Causal inference of multivariate server performance logs

Kewei Zhang

2024/02/20

Introduction

- Data monitoring and visualization
  - Datadog
- Causal Inference
  - Data Imputation
  - Stationary Transformation
  - Correlation Analysis
  - Causal and Inference

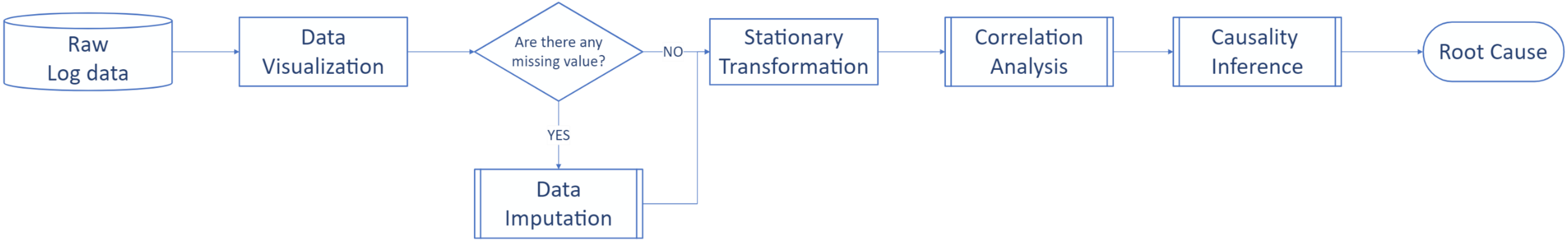
Data Monitoring and Visualization

Datadog

- Logs Dashboard | Datadog (datadoghq.com)

Causal Inference

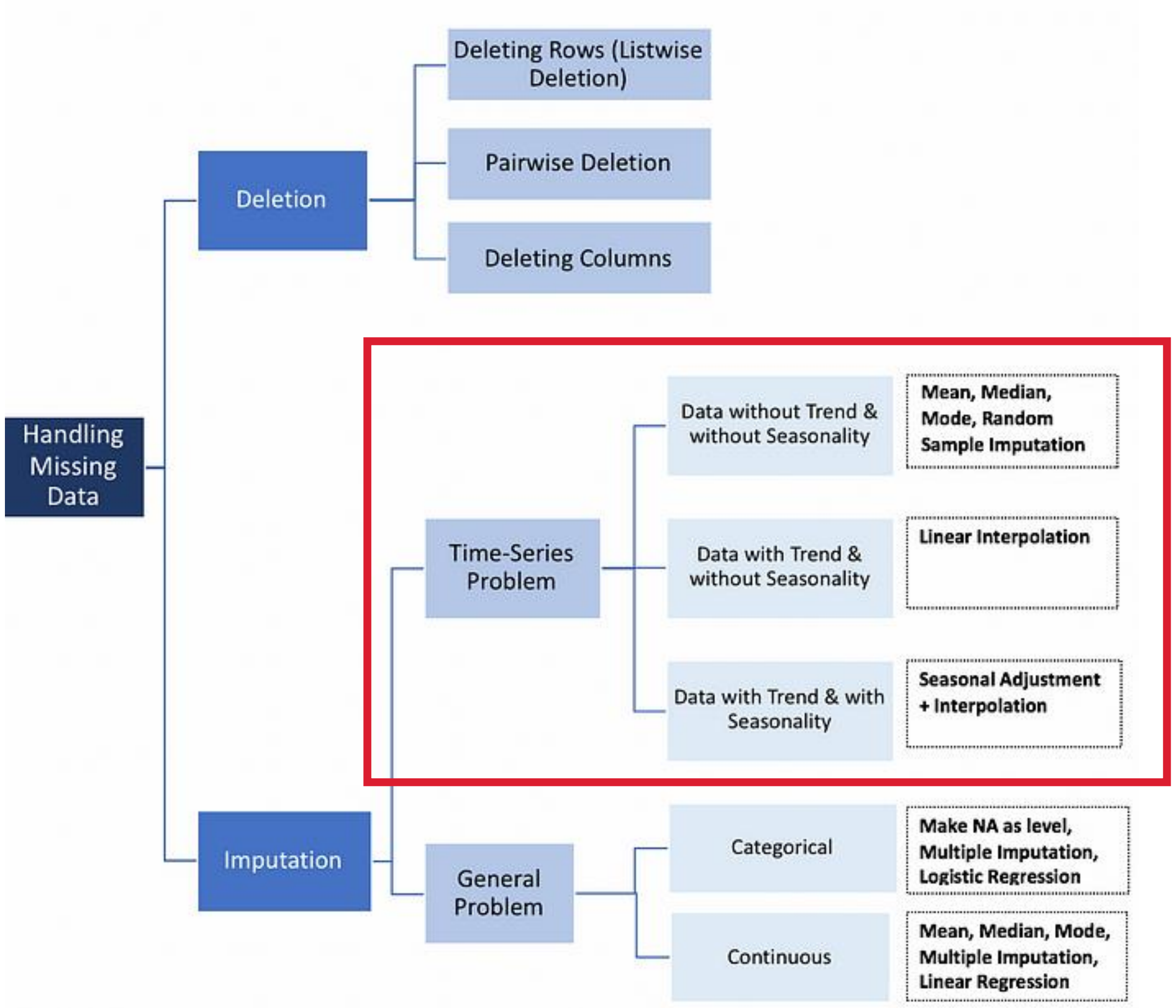
Establish whether and how changes in one variable cause changes in another



Example: Processor Time – Performance Counter (Instance: RapidResponse)

Data Imputation

Imputation for missing value to generate complete datasets


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

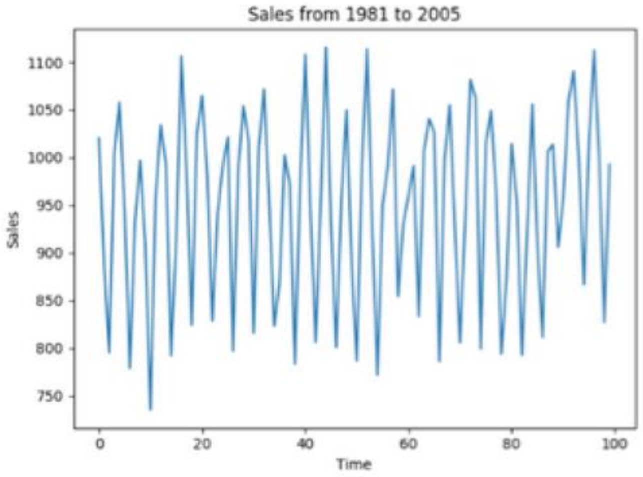
INTERNAL USE

Access limited to internal use only

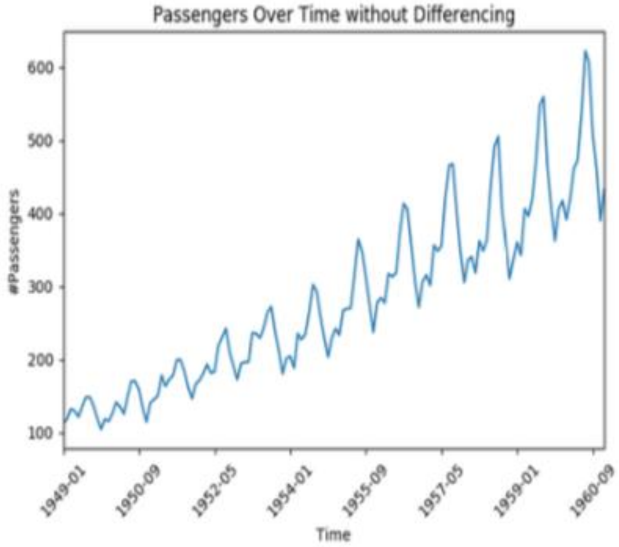
Stationary Transformation

Why stationary is important: most time series models assume that each point is independent of one another.

Stationary data: mean and variance do not vary across time



Stationary data



Non-stationary data

INTERNAL USE

Access limited to internal use only

Correlation Analysis

Evaluate strength and direction of the linear relationship between two variables

Question: We can only retain a month's worth of data, insufficient for any computation

Solution: Employ a variety of mathematical computations

Method:

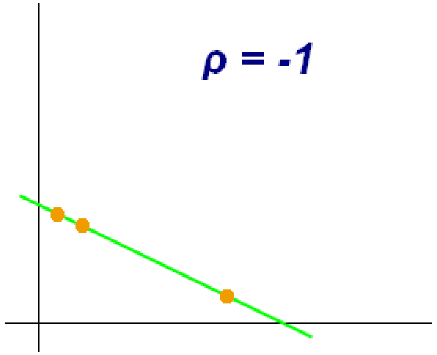
1. Pearson correlation coefficient
2. Spearman correlation coefficient
3. Kendall correlation coefficient

Correlation Analysis

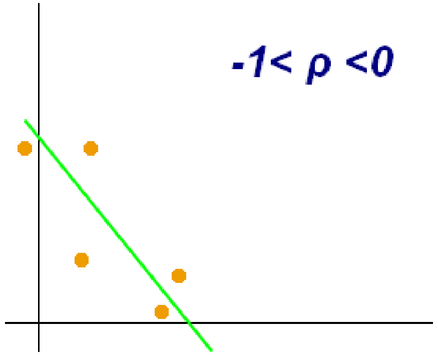
Pearson Correlation Coefficient: [-1,1]

Coefficient = 1 or -1: perfect linear relationship

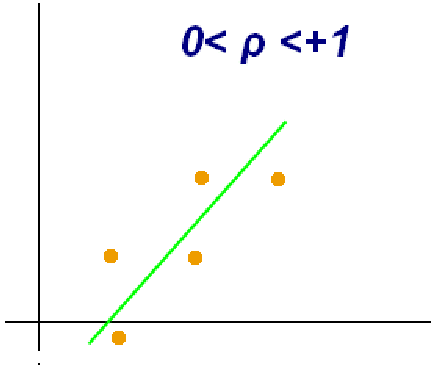
Coefficient = 0 : no linear relationship, could have a non-linear or more complex relationship



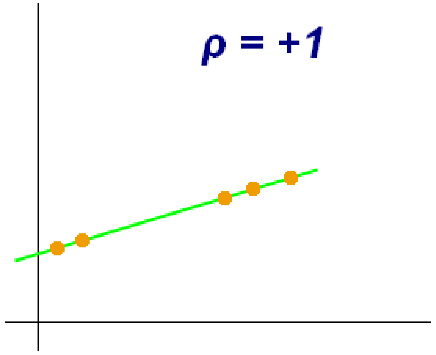
perfect negative correlation



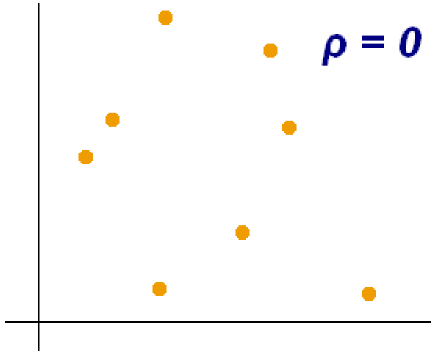
negative correlation



positive correlation

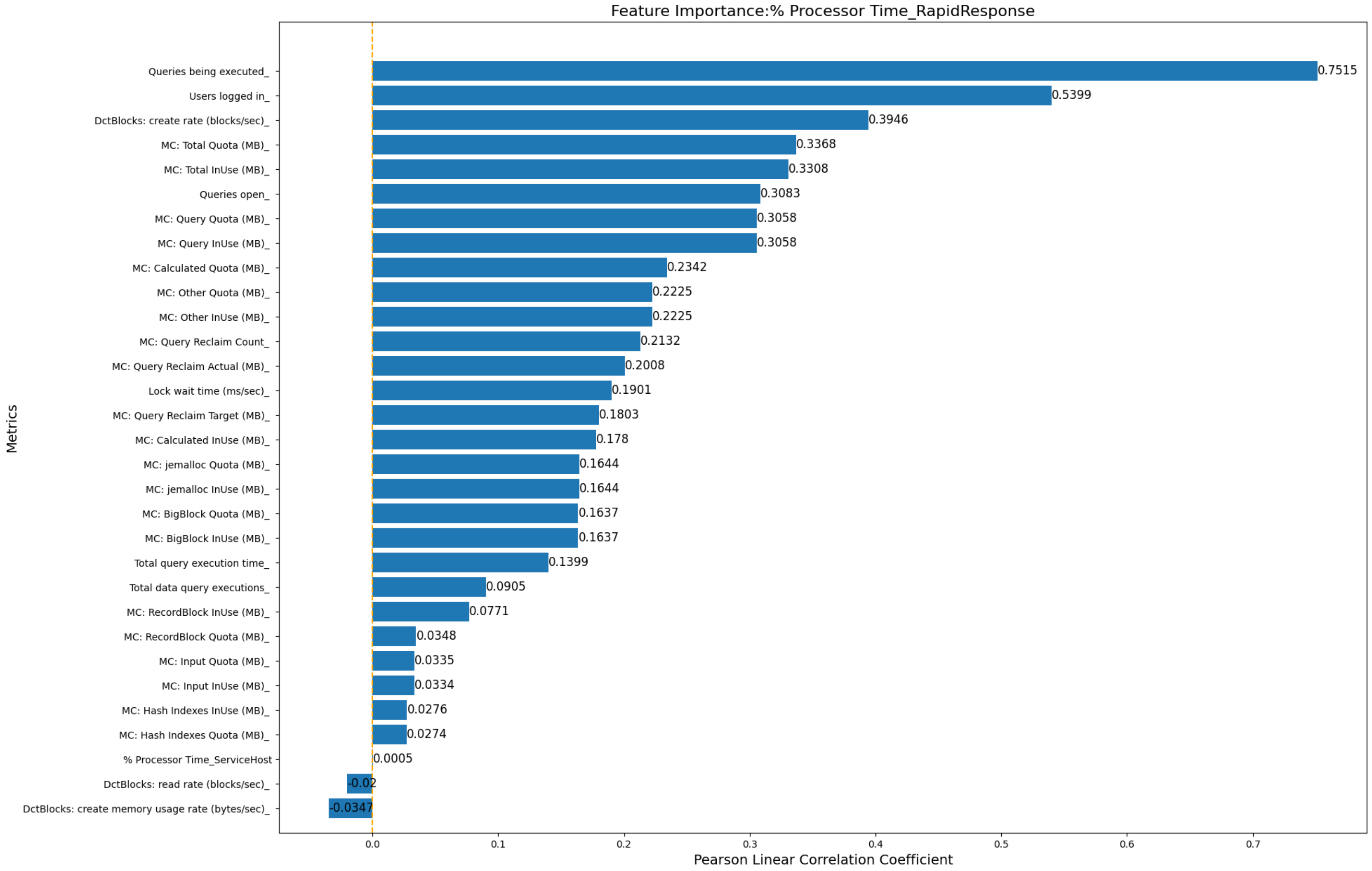


perfect positive correlation



no liner correlation

Pearson Correlation Coefficient

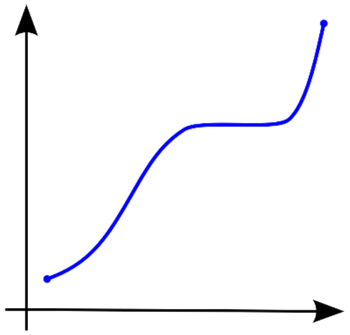


Correlation Analysis

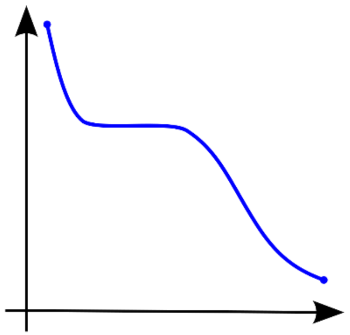
Spearman Correlation Coefficient

coefficient = 1/-1 : perfect positive/negative monotonic relationship

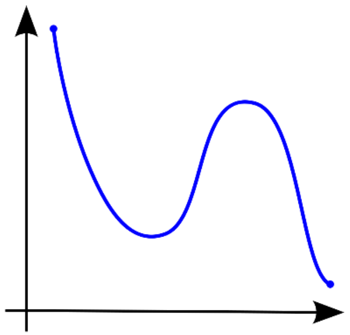
Coefficient = 0 : no monotonic relationship



Monotonic non-decreasing

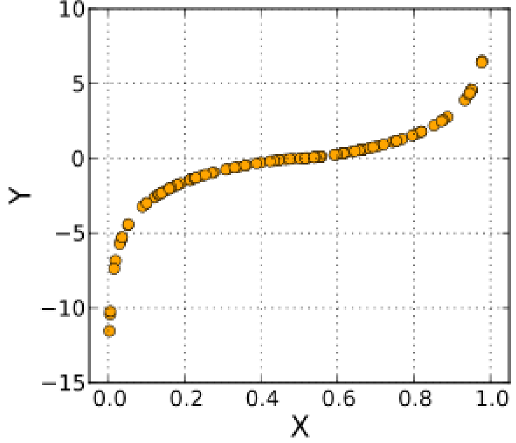


Monotonic non-increasing

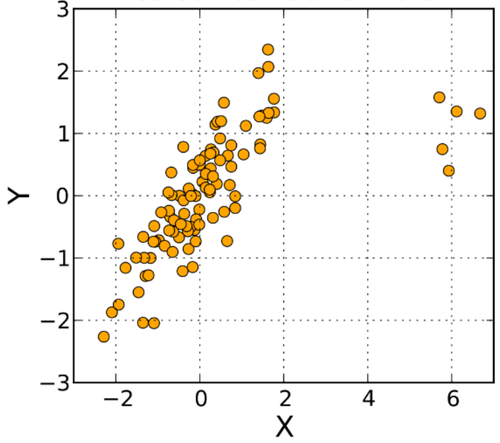


Not monotonic

Spearman correlation=1
Pearson correlation=0.88



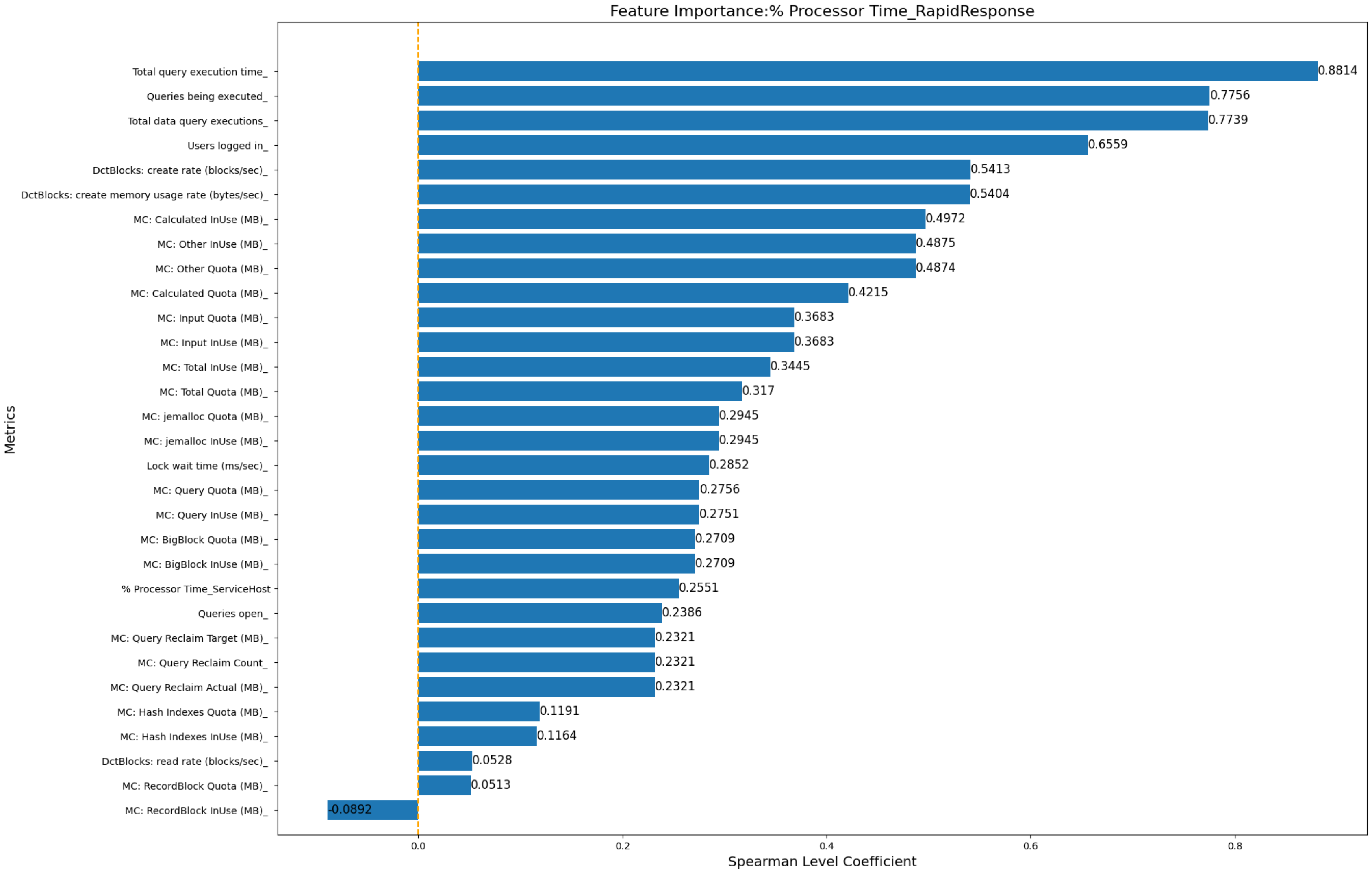
Spearman correlation=0.84
Pearson correlation=0.67



INTERNAL USE

Access limited to internal use only

Spearman Correlation Coefficient



INTERNAL USE

Access limited to internal use only

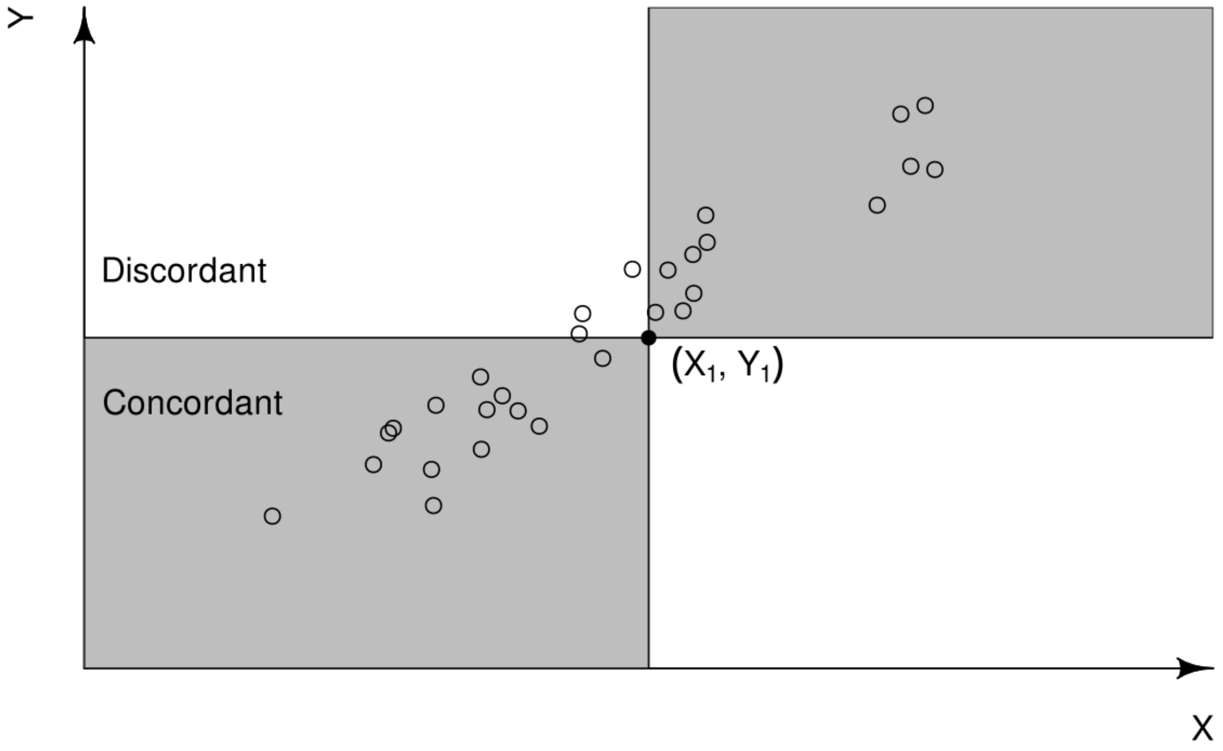
Kendall Correlation Coefficient

coefficient = 1/-1: ranks of corresponding values within each data sample always same

Coefficient = 0 : no association between the ranks of the values

Ex: (xi,yi), (xj,yj), i < j:

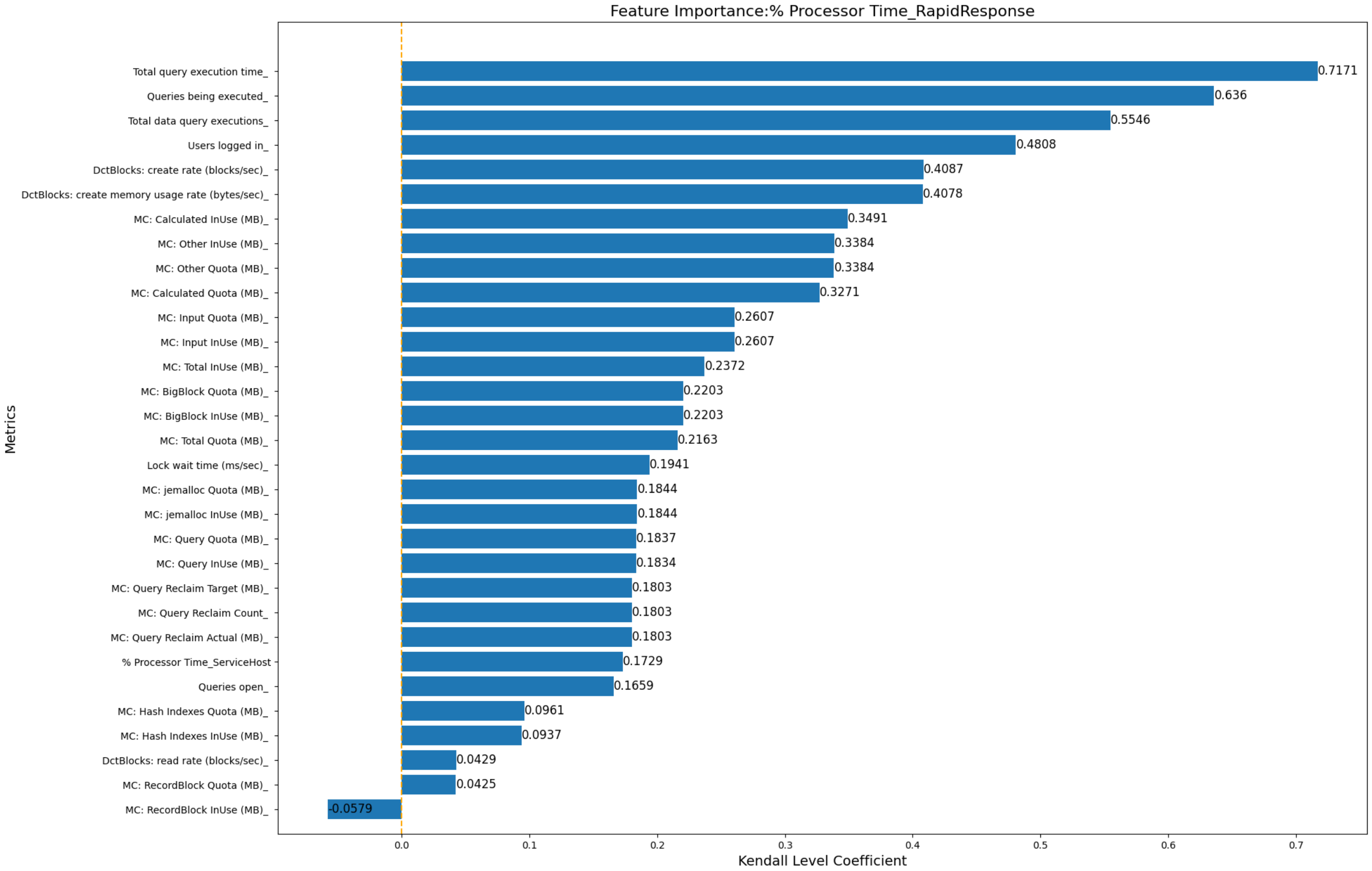
Either both xi > xj and yi > yj or xi < xj, yi < yj



INTERNAL USE

Access limited to internal use only

Kendall Correlation Coefficient

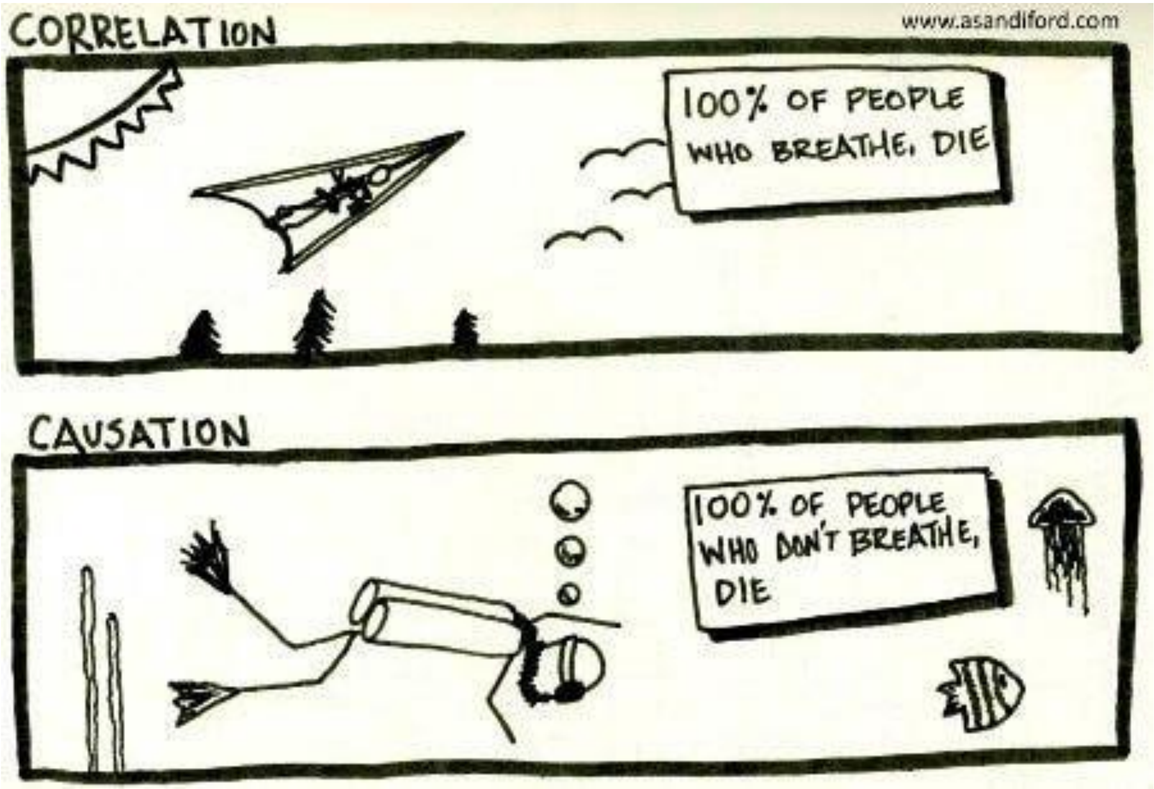


Correlation Coefficient

Top5 Results:

Pearson Correlation:	**Spearman Correlation:**	**Kendall Correlation:**
Queries being executed	Total query execution time	Total query execution time
Users logged in	Queries being executed	Queries being executed
DctBlocks: create rate (blocks/sec)	Total data query execution	Total data query execution
MC: Total Quota (MB)	Users logged in	Users logged in
MC: Total Inuse (MB)	DctBlocks: create rate (blocks/sec)	DctBlocks: create rate (blocks/sec)

Correlation VS Causation


CORRELATION
www.asandiford.com
100% OF PEOPLE WHO BREATHE, DIE
CAUSATION
100% OF PEOPLE WHO DON'T BREATHE, DIE

Transfer Entropy

Features of time series: max, min, mean, median, variance,…

Series A : 1, 2, 1, 2, 1, 2, 1, …

Series B : 1, 1, 2, 1, 2, 2, 1, …

Same mean, same variance, same median, different entropy

Larger the entropy, more chaotic the system

Transfer entropy: transfer of information

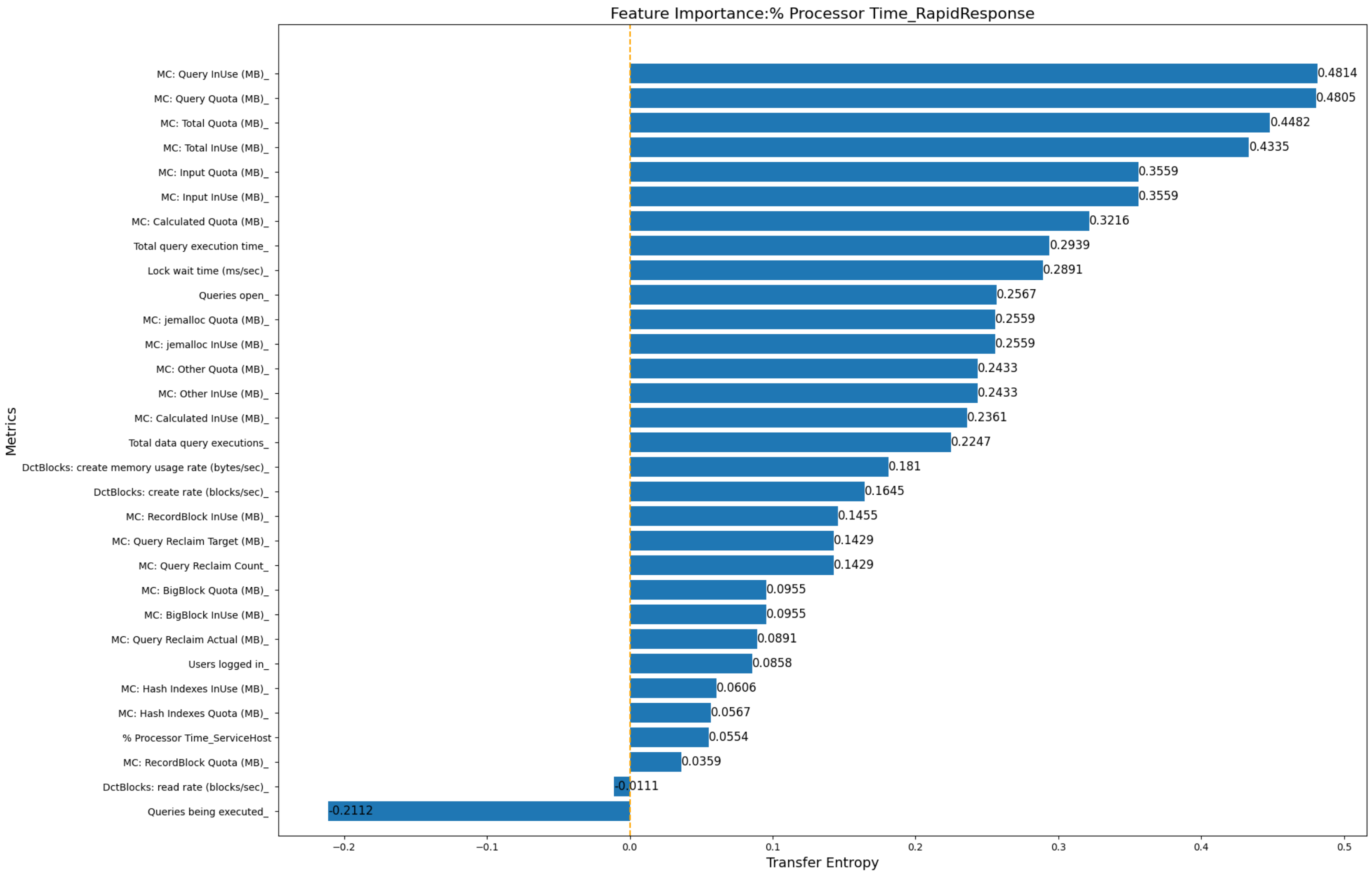
Causal Inference

Transfer Entropy

Symbol: direction of transfer

+:Y -> X

-: X -> Y



Causal Inference

Causal Inference

Neural Network: LSTM(Long-short term memory)

Step 1: Forecasting

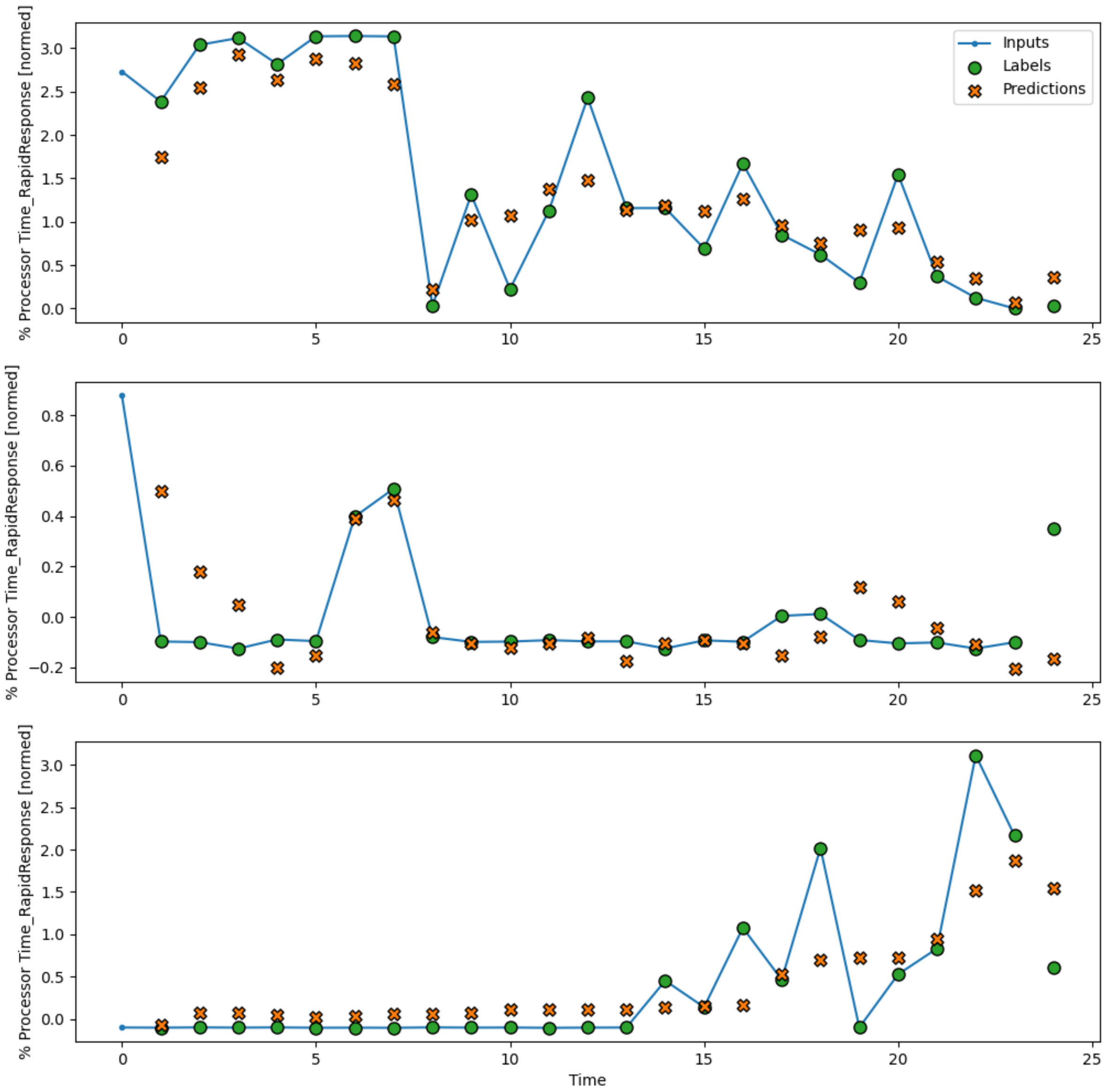
Step 2: Evaluating the prediction model

Step 3: Calculating feature Importance

Causal Inference

Step1: Forecasting

Selected 3 subsections from the entire dataset as examples



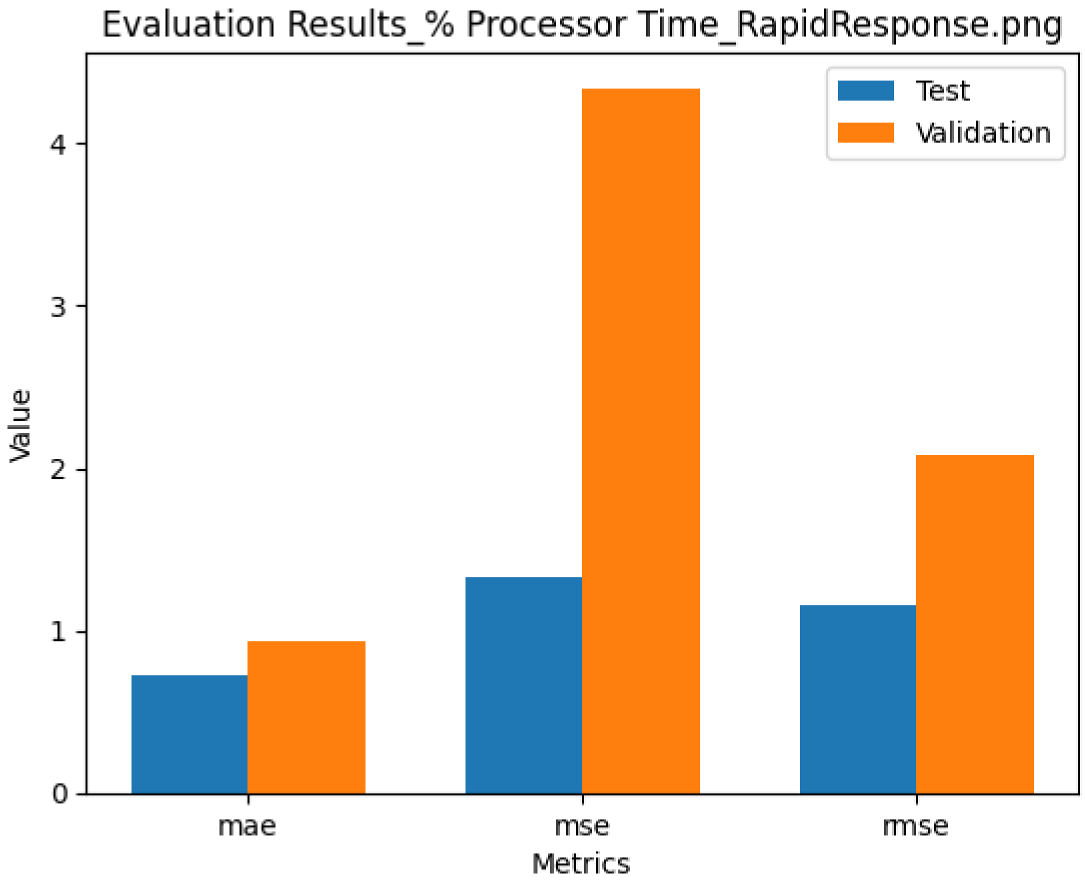
Causal Inference

Step 2: Evaluating the prediction model

MAE: mean absolute error

MSE: mean squared error

RMSE: root mean squared error



INTERNAL USE

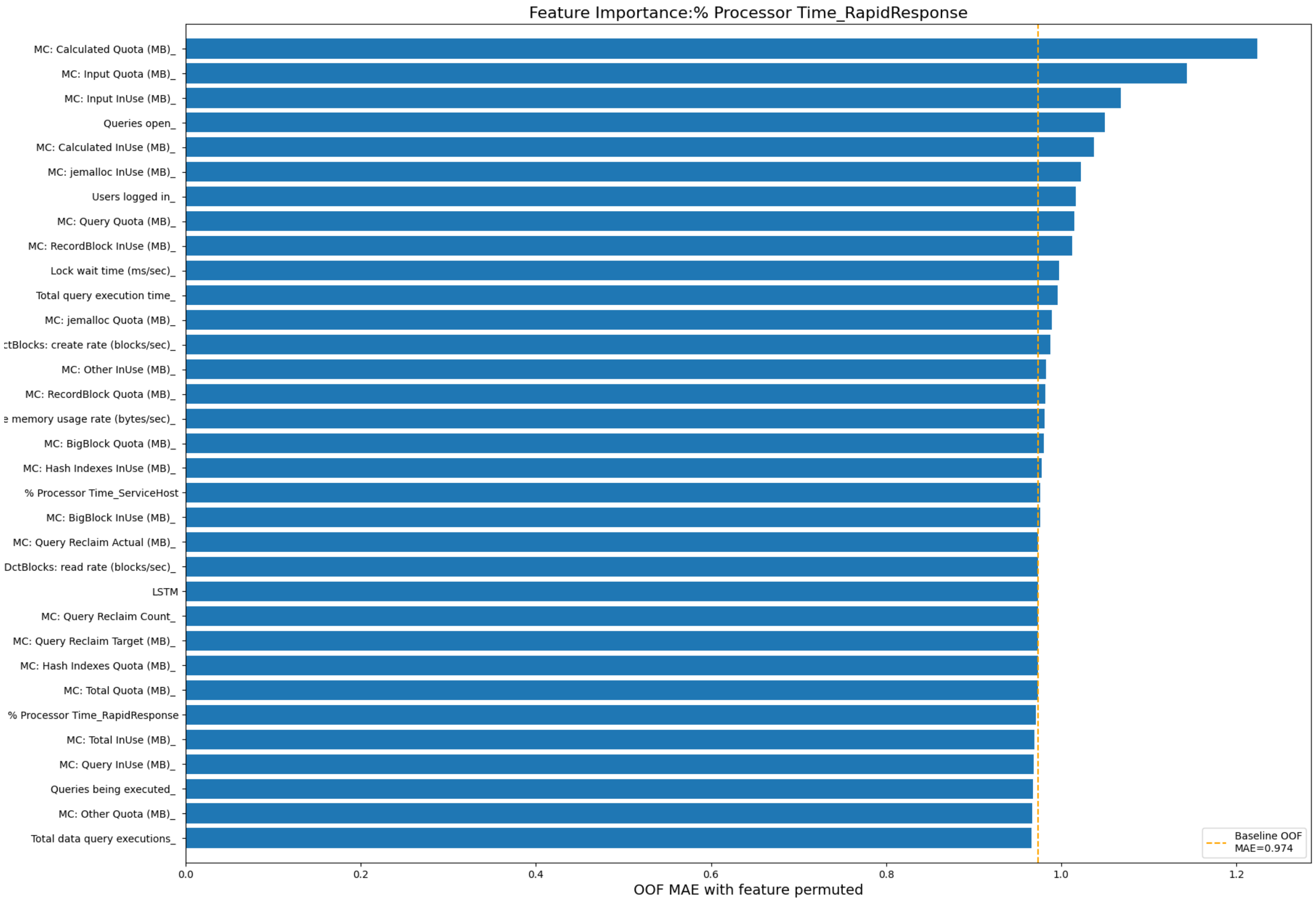
Access limited to internal use only

Causal Inference

Causal Inference

Step 3: Calculating feature Importance

OOF: Out-of-Fold, dividing a dataset into subsets, known as “folds.” Each fold is then used as a validation set once while the remaining folds are utilized for training.



INTERNAL USE

Access limited to internal use only

Conclusion

- Most Correlated metrics with Processor Time:
  - Total query execution time
  - Queries being executed
  - Total data query execution
  - users logged in
  - DctBlocks: create rate (blocks/sec)

- Metrics with the largest Transfer Entropy:
  - MC metrics
  - Total query execution time
  - Lock wait time
  - Queries open
  - Total data query execution

- Metrics with the biggest influence:
  - MC metrics
  - Queries open
  - Users logged in
  - Lock wait time
  - Total query execution time

Conclusion

Possible Reason of error:

- Insufficient data
- Inappropriate model selection
  - Not suitable
  - Overfitting, too complex
- Feature engineering issues
  - Incorrect stationary transformation