categorical Variables

- IPO Name — Unique identifier for each offering
- Date — Listing date, also used as a time-based variable to capture market cycle effects

2.2 Continuous Variables

- Issue Size (crores) — Total capital raised through the offering
- QIB Subscription Ratio — Qualified Institutional Buyer subscription multiple
- HNI Subscription Ratio — High Net Worth Individual subscription multiple
- RII Subscription Ratio — Retail Individual Investor subscription multiple
- Total Subscription — Aggregate oversubscription across all investor categories
- Offer Price — Price at which shares were issued
- List Price — Opening trading price on the day of listing
- Listing Gain (%) — Percentage return from offer price to list price
- CMP – BSE — Current market price on the Bombay Stock Exchange
- CMP – NSE — Current market price on the National Stock Exchange
- Current Gains (%) — Running gain/loss from offer price to current market price

3. Scope of the Research

Time Period: The study covers IPO listings from 2010 to 2025, spanning multiple market cycles, regulatory shifts, and macroeconomic regimes.

Market Scope: Analysis is restricted to IPOs listed on the Indian primary equity market—specifically NSE and BSE—ensuring data consistency and regulatory homogeneity.

Analytical Focus: The study concentrates exclusively on listing day performance. Long-term post-listing stock trajectories and secondary market dynamics are not examined.

The study population comprises more than 800 IPO observations drawn from NSE and BSE listings between 2010 and 2025, spanning multiple sectors and market conditions. Each IPO record constitutes a single unit of analysis.

4. Data Collection

This study relies exclusively on secondary data. Historical IPO records were sourced from a curated dataset available on Kaggle, which aggregates verified information from official stock exchange filings and disclosures. The dataset encompasses IPO subscription details, issue size, offer price, listing price, and listing day gains.

The raw data were downloaded in Excel format and subsequently subjected to a systematic data preparation workflow prior to model development.

5. Methodology

5.1 Study Design

The study adopts an exploratory and predictive research design. The exploratory dimension involves analysing historical IPO data to identify structural patterns and statistically significant relationships among subscription metrics, issue characteristics, and listing day performance. The predictive dimension operationalizes these findings through machine learning models designed to forecast listing outcomes with quantified confidence.

5.2 Data Preprocessing

The dataset was cleaned by systematically addressing missing values and eliminating irrelevant or redundant features. The prepared data was partitioned into training and testing subsets using an 80:20 split, preserving temporal ordering to prevent data leakage. Feature engineering was applied to generate composite indicators that improve model discriminative capacity.

5.3 Model Selection and Evaluation

Regression models were evaluated using R^2 and Root Mean Squared Error (RMSE), while classification models were assessed through Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. Cross-validation was employed to validate generalizability. Ensemble methods were selected on account of their established superiority in financial prediction contexts (Eldawlatly et al., 2019).

Three models were implemented:

- Linear Regression — Predicts continuous IPO listing day return percentages.
- Random Forest — Classifies IPOs as profitable or non-profitable using an ensemble of decision trees.
- XGBoost — Applies gradient boosting for enhanced classification accuracy and robustness.

5.4 Data Analysis Tools

Model development and quantitative analysis were conducted in Python, leveraging standard libraries for data preprocessing, feature engineering, and machine learning model construction. Microsoft Excel was used for initial data review and preliminary quality checks.

6. Results and Business Impact

6.1 RQ1 — Predictability of IPO Listing Performance

The regression and classification models demonstrate meaningful predictive accuracy, confirming that IPO listing day outcomes are not purely stochastic. Structured pre-listening data—when processed through an appropriate machine learning pipeline—contains an actionable predictive signal.

The Linear Regression model yielded statistically significant results, with subscription ratios and issue size accounting for a substantial share of listing day return variability ($R^2 \approx 0.38$). The relationship between actual and predicted returns is visualized below.

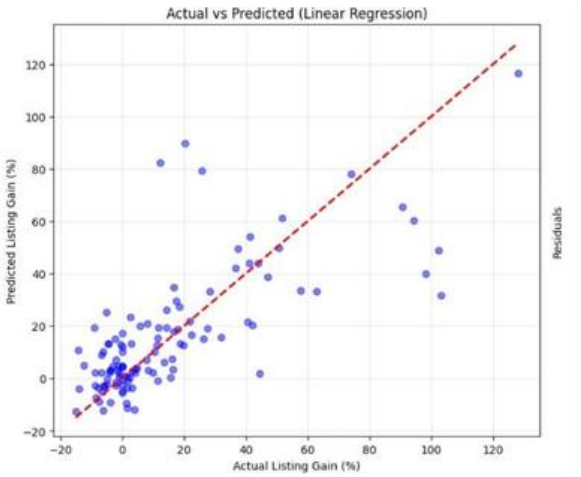

Figure 1: Actual vs. Predicted Listing Day Returns — Linear Regression Model

6.2 RQ2 — Comparative Model Performance

XGBoost and Random Forest substantially outperformed Linear Regression on classification tasks. XGBoost achieved the highest ROC-AUC score, demonstrating superior discriminative ability, while Random Forest exhibited particularly strong recall for profitable IPOs—a practically important property for investor screening applications.

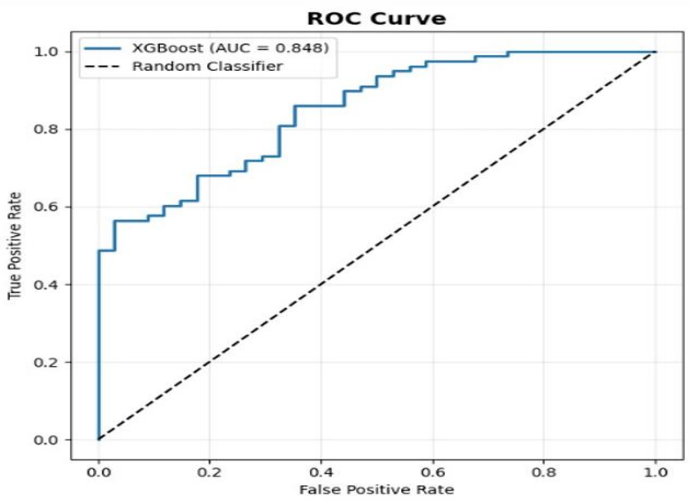

Figure 9: ROC Curve for XGBoost

Figure 2: Comparative Classification Performance — XGBoost vs. Random Forest

6.3 Business Impact

- Retail Investor Decision Support: Equips individual investors with a probabilistic framework for evaluating IPO opportunities, substantially reducing reliance on informal market signals.
- Probability-Based Capital Allocation: Enables systematic, risk-adjusted deployment of investment capital rather than sentiment-driven allocation.
- Brokerage and Advisory Applications: Can be integrated into brokerage platforms and financial advisory tools to deliver data-driven IPO recommendations at scale.
- Conversion of Descriptive Data to Predictive Intelligence: Transforms widely available IPO data into structured, actionable predictions—unlocking value that currently goes unrealized.

6.4 Statistical Evidence

Confusion Matrix — XGBoost Model

The confusion matrix demonstrates high overall accuracy and strong recall for profitable IPO classification, validating the practical utility of the XGBoost model in identifying listing-day winners.

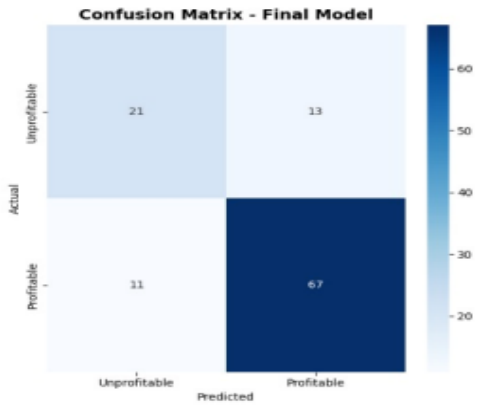

Figure 3: Confusion Matrix — XGBoost Classification Model

Regression Performance Metrics (R^2 and RMSE)

The Linear Regression model achieved the highest R^2 among the regression methods evaluated, confirming that subscription ratios and issue characteristics collectively exert statistically significant influence on listing day returns.

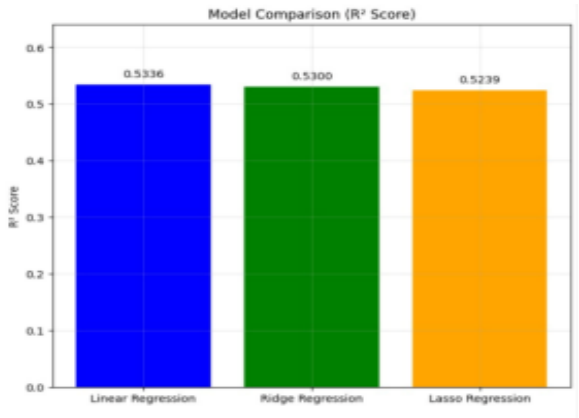

Figure 4: Comparison of Regression Model Performance Metrics

Interpretation: The aggregate statistical evidence confirms that IPO listing day performance is not purely random. Machine learning models—and XGBoost in particular—deliver reliable, actionable predictive insights. On the basis of these findings, the null hypothesis is rejected and the alternative hypothesis is supported: pre-listing subscription data and market indicators are significant determinants of IPO listing performance.

7. Conclusion and Future Scope

7.1 Conclusion

This study demonstrates that structured, pre-listing IPO data can be transformed into reliable predictive signals through a rigorously designed machine learning pipeline. The framework developed here represents a meaningful contribution to evidence-based IPO investing—bridging the gap between readily available financial data and actionable investment intelligence. Both regression and ensemble classification models showed substantive predictive power, with XGBoost emerging as the most effective tool for profit/loss classification and Random Forest excelling in recall for profitable listings.

7.2 Limitations

The study is subject to several important limitations that frame the interpretation of its findings. The model relies exclusively on historical structured IPO data and excludes unstructured data sources such as news sentiment, grey market premiums, and social media signals. While prior research on machine learning–based IPO prediction in the Indian market is limited—which partly motivates this work—the absence of benchmark comparisons warrants caution. Additionally, market volatility and episodic macroeconomic shocks may substantially reduce prediction accuracy during extreme market conditions. The framework is best understood as a decision-support tool, not a guaranteed return predictor.

7.3 Future Research Directions

Several promising avenues exist for extending this research. First, the integration of unstructured data—including financial news, analyst commentary, social media sentiment, and grey market premium signals—through Natural Language Processing (NLP) techniques could substantially improve predictive accuracy and real-time adaptability. Second, the inclusion of macroeconomic indicators (interest rates, index volatility, sector rotation trends) may enhance model robustness across varying market regimes. Third, the framework could be extended to cover SME IPOs and short-term post-listing performance beyond the listing day itself. Finally, advanced deep learning architectures (LSTM networks, transformer-based models) warrant exploration as complements or alternatives to the ensemble methods evaluated here.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785
3. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x
4. Jain, B. A., & Kini, O. (1994). The post-issue operating performance of IPO firms. The Journal of Finance, 49(5), 1699–1726. https://doi.org/10.1111/j.1540-6261.1994.tb04778.x
5. Ritter, J. R. (1991). The long-run performance of initial public offerings. The Journal of Finance, 46(1), 3–27. https://doi.org/10.1111/j.1540-6261.1991.tb03743.x
6. National Stock Exchange of India. (2024). IPO market data and statistics. https://www.nseindia.com
7. Kaggle. (2024). Indian IPO dataset (2010–2025). https://www.kaggle.com