Concise Project Report

Project Title: Intelligent SAP Financial Integrity Monitor (Proof-of-Concept)

Subtitle: Enhancing Financial Controls through Hybrid AI/ML Anomaly Detection on SAP FI/CO Data

Date: April 30 2025
Dataset Origin: https://www.kaggle.com/datasets/sunithasiva/sap-dataset

1. Executive Summary:

This report details the development of the **"Intelligent SAP Financial Integrity Monitor,"** a proof-of-concept (POC) system designed to proactively detect financial anomalies within core SAP FI/CO data (BKPF, FAGLFLEXA). Recognizing that standard SAP checks often miss subtle irregularities in high-volume data, this project leveraged 17+ years of SAP expertise combined with applied AI/ML concepts. Critical data quality issues (duplicates, imbalances) were rigorously addressed, prioritizing the reliable New G/L (FAGLFLEXA) to establish a trustworthy foundation. Exploratory Data Analysis (EDA) informed the engineering of 16 context-specific features. The core innovation lies in a **scalable Hybrid Anomaly Detection strategy**, blending ensemble unsupervised ML models (Isolation Forest, LOF, Autoencoder via Scikit-learn/TensorFlow) with **highly customizable, expert-defined SAP rules (HRFs)**. This approach provides robust, context-aware anomaly prioritization using multi-faceted scores (Priority_Tier, Model_Consensus), presented via an interactive dashboard (Python/Streamlit/Pandas/AgGrid) for efficient investigation. The monitor successfully demonstrates a methodology for enhancing financial controls, reducing risk, and increasing efficiency, with clear **scalability and integration pathways** defined for the operational SAP landscape (SAP BTP AI Core, OData, Fiori/SAC, Workflow), augmenting standard SAP capabilities.

2. Introduction: The Challenge of Financial Integrity in Complex SAP Landscapes

Maintaining financial integrity is critical for organizations relying on SAP. However, the scale and complexity of data in core FI/CO tables (BKPF, FAGLFLEXA, historically BSEG) make identifying subtle errors, potential fraud, or compliance deviations challenging via manual review or standard rule-based checks alone. These undetected anomalies pose significant business risks: inaccurate reporting, reconciliation burdens, audit failures, and hidden operational issues. This project aimed to bridge this gap by creating an intelligent, data-driven monitoring system.

3. Project Objectives

- Validate and prepare SAP FI/CO data (BKPF, FAGLFLEXA) for reliable analysis, addressing common quality issues.
- Understand baseline financial posting patterns through EDA.

- Engineer features quantifying potentially anomalous deviations using EDA insights and SAP domain expertise.
- Develop and apply a **robust, hybrid anomaly detection strategy** combining ML and expert SAP rules.
- Implement an effective prioritization mechanism for detected anomalies.
- Demonstrate the solution's value via an interactive POC dashboard, the **"Intelligent SAP Financial Integrity Monitor."**

4. Overall Methodology & Technical Environment

A phased approach (Data Quality -> EDA/FE -> Modeling/Prioritization -> UI POC) was executed using **Python 3.x** with core libraries: **Pandas, NumPy, Scikit-learn, TensorFlow/Keras, Matplotlib, Seaborn, Joblib, Streamlit, streamlit-aggrid**.

5. Phase 1: Building a Reliable Foundation - Data Quality & Preparation

- **Data Sources & Initial Findings:** Analysis of raw BKPF, BSEG, FAGLFLEXA extracts revealed critical duplicates (~53,000+) and financial imbalance in the classic BSEG view.
- **Strategic Choice:** Prioritized FAGLFLEXA (New G/L) for its financial balance and richer dimensionality, pairing it with BKPF for header context. BSEG was discarded to prevent propagating inaccuracies. *Justification:* Using a balanced, modern ledger structure is fundamental for reliable financial analysis.
- **Cleansing:** Systematically removed exact duplicates based on SAP primary keys using pandas.drop_duplicates().
- **Validation:** Post-cleaning checks confirmed 100% uniqueness, header-item consistency, financial balance (FAGLFLEXA hsl sum = 0), and genuine dimension population. *Outcome:* A trustworthy dataset ready for analysis.

6. Phase 2: Understanding the Data & Engineering Predictive Features

- **Exploratory Data Analysis (EDA):** Analysis of the cleaned, merged dataset revealed typical posting time distributions (business hours, weekly/monthly cycles), user concentration patterns, context dependency of amounts (user, doc type), and confirmed expected process flows (MM/SD integration via TCodes/Doc Types).
- **Feature Engineering (FE):** *Justification:* To quantify deviations identified in EDA and leverage SAP knowledge, 16 features (FE_...) were engineered, categorized as:
    - **Timing:** Deviations from typical work hours/days.
    - **Magnitude:** Log/Absolute transformations of hsl.

    - **User-Based:** Frequency, average amount deviation relative to user history, rare TCode usage.
    - **Account/Dimension-Based:** Account posting frequency, amount deviation relative to account history, check for missing Cost Centers on expenses.
    - **Contextual:** Rarity of Document Type / T-Code.
- **Result:** sap_engineered_features.csv – a feature-rich dataset primed for ML.

7. Phase 3 & 4: Hybrid Anomaly Detection, Prioritization & Evaluation

- **Detection Strategy:** Employed a **scalable Hybrid Anomaly Detection** approach. *Justification:* Combines ML's ability to find novel patterns with the precision of expert rules for known risks, providing comprehensive and context-aware detection.
    - **Ensemble ML:** Utilized unsupervised algorithms: Isolation Forest, Local Outlier Factor (Scikit-learn), and an Autoencoder (TensorFlow/Keras trained on 'normal' data). *Rationale:* Diverse models capture different anomaly types.
    - **Expert Rules (HRFs):** Implemented **highly customizable** boolean flags (HRF_...) based on engineered features exceeding dynamic percentile thresholds or violating contextual rules (e.g., weekend posting, missing cost center). *Rationale:* Directly encodes domain expertise and specific control points.
- **Prioritization:** Implemented a multi-tiered system (Priority_Tier) based on Model_Consensus (number of ML models flagging) and HRF_Count. *Justification:* Focuses investigation on anomalies with strongest evidence.
- **Context Generation (Review_Focus):** Programmatically created text summaries explaining *why* an item was flagged (models involved, HRFs triggered, SAP context).
- **Evaluation:** Included anomaly profiling (comparing feature statistics between normal/anomalous groups for each model) and visual assessment using PCA/t-SNE plots.

8. Solution Demonstration: The "Intelligent SAP Financial Integrity Monitor" POC

- **Technology:** Interactive dashboard built with **Python (Streamlit, Pandas, Plotly Express, AgGrid)**.
- **Key Features:**
    - Secure file upload for anomaly & feature data.
    - **Accurate Multi-Currency KPIs:** Displays "Value at Risk" correctly grouped by Company Code and local currency.

    - Comprehensive interactive filtering.
    - Dynamic visualizations (User/Doc Type/HRF frequencies, time trends).
    - Prioritized anomaly investigation list (AgGrid).
    - Detailed drill-down view integrating anomaly reasons, flags, core SAP data, and additional features.
- **Value:** Provides an intuitive, actionable interface for analysts to efficiently explore, understand, and investigate prioritized financial anomalies.

9. Scalability, Customizability & SAP Landscape Integration

- **Scalability:** The underlying Python/Pandas processing and ensemble ML approach are inherently scalable with appropriate infrastructure. The architecture was designed with large datasets in mind. Integration via BTP (see below) leverages cloud scalability.
- **Customizability:** The rule-based component (HRFs) is highly customizable – new rules can be easily added or thresholds adjusted based on evolving business risks or specific audit requirements. Feature engineering can also be extended.
- **Future Integration Strategy (Beyond POC):** The design explicitly considers seamless integration into the operational SAP landscape:
    - **Data:** Utilize **OData Services (via SEGW/CDS Views)** or **SAP Data Intelligence Cloud** for real-time data feeds.
    - **Model Execution:** Deploy models and logic on **SAP BTP (AI Core/AI Launchpad)** for robust management and scalability, integrating via APIs.
    - **User Interface/Actions:** Replace Streamlit with **Custom Fiori Apps** for native UX, embed insights into **SAC Dashboards**, trigger **SAP Workflows** for automated remediation, and persist results in **Custom SAP Tables**.
- **Benefits:** Integration enables near real-time monitoring, reduces manual effort, accelerates investigation cycles, and provides a unified experience within SAP.

10. Conclusion & Value Proposition

The **"Intelligent SAP Financial Integrity Monitor"** project successfully demonstrates a robust, data-driven methodology for enhancing financial controls within SAP. By strategically cleaning data, engineering insightful features, and applying a **scalable, customizable hybrid detection strategy**, the system effectively identifies and prioritizes potential anomalies. This approach augments standard SAP capabilities, offering:

- **Enhanced Detection:** Uncovering complex issues missed by traditional checks.

- **Increased Efficiency:** Focusing investigative resources on the highest-risk items.
- **Reduced Financial Risk:** Enabling earlier identification of errors or potential fraud.

The POC provides a strong foundation for a powerful, integrated tool to safeguard financial integrity within the enterprise SAP environment.

11. Next Steps

- Formal quantitative model evaluation and hyperparameter tuning.
- Pilot deployment with user feedback collection for threshold/rule refinement.
- Development of the planned SAP integration architecture (BTP, Fiori/SAC, Workflow).
- Exploration of additional features (e.g., NLP on text fields, graph analysis).