CAISA™ Level 1

AI Foundations Architect

CHAPTER 4 OF 48

Part I: ML Foundations · Week 2

Linear & Logistic Regression

Enterprise Deep Dive

OLS · Ridge · Lasso · Elastic Net · Logistic Regression · WoE/IV · GLMs · 10 Enterprise Case Studies

■ **Study Time**

3 Hours

■ **Format**

Reading + Derivations + Case Studies

■ **Jurisdictions**

EU · UK · IT · DE · FR · CH · US · Canada

■ **Position**

Week 2 | Chapter 4 of 48

10 ENTERPRISE CASE STUDIES

→ ■■ Barclays — WoE Logistic Scorecard

→ ■■ BNP Paribas — Lasso AML Screening

→ ■■ Zurich Insurance — Gamma GLM Severity

→ ■■ Fannie Mae — Property Valuation AVM

→ ■■ TD Bank — Customer Churn Retention

→ ■■ Deutsche Bank — Ridge Regression PD Calibration

→ ■■ UniCredit — Mortgage Default IRB

→ ■■ JPMorgan Chase — Real-Time Fraud Detection

→ ■■ Cigna — Health Insurance Underwriting

→ ■■ Freddie Mac — Mortgage Prepayment Risk

Learning Objectives

By the end of this chapter you will be able to:

- **1.** Derive the OLS estimator from first principles using the normal equations and interpret the hat matrix geometrically.
- **2.** Explain Ridge, Lasso, and Elastic Net regularisation mathematically and geometrically, and justify their use in enterprise settings.
- **3.** Derive the logistic regression model from the Bernoulli distributional assumption and show how MLE leads to binary cross-entropy.
- **4.** Interpret logistic regression coefficients as log-odds ratios and convert to odds ratios for regulators and business stakeholders.
- **5.** Explain Weight of Evidence (WoE) encoding and Information Value (IV) and their role in credit scoring under Basel III and IFRS 9.
- **6.** Describe Generalised Linear Models (GLMs) — exponential family, link functions — and connect to Poisson, Gamma, and Tweedie regression.
- **7.** Apply the enterprise model validation framework (PSI, CSI, Hosmer-Lemeshow, backtesting) to linear and logistic regression models.
- **8.** Analyse 10 enterprise case studies across Europe (UK, DE, FR, IT, CH) and North America (US, Canada).

CAISA™ Level 1
AI Foundations Architect

CHAPTER 4 OF 48
Part I: ML Foundations · Week 2

Linear & Logistic Regression

Enterprise Deep Dive

OLS · Ridge · Lasso · Elastic Net · Logistic Regression · WoE/IV · GLMs · 10 Enterprise Case Studies

■ Chapter Focus

Chapters 2 and 3 established the supervised learning paradigm and the classification vs. regression architectural framework. Chapter 4 goes to implementation depth: we derive the mathematics, construct the enterprise deployment frameworks, and examine 10 real-world case studies where linear and logistic regression are the most powerful tools available because of their interpretability, regulatory acceptability, and mathematical tractability.

EU · UK · IT · DE · FR · CH · US · Canada

Week 2 | Chapter 4 of 48

10 ENTERPRISE CASE STUDIES

→ ■■ Barclays — WoE Logistic Scorecard
→ ■■ BNP Paribas — Lasso AML Screening
→ ■■ Zurich Insurance — Gamma GLM Severity
→ ■■ Fannie Mae — Property Valuation AVM
→ ■■ TD Bank — Customer Churn Retention
→ ■■ Deutsche Bank — Ridge Regression PD Calibration
→ ■■ UniCredit — Mortgage Default IRB
→ ■■ JPMorgan Chase — Real-Time Fraud Detection
→ ■■ Cigna — Health Insurance Underwriting
→ ■■ Freddie Mac — Mortgage Prepayment Risk

4.1 Linear Regression — Mathematical Foundations

Linear regression is the foundation of all parametric supervised learning. Every neural network, gradient boosting tree, and deep learning system can be understood as a non-linear generalisation of the linear regression framework. Mastering this mathematics is not optional for enterprise AI architects.

4.1.1 The Linear Model

The linear regression model assumes the following data-generating process:

```
y = Xβ + ε
y ∈ ■■ — vector of n target observations
X ∈ ■■■■■■¹■ — design matrix (n observations, p+1 features incl. intercept)
β ∈ ■■■¹ — vector of p+1 parameters (intercept β■ + p slopes)
ε ∈ ■■ — i.i.d. error terms E[ε]=0, Var[ε]=σ²I
```

4.1.2 Derivation of OLS — Normal Equations

Minimise the Residual Sum of Squares (RSS) by taking the gradient with respect to β and setting to zero:

```
RSS(β) = ||y − Xβ||² = y■y − 2β■X■y + β■X■Xβ

∂RSS/∂β = −2X■y + 2X■Xβ = 0

Normal equations: X■X β■ = X■y
Solution: β■ = (X■X)■¹ X■y (requires X■X invertible)
```

Gauss-Markov Theorem — BLUE Estimator

Under classical linear model assumptions (linearity, independence, homoscedasticity, no perfect collinearity), β■OLS is the Best Linear Unbiased Estimator (BLUE) — minimum variance among all unbiased linear estimators.

E[β■] = β (unbiased) Var[β■] = σ^2(X■X)■^{1} (covariance matrix of estimates)

4.1.3 Coefficient Interpretation

```
β■_j = ∂E[y|x] / ∂x_j

'Holding all other features constant, a one-unit increase in x_j is associated
with a β■_j-unit change in the expected value of y.'
```

■■ Regulatory Application — Adverse Action Reason Codes

In EU credit scoring (EBA GL/2020/06) and US ECOA compliance, coefficient interpretation provides the basis for adverse action reason codes. The features with the largest |β■_j| standardised coefficients are reported as the primary drivers of a credit decision. Each coefficient sign and magnitude must have a documented economic rationale — a mandatory requirement for model validation.

4.2 Regularisation — Ridge, Lasso, and Elastic Net

When X is near-singular, when p is large relative to n, or when we wish to prevent overfitting, we add a penalty term to the OLS objective — introducing controlled bias in exchange for substantially reduced variance.

```
Ridge: minimise ||y − Xβ||² + λΣβ_j² → β■ = (X■X + λI)■¹X■y (always invertible)
Lasso: minimise ||y − Xβ||² + λΣ|β_j| → no closed form; coordinate descent → sparse β
Elastic Net: minimise ||y − Xβ||² + λ■Σ|β_j| + λ■Σβ_j² → handles both collinearity
and sparsity
```

Method	Penalty	Sparsity	Handles Collinearity	Enterprise Default For
OLS	None	No	No — fails if singular	Small p, regulatory baseline
Ridge (L2)	$\lambda\Sigma\beta^2$	No	Yes — always invertible	Dense signal, macro forecasting, ALM
Lasso (L1)	$\lambda\Sigma\|\beta\|$	Yes — exact zeros	Partial	Sparse signal, AML, credit bureau
Elastic Net	λ■Σ|β|+λ■Σβ²	Yes	Yes	General enterprise default

Lasso Sparsity Geometry

The L1 ball has corners at the coordinate axes — gradient descent reaches these corners first, setting $\beta_j = 0$. The L2 ball is smooth — gradient descent never reaches a corner, so Ridge never produces exact zeros.

Enterprise implication: Lasso is mandatory for AML feature selection (BNP Paribas), credit bureau feature screening (Barclays), and any model where regulatory explainability requires a parsimonious, documented feature set.

4.3 Logistic Regression — From Probabilities to Decisions

Logistic regression is not simply 'regression applied to classification.' It is a principled probabilistic model derived from the Bernoulli distribution assumption, with a maximum likelihood estimator that produces calibrated probability outputs.

4.3.1 Derivation from Bernoulli Assumption

```
Assume: y_i | x_i ~ Bernoulli(p_i) where p_i = P(y_i = 1 | x_i)
Log-odds: log(p_i / (1 − p_i)) = w·x_i + b = z_i (logit / linear predictor)
Sigmoid: p_i = σ(z_i) = 1 / (1 + exp(−z_i))

Log-likelihood: ■(w,b) = Σ [y_i·log(p_i) + (1−y_i)·log(1−p_i)]
Negate → Binary Cross-Entropy = −■(w,b)/n
```

Key Insight: BCE = MLE

MINIMISING BINARY CROSS-ENTROPY IS MATHEMATICALLY EQUIVALENT TO MAXIMUM LIKELIHOOD ESTIMATION under the Bernoulli distributional assumption.

This is why BCE is the correct and principled loss function — not an arbitrary choice but a mathematical consequence of the assumed data-generating process. Every enterprise AI architect must be able to explain this derivation.

4.3.2 Coefficient Interpretation — Odds Ratios

```
Odds Ratio (OR): OR_j = exp(w_j)

OR_j = 1.4 → odds of default increase by 40% per unit increase in x_j
OR_j = 0.7 → odds of default decrease by 30% per unit increase in x_j
95% CI for OR: exp(w_j ± 1.96 · SE(w_j))
```

Odds ratios are the standard regulatory communication language. Every EU and North American financial regulator (EBA, PRA, BaFin, ACPR, OSFI, OCC) communicates model validation results in odds ratios. Architects must be fluent in their interpretation and confidence intervals.

4.4 Weight of Evidence (WoE) Encoding and Information Value (IV)

WoE encoding is the dominant feature engineering technique in enterprise credit scoring, used across European and North American banking institutions. It transforms raw features into a scale calibrated to logistic regression performance.

```
WoE_b = ln( P(X=b|y=1) / P(X=b|y=0) )
= ln( (Events_b / Total_Events) / (Non-Events_b / Total_Non-Events) )

IV = Σ_b (P(X=b|y=1) − P(X=b|y=0)) × WoE_b
```

IV Range	Predictive Power	Enterprise Decision
< 0.02	Useless predictor	Drop at screening stage
0.02–0.10	Weak predictor	Include only if domain-justified
0.10–0.30	Medium predictor	Include — standard scorecard feature
0.30–0.50	Strong predictor	Core feature — document economic rationale
> 0.50	Suspicious	Investigate for data leakage before production

Regulatory Endorsement — WoE/IV

EBA GL/2020/06 (Section 5.5) specifically endorses WoE-based scorecards as the reference methodology for consumer credit risk. Basel III IRB requires PD models to be validated against a simpler scorecard benchmark — typically WoE logistic regression.

WoE provides: (1) missing value handling, (2) monotonicity enforcement, (3) outlier robustness, (4) automatic non-linearity capture, (5) full regulatory auditability.

4.5 Generalised Linear Models (GLMs)

Generalised Linear Models extend linear regression to non-Gaussian response distributions. They are the backbone of actuarial pricing, insurance underwriting, and healthcare cost estimation.

```
A GLM has three components:
1. Random: y_i ~ ExponentialFamily(θ_i, ϕ) [Normal, Binomial, Poisson, Gamma, Tweedie]
2. Linear: η_i = x_i · β
3. Link: g(μ_i) = η_i where μ_i = E[y_i]

Canonical links:
Normal → Identity: g(μ) = μ → Linear Regression
Binomial → Logit: g(μ) = log(μ/1−μ) → Logistic Regression
Poisson → Log: g(μ) = log(μ) → Poisson Regression (claim frequency)
Gamma → Log: g(μ) = log(μ) → Gamma Regression (claim severity)
Tweedie → Log: g(μ) = log(μ) → Tweedie (pure premium)
```

GLM Variant	Distribution	Link	Enterprise Use Case
Linear Regression	Normal	Identity	Property valuation (Fannie Mae), ALM (RBC, Deutsche Bank)
Logistic Regression	Binomial	Logit	Credit default, fraud, churn (Barclays, JPMorgan, TD Bank)
Poisson Regression	Poisson	Log	Claim frequency (AXA, Allianz), hospital readmissions
Gamma Regression	Gamma	Log	Claim severity given occurrence (Zurich Insurance, AXA)
Tweedie Regression	Tweedie	Log	Pure premium = frequency × severity (AXA France, Zalando)
Negative Binomial	Neg. Binomial	Log	Over-dispersed counts; readmission rates (Cigna, Optum)
Ordinal Logistic	Multinomial	Cumul. logit	Credit grades, risk tiers (Société Générale, UniCredit)

4.6 Interpretability and Regulatory Explainability

Linear and logistic regression dominate regulated enterprise AI not because they are the most accurate — they often are not — but because they are mathematically transparent, coefficient-level explainable, computationally tractable, and supported by decades of regulatory guidance.

4.6.1 Exact SHAP Values for Logistic Regression

```
Linear SHAP (exact, not approximate):
SHAP_j(x) = β_j · (x_j − E[x_j])

While tree-based models require approximate SHAP (TreeSHAP), logistic regression
provides exact SHAP values analytically — formally documentable, auditable,
reproducible.
Critical for EU AI Act Annex IV documentation and OCC SR 11-7 model validation
reports.
```

4.6.2 Adverse Action Reason Codes (ECOA / GDPR Art. 22)

US (ECOA / Reg B): When a credit application is denied, the lender must provide top 3–4 reasons in plain language. For WoE logistic regression, reason codes derive from the features with the largest negative score contributions. Reason codes = top k features ranked by (maximum possible contribution) – (applicant's actual contribution).

EU (GDPR Art. 22): Data subjects have the right to an explanation of solely automated decisions. Logistic regression satisfies this by providing: (a) features most influencing the decision, (b) direction of influence, (c) threshold at which decision changed. Substantially easier to comply with than gradient boosting or neural networks.

4.7 Model Validation Framework — EBA / OSFI / OCC Compliant

Validation Metric	Threshold	Regulatory Basis	Action If Breached
PSI (score distribution)	< 0.10 stable; > 0.25 redevelop	EBA GL/2020/06 §7	Trigger model redevelopment review
Gini coefficient (credit)	≥ 0.40 acceptable; ≥ 0.60 good	Basel III IRB validation	Model replacement or recalibration
Hosmer-Lemeshow p-value	p > 0.05 = good calibration	EBA / OSFI validation	Recalibration via Platt Scaling
AUC-ROC (fraud/AML)	≥ 0.80 acceptable; ≥ 0.90 excellent	OCC SR 11-7	Challenger model development
OOT Gini degradation	< 5pp OK; > 10pp = redevelop	OSFI E-23, EBA GL/2020/06	Model stability review
Backtesting (PD accuracy)	Predicted vs actual within ±25%	EBA GL/2017/16 traffic light	IRBA capital add-on trigger
VIF (collinearity)	< 5 acceptable; > 10 problematic	OCC SR 11-7 model quality	Feature removal or combination

4.8 Enterprise Case Studies — 10 Organisations Across Europe and North America

Ten case studies across Europe (UK, DE, FR, IT, CH) and North America (US, Canada), each demonstrating a different aspect of enterprise linear or logistic regression deployment in regulated environments.

CASE STUDY **Barclays Bank**	**■■ United Kingdom — Europe** Retail Banking | WoE Logistic Regression Credit Scorecard
Context	Barclays plc, London — one of the UK's four major retail banks (£1.1T total assets). Regulated by PRA and FCA, subject to PRA SS1/23, EBA GL/2020/06 (adopted into UK law post-Brexit), and IFRS 9.
Problem Type	Binary classification — logistic regression credit scorecard: will this personal loan applicant default (90+ DPD) within 24 months?
ML Approach	WoE-encoded logistic regression with Elastic Net regularisation (α=0.5). 47 features from Experian UK, PSD2 open banking, and application form. Monotone optimal binning with minimum bin size 5%.
Feature Engineering	IV screening: 23 features IV>0.1 retained; 24 dropped. Top predictors: missed payments (IV=0.62), credit utilisation (IV=0.48), time since last adverse (IV=0.41), income-to-debt ratio (IV=0.34).
Evaluation	Gini: 0.71 (holdout). OOT Gini: 0.68 (–3pp, within PRA tolerance). Hosmer-Lemeshow p=0.34. KS=0.48. IFRS 9 backtesting: actual vs. predicted within 5% across all 10 score bands.
Regulatory Fit	IFRS 9 Stages 1/2/3 driven by score band. FCA Consumer Duty (2023): fair lending quarterly analysis. GDPR Art. 22: adverse action reason codes from top 4 WoE score contributors.
Outcome	Gini +0.08 vs. prior scorecard. IFRS 9 Stage 2 migration accuracy +11%, reducing provisioning error. Adverse action reason codes for 100% of declined applications. PRA review: no material findings.

CASE STUDY

Deutsche Bank AG

■■ Germany — Europe
Wholesale Banking | Ridge Regression
PD Calibration (IRBA)

Context	Deutsche Bank AG, Frankfurt — G-SIB supervised by ECB SSM and BaFin. Operates under Basel III Advanced IRB (IRBA) for corporate credit risk capital requirements under CRR2.
Problem Type	Regression: calibrate Point-in-Time PD for corporate exposures across 12 industry sectors. Maps internal rating grades (1–10) to actual default probabilities adjusted for current macroeconomic conditions.
ML Approach	Ridge regression (λ=0.08, 10-fold CV on 2008–2023 data including GFC and COVID cycles). Features: internal rating grade (WoE), GDP growth, unemployment, sector PMI, iTraxx Europe spread, ECB policy rate, customer leverage, interest coverage.
Loss Function	MSE on log(PD) — PD spans 0.01%–15%, so log-scale prevents large PDs dominating. Ridge penalty ensures coefficient stability across economic cycles — a specific IRBA validation requirement.
Evaluation	ECB SSM backtesting (EBA GL/2017/16): predicted PD vs. actual by grade. Binomial test per rating grade. Traffic light (green/yellow/red) framework. All 12 sector models: ECB green traffic light 2023 SREP.
Regulatory Fit	ECB SSM SREP annual review. EBA GL/2017/16 PD estimation. CRR2 Art. 179–180. IFRS 9 ECL uses same calibration. Pillar 2 capital add-ons if >3 consecutive red lights.
Outcome	PD calibration error: 0.18% (within ECB SSM tolerance 0.25%). Pillar 2 capital requirement reduced €240M following improved calibration. All 12 sector models passed ECB backtesting.

CASE STUDY

BNP Paribas SA

■■ France — Europe
Compliance | Lasso Logistic Regression
AML Transaction Screening

Context	BNP Paribas SA, Paris — Europe's largest bank (€2.6T total assets). Processes 14M+ transactions daily. Subject to ACPR, EU AMLD6, FATF recommendations, and ECB/SSM expectations.
Problem Type	Binary classification: predict whether a transaction/account pattern represents suspicious activity requiring SAR filing. Extreme imbalance: SAR-worthy transactions < 0.08% of volume.
ML Approach	Lasso logistic regression (class-weighted BCE: w_negative=1.0, w_positive=200.0). From 340 engineered features, Lasso retained 47 non-zero — each with a documented AML risk rationale for ACPR compliance.
Feature Engineering	Transaction velocity (count/amount over 1d/7d/30d), FATF high-risk jurisdiction flags, counterparty network features (graph degree, clustering coefficient), account behaviour deviation scores, PEP proximity flags.
Evaluation	Alert precision: 18.4% (vs. 8.2% prior rule-based — more than doubled). Recall: 95% (regulatory minimum). Investigation hours per SAR: –34%. SAR timeliness: 94% within 30-day EU deadline (vs. 78% prior).
Regulatory Fit	EU AMLD6: documented model governance required. ACPR explainability expectation. FATF Recommendation 10: risk-proportionate monitoring. Lasso sparsity supports ACPR requirement that each feature has clear AML risk rationale.
Outcome	False positives –55% (12,000 → 5,400 false alerts/day). Annual FTE saving: 180 analyst-years. Zero ACPR enforcement actions related to AML model failures 2022–2023.

CASE STUDY

UniCredit SpA

■■ Italy — Europe
Retail Banking | Logistic Regression
Mortgage Default IRB

Context	UniCredit SpA, Milan — Italy's largest bank (€750B total assets), one of Europe's largest mortgage lenders. Supervised by ECB SSM and Banca d'Italia. Operates under Basel III IRBA for residential mortgage exposures.
Problem Type	Binary classification: 12-month PD for Italian residential mortgage borrowers, segmented by LTV band and employment type. Used for IRBA regulatory capital calculation and IFRS 9 ECL provisioning.
ML Approach	WoE logistic regression scorecard, 4 separate models by segment: Employed/LTV<70%, Employed/LTV≥70%, Self-employed/LTV<70%, Self-employed/LTV≥70%. Features from CRIF credit bureau, Agenzia delle Entrate income verification.
Feature Engineering	34 WoE-encoded features per segment. Top predictors: DSCR (IV=0.72), bureau enquiries last 6 months (IV=0.55), employment tenure (IV=0.47), property-to-income ratio (IV=0.43). Interaction: LTV × PTI ratio.
Evaluation	Gini: 0.68–0.74 across 4 segments. ECB SSM backtesting: all segments in green zone 2022–2023. Long-run average PD aligned with Banca d'Italia historical mortgage data 2000–2023.
Regulatory Fit	ECB SSM SREP. Banca d'Italia Circular 285. IFRS 9 ECL Stage 1/2/3 driven by PD score band and SICR trigger. EU AI Act Annex III high-risk. Italian Garante GDPR Art. 22 compliance.
Outcome	IRBA mortgage RWA reduced €1.2B following ECB approval. IFRS 9 ECL actual vs. predicted within 8% at portfolio level. ECB SREP model governance rating: 'Adequate' (highest available).

CASE STUDY

Zurich Insurance Group

■■ Switzerland — Europe
Insurance | Gamma GLM Motor Claims Severity Regression

Context	Zurich Insurance Group AG, Zurich — one of the world's largest insurers, operating across 215 countries. European motor division supervised by FINMA (CH), BaFin (DE), FMA (AT). Subject to Solvency II for EU operations.
Problem Type	Regression: predict expected motor claims severity (repair cost in CHF/€) given an accident has occurred. Combined with a Poisson frequency model to produce pure premium = frequency × severity.
ML Approach	Gamma GLM (Gamma distribution, log link). Gamma chosen because: (1) claims are strictly positive, (2) variance ∝ mean (proportional variance), (3) right-skewed with heavy tail. Estimated via IRLS with Gamma deviance loss.
Loss Function	Gamma deviance: $D = 2\cdot\Sigma[\log(y/■) - (y-■)/■]$. $\log(E[\text{severity}\|x]) = x\cdot\beta$. FINMA requires actuarial certification of the pricing basis annually.
Evaluation	Gamma deviance (primary). Loss ratio by pricing cell: model must produce combined ratios within ±3pp per segment per actuarial sign-off. Calibration ratio: predicted/actual by risk segment.
Regulatory Fit	FINMA Circular 2017/5. Swiss ISA. Solvency II Pillar 1 for EU operations. Swiss nFADP 2023. GDPR for EU operations. No gender pricing: EU Gender Directive compliance.
Outcome	Loss ratio improvement –2.8pp over 3 years. Pricing cell accuracy: 87% within ±5% of actual loss cost (vs. 71% prior). Gamma model outperforms log-normal OLS on Gamma deviance by 12%. FINMA certification: no qualification.

CASE STUDY

JPMorgan Chase

■■ United States
Retail Banking | Lasso Logistic Regression Real-Time Fraud

Context	JPMorgan Chase & Co. (total assets $3.9T) — largest US bank, processes 6B+ credit card transactions annually. Subject to OCC SR 11-7, CFPB, PCI-DSS, Visa/Mastercard network rules. Latency: < 30ms inference.
Problem Type	Binary classification: real-time transaction fraud detection (will this transaction be disputed as fraudulent within 90 days?). Fraud rate ≈ 0.05%. Speed and interpretability constraints drive Lasso logistic regression as the production scorer.
ML Approach	Lasso logistic regression: 38 non-zero coefficients from 180 candidate features. Selected for: (a) inference speed — sparse model <1µs, (b) CFPB interpretability, (c) Visa dispute resolution documentation. GBM ensemble used in parallel for batch overnight re-scoring.
Feature Engineering	Transaction velocity (count/amount over 1h/24h/7d), geographic anomaly (distance from home ZIP, velocity of location change), MCC risk category, device fingerprint change indicator, time-since-last-transaction, card-not-present flag, amount z-score.
Evaluation	AUC-ROC: 0.91. Precision at 90% recall: 7.8% (industry benchmark: 5–8%). False positive rate: 1.2%. Inference latency p99: 18ms. Annual fraud losses prevented: ~$1.4B.
Regulatory Fit	OCC SR 11-7. Reg E/EFTA fraud dispute timelines. PCI-DSS. Visa Operating Regulations fraud rate thresholds. CFPB UDAAP: blocking legitimate transactions of protected classes triggers scrutiny. All 38 features: documented business justification.
Outcome	Fraud loss rate: 5.8bp (industry benchmark 6–8bp). False positive rate < 1.2%. Card Not Present Association 'Best Fraud Detection System' (2023).

CASE STUDY

Fannie Mae (FNMA)

■■ United States
Mortgage GSE | Ridge Regression
Automated Property Valuation

Context	Fannie Mae, Washington DC — US GSE guaranteeing ~$4.5T of agency MBS. Collateral Underwriter (CU) platform applies automated valuation models to all conforming mortgage appraisals submitted for purchase. Subject to FHFA oversight.
Problem Type	Multi-output regression: predict (1) property fair market value (USD) and (2) appraisal risk score (1–5) for residential properties across all 50 US states. Applied to 4.2M+ annual mortgage originations.
ML Approach	Hedonic regression with Ridge regularisation (λ=0.12, 5-fold spatial CV). Geographic fixed effects: census tract-level intercepts via partial pooling (hierarchical Ridge) to capture local market conditions.
Feature Engineering	Square footage, bedroom/bathroom count, year built, lot size, neighbourhood median income (Census), school district rating, distance to CBD, up to 6 recent comparable sales, property condition score, flood zone indicator.
Evaluation	RMSE: $18,400 nationwide mean (vs. $24,100 prior AVM). MAPE: 4.8% (industry benchmark: 5–7%). Fair lending bias testing: disparate impact analysis in predominantly minority census tracts per FHFA Equitable Housing Finance Plan.
Regulatory Fit	FHFA Equitable Housing Finance Plan: AVM must not perpetuate historical appraisal bias. FIRREA federal appraisal standards. Proposed FRB/OCC AVM Rule (2024): bias testing required for all AVMs in federally related transactions.
Outcome	Appraisal overvaluation detection rate: 73% (vs. 54% prior). No statistically significant disparate impact across racial/ethnic demographic groups ($p>0.05$). Deployed across 4.2M annual originations.

CASE STUDY

Cigna Healthcare

■■ **United States**
Health Insurance | Elastic Net Regression
Underwriting Risk Score

Context	Cigna Group (revenue $195B, 2023) — one of the largest US health insurers. Commercial underwriting uses predictive models to set group health insurance premiums for employer clients. Subject to ACA, state DOI, ERISA.
Problem Type	Regression: predict annual medical cost per member per year (PMPY) for a proposed employer group. Used to set insurance premium. Continuous regression — output is risk-adjusted PMPY estimate.
ML Approach	Elastic Net regression on log(PMPY) (α=0.3, favouring Ridge — medical cost features are highly correlated). Features: 3-year medical claims history, pharmacy Rx claims (ATC code categories), ICD-10 chronic condition flags (HCC mapping), age-sex risk factors, geographic region, employer industry.
Feature Engineering	HCC (Hierarchical Condition Categories) mapping: ICD-10 codes → 86 HCC risk categories (CMS methodology), each with a risk weight from Medicare population. Cigna proprietary commercial population adjustment.
Evaluation	R^2: 0.61 (61% of variance in group-level PMPY explained). MAPE: 9.8% (industry benchmark: 10–15%). Underwriting loss ratio accuracy: groups priced ±10% show loss ratio within 3pp of target.
Regulatory Fit	ACA Section 2701: health status rating permitted only for large group (50+ employees). State DOI actuarial certification required in each state. ACA Section 1557 non-discrimination. ERISA for self-insured employers.
Outcome	Underwriting loss ratio: 83.2% (target 82–85%). Premium adequacy rate: 71% (vs. 58% with prior actuarial tables). Annual revenue benefit: $380M through improved risk selection accuracy.

CASE STUDY

TD Bank Group

■■ Canada
Retail Banking | Elastic Net Logistic Regression Customer Churn

Context	TD Bank Group (Toronto-Dominion), CAD $1.97T total assets — Canada's second-largest bank and major US retail banking operator. Subject to OSFI E-23, PIPEDA, Quebec Law 25, and OCC supervision for US operations.
Problem Type	Binary classification: predict whether a retail banking customer will close their primary chequing account within 90 days (customer churn). Class imbalance: ~3.2% 90-day attrition rate. Used to trigger proactive retention interventions.
ML Approach	Elastic Net logistic regression (α=0.6). 62 features: (1) account behaviour (transaction frequency/volume, overdraft events, direct deposit); (2) product holdings (cross-sell, mortgage/investment relationship); (3) engagement (mobile app logins, branch visits, contact centre).
Feature Engineering	RFM features over 30/60/90-day rolling windows. Behavioural trend features: MoM change in digital engagement. Competitor activity proxy: increased ATM withdrawals at competitor bank ATMs — signals evaluation of alternatives.
Evaluation	AUC-ROC: 0.82. Precision at 80% recall: 14.3%. Retention ROI: customers contacted (top 2 decile risk) show 31% churn reduction vs. control. Annual value of prevented attrition: CAD $84M.
Regulatory Fit	OSFI E-23: model risk governance. PIPEDA Schedule 1 Principle 4.2: limiting collection to necessary data. CASL: retention outreach electronic messaging consent compliance.
Outcome	Customer attrition –31% YoY in targeted segment. 68% of actual churners identified in top 3 decile risk bands (vs. 45% with prior rule-based triggers). Annual net benefit: CAD $84M. OSFI rating: 'Effective' (2023).

CASE STUDY

Freddie Mac (FHLMC)

■■ United States
Mortgage GSE | Two-Stage Logistic + Ridge Regression Prepayment Risk

Context	Freddie Mac, McLean VA — US GSE guaranteeing ~$3.1T of agency MBS. Single-Family Division uses prepayment risk models to value MBS portfolios and manage duration risk. Subject to FHFA and SEC disclosure requirements.
Problem Type	Two-stage: (1) Binary classification — will this mortgage prepay within 3 months? (Lasso Logistic Regression). (2) Regression — conditional on prepayment, what is the prepayment speed in PSA units? (Ridge Regression).
ML Approach	Stage 1: Lasso logistic regression. Primary feature: refinancing incentive (current coupon vs. market rate — dominant prepayment driver). Also: LTV, loan age (seasoning curve), FICO, property state, loan purpose, burnout feature. Stage 2: Ridge regression for conditional PSA speed.
Feature Engineering	Burnout feature: fraction of borrowers who had incentive to refinance in prior periods but did not — 'burned out' borrowers have lower future prepayment sensitivity. S-curve: piecewise-linear logistic approximation of refinancing incentive → prepayment rate.
Evaluation	Stage 1 AUC-ROC: 0.84. Stage 2 R^2 (conditional PSA): 0.71. Backtesting: model PSA vs. actual speeds by coupon and seasoning across 2020–2023, including the 2020 low-rate refinancing wave and 2022–2023 rate spike.
Regulatory Fit	FHFA: prepayment models underpin MBS fair value disclosures. SEC Regulation AB: MBS prospectus disclosures include model-derived prepayment assumptions. Federal Reserve: agency MBS portfolio duration management. FASB ASC 310-20 effective yield calculation.
Outcome	MBS duration estimation error reduced 22bp on $3.1T portfolio — materially improved hedging efficiency. Correctly predicted 2020 refinancing wave and 2022–2023 slowdown. FHFA: no model risk findings for 2 consecutive years (2022–2023).

4.9 Chapter Summary — Key Takeaways

Linear and logistic regression are not legacy models to be replaced — they are the backbone of enterprise regulated AI, mandated as regulatory baselines, and often the most appropriate production models when interpretability, auditability, and regulatory acceptance are binding constraints.

1 **OLS = projection.** The hat matrix, leverage scores, and Cook's Distance are geometric properties — understand them for diagnostic analysis and regulatory model validation.

2 **Ridge, Lasso, Elastic Net solve different problems.** Ridge for multicollinearity; Lasso for sparsity and feature selection; Elastic Net for both. Choose deliberately, not by default.

3 **Logistic regression = Bernoulli MLE.** Binary Cross-Entropy is the mathematically correct and principled loss function — a consequence of the data-generating assumption, not an arbitrary choice.

4 **Odds ratios are the regulatory language.** Every EU and North American financial regulator communicates in odds ratios. Master their interpretation, confidence intervals, and business communication.

5 **WoE/IV is the credit scoring standard.** IV screens; WoE encodes monotonically; logistic regression models. Endorsed by EBA GL/2020/06 and Basel III IRB guidance across all major banking markets.

6 **GLMs extend the framework to non-Gaussian responses.** Gamma for severity, Poisson for frequency, Tweedie for pure premium — one framework covers the entire actuarial pricing workflow.

7 **Logistic regression provides exact SHAP values.** Critical advantage for EU AI Act Annex IV documentation and OCC SR 11-7 model validation — analytically exact, not approximate.

8 **PSI and CSI monitoring are mandatory.** PSI > 0.25 triggers mandatory redevelopment under EBA GL/2020/06. Score and characteristic distribution monitoring must be performed monthly.

CAISA™ Level 1 — Chapter 4 of 48 | Part I: Machine Learning Foundations · Week 2