Volatility Prediction Model

JIAJUN LI

July 2024

1 Introduction

The rapid evolution of the cryptocurrency market has presented both unprecedented opportunities and significant challenges for traders and investors. Unlike traditional financial markets, cryptocurrencies are characterized by extreme volatility, driven by factors such as market sentiment, regulatory developments, technological advances, and macroeconomic trends. This volatility, while offering the potential for substantial profits, also poses substantial risks, making effective risk management and strategic planning paramount for market participants.

We try to measure and predict volatility in the high-frequency cryptocurrency market. We used the Btc market to investigate our assumption about a useful indicator to predict volatility. After testing several indicators, we found a highly correlated model, which can predict future volatility well in both test and train data. Additionally, we also introduce VPIN(Volume-Synchronized Probability of Informed Trading) from article [2] and adjust it based on testing different combinations of bucket length and number.

2 Data source

2.1 Data source

Our data are based on the 2023-2024.06 BTC market per minute. The dataframe has component start date, end date, open price, close price, highest price, lowest price, volume, amount, trades, and timestamp(unix). We used data from 2023 to train the model and tested it on the basis of the first half of 2024.

2.2 Variable

We specify all the variables used in the entire article. Here, the return value is defined as $r_t = \frac{p_{t-1}-p_t}{p_{t-1}}$, where p_t is the closing price at time t. The volatility from time i to $i+n$ is defined as $v_t = \sqrt{\frac{1}{n}\sum_{t=i}^{i+n}(v_t - \bar{v})}$. The definition of high minus low(hml) is $hml_t = h_t - l_t$, where h_t is the highest price in time t and l_t is the lowest price in time t.

2.3 Data problem

There is a missing data period from "2024-03-24 19:29:00" to "2023-03-24 22:00:00" due to the closing of the market. To find the problem of missing data, the panda function *dropna*() is a useful tool to drop the missing value.

3 Model

3.1 Model 1

This model only uses the previous volatility(60 minutes) to predict the next volatility(60 minutes). It is the most simple model that first come across everyone's mind. The model can be written as

$$v_t = \alpha + \beta * v_{t-1} + \epsilon_t. \tag{1}$$

This model is also called the AR(1) model in The Analysis of Multiple Stationary Time Series [3] .

3.2 Model 2

The most classical model for predicting volatility is the Garch model from article [1]. Here, we use Garch(1,1), which can be written as

$$r_t = \mu + \epsilon_t, \tag{2}$$

$$v_t^2 = \omega + \alpha * \epsilon_{t-1}^2 + \beta * v_{t-1}^2, \tag{3}$$

$$\epsilon_t = v_t * e_t, e_t \sim N(0,1), \tag{4}$$

where μ is a constant, e_t is random noise with distribution $e_t \sim N(0,1)$ and v_t is the volatility component varying over time. In practice, the short-term constant term μ is rarely known in advance and analysts often use a simplifying assumption of $\mu = 0$.

3.3 Model 3

This model uses three components. It is defined as

$$v_t = a + b * hml_{t-1} + c * volume_{t-1} + d * v_{t-1} + \epsilon_t. \tag{5}$$

The model was derived from basic intuition and several highly correlated variables. hlm is the sum of the minus between the highest price and the lowest price per minute, and the period for the sum is the same as volatility(60 minutes). This range is a direct measure of volatility during that period, capturing price fluctuation. Since volatility is often persistent over short time frames, using the previous period's high-low range helps predict future volatility. $volume_{t-1}$ is the sum of volume in 10 minutes. Trading volume is often correlated with price movement and volatility. Higher volume can signal increased trader activity, which often leads to larger price swings and thus higher volatility. By incorporating this, the model attempts to capture this relationship. v_{t-1} also uses 10 minutes as a period. Volatility tends to cluster, meaning that high volatility periods are often followed by more high volatility periods, and low volatility by low volatility. Using the volatility of the previous period as an input, the model leverages this temporal dependency.

3.4 Model 4

This model is modified based on model 3. Here, we introduce a new concept called VPIN(Volume-Synchronized Probability of Informed Trading), which was first introduced in the article [2]. They transformed the tick-by-tick classification into the possibility of buying or selling to capture the trend of price movement. VPIN uses CDF of the standard normal distribution of the standard price change in a certain volume-based period. The formula can be written as follows.

$$V_\tau^B = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot Z(\frac{P_i - P_{i-1}}{\sigma_{\Delta_P}}), and \tag{6}$$

$$V_\tau^S = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot [1 - Z(\frac{P_i - P_{i-1}}{\sigma_{\Delta_P}})] = V - V_\tau^B, \tag{7}$$

where $t(\tau)$ is the index of the last time bar included in the τ volume bucket, Z is the CDF of the standard normal distribution and σ_{Δ_P} is the estimate of the standard derivation of price changes between time bars. Then, the VPIN flow toxicity metric is defined as

$$VPIN = \frac{\sum_{\tau=1}^{n} |V_\tau^S - V_\tau^B|}{nV}, \tag{8}$$

where n is the bucket number.

The model can be written as

$$v_t = a + b * hml_{t-1} + c * volume_{t-1} + d * v_{t-1} + vpin_{t-1} + \epsilon_t. \tag{9}$$

It has four variables to predict the next volatility(60 minutes). The other three are the same as those in model 3.

3.5 Model 5

This model is also modified based on model 3. Here, we change the VPIN a little by simplifying the bucket volume to just each minute bar. Here, we can show its formula by

$$V_\tau^B = V_i \cdot Z(\frac{r_\tau - \mu}{\sigma_r}), and \tag{10}$$

$$V_\tau^S = V_i \cdot [1 - Z(\frac{r_\tau - \mu}{\sigma_r})], \tag{11}$$

where n is the time step for counting how many minute bars are need to predict. The way to compute adj-VPIN is similar in model 4. It is

$$adj - VPIN_t = \frac{\sum_{\tau=1}^{n} |V_\tau^S - V_\tau^B|}{nV}, \tag{12}$$

where n is *timestep* need to be chosen.

Thus, the model 5 can be written as

$$v_t = a + b * hml_{t-1} + c * volume_{t-1} + d * v_{t-1} + adj - vpin_{t-1} + \epsilon_t. \tag{13}$$

Figure 1: 2023 whole year close price

4 Empirical work

In this section, we divide the data into two parts. The first part is the whole 2023 Btc market treated as train data. The second part is the first half of the 2024 Btc market as test data. In the train part, R square, adjust R square, and the p-value for each coefficient are the measure of the performance of the model and the potential prediction fitness. In the test part, we use model 1 as the basic model and correlation to compare each model.

4.1 Train data

The price change for the entire 2023 is shown in Figure1 and the return value is shown in Figure2. The 2023 Bitcoin market experienced a strong bullish trend with significant price increases from the beginning of the year, peaking toward the end. The market demonstrated typical cryptocurrency volatility, with substantial fluctuations in daily returns. Despite some periods of corrections and dips, overall market sentiment remained positive, leading to substantial gains at year-end.

4.1.1 Model 1

The real and prediction of volatility for Model 1 is shown in Figure4. In Figure 3, Model 1 serves as a basic model with an R-squared of 0.346, indicating that it explains approximately 34. 6 % of the variance in the dependent variable, volatility. The model is statistically significant, with both the F-statistic and p-values showing strong significance. However, the residuals exhibit high skewness and kurtosis, suggesting nonnormality and potential problems with outliers. Although the model generally tracks the real volatility well, it tends to underestimate extreme values, indicating room for improvement as we compare it with more complex models.

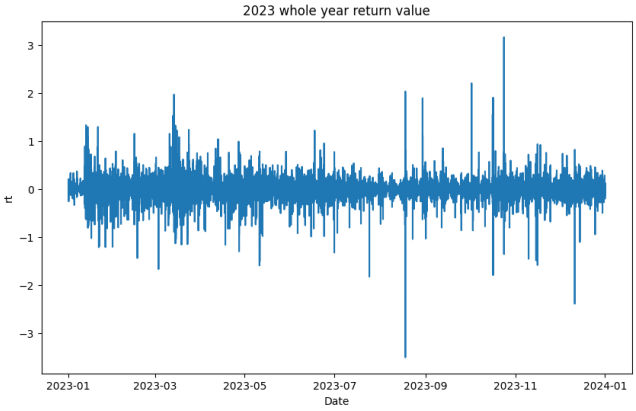

Figure 2: 2023 whole year return value

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.346
Model:	OLS	Adj. R-squared:	0.345
Method:	Least Squares	F-statistic:	4624.
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	0.00
Time:	16:12:56	Log-Likelihood:	18622.
No. Observations:	8759	AIC:	-3.724e+04
Df Residuals:	8757	BIC:	-3.723e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0190	0.001	37.719	0.000	0.018	0.020
volatility_t0	0.5878	0.009	67.997	0.000	0.571	0.605

Omnibus:	10275.398	Durbin-Watson:	2.186
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2818557.839
Skew:	5.838	Prob(JB):	0.00
Kurtosis:	90.101	Cond. No.	28.1

Figure 3: The summary of Model 1

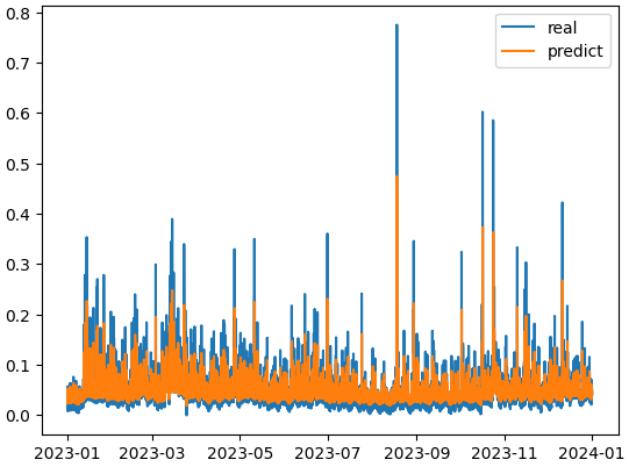

Figure 4: Model 1 real and predicting return value

4.1.2 Model 2

The results of the model summary can be viewed in Figure5. Although the GARCH (1,1) model is theoretically sound for predicting volatility, in this case it appears to be a poor fit for the data, as indicated by the very low value of the R square and the problematic residual statistics (high skewness and kurtosis). The model does not capture volatility dynamics effectively and underperforms compared to Model 1. This might be due to the nature of the data or the need for a more complex model structure, possibly with additional predictors or higher-order GARCH terms. This model is different from the other four models, since it needs to take the square root of the result to get the volatility. If we take the square root of the results, we get R square with 0.123, which is also not as good as model 1.

4.1.3 Model 3

The results of model 3 can be seen in Figure6 and Figure 7. Model 3, which incorporates the high-low price range (60 minutes), the trading volume (10 minutes) and the past volatility (10 minutes) as predictors, demonstrates a higher explanatory power compared to Model 1, with an R-squared value of 0.405. This indicates that it captures more of the variance in volatility. However, while the model does better at explaining the variability in volatility, the residuals still show signs of non-normality, with high skewness and kurtosis. The prediction graph reveals that while Model 3 tracks the general trends in real volatility, it underestimates the volatility spikes, particularly during periods of high market activity. This suggests that although Model 3 provides a better fit overall, it still struggles to capture extreme market movements effectively, a limitation that might be addressed with further refinement or additional predictors.

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.067
Model:	OLS	Adj. R-squared:	0.066
Method:	Least Squares	F-statistic:	312.7
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	6.80e-132
Time:	18:31:55	Log-Likelihood:	27320.
No. Observations:	8759	AIC:	-5.463e+04
Df Residuals:	8756	BIC:	-5.461e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0026	0.000	22.150	0.000	0.002	0.003
return_square_t0	16.1650	1.411	11.458	0.000	13.400	18.930
volatility_t0	0.1234	0.014	8.901	0.000	0.096	0.151

Omnibus:	22917.592	Durbin-Watson:	2.037
Prob(Omnibus):	0.000	Jarque-Bera (JB):	704333946.894
Skew:	30.331	Prob(JB):	0.00
Kurtosis:	1390.884	Cond. No.	1.23e+04

Figure 5: The summary of Model 2

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.405
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	1987.
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	0.00
Time:	18:15:54	Log-Likelihood:	19039.
No. Observations:	8759	AIC:	-3.807e+04
Df Residuals:	8755	BIC:	-3.804e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0192	0.000	40.852	0.000	0.018	0.020
highlow_t0	1.772e-05	5.1e-07	34.737	0.000	1.67e-05	1.87e-05
volume_sum_t0	4.089e-06	3.28e-07	12.463	0.000	3.45e-06	4.73e-06
volatility_t0	0.1889	0.014	13.875	0.000	0.162	0.216

Omnibus:	10934.991	Durbin-Watson:	2.085
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3610147.354
Skew:	6.507	Prob(JB):	0.00
Kurtosis:	101.603	Cond. No.	7.63e+04

Figure 6: The summary of Model 3

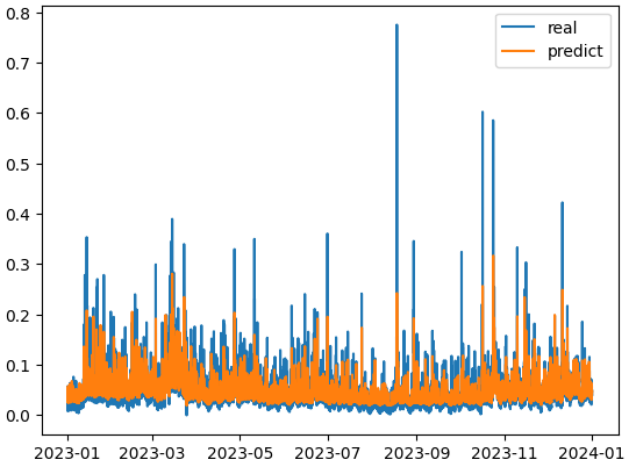

Figure 7: Model 3 real and predicting return value

4.1.4 Model 4

Model 4 builds on the foundation of Model 3 by introducing a new variable, VPIN (Volume-Synchronized Probability of Informed Trading), which is intended to capture the likelihood of informed trading based on volume and price movements. This addition aims to improve the model's predictive power for volatility.

The results of Model 4 based on the volume of the bucket 120 and the number of buckets 10 are shown in Figure8 and Figure9. Here, we choose 120 since the volume for the whole day is nearly 4000 and we take 5% of it. Model 4 shows a slight improvement over Model 3 in terms of explanatory power, as indicated by the higher R-squared value. The inclusion of VPIN adds a new dimension to the model, capturing aspects of informed trading that could impact volatility. However, the model continues to face challenges in accurately predicting extreme volatility spikes, and the residuals still exhibit significant non-normality. Although Model 4 offers some improvements, the overall improvement is marginal and more refinement could be needed to better capture the complexities of market volatility. After changing the volume and number of boxes, we found that a lower volume and higher bucket numbers will produce better results, which helps us introduce model 5.

4.1.5 Model 5

This model was induced from model 4. The results of Model 5 are shown in Figure10 and Figure11. Model 5, with its simplified VPIN adjustment, offers a modest improvement over Model 4 in terms of predictive power, as evidenced by the higher R-squared and F-statistic values. The model appears to balance the complexity introduced by VPIN with the need for accurate minute-level predictions.

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.411
Model:	OLS	Adj. R-squared:	0.410
Method:	Least Squares	F-statistic:	1524.
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	0.00
Time:	18:31:03	Log-Likelihood:	19080.
No. Observations:	8759	AIC:	-3.815e+04
Df Residuals:	8754	BIC:	-3.811e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0161	0.001	27.547	0.000	0.015	0.017
highlow_t0	1.482e-05	6.01e-07	24.658	0.000	1.36e-05	1.6e-05
volume_sum_t0	3.191e-06	3.41e-07	9.344	0.000	2.52e-06	3.86e-06
vpin_t0	0.0337	0.004	9.025	0.000	0.026	0.041
volatility_t0	0.1408	0.015	9.671	0.000	0.112	0.169

Omnibus:	11043.235	Durbin-Watson:	2.092
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3755902.409
Skew:	6.622	Prob(JB):	0.00
Kurtosis:	103.578	Cond. No.	8.24e+04

Figure 8: The summary of Model 4

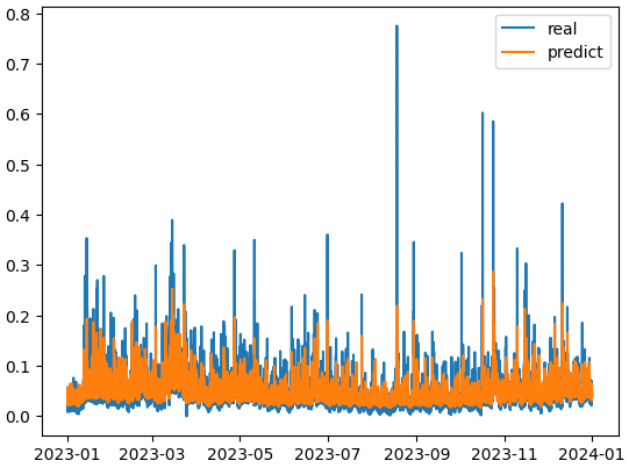

Figure 9: Model 4 real and predicting return value

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.428
Model:	OLS	Adj. R-squared:	0.428
Method:	Least Squares	F-statistic:	1639.
Date:	Sat, 10 Aug 2024	Prob (F-statistic):	0.00
Time:	16:51:22	Log-Likelihood:	19206.
No. Observations:	8755	AIC:	-3.840e+04
Df Residuals:	8750	BIC:	-3.837e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0066	0.001	8.185	0.000	0.005	0.008
highlow_t0	6.436e-06	7.78e-07	8.275	0.000	4.91e-06	7.96e-06
volume_sum_t0	3.94e-06	3.22e-07	12.245	0.000	3.31e-06	4.57e-06
vpin_t0	0.0012	6.31e-05	18.939	0.000	0.001	0.001
volatility_t0	0.1572	0.013	11.688	0.000	0.131	0.184

Omnibus:	11198.379	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4020634.786
Skew:	6.790	Prob(JB):	0.00
Kurtosis:	107.102	Cond. No.	7.69e+04

Figure 10: The summary of Model 5

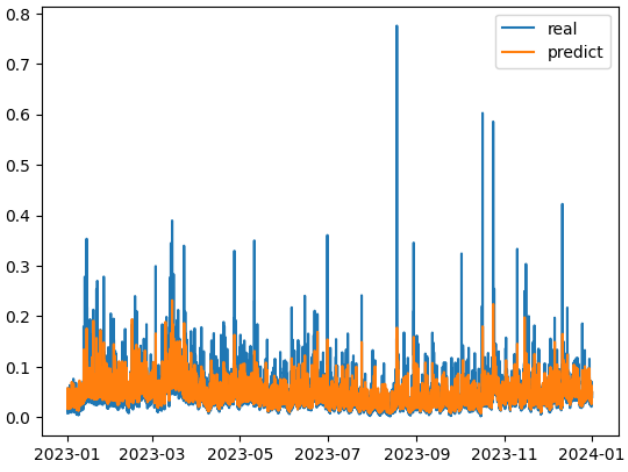

Figure 11: Model 5 real and predicting return value

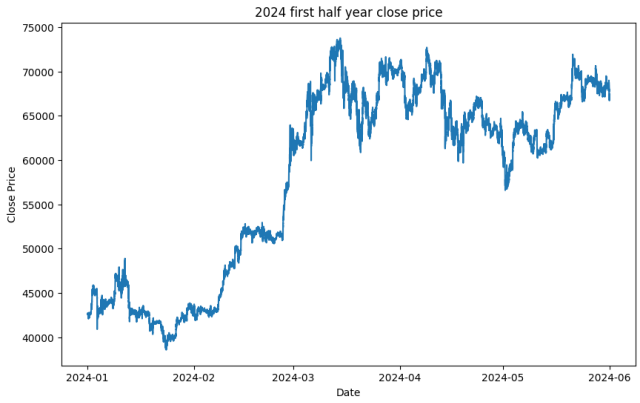

Figure 12: 2024 first half year close price

4.1.6 Comparison

Model 5 demonstrates the best performance with the highest R-squared and Adjusted R-squared values, explaining 42.8 % of the variance in volatility. Each successive model from Model 1 to Model 5 shows incremental improvements in explanatory power, with Model 2 being the least effective. The addition and refinement of variables, particularly the introduction and adjustment of VPIN, contribute to these improvements. All models maintain statistical significance for their included variables, indicating that the chosen predictors are relevant in each context.

Train Model Test			
Model Type	R square	Adjust R square	number of p value(< 0.05) / number of variables
Model 1	0.346	0.345	2/2
Model 2	0.123	0.123	2/2
Model 3	0.405	0.405	3/3
Model 4	0.411	0.410	4/4
Model 5	0.428	0.428	4/4

4.2 Test Data

The price change for the first half of 2024 is shown in Figure12 and the return value is shown in Figure13. The first half of 2024 has been a period of strong growth and volatility for Bitcoin. The market saw a substantial rally in the first quarter, driven by bullish sentiment, followed by a period of price fluctuations and eventual stabilization. Return data reflect market volatility, with several spikes indicating periods of intense trading activity. In mid-2024, the BTC market appeared to be consolidating at a higher price level.

In test data, we used data from 30 days ago to calculate the parameters of the ordinary linear model and used them to predict the volatility per hour the next day. We used the correlation between prediction and the real one to measure the performance of each model.

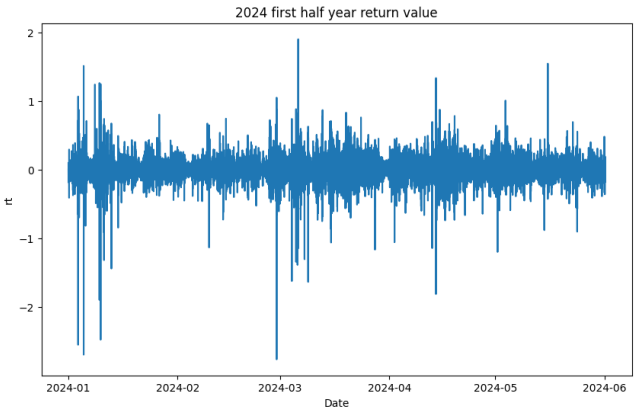

Figure 13: 2024 first half year return value

4.2.1 Comparison

The correlation analysis confirms that Model 5 provides the most accurate predictions of hourly volatility, followed closely by Models 3 and 4. Model 1, while solid as a baseline, lags behind these more complex models, and Model 2 performs the weakest. The incremental improvements in correlation from Model 3 to Model 5 suggest that the enhancements and additional predictors in these models successfully capture more of the volatility dynamics in the Bitcoin market. Therefore, Model 5 is recommended for its superior ability to predict next-day hourly volatility in the 2024 BTC market.

Test Model	
Model Type	$\rho(v_{real}, v_{predict})$
Model 1	0.677988
Model 2	0.637146
Model 3	0.725566
Model 4	0.726023
Model 5	0.729986

5 Conclusion

Through extensive analysis and comparison, Model 5 emerges as the most effective model to predict hourly volatility in the Bitcoin market, based on both training and testing datasets. Model 5, which integrates a refined version of VPIN, demonstrates the highest correlation between predicted and actual volatility, indicating its superior ability to capture the dynamics of the BTC market. The incremental improvements observed from Model 1 to Model 5 suggest that incorporating additional predictors, particularly those related to informed trading activity, significantly enhances the model's predictive power. Although Model 2 (GARCH) performs poorly in this context, the other models show varying degrees of effectiveness, with Model 5 leading the way. This study underscores the importance of model refinement and the inclusion of relevant market indicators in improving volatility predictions, which is crucial for

traders and investors navigating the volatile cryptocurrency landscape. For future improvement, a better indicator of sudden and massive volatility must be introduced. This indicator might be found in other markets or countries' policies.

References

[1] Tim Bollerslev. *Generalized Autoregressive Conditional Heteroskedasticity*. Apr. 1986. URL: https://www.sciencedirect.com/science/article/pii/0304407686900631 (visited on 09/08/2024).

[2] David EasleyMarcos M. López de PradoMaureen O'Hara. *Flow Toxicity and Liquidity in a High-frequency World*. Mar. 2012. URL: https://academic.oup.com/rfs/article-abstract/25/5/1457/1569929?redirectedFrom=fulltext (visited on 07/08/2024).

[3] P. Whittle. *The Analysis of Multiple Stationary Time Series*. 1953. (Visited on 11/08/2024).