MAC0459 - Q3

Leonardo Heidi Almeida Murakami

December 2021

1 Resumo

Um estudo foi feito para entender como os seres humanos utilizam e confiam em algoritmos de Machine Learning para os auxiliarem em duas tarefas razoavelmente difíceis, computar o número de pessoas em uma multidão e identificar animais similares. Além disso, foram utilizados diversos categorias para poder descobrir o quanto entender sobre o modelo impacta a confiança do usuário sobre o modelo enquanto deixa de confiar em seu próprio julgamento para uma tarefa que possui uma alta taxa de acerto.

Foi concluído que, no geral, seguimos as recomendações feitas pelo algoritmo mesmo que elas estejam erradas, embora as errôneas sejam seguidas em menor quantidade. Além disso foi concluído que informações sobre o modelo que nos fariam concluir que o algoritmo teria mais chance de errar (como características e informações adicionais da imagem) no geral impactam na nossa probabilidade de seguir a recomendação mesmo que esteja informado que esta possa estar errada, este impacto é visto mesmo quando a escolha é feita por um expert na área.

2 Metodologia

Foram escolhidos dois datasets com implementações já feitas em Machine Learning, distinguir animais em fotos e escolher a foto com a maior multidão. Fora isso, fora introduzido a todos os participantes algumas informações sobre o modelo (como performance em um subset, dados usados para treino, entre outros), todos estes vídeos foram simplificados de maneira a atender todo o público do estudo.

Além disso, foram selecionados 175 participantes para fazer parte deste estudo, sendo estes de uma gama vasta de perfis, variando suas etnicidades, gênero, idade, nivel de educação e conhecimento de aprendizado de máquina, as tarefas entregues a estes participantes foram então quebradas em diferentes subgrupos que receberiam diferentes informações na tela em que os participantes selecionariam a resposta do dataset mencionado acima.

Estes subgrupos foram criados do seguinte jeito, metade da tarefa seria realizada através de fotos em que o estimador possui alta certeza da predição dada,

o que acabaria resultando em mais respostas corretas (ou uma expectativa de receber a resposta correta maior) e a outra metade da tarefa seria realizada com fotos em que o estimador possui uma incerteza grande, ou seja, possui uma maior chance de estar incorreto em sua recomendação.

Além da quebra da foto fornecida, as informações na tela seriam selecionadas das seguintes opções (sendo estas fixas para cada grupo de fotos com a quebra mencionada acima):

1. Sem recomendação, onde o participante não recebe a recomendação fornecida pelo modelo.
2. Recomendação apenas, onde o participante apenas recebe a recomendação do modelo.
3. Recomendação com vídeos opcionais, onde o participante recebe a recomendação com alguns vídeos adicionais mostrando a performance e informações do modelo, os participantes não são obrigados a ver os vídeos.
4. Recomendação com vídeos obrigatórios, onde o participante recebe a recomendação com alguns vídeos adicionais mostrando a performance e informações do modelo, os participantes são obrigados a ver os vídeos para poder selecionar a resposta.

3 Conclusão

As conclusões obtidas a partir de testes estatísticos foram de que

1. As pessoas geralmente seguem recomendações de modelos em todos os níveis de expertise, seja esta de ML ou de Matemática. As recomendações foram ainda mais seguidas quando eram acompanhadas de informações do modelo.
2. Mesmo que as pessoas completem corretamente as tarefas, existe a tendência de seguir as recomendações do modelo
3. Embora exista a tendência de seguir o modelo, este foi bem menos seguido quando a recomendação estava incorreta e menos ainda quando a recomendação era feita em uma imagem que foi informada que o modelo performaria mal.
4. Nenhuma informação sobre o modelo melhorou a acurácia das pessoas.

4 Análise Crítica

Existem diversos aspectos muito positivos sobre o estudo. A maneira como estes lidaram com os mais diversos cenários, como por exemplo a possibilidade de alguém estar contrariando o modelo de maneira proposital pode ser balanceado

vendo a acurácia quando este não possui a recomendação. O teste foi feito de maneira robusta e extremamente bem feita, a única crítica que possuo para o teste é o n relativamente baixo quando quebramos os perfis de cada pessoa. Como o recrutamento para o teste foi feito através de redes sociais pode acabar tendo enviesado o perfil de pessoas que aceitaram fazer o teste.

Existem outras perguntas que poderiam ser respondidas pelo estudo, será que modelos mais complexos tem a tendencia a serem seguidos mais cegamente pelos usuários, ou seja, a informação explicando o funcionamento do modelo teria menos peso caso a complexidade da explicação seja muito alta, será que modelos muito simplificados podem acabar sendo ignorados por possuirmos um viés de complexidade, onde soluções complicadas acabam sendo preferidas no lugar de uma solução simples, este teste, inclusive, poderia ser realizado mostrando a uma pessoa duas recomendações feitas por dois modelos diferentes e ver qual era seguido na maior parte do tempo e se a expertise em ML impactava na escolha.