PREDICCIÓN DE SENTIMIENTOS EN UN E-COMMERCE

PROYECTO MACHINE LEARNING. BOOTCAMP DATA SCIENCE

ESTHER VEGUILLAS.

1.206 prendas

20 tipos

3 niveles de categorías

22.641 comentarios

Recomendado si/no

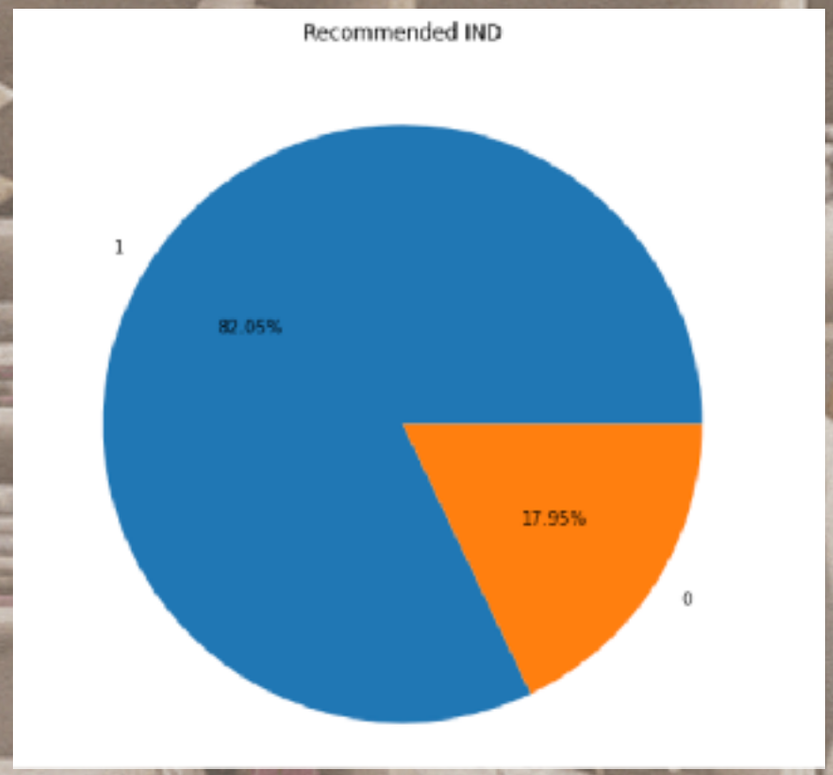
Star rating

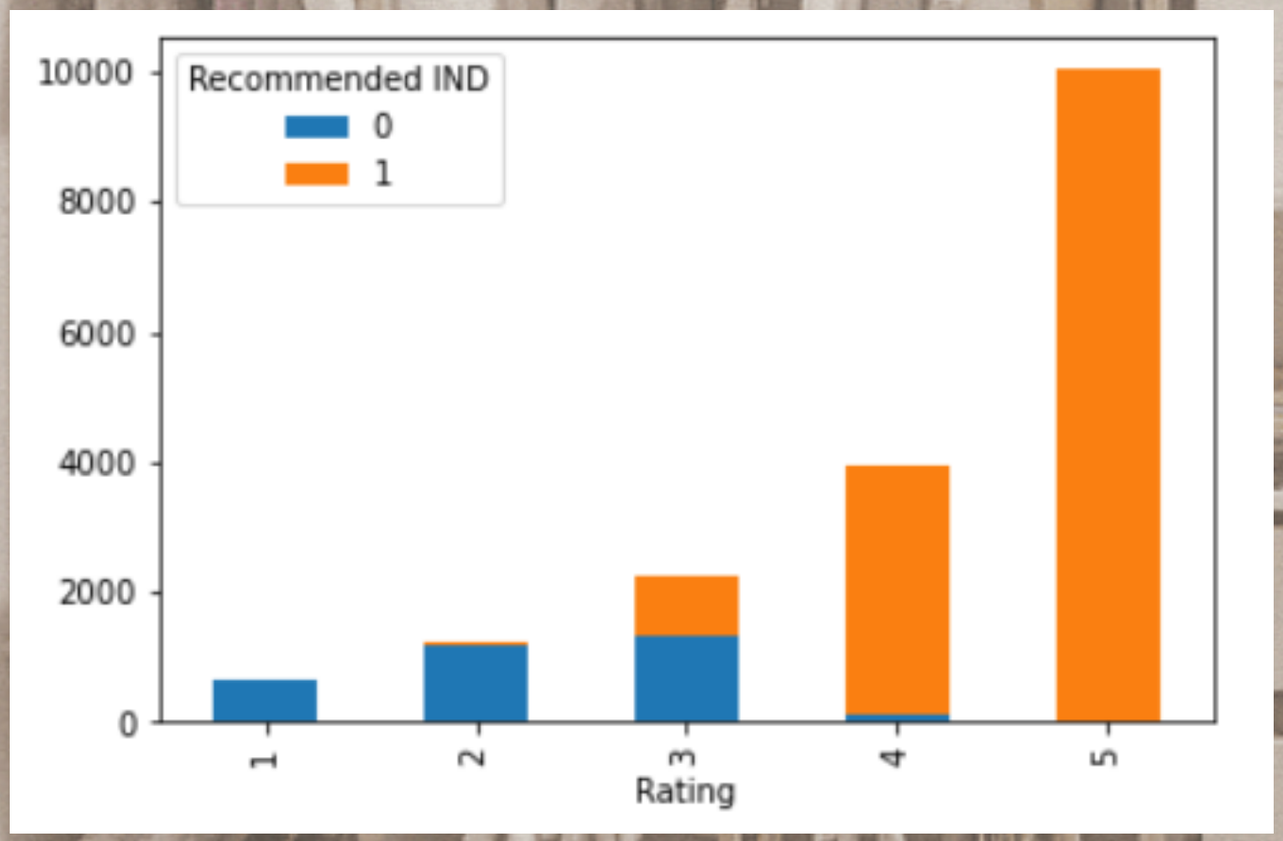
★★★★★

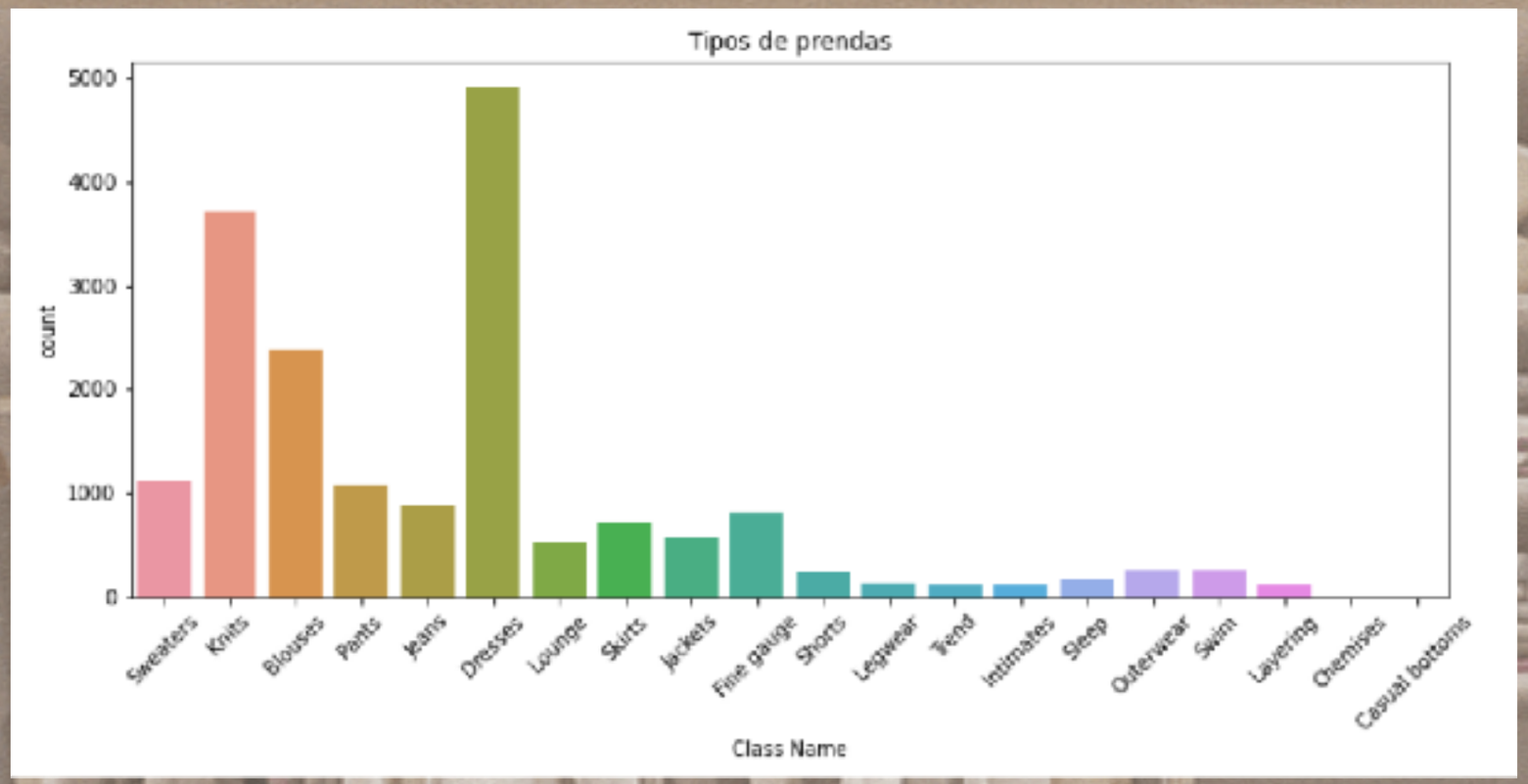
18 a 99 años

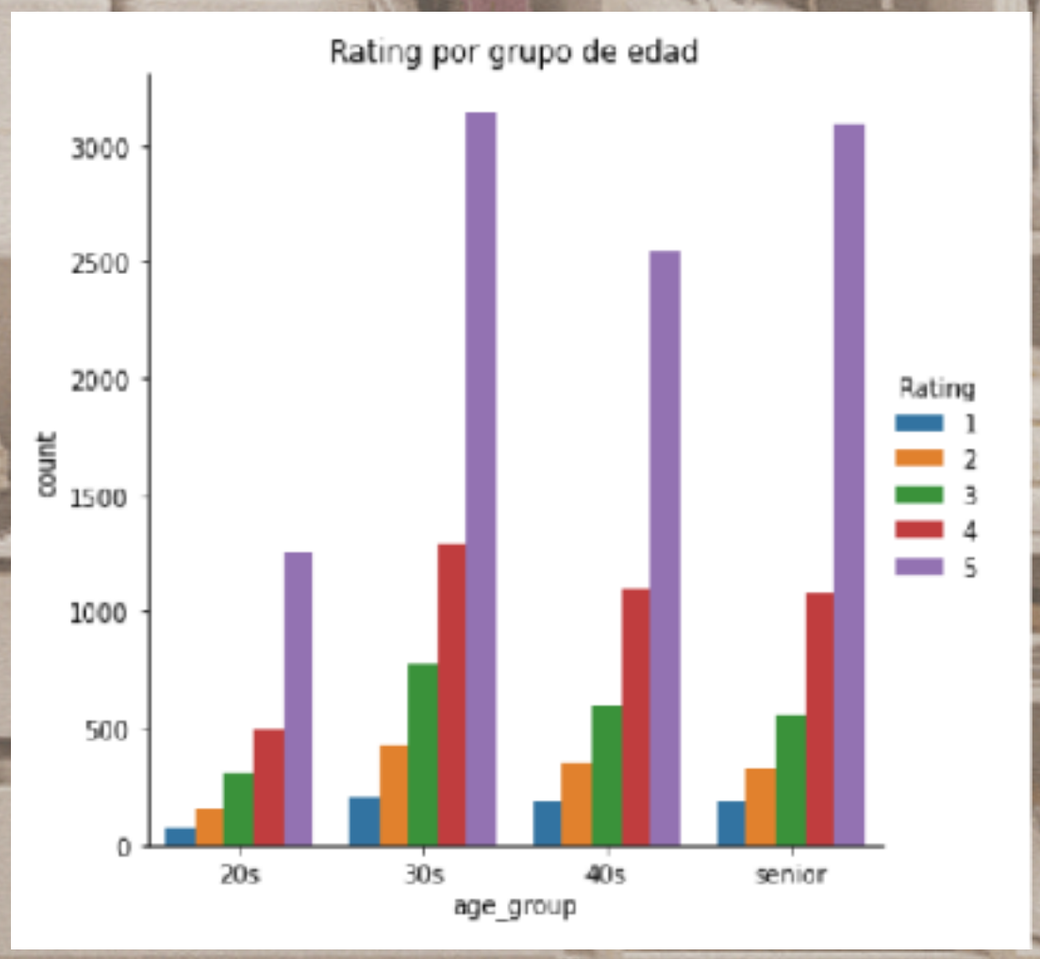
PREPROCESADO PARA EDA

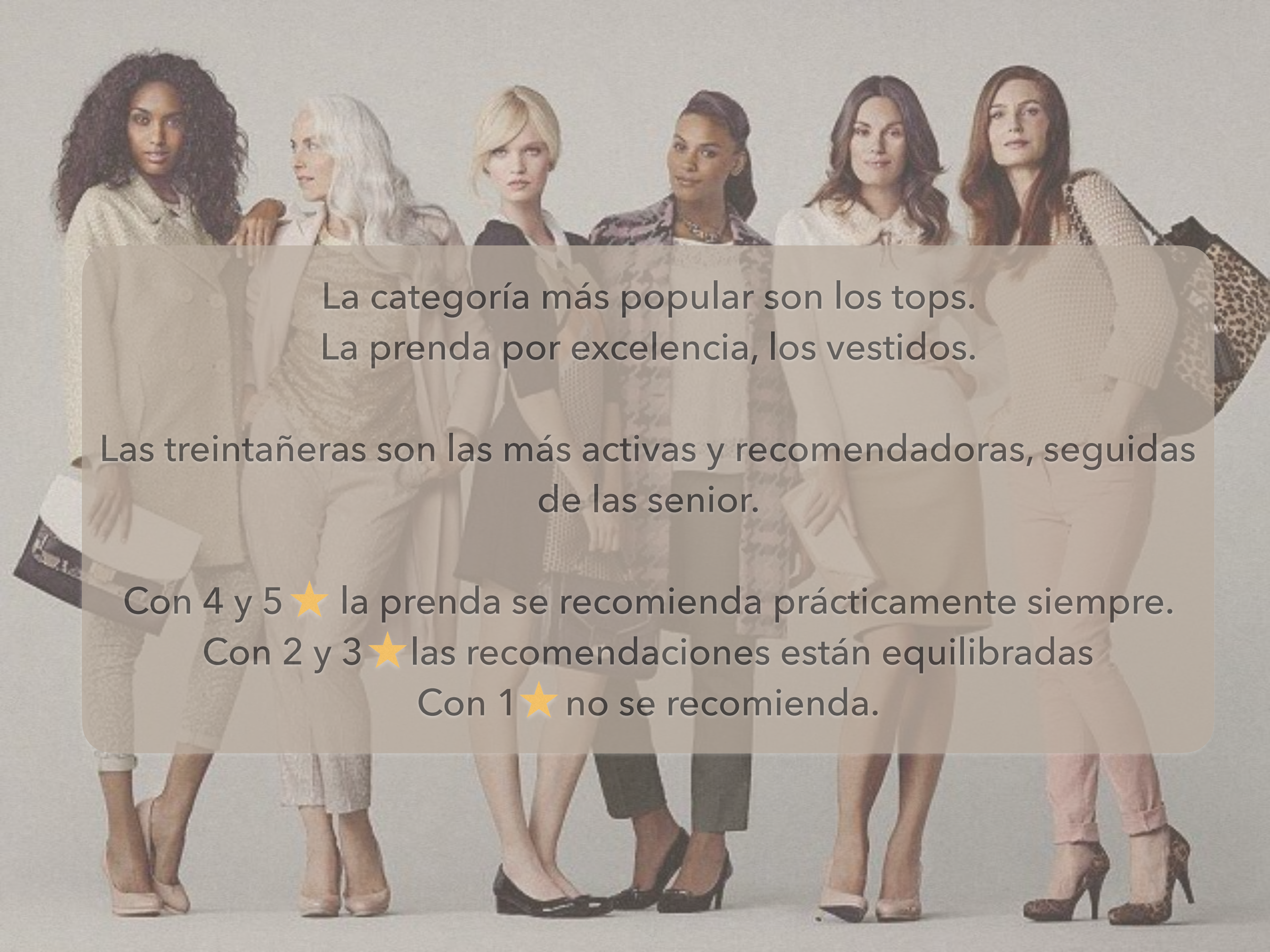
- Elimino registros sin comentario
- Separo train/test
- Relleno categorías con moda (de train)
- Agrupo edades en décadas
- Ignoro outliers


Recommended IND
1
82.05%
17.95%
0


Recommended IND
0
1
10000
8000
6000
4000
2000
0
1
2
3
4
5
Rating


Tipos de prendas
count
5000
4000
3000
2000
1000
0
Sweaters
Knits
Blouses
Pants
Jeans
Dresses
Lounge
Skirts
Jackets
Fine gauge
Shorts
Legwear
Trend
Intimates
Sleep
Outerwear
Swim
Layering
Chemises
Casual bottoms
Class Name


Rating por grupo de edad
count
3000
2500
2000
1500
1000
500
0
20s
30s
40s
senior
age_group
Rating
1
2
3
4
5



La categoría más popular son los tops.
La prenda por excelencia, los vestidos.

Las treintañeras son las más activas y recomendadoras, seguidas de las senior.

Con 4 y 5 ★ la prenda se recomienda prácticamente siempre.
Con 2 y 3 ★ las recomendaciones están equilibradas
Con 1 ★ no se recomienda.

¿Sabrías distinguir cuál es positivo/negativo?


love
fit
color
size
great
wear
fabric
perfect
beautiful
look
comfortable
flattering
cute
would
little
shirt
bought
ordered
material
pretty
nice
i'm
back
one
style
soft
small
work
also
make
think
petite
retailer
white
still
store
wearing


would
fit
look
love
fabric
color
size
like
shirt
material
small
back
cute
great
way
ordered
also
really
large
one
wear
flattering
beautiful
however
model
make
work
even
looked
disappointed
sleeve
much
pretty
design
wanted
return
made
picture
quality
thought
cut
big
online
waist
loved
front
see
think
nice
retailer
thin
huge
tried

PREDICCIÓN DE SENTIMIENTO

- Fusión de título y comentario
- Recomendación como target, cambiando polaridad
- Eliminar prendas de los Wordstops
- CountVectorizer + lemmatizer + stemmer + n-grams + VarianceThreshold

MEJORES MODELOS

Proceso manual:
LogisticRegression
CountVectorizer
n-gram (bigram + trigram)
VarianceThreshold(0.0005)
25.991 features

```
Confusion Matrix:
       0    1
0   3493  186
1    280  569

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.95      0.94      3679
           1       0.75      0.67      0.71       849

    accuracy                           0.90      4528
   macro avg       0.84      0.81      0.82      4528
weighted avg       0.89      0.90      0.89      4528
```

```
Confusion Matrix:
       0    1
0   3481  198
1    283  566

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94      3679
           1       0.74      0.67      0.70       849

    accuracy                           0.89      4528
   macro avg       0.83      0.81      0.82      4528
weighted avg       0.89      0.89      0.89      4528
```

Usando Gridsearch:
SVC,
RandomForest,
LogisticRegression ✓
772.779 features

It looked beautiful on the web, but it is very wide and the fabric looks like plastic

I like the shape of this skirt, but it is a difficult color to match. I don't know if I'm going to be able to wear it a lot

These jeans are great for college or going out with friends. I love how they fit me

It looked beautiful on the web, but it is very wide and the fabric looks like plastic

I like the shape of this skirt, but it is a difficult color to match. I don't know if I'm going to be able to wear it a lot

These jeans are great for college or going out with friends. I love how they fit me

CONCLUSIONES ML

- A pesar de que el dataset tenía un marcado desequilibrio en el target, se ha conseguido un ratio de acierto (recall) para el target desfavorable ("no recomienda") del 67%, obteniendo el target mayoritario ("sí recomienda") un acierto del 95%.
- El modelo que mejor respuesta ha ofrecido ha sido la regresión logística combinada con tratamiento de texto (vectorización, n-gram y reducción de features en función de la desviación típica).
- Para este caso particular, la diferencia en el tratamiento del texto con lemmatizacion o con stemmer no ha sido significativa. Sin embargo, aplicar bigramas y trigramas ha resultado más beneficioso.
- En cuanto a la reducción de features, ha funcionado mucho mejor VarianThreshold que SelectKBest.

Muchas gracias

:)

Git:

https://github.com/Estherveg

LinkedIn:

https://www.linkedin.com/in/eveguillas/

Imágenes: Pexels (christian-diokno, ksenia-chernaya, rosana-solis)