REPLY

ANOMALY DETECTION

TOMMASO AGUDIO - MICHELE AVERSA - NICOLÒ CAPPELLINI

THE CHALLENGE

01

AUGMENTATION

From a “standard” day generate more data that adds some “noise” so that you can feed this data into a Forecasting model.

02

FORECASTING

Train the Forecasting model on the synthetic data and use it to forecast 10 minutes into the future based on the previous number of transaction.

03

ANOMALY DETECTION

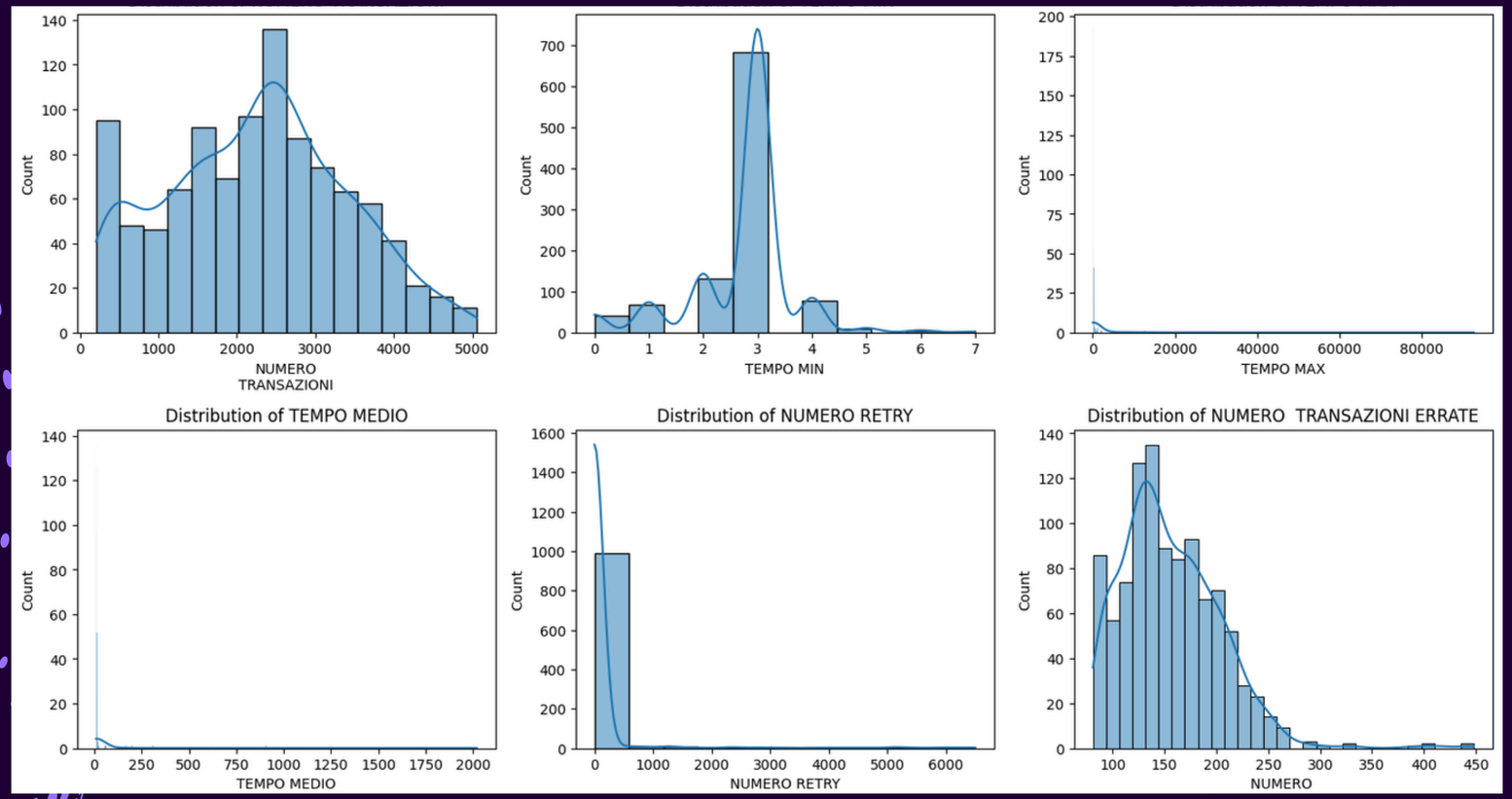
Build an Anomaly detection model to detect anomalies on our forecasted transactions.

OBJECTIVE

CREATE A “PIPELINE OF MODELS” THAT CAN LOOK INTO THE FUTURE AND ALERT BEFORE THE ANOMALIES OCCUR

DATASET

THE DATASET CAPTURES ONE DAY OF A BETTING SERVICE.
THE DAY CONSISTS OF 1018 TRANSACTION RECORDS WITH ONE ROW FOR EACH MINUTE, GOING FROM 7:00 TO 24:00.



STARTING DISTRIBUTION OF TRANSACTION OVERTIME

THE DATASET CONTAINED INFORMATIONS LIKE:

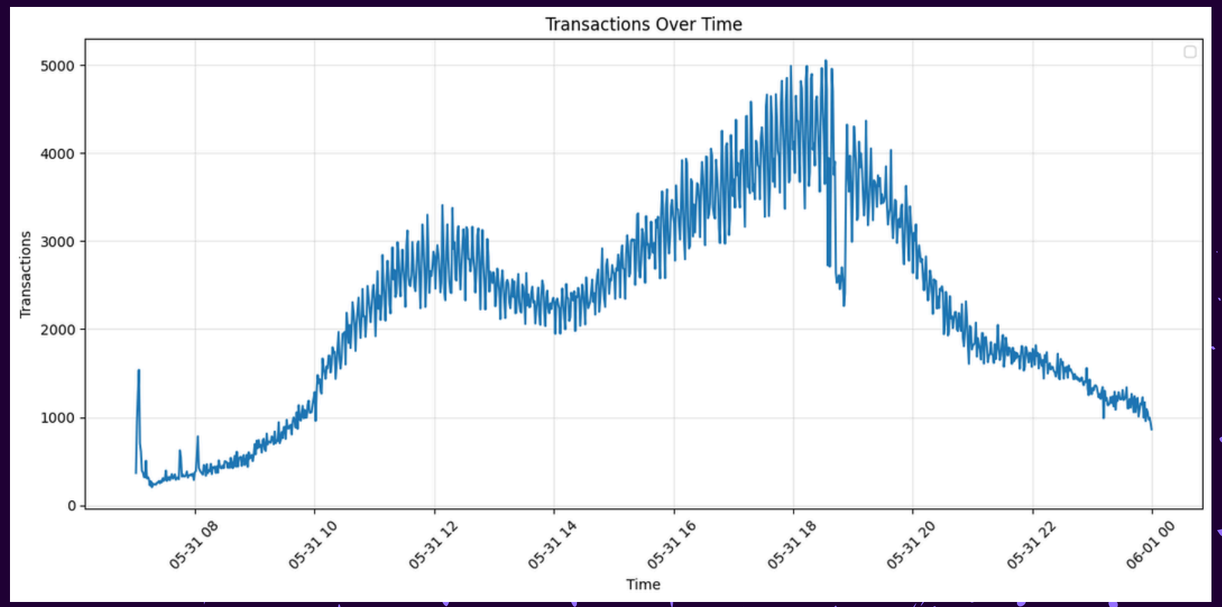
- NUMBER OF TRANSACTIONS
- NUMBER OF WRONG TRANSACTIONS
- NUMBER OF RETRY
- TIME (MEAN, MIN, MAX)

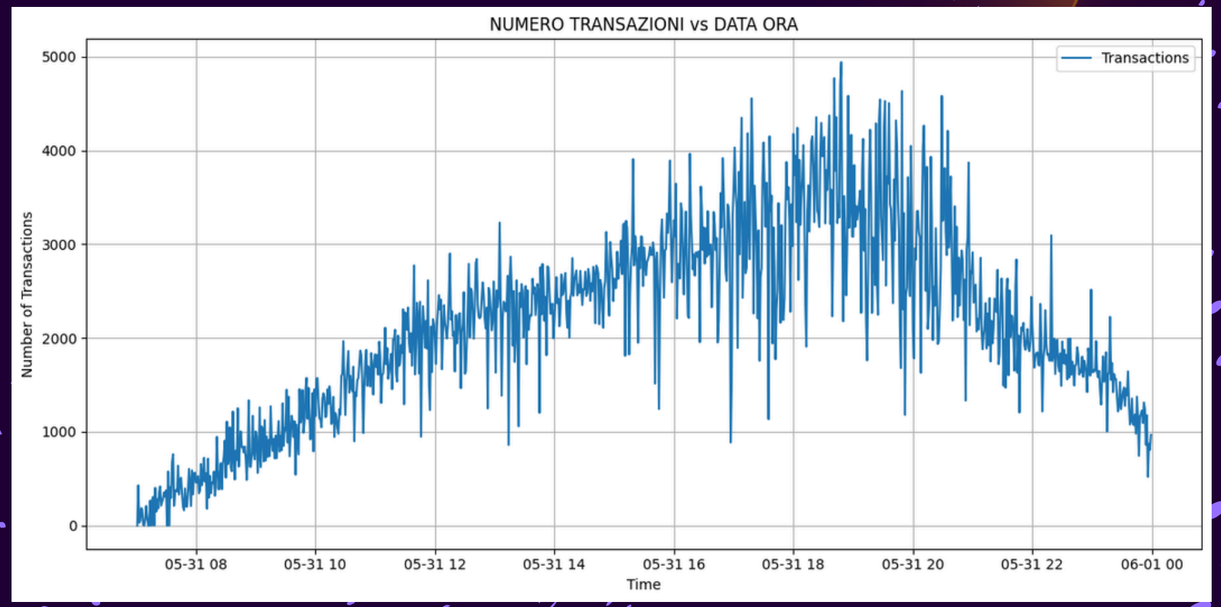
DATA AUGMENTATION

VARIATIONAL AUTOENCODER

AFTER PLAYING AROUND WITH THE LOSS FUNCTION WE WERE ABLE TO GENERATE THIS DISTRIBUTION :

DISTRIBUTION OF TRANSACTIONS OVER THE DAY

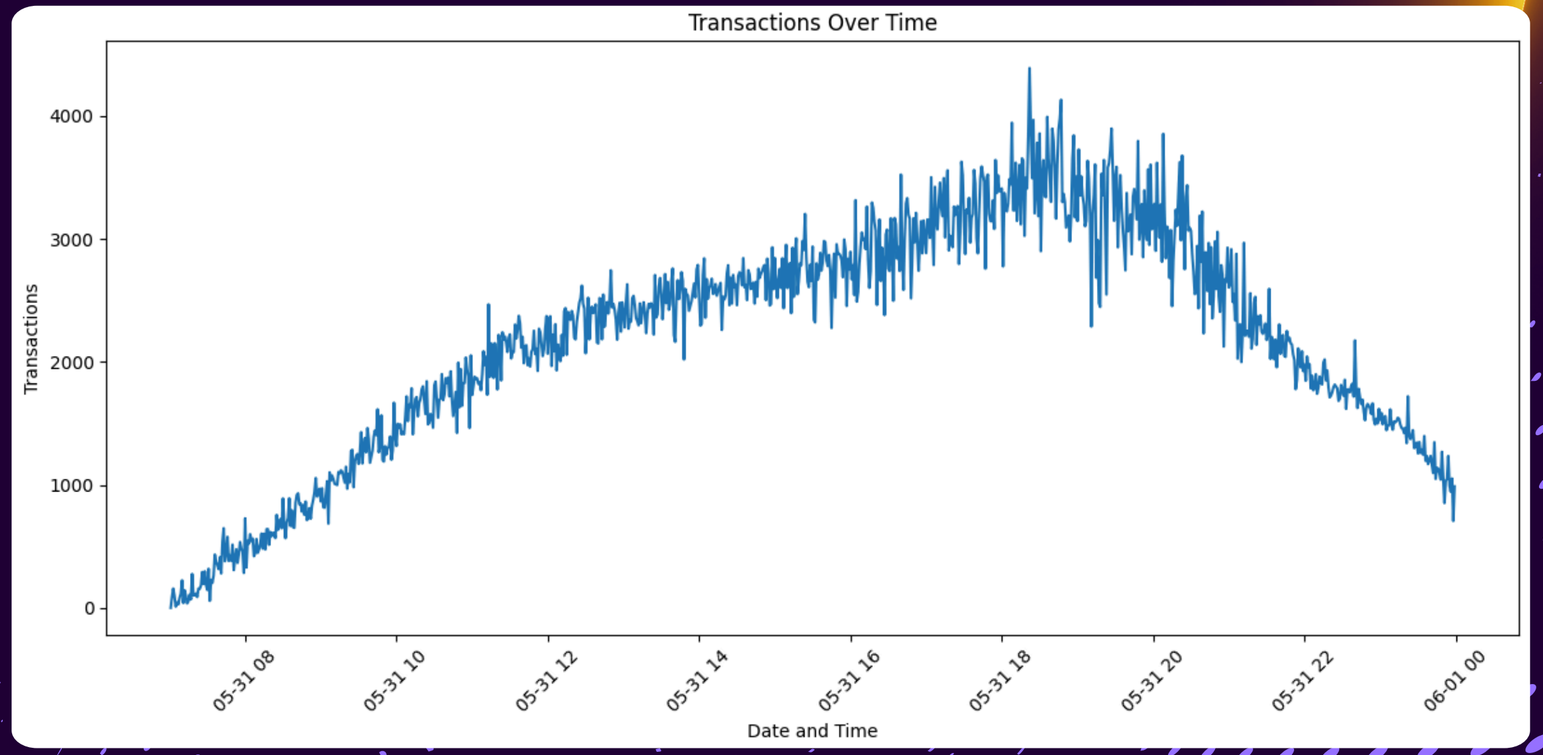




DATA AUGMENTATION

IN ORDER TO AUGMENT OUR DATASET WE USED A VARIATIONAL AUTOENCODER, PERFORMING THE FOLLOWING OPERATIONS:

- GENERATED FIVE DIFFERENT DAYS, AVERAGING EACH ROW TO OBTAIN A SINGLE DAY WITH LESS SPIKES.
- HARDCODED DATA_ORA TO OBTAIN 1 ROW FOR EACH MINUTE.



FINAL DISTRIBUTION OF TRANSACTIONS OVERTIME

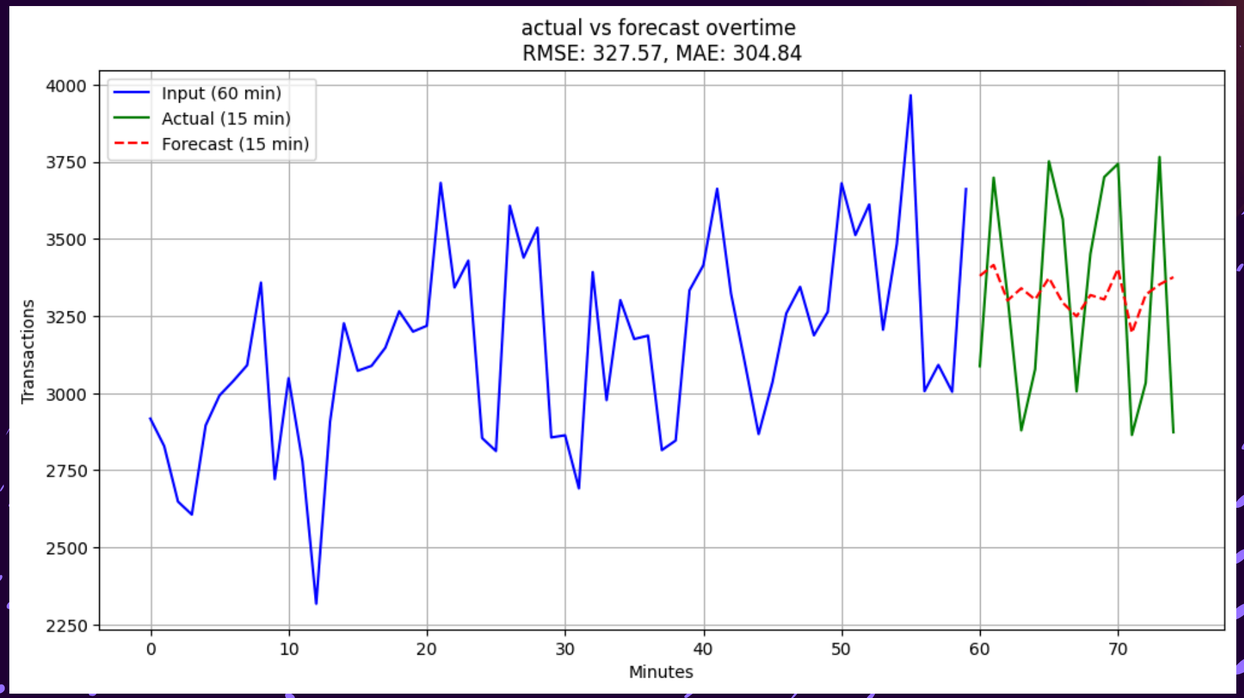
FORECASTING

TO FORECAST THE VALUE OF THE NEXT 15 MINUTES WE USED AN LSTM ON THE SYNTHETIC DATA.
WE WANTED TO HAVE A TREND THAT WAS AS SIMILAR TO THE DATA AS POSSIBLE, SO THAT DOING ANOMALY DETECTION OVER THE FORECASTED VALUE HAD SENSE.
WE FED THE MODEL WITH ONLY ONE DAY OF DATA SINCE OUR AUGMENTED DATA WAS TOO INCONSISTENT AND MORE DAYS CAUSED THE FORECAST TO BECOME EVEN WORSE.

THE VARIABLES THAT WE USED TO FORECAST ARE:

- ‘NUMERO TRANSAZIONI'
- 'NUMERO TRANSAZIONI ERRATE'
- 'NUMERO RETRY'
- 'TEMPO MAX'

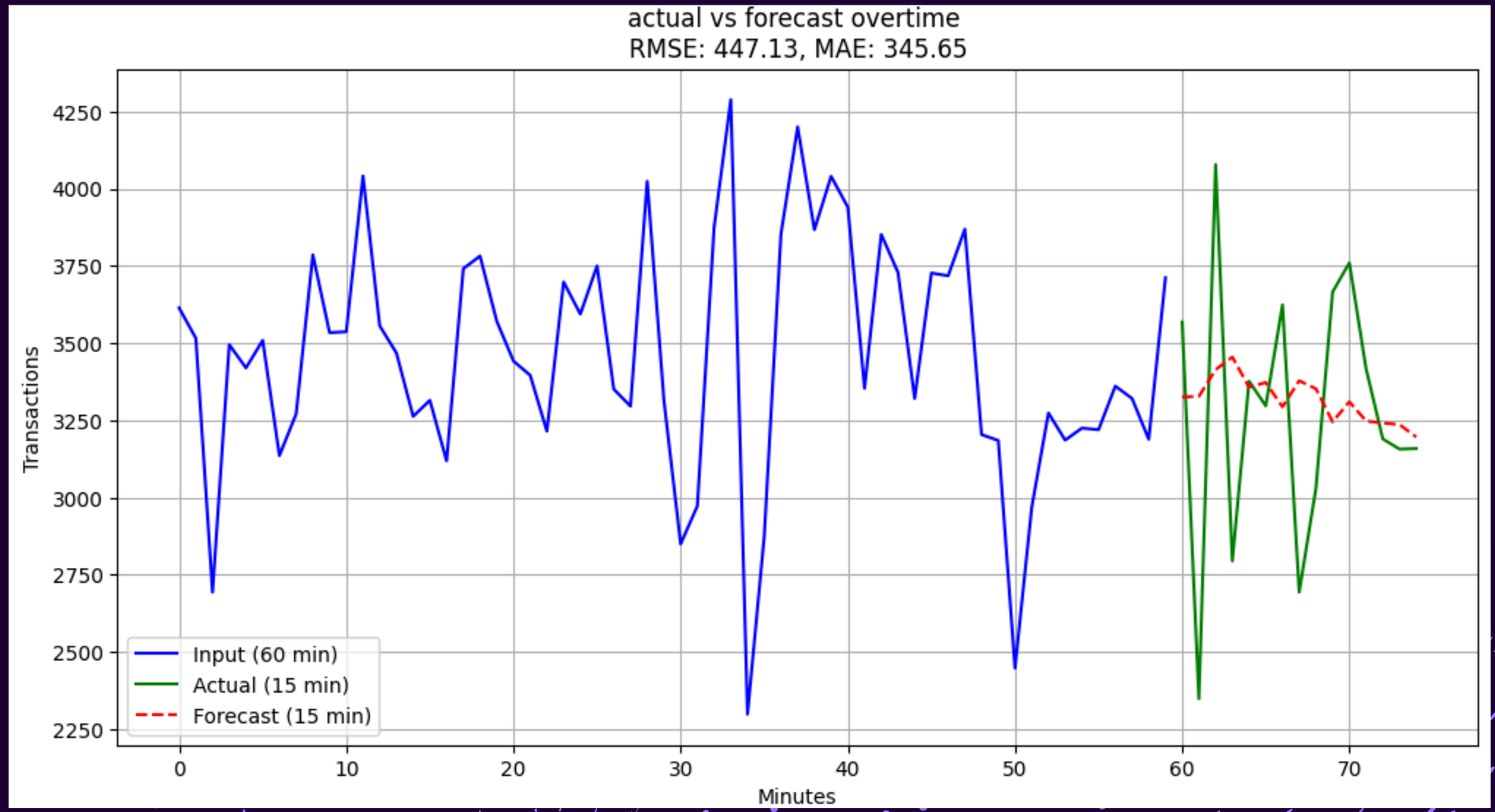
WE CHOSE THOSE BECUASE THEY WERE MEANINGFUL AND DIDN’T HAVE MULTICOLLINEARITY.
WE HAVE DONE THE FORECAST ON TWO DIFFERENT TIME WINDOW OF THE SAME DAY.



FIRST WINDOW FROM 17 TO 17.15 USING THE ENTIRE HOUR OF 16-17

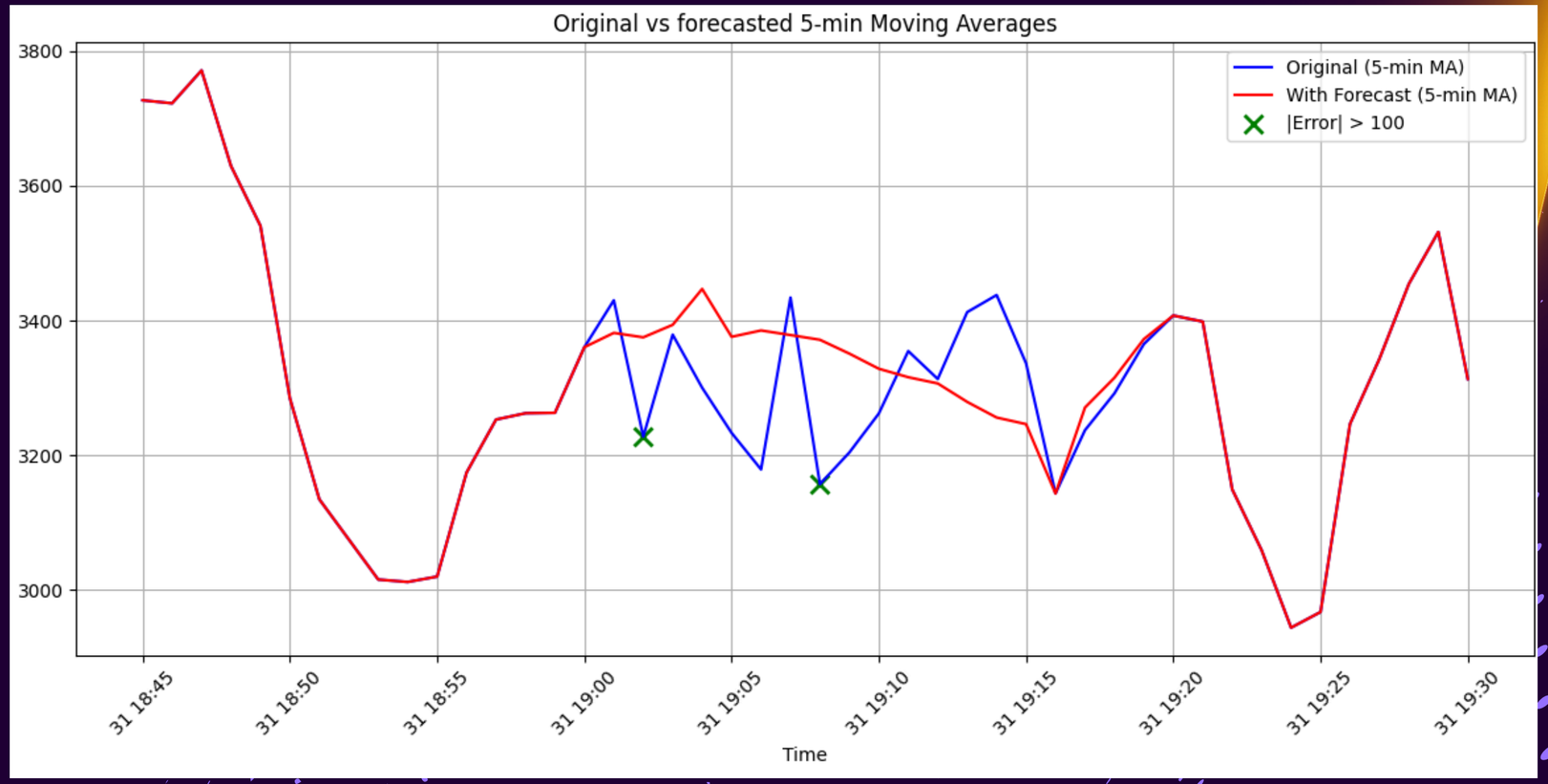
FORECASTING

THIS IS THE SECOND TIME WINDOW, USING FROM 16.45 TO 17.30



ANOMALY DETECTION

FIRST WE TRIED TO DETECT ANOMALIES BY USING THE MOVING AVERAGE.



IN THE END WE DEFINED A TRANSACTION AS ANOMALY IF:

1) LOOKING WHEN ACTUAL VS FORECAST CROSS
2) THE ABSOLUTE ERROR WAS HIGHER THAN 100

ANOMALY DETECTION

FINALLY WE DECIDED TO USE AN ISOLATION FOREST AS WE ARE INTERESTED NOT ONLY IN DETECTING ANOMALIES WITH ROLLING AVERAGES BUT ALSO IN ACTUAL VALUES.

WE PERFORMED ANOMALY DETECTION ON BOTH FORECASTED AND ACTUAL DATA, FINDING FOUR DISTINCT ANOMALIES.

WHAT CAN BE DONE NEXT?

- IMPROVE OUR AUTOENCODER TO BETTER CAPTURE THE INITIAL DISTRIBUTION, SO THAT WE CAN FEED MULTIPLE DAYS TO THE LSTM AND GET A MORE ACCURATE RESULTS.
- PREDICT ALSO OTHER FEATURES LIKE NUMBER OF RETRY SO THAT WE CAN SPOT ANOMALIES ALSO BASED ON THAT.
- SINCE WE LOST THE MAJORITY OF OUR TIME TRYING IN THE AUGMENTATION PROCESS, WE THINK THAT WE COULD'VE TRIED TO IMPLEMENT DIFFERENT SOLUTIONS/MODELS FOR OUR FORECASTING AND ANOMALY DETECTION.

THANK YOU!