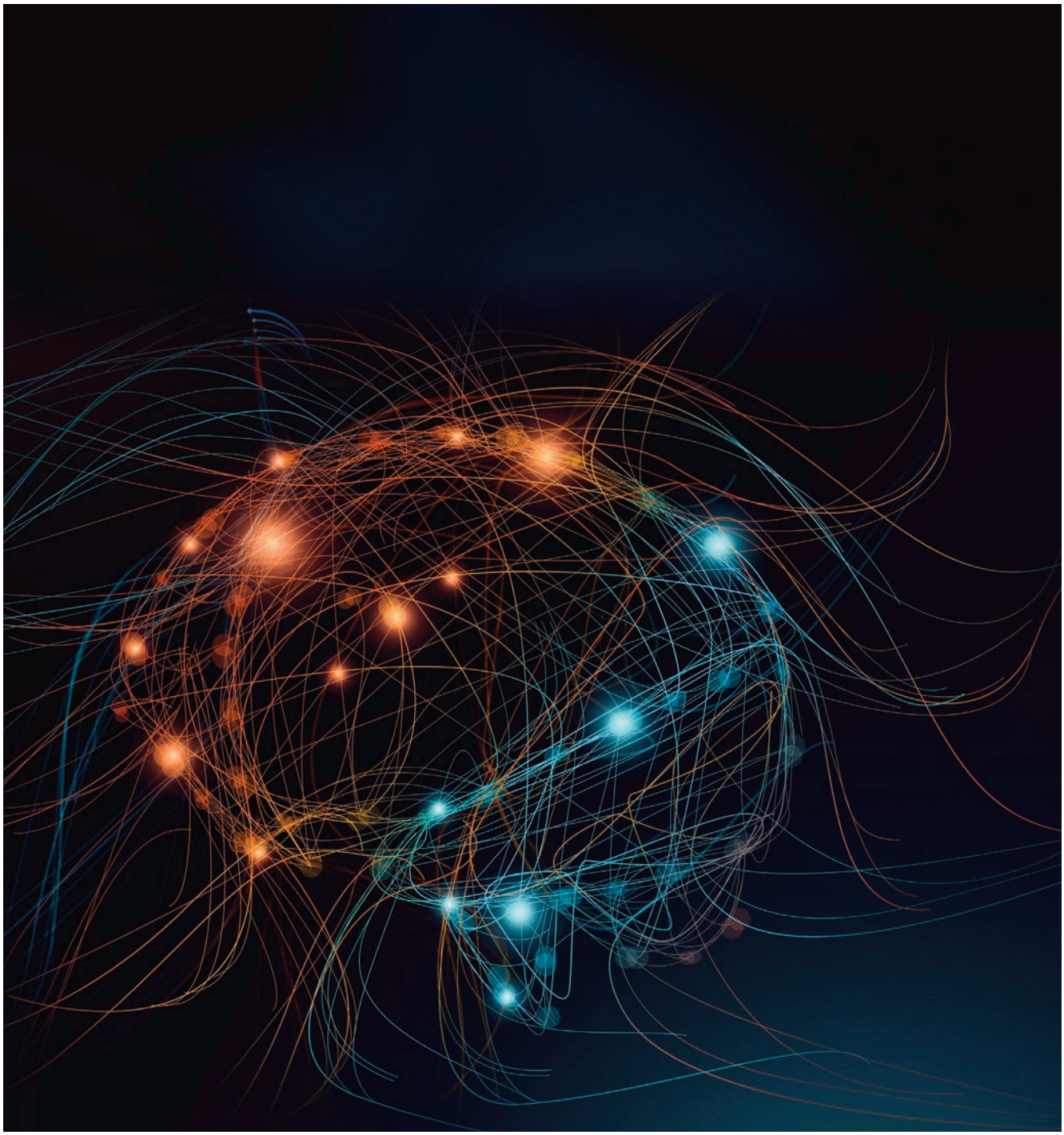
Estudo de caso de interpretabilidade

«Os tolos ignoram a complexidade. Os pragmáticos sofrem com ela. Alguns conseguem evitá-la. Os gênios a eliminam»
Alan Perlis[72]



Abordagem

Esta seção apresenta um estudo de caso de interpretabilidade em inteligência artificial para ilustrar como as técnicas de XAI descritas na seção anterior são aplicadas.

O estudo de caso selecionado aborda o problema da retenção de funcionários em uma organização, concentrando-se em entender e explicar as causas que levam os funcionários a deixar seus empregos. A identificação desses fatores pode permitir que as organizações tomem medidas preventivas e desenvolvam estratégias para melhorar a satisfação no trabalho e a retenção de talentos.

Neste estudo de caso, será usado um conjunto de dados fictício gerado pela IBM e publicado no Kaggle[73]. Esse conjunto de dados contém informações sobre os funcionários de uma organização, incluindo características demográficas, detalhes do cargo e se eles deixaram ou não a empresa.

No exercício atual, a empresa tem uma taxa de fuga de funcionários de 16%, 6% acima da média histórica, e está preocupada em entender as causas para desenvolver um plano de remediação.

As principais variáveis presentes no conjunto de dados incluem:

- Nível de educação (de "secundário" a "doutorado")
- Satisfação com o ambiente de trabalho (de "baixa" a "muito alta")
- Envolvimento no trabalho (de "baixo" a "muito alto")
- Satisfação no trabalho (de "baixa" a "muito alta")
- Classificação de desempenho (de "baixo" a "excelente")
- Satisfação com as relações de trabalho (de "baixa" a "muito alta")
- Equilíbrio entre vida pessoal e profissional (de "ruim" a "ótimo")
- Anos desde a última promoção no emprego (variável numérica)
- Salário mensal (variável numérica)
- Anos no emprego atual (variável numérica)
- Distância até a estação de trabalho (variável numérica)
- Número de empresas nas quais o trabalho foi realizado (variável numérica)
- Cargo atual (variável categórica, inclui "Gerente", "Diretor", "Research Scientist", etc.)

O foco do estudo de caso será treinar e validar diferentes modelos de inteligência artificial para prever o desgaste dos funcionários, usando técnicas de XAI para analisar e entender o comportamento e as decisões dos modelos selecionados.

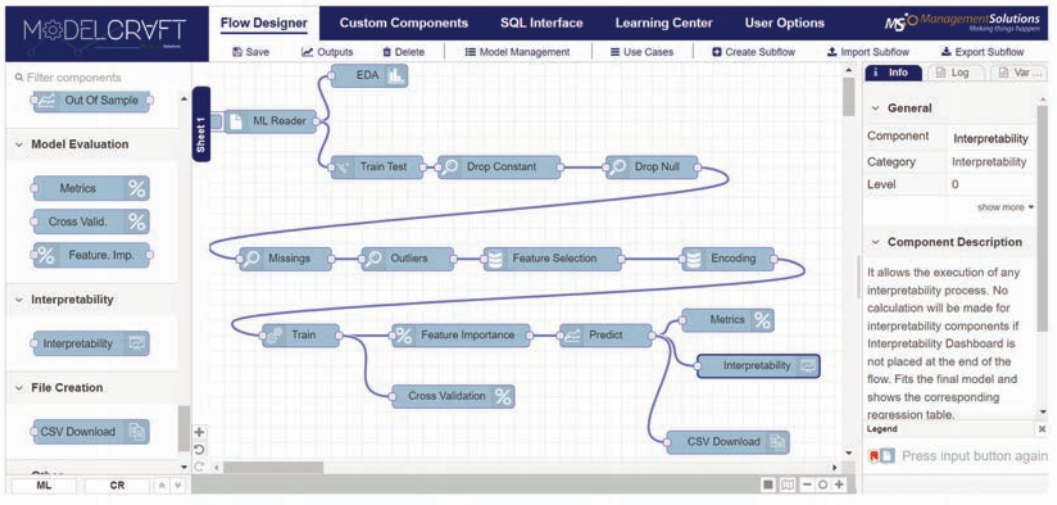
Para simplificar e acelerar o processo, foi usado o sistema de modelagem de componentes ModelCraft™, que contém várias técnicas relevantes de AI e XAI. Esse sistema permitirá que o estudo seja realizado de forma eficiente e sem a necessidade de escrever código.

Ao longo do estudo de caso, as técnicas de interpretabilidade SHAP, LIME e PDP serão aplicadas para analisar os modelos selecionados e entender quais variáveis influenciam as decisões dos funcionários de deixar seus empregos. Além disso, exploraremos como essas variáveis interagem umas com as outras e como elas afetam diferentes segmentos da população de funcionários.

[72] Alan Jay Perlis (1922-1990), cientista da computação americano, PhD em Ciência da Computação pelo MIT e professor da Universidade Purdue, da Universidade Carnegie Mellon e da Universidade da Califórnia em Berkeley, conhecido por seu trabalho pioneiro em linguagens de programação e por ser o primeiro ganhador do Prêmio Turing.

[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figura 13. Fluxo de modelagem no ModelCraft™.



Ao final do estudo de caso, serão avaliadas a eficácia e as limitações das técnicas de interpretabilidade utilizadas. Também será discutido como a combinação de modelos de inteligência artificial e módulos de interpretabilidade pode melhorar a capacidade preditiva e a compreensão dos modelos, facilitando assim a tomada de decisões orientada por dados no ambiente de negócios.

Processo de modelagem

O processo de modelagem é realizado em três fases: engenharia de dados, modelagem e análise de interpretabilidade do modelo.

1. Engenharia de dados

A engenharia de dados é a fase inicial em que o conjunto de dados é preparado e processado para ser usado na criação de modelos de inteligência artificial. Nesse caso, são executadas as seguintes ações:

- Definição do escopo da análise: neste caso, todos os funcionários que estiveram em licença médica nos últimos dois anos são considerados a população.
- Limpeza de dados: a qualidade dos dados é verificada e os registros com informações ausentes ou inconsistentes são removidos ou corrigidos.
- Transformação de variáveis: as variáveis categóricas são convertidas em variáveis numéricas usando técnicas como one-hot encoding ou ordinal encoding. Além disso, as variáveis numéricas são normalizadas ou padronizadas quando necessário.
- Seleção de variáveis: as variáveis mais relevantes para prever o desgaste dos funcionários são identificadas usando técnicas de seleção de variáveis, como correlação de Pearson, significância de recursos em modelos baseados em árvores ou eliminação recursiva de recursos.
- Construção de variáveis: novas variáveis são geradas a partir de variáveis existentes para analisar se elas são melhores preditoras do desgaste dos funcionários, como a "satisfação total", que foi construída como a soma das pontuações das variáveis "Satisfação com o ambiente", "Satisfação com o trabalho", "Avaliação de desempenho", "Equilíbrio entre vida pessoal e profissional", "Envolvimento no trabalho" e "Satisfação com as relações de trabalho".
- Divisão do conjunto de dados: o conjunto de dados é dividido em dois subconjuntos: treinamento e teste. O subconjunto de treinamento é usado para ajustar e otimizar os modelos de inteligência artificial, enquanto o subconjunto de teste é usado para avaliar o desempenho e a capacidade de previsão dos modelos.

2. Desenvolvimento do modelo

Nessa fase, diferentes modelos de inteligência artificial são treinados e validados usando o subconjunto de treinamento. Em particular, vários dos algoritmos de machine learning mais comuns, como regressão logística, árvores de decisão, máquinas de vetor de suporte, redes neurais e random forest, são ajustados e comparados para selecionar o modelo com o melhor desempenho.

Para evitar o treinamento excessivo e otimizar os hiperparâmetros dos modelos, são usadas técnicas de validação cruzada e pesquisa em grade ou aleatória. Além disso, foi dada atenção especial à complexidade do modelo durante o treinamento ao selecionar um algoritmo específico, a fim de facilitar sua interpretação.

Para isso, foi gerado um fluxo de desenvolvimento de modelo no ModelCraft™ (Fig. 13).

Para selecionar o modelo com a melhor capacidade de previsão, seu desempenho no subconjunto de teste é avaliado usando métricas como precisão, sensibilidade, especificidade e área sob a curva ROC (AUC-ROC). Essas métricas nos permitem avaliar a eficácia do modelo selecionado em termos de sua capacidade de prever corretamente o desgaste dos funcionários em dados não vistos anteriormente.

Considerando todos os aspectos, o algoritmo de random forest produz resultados de desempenho superiores, embora represente um desafio de interpretabilidade na compreensão de suas previsões. Esse modelo considerou 300 árvores de decisão e produziu uma precisão de 75% e uma sensibilidade de 84%. Portanto, essas são previsões muito confiáveis e os falsos negativos são raros. Isso é relevante para este estudo de caso, em que a empresa desejaria previsivelmente reduzir esse tipo de erro o máximo possível.

3. Análise de interpretabilidade

Nessa última fase, as técnicas de interpretabilidade são aplicadas para analisar e entender o comportamento e as decisões do modelo selecionado. Especificamente, os objetivos da análise foram:

— Entender quais variáveis são mais importantes na tomada de decisões da empresa em nível global, usando uma comparação por importância de cada variável.

— Entenda como as mudanças nas variáveis mais importantes afetam diferentes faixas populacionais.

— Entender os resultados do modelo em casos específicos em que uma certa probabilidade de desistência é observada.

Neste estudo de caso, as técnicas SHAP, LIME e PDP são usadas para explicar como o modelo toma decisões e como as entradas influenciam as previsões.

O SHAP permite obter resultados de interpretabilidade global, que dão uma interpretação da importância de cada variável, e o LIME permite realizar uma análise intuitiva da interpretabilidade local que permite explicar o resultado do modelo para cada funcionário com base em modelos lineares mais simples. Como complemento, os gráficos PDP permitem visualizar como as alterações em cada variável afetam a previsão do modelo.

Isso resultou na seguinte distribuição da importância de cada variável (Fig. 14).Nesse caso, observa-se que a variável com maior importância na previsão de desligamento (15,65%) é a "satisfação total", um indicador sintético definido como uma média ponderada de seis elementos (ambiente de trabalho, adequação das funções e áreas ao trabalho, classificação interna, conciliação familiar, relacionamento com colegas e supervisores, e cargo e responsabilidade do funcionário).

Esse resultado é intuitivo e mostra que a variável "satisfação total" foi bem projetada. No entanto, as três variáveis seguintes em termos de importância (tempo de serviço, salário e distância de casa ao trabalho) demonstraram ter grande influência na rotatividade de funcionários, que coletivamente é o dobro do indicador "satisfação total".

Para entender como cada variável é influenciada individualmente, os PDPs foram estudados (Fig. 15).

Em termos de tempo de serviço, a tendência se inverte após três anos: os funcionários com tempo de serviço intermediário são, em média, os menos propensos a deixar a empresa. Em relação à satisfação geral, observa-se uma tendência intuitiva: uma maior satisfação relatada em pesquisas internas resulta em uma menor taxa de desligamento.

Para complementar a análise anterior, o LIME foi usado para analisar, caso a caso, os valores das variáveis que influenciam a probabilidade de saída de determinados funcionários. A Fig. 16 mostra dois funcionários com diferentes probabilidades de saída obtidas com o modelo. O LIME mostra uma métrica de

Fig. 14. Interpretabilidade geral do modelo random forest usando SHAP, em que os valores de Shapley são usados para obter a importância das variáveis.

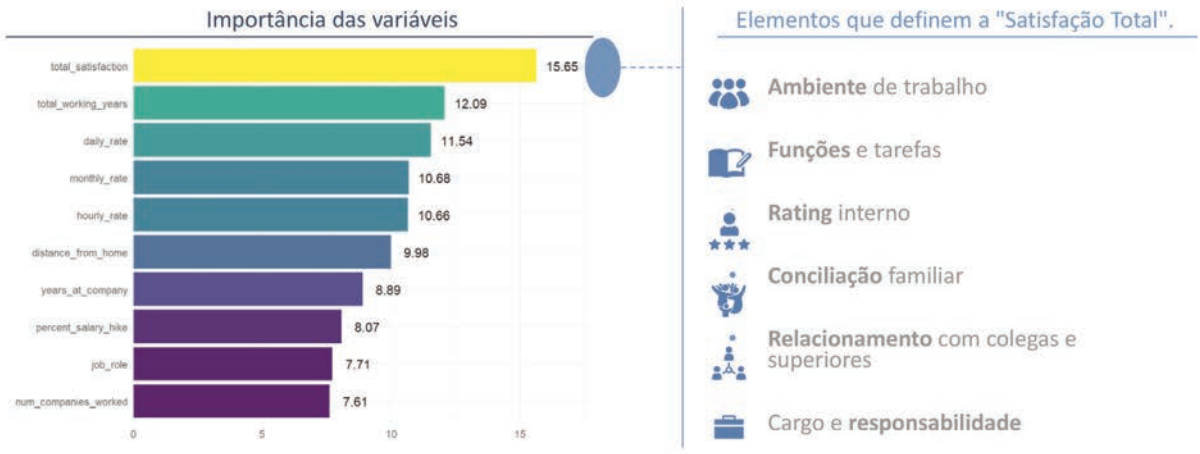
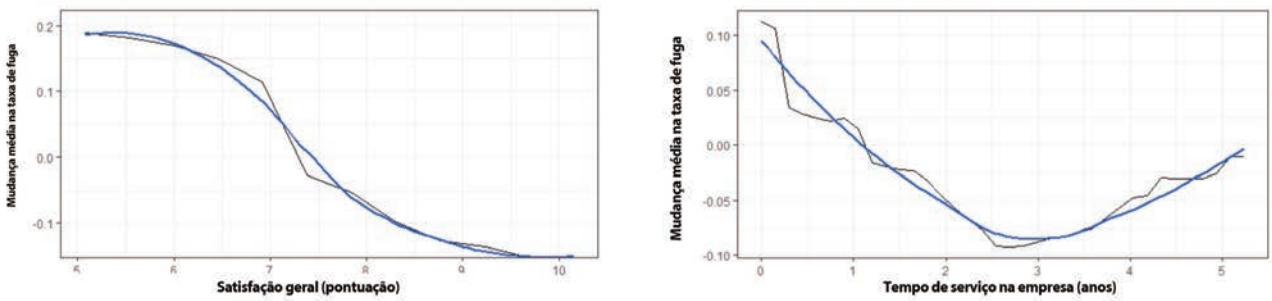


Figura 15. Gráficos PDP para as variáveis "satisfação total" e "tempo de serviço".



explicabilidade que representa a qualidade do ajuste linear obtido usando o modelo substituto local para explicar essas previsões.

Vale ressaltar que as causas mais relevantes de desligamento nesses dois casos não correspondem necessariamente às variáveis mais influentes em nível global. Embora se possa observar que a satisfação total contribui para explicar a probabilidade de saída do funcionário no caso 1, ela não parece ter um impacto significativo no caso 2, em que a probabilidade de saída é maior.

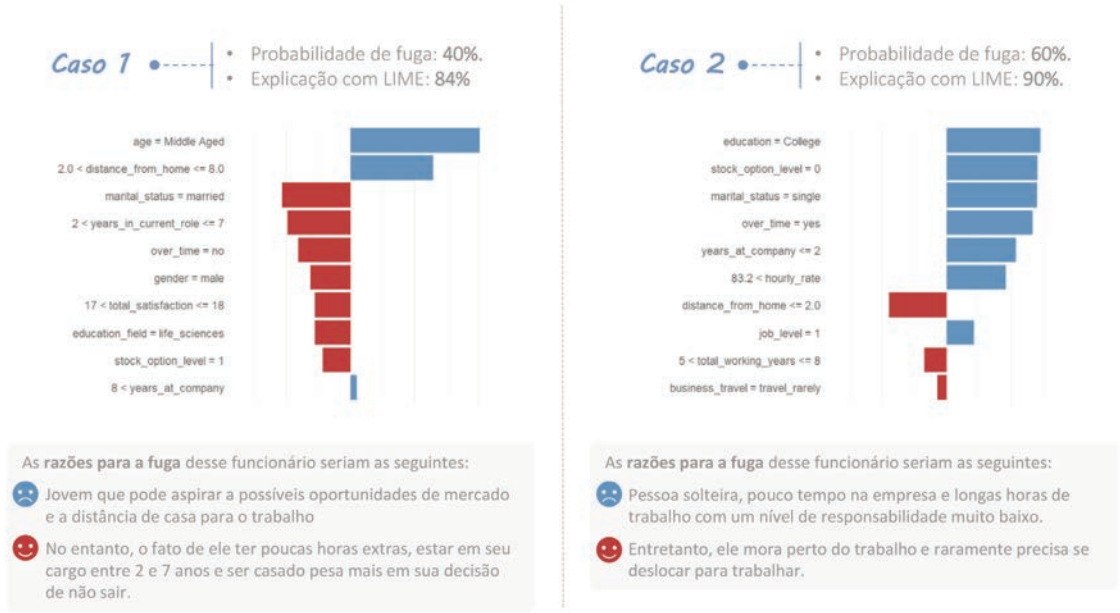
Isso reflete as dificuldades de interpretação desse modelo, que pode ser generalizado para modelos semelhantes: embora a satisfação total possa explicar a probabilidade de desistência em média, essa conclusão é uma generalização; há casos individuais e de grupo em que a desistência é explicada em maior grau por outras variáveis.

Conclusões do estudo de caso

Várias conclusões e lições aprendidas podem ser extraídas do estudo de caso de interpretabilidade de inteligência artificial apresentado, o que pode ser útil em aplicações futuras de modelos de AI e XAI:

- **Aplicação do modelo:** a aplicação e a interpretação corretas do modelo, nesse caso, podem ajudar a prever e evitar o desgaste dos funcionários. Entre os usos que podem ser feitos do modelo está a capacidade de criar diferentes perfis com propensão a sair e de identificar as características desses funcionários com antecedência para

Figura 16. Interpretabilidade local do modelo de random forest usando LIME.



tomar as medidas adequadas, o que, a longo prazo, pode contribuir para reduzir o nível de rotatividade na empresa.

- **Escolha do modelo:** o processo de modelagem mostrou a importância de comparar e validar diferentes algoritmos de machine learning para selecionar o modelo com a melhor capacidade de previsão. Nesse caso, o modelo random forest provou ser o mais adequado para prever a fuga dos funcionários.

- **Importância da interpretabilidade:** a aplicação de técnicas de interpretabilidade, como SHAP, LIME e PDP, proporcionou uma compreensão mais profunda de como o modelo toma decisões e como as entradas influenciam as previsões. Essas informações são cruciais para validar a aplicabilidade do modelo no contexto real e para garantir que as previsões sejam baseadas em recursos relevantes e significativos.

- **Variáveis de influência:** a análise de interpretabilidade identificou as variáveis mais relevantes para prever o desgaste dos funcionários. Essas variáveis podem ser úteis no desenvolvimento de estratégias de retenção e na melhoria da satisfação no trabalho. Além disso, a compreensão de como essas variáveis interagem entre si e como afetam diferentes segmentos da população de funcionários pode enriquecer a análise e facilitar a tomada de decisões com base em dados.

- **Implementação prática:** o estudo de caso demonstra a viabilidade e a utilidade da aplicação de técnicas de AI e XAI em um cenário realista, usando dados fictícios, mas representativos de uma situação comercial. Essa abordagem pode ser adaptada a outros contextos e problemas comerciais, aproveitando a inteligência artificial e a interpretabilidade para melhorar a tomada de decisões e obter resultados mais eficientes e eficazes.

- **Limitações:** ao mesmo tempo, esse caso de uso destacou as limitações e dificuldades na aplicação de técnicas de interpretabilidade post-hoc. É importante reconhecer que os métodos de interpretabilidade não são infalíveis e, às vezes, podem apresentar resultados aproximados ou parciais. Portanto, é essencial aplicar uma abordagem crítica e rigorosa ao interpretar e validar os resultados das técnicas de interpretabilidade.

- **Combinação de modelos de AI e módulos de interpretabilidade:** este estudo de caso mostra como a integração de modelos de AI e módulos de interpretabilidade pode melhorar a capacidade de previsão e a compreensão dos modelos. Isso facilita a adoção de soluções baseadas em AI na tomada de decisões de negócios.

- **Continuidade na análise de interpretabilidade:** por fim, deve-se enfatizar que a análise de interpretabilidade não deve ser um exercício único aplicado durante o desenvolvimento do modelo, mas deve ser realizada de forma contínua, reproduzível e confiável durante toda a vida útil do modelo.

Concluindo, este estudo de caso de interpretabilidade em inteligência artificial proporcionou uma experiência valiosa na aplicação de técnicas de AI e XAI em um contexto comercial e mostrou o potencial da AI e da interpretabilidade para aprimorar a tomada de decisões, revelando, ao mesmo tempo, as limitações e dificuldades associadas a essas técnicas e a necessidade de uma abordagem crítica e rigorosa ao interpretar e validar os resultados da AI.