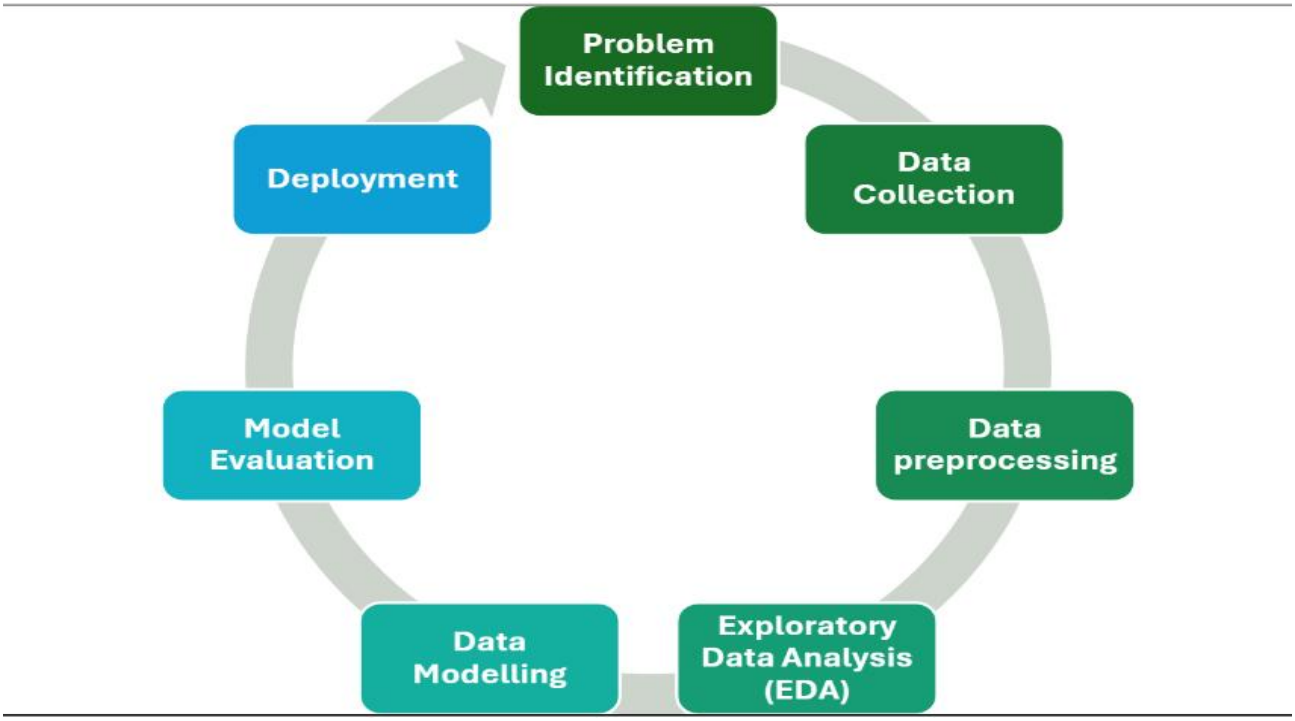
Chapter 3:Research Methodology

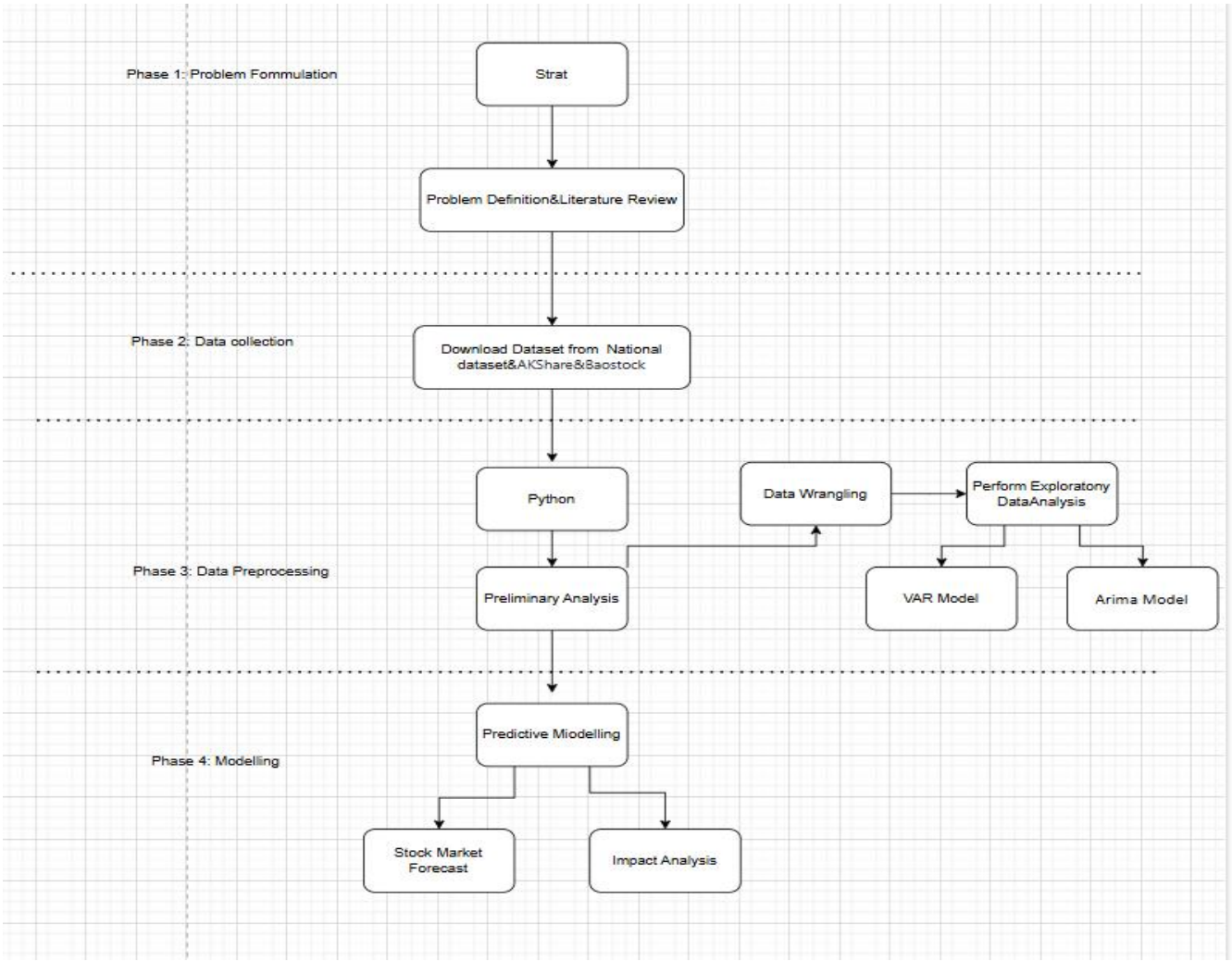
3.0 Data Science Project Life Cycle



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.

3.3Data collection and pre-processing

3.3.1Data 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 pubDate:The date the company releases its earnings report statDate:The last day of the quarter for which the financial report is reported roeAvg:Return on net assets (average) (%) npMargin:Net profit margin (%) gpMargin:Gross profit margin (%) epsTTM:Net profit (RMB) netProfit:Earnings per share MBRevenue:Main operating income (RMB) totalShare:Total share capital liqaShare:Circulating share capital	Get securities data information through Python API and use Baostock to get stock data

Stock Market Dataset	Time:date Code:Stock Code Opening:Starting Price Close:Last Price Highest:Highest Price Lowest:Lowest Price Volume:Transaction Quantity Transaction amount:Transaction amount Amplitude:Highest and Lowest Difference Rise and fall:Percentage increase Change:Changing the amount of money Turnover rate:Percentage of Buying and Selling	AKShare is a Python financial data interface library suitable for various financial data acquisition and processing needs.

3.3.2Data pre-processing

(1) Data cleaning: 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) Data transformation: 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.

3.5Forecasting 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.

3.6Model 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 and unseen 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.