Optimal binning in Logistic Regression

From naive cuts to constrained optimisation in regulated settings

Charaf ZGUIOUAR
zgcharaf@gmail.Com
github : zgcharaf
BNP Paribas
Sorbonne School of Economics
PyData Global Paris
December 10, 2025

Agenda

1. Who I am & why logistic regression?
2. Logistic regression & binning basics
3. Story: Model A vs Model B
4. Case study: one dataset, many binnings
5. Four models on the same dataset
6. Optimal binning & optbinning models
7. Conclusion & further exploration

Who am I?

- Data Scientist at **BNP Paribas**
  - 2023 – Present
- Lecturer at **Sorbonne School of Economics**
  - M2 Finance, Technology & Data (since Sept 2025).
- Past experience **in Treasury risk management and the pharmaceutical industry**
- PyData Paris lurker for the last two years (2024 & 2025), first time speaking.
- Maybe a start a PhD in late 2026?

Disclaimer

- Views are my own and do not represent any employer or institution and all examples are based on public data.

Modeling under operational and regulatory constraints.

- In finance and healthcare, Model and Operational risks can cause severe losses.

JP Morgan trader 'London Whale' blows $13bn hole in bank's value

Shockwaves spread across markets after $2bn trading loss at US bank. Transcript of a JP Morgan Q&A with Jill Treanor, Dominic Rushe and Akshat Tewary

JP Morgan's London office. Photograph: Carl Court/AFP/Getty Images

The City trader at the centre of a $2bn trading loss at JP Morgan Chase had returned to his home in Paris on Friday as the repercussions of the loss spread across the markets.

Some $13bn was wiped off the value of America's largest bank after it admitted the scale of the trading activities of Bruno Iksil - nicknamed the London Whale for his bullish trading - and his colleagues in the bank's little known "chief investment office". The US Securities and Exchange financial watchdog was said to have begun reviewing the losses, the rating agency Standard & Poor's revised its outlook on the bank from stable to negative and Fitch Ratings downgraded it from A-plus to AA-minus.

Contacted by the Guardian, Iksil was reluctant to comment. He was thought to be in Paris and said: "I cannot talk about it. You will have to speak to the bank's representatives."

Figure 1 – Finance: JPMorgan "London Whale" losses.
Source: The Guardian

Duke Suspends Researcher and Halts Cancer Studies

BY NATASHA SINGER JULY 20, 2010 12:43 PM

The Duke University School of Medicine has suspended a researcher and stopped patient enrollment in three cancer studies upon learning of reports that the researcher had overstated his academic credentials.

The lead researcher, Dr. Anil Potti, was placed on administrative leave, said Douglas J. Stokke, a spokesman for Duke, while it investigates allegations that Dr. Potti falsely claimed to have been a Rhodes scholar.

The controversy erupted last week after the Cancer Letter, a weekly publication for cancer specialists, reported that Dr. Potti, an assistant professor of medicine, had padded his résumé on occasion. A spokeswoman at Rhodes House at Oxford University confirmed on Tuesday that Dr. Potti had not received a scholarship.

As news spread of Dr. Potti's problems, the American Cancer Society suspended payments on a five-year, $729,000 grant awarded to Dr. Potti to study the genetics of lung cancer. The society issued the grant based in part on Dr. Potti's résumé, which included a Rhodes scholarship, said Dr. Otis W. Brawley, the chief medical officer of the cancer society.

Figure 2 – Healthcare: Duke cancer genomics trials.
Source: The New York Times

From model risk to modeling choices

What is model risk?

Model risk is the risk of losses or harmful decisions caused by **incorrect, misused, or poorly implemented models** (including data, assumptions, code, and governance).

- In regulated settings, you are not just aiming for **accuracy**:
  - Models must be **stable**, **explainable**, and **documentable**.
  - You need to justify decisions to regulators, independent model reviewers, clinicians,...
- Hence:
  - The effect of each feature can be **challenged**.
  - transformations (such as binning) or combined features are transparent and auditable.
- Historically, **logistic regression** became a workhorse in these domains:
  - Simple, robust, and computationally inexpensive.
  - Coefficients are naturally mapped to odds ratios and risk factors.

Logistic regression recap

Model

$$\Pr(Y = 1 \mid x) = \sigma(\beta_0 + \beta^\top x), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\log \frac{p}{1-p} = \beta_0 + \beta^\top x$$

- The log-odds are a *linear* function of the features.
- Each coefficient β_j is an odds ratio: $\exp(\beta_j)$.
- Great when log-odds vs feature is (almost) linear.

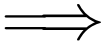
What is binning?

```
(X_train)
✓ 0.0s
```

	age	ldl	tobacco	sbp	adiposity	typea	obesity	alcohol
259	44	7.13	1.80	154	34.04	52	35.51	39.36
328	39	3.20	5.60	106	12.30	49	20.29	0.00
254	46	4.65	9.00	161	15.16	58	23.76	43.20
185	33	3.18	4.40	122	11.59	59	21.94	0.00
152	24	4.19	0.28	122	19.97	61	25.63	0.00
...	...	...	...	...	...	...	...	...
346	53	4.75	4.50	154	23.52	43	25.76	0.00
349	60	4.24	5.60	162	22.53	29	22.91	5.66
166	55	4.99	12.16	110	28.56	44	27.14	21.60
359	32	3.58	1.68	152	25.43	50	27.03	0.00
82	55	6.09	4.80	148	36.55	63	25.44	0.88

323 rows × 8 columns

Raw continuous variables



```
(X_train_B)
✓ 0.0s
```

	age_bin_tree_(14.999, 23.5]	age_bin_tree_(23.5, 31.5]	age_bin_tree_(31.5, 50.5]	age_bin_tree_(50.5, 59.5]	age_bin_tree_(59.5, 64.0]
259	False	False	True	False	False
328	False	False	True	False	False
254	False	False	True	False	False
185	False	False	True	False	False
152	False	True	False	False	False
...	...	...	...	...	...
346	False	False	False	True	False
349	False	False	False	False	True
166	False	False	False	True	False
359	False	False	True	False	False
82	False	False	False	True	False

323 rows × 40 columns

Binned / discretized variables

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

Weight of Evidence (WoE)

- Used to encode binned variables in scorecards and logistic regression.
- For each bin j:

$$\text{WoE}_j = \log\left(\frac{\text{Good}_j/\text{Total Good}}{\text{Bad}_j/\text{Total Bad}}\right)$$

- Interpretable as how much evidence (in log-odds units) that bin gives in favour of "good" vs "bad".

Information Value (IV)

- Global measure of the predictive power of a binned variable.
- Computed by summing over all bins:

$$\text{IV} = \sum_j \left(\frac{\text{Good}_j}{\text{Total Good}} - \frac{\text{Bad}_j}{\text{Total Bad}}\right) \cdot \text{WoE}_j$$

IV as a feature selection metric

Rule of thumb (credit scoring):

- IV < 0.02: not predictive
- 0.02–0.1: weak
- 0.1–0.3: medium
- > 0.3: strong

When log-odds are not linear

- Many risk drivers are:
  - U-shaped (age, blood pressure).
  - Thresholded (lab values, debt-to-income).
  - Very noisy in the tails.
- A single βx cannot capture the plateau + cliff + U-shape.
- We need to "bend" the logarithmic odds in a controlled and explainable way.

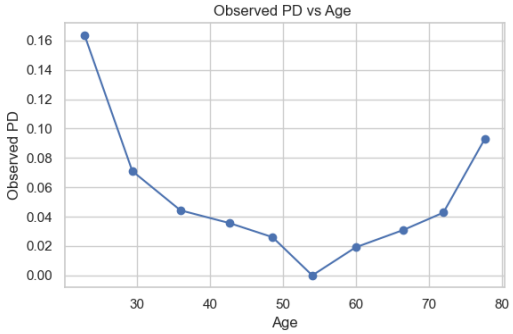


Figure 1: PD vs age (non-linear pattern).

What is binning?

Definition

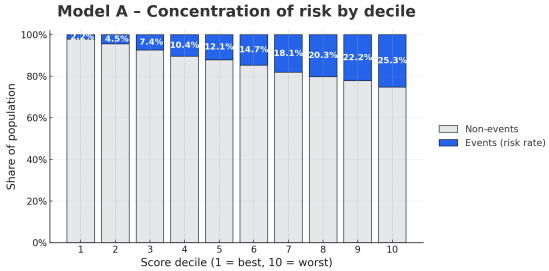
- Map a continuous feature x into K bins:

$$x \mapsto b(x) \in \{1, \ldots, K\}.$$

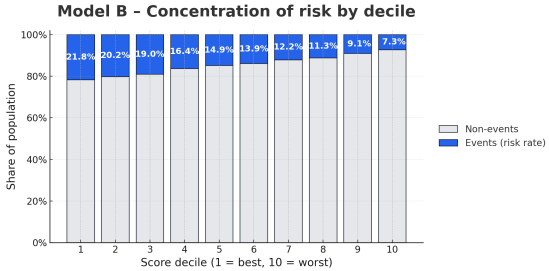
- In the model we use:
  - One-hot dummies for each bin, or
  - Weight-of-Evidence (WoE) encoding of each bin.
- We are approximating the log-odds by a step function.

Model A vs Model B: What is wrong here?

- Context: healthcare-like risk model (readmission probability).



Model A



Model B

Investigating like Sherlock Holmes

- We look at:
  - Data quality and drift.
  - Feature distributions by time period.
  - PD (Probability of default or disease) per decile for key numerical variables.
- We discover:
  - Model B uses naive decile binning everywhere.
  - Several bins have very few events.
  - PD vs bin index is chaotic out-of-time.
- Hypothesis:

Binning choices silently created an unstable model.

Case study dataset

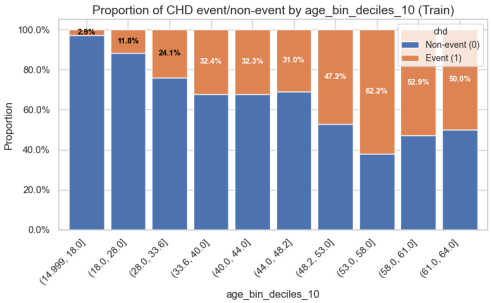
- We are going to use the South African Coronary Heart disease dataset with:
  - 1 = Patient has CHD, 0 = CHD free
- Why this dataset?
  - Resembles a regulated healthcare or credit setting and Clean enough for a talk.

Feature overview

Feature	Min	Max	Mean	Unit / meaning
age	15	64	42.8	years
sbp	101	218	138.3	Systolic blood pressure (mmHg)
tobacco	0.00	31.20	3.64	Cumulative tobacco consumption (kg)
ldl	0.98	15.33	4.74	LDL cholesterol (mmol/L)
adiposity	6.74	42.49	25.41	Adiposity index
famhist	–	–	–	Family history of CHD (Present / Absent)
typea	13	78	53.1	Type A behaviour score
obesity	14.70	46.58	26.04	Obesity index
alcohol	0.00	147.19	17.04	Current alcohol intake (avg. drinks/day)
chd	0	1	0.35	Coronary heart disease (1 = event)

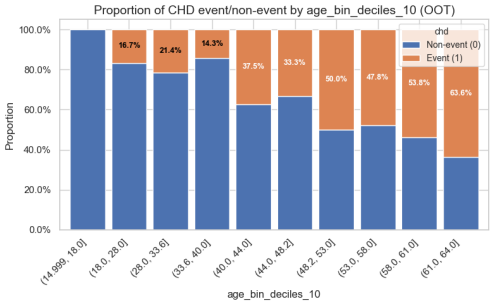
Table 1: Summary of selected features in the heart disease dataset ($n = 462$).

Age vs CHD risk – Decile (Quantile) Binning



Train sample

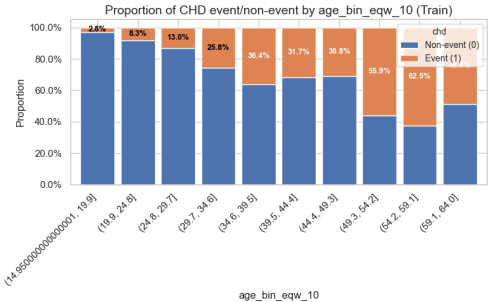
$\Rightarrow$



Out-of-time sample

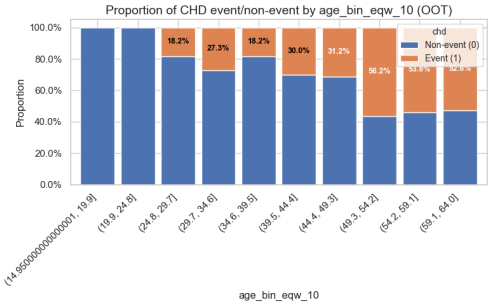
- Each bin has similar size on train, but the risk curve is noisy and not strictly monotone, and this instability carries over to the OOT sample.

Age vs CHD risk – Equal-Width Binning



Train sample

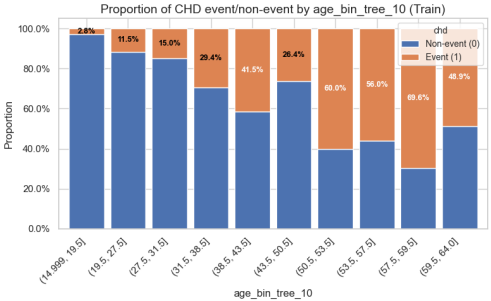
⇒



Out-of-time sample

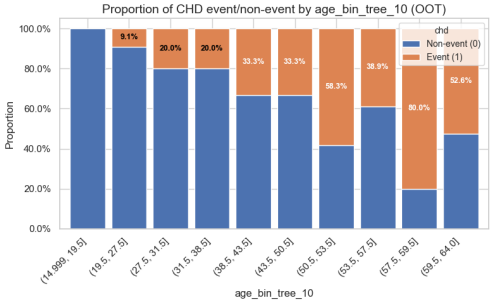
- Bins cover equal age ranges, but sample sizes and risk stability vary a lot, especially at the extremes, and OOT behaviour can diverge in the sparsely populated bins.

Age vs CHD risk – Tree-Based Binning



Train sample

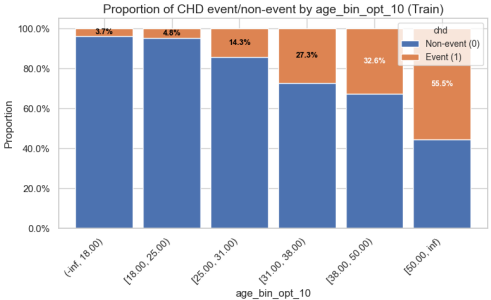
⇒



Out-of-time sample

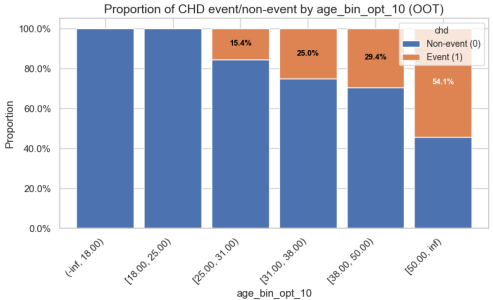
- Splits are placed where risk changes most on train, giving clearer groups, but the pattern is not guaranteed to remain monotone or stable OOT.

Age vs CHD risk – Optimized Binning



Train sample

⇒



Out-of-time sample

- OptimalBinningSketch finds bins that maximise information value under constraints, yielding a smooth, interpretable risk curve that remains much more stable OOT.

Four modelling approaches we will compare

1. **Classic logistic regression** (no binning).
2. **Logistic regression + naive binning**
   - Equal-width or quantile bins everywhere.
3. **LightGBM**
   - Gradient boosting, strong non-linear benchmark.
4. **Logistic regression + optimal binning**
   - Binning done with optbinning / optimal algorithms.

AUC & ROC comparison

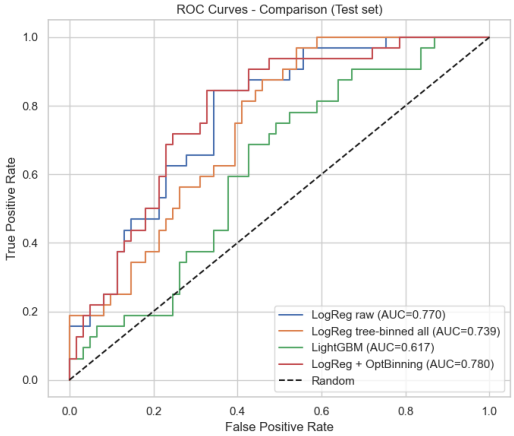


Figure 2: ROC curves (Test set).

Model	AUC_val	AUC_test
LogReg + OptBinning	0.7833	0.7802
LogReg raw	0.7976	0.7700
LogReg tree-binned all	0.6936	0.7387
LightGBM	0.6793	0.6168

Table 2: Indicative performance on validation and test sets.

How Boosting Algorithms Handle "Binning" (1/2)

Implicit binning in tree-based boosting

- Gradient boosting libraries (e.g. XGBoost, LightGBM, CatBoost) use decision trees as base learners.
- Each split of the form $x \leq t$ vs $x > t$ creates *bins* (intervals) on the feature x.
- With many trees and depths, you obtain a very fine, non-linear partition of the feature space:

$$\text{prediction} = \sum_{m=1}^{M} \text{Tree}_m(\text{x})$$

- Some implementations even pre-bucketize features internally (histogram-based algorithms).

How Boosting Algorithms Handle "Binning" (2/2)

Why this is not always perfect

- Bins are **model-dependent**: change the hyperparameters or random seed and you may change the effective "binning".
- No guarantee of **monotonicity** or simple shapes (risk vs feature can oscillate in unintuitive ways).
- Harder to **explain and document**: hundreds of trees and thresholds vs a small, stable set of human-readable bins.

- In regulated settings, we often need:
  - Stable, monotone risk curves;
  - A clear justification for each bin boundary.

Optimal binning as an optimisation problem

For one feature X and binary Y

$$\max_{\text{bins}} \ \text{Score}(\text{bins}) \quad \text{s.t.} \quad \begin{cases} \text{size}(\text{bin}) \geq n_{\min} \\ \text{events}(\text{bin}) \geq e_{\min} \\ \#\text{bins} \leq K_{\max} \\ \text{PD (Probability of Default) is monotone (optional)} \end{cases}$$

- Score can be: Information Value, mutual information, log-likelihood, Gini, etc.
- This is solved by different algorithms, including those in `optbinning`.

MDLP: entropy-based discretisation

- **Minimum Description Length Principle (MDLP)**:
  - See binning as compressing the joint distribution of (X, Y).
  - Choose cutpoints that maximise information gain minus a complexity penalty.
- Algorithm:
  - Greedy, recursive splitting where information gain is highest.
  - Stop when extra splits don't justify added complexity.
- Pros:
  - Target-aware, scale-free, relatively fast.
- Cons:
  - Local optimum.
  - Harder to enforce monotonicity and strict min-event constraints.

Mathematical programming-based optimal binning

- Approach used in **optbinning**:
  - Pre-bin the feature into many small intervals.
  - Use a Mixed-Integer optimisation problem to merge pre-bins.
- Optimisation decides:
  - Which adjacent pre-bins to merge.
  - Under explicit constraints on size, events, trend, #bins.
- Strengths:
  - Global optimum for the given objective.
  - Constraints are guaranteed, not "approximately satisfied".
  - Works naturally with monotonicity (increasing or decreasing PD).

Stochastic optimal binning

- Deterministic optimal binning can still overfit one sample.
- Stochastic variants:
  - Bootstrap the data; fit binning multiple times.
  - Add noise to the objective/constraints and aggregate.
- Why do this?
  - Assess how stable cutpoints are across resamples.
  - Prefer bins that appear consistently across runs.
- In a regulated setting:
  - You can show that cutpoints are not artefacts of a single dataset snapshot.

What "good" looks like: back to our story

- In the case study:
  - Logit + optimal binning matched or approached LightGBM performance.
  - Bins were few, monotone, and stable over time.
- In Model A vs Model B:
  - Model A used thoughtful, constraint-aware binning.
  - Model B used naive decile binning everywhere.
  - Only Model A survived six months of real-world drift.
- Evidence: **good binning is boringly stable and easy to explain**.

Conclusion & how to explore further

- Treat every binning decision as part of the model, not neutral preprocessing.
- Before chasing AUC, decide:
  - Minimum bin size and events.
  - Whether PD must be monotone for that feature.
  - Maximum number of bins you can explain to a human.
- Try this at home:
  - Reproduce the four-model comparison on your own data.
  - Plot PD per bin on at least two time slices (train & OOT).
  - Experiment with `optbinning` and MDLP-style discretisation.

OptBinning library

- **OptBinning** is a Python library for optimal binning and scorecard modelling.
- Created and maintained by **Guillermo Navas-Palencia**.
- Implements mathematical programming formulations for:
  - Binary, continuous and multiclass targets.
  - Monotonicity, minimum size, and other business constraints.

Links

- Documentation: gnpalencia.org/optbinning
- GitHub repository: github.com/guillermo-navas-palencia/optbinning

Stay in touch

Scan to get the slides on GitHub

Scan to connect on LinkedIn

Questions?

Thank you!

Questions, comments ?

(zgcharaf@gmail.com / Github : Zgcharaf)