International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Utilizing Supervised Machine Learning for Stock Market Analysis

1.T.Venkata Sai Hareesh, 2.P.Hemanth Kumar, 3.R.Sai Pavan, 4.M.Pavan Kumar

5.Gudiwaka Vijayalakshmi, Assistant Professor

Computer Science Engineering

Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

Abstract:

The abstract describes an attempt to apply machine learning algorithms to predict future stock prices in the stock market. The paper aims to simplify the complexity of the stock market, which involves various entities such as small ownerships, brokerage corporations, and the banking sector. The motivation for using machine learning is to enhance predictability in this intricate business model.

Key points in the abstract:

Complexity of Stock Market: The stock market is described as one of the most complicated and sophisticated ways of doing business. It involves small ownerships, brokerage corporations, and the banking sector, all relying on the stock market for revenue generation and risk management.

Objective: The primary goal of the paper is to use machine learning algorithms to predict future stock prices. The intention is to leverage open source libraries and existing algorithms to make the unpredictable nature of the stock market more manageable.

Implementation Approach: The paper proposes a simple implementation using machine learning algorithms. It mentions using open source libraries and pre-existing algorithms for the implementation.

Results Expectation: The expected outcome is described as bringing acceptable results. It implies that the machine learning approach has the potential to provide meaningful predictions in the stock market.

Limitations: The abstract acknowledges that the outcome is based on numbers and assumes several axioms. It also notes that these axioms may or may not follow in the real world, indicating an awareness of the limitations and uncertainties associated with predicting stock prices.

Time Dependency: The abstract mentions that the outcome is time-dependent, suggesting that the effectiveness of the predictions may vary over time. It's important to note that the success of such machine learning models in predicting stock prices depends on various factors, including the quality of data, feature selection, model complexity, and the dynamic nature of financial markets. The paper's implementation and results will provide insights into the feasibility and effectiveness of using machine learning in this context.

The introduction of the paper provides a comprehensive overview of the stock market and introduces the concept of using machine learning to predict stock prices. Here are the key points highlighted in the introduction:

Overview of Stock Market:

Describes the stock market as one of the oldest methods for individuals to trade stocks, make investments, and earn money by owning a part of publicly listed companies.

Acknowledges the potential of the stock market as an investment scheme if approached wisely.

Unpredictability of Stock Prices:

Highlights the highly unpredictable nature of stock prices and liquidity in the stock market.

Emphasizes the role of technology, particularly machine learning, in addressing the challenges posed by the unpredictability of the stock market.

Role of Machine Learning:

Presents machine learning as a tool that can assist in understanding and predicting stock market trends.

Indicates the objective of using machine learning algorithms to make the stock market more predictable.

Basic Principles of Stock Market:

Explains the fundamental principles of the stock market, including companies listing their shares as stocks, raising money through initial public offerings (IPOs), and trading on stock exchanges like BSE (Bombay Stock Exchange).

Describes the continuous buying and selling of shares, influencing the price of shares based on market transactions.

Challenges in Predicting Stock Prices:

Recognizes the difficulty in predicting stock exchange prices over time.

Mentions the human brain's ability to extend graphs based on visual interpretation and introduces the concept of crowd computing for predicting outcomes.

Crowd Computing vs. Machine Learning:

Suggests that crowd computing, involving a group of people making predictions, is effective but slow.

Advocates for the use of computers and machine learning to simulate crowd computing with a more scientific and mathematical approach.

Introduction to Linear Regression and Machine Learning:

Introduces the concept of linear regression in statistics and its adaptation in machine learning.

Highlights the importance of selecting the right features and having sufficient data to train a machine learning classifier for accurate predictions.

Preparation for Program Development:

Concludes by expressing readiness to devise a program, combining knowledge of the stock market, graphs, data analysis, and machine learning.

Overall, the introduction sets the stage for the paper, addressing the challenges of stock market predictability and proposing a machine learning approach to enhance understanding and forecasting in this dynamic financial domain.

A. Data Analysis Stage:

In this section, the paper outlines the initial stage of the prediction model, focusing on data analysis. The key steps in this stage include:

Raw Data Examination:

The authors plan to analyse the raw data to identify relevant attributes for predicting the selected label. This involves a thorough examination of the available data to understand its structure and characteristics.

Data Source:

The dataset for the program is sourced from www.quandl.com, described as a premier dataset providing platform. Specifically, the dataset used pertains to GOOGL by WIKI and can be extracted from quandl using the token “WIKI/GOOGL.”

Time Period:

Approximately 14 years of data have been extracted and utilized for the program. The time span covered by the dataset is not explicitly mentioned.

Attributes of the Dataset:

The key attributes of the dataset include:

Open: Opening price of the stock.

High: Highest price possible at an instance of time.

Low: Lowest price possible at an instance of time.

Close: Closing price of the stock.

Volume: Total times traded during a day.

These attributes represent essential financial metrics that are commonly used in stock market analysis.

The absence of specific details regarding the time span covered by the dataset and the exact methods of raw data examination leaves room for further elaboration in the subsequent sections of the paper. The clarity and thoroughness of the data analysis stage will significantly impact the success of the subsequent prediction model.

Data Analysis Stage (Continued):

In this continuation of the data analysis stage, the authors specify the selection of attributes for the prediction model and define the set of features that will be used for the classifier. Key points include:

Selection of Label (Dependent Variable):

The attribute "Close" is chosen as the label, representing the variable that the prediction model aims to predict. This is a common choice in stock market analysis, as the closing price is often used as a key indicator.

Feature Selection:

Features to be used for prediction are selected from adjusted values, specifically "Adj. Open, Adj. High, Adj. Close, Adj. Low, and Adj. Volume." These adjusted values are preferred over raw values due to their processed nature, which helps eliminate common data gathering errors.

Graphing Parameters and Feature Definition:

OHLCV graphs (Open, High, Low, Close, Volume) are commonly used in stock analysis. The same graphing parameters are employed to define features for the classifier.

Defined Features:

Adj. Close: Considered important as it influences the market opening price for the next day and contributes to volume expectancy.

HL_PCT: A derived feature, defined as follows:

$$HL\_PCT = \frac{\text{Adj. High} - \text{Adj. Low}}{Adj.Close} \times 100$$

This derived feature provides information about the percentage change between the highest and lowest prices relative to the closing price.

Further details about the set of features are not provided in the excerpt, and it would be valuable to know if additional features or techniques are considered in the analysis.

Use of Percentage Change:

Percentage change is employed to reduce the number of features while retaining essential information. Specifically, the feature

HL_PCT is chosen as it helps formulate the shape of the OHLCV graph.

PCT_change: This is also a derived feature, defined by:

$$PCT\_Change = \frac{Adj.Close - Adj.Open}{Adj.Open} \times 100$$

Derived Feature - PCT_change:

A new derived feature,

PCT change, is introduced. The exact definition is not provided in the excerpt, but it is mentioned that a similar treatment is applied to Open and Close as with High and Low. This treatment involves the use of percentage change and is presumably designed to capture relevant information for the prediction model.

Importance of Open and Close:

Open and Close are highlighted as crucial features in the prediction model. Similar to High and Low, treating Open and Close with percentage change helps reduce the number of redundant features.

Adj. Volume:

Adj. Volume is considered a very important decision parameter. The rationale is that the volume traded has the most direct impact on future stock prices compared to other features. Therefore, Adj. Volume is retained as is, without undergoing additional treatment or transformation.

Importance of Careful Analysis:

The authors stress the critical nature of the data analysis step, emphasizing that any missing information or small errors in deriving useful information could lead to a failed prediction model and an inefficient classifier.

Subject-Specific Features:

Acknowledges that the features extracted are specific to the subject used (in this case, stock market data) and may vary for different subjects. Generalization is considered possible only if data for another subject is collected with the same coherence as the initial subject.

Treatment of Open and Close:

Similar treatment is applied to Open and Close as with High and Low, emphasizing their relevance in the prediction model. The use of percentage change helps reduce redundancy in features.

Importance of Adj. Volume:

Adj. Volume is considered a crucial decision parameter due to its direct impact on future stock prices. The decision is made to use Adj. Volume in its original form without additional transformation.

Caution in Data Analysis:

The authors stress the importance of careful analysis in this crucial step. Any omission of information or small errors in deriving useful information may lead to a failed prediction model and an inefficient classifier.

Subject-Specific Features:

Acknowledges that the extracted features are specific to the subject (stock market data) and highlights that these features will vary from subject to subject. Generalization is deemed possible only if data for another subject is collected with the same coherence as the initial subject.

B. Training and Testing Stage:

Implementation of Machine Learning Model:

In this stage, the features extracted from the data analysis stage will be implemented in a Machine Learning model. The tools mentioned for this implementation include SciPy, Scikit-learn, and Matplotlib libraries in Python.

Training and Testing:

The model will be trained with the features and labels (presumably the "Close" attribute selected earlier) extracted from the data analysis stage. The same data will then be used to test the model.

Data Preprocessing:

- The data is preprocessed, including the following steps:
- Shifting values of the label attribute by the desired percentage for prediction.
- Conversion of the dataframe format to Numpy array format.
- Removal of all NaN (Not a Number) data values before feeding them to the classifier.
- Scaling of the data so that for any value
- standard deviation
- $(X-\text{mean})/\text{standard deviation}$.
- Splitting the data into test and train data based on their type (label and feature).

Choice of Classifier - Linear Regression:

The chosen classifier is Linear Regression from the Scikit-learn package. The simplicity of Linear Regression aligns with the purpose of the model.

Linear Regression Overview:

Linear Regression is highlighted as a commonly used technique for data analysis and forecasting. It predicts relations between variables based on their dependencies on other features.

Supervised Machine Learning:

Supervised machine learning is explained as a method where labeled data is input, pairing features with their labels. The classifier learns patterns to predict labels based on feature combinations.

Testing in Supervised Machine Learning:

The testing process in supervised machine learning involves inputting feature combinations into the trained classifier and cross-checking the output with the actual label. This step determines the accuracy of the classifier.

Accuracy Requirement:

Emphasizes the crucial role of accuracy in a machine learning model. A classifier with an accuracy less than 95% is considered practically useless. Accuracy is defined as the measure of correct predictions over the total predictions.

Understanding Accuracy:

The authors stress the importance of understanding what accuracy means and indicate that the subsequent subtopic will provide insights into increasing accuracy.

C. Results:

Once the model is ready, we use the model to obtain the desired results in any form we want. In our case, we shall be plotting a graph of our results (fig. 1) as per our requirements which we have discussed earlier in this paper.

The key component of every result is the accuracy it delivers. It should be according to our needs and as stated earlier, a model with accuracy less than 95% is practically useless. There are some standard methods to calculate accuracy in machine learning, some are as follows:

- R2 value of the model.
- Adjusted R2 value
- RMSE Value
- Confusion matrix for classification problems.

Fig. 2. Graph showing the exact amounts of predicted values.

III. HELPFUL HINTS

A. Requirements and Specification:

Thorough Understanding:

Know the problem requirements, machine specifications, and throughput specifications thoroughly at the outset.

Background Check: Conduct a background check on the case, gather ample knowledge, and clearly identify the goals of the program.

B. Careful Function Analysis:

Feature Derivation: Be meticulous in deriving features from the data, ensuring they directly relate to the labels.

Function Minimization: Minimize functions subject to requirement constraints for optimization.

C. Implementation:

Model Selection:

Choose a model compatible with the input data; ensure a proper match between the model and data.

Trial and Error: Experiment with different models simultaneously to identify the most effective one.

Efficient Implementation: Implement the model with efficiency, as the implementation step should take the least amount of time.

D. Training & Testing:

Data Consistency: Ensure consistent, coherent, and abundant training data for a stronger and more accurate classifier.

Testing Guidelines: Test data should be at least 20% of the size of the training data; understand the role of testing in assessing classifier accuracy.

E. Optimization:

Continuous Improvement: It's almost impossible to create a versatile classifier in a single attempt, so continuous optimization is crucial.

Standard Methods: Keep in mind standard methods and basic requirements when optimizing the model.

IV. SOME COMMON MISTAKES

Mentioned common mistakes practitioners should avoid:

- **Bad Annotation:** Ensure accurate annotation of training and testing datasets.
- **Algorithm Assumptions:** Have a clear understanding of algorithms' assumptions.
- **Algorithm Parameters:** Understand algorithms' parameters thoroughly.
- **Objective Understanding:** Failure to understand the objective of the model.
- **Data Understanding**: Lack of understanding of the data.
- Avoid Leakage: Prevent features or information leakage.
- Insufficient Data: Ensure enough data to train the classifier.
- Appropriate Use: Avoid using machine learning where it is unnecessary.

V. CONCLUSIONS

Powerful Tool: Machine learning is a powerful tool with broad applications.

Dependency on Data: Machine learning is highly dependent on data, and data analysis is a challenging task.

Evolution into Deep Learning: Machine learning has evolved into deep learning and neural networks, but the core idea remains the same.

Limitations of the Paper: The paper is limited to supervised machine learning and covers only the fundamentals of the complex process.