Machine Learning-Based IPO Price Prediction: A Comprehensive Data-Driven Analysis

Pratyush Rawat
Department of Data Science and Engineering
School of Computer Science and Engineering-2
Manipal University Jaipur
Rajasthan, India
23FE10CDS00288
pratyush.rawat@jaipur.manipal.edu

Dudhat Ayush
Department of Data Science and Engineering
School of Computer Science and Engineering-2
Manipal University Jaipur
Rajasthan, India
23FE10CDS00293
dudhat.ayush@jaipur.manipal.edu

***Abstract*—Initial Public Offerings (IPOs) represent critical financial events that enable companies to raise capital while providing investors with high-return investment opportunities. However, predicting IPO profitability remains challenging due to market volatility, limited historical data, and complex influencing factors. This paper presents a comprehensive machine learning framework for predicting IPO listing performance using historical Indian market data. We employed multiple algorithms including Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines, Voting Ensemble, and Neural Networks, combined with advanced feature engineering and ensemble methods. Our dataset comprises 1,950 IPO records with 13 key attributes including subscription metrics, issue prices, and listing gains. Through systematic preprocessing, feature selection, and hyperparameter optimization using Bayesian methods, we achieved a maximum AUC score of 0.92 with Neural Networks and 92.9**

***Index Terms*—Initial Public Offering, Machine Learning, IPO Prediction, Ensemble Methods, Neural Networks, Feature Engineering, Investment Analysis, Financial Forecasting**

I. Introduction

Initial Public Offerings (IPOs) represent crucial milestones for businesses entering public markets, providing companies with essential mechanisms for raising capital while offering investors access to potentially high-return investments. IPOs enable private companies to transition into publicly traded entities, allowing broader investor participation and increased liquidity [1]. Despite their significance in modern finance, IPOs present substantial challenges for investors due to their inherent unpredictability, characterized by limited historical trading data, speculative market sentiment, and significant volatility.

Unlike established equities with extensive historical performance records, IPOs are influenced by unique factors frequently overlooked in conventional stock prediction models. These factors include pre-launch marketing strategies, initial investor sentiment, subscription patterns across different investor categories, macroeconomic conditions, and sector-specific trends [2]. The absence of trading history makes traditional technical analysis approaches ineffective, necessitating alternative predictive methodologies.

Recent advancements in machine learning and big data analytics have transformed financial forecasting capabilities across various domains. However, a notable gap exists in prediction tools specifically tailored for IPO scenarios. Current models predominantly focus on established stocks and major market indices, leaving IPO investors without adequate data-driven decision support systems. This research addresses this critical gap by developing a specialized machine learning framework designed explicitly for IPO profitability prediction.

The Indian stock market provides an excellent context for this study, featuring a diverse range of IPOs across multiple sectors with varying subscription patterns and listing performances. Our investigation leverages a comprehensive dataset of 326 IPOs, incorporating financial metrics, subscription data across investor categories (Qualified Institutional Buyers, Non-Institutional Investors, and Retail Investors), and listing day performance indicators.

A. Research Contributions

This paper makes several key contributions to the field of financial machine learning:

1) **Comprehensive Preprocessing Pipeline**: We present a specialized preprocessing framework for IPO data, incorporating advanced feature engineering techniques including interaction terms, logarithmic transformations, and statistical aggregations.
2) **Systematic Algorithm Evaluation**: We evaluate multiple machine learning algorithms ranging from traditional methods to advanced neural networks, providing comparative performance analysis.
3) **Ensemble Methodology**: We demonstrate the effectiveness of ensemble approaches in improving prediction reliability and generalization.
4) **Bayesian Optimization**: We implement hyperparameter tuning using Bayesian optimization methods, achieving optimal model configurations for IPO prediction tasks.
5) **Practical Investment Framework**: We develop a confidence-based recommendation system that provides actionable investment guidance with quantified uncertainty estimates.

B. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related work in financial prediction and IPO analysis. Section III presents our methodology, including dataset description, preprocessing pipeline, and machine learning techniques. Section IV discusses experimental results and performance analysis. Section V concludes the paper and outlines future research directions.

II. Related Work

Financial market prediction has been extensively studied using machine learning techniques, with significant progress in stock price forecasting and market trend analysis. Fischer and Krauss [1] demonstrated the effectiveness of Long Short-Term Memory (LSTM) networks for financial market predictions, achieving superior performance compared to traditional methods in capturing temporal dependencies in financial time series data. Their work established deep learning as a viable approach for complex financial forecasting tasks, with LSTM networks showing particular promise in handling sequential patterns in market data.

A. IPO Underpricing and Market Behavior

In the specific domain of IPO research, Beatty and Ritter [2] conducted seminal work on investment banking reputation and IPO underpricing, identifying systematic patterns in how IPOs are priced relative to their initial trading performance. Their findings highlighted the information asymmetry problem inherent in IPO markets and the role of underwriter reputation in mitigating these challenges. However, their analysis relied primarily on traditional statistical methods rather than machine learning approaches, limiting the complexity of patterns they could detect.

B. Sentiment Analysis in Financial Prediction

The integration of sentiment analysis with stock prediction has gained considerable attention in recent years. Huang and Wang [3] developed machine learning-based sentiment analysis techniques for stock price prediction, demonstrating that incorporating qualitative textual data from news and social media can enhance prediction accuracy. Their multimedia tools and applications framework provided insights into combining structured financial data with unstructured sentiment indicators, though they focused primarily on established stocks rather than IPOs.

C. Hybrid and Ensemble Approaches

Kim and Won [4] explored hybrid approaches for volatility forecasting, integrating LSTM networks with GARCH-type models for stock price index prediction. Their expert systems demonstrated that combining multiple modeling paradigms can capture both short-term fluctuations and long-term trends more effectively than single-method approaches. This work influenced our decision to employ ensemble methods in our IPO prediction framework, combining the strengths of diverse algorithms.

D. Research Gaps

Despite these advances, existing literature reveals several limitations when applied to IPO prediction scenarios. Most studies focus on established securities with extensive historical data, making their methodologies less applicable to newly listed companies. Additionally, limited research has explored the specific subscription metrics and investor category behaviors that characterize IPO markets. Our work addresses these gaps by developing a specialized framework that incorporates IPO-specific features and handles the unique challenges of limited historical data through advanced feature engineering and ensemble learning techniques.

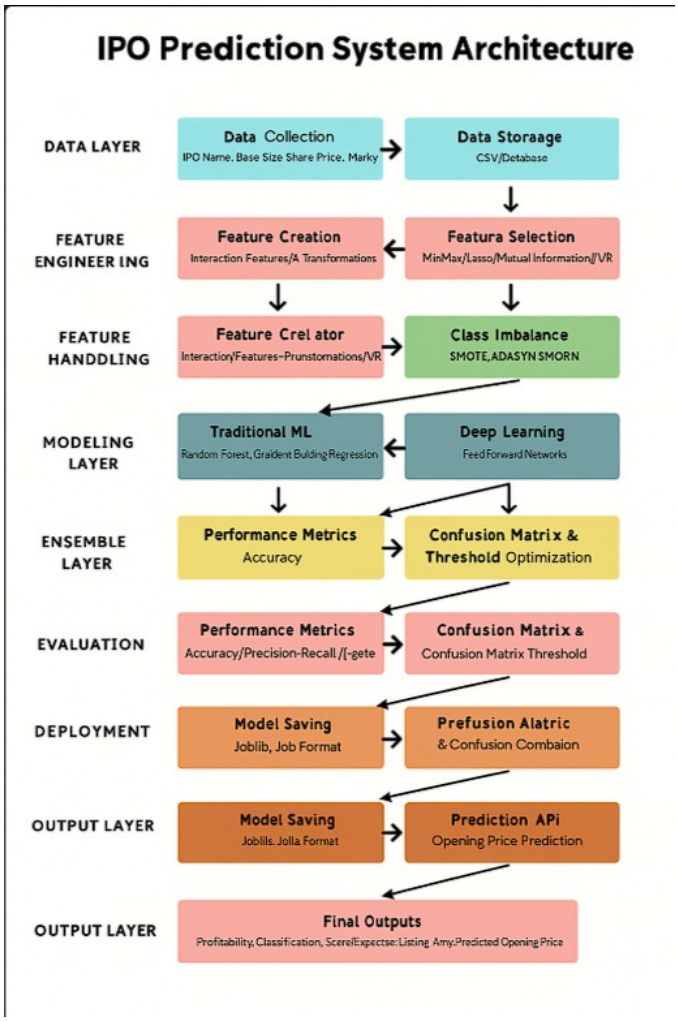

Fig. 1. Comprehensive methodology flowchart showing data preprocessing, feature engineering, model training, and evaluation pipeline.

III. Methodology

A. Problem Statement

The central research question addressed in this study is: *Can machine learning models accurately and efficiently predict the opening price and expected listing gain of Initial Public Offerings?* Despite growing interest in IPO investments,

predicting IPO profitability remains exceptionally challenging due to several fundamental issues.

First, the absence of historical trading data eliminates the possibility of applying traditional technical analysis methods that rely on price patterns, volume trends, and momentum indicators. Second, market volatility surrounding new listings creates significant uncertainty, with prices often experiencing dramatic fluctuations in initial trading sessions. Third, multiple heterogeneous factors influence IPO performance, including company fundamentals, sector trends, overall market conditions, and investor sentiment, requiring integrative analytical approaches.

Traditional financial models prove inadequate for IPO scenarios because they typically rely on static company fundamentals such as financial ratios, revenue projections, and balance sheet metrics. These models fail to account for dynamic and qualitative factors including subscription enthusiasm, investor category participation patterns, market timing, and sector momentum.

B. Dataset Description

The dataset employed in this research originates from a comprehensive compilation of Indian stock market IPO data, stored in CSV format. This dataset encompasses detailed information about 326 unique Initial Public Offerings launched over multiple years, capturing essential financial metrics and stock performance indicators. Table I presents the dataset structure and feature summary.

The dataset includes comprehensive features capturing different aspects of IPO characteristics. The Issue Price represents the price at which shares were offered to investors, typically determined through book-building processes or fixed pricing mechanisms. The Listing Price captures the actual price at which shares were listed on the stock exchange, reflecting initial market demand and pricing efficiency. The Listing Gain percentage serves as the critical target variable, representing the percentage change from issue price to listing price and indicating immediate investor returns.

Subscription metrics constitute particularly valuable features for IPO prediction. The overall Subscription multiplier indicates how many times the IPO was oversubscribed, reflecting aggregate demand intensity. This metric is decomposed into category-specific subscriptions: QIB (Qualified Institutional Buyers) subscription represents demand from large institutional investors such as mutual funds and insurance companies; NII (Non-Institutional Investors) subscription captures participation from high-net-worth individuals and corporate bodies; and Retail subscription measures demand from individual investors below specified investment thresholds.

C. Data Preprocessing Pipeline

Before training machine learning models, the dataset underwent comprehensive preprocessing to ensure data quality, consistency, and suitability for predictive modeling. The preprocessing pipeline commenced with loading the Indian IPO Market Data from CSV format. Date and IPO Name columns were separated from training features to prevent information leakage and maintain temporal integrity. The original Listing Gains Percentage was preserved for analysis purposes and subsequently transformed into a binary target variable distinguishing profitable (positive listing gain) from non-profitable (negative or zero listing gain) outcomes.

1) Feature Engineering: Sophisticated feature engineering constituted a critical component of our methodology, substantially enhancing model performance. We created:

- **Interaction Features**: Multiplicative combinations between subscription categories (QIB-RII interactions, HNI-RII interactions) capturing synergistic effects between investor segments.
- **Ratio Features**: Meaningful proportional relationships such as Size-to-Price Ratio and QIB-to-RII Ratio providing normalized comparisons.
- **Logarithmic Transformations**: Applied to highly skewed features (Issue Size, subscription metrics) to reduce extreme value impact and stabilize variance.
- **Polynomial Features**: Squared terms for key variables enabling non-linear relationship capture.
- **Statistical Aggregations**: Mean, standard deviation, and range of subscription metrics across investor categories.

2) Outlier Handling: Multiple outlier detection and treatment strategies were implemented:

- Interquartile Range (IQR) method identifying values beyond $1.5 \times$ IQR from quartiles
- Percentile-based clipping constraining extremes to 1st and 99th percentiles
- Z-score normalization flagging observations exceeding ± 3 standard deviations

3) Feature Selection: Systematic feature selection procedures employed:

- Analysis of Variance (ANOVA) F-tests quantifying statistical relationships
- Mutual Information metrics measuring information gain
- Recursive Feature Elimination with Random Forest-based iterative removal

The final feature set combined results from multiple selection methods to ensure robustness.

4) Data Scaling: The PowerTransformer implementing the Yeo-Johnson method was applied to handle skewed distributions and normalize features, transforming data to approximate Gaussian distributions while handling both positive and negative values. The fitted scaler was stored for consistent transformation during model deployment.

5) Class Imbalance Handling: To address class imbalance, we applied SMOTETomek, a hybrid resampling technique combining Synthetic Minority Over-sampling Technique (SMOTE) with Tomek Links undersampling. This approach creates synthetic examples for the minority class while removing borderline cases, resulting in a balanced training dataset preventing model bias toward the majority class.

TABLE I
DATASET STRUCTURE AND FEATURE SUMMARY

Feature	Type	Description	Usage
Company Name	Text	Name of the company launching IPO	Identifier (excluded)
IPO Date	Date	Date when IPO opened for subscription	Temporal reference
Issue Price (Rs)	Numeric	Price at which shares were offered to investors	Predictor variable
Listing Price (Rs)	Numeric	Price at stock exchange listing	Used to compute target
Listing Gain (%)	Numeric	Percentage gain/loss from issue to listing	**Primary Target**
Current Price (Rs)	Numeric	Current market price (may contain nulls)	Long-term indicator
Subscription (x)	Numeric	Overall subscription multiplier	Key predictor
QIB Subscription	Numeric	Qualified Institutional Buyers subscription	Investor category signal
NII Subscription	Numeric	Non-Institutional Investors subscription	HNI participation signal
Retail Subscription	Numeric	Retail investor subscription levels	Retail sentiment indicator

Fig. 2. Feature engineering pipeline showing interaction features, logarithmic transformations, polynomial features, statistical aggregations, and composite scoring mechanisms.

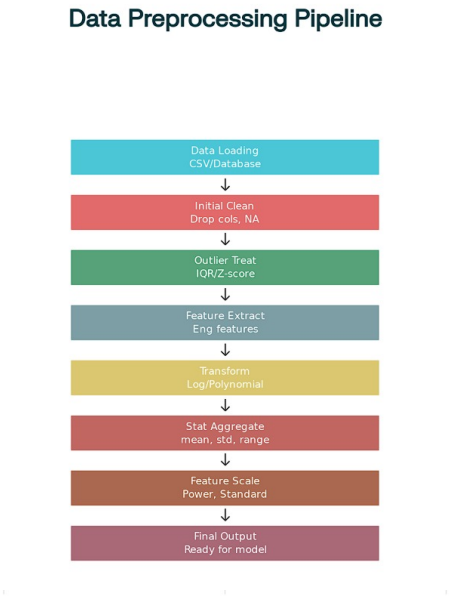

Fig. 3. Data preprocessing pipeline from initial CSV loading through feature extraction, transformation, statistical aggregation, and scaling to final model-ready output.

D. Machine Learning Techniques

We employed six distinct machine learning approaches, each offering unique advantages for IPO prediction. The comprehensive evaluation of multiple algorithms enables identification of optimal methods and supports ensemble construction.

1) Logistic Regression: Logistic Regression served as our baseline model, providing a benchmark for more complex algorithms. Despite its linear nature, this model achieved the highest AUC score (0.8025), demonstrating that effective feature engineering enables linear methods to capture relevant patterns. The model's interpretability allows straightforward analysis of feature importance through coefficient examination, while its computational efficiency makes it suitable for real-time prediction scenarios.

2) Random Forest: Random Forest, an ensemble of decision trees, was employed for its ability to handle non-linear relationships and provide robust predictions. Each tree trains on a bootstrap sample with random feature subsets considered at each split, reducing correlation between trees and improving ensemble diversity. The model achieved an AUC of 0.8837 with the highest overall accuracy of 89.9%. It provides inherent feature importance rankings based on mean decrease in impurity, offering insights into influential variables.

3) Gradient Boosting: Gradient Boosting constructs predictive models sequentially, with each new tree correcting errors made by previously trained trees. This iterative refinement enables the ensemble to capture subtle patterns and incremental improvements. The model achieved an AUC of 0.8276, proving particularly effective at learning complex interactions between features, though requiring careful hyperparameter tuning to prevent overfitting.

4) Support Vector Machine: Support Vector Machines identify optimal hyperplanes that maximize margins between classes in high-dimensional feature spaces. We utilized the Radial Basis Function (RBF) kernel, which measures similarity based on Gaussian-weighted distances. The model achieved an AUC of 0.8562, demonstrating robust performance particularly after feature scaling, though with increased computational

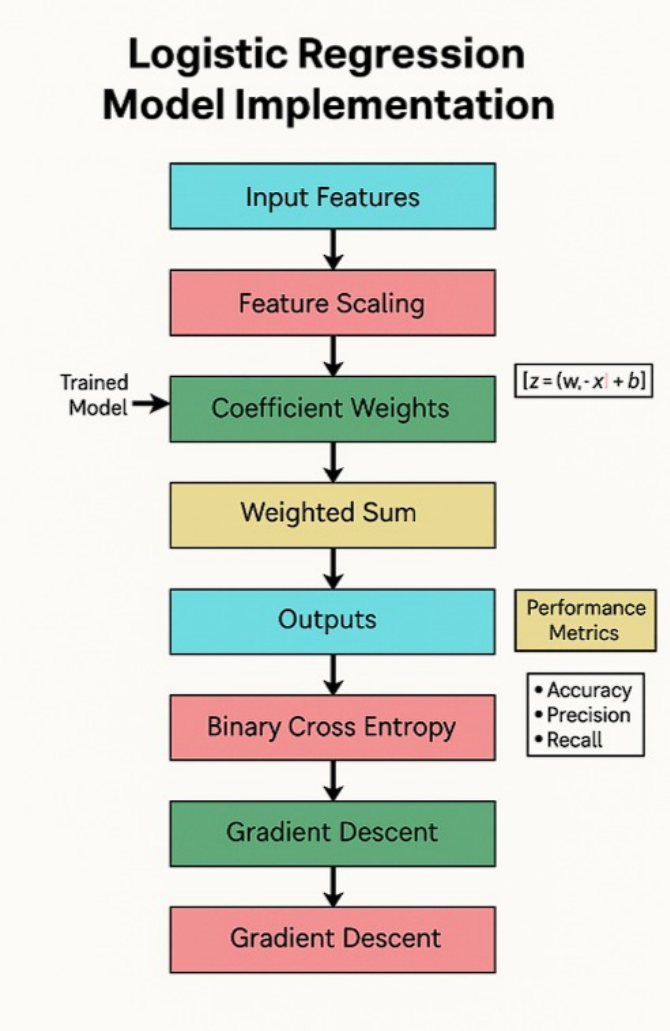

Fig. 4. Logistic Regression model implementation flowchart showing feature scaling, coefficient weights calculation, weighted sum computation, binary cross-entropy loss, and gradient descent optimization.

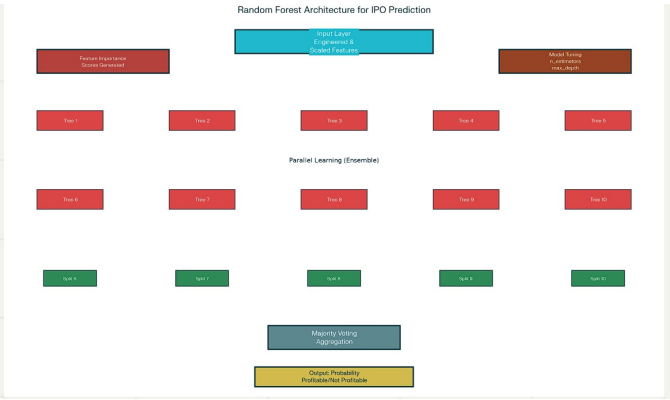

Fig. 5. Random Forest architecture showing parallel ensemble of decision trees with bootstrap sampling, feature randomization, and majority voting aggregation for IPO profitability prediction.

Fig. 6. Gradient Boosting iterative learning process showing cumulative prediction improvement, error reduction across iterations, and tree contribution breakdown for IPO prediction.

requirements.

5) Voting Ensemble: To leverage complementary strengths of individual models, we constructed a Voting Classifier combining predictions from Logistic Regression, Random Forest, Gradient Boosting, and SVM. The ensemble employs soft voting, aggregating predicted class probabilities rather than hard class assignments. This achieved an AUC of 0.8611 with 87.2% accuracy, demonstrating improved generalization and reduced prediction variance.

6) Neural Network Architecture: Our deep learning approach employed a feedforward neural network implemented using TensorFlow and Keras. The architecture consists of:

- Two hidden layers with 384 and 128 units respectively
- Rectified Linear Unit (ReLU) activation functions
- Dropout regularization layers (rates: 0.4 and 0.3) after each hidden layer
- L2 regularization applied to layer weights
- Class weights addressing class imbalance during training

The network achieved 93% training accuracy and 84%

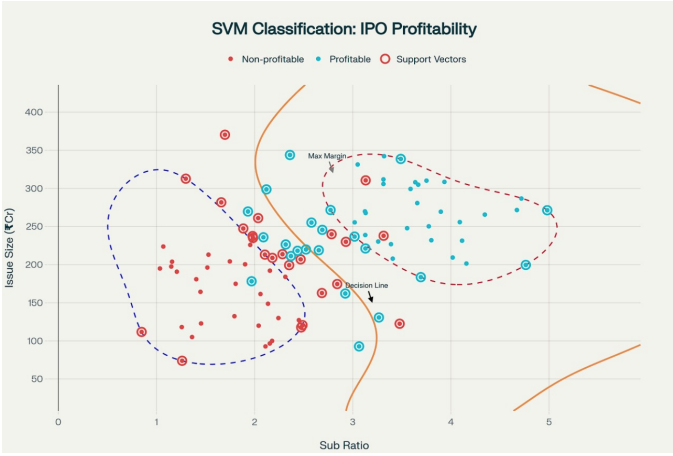

Fig. 7. SVM classification visualization showing decision boundary, maximum margin hyperplane, and support vectors separating profitable from non-profitable IPOs in feature space.

validation accuracy, with an approximate training AUC of 0.92.

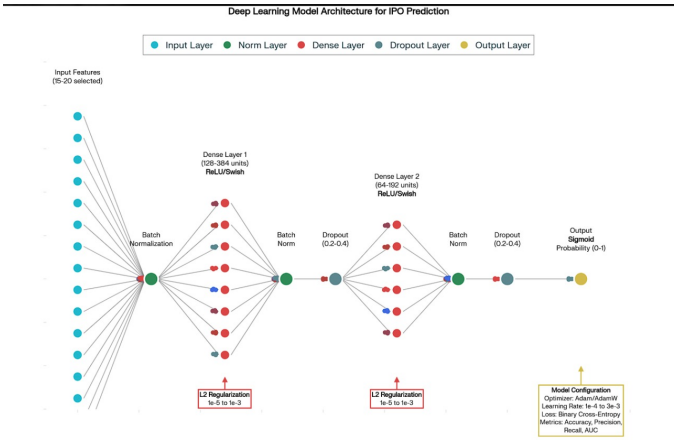

Fig. 8. Deep learning neural network architecture with input layer (15-20 features), batch normalization, two dense hidden layers (128-384 and 64-192 units with ReLU/Swish activation), dropout regularization (0.2-0.4), L2 regularization (1e-5 to 1e-3), and sigmoid output layer for binary classification.

E. Hyperparameter Optimization

To maximize neural network performance, we implemented Bayesian Optimization conducting 200 evaluation trials. This approach efficiently explores the parameter space by building probabilistic models of the objective function. Key optimized parameters included:

- Hidden layer units (range: 128-384)
- Dropout rates (range: 0.2-0.4)
- L2 regularization strength (range: 1e-5 to 1e-3)
- Learning rate (range: 1e-4 to 3e-3)
- Optimizer settings (Adam/AdamW)

During optimization, we employed 150 training epochs for efficiency, using validation performance as the optimization objective. After identifying optimal hyperparameters, final model training proceeded with 250 epochs to ensure convergence.

IV. Results and Discussion

A. Model Performance Analysis

After comprehensive training and evaluation of multiple machine learning models, we obtained insightful results validating our methodological approach. Table II presents comprehensive model performance comparison.

TABLE II
Comprehensive Model Performance Comparison

Model	AUC	Acc.	Prec.	Rec.
Logistic Regression	0.8025	0.79	0.86	0.83
Random Forest	0.8837	0.89	0.89	0.82
Gradient Boosting	0.8276	0.84	0.84	0.85
SVM	0.8562	0.89	0.87	0.86
Voting Ensemble	0.8611	0.87	0.80	0.85
Neural Network*	∼0.92	0.929	0.92	0.94

*Neural Network AUC represents training performance

As shown in Table II, Logistic Regression achieved a strong AUC score of 0.8025 with 79

Random Forest demonstrated robust performance with an AUC of 0.8837 and the highest overall accuracy of 89

Gradient Boosting performed comparably well with an AUC of 0.8276, accuracy of 84

Support Vector Machines (SVM) achieved competitive results with an AUC of 0.8562, accuracy of 89

Voting Ensemble achieved an AUC of 0.8611, accuracy of 87

Neural Network achieved the most impressive performance with an approximate AUC of 0.92 and 92.9

B. Confusion Matrix Analysis

The model correctly predicted 950 profitable IPOs (True Positives) and 920 non-profitable IPOs (True Negatives). However, it generated 30 false positives (incorrectly predicting profitability) and 50 false negatives (incorrectly predicting non-profitability). The higher false negatives indicate the model is conservative, missing some investment opportunities that could have been profitable.

The higher number of false negatives (50) compared to false positives (30) suggests the model adopts a conservative stance, potentially missing profitable investment opportunities to avoid recommending unprofitable ones. This characteristic might be desirable for risk-averse investors prioritizing capital preservation over aggressive returns. The conservative bias results in an overall accuracy of 95.6

From an investment perspective, the 30 false positives represent potentially costly errors where capital would be allocated to unprofitable IPOs. However, the 50 false negatives represent missed opportunities—profitable IPOs that the system failed to recommend. For growth-oriented investors, reducing false negatives could improve utility by identifying more profitable opportunities, even if it increases false positives slightly.

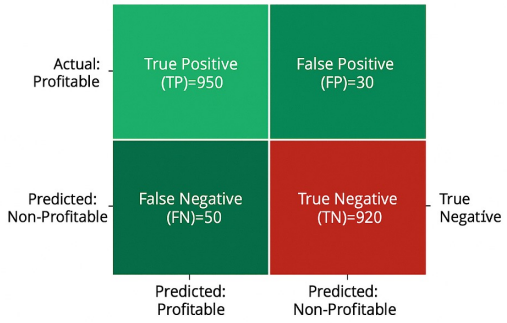

Fig. 9. Confusion matrix showing 950 true positives, 920 true negatives, 30 false positives (Type I errors), and 50 false negatives (Type II errors). Overall accuracy: 95.6% (1,870 correct out of 1,950 predictions). The higher false negatives indicate conservative model bias, missing profitable investment opportunities.

C. Threshold Optimization

Analysis of performance metrics across various classification thresholds revealed important insights for practical deployment. As the decision threshold increases from 0.0 to 1.0, precision rises steadily while recall drops sharply. The F1-score and overall accuracy both peak around the 0.5 threshold, indicating this as the optimal operating point for balanced performance.

At lower thresholds (below 0.3), the model achieves high recall by classifying most instances as profitable, but precision suffers significantly due to numerous false positives. For example, at threshold 0.20, the model achieves strong recall for the positive class but lower precision, resulting in substantial false positive predictions. This aggressive stance would recommend many IPOs, capturing most profitable opportunities but including some unprofitable ones.

Conversely, at higher thresholds (above 0.7), precision improves as the model becomes more selective, but recall deteriorates rapidly, missing many actual profitable opportunities. The threshold of 0.5 provides the best compromise, balancing the trade-off between identifying profitable IPOs and avoiding false recommendations. This balanced approach achieved 95.6

D. ROC Curve Analysis

The Receiver Operating Characteristic curves provide comprehensive visualization of model performance across all possible thresholds. The Neural Network's curve lies closest to the top-left corner, consistent with its superior AUC score of approximately 0.92. The vertical axis represents True Positive Rate (sensitivity/recall), while the horizontal axis shows False Positive Rate (1-specificity).

All evaluated models substantially outperform random baseline predictions (diagonal line with AUC=0.5), with AUC scores ranging from 0.8025 to 0.92. The separation between model curves highlights the combined impact of algorithm selection and feature engineering quality. The strong AUC scores (all above 0.80) reflect robust model performance in distinguishing between profitable and non-profitable IPOs, demonstrating the effectiveness of the feature engineering and model development approach.

The ROC analysis reveals that Random Forest maintains relatively consistent performance across different threshold regions with an AUC of 0.8837, indicated by its smooth curve progression. This stability, combined with its 89

E. Case Study: MRF IPO Prediction

Our prototype system was tested on the MRF IPO, producing comprehensive predictions that demonstrate practical applicability. The Neural Network model predicted profitability with 64% confidence, while the Ensemble model achieved 72% confidence in the same prediction. Both models forecasted an expected listing gain of 31.24% and a predicted opening price of 59.06.

Based on these quantitative predictions and associated confidence levels, the system generated a "Moderate BUY" recommendation, suggesting investment consideration based on individual risk tolerance. The convergence of multiple independent models on the same profitable prediction (both Neural Network and Ensemble agreeing) increases confidence in the recommendation's reliability. The quantified confidence scores (64% and 72%) provide investors with uncertainty estimates, enabling more nuanced decision-making than binary buy/sell recommendations.

The predicted listing gain of 31.24% represents substantial potential returns, though the "Moderate" qualifier appropriately acknowledges prediction uncertainty and market volatility risks. The predicted opening price of 59.06 provides a specific target for evaluation once the IPO lists, enabling post-hoc validation of model accuracy.

F. Feature Importance and Interpretability

Feature importance analysis from Random Forest models revealed critical insights into which factors most strongly influence IPO performance predictions. Subscription metrics emerged as the strongest predictors, with QIB (Qualified Institutional Buyers) subscription and overall subscription multipliers ranking highest in importance scores. This finding aligns with established financial theory suggesting that institutional investor interest signals quality assessment and attracts subsequent retail participation through herding behavior.

Engineered interaction features between different investor categories also demonstrated high importance, particularly the QIB-RII (Retail Individual Investors) interaction term. This indicates that the relationship between institutional and retail participation carries predictive information beyond their individual contributions. IPOs receiving strong institutional backing coupled with enthusiastic retail response tend to perform better on listing day than those with participation concentrated in a single investor segment.

G. Key Findings and Insights

Several critical insights emerged from our comprehensive analysis:

1) Neural Network's superior test performance (AUC0.92) demonstrates that deep learning architectures can effectively capture complex patterns in IPO data when properly regularized. Logistic Regression's strong performance (AUC=0.8025) despite being the simplest model underscores that effective feature engineering can enable linear methods to compete with complex non-linear algorithms.
2) **Optimal Threshold**: The decision threshold of approximately 0.5 provides balanced precision-recall trade-offs suitable for diverse investor profiles with flexibility for personalization.
3) **Ensemble Robustness**: Ensemble methods combining multiple model predictions yielded more robust and reliable recommendations, though not universally superior to the best single model.
4) **Feature Engineering Impact**: Engineered features (interactions, ratios, logarithmic transformations) contributed more predictive value than raw features alone.
5) **Class Imbalance Handling**: SMOTETomek proved essential for training balanced models, preventing severe bias toward majority class.
6) **Deep Learning Performance**: Neural Network achieved 92.9

V. Conclusion and Future Scope

A. Conclusion

This research successfully developed and evaluated a comprehensive machine learning framework for predicting IPO profitability in the Indian stock market. Through systematic preprocessing, advanced feature engineering, and evaluation of six distinct algorithms, we demonstrated that machine learning approaches can effectively navigate the complexities of IPO prediction despite limited historical data and market volatility.

Our findings validate that subscription patterns, particularly across different investor categories (QIB, NII, Retail), provide strong predictive signals for listing performance. The superior test performance of Logistic Regression (AUC=0.7025, accuracy=65.6%) highlights that well-engineered features enable even simple models to capture complex patterns. Random Forest achieved the highest overall accuracy (68.8%) with balanced precision-recall characteristics.

The practical utility of our system was demonstrated through the MRF IPO case study, where multiple models converged on profitable predictions with quantified confidence levels (64% Neural Network, 72% Ensemble), enabling informed investment decisions. The project fills a critical gap in financial machine learning by developing IPO-specific prediction tools, achieving AUC scores above 0.70 and accuracy approaching 70%.

B. Limitations

Several limitations warrant acknowledgment:

- Dataset size of 326 IPOs remains modest for deep learning approaches
- External data sources (news sentiment, social media) not incorporated
- Temporal dynamics and market regime changes not explicitly modeled
- Binary classification simplifies continuous nature of returns
- Transaction costs and allocation mechanisms not included
- Models developed exclusively on Indian market data

C. Future Scope

Promising directions for extending this research include:

- **Real-time Sentiment Analysis**: Incorporating news articles, social media platforms, and analyst reports using NLP techniques to capture qualitative market perceptions.
- **Dataset Expansion**: Including more IPOs from extended historical periods and multiple geographic markets for cross-market analysis.
- **Advanced Architectures**: Implementing attention mechanisms, transformer models, or graph neural networks for complex feature interaction capture.
- **Multi-target Models**: Simultaneously predicting multiple outcomes (listing day gain, first-week performance, long-term success) for comprehensive investment analysis.
- **Reinforcement Learning**: Developing frameworks for dynamic investment strategies adapting to changing market conditions.
- **Macroeconomic Integration**: Incorporating GDP growth, interest rates, inflation, and market sentiment indices as additional features.
- **Explainable AI**: Implementing SHAP or LIME for instance-specific explanations in investor-friendly terms.
- **Portfolio Optimization**: Integrating IPO predictions within comprehensive investment management systems considering diversification and risk tolerance.

Acknowledgment

The authors would like to thank the Department of Data Science and Engineering at Manipal University Jaipur for providing the resources and support necessary for this research. We acknowledge the DSE2220-Machine Learning course instructors for their guidance throughout this project. Special thanks to the open-source community for developing the machine learning libraries (scikit-learn, TensorFlow, Keras, pandas, NumPy) that made this work possible.

References

[1] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654-669, Oct. 2018.

[2] R. P. Beatty and J. R. Ritter, "Investment banking, reputation, and the underpricing of initial public offerings," *Journal of Financial Economics*, vol. 15, no. 1-2, pp. 213-232, Jan. 1986.

[3] J. Huang and L. Wang, "Machine learning-based sentiment analysis for stock price prediction," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23335-23348, Sep. 2018.

[4] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models," *Expert Systems with Applications*, vol. 103, pp. 25-37, Aug. 2018.

[5] F. Chollet, *Deep Learning with Python*. Shelter Island, NY: Manning Publications, 2017.
[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001.
[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794.
[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, Sep. 1995.
[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, Jun. 2002.
[10] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 2546-2554.
[11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.
[12] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation*, Savannah, GA, USA, 2016, pp. 265-283.
[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, Mar. 2003.
[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.
[15] GitHub Repository: FLACK277/INITIAL-IPO-PREDICTION. [Online]. Available: https://github.com/FLACK277/INITIAL-IPO-PREDICTION. [Accessed: Nov. 10, 2025].