International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Leveraging Supervised Machine Learning for Analysis of the Stock Market

1. Gudiwaka Vijayalakshmi (Assistant Professor), 2. T. Venkata Sai Hareesh, 3. P. Hemanth Kumar, 4. R. Sai Pavan, 5. M. Pavan Kumar

Computer Science Engineering, Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

---

Abstract:

This abstract outlines an endeavour to employ machine learning algorithms for forecasting future stock prices within the stock market. It endeavours to streamline the intricacies of the stock market, encompassing diverse entities such as small ownership, brokerage firms, and the banking sector. The motivation behind utilizing machine learning is to augment predictability within this complex business ecosystem.

Key points in the abstract:

Complexity of the Stock Market: The stock market is depicted as one of the most intricate and sophisticated arenas of commerce. It encompasses various stakeholders, including small ownerships, brokerage corporations, and the banking sector, all of which rely on it for revenue generation and risk management.

Objective: The primary aim of the paper is to utilize machine learning algorithms for predicting future stock prices. The intention is to harness open-source libraries and existing algorithms to render the unpredictable nature of the stock market more manageable.

Implementation Approach: The paper advocates for a straightforward implementation using machine learning algorithms. It advocates for the utilization of open-source libraries and pre-existing algorithms for this purpose.

Results Expectation: The anticipated outcome is described as yielding acceptable results, suggesting that the machine learning approach holds promise for generating meaningful predictions in the stock market.

Limitations: The abstract acknowledges that the outcome is grounded in numerical data and relies on several axioms. It also notes that these axioms may or may not hold true in real-world scenarios, highlighting an awareness of the limitations and uncertainties associated with stock price prediction.

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

Overview of Stock Market:

The stock market is portrayed as one of the oldest avenues for individuals to trade stocks, invest, and potentially earn profits by owning shares of publicly listed companies. It recognizes the stock market's potential as an investment vehicle if approached prudently.

Unpredictability of Stock Prices:

The highly erratic nature of stock prices and market liquidity is underscored, highlighting the challenges inherent in stock market dynamics. It stresses the pivotal role of technology, particularly machine learning, in mitigating these challenges.

Role of Machine Learning:

Machine learning is depicted as a valuable tool for comprehending and forecasting stock market trends. It articulates the objective of employing machine learning algorithms to enhance the predictability of the stock market.Basic Principles of Stock Market:

Challenges in Predicting Stock Prices:

Acknowledges the inherent challenge of forecasting stock exchange prices accurately over time.

Introduces the human brain's capacity to extrapolate trends visually from graphs and introduces the concept of crowd computing for predictive purposes.

Crowd Computing vs. Machine Learning:

Proposes that while crowd computing, where a collective makes predictions, can be effective, it tends to be slow.

Advocates for leveraging computers and machine learning to emulate crowd computing more systematically and mathematically.

Introduction to Linear Regression and Machine Learning:

Introduces the concept of linear regression from statistics and its application within the realm of machine learning.

Emphasizes the critical role of feature selection and data adequacy in training a machine learning classifier for precise predictions.

Preparation for Program Development:

Concludes by affirming readiness to develop a program, amalgamating expertise in the stock market, graph analysis, data interpretation, and machine learning.

Overall, the introduction lays the groundwork for the paper, tackling the hurdles of stock market predictability and advocating for a machine learning framework to augment comprehension and forecasting within this ever-evolving financial landscape

A. Data Analysis Stage:

This section delineates the initial phase of the prediction model, focusing on data analysis. Key steps in this stage encompass:

Raw Data Examination:

The authors plan to scrutinize the raw data to pinpoint relevant attributes for predicting the chosen label. This involves a meticulous examination of the available data to comprehend its structure and peculiarities.

Data Source:

The dataset for the program is obtained from www.quandl.com, touted as a premier platform for datasets. Specifically, the utilized dataset pertains to GOOGL by WIKI and can be accessed from Quandl using the token “WIKI/GOOGL.”

Time Period:

Approximately 14 years of data have been procured and employed for the program. However, the specific time span covered by the dataset is not explicitly stated.

Attributes of the Dataset:

Key attributes of the dataset encompass:

Open: Opening price of the stock.

High: Maximum price at a given moment.

Low: Minimum price at a given moment.

Close: Closing price of the stock.

Volume: Total trading activity during a day.

These attributes denote fundamental financial metrics commonly utilized in stock market analysis.

The absence of precise details concerning the dataset's time span and the specific methodologies employed for raw data examination leaves room for further elucidation in subsequent sections of the paper. The clarity and comprehensiveness of the data analysis stage will profoundly influence the efficacy of the ensuing prediction model.

Data Analysis Stage (Continued):

In this extension of the data analysis stage, the authors delineate the selection of attributes for the prediction model and specify the set of features utilized for the classifier. Key points include:

Selection of Label (Dependent Variable):

The attribute "Close" is designated as the label, signifying the variable that the prediction model aims to forecast. This choice is customary in stock market analysis, given the significance of the closing price as a primary indicator.

Feature Selection:

Features for prediction are chosen 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 aids in mitigating common data collection errors.

Graphing Parameters and Feature Definition:

OHLCV graphs (Open, High, Low, Close, Volume) are prevalent 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 offers insights into the percentage change between the highest and lowest prices concerning the closing price. However, additional information regarding the complete set of features utilized in the analysis is not provided in the excerpt. It would indeed be valuable to ascertain if other features or techniques are incorporated into the analysis for a more comprehensive understanding of the prediction model's formulation and efficacy.

Use of Percentage Change:

Percentage change serves as a tool to streamline the number of features while preserving critical information. In particular, the feature HL_PCT is selected due to its ability to contribute to shaping the OHLCV graph. This strategic use of percentage change enhances the model's efficiency by focusing on key aspects of market behavior while minimizing feature redundancy.

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 without an exact definition provided in the excerpt. However, it is noted that a similar treatment is applied to Open and Close as with High and Low, involving the use of percentage change. This treatment is likely designed to capture pertinent information for the prediction model.

Importance of Open and Close:

Open and Close are emphasized as pivotal features in the prediction model. Similar to High and Low, treating Open and Close with percentage change helps streamline the number of redundant features while retaining critical market dynamics.

Adj. Volume:

Adj. Volume is deemed a crucial decision parameter. The rationale behind this is that trading volume has a direct impact on future stock prices compared to other features. Therefore, Adj. Volume is retained in its original form without undergoing additional treatment or transformation for the analysis.

Importance of Careful Analysis:

The authors underscore the critical nature of the data analysis phase, emphasizing that even minor errors or missing information in deriving relevant insights could result in a flawed prediction model and an ineffective classifier.

Subject-Specific Features:

Recognizing that the extracted features are tailored to the specific subject matter (in this case, stock market data), the authors note that these features may differ for distinct subjects. They suggest that generalization is only feasible if data for another subject is collected with comparable coherence to the initial subject.

Treatment of Open and Close:

Similar treatment is applied to Open and Close attributes as with High and Low, underscoring their significance in the prediction model. The utilization of percentage change aids in streamlining feature redundancy, optimizing model performance.

Importance of Adj. Volume:

Adj. Volume is identified as a pivotal decision parameter due to its direct influence on future stock prices. As a result, the decision is made to retain Adj. Volume in its original form without additional transformation, recognizing its intrinsic importance in the analysis.

B. Training and Testing Stage:

Implementation of Machine Learning Model:

During this stage, the features derived from the data analysis phase will be incorporated into a Machine Learning model. The tools specified for this implementation include SciPy, Scikit-learn, and Matplotlib libraries in Python.

Training and Testing:

The model will undergo training using the features and labels (presumably the "Close" attribute chosen earlier) extracted from the data analysis phase. Subsequently, the same dataset will be utilized to evaluate the model's performance.

Data Preprocessing:

The data undergoes preprocessing, encompassing 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 inputting them into the classifier.
- Scaling of the data to ensure uniformity across features, typically by computing the (X−mean)/standard deviation.
- Segregation of the data into test and train sets based on their classification (label and feature).

Choice of Classifier - Linear Regression:

The selected classifier is Linear Regression from the Scikit-learn package. The simplicity of Linear Regression aligns well with the model's objectives.

Linear Regression Overview:

Linear Regression is highlighted as a widely used technique for both data analysis and forecasting purposes. It establishes relationships between variables based on their dependencies on other features.

Supervised Machine Learning:

Supervised machine learning is elucidated as a method where labeled data is provided, associating features with their corresponding labels. The classifier learns patterns from this labeled data to predict labels based on feature combinations.

Testing in Supervised Machine Learning:

The testing phase in supervised machine learning involves feeding feature combinations into the trained classifier and verifying the output against the actual label. This crucial step assesses the accuracy and performance of the classifier.

Accuracy Requirement:

The paramount importance of accuracy in a machine learning model is underscored. A classifier with an accuracy below 95% is deemed practically ineffective. Accuracy, defined as the ratio of correct predictions to total predictions, serves as a vital metric for evaluating model performance.

Understanding Accuracy:

The authors emphasize the necessity of comprehending the concept of accuracy thoroughly. They indicate that the following subtopic will delve into strategies for enhancing accuracy, providing valuable insights into this critical aspect of model evaluation.

C. Results:

Once the model is prepared, we leverage it to generate the desired outcomes in the preferred format. In our case, we will be creating a graph of our results (see Fig. 1) in accordance with the requirements discussed earlier in this paper.

The pivotal aspect of every result lies in the accuracy it yields. It should align with our specifications, and as previously stated, a model with accuracy below 95% is considered practically ineffective. Various standard methods exist to calculate accuracy in machine learning, including:

- R2 value of the model.
- Adjusted R2 value.
- RMSE Value.

Confusion matrix for classification problems.

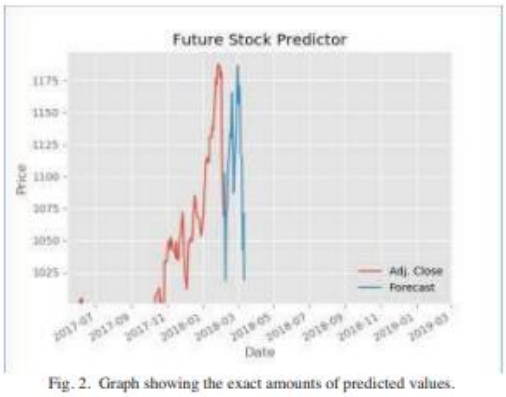

Fig. 2. Graph showing the exact amounts of predicted values.

III. HELPFUL HINTS

A. Requirements and Specification:

Thorough Understanding:

Ensure a comprehensive grasp of the problem requirements, machine specifications, and throughput specifications from the outset.

Background Check:

Conduct thorough research on the case, gather ample knowledge, and clearly define the program's objectives.

B. Careful Function Analysis:

Feature Derivation:

Exercise diligence in deriving features from the data, ensuring they directly correspond to the labels.

Function Minimization:

Minimize functions while adhering to requirement constraints for optimization.

C. Implementation:

Model Selection:

Select a model that aligns with the input data; ensure compatibility between the model and data.

Trial and Error:

Experiment with various models concurrently to determine the most effective one.

Efficient Implementation:

Implement the model efficiently; aim for minimal time consumption during implementation.

D. Training & Testing:

Data Consistency:

Maintain consistency, coherence, and an ample supply of training data for a robust and accurate classifier.

Testing Guidelines:

Allocate test data comprising at least 20% of the training data size; recognize the role of testing in evaluating classifier accuracy.

E. Optimization:

Continuous Improvement:

Acknowledge the necessity for continuous optimization; creating a versatile classifier often requires iterative refinement.

Standard Methods:

Adhere to standard methods and fundamental requirements during the optimization process.

IV. SOME COMMON MISTAKES

Practitioners should be mindful of the following common mistakes and strive to avoid them:

Bad Annotation: Ensure precise and accurate annotation of both training and testing datasets to maintain data integrity.

Algorithm Assumptions: Develop a clear understanding of the assumptions underlying algorithms to prevent erroneous interpretations.

Algorithm Parameters: Thoroughly comprehend the parameters of algorithms to optimize model performance effectively.

Objective Understanding: Failure to grasp the objective of the model can lead to misguided analyses and outcomes; strive for clarity in objectives.

Data Understanding: Lack of comprehension of the data can result in flawed interpretations and modeling; prioritize understanding the data thoroughly.

Avoid Leakage: Take measures to prevent unintended leakage of features or information into the model, which can distort results.

Insufficient Data:

Ensure the availability of adequate data to train the classifier effectively, as insufficient data can compromise model accuracy.

Appropriate Use: Exercise discretion and avoid using machine learning in contexts where it is unnecessary or inappropriate, as it may lead to inefficiencies or inaccuracies.

V. CONCLUSIONS

Powerful Tool:

Machine learning is acknowledged as a potent tool with diverse applications across various domains.

Dependency on Data:

The efficacy of machine learning is heavily reliant on the quality and quantity of data available, making data analysis a formidable undertaking.

Evolution into Deep Learning:

While machine learning has advanced into deep learning and neural networks, the fundamental principle remains unchanged, emphasizing the extraction of patterns from data.

Limitations of the Paper:

The paper is constrained to the realm of supervised machine learning, providing foundational insights into the multifaceted process. It does not delve into more advanced techniques or complex methodologies.

REFERENCES:

[1] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore “A Machine learning approach to Building domain-specific Search engine”, IJCAI, 1999 - Citeseer

[2] Yadav, Sameer. (2017). STOCK MARKET VOLATILITY - A STUDY OF INDIAN STOCK MARKET. Global Journal for Research Analysis. 6. 629-632.

[3] Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[4] Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley. ISBN 0-471-17082-8.

[5] Robert S. Pindyck and Daniel L. Rubinfeld (1998, 4h ed.). Econometric Models and Economic Forecasts

[6] “Linear Regression”, 1997-1998, Yale University http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[7] Agarwal (July 14, 2017). "Introduction to the Stock Market". Intelligent Economist. Retrieved December 18, 2017.

[8] Jason Brownlee, March 2016, “Linear Regression for machine learning”, Machine learning mastery, viewed on December 2018, https://machinelearningmastery.com/linear-regression-for-machine learning

[9] Google Developers, Oct 2018, “Decending into ML: Linear Regression”, Google LLC, https://developers.google.com/machinelearning/crash-course/descending-into-ml/linear-regression

[10] Fiess, N.M. and MacDonald, R., 2002. Towards the fundamentals of technical analysis: analyzing the information content of High, Low, and Close prices. Economic Modelling, 19(3), pp.353-374.

[11] Hurwitz, E. and Marwala, T., 2012. Common mistakes when applying computational intelligence and machine learning to stock market modeling. arXiv preprint arXiv:1208.4429.