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 (Consolidated Report Date)
Dataset Origin: https://www.kaggle.com/datasets/sunithasiva/sap-dataset

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 leverages SAP expertise combined with applied AI/ML concepts. Critical data quality issues (duplicates, imbalances) were rigorously addressed, prioritizing the reliable New G/L (FAGLFLEXA) and removing over 53,000 duplicate records 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.

1. Introduction & Problem Statement

Financial integrity is paramount for any organization utilizing SAP for its core accounting functions. However, the sheer volume and complexity of transactional data generated within tables like BKPF, BSEG, and FAGLFLEXA make manual oversight challenging and traditional rule-based checks insufficient for detecting subtle or novel anomalies. Undetected errors, fraud, compliance breaches, or process deviations within these postings can lead to significant business risks, including:

- Inaccurate financial statements reported to stakeholders.
- Inflated revenue or expense figures distorting performance analysis.
- Severe reconciliation difficulties consuming valuable finance team resources.
- Potential failed audits or compliance penalties.

- Masking of operational errors or fraudulent activity.

This project was initiated to develop a more intelligent, data-driven system capable of proactively identifying potentially anomalous financial postings, thereby enhancing financial controls and mitigating associated risks.

2. Project Objectives

The primary objectives of this project were:

- Assess the quality, completeness, and suitability of standard SAP financial data extracts (BKPF, BSEG, FAGLFLEXA) for reliable anomaly detection.
- Implement a robust data cleaning and preparation strategy to create a trustworthy, financially balanced dataset adhering to SAP structures and accounting principles.
- Perform Exploratory Data Analysis (EDA) to understand baseline patterns and identify characteristics indicative of normal vs. potentially anomalous financial postings.
- Engineer relevant features based on EDA insights and SAP domain expertise to quantify deviations from normal behavior.
- Develop and apply a hybrid anomaly detection strategy combining Machine Learning models and expert-defined rules.
- Create a system for prioritizing detected anomalies to focus investigation efforts.
- (Proof-of-Concept) Demonstrate the findings and investigation process via an interactive dashboard.

3. Overall Methodology

A phased approach was adopted, progressing from foundational data integrity to advanced modeling:

- **Phase 1: Data Quality Assessment & Preparation:** Focused on acquiring, understanding, cleaning, and validating the core SAP financial data extracts.
- **Phase 2: Exploratory Data Analysis & Feature Engineering:** Analyzed the cleaned data to uncover patterns and engineered features to capture potential anomaly indicators.
- **Phase 3: Anomaly Detection Model Building & Evaluation:** Utilized the engineered dataset to train, apply, and evaluate anomaly detection algorithms and rules. (Details inferred from prior discussions).

4. Phase 1: Data Acquisition & Quality Assessment

- **4.1. Data Sources:**

  - bkpf.csv: Accounting Document Header (Table BKPF) - Essential for header context (dates, users, doc types, T-codes).
  - bseg.csv: Accounting Document Segment (Table BSEG - Classic G/L View) - Initially considered for line item details.
  - faglflexa.csv: General Ledger Actual Line Items (Table FAGLFLEXA - New G/L View) - Contains actual postings with enhanced dimensions (e.g., Segment) crucial for modern financial analysis.
- **4.2. Initial Findings (Pre-Cleaning):** Analysis revealed critical issues:
  - **Uniqueness Failures (Critical):** Significant duplicates were found in both BKPF (13,166 headers) and FAGLFLEXA (40,286 line items). *Justification:* Duplicate records fundamentally misrepresent financial activity and invalidate analysis.
  - **Financial Imbalance (Critical):** The BSEG extract showed a significant non-zero sum for local currency amounts (dmbtr), violating the double-entry accounting principle. *Justification:* Imbalanced data cannot be reliably used for financial analysis or reporting; it suggests an incomplete or flawed extract.
  - **Consistency:** Orphaned line items (items without corresponding headers) were a potential risk if not handled during cleaning.
  - **Balance Confirmation:** FAGLFLEXA (based on hsl) summed close to zero, indicating it represented a financially balanced ledger view.
- **4.3. Cleaning Strategy & Implementation:** Based on the findings, a strategic approach was defined:
  - **Discard BSEG:** *Justification:* Due to its critical financial imbalance, relying on BSEG would propagate inaccuracies. It was deemed unsuitable for this analysis.
  - **Focus on BKPF & FAGLFLEXA:** *Justification:* BKPF provides necessary header context, while FAGLFLEXA offers a reliable, balanced source for New G/L line items with richer dimensionality, aligning with modern SAP financial structures.
  - **Handle Duplicates:** Programmatically remove exact duplicate rows using pandas.drop_duplicates(), keeping the first instance based on primary keys (bukrs, belnr, gjahr for BKPF; rbukrs, docnr, ryear, docln for FAGLFLEXA). *Justification:* This ensures each unique business transaction (header or line item) is represented only once, crucial for data integrity and preventing inflated counts or values.

- **4.4. Post-Cleaning Verification Results:** Rigorous checks confirmed the strategy's effectiveness:
  - **Uniqueness:** 0 duplicates found based on primary keys in cleaned BKPF (136,891 unique headers) and FAGLFLEXA (291,648 unique lines).
  - **Consistency:** 0 orphaned FAGLFLEXA items found; all line items linked to a unique header. Redundancy checks between key overlapping fields (Posting Date, User, etc.) in cleaned BKPF and FAGLFLEXA showed 0 mismatches.
  - **Financial Balance:** Cleaned FAGLFLEXA hsl sum remained zero, confirming adherence to double-entry principles.
  - **Dimension Population:** Critical dimensions (rcntr, prctr, segment) remained 100% populated with genuine values (0% placeholders).

5. Phase 2: Exploratory Data Analysis (EDA) & Feature Engineering (FE)

- **5.1. EDA Objectives & Process:**
  - *Justification:* EDA is essential to understand the inherent patterns and characteristics of the *cleaned* financial data before attempting to detect deviations. It helps establish a baseline of "normal" behavior.
  - The cleaned BKPF and FAGLFLEXA datasets were merged. Python (Pandas, NumPy, Matplotlib, Seaborn) was used to analyze distributions, time series patterns, and relationships across dimensions (time, user, account, process context, value).
- **5.2. Key EDA Findings:**
  - **Transaction Timing:** Clear patterns emerged - activity peaks during business hours (8 AM-1 PM), minimal on weekends, and shows monthly cycles (peaks near mid-month/month-end). Deviations are suspicious.
  - **User Activity:** Posting volume is highly concentrated among a few users (BHUSHAN, MOUNIKAPALI). Different users exhibit distinct posting amount profiles (median, spread). This highlights the importance of user context.
  - **Process Context:** Document Types (e.g., WE, RE dominate) and T-Codes (e.g., MB01, MIRO dominate) confirmed expected MM/SD integration points. Rare document types or user-TCode combinations are potential flags. Amount distributions vary significantly by document type, emphasizing context dependency.
  - **GL Accounts & Dimensions:** High activity in expected clearing/recon accounts. Key dimensions fully populated.

  - **Monetary Value (hsl):** Extremely wide distribution concentrated near zero, confirming the need for log transformations or relative deviation features for modeling.

- **5.3. Feature Engineering Strategy:**
  - *Justification:* Raw data fields often don't directly capture anomalous behavior. Feature Engineering transforms the data, informed by EDA and SAP knowledge, to create variables that explicitly quantify potential deviations from normalcy, making it easier for models to detect anomalies.
  - Based on EDA insights and SAP process knowledge, 16 features were engineered.
- **5.4. Engineered Features Overview (Categories & Rationale):**
  - **Timing Features (from BKPF budat):** FE_PostingHour, FE_DayOfWeek, FE_IsOutsideBusinessHours, FE_IsWeekend. *Rationale:* Quantify deviations from typical posting schedules identified in EDA.
  - **Magnitude Features (from FAGLFLEXA hsl):** FE_AbsoluteAmount, FE_LogAmount. *Rationale:* Provide scale-invariant impact and address the wide distribution of hsl.
  - **User-Based Features (from BKPF usnam_bkpf, tcode):** FE_UserPostingFrequency, FE_UserAvgLogAmount, FE_AmountDeviationFromUserMean, FE_IsRareTCodeForUser. *Rationale:* Contextualize activity based on individual user norms (frequency, typical amounts, common processes).
  - **Account/Dimension-Based Features (from FAGLFLEXA racct, rcntr):** FE_AccountPostingFrequency, FE_AccountAvgLogAmount, FE_AmountDeviationFromAccountMean, FE_IsMissingCostCenterForExpense. *Rationale:* Contextualize activity based on G/L account norms and check for known control violations (missing cost center).
  - **Contextual Features (from BKPF blart, tcode):** FE_DocTypeFrequency, FE_TCodeFrequency. *Rationale:* Quantify the rarity of the overall business process being executed.
- **5.5. Resulting Dataset:** The process yielded sap_engineered_features.csv (291,648 rows × 30 columns), containing original identifiers, selected original fields, and the 16 engineered features, ready for modeling.

6. Phase 3: Anomaly Detection Modeling (Based on prior discussions)

- **6.1. Approach - Hybrid Strategy:**

  - *Justification:* A hybrid approach combining ML and expert rules provides comprehensive coverage. ML detects novel/complex patterns, while rules catch known business risks.
  - Both ML model outputs and rule triggers were used to generate anomaly indicators.

- **6.2. Machine Learning Models:**
  - *Justification:* An ensemble of diverse unsupervised algorithms increases the chances of detecting different types of anomalies.
  - Models Applied: Isolation Forest (IF), Local Outlier Factor (LOF), Autoencoder (AE).
- **6.3. Expert Rules (High-Risk Flags - HRFs):**
  - *Justification:* Leverages deep SAP domain knowledge to flag specific, contextually relevant scenarios (e.g., posting times, missing dimensions, deviations from user/account history) that are known indicators of potential issues, regardless of statistical rarity.
  - Features like FE_IsWeekend, FE_IsMissingCostCenterForExpense, FE_AmountDeviation... directly support HRF generation.
- **6.4. Prioritization Logic:**
  - *Justification:* Given potentially numerous flagged items, prioritization is essential to focus limited investigation resources effectively.
  - Metrics like Model_Consensus (how many ML models flagged the item) and Priority_Tier (likely combining ML scores and HRF severity) were implemented.

7. Solution Demonstration: Interactive Anomaly Monitor (Proof-of-Concept)

To effectively present the prioritized anomalies and facilitate investigation, an interactive web-based dashboard was developed as a proof-of-concept (POC) using Python with the Streamlit framework, leveraging Pandas for data manipulation and AgGrid for enhanced table interactivity. The UI, titled "Intelligent SAP Financial Integrity Monitor," was designed to provide a clear workflow for users:

- 7.1. Data Loading:
  - An intuitive File Upload interface allows users to load the necessary CSV files: the Prioritized Anomaly List (containing model scores, HRFs, and priority tiers) and optionally the Full Engineered Features dataset for enhanced detail drill-down.

- 7.2. Dashboard Overview & KPIs:
  - Dataset Snapshot: Upon processing, the dashboard presents high-level KPIs summarizing the entire loaded dataset, including:
    - Total Anomalies Loaded
    - Total Priority 1 Anomalies
    - Priority 1 Value at Risk (by CoCode): *Critically, this KPI accurately displays the sum of absolute hsl amounts for Priority 1 anomalies, correctly grouped and displayed by Company Code and its associated local currency (e.g., EUR (EU01) | USD (USA1) | CAD (C001)), avoiding misleading cross-currency aggregation.*
    - Total Unique Users, Document Types, and Transaction Codes.
    - Overall Data Date Range.
  - Anomaly Explorer: A subsequent section dynamically updates these KPIs based on user-applied filters, providing an immediate view of the filtered subset's scope and risk profile.
- 7.3. Interactive Filtering:
  - A comprehensive sidebar allows users to intuitively filter the anomaly dataset based on multiple criteria:
    - Anomaly Priority (Priority_Tier)
    - Model Consensus (High, Medium, Low - derived from Model_Anomaly_Count)
    - User (usnam_bkpf)
    - Document Type (blart)
    - Transaction Code (tcode)
    - Posting Date Range (budat_bkpf)
    - Document Number (belnr) search
    - Line Item (buzei) *(Added filter capability)*
- 7.4. Anomaly Visualizations (Filtered Data):
  - Based on the active filters, several visualizations provide contextual insights:
    - Top 5 Users by Anomaly Count: Horizontal bar chart clearly showing users associated with the most anomalies in the filtered set.

        - Top 5 Document Types by Anomaly Count: Bar chart highlighting the business processes (via blart) most frequently associated with filtered anomalies.
        - Anomalies Over Time (by Posting Date): Line chart illustrating temporal patterns and spikes in anomaly occurrence within the filtered data.
        - High-Risk Flag Frequency: Bar chart showing the frequency distribution of the different expert-defined HRFs triggered within the filtered dataset, indicating *why* items were flagged by rules.

- 7.5. Anomaly Investigation List:
    - An interactive AgGrid table displays the filtered anomaly list, prioritized for investigation.
    - Key columns are presented clearly (CoCode, Doc No., Line, Year, Consensus, Priority, Anomaly Reason, Amount, Date, User, Doc Type, TCode).
    - Features include pagination, potential sorting/filtering within the grid, and single-row selection to trigger the detail view.
- 7.6. Anomaly Detail Drill-Down:
    - Selecting a row in the AgGrid table dynamically populates a detailed section below, providing comprehensive context for the specific anomaly:
        - Anomaly Reason: Displays the generated text explaining the flags/context.
        - Priority & Consensus: Shows the key prioritization metrics.
        - Key Risk Flags Identified: Lists the specific HRFs triggered for that record.
        - SAP Document Details: Presents core SAP header and item data (Doc No, CoCode, Date, User, Account, Amount - including both raw numeric and formatted local currency values) for immediate reference.
        - Additional Fields: Optionally displays supplementary data (like drcrk, rcntr, prctr, segment) pulled from the full engineered features dataset, providing richer context.
- 7.7. User Experience: The application provides a clean, responsive interface suitable for analysts and managers to explore anomaly data efficiently, moving from a high-level overview down to specific transaction details.

This interactive dashboard effectively translates the complex outputs of the data preparation, feature engineering, and anomaly detection phases into an actionable tool for financial control and investigation teams.

Streamlit UI 2.pdf

8. Conclusion & Key Outcomes

This project successfully demonstrated a robust methodology for detecting financial anomalies in SAP data. Key outcomes include:

- Validation of FAGLFLEXA + BKPF as a reliable data foundation after rigorous cleaning.
- Identification of key behavioral patterns in financial postings through EDA.
- Successful engineering of features quantifying deviations from these patterns.
- Implementation of a hybrid detection strategy combining ML and expert rules.
- Development of a prioritization mechanism and an interactive POC dashboard for investigation.
- Creation of a trustworthy, feature-rich dataset suitable for advanced analysis.

9. Next Steps & Future Enhancements

- **Formal Model Evaluation (Phase 3 Detail):** Conduct rigorous evaluation of the ML models using appropriate metrics (e.g., precision, recall if labels become available, silhouette score for clustering if applicable) and analyze characteristics of detected anomalies. Tune model hyperparameters and thresholds.
- **Integration into SAP Landscape:** Develop a plan and architecture for deploying the solution within the operational SAP environment using technologies like SAP BTP (AI Core, AI Launchpad), OData services, potentially custom Fiori apps or SAC integration, and SAP Workflow for automated routing and remediation tracking.
- **Feature Refinement:** Consider adding more features, such as NLP on text fields (if available and clean), more complex time-series features, or graph-based features analyzing relationships between entities. Refine existing feature calculations based on evaluation results.
- **Threshold Tuning:** Fine-tune thresholds used in HRFs and potentially in model scoring based on business feedback and acceptable false positive rates.
- **Feedback Loop:** Implement mechanisms for user feedback on flagged anomalies to continuously improve model accuracy and rule relevance (supervised learning elements).