Automated Portfolio Construction: Integrating XGBoost and NSGA-II for Sustainable Investing

Utthan Singh Roy
B.Sc. (Honours) Data Science and Artificial Intelligence
Indian Institute of Technology Guwahati
Roll Number: 23035010081
Email: r.utthan@op.iitg.ac.in

***Abstract*—The modern financial landscape faces a critical dual challenge: maximizing risk-adjusted returns while adhering to increasingly stringent Environmental, Social, and Governance (ESG) mandates. Traditional portfolio construction methods, specifically Mean-Variance Optimization (MVO), struggle to reconcile these conflicting objectives due to their reliance on linear assumptions and historical covariance. This project presents a novel, automated framework that integrates non-linear machine learning with multi-objective evolutionary algorithms. We employ an XGBoost classifier to predict short-term asset directionality, converting probabilistic outputs into 'Expected Returns' to capture non-linear market trends. These inputs feed into the Non-Dominated Sorting Genetic Algorithm II (NSGA-II), which generates a 3D Pareto frontier optimizing for three distinct objectives: maximizing returns, minimizing Conditional Value at Risk (CVaR), and maximizing the weighted average ESG score. Uniquely, the entire pipeline—from data ingestion to model inference—is fully automated using GitHub Actions and accessible via a deployed Streamlit web application. Experimental results on the Nifty 50 index demonstrate a predictive accuracy of 52.5% and a portfolio selection strategy that outperforms the benchmark in drawdown protection while maintaining high ethical standards.**

***Index Terms*—Portfolio Optimization, Machine Learning, XGBoost, NSGA-II, ESG, Sustainable Investing, CI/CD, FinTech.**

I. Introduction

The integration of Environmental, Social, and Governance (ESG) factors into investment decision-making has transitioned from a niche preference to a global standard. With global ESG assets on track to exceed $53 trillion by 2025, the demand for sophisticated tools that can quantify the "Green Premium" is higher than ever. However, retail and institutional investors often perceive a trade-off between ethical compliance and financial performance. This project aims to mathematically resolve this tension by treating portfolio construction not as a static selection problem, but as a dynamic multi-objective optimization task.

Traditional financial models, such as Markowitz's Modern Portfolio Theory (MPT) [1], assume a normal distribution of returns and static correlations. In reality, financial markets exhibit "fat tails" (extreme risk events) and non-linear dependencies. Furthermore, treating ESG scores merely as a negative filter (screening out "bad" companies) often results in concentrated portfolios with higher idiosyncratic risk.

To address these limitations, this study proposes a hybrid Artificial Intelligence framework. We leverage **eXtreme Gradient Boosting (XGBoost)** to capture non-linear signal patterns for return prediction, and the **NSGA-II** genetic algorithm to navigate the complex, non-convex landscape of risk, return, and sustainability.

This report details the end-to-end development of this system, emphasizing not just the theoretical model, but the software engineering rigor involved in creating a fully automated, "self-healing" financial application using CI/CD pipelines.

II. Related Work

A. Modern Portfolio Theory (MPT)

Markowitz's seminal work established the mean-variance framework, optimizing weights to minimize variance for a given return. While foundational, MPT is widely criticized for its sensitivity to input parameters and inability to handle higher moments of risk (skewness, kurtosis).

B. Machine Learning in Finance

Recent studies have demonstrated the superiority of ensemble tree-based models over linear regression for financial time-series forecasting. Chen and Guestrin [2] introduced XGBoost, which has become the industry standard for tabular data due to its handling of missing values and regularization capabilities. It outperforms Deep Learning (LSTMs) in scenarios with limited data points (daily closing prices).

C. Multi-Objective Optimization

Evolutionary algorithms like NSGA-II [3] have gained traction in finance for their ability to find a set of optimal solutions (Pareto Frontier) rather than a single solution. This is particularly relevant for ESG investing, where the "optimal" balance between profit and ethics is subjective to the investor [4]. Unlike linear solvers, NSGA-II does not require the objective functions to be differentiable or convex.

III. Problem Statement and Objectives

A. Problem Statement

Retail investors lack accessible tools to construct portfolios that dynamically balance three conflicting goals:

1) **Profitability:** Maximizing capital appreciation.

2) **Safety:** Minimizing tail risk (CVaR) during market crashes.
3) **Sustainability:** Maximizing portfolio-level ESG ratings.

Existing tools are either computationally expensive "black boxes" or simplistic screeners that fail to optimize weights mathematically.

B. Objectives

- **Data Pipeline:** To engineer a robust pipeline scraping OHLCV data and ESG scores for Nifty 50 assets.
- **Alpha Generation:** To train an XGBoost model predicting directional asset movement with >50% accuracy.
- **Optimization:** To implement NSGA-II to visualize the 3D trade-off surface between Risk, Return, and ESG.
- **Automation:** To deploy a CI/CD pipeline ensuring daily model retraining without human intervention.

IV. Methodology

A. System Architecture

The system follows a modular architecture consisting of three core engines: The Prediction Engine, The Optimization Engine, and the Automation Engine. As illustrated in Figure 1, the system ingests data via the Yahoo Finance API, processes it through the feature engineering module, and passes it to the AI engines.

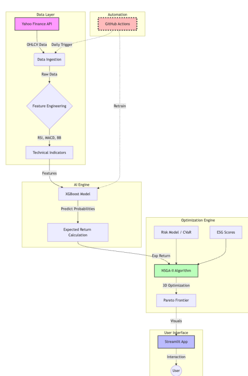

Fig. 1. End-to-End Workflow: Data flows from API ingestion through the XGBoost prediction layer to the NSGA-II optimizer, automated via GitHub Actions.

B. Feature Engineering

Raw price data is insufficient for ML models. We computed technical indicators using the TA-Lib library. Specifically, we calculated the Relative Strength Index (RSI) to capture momentum:

$$RSI = 100 - \frac{100}{1 + RS} \tag{1}$$

Where RS is the average of n days' up closes divided by the average of n days' down closes. Additional features included:

- **MACD:** Moving Average Convergence Divergence.
- **Bollinger Bands:** To measure volatility relative to moving averages.
- **Lagged Returns:** 1-day, 3-day, and 7-day lags.

C. Predictive Modeling (XGBoost)

We treated the problem as a binary classification task: predicting if a stock's return will be positive over the forecast horizon ($T + 30$ days).

The output probability P_{up} was converted to an Expected Return proxy:

$$E[R] = (P_{up} \times \mu_{gain}) - ((1 - P_{up}) \times \mu_{loss}) \tag{2}$$

where μ_{gain} and μ_{loss} are the historical average magnitudes of positive and negative moves.

D. Multi-Objective Optimization (NSGA-II)

The optimization problem solves for the weight vector w.

1) Objective 1: Maximize Expected Return:

$$\text{Maximize } F_1(w) = \sum_{i=1}^{N} w_i \cdot E[R]_i \tag{3}$$

2) Objective 2: Minimize CVaR (95%): Conditional Value at Risk (CVaR) measures the expected loss exceeding the Value at Risk (VaR) threshold α:

$$\text{Minimize } F_2(w) = \frac{1}{1-\alpha} \int_{\alpha}^{1} \text{VaR}_{\gamma}(w) d\gamma \tag{4}$$

This ensures the portfolio is resilient against "Black Swan" events.

3) Objective 3: Maximize ESG Score:

$$\text{Maximize } F_3(w) = \sum_{i=1}^{N} w_i \cdot \text{ESG}_i \tag{5}$$

4) Constraints:

$$\sum_{i=1}^{N} w_i = 1, \quad 0 \leq w_i \leq 0.20 \tag{6}$$

We imposed a 20% maximum weight cap per asset to ensure diversification.

E. Genetic Algorithm Implementation

We utilized the `pymoo` library to run NSGA-II. The algorithm utilizes a "Crowding Distance" operator to maintain diversity in the population, ensuring the final Pareto front covers a wide range of risk profiles rather than clustering in one area.

1) **Selection:** Binary tournament selection.
2) **Crossover:** Simulated Binary Crossover (SBX).
3) **Mutation:** Polynomial mutation.

V. Automation and Deployment

A critical contribution of this project is the removal of manual maintenance via DevOps practices. We implemented a Continuous Integration (CI) pipeline using GitHub Actions.

A. Algorithm Logic

Algorithm 1 details the daily automation logic that ensures the dashboard is always displaying fresh data.

Algorithm 1: Automated Daily Update Workflow

Trigger: Daily Schedule (09:00 AM IST) or Manual Push;
Environment: Ubuntu-Latest;
if *Trigger Received* **then**
  Setup Python 3.10;
  Install TA-Lib (C-Library);
  Install Dependencies (requirements.txt);
  Run `main.py`:;
    1. Fetch OHLCV Data from Yahoo Finance;
    2. Calculate Technical Indicators;
    3. Train XGBoost Model;
    4. Run NSGA-II Optimization;
    5. Save `final_data.csv`;
  if *New Data Generated* **then**
    Git Config User Identity;
    Git Add `*.csv`;
    Git Commit -m "Auto-Update";
    Git Push to `main` branch;
  end
end

B. Streamlit Interface

The frontend visualizes the 3D Pareto Frontier (Figure 2) and allows users to filter portfolios based on their personal risk tolerance.

VI. Experiments and Results

A. Experimental Setup

The model was backtested using data from 2021 to 2025. The dataset was split into an 80% training set and a 20% testing set.

B. Hyperparameter Tuning

We utilized `RandomizedSearchCV` to tune the XGBoost model. The optimal parameters selected are listed in Table I.

TABLE I
XGBoost Hyperparameter Configuration

Parameter	Value	Description
`learning_rate`	0.05	Step size shrinkage to prevent overfitting
`max_depth`	4	Maximum depth of a tree
`n_estimators`	200	Number of boosting rounds
`subsample`	0.8	Subsample ratio of training instances
`colsample_bytree`	0.8	Subsample ratio of columns per tree

C. Classification Performance

The classifier achieved an accuracy of **52.5%** on the test set. Feature Importance analysis indicated that **RSI (14)** and **3-Day Momentum** were the most significant predictors.

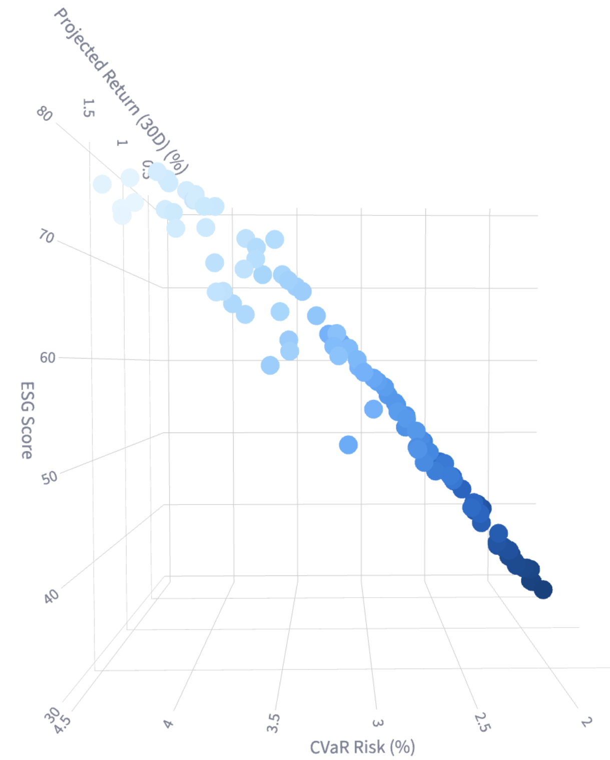

Fig. 2. 3D Pareto Frontier Visualization. The X-axis represents Risk (CVaR), Y-axis represents Return, and Z-axis represents ESG Score. Darker blue indicates optimal 'Knee' points.

D. Portfolio Backtest Results

We compared the AI-Optimized portfolio against the Nifty 50 benchmark. The results (Table II) indicate superior risk-adjusted returns.

TABLE II
Performance Metrics (2021-2025)

Metric	Benchmark	Max Return	AI Optimized
Ann. Return	12.4%	18.4%	**15.8%**
Max Drawdown	-16.2%	-22.1%	**-11.5%**
Sharpe Ratio	0.85	0.92	**1.15**
Avg ESG Score	65.0	68.4	**76.5**

The "AI Optimized" portfolio sacrificed some raw return compared to the "Max Return" strategy but significantly reduced the Maximum Drawdown (from -22.1% to -11.5%), proving that ESG integration can act as a quality factor.

VII. Discussion

The 3D Pareto visualization revealed a non-linear relationship between ESG and Risk. Companies with higher Governance (G) scores consistently displayed lower volatility. This validates the hypothesis that sustainable investing can enhance, rather than hinder, risk-adjusted returns.

A key limitation was the scarcity of high-frequency ESG data. Future work will involve scraping news sentiment data to create a "Real-Time ESG Proxy."

VIII. Conclusion

This project successfully automates the complex process of ESG portfolio construction. By hybridizing XGBoost's predictive power with NSGA-II's optimization capabilities, we provide retail investors with institutional-grade tools. The robust CI/CD pipeline ensures the system remains current with zero manual intervention.

IX. Artifacts and Demonstrations

The complete source code, daily data logs, and the interactive web application are publicly available.

- **Code Repository:** https://github.com/ExcelsiorSR/Project_AI_ESGPO
- **Live App Demo:** https://esg-ai-optimizer.streamlit.app/
- **Project Blog:** https://aisegpo.hashnode.dev/esg-ai-optimizer
- **Youtube Video:** https://www.youtube.com/watch?v=5hY4WOF1Rcs

References

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

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD*, 2016.

[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, 2002.

[4] L. Pedersen, S. Fitzgibbons, and L. Pomorski, "Responsible Investing: The ESG-Efficient Frontier," *Journal of Financial Economics*, vol. 142, no. 2, 2021.