XGBoost-LSTM for Feature Selection and Predictions for the S&P 500 Financial Sector

Ruilin Hu[1,a], **Tianyang Luo**[2, b, *]

[1]*Rotman Commerce, University of Toronto, Toronto, Canada*
[2]*School of Management and Economics, Chinese University of Hongkong Shenzhen, Shenzhen, China*
a. ruilin.hu@mail.utoronto.ca, b. 119020319@link.cuhk.edu.cn
**corresponding author*

Abstract: Financial industry researchers have long been committed to identifying factors that can predict trends in the financial sector of S&P 500, despite these factors often being difficult to discover. This article, through the combination of the Xgboost regressor and the shap summary plot, has mined and continuously optimized a potential excellent factor combination. Also, by utilizing the Xgboost regressor and LSTM models, it has achieved good prediction accuracy on the test set. This research gets the following results: First, the Xgboost regressor, in combination with Shap, has identified the seven most excellent factors from an initial combination of nine factors. Second, after imparting the final seven features to LSTM, the MSEs of the predictions made by Xgboost regressor and LSTM are 0.0003 and 0.0004, while the running times for Xgboost regressor and LSTM are 27 minutes and 16 minutes. Consequently, these results indicate that in the future predictions of finance sector index, investors may use the Xgboost-LSTM model for selecting effective factors and making accurate predictions efficiently.

Keywords: S&P 500, Xgboost regressor, LSTM, Xgboost-LSTM

1. Introduction

Studying the financial sector of the S&P 500 is paramount in understanding market trends, investment potential and risk management. As a significant indicator of economic health, this sector, comprising diverse financial institutions, offers insights into market-wide implications and economic environment shifts. Analysis of this sector is crucial for investors, policymakers, and researchers aiding strategic decision-making and financial planning.

Much research is about applying Xgboost regressor and LSTM models in the prediction area. For the Xgboost regressor, the research of its application in prediction tasks includes environments, business, and medical areas [1-3]. Studies on applying LSTM models for predictive purposes have also explored various domains, including medical, traffic, and industrial applications [4-6]. Additionally, there is research on the applications of the two models in the finance area [7,8]. However, the research on the application of the combined model of Xgboost and LSTM in predicting stock price or index is limited. The outcomes of these investigations consistently affirm the Xgboost regressor's powerful ability in feature selection and the prowess of the LSTM models as a formidable

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

tool for forecasting. To utilize the advantage of these two models, this research implements the combined model of Xgboost regressor and LSTM to forecast the S&P 500 finance sector index.

2. Data

This dataset comprises monthly data from December 2012 to June 2023. It comes from WIND, often referred to as China's Bloomberg, is a highly professional and widely used paid platform which offers various global financial and economic data. We have selected nine feature factors, which are supposed to affect the financial Sector Index of S&P 500, from the perspectives of macro factors, technical indicators, market returns, futures indices, and sub-sector returns (See Table 1 and Table 2).

Table 1: Y value.

SPF.SPI (Financial Sector Index of S&P 500)	It represents the aggregate performance of financial firms listed in the S&P 500 index.

Table 2: Nine features.

RSI Change	Describes the difference in Fast RSI (overbought/oversold indicator) between consecutive time periods. Divergence between price and RSI is considered a strong indicator of an imminent price reversal.
USA Core CPI (NSA: YoY)	Measures the YoY variation in the cost of consumer goods and services, excluding food and energy, thus indicating nationwide inflation trends.
S&P 500	The primary U.S. stock market index, typically used as the benchmark for market returns.
USA: NASDAQ Insurance Index	Showcases the performance of U.S.-listed insurance firms, impacting the overall financial sector.
USA: NASDAQ Bank Index	Reflects the performance of banking firms listed in the U.S., significantly influencing the financial sector.
Gold: SPGSGCTR Index	Captures gold market performance, which often inversely relates to the financial market.
USA: T-Bond YTM (10Y: Monthly Average)	Measured risk-free rate in classical models, tends to correlate with the financial sector.
EFFR Change	Variation in the daily Effective Federal Funds Rate within a month. Rate hikes usually depress markets, while rate cuts stimulate them.
Fear Indicator	Denotes significant market falls due to systemic risk-induced panic in a given month (such as 2020-02/2020-03/2023-03). If no such event occurs, represented as NaN.

3. Method

3.1. Pre-processing

In our data pre-processing, we implement time lagging for all features in X_train, except for EFFR Change and Fear Indicator. These two features are intraperiod variables as their effect manifest within the same period. Including EFFR Change and Fear Indicator features in the current period (t) rather than the previous one (t-1) correspondingly improves the model's fit.

For future EFFR Change projections, if no rate change is expected, we safely input 0 for the upcoming month. If a rate modification is anticipated, widely respected financial analysts usually offer fairly accurate collective forecasts. In uncertain scenarios, we can input an appropriate positive number signifying potential rate increments and compute a more conservative prediction to guide decision making.

Likewise, the Fear Indicator only comes into play when systemic risk events occur, such as a major bankruptcy. We can promptly set the fear indicator to 1 for that month and re-run the model to aid rapid response decisions.

Finally, we use MaxAbsScaler to normalize the X_train and X_test. The MaxAbsScaler is a normalization technique that scales features to a [-1, 1] range by dividing each value by its maximum absolute value.

3.2. Xgboost Regressor

XGBoost uses multiple decision trees trained on different data subsets and combines their predictions to make the final prediction [9].

Key hyperparameters include:

max_depth: controlling overfitting by limiting tree depth;

n_estimators: the number of trees for model robustness;

learning_rate: adjusting the influence of new trees;

subsample: using smaller subsets to add randomness and improve robustness;

colsample_bytree: controlling feature usage to prevent overfitting;

eval_metric: the metric for validation data evaluation;

objective: the loss function to be minimized during training for different problem types. Details are shown in Figure 1.

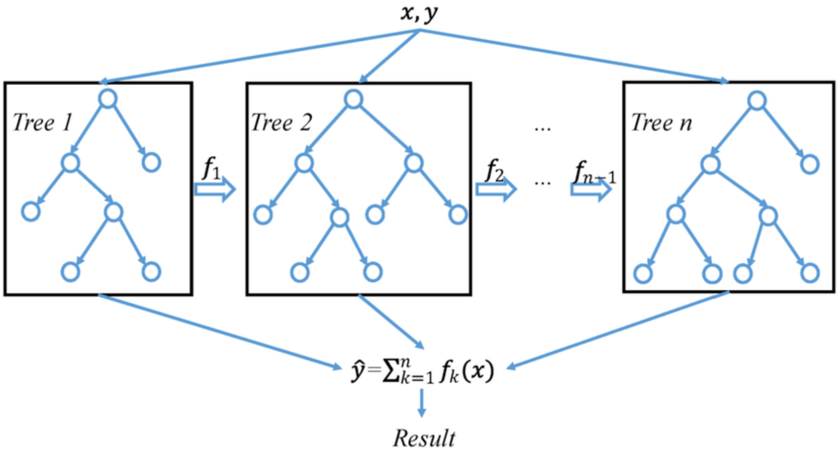

Figure1: Diagrammatic explanation of XGBoost.

We use nine seeds for stable predictions, define an objective using Optuna, and perform hyperparameter optimization through 100 trials to find the least average MSE of validation sets in XGBoost model. With the best hyperparameters, we train our data, make predictions and process them, considering the majority of the seeds. If the conditions are not met, the mean prediction is used. Then, we convert tests and predictions into binary and calculate accuracy. Furthermore, we analyze feature importance with ten groups of average shap values and optimize by eliminating features, iteratively re-running the model.

3.3. LSTM

This research uses the model with two LSTM layers for final prediction. The LSTM model is a type of recurrent neural network that can capture the dependence of the dependent variables on the time series of independent variables. This ability provides the model an advantage in making predictions of industries' performance. Because the trend of the industry heavily depends on the data of related factors over past periods. The full logic of the LSTM model is shown in Figure 2 below.

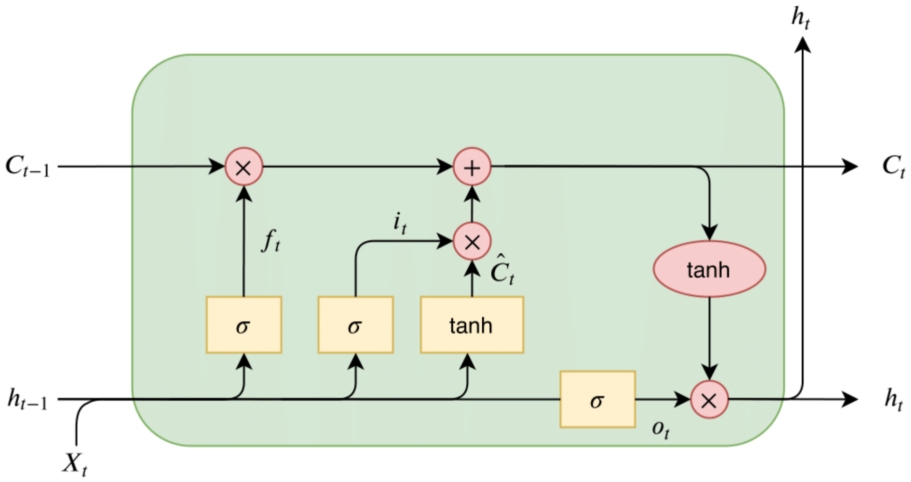

Figure 2: LSTM model.

There are several essential notations important to understanding this model:

ft: The forget gate of the model. It controls whether to discard the information passed to this gate.

Ct: The cell state. It shows the information that the cell is storing.

it: The input gate. It clarifies what information should be passed to the cell.

ot: The output gate. It decides on information that will be passed as output or as the next hidden state.

4. Result

4.1. Xgboost Regressor

In the original shap summary plot, the feature S&P 500 has a trouble. Its major high/low values are clustered around 0, and the points are distributed in an elliptical shape, indicating that this feature has little impact on the Y value (See Figure 3 and Figure 4).

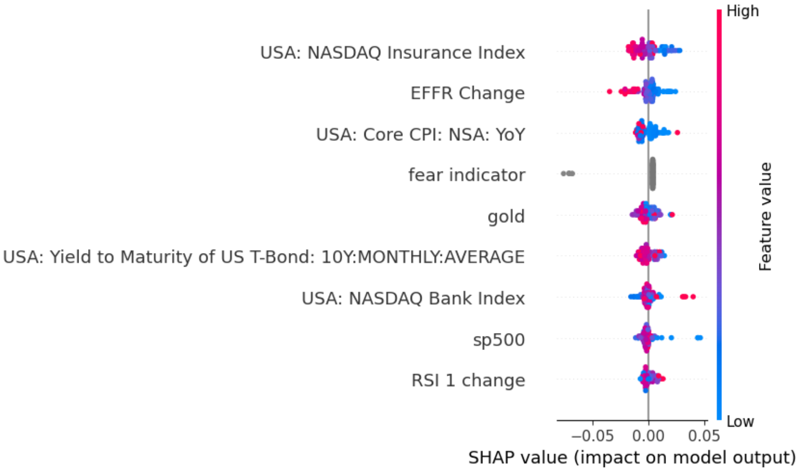

Figure 3: shap summary plot without dropping any feature.

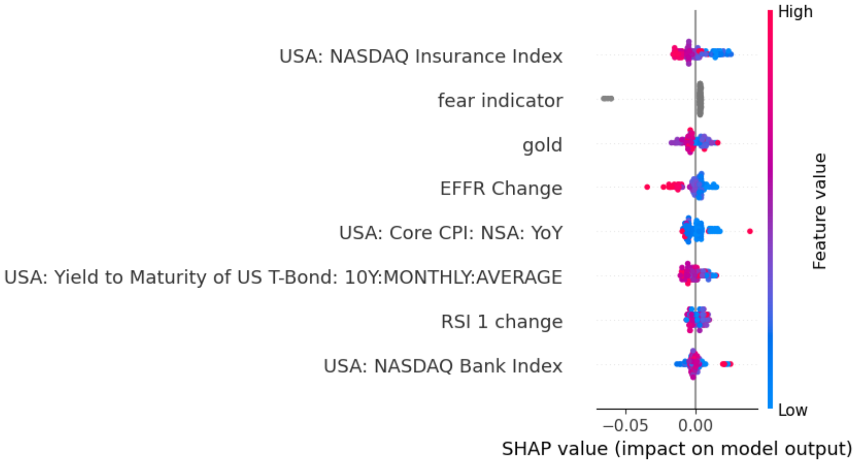

Figure 4: Shap summary plot after dropping S&P 500.

Still, in this plot, USA: NASDAQ Bank Index reflects a problem. Given the premise that the positive and negative values of a feature have opposite effects on the Y value, most high and low values in the Shap summary plot are clustered on the same side of the x-axis. Supposing that this feature does not perform well in the model. After deleting USA: NASDAQ Bank Index, we get the nine seeds' predictions in Table 3:

Table 3: Xgboost regressor's predictions.

	Seed10	Seed20	Seed30	Seed40	Seed50	Seed60	Seed70	Seed80	Seed90
2023-01	0.0302	0.0293	0.0273	0.0315	0.0314	0.0310	0.0317	0.0408	0.0349
2023-02	-0.0331	-0.0188	-0.0308	-0.0306	-0.0300	-0.0306	-0.0246	-0.0259	-0.0198
2023-03	-0.0909	-0.0328	-0.0787	-0.0718	-0.0569	-0.0527	-0.1032	-0.0562	-0.0584
2023-04	0.0234	0.0291	0.0222	0.0248	0.0273	0.0229	0.0268	0.0298	0.0277
2023-05	-0.0278	-0.0180	-0.0261	-0.0248	-0.0246	-0.0300	-0.0212	-0.0207	-0.0170
2023-06	0.0628	0.0416	0.0587	0.0591	0.0521	0.0465	0.0499	0.0581	0.0564

Excluding the outliers in each month, we calculate the average of the rest values for each month and form the final prediction in Table 4:

Table 4: Comparison between the final prediction and actual value.

	Final y pred	Actual y test
2023-01	0.0380	0.0670
2023-02	-0.0322	-0.0244
2023-03	-0.0788	-0.0974
2023-04	0.0369	0.030
2023-05	-0.0267	-0.0448
2023-06	0.0647	0.0652

We get the Mse of 0.0003. After converting the positive values to 1s and negative values to 0s, we obtain the accuracy of 1.

Additionally, we confirm the feasibility of optimizing the Xgboost regressor model through the SHAP summary plot to some extent by comparing the mean squared error (MSE) and accuracy of the three models (See Table 5).

Table 5: The MSE and accuracy of each of the three models.

	Mse	Accuracy
No feature deletion	0.0010	1.0
Delete S&P 500	0.0009	1.0
Delete S&P 500 & USA: NASDAQ Bank Index	0.0003	1.0

We also found that deleting features based on their feature importance might not be applicable in the xgboost regressor. For instance, the feature importance for S&P 500 and USA: NASDAQ Bank Index are not low, but the prediction error of the model increased after removing them. When using 'reg:squarederror' as an objective in xgboost, we probably need to consider more on whether the actual impact aligns with logic, rather than just the contribution.

By far, the best feature group of S&P 500 fianance sector has seven features shown below:

1. USA: Core CPI: NSA: YoY
2. Nasdaq Insurance Index
3. Gold: SPGSGCTR Index
4. RSI Change
5. EFFR Change
6. USA: Yield to Maturity of US T-Bond: 10Y: MONTHLY:AVERAGE
7. Fear Indicator

Details are shown in Figure 5.

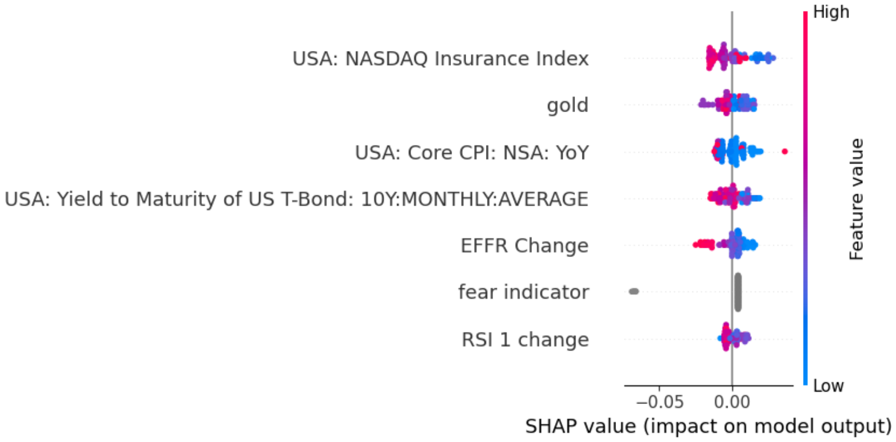

Figure 5: Shap summary plot after the two features.

Based on the SHAP summary plot obtained from the average SHAP value of nine seeds, we can intuitively draw the following interpretations:

Higher values in the insurance industry for the current month are more likely to have a larger negative impact on the next month's S&P 500 financial sector compared to other values in the dataset. This can be explained by the tendency of people to buy insurance as a symbol of economic downturn, and the reverse holds true.

Higher year-on-year data for core CPI in the current month are more likely to have a larger negative impact on the next month's S&P 500 financial sector compared to other values in the dataset. This can be explained by higher CPI indicating greater inflationary pressure, which is detrimental to the economy, and the reverse holds true.

The greater the increase in EFFR Change for the current month, the more likely it is to have a larger negative impact on the next month's S&P 500 financial sector compared to other values in the dataset. This can be explained by the general notion that more rate hikes have a larger negative impact on the financial industry, and the reverse holds true.

The influences and explanations for the current month's gold index, 10-year Treasury bond yield, average of a month, and RSI 1 change on the next month's S&P 500 financial sector are similar to those mentioned above and will not be individually introduced here.

4.2. LSTM

The LSTM model used monthly data from January 2013 to December 2022 as the training set and monthly data from January 2023 to June 2023 as the test set. The LSTM model implements the tanh function as the activation function. Compared to the other common activation function ReLU, the tanh can produce various negative outputs for different negative inputs. But ReLU will only produce zeros for negative inputs. Therefore, the model that implements the tanh function can differentiate the negative inputs better than the model that implements ReLU. Since there are many negative inputs in this research, the LSTM model implements the tanh function to gain better differentiation of these inputs and therefore generate more accurate predictions. The model also used the Adam function as the optimizer because it requires a small amount of time to achieve optimization [10]. This advantage will improve the overall efficiency of the model and reduce the time required to achieve the optimum result.

After setting up the parameters described above, the LSTM model is designed to train with the number of epochs ranging from twenty to two hundred. For each number of epochs, five identical LSTM models are trained. All five models will give their predictions of monthly percentage change in the S&P 500 finance sector index from January 2023 to June 2023. For each of the five models, the Mean squared error (MSE) of predictions will be calculated. The average of these MSEs will be used as the measurement of the accuracy of these predictions. The most accurate predictions will be used as final predictions.

The test set average MSE of the model is shown in Figure 6. From the Figure 6, the best predictions of the model are generated at 150 epochs. The average MSE of the best predictions is 0.0004.

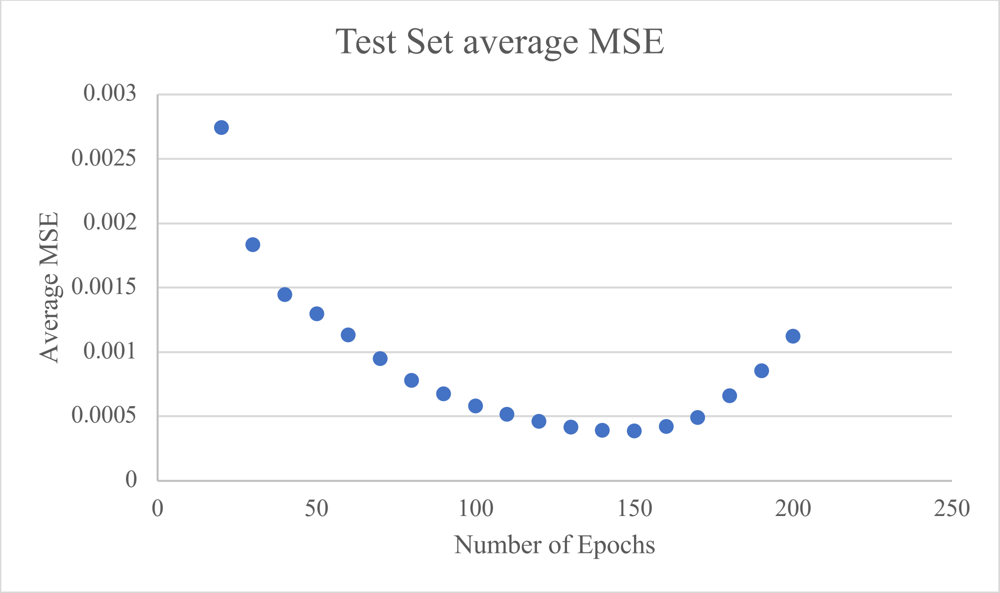

Figure 6: Test Set Average MSE.

The best predictions of the monthly percentage change in the S&P 500 finance sector index are shown in Table 6 and Figure 7. In the figure, the orange line represents the actual value from January

2023 to June 2023, while other lines represent predictions of the model. From the average MSE of the results and the figure, the predictions are very close to the actual value. However, there are still small gaps between the predictions and the actual data.

Table 6: Predictions and actual value.

Dates	Jan-2023	Feb-2023	Mar-2023	Apr-2023	May-2023	Jun-2023
Seed 0	0.0513	-0.0452	-0.0790	0.0577	-0.0247	0.0570
Seed 1	0.0570	-0.0449	-0.0840	0.0620	-0.0225	0.0652
Seed 2	0.0569	-0.0433	-0.0784	0.0604	-0.0233	0.0657
Seed 3	0.0555	-0.0432	-0.0763	0.0586	-0.0205	0.0605
Seed 4	0.0600	-0.0430	-0.0800	0.0654	-0.0207	0.0647
Test	0.0670	-0.0245	-0.0974	0.0302	-0.0448	0.0653

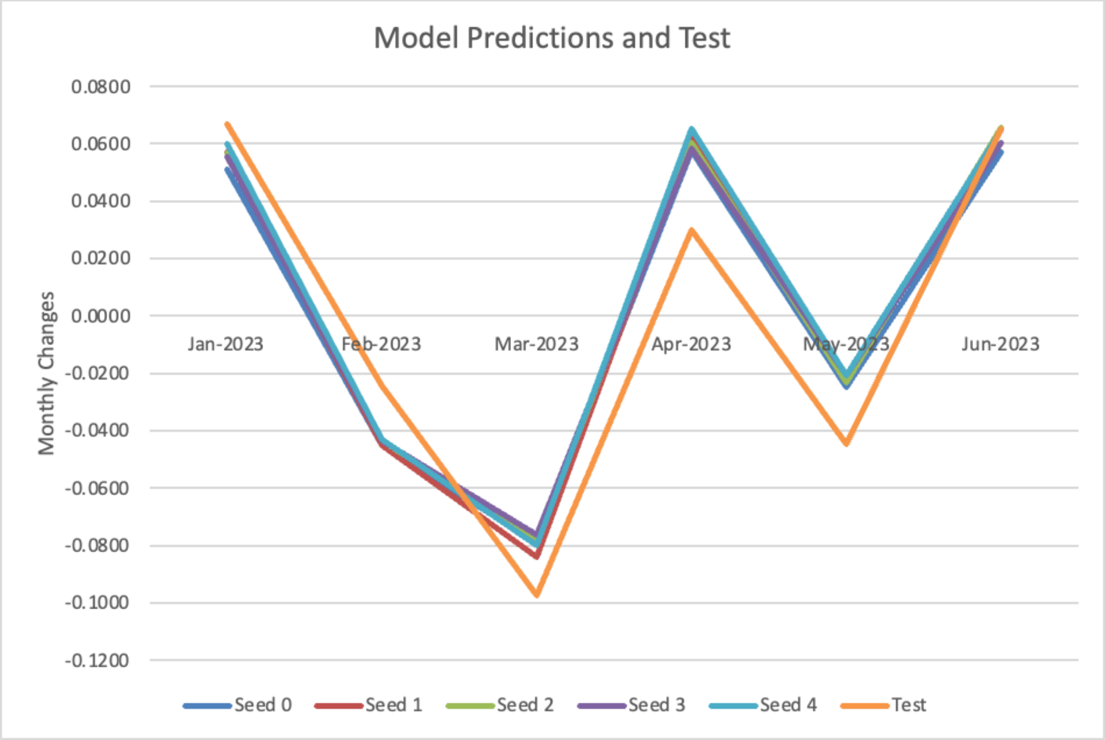

Figure 7: Predictions and the actual data.

The result of the LSTM model is close to the result of the Xgboost model. As shown in Table 7, the MSE of the result of the Xgboost model is 0.0003, which is not far away from the MSE of the best predictions of the LSTM model. With the same level of accuracy, the time for the LSTM model to run through these epochs is around 16 minutes, while the running time for the Xgboost model with nine seeds is 27 minutes. Therefore, the LSTM model is more efficient or less time-consuming than the Xgboost model in making predictions at the same level of accuracy.

Table 7: LSTM and Xgboost's accuracy of results and running times.

Model	MSE	Running Time
LSTM	0.0004	16 minutes
Xgboost	0.0003	27 minutes

5. Conclusion

This research selects features with predictive power by Xgboost regressor and predicts S&P 500 finance sector index by both Xgboost regressor and LSTM.

Firstly, seven of nine important features are successfully selected by Xgboost regressor and shap summary plots. Secondly, after imparting the seven features to LSTM, the running time and MSEs of Xgboost regressor and LSTM are also obtained during the research. It shows that, after obtaining selected features from Xgboost regressor and with almost equivalent precision, LSTM demonstrates significantly higher efficiency in making predictions. Finally, these findings reach the conclusion that for future predictions on finance sector index, investors could consider utilizing the integrated Xgboost-LSTM model to effectively identify crucial factors and achieve precise predictions with efficiency.There are considerations for future explorations. Other methods for choosing factors, such as the Granger Causality test, could be added for comparison to the Xgboost regressor. In addition, time-series model such as Transformer could be utilized to check whether it would also produce a good estimation.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Dai, H., Huang, G., Zeng, H., and Yang, F. (2021). PM2.5 Concentration Prediction Based on Spatiotemporal Feature Selection Using XGBoost-MSCNN-GA-LSTM. Sustainability, 13(21).

[2] Muslim, M. A., and Dasril, Y. (2021). Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. International Journal of Electrical and Computer Engineering (IJECE), 11(6), 5549-5557.

[3] Chen, C., Shi, H., Jiang, Z., Salhi, A., Chen, R., Cui, X., and Yu, B. (2021). DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network. Computers in Biology and Medicine, 136.

[4] Mou, H., and Yu, J. (2021). CNN-LSTM Prediction Method for Blood Pressure Based on Pulse Wave. Electronics, 10(14).

[5] Altché, F., and Fortelle, A. d. L. (2017). An LSTM network for highway trajectory prediction. 2017 IEEE 20th International Conference on Intelligent Transportation Systems, 353-359.

[6] Ren, L., Dong, J., Wang, X., Meng, Z., Zhao, L., and Deen, M. J. (2021). A Data-Driven Auto-CNN-LSTM Prediction Model for Lithium-Ion Battery Remaining Useful Life. IEEE Transactions on Industrial Informatics, 17(5), 3478-3487.

[7] Chen, K., Zhou, Y., and Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. 2015 IEEE International Conference on Big Data (Big Data), 2823-2824.

[8] Li, Y., Stasinakis, C., and Yeo, W. M. (2022). A Hybrid XGBoost-MLP Model for Credit Risk Assessment on Digital Supply Chain Finance. Forecasting, 4(1), 184-207.

[9] Wang, Y., Pan, Z., Zheng, J. et al (2019). A hybrid ensemble method for pulsar candidate classification. Astrophys Space Sci 364, 139.

[10] Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.