Local Interpretable Model-Agnostic Explanations (LIME)

Alysson Machado

Graduação em Engenharia Elétrica
Universidade Federal de Campina Grande

alysson.barbosa@ee.ufcg.edu.br

16 de Março de 2021

Introdução

- Parte do princípio de que a veracidade dos modelos é desconhecida.
- Uma técnica para explicar as previsões de qualquer classificador de aprendizado de máquina.
- Entender a lógica por trás de um modelo ajuda os usuários a decidir quando confiar ou não em suas previsões.

(RIBEIRO; SINGH; GUESTRIN, 2016)

Contextualização

- Quando um engenheiro(a) carrega um modelo para a produção, ele(a) confia implicitamente que o modelo faça previsões sensatas.
- As métricas de avaliação são úteis, mas em alguns casos acabam sendo enganosas em relação ao desempenho real do modelo.
- As vezes, o modelo comete erros embaraçosos demais para serem aceitáveis.
- Considerando que a inteligência e intuição humana é mais valiosa que as métricas de avaliação, é importante assumir uma "etapa de escolha" em que certas previsões representativas são selecionadas para serem explicadas ao humano.

(RIBEIRO; SINGH; GUESTRIN, 2016)

Objetivo do LIME

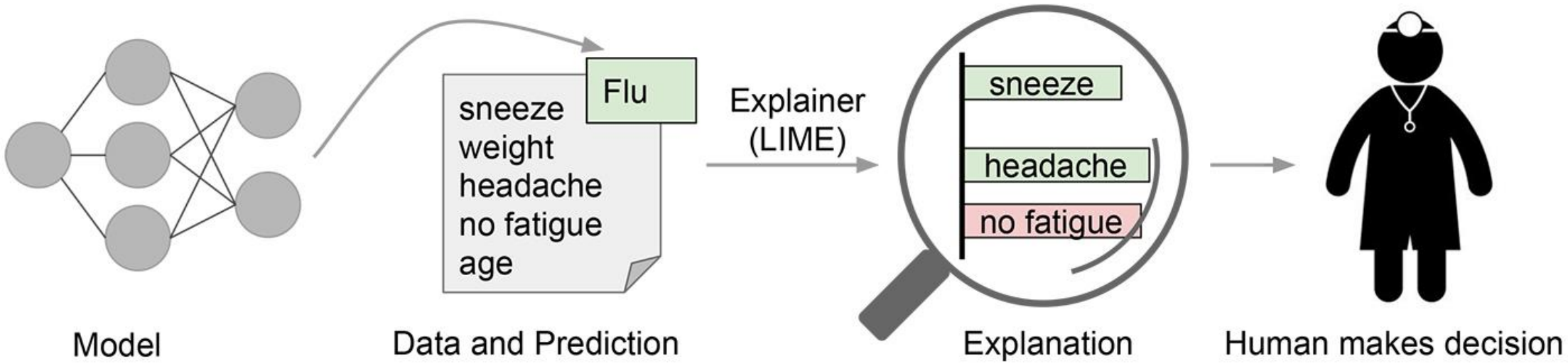


Figura 1 *– O Algoritmo LIME destaca as principais características dos exemplos individuais responsáveis pela predição realizada. Com isso, tal ferramenta pode ajudar o usuário a tomar uma decisão sobre o resultado obtido.* ***Fonte****: Marco Tulio Ribeiro*

(RIBEIRO; SINGH; GUESTRIN, 2016)

Objetivo do LIME

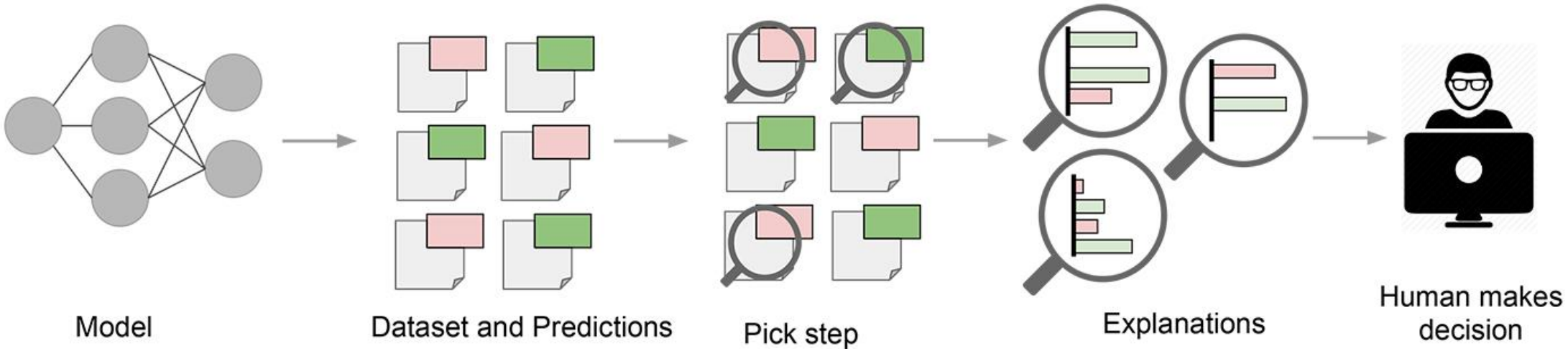


Figura 2 *– Etapa de escolha das previsões representativas para um entendimento mais preciso do desempenho do modelo.* ***Fonte****: Marco Tulio Ribeiro*

(RIBEIRO; SINGH; GUESTRIN, 2016)

Intuição por trás do LIME

- O algoritmo se mantém sempre agnóstico ao modelo.
- Observar mudanças na previsão através de pertubações na entrada, alterando os componentes que fazem sentido para os humanos.

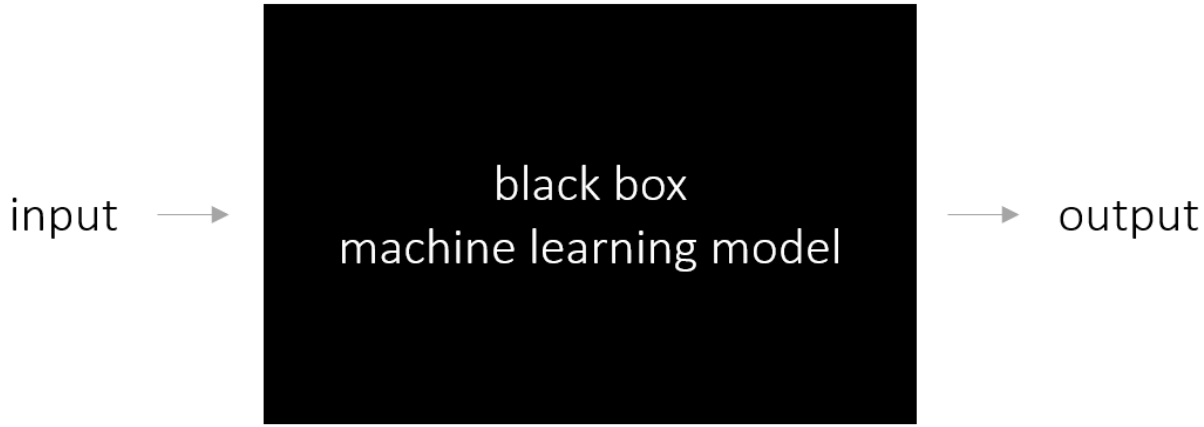


Figura 3 *– O algoritmo LIME não tem interesse em entender como o algoritmo funciona, ele apenas observa a relação entre a entrada e a saída no modelo.* ***Fonte****: Lars Hulstaert.*

(RIBEIRO; SINGH; GUESTRIN, 2016)

Intuição por trás do LIME

- Ponderar as imagens perturbadas por sua semelhança com a instância que está sendo observada.
- Na Figura 1, os três sintomas destacados podem ser uma aproximação fiel do modelo de caixa preta para pacientes com diagnóstico semelhante ao que foi inspecionado, mas provavelmente não representam como o modelo se comporta para todos os pacientes.

(RIBEIRO; SINGH; GUESTRIN, 2016)

Intuição por trás do LIME

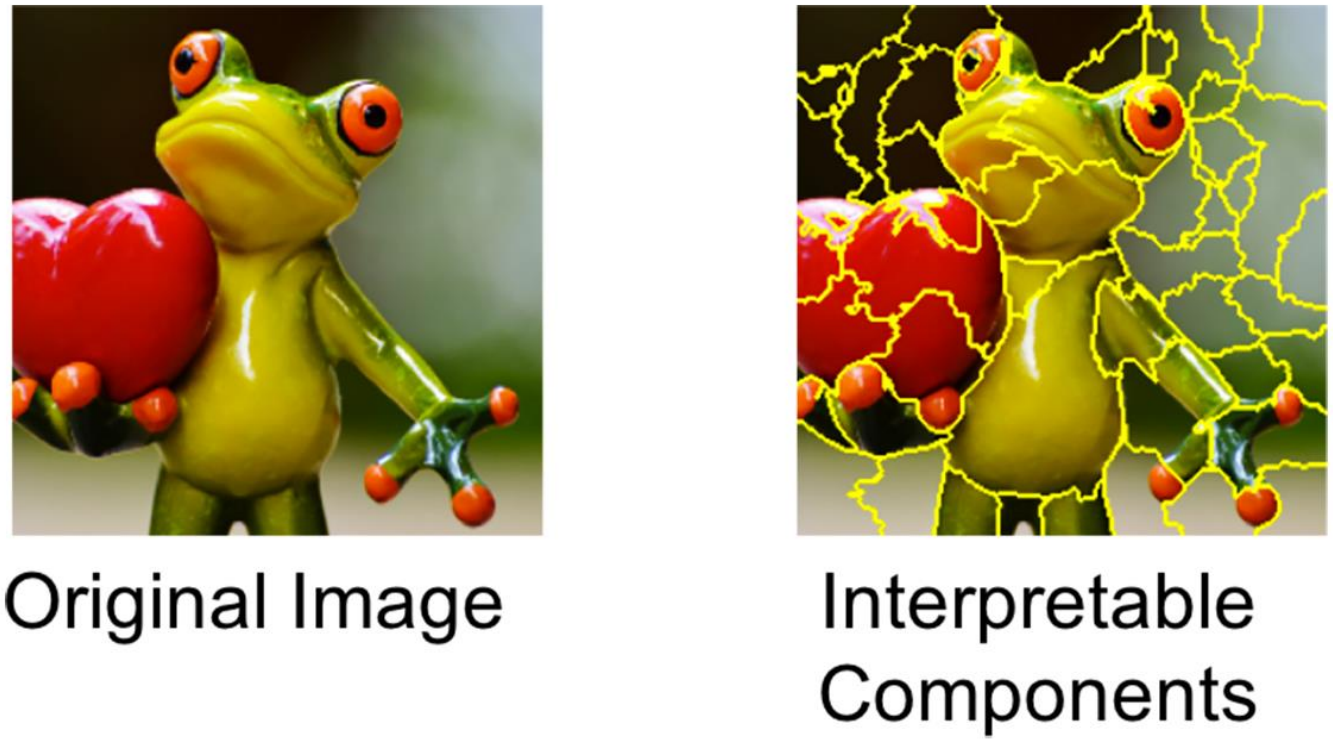


Figura 4 *– Um exemplo com classificadores de imagem é dividí-los em diferentes componentes interpretáveis, de modo analisar quais tem maior prevalência na ativação de uma determinada classe.* ***Fonte****: Marco Tulio Ribeiro.*

(RIBEIRO; SINGH; GUESTRIN, 2016)

Intuição por trás do LIME

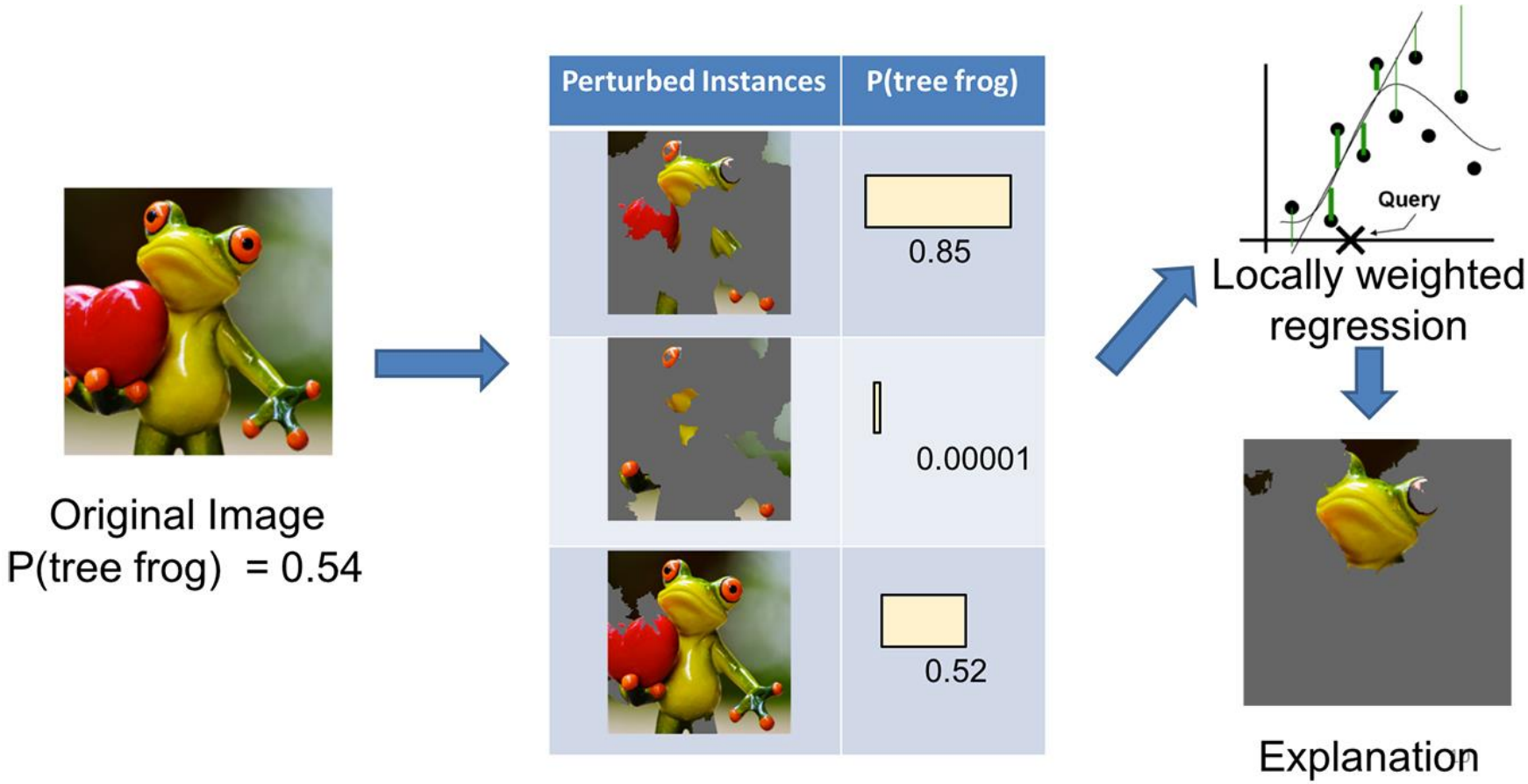


Figura 4 *– Procedimento geral do algoritmo LIME para analisar classificadores de imagens.* ***Fonte****: Marco Tulio Ribeiro.*

(RIBEIRO; SINGH; GUESTRIN, 2016)

Intuição por trás do LIME

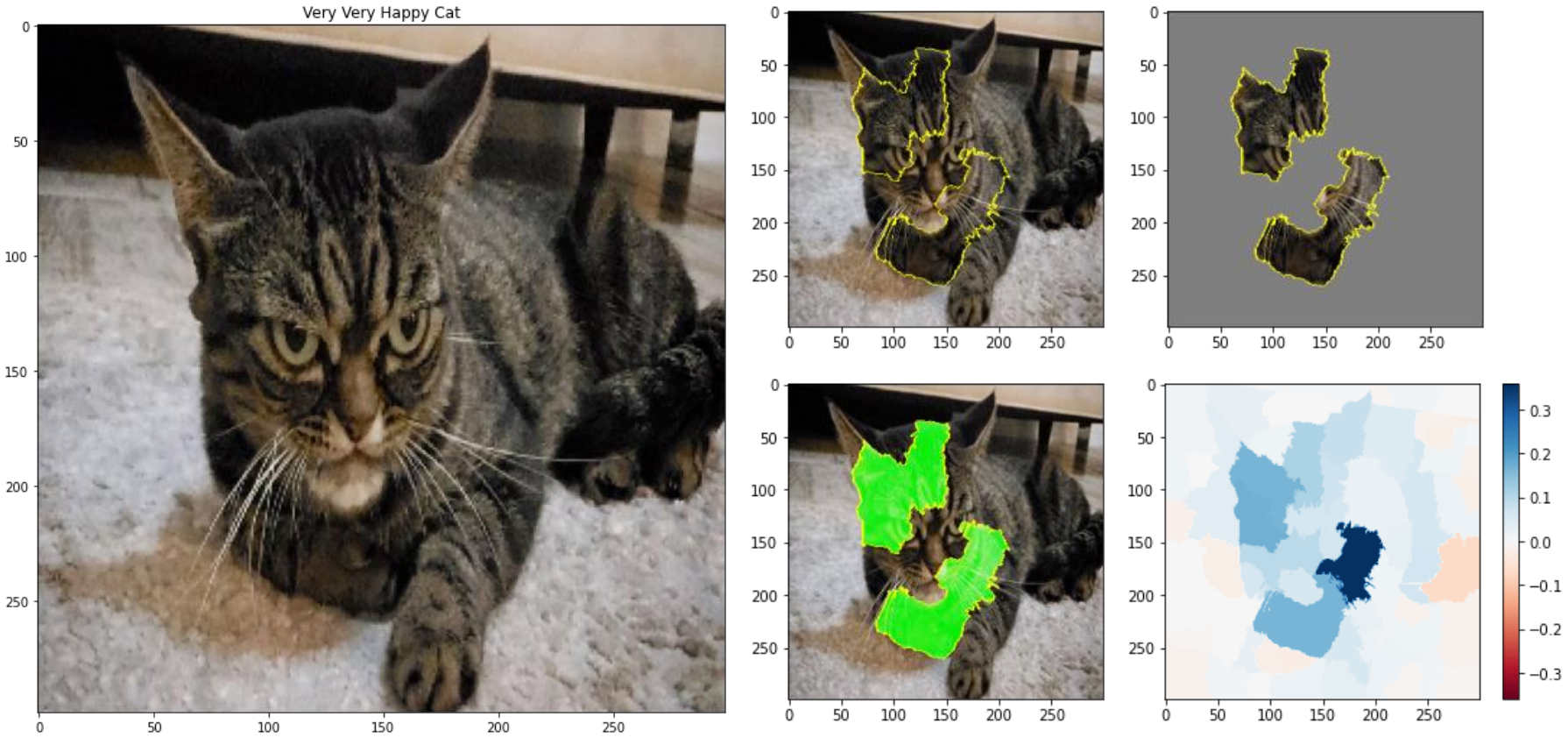


Figura 5 *– Procedimento geral do algoritmo LIME para analisar classificadores de imagens.* ***Fonte****: Alysson Machado.*

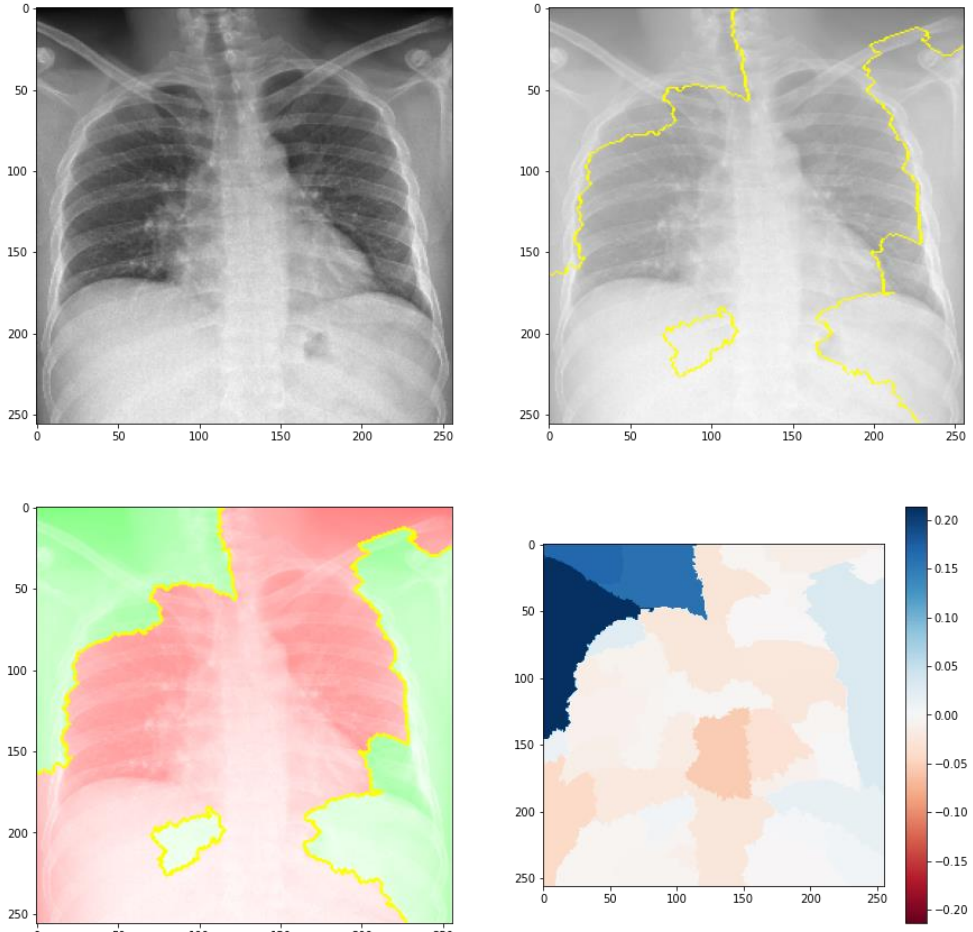
Aplicação do LIME

- A proposta do algoritmo LIME é útil para enteder melhor os modelos com aplicação na medicina, pois esse é um campo que exige uma análise minunciosa das predições realizadas.
- Uma das áreas proeminentes é a classificação ou detecção de distúrbios pulmonares.

(RIBEIRO; SINGH; GUESTRIN, 2016)

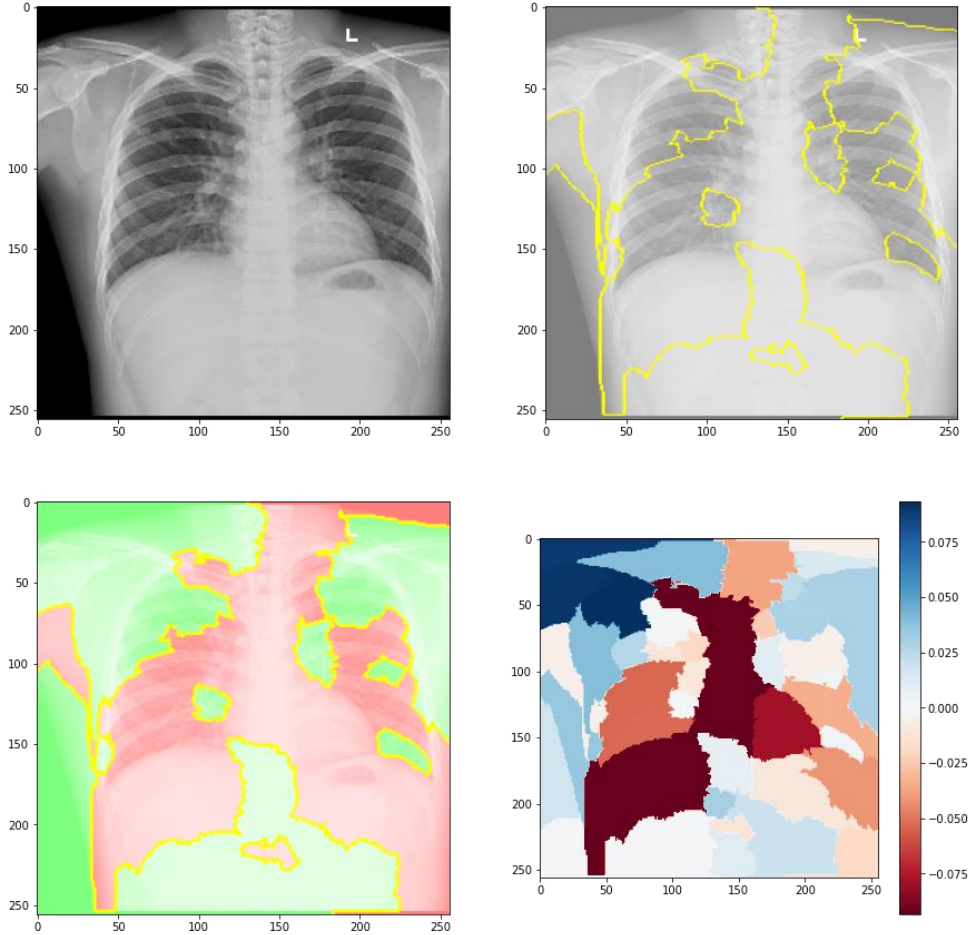
Aplicação do LIME

Radiografia Normal



Fonte: Alysson Machado

Radiografia Anormal



Fonte: Alysson Machado

Conclusão

- A confiança é crucial para uma interação humana eficaz com sistemas de aprendizado de máquina, explicar as previsões individuais é uma forma eficaz de avaliar a confiança do modelo.
- LIME é uma ferramenta eficiente para facilitar essa confiança para praticantes de aprendizado de máquina.

(RIBEIRO; SINGH; GUESTRIN, 2016)

Referência

- RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. " 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**. 2016. p. 1135-1144.