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(a) em um time da Indicium que está trabalhando atualmente junto a um cliente no processo de criação de uma plataforma de aluguéis temporários na cidade de Nova York. Para o desenvolvimento de sua estratégia de precificação, pediu para que a Indicium fizesse uma análise exploratória dos dados de seu maior concorrente, assim como um teste de validação de um modelo preditivo.

**Seu objetivo é desenvolver um modelo de previsão de preços a partir do *dataset* oferecido, e avaliar tal modelo utilizando as métricas de avaliação que mais fazem sentido para o problema.** O uso de outras fontes de dados além do dataset é permitido (e encorajado). Você poderá encontrar em anexo um dicionário dos dados.

# Entregas

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas. Seja criativo!
2. Responda também às seguintes perguntas:
   1. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?
   2. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?
   3. Existe algum padrão no texto do nome do local para lugares de mais alto valor?
3. Explique como você faria a previsão do **preço** 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 apartamento com as seguintes características:

{'id': 2595,

'nome': 'Skylit Midtown Castle',

'host\_id': 2845,

'host\_name': 'Jennifer',

'bairro\_group': 'Manhattan',

'bairro': 'Midtown',

'latitude': 40.75362,

'longitude': -73.98377,

'room\_type': 'Entire home/apt',

'minimo\_noites': 1,

'numero\_de\_reviews': 45,

'ultima\_review': '2019-05-21',

'reviews\_por\_mes': 0.38,

'calculado\_host\_listings\_count': 2,

'disponibilidade\_365': 355}

Qual seria a sua sugestão de preço?

1. Salve o modelo desenvolvido no formato .pkl.
2. A entrega deve ser feita através de um repositório de código público que contenha:
   1. README explicando como instalar e executar o projeto
   2. Arquivo de requisitoscom todos os pacotes utilizados e suas versões
   3. Relatórios das análises estatísticas e EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1 e 2.
   4. Códigos de modelagem utilizados no passo 3 (pode ser entregue no mesmo Jupyter Notebook).
   5. Arquivo .pkl conforme passo 5 acima.
3. Um vídeo curto explicando o desenvolvimento de suas entregas deste desafio, como você planejou e executou as atividades propostas. O vídeo deverá ser entregue via link via Google Drive. Lembre-se de autorizar o acesso para "qualquer pessoa com o link".

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. O não cumprimento deste prazo implica na desclassificação do processo seletivo.
* A Indicium possui ferramentas avançadas de detecção de plágio e inteligência artificial. A utilização de IA implica na desclassificação do processo seletivo.
* Envie o seu relatório dentro da sua data limite para o email: **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 16 colunas. Seus nomes são auto-explicativos, mas, caso haja alguma dúvida, a descrição das colunas é:

id – Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo

nome - Representa o nome do anúncio

host\_id - Representa o id do usuário que hospedou o anúncio

host\_name – Contém o nome do usuário que hospedou o anúncio

bairro\_group - Contém o nome do bairro onde o anúncio está localizado

bairro - Contém o nome da área onde o anúncio está localizado

latitude - Contém a latitude do local

longitude - Contém a longitude do local

room\_type – Contém o tipo de espaço de cada anúncio

price - Contém o preço por noite em dólares listado pelo anfitrião

minimo\_noites - Contém o número mínimo de noites que o usuário deve reservar

numero\_de\_reviews - Contém o número de comentários dados a cada listagem

ultima\_review - Contém a data da última revisão dada à listagem

reviews\_por\_mes - Contém o número de avaliações fornecidas por mês

calculado\_host\_listings\_count - Contém a quantidade de listagem por host

disponibilidade\_365 - Contém o número de dias em que o anúncio está disponível para reserva