Instituto de Matemática e Estatística - USP

MAC0459/MAC5865 - Data and Engineering Science

General Test - 2021 - QUESTÃO 4:

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): Ensaio de revisão sobre o artigo "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" by Arrieta et. al. (2020)

Edilson Pereira dos Santos - N° USP:

RESUMO: Este estudo de revisão buscou analisar criticamente o artigo "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" by Arrieta et. al. (2020), ressaltando os aspectos positivos do documento, suas falhas e considerações para trabalhos futuros. A pesquisa principal do trabalho objeto de análise debruçou-se sobre o tema de inteligência artificial explicável (XAI) definindo seus conceitos, sua importância para o contexto atual de desenvolvimento tecnológico e aprofundamento do termo para um aprimoramento conceitual que desemboca em uma filosofia de 'Inteligência artificial responsável' que não se limita a ser inteligível por parte dos usuários de IA bem como seus desenvolvedores, mas também ética e contextualizada com os valores humanos contemporâneos de modo que o máximo benefício oriundo de algoritmos de IA possam ser alcançados para maior usufruto de todos. Para tal, o trabalho considerou analisar um total de 400 artigos sobre XAI, sintetizando seus aspectos principais e respectivos resultados, considerando as vantagens de sistemas de XAI sustentados pelos autores que são: 1 A interpretabilidade ajuda a garantir a imparcialidade na tomada de decisão, ou seja, a detectar e, consequentemente, corrigir o enviesamento no conjunto de dados de treinamento; 2 A interpretabilidade facilita o fornecimento de robustez, destacando potenciais perturbações adversárias que podem alterar a previsão; e, 3 A interpretabilidade pode atuar como uma garantia de que apenas variáveis significativas inferem a saída, ou seja, garantindo que uma causalidade verdadeira subjacente exista no raciocínio do modelo.

Palavras-Chave: Inteligência artificial explicável, interpretabilidade, responsabilidade.

ARGUMENTO PRINCIPAL

Este ensaio aborda sobre Inteligência Artificial (IA) responsável a partir da análise e síntese de conceitos ligados à ética e transparência. Anterior a isso, o documento apresenta a necessidade de refletir sobre o avanço tecnológico que a área de IA vem sofrendo e consequente aumento do distanciamento em termos de 'explicabilidade' e 'interpretabilidade' para desenvolvedores e usuários leigos. Quando comparado aos modelos algoritmos de IA iniciais, como, árvores de decisão, regressão linear e regressão logística, os atuais apresentam maior complexidade, como os modelos de redes neurais profundas que surgiram recentemente por conta do avanço computacional baseado em melhor utilização de GPU dos computadores.

MÉTODOS

O estudo de Arrieta et al. (2020) é separado em sete seções sendo: seção 1 composta pela introdução ao documento; seção 2 trata da conceituação das terminologias pertinentes ao tema de explicabilidade e interpretabilidade de modelos de IA; seções 3 e 4 que contém uma revisão sobre os principais avanços em XAI seção 5 composta por uma discussão sobre vantagens e cuidados para o futuro das sinergias entre as famílias de métodos , em que é apresentada uma perspectiva de desafios gerais e algumas consequências sobre as quais deve se ter cuidado no futuro; seção 6 desenvolve o conceito de IA Responsável; e, por fim,a seção 7 conclui o estudo com uma perspectiva que procura atrair a comunidade científica em torno do assunto principal do trabalho de revisão.

ANÁLISE CRÍTICA

Os autores tiveram o cuidado de apresentar uma estrutura bem definida que facilita o entendimento sobre o tema, fazendo a junção entre os artigos analisados durante o estudo e os conceitos desenvolvidos. Percebe-se também um cuidado especial com a seção 2 que trata da terminologia aplicada à XAI. Este cuidado reflete-se na divisão dos diversos conceitos em 'sub-tópicos' que são apresentados graficamente de modo a esclarecer o grau de interpretabilidade e explicabilidade de alguns modelos de ML e IA na seção 3, partindo de árvores de decisão até redes neurais profundas. Fica muito claro que os modelos tornaram-se muito complexos a ponto de não poderem ser explicados sem ajuda 'post-hoc' com o passar do tempo. No contexto dos sub-tópicos apresentados na seção 2 e 3, os mais importantes são simulatibilidade, decomponibilidade e transparência algorítmica. A simulatibilidade trata do quão possível é para um ser humano simular o funcionamento de um modelo de aprendizado de máquina sem necessidade de recursos de explicação, por outro lado, a decomponibilidade é capacidade de segmentar a estrutura de um modelo de ML e/ou IA para estudá-los parcialmente e a transparência algorítmica, embora tenha sido colocada como um sub-tópico importante para classificar os modelos em graus de interpretabilidade, parece bastante nebuloso e pouco explicativo Nas palavras dos autores, este termo possui muitas definições e talvez por isso, soe desta forma.

O autor deste presente ensaio julga que seria melhor tratar o tema, portanto, de forma menos abrangente como foi trazido no diagrama de apresentação dos modelos de aprendizado de máquina a fim de poupar a generalidade que o termo reforça, todavia, as seções 2 e 3 são fundamentais para compreensão total da proposta do artigo na totalidade. Em prosseguimento, a seção 4 ocupa-se de explicitar os métodos para interpretação e explicação de modelos com maior teor de complexidade como as redes neurais. Contudo, parece um pouco óbvio que modelos de aprendizado de máquina precisem de diagramas de interpretação para melhor compreensão dos usuários e isso indifere de quão simples ou complicados sejam estes modelos.

Para todos os casos, é sempre de bom-tom, tornar um assunto claro e evidente quando se pretende apresentá-lo, como evidencia Cole NussBaumer Knaflick em seu bestseller *'Storytelling with data: A Data visualization Guide for Business Professionals'* [1] Para ela, é melhor que algo seja 'digerível' e 'facilmente interpretável' do que não seja. Neste contexto, todos os modelos de aprendizado de máquina e IA dependem de ser explicados claramente por mais simples que pareçam. Por outro lado, os autores preocuparam-se em detalhar diversos aspectos de como abordar as técnicas de explicação de modelos de ML e IA baseados em estudos anteriores e isto tem valor, pois, podem ser replicados e utilizados de forma útil no futuro. Há também na seção 4, um cuidado em apresentar o conceito de taxonomia de métodos de explicação para modelos de ML e IA, utiliza-se um diagrama que apresenta hierarquicamente como seriam distribuídos os modelos em termos de técnica de explicação destes. A seção 5 inicia-se revisando conceitos abordados ao longo do artigo e reforça que existe uma relação inversamente proporcional entre desempenho e interpretabilidade de modelos a partir de certo ponto.

Porém, estes eventos de ineficiência do modelo só podem aparecer ao construí-lo pensando em como explicá-lo, se, por outro lado, a construção do modelo for realizada e de forma sequencial houver um foco em como explicá-lo, este efeito poderá ser reduzido ou até mesmo mitigado. Deve-se ainda ressaltar que não se pode simplificar demais uma explicação dado o teor de especificidade de um determinado assunto. Evidenciando o fato, suponhamos que um modelo de ML ou IA seja construído utilizando os parâmetros cabíveis ao problema e deixando-os claramente definidos, qual será o efeito de queda de desempenho deste modelo ao ser explicado depois que este modelo estiver pronto? Neste exemplo, não parece razoável refazer um modelo para que ele se torne inteligível, na contra mão, é mais aceitável que a explicação se adeque ao modelo e não o modelo à explicação. Ainda assim, o artigo se debruça sobre um equilíbrio entre desempenho e explicabilidade digno de nota, pois, de fato, modelos mais complexos exigem maior cuidado para serem explicados, e, portanto, quanto mais complexos, mais difíceis de se interpretá-los e a esta altura, o artigo apresenta uma pequena deficiência.

Afinal, para quem os modelos precisam ser interpretáveis e explicáveis? O que sugere-se pensar é que a especificidade é inerente ao desenvolvimento da humanidade e exemplificando, não se pode esperar que um paciente em estágio terminal entenda todos os aspectos técnicos de sua doença, isto cabe aos médicos.De modo análogo, no campo da ciência de dados, não se pode esperar que o usuário de insights entenda todos os conceitos de um determinado modelo, tendo em vista que ele usufrui apenas dos resultados deste. Ou seja, parece existir um limite para explicabilidade de modelos de ML e IA, claro, é importante que não se caia no âmbito dos modelos de 'Caixa preta' onde nem o criador, nem o usuário seja capaz de interpretá-lo, mas não se pode esperar reduzir ao extremo a complexidade de um modelo para que até leigos o entendam. Isso não ocorre em nenhuma profissão. As seções 6 e 7 do artigo apresentam a necessidade de se pensar em uma inteligência artificial responsável que vá de encontro a valores importantes da sociedade

contemporânea como inclusão e diversidade social entre outros aspectos. Aos olhos do autor deste ensaio, estas seções são as partes de maior relevância do documento,visto que a necessidade principal de se saber como um algoritmo funciona está baseada no fato de que estes algoritmos podem reproduzir questões de cunho muito sensível como, por exemplo, 'racismo algorítmico' em que modelos de processamento natural de imagens podem priorizar etnias em detrimento de outras. Casos como estes aparecem com frequência em veículos de imprensa e em [2], um modelo de IA classifica um ator negro de Hollywood como suspeito de um crime cometido na região nordeste do Brasil, um caso típico de enviesamento de modelos de IA. Existem questões sociológicas que precisam, evidentemente, ser aprofundadas para tratar deste assunto, mas dada a frequência com que eventos deste tipo ocorrem, há uma necessidade de atenção por parte da comunidade científica em prol de algoritmos e modelos mais justos e alinhados com as expectativas sociais contemporâneas. Evidenciando, a ferramenta *Google Ngrams* [3] aponta o crescimento vertiginoso da expressão 'algorithmic discrimination' no mesmo período de tempo da evolução dos principais modelos de ML e IA, que tiveram uma ascensão a partir de 2012, portanto, IA responsável é uma ideologia urgente e necessária.

Referências Bibliográficas

[1] KNAFLIC, C. N. **Storytelling with Data:** A Data Visualization Guide for Business Professionals (English Edition). 1ª ed., 2015. 252 p.

[2] FOTO de astro do cinema Michael B. Jordan aparece em lista de procurados pela polícia do Ceará. **G1**, Ceará, 7 jan. 2022.
Disponível em: https://g1.globo.com/google/amp/ce/ceara/noticia/2022/01/07/astro-do-cinema-michael-b-jordan-aparece-em-lista-de-procurados-pela-policia-do-ceara.ghtml.
Acesso em: 8 jan. 2022.

[3] Algorithmic discrimination. In: **Google Books Ngram Viewer**. 2022. Disponível em:
https://books.google.com/ngrams/graph?content=algorithmic+discriminationyear$_s tart = 1800 year_e nd = 2019 corpus = 26 smoothing = 3 direct_u rl = t1$