Aula 2 – Dados Elegíveis & Amostragens

> **Mentoria em Risco de Crédito & Ciência de Dados** Material de apoio para impressão (PDF) Versão [1] – 13 jul 2025

Sumário

Visão geral

Esta aula aprofunda os fundamentos de **dados elegíveis** e **amostragem** para modelos de risco de crédito. O foco recai sobre:

- diferenças entre dados necessários para PD, LGD e EAD;
- escolha entre amostragem *snapshot* e *painel*;
- efeitos de *right censoring* e *left truncation*;
- uso responsável de *sample weights*;
- boas práticas para definir *targets* (Ever 90 M12, Over 90 M12 etc.).

Todo o conteúdo foi refinado para remover informações sensíveis (nomes próprios, empresas, exemplos internos) e corrigir pequenos deslizes identificados na transcrição.

Conceitos-chave

Modelos PD, LGD, EAD & dados elegíveis

Modelo	Pergunta-chave	Dados elegíveis (***performing***)
PD	Qual a probabilidade de o contrato entrar em *default* nos próximos 12 meses?	Registros **sem** *default* ou *workout* até a data-base. Incluem atrasos < 90 dias.
LGD	Quanto será recuperado após o *default*?	Registros **em** *default* ou **em workout** (negociação/recuperação). Período de *cura* não integra a amostra.
EAD	Qual exposição haverá no momento do *default* futuro?	Mesmos dados de PD **mais** o registro do **primeiro** *default* de cada contrato.

> **Workout & Cura** *Workout* inicia quando o devedor demonstra esforço de pagamento (ex.: acordo, amortização parcial). *Cura* é o intervalo (tipicamente ≥ 3 meses) em que o contrato permanece sem atraso após zerar o saldo vencido, antes de voltar ao status *performing*.

Regimes de política de crédito

Mudanças relevantes de política podem distorcer o modelo. Estratégias recomendadas:

- Manter **máximo de dados** e identificar regimes via **flags** (variável *dummy*);
- Avaliar estabilidade com **KS**, **PSI** ou compressão (**PCA**, **t-SNE**) entre safras;
- EBA recomenda **≥ 5 anos** de histórico para carteiras de varejo (CRD Art. 144–147).

Amostragem

Snapshot

Seleciona uma única observação de cada contrato.

Vantagens	Desvantagens
Implementação simples; menor custo computacional; útil p/ *application score*.	**Viés de recência** (safras recentes super-representadas); ignora dinâmica comportamental; reduz peso de variáveis históricas.

> **Quando usar**: modelos de concessão, testes rápidos, base com vida muito curta.
> **Evitar**: modelos comportamentais (PD dinâmico, LGD, EAD), monitoramento.

Painel

Mantém o filme completo da vida do contrato.

Vantagens	Desvantagens

Vantagens	Desvantagens
Captura evolução de risco e comportamento; adequado a vintage analysis, backtesting & monitoramento; reduz viés de recência.	Demanda maior processamento; requer cuidado com grupos *performing* e maturidade das safras.

> **Quando usar**: PD comportamental, LGD, EAD, análises de safra, modelos de sobrevivência.

Right Censoring & Left Truncation

Conceito	Risco	Boa prática
Right Censoring – safras mais recentes não completaram o horizonte do target.	Subestimar taxa de mau; superestimar recuperação.	Excluir safra < H-target (ex.: remover últimos 12 meses se alvo = 90 M12) ou calibrar.
Left Truncation – falta de histórico pré-corte.	Variáveis históricas truncadas; valores ausentes concentrados.	Iniciar amostra após janelas históricas críticas (ex.: excluir primeiros 6 meses).

Sample Weights (pesos de amostra)

Objetivos principais:

1. **Balancear classes** em problemas desbalanceados (ex.: inadimplência ≈ 10 %);
2. **Ponderar recência** sem duplicar efeitos (não aplicar se amostra já foi recency-sampled);
3. Aplicar pesos manuais (ex.: receita), mantendo prudência regulatória.

```
model.fit(X, y, sample_weight=w)  # vetores customizados > class_weight="balanced"
```

> ✱ **Atenção**: não combine *recency sampling* **e** peso por recência – risco de duplicar o mesmo viés.

Construção de Targets

Target	Definição	Observações
Over 90	90 + dias de atraso no registro atual.	Segmenta *default* corrente (mau na origem).
Ever 90 M12	Algum atraso ≥ 90 d nos 12 m após o registro.	Necessita janela futura completa; sensível a *right censoring*.
Ever 90 M6 (ou *60 M6*, *30 M4*)	Alternativa quando histórico é curto.	Aumenta número de safras maduras.

Exemplo de timeline simplificado

```
┌──────────┬──────┬──────────┬──────────────┐
│ Mês      │ DPD  │ Over 90  │ Ever 90 M12  │
├──────────┼──────┼──────────┼──────────────┤
│ t0       │   0  │ 0        │ ?            │
│ …        │ …    │ …        │ …            │
│ t+7      │ 120  │ 1        │ 1 (retro)    │
│ …        │ …    │ …        │ …            │
└──────────┴──────┴──────────┴──────────────┘
```

Após *workout* e **cura ≥ 3 m**, o contrato volta para o grupo *performing* e ganha um novo *spell_id* (vida 2) para futuros modelos.

Recomendações Práticas

1. **Defenda o maior volume possível** de dados; só reduza após análises de estabilidade e qualidade.
2. **Use flags** de regime de política em vez de descartar safras inteiras.
3. **Corte à direita** conforme horizonte do target para evitar *understatement* de risco.
4. **Evite SMOTE e congêneres** em crédito – riscos de gerar perfis incoerentes. Prefira *sample weights*.
5. **Monitore o modelo** continuamente (input drift, performance drift); todo modelo degrada.
6. **Documente** cada passo (*dataset*, filtros, amostragem, pesos, targets) para auditoria.

Erros Comuns & Armadilhas

- Tratamento igual de variáveis estáticas & históricas em *snapshot*.
- Reutilizar pesos de recência após *recency sampling*.
- Modelar PD com poucas safras maduras (*right censoring*).
- Confiar cegamente no *class_weight="balanced"* – opções manuais são mais transparentes.
- Usar métricas de KS sem verificar *PSI* ou distribuição temporal.

Glossário

Termo	Definição rápida
DPD	*Days Past Due* (dias em atraso).
Ever 90	Indicador se houve atraso ≥ 90 d em qualquer momento.
Over 90	Indicador de atraso ≥ 90 d no momento corrente.
Workout	Fase de renegociação/recuperação após *default*.
Cura	Período sem atraso após zerar saldo vencido (tipicamente ≥ 3 m).
Right Censoring	Safras sem horizonte futuro completo.
Left Truncation	Corte de histórico anterior ao período de análise.
Sample Weight	Peso aplicado a cada observação no treinamento.

Referências Sugeridas

1. European Banking Authority – *Guidelines on PD and LGD estimation* (EBA/GL/2017/16).
2. Basel Committee on Banking Supervision – *IRB Approach: Supporting Document* (2023).
3. Hand, D. J.; Henley, W. E. – *Statistical Classification Methods in Consumer Credit Scoring* (1997).
4. Siddiqi, N. – *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring* (2nd ed.).
5. Brown, I.; Mues, C. – *An Experimental Comparison of Classification Algorithms for Imbalanced Credit Risk Data* (ESWA 2012).
