UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF FINANCE AND BANKING

Final Project Machine learning and artificial intelligence in finance

Machine learning model predicts financial bubbles in the Vietnamese stock market

LECTURER: Master Phan Huy Tâm

STUDENT: Trần Thị Bình Nhi

STUDENT ID: K214140948

CLASS: K21414A

Ho Chi Minh city, Jun 17th 2024

TABELS OF CONTENTS

COMMENTS OF THE SUPERVISOR

I. **Theoretical basis**

1. Bubbles Financial:

A speculative bubble (sometimes also called a "market bubble", "price bubble") is a market phenomenon in which the price of a specific commodity (e.g. food or natural resource) or assets (stocks, real estate, foreign currencies) increase above their basic value, fueled by speculation and after peaking, there are consecutive sharp declines in price. This phenomenon is characterized by the collective belief of investors that the value of the asset will continue to increase indefinitely, leading to increased demand and even higher prices. When a financial bubble occurs, most participants have no doubts about the increase in the price of that asset due to the wave of speculation, but most believe that that is actually their value. .

A stock market bubble that forms in financial markets is a term applied to the self-propagating process of rising or rising stock prices of a particular industry or sector. A bubble occurs when speculators notice a rapid increase in the value of a stock and then decide to buy more of the same stock in anticipation of a further increase rather than because the stock is undervalued. These purchases result in many companies' shares being overvalued, creating a discrepancy between the stock's price and its actual value.PSY method for Bubbles Detection:

2. PSY method for detection bubbles:

The PSY method is a technique introduced by Phillips, Wu, and Yu (2011) that uses recursive regression techniques to investigate the presence of a unit root when faced with an explosive hypothesis. tail must be replaced. The rejection of hypothesis H0 in this test is considered empirical evidence of the existence of financial asset price bubbles. The critical values of these experiments are determined through Monte Carlo simulation, and the results of these experiments help determine the start and end dates of the market bubble.

The BSADF test is enhanced by Phillips, Shi, and Yu (2015) from the traditional Augmented Dickey-Fuller (ADF) test to help detect financial bubbles in time series by applying the ADF test on all possible sub-intervals of the time series, instead of just

applying it once to the entire time series. The BSADF test is used to detect explosive behavior in financial data and is capable of detecting multiple bubbles within the same sampling period.

The ADF test is used to test the existence of a unit root in a time series, that is, to test whether the time series is stationary or not. The regression equation for the ADF test is:

$$\Delta P_t = \alpha + \beta P_{t-1} + \gamma_1 \Delta P_{t-1} + \gamma_2 \Delta P_{t-2} + \cdots + \gamma_p \Delta P_{t-p} + \varepsilon_t$$

ΔP_t is the first-order difference in price at time t.

$\alpha, \beta, \gamma_1, \gamma_2, \ldots, \gamma_p$ are the regression coefficients.

ε_t is the error.

The BSADF test extends the ADF test by applying it to all sub-intervals of the time series, from start point r_1 to the end point r_2.

$$BSADF_{r_2}(r_0) = \max_{r_1 \in [0, r_2 - r_0]} \{ADF_{r1}^{r2}\}$$

$BSADF_{r_2}(r_0)$: BSADF statistical value at time r_2 with time interval length r_0

$r_1 \in [0,\ r_2 - r_0]$: Start interval r_1 ranges from 0 to $r_2 - r_0$

ADF_{r1}^{r2}: The ADF statistical value is calculated for the time period starting at r_1 and ending at r_2.

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

 This formula finds the maximum value of the ADF statistics in all sub-intervals starting at r_1 and ending at r_2, with r_1 ranging from 0 to $r_2 - r_0$. If the BSADF value at time r_2 is greater than the critical value, this indicates the possibility of a financial bubble at that time.

3. Random Forest:

Random Forest is one of the popular and powerful machine learning models, widely used in binary classification problems. Introduced by Leo Breiman in 2001, Random Forest is an ensemble learning algorithm that combines multiple decision trees to create a more powerful and stable model.

Random Forest creates a collection (forest) of decision trees from subsets of training data. Each decision tree in the forest is built from a random sample of the training data with bootstrap sampling. To increase diversity and reduce correlation between trees, each node of the decision tree uses only a random subset of features instead of the entire feature set. This gives the trees in the forest greater diversity, thereby improving the overall performance of the model. For a binary classification problem, each decision tree makes one prediction (0 or 1). Random Forest will aggregate predictions from all trees in the forest using a majority voting method to decide the final label of the input sample.

Random Forest is a powerful and flexible tool for binary classification problems. With the combination of multiple decision trees, it is capable of generating accurate and stable prediction models, although it requires high computational resources and is difficult to explain in detail.

4. Supper Vector Machine:

Support Vector Machine (SVM) is one of the popular and effective machine learning algorithms for binary classification problems. Introduced by Vladimir Vapnik and colleagues in the 1990s, SVM has proven useful and effective in many practical applications.

SVM finds a hyperplane in the feature space to separate samples belonging to two different classes. The goal is to find the optimal hyperplane with the largest distance (margin) to the nearest data points from each class, called support vectors. For data that is not linearly separable in the original space, SVM uses kernel functions to map the data to a higher feature space where the data is linearly separable. To handle cases where data cannot be completely separated, SVM allows some data points to lie within the margin distance or on the wrong side of the hyperplane. This is adjusted via the C parameter, which balances the width of the margin and the number of misclassified data points.

Support Vector Machine (SVM) is a powerful and flexible tool for binary classification problems. With the ability to find the optimal hyperplane and use kernel functions to process nonlinear data, SVM often gives accurate prediction results and good

generalization ability. In addition, thanks to margin maximization, SVM has good generalization ability and minimizes overfitting.

5. Gradient Boosting:

Gradient Boosting is a machine learning method based on ensemble learning, which combines many weak decision trees (weak learners) to form a strong model (strong learner).

Gradient Boosting is an ensemble learning method, meaning it combines multiple weak models into a stronger model. Weak models are often simple decision trees, called decision stumps when there is only one decision node. Each decision tree in Gradient Boosting is built in series, the goal is to improve each tree compared to the previous tree. This process generates better trees step by step, based on the remaining prediction errors from previous trees. Gradient Boosting uses gradient descent technique to adjust the weights during training. It optimizes the loss function by moving in the negative direction of the gradient of the loss function.

Gradient Boosting typically delivers very high performance on binary classification problems when properly trained and tuned. It can handle large and imbalanced data sets well.

6. Stratified K-fold Cross Validation:

Stratified K-fold Cross Validation is a variation of K-fold Cross Validation, designed to ensure each fold (subset) is representative of the entire data set, especially in terms of the distribution of the target variable. This is useful when dealing with imbalanced data sets, where some classes may be less frequent than others.

In K-fold Cross Validation, the dataset is randomly partitioned into k equally (or nearly equally) sized folds. The model is trained k times, each time using k - 1 folds for training and the remaining fold for validation. The performance metric is averaged over the kkk folds to provide an overall estimate of model performance.

In Stratified K-fold Cross Validation, the process of dividing the data set into k sections are taken such that each section has the same class distribution as the entire data set. This helps each fold reflect the true diversity of the original data and helps evaluate model performance more reliably, especially when facing the problem of class imbalance.

The biggest strength of this method lies in its ability to maintain the class distribution within each fold, which is extremely important when working with imbalanced data. In this way, Stratified K-fold Cross Validation minimizes the risk of the model being biased due to a certain class being over-represented in the training or testing set. As a result, this method provides more stable and accurate performance estimates than conventional K-fold Cross Validation, where the class distribution may not be maintained uniformly within each fold.

II. Exploratory Data Analysis (EDA)

1. Describe the dataset:

Features	Description
Close	Closing price of the VN30 index by month from 2009 to 2024 (VND)
GDP	Gross domestic product ($USD): the total value of goods and services produced within a country.
CPI	Consumer Price Index (VND): the average change in prices paid by consumers for goods and services.
Interest Rate	Interest Rate (%): the central bank policy rate
M3	Money supply M3 (VND) (including cash, deposits, and financial instruments)
Unemployment Rate	Unemployment Rate (%): the percentage of the labour that is unemployed at a given date
Inflation	Inflation (%): the continuous increase in the general price level in an economy

Table 1. Describe table of features

To determine the market bubble in the Vietnamese stock market based on the closing price series of the VN30 index, I will use the Extended BSADF test developed by Phillips, Shi, and Yu (2015). The PSY procedure is applied to the monthly closing price series from 2009 to 2024, totaling 185 observations. The model parameters are defined as follows: the

minimum window size $r_0 = T * 0.01 + \frac{1.8}{\sqrt{T}}$ (where T is the length of the price series), a ADF lags of 2 determined by the BIC criterion, and the critical values are computed using the 'psymonitor' library, which is developed in R language. This library implements the PSY approach and has been proven effective in detecting financial bubbles and crises.

Next, I will incorporate key macroeconomic variables in Viet Nam to construct a financial bubble prediction model, including Gross Domestic Product (GDP), Consumer Price Index (CPI), Interest Rate, Money Supply M3 (M3), Unemployment Rate, and Inflation. The Inflation variable is sourced monthly from April 2008 to May 2024 from the Trading Economics website. The remaining variables are sourced quarterly from Q2 2008 to Q1 2024 from the Institute of Banking Technology Development Research (VNUHCM-IBT), and then employed the cubic spline interpolated to convert them into monthly data while preserving their characteristic features.

2. Statistics describe data:

	Close
count	185.000000
mean	782.739405
std	313.609008
min	240.110000
25%	542.520000
50%	647.360000
75%	994.730000
max	1537.590000

Table 2. Statistical table describing close price variables

The descriptive statistics table shows the closing prices of the VN30 index providing an overview of the fluctuations in closing prices. The mean value is higher than the median value indicating that the distribution is likely to be right skewed. A high standard deviation indicates that the closing price changed strongly. The percentiles show that the majority of values lie between 542.52 and 994.73, with some extreme values that may influence the mean.

	GDP	CPI	Interest Rate	M3	Unemployment rate	Inflation Rate
count	64.000000	64.000000	64.000000	64.000000	64.000000	194.000000
mean	56.088922	144.048438	7.128906	6831006.000	2.381094	0.440464
std	27.847768	28.964127	3.030268	4105981.000	0.367199	0.683317
min	17.770000	85.500000	4.000000	1295492.000	1.900000	-1.540000
25%	33.576000	128.000000	6.000000	2947152.000	2.187500	0.070000
50%	49.006500	146.700000	6.500000	6197769.000	2.275000	0.310000
75%	75.189500	167.500000	8.000000	10185360.000	2.432500	0.620000
max	121.211000	188.200000	15.000000	14417410.000	3.980000	3.910000

Table 3. Statistical table describing macroeconomic variables

GDP has a fairly wide distribution, with values from 17.77 to 121.21. A high standard deviation indicates large fluctuations in GDP data. CPI has a high average value (144.05) and wide distribution, from 85.50 to 188.20, showing large fluctuations in consumer prices. The macroeconomic variables in this descriptive statistics table show large fluctuations in their values. GDP and CPI have large fluctuations, interest rates and unemployment rates have smaller fluctuations. In particular, the M3 money supply has a very high mean and standard deviation, indicating significant fluctuations in the money supply. Inflation rates also show large fluctuations, with negative values indicating possible periods of deflation.

3. EDA:

Figure 1. Closing price chart over time

The VN30 index price chart shows a general growth trend from 2009 to 2023 with some periods of strong fluctuations. The price peaks in 2018 and 2022 reflect periods of

rapid growth, while the price declines in 2019 and early 2020 can be attributed to economic factors and exceptional events such as the pandemic. The rapid recovery after periods of price decline demonstrates the market's strong resilience.

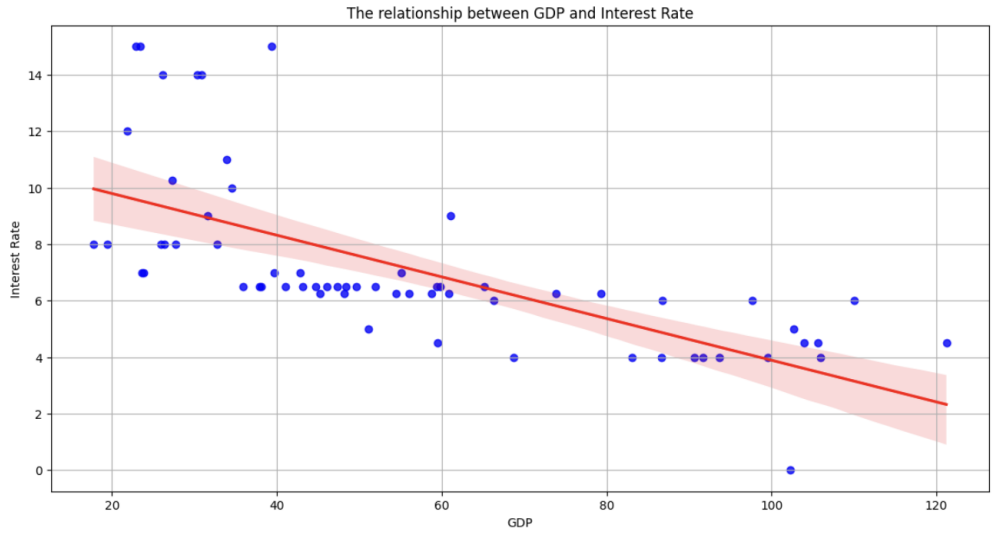

Figure 2. The relationship between GDP and Interest Rate.

The chart shows a clear inverse relationship between GDP and interest rates through the regression line. When GDP increases, interest rates tend to decrease. This can be explained by the fact that when the economy grows strongly (high GDP), the central bank can maintain low interest rates to promote investment and consumption. The data points are quite widely distributed, especially at low GDP levels, showing large fluctuations in interest rates when GDP is low, possibly due to the influence of other factors such as inflation or inconsistent monetary policy. stable. This inverse relationship has important implications in macroeconomics. It suggests that economic stimulus policies through interest rate reductions can boost GDP growth. At the same time, as the economy thrives, maintaining low interest rates could be a strategy to continue supporting growth.

Figure 3. GDP Growth Rate by Quarter

During 2020, GDP grew strongly in Q4/2020 possibly due to economic stimulus measures and recovery after the initial phase of the COVID-19 pandemic. But it dropped sharply to below zero in Q1 2021, possibly due to the lingering impact of the pandemic. However, the recovery was rapid in the following quarters, with significant growth returning in Q4 2021. In 2023, GDP growth continued to be volatile but tended to be more stable than in previous years. last year. After a sharp decline in Q1 2023, GDP has gradually recovered through the remaining quarters of the year, reaching positive growth in Q4 2023. The strong recovery after periods of decline shows the ability to recover. recovery of the economy.

III. Model and Evaluation:

1. Model:

Phase 1, identify financial bubbles using Phillips, Shi and Yu's BSADF test using the psymonitor library. The results determined that in the 185 observations of the data set from 2009 to 2024, there were 5 bubble points from November 2017 to March 2018. The bubble detection time was quite consistent with reality when during this period experts Financial experts commented on the overvaluation of stock prices in the stock market. 2018 stocks recorded the strongest fluctuation in 10 years since the 2008 global financial crisis and was also the first year the stock market declined. Then perform labeling, days with the presence of financial bubbles are assigned 1 and the remaining days are assigned 0. Finally, use the cubic interpolation method to convert the data to the macroeconomic changes.

In phase 2, it can be seen that data about the presence of imbalanced financial bubbles will affect the model's performance. Therefore, to limit the impact of imbalanced data, I used the Stratified K-fold Cross Validation method when dividing the train set and test set to ensure that each train and test set has the same class distribution as the test set. original data. At the same time, using the SMOTE method to generate new samples for the minority class helps improve performance on the minority class and minimize overfitting.

Finally, run the prediction model using Supper Vector Machine, Random Forest and Gradient Boosting with the processed data above, resulting in predicted values over 5 folds after a cross_validation process as follows:

```
Predictions for SVM:
Fold 1: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
Fold 2: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
Fold 3: [0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0]
Fold 4: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0]
Fold 5: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0]

Predictions for Random Forest:
Fold 1: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 2: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 3: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 4: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 5: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Predictions for Gradient Boosting:
Fold 1: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 2: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 3: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 4: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
Fold 5: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

Table 5. Predictions for three model

In this, SVM seems to predict some samples as positive (1) in each fold, with positive predictions concentrated in a small group of samples. Random Forest and Gradient Boosting appear to be more conservative with positive predictions, with only a small number of positive predictions per fold. The models predict differently for the samples in each fold's test data set. The ratio of values 0 and 1 shows the level of confidence and tendency of each model in predicting the positive class. To accurately evaluate the performance of each model, I will consider metrics such as accuracy, precision, recall, F1-score,...

2. Evluation:

	Accuracy	Precision	Recall	F1 Score	AUC	Gini
SVM	0.6548	0.6123	1.0000	0.6239	0.8919	0.7838
Random Forest	0.9674	0.4400	0.6000	0.4667	0.9944	0.9889
Gradient Boosting	0.9728	0.4500	0.6000	0.4800	0.7750	0.5500

Table 4. Evaluate model results

All three models achieved fairly high accuracy, but this does not reflect true performance on layer 1 due to unbalanced data. To evaluate the model accurately, I will look closely at other metrics, especially Precision, Recall, and F1-Score for class 1.

The SVM model has the highest Recall (100%), meaning this model does not miss any financial bubble cases. However, the relatively low Precision (61.27%) shows that this model can make many false positive predictions. AUC (0.9029) shows that this model has good ability to distinguish between classes. This is consistent with the goal of identifying financial bubbles, as the model can distinguish well between bubble and non-bubble cases. A high Gini index indicates good discrimination ability, consistent with the value AUC.

The Random Forest model has the highest AUC and Gini, but the low Recall (60%) shows that it misses 40% of financial bubble cases. Precision is also very low (41.00%), showing that this model is not suitable when the goal is to identify financial bubble cases well.

Boosting model has high accuracy (90.56%) but low Recall (60%), similar to Random Forest. This shows that the model also misses 40% of financial bubbles. There is a lower AUC and Gini, indicating that the ability to discriminate between classes is not as good as SVM and Random Forest.

Thus, with the goal of good prediction for the financial bubble layer (layer 1), the SVM model is the best choice even though there are many false positive predictions, but it ensures that no cases are missed. which financial bubble and it's highly discriminating power. This is important in situations where missing positive cases (financial bubbles) can have serious consequences.

IV. Conclusion:

After implementing the PSY method to determine when financial bubbles appear in the closing price series of the VN30 index and running the financial bubble prediction model, we have noticed some important results. The SVM model shows the ability to identify all cases of financial bubbles (High Recall), but at the same time there are many false positive predictions (False Positives), leading to many false warnings when in fact there are none. balloon. This may cause unnecessary reactions from users or investors.

The cause of these incorrect predictions may stem from a number of data limitations. Main macroeconomic variables can only be collected quarterly and a few monthly, leading to limitations in data detail and frequency. This may reduce the model's accuracy in identifying financial bubble patterns. Financial bubbles occur very rarely, leading to serious data imbalances. Although the SMOTE method has been applied to handle it, this problem has not been completely resolved.

To improve financial bubble prediction models, efforts should be made to collect more detailed macro data, especially monthly or weekly data if possible. This will help the model more accurately identify small fluctuations in the economy that can lead to financial bubbles. In addition, other variables related to financial bubbles can be added, such as investor sentiment index, transaction data of large investors, or political and economic events. important. Use more advanced data imbalance handling techniques, such as ADASYN (Adaptive Synthetic Sampling) or use deep learning algorithms with combined oversampling and undersampling techniques.

By applying the above suggestions, I hope to significantly improve the model's ability to predict financial bubbles, thereby providing more accurate and timely warnings to investors and users.