Stablecoin Depegging Risk Prediction

Yi-Hsi Lee[1]
Department of Financial Engineering and Actuarial Mathematics, Soochow University, Taiwan

Yu-Fen Chiu
Department of Financial Engineering and Actuarial Mathematics, Soochow University, Taiwan

Ming-Hua Hsieh[c,]
Department of Risk Management and Insurance, National Chengchi University, Taiwan

Abstract

This paper extensively reviews empirical literature on stablecoins, systematically identifying key variables that could lead to depegging risks. Based on this, we construct predictive models using three machine learning algorithms (logistic regression, random forest, and XGBoost) that can accurately and timely predict stablecoin depegging events. Our main subjects of study are the top four stablecoins in daily trading volume: USDT, USDC, BUSD, and DAI. Unlike previous literature that used static depegging threshold values, we adopt a dynamic threshold adjusted for trading volume as the criteria for depegging. In addition to traditional on-chain price and volume data, this study is the first to incorporate sentiment indicators from news sources. The empirical period covers from January 1, 2022, to December 31, 2023. The results show that, as indicated in the literature, significant price and volume fluctuations of mainstream cryptocurrencies (BTC and ETH) indeed cause stablecoin depegging. Furthermore, measures of instability from past literature also provide longer-term early warning

[1] Corresponding author.
E-mail address: eclee@scu.edu.tw

effects for stablecoin depegging. However, surprisingly, the sentiment indicators used in this study did not show a significant early warning effect for our research subjects. The models constructed in this study enable crypto asset investors to timely predict the risk of stablecoin depegging and make corresponding investment decisions, thereby reducing investment risks.

JEL classification: G12, G14, G15, G16

Keywords: Stablecoins, Depegging, Machine Learning

Stablecoin Depegging Risk Prediction

Abstract

This paper extensively reviews empirical literature on stablecoins, systematically identifying key variables that could lead to depegging risks. Based on this, we construct predictive models using three machine learning algorithms (logistic regression, random forest, and XGBoost) that can accurately and timely predict stablecoin depegging events. Our main subjects of study are the top four stablecoins in daily trading volume: USDT, USDC, BUSD, and DAI. Unlike previous literature that used static depegging threshold values, we adopt a dynamic threshold adjusted for trading volume as the criteria for depegging. In addition to traditional on-chain price and volume data, this study is the first to incorporate sentiment indicators from news sources. The empirical period covers from January 1, 2022, to December 31, 2023. The results show that, as indicated in the literature, significant price and volume fluctuations of mainstream cryptocurrencies (BTC and ETH) indeed cause stablecoin depegging. Furthermore, measures of instability from past literature also provide longer-term early warning effects for stablecoin depegging. However, surprisingly, the sentiment indicators used in this study did not show a significant early warning effect for our research subjects. The models constructed in this study enable crypto asset investors to timely predict the risk of stablecoin depegging and make corresponding investment decisions, thereby reducing investment risks.

JEL classification: G12, G14, G15, G16

Keywords: Stablecoins, Depegging, Machine Learning,

1. Introduction

The rapid development of crypto assets can be attributed to their unique technological characteristics and increasing market acceptance. However, the volatility of the crypto asset market has led to a demand for stablecoins, a type of cryptocurrency backed by traditional assets (such as fiat currency) or other crypto assets. The core value of stablecoins lies in their ability to provide price stability, which is crucial for the widespread application of crypto assets. Stablecoins bridge traditional finance and cryptocurrency markets, allowing funds to move more smoothly between the two areas and providing a haven during the volatility of crypto assets, hereby enhancing the overall efficiency and acceptability of the crypto asset ecosystem.

Stablecoins were first introduced in 2014, but their actual application and importance in the market did not begin to significantly increase until 2017 with the rise of USDT (Ante et al., 2023). According to statistics from the CoinGecko platform[1] at the end of 2023, there are a total of 98 major circulating stablecoins in the market. The total market capitalization of stablecoins ranks fifth in the overall cryptocurrency asset market value (approximately $135 billion). The two largest stablecoins, USDT and USDC, rank third (with a market cap of about $95 billion) and seventh (with a market value of about $25.5 billion) respectively in the overall cryptocurrency market value ranking. When considering the daily trading volume, USDT and USDC rank first (with a trading volume of about $34 billion) and fourth (with a trading volume of about $7.8 billion) respectively among cryptoassets. This sufficiently demonstrates the importance of stablecoins in the cryptocurrency ecosystem.

Although the primary goal of stablecoins is to maintain the stability of their price with the pegged currency through various design structures (d'Avernas et al., 2021;

[1] https://www.coingecko.com/ and https://www.coingecko.com/en/categories

Gadzinski et al., 2023; Hafner et al., 2023; Jarno & Kołodziejczyk, 2021; Lyons & Viswanath-Natraj, 2023), in recent years, stablecoins have frequently deviated from their pegged value, including the IRON depegging event in 2021 (Clements, 2021), the UST value collapse in 2022 (Uhlig, 2022), and the USDR depegging event in 2023, which has yet to return to its pegged value. Additionally, the market-representative USDT, USDC, and BUSD have also experienced brief depegging due to financial transparency issues and the collapse of Silicon Valley Bank. (De Blasis et al., 2023) found that the depegging of stablecoin UST could even lead to instability in other stablecoins such as BUSD, DAI, USDT, USDC, and the cryptocurrency BTC. These cases and articles show the potential instability of stablecoins and pose serious challenges to market participants. Furthermore, past research and reports on stablecoins have primarily focused on studies of (in)stability, often mentioning the term 'depegging', but without providing a concrete definition.

Despite being called stablecoins, their instability or depegging events have repeatedly shocked the DeFi ecosystem, showing that developing and maintaining high-quality stablecoins is a key challenge in crypto finance (S&P Global, 2023). Before this, there is an urgent need for a timely and effective early warning model for stablecoin depegging, which in the short term can help in the early response to crypto asset investment decisions; in the long term, it can aid in understanding and monitoring the flaws or vulnerabilities in the issuance and operational framework of stablecoins. A timely and effective financial risk management system is crucial for the stability of the financial ecosystem, particularly evident in the still-developing crypto asset and decentralized finance (DeFi) markets (Giokas et al., 2023).

To this end, this, the paper aims to extensively review empirical literature on stablecoins and systematically identify key variables that could lead to depegging risks.

These variables include on-chain price and volume data, as well as instability measures and sentiment indicators proposed in the literature. Furthermore, it integrates multiple risk indicators to form an automated, integrated quantitative model. This benefits participants in the stablecoin ecosystem by enabling them to effectively respond to the risks of stablecoin depegging, thereby enhancing the resilience of the stablecoin system.

The main contributions of this article are twofold: (1) While previous stablecoin research has primarily focused on the (in)stability of stablecoins, there has been limited investigation into the specific phenomenon of stablecoin depegging. This article seeks to concretely define depegging and develop predictive models for it. (2) It introduces instability measures and Sentiment Indicators into the study of stablecoin depegging, facilitating more timely and effective detection of depegging states.

The remainder of the paper is structured as follows: Section 2 reviews the relevant literature on stablecoins. Section 3 describes the data and models used. Section 4 presents the empirical analysis results. Finally, Section 5 concludes.

2. Literature Review

Ante et al. (2023) serves as an excellent starting point for a literature review on stablecoins. They employed a systematic literature review (SLR) methodology to examine academic literature related to stablecoins up to May 2022. Ultimately, from a pool of 2,957 articles, they filtered out 22 peer-reviewed English academic empirical articles. Their study encompassed the top twenty stablecoins at that time, statistically analyzing the research subjects of each document. For the selection of research subjects, they used a list sorted by the top twenty market capitalizations. They created a two-dimensional table based on Metrics (divided into Pricing/Returns, Market Cap/Supply, Trading Volume, and Blockchain Transaction) and Time Interval (Minutes, Hourly,

Daily, and Blocks) as well as Data Source (Market data aggregators, Cryptocurrency exchanges, Blockchain explorers, and Blockchain nodes/clients) to consolidate the empirical literature. Ultimately, they categorized the 22 empirical articles into three main clusters: Cluster 1: The stability and volatility of stablecoins; Cluster 2: The interrelation of stablecoins and crypto markets; and Cluster 3: The relationship of stablecoins with (non-crypto) macroeconomic factors. The primary data frequency in their surveyed empirical literature was Daily, with the main data source being Market Data Aggregators. The empirical Metrics (variables) primarily focused on Pricing/Returns, Market Cap/Supply, and Trading Volume. Moreover, the subjects of study were predominantly the top five by market value at the time of publication, mainly pegged to the US dollar and backed by cash or related collateral. This article, which surveys empirical academic literature, provides a solid foundation for choosing research targets, selecting explanatory variables, acquiring data sources, and deciding on data frequencies.

The literature suggests that the (in)stability of stablecoins is intrinsically linked to the design of their stabilization mechanisms. Bullmann et al. (2019) and European Central Bank (2019) utilized the Crypto-cube Framework to allow for the categorization of current stablecoins into four major types: (1) Tokenized Funds, (2) Off-chain Collateral, (3) On-chain Collateral, and (4) Algorithmic. Jarno & Kołodziejczyk (2021), utilizing Bullmann et al. (2019)'s stablecoin classification, analyzing 20 stablecoins includes the various stablecoin types from the investor's viewpoint. They found that stablecoins differ in their capacity to maintain stable market values, with tokenized funds, typically fiat-collateralized and often backed by USD, being the most effective. These funds ensure full backing by the reference currency and depend heavily on a trusted custodian for collateral. The study indicates that simpler

tokenized fund designs outperform more complex stablecoin structures in reducing volatility. Hafner et al. (2023) categorize stablecoins into four types based on two dimensions: the endogeneity or exogeneity of stablecoin collateral and whether the collateral management is centralized or decentralized. They further propose an agent-based model that employs simulations to assess the collateral and its management mechanisms of five types of stablecoins to ensure stability and minimize risks. Their simulation results indicate that different types of stablecoins, based on their inherent characteristics, face varying impacts on their prices when subjected to changes in variables such as demand shock, demand volatility, fees, and the price of collateral. The above literature explores the stability of stablecoins from the perspective of pre-issuance stability mechanism design, while other studies investigate post-issuance price dynamics models of specific stablecoins and their stability maintenance through Arbitrage (Lyons & Viswanath-Natraj, 2023; Pernice, 2021). Both pre-issuance design and post-issuance arbitrage strategies underscore the varied stability effects across different types of stablecoins, suggesting that any model predicting depegging must be individually tailored for each type.

While stablecoins are not consistently stable, they act as a haven for Bitcoin investors (Baur & Hoang, 2021), with their effectiveness fluctuating with market conditions (Wang et al., 2020). Grobys et al. (2021) observe Bitcoin's volatility significantly influencing stablecoin volatility. Ante et al. (2021a) analyze notable increases in Bitcoin's trading volume and returns following transfers of stablecoins exceeding USD 1 million. Similarly, Ante et al. (2021b) demonstrate the influence of stablecoin issuances of USD 1 million or more on the return of four major cryptocurrencies, i.e., Bitcoin, Ether, XRP, and Litecoin, which differ across individual stablecoins but also note the issuance size does not significantly impact the effect.

Additionally, Griffin & Shams (2020), Wei (2018), and Grobys & Huynh (2022) all principally investigate the relationship between Tether, the operator of USDT, and Bitcoin–specifically Tether's effect on Bitcoin. Griffin & Shams (2020) observe substantial rises in Bitcoin prices during the 2017 'crypto boom', subsequent to USDT acquisitions, typically seen after market declines. In contrast, Wei (2018) uses a VAR model and notes no influence of USDT issuances on ensuing Bitcoin profits, yet records a significant effect on Bitcoin's trading volumes. Furthermore, Grobys & Huynh (2022) detect declines in Bitcoin's price as a response to surges in USDT—a statistically significant price fluctuation within a single day. Kristoufek (2021), utilizing a VAR model, examines the directional overflow between stablecoins and other cryptocurrencies. The researcher reports no substantial proof of stablecoins boosting the prices of other crypto assets; instead, a surge in stablecoin issuances appears to coincide with rises in prices of other crypto assets, suggesting a surge in demand. Thanh et al. (2023) study the interrelation among major stablecoins and note that volatility differs among them. They ascertain that the fluctuations of major stablecoins like USDT and USDC sway the stability of relatively smaller stablecoins, with USDT's pricing exerting influence on the prices of other stablecoins.

These explorations of the price interaction between the "stablecoin market" and the "major cryptocurrency market" (primarily Bitcoin (BTC) and Ethereum (ETH)), as well as the price influence relations among "different stablecoin markets" (mainly the impact of USDT and USDC on other stablecoins), assist in forming the basis for selecting explanatory variables (feature variables) in constructing models to predict the depeggings.

Reviewing existing literature reveals that most focus on measuring or defining (in)stability, with few proposing a theoretically grounded threshold for depegging.

Beyond traditional standard deviation of returns for instability measurement or definition, there are additional methods: Grobys et al. (2021) and Grobys (2021) use Rogers & Satchell (1991)'s method, calculating realized annualized daily volatilities with OHLC data. Kwon et al. (2023) employ two measures: price deviation and downward price deviation. They define stability using the local whittle estimator (LWE) to determine the fractional integration order (*d*) range. Hoang & Baur (2021) test for absolute and relative stability by comparing the standard deviation of individual stablecoins' daily returns to 0.1% and the corresponding benchmark stablecoin's return standard deviation.

In setting the depegging threshold, literature often adopts practical viewpoints, subjective determinations, or empirical distribution statistical definitions. Here, we use 1:1 fiat-pegged stablecoins as an example to illustrate various literature settings for significant instability thresholds (akin to the concept of depegging). From a practical perspective, literature like Giokas et al. (2023) uses the traditional forex market standard, recognizing a price change exceeding 3% of the pegged price as depegging, where prices below $0.97 or above $1.03 are considered depegged (Nicolle, 2023). In empirical distribution statistical literature, such as Duan & Urquhart (2023), using histograms of hourly data for major stablecoins, depegging is defined as a price movement exceeding 1.2%, i.e., prices below $0.988 or above $1.012. Subjective determination article like S&P Global (2023) considers a price variation exceeding 10% as depegging, i.e., prices below $0.9 or above $1.1. Cintra & Holloway (2023) suggest using 5% as the threshold for hourly data and 1% for daily data to minimize noise in trading data. While Kwon et al. (2023) propose equilibrium conditions and upper/lower bounds for five types of stablecoins, they do not provide a concrete calculation method for thresholds. The most convincing definition of a depegging threshold might be from

a Kaiko cryptocurrency research institution, its research report argues that due to varying market trading volumes of different stablecoins, individual depegging thresholds should differ (Carey, 2023). It proposes a dynamic threshold formula inversely proportional to monthly trading volume, arguing that stablecoins with higher trading volumes should have tighter depegging thresholds and vice versa. In the literature, the focus on depegging direction is primarily on downward depegging (where the market price of a stablecoin falls below the lower threshold of its pegged price), but there are also discussions on upward depegging (where the market price of a stablecoin exceeds the upper threshold of its pegged price).

In the literature reviewed above, there has been no consideration of sentiment indicators. A fear and greed index, with roots tracing back to the 1930s ideas of John Maynard Keynes, reflects market driving forces of emotion. Keynes suggested that market behaviors are influenced by 'animal spirits'—spontaneous impulses driving investment or caution. The index's modern incarnation began with CNN Money's 2012 version for traditional markets, but its conceptual foundation lies in behavioral psychology and the notion that optimism (greed) and pessimism (fear) guide financial decisions. Over time, this concept evolved, making fear and greed key elements in trading discussions, illustrating how market confidence or uncertainty influences investor behavior. A fear and greed index uniquely gauges market sentiment, reflecting collective emotions like fear or greed in financial markets, especially in crypto. It assesses not prices or volumes but the overall mood, influencing investor behavior and decision-making. Traders use it to compare personal beliefs with broader market sentiment. This tool, crucial in understanding market psychology, particularly mirrors Bitcoin sentiment in the volatile crypto landscape, highlighting its interpretability and importance alongside other technical analyses (BitDegree, 2023; Tambe & Jain, 2023).

Although the fear and greed index has been applied in empirical analysis of traditional investments, its application in the cryptocurrency asset market is still in the early stages. Lin et al. (2023) explored the interplay of investor sentiment and market risk within the cryptocurrency domain, particularly focusing on bitcoin and 12 dominant cryptos. This study utilized a decomposed and partial connectedness framework to unravel the intricate web of interconnectedness and bi-directional causal relationships. It highlighted a strong correlation between investor sentiment and volatility spillovers, underlining the significance of sentiment contagion in understanding crypto market fluctuations. This offers valuable insights for both market participants and regulators, enriching the sentiment analysis and market dynamics literature in the crypto context. Wang et al. (2024) found a U-shaped pattern was identified between the crypto fear and greed index and the price synchronicity of major cryptocurrencies like Bitcoin, Ethereum, Litecoin, and Monero. This pattern, emerging from intraday data suggests a complex, non-linear interaction between collective investor sentiment and cryptocurrency price movements. Furthermore, the impact of investor fear on Bitcoin prices, especially in the context of the COVID-19 pandemic, was scrutinized (Gaies et al., 2023). The research revealed a non-constant causality between fear sentiment and Bitcoin prices, with both negative and positive interactions observed in various subperiods. The study's approach, involving a bootstrap rolling window Granger causality test, identified significant shifts in these interactions before and during the pandemic. This adds to the growing body of literature on the financial ramifications of the COVID-19 pandemic and contributes to the ongoing debate about Bitcoin's classification as a new asset class, speculative investment, currency, or a safe haven. Collectively, these studies underscore the nuanced and dynamic nature of investor sentiment in the cryptocurrency market, revealing complex patterns and

interactions that challenge traditional market theories and offer new insights for understanding the evolving landscape of digital assets. The application of the FGI in cryptocurrency assets is still in its infancy. As of now, empirical research analyses on stablecoins have not yet incorporated the fear and greed indicator.

As empirical literature unveils more findings, it also inspires the development of theoretical model literature. Building on the theoretical groundwork of Morris & Shin (1998), Kwon et al. (2023) propose a game-theoretical model that establishes equilibrium conditions for five different types of stablecoin mechanisms under various economic states. The theoretical models for stablecoins are still in the initial stages of development currently. Before these models are fully established, an early warning model or system capable of timely and effectively predicting the depegging of stablecoins will be crucial for investors making decisions in cryptocurrency asset investments and for issuers to understand the flaws in their stablecoins.

3. Methodology

3.1. Data

This paper selects the research subjects by 24-hour trading volume from the stablecoin category on the CoinGecko platform[2] (as of December 31, 2023). We chose top five stablecoins are USDT, USDC, FDUSD, BUSD, and DAI. However, since data for FDUSD is only available from July 26, 2023, and the data period is too short, it is excluded from the study. Therefore, the final subjects chosen for this research are USDT, USDC, BUSD, and DAI. All of them are stablecoins pegged to the US dollar. However, with the exception of DAI, which is an on-chain collateralized stablecoin, the

[2] CoinMarketCap. (2023, December 31). Stablecoins ranking. Retrieved from https://www.coingecko.com/en/categories/stablecoins

rest are off-chain collateralized stablecoins.

The empirical variables of this study mainly comprise two categories: on-chain transaction data and sentiment indicators. The data source for the former is CoinMarketCap[3], covering transaction data like pricing/returns, market cap/supply, and trading volume, along with derived explanatory (feature) variables. The latter data sources are AlternativeMe[4] and Alpha Data Analytics[5]. CoinMarketCap's data frequencies include daily and hourly, with data collection starting from January 1, 2018. AlternativeMe's data frequency is daily, starting from February 1, 2018. Alpha Data Analytics (ADA)'s data begins on May 17, 2022, initially daily but changed to hourly from April 2, 2023. Due to varying frequencies and start periods of different data sources, we used daily frequency data for modeling, setting the research period from January 1, 2022, to December 31, 2023[6].

This study aims to construct a predictive model for stablecoin depegging, for which we first need to define depegging, essentially setting the label variable *Y*. The depegging marker is determined jointly by *P* (Price) and *Thresh* (Threshold), where *P* may be derived from OHLC data, and *Threh* represents the threshold value. Regarding the settings of *P* and *Thresh*, our study adopts Carey (2023) 's definition but differs in that we extend the marker to include both downward ($P < Thresh$) and upward ($P > Thresh$) depegging, thus utilizing a bilateral depegging marker that considers both scenarios. Specifically, the label *Y* can be expressed as follows:

---

[3] https://coinmarketcap.com/api/

[4] https://alternative.me/crypto/fear-and-greed-index/

[5] https://adalytica.io/crypto-fear-and-greed-index

[6] Although Alpha Data Analytics' data starts from May 17, 2022, this study begins from January 1, 2022, to ensure a sufficient quantity of depegging labels. For any missing values in this indicator prior to May 17, 2022, they are imputed by backfilling with the values from that date.

$$Y = \begin{cases} 1, & \text{if } P_L \leq Thresh_D \text{ or } P_H \geq Thresh_U \\ 0, & \text{otherwise.} \end{cases}$$

where P_L represents the lowest price for the period, and P_H is the highest price for the period. Moreover, *Thresh* is a dynamic threshold value calculated based on the rolling monthly cumulative trading volume (Trading Volume). Specifically, $Thresh_D = 1 - 10/V^{\alpha}_{\text{monthly}}$ and $Thresh_U = 1 + 10/V^{\alpha}_{\text{monthly}}$, where $V^{\alpha}_{\text{monthly}}$ is the rolling windows sum of trading volume over a 30-day window. The exponent α is set at 1/3, following the optimal setting suggested by Carey (2023). Based on the definition of *Y*, the predictive models we aim to develop in this study belong to a binary classification problem.

To effectively predict the depegging of stablecoins, this paper employs relevant predictive variables (features) from literature. These include the rate of change in trading price and volume, market information change rate, sentiment indicators, and price variables. By analyzing these factors, the paper aims to provide a more accurate prediction of when and how stablecoins might deviate from their pegged values, which is crucial for understanding their stability and reliability in the cryptocurrency market.

The variation in trading price and volume is calculated by comparing past and current transaction prices and volumes, providing insights into the fluctuations in price and volume of cryptocurrency assets. The trading price change rate includes metrics such as the 1-hour, 24-hour, 7-day, and 30-day trading price percentage changes. Additionally, the trading volume change rate encompasses the 24-hour, 7-day, and 30-day trading volume percentage changes, offering a comprehensive view of the market's trading activity over various time frames.

The rate of change in market information is calculated by comparing past and current market data, encompassing aspects such as market capitalization, circulating supply of tokens, and total supply of tokens. This calculation provides an understanding

of the market overview of cryptocurrency assets over a given period. Market capitalization is determined by multiplying the latest trade price by the circulating supply. The market capitalization change rate includes metrics like the 24-hour, 7-day, and 30-day market capital percentage changes. These measures help in assessing the short-term and long-term shifts in the market value of cryptocurrencies. Circulating supply refers to the approximate number of coins currently in circulation. The change rate in circulating supply is tracked over different time frames, including 24-hour, 7-day, and 30-day circulating supply percentage changes. This data is crucial for understanding the liquidity and availability of a cryptocurrency in the market. Total supply represents the approximate total amount of coins currently in existence, minus any coins that have been verifiably burned. The change rate in total supply is also monitored over 24-hour, 7-day, and 30-day periods. This information is vital for gauging the overall scale and potential future supply of a cryptocurrency. Together, these metrics provide a comprehensive view of the cryptocurrency market's performance and trends, aiding investors and analysts in making informed decisions.

In the realm of financial markets, sentiment analysis leverages the capabilities of natural language processing, text analysis, and computational linguistics. This approach is methodically employed to identify, extract, and quantify the emotional states of investors, providing an in-depth perspective of the emotional landscape in the cryptocurrency sector. Such analysis is vital for comprehending investor behavior in this rapidly developing financial field. The study utilizes three key emotional indicators as variables: the Fear and Greed Index, the Sentiment Index, and the Awareness Index. The former comes from AlternativeMe, while the latter two are from Alpha Data Analytics. Each of these indices plays a crucial role in offering a nuanced understanding of investor sentiment and market dynamics.

Fear and Greed Index is designed to capture the nuances of market dynamics by integrating several key factors into a cohesive analysis, with a specific focus on Bitcoin. Fear and Greed Index allocates 25% of its weight to Bitcoin's volatility. Another 25% is dedicated to assessing market momentum and volume. Social media engagement, particularly on Twitter, contributes 15% to the index. Surveys, conducted in collaboration with strawpoll.com, also account for 15%[7]. The dominance of Bitcoin in the market, which forms 10% of the index. The remaining 10% is derived from Google Trends data, where an increase in specific Bitcoin-related queries signals prevailing market emotions. The Fear and Greed Index encapsulates the prevailing sentiment in the Bitcoin market, distilling complex data into a straightforward meter ranging from 0 to 100. Scores from 0 to 24 indicate a state of Extreme Fear, 25 to 46 suggest Fear, 47 to 54 represent a Neutral stance, 55 to 75 signify Greed, and 76 to 100 denote Extreme Greed. This index serves as a concise tool for understanding investor sentiment in the cryptocurrency market.

Sentiment Index generates a sentiment score on a scale from 0 to 100. Scores within the range of 0 to 30 are indicative of a state of Fear, scores from 30 to 70 are considered Neutral, and scores from 70 to 100 are interpreted as Greed. Awareness Index is measured as the ratio of media publications that are specifically focused on a given topic to the total volume of publications. This metric is also scaled from 0 to 100, with positive values indicating a rise in awareness and investor attention. Unlike the sentiment score, the awareness metric does not have a classified range, but rather serves as a continuous indicator of the level of media focus and investor interest in a particular topic.

[7] Currently paused.

Drawing on the extensive research in past empirical academic literature on in(stability), we also utilize these studies to compute measures of in(stability) from the primary literature as predictive variables for depegging. In this paper, three measures are utilized to calculate volatility indicators. The first is realized daily volatility, based on the methodologies of Grobys (2021) and Rogers & Satchell (1991), which measures actual daily price fluctuations. The second method, price deviation, as proposed by Kwon et al. (2023), assesses the deviation from a set benchmark or average price. Lastly, downward price deviation, also by Kwon et al. (2023), specifically quantifies the extent of downward price movements. These methods collectively offer a comprehensive view of the asset's volatility in the market.

In predicting the depegging of stablecoins, the model is customized to include variables specific to the stablecoin being analyzed, as well as to BTC and ETH. For instance, predicting the depegging of USDT involves 21 variables related to the rate of change in trading price and volume for USDT, BTC, and ETH. The model also integrates 27 variables for the rate of change in market information concerning the USDT, BTC, and ETH, along with 15 variables for volatility indicators specific to these currencies.

The model also employs sentiment indicators that are relevant to both stablecoins and cryptocurrencies. These indicators, encompassing both numerical and categorical data, are used selectively. For example, when applying the Fear and Greed Index, the model considers either its numerical variable or its categorical classification. A similar approach is adopted for the Sentiment Index. Thus, sentiment indicators are divided into two groups. The first includes numerical variables for the Fear and Greed Index, Sentiment Index, and Awareness Index. The second group comprises categorical classifications for the Fear and Greed Index and Sentiment Index, along with a

numerical variable for the Awareness Index. This dual approach, blending quantitative and qualitative data, enhances the understanding of market dynamics in cryptocurrency, particularly for stablecoin depegging scenarios.

In total, the study utilizes 66 variables: 21 numeric variables for the rate of change in trading price and volume, 27 numeric variables for the rate of change in market information, 2 numeric (categorical) and 1 numeric variable for sentiment indicators, and 15 numeric variables for volatility indicators. This comprehensive set of variables is meticulously chosen to provide an in-depth analysis of the factors influencing the depegging of stablecoins in the cryptocurrency market. Table 1 summarizes the definitions of the feature (predictive) variables used in this paper. All feature (predictive) variables are lagged by one period to correspond with label variable Y and are then inputted into machine learning algorithms for modeling.

Table 1 Feature (predictive) variable

Variable category	Variable name	Data type	Description
Rate of change in trading price and volume	percent_change_1h	Numeric	1-hour trading price percentage change
	percent_change_24h	Numeric	24-hour trading price percentage change
	percent_change_7d	Numeric	7-day trading price percentage change
	percent_change_30d	Numeric	30-day trading price percentage change
	volume_percent_change_24h	Numeric	24-hour trading volume percentage change
	volume_ percent_change_7d	Numeric	7-day trading volume percentage change
	volume_percent_change_30d	Numeric	30-day trading volume percentage change
Rate of change in market information	market_cap_percent_change_24h	Numeric	24-hour market capital percentage change is calculated by $\ln\left(\frac{\text{24 - hour market capital}}{\text{today's market capital}}\right)$.
	market_cap_percent_change_7d	Numeric	7-day market capital percentage change is calculated by $\ln\left(\frac{\text{7 - day market capital}}{\text{today's market capital}}\right)$.
	market_cap_ percent_change_30d	Numeric	30-day market capital percentage change is calculated by $\ln\left(\frac{\text{30 - day market capital}}{\text{today's market capital}}\right)$.
	circulating_supply_percent_change_24h	Numeric	24-hour circulating supply percentage change is calculated by

Variable category	Variable name	Data type	Description
			$\ln\left(\frac{\text{24 – hour circulating supply}}{\text{today's circulating supplyl}}\right)$).
	circulating_supply_percent_change_7d	Numeric	7-day circulating supply percentage change is calculated by $\ln\left(\frac{\text{7 – day circulating supply}}{\text{today's circulating supplyl}}\right)$).
	circulating_supply_percent_change_30d	Numeric	30-day circulating supply percentage change is calculated by $\ln\left(\frac{\text{30 – day circulating supply}}{\text{today's circulating supplyl}}\right)$.
	total_supply_percent_change_24h	Numeric	24-hour total supply percentage change is calculated by $\ln\left(\frac{\text{24 – hour total supply}}{\text{today's total supplyl}}\right)$).
	total_supply_percent_change_7d	Numeric	7-day total supply percentage change is calculated by $\ln\left(\frac{\text{7 – day total supply}}{\text{today's total supplyl}}\right)$).
	total_supply_percent_change_30d	Numeric	30-day total supply percentage change is calculated by $\ln\left(\frac{\text{30 – day total supply}}{\text{today's total supplyl}}\right)$).
Sentiment indicators*	Fear_and_Greed_Index_Alternative Me	Numeric	Fear and Greed Index value from AlternativeMe
	Fear_and_Greed_Index _Category_AlternativeMe	Ordinal	Fear and Greed Index category from AlternativeMe
	Sentiment_ADA	Numeric	Sentiment Index value from Alpha Data Analytics
	Sentiment_classification_ADA	Ordinal	Sentiment Index category from Alpha Data Analytics
	Awareness_ADA	Numeric	Awareness Index from Alpha Data Analytics

Variable category	Variable name	Data type	Description
Volatility indicators	Realized_Daily_Volatility	Numeric	Referring to the methodology outlined by Grobys (2021). $\sigma_t = \sqrt{ln\frac{P_H}{P_C} ln\frac{P_H}{P_O} + ln\frac{P_L}{P_C} ln\frac{P_L}{P_O}}$ where P_O: open price; P_H: highest price; P_L: lowest price; P_C: close price.
	Price_Deviation_5d	Numeric	Referring to the methodology outlined by Kwon et al. (2023). $\sigma_t = \sqrt{\frac{\sum_{t=0}^{T-1}(P_c-1)^2}{T}}$ where P_C: close price; T is set to a duration of 5 days.
	Price_Deviation_30d	Numeric	Referring to the methodology outlined by Kwon et al. (2023), the variable T is set to a duration of 30 days.
	Downward_Price_Deviation_5d	Numeric	Referring to the methodology outlined by Kwon et al. (2023). $\sigma_t = \sqrt{\frac{\sum_{t=0}^{T-1}(min(P_c-1,0))^2}{T}}$ where P_C: close price; T is set to a duration of 5 days.
	Downward_Price_Deviation_30d	Numeric	Referring to the methodology outlined by Kwon et al. (2023), the variable T is set to a duration of 30 days.

*: The Fear and Greed Index by AlternativeMe and the Sentiment Indicators by Alpha Data Analytics, in addition to their original numerical data, also have corresponding categorical status markers. The latter primarily aids users in understanding the state of the indicator. Additionally, from a data analysis perspective, converting numerical variables into ordinal scales often helps reduce the impact of fluctuations within numerical ranges on the predictive effectiveness of the target variables.

3.2. Methods

This part is divided into three subsections. The first subsection explains the predictive models adopted in this study. The second subsection describes the performance evaluation metrics used for the models. Finally, the last subsection details the approach to handling imbalanced data.

3.2.1. Predictive Models

In our understanding, currently, there are only two pieces of literature that have developed predictive models for the depegging of stablecoins: Giokas et al. (2023) and Cintra & Holloway (2023). The former employs a combination of machine learning methods to build their predictive model. The latter adopts the Bayesian online changepoint detection (BOCD) method. Cintra & Holloway (2023) noted that the empirical results of BOCD indicate that its predictions are relatively sensitive to parameter settings. Therefore, this study adopts machine learning approach for model construction. In this paper, three machine learning methods are employed to develop the predictive model, including logistic regression, random forest, and XGBoost

The reason for choosing these three models is mainly because the Logistic Regression model is a classic model commonly used in both traditional statistics and machine learning, serving as a baseline for performance evaluation. Random Forest, a tree-based classic model, includes Decision Trees and offers the advantage of easily interpretable results. Finally, XGBoost is a frequently winning predictive model in many machine learning competitions. The following briefly introduces the development and advantages of these methods.

3.2.1.1.Logistic regression

Logistic regression, a fundamental statistical analysis method, which is primarily used for predicting the outcomes of a categorical dependent variable based on one or

more predictor variables. This technique extends the concept of linear regression into the realm of classification, making it a pivotal tool in the field of machine learning and data mining. (Cox, 1958).

As described by (Kleinbaum et al., 2010), logistic regression is appreciated for its simplicity and interpretability. Despite its straightforward nature, it requires a thorough understanding of underlying assumptions, including the absence of multicollinearity among predictors and the linearity of independent variables with respect to log odds. Further advancements in logistic regression, such as the adoption of the maximum likelihood estimation technique for more accurate parameter estimation have significantly enhanced its robustness (Aldrich and Nelson, 1984). The core principle of logistic regression is to model the probability of a binary outcome using a logistic function (Hosmer Jr et al., 2013). The logistic model operates by estimating the probability that a given instance falls into a particular category, which is fundamentally different from the linear approach of predicting continuous values. This approach is particularly effective in situations where the dependent variable is dichotomous, such as in medical diagnosis, election outcomes, or credit scoring.

3.2.1.2.Random forest

As an extension of decision trees, Random Forest is a powerful ensemble learning algorithm in machine learning (Breiman, 2001). This algorithm is particularly effective for handling large, high-dimensional datasets and widely used for feature selection (Chen et al., 2020). To improve predictive accuracy and enhance robustness, Random Forest utilizes a collection of decision tree classifiers, each trained on distinct, randomly sampled subsets of the data. This incorporation of randomness diversifies the individual decision trees. It then aggregates the predictions from each tree through majority voting, with the class receiving the majority of votes being selected as the final prediction

(Cutler et al., 2012). This ensemble approach typically results in a model that maintains a low bias similar to individual decision trees but with reduced variance, which is a hallmark of well-performing machine learning models (Strobl, 2007).

A notable benefit of Random Forests is their minimal need for data preprocessing, enabling them to retain high accuracy even when handling incomplete datasets. Similar to their decision tree counterparts, Random Forests utilize criteria such as the Gini index and entropy for split decisions but surpass them in preventing overfitting thanks to their collective learning approach. Nonetheless, the careful calibration of model parameters, including the number of constituent trees and their growth limit, remains essential to maximize efficacy and curb overfitting, as detailed by Probst, Wright, & Boulesteix (2019).

3.2.1.3.eXtreme Gradient Boosting (XGBoost)

XGBoost is an advanced machine learning model based on gradient boosting techniques, has gained significant traction in the field of data science. This algorithm stands out for its efficiency and effectiveness in handling a variety of data types, including large, high-dimensional datasets (Chen & Guestrin, 2016). XGBoost is designed to optimize an objective function, thereby reducing the discrepancy between the predicted outcomes and the actual results. XGBoost integrates regularization terms within its framework, a crucial feature that effectively prevents the common pitfall of overfitting, which is a frequent challenge in machine learning models. This balance of precision and generalization ensures that XGBoost maintains high accuracy and robustness across a spectrum of predictive tasks (Budholiya et al., 2022).

The scalability is one of the key strengths of XGBoost to make is suitable for a wide range of applications, from small to large-scale problems. This flexibility is further enhanced by its implementation of randomization techniques, which not only

reduce overfitting but also increase training speed. Moreover, XGBoost's approach to handling tree complexity and regularization parameters has been a subject of extensive research, offering insights into optimal model tuning for specific applications. These attributes of XGBoost, along with ongoing research and development, underscore its status as a powerful and reliable tool in the arsenal of machine learning techniques (Bentéjac et al., 2021).Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

3.2.2. Performance metrics

This paper employs a machine learning model for a binary classification task, with a specific focus on predicting stablecoin depegging events. The predictive performance of the model is rigorously evaluated using a confusion matrix, which is an essential tool in the assessment of classification models. As depicted in Table 2, this matrix categorizes the predictions into four distinct outcomes: True Positives (TP), indicating instances where the model correctly predicts a depegging event; False Positives (FP), denoting cases where the model predicts depegging inaccurately; False Negatives (FN), representing scenarios where the model fails to identify an actual depegging event; and True Negatives (TN), reflecting situations where the model accurately predicts the absence of a depegging event. These categorizations are pivotal in computing various performance metrics, each offering unique insights into the model's ability to predict depegging events accurately

Table 2 Confusion matrix

	Actual depeg	Actual undepeg
Predict depeg	TP	FP

Predict undepeg	FN	TN

In the context of predicting stablecoin depegging events, the accuracy of the model plays a pivotal role. Defined as the proportion of true results (both *TP* and *TN*) in all evaluated cases, Accuracy is mathematically expressed as $\text{Aaccuracy} = \frac{TP + TN}{TP + TN + FP + FN}$. This metric reflects the model's overall correctness, providing a general overview of its effectiveness across all classifications. High accuracy indicates a model that reliably distinguishes between depegging and non-depegging events, but it may not always be sufficient, especially in imbalanced datasets where one class significantly outweighs the other.

Precision, or the positive predictive value, is particularly crucial in scenarios where the cost of false positives is high. It is calculated as $\text{Precision} = \frac{TP}{TP + FP}$ and indicates the proportion of predicted depegging events that were correctly identified. In the context of stablecoin markets, a high precision means that when the model predicts a depegging event, it is likely to be correct, minimizing the risk of false alarms that could lead to unnecessary market reactions.

Recall, also known as sensitivity, is expressed as $\text{Recall} = \frac{TP}{TP + FN}$. This metric measures the model's ability to correctly identify actual depegging events. High recall is essential in situations where missing an actual depegging event (a FN) could have severe consequences, such as in financial risk management. It highlights the model's effectiveness in capturing all relevant depegging occurrences, ensuring that significant events are not overlooked.

The F1-score, calculated as $\text{F1 - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, is a balanced metric that considers both precision and recall. It is particularly useful when there is a need to find a balance between the reliability of positive predictions (precision) and the importance of not missing actual positive cases (recall). In the domain of stablecoin depegging, the F1 score provides a more nuanced understanding of the model's performance, especially when dealing with imbalanced datasets.

Lastly, Specificity, calculated as $\text{Specificity} = \frac{TN}{TN + FP}$, quantifies the model's ability to correctly identify non-depegging events. High Specificity indicates that the model is effective in recognizing true negatives, which is crucial in avoiding false predictions of stability in the volatile cryptocurrency market.

Accuracy, Precision, Recall, F1-score, and Specificity are common performance evaluation metrics for binary classification models. This paper will reveal these five metrics for the models built, but given that the main focus of this study is on depegging, which involves issues of imbalanced data, Precision and Recall are more appropriate performance evaluation metrics. Moreover, as this paper considers these two metrics equally important, the F1-score, which harmonizes both, is uniformly adopted as the primary indicator for model selection during the empirical analysis in Section 4.

3.2.3. Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Over-sampling Technique (SMOTE) is a widely used approach to address the challenge of imbalanced datasets in machine learning. Imbalanced data refers to situations where the number of observations in one class significantly outweighs the number of observations in another, leading to poor model performance, especially for the minority class (Chawla et al., 2002).

SMOTE is distinct from traditional oversampling methods in that it generates synthetic samples for the minority class rather than simply replicating existing samples. This innovative approach is key in mitigating the overfitting issue commonly associated with duplicating minority class samples. The technique involves selecting a random point from the minority class, identifying its k-nearest neighbors, and then creating synthetic samples along the line segments joining these neighbors. This method not only adds diversity to the data but also maintains critical information, often lost when the majority class is undersampled.

Additional studies have expanded on the foundational work of SMOTE. For example, (Han et al., 2005) introduced a variation of SMOTE that focuses on the

borderline examples of the minority class. Another noteworthy contribution, (He et al., 2008) proposes an adaptive synthetic sampling method to shift the learning bias towards the minority class. (Fernández et al., 2018) provides a comprehensive review of the developments and challenges in the field since the inception of SMOTE.

The significance of SMOTE extends to various applications where the minority class holds particular importance, such as in fraud detection and rare disease diagnosis. Its effectiveness in balancing datasets without compromising on data integrity has been widely recognized and built upon in subsequent research.

4. Empirical Analysis

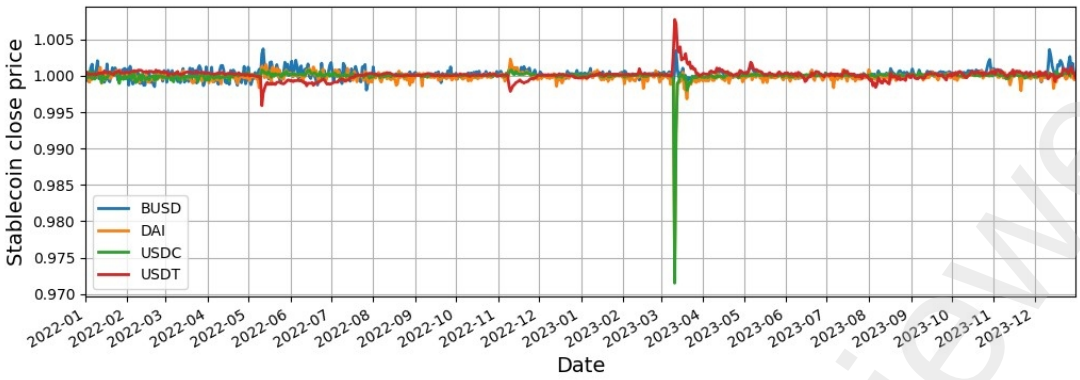
The data in this study are daily frequency data, spanning from January 1, 2022, to December 31, 2023. The stablecoins studied include USDT, USDC, BUSD, and DAI. Additionally, literature indicates that two major cryptocurrencies, BTC and ETH, may impact stablecoin stability. Table 3 presents the descriptive statistics of the closing prices for four major circulating stablecoins and two major cryptocurrencies (BTC and ETH). Judging by the range of minimum (Min) and maximum (Max) values in the table, the data from the source do not contain any unreasonable values. Regarding the average value of stablecoins (Mean), it can be seen that unless the value is rounded to three decimal places, the average value of the stablecoins may not necessarily equal their pegged price of $1. Additionally, the standard deviation (STD) indicates varying degrees of volatility among different stablecoins.

Table 3 Descriptive statistics

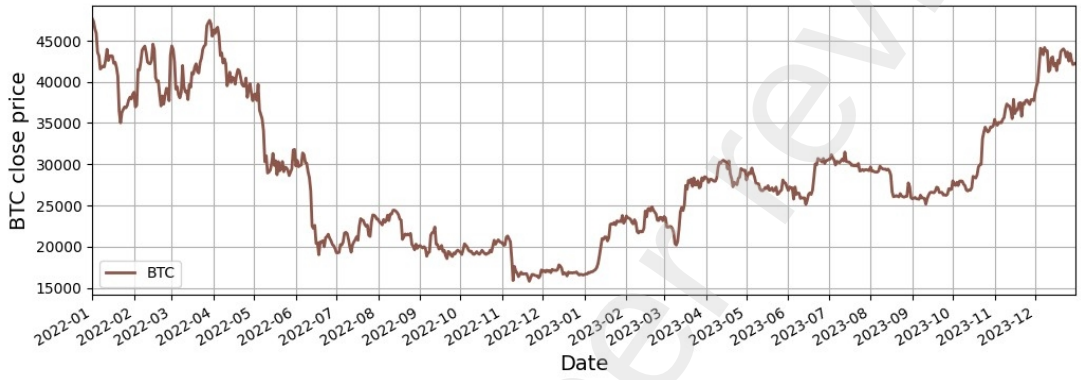
T	Token Symbol	Number of Record	Mean	STD	Min	25%	50%	75%	Max
Stablecoins	USDT	730	1.0001	0.0007	0.9959	1.0000	1.0001	1.0003	1.0077

T	Token Symbol	Number of Record	Mean	STD	Min	25%	50%	75%	Max
	USDC	730	1.0000	0.0011	0.9715	0.9999	1.0000	1.0001	1.0008
	DAI	730	0.9998	0.0011	0.9739	0.9996	0.9999	1.0001	1.0023
	BUSD	730	1.0002	0.0006	0.9980	0.9999	1.0002	1.0005	1.0037
Cryptoassets	BTC	730	28528.7	8332.1	15787.3	21529.6	27272.5	35074.0	47686.8
	ETH	730	1891.3	577.4	993.6	1561.8	1791.0	2044.7	3829.6

Panel A of Figure 1 shows that the price trends of the four major circulating stablecoins exhibit different clustering of volatility during various periods. As indicated in the literature, the clustering of price volatility for stablecoins may stem from the significant downward impact on the prices of major cryptocurrencies, such as BTC and ETH, as referenced in Panels B and C of Figure 1. Interestingly, the trends of different stablecoins are not necessarily the same during different periods. For example, during the significant crypto market downturn in May 2022, the price of USDT notably fell below the pegged price of $1, while the price of BUSD was significantly higher than its pegged price of $1. Additionally, in March 2023, the price of USDC substantially deviated downward from its pegged price of $1 due to the bankruptcy of its collateral custodian bank, SVB. However, during the same period, USDT and BUSD showed the opposite trend. This indicates that the state of the mainstream cryptocurrency market (especially the major ones: BTC and ETH) impacts the stablecoin market, but different stablecoins exhibit varying price trends. Moreover, when individual stablecoins encounter specific situations, they show different patterns in their price trends due to their substitutability.



Panel A The closing price of stablecoins



Panel B The closing price of BTC

Panel C The closing price of ETH

Figure 1 The closing price of stablecoins and major cryptoassets

Table 4 shows the depegging statistics for four major circulating stablecoins. Since the depegging threshold defined in this paper dynamically changes according to the monthly trading volume of each individual stablecoin, the results in Table 4 are somewhat counterintuitive compared to those in Table 3. Although Table 3 indicates that USDC and DAI have higher standard deviations than USDT and BUSD, Table 4

shows the opposite in terms of the depegging ratio, with USDC and DAI having lower depegging rates compared to USDT and BUSD.

Table 4 Depegging statistics

Token Symbol	Number of Record	Number of Depeg Event	Depeg Ratio
USDT	730	168	23.01%
USDC	730	16	2.19%
DAI	730	14	1.92%
BUSD	730	193	26.44%

The depegging ratios shown in Table 4 highlight another issue. Due to the varying number of depegging events for each stablecoin and the definition based on dynamic threshold values, the depegging points in time for each stablecoin may differ. Therefore, in this study, the division of the train dataset and test dataset does not employ a "time-based split" but instead utilizes stratified random sampling based on Label *Y*. This approach ensures that the structure and label proportions in the train and test datasets are consistent with those in the full dataset. Furthermore, despite training and testing sets having depegging ratios close to the overall population, stablecoins like USDC and DAI exhibit depegging ratios below 10%. To address the data imbalance, SMOTE is applied for USDC and DAI, which exhibit lower depegging rates.

Regarding the correspondence between label variables (*Y*) and feature variables (Xs) for each observation. We structure data processing methodology to accurately predict the depegging of stablecoins for the subsequent day. The methodology involves adjusting the dependent variable for a one-day forward prediction, necessitating a dataset shift by one day. The feature set includes 66 variables, either all numeric or a mix of 64 numeric and 2 categorical variables.

To address missing values, a forward-fill strategy substitutes missing data points with the preceding value. Where the preceding value is unavailable, a backward-fill approach is employed, using subsequent data to fill gaps. For categorical features, especially sentiment indicators, label encoding is applied. Numerical variables undergo normalization using a min-max scaling approach, defined as $X_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$, ensuring data scale and distribution uniformity.

This data processing and preparation strategy is designed to enhance the predictive performance of the model, particularly in stablecoin depegging prediction. The study employs 2-fold cross-validation, dividing the dataset for iterative training and validation. In terms of model performance evaluation, setting the verbose option to 2 provides detailed progress updates. The performance metric for model hyperparameter tuning is the F1 score, which is ideal for imbalanced classification scenarios. Higher F1 scores indicate better model performance.

Table 5 presents the results of predictive models built for four stablecoins using three different machine learning algorithms. As described in 3.2.2 Performance metrics, considering the nature of "depegging" in this study and the desire for the predictive models to effectively balance Type I Error (False Positives) and Type II Error (False Negatives), benefiting the practical operability of the models, we use the F1-score as the main indicator of model performance. The F1-score, being the harmonic mean of precision and recall, ranges from 0 to 1, with 1 as the best possible score and 0 as the worst. An F1-score of 0.5 indicates a moderate level of performance, suggesting that the balance between precision and recall is average. This means the model is only half as effective as the best-case scenario in correctly identifying true positives. From this, it is evident that the classic logistic Regression model generally performs poorly, while Random Forest and XGBoost show good predictive effectiveness. Further examining

the Precision and Recall metrics, Logistic Regression exhibits significant fluctuations, but all results for Random Forest and XGBoost consistently exceed 0.5, implying that nonlinear models are more effective in predicting dynamic depegging events. Further analysis shows that the F1-scores for DAI, an on-chain collateralized stablecoin, are relatively lower (0.571) compared to the other three off-chain collateralized stablecoins (USDT, USDC, and BUSD), which almost all have F1-scores above 0.75. This could be due to the differences in the price stability effects inherent in the design of different types of stablecoins, as noted in the literature.

Table 5 Model performance

Model	Toekn Symbol	SMOTE Ratio*	Accuracy	Precision	Recall	F1 Score	Specificity
Logistic Regression	BUSD	NA	0.869	0.783	0.735	0.758	0.921
Random Forest	BUSD	NA	0.891	0.788	0.837	0.812	0.913
XGBoost	BUSD	NA	0.88	0.792	0.776	0.784	0.921
Logistic Regression	USDT	NA	0.829	0.773	0.405	0.531	0.962
Random Forest	USDT	NA	0.886	0.806	0.69	0.744	0.947
XGBoost	USDT	NA	0.897	0.8	0.762	0.78	0.94
Logistic Regression	USDC	0.6	0.928	0.167	0.5	0.25	0.939
Random Forest	USDC	0.6	0.994	1	0.75	0.857	1
XGBoost	USDC	0.6	0.982	0.6	0.75	0.667	0.988
Logistic Regression	DAI	0.6	0.929	0.214	0.75	0.333	0.933
Random Forest	DAI	0.6	0.982	0.667	0.5	0.571	0.994
XGBoost	DAI	0.6	0.982	0.667	0.5	0.571	0.994

*: After hyperparameter tuning, the optimal choice for the SMOTE ratio is determined to be 0.6. Conversely, for stablecoins like BUSD and USDT, with higher depegging ratios, SMOTE is not applied.

Table 6 presents a count statistic of the top ten most important feature variables for the models outlined in Table 5. The higher the frequency of a feature, the more significant it is as a predictive variable for stablecoin depegging. Consistent with previous empirical literature, major cryptocurrencies (like BTC and ETH) are shown to

have a spillover effect on stablecoins, especially in terms of supply and market cap fluctuations. Additionally, it is notable that longer-term instability measures proposed in past literature, such as Realized Daily Volatility, Price Deviation, and Downward Price Deviation, are indicative in predicting stablecoin depegging. However, surprisingly, the consolidated count in Table 6 does not include any of the three sentiment indicators incorporated in this study.

Table 6 Aggregation of feature importance

Rank	Feature	Count
1	BTC_Realized_Daily_Volatility	5
2	ETH_total_supply_percent_change_30d	4
3	ETH_market_cap_percent_change_24h	4
4	BTC_volume_percent_change_30d	4
5	USDT_Realized_Daily_Volatility	3
6	ETH_percent_change_24h	3
7	BUSD_Realized_Daily_Volatility	3
8	BUSD_Price_Deviation_5d	3
9	BUSD_Price_Deviation_30d	3
10	BUSD_Downward_Price_Deviation_30d	3
11	ETH_Realized_Daily_Volatility	3
12	BTC_volume_percent_change_24h	3
13	ETH_volume_percent_change_30d	3
14	BTC_percent_change_24h	3
15	USDT_Price_Deviation_5d	3

5. Conclusion

This paper presents a comprehensive analysis of stablecoin depegging risk prediction, focusing on the top four stablecoins in terms of daily trading volume: USDT, USDC, BUSD, and DAI. The study utilizes a novel approach by incorporating dynamic depegging thresholds based on trading volume and integrating sentiment indicators from news sources, a first in this area of research. Empirical analysis from January 1,

2022, to December 31, 2023, demonstrates the significant influence of major cryptocurrency price and volume fluctuations on stablecoin depegging, aligning with existing literature.

Our predictive models, developed using logistic regression, random forest, and XGBoost machine learning algorithms, reveal the complexity and challenges in predicting stablecoin depegging events. The models indicate that traditional on-chain data such as price, volume, and market capitalization changes are crucial for predicting depegging events. However, sentiment indicators, despite their theoretical relevance, did not show significant predictive power in our models. This suggests that while investor sentiment is a critical aspect of cryptocurrency markets, its direct impact on stablecoin depegging might not be as pronounced as expected.

Interestingly, our results also highlight that stablecoin type plays a role in depegging risk, with on-chain collateralized stablecoins like DAI showing different depegging patterns compared to off-chain collateralized stablecoins. This underscores the importance of considering the underlying collateralization mechanism in stablecoin risk assessment. The inclusion of major cryptocurrencies (BTC and ETH) in the analysis further illustrates the interconnected nature of the crypto asset market, where movements in major cryptocurrencies can significantly impact stablecoins.

In terms of model performance, non-linear models such as random forest and XGBoost outperformed the logistic regression model. This indicates the complexity of factors influencing stablecoin depegging, requiring sophisticated models to capture non-linear relationships and interactions between various variables. The use of SMOTE for addressing imbalanced data sets enhances the models' ability to predict rare depegging events, which is crucial for effective risk management in the crypto asset market.

Overall, this paper contributes to the understanding of stablecoin depegging risks by developing predictive models that combine traditional on-chain data with sentiment indicators and consider the impact of major cryptocurrencies. While the study advances the field, the limited predictive power of sentiment indicators and the varying effects based on the type of stablecoin suggest that further research is needed to refine these models. The findings have practical implications for crypto asset investors, providing tools for better risk assessment and investment decision-making in the volatile world of cryptocurrencies.

The study's implications extend beyond individual investors to the broader financial ecosystem. As stablecoins continue to bridge traditional finance and cryptocurrency markets, understanding their stability and risk factors becomes increasingly important for regulatory bodies, financial institutions, and cryptocurrency platforms. This research provides a foundation for developing more robust risk management strategies and regulatory frameworks, contributing to the overall stability and growth of the crypto asset market.

References

Aldrich, J. H., & Nelson, F. D. (1984). Linear probability, logit, and probit models (No. 45). Sage.

Ante, L., Fiedler, I., & Strehle, E. (2021a). The impact of transparent money flows: Effects of stablecoin transfers on the returns and trading volume of Bitcoin. *Technological Forecasting and Social Change*, *170*, 120851.

Ante, L., Fiedler, I., & Strehle, E. (2021b). The influence of stablecoin issuances on cryptocurrency markets. *Finance Research Letters*, *41*, 101867.

Ante, L., Fiedler, I., Willruth, J. M., & Steinmetz, F. (2023). A Systematic Literature Review of Empirical Research on Stablecoins. *FinTech*, *2*(1), 34-47.

Baur, D. G., & Hoang, L. T. (2021). A crypto safe haven against Bitcoin. *Finance Research Letters*, *38*, 101431.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

BitDegree. (2023). *Crypto Fear and Greed Index*. https://www.bitdegree.org/cryptocurrency-prices/fear-and-greed-index

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University-Computer and Information Sciences, 34(7), 4514-4523.

Bullmann, D., Klemm, J., & Pinna, A. (2019). *In search for stability in crypto-assets: are stablecoins the solution?* E. C. Bank. https://bit.ly/3nqGKhC

Carey, R. (2023). *Defining Depegs: A New Metric for Stablecoin Stability*. Kaiko Research. https://research.kaiko.com/insights/defining-depegs-a-new-metric-for-stablecoin-stability

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 52.

Cintra, T. N., & Holloway, M. P. (2023). Detecting Depegs: Towards Safer Passive Liquidity Provision on Curve Finance. *arXiv preprint arXiv:2306.10612*.

Clements, R. (2021). Built to fail: The inherent fragility of algorithmic stablecoins. *Wake Forest L. Rev. Online*, *11*, 131.

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2), 215-232.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. Ensemble machine learning: Methods and applications, 157-175.

d'Avernas, A., Bourany, T., & Vandeweyer, Q. (2021). Are stablecoins stable? *Working Paper*.

De Blasis, R., Galati, L., Webb, A., & Webb, R. I. (2023). Intelligent design: stablecoins (in) stability and collateral during market turbulence. *Financial Innovation*, *9*(1), 85.

Duan, K., & Urquhart, A. (2023). The instability of stablecoins. *Finance Research Letters*, *52*, 103573.

European Central Bank. (2019). *Stablecoins – no coins, but are they stable?* European Central Bank. https://bit.ly/3yUYkfE

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

Gadzinski, G., Castello, A., & Mazzorana, F. (2023). Stablecoins: Does design affect stability? *Finance Research Letters*, *53*, 103611.

Gaies, B., Nakhli, M. S., Sahut, J.-M., & Schweizer, D. (2023). Interactions between investors' fear and greed sentiment and Bitcoin prices. *The North American Journal of Economics and Finance*, *67*, 101924.

Giokas, Y., Hocking, P., Chapcak, M., & Catala, P. (2023). Digital Asset Monitor.

Global, S. P. (2023). *Stablecoins: A Deep Drive into Valuation and Depegging*. https://www.spglobal.com/en/research-insights/featured/special-editorial/stablecoins-a-deep-dive-into-valuation-and-depegging

Griffin, J. M., & Shams, A. (2020). Is Bitcoin really untethered? *The Journal of Finance*, *75*(4), 1913-1964.

Grobys, K. (2021). When the blockchain does not block: on hackings and uncertainty in the cryptocurrency market. *Quantitative Finance*, *21*(8), 1267-1279.

Grobys, K., & Huynh, T. L. D. (2022). When Tether says "JUMP!" Bitcoin asks "How low?". *Finance Research Letters*, *47*, 102644.

Grobys, K., Junttila, J., Kolari, J. W., & Sapkota, N. (2021). On the stability of stablecoins. *Journal of Empirical Finance*, *64*, 207-223.

Hafner, M., Pereira, M. H., Dietl, H., & Beccuti, J. (2023). The four types of stablecoins: A comparative analysis. *arXiv preprint arXiv:2308.07041*.

Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Berlin, Heidelberg: Springer Berlin Heidelberg.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

Hoang, L. T., & Baur, D. G. (2021). How stable are stablecoins? *The European Journal of Finance*, 1-17. https://bit.ly/3yUn87B

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Jarno, K., & Kołodziejczyk, H. (2021). Does the design of stablecoins impact their volatility? *Journal of Risk and Financial Management*, *14*(2), 42.

Kleinbaum, D. G., Klein, M., Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. Logistic regression: a self-learning text, 1-39.

Kristoufek, L. (2021). Tethered, or Untethered? On the interplay between stablecoins and major cryptoassets. *Finance Research Letters*, *43*, 101991.

Kwon, Y., Pongmala, K., Qin, K., Klages-Mundt, A., Jovanovic, P., Parlour, C., Gervais, A., & Song, D. (2023). What Drives the (In) stability of a Stablecoin? *arXiv preprint arXiv:2307.11754*.

Lin, X., Meng, Y., & Zhu, H. (2023). How connected is the crypto market risk to investor sentiment? *Finance Research Letters*, *56*, 104177.

Lyons, R. K., & Viswanath-Natraj, G. (2023). What keeps stablecoins stable? *Journal of International Money and Finance*, *131*, 102777.

Morris, S., & Shin, H. S. (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, 587-597.

Nicolle, E. (2023). *Crypto's Digital Dollars Are a Calculated Risk*. Bloomberg. Retrieved 2023/12/15 from https://www.bloomberg.com/news/newsletters/2023-11-09/crypto-s-digital-dollars-are-a-calculated-risk

Pernice, I. G. A. (2021). On stablecoin price processes and arbitrage. Financial Cryptography and Data Security. FC 2021 International Workshops: CoDecFin, DeFi, VOTING, and WTSC, Virtual Event, March 5, 2021, Revised Selected Papers 25,

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.

Rogers, L. C. G., & Satchell, S. E. (1991). Estimating variance from high, low and closing prices. *The Annals of Applied Probability*, 504-512.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 1-21.

Tambe, N., & Jain, A. (2023). Fear And Greed Index For Crypto. https://www.forbes.com/advisor/in/investing/cryptocurrency/fear-and-greed-index-crypto/

Thanh, B. N., Hong, T. N. V., Pham, H., Cong, T. N., & Anh, T. P. T. (2023). Are the stabilities of stablecoins connected? *Journal of Industrial and Business Economics*, *50*(3), 515-525.

Uhlig, H. (2022). *A luna-tic stablecoin crash*.

Wang, G.-J., Ma, X.-y., & Wu, H.-y. (2020). Are stablecoins truly diversifiers, hedges, or safe havens against traditional cryptocurrencies as their name suggests? *Research in International Business and Finance*, *54*, 101225.

Wang, J.-N., Liu, H.-C., & Hsu, Y.-T. (2024). A U-shaped relationship between the crypto fear-greed index and the price synchronicity of cryptocurrencies. *Finance Research Letters*, *59*, 104763.

Wei, W. C. (2018). The impact of Tether grants on Bitcoin. *Economics Letters*, *171*, 19-22.