

Technical Note: PajakSight Intelligence Engine

This documents outline technical notes as part of David OnlinePajak technical test

1. Assumptions

- If an entity is missing feature data for a specific month, their operational profile remains consistent with their most recent historical month. However such information is flagged in the output
- The MVP assumes a standard baseline VAT rate of 11%
- The notebook and frameworks assume the first 8 months (Jan-Aug 2025) as the company's asset and that OnlinePajak starting to sell the product at Sep 2025

2. Methodology

- Simulated deployment: 8 months is used as pre-sale database and the output is tested on rest of 4 months
- Steps:
 1. The notebook check 3 scores: Vendor operational efficiency, Vendor financial efficiency, and Audit Risk
 2. Scores are then computed against itself and/or against peers:
 - a. Vendor operational efficiency: detecting slow or inefficient vendor operationally
 - i. Against itself: 2x std deviation from company's mean after data is normalized
 - ii. Against peer: 2x std deviation from peers' mean after data is normalized
 - b. Vendor financial efficiency: detecting potential phantom vendor/frauds
 - i. Against itself: -
 - ii. Against peer: Local Isolation Forest
 - c. Audit Risk: whether data is consistent with benchmark VAT
 - i. Against itself: deviation more than 10%
 - ii. Against peer: -
 3. The 3 scores are then feeded into LLM to summarize and generate report

3. Robustness Guardrails

- Safeguarding against empty data and cold-start by preprocessing the data
- Zero hallucination by using the LLM strictly for summarizing facts