

Introdução a Explicabilidade em Machine Learning

Marcos O Prates

18 de Fevereiro de 2025

Quem sou eu?

Marcos Oliveira Prates

Professor Associado

Departamento de Estatística - UFMG

Áreas de Pesquisa:

- Aprendizado de Máquina
- Estatística Espacial
- Métodos Bayesianos

Contato:

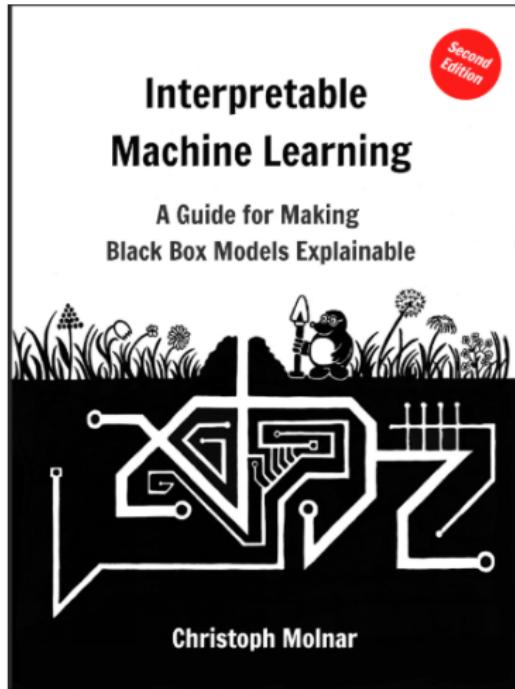
Email: marcosop@est.ufmg.br

Página: www.est.ufmg.br/marcosop



Principal Referência

Christoph Molnar. *Interpretable Machine Learning*. Disponível em:
<https://christophm.github.io/interpretable-ml-book/>
(Molnar, 2020)



O que é explicabilidade em ML e por que é importante



O que é explicabilidade em ML e por que é importante

- ▶ Em um artigo pioneiro, Breiman (2001) balançou a comunidade estatística apontando que os métodos estatísticos tradicionais não são a única forma de se aprender a partir dos dados:
 - ▶ Cultura voltada aos dados (Estatística tradicional):
 - ▶ Regressão Linear, Regressão Logística, Modelos Aditivos, etc.
 - ▶ Eles permitem interpretar como a variável resposta está associada as variáveis de entrada. **Modelos Transparentes**.

O que é explicabilidade em ML e por que é importante

- ▶ Cultura voltada a algoritmos (Aprendizado de Máquina, Ciência de Dados):
 - ▶ Redes Neurais, Florestas Aleatórias, Máquina de Vetores de Suporte ,etc.
 - ▶ Possuem grande capacidade preditiva que comumente performam melhores que os métodos estatística tradicionais nessa característica. **Modelos Caixa Preta**.
- ▶ Isso criou uma aparente separação: capacidade preditiva vs explicabilidade.
- ▶ Breiman reivindicou procedimentos que permitissem uma melhor interpretação do resultados de modelos algorítmicos, sem abrir mão de sua capacidade preditiva.

Problemas Legais



Pedro Domingos
@pmddomingos

Follow



Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

7:59 PM - 28 Jan 2018

188 Retweets 312 Likes



41

188

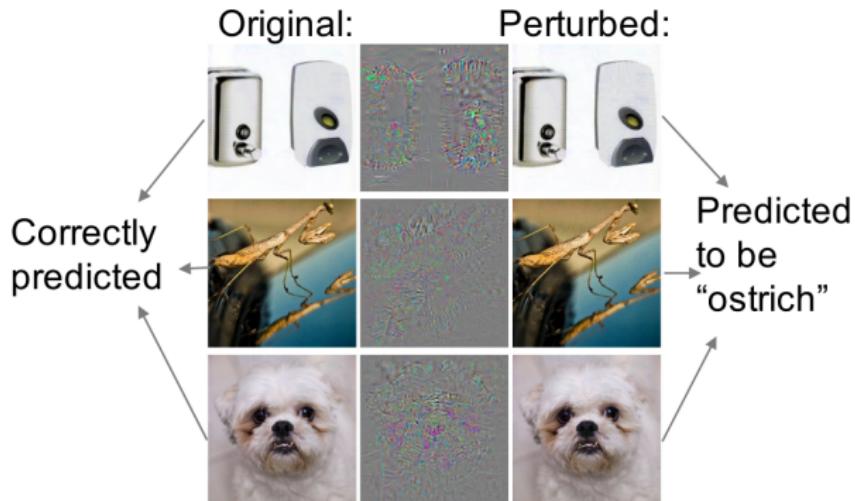
312



Em geral, temos alguns problemas fundamentais com IA

- ▶ Não confiamos no modelo;
- ▶ Não confiamos o que acontece em casos extremos;
- ▶ Erros podem ser danosos/caros;
- ▶ Como corrigir o modelo se algo der errado?

Exemplo de erro (Szegedy et al., 2014)



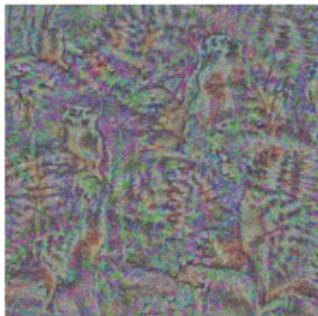
Como entender as máquinas (Zhang et al., 2023)?



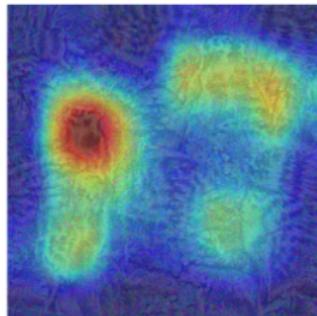
"meerkat" (0.71)



attention map



"meerkat" (0.99)



attention map

Quando não queremos/precisamos de explicação?

- ▶ Sem consequências significativas ou quando as previsões são tudo que você precisa;
- ▶ Problema suficientemente bem estudado;
- ▶ Evite manipular o sistema - objetivos incompatíveis.

O que é explicabilidade?

Não existe uma definição padrão

- ▶ A maioria concorda que é algo diferente de performance;
- ▶ Habilidade de explicar ou apresentar o modelo à um humano (Doshi-Velez and Kim, 2017);
- ▶ Visão cínica – É o que faz você se sentir bem o modelo;
- ▶ Realmente depende do público-alvo.

Explicabilidade pode ser uma maneira de tentar abordar estes problemas.

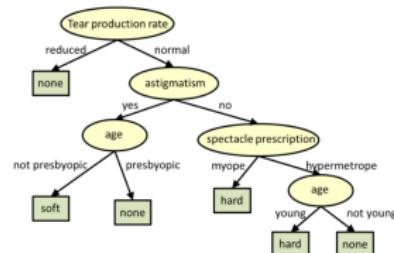
Como era pensada a explicabilidade

Nos modelos de pré-aprendizagem profunda, alguns modelos são considerados “interpretáveis”

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent Variable → Population Y intercept → Population Slope Coefficient → Independent Variable → Random Error term

Linear component → Random Error component

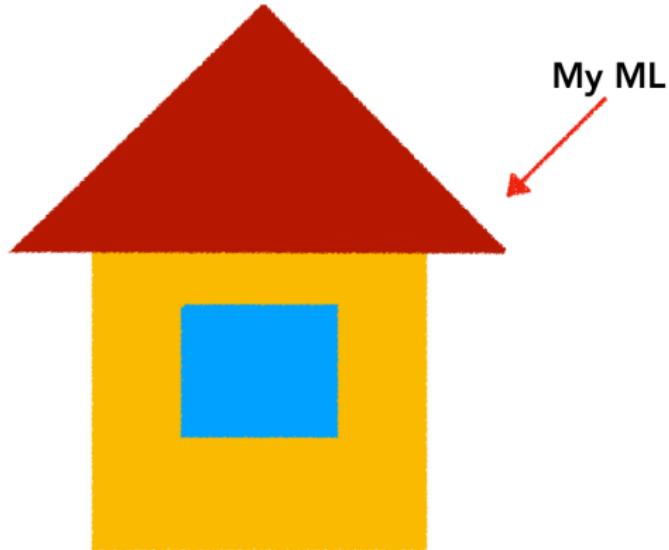


Algumas propriedade de explicabilidade

- ▶ Fidelidade - como fornecer explicações que representem com precisão o verdadeiro raciocínio por trás da decisão final do modelo.
- ▶ Plausibilidade – A explicação está correta ou algo em que podemos acreditar ser verdade, dada a nossa conhecimento atual do problema?
- ▶ Compreensível – Posso colocar em termos que o usuário final sem conhecimento profundo do sistema pode entender?
- ▶ Estabilidade – Instâncias semelhantes têm interpretações semelhantes?

Alguns métodos para explicar modelos

Passos da explicabilidade



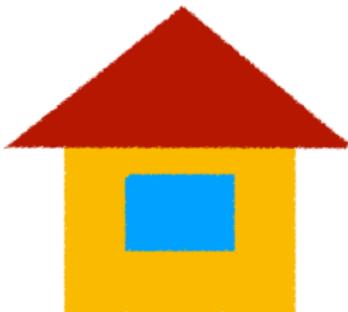
**Before building
any model**



**Building
a new model**



**After
building a model**



Before building
any model



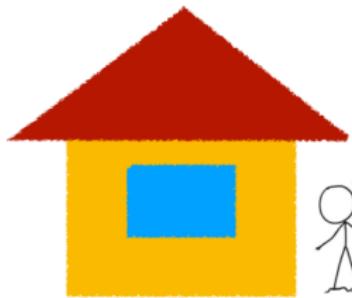
Building
a new model



After
building a model



Common misunderstanding:
Interpretability is always about
machine learning models.



My ML



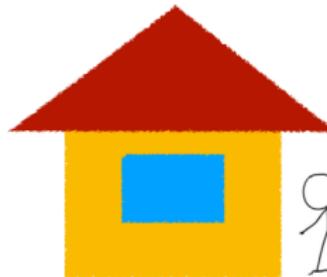
Before building
any model



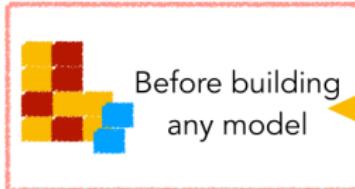
Building
a new model



After
building a model



My ML



Before building
any model

Exploratory data analysis
(e.g, Visualization)



Building
a new model

After
building a model

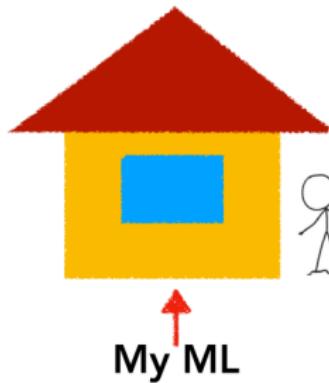


Before building
any model

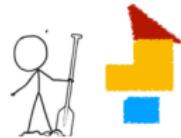


What is the medium and
constraints we use to
explain?

Rules, Examples, Sparsity
and Monotonicity



Before building
any model



Building
a new model



Ablation test

Input-feature importance

Concept importance

Explicabilidade Global vs Local

Global

- ▶ As informações sobre o desempenho global referem-se à determinação qual é o papel de cada variável explicativa na previsão processo sobre todo o suporte das variáveis explicativas.
- ▶ Interpretabilidade global: Medidas de importância variável ou relevância.

Explicabilidade Global vs Local

Local

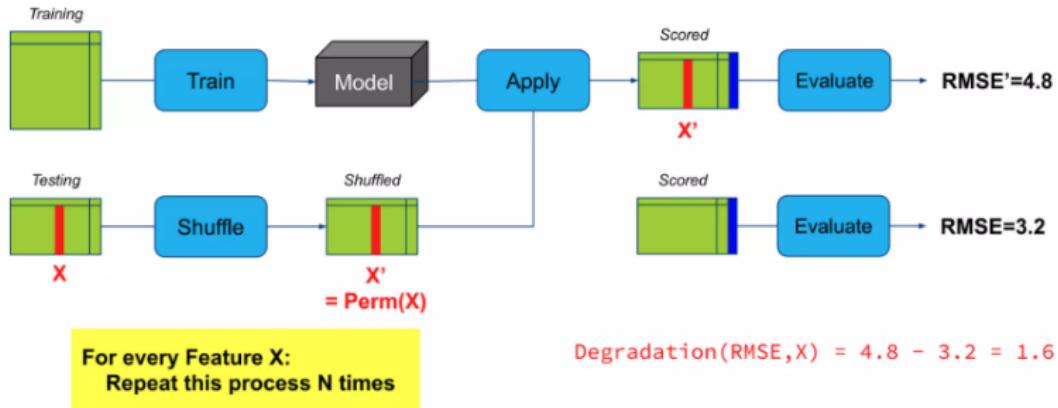
- ▶ Por outro lado, o objetivo de compreender o desempenho local é para fornecer uma explicação significativa de por que o algoritmo retorna um certa previsão, dada uma combinação particular da previsão valores de variáveis.
 - ▶ Interpretabilidade local: por que o modelo de previsão faz uma determinada previsão para um determinado indivíduo?

O aspecto local da interpretabilidade está diretamente relacionado com a experiência dos usuários.

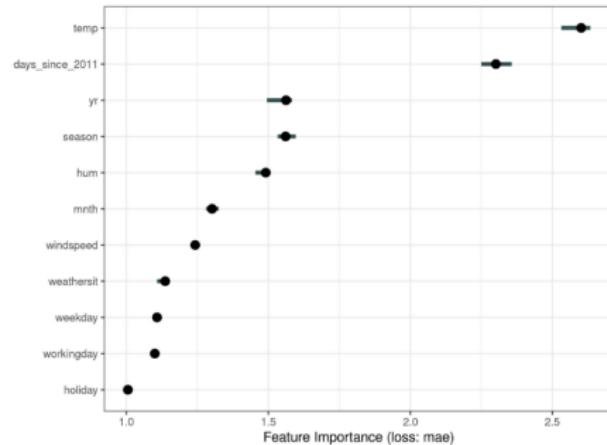
Exemplo: direito à explicação defendido pelo GDPR da UE.

Métodos Globais

Importância do Input



Importância de Inputs



- ▶ Estima a importância do atributo pela mudança de desempenho do algoritmo;
- ▶ Depende de alguma função perda;
- ▶ Cara para ser aplicado em domínios de alta dimensão.

1

¹O conjunto de dados usado nos exemplos contém a contagens diárias de bicicletas alugadas na locadora de bicicletas Capital-Bikeshare, juntamente com informações meteorológicas e sazonais.

Gráficos de dependência parcial (PDPs)

- ▶ Gráficos que mostram a relação marginal entre uma ou mais variáveis independentes e a previsão de um modelo.
- ▶ Utilizados para interpretar o comportamento global do modelo.

Definição

Para uma variável x_S :

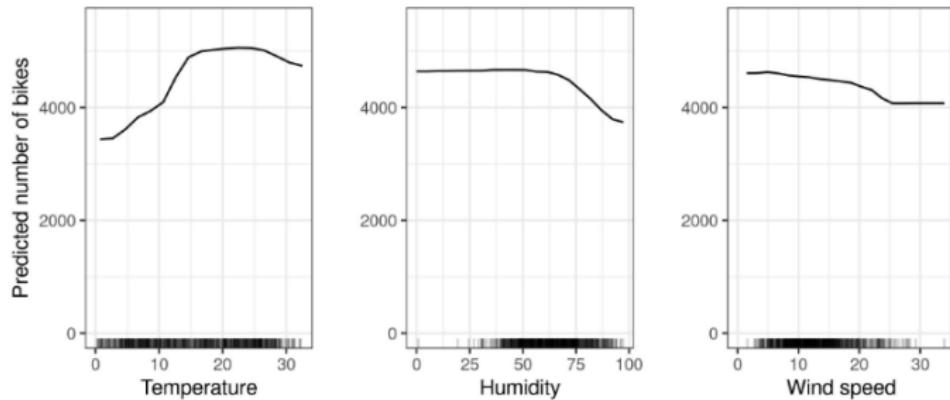
$$\text{PDP}(x_S) = \frac{1}{n} \sum_{i=1}^n f(x_S, x_C^{(i)})$$

Onde:

- ▶ f é o modelo.
- ▶ x_S é o subconjunto de variáveis de interesse.
- ▶ x_C são as demais variáveis do conjunto de dados.

Gráficos de dependência parcial (PDPs)

- ▶ A função parcial nos diz, para determinados valores dos inputs S , qual é o efeito marginal médio na previsão.
- ▶ PDP plano indica que o input não é importante.
- ▶ A importância do recurso baseado em PDP deve ser interpretada com cuidado. Ele captura apenas o efeito principal do recurso e ignora possíveis interações entre os inputs.



Gráficos de Efeitos Locais Acumulados (ALEs)

- ▶ Os PDPs assumem que as variáveis são independentes. Isso pode ser pouco realista.
- ▶ Os ALEs são uma alternativa aos PDPs para lidar com variáveis correlacionadas.
- ▶ Calculam os efeitos acumulados em pequenas janelas do espaço das variáveis.

Definição

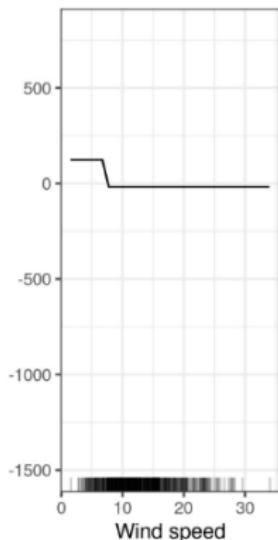
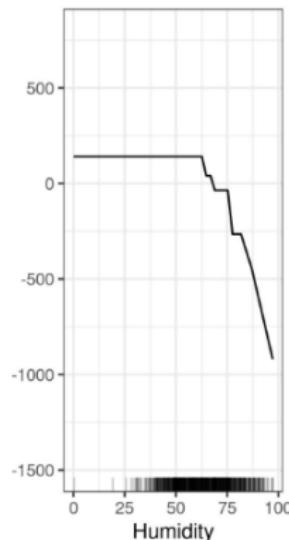
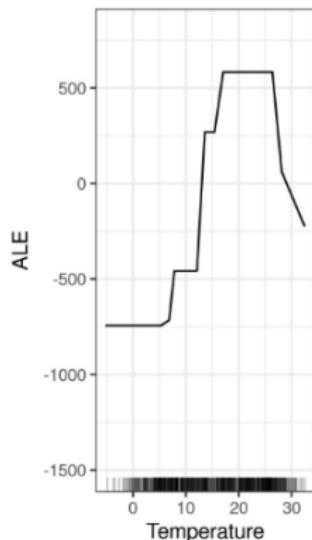
Para uma variável x_j :

$$\text{ALE}_j(x) = \int_{x_{\min}}^x \mathbb{E}[f(x_j, x_{-j}) | x_j = z] dz$$

Onde x_{-j} representa todas as variáveis exceto x_j .

Exemplo de ALE Plot

- ▶ Em resumo, para estimar os efeitos locais, dividimos o input em vários intervalos e calculamos as diferenças nas previsões do modelo.
- ▶ Basicamente, o método ALE calcula as diferenças nas previsões, por meio das quais substituímos o input de interesse pelos valores da grade z.



Resumo sobre os métodos

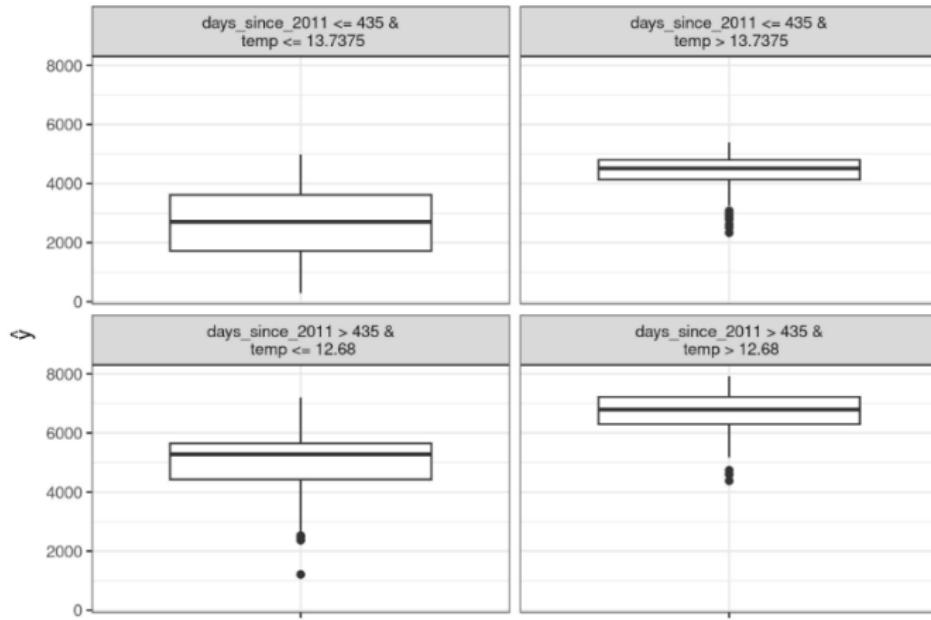
- ▶ Importância do input: “A importância do input é o aumento no erro do modelo quando as informações do input são destruídas.”
- ▶ Gráficos de dependência parcial: “Deixe-me mostrar o que o modelo prevê em média quando cada instância de dados tem o valor v para esse input. Ignoro se o valor v faz sentido para todas as instâncias de dados.”
- ▶ Gráficos ALE: “Deixe-me mostrar como as previsões do modelo mudam em uma pequena “janela” do input em torno de v para instâncias de dados nessa janela.”

Substitutos Globais

- ▶ Um modelo substituto global é um modelo interpretável treinado para aproximar as previsões de um modelo de caixa preta.
- ▶ O objetivo dos modelos substitutos (interpretáveis) é aproximar as previsões do modelo subjacente com a maior precisão possível e, ao mesmo tempo, ser interpretáveis.
- ▶ Uma forma simples de medir quão bem o substituto replica o modelo de caixa preta é a medida R-quadrado

Substitutos Globais

- ▶ Por exemplo, treinamos um substituto (uma árvore de decisão CART) como modelo interpretável para aproximar o comportamento da máquina de vetores de suporte.



Métodos Locais

Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016)

- ▶ Ferramenta para explicação local de previsões de modelos.
- ▶ Modelos substitutos locais: explicam uma previsão substituindo o modelo complexo por um modelo substituto interpretável localmente.
- ▶ Usa modelos lineares simples para aproximar a decisão do modelo em torno de uma instância específica.

Definição Formal

Minimiza:

$$\mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Onde:

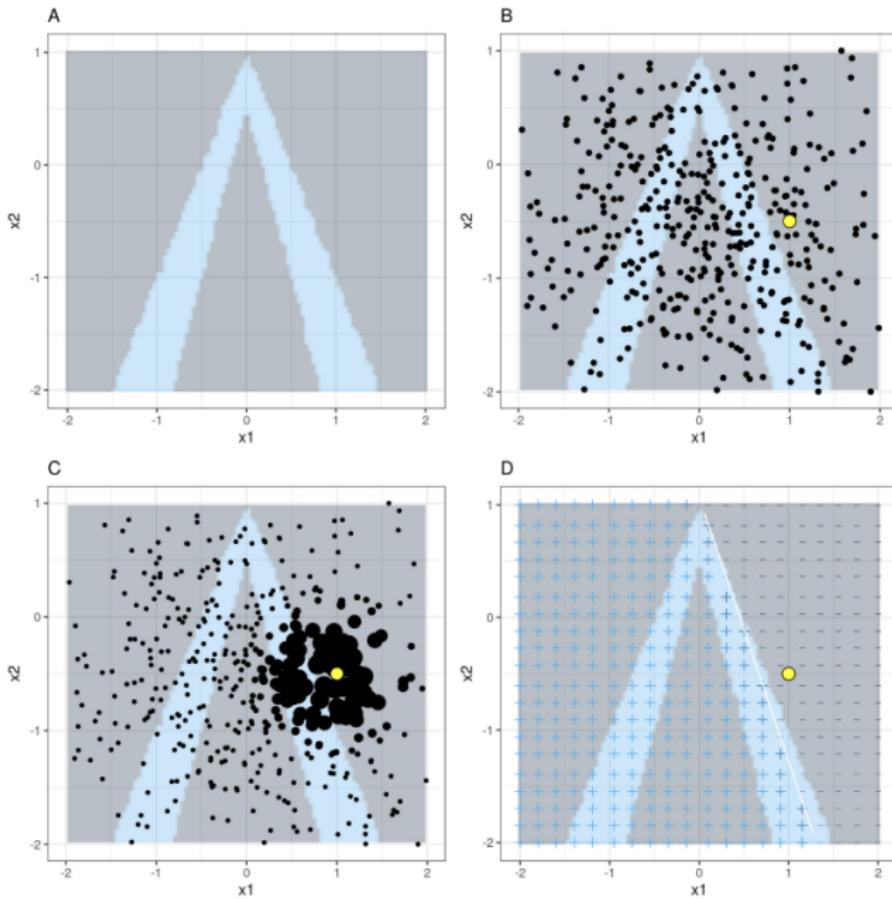
- ▶ \mathcal{L} é a função de perda.
- ▶ f é o modelo original.
- ▶ g é o modelo explicativo local.
- ▶ π_x é a proximidade entre instâncias.
- ▶ Ω penaliza a complexidade de g .

A receita para o LIME

- ▶ Selecione sua instância de interesse para a qual você deseja ter uma explicação de sua previsão de caixa preta.
- ▶ Perturbe seu conjunto de dados e obtenha as previsões de caixa preta para esses novos pontos.
- ▶ Pondere as novas amostras de acordo com sua proximidade à instância de interesse.
- ▶ Treine um modelo ponderado e interpretável no conjunto de dados com as variações.
- ▶ Explique a previsão interpretando o modelo local.

LIME é um dos poucos métodos que funciona para dados tabulares, texto e imagens.

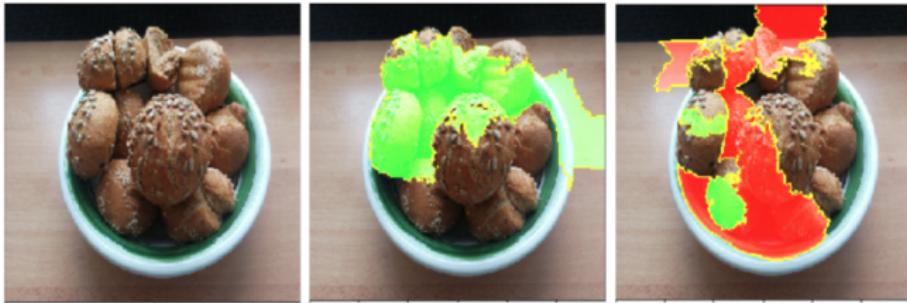
Exemplo de LIME



Exemplo de LIME

- ▶ Neste exemplo, olhamos para uma classificação feita pela rede neural Inception V3. A imagem usada mostra alguns pães que estão em uma tigela.
- ▶ Como podemos ter vários rótulos previstos por imagem (classificados por probabilidade), podemos explicar os rótulos principais. A previsão principal é “Bagel” com uma probabilidade de 77%, seguida por “Morango” com uma probabilidade de 4%.
- ▶ As imagens a seguir mostram para “Bagel” e “Morango” as explicações LIME.

Exemplo de LIME

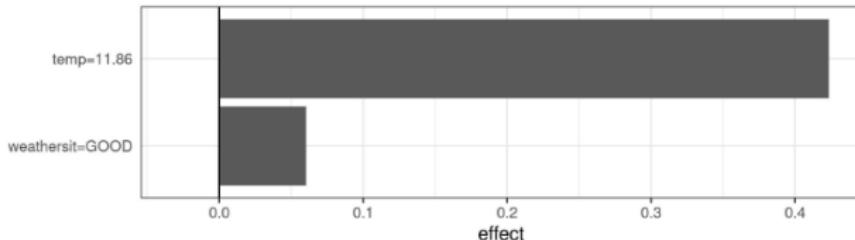


Exemplo de LIME

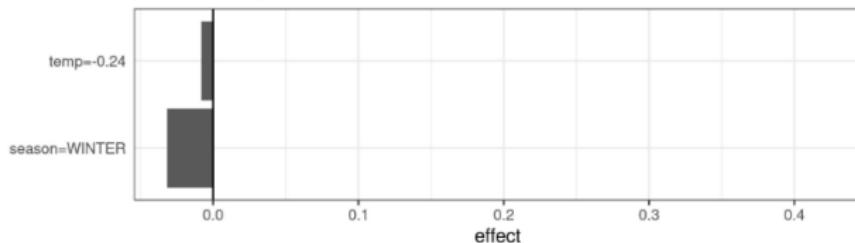
- ▶ Depois de levar em conta a tendência de que o aluguel de bicicletas se tornou mais popular ao longo do tempo, queremos saber em um determinado dia se o número de bicicletas alugadas estará acima ou abaixo da linha de tendência.
- ▶ Primeiro, treinamos uma floresta aleatória com 100 árvores na tarefa de classificação. Em que dia o número de bicicletas alugadas estará acima da média sem tendência, com base nas informações do clima e do calendário?
- ▶ As explicações são criadas com 2 inputs. Os resultados dos modelos lineares locais esparsos treinados para duas instâncias com diferentes classes previstas:

Exemplo de LIME

Actual prediction: 0.89
LocalModel prediction: 0.44



Actual prediction: 0.01
LocalModel prediction: -0.03



SHapley Additive exPlanations (SHAP, Lundberg, 2017)

- ▶ Explicação baseada na teoria dos valores de Shapley (Shapley, 1953).
- ▶ Atribui contribuições justas de cada variável a uma previsão específica.

Definição Formal

Valor de Shapley para a variável j :

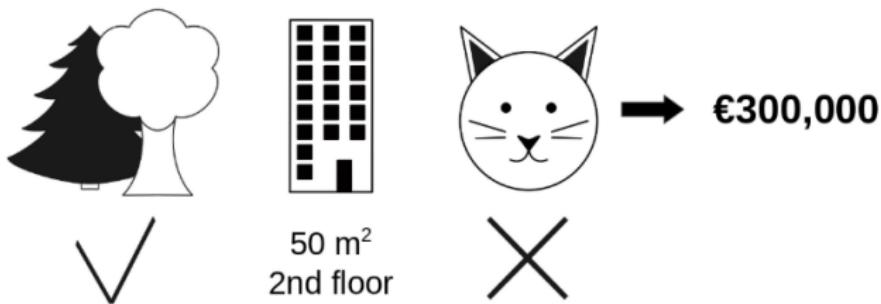
$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{j\}) - v(S)]$$

Onde:

- ▶ N é o conjunto de todas as variáveis.
- ▶ $v(S)$ é a função de valor do modelo para o subconjunto S .

Intuição SHAP

- ▶ A definição formal vem da teoria de jogos e é complexa. Vamos tentar entender a intuição do que é feito.
- ▶ Você treinou um modelo de aprendizado de máquina para prever preços de apartamentos. Para um certo apartamento, ele prevê €300.000 e você precisa explicar essa previsão. O apartamento tem uma área de $50m^2$, está localizado no 2º andar, tem um parque próximo e gatos são proibidos:

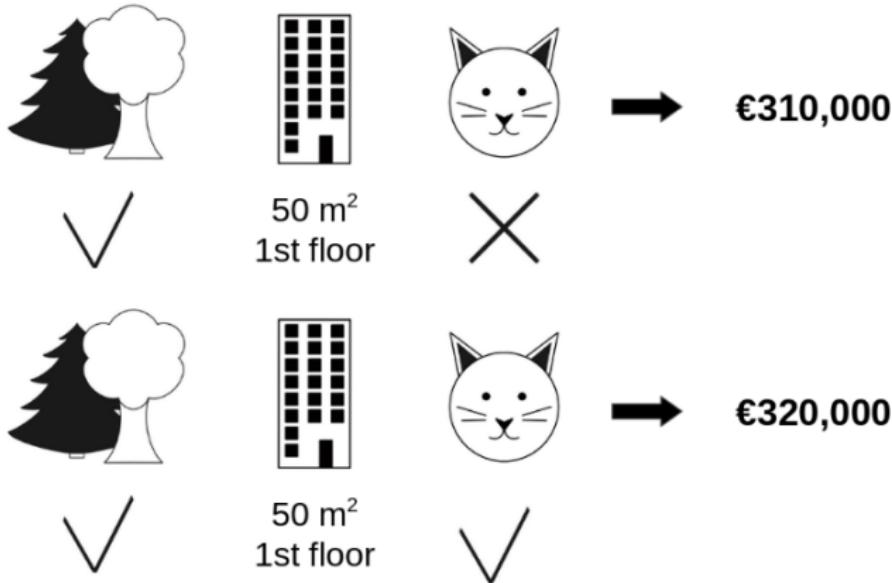


- ▶ A previsão média para todos os apartamentos é de €310.000. Quanto cada valor de característica contribuiu para a previsão em comparação com a previsão média?
- ▶ Jogadores? Jogo? Pagamento? Qual é a conexão com as previsões de aprendizado de máquina e interpretabilidade?
- ▶ O “jogo” é a tarefa de previsão para uma única instância do conjunto de dados.
- ▶ O “ganho” é a previsão real para esta instância menos a previsão média para todas as instâncias.
- ▶ Os “jogadores” são os valores de recursos da instância que colaboraram para receber o ganho (= prever um determinado valor).

- ▶ Em nosso exemplo de apartamento, os valores de recursos park-nearby, cat-banned, area-50 e floor-2nd trabalharam juntos para atingir a previsão de €300.000.
- ▶ Nosso objetivo é explicar a diferença entre a previsão real (€300.000) e a previsão média (€310.000): uma diferença de -€10.000.
- ▶ Uma possível resposta é: O parque-próximo contribuiu com €30.000; área-50 contribuiu com €10.000; andar-2º contribuiu com €0; gato-proibido contribuiu com -€50.000. As contribuições somam -€10.000, a previsão final menos o preço médio previsto do apartamento.

Como calculamos o valor de Shapley para um input?

- ▶ O valor de Shapley é a contribuição marginal média de um valor de característica em todas as coalizões possíveis. **Está claro?**



Esse processo precisa ser feito para todas as combinações



$$\left\{ \begin{array}{c} \text{Tree} \\ \checkmark \end{array} \right\}$$

$$\left\{ 50m^2 \right\}$$

$$\left\{ \begin{array}{c} \text{Building} \\ \text{2nd floor} \end{array} \right\}$$

$$\left\{ \begin{array}{c} \text{Tree} \\ \checkmark \end{array} \quad 50m^2 \right\}$$

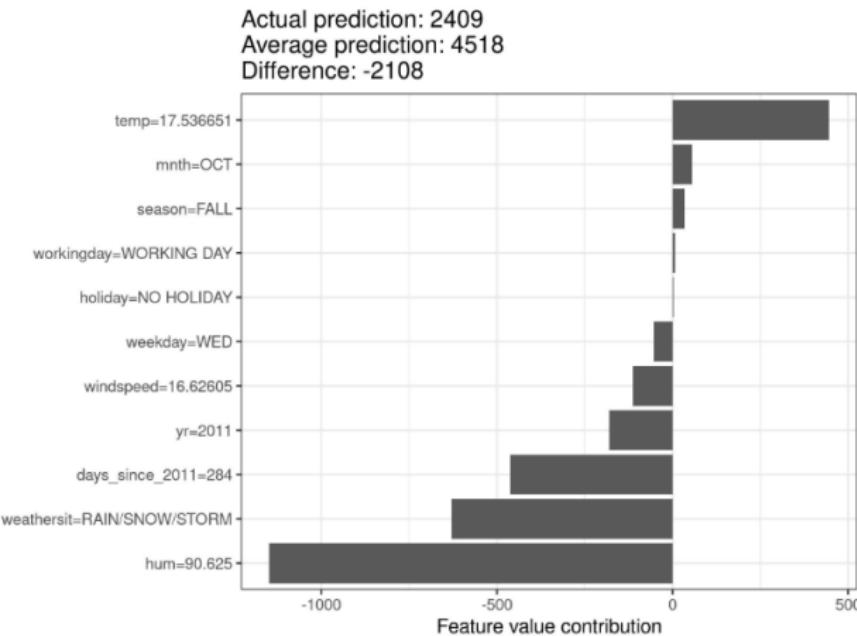
$$\left\{ \begin{array}{c} \text{Tree} \\ \checkmark \end{array} \quad \begin{array}{c} \text{Building} \\ \text{2nd floor} \end{array} \right\}$$

$$\left\{ \begin{array}{c} 50m^2 \\ \text{Building} \\ \text{2nd floor} \end{array} \right\}$$

$$\left\{ \begin{array}{c} \text{Tree} \\ \checkmark \end{array} \quad 50m^2 \quad \begin{array}{c} \text{Building} \\ \text{2nd floor} \end{array} \right\}$$

- ▶ Para cada uma dessas coalizões, calculamos o preço previsto do apartamento com e sem o valor do input cat-banned e pegamos a diferença para obter a contribuição marginal.
- ▶ O valor de Shapley é a média das contribuições marginais.
- ▶ Se estimarmos os valores de Shapley para todos os valores de inputs, obteremos a distribuição completa da previsão (menos a média) entre os valores de inputs.

Exemplo de SHAP



Referências I

- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16(3), 199–231.
- Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- Ribeiro, M. T., S. Singh, and C. Guestrin (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
- Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games* 2.

Referências II

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2014). Intriguing properties of neural networks.

Zhang, M., Y. Chen, and C. Qian (2023). Fooling examples: Another intriguing property of neural networks. *Sensors* 23(14), 6378.