📄 CreditPathAI — Project Progress Report

Overview

CreditPathAI is a machine-learning driven credit risk and loan recovery intelligence system designed to predict borrower default probability and support collection agents with actionable insights.
The project focuses on building a transparent, scalable, open-source pipeline using Python and modern ML tools.

Up to this stage, the system has successfully completed:

- Dataset acquisition and consolidation
- Data cleaning and integrity checks
- Feature engineering with domain-driven transformations
- Leakage detection and correction
- Exploratory analysis
- Baseline model development with Logistic Regression
- Full evaluation and interpretation of model behavior

The foundation for advanced modeling and next modules is now secure and reliable.

Dataset Summary

The dataset originates from Kaggle's loan repayment/default records and contains borrower-level credit, income, payment, and behavioral indicators. After cleaning and preparing the data, the final dataset contains:

- **9,578 rows**
- **24 cleaned and validated features**
- Fully numeric, leak-free, model-ready structure

Key original fields include:

- **fico:** credit score
- **dti:** debt-to-income ratio
- **installment:** monthly payment
- **revol_bal / revol_util:** credit card balance and utilization
- **log.annual.inc:** income proxy
- **days.with.cr.line:** credit history length
- **inq.last.6mths:** recent credit inquiries
- **delinq.2yrs:** past delinquencies
- **purpose:** reason for the loan

- **credit.policy:** lending eligibility flag

Feature Engineering

To improve predictive power, several meaningful features were engineered.

Feature	Purpose	Description
repayment_velocity_proxy	Reliability measure	Ratio combining FICO & DTI to estimate borrower repayment behavior
approx_credit_limit	Derived limit	Approximates credit card limit from balance & utilization
credit_utilization	Risk indicator	Measures percentage of revolving credit used
annual_inc	Income recovery	Derived from log income (exp transformation)
debt_to_income_calc	Payment burden	Installment-to-income calculation for real DTI
credit_age_years	Stability marker	Converts credit days into years
has_recent_inq	Stress signal	Recent credit inquiries → higher risk
has_past_delinquency	Past behavior	Flags borrowers with delinquency history
purpose_*	One-hot vectors	Captures loan-type based risk differences
target_default	Target label	Final binary variable used for supervised ML

These engineered features were validated through distribution analysis and correlation checks.

Data Quality & Leakage Correction

During early modeling, the system detected **perfect prediction artifacts**, caused by an original field:

```
not_fully_paid
```

This column was a direct duplicate of the target (`target_default`), causing the model to “cheat” and achieve false perfect accuracy.

Actions taken:

- Removed **not_fully_paid** entirely
- Converted all **True/False** and mixed-type columns into **numeric 0/1**
- Scanned for perfectly predictive columns

- Dropped fields with **|correlation| ≥ 0.98** with the target
- Rechecked the dataset with information-leakage tests
- Rebuilt a fully clean dataset

The result is a **trusted**, leak-free dataset suitable for real modeling.

Exploratory Data Analysis

EDA was performed using histograms, boxplots, correlation heatmaps, and categorical performance charts.

Key observations:

- Defaulting borrowers tend to have **higher credit utilization**, **lower FICO**, and **younger credit history**.
- Loan purpose categories show natural risk variation.
- The target variable is imbalanced, which requires special handling during modeling.
- No obvious outliers were removed, but skew patterns were noted for future model tuning.

The EDA notebook is complete and reproducible.

Baseline Model — Logistic Regression

A clean baseline model was trained using:

- Stratified train-test split (75/25)
- Feature scaling (StandardScaler)
- Balanced class weighting to counter imbalance
- Logistic Regression (max_iter=2000)

Model Performance (from Colab run)

Metric	Value
Accuracy	**0.6459**
Precision	**0.2408**
Recall	**0.5640**
F1 Score	**0.3375**
AUC-ROC	**0.6767**

Interpretation

- Accuracy is modest due to class imbalance (normal).

- Precision is reasonable for a simple baseline.
- **Recall is strong (56%)**, meaning the model detects more than half of real defaulters — excellent for Logistic Regression.
- AUC-ROC near **0.68** indicates meaningful ranking ability.

This is a solid and realistic foundation before advancing to more powerful models like XGBoost or LightGBM.

Current Project Status

Your pipeline now includes:

- ✓ Clean, validated, leak-free dataset
- ✓ Documented feature engineering
- ✓ Exploratory data analysis
- ✓ Baseline Logistic Regression model
- ✓ Clear understanding of model behavior
- ✓ Ready-to-extend notebooks for further modeling

The system is now ready to move into advanced ML techniques, hyperparameter tuning, and model comparison.
