PROVA PRÁTICA
CIENTISTA DE DADOS I

OBSERVATÓRIO DA INDÚSTRIA

O que queremos receber?

Queremos receber nos e-mail a seguir a resposta da auto avaliação e os artefatos desenvolvidos no desafio: fjfilho@sfiec.org.br, dgasilva@sfiec.org.br, elgomes@sfiec.org.br, t_msaraujo@sfiec.org.br.

1) Auto avaliação

Auto-avalie suas habilidades nos requisitos de acordo com os níveis especificados usando o link abaixo.

Qual o seu nível de domínio nas técnicas/ferramentas listadas abaixo, onde:

- 0, 1, 2 - não tem conhecimento e experiência;
- 3, 4 ,5 - conhece a técnica e tem pouca experiência;
- 6 - domina a técnica e já desenvolveu vários projetos utilizando-a.

Tópicos de Conhecimento:

- Manipulação e tratamento de dados com Python:
- Manipulação e tratamento de dados com Pyspark e Pandas:
- Elaboração de Modelos de Machine Learning;
- Otimização de performance de modelos de Machine Learning;
- Desenvolvimento de data workflows com Airflow:
- Desenvolvimento de experimentos no MLFlow
- Desenvolvimento de data workflows em Ambiente Azure com databricks:
- Manipulação de bases de dados NoSQL:
- Web crawling e web scraping para mineração de dados:

2) Desafio Cientista de dados II.

Instruções

Imagine que você é Cientista de Dados e está responsável por um projeto no qual a Federação das Indústrias está prestando consultoria para um e-commerce. O Cliente está querendo aumentar o seu faturamento e devido a questões de negócios como problemas com fornecedores, promoções mal planejadas, incidência de impostos e afins, algumas vendas podem resultar em prejuízo Atualmente, essas informações não estão disponíveis previamente. Dessa forma, faça um modelo que preveja o Lucro (Profit) ou prejuízo de vendas feitos no e-commerce. Caso o algoritmo identifique que haverá prejuízo, o site irá indeferir a compra, enviando-a para que um analista possa verificar o que está ocorrendo.

Questão A: Faça um Jupyter notebook, com a análise, exploração, visualização dos dados e elaboração de um modelo que preveja a coluna Profit do Dataset disponibilizado.

Questão Extra: Faça um experimento no MLFlow ou uma DAG usando Airflow na qual as etapas de extração, transformação, treino e previsão do modelo sejam separadas em scripts/tasks.

Sugerimos utilizar boas práticas de programação na sua solução, documente e escreva um README.md explicando suas escolhas.