



#### **INSTAGRAM**

Editeurs et Libraires : Prédiction du nombre de likes

Janvier 2020

#ironhack #machinelearning #instagram #seaborn #webscraping #xgboost #prediction #datacleaning #bookstagram #editeur #book #bookaddict #livrestagram #livre #editions #booklovers #bookslover #lovebooks #NLP #lovereading #librairies #librairies #emoji











# **Sommaire**

# Contexte

Le secteur du livre vs l'univers numérique

La base de données Web Scraping & Cleaning

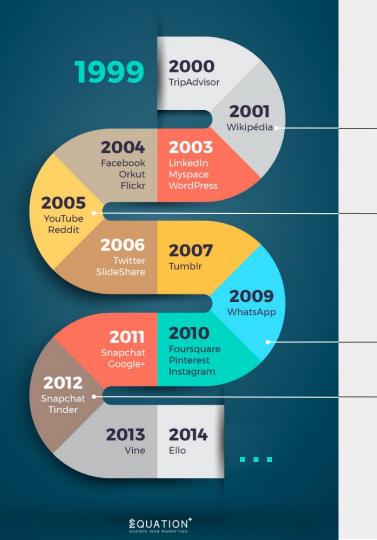
Sélection des variables

Machine Learning
Cross-validation & Grid-search

Feature engineering NLP & Emoji

Conclusion





### **Contexte**

Etude sur les communautés

virtuelles

# Marketing et communication sur Internet deviennent une évidence :

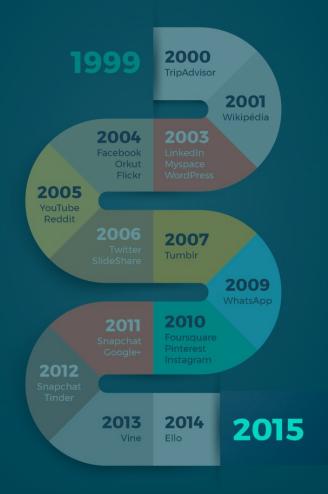
apparition de sites d'éditeur, assurer la présence des auteurs sur les médias en ligne ...

## Prescription 2.0:

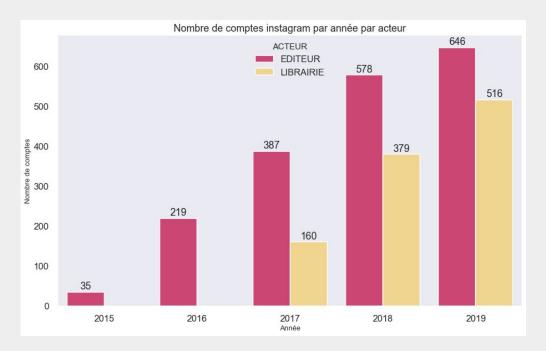
Avènement des réseaux sociaux et bientôt des bookTubers / bookstagramers

## Crainte de disruption :

La place d'Amazon grandit et le livre numérique fait de plus en plus parler de lui (Kindle en 2007)



# Reach & Read un projet de base de données

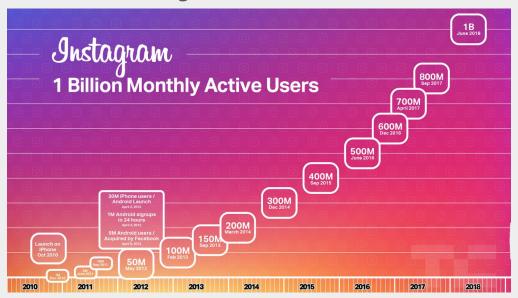


+3226%

Editeurs / 2015-2019 : +1746% Libraires / 2017-2019 : +223%



### Instagram en chiffres (2019)



#### Monde:

1 milliard d'utilisateurs mensuel dans la monde 1 utilisateur sur 2 utilisent l'onglet Explorer France :

17 millions d'utilisateurs en France 54% sont des femmes **4,2 milliards de likes sont postés chaque jour**  

# Webscraping

1162 comptes scrappés

15 Octobre-15 Novembre

Selenium, BeautifulSoup

ntry\_data":{"ProfilePage":{("logging\_page\_id":"profilePage\_2051081056","show\_suggested\_profiles":false,"show\_follow\_dialog":false,"grapl ":/"user":{"biography":"\ud83d\udcda Art de vivre, bien-\u00eatre, sant\u00e9, \u00e9ducation, fiction, sport, loisirs... Bienvenue dans univers des \u000c9ditions Marabout ! \ud83c\udf1e\ud83c\udf74\ud83e\udd8\ud83c\udffb\u200d\u2640\ufe0f\ud83c\udfa8 u26bd\ufe0f\ud83d\udeb4\ud83c\udffc\u200d\u2640\ufe0f","blocked\_by\_viewer":false,"country\_block":false,"external\_url bout.com/","external\_url\_linkshimmed":"https://l.instagram.com/?u=http%3A%2F%2Fwww.marabout.com%2F\u0026e=ATOdLaT AvNmDSwn\_odilRcfUvbRKoAaYG9I7IZcmRsE2Y1ox7ccIDKum399a19Fuhbv85K4"." uanarbelaezchef auteur du \u00ab\u00a0Cuisinez, partagez \u00bb\n\nSi vous avez aim\u00e9 son histoire et que vous souhaitez gagne in livre d\u00e9dicac\u00e9, il vous suffit de participer :\u1\ufe0f\u20e3 Abonnez-vous \u00e0 notre compte @edition\_marabout u00e0 celui de @iuanarbelaezchef \n2\ufe0f\u20e3 Commentez en taguant 2 de vos ami.e.s et dites nous si yous aussi yous vous \u00eate \u00e9j\u00e0 d\u00e9guis\u00e9 en p\u00e8re No\u00ebl ?\n.\nA vous de jouer! Vous avez jusqu\u2019au 25 d\u00e9cembr nt":{"count":63},"comments\_disabled":false,"taken ("height":750, width":750), "display\_url":"https://scontent-cdt1-1.cdninstagram.com/v/t51.2885-15/e35/7541271 9250805 4087998578897598369 n.jpg? nc ht=scontent-cdt1-1.cdninstagram.com\u0026 nc cat=109\u0026 nc ohc=AiMwAd5L 8AX-NvRzM\u0026oh=47152e47eb899aacc33c6e77bd218b6d\u0026oe=5E07AF7B","edge\_liked\_by":{"count": ocation" null, "gating info":null, "fact check overall rating":null, "fact check information":null, " DMT/739Klm0DcAV1wwyO4qtcxeVGzxllKjOAxxx2qyOlMuyVjYjg47VJoc5cXTSqFLFu+T/LoPx/SqVBpK02OVvr edition marabout"), "thumbnail src"; "https://scontent-cdt1-1.cdninstagram.com/v/t51,2885-15. t-cdt1-1.cdninstagram.com/v/t51.2885-15/e35/s150x150/75412712 250534279250805 4087998578897598369





# **Cleaning**

DataSet initial: ~322K publications

2 parties exploitées:

infos des comptes
(ID, Nom, nb de followers)
infos des publications
(ID, text, nb de like, format de la publication,...)
ajout de colonnes :

- nb de mots, de #, traitement du datetime...

	page_id	page_username	date_time	date	code_jour_semaine	day	month	year	
0	4741863243	librairielestempsmodernes	2017-03-09 22:16:32	2017-03-09	4	9	3	2017	;
1	4741863243	librairielestempsmodernes	2017-03-01 20:16:30	2017-03-01	3	1	3	2017	:
2	3655964171	librairielibellule	2019-10-11 16:56:30	2019-10-11	5	11	10	2019	

suppression des colonnes de localisation : Drop nan & duplicates

DataSet final: ~23K Publications and 29 colonnes



# **Analyse du** dataset

Quelques tendances:

**75%** 

des comptes



ont moins de 3000 followers



sont des photos/images

des

Aa

ont jusqu'à 64 mots/expressions.

publications

ont jusqu'à 11 #



ont moins de 93 likes

Pic d'activité (publication)

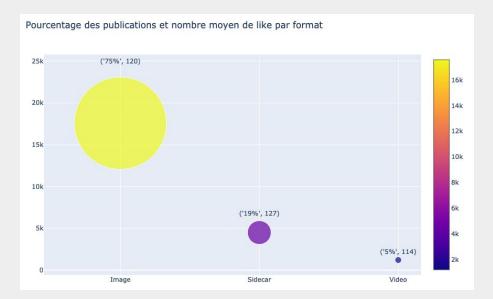


en semaine, plus particulièrement en 2ème partie de semaine

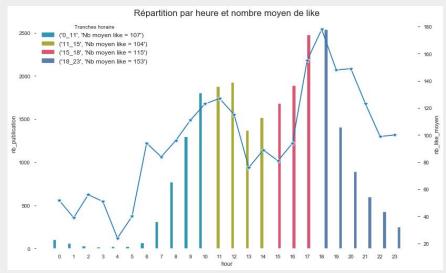


vers 14h/15h avec Identification de 4 tranches horaires : avant 11h, entre 11h et 15h, entre 15h et 18h et après 18h.

### **Datavisualisation**

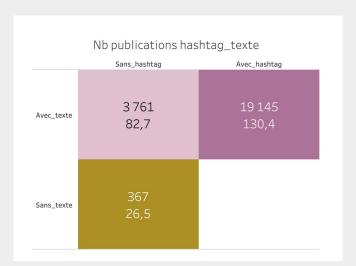


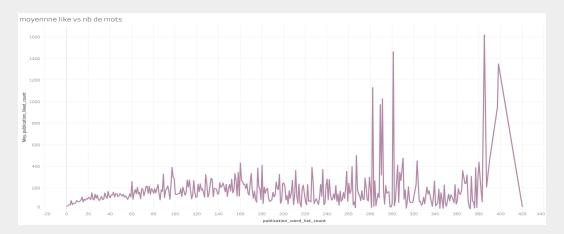


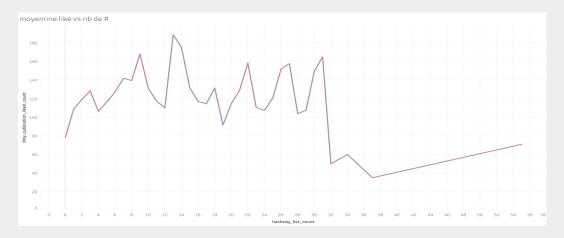


# Datavisualisation

Pour plus de likes :
Une publication avec du texte et des hashtags.
Un texte de plus de 80 mots
mais pas de surenchère de hashtags







# Matrice de corrélation



### Coefficients de corrélation de la target choisie "publication\_liked\_count" (nombre de like pour une publication).

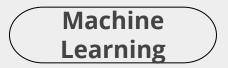
Variable	Coefficient			
Nombre de fans de la page	0.496264			
Nombre de mots	0.104775			
Largeur de la publication	0.048080			
Nombre de hashtags	0.041036			
Hauteur de la publication	0.038913			

# 16 Features retenues

- » Activité de l'utilisateur
- » # de followers
- » Jour de la semaine
- » # de mots
- » # de hashtag
- » Type de la publication :
  - image
  - vidéo
  - diaporama



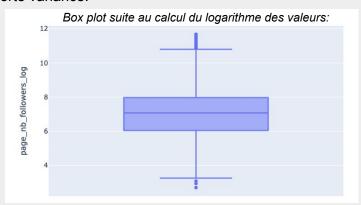
- » Hauteur de la publication
- » Heure
- » Tranche horaire :
  - **\*** 0-11
  - **\*** 11-15
  - 15-18
  - **\*** 18-23
- » # d'emoji
- » Polarité
- » Subjectivité



# Préparation du dataset

- Transformation des variables catégorielles avec la méthode get.dummies()
- Split: 80% train / 20% test
- Cross-validation

Pour la colonne "page\_nb\_followers" afin de faciliter la phase d'apprentissage du modèle, je choisis de calculer le logarithme des valeurs car cette variable a une trop forte variance.



#### Cross-validation avec 9 modèles de prédiction

	Model	Mean	Min	Max	Std	Mad
3	Random Forest	77.099306	0.742490	0.783751	0.012489	0.009794
6	Bagging Regressor	77.090864	0.748421	0.787427	0.011838	0.009658
4	Gradient Boosting	72.769696	0.697914	0.750937	0.014133	0.011223
8	XGB Regressor	72.729981	0.698563	0.750518	0.014065	0.011495
2	Ridge Regression	68.406413	0.655149	0.708757	0.013032	0.009310
7	Linear Regression	68.406391	0.655142	0.708756	0.013034	0.009312
5	AdaBoost	59.668860	0.560612	0.629102	0.021804	0.019576
1	Decision Tree	59.207709	0.534645	0.635184	0.033377	0.028008
0	K-Nearest Neighbor	34.554129	0.291033	0.396742	0.031989	0.025269

### Grid Search + cross validation pour le XGBoost, Gradient Boosting, Bagging Regressor et Random Forest

	Model	Mean	Min	Max	Std	Mad
3	XGB Regressor_grid	78.181934	0.775991	0.788672	0.004708	0.004202
2	RF Regressor_grid	76.264320	0.753880	0.772806	0.008211	0.007914
0	Gradient Boosting	73.398347	0.724960	0.741862	0.005727	0.004692
1	Bagging Regressor	68.126199	0.591730	0.727472	0.046936	0.035813

# Machine Learning

# Je retiens donc le modèle xgboost



#### Création d'une colonne "Tranches\_horaire"

Teste de l'éventualité qu'un apprentissage suivant la tranche horaire soit plus pertinente que simplement suivant l'heure.



```
R2_score_train 87.0 R2_score_test 78.0
```

rmse\_score\_train 0.43204563352538794 rmse score test 0.566950640629682

real\_rmse\_score\_train 144.2308206349386 real rmse score test 189.06892071329463



```
R2_score_train 87.0 R2_score_test 80.0
```

rmse\_score\_train 0.4318626797438654 rmse\_score\_test 0.5526246942304651

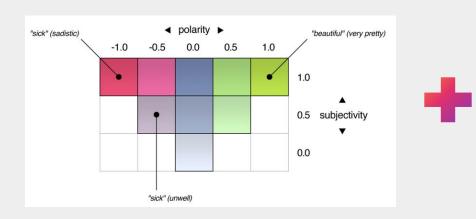
real\_rmse\_score\_train 159.01949660828788 real rmse score test 162.37744761101908

**Résultat**: Avec ce changement de variable, le R2 score gagne 2 point.

# Feature engineering

# Polarité / Subjectivité

#### librairie TextBlob



# Nombre d'emoji

#### librairie Demoji



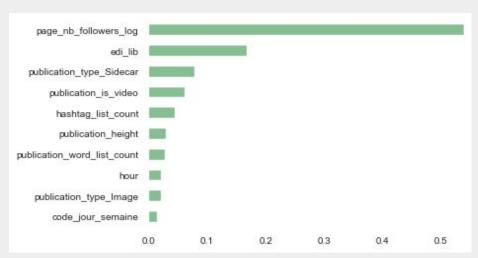
**Résultat**: Un R2 score légèrement meilleure qu'avec l'apprentissage initial.

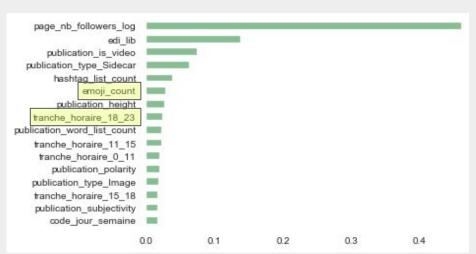
# Features importance





# Features importance après feature engineering

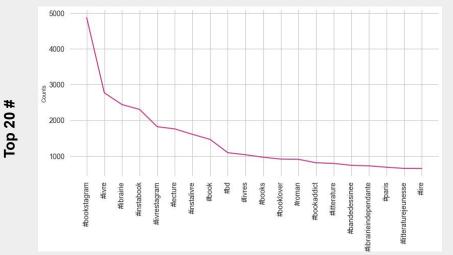


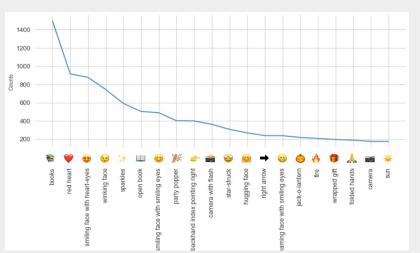


### **Datavisualisation**

#### Nuage de mots :







Top 20 emoji



# **Conclusion**

#### Dataset

Avoir un historique plus long afin de tenir compte de toutes les saisonnalités propre au secteur du livre.

#### **Variables**

Ajouter les variables relative aux visuels : analyse des images identifications des publications citant des influenceurs, auteurs etc..

#### **Machine Learning**

Réaliser un apprentissage distinct par type d'acteur Sélection d'hyper-paramètres plus poussée

#### NLP

Analyse et identification de typologies de publications

### Bonus : Testez votre publication !





# Lecture

L'utilisation des réseaux sociaux modifie le milieu du livre : Booksquad 🖱

La communication numérique : un enjeu pour l'attractivité des librairies : Actualitte

Vente en ligne et réseaux sociaux : libraires, vos (livres) papiers ! : Actualitte

Prix, rencontres, fidélisation : penser le client autour de la librairie : Actualitte 🖱

Instagram, les chiffres essentiels en 2019 en France et dans le monde : Digimind 🖔

Guide des Dimensions des Images sur Instagram : Webmarketing-conseil 🖱





## **MERCI!**

#ironhack #machinelearning #instagram #seaborn #webscraping #xgboost #prediction #datacleaning #bookstagram #editeur #book #bookaddict #livrestagram #livre #editions #booklovers #bookslover #lovebooks #NLP #lovereading #libraires #librairies #emoji







