

LGBM on Stock Returns Prediction and Portfolio Construction

Hongyu Zhu^a

Business and Leadership School, Monash University, Wellington Rd, Clayton VIC 3800, Australia

Keywords: LGBM, Stock Returns Prediction, Portfolio Construction.

Abstract: Investors need accurate stock price forecasts because they raise the chances of successful investments. This research assesses how machine learning methods predict the next day's daily returns of S&P 500 companies based on the past 10 days' data. The study also focuses on building portfolios based on forecasted returns. The study's dataset contains stock prices for S&P500 index companies and S&P500 index prices from 2014 to 2017, while including 497,472 total data points. The extended dataset includes 18 features which cover stock market relationship indicators as well as technical analysis indicators. This study assessed predictive models which consist of Random Forest, eXtreme Gradient Boosting (XGBoost) as well as Long Short-Term Memory (LSTM) and Light Gradient Boosting Machine (LGBM). The analysis results show that the LGBM model shows superior accuracy in forecasting the next day's daily returns since it achieves an R-square value of 0.7377. Concurrently, this study utilized the predicted daily returns to construct a portfolio comprising 20 companies from the S&P500 index companies. Based on the objective of maximizing the portfolio alpha, the optimal portfolio result contains 20 stocks of S&P500 companies, achieving an alpha over 10% and a total return of by 20%.

1 INTRODUCTION

Accurate stock price predictions enable investors to determine market conditions which lead to better stock market decisions and ultimately results in increased returns. Both individual and institutional investors who master precise stock movement predictions gain significant competitive advantages in the present uncertain financial market (Soni, Tewari, & Krishnan, 2022). Research shows that predicting stock prices requires using multiple analytical tools, including fundamental and technical indicators, which help create investment strategies. According to investigation sophisticated investors now focus on using predicted daily returns to boost alpha while managing risk exposure (Mohammed et al., 2023). Also, according to Di's research, the employment of technical analysis indicators in the context of price forecasting has been demonstrated to achieve a maximum accuracy of over 70% (Di, 2014). This efficacy can be attributed to the principles underlying the efficient market hypothesis and the random walk theory.


This study evaluates the effectiveness of machine learning in forecasting next-day daily returns for S&P 500 firms and formulates an optimal portfolio using these return predictions. Utilising data from 2014 to 2017 with over 497,000 points, the study integrates technical and market features to enhance accuracy. The research aims to optimize portfolio management, improve returns, and refine risk control, offering valuable insights for investors.

2 MODELS

The present study appraised five distinct models: LSTM, LGBM, XGBoost, Random Forest and Decision Tree. These models function as discrete strategies for time series data processing and stock market movement prediction, and machine learning methods have been shown to be capable of identifying complex patterns and connections within voluminous datasets (Mohammed et al., 2023).

2.1 Long Short-Term Memory (LSTM)

LSTM demonstrates a particular aptitude for sequence prediction tasks and the capture of long-

^a <https://orcid.org/0009-0002-3023-357X>

term dependencies, rendering it well-suited for applications such as time series analysis, machine translation, and speech recognition. LSTM networks excel at sequence prediction and long-term dependency modeling which makes them ideal for time series analysis and machine translation applications (Mehtab, Sen, & Dutta, 2020; Wen et al., 2022).

2.2 LightGBM (LGBM)

LGBM is a fast, distributed framework based on gradient boosting that uses leaf-wise growth to optimize data partitioning and significantly reduce the consumption of memory and computational resources. Unlike the level-wise growth of XGBoost, LGBM uses leaf-wise growth. At each iteration, LGBM selects the node that yields the greatest gain among all the current leaves for splitting, as opposed to splitting by layer. This strategy is more effective in reducing errors and enhancing model performance (Ke et al., 2017).

2.3 Decision Tree

Decision tree is a data structure that conditionally partitions data. Each node in the tree represents a test of a specific feature, and the outcome of each test is represented by a branch in the tree. The merits of decision trees include their simplicity, intuition, ease of interpretation, and rapid computation. However, it should be noted that decision trees are sensitive to noise in the data.

2.4 Random Forest

Random Forest is an integrated learning method that improves model stability and accuracy by constructing multiple decision trees and averaging their outputs. In comparison with a solitary decision tree, Random Forest exhibits a reduced risk of overfitting, whilst simultaneously demonstrating superior accuracy and learning capability in scenarios characterized by noise and anomalies in the data.

2.5 XGBoost

The XGBoost algorithm develops from gradient boosting by building sequential regression trees which work to reduce errors that arise in previous rounds. The addition of regularisation (L1 and L2) helps enhance model generalisation by controlling complexity and avoiding overfitting. XGBoost achieves better efficiency and scalability using parallelisation techniques combined with the automatic handling of missing values and

approximate tree learning methods. XGBoost is effective for extensive datasets and complex dimensional challenges while delivering robust predictive results (Song et al., 2022).

3 DATA PREPROCESSING

3.1 Features Prepared

The datasets utilized in this study encompass S&P 500 company stock prices and S&P 500 index prices ("Stock Prices," 2021) ("S&P 500 Historical Data," 2020). These datasets encompass the opening price, closing price, high price, low price, and trading volume of all S&P 500 companies and the S&P 500 index, respectively, for each day between 2014 and 2017.

Traditional stock price prediction methods often rely solely on historical price and volume data, which limits their accuracy due to the reliance on a single data source. This experiment aims to improve prediction accuracy by incorporating the S&P500 index, which reflects overall market sentiment, volatility, and the correlation of a stock with broader market returns. By adding the S&P500 index as an additional feature, the model leverages more comprehensive information. At the same time, technical analysts believe that by analyzing historical data to create indicators that react to market trends and reversals, they can help predict future price movements. This study also explores the use of specific technical analysis indicators to provide more non-linear relationships for model training, enhancing the model's explanatory power and forecasting accuracy (Chong & Ng, 2008) (Agrawal et al., 2021). Table 1 all the technical analysis indicators in the dataset.

Table 1: Features prepared

Indicators	Formula	Application
Moving average (MA5&20)	$MA_n = \frac{1}{n} \sum_{i=t-n+1}^t P_i ; n = 5$ $MA_n = \frac{1}{n} \sum_{i=t-n+1}^t P_i ; n = 20$ <i>P_i: the close price of n day</i> <i>n: window size for the moving average</i> <i>t: current time point</i>	Moving average determines the overall market trend and measures the decrease in market noise.
Relative Strength Index (RSI)	$RSI = 100 - \frac{100}{1 + RS}$ $RS = \frac{\text{Average gain}}{\text{Average loss}}$ <i>Average gain: average price increase over 20 da</i> <i>Average loss: average price decrease over 20 da</i>	RSI is used to anticipate potential price reversals in the market.
On-Balance Volume (OBV)	$OBV_t = OBV_{t-1} + \text{Sign}(P_t - P_{t-e}) * V_t$ <i>Sign(P_t - P_{t-e}): +1 if P_t > P_{t-e} - 1 if P_t < P_{t-e}; and 0 otherwise</i> <i>P_t: Current price</i> <i>P_{t-e}: Price of e days ago</i> <i>V_t: Current trading volumn</i>	OBV integrates trading volume and price shifts to help understand how capital enters and exits the market cap.
Bollinger band upper & lower	$\text{Middleline(MA)} = \frac{1}{N} \sum_{i=1}^N P_i$ $\text{Upper Band} = MA + 2 * \sigma$ $\text{Lower Band} = MA - 2 * \sigma$ <i>P_i: the close price of n day</i> <i>σ: standard deviation of the price</i>	The stock reaches overbought or oversold conditions when its price moves beyond the upper or lower band respectively.
Average Directional Index (ADX)	$+DI = \frac{(\text{CurrentHigh} - \text{PreviousHigh})}{\text{True Range}}$ $-DI = \frac{(\text{PreviousLow} - \text{CurentLow})}{\text{True Range}}$ $DX = \frac{ +DI - -DI }{+DI + -DI}$ $ADX = \text{Smooth}(DX)$ <i>True Range: the maximum price range of a period</i>	The ADX tool measures the strength of current market trends. When the ADX reads above 25 it typically shows that the market is in a strong trend regardless of direction.

3.2 Time-Series Split

When dealing with financial data, time series split is essential as stock data is usually organized in chronological order. The underlying logic of forecasting is that it is impossible to predict past data using future data (LeBaron & Weigend, 1996).

The experiments used a time series split method where the data was divided into 10 parts for 10 cross-validations in chronological order. The main advantage of this method is that it ensures that the chronological order is maintained, thus avoiding the influence of future data on model training. It unveils a two- to threefold augmentation in the discrepancy

of model performance when subjected to varying segmentations in comparison to the variations in neural network parameters (LeBaron & Weigend, 1996). This underscores the necessity for a dynamic validation framework when appraising models. With LSTM, if the training set contains future data, the effectiveness of LSTM can be inflated by data leakage (Wang & Guo, 2020).

In addition, time series splitting facilitates incremental training of the model, which is each split progressively expanding the training set. This also makes it possible to simulate a real-world situation where more and more data is obtained over time.

4 DAILY RETURN PREDICTION AND PORTFOLIO CONSTRUCTION

Following an investigation into the efficacy of mechanical learning in predicting the subsequent day's closing price of S&P 500 company stocks, the subsequent objective of this study is to predict the subsequent day's daily return of S&P 500 company stocks and to construct a portfolio that maximizes the alpha based on the predicted daily return in S&P 500 company stocks. In addition, this segment of the dataset has undergone a uniform preprocessing procedure, identical to the one applied to the dataset intended for the prediction of closing prices.

4.1 Prediction Results

The results of all the models investigated in this section are compared in Table 2. Upon evaluation of

the mean absolute error (MAE) across all models the LGBM model showed the lowest figure at 0.0013 which indicates its predictions were closest to the true values. The Random Forest model comes next with an MAE of 0.0014 while the LGBM model showed better performance through its R^2 score and explained variance which stood at 0.8384 and 0.8385 respectively. The model demonstrates an explanatory power for the target variable variance of about 83.8%.

The LGBM model achieves the best results across all evaluation metrics by having minimal error and maximal explanatory power for the target variable. Among the models tested Random Forest secures second place in performance. The Decision Tree model ranks below Random Forest and LGBM yet remains a viable option. The performance of XGBoost falls short of other models across all evaluation metrics with marked weaknesses in error rates and explanatory power which makes it an unsuitable model for this dataset.

Table 2: Stock daily return prediction evaluation

	MSE	MAE	R2	Explain variance
LGBM	0.0000	0.0017	0.7377	0.7377
XGBoost	0.0000	0.0024	0.4959	0.4959
Decision tree	0.0000	0.0024	0.5352	0.5352
Random forest	0.0000	0.0022	0.6108	0.6108
LSTM	0.0000	0.0025	0.4905	0.4906

4.2 Portfolio Construction

In light of the predicted returns for S&P 500 companies the following day, it can be concluded that the best-performing model is the LGBM. Therefore, the subsequent section of this paper will utilize the trained LGBM model to predict returns for all subsequent days and construct a portfolio based on the predicted returns in a manner that maximizes alpha.

4.2.1 Portfolio Analysis

Firstly, the LGBM model, which was trained in the previous section, is employed to predict the daily returns for each day in the dataset. The predicted daily returns are then converted into annual returns, and the covariance matrix for each company within S&P500 is calculated.

The formula below is utilized in this study to calculate the cumulative return, defined as the annualized return for months. This is achieved by multiplying the cumulative product of the returns of all the trading days of each company each month.

$$\text{Annualreturn}(R_{i,m}) = \prod_{d=1}^t (1 + r_{i,d,m}) - 1$$

$R_{i,m}$: the annual return of asset for month

$r_{i,d,m}$: the return of asse on day within month
 t : the total number of days in month (1)

The construction of the portfolio is informed by a maximum alpha construction strategy, incorporating 20 stocks to achieve optimal diversification. The weighting of each stock in the portfolio is determined by a risk parity model, a strategic approach aimed at minimizing the risk contribution of assets in the portfolio.

$$\text{Marginalriskcontribution}(MRC_i) = \frac{\partial \sigma_p}{\partial w_i} = \frac{\partial}{\partial w_i} \sqrt{w^T \Sigma w}$$

w_i : the weight of asset in the portfolio

W : the vector of all asset weights in the portfolio

$$\text{Risk contribution } (RC_i) = w_i * MRC_i \quad (2)$$

Risk parity objective

$$= \min \sum_{i=1}^n (RC_i - \text{mean}(RC_i))^2$$

$$\text{Alpha} = R_{\text{portfolio}} - (R_{\text{risk-free}} + \beta * (R_{\text{market}} - R_{\text{risk-free}})) \quad (3)$$

After the initial period, in order to assess the overall risk of the portfolio, the utilization of Value at Risk (VARs) and Conditional Value at Risk (CVARs) was introduced to evaluate the risk to which the portfolio is exposed. Assuming the confidence interval which is α is 95%.

$$\text{Value at Risk (VARs)} = \text{percentile}(r, 5)$$

$$\text{Conditional Value at Risk (CVARs)} = \frac{1}{\alpha} \int_{-\infty}^{\text{VaR}_\alpha} r(x) dx \quad (4)$$

It is on this basis that the present study introduces the Kelly criterion, which is utilized for the purpose of calculating the optimal investment amount in a portfolio, as determined by the market risk-free rate. The residual amount is then invested at the risk-free rate. Within the parameters of this study, the Kelly criterion is calculated by assuming the probability of success is 70%, according to the R2 of the LGBM model, which is 0.7377.

$$\text{KellyCriterion}(f^*) = \frac{p * b - (1 - p)}{b}$$

p : the probability of winning
 b : total return of the optimal portfolio. (5)

4.2.2 Portfolio Results

Table 3: Portfolio construction

Symbol	Weight	Symbol	Weight
DOV	0.05	NWSA	0.05
EXPD	0.05	MA	0.05
DISCK	0.05	UA	0.05
HLT	0.05	WMT	0.05
KORS	0.05	TPR	0.05
AOS	0.05	DVN	0.05
UAL	0.05	BEN	0.05
HBI	0.05	MAT	0.05
GE	0.05	HP	0.05
LH	0.05	SNI	0.05

As shown in Table 3, this is a portfolio that is a total of 20 stocks that make up this portfolio while each one holds a 5% share of the overall investment weight. The portfolio contains businesses across multiple segments including consumer products, financial services, technology firms and industrial companies. The portfolio holds significant stocks such as Disney (DISCK), Walmart (WMT), General Electric (GE), and additional companies like Maersk (EXPD), Hanesbrands (HBI), United Airlines (UAL). Through diversification the portfolio reduces risk exposure and offers protection from volatility of individual market sectors and stocks. The portfolio composition provides opportunities for sustained growth while distributing investment across multiple market sectors. Equal weight distribution across all assets prevents excessive investment in one stock or industry, thus maintaining portfolio stability while reducing risk from sector-specific declines.

The provided Table 4 delivers insight into the portfolio's risk and return characteristics. The portfolio's risk level is measured through both VaR and CVaR at 95% confidence interval to determine potential losses under extreme market conditions. The

portfolio has a VaR of 0.1780 meaning it could lose up to 17.80% in value with 95% confidence. When losses exceed the threshold established by VaR, the CVaR figure of 0.1773 indicates the average loss reaches 17.73% in such cases with 95% confidence. The figures given demonstrate how much the portfolio is at risk of losing value.

Table 4: Portfolio evaluation

Metric	Value
VaR (95%)	0.1780
CVaR (95%)	0.1773
Alpha	0.1061
Total Return	0.1914
Optimal investment fraction (Kelly Criterion)	-0.6327

The Alpha metric shows the portfolio's performance relative to the market benchmark. The portfolio delivered a positive Alpha of 0.1061, which reveals its superior performance by 10.61% above the market benchmark's results. The portfolio achieved overall profitability, which is demonstrated by the Total Return of 19.14%, that represents the combined gain during the period covered.

According to the Kelly formula investors should reduce their portfolio exposure by lowering the amount of capital they allocate to it. The most suitable investment for the portfolio is -63.27%. The Kelly formula suggests that, in this case, the portfolio reveals excessive risk and may lead to losses and is therefore not recommended for investment. This is because the portfolio's winning percentage and expected return are not satisfactory. By calculation, assuming a constant portfolio win rate of 0.7, the expected return on the portfolio would be a minimum of 42.86%, if the objective is to achieve an investment worthy of the Kelly model. This indicates that, with the current portfolio win rate, the expected return would need to increase by a minimum of another 23.72% to meet the requirement of being worthy of investment.

Although the portfolio exhibits strong overall alpha performance, it presents two critical shortcomings when assessed in terms of win rate and risk reporting: The model demonstrates limited predictive accuracy through its low win rate while also producing suboptimal returns. The portfolio's practical application and dependable performance in real-world scenarios are severely impacted by these limitations.

5 CONCLUSION

The study shows how machine learning techniques can successfully predict stock returns and develop effective investment portfolios. The LSTM, LGBM, and XGBoost models successfully forecast daily returns for S&P 500 companies where LGBM shows superior performance in next day return predictions and Random Forest leads in closing price predictions.

The portfolio delivers a strong return of 19.14% while generating an alpha of 10.61% but presents risk factors that create investor concerns. The Kelly Criterion demonstrates that although the portfolio delivers positive performance results its limited win rate cannot merit the associated risk level which renders it inappropriate for risk-averse investors. Potential drawdowns and volatility undermine the long-term sustainability of the current allocation strategy.

Upcoming studies need to concentrate on improving the risk-return trade-off while developing advanced methods to control risk and increase the precision of market predictions. Investment portfolio construction improves when investment strategies adapt to investor preferences through personalized risk profiles. The inclusion of various asset types such

as gold, cryptocurrencies, and real estate into machine learning models enhances portfolio resilience and diversification. Ongoing advancements in machine learning enable the creation of portfolio optimization strategies that are adaptive, dynamic and robust.

REFERENCES

- Agrawal, M. 2021. Stock prediction based on technical indicators using Deep Learning Model. *Computers, Materials & Continua*.
- Chong, T. T. L., & Ng, W. K. 2008. Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30. *Applied Economics Letters*, 15(14), 1111–1114.
- Di, X. 2014. Stock trend prediction with technical indicators using SVM.
- Hanfei Wen, Jun Yu, Guangjin Pan, Xiaojing Chen, Shunqing Zhang, & Shugong Xu. 2022. A Hybrid CNN-LSTM Architecture for High Accurate Edge-Assisted Bandwidth Prediction. *IEEE Wireless Communications Letters*.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- LeBaron, B., & Weigend, A. 1996. A Bootstrap Evaluation of the Effect of Data Splitting on Financial Time Series. *Capital Markets eJournal*.
- Mehtab, S., Sen, J., & Dutta, A. 2020. Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. *ArXiv*.
- Mohammed, S. H. Krishna, P. Mudalkar, Narinder Verma, P. Karthikeyan, & Ajay Singh Yadav. 2023. Stock Market Price Prediction Using Machine Learning. *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*.
- Song, Y., Li, H., Xu, P., & Liu, D. 2022. A Method of Intrusion Detection Based on WOA-XGBoost Algorithm. *Discrete Dynamics in Nature and Society*.
- Soni, P., Tewari, Y., & Krishnan, D. 2022. Machine Learning Approaches in Stock Price Prediction: A Systematic Review. *Journal of Physics: Conference Series*.
- Wang, Y., & Guo, Y. 2020. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*.
- Stock prices. 2021, September 11. Kaggle. <https://www.kaggle.com/datasets/mysarahmadbhat/stock-prices>
- S&P 500 historical data. 2020, November 5. Kaggle. <https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data>