IA eficiente começa com uma boa governança dos dados

O grande desafio das companhias ao desenvolverem projetos de inteligência artificial começa em zelar por uma boa governança de dados. Como o dado é a matéria-prima para qualquer ferramenta de analytics, sem uma coleta e um tratamento corretos, não há IA eficiente e, mais, aumenta-se o risco de insights com vieses. Portanto, zelar pela qualidade dos dados e ter estruturada uma boa governança são peças fundamentais para, entre outros tópicos, conseguir rastrear o caminho dos dados para entender a fonte primária das decisões dos algoritmos.

Mas como assegurar que o processo de coleta, armazenamento e uso dos dados está sendo feito de forma a garantir resultados fidedignos? Segundo especialistas ouvidos para esta reportagem, tudo começa com a definição dos objetivos de negócios e pelo entendimento dos problemas que se quer resolver com o uso dos dados.

Pode parecer simples, mas, na prática, muitas empresas, por ainda não terem claras suas metas, acabam, por exemplo, guardando tudo quanto é tipo de dado e não sabendo, depois, o que fazer com uma enorme quantidade de informações — e muito menos conseguindo gerenciar tudo isso. Assim, definir o propósito deve vir antes de qualquer outra estratégia.

“O dado em si não vale quase nada, se não estiver atrelado ao objetivo do negócio. Este é o cerne da questão”, diz Carlos Abdalad, fundador da consultoria CRMWise Analytics. Para ele, a governança de dados é o segundo passo mais importante da IA: o primeiro é reconhecer o problema que se quer resolver.

“A boa governança de dados é uma jornada, na qual os dados são tratados do princípio ao fim, desde o momento quando se coleta, armazena e se produz algum insight até o momento de descarte, afinal, você não pode guardar infinitamente”, assinala César Patiño, principal advisor da JCP Consulting. Isso pressupõe também entender quais dados estão sendo armazenados, quais autorizações se têm para usá-los, quem pode acessá-los e como está sendo feita a segurança deles, garantindo que estejam bem protegidos contra algum ataque hacker.

A governança, portanto, versa sobre todo o processo que garante as melhores práticas desde a coleta até o descarte. O ideal é estabelecer uma cultura data driven, que começa com o patrocínio da alta gestão e conta com investimentos tanto em equipe como em ferramentas para o melhor manuseio dos dados. As equipes

devem ser multidisciplinares e buscar fontes internas e externas de dados para trabalhar. “Depois, vem a parte de catalogar os dados, entender quais são as fontes de dados que serão trabalhadas, qual é a taxonomia, como serão normatizados os nomes de arquivos, dos campos... tudo isso está organizado no metadados, tendo padrões para como os dados serão compilados e tratados”, aponta Patiño.

Assim, uma IA eficiente começa com a boa governança dos dados. É, segundo Ronald Rowlands, gerente de prévendas do SAS, até possível ter insights sem ter dado governado, mas isso gera insegurança. “Você não sabe quem está acessando a informação ou o que foi trabalhado em cima do dado. A governança de dados vem para mitigar esse problema, fornecer padrões, limpeza das informações para manter o dado disponível e atualizado

e zelar pela privacidade e pela integridade dos dados", diz Rowlands.

A governança está em todo este processo e ela existe para garantir que tudo funcione como deve ser, em conformidade com as leis e seguindo as padronizações e os processos definidos. Se a qualidade de dados não é boa, haverá impacto nos modelos e nos algoritmos, sejam eles machine learning, preditivo ou qualquer outro, levando a um resultado ruim ou insatisfatório na ponta. "Quando falamos de ciência de dados, entre 50% e 80% do tempo do trabalho está na manipulação dos dados", ressalta César Patiño.

A estratégia para governar o dado deve contemplar, de acordo com o gerente de prévendas do SAS, o mapeamento das fontes que estão sendo usadas, a documentação dos processos, o fluxo de trabalho e a criação de regras para fazer o tratamento das informações.

"Deve-se fazer a limpeza e a padronização das informações, mas, primeiro, você precisa entender os dados, ter regras para saber o que faz sentido, documentar e criar dicionário de redes de dados para saber o que tem de informação em cada campo do data lake e do data warehouse", explica Ronald Rowlands.

Um erro comum é as empresas começarem querendo fazer de tudo, em vez de iniciar por partes. Igualmente importante é ter uma estratégia para comunicar a companhia sobre a governança de dados, mapear quem são as partes interessadas e ter um plano de comunicação efetivo para ficar claro para todos o que pode (e não pode) ser feito com os dados. Tudo de forma transparente.

Base para uma IA ética

A governança também está, intrinsecamente, ligada à ética na inteligência artificial. "Um dos princípios para garantir a ética é saber explicar as decisões que a IA tomou e isso é parte da governança", explica Daniel Arraes, diretor de desenvolvimento de negócios para América Latina da Fico. "Não se consegue garantir que decisões de inteligência artificial sejam éticas, se você não tiver governança de dados", acrescenta ele.

É necessário rastrear o caminho dos dados, desde a coleta, permite a identificação de fontes que levaram a vieses e a entender como o algoritmo tomou determinada decisão. "Tem a rastreabilidade física, que é de metadados, e isso é importante para quando alguém questiona de onde apareceu determinado dado, porque os metadados rastreiam tudo que aconteceu com dado. Já o uso que você faz com o dado tem a ver com ética e IA", explica Carlos Abdalad.

Assim, são igualmente importantes as rastreabilidades do dado e do processo que gerou o resultado. "Você tem de acompanhar a história do dado e o que fez com ele para virar decisão", completa Abdalad.

O grande desafio das companhias ao desenvolverem projetos de inteligência artificial é zelar por uma boa governança de dados. Nesse caminho, quais são os principais erros e acertos das companhias nesta jornada? "Quando se fala em governança, ela extrapola não só a questão dos dados, mas a governança deve atuar nas decisões e nos dados que alimentaram essas decisões. Quando se fala em IA, o mundo decisional fica complexo, porque os modelos podem ser bastante complexos, usando um grande número de informações. Então, como você retroage e explica cada etapa da decisão para saber se cada decisão tomada foi justa?", analisa Daniel Arraes, da Fico.

É por isso que as companhias precisam garantir o princípio da rastreabilidade da informação, sendo capaz de rastrear o dado desde quando foi gerado. "Blockchain é tecnologia interessante para isso, porque garante que se consegue buscar a origem da informação de forma fidedigna", acrescenta Arraes.

Tudo isso também tem de ser feito cumprindo os preceitos da legislação em vigor. No Brasil, a Lei Geral de Proteção de Dados determina o que pode ou não ser feito com que tipo de dados.

Já há algum tempo, algumas empresas vêm investindo na função de alocar um profissional responsável pela governança de dados e essa tendência ganha cada vez

mais força devido ao volume de dados e de legislações. “Primeiro, se trata de gestão de risco. Quem não se dedicar a cumprir legislação está sujeito a riscos, principalmente, de imagem e isso custa caro. Então, a preocupação não é opcional; é algo que tem de acontecer. E depois tem a tecnologias. Hoje, as soluções que as empresas usam já tem incorporado o conceito de security by design e, quando falamos em segurança, isso extrapola o ‘não deixar o dado vazar’; é garantir compliance”, destaca o diretor de desenvolvimento de negócios para América Latina da Fico.

Armazenamento infinito

A diminuição nos custos de armazenamento tem feito muitas empresas adotarem a prática de guardar tudo quanto é tipo de dado. Esta metodologia divide opiniões, mas um ponto é consenso: a máxima de que se entra lixo e sai lixo segue valendo.

Carlos Abdalad defende que não se deve armazenar todos os dados para se e quando um problema aparecer. “O data lake virou data swamp; jogam tanta porcaria que não tem onde achar a parte limpa. O dado tem valor quando você usa ele; quando só armazena é custo”, diz o fundador da consultoria CRMWise Analytics.

Na mesma linha, César Patiño, da JCP Consulting, aponta que os dados armazenados que não são utilizados resultam consumo de recursos de datacenter e energia para algo que não servirá para nada. “Hoje, há algumas estimativas de que dois terços dos dados são armazenados, mas não usados. Se você não usa, eles consomem espaço, é dinheiro jogado fora”, aponta.

Daniel Arraes, da Fico, assinala que, ao mesmo tempo em que o custo do armazenamento caiu muitíssimo, cresceu o volume de informações. “O custo de armazenamento ficou mais barato, então, se armazena tudo e se passa a ter mais cuidado na hora de extrair o dado que importa para utilizar os modelos de IA”, diz Arraes. “Quando eu comecei a trabalhar com dados, o uso era bem limitado. Era definir a informação, limpar a informação e guardar a informação que tinha certeza que queria usar. Hoje, o enfoque é diferente, falamos em informação não-estruturada, é um volume de informação extremamente alto no mundo”, acrescenta o diretor.

As empresas precisam ter uma estratégia bem definida, enquanto companhia, e saber a necessidade ou problema de negócio que querem resolver. Se isso está claro, boa parte do caminho está trilhado. “Você tem regras bem definidas, você não precisa armazenar todas as informações. Isso depende da estratégia da companhia. Agora, tendo o custo de armazenamento mais barato, as empresas podem se dar ao luxo de armazenar tudo, mas depende muito da estratégia da empresa. Já vi empresas que armazenavam tudo e outras que só limitavam ao escopo definido. Não vejo um certo ou errado”, pontua Ronald Rowlands, gerente de pré-vendas do SAS.

Fugindo do caos

Os especialistas concordam que apenas armazenar todas as informações e não analisá-las significa não gerar benefícios para os negócios — apenas custos. E, uma vez armazenados, os dados, para serem considerados bons, precisam ser confiáveis, íntegros (ter rastreabilidade), documentados (saber o que se tem) e protegidos. “A IA necessita de uma base de dados robusta, organizada, segura, confiável e atualizada para ser eficiente”, contextualiza Geraldo Urbaneca Ozorio Filho, diretor-sênior de engenharia de soluções na Tableau.

“Vivemos a época do ‘caos de dados’, na qual temos diversas fontes e diversas plataformas dentro das empresas. Manter a governança nesse caos de dados é hoje um dos maiores desafios das empresas. Por isso, temas como orquestração de dados estão sendo muito falados e utilizados hoje em dia. Devem existir políticas

de orquestração e governança nas empresas para garantir que o dado tenha sido preparado, limpo, organizado e certificado, conforme comentado acima", acrescenta.

Essa época do caos de dados apontada por Ozorio leva as empresas a enfrentarem muita dificuldade e terem receio do alto custo de se trabalhar todos os dados de uma vez, devido à grande quantidade de dados e à diversidade de locais onde estão armazenados. Por essa razão, ele orienta que o ideal é focar nas metas principais da empresa, definir os processos e, então, iniciar o trabalho e a governança dos dados, sempre priorizando os objetivos principais.

"A IA não pode ser eficiente em um ambiente que não tenha dados confiáveis, organizados, que tragam informações relevantes sobre seu produto, seu mercado, preços, eficiência etc. O software trabalha alimentado por essas informações. É sempre possível melhorar, tornar dados de ambientes diferentes compatíveis, acelerar a limpeza e observar o que a IA está entregando, atualizando e corrigindo. Com o tempo, aprende-se a tirar os maiores benefícios possíveis e a IA, junto com processos automatizados, passa a entregar os melhores insights e fazer análises que não seriam levantadas por analistas, que só o sistema é capaz de trazer", aponta o diretor-sênior da Tableau.

Para ele, uma boa governança implica organizar informações de modo a distinguir quais são os dados críticos da empresa, mapear os fluxos de dados sensíveis e colocar maior restrição e segurança para áreas específicas. "No meio do caos de dados das empresas, temos fontes de dados que exigirão maior complexidade de tratamento e outras fontes com menor complexidade de tratamento. Para as de alta complexidade, nas quais o dado necessita aperfeiçoamento e entendimento, deve-se utilizar equipes especializadas de dados para o correto ajuste", finaliza.

Metodologia CRISP-DM

Com 20 anos de experiência em análise de dados, Carlos Abdalad, fundador da consultoria CRMWise Analytics, recomenda usar a metodologia CRISP-DM (sigla em inglês para cross-industry standard process for data mining), um modelo de processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de

dados para atacar problemas. "É uma metodologia muito intuitiva. A primeira fase é entender o negócio e o processo de negócio que você quer melhorar com analytics. Para isso, você tem de ter uma pessoa que entenda do negócio e outra que conheça dado. Você vai ter de levantar as hipóteses que você quer testar; e esta é uma discussão boa, porque quem entende do negócio precisa ser provocado com a metodologia", explica.

O entendimento do negócio (e dos problemas que precisam ser resolvidos) é o primeiro passo e é a base de tudo. Em cima disso, vem o entendimento dos dados que são necessários e a preparação deles. Depois, desenvolve-se a modelagem, avalia-se e implanta-se. "Com o entendimento do processo de negócio, começa-se a entender se os dados que eu tenho têm valor. Normalmente, as empresas têm dados suficientes na mão", diz.

A próxima fase é a preparação dos dados, que inclui a análise de se os dados existentes têm qualidade para aplicação. Essa qualidade está ligada ao objetivo de negócio, ao problema mapeado e que se quer resolver. "É tentar transformar um problema de negócio em um problema de dados. Hoje, conseguimos resolver quase tudo usando dados", aponta Abdalad.

Com o objetivo definido de qual questão de negócio se quer resolver, parte-se para arrumar a informação para o algoritmo aprender de jeito mais fácil. "Você preparou os dados, modelou — ou seja, implementou algoritmo que responda à questão de negócio — e, assim, vai criando ciclos, confiando no processo e gerando resultados no modelo. Quando chegamos ao modelo satisfatório, avalia-se se o modelo está respondendo às perguntas e depois começa a colocá-lo na vida real", detalha. •