Unsupervised Anomaly Detection in Financial Time Series Using Transformer and Temporal Convolutional Networks

Deroo Florian
florianderoo@outlook.fr

Abstract

This study presents an unsupervised approach for anomaly detection in financial time series using two complementary architectures: an enhanced temporal convolutional autoencoder (TCN-AE) and a modified transformer model. Leveraging 50 years of market data (1973-2023) from CRSP covering 36,328 different assets, we compare three training strategies: one using all available data, another focusing exclusively on windows preceding bullish movements, and a third targeting windows before bearish trends. Results demonstrate that models trained on bearish windows present superior and more consistent discriminative capacity, effectively identifying market configurations preceding upward movements with average returns reaching 1.61% for the transformer model and 0.76% for the TCN-AE. This asymmetric behavior offers promising perspectives for algorithmic trading strategies, as validated by our back-testing results showing average returns of 4.40% and 3.07% respectively when applying a simple trading rule based on the detected anomalies.

Index Terms

anomaly detection, unsupervised learning, financial time series, temporal convolutional networks, transformer models, autoencoder, algorithmic trading

I. Introduction

Anomaly detection in financial markets represents a major challenge, both for risk management and for identifying investment opportunities. In this context, machine learning techniques, particularly unsupervised learning, are emerging as promising tools for identifying abnormal patterns in financial time series. These anomalies can signal various situations of interest: unusual price movements, exceptional transaction volumes, or atypical market behaviors preceding significant events. The unsupervised approach offers several decisive advantages for this task. It allows for more flexible and robust detection of anomalies, not being constrained by pre-established definitions of what constitutes an anomaly. This flexibility is crucial in financial markets, where the nature of anomalies can evolve over time and vary according to market conditions. Unsupervised learning can potentially identify previously unknown types of anomalies, paving the way for new investment strategies. Our study proposes a comparison between two distinct architectures for unsupervised anomaly detection in market data: a temporal convolutional autoencoder and an adapted transformer model. These methods aim to capture the complex and multi-scale dynamics of financial time series, while adapting to the specificities of market data, particularly their non-stationarity and high volatility. By comparing different training strategies across both models, we seek to determine which approach better identifies market configurations with particular predictive value when flagged as anomalies.

II. The Data

The Center for Research in Security Prices (CRSP) is a major reference in the collection and maintenance of stock market data, particularly for the NYSE, AMEX, and NASDAQ markets. For this study, we specifically exploited stock price data over a period extending from 1973 to 2023, focusing on six key indicators: Bid, Ask, Low, High, Close, and transaction volume. This CRSP database is particularly reliable due to its methodological rigor and comprehensive coverage of the American market, making it a reference tool for academic research and financial analysis. Our dataset includes 36,328 different assets, all listed on the NYSE, AMEX, and NASDAQ markets, representing a total of 60,555,699 daily observations. The choice of daily granularity allows capturing significant market dynamics while minimizing the noise inherent in higher-frequency data.

III. Normalization

Data normalization constitutes a crucial preliminary step in our analysis, particularly important for financial data where price scales between different assets can vary considerably. This step ensures relevant comparability between different time series. For transaction volume, we apply logarithmic normalization followed by standardization, defined by the following equation:

$$z_t = \frac{\log(1 + x_t) - \mu_{\log,w}}{\sigma_{\log,w} + \epsilon}$$

x_t is the daily volume at time t

$\mu_{\log,w}$ is the moving average of log-volumes over a window of size $w = 50$

$\sigma_{\log,w}$ is the moving standard deviation of log-volumes over the same window

$\epsilon = 10^{-8}$ is a term for numerical stabilization

This logarithmic approach effectively manages the strong asymmetry and high variability typically observed in volume data. For price series, we employ a different normalization, more adapted to their nature:

$$z_t = \frac{x_t - \mu_w}{\sigma_w + \epsilon}$$

x_t is the value of the series at time t
μ_w is the moving average over a window of size w = 50
σ_w is the moving standard deviation over the same window
$\epsilon = 10^{-8}$ is a term for numerical stabilization

This normalization method was chosen considering that the maximum prices of assets are not known a priori and evolve over time. The moving window allows dynamic adaptation to regime changes in the time series. The two normalization approaches presented above allow obtaining standardized and comparable time series, while preserving the essential characteristics of the underlying market dynamics.

The bimodal distributions observed (figure 15) in the data normalization are a natural consequence of using a rolling window for standardization, where only the past context is taken into account. This characteristic, far from being problematic, reflects the dynamic nature of our normalization approach, where each point is standardized relative to its local temporal past context. This bimodality illustrates the distinction between values that are above and below their respective moving averages calculated on past data, thus capturing the intrinsic local variations in financial time series.

IV. Transformer Architecture adapted for anomaly detection

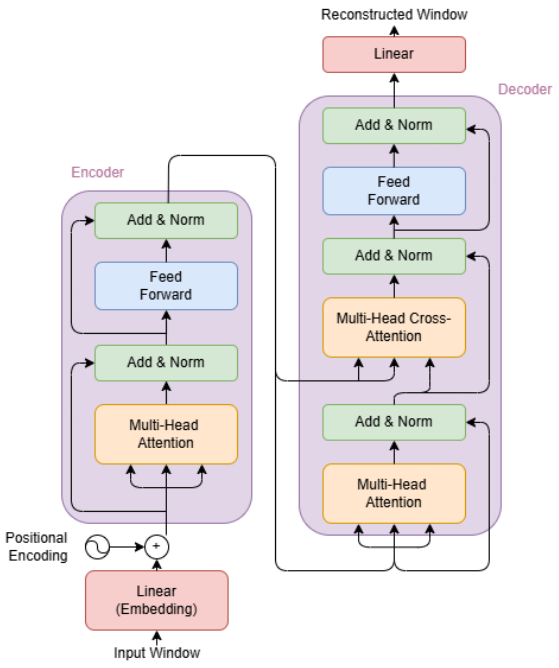


Fig. 1: Transformer Architecture adapted for anomaly detection

The presented architecture (figure 1), based on a Transformer, is particularly well-suited for an autoencoder in the context of anomaly detection in time series, as it leverages the power of attention mechanisms to capture complex dependencies within sequences, while enabling precise reconstruction of input data.

Unlike the classical Transformer architecture, as described by Vaswani et al., 2017 [2], which is designed for translation tasks with a decoder generating a target sequence autoregressively, this architecture adapts the decoder for a reconstruction task by using the encoder output as the target input, without causal masking. This absence of masking allows bidirectional processing of sequences, where each position can access the entire context in both directions, a significant advantage for anomaly detection where patterns may depend on information located both before and after a given point in the sequence.

The combined use of an encoder and decoder, rather than an encoder alone, is motivated by the need to introduce a functional separation: the encoder extracts a latent representation of the data, while the decoder learns to reconstruct from this representation. Experiments with an encoder-only approach demonstrated insufficient performance, confirming the importance of this dual architecture. Indeed, when the encoder output is used both as the source sequence and as the cross-attention context for the decoder, it creates an information bottleneck that forces the model to learn more robust and generalizable representations of normal patterns in the data.

This encoder-decoder separation promotes better generalization and more robust anomaly detection by precisely identifying the discrepancies between the input and reconstructed output. The decoder, by receiving the encoded representation as an attention context, can focus on reconstructing the normal parts of sequences while highlighting abnormal segments that resist faithful reconstruction, a crucial aspect for anomaly detection applications where a simple encoder would lack the capacity to explicitly model this differentiated reconstruction process.

V. Auto-Encoder Architecture

The proposed architecture builds upon the foundational work presented in "Temporal convolutional autoencoder for unsupervised anomaly detection in time series" by Thill et al., 2021 [1]. This unsupervised approach to anomaly detection in time series relies on an autoencoder using temporal convolutional networks (TCN) with dilated convolutions. The effectiveness of this method, designated TCN-AE, has been demonstrated on electrocardiogram (ECG) data for the detection of cardiac arrhythmias, surpassing the performance of state-of-the-art unsupervised methods in 2021.

The central mechanism of the architecture relies on expansion and compression operations of our features (dConv), allowing the model to autonomously learn relevant attributes at different temporal scales. This multi-scale analysis is made possible thanks to a dilation factor d increasing exponentially according to the relation $d = 2^n$.

Formally, a dConv operation can be expressed as:

$$\text{dConv}(X, d) = \text{Conv1D}_{\text{compression}}(\text{Conv1D}_{\text{expansion}}(X, d))$$

where $\text{Conv1D}_{\text{expansion}}$ applies a one-dimensional convolution with a dilation step d to capture long-distance dependencies, and $\text{Conv1D}_{\text{compression}}$ uses a standard convolution without dilation to reduce the dimensionality of the extracted features.

The compressions obtained at each temporal scale (thus at each dConv output) are then concatenated before a final compression of the features.

Our main contribution lies in the introduction of progressive temporal compressions (pConv). This innovation aims to smooth the temporal compression process by combining a 1D Max Pooling operation with an expansive 1D convolution. This approach allows a more gradual reduction of the temporal dimension while preserving the essential characteristics of the signal.

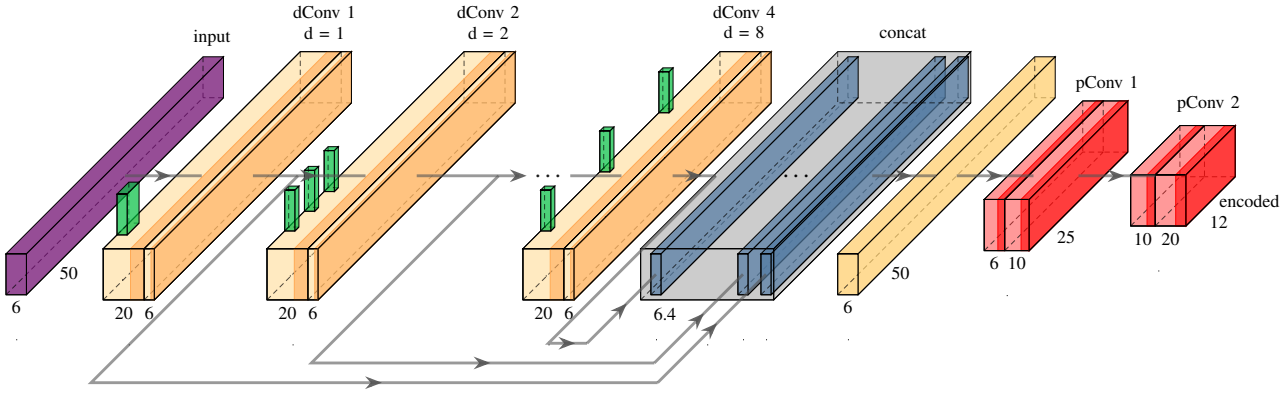


Fig. 2: Architecture of the TCN-AE encoder with smooth temporal compression

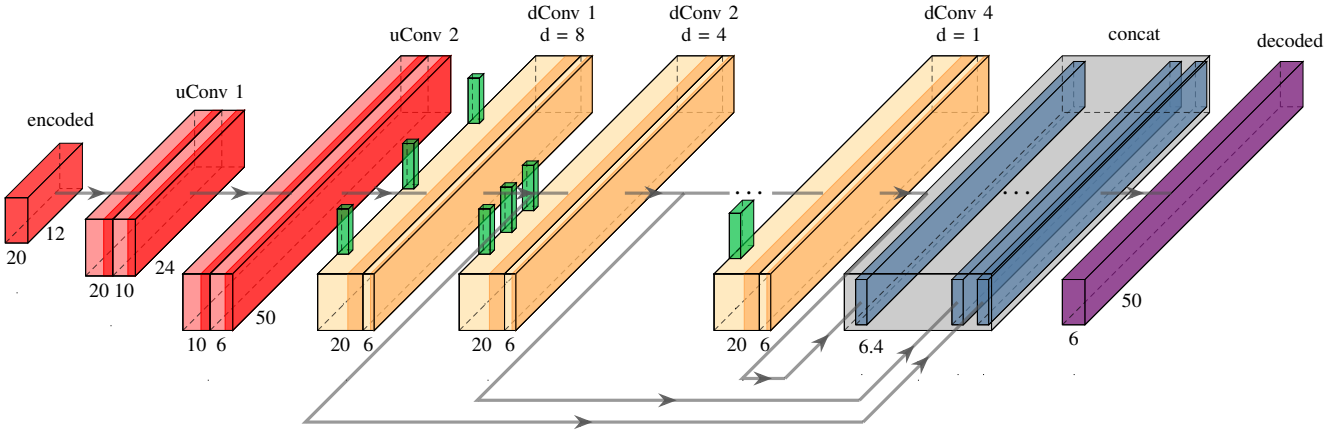


Fig. 3: Architecture of the TCN-AE decoder with smooth temporal decompression

VI. Models Training

The model takes as input time series of 50 market days, comprising 6 features for each time point: Bid, Ask, Low, High, Close, and Volume. This 50-day window was chosen to capture a sufficient temporal context while maintaining relevant granularity for anomaly detection.

In our experimental approach, we explored three distinct training strategies.

The first consists of training the model on all available time windows regardless of future market evolution; models trained on this dataset will be labeled as ”All” or ”All Windows.”

The second focuses exclusively on windows where the closing price 20 days after the end of the window is higher than the closing price of the last day of the window; models trained on this dataset will be labeled as ”Positive” or ”Positive Windows”

The 20-day horizon was chosen as a compromise between a period long enough to capture significant market movements, but not too extended to avoid diluting the signal with economic events or news unrelated to the detected configurations.

The third, which is a model trained on negative future outcomes, focuses on periods where the market continues to decline 20 days after the closing of the last candle of the window; models trained on this dataset will be labeled as ”Negative” or ”Negative Windows.”

This triple approach aims to evaluate whether the model

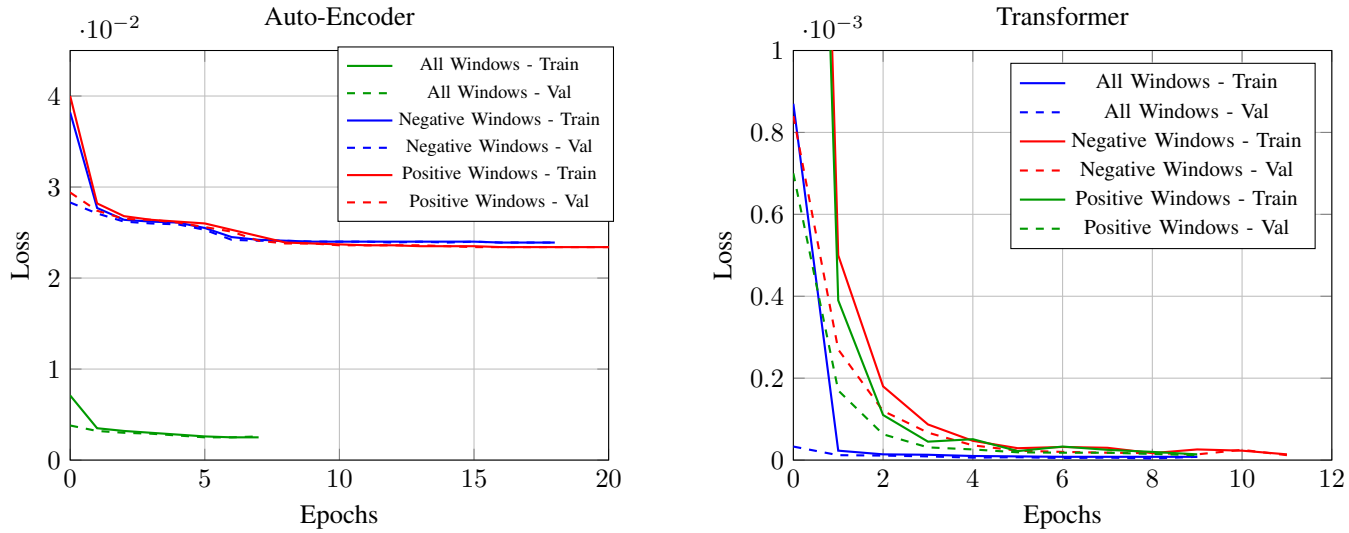
develops different reconstruction capabilities depending on the market context used for its learning.

The interest of this multiple training strategy lies in the possibility of analyzing potential differences in the model's reconstruction capabilities according to market context. This analysis could reveal interesting perspectives on the nature of detected anomalies and their relationship with future market movements.

TABLE I: Parameters of dilated convolutional autoencoder and transformer models

Autoencoder	
Parameter	**Value**
Input dimension	6
Sequence length	50
Expansion channels	20
Compression channels	6
Post-pooling channels (pConv 1,2,3)	[20, 40, 60]
Convolution kernel (expansion)	3
Activation	ReLU
Training parameters	
Batch size	32
Learning rate	0.001
Optimizer	Adam

Transformer	
Parameter	**Value**
Input dimension	6
Sequence length	50
Embedding dimension	64
Number of attention heads	4
Number of encoder layers	2
Number of decoder layers	2
Feed-forward network dimension	128
Dropout	0.1
Training parameters	
Batch size	32
Learning rate	0.001
Optimizer	Adam



(a) Learning curves for the Auto-Encoder

(b) Learning curves for the Transformer

Fig. 4: Comparison of training performances between Auto-Encoder and Transformer models

VII. Evaluation Methodology

To evaluate the performance of our model, we defined an evaluation methodology based on anomaly detection across our historical data covering the period from 1973 to 2023.

A. Definition of the Anomaly Threshold

Since our model functions as an unsupervised anomaly detector, it is necessary to define a threshold above which a time window is considered anomalous. For each 50-day window analyzed by our autoencoder, we calculate the reconstruction error (MSE) which represents the anomaly score.

To establish a relevant threshold, we opted for an approach based on quantiles (0.98, 0.99, etc.) of the reconstruction error distribution. Specifically, we submit all available windows to our model and select those with the highest reconstruction errors, considering that they correspond to the most atypical market configurations.

Table II presents the number of windows identified as anomalies according to different quantile thresholds:

Quantile	Number of windows
0.98	11938
0.99	5973
0.994	3580
0.995	2986
0.996	2388
0.997	1790

TABLE II: Number of windows identified as anomalies by quantile

This method has the advantage of being directly applicable in a production context, by maintaining the threshold determined during the evaluation phase.

B. Evaluation of the Relevance of Detected Anomalies

To determine whether the detected anomalies have predictive value, we analyzed the market behavior following each window identified as anomalous. Specifically, we compared the closing price approximately 20 days after the end of the window with that of the last day included in the window.

This methodology allows us to verify whether market configurations considered anomalous by our model indeed precede significant price movements, whether bullish or bearish.

C. Comparative Analysis of Training Strategies for the 0.98 Quantile

Our three training strategies revealed significant differences in anomaly detection capability and their implications for associated returns:

TABLE III: Comparison of return distributions for detected anomalies

Statistics	Transformer			Auto-encoder			Full Dataset
	Generalist	Bullish	Bearish	Generalist	Bullish	Bearish	
Mean (%)	0.52	0.66	**1.61**	-0.21	0.77	**0.76**	0.53
Median (%)	0.02	0.23	**0.67**	0.00	0.06	**0.55**	0.19
Standard deviation (%)	13.75	14.7	16.65	13.89	16.34	14.73	13.9

It is important to note that, to prevent the mean from being too influenced by a few extreme values, we capped positive returns at 100%. This limit allows for a more balanced evaluation of the performance of different models by reducing the impact of exceptionally profitable transactions.

- The generalist model (trained on all windows) produces anomalies associated with returns close to the average of the complete dataset (0.52% for the Transformer, compared to 0.53% for all data), not demonstrating significant directional predictive capability.
- The model trained on windows preceding declines (bearish model) remarkably identifies configurations followed by significant increases. This phenomenon is particularly pronounced with the Transformer, which achieves an average return of 1.61% and a median of 0.67%, significantly higher than the values of the complete dataset (0.53% and 0.19% respectively). The bearish Auto-encoder also performs well with average returns of 0.76% and a median of 0.55%.
- The model trained on windows preceding rises (bullish model) fails to effectively detect bearish configurations, contrary to our hypothesis. The performances remain similar to those of the generalist model for the Transformer (0.66% versus 0.52%), and although the bullish Auto-encoder presents an average return of 0.77%, its very low median (0.06%) indicates that this average is influenced by a few extreme values, without true systematic predictive capability.

These results partially validate our hypothesis: targeted training on bearish configurations indeed allows models to develop sensitivity to specific temporal structures, thus giving them the ability to identify, by contrast, market configurations likely to evolve in the opposite direction (upward). However, this phenomenon is not symmetric, as training on bullish configurations fails to effectively identify structures preceding declines. This asymmetry is particularly visible with the Transformer architecture, which seems to better capture the complex temporal dependencies preceding certain trend reversals, but only in the bearish-to-bullish direction.

D. Temporal Analysis and Strategy Development

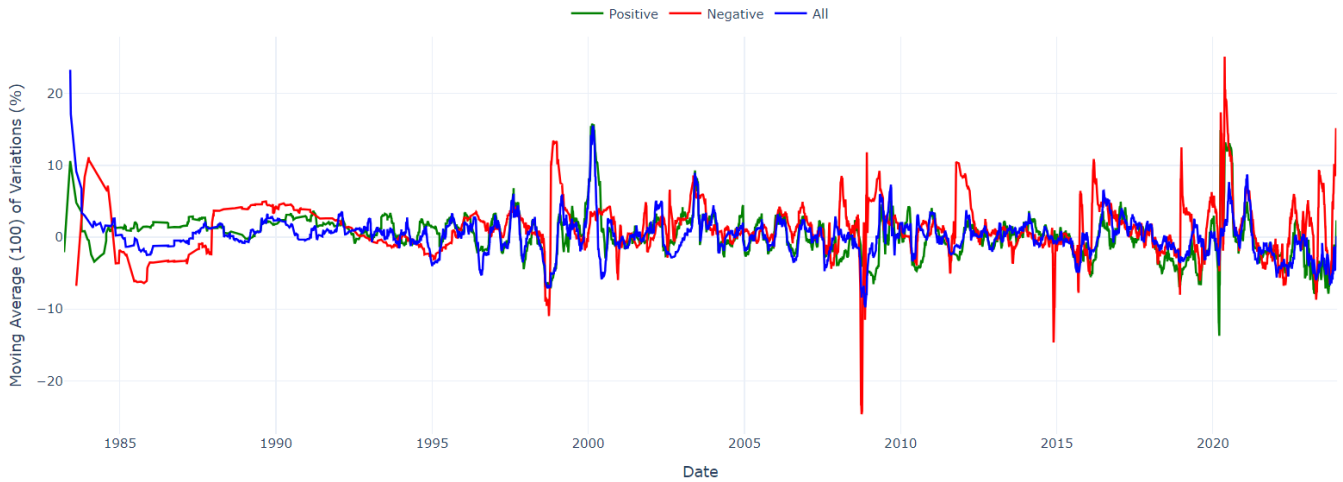


Fig. 5: Moving average of returns at horizon d+20 for anomalies detected by the Transformer (98th quantile). Comparison of models: bearish (red), bullish (green), and generalist (blue)

The chronological analysis of anomalies detected by the model trained on windows preceding bearish movements reveals remarkable patterns. By calculating a moving average over 100 price variations associated with identified anomalies (for the 98th quantile), we observe significant temporal persistence in the generated signals.

This persistence is characterized by extended periods, ranging from a few days to more than a year, during which the detected anomalies are systematically followed by price movements in the same direction. This temporal stability suggests that the model effectively captures fundamental structures in the market data, rather than simply reacting to isolated events or statistical noise.

To exploit this characteristic and concretely evaluate the predictive capacity of the model, we defined a simple trading strategy:

- When the moving average of price variations (i.e., realized trades) over the last 100 detected anomalies rises above zero, we take long positions on newly identified anomalies
- When this moving average falls below zero, we suspend any new positions

This approach, deliberately simplified, aims to evaluate the robustness of the signal generated by our model without introducing additional complexity that could mask its intrinsic performance. We deliberately chose not to introduce short positions when the moving average becomes negative, in order to maintain clarity in the analysis and avoid biases related to specific constraints associated with short selling.

The results of this strategy applied to anomalies detected by our two model architectures are presented in Table IV:

Metrics	Auto-Encoder	Transformer
Mean (%)	3.07	4.40
Median (%)	1.66	2.38
Standard deviation (%)	13.7	16.71
Trade count	7801	7328

TABLE IV: Comparative performance of the trading strategy based on detected anomalies

These results demonstrate a significant predictive capacity for both architectures, with a notable advantage for the Transformer model which shows higher mean and median returns, albeit at the cost of slightly higher volatility.

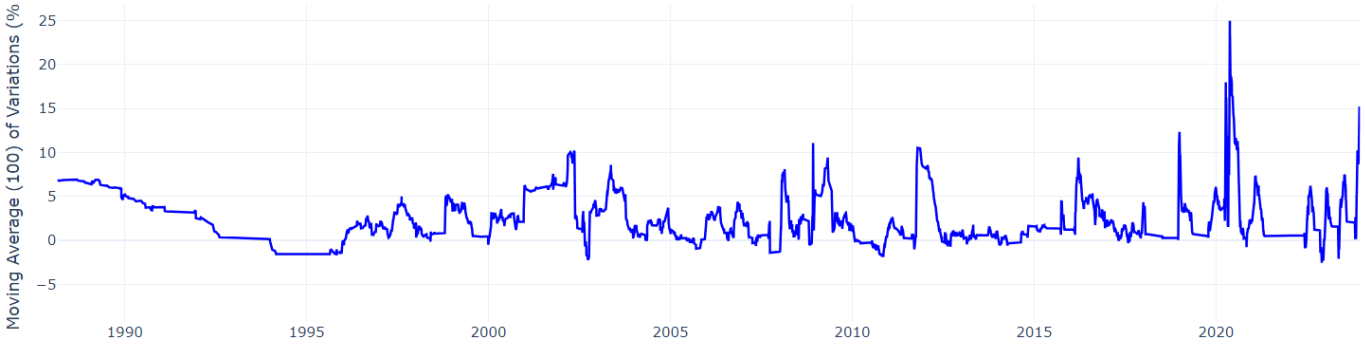


Fig. 6: Moving average of strategy returns at horizon d+20 for anomalies detected by the Transformer (98th quantile).

The ”flat” periods simply mean that the model has stopped trading. Detailed analysis of the trading periods reveals that the moving average of price variations only enters negative territory during short periods. This stability suggests a real capacity to capture temporal structures preceding significant market movements, rather than a mere statistical coincidence.

VIII. Conclusion

This study explores unsupervised learning approaches for anomaly detection in financial time series, comparing a temporal convolutional autoencoder (TCN-AE) with a modified transformer architecture. Our results suggest these models can identify potentially valuable market configurations for trading applications.

An interesting observation from our experiments is the asymmetric behavior during model training: architectures trained on windows preceding bearish movements appear to develop stronger discriminative capabilities for identifying subsequent upward movements. This may suggest differences in the temporal structures between bullish and bearish market phases, though further investigation is needed to confirm this hypothesis.

The transformer model showed better performance (average returns of 1.61% compared to 0.76% for the TCN-AE), suggesting attention mechanisms may offer advantages for capturing dependencies in financial time series. When implemented within a trading strategy, both models demonstrated potential, with average returns reaching 4.40% and 3.07% respectively.

The temporal persistence of signals generated by our models suggests they may be capturing more than random noise, though it’s important to note that our backtest does not account for transaction costs, slippage, or liquidity constraints that

would impact real-world implementation. Additionally, financial markets evolve continuously, and patterns detected historically may not persist in the future.

This work adds to the research on unsupervised anomaly detection in financial markets, while acknowledging that our approach may contain biases or limitations we have not identified. The observed asymmetry between models trained in different market contexts might warrant further investigation into market structures and their predictability.

Disclaimer

> These results should be interpreted with caution and in no way claim to offer an infallible formula for generating profits. The observed historical performances do not guarantee similar results in a real-time market context. This study primarily demonstrates the discriminative capacity of the model in a controlled experimental framework.

APPENDIX

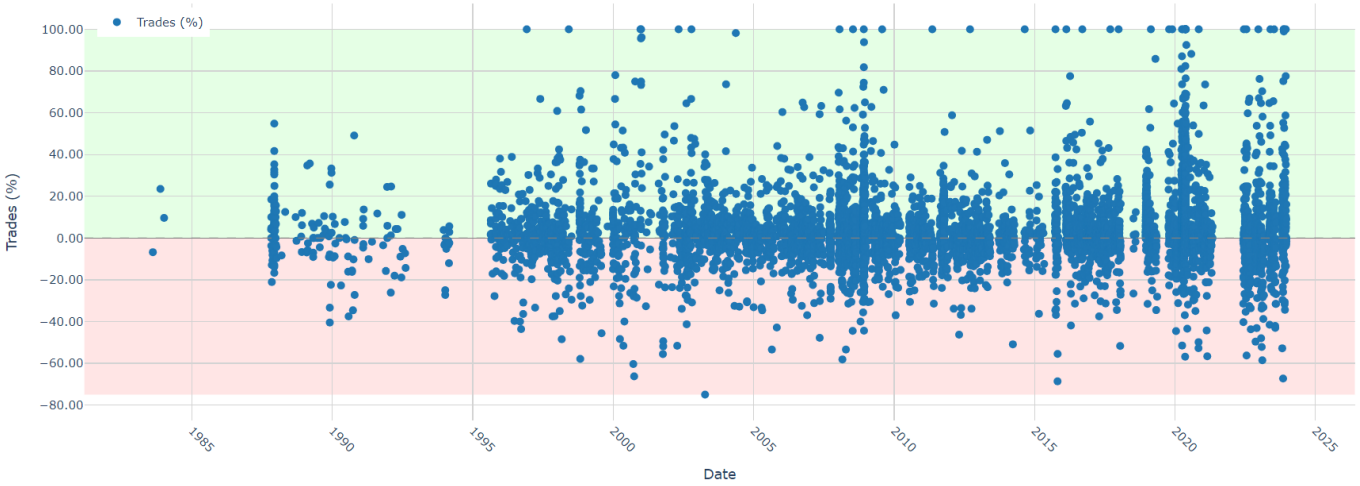


Fig. 7: Trade returns (%) generated by the anomaly detection strategy across market regimes (1985-2023) (Transformer Quantile 0.98)

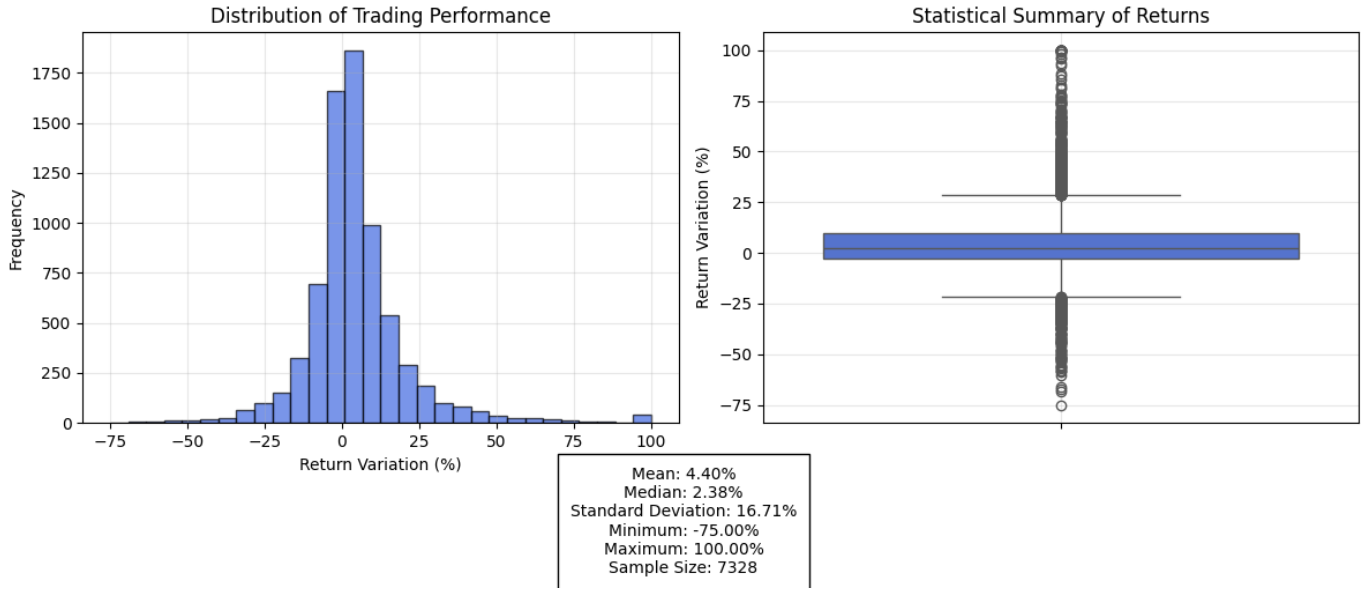


Fig. 8: Distribution and statistical analysis of returns from Transformer-based anomaly detection strategy (Quantile 0.98)

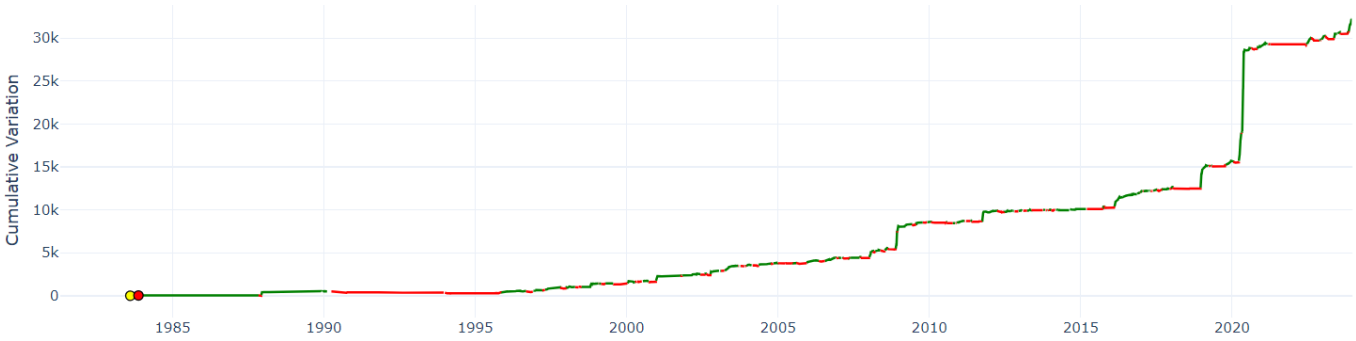


Fig. 9: Cumulative performance from Transformer-based anomaly detection strategy (Quantile 0.98)

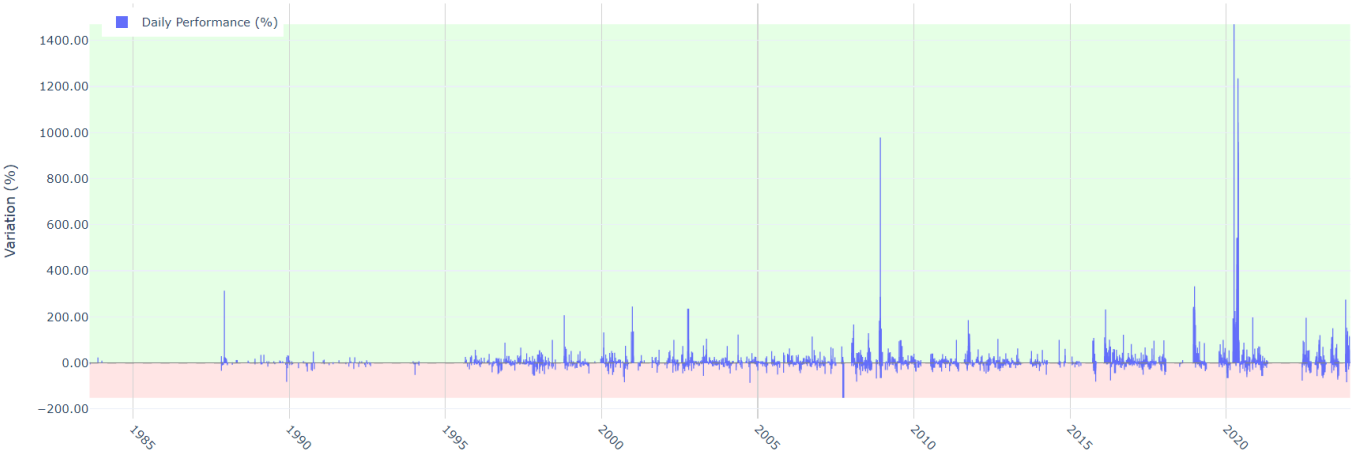


Fig. 10: Strategy Performance Transformer-based anomaly detection strategy: Daily Variation Analysis (Quantile 0.98)

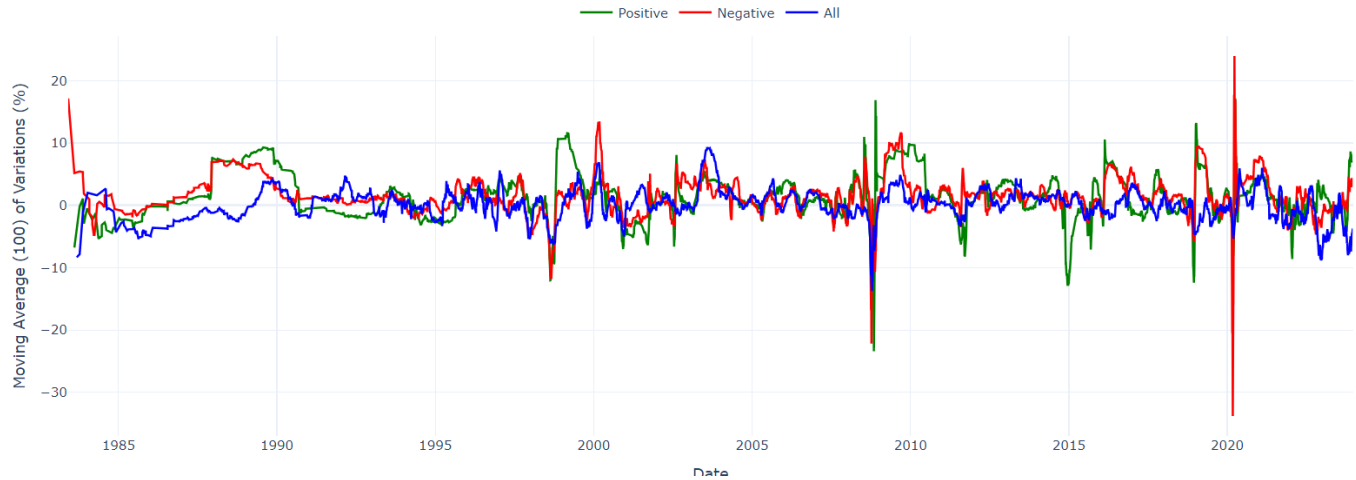


Fig. 11: Moving average of returns at horizon d+20 for anomalies detected by the Auto-Encoder (98th quantile). Comparison of models: bearish (red), bullish (green), and generalist (blue)

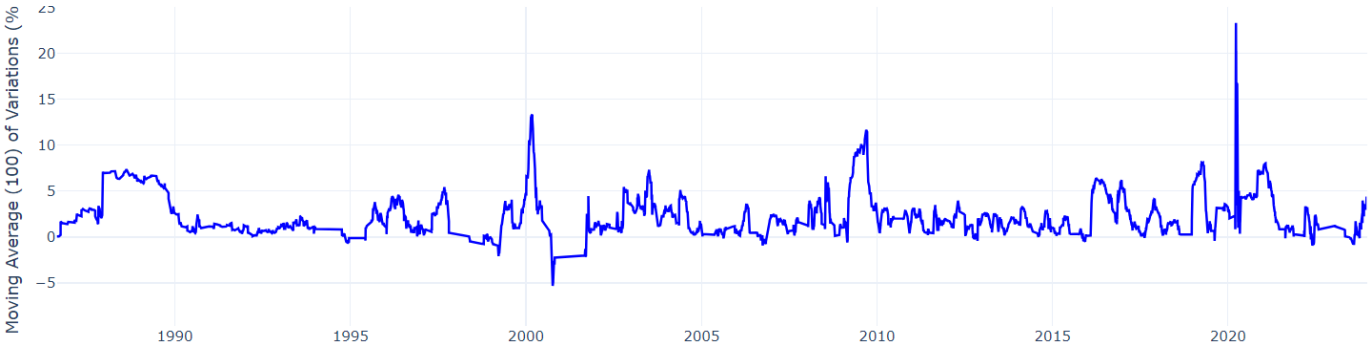


Fig. 12: Moving average of strategy returns at horizon d+20 for anomalies detected by the Auto-Encoder (98th quantile).

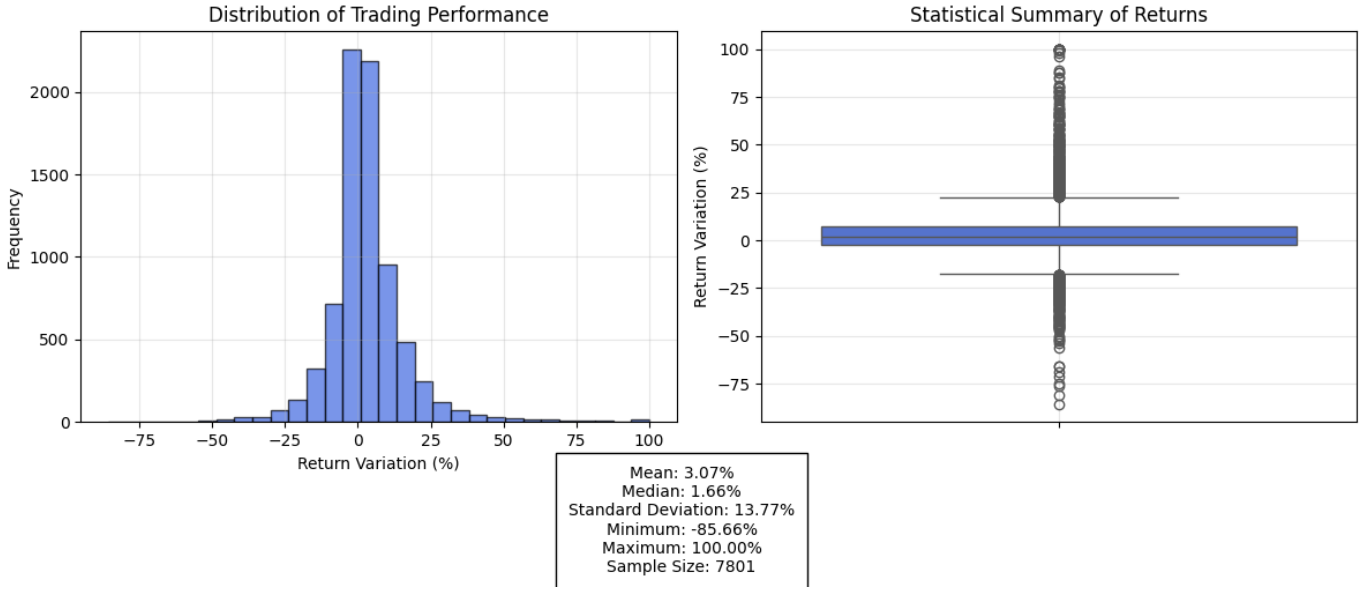


Fig. 13: Distribution and statistical analysis of returns from Auto-Encoder-based anomaly detection strategy (Quantile 0.98)

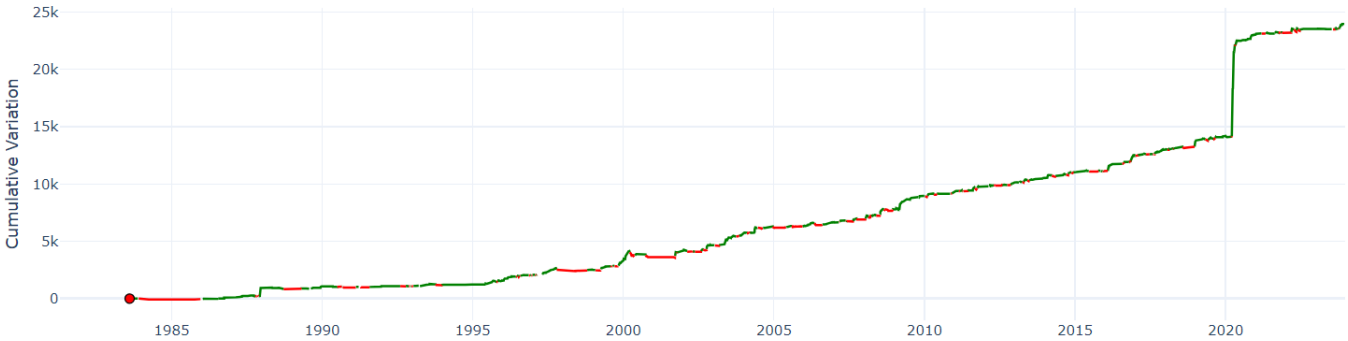
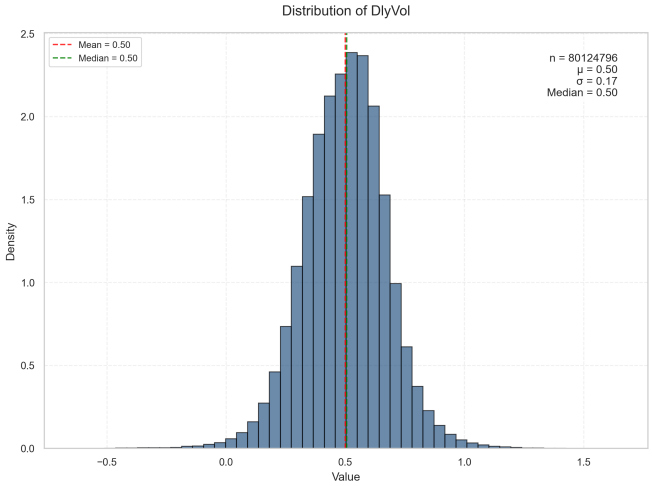
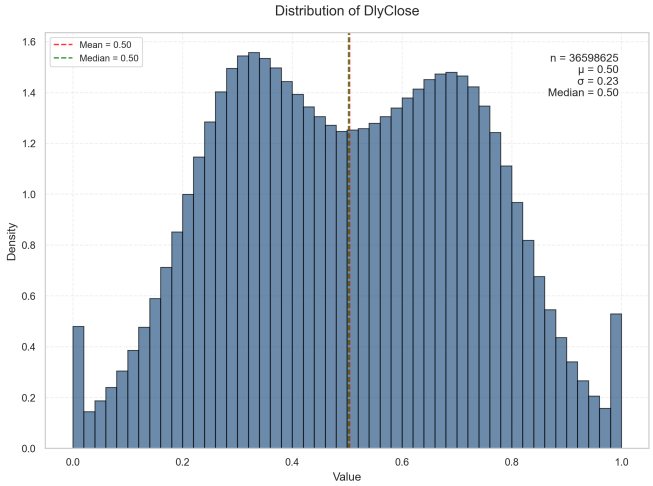


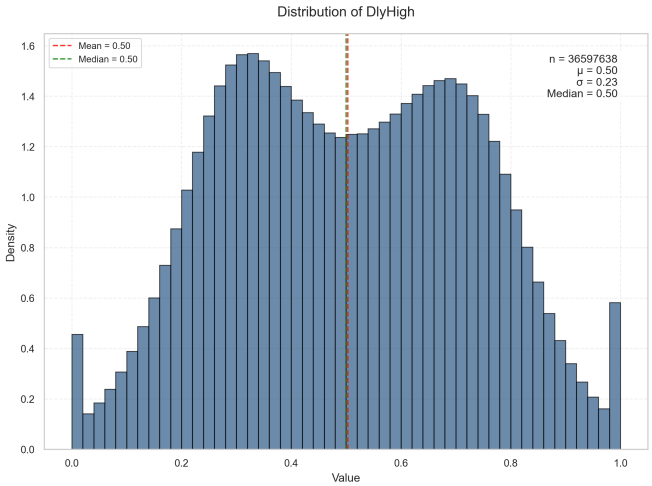
Fig. 14: Cumulative performance from Auto-Encoder-based anomaly detection strategy (Quantile 0.98)



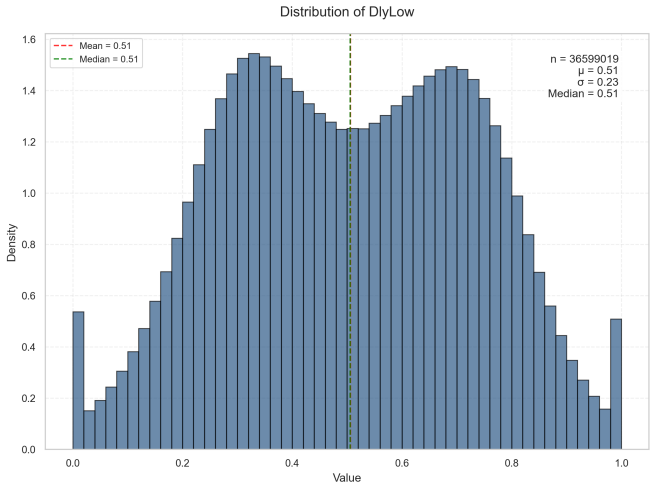
(a) Volume distribution



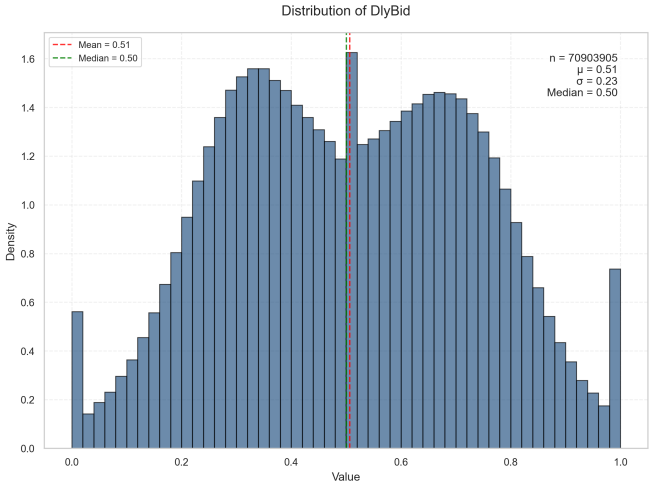
(b) Closing price distribution



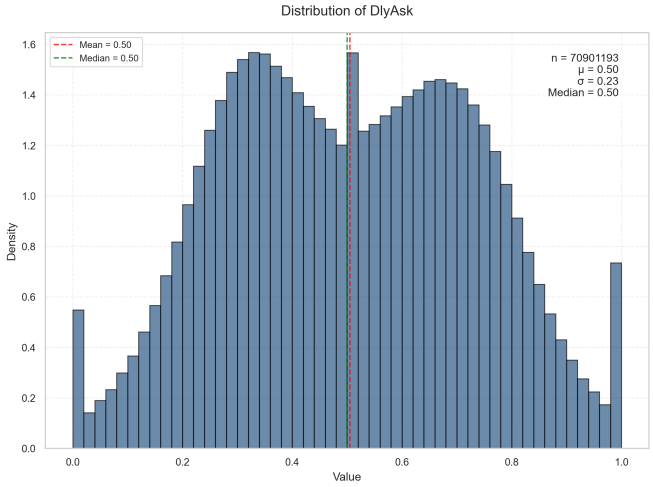
(c) High price distribution



(d) Low price distribution



(e) Bid price distribution



(f) Ask price distribution

Fig. 15: Distributions of different variables

REFERENCES

[1] Markus Thill, Wolfgang Konen, Hao Wang, and Thomas Bäck. Temporal convolutional autoencoder for unsupervised anomaly detection in time series. *Applied Soft Computing*, 112:107751, 2021.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008, 2017.