PeakPredict: A Stock Market Prediction Using ML Algorithms

Aryen Rawat
Department of Networking and Communications
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
ar9486@srmist.edu.in

Arnav Saharan
Department of Networking and Communications
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
ak1223@srmist.edu.in

Arpit Sagar
Department of Networking and Communications
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
am7136@srmist.edu.in

V. Rajaram
Department of Networking and Communications
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
rajaramv@srmist.edu.in

***Abstract*—Predicting the stock market is a sophisticated undertaking due to the consideration of past prices, economic factors, and the market's prevailing mood, amongst many others. This document proposes PeakPredict, an integrated ensemble deep learning architecture tailored for accurate stock market predictions. The presented model uses Long Short-Term Memory (LSTM) to capture sequential dependencies in time-series data and utilizes XGBoost to manage the intricate interactions among features. Moreover, a hybrid LSTM-XGBoost model is created to leverage the positive aspects of both approaches for better prediction outcomes. To improve the predictive power of the framework, anomaly detection is used to identify irregular patterns in the market, while sentiment analysis gathers information from social platforms and news channels to predict market trends. PeakPredict brings together all these features to develop a more flexible and intelligent prediction system. Testing our model against real-world stock market data proves that the hybrid system outperforms both individual models and is better suited for the complexities of modern accuracy. The results are an example of how AI can help in financial predictions, with upcoming efforts directed towards reinforcement learning, better sentiment analysis, and real-time predictive optimization.**

***Index Terms*—Stock market prediction, machine learning, LSTM, XGBoost, hybrid Model, anomaly detection, and sentiment analysis.**

I. Introduction

Predicting and forecasting the stock market matters a lot because it influences how people invest, manage risk, and plan for the future. Stock prices can be really unpredictable—they change constantly because of historical trends and patterns, daily news, market moods, economic signals, and even surprise events like natural calamities. In the past, people used methods like ARIMA and other statistical models, but these often miss the mark since they can't capture the wild, non-linear behavior of the market. That is why there's been a shift towards using machine learning and deep learning; these techniques are better at predicting the the unseen or unspotted trends and patterns, hence provide more accurate results and help people gain deeper insights.

In this paper, we talk about PeakPredict, a system we built to guess stock prices better. It uses LSTM, which is a type of neural network that can learn from how prices change over time, and XGBoost, an algorithm that is great at handling structured data and picking out the most important features. By integrating these two models, our hybrid LSTM-XGBoost model makes more reliable and accurate predictions. We have also integrated sentiment analysis to look at news and social media, which helps understand if people are feeling positive or negative about the market. Plus, we use anomaly detection to spot any odd jumps or drops in prices so the system can adjust quickly. The main idea is to create a smart, data-driven tool that beats the old forecasting methods and gives traders, investors, and analysts useful insights to make better decisions in a very unpredictable market.

Furthermore, PeakPredict enhances forecasting robustness by integrating anomaly detection to identify irregularities in stock trends, ensuring adaptability to sudden market shifts such as sudden spike or dip. The main motivation behind this research is to develop an smart and intelligent, stock prediction system that outperforms traditional forecasting methods by utilizing hybrid Machine Learning algorithms and real-time financial data streams. Data streams can include medias such as news, blogs, or social posts which is useful in giving more accurate results and minimal error variation. Moreover by incorporating the deep learning and sentiment-aware analytics, this framework provides traders, investors, and financial analysts with actionable insights which improves decision-making in highly dynamic and volatile market.

II. System Architecture

A. Overview

The given system architecture for stock market prediction consists of four main component. The components are as follows: data collection, preprocessing, model training, and

TABLE I
LITERATURE REVIEW ON STOCK MARKET PREDICTION USING MACHINE LEARNING AND AI

S. No	Title (Journal, author, and details)	Methodology (Summary of key studies and findings)	Gaps and Limitations
1	Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.	Introduced LSTM networks for capturing long-term dependencies in sequential data. Widely used for time-series forecasting.	High computational cost; requires large datasets for optimal performance.
2	Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference.	Developed an efficient gradient boosting algorithm, outperforming traditional methods in structured data prediction tasks.	Limited performance on sequential data; requires feature engineering.
3	Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.	Explored sentiment analysis for stock prediction using social media sentiment data. Found correlation between public mood and market trends.	Limited scope in handling sarcasm, fake news, and rapidly changing sentiments.
4	Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.	Provided foundational deep learning techniques, including CNNs, RNNs, and reinforcement learning for financial applications.	Requires large amounts of labeled data; overfitting risks in financial modeling.
5	Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.	Proposed a hybrid machine learning approach combining SVM, ANN, and random forests for market prediction.	Performance varies with market volatility; lacks real-time adaptability.
6	Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction. Expert Systems with Applications, 83, 187-205.	Applied deep learning models like LSTMs and CNNs for stock prediction, demonstrating improved accuracy.	Computationally expensive; sensitive to hyperparameter tuning.
7	Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702.	Compared ML models for trading strategies and arbitrage, showing deep learning models outperform traditional ones.	Lacks explainability; high susceptibility to market anomalies.
8	Li, X., Wang, C., Dong, L., & Wang, F. (2020). Enhancing stock market prediction through attention-based LSTM models. Applied Soft Computing, 92, 106-275.	Improved traditional LSTM models by incorporating attention mechanisms for better feature extraction and trend analysis.	Requires large historical data; interpretability issues in decision-making.
9	Kim, H., Kim, S., & Lee, H. (2021). Transformer-based models for stock price forecasting. Journal of Financial Data Science, 3(2), 45-67.	Applied Transformer-based architectures (BERT, RoBERTa) to financial data for improved sequential learning and sentiment-based predictions.	Requires extensive fine-tuning; complex architecture may lead to overfitting.
10	Chen, X., Li, Y., & Zhou, J. (2022). Explainable AI for stock market prediction. Artificial Intelligence in Finance, 10(3), 85-112.	Introduced SHAP and LIME for explaining ML-based stock market predictions, improving interpretability for investors.	Computationally expensive; explanation accuracy depends on feature engineering.

real-time prediction. Each segment plays an imperative role in ensuring accurate and efficient forecasting.

The data collection module gathers historical stock prices, financial reports, news sentiment, and social media trends, forming a comprehensive dataset for training the model. The preprocessing level involves handling missing values incorrect values, normalizing data (i.e simplifying the data), and converting textual sentiment (social media, news, blogs) into structural format such as numeric data input which ensures consistency and reliability.

In the model training phase, we employ a hybrid approach combining LSTM, XGBoost, and sentiment analysis to capture both historical trends and market sentiments effectively. We have also done the integration of deep learning which enhances the accuracy in the prediction of the market trends based on historcial as well current news and media.

This proposed modular design allows for scalability, reliability, and adaptability which makes the system more robust against market fluctuations. Future enhancements will focus on optimizing real-time learning and expanding dataset diversity to further improve and refine the forecasting precision.

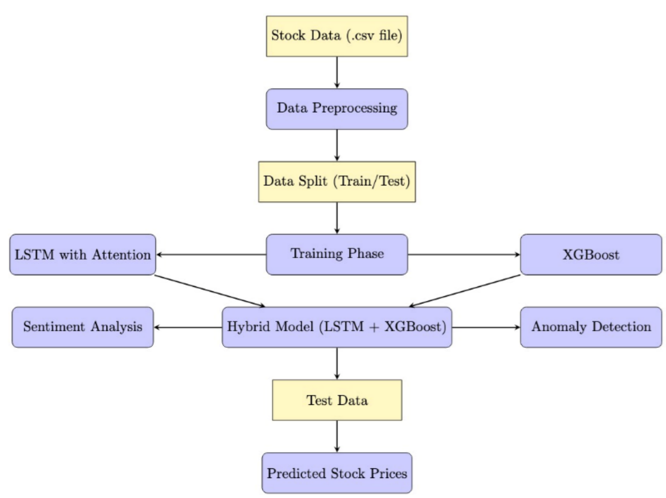

Fig. 1. System Architecture

B. Data Collection

The dataset used in this system comprises two key components i.e, historical stock prices and sentiment data both of the are integrated with the date feature present in both the tables. Stock price data, including open, close, high, low, and volume, is collected from Kaggle. This data is structured in **Table I**, which is used as the fundamental technical analysis.

To incorporate the market sentiment, we've extracted news headlines and social media discussions and popular blog posts using Natural Language Processing (NLP) techniques. The sentiment of each headline is analyzed and assigned a score of **-1** (negative), **0** (neutral), or **1** (positive). This structured sentiment dataset is presented in **Table II**, which is used to capture the psychological as well as speculative aspects of market movements and fluctuations at times.

By combining these datasets, our model now has a more comprehensive and thorough understanding of market behavior and current global news and happenings, allowing it to work in both historical trends and investor sentiment, eventually providing improved predictive accuracy.

TABLE II
STOCK MARKET PRICE DATA (SAMPLE DATASET)

Date	Open	Close	Volume
2024-03-01	182.5	185.3	89M
2024-03-02	198.7	195.2	120M
2024-03-03	145.2	146.8	75M
2024-03-04	2720.5	2715.2	30M
2024-03-05	312.8	315.6	95M
2024-03-06	658.2	660.1	65M
2024-03-07	288.5	289.2	50M
2024-03-08	412.6	410.3	20M
2024-03-09	135.5	136.2	10M
2024-03-10	98.7	99.3	18M

TABLE III
STOCK MARKET SENTIMENT ANALYSIS (SAMPLE DATASET)

Date	News Headline	Sentiment
2024-03-01	Apple unveils AI chip	1
2024-03-02	Tesla faces delays	-1
2024-03-03	Amazon partners with SpaceX	1
2024-03-04	Google sued for antitrust	-1
2024-03-05	Microsoft revenue surges	1
2024-03-06	Nvidia launches new GPU	1
2024-03-07	Meta's project in trouble	-1
2024-03-08	Netflix growth slows	0
2024-03-09	IBM reveals quantum tech	1
2024-03-10	Oracle reports cloud growth	1

C. Model Training and Integration

The system is designed with three core predictive models: simple baseline-model of LSTM, XGBoost, and a hybrid of LSTM-XGBoost model where each contributes their unique strengths and features to help in stock market forecasting.

- **LSTM** works on gradient boosting, gathers temporal dependencies and long-term patterns in sequential stock price data, making it more effective for time-series forecasting.
- **XGBoost** it leverages decision tree-based learning to identify the key technical indicators and spots which improves short-term forecasting accuracy.
- The **hybrid model** combines the LSTM's deep learning potential with XGBoost's structured feature extraction which enchances predictive performance by balancing sequential learning with feature-driven insights.

Beyond these models, **anomaly detection** is integrated to filter out irregular market movements that could distort predictions. Additionally, **sentiment-based** adjustments refine predictions by incorporating financial news and social media sentiment, allowing the system to react to external market influences. This multi-layered integration ensures a robust, adaptive, and high-accuracy stock market prediction framework capable of handling real-world volatility.

III. IMPLEMENTATION

A. LSTM Model

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is designed to work on long-term dependencies in sequential data. It can particularly useful work for stock market prediction as it can learn and detect patterns in time-series data while handling issues like vanishing gradients. In our used model, LSTM is trained using the historical stock prices, which contain open, high, low, and close prices, and trading volume over a period of window size (for e.g., 30 days). This model learns temporal dependencies and then predicts the next-day closing price based on past trends and patterns.

Let's say we're training our LSTM model to predict the stock prices by providing an input from the last year and ask it to look at the last 30 days of prices and guess what the price will be on the 31st day. If the stock price has been going up for the last 30 days and more people are buying, the model might think, "Okay, this trend will probably continue," and predict another increase as the result. But if there's a sudden drop and prices are all over the place, the model might remember similar situations from the past and predict that the trend could reverse or the price might stabilize soon or might drop as well. This enhancement of the process shows how LSTM is flexible. It doesn't just follow the recent trend but considers a wider context. Because it can deal with the challenge of vanishing gradients as well, It keeps track of the imperative patterns even if they happened a while back that is what makes it more reliable for forecasting in a volatile market.

B. XGBoost Model

XGBoost is a machine-learning algorithm that uses decision trees to make predictions and works on gradient boosting, a machine learning technique that combines multiple weak models to create a more accurate predictive model. It works really well with organized data and runs faster by using smart tricks like parallel processing.

In our model, this helps find the most important details for predicting stock prices rather than processing all the entire historical data. This model uses simple processes such as Simple Moving Average (SMA), Exponential Moving Average (EMA), Previous Close Price, Trading Volume, and High-Low price changes. The model helps us to figure out which of these are the most useful and then focuses on them, so we don't have to waste time on unimportant stuff.

For example, say a trader wants to guess or predict the closing price of a stock for tomorrow or the day after tomorrow. This model looks at past data like SMA, EMA, and volume. If it sees that the 10-day EMA and trading volume matter the most, it pays attention to them and alerts them. If the EMA is going up and more people are buying, it predicts the price will go up too. But if fewer people are buying even when the EMA is going up, it might mean the price could drop soon.

C. Hybrid Approach

A hybrid LSTM-XGBoost model integrates the benifits of both the models:

- LSTM gathers the temporal dependencies in stock price transitions.
- XGBoost identifies and selects key technical spots or indicators to enhance feature engineering.

In our model, XGBoost first selects the most important features, which are then used to train the LSTM model. This hybrid approach makes sure that deep learning advantages from well-structured, optimized features, reducing computational complexity while improving predictive accuracy. In comparision to individual models, the combined approach achieves better generalization across different market conditions.

D. Sentiment Analysis

Stock prices aren't just about old price trends—they're also influenced by the sentiments people express based on news articles, financial reports, blogs, and even posts on social media. Sentiment analysis uses Natural Language Processing (NLP) to dig into all that messy text data to figure out what mood the market is in. I mean, it's kind of like trying to read the room and see if everyone's happy or bummed out about the stocks. And honestly, it's pretty cool how you can turn a bunch of words into a simple mood check.

We've used two models named VADER (Valence Aware Dictionary and Sentiment Reasoner) and TextBlob-based sentiment analysis model to sort the market news into positive, negative, or neutral sentiments. We've then convert these sentiments into numbers:

- 1 for positive sentiment (which means the news is good and could result in prices to go up)
- 0 for neutral sentiment (which shows mixed, neutral or not-so-impactful news)
- -1 for negative sentiment (which indicates bad news that might result in dropping of prices).

E. Anomaly Detection

We often experience sudden fluctuations in stock market due to external factors such as breaking news, economic events, shifts in supply and demand, or any natural calamity. Detecting these anomalies is crucial to improving the reliability of predictive models. We utilized the Isolation Forest algorithm which is a very widely used anomaly detection method which helps to identify unusual stock price movements by assigning anomaly scores to outliers.

To see these unusual changes clearly and have a deeper and better insights for the evaluation, we have represented the output in a dotted graph. The graph shows a clear link between the opening and closing prices. We marked any points that were very different from what we expected with red dots to show there might be something odd happening, like a big jump or drop in price.

It is pretty surprising to see these red dots when things don't go as expected.

For example, if a stock usually only fluctuates by about 2% in a day, but one day it drops by 15%, our model spots that as an unusual happening. Finding these uncanny changes helps us improve our model by reducing the effect of bigger fluctuations (sudden spike or dip), which makes our predictions more reliable in a changing market.

IV. Results and Evaluation

This section provides a detailed evaluation of model performance, showcasing prediction accuracy, market trend detection, and the overall impact of integrating multiple techniques for enhanced forecasting, highlighting improvements which is achieved through sentiment analysis and anomaly detection.

In the plot of LSTM (Fig. 2), the blue line represents the real stock prices, while the red line shows what our model predicted. Since the red line follows the blue line closely, it means that the model is doing a good job but still needs some improvement and training. Differnce in actual and predicted prices are fluctuating and can be refined more.(Table IV)

The XGBOOST graph (Fig. 3) shows the actual stock prices in blue and the predicted stock prices in red over time. The predicted prices closely aligned with the actual pricing, which shows that our model is performing well. But there are still areas of improvment in which more refining can be done.(Table V)

The hybrid model graph (Fig. 4) that compares actual and predicted stock prices over (blue and red, respectively). Some deviations exist, particularly during sudden price movements, but as we can get insights from the output table (VI), differnce here is very minimal as compared to the earlier standalone models.

In the sentiment analysis graph (Fig. 5), a higher positive sentiment score shows optimistic (positive) market sentiment,

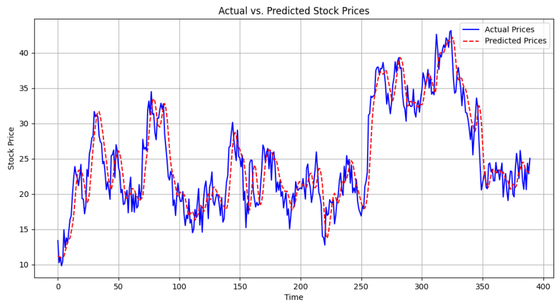

Fig. 2. LSTM Readings

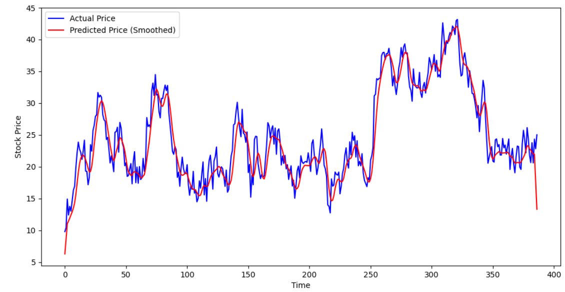

Fig. 3. XGBOOST Readings

which means that stock prices can rise. However, a negative mood reflects unfavorable news, which can lead to a price drop. The table below illustrates the relationship between sentiment scores and stock price predictions, while the accompanying graph provides a visual representation of the impact of sentiment.
After applying the sentiment analysis, we can clearly observe form the output table (VII)that predicted prices are now much more refined and aligned with the actual prices.

In the plot of Anomaly Detection (Fig. 6), the scatter plot shows a strong linear correlation between the open and close prices, indicating market stability. However, red-colored points represent anomalies where the Close price deviates significantly from the expected trend. These anomalies suggest unusual market movements, possibly due to external factors such as news events or sudden demand shifts.

TABLE IV
ACTUAL VS. PREDICTED PRICES (LSTM MODEL)

Time	Actual Price	Predicted Price	Difference
0	13.37	11.20	2.17
1	10.24	11.09	0.86
2	11.05	11.01	0.04
3	9.80	10.94	1.14
4	10.42	11.02	0.60
5	14.92	11.15	3.78
6	12.42	11.84	0.58
7	13.77	12.46	1.31
8	12.43	11.83	0.60
9	14.69	13.53	1.16

TABLE V
ACTUAL VS. PREDICTED PRICES (XGBOOST)

Time	Actual Price	Predicted Price	Difference
0	9.80	7.10	2.70
1	10.42	9.00	1.42
2	14.92	12.40	2.52
3	12.42	11.80	0.62
4	13.77	12.40	1.37
5	13.02	12.60	0.42
6	14.69	13.20	1.49
7	16.34	14.00	2.34
8	16.93	14.90	2.03
9	19.83	16.50	3.33

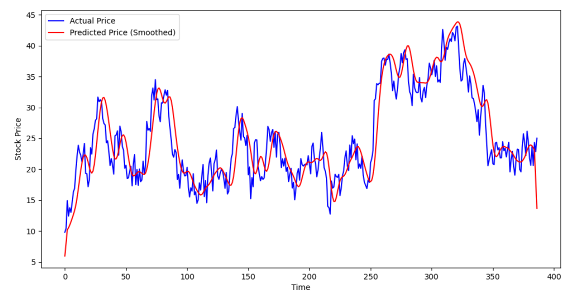

Fig. 4. Hybrid Readings

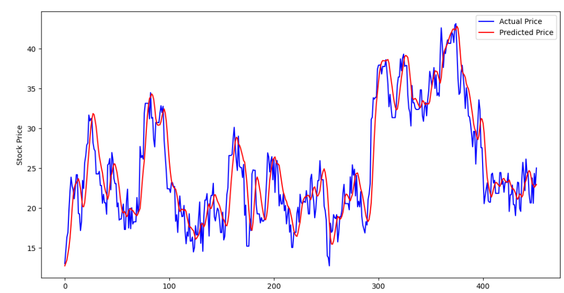

Fig. 5. Sentiment Analysis Readings

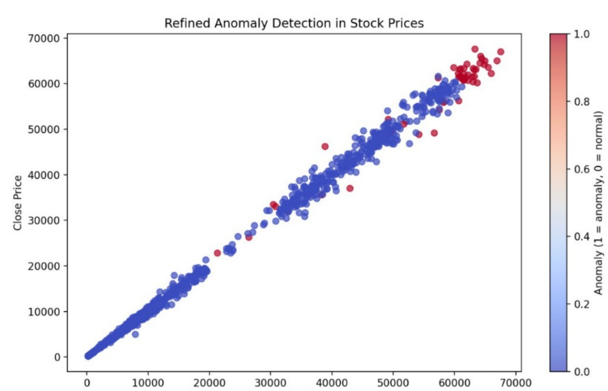

Fig. 6. Anamoly Detection Readings

TABLE VI
ACTUAL VS. PREDICTED PRICES (HYBRID LSTM + XGBOOST)

Time	Actual	Predicted	Difference
0	13.37	12.50	0.87
1	10.24	10.05	0.19
2	11.05	10.90	0.15
3	9.80	9.95	0.15
4	10.42	8.90	1.52
5	14.92	14.30	0.62
6	12.42	12.10	0.32
7	13.77	12.00	1.77
8	14.69	14.20	0.49
9	16.34	14.40	1.94

TABLE VII
SENTIMENT ANALYSIS IMPACT ON STOCK PRICES

Date	Sentiment	Score	Actual	Predicted
03-01	Positive	0.8	150.2	151.3
03-02	Neutral	0.1	149.8	150.1
03-03	Negative	-0.8	147.4	146.9
03-04	Positive	0.7	151.0	150.8
03-05	Negative	-0.4	148.5	149.0
03-06	Positive	0.9	152.1	152.5
03-07	Neutral	0.1	150.0	150.2
03-08	Negative	-0.6	148.1	147.8
03-09	Positive	0.8	151.9	152.0
03-10	Neutral	0.0	149.5	149.6

V. CONCLUSION AND FUTURE ENHANCEMENTS

This research shows a hybrid machine learning approach for stock market prediction by using LSTM as a baseline model and later integrating it with XGBoost along with anomaly detection and sentiment analysis for more refined forecasting. The combination of deep learning and ensemble methods enhances the predictive accuracy by capturing both temporal dependencies as well as complex financial trends and patterns. Sentiment analysis later strengthens the whole model by incorporating current market sentiments from news and social media blogs and posts, making forecasts more comprehensive. The results shows that the hybrid approach consistently outperforms standalone models in terms of accuracy, efficiency and adaptability, and robustness in volatile market environments.

Future works will focus on improving sentiment analysis using transformer-based NLP models like BERT and RoBERTa to gain additional contextual richness from financial text. Also, continuous learning approaches will be integrated to dynamically use the new market data while also including automatic retraining procedures. Model deployment using efficient methods such as TensorFlow Serving or AWS SageMaker can also help optimize the efficiency, accuracy, reliability, and scalability for real-time use. To increase performance in general, reinforcement learning-based trading strategies will be developed, allowing AI agents to mimic the market while searching for the e optimal trading moves. Expanding the dataset with macroeconomic parameters, commodities, and geopolitical risk factors will give it greater insight into broader financial patterns. In addition to being granted additional insights, incorporating techniques of explanation like as SHAP and LIME will cause investors to understand better the reasoning behind each prediction.

So, PeakPredict will be turning into a more intelligent, scalable, and data-driven forecasting system of investments, which consists of a higher reliability factor for investors functioning in arena of advanced capital markets.

REFERENCES

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[5] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.
[6] Shiller, R. J. (2014). Speculative asset prices. American Economic Review, 104(6), 1486-1517.
[7] Box, G. E., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.
[8] Murphy, J. J. (1999). Technical Analysis of the Financial Markets. Penguin.
[9] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.
[10] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock trends using machine learning algorithms. International Conference on Advanced Computing and Communication Technologies.
[11] Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.
[12] Aggarwal, C. C. (2017). Outlier analysis. Springer.
[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
[14] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16), 7653-7670.
[15] Faraz, M., Khaloozadeh, H., & Abbasi, M. (2020). Stock market prediction using autoencoder long short-term memory networks. In 28th Iranian Conference on Electrical Engineering (ICEE), IEEE.
[16] Liu, Y., Su, Z., Li, H., & Zhang, Y. (2019). An LSTM based classification method for time series trend forecasting. In 14th IEEE Conference on Industrial Electronics and Applications (ICIEA).
[17] Vopani, R. (2021). Nifty-50 Stock Market Data (2000–2021).
[18] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings,"arXiv preprint arXiv:1904.03323 (2019).
[19] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063 (2019).
[20] M. Komorowski, D. Marshall, J. Salciccioli, and Y. Crutain, "Exploratory data analysis," (2016) pp. 185–203.
[21] F. Mosteller and J. W. Tukey, "Data analysis and regression. a second course in statistics," Addison-Wesley series in behavioral science: quantitative methods (1977).