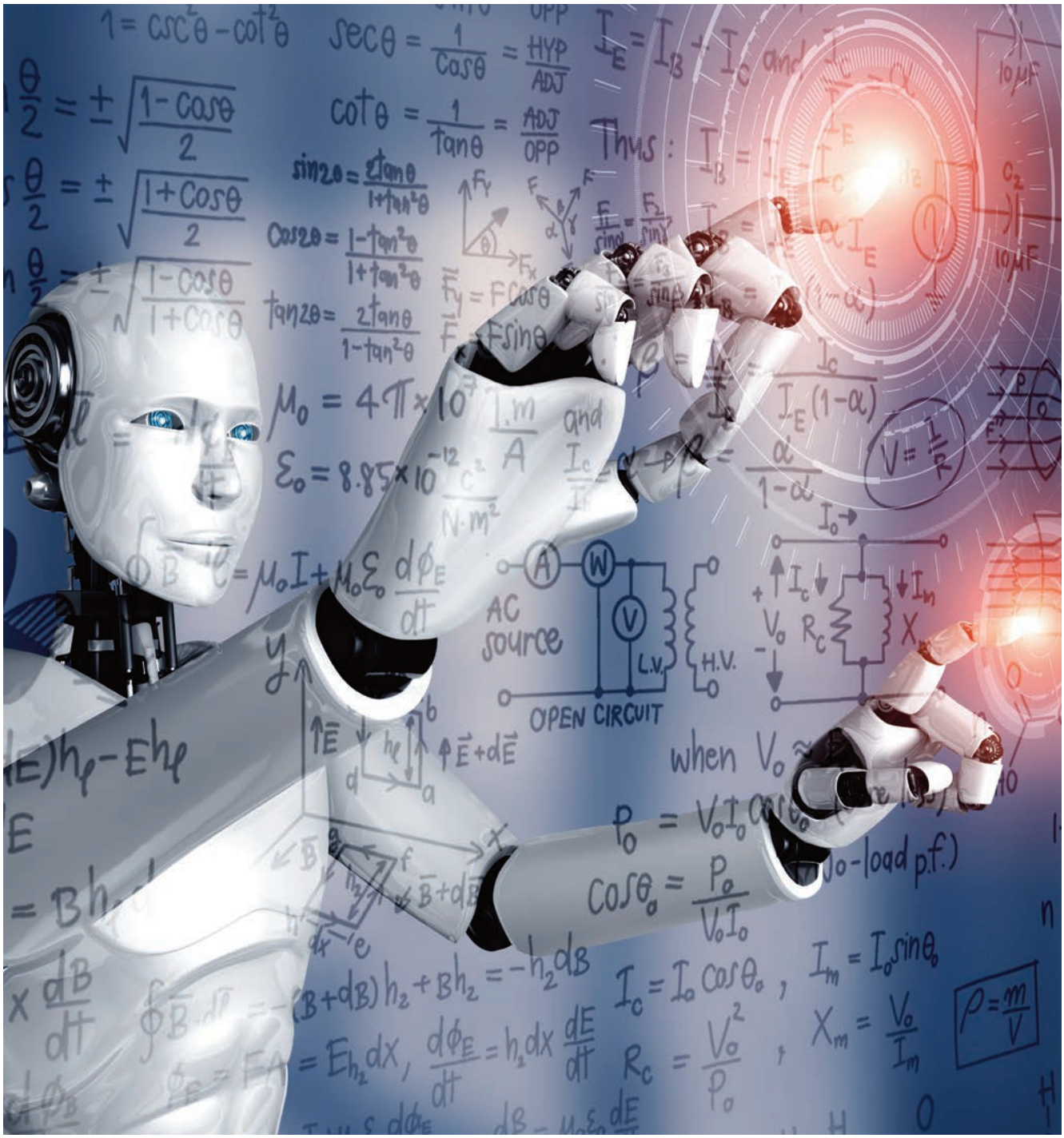
Resumo executivo

"Toda tecnologia deve ser acompanhada de um manual especial: não sobre como usá-la, mas por quê, quando e para quê."
Alan Kay[12]



Contexto e fundamentos da XAI

1. A transformação digital possibilitou o acesso e a exploração de uma grande quantidade de dados estruturados e não estruturados, impulsionando o uso de técnicas de machine learning e inteligência artificial em todos os setores.

2. Os modelos de AI oferecem maior poder de previsão, mas também apresentam riscos, como a presença de vieses inadvertidos, a falta de compreensão do modelo ou erros em sua aplicação decorrentes de causas como o treinamento excessivo, o que pode levar à desconfiança em relação ao modelo. Isso levanta a questão de saber se é possível entender os resultados dos algoritmos de AI suficientemente bem para tomar as decisões adequadas.

3. A Inteligência Artificial Explicável (XAI) é um conjunto de processos e métodos que permite que os usuários entendam e confiem nos resultados e saídas criados pelos algoritmos de machine learning. Essa disciplina é fundamental para que uma organização crie confiança ao empregar modelos de AI, ajudando a caracterizar a precisão do modelo, a imparcialidade, a transparência e a compreensão dos resultados na tomada de decisões baseada em AI.

4. O interesse acadêmico e profissional na XAI aumentou exponencialmente nos últimos anos, devido à capacidade da disciplina de abordar uma série de preocupações do setor no uso da AI, como requisitos regulatórios, falta de confiança, potencial uso indevido, impacto reputacional, impactos sociais ou humanos e outros riscos.

5. Isso levou os órgãos reguladores e supervisores de diferentes jurisdições a estabelecer regulamentos e diretrizes para o uso adequado da AI, incluindo aspectos de interpretabilidade do modelo.

6. Na Europa, o Regulamento Geral sobre a Proteção de Dados (GDPR) do Parlamento Europeu, que entrou em vigor em 2018, estabelece um "direito a uma explicação" para os cidadãos, exigindo que as empresas possam explicar por que um modelo de AI retornou um determinado resultado. Isso tem implicações críticas para o design e a análise de interpretabilidade dos modelos de AI.

7. Além disso, o Parlamento Europeu propôs em 2021 o Artificial Intelligence Act (AI Act) para regulamentar o uso da inteligência artificial na União Europeia. Esse regulamento proposto estabelece um framework regulatório para sistemas de AI, incluindo requisitos de desenvolvimento ético, transparência, segurança e precisão, bem como um sistema de governança e supervisão. A Lei de AI classifica os aplicativos de AI em níveis de risco (práticas inaceitáveis, sistemas de alto risco e sistemas de risco baixo ou limitado) e estabelece requisitos de transparência e supervisão humana para sistemas de alto risco, que serão aplicáveis em toda a União. É provável que isso desencadeie iniciativas de adaptação ao Regulamento, como documentação abrangente de modelos, técnicas de interpretabilidade, dashboards de monitoramento e alertas de modelos, entre outros.

8. Além disso, a Comissão Europeia formulou em 2019 as Diretrizes Éticas para Inteligência Artificial Confiável, que propõem sete requisitos principais para que os sistemas de AI sejam considerados confiáveis: (i) ação e supervisão humanas, (ii) robustez técnica e segurança, (iii) gestão da privacidade e dos dados, (iv) transparência, (v) diversidade, não discriminação e equidade, (vi) bem-estar ambiental e social e (vii) responsabilização. Dentro do requisito de transparência, é estabelecida a necessidade de explicabilidade dos modelos de AI. As Diretrizes propõem critérios de avaliação para determinar até que ponto um modelo de AI atende a esses requisitos.

9. Nos Estados Unidos, em 2022, a Casa Branca propôs uma minuta da Declaração de Direitos sobre a AI (AI Bill of Rights), promovida pelo presidente Joe Biden. Esse projeto de lei estabelece cinco princípios ou direitos dos cidadãos com relação à AI, incluindo sistemas seguros e eficazes, proteção contra discriminação dos algoritmos, privacidade de dados, notificação e explicação, e avaliação e correção humanas em caso de falha da AI (fallback). Esses princípios incluem a explicabilidade dos modelos de AI, que exige

[12] Alan Kay (nascido em 1940), cientista da computação americano ganhador do Prêmio Turing, considerado o "pai dos computadores pessoais".

documentação em linguagem simples, explicações tecnicamente válidas, significativas e úteis, e notificações de uso comprovadamente claras, oportunas, compreensíveis e acessíveis.

10. Os Princípios da OCDE sobre Inteligência Artificial de 2019 promovem o uso de AI que seja confiável e respeite os direitos humanos e os valores democráticos. Eles foram adotados por todos os 38 países membros da OCDE e exigem, entre outros, transparência e divulgação responsável dos sistemas de AI para que as pessoas afetadas por um sistema de AI possam compreender o resultado.

11. O Discussion Paper on Machine Learning for IRB Models[da Autoridade Bancária Europeia (EBA), publicado em 2021, analisa a relevância de possíveis barreiras à implementação de técnicas de machine learning na abordagem IRB para o cálculo de capital em instituições financeiras. O documento estabelece princípios e recomendações para tornar o uso dessas técnicas compatível com a conformidade com a regulação europeia de requisitos de capital (CRR). Essas recomendações incluem análise estatística e econômica da relação entre as variáveis de entrada e a variável de saída, documentação que explique o modelo de forma simples e a necessidade de detecção de possíveis vieses no modelo.

12. Um princípio básico da XAI é a necessidade de integrar a interpretabilidade e a explicabilidade à organização e aos processos de uma empresa. Isso é feito por meio de um framework de XAI que consiste em quatro elementos: técnicas de interpretabilidade de modelos de AI, integração em processos de gestão de risco de modelo (MRM), suporte tecnológico e fatores humanos.

13. Técnicas: o núcleo do framework de XAI baseia-se em três aspectos principais de interpretabilidade: a explicação do desenho do modelo, a explicação dos resultados do modelo e outros aspectos, como a detecção de viés e o monitoramento periódico do modelo.

14. MRM: A interpretabilidade dos modelos de AI é uma característica que afeta toda a cadeia do ciclo de vida do modelo e, portanto, a gestão do risco do modelo. Para incorporar elementos XAI, a estrutura organizacional e de governança, as políticas e os procedimentos para desenvolvimento, monitoramento, validação, implementação e uso de modelos, bem como a estrutura de auditoria, precisam ser revisados e atualizados.

15. Suporte tecnológico: para implementar um framework de XAI, são necessárias soluções tecnológicas profissionais para dar suporte aos aspectos de interpretabilidade dos modelos de AI, como ferramentas de interpretabilidade e de governança de modelos, sistemas de análise de dados, APIs, mecanismos de segurança e auditoria e protocolos para garantir a conformidade com os padrões de qualidade e explicabilidade.

16. Fator humano: a integração da XAI deve considerar o fator humano, incluindo o recrutamento e a retenção de talentos especializados, programas de treinamento, a criação de uma cultura que aprimore o uso da explicabilidade e interpretabilidade dos modelos de AI e programas de gestão de mudanças para garantir a adoção adequada da XAI.

17. Além disso, um quinto elemento central para a AI e a XAI são os dados, pois sua governança, qualidade, integridade, consistência, rastreabilidade e ausência de viés determinam a qualidade do modelo de AI e, em última análise, das decisões tomadas com base nele. No entanto, as questões relacionadas aos dados e sua relevância para os modelos não são o tema deste documento, pois já foram amplamente abordadas em publicações anteriores[13].

Técnicas de interpretabilidade: estado da arte

18. O uso de técnicas de AI se espalhou por todos os setores e domínios, oferecendo maior poder de previsão em troca de maior complexidade. Isso gerou a necessidade de explicar os resultados dos modelos de AI, o que levou ao surgimento de técnicas cada vez mais sofisticadas de interpretabilidade local e global. Essas técnicas não resolvem completamente o problema, e diferentes abordagens para garantir a interpretabilidade dos modelos de AI continuam a ser desenvolvidas, como o desenvolvimento de modelos inerentemente interpretáveis ("caixas brancas").

19. As abordagens mais comuns para tratar o problema da interpretabilidade podem ser classificadas em dois grupos: interpretabilidade post-hoc (técnicas de interpretabilidade global e local) e modelos inerentemente interpretáveis. Além disso, há estratégias complementares, como a simplificação do modelo, o uso de variáveis de senso comercial, a análise de dados para identificar parcialidade ou imparcialidade e a análise da reprodutibilidade do desenvolvimento do modelo.

20. O LIME (Local Interpretable Model-agnostic Explanations) permite que um modelo seja explicado de forma local e agnóstica, ou seja, ele pode gerar explicações para uma previsão específica sem a necessidade de entendimento do modelo subjacente.

21. O SHAP (SHapley Additive exPlanations) explica o modelo de forma global, avaliando a contribuição de cada variável de entrada para a previsão de saída.

22. Os PDPs (Partial Dependence Plots, gráficos de dependência parcial) são usados para visualizar como o resultado de um modelo muda quando os valores das variáveis de entrada são alterados.

[13] Ver Management Solutions (2020, 2018 e 2015): "Auto machine learning, rumo à automação de modelos", "Machine learning, uma peça-chave na transformação dos modelos de negócios" e "Data science e a transformação do setor financeiro".

23. Os modelos white box baseiam-se no desenvolvimento de algoritmos que, por sua concepção, são inerentemente interpretáveis. Esses modelos são agrupados de acordo com o tipo de algoritmo usado, e os parâmetros a serem otimizados geralmente são limitados para obter maior interpretabilidade. Isso resulta em resultados mais precisos, pois permite uma melhor compreensão das informações, o que, por sua vez, resulta em uma melhor tomada de decisão, especialmente nos setores em que a interpretabilidade é um fator crítico.

24. Apesar dos avanços na interpretabilidade dos modelos de AI, ainda há desafios, como a reprodutibilidade dos resultados, a explicação da sequência de previsões mais prováveis, vieses nos dados de entrada, imparcialidade (fairness) e precisão da explicação. Além disso, há espaço para melhorias no desenvolvimento de modelos white box para competir em precisão com modelos black box em problemas complexos, bem como no desenvolvimento de novas técnicas para explicar modelos mais complexos.

Estudo de caso de interpretabilidade

25. Para demonstrar a aplicação das técnicas de interpretabilidade descritas acima, é realizado um exercício ilustrativo usando dados fictícios gerados pela IBM e publicados no Kaggle[14]. O objetivo do estudo é entender as causas que levam os funcionários a deixar seus empregos, usando técnicas de AI e XAI nos dados fictícios propostos.

26. O exercício foi realizado com a ajuda de um sistema de modelagem de componentes, o ModelCraft™[15], que contém múltiplas técnicas relevantes de AI e XAI, permitindo que o estudo seja concluído em muito menos tempo do que o normal e sem a necessidade de escrever código.

27. Para explicar o desgaste dos funcionários, diferentes modelos foram treinados e validados, entre os quais o algoritmo de random forest demonstrou a melhor capacidade de previsão.

28. Para explicar os resultados do modelo, foram aplicadas as técnicas de interpretabilidade SHAP, LIME e PDP para entender quais variáveis explicam melhor a fuga dos funcionários, como as mudanças nas variáveis mais importantes afetam diferentes faixas populacionais e os resultados do modelo em casos individuais.

29. A aplicação e a interpretação corretas do modelo nesse estudo de caso permitiriam antecipar e evitar a rotatividade de funcionários, criar perfis com diferentes propensões à fuga e identificar antecipadamente as características desses funcionários para tomada de medidas adequadas. Além disso, esse caso de uso destaca as limitações e as dificuldades na aplicação de técnicas de interpretabilidade post-hoc, bem como o fato de que o uso de modelos de AI juntamente com um módulo de interpretabilidade pode aumentar a capacidade preditiva do modelo.

Conclusão

30. A Inteligência Artificial Explicável (XAI) é uma disciplina emergente que busca melhorar a interpretabilidade dos modelos de AI usando técnicas específicas para entender e explicar os resultados dos modelos de AI, e é especialmente importante em âmbitos de alta sensibilidade, como saúde, segurança, serviços financeiros e de energia, entre outros.

31. A XAI se tornou uma prioridade para muitos setores à medida que os modelos de AI se tornam cada vez mais complexos e cada vez mais regulações exigem sua interpretabilidade. Um estudo de caso desenvolvido com o ModelCraft™ demonstrou como essas técnicas podem ser usadas para entender e explicar os modelos de AI.

32. Nos próximos anos, espera-se que a XAI continue a se desenvolver e a crescer em importância à medida que os modelos de AI se tornam mais complexos, a regulação continue a proliferar e seu uso se estenda para mais áreas de alta sensibilidade.

[14] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.
[15] Ferramenta de AutoML e de modelagem por componentes proprietária da Management Solutions. Ver Management Solutions (2023).