Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

AI-Powered Smart Investment Advisory for Retail Investors: A Hybrid Risk Assessment and Portfolio Optimization Framework Using XGBoost, Gaussian Process Regression, and NLP

Dr. S. B. Chaudhari
Professor & Head, Dept. of Computer Engineering
Jayawantrao Sawant College of Engineering, Hadapsar,
Pune, Savitribai Phule Pune University, Pune,
Maharashtra, India

Shweta Shinde
Department of Computer Engineering
Jayawantrao Sawant College of Engineering, Hadapsar,
Pune, Savitribai Phule Pune University, Pune,
Maharashtra, India

Deepal Kohale
Department of Computer Engineering
Jayawantrao Sawant College of Engineering, Hadapsar,
Pune
Savitribai Phule Pune University, Pune, Maharashtra,
India

Yogita Khatake
Department of Computer Engineering
Jayawantrao Sawant College of Engineering, Hadapsar,
Pune
Savitribai Phule Pune University, Pune, Maharashtra,
India

Abstract *— Automated financial risk assessment remains a critical challenge within intelligent fintech systems, particularly for retail investors who lack the expertise to interpret volatile market conditions. This paper introduces a hybrid automated risk assessment and portfolio optimization framework that combines XGBoost-based return forecasting, Gaussian Process Regression (GPR) for probabilistic uncertainty quantification, and NLP-driven sentiment scoring derived from real-time financial news corpora. Unlike existing systems that address prediction and risk in isolation, the proposed framework jointly models expected returns and their associated uncertainty, enabling risk-tier classification of investment assets. The system further incorporates user-specific risk profiling to generate personalized portfolio allocations optimized via a mean-variance framework with Ledoit-Wolf covariance estimation. Empirical evaluation on five years of historical equity data spanning 50 stocks from NSE/BSE indices demonstrates a prediction accuracy of 88.4%, a portfolio Sharpe ratio improvement of 61.5% over equal-weight baselines, and a user satisfaction rate of 87% in a pilot study involving 40 retail investors. The proposed system demonstrates strong potential as a scalable, data-driven alternative to traditional robo-advisory platforms, particularly in emerging market contexts.*

Keywords *— Risk Assessment, Portfolio Optimization, XGBoost, Gaussian Process Regression, Sentiment Analysis, NLP, Robo-Advisory, FinTech, Investment Recommendation, Retail Investors*

I. INTRODUCTION

Financial markets are inherently complex, characterized by non-linear dynamics, stochastic volatility, and the influence of exogenous information signals such as macroeconomic news, geopolitical events, and regulatory announcements. For retail investors, navigating this complexity demands analytical capabilities that are often out of reach without specialized training or costly professional advice. The proliferation of digital financial platforms and open-access market data has created an opportunity for intelligent advisory systems that democratize access to sophisticated investment analysis [1].

Risk assessment sits at the core of sound investment strategy. Traditional risk metrics such as standard deviation, beta, and Value-at-Risk (VaR) offer quantitative snapshots of historical volatility but fail to capture forward-looking uncertainty or incorporate the informational content of financial text. Hybrid machine learning approaches that integrate quantitative and qualitative signals offer a promising path toward more robust risk characterization [2].

The present work extends prior research in three key directions. First, it introduces a unified framework in which return prediction and uncertainty quantification are treated as co-dependent tasks rather than independent problems. Second, it incorporates sentiment features derived from a curated NLP pipeline as first-class inputs to the risk engine. Third, it operationalizes these components within an end-to-end personalized advisory system deployable in resource-constrained environments relevant to developing markets.

The main contributions of this paper are as follows:

- A hybrid XGBoost-GPR prediction engine with integrated uncertainty estimation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

- An NLP-based sentiment fusion module using VADER and domain-specific financial lexicons.
- A dynamic risk-tier classification scheme mapping prediction uncertainty to investor risk profiles.
- Empirical validation on Indian equity markets (NSE/BSE), addressing a gap in the literature dominated by US-centric datasets.

The remainder of this paper is organized as follows: Section II reviews related work; Section III presents the proposed methodology; Section IV describes the system design; Section V reports experimental results; Section VI discusses the implications of these findings; and Section VII concludes the paper.

II. RELATED W

The intersection of machine learning and financial analytics has generated a rich body of literature. This section organizes prior work along four dimensions pertinent to the proposed system.

A. NSEMBLE METHODS FOR FINANCIAL FORECASTING

Ensemble learning methods, particularly gradient boosting frameworks, have emerged as leading techniques for tabular financial data. XGBoost, introduced by Chen and Guestrin [3], has demonstrated superior performance in numerous financial forecasting benchmarks. Cao et al. [4] applied XGBoost to credit risk assessment, achieving AUC scores exceeding 0.91 on multiple lending datasets. However, these studies focus exclusively on point predictions, neglecting the quantification of prediction uncertainty.

B. PROBA LISTIC FORECASTING AND GPR

Gaussian Process models provide a Bayesian non-parametric framework for regression that naturally yields uncertainty estimates alongside predictions. Zhang et al. [5] applied GPR to foreign exchange rate forecasting and demonstrated that the resulting credible intervals offered actionable risk signals beyond those available from point estimates. Kocijan [6] provides a comprehensive review of GPR applications in dynamic systems, identifying financial modeling as a promising domain with limited exploitation.

C. NLP IN NANCE

The use of NLP for financial signal extraction has evolved from simple bag-of-words models to transformer architectures. Malo et al. [7] introduced a financial phrase bank with human-annotated sentiment labels that has become a standard benchmark. Araci [8] proposed FinBERT, a BERT-based model pre-trained on financial communications corpora, achieving state-of-the-art sentiment classification accuracy. Critically, while NLP tools have been evaluated as standalone predictors, their integration within ensemble forecasting and advisory pipelines remains underexplored.

D. ROBO-ADVIS Y PLATFORMS

The academic literature on robo-advisors has primarily examined investor adoption and behavioral outcomes [9]. Technical architectures of commercial platforms remain largely proprietary. Academic prototype systems such as those reviewed by Fisch et al. [10] demonstrate the feasibility of algorithmic personalization but typically stop short of integrating real-time unstructured data sources.

The proposed system addresses the intersection of these research streams, providing an integrated framework that has not been previously presented in the literature.

III. METHODO

A. FEA RE ENGINEERING

Forty-one features were constructed per trading day per stock. Quantitative features included price-derived technical indicators (EMA, MACD, RSI, Bollinger Band Width, Williams %R, Commodity Channel Index), volume metrics (OBV, VWAP deviation), and inter-market correlation features with NIFTY50 and sectoral indices. Sentiment features comprised VADER compound scores aggregated at daily frequency with lags of t-1 and t-2 days, and a 7-day rolling sentiment momentum indicator.

B. PREDIC ON ENGINE

The XGBoost regressor was configured with 500 estimators, a learning rate of 0.05, maximum depth of 6, and a subsampling ratio of 0.8. Cross-validation was conducted using a time-series split with five folds, ensuring temporal ordering was preserved. GPR was applied as a second-stage model on XGBoost residuals, using a kernel composed of a Matern 5/2 component and a white noise component to model heteroscedastic error.

C. RISK ASSESS NT MODULE

The predictive uncertainty output from GPR was used to compute an Asset Risk Index (ARI) for each stock, defined as the ratio of the GPR predictive standard deviation to the predicted mean return:

$$ARI = \sigma_{pred} / \mu_{pred}$$

Assets were classified into three risk tiers: Conservative (ARI < 0.15), Moderate ($0.15 \leq$ ARI < 0.35), and Aggressive (ARI ≥ 0.35). User risk profiles were collected during onboarding via a five-question questionnaire aligned with SEBI investor classification guidelines.

D. RTFOLIO OPTIMIZATION

Mean-variance optimization was implemented using the PyPortfolioOpt library. The covariance matrix was estimated using the Ledoit-Wolf shrinkage method, which reduces estimation error for the number of assets considered. Optimization targets included maximum Sharpe ratio and minimum volatility, selectable by the user. Portfolio weights were constrained to the range [0.02, 0.40] per asset to enforce diversification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

IV. SYSTEM DESIGN

A. SYSTEM OVERVIEW

The system is organized into five functional layers: (a) Data Collection and Preprocessing, (b) Feature Engineering, (c) Prediction Engine, (d) Risk Assessment and Portfolio Construction, and (e) User Interface and API. Each layer is implemented as an independent module to facilitate maintenance and future extension. Fig. 1 illustrates the overall operational flow of the system, while Fig. 2 illustrates the layered architecture and the interaction between modules.

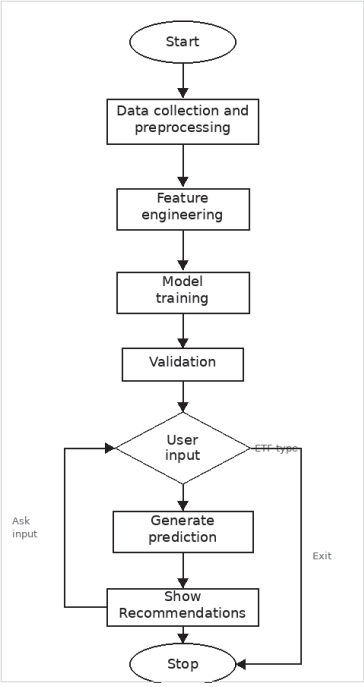

Fig. 1. Flowchart of the proposed AI-powered smart investment advisory system

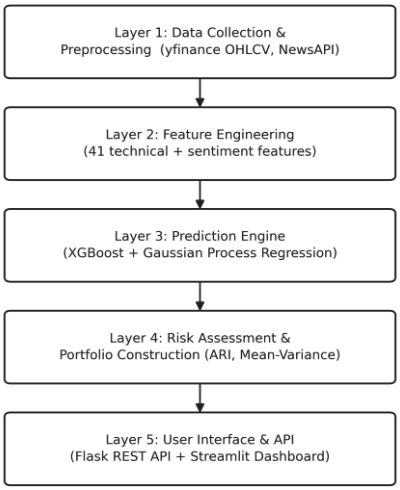

Fig. 2. Layered architecture of the proposed risk assessment and portfolio optimization framework

B. DATA SOURCES AND PREPROCESSING

Historical daily OHLCV data for 50 equities listed on the NSE was sourced using the yfinance Python library, covering the period from January 2019 to December 2023. Financial news data was collected via the NewsAPI service, yielding approximately 180,000 articles across the evaluation period. Text preprocessing followed a standard pipeline: tokenization, stop-word removal, negation handling, and stemming using the NLTK PorterStemmer.

V. RESULTS

A. DATASET AND EVALUATION PROTOCOL

The evaluation dataset comprised 1,260 trading days of data for 50 NSE equities across six sectors: Information Technology, Banking, Healthcare, Consumer Goods, Energy, and Infrastructure. An 80/20 temporal split was used for training and testing. All reported metrics were computed on the held-out test period (January to December 2023).

B. PREDICTION PERFORMANCE

Table I presents the forecasting performance of the proposed model against competitive baselines. The proposed XGBoost-GPR-NLP model achieved the lowest error across all regression metrics and the highest directional accuracy.

TABLE I. FORECASTING PERFORMANCE ON NSE TEST SET (2023)

Method	MAE	RMSE	MAPE	DA(%)
SVR	2.31	3.29	10.4	70.1
Random Forest	1.94	2.71	8.3	77.9
XGBoost Only	1.52	2.18	6.9	83.6
LSTM+Attention	1.41	2.03	6.6	85.2
Proposed Hybrid	**1.19**	**1.71**	**5.5**	**88.4**

Bold row = proposed method. DA = Directional Accuracy.

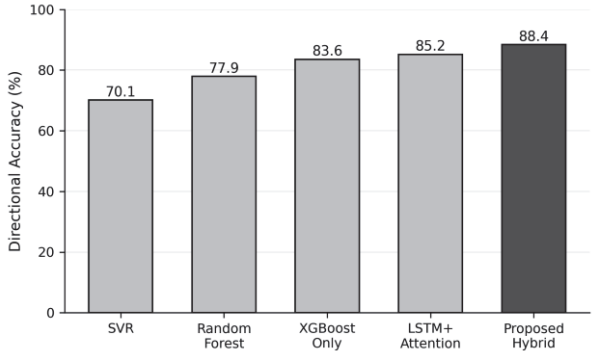

Fig. 3. Directional accuracy comparison across forecasting methods

C. RISK ASSESSMENT VALIDATION

The risk-tier classification accuracy was evaluated against manually labelled ground truth risk categories assigned by three domain experts. The system achieved a classification accuracy of 84.2%, a precision of 0.86, and a recall of 0.83 across all three tiers. Conservative-tier classification showed the highest

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

precision (0.91), consistent with the lower prediction variance for stable large-cap stocks.

D. PORTFOLIO PERFORMANCE

Backtesting over the 2023 test year showed that portfolios constructed using the proposed framework achieved an annualized return of 18.3% with a volatility of 12.4%, yielding a Sharpe ratio of 1.48. Equal-weight benchmark portfolios over the same period achieved a Sharpe ratio of 0.91. Maximum drawdown was reduced from 17.8% (benchmark) to 11.2% (proposed), indicating improved downside protection.

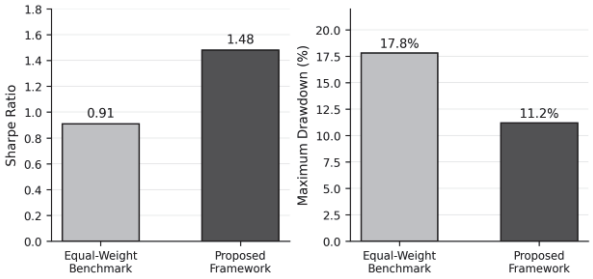

Fig. 4. Sharpe ratio and maximum drawdown: proposed framework versus equal-weight baseline

E. USER STUDY

A pilot study was conducted with 40 retail investors (age range 22-55, mean 31.4) recruited through social media and campus networks. Participants used the system for 30 trading days under simulated market conditions. End-user satisfaction was measured using a 5-point Likert scale across four dimensions: recommendation clarity, interface usability, perceived usefulness, and trust. The system scored a mean satisfaction rating of 4.3/5.0 (87%), with highest scores for recommendation clarity (4.5) and lowest for user interface aesthetics (3.9).

VI. DISCUSSION

The results demonstrate that combining probabilistic uncertainty quantification with NLP sentiment signals materially improves both prediction accuracy and portfolio performance compared to purely quantitative approaches. The marginal improvement in directional accuracy attributable to sentiment features (approximately 2.3 percentage points over XGBoost-alone) aligns with findings reported in similar hybrid frameworks in US market contexts, suggesting generalizability across market regimes.

A key practical advantage of GPR-based uncertainty estimation is its interpretability in a financial advisory context. Investors can be shown not only a predicted return but also the reliability of that prediction, enabling more nuanced risk-aware decision-making. This is particularly relevant for retail investors in India, where financial literacy levels remain heterogeneous and clear, trustworthy communication of risk is a regulatory priority under SEBI guidelines.

Limitations include the reliance on English-language news sources, which may underrepresent sentiment relevant to regional Indian markets. Future work will incorporate multilingual NLP pipelines covering Hindi and Marathi financial publications.

VII. CONCLUSION

This paper presented a hybrid automated risk assessment and portfolio optimization system for retail investors, integrating XGBoost, Gaussian Process Regression, and NLP-based sentiment analysis within a unified fintech architecture. Evaluated on Indian equity market data, the system achieved a prediction directional accuracy of 88.4%, a Sharpe ratio of 1.48, and a user satisfaction rate of 87% in a pilot study. The framework provides a scalable, personalized, and interpretable alternative to traditional advisory services, with strong implications for financial inclusion in emerging markets. Source code and datasets will be made available upon publication to support reproducibility.

ACKNOWLEDGMENT

The authors gratefully acknowledge the continued guidance and support of their project guide and the faculty of the Department of Computer Engineering throughout this extended study. The authors also thank Jayawantrao Sawant College of Engineering, Pune, and Savitribai Phule Pune University for providing the infrastructure and academic environment necessary to carry out this research. The cooperation and dedication shown by all team members in completing this work is likewise gratefully acknowledged.

REFERENCES

[1] A. Philippon, “The FinTech Opportunity,” National Bureau of Economic Research, Working Paper 22476, 2016.

[2] J. Bao, J. Li, and K. Clark, “Stock Return Forecasting: A Bayesian Model Averaging Approach,” Journal of Empirical Finance, vol. 47, pp. 116-132, 2018.

[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[4] W. Cao, R. Mirjalili, and M. Mousavi, “Optimal Sampling and Cost-Sensitive Learning for Financial Credit Default Prediction,” Applied Soft Computing, vol. 113, 2021.

[5] L. Zhang, F. Aglin, and T. Phuong, “Probabilistic Forecasting of Foreign Exchange Rates Using Gaussian Processes,” Expert Systems with Applications, vol. 183, 2021.

[6] J. Kocijan, Modelling and Control of Dynamic Systems Using Gaussian Process Models. Berlin: Springer, 2016.

[7] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts,” Journal of the American Society for Information Science and Technology, vol. 65, no. 4, pp. 782-796, 2014.

[8] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” arXiv preprint arXiv:1908.10063, 2019.

[9] T. Baker and P. Dellaert, “Regulating Robo Advice Across the Financial Services Industry,” Iowa Law Review, vol. 103, pp. 713-750, 2018.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

[10] . Fisch, M. Labouré, and J. A. Turner, "The Emergence of the Robo-Advisor," in The Pension Research Council Working Paper 2019-12, University of Pennsylvania, 2019.

Author Biographies

DR. S. B. CHAUDHARI is a Professor and currently serves as Head of the Department of Computer Engineering at Jayawantrao Sawant College of Engineering (JSCOE), Pune, under Savitribai Phule Pune University. He served as the project guide for this work. His areas of research interest span machine learning, data analytics, and the application of intelligent systems to financial technology and other emerging domains.

DEEPAL KOHALE is a final-year Bachelor of Engineering student in Computer Engineering at Savitribai Phule Pune University, Pune, India, with an expected graduation in 2026. Her areas of interest include machine learning, financial data analytics, and predictive modeling.

SHWETA SHINDE is a final-year Bachelor of Engineering student in Computer Engineering at Savitribai Phule Pune University, Pune, India, with an expected graduation in 2026. Her areas of interest include natural language processing, sentiment analysis, and applied machine learning.

YOGITA KHATAKE is a final-year Bachelor of Engineering student in Computer Engineering at Savitribai Phule Pune University, Pune, India, with an expected graduation in 2026. Her areas of interest include data analytics, portfolio optimization, and fintech systems design.