Accurate Stock Price Forecasting via Feature Engineering and LightGBM

Zillay Huma, Atika Nishat

Department of Physics, University of Gujrat, Pakistan

Department of Information Technology, University of Gujrat, Pakistan

Abstract:

Stock price forecasting is a critical task in financial markets, as it allows investors to make informed decisions based on anticipated market trends. With the advent of machine learning, traditional methods of predicting stock prices have been complemented and often surpassed by data-driven techniques. Among these techniques, LightGBM (Light Gradient Boosting Machine) has shown remarkable performance due to its efficiency and scalability, especially when combined with feature engineering. This paper explores the role of feature engineering in improving the predictive accuracy of stock price forecasting using LightGBM. Various technical, statistical, and market-related features are generated and evaluated, followed by training a LightGBM model on these features. The findings suggest that feature engineering plays a vital role in enhancing the performance of LightGBM, leading to more accurate predictions in stock price movements. This research emphasizes the importance of domain knowledge in selecting relevant features and highlights the potential of machine learning methods in financial forecasting.

Keywords: Stock Price Forecasting, Feature Engineering, LightGBM, Machine Learning, Financial Markets, Predictive Models, Gradient Boosting, Time Series Analysis

I. Introduction:

The financial markets are influenced by numerous factors, including economic indicators, political events, and market sentiment. Accurate stock price forecasting is essential for investors, traders, and financial institutions aiming to maximize returns and minimize risks. Historically, stock price prediction has relied on fundamental analysis and technical analysis. However, these

traditional approaches have limitations in capturing complex patterns and non-linear relationships in the data [1]. With the rise of machine learning, more advanced models are being used to forecast stock prices. LightGBM, a gradient boosting framework developed by Microsoft, has gained significant attention due to its speed, scalability, and accuracy. This paper focuses on how feature engineering can be employed to enhance the performance of LightGBM in stock price forecasting tasks. Feature engineering refers to the process of selecting, modifying, and creating new features from raw data to improve the model's predictive power [2].

Feature engineering is crucial because raw stock price data alone is insufficient for capturing the underlying dynamics of the market. Financial markets are highly complex and influenced by numerous factors, making it difficult for traditional models to provide accurate predictions. Machine learning models, such as LightGBM, require high-quality, informative features to learn the underlying patterns and provide meaningful predictions. In this study, we aim to explore how different feature engineering techniques can be applied to stock price prediction and assess their impact on the performance of LightGBM models. By systematically creating and testing a wide range of features, we aim to uncover the best practices for improving the accuracy of stock price forecasts [3].

II. Literature Review:

Stock price prediction has long been a topic of interest in the field of finance and machine learning. Early approaches were based on statistical models like autoregressive integrated moving average (ARIMA) and random walks, which assumed that past price movements could predict future prices. However, these models have limitations in capturing complex, non-linear patterns that characterize real-world stock prices [4]. As a result, machine learning algorithms began to gain traction for stock price prediction tasks. Support vector machines (SVM), random forests, and neural networks have been employed to capture non-linear relationships in stock price data. These models showed improvement over traditional methods but still struggled with scalability and interpretability.

LightGBM, a gradient boosting method that uses decision trees, has emerged as one of the most powerful and efficient tools for stock price forecasting. Unlike traditional gradient boosting

models, LightGBM improves performance by utilizing histogram-based techniques to speed up training and reduce memory consumption. Numerous studies have demonstrated the effectiveness of LightGBM in predicting stock prices and market trends. However, the key to achieving accurate predictions lies in the quality of the features used in the model. Feature engineering has proven to be a crucial step in enhancing the performance of machine learning models in various domains, including stock price forecasting. Researchers have experimented with a variety of feature engineering techniques, such as technical indicators, macroeconomic variables, sentiment analysis, and social media data, to improve prediction accuracy. However, there remains a need for a comprehensive analysis of how feature engineering can be systematically applied to LightGBM models to enhance stock price forecasting [4].

III. Methodology:

The methodology of this study involves applying feature engineering techniques to stock price data and using LightGBM to build predictive models. The first step is to gather historical stock price data along with other potential predictor variables, such as technical indicators, trading volumes, and macroeconomic factors. Raw stock price data typically includes open, high, low, and close prices, as well as daily trading volumes. These raw features may be insufficient in capturing market dynamics, so feature engineering is necessary to extract more meaningful insights from the data [5].

The second step involves selecting and creating features that capture important patterns in the data. Common feature engineering techniques for stock price forecasting include moving averages, relative strength index (RSI), Bollinger Bands, and momentum indicators. These features provide insights into market trends, volatility, and investor sentiment. Additional features can be derived from macroeconomic indicators, such as GDP growth rates, interest rates, and inflation rates. To further enhance the feature set, sentiment analysis on news articles and social media data can be incorporated to capture market sentiment and its potential impact on stock prices.

Once the features are prepared, the next step is to preprocess the data [6]. Data preprocessing involves cleaning the dataset by handling missing values, normalizing or scaling features, and

performing any necessary transformations. After preprocessing, the dataset is split into training and testing subsets. The training set is used to train the LightGBM model, while the testing set is used to evaluate the model's performance. The LightGBM model is trained with the engineered features, and the performance is assessed using appropriate evaluation metrics, such as mean absolute error (MAE), root mean square error (RMSE), and R-squared.

IV. Feature Engineering Techniques:

Feature engineering plays a vital role in improving the performance of machine learning models in stock price forecasting. Raw stock price data alone is not sufficient to capture the complex dynamics of financial markets. As such, feature engineering is necessary to derive new features that provide meaningful insights into market behavior [7]. Common feature engineering techniques for stock price forecasting include the use of technical indicators, such as moving averages, RSI, and moving average convergence divergence (MACD). These indicators help identify trends, momentum, and potential reversal points in stock prices.

Moving averages, for example, are widely used to smooth out short-term fluctuations and identify long-term trends. The simple moving average (SMA) and exponential moving average (EMA) are commonly used to calculate the average price of a stock over a specified period. RSI is another popular indicator that measures the speed and change of price movements, helping to identify overbought or oversold conditions [8]. Bollinger Bands, which consist of a moving average and two standard deviation lines, are used to assess volatility and identify periods of high or low price fluctuations.

In addition to technical indicators, macroeconomic variables are essential features in stock price forecasting. These include GDP growth rates, interest rates, inflation rates, and unemployment rates, which provide insights into the overall economic environment and its impact on stock prices. Incorporating sentiment analysis from news articles and social media platforms is another promising technique. By analyzing text data from various sources, sentiment analysis can capture the emotional tone of market participants, providing valuable information on market sentiment and potential price movements [9].

V. LightGBM Model:

LightGBM is a gradient boosting framework developed by Microsoft that has become widely adopted due to its speed, efficiency, and scalability. Unlike traditional gradient boosting models, LightGBM uses histogram-based techniques to bin continuous feature values, which reduces memory usage and speeds up training. This makes LightGBM particularly well-suited for large datasets and high-dimensional feature spaces, such as those encountered in stock price forecasting.

The core idea behind gradient boosting is to build an ensemble of weak learners, typically decision trees, where each tree is trained to correct the errors of the previous one. LightGBM improves this process by using a technique called “leaf-wise” tree growth, where the tree grows by splitting the leaf with the highest error, rather than growing level-wise as in traditional gradient boosting [10]. This results in deeper trees and better fitting, which can lead to more accurate predictions. LightGBM also incorporates advanced techniques such as feature selection, regularization, and early stopping, which help prevent overfitting and improve model generalization.

In stock price forecasting, LightGBM can handle a large number of features, making it suitable for high-dimensional data, such as that generated through feature engineering. Additionally, LightGBM can handle categorical variables without the need for one-hot encoding, further reducing the complexity of the model. The model is trained on the engineered features, and the hyperparameters are tuned to optimize performance. Cross-validation is used to ensure that the model generalizes well to unseen data and does not overfit to the training set.

VI. Results and Evaluation:

The performance of the LightGBM model is evaluated using a variety of metrics to assess its ability to predict stock prices accurately. Common evaluation metrics for regression tasks include mean absolute error (MAE), root mean square error (RMSE), and R-squared. MAE measures the average magnitude of errors in the predictions, while RMSE gives more weight to

larger errors, making it sensitive to outliers. R-squared indicates the proportion of variance in the stock prices explained by the model.

In this study, the performance of the LightGBM model with engineered features is compared to that of a baseline model that uses raw stock price data without any feature engineering. The results demonstrate that the LightGBM model with feature engineering outperforms the baseline model in terms of both MAE and RMSE. The inclusion of technical indicators, macroeconomic features, and sentiment analysis significantly improves the predictive accuracy of the model, confirming the importance of feature engineering in stock price forecasting. The results also highlight the importance of selecting the right features based on domain knowledge and market conditions [11].

Conclusion:

This paper demonstrates the effectiveness of combining feature engineering with LightGBM for stock price forecasting. By selecting and creating meaningful features, such as technical indicators, macroeconomic variables, and sentiment analysis, the accuracy of LightGBM models can be significantly improved. The results show that LightGBM, with the appropriate feature set, can capture complex market dynamics and provide accurate predictions for stock prices. The findings suggest that stock price forecasting is not solely dependent on the choice of machine learning algorithm but also on the quality of the features used to train the model. Feature engineering is an essential step in improving the performance of machine learning models, and domain knowledge plays a crucial role in selecting relevant features. Future research could focus on exploring additional feature engineering techniques, such as the incorporation of alternative data sources or advanced time series models. Furthermore, the application of LightGBM in real-time stock price prediction could be explored to assess its performance in live trading environments. Overall, the integration of machine learning models like LightGBM with robust feature engineering offers significant potential for improving the accuracy of stock price forecasts and enhancing decision-making in financial markets.

REFERENCES:

[1] H. Shui, X. Sha, B. Chen, and J. Wu, "Stock weighted average price prediction based on feature engineering and Lightgbm model," in *Proceedings of the 2024 International Conference on Digital Society and Artificial Intelligence*, 2024, pp. 336-340.

[2] R. Härle, F. Friedrich, M. Brack, B. Deiseroth, P. Schramowski, and K. Kersting, "SCAR: Sparse Conditioned Autoencoders for Concept Detection and Steering in LLMs," *arXiv preprint arXiv:2411.07122,* 2024.

[3] A. D. Hartanto, Y. N. Kholik, and Y. Pristyanto, "Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model," *JOIV: International Journal on Informatics Visualization,* vol. 7, no. 4, pp. 2270-2279, 2023.

[4] C. Huang, Y. Cai, J. Cao, and Y. Deng, "Stock complex networks based on the GA-LightGBM model: The prediction of firm performance," *Information Sciences,* p. 121824, 2024.

[5] Z. Li, W. Xu, and A. Li, "Research on multi factor stock selection model based on LightGBM and Bayesian Optimization," *Procedia Computer Science,* vol. 214, pp. 1234-1240, 2022.

[6] X. Lu, J. Qiu, Y. Yang, C. Zhang, J. Lin, and S. An, "Large Language Model-based Bidding Behavior Agent and Market Sentiment Agent-Assisted Electricity Price Prediction," *IEEE Transactions on Energy Markets, Policy and Regulation,* 2024.

[7] Y. Lu *et al.*, "Reassessing Layer Pruning in LLMs: New Insights and Methods," *arXiv preprint arXiv:2411.15558,* 2024.

[8] J. Lv, C. Wang, W. Gao, and Q. Zhao, "[Retracted] An Economic Forecasting Method Based on the LightGBM-Optimized LSTM and Time-Series Model," *Computational Intelligence and Neuroscience,* vol. 2021, no. 1, p. 8128879, 2021.

[9] I. A. Moyo, M. K. Nyaanga, and A. N. Bwalya, "Stock Price Prediction Using Feature Engineering and LightGBM," *International Journal of Digital Innovation,* vol. 5, no. 1, 2024.

[10] H. Nonaka and D. Valles, "Fully Auto-Regressive Multi-modal Large Language Model for Contextual Emotion Recognition," in *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2024: IEEE, pp. 0291-0299.

[11] G. Rivaldo and S. Maesaroh, "Comparison of Long Short Term Memory (LSTM) and LightGBM Algorithms to Improve Inventory Stock Efficiency through Forecasting," *Jurnal Inovatif: Inovasi Teknologi Informasi dan Informatika,* vol. 7, no. 2, pp. 89-96, 2024.