REPLY X LUISS

Needle in Haystack

Anomaly Detection with Codeine

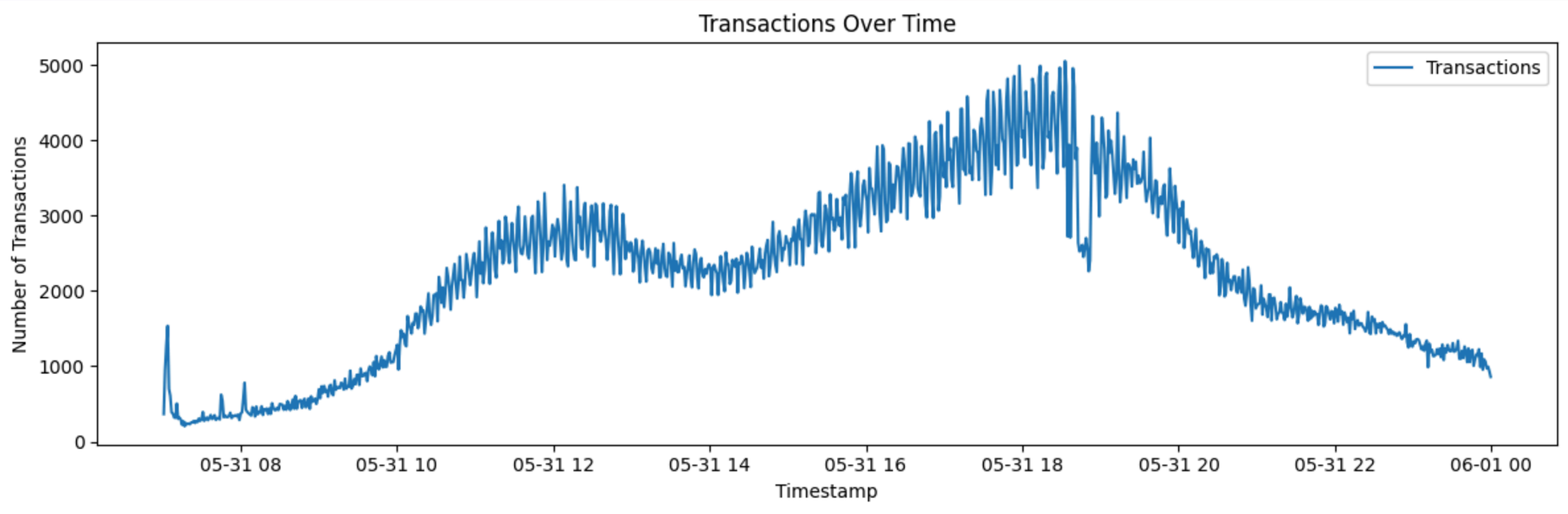
Shayan

Mohammad

Sergio

Codeine is a good painkiller, it promises to **REPLY** fast, not like many, thousand years later!

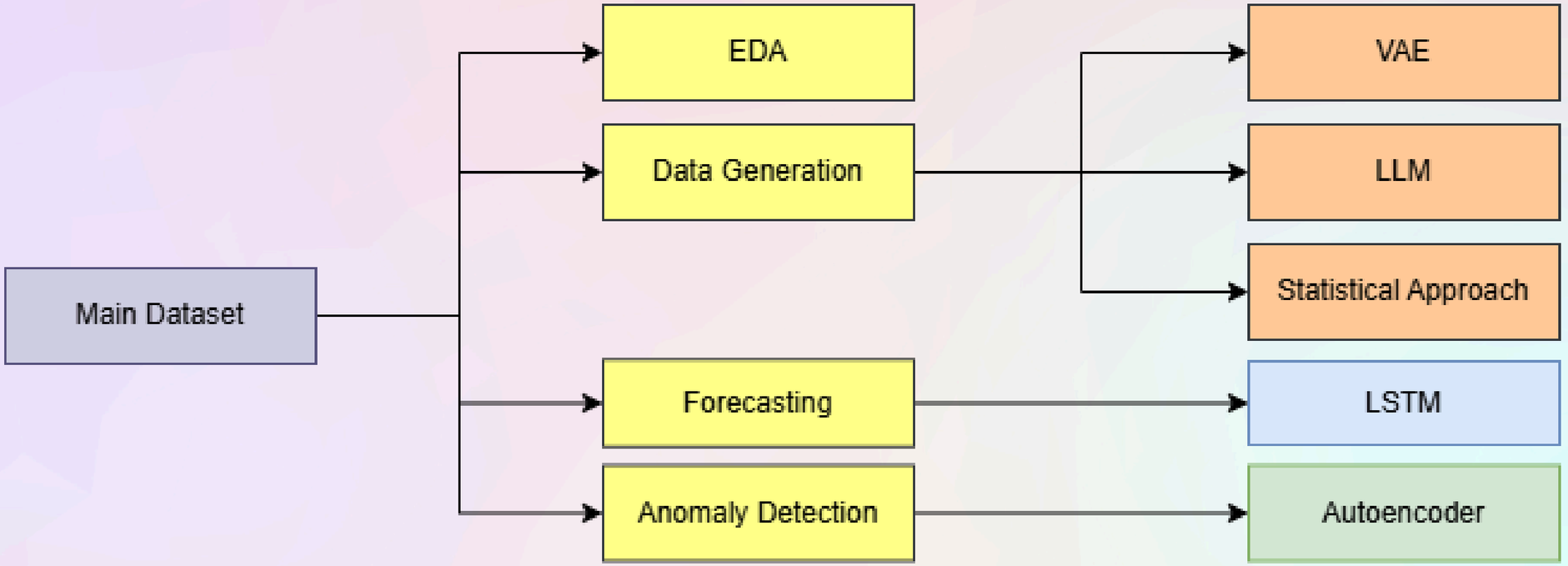
Introduction

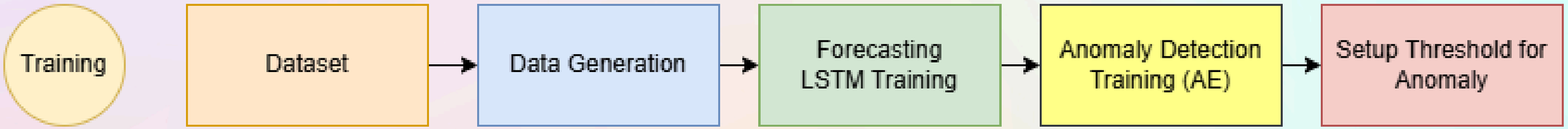


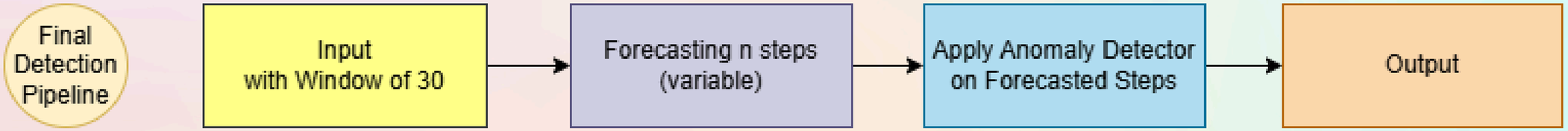
What should we do to become better detectives in this enviorment?

- **Understand Normal Patterns**:
  Learning how the system behaves under typical conditions
- **Spot Suspicious Signals**:
  Identifying unusual deviations that may indicate anomalies
- **Combine forecasting and anomaly detection models**:
  Integrating both approaches for more accurate and robust insights
- **Forecast Future Trends:**
  expected/unexpected behaviors

Overview

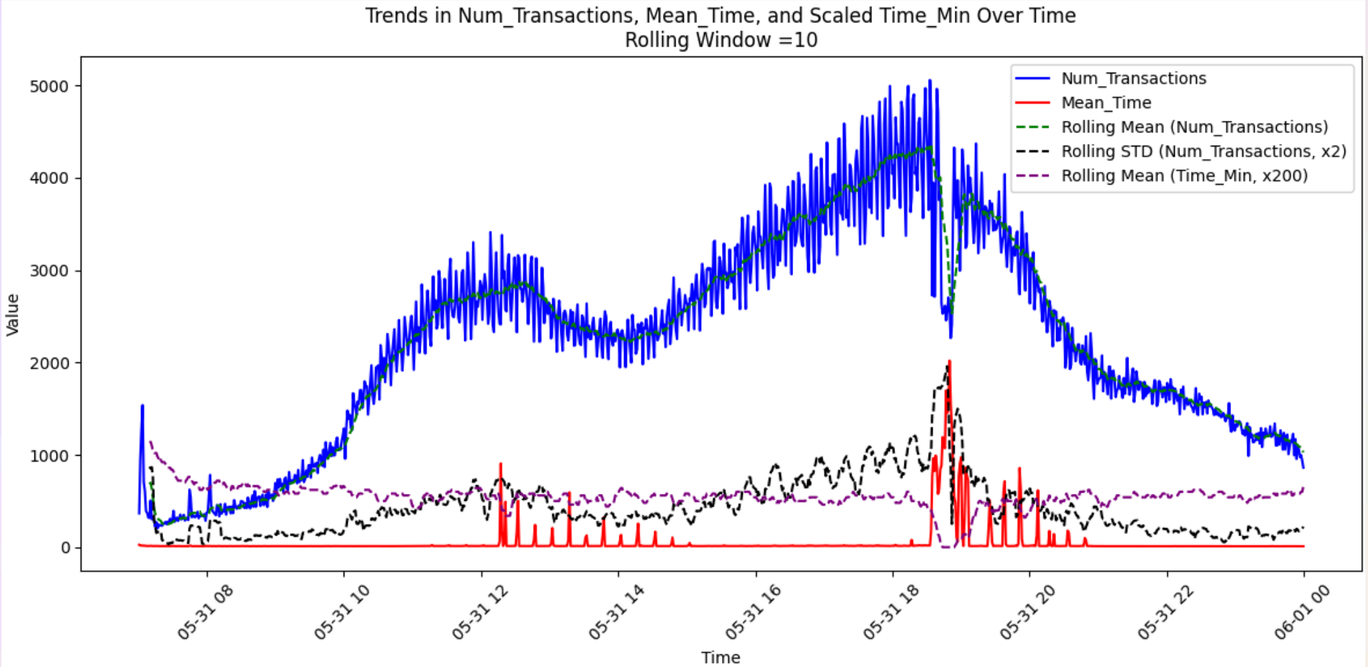

Main Dataset
EDA
Data Generation
Forecasting
Anomaly Detection
VAE
LLM
Statistical Approach
LSTM
Autoencoder


Training
Dataset
Data Generation
Forecasting
LSTM Training
Anomaly Detection
Training (AE)
Setup Threshold for
Anomaly


Final
Detection
Pipeline
Input
with Window of 30
Forecasting n steps
(variable)
Apply Anomaly Detector
on Forecasted Steps
Output

Exploratory Data Analysis

Exploratory Data Analysis



We observed that the number of transactions over time does not follow a consistent trend. Instead, it fluctuates in a **zigzag pattern** between intervals.

`Time_Min` shows a moderate negative correlation with most other features**,** This implies that when the minimum transaction time decreases, (crowded times of system), the system tends to perform worse, with more retries and higher average or maximum durations.

Num_Transactions

Clear trends in times of day
Peaking around 18:00.
Patterns in number

Num_Transactions (Rolling Variance)

Zigzag pattern of transactions
Patterns in fluctuations

Time_Min (Rolling Mean)

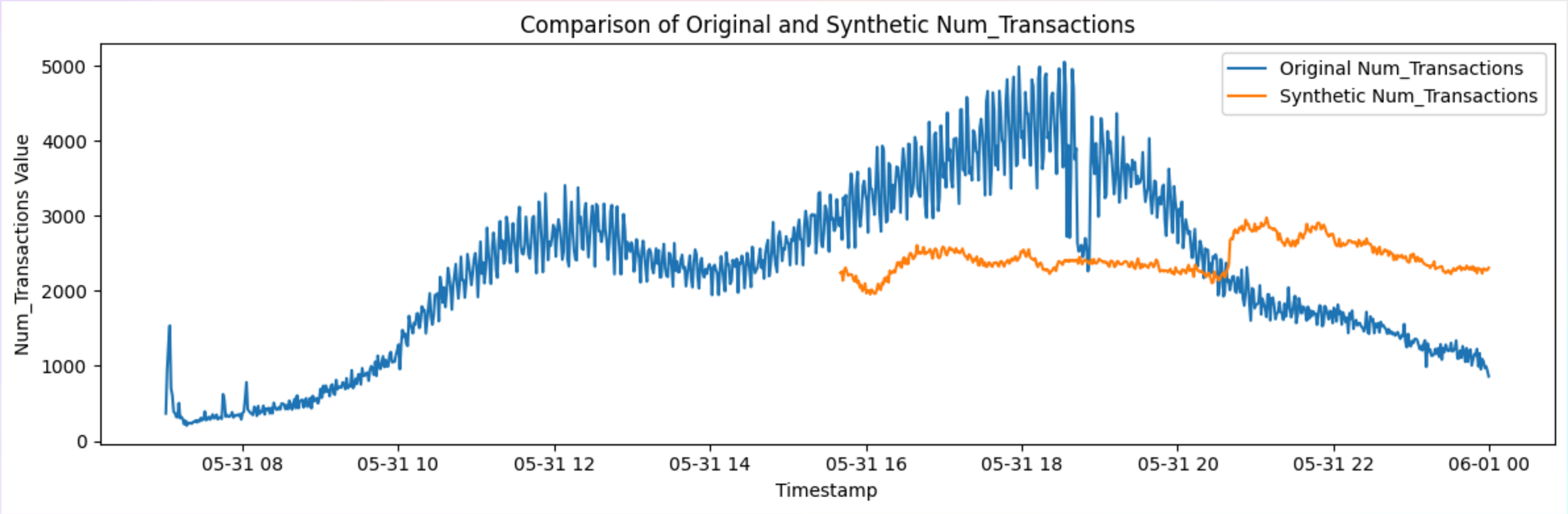
Under 1 vs other metrics
Negative correlation
Lower `Time_Min` and potential instability

Mean_Time

Remains low and stable for much of the day
sudden sharp spikes, particularly after 18
along with a steep decline with Time_Min

Data Generation

Data Generation – VAE



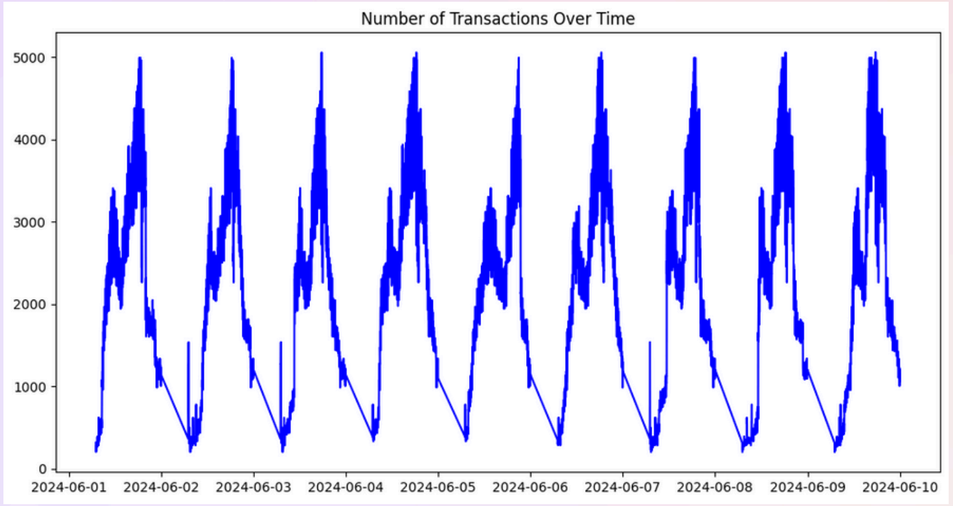
We first examined **VAE** for data generation, but encountered significant **difficulties.**

.

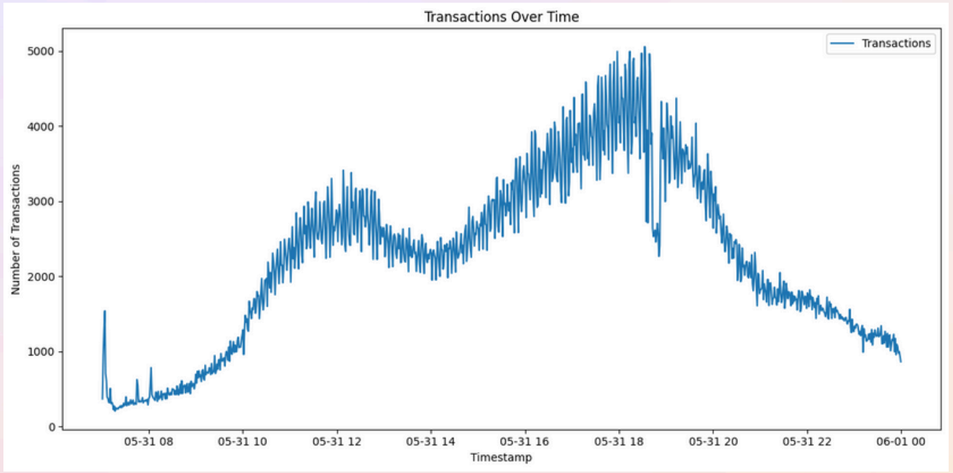
- Tuning the VAE was both **challenging** and **time-consuming,**
- Despite our efforts, we were only able to generate a **limited number of samples**.
- More critically, the generated data **lacked the crucial variance** found in the real data, rendering it ineffective for our purposes.

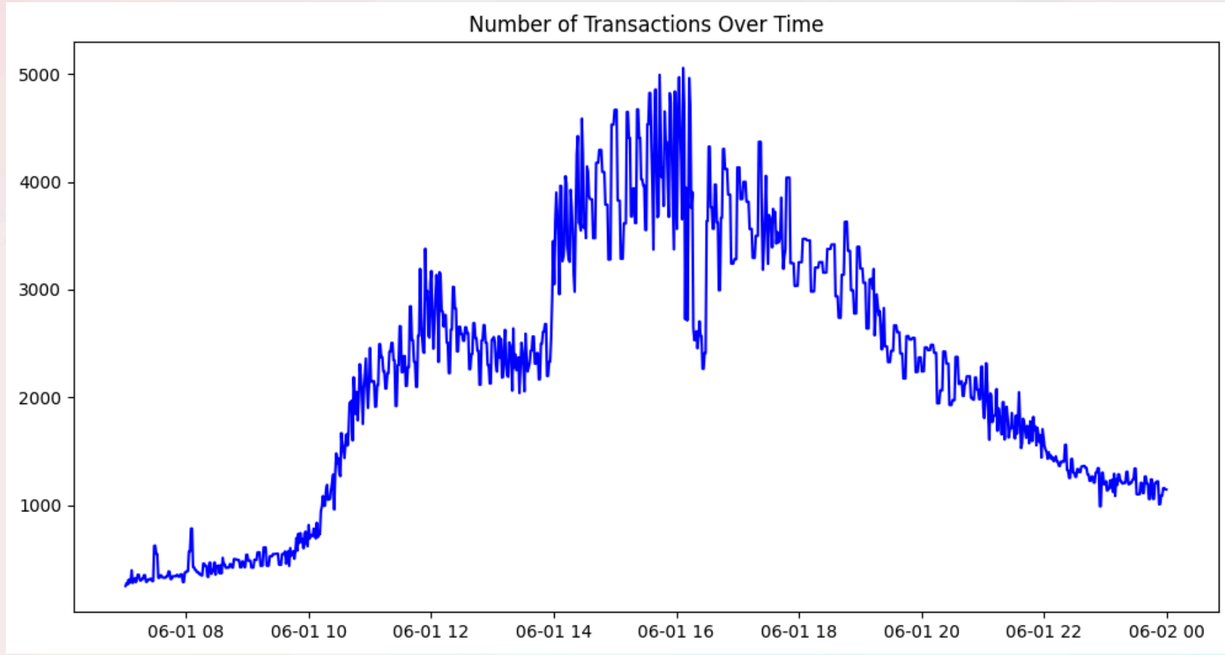
Data Generation – Statistical

Generated



Real





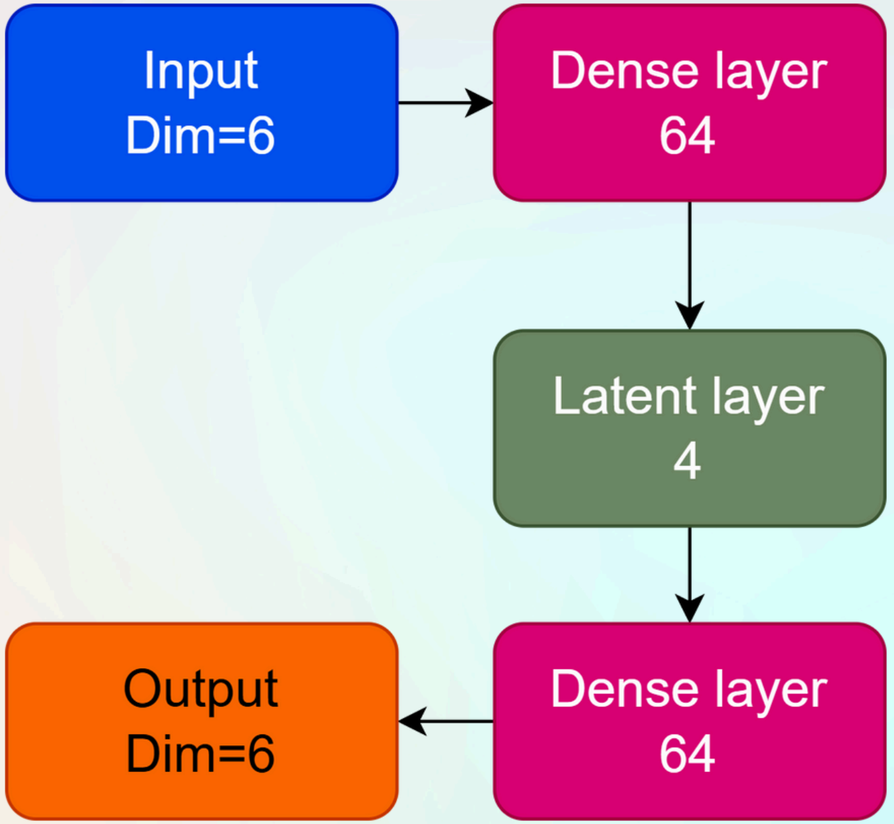
We initially attempted to run **LLMs locally**, but **hardware limitations** prevented effective execution or good outputs. We then switched to an **iterative statistical approach** to generate similar data, adding random noise, variations, and random distribution settings.

Anomaly Detection

Anomaly Detection – Autoencoder

This step involved using an **Autoencoder (AE)** to reconstruct data from a compressed state.

- The reconstruction error and a set threshold were used to classify data points as anomalous.
- We also tried adding 2 noisy sample for each record. But this only added computational cost because it would only lower the threshold sensitivity.
- Training this model was the easiest part of the project.



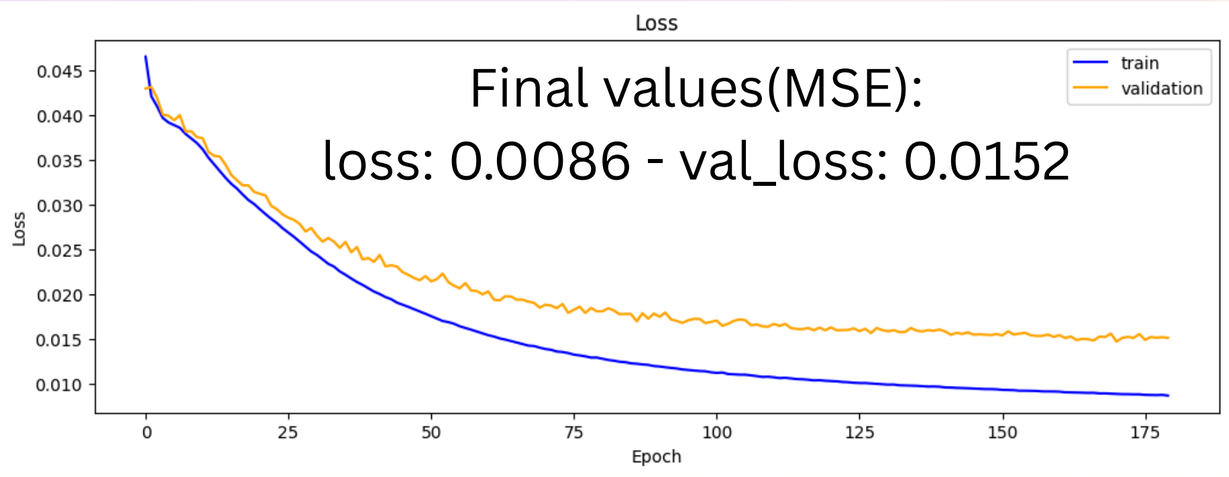
The overal reconstruction error: **3.4436e-5**

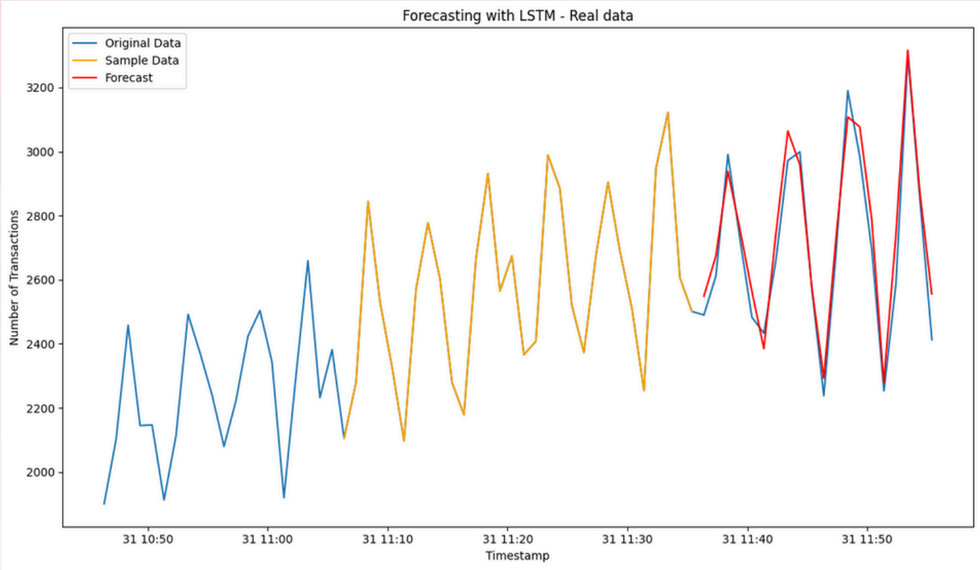
Forecasting

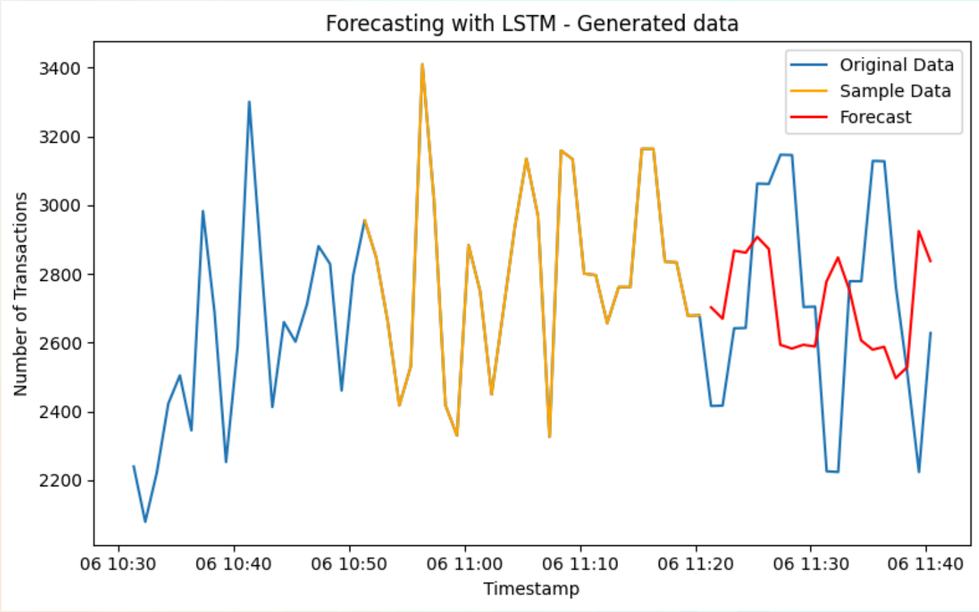
Forecast – LSTM

We used **LSTM** for future prediction.

- We set a **low learning rate**, as a high one would make the model bounce over the curves. This made the learning slow but good.
- The model was still c**onverging even after 180 epochs**. We also added **2 noisy samples** of each record for training to prevent overfitting.
- A notable point about our model is that it **generates only one point in the future and moves the window one step ahead at a time,** so we can use it to produce a v**arying number of future points.**

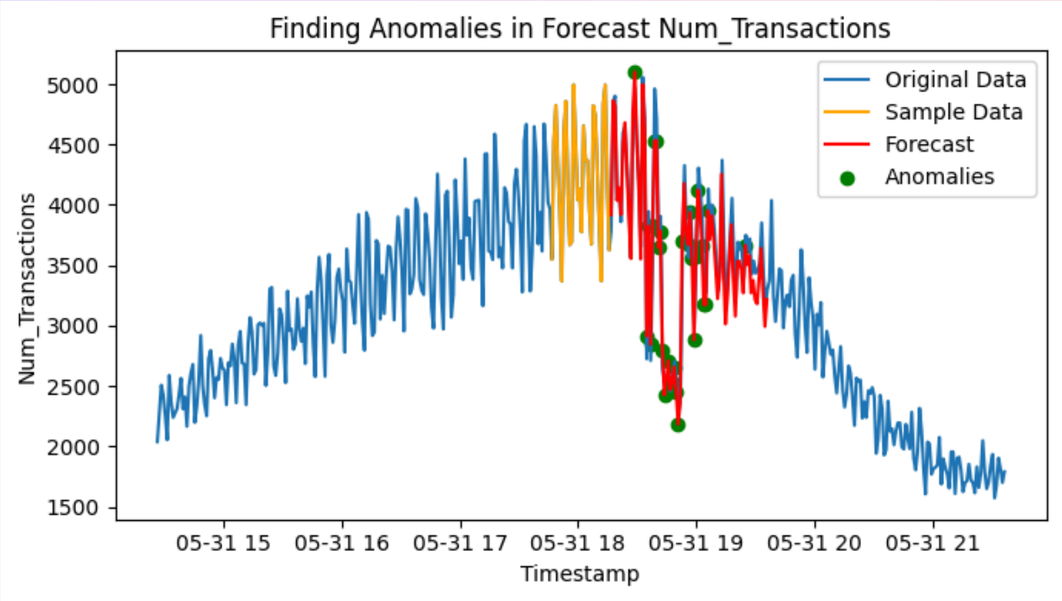


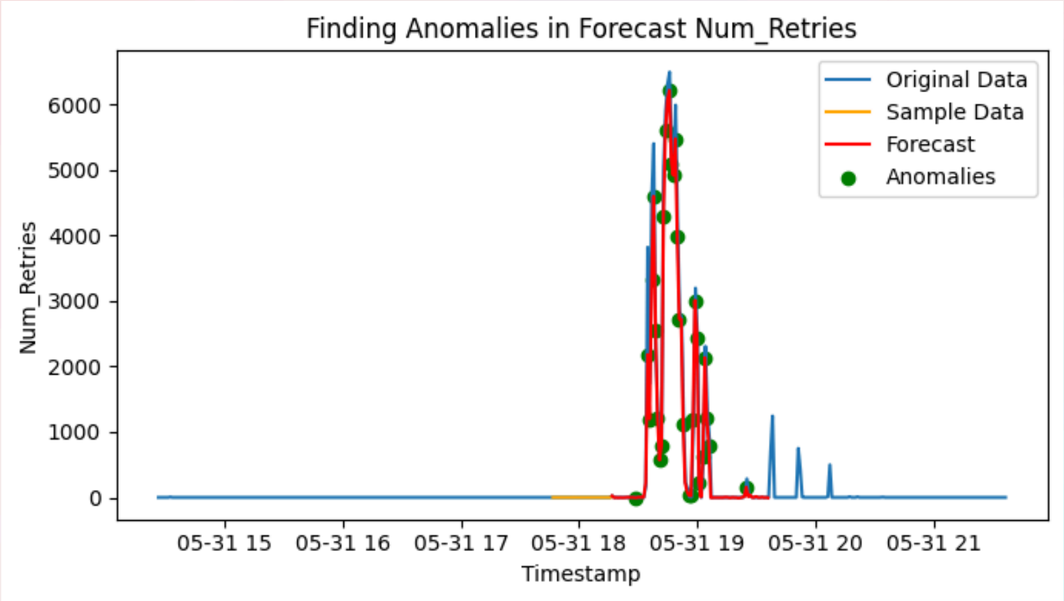


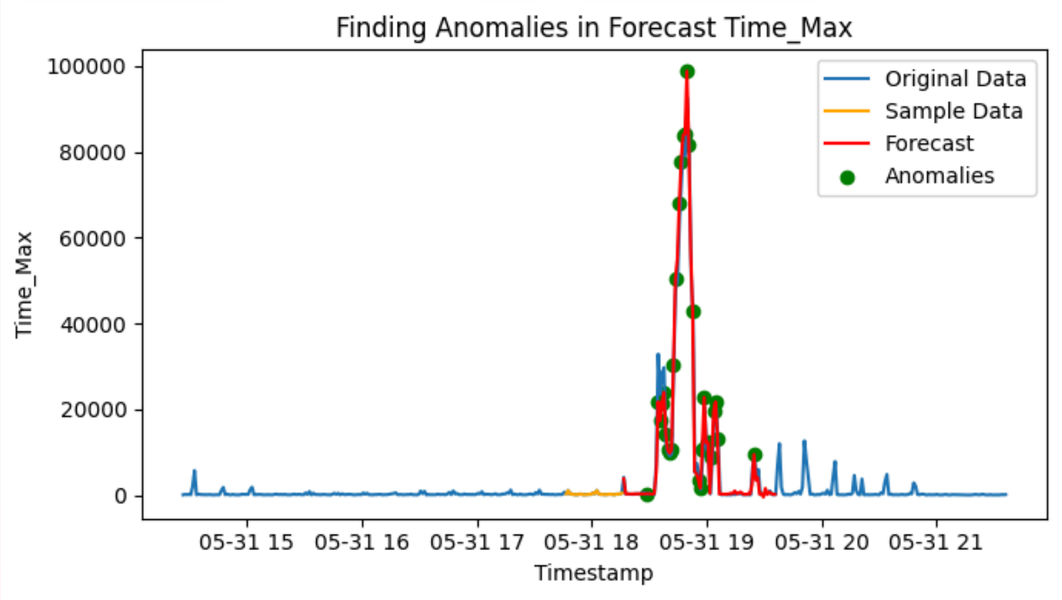


Results Evaluation

Results Evaluation



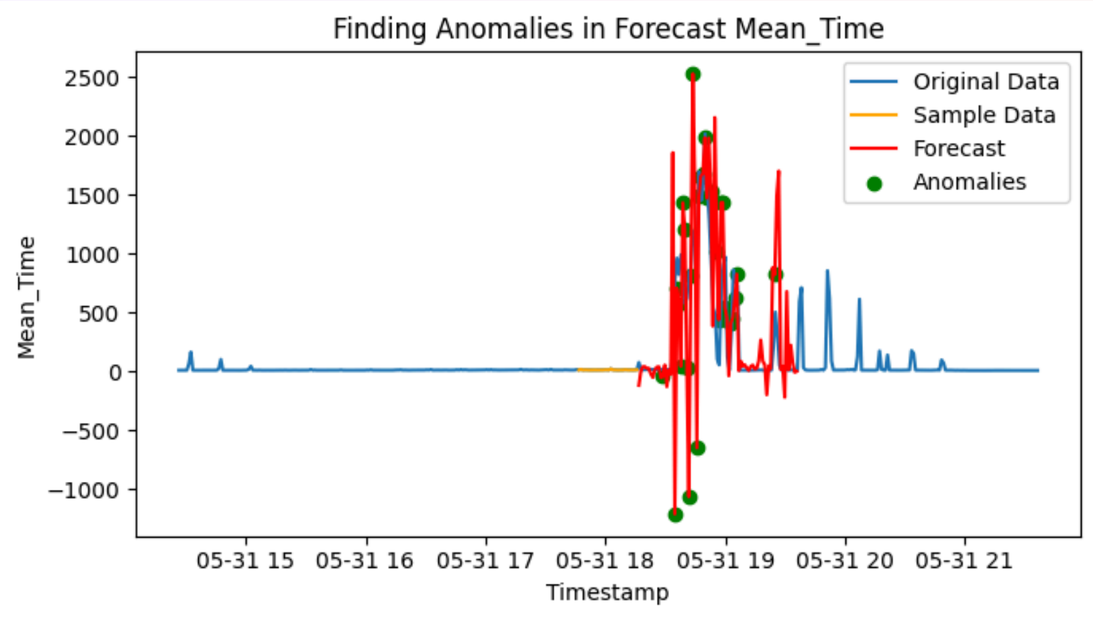


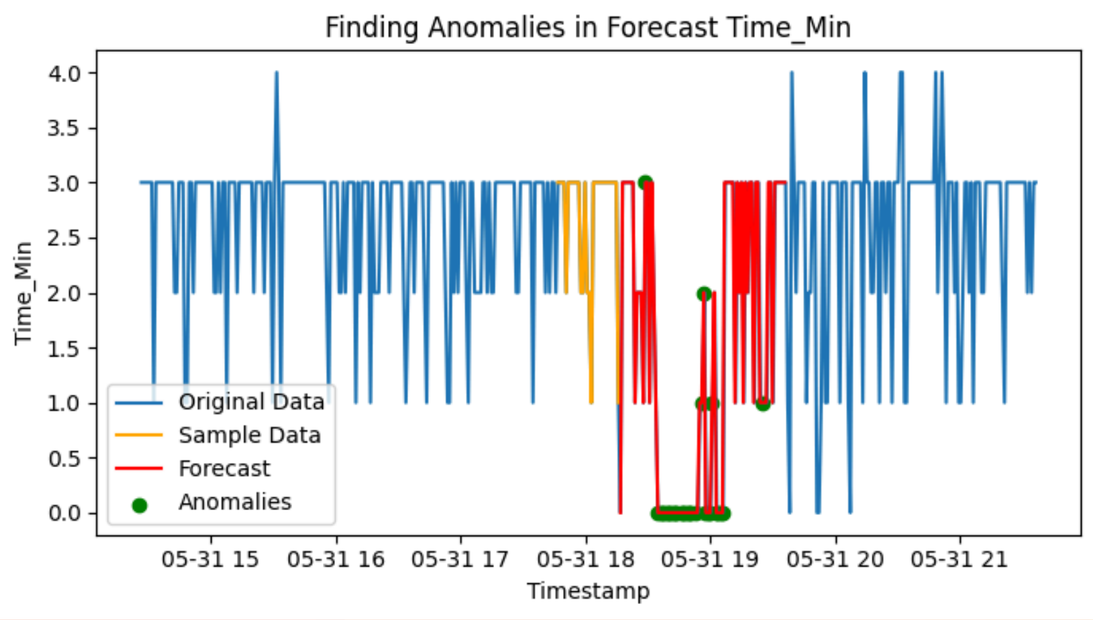


We believe a **high value in number of transactions** is the main reason for being anomaly that **could lead to the system's instability**.

The second reason would be a **high number of retries** and **high value for maximum transaction time**.

Results Evaluation





- High number of retries and mean time are also involved. Although our model had some prediction errors in Mean Time.

- For **Minimum time**, We could not perfectly understand its nature and relation between a low value with being anomalies.

- We believe there is a special meaning behind the value of 0 in minimum time.
- e.g. No complete transaction. Which most probably happens during data aggregation.

Lessons learned

Conducting these kinds of projects is time-consuming, especially if one is not experienced.

These task also need time and resources (Fast machines to compute, tuning the parameters and the retries).

However, we learned that data generation is a tricky task.

Adding some noise to data makes the forecast more stable.

For future work, we would use LLM for generating data and aslo we would use the variance of features as input

Thank You.

Thank You.

Thank You.

REPLY x LUISS