Elo Bank Fraud Dataset (Data Quality Monitoring) Report – Week 5

Name: Joori Ahmed Shareef

Company: Jedco

Department: Business Applications – Data Center 2

Duration: 20 hours

1. Objectives

The main objective of this week was to get hands-on experience with the Dataiku platform and start exploring a credit card transaction dataset provided in the Elo Bank fraud detection use case. This task aligned with item 5 of the internship plan (Monitoring Data Quality in Aviation Datasets) as it introduced us to identifying and resolving real-world data quality issues, especially those related to financial fraud detection.

2. Context & Problem Statement

Elo Bank is facing an increase in fraudulent transactions across U.S. states, leading to a rise in insurance costs. Our mission was to create a map identifying states with the highest number of fraudulent transactions, which required a clean, unified dataset built from multiple sources.

3. Activities Completed

- Attended the onboarding session with Dataiku's team, which covered:
  1. Overview of the Elo Bank use case.
  2. Core DSS (Dataiku Data Science Studio) concepts.
  3. Project creation and dataset integration steps.

- Connected four datasets to the platform: RANSACTIONS_2021, TRANSACTIONS_2022, MERCHANT_INFO, CARDHOLDER_INFO.

- Applied the Stack visual recipe to consolidate TRANSACTIONS_2021 and TRANSACTIONS_2022 into a single dataset (TRANSACTIONS_STACKED).

- Used the Join recipe (Left Join) to merge TRANSACTIONS_STACKED with merchant and cardholder info.

- Analyzed column-level issues, including:
  - Invalid formatting in signature_provided column (error values instead of integers).
  - Mixed datetime formats in purchase_date.
  - Inconsistent data types across datasets (e.g., authorized_flag as string vs. integer).

4. Key Learnings

- Gained hands-on experience using Dataiku visual flows to combine and inspect datasets.
- Understood how stacking helps unify datasets across different years.
- Learned how left joins allow enriching transactions with related metadata from merchants and cardholders.
- Identified common real-world data quality problems, especially in financial fraud detection contexts:
    - Formatting mismatches.
    - Null or erroneous entries.
    - Type conflicts across sources.

5. Reflection

This week gave me a deep introduction to how business data from multiple sources must be cleaned and combined before any advanced analytics or modeling can take place. Even though the Elo Bank case is financial, the same challenges exist in aviation data, where accuracy and consistency are critical. The tools in Dataiku made it easy to spot issues visually and design a pipeline that prepares the data for downstream tasks like fraud detection or dashboarding. I'm excited to build on this foundation next week as we move into data validation and cleaning.