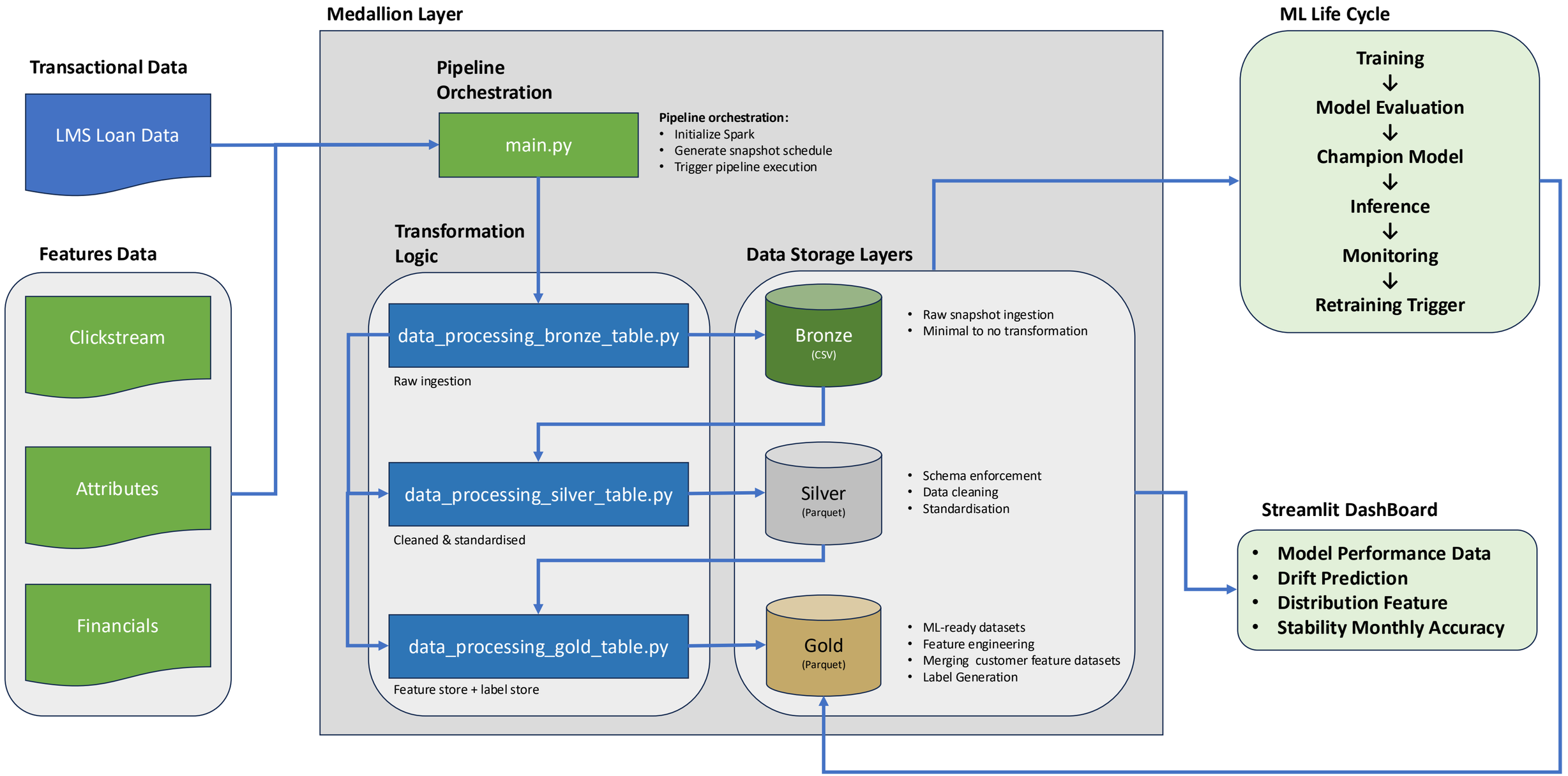
ML Pipeline

The Big Picture



DAG Pipeline – Medallion Architecture

The Medallion pipeline transforms raw monthly data into clean, validated, and ML-ready feature and label stores for downstream model training and inference.

Medallion Architecture

- Implements a **Bronze–Silver–Gold** architecture to prepare data for machine learning.
  - **Bronze:** Ingests and preserves immutable raw snapshots for reproducibility and auditing.
  - **Silver:** Cleans, validates, standardises, and enforces data quality.
  - **Gold:** Generates feature and label stores through feature engineering and data integration.
- Features are engineered using historical information only to prevent temporal and target leakage.
- Feature and label pipelines are processed independently before being combined for machine learning.
- Monthly pipeline runs produce validated feature and label stores for downstream ML workflows.

Feature Engineering Highlights (Gold)

- **Debt-to-Income Ratio** – Measures customer debt burden.
- **EMI-to-Salary Ratio** – Measures repayment affordability.
- **High Credit Utilization Flag** – Identifies customers with high credit usage.
- **Credit History Age (Months)** – Converts credit history into a numeric feature for modelling.

DAG Pipeline – Training

Once the Gold datasets are prepared, the training pipeline builds and evaluates candidate models before selecting the best-performing model for deployment. AUC is the primary evaluation metric because the cost of misclassifying loan defaulters is high, making strong class discrimination important across different decision thresholds.

model_version	train_date	auc_train	auc_test	auc_oot	gini_train	gini_test	gini_oot	brier_train	brier_test	brier_oot	log_loss_train	log_loss_test	log_loss_oot	calibration_method	calibration_selection_metric	prediction_threshold	threshold_selection_metric	threshold_selection_score	oot_threshold_selection_score	champion	challenger
credit_model_2024_09_01_v1	1/9/24	0.87024759	0.84151873	0.8088488	0.74	0.683	0.618	0.1274039	0.13403054	0.15234736	0.4046904	0.42521712	0.47240172	sigmoid	fixed_sigmoi	0.23	youden_j	0.57493536	0.49866308	1	0
credit_model_2024_12_01_v1	1/12/24	0.87540798	0.79627616	0.77977473	0.751	0.593	0.56	0.12561014	0.15210895	0.15910007	0.40552814	0.47310634	0.49024766	sigmoid	fixed_sigmoi	0.28	youden_j	0.47966281	0.45488546	0	1

Training

- **Condition:** Model training begins only after both the Gold feature store and label store are successfully generated.
- **Frequency:** Training is performed on a scheduled **3-month cadence** (at least 3 months from latest training run) to ensure sufficient historical data is available while reducing unnecessary retraining.

Training Pipeline

1. Joins Gold features with labels.
2. Performs a **chronological split** into training, testing and **out-of-time (OOT)** datasets to evaluate performance on unseen future data.
3. Trains multiple XGBoost and Logistic Regression candidate models.
4. Applies probability calibration (Sigmoid) and threshold optimisation.
5. The best-performing model (Test + OOT AUC) is saved to the **Model Bank** *(model_log.csv),* together with its preprocessing pipeline, evaluation metrics, and model metadata.
6. If retraining is not scheduled, the existing Champion model is reused for downstream inference.

DAG Pipeline – Champion Model

The Champion Model stage governs model versioning, approval, and deployment to production.

model_version	train_date	auc_train	auc_test	auc_oot	gini_train	gini_test	gini_oot	brier_train	brier_test	brier_oot	log_loss_train	log_loss_test	log_loss_oot	calibration_method	calibration_selection_metric	prediction_threshold	threshold_selection_metric	threshold_selection_score	oot_threshold_selection_score	champion	challenger
credit_model_2024_09_01_v1	1/9/24	0.87024759	0.84151873	0.8088488	0.74	0.683	0.618	0.1274039	0.13403054	0.15234736	0.4046904	0.42521712	0.47240172	sigmoid	fixed_sigmoid	0.23	youden_j	0.57493536	0.49866308	1	0
credit_model_2024_12_01_v1	1/12/24	0.87540798	0.79627616	0.77977473	0.751	0.593	0.56	0.12561014	0.15210895	0.15910007	0.40552814	0.47310634	0.49024766	sigmoid	fixed_sigmoid	0.28	youden_j	0.47966281	0.45488546	0	1

Champion Model

- Compare newly trained models against the current Champion using evaluation metrics.
- Register the best-performing candidate as a **Challenger** for review.
- Promote the Challenger to **Champion** only after approval.
- Store model artefacts, metadata, and version history in the **Model Bank.**
- Each ML algorithm will have **at most 1 Champion and 1 Challenger.**
- Deploy the latest Champion model for downstream inference.

Promotion Criteria:

✅ Higher Test AUC
✅ Higher OOT AUC
↓
Challenger
↓
✅ Manual Approval
↓
Champion

Promotion Logic

- A newly trained model is evaluated against the current Champion using **Test AUC** and **Out-of-Time (OOT) AUC**.
- The model is registered as a **Challenger** only if it outperforms the Champion on **both** evaluation metrics.
- The Challenger requires **manual approval** before being promoted to Champion, providing a governance checkpoint.
- Once approved, the Champion status is updated in the **Model Bank** and becomes the default model for inference.

DAG Pipeline – Inference

The inference behaves like a volume trigger. If a new gold feature partition appears and the selected model has not scored it yet, inference runs for that partition.

Inference

- Inference begins after the training stage, regardless of whether a new model was trained or the existing Champion model was reused.
- The pipeline loads the latest Champion model from the **Model Bank**.
- Only new Gold feature partitions that have not been scored by the current model version are selected for inference.
- The Champion model generates calibrated prediction probabilities and binary default predictions.
- Prediction results are stored as a Gold Prediction Table partitioned by snapshot date for downstream monitoring and reporting.

DAG Pipeline – Monitoring and Retraining

The monitoring pipeline tracks model performance and data drift to determine whether retraining is required.

Monitoring

- Monitoring begins after inference is completed.
- Compare the current inference data against the training reference dataset.
- Measure model and feature drift using **Population Stability Index (PSI)** and **Characteristic Stability Index (CSI)**.
- Record monitoring metrics and retraining recommendations for model governance.
- Trigger retraining when drift exceeds PSI thresholds.

PSI Threshold to Retrain

- **Healthy:** PSI < 0.1
- **Warning:** 0.1 <= PSI < 0.25
- **Retrain_required = True:** PSI >= 0.25

Performance drop in prediction score such as AUC are to be investigate and training done manually as it computational costly.

Automatic

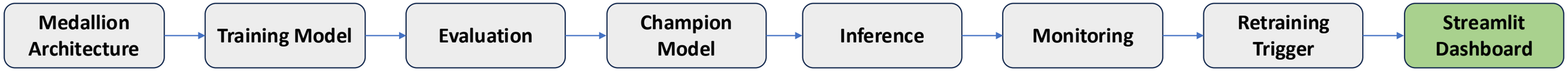
PSI
CSI
↓
Retraining Trigger for Retrain

Manual Review

AUC
Precision
Recall
↓
Retrain Decision

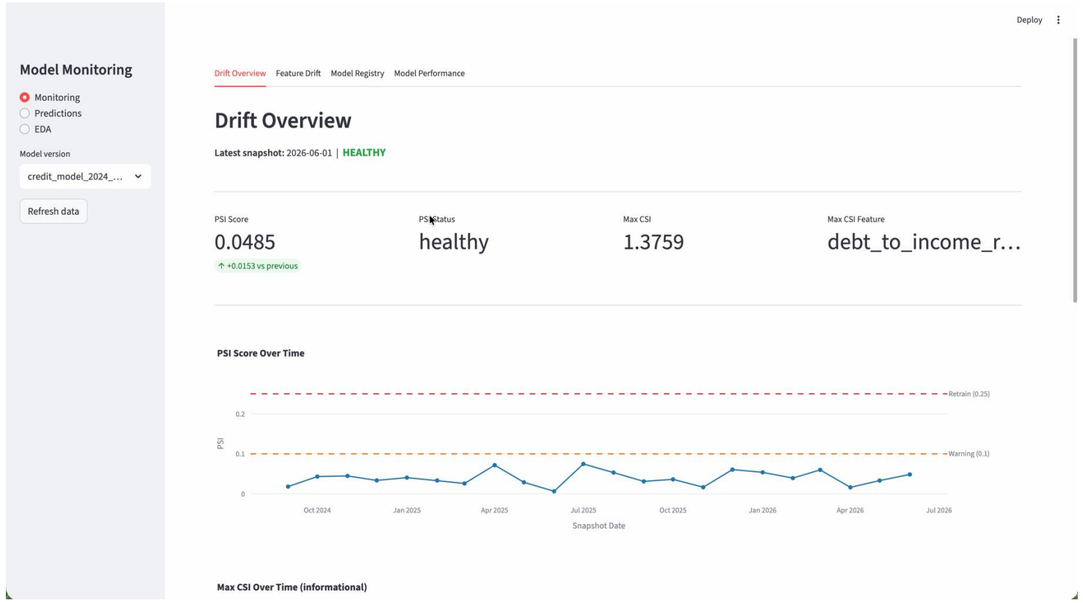
DAG Pipeline – Streamlit Dashboard

The Streamlit dashboard acts as the control room for the ML pipeline, helping users track the active model, predictions, drift, performance, and data health after the Airflow DAG has run.



Main Views

- Pipeline outputs from the **Model Bank** and **Gold datamart**.
- Active **Champion and any Challenger models**.
- **Visualises prediction** trends, default rates, and score distributions.
- **Data and model drift** using PSI and CSI.
- Tracks model performance once labels mature.
- **EDA views** to check Bronze, Silver, and Gold data health.



Launching the Dashboard

Launch via running **streamlit run dashboard.py** in terminal.

End-to-End Simulation with Synthetic Data

Synthetic data extends the observable pipeline from **24 months to 42 months**, allowing the full ML lifecycle to be demonstrated beyond the available real dataset.

Synthetic Data Generation:

1. Extend the ML pipeline beyond **Dec 2024** using synthetic Gold data.
2. Generate monthly feature and label stores from historical Gold distributions with controlled random variation.
3. Preserve engineered features and historical default behaviour.
4. Reuse the existing inference, monitoring, and evaluation pipeline without modifying the Bronze or Silver layers.

End-to-End Simulation Workflow:

1. Execute the main ML pipeline.
2. Generate synthetic Gold feature and label stores.
3. Run inference using the Champion model. Monitor drift with **PSI** and **CSI**.
4. Evaluate predictions using matured labels (**6-month maturity rule**).

Consideration and limitations:

- Gold is simulated while Bronze and Silver remain unchanged to preserve the production data ingestion pipeline.
- Synthetic data supports pipeline validation but cannot replace real-world monitoring.

Model Performance and Stability

XGBoost achieved the highest predictive performance while maintaining good generalisation, with only a small reduction in AUC from the training to the out-of-time dataset.

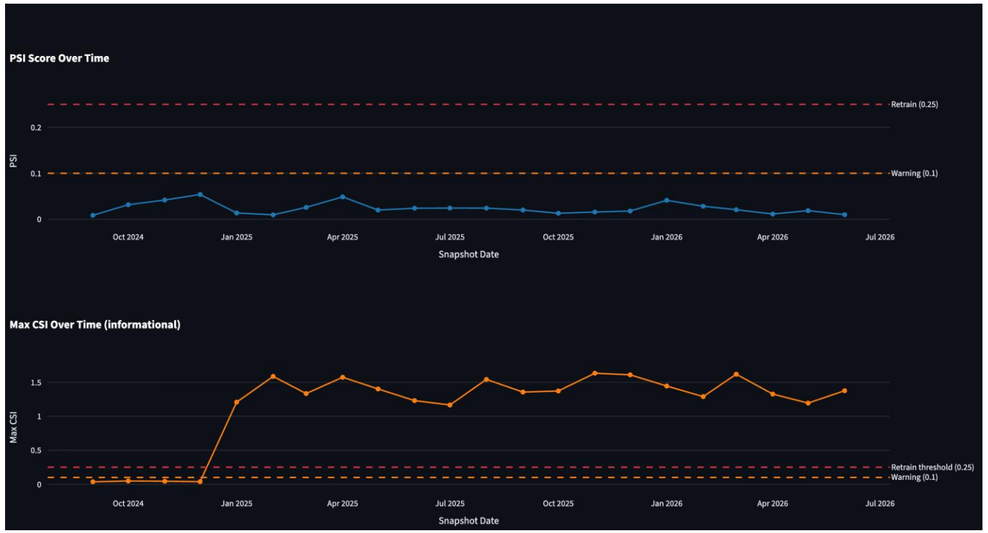


Fig 10.1

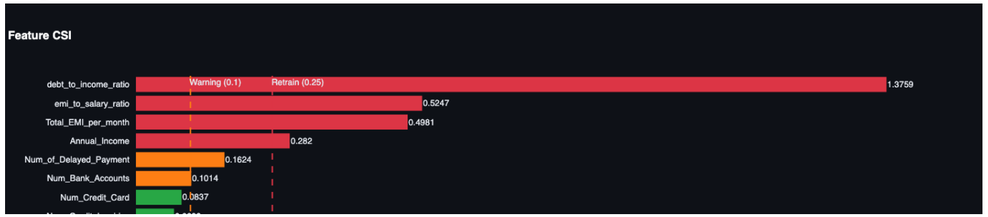
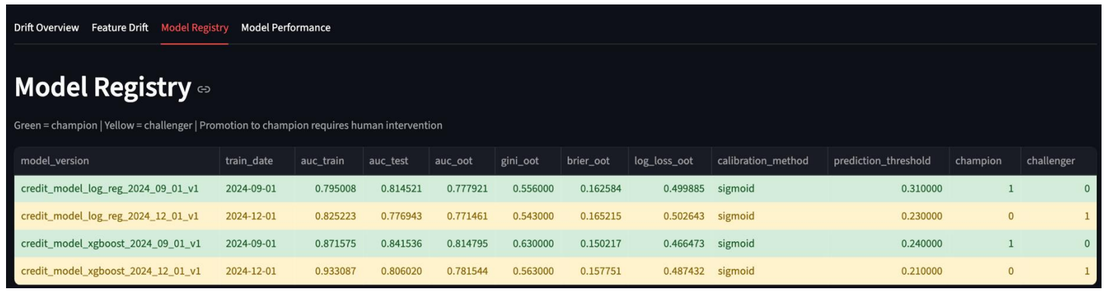


Fig 10.2



model_version	train_date	auc_train	auc_test	auc_oot	gini_oot	brier_oot	log_loss_oot	calibration_method	prediction_threshold	champion	challenger
credit_model_log_reg_2024_09_01_v1	2024-09-01	0.795008	0.814521	0.777921	0.556000	0.162584	0.499885	sigmoid	0.310000	1	0
credit_model_log_reg_2024_12_01_v1	2024-12-01	0.825223	0.776943	0.771461	0.543000	0.165215	0.502643	sigmoid	0.230000	0	1
credit_model_xgboost_2024_09_01_v1	2024-09-01	0.871575	0.841536	0.814795	0.630000	0.150217	0.466473	sigmoid	0.240000	1	0
credit_model_xgboost_2024_12_01_v1	2024-12-01	0.933087	0.806020	0.781544	0.563000	0.157751	0.487432	sigmoid	0.210000	0	1

Fig 10.3

Model Stability

1. The Models are generally stable with a PSI below during Sep 2024 to Jun 2026 (Fig 10.1).
2. CSI are observed to be higher on Jan 2025 (actual data available till end 2024) and its mainly coming from the engineered features (Fig 10.2)

Model Performance

1. XGBoost achieved the highest predictive performance while maintaining good generalisation.
2. Although the AUC decreases slightly from the training, test, and out-of-time datasets, the performance gap is small, indicating only minor overfitting and good robustness on unseen future data (Fig 10.3).

- **XGBoost:** auc_train:0.87, auc_test: 0.84, auc_oot: 0.81
- **Logistic Regression:** auc_train:0.79, auc_test: 0.81, auc_oot: 0.77

Jan 2025 onwards are running on synthetic data and is only used for simulation