SemeAd 2024
XXVII Seminários em Administração

05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

ABRINDO A CAIXA PRETA DOS MODELOS DE INTELIGÊNCIA ARTIFICIAL: APLICAÇÃO DE MÉTODO DE EXPLICABILIDADE PARA REDUZIR VIESES

CAROLINA ROBLEDO VELINI DE ANDRADE
UNIVERSIDADE DE SÃO PAULO (USP)

DAIELLY MELINA NASSIF MANTOVANI
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

CELSO MACHADO JR.
UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL (USCS)

GUILHERME AREVALO LEAL
UNIVERSIDADE DE SÃO PAULO (USP)

SemeAd 2024
XXVII Seminários em Administração

05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

ABRINDO A CAIXA PRETA DOS MODELOS DE INTELIGÊNCIA ARTIFICIAL: APLICAÇÃO DE MÉTODO DE EXPLICABILIDADE PARA REDUZIR VIESES

Introdução

Diante o cenário atual de Big Data onde uma grande quantidade de dados é gerada diariamente, muitas empresas e organizações recorrem a variados métodos para melhor analisar dados e obter informações relevantes para os processos de tomada de decisão. Os modelos black-box são aplicados por demonstrarem ótimos resultados, porém, esses modelos são aqueles que não podem ser facilmente compreendidos, não permitindo que o usuário tenha a transparência sobre o que está acontecendo para que se gere determinada predição, o que levanta questionamentos éticos acerca de sua aplicação.

Contexto Investigado

Logo, é necessária a aplicação de métodos de interpretabilidade para que se obtenha uma aproximação do que está influenciando a predição gerada por um modelo black-box, para se garantir que os resultados gerados pelo modelo não sejam fruto de propagações em padrões discriminatórios nos dados, por exemplo.

Diagnóstico da Situação-Problema

Foram selecionadas base de dados em contextos de decisões de alto e baixo risco de se cometer uma falha ética: uma sobre uma campanha de marketing e a aderência dos usuários ao produto, e outra sobre o risco de crédito perante calotes, nos quais foram aplicados modelos black-box de classificação binária que tiveram predições interpretadas pelo método LIME (Local Interpretable Model-Agnostic Explanations).

Intervenção Proposta

Foram desenvolvidos três modelos black-box para cada base de dados (XGBoost, CatBoost e Light GBM), aferindo-se suas métricas de qualidade e taxa de acerto. Aplicou-se então o método XAI LIME para três casos em cada modelo, dois classificados corretamente e um caso classificado incorretamente, avaliando-se quais variáveis impactaram na decisão do modelo em classificar o caso como sucesso/fracasso.

Resultados Obtidos

Os resultados permitiram gerar explicações locais que evidenciaram as variáveis mais importantes para a classificação de determinadas observações no grupo positivo e negativo, além disso, foi possível observar que as variáveis significantes estavam ligadas ao contexto do problema e não às características socioeconômicas dos indivíduos, indicando uma baixa probabilidade de viés discriminatório em relação a esses fatores. Uma preocupação das organizações em relação à IA é que sua aplicação é promissora mas, o processo analítico artificial difere do raciocínio humano, focalizando no resultado.

Contribuição Tecnológica-Social

Essa característica pode levar a graves vieses no processo decisório, por exemplo, cometendo-se discriminação de grupos sociais ainda que de forma não intencional, ou realizando recomendações de ações inadequadas (recomendar um diagnóstico equivocado a um paciente e com isso o tratamento errado). O estudo demonstrou que o XAI permite verificar com clareza as regras de classificação permitindo identificar vieses e aumentando a confiança nos modelos de IA. Destaca-se que a intervenção humana em alguma medida é relevante para que se possa confiar nos modelos e

05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

reduzir vieses e problemas éticos.