Article

Machine Learning to Forecast the Financial Bubbles in Stock Market: Evidence in Vietnam

Kim Long Tran , Hoang Anh Le * , Cap Phu Lieu , Duc Trung Nguyen

Posted Date: 30 September 2023

doi: 10.20944/preprints202309.2150.v1

Keywords: financial bubbles; machine learning algorithms; Vietnamese stock market doi:10.20944/preprints202309.2150.v1

Article

Machine Learning to Forecast the Financial Bubbles in Stock Market: Evidence in Vietnam

Kim Long Tran [1], Hoang Anh Le [1,*], Cap Phu Lieu [1] and Duc Trung Nguyen [1]

[1] Ho Chi Minh University of Banking, No. 36 Ton That Dam Street, Nguyen Thai Binh ward, District 1, Ho Chi Minh City, 700000, Vietnam; longtk@hub.edu.vn (K.L.T.); phulc@hub.edu.vn (C.P.L.); trungnd@hub.edu.vn (D.T.N.)

* Correspondence: anhlh_vnc@hub.edu.vn

Abstract: Financial bubble prediction has been a significant area of interest in empirical finance, garnering substantial attention in the literature. This study aimed to detect and forecast financial bubbles in the Vietnamese stock market from 2001 to 2021. The PSY procedure, which involves a right-tailed unit root test to identify the existence of financial bubbles, was employed to achieve this goal. Machine learning algorithms were then utilized to predict real-time financial bubble events. The results revealed the presence of financial bubbles in the Vietnamese stock market during 2006-2007 and 2017-2018. Additionally, the empirical evidence supported the superior performance of the Random Forest and Artificial Neural Network algorithms over traditional statistical methods in predicting financial bubbles in the Vietnamese stock market.

Keywords: financial bubbles; machine learning algorithms; Vietnamese stock market

1. Introduction

Financial bubbles can have profound effects on a economy and the livelihoods of its citizens. Consequences such as financial and debt crises, as seen in the dot-com bubble of 1999-2001 and the U.S. subprime mortgage crisis of 2007-2009, often originate from the abnormal escalation of asset prices in the stock market or the real estate market. When a bubble bursts, it can lead to the collapse of major financial institutions, pushing countries to the brink of bankruptcy and triggering comprehensive financial and economic crises. Governments are forced to allocate substantial resources to bailout packages and recovery programs in the wake of such crises.

Moreover, the social costs remain significant, and the restoration of public confidence in the market poses a particularly formidable challenge. Inexperienced investors, lacking the knowledge to manage risks, are among the most severely impacted [1]. They often hold assets during the late stages of a bubble, making them particularly vulnerable to the adverse effects of the bubble's burst. Therefore, for governments and market supervisors, researching and forecasting the status of financial bubbles is extremely important. This enables the implementation of necessary interventions to mitigate adverse impacts of financial bubbles on the market and society, especially in the context of globalization where risks can easily spread between markets. However, research relating to financial bubbles in the context of Vietnam remains relatively limited, especially in studies employing quantitative methodologies and machine learning algorithms. The effectiveness of forecasting tools utilizing machine learning algorithms continues to be a subject of skepticism and debate within the academic community.

The main objective of our research is to detect financial bubbles in the Vietnamese stock market from 2001 to 2021 and subsequently forecast financial bubble occurrences based on macroeconomic indicators. We employed the PSY procedure to identify market phases with bubble-like characteristics and use machine learning algorithms for prediction. Furthermore, we have implemented the SMOTE method to rectify data imbalances within the dataset and leveraged the PCA method to reduce data dimensionality, thereby enhancing the quality of our forecasting outcomes. We aim to determine the best-performing model among the methods used. We anticipate that our study can contribute empirical evidence regarding the application of machine learning for real-time forecasting of financial bubbles. This can provide early warning signals for policymakers and investors to make informed financial decisions.

The remaining sections of this research paper are organized as follows: Section 2 introduces the theoretical framework for detecting financial bubbles and the application of machine learning in forecasting. Section 3 outlines our data sources and methodologies, including the PSY procedure and machine learning techniques. Section 4 presents our research findings and thoroughly discusses their implications. Finally, Section 5 provides a concise conclusion, summarizing key insights and suggesting avenues for future research.

2. Literature Review

2.1. Definition of Financial Bubbles

In theory, the concept of financial bubbles is often referred to by various terms like asset price bubbles or speculative bubbles, and is an intriguing research topic. There exist various perspectives on financial bubbles and their identification. However, attempting to classify and provide a clear-cut definition remains a controversial subject within the academic community. Generally, the classification of financial bubbles falls into two primary categories: classical bubbles and modern bubbles.

Classical bubbles are primarily driven by irrational investor behavior. Shiller [2] posits that bubbles in the market are a psychological phenomenon. He suggests that the occurrence of these bubbles is a result of amplified feedback-trading tendencies, which are caused by the attention given to them by news media. The reason for this is that as more investors show interest in a particular asset, news media tend to expand their coverage of it, which in turn attracts even more potential investors. This leads to an increase in demand for the asset, which causes its price to rise, thereby attracting even more attention from the news media. This cyclical process reinforces the feedback-trading tendency in the market, ultimately leading to the occurrence of bubbles. This phenomenon is often referred to as 'herd mentality', and its consequences can lead to a severe market collapse, subsequently exerting a profound impact on the overall economy. Kindleberger, Aliber and Solow [3] proposed an approach to understanding financial bubbles from the perspectives of irrational exuberance and psychological expectations. According to these authors, financial bubbles were created by the irrational exuberance and blind faith of investors, leading to a series of reckless investment decisions and ultimately culminating in a market collapse and asset value correction. Stiglitz [4] suggested that the phenomenon of a financial bubble occurs when investors believe that current prices already reflect high levels of expectations, and the fundamental factors supporting those prices are no longer in place. In other words, when investors believe that current prices no longer offer them the potential for future profits and this sentiment becomes widespread, a bubble begins to form. When investors have faith that the upward trend will continue and fear missing out on potential gains if they do not buy at the present moment, the bubble inflates. However, a bubble will be prone to burst when investors start to believe that prices can no longer rise further, demand wanes, and this can trigger a significant sell-off, causing prices to fall rapidly [5].

The second approach is the modern bubble, described by Tirole [6] as a situation in which the price of an asset exceeds its fundamental value. The fundamental value of an asset is typically based on its expected future cash flows, such as dividends, coupon payments, or rental income. According to the author, bubbles occur when investors are willing to pay a higher price for an asset that can be resold immediately than if they were obligated to hold onto it for a longer period. This view recognizes that an asset's perceived value is not always tied to its true value, but rather to its potential for short-term profits. In the derivatives market, a bubble is considered to exist if the market value of a derivative consistently exceeds the cost of creating similar derivatives. This means that the price at which the derivative is trading in the market is higher than the cost of creating a comparable derivative. An illustration of such a bubble can be observed in the price disparity in option pricing. Specifically, a bubble may occur when a combination of put and call options, designed to replicate the movements of a stock, is traded at a price differential compared to the price of the underlying stock. This price differential must also take into account factors such as interest rates and the cost of borrowing the stock. Therefore, the existence of bubbles in various forms highlights the complexity of market dynamics and raises the challenges associated with maintaining economic stability.

Additionally, the concept of bubbles can be classified into two categories: rational bubbles and partially rational bubbles. The rational bubble theory proposes that investors knowingly purchase overpriced assets with the understanding that they can sell them at a profit in the future. This theory posits that even when faced with prices that are clearly overvalued, expectations of future profits can drive investment behavior. In other words, investors in rational bubbles willingly engage in the bubble, motivated by the prospect of profiting from price increases before the bubble eventually bursts. Shiller [7] introduced the concept of partially rational bubbles, which posits that stock prices are influenced by a combination of rational and irrational behavior among investors. He suggested that individual investors were prone to irrational exuberance, often driven by sensationalized media reports, but this does not mean that investors are consistently irrational or 'crazy'. Rather, the stock market is influenced by social trends and short-term desires, which may either lead to the formation of bubbles or not. In essence, partially rational bubbles recognize that market behavior can be influenced by both rational and irrational elements that stem from societal trends and collective beliefs, and these factors contribute to the persistence of bubbles. This perspective provides a more nuanced understanding of the dynamics that drive financial bubbles.

Fama [8], who advocates the Efficient Market Hypothesis (EMH), offers an alternative perspective on financial bubbles. According to Fama, the extreme price fluctuations of assets can be anticipated, and as a result, there are no bubbles in asset prices. While this stance is still subject to debate, it provides a framework for delving into the identification of factors or causes that could lead to the predictable formation of bubbles. Fama's view challenges the perception that financial bubbles are irrational and uncontrollable, and proposes that there may be underlying patterns and elements that can be used to predict or understand the emergence of bubbles. This perspective has given rise to further research into the predictability of bubbles and the factors that contribute to their formation.

In this article, we approach bubbles from the perspective of irrational bubbles, which are characterized by a sudden surge in prices within a short time frame followed by a rapid decline in the VNINDEX. Besides, we also recognize the presence of external factors that influence market dynamics beyond investor behavior, as posited by Fama's perspective.

2.2. Literature review on Detecting Financial Bubbles

Throughout history, a multitude of research studies have been conducted with the objective of identifying bubbles in financial markets that are characterized by speculation, including but not limited to stock markets, foreign exchange markets, real estate markets, and more recently, cryptocurrency markets. However, within the context of this article, we will provide a concise overview of studies conducted solely within the sphere of the stock market, with a particular focus on those that utilize statistical models on time series data.

Shiller [9] introduced a novel method called Variance Bounds Tests, which he applied to sample data of the S&P price index from 1871 to 1979, revealing evidence of bubble existence. However, Shiller's approach is often deemed less reliable when applied to small sample sizes. Blanchard and Watson (1982) explained rational bubbles as the discrepancy between asset prices and their fundamental values, suggesting that speculative bubbles do not adhere to rational behavior. The impact of irrational factors adds complexity to assessing bubbles in the stock market.

West [10] employed Euler equations and ARIMA models on annual stock price and dividend data of the S&P 500 from 1871 to 1980 and the Dow Jones index from 1928 to 1978, providing robust statistical evidence of the existence of stock market bubbles in the United States.

Phillips, Wu and Yu [11] proposed the Sup Augmented Dickey-Fuller test (SADF), also known as the PWY method, to assess the presence of rational bubbles in financial markets. This approach is based on the null hypothesis of a unit root, analogous to the conventional Dickey-Fuller test, but with a right-tailed alternative hypothesis. Rejecting the null hypothesis in this test indicates the presence of explosive behavior in the price series, thereby providing empirical evidence for the existence of a bubble. The right-tailed SADF unit root tests are conducted using rolling window forms. Homm and Breitung [12] applied this test to detect stock market bubbles, and after a process of simulation and evaluation criteria comparison, the authors found that the SADF test was the most optimal among the methods employed. The SADF test is effective when there is a single bubble event, but in practical applications, there may be multiple bubbles appearing in sufficiently large samples. While this method successfully identified famous historical bubbles, the SADF test failed to detect the 2007-2008 debt crisis bubble.

Phillips, Shi and Yu [13] developed the Generalized sup ADF (GSADF) test as an improvement over the SADF method, also referred to as the PSY procedure, to overcome its limitations. The GSADF test is an iterative application of the right-tailed ADF test based on the rolling-window SADF test that aims to detect explosive patterns in sample sequences. Compared to the SADF, the GSADF is more flexible in terms of rolling windows, making it a valuable tool for investigating price explosion behavior and confirming the presence of market bubbles. In this study, they applied both the SADF and GSADF tests to the S&P 500 index from 1871 to 2010, revealing that the GSADF successfully identified two bubble periods: the Panic of 1873 (from October 1879 to April 1880) and the Dot-com bubble (from July 1997 to August 2001). When restricting the bubble duration to over 12 months, the results showed that there were three existing bubble periods: the post-1954 war period, Black Monday in 1987, and the Dot-com bubble in 2000.

Based on the findings outlined above, it is evident that the GSADF test is an effective approach for detecting the presence of market bubbles. Hence, in this study, we have employed the PSY procedure to identify the existence of bubbles in the Vietnamese stock market.

2.3. Literature Review on Machine Learning Applied in Economic Forecasting

In macroeconomics, machine learning techniques are commonly used for classification problems such as predicting stock market trends and corporate bankruptcies. While the use of machine learning for identifying financial bubbles in the stock market is a relatively new approach and has been the subject of limited research, most studies on financial crises in banking, securities markets, and public debt have focused on predicting general economic crises rather than forecasting individual financial bubbles. To date, only one recent study by Başoğlu Kabran and Ünlü [14] has been conducted on predicting financial bubbles. The authors used a support vector machine (SVM) approach to forecast bubbles in the S&P 500 index and compared it with other methods. The results demonstrated that SVM has superior performance in predicting financial bubbles.

For forecasting financial crises, several notable studies have been conducted. Alessi and Detken [15] constructed a warning system using random forest to identify systemic risk from a dataset of banking crises in the EU, using macroeconomic indicators as predictors. Their results demonstrated that the random forest model provided excellent predictive performance and holds promise for macroeconomic forecasting. Beutel, List and von Schweinitz [16] compared the out-of-sample predictive performance of various early warning models for systemic banking crises in advanced economies and found that while machine learning methods often exhibit high in-sample fits, they were outperformed by the logit approach in recursive out-of-sample evaluations. Chatzis, Siakoulis, Petropoulos, Stavroulakis and Vlachogiannakis [17] utilized a wide range of machine learning algorithms to forecast economic risks across 39 countries. They demonstrated that deep neural networks significantly improved classification accuracy and provided a robust method to create a global systemic early warning tool that is more efficient and risk-sensitive compared to established methods. Ouyang and Lai [18] proposed an Attention-LSTM neural network model to assess systemic risk early warning in China. They found that the model exhibited superior accuracy compared to other models, suggesting that it could be a valuable tool for systemic risk assessment and early warning in the Chinese context.

In default prediction, machine learning models have also demonstrated superiority in handling non-linear relationships compared to traditional models. Shin, Lee and Kim [19] employed SVM to forecast bankruptcy for 2,320 medium-sized enterprises in the Korean Credit Guarantee Fund from

1996 to 1999. The study's outcomes exhibited that SVM provided superior predictive outcomes compared to other models, including artificial neural network (ANN) models. Zhao, Xu, Kang, Kabir, Liu and Wasinger [20] conducted a research to develop a credit scoring system based on ANN employing a credit dataset from Germany. The study's results demonstrated that ANN could forecast credit scores more accurately than traditional models, obtaining an accuracy rate of 87%. Geng, Bose and Chen [21] utilized machine learning models to predict financial distress for companies listed on the Shanghai and Shenzhen stock exchanges from 2001 to 2008. The study's findings revealed that the ANN model yielded better results when compared to decision trees, SVM, and random forest models. Barboza, Kimura and Altman [22] applied SVM, ensembles, boosting, and random forest methods to predict bankruptcy for 10,000 companies in the North American market from 1985 to 2013. The authors contrasted these models with traditional statistical models and examined their predictive performance. The results indicated that ensemble methods like bagging, boosting, and random forests outperformed other approaches. Specifically, machine learning models achieved an average accuracy that was approximately 10% higher than traditional models. The random forest model displayed the highest accuracy, reaching up to 87%, whereas traditional models ranged from 50% to 69% accuracy. Fuster, Goldsmith-Pinkham, Ramadorai and Walther [23] examined mortgage default cases in the United States and found that the random forest model achieved higher predictive accuracy than logistic regression. These results highlight the effectiveness of ensemble methods, particularly random forests, in financial distress prediction compared to traditional models. A recent study conducted by Tran, Le, Nguyen and Nguyen [24] incorporated the Sharpe ratio to clarify the forecasting outcomes of complex machine learning models on a dataset of listed companies in Vietnam from 2010 to 2021. The study's results showed that the extreme gradient boosting and random forest models outperformed other models. One interesting point is that based on Shapley values, the authors can determine the influence of each feature on the prediction results and provide additional insights into relationships that are not present in the theory.

The literature suggests that machine learning algorithms have exhibited better outcomes than statistical models in both classification and time series regression problems. Nevertheless, the forecasting results across models can differ considerably depending on the dataset used, and there is no one-size-fits-all superior approach. Several factors can significantly impact forecasting results, such as data imbalance, the presence of outliers in the dataset, and the selection of parameters within the models. In this case, the choice of the forecasting model is of utmost importance and can significantly impact the output results. Therefore, the primary aim of this research is to identify and select the most appropriate model for forecasting financial bubbles in the Vietnamese stock market. To our best knowledge, there has been no prior research conducted on this topic.

3. Methodology

3.1. Research Design and Data Preprocessing

Our study is divided into two primary phases. Phase 1 involves the detection of financial bubbles in the stock market, while Phase 2 employs machine learning algorithms to predict the appearance of financial bubbles. The details of each phase are as follows:

In Phase 1, we focus on identifying financial bubbles in the Vietnamese stock market, utilizing historical stock market data, specifically the VNINDEX, from 2001 to 2021. This timeframe encompasses significant economic events and global economic fluctuations, particularly during the 2006-2008 and 2017-2018 periods, as noted by economic experts. Our goal is to identify financial bubble phenomena on a monthly basis during this time frame. We obtained daily stock index data from FiinPro, a reputable financial data provider in Vietnam.

In Phase 2, we use various machine learning techniques to predict the presence of financial bubbles. We utilize macroeconomic indicators of Vietnam during the same timeframe to forecast financial bubble occurrences. The target variable is the financial bubble status for each month, as detected in Phase 1. We label months as 1 if a financial bubble occurred and 0 for other months. The forecasting time frame is one month prior. The features include macroeconomic variables collected from the World Bank data, representing crucial aspects of the economy, such as GDP growth, Consumer Price Index, broad money, foreign direct investment, interest rate, etc. In total, we selected 55 important macroeconomic indicators extracted from the World Bank dataset, considering them as features for forecasting. Most of the data is reported quarterly or annually, so we employed the cubic spline interpolation method to convert them into monthly data while preserving the characteristic features of the variables.

The dataset utilized in our study was collected over a period of twenty years, spanning from December 2001 to December 2021, and comprises a total of 252 data points. Out of these, 33 months were identified as having financial bubbles. To facilitate the training and testing of our model, the data was partitioned into two main sets - the training set and the test set. The time chosen to split the dataset was January 2018, ensuring that the proportion of months with bubble phenomena in the training and test sets was 2/3 and 1/3, respectively. The objective here is to ensure that the training and testing datasets contain sufficient data for model development and evaluation.

In financial datasets, class imbalance is a commonplace phenomenon where the occurrence of financial bubbles is relatively infrequent compared to non-bubble periods. Our dataset also exhibited class imbalance, and to mitigate this issue, we employed the Synthetic Minority Over-sampling Technique (SMOTE). This approach was chosen due to its suitability for financial datasets where rare class occurrences are prevalent. SMOTE generates synthetic samples for the minority class, addressing the imbalance while preserving the data distribution characteristics. In financial markets, the accurate detection of financial bubbles is of utmost importance, and SMOTE assists in reducing bias, improving model generalization, and enhancing performance metrics such as precision and recall. This technique ensures that the model can reliably identify critical financial bubble events with greater accuracy.

Given the presence of 55 macroeconomic indicator features in our dataset, we recognized the need to adopt a judicious strategy to mitigate the risk of overfitting. In this regard, we employed Principal Component Analysis (PCA) as a systematic means to condense the dimensions of our dataset. By transforming the original features into a smaller set of principal components, we retained the critical information essence while addressing the perils associated with excessive dimensions. This approach not only provides a practical solution to tackle the dimensionality conundrum but also serves as a catalyst for enhancing the efficacy of our machine learning model training, thereby augmenting prediction accuracy. It is important to note that during the PCA process, feature scaling is applied through standardization using the StandardScaler tool from scikit-learn.

The identification of financial bubbles in the VNINDEX series was carried out by using the 'psymonitor' package in R to execute the PSY procedure as proposed by Phillips, Shi, Caspi and Caspi [25]. Furthermore, for data analysis, Python and several associated packages such as Numpy, Pandas, Scikit-learn, and Seaborn [26-29] were utilized.

3.2. PSY Method for Bubbles Detection

Within the scope of our research, we utilize the Phillips, Shi, and Yu (PSY) methodology, which employs recursive regression techniques to investigate the presence of a unit root in the face of an alternative right-tailed explosive hypothesis. This technique, which was introduced by Phillips, Shi and Yu [30], has been tailored to identify multiple bubble periods in a time series dataset. Rejecting the null hypothesis in this test is regarded as empirical evidence of the existence of financial asset price bubbles. The critical values for these tests are determined through Monte Carlo simulations, and the outcomes of these tests aid in defining the start and end dates of the bubbles.

The objective is to compute statistical measures on the right tail of the Augmented Dickey-Fuller (ADF) test with regard to a time series. Through a comparison of the maximum values that are derived from the test statistics with corresponding threshold values obtained from the distribution, analysts can draw inferences regarding the explosiveness of the observed values. The null hypothesis assumes that a time series follows a random walk process with an extremely small drift coefficient, as expressed by the equation:

$$\Delta y_t = \mu + \sigma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-1} + e_t \tag{1}$$

Where y_{t-1} represents stock prices at time t; μ is the intercept; p is the maximum lag; ϕ_i are the regression coefficients corresponding to different lags; and e_t is the error term.

The recursive model is proposed as follows:

$$\Delta y_t = \alpha_{r_1,r_2} + \beta_{r_1,r_2} y_{t-1} + \Sigma_{\mathrm{i}=1} \phi^i_{r_1,r_2} \Delta y_{t-i} + \varepsilon_t, \varepsilon_t \sim^{i.i.d} N(0, \sigma^2_{r_1,r_2}) \tag{2}$$

When conducting the test, the authors assume a sample time frame of [0,1]. The symbols δ_{r_1,r_2} and ADF_{r_1,r_2} represent the estimated coefficients according to Equation (2) and the corresponding ADF test within the data window $[r_1, r_2]$. Let r_w denote the window size, so $r_w = r_2 - r_1$. The starting points r_1 can vary within the range $[0, r_2 - r_1]$.

The maximum ADF test statistic for the right tail is expressed by the following formula:

$$BSADF_{r_2}(r_0) = \sup_{\mathrm{r}_1 \in [0, \mathrm{r}_2 - r_0]} (ADF^{r_2}_{r_1} \} \tag{3}$$

The PSY procedure assumes that the errors in the regression model, represented by ϵ, have constant error variance. Based on this, the estimates of bubble periods using GSADF are determined by the following formulas:

$$r_e = \inf_{r_2 \in [r_0, 1]} (r_2 \colon BSADF_{r_2} > cv^{\beta_T}_{r_2}) \tag{4}$$

$$r_f = \inf_{r_2 \in [r_e, 1]} (r_2 \colon BSADF_{r_2} > cv^{\beta_T}_{r_2}) \tag{5}$$

Where $cv^{\beta_T}_{r_2}$ is the $100(1 - \beta_T)\%$ critical value of the SADF statistic based on observations T, r_2.

$BSADF_{r_0}$ with $r_2 \in [r_0, 1]$ is the delayed SADF statistic related to the GSADF statistic by the following relationship:

$$BGSADF(r_0) = \sup_{r_2 \in [r_0, 1]} (BSADF_{r_2}(r_0)) \tag{6}$$

In the PSY method, r_0 represents the minimum size, r_1 is the starting point, and r_2 is the ending point for each sample. The starting point r_1 is fixed, and r_2 varies from 0 to $r_2 - r_0$.

3.3. Machine Learning Methods to Predict Financial Bubbles

3.3.1. Logistic regression

Logistic regression is a widely used statistical method for predictive modelling in scenarios where the response variable is binary. In the present study, the focus is on predicting financial distress. Based on the input features, the model generates a probability estimate of financial distress. This probability is computed through equation (7).

$$P_n(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}} \tag{7}$$

Logistic regression is a commonly employed benchmark in research studies that investigate different forecasting methods. One of the key advantages of this technique is its ease of interpretation, which makes it accessible to most users. Therefore, logistic regression is frequently utilized in practical applications within financial institutions, owing to its high explanatory power.

3.3.2. Support vector machine

Support vector machines (SVMs) are a type of machine learning algorithm that relies on the concept of defining hyperplanes to partition observations in high-dimensional feature spaces. Linear SVM models prioritize the maximization of the margin between positive and negative hyperplanes. The classification decision is made using equation (8).

$$\mathrm{y_i} = \begin{cases} +1 \text{ if } \mathrm{b} + \alpha^{\mathrm{T}} x \geq +1 \\ -1 \text{ if } \mathrm{b} + \alpha^{\mathrm{T}} x \leq -1 \end{cases} \tag{8}$$

where b is the bias.

In cases where the relationship between features and outcomes is nonlinear, a kernel function is employed to transform the features into a higher dimensional space. A commonly used kernel function is the Gaussian radial basis function, which is expressed as equation (9).

$$K(x, x_i) = \exp\left(-\gamma\left||x - y|\right|^2\right) \tag{9}$$

One of the strengths of Support Vector Machines (SVMs) is their ability to avoid overfitting with small sample sizes and to remain robust to unbalanced distributions.

3.3.3. Decision tree

Decision tree algorithms use a tree structure to extract insights from data and derive decision rules. The algorithm determines the optimal allocation for each split with maximum purity, using a measure such as the Gini coefficient or Entropy. The root node, representing the most distinguishing attribute, is located at the top of the tree, while the leaf nodes represent classes that correspond to the remaining attributes.

One of the main advantages of decision tree models is that they are intuitive and easy to interpret. However, there is a risk of overfitting during the feature domain division or branching process, which represents a potential drawback.

3.3.4. Random forest

The random forest technique, which was developed by Breiman [31], is based on the decision tree model. In random forests, decision trees are grown using a subset of randomly selected features. This random selection of both samples and features ensures the diversity of the basic classifiers. The forest is constructed from multiple subsets that generate an equal number of classification trees. The preferred class is identified based on a majority of votes, which results in more precise forecasts and, importantly, protects against overfitting the data.

3.3.5. Gradient boosting (GB)

Gradient boosting is a machine learning technique commonly used in regression and classification tasks. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient Boosting is a robust boosting algorithm that combines several weak learners into strong learners. In this technique, each new model is trained to minimize the loss function, such as mean squared error or cross-entropy, of the previous model using gradient descent. During each iteration, the algorithm calculates the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and this process is repeated until a stopping criterion is met. This method has been proven to be effective in improving the prediction accuracy of models.

3.3.6. Artificial neural network

An artificial neural network, also known as a neural network, is a machine learning algorithm that is designed based on the structure and connections between neurons in the brain. This algorithm is capable of solving complex problems by mimicking the brain's structure. An artificial neural network consists of many layers of artificial neurons, which are connected to one another. Each layer is composed of an input layer, output layer, and hidden layer. These artificial neurons simulate the role of real neurons through mathematical models. Each artificial neuron receives an input signal, $x_1, x_2, \ldots, x_j$, consisting of binary numbers (0 or 1). It then calculates the weighted sum of the signals it receives based on their weights, $w_1, w_2, \ldots, w_j$. The signal is only transmitted to the next artificial neuron when the sum of the weights of the received signals exceeds a certain threshold. An artificial neuron can be represented as the equation (10).

$$\mathrm{y_i} = \text{output} = \begin{cases} 0 \text{ if } \sum_{\mathrm{j}} w_j x_j \leq \text{threshold} \\ 1 \text{ if } \sum_{\mathrm{j}} w_j x_j > \text{threshold} \end{cases} \quad (10)$$

Based on historical data, neural network optimization is performed by determining the appropriate weights and activation thresholds. This process involves training the network using a set of input data and corresponding output data, which is used to adjust the weights and thresholds of the artificial neurons until the desired level of accuracy is reached. Through this iterative process, the neural network can learn to accurately predict outcomes for new input data.

3.4. Evaluation of model performance

To evaluate the performance of classification algorithms, it is important to avoid focusing on a single class. Instead, a more comprehensive approach is preferred, which involves analyzing multiple metrics. The metrics are accuracy, precision, and sensitivity (recall).

Accuracy measures the proportion of correct classifications in the evaluation data and is calculated as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (11)$$

Precision measures the proportion of true positives among the predicted positives and is calculated as follows:

$$Precision = TP / (TP + FP) \quad (12)$$

Sensitivity (recall) measures the proportion of positives that are correctly predicted and is calculated as follows:

$$Sensitivity\ (recall) = TP / (TP + FN) \quad (13)$$

In these equations, TP represents true positives, FP represents false positives, FN represents false negatives, and TN represents true negatives. By analyzing these metrics, the performance of different classification algorithms can be compared and evaluated.

In the case of imbalanced datasets, accuracy alone may not be a reliable metric for evaluating classification models. To provide a more comprehensive evaluation, the F1-score and the ROC curve are often used.

The F1-score is the harmonic mean of precision and sensitivity, ensuring that the F1-score is higher only when both components are higher. The F1 Score is calculated as follows:

$$F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (14)$$

The ROC curve plots the true positive rate against the false positive rate, with the horizontal coordinate representing the false positive rate and the vertical coordinate representing the true positive rate. To quantify the characteristics of the ROC curve, the area under the curve (AUC) is introduced. The AUC value represents the area of the shadow part at the lower right of the ROC curve. The larger the shadow area, the greater the AUC value and the closer the ROC curve is to the upper left, indicating that the classification model is more accurate.

4. Result and Discussion

4.1. Overview of the Vietnamese Stock Market

The Vietnamese stock market was established in July 2000 with the launch of the Ho Chi Minh City Stock Exchange. The market experienced significant growth from 2006 to 2007, thanks to foreign investment inflows following Vietnam's entry into the WTO and the enactment of the Securities Law in 2006. However, there was a sharp decline in 2008, followed by a recovery in 2009, leading to stable growth until 2016. From 2017 to the present, the Vietnamese stock market has been characterized by significant volatility, with the VNIndex surpassing the 1000-point mark in 2018, followed by a sharp decline in 2020, and continued fluctuations from 2021 to 2022. Vietnam Stock Market Capitalization accounted for 205.153 USD billion in July 2023, which is equivalent to 65% of GDP. There are over 1600 listed companies in the market, with the majority operating in the financial, real estate, and essential consumer goods sectors, accounting for over 80% of the total market capitalization.

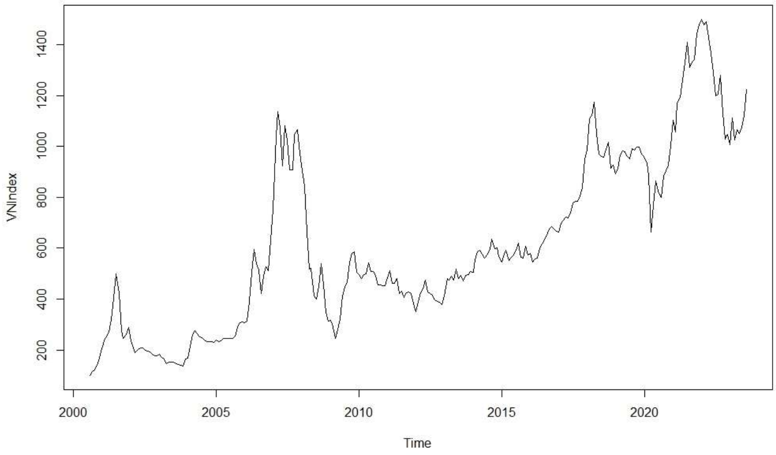


Figure 1. The Vietnamese Stock Market from 2001 to 2021. Source: Author (2023). Based on data from FiinPro (2023).

4.2. Results of Financial Bubble Detection

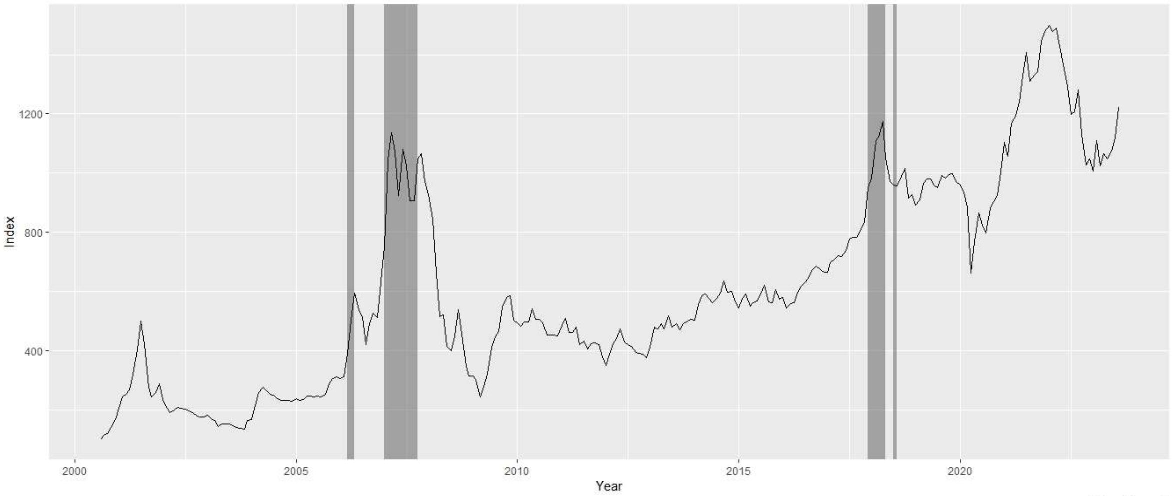
In this study, we employed the PSY procedure to detect the presence of financial bubbles in the Vietnamese stock market from January 2001 to December 2021, on a monthly basis. The VNINDEX was determined on the last trading day of each month. The PSY procedure has the minimum window that includes at least 38 observations, as determined by the rule of $tmin = 0.01T + 1.8\sqrt{T}$. The monitoring process starts from January 2001 onwards.

Table 1 displays the results of the financial bubble identification, providing the start and end dates for each bubble. The analysis identified a total of 33 months as having experienced financial bubbles, which are highlighted with dark gray vertical lines in Figure 2. Notably, two significant periods with multiple bubbles emerged during 2006-2007 and 2017-2018. These findings confirm the observations of financial experts regarding the overvaluation of stock prices in the stock market during these periods. Following the obtained results, we proceeded to label the months that were identified as bubble, thereby preparing the dataset for the subsequent stage of supervised machine learning.

Table 1. Statistics of Financial Bubbles Occurrences.

	Start Date	End Date
1	2006-02-28	2006-04-28
2	2006-06-30	2006-06-30
3	2006-12-29	2007-09-28
4	2017-11-30	2018-04-27
5	2018-06-29	2018-07-31

Source: Author (2023).



Notes: The solid line is the index and the shaded areas are the periods where the PSY statistic exceeds its 95% bootstrapped critical value.

Figure 2. Bubbles in Vietnam Stock market from January 2001 to December 2021.

4.3. Results of Forecasting Financial Bubbles Using Machine Learning Algorithms

Our study involved the implementation of various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Neural Networks, Gradient Boosting, and Support Vector Machine (SVM). The following table summarizes the performance of these models:

The results from Table 2 show that both Random Forest and Neural Network models outperform other algorithms in terms of AUC, accuracy, and F1 score. The Neural Network model stands out with the highest AUC (0.968), accuracy (0.915) and sensitivity (1.00), indicating its proficiency in accurately classifying periods marked by financial bubbles. However, the relatively lower F1 score (0.75) suggests a potential trade-off between precision and recall. This observation implies that while the Neural Network excels in correctly classifying positive instances, it may benefit from adjustments to maintain a balance between precision and recall.

Table 2. The performance results of classifiers.

Algorithms	Hyper-Parameter	AUC	F1 score	Accuracy	Precision	Sensitivity
Neural Networks	hidden_layer_sizes =100, max_iter = 300, activation = "relu", solver = 'adam', alpha = 0.0001	0.968	0.750	0.915	0.600	1.000
Random Forest	max_depth=5, n_estimators= 50	0.953	0.800	0.894	1.000	0.667
Gradient Boosting	max_depth=3, learning_rate=0.1, n_estimators= 100	0.957	0.727	0.872	0.800	0.667
Logistic Regression	C = 1	0.943	0.700	0.872	0.700	0.700
Support Vector Machine	C = 1, kernel = 'rbf', class_weight = 'balanced'	0.965	0.696	0.851	0.800	0.615
Decision Trees	max_depth = 5	0.819	0.667	0.830	0.800	0.571

Source: Author (2023).

Besides, the Random Forest algorithm delivers commendable results with a high AUC (0.953) and perfect precision (1.00), indicating its capacity to identify periods marked by financial bubbles precisely. However, its slightly lower sensitivity (0.667) implies a higher likelihood of missing some actual instances of financial bubbles. This trade-off underscores the need to consider the practical implications of false positives and false negatives in real-world financial decision-making.

Decision Trees underperformed in comparison, with an AUC of 0.82 and lower F1 scores. This serves as a baseline for evaluating the machine learning models' performance.

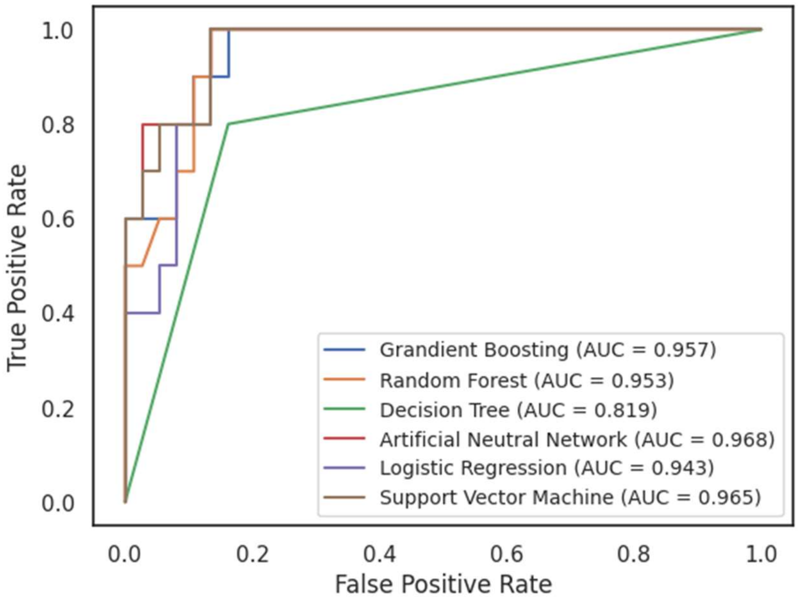


Figure 3. ROC curve of classifiers.

The AUC ROC graph also provides similar outcomes. Among the machine learning algorithms we employed, Artificial Neural Network emerged as the top-performing model with an AUC score of 0.968. This score reflects its ability to achieve a high true positive rate while keeping the false positive rate low. In essence, it excels at distinguishing bubble periods from non-bubble periods in the stock market. Other models, including Logistic Regression, and Decision Trees, demonstrated competitive but relatively lower AUC scores. While these models may still provide useful insights, their performance fell slightly behind that of Neural Networks, Random Forest, Gradient Boosting and SVM in this specific task.

Our research results differ from the findings of Başoğlu Kabran and Ünlü [14], which demonstrated that SVM yielded the best forecasting results for the S&P 500 index in the United States. We attribute this discrepancy to differences in the selected features incorporated into the models, as well as variations in the scale of the datasets used in the two studies. However, it is noteworthy that this study represents the first attempt to apply machine learning to forecast bubbles in an emerging market. In the future, we aspire to conduct empirical research across a broader range of markets.

Our research demonstrates the potential of machine learning algorithms in forecasting financial bubbles. The choice of the most suitable algorithm depends on the specific goals and constraints of the application, with Neural Network and Random Forest showing strong promise in this context. The results of this study have practical implications for financial policymakers, investors, and institutions. The ability to forecast financial bubbles can serve as an early warning system, enabling timely interventions to mitigate the adverse effects of market crashes. Policymakers can use these insights to develop more informed regulatory strategies, and investors can adjust their portfolios to reduce risks associated with bubble bursts.

In terms of future research directions, further studies could enhance the interpretability of forecasting results and analyze causal relationships to uncover the root causes of financial bubbles. Additionally, comparing and investigating the duration as well as the impact of financial bubbles on different asset markets is another promising research avenue.

5. Conclusions

In this study, we employed the PSY procedure to identify the presence of financial bubbles and forecast this phenomenon in the Vietnamese stock market using data from 2001 to 2021. The results indicated that financial bubbles occurred during the periods from 2006 to 2007 and from 2017 to 2018. Regarding predictive results, the Neural Network and Random Forest models exhibited superior forecasting performance with high F1 scores of 0.80 and 0.75, respectively. Based on the study's findings, policymakers, governments, and market regulatory agencies now have a valuable tool to detect and predict the emergence of financial bubbles on the real time basis, enabling them to formulate appropriate policies to mitigate their adverse effects. From an academic perspective, this research also opens up potential avenues for applying machine learning tools to prediction tasks in the field of economics.

Author Contributions: Assoc Prof. Duc Trung Nguyen conceived the idea and wrote the Introduction. Dr. Hoang Anh Le wrote Literature review. Dr. Kim Long Tran and Dr. Cap Phu Lieu wrote the Methodology, Results and discussions, and Conclusion.

Funding: The research topic was supported by The Youth Incubator for Science and Technology Programme, managed by Youth Promotion Science and Technology Center - Ho Chi Minh Communist Youth Union and Department of Science and Technology of Ho Chi Minh City. The contract number is "25/2022/ HĐ-KHCNT-VU" signed on 30th December, 2022.

Data Availability Statement: The data for this study can be found on our GitHub page: https://github.com/kimlongdhnh/long.tran.git (accessed on 02 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Galbraith, J.K., S. Hsu, and W. Zhang, *Beijing bubble, Beijing bust: Inequality, trade, and capital inflow into China.* Journal of Current Chinese Affairs, 2009. **38**(2): p. 3-26.
2. Shiller, R.J., *Bubbles, human judgment, and expert opinion.* Financial Analysts Journal, 2002. **58**(3): p. 18-26.
3. Kindleberger, C.P., R.Z. Aliber, and R.M. Solow, *Manias, panics, and crashes: A history of financial crises*. Vol. 7. 2005: Palgrave Macmillan London.
4. Stiglitz, J.E., *Symposium on bubbles.* Journal of economic perspectives, 1990. **4**(2): p. 13-18.
5. Case, K.E. and R.J. Shiller, *Is there a bubble in the housing market?* Brookings papers on economic activity, 2003. **2003**(2): p. 299-362.
6. Tirole, J., *Liquidity shortages: theoretical underpinnings.* Banque de France Financial Stability Review: Special Issue on Liquidity, 2008. **11**: p. 53-63.
7. Shiller, R.J., *Irrational exuberance*, in *Irrational exuberance*. 2015, Princeton university press.
8. Fama, E.F., *Two pillars of asset pricing.* American Economic Review, 2014. **104**(6): p. 1467-1485.
9. Shiller, R.J., *Do stock prices move too much to be justified by subsequent changes in dividends?* 1981.
10. West, K.D., *A specification test for speculative bubbles.* The Quarterly Journal of Economics, 1987. **102**(3): p. 553-580.
11. Phillips, P.C., Y. Wu, and J. Yu, *Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values?* International economic review, 2011. **52**(1): p. 201-226.
12. Homm, U. and J. Breitung, *Testing for speculative bubbles in stock markets: a comparison of alternative methods.* Journal of Financial Econometrics, 2012. **10**(1): p. 198-231.
13. Phillips, P.C., S. Shi, and J. Yu, *Testing for multiple bubbles: Limit theory of real-time detectors.* International Economic Review, 2015. **56**(4): p. 1079-1134.
14. Başoğlu Kabran, F. and K.D. Ünlü, *A two-step machine learning approach to predict S&P 500 bubbles.* Journal of Applied Statistics, 2021. **48**(13-15): p. 2776-2794.
15. Alessi, L. and C. Detken, *Identifying excessive credit growth and leverage.* Journal of Financial Stability, 2018. **35**: p. 215-225.
16. Beutel, J., S. List, and G. von Schweinitz, *Does machine learning help us predict banking crises?* Journal of Financial Stability, 2019. **45**: p. 100693.
17. Chatzis, S.P., et al., *Forecasting stock market crisis events using deep and statistical machine learning techniques.* Expert systems with applications, 2018. **112**: p. 353-371.
18. Ouyang, Z.-s. and Y. Lai, *Systemic financial risk early warning of financial market in China using Attention-LSTM model.* The North American Journal of Economics and Finance, 2021. **56**: p. 101383.

19. Shin, K.-S., T.S. Lee, and H.-j. Kim, *An application of support vector machines in bankruptcy prediction model.* Expert systems with applications, 2005. **28**(1): p. 127-135.
20. Zhao, Z., et al., *Investigation and improvement of multi-layer perceptron neural networks for credit scoring.* Expert Systems with Applications, 2015. **42**(7): p. 3508-3516.
21. Geng, R., I. Bose, and X. Chen, *Prediction of financial distress: An empirical study of listed Chinese companies using data mining.* European Journal of Operational Research, 2015. **241**(1): p. 236-247.
22. Barboza, F., H. Kimura, and E. Altman, *Machine learning models and bankruptcy prediction.* Expert Systems with Applications, 2017. **83**: p. 405-417.
23. Fuster, A., et al., *Predictably unequal.* The Effects of Machine Learning on Credit Markets. Revise & Resubmit in Journal of Finance, 2018.
24. Tran, K.L., et al., *Explainable machine learning for financial distress prediction: evidence from Vietnam.* Data, 2022. **7**(11): p. 160.
25. Phillips, P.C., et al., *Package 'psymonitor'.* Biometrika, 1984. **71**: p. 599-607.
26. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
27. McKinney, W. *Data structures for statistical computing in python*. in *Proceedings of the 9th Python in Science Conference*. 2010. Austin, TX.
28. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
29. Waskom, M., et al., *Mwaskom/Seaborn: V0. 8.1 (September 2017).* Zenodo, 2017.
30. Phillips, P.C., S. Shi, and J. Yu, *Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500.* International economic review, 2015. **56**(4): p. 1043-1078.
31. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.