DataScientest • com

Challenge data

Rakuten Institute of Technology

TEAM :

FATOUMATA BARRY

ERIC MARCHAND

EDGAR HIDALGO LOPEZ

EMMANUEL BONNET

Présentation Projet

CHALLENGE RAKUTEN 2020 @ COMMERCE

PRÉSENTATION DU CHALLENGE ET DES ENJEUX

PROJET : **prédire le code type des produits** sur la base de **données textuelles** (désignation et description & de **données images**

ENJEUX : Amélioration du **catalogage des produits** selon des **données différentes** (textes et images) pour réaliser des applications diverses -> **recommandation de produits et recherche personnalisée**

Rang	Marques	Visiteurs uniques moyens par mois	Couverture France mensuelle (en % de la pop. Française)	Visiteurs uniques moyens par jour
1	Amazon *	32 566 000	52,0%	6 141 000
2	Cdiscount *	20 965 000	33,5%	2 317 000
3	Fnac *	16 568 000	26,4%	1 313 000
4	E.Leclerc *	13 156 000	21,0%	1 529 000
5	Booking.com	13 109 000	20,9%	1 190 000
6	Carrefour *	12 987 000	20,7%	1 591 000
7	Veepee *	12 658 000	20,2%	2 767 000
8	Vinted *	12 496 000	19,9%	3 395 000
9	Wish	12 327 000	19,7%	2 388 000
10	OUI.sncf *	11 319 000	18,1%	1 237 000
11	eBay *	10 729 000	17,1%	1 234 000
12	Leroy Merlin *	10 666 000	17,0%	802 000
13	Groupon *	9 631 000	15,4%	1 294 000
14	Rakuten (B)*	9 400 000	15,0%	926 000
15	Airbnb	9 088 000	14,5%	1 030 000

https://www.fevad.com/barometre-trimestriel-de-laudience-du-e-commerce-en-france-enquete-e-commerce-et-confinement/

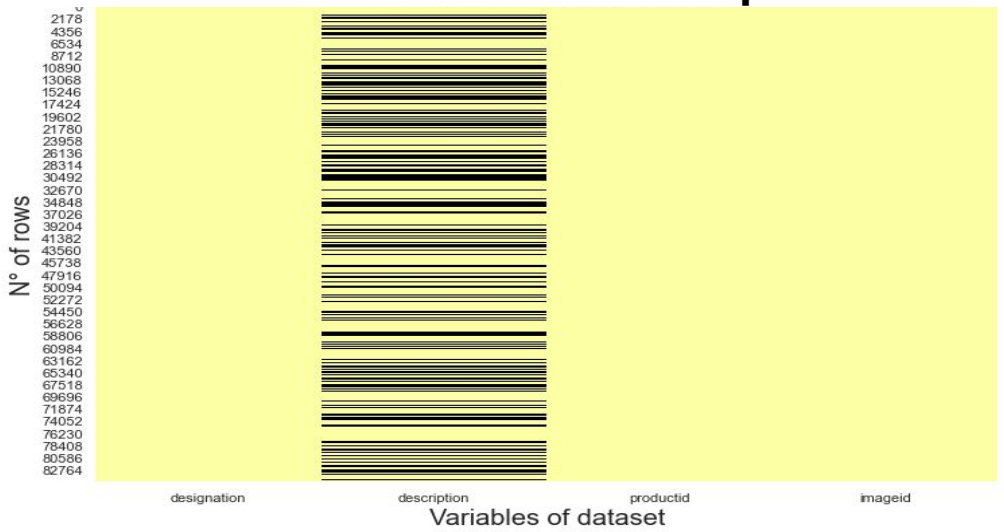
PRÉSENTATION DES DONNÉES

27 Classes sur **6 thèmes principaux** !

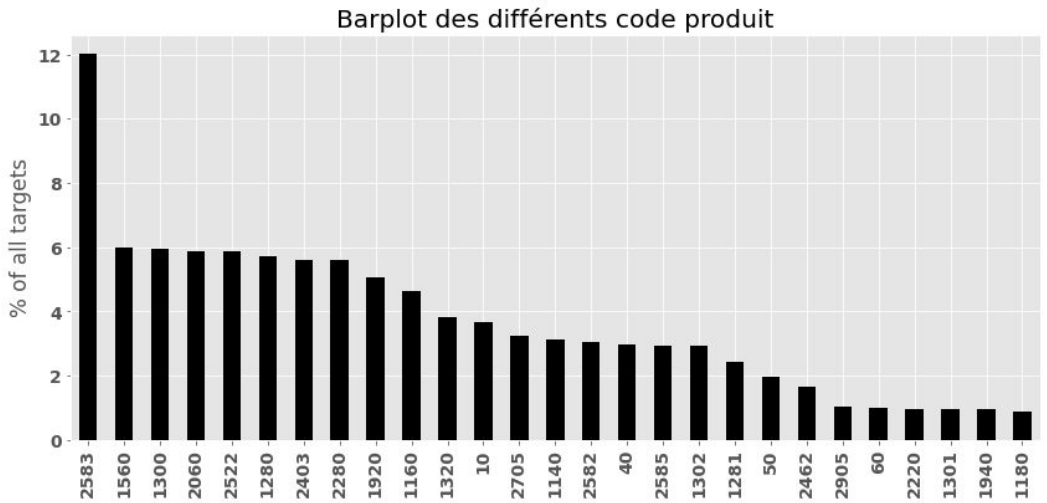
	y	Nombre	Caractéristiques des images
Livres	**10**	3116	Livres occasion
	2280	4760	Journaux et revues occasion
	2403	4774	Livres, BD et magazines
	2522	4989	Fournitures papeterie et accessoires bureau
	2705	2761	Livres neufs
Jeux	**40**	2508	Jeux videos, CDs, équipements, câbles, neufs
	50	1681	Accessoires gaming
	60	832	Consoles de jeux
	2462	1421	Jeux vidéos occasion
	2905	872	Jeux vidéos pour PC
Jouets & figurines	**1140**	2671	Figurines, objets pop culture
	1160	3953	Cartes de jeux
	1180	764	Figurines et jeux de rôles
	1280	4870	Jouets enfants
	1281	2070	Jeux société enfants
	1300	5045	Modélisme
	1302	2491	Jeux de pleins air, Habits
Meubles	**1560**	5073	Mobilier général : meubles, matelas, canapés lampes, chaises
	2582	2589	Mobilier de jardin : meubles et outils pour le jardin
Equipements divers	**1320**	3241	Puériculture, accessoire bébé
	2220	824	Animalerie
	2583	10209	Piscine et accessoires
	2585	2496	Outillages de jardin, équipements technique extérieur maison et piscines
Déco	**1920**	4303	Linge de maison, oreillers, coussins
	2060	4993	Décoration
Autres	**1301**	807	Chaussettes bébés, petites photos
	1940	803	Confiserie

https://challengedata.ens.fr/participants/challenges/35/

NaN's dans la colonne '**description**' !

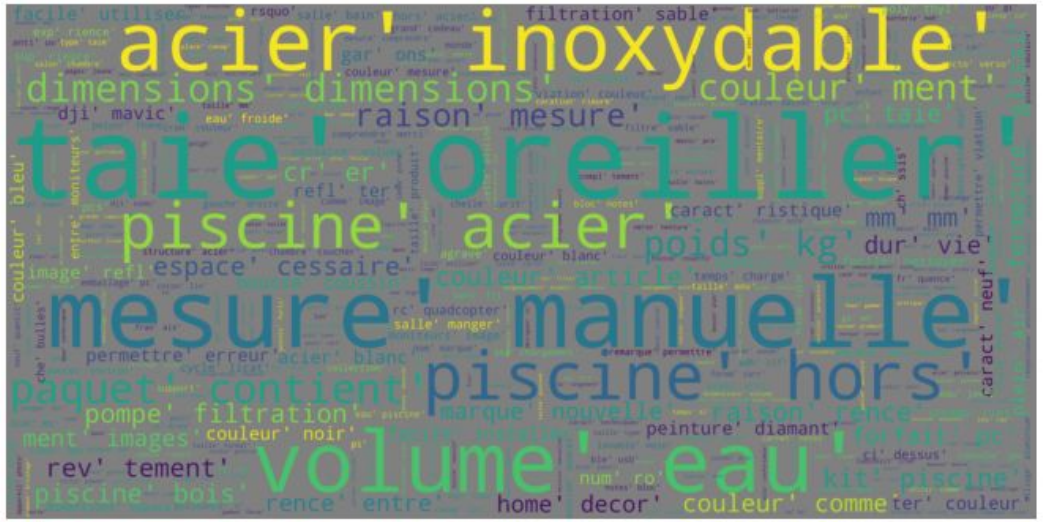


Imbalanced datas !

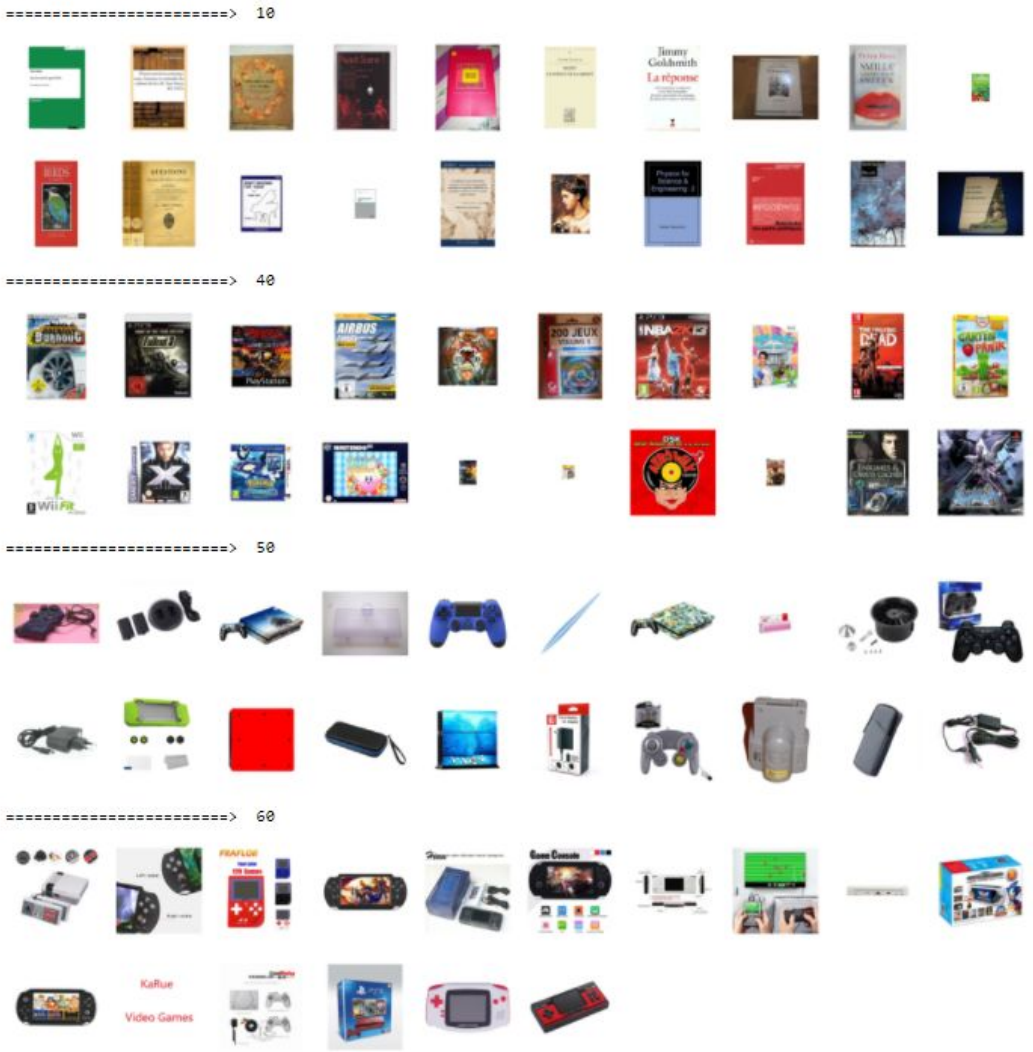


PRÉSENTATION DES DONNÉES

TEXTE : Fortes **disparités** dans la **fréquence** d'apparition des **mots** !

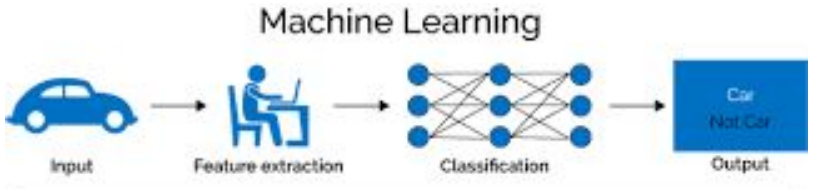


IMAGES : Échantillonnage du dataset



PRÉPARATION DES DONNÉES TEXTE

MACHINE LEARNING TEXTE

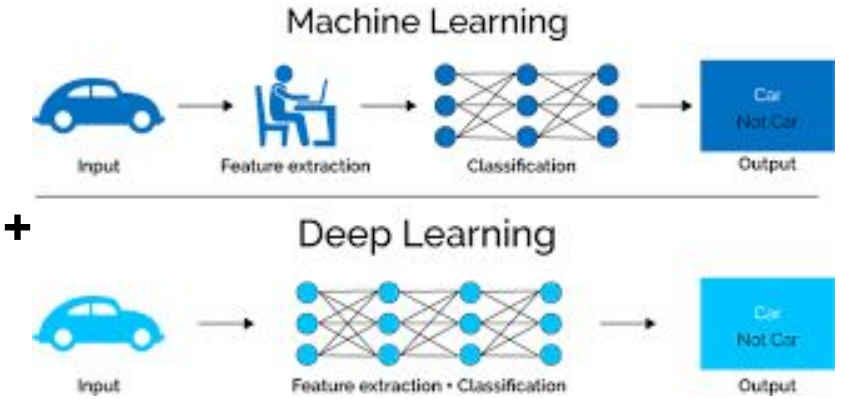


-> **Lemmatizer**, **Steeming**, **Stopword**

-> Vectorization **TfidVectorizer**

DEEP LEARNING TEXTE

-> **idem Partie Machine Learning**



-> **+ Tf Tokenizer avec Embedding**

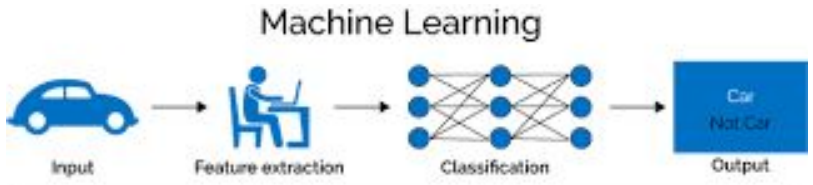
-> **+ Tf hub** universal-sentence-encoder

-> **+ Spacy Token** + Lem

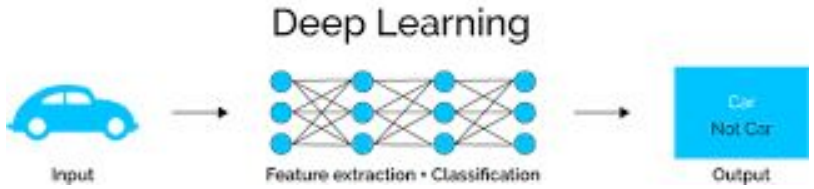
Keras Tokenizer 70000 mots max Phrases < 600 mots	Embed Multilingual-large	Spacy Token+Lem | SVC (<50000)

PRÉPARATION DES DONNÉES IMAGES

MACHINE LEARNING IMAGES - Classif non supervisée

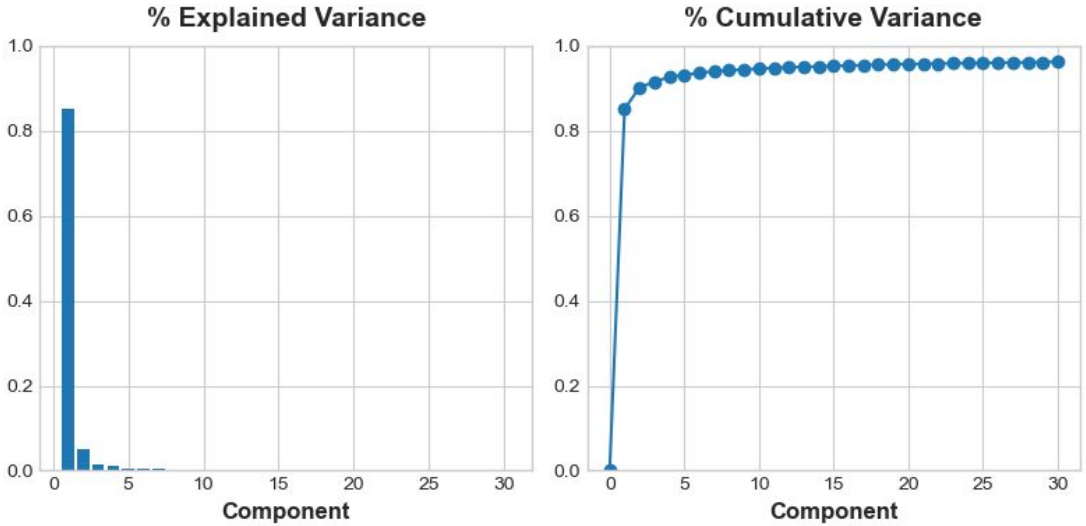


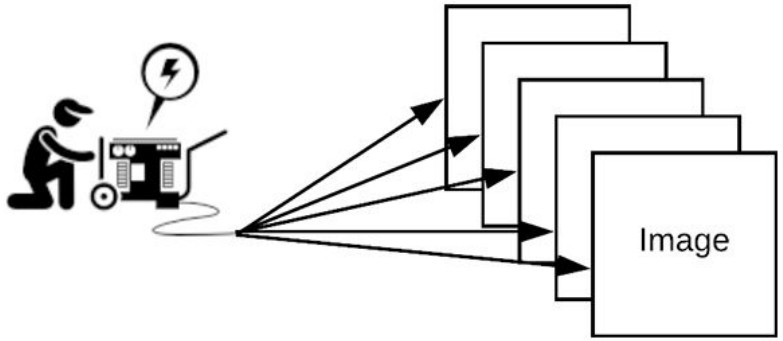
DEEP LEARNING IMAGES



-> **Keras/TF Images data generator**

-> **PCA**





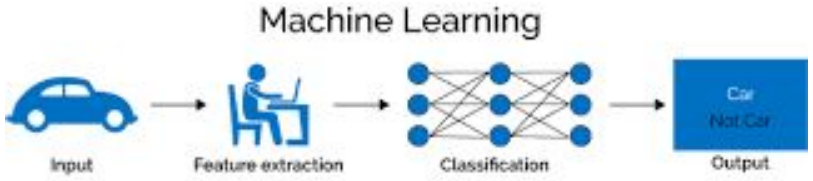
-> **Transfer Learning method**

-> **Clustering**

clusters = 3
silhouette score 0.61

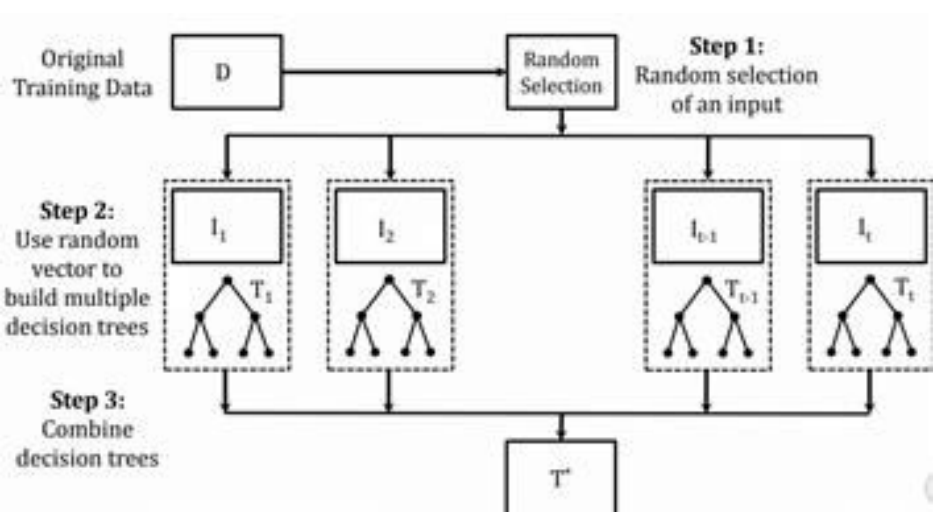
MODELES MACHINE LEARNING IMAGES

Screening x modèles Machine Learning - Classification supervisée



Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
ExtraTreesClassifier	0.40	0.31	None	0.37	34.78
RandomForestClassifier	0.40	0.31	None	0.37	55.69
LGBMClassifier	0.39	0.30	None	0.36	326.00
XGBClassifier	0.38	0.29	None	0.35	800.33
BaggingClassifier	0.33	0.27	None	0.32	233.55
SVC	0.34	0.25	None	0.30	1520.29
DecisionTreeClassifier	0.27	0.22	None	0.26	38.28
KNeighborsClassifier	0.27	0.22	None	0.25	579.98
LinearDiscriminantAnalysis	0.27	0.22	None	0.26	359.96
RidgeClassifier	0.27	0.21	None	0.24	5.29
RidgeClassifierCV	0.27	0.21	None	0.24	21.89
LogisticRegression	0.26	0.21	None	0.24	28.13

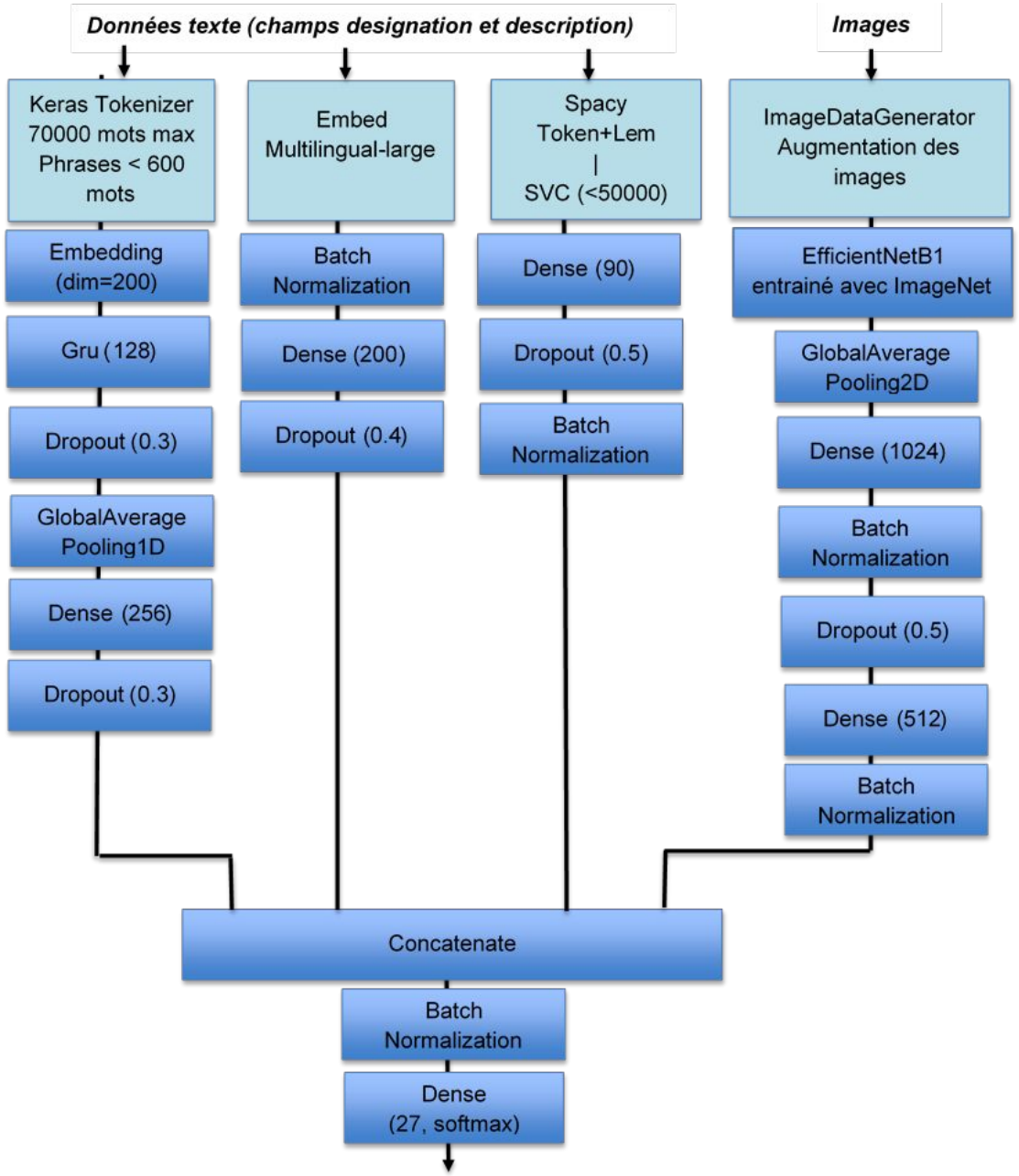
Extremely Randomized Trees



Pour positionnement comparatif vs (Deep) F1 score

Vgg16 : 0.26 , Xception : 0.51 , EffNetB5 : 0.66

MODÈLES CRÉÉS PAR CONCATÉNATION MODELES TEXTE & IMAGE



RESULTATS MEILLEURS MODELE

Scoring f1 weighted sur les données test de notre modèle final Concatenate: 0.863

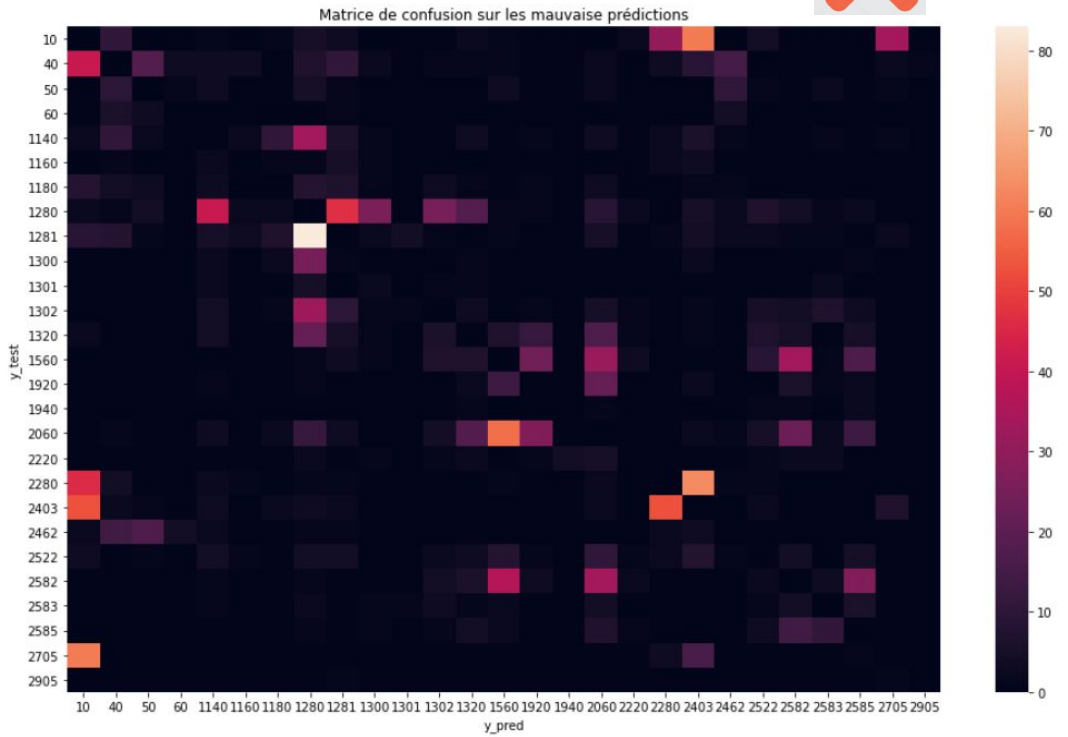
Analyse bonnes prédictions

	f1 weighted																											
Models	weighted	10	1140	1160	1180	1280	1281	1300	1301	1302	1320	1560	1920	1940	2060	2220	2280	2403	2462	2522	2582	2583	2585	2705	2905	40	50	60
EmbedRNN, OneHot, Multilingu, EfficientNetB1 **TEST 1**	**0,8776**	0,78	0,86	0,98	0,7	0,76	0,66	0,95	0,92	0,83	0,84	0,85	0,91	0,92	0,81	0,87	0,89	0,87	0,86	0,93	0,77	0,98	0,86	0,91	0,99	0,83	0,89	0,93
EmbedRNN, OneHot, Multilingu, EfficientNetB1 **TEST 2**	**0,878**	0,76	0,84	0,98	0,68	0,76	0,67	0,97	0,94	0,84	0,86	0,86	0,93	0,95	0,82	0,88	0,89	0,85	0,84	0,94	0,77	0,98	0,86	0,9	0,99	0,82	0,88	0,92
EmbedRNN, OneHot, Multilingu, EfficientNetB1 **TEST 3**	**0,8783**	0,72	0,84	0,98	0,72	0,77	0,67	0,96	0,94	0,86	0,87	0,86	0,93	0,97	0,83	0,88	0,88	0,84	0,85	0,94	0,78	0,98	0,88	0,88	0,99	0,81	0,87	0,93

Certaines classes compliquées à classer (score f1-weighted < 0.8) :

10 (livres d'occasion)
1180 (figurines et jeux de rôles)
1280 (jouets enfants)
1281 (jeux de société enfant)

Analyse mauvaises prédictions



Plusieurs classes + problématiques -> génèrent x erreurs dans les autres classes de thèmes

40 (jeux vidéos), **1280** (jouets enfants), **1320** (puériculture), , **1560** (mobilier général), **1920** (Linge de maison), **2060** (Décoration), **2582** (mobilier de jardin)

Bilan

Panel de compétences & intégration du top 10

Computer Vision

Natural Language Processing

Machine Learning

Deep Learning

Réduction de dimension

Classé 9ème

F1-weighted score : 0.86

F1-weighted score image : 0.66

F1-weighted scores texte : [0.81 , 0.85]

Etendu F1-weighted scores: [0.67 , 0.99]

Ouverture

Limites & perspectives

Images similaires

4 classes mal-prédites

Décalage formation – projet

Matériel (GPU)

Temporel

Nouveaux modèles pré-entraînés

(versions, données d'entraînements)

Amélioration rétrospective grâce à l'interprétabilité

Inclusion de modèle(s) entraîné(s) uniquement sur classes mal prédites

Techniques d'assemblage de modèles

Procédure d'hyperparamétrisation et de paramétrisation

Rééquilibrage des classes

DataScientest • com

Challenge data

Rakuten Institute of Technology

TEAM :

FATOUMATA BARRY

ERIC MARCHAND

EDGAR HIDALGO LOPEZ

EMMANUEL BONNET

Merci de votre écoute et du soutien !

REGARD CRITIQUE & PERSPECTIVES

DONNÉES DU DATASET CHALLENGE RAKUTEN 2020

-> Images à fortes similitudes visuelles dans différentes classes, même à l'œil humain !

-> Doute sur la qualité et/ou uniformité des données textes et images mises à disposition pour les 4 classes les + mals scorées

IMPLÉMENTATION DES MODÈLES DE DEEP

-> La formation en deep learning était assez loin dans la formation => temps limité pour assimiler et appliquer ces connaissances pratiques au projet

-> Limites en disponibilité machine et pas de GPU (limites dans le nombre d'époques d'entraînement)

PERSPECTIVES

-> Tentatives d'améliorations modèle par augmentation des échantillons des classes mal prédites, ceci dans le but d'améliorer leurs scores & tests complémentaires sur la base du meilleur modèle CONCATENATE

-> Autres approches possible pour ce type de projet : tester la faisabilité d'implémentation de transformeurs viT dans la partie Computer Vision (évaluation si gain versus CNN)

REGARD CRITIQUE & PERSPECTIVES

DONNÉES DU DATASET CHALLENGE RAKUTEN 2020

-> Images à fortes similitudes visuelles dans différentes classes, même à l'œil humain !

-> Doute sur la qualité et/ou uniformité des données textes et images mises à disposition pour les 4 classes les + mals scorées

IMPLÉMENTATION DES MODÈLES DE DEEP

-> La formation en deep learning était assez loin dans la formation => temps limité pour assimiler et appliquer ces connaissances pratiques au projet
-> Limites en disponibilité machine et pas de GPU (limites dans le nombre d'époques d'entraînement)

PERSPECTIVES

-> Tentatives d'améliorations modèle par augmentation des échantillons des classes mal prédites, ceci dans le but d'améliorer leurs scores & tests complémentaires sur la base du meilleur modèle CONCATENATE
→ Modèles pré-entraînés non testés
→ Modèles entraînés sur classes mal prédites puis joint à la concaténation
→ Comprendre les features sur lequel les données images sont entraînés et paramétrisés
→ Comprendre les features sur lequel les données textes …
-> Autres approches possible pour ce type de projet : tester la faisabilité d'implémentation de transformeurs viT dans la partie Computer Vision (évaluation si gain versus CNN)

CONCLUSION

Bilan technique :

-> Mise en oeuvre d'une partie significative des modules/connaissances enseignées dans le master Datascientest

-> Résultat final de scoring dépassant le benchmark, nous positionnant dans le top 10

Bilan projet :

-> Très bonne collaboration et entente entre les différents membres de l'équipe projet

Ranking	Date	User(s)	Public score
1	Dec. 10, 2020, 9:35 p.m.	elieS	0.9037
2	April 7, 2020, 1:05 p.m.	Shiro	0.8957
3	Nov. 15, 2020, 10:03 a.m.	kobayashi_shu	0.8940
4	Oct. 18, 2020, 9:51 p.m.	julienC	0.8903
5	March 8, 2020, 3:38 p.m.	Binouze & BlueDrey & JojoFlower	0.8852
6	July 16, 2021, 6:30 p.m.	tbierlaire & meriem_e & vincent__	0.8841
7	Jan. 22, 2021, 3:46 p.m.	NadirEM	0.8714
8	Jan. 25, 2021, 2:08 a.m.	EmmanuelJunior.WafoWembe	0.8708
9	Aug. 9, 2021, 11:01 a.m.	FEEEScientest	0.8628

https://docs.google.com/presentation/d/1zKsGADbbwGEzlVldO6dcjASplksndgI3Wf2Wx-7p0OM/edit#slide=id.ge8861749c3_0_3