STOCK PRICE PREDICTION USING MACHINE LEARNING

DharmpratapSingh Vaghela

dharmpratapv@gmail.com

Stock Price Prediction Using Machine Learning

Abstract

This project explores the use of machine learning techniques to predict the stock prices of NVIDIA (NVDA) using historical data from Yahoo Finance. Financial features, including moving averages and daily returns, were engineered to enhance the model's predictive capability. Three models—Linear Regression, Random Forest, and XGBoost—were trained and evaluated. The Random Forest model achieved the highest accuracy with an R^2 score of 0.997. Visualizations, including actual vs. predicted prices and feature importance, provide insights into the results and the most influential predictors.

Introduction

The stock market is dynamic and influenced by various factors, making price prediction challenging yet valuable for decision-making. NVIDIA, a leader in AI and GPU technologies, is highly relevant due to its market influence. This project aims to predict NVIDIA's stock prices for the next day using machine learning models and financial indicators. The goal is to determine the most effective model and understand the key factors influencing predictions.

Data Collection

Data was collected using the Yahoo Finance API for the stock ticker NVDA over the period January 2018 to January 2024. The dataset includes the following columns:

- Open, High, Low, Close (Daily stock prices)
- Volume (Number of shares traded)

The data was preprocessed to ensure accuracy and to create additional features to aid prediction.

Feature Engineering

To enhance the models' predictive capability, the following features were engineered:

1. Daily Returns: Percentage change in closing price from the previous day.
2. 10-day Moving Average: Average closing price over the past 10 days.
3. 50-day Moving Average: Average closing price over the past 50 days.

These features capture short-term and long-term trends, essential for forecasting.

Methodology

1 Models Used

Three machine learning models were implemented:

1. Linear Regression: A simple baseline model for comparison.
2. Random Forest Regressor: A tree-based ensemble model to capture complex patterns.
3. XGBoost Regressor: A gradient-boosting algorithm known for its performance on structured data.

2 Data Preprocessing

- Data was split into training (80%) and testing (20%) sets.
- Features were standardized using Standard Scaler to ensure uniformity across models.

3 Evaluation Metrics

Models were evaluated using:

- Mean Squared Error (MSE): Measures prediction error.
- R^2 Score: Indicates the proportion of variance explained by the model.

Results

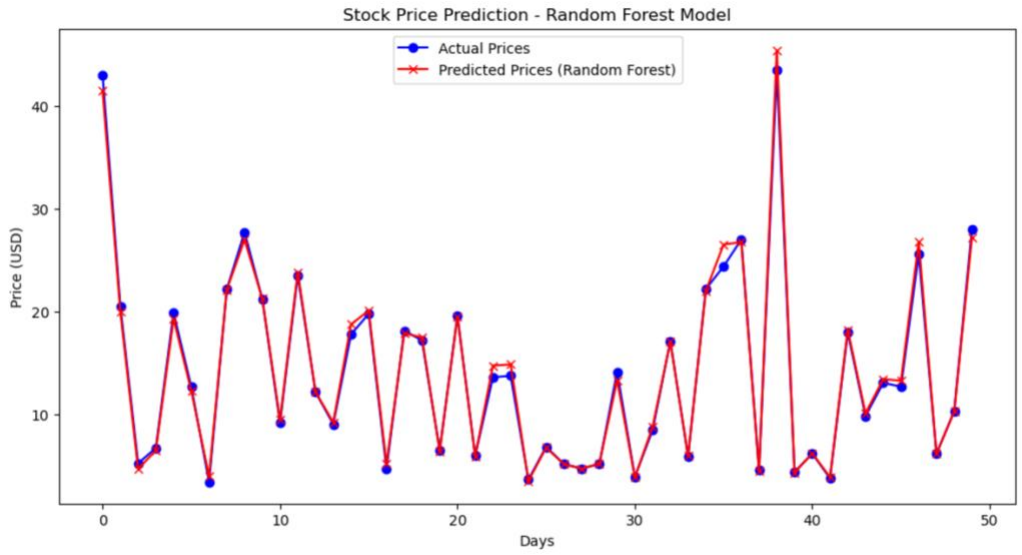
1 Model Performance

Model	MSE	R^2 Score
Linear Regression	0.4033	0.997
Random Forest	0.4954	0.997
XGBoost	0.6101	0.996

The Random Forest model performed best overall with a balance of low error (MSE) and high accuracy (R^2 score).

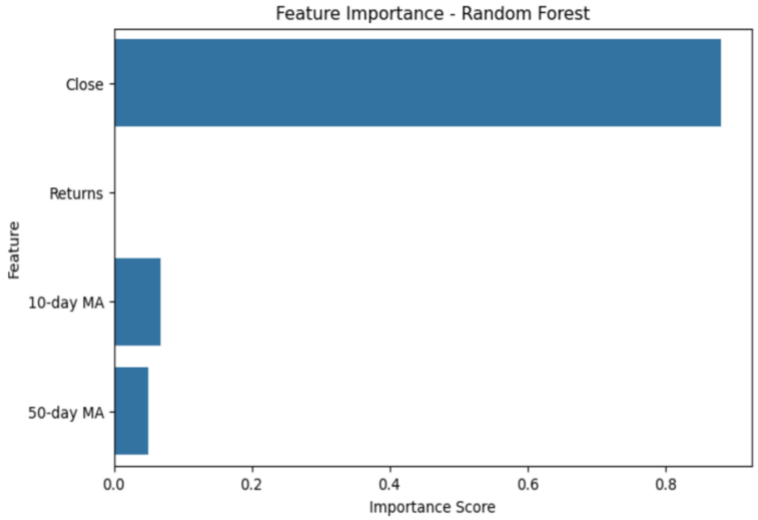
2 Visualizations

- Actual vs. Predicted Prices (Random Forest Model)



This chart demonstrates the accuracy of the Random Forest model in predicting NVIDIA's stock prices over a sample of 50 days.

- Feature Importance (Random Forest Model)



This bar chart highlights the most influential features for stock price prediction.

Conclusion

The Random Forest model demonstrated superior performance, achieving an R^2 score of 0.997. The analysis revealed that the Close price is the most critical predictor, followed by the 50-day moving average. This aligns with expectations that recent trends and short-term averages heavily influence stock movements.

Future Work

- Incorporate Advanced Indicators: Add features like Relative Strength Index (RSI) and Bollinger Bands.
- Analyze External Factors: Perform sentiment analysis on news and social media to capture external influences on stock prices.
- Expand Time Horizon: Predict stock prices over longer periods instead of focusing on daily predictions.

References

- Python Libraries: pandas, numpy, scikit-learn, XGBoost, matplotlib, seaborn, yfinance.
- NVIDIA Stock Data: Yahoo Finance API.