Auditing & Forecasting of Wafer

IE3100R System Design Project AY 2025/26

Department of Industrial & Systems Engineering

Group 3: Nur Aliyyah Ayub Mansul, Yang Xueqi, Mio Zheng An, Ho Yuheng Tristan, Lee Hong Ee, Gabriel

Course Advisor: Qin Hanzhang **Industry Supervisor:** Lim Her Phay

1. Problem Description

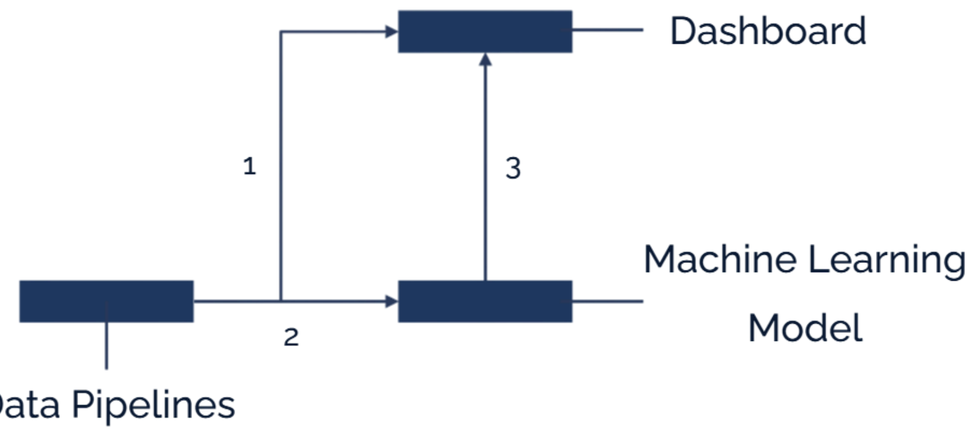
The current system lacks a clear and standardised process to audit forecasting performance across production stations. There is limited visibility in comparing MOR (forecast) and POR (actual) outputs, making it difficult to assess forecast accuracy. Additionally, auditing is manual and time-consuming, and abnormal forecast behaviour is not easily detected from raw data.

2. Project Aim

The aim is to design a data-driven dashboard that enables efficient evaluation of forecasting performance, improves visibility of MOR and POR trends, and supports better capacity planning decisions.

3. Solution Overview & System Architecture

To address the issues identified, our team developed a three-stage approach to transform raw production data into actionable insights.



1st Stage: Data Piplines

- Raw data is extracted, cleaned, and structured
- Ensures data quality and consistency for analysis

2nd Stage: Machine Learning

- Forecasting evaluation and machine learning models applied
- Identifies patterns, errors, and performance gaps

3rd Stage: Dashboard

- Insights presented through an interactive dashboard
- Enables monitoring of MOR vs POR trends and anomaly detection

4. 1st Stage: Data Pipelines - Test Wafer Data Set

Raw Data: Raw data was first extracted from the source system. As the original dataset contained inconsistencies, pre-processing was required before further evaluation could be carried out.

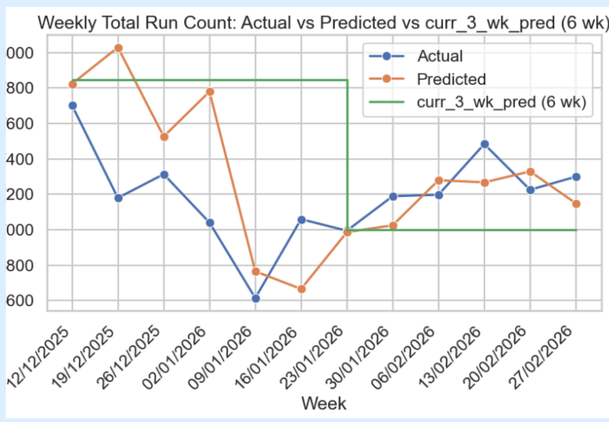
Data Cleaning: The extracted data was cleaned and transformed by removing inconsistencies, handling missing values ensuring that the dataset was accurate, consistent, and suitable for downstream analysis.

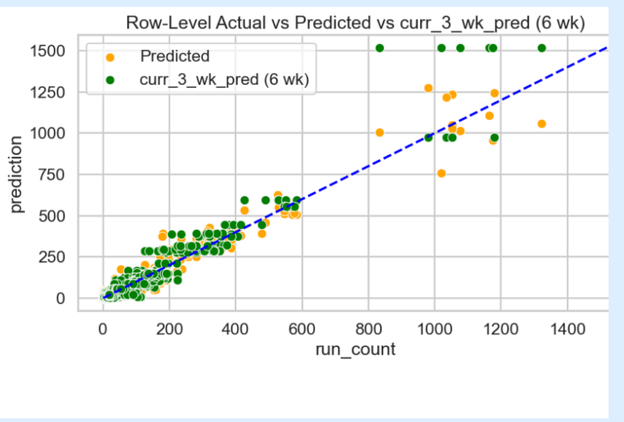
Structured Data: The cleaned dataset was then structured into an analysis-ready format, allowing efficient comparison of MOR and POR outputs and supporting subsequent modeling, visualization, and decision-making processes.

Software Used: Tableau, Python

5. 2nd Stage: Machine Learning - Forecasting of Wafer

Machine Learning Forecasting: Enhances accuracy, efficiency, and agility by reducing planning errors, adapting quickly to changes in run count usage, and improving capacity utilisation and resource allocation. It also provides a scalable, future-ready solution that enables the adoption of advanced analytics and staying aligned with evolving AI capabilities.



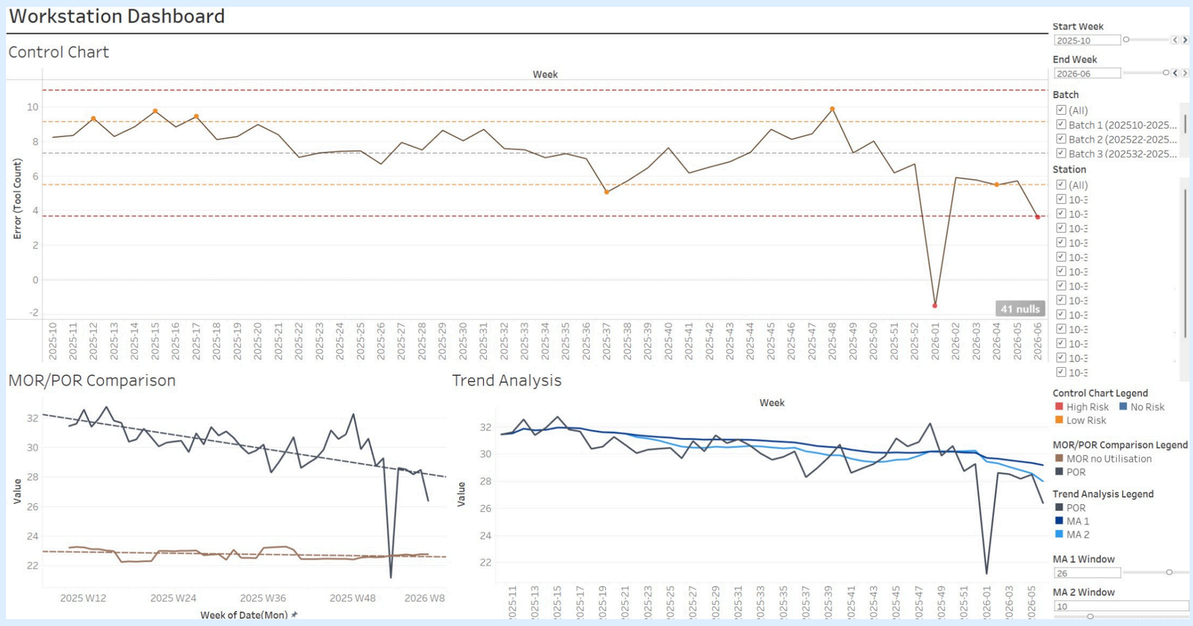


Feature Engineering: From the base dataset (Date, Workstation, Loading), we engineered decomposed workstation components and lag-based features, enabling the model to capture structural patterns and time-series dependencies for improved predictive performance.

Model: We developed a supervised XGBoost regression model incorporating time-based features, with tuned hyperparameters for optimal performance. The model achieved a **~5% reduction** in MAE compared to the baseline approach.

6. 3rd Stage: Dashboard - Visual Comparison

The dashboard provides clear insights into forecast accuracy and gaps, improves audit efficiency and decision support, and reduces manual checks, enabling faster actions.



Control Charts for Anomaly Detection: Identify unusual error patterns and outliers efficiently.

Comparison of MOR and POR Trends: Track alignment between forecast and actual performance.

Trend Analysis: Highlights gradual changes and potential risks across the work weeks and stations.

7. Future Improvements

Automated Alert System: Outliers identified through control charts can trigger automated alerts, allowing the team to respond promptly and adjust operations accordingly.

Automated Interpretation of Dashboard Insights: The system can automatically highlight important trends and anomalies, helping the team to generate insights for decision-making.

Automated ML Model Retraining Pipeline: New data can be automatically ingested and used to retrain the ML model on a scheduled basis, ensuring predictions remain accurate and adapt to evolving time-series patterns.