RELATÓRIO DESAFIO CIENTISTA DE DADOS

PROGRAMA LIGHTHOUSE INDICIUM

Nome Ayrton Bruno Silva Bezerra

1. INTRODUÇÃO

O presente relatório tem como objetivo demonstrar o raciocínio e métodos utilizados para solucionar o Desafio de Ciência de Dados proposto como uma das etapas do processo seletivo para a vaga de estágio.

A finalidade do desafio é, a partir da análise dos dados de um concorrente da empresa a qual eu faria parte, determinar insights sobre o mercado imobiliário de Nova York, especificamente sobre imóveis para aluguel temporário, de modo a possibilitar fazer recomendações, identificar oportunidades e riscos. Além disso, utilizando técnicas matemáticas e estatísticas relacionadas a aprendizado de máquina, treinar um modelo capaz de fazer a previsão do preço do aluguel de novos imóveis que forem adicionados à base de dados.

2. ANÁLISE EXPLORATÓRIA DE DADOS

Esta etapa foi utilizada para entender as variáveis presentes na base e visualizar as correlações entre elas. Não houve problemas em identificar o que representaria cada uma das colunas, visto que seus nomes são autoexplicativos e foi fornecida uma descrição para cada uma delas.

O primeiro passo foi verificar como as características dos imóveis estavam distribuídas em relação a quantidade de dados, variáveis categóricas, numéricas, dados nulos, duplicados e como deveria ser o tratamento adequado para cada uma.

Dado o contexto do problema, dados duplicados em qualquer uma das colunas não seria necessariamente um problema, pois podem representar usuários fazendo mais de um anúncio ou diferentes imóveis que compartilham características em comum como localização, preço ou disponibilidade que são informações valiosíssimas para mapear padrões e preferências dos clientes.

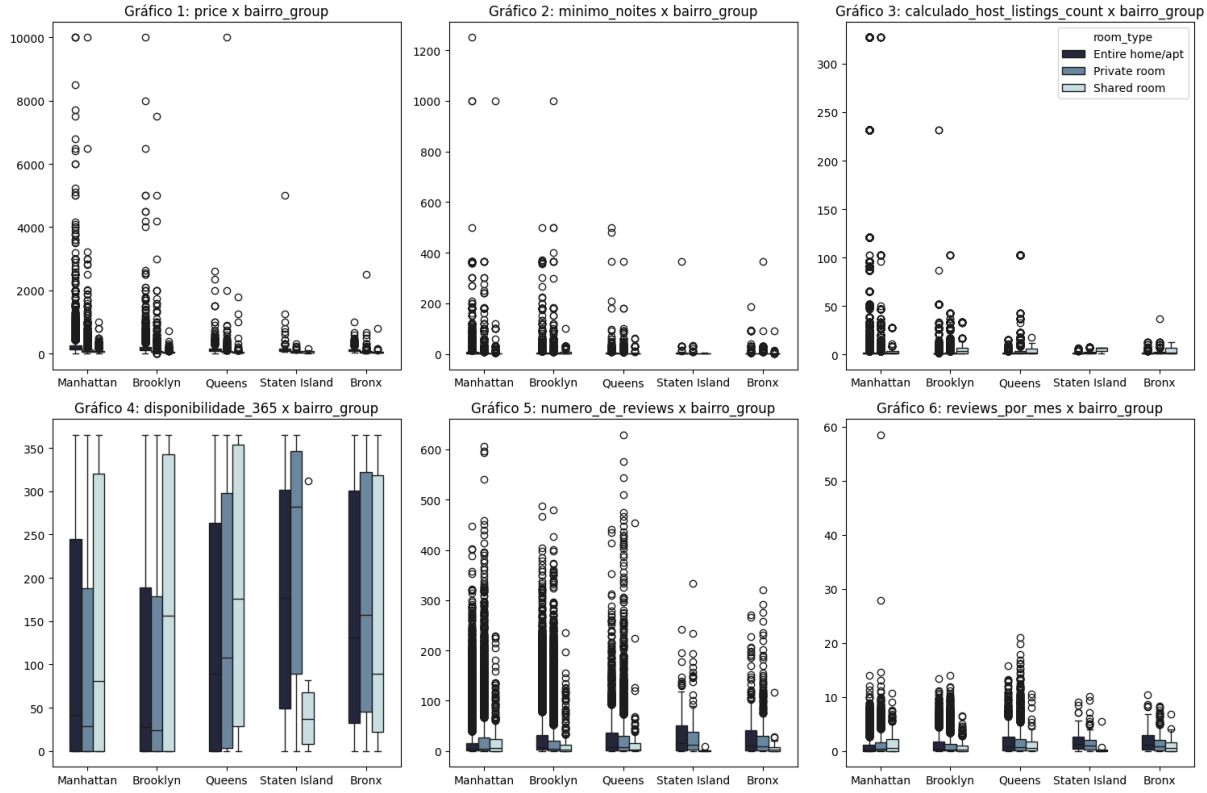
Sobre os dados vazios, os únicos que foram preenchidos foram os da coluna “reviews_por_mes” com o valor 0, para evitar algum viés nas etapas seguintes em que esta variável seria usada para identificar correlações com outras.

2.1.Correlação entre variáveis categóricas e numéricas

Feito esse tratamento inicial, o próximo passo foi observar a distribuição dos dados em relação a variáveis categóricas que indicariam o perfil de imóveis que os clientes preferem. Para isso, inicialmente foi assumido que o preço seria o fator com mais peso no momento de escolha do imóvel a ser alugado e, portanto, localização e o tipo de imóvel foram escolhidos como

parâmetros guia para se observar a distribuição dos dados devido a tradicionalmente serem fatores que influenciam diretamente no preço. Dito isso, a imagem a seguir contém gráficos do tipo boxplot para observar a distribuição das variáveis numéricas em relação aos bairros que os anúncios estão localizados e segmentados pelo tipo de imóvel.

Imagem 1: Grade de gráficos para se observar a distribuição de dados em relação a localização e segmentado por tipo de imóvel



Fonte: Autor

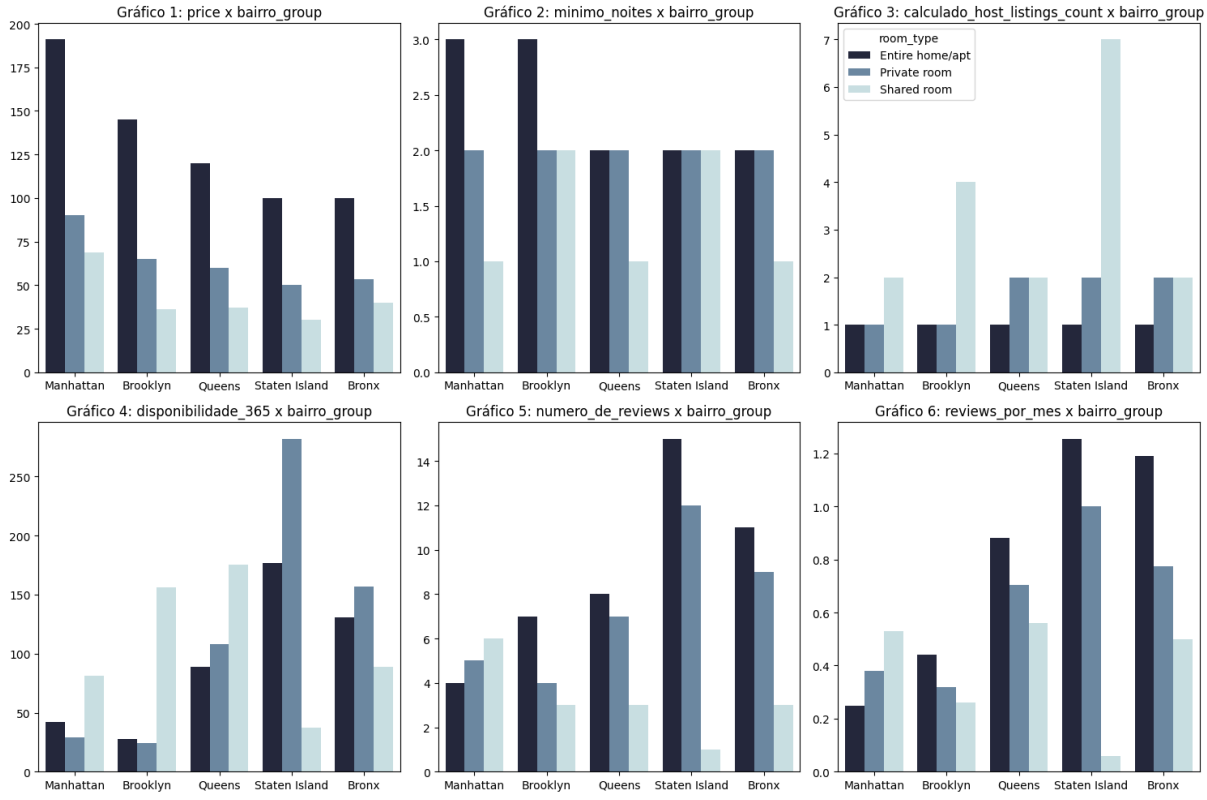
Gráficos do tipo boxplot são muito úteis pois permitem visualizar a distribuição dos dados junto com alguns parâmetros de referência como os quartis e a mediana (2º quartil) além de também ser possível visualizar a presença ou não de outliers, que seriam aquelas entradas que estariam além dos limites superiores e inferiores de cada elemento do gráfico. Estes limites são determinados geralmente como sendo 1,5 vezes o 1º e 3º quartis. Dito isso, da plotagem acima é nítida a presença de uma grande quantidade de outliers em praticamente todas as variáveis com exceção de “disponibilidade_365” junto de uma grande variação nos valores dos dados, o que impede uma boa visualização das informações.

Não foi possível visualizar de forma satisfatória a distribuição dos dados utilizando boxplot, entretanto alguns pontos importantes devem ser considerados:

- Estimativas de localização como média são sensíveis aos outliers, portanto não são aconselháveis de se usar nesse caso sendo mais interessante estimativas de localização como mediana, que não são afetadas pela presença de outliers.
- Em casos de erros de medição, por exemplo, que acabam gerando outliers é possível e aconselhável removê-los do conjunto de dados por não representarem fielmente aquele contexto, porém neste caso eles não serão removidos por serem informações representativas e que podem gerar insights valiosos para a estratégia de negócios da empresa.

Continuando a análise, a etapa seguinte foi destinada a tentar determinar o perfil de imóvel com maior preferência pelos clientes mantendo o cruzamento de variáveis realizado no boxplot, mas agora utilizando gráficos de barra cujas alturas seriam as medianas das variáveis.

Imagem 2: Grade de gráficos para se comparar o perfil dos imóveis alugados nos bairros em relação a diferentes variáveis numéricas



Fonte: Autor

A partir dos gráficos acima, é possível observar diversas coisas interessantes. Primeiramente em relação aos preços dos aluguéis, é notório que localização e o tipo de imóvel são determinantes na composição do preço, visto que há uma grande variação entre eles. Percebe-se que Manhattan possui imóveis mais caros e, portanto, deve ser uma área nobre de NY. Também é observado um padrão que se observa em todas as outras localidades, em que o tipo “Entire home/apt” é o mais caro seguido de “Private room” e “Shared room”. Ainda neste gráfico é importante destacar que o tipo de imóvel mais barato são as salas compartilhadas presentes em Staten Island.

No gráfico 2 temos a relação de localização e tipo de imóvel com o número mínimo de noites que o usuário deve reservar. Entre as salas privadas, o número de noites se mantém o mesmo em todos os bairros e para os outros tipos, é observado variações com destaque para as casas/apartamentos inteiros que em Manhattan e Brooklyn possuem a maior exigência por parte dos anfitriões. Isso pode estar relacionado com a hipótese levantada no gráfico 1, em que as regiões mais caras seriam mais nobres, logo pode haver custos maiores para manter o local em bom estado como impostos, contas, manutenção etc. Outra hipótese pode ser um modelo de negócio que explora um provável poder aquisitivo maior por parte dos clientes que alugam nesta região.

O gráfico 3 mostra que entre os proprietários de salas compartilhadas, os de Staten Island são os que possuem maior quantidade destes anúncios, que poderia indicar uma demanda acentuada por estes modelos devido a quantidade de anúncios e consequentemente uma concorrência acirrada nesta modalidade, entretanto ao agrupar os dados por tipo de imóvel e bairro, as salas compartilhadas nesta região são apenas 9, ou seja, a maioria dos anúncios se concentram na mão de uma única pessoa, podendo talvez ser uma oportunidade.

Seguindo para o gráfico 4, nele podemos identificar a demanda por cada tipo de imóvel em cada uma das regiões, basta olhar para os que possuem menos dias disponíveis ao longo do ano. Além disso, percebe-se que as salas compartilhadas, tipo mais barato de imóvel, são os menos procurados nas regiões mais caras enquanto que as casas/apartamentos inteiros (tipo mais caro) são os mais alugados, porém o contrário é observado nos bairros com imóveis mais baratos, logo é provável que o perfil de clientes varie em cada bairro em função de seu poder aquisitivo, reforçando a hipótese elaborada no gráfico 2 em que os imóveis mais caros dos

bairros mais caros possuem um número mínimo de noites maior porque os clientes podem pagar.

Por fim nos gráficos 5 e 6 temos as relações entre o total de reviews e reviews mensais por bairro/tipo de imóvel respectivamente. Nota-se que comparado com a quantidade de anúncios, a quantidade de reviews é bem baixa, ou seja, a maior parte das pessoas que utilizam dos serviços da concorrente não fazem uma avaliação, reforçando ainda mais a tese de que o preço é um dos principais fatores que justificam para o cliente alugar determinado imóvel, uma vez que com a ausência de avaliações, restam poucos recursos para julgar a qualidade do serviço.

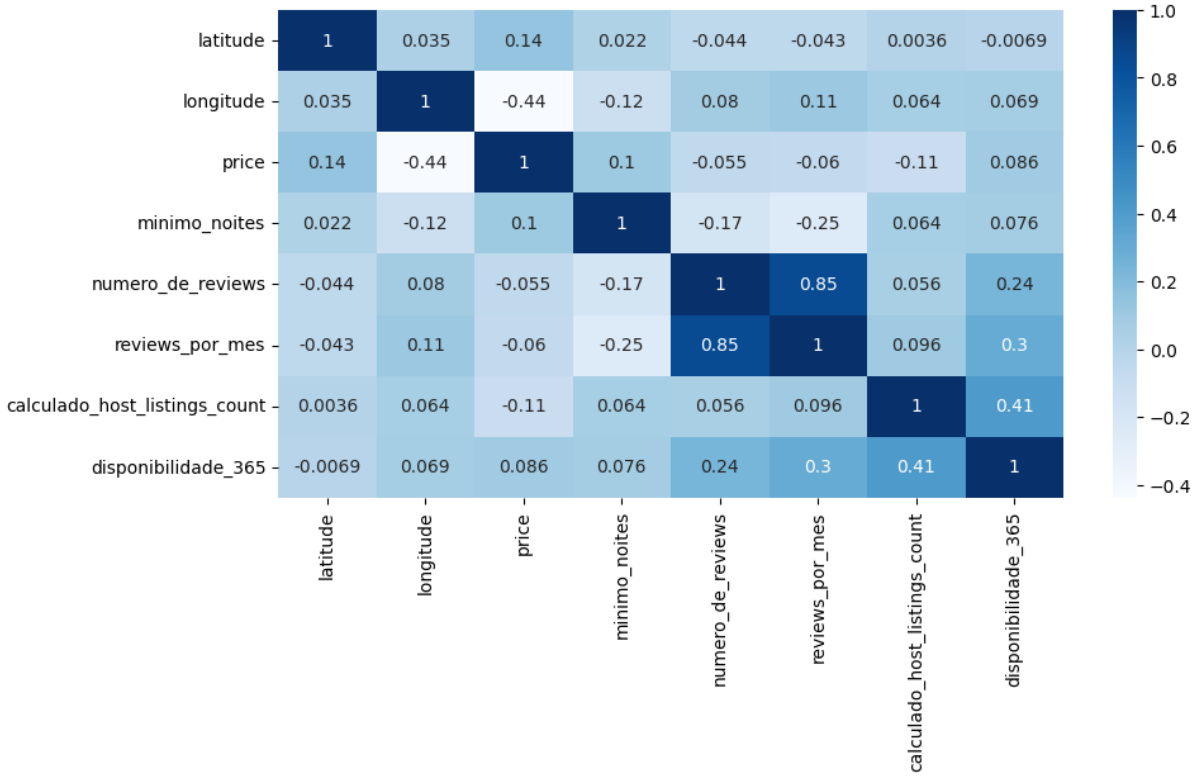
2.2.Correlação de variáveis numéricas

Com as análises anteriores foi possível perceber que o preço dos aluguéis é fundamental na escolha do imóvel que será alugado por parte dos clientes, dessa forma este é um ponto que será explorado a seguir. Para isso será verificada a correlação do preço com as variáveis da base de dados de modo que será possível quantificar o quanto elas estariam correlacionadas, com exceção de id e host_id por serem apenas valores numéricos de identificação dos anúncios e não representarem alguma característica do imóvel.

O método utilizado foi a visualização de uma matriz com os coeficientes de correlação de Spearman, que é um método estatístico mata medir a correlação entre variáveis cuja distribuição não é normal. Apenas com os boxplots é possível identificar que os dados não possuem uma distribuição normal pela grande presença de outliers, que fazem com que a média seja muito distante da mediana.

A interpretação dos resultados é relativamente simples: os valores variam de -1 a 1 e quando mais próximos de 1, mais forte é a correlação entre as variáveis, ou seja, se uma aumenta a outra também. O contrário é correto para valores negativos próximos de -1. Entretanto, quanto mais próximo de 0 menor é a correlação entre as variáveis. Dito isso, a imagem a seguir mostra uma matriz cujos valores são os coeficientes de Spearman:

Imagem 3: Matriz de correlação de coeficientes de Spearman



Fonte: Autor

A matriz acima nos permite validar e complementar alguns pontos da análise anterior. Primeiramente em relação aos preços e as características do anúncio, os coeficientes mais distantes de zero são aqueles relacionados a latitude e longitude, ou seja, indicando que a localização dos imóveis é o fator que mais impacta no preço do aluguel, seguido pelo número mínimo de noites que devem ser reservadas que tornariam o imóvel mais caro. Além disso observa-se uma correlação negativa entre os preços e a quantidade de anúncios feitos por hosts, que nos diz que os que possuem mais imóveis são do tipo mais barato, indicando uma maior concorrência neste nicho.

Em relação ao número de reviews mensal e total, observa-se uma correlação positiva dessas variáveis com a disponibilidade do imóvel ao longo do ano, indicando que grande parte dos reviews feitos são negativos visto que os que possuem a maior quantidade são os que são menos alugados e que geralmente são os hosts que possuem mais anúncios publicados. O restante dos coeficientes não apresenta correlações suficientemente grandes para indicar alguma tendência e, portanto, esta etapa termina.

Concluindo, o aluguel temporário de imóveis em Nova York é altamente impactado pela localização que eles se encontram e o tipo a que pertencem pois isso impacta diretamente no preço, que seria o principal critério usado pelos clientes. Também é razoável imaginar que o perfil dos clientes varia conforme a região e, portanto, o que seria um bom investimento pode variar também, baseado nos dados de disponibilidade e quantidade de anúncios feitos.

2.3.Respostas

Depois desta investigação realizada nos dados é possível responder às perguntas feitas.

Pergunta 2.a:

Como recomendação de investimento, eu diria a pessoa que a melhor opção seria comprar um imóvel do tipo “Entire home/apt” em Manhattan por serem imóveis que possuem a mediana de preço e o número mínimo de noites necessárias para o aluguel mais altos, o que tornaria o faturamento bem alto e possivelmente um retorno pelo investimento mais rápido que todas as outras opções, pois além destes fatores este tipo de imóvel também possui a maior demanda entre todos os outros empatando com as salas compartilhadas de Staten Island, que apesar de praticamente não terem muita concorrência, o retorno seria menor.

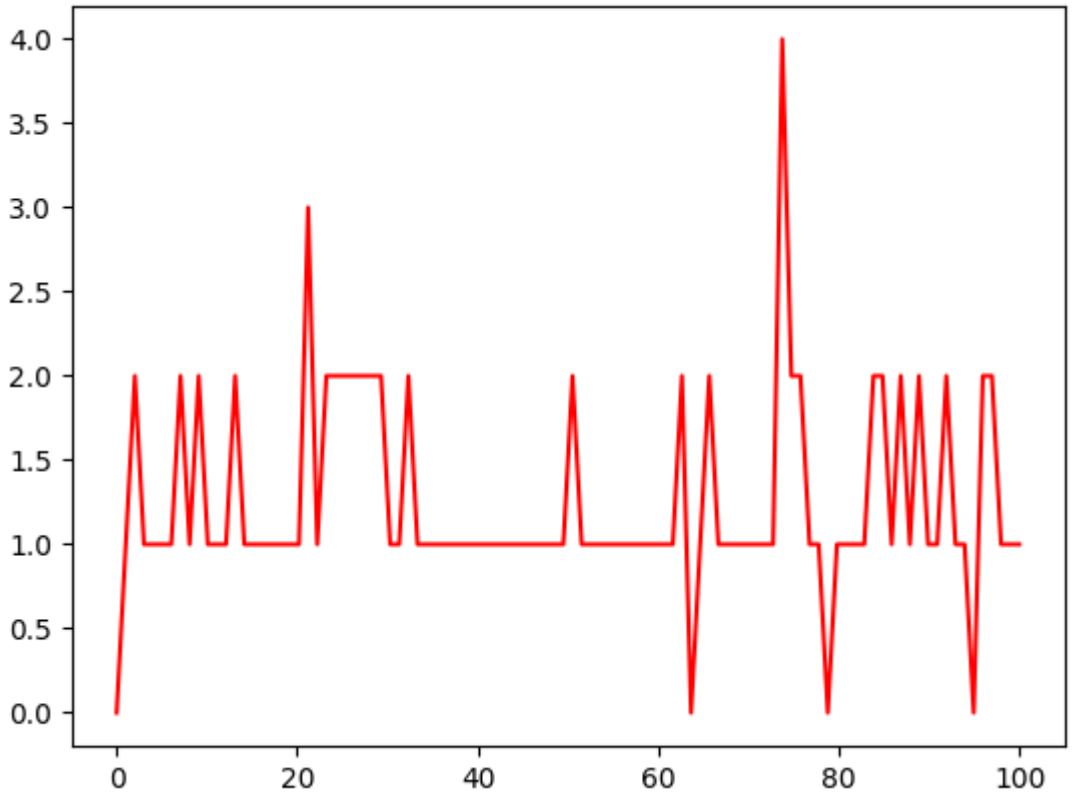
Pergunta 2.b:

A partir da matriz de correlação, observa-se uma correlação positiva entre o preço e número mínimo de noites, o que faz sentido visto que quanto mais noites são necessárias para o cliente alugar, mais caro será a reserva. Em relação a disponibilidade ao longo do ano, a correlação é praticamente nula, indicando que não há uma correlação direta entre estas duas variáveis.

Pergunta 2.c:

Os textos dos nomes dos bairros com locais mais caros não me pareceram ter um padrão específico, porém tentei encodar os valores e plotar um gráfico para tentar visualizar algo.

Imagem 4: Padrão no nome dos bairros



Fonte: Autor

Percebe-se um padrão oscilatório de valores quando se olha os 100 imóveis mais caros. Aumentando a quantidade de gráficos não ajuda muito, apenas atrapalha a visualização.

3. PREVISÃO DE PRECOS

Para fazer a previsão de preços utilizando os dados, é necessário fazer uma regressão, que seria basicamente identificar uma curva (função) que se aproxima o máximo possível de todos os pontos presentes na base de dados, no caso os preços. Para este caso, a regressão usada será linear utilizando múltiplas variáveis (features) cujas métricas de avaliação usadas foram o coeficiente de determinação e o erro mínimo quadrado.

O coeficiente de determinação ou R^2 indica o quão bom seu modelo é em relação a utilizar a média dos dados, ou seja, caso a previsão fosse feita utilizando simplesmente a média dos preços o quão melhor seu modelo consegue prever dados em relação a este método. O erro mínimo quadrado é a soma do quadrado da diferença entre o valor previsto e o valor real, de modo que o erro sempre se acumule e não seja cancelado com previsões que são maiores ou menos que o valor real.

Para fazer a regressão, todas as features com exceção de latitude, longitude, id, host_id, host_name, bairro, ultima_review e nome foram utilizadas, entretando as variáveis categóricas precisaram ser tratadas. O método utilizado foi OneHotEncoding que cria uma relação binária entre as categorias daquela coluna específica, de modo que 0 aquela entrada não pertence a categoria e 1 pertence. Além disso, as variáveis numéricas precisaram ser normalizadas, visto que havia uma grande variação na escala dos valores, dessa forma todas as features tiveram seus valores transformados para uma faixa de 0 a 1.

Dito isso, a previsão de novos dados de entrada é feita utilizando os mesmos tratamentos de dados na nova entrada e utilizando o método “predict”. Para o novo anúncio sugerido, o preço previsto foi de $ 289,00.