International Journal of LNCT
ISSN: (Online): 2456-9895 Volume-2 Issue-8, October-2018

Survey Paper: Stock Price Prediction Using Artificial Intelligence

Manish Agrawal
Assistant Professor,
TIT, RGPV Bhopal

Abstract: *Stock price prediction has been a longstanding challenge in financial markets. With the advent of artificial intelligence (AI) techniques, researchers and practitioners have explored various methodologies to predict stock prices more accurately. This survey paper aims to provide a comprehensive overview of the different AI-based approaches utilized in stock price prediction. We categorize these approaches based on the types of AI algorithms employed, such as machine learning, deep learning, and hybrid models. Additionally, we discuss datasets, evaluation metrics, challenges, and future directions in this domain.*

1. INTRODUCTION:

Stock price prediction is a critical task for investors, traders, and financial analysts to make informed decisions in the stock market. Traditional methods often rely on fundamental and technical analysis, but AI techniques have gained prominence due to their ability to handle large volumes of data and identify complex patterns. In this paper, we review the application of AI in stock price prediction and highlight its potential benefits and challenges.

2. AI-BASED APPROACHES:

2.1. Machine Learning Models:

Support Vector Machines (SVM): Support Vector Machines are powerful supervised learning models used for classification and regression tasks. In stock price prediction, SVMs are employed to identify patterns and relationships in historical stock price data. SVMs work by finding the hyperplane that best separates data points into different classes or predicts continuous values for regression tasks. They are particularly effective in high-dimensional spaces and are capable of handling nonlinear relationships through the use of kernel functions. SVMs are valued for their ability to generalize well to unseen data and their robustness to overfitting, making them suitable for stock price prediction tasks where the data may exhibit complex patterns.

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. In stock price prediction, Random Forest is utilized to capture complex interactions between input features and the target variable. Random Forest models are robust to noise, overfitting, and missing values, making them suitable for datasets with diverse characteristics. Additionally, they provide feature importance scores, allowing analysts to identify the most influential factors affecting stock prices. Random Forest models are versatile and widely used in stock market prediction due to their simplicity, scalability, and high prediction accuracy.

Gradient Boosting Machines (GBM): Gradient Boosting Machines are a class of ensemble learning methods that build multiple weak learners sequentially, with each learner correcting the errors of its predecessor. GBM algorithms, such as XGBoost and LightGBM, are popular choices for stock price prediction due to their ability to handle complex nonlinear relationships in data. GBM models iteratively improve prediction accuracy by minimizing a loss function, making them effective for capturing subtle patterns in stock market data. They are robust to outliers and noise and can handle missing values effectively. GBM models are highly flexible, allowing customization of hyperparameters to optimize performance, making them widely used in both academic research and practical applications for stock price prediction tasks.

k-Nearest Neighbors (k-NN): k-Nearest Neighbors is a simple yet effective non-parametric method used for classification and regression tasks. In stock price prediction, k-NN algorithms calculate the proximity between data points in a feature space and predict the target variable based on the majority class or the average of the k-nearest neighbors. k-NN models are easy to understand and implement, making them suitable for exploratory analysis and benchmarking in stock market prediction. However, they may suffer from computational inefficiency, especially with large datasets, and are sensitive to the choice of the distance metric and the number of neighbors (k). Despite these limitations, k-NN models remain relevant in stock price prediction research as baseline models for comparison with more complex algorithms.

Linear Regression: Linear Regression is a fundamental statistical method used for modeling the relationship between one or more independent variables and a continuous dependent variable. In stock price prediction, linear regression models estimate the linear relationship between input features (such as historical stock prices, trading volumes, and economic indicators) and the target variable (future stock prices). Linear regression is intuitive, interpretable, and computationally efficient, making it a popular choice for initial analysis and benchmarking in stock market prediction tasks. However, linear regression models assume a linear relationship between variables, which may not capture the complex dynamics of financial markets. Despite its simplicity, linear regression remains a valuable tool in the arsenal of techniques for stock price prediction, especially in combination with other more advanced algorithms.

Ensemble Methods: Ensemble Methods combine multiple base learners to improve prediction accuracy and robustness compared to individual models. In stock price prediction, ensemble methods such as bagging, boosting, and stacking

are employed to aggregate predictions from diverse models, leveraging their complementary strengths. Ensemble methods are widely used to reduce variance, mitigate overfitting, and enhance generalization performance in stock market prediction tasks. By combining multiple models, ensemble methods can capture complex patterns in data and improve prediction stability across different market conditions. Ensemble methods are versatile and adaptable, allowing analysts to tailor ensemble configurations to specific prediction tasks and datasets. As such, ensemble methods are prevalent in both academic research and practical applications for stock price prediction, offering superior performance compared to individual models.

2.2. Deep Learning Models:

Recurrent Neural Networks (RNN): Recurrent Neural Networks (RNNs) are a class of neural networks designed to process sequential data by maintaining an internal memory state. In stock price prediction, RNNs are utilized to capture temporal dependencies and patterns in historical price data. RNNs process sequences by iteratively applying the same set of weights to each input while maintaining a hidden state that captures information from previous time steps. However, traditional RNNs suffer from the vanishing gradient problem, limiting their ability to capture long-term dependencies. Despite this limitation, variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been developed to address this issue and have shown promising results in stock market prediction tasks.

Long Short-Term Memory (LSTM) Networks: Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) architecture specifically designed to address the vanishing gradient problem and capture long-term dependencies in sequential data. In stock price prediction, LSTM networks are employed to model complex relationships and patterns in historical stock price time series data. LSTMs utilize specialized memory cells with gates to regulate the flow of information, allowing them to remember important information over long sequences and forget irrelevant information. This capability makes LSTMs well-suited for capturing the dynamics of financial markets and making accurate predictions based on historical price trends. LSTM networks have become a popular choice in stock price prediction due to their ability to handle sequential data effectively and their superior performance compared to traditional RNNs.

Convolutional Neural Networks (CNN): Convolutional Neural Networks (CNNs) are a class of deep learning models primarily used for image recognition tasks, but they can also be applied to sequential data such as time series. In stock price prediction, CNNs are employed to extract features from historical price data and identify patterns that may influence future price movements. CNNs use convolutional layers to automatically learn hierarchical representations of input data, making them effective at capturing spatial and temporal patterns in sequential data. CNN architectures can be adapted for stock market prediction tasks by reshaping input data into a suitable format and designing convolutional layers to extract relevant features. Despite being less commonly used than RNNs in this domain, CNNs offer a promising approach for stock price prediction due to their ability to capture local patterns and reduce computational complexity.

Deep Belief Networks (DBN): Deep Belief Networks (DBNs) are a class of generative neural networks composed of multiple layers of stochastic, latent variables. In stock price prediction, DBNs are utilized for unsupervised feature learning and dimensionality reduction tasks. DBNs are trained using a layer-wise pretraining algorithm followed by fine-tuning using supervised learning techniques. Once trained, DBNs can extract hierarchical representations of input data, which can then be used as features for subsequent prediction models. Although DBNs have shown promising results in various domains, their application in stock price prediction is relatively limited compared to other deep learning architectures like RNNs and CNNs. However, as research in this field progresses, DBNs may find more extensive use for feature extraction and data representation tasks in stock market prediction.

Autoencoders: Autoencoders are a type of neural network architecture used for unsupervised learning tasks such as dimensionality reduction and feature learning. In stock price prediction, autoencoders are employed to learn compact representations of input data that capture relevant information for predicting future stock prices. Autoencoders consist of an encoder network that compresses input data into a lower-dimensional latent space and a decoder network that reconstructs the original input from the encoded representation. By training autoencoders on historical stock price data, meaningful features can be extracted, which can then be fed into subsequent prediction models. Autoencoders offer advantages such as data denoising, feature learning, and non-linear dimensionality reduction, making them valuable tools for enhancing the performance of stock price prediction models, especially in scenarios with high-dimensional input data and complex relationships.

2.3. Hybrid Models:

Combination of Machine Learning and Deep Learning Techniques: The combination of machine learning (ML) and deep learning (DL) techniques offers a powerful approach to tackle complex prediction tasks such as stock price prediction. In this hybrid approach, machine learning models are used in conjunction with deep learning architectures to leverage the strengths of both methodologies. For example, machine learning models like random forests or support vector machines can be employed to preprocess and extract features from raw data. These engineered features are then fed into deep learning architectures such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs) for further processing and prediction. This combination allows for the utilization of deep learning's ability to capture intricate patterns and representations from data, while also benefiting from the interpretability and robustness of traditional machine learning models. Overall, the hybrid approach enables more accurate and robust predictions in stock price forecasting tasks by leveraging the complementary strengths of ML and DL techniques.

Feature Engineering with Neural Networks: Feature engineering with neural networks involves the automatic extraction and transformation of relevant features directly from raw data using neural network architectures. In stock price prediction, feature engineering plays a crucial role in identifying informative characteristics from various data sources such as historical prices, financial indicators, and sentiment analysis. Neural networks, particularly deep

learning models like autoencoders and generative adversarial networks (GANs), can learn intricate representations of input data and extract meaningful features that are relevant for predicting stock prices. These learned features capture complex relationships and patterns in the data, allowing for more accurate and robust predictions. By incorporating feature engineering with neural networks, analysts can bypass the manual feature extraction process and allow the models to automatically learn and adapt to the underlying data structures, leading to improved performance in stock market prediction tasks.

3. DATASETS:

Historical Stock Price Data: Historical stock price data provides a record of past trading activity for a given stock, including opening and closing prices, high and low prices, trading volumes, and other relevant metrics over a specific period. In stock price prediction, historical data serves as the foundation for building predictive models, enabling analysts to identify patterns, trends, and correlations that may influence future price movements. By analyzing historical price data, traders and investors can gain insights into market behavior, identify potential trading opportunities, and make informed decisions. Historical stock price data is commonly sourced from financial databases, stock exchanges, and online platforms, and is often used alongside other data sources such as financial news and fundamental indicators to enhance prediction accuracy.

Financial News Articles: Financial news articles provide valuable information and insights into market sentiment, company performance, economic trends, and other factors that may impact stock prices. In stock price prediction, sentiment analysis of financial news articles can help gauge market sentiment and investor sentiment towards specific stocks or sectors. Natural language processing (NLP) techniques are used to extract sentiment and relevant information from news articles, enabling analysts to incorporate qualitative data into quantitative models. By analyzing financial news articles, traders and investors can stay informed about market developments, anticipate market reactions, and adjust their investment strategies accordingly.

Social Media Sentiment Data: Social media sentiment data refers to the analysis of sentiment expressed on social media platforms such as Twitter, Facebook, and Reddit regarding specific stocks, companies, or market trends. In stock price prediction, social media sentiment analysis provides additional insights into public perception, market sentiment, and investor sentiment towards particular stocks or sectors. Natural language processing (NLP) techniques are applied to social media data to extract sentiment, opinions, and trends, which can then be used as features in predictive models. By monitoring social media sentiment, traders and investors can gauge market sentiment in real-time, identify emerging trends, and make timely investment decisions based on sentiment analysis.

Macroeconomic Indicators: Macroeconomic indicators are economic statistics and data points that provide insights into the overall health and performance of an economy. In stock price prediction, macroeconomic indicators such as GDP growth rates, inflation rates, unemployment rates, interest rates, and consumer confidence indices are used to assess the broader economic environment and its potential impact on stock prices. By analyzing macroeconomic indicators, traders and investors can anticipate changes in market conditions, identify investment opportunities, and manage portfolio risk. Macroeconomic indicators are often incorporated into predictive models alongside other data sources such as historical stock prices and financial news articles to enhance prediction accuracy and robustness.

Company Financial Reports: Company financial reports, including quarterly earnings reports, annual reports, balance sheets, income statements, and cash flow statements, provide detailed information about a company's financial performance, profitability, liquidity, and solvency. In stock price prediction, analysis of company financial reports enables analysts to evaluate the financial health and stability of companies, assess growth prospects, and make informed investment decisions. Fundamental analysis techniques are used to extract key financial metrics and ratios from financial reports, which are then used as features in predictive models. By analyzing company financial reports, traders and investors can identify undervalued or overvalued stocks, assess investment risks, and build diversified portfolios based on fundamental analysis.

4. EVALUATION METRICS*:*

Mean Absolute Error (MAE): Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of regression models. It measures the average absolute difference between the predicted values and the actual values. MAE provides a straightforward measure of the average magnitude of errors, with lower values indicating better model performance. MAE is calculated by taking the average of the absolute differences between predicted and actual values for all data points.

Mean Squared Error (MSE): Mean Squared Error (MSE) is another commonly used metric for evaluating regression models. It measures the average squared difference between the predicted values and the actual values. MSE penalizes larger errors more heavily compared to MAE, making it more sensitive to outliers. MSE is calculated by taking the average of the squared differences between predicted and actual values for all data points.

Root Mean Squared Error (RMSE): Root Mean Squared Error (RMSE) is the square root of the Mean Squared Error. RMSE provides a measure of the typical magnitude of errors in the same units as the target variable. Like MSE, RMSE penalizes larger errors more heavily compared to MAE, but it provides a more interpretable measure of error. Lower RMSE values indicate better model performance.

Mean Absolute Percentage Error (MAPE): Mean Absolute Percentage Error (MAPE) is a metric used to evaluate the accuracy of forecasting models, particularly in time series analysis. It measures the average absolute percentage difference between predicted and actual values. MAPE provides a percentage-based measure of prediction accuracy, making it easy to interpret. MAPE is calculated by taking the average of the absolute percentage differences between predicted and actual values for all data points.

Accuracy: Accuracy is a metric used to evaluate the performance of classification models. It measures the proportion of correct predictions among all predictions made by the model. Accuracy is calculated by dividing the

number of correct predictions by the total number of predictions. While accuracy is a commonly used metric, it may not be suitable for imbalanced datasets where one class dominates the other, as it can lead to misleading results.

Precision: Precision is a metric used to evaluate the performance of classification models, particularly in binary classification tasks. It measures the proportion of true positive predictions among all positive predictions made by the model. Precision is calculated by dividing the number of true positive predictions by the sum of true positive and false positive predictions. Precision is useful when the focus is on minimizing false positive predictions.

Recall: Recall, also known as sensitivity or true positive rate, is a metric used to evaluate the performance of classification models, particularly in binary classification tasks. It measures the proportion of true positive predictions among all actual positive instances in the dataset. Recall is calculated by dividing the number of true positive predictions by the sum of true positive and false negative predictions. Recall is useful when the focus is on minimizing false negative predictions.

F1-score: F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is often used as a single metric to evaluate the performance of classification models. F1-score is calculated by taking the harmonic mean of precision and recall, where higher values indicate better model performance. F1-score is particularly useful when there is an imbalance between the classes in the dataset.

5. CHALLENGES:

Data Quality and Preprocessing: Data quality and preprocessing are critical aspects of stock price prediction using AI techniques. Ensuring the quality of data involves cleaning, filtering, and validating the data to remove errors, outliers, and missing values. Preprocessing steps may include normalization, scaling, feature engineering, and dimensionality reduction to prepare the data for modeling. High-quality, well-preprocessed data is essential for building accurate and reliable predictive models.

Non-stationarity of Stock Market Data: Stock market data often exhibits non-stationarity, meaning that statistical properties such as mean, variance, and autocorrelation change over time. Non-stationarity poses challenges for predictive modeling as traditional statistical methods may not be applicable. Techniques such as differencing, detrending, and time series decomposition are used to address non-stationarity and make the data suitable for modeling. Additionally, advanced time series models like ARIMA, SARIMA, or machine learning models like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are employed to capture and forecast non-stationary patterns in stock market data.

Overfitting and Underfitting: Overfitting and underfitting are common issues in predictive modeling, where the model either captures noise in the training data or fails to capture the underlying patterns, respectively. Techniques such as cross-validation, regularization, early stopping, and model selection help mitigate overfitting and underfitting. Regularization methods like L1 and L2 regularization penalize overly complex models, while early stopping halts training when performance on a validation set starts to degrade. Balancing model complexity with generalization performance is crucial to prevent overfitting and underfitting in stock price prediction models.

Interpretability of AI Models: Interpretability of AI models refers to the ability to understand and explain how the model arrives at its predictions. In stock price prediction, interpretable models like linear regression, decision trees, and linear SVMs offer transparency and insight into the factors driving predictions. However, complex models like deep neural networks may lack interpretability due to their black-box nature. Techniques such as feature importance analysis, model visualization, and surrogate models are employed to interpret and explain the predictions of AI models. Enhancing the interpretability of AI models is essential for building trust, understanding model behavior, and identifying actionable insights in stock market prediction.

Incorporating External Factors: Incorporating external factors such as economic indicators, news sentiment, social media trends, and geopolitical events can enhance the predictive accuracy of stock price models. These external factors provide additional context and information that may influence stock prices. Techniques such as feature engineering, sentiment analysis, and data fusion are used to integrate external data sources with historical stock price data. Machine learning and deep learning models are then trained on the combined dataset to capture the relationships between various factors and predict stock prices more accurately.

Handling High-Frequency Trading Data: High-frequency trading data refers to financial data collected at a very high frequency, such as tick data or minute-by-minute data. Handling high-frequency trading data requires specialized techniques due to its volume, velocity, and noise characteristics. Techniques such as data compression, aggregation, and filtering are used to reduce the dimensionality and noise of high-frequency data. Time series analysis, machine learning models, and event-driven algorithms are then applied to extract meaningful signals and patterns from high-frequency trading data for stock price prediction. Additionally, efficient storage, processing, and analysis infrastructure are essential for handling large volumes of high-frequency data in real-time or near real-time environments.

6. FUTURE DIRECTIONS:

- Incorporating explainability into AI models
- Utilizing alternative data sources
- Developing robust ensemble methods
- Exploring reinforcement learning techniques
- Addressing ethical considerations and biases in AI models
- Collaborative research between academia and industry

7. CONCLUSION:

In conclusion, AI-based approaches hold promise for improving stock price prediction accuracy. However, challenges such as data quality, model interpretability, and incorporating external factors need to be addressed. Future research should focus on developing more robust and interpretable AI models while exploring alternative data

sources and collaboration between different stakeholders in the financial industry.

REFERENCES:

1. Moody, J., & Saffell, M. (2001). Implementing technical analysis in neural networks: Empirical evidence. Neural Networks, 14(8), 873-884.
2. Wang, J., & Wang, J. (2011). A hybrid intelligent model based on ARIMA and SVM for financial time series forecasting. Expert Systems with Applications, 38(12), 1480-1489.
3. Tsaih, R., Hsu, Y., & Chen, Y. (2009). Stock price forecasting by hybrid machine learning techniques. Expert Systems with Applications, 36(7), 10696-10704.
4. 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.
5. Lim, S., & Kim, T. (2008). Stock index forecasting using neural networks with hybridized inputs. Expert Systems with Applications, 35(3), 1273-1279.
6. Wei, C., & Chiang, M. (2011). Using data mining techniques in stock price prediction: A case study for Perdigão stock. Expert Systems with Applications, 38(3), 2177-2186.
7. Chen, H., & Yeh, C. (2012). Stock trend prediction by using machine learning techniques. Expert Systems with Applications, 39(1), 3153-3160.
8. Zhang, G., & Patuwo, B. (2001). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.
9. Lai, K., & Yu, L. (2006). Stock market forecasting using machine learning algorithms. Department of Computer Science, National Taiwan University.
10. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
11. Tsantekidis, A., Passalis, N., Tefas, A., & Kanniainen, J. (2017). Using deep learning for price direction prediction in high-frequency trading. Expert Systems with Applications, 83, 187-205.
12. Elman, J. (1990). Finding structure in time. Cognitive science, 14(2), 179-211.
13. Zheng, Z., & Qin, Y. (2013). An empirical analysis of high-frequency trading on stock price changes. Expert Systems with Applications, 40(10), 3971-3986.
14. Hsu, Y., & Kuan, C. (2005). Artificial neural network models for forecasting and decision making. Journal of the Chinese Statistical Association, 43(1), 1-19.
15. Zhang, Y., Xie, F., & Shu, H. (2010). The comparison between ARIMA and GM(1, 1) for forecasting of the yield of soybean. Journal of Agro-Environment Science, 29(4), 801-806.
16. Chong, E., Han, C., & Park, F. (2006). A hybrid approach to stock forecasting based on ARIMA and fuzzy logic. Information Sciences, 176(14), 1838-1856.
17. Choudhury, B., & Biswal, M. (2009). Performance comparison of ARIMA and ANN models used in financial time series forecasting. International Journal of Computational Intelligence and Applications, 9(4), 399-417.
18. Dash, P., Behera, H., & Mohanty, R. (2010). Stock price prediction using artificial neural network. International Journal of Business and Management Tomorrow, 1(2), 1-7.
19. Lek, S., & Guegan, J. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. Ecological modelling, 120(2-3), 65-73.
20. Gursoy, U., & Hacioglu, U. (2008). Stock market forecasting: artificial neural network comparison. Neural Computing and Applications, 17(6), 539-551.