From Data to Decisions: A Production-Ready ML Pipeline for Robust Loan Default Prediction

Lim Kai Zhuo

Nov 2025

Executive Summary

End-to-end, production-ready ML pipeline featuring a robustly trained and tuned model for accurate, scalable, and governed credit-risk prediction.

Problem

Loan defaults have risen sharply, impacting profitability and increasing credit risk exposure.

Data scattered across multiple systems, inconsistent preprocessing, and manual model refreshes have made existing credit models unstable and slow to respond to new borrower behavior.

Without a unified and monitored pipeline, prediction accuracy declines over time, leading to poor credit decisions.

Solution

A fully automated Machine Learning Pipeline was built using Medallion Architecture, Docker, and Apache Airflow to manage the complete model lifecycle.

Process Data: Bronze–Silver–Gold layers standardize and clean raw financial, attribute, and behavioral data, creating validated, leakage-free features.

Develop Model: A 3-stage framework automates hyperparameter tuning, SHAP-based feature selection, and final model refinement to identify the best XGBoost model.

Deploy: The best model is stored in a centralized Model Registry and automatically scheduled by Airflow for monthly inference on new data snapshots.

Monitor: Continuous tracking of feature drift and AUC drift detects performance degradation; automated email alerts and governance rules trigger model retraining.

Impact

Accuracy: Achieved stronger predictive performance with improved model stability across time, enabling earlier and more accurate identification of high-risk or potential bad loans.

Efficiency: Streamlined data processing and retraining cycles, allowing faster and scalable deployment of refreshed models to keep predictions current with evolving borrower behavior.

Governance: Established transparent, auditable model management aligned with enterprise risk and regulatory standards, ensuring confidence and accountability in production decisions.

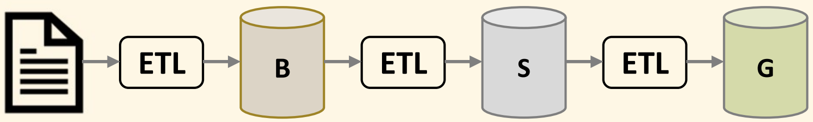
Business Value: Improved detection of default-prone borrowers reduces bad-loan approvals, safeguards credit portfolio quality, and enhances overall profitability through better risk-adjusted lending.

Overall Architecture

End-to-end machine learning pipeline: from raw data ingestion to feature engineering, model tuning, deployment, and monitoring.

Process Data → Develop Model → Deploy → Monitor

Process Data



Bronze
Extract raw data from backend data source

Silver
Data cleaning of raw Bronze data

Gold
Feature engineering for new, insightful, and non-linear features

Silver Overview

Numerical: Outliers handling, null/ placeholder coercion, string to float/int cleaning
Categorical: Regular expression cleaning, binning, binary/indicator variables (e.g. Type of loan split into 1 column for 0/1 for each loan type)
Imputation: Missing values fill with mean of group (e.g. NA salary filled with avg. of job)

Gold Overview

Ratio Metrics: Taking ratios of variables of absolute monetary values into new scale-independent financial ratios. E.g. Debt-to-income, Dept-per-loan, EMI-to-income
Non-linear Transformations: Log, squared, cubed
Threshold / Binning: Simplify continuous variables into bins or ordinal categories

Medallion Data Framework

Develop Model

Initial Tuning

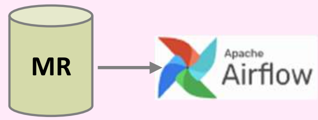
SHAP Feature Importance

Final Tuning + Feature Selection

3 Stage Automatic Hyperparameter + Feature Selection Framework

Model Training & Tuning

Deploy



Model Registry
Stores the best XGB model after tuning

Airflow
Schedule monthly inference on new snapshot data
(or manual trigger)

Deploy with Airflow Monthly Scheduler

Monitor

X | AUC

Feature Drift
Monitors distribution of reference vs new data with KS Statistical Test

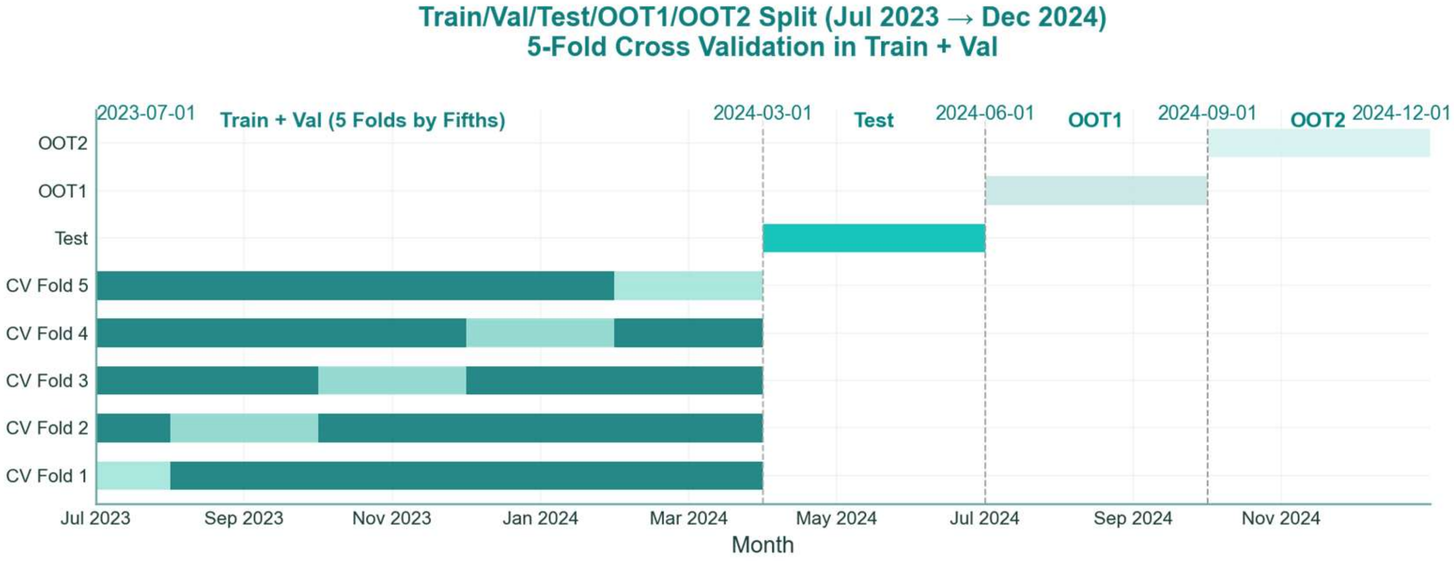
Prediction AUC Drift
Monitors new inference performance across time

Email Alerts
Sends alert if model performance worsens

Monitoring & Governance

Data Splitting Strategy

Temporal data splits **ensure sufficient tuning data** and independent validation windows, **enabling robust evaluation and confidence** in model performance and inference stability before deployment.



- 5 fold cross validation used for more accurate estimation of model performance for each hyperparameter trial
- Robust train/test/OOT split to ensure confidence in model performance & deployment

Dataset	Period	Purpose / Usage	Description
Train / Val	Jul 2023 – Mar 2024	Model training, tuning, and feature selection	Used to train models and optimize hyperparameters through cross-validation.
Test	Apr 2024 – Jun 2024	Final model evaluation	Held out for unbiased assessment of model performance before deployment.
OOT 1	Jul 2024 – Sep 2024	Out-of-time validation	Simulates model behavior on unseen, time-shifted data to check temporal stability.
OOT 2	Oct 2024 – Dec 2024	Production inference simulation	Represents real-world performance via the deployed inference pipeline.

Model Training & Tuning

3-Stage Integrated Hyperparameter Tuning and Feature Selection. Enables automatic and efficient tuning of model together with feature selection. Critical since feature set will affect optimal hyperparameter and vice versa.

Productivity Gain with **Automatic** Hyperparameter tuning + Feature Selection

Optimality where feature set and hyperparameters are **tuned concurrently**

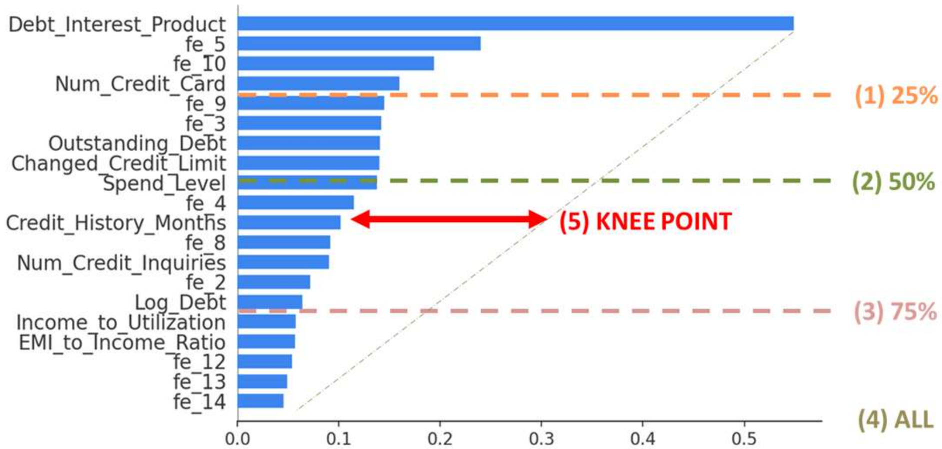
Stage 1
Initial Tuning of Hyperparameters

Learning Capacity
(LR, depth, γ)

Subsampling
(subsample col, tree)

Regularization
(L1, L2, α)

1) Initial Tuning
Feature Set: ALL
Hyperparameters: Tree hparams
Find rough set of hyperparameter that performs relatively well with all features



2) SHAP Feature Importance
Break down feature set into 5 groups
4 groups based on cumulative percentile
Last group based on elbow method (knee pt.)

Stage 3
Final Tuning with Feature Group Choice

Feature Group
(percentile, knee)

Learning Capacity
(LR, depth, γ)

Subsampling
(subsample col, tree)

Regularization
(L1, L2, α)

3) Final Integrated Tuning
Hyperparameters: Tree hparams + Feature Groups
Optuna also optimizes for feature group (Categorical Parameter)
Automatic Feature Selection + Hyperparameter Tuning

* Tuned using TreeBasedSampler in Optuna for cost-efficient hyperparameter tuning

Model Monitoring

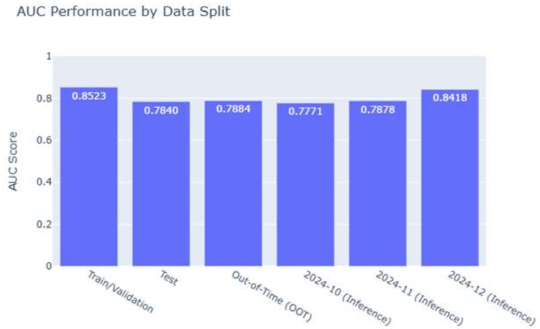
Automated monitoring of model performance and feature drift to safeguard long-term predictive integrity.

Summarized Monitoring Reports *(reports/aggregated_manual/aggregated_plotly_report.html)*

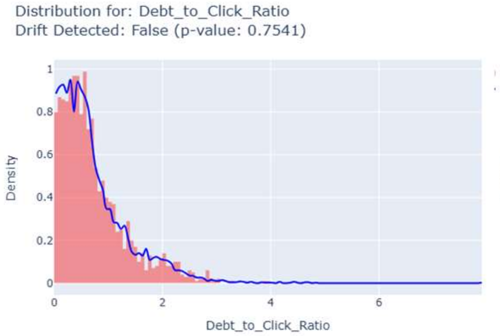
1 Overall Status Summary
Birds-eye view of status of inference data & model performance

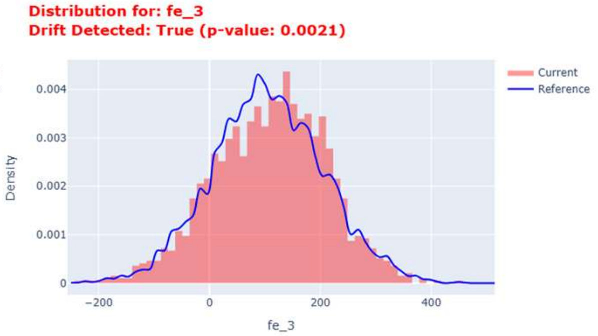
Metric	Status
Feature Drift	1 / 32 features drifted
Model AUC Performance	Good

2 Inference Model Performance
Compares model train, test, and OOT performance to the new inference performance
Highlights any model AUC performance degradation for investigation (3 Tiered Alert)



3 Feature Drift Comparison
Compares the input features distribution for the train, test, and OOT against the new inference data
Highlights any drift (KPSS statistical test p = 0.05) for investigation





Model Governance

Automated governance framework that continuously monitors model health, detects performance degradation, and triggers tiered alerts for retraining actions.

Level	Condition (AUC Drop vs OOT)	Action
Info / Warning	≥ 5% relative drop	Log drift event; email summary; continue monitoring next cycle
Alert / Investigate	≥ 10% relative drop	Deep-dive diagnostics (feature drift, calibration, segment AUCs)
Critical / Retrain	≥ 15% relative drop	Trigger model retraining & shadow test new model before promotion

Simulated email *(reports/aggregated_manual/alert_simulation.html)*

Model Monitoring Alert

Issues were detected during the monitoring run for the period ending **2024-12-01**.

Significant feature drift detected:

- - fe_3 (p-value: 0.0021)

AUC Performance Degradation:

- **Warning:** 2024-10 (Inference) AUC (0.7771) is >5% below OOT AUC (0.7884). Closer observation is advised.
- **Warning:** 2024-11 (Inference) AUC (0.7878) is >5% below OOT AUC (0.7884). Closer observation is advised.
- **Warning:** 2024-12 (Inference) AUC (0.8418) is >5% below OOT AUC (0.7884). Closer observation is advised.

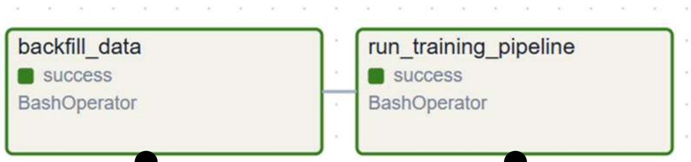
*Threshold manually set to 120% for simulation purposes

1. **Email Alert for Feature Drift**
   Highlights drift in features if KPSS test significant ($p<0.05$).
   Feature plot attached at the bottom of the HTML.

2. **Email Alert**
   Compares model train, test, and OOT performance to the new inference performance
   Highlights any model AUC performance degradation for investigation (3 Tiered Alert)

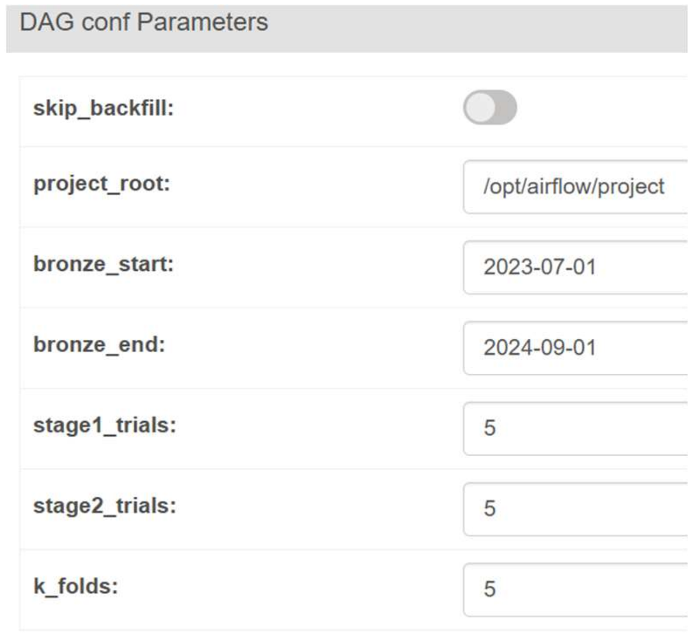
Model Training & Tuning DAG

Automated Airflow DAG orchestrating data backfill, model training, and multi-stage hyperparameter tuning for optimal model selection.



Passes selected months through
bronze – silver – gold ETLs
Stores data in datamart

Runs 3-stage Hyperparameter Tuning
+ Feature Selection
Stores best model in model registry



To skip data pipeline ETL
(if datamart already exists)

Project dir

Start month for ETL job

End month for ETL job

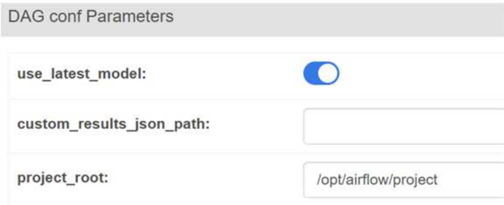
Number of trials for stage 1

Number of trials for stage 3

Number of folds for train/val
crossfold validation

Model Inference, Monitoring, & Governance DAG

Automated Airflow DAG orchestrating model inference, performance monitoring, and governance alerts for continuous production assurance.



Selector to choose latest model for inference (or manual input)

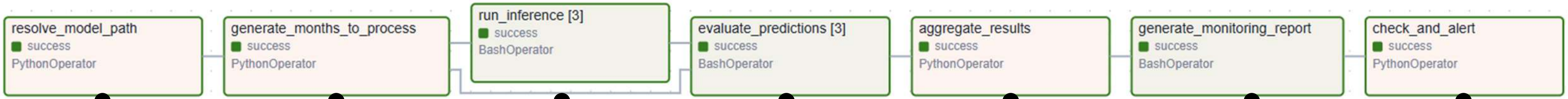
If manual input, enter model dir here

Project root dir

Simluated alert email

Classification threshold

Inference month to run
(Use 2024-12-01 for demo)



Pick the latest tuned
model (or any manual
user-input model path)

Process the snapshot
months to run inference
on based on user-input

Pass new snapshot months through bronze –
silver – gold ETL
Then pass to model for inference
Dynamic task generation based on number
of snapshot months for inference

Evaluate predictions
against actual labels
Calculate AUC

Aggregate predictions results
from various inference months
into a single dataframe

Generate plotly HTML monitoring report
1) Overall Summary
2) Model AUC performance trend
3) Feature Drift Monitoring

Generate simulated email
alert HTML report with alert
and attaches relevant plots
3 Level Alert Structure

Conclusion

A **production-ready, end-to-end ML** pipeline delivering robust, tuned, and continuously monitored models for **accurate and governed credit-risk prediction**.

- **End-to-end automation** across data, model, and monitoring pipelines.

- **Robustly 3-Stage Tuned XGBoost model** delivering higher accuracy and stability.

- **Governed, scalable deployment** with automated drift alerts and retraining.