Dean&Francis
ISSN 2959-6130

Application of Linear Regression in the Stock Market in Machine Learning

Yunwei Duan

Abstract:

This paper explores the application of linear regression (a machine learning technique) in stock market analysis,focusing on Apple (AAPL) stock and the S&P 500 Index (SPY),aiming to predict AAPL's stock returns and assess the model's value for investment decisions. It includes data processing,training the linear regression model (with SPY as the independent variable and AAPL as the dependent variable),model evaluation via cross-validation,and hedging return calculation.Key results show AAPL underperformed SPY during the test period,the model effectively shows their correlation,and hedging returns help assess investment outcomes.The study confirms linear regression's feasibility in stock analysis,notes its limitation in handling nonlinear relationships,and suggests using more complex algorithms like neural networks for better accuracy in the future.

Keywords: machine learning; linear regression; stock market

1. Project Framework and Introduction

1.1 Definition of Machine Learning

Machine learning is a rather crucial branch within artificial intelligence. It mainly focuses on how to enhance the performance of computers. To improve computer performance, it is achieved by enabling computers to learn from data rather than directly relying on manual programming. To elaborate more, Algorithms in machine learning can automatically identify patterns, rules, and some valuable insights that may exist in large datasets. Afterwards, the algorithm will utilize this learned knowledge to predict or make judgments on new and untested data. This process is actually imitating human learning. Human problem-solving abilities also gradually improve by constantly summarizing and refining rules from past experiences.

In the field of stock market analysis, machine learning has proved to be a very useful tool. It can screen complex and volatile stock data, extract useful information from it, and this extracted information can serve as a basis to help people make wiser investment decisions.

1.2 Workflow of Machine Learning

The workflow of machine learning usually includes the following key steps:

Data collection:This actually represents a fundamental stage of machine learning. At this rather critical juncture, it is essential to collect a large amount of data related to the issue under investigation and

ISSN 2959-6130

research. Take this special project as an example. Data collection includes data such as the adjusted closing price of Apple's stock and the adjusted closing price of the S&P 500 Index.

Data preprocessing:It refers to a series of operations such as cleaning, transforming, and merging the collected data. The aim of these operations is to remove the noise existing in the data, handle the missing and abnormal data points, and make the data meet the standards for model training. In this project, we will process the data related to extreme values in the dataset. And data related to holidays.

Model training. At the very beginning, we will select a suitable machine learning algorithm. Then, we will use the preprocessed training data to train the model. During this training process, we will make some fine-tuning to the model's parameters to enable the model to approximate the data pattern as accurately as possible. In this specific project, we choose the linear regression algorithm to train the training dataset and determine the parameters of the model.

Model evaluation and improvement: We will use test data and examine its performance indicators, such as error rates, to evaluate the performance of the trained model. If the model fails to provide satisfactory results, then some adjustments and improvements need to be made to the model. Such adjustments and improvements may involve replacing the existing algorithm with a more suitable one or adjusting the model's parameters Or perhaps expand the dataset. In the case of this project, we used the cross-validation method. With these approaches, we evaluated the model and made corresponding improvements.

In terms of projection, we use an improved model to generate projections on brand-new and previously unobserved data, and then obtain the projection results. In an environment like the stock market, to achieve this projection, it is necessary to rely on a trained linear regression model to predict the return rate of stocks.

This article will carry out model construction. All the data used throughout the process, the calculation methods adopted, and the involved variables will be arranged in order to create a relatively comprehensive model. This pattern will be designed to make it more convenient to utilize and more smoothly spread in the future.

1.3 Applications of Machine Learning

Machine learning has a wide range of applications in many different fields. In the financial sector, it can not only analyze the stock market but also be used for credit assessment, identifying fraud, and mitigating various risks, etc. In the medical field, machine learning can help diagnose diseases. It can also be involved in the research and development of drugs. In the field of transportation, machine learning can play a certain role in promoting the development of autonomous driving technology.

In the stock market, machine learning can predict the trend of stock prices by reviewing historical stock data, and at the same time, it can assess the risks existing in stocks, providing certain basic conditions for investors to make decisions.

This rather special project actually represents the specific implementation process of machine learning in stock market analysis.

2. Methodology and Linear Regression

Linear regression is a statistical modeling method with a long history. Its main purpose is to describe the linear relationship between the independent variable (X) and the dependent variable (Y). The basic concept behind linear regression is to formulate a linear equation that can accurately match the data situation. In this way, it is possible to predict the dependent variable. The connection between the independent variable and the dependent variable can also be made clear.

In linear regression, the basic model assumption can be expressed as $Y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + ...$ In this expression, Y is the response variable, while X_1, X_2,... X exists as a predictor variable, β0 represents the constant term, β 1, β 2,... β a belongs to the regression coefficient. These coefficients quantify the extent to which each predictor variable affects the response variable. That is to say, relying on these coefficients, we can know to what extent each predictor variable will influence the response variable.

The ε term is called the residual term. The residual term follows a normal probability distribution and is denoted as ε ~ N. This indicates that the mean of the residual terms is equal to zero, its variance remains constant, and there is no correlation among the residual terms. That is to say, the situation of one residual term will not affect the others.

In the field of linear regression learned in high school, the optimal fitting line is determined by the ordinary least squares method. The basic concept behind the ordinary least squares technique is actually to minimize the sum of the differences of squares. To be more precise, in order to determine the set of regression coefficients, that is, β 0, β 1,... , beta $_p$, makes the ∑ (I = 1 to n) epsilon $_i$ squared, namely ∑ (from I = 1 to n) [Y $_i$ - (beta zero + beta if x1 $_{i\ 1}$ +... + beta $_p$ X '$_p$)] squared reached the lowest value. By solving the problem of finding the minimum value, we can obtain the estimated value of the regression coefficient. After obtaining the estimated value of the regression coefficient, we can establish the linear regression equation.

3. Data Acquisition and Processing

3.1 Data Acquisition

This project uses Python programming to collect data related to Apple's stock and the S&P 500 index. Our focus is on the adjusted closing price, as it has taken into account factors such as stock splits and dividend payments, and it can more accurately reflect the actual value of the stock. After a preliminary analysis of the obtained raw data, it was found that these data contain specific information on relevant prices starting from January 1, 2021. The specific situation is as follows:

Date	AAPL_Adj Close	SPY_Adj Close
2021-01-01	150.496714	398.931725
2021-01-04	150.358450	399.235883
2021-01-05	151.006138	399.724351
2021-01-06	152.529168	400.172184
2021-01-07	152.295015	401.036808

3.2 Data Splitting

This article will divide the collected data into training datasets and test datasets. To elaborate, December 30, 2022 was the last trading day before the test began. This article will use this date as a dividing point. The data before this date will be used as the training dataset in this article, and the data after this date This article will take it as the test dataset. The following is the specific situation of the test data:

Date	AAPL_Adj Close	SPY_Adj Close
2022-12-30	149.999415	446.742943
2023-01-02	150.542775	446.602975
2023-01-03	149.880151	445.998379
2023-01-04	150.450750	446.427587
2023-01-05	149.687491	445.708813

3.3 Data Processing

In this dataset, we need to handle outliers and data related to holidays. Outliers may have a negative impact on the training and prediction of the model. We will use specific statistical techniques to accurately identify and manage these outliers to ensure the reliability of the data, because the stock market does not have trading activities during holidays. We have to eliminate the holiday data. This step is crucial for ensuring that the dataset only contains information about trading days. After processing, the date span of the filtered dataset has been expanded from December 31, 2021 to December 29, 2023. The following is a preview of this dataset:

Date	AAPL_Adj Close	SPY_Adj Close
2021-12-31	151.053034	431.589158
2022-01-03	150.993509	432.452103
2022-01-04	147.752241	432.421296
2022-01-05	146.727854	432.283194
2022-01-06	146.475286	432.990122

In addition to the content mentioned earlier, this article also calculated the return values of the test data. By calculating the returns, the situation of stock price fluctuations can be reflected more accurately. The following is the return situation of the test data:

Date	AAPL_Adj Close Return	SPY_Adj Close Return
2023-01-02	0.003622	-0.000313
2023-01-03	-0.004402	-0.001354
2023-01-04	0.003807	0.000962

ISSN 2959-6130

Date	AAPL_Adj Close Return	SPY_Adj Close Return
2023-01-05	-0.005073	-0.001610
2023-01-06	-0.012058	0.000051

4. Model Training and Evaluation

4.1 Training the Linear Regression Model

This paper selects the relevant data of the S&P 500 index as the independent variable (X), and takes the corresponding data of Apple stock as the dependent variable (Y). These data are used to train the linear regression model. During the training of this linear regression model, this paper will calculate the model parameters β 0, which is the intercept term, and β 1, which is the regression coefficient. These calculated parameters can reflect the extent to which the S&P 500 Index affects the return of Apple's stock. In this paper, by integrating the returns of the test data, the derived linear equation can be used to predict the return of Apple's stock. For example, If the return rate of the S&P 500 index on January 2, 2023 is -0.000313, then this article can estimate what kind of return Apple's stock is likely to have on this day.

4.2 Cross-Validation Evaluation

This project aims to mitigate the potential uncertainties that may arise from a single training dataset and enhance the reliability of the model. It adopts a cross-validation method to evaluate the model. This method operates by dividing the training dataset into several subsets, and then iteratively designates one of these subsets as the validation dataset. The remaining subsets are used as training datasets to carry out the training and evaluation of the model. Finally, the average value of multiple evaluation results will be adopted as the performance index of the model.

The testing period of this instance was from December 30, 2022 to December 29, 2023, which included 261 trading days of data. By using the cross-validation method, we can have a more comprehensive understanding of the model's performance in different subsets of these 261 trading days of data. In this way, we can measure the stability and generalization ability of the model.

5. Hedge Returns and Analysis

The hedging yield refers to the specific difference between the actual yield and the predicted yield. This difference can reflect the final returns generated by the hedging activities carried out based on model predictions. In this project, we obtained the following detailed information by using the data from the test period:

Performance of Apple Stock (AAPL)
Starting Price: 150.00
Ending Price: 142.52
Total Return: -4.99%
Daily Average Return: -0.0172%
Return Volatility: 0.0071
Maximum Daily Gain: 1.97%
Maximum Daily Loss: -1.85%
Performance of the S&P 500 Index (SPY)
Starting Price: 446.74
Ending Price: 464.23
Total Return: 3.92%
Daily Average Return: 0.0149%
Return Volatility: 0.0018
Maximum Daily Gain: 0.46%
Maximum Daily Loss: -0.49%
Relative Performance
Apple stock (AAPL) underperformed the S&P 500 Index (SPY) by 8.90%.

We aim to assess the scenarios that may arise during market fluctuations by calculating hedging returns. For instance, on January 2, 2023, the actual increase of Apple's stock was 0.003622. By integrating this actual increase with the expected return, we can determine the hedging return of Apple's stock on a specific date. By doing these calculations, we can examine what kind of difference exists between the actual return of Apple's stock and the return predicted by the linear regression model when the overall market, represented by the S&P 500 index, experiences fluctuations.

If the hedging return is calculated to be positive, it indicates that the hedging behavior based on the model prediction has produced a relatively positive result. Conversely, if the hedging return is negative when calculated, it indicates that this hedging behavior has led to an unfavorable outcome. This indicator of hedging return can help investors assess the role of the model in managing investment risks and achieving investment returns, and also provide some reference content for investors to formulate investment strategies.

6. Conclusions

6.1 Summary of the Machine Learning Workflow

This measure fully implemented the operation process of

machine learning. All of this started with data collection. We collected data such as the adjusted closing prices of Apple stocks and the S&P 500 index. After collecting the data, we carried out data preprocessing work, which included data segmentation. Handle those abnormal data points and manage data related to holidays. Next, we use linear regression technology to train the model to determine its parameters. Then, we evaluate the model and improve it by means of cross-validation calculation. Finally, we use this model to make predictions and calculate the hedging returns. The entire operation process is very clear and well-organized, which provides a feasible approach for analyzing the stock market.

6.2 Connection with Financial Data

The connection between machine learning and financial data is very close. Financial data has many characteristics, such as a very large volume, a very fast dynamic change speed, and complex factors influencing it. Machine learning has the ability to effectively process these data and can discover potential patterns within them. This can be seen from this project. Linear regression models can establish the correlation between stock returns and market indices, which provides us with a brand-new perspective to understand the fluctuations of stock prices.

6.3 Application Scenarios

Linear regression has found a wide variety of different application scenarios in the stock market. In this specific project, linear regression can not only be used to predict stock returns and calculate hedging returns, but also play a role in areas such as stock risk assessment and portfolio optimization. By analyzing the linear relationship between various stocks and market indices, Then we can assess the systemic risk of stocks. Doing so later will enable us to build a more reasonable investment portfolio.

6.4 Insights and Experiences

Through the implementation of this project, we have truly experienced the significant impact that machine learning has brought to the financial field. From processing data to cultivating models and evaluating results, each stage requires rigorous methods and solid professional knowledge as support. We have gradually begun to understand that the financial market is very complex. A single linear regression model may not be able to explain all the fluctuations in the market clearly. We need to integrate multiple techniques and metrics to conduct a comprehensive assessment.

6.5 Model Limitations and Future Outlook

The linear regression model used in this project has specific constraints. It is based on the premise that there is a linear correlation between the independent variable and the dependent variable. However, in the real stock market, many factors can affect stock returns, and the relationship between these factors influencing stock returns and stock returns may not be a simple linear one. Instead, there exists a nonlinear relationship. For models like linear regression, it is very difficult to describe such complex relationships with particular precision.

In the future, learning approaches can focus on more complex machine learning algorithms, such as neural networks. Neural networks have a strong nonlinear fitting ability, which enables them to handle more complex financial data and extract more market insights from it. If research is conducted on algorithms like neural networks and they are put into use, This can enhance the accuracy and reliability of stock market predictions and provide a more powerful basis for investment decisions.

Besides, it is crucial to deepen the understanding of financial market theories and real-world practices, and to more effectively integrate machine learning techniques with financial expertise to address the complex challenges brought about by financial markets.

References

1.Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

2.Zhihua Zhou. (2016). Machine learning.Tsinghua University Press .

3.Yahoo Finance. (2023). Apple Inc. (AAPL) adjusted closing prices [Data set].From https://finance.yahoo.com/quote/AAPL/history/

4.Yahoo Finance. (2023). S&P 500 ETF (SPY) adjusted closing prices [Data set].From https://finance.yahoo.com/quote/SPY/history/