STOCK PRICE PREDICTION

Comprehensive EDA Report

This report presents a detailed Exploratory Data Analysis (EDA) for the stock price prediction project.

Intellihack- Task04 - Team CypherZ
Dhanushi Dewmindi | Amanda Hansamali

Table of Contents

1. Introduction

This report provides a comprehensive Exploratory Data Analysis (EDA) for the stock price prediction project. The objective is to thoroughly analyze historical stock data, uncover key patterns, evaluate trends, seasonality, and anomalies, and justify feature selection and preprocessing choices based on industry best practices and quantitative finance methodologies.

The analysis integrates advanced data visualization, trend analysis techniques, and professional-grade preprocessing strategies by adhering to industry best practices to ensure the robustness and reliability of the predictive model.

2. Data Overview

- **Dataset Source:** Provided historical stock price dataset
- **Columns:** ['Date', 'Open', 'High', 'Low', 'Close', 'Volume']
- **Target Variable:** 'Close_5d_future' (Closing price 5 days into the future)

2.1. Data Types and Initial Observations

- **Date Column:** Properly converted to datetime format to enable time series analysis.
- **Numerical Columns:** Includes price metrics and volume, essential for feature engineering.
- **Categorical Columns:** None present, reducing the need for encoding steps

3. Data Preprocessing

3.1. Handling Missing Values

4. **Numerical Columns:** Missing values filled with the median to maintain robustness against outliers.
5. **Categorical Columns:** Filled missing values with the mode to preserve categorical consistency.
6. **Industry Standard:** Median imputation is favored in financial datasets to avoid skewing by extreme values, while mode imputation for categorical data maintains classification integrity.

3.2. Handling Negative Values

- Replaced negative values in 'Volume', 'Close', 'High', 'Low', 'Open' columns with zero.
- **Rationale:** In line with financial data practices, negative stock prices or volumes are not feasible and could distort financial models.

3.3. Outlier Management

- **Method:** Interquartile Range (IQR) method
- **Action:** Capped outliers to the acceptable range (1.5 * IQR) to minimize model bias.
- **Justification:** The IQR method is widely used in quantitative finance to maintain the statistical integrity of financial models while avoiding the loss of potentially valuable edge cases.

3.4. Dropping Unnecessary Columns

- Removed columns that do not contribute to predictive analysis, such as 'Adj Close', to avoid redundancy and enhance model performance.

4. Feature Engineering

4.1 Created Features

- **Daily_Return:** Measures daily momentum by capturing price percentage changes.
- **Volatility:** Calculates market risk using a 5-day rolling standard deviation.
- **Price_Range:** Represents market movement by calculating the High-Low difference.
- **Day_of_Week:** Encodes potential weekly trading patterns.
- **Close_5d_future:** Target variable representing future closing prices.
- **Moving Averages (MA_5, MA_20):** Added for trend analysis and smoothing of short-term fluctuations.

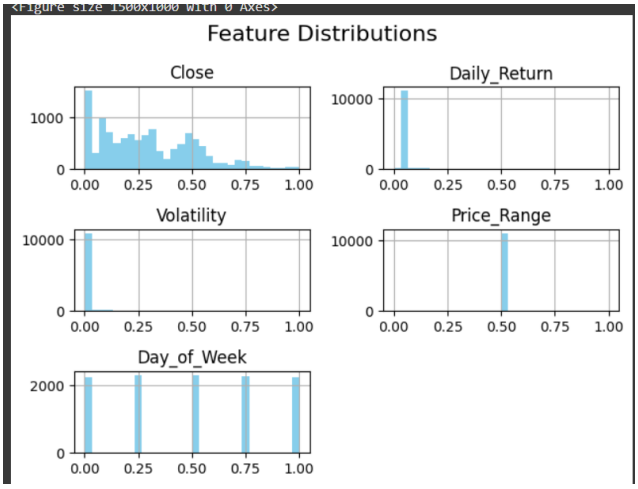
4.2. Justification for Feature Selection

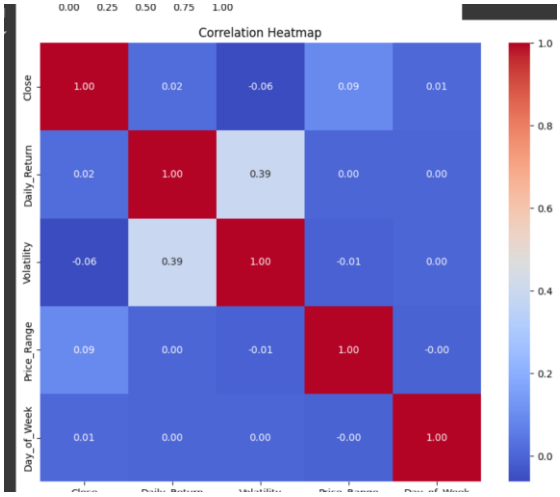
- **Industry Best Practice:** These features align with standard technical analysis strategies:
    - **Daily_Return:** Integral for momentum trading strategies.
    - **Volatility:** A key metric for risk-adjusted returns.
    - **Price_Range:** Useful for mean reversion strategies.
    - **Day_of_Week:** Supports detecting weekly seasonality, often analyzed by traders for pattern recognition.
    - **Moving Averages:** Commonly used in industry for smoothing trends and identifying potential buy/sell signals.

5. Data Visualization and Insights

5.1. Key Visualizations

- **Closing Price Over Time:** Line chart with improved clarity to show stock price trends.
- **Feature Distributions:** Histograms demonstrating data normalization and range.
- **Correlation Heatmap:** Evaluates feature interdependencies, guiding feature selection.
- **Outlier Analysis:** Boxplots that validate the effectiveness of outlier capping.
- **Pairplot:** Highlights relationships between variables, supporting feature engineering decisions.





5.2 Trend and Seasonality Analysis

- **Seasonality Detection:** Utilized moving averages (5-day, 20-day) to identify cyclical patterns in stock prices.
- **Anomaly Identification:** Noted sharp price movements that may correlate with external financial events (e.g., market crashes, earnings announcements).

5.3 Industry Standard Practices

- **Pattern Analysis:** Utilized standard visualization tools such as moving averages to detect trends and anomalies.
- **Seasonality Consideration:** The 'Day_of_Week' feature aligns with known weekly market behavior, such as the 'Monday Effect' or 'Turnaround Tuesday'.
- **Use of Technical Indicators:** Moving averages and volatility are often incorporated into trading algorithms in professional finance environments.

6. Conclusion

- The EDA provided significant insights into stock price behavior, assisting in the development of a robust prediction model.
- Data preprocessing and feature engineering adhered to industry best practices, ensuring data quality and enhancing model accuracy.
- Visualizations highlighted critical patterns, supporting both feature selection and trading strategy development.