IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Stock Price Prediction using Machine Learning

Surbhi Doliya **Priti Sharma**

Computer Science Department
Birla Institute of Technology, Mesra

Abstract

In the realm of market prediction, investors have historically relied on the analysis of stock prices, indicators, and related news to anticipate market movements, underscoring the significance of news in influencing stock prices. Previous studies in this field have largely focused on categorizing market news as positive, negative, or neutral and examining their impact on stock prices, or on analyzing historical price data to forecast future movements. In our research, we present an automated trading system that amalgamates mathematical functions, machine learning techniques, and external factors such as sentiment analysis of news to enhance stock prediction accuracy and facilitate profitable trades. Specifically, our objective is to forecast the price or trend of a given stock by the end of the trading day based on its performance during the initial trading hours. To accomplish this objective, we have trained conventional machine learning algorithms and developed multiple deep learning models, taking into account the significance of relevant news.

1. Introduction

The financial market operates as a multifaceted and dynamic system where individuals engage in the buying and selling of currencies, stocks, equities, and derivatives through virtual platforms facilitated by brokerage firms. Investors participate in the stock market to acquire ownership in public companies via trading conducted either through exchanges or over-the-counter markets. This avenue has provided investors with opportunities for wealth accumulation and financial prosperity by investing modest initial sums of money, which presents lower risk compared to venturing into new business ventures or pursuing high-salary careers. The stock market is subject to various factors that contribute to its uncertainty and high volatility.

While humans are capable of executing orders and submitting them to the market, automated trading systems (ATS), driven by computer programs, often exhibit superior performance and efficiency in order execution. However, the evaluation and management of ATS performance require the implementation of risk management strategies and safety measures informed by human judgment. Developing an ATS involves consideration of numerous factors, including the trading strategy to be adopted, complex mathematical functions reflecting the state of specific stocks, machine learning algorithms enabling the prediction of future stock values, and relevant news pertaining to the stocks under analysis.

Numerous studies have investigated the prediction of stock price trends, primarily within daily timeframes, utilizing models that integrate diverse data sources such as news articles, Twitter data, Google Trends, and Wikipedia data. Integration of these external factors with stock prices and technical indicators has demonstrated their impact on stock price movements. The stock market is recognized as volatile due to the influence of external factors, market dynamics, and the complexity of market dimensionality, posing challenges to the prediction of stock trends/prices even with the application of deep learning models.

External factors affecting the stock market can be categorized into fundamental factors, technical factors, and market sentiments. For instance:

- Supply and demand dynamics play a crucial role; if traders exhibit a preference for buying a particular stock over selling it, this can lead to price increases as demand exceeds supply.

- Stock prices can experience unexpected fluctuations due to a single news event, leading to artificial highs or lows. This unpredictability is often attributed to market sentiment factors, which encompass various elements such as company news, economic indicators, and global events.

- Global Economy: The movement of money and transactions within the stock market is influenced by the overall economic conditions, both domestically and internationally. Changes in economic indicators, such as GDP growth, inflation rates, or interest rates, can significantly impact investor sentiment and consequently affect stock prices.

- Stock Historical Prices: Each stock has its own trading range, within which its price moves based on historical data and investor behavior. Chart patterns and past price movements are often analyzed to anticipate future trends, although they do not guarantee future performance.
- Public Sentiment and Social Media: Public sentiment, as reflected in social media platforms and news outlets, can have a profound effect on stock prices. A single tweet from a notable figure or the release of an influential article can cause sudden shifts in investor perception and drive buying or selling activity. For instance, an unofficial announcement or resignation of a CEO on a platform like Twitter can lead to significant fluctuations in the stock price of the company involved.

The complexity of various factors influences price movements, making stock prediction challenging. Scholars contend that market prediction follows non-random patterns (Schumaker, R. et al. 2009). Numerous studies aim to enhance the precision of future price forecasts. For instance, Mark L. et al. (1994) investigated the impact of public information from Dow Jones and found a clear correlation between news releases and stock market behavior.

2. Related work and background

Early research on stock market prediction introduced the Efficient Market Hypothesis (EMH) by Fama, E. F. (1970) and the Random Walk theory by Horne, J. C., & Parker, G. G. (1967). These theories posited that market prices are influenced by various information beyond historical data, making price prediction impractical. According to EMH, stock prices solely reflect market information, leading to price adjustments in response to new information. It asserts that stocks are always traded at their fair value, making it impossible for traders to profit from undervalued or inflated prices unless they increase their risk. EMH delineates three variations—Weak Form, which considers only historical data; semi-Strong Form, which adds current public data; and Strong Form, which includes private data. It contends that price movements stem from new information or random fluctuations, thwarting prediction models.

Contrastingly, the Random Walk Hypothesis, proposed by Horne, J. C., & Parker, G. G. (1967), suggests that stock prices change randomly, with past movements having no bearing on current ones. This hypothesis, focusing on short-term market patterns, aligns with EMH to some extent but emphasizes the unpredictability of stock prices. Combining these theories implies that stock market movements follow a random path, limiting prediction accuracy to no more than 50%. However, recent studies challenge this notion by demonstrating some predictability in stock prices.

These studies employ two primary methods for stock market prediction:

- Fundamental Analysis: This approach evaluates a company's health using qualitative and quantitative factors such as interest rates, return on assets, and price-to-earnings ratio to assess long-term investment potential.
- Technical Analysis: Utilizing time series data, traders analyze historical price trends and chart patterns, considering time as a critical factor in prediction. Technical analysis relies on observing stock price movements, identifying historical trends, and incorporating relevant information to forecast future price movements.

Recent research incorporates various machine learning techniques for stock price prediction, leveraging their ability to identify generalized patterns in data. These models and risk strategies aim to forecast price direction across different time frames by considering diverse factors influencing market prices.

3. Data Sources

We require two primary sources of information for our study: news sentiments and historical prices. Over a span of ten years, tick data and news data were sourced from the Reuters platform, covering the period from January 01, 2008, to December 31, 2017. This dataset encompasses five different stocks: AAPL (Apple), GOOGL (Google), AMZN (Amazon), and FB (Facebook).

Each tick represents the minimum movement, either upward or downward, in price. Within a one-second timeframe, numerous ticks may occur, often ranging from 20 to 30 ticks. The collected tick data includes essential details such as open bid, close bid, high bid, low bid, and a corresponding timestamp. This high-frequency data is crucial for conducting intra-day short-term predictions.

To fulfill the requirements of our model, which operates on an hourly basis, we ensure that at least one tick is recorded per hour. Given the substantial volume of data (approximately 25,200 tick prices per day, considering 7 trading hours), preprocessing becomes necessary. This preprocessing accounts for both the large volume of data and the variability in tick intervals. It addresses scenarios where multiple prices may be released within the same second, while some seconds might lack recorded ticks altogether.

In addition to tick data, we have collected news sentiments. This dataset includes details such as the stock symbol, date and time of issuance, news source, headline, sentiment (0 for neutral, 1 for positive, and -1 for negative), and polarity of sentiments (positive, negative, and neutral). The polarity of news sentiment is determined based on the frequency of positive and negative words within the news article.

4. Data Preprocessing

To enhance data manipulation and management of the extensive tick data, we migrated our dataset to a MySQL database. This database facilitates sorting data during queries, streamlining the process. The initial phase involved addressing missing tick entries within the dataset. Due to varying time intervals between ticks, some intervals lacked recorded data, leading to irregularities. For instance, certain seconds may contains multiple recorded prices, while others may have none.

To address this, we implemented a strategy to fill in missing ticks by identifying the nearest available tick data to populate the vacant seconds. Following the data import and the filling of missing ticks, we proceeded to aggregate our dataset into one-minute interval. During this aggregation, we extracted the latest tick received for each minute captured in our dataset. Subsequently, we filtered the data to exclude weekends and ticks falling outside of market opening hours.

This refined one-minute dataset, devoid of irregularities and aligned with market hours, was then stored in a new table for further analysis and utilization.

5. Feature Generation

Various window sizes have been experimented with in our models, allowing us to specify how many hours to consider when training. This process generates our input data in a structured format, represented as (window size * features).

The features incorporated in our models encompass:

- Maximum: The highest price recorded during the selected hour.
- Minimum: The lowest price recorded during the selected hour.
- Average: The mean price observed throughout the selected hour.
- Standard Deviation: The degree of dispersion of prices recorded during the selected hour, indicating volatility.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{\sum p_i}{count\,(p_i)}\right)^2}$$

- Pseudo Log Return: The logarithmic difference between the average prices of two consecutive hours, providing insight into price movements.

$$\ln\left(\frac{p_t}{p_{t-1}}\right)$$

where p_t *is the average price at time t*

- Trend Indicator: The slope of a linear model applied to the tick data within the respective hour, offering an indication of recent price trends.
- Price: The last tick received within the selected hour.
- Sentiment: News sentiment analysis.

Thus, our input data comprises 8 features, with the number of features determined by the formula:

Features = 8n, where 'n' represents the window size.

The output generated by our model corresponds to the end-of-day price, providing valuable insights for forecasting and decision-making.

6. Data Normalization

To ensure consistency and efficient convergence in our modeling process, it's essential to normalize the input data since the features extracted come from different units and scales. Normalization scales the data between 0 and 1, facilitating faster convergence. To achieve this, we utilize the min-max scaling technique provided by the scikit-learn framework. This function computes the maximum and minimum values for each column and applies the following formula to normalize the data:

$$\frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Following data normalization, we proceed to experiment with various models to determine the most suitable approach for our task. These models include Recurrent Neural Network (RNN), Deep Neural Network (DNN), Support Vector Machine (SVM), and Support Vector Regression (SVR). Each model offers unique characteristics and capabilities, and through experimentation, we aim to identify the optimal model for our specific requirements.

7. Models

We conducted a comprehensive comparison of four different models—Recurrent Neural Network (RNN), Feed Forward Neural Network (FFNN), Support Vector Machines (SVM), and Support Vector Regression (SVR)—to evaluate their effectiveness in predicting the direction of today's closing price relative to yesterday's closing price.

FFNN, which is extensively employed across various industries including finance, trading, analytics, and product maintenance, is adept at tasks such as classification, regression, and pattern recognition.

RNN, designed for sequence data, distinguishes itself from FFNN by retaining information from previous steps. The memory of an RNN can be represented by various cell types, including Vanilla RNN (for short-term memory), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) which enhance short-term memory using gating mechanisms. We experimented with different network structures, varying the number of layers (between 3 and 7) and the number of neurons per layer (ranging from 250 to 5 neurons). Our testing revealed that networks with 3 and 4 layers yielded the best results. Additionally, we explored different RNN cell types provided by TensorFlow, including Basic RNN, LSTM, and GRU cells. The models were trained for 100 epochs, utilizing the ADAMOptimizer for optimization.

SVM, a supervised learning algorithm, can be utilized for both regression and classification tasks. It employs a kernel trick technique to transform data and identify optimal boundaries between outputs. SVM has demonstrated efficacy in stock market prediction, particularly on non-linear datasets. We experimented with different kernel algorithms (Linear, Polynomial, and Radial Basis Function) and tuned parameters for each model. The output of the SVM model is a binary value: 0 when yesterday's closing price decreases compared to today's, and 1 when the price increases. We utilized the scikit-learn library for model construction, employing GridSearchCV to select the best parameters.

SVR shares similarities with SVM but is tailored for regression instead of classification. It leverages the same principles and functionalities as SVM to predict continuous values, with the output representing the end-of-day price. The modeling process for SVR mirrors that of SVM, with the only difference being the nature of the output, which is not a class but rather a continuous value.

8. Conclusion and Future Works

This paper presents the development of a stock price trend prediction system, utilizing data gathered from two distinct sources: (i) Historical stock market data obtained from Reuters, and (ii) news sentiment related to specific stocks. The dataset spans a period of 10 years and encompasses information from four different stocks. Technical features derived from the historical stock market data were computed and utilized as input for our predictive models. Additionally, three scenarios were considered when incorporating sentiment analysis into the calculated features.

Our AI framework primarily integrates Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), Support Vector Regression (SVR), and Support Vector Machines (SVM) for prediction purposes. These models were trained and tested using the combined dataset consisting of historical market data and sentiment analysis, with the aim of forecasting stock price trends accurately.

The development of this system involved careful preprocessing of the data, feature engineering, model selection, and optimization to achieve robust and reliable predictions. By leveraging advanced machine learning techniques and sentiment analysis, we aimed to provide valuable insights into stock market behavior and improve decision-making processes for investors and traders.

References

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. Expert Systems with Applications,79,153-163. doi:10.1016/j.eswa.2017.02.041

Li, X., Huang, X., Deng, X., & Zhu, S. (2014). Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. Neurocomputing, 142, 228-238. doi:10.1016/j.neucom.2014.04.043

Arévalo, A., Niño, J., Hernández, G., & Sandoval, J. (2016). High-Frequency Trading Strategy Based on Deep Neural Networks. Intelligent Computing Methodologies Lecture Notes in Computer Science, 424436. doi:10.1007/978-3-319-42297-8_40

Horne, J. C., & Parker, G. G. (1967). The RandomWalk Theory: An Empirical Test. Financial Analysts Journal, 23(6), 87-92. doi:10.2469/faj.v23.n6.87

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383. doi:10.2307/2325486

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. Information Processing & Management, 45(5), 571- 583. doi:10.1016/j.ipm.2009.05.001

Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. Computer,44(10), 91-94. doi:10.1109/mc.2011.323 Arora, A., et al.: Deep Learning with H2O (2015)

learning process 20. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method, 6 (2012)

learning process 20. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method, 6 (2012)

Investopedia. “World's Greatest Investors.” Investopedia, Investopedia, 9 July 2008, www.investopedia.com/slide-show/worldsgreatest-investors/.

Singh, Aishwarya. “Predicting the Stock Market Using Machine Learning and Deep Learning.” Analytics Vidhya, 26 July 2019, www.analyticsvidhya.com/blog/2018/10/predicting -stock-price-machine-learningnd-deep-learningtechniques-python/.