Big Data Analysis of Yahoo Finance Stock Dataset Using Hadoop and PySpark

Krishna Kant Mishra (kmishra2, ID: 50681760)
Smit Mahajan (smitdhan, ID: 50682427)
Dhvani Dave (dhvanigu, ID: 50670703)
Partha Chakraborty (parthach, ID: 50669434)

Introduction

A large multiyear Yahoo Finance stock dataset was analyzed to understand stock behavior and market patterns. The data was cleaned, engineered, explored and stored in HDFS to support distributed processing. PySpark was then used to perform parallel EDA and implement four machine learning tasks: egression, classification, clustering and forecasting to study stock trends and market behavior. This combined Hadoop and PySpark workflow enables efficient handling of large financial datasets and supports accurate, data-driven insights into market patterns.

Project Objectives (N = 4 Problems, 8 Goals)

Problem 1: Regression (Forecasting Exact Price)

Goal 1: Identify the predictive influence of lagged price features on next-day closing price using feature-importance analysis.

Goal 2: Evaluate regression accuracy using RMSE and R^2 to measure model error and explanatory power.

Problem 2: Classification (Predicting Price Direction)

Goal 1: Determine which engineered indicators (Daily Return, Volatility Index, etc.) provide the strongest discriminative power for predicting next-day price movement.

Goal 2: Compare classification performance using **Accuracy, Precision, Recall, F1-score** and **AUC-ROC** to evaluate model effectiveness.

Problem 3: Clustering (Market Segmentation)

Goal 1: Segment companies into clusters with distinct risk--return profiles using K-Means and PCA.

Goal 2: Interpret each cluster in financial terms and relate it to potential portfolio strategies.

Problem 4: Forecasting (Sequential Trend Prediction)

Goal 1: Capture temporal dependencies in stock prices using lag-based models that can be applied iteratively.

Goal 2: Produce 5-day forecasts and quantify forecasting error (e.g: MAE), or at minimum demonstrate a working multi-step forecasting pipeline.

Exploratory Data Analysis (EDA)

- Analyzed a large multiyear Yahoo Finance dataset with daily prices and volume for hundreds of stocks.
- Figure 1: Comparative closing-price plots showed that major tech stocks follow similar macro trends (e.g., pandemic shock), while NVIDIA and Microsoft displayed the highest long-term appreciation.
- Figure 2: Weekly aggregated PySpark analysis of MSFT showed clear growth phases, with visible periods of acceleration and consolidation across different market cycles.Compared major companies and examined correlations to understand price behavior and liquidity patterns.
- Figure 3: The correlation matrix shows that Open, High, Low, and Close prices are almost perfectly positively correlated, meaning these price variables move together consistently across all companies. In contrast, Volume, Dividends, and Stock Splits exhibit very weak or near-zero correlations with price metrics, indicating they behave independently of daily price movements.

Figure 1: Closing Prices of Major Companies Over Time

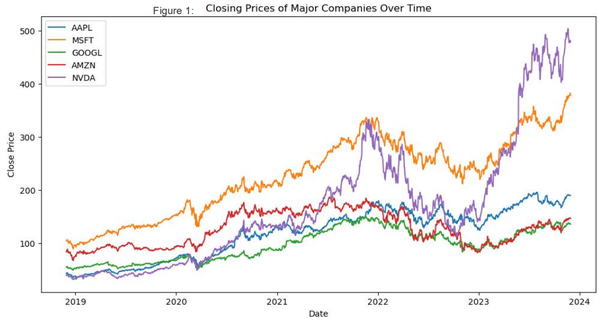


Figure 2: MSFT - Weekly Average Close

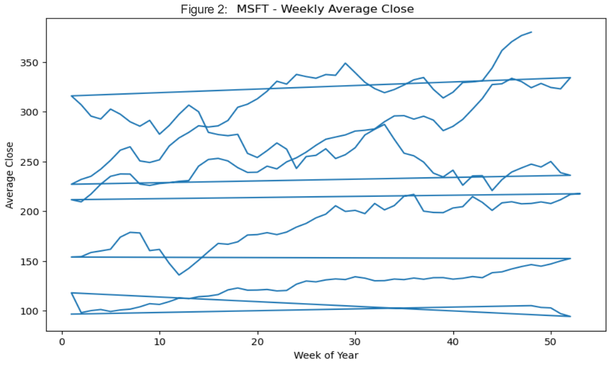
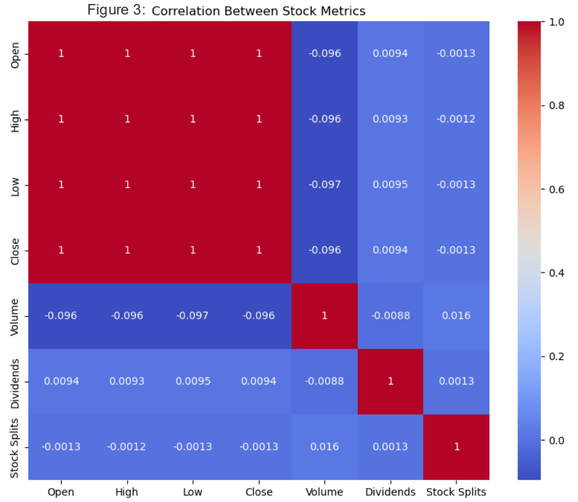


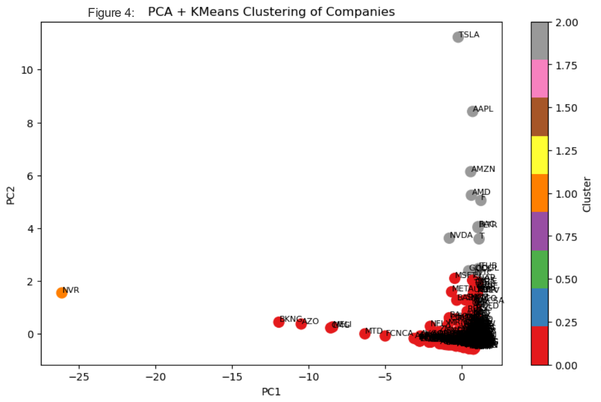
Figure 3: Correlation Between Stock Metrics



Machine Learning Models

- **Regression (GBT Regressor):** Predicted next-day closing prices using 7-day lagged features, achieving $R^2 \approx 0.79$, and enabled a functional 5-day forecasting pipeline.
- **Forecasting (Iterative Modeling):** Used the regression model iteratively to generate 5-day ahead price forecasts, demonstrating the ability to extend single-step predictions into short-term trend projections.
- **Clustering (PCA + K-Means):** Grouped companies into stable, aggressive growth, and intermediate clusters based on return, volatility, and volume characteristics (Figure 4)

Figure 4: PCA + KMeans Clustering of Companies



Key Findings

- **Data Preparation and Analysis:** The dataset was thoroughly cleaned and explored using Python and Py Spark, providing high-quality inputs and actionable insights.
- **Modeling Results:** Regression and clustering objectives were fully achieved with interpretable outputs; classification showed no strong predictive signal, highlighting market efficiency, while forecasting is implemented but pending full evaluation.
- **Limitations and Future Work:** Models rely mainly on technical indicators . Classification and forecasting performance require additional metrics and broader feature selection.

Future Directions

- **Enhanced Feature Engineering:** Incorporate technical indicators, fundamental metrics, and text-based sentiment data to improve classification accuracy and strengthen regression and forecasting models.
- **Advanced Modeling and Evaluation:** Explore deep learning models (RNNs, LSTMs, GRUs, Temporal CNNs) for sequential forecasting and implement robust evaluation using rolling-window MAE/MAPE and statistical tests against simple baselines.
- **Practical Applications and Expansion:** Leverage clustering for portfolio construction (conservative vs. aggressive) and extend the analysis to longer time horizons and additional international markets to assess generalizability and investment strategies.
- **Overall Recommendation:** Combining richer features, advanced models, and broader market coverage can enhance predictive performance and make the analytics pipeline more actionable for real-world investment decisions.

References

1. Yahoo Finance Dataset, Kaggle. https://www.kaggle.com/
2. Apache Hadoop Documentation. https://hadoop.apache.org/
3. Apache Spark MLlib Documentation. https://spark.apache.org/
4. Python Pandas Documentation. https://pandas.pydata.org/
5. Matplotlib and Seaborn Documentation. https://matplotlib.org/

University at Buffalo The State University of New York

Department of Computer Science
State University of New York, Buffalo
buffalo.edu