

# Engenharia de Machine Learning

2025-03-26

# Cenário atual

- Dados devidamente rastreados, imutáveis e acessados por catálogo (como o do Kedro)
- Reprodutibilidade do Treinamento como os cuidados de estruturação de código e documentação
- Métricas de treinos devidamente rastreadas com MLFlow
- Modelos versionados e servidos por API

# O que auditar e monitorar?

## Mudanças

Diversos fatores pode fazer com que haja mudança na performance dos modelos:

- **Data drift:** mudança na distribuição dos dados, ou na relação entre as features
- **Feature drift:** mudanças na semântica das features
- **Concept drift:** mudança na relação entre features e saída do modelo

## Viés de Modelos

- Problemas durante o treino
- Diferenças de representatividade entre dados de treino e vida real
- Refletir viés humano

## Transparência de Modelos

- Qualidade de dados de treino
- Rastreabilidade de fonte de dados e processo de anotação
- Transparência no processo de treino e tratamento dos dados
- Rastreabilidade de qual modelo gerou qual resultado

## Explicabilidade

Técnicas para explicar como o modelo chegou a uma determinada decisão

- **Local:** explicar a saída do modelo para uma entrada específica
- **Global:** entender como o modelo se comporta globalmente

# Monitoramentos em Produção

Coletar dados fornecidos para inferência para determinar mudanças:

- **Data drift:**

- Testes de hipótese (por exemplo, Kolmogorov-Smirnov com duas amostras)
- Tentativa de treinar um discriminador de dados velhos e novos

- **Model drift:**

- Observar distribuição da saída do modelo
- Monitorar a importância das features ao longo do tempo
- Retreinar periodicamente e comparar resultados

## Processo de Anotação Contínua e Retreino

- Monitorar métricas de desempenho do modelo ao longo do tempo (erro médio, F1, etc.)
- Periodicamente reanotar amostra de dados para avaliação

# Explicabilidade de Modelos

## Objetivo

Entender relevância de variáveis, e o que o modelo está levando em consideração para dar uma determinada classificação

## Conceitos

- Explicabilidade Global: mede a influência de cada variável na predição do modelo levando em consideração toda a base de dados conhecidos
- Cohort: Leva em consideração somente um stratum dos dados
- Explicabilidade Local: mede influências para a classificação de uma amostra específica

## Métodos

- **Específicos do modelo:** alguns modelos tem importâncias diretamente interpretáveis. Por exemplo: em Random Forests, as variáveis mais importantes tendem a aparecer como primeira decisão
- **Método Ablativo:** treinar  $n$  novos modelos candidatos usando  $n-1$  features, cada hora removendo uma variável. Medir importância pelo impacto na performance
- **Métodos Locais:** avaliar localmente cada exemplo

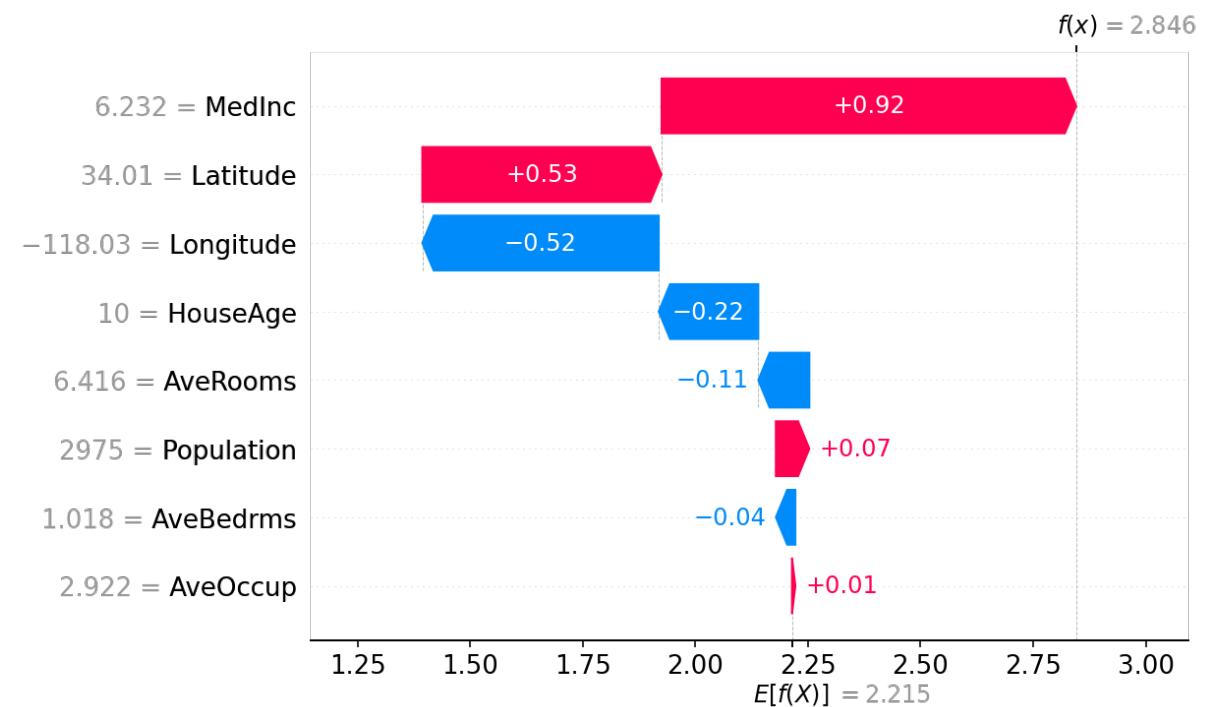
# Explicabilidade de Modelo (métodos locais)

## LIME (Local Interpretable Model-Agnostic Explanation)

- Mede o grau de influência de cada variável no ponto de operação
- Análogo a uma derivada parcial da saída em relação as entradas, no ponto da amostra local dos dados

## SHAP (SHapley Additive exPlanations) Values

- Análogo ao método ablativo, mas levando em consideração a interação e correlações entre variáveis



# Métodos Locais - SHAP Values



por Lucas Murakami (2022)

# Métodos Locais - Titanic

