FRAUDES NO CONSUMO DE ÁGUA: IDENTIFICAÇÃO E MODELAGEM

João Pedro Felicio Prudencio, Elisa Henning, Andreza Kalbusch

INTRODUÇÃO

As perdas em sistemas de abastecimento representam um desafio econômico e ambiental para as companhias de saneamento. O volume global de água não faturada ultrapassa US$ 39 bilhões por ano (Liemberger; Wyatt, 2019), sendo o consumo não autorizado um de seus principais componentes. Sua detecção por métodos tradicionais é dificultada pelo comportamento irregular e não linear do consumo clandestino, o que motiva o uso de métodos computacionais mais robustos (Vargas; Azevedo, 2021).

No contexto da cidade de Joinville, Stramari, Kalbusch e Henning (2022) aplicaram o algoritmo Random Forest, combinado ao SMOTE (uma técnica de balanceamento), a dados com 12 casos de consumo não autorizado de água dentre 1.400 unidades consumidoras de um Distrito de Medição e Controle, obtendo resultados promissores. Naquele estudo, contudo, apenas uma técnica de balanceamento foi empregada, e os próprios autores indicam a avaliação de outros métodos como possibilidade de continuidade.

Este trabalho tem como objetivo avaliar comparativamente seis condições de balanceamento de dados na detecção de consumo não autorizado: SMOTE, SMOTE-N, sobreamostragem aleatória, ADASYN, SMOTETomek e sem balanceamento. Como objetivos específicos, busca-se quantificar o efeito de cada técnica de balanceamento sobre o desempenho de diferentes algoritmos de classificação: Random Forest, Regressão Logística, K-Nearest Neighbors, Naive Bayes, Support Vector Machine e Rede Neural. Além disso, busca-se identificar quais técnicas de balanceamento e quais algoritmos de classificação são mais adequados ao tipo de dado utilizado.

DESENVOLVIMENTO

A base de dados, a mesma utilizada por Stramari, Kalbusch e Henning (2022), e contém informações do histórico de consumo mensal de 1.400 unidades consumidoras ao longo de 76 meses, entre fevereiro de 2016 e maio de 2022. Nesse período, doze casos de consumo não autorizado foram confirmados em campo, correspondendo a 0,86% da amostra.

Foram utilizadas as mesmas sete variáveis preditoras definidas por Stramari, Kalbusch e Henning (2022): tipo de hidrômetro, conformidade da ligação, redução progressiva do consumo identificada pelo coeficiente de correlação de Pearson, dois limiares de queda súbita medidos pelo Z-score, queda súbita pelo Z-score modificado e ocorrência de consumo nulo por períodos prolongados. Todas as variáveis são binárias.

Adotou-se o mesmo protocolo de avaliação do estudo de referência: o balanceamento foi aplicado sobre o conjunto, elevando a classe minoritária a 1.212 registros e reduzindo a classe majoritária ao mesmo total (Chawla et al., 2002), seguido da divisão em 70% para treinamento e 30% para teste. Esse procedimento foi mantido para permitir a comparação direta com os resultados publicados por Stramari, Kalbusch e Henning (2022).

Foram avaliadas seis condições de balanceamento: SMOTE, SMOTE-N e sobreamostragem aleatória (Chawla et al., 2002), ADASYN (He et al., 2008), SMOTETomek (Batista; Prati; Monard, 2004) e sem balanceamento. Cada condição foi testada com seis algoritmos de classificação: Random Forest (Breiman, 2001), Rede Neural (Mitchell, 1997), Regressão Logística, K-Nearest Neighbors, Naive Bayes e Support Vector Machine (Pedregosa et al., 2011), totalizando 36 configurações.

Os hiperparâmetros de cada modelo foram ajustados por otimização bayesiana (Frazier, 2018), conduzida com a ferramenta Optuna (Akiba et al., 2019), e as implementações utilizaram a biblioteca

Scikit-learn (Pedregosa et al., 2011). O desempenho foi medido por sensibilidade, especificidade, precisão e F1-Score (Pedregosa et al., 2011), exatidão balanceada (Diego et al., 2022) e área sob a curva ROC (Hastie; Tibshirani; Friedman, 2009).

RESULTADOS

A Tabela 1 apresenta a exatidão balanceada obtida em cada uma das 36 configurações. Houve uma separação clara entre dois grupos de técnicas. As abordagens baseadas em interpolação (SMOTE, ADASYN e SMOTETomek) apresentaram desempenho equivalente entre si, com uma exatidão balanceada em torno de 0,95, e superior ao das técnicas que preservam os valores categóricos originais, sobreamostragem aleatória e SMOTE-N, situadas entre 0,81 e 0,85. Na ausência de balanceamento, todos os modelos classificaram a totalidade das amostras como “Não fraudes”, resultando em exatidão balanceada de 0,500.

Essa diferença está ligada à forma como as amostras sintéticas são geradas. A interpolação linear empregada pelo SMOTE e por suas variantes produziu valores fracionários em atributos que admitem apenas os valores zero e um: foram contabilizados 949 registros com valores fracionários no SMOTE e 977 no ADASYN, contra nenhum na sobreamostragem aleatória e no SMOTE-N. Esses valores não existem nos dados reais e dão aos classificadores uma granularidade que os atributos originais não têm. Chawla et al. (2002) propuseram o SMOTE-N justamente para atributos nominais, substituindo a interpolação por votação majoritária entre vizinhos.

Um segundo resultado aponta na mesma direção. Nas duas técnicas que preservam os valores originais, três algoritmos de naturezas distintas convergiram para resultados idênticos: 0,853 na sobreamostragem aleatória e 0,806 no SMOTE-N. Verificou-se que as sete variáveis binárias produzem apenas 52 combinações distintas entre as 1.400 unidades, de modo que existe uma única forma ótima de separar as classes, encontrada por qualquer classificador com flexibilidade suficiente. A diferenciação entre algoritmos observada nas técnicas baseadas em interpolação decorre, portanto, da granularidade artificial introduzida, e não de capacidade preditiva distinta.

A Figura 1 compara a melhor configuração obtida (Random Forest otimizada com ADASYN) aos valores relatados por Stramari, Kalbusch e Henning (2022). O desempenho foi superior em todas as métricas avaliadas, o que se deve tanto à técnica de balanceamento quanto à otimização de hiperparâmetros, não realizada no estudo original. Destacam-se, então, a especificidade, que passou de 0,590 para 0,920, e a área sob a curva ROC, de 0,736 para 0,984. A matriz de confusão do conjunto de teste registrou 358 verdadeiros positivos, 335 verdadeiros negativos, 29 falsos positivos e 6 falsos negativos.

Observa-se, ainda, que, entre os algoritmos de melhor desempenho (Random Forest, Support Vector Machine, K-Nearest Neighbors e Rede Neural), a variação entre grupos de técnicas de balanceamento, aproximadamente 0,12, supera a variação entre os classificadores submetidos a uma mesma técnica, inferior a 0,02. Dentre os resultados mais promissores, portanto, a escolha do procedimento de reamostragem exerce influência maior sobre o resultado do que a escolha do classificador. Regressão Logística e Naive Bayes apresentaram desempenho inferior em todas as condições, o que sugere limitações estruturais desses classificadores nesta aplicação.

CONSIDERAÇÕES FINAIS

A avaliação das seis técnicas indicou que a escolha do método de balanceamento afeta o desempenho tanto quanto a escolha do algoritmo de classificação. Técnicas baseadas em interpolação apresentaram os melhores resultados quando aplicadas a atributos binários, embora parte dessa vantagem venha de valores que não existem nos dados reais.

Os resultados sugerem que a adequação de uma técnica de reamostragem depende da natureza dos atributos e que essa adequação deve ser verificada quando os dados são categóricos. A convergência

de algoritmos distintos para resultados idênticos, observada nas técnicas que preservam os valores originais, sugere ainda que o conjunto de variáveis utilizado pode estar limitando o desempenho alcançável. A ampliação desse conjunto e a avaliação sob protocolos alternativos são os próximos passos deste trabalho.

Palavras-chave: perdas de água; aprendizado de máquina; balanceamento de dados; detecção de fraudes; SMOTE.

ANEXOS

Tabela 1. *Exatidão balanceada por técnica de balanceamento e algoritmo de classificação.*

Técnica	Random Forest	KNN	SVM	Rede Neural	Reg. Logística	Naive Bayes
ADASYN	0,952	0,935	0,945	0,941	0,702	0,692
SMOTE	0,951	0,945	0,949	0,934	0,699	0,703
SMOTETomek	0,951	0,945	0,949	0,934	0,699	0,703
Sobreamostragem aleatória	0,853	0,853	0,853	0,835	0,665	0,610
SMOTE-N	0,806	0,805	0,806	0,806	0,673	0,712
Sem balanceamento	0,500	0,500	0,500	0,500	0,500	0,500

Fonte: o autor.

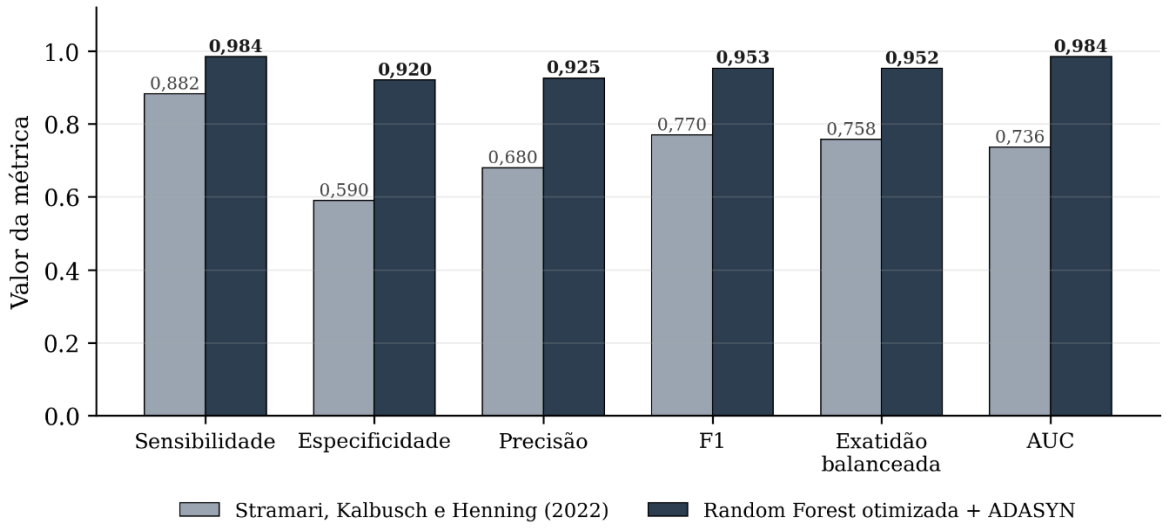


Figura 1. *Comparação entre a melhor configuração obtida e os resultados do estudo de referência.*

Fonte: o autor.

REFERÊNCIAS BIBLIOGRÁFICAS

AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. Optuna: a next-generation hyperparameter optimization framework. **In**: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING, 25., 2019, Anchorage. Anais [...]. New York: ACM, 2019. p. 2623-2631.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20-29, 2004.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002.

DIEGO, I. M. de et al. General performance score for classification problems. **Applied Intelligence**, v. 52, n. 10, p. 12049-12063, 2022.

FRAZIER, P. I. A tutorial on Bayesian optimization. **arXiv**,1807.02811, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. **In**: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2008, Hong Kong. Anais [...]. Hong Kong: IEEE, 2008. p. 1322-1328.

LIEMBERGER, R.; WYATT, A. Quantifying the global non-revenue water problem. **Water Supply**, v. 19, n. 3, p. 831-837, 2019.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

STRAMARI, M. R.; KALBUSCH, A.; HENNING, E. Random forest for the detection of unauthorized consumption in water supply systems: a case study in southern Brazil. **Urban Water Journal**, v. 20, n. 3, p. 394-404, 2022.

VARGAS, J. C. B.; AZEVEDO, B. B. Complexidade, leis de escala urbana e perdas na distribuição de água potável: análise da rede de cidades do sul do Brasil. **Ambiente Construído**, v. 21, n. 3, p. 65-78, 2021.