Methods and Algorithms – MSc Data Science

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

1. Explain how logistic regression models binary outcomes
2. Derive the maximum likelihood estimation for logistic regression
3. Interpret classification metrics (precision, recall, AUC)
4. Apply logistic regression for credit scoring decisions

Finance Application: Credit default prediction

These objectives span Bloom's levels: Understand, Apply, Analyze

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

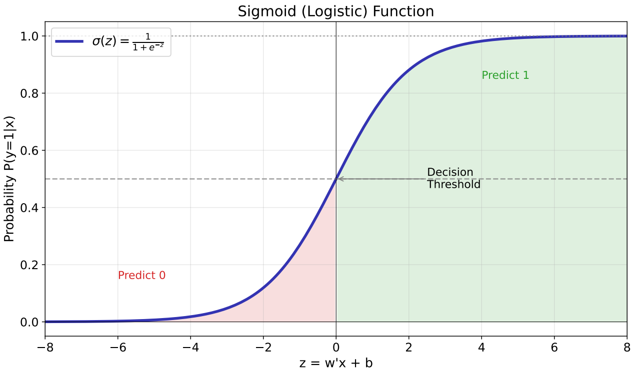
Why Not Linear Regression?

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

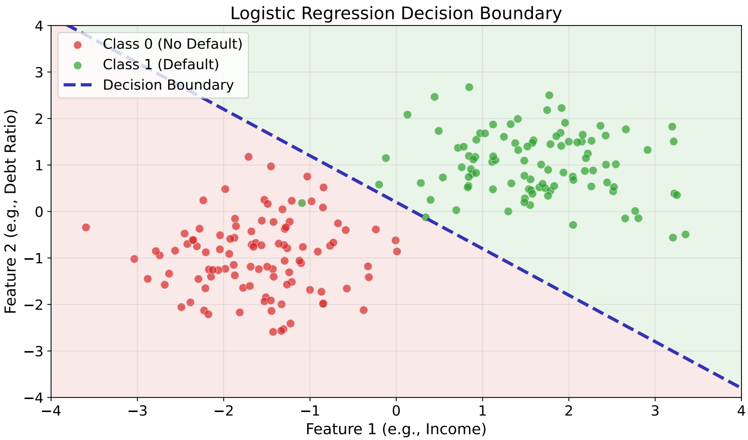
Logistic regression: the industry standard for credit scoring since 1980s

From Linear to Probability

- Maps any real number to (0, 1) range
- Smooth, differentiable, interpretable



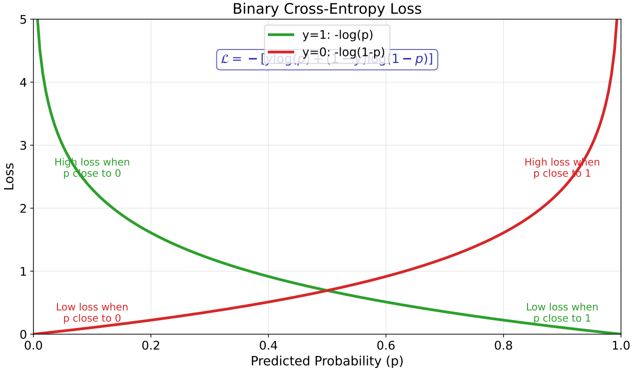
$\sigma(z) = 1/(1 + e^{-z})$ *transforms linear combination to probability*



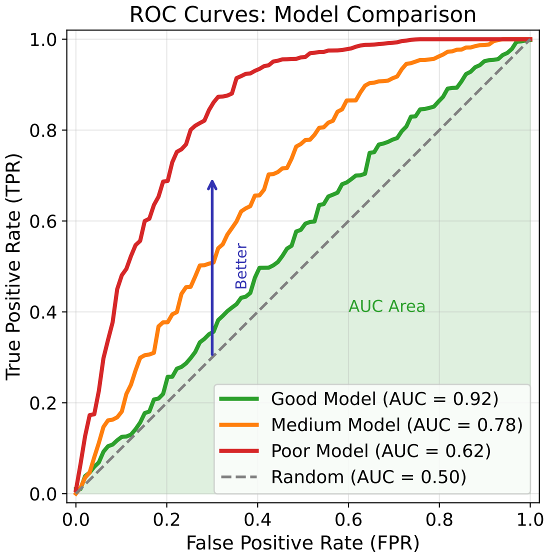
The decision boundary is where $P(y = 1|x) = 0.5$*, i.e.,* $w'x + b = 0$

Why Not MSE?

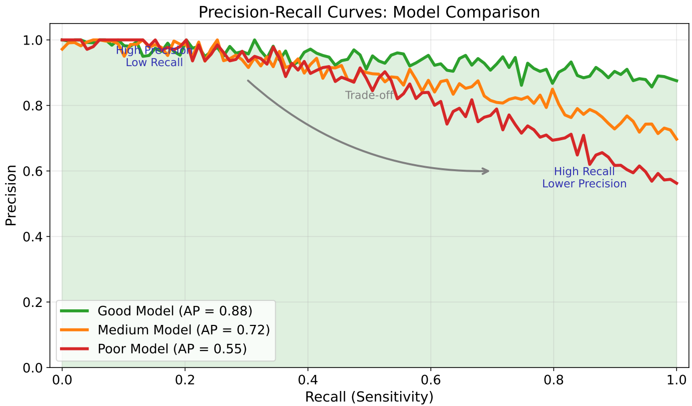
- MSE with sigmoid creates non-convex loss landscape
- Cross-entropy is convex, guarantees global optimum



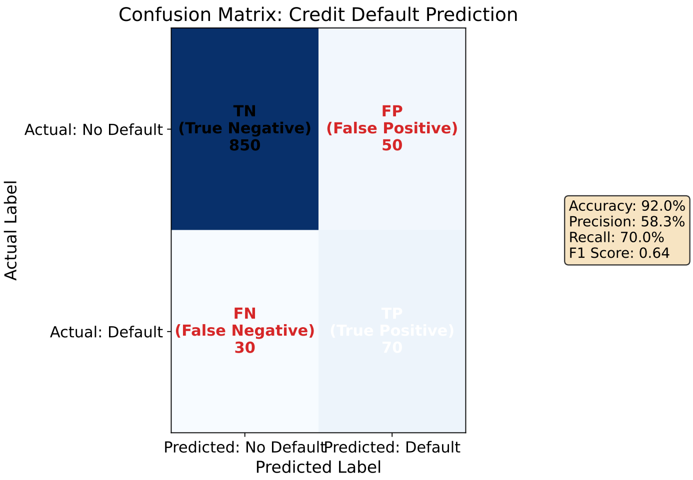
Heavily penalizes confident wrong predictions



AUC = probability random positive ranks higher than random negative

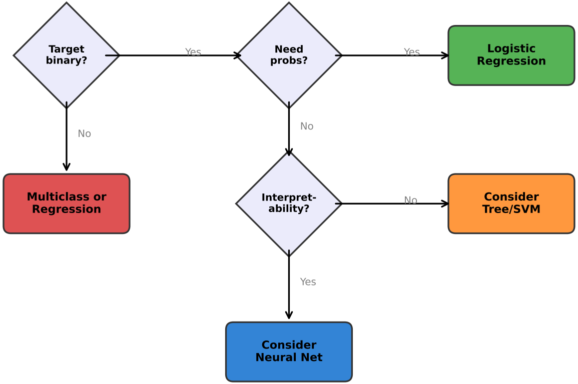


Use PR curve when classes are imbalanced (common in fraud detection)



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

Logistic Regression Decision Guide



Key strengths: interpretable coefficients, probability outputs, fast training