Predictive Analytics for NVIDIA Stock Market Trends

Hunter Fernandez, Dexter Schincke

Bellevue University

DSC 630-T301

**Milestone 4**

Data Preparation

In Milestone 4, we built upon the foundational work completed in Milestone 3, focusing on data preparation, model building, and evaluation. We began by gathering and preprocessing the data on NVIDIA's stock movements from the provided CSV file, utilizing Polars for efficient data handling. After loading the dataset, we ensured it was sorted by date and checked for any missing values, confirming the data's integrity.

To enhance our predictive capabilities, we created a lagged feature called prev\_close, representing the previous day's closing price. This was accomplished using the shift() function, followed by the removal of rows without valid previous close values. To standardize our features, we normalized the Close and Volume columns using MinMaxScaler from scikit-learn. Additionally, we constructed a binary target variable, target, indicating whether the stock price increased from the previous day.

Model Building and Evaluation

For modeling, we selected key features: prev\_close and Volume. We then split the data into training (80%) and testing (20%) sets to evaluate our models effectively. We implemented three different models: Logistic Regression, Random Forest, and LSTM.

The Logistic Regression model achieved an accuracy of 0.52, demonstrating perfect recall (1.00) but struggling with false positives, indicating it could mislead investors about potential increases. The Random Forest model, optimized using GridSearchCV, resulted in an accuracy of 0.51, reflecting similar challenges in distinguishing between classes. The LSTM model mirrored the performance of the Logistic Regression model, highlighting its ability to capture trends but with limitations in precision and overall accuracy. These results collectively revealed a common trend: while the models could identify actual price increases, they struggled with precision, emphasizing the need for improved reliability in predictions.

Interpretation of Results

The results from our modeling efforts indicate a critical insight: while our models show potential in predicting stock price movements, their high recall coupled with low precision presents a challenge. This suggests that while the models can successfully identify days when the stock price increased, they often misclassify days when it did not. This misclassification can lead to investors being misled about potential upward movements, ultimately affecting their investment decisions.

The performance analysis of the three models provides valuable insights into their ability to predict NVIDIA’s stock price movements. The Logistic Regression model shows a perfect recall of 1.00, which means it captures every actual stock price increase. However, it falls short in accuracy and precision, both sitting at 0.52. This indicates that while the model identifies positive cases effectively, it misclassifies a significant number of negative instances, leading to a high rate of false positives. This kind of misclassification can be a real headache for investors, as it might give them the wrong impression that a stock price increase is likely when it isn’t.

On the other hand, the Random Forest model records an accuracy of 0.51, a precision of 0.52, and a recall of 0.53. These numbers suggest that it struggles to accurately classify stock price movements, failing to recognize both actual price increases and avoiding false positives. Its relatively low scores in precision and recall indicate that it’s not performing efficiently when it comes to identifying true positives.

The F1 Score is a key metric that helps us balance precision and recall, which is especially important in stock price prediction where we often deal with imbalanced data. In this visualization, we can see the F1 Scores for our three models: Logistic Regression, Random Forest, and LSTM. Both the Logistic Regression and LSTM models have an F1 Score of 0.68, showing they do a solid job of identifying positive stock price movements while keeping false positives in check. On the other hand, the Random Forest model lags behind with an F1 Score of 0.51, suggesting it has a tougher time balancing the detection of true positives and avoiding false positives. This really emphasizes the need to look at the F1 Score along with other metrics when we assess how well our models perform, especially since accurate predictions of stock price increases are crucial for our investment strategies.

A graph showing different colored squares

Description automatically generated

The LSTM model mirrors the performance of Logistic Regression, boasting a recall of 1.00 but also showing the same low accuracy and precision scores. Like Logistic Regression, it effectively identifies all stock price increases but misclassifies many downward movements, as reflected in its F1 score. While LSTM models are usually great at capturing sequential patterns in time-series data, in this case, it seems like it’s not fully utilizing that capability.

Conclusions and Recommendations

There’s plenty of room for improvement across all models, as demonstrated in the graph comparing their precision, recall, and accuracy. This visualization clearly highlights the strengths and weaknesses of each model: the Logistic Regression and LSTM models show high recall, successfully capturing price increases, but their precision and accuracy reflect substantial misclassification of negative cases. Meanwhile, the Random Forest model shows a more balanced yet lower performance across all metrics, suggesting that none of the models currently provide an optimal solution for stock prediction.

A graph of different colored bars

Description automatically generated

Additionally, utilizing hyperparameter tuning and optimizing model parameters could significantly enhance the Random Forest and LSTM models. Techniques like grid search or random search could help pinpoint the most effective configurations for these models, making them more powerful.

From a business perspective, it’s crucial to prioritize a higher precision score, especially if the goal is to make financial decisions based on predicted stock price increases. A model with high precision means that when it forecasts a price increase, it’s more likely to be correct, which helps minimize the risk of overly optimistic predictions that could lead to poor investment choices. However, if the focus shifts to minimizing missed opportunities, we should aim to improve recall to ensure that the model captures a wider range of stock price increases.

Among the models tested, Logistic Regression stands out as the most promising choice for stock movement predictions, despite its limitations. Its high recall (1.00) ensures that it captures every actual stock price increase, a crucial metric for investment strategies focused on upward trends. While its lower precision and accuracy require caution, Logistic Regression’s reliable identification of increases suggests it could be optimized with further feature engineering and data refinement to improve its precision. This model’s performance makes it the most practical starting point, with adjustments, for investors seeking a balance between caution and opportunity in forecasting NVIDIA’s stock trends.

**Milestone 3**

**Introduction**

The semiconductor industry is crucial for pushing technology forward, with NVIDIA standing out as a major company in GPUs and AI hardware. Their innovations, especially with AI-focused chips, have significantly impacted everything from gaming to AI research. Launching new GPU generations, like the GTX and RTX series, has caused noticeable shifts in NVIDIA's stock price.

In this project, we will investigate how technological breakthroughs, particularly NVIDIA's AI-powered RTX chips, have influenced their stock price movement. Rather than predicting exact future prices, we will focus on whether the stock price will likely go up or down based on past data. We will also look at how big events, such as the release of the RTX series compared to the earlier GTX series, affect stock trends. By exploring these connections, our study will shed light on how tech advancements impact investor behavior and market reactions, offering valuable insights for financial analysts and those following the tech sector.

**Data**

The dataset for this project is centered on historical stock data for NVIDIA (NVDA). The data includes key stock market variables such as:

-          Date: The trading date.

-          Open: NVIDIA stock's opening price.

-          High: The highest price recorded during the trading day.

-          Low: The lowest price recorded during the trading day.

-          Close: The final price at the close of the trading day.

-          Adjusted Close: The closing price is adjusted for corporate actions, such as stock splits and dividends.

-          Volume: The number of shares traded on that day.

**Model Selection**

To predict NVIDIA's stock price movement, we will evaluate the following models:

-          Logistic Regression: This will classify price movements as increases or decreases based on historical data, serving as a baseline for binary classification.

-          LSTM (Long Short-Term Memory): This recurrent neural network will analyze historical trends to forecast future price direction, considering short-term and long-term patterns.

-          Random Forest Classification: This algorithm will handle complex interactions and non-linear relationships to predict stock price movements using features like historical prices, trading volume, and price volatility.

These models will be tested to assess their effectiveness in predicting NVIDIA's stock price direction, offering a range of traditional and advanced methods.

**Why these Models?**

We chose these models because of their unique strength in predicting stock price movements. Logistic Regression gives us a simple baseline for classifying price changes. LSTM is excellent at capturing long-term patterns and time-based trends through deep learning. Random Forest handles complex, non-linear relationships and interactions, like price volatility and trading volume. By testing these models, we can compare their performance and determine which approach works best for forecasting stock prices.

**Evaluation of Results**

We will use several metrics to evaluate and ensure accurate predictions of NVIDIA's stock price movements. Accuracy will measure the proportion of correct predictions for classification models like Logistic Regression and Random Forest. Precision and recall will assess how well the models predict upward or downward movements, with precision focusing on the correctness of predicted movements and recall on identifying actual movements. The F1 score will balance precision and recall, particularly for imbalanced classes. AUC-ROC will evaluate how well the models distinguish between price increases and decreases. Lastly, cross-validation will help prevent overfitting by assessing model performance across different data splits. These metrics will guide us in selecting the most effective model for predicting stock price directions.

**Learning Objectives**

This project aims to deepen our understanding of the factors influencing NVIDIA's stock price movements by analyzing fundamental variables such as trading volume, volatility, and major product releases. Through this analysis, we seek to identify the elements that significantly drive stock price increases or decreases. Additionally, by comparing models like Logistic Regression, LSTM, and Random Forest, we aim to determine which algorithms are most effective for predicting stock price movements. The project will also help refine our skills in classification-based modeling and machine learning techniques, enabling us to enhance our predictive capabilities within financial markets. Ultimately, we expect to develop a model that accurately predicts stock price movements while providing insights into the behavior of NVIDIA's stock about technological advancements.

**Risks**

This project faces several risks that must be carefully managed. One critical risk is overfitting, where the model may perform well on historical data but needs to generalize to new, unseen data. To address this, we will employ cross-validation techniques and avoid using overly complex models. Another concern is data quality; inaccuracies or missing data in the historical stock dataset could lead to flawed predictions. We will use thorough data preprocessing and validation steps to mitigate this risk and improve the model's reliability.

**Ethical Concerns**

There are also ethical considerations to keep in mind for this project. One significant concern is market manipulation. The developed predictive models could be used for unethical purposes, such as manipulating stock prices. While this project is for educational purposes, it is essential to consider the broader implications of sharing predictive models. Additionally, data privacy becomes another concern if we decide to integrate sentiment analysis from social media or financial reports. We must ensure that any external data sources comply with privacy regulations to avoid misuse of personal information.

**Contingency Plan**

Suppose the initial project fails to deliver the expected results. In that case, we will pivot to a predictive analytics project using an NCAA dataset containing information on sports participation, revenues, expenses, and enrollment at various institutions. The alternative project will focus on predicting the financial performance of sports programs based on participation numbers and other relevant factors. Similar preprocessing steps will be applied, such as handling missing data and feature engineering and testing models like linear Regression and random forest to forecast revenue and expenses. This pivot will allow us to apply the same machine-learning techniques in a different domain.

**Conclusion**

This project aims to predict NVIDIA's stock price movements by leveraging predictive analytics models, such as Logistic Regression, LSTM, and Random Forest Classification. By focusing on how technological innovations, like the release of NVIDIA's AI-powered RTX chips, have influenced investor behavior, we seek to uncover patterns in stock market reactions. The project will evaluate short- and long-term trends using key metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. These evaluations will guide us toward identifying the most effective model for predicting stock price movements.

We have carefully considered risks such as overfitting and data quality, with plans for cross-validation and robust data preprocessing. Ethical concerns, especially around market manipulation and data privacy, are also addressed, ensuring the models remain responsibly used. Suppose the stock price prediction project does not yield satisfactory results. In that case, we are prepared to pivot to a backup project focused on predicting the financial performance of NCAA sports programs, applying similar machine learning techniques.

Overall, this project will enhance our understanding of stock market dynamics and technological advancements, refining our forecasting techniques and providing actionable insights for financial analysts and investors.

Sources

HRTERHRTER. (2024). \*Nvidia stock historical data\* [Data set]. Kaggle.

<https://www.kaggle.com/datasets/programmerrdai/nvidia-stock-historical-data>

*Nvidia announces financial results for second quarter fiscal 2025.* NVIDIA Newsroom. (2024,

August 29). <https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for->second-quarter-fiscal-2025