Desafio Cientista de Dados

Introdução

Olá candidato(a), o objetivo deste desafio é testar os seus conhecimentos sobre a resolução de problemas de negócios, análise de dados e aplicação de modelos preditivos. Queremos testar seus conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning. É importante deixar claro que não existe resposta certa e que o que nos interessa é sua capacidade de descrever e justificar os passos utilizados na resolução do problema.

Desafio

Você foi alocado em um time da Indicium contratado por um estúdio de Hollywood chamado *PProductions,* e agora deve fazer uma análise em cima de um banco de dados cinematográfico para orientar qual tipo de filme deve ser o próximo a ser desenvolvido. Lembre-se que há muito dinheiro envolvido, então a análise deve ser muito detalhada e levar em consideração o máximo de fatores possíveis (a introdução de dados externos é permitida - e encorajada).

Entregas

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!
2. Responda também às seguintes perguntas:
    a. Qual filme você recomendaria para uma pessoa que você não conhece?
    b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?
    c. Quais insights podem ser tirados com a coluna *Overview*? É possível inferir o gênero do filme a partir dessa coluna?
3. Explique como você faria a previsão da **nota do imdb** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?
4. Supondo um filme com as seguintes características:

```
{'Series_Title': 'The Shawshank Redemption',
 'Released_Year': '1994',
 'Certificate': 'A',
 'Runtime': '142 min',
 'Genre': 'Drama',
 'Overview': 'Two imprisoned men bond over a number of years,
finding solace and eventual redemption through acts of common
decency.',
 'Meta_score': 80.0,
 'Director': 'Frank Darabont',
 'Star1': 'Tim Robbins',
 'Star2': 'Morgan Freeman',
 'Star3': 'Bob Gunton',
 'Star4': 'William Sadler',
 'No_of_Votes': 2343110,
 'Gross': '28,341,469'}
```

Qual seria a nota do IMDB?

5. Salve o modelo desenvolvido no formato .pkl.
6. A entrega deve ser feita através de um repositório de código público que contenha:
    a. README explicando como instalar e executar o projeto
    b. Arquivo de requisitos com todos os pacotes utilizados e suas versões
    c. Relatórios das análises estatísticas e EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1 e 2.
    d. Códigos de modelagem utilizados no passo 3 (pode ser entregue no mesmo Jupyter Notebook).
    e. Arquivo .pkl conforme passo 5 acima.

    Todos os códigos produzidos devem seguir as boas práticas de codificação.

Prazo

- Você tem até **7 dias corridos** para a entrega, contados a partir do recebimento deste desafio.
- Envie o seu relatório dentro da sua data limite para o e-mail: **selecao.lighthouse@indicium.tech**
- O arquivo de entrega deve ser nomeado como: **LH_CD_SEUNOME**

Bom trabalho!

Dicionário dos dados

A base de dados de treinamento contém 15 colunas. Seus nomes são auto-explicativos, mas, caso haja alguma dúvida, a descrição das colunas é:

Series_Title – Nome do filme
Released_Year - Ano de lançamento
Certificate - Classificação etária
Runtime – Tempo de duração
Genre - Gênero

IMDB_Rating - Nota do IMDB

Overview - *Overview* do filme

Meta_score - Média ponderada de todas as críticas

Director – Diretor

Star1 - Ator/atriz #1

Star2 - Ator/atriz #2

Star3 - Ator/atriz #3

Star4 - Ator/atriz #4

No_of_Votes - Número de votos

Gross - Faturamento