L02: Logistic Regression

Classification with Probability Estimates

Methods and Algorithms

Spring 2026

Outline

Learning Objectives

By the end of this lecture, you will be able to:

1. Derive the MLE for logistic regression via gradient of the log-likelihood
2. Analyze model fit using deviance, LRT, AIC/BIC, and Hosmer-Lemeshow
3. Evaluate classification performance using ROC, calibration, and cost-sensitive metrics
4. Apply logistic regression to credit scoring with regulatory interpretation (Basel PD)

Finance Application: Credit scoring and probability of default (PD)

Bloom's Levels 4–5: Analyze, Evaluate, Create

Why Logistic Regression?

The Business Problem

- Banks must decide: approve or reject loan applications
- Need probability of default, not just yes/no prediction
- Regulatory requirement: interpretable, auditable models

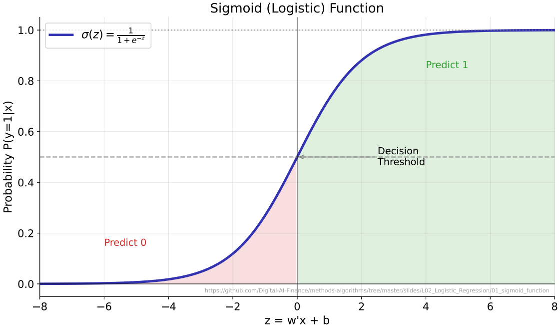
Why Not Linear Regression?

- Linear regression can predict values outside [0,1]
- Binary outcomes need probability-based approach
- Logistic regression outputs calibrated probabilities

Logistic regression: the industry standard for credit scoring since 1980s

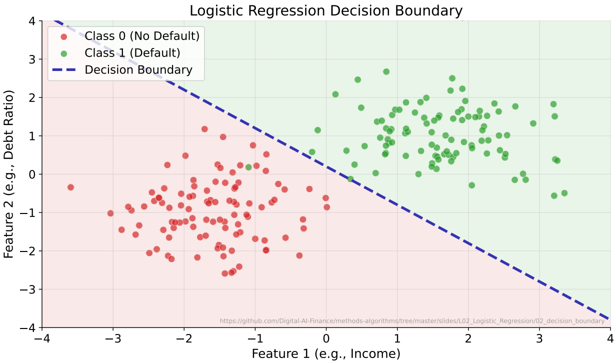
From Linear to Probability
Logistic function (probability of class 1):

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x})}} \tag{1}$$



Maps any real number to $(0, 1)$; smooth, differentiable, invertible

Decision Boundary



Log-odds (logit):

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

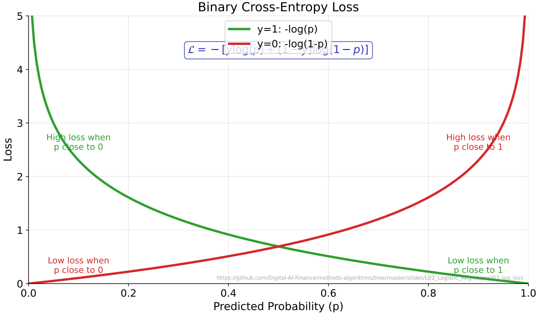
Decision boundary: $P(y = 1|\mathbf{x}) = 0.5 \iff \mathbf{w}'\mathbf{x} + b = 0$

The boundary is a hyperplane; points on one side predict class 1, the other class 0

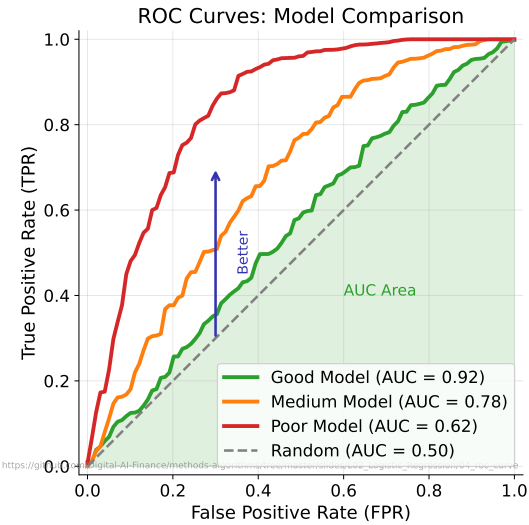
Maximum Likelihood Estimation:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right] \tag{3}$$

where $p_i = p(y_i = 1|\mathbf{x}_i)$.

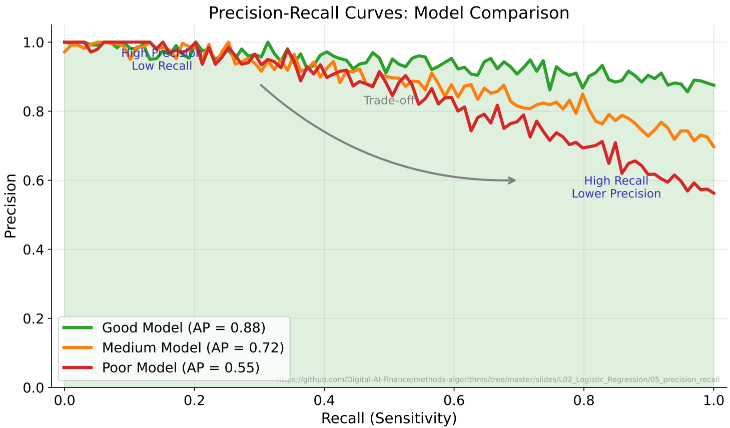


Maximize log-likelihood = minimize cross-entropy loss (convex, guaranteed global optimum)



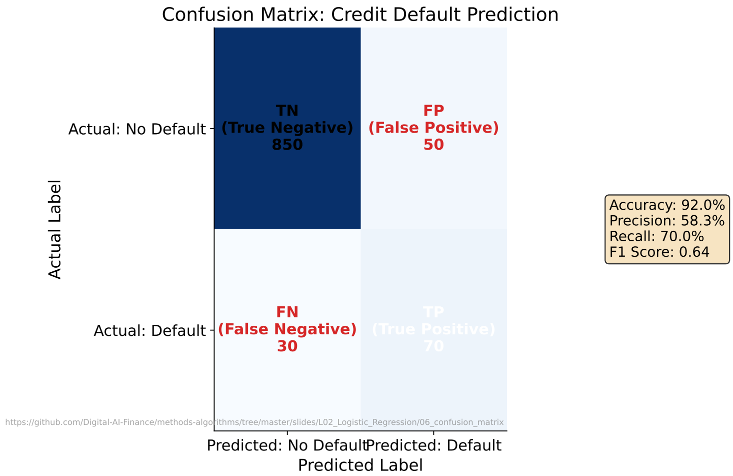
AUC = probability random positive ranks higher than random negative

Precision-Recall Trade-off



Use PR curve when classes are imbalanced (common in fraud detection)

Confusion Matrix: Reading the Results



FP = approve bad loans (costly), FN = reject good customers (lost revenue)

Hands-on Exercise

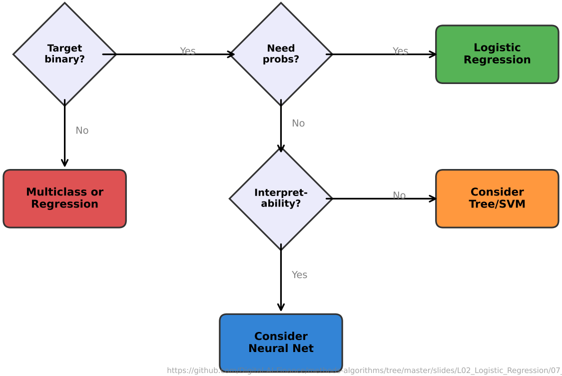
Open the Colab Notebook

- Exercise 1: Implement logistic regression from scratch
- Exercise 2: Train model on credit scoring data
- Exercise 3: Evaluate with ROC curve and confusion matrix

Link: See course materials on GitHub

When to Use Logistic Regression

Logistic Regression Decision Guide



Key strengths: interpretable coefficients, probability outputs, fast training

References

- James et al. (2021). *Introduction to Statistical Learning*. `https://www.statlearning.com/`
- Hastie et al. (2009). *Elements of Statistical Learning*. `https://hastie.su.domains/ElemStatLearn/`