

Project 01: AI-Powered Loan Default Prediction System

[Week 01 Report]

Name: Ahetasham Shifat

E1 ID: 23201216

Wings: Intelligence & Perception, System & Application, General

1. Project Initiation and Business Understanding

1.1 Problem Definition:

Banks and fintech companies often struggle to identify which customers are likely to default on their loans. This leads to financial loss, poor credit decisions, and unnecessary risks.

Our goal is to build an AI model that can analyze a customer's financial and personal information and predict the likelihood of loan default. This prediction will help the bank make safer, faster, and more accurate lending decisions.

1.2 Business Objective:

The AI company aims to support the bank by improving its loan approval process, reducing default rates, and increasing overall profitability.

For the bank, an accurate prediction system means:

1. Fewer risky loans
2. Better customer selection
3. More efficient operations and
4. Stronger long-term financial stability.

By aligning the AI solution with the bank's goals, both organizations benefit from improved decision-making and reduced financial losses.

1.3 Feasibility Assessment:

From a technical point of view, this project is very feasible.

1. Machine learning models like Logistic Regression, Random Forest, or Gradient Boosting are commonly used for loan default prediction.
2. The required data (income, credit history, loan amount, repayment records, etc.) is usually available in banking systems.
3. The processing power needed is moderate, and many open-source tools already exist to help build such models.

Overall, both the data requirements and technical workload make the project achievable within a realistic timeline.

2. Data Strategy and Collection:

To write this part, the Epoch One Data Ethics Guideline is fully considered.

How We Would Collect Data in a Real Project

In a real banking project, the data would come from secure internal systems such as:

1. Loan application forms
2. Customer financial history
3. Previous repayment records
4. Credit bureau reports
5. Verified documents like income statements or employment details

All data would be collected legally, with proper consent, and only for the purpose of building the loan default prediction model.

We would avoid collecting unnecessary or sensitive information unless required by law or directly relevant to risk assessment. Personal identifiers would be removed or converted into anonymous IDs before the data enters the machine-learning pipeline.

Using Tax Indicators Without a Training Dataset: Banks can also safely use applicant-provided tax documents (with consent) by converting them into simple numerical features like filing consistency, late filings, income-to-tax alignment, and declared income level. These features can adjust baseline ML risk scores, support rule-based checks for borderline cases, and be gradually collected (anonymized) to build an internal dataset for future model training, all without exposing private tax data.

Using Kaggle Data and Addressing Ethical Issues:

For practice and modeling, we are using the Loan Default Dataset from Kaggle. However, Kaggle datasets sometimes include information that may be sensitive, such as usernames or personal details from the uploader.

To stay ethical and follow data-privacy standards, we would:

1. Review the dataset carefully and remove any columns that contain personal or identifying information.
2. Anonymize or mask any values that can reveal someone's identity.
3. Use the dataset strictly for learning and model development, not for real-world decisions.
4. Ensure that no private information is stored, shared, or used beyond the project scope.

This keeps our work aligned with the Epoch One Data Ethics Guideline while allowing us to train and test our model responsibly.