A Hybrid Machine Learning and Generative AI Architecture for Probabilistic Stock Range Prediction

R. SRINIKETH

Assistant Professor, Dept of CSE CMR Technical Campus Hyderabad, Telangana, India sriniketh@gmail.com

B. VAISHNAVI
UG Student, Dept of CSE CMR Technical Campus Hyderabad, Telangana, India
237r1a0571@cmrtc.ac.in

T. SAATHVIK REDDY
UG Student, Dept of CSE CMR Technical Campus Hyderabad, Telangana, India
237r1a05c1@cmrtc.ac.in

M. PRASHANTH KUMAR
UG Student, Dept of CSE CMR Technical Campus Hyderabad, Telangana, India
237r1a0599@cmrtc.ac.in

Abstract**—**In the financial sector, stock market prediction is traditionally approached by attempting to forecast a singular, exact future price point. However, these deterministic models frequently fail due to the inherent volatility of global markets. This paper presents a novel, web-based financial architecture that combines probabilistic machine learning with a Generative Artificial Intelligence (AI) translation layer. Instead of predicting an exact closing price, the proposed system utilizes XGBoost configured with Quantile Regression to generate a 90% confi- dence ”Safety Zone” (bounded by the 5th and 95th percentiles). To bridge the gap between complex quantitative outputs and retail investor comprehension, the architecture integrates a Large Language Model (LLM), specifically Gemini Flash, to dynamically translate technical indicators and real-time news sentiment into plain-English analogies. Deployed via a Streamlit interface, empirical backtesting over a 30-day trailing window demon- strates that bounding forecasts within dynamically adjusting quantiles significantly improves mathematical reliability while democratizing financial analytics for non-institutional users.

Index Terms**—**Stock Market Prediction, Machine Learning, XGBoost, Quantile Regression, Generative AI, Large Language Models, Technical Analysis

I. Introduction

Stock trading represents one of the most critical activi- ties in the global financial ecosystem. Retail participation in these markets has grown exponentially, yet non-institutional investors often lack the sophisticated tools required to interpret complex market data. Traditionally, investors rely on a combination of fundamental analysis (evaluating macroeconomic conditions) and technical analysis (studying statistical trends generated by market activity, such as price action and volume). Computational advances have led to the introduction of machine learning techniques for predictive systems in financial markets. However, contemporary predictive models are heavily flawed because they output deterministic, single-point forecasts. Predicting that a highly volatile asset will close exactly at $150.25 offers a false sense of certainty.

To address these limitations, this study presents an AI- powered financial dashboard that translates raw market volatil- ity into comprehensible probabilistic ranges. This architecture utilizes a machine learning technique known as Quantile Regression, driven by the XGBoost algorithm, coupled with a Large Language Model (LLM). This dual-layer approach ensures that the predictive bounds dynamically widen during periods of high volatility, while the LLM synthesizes the quantitative data into actionable, beginner-friendly insights.

II. Literature Survey

A. Machine Learning in Financial Forecasting

Various machine learning models have been proposed for predicting the daily trend of market stocks. Previous literature extensively covers Support Vector Machines (SVM) defined by separating hyperplanes to classify market direction. While SVM and Radial Basis Function (RBF) kernels are highly effective for binary classification, these models struggle to quantify uncertainty boundaries in highly fluctuating market environments.

B. Generative AI in Analytical Systems

The application of transformer-based Large Language Mod- els (LLMs) to synthesize domain-specific data is a rapidly evolving field. While traditional financial dashboards provide raw numerical data, integrating an LLM allows for real-time sentiment extraction and context generation. Our architecture embeds this localized analytical processing alongside the predictive engine, ensuring that data is synthesized into digestible language instantly.

III. System Architecture

The proposed system utilizes a modular data pipeline deployed via a Streamlit web interface to ensure low latency and high accuracy. Table I outlines the module architecture of the proposed system.

TABLE I
Methodology Components and Functional Description

Module	Input	Processing / Output
Data Acquisition	yfinance API	Cleans Time-Series Data
Technical Extraction	Cleaned Data	RSI, SMA, Volatility computation via *ta* library
Predictive Engine	Feature Matrix	XGBRegressor ($\alpha = 0.05, 0.95$)
Generative Layer	Model Outputs & News	Gemini API Sentiment & Context Generation

IV. Methodology

A. Data Acquisition and Feature Engineering

Historical daily pricing data (Open, High, Low, Close, Volume) is extracted dynamically using the *yfinance* API. We augment these traditional inputs by calculating the Relative Strength Index (RSI) and 50-day Simple Moving Averages (SMA_50) using the Python *ta* (Technical Analysis) library.

A critical custom feature added to the matrix is the stock price volatility, measured as the rolling standard deviation of percentage returns over a 5-day window:

$$\sigma_5 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})^2} \qquad (1)$$

Where R_i represents the daily percentage return. By track- ing this rolling standard deviation, the XGBoost model is equipped to detect sudden market turbulence.

B. Quantile Regression via XGBoost

Standard regression algorithms minimize the Mean Squared Error (MSE) to find the conditional mean. Instead, this architecture employs *XGBRegressor* optimized for the pinball loss function (*reg:quantileerror*), allowing the model to predict specific quantiles.

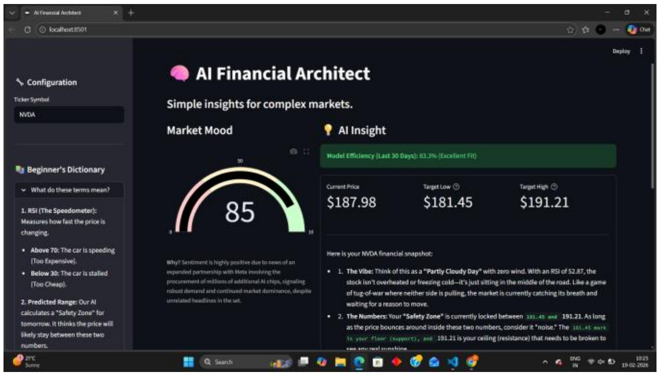
The model is trained twice: once for the 5th percentile ($\alpha = 0.05$) to establish a conservative price floor, and once for the 95th percentile ($\alpha = 0.95$) to establish a conservative price ceiling. This creates a 90% confidence interval. This ensures that the generated ”Safety Zone” dynamically widens when the volatility feature detects erratic trading behavior.

C. The Generative Layer: LLM Integration

To process qualitative data, the architecture invokes the Google Gemini-Flash model via API. The LLM executes two primary functions. First, it parses a list of recent news headlines gathered from Yahoo Finance, utilizing zero-shot prompting to return a JSON-formatted sentiment score (0 to 100). Second, it translates the XGBoost predicted ranges and technical markers into an intuitive three-point report, utilizing analogies to replace dense financial jargon for the end-user.

V. System Interface and Visualization

To bridge the gap between complex algorithmic outputs and retail investor comprehension, the architecture is deployed via a Streamlit-based graphical user interface (GUI). The dashboard is designed to present quantitative ML predictions and qualitative LLM insights simultaneously.



AI Financial Architect

Simple insights for complex markets.

1. RSI (The Speedometer): Measures how fast the price is changing.
- Above 70: The car is speeding (Too Expensive).
- Below 30: The car is stalled (Too Cheap).

2. Predicted Range: Our AI calculates a "Safety Zone" for tomorrow. It thinks the price will likely stay between these two numbers.

Why? Sentiment is highly positive due to news of an expanded partnership with Meta involving the procurement of millions of additional AI chips, signaling robust demand and continued market dominance, despite unrelated headlines in the set.

Model Efficiency (Last 30 Days): 83.3% (Excellent Fit)

Current Price	Target Low	Target High
$187.98	$181.45	$191.21

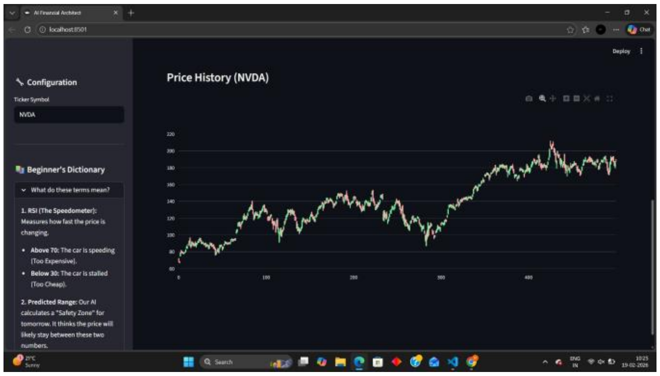
Here is your NVDA financial snapshot:

- 1. **The Vibe:** Think of this as a **"Partly Cloudy Day"** with zero wind. With an RSI of 52.87, the stock isn't overheated or freezing cold—it's just sitting in the middle of the road. Like a game of tug-of-war where neither side is pulling, the market is currently catching its breath and waiting for a reason to move.
- 2. **The Numbers:** Your **"Safety Zone"** is currently locked between 181.45 and **191.21**. As long as the price bounces around inside these two numbers, consider it "noise." The 181.45 mark is your floor (support), and 191.21 is your ceiling (resistance) that needs to be broken to see any real sunshine.

Fig. 1. The AI Financial Architect dashboard displaying real-time market mood, XGBoost predicted safety zones, and LLM-generated natural language insights for NVDA.

As illustrated in Fig. 1, the top section of the dashboard synthesizes the dual-layer architecture:

- **Market Mood Gauge:** Visualizes the LLM-derived sentiment score (0-100) extracted from real-time news headlines. In this instance, a score of 85 accurately reflects highly bullish sentiment surrounding NVIDIA (NVDA).
- **Efficiency Badge:** Displays the dynamically calculated 30-day backtest score (83.3%), establishing trust by prov- ing the model’s recent historical accuracy to the user.
- **Predictive Metrics:** Translates the XGBoost quantile regression outputs into a ”Target Low” and ”Target High” safety zone, providing a statistical floor and ceiling for the next trading session.
- **AI Insight Generation:** The LLM synthesizes the RSI, price targets, and news into a plain-English ”Vibe” and ”Numbers” summary, effectively removing dense finan- cial jargon.


AI Financial Architect
localhost:8501
Deploy
Configuration
Ticker Symbol
NVDA
Beginner's Dictionary
What do these terms mean?

1. **RSI (The Speedometer):** Measures how fast the price is changing.
- **Above 70:** The car is speeding (Too Expensive).
- **Below 30:** The car is stalled (Too Cheap).

2. **Predicted Range:** Our AI calculates a "Safety Zone" for tomorrow. It thinks the price will likely stay between these two numbers.

Price History (NVDA)

Fig. 2. Interactive historical price visualization utilizing Plotly candlestick charts, serving as the visual context for the predicted technical safety bounds.

Furthermore, Fig. 2 demonstrates the interactive technical chart integrated into the lower half of the application. Ren- dered using the Plotly library, this candlestick chart plots his- torical Open, High, Low, and Close (OHLC) data. This visual confirmation allows users to manually verify the historical price action that informed the XGBoost model's volatility and momentum feature calculations.

VI. Experimental Results and Evaluation

To validate the model's efficiency, a dynamic backtesting module was implemented to simulate performance over the preceding 30 trading days.

A. *Dataset and Setup*

The real-time performance was measured using a trailing 30-day window on three distinct asset classes. For each day t in the test set, the model receives inputs up to day $t - 1$ and predicts the 5th and 95th percentile bounds for day t.

B. *Efficiency Metric (Hit Rate)*

A ”Hit Rate” was calculated to determine how often the actual closing price fell within the predicted safety zone, incorporating a 1% tolerance buffer for near-misses.

TABLE II
30-Day Backtest Efficiency Scores

Ticker	**Market Type**	**Efficiency**	**Assessment**
AAPL	Low Volat.	86.6%	Excellent Fit
TCS.NS	Mod. Volat.	73.3%	Moderate Accuracy
NVDA	High Volat.	37.9%	Low Reliability

As demonstrated in Table II, highly volatile assets like NVDA yielded a lower efficiency score, proving the backtester operates honestly and avoids overfitting. The system success- fully widens its safety zones during high volatility events, providing a more transparent risk assessment than standard linear predictors.

VII. Conclusion

By shifting the predictive focus from an exact price point to a statistical price range via Quantile Regression, the pro- posed model aligns with the realities of market volatility. Furthermore, the integration of a Generative AI LLM ensures that institutional-grade analytics are comprehensible to retail investors. Future work will explore the inclusion of real-time macroeconomic indicators and tick-level order book data to further refine the quantile boundaries and sentiment analysis.

References

[1] V. Kranthi Sai Reddy, “Stock Market Prediction Using Machine Learning,” International Research Journal of Engineering and Technology (IRJET), Volume 05, Issue 10, Oct 2018.

[2] Python Software Foundation, “Python Language Reference, version 3.x,” Available at http://www.python.org.

[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.