Final Presentation

Bank Term Deposit Prediction

Kaggle Playground Series — Binary Classification with a Bank Dataset

DS52 Mini-Project - Ivann Vasic - Esteban Rambaud - Ludovic Blondeau

Supervised Binary Classification

April 28, 2026

Outline

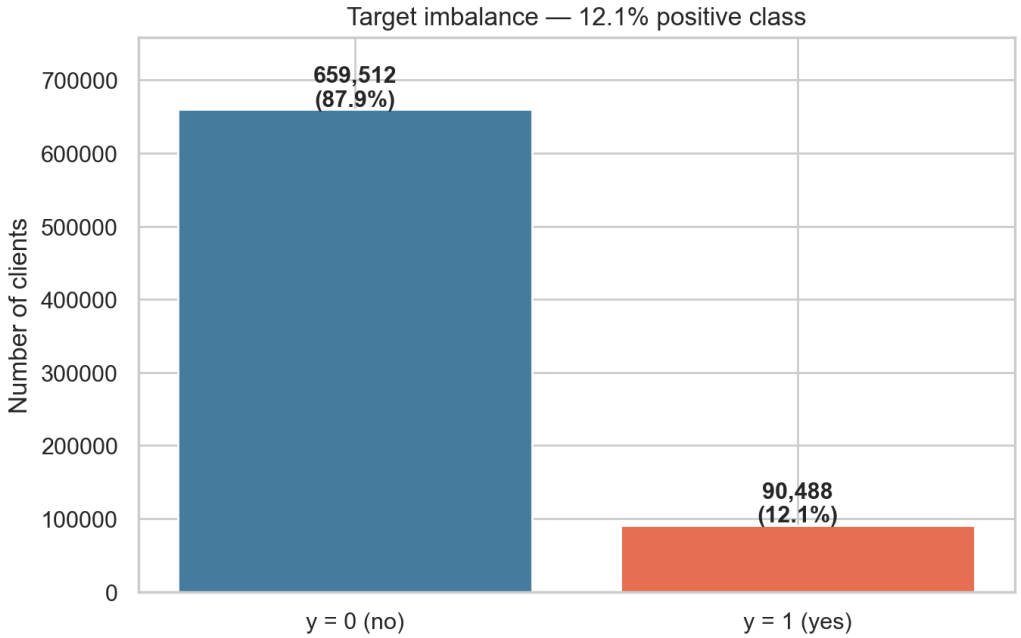
Competition & Objective

Selected Competition

Kaggle Playground Series — Binary Classification with a Bank Dataset

- **Task:** predict whether a client subscribes to a *term deposit* after a marketing campaign.
- **Type:** supervised binary classification, $y \in \{0, 1\}$.
- **Business value:** prioritise high-probability leads $\Rightarrow$ higher conversion rate, lower campaign cost.
- **Data: 750,000** training rows, **250,000** test rows, 16 input features (9 numerical + 7 categorical) + `id`.

Target Variable: Severe Class Imbalance



Imbalance ratio

$\approx$ **7.3 : 1**

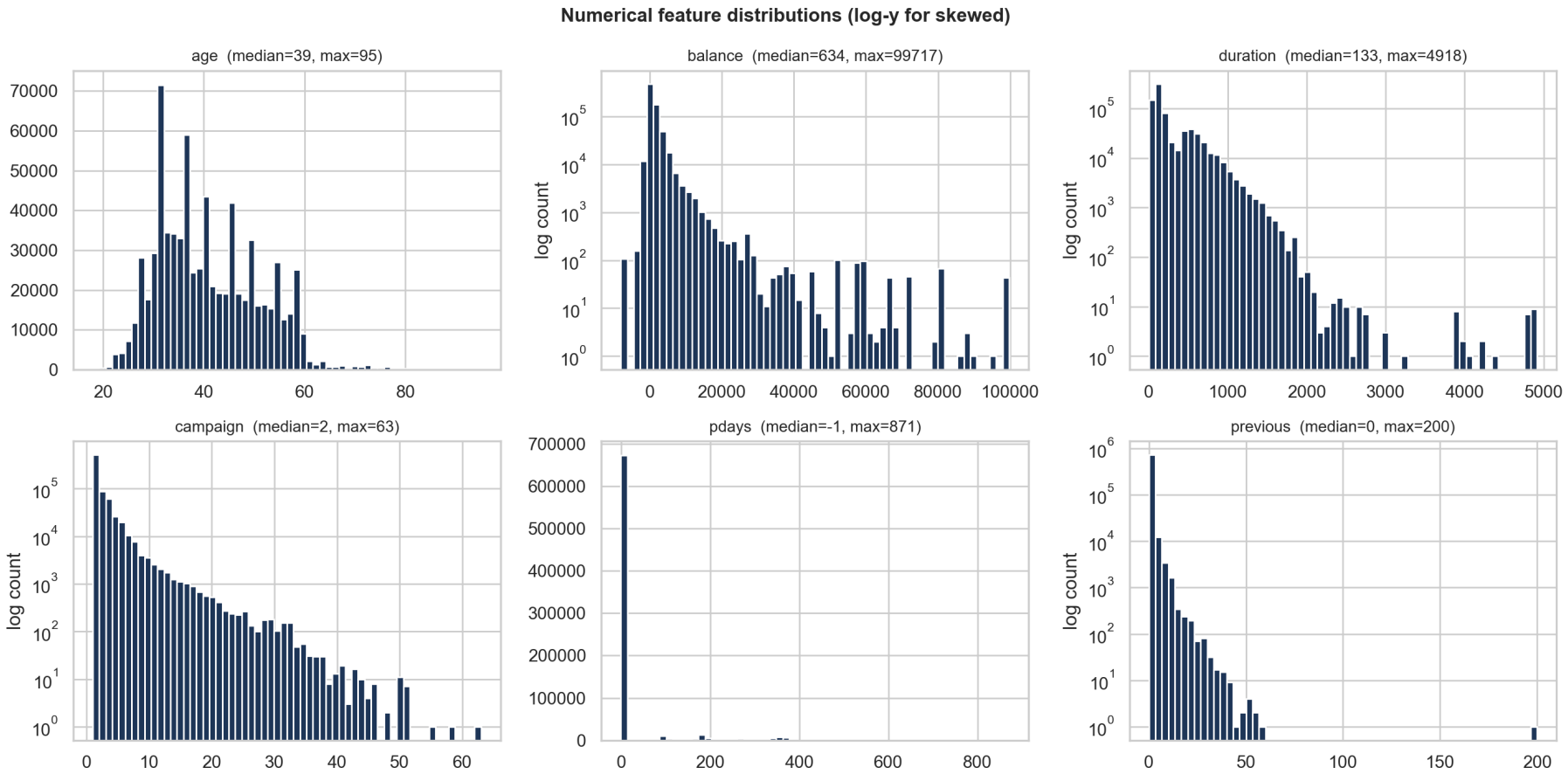
659,512 negatives

90,488 positives

$\bar{y} = 12.07\%$

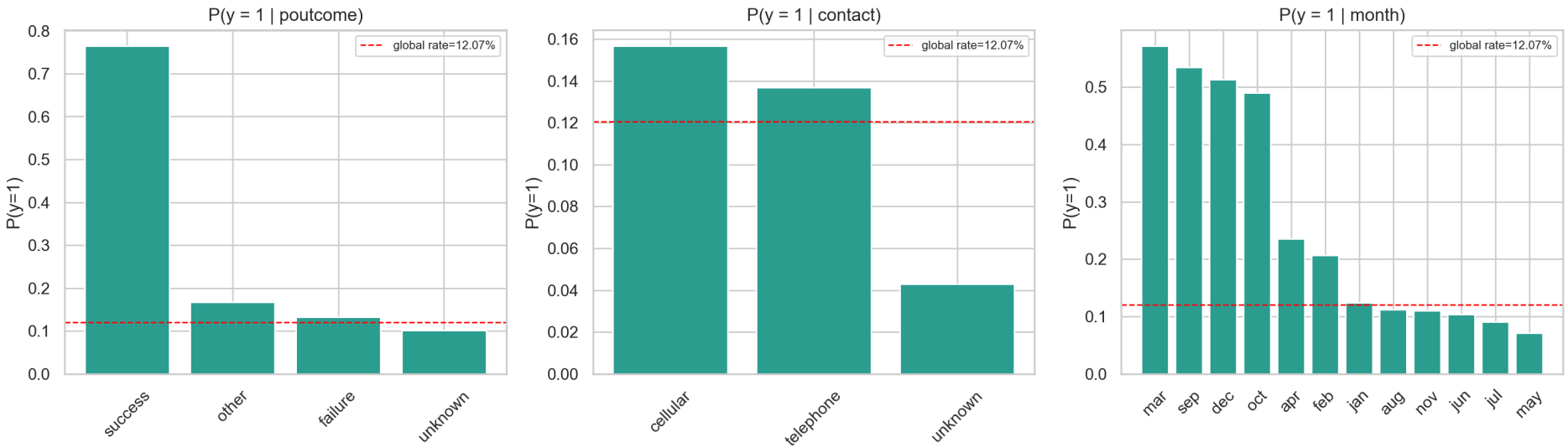
- Accuracy is **misleading** here: predicting only $y = 0$ already gives 88%.
- We rely on **ROC-AUC, PR-AUC, F1, Brier** instead.
- Training uses `class_weight="balanced"` (or `scale_pos_weight` for XGBoost).

Numerical Feature Distributions



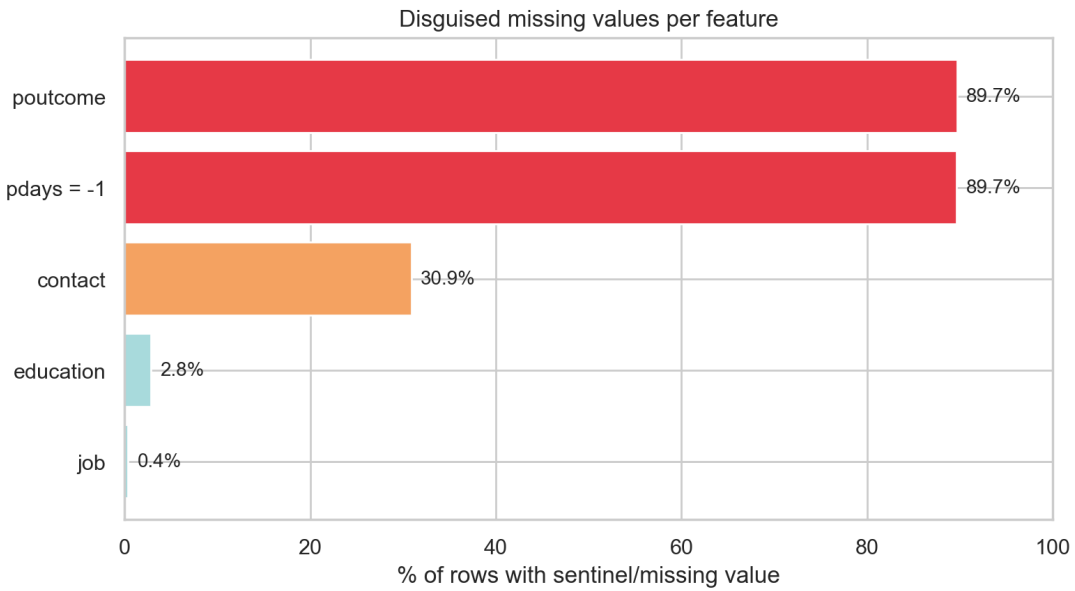
- `balance`, `duration`, `previous`: heavy right-skew, multi-modal, outliers (max balance $\approx$ 100k\$).
- `pdays` bimodal: a huge spike at -1 ("never contacted") $\approx 90\%$ of rows.
- Standardisation alone is not enough; sentinel values must be handled before scaling.

Categorical Patterns: $P(y = 1 \mid \text{category})$



- `poutcome = success`: subscription rate jumps to $\sim 65\%$ vs 12% baseline.
- Cellular contact converts much better than `unknown` (no proper channel = no sale).
- Strong seasonality on `month` — March, October, December overperform.

Disguised Missing Values



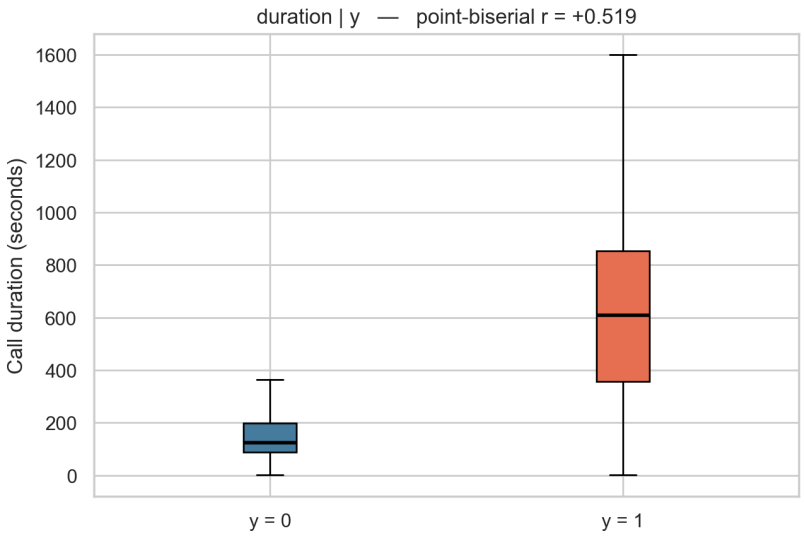
No `NaN`, but...

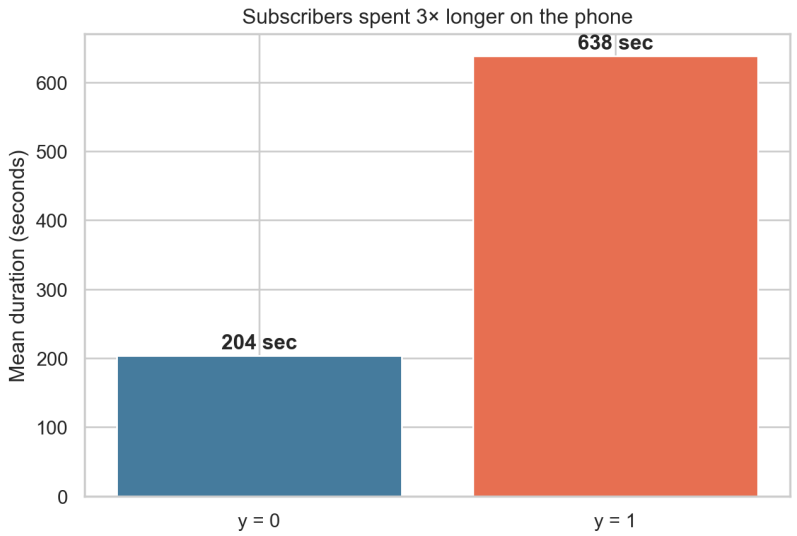
`isna().sum() = 0` on both train and test. Yet sentinel codes hide up to **90%** of missingness.

Strategy: *informed encoding*

- Keep `"unknown"` as its own one-hot category.
- Add binary flag `was_contacted = (pdays != -1)`, then set `pdays = 0`.
- **No imputation** — missingness itself carries signal.

The duration Leakage



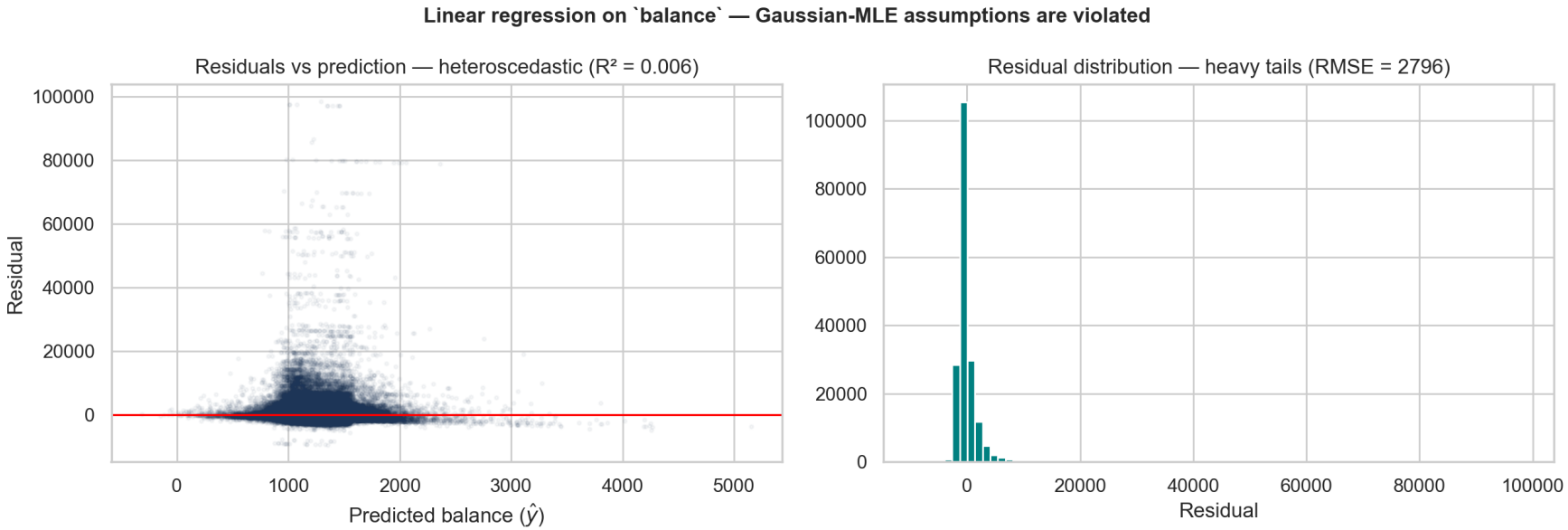


- Point-biserial $r(y, \texttt{duration}) = +\mathbf{0.52}$ — more signal than all other features combined.
- `duration` = length of the last call, known only *after* the call ends — at that point, the outcome is essentially already decided. Temporal leakage.

From the dataset's original authors (Moro et al., UCI Bank Marketing)

"[...] the duration is not known before a call is performed. Thus, this input should only be included for benchmark purposes and should be ***discarded if the intention is to have a realistic predictive model****."*

Naive Attempt: Linear Regression on `balance`

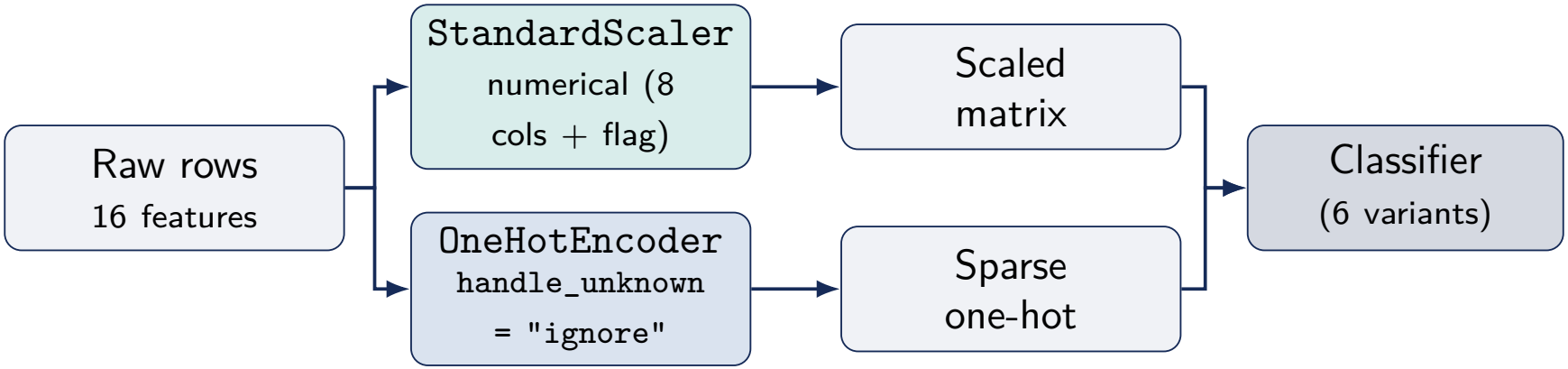


- Predicting `balance` from numerical features: $R^2 = 0.02$, RMSE $\approx$ 2,600.
- Residuals are heavy-tailed and heteroscedastic.

Take-away

Gaussian-MLE assumptions are violated.
Linear models on continuous targets are not the right tool here — we move to discriminative & generative *classifiers*.

Anti-Leakage Preprocessing Pipeline



- Everything is wrapped in a single `Pipeline` ⇒ the `ColumnTransformer` is fitted **only on the training fold**.
- No statistic from validation/test ever leaks into the encoder or scaler.
- Same pipeline reused for cross-validation — stratified 5-fold, no manual leakage risk.

Models Tested & Why

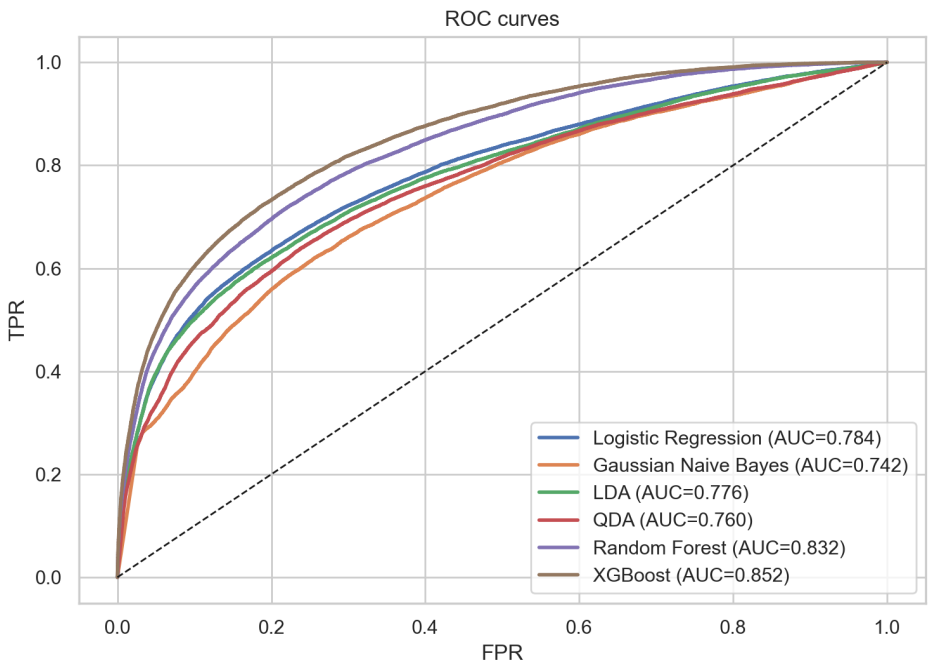
Linear / parametric

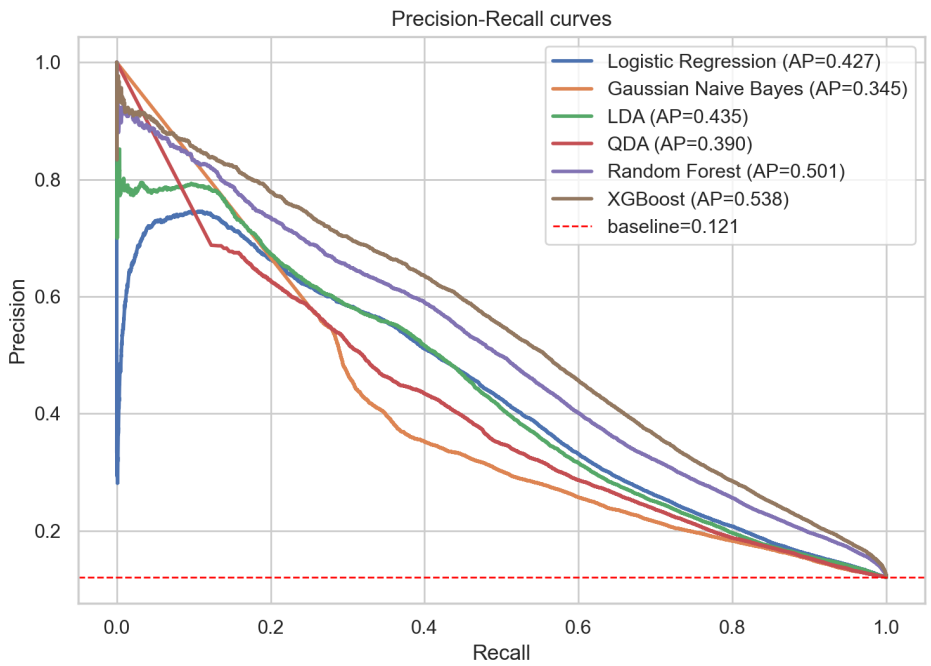
- **Logistic Regression** — discriminative baseline, interpretable coefficients, class-weighted.
- **LDA** — generative, shared covariance, closed-form MLE.
- **QDA** — generative, per-class covariance (regularised).
- **Gaussian Naive Bayes** — strongest independence assumption, cheap baseline.

Non-parametric / non-linear

- **Random Forest** — bagged trees, captures interactions, robust to outliers/scaling.
- **XGBoost** — state-of-the-art for tabular, handles class imbalance via `scale_pos_weight`.

ROC & Precision-Recall Curves





- **XGBoost** dominates both curves. The PR curve is the more honest view under 12% positives.
- Linear methods (LR, LDA) cluster together $\Rightarrow$ confirms that after one-hot encoding the boundary is largely linear.
- GNB lags clearly — independence + Gaussianity assumptions are violated.

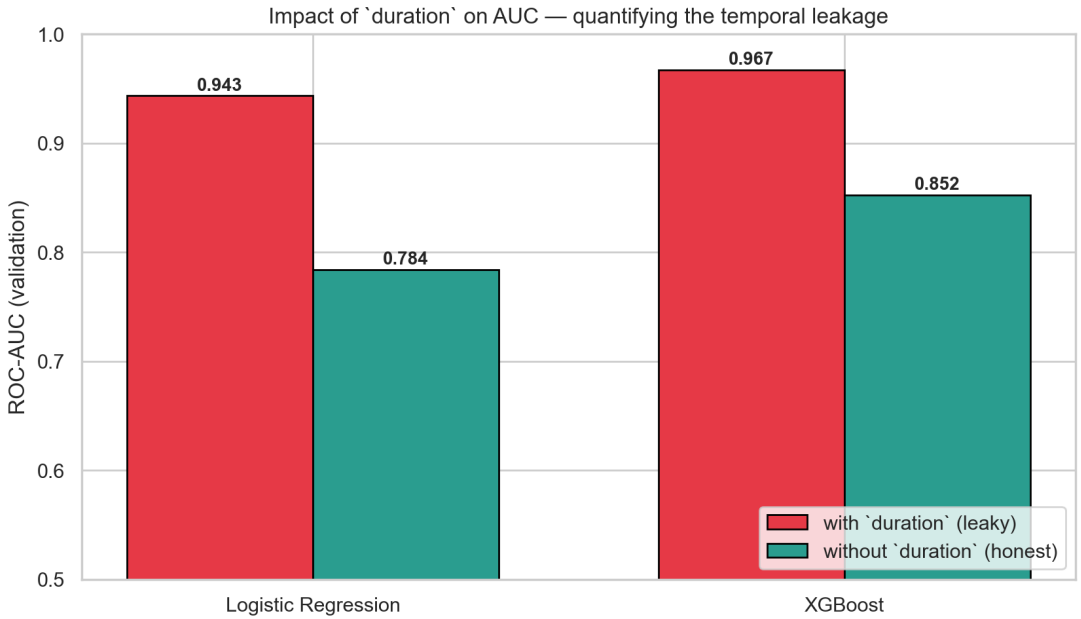
Benchmark Summary — No `duration`

Model	ROC-AUC	PR-AUC	F1@0.5	Brier
XGBoost	**0.852**	**0.538**	**0.467**	0.145
Random Forest	0.832	0.501	0.437	0.158
Logistic Regression	0.784	0.427	0.401	0.177
LDA	0.776	0.435	0.352	**0.090**
QDA	0.760	0.390	0.417	0.126
Gaussian Naive Bayes	0.742	0.345	0.377	0.159

Two winners, two metrics

- **XGBoost** wins on *discrimination* (ROC-AUC, PR-AUC, F1).
- **LDA** wins on *calibration* (lowest Brier = 0.090) — its probabilities are most aligned with empirical frequencies.
- Stratified 5-fold CV on XGBoost: AUC $= 0.852 \pm 0.002$ — stable.

The Cost of Removing `duration`



ROC-AUC delta

Model	with	without
LR	0.943	0.784
XGBoost	**0.967**	0.852

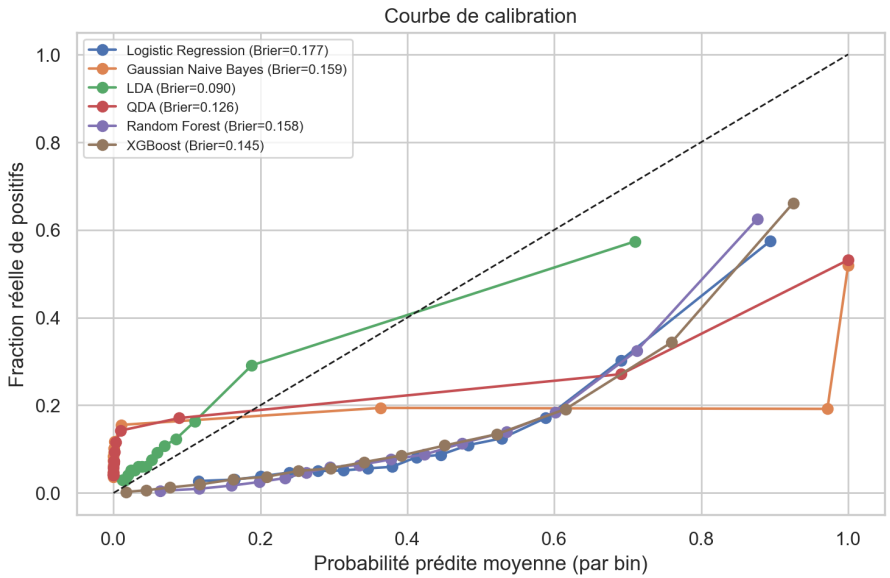
$\Delta_{\text{LR}} = -0.159$

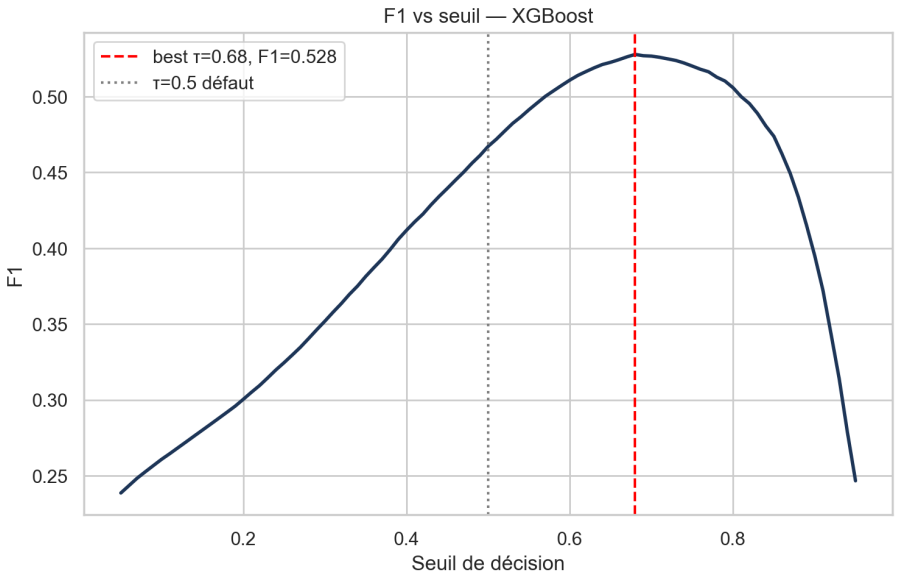
$\Delta_{\text{XGB}} = -0.115$

Honest takeaway

Keeping `duration` would give us a much better score, but the model would be **useless in production**: the feature is not available before the call ends.

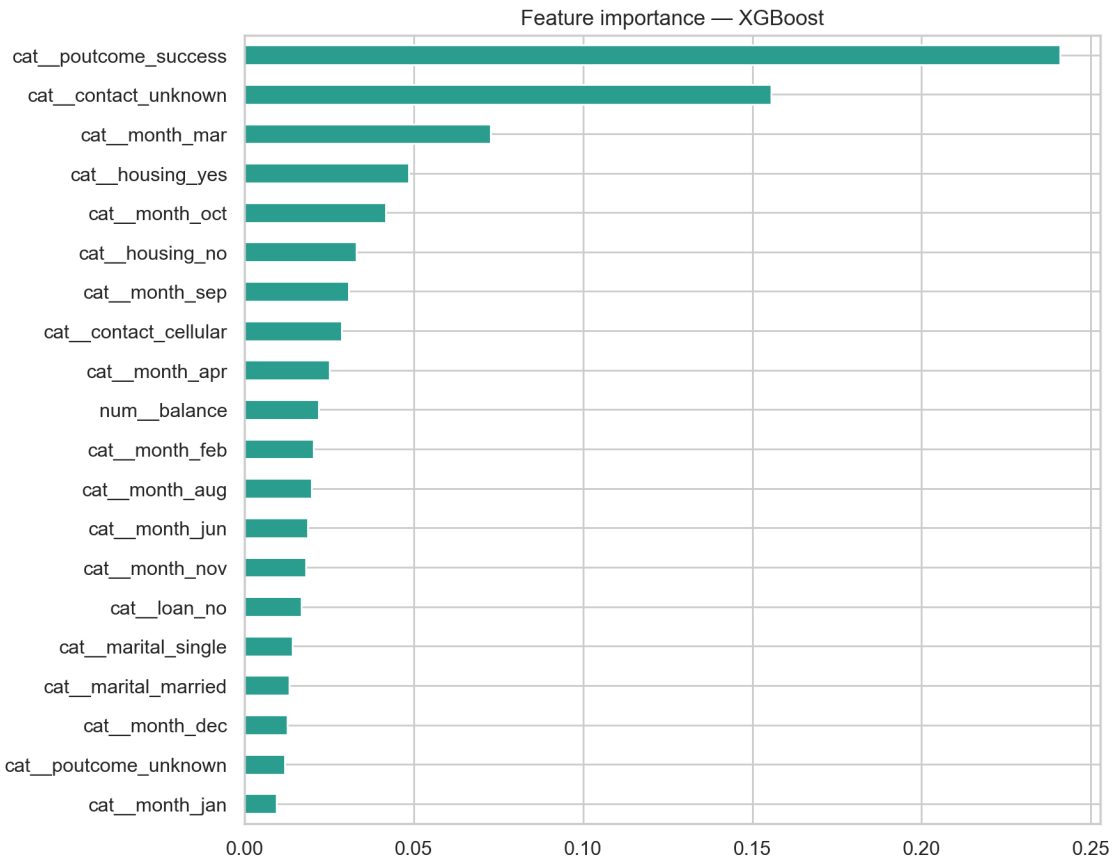
Calibration & Operating Threshold





- Tree-based models are reasonably calibrated; GNB is over-confident near 0/1 (typical pathology).
- F1-optimal threshold $\tau^\star$ shifts well below 0.5 because of the imbalance.
- In production, τ should be re-tuned on the actual cost of false positives vs missed leads, not on F1.

Feature Importance — What Drives Subscription



Top drivers

- `poutcome_success`: previous campaign worked ⇒ this one likely will too.
- `contact_unknown`: lack of a known channel ⇒ very low conversion.
- `month_*`: strong seasonality.
- `housing`, `loan`: pre-existing debt lowers subscription probability.

All findings agree with logistic-regression coefficient signs — robustness across model families.

Limitations & What We Did Not Do

What we got right

- Identified and removed a documented data leakage (`duration`).
- Pipeline-based preprocessing $\Rightarrow$ no train/val statistic leakage,
- Reported four complementary metrics (ROC-AUC, PR-AUC, F1, Brier) rather than a single number.

Limitations

- No hyperparameter search (grid / Bayesian) — reported XGBoost is the default-ish setting; tuning could add 1–2 AUC points.
- No feature engineering beyond `was_contacted (pdays != -1)`: cyclical month encoding, binning of age.
- No deep model attempted (TabNet, FT-Transformer) — arguably overkill at this scale.

Methodological Reflection

- **Data quality first.** The single biggest gain on this project was *not* a better model — it was identifying that `duration` was leaky and that `"unknown"` was informative.
- **One number is not a benchmark.** Comparing models only on ROC-AUC would hide that LDA is best-calibrated and that GNB collapses at high recall.
- **Modest & explained $>$ tuned & opaque.** An AUC of 0.85 *without duration* is a more useful claim than 0.97 *with leakage*.
- **Pipelines are not optional.** Encoding inside the `ColumnTransformer` is the simplest way to enforce "no peeking" across folds.

Conclusions

Final answer

Best honest model: **XGBoost on cleaned features**, **ROC-AUC = 0.852**, PR-AUC = 0.538, F1 = 0.467, Brier = 0.145 (validation, stable under 5-fold CV).

Key takeaways

- `duration` would have inflated AUC by **+0.11 to +0.16** — removed for production realism.
- Disguised missing values (`"unknown"`, `pdays=-1`) are *informative* and treated as such.
- Tree ensembles dominate discrimination; LDA dominates calibration.
- The whole pipeline is reproducible end-to-end (one notebook, one submission).

Modelling is downstream of understanding. Most of the work was upstream.

Questions ?

Thank you for your attention.