NVIDIA Stock Movement Prediction – Final Report

Fabien SPORTOUCH

Guillaume ROUGIER-CAZANAVE-PIN

Allan VU

Julie WALLET

Table of Contents

1. Introduction and Objective

This project investigates the short-term prediction of NVIDIA's stock price direction—specifically whether the stock will go up or down the next trading day. Given the volatility of stock markets, even marginal predictive power can offer substantial value to traders and analysts. Using historical price data, technical indicators, and external market variables, we apply machine learning classification models to identify actionable patterns.

Our goal is not to forecast exact stock prices, but to offer directional insight that could assist in strategic decision-making. The approach follows the CRISP-DM methodology and compares several models in terms of accuracy, F1-score, and interpretability.

2. CRISP-DM: Business Understanding

The core business question is: *Can we predict NVIDIA's short-term stock movement (up/down) using machine learning?* This is a classification problem with direct applications in finance, especially in algorithmic and tactical trading strategies.

While markets are influenced by news, sentiment, and macroeconomic events, technical indicators and correlated asset movements often contain informative signals. If captured effectively, these signals could improve the quality of trading decisions.

Key stakeholders who might benefit from this analysis include retail traders, portfolio managers, and fintech platforms offering predictive analytics.

Why classification?

- Predicting "Up" or "Down" is more interpretable and realistic for short-term decision-making
- It avoids overfitting compared to noisy price prediction
- Easier to evaluate using metrics like **accuracy** and **F1-score**

3. CRISP-DM: Data Understanding

We utilized two main categories of data:

1. **NVIDIA Historical Stock Data (2014–2024)**
   Sourced from Yahoo Finance, this includes OHLCV data (Open, High, Low, Close, Volume).
   Source: https://finance.yahoo.com/quote/NVDA/history
2. **External Features**
   Using the yfinance library, we collected returns for:
   - Major indices (S&P 500, Nasdaq, CAC 40, DAX)
   - Commodities (Gold, Oil)
   - Currencies (EUR/USD, BTC/USD)
   - Peer stocks (Apple, AMD, Intel, QQQ)

We computed technical indicators with the ta library:

- Relative Strength Index (RSI)
- MACD and Signal Line
- Bollinger Bands
- Commodity Channel Index (CCI)

All data was daily and aligned to the same time range. External variables were added to capture broader market dynamics potentially influencing NVIDIA's stock behavior.

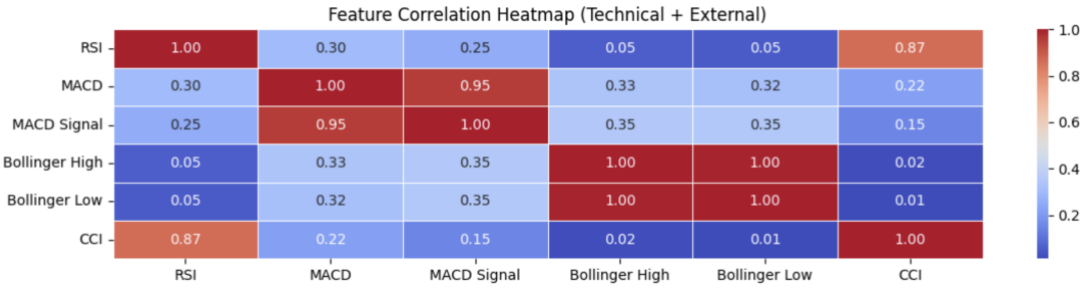
4. Preprocessing and Feature Engineering

The following preprocessing steps were applied:

- Missing values were forward-filled; rows with insufficient data post-indicator calculation were dropped.
- Daily returns were calculated for all external assets and merged into the main dataset.
- Features were normalized where necessary, particularly for models sensitive to scale (e.g., SVM and Logistic Regression).
- A new binary target variable was created:
  - **1** if the next day's close price was higher than the current day's
  - **0** otherwise

We retained only features known at the close of each day to avoid lookahead bias.

Correlation Heatmap of Features:
A visual showing correlations among technical indicators and external features to detect redundancy or spurious relationships.



5. Modeling

We trained three mandatory models plus an optional ensemble model:

Model	Type	Scaling Applied	Strengths
Logistic Regression	Linear	Yes	Fast, interpretable
Decision Tree	Non-linear	No	Handles non-linearity well
Support Vector Classifier (SVC)	Non-linear	Yes	Robust to overfitting (with tuning)
Random Forest (optional)	Ensemble	No	Captures complex patterns

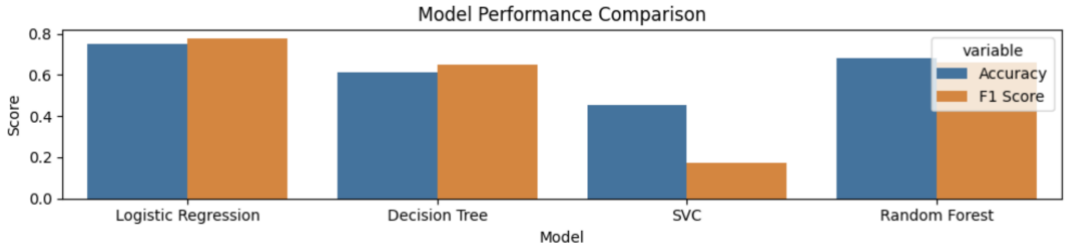
6. Evaluation Metrics and Results

We used the following evaluation metrics:

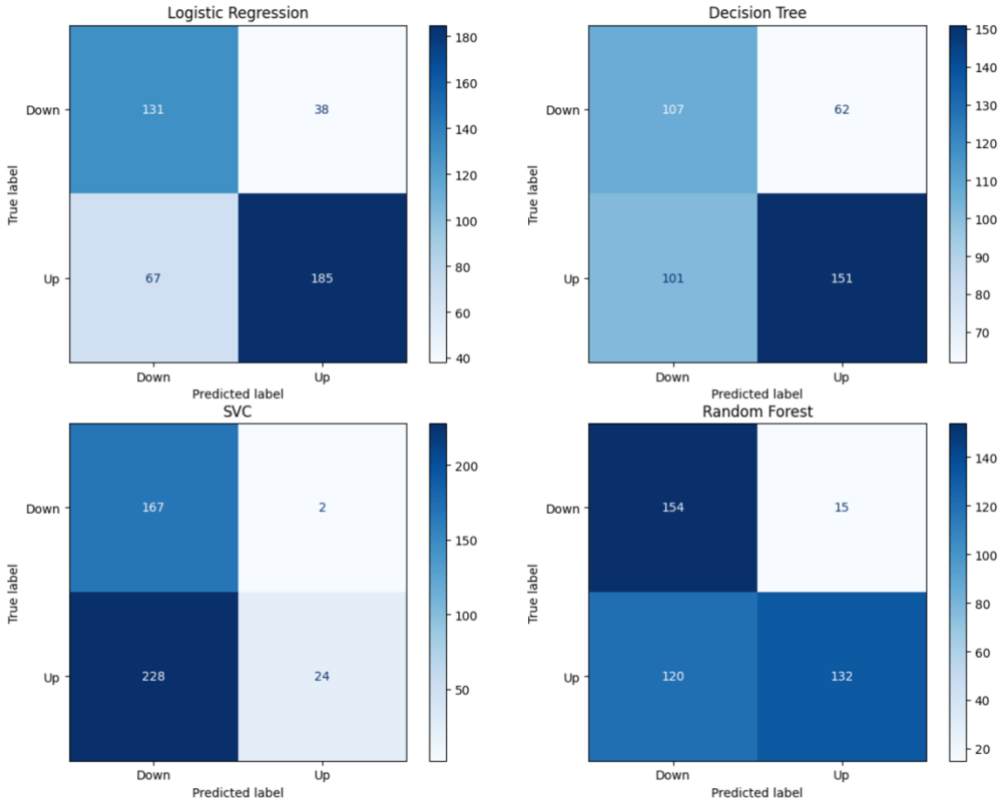
- **Accuracy**: Measures overall correctness.
- **F1 Score**: Balances precision and recall, crucial in imbalanced datasets.
- **Confusion Matrix**: Provides insight into true positives/negatives and model biases.

Model Performance Summary

Model	Accuracy	F1 Score
Logistic Regression	0.75	0.78
Decision Tree	0.61	0.65
SVC	0.45	0.17
Random Forest	0.68	0.66



Confusion Matrices

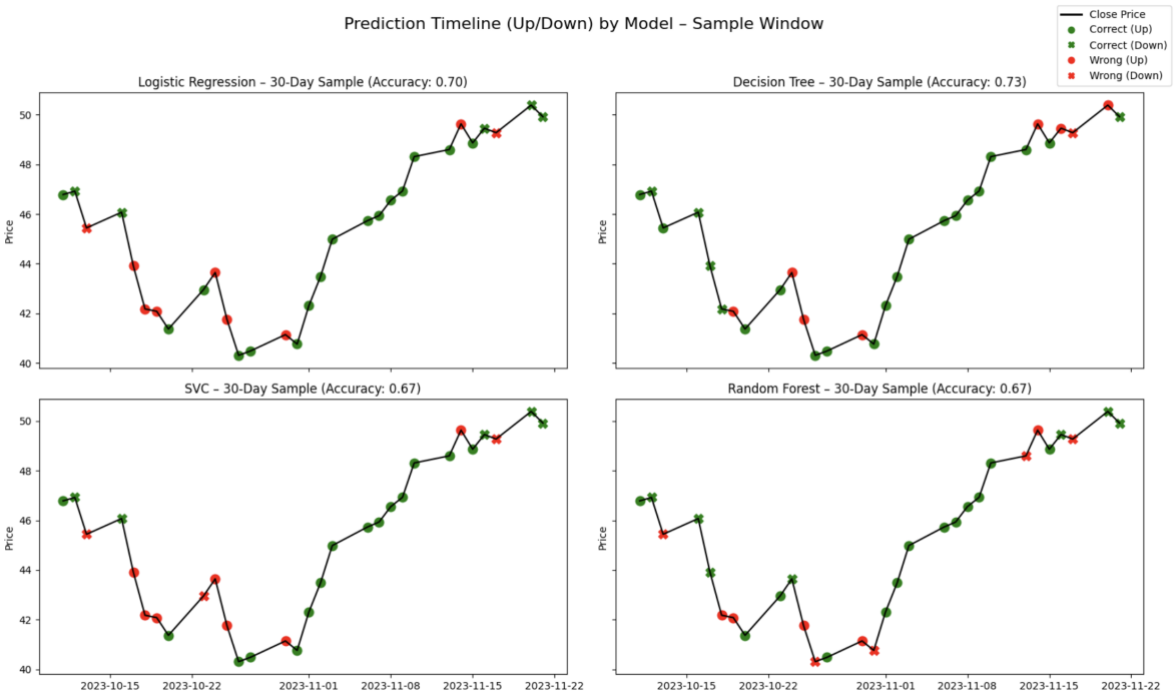


7. Visualizing Predictions

To validate the model behavior visually, we selected a 30-day sample window and plotted predicted vs. true movements for each model.

Each point represents a day:

- Green: Correct prediction
- Red: Incorrect prediction



This qualitative analysis reveals that while models generally follow the price trend, misclassifications often occur during flat or reversal phases.

8. Insights and Conclusions

Our analysis shows that predicting short-term stock price direction is feasible using a mix of technical indicators and external market signals. Among the models tested, Logistic Regression consistently outperformed the others, offering both accuracy and robustness. Its simplicity allowed it to generalize well on unseen data, despite the non-linear nature of financial markets.

Random Forest also performed well, reinforcing the idea that ensemble methods can capture more complex patterns without significant overfitting. Conversely, Decision Trees alone lacked generalization power, and SVC underperformed, likely due to its sensitivity to feature scaling and parameter tuning.

The feature importance analysis highlighted the predictive strength of indicators like MACD and RSI, as well as external signals such as Bitcoin and QQQ returns, confirming that short-

term movements are influenced not just by historical price patterns but also by broader market and sentiment signals.

Overall, our results suggest that even simple classification models can provide actionable insights when engineered with relevant, well-structured features. This opens the door to further experimentation with more advanced models and real-time trading strategies.

9. Ethical and Practical Considerations

While machine learning models can support trading decisions, several limitations must be acknowledged. First, all models rely on historical data and thus offer no guarantee of future performance—market dynamics may shift in ways that models cannot anticipate. Furthermore, our approach is pattern-based and does not imply any causal relationship between the features and the target, meaning predictions should not be interpreted as grounded in economic theory.

Model transparency also plays a role in practical deployment. While linear models like Logistic Regression offer a degree of interpretability, ensemble methods such as Random Forest require additional techniques to ensure explainability, especially in high-stakes environments like finance.

On the ethical side, our project avoided the use of any personal data, thereby mitigating privacy concerns. Ultimately, these models should be seen as decision-support tools—useful for augmenting analysis, but not as a replacement for financial expertise or rigorous due diligence.

10. Team Contribution

Guillaume: Feature engineering, modeling, and hyperparameter tuning

Julie: Report writing, structure, and results analysis

Fabien: Exploratory data analysis, correlation insights

Allan**:** External data integration, support with evaluation metrics and confusion matrices