

Na rota dos sinistros: predição de roubos de carga no transporte rodoviário brasileiro usando métodos de Machine Learning

JESSICA ALMINO

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

JOÃO VINÍCIUS DE FRANÇA CARVALHO

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E ATUÁRIA DA UNIVERSIDADE DE SÃO PAULO - FEA

NATHALIA COSTA FONSECA

UNIVERSIDADE DE SÃO PAULO (USP)

Introdução

O setor de transporte é estratégico para a economia brasileira, pois se destaca como fator que influencia as diversas atividades produtivas, por ser um elo de integração territorial que liga oferta a demanda e polos produtores aos grandes centros urbanos. Isto ocorre porque o Brasil é o país com a maior distância norte-sul e possui a quarta maior rede rodoviária do mundo. Devido aos altos índices de roubo de carga no Brasil e seu status como um dos principais exportadores globais de alimentos, o Brasil é um caso de estudo relevante para a área de seguro de transporte de cargas.

Problema de Pesquisa e Objetivo

O objetivo deste trabalho é comparar modelos tradicionais (GLM) e modernas (Machine Learning e Deep Learning) para estimar a ocorrência de roubo de cargas, utilizando métricas apropriadas de desempenho. A inclusão dessas métricas permite avaliar os efeitos do forte desbalanceamento amostral dos casos em relação aos expostos, oferecendo uma visão ampla de como diferentes aspectos do desempenho são influenciados. Além disso, o estudo buscará não apenas identificar os fatores de entrada do modelo, mas também determinar a importância relativa desses fatores para cada abordagem.

Fundamentação Teórica

A indústria de seguro é peculiar por não se saber de antemão o custo associado ao seu serviço. Para estimar a potencial perda, é necessário analisar duas dimensões do risco: severidade e frequência. A severidade está relacionada ao montante financeiro da perda econômica e a frequência, ao número de vezes em que se incorre em perdas. A convolução dessas dimensões é relevante para calcular um prêmio que equilibre a expectativa da despesa futura, dimensionar o capital de solvência, mantendo a probabilidade de ruína da seguradora em nível aceitável.

Metodologia

Esta pesquisa utiliza técnicas estatísticas e de aprendizado de máquina, incluindo Árvore de Decisão, Random Forest, Deep Learning, XGBoost, Naïve Bayes e o tradicional GLM, para realizar uma análise aprofundada do risco de roubo de carga. O conjunto de dados abrange informações sobre o transporte de mercadorias, totalizando 57.450.563 observações de apólices expostas, que representam os registros realizados pelo segurado para garantir que o seguro esteja ativo durante o transporte da carga. Todo o processamento e análise dos dados foram realizados utilizando o software R.

Análise dos Resultados

Os resultados mostraram que o modelo de Random Forest apresentou o melhor desempenho, destacando-se pela sensibilidade e F1-Score, essenciais em cenários com forte desbalanceamento de classes. As métricas AUC-PR e F1 são recomendadas para esse tipo de dado, pois equilibram precisão e sensibilidade na classe minoritária, ao contrário da AUC tradicional, que pode superestimar o desempenho. Entre as variáveis mais influentes, destacam-se “Rota: MG->SP,” “Periféricos de Computadores” e “Estrada Não Urbanizada,” refletindo riscos elevados nessas condições.

Conclusão

O impacto do roubo de carga na economia vem sendo amplamente discutido na literatura acadêmica. Contudo, análises baseadas em dados de roubo de carga com o uso de técnicas modernas de modelos preditivos têm recebido atenção apenas mais recentemente. Este trabalho contribui ao comparar o desempenho de modelos (Árvore de Decisão, Random Forest, Redes Neurais, XGBoost e GLM) na detecção de ocorrência de sinistros de roubos de cargas. Além disso, avaliaram-se os desempenhos dos modelos a partir de métricas relevantes, assim como encontraram-se as variáveis mais determinantes para esta finalidade.

Contribuição / Impacto

As principais contribuições deste estudo recaem sobre dois pilares: (i) uma análise aprofundada das variáveis de risco no transporte de carga, com a identificação de fatores críticos, como tipo de mercadoria e rota, que elevam o risco de roubo, e; (ii) a aplicação de modelos preditivos de ML e DL em dados desbalanceados, evidenciando desempenho superior em métricas específicas de classificação, especialmente em relação a modelos tradicionais como o GLM.

Referências Bibliográficas

- Frees, E. W. (Jed), & Huang, F. (2023). The Discriminating (Pricing) Actuary. *North American Actuarial Journal*, 27(1), 2-24.
- Silva, B. T., & Sampaio, M. (2023). Factors influencing cargo robbery in last-mile delivery of e-commerce: an empirical study in Brazil. *Journal of Transportation Security*, 16(1), 10.
- Song, R., Huang, L., Cui, W., Óskarsdóttir, M., & Vanthienen, J. (2020). Fraud Detection of Bulk Cargo Theft in Port Using Bayesian Network Models. *Applied Sciences*, 10(3), 1056.