Desafio para área de Data Science

Parabéns pela aprovação na primeira etapa do processo seletivo para a área de Data Science da DHAUZ!

Seguindo com o nosso processo, nesta etapa você deverá realizar um desafio técnico, detalhado abaixo. O objetivo é avaliar a sua capacidade de exploração de dados, criatividade na elaboração de hipóteses, features e metodologia na resolução de problemas.

Você pode utilizar a linguagem, ferramentas e frameworks que se sentir mais confortável para elaborar a solução. Recomendamos o desenvolvimento do racional e exploração em uma ferramenta como por exemplo o Jupyter Notebook para facilitar a inclusão de comentários e gráficos explicativos.

Sinta-se livre para utilizar qualquer outra funcionalidade não listada acima que demonstre suas habilidades!

O projeto deve ser feito em um repositório no Github e o seu link enviado no final do desafio.

1. Previsão de cancelamentos em Hotéis

Você foi contratado pela DHAUZ como cientista de dados para analisar uma base de dados de clientes de uma rede de Hotéis e sua tarefa é investigar os dados em busca de *insights* que possam ajudar a empresa a evitar cancelamentos e também construir um modelo preditivo que possa antecipar esses cancelamentos, de modo que a empresa tenha tempo hábil para agir com ações de retenção.

Informações sobre o dataset:

- As informações são anonimizadas por questões de privacidade
- Base de dados no link:
    - https://dhauz-challenges.s3.amazonaws.com/cancellation_prediction.rar

Comece respondendo as seguintes questões:

a) Elabore hipóteses e visualizações envolvendo a variável *cancellation* e, pelo menos, outras duas variáveis presentes no dataset;

b) Desenvolva um modelo preditivo de classificação para identificar cancelamentos e utilize métricas adequadas para argumentar a efetividade do modelo;

c) Ao realizar a validação cruzada do modelo de classificação, discuta sobre as diferenças entre utilizar uma separação entre treino e teste aleatória e uma separação temporal (Ex: treino em 2015 e 2016 e validação em 2017). Os resultados são diferentes? Qual o mais indicado?