



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

Formação Cientista de Dados

Projeto com Feedback 1

Detecção de Fraudes no Tráfego de Cliques  
em Propagandas de Aplicações Mobile

O risco de fraude está em toda parte, mas para as empresas que anunciam online, a fraude de cliques pode acontecer em um volume avassalador, resultando em dados de cliques enganosos e dinheiro desperdiçado. Os canais de anúncios podem aumentar os custos simplesmente quando pessoas ou bots clicam nos anúncios em grande escala, o que na prática não gera o resultado esperado. Com mais de 1 bilhão de dispositivos móveis em uso todos os meses, a China é o maior mercado móvel do mundo e, portanto, sofre com grandes volumes de tráfego fraudulento.

A TalkingData (<https://www.talkingdata.com>), a maior plataforma de Big Data independente da China, cobre mais de 70% dos dispositivos móveis ativos em todo o país. Eles lidam com 3 bilhões de cliques por dia, dos quais 90% são potencialmente fraudulentos. Sua abordagem atual para impedir fraudes de cliques para desenvolvedores de aplicativos é medir a jornada do clique de um usuário em todo o portfólio e sinalizar endereços IP que produzem muitos cliques, mas nunca acabam instalando aplicativos. Com essas informações, eles criaram uma lista negra de IPs e uma lista negra de dispositivos.

Embora bem-sucedidos, eles querem estar sempre um passo à frente dos fraudadores e pediram a sua ajuda para desenvolver ainda mais a solução. Você está desafiado a criar um algoritmo que possa prever se um usuário fará o download de um aplicativo depois de clicar em um anúncio de aplicativo para dispositivos móveis.

Em resumo, neste projeto, você deverá construir um modelo de aprendizado de máquina para determinar se um clique é fraudulento ou não.

Para a construção desse projeto, recomendamos a utilização da linguagem R e o dataset disponível no Kaggle em:

<https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/data>

Quando concluir o projeto, envie os scripts e datasets para [projeto@dsacademy.com.br](mailto:projeto@dsacademy.com.br). Caso os datasets usados sejam muito grandes, armazene em um diretório virtual (existem vários na internet, como Google Drive ou Dropbox) e envie o link para que nossa equipe possa baixar os datasets. Se os arquivos foram pequenos (uma amostra do dataset original), envie no anexo junto com o script. Documente seu script tanto quanto possível.

Caso prefira, disponibilize seu projeto no Github e envie o link do seu repositório para nossa equipe no e-mail [projeto@dsacademy.com.br](mailto:projeto@dsacademy.com.br). Nesse caso, o Readme do repositório deve constar que este trata-se de um projeto da Formação Cientista de Dados da Data Science Academy.

Em até 24 horas, daremos o feedback!

Bom trabalho!