Project Report

Analyzing and Predicting VN-Index Fluctuations Using Data Science Techniques

Team Members:

Nguyễn Duy Đạt, Chu Thiên Hải, Vũ Khắc Long, Nguyễn Minh Đức, Phùng Thanh Đăng, Nguyễn Việt Anh

Abstract:

This study explores the use of data science methodologies to analyze and predict fluctuations in VN-Index, a benchmark index for Vietnam's stock market. By integrating macroeconomic indicators, international market data, and investor sentiment, the research develops a comprehensive dataset spanning over two decades (2002-2024). The data is processed using advanced techniques, including interpolation, normalization, and dimensionality reduction, to prepare it for predictive modeling. The analysis leverages time-series modeling approaches, focusing on feature engineering to derive meaningful insights. The findings aim to support investors and policymakers by providing a deeper understanding of the factors driving VN-Index movements, enabling informed decision-making. This report highlights challenges in data collection and integration, while also emphasizing the importance of robust preprocessing techniques in financial data analysis.

Keywords:

VN-Index, Stock Market Analysis, Data Science, Time-Series Modeling, Macroeconomic Indicators, Investor Sentiment, Data Integration, Predictive Analytics.

1. Introduction to VN-Index

VN-Index is a benchmark stock market index that tracks the performance of all companies listed on the Ho Chi Minh City Stock Exchange (HOSE). It serves as an indicator of the Vietnamese stock market's overall health and is widely used by investors to gauge market trends and sentiments. VN-Index provides a crucial insight into how different sectors of the economy perform over time, and it reflects the investment environment's dynamic nature.

The index is calculated using the market capitalization-weighted method, ensuring that larger companies have a greater influence on their movements. This method allows for a more accurate representation of market performance, as it gives weight to companies based on their size and market value. Key factors affecting VN-Index include macroeconomic indicators, foreign investment, sectoral performances, and investor sentiment. As a critical tool for decision-making, VN-Index reflects the dynamic interactions between Vietnam's economic conditions and its financial markets, making it indispensable for market analysts and policymakers alike.

Additional Information:

1. **Historical Background**: The VN-Index was first introduced in 2000, marking the establishment of Vietnam's modern stock market. Its inception coincided with Vietnam's efforts to attract foreign investment and integrate into the global financial system.
2. **Current Scope**: As of 2024, VN-Index includes companies across diverse sectors such as banking, real estate, and manufacturing, making it a comprehensive representation of Vietnam's economic landscape. [1]
3. **Relevance in Global Context**: VN-Index has gained international recognition as a measure of emerging market potential, attracting attention from global investors.

Formula of calculating VN-Index:

VN-Index is calculated based on the market capitalization value of stocks listed on the Ho Chi Minh City Stock Exchange (HOSE), with the following specific formula:

VN-Index = (Current market capitalization / Base market capitalization) x 100

In there:

- The total market capitalization of listed stocks is calculated by multiplying the stock price by the number of outstanding shares of each company listed on the HOSE.
- Base total market capitalization is the total market capitalization on the base date (July 28, 2000).
- The base index value is usually set at 100 points or some other pre-selected specific number.

The detailed formula is as follows: **VN-Index =** $$\frac{100 * \sum_{i=1}^{N} P_{1i} Q_{1i}}{\sum_{i=1}^{N} P_{0i} Q_{0i}}$$

In which:

P_{1i}: Current price of i

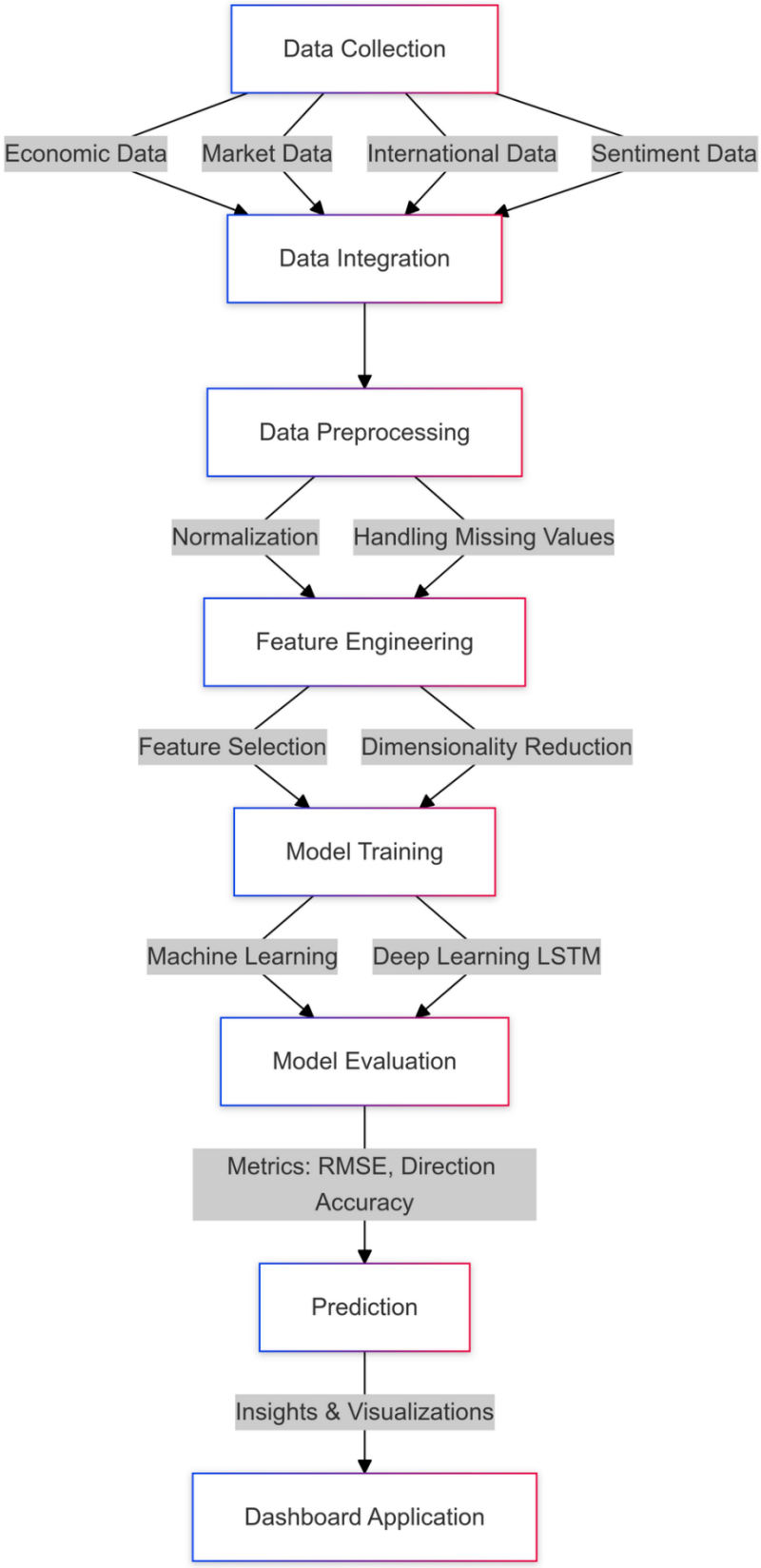
Q_{1i}: Circulating volume (listed volume) of i

P_{0i}: Price of stock i in base period

Q_{0i}: Volume of stock i at base period

This formula allows calculating the change of VN-Index based on the fluctuations of stock prices and the number of outstanding shares. The VN-Index will increase when the total market capitalization increases, which usually happens when the stock prices of one or more listed companies increase. Conversely, this index will decrease when the stock prices decrease. Changes in the index **VN-Index** are closely monitored throughout the trading session to most accurately reflect the market situation at all times. [2]

Pipeline Overview:



2. Data Collection

2.1 Potential Influencing Factors:

1. **Macroeconomic Factors:**
    a. GDP growth
    b. Interest rates (bank interest rates, lending interest rates)
    c. Inflation (CPI)
    d. Exchange rates (USD/VND)
2. **Stock Market Factors:**
    a. Trading volume
    b. Transaction values of foreign investors
    c. Dividends of major companies in the VNINDEX
3. **International Factors:**
    a. Oil prices, gold prices
    b. International stock market indices (S&P 500, Nikkei, Dow Jones)
4. **Investor Sentiment Factors:**
    a. News regarding the stock market and major companies
    b. Investor attitudes reflected through forums and social media platforms

2.2 Data Sources

To understand and predict VN-Index fluctuations, data was collected from multiple sources, ensuring comprehensive coverage of economic, market, and international factors. This multi-source approach allows for a holistic understanding of the factors influencing VN-Index, enabling more robust predictions.

1. **VN-Index Historical Data:**
    a. Source: Vietstock
    b. Period: January 2002 to November 2024
    c. Features: Daily open, high, low, close (OHLC) prices, trading volumes, and percentage changes. This data serves as the foundation for time-series analysis.

	Ngày	Lần cuối	Mở	Cao	Thấp	KL	% Thay đổi
0	17/10/2024	1286.52	1,282.65	1,286.52	1,271.82	357.35K	0.55%
1	16/10/2024	1279.48	1,279.86	1,283.33	1,276.4	357.35K	-0.12%
2	15/10/2024	1281.08	1,287.81	1,294.05	1,279.81	357.35K	-0.41%
3	14/10/2024	1,286.34	1,288.39	1,297.67	1,286.13	795.71M	-0.16%
4	11/10/2024	1,288.39	1,286.36	1,289.37	1,283.56	531.74M	0.16%

2. **Macroeconomic Indicators:**
    a. Source: General Statistics Office of Vietnam
    b. Data Includes:
        i. GDP growth rates (2015 onward)
        ii. Inflation rates (CPI)
        iii. Bank interest rates
        iv. Foreign exchange rates (USD/VND)
    c. These indicators are pivotal in understanding the broader economic conditions affecting stock market performance.

Chỉ số giá tiêu dùng các tháng trong năm chia theo Các tháng (tháng trước = 100) và Năm									
	2015	2016	2017	2018	2019	2020	2021	2022	2023
Tháng 1	99,8	100	100,46	100,51	100,1	101,23	100,06	100,19	100,52
Tháng 2	99,95	100,42	100,23	100,73	100,8	99,83	101,52	101	100,45
Tháng 3	100,15	100,57	100,21	99,73	99,79	99,28	99,73	100,7	99,77
Tháng 4	100,14	100,33	100	100,08	100,31	98,46	99,96	100,18	99,66
Tháng 5	100,16	100,54	99,47	100,55	100,49	99,97	100,16	100,38	100,01
Tháng 6	100,35	100,46	99,83	100,61	99,91	100,66	100,19	100,69	100,27
Tháng 7	100,13	100,13	100,11	99,91	100,18	100,4	100,62	100,4	100,45
Tháng 8	99,93	100,1	100,92	100,45	100,28	100,07	100,25	100	100,88
Tháng 9	99,79	100,54	100,59	100,59	100,32	100,12	99,38	100,4	101,08
Tháng 10	100,11	100,83	100,41	100,33	100,59	100,09	99,8	100,15	100,08
Tháng 11	100,07	100,48	100,13	99,71	100,96	99,99	100,32	100,39	100,25
Tháng 12	100,02	100,23	100,21	99,75	101,4	100,1	99,82	99,99	100,12

A	B	C	D	E	F
Ngày	Lần cuối	Mở	Cao	Thấp	% Thay đổi
07/11/2024	25,380.0	25,415.0	25,415.0	25,380.0	0.04%
06/11/2024	25,370.0	25,365.0	25,410.0	25,350.0	0.16%
05/11/2024	25,330.0	25,310.0	25,360.0	25,285.0	0.08%
04/11/2024	25,310.0	25,260.0	25,325.0	25,250.0	0.12%
01/11/2024	25,280.0	25,265.0	25,330.0	25,250.0	0.04%
31/10/2024	25,270.0	25,325.0	25,345.0	25,260.0	-0.04%
30/10/2024	25,280.0	25,335.0	25,400.0	25,280.0	-0.06%
29/10/2024	25,295.0	25,345.0	25,345.0	25,270.0	-0.20%
28/10/2024	25,345.0	25,390.0	25,430.0	25,340.0	-0.04%
25/10/2024	25,355.0	25,400.0	25,425.0	25,355.0	-0.12%
24/10/2024	25,385.0	25,391.0	25,440.0	25,380.0	-0.06%
23/10/2024	25,400.0	25,400.0	25,420.0	25,370.0	0.04%
22/10/2024	25,390.0	25,375.0	25,390.0	25,345.0	0.42%
21/10/2024	25,285.0	25,150.0	25,285.0	25,150.0	0.54%
18/10/2024	25,150.0	25,210.0	25,260.0	25,150.0	-0.12%
17/10/2024	25,180.0	25,010.0	25,195.0	25,010.0	0.84%

3. **International Market Factors:**
    a. Source: Yahoo Finance API
    b. Data Includes:
        i. Crude oil and gold prices (2000 onward)
        ii. Global stock indices (e.g., S&P 500, Nikkei, Dow Jones)
    c. These factors provide a global perspective on market trends, linking VN-Index to international economic shifts.

Price	Adj Close	Close	High	Low	Open	Volume
Ticker	GC=F	GC=F	GC=F	GC=F	GC=F	GC=F
Date						
2000-08-3	273.9	273.9	273.9	273.9	273.9	0
2000-08-3	278.3	278.3	278.3	274.8	274.8	0
2000-09-0	277	277	277	277	277	0
2000-09-0	275.8	275.8	275.8	275.8	275.8	2
2000-09-0	274.2	274.2	274.2	274.2	274.2	0
2000-09-0	274	274	274	274	274	125
2000-09-0	273.3	273.3	273.3	273.3	273.3	0
2000-09-1	273.1	273.1	273.1	273.1	273.1	0
2000-09-1	272.9	272.9	272.9	272.9	272.9	0
2000-09-1	272.8	272.8	272.8	272.8	272.8	0
2000-09-1	272.4	272.4	272.4	272.4	272.4	0
2000-09-1	272.3	272.3	272.3	272.3	272.3	0
2000-09-1	271.4	271.4	271.4	271.4	271.4	0
2000-09-1	271.9	271.9	271.9	271.9	271.9	0
2000-09-2	269	269	269	269	269	0
2000-09-2	270.3	270.3	270.3	270.3	270.3	0

Price	Adj Close	Close	High	Low	Open	Volume
Ticker	CL=F	CL=F	CL=F	CL=F	CL=F	CL=F
Date						
2000-08-2	32.05	32.05	32.8	31.95	31.95	79385
2000-08-2	31.63	31.63	32.24	31.4	31.9	72978
2000-08-2	32.05	32.05	32.1	31.32	31.7	44601
2000-08-2	32.87	32.87	32.92	31.86	32.04	46770
2000-08-2	32.72	32.72	33.03	32.56	32.82	49131
2000-08-3	33.4	33.4	33.4	32.1	32.75	79214
2000-08-3	33.1	33.1	33.7	32.97	33.25	56895
2000-09-0	33.38	33.38	33.45	32.75	33.05	45869
2000-09-0	33.8	33.8	33.99	33.42	33.95	55722
2000-09-0	34.95	34.95	34.95	33.83	33.99	74692
2000-09-0	35.33	35.33	35.5	34.45	34.5	74105
2000-09-0	33.7	33.7	34.78	33.4	34.55	88415
2000-09-1	35.1	35.1	35.85	33.75	33.8	101518
2000-09-1	34.2	34.2	35.5	34.1	35.45	91911
2000-09-1	33.8	33.8	34.74	33.5	34	94630
2000-09-1	34.1	34.1	34.5	33.12	33.78	98068

4. **Investor Sentiment:**
    a. Source: AAII Sentiment Survey
    b. Period: 1987 onward
    c. Data Captures: Weekly sentiment trends among individual investors in the U.S., often correlated with global market movements. Understanding sentiment aids in predicting market behavior during uncertain times.

Reported Date	Bullish	Neutral	Bearish	Total	Bullish 8-week Mov Avg	Bull-Bear Spread	Bullish Average	Bullish Average +St. Dev.	Bullish Average - St. Dev.
3-21-24	43.20%	29.61%	27.19%	100%	46.48%	16.0%	37.7%	47.8%	27.6%
3-28-24	50.00%	27.55%	22.45%	100%	46.60%	27.6%	37.7%	47.8%	27.6%
4-4-24	47.29%	30.49%	22.22%	100%	46.38%	25.1%	37.7%	47.8%	27.6%
4-11-24	43.44%	32.51%	24.04%	100%	46.54%	19.4%	37.7%	47.8%	27.6%
4-18-24	38.27%	27.76%	33.96%	100%	45.79%	4.3%	37.7%	47.8%	27.6%
4-25-24	32.13%	33.93%	33.93%	100%	44.00%	-1.8%	37.7%	47.8%	27.6%
5-2-24	38.49%	29.02%	32.49%	100%	42.34%	6.0%	37.7%	47.8%	27.6%
5-9-24	40.82%	35.37%	23.81%	100%	41.71%	17.0%	37.7%	47.8%	27.6%
5-16-24	40.86%	35.88%	23.26%	100%	41.41%	17.6%	37.7%	47.8%	27.6%
5-23-24	47.04%	26.64%	26.32%	100%	41.04%	20.7%	37.7%	47.8%	27.6%
5-30-24	39.04%	34.25%	26.71%	100%	40.01%	12.3%	37.7%	47.8%	27.6%
6-6-24	38.97%	29.04%	31.99%	100%	39.45%	7.0%	37.7%	47.8%	27.6%
6-13-24	44.59%	29.73%	25.68%	100%	40.24%	18.9%	37.7%	47.8%	27.6%
6-20-24	44.37%	33.12%	22.51%	100%	41.77%	21.9%	37.7%	47.8%	27.6%

2.3 Challenges in Data Collection

- **Data Completeness**: Older datasets required significant cleaning due to missing values. For example, historical GDP data had gaps that needed to be filled using estimation techniques.
- **Diverse Formats**: Data from different sources were in varied formats (e.g., daily vs. quarterly). Standardizing these formats was critical for effective integration.
- **Timeliness**: Some data sources had delays in updates, especially during volatile economic periods. Ensuring up-to-date information required additional data scraping techniques.
- **Data Volume**: Managing and processing a large volume of data from 2002 to 2024 demanded robust storage and processing systems.

Additional Information:

1. **Correlation Between Indicators**: Macroeconomic and international factors often exhibit strong correlations with VN-Index trends, underscoring the need for integrated analysis.
2. **Technological Tools**: Advanced tools like Python libraries (Pandas, NumPy) and APIs were crucial for handling and analyzing diverse datasets. [7]

3. Data Processing and Integration

3.1 Preprocessing Steps

Data preprocessing is a critical phase in any data analysis project. This stage involves a series of systematic steps to ensure that the data is standardized, cleaned, and integrated into a form suitable for subsequent analysis and modeling. The objective is to make the dataset consistent, reliable, and free from irregularities that could compromise the accuracy of the models. The steps involved include handling missing data, removing outliers, normalizing data formats, and merging datasets from multiple sources.

1. **Handling Missing Data**:
   Addressing missing values is a common challenge in data preprocessing. Missing data, if not handled correctly, can lead to biased or inaccurate results. The approach to handling missing data must be informed by the nature of the dataset and the analysis objectives.
    a. Methods: Interpolation: Linear interpolation techniques were employed to estimate missing values based on surrounding data points. This method is particularly effective for filling gaps in time-series data, where values are assumed to follow a logical progression over time.

Backfill Strategies: In cases where interpolation was not feasible, backfill strategies were used. This involved filling missing values with the nearest subsequent available data point, ensuring no interruptions in the sequence of data.

b. Example: In a specific instance, historical GDP data contained several missing quarterly values, which posed a challenge to accurately analyze macroeconomic trends. To address this:
    - Missing GDP values were interpolated based on available growth rates from surrounding quarters, ensuring a smooth and realistic progression in the dataset.
    - In situations where interpolation was not viable, backfill strategies were applied using subsequent data points to maintain continuity.

    This approach ensured that the dataset faithfully represented macroeconomic trends, providing a robust foundation for predictive modeling and trend analysis.

By implementing these preprocessing techniques, the data was transformed into a high-quality format, laying the groundwork for reliable and meaningful analytical outcomes..

2. **Normalization**:

Normalization ensures that data from diverse sources and scales can be compared and analyzed consistently. This is particularly important when integrating datasets with different measurement units and frequencies.

a. Macroeconomic data

Macroeconomic indicators such as the Consumer Price Index (CPI) and interest rates were aggregated to a uniform monthly or quarterly frequency. This step eliminated inconsistencies due to differing reporting periods across sources, allowing for seamless comparison and integration.

b. Stock Price Data:

Stock price data, which often contains significant noise due to daily fluctuations, was normalized to provide a clearer picture of performance trends. The normalization process used the following formula:

Stock price data was normalized using the formula:

$$\text{Price} = \frac{\text{Open} + \text{High} + \text{Low} + \text{Close}}{4}$$

This formula transforms stock price data into a relative scale, simplifying the comparison of performance across different time periods or companies.

c. The normalization process reduced the noise from daily price movements, enabling analysts to focus on longer-term trends and patterns. This approach enhanced the reliability of analyses involving both macroeconomic and financial datasets, ensuring consistency and clarity in the results.

3.2 Data Integration

After preprocessing, the next step was integrating all datasets into a unified format that facilitated consistent and efficient analysis. This process involved aligning data across different sources and transforming it into a structure suitable for time-series analysis. Key components of the integration phase included feature engineering, dimensionality reduction, and the preparation of a final dataset.

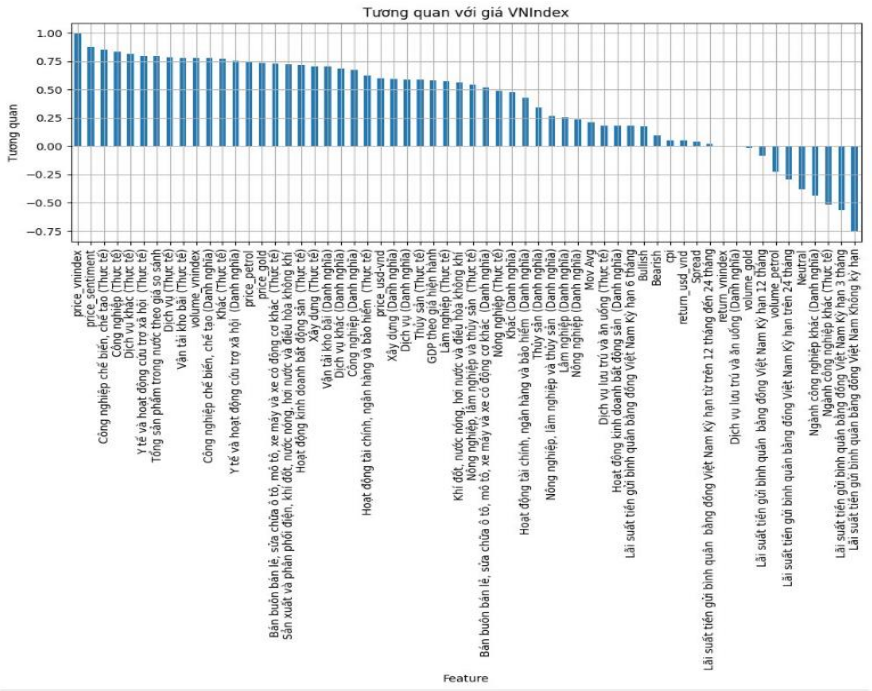
a. **Unified Format for Time-Series Analysis:**

   All datasets were merged based on daily timestamps, ensuring that every observation across different sources corresponded to the same time point. This alignment allowed for seamless integration of macroeconomic indicators, stock market data, and other relevant variables into a cohesive time-series framework.

b. **Feature Engineering**:
   To enrich the analysis, additional features were derived from the raw data:
   - Daily Returns: Calculated as the percentage change in stock prices from one day to the next, providing insight into short-term performance trends.
   - Moving Averages: Computed over various time windows (e.g., 5-day, 10-day) to smooth out daily price fluctuations and highlight longer-term trends.
   - Volatility Measures: Captured the variability of stock prices over time, offering valuable information about market risk and sentiment.
     These engineered features added depth to the dataset, enabling more nuanced insights and improving the predictive power of subsequent models.

   Derived additional features such as daily returns, moving averages, and volatility measures. These features added depth to the analysis, allowing for more nuanced insights.

c. **Dimensionality Reduction**:
   With a large number of features initially included in the dataset, correlation analysis was employed to identify and eliminate redundant or less impactful variables. This process reduced the dataset to 60 key features, each chosen for its significant contribution to the analysis. By focusing on these critical features, the dimensionality reduction step improved computational efficiency without sacrificing analytical accuracy



d. **Final Dataset**:
   - Rows: 2,519
   - Columns: 60

4. Feature Extraction and Training Model

4.1 Model Selection

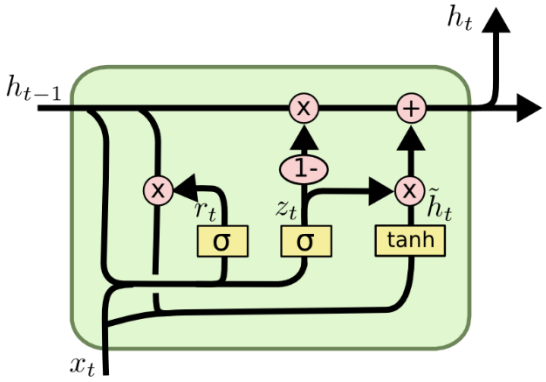
The team decided to use deep learning models, specifically LSTM (Long Short-Term Memory), instead of basic machine learning models for the following reasons:

1. **Ability to Handle Time-Series Data:**
   LSTM is designed to learn and predict time-series data, which is crucial for stock market analysis. Stock price movements are inherently sequential, and understanding temporal dependencies is key to making accurate predictions. LSTM excels at capturing these temporal patterns, making it a more suitable choice compared to traditional machine learning models.

2. **Capability to Retain Long-Term Information:**
   One of the standout features of LSTM is its ability to maintain information over long time steps. This allows it to leverage long-term historical data to forecast stock price fluctuations. In contrast, machine learning models like XGBoost or Random Forest generally struggle with such sequential dependencies, as they lack mechanisms for retaining and processing long-term context.

3. **Adaptability to Unstructured Data:**
   Stock market data can be unpredictable and often lacks a clear structure for straightforward analysis. LSTM is capable of identifying patterns in such unstructured data, enabling it to perform well even when the input data exhibits unexpected variations. Traditional machine learning models, on the other hand, typically require well-prepared and structured datasets, limiting their flexibility in handling real-world stock market scenarios.

By leveraging the strengths of LSTM, the team aims to build a robust and effective model that can navigate the complexities of stock market prediction with greater accuracy and reliability.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

4.2 Training process

The data is split into training and testing sets, with 90% allocated for training and 10% for testing:

1. **Data Splitting:**
   - The first 90% of the data is assigned to the training set (X_train, y_train), while the remaining 10% is used for the testing set (X_test, y_test). This ensures that the model is trained on the majority of the dataset while preserving a separate portion for evaluation.
2. **LSTM Model Architecture:**
   - **LSTM Layers:** The model consists of three LSTM layers, each designed to extract sequential patterns in the data. The number of units in each LSTM layer decreases progressively, helping the model focus on extracting increasingly abstract features. To mitigate overfitting, each LSTM layer is paired with a

Dropout layer, which randomly drops a fraction of the connections during training.

- **Dense Layers:** After the LSTM layers, the model includes a Dense layer with 25 units to process the output of the LSTM stack. Finally, a single-unit Dense layer serves as the output layer, responsible for predicting the target value.

3. **Model Compilation:**
    - The model is compiled using the **Adam optimizer**, which adjusts the learning rate dynamically for efficient training. The loss function is set to **mean_squared_error**, which measures the average squared difference between predicted and actual values, making it suitable for regression tasks like stock price prediction.
4. **Callbacks:**
    - **EarlyStopping:** This callback halts training if the validation loss (val_loss) does not improve for 20 consecutive epochs, preventing the model from overfitting and saving training time.
    - **ModelCheckpoint:** This callback saves the best model during training based on the lowest validation loss. It ensures that the optimal version of the model is retained, even if subsequent epochs lead to overfitting or performance degradation.

By combining these elements, the model is designed to effectively learn from time-series data while minimizing overfitting and maximizing generalization to unseen data.

4.3 Grid Search

Grid Search for the LSTM network will explore combinations of the following hyperparameters to optimize the model for stock price prediction:

1. **Start Dates:**
    - The starting points for training data are varied across different time periods from 2015 to 2023. This ensures that the model is evaluated on datasets with diverse temporal characteristics, accounting for different market conditions over time.
2. **Time Steps Options:**

- The number of time steps (sequence length) considered by the LSTM model will include values of **5, 15, 30, 60, and 120**. These options represent varying levels of historical context, allowing the model to identify patterns over short-term and long-term trends.

3. **Thresholds:**
    - Classification thresholds of **0.5, 0.7, and 0.8** will be tested. These thresholds define the decision boundary for making predictions, helping identify the most suitable value for accurately classifying stock movements.
4. **LSTM Units Options:**
    - The number of units (neurons) in the LSTM layers will be tested with values of **150 and 200**. This parameter controls the capacity of the model to capture complex patterns in sequential data.
5. **Learning Rate Options:**
    - Learning rates of **0.0003** and **0.0007** will be evaluated to determine the best rate at which the model updates weights during training. A smaller learning rate ensures stable convergence, while a larger one may speed up training but risks overshooting the optimal solution.
6. **Epochs Options:**
    - The number of training epochs is fixed at **100** for all configurations. This allows sufficient training time to evaluate the performance of different hyperparameter combinations consistently.

Objective:

The goal of this Grid Search process is to identify the optimal combination of hyperparameters that enables the LSTM model to achieve the best performance in predicting stock values. By systematically evaluating all possible configurations, the approach ensures that the selected model configuration is both effective and well-suited for the task. Here is an example of a model configuration:

Model: "sequential"

Layer (type)	Output Shape
lstm (LSTM)	(None, 60, 150)
dropout (Dropout)	(None, 60, 150)
lstm_1 (LSTM)	(None, 60, 150)
dropout_1 (Dropout)	(None, 60, 150)
lstm_2 (LSTM)	(None, 50)
dropout_2 (Dropout)	(None, 50)
dense (Dense)	(None, 25)
dropout_3 (Dropout)	(None, 25)
dense_1 (Dense)	(None, 1)

Total params: 321,701 (1.23 MB)
Trainable params: 321,701 (1.23 MB)
Non-trainable params: 0 (0.00 B)

5. Result

5.1 Training result

The **direction_accuracy** function is used to calculate the accuracy of the predicted price direction compared to the actual price direction. This function evaluates whether the model correctly predicts the trend (upward or downward movement) of the stock prices rather than focusing solely on the absolute values. The result is expressed as the percentage of correct predictions over the total predictions.

Key Steps in the direction_accuracy Function:

1. **Compute the Direction of Change:**

For both actual and predicted prices, the function determines whether the price increased or decreased compared to the previous time step.

2. **Compare Directions:**

   The predicted and actual directions of change are compared at each step.

3. **Calculate Accuracy:**

   The function calculates the percentage of correct predictions by dividing the number of matches by the total number of predictions.

Best Model Performance:

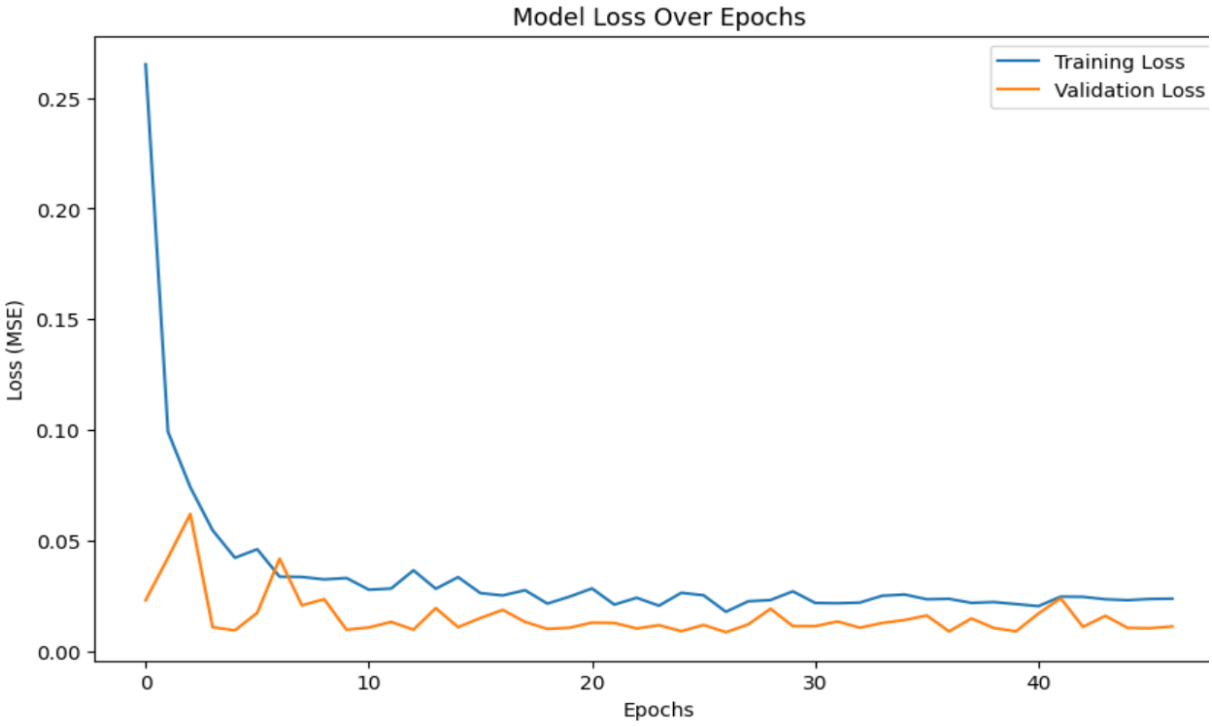
The best model achieved a **direction accuracy** of **65.7%**, indicating that the model correctly predicted the direction of stock price changes in 65.7% of cases.

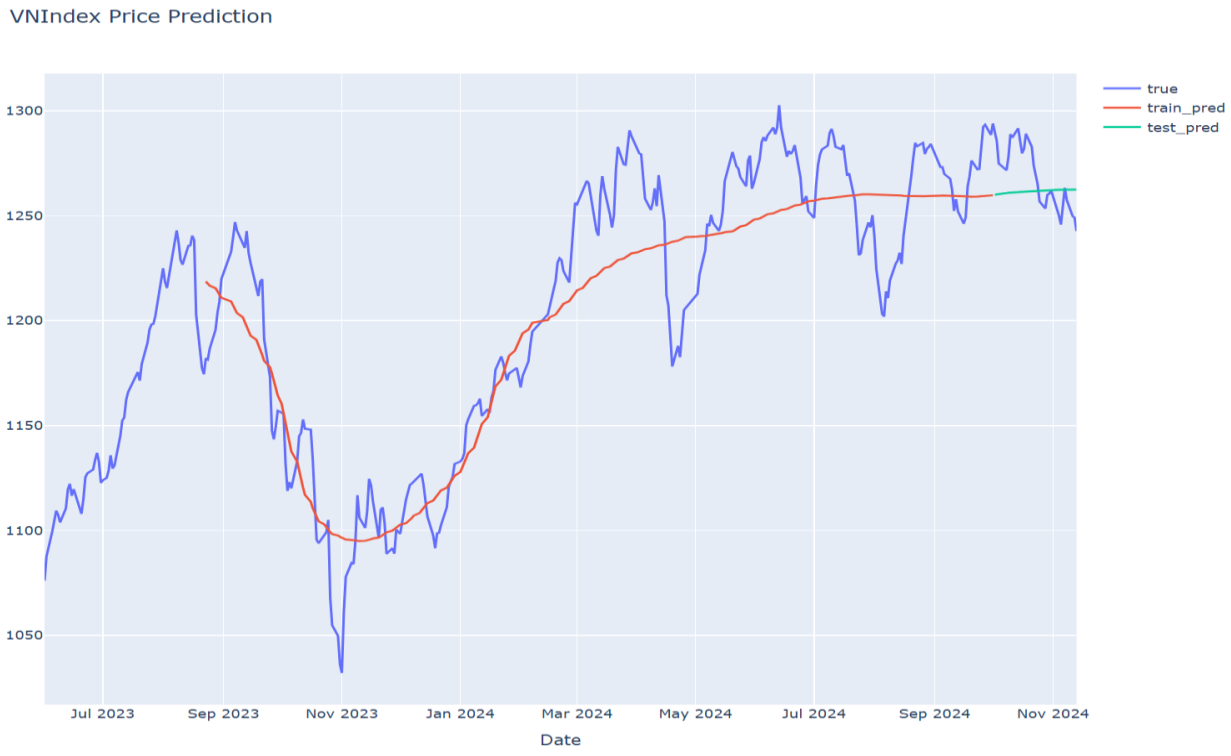
Best Model Configuration:

The optimal hyperparameter configuration for this performance was:

- **Start Date:** 2023-06-01
- **Time Steps:** 60
- **Threshold:** 0.5
- **LSTM Units:** 150
- **Learning Rate:** 0.0003
- **Epochs:** 100

This configuration suggests that using 60 time steps (representing a long-term historical context) and a smaller learning rate allowed the model to effectively learn and generalize the trends in the data. The threshold of 0.5 served as an effective decision boundary for determining the direction of price movement.





5.2 Key Factors

1. **Market Trend Factors:**

The overall market sentiment and trends play a significant role in determining the movement of the VN-Index. Positive investor sentiment, market liquidity, and major buying or selling pressure from institutional investors can drive the index up or down. Additionally, global market trends and regional influences also contribute to fluctuations in the VN-Index.

2. **Actual GDP of Industry Groups:**

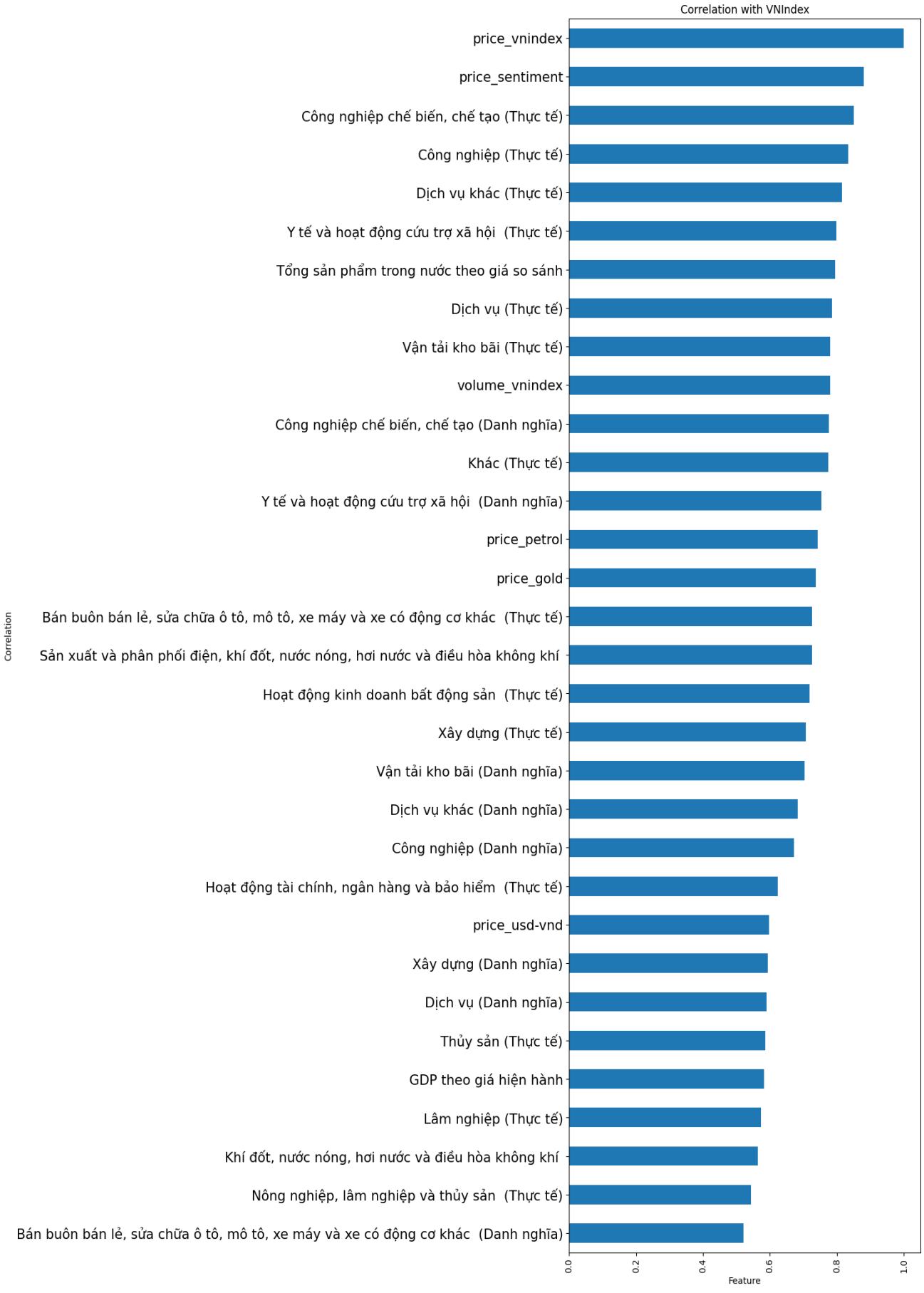
   The VN-Index is closely tied to the performance of various sectors within the economy. Real GDP growth in key industries such as banking, manufacturing, real estate, and technology can significantly impact the index. Strong growth in these sectors indicates a healthy economy, often leading to an increase in stock prices, while weak performance can have the opposite effect.

3. **Prices of Oil, Gold, and Exchange Rates (USD-VND):**

   Oil Prices: Fluctuations in global crude oil prices affect energy-related stocks and industries that rely on oil. Rising oil prices can increase costs for businesses, potentially negatively impacting the VN-Index.

   Gold Prices: Gold is often seen as a safe-haven asset. An increase in gold prices might indicate economic uncertainty, which could result in a decline in stock market investments, affecting the VN-Index.

   Exchange Rates (USD-VND): The value of the Vietnamese Dong (VND) relative to the US Dollar (USD) influences export-driven industries and foreign investment flows. A weakening VND might benefit exporters but deter foreign investors, while a strong VND can have the opposite effect.


Correlation with VNIndex
price_vnindex
price_sentiment
Công nghiệp chế biến, chế tạo (Thực tế)
Công nghiệp (Thực tế)
Dịch vụ khác (Thực tế)
Y tế và hoạt động cứu trợ xã hội (Thực tế)
Tổng sản phẩm trong nước theo giá so sánh
Dịch vụ (Thực tế)
Vận tải kho bãi (Thực tế)
volume_vnindex
Công nghiệp chế biến, chế tạo (Danh nghĩa)
Khác (Thực tế)
Y tế và hoạt động cứu trợ xã hội (Danh nghĩa)
price_petrol
price_gold
Bán buôn bán lẻ, sửa chữa ô tô, mô tô, xe máy và xe có động cơ khác (Thực tế)
Sản xuất và phân phối điện, khí đốt, nước nóng, hơi nước và điều hòa không khí
Hoạt động kinh doanh bất động sản (Thực tế)
Xây dựng (Thực tế)
Vận tải kho bãi (Danh nghĩa)
Dịch vụ khác (Danh nghĩa)
Công nghiệp (Danh nghĩa)
Hoạt động tài chính, ngân hàng và bảo hiểm (Thực tế)
price_usd-vnd
Xây dựng (Danh nghĩa)
Dịch vụ (Danh nghĩa)
Thủy sản (Thực tế)
GDP theo giá hiện hành
Lâm nghiệp (Thực tế)
Khí đốt, nước nóng, hơi nước và điều hòa không khí
Nông nghiệp, lâm nghiệp và thủy sản (Thực tế)
Bán buôn bán lẻ, sửa chữa ô tô, mô tô, xe máy và xe có động cơ khác (Danh nghĩa)
Correlation
0.0
0.2
0.4
0.6
0.8
1.0
Feature

6. Conclusion

By harnessing historical data, predictive modeling, and advanced visualization techniques, this project has illuminated the intricate relationships among macroeconomic factors, market trends, and stock price movements. The results underscore the value of data science as a robust tool for analyzing and predicting VN-Index fluctuations, bridging the gap between raw data and actionable insights

6.1 Strengths:

1. **Comprehensive Analysis:**
   This project successfully integrated a wide array of macroeconomic indicators and market data, including GDP growth, inflation rates (CPI), interest rates, and exchange rates. By doing so, it offered a holistic and multi-dimensional perspective on the factors influencing VN-Index trends.
2. **Effective Use of Modern Tools and Techniques:**
   The adoption of Python and its rich ecosystem of data analysis libraries, coupled with machine learning models and advanced data visualization frameworks, facilitated efficient data processing, accurate modeling, and intuitive result presentation. These technologies not only enhanced the depth of analysis but also improved the accessibility and interpretability of findings for diverse stakeholders.
3. **Practical Applications:**
   The insights derived from this project have significant real-world implications. Investors, financial institutions, and policymakers can utilize these findings to make informed decisions, optimize portfolio strategies, and anticipate market shifts more effectively.

6.2 Limitations:

1. **Data Availability and Scope:**
   Despite extensive efforts, the analysis was limited by the availability and quality of certain datasets. Key factors, such as international economic trends or high-frequency trading data, were not included due to data constraints. These omissions may have led to an incomplete representation of all potential influences on VN-Index performance.
2. **Model Dependency and Sensitivity:**
   The accuracy and reliability of the predictive models were inherently dependent on the completeness and quality of the input data. Additionally, the models might underperform in capturing sudden or rare market disruptions that deviate from historical patterns.

6.3 Future Directions:

To address these limitations and further expand the scope and impact of the analysis, the following directions are proposed:

1. **Expanding Data Sources:**
   Incorporate international economic data, geopolitical events, and real-time market information to capture a more comprehensive range of factors affecting VN-Index fluctuations. This could include global indices, commodity prices, and cross-border investment flows.
2. **Exploring Advanced Modeling Techniques:**
   Experiment with ensemble methods and hybrid machine learning models that combine the strengths of different algorithms to improve forecasting accuracy. Additionally, integrating time-series-specific methods like ARIMA, Prophet, or LSTM-based models may provide deeper insights into temporal trends.
3. **Developing Real-Time Analytical Systems:**
   Build an automated system for real-time data collection, processing, and model updating. Integrating MLOps practices would ensure the continuous improvement of models, scalability of the system, and efficient deployment of insights to end-users.
4. **Enhancing Interpretability and User Engagement:**
   Prioritize the development of user-friendly dashboards and explainable AI (XAI) methods to make predictions and insights more accessible and actionable for non-technical stakeholders.

In conclusion, this project lays a solid foundation for leveraging data science in financial market analysis. By addressing its current limitations and embracing future advancements, it has the potential to evolve into a powerful framework for predicting market behaviors, supporting strategic investment decisions, and fostering resilience in the face of economic uncertainties. Such initiatives not only benefit individual investors and institutions but also contribute to the overall stability and efficiency of financial markets.

References

1. Ho Chi Minh City Stock Exchange: https://www.hsx.vn
2. WHAT IS VN-INDEX AND ITS IMPORTANCE TO INVESTORS - HVA Group
3. Vietstock Financial Data: https://vietstock.vn
4. General Statistics Office of Vietnam: https://www.gso.gov.vn/
5. Yahoo Finance API: https://pypi.org/project/yfinance/
6. AAII Sentiment Survey: https://www.aaii.com/sentimentsurvey/sent_results
7. Springer Research: https://link.springer.com/article/10.1007/s10203-021-00328-8