ANOMALY DETECTION

Business Case 3

Agostini Chiara
Di Blasio Giuseppe
Ferro Andrea
Gattinoni Valentina
Venturellli Matteo

THE PROBLEM

Our project focused on evaluating the anomaly detection model's performance using a practical and **business-oriented approach.**

A labeled test set obtained by splitting the dataset would contain around 100-200 samples. Given that anomalies are much rarer than normal data, even one misclassified anomaly could affect the model's performance. Hence, two important decisions were made accordingly.

OUR PROJECT

1. We implemented multiple anomaly detection techniques, then the most effective models were combined to create a single, more robust model through **model averaging**. This promoted diversification and error mitigation.

2. The predictions generated by the anomaly detection model on additional unlabeled data were utilized to regulate the risk tolerance of a **simulated investment portfolio**. Our customized portfolio outperformed all benchmarks.

DATASET

80%

The different models that we have decided to use, have been trained considering the selected features as input.

TRAINING SET

20%

The labeled data is used to assess the performance of the trained model and optimize the best threshold based on the highest F1/RECALL score

VALIDATION SET

The models were evaluated on a dataset composed by the same features with values ranging from 04-2021 to 05-2023. These historical asset data were downloaded from Bloomberg and then converted into the same currency (USD)

TEST SET

The idea behind the usage of an external, unlabeled test set was to provide a realistic connotation to the project, as in practice labeled data is difficult to obtain since it requires a vast knowledge of the field.

DATA PREPROCESSING AND EXPLORATION

Feature Analysis

Stationarity Analysis

FEATURE ANALYSIS

Asset selection

Starting from the original dataset, we remove highly correlated features and redundant assets (i.e. features that provide the same information)

Weekly Returns

We transform non-index and non-label data into weekly returns. This operation not only rescales data of different features to have similar values, but also allows to obtain stationary data.

S&P500/Gold

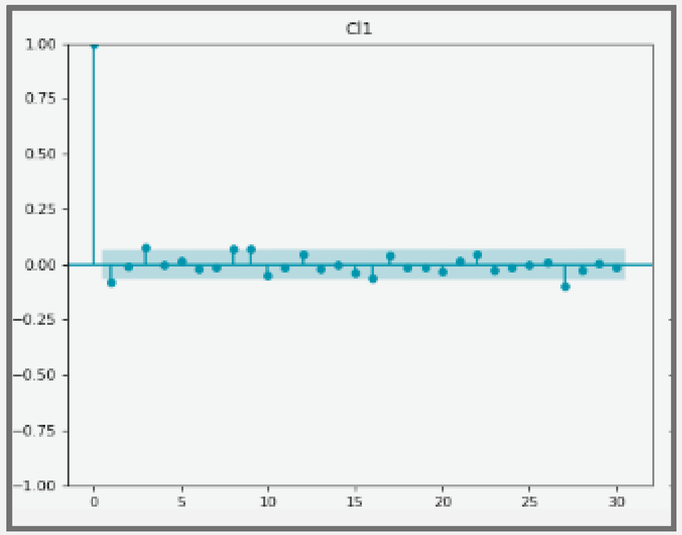
We add the new index S&P500/Gold. This index provides a relative measure of risk appetite in the market. When the ratio is rising, it suggests that the stock market is outperforming gold, indicating a risk-on sentiment.

STATIONARITY ANALYSIS

Statistical Test

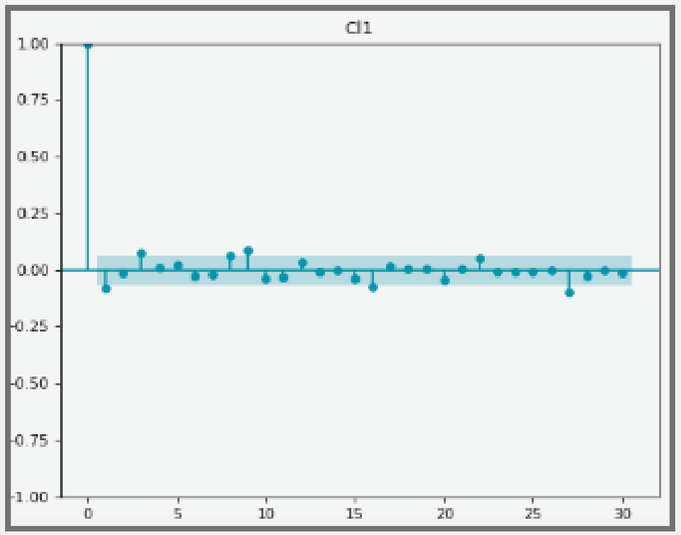
ACF

Autocorrelation Function plot shows no significant autocorrelation in the data after the first lag.



PACF

Partial Autocorrelation Function plot suggests no autoregressive behavior: the only significant lags are meaningless from a financial perspective.



Augmented Dickey-Fuller

This test outputs a p-value smaller than 0.05. The null hypothesis of a Unit Root is then rejected. There is sufficient evidence to support the alternative hypothesis of stationarity.

ANOMALY DETECTION MODELS

1

SUPPORT VECTOR MACHINES

2

LSTM

3

CONVOLUTIONAL AUTOENCODER

4

COPULA

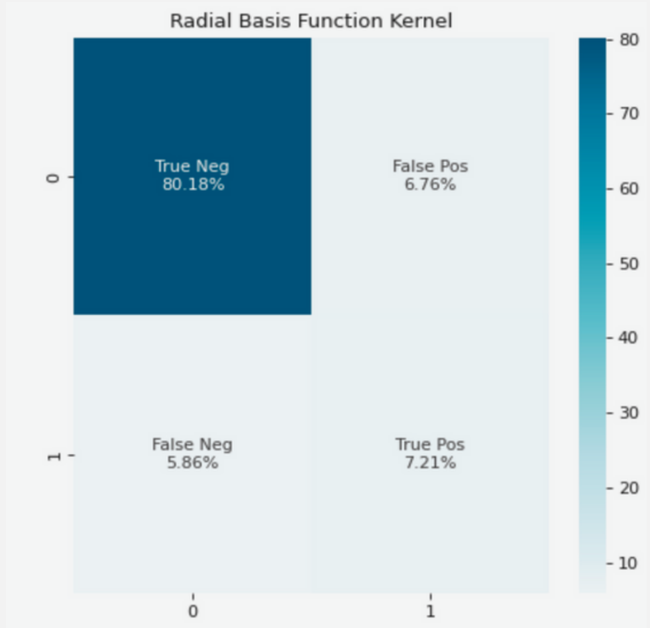
SVM

SUPERVISED BINARY CLASSIFICATION APPROACH

- Robust to noise and outliers
- Effectively captures nonlinear relationships

- Difficulty in handling temporal dependencies in data

Confusion Matrix



Metrics

Precision: 0.516
Recall: 0.552
F1 Score: 0.533

Data from validation set

- **Training set:** the model is trained on single time step instances of the normalized dataset.
- **Procedure:** linear and radial basis function kernels are tested. A larger weight is given to anomaly samples with respect to normal samples due to dataset imbalance. Parameter selection is done by maximizing validation recall.
- **SVM output:** the output of the SVM model is a binary label {0, 1}. This model does not allow to output a class probability.

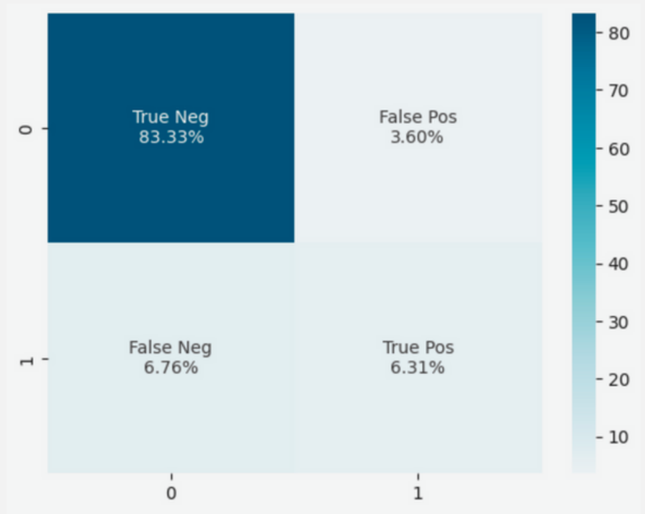
LSTM

SUPERVISED BINARY CLASSIFICATION APPROACH

- Exploits information on past observations
- Designed for time series

- Complex network models have a high risk of overfitting

Confusion Matrix



Metrics

Precision: 0.636
Recall: 0.483
F1 Score: 0.549

Data from validation set

- **Training set:** the model is trained on sliding windows made of 50 time steps. Each window is associated with the label corresponding to the last time step of the window.

- **Procedure:** the parameters of the model and the binary classification threshold were selected to maximize the F1 score on the validation set. The number of trainable parameters was reduced (~1000) to prevent overfitting.

- **LSTM output:** the output layer has 1 unit and sigmoid activation to compute the probability for a sample to be classified as an anomaly.

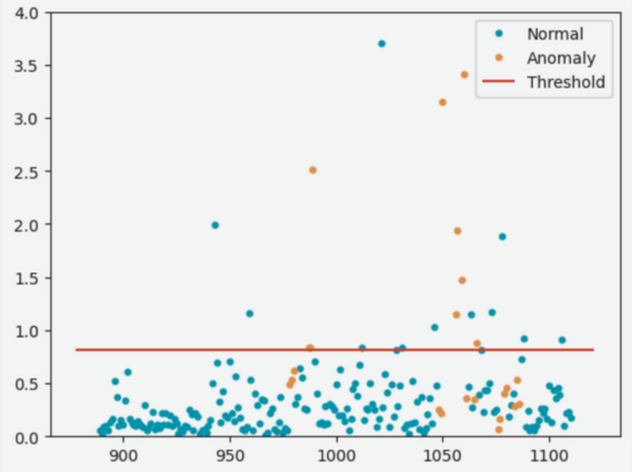
AUTOENCODER

SEMI-SUPERVISED RECONSTRUCTION ERROR APPROACH

- Feature extraction via latent representation
- Effective capture of nonlinear dependencies

- Sensitive to noisy data
- Low data interpretability

Reconstruction Errors



Metrics

Precision: 0.577
Recall: 0.517
F1 Score: 0.545

Data from validation set

- **Definition of anomaly**: point whose reconstruction error exceeds the specified threshold.
- **Training set:** the model is trained on sliding windows made of 52 time steps containing only "normal" data.
- **Procedure:** the model is fit on normal data and then used to evaluate the reconstruction errors for the validation set. Threshold calibration is made by maximizing the F1 score on the validation set.
- **Autoencoder output:** reconstruction errors are converted to "anomaly probabilities" by using the sigmoid function (centered in the threshold).

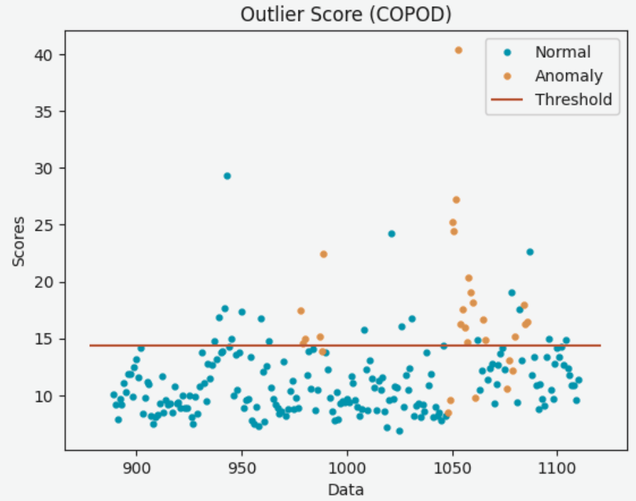
COPOD

SEMI-SUPERVISED EMPIRICAL COPULA APPROACH

- Parameter-free
- Highly interpretable
- Computationally efficient

- Marginal distributions of variables are considered independent

Outlier scores



Data from validation set

Metrics

Precision: 0.55
Recall: 0.7586
F1 Score: 0.6377

- **Definition of outlier**: point differing from the distribution of the training data within a certain threshold.
- **Training set:** observations describing normal behavior only
- **Procedure:** the model is fit on training data and then used to evaluate new observations. The threshold calibration score is made maximazing the F1 score on the validation set.
- **COPOD output:** scores determine the level of "extremeness" for each point. The model identifies scores above the threshold (red line) as anomalies.

BUSINESS ORIENTED TEST

Objective:

Due to the absence of explicit labels in our test set, we evaluate our models by using their outputs to gain insights into their effectiveness in identifying abnormal patterns in financial markets.

Test Data:

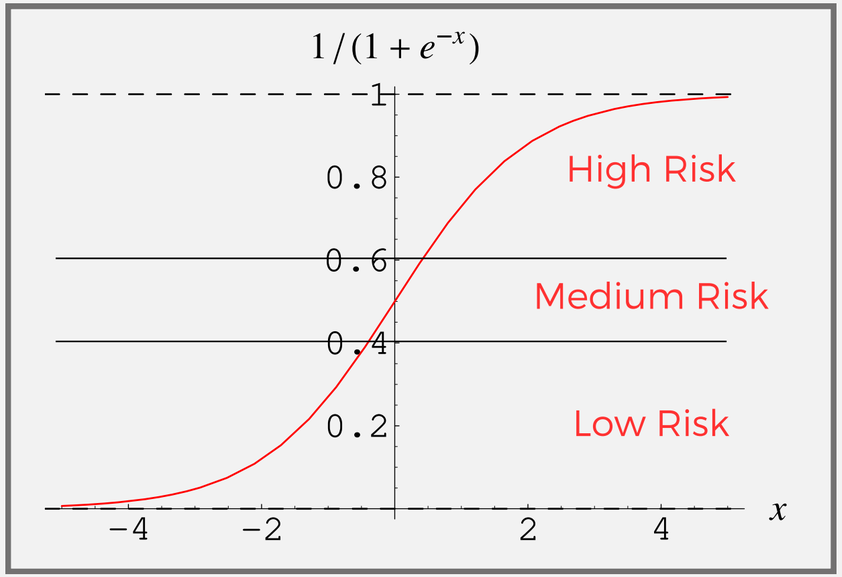
Our test set is downloaded from Bloomberg and contains data up to the most recent dates available. Data are scaled according to the training and validation scaling process.

ENSAMBLE AVERAGING

- **Uniformity in model outputs:** We transformed the outputs using a Sigmoid function to convert them into "probability"-like values ranging from 0 to 1. This enables consistent interpretation and comparison of the outputs across all models.

- **Weighted averaging:** We used F1-Score to assign weights in the voting phase, prioritizing models with higher performance for a more accurate outcome.

- **Result interpretation:** The voting process assigns "probabilities" to market risk labels (Low, Medium, High). We set arbitrary thresholds: given that with a sigmoid the threshold for anomaly detection set at 0.5, data between 0.4 and 0.6 is labeled as Medium Risk, indicating neither risk-on nor risk-off periods.

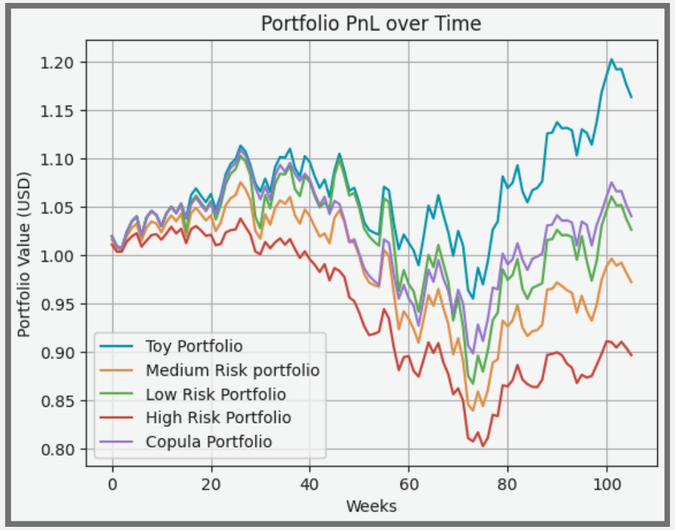


Note: what we refer to as "probability" is not a proper probability object, but a rescaled score that, for our purposes, can be interpreted as such.

PERFORMANCE EVALUATION: TOY PORTFOLIO

ENSAMBLE OF MODELS

TEST DATA



TOY- PORTFOLIO

No trading or speculation is to be intended. It is more *nowcasting* than forecasting

- **Idea:** Mimic a portfolio dynamics evolution in time, whose rebalancing strategy is entirely based on the voting risk-on and risk-off period detection.

- **Assumptions and simplifications:** We have considered a small basket of assets as components of our toy portfolio. The portfolio is exposed to both stock and bond markets, and it also includes commodities. In this way and by means of EFTs and financial indices, the diversification principle has not been completely disregarded. Furthermore, no transaction fees or changes are involved.

- **Time window :** from April 20, 2021 to May 5, 2023 (same as Test set. Here the real risk on- risk off labels are unknow).

- **Currency:** our portfolio value is in USD. Assets prices that are not traded in this currency were converded properly.

Assets

- Stock Market
  - S&P 500 [*^GSPC*]
  - Invesco EURO STOXX 50 UCITS ETF *[SX5E.SW]*

- Bond Market
  - Vanguard EUR Eurozone Government Bond UCITS ETF *[VETY.L]*
  - Vanguard Total Bond Market Index Fund *[VBTLX]*

- Commodities
  - S&P GSCI Index *[^SPGSCI]*
  - Crude Oil Jul 23 *[CL=F]*

ASSET ALLOCATION

Assumption: one-to-one correspondance between an investor's risk exposure and the market risk labels, determined by the asset class that is expected to deliver higher returns during a risk-on/ risk-off period.

Optimizing portfolio allocation weights involves assessing the risk levels associated with various asset classes or investment options. The allocation strategy employed will be contingent upon the risk tolerance, investment objectives of the individual or institution involved, as well as the market label derived from our voting analysis.

	Equity	Bond	Commodities	
Conservative	0.25	0.65	0.05	**High Risk Label**
Moderate	0.50	0.35	0.15	**Medium Risk Label**
Aggressive	0.65	0.15	0.20	**Low Risk Label**

A summary of the weights assigned to each macro-component of our portfolio. The risk labels correspond to the outcomes generated by the voting analysis: in a high risk label, date bonds are expected to yield higher returns; while in a low risk scenario, equity should be the primary asset class to monitor.

EVOLUTION AND COMPARISON

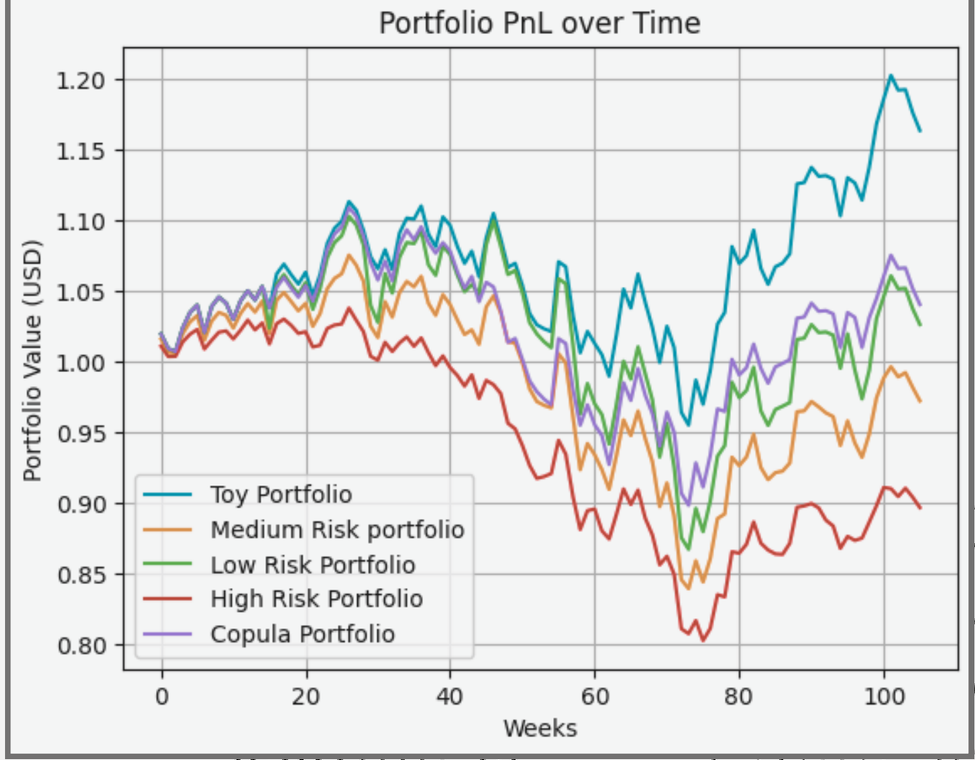
- **Benchmark portfolios:** three portfolios are taken as benchmarks for low, medium and high risk market labels, respectively. They mantain a fixed asset allocation over time.

- **Toy portfolio:** the asset weights are determined by changes in market risk labels, coming from the voting.

- **Copula portfolio:** its composition is determined by changes in market risk labels, obtained via the copod algorithm on the test set. COPOD was the method with the best F1 score according to our analysis.

Comparison between the market performance of our toy portfolio and some benchmarks

Profit and Loss portfolio evolution



FINAL RESULTS AND CONCLUSION

- As expected, the three fixed portfolio benchmarks follow the same trend and their values are just rescaled.

- Despite following the best performing method, the copula portfolio does not outperform its competitors. It seems to replicate the evolution of a low risk portfolio.

- On the other hand, our toy portfolio outperforms the market, suggesting that the voting scheme captures market insights missed by COPOD. This could be the reason why the toy portfolio beats the copula one. Moreover, this highlights the importance of considering time (or even dependance over time, if any) in anomaly detection. This trade-off differentiates more statistical-like approaches like COPOD from artificial neural networks.


THANKS