REPLY
WHITEHALL

Anomaly Detection

Gaia D’Amico
Anastasia Farinaro
Alessia Lorenzini

Project Overview

- Data Exploration
- Augmentation
- Forecasting
- Anomaly Detection

Dataset Overview

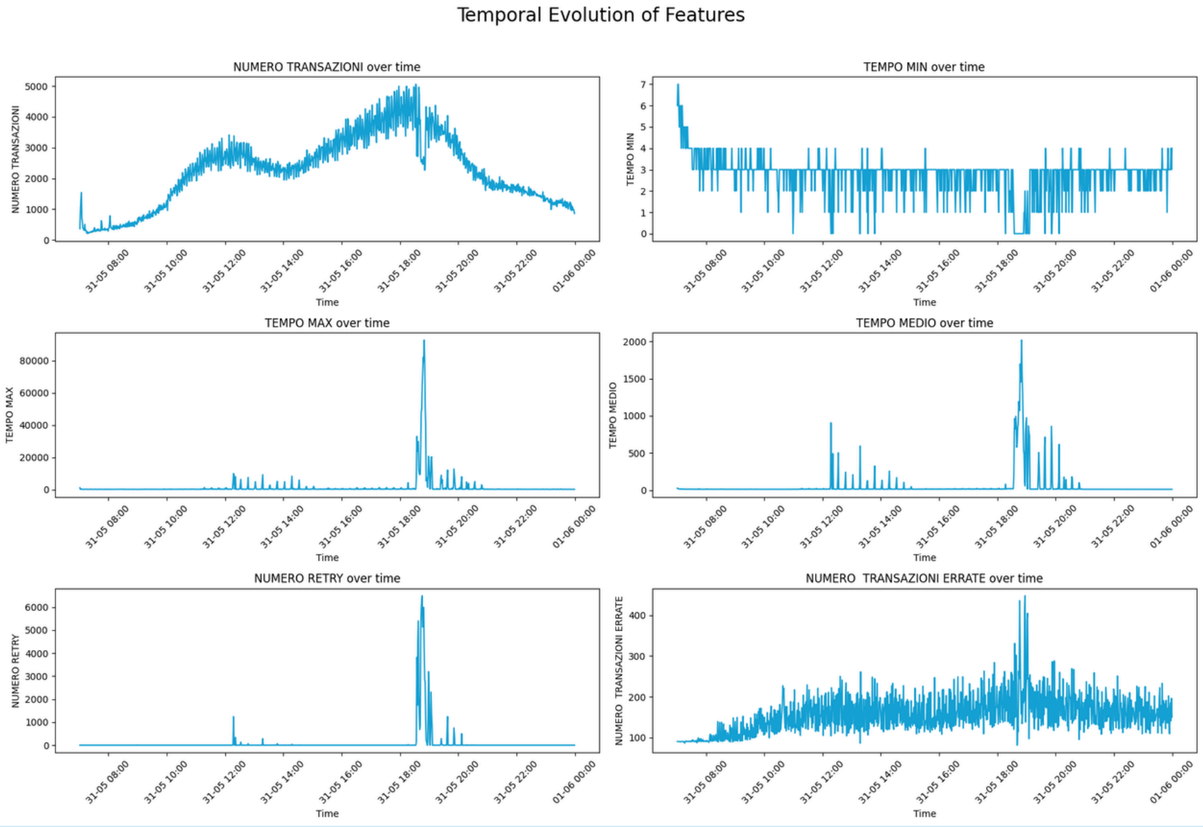
- Dataset: FullDayWithAlarms.xlsx
- Duration: 1 single day – May 31st, 2024
- Frequency: 1 row every minute → 1,018 records

	Code_ID	Timestamp	Acquisition_Interval	Number_Transactions	Time_Min	Time_Max	Time_Mean	Number_Retries	Number_Wrong_Transactions
0	8	31/05/2024 07:01:11	60	366	6	1019	25.907562	0	90
1	8	31/05/2024 07:02:11	60	948	7	90	18.181719	0	90
2	8	31/05/2024 07:03:11	60	1273	7	408	18.813356	0	90
3	8	31/05/2024 07:04:11	60	1538	6	70	16.607435	0	90
4	8	31/05/2024 07:05:11	60	703	5	85	16.645409	0	90

Main features:

- **Number of Transactions** – count per minute
- **Time Min, Time Max, Time Mean** – processing time indicators
- **Number of Retries** – how many retries were attempted
- **Wrong Transaction**s – failed operations

Exploratory data analysis



This chart shows the temporal evolution of the main numerical features in the dataset across the whole day (31/05/2024).

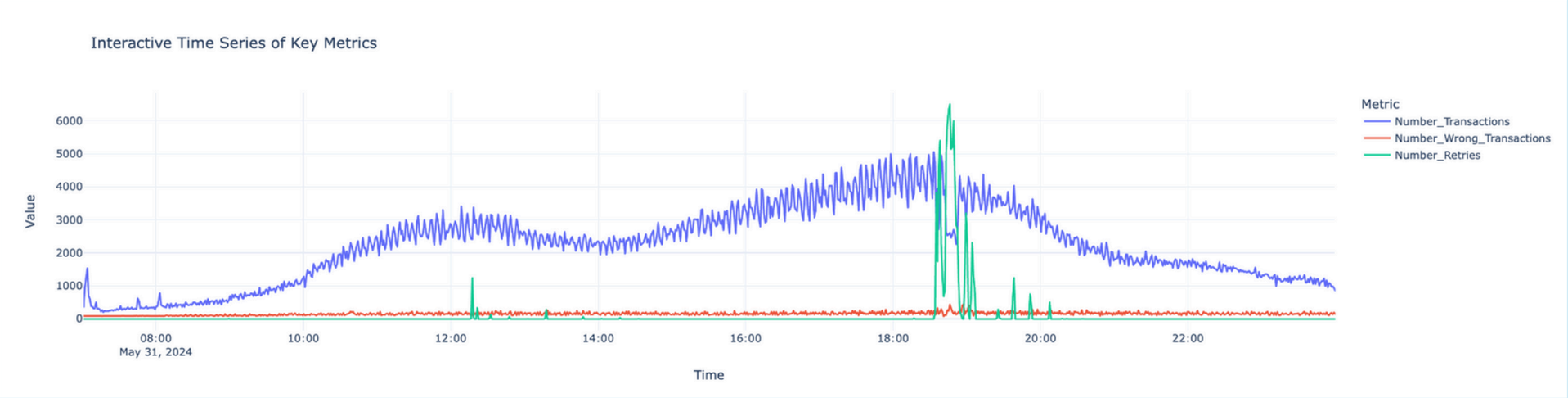
- **Number_Transactions** increases during the morning and reaches a peak around 17:00, then drops off gradually.

- **Time_Min and Time_Max** show occasional spikes that may indicate delays or irregular system responses.

- **Time_Mean** follows a similar trend, with a very high peak around 17:40, potentially highlighting a critical moment.

- **Number_Retries and Number_Wrong_Transactions** both exhibit bursts in the late afternoon, which may signal anomalous behavior or overload.

Exploratory data analysis



This interactive time series chart shows the behavior of three key metrics throughout the day on May 31st, 2024:

- **Number_Transactions (blue)**: represents the total number of transactions per minute. We observe a gradual increase during the day, peaking around 18:00, followed by a decrease in the evening.

- **Number_Retries (green)**: represents the number of retry attempts. There are distinct bursts between 17:30 and 19:30, which strongly correlate with the peak in transaction volume. These spikes may indicate system overload or instability.

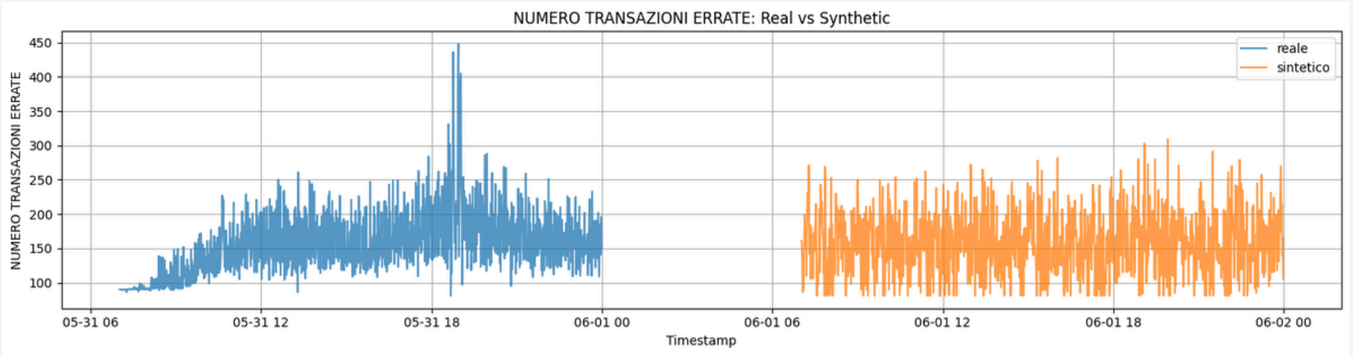
- **Number_Wrong_Transactions (red)**: tracks the number of incorrect or failed transactions. Although it stays relatively stable throughout the day, it slightly increases around peak load hours, suggesting that errors become more frequent when the system is under pressure.

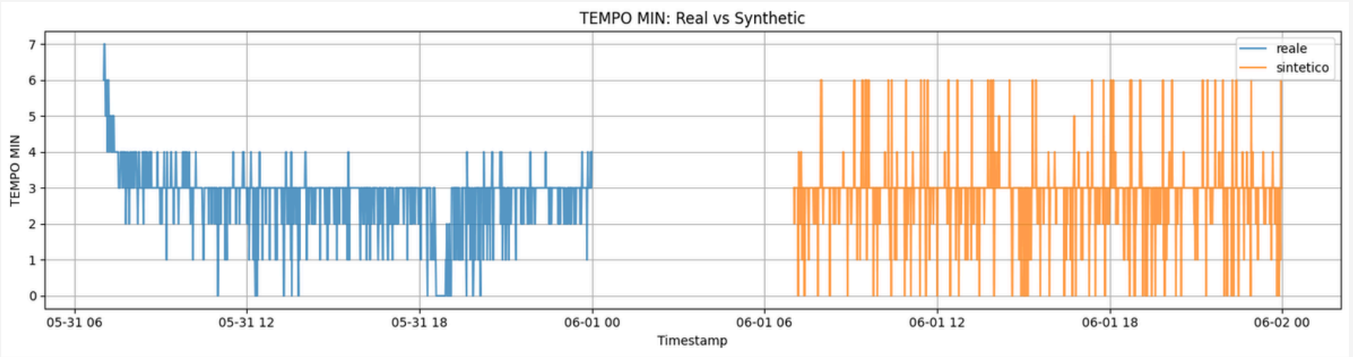
Data augumentation

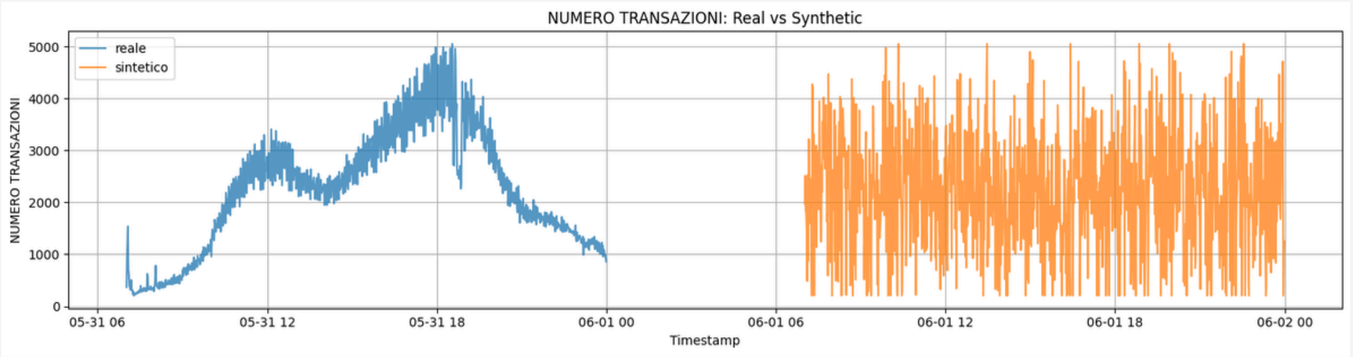
Synthetic data via SDV (Gaussian Copula)

Ensure that the synthetic data generated through augmentation is realistic, coherent, and suitable for training robust models.

Verify temporal consistency



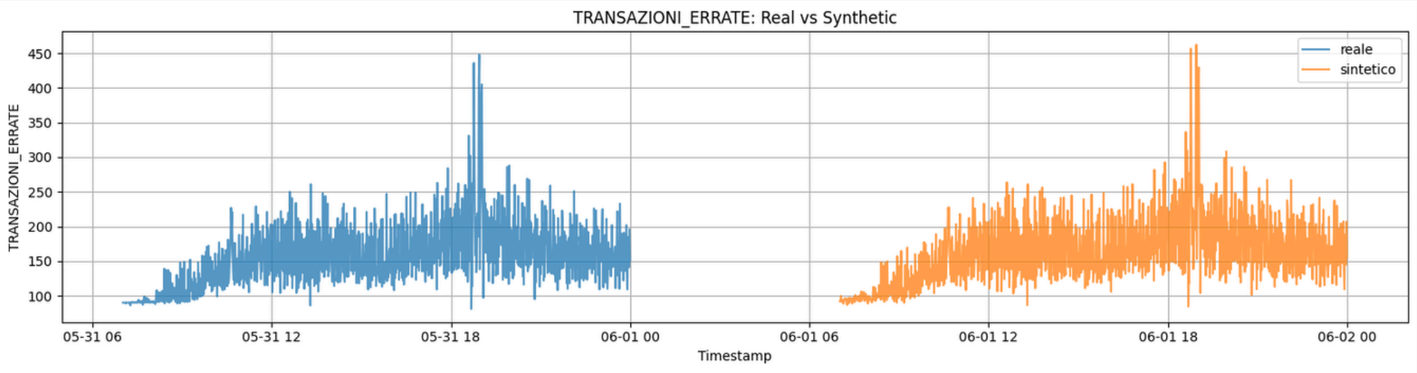


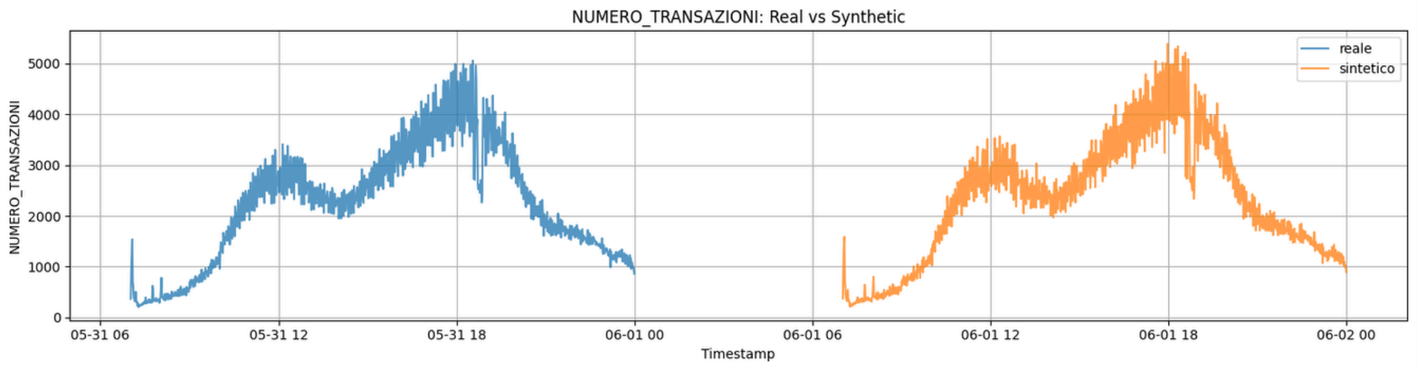


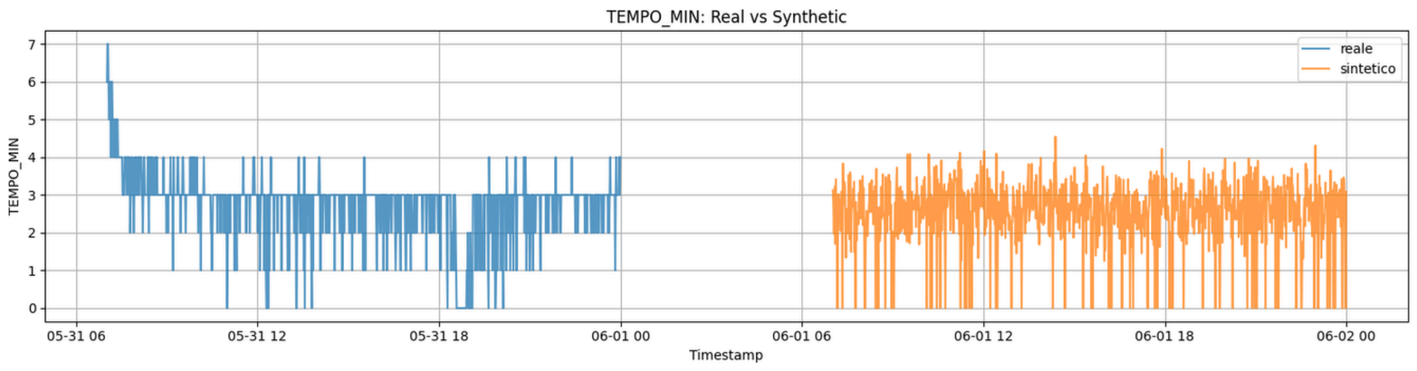
Data augumentation

Using LLM

Verify temporal consistency



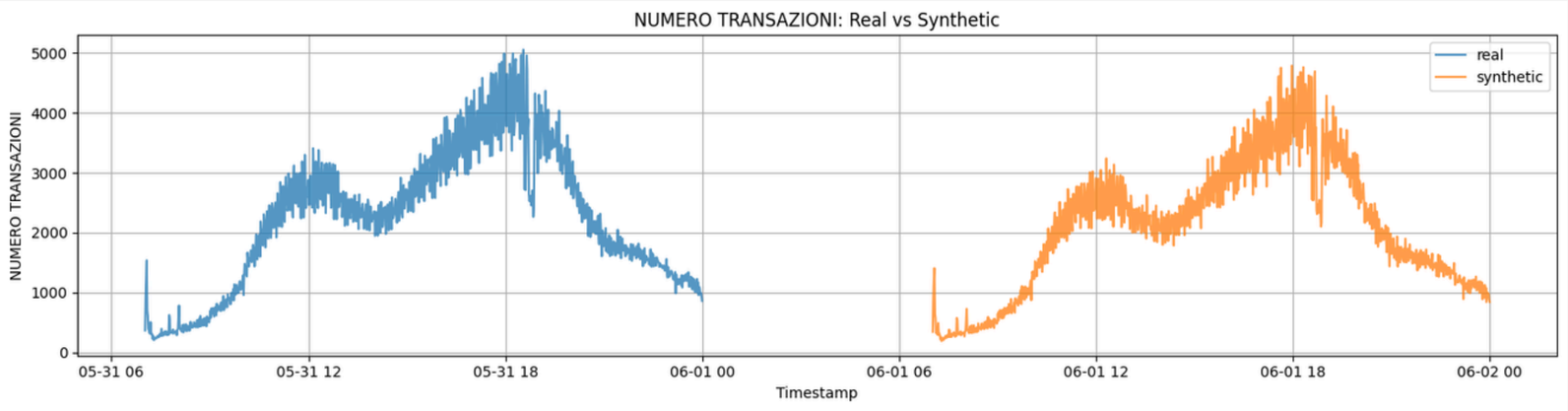


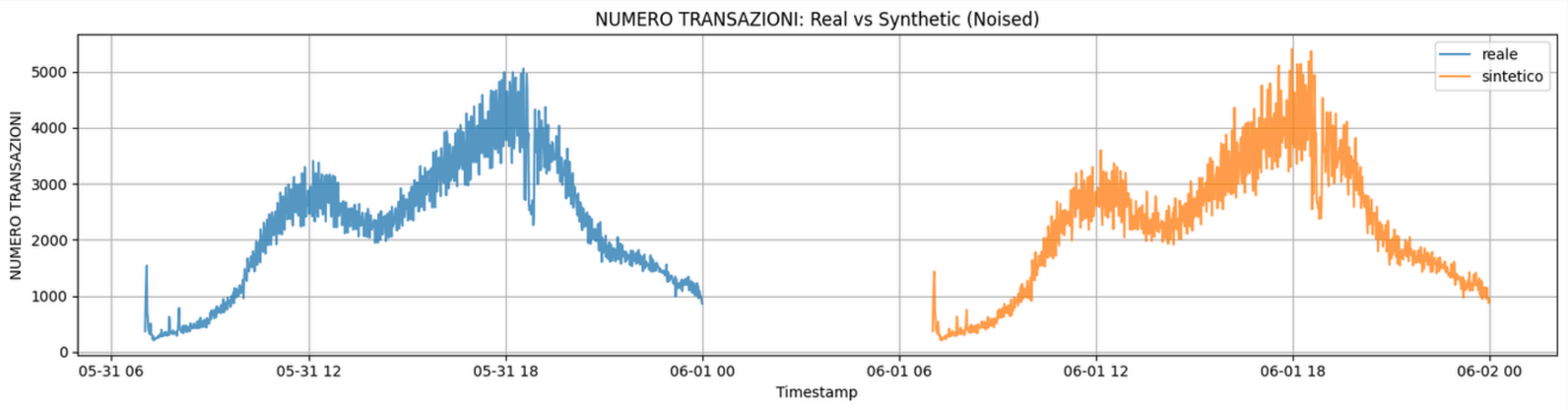


Ensure that the synthetic data generated through augmentation is realistic, coherent, and suitable for training robust models.

Data augumentation

Timestamp shifting + noise injection





To extend the dataset, we replicated the original real-day data (May 31, 2024) three times:

- Day_1 → June 1, 2024
- Day_2 → June 2, 2024
- Day_3 → June 3, 2024

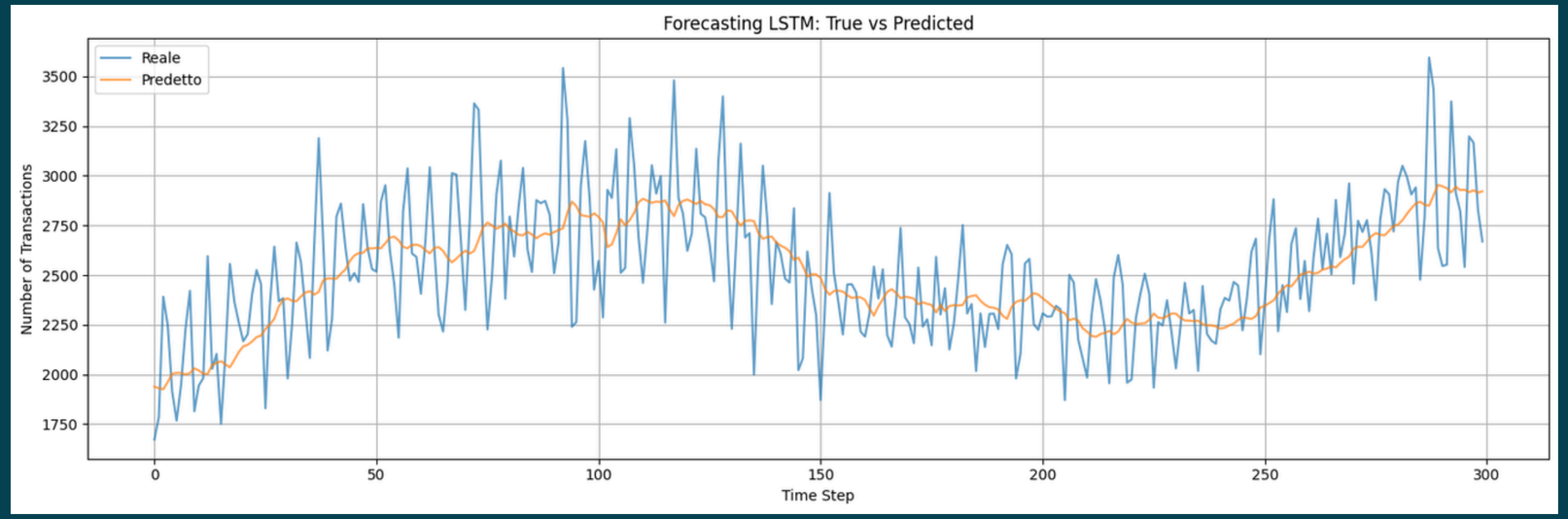
Each copy preserved the original temporal structure (07:01–23:59) and was shifted by one day.

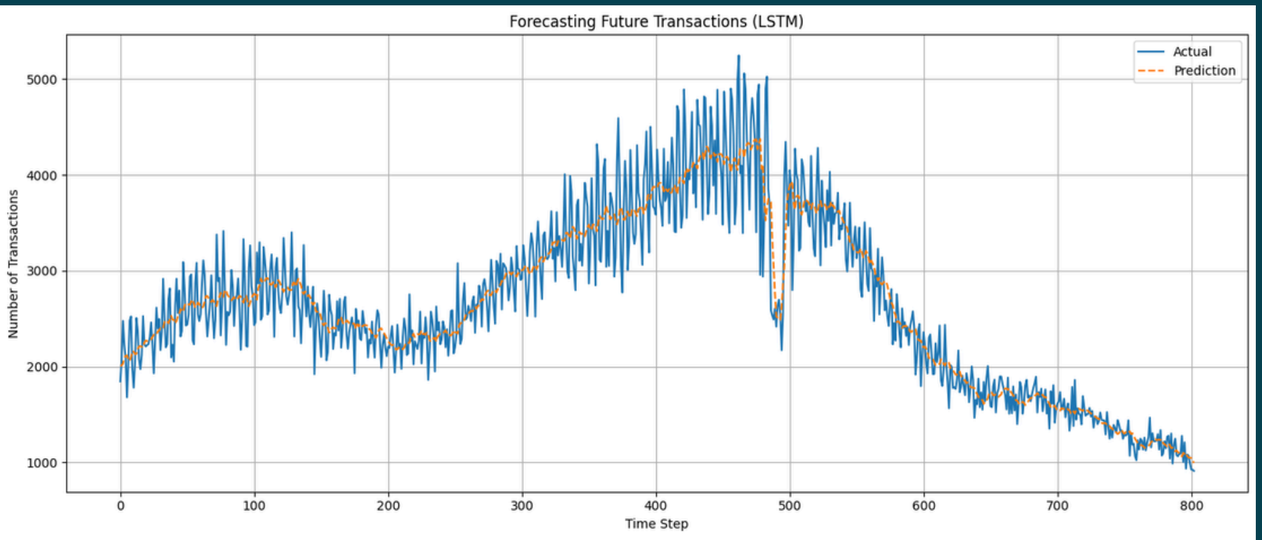
To improve variability, we added light Gaussian noise to the synthetic days.
This process allows us to test forecasting and anomaly detection on a richer, more realistic time series.

Forecasting- LSTM

Residual Analysis

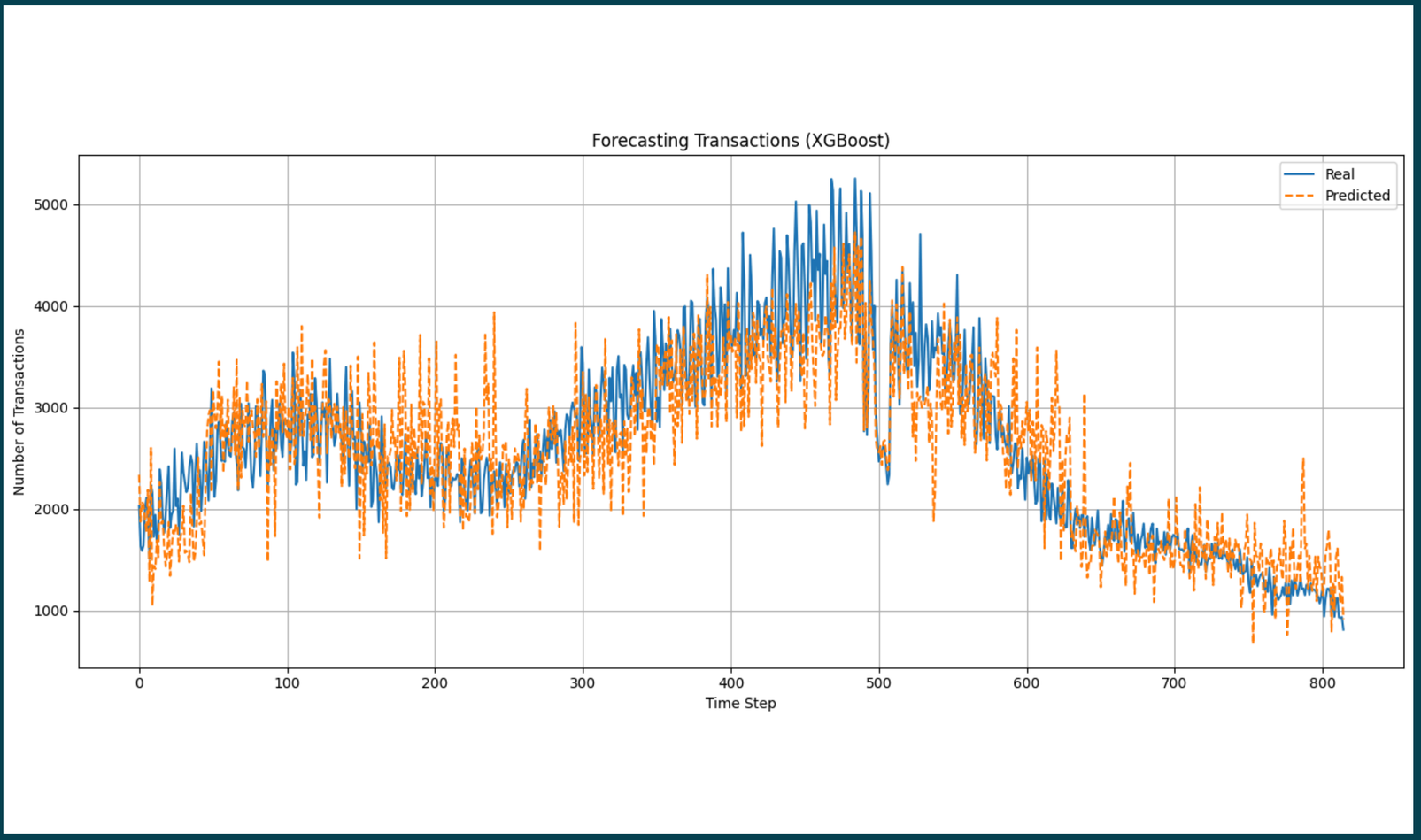
The LSTM model effectively captures the overall trend of the series, smoothing out high-frequency fluctuations. Predictions remain well aligned with the true signal.





LSTM (Long Short-Term Memory) is a type of recurrent neural network specifically designed to capture long-range temporal dependencies in sequential data. This makes it especially well-suited for tasks like time series forecasting, where patterns unfold over time.

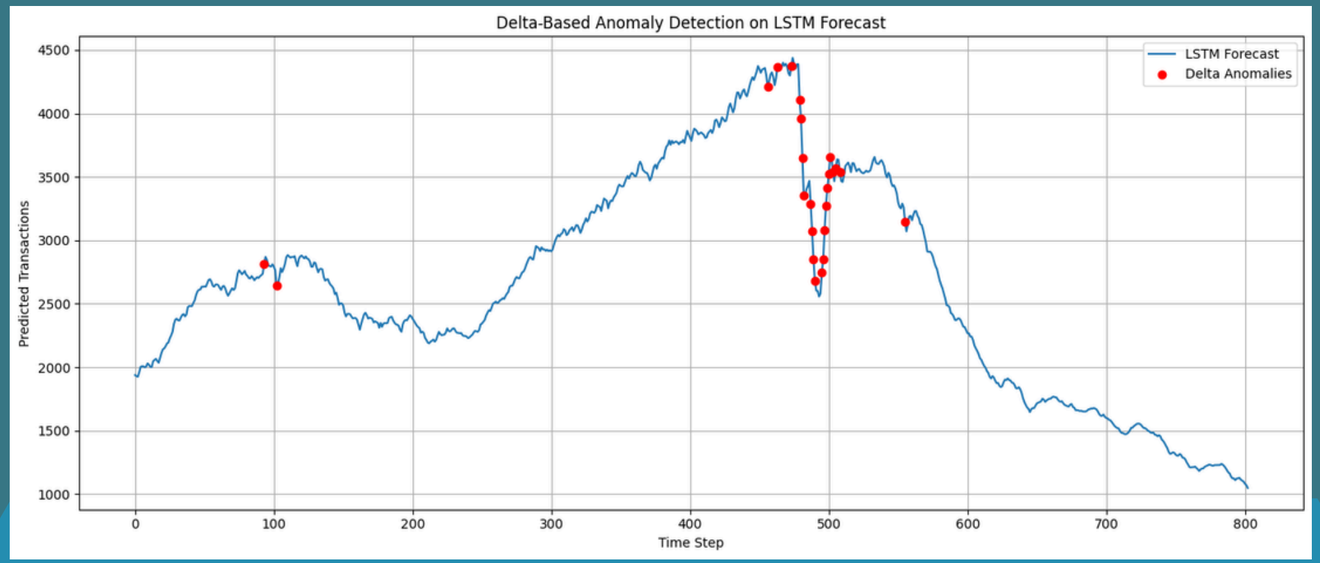
Forecasting - xgboost

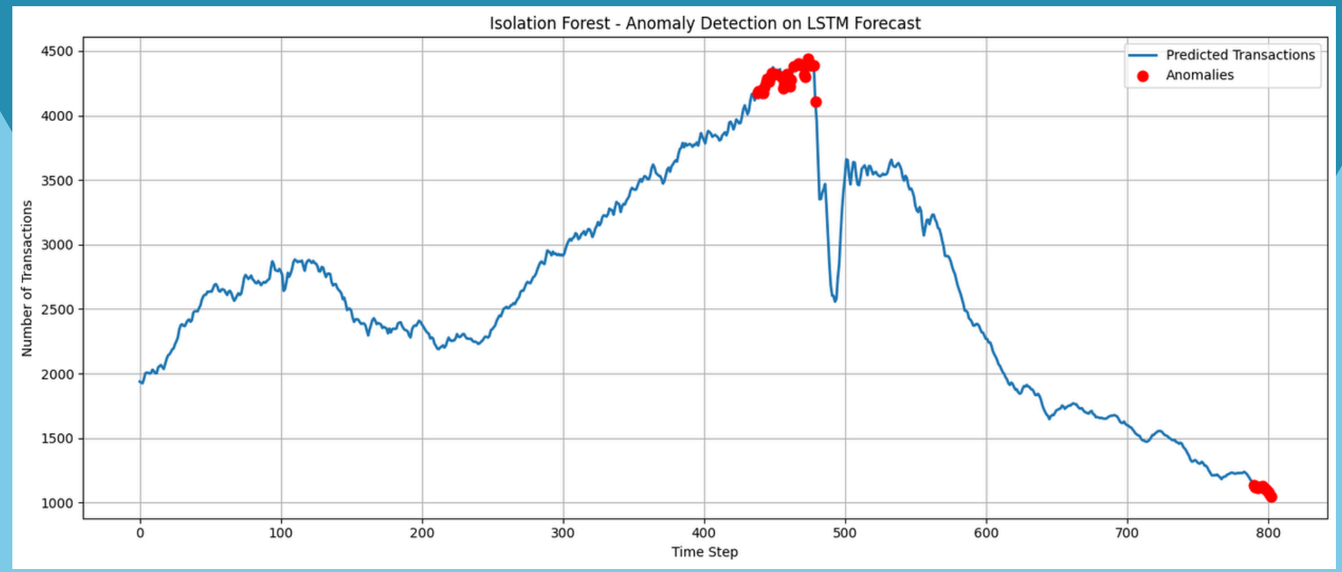


- XGBoost was chosen for its **flexibility** and robustness with **tabular** data.
- **Feature Engineering:** Included lag variables and rolling statistics to capture temporal patterns.
- Evaluation: Temporal **train-test split**
- Metrics: **RMSE**, MAE, R^2.
- Result: **Realistic** prediction with strong **trend adherence** and good **peak** tracking.

Anomaly Detection using LSTM

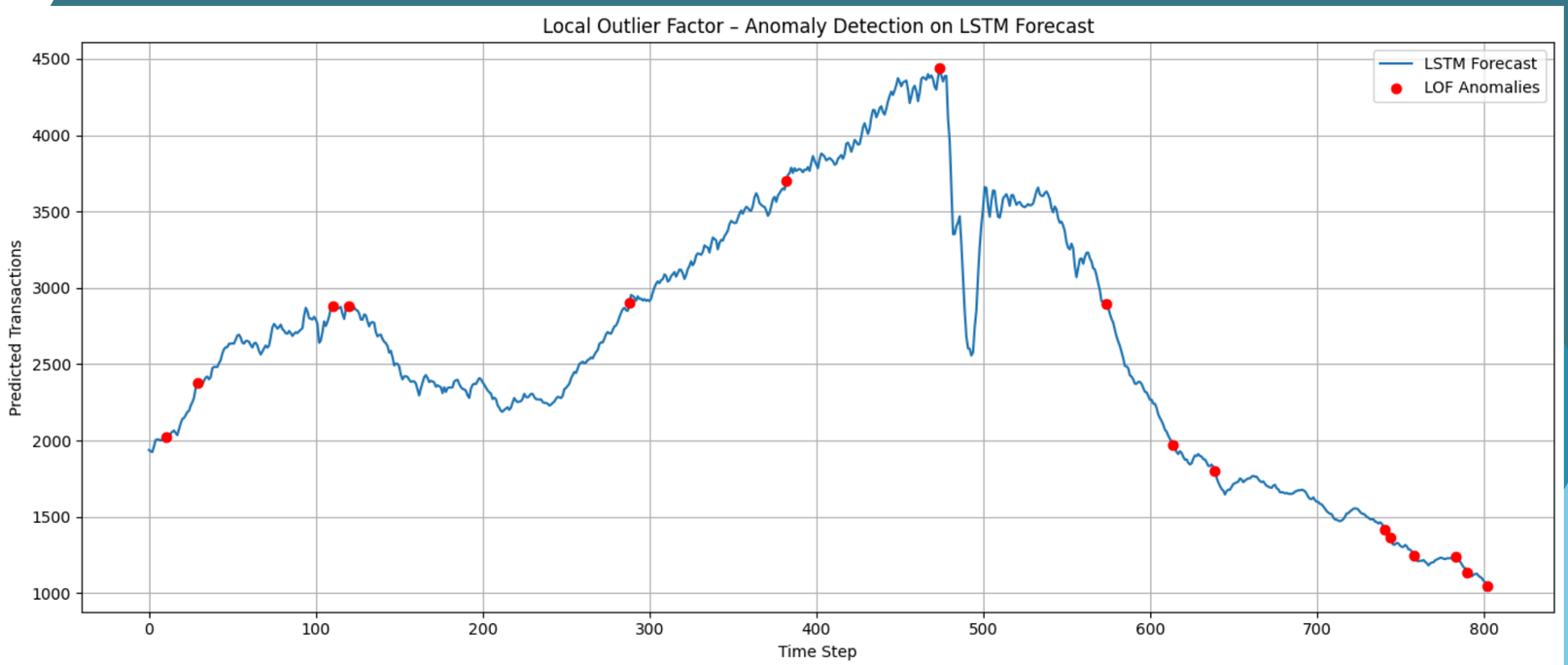
- **Delta Based**
- **Isolation Forest**





Anomaly Detection using LSTM

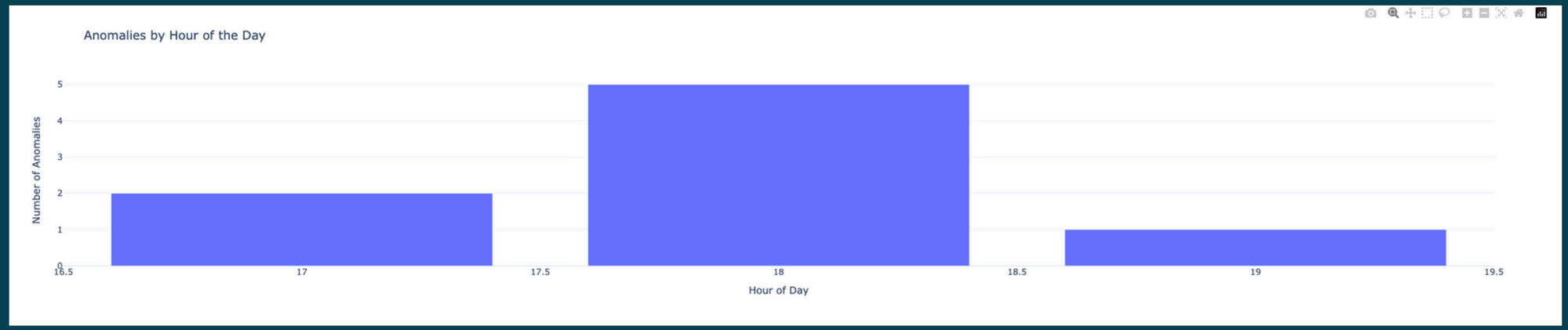
Local Outlier Factor (LOF)
LOF compares the local density of a point to that of its neighbors. If the density is significantly lower, the point is labeled as an anomaly.



Conclusion

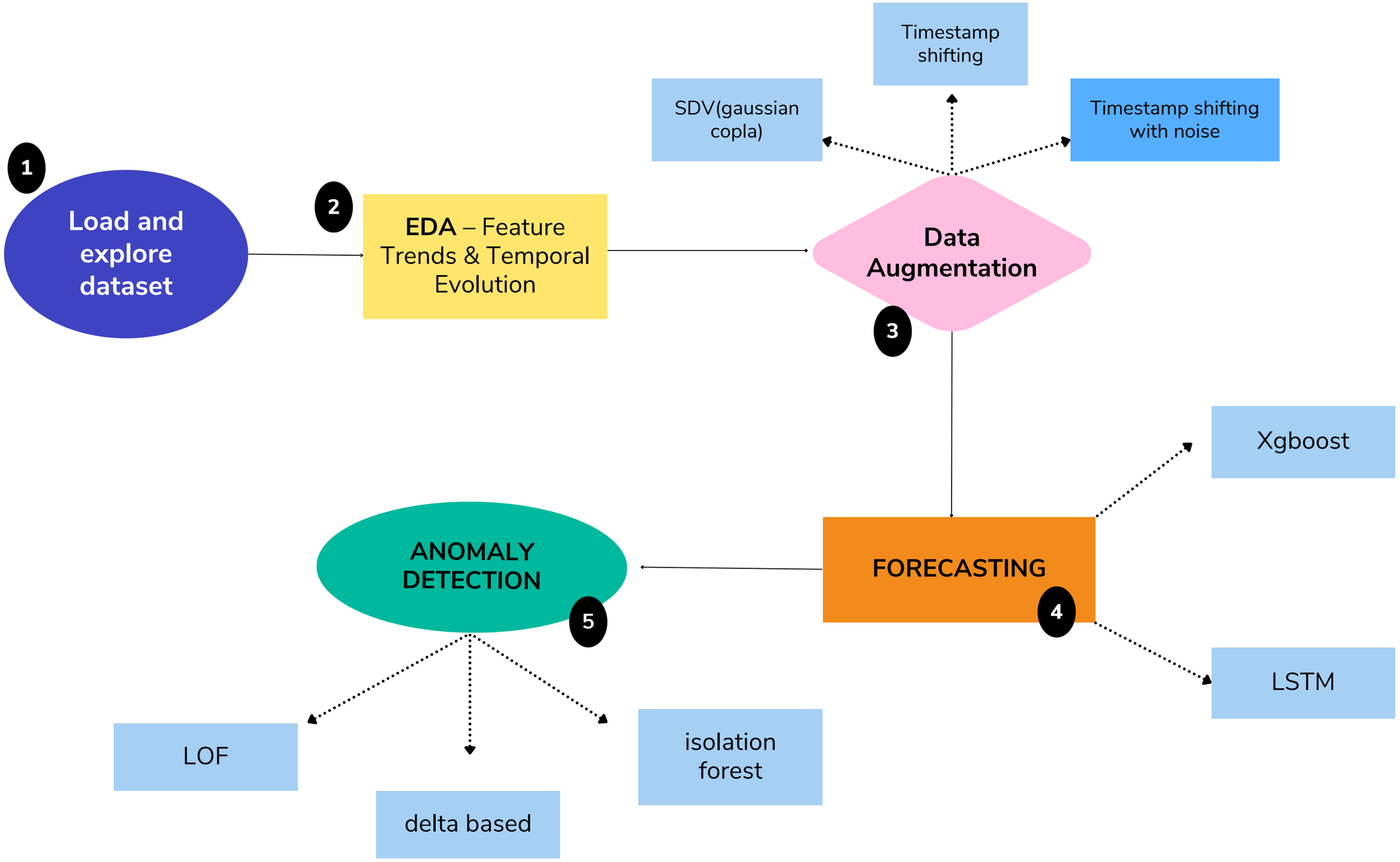
The combination of forecasting and anomaly detection enabled us to uncover both global trends and critical local deviations in transaction behavior.

- The XGBoost model proved effective in capturing the dynamics of the signal.
- Anomalies consistently clustered around **17:30–19:00**, revealing a time-sensitive instability window.
- Different detection techniques complemented each other, offering both robustness and granularity.



This analysis provides a foundation for real-time monitoring, alert systems, or capacity planning in future applications.

THANK YOU
FOR THE ATTENTION!


1
Load and explore dataset
2
EDA – Feature Trends & Temporal Evolution
Timestamp shifting
SDV(gaussian copla)
Timestamp shifting with noise
Data Augmentation
3
FORECASTING
4
Xgboost
LSTM
ANOMALY DETECTION
5
LOF
delta based
isolation forest