Stock Price Prediction

Report 01

Stock Price Prediction: Comprehensive EDA Report

intelliHack 5.0 - IEEE Computer Society UCSC

Team Name : Code Voyagers

Date : 10/03/2025

Task Number : 04

Introduction

This report presents an exploratory data analysis (EDA) of the provided historical stock price dataset. The goal of this analysis is to identify relevant patterns and features that can help predict the stock's closing price 5 trading days into the future. The analysis focuses on understanding price trends, seasonality, anomalies, and relationships between different variables to inform feature engineering and model selection decisions.

Dataset Overview

The dataset spans from March 17, 1980, to December 27, 2024, comprising 11,291 trading days. It contains the following columns:

- Date
- Adjusted Close
- Close
- High
- Low
- Open
- Volume

Key statistics:

- Price Range: $3.24 to $254.77
- Average Closing Price: $72.03
- Missing Values:
    - Open prices: 27.02%
    - Volume data: 2.44%
    - Close prices: 1.04%

Data Preprocessing Decisions

Handling Missing Values

1. **Open Prices**: 27.02% of open prices were missing (value of 0). These were replaced with the previous day's closing price, which is a reasonable approximation since stocks typically open near their previous closing price.
2. **Volume Data**: 2.44% of volume entries were missing or zero. These were replaced with the median non-zero volume to maintain the time series continuity without skewing the distribution.
3. **Close Prices**: While only 1.04% of closing prices were missing, these were critical for our analysis. Rather than imputing values, we excluded these dates from return calculations to avoid introducing bias.

Date-Based Feature Creation

To capture temporal patterns, we created the following date-based features:

- Day of week (0-4 for Monday-Friday)
- Month (1-12)
- Year

Time Series Features

We calculated various time-based metrics to capture price dynamics:

- Daily returns (percentage change from previous day)
- Rolling 20-day volatility (standard deviation of returns)
- Moving averages (5, 20, 50, and 200 days)
- Price momentum over different timeframes (1, 5, 10, 20 days)

Visualizations of Key Patterns and Relationships

1. Long-Term Price Trends

The stock shows significant growth over the 44-year period, with several distinct market cycles. Most notably, there was substantial growth in recent years, with the price increasing from under $50 in the early 2000s to over $200 by the end of 2024.

2. Daily Returns Distribution

The daily returns follow a leptokurtic distribution (higher peak and fatter tails than normal), indicating more extreme price movements than would be expected under a normal distribution. Key statistics:

- Average daily return: 0.0545%
- Standard deviation: 1.8259%
- Maximum daily gain: 19.35%
- Maximum daily loss: -16.52%

The normality test confirms that returns are not normally distributed, which has implications for our prediction modeling approach.

3. Volatility Patterns

The 20-day rolling volatility analysis reveals:

- Periods of high volatility clustering together
- Recent volatility (1.9002%) is slightly higher than historical volatility (1.8259%)

- Volatility ratio (recent/historical) of 1.0407 suggests current market conditions are marginally more volatile than the long-term average

4. Seasonality Effects

Monthly returns heatmap shows:

- Certain months consistently outperform others
- January and April tend to have positive returns across years
- September and October often show negative returns

Day-of-week analysis reveals:

- Monday and Friday show more extreme return patterns
- Mid-week trading days (Tuesday-Thursday) are relatively more stable

5. Volume-Price Relationship

Volume analysis shows:

- Correlation between volume and absolute daily returns of approximately 0.3, indicating higher volume on days with larger price movements
- Average volume of 216,966 shares
- Monday typically shows the highest average volume, while Friday shows the lowest

6. Technical Indicators

The RSI, MACD, and Bollinger Bands analysis reveals:

- Several overbought and oversold conditions that preceded price reversals
- Bollinger Band squeezes followed by significant price movements
- MACD crossovers often signaling trend changes

7. Price Gap Analysis

There were numerous significant price gaps (>2% difference between previous close and current open):

- Most gaps were associated with earnings announcements or market-wide events
- Positive gaps had a higher tendency to continue in the same direction
- Negative gaps showed more mean reversion

Analysis of Trends, Seasonality, and Anomalies

Long-Term Trends

The data shows multiple bull and bear market cycles. Recent yearly performance:

- 2020: +8.67%
- 2021: +20.88%
- 2022: -26.68%
- 2023: +32.61%
- 2024: +10.01%

The moving average analysis shows strong uptrends when the 50-day MA crosses above the 200-day MA, and significant downtrends when it crosses below.

Seasonality Patterns

Time series decomposition confirmed the presence of:

1. **Annual seasonality**: Certain months consistently outperform others
2. **Weekly seasonality**: Return patterns differ by day of the week
3. **Quarterly effects**: Possible influence of earnings announcements

Anomalies Detection

The analysis identified:

- 73 return outliers (beyond 3 standard deviations)
- 286 significant price gaps exceeding 2%
- 124 days of abnormally high volatility (2x historical average)

These anomalies were often associated with:

- Earnings announcements
- Market-wide events
- Industry-specific news

Feature Selection Justification

Based on the analysis, the following features were identified as most predictive for the 5-day future return prediction:

1. **Momentum Features**:
    - 5-day and 20-day momentum (correlation: 0.21 and 0.18 with future returns)
    - Rate of change over 5, 10, and 20 days
    - These capture short-term price trends that tend to continue
2. **Technical Indicators**:
    - RSI (identifies overbought/oversold conditions)
    - Moving average crossovers (MA5/MA20 and MA20/MA50)

    - Bollinger Band positions (relative to upper and lower bands)
    - These indicators have proven value in forecasting price movements
3. **Volatility Measures**:
    - 20-day volatility
    - ATR (Average True Range)
    - These help identify potential for larger price moves
4. **Volume Indicators**:
    - Volume pressure (ratio to 20-day average)
    - These capture unusual trading activity that often precedes price movements
5. **Gap Analysis**:
    - Price gaps as percentage of previous close
    - These often signal continued momentum

Using Random Forest feature importance analysis, the top 5 predictive features were:

1. 20-day momentum (importance: 0.182)
2. RSI (importance: 0.167)
3. 5-day momentum (importance: 0.154)
4. Volatility (importance: 0.132)
5. 5-day rate of change (importance: 0.121)

Risk Analysis

The stock exhibits:

- Maximum historical drawdown of 72.46%
- Sharpe Ratio of 0.76 (assuming 2% risk-free rate)
- Sortino Ratio of 1.12
- Beta fluctuations between 0.7 and 1.4 (against synthetic market data)

These risk metrics suggest moderate volatility with positive risk-adjusted returns over the long term.

Autocorrelation Analysis

The autocorrelation analysis reveals:

- Statistically significant autocorrelation at lags 1 and 2
- Durbin-Watson statistic of 1.92 (close to 2, suggesting limited serial correlation)
- Some predictability in short-term returns, supporting the use of recent price data for predictions

Conclusion and Recommendations

This exploratory analysis reveals several important patterns and relationships in the stock price data that can be leveraged for prediction:

1. The stock shows clear trends, with momentum effects persisting over 5-20 day periods
2. Technical indicators (RSI, MACD, Bollinger Bands) provide valuable signals
3. Volume-price relationships offer additional predictive power
4. Seasonal patterns exist both within weeks and across months of the year
5. Volatility clustering suggests the need for regime-aware modeling

For the prediction task, I recommend:

1. Using features from multiple timeframes (5, 10, 20 days)
2. Including both price-based and volume-based indicators
3. Incorporating technical analysis signals
4. Accounting for seasonal patterns
5. Developing ensemble models that can handle different market regimes

The preprocessing steps outlined in this report provide a solid foundation for feature engineering, while the identified patterns justify the selection of specific features for the prediction model.