Stock Market Pattern Mining

Arnav Purushotham
University of Colorado Boulder
USA

Meghasrivardhan Pulakhandam
University of Colorado Boulder
USA

Sneha Nagaraju
University of Colorado Boulder
USA

Abstract

The stock market generates massive amounts of financial data daily, ranging from price movements to trading volumes. While the data is abundant, it is also extremely noisy, making it challenging to extract useful patterns or build accurate predictive models. Most prior efforts focus on static prediction using historical prices alone, but these approaches fail to generalize in real-world scenarios. In this project, we propose a hybrid system that integrates machine learning models trained on historical data with live stock prices obtained through the FinnHub API and contextual analysis from the GPT-4 API. Our goal is not to "beat the market" but to design a comprehensive data mining framework that uncovers patterns, evaluates their reliability, and enriches predictions with external knowledge. This project highlights both technical and educational contributions by combining pattern mining, model training, real-time inference, and contextual aggregation.

Keywords

Data Mining, Stock Market, Machine Learning, Pattern Mining, Real-Time Inference, Financial Analytics

1 Introduction

The stock market is a cornerstone of the global economy, attracting analysts, traders, and researchers who all seek to understand its underlying patterns. Despite the availability of vast datasets, stock prices are highly volatile and influenced by both internal (e.g., technical indicators, trading volumes) and external (e.g., political news, earnings reports) factors. This makes the task of stock prediction particularly difficult.

Most student-level projects focus solely on predicting stock movements using historical prices. However, accuracy often hovers around 50%, which is no better than random guessing. Our project takes a broader view of data mining by asking: can we extract interpretable patterns from historical data, evaluate their reliability, and combine them with real-time financial and contextual information? By addressing this question, we shift the focus from achieving high predictive accuracy to generating actionable insights and demonstrating the educational value of the data mining pipeline.

The specific objectives of this project are:

- To preprocess and transform raw stock market data into structured features such as moving averages, volatility measures, and candlestick indicators.
- To mine frequent and co-occurring patterns, evaluating their reliability using support, confidence, and lift.
- To train and compare multiple ML models, ranging from interpretable baselines to advanced classifiers.
- To integrate real-time stock price feeds through FinnHub and contextual news summaries through GPT-4, producing combined analysis results.
- To present findings through visual dashboards that emphasize interpretability and transparency.

2 Related Work

The task of stock prediction has attracted decades of research. Early approaches relied on technical indicators such as Simple Moving Averages (SMA), Exponential Moving Averages (EMA), Relative Strength Index (RSI), and Bollinger Bands. These rule-based strategies are still widely used but often fail to adapt to sudden market shifts.

Machine learning introduced models such as Decision Trees, Random Forests, and Support Vector Machines (SVMs) that can capture non-linear relationships in financial data. Neural networks and deep learning methods such as LSTMs have also been applied, with some success in modeling sequential dependencies. However, these methods suffer from limited interpretability and often require large-scale proprietary datasets that are not available to academic projects.

Another strand of research examines sentiment analysis and natural language processing (NLP). Studies have shown that news headlines, analyst reports, and even Twitter posts can influence short-term price movements. These methods are usually applied separately from technical analysis, leading to siloed results.

Recent work has sought to address these gaps. Li et al. (2024) introduced the MASTER model, a market-guided transformer that captures both cross-time and cross-stock correlations while incorporating market-level guidance [1]. Their model demonstrated strong improvements in short-horizon forecasting compared to RNN and transformer baselines, showing the importance of cross-asset modeling. However, MASTER relies solely on numerical stock data and does not integrate external information sources such as real-time APIs or news.

Cao et al. (2023) proposed a large-scale forecasting framework that improves generalization by introducing an invariant-learning regularizer [2]. Their results showed more robust performance across different market regimes, highlighting the importance of evaluation beyond in-sample accuracy. Inspired by this, our work also emphasizes pattern reliability and interpretability, not only predictive accuracy.

Chen and Kawashima (2024) explored the use of large language models (LLMs) for stock prediction via sentiment analysis [3]. Their findings reinforce that textual and sentiment signals can complement numerical models. Our project incorporates a practical extension of this idea by querying GPT-4 for contextual insights and combining them with numerical ML predictions.

In summary, prior research has demonstrated the value of technical indicators, classical ML, deep sequence models, and more recently transformers and LLMs for financial forecasting. Our work builds on these efforts but contributes a novel system that combines (1) traditional ML on historical price data, (2) real-time stock price integration via the FinnHub API, and (3) GPT-4–based contextual analysis. This integration allows us to move beyond static prediction to a multi-source, interpretable system. The focus on reliability of patterns and explainability sets this project apart from prior student-level work.

3 Proposed Work

We will design and implement an end-to-end pipeline consisting of the following steps:

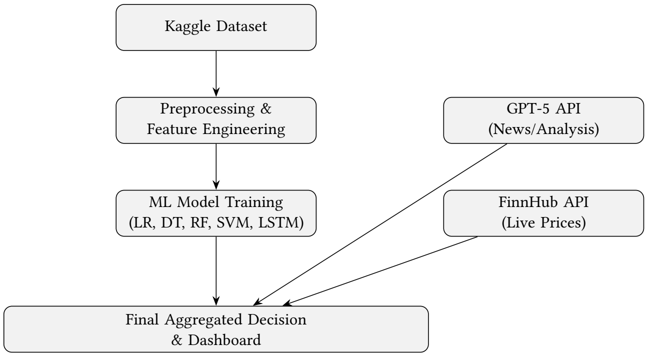

Figure 1: Proposed system architecture integrating historical data (Kaggle), ML training, FinnHub live prices, and GPT-4 contextual analysis.

3.1 Data Preparation

We will use the Kaggle Stock Market Dataset, which includes open, high, low, close, and volume data for multiple stocks. Data cleaning will involve handling missing values, aligning trading dates, and adjusting for splits or anomalies. Feature engineering will generate moving averages, returns, volatility measures, and volume ratios, along with derived features like candlestick shapes.

3.2 Pattern Mining

Using the engineered features, we will identify frequent patterns such as moving average crossovers, volume surges, and volatility spikes. Association rule mining will be applied to detect combined signals (e.g., crossover + high volume). Each pattern will be evaluated for reliability using support, confidence, and lift.

3.3 Model Training

We will train and compare several ML models:

- Logistic Regression: interpretable baseline.
- Decision Trees: captures non-linear patterns, easy to visualize.
- Random Forests: ensemble approach for improved robustness.
- SVM: handles complex boundaries in feature space.
- LSTM (optional): sequence model to capture time dependencies.

3.4 Real-Time Inference

The FinnHub API will provide live stock prices and recent historical windows. These will be fed into trained models to produce predictions (up or down) with confidence scores. This allows us to test the models in a streaming, real-world environment.

3.5 External Context Integration

The GPT-4 API will be queried to summarize recent news or analysis about the chosen stock. The ML prediction and GPT-4 context will then be concatenated and re-processed by GPT-4 to produce a final combined decision. This step simulates how human decision-making combines numerical data with qualitative context.

3.6 Visualization

We will build a dashboard to visualize mined patterns, model predictions, and GPT-4 summaries. This dashboard will highlight interpretability and demonstrate the integration of multiple data sources.

4 Evaluation

Evaluating the effectiveness of our proposed system requires going beyond simple notions of accuracy or profitability. Since financial markets are highly efficient and noisy, a strong evaluation must measure not only whether predictions are correct, but also how robust, interpretable, and informative the system is. We outline five complementary evaluation dimensions.

4.1 Pattern Reliability

One core contribution of this project is mining and quantifying recurring patterns in historical data. We will evaluate each pattern using classical data mining measures:

- **Support:** the proportion of days a given pattern (e.g., SMA crossover + volume surge) occurs in the dataset.
- **Confidence:** the conditional probability that a price increase (or decrease) follows the pattern within a specified time horizon (e.g., next 3 trading days).
- **Lift:** the ratio of observed confidence to baseline probability of an up or down move, showing how much the pattern improves upon chance.

These measures allow us to claim whether a pattern is meaningful or just noise. For example, if a crossover pattern occurs often but has a lift close to 1.0, it adds little predictive value.

4.2 Model Metrics

We will evaluate our machine learning models (Logistic Regression, Decision Trees, Random Forests, SVM, and optionally LSTM) using well-established metrics:

- **Accuracy:** proportion of correct predictions, compared against a 50% random baseline.
- **Precision and Recall:** to assess whether the model correctly identifies up-moves (precision) and how many up-moves it captures (recall).

- **F1-score:** harmonic mean of precision and recall, balancing the two.

Rather than focusing only on accuracy, which may be misleading in imbalanced datasets, we will use these complementary metrics to present a fairer picture of model quality.

4.3 Feature Importance

Interpretability is key for educational and practical relevance. For tree-based models (Decision Trees, Random Forests), we will extract feature importances to identify which indicators drive predictions most strongly. For Logistic Regression, coefficient magnitudes will be analyzed. We will compare the relative importance of features such as short-term moving averages, long-term moving averages, volatility, and volume ratios. This analysis will provide insights into which engineered features contribute most to model behavior, even if overall accuracy remains close to random.

4.4 System Evaluation

Beyond individual models, we will test the end-to-end system with live inputs. FinnHub's API will provide rolling sequences of recent prices, and we will compare model outputs against the actual market direction over subsequent days. The key here is to assess robustness: does the system behave consistently across stable, volatile, and high-volume regimes? We will also evaluate whether integrating GPT-4's contextual summaries improves clarity and interpretability of results. For example, if the model predicts "up" but GPT-4 highlights negative earnings news, the combined output may present a more nuanced and realistic analysis.

4.5 Visualization Quality

Finally, we will evaluate the usability and interpretability of our visual dashboard. Effective visualizations should make complex signals understandable:

- Overlaying technical indicators and highlighting mined patterns on price charts.
- Comparing predicted vs. actual outcomes in time-series plots.
- Presenting model performance side-by-side in bar charts.
- Integrating GPT-4 textual summaries alongside numeric predictions.

We will informally test whether these visualizations make sense to an audience with limited financial background. The success criterion is that patterns and outcomes can be communicated clearly, supporting the educational goals of the project.

5 Milestones

The project will follow this timeline:

- Weeks 1–2: Collect and clean Kaggle dataset; implement preprocessing pipeline.
- Weeks 3–4: Engineer features; implement pattern mining; run initial statistics.
- Week 5: Train baseline ML models (Logistic Regression, Decision Tree).
- Week 6: Train advanced models (Random Forest, SVM, optional LSTM); compare results.
- Week 7: Integrate FinnHub API for live inference; integrate GPT-4 API for contextual analysis.
- Week 8: Develop dashboard; finalize experiments; prepare final report and presentation.

6 Discussion and Expected Contributions

This project will contribute in three ways:

(1) Demonstrate the application of multiple data mining techniques on financial data, including preprocessing, pattern mining, model training, and evaluation.
(2) Highlight the reliability of common trading patterns using support and confidence measures, bridging technical analysis and machine learning.
(3) Build a novel integration of live stock prices and external news analysis, showcasing how multiple information sources can be combined into a coherent decision-making system.

7 Conclusion

In summary, our project will not attempt to produce trading profits but instead aims to showcase the educational and analytical value of data mining in finance. By combining historical ML models, pattern mining, live data, and contextual analysis, we aim to deliver an interpretable and realistic system that demonstrates the strengths and limitations of predictive modeling in noisy real-world domains.

References

[1] T. Li *et al.*, "MASTER: Market-Guided Stock Transformer for Stock Price Forecasting," in *Proc. AAAI Conf. Artificial Intelligence (AAAI'24)*, 2024.

[2] D. Cao *et al.*, "Large-Scale Financial Time Series Forecasting with a Multi-Faceted Model and Invariant Learning," in *Proc. 4th ACM Int. Conf. AI in Finance (ICAIF '23)*, New York, NY, USA, 2023.

[3] Q. Chen and H. Kawashima, "Stock Price Prediction Using LLM-Based Sentiment Analysis," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Washington, DC, USA, 2024. (to appear)