Stock Price Prediction Using Time Series & Machine Learning

Role: Data Science Intern – Hedge Fund

Executive Summary

This report presents an end-to-end data science pipeline for predicting stock prices using statistical and machine learning models. Using daily OHLCV data from Yahoo Finance for multiple equities, ARIMA and Gradient Boosting models were developed and evaluated. Gradient Boosting demonstrated superior predictive performance and stronger applicability for trading strategies.

Data & Methodology

Daily OHLCV data for AAPL, MSFT, GOOGL, AMZN, and META from 2015–2024 was used. Data was cleaned, validated, indexed by date, and prepared for time-series modeling. Missing values and corrupt records were removed to ensure data integrity.

Exploratory Data Analysis

EDA revealed strong long-term trends, volatility clustering, and regime shifts during macroeconomic events. Returns exhibited non-normal distributions, supporting the need for advanced models.

Feature Engineering

Lagged prices, rolling means, rolling volatility, percentage returns, and volume changes were engineered to capture momentum, mean reversion, and liquidity effects.

Model Development

ARIMA: Used as a statistical baseline after ensuring stationarity. Effective in stable markets but weak during volatile regimes.

Gradient Boosting: Leveraged engineered features to capture non-linear patterns. Hyperparameter tuning significantly improved accuracy.

Model Evaluation

Models were evaluated using RMSE, MAE, and MAPE. Gradient Boosting consistently outperformed ARIMA across all metrics, achieving lower error rates and better tracking of market movements.

Trading Strategy Implications

ARIMA models are suitable for short-term, mean-reversion strategies in low-volatility markets. Gradient Boosting models are better suited for alpha generation, momentum strategies, and risk-adjusted trading systems due to their adaptability and accuracy.

Recommendations

Use Gradient Boosting as the primary predictive model, retain ARIMA as a benchmark, and enhance the system with additional technical and macroeconomic features. Future work should focus on signal-based trading and portfolio-level optimization.

Conclusion

This project demonstrates a hedge-fund-grade modeling workflow combining statistical rigor with machine learning, offering a strong foundation for real-world trading applications.