Name : Ittehad Ahmed Tausif
E1 ID : E1-23201642-MHUN9UHK
Wings : Intelligence & Perception, Systems & Application, Development

Project 01: AI-Powered Loan Default Prediction System for Banking and Fintech

AI Project Management Lifecycle:

1. Project Initiation and Business Understanding
2. Data Strategy and collection
3. Data Preprocessing and Feature Engineering
4. Model Selection and Development
5. Model Training, Evaluation, and Validation
6. Deployment and Integration
7. Monitoring and Maintenance

[Week 01 Report]

Date: November 18, 2025

1. Project Initiation and Business Understanding

1.1 Problem Definition: The Modern Credit Risk Challenge

The core operational imperative for any lending institution—whether a Tier-1 bank or a Fintech disruptor—is the accurate quantification of credit risk. In a financial landscape defined by volatile interest rates and rapid digitization, traditional credit scoring mechanisms have become existential liabilities.

Historically, risk assessment relied on the "5C" model (Character, Capacity, Capital, Collateral, Conditions) implemented via rigid logistic regression scorecards. These legacy systems primarily utilize static, backward-looking data and fail to capture the complex, non-linear behavioral patterns of modern borrowers (Zhou et al., 2022). This rigidity results in two costly types of errors:

- **Type I Errors (False Positives):** Incorrectly classifying creditworthy borrowers as high risk. This causes lost revenue and market share attrition to competitors with more sophisticated underwriting.
- **Type II Errors (False Negatives):** Classifying defaulters as creditworthy. This leads to direct capital destruction through Non-Performing Assets (NPAs) and increased provisioning requirements under standards like IFRS 9.

Furthermore, regulatory pressure is intensifying. Frameworks like the **EU AI Act** now designate credit scoring as "high-risk," mandating that systems be not only accurate but also explainable and free from systemic bias (KPMG, 2024). The problem, therefore, is to engineer a regulatory-compliant AI system that outperforms traditional models in predicting default without becoming an opaque "black box."

1.2 Business Objective: Strategic Alignment

The deployment of this AI system is a strategic initiative designed to drive capital efficiency and operational excellence.

- **Minimizing Expected Credit Loss (ECL):** The primary objective is to reduce the NPA ratio. By employing Gradient Boosting Machines (e.g., XGBoost) or Deep Learning, the bank aims to identify subtle delinquency patterns that linear models miss. A precise Probability of Default (PD) calculation directly reduces the capital reserves locked away as provisions (Ptak-Chmielewska, 2022).
- **Operational Efficiency (Straight-Through Processing):** Legacy underwriting is labor-intensive. An AI-driven engine enables Straight-Through Processing (STP), allowing instant approvals for low-risk applicants and instant rejections for high-risk ones. This allows human underwriters to focus solely on complex, borderline cases, drastically reducing the "Time to Yes" (ET Edge Insights, 2024).
- **Financial Inclusion:** Traditional models often reject "thin-file" customers who lack a credit

history but are financially responsible. By analyzing alternative data points, the AI model can score these underserved segments, expanding the total addressable market without proportionally increasing risk (Sullivan, 2025).

1.3 Feasibility Assessment

- **Technical Feasibility:** The project is technically viable. The bank possesses the necessary historical loan performance data and transaction logs. Research indicates that machine learning models, particularly XGBoost, consistently outperform traditional logistic regression in terms of accuracy and discriminatory power (Khatib & Al-Naji, 2024). Furthermore, Explainable AI (XAI) techniques like SHAP values resolve the "black box" regulatory concern by providing transparent decision logic (Grant Thornton, 2023).
- **Operational Feasibility:** The primary challenge is integrating a Python-based AI model with a legacy Core Banking System (often running on COBOL mainframes). This will be addressed using an API-first architecture, where the risk engine is decoupled from the core ledger. Success relies on change management—training loan officers to interpret and trust AI-assisted risk scores.
- **Financial Feasibility:** While the initial investment in data engineering is significant, the ROI is compelling. The cost of a single default (principal loss + recovery costs) far outweighs the cost of computing power. Even a modest 5-10% reduction in the default rate translates to millions in preserved capital (ET Edge Insights, 2024).

---

2. Data Strategy and Collection

Success in banking AI depends less on the algorithm and more on the integrity of the data pipeline. Our strategy is built on a rigorous ethical framework derived from the **Epoch One Data Ethics Guideline** to ensure compliance and trust.

2.1 Strategic Data Governance and Ethics

We will adhere to the following principles to manage legal and reputational risks:

- **Purpose & Legality:** Data collection is strictly limited to assessing creditworthiness. We rely on "Contractual Necessity" and "Legitimate Interest" as our legal bases (Epoch One, n.d.). Every data field must map directly to the risk assessment function.
- **Data Minimization:** We will collect only the minimum data necessary for predictive accuracy. "Data hoarding" is strictly prohibited. For instance, we will exclude social media data unless it has proven, non-discriminatory predictive value and explicit consent is obtained (Epoch One, n.d.).
- **Fairness & Bias Prevention:** Historical lending data often contains embedded biases. We will explicitly exclude protected characteristics (race, religion, gender) from training datasets.

Furthermore, we will screen for "proxy variables"—features like zip codes that may correlate with demographics—to prevent indirect discrimination (Consumer Financial Protection Bureau, 2014).

- **Transparency & Consent:** Borrowers must be informed of how their data determines their score. We will implement mechanisms to provide "Counterfactual Explanations," informing rejected users exactly which financial behaviors led to the decision (European Commission, n.d.).
- **Ethical Oversight (DPIA):** Given the high impact of credit denial on individuals, we will conduct a mandatory **Data Protection Impact Assessment (DPIA)**. This process will proactively identify risks regarding bias, security, and privacy before any model goes into production (Epoch One, n.d.).

2.2 Real-World Data Collection Strategy

2.2.1 Internal Data Sources (First-Party Data)

The bank's proprietary data is the most valuable asset for prediction.

- **Application Data:** Income, employment duration, loan purpose, and residential status from the Loan Origination System (LOS).
- **Transaction Logs:** High-frequency data from the Core Banking System, such as average daily balance, cash flow volatility, and overdraft frequency. These are powerful behavioral indicators.
- **Repayment History:** Historical performance on previous loans, including Days Past Due (DPD) and cure rates.

2.2.2 External Data Sources (Third-Party Data)

- **Credit Bureau Reports:** Aggregated history from agencies (Equifax, Experian) providing FICO scores, active trade lines, and recent hard inquiries (Defi Solutions, n.d.).
- **Trended Credit Data:** Unlike static snapshots, this reveals the trajectory of debt (e.g., is the borrower paying down balances or making minimum payments while debt grows?).
- **Public Records:** Bankruptcy filings and court judgments for identifying severe derogatory history.

2.2.3 Alternative Data (Strategic Innovation)

To address "thin-file" applicants, we may incorporate alternative sources, subject to strict ethical review:

- **Utility & Telecom Data:** Consistent payment of electricity or mobile bills demonstrates "willingness to pay."

- **Open Banking APIs:** With user consent, accessing transaction data from competitor banks via aggregators (e.g., Plaid) provides a holistic view of the applicant's financial health (Sullivan, 2025).

2.3 Handling Ethical Problems in Data Collection (Kaggle vs. Reality)

Using public datasets like those from Kaggle introduces specific risks that must be managed to simulate a professional environment.

- **The "Provenance Gap":** Public datasets often lack clear consent metadata. In a real project, using data without verifiable consent violates the "Lawful Purpose" guideline. For this project, we treat the data as "Simulated/Synthetic," acknowledging that in production, every record would require a traceable consent trail.
- **Re-identification Risk:** Even anonymized datasets can be compromised if they contain unique combinations of attributes (e.g., specific income + zip code). We will apply **Pseudonymization** and **k-anonymity** techniques, generalizing quasi-identifiers (e.g., converting exact age to an age range) to protect individual privacy (El Emam et al., 2012).
- **Synthetic Data Generation:** To mitigate privacy risks during development, we will prioritize the use of synthetic data—artificial data that mimics the statistical properties of real customers without containing any Personal Identifiable Information (PII). This allows developers to test models aggressively without exposing sensitive client data (American Bankers Association, 2023).

2.4 Feature Engineering Strategy

Raw data must be transformed into meaningful signals to be predictive.

- **Debt-to-Income (DTI) & Payment-to-Income (PTI):** Fundamental capacity metrics.
- **Velocity Features:** Measuring the speed of behavior changes, such as a sudden spike in credit inquiries over 30 days, which often signals financial distress.
- **Trended Utilization:** Calculating the slope of credit usage over 6 months. A borrower whose utilization is creeping up month-over-month is a higher risk than one with stable high utilization (Defi Solutions, n.d.).
- **Interaction Effects:** Capturing non-linear risks, such as the compounded risk of high utilization combined with variable income, which tree-based models (XGBoost) are particularly adept at identifying (Khatib & Al-Naji, 2024).

References

American Bankers Association. (2023). *Harnessing synthetic data: Advancing innovation with privacy-enhanced insights*.

Consumer Financial Protection Bureau. (2014). *Using publicly available information to proxy for unidentified race and ethnicity*.

Defi Solutions. (n.d.). *How trended data credit reports improve loan origination decisions*.

El Emam, K., et al. (2012). *De-identification methods for open health data: The case of the Heritage Health Prize Claims Dataset*. Journal of Medical Internet Research.

Epoch One. (n.d.). *Epoch One Data Ethics Guideline for Data Collection in ML Projects*.

ET Edge Insights. (2024). *How AI drives precision in borrower management and cuts NPAs for lenders*.

European Commission. (n.d.). *Rules for business and organisations: Automated decision-making*.

Grant Thornton. (2023). *Fair lending in the digital age: Managing model risk*.

Khatib, T., & Al-Naji, A. (2024). *Technical feasibility of machine learning for loan default prediction*. MDPI.

KPMG. (2024). *Setting the ground rules: The EU AI Act and its impact on financial services*.

Ptak-Chmielewska, A. (2022). *Banking credit risk project initiation: Probability of Default (PD) estimation*. MDPI.

Sullivan, T. (2025). *6 types of alternative credit data for better loan decisions*. Plaid.

Zhou, H., et al. (2022). *Identifying high-risk customers: From 5C to AI*. PMC.