Who Pays the Price?

An Explainable & Responsible AI System for Credit Risk Decisions

Project Motivation

Machine Learning models are increasingly used to automate high-stake decisions such as loan approvals, hiring, admissions and even resume screening. These systems are often evaluated using technical metrics like accuracy or precision. However, such metrics do not answer a more important question

When a model makes mistakes, who is harmed?

In real – world decision systems, not all the errors are similar or equal. Some mistakes cause financial loss to institutions, while others directly affect individuals by denying opportunities. This project focuses on understanding and quantifying the human cost of model errors, rather than optimizing accuracy alone.

Project Objective

The goal of the project is to design and evaluate an AI system that:

- Predicts credit risk using real-world data
- Goes beyond accuracy to analyze error impact
- Identifies which groups bear the cost of mistakes
- Uses explainable AI techniques to make important decisions

Dataset:

German Credit Risk Dataset: Sourced from UCI Machine Learning Repository

This data set includes an official cost matrix, making it ideal for cost- sensitive analysis

In this project, model errors are not treated equally.

Error Type	Meaning	Who pays the Price
False Positive	Approving a risky applicant	Financial institution
False Negative	Rejecting a safe applicant	Human applicant

Rejecting a low-risk applicant represents direct human harm, including lost opportunity, financial stress, and delayed progress. The project therefore focuses heavily on false positives and false negatives, rather than accuracy alone.

Data Preparation:

- Parsed raw space-separated academic data correctly
- Assigned meaningful column names
- Converted the target variable into binary risk label
- Created an age_group feature to analyze fairness
    - Young : age < 25
    - Older : age > = 25

Final Feature Matrix:

- 49 features after encoding

- 1000 rows

Model Selection

Chosen Model: Logistic Regression

Logistic regression was selected because:

- It is interpretable
- Commonly used in financial risk modeling
- Suitable for explainable and responsible AI

Two versions evaluated:

1. Logistic Regression (unscaled features)
2. Logistic Regression (scaled features)

Model Evaluation

Baseline Model (Unscaled features)

Confusion Matrix:

158	17
38	37

Key Metrics:

- Accuracy: 78%
- Recall (high risk) : 49%
- False Positives (human harm): 17
- False Negatives (financial risk) : 38

Scaled Model

Confusion Matrix:

151	24
38	37

Key Metrics:

- Accuracy: 75%
- Recall (high risk) : 49%
- False Positives (human harm): 24
- False Negatives (financial risk) : 38

Model Comparison & Trade-offs:

Feature scaling increased the number of false positives (meaning more low-risk applicants were incorrectly rejected) without improving the detection of high-risk applicants.

Key Insight: Technical improvements do not always reduce harm. In some cases, they increase it.

The unscaled model was therefore selected as the preferred model, as it balanced the accuracy and human impact better.

Who pays the price?

False positives were analyzed across age groups to determine whether harm was evenly distributed.

Findings:

- Model errors were not evenly distributed
- Younger applicants experienced a higher false rejection rate
- Accuracy alone masked these disparities

This demonstrates how automated systems can unintentionally disadvantage certain groups if fairness is not explicitly evaluated.

Cost – Sensitive perspective

The dataset includes an official cost matrix:

Actual \ Predicted	Good	Bad
Good	0	0
Bad	5	1

This helps us to understand that:

- Approving a bad applicant is costly
- Rejecting a good applicant still causes harm

In this project, I used this framework to emphasize decision impact, not just prediction accuracy.

Based on the findings, the following are recommended:

- Use AI predictions as decision support, not as final authority.
- Flag low-confidence predictions for human review
- Monitor false rejection rates across demographic groups
- Regularly audit models for fairness drift

Tools & Technologies:

- Python
- Pandas & Numpy – data processing
- Scikit-learn – modeling & evaluation
- Matplotlib & Seaborn - Visualization
- VS Code