Neuralwave Hackathon Causal Analysis Bosch Project

Oliver Tryding, Christian Berchtold, Nathan Margni

October 27, 2024

link to video: https://www.loom.com/share/2d51337bf0a64bcab2ff7759fe9bcb00?sid=a23df8d7-90b5-45aa-a68d-8217274b885d

link to github: https://github.com/Neural-Wave/project-OgD3A

1 Introduction

The manufacturing plant experienced an unexpected decline in product quality, resulting in increased scrap rates. Factory workers requested an analysis to identify the root causes of this quality degradation. Two datasets were provided, representing measurements from 2,500 parts before and after the quality decrease.

2 Data Description

2.1 Datasets

- **Low_scrap.csv**: Measurements before the decrease in quality.
- **High_scrap.csv**: Measurements after the decrease in quality.

Each dataset contains rows representing individual parts and columns representing specific physical quantities measured at different stations in the production process. The target parameter of interest is `Station5_mp_85`, where higher values correlate with reduced quality and increased scrap likelihood.

3 Methodology

To uncover the causal relationships between the physical quantities and identify the root causes, we employed three causal discovery methods:

3.1 Notears

Notears is a gradient-based optimization algorithm that learns a directed acyclic graph (DAG) by formulating the structure learning problem as a continuous optimization problem.

3.2 PC Algorithm

The PC Algorithm is a constraint-based method that uses conditional independence tests to infer the causal structure among variables.

3.3 Bayesian Networks

Bayesian Networks represent the joint probability distribution of variables using a DAG, where edges indicate probabilistic dependencies.

4 Results

To understand the linear relationships between each physical quantity and the target parameter `Station5_mp_85`, we calculated the Pearson correlation coefficients. Table 1 lists the physical quantities and their correlation strengths with `Station5_mp_85`.

Table 1: Correlation of Physical Quantities with `Station5_mp_85`

Physical Quantity	**Correlation with `Station5_mp_85`**
⋮	⋮
`mp_28`	0.40
`mp_13`	0.44
`mp_85`	1.00

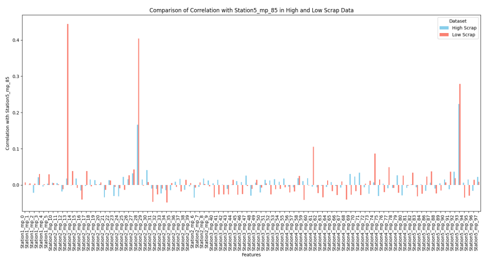

Figure 1: Correlation between the components and `Station5_mp_85` in the low-scrap and high-scrap datasets.

From figure 1 we observe that `mp_13` and `mp_28` have the strongest correlations with `Station5_mp_85` in both datasets. Notably, the correlation of `mp_13` with `Station5_mp_85` decreases drastically from the High Scrap dataset to the Low Scrap dataset. This significant change suggests that `mp_13` is a key factor influencing the increase in `Station5_mp_85` values in the High Scrap scenario, indicating it as a root cause of the quality issues.

4.1 Root Cause Analysis

All three methods consistently identified `Station2_mp_13` as a root cause for the increase in `Station5_mp_85` values. Notears additionally identified `Station2_mp_28` as a significant contributing factor.

The influence number for each ancestor variable quantifies the magnitude of its effect on the expected value of the target variable, `Station5_mp_85`, within our Bayesian Network model. Specifically, it measures how much the expected value of `Station5_mp_85` changes when the ancestor variable transitions from its low state (minimum discretized value) to its high state (maximum discretized value). A higher influence number indicates a stronger impact on the target variable, allowing us to identify and rank potential root causes affecting product quality based on how significantly they influence `Station5_mp_85`. Since we only find a single root cause, we report its value of `Station5_mp_13` as 1.11.

4.2 Adjacency Matrix Comparison

Comparing the adjacency matrices derived from each method to the ground truth revealed that the Bayesian Network provided the most accurate representation. However, this higher accuracy may be attributed to the Bayesian Network's sparser adjacency matrix, containing fewer entries.

4.3 Visualization

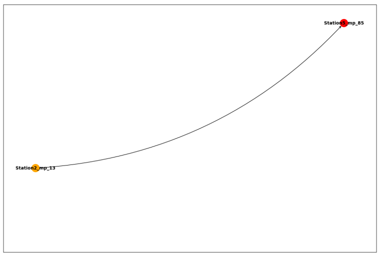

Figure 2: Causal graph illustrating the relationships between physical quantities.

The graph in figure 2 illustrates the causal relationship of the root causes as produced by the Bayesian model. As can be seen, only `Station2_mp_13` significantly effects `Station5_mp_85`.

5 Conclusion

The analysis successfully identified `mp_13` as the primary root cause of the decreased product quality, with Notears also pointing to `mp_28` as a secondary cause. The Bayesian Network provided the most accurate adjacency matrix when compared to the ground truth, although its sparsity may have influenced this outcome.