



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

ISP – Time Series for the banking sector

*ISP: Paola Barone, **Riccardo Crupi**, Rachele Desiante, Alessandro Sabatino
INAF Capodimonte
Tor Vergata*

Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

Scientific Rationale

Anomaly detection plays a crucial role in ensuring data quality for time series datasets. The objective of anomaly detection on time series data is to identify unusual patterns or outliers that deviate from the expected behavior. By detecting anomalies, we can ensure data quality, identify data integrity issues, and mitigate potential risks.

By identifying and managing anomalies, we can improve the quality of data in our systems by helping to identify:

- Data measurements errors
- Data entry mistakes
- Data corruption
- Network disruption
- Intercept fraudulent transactions.

Technical Objectives, Methodologies and Solutions

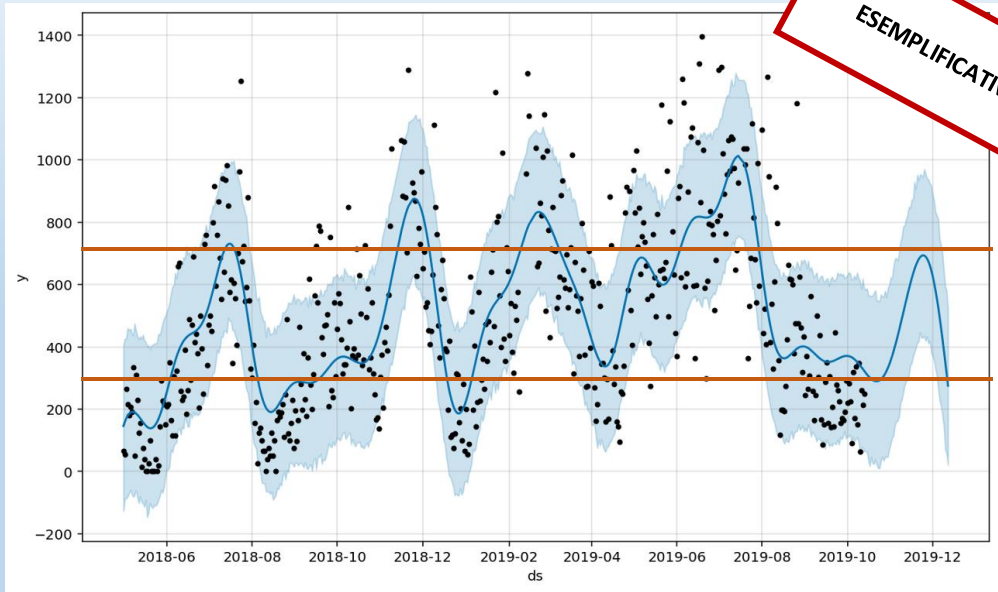
Traditional Data Quality

- **Deteministic control** that verifies the percentage deviation between day T and day T-1.
- **The threshold** for comparison is set by the **Data Owner** based on their own experience.



AI Data Quality basata sull'A.I.

- **Anomaly detection** with AI libraries (Fbprophet, Pyod) reporting warning in case of significant deviations
- **Dynamic threshold** shall be corrected over time on the basis of variation in distribution, taking into account seasonality and other regressions



Legenda

- Soglia statica definita dal DOW
- Soglia dinamica dell' algoritmo

Technical Objectives, Methodologies and Solutions

The time series analyzed in this context are:

1. credit current account balances (positive balances)
2. debit current account balances (negative balances)
3. number of insurance policies (11 different products)
4. amounts of insurance policies (11 different products)

CASE STUDIES

- **Single** time series: sum of daily balances
- **Multi Time series:** insurance policies – a time series for insurance policies typologies

Lenght Single TS --> 2 Years

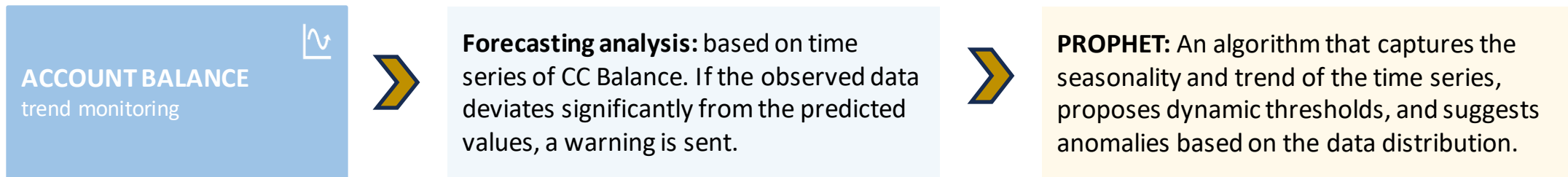
Lenght Multi TS --> may vary from few days to years

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.

Technical Objectives, Methodologies and Solutions

Credit and Debit current account balances



Dataset

COD_ANNO	COD_MESE	IMP_SALDO_CTB_01	IMP_SALDO_CTB_02	IMP_SALDO_O_CTB_	IMP_SALDO_CTB_30	IMP_SALDO_CTB_31	COD_CONTRATTO
2023	3	0	0	0..		0	0X1
2023	2	0	0	0..		0	0X2
2023	3	0	0	0..		0	0X3
2023	2	0	0	0..		0	0X4
2023	3	-1083.22	-1135.22..			0	0X5
2023	2	-236.98	-236.98..			0	0X6
2023	1	-1.89	-285.97..			0	0X7
2023	3	-10.23	-10.23..		-10.23	-10.23	0X8
2023	1	0	0	0..		0	0X9
2023	1	-56.4	-56.4..		-70.9	-70.9	0X10
2023	3	0	0	0..	-6.01	-6.01	0X11
2023	1	0	0	0..		0	0X12
2023	3	-198.97	-221.56..			0	0X13

Technical Objectives, Methodologies and Solutions

Credit and Debit current account balances



The trend component captures long-term variations in the historical series.



The holiday component captures the effects of specific holiday dates.



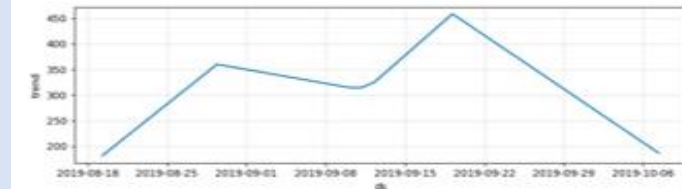
The seasonal component captures short-term variations that repeat regularly.



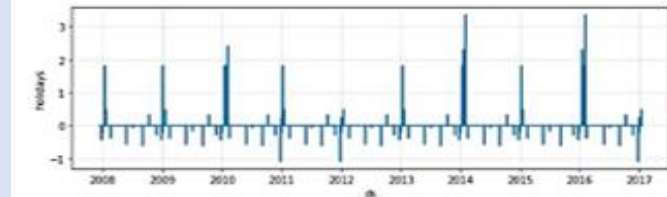
Regressors can provide additional information to the model to improve the accuracy of the forecasts.

Plot components

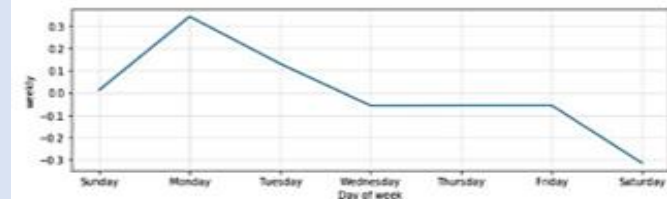
Trend
 $g(t)$



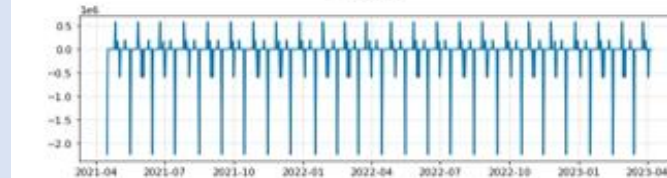
Events
 $h(t)$



Seasonality
 $s(t)$

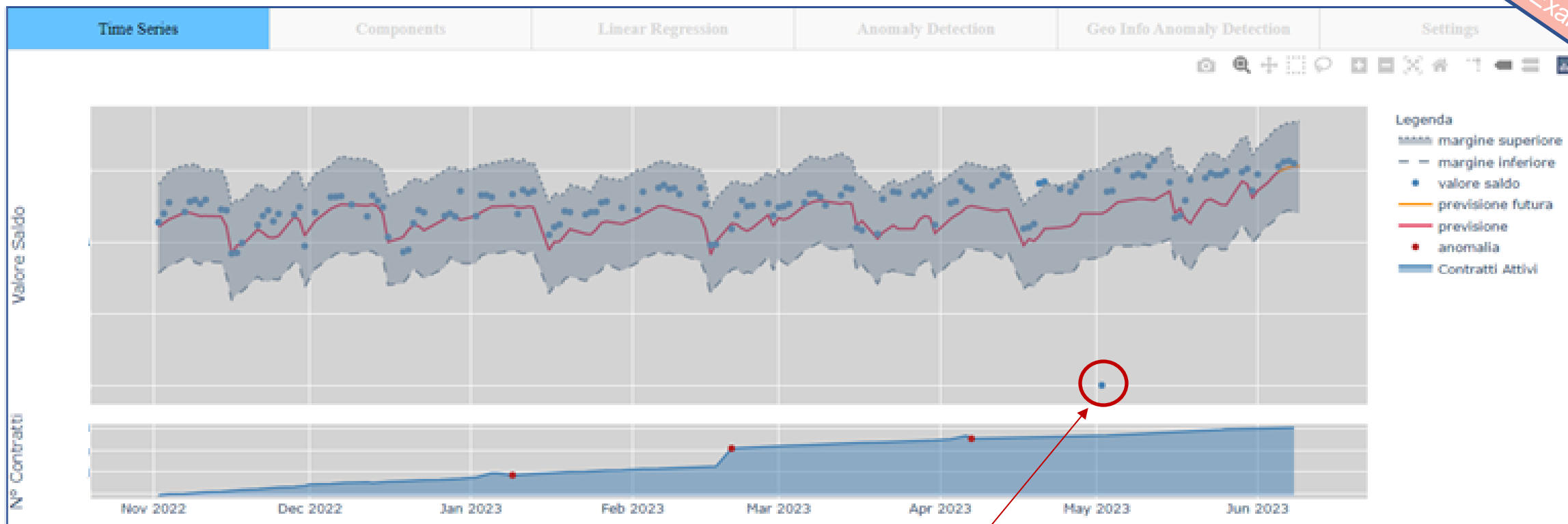


Regressors
 $x(t)$



Technical Objectives, Methodologies and Solutions

Credit and Debit current account balances



Example

anomaly

Technical Objectives, Methodologies and Solutions

Insurance policies

INSURANCE POLICIE
trend monitoring



Anomaly detection on time series of number of insurance policies and daily amount. If what is observed deviates significantly from what is expected, a warning is sent.



PyOD: It is a comprehensive toolbox that offers a wide range of algorithms to detect anomalies in data. SUOD (Scalable Unsupervised Outlier Detection) is a module within PyOD that focuses on scalable unsupervised outlier detection algorithms.

Dataset

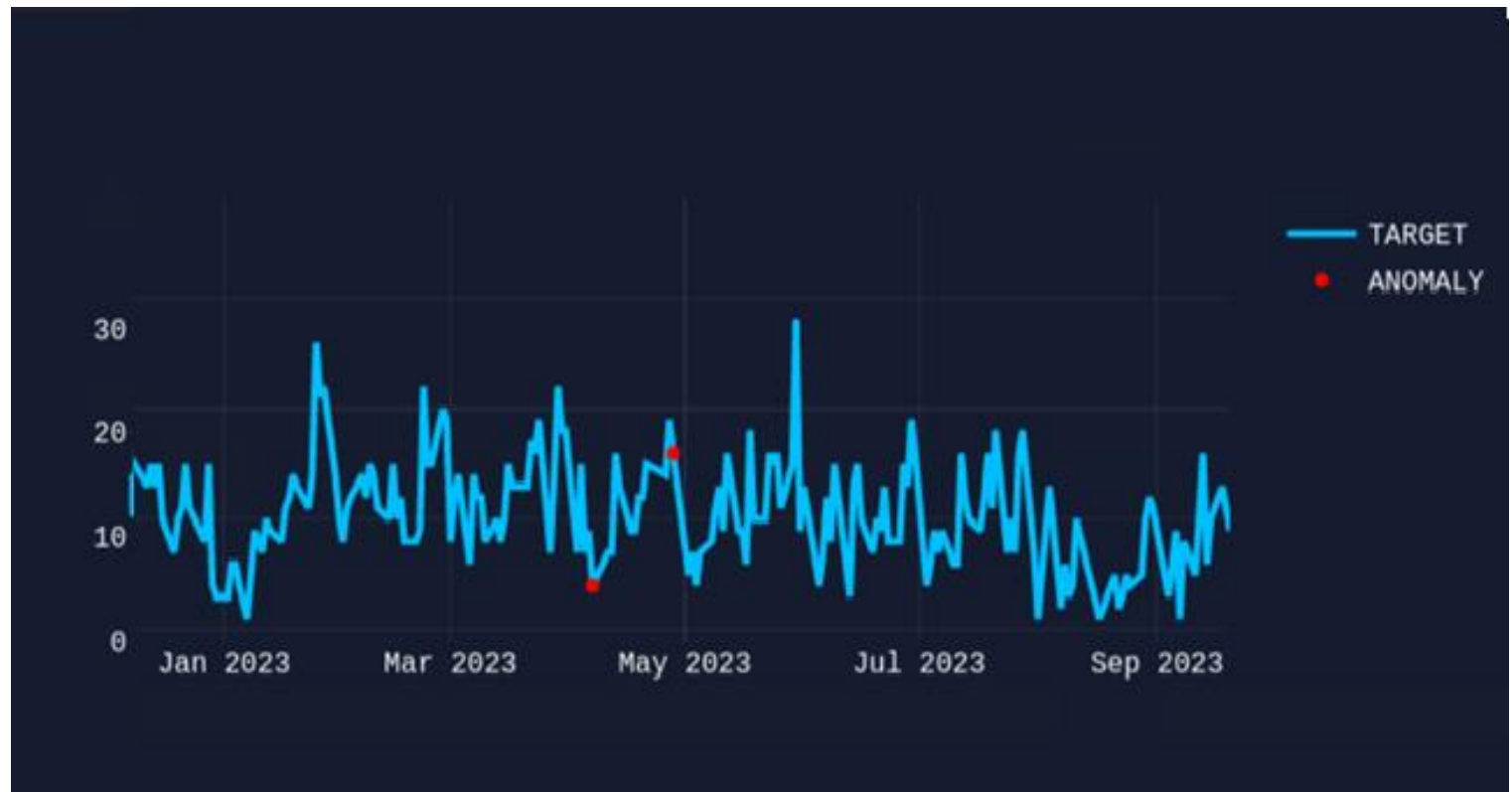
DES_PRODOTTO	DAT_DISSUEPOL	COUNT_DISTINCT_of_NOM_CNUMP	COD_CCODESTATUS	DAT_DEFF	NUM_NNETAMOUNT	DES_PRODOTTO
COLLETTIVA INFORTUNI	01-feb-23	8	4	01JAN2023	19.73	TUTELA BUSINESS - UFFICI E STUDI
COLLETTIVA INFORTUNI	01-mar-23	8	1	01AUG2023	43.76	POLIZZA XME PROTEZIONE
COLLETTIVA MALATTIE GRAVI	01-feb-23	10	4	01JAN2023	2589.37	TUTELA BUSINESS AGRICOLTURA
CYBER PROTECTION BUSINESS	09-feb-23	24	1	01AUG2023	625.04	TUTELA BUSINESS COMMERCIO
CYBER PROTECTION BUSINESS	09-mar-23	16	1	01AUG2023	41.92	POLIZZA XME PROTEZIONE
PIANO SANITARIO RIMBORSO			1	01AUG2023	28.54	POLIZZA XME PROTEZIONE
SPESE MEDICHE	20-feb-23	4	1	01AUG2023	80.65	POLIZZA XME PROTEZIONE
PIANO SANITARIO RIMBORSO			4	01JAN2023	203.8	TUTELA BUSINESS COMMERCIO
SPESE MEDICHE	20-mar-23	4	1	01AUG2023	30.66	POLIZZA XME PROTEZIONE
PIANO SANITARIO RIMBORSO			1	01AUG2023	40.71	POLIZZA XME PROTEZIONE
SPESE MEDICHE	21-apr-23	11	1	01AUG2023	807.44	TUTELA BUSINESS COMMERCIO
PIANO SANITARIO RIMBORSO			1	01AUG2023	50.38	POLIZZA XME PROTEZIONE
SPESE MEDICHE	21-feb-23	5	1	01AUG2023	36.27	POLIZZA XME PROTEZIONE
			4	01JAN2023	1465.33	TUTELA BUSINESS COMMERCIO

Technical Objectives, Methodologies and Solutions

Insurance policies

```
detector_list = [KNN(), LOF(), IForest(random_state=42)]  
if_model = SUOD(base_estimators=detector_list, n_jobs=-1, combination='majority score', verbose=False)
```

	ds	y	score	prediction
299	2023-04-07 00:00:00	4	0.023538	1
300	2023-04-11 00:00:00	7	-0.069881	0
301	2023-04-12 00:00:00	7	-0.061435	0
302	2023-04-13 00:00:00	16	-0.062165	0
303	2023-04-14 00:00:00	13	-0.080407	0
304	2023-04-17 00:00:00	9	-0.060340	0
305	2023-04-18 00:00:00	9	-0.068054	0
306	2023-04-19 00:00:00	12	-0.059391	0
307	2023-04-20 00:00:00	12	-0.066804	0
308	2023-04-21 00:00:00	15	-0.067952	0
309	2023-04-26 00:00:00	14	-0.042000	0
310	2023-04-27 00:00:00	19	-0.063753	0
311	2023-04-28 00:00:00	16	0.008082	1



Timescale, Milestones and KPIs

The work package in the Spoke 3 corresponds to WP3. The involved partners are listed in section 3.2.

Deliverable:

- M6 Data delivery – completed 100%
- M7 Preliminary results on data exploration and how to format the dataset, preprocessing the data (e.g. feature engineering, feature selection) - completed 0%
- M8 first prototype application with ML to one use case – completed 0%
- M9 Prototype with explainability - completed 0%
- M10 Python library that works on the banking dataset and environment – completed 0%

Timescale, Milestones and KPIs

The work package in the Spoke 3 corresponds to WP3. The involved partners are listed in section 3.2.

Deliverable:

M7 Preliminary results on data exploration and how to format the dataset, preprocessing the data (e.g. feature engineering, feature selection)

In the initial stages of the project, the team undertook several critical actions to prepare the data effectively. They began by conducting preliminary data exploration, delving into the dataset to gain insights and identify patterns. Simultaneously, they worked on formatting the dataset, ensuring it was organized and structured in a way that would facilitate further analysis. Additionally, they engaged in preprocessing the data, a multifaceted task that encompassed activities such as feature engineering and feature selection. These efforts were aimed at refining and enhancing the dataset, optimizing it for subsequent machine learning and analytical tasks.

Need: kick-off meeting with counterparts for defining the technical objective and sharing the data.

KPI: 2x100. Link experiments and methods for data exploration and/or preparation in a repository. Measure: 100% present, 0% not delivered. 1 h presentation of the discovery and the data preparation (ppt, technical report?). Measure: 100% done, 0% not done.

Timescale, Milestones and KPIs

The work package in the Spoke 3 corresponds to WP3. The involved partners are listed in section 3.2.

Deliverable:

M8 first prototype application with ML to one use case. Prepare paper to submit

The team commenced the project by developing the M8 prototype application, incorporating machine learning techniques to address a specific use case. They constructed a solution informed by the extensive expertise and domain knowledge of their counterparts, ensuring a highly specialized and effective approach.

Project written in Python in a Git environment, commented, modular (developed in class) such that the pipeline could consist in:

1. `DF = Data_preparation(data, ...)`
2. `M = Train_model(DF, develop=False, ...)`
3. `M.predict(DF)`

KPI: 2x100. Working code in the Git repository: done 100% otherwise 0%. Tutorial 1h presentation 100%, 0% otherwise.

Timescale, Milestones and KPIs

The work package in the Spoke 3 corresponds to WP3. The involved partners are listed in section 3.2.

Deliverable:

M9 Prototype with explainability. Submit paper to be reviewed represents a significant advancement, as it not only showcases the prototype's functionality but also emphasizes the crucial aspect of interpretability, enabling a deeper understanding of the machine learning models' decision-making processes.

Add in pipeline:

1. M.explain(sample)

KPI: 2x100. feature implemented (or present by design) 100%, not implemented 0%. Submission of the paper to a journal/high-level-conference 100%, 0% otherwise.

Timescale, Milestones and KPIs

The work package in the Spoke 3 corresponds to WP3. The involved partners are listed in section 3.2.

Deliverable:

M10 Python library that works on the banking dataset and environment.

The M10 Python library has been designed to operate seamlessly within the banking dataset and environment, facilitating robust and tailored data analysis and processing.

pip install coolestProject

`CoolestProject.prepare_dataset()`

`CoolestProject.train()`

`CoolestProject.predict()`

`CoolestProject.explain()`

KPI: 4x100. if NAME SURNAME can install it on his PC and use it on a toy dataset 100%, 0% otherwise. Quick start on GitHub page 100%, 0% otherwise. Documented page on GitHub (readthedocs) 100%, 0% otherwise. Paper published in a journal/high-level-conference 100%, 0% otherwise.

Timescale, Milestones and KPIs

Rome Tor Vergata: the Rome "Tor Vergata" unit, composed of Alessandro Casini, Alessio Farcomeni (Fraud detection manager), Marco Patacca and Tommaso Proietti (data quality manager), will contribute to the project by developing methodologies for identifying anomalies, fraud and structural changes in supervised and unsupervised environments, for panel and time series data, based on:

- robust univariate and multivariate filtering techniques for state-space models (e.g., Kalman Filter)
- statistical learning methods for anomaly detection based on trimming and snipping procedures

References:

Marczak, M., Proietti, T. and Grassi S. (2017) A Data-Cleaning Augmented Kalman Filter for Robust Estimation of State Space Models
Article reference: *Econometrics and Statistics*.

Marczak, M. and Proietti, T. (2016). Outlier Detection in Structural Time Series Models: the Indicator Saturation Approach. *International Journal of Forecasting*, 32, 180–202.

Proietti T. and Pedregal D. (2023). Seasonality in High Frequency Time Series. *Econometrics and Statistics*, in Press,
<https://doi.org/10.1016/j.ecosta.2022.02.001>.

Rome Tor Vergata - Gordaliza, A. (1991) Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, 64, 162–180

Rome Tor Vergata - Farcomeni, A. (2014) Snipping for robust k-means clustering under component-wise contamination, *Statistics and Computing*, 24, 909-917

Timescale, Milestones and KPIs

The INAF Capodimonte unit, which is part of the Capodimonte Astronomical Observatory and consists of Stefano Cavuoti and Giuseppe Riccio, will contribute to the project by leveraging their expertise in Machine Learning [5][6]. Their primary focus will be on developing and applying algorithms to detect anomalies in time series data. Specifically, their approach could involve manually extracting features from the time series data and utilizing tools such as [1][2][3]. Subsequently, an anomaly detector (for example, one provided in [4]) can assign anomaly scores to individual clients, groups of clients, or specific time periods (such as weeks).

References:

[1] <https://tsfresh.readthedocs.io/en/latest/>

Christ, Maximilian, et al. "Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package)." *Neurocomputing* 307 (2018): 72-77.

[2] <https://tsfel.readthedocs.io/en/latest/>

Barandas, Marília, et al. "TSFEL: Time series feature extraction library." *SoftwareX* 11 (2020): 100456.

[3] <https://www.featuretools.com/>

[4] <https://github.com/yzhao062/pyod>









Zhao, Yue, Zain Nasrullah, and Zheng Li. "Pyod: A python toolbox for scalable outlier detection." *arXiv preprint arXiv:1901.01588* (2019)

[5] D'Isanto, Antonio, et al. "An analysis of feature relevance in the classification of astronomical transients with machine learning methods." *Monthly Notices of the Royal Astronomical Society* 457.3 (2016): 3119-3132.

[6] De Cicco, D., et al. "A random forest-based selection of optically variable AGN in the VST-COSMOS field." *Astronomy & Astrophysics* 645 (2021): A103.

Accomplished Work, Results

- First shot of a dataset for Data Quality

	dtf_cc_aver_2021.csv.gz
	dtf_cc_aver_2022.csv.gz
	dtf_cc_aver_2023.csv.gz
	dtf_cc_dare_2021.csv.gz
	dtf_cc_dare_2022.csv.gz
	dtf_cc_dare_2023.csv.gz
	ts_importi_polizze_bfd.csv.gz
	ts_numeriche_polizze.csv

Next Steps and Expected Results (by next checkpoint: April 2024)

Data Quality:

- **Expected benefits: increase precision and recall by 10%. This will help the ISP Database maintainers to anticipate and focus only on the most critical data flux.**

Fraud Detection:

- **On a small scale, we aim to achieve a more precise automated detection of fraud, leading to a better customer experience, and a decreased monetary loss for the company.**
- **On a larger scale, by better intercepting fraud in our financial transactions, we could increase our commitment to the National and European supervisory authorities, reinforcing their capabilities in the fight against financial crimes.**
- **For Fraud Detection reduce the False Positive by 10% of the standard process (XGBoost of features extracted from the TS).**
- **Reducing False Positive avoids inconvenience related to the interruption of customer operations, and thus increasing customer satisfaction and preventing reputational damage.**
- **Moreover, the environmental footprint of quantum computation is less demanding compared to HPC data centers that are generally used to address this kind of problem.**