Insper

Projeto Spark

Fernando Cesar Furtado Ballesteros Fincatti

Gabriela Moreno Boriero

Lais Nascimento da Silva

Professor Fábio Ayres

São Paulo

Dezembro/2021

Tarefa 1

• Quantos reviews existem?

Existem um total de 150962278 reviews.

• Quantos clientes existem?

Existem 33497620 clientes.

• Quantos produtos existem?

Existem 21390118 produtos.

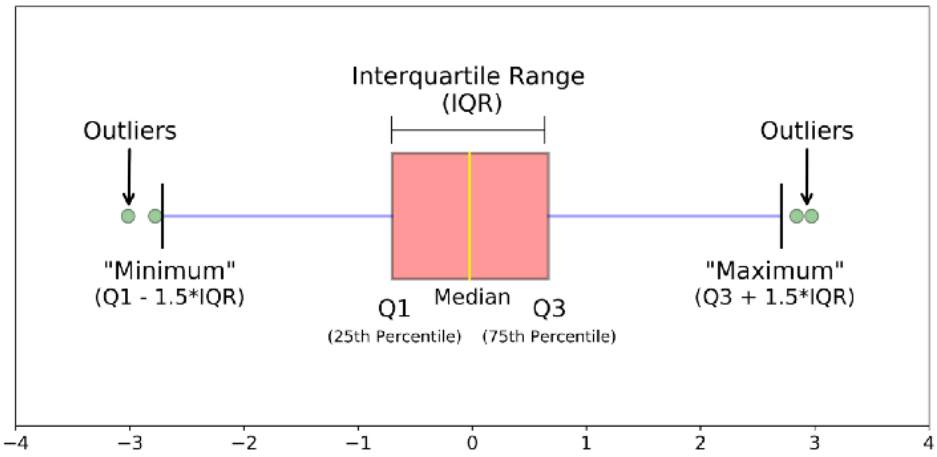
• Quantos reviews existem para cada “star_rating” (de 1 a 5 estrelas)?

Para o “star_rating” 1 existem 12099639 reviews, para o 2 existem 7304430 reviews, para o 3 12133927 reviews, para o 4 26223470 e para o “star_rating” 5 93200812 reviews.

Tarefa 2

Tendo em vista que um alto número de reviews a chance de ser um bot é alta, para definir se foram bots inicialmente utilizamos o esquema do boxplot para definir um valor de reviews o qual provavelmente estará associado a um bot.

Ou seja, iremos pegar os outliers “Maximum”, ou seja, o que possui número de reviews igual a Q3 + 1,5 x (Q3-Q1). Como é possível observar na imagem abaixo:



Fonte: https://ichi.pro/pt/compreendendo-boxplots-103551522880849

Com a analise Q1 foi de 1 review e o Q3 foi de 4 review, desta forma obtivemos um Outlier maior de 18 reviews. No entanto 18 reviews ainda é um número possível para um ser humano. Ademais tem-se que estes dados são de alguns anos e como a Amazon é um site super famoso, requisitado e muito utilizado, nele existem compradores empolgados, que fazem muitas aquisições no site e em todas fazem questão de deixar suas reviews. Sendo assim, consideramos que um valor superior a 15 vezes o Outilier, ou seja, 15 vezes o 18 seria um bom número de corte. Portanto os clientes que deram mais de 270 reviews foram considerados bots.

Desse modo, tem-se que o total de bots é de 10550. Analisando os bots tem-se que eles fazem reviews de diferentes tipos de produtos, a grande maioria, 1693098, faz do tipo Books, em segundo lugar, com 1159694 avaliações feitas por bots, vem o tipo Digital_Ebook_Purchase. Em seguida vem Video DVD, com 479017 e Music com 447649. Essa lista continua como é mostrado na imagem abaixo:

_c6	count
Books	1693098
Digital_Ebook_Pur...	1159694
Video DVD	479017
Music	447649
Home	128111
Health & Personal...	123344
Toys	119022
Kitchen	115491
Beauty	114688
PC	104216
Grocery	95887
Wireless	84966
Office Products	69379
Apparel	66722

Diante disso, é possível verificar um comportamento padrão de bots, visto que a grande maioria realiza reviews de produtos de entretenimento, com livros, vídeos e músicas. AS avaliações são bem distribuídas em negativas, positivas e neutras, considerou-se que as avaliações com 5 estrelas são positivas, com 4 estrelas são neutras e 3 ou menos são avaliações negativas, os resultados obtidos foram: 1520705 avaliações neutras, 3114262 avaliações positivas e 1011678 avaliações negativas. Dessa forma é possível dizer que a tendência de avaliação é positiva.

Assim, tem que se o cliente realiza mais de 270 reviews, e a maioria dos reviews são de produtos do tipo entretenimento com livros, vídeos e músicas e além disso, a as avalições são majoritariamente positivas, há grande probabilidade de ser bot.

Tarefa 3

Foram separados os grupos entre positivos (2), neutro (1) e negativos (0), depois disso, houve a separação em treino e teste. O modelo naive-Bayes foi treinado. Após esse procedimento de a acurácia de teste foi de 74.17%.