

Modelos de Credit Scoring

Walter C. Neto

8 de dezembro de 2025

Resumo

Este documento descreve a construção de um modelo de *credit scoring* de contratos PJ a partir da base PEJOTA.xls . São apresentadas as etapas de exploração da base, preparação dos dados, cálculo de Information Value (IV), binning e transformação em Weight of Evidence (WoE), imposição de monotonicidade e agrupamento de bins, estimação de regressão logística, construção do scorecard na escala de pontos e avaliação do desempenho (AUC, KS e PSI). Incluem-se as principais equações teóricas e gráficos gerados no processo (ROC, KS, PSI e WoE por variável). Esse documento apenas auxilia em algumas etapas, não exaustivo.

1 Exploração Inicial da Base

No primeiro script em `Python`, foi feita a leitura do arquivo `PEJOTA.xls` e listadas as abas disponíveis. A aba **VARIÁVEIS** contém o dicionário; a aba **PEJOTA** contém os dados.

Principais passos:

- visualização do dicionário (tipo, período, descrição);
- visualização das primeiras linhas da base;
- inspeção dos *dtypes* (numérico, texto, etc.);
- cálculo do percentual de missing por variável.

A coluna **STATUS** é o rótulo original, com valores **BOM** e **MAU**. Criou-se a variável:

$$\text{TARGET} = \begin{cases} 1, & \text{se STATUS} = \text{MAU}, \\ 0, & \text{se STATUS} = \text{BOM}. \end{cases}$$

Essa codificação é padrão em modelos de inadimplência: 1 representa o *evento de risco* (ativo problemático).

2 Preparação dos Dados

Os principais tratamentos efetuados foram:

1. **Padronização de nomes de colunas:** conversão para maiúsculas, remoção de espaços e substituição por “_”.
2. **Conversão de variáveis financeiras** (IFIN1, IFIN2, IFIN4, IFIN5, IFIN7, IFIN8) que vinham como texto com vírgula decimal. Foi feita a troca vírgula \rightarrow ponto e extraída a parte numérica.
3. **Tratamento de missing:** para IFIN1 e IDADE, utilizou-se a mediana da amostra, apenas para permitir o cálculo de WoE posteriormente.
4. **Codificação de UF:** a variável UF foi transformada em dummies, resultando em indicadores como UF_RJ, UF_SP, etc., com *drop first* para evitar multicolinearidade exata.
5. **Remoção de identificadores** (COD_CLI) e do próprio STATUS, mantendo apenas TARGET.

Ao final, obtém-se uma matriz de entrada com variáveis numéricas e dummies, prontas para a etapa de *binning*.

3 Information Value (IV)

O Information Value mede a força de separação de cada variável individualmente entre *bons* e *maus*. Para uma variável categorizada em K grupos, tem-se:

$$\text{WoE}_k = \ln \left(\frac{\text{Good}\%_k}{\text{Bad}\%_k} \right),$$

$$\text{IV} = \sum_{k=1}^K (\text{Good}\%_k - \text{Bad}\%_k) \cdot \text{WoE}_k.$$

Valores típicos de interpretação (Siddiqi, 2006):

- $\text{IV} < 0,02$: variável fraca;
- $0,02 \leq \text{IV} < 0,1$: fraca a média;
- $0,1 \leq \text{IV} < 0,3$: boa;
- $\text{IV} \geq 0,3$: muito forte (pode sinalizar *overfitting* ou variável “quase determinística”).

Neste trabalho, foram inicialmente selecionadas as variáveis com $\text{IV} > 0,02$. Essa lista foi filtrada posteriormente pela monotonicidade.

4 Binning e Weight of Evidence (WoE)

Para cada variável X_j selecionada, aplica-se o algoritmo `woebin` do `scorecardpy`, que:

- encontra pontos de corte para dividir a distribuição em bins;
- garante frequência mínima por grupo;
- calcula os valores WoE e IV de cada variável.

Seja $b = 1, \dots, K_j$ os bins da variável X_j :

$$\text{WoE}_{jb} = \ln \left(\frac{\text{Good}\%_{jb}}{\text{Bad}\%_{jb}} \right).$$

A transformação de WoE tem duas vantagens principais:

1. lineariza (aproximadamente) a relação entre a variável e o log-odds da inadimplência;
2. produz coeficientes de fácil interpretação: sinais e magnitudes consistentes.

5 Imposição de Monotonicidade

Para variáveis de risco, espera-se que, à medida que o valor do bin “piora” (por exemplo, mais atraso, mais restrições, menor renda), o WoE ou o *bad rate* cresça (ou decresça) de forma consistente. Esse ajuste deve ser analisado em conjunto com o sentido econômico esperado da variável, mantendo coerência de ordenação risco para os contratos escorados.

Para cada variável, o algoritmo implementado faz:

1. separa explicitamente o bin *missing* (não entra na checagem de monotonicidade);
2. calcula, para cada bin não-missing, o bad rate

$$\hat{p}_{jb} = \frac{\text{Bad}_{jb}}{\text{Total}_{jb}};$$

3. determina a tendência global (crescente ou decrescente) comparando $\hat{p}_{j\text{último}}$ com $\hat{p}_{j\text{primeiro}}$;
4. enquanto houver violações à tendência global, combina bins adjacentes (soma de contagens e de maus), atualiza \hat{p}_{jb} e reavalia;
5. se, ao final, restarem menos de dois bins não-missing, a variável é descartada.

Esse procedimento gera um conjunto de *breaks* monotônicos por variável, usado em uma segunda chamada de `woebin`, agora forçando a estrutura monotônica.

6 Avaliação da Monotonicidade e Penalização de Missing

Após o binning monotônico, foi gerada uma tabela de avaliação econômica para cada variável, contendo:

- vetor de pontos por bin (após escala para score);
- orientação: “maior valor melhor” ou “maior valor pior”;
- valor de pontos atribuído ao *missing*;
- indicador se o missing foi de fato penalizado em relação ao pior bin observado.

Regra adotada:

Maior score = melhor cliente. Portanto, o missing deve receber o menor número de pontos da variável.

O código ajusta automaticamente os pontos dos bins `missing` para garantir essa propriedade.

7 Modelo Logístico em WoE

Seja o vetor de variáveis WoE para a observação i denotado por $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})$. O modelo logístico é:

$$\log \left(\frac{\Pr(\text{TARGET}_i = 1 \mid \mathbf{w}_i)}{\Pr(\text{TARGET}_i = 0 \mid \mathbf{w}_i)} \right) = \beta_0 + \beta_1 w_{i1} + \dots + \beta_p w_{ip}.$$

Equivalente a:

$$\Pr(\text{TARGET}_i = 1 \mid \mathbf{w}_i) = \frac{1}{1 + \exp(-z_i)}, \quad z_i = \beta_0 + \sum_{j=1}^p \beta_j w_{ij}.$$

A base transformada em WoE foi dividida em:

- 70% para treino;
- 30% para teste;
- estratificação pelo TARGET para preservar o *bad rate*.

A estimação foi feita via *maximum likelihood*. O vetor de coeficientes resultante indicou, por exemplo, efeitos relevantes para NUM_SOCIOS, REST_TIPO_1, REST_TIPO_5, IFIN3, IFIN4, IFIN5 e dummies de UF.

8 Escala de Score

A partir do modelo logístico, constrói-se o scorecard na escala de pontos. A relação clássica entre log-odds e score é:

$$\text{Score} = \text{offset} + \text{factor} \cdot \log\left(\frac{\text{odds}}{1}\right),$$

onde

$$\text{odds} = \frac{\text{Pr}(\text{bom})}{\text{Pr}(\text{mau})}.$$

Com os parâmetros:

- $\text{points}_0 = 600$ (pontos na odds de referência);
- $\text{odds}_0 = 20$ (20 bons para 1 mau);
- $\text{PDO} = 50$ (cada +50 pontos dobra a odds).

Tem-se:

$$\text{factor} = \frac{\text{PDO}}{\ln(2)}, \quad \text{offset} = \text{points}_0 - \text{factor} \cdot \ln(\text{odds}_0).$$

O `scorecardpy` distribui essa estrutura entre um *basepoints* (constante) e pontos por bin de cada variável, gerando uma tabela do tipo:

Variável	Bin	Pontos
IFIN3	$[-\infty, 10)$	43
IFIN3	$[10, 35)$	4
IFIN3	$[35, 65)$	-15
IFIN3	$[65, 85)$	-43
IFIN3	$[85, \infty)$	-55

A pontuação total do cliente é:

$$\text{Score}_i = \text{basepoints} + \sum_j \text{pontos}_{j, \text{bin}(x_{ij})}.$$

9 Métricas de Desempenho: AUC, KS e PSI

9.1 AUC e Curva ROC

A Curva ROC é o gráfico entre TPR (sensibilidade) e FPR ($1 - \text{especificidade}$) para todos os limiares possíveis da probabilidade prevista \hat{p}_i . A AUC (*Area Under the Curve*) é a área sob essa curva, com interpretação:

Probabilidade de um mau ter score (probabilidade prevista) maior do que um bom, escolhidos aleatoriamente.

9.2 KS

O KS (*Kolmogorov-Smirnov*) é definido por:

$$KS = \max_x |F_{\text{maus}}(x) - F_{\text{bons}}(x)|,$$

onde F são as distribuições acumuladas das probabilidades previstas para maus e bons, respectivamente. Em dados de varejo, valores entre 0,3 e 0,6 costumam ser considerados bons.

9.3 PSI

O PSI (*Population Stability Index*) mede a mudança de distribuição entre dois momentos ou amostras (treino vs teste, por exemplo). Seja i um bin de score ou probabilidade:

$$\text{PSI}_i = (\text{act}\%_i - \text{exp}\%_i) \cdot \ln \left(\frac{\text{act}\%_i}{\text{exp}\%_i} \right),$$

$$\text{PSI} = \sum_i \text{PSI}_i.$$

Regra prática:

- $\text{PSI} < 0,10$: estável;
- $0,10 \leq \text{PSI} < 0,25$: atenção;
- $\text{PSI} \geq 0,25$: instável (rever modelo).

No modelo estimado, o PSI treino vs teste ficou em torno de 0,04, indicando boa estabilidade.

10 Resultados Gráficos

10.1 Curva ROC – Base de Teste

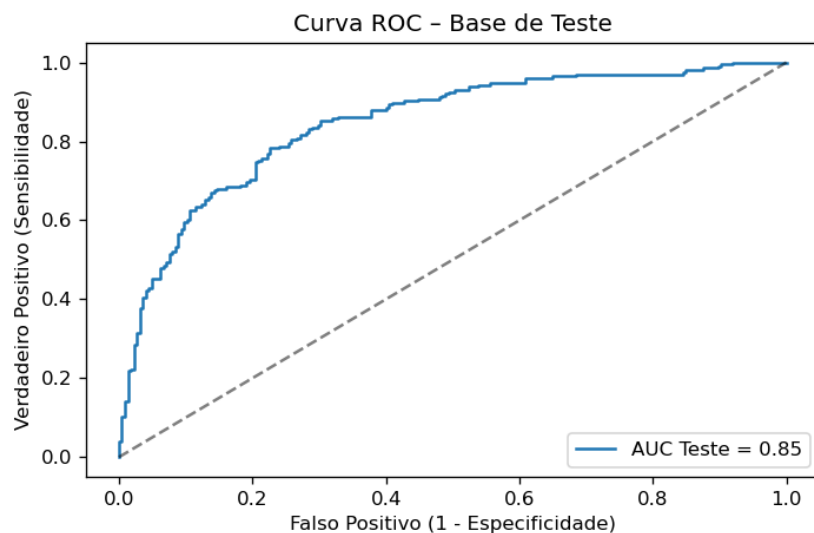


Figura 1: Curva ROC – Base de Teste ($AUC \approx 0,85$).

10.2 Curva KS – Base de Teste

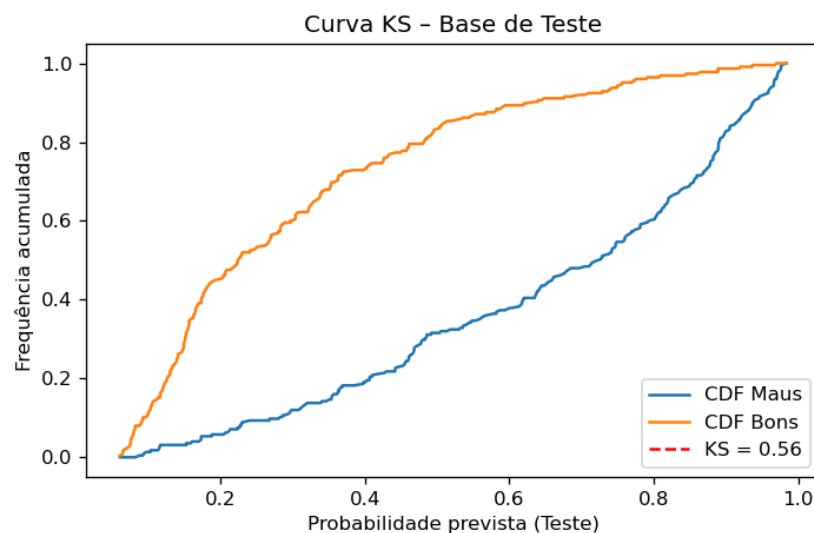


Figura 2: Curva KS – Base de Teste ($KS \approx 0,56$).

10.3 PSI Treino vs Teste

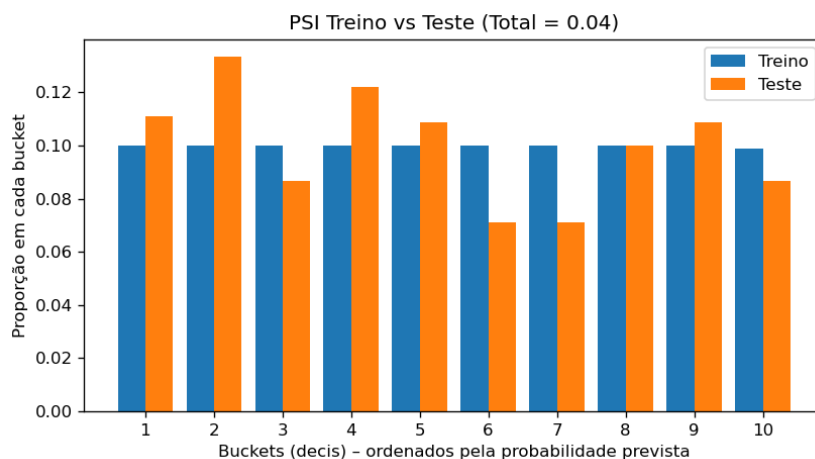


Figura 3: PSI Treino vs Teste – Probabilidades Previstas (PSI \approx 0,04).

10.4 Split Treino vs Teste

Proporção de Registros – Treino vs Teste

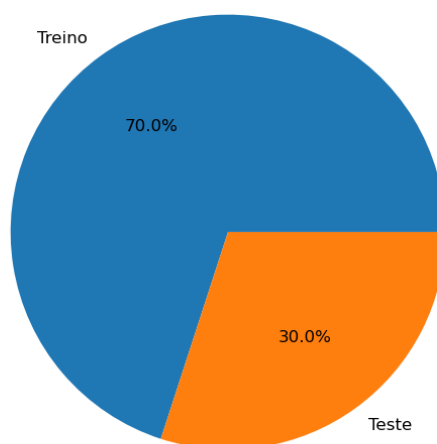


Figura 4: Proporção de Registros – Treino (70%) e Teste (30%).

11 Gráficos de WoE e Monotonicidade por Variável

Os gráficos a seguir mostram o WoE por faixa de cada variável após o binning monotônico. Todos foram gerados pelo script Python a partir da lista `bins` do `scorecardpy`.

Em cada figura, observa-se:

- eixo x : faixas (bins) da variável, com limites truncados em duas casas decimais;
- eixo y : valor de WoE;
- monotonicidade aproximada (crescente ou decrescente) do WoE.

IFIN1, IFIN2, IFIN3

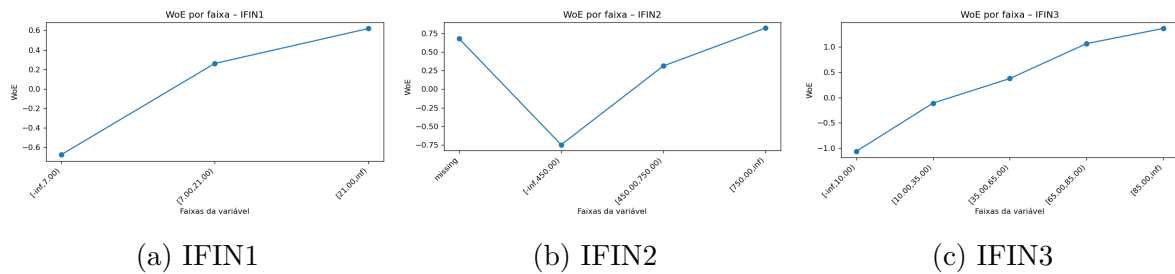


Figura 5: WoE por faixa – IFIN1, IFIN2, IFIN3.

IFIN4, IFIN5, IFIN6

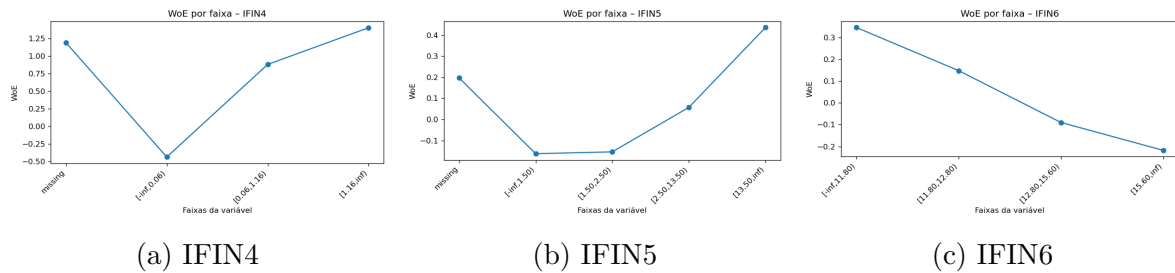


Figura 6: WoE por faixa – IFIN4, IFIN5, IFIN6.

IFIN7, IFIN8, IFIN9

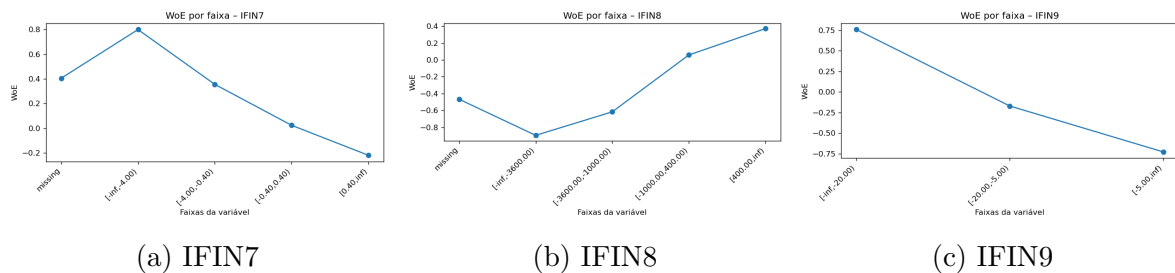


Figura 7: WoE por faixa – IFIN7, IFIN8, IFIN9.

IFIN10, FATLIQ, IDADE

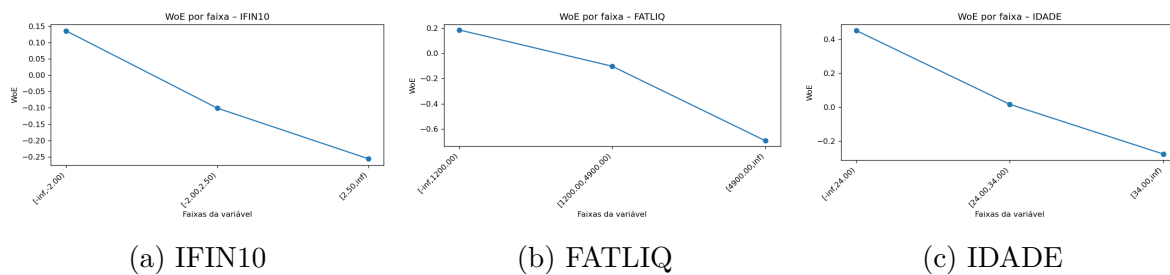


Figura 8: WoE por faixa – IFIN10, FATLIQ, IDADE.

NUM_SOCIOS, RAMO, UF

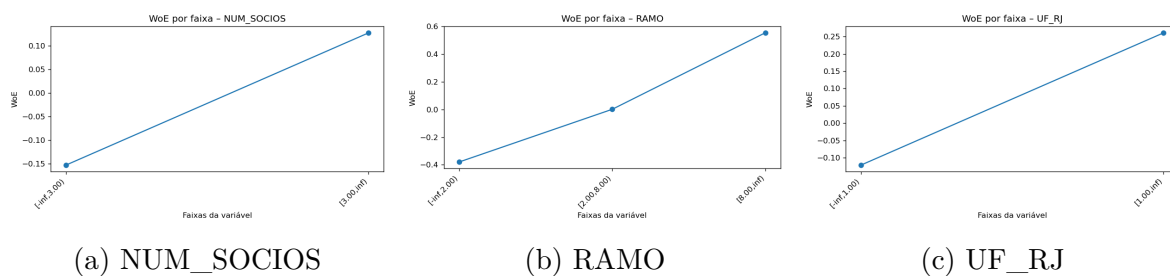


Figura 9: WoE por faixa – NUM_SOCIOS, RAMO, UF_RJ.

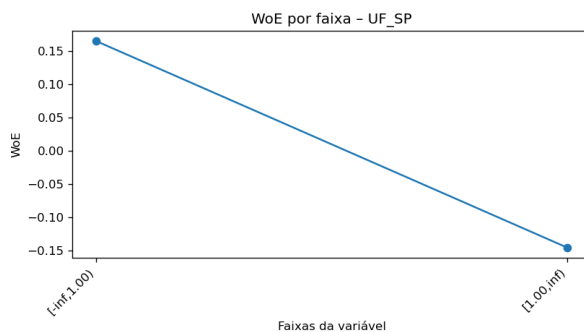


Figura 10: WoE por faixa – UF_SP.

REST_TIPO_1 a REST_TIPO_5

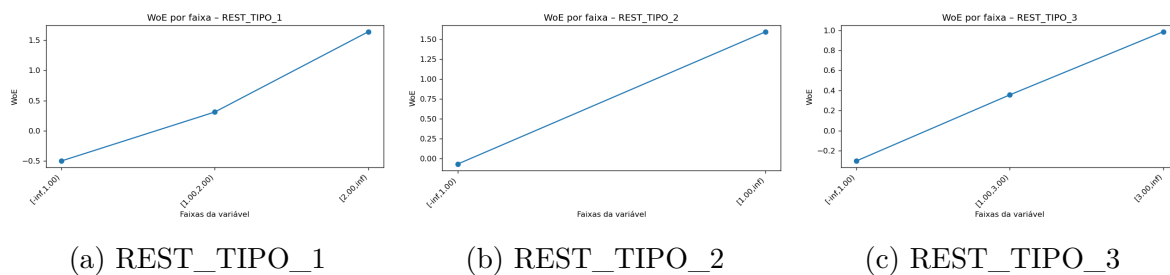


Figura 11: WoE por faixa – REST_TIPO_1, REST_TIPO_2, REST_TIPO_3.

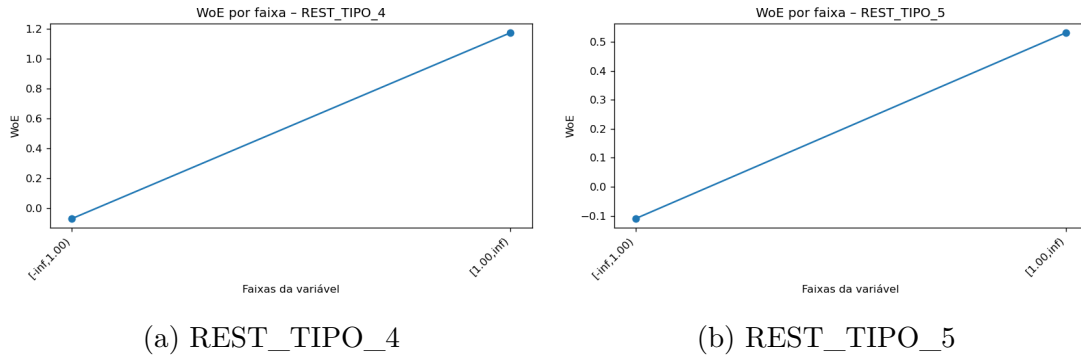


Figura 12: WoE por faixa – REST_TIPO_4, REST_TIPO_5.

Esses gráficos confirmam, visualmente, a monotonicidade imposta e permitem interpretar economicamente cada variável: por exemplo, aumentos de determinadas razões financeiras (IFIN3, IFIN4, IFIN5) estão associados a queda de WoE (pior risco) em certas faixas, enquanto o aumento da idade do negócio tende a melhorar o score.

12 Simulação de Novos Clientes

Por fim, foi gerada uma massa de teste com 10 clientes fictícios, amostrados com reposição da base original, mas tratados como novas solicitações de crédito. Para cada cliente foram calculados:

- valores das variáveis de entrada (IFINs, REST_TIPO, UF, etc.);
- score total segundo o scorecard monotônico;
- uma decisão de crédito exemplificativa: *rejeitar*, *análise manual* ou *aprovar*, com base em tercís do score.

O arquivo `simulacao_10_clientes_PEJOTA.xlsx` contém essa saída, que pode ser usada para testar regras de negócio.

Apêndice A — Metadados da Base PEJOTA

Tabela 1: Dicionário de Variáveis da Base PEJOTA (Sicsú, 2010)

Tipo	Variável	Comentário	Período / Observação
CADASTRAL	COD_CLI	Código interno do cliente	—
CADASTRAL	STATUS	BOM / MAU	—
FINANCEIRA	IND_FIN_1	Aging último balanço	1 mês

Continua na próxima página

Tipo	Variável	Comentário	Período / Observação
FINANCEIRA	IND_FIN_2	Aging último balanço	2 meses
FINANCEIRA	IND_FIN_3	Aging último balanço	3 meses
FINANCEIRA	IND_FIN_4	Aging último balanço	4 meses
FINANCEIRA	IND_FIN_5	Aging último balanço	5 meses
FINANCEIRA	IND_FIN_6	Aging último balanço	6 meses
FINANCEIRA	IND_FIN_7	Aging último balanço	7 meses
FINANCEIRA	IND_FIN_8	Aging último balanço	8 meses
FINANCEIRA	IND_FIN_9	Aging último balanço	9 meses
FINANCEIRA	IND_FIN_10	Aging último balanço	10 meses
FINANCEIRA	FATR_LIQ	Faturamento líquido	últ. ref
RESTRIÇÃO	REST_TIPO_1	Apontamentos	1 mês
RESTRIÇÃO	REST_TIPO_2	Apontamentos	2 meses
RESTRIÇÃO	REST_TIPO_3	Apontamentos	3 meses
RESTRIÇÃO	REST_TIPO_4	Apontamentos	4 meses
RESTRIÇÃO	REST_TIPO_5	Apontamentos	5 meses
CADASTRAL	UF	Localização da matriz do cliente	Estado
CADASTRAL	IDADE	Idade da empresa (meses)	–
CADASTRAL	NUM_SOCIOS	Número de sócios	–
CADASTRAL	RAMO	Ramo de atividade	CNAE

Conclusão

O processo descrito reproduz de forma simplificada a lógica de desenvolvimento de modelos de *credit scoring*, conforme Sicsú (2010) e a literatura em geral:

- uso de WoE/IV para seleção e transformação de variáveis;
- imposição de monotonicidade para garantir coerência de risco;
- regressão logística em variáveis WoE;
- calibragem em pontos (600–20–50);
- avaliação por AUC, KS e PSI;
- scorecard.

Os resultados numéricos ($AUC \approx 0,85$, $KS \approx 0,56$, $PSI \approx 0,04$) são consistentes com um modelo de boa capacidade discriminante e estabilidade aceitável. Existe um fator temporal não analisado nesse exemplo que é importante no desenvolvimento desses modelos em instituições financeiras.

Referências

- Sicsú, Abraham Laredo (2010). *Credit Scoring: Desenvolvimento, Implantação, Acompanhamento*. São Paulo: Blucher.
- Siddiqi, Naeem (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons.
- Thomas, L. C.; Edelman, D. B.; Crook, J. N. (2002). *Credit Scoring and Its Applications*. SIAM.
- Anderson, R. (2007). *The Credit Scoring Toolkit*. Oxford University Press.