



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Spoke 3 - Astrophysics and Cosmos Observations

Intesa Sanpaolo

Spoke 3 General Meeting

12-14 Giugno 2023

Dipartimento di Fisica e Astronomia – Università di
Catania

Use cases:

- Intra Spoke 2, 3 e 10 –
Fraud Detection
- Spoke 3 – Banking Time Series
(Data Quality)



Projects steps

- **Data Exploration** of the dataset
- **Formalization of the problem** and study of the state of the art of the algorithm that fits the goal, with a cutting-edge solution and in the context of Time Series/Fraud detection data.
- Time series/Data **preprocessing** and standardization: build one or more input/output formats for Machine Learning algorithm.
- One step further in the application of the **ML algorithm** is to implement an ad-hoc technique of explainability. It is suggested a State of the art study on **Explainability on TimeSeries** algorithm or classic algorithm.
- Realization of **prototype** in **Python** versioned in **GitHub**.
 - Study the data as the step 1
 - Reframe the problem as in step 2
 - Preprocess the data as in step 3
- Build an explainability algorithm (if not already explainable by design) integrated within the package. Look step 4.



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Fraud Detection



Fraud Detection – spoke context

Contribution from various spoke:

- Spoke 2 - Cutting-edge classical approach on fraud detection (to be selected among anomaly detection, supervised, Time series, ...)
- Spoke 3 - Time series approach and oriented to anomaly detection (not strictly necessary)
- Spoke 10 - Quantum approach to fraud detection



Fraud Detection - description

There is already a **dataset ready** to be used in the experimentations, composed of **515651 rows** and **17 columns**. It represents synthetic data based on real **transactions** alerted by **Intesa Sanpaolo** Transaction Monitoring Engine.

- The data format is CSV.
- The data represents a sample of transactions that occurred during the year 2022.
- **Sensible information has been hashed** with the SHA-512 algorithm. The **first column**, defined by the description “chiave_univoca”, is the transaction **ID**. The **second column** represents the transaction event **timestamp**. The third column represents the economic amount of the transaction. Columns from fourth to sixth represent the year, month, and day.
- The **seventh column** represents the **target** for the supervised learning task. The target is a binary variable, where a fraudulent transaction is encoded with 1 and a genuine transaction is encoded with 0.
- **The dataset is highly imbalanced i.e., ratio around 2%.**
- The remaining columns represent hashed categorical features of the transactions. The data can contain missing values, that should be treated in the pre-processing phase.



Fraud Detection – description

Timestamp	1	10	12	20	40	100	130
T-type	AA	CC	BB	BB	BB	AA	BB
Beneficiario	1235	NaN	1235	NaN	NaN	1111	1235
Importo	10,00	100,54	1,00	30,82	455,40	1000,00	15,00
...							
Fraud	0	0	0	0	0	1	0



Fraud Detection – Challenges/Innovation

- Performance evaluation proposed:
 - Most important metric to catch all frauds: **Recall**.
 - Particular attention goes **Precision**. The **false positive** (transactions that are not frauds) that can lead to inconvenience related to the interruption of customer operations.
- From the point of view of a standard approach, the challenge here is to **leverage on the history of the client's transaction** (in the form of a Time Series) without explicitly perform a feature engineering.
- The target variable is **highly unbalanced**. Undersampling/oversampling?
- Possibility of adapting to the constantly and rapidly **evolving phenomenon**. Customization of alerts according to the habits of the individual customer.
- **Expainability** on TS?



Fraud Detection – business impact

- **decreased monetary loss** for the company
- **better customer experience**, avoids inconvenience related to the interruption of customer operations, and thus increasing customer satisfaction and preventing reputational damage
- Intercepting fraud for European supervisory authorities, reinforcing their capabilities in the **fight against financial crimes**.



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Banking time series



Corporate Credit Risk - description






[Optional]

Problem definition:

Predict the *status* of a «**performing**» client after 6 months from the reference time of observation (i.e. predict if the client will be able to repay his loans/debts).



Corporate Credit Risk - description

 ANAGRAFICA_PERIMETRO_ESTESO.csv	729.610 KB
 BON_ESTESO.csv	1.136.281 KB
 MOCRED_PERIMETRO_ESTESO.csv	77.043 KB
 RISK_ANAG_PG_SEGREG.csv	115.889 KB
 RRC_SNDG_SCONFINI_PERIMETRO_ESTESO.c...	71.014 KB



Corporate Credit Risk - description

- **ANAG** = This is basically a table version of a customer registry. This will enable us to do an initial analysis on the customers based on their origin (ITALY/FOREIGN), industrial activity (Trade, construction, hotel etc), province and many more. Hence we could understand the portfolio of customers sector.
- **BONIFICI** = The details of bank transactions extracted which is an essential data to build the network of customer and suppliers
- **SCONFINI** = This contains the information about the overdrafts of companies. The companies withdraw money from the bank up to certain limit to fund their short term cash requirements.
- **Stato Amministrativo (target)** = This contains the data of customers who obtained a classification code by the bank based on their financial distress. The COD_CPSR of raw data is transformed to a code of 1 or 0 in such a way that the real defaulted (liquidated) companies, the customers who haven't paid the bank for 90 to 180 days and the customers who are in the risk of being default (as identified by the bank) are given a code 1 while the other Codes a zero.

Corporate Credit Risk – challenges/innovation

Challenges/innovation:

- **Default rate** of «performing» clients is low, ranging from **1 to 3%**.
- **Feature distribution can vary significantly** in test set if compared to training set due to unexpected major events (e.g. covid crisis, bank merge and acquisitions, ...)
- The nature of clients in **perimeter is heterogeneous** (firms of different size, industries, ...)
- The output of the model **must be explainable**



Corporate Credit Risk – challenges/innovation

Current approach: classification model (XGBoost), binary target, tabular data

Possible alternative approaches :

Use a generative model to assess **covariate shift**

Clustering on clients (graph on clients or in a latent space)

and subsample on cluster

Explainable Time series



Corporate Credit Risk – business impact

Increase in Recall can keep the FPR of defaults stable, in order to **concentrate resources on really risky cases** avoid loss.



Data Quality - description

The goal of this project is to identify anomalies on the balances of the current accounts of some bank customers in order to allow the remediation activity.

The values to be checked are:

- Sum of positive daily balance
- Sum of negative daily balance
- Daily account balance of individual contracts and difference from the previous day

Data Quality - Description

DATASET

We deal with two main different types of timeseries:

1. **Single** time series - sum of daily balances
2. **Multiple** timeseries (~ 15 millions) - related to the daily balance of several clients' current accounts

Lenght Single TS

~ 2 years

Lenght Multi TS

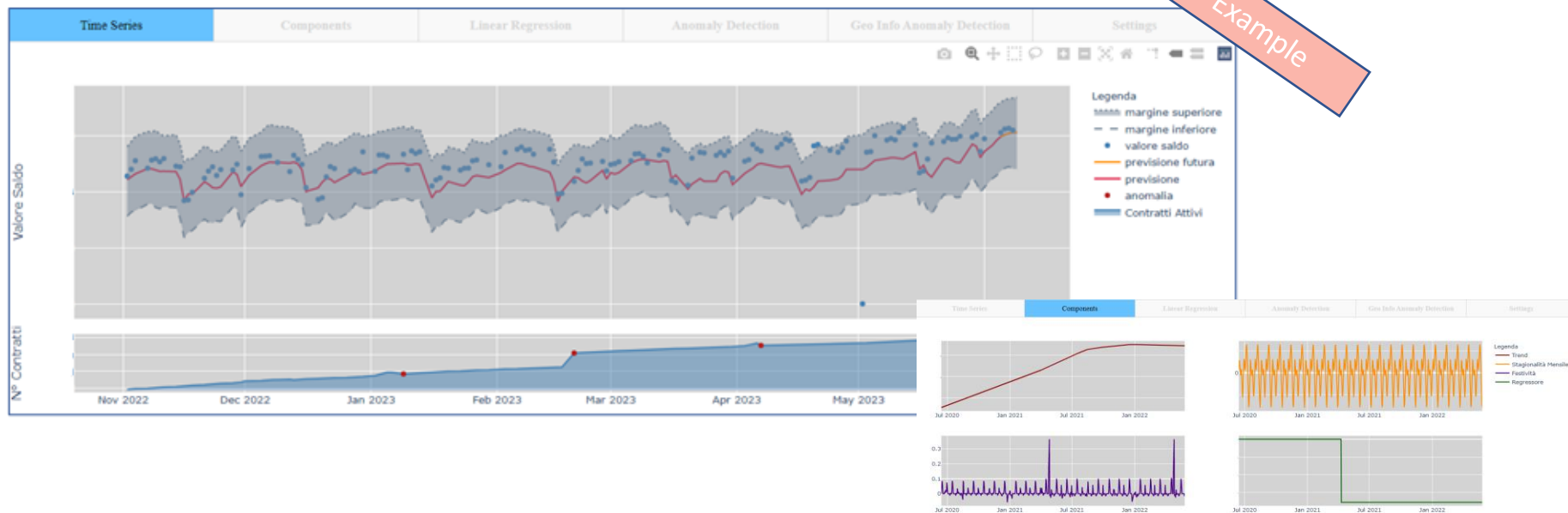
may vary from few days to years

The need is to study its evolution over time and to highlight possible anomalies in data recording and processing.

- The main cause of anomalies are errors during data transfer processes between systems or caused by customers' illicit behaviours.

Data Quality - Description

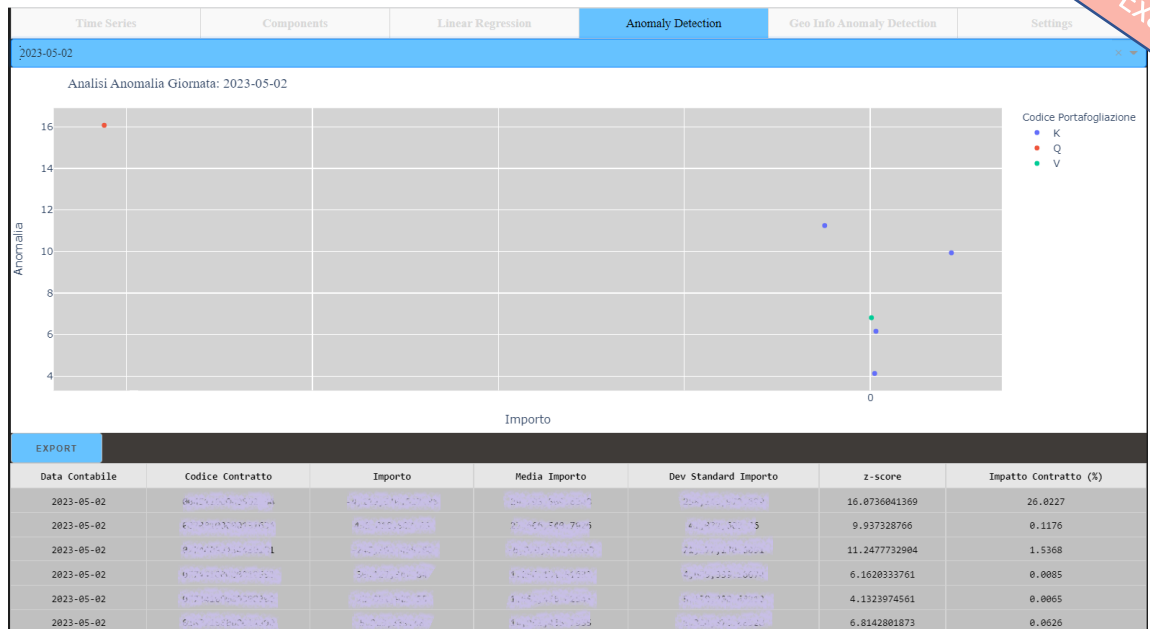
Anomaly detection on single time series – negative balance



Data Quality - Description

Anomaly detection on multi time series

Example





Data Quality - Challenges/Innovation

Actual approach: Prophet, z-score

Challenges: really unbalanced dataset or not target available.

Proposed approach: Forecast on time series with ML/DL approach, Anomaly Detection on Time Series, explainability to understand the root cause of the anomalies.



Data Quality – Expected benefits

Expected Benefits: increase precision and recall to reduce the number of false positive reported.
Report customers' illicit behaviours.
Report systematic errors in database.



Questions?



Data Quality



Fraud Detection



CCR



Know Nothing