Part A: Project Details

1) Team Information

- **Team Name:** Drift Detectives
- **Members:**
    - Pedram Yazdinia - pedram.yazdinia@mail.utoronto.ca
    - Ben Hartwick - ben.hartwick@mail.utoronto.ca
    - Ahanaf Hassan Rodoshi - ahanaf.rodoshi@mail.utoronto.ca
    - Tausif Ahmed - tsf.ahmed@mail.utoronto.ca

2) Project Choice

- **Type:** Kaggle Competition
- **Link:** Home Credit Default Risk
- **Description Summary:** Home Credit Group expands fair access to credit for people without formal credit histories by using alternative data such as telco and transactional signals to predict repayment ability. We'll first build strong ML models to better identify capable borrowers and tailor loan amounts, terms, and schedules. Then we'll apply MLOps to make end-to-end reproducible data pipelines, automated training/CI/CD, experiment tracking, model registry, containerized deployment, and monitoring for drift and feedback-driven improvement.

3) Data Sources & Description

- **Source/Permission:** Public Kaggle competition data (anonymized; permitted for academic use).
- **Core tables (keys):**
    - application_train/test (target in train) - one row per current loan (SK_ID_CURR).

  - bureau + bureau_balance - prior third-party credits + monthly status (SK_ID_CURR).

  - previous_application - prior Home Credit apps (SK_ID_PREV, links to SK_ID_CURR).

  - POS_CASH_balance, credit_card_balance, installments_payments monthly behaviors for prior products.

  - HomeCredit_columns_description - schema docs.

4) Data Size

- Train: ~307k rows, Test: ~49k rows, positive class ~8%.

- Total download ~2.7 GB

5) Streaming/Changing Data Simulation

- **Time-staged reveals:** Feed monthly rows from panel tables (MONTHS_BALANCE) in batches to mimic new data arrival.

- **Covariate drift:** Inject controlled noise/shifts (e.g., feature distribution drift through multivariate covariate shift).

- **Prior drift:** Vary class ratio in backtests (e.g., 8% to 12%) to test calibration/thresholding.

- **Concept drift:** Policy-threshold change scenario to verify retrain triggers.

- **SMOTE:** Use **only inside training** to address imbalance (tagged runs); **not** for drift simulation.

6) EDA Plan

- Schema/dtype checks; join-key integrity.

- Target imbalance review; baseline dummy vs logistic.

- Missingness map & imputation strategy.

- Leakage checks (time/“future” features).
- Feature importance (permutation + SHAP) to prioritize stable, useful features.

7) Tech Stack

- **Data/Orchestration:** S3 (bronze/silver/gold), Glue/Athena, Step Functions/EventBridge.
- **Compute/ML: SageMaker** (Processing, Training, Pipelines, Experiments, Clarify, **Model Registry**, Endpoints).
- **Models:** Baseline logistic; XGBoost/LightGBM for tabular; optional SMOTE variant.
- **Feature Store:** SageMaker Feature Store (offline/online).
- **Monitoring:** Great Expectations, Evidently, CloudWatch; shadow + canary.
- **CI/CD & Repro:** GitHub Actions, Docker, pinned deps, IaC (CDK/Terraform).
- **Security:** IAM least-privilege, KMS encryption, signed images.

Part B: Team Charter

1) Team Goals & Success Metrics

- **Primary goal:** ship a reliable, monitored credit-risk pipeline.
- **Success Metrics**
    - **Deployed:** SageMaker endpoint + CI/CD with approval gates.
    - **Data contracts:** Blocking bad inputs; **>95%** nightly pipeline success.
    - **Model registry** with at least 2 promoted versions and rollback tested.
    - **Observability**: Drift and perf dashboards with alerts.
    - **Reproducibility**: One-command environment spin-up; experiments tracked.

    - **Team**: Each member controls at least one major project area and merges 3-4 substantive pull requests; 100% code reviewed.
    - **Presentation**: 10–12 min demo (ops flow, failures handled, lessons learned).

2) Communication Plan

- **Weekly in-person: Mon 6:00 pm (ET)** (core decisions, blockers).
- **Ad-hoc online:** Discord call later in the week as needed.
- **Response:** 24h for messages (9–7 ET).
- **Tracking:** Trello (Backlog / In Progress / Review / Done)

3) Role Distribution

- **Pedram** - *Pipeline & Orchestration* (SageMaker Pipelines, Step Functions; backup: **Ben**)
- **Tausif** - *Modeling & Feature Eng.* (baselines, XGB/LGBM, SHAP; backup: **Ahanaf**)
- **Ahanaf** - *Data Validation & Model Monitoring* (Great Expectations, Evidently, drift/calibration; backup: **Ben**)
- **Ben** - *DevOps & Deployment* (CI/CD, infra as code, endpoints, security; backup: **Pedram**)
- **Everyone:** tests, docs, code reviews, demo prep.

4) Decision-Making Process

1. Post options with a short 1–2 paragraphs in Discord or Trello.
2. 10-minute discussion on pros/cons + impact on scope/timeline.
3. Vote; if tied, the area lead decides; if cross-area, ***Pedram*** (pipeline lead) is the tie-breaker.
4. Record the decision and rationale on the issue and proceed.

5) Problem Resolution Protocols

- **Disagreements:** schedule a 30-minute call within 24 hours; each presents their view + evidence; seek a compromise.

- **Poor communication / silent:** DM after 48h; loop in another member at 4 days.

- **Missed deadlines:** 24h advance notice required; redistribute tasks in meeting; repeated misses trigger scope cut or role swap.

- **Unequal contribution:** weekly check-ins with task burndown.

6) Quality Bar & Deliverables

- **Tests:** unit and integration in CI; target **≥70%** coverage on core libs.

- **Reproducibility:** Setup and train work from a clean repo.

- **Docs:** README (runbook), ARCHITECTURE.md (data flow, DAG), OPS.md (monitoring/alerts, rollback).

- **Demo:** failure-injection scenario (bad schema + drift) and recovery.

7) Timeline

- **W1–W3:** data contracts, baseline model, repo scaffolding, CI.

- **W4–W6:** feature pipelines, experiments tracking, registry.

- **W7-W8:** deployment, monitoring, canary/shadow.

- **W9:** hardening, demo, report.