POLITECNICO
MILANO 1863

Early Warning System

FinTech final project (Prof. R. Zenti - Business Case 3)

Group n.9: M. Bersani, L. Corazza, A. Grignani, G. Radaelli, S. Schembri

Table of contents

Problem Description

Financial Overview

To enhance financial performances and mitigate risks, it is crucial to identify anomalous behavior in financial markets, which periodically tend to crash.

- During normal periods **(risk-on periods)**: investors exhibit a high-risk appetite, driving up the prices of risky assets.
- During crises **(risk-off periods)**: risk premia and financial assets display abnormal behavior as investors become more risk-averse, leading to the sale of risky assets, prices declines, and a shift toward lower-risk investments.

Anomalies in Financial Time Series

Definition: An anomaly in a financial time series refers to an observation or a set of observations that deviate significantly from the expected pattern or behavior of the data.

Causes of anomalies: Market events, economic news, earnings reports, political events.

Goal

Develop an Early Warning System to detect anomalies in financial markets, helping to prevent financial crises and improve investment performances.

Project Structure

The **pipeline** (fig. 1) for the classification of anomalies is divisible into two macro parts: training and actual use of the final model for classification.

The main reason why to create a pipeline is to make each section independent without having to re-execute the individual phases from scratch, but each phase takes as input the product of the previous one saved in intermediate files.

Project Structure

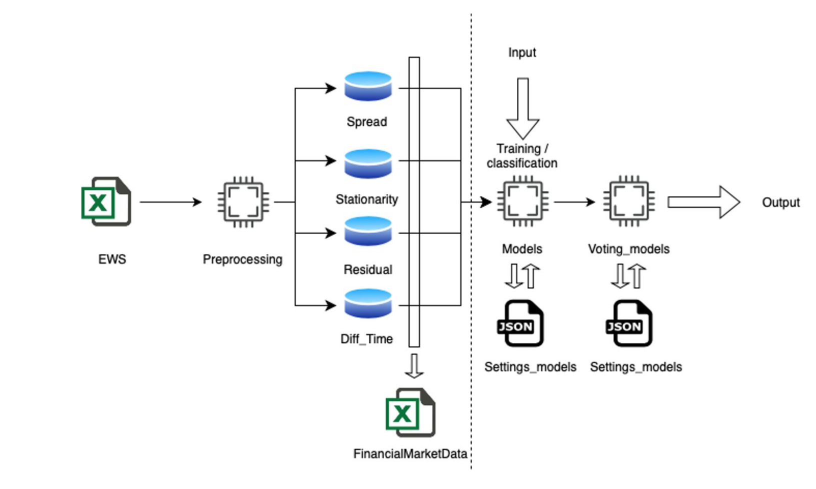

Figure 1: Pipeline

Data

Data Overview - Covariates

Weekly data from Bloomberg of various financial instruments and indices, from 11/01/2000 to 20/04/2021, that provide insights into global economic conditions, market trends, and investment opportunities:

- Key equity indices;
- Bond indices (Global, Corporate IG/HY, Inflation-linked, Municipals, Mortgages);
- Short/medium/long term interest rates;
- Key exchange rates;
- Commodities;
- Leading indicators (Economic surprise, Baltic Dry Index);
- VIX (option implied volatility);
- **A label 'abnormal/normal'**.

Data Overview - Issues

- Unbalanced dataset with nearly an 80/20 ratio between classes;
- Too many features to be able to construct an effective and interpretable model;
- Non-stationary behaviors of the most part of the time series in the dataset.

Data Preprocessing

The objective of the preprocessing is to obtain as much information as possible from the data, in order to boost the performances of the implemented algorithms. For this purpose, new datasets have been constructed, though only a selection has been incorporated into our analysis:

- **Cointegrated data (Stationarity):** After a cointegration analysis between the original variables (fig. 2), the parameters to combine cointegrated variables have been estimated through Ordinary Least Squares.
- **Spread data:** Following the cointegration analysis, a spread variable has been created for each feature w.r.t. two variables ('ECSURPUS' and 'VIX') using the same linear combination for all of them (for consistency). Then, only the stationary ones have been selected.

Data Preprocessing

- **Differences data (DiffTime):** To exploit the problem of independence hypothesis in each model, each observation has been related to its previous and subsequent record. This strategy allows to take into consideration local distributions.
- **Residual data:** The objective of this dataset is to ignore the seasonality and the trend, indeed the provided data is distributed on a long period of time.

When referring to **EWS** dataset, it is intended the original one, provided in the course.

Data Preprocessing

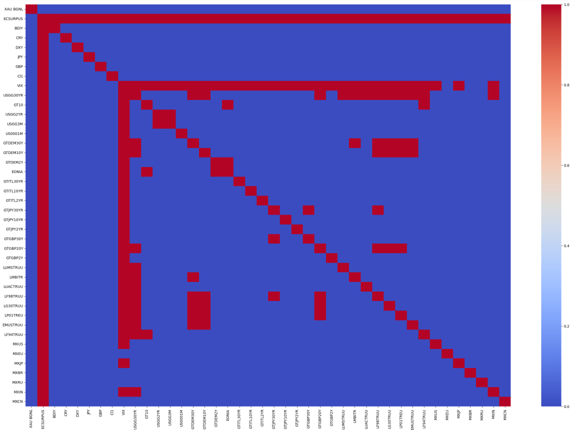

Figure 2: Cointegration matrix: the value is equal to 1 if two coordinates are cointegrated ($p - value \leq threshold$), 0 otherwise

Models

Selected Models

Different models have been tested on the datasets. Their performances (fig. 3) have been evaluated on the F1-score, in order to minimize the number of anomalies not detected. The developed algorithms are:

- **Isolation Forest:** It is commonly used for outlier detection. However, it is unsupervised.
- **Random Forest:** It is an ensemble learning method that builds multiple decision trees to make more accurate and stable predictions.
- **XGBoost:** It is a highly efficient and scalable implementation of gradient boosting that offers advanced regularization to prevent overfitting.
- **KNN:** Through this model, it is possible to identify anomalies as observations with highest distance w.r.t. the others.

For what concerns the hyperparameters of these models, they have been optimized through a Bayesian approach.

Models Performances

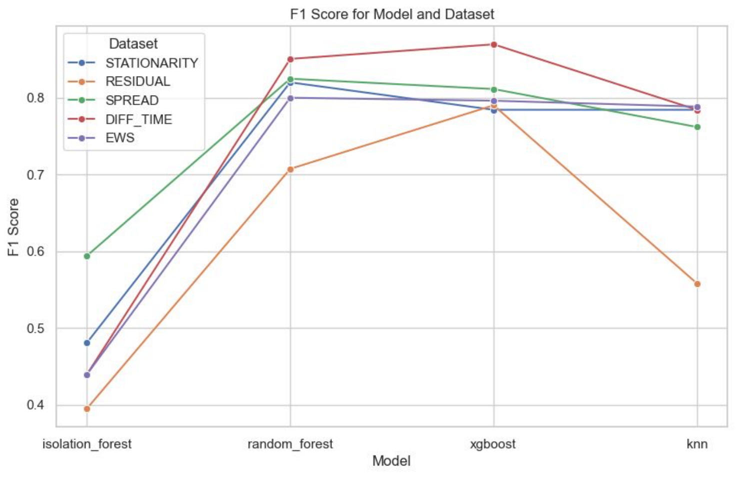

Figure 3: Models performances for every dataset

Early Warning System

Voting System

Idea: Construct a model that is as robust as possible, taking into account the outcomes achieved by each dataset-model pairing. This strategy ensures that if certain models exhibit superior performance during specific periods, their influence will be equitably balanced over the long term.

For this purpose, two possible systems are considered:

- **Democratic:** Each model votes and the majority wins.
- **Meritocratic:** Higher weights are assigned to algorithms with better performances.

EWS 1 - Models and Datasets Selection

The Early Warning System is implemented with models trained on statics datasets (which do not change over time). As of the results in sec. 3 (fig. 3), the best three models were selected. Then, for each of these models, the best two dataset-model combination (only stationary data i.e. not 'Difference' Dataset) have been chosen:

- Random Forest - Spread
- Random Forest - Stationarity
- XGBoost - Spread
- XGBoost - EWS
- KNN - EWS
- KNN - Stationarity

Both the voting systems have been tried on this EWS: the one with better performances (fig. 6 and fig. 7) is the **democratic voting system**.

EWS 1 - Performances

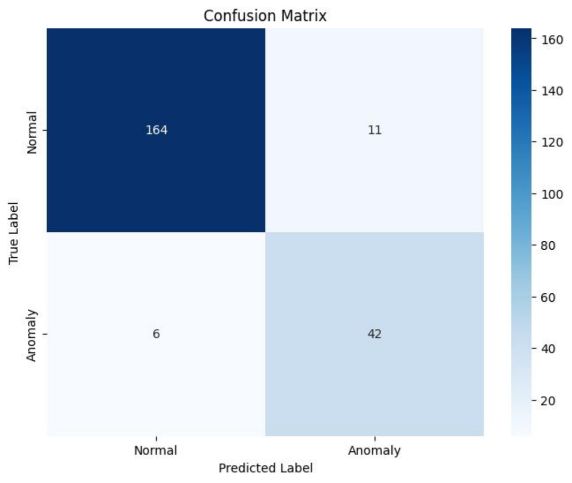

Figure 4: EWS 1: Confusion Matrix

EWS 1 - Performances

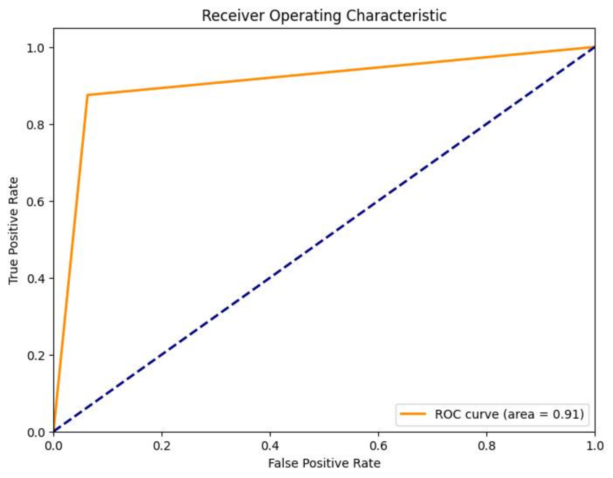

Figure 5: EWS 1: ROC-AUC

Improvements: Temporal Dependence

- **Temporal Dependency Consideration:** Implement a rolling window to account for temporal dependencies to ensure stable model performance over time.
- **Implementation of a Temporal Window:** Employ a temporal window for training and testing on subsets of the dataset, prioritizing the most recent data to enhance classification accuracy.
- **Data Interconnectivity:** Utilize a dataset where individual observations are interconnected, like the 'Difference' dataset, to facilitate a deeper analysis of data dynamics and improve predictive capabilities in financial technology research.

EWS 2 - Models and Datasets Selection

The second Early Warning System is implemented through a rolling window approach. For this purpose, only datasets which consider the dependence of data with time are evaluated:

- Random Forest - DiffTime
- Random Forest - Stationarity
- XGBoost - DiffTime
- XGBoost - Stationarity
- KNN - DiffTime
- KNN - Stationarity

Both the voting systems have been tried on this EWS: the one with better performances (fig. 6 and fig. 7) is the **democratic voting system**.

EWS 2 - Performances

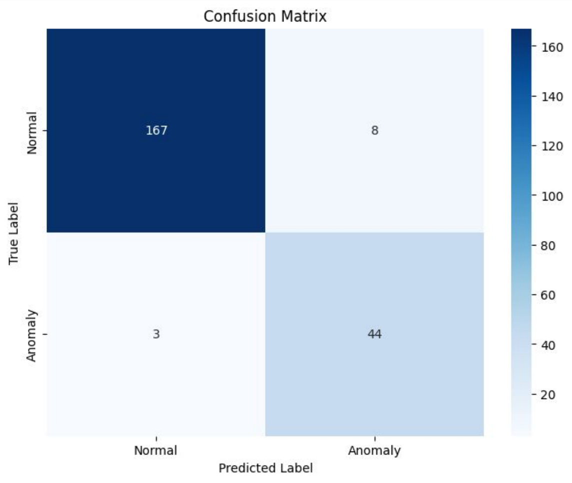

Figure 6: EWS 2: Confusion Matrix

EWS 2 - Performances

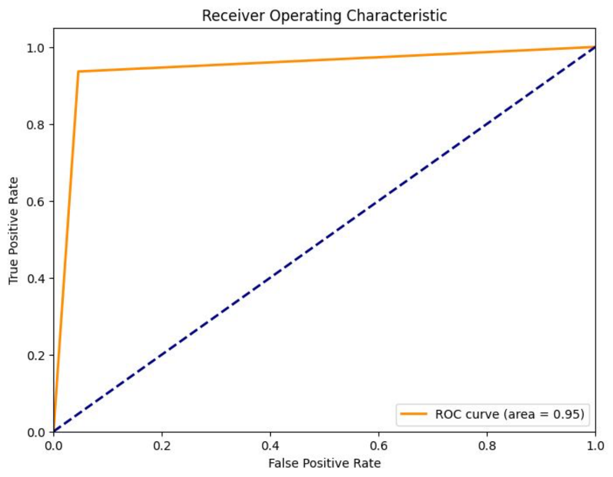

Figure 7: EWS 2: ROC-AUC

Conclusions

Conclusions

In both methodologies, we observe the achievement of the predefined objective: to identify a model that performs well and minimizes the incidence of false negatives. Regarding the two voting systems, the democratic algorithm consistently outperforms in both approaches. This phenomenon can be attributed to the presence of fewer sub-models, potentially fostering a form of 'collaboration' among the algorithms without allowing any single model to dominate the others.

In terms of the two different approaches, one using a static dataset and the other employing a temporal window, it is apparent that the latter exhibits superior performance. This enhanced efficacy can be explained by the fact that the hyperparameters in the temporal window approach are trained across multiple time-evolving datasets, thereby rendering it more adaptable and effective over time.

Conclusions

The two Early Warning Systems could be used synergistically: the first to react promptly to market variations, and the second (which includes subsequent observations) to confirm the decisions based on the first. The initial EWS allows for immediate responsiveness, crucial for minimizing losses and seizing opportunities. The second EWS provides additional data and analysis, verifying initial reactions and offering deeper market insights. This dual approach enhances decision-making and risk management.