ManagementSolutions
Making things happen

Auto Machine Learning, rumo à automação dos modelos

P&D

www.managementsolutions.com

Design e diagramação
Departamento de Marketing e Comunicação
Management Solutions - Espanha

Fotografias
Arquivo fotográfico da Management Solutions
iStock

Índice

Introdução

"A half-dozen monkeys provided with typewriters would, in a few eternities, produce all the books in the British Museum"
– Jorge Luis Borges[1]

Um modelo matemático é, de certa forma, uma simplificação da realidade que tira proveito das informações disponíveis para sistematizar a tomada de decisões. Essa simplificação permite que hipóteses sobre o comportamento de variáveis e sistemas sejam avaliadas através de sua representação sumária sob um conjunto de postulados, geralmente com base em dados e aplicando critérios de inferência. Seu principal objetivo é explicar, analisar ou prever o comportamento de uma variável.

A revolução nas técnicas de modelagem, combinada com maior poder computacional, maior acessibilidade e maior capacidade de armazenamento de dados, mudou radicalmente a forma como os modelos são construídos nos últimos anos. Essa revolução foi um fator-chave que estimulou não apenas o uso dessas novas técnicas nos processos de tomada de decisão, onde as abordagens tradicionais eram usadas, mas também em áreas onde o uso de modelos não era tão comum. Por fim, em alguns setores, como o setor financeiro, o uso de modelos também foi impulsionado pela regulamentação. Normas como IFRS 9 e 13 ou Basileia II promoveram o uso de modelos internos com o objetivo de aumentar a sensibilidade e melhorar a sofisticação do cálculo de deterioração contábil ou determinação de riscos financeiros.

Embora possa parecer o contrário, as técnicas de modelagem mais comuns aplicadas no campo de negócios não têm uma origem recente. Especificamente, as regressões lineares e logísticas datam do século XIX. No entanto, há algum tempo, há um desenvolvimento significativo de novos algoritmos, cujo objetivo é refinar a maneira como os padrões são encontrados nos dados, mas também apresenta novos desafios, como melhorar as técnicas de interpretabilidade. A aplicação desses novos modelos matemáticos à computação é uma disciplina científica conhecida como aprendizado automático ou *machine learning,* pois permite que os sistemas aprendam e encontrem padrões sem serem explicitamente programados para isso.

Existem várias definições de *machine learning*. Entre elas, as duas mais ilustrativos são os de Samuel e Mitchell. Para Arthur Samuel[2], o *machine learning* é "o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados", enquanto para Tom Mitchell[3] é definido como "um programa que aprende com a experiência *E* com relação a alguma classe de tarefas *T* e com base em uma medida de desempenho *P*, se esse desempenho nas tarefas em *T*, de acordo com a medida de *P*, melhorar com a experiência *E*". Essas duas definições geralmente estão relacionadas a aprendizado não supervisionado e aprendizado supervisionado, respectivamente[4].

Como conseqüência, o apetite para entender e tirar conclusões dos dados aumentou dramaticamente. Mas, paralelamente, a implementação desses métodos exigiu modificações em múltiplos aspectos nas organizações[5], e é, por sua vez, o foco de possíveis riscos decorrentes de erros em seu desenvolvimento ou implementação, ou seu uso inadequado.

A modelagem avançada melhora os processos comerciais e operacionais, ou até facilita o surgimento de novos modelos de negócios. Um exemplo pode ser encontrado no setor financeiro, onde novas metodologias, no contexto da digitalização, estão modificando a proposta de valor atual, mas também adicionando novos serviços. De acordo com uma pesquisa realizada pelo Banco da Inglaterra e pela Autoridade de Conduta Financeira de quase 300 empresas do setor financeiro e de seguros, dois terços dos participantes usam o *machine learning* em seus processos[6]. As técnicas de *machine learning* são frequentemente usadas em tarefas de controle típicas, como prevenção à lavagem de dinheiro (AML), análise de ameaças relacionadas à ciberssegurança ou detecção de fraude, bem como em processos de negócios, como a classificação de clientes, sistemas de recomendação ou atendimento ao cliente através do uso de *chatbots*. Também são utilizados no gerenciamento de risco de crédito, precificação, na execução de operações ou na subscrição de seguros.

[1] Jorge Luis Borges, "La biblioteca total" (1939). Escritor, poeta, ensaísta e tradutor argentino, autor, entre outros, de "Ficciones" e "El Aleph".
[2] Samuel, 1959.
[3] Mitchell, 1997.
[4] Management Solutions, 2018.
[5] Ibídem.
[6] Bank of England, 2019.

Um grau semelhante de desenvolvimento pode ser observado em outros setores. O uso de modelos de *machine learning* é comum em setores como manufatura, transporte, medicina, justiça ou nos setores de varejo e bens de consumo. Isso fez com que o investimento em empresas dedicadas à inteligência artificial aumentasse de US $ 1,3 bilhão em 2010 para US $ 40,4 bilhões em 2018 no mundo[7] (ver figura 1). O retorno esperado justifica esse investimento: 63% das empresas que adotaram o uso de modelos de *machine learning* em suas unidades de negócios relatam um aumento na receita, sendo mais de 6% para aproximadamente metade delas. Da mesma forma, 44% das empresas relatam economia de custos, sendo mais de 10% para aproximadamente metade delas[8].

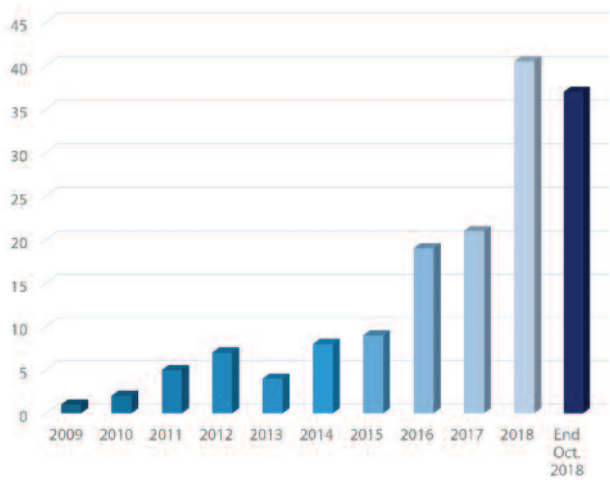
Das diferentes mudanças registradas nas organizações para se adaptar a esse novo paradigma, o recrutamento e a retenção de talentos ainda são dos elementos centrais. Em um primeiro momento, foi necessário um aumento nas equipes de especialistas em *machine learning.* A demanda por profissionais nesse campo aumentou 728% entre 2010 e 2019 nos Estados Unidos[10] (ver figura 2), também registrando uma mudança qualitativa na demanda por habilidades e conhecimentos dos cientistas de dados.

Mas essa demanda não é genérica: com a intenção de explorar a quantidade crescente de dados disponíveis por meio de ferramentas cada vez mais sofisticadas, os requisitos se tornaram mais específicos (incluindo o conhecimento de diferentes linguagens de programação, como Python , R, Scala ou Ruby, capacidade de tratamento de bancos de dados em arquiteturas de *big data,* conhecimento em computação em nuvem, conhecimento avançado em matemática e estatística, posse de cursos de pós-graduação especializados, etc.), com grande diversidade de posições, com requisitos muito específicos e, portanto, difíceis de atender. Além disso, o grande aumento no volume de geração de dados pelas empresas significa que, mesmo com um suprimento estável de cientistas de dados, a solução atual de recrutamento de recursos não é escalável.

Mas não é apenas necessário estabelecer equipes especializadas, mas também o uso de novos procedimentos de desenvolvimento, a revisão dos métodos de validação, revisão e avaliação dos modelos nas áreas de validação e auditoria, além de uma mudança cultural importante nas outras áreas para alcançar uma implementação eficaz. A inclusão desses novos processos gera uma reação em cadeia que afeta todo o ciclo de vida dos modelos, destacando entre eles a identificação e o gerenciamento de riscos do modelo, bem como sua governança[11]. Muitos desses modelos exigem adicionalmente a

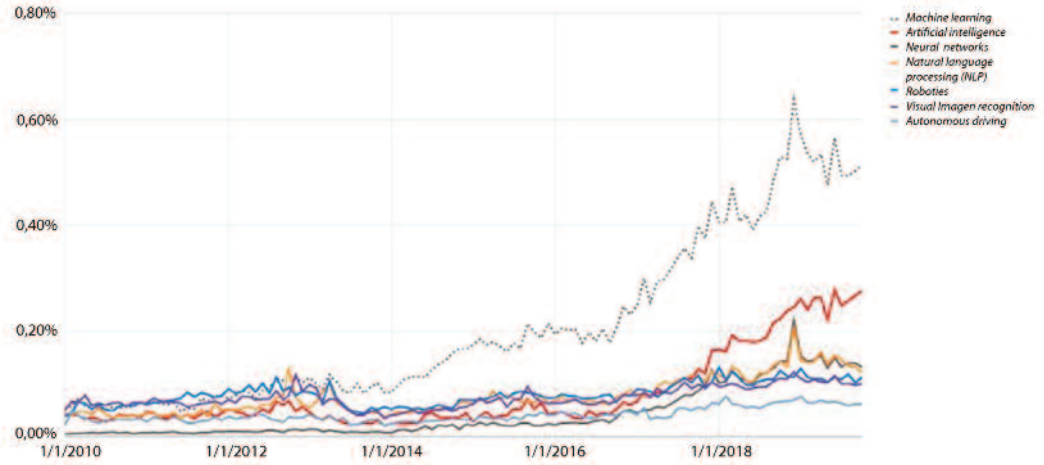
[7] Inclui unicamente investimentos de valor superior a 400.000 dólares, Stanford University, 2019.
[8] Statista, 2019.
[9] Stanford University, 2019.
[10] Ibídem.
[11] Management Solutions, 2014.

Figura 1: investimento anual (bilhões de dólares) em empresas de IA[9].



Fonte: Stanford University, 2019.

Figura 2: maior demanda por perfis com conhecimento em machine learning e inteligência artificial.



Fonte: Burning Glass, 2019.

aprovação dos órgãos de supervisão, como ocorre no setor financeiro (por exemplo, nos modelos de capital ou de provisão) ou na indústria farmacêutica, o que acrescenta desafios adicionais aos já existentes, como é necessário garantir a interpretabilidade dos modelos utilizados, bem como desenvolver os demais elementos de confiança dos modelos.

Outro aspecto notável do investimento em métodos de *machine learning* é que ele tem um desenvolvimento desigual: a obrigação de passar nos processos de validação, auditoria e aprovação, de acordo com a regulamentação estabelecida ou a exigência de manter padrões específicos de documentação, está gerando diferenças na implantação de modelos internos das empresas. De acordo com o relatório sobre *big data* e *analytics* da EBA[12], as instituições financeiras estão adotando programas de transformação digital ou promovendo o uso de técnicas de machine learning em aspectos como mitigação de riscos (incluindo pontuação automática, gerenciamento de riscos operacionais ou fraude) e nos processos de *Know Your Client*. No entanto, "embora a aplicação do *machine learning* possa representar uma oportunidade para otimizar capital, da perspectiva de uma estrutura prudencial, é prematuro considerar o uso de técnicas de *machine learning* apropriadas para determinar os requisitos de capital"[13].

[12] European Banking Authority, 2020.
[13] Ibídem.

Também existem riscos operacionais difíceis de detectar, como os de natureza humana durante o processo de implementação de um modelo ou os relacionados à segurança do armazenamento de dados, que devem ser convenientemente gerenciados para garantir o uso desses sistemas em um ambiente adequado. Isso adquire relevância significativa para empresas que operam em ambientes considerados de alto risco. Um exemplo disso é o *framework* estabelecido pela Comissão Europeia nestes casos e que engloba diferentes aspectos do processo de modelagem[14]. Por fim, e também devido a critérios regulatórios e de gestão, os modelos devem funcionar de forma confiável e devem ser usados de forma ética, para que o usuário possa confiar neles para uso nos processos de tomada de decisão. Nesta linha, é de especial interesse a proposta da EBA baseada em sete pilares de confiança[15]: ética, interpretabilidade, eliminação da discriminação, rastreabilidade, proteção e qualidade dos dados, segurança e proteção do consumidor. Essas questões foram identificadas como elementos-chave em universidades como também por empresas[16].

Nesse contexto, as tarefas de desenvolvimento de modelos exigem tempos muito desiguais: as tarefas anteriores e complementares à análise também exigem uma grande quantidade de tempo e recursos destinados à preparação, limpeza e tratamento geral dos dados; 60% do tempo de um cientista de dados é gasto limpando dados e organizando informações, enquanto 9% e 4% se concentram em tarefas de descoberta de conhecimento e refinamento de algoritmos, respectivamente[17]. Tudo isso leva à necessidade de mudar a maneira de abordar o desenvolvimento, a validação e a implementação de modelos, para que sejam exploradas as vantagens de novas técnicas, mas resolvendo as dificuldades associadas ao seu uso, além de mitigar seus possíveis riscos.

Em decorrência das razões acima mencionadas, há uma clara tendência em direção à automação de processos relacionados à aplicação de técnicas avançadas de análise, que tem sido geralmente chamada de aprendizado automático de máquina (AutoML ou automated *machine learning*, de forma intercambiável), cujo objetivo não é apenas automatizar as tarefas em que os processos heurísticos são limitados e facilmente automatizáveis, mas também permitir a geração de processos e algoritmos de pesquisa de padrões mais automáticos, ordenados e rastreáveis. De acordo o Gartner[18], mais de 50% das tarefas de ciência de dados serão automatizadas até 2025.

Essa tendência para a automação é explicada não apenas pelas questões levantadas acima, mas também pelas oportunidades oferecidas pela arquitetura dos sistemas utilizados, em termos de design de fluxo de trabalho, inventário de modelo ou validação de componentes. Os sistemas de AutoML integram várias ferramentas para desenvolver modelos, reduzindo também custos, tempo de desenvolvimento e erros na implementação de tais sistemas.

[14] European Comission, 2020.
[15] European Banking Authority, 2020.
[16] Por exemplo, a cadeira iDanae, resultado da colaboração entre a Universidade Politécnica de Madri e a Management Solutions, publicou boletins informativos sobre interpretabilidade (cadeira iDanae, 3T-2019) e ética em inteligência artificial (cadeira iDanae, 4T-2019).
[17] De acordo com uma pesquisa realizada pela plataforma em Inteligência Artificial CrowdFlower (CrowdFlower, 2017).
[18] Gartner, 2019.

Os sistemas e métodos de AutoML buscam, entre outras coisas:

- Reduzir o tempo gasto pelos cientistas de dados no desenvolvimento de modelos por meio de técnicas de *machine learning* e até mesmo permitir o desenvolvimento de algoritmos de *machine learning* e por equipes não especializadas em ciência de dados.
- Melhorar o desempenho dos modelos desenvolvidos, bem como a rastreabilidade e comparabilidade dos modelos obtidos com as técnicas de busca manual por hiperparâmetros.
- Permitir questionar os modelos desenvolvidos por outras abordagens.
- Reutilizar o investimento feito em tempo e recursos para desenvolvimento de códigos, melhorar e refinar os componentes incluídos nos sistemas de forma eficiente e com maior rastreabilidade.
- Simplificar a validação dos modelos e facilitar seu planejamento.

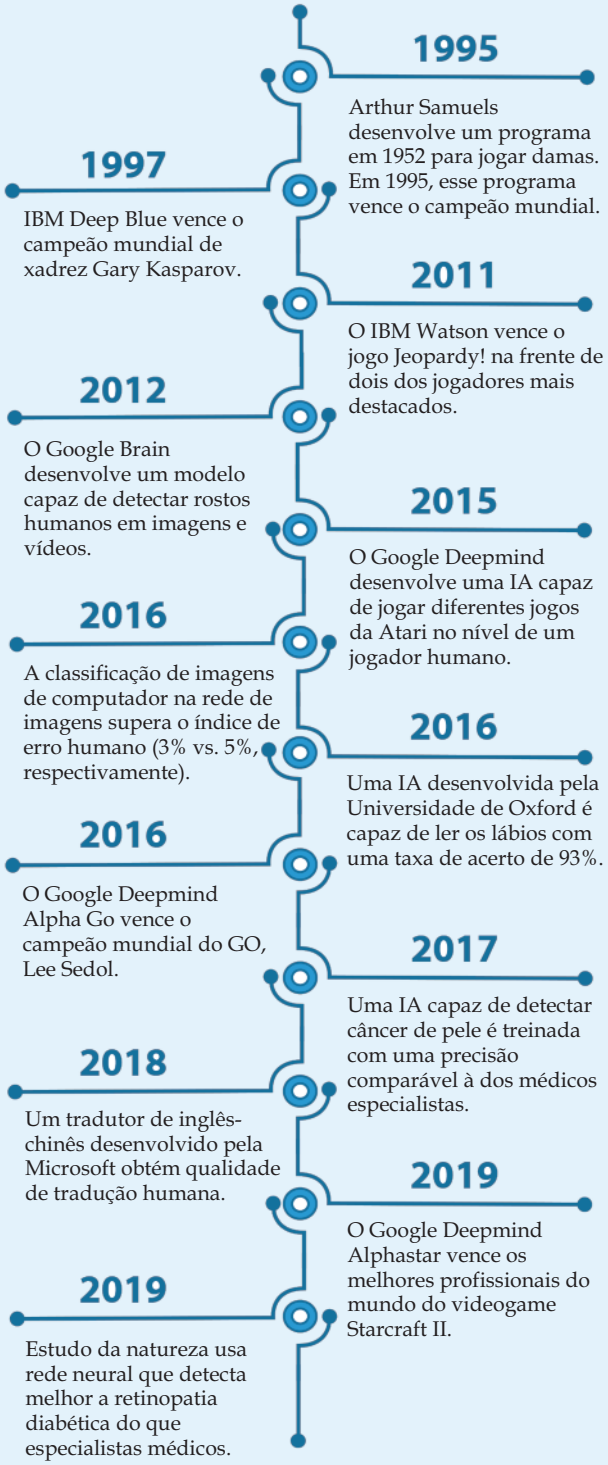
Neste contexto, este documento tem como objetivo descrever os principais elementos sobre os sistemas de AutoML. Para isso, foi estruturado em três seções, que por sua vez correspondem a três objetivos:

- No primeiro bloco, a evolução na automação dos processos de machine learning é analisada, assim como os motivos subjacentes no desenvolvimento de sistemas de AutoML.

- O segundo bloco fornece uma visão descritiva das principais estruturas do AutoML e explica quais abordagens estão sendo seguidas, tanto academicamente quanto em experiências práticas destinadas a automatizar processos de modelagem por meio de técnicas de *machine learning*.

- Por fim, o terceiro bloco tem como objetivo ilustrar os resultados do desenvolvimento de sistemas de AutoML, apresentando como estudo de caso um campeonato organizado pela Management Solutions no início de 2020, dirigido aos profissionais da firma e cujo objetivo foi o desenho de um modelo de *Automated Machine Learning*.

Principais marcos no desenvolvimento do ML



Reporting sobre *Big Data* e *Advanced Analytics* da EBA

A Autoridade Bancária Europeia (EBA) publicou um relatório sobre *big data* e *advanced analytics,* com o objetivo de divulgar seu uso no setor financeiro europeu, além de fornecer seu entendimento sobre (i) a identificação de quatro pilares fundamentais para seu desenvolvimento, implementação e adoção, (ii) os principais elementos de confiança nos quais uma estrutura de *big data* e de *advanced analytics* deve se basear; e (iii) apontar as principais observações, oportunidades e riscos decorrentes da aplicação dessas soluções.

I. Pilares principais de uma estrutura de *big data* e de *advanced analytics*

II. Elementos de confiança

III. Principais observações, oportunidades e riscos no uso

Pilares principais de uma estrutura de *big data* e *advanced analytics*

GESTÃO DE DADOS

- Tipos de dados e fontes de dados: dados estruturados, não estruturados e semiestruturados.
- Segurança e proteção dos dados: protege a confidencialidade, integridade e disponibilidade dos dados.
- Qualidade dos dados: assegura a qualidade dos dados durante todo o ciclo de vida da *Big Data* e de *Advanced Analytics* (BD&AA).

INFRAESTRUTURA TECNOLÓGICA

- Infraestrutura: conecta recursos para transmissão *Big Data*
- Plataforma de dados: gerencia todos os dados usados por um sistema de análise avançado
- Processamento: fornece o software necessário para suportar a implementação de aplicativos analíticos avançados

METODOLOGIA DE ANÁLISE

- Preparação dos dados: inclui a engenharia de dados e a análise de relevância
- Analítica: treinamento de modelos, calibração, validação e seleção, assim como testes e implantação
- Operações: controle e atualização de modelos

ORGANIZAÇÃO E GOVERNANÇA

- Estrutura e medidas internas de governança: estrutura de governança, estratégia e gerenciamento de riscos, transparência, desenvolvimento externo e terceirização.
- Habilidades e conhecimentos: nível de entendimento do órgão de administração e da gerência sênior, segunda e terceira linhas, desenvolvedores de sistemas avançados para análise.

Elementos de confiança

Uma parte fundamental para a construção de modelos confiáveis é que sejam explicáveis e interpretáveis, transparentes, compreendidos corretamente e que apresentem justificativas claras.

O uso de soluções rastreáveis auxilia no rastreamento de todas as etapas, critérios e decisões ao longo do processo, o que permite a réplica dos processos que resultam nas decisões tomadas pelo modelo e ajuda a garantir a auditabilidade do sistema.

É importante manter uma vigilância técnica sobre os ataques de segurança mais recentes e técnicas de defesa relacionadas e garantir que a governança, a supervisão e a infraestrutura técnica estejam em vigor para um gerenciamento eficaz dos riscos de ICT *(Information and communications technology)*.

O desenvolvimento, implantação e uso de qualquer solução de Inteligência Artificial (IA) deve aderir a alguns princípios éticos fundamentais.

A justiça exige que o modelo garanta a proteção dos grupos contra a discriminação e, para garantir isso, o modelo deve estar livre de vieses.

Os dados devem ser adequadamente protegidos com um sistema confiável de BD&AA e sua qualidade precisa ser levada em consideração durante todo o processo do BD&AA.

Um sistema confiável de BD&AA deve respeitar os direitos dos consumidores e proteger seus interesses.

Principais observações, oportunidades e riscos no uso

Observações chave

- As instituições estão em diferentes estágios do desenvolvimento da BD&AA. Alguns dos usos mais comuns são a detecção de fraudes, CRM e automação de processos.
- Existe uma maior dependência de dados internos, em vez de dados externos ou mídias sociais. Incorporação de soluções de código aberto. Existe um uso limitado de algoritmos complexos.
- Existem níveis diferentes de integração e governança de análises avançadas na instituição.
- Pode-se observar um aumento da dependência em empresas de tecnologia para a provisão de serviços de infraestrutura e de computação na nuvem.

Principais oportunidades

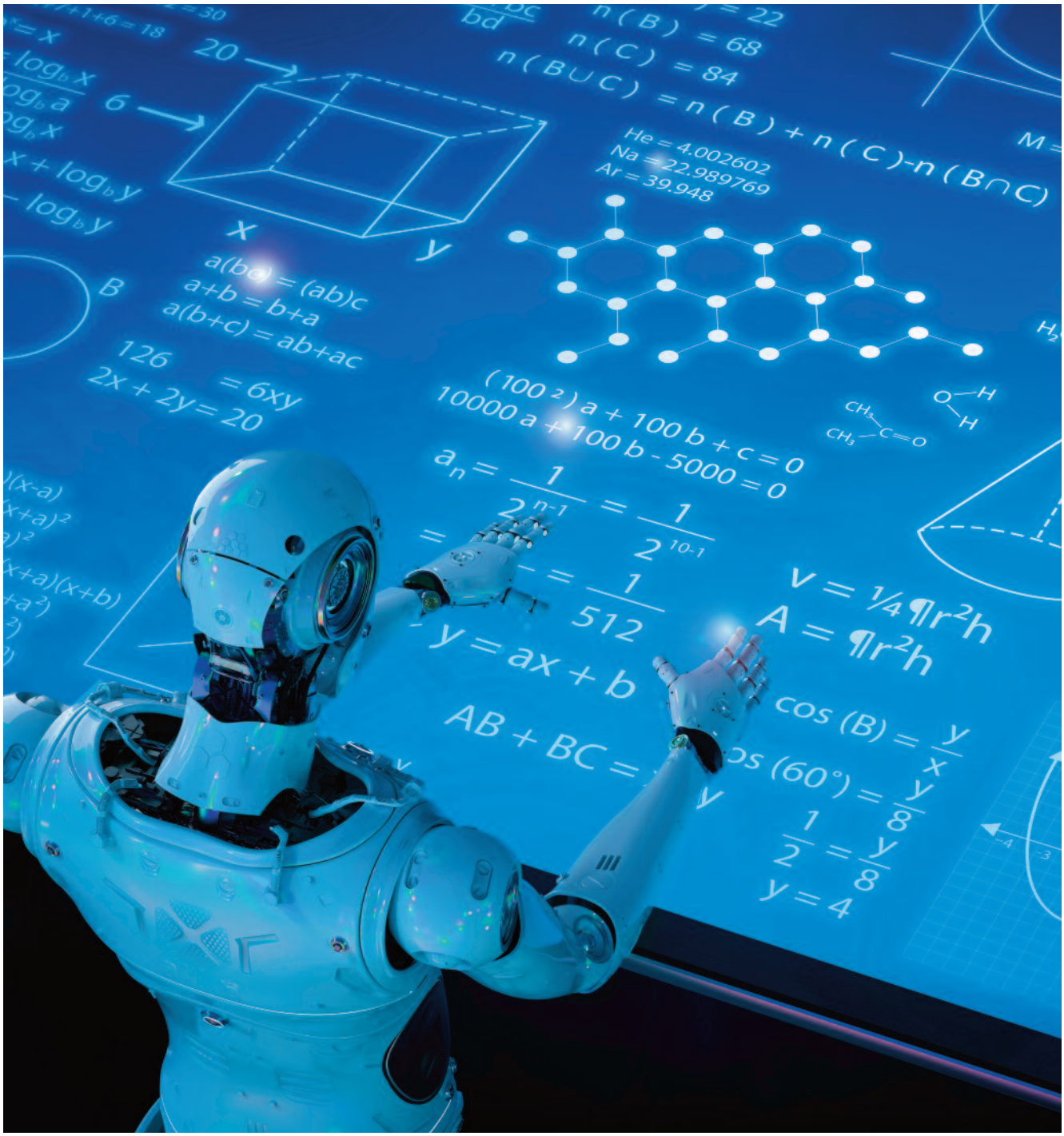
- Os clientes dos serviços financeiros dos setores de varejo e lazer esperam um serviço mais personalizado. Existe confiança no setor financeiro em relação ao cumprimento das leis de proteção de dados.
- O aumento da satisfação do consumidor e o uso de informação para melhorar a oferta, reduzir a perda de clientes, otimizar processos e ajudar na mitigação do risco e da detecção de fraude.
- Existem muitos usos e oportunidades possíveis que surgem do uso de modelos interpretáveis.

Principais riscos e orientação proposta

- O resultado dos modelos pode ser complexo, não determinístico e correto de acordo com uma medida de probabilidade, que pode prejudicar a instituição ou seus clientes. Deve-se garantir que os resultados desses sistemas não violem os padrões éticos das instituições. Além disso, um ser humano deve estar envolvido no ciclo de tomada de decisão, por isso é necessário realizar treinamento do pessoal.
- A implementação de um *framework* de governança e metodologia de BD&AA poderia promover o seu uso responsável, que deveria incluir documentação adequada, uma justificação suficiente e outras técnicas explicativas e de monitoramento, incluindo o uso de soluções rastreáveis. A explicação deve ser baseada em uma abordagem baseada no risco.
- Há a necessidade de modelos precisos e verificações regulares.
- O uso de soluções de *machine learning* pode levar a riscos de ICT: segurança de dados, segurança de modelos, qualidade de dados, gestão de mudanças e continuidade e resiliência do negócio.
- Como resultado da dependência em *frameworks* de código aberto, ou em ferramentas e sistemas desenvolvidos por terceiros, tanto seus riscos potenciais (falta de controle e conhecimento do terceiro, alta dependência de um fornecedor, risco de concentração, manutenção de um modelo, etc.) como a responsabilidade que deve ser sempre mantida na instituição, deve ser avaliado.
- Por fim, é enfatizada a importância da qualidade, proteção e segurança dos dados, tanto para propósitos regulatórios (incluindo o cumprimento de leis de proteção de dados) e para assegurar a adequação do modelo.

Resumo executivo

"Strictly speaking, one immortal monkey would suffice"
Jorge Luis Borges



O contexto da automação dos modelos de machine learning

1. A incorporação de técnicas de *big data* e de *advanced analytics* na economia está mudando a maneira como as informações são usadas. Com base na combinação de diferentes conhecimentos relacionados à exploração de dados e negócios, a capacidade de análise aumentou radicalmente, embora ao mesmo tempo seja o foco de possíveis riscos derivados de erros no seu desenvolvimento ou implementação, uso inadequado ou confiança excessiva em seu aplicativo.

2. Para aproveitar o potencial dessas novas técnicas, as instituições estão transformando sua maneira de trabalhar. Essas mudanças afetam diretamente o desenvolvimento e a validação dos modelos, mas também outros processos, como os relacionados a estruturas tecnológicas, a seleção, o treinamento e a retenção de perfis especializados ou, de maneira mais ampla, a cultura do trabalho.

3. Também existem riscos operacionais difíceis de detectar e, em alguns casos, regulamentos relacionados ao uso, qualidade e processos relacionados aos dados, bem como a aplicação ou interpretabilidade de modelos que geram a necessidade de processos rastreáveis, validações e análises recorrentes dos modelos.

4. Nesse contexto, há uma clara tendência em direção à automação de processos relacionados à aplicação de técnicas avançadas de análise, cujo objetivo não é apenas automatizar as tarefas em que os processos heurísticos são limitados e facilmente automatizáveis, mas também permitir uma geração mais automática, ordenada e rastreável de modelos.

5. Com tudo isso, é favorecida a redução do tempo dedicado a tarefas complementares e repetitivas; acesso a essas técnicas por equipes não especializadas; o desempenho, rastreabilidade e comparabilidade dos modelos; reutilização de desenvolvimentos de código em projetos específicos, aprimoramento e aprimoramento de técnicas; e até a melhoria dos processos de validação, incluindo a geração de modelos *challenger*.

Rumo à automação de modelagem

6. No processo de automação do fluxo de trabalho *(workflow)* de modelagem, existem vários desafios. Entre eles, a necessidade de ter processos que garantam a adequação da carga de dados e aqueles relacionados ao desenvolvimento, validação, implementação e monitoramento dos modelos. A rastreabilidade dos processos de construção também deve ser garantida, assim como sua interpretabilidade e governança adequada, que permitam sua integração na gestão, além de cumprir as regulamentações existentes.

7. Com relação ao processo de tratamento de dados, 60% do tempo de um cientista de dados é dedicado à limpeza e organização das informações, com um longo caminho a percorrer em termos de automação desses processos.

8. No fluxo de trabalho de modelagem, é possível automatizar esse processo por meio de duas opções, que geralmente são combinadas: a componentização dos diferentes processos em elementos segregados e a execução automática desses componentes, automatizando-os por meio de regras pré-estabelecidas e técnicas estatísticas.

9. A componentização é baseada na separação das tarefas de modelagem em diferentes partes, e sua programação e seu desenvolvimento de forma independente. Cada um desses componentes recebe uma determinada entrada e executa uma tarefa específica.

10. As vantagens da componentização são a padronização de processos, aumentando a qualidade e a eficiência, especializando-se em desenvolvimento, melhorando a usabilidade e promovendo a escalabilidade.

11. Por outro lado, a automação do processo de construção do modelo é baseada no uso de critérios automáticos para selecionar seus atributos, para que o procedimento possa ser replicado e auditado. Ele também garante que a seleção final tenha sido feita através de um processo que garante o poder preditivo ideal, dadas as restrições.

12. As vantagens da automação de pesquisa são a otimização do processo de seleção de hiperparâmetros, a generalização de problemas de modelagem, a adaptação dos espaços de pesquisa de parâmetros para cada problema e a possibilidade de experimentação fora dos limites habituais.

Estruturas de automação para processos de machine learning

13. Na prática, a maneira de automatizar esses processos foi baseada em (i) automatizar a maioria dos aspectos relacionados à análise e tratamento prévio dos dados, incluindo a transformação de variáveis e sua pré-seleção, (ii) gerar um espaço de busca para possíveis modelos e parâmetros, bem como um processo de desenvolvimento e seleção de modelos que evite tanto o *overfitting*[19] quanto o *underfitting*[20] e (iii) automatizar a aplicação de técnicas de interpretabilidade.

14. Há uma grande variedade de opções para colocar esses sistemas em produção, coletados em abordagens baseadas em modelo *(model-based schemes)* e em abordagens orientadas a dados *(data-driven approaches)*.

15. Os processos para gerar e avaliar modelos são fundamentalmente baseados em dois componentes: o otimizador e o avaliador.

16. O otimizador gera e atualiza combinações de parâmetros dentro do limite de possibilidades definidas de acordo com o modelo e os dados utilizados. Posteriormente, o avaliador é responsável por medir o desempenho das opções propostas pelo otimizador e pode influenciar a estratégia de busca com base nos resultados.

17. O otimizador usa várias técnicas para encontrar a melhor configuração. Essas técnicas podem ser classificadas como simples (como *grid search, random search,* algoritmos evolutivos ou otimização bayesiana) ou com base na experiência (como *meta-learning* ou *transfer learning*).

18. O avaliador é responsável por verificar se a configuração fornecida pelo otimizador é ideal. Também existem diferentes abordagens de otimização, como (i) *early stop,* na qual o avaliador para de avaliar se o desempenho é muito baixo nas primeiras iterações, (ii) reutilização, com base no uso das configurações usadas no treinamento anterior ou (iii) o uso de modelos substitutos na avaliação.

19. Nesse tipo de sistema, os desafios existentes são a inclusão de conhecimento anterior, como conhecimento de negócios ou critérios de especialistas, bem como o desenvolvimento de construção de sistemas que cubram todo o processo de construção de modelo.

20. Uma alternativa ao uso de um sistema baseado na interação de um otimizador e um avaliador é a busca de arquiteturas neurais (NAS). Essa técnica, utilizada na modelagem de linguagem ou classificação de imagens, executa simultaneamente as três tarefas necessárias para a automação: a determinação do espaço de busca, a estratégia de busca nesse espaço e a estimativa dos modelos obtidos em cada estimativa.

21. Nesta abordagem, embora o desempenho seja alto, é mais difícil explicar por que determinadas configurações são atingidas e se elas servem para estender seu uso a outros tipos de problemas.

22. Atualmente, e apesar de ainda haver muito espaço para melhorias, os sistemas de AutoML alcançaram um estágio de desenvolvimento que pode competir e derrotar especialistas humanos em *machine learning*, configurando-se como uma ferramenta fundamental que pode modificar o tipo de trabalhos desenvolvidos.

[19] Característica de um modelo que ocorre quando ele é ajustado muito próximo da amostra de treinamento, para que ele não atinja resultados satisfatórios em amostras diferentes desta.

[20] Característica de um modelo que ocorre quando não foi ajustado suficientemente para a amostra de treinamento, para que ele não atinja resultados satisfatórios em amostras diferentes desta.

23. Os desafios pendentes, tanto para sistemas baseados na otimização de hiperparâmetros quanto para métodos NAS, referem-se a questões relacionadas à interpretabilidade, reprodutibilidade, bem como à reutilização nas configurações de exercícios anteriores e para facilitar uma melhor interação com o usuário.

Campeonatos de AutoML: Uma ferramenta de exploração de abordagem de AutoML

24. Para aprofundar o entendimento e a implementação das abordagens de AutoML, foram criados e realizados campeonatos que confrontam metodologias. Um exemplo destes campeonatos é discutido no capítulo "Campeonatos de AutoML: uma ferramenta de exploração de enfoques de AutoML".

25. No caso do campeonato de AutoML realizado pela Management Solutions, os participantes adotaram abordagens semelhantes às discutidas acima, usando abordagens de *grid, random search,* algoritmos genéticos ou pesquisas bayesianas para realizar a geração de modelos. Técnicas de cross validation foram usadas para avaliar a configuração.

26. A partir deste exercício, algumas conclusões úteis foram tiradas: i) o processamento de dados foi bastante homogêneo, mostrando que existem várias técnicas padronizadas no setor, ii) a redução da dimensionalidade foi realizada pela maioria dos Os participantes, devido à grande redução no custo computacional que isso implica, iii) os sistemas de AutoML melhoram, otimizando todo o pipeline, usando modelos de *stacking* ou tarefas paralelas em vários núcleos.

Reflexões Finais

27. Atualmente, a configuração dos modelos de *machine learning* depende significativamente de ajustes a priori e manuais, o que pode levar a um nível abaixo do ideal, como consequência tanto do *overfitting* quanto do *underfitting*, dependendo do tamanho do *dataset* e das técnicas utilizadas. Em algumas técnicas, o *overfitting* de modelos ainda é comum, o que parece indicar que há espaço para melhorias na geração de sistemas de AutoML em casos específicos.

28. Embora as abordagens de AutoML tenham atingido um alto nível de desenvolvimento, ainda existem limitações relacionadas ao fato de o *pipeline* não ser totalmente automatizado ou à ausência ou escassez de objetificação em algumas das decisões e no espaço de busca.

29. Outro desafio é permitir que perfis de não especialistas acessem os ambientes de AutoML, para que eles possam interagir diretamente com esses métodos e sistemas, para que a intuição do negócio possa ser incorporada ou a interpretabilidade dos modelos possa ser avaliada diretamente. Finalmente, e no que diz respeito à interpretabilidade, essa continua sendo uma das questões em aberto nos sistemas de AutoML.

30. Em termos de avanços recentes, estes são mais comuns na otimização da *feature engineering,* bem como na seleção de modelos, em detrimento do processamento ou preparação dos dados.

31. Por fim, espera-se que os sistemas de AutoML sejam configurados como uma ferramenta fundamental, que modifique o trabalho realizado pelos *data scientists,* para que eles se concentrem nas análises anteriores ou subsequentes do desenvolvimento de modelos, na geração de componentes e sistemas de AutoML, bem como na resolução de problemas específicos em que um sistema de AutoML não alcança bons resultados.

Rumo à automação da modelagem

"And if you play it for a hundred years, or a thousand years or a hundred thousand, the law of chances tells us that a poem will probably come out. And if you play it forever, every possible poem and every possible story will have to come out"

Michael Ende[21]

O desenvolvimento e implementação no gerenciamento de modelos de *machine learning* gera um conjunto de benefícios, que são consequência tanto da melhoria dos processos de tomada de decisão quanto da automação de tarefas no desenvolvimento de modelos. Esses benefícios se materializam, por exemplo, através de uma previsão mais precisa da demanda, em melhorias no gerenciamento de estoque, em estratégias de pricing, em aumento da fidelidade do cliente ou em melhorias na eficiência e na redução de custos de produção, entre outros. Isso, por sua vez, implica melhores resultados no desenvolvimento de produtos ou na prestação de serviços, em uma distribuição mais eficiente de recursos ou em um melhor posicionamento no mercado, podendo gerar vantagens competitivas sobre os concorrentes que utilizam menos essas técnicas.

No entanto, no processo de construção de modelos, também existem vários desafios relacionados ao desenvolvimento e implementação desses novos métodos:

- Por um lado, em muitas ocasiões, os modelos de *machine learning* exigem grandes quantidades de dados para evitar o *overfitting*, o que implica a necessidade de investir na obtenção, ingestão, armazenamento e gerenciamento de fontes de dados e arquiteturas tecnológicas, com soluções *in-house* ou *cloud* para garantir a disponibilidade e a qualidade dos dados utilizados.

- Por outro lado, é necessário investir no desenvolvimento dos modelos, sua validação, implementação nos processos de gestão e monitoramento e manutenção dos algoritmos .

- Da mesma forma, a rastreabilidade dos processos de construção deve ser considerada, além de garantir a interpretabilidade dos algoritmos e os resultados obtidos, uma vez que decisões baseadas, ainda que parcialmente, em algoritmos devem ser apoiadas por esse conhecimento.

- Todas as opções acima requerem governança adequada, o que garante a consideração de elementos gerenciais e éticos no uso de modelos e requisitos regulatórios. Esses impactos são ainda maiores para entidades que atuam em setores regulados, uma vez que existem limitações na implementação e no uso desses modelos para determinados fins.

- Por fim, para garantir o cumprimento do exposto, é necessário ter perfis especializados, seja através da contratação direta de *data scientists* e desenvolvedores de dados, seja terceirizando o processo com empresas especializadas, além de transformar a estrutura organizacional e adaptá-la de acordo com necessidades de desenvolvimento de modelos, incluindo novas formas de trabalhar (por exemplo, através de organizações *Agile*[22]).

Esses desafios motivaram o surgimento e o desenvolvimento de sistemas de AutoML, pois seu uso pode responder às dificuldades colocadas, tirar proveito dos benefícios da automação e contribuir para a democratização dos processos de modelagem, facilitando o uso por usuários não especialistas.

Noções básicas de automação

O princípio *do no free lunch*[23] sugere que não existe um modelo cujo desempenho seja sempre melhor que todos os outros, de modo que, dependendo do conjunto de dados analisados, o tipo de modelo que melhor os prevê ou explica pode ser diferente. Estendendo essa ideia ao campo do *machine learning*, esse princípio pode ser interpretado como a inexistência de estimadores ou combinações de configurações, hiperparâmetros ou arquiteturas de rede sempre melhores do que outras alternativas. Embora existam estudos acadêmicos ao lidar com problemas específicos que mostram que a seleção de

[21] Michael Ende, "A história sem fim" (1979). Escritor alemão do século XX, mais conhecido por suas obras de ficção infantis.
[22] Os detalhes da transformação para organizações agile foram amplamente descritos na publicação "De projetos Agile, a organizações Agile", Management Solutions, 2019.
[23] Wolpert & Macready, 1997.

valores para hiperparâmetros em faixas reduzidas pode garantir modelos ótimos[24], essa ideia não pode ser extrapolada para todos os problemas possíveis. Essa situação leva à necessidade de encontrar métodos para garantir que um algoritmo de *machine learning* execute uma pesquisa adequada por possíveis configurações para maximizar seu desempenho.

Dado um problema que você deseja resolver usando técnicas de *machine learning*, a maneira de resolvê-lo é baseada no estabelecimento das diferentes opções de parâmetros e configurações que podem ser escolhidas ao longo do processo. Para isso, primeiro é necessário identificar as características dos dados a serem processados, bem como as técnicas utilizadas.

Posteriormente, deve ser estabelecida a abordagem de modelagem a ser usada, bem como as métricas com as quais os modelos serão selecionados e, finalmente, as restrições que possam existir com base no conhecimento do problema são incorporadas (por exemplo, o sinal de variáveis). Dessa forma, é configurado um *workflow* de modelagem que servirá para obter um grupo de modelos ordenados de acordo com seu desempenho. Nesse fluxo, é possível realizar a automação com base na componentização, ou seja, a separação dos diferentes processos de construção do modelo em componentes que podem ser realizados de maneira modular.

Componentização e otimização do workflow de modelagem

No processo de modelagem, há um amplo conjunto de opções em cada uma das fases que compõem o workflow de desenvolvimento, resultado da combinação das diferentes técnicas usadas em cada seção. Embora a seleção dos diferentes parâmetros e a configuração das técnicas (quais aplicar, em que ordem, em que parte do *dataset*, etc.) variem dependendo do problema, é possível desenvolver técnicas para obter a automação de várias tarefas. Entre os principais objetivos alvos por essa automação estão a redução de custos e possíveis erros operacionais derivados do desenvolvimento end-to-end para cada problema de *machine learning*, além de melhorar a eficiência do processo de modelagem.

Três das causas que geram a necessidade de investir em automação são as seguintes:

- **Redundância de desenvolvimento:** algumas tarefas e funções programáveis para gerar o processo de modelagem podem ter sido desenvolvidas em outros processos anteriores, internamente por equipes especializadas da empresa ou pela comunidade *data science*.

- **Existência de erros:** o desenvolvimento de um novo código pode acarretar uma maior probabilidade de conter erros; portanto, é necessário realizar testes e processos de testes, o que implica maior esforço de tempo e recursos.

- **Necessidade de procurar com eficiência estratégias** que excluam explicitamente combinações de configurações e intervalos de hiperparâmetros considerados inadequados ou que possam levar a erros de implementação[25].

[24] Ver, por ejemplo, Segal, 2004.

[25] No entanto, embora a generalização do problema ajude a obter uma resolução eficiente, em certos casos é necessário estabelecer mecanismos para que o modelador possa testar determinadas configurações ou definições de hiperparâmetros fora do espaço de busca.

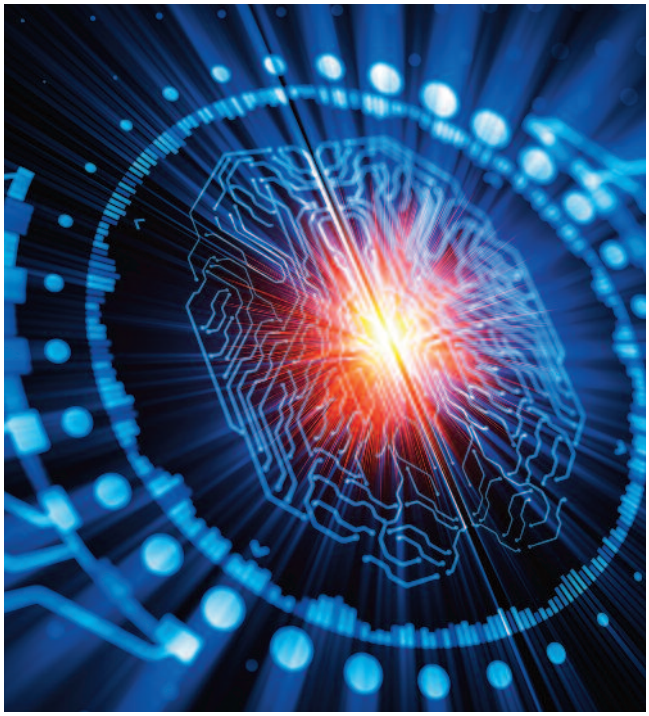
Duas das maneiras de aumentar a automação do processo de modelagem são a componentização das tarefas e a automação da busca de configurações e hiperparâmetros ideais.

Componentização do workflow

Em primeiro lugar, a segregação das tarefas de modelagem em diferentes partes, e sua programação e desenvolvimento independentemente, permitem que o modelador use automaticamente cada parte, na forma de chamadas para o código desenvolvido, adaptando apenas os parâmetros e configurações, dentro das opções possíveis, para resolver uma tarefa específica. Esse tratamento, que é análogo ao aplicado no desenvolvimento de bibliotecas em ambientes de programação ou à programação orientada a objetos, permite que a tarefa de desenvolver o código, bem como a linguagem de programação específica, seja isolada de sua aplicação subsequente. Isso permite um ambiente de modelagem ágil. Cada um desses componentes recebe um *input* específico (geralmente um *dataset* e um conjunto de parâmetros) e executa uma tarefa específica, retornando como saída outro *dataset* com o resultado da tarefa aplicada.

A componentização dos processos de modelagem gera um conjunto de vantagens sobre o desenvolvimento de um *workflow* para cada problema, como as seguinte:

- **Padronização:** o desenvolvimento de componentes utilizados na modelagem, o que reduz a frequência de erros e melhora a comparabilidade.
- **Melhoria da qualidade:** no desenvolvimento dos componentes, bem como na sua aplicação.



Elementos de um sistema automatizado de *machine learning*

- **Parâmetro:** propriedade interna do modelo aprendido durante o processo de aprendizado, sendo necessário para fazer previsões.
- **Hiperparâmetro:** parâmetro que não pode ser obtido durante o processo e deve ser definido com antecedência. Os valores que os hiperparâmetros devem adotar para resolver um problema específico são desconhecidos. O número de árvores em uma *Random Forest* ou o número de *clusters* em um *K-Means* são exemplos de hiperparâmetros.
- **Configuração:** as possíveis combinações de valores que os hiperparâmetros podem assumir.
- **Espaço de configuração / pesquisa:** conjunto de todas as configurações possíveis de hiperparâmetros nas quais a configuração ideal é pesquisada para obter a melhor previsão possível.
- **Arquitetura de rede neural:** refere-se em conjunto ao número de camadas e neurônios presentes em cada uma delas, bem como à maneira como elas estão conectadas. Em alguns casos, a maneira como eles são treinados também está incluída no conceito.
- **Função de custo:** função cujos mínimos correspondem às configurações ideais. Procurar a configuração ideal é equivalente a encontrar os mínimos da função de custo. Algumas funções de custo podem ser o erro quadrático médio ou a entropia cruzada, entre outras opções.

- **Maior eficiência:** na aplicação dos componentes, na revisão pelas áreas internas de validação e auditoria, bem como nos processos de aprovação.

- **Especialização:** no desenvolvimento de cada componente por especialistas em cada disciplina.

- **Usabilidade aprimorada:** em seu aplicativo, pois esses pacotes podem ser usados por diferentes tipos de usuários, incluindo aqueles que não possuem conhecimento de programação.

- **Escalabilidade:** em desenvolvimento, que pode ser baseada em um host interno ou em cloud, tanto para cada subsidiária da empresa quanto por áreas geográficas.

Automatizando a busca por configurações ideais

Depois que as diferentes etapas do processo foram separadas em componentes, uma seleção deve ser feita dos melhores parâmetros que configuram o processo ideal para realizar a modelagem. Uma abordagem que permite abordar essa seleção é a automação de sua pesquisa, definindo diferentes estratégias que abordam esse problema de um ponto de vista sistemático e ordenado e deixando um rastro do processo que gera cada combinação possível, mas que, por sua vez, permite avaliar o impacto das decisões tomadas em cada etapa do processo no desempenho do modelo final.

No entanto, durante o processo de automação, a avaliação de todo o conjunto de opções é frequentemente muito complexa e geralmente é condicionada por um tempo de computação limitado, tanto pelo número possível de opções e combinações, quanto à complexidade do modelo e pela quantidade de dados analisados. Apesar do exposto, a automação na busca de configurações, embora limitada, gera um conjunto de vantagens sobre a busca manual, como:

- **Otimização de busca:** pois permite gerar um conjunto de combinações que serão avaliadas e selecionar as que geram o melhor desempenho.

- **Generalização dos problemas:** pois permite gerar amplos espaços de pesquisa quando não há informações anteriores que permitam antecipar quais subespaços de pesquisa devem ter maior probabilidade de gerar modelos com desempenho superior.

- **Adaptação do espaço de busca:** para os problemas em que há alguma informação sobre qual deve ser o espaço ideal de pesquisa; é fácil adaptar isso para melhorar os resultados com base em restrições computacionais.

- **Experimentação:** pois permite avaliar o impacto das microdecisões em cada um dos componentes incluídos na automação no desempenho final. Por exemplo, permite avaliar a alteração em vários parâmetros do modelo, como a profundidade máxima das árvores em um algoritmo de *random forest*.

Nesse sentido, um sistema de AutoML pode ser definido como um método que permite a construção de modelos de *machine learning* sem a necessidade de intervenção humana e sujeito a certas limitações computacionais[26]. O papel da pessoa que

[26] Yao, e outros, 2018.

desenvolve um modelo em um sistema de AutoML se concentra na escolha dos dados, na seleção dos critérios de validação para os dados e na escolha das métricas a serem usadas para estimar e selecionar os modelos, em vez de gastar tempo no processamento de dados e na otimização do hiperparâmetro iterativamente com base nos resultados do modelo. Tudo isso determina o espaço de busca, para que seja gerado um conjunto de opções que o algoritmo avalia, sujeito às condições estabelecidas. Finalmente, como resultado desse processo, é obtido um conjunto de modelos ordenados de acordo com seu desempenho.

A seção a seguir abordará a abordagem de AutoML e seu impacto na solução dos desafios descritos nesta seção e na transformação que ela está assumindo na maneira de modelar e, em geral, de todo o workflow de um processo de *machine learning*.

Workflow de modelagem

A definição de um workflow depende do problema a ser resolvido e do tipo e qualidade dos dados utilizados. Existem diferentes metodologias para o desenvolvimento de projetos de *machine learning*, como KDD *(Knowledge Discovery in Databases)*, CRISP-DM *(Cross-Reference Industry Standard Process for Data Mining)* ou SEMMA *(Sample, Explore, Modify, Model and Assess)*, entre outras. Embora existam diferenças entre eles, existem elementos comuns em todos eles, que são os eixos fundamentais para a construção de modelos de *machine learning*. O processo de modelagem está resumido abaixo:

Identificação do problema e planejamento: primeiro, os objetivos de negócios devem ser determinados e o problema a ser resolvido deve ser entendido com um algoritmo de *machine learning*, bem como os KPIs que servirão para medir o sucesso do projeto. Com isso, o planejamento do projeto é realizado.

Preparação de dados: esta fase do processo refere-se a um tratamento prévio dos dados e inclui as fases de coleta (obtenção, rotulagem e classificação dos dados coletados e aprimoramento dos dados existentes), limpeza e preparação (tratamento dos dados para torná-los utilizáveis), análise (detecção de padrões e desenvolvimento de hipóteses), visualização (representação gráfica para identificar tendências, *outliers* ou padrões), integração (combinação de *datasets* diferentes para ter uma visão unificada) e *feature engineering* (conversão de dados brutos em dados com a forma desejada, geração de novas variáveis e seleção de variáveis a serem incluídas no modelo).

Desenvolvimento do modelo: esta fase do processo refere-se à seleção do tipo de modelo utilizado, seu treinamento, avaliação de desempenho e critérios estatísticos, bem como o ajuste de parâmetros, e inclui a escolha (avaliação da modelos diferentes disponíveis para encontrar o que melhor se encaixa), treinamento (avaliação das diferentes configurações para converter as informações fornecidas em padrões e relacionamentos), avaliação (análise do desempenho do modelo por métricas usando dados que não foram usados durante o treinamento) e ajuste de parâmetros (revisão da possibilidade de melhorar a previsão do modelo reajustando os hiperparâmetros).

Avaliação, validação e aprovação: avaliar se os objetivos de negócios estabelecidos no início foram atingidos e se as expectativas iniciais foram atendidas. Da mesma forma, dependendo da governança definida e da classificação dos modelos (*tiering*) estabelecidos na estrutura de gerenciamento de risco do modelo da instituição[27], pode haver fases adicionais nas quais as equipes de validação e auditoria, independentes dos desenvolvedores, executam uma revisão dos diferentes aspectos do modelo (dados utilizados, metodologia, resultados, documentação, etc.). Essa validação pode incluir técnicas de interpretabilidade[28], a fim de entender os relacionamentos subjacentes que explicam o resultado do modelo. Da mesma forma, os processos de aprovação estabelecidos nas estruturas de governança da instituição devem ser realizados. No caso de modelos regulatórios, um processo de aprovação final deve ser realizado pelo supervisor.

Implantação e integração no gerenciamento: finalmente, passadas as fases anteriores, o modelo é integrado ao gerenciamento, através da implementação de arquiteturas tecnológicas, do início da produção e dos processos de monitoramento e monitoramento periódicos dos resultados.

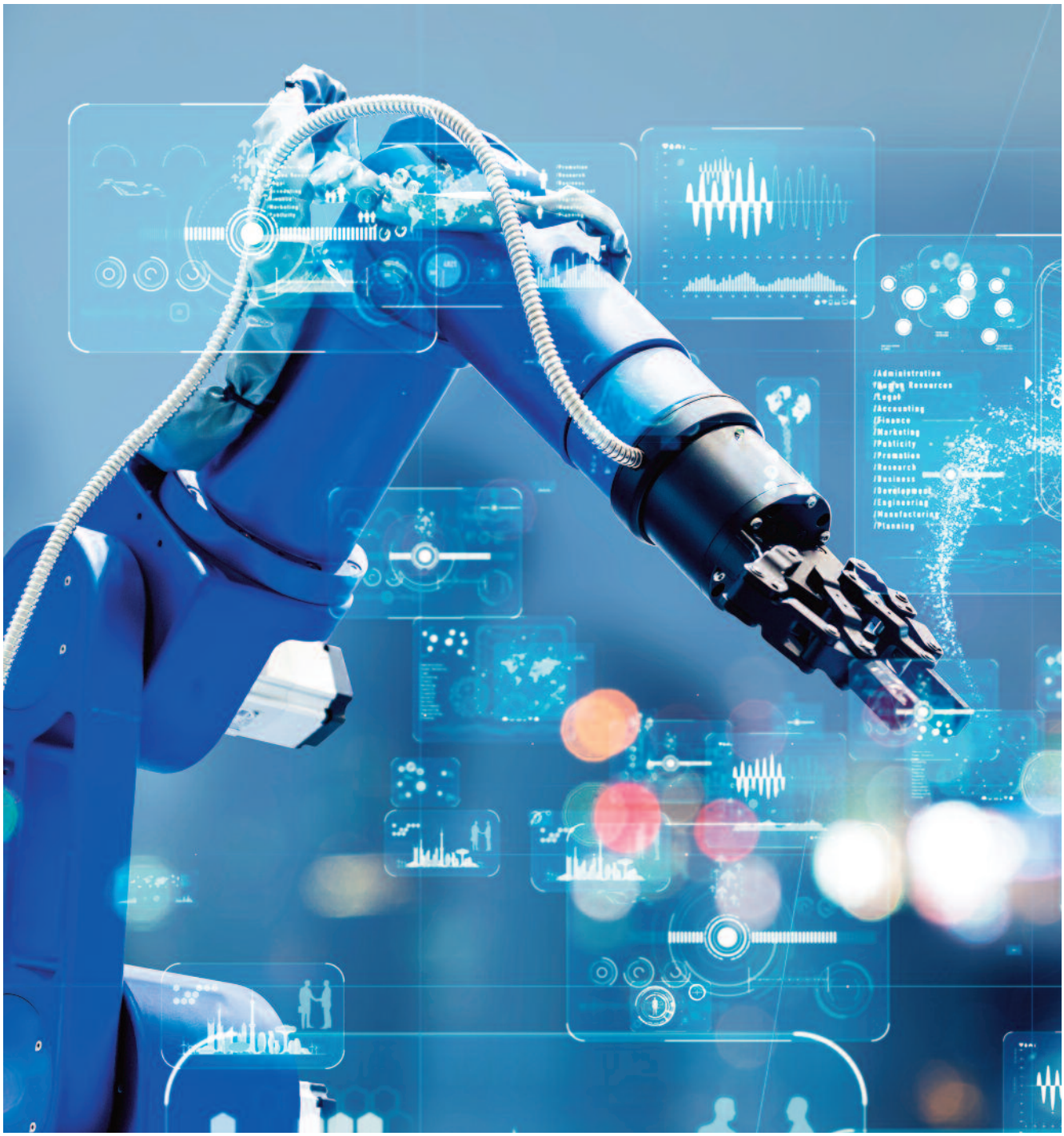
[27] Um detalhe dessas estruturas de gerenciamento pode ser encontrado na publicação "Model Risk Management: Aspectos quantitativos e qualitativos da gestão do risco de modelo", Management Solutions, 2014.

[28] Um detalhe dessas técnicas pode ser encontrado na publicação Cátedra iDanae "Interpretabilidad de los modelos de Inteligencia Artificial", UPM y Management Solutions, 2019.

Estruturas de automação de processos de machine learning

"We've heard that a million monkeys at a million keyboards could produce the complete works of Shakespeare; now, thanks to the Internet, we know that it is not true"

Robert Wilensky[29]



Uma vez discutidos os motivos que levam a componentização e automação dos *workflows* e algoritmos de *machine learning*, surge a principal questão sobre a melhor abordagem para realizá-lo. Especificamente, quando se trata de automatizar o processo de desenvolvimento de modelos de *machine learning*, é necessário responder às seguintes perguntas:

- Quais são as etapas preliminares necessárias para preparar os dados antes do processo de modelagem?
- Como os algoritmos mais adequados para o conjunto de dados a serem avaliados devem ser selecionados?
- Como deve ser determinado o espaço de busca para hiperparâmetros e possíveis configurações?
- Deve-se seguir uma abordagem para reduzir o tamanho do espaço de busca?

Nas abordagens tradicionais, a maneira de responder a essas perguntas tem sido através da seleção manual desses critérios por um analista, com base em configurações a priori e que funcionaram no passado, bem como por tentativa e erro que inclui algum componente da aleatoriedade.

Na prática, tanto no desenvolvimento manual quanto no automático, o problema de seleção de parâmetros obedece aos seguintes desafios:

- Uma abordagem maximalista baseada na revisão exaustiva de todas as combinações possíveis é imensurável no tempo e nos requisitos de recursos computacionais. Mesmo para *datasets* relativamente pequenos ou incorporando restrições de pesquisa como resultado da experiência, essa tarefa ainda é inacessível, o que implica a obrigação de renunciar à otimização em algumas partes do processo.
- As configurações usadas são altamente dependentes dos apriorismos dos analistas e dos ajustes manuais, tornando necessário programar explicitamente uma grande quantidade de código. Portanto, a escolha e o desempenho de muitos dos métodos de *machine learning* usados dependem de um grande número de decisões sobre seu design, feitas manualmente ou com base em hipóteses anteriores.
- Se se trata de gerar uma função de avaliação que permita conhecer a relação entre as alterações nos hiperparâmetros e o desempenho do modelo, a geração disso pode ser muito cara e, às vezes, essa relação não é clara ou não permite inferência sobre os resultados obtidos.
- Essa restrição não é apenas interpretada globalmente, uma vez que o impacto na função de perda nas alterações nos hiperparâmetros não pode ser adequadamente inferido, mesmo localmente.
- Não pode ser otimizado diretamente quando os datasets são grandes, pois os tempos de execução são longos.

Portanto, embora haja incentivos para realizar procedimentos de busca sistemática e automática; a configuração desses sistemas envolve resolver como avaliar as possíveis configurações, dadas as restrições existentes.

Considerando o exposto acima, uma visão que geralmente sustenta o desenvolvimento dos componentes de um sistema de AutoML se baseia em:

1. Automatizar a maioria dos aspectos relacionados à análise e pré-tratamento dos dados, gerando sistemas que permitem que os dados sejam processados e as variáveis sejam transformadas usando as técnicas mais comuns no tratamento manual.

[29] Robert Wilensky durante um discurso em 1996. Professor da Escola de Informação da Universidade da Califórnia em Berkeley, seu principal campo de pesquisa era a inteligência artificial.

2. Gerar um espaço de busca para possíveis modelos e parâmetros em que um conjunto de opções é configurado para sua geração e, através de um critério que percorre esse espaço, os melhores modelos podem ser obtidos, comparados e selecionados.

3. Por fim, automatizar as técnicas de interpretabilidade, embora separadamente do modelo de otimização anterior, para gerar relatórios que sejam mais compreensíveis por usuários diferentes.

Por fim, o objetivo é obter um sistema que, automaticamente, permita encontrar padrões nos dados, selecionando uma maneira de responder a uma pergunta do usuário e capaz de explicar adequadamente os resultados. Com isso, as tarefas com maior complexidade e menos relacionadas ao negócio são substituídas, o acesso a perfis de especialistas no negócio é permitido com um treinamento menos aprofundado nas áreas de data science, realizando todos os processos com eficiência e robustez e tendo levado em consideração as restrições de tempo de computação e execução. Portanto, idealmente, o sistema de AutoML deve permitir automatizar:

- O processo de processamento de dados quando eles têm *missing, outliers,* são mal categorizados ou há erros neles.

- A possibilidade de combinar, reduzir, transformar, criar ou eliminar variáveis com base em critérios estatísticos.

- O processo de seleção de variáveis.

- Seleção de um modelo que tenta evitar o *overfitting* (ajuste excessivo nos dados de treinamento, distorção da previsão de dados desconhecidos) e *underfitting* (o conceito oposto ao ajuste excessivo: quando um modelo não encaixa dados suficientes, como prever corretamente).

- A explicação dos padrões identificados nos dados para o usuário, para que um humano possa entendê-los.

O objetivo do sistema é realizar todos esses processos de maneira eficiente e robusta, levando em consideração restrições computacionais e tempos de execução. Atualmente, existem muitas soluções propostas, incluindo estruturas que permitem o uso centralizado, distribuído ou na nuvem. Embora o grau de desenvolvimento dessas abordagens possa competir e derrotar especialistas humanos em *machine learning*, ainda há muitas questões que precisam ser resolvidas para que sejam aplicadas corretamente.

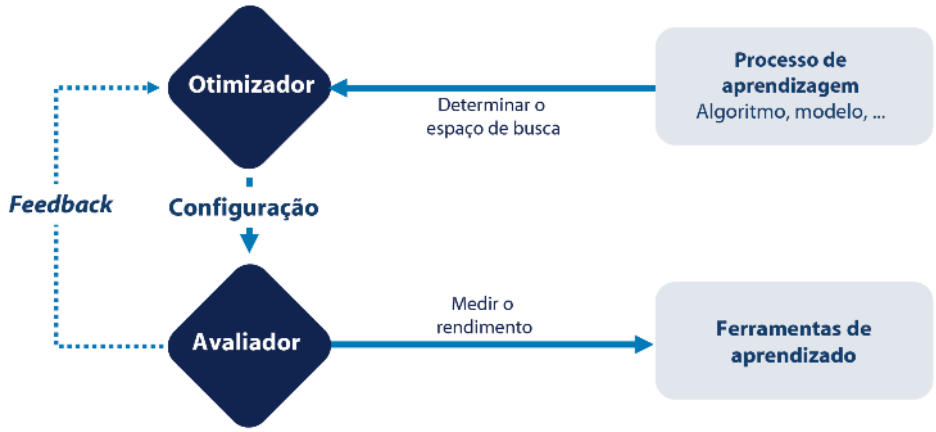
Uma *framework* geral que engloba todas as possíveis partes para automatizar estão indicadas na Figura 3[30]. Este *framework* baseia-se na interação de dois componentes fundamentais: um otimizador, que funciona em um espaço de busca definido, e um avaliador.

Por um lado, o otimizador gera e atualiza as configurações usando um espaço de busca determinado de acordo com o modelo escolhido e o tratamento dos dados que foram executados anteriormente. Posteriormente, o avaliador é responsável por medir o desempenho das configurações propostas pelo otimizador. Dependendo da abordagem selecionada, o avaliador pode afetar a estratégia de pesquisa do otimizador.

Em geral, os componentes que são automatizados no fluxo são algumas fases do processamento de dados (pré-processamento, *feature enginerring,* tratamento de *missings*, dimensionamento etc.), modelagem (seleção de algoritmos,

[30] Yao, e outros, 2018.

Figura 3: framework geral para un sistema de AutoML.



Fonte: Yao, e outros, 2018.

otimização de hiperparâmetros etc.) e, finalmente, a avaliação dos resultados. Algumas partes do processamento de dados geralmente são deixadas de fora do processo de automação, pois dependem mais da percepção dos negócios. Da mesma forma, a interpretabilidade do modelo não é avaliada automaticamente, embora geralmente sejam incluídas ferramentas que ajudam a entender os resultados.

Embora exista uma grande variedade de opções, é necessário gerar um sistema de AutoML no qual o método, descrito como uma estrutura teórica, se torne um conjunto de tarefas (na forma de programas) relacionadas entre si (separando o tarefas em componentes ou através de um design *end-tot-end)*. Dessa forma, esse sistema consiste em um *workflow* que automatiza o *design* do fluxo de trabalho de modelagem.

Em geral, existem abordagens mistas no mercado, nas quais algumas fases são separadas (como preparação de dados e tratamento variável, além de explicações explicativas pós-modelo), enquanto os componentes *feature engineering* , seleção de algoritmo e avaliação de modelo são incluídas em um modelo de otimização.

Por sua vez, as abordagens de modelagem no design de um fluxo desse tipo são classificadas em abordagens que enfatizam o processo de modelagem *(model-based schemes)* e aquelas que o fazem sobre dados *(data-driven approaches)*. No primeiro, a modelagem requer conhecimento a priori do componente de negócios e da estatística-matemática que suporta o modelo. No caso da segunda abordagem, a alternativa é processar as informações de dados diretamente, sem particionar por componente no processo de modelagem.

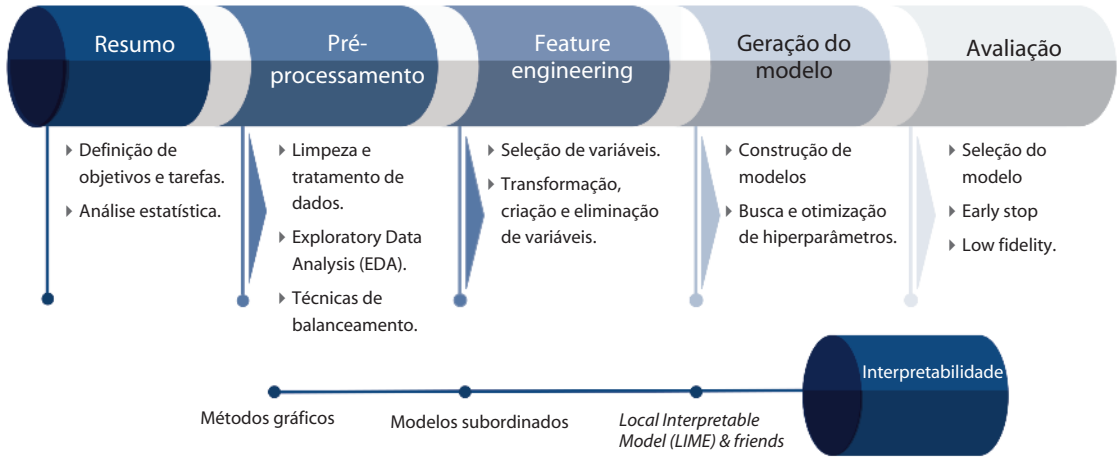
Em geral, existem abordagens mistas no mercado, nas quais algumas fases são separadas (como preparação de dados e tratamento variável, além de explicações explicativas pós-modelo), enquanto os componentes de *feature engineering*, seleção de algoritmo e avaliação de modelo são incluídas em um modelo de otimização.

Componentes de um sistema de AutoML

De acordo com o acima, os diferentes componentes de um sistema de AutoML podem ser separados em diferentes componentes, de modo que, como pode ser visto na Figura 4, a anatomia básica contém os seguintes módulos:

- **Resumo:** fase exploratória do conjunto de dados que definirá a maior parte do conjunto de opções que o processo de AutoML terá que enfrentar.
- **Pré-processamento:** estágio de limpeza e transformação de dados brutos antes do processamento e análise.
- **Feature engineering:** processo no qual o conhecimento fornecido pelos dados é usado para gerar variáveis que permitem um melhor desempenho dos algoritmos de *machine learning*.
- **Geração de modelo:** processo de busca por hiperparâmetro e otimização de modelo
- **Avaliação de modelo:** conjunto de métricas que permitem avaliar a precisão dos modelos obtidos.
- **Interpretabilidade:** combinação de técnicas ou modelos que permitem a interpretação do resultado obtido.

Figura 4: diferentes componentes de um sistema de AutoML[31].



Fonte: He, Zhao, & Chu, 2019.

Otimização de hiperparâmetros

O otimizador usa várias técnicas para encontrar a melhor configuração dos hiperparâmetros, para que o desempenho do modelo seja o melhor possível. Do ponto de vista técnico, a função do otimizador é procurar configurações ideais no espaço de busca, para encontrar o mínimo global ou, pelo menos, um mínimo local, da função de custo. Uma distinção pode ser feita entre técnicas simples (como *grid search, random search,* algoritmos evolutivos ou otimização bayesiana) ou com base na experiência (como *meta-learning* ou *transfer learning)*.

O avaliador, por sua vez, utiliza várias técnicas para estimar o desempenho das configurações propostas pelo otimizador, sendo a mais simples a formação do modelo. Quando isso é muito caro computacionalmente, pode ser necessário usar subamostras ou incluir um *early stop.*

Técnicas para o otimizador: dos métodos greedy ao meta-learning

Uma vez definido o espaço de busca, é necessário estabelecer um otimizador que permita buscas de configurações no espaço. Duas das abordagens mais comuns são o *Grid Search* e o *Random Search,* nas quais nenhuma suposição é feita sobre o espaço de busca.

Uma busca *Grid*, ou de força bruta, estabelece uma grade no espaço de busca e avalia a combinação dada por cada ponto da rede. Esse tipo de busca, que foi aplicada pela primeira vez com uma abordagem de AutoML em 1990, não garante que uma boa configuração seja alcançada (ou seja, um mínimo local) e

[31] He, Zhao, & Chu, 2019.

pode ser computacionalmente cara para um grande número de hiperparâmetros. A partir dessa abordagem, foram desenvolvidos outros que melhoram o processo com base no uso de uma grade inicial para explorar todas as regiões do espaço e, posteriormente, uma grade mais fina nas regiões com melhor comportamento, podendo iterar o processo até que seja encontrado um mínimo local. No entanto, embora os resultados sejam aprimorados, o custo computacional desse tipo de técnica permanece alto.

Uma das primeiras soluções que aprimora os resultados de uma busca *Grid* é a *Random Search,* que se baseia na seleção de um ponto no espaço de busca aleatoriamente. Isso permite que sejam feitas buscas em áreas do espaço que não são igualmente distribuídas e, portanto, podem avaliar áreas com maior desempenho (consulte a figura 5). Essa técnica ainda é computacionalmente cara, embora, como solução, atenda à condição de convergência: quanto maior o tempo de busca, maior a probabilidade de encontrar o conjunto ideal de hiperparâmetros.

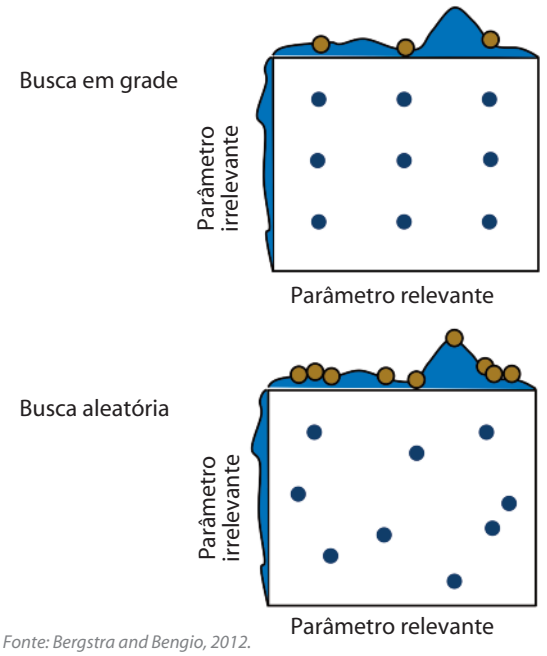
Algumas abordagens mais elaboradas de algoritmos incluem, por exemplo, algoritmos evolutivos (incluindo algoritmos genéticos). Esses algoritmos criam, em uma primeira fase, uma população inicial de configurações aleatoriamente. Eles então avaliam o desempenho de todos os indivíduos da população e selecionam os melhores desempenhos para criar uma nova geração com base na primeira. Além disso, é possível adicionar mutações às novas gerações, para que elas sejam diferentes da geração anterior. Esse tipo de algoritmo permite otimizar uma ampla variedade de problemas, mas ainda não é muito eficiente em termos de custo computacional, pois ainda é necessário avaliar todos os indivíduos de todas as gerações.

Tanto os métodos de busca por grade ou busca aleatória como os algoritmos evolutivos têm o risco de que possam investigar repetidamente regiões com desempenho muito baixo do espaço de configuração sem poder incluir adequadamente uma condição na programação do algoritmo. A otimização bayesiana (utilizada ao menos desde 2005[33]) resolve esse problema criando um modelo probabilístico da função de custo, através do qual ele seleciona as melhores configurações possíveis do hiperparâmetros para avaliá-los e estimar a verdadeira função do custo. A otimização bayesiana pode atualizar o modelo iterativamente, rastreando os resultados de avaliações anteriores. Isso permite atualizar o modelo probabilístico em cada cálculo.

Há casos em que os processos acima não podem ser aplicados, por exemplo, devido à falta de dados. Em outros casos, onde os conjuntos de dados podem ser semelhantes aos outros

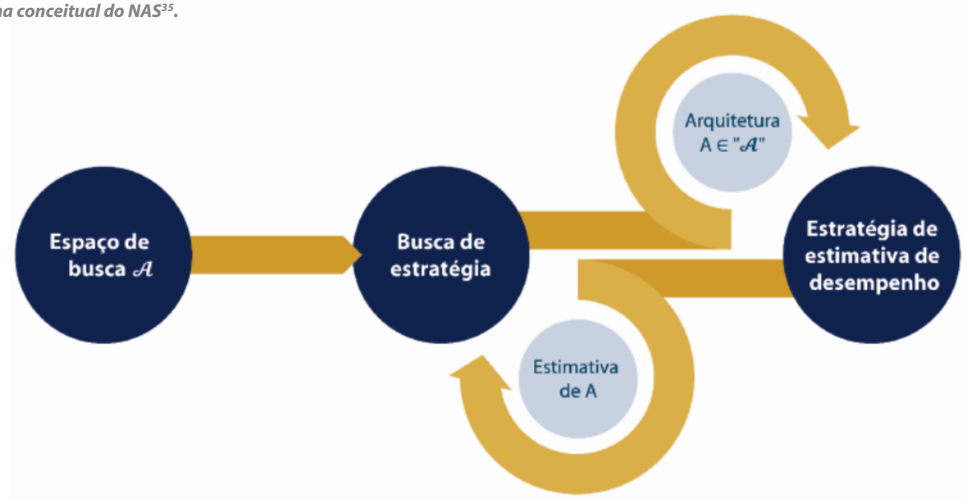
[32] Michie, Spiegelhalter, Taylor, & Campbell, 1994.
[33] Fröhlich & Zell, 2005.
[34] Bergstra and Bengio, 2012.

Figura 5: comparação de buscas grid e random [34].



Fonte: Bergstra and Bengio, 2012.

Figura 6: esquema conceitual do NAS[35].



Fonte: Elsken, Metzen, & Hutter, 2019.

estudados anteriormente, esse conhecimento não será aplicado. Com esses objetivos, foi desenvolvida a abordagem de meta-aprendizagem, também conhecida como "aprender a aprender", que consiste em projetar modelos de *machine learning* capazes de imitar o comportamento humano, aprendendo novos conceitos e habilidades rapidamente, usando um número reduzido de amostras. Ou seja, o objetivo é projetar modelos que possam adquirir novas habilidades e que possam se adaptar rapidamente a novos ambientes em poucos casos.

Técnicas para o avaliador

A maneira mais simples de avaliar as configurações fornecidas pelo otimizador é a avaliação direta dos dados de treinamento e teste. Devido ao grande número de configurações que o otimizador deve fornecer ao avaliador em um processo de AutoML, esse método pode consumir muito tempo ou ser computacionalmente caro. É por isso que existem certas abordagens para acelerar o processo de avaliação, embora isso geralmente signifique uma perda de capacidade preditiva nos modelos obtidos. Essas técnicas incluem a avaliação de subconjuntos de dados de treinamento; processos de parada antecipada, nos quais o avaliador para de avaliar se o desempenho é muito baixo nas primeiras iterações; reutilização de parâmetros treinados em modelos anteriores para inicializar o novo modelo; ou, finalmente, o uso de modelos substitutos para prever o desempenho, geralmente usando a experiência de avaliações anteriores.

Neural Architecture Search (NAS)

Devido ao aumento da aplicação de técnicas de aprendizado profundo em aspectos como reconhecimento de imagem, reconhecimento de voz e tradução automática, uma das áreas em que mais interesse foi dedicado à configuração de arquiteturas de redes neurais. De maneira análoga à mencionada anteriormente, essas configurações são geralmente estabelecidas manualmente por especialistas humanos, o que acarreta os erros mencionados anteriormente.

Como alternativa, a busca por arquiteturas neurais (NAS) baseia-se no uso de diferentes técnicas para automatizar o design de redes neurais. Os aspectos sobre os quais existem parâmetros são análogos aos comentados anteriormente: espaço de busca, estratégia de busca e estimativa de desempenho. Usando essas técnicas, todo o processo é determinado simultaneamente, conforme indicado na Figura 6.

[35] Elsken, Metzen, & Hutter, 2019.

Abordagem de implementação do AutoML

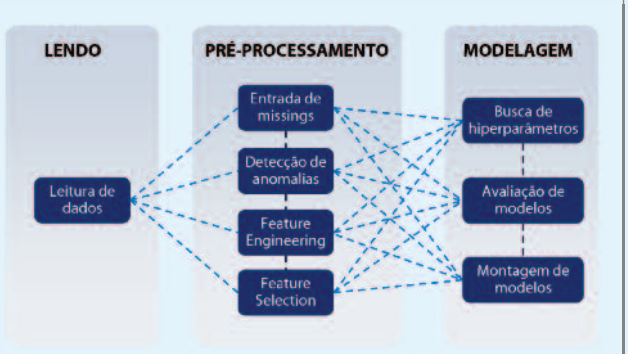
Ao implementar um sistema AutoML na prática, é necessário levar em consideração algumas considerações, como o perfil do usuário que vai usar o sistema ou a profundidade e personalização da análise necessária. No entanto, é possível incluir essas implementações em duas abordagens principais:

- Uma abordagem é projetar um fluxo parcialmente ou totalmente editável (figura 7), onde o usuário pode definir o fluxo que o processo de processamento de dados seguirá, bem como as técnicas que serão aplicadas em cada fase desse processo. Nesse caso, o nível de automação é mais baixo, pois só se aplica à execução dos módulos após a definição de sua ordem. Devido a essas características, essa abordagem é mais adequada quando o usuário possui conhecimento técnico avançado.
- Uma abordagem alternativa é aplicar a automação de ponta a ponta, com um fluxo predefinido (Figura 8). Os dados seguem um processo em que a ordem de cada componente do AutoML é definida de acordo com o pipeline geral de construção do modelo de aprendizado de máquina. Dessa forma, o usuário não precisa modificar a ordem de execução dos componentes em desenvolvimento. Isso pode escolher os tipos de técnicas que se aplicam em cada componente, mas sempre seguindo a ordem predefinida. Devido a essas características, essa abordagem é mais adequada quando o usuário não possui conhecimento técnico avançado, o que é comum em perfis focados nos negócios.

Atualmente, nenhuma das abordagens automatiza a geração de novas variáveis a partir das originais. Os motivos são computacionais (a criação de transformações aleatórias de variáveis gera um custo computacional muito alto) e comerciais (o conhecimento especializado do tipo de problema que está sendo tratado nos permite saber qual transformação é a mais apropriada e nos permite dar um significado mais adequado ao tempo para interpretar o resultado).

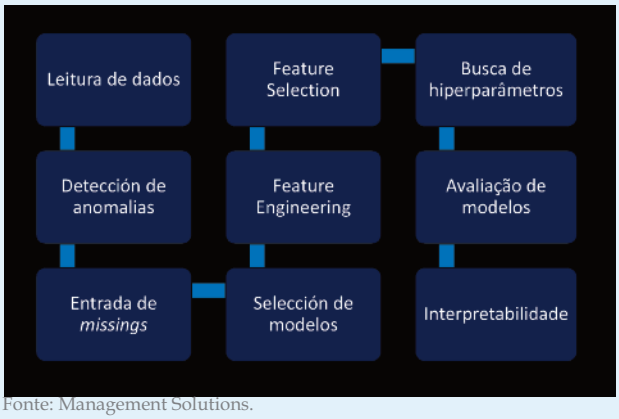
[36] A ferramenta de modelagem incorpora um módulo, *Model Creator*, baseado na automação de ponta-a-ponta, e um módulo alternativo, *Model Component*, que permite a geração de fluxo pelo usuário.

Figura 7: fluxo parcial ou totalmente editável.



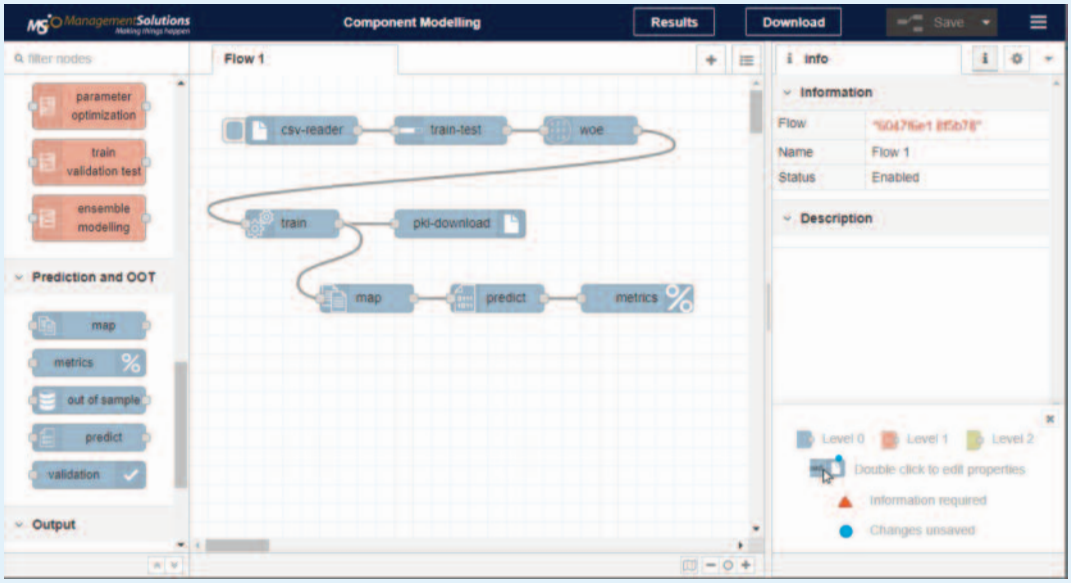
Fonte: Management Solutions.

Figura 8: fluxo pré-estabelecido.



Fonte: Management Solutions.

Figura 9: Workflow projetado na ferramenta de modelagem por componentes criada pela Management Solutions[36].



Fonte: Management Solutions.

Embora ao aplicar abordagens de NAS, os elementos do processo sejam determinados simultaneamente, geralmente são necessários pré-processos, como os relacionados à preparação de dados ou engenharia de recursos, e sua aplicação correta resulta em melhorias no poder preditivo.

Vários elementos estão incluídos no espaço de busca, como o número de camadas do algoritmo, o tipo de operação que cada camada executa, bem como os hiperparâmetros associados a essas operações (como o número de filtros ou o tamanho do kernel associado), bem como o relacionamento e a hierarquia existente entre as diferentes camadas, dependendo do algoritmo usado. Se você tiver informações a priori sobre possíveis arquiteturas que normalmente funcionam corretamente para uma determinada tarefa, poderá reduzir o tamanho do espaço, simplificando a busca. Nessa mesma linha, também são usadas abordagens para reduzir o espaço de busca para estabelecer isso através de blocos de camadas, em vez de toda a arquitetura.

Em relação às estratégias de busca, e como mencionado anteriormente, existem várias estratégias, incluindo otimização bayesiana, busca aleatória, métodos evolutivos ou com base em experiências anteriores, como aprendizado por reforço. Na prática, essas técnicas não apresentam melhores resultados do que a busca baseada em buscas aleatórias. No campo acadêmico, um estudo recente[37] destaca motivos como o uso de espaços de busca restritos por esses algoritmos, bem como a distribuição do peso das diferentes camadas na decisão final como os elementos que limitam os resultados.

Finalmente, em relação às estratégias para otimizar o desempenho do NAS, existem abordagens de "baixa fidelidade", nas quais são utilizados tempos de treinamento mais curtos, treinamento em subconjuntos de dados ou com menos filtros por camada, com o problema de uma subestimação de desempenho. Outras estratégias consistem em extrapolar a curva de aprendizado, usar modelos substitutos para prever o desempenho de novas arquiteturas, inicializar a rede com pesos obtidos em redes previamente treinadas ou usar a teoria dos grafos.

Embora a busca por arquiteturas neurais tenha atingido um nível de desempenho que possa competir com a configuração manual, as razões pelas quais as arquiteturas selecionadas funcionam bem não são claras no momento. Da mesma forma, é necessária uma verificação empírica para determinar se os motivos que fazem uma configuração funcionar podem ser generalizados para diferentes problemas.

Desafios atuais dos sistemas de AutoML

Atualmente, e apesar de ainda haver espaço para melhorias, os sistemas de AutoML atingiram um estágio de desenvolvimento que lhes permite competir e até vencer especialistas humanos em aprendizado de máquina, configurando-se como uma ferramenta fundamental que modifica o tipo de trabalho realizado pelos profissionais envolvidos. Desta forma, os data scientists são distribuídos para tarefas mais relacionadas à análise antes e depois do desenvolvimento dos modelos em si, bem como para a manutenção desses métodos e sistemas.

[37] Sciuto, 2019.

Alguns dos desafios atuais consistem em melhorar o processo, além de incorporar a interpretabilidade e facilitar a interação dos especialistas:

- Atualmente, a maioria das inovações visa à seleção e otimização de modelos, prestando menos atenção ao tratamento e preparação de dados. Isso se deve à dificuldade de automatizar alguns processos sem incorrer em um alto custo computacional.

- Outra questão em aberto é como lidar com elementos com uma interpretabilidade muito baixa, conhecidos como caixas-pretas, pois podem acarretar problemas legais, éticos e técnicos na incorporação das decisões. Nesse sentido, algumas das principais linhas de pesquisa têm como objetivo a *Explainable AI*, a interpretabilidade e o aprimoramento da rastreabilidade e transparência dos modelos. Esse problema também é compartilhado pela maneira tradicional de desenvolver modelos de *machine learning*, no entanto, a maior automação oferecida por um sistema de AutoML torna seu uso nesse processo mais enfatizado.

- Por outro lado, para que os sistemas de AutoML sejam eficazes, eles devem permitir que o usuário interaja com o sistema, modificando e substituindo as decisões que tomam, incorporando o conhecimento dos especialistas em negócios em relação a vários aspectos do processo, como as previsões feitas ou em relação à complexidade e interpretabilidade dos modelos obtidos.

  Por fim, destaca a necessidade de estabelecer benchmarks que sirvam como padrões para poder comparar o desempenho entre as diferentes soluções propostas, além de ter uma definição clara das métricas usadas para medir esse desempenho.

Augmented machine learning

Uma das abordagens que estão atraindo mais atenção derivada da generalização da aplicação dos métodos e sistemas de AutoML é o chamado *Augmented machine learning*. Nessa abordagem, a automação de certos processos visa permitir que os sistemas de AutoML lidem com a complexidade derivada do aumento de possíveis arquiteturas, opções de hiperparâmetro e opções de treinamento, mas continuem a ter um especialista que use os resultados da ferramenta e avaliar as alternativas e combinações lançadas de maneira holística. Isso é explicado por vários motivos:

- A primeira delas é que esses sistemas não podem incorporar o contexto que o usuário possui nos dados, portanto, parece que o processo que o usuário guia o sistema na busca de padrões nos dados ainda melhora. Esse conceito, conhecido como engenharia de representação, é, por exemplo, comum em áreas como a interpretação de pesquisas na Internet[38].

- Por outro lado, a análise de informações em silos por meio dessas ferramentas reduz o valor esperado que pode ser extraído usando técnicas avançadas de análise; portanto, o papel de um especialista em ciência de dados que toma decisões sobre questões é essencial. como quando as diferentes fontes de dados devem ser combinadas ou em quais casos aplicar técnicas de transferência de aprendizado, entre outras, já que não é possível, no momento, que os sistemas de AutoML tratem da análise de todas as opções possíveis antes de tomar a decisão .

- Por fim, e conforme mencionado no ponto anterior, questões éticas relacionadas ao objetivo, aos dados utilizados e aos possíveis vieses gerados no processo de decisão exigem que um analista avalie a relevância do uso do modelo em um processo de tomada de decisão, bem como avaliar as limitações do modelo. Em geral, o uso da percepção funciona adequadamente, enquanto na aplicação do julgamento automático é imperfeito, mas está melhorando. No caso da previsão do comportamento humano, seus resultados são fundamentalmente duvidosos[39].

[38] Abbasi, Kitchens, & Ahmad, 2019.
[39] Narayanan, 2019.

Campeonatos de AutoML: uma ferramenta de exploração do enfoque de AutoML

"Ford! he said, there's an infinite number of monkeys outside who want to talk to us about this script for Hamlet they've worked out"

Douglas Adams[40]

Como foi visto nas seções anteriores, e apesar dos avanços observados ultimamente nesta disciplina, ainda não existe uma abordagem preferível sobre as demais alternativas em questões como pré-processamento de dados antes da modelagem, sobre como selecionar algoritmos ou como configurá-los corretamente. No entanto, algumas tendências e taxas estão começando a ser traçadas que qualquer processo que integre as técnicas do *Advanced Analytics* e, em particular, um sistema de AutoML deve incorporar.

Uma abordagem comum para avaliar diferentes abordagens de AutoML tem sido o desenvolvimento de campeonatos de cientistas de dados com o objetivo de criar sistemas de AutoML. Essa é uma boa referência, pois permite que as diferentes abordagens sejam enfrentadas em condições iguais e, portanto, extraia de seus resultados se houver algumas configurações preferíveis e em quais circunstâncias elas funcionam melhor. Inicialmente, essas competições foram baseadas na avaliação da seleção de modelos e hiperparâmetros[41]. Posteriormente, esse tipo de exercício foi aprimorado, de modo que os participantes devem desenvolver um sistema automático e computacionalmente eficiente, capaz de treinar e avaliar modelos sem nenhuma intervenção humana[42].

Em geral, o objetivo principal dessas competições é responder a uma série de perguntas como: i) conhecer o efeito das restrições de tempo no design de algoritmos; ii) identificar quais tarefas são mais difíceis e para qual tipo de participantes. iii) saber se existem certas configurações que geralmente funcionam melhor para certos tipos de conjuntos de dados ou problemas; iv) avaliar o impacto da otimização de hiperparâmetros e configurações no desempenho dos modelos finais.

Uma revisão dos campeonatos de AutoML

Nos diferentes campeonatos analisados, alguns padrões podem ser observados[43]:

- Em geral, é comum o uso de abordagens heurísticas ou de grade ou buscas uniformes no espaço de busca definido por meio de uma definição linear ou logarítmica.
- Em alguns casos, o método acima é aprimorado através do uso de métodos de regularização.
- O overtraining é controlado incluindo condições de parada nos métodos de otimização iterativa.
- A separação entre a amostra de treinamento e validação geralmente não é otimizada.

Como resultado, percebe-se que em nenhum caso é possível automatizar todo o processo e deve haver intervenção humana nas tarefas mais relacionadas à definição de exercício. Ainda é difícil selecionar um sistema de acordo com o tipo de problema e adaptá-lo ao conjunto de dados existente.

Em relação às variáveis, a dificuldade está diretamente relacionada à existência de certos atributos nos conjuntos de dados analisados, como a existência de dados não formatados, a escassez de dados, a existência de missings ou a existência de variáveis categóricas. Nesses casos, a intervenção para identificar, tratar e avaliar o impacto do tratamento no processo é maior.

Em relação ao processo de configuração e seleção de hiperparâmetros, os principais problemas decorrem do uso de técnicas ad hoc que pioram o desempenho do modelo, como a separação por técnicas não sofisticadas da amostra em treinamento e teste, seleção inadequada da complexidade da modelo, seleção de hiperparâmetro selecionando a amostra de teste, não usando todos os recursos computacionais ou definindo métricas de desempenho inadequadas.

[40] Douglas Adams, "The Hitchhiker's Guide to the Galaxy" (1979). Escritor e roteirista inglês, especialmente conhecido pela saga do nome homônimo.
[41] Veja por exemplo, NIPS 2005.
[42] NIPS 2016, ICML 2016 y PAKDD 2018.
[43] Hutter, Kotthoff, & Vanschoren, 2019.

Do ponto de vista das técnicas de busca, existe um amplo uso de técnicas baseadas em grid ou distribuições uniformes nos parâmetros do espaço de busca, embora existam algumas sofisticações baseadas em métodos de regularização ou abordagens bayesianas que evitam o *overfitting* incorporando condições de parada.

Campeonato de AutoML Management Solutions

Objetivo e definição

Em um espírito semelhante, a Management Solutions projetou e realizou um campeonato, dirigido aos profissionais da Firma, com o objetivo de gerar um algoritmo de AutoML capaz de fazer previsões em diferentes conjuntos de dados sem fazer modificações no código, com uma limitação de tempo para incentivar a eficiência computacional. O exercício proposto foi baseado na resolução, através da aplicação de abordagens supervisionadas, de problemas de resposta binária nas seguintes condições:

- 3 datasets com tamanhos distintos (<100 kb, <1 Mb e <5 Mb), todos eles com uma amostra balanceada
- Sem valores missings, com variáveis categóricas e contínuas, e incluindo variáveis irrelevantes
- Com uma limitação de recursos computacionais: computador com Windows 10, processador Intel Core i5-6300 CPU @ 2.40GHz 2.50GHz e 8 Gb de memória RAM, e com um tempo de execução máximo de 20 minutos para cada dataset

Avaliação

A função enviada foi avaliada com três datadets diferentes, semelhantes aos enviados como amostras de treinamento. Para isso, os seguintes aspectos foram levados em consideração:

- Métrica área sob a curva (AUC) (50%)
- Qualidade e limpeza do código e utilização de padrão PEP8 (20%)
- Utilização de Programação Orientada a Objetos (10%)
- Originalidade (20%)

Resultados

O número de participantes foi superior a cem, integrados em mais de setenta equipes, com perfis muito diversos, havendo entre os participantes tanto físicos e matemáticos, como engenheiros e economistas. Muitos dos participantes, possuem ou estão cursando alguma pós-graduação em *data science*. A origem geográfica dos participantes também foi muito diversa, com participantes do Peru, Chile, Colombia, Brasil, Alemanha, Estados Unidos e Espanha.

Durante toda a competição, os participantes enfrentaram várias opções em relação ao processamento de dados, escolha de modelos e otimização de hiperparâmetros. A maioria das equipes realizou um tratamento de dados com base na eliminação de possíveis outliers e variáveis correlacionadas, na normalização de variáveis, na redução da dimensionalidade e da entrada de *missings*. Algumas equipes usaram técnicas WOE ou one-hot-encoding para variáveis categóricas e tratamento

específico no caso de haver conjuntos de dados desequilibrados, bem como a consideração de interações entre variáveis para aumentar a capacidade preditiva; ou a eliminação de variáveis irrelevantes, como variáveis constantes, com muito pouca variação ou variáveis categóricas com um número muito grande de categorias em relação ao número total de entradas.

O objetivo por trás desses tratamentos é claro: por um lado, prepara os dados para serem lidos corretamente pelos modelos utilizados e, por outro, reduz a dimensionalidade do espaço de busca, para que seja necessário menos tempo para encontrar configuração ideal. Outros participantes adotaram uma abordagem diferente para lidar com esse problema, limitando o número de execuções de algoritmos a um número específico e constante ou limitando o número de modelos que são avaliados pelo sistema.

A otimização do hiperparâmetro foi focada, na maioria dos casos, pelo uso de pesquisas grid. Algumas equipes usaram pesquisa aleatória, algoritmos genéticos ou pesquisa bayesiana. Deve-se observar que um participante implementou o uso de *random search* para subsequentemente realizar uma pesquisa em um ambiente com a configuração ideal encontrada, para tentar melhorar a métrica com essas configurações, caso o resultado fornecido pelo random search não superasse um pontuação determinada.

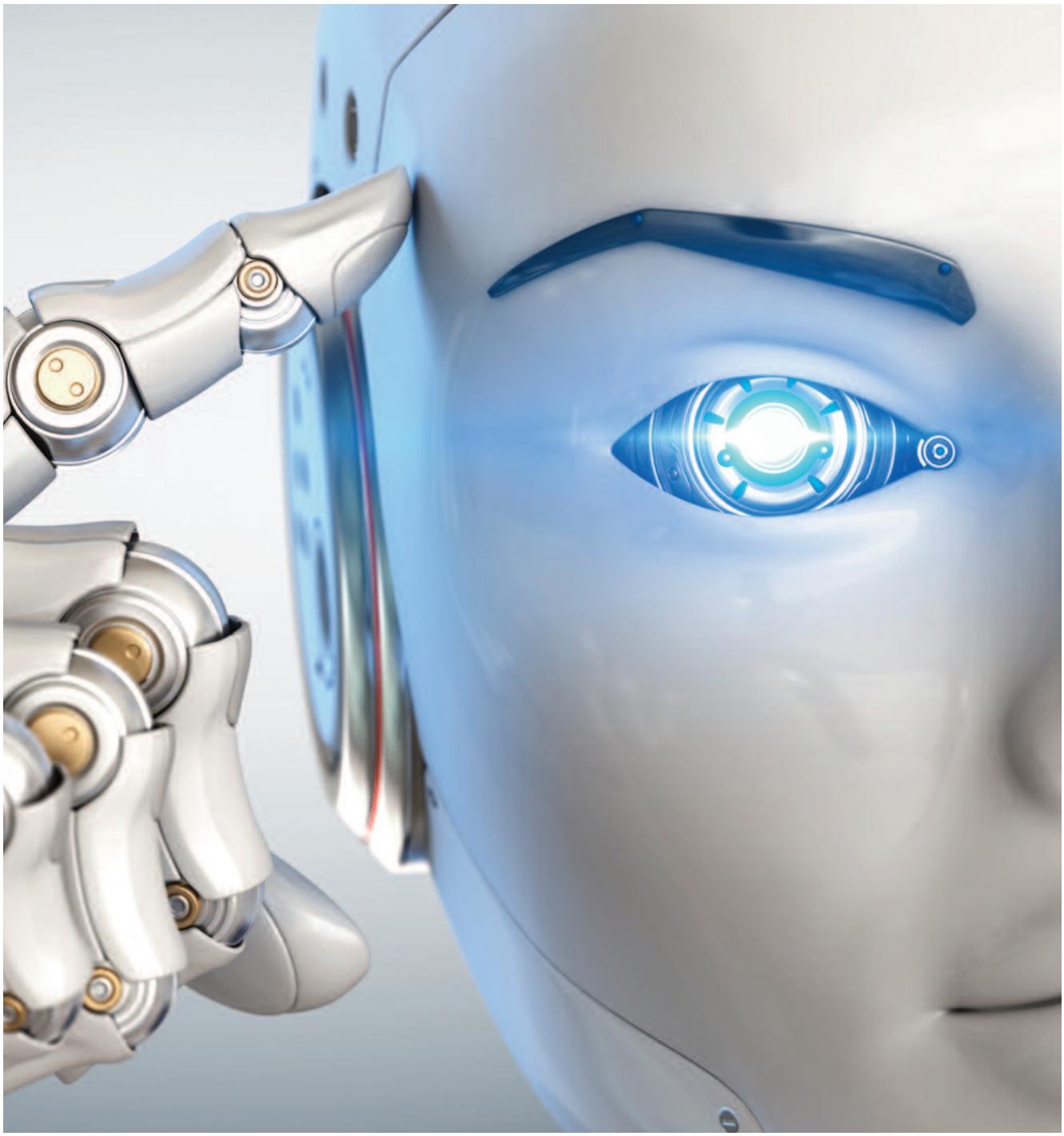
Para avaliar o desempenho da configuração proposta, as equipes usaram cross validation, e os modelos implementados foram, em grande parte, obtidos na biblioteca *scikit-learn,* com algumas exceções, como o uso de keras, *lightgbm* ou *xgboost*. Para otimizar o tempo computacional, alguns participantes realizam um estudo preliminar das variáveis mais preditivas para trabalhar apenas com elas, enquanto outros avaliaram uma lista de modelos e pararam de avaliar quando o tempo máximo definido foi atingido, podendo haver modelos estimados, mas não avaliados na amostra.

Cabe destacar que, em geral, foi realizada uma otimização de todo o pipeline, o uso de modelos stacking ou a paralelização de tarefas em vários núcleos, bem como a inclusão de módulos de interpretabilidade por algumas das equipes participantes, seja para interpretar os datasets ou para interpretar o processo de AutoML, como pode ser a seleção de um modelo frente a outros.

A limitação no tempo de execução de cada dataset não teve um grande impacto em geral, uma vez que os arquivos de avaliação eram pequenos e o AutoML dos participantes executaram sem problemas dentro do tempo. Apenas em alguns casos, alguns participantes limitaram o número de modelos a serem avaliados para evitar levar mais tempo do que o estipulado.

Reflexões Finais

Well? What do you think of my new poem?
I once read that given infinite time, a thousand monkeys with typewriters would eventually write the entire works of Shakespeare
But what about my poem?
Three monkeys, ten minutes
– Scott Adams[44]



Situação atual e desafios do AutoML

As configurações usadas para obter modelos de *machine learning* dependem significativamente dos conhecimentos prévios do analista e ajustes manuais, o que significa que o desenvolvimento de modelos usando técnicas de *machine learning* requer programação explícita de uma grande quantidade de código. Dessa forma, a escolha e, portanto, o desempenho de muitos dos métodos de *machine learning* usados depende de um grande número de decisões sobre seu desenho, que são tomadas manualmente ou com base em hipóteses anteriores e, portanto, pode ocorrer overfitting nos modelos desenvolvidos com *datasets* pequenos e *underfitting* para *datasets* maiores[45], indicando que ainda são necessárias melhorias para garantir o uso adequado desses sistemas para sua aplicação industrial.

Embora as abordagens de AutoML tenham atingido um estágio de desenvolvimento que possa competir e derrotar os especialistas humanos em *machine learning,* ainda há muitos problemas que precisam ser resolvidos para que sejam aplicados corretamente. O principal desafio que os sistemas de AutoML atuais enfrentam é que as decisões de design são tomadas com uma abordagem data-driven, de maneira objetiva e automática.

De qualquer forma, o precedente não é incompatível com o usuário com a possibilidade de interagir com o sistema e com a possibilidade de modificar e substituir as decisões que toma. Nesse sentido, o desenvolvimento de modelos de *machine learning* é realizado por meio de uma indústria artesanal, na qual especialistas enfrentam problemas através do design de soluções manuais, que em muitos casos são tomadas *ad hoc* para esse projeto, bem como preferências e conhecimentos prévios dos especialistas, mas muitas vezes não incorpora a sensibilidade dos especialistas em negócios. Uma interface compreensível para analistas de negócios que permite a execução de um sistema de AutoML evita a decisão manual nas configurações, mas, por sua vez, permite a incorporação de decisões de negócios quanto ao sinal ou importância das variáveis ou à seleção de modelos com base em uma interpretação das projeções, a sensibilidade aos cenários ou a complexidade e interpretabilidade dos modelos obtidos.

Outra questão em aberto é como lidar com os elementos que são *black boxes,* uma vez que limitam sua interpretabilidade e podem acarretar problemas legais, éticos e técnicos na incorporação das decisões. Nesse sentido, algumas das principais linhas de pesquisa visam o *Explainable AI,* explicável, interpretabilidade e melhoria da rastreabilidade e transparência dos modelos.

Finalmente, alguns aspectos, como a eficiência dos processos de busca, estão sendo constantemente aprimorados, como pode ser observado nas diferentes competições[46].

Grau de desenvolvimento

Os avanços no AutoML são desiguais: a maioria das inovações visa a algumas técnicas de *feature engineering* e seleção de modelos versus preparação e tratamento de dados[47], onde ainda há um longo caminho a percorrer. Isso tem efeitos tanto no tipo de tarefas que devem ser realizadas nas organizações quanto no volume de emprego.

Por um lado, substitui as tarefas por maior complexidade e menos relacionadas aos negócios relacionados ao design de pipelines para cada problema específico, permitindo acesso ao design de um *pipeline* completo para perfis com menos conhecimento de *machine learning* e, portanto, dedicar estas tarefas a experts no negócio com uma formação menos profunda em âmbitos de *data science*.

Por outro lado, requer uma infraestrutura que permita a execução desses processos, além de mantê-los atualizados, seja através da terceirização ou contratação de serviços de AutoML, seja através da geração de um AutoML próprio que exija equipamento especializado que o fluxo de trabalho funcione corretamente.

Além disso, existem problemas que dificilmente foram abordados pelos sistemas de AutoML, como tarefas como integração ou limpeza de dados, geração de variáveis ou tratamento delas, além de algumas abordagens de *machine learning*, como, por exemplo, o aprendizado não supervisionado ou o reinforcement learning não são rotineiramente integrados a esses sistemas.

Dessa forma, espera-se que os sistemas de AutoML sejam configurados como uma ferramenta fundamental, capaz de modificar o tipo de trabalho realizado, para que os recursos de *data science* sejam distribuídos em tarefas mais relacionadas à análise, antes e depois do desenvolvimento dos próprios modelos, como a geração dos sistemas de AutoML, bem como a resolução de problemas em que as ferramentas genéricas do AutoML não permitem uma configuração adequada.

[44] Scott Adams em uma tira de Dilbert de 1989. Desenhista, autor da tira diária homônima.
[45] Por exemplo, no caso de métodos HPO. Ver Hutter, Kotthoff, & Vanschoren, 2019.
[46] Hutter, Kotthoff, & Vanschoren, 2019.
[47] Ibídem.

Bibliografia

Abbasi, A., Kitchens, B., & Ahmad, F. (2019). The Risks of AutoML and How to Avoid Them. Harvard Business Review.

Bank of England. (2019). Machine learning in UK financial services. Bank of England.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of machine learning research.

Cátedra iDanae. (3T-2019). Interpretabilidad de los modelos de Machine Learning. Cátedra iDanae.

Cátedra iDanae. (4T-2019). Ética e Inteligencia Artificial. Cátedra iDanae.

CrowdFlower. (2017). Data Scientist Report. CrowdFlower.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural Architecture Search: A Survey. Journal of Machine Learning Research.

European Banking Authority. (2020). EBA report on Big Data and Advanced Analytics. European Banking Authority.

European Comission. (2020). White paper on Artificial Intelligence - A European approach to excellence and trust. European Comission.

Fröhlich, H., & Zell, A. (2005). Efficient Parameter Selection for Support Vector Machines in Classification and Regression via Model-Based Global Optimization. IEEE Xplore.

Gartner. (2019). How Augmented Machine Learning Is Democratizing Data Science. Gartner.

He, X., Zhao, K., & Chu, X. (2019). AutoML: A Survey of the State-of-the-Art. arXiv preprint arXiv:1908.00709.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated Machine Learning: Methods, Systems, Challenges. Springer.

Management Solutions. (2014). Model Risk Management: Aspectos quantitativos e qualitativos da gestão do risco de modelo. Management Solutions.

Management Solutions. (2018). Machine Learning, uma peça-chave na transformação dos modelos de negócio. Management Solutions.

Management Solutions. (2019). De projetos Agile, a organizações Agile. Management Solutions.

Michie, D., Spiegelhalter, D., Taylor, C., & Campbell, J. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Narayanan, A. (2019). How to recognize AI snake oil.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development. IBM J. Res.

Sciuto, C. &. (2019). Evaluating the Search Phase of Neural Architecture Search. Sciuto, Christian & Yu, Kaicheng & Jaggi, Martin & Musat, Claudiu & Salzmann, Mathieu.

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression.

Stanford University. (2019). AI Index Report. Stanford University.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. IEEE transactions on evolutionary computation.

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., . . . Yu, Y. (2018). Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv preprint arXiv:1810.13306.

Glossário

Cloud Computing: disponibilidade de recursos, como armazenamento de dados e poder computacional, sem gerenciamento ativo pelo usuário.

Configuração: as possíveis combinações de valores que os hiperparâmetros podem ter .

Cross Validation: processo de validação de amostra cruzada que consiste em dividir a amostra em grupos k e usar iterativamente cada grupo para validação e o restante para construção, alterando o grupo de validação em cada iteração.

Early Stop: técnica que consiste em interromper o processo de pesquisa antes do planejado, se determinados requisitos forem atendidos.

Espaço de configurações: conjunto de todas as configurações possíveis nas quais a configuração ideal é procurada para fazer a melhor previsão possível.

Feature Engineering: processo de extração de características dos dados mediante o uso de técnicas de *data mining* e conhecimento de um âmbito concreto.

Função de custo: função cujos mínimos correspondem às configurações ideais. Procurar a configuração ideal é equivalente a encontrar os mínimos da função de custo. Algumas funções de custo podem ser o erro quadrático médio raiz ou a entropia cruzada, entre outras opções.

Grid / random / evoluções / bayes / meta / transfer: diferentes métodos usados para procurar otimizações de hiperparâmetros.

Hiperparâmetro: parâmetro que não pode ser obtido durante o processo e deve ser definido anteriormente. Os valores que os hiperparâmetros devem adotar para resolver um problema específico são desconhecidos.

Machine learning: campo da ciência da computação que se concentra no desenvolvimento de técnicas que permitem que um programa aprenda a encontrar padrões em um conjunto de dados.

Métrica: medida para avaliar o desempenho de um modelo.

Missings: valores que faltam dentro de um dataset.

Modelo substituto: modelo geralmente mais simples, que tenta emular um modelo mais complexo em determinados ambientes ou situações.

Normalização: tratamento de dados que consiste em fazer a média dos valores de uma variável centrada em zero e mover-se entre -1 e 1.

Outliers: valores que, por terem sido mal medidos ou por serem um comportamento atípico, estão numericamente distantes do restante dos dados.

Overfitting / underfitting: característica de um modelo que ocorre quando foi ajustado muito / pouco à amostra de treinamento, não atingindo resultados satisfatórios em amostras que não sejam essa (por exemplo, na amostra de validação)

Parâmetro: propriedade interna ao modelo aprendido durante o processo de aprendizado, sendo necessário para fazer previsões.

Redução de dimensionalidade: processo pelo qual o espaço de busca é reduzido, por combinação de variáveis, eliminação ou outros métodos.

Regularização: técnica matemática que consiste em adicionar um componente à função de custo para detectar as variáveis que não estão fornecendo ao modelo informações significativamente diferentes. É usado para evitar problemas de overfitting (como o caso de redes elásticas)

Variável contínua / categórica: uma variável contínua é uma variável numérica que pode assumir qualquer valor entre dois valores-limite. Uma categórica pode ser uma variável numérica discreta ou pode ser palavras ou outro tipo de variável.

Variáveis correlacionadas: variáveis que possuem um comportamento similar.

WOE (Weight of Evidence): tratamento de dados para variáveis categóricas.

Nosso objetivo é superar as expectativas dos nossos clientes sendo parceiros de confiança

A Management Solutions é uma firma internacional de serviços de consultoria com foco em assessoria de negócios, riscos, organização e processos, tanto sobre seus componentes funcionais como na implementação de tecnologias relacionadas.

Com uma equipe multidisciplinar (funcionais, matemáticos, técnicos, etc.) de 2.500 profissionais, a Management Solutions desenvolve suas atividades em 31 escritórios (15 na Europa, 15 nas Américas e um na Ásia).

Para atender às necessidades de seus clientes, a Management Solutions estruturou suas práticas por setores (Instituições Financeiras, Energia e Telecomunicações) e por linha de negócio (FCRC, RBC, NT), reunindo uma ampla gama de competências de Estratégia, Gestão Comercial e Marketing, Gestão e Controle de Riscos, Informação Gerencial e Financeira, Transformação: Organização e Processos, e Novas Tecnologias.

A área de P&D presta serviço aos profissionais da Management Solutions e a seus clientes em aspectos quantitativos necessários para realizar os projetos com rigor e excelência, através da aplicação das melhores práticas e da prospecção contínua das últimas tendências em *data science, machine learning, modelagem* e *big data*.

Javier Calvo Martín
Sócio
javier.calvo.martin@msgermany.com.de

Manuel Ángel Guzmán
Diretor P&D
manuel.guzman@managementsolutions.com

Daniel Ramos García
Supervisor P&D
daniel.ramos.garcia@managementsolutions.com

Segismundo Jiménez
Supervisor P&D
segismundo.jimenez@managementsolutions.com

Carlos Alonso Viñas
Consultor
carlos.alonso.vinas@msspain.com

Management Solutions, serviços profissionais de consultoria

A Management Solutions é uma firma internacional de serviços de consultoria focada na assessoria de negócio, riscos, finanças, organização e processos

Para mais informações acesse: **www.managementsolutions.com**

Nos siga em:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña London Frankfurt Paris Amsterdam Copenhaguen vOslo Warszawa Zürich Milano Roma Lisboa Beijing New York Boston Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Medellín Bogotá Quito São Paulo Lima Santiago de Chile Buenos Aires