Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Wesley Oliveira Furriel

Avaliação do impacto de conjuntos de dados desbalanceadas em modelos de classificação para risco de crédito

Curitiba
2023

Wesley Oliveira Furriel

Avaliação do impacto de conjuntos de dados desbalanceadas em modelos de classificação para risco de crédito

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Wagner Hugo Bonat

Curitiba
2023

Avaliação do impacto de conjuntos de dados desbalanceadas em modelos de classificação para risco de crédito

Evaluation of the Impact of Imbalanced Datasets on Credit Risk Classification Models

Wesley Oliveira Furriel[1], Wagner Hugo Bonat[2]

[1]Aluno do programa de Especialização em Data Science & Big Data, wesleyofurriel@gmail.com
[2]Professor do Departamento de Estatística - DEST/UFPR, wbonat@gmail.com

Este trabalho teve como objetivo avaliar o desempenho de modelos de regressão e classificação binária em conjuntos de dados com diferentes níveis de desbalanceamento via *oversampling* e *undersampling* aleatório em um problema de risco de crédito. Os modelos de regressão logística, *random forest*, *catboost* e *lgbm* foram treinados e validados inicialmente pelas amostras obtidas pelo método *Stratified K-Fold*. Em seguida, foram avaliados em uma amostra com a distribuição original da variável resposta, permitindo uma comparação entre os resultados obtidos. Desse modo, observou-se, que não ocorreram ganhos relevantes ao balancear os conjuntos de dados, ainda assim, foram constatadas quedas nos valores de *F1-Score* e *LogLoss* para os balanceamentos de 50%. Além disso, para os experimentos com maior desbalanceamento, foram identificados maiores níveis de variabilidade entre as amostras de treinamento e assimetrias mais acentuadas na distribuição de probabilidade predita.
Palavras-chave: Balanceamento, desbalanceamento, regressão, classificação, binária, risco de crédito

This study aimed to evaluate the performance of binary classification models on datasets with different levels of imbalance using oversampling and undersampling techniques in a credit risk problem. The models were initially trained and validated using samples from the Stratified K-Fold method. They were then evaluated on a sample with the original distribution of the response variable, allowing for a comparison of the results obtained. It was observed that there were no significant gains from balancing the response variable, only decreases in the values of F1-Score and LogLoss for the 50% balancing. Furthermore, for experiments with higher imbalance, higher levels of variability were identified among the training samples and more pronounced asymmetries in the predicted probability distribution.
Keywords: Balanced,imbalanced,regression, classification, binary, credit risk

1. Introdução

Modelos de classificação com amostras desbalanceados, nos quais as classes da variável resposta apresentam uma distribuição desigual são um problema bastante frequente na modelagem estatística e no aprendizado de máquina. Em áreas como, risco de crédito, detecção de fraudes, diagnósticos médicos entre outras, é comum encontrar conjuntos de dados em que uma classe é consideravelmente menor que as demais. Desse modo, existe a discussão se tal desequilíbrio pode causar algum impacto no desempenho do ajuste dos modelos, gerando resultados equivocados.

A discussão em questão se mostra relevante, uma vez que, em problemas como risco de crédito, o evento de interesse identificado como mau pagador é em geral a classe minoritária e que desejamos identificar com maior precisão. Se a classe de interesse for pouco representada, o modelo pode ter dificuldade em identificar corretamente o mau pagador, gerando prejuízos às instituições financeiras.

O desbalanceamento também pode afetar as medidas de avaliação tradicionais. A acurácia, por exemplo, pode ser enganosa em problemas desbalanceados, uma vez que um modelo pode obter uma alta taxa de acertos prevendo predominantemente a classe majoritária, gerando uma falsa percepção de sucesso.

Algumas das abordagens mais usuais para resolver esse problema são as técnicas de *Oversampling* e *Undersampling*, que visam melhorar a distribuição da variável resposta, permitindo uma aprendizagem mais equilibrada entre as categorias. Todavia, tais técnicas alteram a estrutura dos dados, gerando uma distribui-

ção que não reflete o comportamento real da população modelada.

Tendo em vista tais problemas e o teor prático deste trabalho, iremos avaliar o desempenho de modelos de regressão e classificação binária, como Regressão Logística, *LightGBM*, *Random Forest* e *CatBoost*, em um conjunto de dados com diferentes níveis de desbalanceamento.

Cada um dos modelos selecionados será treinado levando em consideração o ajuste dos respectivos hiperparâmetros e as validações serão realizadas primeiramente, durante a etapa de modelagem nas amostras geradas pelo método *Stratified K-Fold*, a fim de garantir a qualidade e generalização do ajuste. Posteriormente, os modelos serão avaliados em uma amostra com o balanceamento original, com o objetivo de comparar os resultados entre modelos.

Nosso trabalho foi dividido em três seções, na seção 2 fizemos uma breve discussão acerca dos métodos de *oversampling* e *undersampling*, bem como, dos modelos selecionados para o estudo. Na seção 3 apresentamos o delineamento dos experimentos criados para o estudo e a descrição do processo de modelagem e avalição dos modelos, por fim na seção 4 foram discutidos os resultados obtidos.

2. Discussão

Ao considerarmos um problema supervisionado, com resposta binária, é comum avaliar e comparar diferentes algoritmos de regressão e classificação, buscando o melhor resultado de ajuste. Entre os principais modelos utilizados na área de crédito, consideramos para este trabalho a Regressão Logística, o *Random Forest*, o *LightGBM* e o *CatBoost*, com o intuito de avaliar técnicas com distintos níveis de complexidade e graus de interpretabilidade.

A regressão logística é um modelo linear generalizado mais simples, de caráter paramétrico, assumindo uma relação linear entre as variáveis independentes e a dependente, sendo adequado para problemas em que se busca entender o impacto direto das variáveis preditoras no resultado final [2]. Por outro lado, tanto o *Random Forest* quanto o *LightGBM* e o *CatBoost* são algoritmos da família *ensemble*, que permitem capturar relações não lineares e interações complexas entre as variáveis preditoras e seu impacto na resposta, combinando modelos básicos para obter um resultado mais robusto e preciso.

O *Random Forest* cria um conjunto de árvores de decisão mais simples, nas quais cada árvore é construída de forma independente, usando uma amostra dos dados de treinamento. Logo, o resultado final é obtido por meio da média ou de uma votação das previsões das árvores individuais [1]. Já o *LightGBM* e o *CatBoost* são algoritmos de *boosting*, que constroem um modelo de forma sequencial, corrigindo os erros dos modelos anteriores, para gerar resultados mais precisos. O *LightGBM* é conhecido por sua eficiência e velocidade [15], enquanto o *CatBoost* se destaca por sua capacidade de lidar diretamente com variáveis categóricas e também, pela eficácia com dados desbalanceados[12].

No que diz respeito às técnicas de balanceamento, o *oversampling* visa aumentar o número de observações da classe minoritária, criando amostras artificiais desta classe. De outra forma, o *undersampling* é uma técnica que consiste em reduzir a quantidade de observações da classe majoritária, equilibrando a representação entre as classes. [6] [7] Existem ainda variantes dessas técnicas como *SMOTE (Synthetic Minority Oversampling Technique) e ADASYN (Adaptive Synthetic Sampling)*. No entanto, aqui nos ocupamos apenas em realizar o ajuste dos dados pela abordagem aleatória, conhecida como *naive random sampling*.

3. Materiais e métodos

A base de dados utilizada para o estudo foi obtida no site *Kaggle*, oriunda do desafio *Give Me Some Credit* [10], tendo $n = 72.247$, 10 variáveis preditoras que descrevem principalmente comportamentos relacionados a transações e contratações de produtos financeiros e informações cadastrais dos clientes de uma instituição financeira. Como o objetivo foi identificar os indivíduos maus pagadores, ou seja, que atrasaram ou não pagaram suas dívidas, a variável resposta empregada indica se o cliente pagou ou não o empréstimo adquirido dentro de um determinado período de tempo, de modo que, clientes que cumpriram com seus pegamentos são chamados de bons e clientes que entraram em atraso são chamados de maus ou *default*, gerando uma resposta binária.

A fim de investigar os efeitos do balanceamento e desbalanceamento dos dados, foram geradas amostras a partir do conjunto original de dados, que possuía uma proporção do evento de interesse, os clientes maus de 7%. Portanto, foram criadas duas amostras utilizando a técnica *undersampling* em que foram reduzidos o volume de bons mantendo o de maus e cinco

amostras a partir do *oversampling*, destas três foram inflados o volume de maus e duas os casos foram reduzidos, mantendo fixo o volume de bons. Partindo disso, o intuito foi avaliar se existiam diferenças relevantes nos resultados entre tais experimentos.

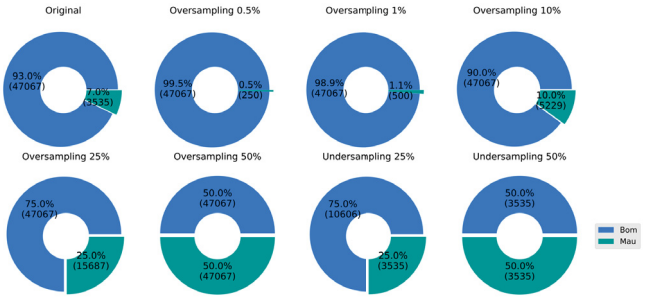

Figura 1: Definição dos experimentos.

Pela Figura 1, observamos a distribuição dos experimentos criados, que variam desde amostras bastante desbalanceadas, com 0.5%, 1% e 5% de ocorrência de maus, até amostras mais balanceadas, com 10%, 25% e 50% utilizando oversampling, e 25% e 50% utilizando undersampling. Embora seja possível criar cenários de maior complexidade, este trabalho limitou-se aos mencionados devido a restrições de tempo e recursos computacionais. Ainda assim, os experimentos propostos abrangem uma gama parcimoniosa de proporções desbalanceadas e balanceadas, permitindo uma análise abrangente desses efeitos nos resultados dos modelos.

Para realizar uma comparação justa após a modelagem, dividimos a população em duas partes, sendo 70% separada para o treino do modelo, no qual foram realizadas as validações de qualidade de ajuste pelo AUC a partir do *Stratified K-Fold* [11], fixando $K = 5$. Com exceção da regressão logística, para cada um dos modelos e em cada cenário, foi realizada a otimização bayesiana para encontrar os melhores conjuntos de hiperparâmetros pelo *Optuna* [14], com o intuito de garantir o melhor e mais parcimonioso modelo. Por fim, aplicamos os modelos ajustados na amostra de validação (30%), utilizando as métricas de Acurácia, *Recall*, *Precision* e *F1-Score* considerando o ponto de corte ótimo, AUC, *Logloss* e o coeficiente de variação dos AUC_k durante o processo de modelagem. Um resumo da lógica empregada pode ser visto na Figura 2.

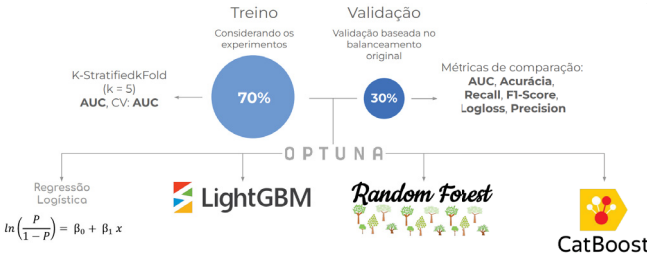

Figura 2: Fluxograma de modelagem dos experimentos.

Por fim, para atingir os objetivos desejados utilizamos a linguagem *Python* como ferramenta de análise e os pacotes *pandas*, *numpy*, *scipy*, *sklearn*[11], *matplotlib* [13], *lightgbm* [15], *catboost* [12] *imblearn* e *optuna*[14], que contam com a implementação dos algoritmos e avaliações que foram necessárias para o estudo.

4. Resultados

No que tange aos resultados obtidos, pela Figura 3 avaliamos primeiramente a distribuição da probabilidade predita $\hat{P}$ transformada pela função *logit* $ln(P/1-P)$ para suavizar os picos de concentração gerados pelos modelos desbalanceados e torná-los visualmente comparáveis. Dessa forma, constatamos que o *Random Forest* retorna curvas mais assimétricas e com comportamentos bimodais, enquanto que os modelos logístico e de *Boosting* têm distribuições mais suaves e menos assimétricas.

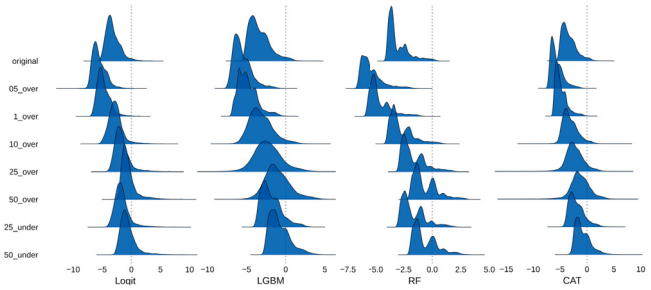

Figura 3: Distribuição da probabilidade predita via logit na amostra de validação.

Além das distribuições avaliamos também as curvas de ordenação, isto é, a proporção de maus observados dentro de faixas construídas por decil aplicadas da probabilidade predita. Desse modo, notamos que todos os experimentos e modelos apresentam ordenação satisfatória, sem quaisquer inversões no decorrer das dez faixas.

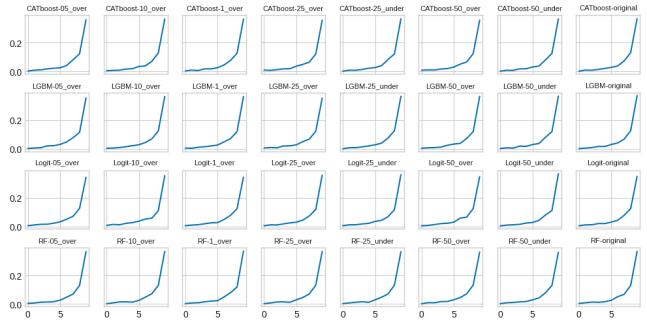

Figura 4: Ordenação da proporção de maus observados, segundo decil da probabilidade predita.

No Heatmap exposto na Figura 5 padronizamos a escala de cor segundo o método de normalização Min-Max, dado por $X_{\text{norm}} = (X - X_{\text{min}})/(X_{\text{max}} - X_{\text{min}})$, de modo que tons de azul mais escuros próximos a 1 indicam medidas de diagnóstico melhores, enquanto ctons mais claros próximos a 0 representam medidas mais fracas. Essa codificação permite uma comparação eficiente tanto entre diferentes experimentos quanto entre algoritmos, facilitando a análise e interpretação dos resultados obtidos. Dessa forma, observamos que ocorreram distinções entre os modelos e os experimentos. Porém, ainda assim, para todos os casos foram obtidos resultados de ajustes satisfatórios, ou seja, em qualquer uma das medidas empregadas não foram identificados valores que indiquem problemas de ajuste.

Dito isto, quando consideramos a Acurácia e o AUC, em geral, o aumento ou diminuição da proporção do evento de interesse não resultou em melhorias ou quedas relevantes na performance dos modelos. Ambas as medidas oscilaram no máximo dois pontos para baixo ou para cima entre os experimentos para todos os modelos em questão, quando comparados aos resultados da amostra original. Já na comparação entre modelos, constatamos que os algoritmos de *ensemble* obtiveram performance um pouco acima das do modelo de Regressão Logística.

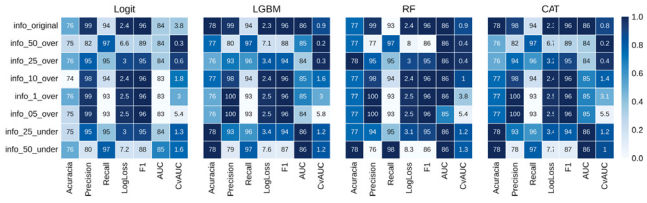

Figura 5: Medidas de validação

As medidas que mostraram maior discrepância entre si, foram as de *Precision*, *Recall*, *Logloss*, *F1-Score* e o Coeficiente de Variação da AUC durante o treinamento. Desse modo, pela Figura 6 conseguimos olhar tais aspectos com maiores detalhes. Primeiramente ao avaliar o CV da AUC, é possível verificar maior variabilidade nos experimentos mais desbalanceados de 1% e 0.5%, possivelmente devido a baixa proporção de maus nesses casos o *Stratified K-Fold*, encontrou dificuldaden em gerar populações igualmente distribuídas para cada uma de suas amostras, gerando maior heterogeneidade entre os resultados. Em seguida, ao avaliar o *Logloss* e o *F1-Score* notamos que as amostras de *Oversampling* e *Undersampling* com balanceamento de 50% apresentaram piores resultados quanto a tais medidas. Na Figura 6 conseguimos observar as diferenças com maior detalhamento.

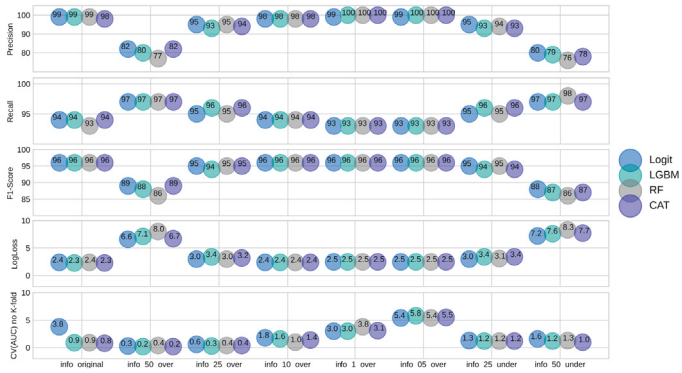

Figura 6: Medidas de validação dissidentes

Quando avaliamos os resultados do *Precision* e o *Recall*, constatamos que ao aumentar nosso evento de interesse ganhamos um pouco em termos de *Recall*, mas aumentamos consideravelmente as chances de cometer o erro Tipo I de falsos positivos, como indica a queda do *Precision*. O F1-Score leva em consideração as duas medidas anteriores em sua construção e se mostra útil para uma avaliação mais direta [9], mesmo que não evidencie a direção do problema.

5. Conclusão

Portanto, ao analisar os resultados obtidos nos diferentes tipos de experimentos e modelos, constatamos que o desbalanceamento pode afetar a estabilidade da qualidade do ajuste no processo de modelagem, assim como a forma da distribuição da probabilidade predita. No entanto, apesar das variações observadas, poucas diferenças significativas foram encontradas para justificar a necessidade de alterar a distribuição original dos dados.

As abordagens de *oversampling* e *undersampling* aqui adotadas podem ser úteis em alguns casos para

melhorar o desempenho dos modelos, principalmente em dados extremamente desbalanceados, porém, é importante considerar o impacto dessas técnicas nos resultados finais. Constatamos que o balanceamento de 50% resultou em um desempenho inferior aos demais em métricas de avaliação como *F1-Score* e *LogLoss*, fato que não verificamos na AUC e acurácia. Ou seja, a utlização de diagnósticos plurais podem nos ajudar não só avaliar a qualidade de um ajuste, mas também compreender as nuances do impacto do balanceamento nos modelos ajustados e validar se de fato existe a necessidade desse tipo de alteração na amostra original.

Por fim é importante considerar que outros algoritmos de aprendizado de máquina, ou mesmo conjuntos de dados distintos, além do analisado neste estudo, podem apresentar resultados divergentes e, portanto, devem ser considerados em trabalhos futuros.

6. Agradecimentos

Gostaria de expressar meus agradecimentos a todos os docentes do curso de especialização em Data Science e Big Data que contribuíram para a realização do meu Trabalho, em especial ao meu orientador Wagner Bonat, aos avaliadores Anderson Ara e Paulo Justiniano Ribeiro Jr. Agradecer também, meus amigos e família que me ajudaram e incentivaram a continuidade dos meus estudos, em especial minha noiva Jaqueline Yumi.

Referências

[1] Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

[2] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[3] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (pp. 6638-6648).

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Yu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[6] Ricardo Barandela, José Salvador Sánchez, V Garca, and Edgar Rangel. Strategies for learning in class imbalance problems. Pattern Recognition, 36(3):849–851, 2003.

[7] Chao Chen, Andy Liaw, Leo Breiman, and others. Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12):24, 2004.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., *SMOTE: synthetic minority oversampling technique. Journal of artificial intelligence research, 16, 321-357*

[9] Hand, David J., Peter Christen, and Nishadi Kirielle. F*: an interpretable transformation of the F-measure. Machine Learning 110.3 (2021): 451-456.

[10] Give Me Some Credit. Kaggle, 2011. Disponível em: <https://www.kaggle.com/c/GiveMeSomeCredit>. Acesso em: 11/06 /2023.

[11] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python."the Journal of machine Learning research 12 (2011): 2825-2830, url *https://scikit-learn.org/stable/index.html*

[12] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features."Advances in neural information processing systems 31 (2018).

[13] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[14] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta,and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.

[15] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.