Data Validation & Preparation for Modeling Report – Week 6

Name: Joori Ahmed Shareef

Company: Jedco

Department: Business Applications – Data Center 2

Duration: 20 hours

1. Objectives

The objective of this week was to apply data validation and preparation techniques using Dataiku to ensure data quality before model training. This directly aligns with item 6 in the internship plan: 'Participation in data validation discussions and team review meetings (as observer).' The focus was on using Dataiku's built-in tools to validate, clean, and structure the dataset prepared in Week 5.

2. Activities Completed

- Participated in the second hands-on session with the Dataiku team focused on data validation workflows.

- Reviewed issues from the previous week such as format inconsistencies and type mismatches.

- Used the Prepare recipe to:
  • Remove 983 invalid rows in the column `signature_provided` due to non-integer entries.
  • Replace values in `authorized_flag` with labels: 1 → Authorized, 0 → Fraud.

- Parsed geographical columns to extract `state` values from latitude/longitude.

- Flagged spending types as essential or discretionary and added a calculated column for risk levels.

- Generated summary statistics and used visual profiling tools to validate data distribution and spot remaining issues.

- Confirmed readiness of the cleaned dataset (`TRANSACTIONS_prepared`) for the modeling stage in Week 7.

3. Key Learnings

- Understood the role of data validation as a prerequisite to accurate modeling.

- Learned how to use Dataiku's no-code recipes (Prepare, Filter, Replace, Generate Features) to automate data cleanup.

- Identified and handled multiple data quality problems including nulls, type errors, and formatting issues.

- Explored how data enrichment (e.g., adding risk scores and category flags) helps in feature engineering.

- Realized the impact of clean data on model reliability and future predictions.

4. Reflection

This week served as a critical bridge between raw data and AI modeling. I saw how unclean data could introduce errors into predictive models, and how strategic cleaning steps can significantly improve performance. The hands-on work in Dataiku helped reinforce my understanding of real-world data preparation. I'm now more confident in moving forward to model training in the upcoming phase.