L02: Logistic Regression

Classification with Probability Estimates

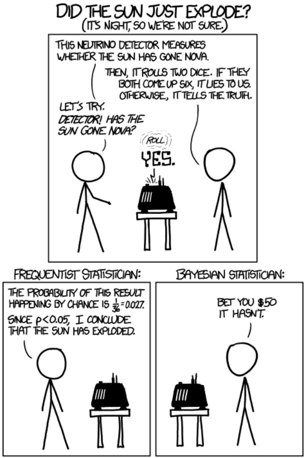
Methods and Algorithms

MSc Data Science

Spring 2026

Outline

The Classification Challenge



XKCD #1132 by Randall Munroe (CC BY-NC 2.5)

Why Logistic Regression?

The Business Problem

- Banks process millions of loan applications every year – each one is a **yes/no decision**
- A wrong "yes" costs the bank the entire loan amount; a wrong "no" loses a profitable customer
- Regulators demand models that are **interpretable**, **auditable**, and produce **calibrated probabilities**

The Standard Tool

- Logistic regression has been the **industry standard for credit scoring** since the 1980s
- It is fast to train, easy to explain, and directly outputs the probability of default

Every major bank uses logistic regression in its credit risk pipeline

From Prediction to Classification

What Changes When the Output is Yes/No?

- In regression, we predicted a continuous number (house price, stock return)
- In classification, the target is a **category**: default vs. no default, fraud vs. legitimate
- The model must output a **probability** between 0 and 1

The Problem with Linear Regression for Classification

- A straight line can predict values below 0 or above 1 – nonsensical as probabilities
- It treats the gap between 0.01 and 0.02 the same as between 0.49 and 0.50
- We need a function that **bends** the line to stay within valid bounds

---

Linear regression is unbounded – classification requires outputs in [0,1]

What Banks Need: Probabilities

Why a Simple Yes/No Is Not Enough

- Regulators (Basel framework) require banks to estimate the **Probability of Default** for every borrower
- These probabilities feed into capital calculations – how much reserve the bank must hold
- A model that only says “default” or “no default” cannot do this

From Probability to Scorecard

- Banks convert model probabilities into credit scores (e.g., 300–850 range)
- Higher score means lower probability of default means better lending terms
- Every coefficient must be **explainable** to auditors and regulators

Basel II/III: banks must produce PD estimates for all credit exposures

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, Apply

Why Not Linear Regression?

The Fundamental Issue

- Linear regression predicts $\hat{y} = \mathbf{x}^\top \boldsymbol{\beta}$, which is unbounded
- For binary classification, we need $P(y = 1|\mathbf{x}) \in (0, 1)$

The Logistic Solution

- Wrap the linear predictor in a **sigmoid function**:

$$P(y = 1|\mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} \tag{1}$$

- Output is always a valid probability – bounded, smooth, differentiable

Example: If $z = 0$, then $\sigma(0) = 0.5$ (50-50 chance). If $z = 2$, then $\sigma(2) = 0.88$ (88% likely).

The sigmoid "squashes" any real number into $(0, 1)$

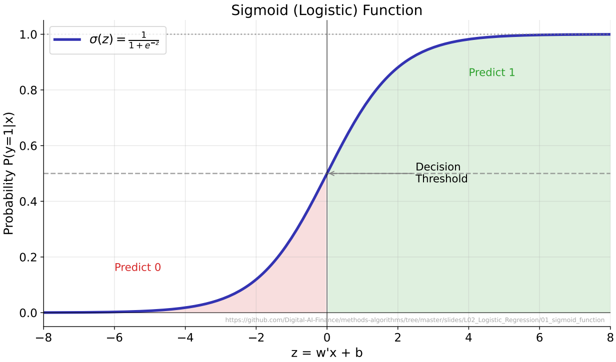
The Sigmoid Function

Key Properties

- $\sigma(0) = 0.5$ (the decision point)
- Symmetric: $\sigma(-z) = 1 - \sigma(z)$
- Derivative: $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Interpretation

- Large positive z: probability near 1
- Large negative z: probability near 0
- Steepness controlled by coefficient magnitude



The sigmoid maps $(-\infty, +\infty) \to (0, 1)$ – the foundation of logistic regression

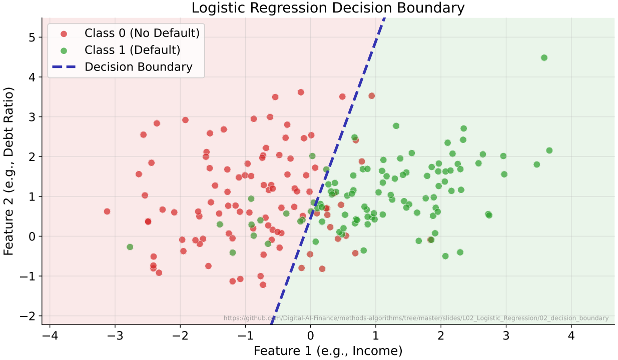
Decision Boundary

How Classification Works

- Predict class 1 if $P(y = 1|\mathbf{x}) \geq 0.5$
- Equivalently: predict 1 if $z \geq 0$
- The boundary is a **hyperplane** in feature space

Threshold Choice

- Default threshold $= 0.5$ is not always optimal
- Adjust based on costs of false positives vs. false negatives



Decision boundary: $\beta_0 + \boldsymbol{\beta}^\top \mathbf{x} = 0$ **– a linear separator in feature space**

Maximum Likelihood Estimation

The Likelihood Function

- Each observation contributes: $P(y_i|\mathbf{x}_i) = p_i^{y_i}(1-p_i)^{1-y_i}$
- Full likelihood: $L(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$

Log-Likelihood (what we maximize)

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right] \tag{2}$$

- No closed-form solution – must use **iterative optimization** (gradient ascent, Newton-Raphson)
- The log-likelihood is **concave** – guaranteed to find the global maximum

MLE: find the parameters that make the observed data most probable

Binary Cross-Entropy Loss

From Likelihood to Loss

- Minimizing the **negative** log-likelihood is equivalent to maximizing the log-likelihood
- The loss function for a single observation:

$$\mathcal{L}(y_i, p_i) = -\left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] \tag{3}$$

Intuition

- If $y_i = 1$ and $p_i \approx 0$: loss is very large (confident and wrong)
- If $y_i = 1$ and $p_i \approx 1$: loss is near zero (confident and correct)
- The loss **penalizes confident mistakes** more heavily than uncertain ones

Cross-entropy loss is convex in β – optimization is well-behaved

Gradient Derivation

Taking the Derivative

- The gradient of the log-likelihood with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} (y_i - p_i)\,\mathbf{x}_i = \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) \tag{4}$$

Key Insight

- The gradient has the same form as in linear regression: residual times feature
- Each update pushes predictions closer to the true labels
- Set to zero and solve iteratively (no closed-form solution)

Hessian (for Newton-Raphson)

$$\mathbf{H} = -\mathbf{X}^\top \mathbf{W} \mathbf{X}, \quad W_{ii} = p_i(1 - p_i) \tag{5}$$

Newton-Raphson converges in 5–10 iterations for typical credit scoring data

Odds and Log-Odds Interpretation

The Logit Link

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{6}$$

Coefficient Interpretation

- β_j: a one-unit increase in x_j changes the **log-odds** by β_j
- e^{β_j}: the **odds ratio** – multiplicative effect on the odds
- Example: if $\beta_{\text{income}} = -0.3$, then $e^{-0.3} = 0.74$

Credit Scoring Example

- "Each additional 10K income **multiplies** the odds of repayment by 1.35"
- This is exactly what regulators and auditors want: clear, directional, quantified effects

Odds ratio interpretation is the reason logistic regression dominates credit scoring

Model Fit and Comparison

Deviance and the Likelihood Ratio Test

- **Deviance**: $D = -2\,\ell(\hat{\beta})$ – analogous to residual sum of squares
- **LRT**: compare nested models via $\Delta D = D_{\text{reduced}} - D_{\text{full}} \sim \chi^2_{df}$

Information Criteria

- **AIC** $= -2\ell + 2k$ – penalizes model complexity (prefer smaller)
- **BIC** $= -2\ell + k \log N$ – stronger penalty, favors simpler models

Calibration: Hosmer-Lemeshow Test

- Groups observations into deciles of predicted probability
- Compares predicted vs. observed event rates – are the probabilities trustworthy?

AIC for prediction, BIC for model selection, Hosmer-Lemeshow for calibration

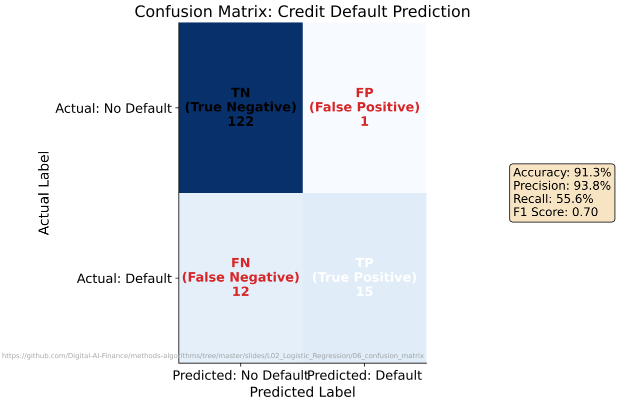
Confusion Matrix: Reading the Results

The Four Outcomes

- **TP**: correctly predicted default
- **TN**: correctly predicted repayment
- **FP**: predicted default, actually repaid (lost business)
- **FN**: predicted repayment, actually defaulted (lost money)

Banking Asymmetry

- FN is far more costly than FP
- A single default can wipe out profit from many good loans



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

Classification Metrics

Core Metrics from the Confusion Matrix

- **Accuracy** $= \frac{TP+TN}{TP+TN+FP+FN}$ – misleading with imbalanced classes
- **Precision** $= \frac{TP}{TP+FP}$ – of those flagged as default, how many truly defaulted?
- **Recall** $= \frac{TP}{TP+FN}$ – of all actual defaults, how many did we catch?

The F1 Score

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

- Harmonic mean – penalizes models that sacrifice one metric for the other
- Use when you care about **both** catching defaults and avoiding false alarms

Always report multiple metrics – no single number tells the whole story

ROC Curve and AUC

Receiver Operating Characteristic

- Plots True Positive Rate vs. False Positive Rate at every threshold
- **AUC** (Area Under the Curve): probability that a random positive ranks higher than a random negative
- AUC $= 0.5$: random guessing; AUC $= 1.0$: perfect separation

The Gini Coefficient

$$\text{Gini} = 2 \cdot \text{AUC} - 1 \tag{8}$$

- Ranges from 0 (no discrimination) to 1 (perfect)
- **Industry standard** for comparing credit scoring models
- Typical production models: Gini 0.4–0.7 depending on portfolio

---

ROC/AUC is threshold-independent – it evaluates the model's ranking ability

Credit Scoring in Practice

From Model to Scorecard

- Logistic regression coefficients are converted to **scorecard points**
- Each feature contributes points: higher total score = lower default risk
- Basel framework requires PD estimates for regulatory capital calculation

Regulatory Requirements (Basel II/III)

- **PD** (Probability of Default): direct output of logistic regression
- Models must be validated annually with out-of-sample testing
- **Discrimination** (Gini/AUC) and **calibration** (predicted vs. observed PD) both matter

Why Logistic Regression Dominates

- Transparent coefficients satisfy explainability requirements
- Well-calibrated probabilities without post-hoc adjustment

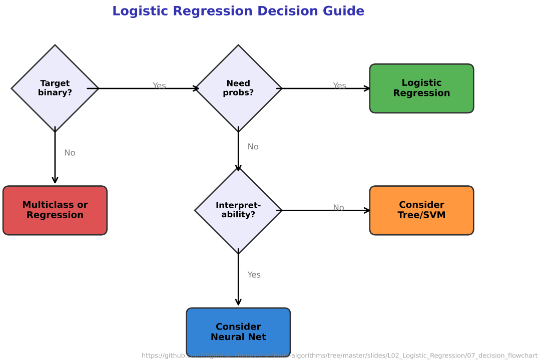
Basel II IRB approach: banks must estimate PD, LGD, EAD for every exposure

Best When

- Binary outcome with linear decision boundary
- Interpretability is required
- Calibrated probabilities are needed

Consider Alternatives When

- Highly non-linear boundaries
- Many feature interactions
- Prediction accuracy matters more than interpretability



Key strengths: interpretable coefficients, probability outputs, fast training

Hands-on Exercise

Open the Colab Notebook

- **Exercise 1:** Implement logistic regression from scratch (sigmoid, log-likelihood, gradient)
- **Exercise 2:** Train model on credit scoring data and interpret coefficients as odds ratios
- **Exercise 3:** Evaluate with confusion matrix, ROC curve, and Gini coefficient

What to Look For

- How do coefficients change when you add/remove features?
- What threshold gives the best trade-off for a bank's cost structure?
- Is the model well-calibrated (Hosmer-Lemeshow)?

Link: See course materials on GitHub

Estimated time: 45–60 minutes for all three exercises

Key Takeaways

Mathematical Foundation

- Sigmoid wraps a linear predictor to produce valid probabilities
- MLE via gradient ascent – concave log-likelihood guarantees convergence
- Coefficients have direct odds-ratio interpretation

Evaluation Toolkit

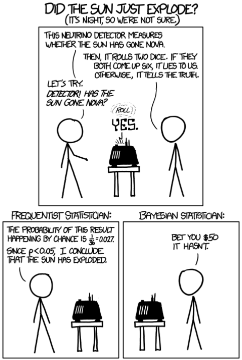
- Confusion matrix, precision, recall, F1 for threshold-specific performance
- ROC/AUC and Gini for threshold-independent model comparison
- Hosmer-Lemeshow for calibration quality

Practical Impact

- Industry standard for credit scoring – interpretable, auditable, calibrated
- Basel PD estimation relies on logistic regression in most banks

Logistic regression: simple enough to explain, powerful enough to deploy

Until Next Time...



"Is the sun going to explode?"

Now you have the tools to answer with a probability, not just yes/no.

Next Session: L03 – KNN & K-Means (from parametric to non-parametric)

XKCD #1132 by Randall Munroe (CC BY-NC 2.5) – classification is about probabilities, not certainties

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*, 2nd ed. Chapter 4. `https://www.statlearning.com/`
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Chapter 4. `https://hastie.su.domains/ElemStatLearn/`
- Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression*, 3rd ed. Wiley.
- Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards* (Basel II).

Primary textbook: ISLR Chapter 4 – Logistic Regression