STOCK ANALYSIS AND S&P 500 TREND PREDICTION MIDTERM - REPORT

Ahemed Bullo
Abidul Islam
Abdullahi Nur

PURPOSE

The main goal of this project is to analyze the relationship between individual stocks and the S&P 500 index to identify key influencers on market trends.

By examining historical stock data, we aim to identify patterns and correlations that reveal how specific stocks affect the S&P 500's overall performance.

OBJECTIVES

Data Collection and Processing:

- Gather historical stock data for major companies listed in the S&P 500, including daily price changes and percentage changes.
- Clean and preprocess the data to ensure accuracy and consistency, handling missing values and outliers.

Correlation Analysis:

- Calculate correlations between daily percentage changes in individual stocks and the S&P 500 index.
- Identify which stocks serve as leading indicators for market movements, helping investors make informed decisions.

Predictive Modeling:

- Develop statistical models, such as linear regression, to predict S&P 500 movements based on the performance of correlated stocks.
- Evaluate the effectiveness of these models in forecasting market trends.

Data Visualization:

- Create visual representations of stock performance and correlations to facilitate understanding and communication of findings.
- Use plots and heatmaps to illustrate trends and relationships within the data.

Modeling:

- Implement linear regression to quantify relationships between selected stocks and the S&P 500.
- Analyze feature importance to identify significant stocks influencing the S&P 500.
- Refine models based on evaluation metrics and incorporate additional data as needed.

TECH STACK AND TOOLS

Programming Language:

Python

Libraries and Frameworks:

Data Manipulation

Pandas

Numerical Operations

NumPy

Data Visualization

Matplotlib

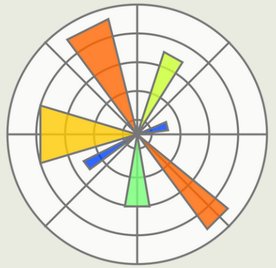
Seaborn

Machine Learning

Scikit-learn

Web Scraping and Data Retrieval

yfinance



DATA LOADING AND CLEANING

Data Loading

Historical stock data is loaded from a Yahoo Finance API as CSV files located in the data/raw/ directory.

```
# Load data for Microsoft
msft_data = pd.read_csv('data/raw/MSFT.csv')
# Load data for NVIDIA
nvda_data = pd.read_csv('data/raw/NVDA.csv')
```

↓

Handling Missing Values

Identified and addressed missing values to maintain data integrity.
Used dropna() to remove rows with missing values in critical columns.

```
# Remove rows with missing values
msft_data.dropna(inplace=True)
nvda_data.dropna(inplace=True)
```

→

Data Type Conversion

Converted date strings to datetime objects for accurate time series analysis

```
# Convert 'Date' column to datetime
msft_data['Date'] = pd.to_datetime(msft_data['Date'])
nvda_data['Date'] = pd.to_datetime(nvda_data['Date'])
```

FEATURE ENGINEERING AND MODEL PREPARATION

Feature Engineering

Calculated additional features such as:

- **Daily Return:** Percentage change in closing prices, which is crucial for understanding stock performance.

```
# Calculate Daily Return
msft_data['Daily Return'] = msft_data['Close'].pct_change()
nvda_data['Daily Return'] = nvda_data['Close'].pct_change()
```

Preparing Data for Modeling

Daily Return Matrix: Constructed a DataFrame containing daily returns for all stocks, which serves as the basis for model training

Top Correlated Stocks: Identified the top stocks correlated with the S&P 500 to use as features in the model.

```
top_stocks = get_top_correlated_stocks(return_df,
target_col="SP500", top_n=6)
```

Model Training

Linear Regression Model: Built a linear regression model to predict S&P 500 movements based on the selected features

Model Evaluation: Evaluated the model using metrics such as Mean Squared Error and R^2 Score to assess performance.

```
top_stocks = get_top_correlated_stocks(return_df,
target_col="SP500", top_n=6)
```

Data Saving

Saved the cleaned and processed data to the data/processed/ directory in CSV format for further analysis.

```
# Save cleaned data
msft_data.to_csv('data/processed/MSFT_cleaned.csv',
index=True)
nvda_data.to_csv('data/processed/NVDA_cleaned.csv',
index=True)
```

PRELIMINARY RESULTS

Findings
Correlation Analysis:

After running the model and computing Pearson correlation coefficients, the following stocks were found to have the strongest positive relationships with the S&P 500 index:

Correlation Values:

^GSPC: 1.000

MSFT: 0.828

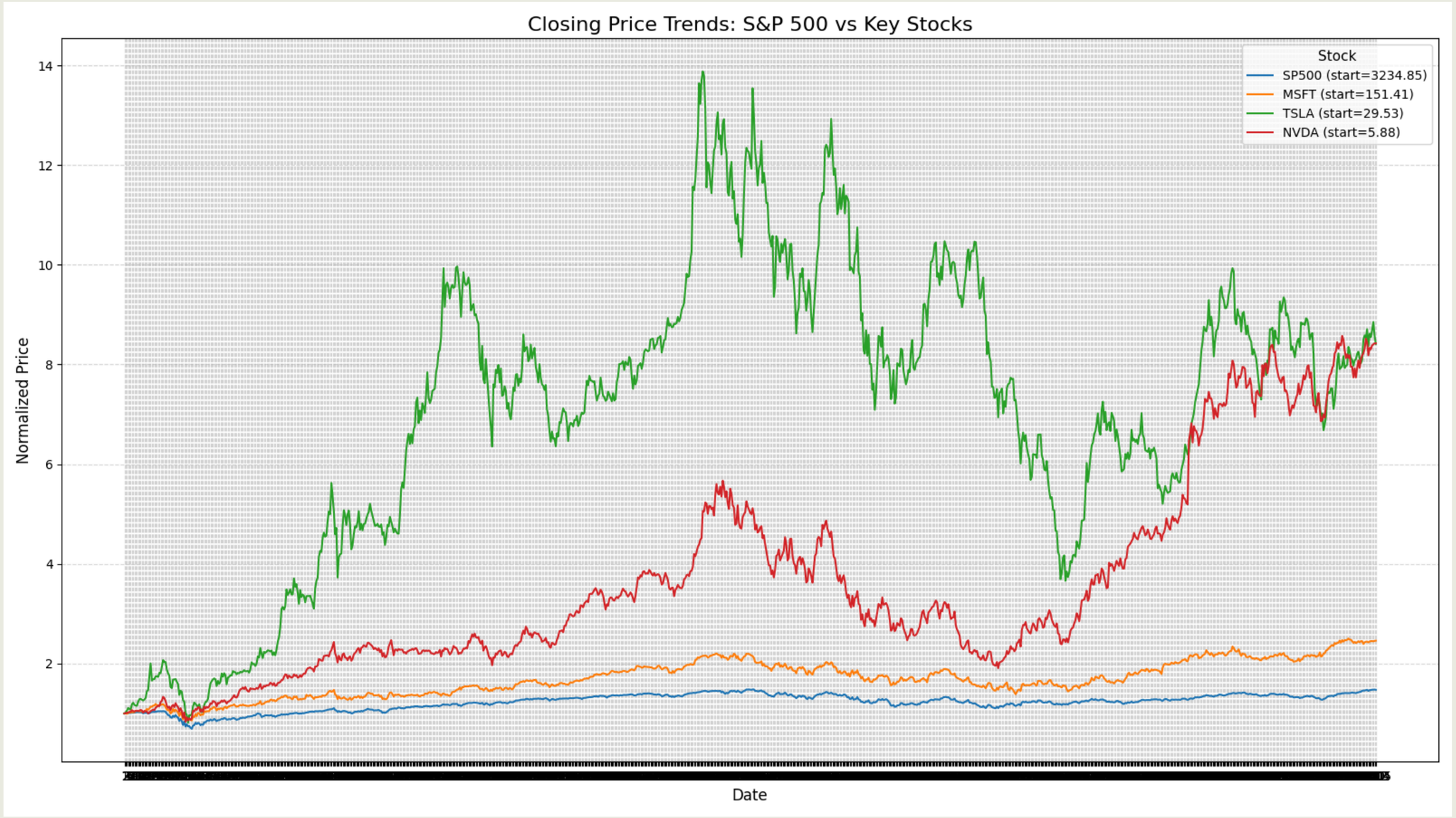
NVDA: 0.717

TSLA: *0.513*

Model Evaluation:

Mean Squared Error (MSE): 0.000052

R^2 Score (Coefficient of Determination): 0.0871


Closing Price Trends: S&P 500 vs Key Stocks
Normalized Price
Date
Stock
SP500 (start=3234.85)
MSFT (start=151.41)
TSLA (start=29.53)
NVDA (start=5.88)
14
12
10
8
6
4
2

NEXT STEPS

Additional Data Collection	We''ll gather more historical stock data for additional companies to enhance our analysis
Further Analysis and Modeling	We'll perform a more thorough correlation analysis to find relationships between the market indexes and stocks. More advance modeling techniques such as: • Time series forecasting • Machine learning models
Improvements to Visualizations	Create interactive visualizations with libraries such as Plotly or Dash that let people to examine data in real time. Enhance the existing plots with extra features using: • Annotations for major events, such as earnings report, market crashes • Comparative visualizations to improve understanding
Reporting and Documentation	Update the project documentation with any updated results and methodology Create a detailed report detailing the study, findings, and insights got.

Thank you.