Explainable AI: como tornar modelos complexos compreensíveis

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Por que precisamos de modelos explicáveis?

melhor desempenho ≠ melhor entendimento

Se você não consegue explicar, o modelo serve pra quem?

Por que precisamos de modelos explicáveis?

Do ponto de vista do analista de dados:

- **Modelos são cada vez mais complexos:** algoritmos como XGBoost, redes neurais e ensembles oferecem alto desempenho, mas são verdadeiras "caixas-pretas"
- **Facilita o diagnóstico e melhoria do modelo:** entender quais variáveis mais influenciam o resultado permite detectar erros, ajustar o modelo e construir melhores features.
- **Evita decisões erradas:** se uma variável irrelevante estiver influenciando muito, o analista pode detectar isso com SHAP ou LIME.
- **Mais confiança no que entrega:** você não está apenas dizendo "o modelo prevê bem", mas mostrando o porquê.

Por que precisamos de modelos explicáveis?

Do ponto de vista do stakeholder / tomador de decisão:

- **Precisa confiar para agir:** não basta saber que “o modelo prevê churn com 89% de acurácia”. O stakeholder quer saber: "Quem são os clientes que vão sair? Por que eles vão sair? O que eu posso fazer?"
- **Impacta decisões críticas:** em crédito, saúde, RH ou precificação, previsões sem explicação são arriscadas.
- **Permite alinhamento estratégico:** entender os principais fatores que levam a uma previsão ajuda na criação de planos de ação
- **Ajuda a contar a história certa:** o stakeholder consegue levar o insight para frente para comitês, investidores, clientes com clareza e segurança.

O que é Explainable AI (XAI)?

Conceito: conjunto de técnicas que tornam previsões de IA interpretáveis.

- Interpretabilidade: transparência do modelo (ex: regressão linear).
- Explicabilidade: capacidade de explicar um modelo não transparente (ex: XGBoost).

Benefícios:

- Confiabilidade
- Diagnóstico
- Aderência a compliance
- Confiança de stakeholders

Níveis de Explicabilidade

1. Explicabilidade Global (Global Explainability)

O que é: Entendimento do comportamento geral do modelo. Ou seja, como o modelo se comporta em média, quais variáveis ele mais considera, como ele toma decisões em geral.

Para que serve:

- Avaliar se o modelo está aprendendo padrões corretos.
- Comunicar fatores de influência para áreas de negócio.
- Garantir compliance e alinhamento com valores éticos (ex: evitar viés sistêmico)

Exemplo prático: Em um modelo de previsão de churn, descobrir que “tempo de contrato” e “número de chamadas no suporte” são os principais preditores em média, no conjunto todo.

Níveis de Explicabilidade

2. Explicabilidade Local (Local Explainability)

O que é: Explicação da previsão para um indivíduo específico. Não responde "como o modelo funciona", mas "por que essa previsão foi feita".

Para que serve:

- Tomar decisão individual (ex: aprovar crédito ou intervir num cliente específico).
- Justificar uma decisão com base em evidências.
- Fazer debugging de previsões fora do esperado.

Exemplo prático: Um cliente específico foi classificado com alto risco de churn. O SHAP mostra que isso aconteceu porque ele está há pouco tempo na base, teve 3 chamados no mês e não usou o app nos últimos 10 dias.

Técnicas de Interpretação

1. Feature Importance (Importância das Variáveis)

- 🔹 O que é: Mede quanto cada variável contribui para as previsões do modelo. Pode ser global (em média) ou local (por instância).
- 🔹 Como funciona:
  - Em modelos de árvore (Random Forest, XGBoost): mede a redução de impureza.
  - Com permutation importance: vê o impacto de embaralhar uma variável no desempenho do modelo.
- 🔹 Uso prático: Entender se "renda" influencia mais que "idade" na concessão de crédito, por exemplo.

Técnicas de Interpretação

2. SHAP (SHapley Additive exPlanations)

- 🔸 **O que é:** Explica a contribuição de cada variável para a previsão feita (local ou global). Baseado em Teoria dos Jogos de Shapley.

- 🔸 **Diferencial:**
  - Mostra o efeito positivo ou negativo de cada feature.
  - Permite entender por que aquele valor foi previsto.

- 🔸 **Tipos de visualizações:**
  - Summary plot: visão global
  - Force plot: explicação individual
  - Waterfall plot: decomposição da previsão

- 🔸 **Uso prático:** O SHAP mostra que o cliente foi classificado como risco porque teve muitas reclamações e está há pouco tempo na empresa.

Técnicas de Interpretação

3. LIME (Local Interpretable Model-Agnostic Explanations)

- 🔹 O que é: Explica a decisão do modelo para um único exemplo, usando um modelo interpretable próximo (como uma regressão linear simples).

- 🔹 Como funciona:
  - Cria perturbações no ponto analisado.
  - Treina um modelo linear local com essas variações.
  - Estima os pesos das variáveis para aquela decisão.

- 🔹 Uso prático: Entender por que uma imagem foi classificada como "cão" ou um texto foi considerado "spam".

- 🔹 Limitações: Menos estável que SHAP. Foca mais em "como" do que em "por que".

SHAP vs LIME

Critério	SHAP	LIME
Tipo de explicação	Local + Global	Local
Estabilidade	Alta	Média (pode variar)
Interpretação	Mais robusta	Mais leve e rápida
Visualização	Variada e rica	Simples (tabelas, texto)
Tempo de processamento	Maior	Menor

Exemplos de XAI: Finanças 💰

Situação: Um cliente teve um pedido de empréstimo negado por um modelo de scoring baseado em XGBoost.

Explicação com SHAP: O modelo mostra que as variáveis que mais contribuíram para a negativa foram:

- Renda mensal baixa (-20 pontos)
- Histórico de inadimplência (-35 pontos)
- Tempo de emprego < 6 meses (-10 pontos)

Impacto: Ajuda o cliente a entender o que precisa melhorar para ser aprovado no futuro, e a instituição cumpre a exigência regulatória de transparência.

Exemplos de XAI: Varejo🛒

Situação: Um cliente foi classificado como alto risco de churn (cancelamento de assinatura) em uma plataforma de streaming.

Explicação com LIME ou Feature Importance: Principais variáveis para o churn:

- Tempo sem assistir nenhum conteúdo > 30 dias
- Não interagiu com sugestões personalizadas
- Cancelou uma renovação automática anterior

Impacto: A plataforma consegue acionar uma oferta personalizada e mostra:

> *"Notamos que você não consome conteúdo há um tempo. Usuários com esse padrão normalmente deixam a plataforma, por isso estamos oferecendo um benefício especial."*

Gera ação direta para retenção e entendimento do comportamento do cliente.

Exemplos de XAI: Saúde 🏥

Situação: Uma IA hospitalar gera um alerta de risco de sepse em um paciente internado.

Explicação com SHAP: Variáveis que mais contribuíram para o alerta:

- Frequência cardíaca elevada
- Temperatura corporal acima de 38,5°C
- Contagem de leucócitos alta
- Pressão arterial baixa

Impacto: Um painel com visualização SHAP mostra o impacto individual de cada indicador. O médico vê que a IA "puxou" o alarme por alterações rápidas nos sinais vitais, combinadas com exames laboratoriais anormais.
Aumenta a confiança do médico na IA e possibilita ação clínica rápida e justificada.

Exemplos de XAI: RH👥

Situação: Um modelo de IA não selecionou um candidato para a próxima etapa.

Explicação com LIME: Variáveis que mais contribuíram:

- Falta de experiência na ferramenta-chave
- Desalinhamento com a pontuação de fit cultural
- Nota técnica abaixo da média nas provas online

Impacto: Ao oferecer um feedback automatizado, a empresa pode mostrar: “A IA identificou lacunas técnicas e culturais que não atendem aos requisitos da vaga. Recomendamos estudar as ferramentas X e Y para futuras candidaturas.”
Transparência no processo seletivo, evita viés oculto e melhora a percepção do candidato.

Como comunicar explicações para o time de negócio

- Evitar termos técnicos (shapley, log-odds, etc.)
- Dizer: "O cliente X foi classificado como risco alto porque ele teve atraso recente e tem baixo histórico de compras."
- Criar storytelling com gráficos simples
- Usar heatmaps, rankings, frases do tipo “as 3 variáveis mais relevantes foram…”

Cuidados ao usar XAI

- Explicabilidade não é justificativa: pode gerar falsa confiança
- Importância ≠ causalidade
- Ferramentas diferentes geram interpretações diferentes
- XAI não substitui entender o problema de negócio

Conclusão

- **Checklist para aplicar XAI no seu projeto:**
  - **Seu modelo é complexo?**
  - **O negócio exige explicações?**
  - **Há impacto direto na decisão?**