Predicting Stock Price Directions using Support Vector Machines: A Data Science Approach

Bhuvan Chandra Sarakam
bhuvansarakam@gmail.com
https://doi.org/10.34293/ICETCS2024.ch004

Abstract-*Predicting stock price movements remains a formidable challenge due to the inherent volatility and complexity of financial markets. This paper investigates the application of Support Vector Machines (SVM) for forecasting stock price directions in the technology sector, leveraging historical data from 34 mature technology companies over the period 2007 to 2014. The study employs four key features: price volatility and momentum for individual stocks and sector-wide metrics, with the goal of uncovering patterns in historical data that can inform future price movements. Results indicate that short-term predictions, such as daily price directions, are only marginally better than random guessing, aligning with the Efficient Markets Hypothesis (EMH), which posits that stock prices fully reflect all available information. However, the model exhibits significantly better performance in long-term forecasts, achieving prediction accuracies of 55%–60% for horizons of 90 days or more. This suggests that while short-term price movements are largely unpredictable, certain long-term trends, particularly at the sectoral level, may provide actionable insights. The findings emphasize the limitations of relying solely on daily closing price data, particularly for short-term predictions, and highlight the potential of incorporating intra-day data to enhance model performance. Additionally, the study underscores the importance of feature selection and parameter optimization in improving predictive accuracy. Expanding the feature set to include company- specific financial metrics and broader macroeconomic indicators could further refine the model. This research contributes to the growing field of machine learning in financial forecasting by demonstrating the potential and limitations of SVM-based approaches for stock prediction. While challenges remain in translating predictive accuracy into trading profits due to market efficiency and transaction costs, the results provide a foundation for future work exploring more granular data, diverse sectors, and advanced feature engineering techniques.*

INTRODUCTION

The prediction of stock prices has long been recognized as one of the most challenging problems in finance, attracting researchers from diverse fields such as economics, finance, and computer science. The inherent complexity and volatility of financial markets make accurate forecasting particularly difficult. Stock prices are influenced by a wide range of factors, including macroeconomic conditions, sectoral trends, company-specific developments, and investor sentiment, many of which are difficult to quantify or anticipate. Despite these challenges, stock price prediction remains a highly studied problem due to its practical significance: even small improvements in prediction accuracy can translate into substantial financial gains.

Traditional approaches to stock price prediction often rely on statistical and econometric methods, such as time-series models like ARIMA [1]. These linear models, while effective in capturing trends in stationary data, struggle to account for the non-linear and stochastic nature of stock price movements. As a result, more advanced models, such as ARCH and GARCH [2], have been developed to capture the dynamic volatility of financial time series. However, these models also face limitations when dealing with the high dimensionality and complexity of modern financial markets.

In recent years, the advent of machine learning and big data techniques has revolutionized stock price prediction. Machine learning models, particularly those that can capture non-linear relationships and interactions, have shown promise in improving predictive accuracy. These techniques have found applications in algorithmic trading and high-frequency trading, where rapid and accurate predictions are critical. Support Vector Machines (SVMs) have gained attention due to their ability to handle high-dimensional data and find optimal decision boundaries for classification tasks. In the context of stock price prediction, SVMs offer a robust framework for predicting directional movements, i.e., whether a stock's price will rise or fall over a given time horizon.

This study focuses on applying SVMs to predict the directional movement of stock prices in the technology sector. Using historical data from 34 mature technology companies between 2007 and 2014, I aim to evaluate the effectiveness of SVMs in this context. my model incorporates four key input parameters: the recent price volatility and momentum of individual stocks and the corresponding metrics for the technology sector. By analyzing these features, I seek to uncover patterns that may provide insight into future price movements.

The Efficient Markets Hypothesis (EMH) presents a significant theoretical challenge to stock price prediction. According to EMH, stock prices fully reflect all available information, and future price movements are determined by new, unpredictable information. As a result, any attempt to predict stock prices using historical data should, in theory, be no better than random guessing my findings, however, reveal a more nuanced picture. While short-term price movements remain difficult to predict, consistent with EMH, my model demonstrates modest predictive power over longer time horizons. Specifically, I achieve prediction accuracies of 55%-60% when forecasting price directions over medium to long-term periods, such as 90 days or longer.

These results suggest that while short-term stock price movements are largely random, there may be underlying trends or patterns that emerge over longer periods, particularly at the sectoral level. At the same time, my study highlights the limitations of using daily closing price data for short-term predictions. Financial institutions often utilize intra-day data, which captures price movements at a much finer granularity, to enhance predictive accuracy and profitability. Incorporating such data into future models may provide significant improvements in short-term forecasting performance.

In this paper, I aim to contribute to the growing body of literature on machine learning applications in finance by exploring the potential of SVMs for stock price prediction. In addition to evaluating the model's predictive accuracy, I analyze the impact of key input parameters and prediction horizons on performance. My findings underscore both the opportunities and limitations of SVM-based approaches in financial forecasting, paving the way for further research in this domain.

RELATED WORK

The prediction of stock price movements has been a significant research area, spanning disciplines such as finance, economics, and computer science. Traditional approaches have predominantly relied on statistical models like ARIMA and time-series analysis, which perform well on stationary data but struggle to capture the non-linear and stochastic nature of financial markets. Advanced models like ARCH and GARCH have been developed to handle dynamic volatility, yet they face limitations in high-dimensional scenarios and non-linear dependencies.

The advent of machine learning has provided new opportunities to enhance predictive accuracy by identifying complex patterns in financial data. Artificial Neural Networks (ANNs) have been widely adopted in this domain. Kar and Patel explored the use of ANNs for stock prediction, demonstrating their capacity to model intricate relationships in financial datasets. However, the performance of ANNs can be hindered by overfitting, instability, and computational inefficiency in highly volatile markets.

Support Vector Machines (SVMs) have emerged as a promising alternative, particularly for classification tasks in financial prediction. Kim highlighted the robustness of SVMs in forecasting financial time series, emphasizing their ability to find global optima and handle high-dimensional data. Shah corroborated these findings, showcasing the superior performance of SVMs over traditional neural networks in stock price direction prediction. The inclusion of kernel functions, such as the Radial Basis Function (RBF), enables SVMs to model non-linear relationships effectively, making them particularly suited for financial time-series data.

Machine learning techniques have also been leveraged to challenge the Efficient Markets Hypothesis (EMH). While EMH asserts that stock prices reflect all available information, thereby rendering future movements' unpredictable, studies such as Jegadeesh and Titman, Jacobsen and Zhangidentified market anomalies, including momentum and seasonal trends. These findings suggest that machine learningmodels can exploit such anomalies to achieve predictive accuracies above random guessing, thereby challenging traditional views of market efficiency.

Despite these advances, the practical application of machine learning in stock price prediction remains constrained by several factors. Transaction costs, market liquidity, and the variability in model performance across different stocks and sectors limit the real-world applicability of these models. Additionally, the reliance on daily closing price data poses challenges for short-term predictions. Incorporating intra-day trading data and additional features, such as macroeconomic indicators and company-specific metrics, has been suggested as potential avenues for improvement.

This study builds on the existing body of work by employing SVMs to predict stock price directions in the technology sector, focusing on the interplay between individual stock features and broader sectoral trends. By examining prediction accuracies across various horizons, this research aims to contribute to the understanding of both the potential and limitations of SVM-based approaches in financial forecasting.

BACKGROUND INFORMATION

A. Stock Market Efficiency

The Efficient Markets Hypothesis (EMH) posits that stock prices fully reflect all available information, making them unpredictable [3]. According to EMH, prices follow a random walk, responding only to new information. This unpredictability signifies market efficiency since predictable movements would indicate unreflected information.

EMH has three variants: weak, semi-strong, and strong. The semi-strong form, widely supported by research, states that prices reflect all publicly available information, while private information can lead to unfair profits, forming the basis for insider trading laws.

However, market anomalies challenge EMH. For instance, Jegadeesh and Titman observed short-term momentum in stock prices [4], and Jacobsen and Zhang identified seasonal trends, such as higher winter returns [5]. These phenomena suggest some predictability, contradicting EMH. Studies, including ours, show that machine learning models can exploit such trends to achieve prediction accuracies above 50%, challenging the notion of perfect market efficiency.

B. General Machine Learning

Machine learning techniques are divided into supervised and unsupervised learning. In supervised learning, models are trained on labeled data to predict outputs for unseen data [6]. This can be further categorized into classification problems (discrete outputs) and regression problems (continuous outputs). In this study, I treat stock price forecasting as a binary classification problem, predicting whether a stock's price will rise (+1) or fall (−1) after m days based on features such as volatility and momentum.

C. Previous Research

Most stock forecasting studies have used Artificial Neural Networks (ANNs), which model patterns through interconnected layers of neurons. While effective for stable data, ANNs struggle with the non-linear and volatile nature of stock markets [7], [8]. Support Vector Machines (SVMs) offer a robust alternative by finding a global optimum and focusing on maximizing the margin between classes. Research has shown that SVMs outperform ANNs in stock prediction, achieving accuracies up to 60% [9], [10]. Consequently, I employ SVMs for my analysis.

D. Support Vector Machines

SVMs are powerful binary classifiers that construct a decision boundary (hyperplane) to separate data into two categories. The optimal hyperplane maximizes the margin, defined as the distance between the boundary and the nearest data points (support vectors). For non-linear separations, SVMs use kernel functions to map data into higher-dimensional spaces for linear classification. The radial basis function (RBF) kernel, defined as

$$K(x_i, x_k) = \exp\left(-\frac{1}{\delta^2}\sum_{j=1}^{n}(x_{ij} - x_{kj})^2\right)$$

is particularly suited for time-series data due to its ability to capture complex patterns [9], [11].

MODEL CREATION AND EVALUATION

A. Data Collection

The study focuses on 34 stocks from the technology sector during the period 2007 to 2014. This timeframe encompasses significant market events, including the Great Recession and the subsequent recovery, providing a diverse dataset with both high and low volatility periods. Daily closing price data for the selected stocks were obtained from the Center for Research in Security Prices (CRSP) and Yahoo Finance.To ensure robust model training and testing, the dataset was split into two parts:

Training Set (2007–2011): Comprising 62.5% of the dataset, this period was used to train the SVM model. It includes data from both recessionary and recovery phases.

Testing Set (2012–2014): Comprising 37.5% of the dataset, this period evaluates the model’s performance on unseen data, capturing a stable market phase and some volatility.

This temporal split ensures the model’s ability to generalize across different market conditions.

B. Feature Selection

To capture the dynamics of both individual stocks and the broader technology sector, four features were derived from the dataset:

Stock Volatility: Measures the average daily price change of an individual stock over the past *n* days.

Stock Momentum: Captures the directional trend of an individual stock over the past *n* days.

Sector Volatility: Represents the average daily price change of the technology sector index over the past *n* days.

Sector Momentum: Tracks the directional trend of the sector index over the past *n* days.

These features were calculated over varying lookback periods, $n \in \{5, 10, 20, 90, 270\}$, to examine short-term, medium-term, and long-term trends. The feature definitions and formulas are summarized in Table II.

C. Methodology

The predictive model uses a Support Vector Machine (SVM) algorithm with a radial basis function (RBF) kernel. This choice balances computational efficiency with the ability to capture non-linear relationships in the data. The training and evaluation process is as follows:

Feature Engineering: For each stock, the four features (stock volatility, stock momentum, sector volatility, sec- tor momentum) were computed for varying lookback periods n. The SVM model was trained using combinations of n_1 (for stock features) and n_2 (for sector features) to explore the effect of short-term and long- term trends.

Prediction Horizon: The model predicts the stock price direction for $m \in \{1, 5, 10, 20, 90, 270\}$ days into the future, enabling an analysis of short-term and long-term forecasting performance.

Data Splitting: Each stock's data was split into training (70%) and testing (30%) subsets, ensuring no data leakage between these phases.

Evaluation Metrics: Prediction accuracy was calculated as the percentage of correct directional predictions (upward or downward movement) across the testing set. Results were analyzed across different values of n_1, n_2, and m.

D. Evaluation and Results

The model's performance was evaluated across varying prediction horizons. Figure 3 illustrates the mean prediction accuracy for different values of m. The results highlight those short-term predictions (m = 1 or m = 5) are only marginally better than random guessing, with accuracies around 50%. However, for long-term predictions (m = 90 or m = 270), certain parameter combinations achieved accuracies exceeding 60%.

Figure 4 shows the variation in mean prediction accuracy for different combinations of n_1 (stock features) and n_2 (sector features) when m = 90. These results suggest that long-term trends at the sector level play a critical role in improving prediction accuracy.

The results confirm that the Efficient Markets Hypothesis (EMH) holds for short-term predictions, where price movements appear random. For long-term predictions, incorporating sector-wide trends and selecting optimal feature parameters significantly improves performance, indicating the potential for actionable insights in long-horizon forecasts.

Table 1: Features used in SVM

Feature Name	Description	Formula
σ_s	Stock price volatility. These measures the average daily percent change in the stock's closing price over the past *n* days.	$\frac{\sum_{i=t-n+1}^{t} \frac{C_i - C_{i-1}}{C_{i-1}}}{n}$
Stock Momentum	This measures the average directional movement of the stock's price over the past *n* days. A value of +1 is assigned if the closing price increases, and −1 if it decreases, for each day.	$\frac{\sum_{i=t-n+1}^{t} y_i}{n}$
σ_i	Index volatility. These measures the average daily percent change in the index's closing price over the past *n* days.	$\frac{\sum_{i=t-n+1}^{t} \frac{I_i - I_{i-1}}{I_{i-1}}}{n}$
Index Momentum	This measures the average directional movement of the index's price over the past *n* days. A value of +1is assigned if the index price increases and −1 if it decreases, for each day.	$\frac{\sum_{i=t-n+1}^{t} d_i}{n}$

Where: *Ct* is the stock's closing price at time *t* and *It* is the index's closing price at time *t*. The stock's daily directional change is labeled as $y \in \{-1, 1\}$, and the index's daily directional change is labeled as $d \in \{-1, 1\}$. These features are used to predict the direction of the stock price change over a specified period of $m \in \{1, 5, 10, 20, 90, 270\}$ days.

Table 2: Features used in SVM Model

Feature Name	Formula and Description
Stock Volatility	Volatility = $\frac{\sum_{i=t-n+1}^{t} \left\| \frac{C_i - C_{i-1}}{C_{i-1}} \right\|}{n}$ Captures the average daily price change of a stock over the past *n*days.
Stock Momentum	Momentum = $\frac{\sum_{i=t-n+1}^{t} y_i}{n}$ Measures the directional trend of a stock over *n* days, where $y_i$is the price direction (+1 for an upward movement, −1 for downward).
Sector Volatility	Volatility = $\frac{\sum_{i=t-n+1}^{t} \left\| \frac{I_i - I_{i-1}}{I_{i-1}} \right\|}{n}$ Tracks average daily price changes for the technology sector index over *n* days.
Sector Momentum	Momentum = $\frac{\sum_{i=t-n+1}^{t} d_i}{n}$ Measures the sector's directional trend over *n* days, where $d_i$is the sector index direction (+1 for upward movement, −1 for downward).

DISCUSSION

The SVM model's accuracy for predicting one-day price direction (m = 1) is only slightly better than random guessing, supporting the Efficient Markets Hypothesis (EMH).

Prices appear to follow a random walk, where tomorrow's direction depends only on new information, making it difficult for historical data to offer predictive insights.

As m increases, mean prediction accuracy initially improves (m = 5, 10, 20) but declines for m = 90, 270 (Figure 3). Notably, when m = 1, short-term (n_1, n_2 = 5) and long-term (n_1, n_2 = 90, 270) trends yield the best results, while intermediate trends perform worse than random guessing. For larger m, input parameters significantly influence results. For example, when m = 90, accuracies range from below 50% (e.g., n_1 = 10, n_2 = 90) to over 60% (e.g., n_1 = 270, n_2 = 5), indicating that parameter choice becomes critical (Figure 4).

Sector trends (n_1 = 270) generally outperform short-term stock trends (n_2 = 5), particularly for longer-term predictions. This suggests that macroeconomic conditions have greater predictive power than idiosyncratic stock behavior for long- term trends. Conversely, for short-term predictions, recent momentum (n_2 = 5) is more influential than sector trends. This aligns with EMH, as long-term predictions depend more on broader market conditions than individual stock trends.

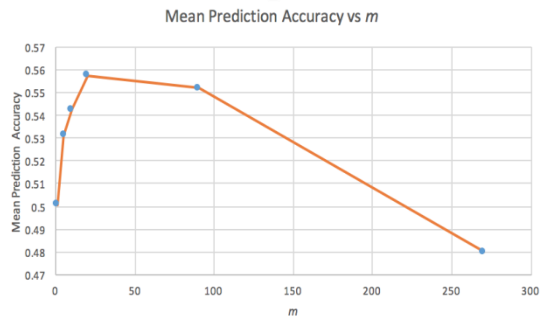

Figure 1 Mean Prediction Accuracy for Different Prediction Horizons (m)

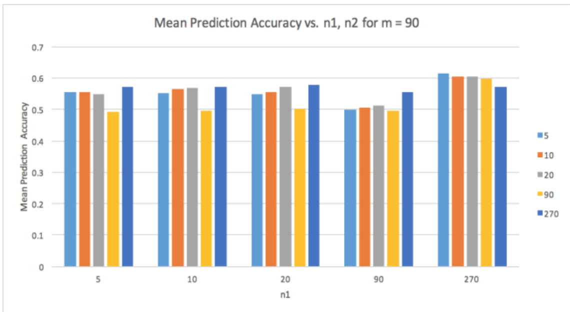

Figure 2 Mean Prediction Accuracy for Different Combinations of n_1 and $n_2$when m = 90.

While higher prediction accuracy for longer horizons (m = 90, 270) is promising, translating these into trading profits is challenging. Short-term accuracy may not account for trans- action costs, and long-term predictions suffer from significant variation across stocks. Some stocks exceed 80% accuracy, while others fall below 30%, and identifying which stocks the model predicts well remains an open question. Overall, while trends exist, they do not consistently translate to actionable trading strategies.

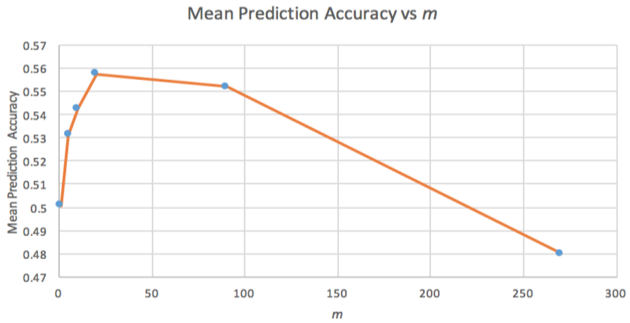

Figure 3 The Mean Prediction Accuracy for Each Parameter *m*

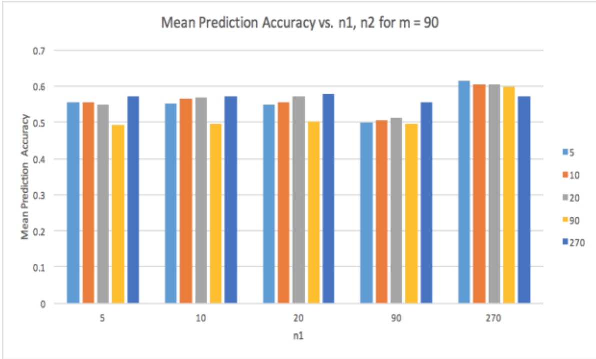

Figure 4 The Mean Prediction Accuracy for Each Combination of n_1, n_2 when$m$ = 90

FUTURE WORK

A. Granularity of Data

This study is limited to daily stock price data, with momentum and volatility calculated over weeks. Future work could analyze intra-day trading data, enabling models to capture price swings occurring over minutes or seconds. By studying intra-day trends, I can create models that capitalize on rapid momentum shifts, potentially enhancing short-term predictive power.

B. Feature Expansion

Expanding the feature set is a key avenue for improvement. In addition to volatility and momentum, future models could incorporate company-specific metrics such as P/E ratio, market cap, and working capital ratio, as well as macroeconomic indicators like interest rates, inflation, and GDP growth. Such features could provide deeper insights into stock price movements.

C. Broader Application

This study focused on the technology sector and mature companies. Future work could apply the model to other sec- tors, such as healthcare or retail, and analyze its performance across companies of varying sizes, including startups and mid- sized firms. This would help evaluate how sector-specific or size-related factors influence predictive accuracy.

CONCLUSION

This study highlights the inherent challenges in predicting stock price direction using historical data and provides empirical support for the Efficient Markets Hypothesis (EMH). The findings reveal that short-term predictions, such as forecasting the price direction for the next trading day, are not significantly better than random guessing. This aligns with the EMH, which posits that prices reflect all available information and that future price movements are determined by new, unpredictable information.

As the prediction horizon increases, the results demonstrate modest improvements in predictive accuracy. Specifically, long-term trends, influenced by sectoral and broader market conditions, exhibit some predictive potential. This indicates that while short-term stock price movements are largely stochastic, long-term price directions may reflect structural or macro economic factors that manifest over time. However, the effectiveness of predictions is highly sensitive to the choice of input parameters, such as the historical time windows n_1 and n_2, and the prediction horizon m. For instance, longer sector trends (n_1 = 270) and shorter individual stock momentum (n_2 = 5) are particularly effective in predicting longer-term price directions.

Despite the observed trends, translating predictive ability into profitable trading strategies remains a significant challenge. Short-term trading profits are constrained by transaction costs and the small magnitude of typical price movements. Similarly, for long-term predictions, the variability in model accuracy across stocks introduces uncertainty, as some stocks show high predictive accuracy while others perform worse than chance. Without prior knowledge of which stocks the model will predict accurately, building a reliable trading strategy becomes difficult.

The study also underscores the limitations of using a static feature set. Although the inclusion of stock-specific momentum and volatility alongside sectoral trends provides a useful starting point, the model's predictive performance could benefit from additional features. Incorporating company-specific financial metrics, such as P/E ratio, market capitalization, or working capital ratio, as well as macroeconomic indicators like interest rates, inflation, or GDP growth, could provide a more comprehensive understanding of price dynamics. Additionally, considering intra-day trading data could enhance the model's granularity and responsiveness to rapid price changes.

Future research should aim to address these limitations by refining the model and exploring its applicability in diverse scenarios. Expanding the dataset to include stocks from other sectors, such as healthcare or retail, and analyzing companies of varying sizes, including startups and mid-cap firms, would provide a broader perspective on the model's utility. Moreover, adopting a more granular approach, such as using minute- level data, could uncover patterns that daily data may obscure, especially for short-term predictions.

In conclusion, this study provides valuable insights into the complexities of stock price prediction and its limitations within the context of historical data analysis. While the results reaffirm the challenges posed by the EMH, they also suggest opportunities for improvement through the inclusion of richer features, finer-grained data, and application across diverse sectors and firm sizes. By addressing these areas in future research, I can better understand the interplay between historical trends and stock price movements and potentially enhance the predictive capabilities of machine learning models in finance.

REFERENCES

[1] Bontempi, G., Taieb, S. B., & Borgne, Y, L. (2013). Machine Learning Strategies for Time Series Forecasting. *Lecture Notes in Business Information Processing*, Springer.

[2] Zhang, J. (2009). Applying time series analysis builds stock price forecast model. *Modern Applied Science*, *3*(5).

[3] Bodie, Z., Kane, A., &Marcus, A. (2014). *Investments*. McGraw-Hill.

[4] Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65-91.

[5] Jacobsen, B., & Zhang, C. (2012). Are monthly seasonals real? A three century perspective. *Review of Finance*.

[6] Brownlee, J. (2023). *A tour of machine learning algorithms*.

[7] Kar, A. *Stock prediction using artificial neural networks*.

[8] Patel, M., Yalamalle, R. S. & Mayankkumar, Y. (2014). Stock prediction using artificial neural networks. *International Journal of Innovative Research in Science, Engineering, and Technology*.

[9] Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*, 307-319.

[10] Shah, V. H. (2007). *Machine learning techniques for stock price prediction*.

[11] Buhmann, M. (2010). *Radial basis functions*.