CreditKarma Suite

JUNE 2025 UPDATE – ML PIPELINE

Jennifer Poernomo

RECAP: CREDITKARMA SUITE

A unified ecosystem of tools that aggregates data from the organization's diverse systems to assess and analyze customer credit scores.

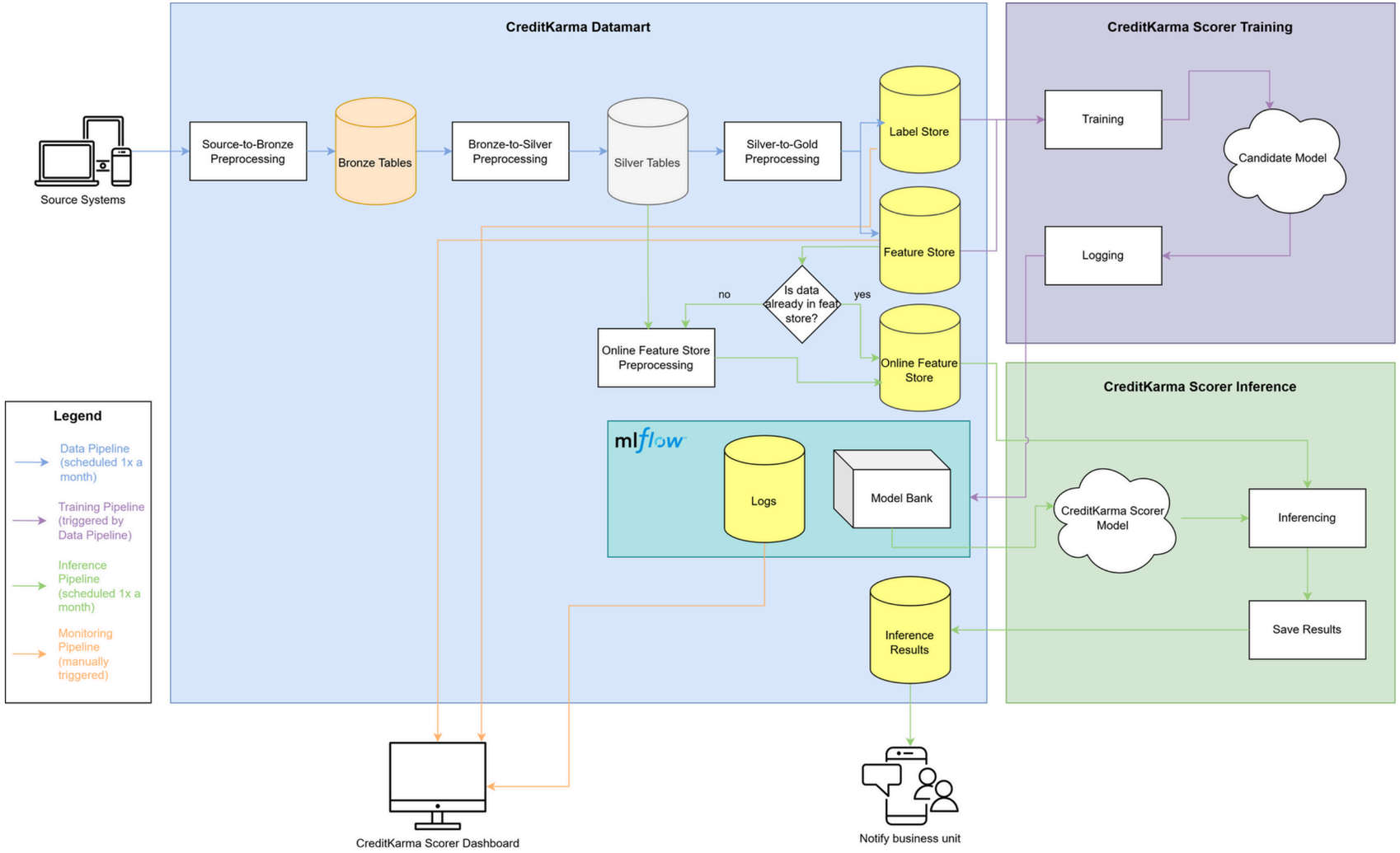
Currently comprises of two main components:

1. **CreditKarma Datamart**: Data repository aggregated from the organizations' systems that serves as a single source of truth applications can easily refer to.
2. **CreditKarma Scorer**: Binary classifier model that predicts the probability of customer default.

PROGRESS UPDATE

1. Deployed end-to-end machine learning pipeline using Airflow.
    a. Monthly automated data processing pipeline.
    b. Automated model training pipeline.
    c. Model inferencing pipeline.
2. Created central model registry for CreditKarma Scorer versioning.
3. Created prototype model monitoring dashboard using Jupyter Notebook.

CREDITKARMA SUITE ARCHITECTURE

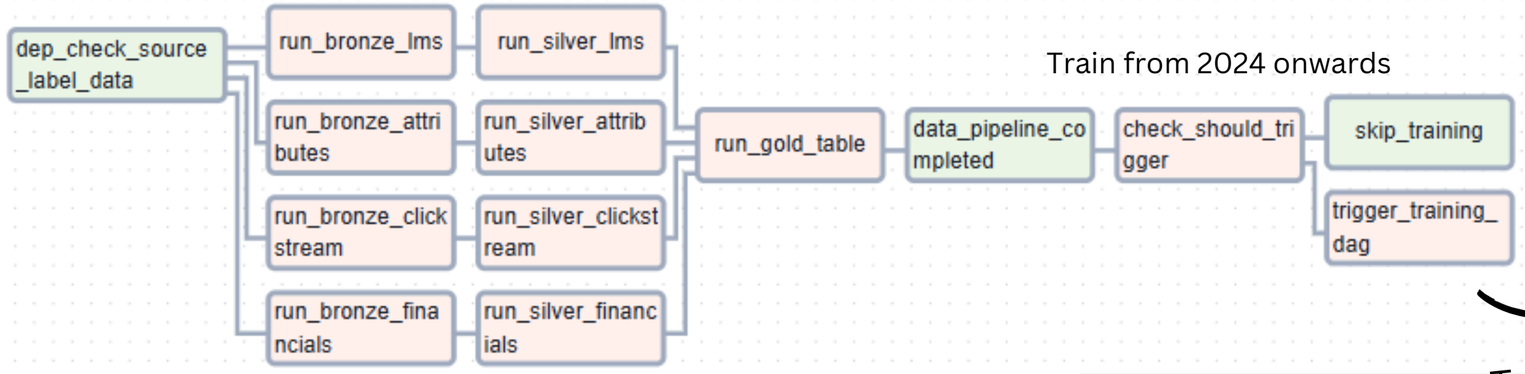


Pipeline is orchestrated using **Airflow**

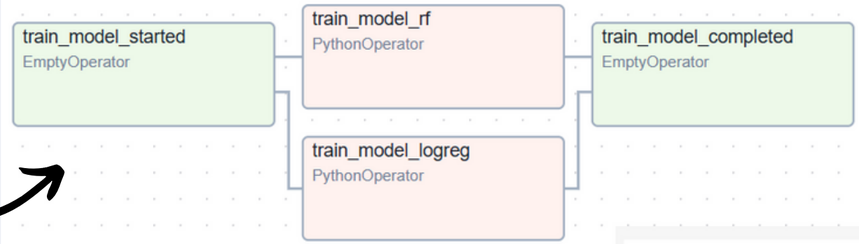
- Data - Training pipeline is triggered on the 1st day of each month
- Inference pipeline is triggered on the 7th day of each month in case the team needs to analyze the metrics to choose the champion model

AIRFLOW PIPELINE

Data Pipeline DAG

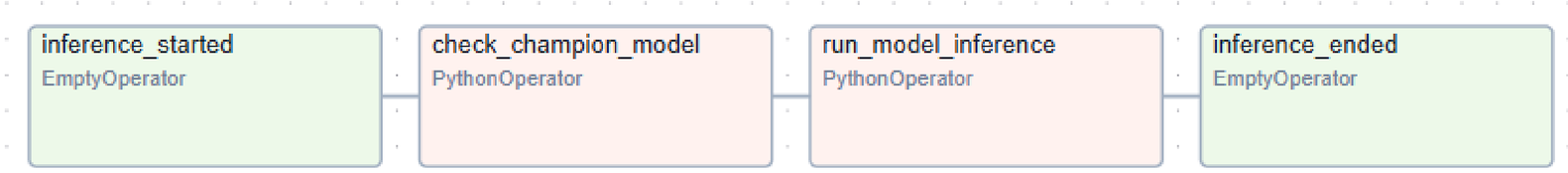


Training Pipeline DAG



Triggers

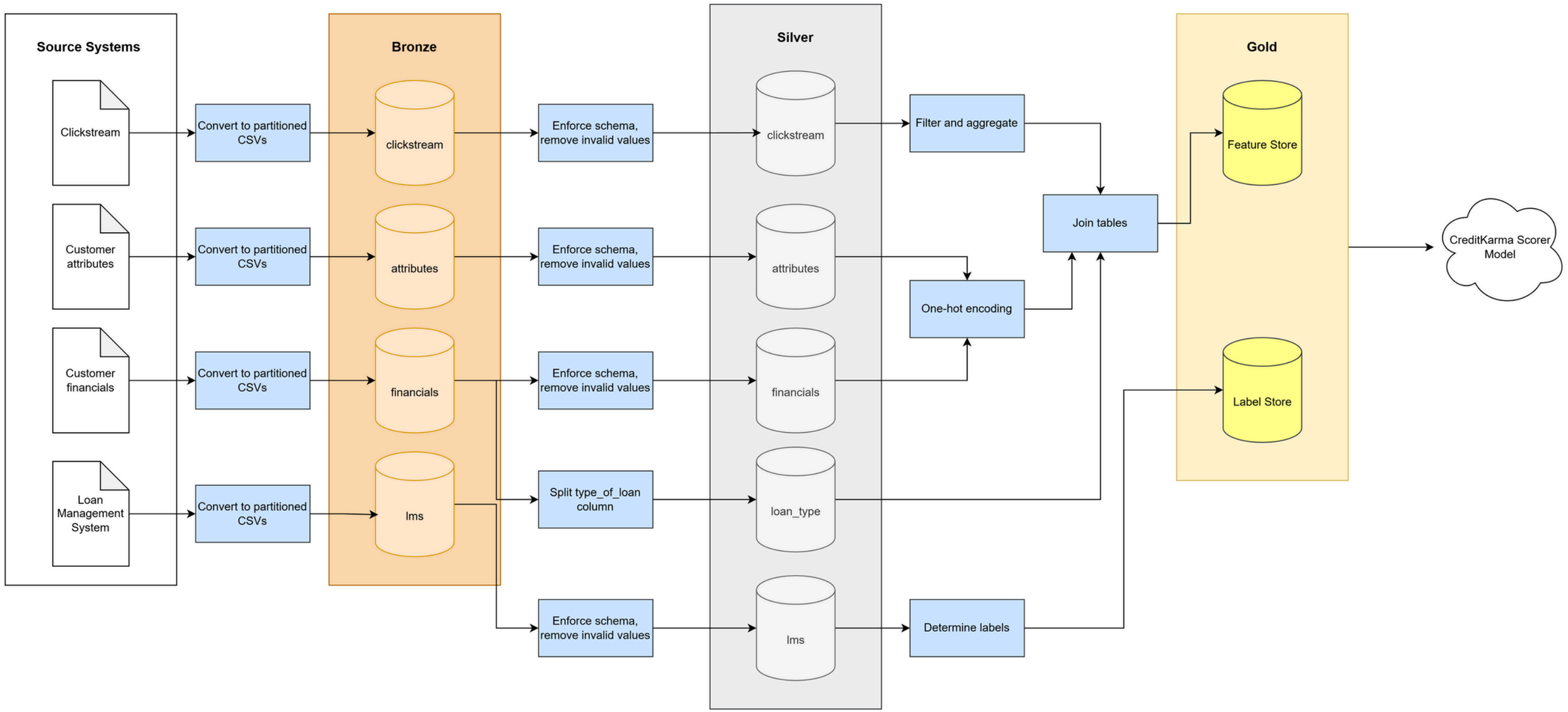
Inference Pipeline DAG



Manually triggered after new champion selected

CREDITKARMA DATAMART

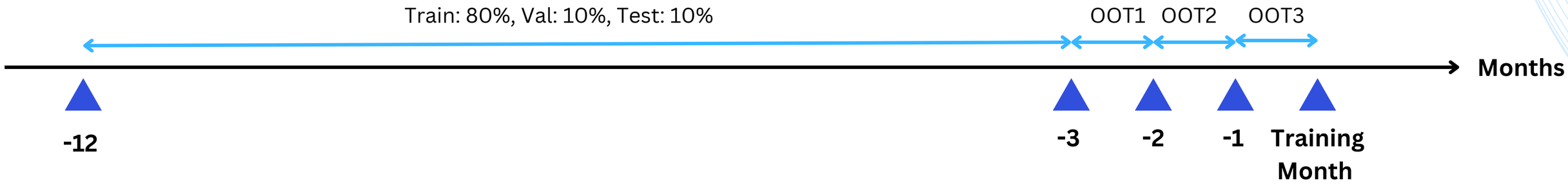
Follows the Medallion Architecture for consistency across applications and ease of maintainability



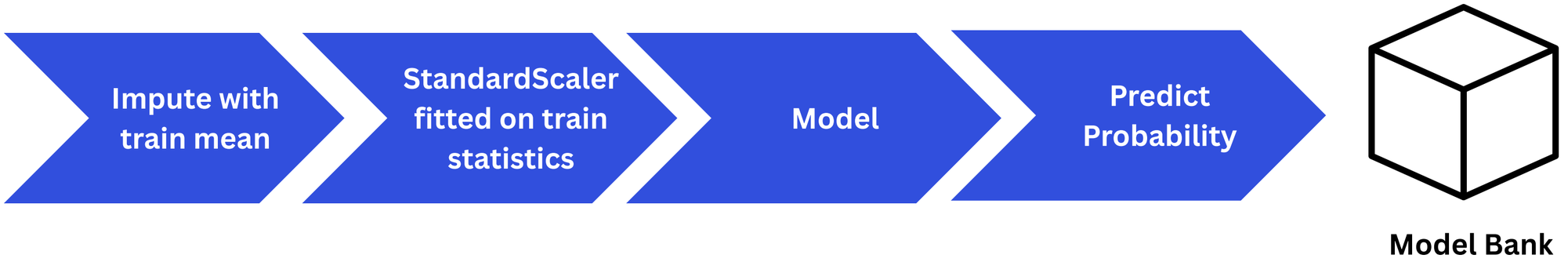
MODEL TRAINING

CreditKarma Scorer is trained on **the first day of each month** starting from 2024. The training pipeline performs automatic hyperparameter validation for **logistic regression** and **random forest classifier.** Model artifacts, alongside its data preprocessing pipeline, are logged into a model registry hosted using **mlflow**.

Dataset Split



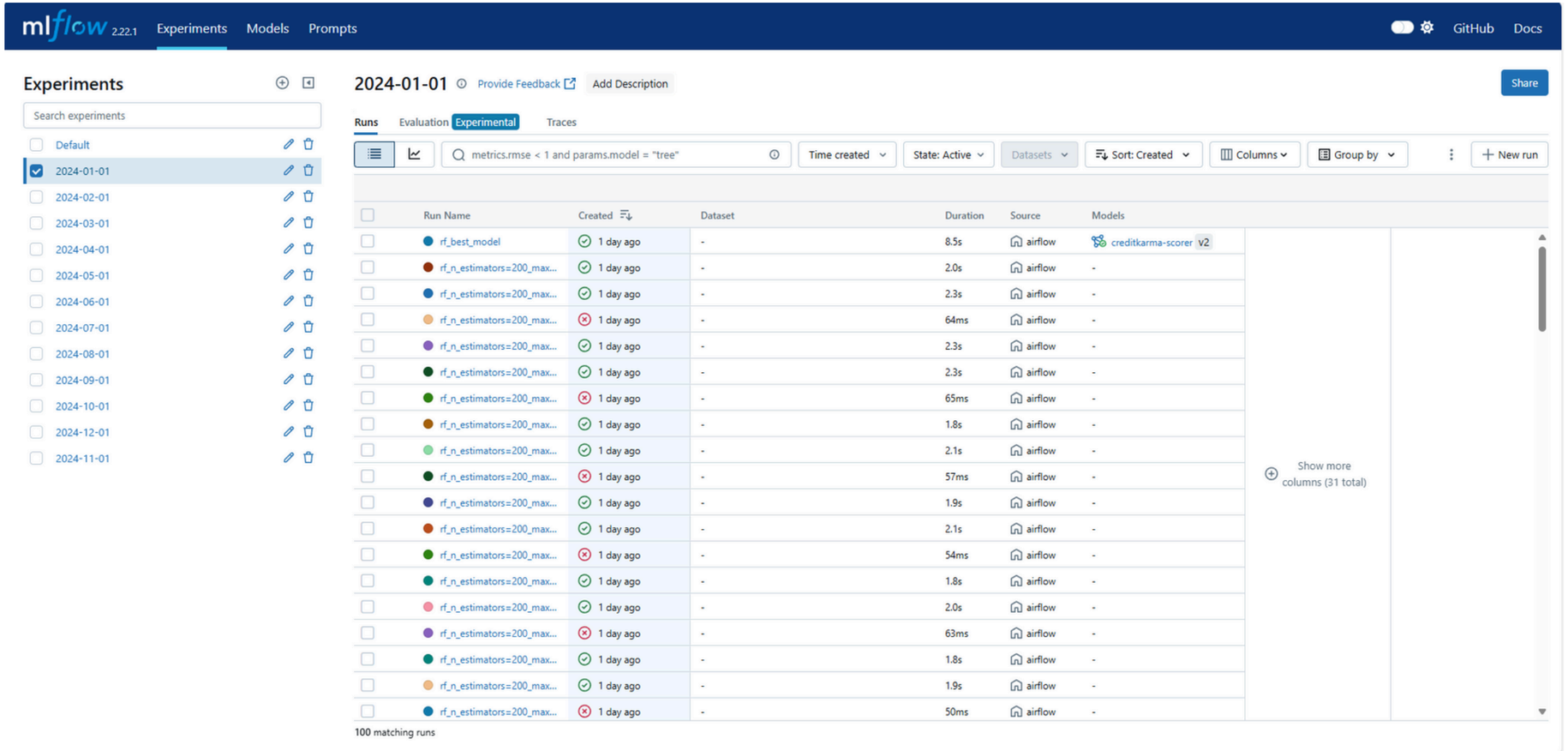
Pipeline



MODEL REGISTRY

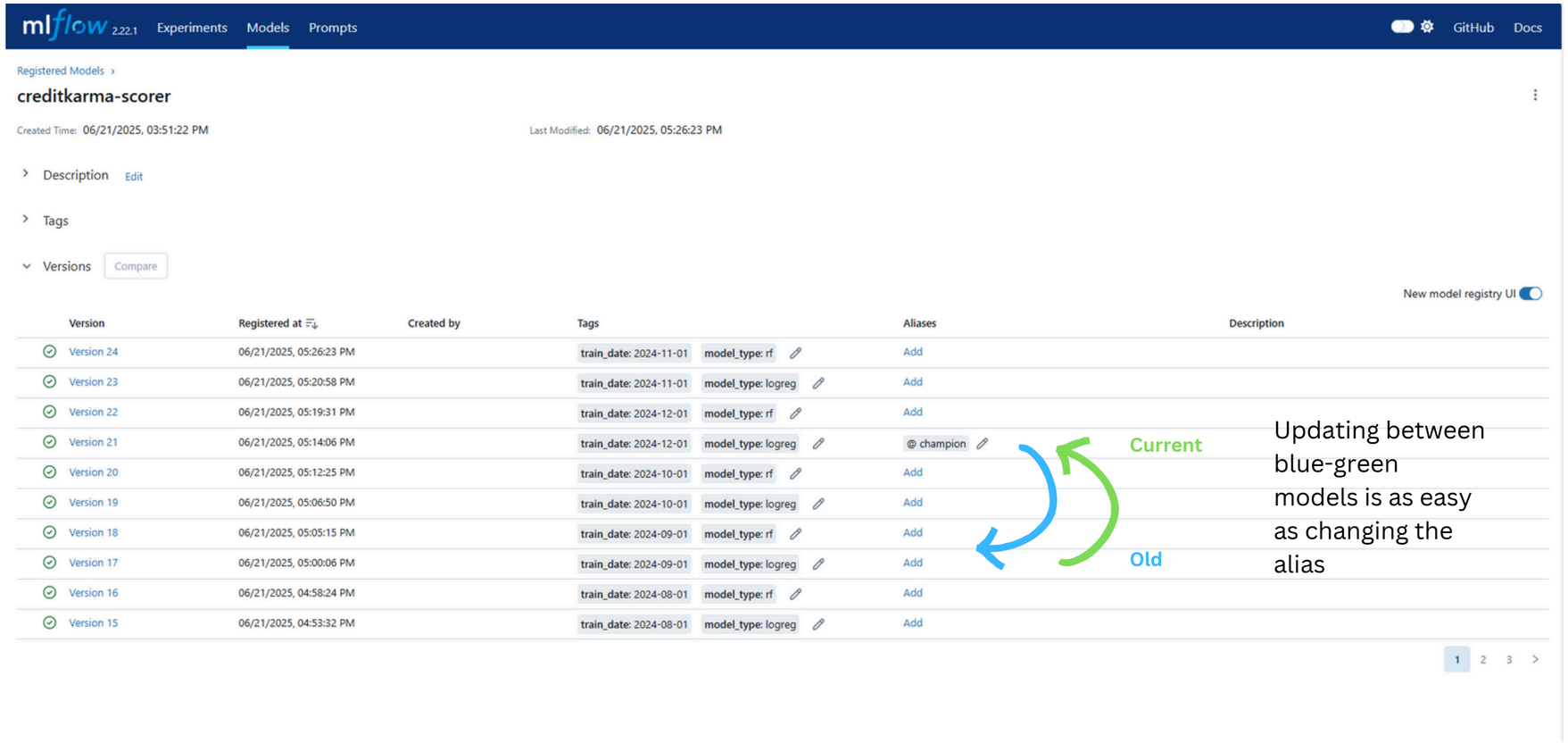
mlflow logs hyperparameter tuning results for each month of model training run and saves the best model into the **model_bank**.

GUI is served on localhost:5001 for the maintainers of CreditKarma Scorer to collaborate on.



MODEL DEPLOYMENT

Model deployment is also done via **mlflow**. CreditKarma maintainers must assign the **“champion”** alias to the best model. This allows for easy **blue-green deployment**, which is safe enough for the CreditKarma Scorer as its primary users are the internal business unit who maintain active communication with the maintainers.

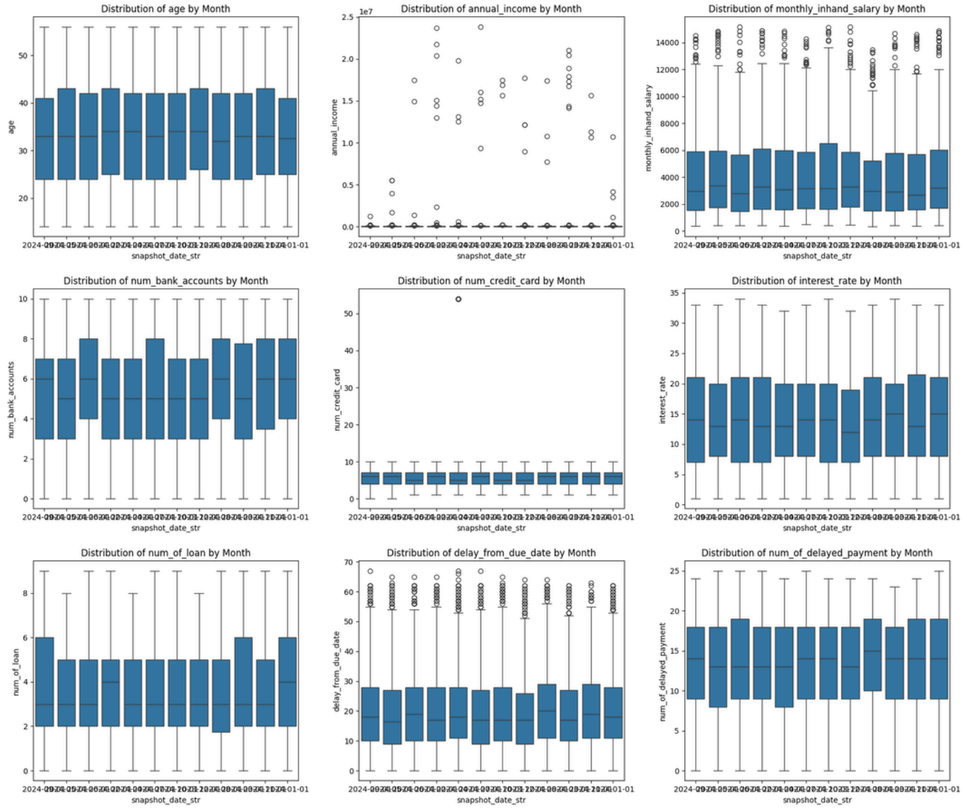


Updating between blue-green models is as easy as changing the alias

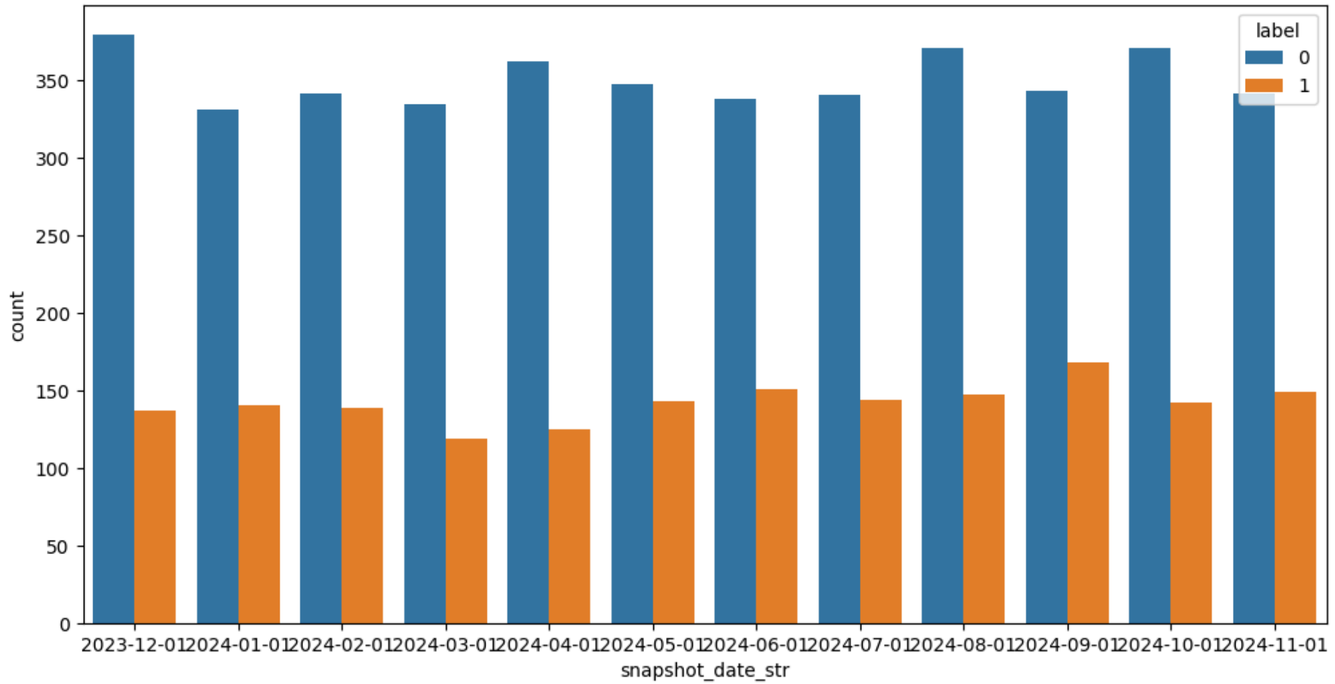
DATA MONITORING

Monitors the distribution of data over time (whether or not there is data / concept drift to be aware of).
Currently done using Jupyter Notebook, but planned to migrate to a Streamlit app in the future.

Feature box plots over months

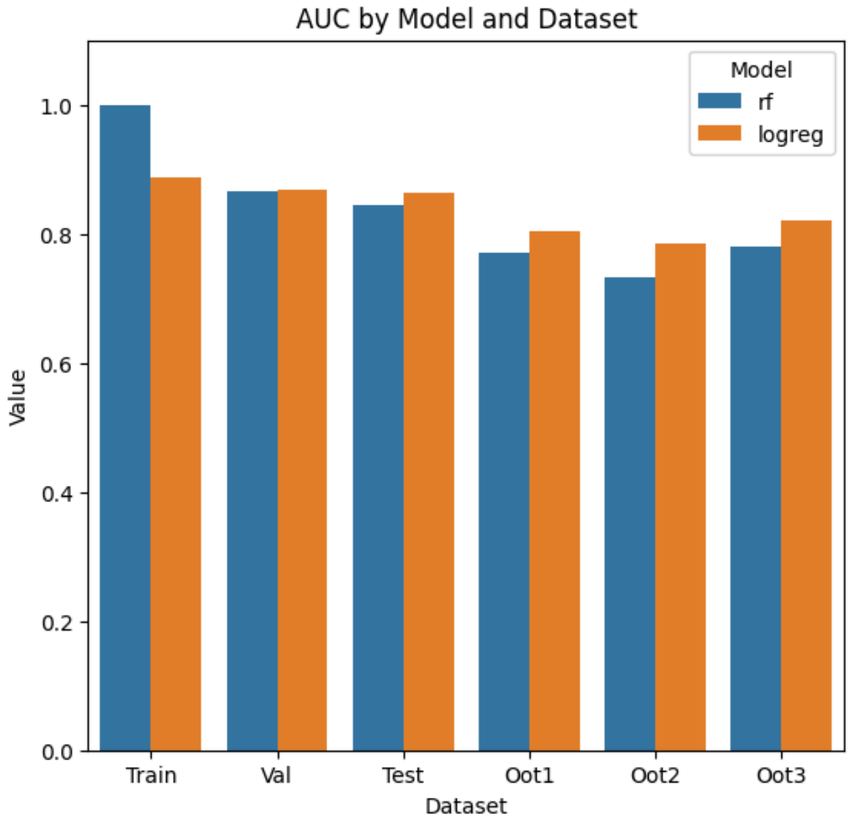


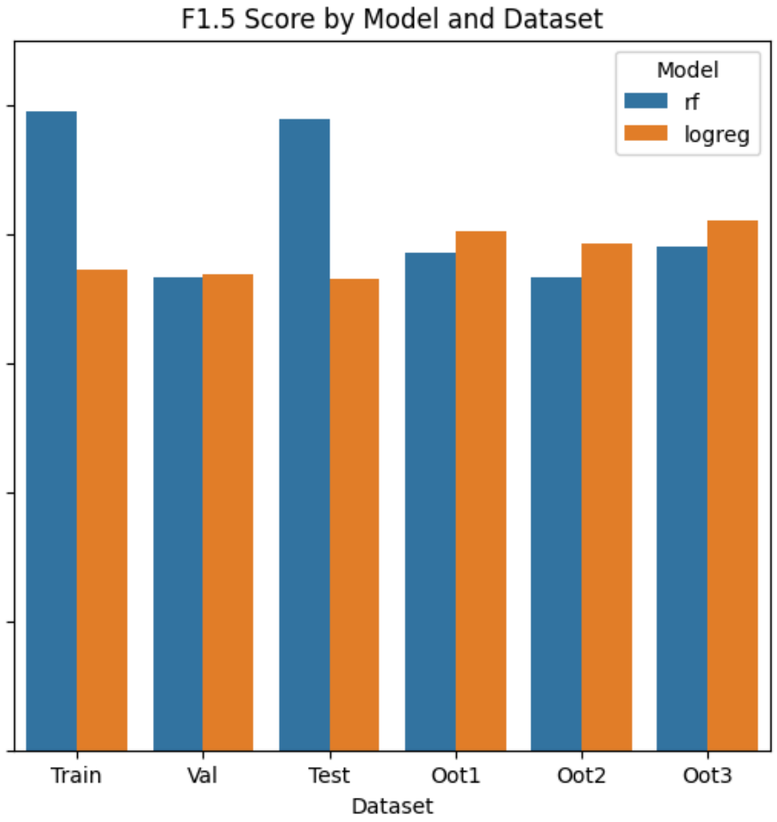
True label distribution over months



MODEL PERFORMANCE MONITORING

Monitors the performance of newly-trained models to select the best one.



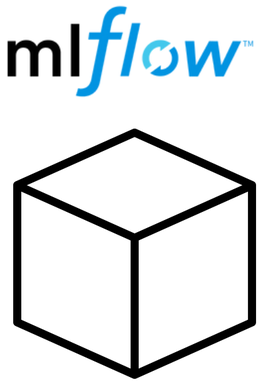


- Random Forest Classifier seems to overfit on the train data, with a steep drop in performance in every set other than Test
- Therefore, **Logistic Regression** is chosen as the model of choice for CreditKarma Scorer

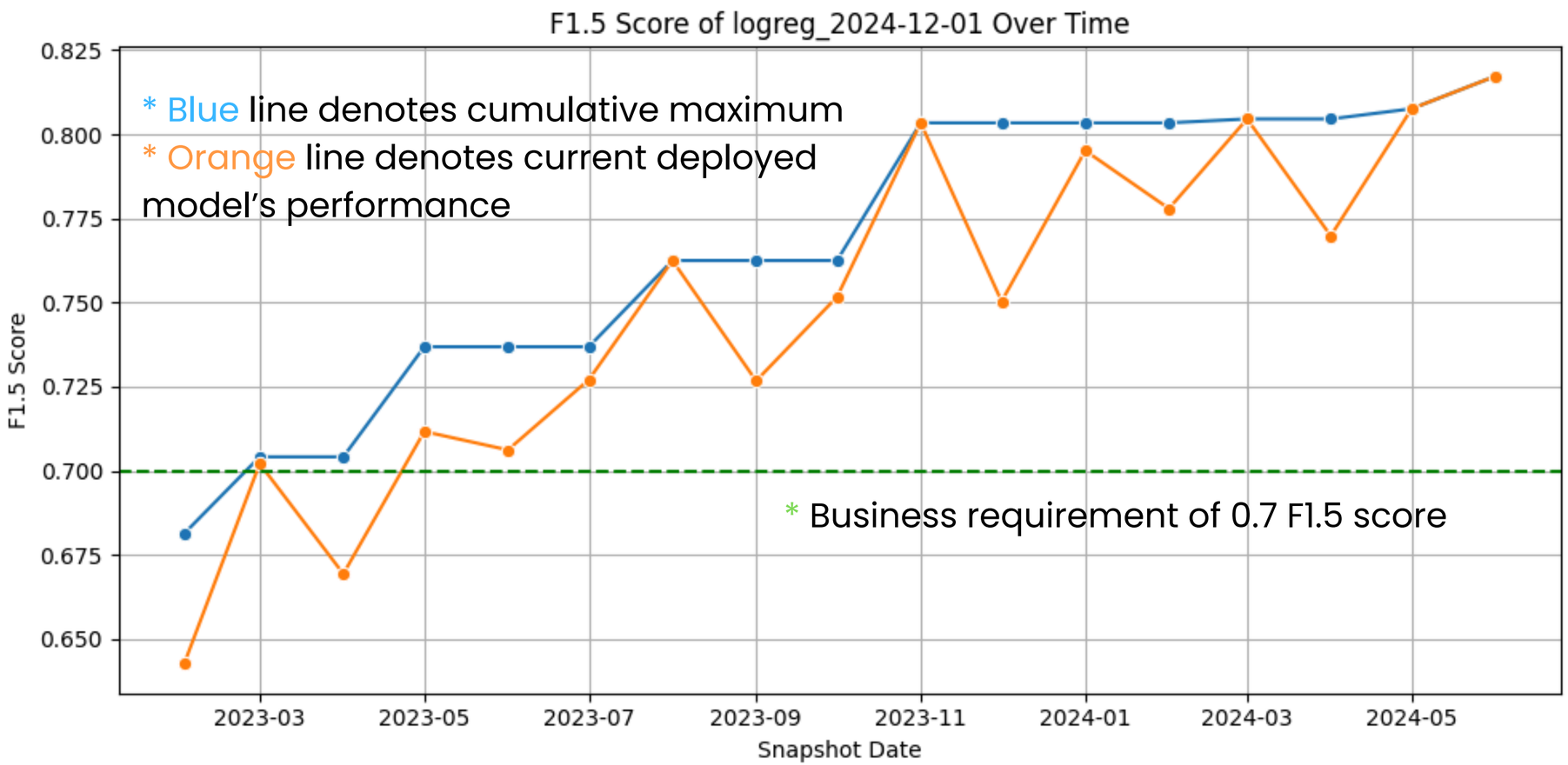
* **F1.5 Score** is chosen as the primary validation metric because we want to slightly prioritize recall over precision. Giving out loans to customers who default will result in greater loss, but we don't want to be too conservative.

MODEL STABILITY MONITORING

Monitors the stability of the deployed model version over time to alert for potentially having to deploy a new version.



Pull "champion" model



* Inference results are backfilled by Airflow when a new champion model is deployed