XXVI SEMAT
XVI SEMEST

20 a 24 de outubro de 2026

Análise discriminante linear aplicada à classificação de inadimplência em clientes de cartão de crédito

Matheus de Moraes, Lúcio Borges (orientador)
Universidade Federal de Uberlândia, Instituto de Matemática e Estatística

Introdução e objetivo

O risco de inadimplência é relevante na concessão de crédito, pois a classificação incorreta de um cliente pode gerar perdas financeiras. Neste trabalho, foi utilizada a base *Default of Credit Card Clients*, disponibilizada no UCI Machine Learning Repository [1]. A variável resposta indica se o cliente apresentou inadimplência no mês seguinte.

A base contém 30.000 clientes, sendo 23.364 não inadimplentes e 6.636 inadimplentes. Como a classe 1 representa apenas 22,12% das observações, um modelo pode alcançar uma taxa global elevada e ainda identificar poucos inadimplentes.

Classe	Quantidade	Proporção
0: não inadimplente	23.364	77,88%
1: inadimplente	6.636	22,12%

Objetivo: aplicar a análise discriminante linear para classificar clientes quanto à inadimplência e comparar o desempenho do modelo antes e depois do uso do SMOTE.

Metodologia

As variáveis categóricas foram transformadas em fatores. Em seguida, os dados foram embaralhados com semente fixa e separados em 75% para treino e 25% para teste. O conjunto de treino foi usado no ajuste dos modelos, enquanto as 6.000 observações do teste foram reservadas para a avaliação final.

- Ajuste da análise discriminante linear nos dados originais de treino.
- Avaliação pela matriz de confusão e pelas taxas de acerto de cada classe.
- Aplicação do SMOTE somente no conjunto de treino.
- Novo ajuste da análise discriminante e comparação no mesmo conjunto de teste.

Definição

O *SMOTE* [3] cria observações sintéticas da classe minoritária por interpolação. Para uma observação minoritária $\boldsymbol{x}_i$ e um de seus vizinhos mais próximos $\boldsymbol{x}_{viz}$, a nova observação é

$$\boldsymbol{x}_{novo} = \boldsymbol{x}_i + \lambda(\boldsymbol{x}_{viz} - \boldsymbol{x}_i), \qquad 0 \leq \lambda \leq 1.$$

Usou-se `step_smote()`, do pacote `themis` do R, com $k = 5$ vizinhos e `over_ratio = 1`, o que deixou o treino com 50% de cada classe. O teste manteve a distribuição original. O código está disponível em [8].

Análise discriminante linear

A análise discriminante compara as distribuições das classes e constrói uma regra de classificação. Admitindo que o vetor de características da classe k segue uma distribuição normal multivariada, sua densidade é

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}.$$

Na LDA, as classes compartilham a matriz de covariância $\boldsymbol{\Sigma}$. Ao comparar o logaritmo das densidades e incorporar a probabilidade a priori π_k, obtém-se a função discriminante [2].

Definição

A *função discriminante linear* da classe k é

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k).$$

A regra atribui o cliente à classe com maior escore:

$$\widehat{G}(\boldsymbol{x}) = \underset{k\in\{0,1\}}{\operatorname{argmax}}\ \delta_k(\boldsymbol{x}).$$

Para duas classes, a direção de Fisher pode ser escrita como

$$\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \qquad Z = \boldsymbol{a}^T\boldsymbol{x}.$$

Nessas expressões, $\boldsymbol{\mu}_k$ é o vetor médio da classe k, $\boldsymbol{\Sigma}$ é a matriz de covariância comum e π_k é a probabilidade a priori. O escore Z é comparado a um ponto de corte. A QDA foi descartada por instabilidade numérica.

Modelo antes do SMOTE

O modelo original apresentou elevada taxa de acerto global, mas reconheceu apenas 26,05% dos inadimplentes. A distribuição dos escores mostra sobreposição entre as classes, especialmente próximo ao ponto de corte.

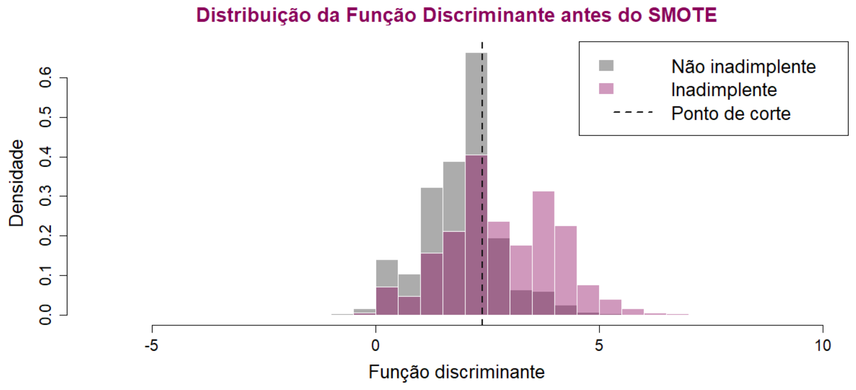


Figura: Distribuição da função discriminante antes do SMOTE.

Matriz de confusão antes do SMOTE

	Classe predita	
Classe real	0	1
0	4560	150
1	954	336

Código e materiais

Modelo depois do SMOTE

Após o balanceamento do conjunto de treino, o modelo passou a reconhecer uma parcela maior dos clientes inadimplentes. A taxa de acerto da classe 1 aumentou para 61,78%, embora a separação visual entre as distribuições ainda apresente sobreposição.

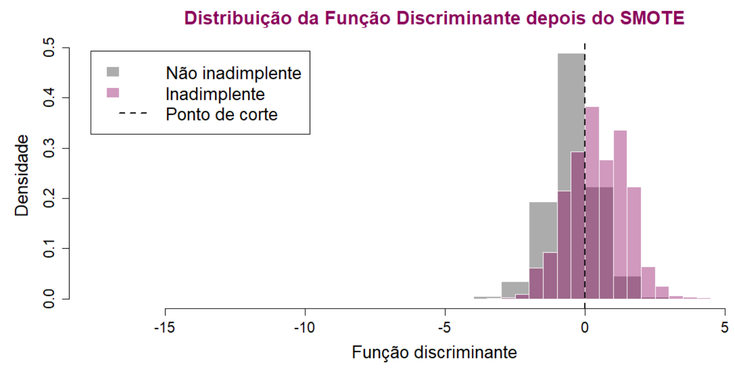


Figura: Distribuição da função discriminante depois do SMOTE.

Matriz de confusão depois do SMOTE

	Classe predita	
Classe real	0	1
0	3380	1330
1	493	797

Medidas de avaliação

Considerando o inadimplente como a classe positiva, denotam-se por VP e VN os verdadeiros positivos e negativos, e por FP e FN os falsos positivos e negativos. As medidas utilizadas foram

$$\text{Acerto global} = \frac{VP + VN}{VP + VN + FP + FN},$$
$$\text{Acerto da classe 0} = \frac{VN}{VN + FP},$$
$$\text{Acerto da classe 1} = \frac{VP}{VP + FN},$$
$$\text{Taxa de falsos positivos} = \frac{FP}{VN + FP},$$
$$\text{Taxa de falsos negativos} = \frac{FN}{VP + FN}.$$

Essas medidas são necessárias porque a taxa global pode ocultar um desempenho fraco na classe minoritária.

Comparação dos modelos

Medida	Antes	Depois
Taxa de acerto global	81,60%	69,62%
Taxa de acerto da classe 0	96,82%	71,76%
Taxa de acerto da classe 1	26,05%	61,78%
Taxa de falsos positivos	3,18%	28,24%
Taxa de falsos negativos	73,95%	38,22%
Taxa média entre as classes	61,43%	66,77%

O modelo sem SMOTE favoreceu a classe majoritária. Com o SMOTE, a taxa global diminuiu, mas a identificação de inadimplentes melhorou de forma expressiva. Também houve redução dos falsos negativos, que correspondem aos inadimplentes classificados como não inadimplentes. Em contrapartida, aumentou a proporção de clientes não inadimplentes classificados como inadimplentes.

26% → 62%

acerto na classe de inadimplentes, antes e depois do SMOTE

O SMOTE não melhorou todas as medidas simultaneamente. Ele produziu uma regra mais equilibrada entre as classes e mais adequada quando a prioridade é reconhecer clientes com risco de inadimplência.

Conclusão

A taxa de acerto global, isoladamente, não descreveu adequadamente o desempenho do modelo em uma base desbalanceada. Antes do SMOTE, o modelo acertou a maioria dos clientes, mas deixou de reconhecer grande parte dos inadimplentes. Depois do SMOTE, houve perda de acerto na classe 0 e ganho relevante na classe 1.

Assim, a escolha do modelo depende do custo associado a cada tipo de classificação incorreta. Em aplicações de crédito nas quais o falso negativo representa maior risco financeiro, o modelo com SMOTE oferece uma alternativa mais coerente com o objetivo da análise.

Como trabalhos futuros, podem ser avaliados modelos específicos para dados desbalanceados, como Floresta Aleatória Balanceada [4], Naive Bayes com Estrutura de Árvore [5], Naive Bayes Complementar [6] e variações da regressão logística [7].

Referências Bibliográficas

1. Yeh, I. C. *Default of Credit Card Clients*. UCI Machine Learning Repository, 2009.
2. Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*. 6. ed. Pearson, 2007.
3. Chawla, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.
4. Chen, C.; Liaw, A.; Breiman, L. *Using Random Forest to Learn Imbalanced Data*. University of California, Berkeley, 2004.
5. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Machine Learning*, v. 29, p. 131–163, 1997.
6. Rennie, J. D. M. et al. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *ICML*, p. 616–623, 2003.
7. King, G.; Zeng, L. Logistic Regression in Rare Events Data. *Political Analysis*, v. 9, n. 2, p. 137–163, 2001.
8. Código e materiais: `https://github.com/matsmoraes/AD_SEMAST2026`.

matheus.neves1@ufu.br, lucio.araujo@ufu.br