Desafio Cientista de Dados

Vinícius Guimarães Ferreira

# 

# Desafio

**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.**

# 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!

Resposta: A análise exploratória dos dados encontra-se no Jupyter Notebook dentro do repositório do git-hub:

https://github.com/FerreiraVini/INDICIUM\_DataScience\_2024/tree/main

Análises simples como histogramas foram utilizadas para avaliar a distribuição dos dados, bem como promover embasamento para a exclusão de outliers. Em seguida foram gerados boxplots e barplots de todas as variáveis em relação ao preço, variável dependente, em busca de visualizar possíveis correlações lineares. Foram promovidas análises específicas para cada uma das features. Na análise do mínimo de noites foram excluídas todas as locações com mínimo acima de 30 dias, já que o site possui o foco em locação de temporada (short term rental), e a lei de NY diz que para se enquadrar nesta categoria, o limite de locação é de 30 dias. Outro exemplo é a análise da localização das listings no mapa de NY. Em seguida, foi utilizada uma biblioteca baseada no Open Street Map, em busca de verificar a proximidade das listings de maior valor agregado com pontos turísticos da cidade. Por fim, foi criada uma feature baseada na distância de cada listing da ilha do governador, onde fica a estátua da liberdade.

Outra análise interessante foi promovida com os nomes de cada listing. Foi feita a análise das palavras (tokenization) removendo-se as stopwords, palavras em outras línguas, e reduzindo cada palavra a sua raiz, para minimizar a análise duplicada de palavras semelhantes. Após estas etapas, foi criado um gráfico de Nuvem de palavras, e avaliadas as palavras associadas à maior média de preços dos imóveis anunciados (listings). Foi criada uma feature baseada na soma dos ‘preços médios associados’ a cada palavra. Por fim, foi também utilizado um embedding para cada token, a fim de extrair mais informações sobre cada anúncio. As demais features foram também avaliadas, e deram origem a novas features, como por exemplo a disponibilidade\_365 que originou features baseadas na faixa de disponibilidade de cada imóvel.

Após a EDA, foi feito o one-hot-encoding das variáveis categóricas, bem como a transformação (utilizando a raiz quadrada) de alguns features, a fim de melhorar a distribuição dos dados. Data scaling foi promovida em todas as variáveis independentes. Por fim, foi feita a divisão dos dados em matrizes de teste e de treinamento, e avaliados diferentes modelos estatísticos para previsão dos preços, incluindo regressão linear, random forest regression, XGBoost e LightGBM.

1. 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?

Resposta: Desconsiderando o possível valor do imóvel como um possível impedimento, e desconsiderando os outliers (valores muito altos), o bairro com a maior média de preços de aluguel é Willowbrook, em Staten Island. Apesar de Staten Island não ser o grupo de vizinhança com a maior média de preços, o bairro de Willowbrook se destaca entre os demais. Uma provável explicação para os altos valores médios de aluguel seria o fato de o bairro ser composto por casas térreas, com bom espaço, e com uma localização diferenciada. No entanto, valores médios de aluguel semelhantes podem ser obtidos também em Neponsit (Queens), Tribeca e NoHo (Manhattan).

* 1. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Resposta: Uma análise gráfica de correlação é incapaz de mostrar uma clara correlação entre mínimo de noites, disponibilidade\_365 e preço. Porém, ambas as features obtiveram um alto impacto no melhor modelo preditivo gerado, indicando que possuem sim um efeito no preço.

* 1. Existe algum padrão no texto do nome do local para lugares de mais alto valor?

Resposta: Os gráficos de nuvem de palavras (word cloud) possibilitaram visualizar de forma clara quais palavras estavam correlacionadas com maior média de preços dos imóveis. Palavras enfatizando qualidades nobres dos imóveis, ou ressaltando ótimas localizações obtiveram maior importância em relação ao preço. Palavras como: Club, Epic, Statu (de Statue of Liberty), VIP, Stylish, Village, Chelsea, SoHo, Luxurious, duplex, entre outras.

Para o modelo preditivo gerado ao final, as palavras Studio, room e luxuri (de Luxurious) foram as mais influentes no modelo.

1. 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?

Resposta: Na análise apresentada no Jupyter Notebook foram gerados diversos modelos preditivos. Tendo em vista a necessidade de promover a previsão de um valor contínuo, todos os modelos utilizados foram modelos de regressão. Após as análises geradas, e o treinamento de diversos modelos, as cinco features com maior impacto na previsão dos preços foram: Tipo de imóvel (em específico se o imóvel era ou não a casa inteira privada); longitude, latitude, disponibilidade\_365 e a palavra ‘studio’ no anúncio. No Jupyter Notebook são apresentadas as 20 principais features para o melhor modelo gerado, LightGBM, um modelo de gradient boosting que aplica arvores de decisão. No entanto, o modelo não foi capaz de predizer com alta precisão os preços, apresentando um MAPE (Mean Absolute Percentage Error) de 25%, que equivale a uma variação de 43 dolares em média dos valores reais (dado pelo RMSE – Rooted Mean Squared Error).

1. 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',

'price': 225,

'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?

Resposta: Na realidade esta amostra faz parte dos dados de treinamento, e seu valor é de 225 dólares. No entanto, para fazer a previsão desta amostra a partir do modelo de LightGBM gerado, seria necessário promover uma alteração no código de pré-processamento dos dados, como por exemplo salvar o valor de preço médio atribuído a cada token de palavra, algo que infelizmente não foi possível ser feito em tempo hábil. Porém, vale ressaltar que não devemos fazer previsão de dados utilizados no treinamento.

1. Salve o modelo desenvolvido no formato .pkl.

Resposta: Feito e demonstrado no final Jupyter Notebook, logo antes do item 7 – Conclusion. Foram também definidas funções para o pré-processamento direto dos dados, função esta contendo todos os filtros, transformações e escalonamento dos dados. Por fim foi gerada também uma função para separar os dados de treino, escalonar as features e gerar o modelo LightGBM com o dataframe, e salvá-lo como pickle.

1. 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.

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

Github: https://github.com/FerreiraVini/INDICIUM\_DataScience\_2024

# 

# 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 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