

# Desafio de Previsão de Precificação - Plataforma de Aluguéis Temporários

## Introdução

O objetivo do desafio é construir um modelo preditivo capaz de prever preços de aluguéis temporários com base nos dados fornecidos. Nesse documento consta o seguinte:

1. Análise Exploratória de Dados (EDA - Exploratory Data Analysis)
2. Respostas para perguntas específicas sobre as variáveis e seus impactos nos preços.
3. Modelagem e avaliação de modelos preditivos.
4. Sugestão de preço para um apartamento específico.
5. Armazenamento do modelo final.

## 1. Análise Exploratória dos Dados (EDA)

### Características Gerais:

- O conjunto de dados possui **16 colunas**, incluindo informações sobre a propriedade, anfitrião e disponibilidade de aluguel.
- Algumas colunas continham valores ausentes, como `ultima_review` e `reviews_por_mes`, que foram tratados.

### Insights Principais:

1. **Distribuição dos Preços:**
  - A maior parte dos preços está abaixo de \$300 por noite.
  - Outliers elevados foram identificados e tratados (valores acima de \$10.000).
2. **Preço por Grupo de Bairro:**
  - **Manhattan** apresenta a maior média de preço (\$178,94), seguido por **Brooklyn** (\$117,81).
  - **Bronx** possui os preços mais baixos (\$85,28).
3. **Correlação:**
  - `disponibilidade_365` possui uma correlação moderada positiva com o preço (0.23).
  - Outras variáveis numéricas possuem baixa correlação direta com o preço.

## 2. Respostas a perguntas específicas

### a. Onde seria adequado investir em um apartamento para alugar?

- **Manhattan** se destaca como uma escolha melhor devido à maior lucratividade do aluguel.
- Em Manhattan, **Midtown** e **Upper East Side** são áreas muito procuradas.

### b. Como o mínimo e a disponibilidade ao longo do ano afetam os preços?

- O número mínimo de noites impacta um pouco os preços, mas valores extremos (>30 noites) foram modelados para evitar distorções.
- `Availability_365` é moderadamente correlacionado com os preços no sentido de que, em menor extensão, propriedades com mais dias de disponibilidade têm valores mais altos.

### c. Existe alguma correlação entre o nome do local e as propriedades mais caras?

- Sim. A análise textual indica que palavras como **"luxo"**, **"cobertura"** e **"inteiro"** são correlacionadas com preços altos.

## 3. Modelagem preditiva

### Modelo de teste:

#### Regressão linear (linha de base):

- **MSE:** 13.325,61
- **MAE:** 76,97
- **R²:** 0,032
- Desempenho abaixo do padrão, pois não conseguiu capturar relacionamentos não lineares.

#### Gradient Boosting Regressor:

- **MSE:** 11.909,65
- **MAE:** 71,47
- **R²:** 0,135
- Melhor desempenho devido a relações não lineares entre outras variáveis.

Modelo Final Usado:

O **Gradient Boosting Regressor** foi escolhido explicando **13,5% da variação de preço** ao atingir um erro absoluto médio de **\$71,47**.

---

4. Preço Sugerido para Apartamento de Exemplo

Descrição do apartamento:

- **Localização:** Midtown, Manhattan
- **Tipo:** Casa/apto inteiro
- **Disponibilidade:** 355 dias por ano
- **Número de avaliações:** 45

Previsão: O modelo **Gradient Boosting** deu um preço previsto de **\$178,79 por noite**.

---

5. Armazenamento do modelo

O modelo final foi salvo no formato **.pkl** para previsões futuras e reutilização.

---

6. Conclusão

O modelo **Gradient Boosting** demonstrou melhor capacidade de capturar relacionamentos não lineares nos dados, resultando em uma previsão de mais precisa.