UDC 004.94

R. Piznak, T. Likhouzova

MODELS FOR ANALYZING AND FORECASTING SHARE PRICES ON THE STOCK EXCHANGE

Abstract: The work is devoted to the analysis and forecasting of share prices for four leading technology companies: Nvidia, Apple, Google, and Netflix. These companies are leaders in their fields and have a significant impact on the global economy. The goal is to study the dependencies affecting the share prices of companies, as well as to develop models for forecasting future trends. In the work, a thorough analysis of historical data on company share prices and their macroeconomic indicators was carried out. The study was based on the fundamental concepts of economic science. The study results are expected to provide a deeper understanding of the prospects of these companies.

Keywords: intelligent data analysis, prediction model, time series, LSTM, decision tree, ARIMA, accuracy metrics.

Introduction

The stock market plays an important role in the global economy, providing an opportunity for companies to raise capital and for investors to earn income from their investments. The dynamics of share prices are an indicator of the company's financial health and reflect investors' expectations regarding its prospects.

Analyzing and forecasting stock prices is a critical task for both investors and companies themselves, as it allows for informed investment and strategic planning decisions.

According to the World Federation of Exchanges for 2023, the average daily volume of stock trading on all stock exchanges in the world was about 1.15 trillion US dollars. [1] Interestingly, most of the financial transactions on the stock exchange are now carried out precisely by intelligent systems, which are based on modern methods of machine learning and artificial intelligence. Therefore, these methods are becoming increasingly important tools in stock price analysis and forecasting.

Analyzing and forecasting the course on the stock market is a complex and multifaceted task that requires careful study of various factors and the use of a wide range of methods and tools. The following key aspects should be considered: share trading volumes, pricing, and share volatility [2].

ISSN 1560-8956

Materials and methods

For the given task, data was selected in a format standard for any stock exchange: Nvidia [3], Google [4], Apple [5], and Netflix [6]. The datasets contain data on the dynamics of share price changes for each of the studied companies since the IPO (first sale of shares). Each of the datasets has the same structure (Table 1).

Table 1

Name	Description
Date	the date for which the stock data is recorded
Open	the price at the time of opening of trades on this date
High	the maximum price during trading on that date
Low	the minimum price during trading on that date
Close	the stock price at the close of trading on that date
Adj Close	the adjusted (final) price at the close of trading on that date
Volume	the number of shares sold that day

Data pre-processing

The Pandas library [7] was used to work with datasets, and the software application for analysis and modeling was implemented with Python 3.11.

We read each of the datasets into a data frame and check the correspondence between the expected and real data types and the number of rows. In case of a mismatch of data types, we apply type casting. After making sure that there are no missing values, we delete all duplicate lines.

After that, for each of the data frames, we will set the Date column as an index and set the daily frequency for it. This is necessary to increase the convenience of further analysis of the time series. We will add a Daily Return column to each of the data frames, which will indicate the difference between the adjusted price for the current date and the previous one.

At this stage, data preprocessing can be considered complete.

Primary analysis of input data

At this stage, the main task is to understand how some characteristics of trades for each of the companies change over time. It is important to understand whether there is a relationship between these characteristics and how they are distributed in general. We can draw some conclusions from the obtained results:

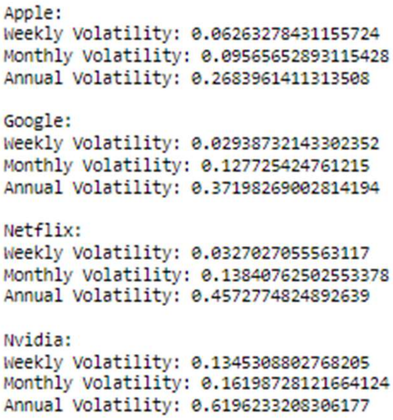
– Apple has been selling its shares for a long time;

– the record maximum price among all companies is held by Nvidia;

– the biggest positive changes in the price for the day – Nvidia;

- the biggest negative price change of the day - Apple.

For each of the companies, we will calculate the average volatility for different periods (Fig. 1). Volatility is a statistical indicator that characterizes the degree of variation in the share price. This is an important indicator for investors and traders, as it is the basis for assessing risks, pricing options, and choosing a trading strategy.



```
Apple:
Weekly Volatility: 0.06263278431155724
Monthly Volatility: 0.09565652893115428
Annual Volatility: 0.2683961411313508

Google:
Weekly Volatility: 0.02938732143302352
Monthly Volatility: 0.127725424761215
Annual Volatility: 0.37198269002814194

Netflix:
Weekly Volatility: 0.0327027055563117
Monthly Volatility: 0.13840762502553378
Annual Volatility: 0.4572774824892639

Nvidia:
Weekly Volatility: 0.1345308802768205
Monthly Volatility: 0.16198728121664124
Annual Volatility: 0.6196233208306177
```

Fig. 1. Calculation of stock volatility

Based on the results, we can draw some conclusions. For investors seeking stability:

– Apple: The lowest annual volatility makes it a relatively stable long-term investment.

– Google: Weekly volatility is at its lowest, making Google stock attractive for short-term investors.

For investors ready for high risk:

– Nvidia: The highest volatility on both the weekly and yearly levels suggests the potential for both significant gains and losses. Investing in Nvidia can be interesting for aggressive investors.

For investors looking for medium risk:

– Netflix: Monthly and annual volatility are at an average level among the companies represented, which can be a balanced option for investors.

For a better understanding of changes in the characteristics of trades over time, we will create appropriate visualizations. We visualize the dynamics of changes in the adjusted price using the matplotlib library [8] (Fig. 2). From the graphs, we can see that over time, the price of shares of companies is increasing, especially this trend is observed for the last few years.

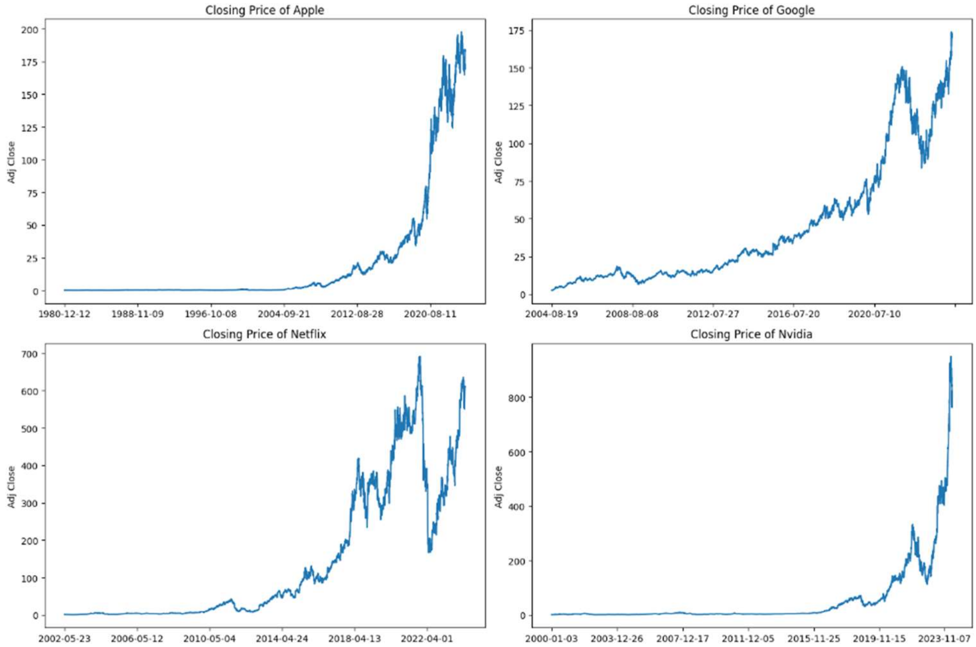


Fig. 2. The dynamics of changes in the adjusted price for each of the companies

To have an idea of the profitability of shares over time, let's analyze the daily price change for each of the companies using histograms. The distributions are nearly symmetric about the point of zero daily income.

For each of the companies, let's examine the relationship between the factors using correlation matrices. There is a weak negative correlation between adjusted price and trading volume for Apple, Google, and Netflix: as trading volume increases, stock price decreases; and for Nvidia, there are no potentially good predictors of adjusted price. Similar to the previous data sets, there is no significant correlation between the trading volume and daily return factors, meaning that the number of shares sold on a given day does not affect the return of the stock on that day.

Let's check whether data about one company can be potential predictors of dynamics for another company from the set. To do this, we will create a correlation matrix for the main characteristics of the available data sets (Fig. 3). We see that the growth of one 'tech giant' is a potentially good predictor of growth for other companies on the list. The correlation between Apple and Google is particularly strong, while the correlation between Netflix and Nvidia is the weakest, but still at a level sufficient for a good prediction.

Let's examine the correlation between companies within the framework of share trading volumes. From the results, we can see that the trading volume of companies is not as good a predictor for other companies as the adjusted price. Again, we see that

Apple and Google have the most similarities, they are potentially good predictors of each other.

Let's investigate the dependence of daily profit for the companies under study. We can see that almost all correlation indicators are positive and moderate, that is, the level of profit from one company's share is a potential predictor of the return on another company's share.

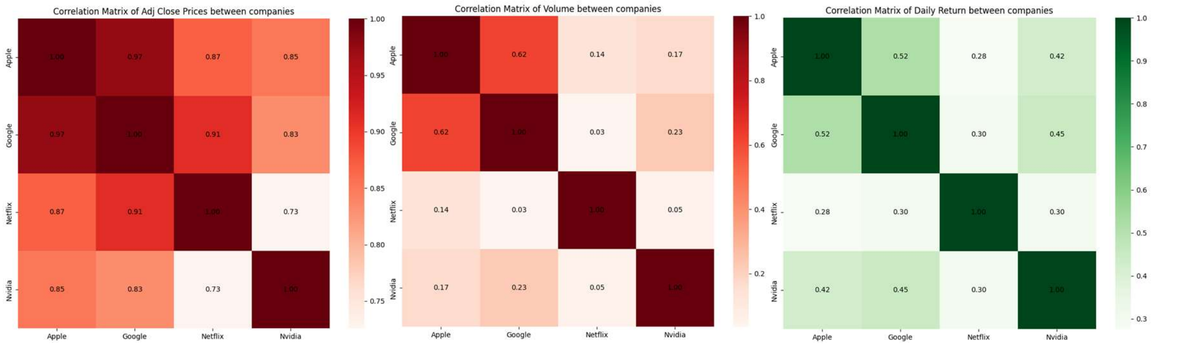


Fig. 3. Factor search

The next step is to decompose the time series. Seasonality will be studied quarterly (90 days), this is a classic term for this kind of problem in the economy. In all cases, there is an upward trend, while there is no clear seasonality within quarters. To test the hypothesis about the stationarity of the series, the extended Dickey-Fuller test was performed, and 120 lags were taken into account. The result showed non-stationarity. For further analysis and modeling of such a non-stationary series, it is necessary to apply methods that ensure stationarity, for example, taking differences.

This is the end of the time series decomposition analysis. Let's add some concepts from classical economics - the "Risk-Profit" graph. This chart is one of the most important and widely used tools for visualizing and analyzing investment portfolios. It demonstrates the relationship between expected return and risk for various investment assets or portfolios.

The X-axis plots the expected return, which is typically measured as the average of historical returns over some time. The Y-axis shows risk, which is usually measured as the standard deviation of returns. Thus, each point on the graph represents a specific investment asset or portfolio with corresponding values of expected return and risk.

The chart helps investors visually assess the trade-off between risk and potential reward for different investment options. Investors generally seek to maximize expected return for a given level of risk or to minimize risk for a given level of expected return. Therefore, the investment opportunities located in the lower right part of the graph are considered the most attractive because they offer a higher expected return with lower risk. Let's build such a graph for our data sets (Fig. 4). From the graph, we can see that the shares of Nvidia and Netflix have the highest average return, but they are accompanied by a greater risk for investors. Google shares have the lowest average return, with the least risk.

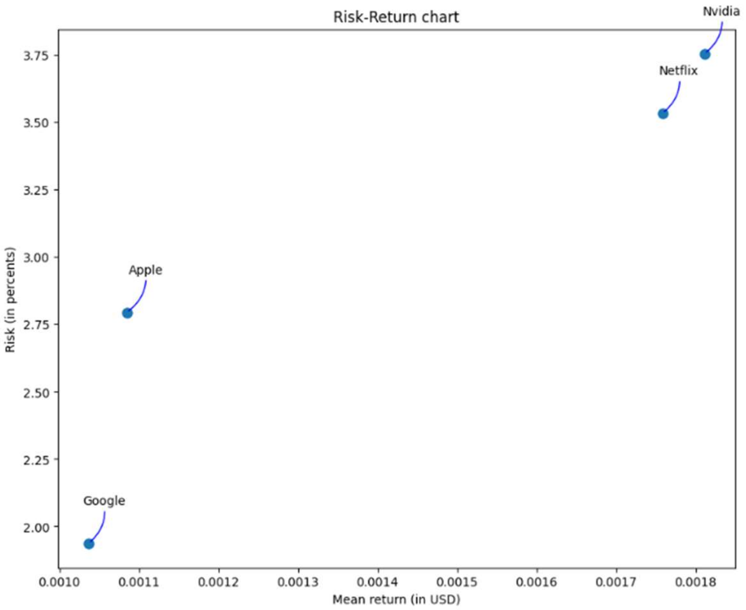


Fig. 4. Risk-Reward

Summarizing the obtained results, we understood the general distribution of trading characteristics, discovered some regularities, and investigated how the factors correlate with each other. For each of the companies, the optimal trading strategy was determined, taking into account the risks and volatility of shares. Based on these results, we can approach the choice of forecasting models with greater understanding and determine the time for which we will create a forecast. Therefore, we will perform a forecast for a period that will be 5% of the entire period of existence of each of the companies. Determining this size as a percentage will be fair, because all companies were founded in different years and, accordingly, there are different numbers of records about each company.

Results and discussion

For the task of forecasting the share price on the stock market, the following methods of intelligent data analysis were chosen: LSTM [9], Decision Tree [10], and ARIMA [11]. These methods complement each other and allow you to get a comprehensive approach to the analysis and forecasting of financial data.

Application of the LSTM model

As part of the application of the model, a study was conducted to identify the optimal structure of the network and find the optimal values of hyperparameters. Networks with different numbers of layers were studied. The number of neurons in each of the layers was also investigated. Special attention was paid to the selection of hyperparameters batch_size and epochs, which determine the size of data processed simultaneously and the number of epochs for training. MSE functions were chosen for loss estimation. It was found that the optimal structure of the neural network was a network with 2 consecutive LSTM layers, with the first layer with 128 neurons extracting low-level features of the sequence, while the second layer with 64 neurons processes them and detects higher-level patterns. After the LSTM layers, two fully connected Dense layers were added, the first with 25 neurons, and the second with 1 output neuron. The optimal packet size is 64, the optimal number of epochs is 5. A smaller packet size does not improve the accuracy, but greatly affects the training time. A larger number of epochs often provokes retraining of the model, and accordingly, a decrease in its accuracy on previously unseen data. We visualize the forecasts obtained with the help of this model. Figure 5 shows a graph of real data and predicted data obtained with the help of a neural network for companies. As you can see, the model predicts the rate quite well, because in many places the graphs coincide, or have a slightly shifted common trend.

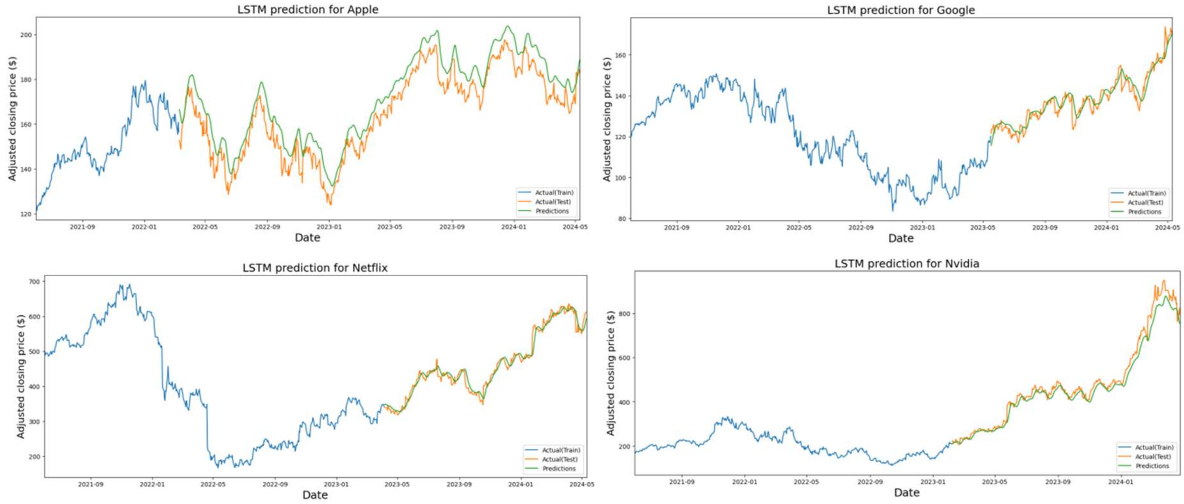
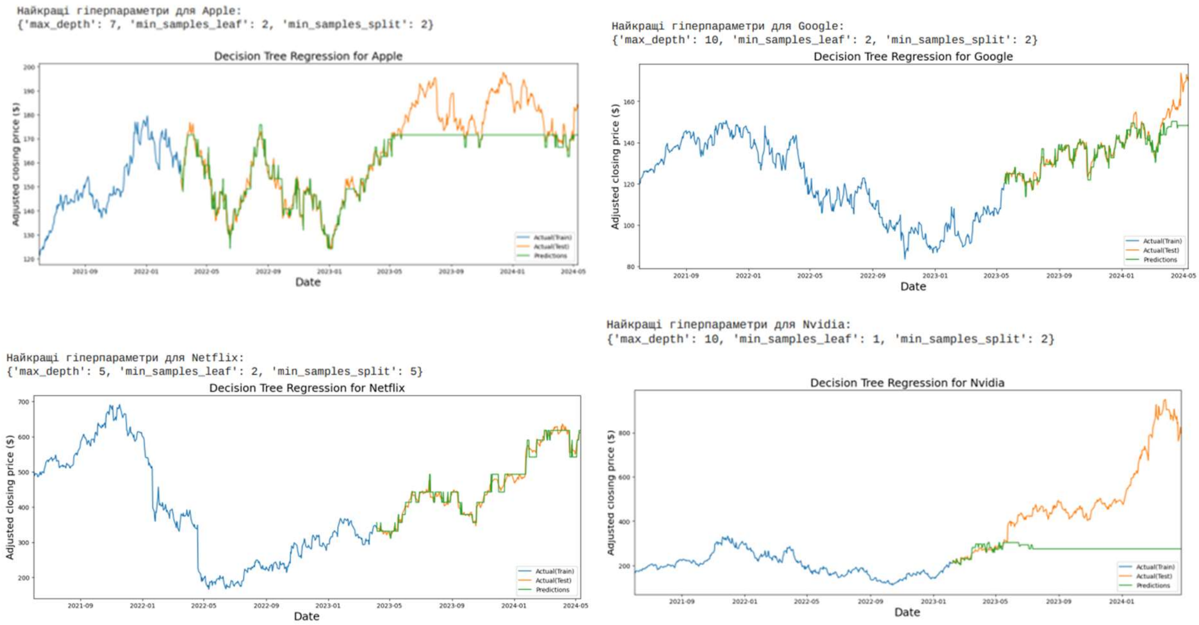


Fig. 5. Graphs of real and forecasted rate, LSTM

Application of the decision tree model

Before obtaining the results, a grid search was used to select the optimal hyperparameters of the model. Hyperparameters of maximum depth 3, 5, 7, 10 were studied; the minimum number of samples required to split a node with possible values of 2, 5, 10; the minimum number of samples that must be present in a leaf node can be 1, 2, 4. For each of the data sets, an optimal set of hyperparameters was chosen. We visualize real and predicted quotes (Fig. 6). Good predictions are observed for some data sets, but there are inaccuracies for data sets with sharp declines/rises in rates.



```
Найкращі гіперпараметри для Apple:
{'max_depth': 7, 'min_samples_leaf': 2, 'min_samples_split': 2}
```

```
Найкращі гіперпараметри для Google:
{'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2}
```

```
Найкращі гіперпараметри для Netflix:
{'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 5}
```

```
Найкращі гіперпараметри для Nvidia:
{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}
```

Fig. 6. Graphs of the real and predicted rate, decision tree

Application of the ARIMA model

The optimal hyperparameters of the model were selected using the auto_arima function from the pmdarima module. For each data set, the optimal set of parameters was saved and a model was created with these parameters. We visualize the real and predicted quotations (Fig. 7). We can conclude that the accuracy of the forecast is low.

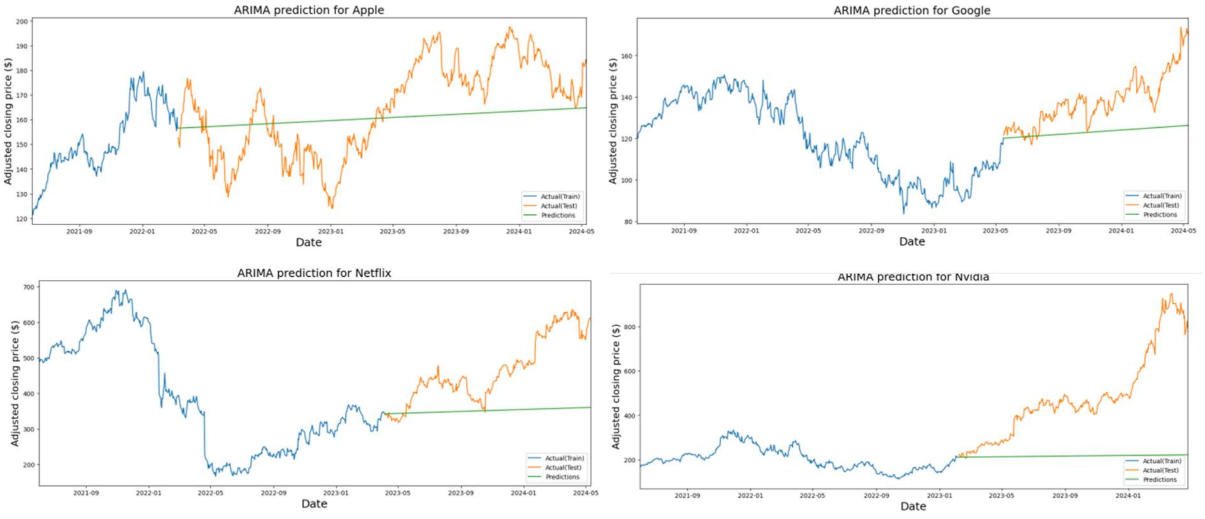


Fig. 7. Application

Model comparison

We will compare all the above methods. We will evaluate the models using the following metrics: RMSE, MAPE, MAE, and R-squared.

The LSTM model showed the best results for forecasting the stock price of Google. This is confirmed by the lowest values of RMSE (3.84), MAPE (2.14%), and MAE (2.91), as well as a high value of R-squared (0.90), which indicates a high quality of forecasting.

For Apple stock, the model performed worse than Google, but still satisfactory, with an RMSE of 8.24, a MAPE of 4.50%, an MAE of 7.27, and an R-squared of 0.80.

The model showed the worst results for Nvidia shares, with high error values: RMSE 31.69, MAPE 4.26%, MAE 22.75. However, the R-squared remains high (0.97), which may be due to the high volatility of share prices.

For Netflix stock, the model performed well in terms of MAPE (2.38%) and R-squared (0.97), but high RMSE (14.92) and MAE (10.82) values, which may be due to the presence of outliers or sharp changes in stock prices.

In general, the LSTM model has demonstrated a satisfactory ability to predict stock prices for most companies, although accuracy varies depending on the specific data and properties of the time series.

Overall, the decision tree model performed worse than the LSTM for most companies, indicating the difficulty of extracting temporal dependencies in stock price data using this method.

For Apple stock, the model showed RMSE of 8.89, MAPE of 3.34%, MAE of 5.94, and R-squared of 0.77, which is acceptable but worse than LSTM.

The results for Google are also inferior to LSTM, with an RMSE of 5.6, a MAPE of 2.24%, an MAE of 3.31, and an R-squared of 0.78.

Netflix has the best performance among all companies, with an RMSE of 13.58, a low MAPE of 2.24%, an MAE of 10.35, and a high R-squared of 0.98.

However, for Nvidia stock, the decision tree model was completely ineffective, with very high errors (RMSE 280.69, MAPE 35.63%, MAE 210.26) and a negative R-squared of -1.12, indicating poor predictive quality. Unsatisfactory results for Nvidia may be due to the high volatility and complexity of the stock price time series, which is difficult to capture with a simple decision tree model.

Hence, a decision tree model may be appropriate for forecasting less volatile time series of stock prices, as in the case of Netflix. However, for most companies, it is inferior in performance to more sophisticated methods such as LSTM, which are better suited to detect temporal dependencies. The decision tree is particularly bad at dealing with highly volatile and complex time series, as in the case of Nvidia.

The ARIMA model performed very poorly for all companies compared to the LSTM and the decision tree model. This indicates that the linear statistical ARIMA model is not able to adequately capture the complex nonlinear dependencies and volatility in stock price time series.

Apple stock performs relatively well with RMSE 17.44, MAPE 9.18%, and MAE 15.06, but a very low R-squared of 0.11, indicating poor forecasting quality.

For Google and Netflix, the results are even worse, with high errors (RMSE around 18 and 141, respectively, MAPE 10.22% and 22.51%, respectively) and negative R-squared values around -1.25 and -1.5, respectively.

The worst ARIMA results were for Nvidia stock with catastrophic errors: RMSE 326.09, MAPE 48.01%, MAE 265.3, and a very low R-squared of -1.87.

The negative R-squared values for most companies indicate that the sample mean of stock prices is a better predictor than the ARIMA model itself.

Such poor results may be because ARIMA is based on linear assumptions and cannot adequately model complex stock price time series with non-linear effects, outliers, and high volatility.

Hence, the ARIMA model proved to be ineffective for forecasting stock prices compared to LSTM and decision trees. Its linear nature and assumptions about stationarity and normality do not match the properties of real financial data. More sophisticated nonlinear methods capable of extracting complex temporal dependencies, such as LSTM or tree-based ensemble models, must be used to obtain accurate stock price forecasts.

In this work, a comprehensive analysis and forecasting of stock prices for four leading technology companies: Nvidia, Apple, Google, and Netflix are carried out. The work aimed to develop effective models for predicting future trends.

The results showed that the LSTM model showed the best performance for forecasting stock prices, especially for companies with relatively stable dynamics like Google. Decision trees also showed acceptable results for some companies but were inferior to LSTMs for more volatile time series. The ARIMA model proved ineffective for this task due to its linear nature and inability to capture complex nonlinear effects in financial data. The obtained results can be used both by investors and by the companies themselves to make more informed decisions and develop effective strategies.

Conclusion

Developing models to forecast stock prices, applying machine learning algorithms, processing big data, and discovering hidden patterns can significantly improve the accuracy of forecasts and help make more informed investment decisions. Finally, effective portfolio management, which includes risk assessment and hedging strategies, as well as constant portfolio monitoring and re-evaluation, is an integral part of successfully investing in stocks of technology giants such as Nvidia, Netflix, Google, and Apple.

References

1. Prospects of the world economy in 2023. – URL: https://niss.gov.ua/doslidzhennya/mizhnarodni-vidnosyny/perspektyvy-svitovoyi-ekonomiky-u-2023-rotsi
2. Best Overall: Economics in One Lesson. Henry Hazlitt 1988. – URL: https://books.google.com.ua/books/about/Economics_in_One_Lesson.html?id=TOCNEAAAQBAJ&redir_esc=y
3. Kaggle.Datasets – URL: https://www.kaggle.com/datasets/prajwaldongre/nvidia-corp-share-price-2000-2024?rvi=1
4. Kaggle.Datasets – URL: https://www.kaggle.com/datasets/nikhil1e9/netflix-stock-price?select=GOOGLE_daily.csv
5. Kaggle.Datasets – URL: https://www.kaggle.com/datasets/nikhil1e9/netflix-stock-price?select=APPLE_daily.csv
6. Kaggle.Datasets – URL: https://www.kaggle.com/datasets/nikhil1e9/netflix-stock-price?select=NETFLIX_daily.csv
7. Documentation of the library Pandas. – URL: https://pandas.pydata.org/docs/
8. Documentation of the library Matplotlib. – URL: https://matplotlib.org/stable/
9. LSTMs Explained: A Complete, Technically Accurate, Conceptual Guide with Keras. – URL: https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2
10. Decision Trees in Machine Learning. – URL: https://www.coursera.org/articles/decision-tree-machine-learning
11. What is an ARIMA model? – URL: https://towardsdatascience.com/what-is-an-arima-model-9e200f06f9eb