

1. INTRODUÇÃO (Max 1000 carac)

Em virtude do crescimento da complexidade e do volume dos dados atualmente disponíveis, métodos algorítmicos de aprendizagem de máquina passaram a oferecer soluções muito bem sucedidas, sobretudo para as atividades de previsão, de classificação e de agrupamento. Entretanto, tais benefícios trouxeram consigo perdas substanciais na capacidade de abstração, de interpretação e de generalização. Neste estudo, realizamos uma extensa comparação entre os resultados obtidos por modelos probabilísticos e de aprendizagem de máquina no ajuste de um conjunto de dados reais. Tal exercício ilustrou o uso de alguns dos principais métodos de interpretação e de inferência estatística em modelos preditivos.

2. METODOLOGIA (Max 1000 carac)

Foi realizada uma comparação de capacidade de previsão, por meio dos coeficientes R^2 e o EQM (Erro quadrático médio), entre um modelo probabilístico de Regressão Linear Múltipla, modelos de Redes Neurais Artificiais e um modelo XGBoost. Com o intuito de interpretar os modelos de aprendizagem de máquina, foram utilizadas simulações de Monte Carlo para obter coeficientes que representam a importância de cada variável na previsão em cada modelo. Os valores obtidos foram comparados com os coeficientes da Regressão Linear e com os valores de SHAP (SHapley Additive exPlanations) da Rede Neural. A metodologia apresentada foi ilustrada em um conjunto de dados do censo norte americano de 1990, conhecido como "Califórnia Housing", no qual há informações sobre a demografia nos bairros, a localização (latitude e longitude) e informações gerais sobre as casas. A variável de interesse é o valor médio das casas. As análises foram feitas com a linguagem de programação Python.

3. RESULTADOS (Max 1000 carac)

Pôde-se observar, por meio dos coeficientes R^2 e o EQM, uma melhor capacidade preditiva dos modelos de aprendizado de máquina em relação ao probabilístico. Um dos motivos para essa diferença de desempenho foi a falta de capacidade do modelo probabilístico para captar as não-linearidades de algumas das variáveis. Com os coeficientes obtidos pelo método de simulação de Monte Carlo, foi possível observar qual importância os modelos atribuíram a cada variável, uma vez que todos os modelos obtiveram as mesmas variáveis como mais relevantes para a previsão. Contudo, graus de relevância diferentes foram apresentados, fato evidenciado quando comparado aos valores de SHAP e coeficientes da Regressão Linear, pois todos estavam de acordo. Outrossim, observando os coeficientes e os p-valores obtidos pela simulação de Monte Carlo e pelo modelo probabilístico, todos os modelos foram capazes de identificar a irrelevância das duas variáveis que foram artificialmente adicionadas aos dados.

4. CONCLUSÃO (Max 1000 carac)

O presente trabalho comparou modelos probabilísticos com os de aprendizagem de máquina, buscando assim, explorar potenciais técnicas para a interpretação dos modelos conhecidos como "caixa-preta". Nesse sentido, observa-se que, para dados mais complexos, a Rede Neural e o XGBoost tiveram maior capacidade de previsão comparados à Regressão Linear Múltipla e, por isso, interpretar estes modelos é de suma importância. Assim, com as técnicas de simulação de Monte Carlo e os valores SHAP, foi possível esclarecer as previsões dos algoritmos "caixa-preta" e ter um ganho muito significativo na interpretabilidade desses modelos.

5. PALAVRAS-CHAVE

Regressão Linear

Redes Neurais Artificiais

XGBoost

Inteligência Artificial Interpretável

Simulações de Monte Carlo

Valores de SHAP