Estudo de inferência causal sobre a maximização preditiva do índice de satisfação dos usuários no Airbnb

Rian Freitas da Silva[a]

[a]*Escola de Matemática Aplicada, Fundação Getúlio Vargas,*

Abstract

Este artigo é resultado de uma modelagem de experiências de hóspedes em acomodações reservadas pela plataforma *Airbnb* em 10 cidades europeias. O objetivo é entender de que forma a correlação entre os preços ofertados e o índice de limpeza em que a acomodação se encontrava ao hóspede para sua avaliação final. A partir disso, utilizando uma função a ser otimizada, propor que valores para esses condicionantes são importantes para uma expectativa de satisfação máxima do usuário. Os resultados demonstram um aumento, em média, de 4,82% na satisfação dos usuários após a aplicação desse valores ótimos, para um conjunto de dados de 51708 pontos.

Keywords: airbnb, modelagem, otimização, BFGS, Métodos Quasi-Newton

1. Introdução

O *Airbnb*, um conhecido aplicativo de hospedagem compartilhada em que anfitriões cedem imóveis ou quartos em suas casas para acomodação de usuários, é um exemplo muito forte do movimento de economia compartilhada. Essa nova forma de transação, possibilitada principalmente pela Internet, se caracteriza ”incentivando os consumidores a trocarem entre si o que já possuem e, consequentemente, reduzindo a produção e comercialização em larga escala dos produtos mais variáveis”(COHEN, Lucas; 2021).

Apesar dessa nova tendência, é possível entender o impacto negativo de sua ampliação a partir de algumas perspectivas. Seja a concorrência desleal e *dumping* dessas platarformas com a economia hospitaleira local, seja a precarização da classe trabalhadora que a terceirização causa pelo estabelecimento do chamado trabalho por demanda (CARDOSO, Ana; DE OLIVEIRA, Marcela; 2022). Contudo, este trabalho se origina a partir de uma outra problemática: a falta de instrução dos anfitriões no que tange à precificação. Em se tratando de um sistema de autoveiculação, os usuários não possuem qualquer documentação ou suporte. Existem as taxas próprias do *Airbnb*, que totalizam em média entre 14 e 16% do valor da acomodação por noite, mas custo da reserva é determindado pelo dono no imóvel.

1.1. Base de dados escolhida

Disponível no site *Kaggle*, a base de dados que será usada neste trabalho é um conjunto de *datasets* isolados do preço de acomodações em algumas cidades europeias. Para essa análise, unifiquei todos os dados em uma única fonte utilizando o editor do *Power Query*, da *Microsoft*, totalizando 51.708 linhas.

A base contém colunas muitas colunas importantes para esse estudo de caso: informações sobre localização (proximidade do metrô, do centro da cidade), sobre a acomodação (número de quartos, se a oferta é um quarto ou toda a casa), o preço da diária e o índice de satisfação do cliente, a variável mais importante para o artigo.

Uma primeira premissa que é fundamental para continuidade da argumentação é a existência de dois tipos de covariável: de caráter **fixo**; isto é, representam informações sobre a oferta que não podem ser modificadas pelo anfitrião; e de caráter mutável, que representam suas escolhas. Infelizmente, uma limitação dessa base de dados é a carência de covariáveis mutáveis. Só foi possível considerar duas: o preço e o índice de limpeza do ambiente.

1.2. Objetivo

Com a análise causal do índice de satisfação dos usuários, pode-se extrair que ações devem ser tomadas de modo a maximizar essa predição. Portanto, traduz-se em um preço e um grau de limpeza que podem ser chaves para uma boa experiência do hóspede, dado as condições imutáveis.

É evidente que a modelagem responsável por essa predição não pode permitir valores extremos ou irreais. Não se pode permitir a conclusão de que um preço zero implica alta satisfação do cliente. Aliás, a análise exploratória que será disponibilizada na próxima seção vai de encontro a esse senso comum.

Por conseguinte, prioriza-se a experiência do hóspede, que é o desfecho desse estudo, pois garantindo essa boa devolutiva o anfitrião possui uma série de benefícios. Ele mantém uma boa pontuação na plataforma, podendo atingir o *status* de *SuperHost*, suas ofertas são melhores colocadas no algoritmo de busca e é possível fidelizar seus clientes de modo a manter seus imóveis sempre sob demanda alta.

2. Metodologia

Para prover uma maximização preditiva da satisfação dos usuários, será necessário combinar uma modelagem estatística

que se adéque aos dados disponibilizados e, a partir dessa escolha, resolver um problema de otimização, cuja função objetivo; ou seja, a função que se deseja maximizar nessa análise, será advinda do modelo.

A seguir, será mostrado os passos a serem realizados a fim de se chegar à resposta desejada.

2.1. Subconjuntos de treino e validação

Para que se possa avaliar qual modelo deverá ser escolhido, o *dataset* será repartido em subconjuntos de treino e validação, de tamanho equivalente a 70% e 30% do conjunto total, respectivamente. Assim, poderemos ver a capacidade preditiva de cada opção de modelo, o que será um critério chave para a escolha do melhor.

2.2. Escolha de um modelo adequado

A partir da fórmula geral do modelo linear generalizado, dado por

$$\mathbb{E}[y] = f(X^T\beta), \tag{1}$$

cuja função $f(\cdot)$ poderá ou não ser uma função identidade. Serão testados na parte de ajuste alguns modelos para diferentes famílias de $f(\cdot)$. Uma vez escolhida a família de melhor resultado, serão testadas algumas combinações de covariáveis, sobretudo removendo as que tiverem um intervalo de confiança que contenha zero. A partir dessas duas etapas, será decidido o melhor modelo.

O critério de escolha será uma junção de dois fatores: a capacidade preditiva do modelos e a bonança do ajuste. Será calculada a média de erro quadrático no conjunto de validação, dada por:

$$\frac{1}{N}\sum_{n=1}^{N}(y_{pred} - y)^2, \tag{2}$$

em que y_{pred} é a saída do modelos e y é o índice de satisfação real. Além disso, será usado o *AIC* (Akaike Information Criterion), que "mede a qualidade relativa de um modelo estatístico para um determinado conjunto de dados"(KANDEKAR, Supriya; SMITH, Emma; 2023). No entanto, como o *AIC* penaliza em função da complexidade do modelo (KANDEKAR, Supriya; SMITH, Emma; 2023), a prioridade será o erro de validação; ou seja, será priorizada capacidade de predição à simplicidade do modelo. Como consequência disso, o método de ajuste será por máxima verossimilhança.

Ademais, outros métodos foram usados na avaliação, como a distribuição dos resíduos dos modelos e da densidade da predição realizada por ele. Esses dois métodos foram chave para a escolha.

2.3. Delimitação do espaço de busca

A partir daqui, é importante definir algumas variáveis de interesse. Sabendo que

$$X^T\beta = \beta_0 + x_1\beta_1 + ... + x_{n-1}\beta_{n-1} + x_n\beta_n, \tag{3}$$

pode-se chamar x_{n-1} e x_n como o preço e o índice de limpeza da hospedagem, respectivamente. Como definido anteriormente, essas são as variáveis mutáveis, que terão um espaço de busca para se encontrar seu valor que maximiza a satisfação do cliente. Esses termos são chamados de **interventores**. Como temos uma função $f(X^T\beta) = \mathbb{E}[y]$ e dados $x_1, x_2, ..., x_{n-2}$, precisa-se buscar x_{n-1} e x_n que maximizam $\mathbb{E}[y]$.

Contudo, não se pode permitir que essa busca não tenha restrições. Por exemplo, é senso comum considerar que uma acomodação com valor de 0 reais por noite e índice de limpeza 10,0 terá uma avaliação 100,0. Nessa perspectiva, nosso problema de otimização deverá ter restrições.

Um meio escolhido para tal foi impor que x_{n-1} terá como limites $q_{10}(x_{n-1})$ e $q_{90}(x_{n-1})$. Da mesma forma, x_n terá $q_{10}(x_n)$ e $q_{90}(x_{n-1})$. Considere $q_{10}(x)$ o quantil que separa os 10% dos valores ordenados de x. Dessa mesma forma, o q_{90} determina os 90% dos dados ordenados. A escolha inicial seria a função *max* ou *min*, porém constataram-se *outliers* no conjunto; isto é, valores muito destoantes da distribuição. Portanto, usar medidas que tenham mais estabilidade é a melhor decisão.

Deve-se alertar que essa decisão acarreta um viés considerável para o cálculo da predição, pois já se pressupõe que $q_{10}(x) < x_{pred} < q_{90}(x)$. Porém, é a melhor escolha de estimativa, já que se apoia em um valor real e observado de preço e índice de limpeza. Outras estimativas minimamente embasadas requereriam um estudo econômico muito aprofundado.

2.4. Problema de otimização restrita

Como vimos na equação 1, temos uma função $f(\cdot)$ que recebe a matriz de covariáveis X. Podemos abrir como fizemos na 3, de modo que:

$$\mathbb{E}[y] = f(\beta_0 + x_1\beta_1 + ... + x_{n-1}\beta_{n-1} + x_n\beta_n). \tag{4}$$

Ainda considerando os interventores x_{n-1} e x_n, Podemos, então, fazer uma decomposição da matriz de desenho X de modo que:

$$\mathbb{E}[y] = f\left((X')^T\beta' + x_{n-1}\beta_{n-1} + x_n\beta_n\right), \tag{5}$$

em que X' é a matriz de desenho cujas colunas são $x_1, ..., x_{n-2}$ e $\beta' = (\beta_0, \beta_1, ..., \beta_{n-2})$.

Uma vez tendo o modelo ótimo em mãos, para cada linha do *dataset*, queremos calcular x_{n-1} e x_n que maximizem y_{pred}. Ou seja, dada uma acomodação, $(X')^T\beta'$ é uma constante, que será chamada c.

Logo, para uma hospedagem, temos o seguinte problema de otimização:

$$\underset{x_n, x_{n-1} \in \Theta}{argmax}\ f(c + x_{n-1}\beta_{n-1} + x_n\beta_n), \tag{6}$$

para $\Theta = (min(X_n), max(X_n)) \times (min(X_{n-1}), max(X_{n-1}))$.

Para esta implementação, eu usarei um método Quasi-Newton de otimização disponível na linguagem de programação R, chamado "L-BFGS-B". O algoritmo "BFGS"foi desenvolvido para problemas de otimização em funções objetivo não lineares ou não unimodais — em que não é um único ponto crítico. Esse método usado na análise é uma expansão desse algoritmo, especializado em problemas com restrição de espaço de busca, o que é esse caso.

3. Resultados

Dada toda a metodologia detalhada acima, ela será implementada no conjunto de treino. Para ajudar na escolhar do modelo ideal, algumas visualizações serão mostradas de modo a provar as tendências perceptíveis nos dados.

3.1. Análise exploratória de dados

Primeiramente, é importante dar destaque à distribuição dos valores de índice de satisfação.

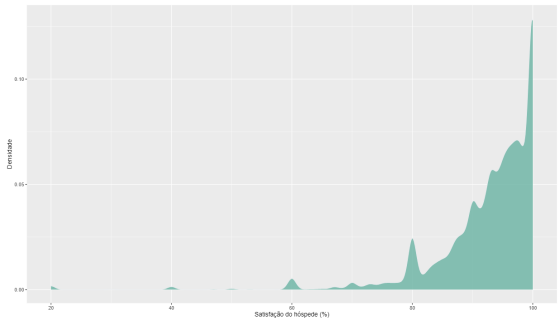

Figura 1: Gráfico da distribuição de densidade do índice de satisfação do cliente.

Percebe-se de primeira que o dado é assimétrico; ou seja, temos muito mais informações sobre acomodações com uma boa nota dos hóspedes em dentrimento das com avaliações ruins. Contudo, como nosso problema final é uma maximização desse índice, essa consequência não compromete a análise. Ainda assim é fundamental levar essa assimetria em consideração na análise dos próximos gráficos.

3.1.1. Matriz de correlação

covariável	satisfação do cliente
preço	-0,0019
limpeza	0,714
dist. ao centro	-0,0042
dist. ao metrô	-0,0298

Tabela 1: índice de correlação entre as covariáveis numéricas e a variável resposta. Alguns nomes foram simplificados.

Outro fator preocupante é acerca da correleção dos dados. À exceção do índice de limpeza da hospedagem, todas as outras covariáveis possui uma correlação próxima à zero, o que inviabiliza uma modelagem usando as variáveis originais, sem nenhuma transformação. A parte de adição de interações de transformações nos dados será tratada mais à frente.

-	preço	limpeza	dist. ao centro	dist. ao metrô
preço	1	-0,0061	-0,0447	-0,061
limpeza	-0,0061	1	-0,0302	0,0104
dist. ao centro	-0,0447	-0,0302	1	0,5581
dist. ao metrô	-0,061	0,0104	0,5581	1

Tabela 2: índice de correlação entre as covariáveis numéricas e a variável resposta. Alguns nomes foram simplificados.

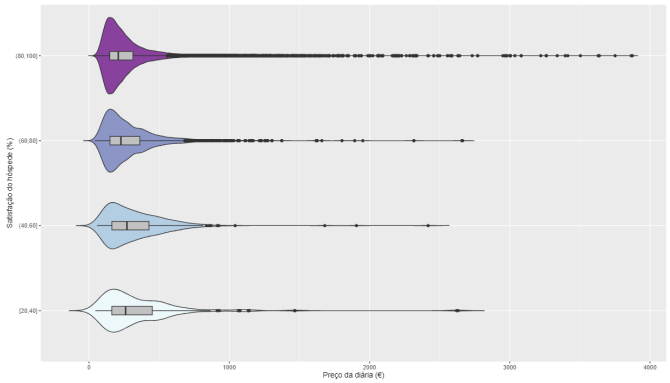

Figura 2: Gráfico da distribuição de preço a partir de diferentes níveis de satisfação do cliente.

Entretanto, uma constatação positiva é a matriz de correlação das covariáveis numéricas entre si. Não há um valor muito alto na tabela 2; ou seja, não uma correlação notável. A única exceção é a interação entre a distância ao centro da cidade e a distância ao metrô. Faz bastante sentido esse valor, uma vez que, em modo geral, só há estações de metrô nos centros urbanos. É um *dataset* de cidades europeias, onde a malha ferroviária é o principal meio de transporte que atende às regiões afastadas (NASH, Chris; 2005).

3.1.2. Distribuição dos interventores

Para fundamentar a relação causal dos interventores na predição do modelo, é ideal entender de que forma essas covariáveis ficam distribuídas ao longo de diferentes valores de índice de satisfação. Uma vez tendo como premissa essa causalidade, é necessário levantar que valores estão ligados a uma satisfação dos clientes (e a insatisfações também).

O gráfico em 2 mostra a distribuição do valor da diária estratificada por níveis de avaliação. Em um primeiro momento, é possível concluir que o preço se concentra cada vez mais à medida em que se eleva o índice; ou seja, avaliações ótimas (entre 80 e 100%) e boas (entre 60 e 80%) têm uma concentração

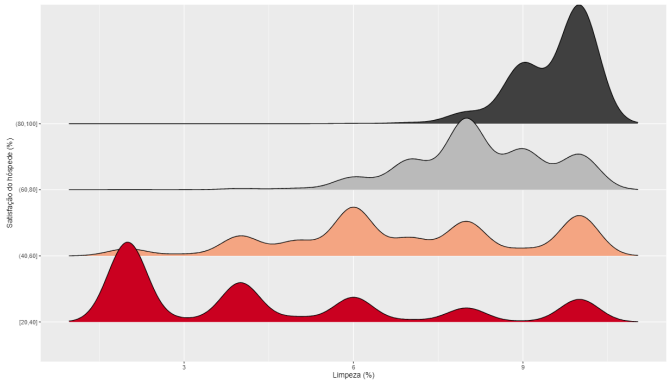

Figura 3: Gráfico da distribuição de índice de limpeza a partir de diferentes níveis de satisfação do cliente.

maior em preços mais baixos do que em avaliações medianas (40 e 60%) e ruins (20 e 40%, pois o que tudo indica é que a nota mínima é 20

Entretanto, deve-se relembrar da conclusão do gráfico 1. Há muito mais registros de boa avaliação do que ruins, o que influenciou nesse gráfico. Além disso, o valor de correlação na tabela 1 conclui que esse aumento de concentração de preço não indica uma nova descoberta. Na realidade, a média dos *boxplots* na figura corrobora a hipótese de que preço e satisfação do cliente não tem uma conexão forte.

Diferentemente do que se pode aferir acerca da porcetagem de limpeza. A figura 3 dá uma evidência forte de correlação, do quanto uma hospedagem bem limpa resulta em uma avaliação positiva.

3.1.3. Distribuição de variáveis categóricas

O termo *Superhost* é uma classificação de anfitriões dentro do *Airbnb*, para aqueles que têm um histórico considerável na plataforma, hospedando clientes e recebendo bons *feedbacks*. Dessa forma, ofertas cujo criador é um *Superhost* possuem este selo quando um usuário as encontra, além de terem prioridade nos resultados de pesquisa e outros incentivos. Portanto, é fundamental analisar o impacto dessa classificação no índice de satisfação dos clientes. A figura 4 demonstra como hóspedes instalados em acomodações cujo proprietário é *Superhost* tende a ter uma boa experiência muito mais, em dentrimento das cujo dono não é, que têm uma distribuição com uma variância maior.

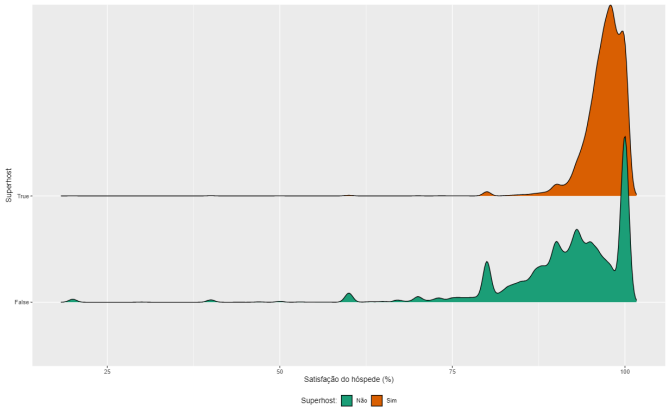

Figura 4: Gráfico da distribuição de do índice de satisfação dependendo se o anfitrião é *Superhost.*

No que tange à cidade da acomodação, o *boxplot* da figura 5 demonstra uma variação do índice em diferentes cidades. Ou seja, há sim um efeito computável. Contudo, em se tratando da modelagem do problema, esse efeito são se mostra substancialmente forte ao ponto de ser necessário um modelo hierárquico, que dê um peso maior à origem da acomodação.

Por fim, em relação ao tipo de acomodação, a figura 5 somente há uma diferença entre se a oferta é de um quarto compartilhado ou não. O caso afirmativo tem uma variedade de avaliações maior do que o negativo.

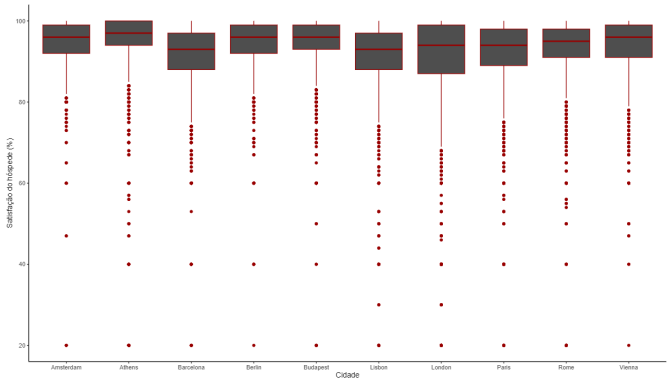

Figura 5: *Boxplot* representando a distribuição do índice de satisfação em diferentes cidades da Europa.

3.2. Escolha da melhor família

A seguir, a tabela 3 mostra os valores de AIC e MSE para cada família e especificação de família escolidas previamente. Como dito anteriormente, estamos trabalhando com a versão expandida do modelo linear, que possui uma função de ativação $f(\cdot)$, chamada `link` na função `GLM` do R.

A primeira coisa a se notar é que os dados não parecem se adequar bem a modelos da forma $f(X^T\beta)$ para $f(\cdot)$ diferente da função identidade.

Em um momento inicial, acreditei que uma função log seria adequada ao modelo, devido à assimetria. Contudo, o que constatei nesse modelo é que a proporção se manteve em comparação a um modelo com função identidade, apenas a escala foi modificada, tendo um espaço de predição muito reduzido. Isso explica o tamanho do MSE para eles.

Embora a decisão mais correta seria escolher a distribuição gaussiana para Y, de acordo com o melhor AIC e MSE, essa decisão será melhor baseada a partir do gráfico 6. Essa comparação se deu porque, inicialmente, conhecendo teoricamente esse modelos, eu tinha uma preconcepção de que a distribuição gaussiana inversa seria melhor. A razão para acreditar nessa hipótese é que, diferente da normal comum, ela não

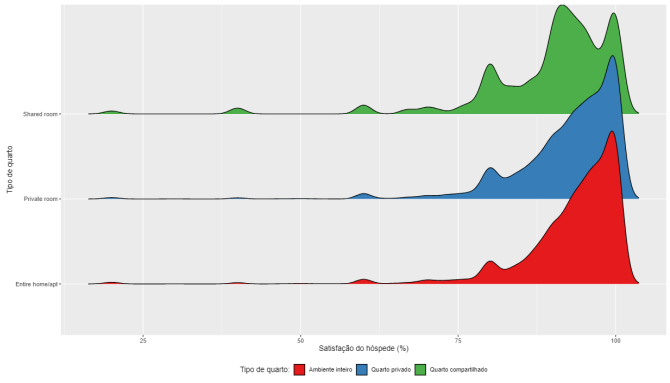

Figura 6: Distribuição do índice de satisfação do usuário para cada tipo de acomodação.

Modelo	AIC	MSE
`gaussian(link = "identity")`	234083	38,4189
`gaussian(link = "log")`	235510	7840.158
`gaussian(link = "inverse")`	236779	8657.133
`Gamma(link = "identity")`	250005	38.76727
`Gamma(link = "log")`	252641	7840.124
`Gamma(link = "inverse")`	254505	8657.132
`poisson(link = "identity")`	247911	38.42494
`poisson(link = "log")`	248885	7840.167
`poisson(link = "inverse")`	249705	8657.133
`inverse.gaussian(link = "identity")`	264737	39.83357

Tabela 3: Valores de AIC de MSE para cada modelo em avaliação.

adimite valores negativos, o que dá um conhecimento prévio à modelagem que é condizente com a realidade.

Medidas de erro como o MSE não são muito claros acerca de que regiões a predição está sendo subestimada ou superestimada, então analisar graficamente esses dois modelos é a melhor opção. A partir do gráfico de dispersão dos resíduos, pode-se perceber que a predição no modelo gaussiano tem um resíduo maior do que o gaussiano inverso nas áreas de satisfação baixas e menor em avaliações boas. A ligeira diferente no MSE dos dois modelos se dá, muito provavelmente, ao fato de se ter mais avaliações positivas, como discutido anteriormente. Logo, a escolha, no fim, deve ser baseada em qual seria o objetivo dessa modelagem.

O fim último desse estudo é maximizar essa predição. Logo, o ideal será o modelo que acerte suas predições em valores altos de satisfação. Nessa perspectiva, será escolhido o modelo gaussiano, representando uma regressão linear simples.

3.3. Escolha do melhor conjunto de covariáveis

Nesse parte da análise, serão aplicadas todas as conclusões que foram identificadas na seção 3.1.

Em relação às covariáveis numéricas, temos um caso de modificação. A tabela 1 mostra que a distância ao metrô e ao centro da cidade. Testando algumas combinações de interações; isto é, termos que relacionam e unem duas ou mais covariáveis,

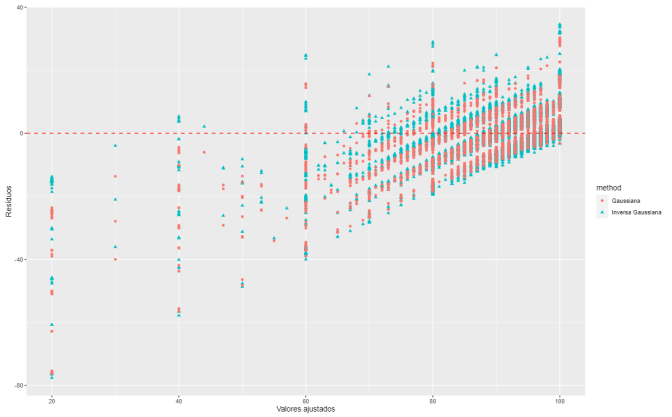

Figura 7: Gráfico de dispersão dos resíduos para o modelo gaussiano e gaussiano inverso.

houve a constatação de que, sendo A = dist. ao metrô e B = dist. ao centro, o tempo $-\frac{A}{B}$ aumentou ligeiramente a correlação com a variável resposta, com um valor de -0.03078343.

Embora o preço tenha pouca correlação com a satisfação dos clientes, essa covariável será mantida para a aplicabilidade da análise, já que ela é um interventor a ser otimizado.

Acerca das covariáveis discretas, será feita uma modificação na coluna de tipo de quarto. Agora, ao invés de 3 tipos, haverá apenas dois, avaliando se é quarto compartilhado ou não.

Além disso, foram removidos do *dataset* final outras variáveis cuja definição não estava bem especificada ou que eram apenas índices da plataforma.

3.4. Resultados final do modelo

O modelo final tem um AIC de 234146 e um MSE de 38,55814 no conjunto de teste. A tabela 4 dá alguns resultados interessantes. No entanto, há 4 parâmetros cujo intervalo de confiança contém o número zero. Desses, não há como inferir uma interpretação, pois possuem uma expressão fraca no modelo.

Os parâmetros fazem muito sentido com a análise exploratória de dados realizada em 3.1. Os parâmetros para as cidades implicam uma redução que é bem semelhante ao *boxplot* na figura 5. Fica evidente que as cidades cuja a média de satisfação é menor em comparação às outras impleca em um efeito negativo maior na predição do modelo. Sobre o *Superhost*, assim como demonstrado anteriormente, possuir esse selo causa um aumento da predição, em média de 1,7536, no índice de satisfação do cliente.

parâmetro	estimativa	erro padrão	intervalo de confiança 96%
intercepto	33,3174	0,3725	(32,5524;34,0823)
cidadeAtena	-0,7335	0,1950	(-1,1341;-0,3330)
cidadeBarcelona	-2,0058	0,2148	(-2,4469;-1.5648)
cidadeBerlim	0,0646	0,2239	(-0,3953;0,5245)
cidadeBudapeste	0,1143	0,2025	(-0,3015;0.5302)
cidadeLisboa	-2,5601	0,1895	(-2,9494;-2,1710)
cidadeLondres	-1,4923	0,1821	(-1,8663;-1,1184)
cidadeParis	-0,6584	0,1912	(-1,0511;-0,2656)
cidadeRoma	-1,5357	0,1822	(-1,9099;-1,1616)
cidadeViena	-0,6982	0,2078	(-1,1249;-0,2715)
preço	0,0001	0.0001	(-0,00091;0,00032)
SuperhostTrue	1,7536	0.0781	(1,5930;1,9142)
limpeza	6,3800	0,0351	(6,3081;6,4520)
-metro/centro	-0,2801	0,1238	(-0,5343;-0,0258)
quartoCompTrue	0,4298	0,3843	(-0,3594;1,2190)

Tabela 4: Estimativa dos parâmetros do modelo escolhido, juntamente com seu erro padrão e intervalo de confiança. Valores aproximados em 4 casas decimais

A conclusão mais importante dos parâmetros é o efeito no índice de limpeza. É a covariável com maior peso do modelo e o protagonista da nossa análise. Isso corrobora a importância desse estudo, pois, como esse vínculo é tão forte, entender que nível de limpeza pode maximizar essa predição.

Na construção do modelo, foi testado remover *outliers*, pois poderia ser que a amplitude do preço (aproximadamente entre 100 e 18,000 euros) em detrimento da satisfação (20 a 100%) cause esse valor baixo de estimativa. Contudo, fizeram-se testes removendo valores extremos e normalizando essa coluna, o que

não gerou um resultado melhor. Além disso, como queremos usar essa covariável como interventor, a normalização alterará os resultados práticos.

3.5. Maximização preditiva

Para a tarefa a seguir, criei o seguinte código para efetuar a busca pelos parâmetros ótimos.

```
pred_max <- function (realSum, cleanliness_rating
    , city, host_is_superhost, guest_satisfaction
    _overall, metroAndCenter, isSharedRoom) {

    f <- function(realSum, cleanliness_rating) {
        return(predict(model_final, data.frame(
            realSum = realSum,
            cleanliness_rating = cleanliness_
    rating,
            city = city,
            host_is_superhost = host_is_superhost
    ,
            metroAndCenter = metroAndCenter,
            isSharedRoom = isSharedRoom
            )))
        }

    max <- optim(par = c(0, 0), fn = function(x)
    -f(x[1], x[2]), lower = c(q1_price, q1_clean)
    , upper = c(q9_price, q9_clean), method = "L-
    BFGS-B")$par

    return (max)
}
```

Listing 1: Código de otimização da predição

Precisava-se de uma função que recebesse apenas como parâmetros os interventores para que o otimizador funcionasse. Por isso, há uma composição de funções no código. O objetivo que é, para cada linha, a função `pred_max` seja chamada e retorne os interventores ótimos para essa linha.

Para, por fim, ter os resultados finais da maximização preditiva, decidiu-se por unir os conjuntos de treino e validação, ajustar o modelo a todos os dados possíveis e, desses dados, extrair uma porção aleatória de 1000 pontos para calcular os interventores ótimos. Essa decisão se dá porque o `dataset` é muito extenso e exigiria um custo computacional muito grande.

Além disso, não é necessário rodar em todos os pontos porque, após esse teste, é possível notar que basta rodar em um único ponto. Os parâmetros finais são os mesmos, como mostra na tabela 5.

interventor	valor ótimo
preço	500,88
limpeza	10,0

Tabela 5: Valores ótimos dos interventores calculados pela função `pred_max`.

Faz muito sentido existir um valor ótimo para todos os pontos, uma vez que a função $f(X) = \mathbb{E}[Y|X]$ é linear. Como os pesos são lineares, uma vez observado o ponto x, a função de predição é uma reta n-dimensional, sendo n o número de pesos.

E como se pode supor também, por ser uma reta, os parâmetros ótimos são justamente os quantis dados como restrição. É interessante notar que os dados apontam o oposto do senso comum, pois o aumento do preço acarreta no aumento da satisfação. Contudo, devido à relação fraca no modelo, não é uma conclusão forte. Sobre o índice de limpeza, o resultado era esperado, devido à estimativa de seu efeito no modelo.

Entretanto, deve-se relembrar que esse é um estudo de inferência causal. Ou seja, análise é sobre o efeito que esses parâmetros ótimos têm na predição. Assim, comparam-se os valores da predição real, calcada nas informações reais da acomodação, com essa nova predição, que reflete qual seria a satisfação do hóspede caso esses valores de preço e índice de limpeza fossem implementados.

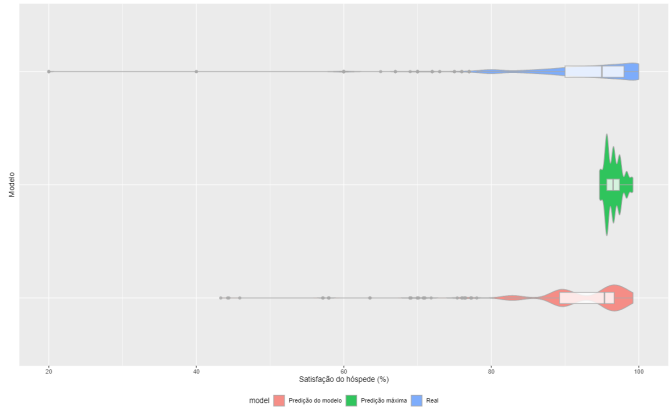

Figura 8: Gráfico de densidade para os valores reais, as predições do modelo e as predições máximas a partir dos valores ótimos.

A figura 8 mostra como as predições melhoraram no geral com os novos valores de interventores. A concentração em valores altos de satisfação confirma que o objetivo final foi atingido. Além disso, ainda há uma variância na expectativa de satisfação, que é de acordo com os outros parâmetros do modelo, representado pela constante c.

preço	limpeza	satisfação	preço*	limpeza*	satisfação*
143,0809	2,0	20	500,88	10,0	95,85139
197,0801	4,0	40	500,88	10,0	96,58164
170,2627	2,0	20	500,88	10,0	94,74812
208,9110	2,0	20	500,88	10,0	97,32909
245,7725	10,0	40	500,88	10,0	95,64437
554,1052	2,0	20	500,88	10,0	95,64421

Tabela 6: Resultados práticos dos 6 registros de pior avaliação no *dataset*. Mostra-se uma comparação entre os valores reais e a expectativa de predição a partir da alteração dos interventores.

Na tabela 6, há alguns exemplos práticos dessa otimização. Percebe-se que, à exceção do quinto registro, todos os dados presentes têm um índice de limpeza muito baixo. Assim, revela-se o impacto de boas práticas na manutenção da hospedagem e no quanto isso pode ter de retorno.

4. Considerações Finais

Esse estudo é apenas uma demonstração sobre a importância do cuidado na hospedagem e no quanto isso interfere na

avaliação do usuário. Sobre o preço, a falta de correlação entre ele e a avaliação não é forte o suficiente para se extrair uma boa conclusão.

O modelo tem problemas, como mostra o alto valor de MSE. Deve-se levar em consideração que o índice de satisfação está discretizado, não assumindo valores contínuos. Isso interfere a modelagem linear, que são ideais para variáveis contínuas. Contudo, deve-se ressaltar que um modelo categórico, que talvez seja uma opção mais viável para o problema, será complexo ao nível de causar um provável *overfitting*, já que o número de categorias seria muito alto e específico para poucos casos. Nessa perspectiva, algum método para discretizar a saída do modelo pode ser uma solução que melhoraria o desempenho. Consequentemente, deveria rever o método de otimização, pois os métodos Quasi-Newton só funcionam em funções contínuas.

Ainda assim, o estudo traz conclusões importantes que servem de orientação para anfitriões que querem melhorar a experiências de seus hóspedes e até mesmo conseguir o sela de *Superhost*.

Referências

[1] COHEN, L. M. *Economia compartilhada e precarização do trabalho: Onde estamos na discussão das mudanças no mundo do trabalho*. **Revista Avesso: Pensamento, Memória e Sociedade**, [S. l.], v. 2, n. 2, 2022. DOI: 10.23925/2675-8253.2021v2n2A1. Disponível em: https://revistas.pucsp.br/index.php/avesso/article/view/53075. Acesso em: 17 jun. 2023.

[2] CARDOSO, A. C. M.; DE OLIVEIRA, M. C. B. *Como avança a uberização no setor de turismo*.**Outras Palavas**, São Paulo. 20 jan. 2022. Disponível em: https://outraspalavras.net/trabalhoeprecariado/como-avanca-a-uberizacao-no-setor-de-turismo. Acesso em 17 jun. 2023.

[3] NASH, C. *Rail Infrastructure Charges in Europe*. **Ingenta Connect**. Journal of Transport Economics and Policy (JTEP). Vol. 39, No. 3, Set. 2005, pp. 259-278(20). Disponível em: https://www.ingentaconnect.com/content/lse/jtep/2005/00000039/00000003/art00002. Acesso em 25 jun. 2023.

[4] KANDEKAR, Supriya; SMITH, Emma. *AIC vs BIC: Diferença e Comparação*.**Ask Any Difference**. Última atualização: 11 jun. 2023. Disponível em: https://askanydifference.com/pt/difference-between-aic-and-bic/. Acesso em 24 jun. 2023.