Sentiment Analysis

Cyril Lagneau
Salma Fetjah
Matthis Assalier

SOMMAIRE:

Introduction : Présentation du Projet

1 - Démonstration de l’application

2 - Présentation du jeu de donnée

3- Exploration de l’analyse des données

4- Préprocessing / Models Machine learning

5- Preprocessing / Model Deep

Conclusion: Pistes d’amélioration / Next steps

Présentation du projet

BESOIN DU CLIENT

Améliorer son service d'e-commerce en comprenant les avis négatifs

OBJECTIF

Créer un model capable d'analyser les sentiments

DONNÉES

Commentaire laissé par des femmes sur un e-commerce

1 Démonstration de l'application

https://sheltered-meadow-42281.herokuapp.com

2 - Présentation du jeu de donnée

FEATURES

9 colonnes

TARGET

Recommended IND

0 = non recommandé

1 = recommandé

DIMENSION

23486 observations

10 collones

DONNÉES MANQUANTES

845

DOUBLONS

Aucun

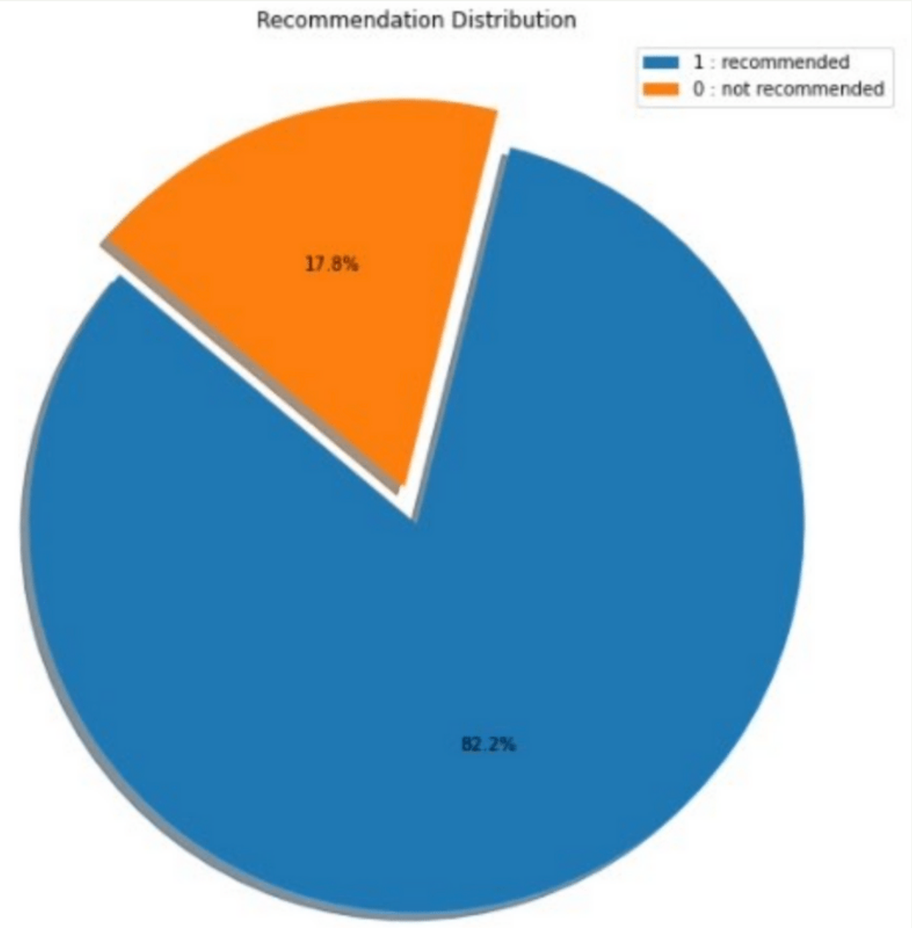
CLEANING

Suppression des valeurs nulles

EDA - Features Information

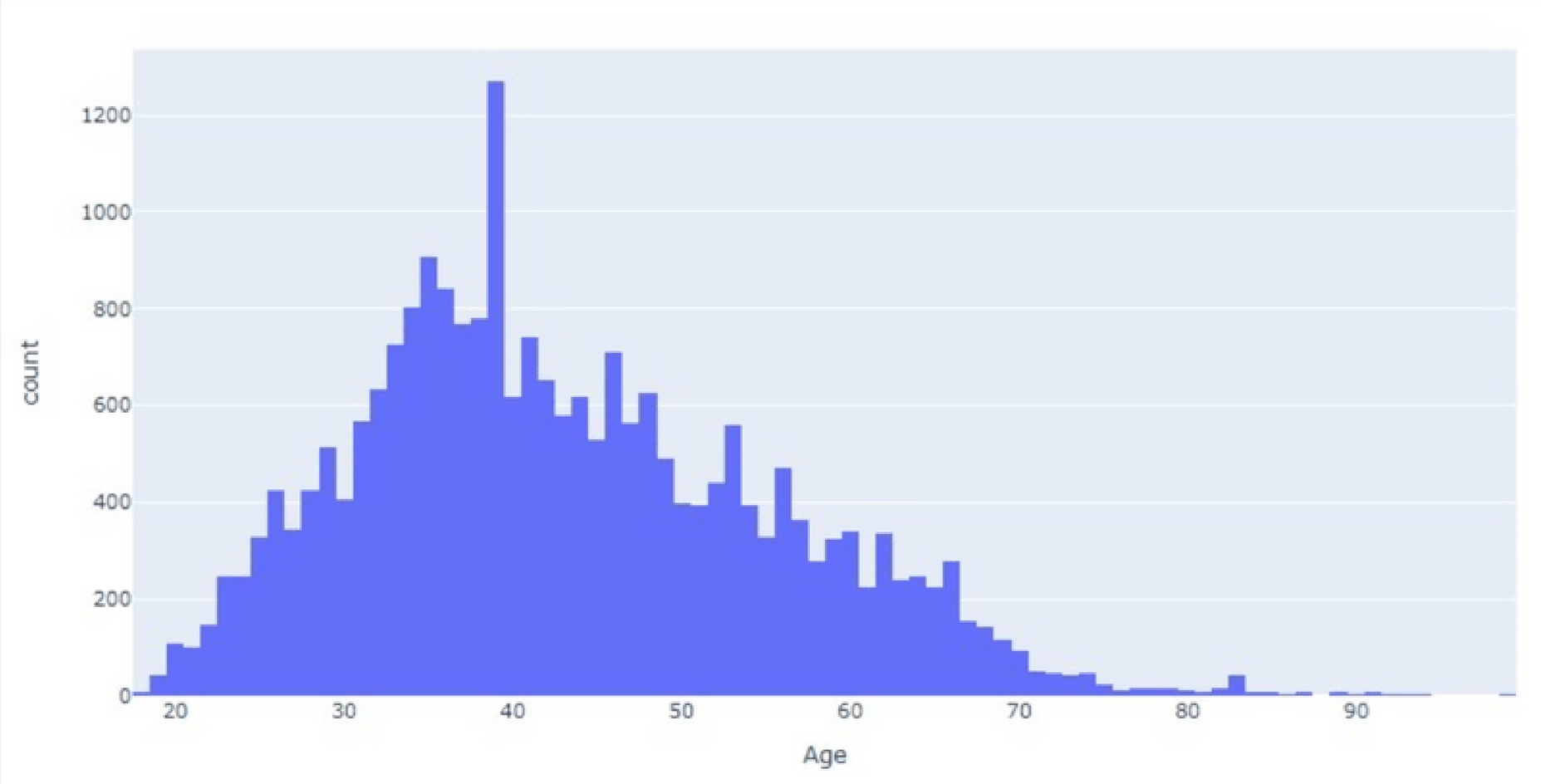
- **Clothing ID**: Integer Categorical variable that refers to the specific piece being reviewed.
  Age: Positive Integer variable of the reviewers age.
- **Title**: String variable for the title of the review.
- **Review Text**: String variable for the review body.
- **Rating**: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- **Recommended IND**: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count**: Positive Integer documenting the number of other customers who found this review positive.
- **Division Name**: Categorical name of the product high level division.
- **Department Name**: Categorical name of the product department name.
- **Class Name**: Categorical name of the product class name.

EDA - Target



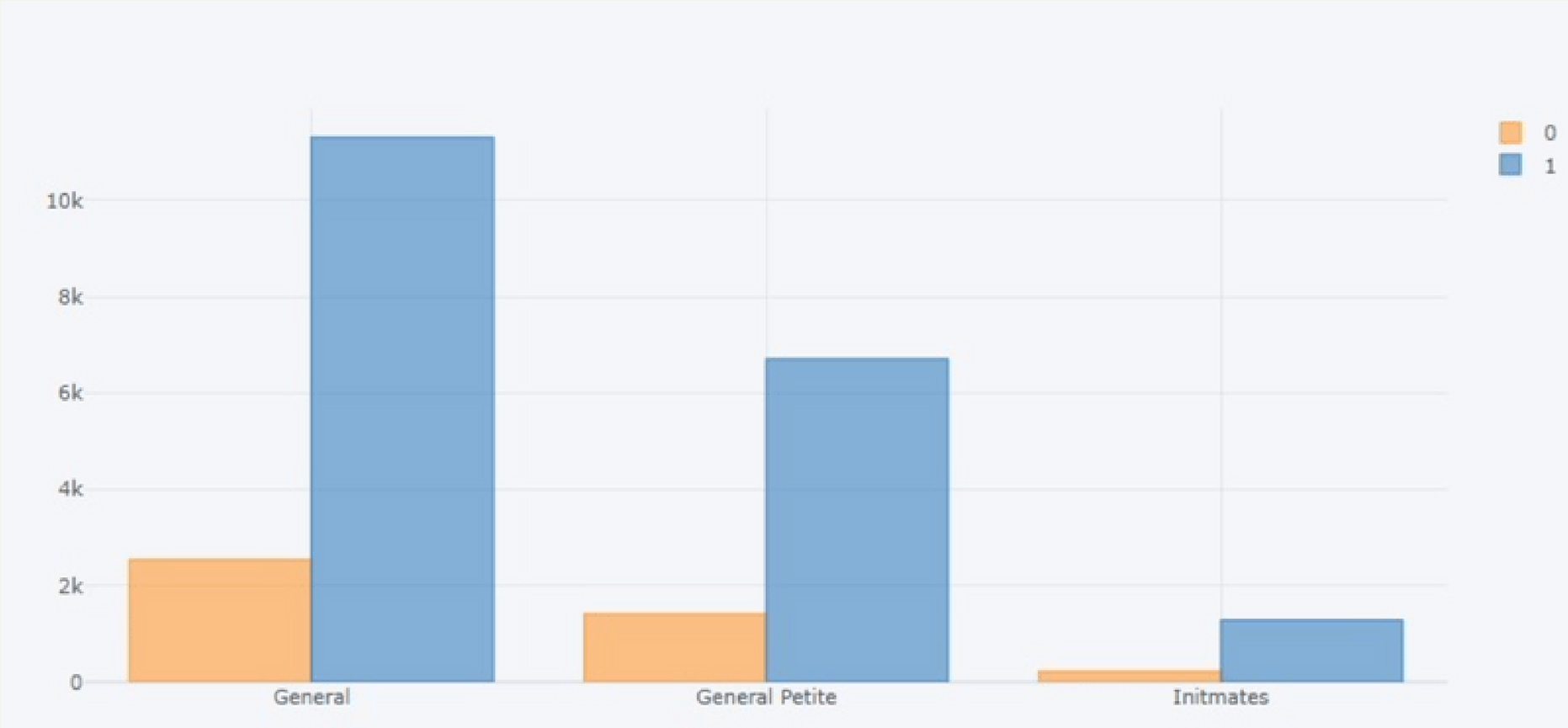

```
1    19314
0     4172
Name: recommended, dtype: int64
```

EDA - Age

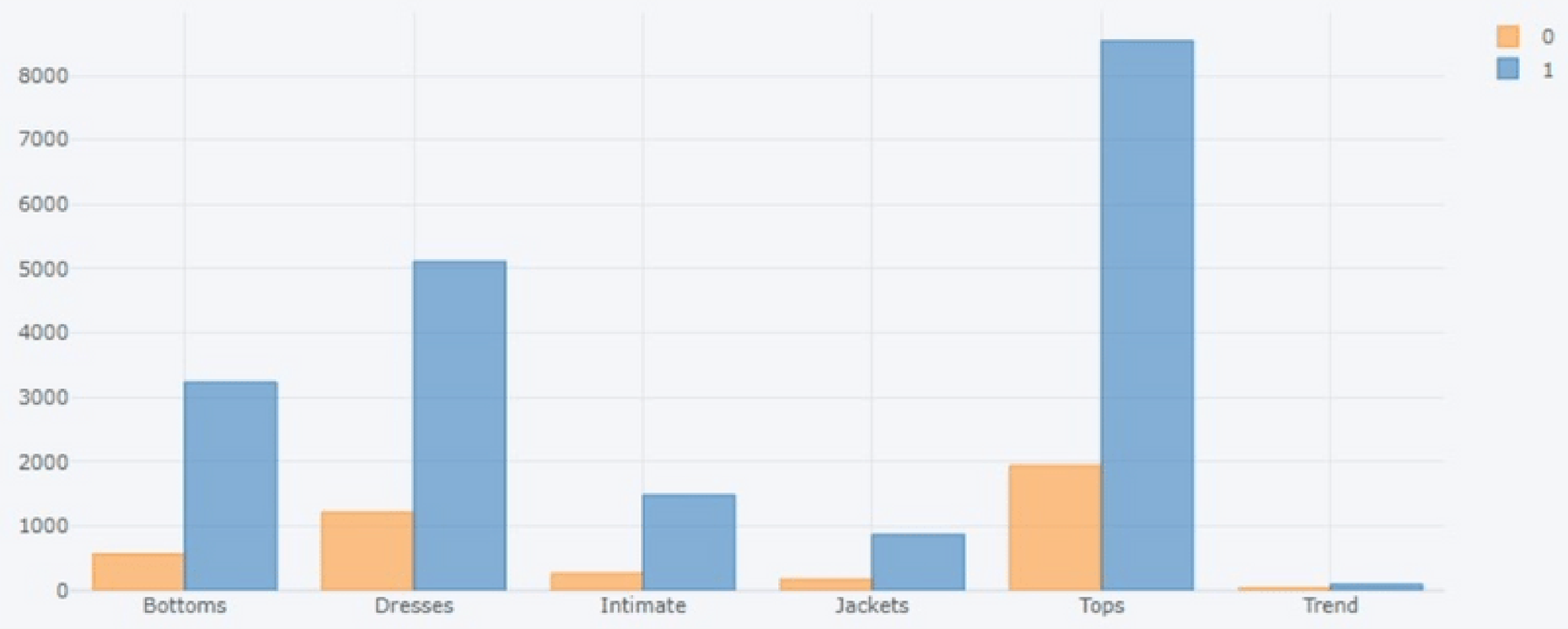


EDA - Rating

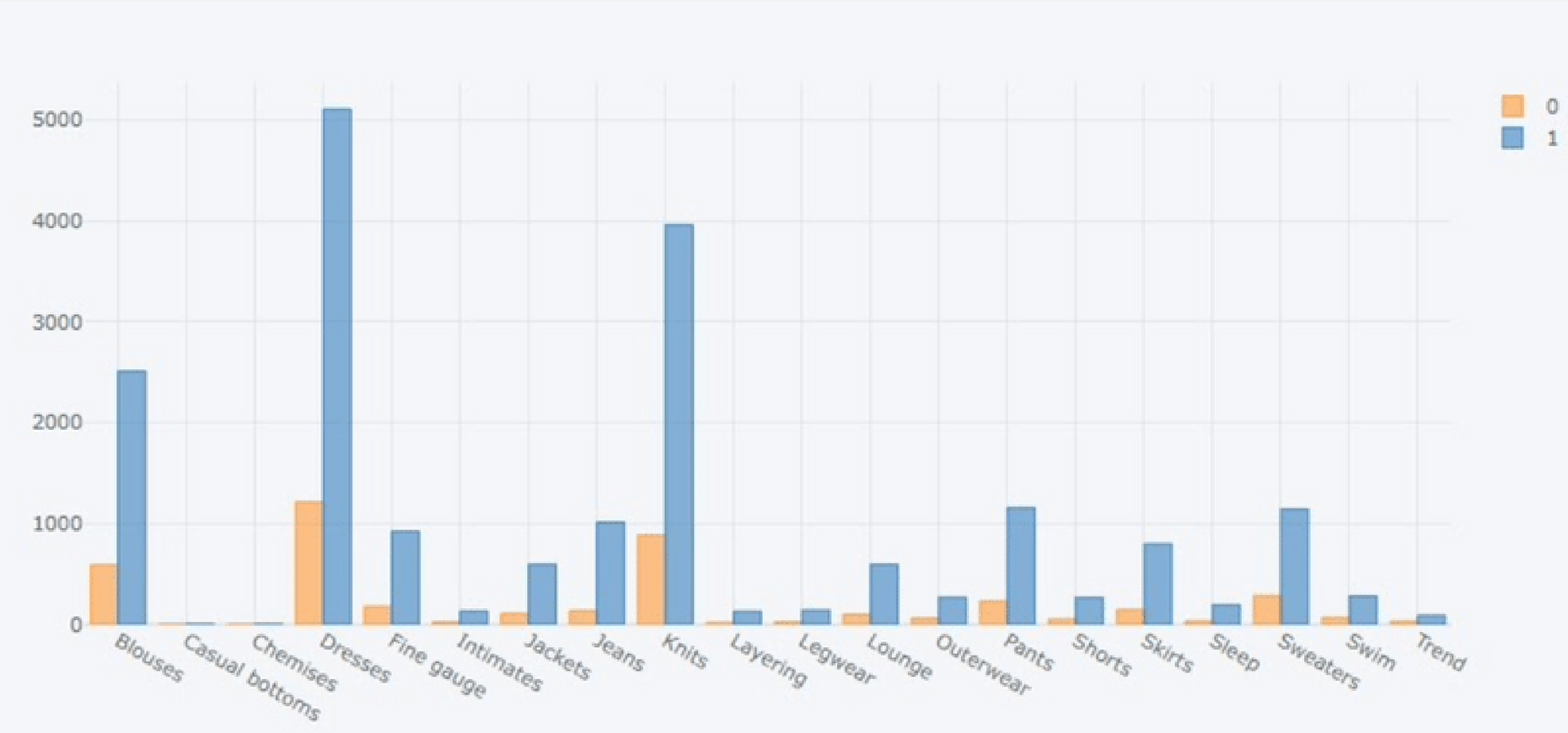
EDA - Division Name



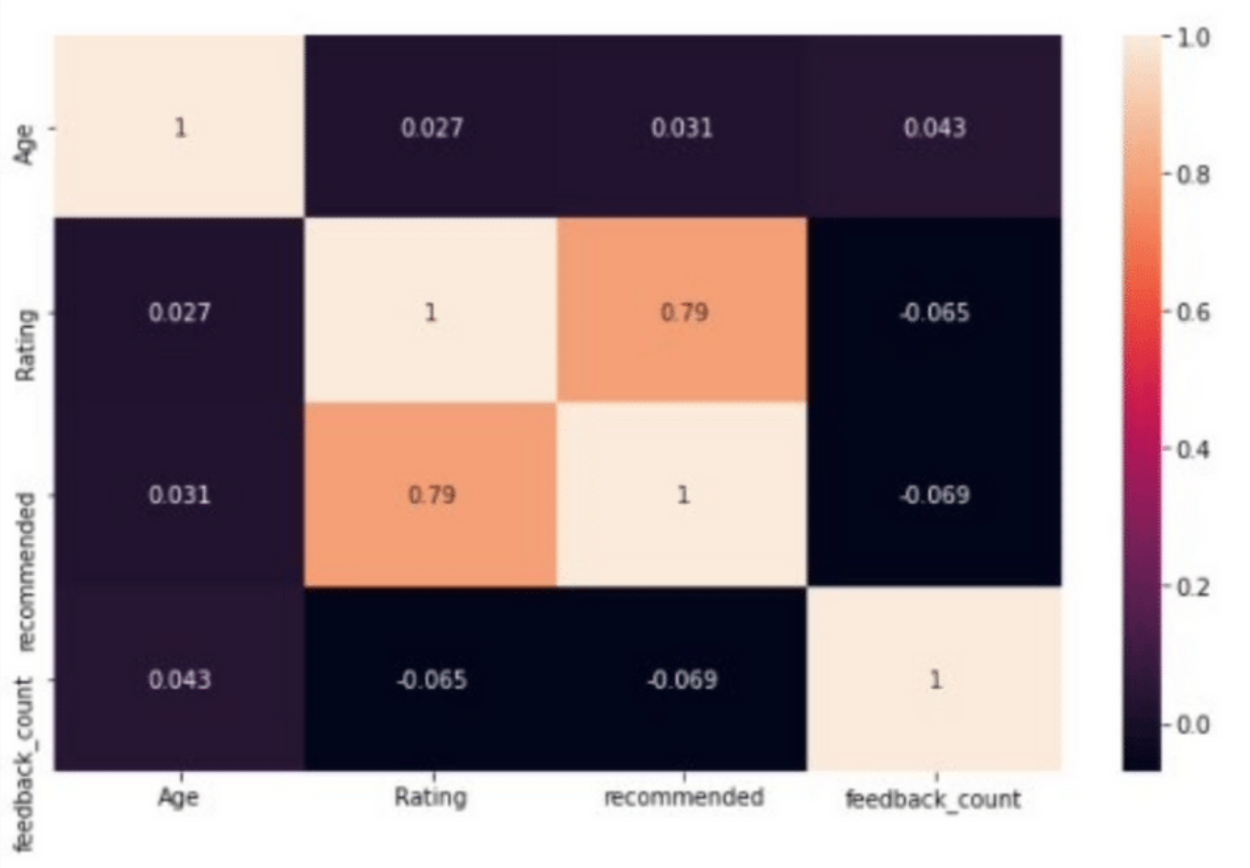
EDA - Département



EDA - Class


5000
4000
3000
2000
1000
0
0
1
Blouses
Casual bottoms
Chemises
Dresses
Fine gauge
Intimates
Jackets
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
Pants
Shorts
Skirts
Sleep
Sweaters
Swim
Trend

EDA - Corrélation



Text Mining - Préprocessing

Tokenization, Noise Removal, Lexicon Normalization

```
stop_words = stopwords.words('english')
```

```
def cleaning(data):

    #1. Tokenize
    text_tokens = word_tokenize(data.replace("'", "").lower())

    #2. Remove Puncs
    tokens_without_punc = [w for w in text_tokens if w.isalpha()]

    #3. Removing Stopwords
    tokens_without_sw = [t for t in tokens_without_punc if t not in stop_words]

    #4. lemma
    text_cleaned = [WordNetLemmatizer().lemmatize(t) for t in tokens_without_sw]

    #joining
    return " ".join(text_cleaned)
```

Text Mining - Préprocessing

1

```
df_ml.head()
```

	text	recommend
0	Absolutely wonderful - silky and sexy and comfortable	1
1	Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's petite. i bought a petite and am 5'8". i love the length...	1
2	I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual size) but i found this to be outrageously small. so small in fact that i co...	0
3	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!	1
4	This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to wear with leggings and it is sleeveless so it pairs well with any cardigan. love this shirt!!!	1

2

```
df_ml["text"] = df_ml["text"].apply(cleaning)
df_ml["text"].head()
```

```
0
absolutely wonderful silky sexy comfortable
1                        love dress sooo pretty happened find store im glad bc never would ordered online bc petite bought petit
e love length hit little knee would definitely true midi someone truly petite
2    high hope dress really wanted work initially ordered petite small usual size found outrageously small small fact could zip
reordered petite medium overall top half comfortable fit nicely bottom ha...
3                                                                                                                  love love l
ove jumpsuit fun flirty fabulous every time wear get nothing great compliment
4                                                                                                  shirt flattering due adjustable f
ront tie perfect length wear legging sleeveless pair well cardigan love shirt
Name: text, dtype: object
```

3

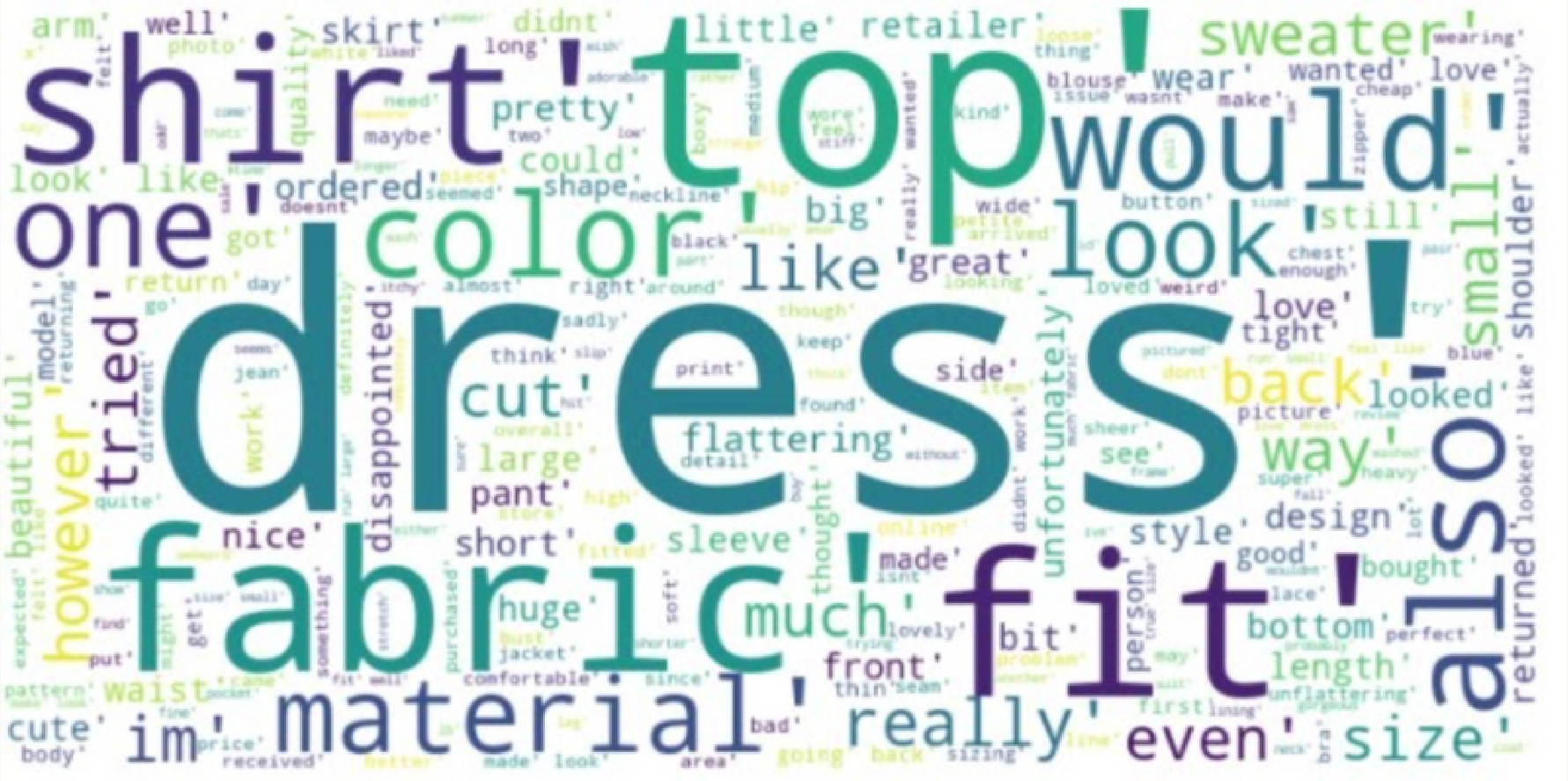
```
" ".join(df_ml["text"]).split()
```

```
['absolutely',
 'wonderful',
 'silky',
 'sexy',
 'comfortable',
 'love',
 'dress',
 'sooo',
 'pretty',
 'happened',
 'find',
 'store',
 'im',
 'glad',
 'bc',
 'never',
 'would',
 'ordered',
 'online',
```

Text Mining - Recommend

dress
top
love
like
fabric
shirt
one
skirt
sweater
also
perfect
color
think
im
great
comfortable
style
true
size
waist
pretty
beautiful
make
go
gorgeous
tight
fun
fitted
overall
casual
store
try
come
bottom
long
soft
without
since
flattering
shoulder
different
got
work
look
jean
side
unique
could
retailer
ordered
front
dress
design
length
however
nice
fall
feel
pant
tried
little
button
fit
well
find
big
black
definitely
wearing
lovely
jacket
bought
sleeve
material
detail
white
found
cute
though
see
made
short
purchased
bit
loved
may
print
back
even
usually
wear
right
enough
get
really
much
blouse
would
still
way
pattern
cut
fit
great
online
issue
looked
neckline
tee
first
piece
give
quite
flowy
sizing
arm
cardigan
versatile
well
made
love
thing
keep
actually
perfect
pocket
top
pair
small
size
show
quality
adorable
gorgeous
comfy
day
stretch
shape
thin
light
large
run
loose
looking
didnt
dont
wore
blue
especially
summer
look
like
fabric
fit
true
wear
weight
lace
usual
feel
like
small
fit
perfectly
fit
find
love
pocket
petite
wanted
product
pair
perfect
fit
small
show
keep
thing
love
love
well
made
versatile
right
blue
wore
look
size
great
fit
soft
shape
love
dress
design
fall
nice
however
tried
little
button
fit
well
though
cute
see
found
white
may
loved
bit
detail
sizing
material
sleeve
bought
jacket
flowy
dont
made
piece
purchased
short
quite
give
arm
cardigan
enough
actually
well
comfortable
style
great
even
usually
wear
right
much
really
back
print
cut
look
like
pattern
still
way
would
blouse

Text Mining - Not Recommend


dress
fabric
fit
material
shirt
top
would
one
color
look
also
sweater
small
tried
however
beautiful
cut
back
way
flattering
large
pant
nice
short
sleeve
huge
much
design
style
bought
bottom
length
returned
shoulder
looked
love
tight
still
great
like
ordered
pretty
could
shape
big
unfortunately
disappointed
model
waist
cute
im
really
even
size
front
bit
person
made
good
side
picture
see
wear
wanted
retailer
little
skirt
arm
well
didnt
long
quality
look
maybe
photo
white
liked
adorable
other
need
medium
boxy
wore
feel
stiff
kind
wasnt
make
blouse
issue
thing
loose
wearing
cheap
zipper
actually
button
wide
petite
arrived
chest
enough
weird
loved
looking
sadly
keep
print
found
work
item
sheer
super
heavy
fall
lace
perfect
unflattering
first
linsing
sizing
back
going
area
bad
made
look
better
price
received
body
comfortable
since
thin
seam
jacket
lovely
trying
probably
may
bra
neck
coat
pattern
put
expected
felt
fit
well
fine
jean
work
get
something
purchased
soft
high
store
online
fitted
isnt
didnt
lot
shoulders
quite
different
definitely
returning
go
day
return
got
doesnt
seemed
piece
neckline
black
part
itchy
almost
right
around
though
think
slip
detail
without
overall
large
high
store
quality
boxy
medium

Sentiment Classification with Machine Learning

Preprocessing

4-Préprocessing Machine learning

- Columns: Text review & Recommend
- Droping Missing Values
- Split the Data
- CountVectorizer()
- TfidVectorizer()

```
text            845
recommended       0
dtype: int64
```

```
X = df_ml["text"].values
y = df_ml["recommend"].map({0:1, 1:0}).values
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, stratify = y, random_state = 101)
```

4-Préprocessing Machine learning

```
from sklearn.feature_extraction.text import CountVectorizer
```

```
vectorizer = CountVectorizer()

X_train_count = vectorizer.fit_transform(X_train)
X_test_count = vectorizer.transform(X_test)
```

```
pd.DataFrame(X_train_count.toarray(), columns = vectorizer.get_feature_names())
```

	aa	ab	abby	abdomen	ability	able	abo	absolute	absolutely	absolutley	abstract	absurd	abt	abundance	ac	accent	accented	accentuate	accen
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
18107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18110	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
18111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

18112 rows × 5468 columns

4-Préprocessing Machine learning

```
tf_idf_vectorizer = TfidfVectorizer()

X_train_tf_idf = tf_idf_vectorizer.fit_transform(X_train)
X_test_tf_idf = tf_idf_vectorizer.transform(X_test)
```

```
pd.DataFrame(X_train_tf_idf.toarray(), columns = tf_idf_vectorizer.get_feature_names())
```

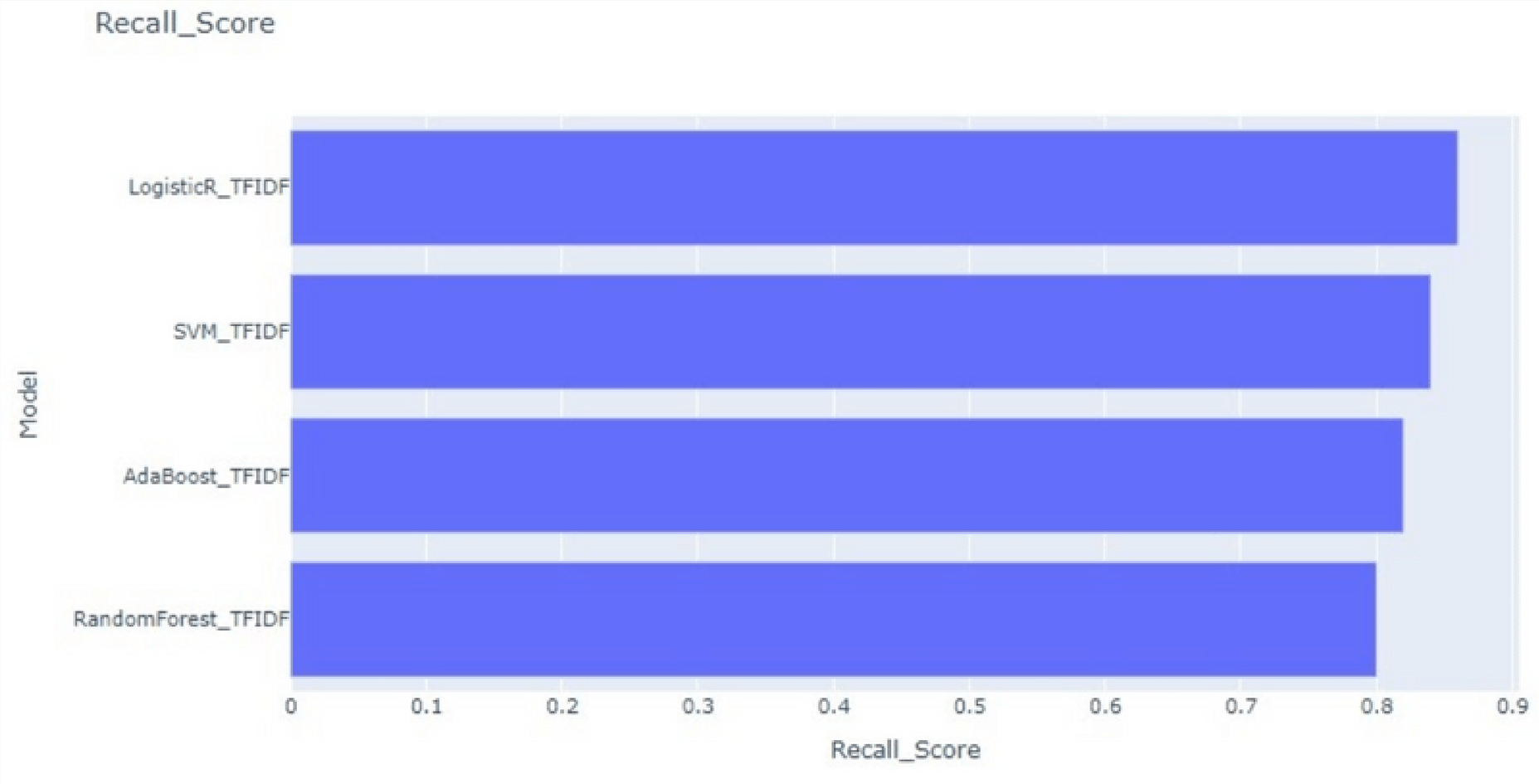
	aa	ab	abby	abdomen	ability	able	abo	absolute	absolutely	absolutley	abstract	absurd	abt	abundance	ac	accent	accented
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.154	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18107	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18108	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18109	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18110	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.324	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18111	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

18112 rows × 5468 columns

Models Machine Learning

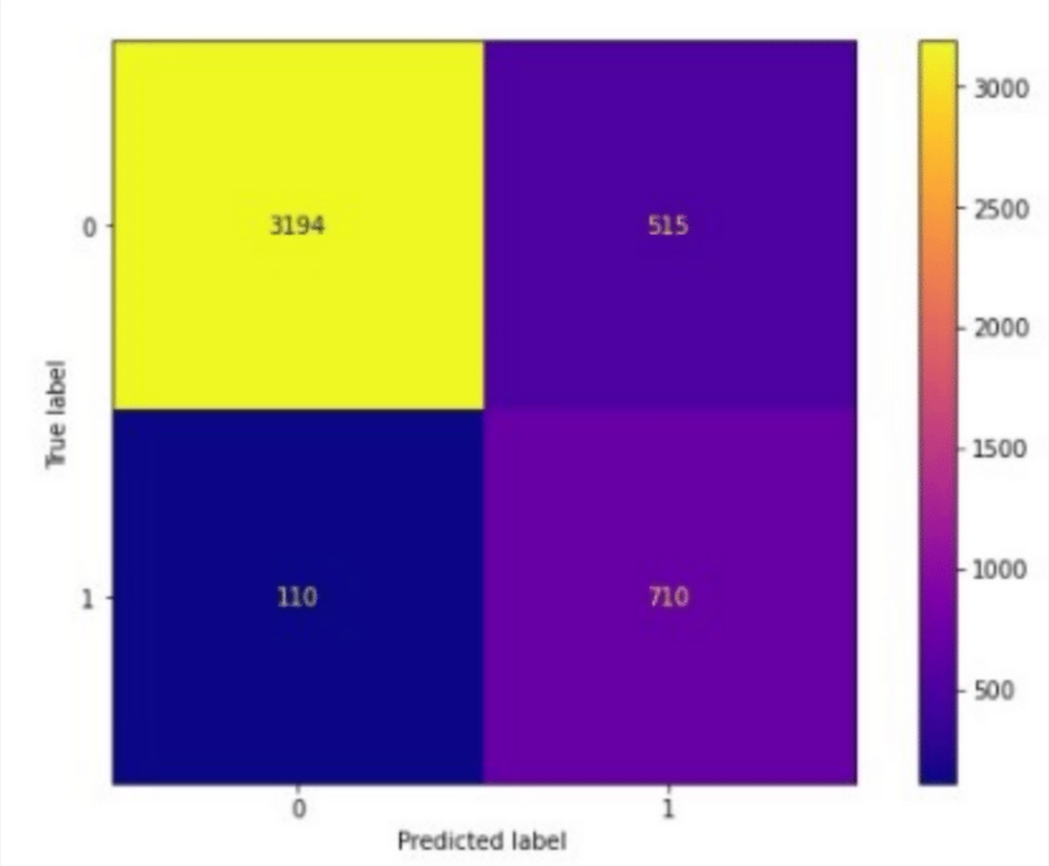
- LogisticRegression
- Support Vector Machine
- Random Forest
- AdaBoost

Résultats - Recall Score



Résultats - Matrice de confusion

```
LogReg_TFIDF Model
------------------
[[3194  515]
 [ 110  710]]
Test_Set
              precision    recall  f1-score   support

           0       0.97      0.86      0.91      3709
           1       0.58      0.87      0.69       820

    accuracy                           0.86      4529
   macro avg       0.77      0.86      0.80      4529
weighted avg       0.90      0.86      0.87      4529

Train_Set
              precision    recall  f1-score   support

           0       0.98      0.89      0.93     14831
           1       0.64      0.93      0.76      3281

    accuracy                           0.89     18112
   macro avg       0.81      0.91      0.84     18112
weighted avg       0.92      0.89      0.90     18112
```



Test

	0
0	I ordered this dress in 0p since i am 5ft. it fits great its not too short just above the knee.
1	Love the top design, not so much the fabric. i should've read the top was 100% polyester. returning and finding a better quality top.
2	Love this henely, the lace-up is cute. the shirt is comfy. so versatile. beach coverup or over jeans
3	Loved this dress online, but my small pettie stature made me looks like an orphan from the prairie.
4	I tried it in the store but was not the true size so i ordered online. i wore it only once so far and got lots of compliments. very unique design.
5	Nice shirt seems well made. good just not a great fit for me.
6	I tried on this dress in store and was amazed by the quality and simple structure of the dress. bought it with no hesitation.
7	This top has a bit of a retro flare but so adorable on. looks really cute with a pair of faded boot cut jeans.
8	Great quality top. i do wish it fit me...i'm really long waisted and this top was too short. had to return
9	A serious joke. i struggled with the buttons for a good 10 minutes and gave up after the 3rd button. i'm not sure what they were thinking.

Good
Bad
Good
Bad
Good
Good
Good
Good
Bad
Bad

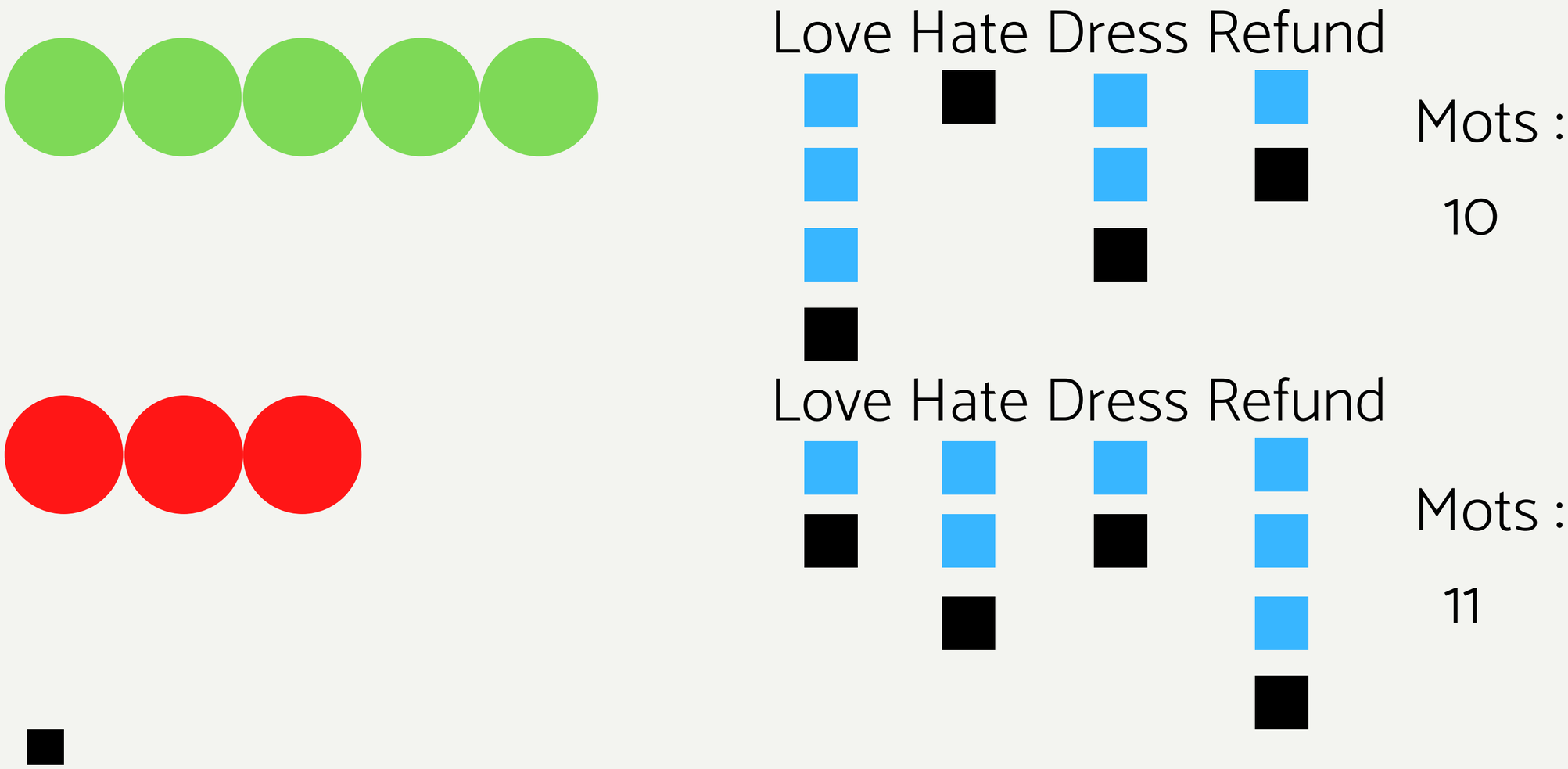
4- Model Naive Bayes

Bonnes recommandations

Mauvaises recommandations

Analyse des fréquences d'apparition de ces mots :

4- Model Naive Bayes



4- Model Naive Bayes

Love	Hate	Dress	Refund
0,4	0,1	0,3	0,2

Love	Hate	Dress	Refund
0,18	0,27	0,18	0,36

4- Model Naive Bayes

Love	Hate	Dress	Refund
0,4	0,1	0,3	0,2

Love Dress

p(B) x p(Love | B) x p(Dress | B)

0,48 x 0,4 x 0,3 = 0,0576

4- Model Naive Bayes

p(M) x p(Love | M) x p(Dress | M)

0,52 x 0,18 x 0,18 = 0,0168

Love Dress

Love	Hate	Dress	Refund
0,18	0,27	0,18	0,36

4- Model Naive Bayes

Love Dress

0,0576 > 0,0168 --> Donc, bonne recommendation

4- Model Naive Bayes

Modèle NB :

	precision	recall	f1-score	support
0	1.00	0.83	0.90	3914
1	0.03	1.00	0.05	19
accuracy			0.83	3933
macro avg	0.51	0.91	0.48	3933
weighted avg	1.00	0.83	0.90	3933

4- Preprocessing Deep Learning

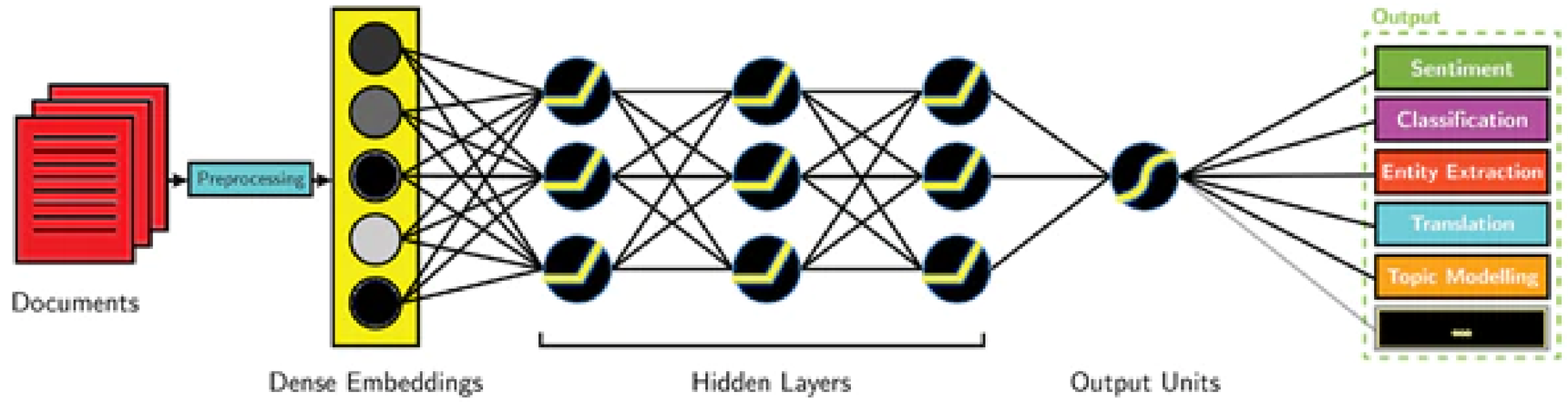
Méthode de tokenisation et séquençage

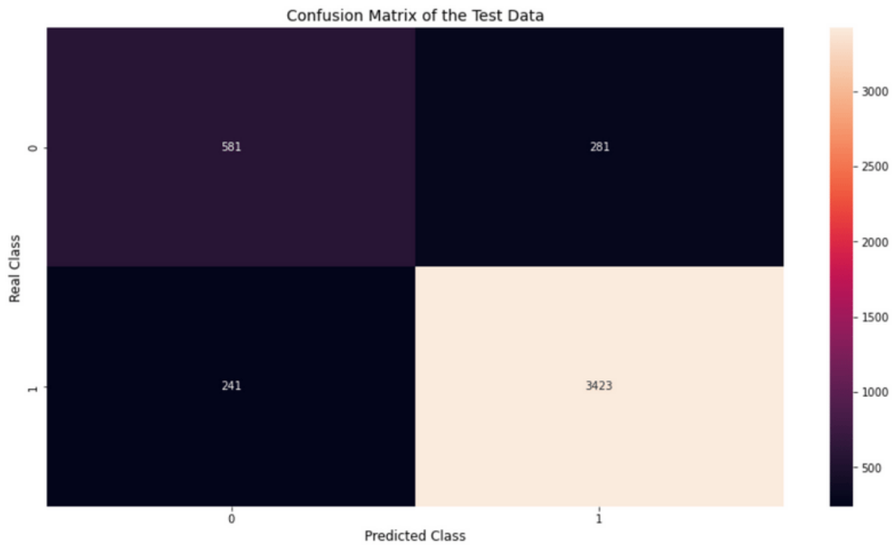
```
Non-tokenized Version:  absolutely wonderful silky sexy comfortable
Tokenized Version:  [[161, 366, 748, 445, 33]]
----------------------------------------
Non-tokenized Version:  usually petite since dress not come petites try fit lbs dress hit knee hem bite not overwhelm dress loo
k stun great vibrant color dark hair make classic elegant dress look contemporary stylish try store salesperson others happen s
ee rave tell grab glad plan wear spring daughte
Tokenized Version:  [[61, 47, 150, 2, 7, 109, 769, 23, 3, 68, 2, 146, 269, 223, 38, 7, 746, 2, 8, 397, 12, 356, 11, 278, 1025,
18, 342, 459, 2, 8, 2151, 344, 23, 57, 1885, 350, 601, 62, 1723, 432, 631, 218, 313, 6, 212, 6315]]
```

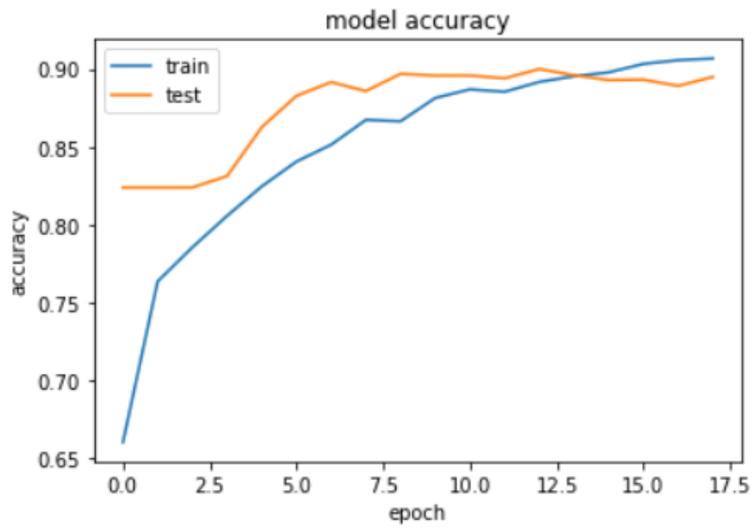
Padding

```
array([[  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0, 161, 366, 748, 445,  33]])
```

4- Model Deep Learning







Précision total: 90%

Score de reconnaissance des commentaires négatif: 70%

Conclusion

Les besoins du client ont été respectés

Next Steps & pistes d'amélioration :

- **EDA** : création d'un dashbord Sentiment Analysis, avec une analyse plus détaillée des profils des clients...
- **Text Mining**:nettoyage du nuage de mots
- **Models:** tunning des models SVM, Adaboost...
- **Features Selection** sur les mots apres TFIDF
- **Déploiement** : Améliorer l'interface streamlit HEROKU