Stock price Prediction using Hybrid LSTM & XGBoost Model

Deng Li, Xinhao Zhou

1 INTRODUCTION

Stock is a fundamental component of the finance system, and an essential part of personal wealth. Prediction of stock prices using deep neural networks classification models has been developed very well to help investors to gain financial benefits[4]. However, stock prices comprise a sequence of data points arranged in a chronological order, necessitating their analysis as time series data. Stocks also contain strong trend-related characteristics that are essential for price prediction.

In this project, we will use two classic models of machine learning, LSTM[5], which is based on recurrent neural networks, and XGBoost[1], which is based on boosting, to predict the Apple (AAPL) stock price in 10 years period (from 01/01/2013 to 01/01/2023). In order to reveal generosity, we will use historical open-end high-low price of stocks and auction book data (volume), which are directly price-correlated. After that, we will expand our experiment to a bucket of stocks and a longer period of time to test the overall performance and robustness of our model.

The scope our study is to introducing the functionality of LSTM and XGBoost in stock price prediction, and further construct an advanced ensemble model that leverage the advantage of both models combine them to have an accurate, efficient, and robust performance on the stock price prediction task.

2 LITERATURE REVIEW

2.1 LSTM

Long Short-Term Memory[2] is a type of recurrent neural network (RNN) architecture used in the field of deep learning. Unlike traditional RNNs, LSTMs are designed to avoid the long-term dependency problem, making them effective for learning from sequences of data where there are long gaps of relevance. They accomplish this through a unique configuration of gates that regulate the flow of information. These gates enable LSTMs to selectively retain or discard information over extended periods, contributing to their robust memory capabilities.

Contrasting with the single state in traditional RNNs, the hidden layer of LSTMs comprises two distinct states: one that is highly responsive to short-term input data (the hidden state), and another, known as the cell state, that maintains long-term data changes. This dual-state architecture is central to the LSTM's proficiency in managing long-range dependencies in sequential data. The graph below represents the architecture of an LSTM cell.

The advantages of LSTMs makes them particularly powerful in stock price prediction, which is significantly time dependent. However, their complex structure and large data demand also make them prone to overfit, sensitive to noise in the stock price data and computationally intensive.

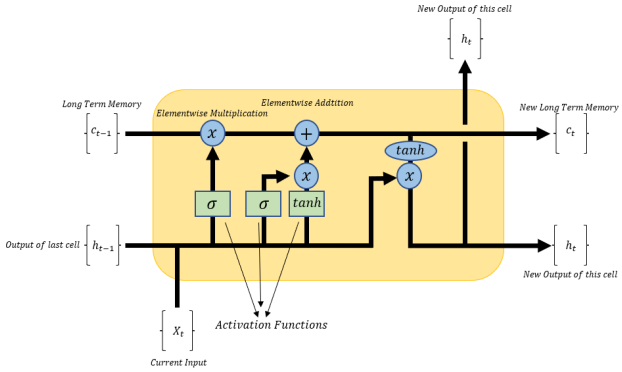

Figure 1: LSTM Model Cell

2.2 XGboost

XGBoost[1] is an advanced, efficient, and scalable end-to-end tree-based boosting system. It's known for handling large datasets and has become a popular choice in predictive tasks due to its speed and performance. The Regularized Learning Objective is shown below. it penalizes complexity by adding a cost associated with the number of leaves, and smooths the output weights. This dual approach helps to prevent overfitting, even when dealing with feature-rich structured data.

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

However, XGBoost has limitations, particularly in processing time-dependent data. Its standard form lacks the inherent architecture to naturally account for temporal dependencies, which is crucial in time-series predictions like stock price forecasting. As a result, XGBoost models may not yield optimal results in such applications, as temporal dynamics are a significant aspect of stock price movements.

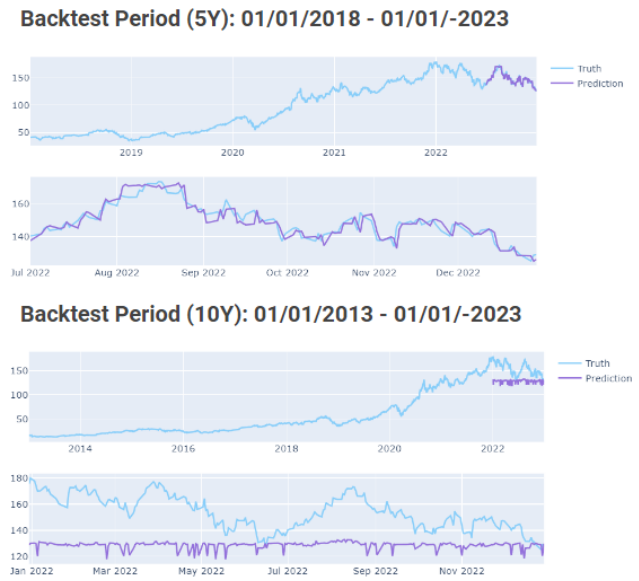

Figure 2: Backtest of XGboost Model

2.3 Previous Work

2.3.1 Single model. A study[6] using LSTM for stock price prediction, employing Long Short Term Memory (LSTM) and Bi-Directional Long Short Term Memory (BI-LSTM). The framework focuses on hyper-parameter tuning of these models to achieve high accuracy in forecasting future stock trends. Evaluations using a publicly available dataset of stock market prices, including open, high, low, and closing prices, demonstrated the effectiveness of the LSTM and BI-LSTM models in predicting stock prices with minimized root mean square error (RMSE). Nevertheless, using only LSTM lacks feature analysis which is important in practical finance problems. This work also did not consider scalability, considering the very large magnitude of financial data in reality.

2.3.2 Hybrid Model. Another existing work[3] gives a novel LSTM-BO-XGBoost model for stock market prediction, employing correlation analysis, Bayesian optimization, and a combination of LSTM and XGBoost techniques. The model is tested on ten different stock rates, showing superior performance in accuracy and stability compared to traditional LSTM and RNN models. Evaluation metrics like RMSE, MAE, accuracy rate, and F1-score confirm the effectiveness of the LSTM-BO-XGBoost model in predicting stock prices. However, this work mainly focused on the model construction, using raw open close high low, which contains lots of noise, for correlation analysis and lack of feature engineering.

3 OUR METHODOLOGY

In this paper, we introduce a new model that combines LSTM and XGBoost. This model is designed to effectively analyze the complex time-dependent characteristics of stock data, leveraging LSTM's sequential data processing capabilities and XGBoost's efficiency, scalability, and generalization ability.

3.1 Feature Engineering

Our methodology departs from conventional practices of utilizing raw stock data. Instead, it adopts a comprehensive feature engineering approach. This includes price change features like close-open-high-low, Moving Average and Ratios like EMA over four window sizes: 5 days, a month, two months, and a year. Additionally, we incorporate Volume features like normalized volume to train our LSTM model. Furthermore, the model includes technical indicators, RSI (Relative Strength Index) and MACD (Moving Average Convergence Divergence) to reveal short-term momentum and moving trends.

By using these engineered features, our LSTM model is designed to uncover underlying patterns and trends in feature changes that are not immediately apparent in raw data. Taking full consideration of time-dependent characteristics of features, LSTM model is introduced to give better feature purification for the usage of XGBoost algorithm. Figure 3 visualizes the comprehensive set of features that are harnessed to train our predictive model.

Feature Category	Feature Name
Price change features	open-close, high-low
Moving Averages and Ratios	ema5, ave5, ema30, ave30, ema60, ave60, ema252, ave252
Volume Features	norm_vol, vol_change
RSI	Relative Strength Index
MACD	Moving Average Convergence/Divergence

Figure 3: Summary of Financial Features

3.2 Model Architecture

The data pre-processing phase is constructed with 15 features, aiming to enhance the precision and relevance of our predictions. This is particularly vital given the inherent noise and fluctuating nature of stock market data.

Within our model's architecture, the LSTM network is constructed with dual layers to optimize learning efficiency and prevent overfitting. It consists of 128 hidden states with a dropout layer, followed by a secondary LSTM layer comprising 64 hidden states and an additional dropout layer, culminating in a dense layer for final prediction output.

The model training involved a comprehensive dataset of 2205 samples, each characterized by 15 features, processed over 60 time steps. These samples were intricately passed through 15 individual LSTM models, each tailored to predict a specific feature to better capture the temporal dependencies within the data. The shape of the input for these models is (2205, 60, 15), which indicates the number of samples, time steps, and features, respectively.

Following the LSTM training phase, the predictions for the training set are utilized as inputs to the XGBoost model, ensuring a fusion of temporal dynamics and ensemble learning. The XGBoost model's predictions are based on the features engineered by the LSTM models for the test set.

Notably, we chose to employ 15 independent LSTM models for each feature rather than a unified model for all features. This decision was informed by empirical evidence suggesting that individual models, under a similar structural framework, performed more effectively than a single comprehensive model. A unified model, when scaled in complexity to encompass all 15 features, posed a risk of overfitting, particularly given the relatively limited data available for an individual stock. This risk was mitigated by the use of separate models for each feature. However, it's important to

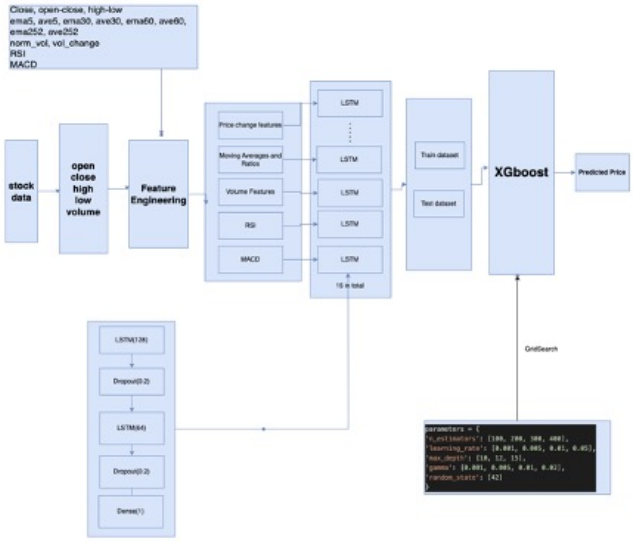

Figure 4: Construction of Our Model

acknowledge the trade-off in this approach: while individual feature models reduce overfitting, they demand extensive training time and consume significant computational resources.

In the final XGBoost training stage, we employed grid search for optimal hyperparameter selection, as shown in the figure 4, focusing on minimizing squared error. This approach ensures our model's robustness and accuracy.

n_estimators	100, 200, 300, 400
learning_rate	0.001, 0.005, 0.01, 0.05
max_depth	10, 12, 15
gamma	0.001, 0.005, 0.01, 0.02
random_state	42

Figure 5: Hyparameters of XGBoost Grid Search

4 EXPERIMENT

Our experiment involves Apple Inc.'s stock, focusing on the period from 01/01/2013 to 01/01/2023.

4.1 Feature Importance

Our analysis initiates with a meticulous assessment of the correlations between the closing price and the array of engineered features. From the heat map, it indicates that the 'high-low' spread and several long-term moving averages (such as EMA252 and ave60) display a

notable correlation with the closing price. These correlations suggest that while 'high-low' serves as a indicator for short-term market volatility and momentum, the long-term moving averages explain the underlying trends over extended periods.

The trend graphs further illustrate these correlations, with overlays of closing prices and select feature trends providing a visual insight of their synchronized movements.

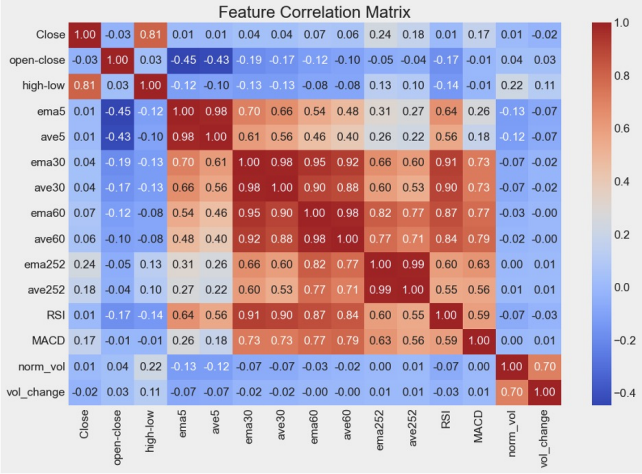

Figure 6: Heat Map of Features

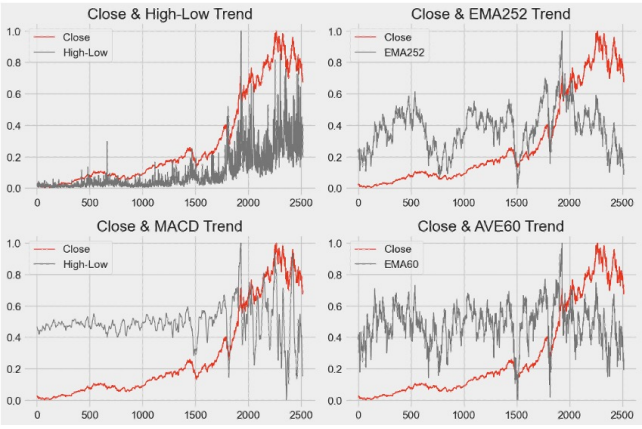

Figure 7: Trend of Features

However, we found an interesting emphasis of our model on MACD and RSI. MACD tracks both the momentum through the convergence and divergence of moving averages, as well as the potential reversal points in market trends. RSI indicates overbought or oversold

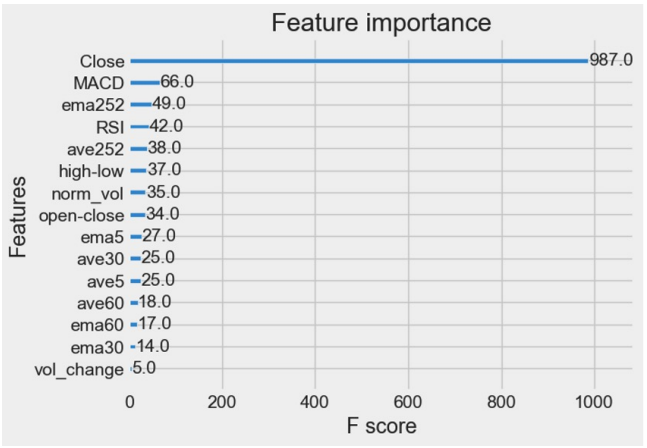

Figure 8: Feature Importance

conditions. The model's emphasis on these indicators highlights their utility in capturing market trends and shifts, potentially offering predictive insights into future stock movements.

4.2 Individual Stock Experiment

Performance Metrics	Hybrid Model	LSTM	XGBoost
RMSE	20.39156	69.04045	751.85262
MAE	3.60439	7.02835	24.47378
R2	0.787382	0.756179	-3.625562

Figure 9: Metrics for Individual Stock

We introduce similar structure settings for the comparable LSTM and Xgboost algorithms. Our empirical analysis indicates that the hybrid model outperforms in terms of visualization as well as performance metrics, as shown in Figure 9 and Figure 10. The model shows a reduced generalization error, and it demonstrates a greater explanatory power, affirming its capacity to capture and interpret data patterns more effectively.

4.3 Multiple Stock Experiments

Our comprehensive evaluation of model performance, conducted on the top ten stocks by market capitalization, reveals notable outcomes. The performance metrics—RMSE, MAE and R2—are averaged, as shown in

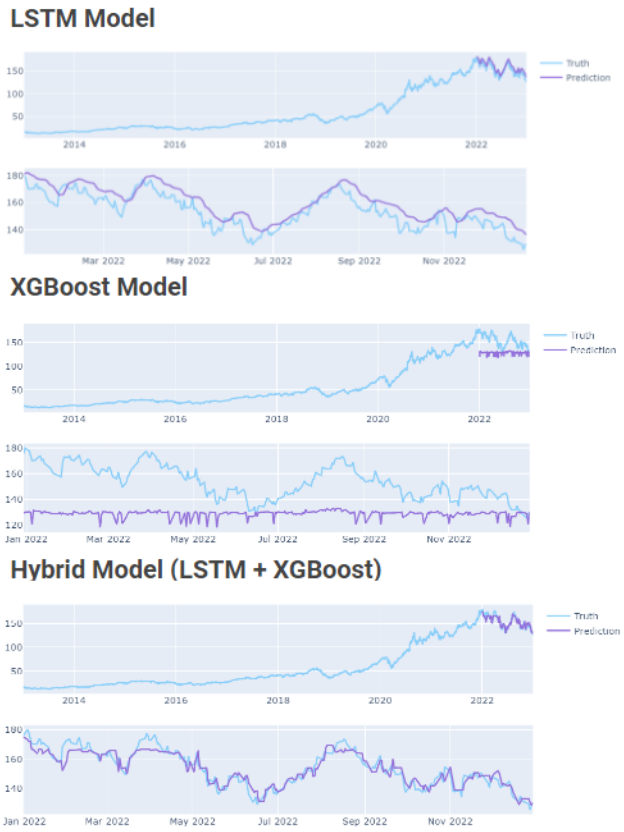

Figure 10: Model Comparison Backtest

Figure 11. From the box plots of Figure 12, the distribution of these metrics across the dataset illustrates the variability and consistency of each model.

Model	RMSE	MAE	R2
Hybrid	119.903233	7.178483	0.890667
LSTM	222.284008	10.852539	0.763305
XGBoost	2023.841348	30.234418	-1.375534

Figure 11: Averaged Metrics for Multi-Stock

The Hybrid model exhibits a dominant performance with the lowest RMSE and MAE, suggesting a better accuracy in prediction. Its R2 score is considerably higher than the other models, indicating a strong explanatory power. In contrast, the LSTM model, while outperforms the XGBoost, falls short of the Hybrid model's benchmark. The XGBoost model, with the highest RMSE and MAE and a negative R2 score, demonstrates considerable variance from the observed values, considering its challenges with this long timwe-span dataset.

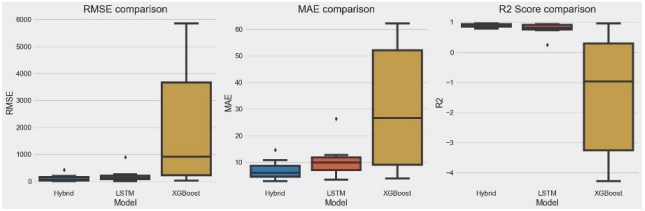

Figure 12: Box Plots of Metrics Distribution

Statistical validation of these differences was pursued through ANOVA and Tukey's HSD tests. While HSD tests indicate no significant performance differences between the Hybrid and LSTM models, they reveal a significant divergence when comparing the Hybrid to the XGBoost model, and similarly, the LSTM to the XGBoost.

```
ANOVA table for RMSE:
                sum_sq    df         F    PR(>F)
Model     1.834952e+07   2.0  5.132908  0.015313
Residual  3.753623e+07  21.0       NaN       NaN
Performing Tukey's HSD test for RMSE...

    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
group1  group2  meandiff p-adj    lower      upper   reject
-------------------------------------------------------------
Hybrid    LSTM  102.3808 0.9872 -1582.5594  1787.321  False
Hybrid XGBoost 1903.9381  0.025   218.9979 3588.8783   True
  LSTM XGBoost 1801.5573 0.0347   116.6172 3486.4975   True
-------------------------------------------------------------
```

Figure 13: ANOVA & HSD Test of RMSE

These findings highlight capability of the Hybrid model in stock market prediction. It is clear that the synergy of LSTM's temporal pattern recognition with XGBoost's structured data handling creates a robust predictive framework, outperforming the individual strengths of the singular models.

5 CONCLUSION AND FUTURE WORK

In conclusion, our project demonstrates the potential combination of LSTM and XGBoost for stock price prediction. By leveraging the strengths of both LSTM's time-dependency analysis and XGBoost's scalability

and generalization, we developed a Hybrid model that outperformed the single model. As shown in the Experiment, our model showed high accuracy in the prediction tasks. Future work included integrated a automated hyperparameter generation algorithm based on the training error in the model instead of using pre-selected parameters which need fine-tuning in order to have a stable performance on different dataset. Moreover, further actions are needed to reduce the computaional resources that training multiple LSTM model requires.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM. https://doi.org/10.1145/2939672.2939785

[2] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation* 12, 10 (2000), 2451–2471.

[3] Tian Liwei, Feng Li, Sun Yu, and Guo Yuankai. 2021. Forecast of lstm-xgboost in stock price based on bayesian optimization. *Intell. Autom. Soft Comput* 29, 3 (2021), 855–868.

[4] Ritika Singh and Shashi Srivastava. 2017. Stock prediction using deep learning. *Multimedia Tools and Applications* 76 (2017), 18569–18584.

[5] Ralf C Staudemeyer and Eric Rothstein Morris. 2019. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586* (2019).

[6] Md Arif Istiake Sunny, Mirza Mohd Shahriar Maswood, and Abdullah G Alharbi. 2020. Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. In *2020 2nd novel intelligent and leading emerging sciences conference (NILES)*. IEEE, 87–92.

A APPENDIX

A.1 Code

Link to the GitHub repo