FINTECH FINAL PROJECT A.Y. 2024/25

Early Warning System

POLITECNICO MILANO 1863

GROUP 21:

Vincenzo Martino Pio Arena
Micol Cavazzoli
Marco De Luca
Alessandro Morra
Federico Savini

TABLE OF CONTENTS

AIM OF THE PROJECT: PIPELINE

In this project we applied machine learning techniques to identify shifting market regimes, offering a robust tool for detecting financial stress and adapting strategies to changing macroeconomic conditions.

- Uses macroeconomic indices and anomaly labels
- Tests multiple models for best performance
- Classifies risk-on vs. risk-off phases
- Investigates local and global explainability

AIM OF THE PROJECT: WHY ANOMALY DETECTION?

Early detection of anomalies in the market is a key issue in finance:

- Support risk management by providing early warnings of volatility spikes or financial stress
- Improve decision-making by flagging outliers that require attention or deeper analysis
- Safeguard financial stability: anomalies can precede market crashes, liquidity crises, or bubbles. Detecting them early helps regulators and institutions take preventive action.

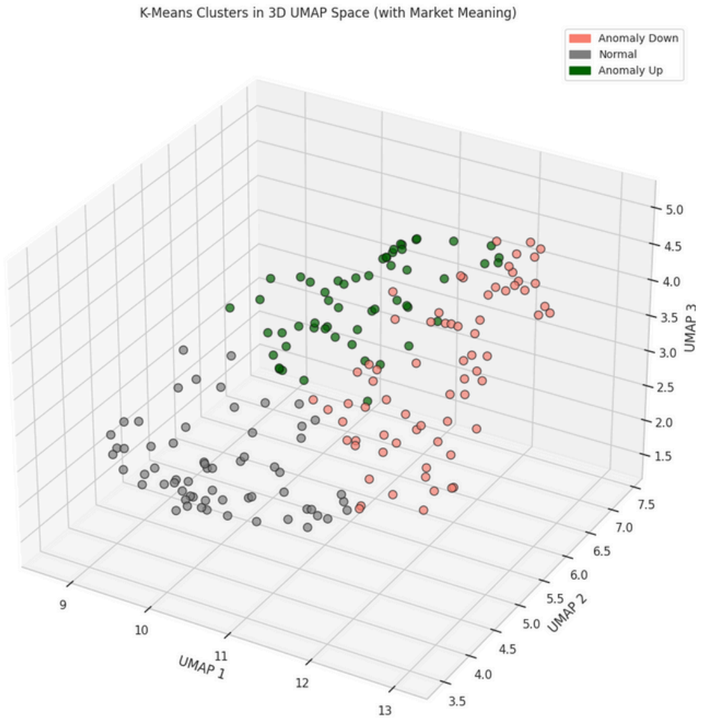
DATA PREPROCESSING: STATIONARITY & SCALING

- We used Augmented Dickey-Fuller test to check statistical evidence for non-stationarity. Non-stationary data were transformed, applying log or simple differences, in order to avoid misleading models.

  Only the following indices were already stationary:
  ECSURPUS | ADF Statistic: -6.492 | p-value: 0.000| ✓ Stationary
  US0001M | ADF Statistic: -2.944 | p-value: 0.040| ✓ Stationary
  VIX | ADF Statistic: -3.971 | p-value: 0.002| ✓ Stationary

- To ensure consistent input ranges and improve algorithm performance, all features were standardized using z-score normalization.

DATA VISUALIZATION: UMAP & UP-DOWN ANALYSIS



From the UMAP visualization, we can observe an apparent cluster structure: the three groups of data are each concentrated in distinct regions of the UMAP space.

Visualization is a fundamental part of a good analysis. We decided to use 3D UMAP, a dimensionality reduction technique which deals with non-linearity.

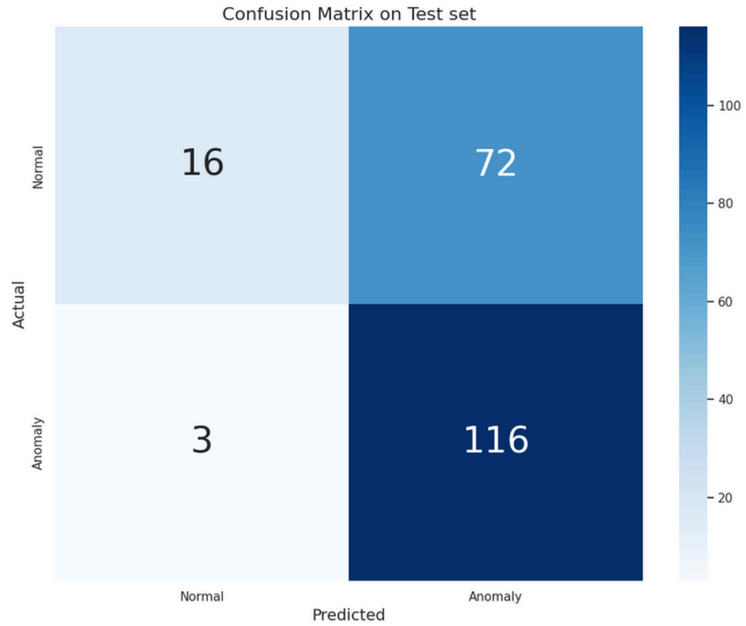
We built a global market index as a linear combination of regional indices, using their market capitalizations as weights, and of the volatility index to consider market uncertainty.
By applying this composite index, we labeled each anomaly as either “up” (positive market stress) or “down” (negative market stress), depending on whether the adjusted market signal was above or below the median threshold.

BENCHMARK MODEL: MVG ANOMALY DETECTOR

As a benchmark model, we use a Multivariate Gaussian anomaly detector. This model identifies anomalies as data points that are unlikely according to a Gaussian distribution fitted to the training set.
The detection threshold is selected to maximize the F1 score.

Evaluation metrics	
Precision	0.6170
Recall	0.9748
F1 score	0.7557

While the model achieves excellent recall, its precision is still lacking. There's definitely room for improvement.

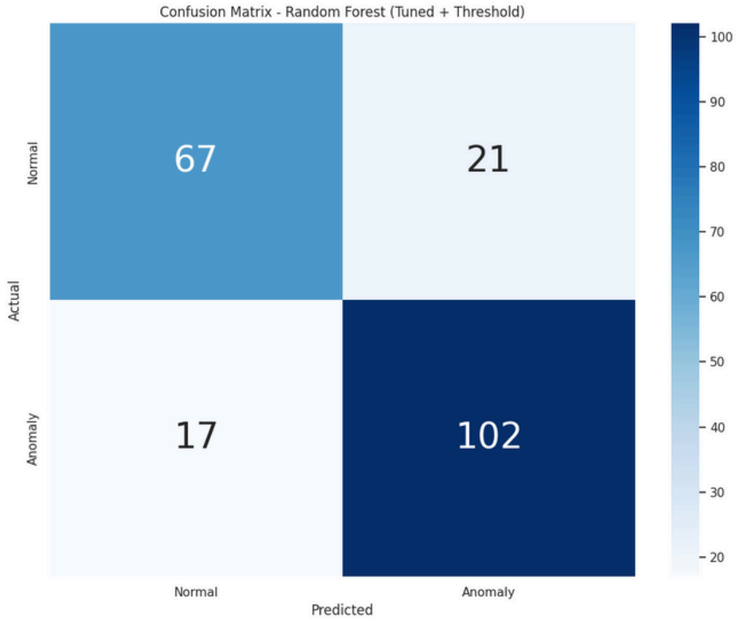


OUR MODELS: RANDOM FOREST

Random forest is a supervised learning model that build multiple decision trees and combine their outputs to improve prediction accuracy. We tuned this model using Optuna, an automatic hyperparameter optimization framework. We obtained the best results maximizing Matthews Correlation Coefficient (MCC), an alternative to F1 score which considers the whole confusion matrix.

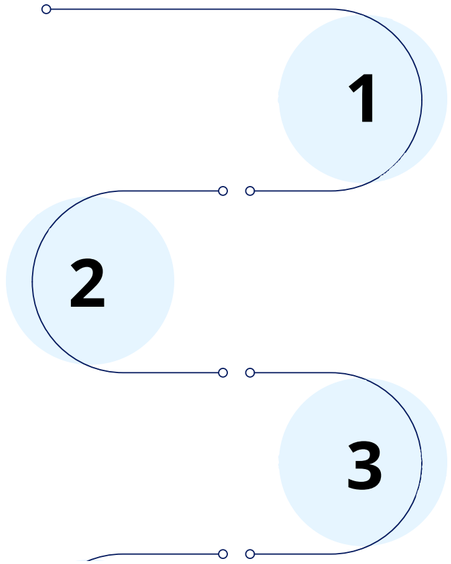
Evaluation metrics	
Precision	0.8293
Recall	0.8571
F1 score	0.8430
MCC	0.6227

This is a solid score, we improved precision, slightly sacrifying recall with respect to benchmark model.



OUR MODELS: ELLIPTIC ENVELOPE

Based on the Gaussianity assumption, Elliptic Envelope is an unsupervised technique that fits an ellipsoid around the data, classifying as anomalies the points beyond the estimated boundary.



Gaussianity check: verified comparing both Shapiro and Jera-Barque p-values

PowerTrasformer: we applied Yeo-Johnson method to make data gaussian-like

Implementation: we run the Elliptic algorithm

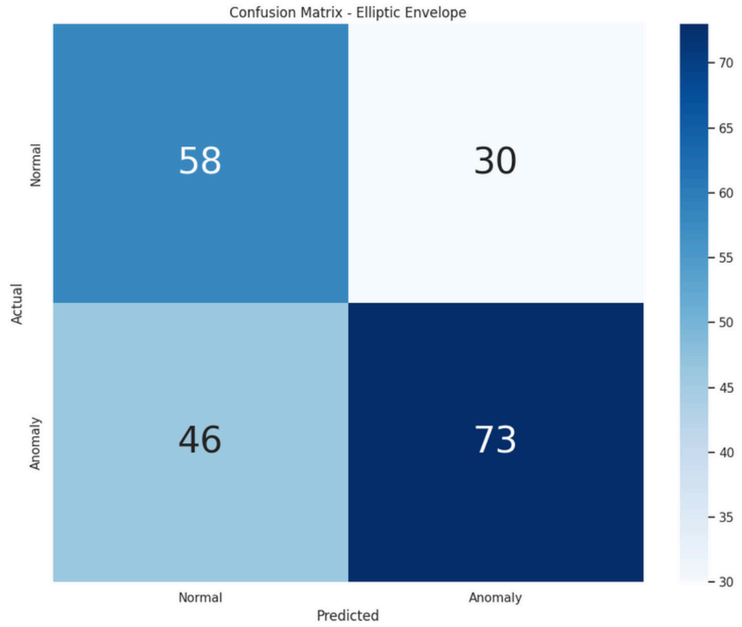
We did not manage to recover normality, probably due to high-dimensionality of the dataset. We decided to look at model performance anyway.

OUR MODELS: ELLIPTIC ENVELOPE

For this and the following models, which are (at least partially) unsupervised we had to estimate the expected anomaly rate from validation set.
This is a fundamental information, the model a priori knows what to expect from the data.

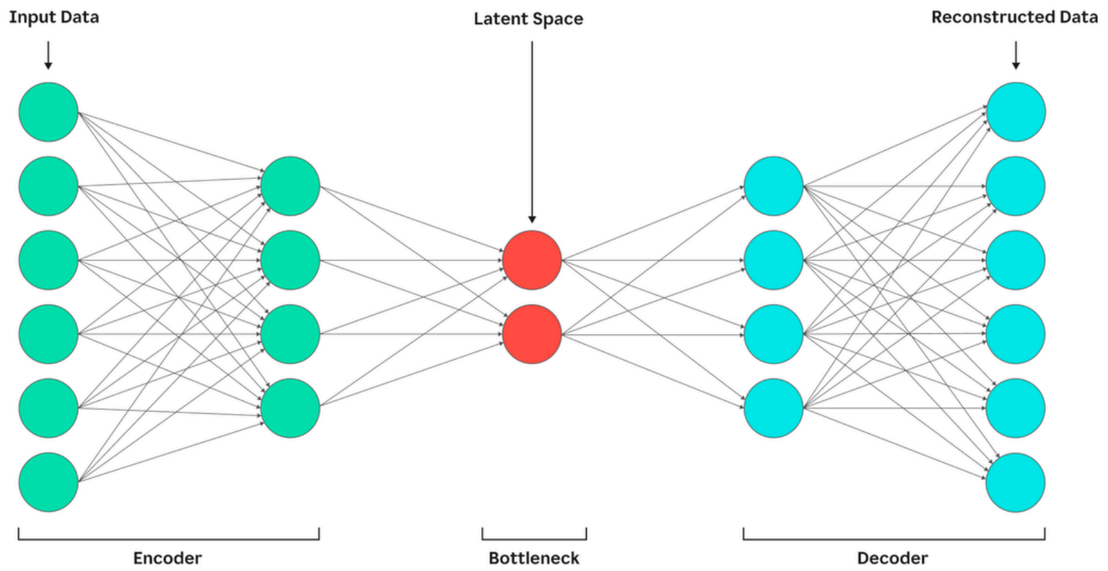
Evaluation metrics	
Precision	0.7087
Recall	0.6134
F1 score	0.6577

Results are not amazing, in particular the model often fails in flagging true anomalies, which is probably the most important task of the project.



OUR MODELS: AUTOENCODER

Autoencoders learn to compress input data into a compact latent representation and then reconstruct it, capturing its underlying structure. Anomalies are flagged based on the reconstruction error - if the model can't recreate it well, it's likely an outlier.

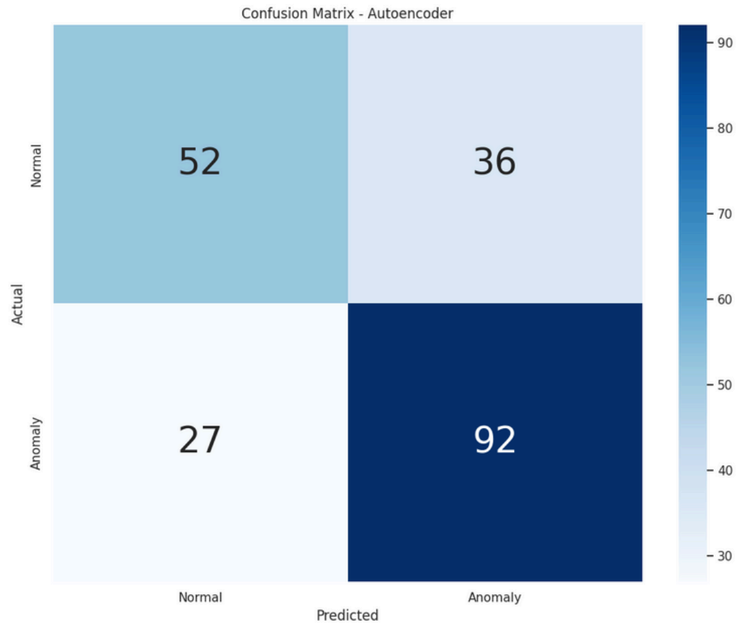


OUR MODELS: AUTOENCODER

We adopted a symmetric architecture, four encoding and four decoding layers surround the bottleneck layer which captures the compressed representation. Batch Normalization was applied after each linear layer to stabilize learning, reduce internal covariate shift, and speed up convergence. It also helps regularize the model and mitigate vanishing gradients in deep architectures.

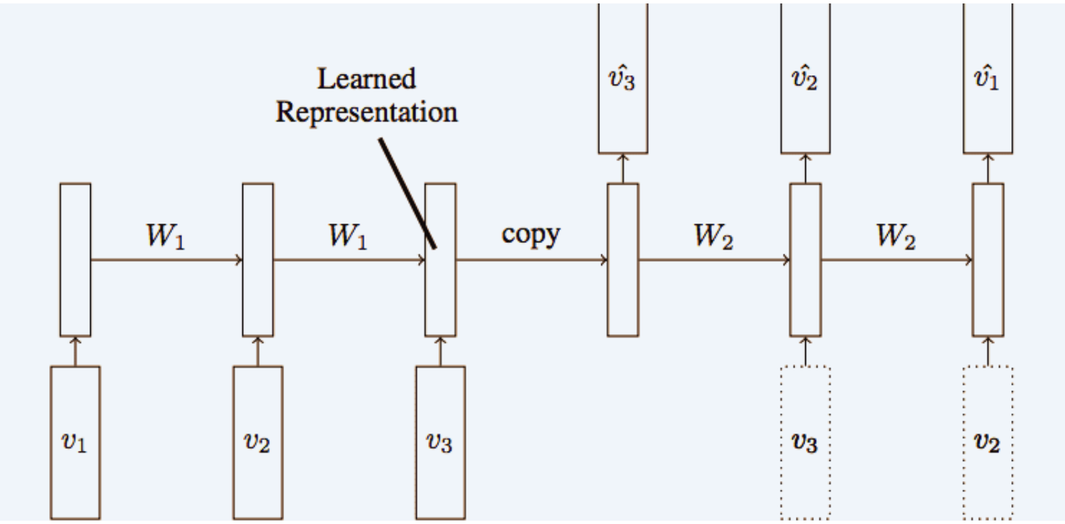
Evaluation metrics	
Precision	0.7188
Recall	0.7731
F1 score	0.7449

Results, despite reasonably goodness, are not enough considering also model complexity. A “simple” random forest performed better on our dataset.



OUR MODELS:
LSTM AUTOENCODER

The natural extension of standard autoencoders for time series data is the LSTM (Long Short-Term Memory) autoencoder.
While the core idea of compressing and reconstructing information remains the same, this model is specifically designed to handle sequential data, enabling it to learn and capture temporal patterns.



We divided our training test and validation sets in several 20-steps rolling windows.
If a particular window is not reconstructed “sufficiently well”, we flag its last time instant as an anomaly.

OUR MODELS:
LSTM AUTOENCODER

- But how to define “sufficiently well”?

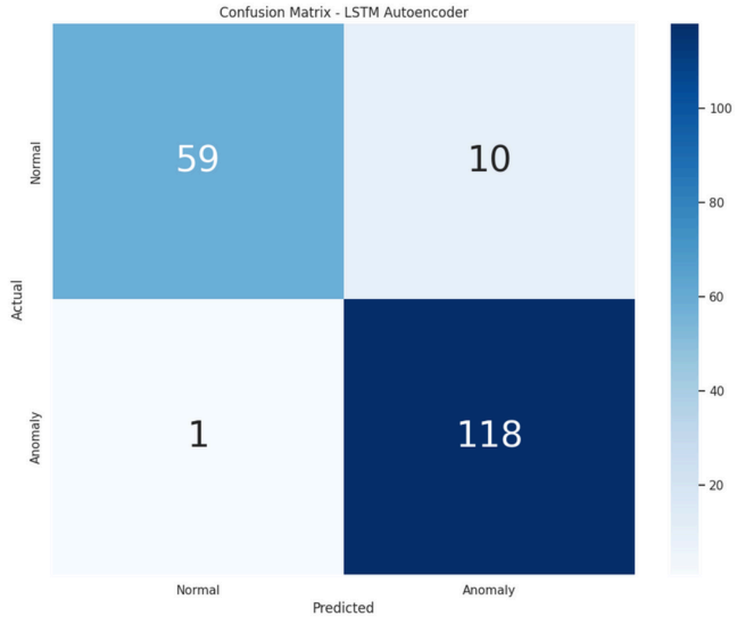
We tuned a reconstruction error treshold using the contamination rate of the validation set (supervised part).

- How to avoid overfitting?

We introduced a dropout coefficient (0.3) between layers and adopted an early stopping rule based on patience.

Evaluation metrics	
Precision	0.9219
Recall	0.9916
F1 score	0.9555

The model performs excellently in both precision and recall, suggesting that information from recent past observations is a key factor in detecting anomalies.



CHOICE OF THE BEST MODEL: COMPARISON

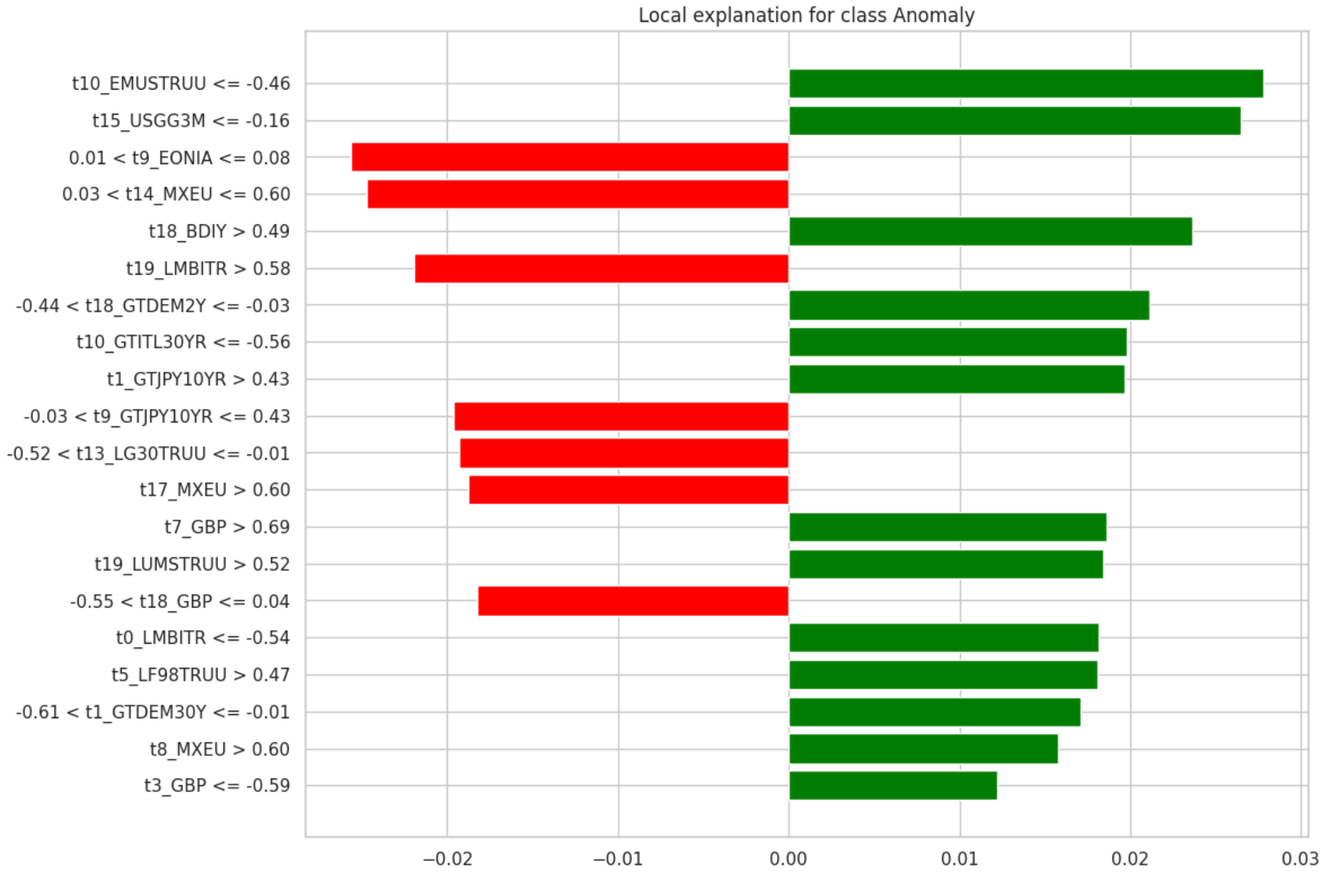
	Precision	Recall	F1 Score
MVG (Benchmark)	0.6170	0.9748	0.7557
Random Forest	0.8293	0.8571	0.8430
Elliptic Envelope	0.7087	0.6134	0.6576
Autoencoder	0.7188	0.7731	0.7449
LSTM	0.9219	0.9916	**0.9555**

As we can observe from the table above, **LSTM autoencoder** is the only model that manages to outperform the benchmark in terms of both precision and recall. For this reason, we selected it as our reference model.
In the following slides, we present the interpretability analyses we conducted on its predictions and feature contributions.

LOCAL EXPLAINABILITY: LIME

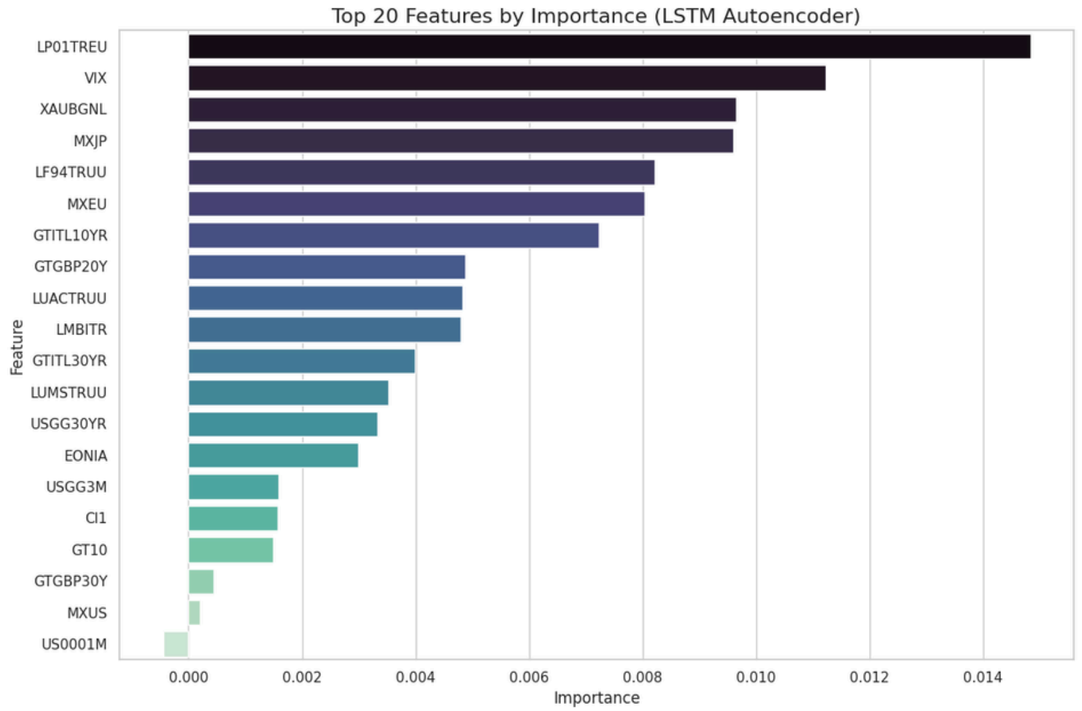
LIME is a local interpretability technique that approximates locally models with simpler ones in order to show which features influence the most a specific prediction. We applied LIME to our LSTM model and extract the 20 most influential features.

RED = Normality
GREEN = Anomaly



GLOBAL EXPLAINABILITY: FEATURE PERTURBATION

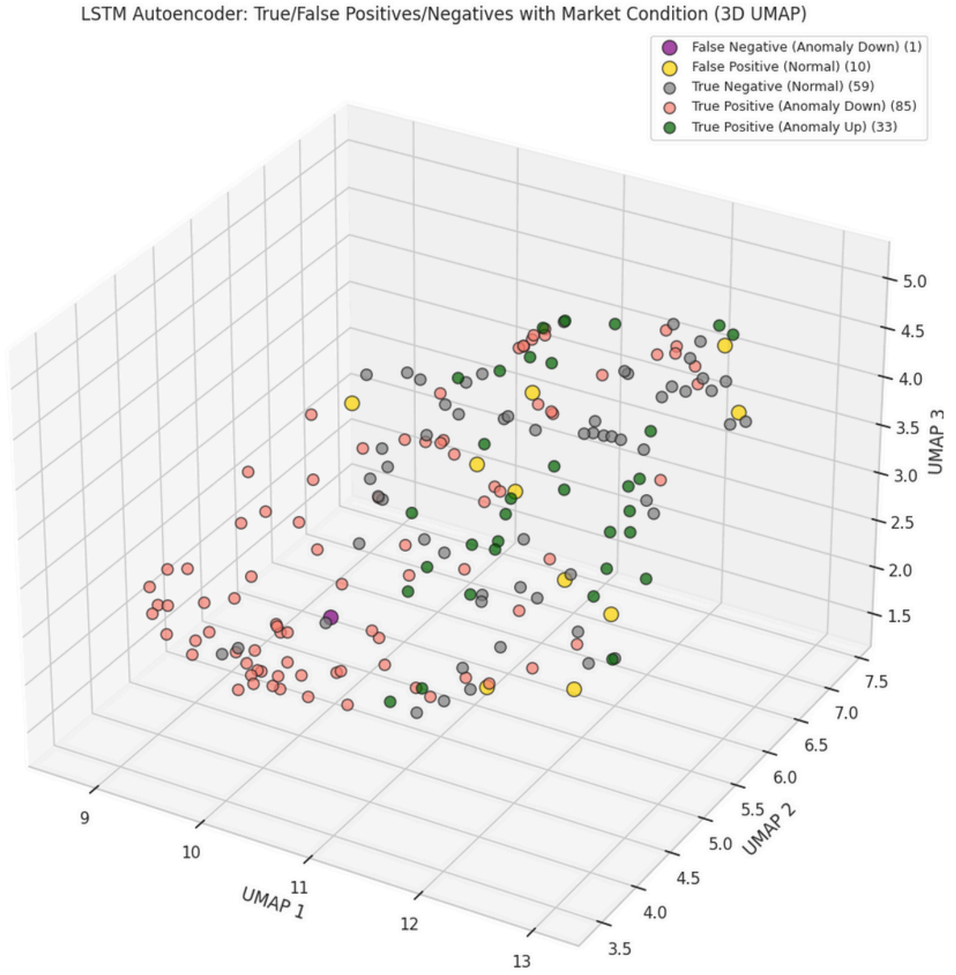
We analyzed global feature significance in the LSTM model using the perturbation-based feature importance approach.
This method evaluates the impact of each feature by measuring the increase in reconstruction error when that feature is removed across all time steps.



Economically relevant indices, such as MXEU (MSCI Europe Equity Index), emerged as significant in both global and local feature importance analyses. This consistency with financial intuition is a positive sign that the model is capturing meaningful patterns.

VISUALIZATION: UP & DOWN LSTM

As can be observed from the graph, our model demonstrates strong classification power in detecting both “up” and “down” anomalies, indicating no significant directional bias in its performance. Furthermore, nearly all misclassified points appear to be concentrated near the central region of the UMAP space, suggesting that these instances may be inherently more ambiguous or harder to separate.



FINAL CONSIDERATIONS

Most relevant findings and results of our project:

- Anomalies showed a cluster-like pattern when data were visualized in the UMAP space, moreover misclassified observations appear more concentrated in the central region.

- The LSTM autoencoder achieved a near-perfect performance on test set. The combination of data compression and reconstruction, together with the LSTM architecture's ability to capture temporal dependencies, turned out to be extremely effective.

- The results of the feature importance analysis seem to align with those that could be expected by financial reasoning.