Machine Learning and Fama-French Three-Factor Model Applications for Analyzing Stock Price in Technology Enterprises

Jiaxiang He *

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suchou, 215123, China

* Corresponding author: Jiaxiang.He21@student.xjtlu.edu.cn

Abstract. This study endeavors to forecast the stock prices of the leading U.S. technology entities - Google, Microsoft, Amazon, Meta, and Apple - through the application of diverse machine learning models, complemented by the traditional Fama-French three-factor model. The employed models encompass Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), Support Vector Machines (SVM), and Decision Tree models. Initially, historical stock price data is utilized to train these machine learning models, enabling the identification of potential price trends. Subsequently, the integration of the Fama-French three-factor model enhances the analysis by scrutinizing the impacts of market risk, company size, and book-to-market value on stock prices. The outcomes illuminate both the effectiveness and limitations of various models in stock price prediction, highlighting the advantages of machine learning methodologies over traditional financial theories. This research provides financial market analysts and investors with a fresh perspective on the amalgamation of machine learning and traditional financial theories for enhanced stock price prediction.

Keywords: Technology Stock, Machine Learning, Fama-French Three-Factor Model.

1. Introduction

In the contemporary digital landscape, technology companies, particularly the top five U.S. entities—Google, Microsoft, Amazon, Meta, and Apple—have emerged as pivotal drivers of the global economy. Innovations in internet technology, software development, e-commerce, and social media position these companies at the forefront of technological advancements. Beyond technological leadership, their stock market influence is substantial, with fluctuations in their stock prices serving as indicators of the health of the tech industry and the broader stock market. Various factors, including financial performance, market trends, global economic conditions, and policy changes, influence the stock prices of these companies [1]. With the rapid progression of information technology, big data analysis and artificial intelligence play an increasingly prominent role in stock price prediction [2]. This study endeavors to explore and validate the efficacy and reliability of different models in predicting actual stock prices for these prominent tech companies, offering valuable insights for investors, market analysts, and paving the way for future research and applications in the fintech sector.

Stock price analysis constitutes a fundamental topic in financial research. Traditional methods like fundamental analysis and technical analysis, while successful to some extent, reveal limitations in the face of complex financial markets and large-scale data dynamics [3]. In recent years, the application of machine learning technologies in financial market analysis has gained prominence, particularly in handling large volumes of data and navigating the nonlinear, high-dimensional, and dynamic features of the stock market [2]. This study employs advanced machine learning models, such as Convolutional Neural Networks (CNN) for complex data patterns, Long Short-Term Memory Networks (LSTM) for long-term dependencies in time series data, Support Vector Machines (SVM) and Decision Tree models for classification and interpretability. Additionally, the study integrates the traditional Fama-French three-factor model, providing a theoretical perspective on stock price movements. The interdisciplinary approach aims not only to enhance the accuracy of stock price prediction but also to offer new insights into the complexity and dynamism of financial markets.

The primary objective of this study is a comprehensive evaluation and comparison of various advanced models for predicting the stock prices of the top five U.S. technology companies. This involves collecting detailed stock price data from recent years, including daily opening, highest, lowest, and closing prices, as well as trading volumes. Various machine learning models are then applied for training and testing, including CNN for feature extraction, LSTM for time dependencies, SVM for nonlinear classification, and Decision Tree models for interpretability. The study also incorporates the Fama-French three-factor model to analyze factors like market risk, company size, and book-to-market value. Performance evaluation metrics, including prediction accuracy, Mean Squared Error (MSE), and Absolute Percentage Error, are employed for comprehensive analysis. This methodological approach aims to provide an innovative analytical framework by combining modern machine learning technology with traditional financial theory, expanding research horizons and contributing to more accurate prediction tools for investors and market analysts.

2. Data Collection

2.1. Data Selection

This study utilizes stock price data from January 1, 2020, to January 1, 2023, for the top five U.S. technology companies: Google, Microsoft, Amazon, Meta (formerly Facebook), and Apple. The chosen timeframe captures market cycles, including the COVID-19 pandemic's impact and subsequent economic recovery. Daily trading data, encompassing opening, highest, lowest, closing prices, and trading volumes, provides a comprehensive view of intraday fluctuations and overall trends. The selection aligns with the companies' sustained global economic and technological influence during the period.

2.2. Data Pre-Processing

Before model training, data preprocessing is an essential step to ensure the quality and usability of data. The data preprocessing process in this study includes several key steps:

(1) Handling missing and outlier values is critical. Missing values on non-trading days are removed, ensuring a dataset with valid trading day information. The Box Plot method identifies and addresses outliers, improving model accuracy by reducing noise in the data [4].

(2) Normalization using min-max scaling is applied to all data, ensuring numerical stability during model training [5]. Feature engineering involves creating new features based on stock market data characteristics and model requirements. Historical average prices, moving averages, and technical indicators like the Relative Strength Index (RSI) enhance the model's ability to identify trends and momentum [6]. For time series models like LSTM, data is divided into fixed-size windows, and lagged features are generated to capture time dependencies.

(3) The dataset is divided into training, validation, and test sets in a 7:1:2 ratio or a 7:3 ratio, depending on the model. This segmentation ensures adequate data for model training, parameter tuning, and evaluation of real-world performance. The training set facilitates initial model learning, the validation set assists in parameter tuning and preventing overfitting, and the test set evaluates the model's predictive performance [7]. Maintaining representativeness and consistency across splits ensures accurate assessment of the model's effectiveness.

Through these preprocessing steps, the dataset is optimized for machine learning modeling and analysis, enhancing accuracy and generalization capabilities. This rigorous approach ensures the reliability and validity of research findings.

3. Modelling

3.1. Support Vector Machine (SVM):

Support Vector Machine (SVM) serves as a generalized class of linear classifiers for binary classification in supervised learning [8]. Assuming linearity, SVM uses a straight line or hyperplane to separate data into distinct classes [9, 10]. In cases of non-linear data, SVM employs the kernel trick to transform it into a higher-dimensional space [11]. In stock market prediction, SVM categorizes stocks based on historical features, efficiently handling non-linear relationships through kernel tricks. Emphasizing maximizing the margin between categories enhances its generalization for unseen data. SVM's merits for stock market prediction include effective handling of non-linearity and suitability for high-dimensional datasets influenced by multiple factors.

3.2. Decision Trees

Decision trees, structured like trees, utilize non-leaf nodes for feature attribute tests and branches for attribute values, with leaf nodes representing classes [12]. The decision process begins at the root node, testing each feature attribute, and navigating to leaf nodes for the decision result. Decision trees are valued in stock market prediction for their interpretability, providing understandable decision-making processes. Their effectiveness in capturing non-linear relationships and patterns in stock data further contributes to their utility.

3.3. Fama-French Three-Factor Model

The FAMA-FRENCH Three-Factor Model explains stock returns based on market risk, size effect (SMB), and value effect (HML). Extending the Capital Asset Pricing Model (CAPM), it introduces SMB and HML as additional risk factors [13]. Market risk (Mkt-RF) reflects the market's excess return rate over the risk-free rate. SMB addresses the historical outperformance of small-cap stocks, while HML captures differential returns between high and low book-to-market ratio stocks [13]. Utilizing these factors, the model aims for a comprehensive explanation of stock returns. Reasons for its use in stock market prediction include risk-adjustment and enhanced explanatory power based on market conditions, size, and value factors.

3.4. Long Short-Term Memory (LSTM)

As a variant of Recurrent Neural Networks (RNN), LSTM specializes in processing and predicting sequence data, overcoming issues like vanishing or exploding gradients in traditional RNNs through gating mechanisms [14]. These mechanisms include input, forget, and output gates, controlling information influx, memory retention, and output determination. LSTM's architecture excels in capturing and retaining long-term dependencies, making it effective for sequence data tasks [15]. In stock market prediction, LSTM proves valuable for its proficiency in sequence pattern recognition and adeptness at handling time dependencies crucial for financial time series analysis.

3.5. Machine Learning - Gradient Boosting

Gradient Boosting, a supervised machine learning algorithm, employs sequentially arranged decision trees correcting errors of predecessors [16]. Each tree in the sequence is trained on residual errors of the previous one, focusing on areas of poor performance. The final prediction is a weighted sum of all trees' predictions. Reasons for using gradient boosting in stock market prediction include its proficiency in predicting stock price trends and efficient handling of non-linear features and patterns common in financial stock data.

3.6. Convolutional Neural Networks (CNN)

CNNs, a type of deep learning algorithm, consist of convolutional, pooling, and fully connected layers [17]. Convolutional layers apply filters to create feature maps, pooling layers reduce spatial size, and fully connected layers generate the output [17]. CNNs are suitable for stock market

prediction due to their effectiveness in pattern recognition, particularly in identifying patterns in stock price trends using one-dimensional convolutions. They excel at handling complexity and noise in financial time series data.

3.7. ARIMA

ARIMA, a statistical model for time series forecasting, comprises Autoregressive (AR), Integrated (I), and Moving Average (MA) parts [18]. AR utilizes dependencies between observations and lagged observations, I make the time series stationary, and MA uses residual errors of a moving average model. The model requires parameters p, d, and q for AR, I, and MA parts. ARIMA is chosen for stock market prediction due to its effectiveness with time series data showing trends or seasonality and its interpretability, with modifiable parameters for model adjustments [18].

4. Empirical Results

4.1. MAE

MAE stands for Mean Absolute Error. It is a metric that measures the average absolute difference between the predicted and actual values in a set of data [19]. The formula for the mean absolute error is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\text{actual}_i - \text{predicted}_i| \quad (1)$$

Where: n is the number of observations or data points, actual is the actual value and predicted is the predicted value of the observations.

MAE provides a straightforward way to assess the accuracy of a predictive model by measuring the size of the average error between predicted and actual values.

Table 1. Results of MAE

MAE/Models	Machine Learning	CNN	ARIMA	LSTM
Google	0.00569	0.00315	0.02912	0.03261
Microsoft	0.01213	0.00385	0.03770	0.03054
Apple	0.01201	0.00330	0.02559	0.02666
Meta	0.00307	0.00291	0.02778	0.01993
Amazon	0.00952	0.00540	0.04348	0.04092

According to the definition of MAE, a smaller MAE value indicates that the model's prediction is closer to the actual value. From Table 1, it is intuitively clear that the CNN model outperforms the other models with respect to the simulation of the stocks of Microsoft, Apple and Amazon. However, in the Google and Meta stock simulations, the CNN model and the Machine Learning model perform about the same. the ARIMA and LSTM models perform consistently overall but the results are not very positive in terms of MAE.

4.2. Confusion Matrix

The confusion matrix is a visualization tool that is used in supervised learning to visually compare classification results with actual measurements. Table 2 shows the results of confusion matrix. Accuracy, Precision and Recall can be easily calculated by using the confusion matrix in Table 3.

Table 2. the results of Confusion Matrix

CM/Models	Machine Learning	CNN	ARIMA	Decision Tree	SVM
Google	[[62 45] [47 42]]	[[73 34] [60 29]]	[[60 47] [49 40]]	[[41 41] [35 35]]	[[36 46] [29 41]]
Microsoft	[[55 51] [44 46]]	[[59 47] [54 36]]	[[41 66] [61 28]]	[[44 40] [39 29]]	[[37 47] [26 42]]
Apple	[[46 56] [52 42]]	[[54 48] [49 45]]	[[54 53] [50 39]]	[[39 41] [37 35]]	[[20 60] [14 58]]
Meta	[[54 46] [51 45]]	[[57 43] [56 40]]	[[48 59] [54 35]]	[[26 52] [28 46]]	[[26 52] [29 45]]
Amazon	[[56 48] [47 45]]	[[61 43] [57 35]]	[[49 58] [54 35]]	[[27 54] [24 47]]	[[21 60] [19 52]]

Accuracy is the number of correct predictions as a percentage of the total number of predictions.

$$\text{Accuarcy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2}$$

Recall (Sensitivity) shows the ratio of true positives to everything positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

Precision is the fraction of positive predicitions is actually correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

Table 3. Results of accuracy, precision and recall

CM/Models	Machine Learning	CNN	ARIMA	Decision Tree	SVM
Google	Accuracy:0.53 Precision:0.48 Recall:0.47	Accuracy:0.52 Precision:0.46 Recall:0.33	Accuracy:0.51 Precision:0.46 Recall:0.46	Accuracy:0.53 Precision:0.46 Recall:0.50	Accuracy:0.51 Precision:0.47 Recall:0.59
Microsoft	Accuracy:0.52 Precision:0.47 Recall:0.51	Accuracy:0.48 Precision:0.43 Recall:0.40	Accuracy:0.35 Precision:0.30 Recall:0.31	Accuracy:0.48 Precision:0.42 Recall:0.42	Accuracy:0.52 Precision:0.47 Recall:0.62
Apple	Accuracy:0.45 Precision:0.43 Recall:0.45	Accuracy:0.51 Precision:0.48 Recall:0.48	Accuracy:0.47 Precision:0.42 Recall:0.44	Accuracy:0.48 Precision:0.46 Recall:0.49	Accuracy:0.51 Precision:0.49 Recall:0.80
Meta	Accuracy:0.51 Precision:0.49 Recall:0.47	Accuracy:0.49 Precision:0.48 Recall:0.42	Accuracy:0.42 Precision:0.37 Recall:0.39	Accuracy:0.47 Precision:0.47 Recall:0.62	Accuracy:0.48 Precision:0.46 Recall:0.61
Amazon	Accuracy:0.52 Precision:0.48 Recall:0.49	Accuracy:0.49 Precision:0.45 Recall:0.38	Accuracy:0.43 Precision:0.38 Recall:0.39	Accuracy:0.48 Precision:0.47 Recall:0.66	Accuracy:0.48 Precision:0.46 Recall:0.73

4.3. F1-Score

In order to measure and compare the accuracy of each model, the F-score (F1 score), which takes into account both the accuracy and recall of the classification model, was used [20]. The F1 score is the harmonic mean of precision and recall:

$$F_1 score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

Table 4. Results of F-score

F1Score/Models	Machine Learning	CNN	ARIMA	Decision Tree	SVM
Google	0.48	0.38	0.45	0.54	0.48
Microsoft	0.49	0.42	0.31	0.47	0.50
Apple	0.44	0.48	0.43	0.52	0.35
Meta	0.48	0.45	0.38	0.39	0.40
Amazon	0.49	0.41	0.38	0.43	0.34

According to Table 4 and the formula of F-score, it is easy to get the F-score of the models. And Table 4 provides a visual comparison of the evaluation results of the different models in terms of F-score.

Machine Learning performs consistently and has the highest F-score of 0.49 on Microsoft and Amazon. The CNN performs inconsistently, getting the highest F-score of 0.49 on apple predictions. however, it gets the lowest F-score of 0.38 on Google predictions. The F-score of ARIMA is stable but not high, getting 0.38 on META and AMAZON. While the Decison Tree was not stable enough but got the highest value of all F-scores 0.54. SVM has the best F-score value on Microsoft.

For Google and Apple stock predictions, the Decision Tree model is available. For Meta and Amazon stock prediction you can choose Machine Learning model. For Microsoft stock prediction you can choose SVM model.

Upon integrating MAE, F1 scores, and confusion matrix data, Gradient Boosting demonstrates robust performance in predicting Google and Amazon stock prices, striking a balance between high accuracy and the ability to classify stock price trends. Meanwhile, CNN excels in capturing time-series patterns in the Meta dataset but falls slightly short in predicting Microsoft data. Conversely, ARIMA performs the least favorably among all models, particularly evident in the low F1 score, especially in Microsoft data, revealing its limitations in handling complex datasets. Support Vector Machines exhibit a better balance in the Google and Microsoft data but struggle in the Amazon dataset, underscoring sensitivity to data features. Decision Trees exhibit higher accuracy in Google and Apple datasets, though overall performance in classifying stock price trends remains uncertain due to the lack of F1 score data. In summary, Gradient Boosting and CNN outperform in stock price prediction, while ARIMA and SVM yield mixed results, necessitating additional data for a definitive assessment of decision tree perfo.

4.4. Fama-French Three-Factor Model

The Fama-French results in Table 5 suggest that the Market (Mkt), Size (SMB), and Value (HML) factors have varying effects on the stock returns of Google, Meta, Apple, Amazon, and Microsoft. The coefficient of MKT (Market Factor) is statistically insignificant for the stock returns of Google, Meta, Apple, Amazon and Microsoft. This indicates that the market factor has no significant effect on the stock performance of these companies. Similarly, SMB (Firm Size Factor) is also not significant in terms of stock returns of these companies. This means that firm size has no statistically significant effect on the stock. It is worth noting that the HML factor for Amazon is the most

statistically significant, and the HML factors for the other companies are also significant to some extent, suggesting a positive correlation with their return. What's more, for Google, Meta, Apple, Amazon, and Microsoft, the values of R-squared are 0.056, 0.100, 0.171, 0.365, and 0.144, respectively. These values indicate the proportion of stock return variance that the model is able to explain. Higher R-squared values usually indicate that the model fits the data better, but do not necessarily mean that the model is valid or has predictive power. Here, lower R-squared values may indicate that the Fama-French model explains relatively little of the variation in stock returns and that there may be other unconsidered factors affecting stock performance.

Table 5. Results of Fama-French

Variables/Models	Google	Meta	Apple	Amazon	Microsoft
Mkt	-0.8803 (0.693)	-1.4668 (1.526)	-0.4787 (0.712)	-0.4714 (0.645)	-1.3386 (1.055)
SMB	-0.9848 (1.328)	-1.8291 (2.927)	-0.9745 (1.365)	-0.3627 (1.236)	-0.8501 (2.023)
HML	1.5467* (0.858)	4.4960** (1.894)	1.8577** (0.883)	2.1881*** (0.800)	2.8045** (1.309)
Observations	417	417	417	417	417
R-squared	0.056	0.100	0.171	0.365	0.144

Overall, for these companies, the effects of the Market Factor (MKT) and Firm Size Factor (SMB) are not significant. This implies that the stock performance of these companies is less influenced by the market as a whole and the size of the company, and that investors should pay more attention to other factors in their decision-making. It is worth paying attention to the value factor (HML). Therefore, investors may consider focusing on the fundamental value of these companies. The relatively low R-squared value indicates that the model has limited ability to explain the variance of stock returns. In making investment decisions, it is important to consider not only the results of the model, but also other factors not included in the model to obtain more comprehensive information.

5. Conclusion

In light of the inherent limitations of traditional financial analysis methods in coping with the intricacies of dynamic financial markets and extensive data, this study adopts machine learning methodologies to scrutinize and substantiate the efficacy and dependability of diverse models in predicting the actual stock prices of prominent technology companies. Leveraging data spanning from January 1, 2020, to January 1, 2023, encompassing share prices of Google, Microsoft, Amazon, Meta, and Apple, the study embarks on a multifaceted exploration.

The initial step involves normalizing the data through min-max scaling, fostering numerical stability during model training. Subsequently, the dataset undergoes segmentation into training, validation, and test sets, following a 7:1:2 or 7:3 ratio, contingent on the specific model requirements. This meticulous preprocessing lays the foundation for comprehensive model experimentation.

The study employs a spectrum of machine learning algorithms, including SVM, decision tree, LSTM, Gradient Boosting, CNN, and ARIMA, to construct predictive models. The Fama-French three-factor model is then employed to assess the impact of market risk, company size, and book-to-market value on the stock prices of the five technology giants. Finally, the evaluation phase employs metrics such as MAE, Confusion Matrix, and F-score to gauge the performance and effectiveness of the models.

In conclusion, this research endeavor synthesizes cutting-edge machine learning techniques with traditional financial theory, providing a holistic approach to stock price prediction for technology companies. The integration of diverse models and rigorous evaluation metrics contributes to a nuanced understanding of the predictive capabilities of these methodologies. This interdisciplinary approach not only broadens the scope of research but also furnishes valuable insights for investors and market analysts navigating the dynamic landscape of technology stock markets.

References

[1] Li, Y., & Pan, Y. (2020). A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics, 13, 139 - 149.

[2] Nguyen, T. H., Shirai, K., & Velcin, J. (2011). Expert Systems with Applications.

[3] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2019). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53, 3007-3057.

[4] Smiti, A. (2020). A critical overview of outlier detection methods. Comput. Sci. Rev., 38, 100306.

[5] Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Appl. Soft Comput., 97, 105524.

[6] Hari, Y., & Dewi, L. P. (2018). Forecasting System Approach for Stock Trading with Relative Strength Index and Moving Average Indicator. Journal of Telecommunication, Electronic and Computer Engineering, 10, 25 - 29.

[7] Rajaraman, S., Jaeger, S., & Antani, S. K. (2019). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ, 7.

[8] Vapnik, V. N. (2000). The Nature of Statistical Learning Theory. In Statistics for Engineering and Information Science.

[9] Cai, Z., & Jiang, G. (2008). Application of multiple SVM classifier fusion technique in freeway automatic incident detection, 581 - 585.

[10] Sanjaa, B., & Chuluun, E. (2013). Malware detection using linear SVM. In Ifost Ulaanbaatar, 136 - 138.

[11] Bhaskar, S., Singh, V. B., & Nayak, A. K. (2014). Managing data in SVM supervised algorithm for data mining technology, 1 - 4.

[12] Abdelhalim, A., & Traore, I. (2009). A New Method for Learning Decision Trees from Rules. In 2009 International Conference on Machine Learning and Applications, 693 - 698.

[13] Womack, K. L., & Zhang, Y. (2003). Understanding Risk and Return, the CAPM, and the Fama-French Three-Factor Model. ESADE Business School Research Paper Series.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9, 1735 - 1780.

[15] Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. ArXiv, abs/1808.03314.

[16] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7.

[17] O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. ArXiv, abs/1511.08458.

[18] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 106 - 112.

[19] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79 - 82.

[20] Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. In International Conference on Language Resources and Evaluation.