

An

Doriedson

09-11-2024

Os dados

Trata-se de um conjunto extenso, com mais de 20 mil observações e 680 variáveis, sobre receitas culinárias do site *Epicurious*. Houve uma filtragem inicial no conjunto, restaram apenas 6 variáveis explicativas, as consideradas mais importantes: nota, a quantidade de calorias, de proteína, de gordura e de sódio. Há também o nome da receita e a variável resposta, que no caso é binária: trata-se de uma sobremesa? No site em questão, encontra-se categorias de pratos como *café da manhã*, *almoço*, *jantar*, *drinks* e, naturalmente, o objeto de interesse do trabalho, as *sobremesas*.

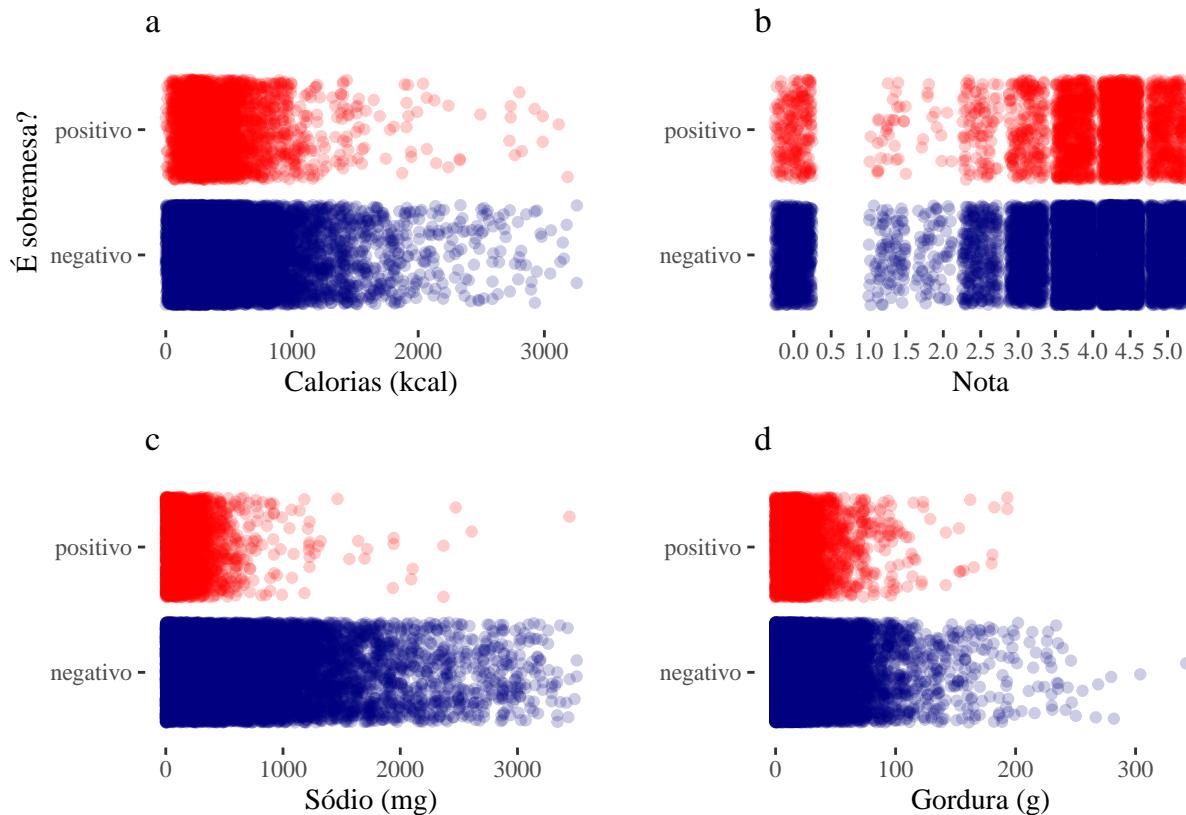
Em seguida, foi feita a exclusão de linhas que possuíam valores faltantes. Por fim, houve a retirada de valores absurdos (descrita mais detalhadamente abaixo). O conjunto final possui 15706 observações e 7 variáveis.

Obs.: a variável calorias está na unidade 'kcal'.

```
## [1] "title"          "rating"         "calories"       "protein"  
## [5] "fat"            "sodium"         "dados$dessert"  
  
##                                     titulo  nota calorias protein gordura  
## 1           Lentil, Apple, and Turkey Wrap 2.500    426     30      7  
## 2 Boudin Blanc Terrine with Red Onion Confit 4.375    403     18     23  
## 3           Potato and Fennel Soup Hodge 3.750    165      6      7  
## 5           Spinach Noodle Casserole 3.125    547     20     32  
## 6           The Best Blts 4.375    948     19     79  
## 9           Korean Marinated Beef 4.375    170      7     10  
##   sodio  sobremesa  
## 1   559      0  
## 2  1439      0  
## 3   165      0  
## 5   452      0  
## 6  1042      0  
## 9  1272      0  
  
## [1] 15864      7
```

Observando o percentil 99 tem-se que 99% dos valores são menores ou iguais a 3257 kcal. Tomarei como limite das observações de caloria, pois é improvável/inviável que receitas ultrapassem de maneira tão acentuada esse valor; receitas com milhões de kcal são claramente erros nessa base de dados. Foi feito um tratamento semelhante para a variável *sódio*.

Análise descritiva



Modelo base

O modelo principal será analisado da seguinte maneira: o conjunto de dados selecionados será dividido em *conjunto_treino* (70%) e *conjunto_teste* (30%)

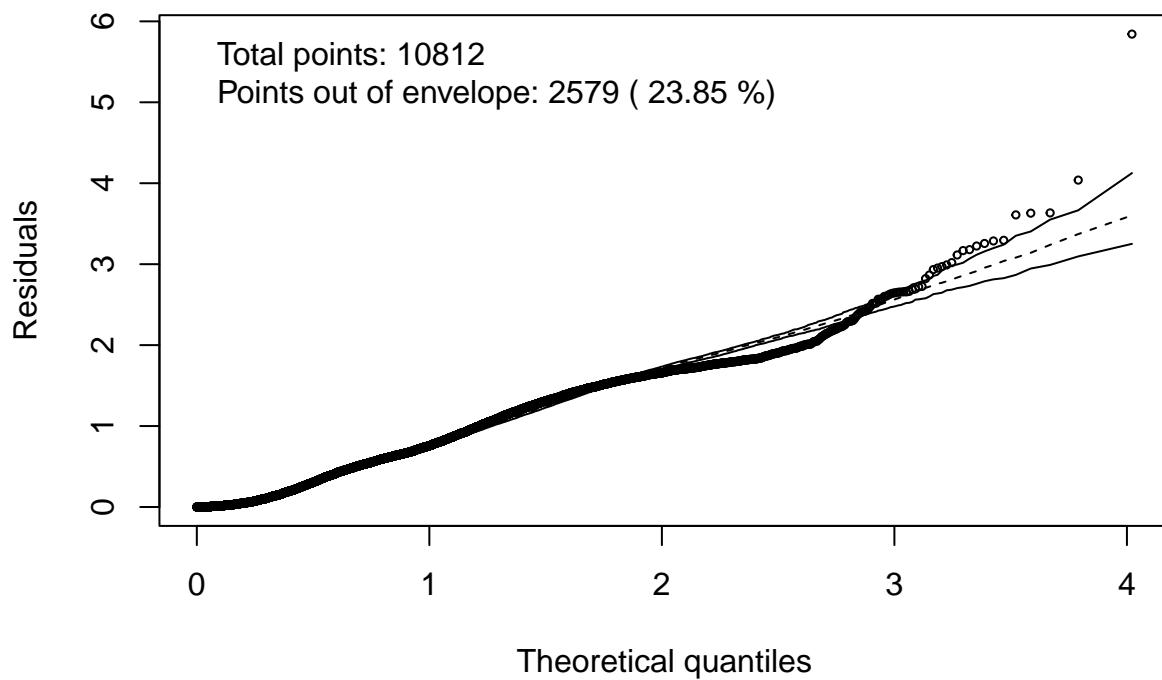
```
##
## Call:
## glm(formula = sobremesa ~ calorias + proteina + gordura + sodio,
##      family = binomial(), data = df1_treino)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.4896170  0.0496225 -30.02 <2e-16 ***
## calorias     0.0081041  0.0002651  30.57 <2e-16 ***
## proteina    -0.1654623  0.0071855 -23.03 <2e-16 ***
## gordura     -0.0347271  0.0028803 -12.06 <2e-16 ***
## sodio       -0.0031312  0.0001715 -18.25 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 10510.4  on 10811  degrees of freedom
```

```

## Residual deviance: 7101.2 on 10807 degrees of freedom
## AIC: 7111.2
##
## Number of Fisher Scoring iterations: 8

## Binomial model

```



```

##
## Call: glm(formula = sobremesa ~ calorias + proteina + gordura + sodio,
##           family = binomial())
##
## Coefficients:
## (Intercept)    calorias     proteina     gordura      sodio
## -1.4194366    0.0008484   -0.0019997   -0.0078761   -0.0004182
##
## Degrees of Freedom: 15863 Total (i.e. Null); 15859 Residual
## Null Deviance: 15250
## Residual Deviance: 14440      AIC: 14450

```

Parte final (teste com outros tamanhos amostrais)

Para esse conjunto de dados, realmente é necessário utilizar todos dados selecionados para ajustar um modelo satisfatório?

Referências

<https://rpubs.com/mpfoley73/527573>

<https://www.kaggle.com/code/rtatman/regression-challenge-day-1>

[https://www.kaggle.com/code/rtatman/datasets-for-regression-analysis#Poisson-regression-\(predicting-a-count-value\)](https://www.kaggle.com/code/rtatman/datasets-for-regression-analysis#Poisson-regression-(predicting-a-count-value))

<https://www.kaggle.com/datasets/hugodarwood/epirecipes?resource=download>