

# MACHINE LEARNING

Santander Coders 2024



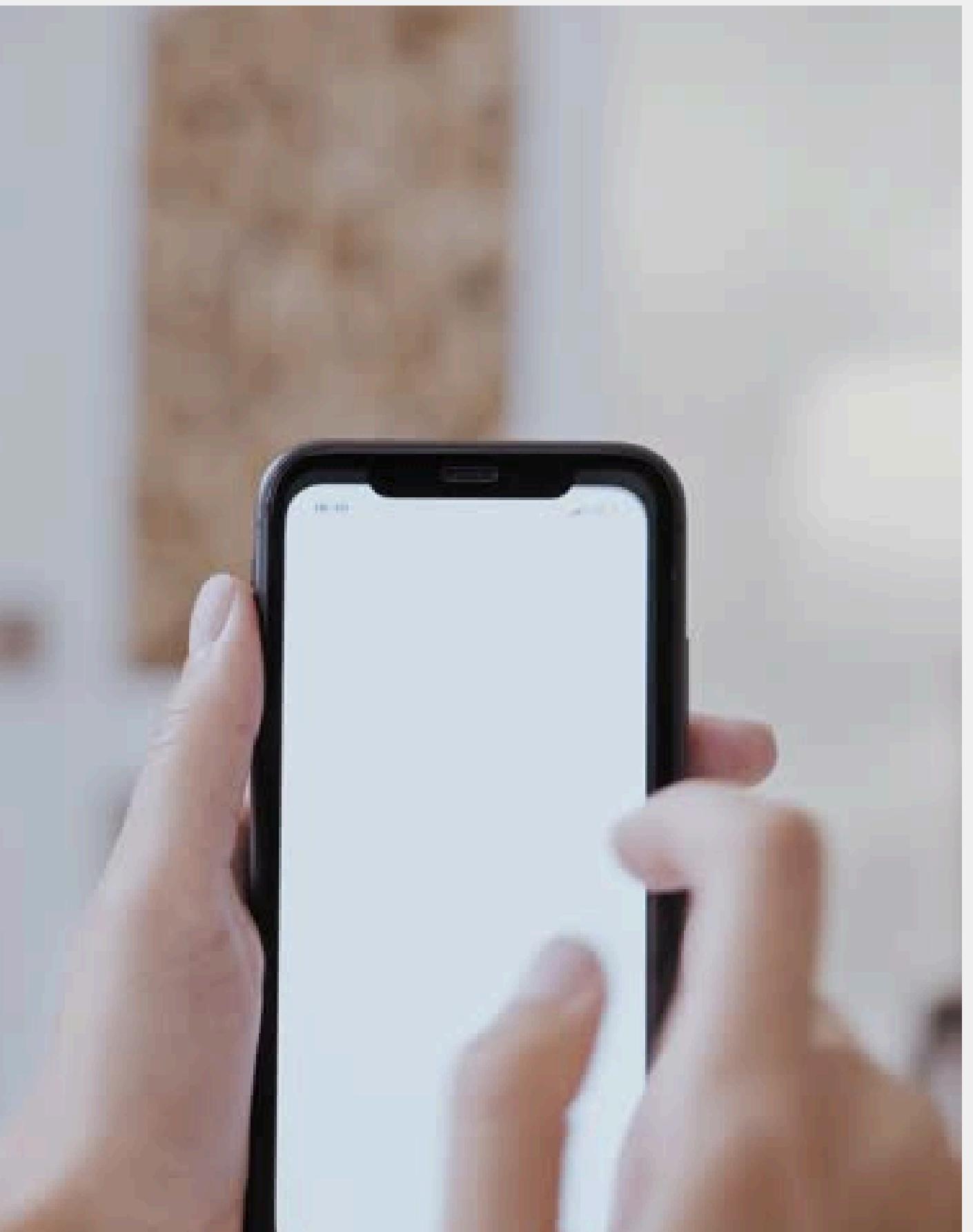
Profesor: Alex Lima  
Marcos Carvalho  
Mateus Gouveia  
Michael Florentino  
Murilo Silva  
Olga Osorio

# INTRODUÇÃO

Hoje vamos apresentar um projeto de Machine Learning focado em previsão de preços para hospedagens, usando um conjunto de dados que inclui variáveis como localização, tipo de quarto, número de avaliações e outros fatores que influenciam no valor das diárias. Utilizei diversos modelos de aprendizado de máquina, incluindo Decision Tree, Random Forest, Gradient Boosting e XGBoost, para comparar o desempenho e encontrar a melhor abordagem para o problema.

Além disso, buscamos otimizar alguns desses modelos para melhorar sua precisão e reduzir o erro de previsão.

Objetivo: Desenvolver um modelo de regressão para prever o preço de uma hospedagem com base em suas características .



# SOBRE O DATASET



Airbnb listings and metrics in NYC, NY, USA (2019)

- Fonte: Kaggle.
- Este dataset contém informações sobre propriedades de aluguel do AirBnB em Nova Iorque. Ele abrange uma variedade de dados sobre as propriedades e os anfitriões, incluindo características de localização, preço, tipo de quarto, avaliações e disponibilidade.
- Tamanho do Conjunto de Dados: 48.895 amostras.
- Principais Colunas:
- id: Identificador único da propriedade.
- name: Nome da propriedade (16 valores faltantes).
- host\_id: Identificador do anfitrião.
- host\_name: Nome do anfitrião (21 valores faltantes).
- neighbourhood\_group: Grupo de bairros em Nova Iorque.
- neighbourhood: Nome do bairro (ex.: Manhattan, Brooklyn).
- latitude e longitude: Coordenadas geográficas da propriedade.
- room\_type: Tipo de quarto disponível (ex.: inteiro, privado, compartilhado).
- price: Preço da diária da propriedade (nenhum valor faltante).
- minimum\_nights: Número mínimo de noites para reserva (nenhum valor faltante).
- number\_of\_reviews: Número total de avaliações da propriedade (nenhum valor faltante).
- last\_review: Data da última avaliação (10.052 valores faltantes).
- reviews\_per\_month: Avaliações mensais (10.052 valores faltantes).
- calculated\_host\_listings\_count: Número total de propriedades listadas pelo anfitrião.
- availability\_365: Disponibilidade do imóvel durante o ano (365 indica o ano todo disponível).



# NOVA IORKE AIRBNB

## Análise de Valores Faltantes:

Algumas colunas possuem valores faltantes, como name (16 valores faltantes), host\_name (21 valores faltantes), last\_review e reviews\_per\_month (10.052 valores faltantes).

## Tratamento de Valores Faltantes.

Dentro do tratamento, várias etapas de pré-processamento e modelagem de dados foram realizadas para prever preços com base em variáveis de um dataset. Primeiramente, são tratados valores ausentes, variáveis categóricas são convertidas em numéricas e os dados são normalizados. Em seguida, são treinados e avaliados diferentes modelos de regressão (Árvore de Decisão, Random Forest, XGBoost, Gradient Boosting), ajustando-se os hiperparâmetros com RandomizedSearchCV. Também é realizada a visualização da importância das variáveis e explorada a relação entre preço e características como bairro e tipo de quarto. Por fim, realiza-se uma análise de classificação binária com Random Forest, incluindo a Curva ROC e a matriz de confusão.

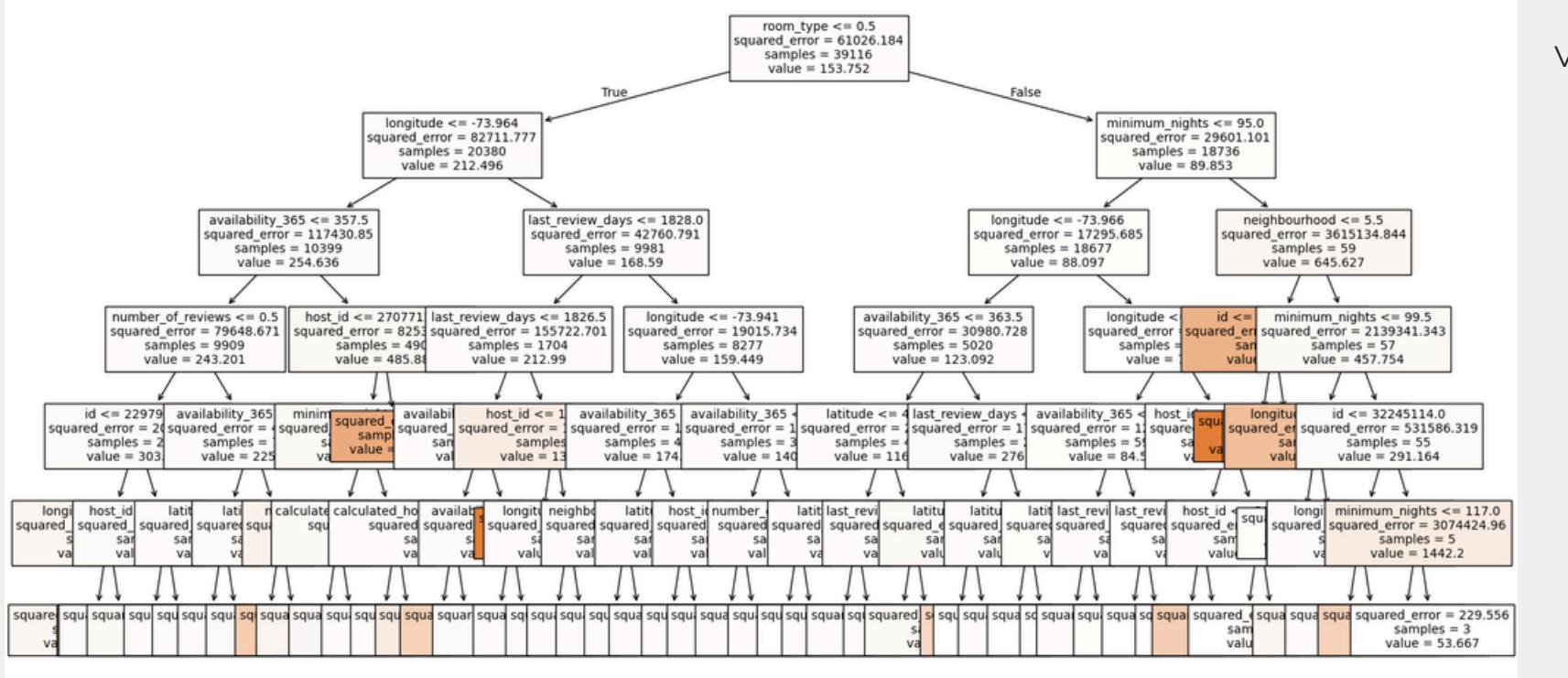
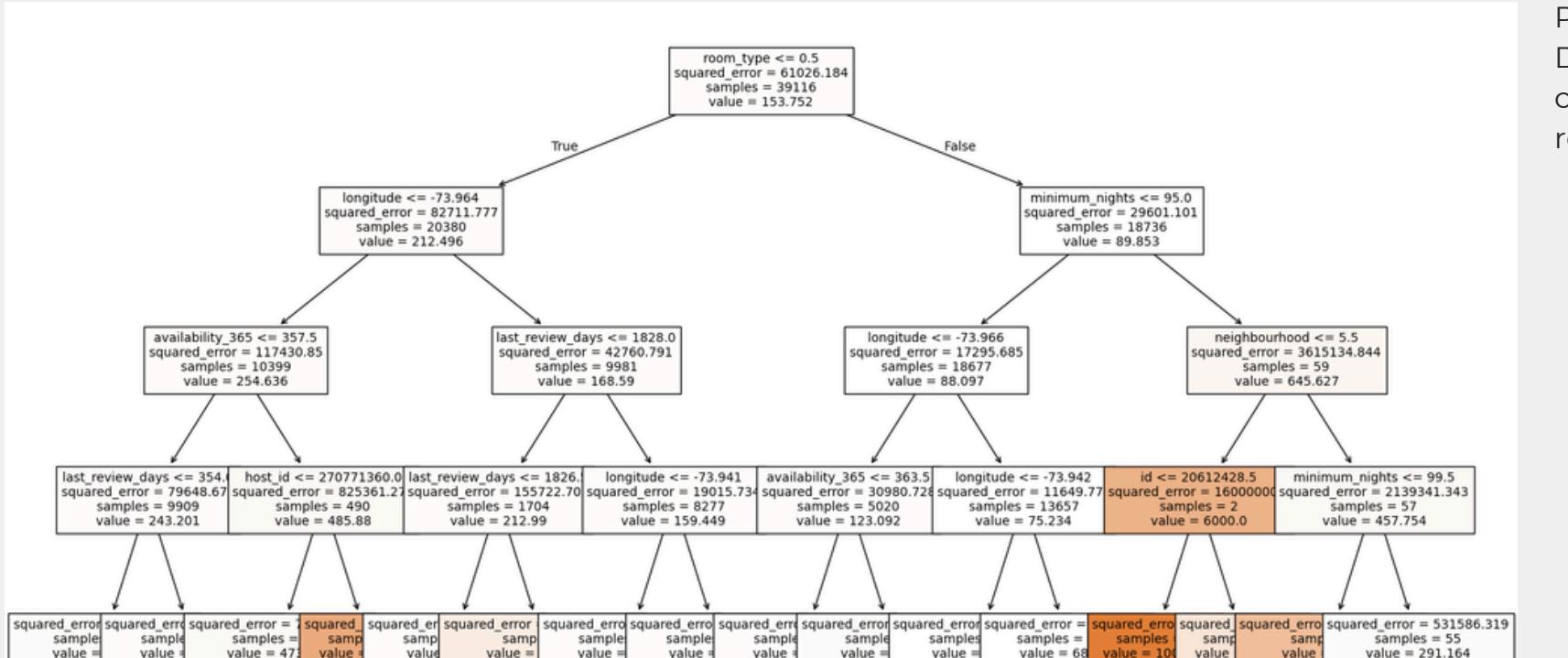
## Análise de Dados Categóricos

No tratamento de dados categóricos, variáveis como "bairro" e "tipo de quarto" foram convertidas para formato numérico utilizando Label Encoding. A análise exploratória revelou como essas variáveis se relacionam com o preço, identificando padrões significativos. A visualização da importância das variáveis e a análise das médias de preço por categoria aprimoraram a modelagem. Esse processo garantiu que as variáveis categóricas fossem integradas de forma eficiente aos modelos.

# MODELOS DE MACHINE LEARNING

## DECISION TREE

Santander Coders 2024



A prever os preços das hospedagens, utilizamos o modelo `DecisionTreeRegressor` com profundidades máximas de 6 e 4, e comparamos o desempenho entre ambas. Abaixo, estão os resultados obtidos:

Profundidade Máxima = 6

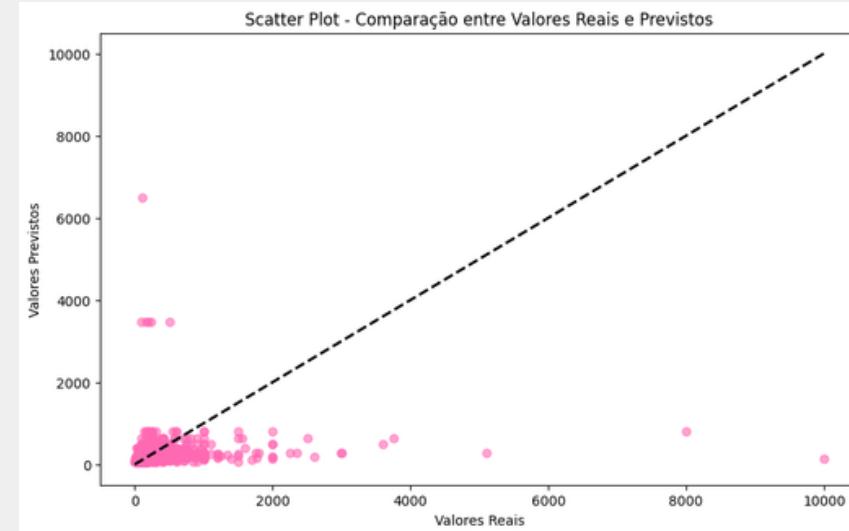
- Erro Quadrático Médio (MSE): 46.319
  - Com uma profundidade de 6, o modelo conseguiu capturar mais detalhes, porém, apresentou um ligeiro aumento no MSE devido ao possível overfitting. A árvore de decisão mostra uma maior complexidade, com mais divisões, o que melhora as previsões para alguns dados, mas pode gerar imprecisões em outros casos.

Profundidade Máxima = 4

- Erro Quadrático Médio (MSE): 43.021
  - Com profundidade reduzida, o modelo simplifica as decisões, capturando apenas as divisões principais. Isso ajuda a evitar o overfitting e gera um MSE mais baixo, o que indica uma melhor generalização para dados novos.

#### Visualização da Árvore de Decisão:

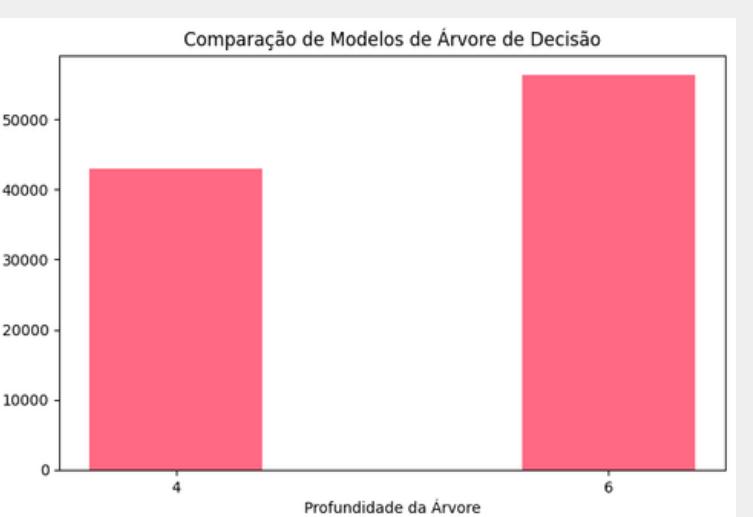
- Exibe as divisões principais feitas pelo modelo para ambas as profundidades. Na apresentação, destaque como a profundidade impacta na granularidade das previsões.



Scatter Plot (Valores Reais vs. Previstos):

- Um gráfico de dispersão compara os valores reais e previstos pelo modelo, permitindo observar a proximidade entre as previsões e os valores reais. A linha  $y = x$  serve como referência, onde pontos próximos a ela indicam previsões mais precisas.

Profundidade Máxima = 6



# MODELOS DE MACHINE LEARNING RANDOM FOREST

Santander Coders 2024

Para prever o preço das hospedagens, utilizamos o modelo Random Forest Regressor. Esse modelo combina várias árvores de decisão independentes para melhorar a precisão e reduzir o risco de overfitting, resultando em previsões mais robustas. Aqui está o processo e os resultados obtidos:

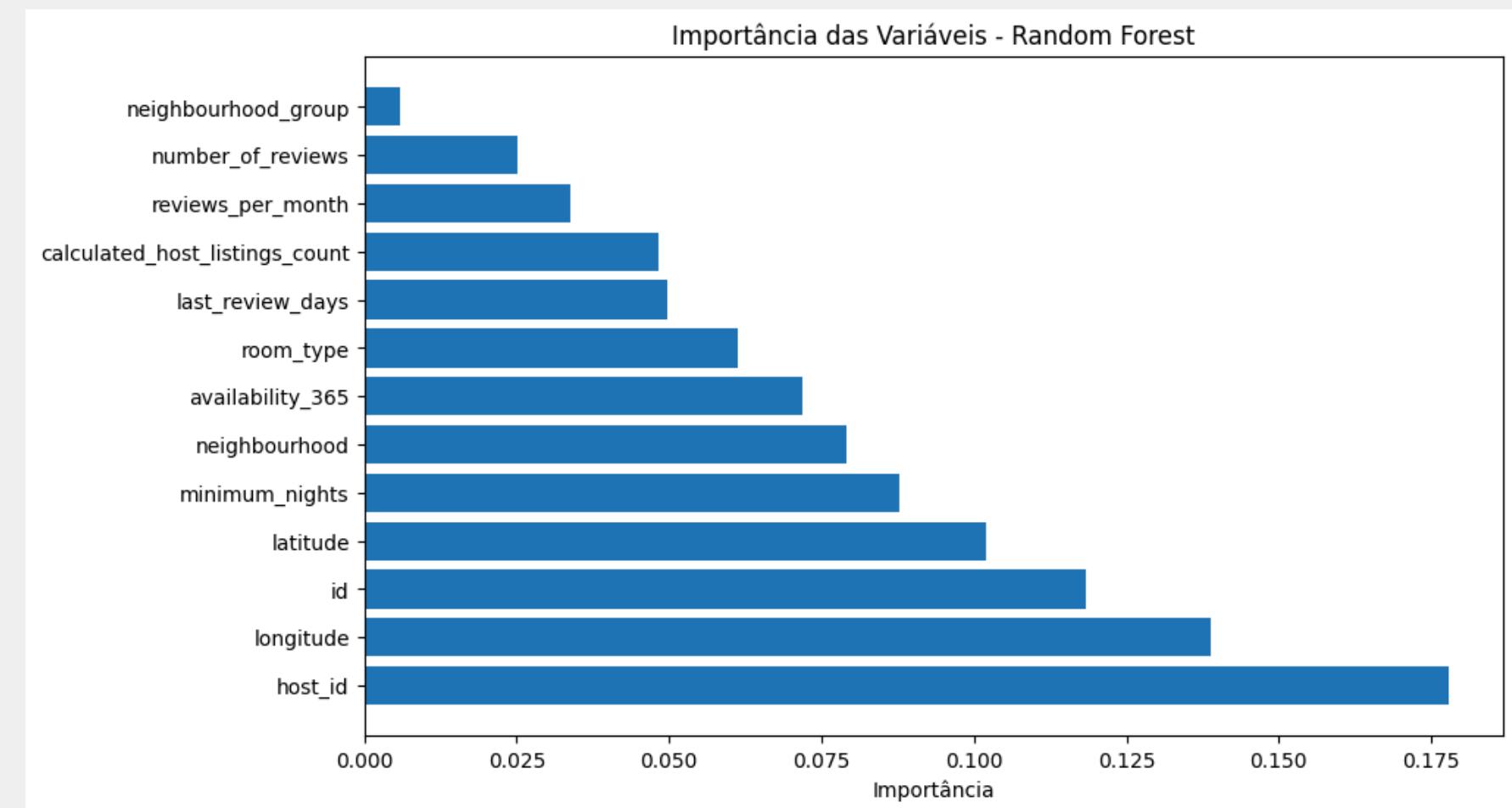
## 1. Treinamento Inicial do Random Forest:

- Configuração: Treinamos o modelo inicialmente com 100 árvores.
- Erro Quadrático Médio (MSE): O modelo apresentou um MSE de aproximadamente 35.146, o que mostra uma precisão melhor do que a Decision Tree isolada.

## 2. Ajuste de Parâmetros do Random Forest:

- Configuração Ajustada: Alteramos os parâmetros para incluir 200 árvores (`n_estimators=200`), profundidade máxima de 10 (`max_depth=10`) e um mínimo de 20 amostras para dividir um nó (`min_samples_split=20`).
- Erro Quadrático Médio (MSE) do Modelo Ajustado: O MSE foi 35.230. Apesar do ajuste, o MSE permaneceu próximo do valor inicial, indicando que o modelo já estava próximo de seu desempenho ótimo com a configuração inicial.

Esses valores de MSE sugerem que o Random Forest consegue capturar bem as variações dos dados sem superajustar muito, o que o torna uma boa escolha para previsões com dados de média complexidade.



Um dos pontos fortes do modelo Random Forest é a capacidade de identificar as variáveis mais importantes para a previsão. No gráfico de Importância das Variáveis, observamos que:

- `id` e `longitude` aparecem como as variáveis mais influentes na determinação do preço.
- `neighbourhood_group` tem uma importância menor, indicando que, embora o bairro influencie o valor, ele não é tão determinante quanto as coordenadas exatas (`longitude`) e o identificador específico (`id`) do imóvel.

Essa análise de importância ajuda a entender que as características de localização exata (`latitude` e `longitude`) e características específicas dos imóveis são fatores mais decisivos na especificação do que variáveis como o grupo do bairro.

# MODELOS DE MACHINE LEARNING

# GRADIENT BOOSTING

Santander Coders 2024

Utilizamos o Gradient Boosting Regressor para prever os preços das hospedagens. Esse modelo cria árvores de decisão sequenciais, onde cada nova árvore tenta corrigir os erros da árvore anterior.

## 1. Erro Quadrático Médio (MSE):

- O Gradient Boosting alcançou um MSE de 36.092, o que representa uma melhoria em relação aos modelos anteriores de Decision Tree e Random Forest. Isso sugere que o modelo conseguiu captar padrões adicionais nos dados, embora ainda com um erro considerável.

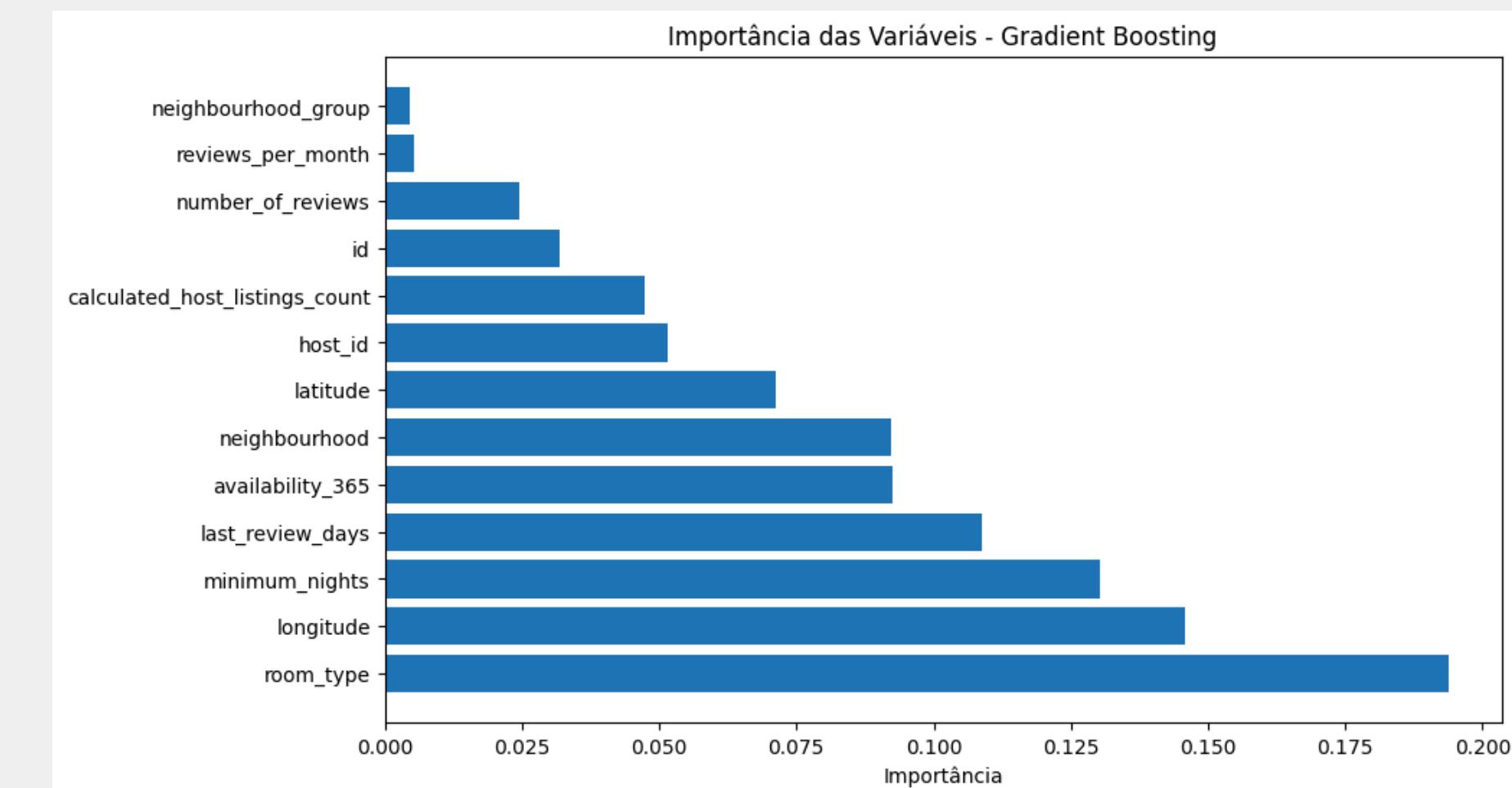
## Importância das Variáveis

O gráfico de importância das variáveis do Gradient Boosting revela quais variáveis mais influenciam o modelo na previsão dos preços:

**room\_type** (tipo de quarto): É a variável mais importante, indicando que o tipo de quarto tem um impacto significativo no valor das hospedagens.

**longitude**: A posição geográfica também se destaca, possivelmente porque ela ajuda a identificar regiões mais ou menos valorizadas dentro dos bairros.

**minimum\_nights** e **neighbourhood\_group**: Embora também relevantes, têm menor impacto no modelo em comparação com as variáveis acima.



# CLASSIFICAÇÃO BINÁRIA

Santander Coders 2024

## Classificação Binária com Curva ROC e AUC

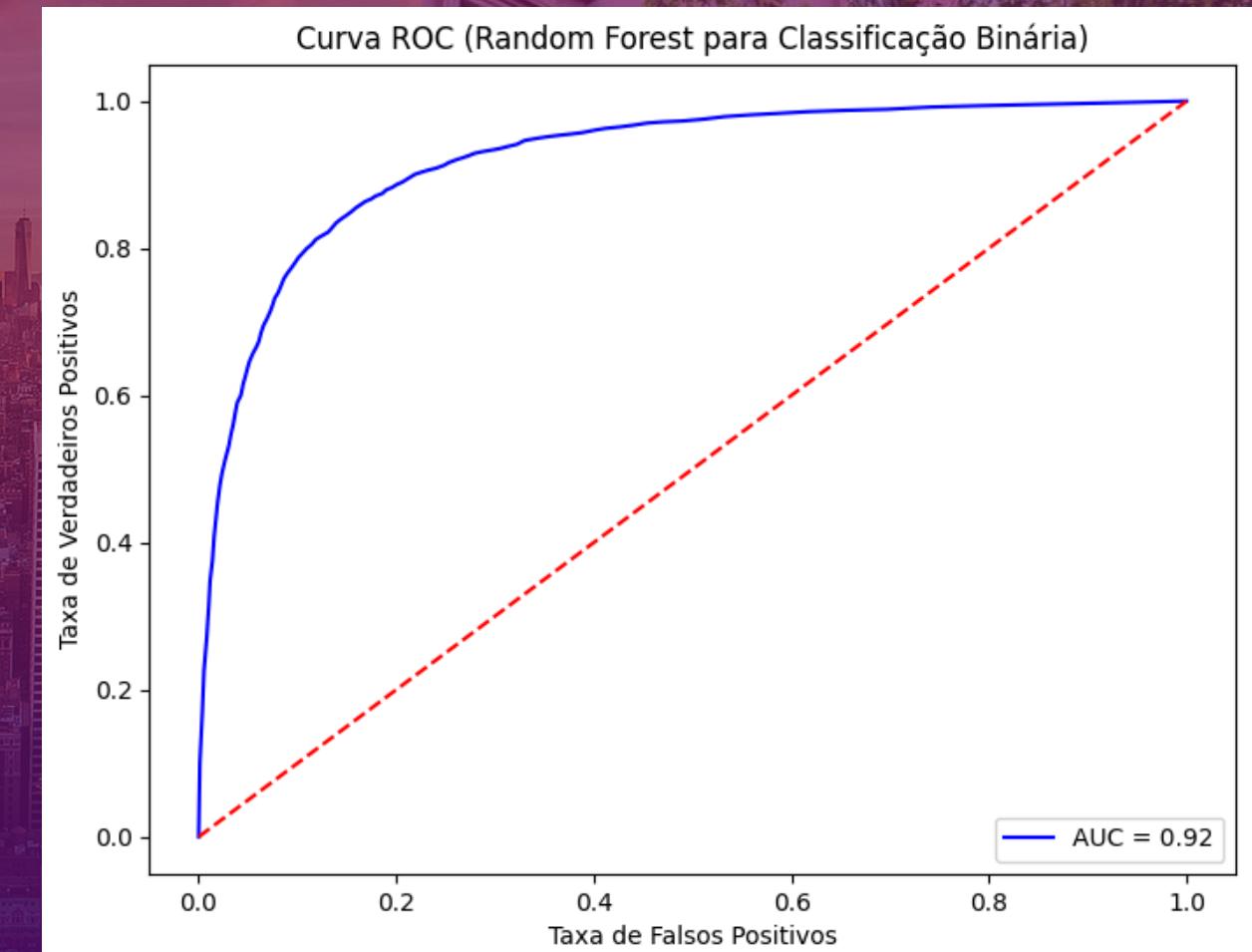
Para simplificar a análise, fizemos uma classificação binária, onde definimos os preços maiores que um certo valor como “altos” (classe 1) e os preços menores como “baixos” (classe 0). Utilizando o Random Forest para Classificação Binária, geramos a Curva ROC e calculamos a AUC.

### 1. Curva ROC:

- A Curva ROC mostra a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos. Ela ajuda a visualizar a performance do modelo em diferentes limiares de decisão.
- A curva inclinada para o canto superior esquerdo indica que o modelo consegue distinguir bem entre as classes “alto” e “baixo”.

### 2. AUC:

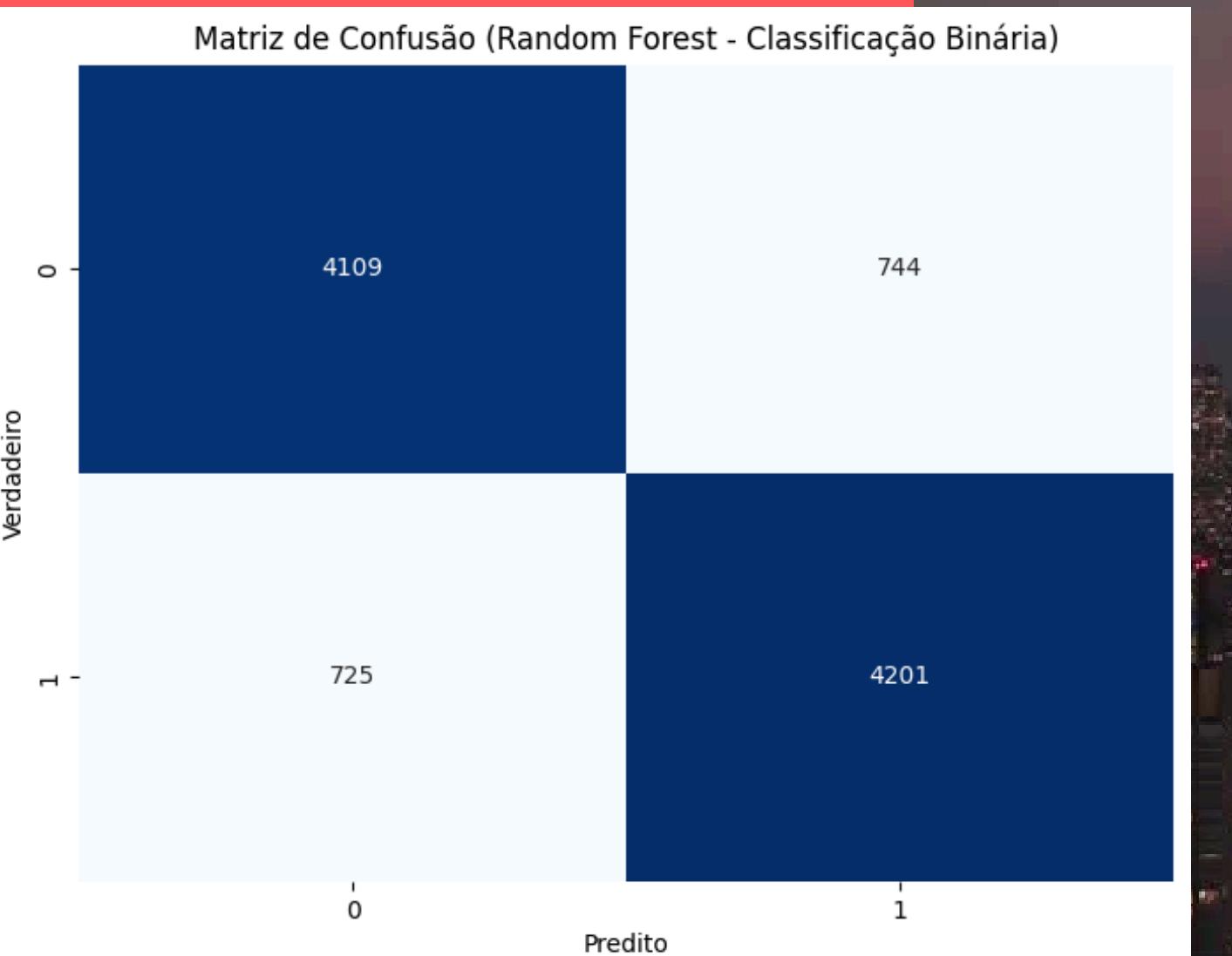
- AUC (Área Sob a Curva) foi de 0,92, indicando que o modelo é capaz de separar bem as classes, com uma precisão alta. Valores de AUC próximos de 1 indicam que o modelo é excelente na tarefa de classificação.



# Random Forest

# MATRIZ DE CONFUSÃO

- Descrição do Modelo:
- Tipo de Classificação: Classificação Binária usando Random Forest
- Parâmetros Importantes:  
n\_estimators=100, random\_state=42
- Critério de Classificação: Limiar de 100 para a variável dependente y
- Matriz de Confusão:
- Verdadeiro Positivo (4201) e Verdadeiro Negativo (4109)
- Falso Positivo (744) e Falso Negativo (725)
- Métricas de Avaliação:
- Precisão: 85%
- Revocação: 85%
- F1-Score: 85%
- Acurácia Geral: 85%
- Observações Finais:
  - O modelo teve um desempenho equilibrado nas classes, sem viés evidente entre elas.
  - A acurácia de 85% indica um bom desempenho, mas os valores de Falsos Positivos e Falsos Negativos mostram que ainda há margem para melhorias.



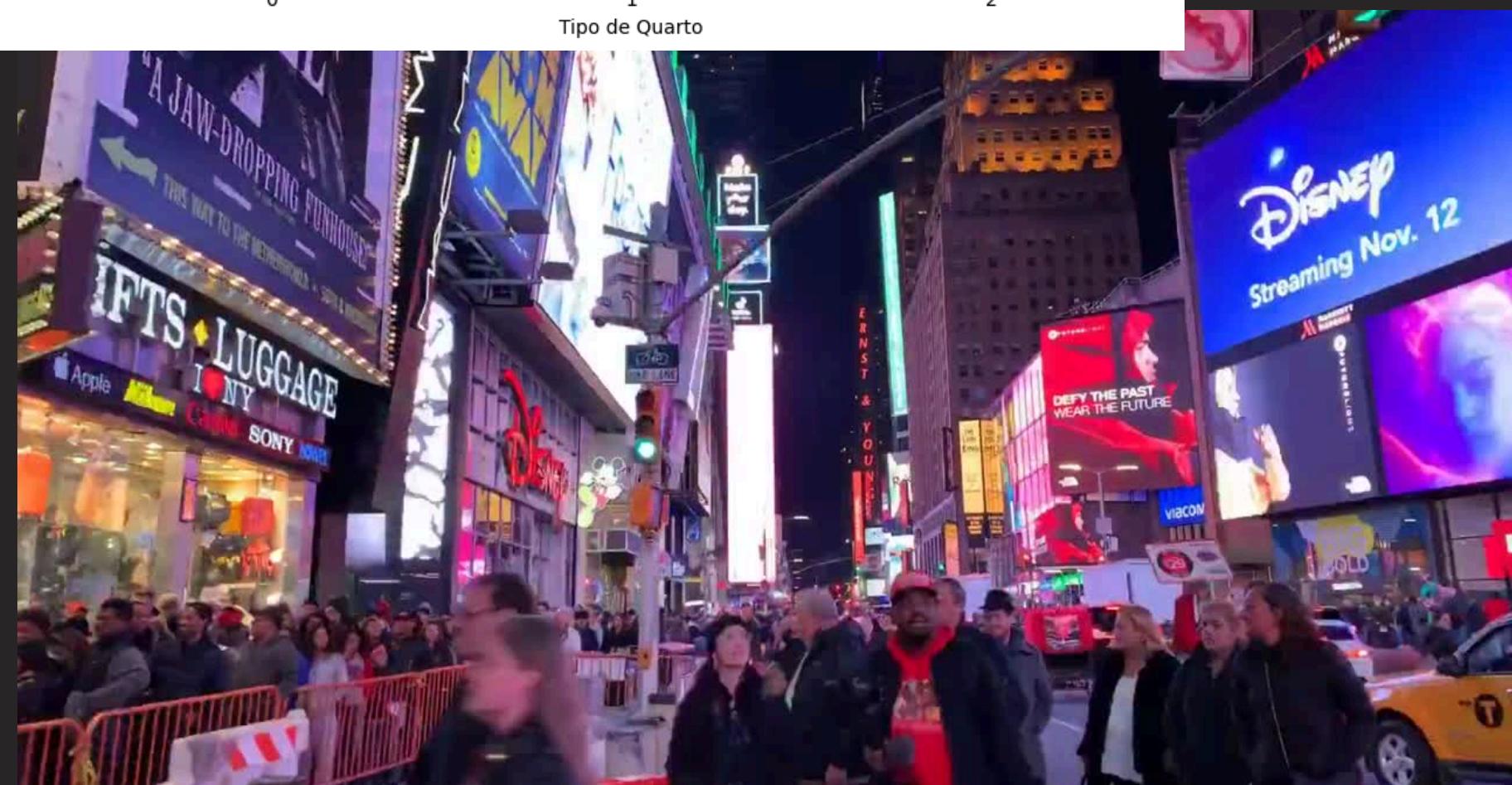
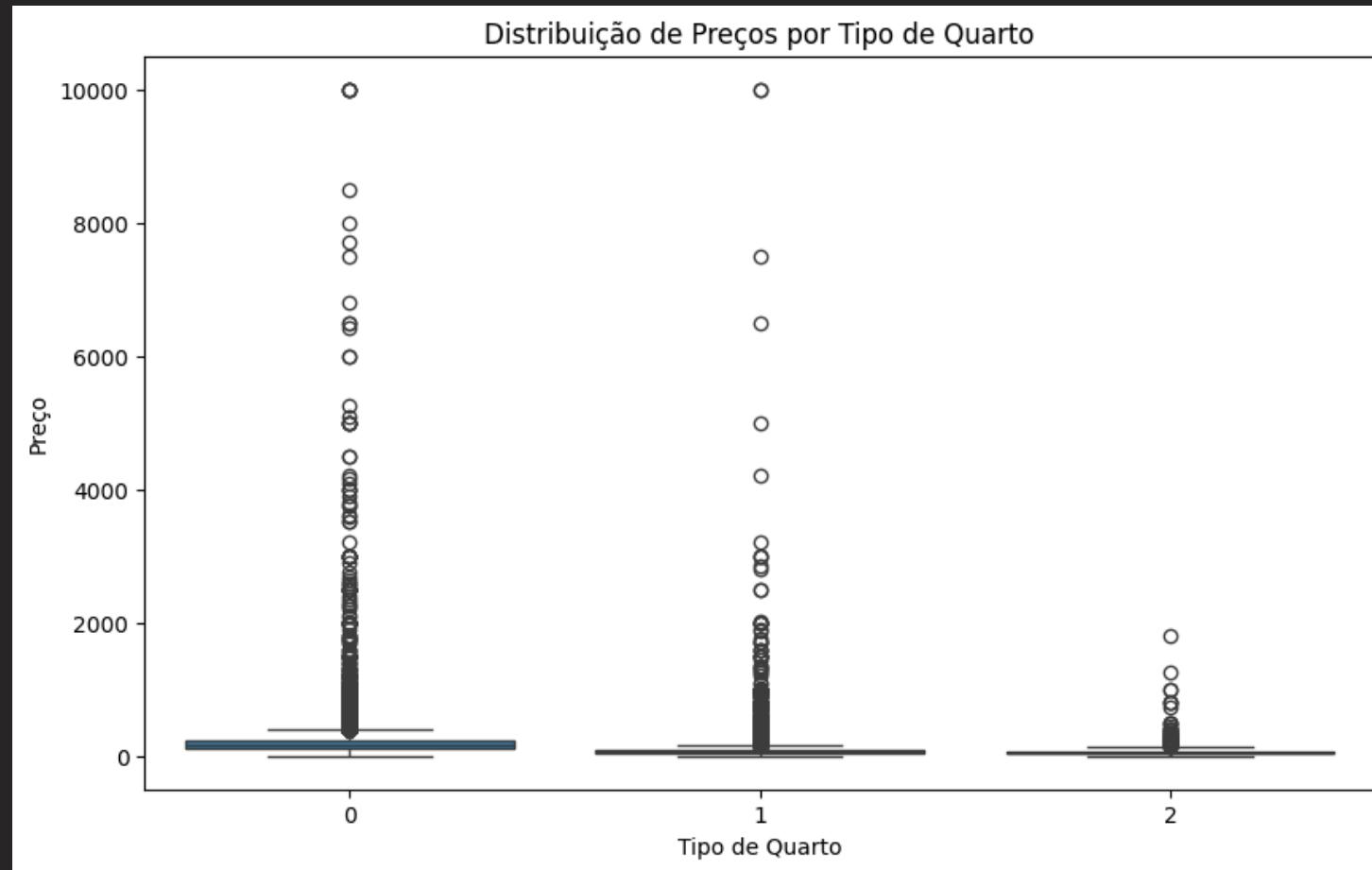
Relatório de Classificação:				
	precision	recall	f1-score	support
0	0.85	0.85	0.85	4853
1	0.85	0.85	0.85	4926
accuracy			0.85	9779
macro avg	0.85	0.85	0.85	9779
weighted avg	0.85	0.85	0.85	9779

## 1. Gráfico de Dispersão: Relação entre Reviews por Mês e Preço

Este gráfico mostra a relação entre o número de reviews por mês e o preço de acomodações.

- Análise do Gráfico:

- A maioria das acomodações tem preços mais baixos (entre 0 e 2000) e um número pequeno de reviews por mês (entre 0 e 10).
- Há poucos pontos com preços e número de reviews altos, indicando que a maioria das acomodações com preços elevados não recebe tantos reviews mensais.
- Esse padrão sugere que preços mais acessíveis estão associados a uma demanda maior, enquanto acomodações mais caras tendem a ter menos reviews.

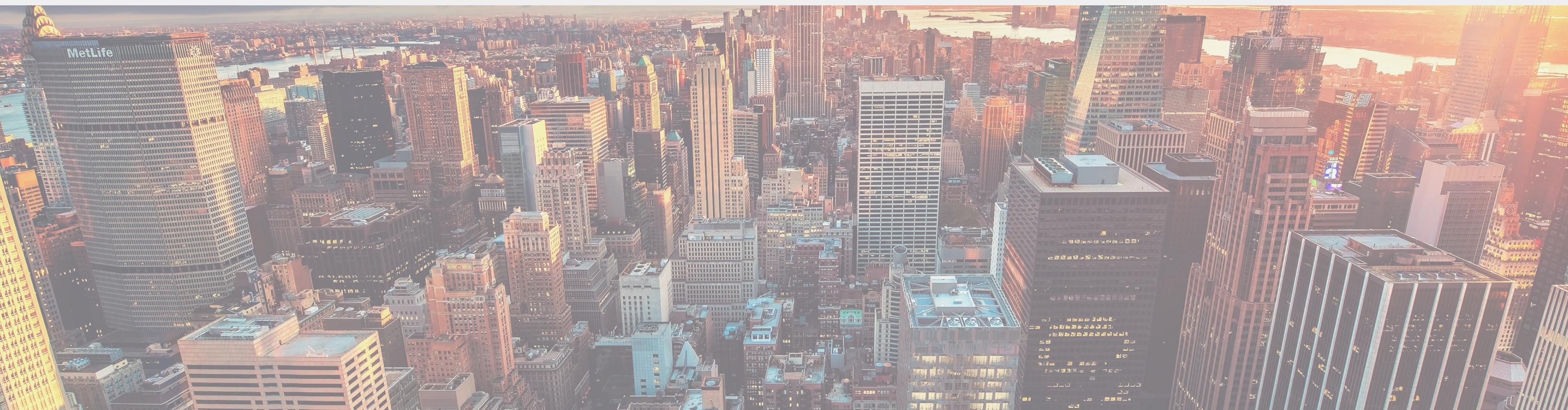


# Airbnb GRÁFICOS

# CONCLUÕES

Airbnb

- Insights: Modelos de ensemble (Random Forest, XGBoost) apresentaram melhor precisão.
- Limitações: Dados de 2019, possivelmente desatualizados; viés em áreas com maior concentração de anúncios e falta de dados ou dados nulos.
- Próximos Passos: Coletar dados mais recentes e explorar variáveis adicionais para aprimorar a previsão.
- O Random Forest Regressor com 100 árvores foi o melhor modelo para prever os preços das hospedagens, oferecendo o menor MSE e uma boa robustez contra overfitting. Esse resultado mostra que o modelo conseguiu capturar as variações nos dados com um bom equilíbrio entre precisão e generalização.



OBRIGADO

*Obrigado*

