Finshield – Progress Report

Project: Alternative Credit Risk Scoring Using Synthetic Data and Transformer-Based Modeling

Status: Ongoing Development (Hackathon Prototype Stage)
GitHub Repository: https://github.com/SuryaAshish1404/Finshield_Hackathon

1. Overview
The project aims to develop a privacy-compliant, explainable, and scalable credit scoring system using:
- Synthetic data generation inspired by Banksformer.
- Transformer-based tabular modeling approaches such as TabTransformer, SAINT, and TabLLM.
- Fidelity and fairness validation before deployment.

The primary objective is to enable financial inclusion for underbanked populations while maintaining high predictive accuracy and compliance with data privacy regulations.

2. Completed Components
A. Data Ingestion and Preparation – 100%
Integrated Home Credit Default Risk and PaySim datasets; complete ingestion pipeline; missing data handling.

B. Synthetic Data Generator – 95%
Transformer-based synthetic data generator implemented; preserves feature distributions; supports missing value imputation.

C. Transformer-Based Modeling – 90%
Completed training pipeline for TabLLM-style architecture; multi-dataset CLI; checkpoint saving/loading.

D. Inference Pipeline – 85%
CLI inference script developed; generates synthetic rows; saves to CSV.

3. Component Completion Status

Component	Completion %	Notes
Fidelity Testing	40%	Basic setup, validation not yet integrated
Fairness Auditing	35%	Fairness metrics framework prepared
Feature Engineering	60%	Basic handling implemented
Error Handling in Inference	70%	Tokenizer–schema mismatch handling ongoing
Evaluation Metrics	50%	Banking-standard metrics pending

4. Pending Tasks
- Implement full fidelity and accuracy validation for synthetic data.
- Automate feature engineering with domain-specific metrics.
- Add SHAP-based explainability layer.
- Complete fairness auditing process.
- Integrate evaluation metrics into training and inference pipelines.
- Resolve token map mismatches in inference.
- Develop Streamlit dashboard for demo and deployment.

5. Achievements to Date
- Functional synthetic data generation pipeline established.
- Trained transformer-based scoring models on large datasets.
- Functional CLI for training and inference.
- Infrastructure for scalable dataset handling.
- Research-aligned approach with Banksformer and TabLLM methodologies.

6. Next Steps
- Stabilize inference by resolving token mismatches.
- Enhance feature engineering automation.
- Add fidelity, fairness, and explainability layers.
- Build Streamlit-based user interface.
- Perform stress testing and temporal validation.

7. Current Readiness

- Backend (Model + Data): ~85% complete.
- Validation + Fairness Layers: ~45% complete.
- UI + Deployment: Not yet started.
- Overall Progress: ~65% complete.

A functional prototype is available for the hackathon, demonstrating synthetic data generation and transformer-based scoring.