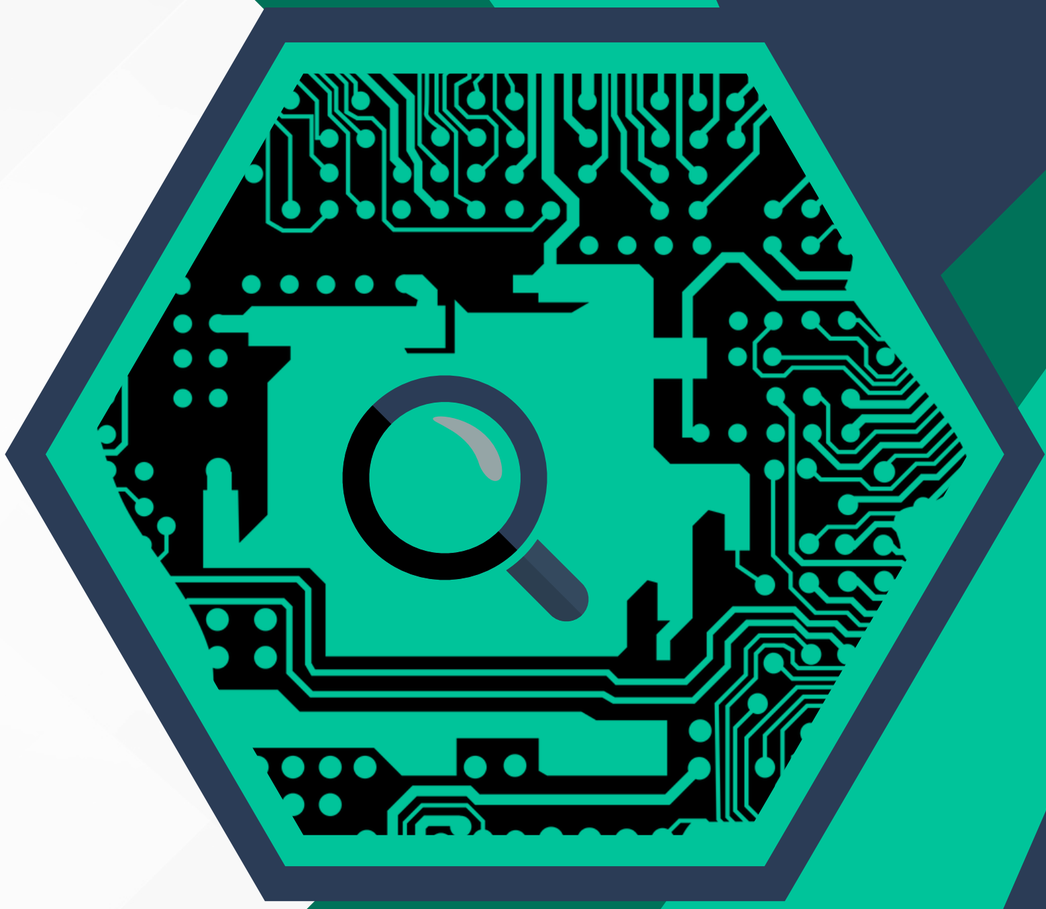
ANOMALY DETECTION



GROUP 8: ***Keras-trophic***

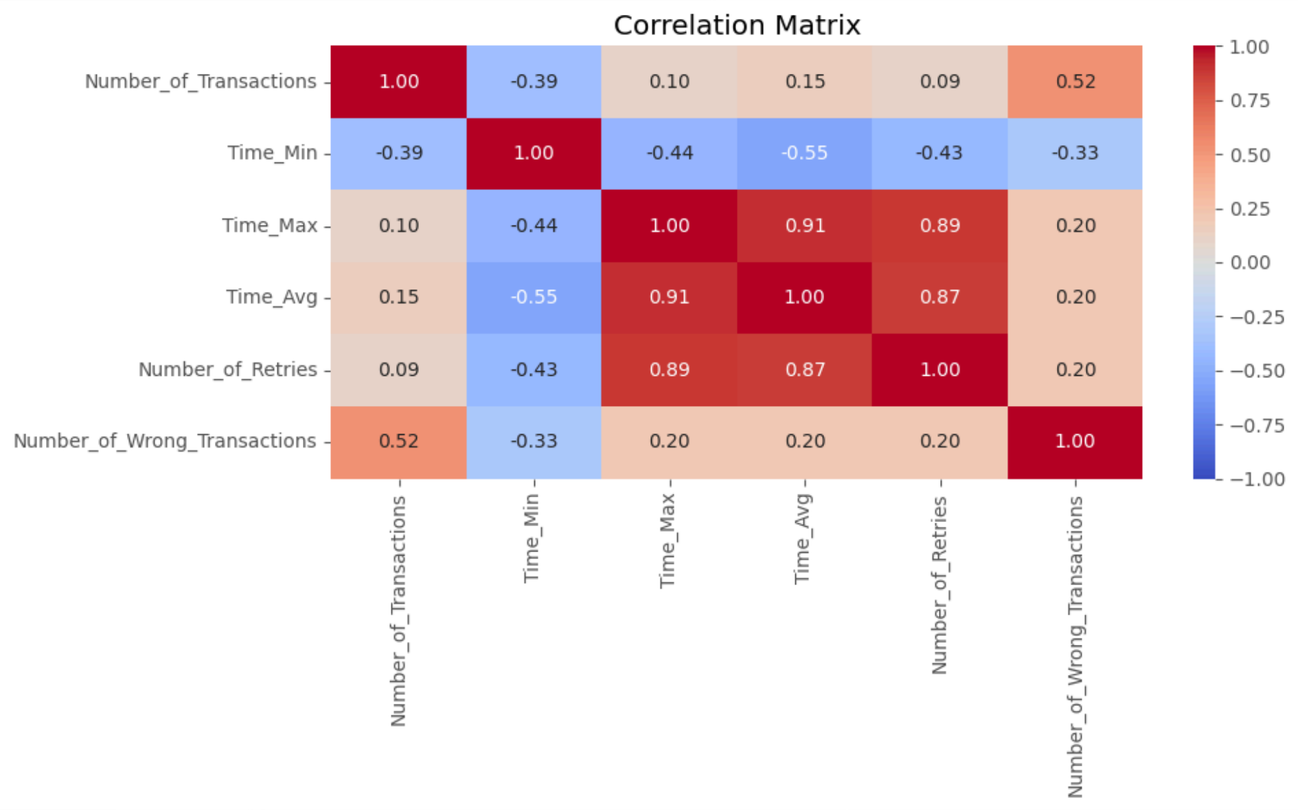
Flavia Fuscaldi
Giacomo Guerra
Caterina Perrotti

REPLY

PROJECT GOAL

- Develop a time series model that combines forecasting and anomaly detection to identify unusual patterns or outliers.
- Enable proactive issue detection to improve operational efficiency, reliability, and system security.

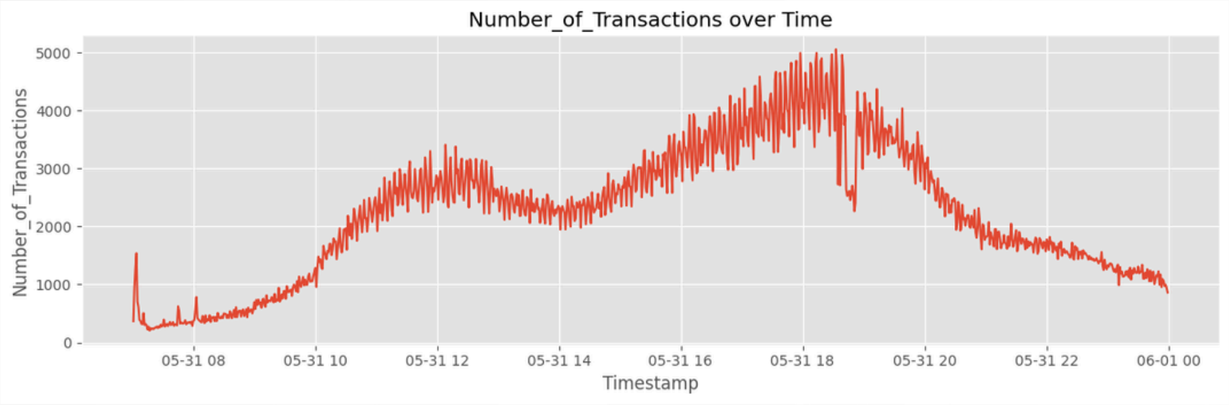
EDA

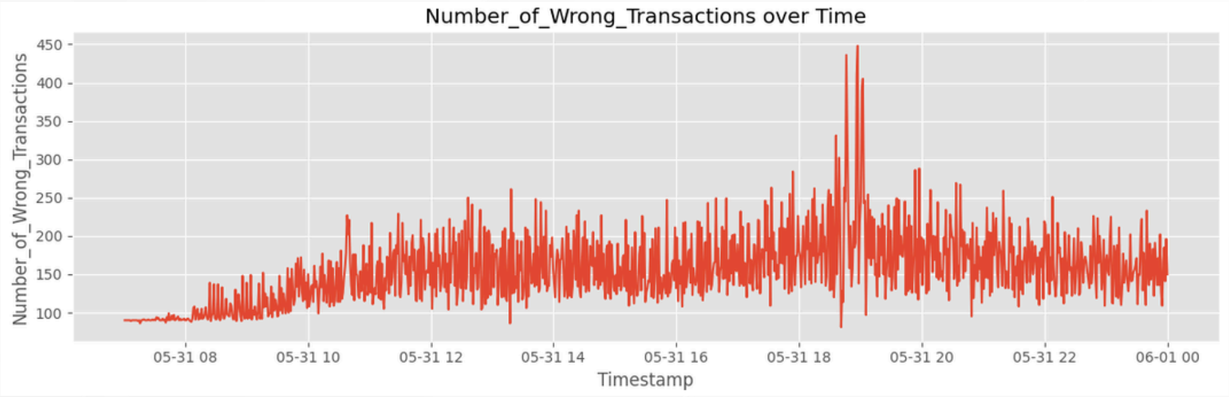


Feature Cleaning

- Understood column meaning through initial data inspection
- Dropped columns not useful for our analysis:
  - Intervallo di acquisizione
  - Code_Id
- Due to high correlation and low interpretability, excluded:
  - Time_Max
  - Time_Min
- Kept Time_Avg as the only representative time feature

EDA





By plotting the distribution of features over time

Anomaly in Input Data around 7:00 PM

Including it in training could lead the model to:

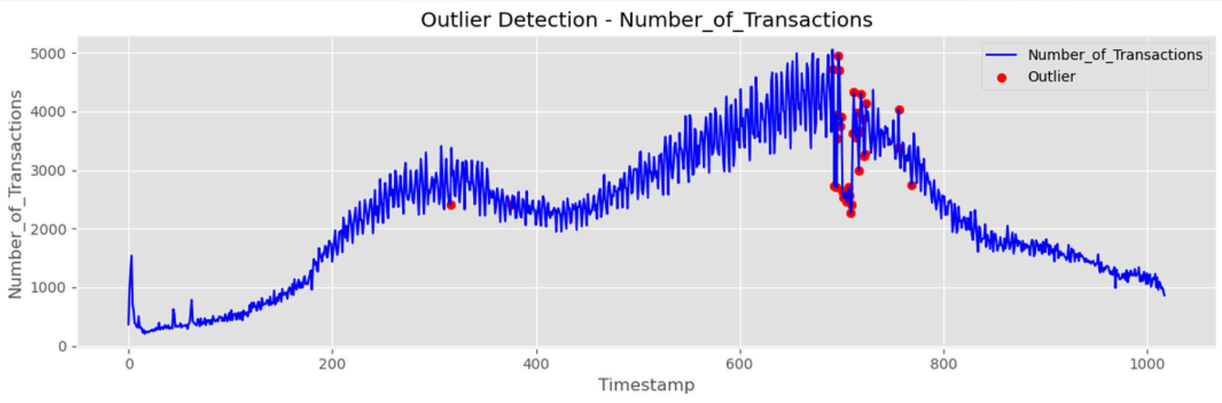
Interpret the anomaly as part of the normal pattern

Fail to detect similar anomalies in future data

Key Design Choice:
Filter or handle anomalies before training to preserve model reliability.

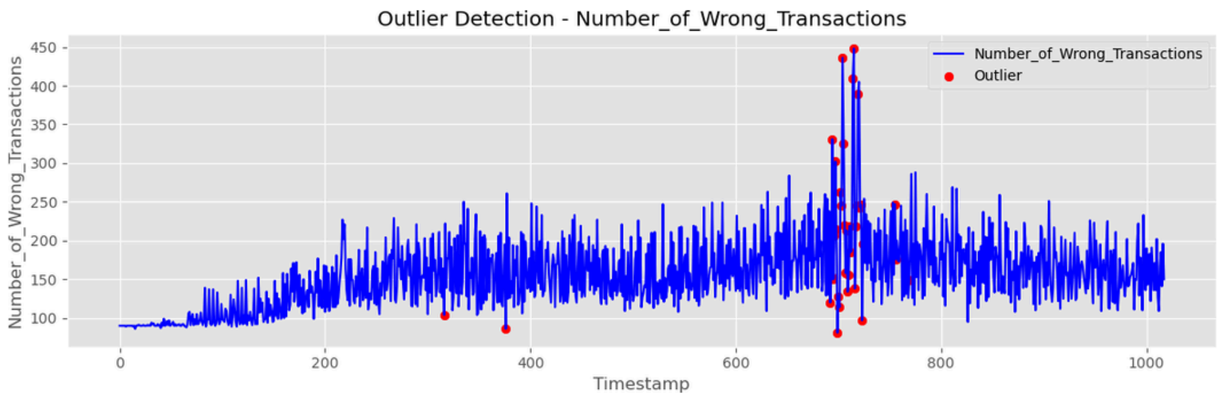
OUTLIERS REMOVAL

We decided to use
ISOLATION FOREST
to flag the outliers



Next Steps: Handling Outliers

- we created a copy of the original day's data
- we first replaced them with NaN values



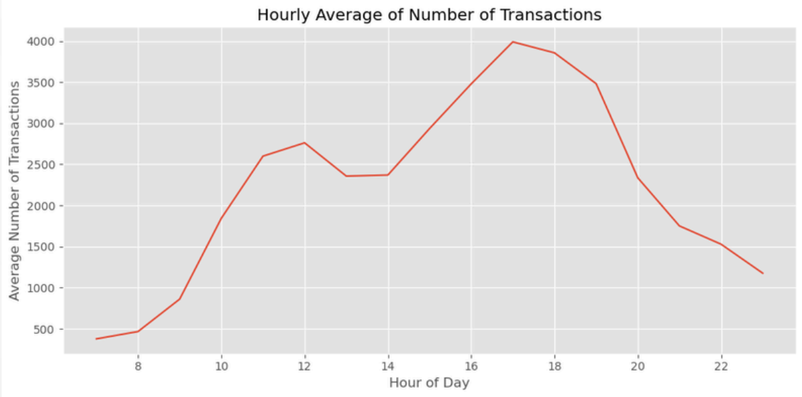
OUTLIER IMPUTATION

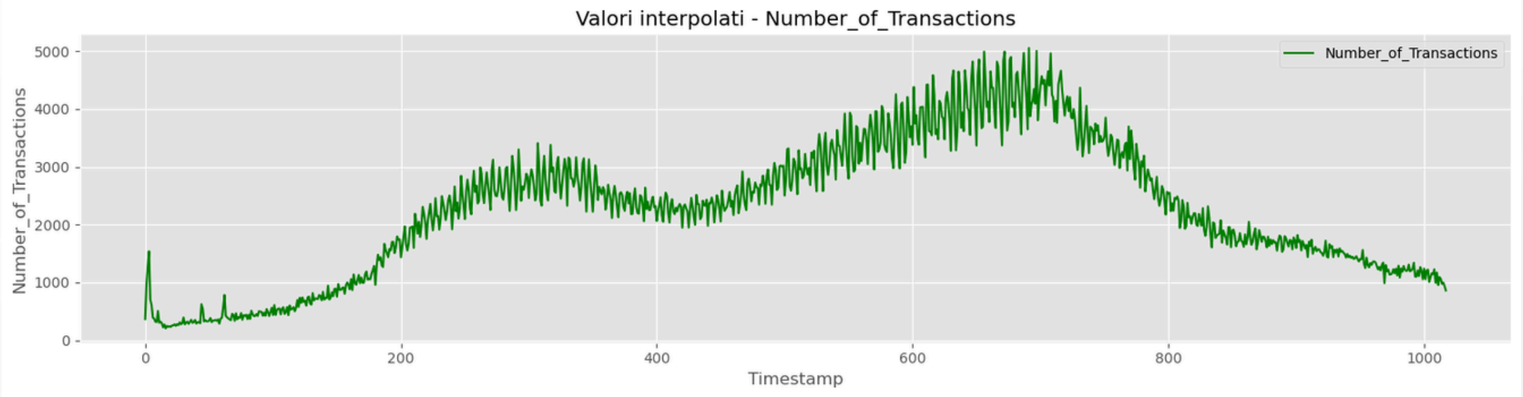
Initial attempts to fill values using:

- Linear & spline interpolation
- Predictive models

Did not yield reliable results

So missing values were imputed using an **hourly rolling mean** based on the Timestamp column, with **added noise** scaled by the rolling standard deviation, and clipped to predefined variable-specific limits to ensure realistic value ranges.



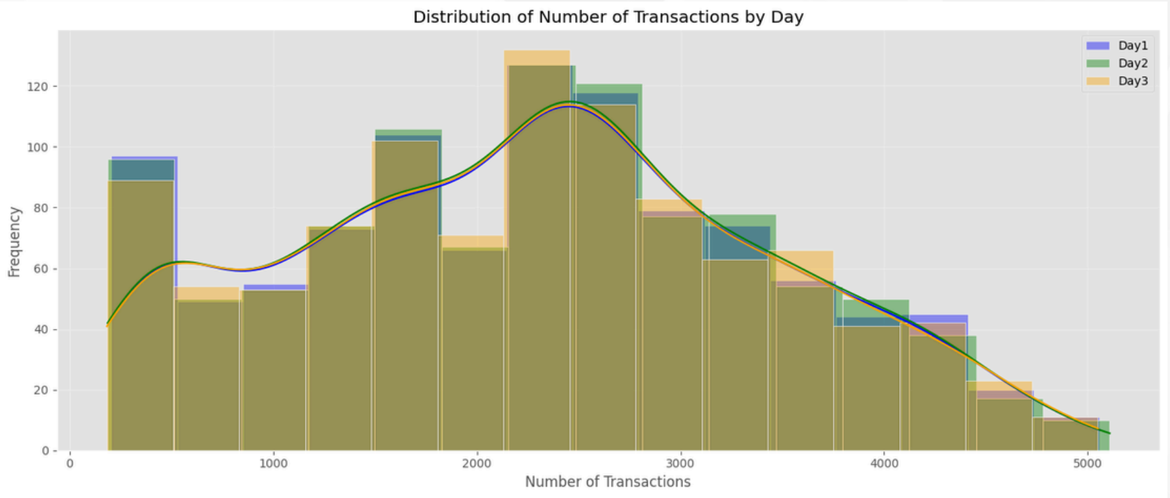


DATA AUGMENTATION

Since the dataset provided was about one day we needed to **increase** the dimension

DIFFERENT APPROACHES:

- More advanced ones failed (TimeGAN, VAE)
- At the end we opted for the inection of gaussian noise.



we defined a function able to recreate one day introducing a random noise that could go from.
0.1 to 0.5
and recalled the function 2 times obtaining a total of 3 days usable.

FEATURE ENGINEERING

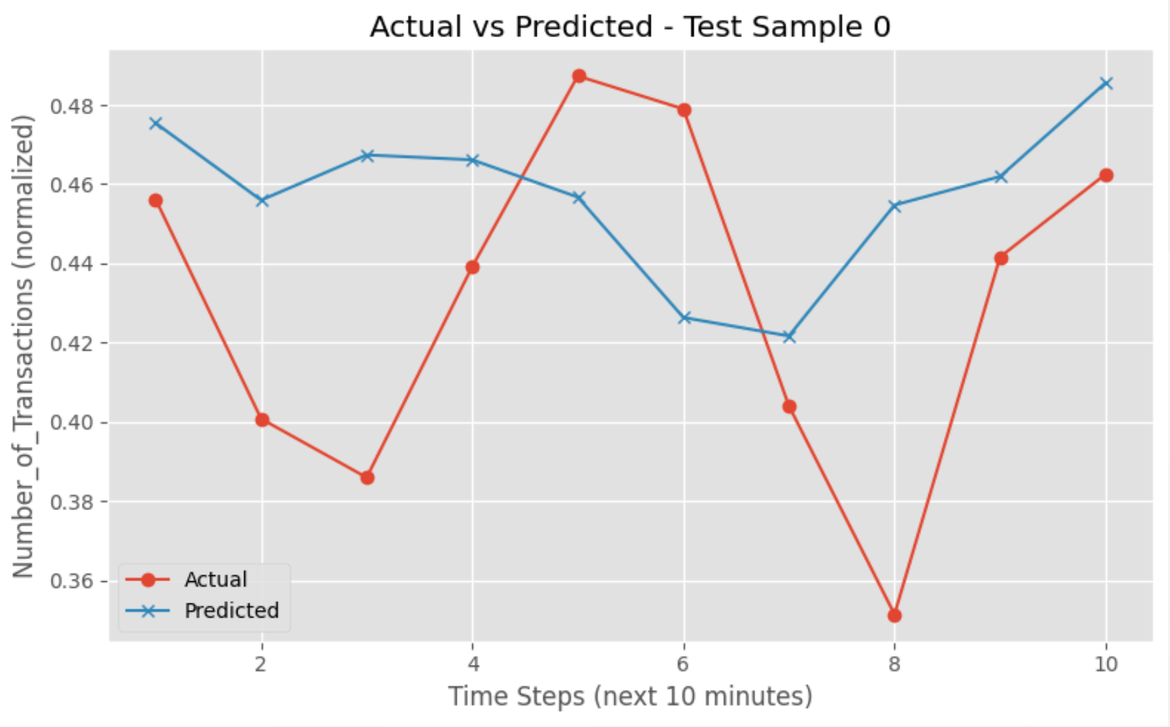
Before concatenating the three days of data, we engineered new daily features.

- **Lagged values** to provide historical context
- **Rolling statistics** (e.g., mean, standard deviation) for smoothing and trend detection - shifted to avoid leakage
- **Step-wise changes** to highlight abrupt variations
- **Ratios** related to errors and retries for performance insights
- **Time of day** to incorporate cyclical patterns.

We defined a function to generate daily meaningful features that enhance the model's ability to capture **patterns**, **trends**, and **temporal behaviors.**
This function adds:
After incorporating these features, we combined the three days into a single dataset for downstream analysis.

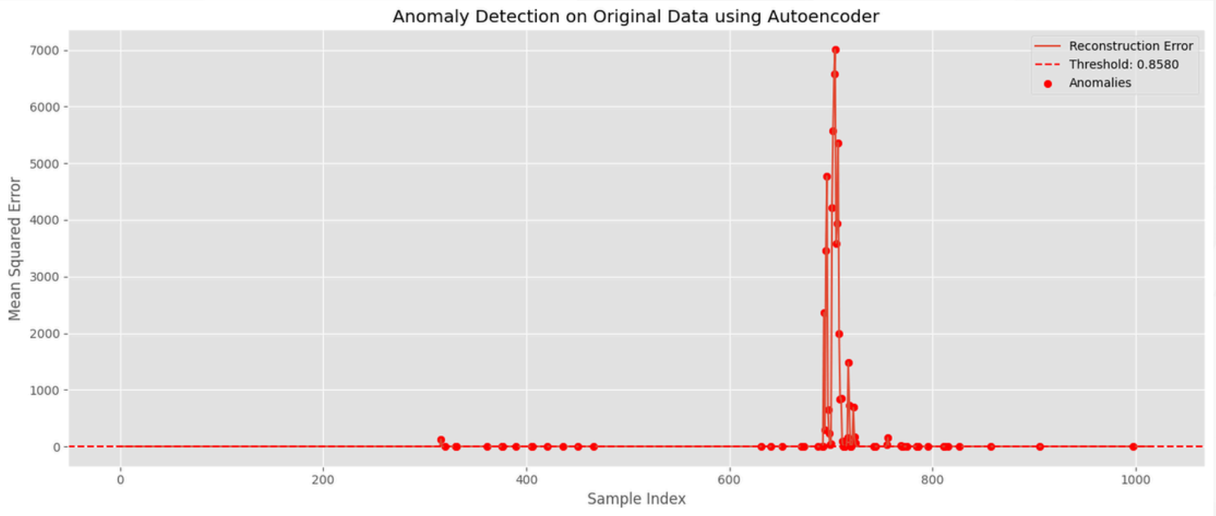
MODEL: UNIVARIATE LSTM

- We used a **univariate LSTM** model able to learn temporal dependencies in the transaction series and forecast future values based solely on its past behavior.
- We built the model considering an input window of 30 minutes and an output window of 10.
  - Target variable: Number of transactions
- The model achieved strong performance, with:
  - RMSE: 268.98
  - R^2: 0.93
- This indicates that the LSTM captured most of the underlying patterns in the data, producing accurate forecasts and explaining 93% of the variance in the target variable



ANOMALY DETECTION

- Coming back to the original version of the dataset we used that to check the abilities of the model to detect real anomalies
- Threshold fixed based on the 99th percentile of error distribution
- When comparing our forecasting-based approach to Isolation Forest, the results were largely similar.
- However, some differences arose because our forecasting model is univariate, while Isolation Forest operates as a multivariate method.

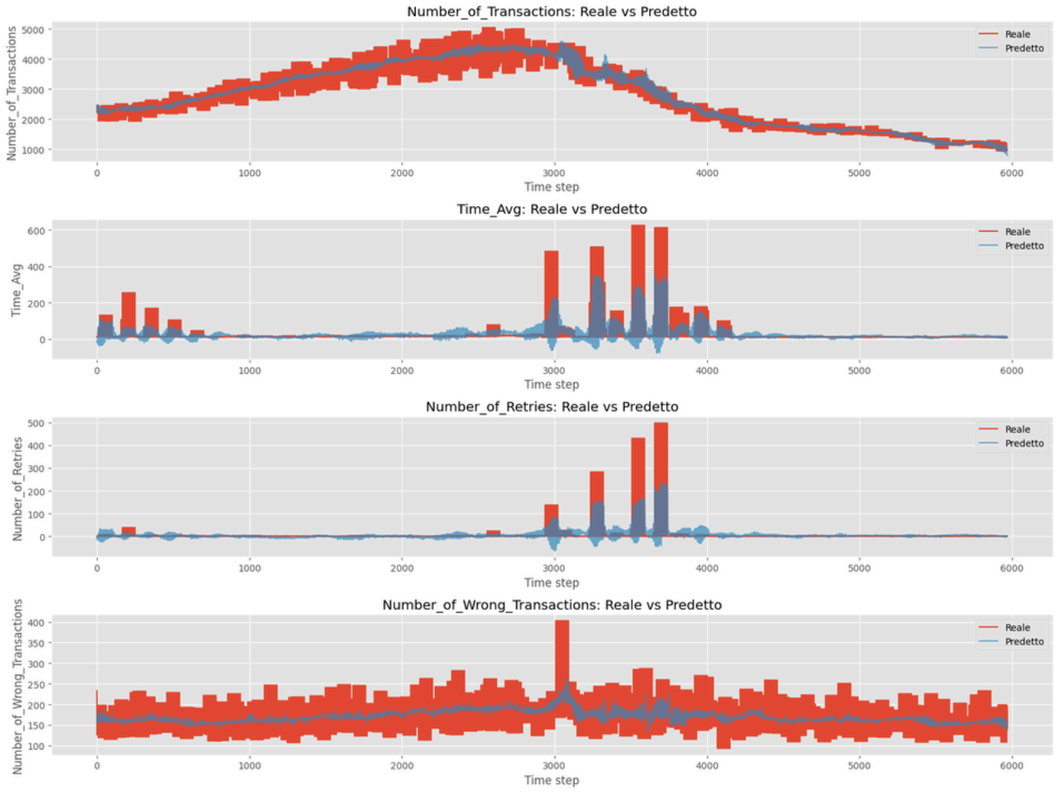


Top 10 detected anomalies:

	Timestamp	Actual	Predicted	Error
61	2024-05-31 19:05:05	3763.0	54067.957031	107.007286
60	2024-05-31 19:04:04	4326.0	50100.679688	102.424473
57	2024-05-31 19:01:03	2264.0	43966.851562	98.429136
59	2024-05-31 19:03:03	3637.0	50577.785156	97.454547
62	2024-05-31 19:06:05	3563.0	54384.019531	97.224899
58	2024-05-31 19:02:03	2411.0	45291.417969	96.619967
56	2024-05-31 19:00:03	2570.0	47127.546875	95.465488
55	2024-05-31 18:59:02	2704.0	45857.910156	94.808368
54	2024-05-31 18:58:02	2560.0	48150.933594	94.744343
49	2024-05-31 18:53:00	2641.0	48952.500000	88.602761

MULTIVARIATE LSTM MODEL

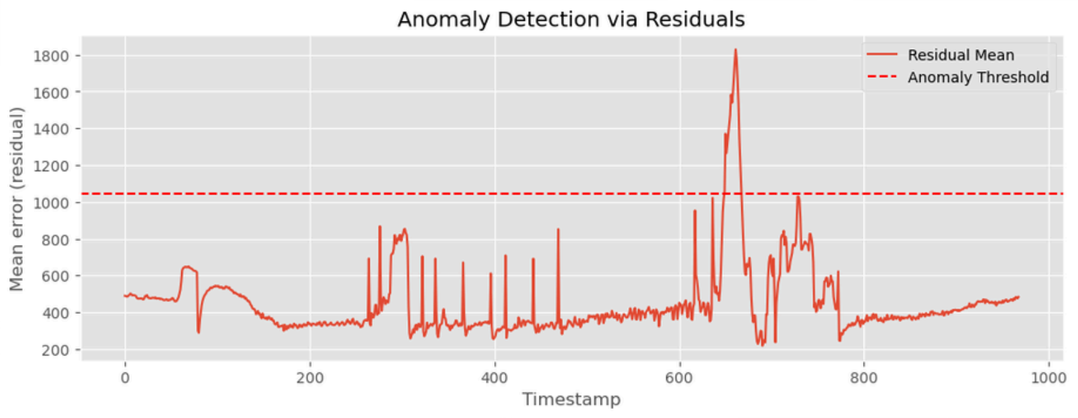
Since our starting point was the Isolation Forest, which employs a multivariate approach, we also chose a **multivariate approach** (Number of Transactions, Time_Avg, Number of Retries, Number of wrong Transactions) for the LSTM model to evaluate whether the anomaly detection results would align more closely




```
Forecast Evaluation Metrics:

Number_of_Transactions:
  RMSE = 267.7101
  MAE  = 196.5568
  R²   = 0.9345

Time_Avg:
  RMSE = 44.8630
  MAE  = 19.2586
  R²   = 0.2351

Number_of_Retries:
  RMSE = 26.6291
  MAE  = 9.4213
  R²   = 0.2082

Number_of_Wrong_Transactions:
  RMSE = 36.4557
  MAE  = 30.1507
  R²   = 0.1143
```

COMPARISON WITH THE ISOLATION FOREST

- with the same approach of the univariate model, to detect anomalies, we applied the model to the original day containing potential anomalies and computed the residuals between predictions and actual values. Anomalies were identified where the average residual (between the four target variables) per prediction window exceeded a threshold based on the normal residual distribution.
- we compared the anomalies with the ones identified by the Isolation Forest



Detected by LSTM	False	True
Isolation Forest Outlier		
False	980	2
True	11	25

	Timestamp	Number_of_Transactions	Time_Min	Time_Max	Time_Avg	Number_of_Retries	Number_of_Wrong_Transactions
710	2024-05-31 18:51:59	2411	0	52047	1298.169300	2734	156
711	2024-05-31 18:53:00	3637	0	40480	902.315250	1162	219
712	2024-05-31 18:54:00	4326	0	5350	541.213750	230	185
714	2024-05-31 18:56:01	3563	1	4762	105.871506	5	410
715	2024-05-31 18:57:01	3973	2	1352	57.112330	1	448

THANK YOU !

REPLY