Interpretabilidade de Modelos de AM

Ricardo Prudêncio

Feature Importance

- Métodos para explicações globais baseados em medir a importância das variáveis para as predições do modelo a ser explicado

- Pergunta: qual o impacto de cada variável para predições de instâncias não usadas no treinamento?

- Existem métodos gnósticos e agnósticos

Feature Importance para Random Forest

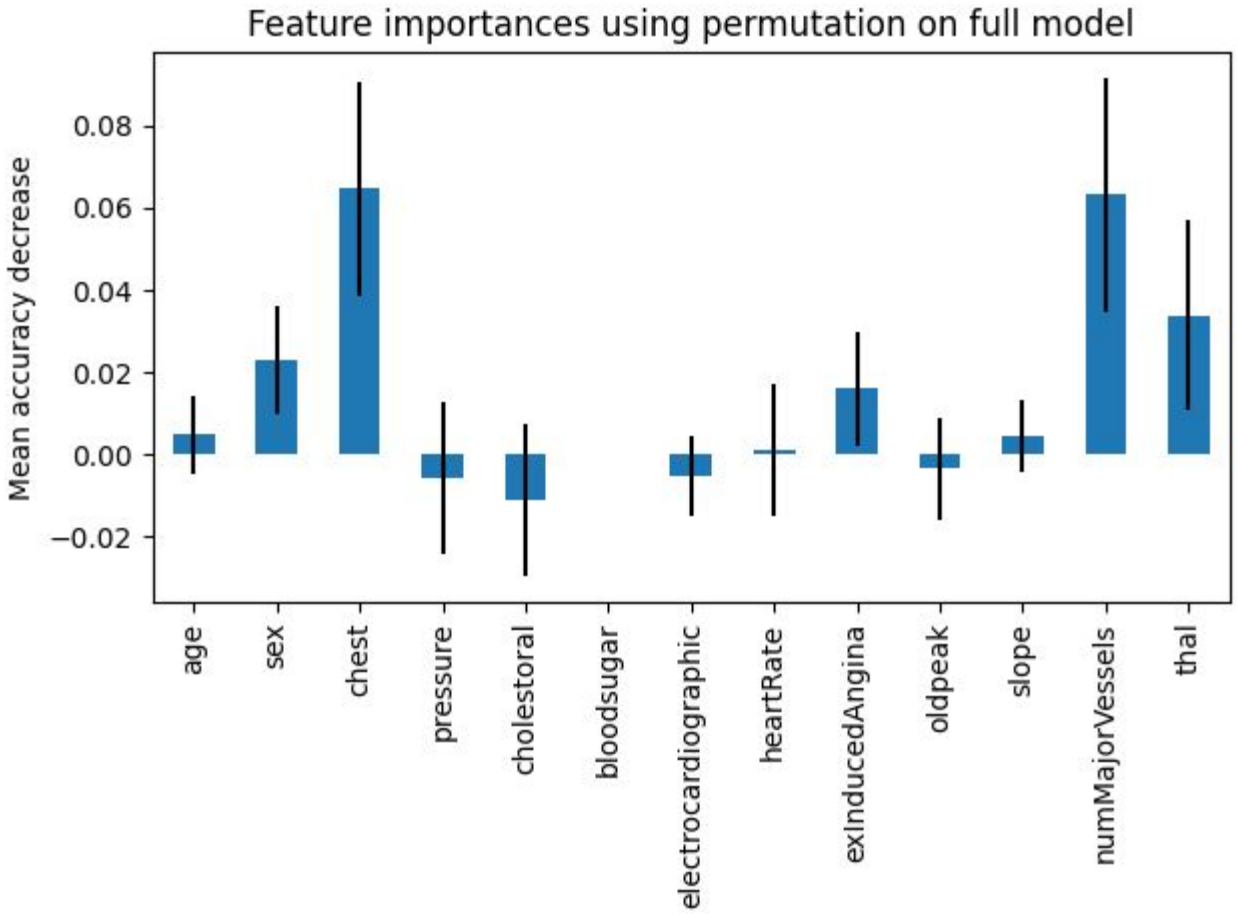
- Baseado na redução do critério de impureza
  - Considerar todas as árvores que usam o atributo
  - e medir a redução média de impureza causada pelo atributo

- Baseado no número de vezes em que o atributo é selecionado entre todas as árvores de decisão

Permutation Feature Importance (PFI)

1. Estimar o erro do modelo original para cada instância
2. Para cada feature:
    Gere uma nova matriz de dados, permutando os valores do atributo
    Estime o erro do modelo para os novos dados com a feature permutada

3 Calcule a importância como a diferença do erro original e o novo erro


Feature importances using permutation on full model
Mean accuracy decrease
0.08
0.06
0.04
0.02
0.00
−0.02
age
sex
chest
pressure
cholestoral
bloodsugar
electrocardiographic
heartRate
exinducedAngina
oldpeak
slope
numMajorVessels
thal

Observações Importantes

- PFI é calculada usando dados de teste não vistos no treinamento para evitar calcular importância de modelos com overfitting

- Pode gerar instâncias de teste que não são factíveis, i.e., fora da distribuição real dos dados ou inconsistentes

- Pode apresentar problemas para variáveis correlacionadas