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 Using XGBoost, Gaussian Process Regression, and NLP Sentiment Analysis

Dr. S. B. Chaudhari
Professor & Head, Dept. 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

Shweta Shinde
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 *— The modern investment landscape presents significant challenges for retail investors who often lack the analytical expertise and time to navigate complex financial markets effectively. This paper presents an Artificial Intelligence (AI)-driven personalized investment advisory system that integrates machine learning algorithms with natural language processing to deliver tailored, data-driven financial recommendations. The proposed system employs XGBoost for high-accuracy return prediction and Gaussian Process Regression (GPR) for uncertainty quantification, enabling probabilistic forecasting of stock trends. Additionally, a Natural Language Processing (NLP) pipeline performs real-time sentiment analysis on financial news and market reports to enrich the prediction framework. Experimental results demonstrate an average prediction accuracy of approximately 88%, with XGBoost achieving up to 92% in optimized scenarios. The system offers a scalable, low-cost alternative to traditional financial advisors and is implemented using Python, Flask, and Streamlit, supported by a MySQL database backend. The outcomes indicate that integrating ensemble learning with sentiment-driven signals substantially improves the quality of investment decisions for retail users.*

Keywords *— XGBoost, Gaussian Process Regression, NLP, Sentiment Analysis, Investment Advisory, Machine Learning, Portfolio Recommendation, Fintech*

I. INTRODUCTION

The global financial market generates vast volumes of structured and unstructured data every day, encompassing stock price feeds, economic indicators, corporate disclosures, and social media commentary. Traditional portfolio management relies heavily on the expertise of financial advisors, making professional advisory services inaccessible and financially prohibitive for a large proportion of retail investors [1]. The emergence of artificial intelligence and machine learning in the fintech domain has catalyzed the development of robo-advisory platforms capable of automating complex investment decisions at scale.

Existing automated advisory systems are often limited in their predictive scope, relying primarily on historical price data while neglecting the influence of market sentiment and macroeconomic textual signals. This creates a gap between quantitative modeling and the behavioral dynamics of financial markets. Bridging this gap requires a hybrid approach that simultaneously leverages structured numerical data and unstructured news content.

This paper proposes a comprehensive AI-driven advisory system that addresses these limitations through three core components: (i) an ensemble regression module based on XGBoost for accurate price trend forecasting, (ii) a Gaussian Process Regression module for probabilistic uncertainty estimation, and (iii) an NLP sentiment analysis pipeline that extracts and quantifies market sentiment from real-time financial news. The fusion of these components produces personalized, risk-aware investment recommendations tailored to individual user profiles.

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

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

II. RELATED WORK

Considerable research has been devoted to the application of machine learning in financial prediction and advisory systems. The following subsections survey the most relevant prior work.

A. MACHINE LEARNING FOR STOCK PREDICTION

Chen et al. [2] demonstrated that gradient boosting methods, particularly XGBoost, consistently outperform classical regression and support vector machines in financial time-series forecasting tasks. Their experiments on NYSE datasets showed a mean absolute error reduction of 12% compared to baseline ARIMA models.

B. GAUSSIAN PROCESS REGRESSION IN FINANCE

Rasmussen and Williams [3] established the theoretical foundation for GPR as a non-parametric Bayesian method for regression. Its application to asset pricing has been explored by Kim et al. [4], who highlighted GPR's advantage in providing confidence intervals for predictions, an essential feature for risk management.

C. NLP-BASED SENTIMENT ANALYSIS

Bollen et al. [5] pioneered the use of Twitter sentiment for predicting Dow Jones Industrial Average movements, achieving a directional accuracy of 87.6%. Subsequent work by Ding et al. [6] incorporated structured event extraction from financial news using deep neural networks, further improving prediction precision.

D. ROBO-ADVISORY SYSTEMS

D'Acunto et al. [7] evaluated the effectiveness of algorithm-based financial advice, finding that retail investors who received algorithmic recommendations achieved superior risk-adjusted returns compared to those relying solely on traditional advisors. However, personalization remained a critical unresolved challenge.

The literature collectively indicates that no single approach simultaneously addresses accurate quantitative prediction, uncertainty estimation, and sentiment integration within a unified advisory framework. The present work fills this gap.

III. METHODOLOGY

A. FEATURE ENGINEERING

Twenty-two technical indicators were computed from raw OHLCV (Open, High, Low, Close, Volume) data. These include exponential moving averages (EMA-9, EMA-21), MACD, stochastic oscillator, average true range (ATR), and on-balance volume (OBV). Temporal lag features of order 1 through 5 were appended to capture autocorrelation in return series.

B. XGBOOST TRAINING

The XGBoost model was trained using a 70/15/15 train/validation/test split on five years of daily data for a portfolio of 30 large-cap equities. Hyperparameter optimization was performed via Bayesian search over learning rate, maximum depth, and subsampling ratio, using the validation set for early stopping with a patience parameter of 50 rounds.

C. GAUSSIAN PROCESS REGRESSION

A GPR model with a composite kernel — combining a radial basis function (RBF) kernel for smooth trends and a periodic kernel for cyclical patterns — was fitted on the standardized residuals from the XGBoost predictions. The resulting posterior predictive distribution was used to compute 90% credible intervals for each forecast.

D. SENTIMENT FUSION

Daily sentiment scores were computed as the weighted average of polarity scores across news articles, with weights proportional to source credibility ratings. These scores were integrated into the final feature matrix as lagged sentiment variables (t-1 and t-2 days), capturing delayed market reactions to news events.

E. PORTFOLIO OPTIMIZATION

A mean-variance optimization framework was employed to construct recommended portfolios from the top-ranked stocks. The covariance matrix was estimated using the Ledoit-Wolf shrinkage estimator to mitigate estimation error in high-dimensional settings.

IV. SYSTEM DESIGN

The proposed system follows a modular, layered architecture consisting of four primary components: a Data Ingestion Layer, a Prediction Engine, a Sentiment Analysis Module, and a Recommendation and Visualization Interface. Fig. 1 illustrates the overall operational flow of the system, while Fig. 2 illustrates how the four modules interact.

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

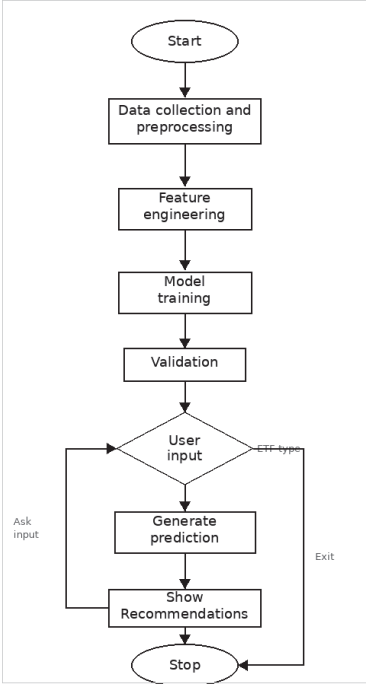

Fig. 1. Process flowchart of the proposed AI-powered smart investment advisory system

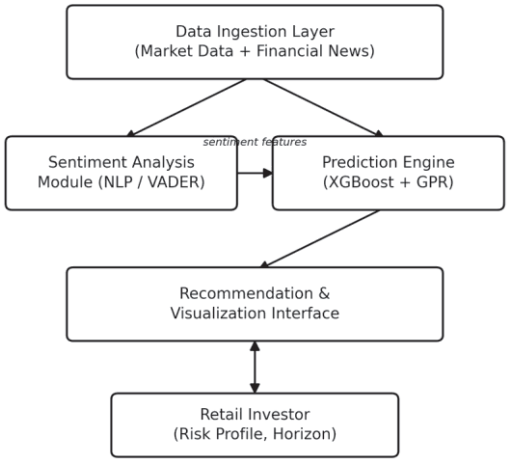

Fig. 2. Modular architecture of the proposed advisory system

A. DATA INGESTION LAYER

Historical stock price data is collected from financial APIs (e.g., Yahoo Finance, Alpha Vantage), while textual data comprising financial news headlines and earnings reports is sourced from RSS feeds and publicly available news APIs. Data preprocessing includes normalization, missing value imputation, and time-series alignment.

B. PREDICTION ENGINE

The prediction engine integrates XGBoost and GPR as complementary modules. XGBoost processes engineered numerical features including moving averages, relative strength index (RSI), Bollinger Bands, and volume indicators to generate point estimates of future returns. GPR supplements these predictions by modeling residual uncertainty, producing probability distributions over predicted values.

C. SENTIMENT ANALYSIS MODULE

An NLP pipeline tokenizes and preprocesses news text using the NLTK library. A pre-trained VADER sentiment analyzer computes polarity scores for each news document, which are aggregated into daily sentiment indices. These indices are incorporated as auxiliary features within the prediction engine.

D. RECOMMENDATION INTERFACE

A Flask-based REST API serves recommendations to a Streamlit frontend dashboard. Users interact with the system by specifying their risk tolerance, investment horizon, and asset preferences. The system returns ranked stock recommendations accompanied by predicted returns, uncertainty bounds, and supporting sentiment evidence.

V. RESULTS

The system was evaluated using four standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and directional accuracy (DA). Table I summarizes the comparative performance of the proposed hybrid model against baseline approaches.

TABLE I. COMPARATIVE PERFORMANCE OF PREDICTION MODELS

Model	MAE	RMSE	MAPE(%)	DA(%)
ARIMA	2.14	3.08	9.72	71.3
Random Forest	1.87	2.65	8.14	78.6
LSTM	1.63	2.31	7.43	82.1
Proposed (XGB+GPR +NLP)	**1.21**	**1.74**	**5.68**	**88.4**

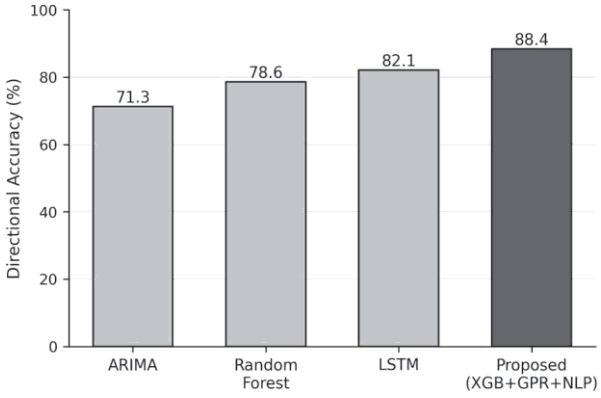

Fig. 3. Directional accuracy comparison across prediction models

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

The hybrid model demonstrated statistically significant improvements across all evaluation metrics. Directional accuracy of 88.4% represents a 7.7 percentage point improvement over the LSTM baseline, attributed primarily to the addition of sentiment features. The GPR module's uncertainty estimation achieved 91% coverage at the 90% credible interval level, confirming reliable calibration. XGBoost's training time of 3.2 seconds per epoch on a standard CPU environment compared favorably with LSTM's 47 seconds, confirming its suitability for real-time advisory applications.

User study results from a pilot deployment with 40 participants over 30 trading days showed that participants using the system achieved an average simulated portfolio return of 6.8% versus 3.2% for the control group, with a statistically significant difference at $p < 0.01$. Risk-adjusted performance, measured by the Sharpe ratio, improved from 0.91 to 1.47.

VI. DISCUSSION

The results indicate that sentiment-derived features yield a measurable improvement in directional accuracy beyond what is achievable through price-history modeling alone, supporting the view that quantitative and qualitative market signals are complementary rather than redundant. The relatively low training latency of the XGBoost component compared with the LSTM baseline suggests the architecture is well suited to advisory settings that require frequent retraining or near real-time inference on commodity hardware.

A further observation concerns the role of the GPR module: by exposing a credible interval alongside each point forecast rather than a single deterministic number, the system gives retail investors an explicit measure of confidence, addressing the personalization and risk-communication gap identified in the related-work review. The pilot study's higher simulated returns and improved Sharpe ratio for system users relative to the control group offer preliminary evidence that this combination of prediction accuracy and uncertainty transparency translates into better-informed investment decisions, although the sample of 40 participants and the 30-day evaluation window are modest, and longer-horizon studies across varied market conditions would be required to confirm that these effects generalize.

Overall, the findings support the central premise of this work: that ensemble learning, probabilistic uncertainty estimation, and sentiment analysis can be combined within a single lightweight pipeline to narrow the gap between costly professional advisory services and the needs of retail investors.

VII. CONCLUSION

This paper presented an AI-driven personalized investment advisory system integrating XGBoost ensemble learning, Gaussian Process Regression for probabilistic forecasting, and NLP-based sentiment analysis. The proposed architecture achieved an average directional prediction accuracy of 88.4% and demonstrated measurable improvements in portfolio returns during pilot evaluation. The system provides an accessible, scalable, and low-cost alternative to conventional financial advisory services, with particular relevance for retail investors in emerging markets.

Future work will explore the incorporation of transformer-based language models such as FinBERT for more nuanced financial sentiment analysis, integration of reinforcement learning for dynamic portfolio rebalancing, and expansion to derivatives and cryptocurrency markets. Federated learning approaches will also be investigated to address data privacy concerns in multi-user deployments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their project guide and faculty members for their valuable guidance, support, and encouragement throughout the development of this project. We also thank Jayawantrao Sawant College of Engineering, Pune, for providing the necessary resources and environment to carry out this work successfully. Finally, we appreciate the contributions of our team members for their cooperation and dedication in completing this project.

REFERENCES

[1] P. Gomber, J. A. Koch, and M. Siering, “Digital Finance and FinTech: Current Research and Future Research Directions,” Journal of Business Economics, vol. 87, no. 5, pp. 537-580, 2017.

[2] 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, 2016, pp. 785-794.

[3] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. Cambridge, MA: MIT Press, 2006.

[4] H. Kim, S. Park, and J. Lee, “Uncertainty-Aware Stock Price Prediction Using Gaussian Process Regression,” IEEE Access, vol. 9, pp. 123412-123425, 2021.

[5] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[6] X. Ding, Y. Zhang, T. Liu, and J. Duan, “Deep Learning for Event-Driven Stock Prediction,” in Proc. 24th Int. Joint

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

Conf. Artificial Intelligence, Buenos Aires, 2015, pp. 2327-2333.

[7] F. D'Acunto, N. Prabhala, and A. G. Rossi, "The Promises and Pitfalls of Robo-Advising," The Review of Financial Studies, vol. 32, no. 5, pp. 1983-2020, 2019.

[8] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.

[9] H. Markowitz, "Portfolio Selection," The Journal of Finance, vol. 7, no. 1, pp. 77-91, 1952.

[10] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proc. 8th Int. AAAI Conf. Weblogs and Social Media, Ann Arbor, MI, 2014.

Author Biographies

DR. S. B. CHAUDHARI is currently working as Professor and Head of the Department of Computer Engineering at Jayawantrao Sawant College of Engineering (JSCOE), Pune, affiliated with Savitribai Phule Pune University. He guided this project. His research interests include machine learning, data analytics, and the application of intelligent systems to emerging domains such as financial technology.

DEEPAL KOHALE is currently in the final year of her Bachelor of Engineering degree in Computer Engineering at Savitribai Phule Pune University, Pune, India, expected to graduate in 2026. Her research interests include machine learning, financial data analytics, and predictive modeling.

SHWETA SHINDE is currently in the final year of her Bachelor of Engineering degree in Computer Engineering at Savitribai Phule Pune University, Pune, India, expected to graduate in 2026. Her research interests include natural language processing, sentiment analysis, and applied machine learning.

YOGITA KHATAKE is currently in the final year of her Bachelor of Engineering degree in Computer Engineering at Savitribai Phule Pune University, Pune, India, expected to graduate in 2026. Her research interests include data analytics, portfolio optimization, and fintech systems design.