# Market Analysis Utilizing Time Series Analysis to Forecast Future Market Trends

LI QIZHI

UNIVERSITI TEKNOLOGI MALAYSIA

**ABSTRACT**

This project incorporates time series analysis in making projections for future financial markets. Building a model based on historical market data, with the unique determinants, we will construct a predictive model explaining the market trends, which gives useful information to decision-makers and investors. We will surpass the challenges of the regular market analysis by mental labs of the trends in the financial market and using fore-developed time series models.

The project will involve several key steps: Data collection and cleaning: To this end, we will obtain historical market data, clean and preprocess these data to ensure data quality during the analysis process. Determinant analysis: We are going to explore key economic indicators, company performance indicators, as well as investor sentiment indicators, to determine the factors most influential on market movements. Model selection and development: We will apply several time series models, such as Arima, SARIMA, and Exponential Smoothing, to predict market trends and choose the best one based on the results. Model evaluation: Evaluating the fact of predicting the market with historical data for backtesting will be done as well as the use of different indicators to determine the accuracy of the prediction. Actionable insights and advice: The model results, trends, and risk factors will be monitored to give advice and recommendations for the businesses and investors in the market.

Through the project work, a chronological dynamics will be analyzed in spoiling the gamble. Thus, better investment options will be made basing on established knowledge.

Click or tap here to enter text.

**TABLE OF CONTENTS**

**TITLE** **PAGE**


**ABSTRACT** **I**

**TABLE OF CONTENTS** **II**

**CHAPTER 1 INTRODUCTION** **1**

1.1 Introduction 1

1.2 Background of the Problem 1

1.3 Statement of the Problem 2

1.4 Research Questions 2

1.5 Objectives of the Research 3

1.6 Scope of the Study 3

1.7 Significance of the study 4

**CHAPTER 2 LITERATURE REVIEW** **4**

2.1 Introduction 4

2.2 Economic Indicators 5

2.3 Company Performance Indicators 6

2.4 Investor Sentiment Indicators 8

2.5 ARIMA, SARIMA and exponential smoothing for accurate forecasting 8

2.6 Comparative Performance of ARIMA, SARIMA, and Exponential Smoothing 9

**CHAPTER 3: RESEARCH METHODOLOGY** **10**

3.0 Data Science Project Life Cycle 10

3.1 Research design 11

3.2 Problem identification 12

3.3 Data collection and pre-processing 12

**Chapter 4 INITIAL RESULT** **16**

4.1 Exploratory Data Analysis (EDA) 16

**CHAPTER 5 DISCUSSION 32**

5.1 Summary 32

5.2 Future Work 33

# CHAPTER 1 INTRODUCTION

## 1.1 Introduction

The financial market trades with prevailing economic conditions and is hence vital for businesses and investors because of its volatility. More accurate forecasting of the market's dynamics can help businesses establish fairer and more efficient strategies and may also enable the investors to put their money in the place that gives them the best return. Nonetheless, the complexities of the financial market make it hard for any offline prediction mechanism to understand the heightened interactions. These days, while data science and machine learning obviously proceed at a rapid pace, financial forecasting turns to time series analysis more and more. the help of statistical methods called as time series, an appraisal or forecasting is possible from sequenced time ordered data. It is time series that will help us in finding the patterns, such as trends and seasonality, and also in getting the information about the cyclical structure of the data and in building models for future values prediction. This research will focus on applying time series analysis to forecast ongoing financial market trends. Through the use of analytics on historical data of the market that includes some factors, including fundamental, we can be able to build a predictive model to offer insightful information to both investors and businesses, where they can understand the diversity of market dynamics and further make more informed decisions.

## 1.2 Background of the Problem

It is essential to be very accurate in trend identification in order to make well-informed decisions in line with businesses and stock dealers. The time series analysis is thus regarded as the most important tool for studying how markets

fluctuate from one period to the next. Nevertheless, traditional approaches to tend studying often neglect the entirety of the financial industry.

## 1.3 Statement of the Problem

A good financial forecasting model for the financial market is a complex job in which one has to take stock of the fundamental factors: Data Quality: These elements are the necessary statistics, compliance, and confirmation for making an accurate collection of predictions. Determinants: Market behavior depends on a number of different factors, and the selection of the key elements is one of the most important things to consider. Model Selection: It is really important to select an appropriate model considering not only the data characteristics but also the main goals of the prediction. Model Evaluation: Evaluation of the selected model involves a consideration of the specified outcomes and (or) use of the additional outcome indicators.

## 1.4 Research Questions

What factors are of more weight in the move of stock markets? We will also specify the variables related to economy, company, and investors' opinion as we try to find the drivers of the markets themselves. How to set up a solid and accurate time series model that will advise us about market trends? We will be introduced to the Arima, SARIMA model and Exponential Smoothing method and find out the most suitable way to predict market trends. How trust can we describe the model in terms of future related market movements? Historical data will be used for the back testing approach, and different parameters and the prediction accuracy level of the model will be done by using the selected indicators. To convey the model prognosis as advice being addressed to entrepreneurs and their funds: Businesspeople and investors will

receive advice based on market statistics, and potential risks, and outlook based on model prediction, and take appropriate action.

## 1.5 Objectives of the Research

Learn in depth the time series forecasting methods both theoretically and practically: We will study the most basic principles and techniques of time series, and also we implement different time series forecasting models. Acquiring market data and handling it: We will collect the data that would be necessary through data scraping, clean it, and will carry-out other data management tasks. Discerning the main market directions and the key influences: Through application of statistical analysis methods, it is possible to answer questions about market trends and writings of the same. Establishing a time series prediction model and validating it: We will consider the knowledge base and the intended goals as we choose a model and apply historical data for testing its efficiency. Results Revealed from Model Forecasts traversing several areas including business and funds: We are going to analyze market trends and judge the risks related to it with the help of model prediction results, such that advice and suggestions could be offered to business men and investors.

## 1.6 Scope of the Study

The study includes the selection of a specific area (e.g., stock market, real estate market) and the development of the prediction model based on historical data. The model is constructed at this stage by treating a different set of test parameters to test for the provision of accurate and reliable forecast.

## 1.7 Significance of the study

This analysis will offer a comprehensive plan for market analysis, which will be devoted to predicting the market movements through an analysis of trends and determinants. It also takes a step to put up a robust and declarative time series model, which would give helpful suggestions on the future prices. Besides, we may offer valuable advice and solutions for entrepreneurs and fund managers backed up with the help of model forecasts, which can help them in searching for a niche in the market and making the right decisions.

# CHAPTER 2 LITERATURE REVIEW

## 2.1 Introduction

The financial market trades with prevailing economic conditions and is hence vital for businesses and investors because of its volatility. More accurate forecasting of the market's dynamics can help businesses establish fairer and more efficient strategies and may also enable the investors to put their money in the place that gives them the best return. Nonetheless, the complexities of the financial market make it hard for any offline prediction mechanism to understand the heightened interactions.Because the key factors that affect market trends include: economic indicators, company performance indicators and investor sentiment indicators. This literature review discusses the impact of these factors and uses time series models such as ARIMA, SARIMA and exponential smoothing for accurate forecasting.

## 2.2 Economic Indicators

Among economic indicators, Macroeconomic factors like GDP, inflation, and interest rates are crucial for understanding market trends. (Levitt, A. 2008) These indicators shed light on the viable and developing prospects of the ultra-modern economy influencing a vast range of sectors, businesses, and consumer behavior.

**Gross Domestic Product (GDP):** GDP represents the total market value of all final goods and services produced within a country in a given period (Samuelson & Nordhaus, 2010) The growing GDP indicates a vibrant economy where people are employed and spend more. A sustained decline in GDP could mean the economy is not growing, stagnating, or going into a recession stage (Romer, 2012). Not only does this knowledge of GDP movement spur business and policymakers in anticipating shifts, it also helps in demand and investment pattern projection (Mankiw, 2014).

**Inflation:** Inflation, that is, the rate at which an average price level of commodities and services generally rises, is a key element in all economic analysis (Blanchard, 2017). Normal inflation rates demonstrate that the economy is advancing, while room for folly is identified by those who despise excessive inflations and how it influences the purchasing power and confidence of the consumer (Friedman, 1968). Central banks engage in the implementation of monetary policy as a way of influencing the economy by controlling the interest rates and the money supply (Mishkin, 2015).

**Interest Rates:** Interest rates, which are mainly the concern of the central banks, are the most significant reason in the running of the economy through determining borrowing, investing, and saving trends (Bernanke, 2010). The result of high-interest rates is to make borrowing unattractive so as reduce the money circulation in the economy, while low-interest rates might encourage spending and borrowing which in return may stimulate the growth of the economy (Clarida, Gali, & Gertler, 1999). The relationship between interest rates and inflation is key in the management of the macro economy. Given, central banks have to adjust their rates

accordingly in order to curb inflation to ensure that the economy is stable (Taylor, 1993).

### 2.3 Company Performance Indicators

Among company performance indicators, capital market research should focus on how to extract effective signals that can predict market reactions from accounting information. (Kothari, S. P. 2001) Evaluating a firm's performance involves comparing the financial and operational indicators and evaluating the correlation with the company's overall market value. Therefore, these indicators display the financial health of the firm and at the same time provide predictive signals and sentiments about its future performance and the market in general. KPIs like profitability, liquidity, efficiency, and leverage ratios are often used to determine a company's ability to survive in the market and to guide the investors in their investment decisions.

**Profitability Ratios:** The profitability ratios, i.e., return on equity (ROE), return on assets (ROA), and profit margins, are the indicators of a company's (CCIE) generation of the profits in the relation to the revenue, assets, and equity. Such ratios seem to be used by investors as a means of assessing the extent to which a company converts its resources into profit. The information extracted from such performance indicators can have a significant impact on stock price movements, as high profitability often signals strong growth potential, positively influencing investor sentiment (Penman, 2013).

**Liquidity Ratios:** Liquidity ratios, which include the current ratio and quick ratio, show how well a company can pay its short-term obligations due to liquid assets. These ratios are key indicators for understanding debtors' financial soundness, as companies with sufficient liquidity are better positioned to withstand uncertainty and unexpected risks, which makes them more attractive to the creditors (Pervan et al. 2019).

**Efficiency Ratios:** Efficiency ratios, for instance, asset turnover ratio and the inventory turnover rate, are surely very pivotal criteria to gauge how effectively the company is profiting from its assets. Increased efficiency fairly reveals better operational and potential for profits, while inefficiency leads to increased operational costs and hindrance of profit. Market participants closely watch efficiency indicators as they offer insights into the company's cost structure and capacity to deliver consistent earnings (Graham & Dodd, 2009).

**Leverage Ratios:** The like of debt-to-equity ratio is the one used to determine how much of a company evades structural decisions that they can derive profits from. While adequate leverage can be the driver for increased returns and insight on the success of the company due to vision for growth, excessive leverage poses a serious demand of high financial risk, which in turn may impair companies' capability to cope with market wavering; hence recommended to the investors with a cautionary voice. The market's interpretation of a company's leverage influences stock price fluctuations and signals investors regarding the level of risk they might face (Miller & Modigliani, 1961).

Apart from traditional financial indicators, performance indicators of a firm, when derived from accounting data also essential to understanding the company's financial health and its evolution. Capital market research makes use of those indicators to produce ahead of time those predictive signals that will aid investors in forecasting what will be the market reaction afterward, helping them to make informed investment decisions and to manage their risks. By analyzing how market participants respond to changes in these performance indicators, analysts can better forecast future movements in stock prices and overall market behavior (Kothari, S. P. 2001).

## 2.4 Investor Sentiment Indicators

There are sentiment indicators that help determine the sentiment of investors-VIX is the one such instrument used in the stock market. VIX index is also acknowledged as a "fear gauge" which usually goes up during times of market hesitation，uncertainty or crisis, reflecting the sharp growth of the investor's fear that was noted in the market. During these times, investors anticipate higher volatility, and the demand for options as a hedge against market downturns increases, causing the VIX to climb (Whaley, R. E. 2000). In contrast, when the market is relatively stable, and investor confidence is high, the VIX tends to be lower, indicating a calm and optimistic outlook among market participants (CBOE, 2023).

**VIX and Market Movements:** Because the correlation between the VIX and the overall market is most often observed in the negative cases, it is a certainty that the score of the market against the volume of VIX is inversely related. In case the market becomes uncertain or fear is getting higher, and consequently the VIX grows, it is normally a sign that the market is heading towards a drop or the risk for the investments is getting higher. The VIX can also serve as a leading indicator, as sharp increases in volatility often precede downturns or corrections, while declines in the index can signal periods of economic recovery or market stability (Fear & Gustafsson, 2011).

## 2.5 ARIMA, SARIMA and exponential smoothing for accurate forecasting

Time-series forecasting plays a crucial role in a number of disciplines, such as finance, logistics, and energy management. Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing are but some of the techniques that offer powerful tools for analytical data used in making decisions. Models are defined based on the time-series data of different patterns observed in the data, and whose main application is dependent on those patterns.

ARIMA is especially suitable for data that needs stabilization before forecasting, such as economic or stock price time-series (Box et al., 2015).For example, Akaike (1974) demonstrated ARIMA's power in financial forecasting through its ability to analyze and predict stock market trends using historical price data. An example includes its use for predicting daily stock price indices by combining ARIMA with additional autoregressive terms for error adjustment.

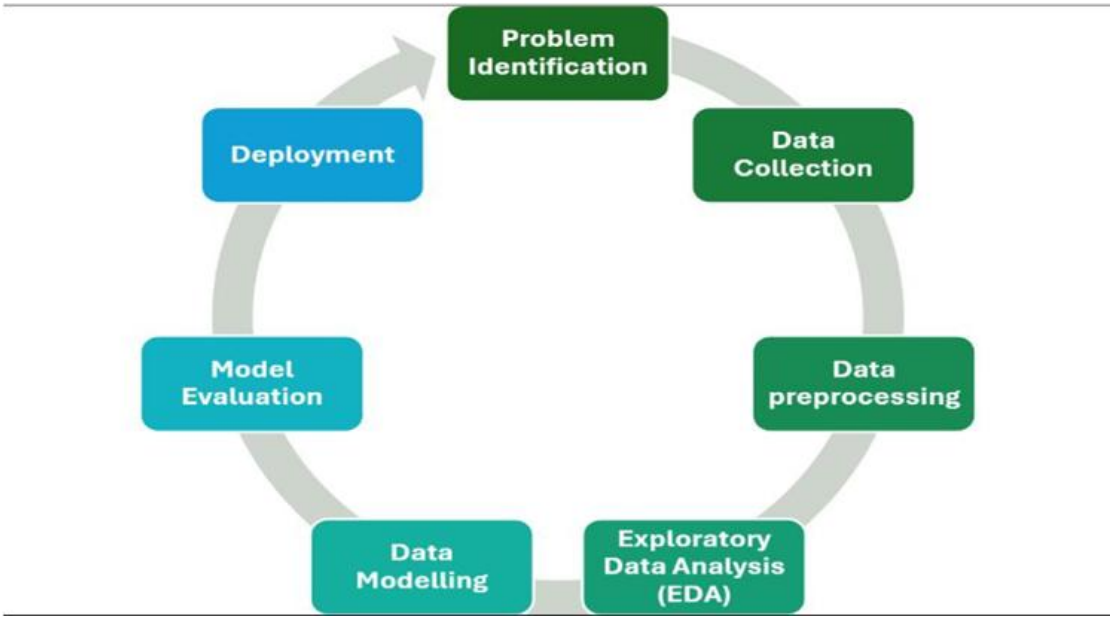
### 2.6 Comparative Performance of ARIMA, SARIMA, and Exponential Smoothing

| Aspect | ARIMA | SARIMA | Exponential Smoothing |
|---|---|---|---|
| Seasonality Handling | Not explicitly modeled | Explicitly modeled | Modeled in Holt-Winters variant |
| Trend Management | Captures through differencing | Combines differencing and seasonal components | Captured in Holt's method |
| Computational Cost | Moderate | High due to additional parameters | Low to moderate |
| Forecasting Horizon | Short-to-medium term | Medium-to-long term | Short-term |

Conclusion:Each model (ARIMA, SARIMA, and Exponential Smoothing) has its strong and weak points among the applications of forecasting in time series. ARIMA works exceptionally well in a scenario that captures non-seasonal elements, SARIMA provides an extremely good fit for datasets that depict cyclical patterns, on the contrary, Exponential Smoothing is known for its approach of having an easy and flexible application in real-time systems

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.0 Data Science Project Life Cycle




In this study, we adopted a systematic framework of data science lifecycle, which is developed to ensure the quick and easy flow of data management, processing, and model development. There are seven fundamental stages in this lifecycle whose implementation would lead to the project's growth and success, and hence, getting accurate and reliable results. Such stages are:

1. Problem identification: Soliciting information in this step involves delineating the question or query intended for our project to confirm that our line of action brings to the table practicality and relevance.

2. Data collection: We acquire the necessary data using a variety of (or form) means and methods, which will form the basis for further analysis and modeling.

3. Data preprocessing: The information is cleaned, formatted, missing data is dealt with, and the data is then transformed in order to improve the quality and the feasibility level of the data.

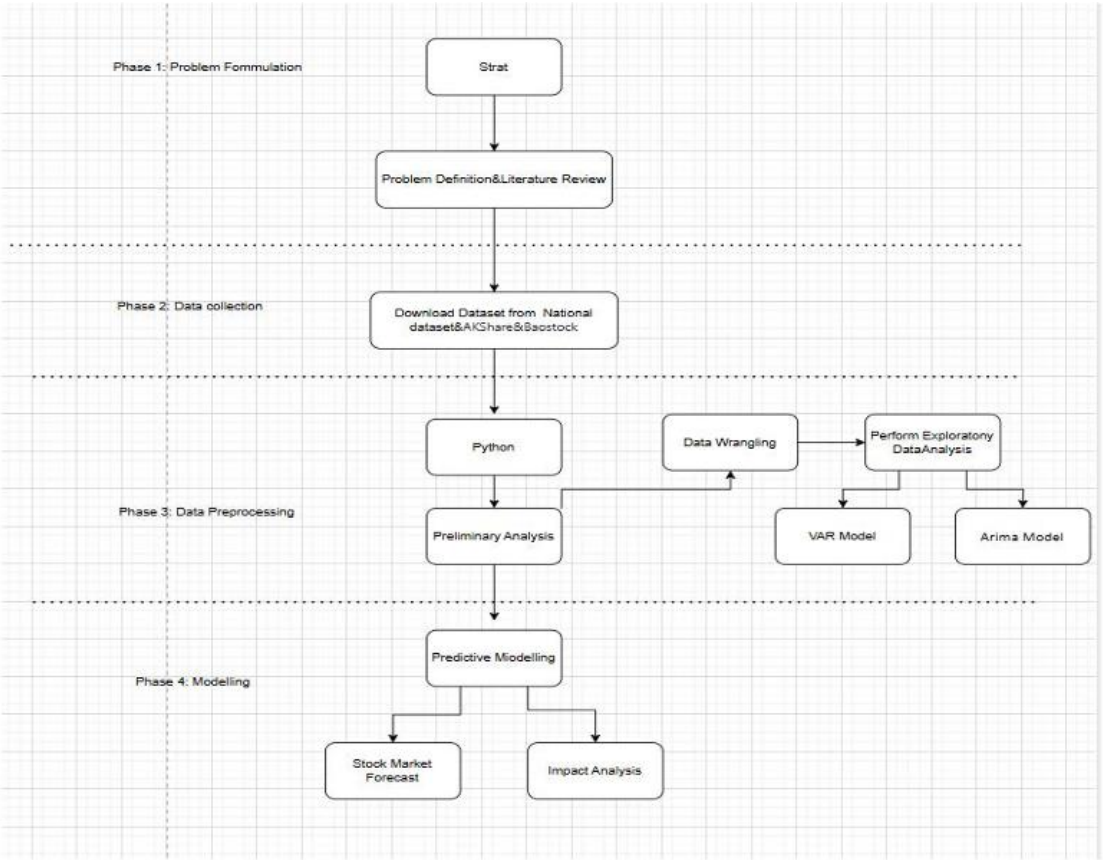
4. Analysis: Around the exploratory data analysis (EDA), we explore in the data attributes to understand it more emotively and find the patterns and relationships between each other.

5. Data modeling: This stage is marked by the usage of specific algorithms in the building of a model to predict or describe the scientific problem.

6. Evaluate model performance: The model's efficiency is validated through cross-validation, test data evaluation, and the use of other relevant methods.

7. Deployment: Last but not least, we develop the application and realize the implementation of the verified models to be used at the aim of the project

### 3.1 Research design



### 3.2 Problem identification

| Research questions | Research Objectives | Proposed solutions |
|---|---|---|
| 1.Complexity of financial market forecasting: Financial markets are influenced by a variety of factors, including macroeconomic indicators, company performance indicators, and investor sentiment indicators. The interaction and dynamic changes of these factors make it extremely challenging to accurately predict market trends. | Determine whether GDP, company performance indicators and sentiment indicators have an impact on stock market fluctuations | The vector autoregression (VAR) model is a commonly used econometric model used to predict and analyze the dynamic relationship between multiple time series variables. When studying the factors affecting the stock price fluctuations of listed companies, the VAR model can be used to explore the relationship between stock prices and variables such as GDP, company performance indicators, and sentiment indicators. |
| 2.Combined with correlation factors, the time series model is used to predict stock market price fluctuations and determine the accuracy of the model. | Determine whether the factors affecting stock price fluctuations are reasonable | Develop and train traditional time series models to predict stock price movements based on relevant influencing factors. |

### 3.3 Data collection and pre-processing

#### 3.3.1 Data collection

The key data required for this project include: GDP values, company performance indicators, sentiment indicators and stock market fluctuations.

| Dataset | Description | Data source |
|---|---|---|
| China GDP Dataset | Time:2014-2023 | National data: https://data.stats.gov.cn/easyquery.htm?cn=C01 |
| Company Performance Dataset | Code: Stock Code<br>pubDate: The date the company releases its earnings report<br>statDate: The last day of the quarter for which the financial report is reported<br>roeAvg: Return on net assets (average) (%)<br>npMargin: Net profit margin (%)<br>gpMargin: Gross profit margin (%)<br>epsTTM: Ne tprofit (RMB)<br>netProfit: Earnings per share<br>MBRevenue: Main operating income (RMB)<br>totalShare: Total share capital<br>liqaShare: Circulating share capital | Get securities data information through Python API and use Baostock to get stock data |
| Stock Market Dataset | Time: date<br>Code: Stock Code<br>Opening: Starting Price<br>Close: Last Price<br>Highest: Highest Price<br>Lowest: Lowest Price<br>Volume: Transaction Quantity Transaction<br>amount: Transaction amount | AKShare is a Python financial data interface library suitable for various financial data acquisition and processing needs. |

| | | |
|---|---|---|
| | Amplitude: Highest and Lowest Difference<br>Rise and fall: Percentage increase<br>Change: Changing the amount of money<br>Turnover rate: Percentage of Buying and Selling | |

**3.3.2Data pre-processing**

(1) Datacleaning: Deduplication: Get rid of duplicate records from data. Handling missing values: Replace, throw away, or use interpolation techniques for missing data. Correcting erroneous data: Deal with faulty or inaccurate data, which means identifying and amending mistakes found in a dataset.

(2) Datatransformation: Data normalization: Restriction of the data values to the particular, fairly small interval, typically [0,1] or [-1,1]. Data standardization: Alteration of the data parameters to the properly chosen form with mean = 0 and variance = 1. Data encoding: Translate the books to the data language that makes it understandable; for example, one-hot encoding, label encoding, etc.

(3) Feature Engineering: Feature Selection: Determine the chief attributes from a current list of attributes for further analysis. Feature Extraction: Extract fresh, additional characteristics (features) coming from raw data, an example could be PCA.

3.4 Exploratory Data Analysis (EDA)

**Var Model:** VAR is a statistical model used to analyze the relationship between multiple time series variables. As mentioned above, the correlation between the three influencing factors and the stock market is determined by using the VAR model. The correlation between GDP, company performance indicators and sentiment indicators and the stock market is determined.

**Arima Model:** Substitute three possible influencing factors into the time series model to predict future stock market trends.

3.5 Forecasting Modeling

**Stock Market Forecast:** Use ARIMA to perform time series analysis to predict the future stock market based on historical data. Use the VAR model to

determine whether GDP, company performance, and sentiment indicators have an impact on the stock market. Then use the time series model to bring in the impact shadow to predict the future stock market.

3.6 Model evaluation and validation

**Model Evaluation:**

**Residual Analysis:** It is important to check the residuals of the ARIMA model (the discrepancies between observed and forecasted values) after the model is succeeded in being implemented. Model Fit Statistics: For comparing different models, use these fit statistics- Akaike Information Criterion (AIC), Bayesian Information Criteria (BIC), and R-squared. Diagnostic information related to AIC and BIC is known: the lower the values, the better the model.

**Model Validation:**

**Hold-Out Validation:** Partition the data into two distinct parts: training set and testing set. The model is run on training data. After that, its performance on testing data is evaluated to give importance to the accuracy. So, the predictive power of the data model on new andunseen data is depicted more accurately.

**Cross-Validation:** Time-series cross-validation often becomes more complex, but it is even more necessary as the data ordering in time is sequential. One approach could be a gradient boosting machine (GBM) implemented with rolling forecasts in which the model recursively trained using a learning period to update predictions.

**Steps for Model Validation**

**Fit the Model:** Ensure that the presented ARIMA model is fit to the training data.

**Make Forecasts**: Forecasts using the model are made for the test set.

**Calculate Errors:** You calculate the forecast errors by matching the forecasts to the data in the test set.

**Assess Accuracy:** Forecasts are compared against observed data. There are many ways to do so; you can employ a more formal statistical testing for the model's predictions or simply comparing actual values with the model's predictions.

**Iterate:** This allows for rectifying the model to hit the desired accuracy or to go for another model in case the present one proves to be unsatisfactory.

# Chapter 4 INITIAL RESULT

## 4.1 Exploratory Data Analysis (EDA)

### 4.1.1 Data processing

Step 1: Read the data of the three datasets

```
[5]: import pandas as pd

stock_data = pd.read_csv('stock_data.csv')
profit_data = pd.read_csv('profit_data.csv')
china_data1_0 = pd.read_csv('china_data1_0.csv')

stock_data_head = stock_data.head()
stock_data_head
```

[5]:

| | 日期 | 股票代码 | 开盘 | 收盘 | 最高 | 最低 | 成交量 | 成交额 | 振幅 | 涨跌幅 | 涨跌额 | 换手率 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2011-06-30 | 2594 | 22.00 | 25.45 | 26.19 | 22.00 | 562925 | 1.349342e+09 | 23.28 | 41.39 | 7.45 | 87.96 |
| 1 | 2011-07-01 | 2594 | 25.77 | 28.00 | 28.00 | 25.00 | 335613 | 8.986792e+08 | 11.79 | 10.02 | 2.55 | 52.44 |
| 2 | 2011-07-04 | 2594 | 28.88 | 30.80 | 30.80 | 28.58 | 219646 | 6.535179e+08 | 7.93 | 10.00 | 2.80 | 34.32 |
| 3 | 2011-07-05 | 2594 | 32.78 | 33.05 | 33.88 | 31.89 | 381081 | 1.257046e+09 | 6.46 | 7.31 | 2.25 | 59.54 |
| 4 | 2011-07-06 | 2594 | 32.80 | 33.40 | 35.55 | 32.11 | 312967 | 1.056614e+09 | 10.41 | 1.06 | 0.35 | 48.90 |

Figure 4.1 Reading Data

```
[7]: profit_data_head = profit_data.head()
profit_data_head
```

[7]:

| | code | pubDate | statDate | roeAvg | npMargin | gpMargin | netProfit | epsTTM | MBRevenue | totalShare | liqaShare |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | sz.002594 | 2011-08-23 | 2011-06-30 | 0.014288 | 0.017500 | 0.157585 | 3.945320e+08 | 0.160401 | 2.183512e+10 | 2.354100e+09 | 64000000.0 |
| 1 | sz.002594 | 2011-10-29 | 2011-09-30 | 0.018285 | 0.014911 | 0.158514 | 5.119430e+08 | 0.188447 | NaN | 2.354100e+09 | 79000000.0 |
| 2 | sz.002594 | 2012-03-26 | 2011-12-31 | 0.069957 | 0.032668 | 0.171793 | 1.595076e+09 | 0.588176 | 4.728200e+10 | 2.354100e+09 | 79000000.0 |
| 3 | sz.002594 | 2012-04-26 | 2012-03-31 | 0.001279 | 0.004568 | 0.174024 | 5.360600e+07 | 0.486353 | NaN | 2.354100e+09 | 79000000.0 |
| 4 | sz.002594 | 2012-08-28 | 2012-06-30 | 0.000770 | 0.004612 | 0.156560 | 1.041540e+08 | 0.478115 | 2.198346e+10 | 2.354100e+09 | 79000000.0 |

Figure 4.2 Reading Data

```
[9]: china_data1_0_head = china_data1_0.head()
     china_data1_0_head
```

[9]:

| | 年份 | 中国 | GDP(美元) | 占世界% |
|---|---|---|---|---|
| 0 | 2023 | 17.79万亿 (17,794,781,986,104) | 16.8775% | NaN |
| 1 | 2022 | 17.88万亿 (17,881,783,387,000) | 17.6654% | NaN |
| 2 | 2021 | 17.82万亿 (17,820,459,508,852) | 18.2723% | NaN |
| 3 | 2020 | 14.69万亿 (14,687,744,162,801) | 17.1630% | NaN |
| 4 | 2019 | 14.28万亿 (14,279,968,506,271) | 16.2373% | NaN |

Figure 4.3 Reading Data

Step 2: Combine the three datasets

```
[11]: # Remove the market prefix from the 'code' column in profit_data
profit_data['股票代码'] = profit_data['code'].str.replace('sz.', '').astype(int)

# Ensure the '日期' column in stock_data and 'statDate' in profit_data are in datetime format
stock_data['日期'] = pd.to_datetime(stock_data['日期'])
profit_data['statDate'] = pd.to_datetime(profit_data['statDate'])

# Merge stock_data and profit_data on the common columns '股票代码' and 'statDate'
merged_data = pd.merge(stock_data, profit_data, how='left', left_on=['股票代码', '日期'], right_on=['股票代码', 'statDate'])

# Convert the '年份' column in china_data1_0 to integer format
china_data1_0['年份'] = china_data1_0['年份'].astype(int)

# Add the china_data1_0 to the merged data, matching the '年份' with the year part of '日期' in merged_data
# First, extract the year from the '日期' column in merged_data
merged_data['年份'] = merged_data['日期'].dt.year

# Now merge the china_data1_0 into the merged_data
final_data = pd.merge(merged_data, china_data1_0, how='left', on='年份')

# Display the first few rows of the final_data
final_data.head()
```

[11]:

| | 日期 | 股票代码 | 开盘 | 收盘 | 最高 | 最低 | 成交量 | 成交额 | 振幅 | 涨跌幅 | ... | gpMargin | netProfit | epsTTM | MBRevenue | totalShare | liqaShare |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2011-06-30 | 2594 | 22.00 | 25.45 | 26.19 | 22.00 | 562925 | 1.349342e+09 | 23.28 | 41.39 | ... | 0.157585 | 394532000.0 | 0.160401 | 2.183512e+10 | 2.354100e+09 | 64000000.0 |
| 1 | 2011-07-01 | 2594 | 25.77 | 28.00 | 28.00 | 25.00 | 335613 | 8.986792e+08 | 11.79 | 10.02 | ... | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | 2011-07-04 | 2594 | 28.88 | 30.80 | 30.80 | 28.58 | 219646 | 6.535179e+08 | 7.93 | 10.00 | ... | NaN | NaN | NaN | NaN | NaN | NaN |

Figure 4.4 Merge Data

Step 3: Process null values and delete unnecessary data

```
final_data_cleaned.head()
```

| | TIME | Stock_Price | Turnover rate | GDP(Trillion) | npMargin |
|---|---|---|---|---|---|
| **0** | 2011-06-30 | 25.45 | 87.96 | 7.55 | 0.017500 |
| **64** | 2011-09-30 | 19.30 | 6.84 | 7.55 | 0.014911 |
| **366** | 2012-12-31 | 20.35 | 2.66 | 8.53 | 0.004544 |
| **543** | 2013-09-30 | 39.36 | 1.87 | 9.57 | 0.016119 |
| **604** | 2013-12-31 | 37.68 | 1.22 | 9.57 | 0.014677 |

Figure 4.5 Cleaning Data

| Variable | Describe |
|---|---|
| TIME | Time Sequence |
| Stock_price | Stock Price |
| Turnover rate | Represents high and low emotions |
| GDP(Trillion) | Gross Domestic Product |
| Npmargin | Represents the company's profitability |

### 4.1.2 VAR Model

The vector autoregression model, referred to as the VAR model, is a generalization of the AR model and is a commonly used econometric model. It is used to determine the correlation.

**Step 1: Define variables**

Variable definition:

1. Turnover rate (represents the public sentiment towards this stock): Turnover rate
2. Company net profit margin (represents the company's operating conditions): npMargin
3. China's GDP (represents the value of GDP): GDP(Trillion)
4. Stock price: Stock_Price

**Step 2: Stationarity test**

The stability test method used is the ADF unit root test. Generally speaking, the P value of the unit root test t statistic should be significant at the 5% significance level. If the P value is lower than 0.05, the time series can be considered stable. If it is higher than 0.05, it is unstable. The unstable time series can be processed by first-order difference to test the stability of its first-order difference terms. If it is still unstable, the second-order difference method can be used to test until it is stable.

Null Hypothesis: LNTURNOVER RATE has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -5.488289 | 0.0001 |
| Test critical values: | 1% level | -3.653730 | |
| | 5% level | -2.957110 | |
| | 10% level | -2.617434 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNNPMARGIN has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -2.739282 | 0.0786 |
| Test critical values: | 1% level | -3.653730 | |
| | 5% level | -2.957110 | |
| | 10% level | -2.617434 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNGDP TRILLION has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -2.472742 | 0.3384 |
| Test critical values: | 1% level | -4.273277 | |
| | 5% level | -3.557759 | |
| | 10% level | -3.212361 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNSTOCK PRICE has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -1.989035 | 0.5850 |
| Test critical values: | 1% level | -4.273277 | |
| | 5% level | -3.557759 | |
| | 10% level | -3.212361 | |

*MacKinnon (1996) one-sided p-values.

Figure 4.6 ADF test results

| variable | ADF test value | 5% critical value | P-value | Test results |
|---|---|---|---|---|
| lnTurnover rate | -5.488289 | -2.957110 | 0.0001 | Stable |
| lnnpMargin | -2.739282 | -2.957110 | 0.0786 | Unstable |
| lnGDP(Trillion) | -2.472742 | -3.557759 | 0.3384 | Unstable |
| lnStock_Price | -1.989035 | -3.557759 | 0.5850 | Unstable |

Table 4 shows the unit root test results, where d shows the first difference of the variable. According to the test results, here we must note that the original variable sequences in (np margin), in (GDP-USD (T.)), and in (Stock Price-USD) are non-stationary (i.e., the levels are non-stationary), and their p-values are all much greater than the 0.05 critical value, which cannot be rejected. After the differencing of the indicators and comparing the sequence of variables of the corresponding ones at the 5% significance level, it was noted that all the P values were less than 0.05, and upon passing the stationary test, we accepted
them.

Null Hypothesis: D(LNTURNOVER RATE) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -6.474323 | 0.0000 |
| Test critical values: | 1% level | -3.661661 | |
| | 5% level | -2.960411 | |
| | 10% level | -2.619160 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNNPMARGIN) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -6.780305 | 0.0000 |
| Test critical values: | 1% level | -3.661661 | |
| | 5% level | -2.960411 | |
| | 10% level | -2.619160 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNGDP_TRILLION) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -6.345588 | 0.0001 |
| Test critical values: | 1% level | -4.284580 | |
| | 5% level | -3.562882 | |
| | 10% level | -3.215267 | |

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNSTOCK_PRICE) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -6.416630 | 0.0000 |
| Test critical values: | 1% level | -4.284580 | |
| | 5% level | -3.562882 | |
| | 10% level | -3.215267 | |

*MacKinnon (1996) one-sided p-values.

Figure 4.7 Data processed by first-order difference

| Variable | ADF test value | 5% critical value | P-value | Test results |
|---|---|---|---|---|
| dlnTurnover rate | -6.474323 | -2.960411 | 0.0000 | Stable |
| dlnnpMargin | -6.780305 | -2.960411 | 0.0000 | Stable |
| dlnGDP(Trillion) | -6.345588 | -3.562882 | 0.0001 | Stable |
| dlnStock_Price | -6.416630 | -3.562882 | 0.0000 | Stable |

**Step 3: Determine the lag order of the model**

Before conducting a cointegration test, it is necessary to reasonably determine the lag order of the model to avoid problems such as too little freedom or autocorrelation caused by a lag order that is too large or too small.

VAR Lag Order Selection Criteria
Endogenous variables: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LN...
Exogenous variables: C
Date: 01/02/25 Time: 18:19
Sample: 2011Q2 2023Q2
Included observations: 31

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -63.09299 | NA | 0.000891 | 4.328580 | 4.513610 | 4.388895 |
| 1 | 26.47936 | 150.2504* | 7.83e-06* | -0.418023* | 0.507130* | -0.116446* |
| 2 | 38.62415 | 17.23777 | 1.06e-05 | -0.169300 | 1.495976 | 0.373538 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Figure 4.8 Lag order

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -63.09299 | NA | 0.000891 | 4.328580 | 4.513610 | 4.388895 |
| 1 | 26.47936 | 150.2504* | 7.83e-06* | -0.418023* | 0.507130* | -0.116446* |
| 2 | 38.62415 | 17.23777 | 1.06e-05 | -0.169300 | 1.495976 | 0.373538 |

In this section, we establish a Vector Auto-regression (VAR) model to analyze factors that contribute to stock price changes. The AIC/SC lines with the least amount of * on the graph above indicate the model that has the least amount of lag order. In a nutshell, the calculated lag orders for LR, FPE, AIC, SC, and HQ in the Lag Order are 1. A VAR (1) model still needs to be put in place.

**Step 4: Cointegration test**

The unit root stationary test mainly assumes that the processes of these variables are all first order, and the integrated stationary sequence. The Johansen Analysis can be employed to test whether cointegration is possible, indicating the possibility that the variable sequences share a long-term relationship.

Date: 01/02/25 Time: 18:19
Sample (adjusted): 2011Q3 2023Q2
Included observations: 32 after adjustments
Trend assumption: Linear deterministic trend
Series: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LNTURNOVER...
Lags interval (in first differences): No lags

Unrestricted Cointegration Rank Test (Trace)

| Hypothesized No. of CE(s) | Eigenvalue | Trace Statistic | 0.05 Critical Value | Prob.** |
|---|---|---|---|---|
| None * | 0.637007 | 54.03092 | 47.85613 | 0.0118 |
| At most 1 | 0.348164 | 21.60303 | 29.79707 | 0.3211 |
| At most 2 | 0.161723 | 7.908236 | 15.49471 | 0.4753 |
| At most 3 | 0.068283 | 2.263233 | 3.841466 | 0.1325 |

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Figure 4.8 Cointegration Test

| hypothesis | Eigenvalue | Trace Statistics | 5% critical value | P-value |
|---|---|---|---|---|
| None* | 0.637007 | 54.03092 | 47.85613 | 0.0118 |
| At most 1 | 0.348164 | 21.60303 | 29.79707 | 0.3211 |
| At most 2 | 0.161723 | 7.908236 | 15.49471 | 0.4753 |
| At most 3 | 0.068283 | 2.263233 | 3.841466 | 0.1325 |

The above table is the result of cointegration test. The trace statistic test shows that the P value rejects the null hypothesis of no cointegration relationship at the 5% significance level, which means that there is a cointegration relationship. Therefore, there is a long-term equilibrium relationship between the time series. The original sequence lnTurnover rate, lnnpMargin, lnGDP(Trillion), and lnStock_Price can be used to build a model.

**Step 5: AR characteristic root test**

The AR eigenvalue test is a commonly used method to check whether the VAR model is stable (the model is referred to as Video no EF No EF as stability check). The issue at hand while determining the model stability is to consider the inversion of the module of the AR unit root, making it 1. When the underlying AR unit roots are all less than one, that is to say, that they are all positioned within the unit circle area, which looks like this graphically.

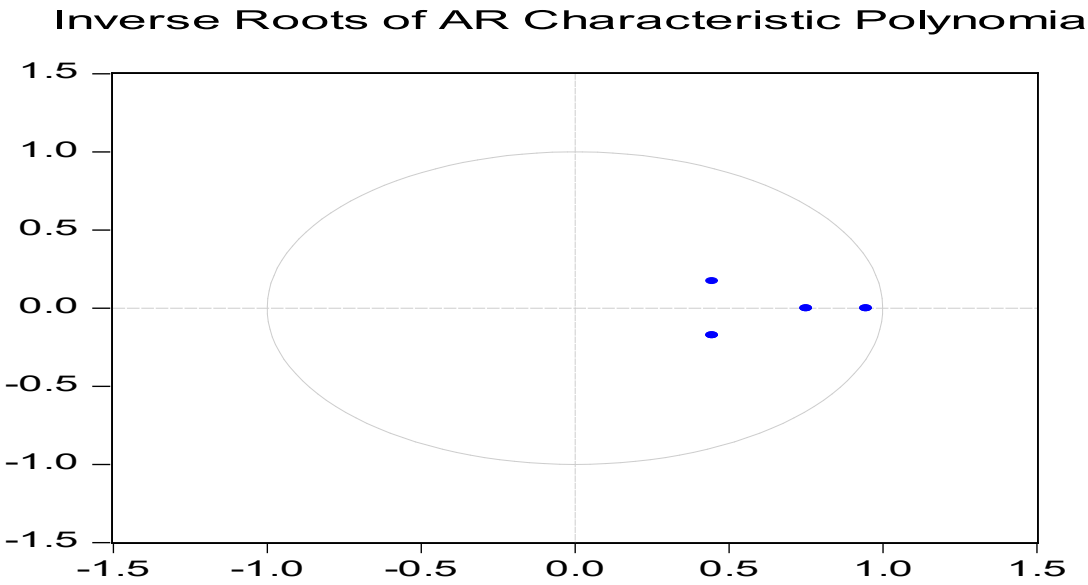



Figure 4.9 The AR eigenvalue test

The figure above, in some respects, demonstrates the AR characteristic roots test result, and we can recognize that the test points also make up a circle area terrace. Having obtained that result, the VAR models also, hence, became confirmed, as well. Additionally, it also typifies the fact that a long-run and theoretically stable link exists between the variables, which is obviously an important precondition for the VAR analysis.

**Step 6: Granger causality test**

Under the Granger causality test, if we test that the variances of Turnover and Net Profit Margin to Stock Price demonstrates greater than 0 percent and their hypotheses are to be rejected, and in traditional probability, the p-value test for Turnover and Net Profit Margin, as a read, is 0.9362 and 0.1088, thus, there are no greater than 0.05, which means the goal not to accept those two null hypotheses at 95% significance level.

Pairwise Granger Causality Tests
Date: 01/02/25 Time: 18:21
Sample: 2011Q2 2023Q2
Lags: 1

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| LNGDP TRILLION does not Granger Cause LNSTOCK PRICE | 32 | 7.65775 | 0.0097 |
| LNSTOCK PRICE does not Granger Cause LNGDP TRILLION | | 0.04856 | 0.8271 |
| LNNPMARGIN does not Granger Cause LNSTOCK PRICE | 32 | 2.73783 | 0.1088 |
| LNSTOCK PRICE does not Granger Cause LNNPMARGIN | | 1.11490 | 0.2997 |
| LNTURNOVER RATE does not Granger Cause LNSTOCK PRICE | 32 | 0.00651 | 0.9362 |
| LNSTOCK PRICE does not Granger Cause LNTURNOVER RATE | | 0.00904 | 0.9249 |

Figure 4.10 The cointegration relationship test

| Null hypothesis | Statistics | P-value |
|---|---|---|
| Turnover rate is not the Granger cause of stock price | 0.00651 | 0.9362 |
| The company's net profit margin is not the Granger cause of the stock price | 2.73783 | 0.1088 |
| China's GDP is not a Granger cause for stock prices | 7.65775 | 0.0097 |

The Granger cause test result shows that the p-value is more than 0.05, which means that the null hypothesis of China's GDP is not the Granger cause of stock price is possible to be rejected. In the study Granger causality aspect, the fluctuation of China's GDP has largely influenced the stock price, however, the Granger causality test is a realization of the statistical time series, which doesn't signify that the actual material cause exists. The relationship ought to be the subject of further research with regard to the variance decomposition and pulse response of the VAR model.

**Step 7: Impulse response**

The first method of impulse response analysis, the comparison method, seems to be more intuitive. Via the impulse function plot, we would be able the effect on the return on investment due to the differing condition. The figures below show the impulse response plots in the case of China's GDP, corporate net profit margin, turnover rate, and stock price variability, with the time difference of the stimulus situation plotted on the left by the letter Y, the intensity of the impact on the variable on the right by the up and down designation on the axis Y, the continuum impression by the continuous line, and the border impression highlighted by the fine line at a 95% confidence limit.

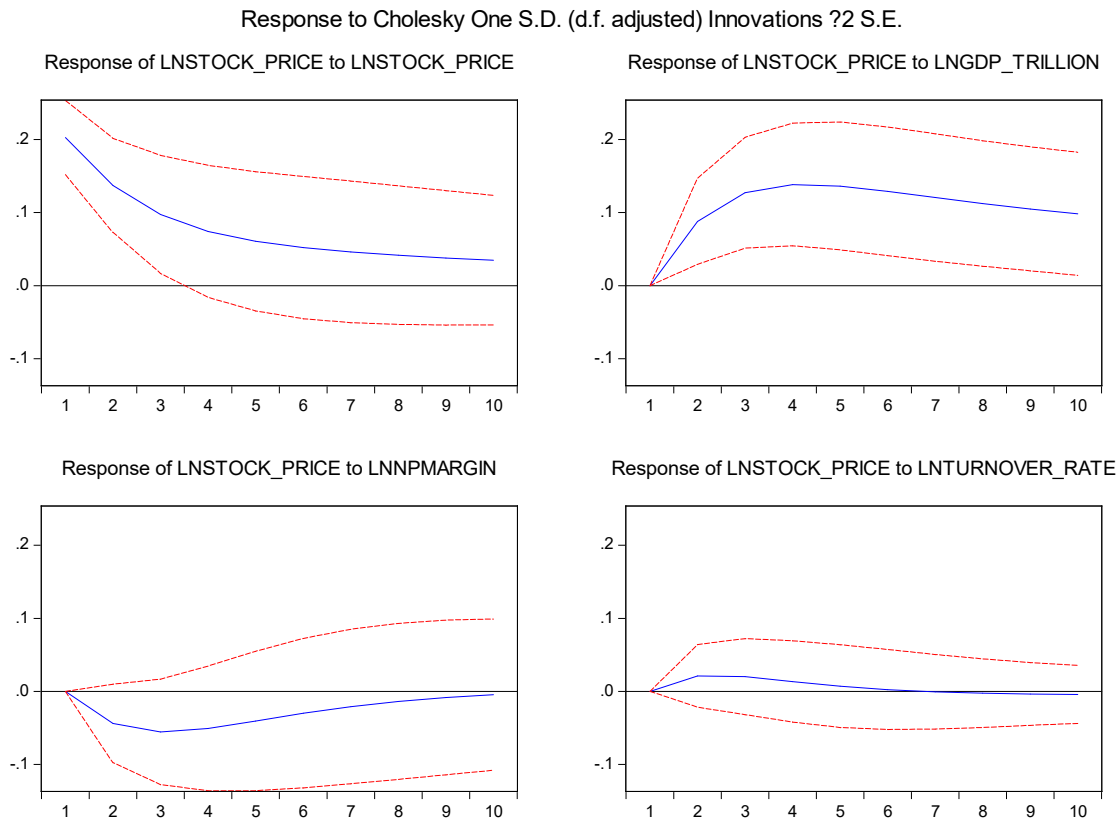



Figure 4.11 Pulse diagram

From the impulse response diagram in the figure above, it can be seen that when a positive shock is given to the stock price, China's GDP will have a greater impact on the fluctuation of the stock price, and it is a positive shock from the first to the tenth period, reaching the maximum value of the positive shock in the fourth period, and the impact gradually converges in the long-term development. This shows that China's economic development level will greatly affect the fluctuation of the stock price, and will have a positive impact on the stock price. The company's net profit margin will also have a certain impact on the fluctuation of the stock price, but the impact from the first to the tenth period is negative, reaching the maximum value of the negative shock in the third period, and the impact between the seventh and tenth periods gradually approaches the horizontal axis. This means that the company's net profit margin will have a negative impact on the fluctuation of the stock price. The turnover rate will have a relatively small impact on the fluctuation of the stock price, and will have a positive impact from the first to the fifth period, reaching the maximum value of the positive shock in the second period, and the

impact from the sixth to the tenth period is not obvious. This shows that the public's sentiment towards this stock will also have a certain degree of positive impact on the fluctuation of the stock price, and the rise of the public's sentiment towards the stock will drive the stock price to rise.

**Step 8: Variance Decomposition**

In order to further analyze the impact relationship between China's GDP, corporate net profit margin, turnover rate and stock price fluctuations, a model was established using the variance decomposition method to analyze the contribution of the impact.

| Period | S.E. | LNSTOC... | LNGDP ... | LNNPMA... | LNTURN... |
|---|---|---|---|---|---|
| 1 | 0.202681 | 100.0000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | 0.264501 | 85.56577 | 11.04669 | 2.752281 | 0.635261 |
| 3 | 0.314699 | 69.99067 | 24.10792 | 5.042350 | 0.859058 |
| 4 | 0.355563 | 59.17305 | 34.01390 | 5.995267 | 0.817783 |
| 5 | 0.387738 | 52.19708 | 40.94219 | 6.139189 | 0.721547 |
| 6 | 0.413006 | 47.58975 | 45.83528 | 5.935484 | 0.639485 |
| 7 | 0.433171 | 44.39221 | 49.40089 | 5.625380 | 0.581524 |
| 8 | 0.449618 | 42.05672 | 52.08737 | 5.313128 | 0.542782 |
| 9 | 0.463311 | 40.27431 | 54.17263 | 5.035888 | 0.517171 |
| 10 | 0.474904 | 38.86631 | 55.83166 | 4.801850 | 0.500184 |

Cholesky Ordering: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LNTURNOVER RATE

Figure 4.11 Variance Decomposition Method

| Period | S.E. | lnStock_Price | lnGDP(Trillion) | lnnpMargin | lnTurnover rate |
|---|---|---|---|---|---|
| 1 | 0.202681 | 100.0000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | 0.264501 | 85.56577 | 11.04669 | 2.752281 | 0.635261 |
| 3 | 0.314699 | 69.99067 | 24.10792 | 5.042350 | 0.859058 |
| 4 | 0.355563 | 59.17305 | 34.01390 | 5.995267 | 0.817783 |
| 5 | 0.387738 | 52.19708 | 40.94219 | 6.139189 | 0.721547 |
| 6 | 0.413006 | 47.58975 | 45.83528 | 5.935484 | 0.639485 |
| 7 | 0.433171 | 44.39221 | 49.40089 | 5.625380 | 0.581524 |
| 8 | 0.449618 | 42.05672 | 52.08737 | 5.313128 | 0.542782 |
| 9 | 0.463311 | 40.27431 | 54.17263 | 5.035888 | 0.517171 |
| 10 | 0.474904 | 38.86631 | 55.83166 | 4.801850 | 0.500184 |

The variance decomposition of stock price fluctuations shows that in the short term, stock price fluctuations are most affected by their own changes, and in the long term, they are most affected by China's GDP. Specifically, the impact of China's GDP on stock price fluctuations was 0 in the first period, and it increased rapidly to about

11% in the second period. After that, with the continuous rise of China's economic level in the long term, the impact on stock prices has greatly increased, and the contribution has stabilized at about 55% in the long term. The impact of the company's net profit margin on stock price fluctuations has generally increased first and then decreased, and the overall impact has not changed much. The contribution has stabilized at about 5% in the long term. In contrast, the turnover rate contributes less to stock price fluctuations, and the impact does not exceed 1% in both the long term and the short term. Overall, China's GDP, company net profit margin, and turnover rate will have a certain degree of impact on stock price fluctuations, among which China's GDP has the greatest impact on stock price fluctuations.

### 4.1.3 Arima Model

**1.Convert the TIME column to date format.**

```
[3]: import pandas as pd
new_file_path = 'merged_data.csv'
new_df = pd.read_csv(new_file_path)
new_df.head()
```

[3]:

| | TIME | Turnover rate | npMargin | GDP(Trillion) | Stock_Price |
|---|---|---|---|---|---|
| 0 | 2011/6/30 | 87.96 | 0.017500 | 7.55 | 25.45 |
| 1 | 2011/9/30 | 6.84 | 0.014911 | 7.55 | 19.30 |
| 2 | 2012/12/31 | 2.66 | 0.004544 | 8.53 | 20.35 |
| 3 | 2013/9/30 | 1.87 | 0.016119 | 9.57 | 39.36 |
| 4 | 2013/12/31 | 1.22 | 0.014677 | 9.57 | 37.68 |

Figure 4.12 Reading Data

**2.Calculate lnTurnover rate as a new column**

```
•[7]:
new_df['TIME'] = pd.to_datetime(new_df['TIME'])
new_df['lnTurnover rate'] = np.log(new_df['Turnover rate'])
new_df.head()
```

[7]:

| | TIME | Turnover rate | npMargin | GDP(Trillion) | Stock_Price | lnTurnover rate |
|---|---|---|---|---|---|---|
| 0 | 2011-06-30 | 87.96 | 0.017500 | 7.55 | 25.45 | 4.476882 |
| 1 | 2011-09-30 | 6.84 | 0.014911 | 7.55 | 19.30 | 1.922788 |
| 2 | 2012-12-31 | 2.66 | 0.004544 | 8.53 | 20.35 | 0.978326 |
| 3 | 2013-09-30 | 1.87 | 0.016119 | 9.57 | 39.36 | 0.625938 |
| 4 | 2013-12-31 | 1.22 | 0.014677 | 9.57 | 37.68 | 0.198851 |

Figure 4.13 Change Name

**3.Adjust each feature according to the new impact**

Now that we have the lnTurnover rate column, I will adjust each feature according to the new impact. The new impacts are:

GDP(Trillion): 0.55

npMargin: 0.04

lnTurnover rate: 0.005

I will create a new DataFrame with these adjusted features and use it to train the ARIMA model. Next, let's adjust the features and prepare the data.

```
•[9]:
new_impact_degrees = {
    'GDP(Trillion)': 0.55,
    'npMargin': 0.04,
    'lnTurnover rate': 0.005
}

new_adjusted_features_df = new_df[['TIME']].copy()
for feature, impact in new_impact_degrees.items():
    new_adjusted_features_df[f'adj_{feature}'] = new_df[feature] * impact


new_adjusted_features_df.head()
```

[9]:

| | TIME | adj_GDP(Trillion) | adj_npMargin | adj_lnTurnover rate |
|---|---|---|---|---|
| 0 | 2011-06-30 | 4.1525 | 0.000700 | 0.022384 |
| 1 | 2011-09-30 | 4.1525 | 0.000596 | 0.009614 |
| 2 | 2012-12-31 | 4.6915 | 0.000182 | 0.004892 |
| 3 | 2013-09-30 | 5.2635 | 0.000645 | 0.003130 |
| 4 | 2013-12-31 | 5.2635 | 0.000587 | 0.000994 |

Figure 4.14 Add impact factor

**4.Forecast using the ARIMA model**

Stock price prediction using ARIMA model. Before applying the ARIMA model, the parameters (p, d, q) of the ARIMA model need to be determined. This is usually done through ACF (autocorrelation function) and PACF (partial autocorrelation function) plots. Since the adjusted features are being used to predict stock prices, a new time series is first created, which is the weighted sum of the adjusted features, and then the parameters of the ARIMA model will be determined based on this time series.

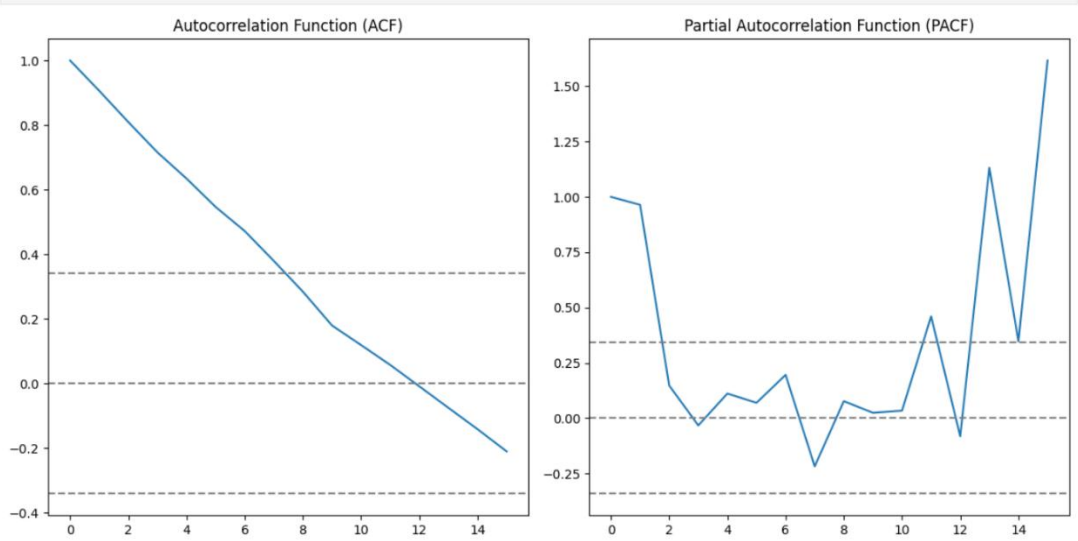



Figure 4.15ACF and PACF

From the ACF plot, you can see that the lag for the first significant crossing of the confidence interval (blue dashed line) is 1, which is probably the p-value in the ARIMA model. From the PACF plot, you see that the lag for the first significant crossing of the confidence interval is also 1, which may be the q value in the ARIMA model. As for the d value, we need to determine the difference order of the time series to make it stationary. Typically, we can check stationarity by observing a time series plot or using the ADF (Augmented Dickey-Fuller) test.

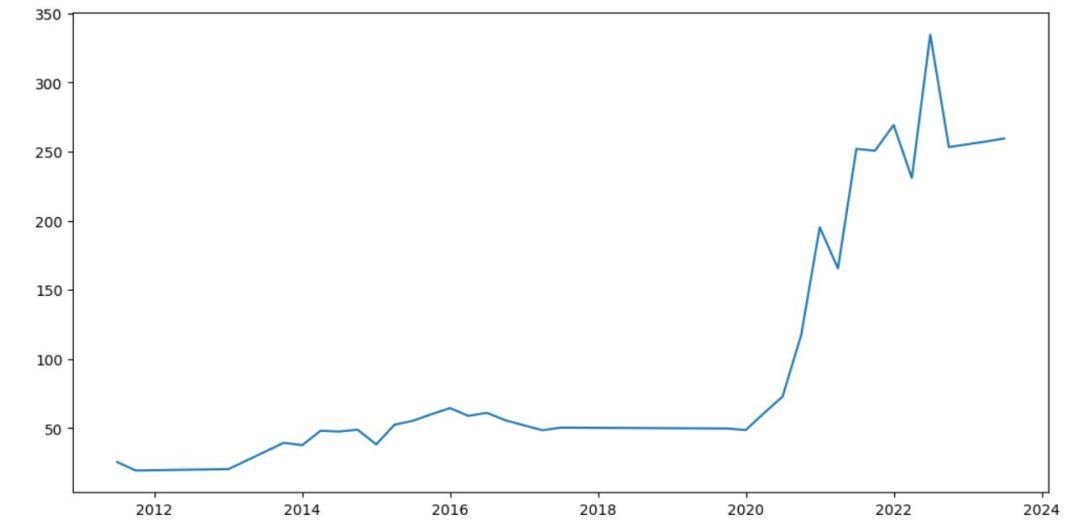



Figure 4.16 ADF

**5.Visualize the forecast results**

Now we have a visualization of the actual and predicted stock prices. As can be seen from the chart, the predicted stock price (red line) tries to follow the trend of the actual stock price (blue line).

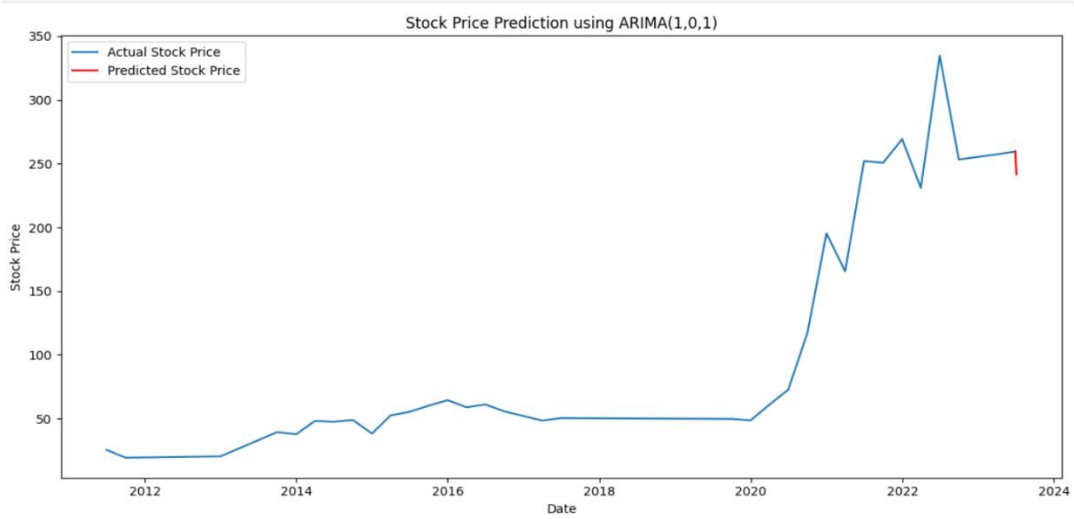



Figure 4.17 Prediction results

# CHAPTER 5 DISCUSSION

## 5.1 Summary

Data preprocessing: First, the researchers merged the three data sets into a complete data set, which contains information such as time series, stock prices, turnover rates, sentiment indicators, China's GDP, and company net profit margins. Subsequently, the data was cleaned, missing values were processed, and unnecessary variables were deleted to lay the foundation for subsequent analysis.

**VAR model:**

a) Stationary test: The researchers used the ADF unit root test to test the stationary nature of the variables and found that the original sequences of stock prices, net profit margins and GDP were non-stationary, while their first-order difference sequences were stationary.
b) Determination of lag order: The lag order of the VAR model was determined to be 1 through the AIC, SC and HQ criteria.
c) Cointegration test: The Johansen cointegration test shows that there is a long-term equilibrium relationship between the variables, and a VAR model can be established.
d) Model stability test: The AR characteristic root test shows that the model is stable.
e) Granger causality test: The Granger causality test shows that China's GDP is the Granger cause of stock prices, while turnover rate and net profit margin are not.
f) Pulse response analysis: The pulse response analysis shows that China's GDP has the greatest impact on stock prices, followed by net profit margins, and turnover rate has the least impact.

g) Variance decomposition analysis: Variance decomposition analysis shows that in the short term, stock price fluctuations are mainly affected by their own changes, while in the long term, China's GDP has the greatest impact.

**ARIMA model:**

a) Data transformation: The turnover rate is converted to logarithmic form and adjusted according to its impact on China's GDP, net profit margin and stock price.
b) Model parameter determination: The parameters (p, d, q) of the ARIMA model are determined through ACF and PACF plots.
c) Model training and prediction: The ARIMA model is trained using historical data and future stock prices are predicted.
d) Result visualization: The actual stock price is compared with the predicted stock price to evaluate the prediction effect of the model.

## 5.2 Future Work

In the future, we can further study the mechanism by which China's GDP affects stock prices, and try to use other time series models for prediction, such as the SARIMA model or the LSTM model, to improve the accuracy of the prediction. In addition, the model can also be applied to other stocks or markets for analysis and prediction, providing investors with more effective decision support.