Portfolio Optimization Model for Indian Market Using Modern Portfolio Theory

Ramish Jamal
IIIT Delhi
ramish22395@iiitd.ac.in

Mohd Masood
IIIT Delhi
masood22299@iiitd.ac.in

Akshat Karnwal
IIIT Delhi
akshat22052@iiitd.ac.in

Keshav Chhabra
IIIT Delhi
keshav22247@iiitd.ac.in

I. Motivation

The Indian financial market presents unique opportunities and challenges, necessitating sophisticated strategies for investors. The motivation for this project can be summarized in the following points:

- The Indian financial market offers distinct opportunities and challenges, requiring advanced investment strategies.
- The project aims to develop a portfolio optimization model combining Modern Portfolio Theory (MPT) with machine learning techniques specifically tailored for the Indian market.
- A critical aspect of this project is the creation of a predictive model to estimate stock price volatility over a 90 and 180 trading day span, providing valuable insights into future trends.
- This model will facilitate diversification across large-cap, mid-cap, and small-cap stocks, as well as various sectors of the Indian stock market.
- The ultimate goal is to empower investors with actionable insights and optimized strategies that enhance risk management while maximizing returns.
- Our primary objective is to predict stock price volatility (a measure of risk) to create an optimized portfolio.

***Index Terms*—Portfolio optimization, machine learning, stock price volatility, modern portfolio theory, financial market**

II. Introduction

The Indian financial market is rapidly evolving, presenting unique opportunities and challenges for investors. This project aims to develop a portfolio optimization model that leverages Modern Portfolio Theory (MPT) combined with machine learning techniques specifically designed for the Indian market. By predicting stock price volatility over different time frames (90 and 180 trading days), we seek to empower investors with actionable insights that optimize their portfolios while effectively balancing risk and returns.

III. Literature Survey

Previous studies have shown that machine learning significantly improves stock price predictions by identifying hidden patterns and non-linear relationships. Traditional methods, such as technical and qualitative analysis, often miss such patterns. Research on models like Artificial Neural Networks (ANN) and Random Forest (RF) has demonstrated better accuracy in predicting stock trends. These models have been applied to data from different sectors, making them an appropriate choice for this project. We have reviewed two papers that discuss the intricacies of applying models on the stock market for prediction tasks.

1st paper

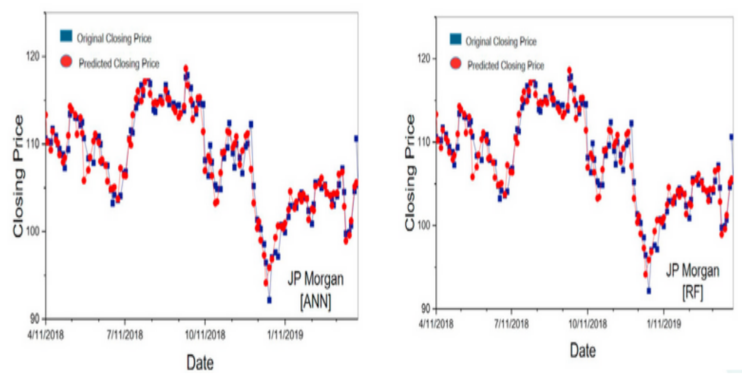

Fig. 1. Histogram of Price Distribution

- Artificial Neural Networks (ANN) and Random Forest (RF) for predicting stock closing prices.

- Stock Price Indicators: The High minus Low (H-L) captures daily price fluctuation by subtracting the lowest price from the highest. Close minus Open (O-C) measures the price change from market open to close, indicating daily momentum. Moving averages (7, 14, and 21 days) smooth short-term fluctuations, track medium-term trends, and highlight long-term trends, respectively. The 7-day standard deviation assesses price volatility by measuring variation over the past week.
- Dataset
  The dataset comprises 10 years of stock market data (2009-2019) from five companies: Nike, JP Morgan, Goldman Sachs, Johnson & Johnson, and Pfizer. The data source is Yahoo Finance, which includes stock prices such as Open, High, Low, Close, Adjusted Close, and Volume.

2nd paper

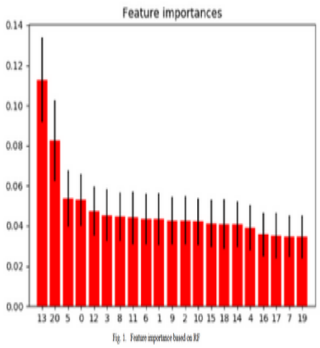

TABLE III. TOP SIX FEATURES SELECTED BY RF

No.	Feature Name
13	PB
20	Relative Return
5	Book Value
0	PE
12	Capital Expenditure
3	Liability

Fig. 2. Feature Selection

- Feature Selection
  The six most important features identified were Price-to-Book ratio (PB), Relative Return, Book Value, Price-to-Earnings ratio (PE), Capital Expenditure, and Liabilities. Feature selection significantly boosted FNN's prediction accuracy, with a notable improvement in the Portfolio Score from 0.202 to 0.345, unlike the modest effect on RF.
- Dataset
  The dataset comprises 22 years of quarterly financial data (1996-2017) for 70 companies from the SP 100 index. Data were sourced from SEC 10-Q filings, with features derived from financial statements, such as revenue, earnings, capital expenditure, and book value.
- Challenges Encountered
  The limited amount of historical financial data and quarterly data frequency restricted the volume available for training. The Efficient Market Hypothesis (EMH) implies that publicly available information is already reflected in stock prices, making predictions challenging.
- Conclusion
  Models based on fundamental analysis can be effective in constructing portfolios that outperform the market without expert intervention. Among the tested algorithms, Feed-forward Neural Network (FNN) achieved the best performance, especially with feature selection. Aggregation of multiple models further enhanced accuracy and portfolio stability.

IV. DATA DESCRIPTION

The dataset used in this project includes stock data from multiple companies spanning the years 2000 to 2023. Key features include:

- **DataSet Link :** https://www.kaggle.com/datasets/chiragb254/indian-stock-market-complete-dataset-2024
- **Date:** Trading date (YYYY-MM-DD).
- **Open, High, Low, Close:** Stock prices for each trading day.
- **Adjusted Close:** Price adjusted for dividends and stock splits.
- **Volume:** The total number of shares traded on the given date.

V. EXPLORATORY DATA ANALYSIS

We performed exploratory data analysis (EDA) to uncover significant trends and insights in the HP Cotton stock data.

A. Statistical Summary

The statistical summary of the data reveals the following insights:

- **Price Features:** The price distribution is highly skewed towards lower values, with outliers reaching up to 200.
- **Volume Distribution:** The trading volume is mostly low, with occasional extreme spikes indicating significant market events.
- **Correlation Analysis:** Price features (Open, High, Low, Close, Adjusted Close) exhibit perfect correlation, while the correlation between price and volume is relatively low (around 0.3).
- **Price Trend:** A clear upward trend is observable post-2020, suggesting positive market sentiment.
- **Volume Spikes:** Sporadic volume spikes may indicate critical market events or announcements.
- **Statistical Metrics:** The median price is approximately 23, with a maximum price exceeding 200, reflecting high variance and large outliers in volume data.
- **Redundant Information:** To reduce dimensionality, focusing on a single price feature may be sufficient due to redundancy in price-related data.

B. Visualizations

The dataset was visualized using various plots to facilitate understanding:

- Histograms to illustrate the distribution of price and volume features.
- Correlation heatmaps to explore relationships between different stock attributes.
- Time-series line plots to track trends in stock prices over the observed period.

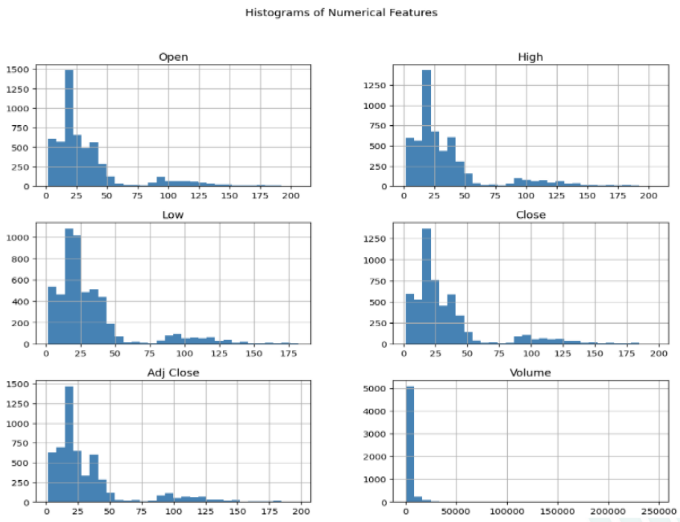

Fig. 3. Histogram of Price Distribution

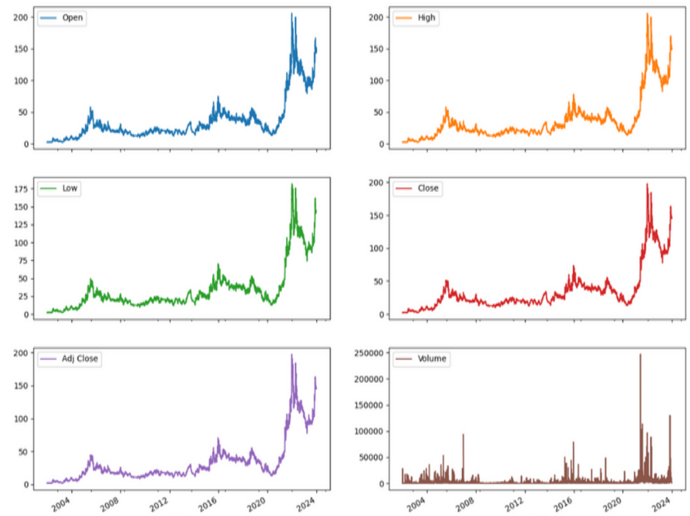

Fig. 5. Time-Series Trend of Stock Prices

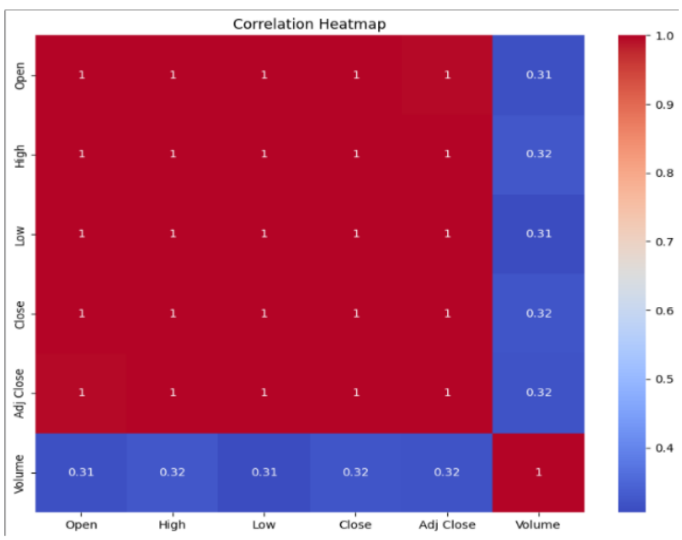

Fig. 4. Correlation Heatmap of Stock Features

VI. Data Preprocessing and Feature Engineering

This section outlines the preprocessing and feature engineering steps taken to prepare the dataset for analysis.

A. Preprocessing Steps

- **Combining Data:** Consolidated all yearly data for each stock into a single file.
- **Filling NaN Values:** Replaced missing values with column averages.
- **Normalization:** Scaled numerical features for uniformity.
- **Handling Discrete Dates:** Filled non-trading days to ensure a continuous dataset.
- **Creating Target Variables:** Generated targets for stock prices after 90 and 180 days.

B. Feature Engineering Steps

- **Volume-based Features:** Calculated percentage changes in volume (quarterly, monthly, weekly, half-yearly, yearly) to assess market activity.
- **Close-based Features:** Evaluated percentage changes in closing prices over various time frames (quarterly, monthly, weekly, half-yearly, yearly).
- **Day and Date-related Features:** Included temporal attributes such as day name, day of week, month, quarter, and week of year to capture seasonal patterns.
- **Flags:** Created indicators for critical milestones (e.g., start and end of years, months, and quarters) to identify significant market events.

C. Date-related Features

- **Day of the Week, Month, Quarter:** Features for capturing temporal patterns.
- **Year Start/End Flags:** Identifies the beginning and end of financial periods.

VII. Methodology

To build an optimized portfolio with carefully selected stocks across various sectors, we follow a two-step approach:

A. Volatility Prediction

We begin by predicting the volatility of a diverse set of stocks and sectors for the upcoming period. Volatility is a critical measure of investment risk, guiding our understanding of potential fluctuations in stock prices.

B. Portfolio Optimization

Using the predicted volatilities, we construct an optimized portfolio that balances risk and return. By selecting stocks with suitable risk profiles, we aim to maximize returns while minimizing overall portfolio risk.

VIII. Results and Analysis

To evaluate the accuracy of our stock prediction model, we conducted out-of-sample testing on financial data from 2021, 2022, and 2023. This involved: Predicting Volatility: Using the trained model to forecast volatility for each year. Comparing Predictions vs. Actuals: Analyzing discrepancies between predicted volatility and actual market volatility to assess the model's reliability. The goal is to validate how well the model generalizes to new data and to understand its performance in real-world scenarios beyond the training set.

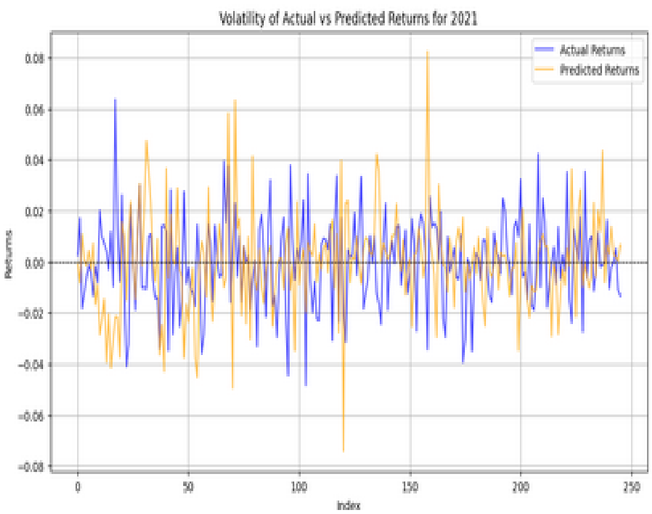

Fig. 6. Volatility of Actual vs Predicted Returns 2021

Training on data up to 2020, Testing on 2021, Test MSE: 41655.76, Test MAPE: 9.34%, Direction Accuracy: 53.25% Volatility (Actual) for 2021: 0.0176 Volatility (Predicted) for 2021: 0.0189

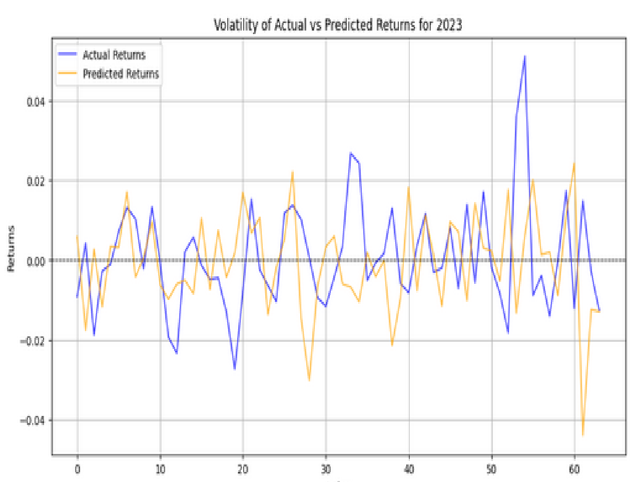

Fig. 7. Volatility of Actual vs Predicted Returns 2023

Training on data up to 2022, Testing on 2023, Test MSE: 28019.87, Test MAPE: 10.61%, Direction Accuracy: 42.19% Volatility (Actual) for 2023: 0.0136 Volatility (Predicted) for 2023: 0.0123

Training on data up to 2020, Testing on 2021, Test MSE: 41655.76, Test MAPE: 9.34, Direction Accuracy: 53.25

A. *Directional Accuracy*

The model achieved a directional accuracy of approximately 50

B. *Volatility Comparison for Different Stocks (KMB & INSY)*

- **2021:**
  - Stock A: Predicted volatility: 0.0189, Actual: 0.0176
  - Stock B: Predicted volatility: 0.0095, Actual: [0.0176]
- **2022:**
  - Stock A: Predicted volatility: 0.0115, Actual: 0.0109
  - Stock B: Predicted volatility: 0.0157, Actual: 0.0149

C. *Key Findings*

- Predicting volatility is a complex task but essential for effective portfolio optimization.
- Our model provides valuable insights into the potential fluctuations of selected stocks.
- The optimization process allowed us to identify and allocate funds to stocks that offer favorable risk-reward ratios.

IX. Conclusion and Future Work

Our model demonstrated a reasonable level of accuracy in predicting stock price trends. However, further refinements are needed to improve consistency across different stocks and market conditions. Future work will focus on:

- **Enhance Features:** Add lag features, technical indicators (RSI, EMA), and sentiment analysis from news.
- **Explore Models:** Test algorithms like XGBoost and implement ensemble methods.
- **Hyperparameter Tuning:** Adjust model parameters manually based on trial and error.
- **Evaluate Performance:** Use metrics like MAE, RMSE, Sharpe Ratio, and perform backtesting.
- **Document Findings:** Record model performance and feature importance for future improvements.
- **Contribution**
  - Akshat Karnwal (2022052), Keshav Chhabra (2022247): Literature Review , Model Building
  - Ramish Jamal (2022395): Model building , Evaluation
  - Mohd Masood (2022299): Data preprocessing , Model Building

X. Appendix

- **Code Snippets:** Key algorithms and implementations used in this project.
- **Data Visualization:** Additional visualizations that support our findings.

References

[1] Y. Huang, L. F. Capretz, and D. Ho, "Machine Learning for Stock Prediction Based on Fundamental Analysis," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-10, doi: 10.1109/SSCI50451.2021.9660134.

[2] H. Ma, G. E. Ab, V. K. Menon, S. K. P. et al., "NSE Stock Market Prediction Using Deep-Learning Models," in *Proceedings of the 2018 International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, pp. 1–x (page numbers to be filled in).

[3] Y. Deshmukh, D. Saratkar, H. Hiratkar, S. Dhopte, S. Patankar, T. Jambhulkar, and Y. Tiwari, "Stock Market Prediction Using Machine Learning,"