

An Intelligent Stock Market Recommendation System Based on Financial Behaviour and Investor Profiling

Ms. Rakshita Sanjay Mutha¹ and Ms. Asha Mane²

Sadhu Vaswani Institute of Management Studies for Girls, Pune

Savitribai Phule Pune University, Pune

rakshitamutha18@gmail.com and ashamane90@gmail.com

Abstract: Choosing suitable stocks can be difficult for individual investors, especially those without professional financial knowledge. Many stock-analysis approaches focus on market prices, technical indicators or company fundamentals, while an investor's financial behaviour, risk tolerance and investment horizon may receive less attention. This paper presents an intelligent stock market recommendation framework that connects investor profiling with machine-learning-based analysis. The project materials describe a tabular dataset containing 16,203 records and 36 columns. The workflow includes data inspection, exploratory data analysis, feature preparation, investor-profile classification and recommendation analysis. The project materials also report an Extra Trees result of 77.4% accuracy and a weighted F1-score of 0.748 using an 80:20 train-test split. These figures are treated as reported project results and must be reproduced from the final implementation before being considered verified. The dataset has an uneven distribution of investor profiles and recommendation labels, making per-class evaluation important. The framework is intended as decision support, not as a guarantee of future stock performance or investment returns.

Keywords: stock recommendation, investor profiling, financial behaviour, machine learning, Random Forest, Gradient Boosting, risk tolerance, personalized recommendation

I. INTRODUCTION

Online trading platforms have made stock-market information easier to access, but having more information does not always make investment decisions easier. Investors differ in financial capacity, experience, preferred investment period and willingness to accept losses. A stock that may be suitable for one person may not be appropriate for another. For example, a conservative investor may prefer lower volatility, while an aggressive investor may accept greater fluctuations in pursuit of higher potential returns.

Machine learning can help identify patterns in financial data and classify records into meaningful categories. However, predicting a stock or market movement is not the same as making a recommendation tailored to a particular person. This project brings investor profiling and stock recommendation into a common workflow. Its intended users include individual and non-professional investors who may benefit from clearer, profile-aware decision support.

II. LITERATURE REVIEW AND RESEARCH GAP

Earlier research has explored machine learning for stock forecasting, sentiment analysis, financial risk assessment and recommendation systems. Studies of investor sentiment show that opinions and information from news or social networks can add context to market analysis. Other work examines risk profiling and investor preferences, while stock-prediction studies compare algorithms using historical financial data. Research on personalized recommendation also suggests that users' preferences and characteristics can influence the usefulness of suggested options.



These research areas are relevant, but they are not always brought together in one practical workflow. A model may predict a market pattern without considering the investor's risk tolerance, or a profiling tool may classify an investor without explaining how that profile affects the stock recommendation. The research gap addressed here is the need for a straightforward framework linking investor characteristics and financial behaviour with stock-related features and recommendation categories. This gap is the interpretation adopted for this project and should be supported by the complete, verified literature review in the final submission.

III. OBJECTIVES

- To examine investor-related and stock-related features available in the project dataset.
- To explore the distribution of investor profiles and recommendation labels through exploratory data analysis.
- To investigate machine-learning classifiers for investor-profile classification.
- To connect investor profiles with recommendation categories in an understandable way.
- To discuss model performance carefully, including class imbalance, data leakage and the limits of classification scores as evidence of investment success.

IV. RESEARCH METHODOLOGY

The study follows an applied, quantitative machine-learning approach. It begins with inspection of the supplied tabular dataset, followed by data-quality checks and exploratory analysis. Suitable input variables are then identified, numerical and categorical features are prepared, classifiers are trained and their outputs are evaluated. The exact feature list, preprocessing steps and model settings must match the final source code.

4.1 Dataset

The project report and presentation describe a dataset of 16,203 records and 36 columns. It contains investor-profile information and fields related to financial behaviour and stock recommendations. The project materials report no missing cells and no exact duplicate rows. These figures should be checked against the exact CSV used for training. Complete cells do not guarantee correct values, representative sampling or reliable target labels.

4.2 Data Preparation and Exploratory Analysis

Exploratory data analysis helps show how investor profiles and recommendation labels are distributed. Numerical summaries and correlation plots can reveal ranges and relationships among financial variables. During preprocessing, categorical variables may be encoded and numerical fields prepared for the chosen model. Preprocessing steps that learn from the data should be fitted on the training set only and then applied to the test set to reduce data leakage.

4.3 Machine-Learning Models

Random Forest combines predictions from multiple decision trees and can represent non-linear relationships. Gradient Boosting builds learners in sequence, with later learners attempting to correct earlier errors. Both are candidate methods for structured tabular data. The project materials also report an Extra Trees classifier result. The final paper should distinguish models that were actually trained from models that were only proposed, and compare scores only when they come from the same evaluation setup.

4.4 General Workflow

- Load the dataset and inspect records, columns, data types and missing values.
- Check duplicates, invalid values and the meaning of target labels.
- Explore profile and recommendation distributions and visualize numerical features.
- Separate the target from the input features and remove fields that could cause target leakage.
- Split the data into training and testing sets, preserving class proportions where appropriate.
- Fit preprocessing and train the selected model or models on training data.
- Generate predictions for held-out test records and calculate evaluation metrics.
- Review the confusion matrix and per-class precision, recall and F1-score before interpreting results.



V. EVALUATION MEASURES

Accuracy measures the proportion of correct predictions. Precision indicates how many records predicted as a class truly belong to that class, while recall shows how many actual records of the class were identified. The F1-score balances precision and recall. A confusion matrix shows correct and incorrect classifications for each class. When categories are unevenly represented, macro F1 and per-class recall should be reviewed alongside accuracy and weighted F1 because an overall score can hide weak performance on a smaller class.

VI. RESULTS AND DISCUSSION

The supplied project materials report 16,203 records across 36 columns. They report 12,003 Conservative investor-profile records (74.1%), 3,394 Moderate records (20.9%) and 806 Aggressive records (5.0%). This distribution is imbalanced, with Conservative profiles forming the largest group. It describes this dataset and should not be treated as proof that the same proportions apply to all real-world investors.

The recommendation labels are reported as 7,609 SELL records (47.0%), 7,102 HOLD records (43.8%) and 1,492 BUY records (9.2%). BUY is the smallest category, so its precision and recall should be inspected separately if recommendation classification is evaluated. The label distribution does not establish that SELL is the best market action; it only describes the labels present in the dataset.

The project presentation reports an Extra Trees accuracy of 77.4% and weighted F1-score of 0.748 using an 80:20 train-test split. These are reported figures rather than independently reproduced results. Before final submission, they should be confirmed from the final code, test predictions and confusion matrix. The report should state which target was predicted, which features were used and whether any variables directly or indirectly encoded the target. Without these checks, the score cannot be interpreted as proof of reliable investment recommendations.

The preliminary analysis highlights the importance of class balance, transparent feature selection and careful evaluation. Classification performance does not prove that recommendations would earn positive returns in a live market; that claim would require a separate time-based backtest and risk-adjusted performance analysis.

6.1 Visual Results and Application Screenshots



Figure 1. Distribution of investor and recommendation categories.

This plot compares category counts in the dataset. It helps show that the labels are not evenly represented, so overall accuracy should be interpreted alongside per-class metrics.



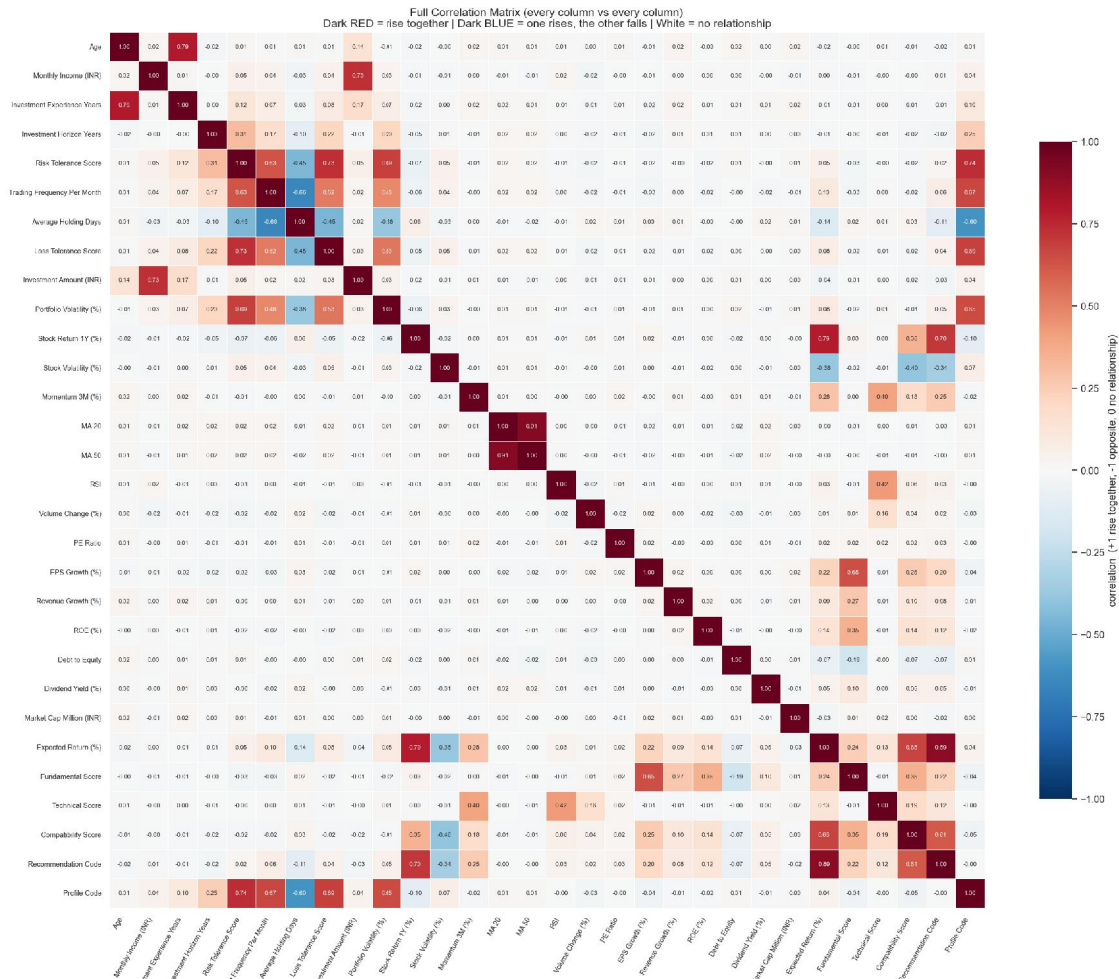


Figure 2. Correlation heatmap of selected numerical features.

The heatmap visualises pairwise relationships among numerical variables. Correlation can help identify related features, but it does not prove causation and should not replace domain-based feature checks.

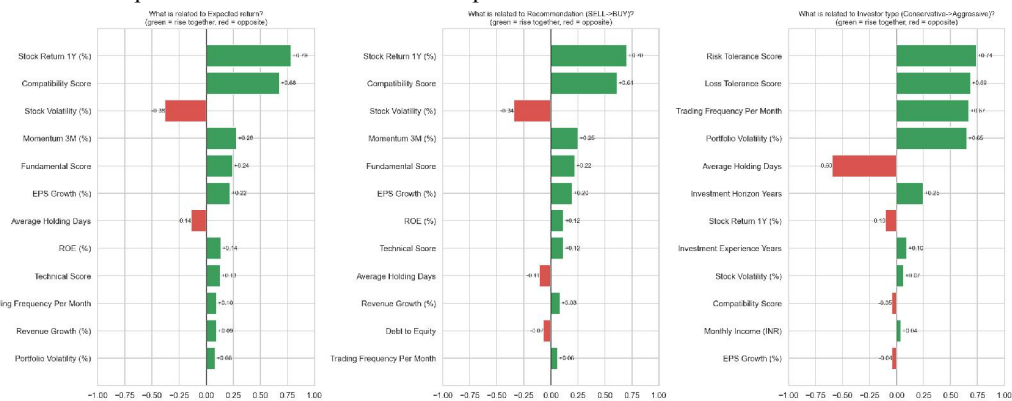


Figure 3. Feature-importance visualisation.



The feature-importance chart summarises which inputs contributed most to the fitted model according to its importance measure. Importance is model-specific and does not by itself establish a causal effect.

Confusion matrix (counts) - Gradient Boosting



Figure 4. Confusion matrix for investor-profile classification.

The confusion matrix compares actual and predicted investor-profile classes. It helps identify which profiles are classified correctly and where errors occur, including errors involving smaller classes.

Confusion matrix (counts) - Gradient Boosting



Figure 5. Confusion matrix for recommendation classification.

This matrix compares predicted and actual BUY, HOLD and SELL labels. It provides more detail than a single accuracy value, especially for the less frequent BUY class.



Recommendation distribution

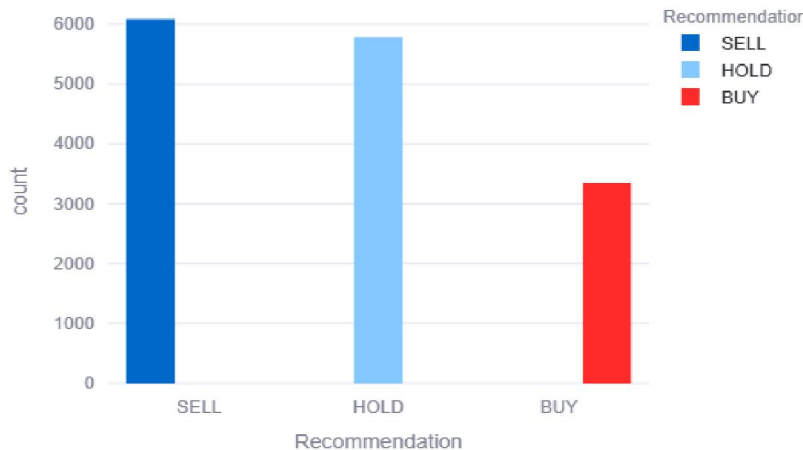
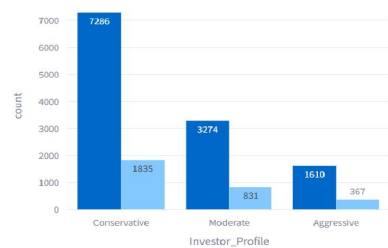


Figure 6. Recommendation-label distribution.

The chart shows the frequency of BUY, HOLD and SELL labels represented in the module data. These frequencies describe the dataset labels and should not be interpreted as current investment advice.

Records	Unique investors	Stocks	Columns
15,203	15,058	12	37

Investor profile: train vs test



Recommendation distribution

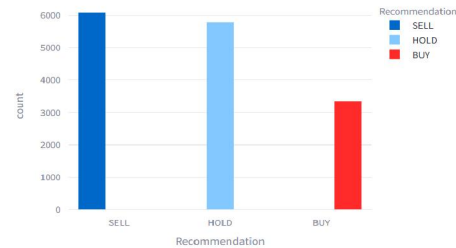


Figure 7. Stock recommender dashboard overview.

The dashboard screenshot presents summary counts and visual summaries from the application. It gives users a quick overview of profile and recommendation information before exploring an individual prediction.



Manual Prediction

All of the model's input parameters are here, grouped into sections. Default = dataset median; ranges come from the dataset. Change the values and press Predict.

1. Investor Profile 2. Expected Stock Return (%) 3. Stock Recommendation 4. ★ Top-5 Stocks for Me

Model: Gradient Boosting | **Sabse important inputs:** Loss Tolerance Score, Trading Frequency Per Month, Risk Tolerance Score, Portfolio Volatility (%), Average Holding Days

The model was trained on a dataset where Conservative, Moderate and Aggressive investors have equal row counts.

Personal

Age: 34 Monthly Income (INR): 52100.00 Investment Experience Years: 4 Investment Amount (INR): 15016

Goals & horizon

Investment Horizon: Medium Term Investment Horizon Years: 5 Investment Goal: Capital Growth Preferred Sector: Infrastructure

Risk behaviour

Risk Tolerance: Medium Risk Tolerance Score: 56.80 Loss Tolerance Score: 45.30 Portfolio Volatility (%): 12.89

Figure 8. Manual prediction input screen.

The manual prediction form lets a user enter investor-related information and financial preferences. These inputs form part of the application workflow that connects investor characteristics with the model output.

Predict

Predicted expected return

9.45 %

Figure 9. Predicted expected-return output.

This screen displays an expected-return estimate produced by the module for the selected inputs. It should be treated as a model estimate, not a guaranteed return or a substitute for professional financial advice.

Manual Prediction

All of the model's input parameters are here, grouped into sections. Default = dataset median; ranges come from the dataset. Change the values and press Predict.

1. Investor Profile 2. Expected Stock Return (%) 3. Stock Recommendation 4. ★ Top-5 Stocks for Me

★ **Top-5 stock recommendations for your risk & behaviour**

Fill in the investor's risk, behaviour and goal. The trained models first read the investor's type, then every stock is scored for THIS investor and the 5 best matches are shown.

Personal

Age: 34 Monthly Income (INR): 52100.00 Investment Experience Years: 4 Investment Amount (INR): 15016

Goals & horizon

Investment Horizon: Medium Term Investment Horizon Years: 5 Investment Goal: Capital Growth Preferred Sector: Infrastructure

Risk behaviour

Risk Tolerance: Medium Risk Tolerance Score: 56.80 Loss Tolerance Score: 45.30 Portfolio Volatility (%): 12.89

Figure 10. Top-five stock recommendations for a profile.

The recommendation screen lists stocks selected for the entered investor preferences and risk/behaviour information. Suitability scores and model outputs depend on the data and rules implemented in the module.



 Your Top-5 ⇌

Rank	Stock	Sector	Suitability (0-100)	Model says	P(BUY)	Expected return %	Volatility %
1	Axis Bank	Banking	79.2	HOLD	0.02	9.42	17.1
2	ITC	FMCG	76.0	HOLD	0.02	9.34	17.18
3	Maruti Suzuki	Automobile	56.1	HOLD	0.02	9.21	17.84
4	HCL Tech	IT	50.6	HOLD	0.02	9.11	17.61
5	HDFC Bank	Banking	47.0	HOLD	0.02	9.12	17.36

Figure 11. Top-five recommendation table.

The table compares recommended stocks using fields such as sector, suitability, model output, expected return and volatility. It helps users compare options, but estimates should be validated against current market data before any real investment decision.

VII. MAJOR FINDINGS

Investor-profile labels are unevenly distributed, with Conservative profiles forming approximately three-quarters of the reported records.

SELL and HOLD together account for most reported recommendation labels, while BUY is a minority category.

Accuracy and weighted F1 alone are not sufficient when classes are imbalanced.

Investor profiling can provide useful context, but the relationship between a profile and a specific stock should be explicitly defined and evaluated.

The system should be presented as decision support; it does not guarantee returns or replace independent financial judgement.

VIII. LIMITATIONS

The conclusions depend on the dataset's source, representativeness, label-generation rules and selected features. Class imbalance may affect model performance. The reported Extra Trees metrics have not been independently reproduced within this paper. If the dataset is synthetic or uses labels generated from predefined rules, the model may learn those rules rather than real market relationships. This paper does not establish live trading performance, causality or guaranteed profitability. Market outcomes are affected by changing economic conditions and unexpected events that may not be represented in the data.

IX. FUTURE SCOPE

Future work can test the model on a larger and more representative dataset, use chronological train-validation-test splits and conduct historical backtesting with transaction costs. The system could incorporate updated market prices, company fundamentals and carefully sourced financial-news sentiment. Investor profiles could be refreshed when goals or financial circumstances change. Explainability methods and a simple interface may help users understand why a recommendation was produced. These extensions should be evaluated for accuracy, fairness, stability and risk.

X. CONCLUSION

This paper presents a practical direction for an intelligent stock market recommendation system that considers investor profiling and financial behaviour together. Its central idea is that stock analysis becomes more relevant when considered alongside the needs and risk preferences of the person making the investment. The supplied project materials provide a starting point through exploratory analysis and machine-learning classification. However, the imbalanced labels and the need to reproduce reported model scores show why careful validation is essential. The



system should be understood as a decision-support framework, not a promise of future returns. Verified evaluation, transparent recommendation rules and testing on time-separated market data would help develop it into a more reliable and understandable tool for individual investors.

REFERENCES

1. Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156, Article 113464. <https://doi.org/10.1016/j.eswa.2020.113464>
2. Cavezzali, E., & Rigoni, U. (2012). Know your client! Investor profile and tailor-made asset allocation recommendations. *Journal of Financial Research*, 35(1), 137–158. <https://doi.org/10.1111/j.1475-6803.2011.01312.x>
3. Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, 197, Article 116659. <https://doi.org/10.1016/j.eswa.2022.116659>
4. Nazareth, N., & Ramana Reddy, Y. V. (2023). Financial applications of machine learning: A literature review. *Expert Systems with Applications*, 219, Article 119640. <https://doi.org/10.1016/j.eswa.2023.119640>
5. Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, 144(1), 273–297. <https://doi.org/10.1016/j.jfineco.2021.06.002>
6. Shah, J., Vaidya, D., & Shah, M. (2022). A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications*, 16, Article 200111. <https://doi.org/10.1016/j.iswa.2022.200111>
7. Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: A systematic review. *Journal of Physics: Conference Series*, 2161, Article 012065. <https://doi.org/10.1088/1742-6596/2161/1/012065>
8. Qin, C., Chang, J., Tu, W., & Yu, C. (2024). FollowAKOInvestor: Stock recommendation by hearing voices from all kinds of investors with machine learning. *Expert Systems with Applications*. Verify volume, article number and DOI from the publisher before submission.

