IV Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Crítica metodológica a estudos de previsão do tempo de estadia de navios em portos brasileiros

Eduardo Caruso Barbosa Pacheco[1], Alexandre dos Santos Gualberto[2], Kaique Antunes dos Santos [3], Marcos Antonio Alves Bezerra Júnior [4], Reynaldo Pereira Martins [5]
Milton Miranda Neto [6] Francisco Louzada [7]
ICMC-USP

1 Estudo de caso do tempo de estadia de navios em portos brasileiros

A Agência Nacional de Transportes Aquaviários(ANTAQ) mantém um dataset contendo dados a respeito da estadia de navios em portos que permitem explorar gargalos operacionais e oportunidades de melhoria de eficiência logística para o transporte de cargas.

Um primeiro estudo, intitulado *"A decision tree model for the prediction of the stay time of ships in Brazilian ports"* [1] analisou os dados da ANTAQ do ano de 2018 e comparou diferentes modelos preditivos para o tempo de estadia dos navios em portos. Os autores selecionaram 35 variáveis independentes e a variável alvo é uma classificação arbitrária sobre o tempo de estadia: Baixo, Médio, Alto e Muito Alto. Esse trabalho disponibilizou seu dataset já tratado e descreveu brevemente o processo de tratamento e limpeza dos dados. Apresentou uma tabela comparando diversos modelos, com superioridade de um modelo de RandomForest para a previsão da classe do tempo de estadia, com acurácia de 0.739 medida via 10fold-cross-validation. Os autores explicaram o peso das features com feature importances das Florestas Aleatórias.

Outro artigo, *"Predictive Analysis for Optimizing Port Operations"* [2], se baseou no trabalho anterior e trouxe novas contribuições. Partiu da mesma base de dados e selecionou 24 das 35 features. Quanto à variável alvo, explorou algumas variações, envolvendo desde problemas de regressão (para o tempo total de estadia e o tempo de ateaso) quanto de classificação (com a mesma classificação em 4 categorias e uma variação com 2 categorias). Uma contribuição importante foi separar o tempo total de estadia do tempo de atraso. Fizeram também modificações no design das

[1] eduardo.caruso.pacheco@usp.br
[2] alexandre.gualberto@usp.br
[3] kaiqueas@usp.br
[4] marcosbezerra85@usp.br
[5] reynaldo.martins@usp.br
[6] miltonmirandaneto@usp.br
[7] louzada@icmc.usp.br

bases de dados pois dividiram a base em training set e test set e depois disso compararam diversos outros modelos (como Extreme Gradient Boosting, Extra Trees e ResNet). Obtiveram um modelo de classificação com acurácia de 0.842 no 10fold-cross-validation para tempo total de estadia e 0.951 para tempo de atraso. Esse segundo trabalho explicou o impacto das features por meio de SHAP.

O presente artigo foi desenvolvido na disciplina MAI5003- Probabilidade e Estatística, sob orientação do Prof. Dr. Francisco Louzada, no programa de mestrado profissional em Matemática, Estatística e Computação Aplicadas à Industria(MECAI) do ICMC-USP.

2 Metodologia

Nosso trabalho analisa os dois artigos anteriores, contrastando as descrições metodológicas com os dados disponibilizados já tratados usados por ambos os autores e os dados originais fornecidos pela ANTAQ. Pelo contraste foi possível entender - e questionar - algumas decisões de modelagem que merecem, a nosso juízo, aprimoramento.

3 Síntese dos problemas metodológicos

Embora o segundo trabalho tenha trazido avanços ao primeiro, alguns erros persistem e prejudicam as conclusões de ambos os trabalhos. O motivo disso é que o dataset usado no segundo artigo é mesmo do primeiro, de modo que falhas originárias sejam propagadas ao segundo estudo. Os problemas em síntese são o vazamento de dados, a remoção de valores nulos, a remoção de variáveis com alta cardinalidade e a limitação temporal ao ano de 2018. Trataremos ponto a ponto.

4 O vazamento de dados

O cerne dos dados da ANTAQ é a tabela de atracação. A atracação significa a parada de um navio em um porto e carrega informações relevantes sobre o tempo de estadia, dentre outras. Uma outra tabela relevante é a de Carga, definida pela expressão concreta de uma determinada Mercadoria. Ou seja: o minério de ferro é a mercadoria, um volume específico de minério de ferro é a carga e uma atracação pode envolver muitas cargas que estão nesse navio sendo transportadas. A relação entre carga e atracação, então, é muitos-para-um.

Essa relação é central para entender o vazamento de dados. Enquanto a carga possui variáveis como porto de origem e porto de destino, a Atracação registra os tempos de espera dos navios dos portos. Para ter ambas as features no dataset os autores originais uniram as duas tabelas por meio de um campo chave. Isso gerou o vazamento de dados pois propagou a variável alvo em diversas linhas. Assim, imagine que um navio (e sua respectiva atracação) tenha 10 cargas. Da maneira como foi feito, isso irá gerar 10 linhas no dataset, uma para cada carga, cada uma delas carregando informações sobre a atracação. Dessas 10 linhas, vamos supor que 8 fiquem no treino e 2 no teste. Estaríamos vazando, no teste, casos já vistos no treino. A figura 1 mostra a repetição de dados por criar uma linha por carga indevidamente.

Esse vazamento afeta tanto o primeiro quanto o segundo artigo. No caso do primeiro, os modelos foram avaliados com 10fold-cross-val e no segundo, os hiperparâmetros foram encontrados

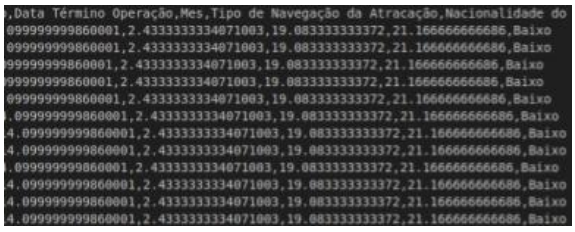

Figura 1: Demonstração do vazamento de dados.

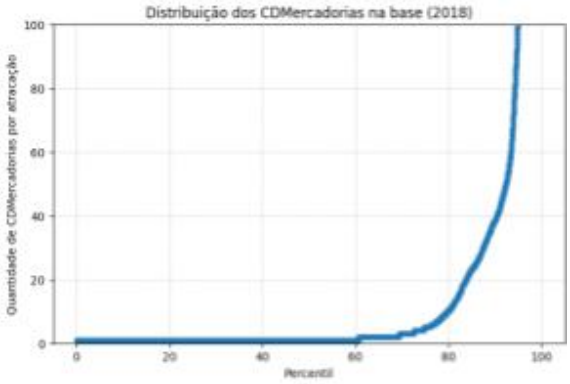

Figura 2: Distribuicao do vazamento de dados

com 10fold-cross-val e depois separaram treino e teste. Para corrigir o problema, é preciso agregar os dados sobre as cargas, de modo a ter uma linha por atracação. Isso faz mais sentido quando se pretende ler o tempo de atracação e evita esse vazamento. Do jeito que está modelaram a carga-atracação, o que é de difícil interpretação e contém esse vazamento, não o tempo de atracação. Esse tipo de erro metodológico [4] compromete a validade empírica dos resultados e é reconhecido como uma das principais causas de overfitting em modelagem preditiva. A figura 2 mostra que 40 porcento da base tem mais de 1 Mercadoria (e portanto, carga, já que uma carga só pode ter 1 mercadoria), mostrando o grande impacto do vazamento.

Corrigir esse problema leva a alguns desafios metodológicos a respeito da agregação de cargas, assunto que trataremos a seguir. O segundo trabalho, ainda, encontra os hiperparâmetros por meio de kfold-cross-val para depois separar em treino em teste. Essa escolha enviesa os hiperparâmetros, que terão sido definidos tendo visto os dados de teste indevidamente. Melhor seria separar originalmente em validação (para encontrar os hiperparâmetros), treino (para efetivamente treinar) e teste (para avaliar o desempenho do modelo sem contaminação).

5 Remoção de valores nulos

A remoção dos nulos no dataset usado nos artigos cria um viés de seleção (na modalidde viés de sobrevivência [3]) porque esses nulos podem carregar informação sobre o processo de produção dos casos que podem estar relacionados com a variável alvo de maneira diferencial.

6 Remoção de variáveis com alta cardinalidade

O dataset usado nos artigos analisados também passou pela remoção de variáveis com alta cardinalidade. Essa escolha pode ter retirado features muito informativas para o problema. Por exemplo: a Mercadoria, que é o ponto principal da Carga, foi removida. Se está se tentando modelar o tempo de permaência de navios em portos, o tipo de mercadoria não faz nenhuma diferença? Será que não há distinção nenhuma entre tomates, minério de ferro, material explosivo ou corrosivo e chips eletrônicos? Nesse caso em particular defendemos incluir a informação da mercadoria ao mesmo tempo que se agrega as cargas (e mercadorias) para que se tenha uma linha por atracação. One-Hot-Encoding não é uma boa opção porque isso criaria um número muito grande de colunas. Nossa proposta, ainda em desenvolvimento a essa altura, é representar a mercadoria por seus embeddings, e agregá-los pela média ponderada dos pesos das Cargas e pelo valor declarado da Carga transportada. Assim, dois campos resumem uma miríade de mercadorias transportadas em um navio que está atracado em um porto, e aquelas mercadorias com maior participação (em peso ou em valor) ganham também maior relevância no embedding pela ponderação.

7 Limitação temporal

As conclusões seriam mais sólidas se considerassem dados de mais anos além de 2018.

8 Conclusões

As escolhas de modelagem do dataset feitas no primeiro trabalho analisado comprometem seus resultados e o do trabalho posterior. É preciso corrigir o vazamento de dados agregando as cargas, de modo a produzir uma linha por atracação. Também é preciso evitar o viés de sobrevivência ao dar soluções melhores para os nulos que sua simples remoção. Também pode-se incluir variáveis que foram excluídas por alta cardinalidade mas cujos desafios técnicos são contornáveis. Por fim, é possível ter uma amostra mais representativa incluindo dados de mais anos.

Referências

[1] Abreu, L. R., Maciel, I. S. F., Alves, J. S., Braga, L. C., & Pontes, H. L. J. (2023). *A decision tree model for the prediction of the stay time of ships in Brazilian ports*. Engineering Applications of Artificial Intelligence, 117, 105634. https://doi.org/10.1016/j.engappai.2022.105634

[2] Rao, A. R., Wang, H., Gupta, C. (2024). *Predictive Analysis for Optimizing Port Operations*. arXiv preprint arXiv:2401.14498. Available at: https://arxiv.org/abs/2401.14498

[3] Brown, S. J., Goetzmann, W. N., Ibbotson, R. G., & Ross, S. A. (1992). *Survivorship Bias in Performance Studies*. *Review of Financial Studies*, 5(4), 553–580.

[4] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2011). *Leakage in Data Mining: Formulation, Detection, and Avoidance*. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), 1–21. https://doi.org/10.1145/2048620.2048623