Desafio Cientista de Dados

Modelo de previsão de preços para aluguéis temporários na cidade de Nova York.

Autor: Lucas Oliveira

João Monlevade
Fevereiro de 2024

1 Introdução

Este relatório aborda a criação de uma plataforma de aluguéis temporários em Nova York, com foco na estratégia de precificação. O trabalho envolveu uma análise exploratória dos dados do maior concorrente, seguida pelo desenvolvimento e validação de um modelo preditivo para prever os preços dos aluguéis. A abordagem busca não apenas compreender os padrões do mercado, mas também avaliar a eficácia do modelo utilizando métricas específicas, visando melhorar a estratégia de precificação do cliente.

2 Objetivo

O objetivo deste trabalho é desenvolver um modelo de previsão de preços para aluguéis temporários em Nova York utilizando o algoritmo *Linear Regression*. Inicialmente, realizaremos uma análise exploratória dos dados do concorrente para identificar padrões. Em seguida, implementaremos o modelo Linear Regression, visando ajustá-lo aos dados disponíveis. A avaliação do modelo será conduzida por meio das métricasMAE e MSE, proporcionando uma análise abrangente da capacidade preditiva e contribuindo para aprimorar a estratégia de precificação do cliente.

3 Análise dos Dados

Nesta seção, iremos explorar as correlações entre as variáveis dos aluguéis temporários em Nova York, considerando fatores como localização, tamanho, comodidades e publicações. Utilizaremos medidas estatísticas, incluindo o cálculo da média, desvio padrão e coeficientes de correlação, para identificar padrões e relações significativas. Além disso, empregaremos visualizações gráficas, como mapas de calor, para tornar a análise mais acessível. Essas abordagens estatísticas proporcionarão insights valiosos para aprimorar a precisão do modelo de previsão e informar a estratégia de precificação.

3.1 Heatmap e Matriz de correlação.

A matriz de correlação é uma tabela que mostra as correlações entre variáveis. Ela destaca o grau e a direção da relação linear entre duas variáveis, variando de -1 (correlação negativa perfeita) a 1 (correlação positiva perfeita). Um valor próximo a 0 indica uma correlação fraca.

O heatmap, ou mapa de calor, é uma representação gráfica da matriz de correlação. Ele utiliza cores para destacar visualmente os diferentes níveis de correlação. Valores mais altos podem ser representados por cores mais intensas, como vermelho, enquanto valores mais baixos podem ser representados por cores mais suaves, como azul. O heatmap facilita a identificação rápida de padrões e relações entre variáveis em conjuntos de dados extensos, sendo uma ferramenta valiosa na análise exploratória de dados.

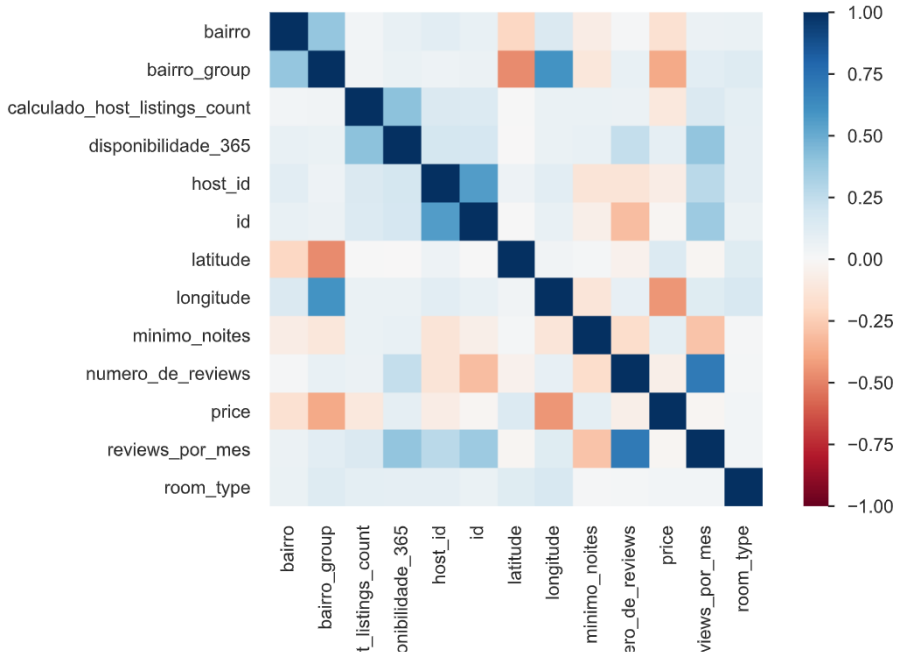

Figura 1: Heatmap das features do conjunto de dados.

Variáveis	bairro	bairro_group	host_listings_count	disponibilidade	host_id	id	latitude	longitude	minimo_noites	numero_de_reviews	price	reviews_por_mes	room_type
bairro	1.000	0.386	0.029	0.074	0.108	0.072	-0.206	0.146	-0.076	0.010	-0.150	0.058	0.066
bairro_group	0.386	1.000	0.037	0.065	0.054	0.055	-0.475	0.600	-0.112	0.078	-0.378	0.103	0.126
host_listings_count	0.029	0.037	1.000	0.407	0.147	0.135	0.004	0.064	0.064	0.056	-0.106	0.147	0.097
disponibilidade	0.074	0.065	0.407	1.000	0.173	0.166	-0.007	0.069	0.076	0.237	0.086	0.392	0.087
host_id	0.108	0.054	0.147	0.173	1.000	0.559	0.050	0.109	-0.130	-0.128	-0.072	0.268	0.092
id	0.072	0.055	0.135	0.166	0.559	1.000	0.005	0.071	-0.058	-0.308	-0.021	0.360	0.070
latitude	-0.206	-0.475	0.004	-0.007	0.050	0.005	1.000	0.035	0.022	-0.044	0.136	-0.023	0.117
longitude	0.146	0.600	0.064	0.069	0.109	0.071	0.035	1.000	-0.119	0.080	-0.438	0.119	0.158
minimo_noites	-0.076	-0.112	0.064	0.076	-0.130	-0.058	0.022	-0.119	1.000	-0.175	0.101	-0.289	0.012
numero_de_reviews	0.010	0.078	0.056	0.237	-0.128	-0.308	-0.044	0.080	-0.175	1.000	-0.055	0.706	0.021
price	-0.150	-0.378	-0.106	0.086	-0.072	-0.021	0.136	-0.438	0.101	-0.055	1.000	-0.019	0.025
reviews_por_mes	0.058	0.103	0.147	0.392	0.268	0.360	-0.023	0.119	-0.289	0.706	-0.019	1.000	0.029
room_type	0.066	0.126	0.097	0.087	0.092	0.070	0.117	0.158	0.012	0.021	0.025	0.029	1.000

Tabela 1: Matriz de Correlação

Ao analisar o heatmap e a tabela de correlação, é possível identificar padrões de correlação entre as variáveis.

Correlações Mais Fortes:
bairro e bairro_group: Correlação positiva de 0.386.
bairro_group e longitude: Correlação positiva mais forte de 0.600.
numero_de_reviews e reviews_por_mes: Correlação positiva mais forte de 0.706.
Correlações Negativas Mais Fortes:
bairro e latitude: Correlação negativa de -0.206.
bairro_group e latitude: Correlação negativa mais forte de -0.475.
minimo_noites e longitude: Correlação negativa de -0.119.
Correlações Próximas a Zero:
calculado_host_listings_count e disponibilidade_365: Correlação próxima a zero (0.037).
id e price: Correlação próxima a zero (-0.021).
Essas observações indicam quais variáveis têm uma forte relação linear positiva ou negativa entre si, bem como aquelas que não possuem correlação significativa. No entanto, vale ressaltar que correlação não implica causalidade, e análises mais avançadas podem ser necessárias para compreender completamente as relações entre as variáveis.

4 EAD e Questionamentos

4.1 Investimento em imóveis

A primeira pergunta feita foi "Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?"

A decisão de compra seria mais apropriada em regiões onde os imóveis apresentam uma menor disponibilidade ao longo do ano, inclusive podendo alcançar zero datas disponíveis. Essa condição sugere a possibilidade de o local ficar menos tempo ocioso, indicando potencialmente uma demanda mais consistente e uma maior atratividade para investimentos imobiliários na plataforma.

4.1.1 Interferencia no preço

A segunda pergunta foi se "O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?"

Ao observar o Heatmap e a matriz de correlação, percebe-se que tanto o número mínimo de noites quanto a disponibilidade ao longo do ano apresentam correlações relativamente baixas com o preço (0.101 e 0.086, respectivamente). Esses resultados indicam que essas variáveis têm uma influência limitada nos preços dos aluguéis. Por outro lado, as correlações mais expressivas estão associadas a dados geográficos, como group_bairro e longitude. Isso sugere que, na determinação dos preços, a localização exerce uma influência mais significativa do que o tempo mínimo de estadia ou a disponibilidade ao longo do ano..

4.1.2 Padrão em anuncios

Respondendo ao terceiro questionamento se "Existe algum padrão no texto do nome do local para lugares de mais alto valor?"

Em 190 instancias analizadas, as palavras mais relevantes que foram publicadas foram loft 24 repetições (2.2%) , townhouse com 22 repetições (2.0%),luxury com repetições 18 (1.6%) e private 14 repetições (1.3%). As demais palavras não possuem muita significancia nesta analize.

4.2 Desenvolvimento do modelo de *Regressão Linear*

O modelo de regressão linear é uma técnica estatística que busca modelar a relação linear entre uma variável dependente e uma ou mais variáveis independentes. Na forma simples, a regressão linear assume que a relação entre as variáveis pode ser representada por uma linha reta. O objetivo é encontrar os coeficientes que minimizam a diferença entre os valores previstos pelo modelo e os valores reais observados. Esse modelo é amplamente utilizado para prever ou explicar o valor de uma variável dependente com base em uma ou mais variáveis independentes, proporcionando insights sobre a relação entre elas..

4.2.1 Escolha de variáveis

Para realizar a previsão do preço com base nos dados fornecidos, foram selecionadas as seguintes variáveis para treinamento do modelo: bairro_group, bairro, latitude, longitude, room_type, price, minimo_noites, numero_de_reviews, reviews_por_mes, calculado_host_listings_count, e disponibilidade_365. Foram excluídas as features relacionadas à identificação para focar nas características mais relevantes.

4.2.2 Transformações em variáveis

Transformações foram aplicadas para lidar com variáveis categóricas, incluindo o mapeamento de valores únicos nas features bairro_group e bairro. Além disso, as variáveis foram normalizadas utilizando o MinMaxScaler e padronizadas com o StandardScaler. Essas etapas são essenciais para garantir que as variáveis estejam em escalas apropriadas para o modelo de regressão linear.

4.2.3 Tipo do Problema

O tipo de problema abordado é de regressão, onde buscamos prever valores contínuos, como o preço dos aluguéis. O modelo escolhido foi o LinearRegression do Scikit-learn, treinado utilizando dados normalizados.

4.2.4 Avaliação do modelo

Ao avaliar o desempenho do modelo, optamos por métricas comuns em problemas de regressão, como o Mean Absolute Error (MAE) e o Mean Squared Error (MSE). Essas métricas proporcionam uma medida quantitativa da precisão do modelo em relação aos valores reais, sendo essenciais para a validação do desempenho do modelo preditivo.

Em resumo, a abordagem adotada busca otimizar a previsão do preço, incorporando variáveis relevantes, transformações adequadas e a escolha de métricas apropriadas para a avaliação do modelo de regressão linear..

4.3 Análise do desempenho do modelo de *Regressão Linear*

A análise das métricas de avaliação dos modelos revela resultados consistentes entre o modelo sem pré-processamento, o modelo com MinMaxScaler e o modelo com StandardScaler. Vamos considerar cada métrica:
Média do Erro Absoluto (MAE):
Sem pré-processamento: 61.51
Com MinMaxScaler: 61.04
Com StandardScaler: 61.04
Observa-se uma ligeira melhoria na média do erro absoluto ao utilizar o pré-processamento, mas a diferença é mínima entre os modelos.

Erro Quadrático Médio (MSE):
Sem pré-processamento: 32040.14
Com MinMaxScaler: 32036.88
Com StandardScaler: 32036.88
Assim como na MAE, não há uma mudança significativa nos resultados do MSE após o pré-processamento dos dados.

Análise Geral:
A variação nas métricas entre os modelos é pequena, indicando que o pré-processamento dos dados com MinMaxScaler e StandardScaler não teve um impacto expressivo na performance do modelo em relação aos dados sem pré-processamento. O MSE, que penaliza mais os erros maiores, apresenta valores relativamente elevados, sugerindo que o modelo pode estar sujeito a

outliers ou variações significativas nos dados. A média do erro absoluto em torno de 61.0 indica que, em média, as previsões estão desviando em cerca de $61 unidades da variável de resposta (preço). Recomendações:

Considerando a consistência nos resultados, pode ser interessante no futuro explorar outros modelos ou técnicas de ajuste de hiperparâmetros para buscar melhorias adicionais na performance. Investigar a presença de outliers nos dados ou explorar técnicas avançadas de feature engineering pode contribuir para aprimorar a qualidade das previsões. Em resumo, embora os modelos apresentem resultados razoáveis, há espaço realização de mais estudos para otimizações adicionais e exploração mais aprofundada das características dos dados para aprimorar a qualidade das previsões.

5 Conclusão

O presente trabalho buscou desenvolver um modelo de previsão de preços para a plataforma de aluguéis temporários na cidade de Nova York, utilizando uma abordagem de regressão linear. Ao longo do processo, foram realizadas análises exploratórias de dados, pré-processamento, seleção de variáveis relevantes, e treinamento do modelo com diferentes estratégias de normalização, incluindo MinMaxScaler e StandardScaler.

A análise exploratória permitiu identificar correlações entre variáveis, fornecendo insights valiosos sobre a influência de fatores como localização e características relacionadas a avaliações na determinação dos preços dos aluguéis. A seleção de variáveis visou incluir elementos cruciais para a previsão, enquanto o pré-processamento buscou ajustar as escalas e lidar com variáveis categóricas de maneira adequada.

Os resultados das métricas de avaliação, incluindo a média do erro absoluto (MAE) e o erro quadrático médio (MSE), revelaram uma consistência nos desempenhos entre o modelo sem pré-processamento e os modelos com MinMaxScaler e StandardScaler. A média do erro absoluto em torno de 61.0 indica uma razoável precisão nas previsões, enquanto os valores elevados de MSE sugerem que o modelo pode ser sensível a outliers ou variações significativas nos dados.

Em resumo, o trabalho proporcionou uma compreensão abrangente da relação entre as variáveis e os preços de aluguéis na plataforma, culminando em um modelo de regressão linear que, embora apresente resultados razoáveis, sugere a necessidade de investigações adicionais para refinamento e otimização.