Building a Credit Risk data pipeline: from raw data to Machine Learning ready data

Presented by Min Htet Hein

Purpose: Documentation of design, implementation and output format of a data pipeline for **loan default prediction of each user when applying loan**.

We have built a scalable, reliable pipeline that transforms raw, disparate data sources into a clean, structured, and analysis-ready format for machine learning engineers.

Project leverages the **Medallion Architecture** (Bronze -> Silver -> Gold layers) to ensure data quality and integrity at each state of the pipeline.

Data scope: 4 raw datasets, monthly snapshots

Problem type: Binary classification (will default or not)

Project Structure

```
project/
├── main.py
├── data/ (raw CSV files)
├── datamart/
│   ├── bronze/ (partitioned CSV files)
│   ├── silver/ (cleaned Parquet files)
│   └── gold/ (ML-ready Parquet files)
└── utils/
    ├── data_processing_bronze.py
    ├── data_processing_silver.py
    ├── data_processing_gold.py
    └── date_utils.py
```

Business Problem

Class: CS611 G1

The problem: raw data from loans, customer attributes, financials, and website clicks were in inconsistent format not yet suitable for being fed into Machine Learning pipeline to build predictive models. There were data silos across loan management (LMS), customer attributes, financials and clickstream.

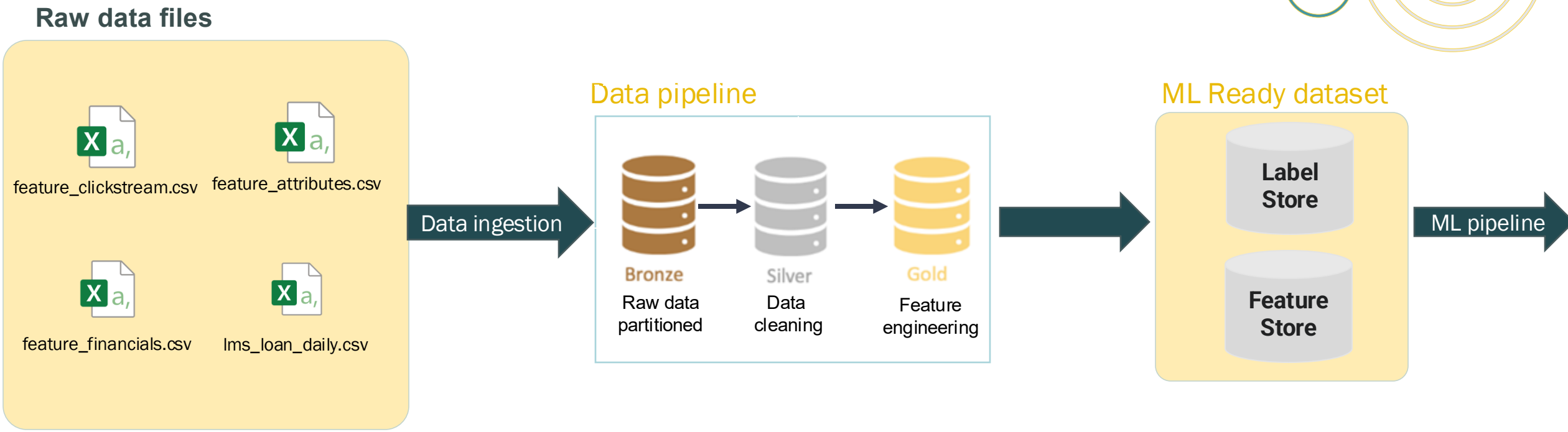
Solution: a unified data pipeline that automatically processes this data on a monthly basis (first day of the month used as snapshot date in the data), creating a single source of truth for loan default prediction

End output:

- Feature Store: a curated set of 8 key customer attributes (eg: Income, Debt, Payment Behaviour) for customers with active loans
- Label Store: a clear definition of “bad” loans (30+ days past due after 6 months) to train and evaluate models

Business Impact: this datamart generated enables us to build accurate models to predict customer default, leading to better lending decisions, reduced risk, and thus increased profitability in the long run.

Data Journey



- The pipeline ingests raw data and prepares the data into Gold standard for modelling.
- A centralized Feature Store ensures consistent features are used for both model training and deployment, promoting reliability and reusability.
- Bronze layer: As-Is Archive ; csv format files
- Silver layer: cleaned data; standardized; validated; parquet format files
- Gold layer: Business level; **Features** & **Label** clearly produced; Machine Learning Ready

Raw Datasets Provided

These are the **columns** of each of 4 datasets which are examined one by one to see which column needs data cleaning or can be used for feature engineering.

feature_financials

- Customer_ID
- Annual_Income
- Monthly_Inhand_Salary
- Num_Bank_Accounts
- Num_Credit_Card
- Interest_Rate
- Num_of_Loan
- Type_of_Loan
- Delay_from_due_date
- Num_of_Delayed_Payment
- Changed_Credit_Limit
- Num_Credit_Inquiries
- Credit_Mix
- Outstanding_Debt
- Credit_Utilization_Ratio
- Credit_History_Age
- Payment_of_Min_Amount
- Total_EMI_per_month
- Amount_invested_monthly
- Payment_Behaviour
- Monthly_Balance
- snapshot_date

feature_attributes

- Customer_ID
- Name
- Age
- SSN
- Occupation
- snapshot_date

feature_clickstream

- fe_1 , fe_2, ….., fe_19, fe_20
- Customer_ID
- snapshot_date

lms_loan_daily

with datatypes

- loan_id - String
- Customer_ID - String
- loan_start_date - Date
- tenure - Integer
- installment_num - Integer
- loan_amt - Float
- due_amt - Float
- paid_amt - Float
- overdue - Float
- balance - Float
- snapshot_date - Date

Data about each user

Data about loans

Data foundation: from Raw source to cleaned data

Bronze layer

- Create an unaltered archive of all source data for tracebility
- **Data input (ingestion)**
  - 4 Source Tables ingested:
    - **bronze_loans**
      - Main loan account data
    - **bronze_financials**
      - Customer financial health indicators
    - **bronze_attributes**
      - Customer demographic data
    - **bronze_clickstream**
      - Customer website engagement metrics
- **Output:**
  - csv files partitioned by month in datamart

Silver layer

- Apply data quality rules, standardize formats, validate rules
- **Data input**
  - csv files for 4 raw datasets partitioned by month in datamart
- **Data cleaning**
  - Column data type formatting
  - clean columns with invalid values in them
- **Output:**
  - File organization similar to bronze layer but data output in parquet file format with cleaned data in datamart

Gold layer: features and label

Feature store engineering

- Created a curated set of business-ready features specifically for users with active loans
- **Core features**
  1. Annual_Income
  2. Outstanding_Debt
  3. Credit_Utilization_Ratio
  4. Total_EMI_per_month
  5. Num_of_Delayed_Payment
- **Derived features**
  1. Functions Debt_to_Income_Ratio
  2. EMI_Burden_Ratio
  3. Credit_Mix_encoded

Label store engineering

- Created a binary target variable for prediction
- **Label definition**
  - “bad” loan or “likely to default” as a customer who is 30+ days past due (DPD) when their loan is 6 months old (MOB)
  - Label = 1 if *DPD >= 30* at MOB = 6
    - Among the loans that are exactly 5 months om Boo, Flag as “bad” if 30+ days past die (DPD)
    - Likely to default
  - Label = 0 otherwise

8 features are selected to predict a customer’s ability and willingness to repay (whether or not user will default). Not many columns in the dataset given are used to avoid the “curse of dimensionality” which is a concept where the number of records required increases exponentially with every increase in number of features (dimensions).

Features selection

Feature	Base or Derived?	If derived, formula	Reason
Annual_Income	Base	NA. from raw data	Direct measure of repayment ability. Higher income could mean lower default risk.
Outstanding_Debt	Base	NA. from raw data	Total existing debt. Higher debt could increase default risk
Credit_Utilization_Ratio	Base	NA. from raw data	Credit usage behaviour. High utilization could mean financial stress
Total_EMI_per_month	Base	NA. from raw data	Monthly debt oligations. Higher EMIs could increase default risk
Num_of_Delayed_Payment	Base	NA. from raw data	Historical payment performance. Past delays could be predict future defaults
Debt_to_Income_Ratio	Derived	**Outstanding_Debt** / **Annual_Income**	Measures overall debt burden relative to income. High ratio could be problem
EMI_Burden_Ratio	Derived	**Total_EMI_per_month** / **Monthly_Inhand_Salary**	Cash flow pressure. High ratio could mean higher default risk
Credit_Mix_encoded	Derived	Encoded from **Credit_Mix**	Summarizes credit history quality. “Good” mix indicates experience managing different credit types well

Pipeline output metrics

Processing Statistics

- **Time Period:**
  - 24 months
  - January 2023 – December 2024
- **Total bronze files:**
  - 96 (4 datasets x 24 months)
- **Total silver tables**
  - 96 cleaned Parquet files
- **Gold feature store:**
  - 24 monthly snapshots
- **Gold label store:**
  - 24 monthly snapshots

Machine Learning ready dataset

- Total Records:
  - 23,948 customer-month observations
- Dimensionality:
  - 8 features + 1 label
- Data types:
  - Mainly float or integer
- Format:
  - Apache Parquet
    - Optimized for fast querying and storage efficiency

Future Work

1. **Machine Learning Pipeline** (next assignment)
2. **MLOps**

Thank You

Presented by Min Htet Hein