



Sandrine Henry



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 #lovereadng
#libraires #librairies #emoji





Sommaire

Contexte

Le secteur du livre vs l'univers numérique

La base de données

Web Scraping & Cleaning

Sélection des variables

Machine Learning

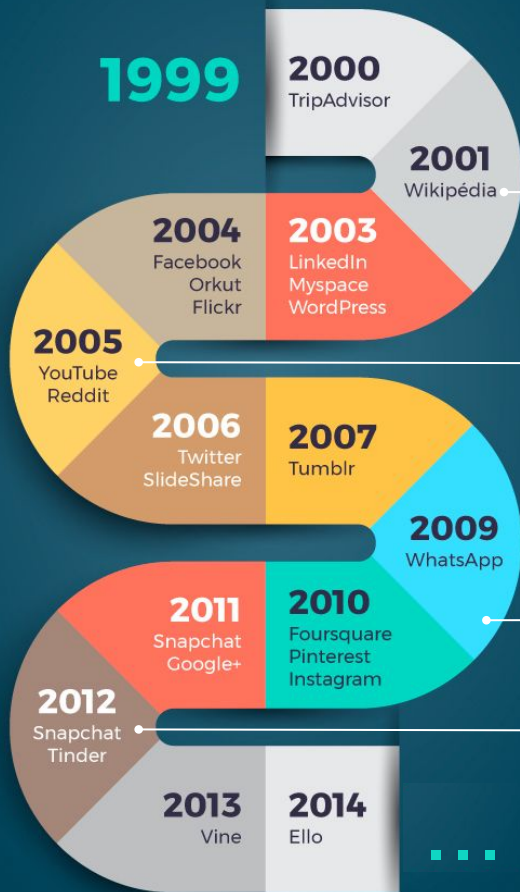
Cross-validation & Grid-search

Feature engineering


NLP & Emoji

Conclusion





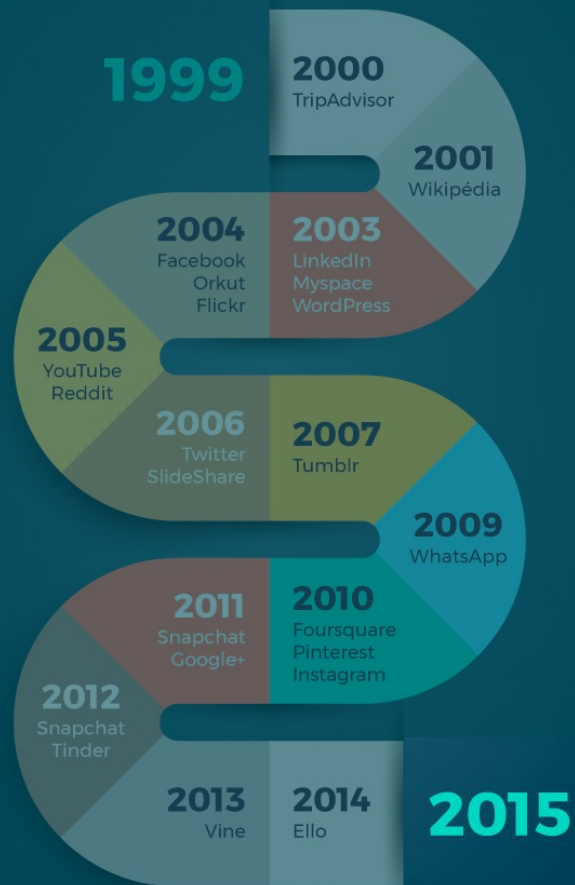
Contexte

En licence : 
 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 ...

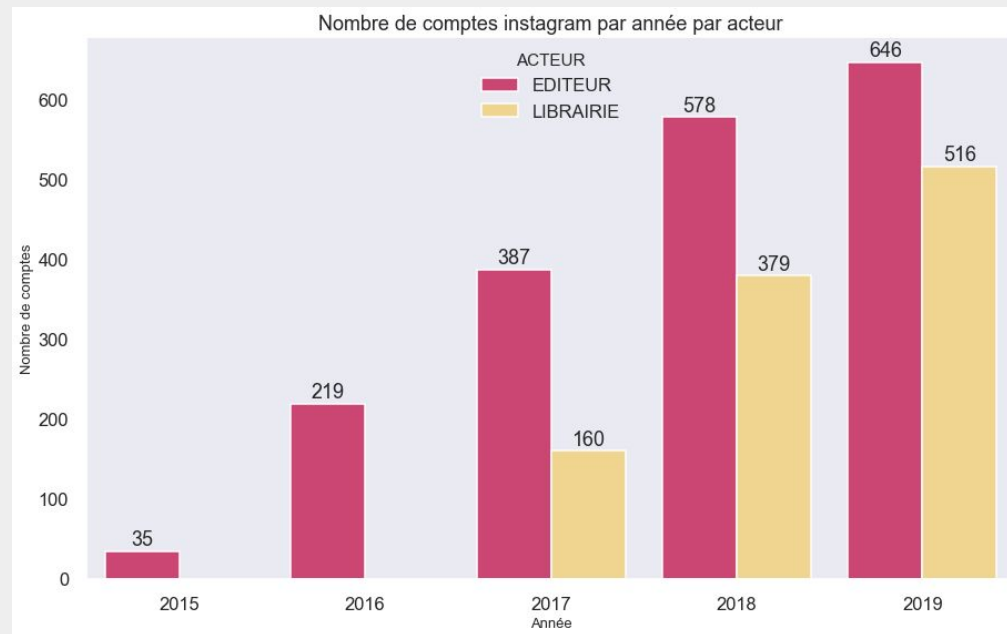
Crainte de disruption :
 La place d'Amazon grandit et le livre numérique fait de plus en plus parler de lui (Kindle en 2007)

Prescription 2.0 :
 Avènement des réseaux sociaux et bientôt des bookTubers / bookstagramers



Reach & Read

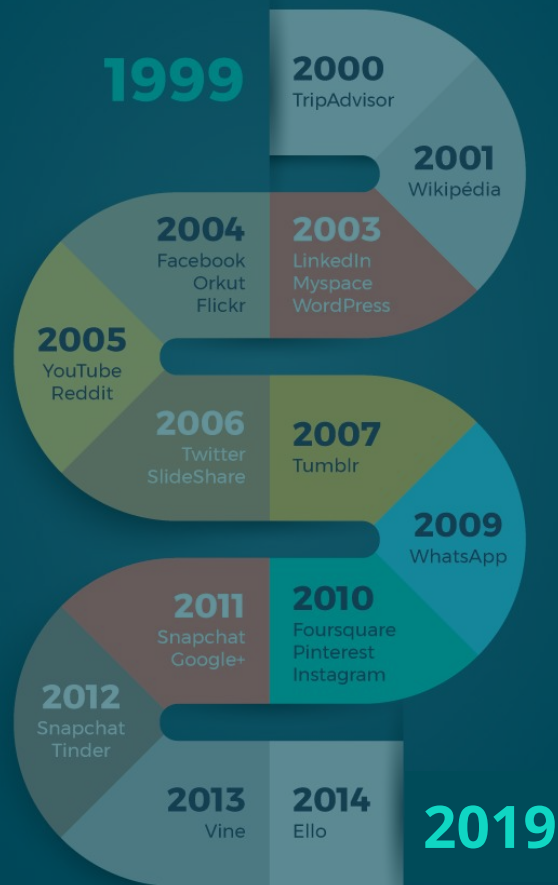
un projet de base de données



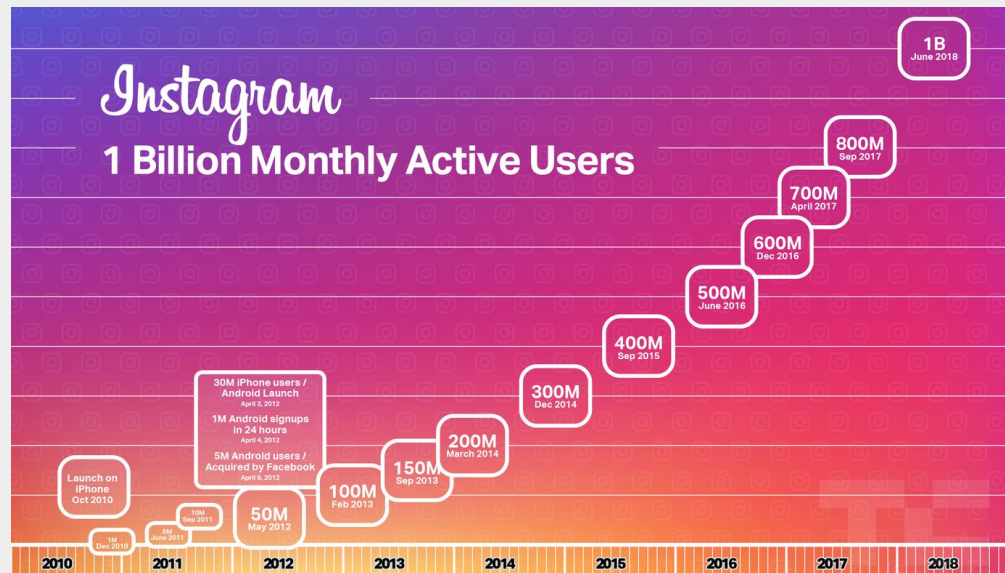
+3226%

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

2020





Analyse du dataset

Quelques tendances :
75 %

des comptes



ont moins de 3000 followers



sont des photos/images

des
publications



ont jusqu'à 64 mots/expressions.



ont jusqu'à 11 #



ont moins de 93 likes

Pic d'activité (publication)



en semaine, plus particulièrement
en 2ème partie de semaine

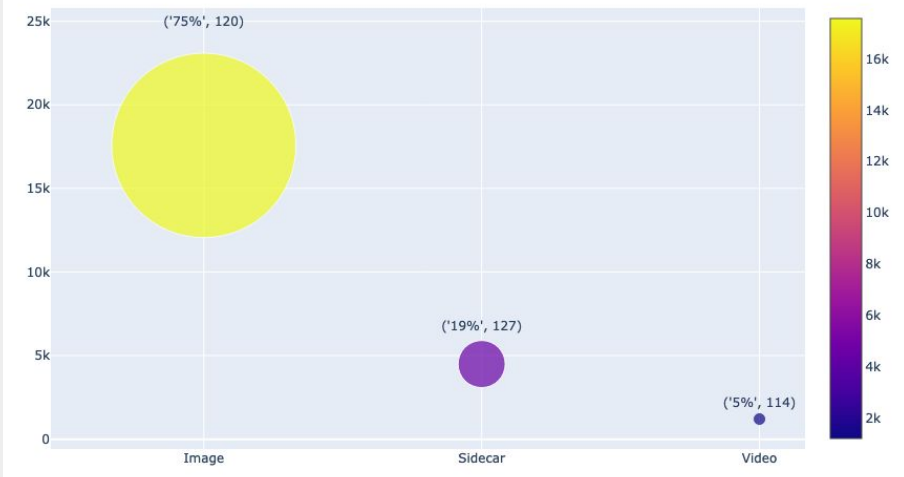


vers 14h/15h avec

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

Datavisualisation

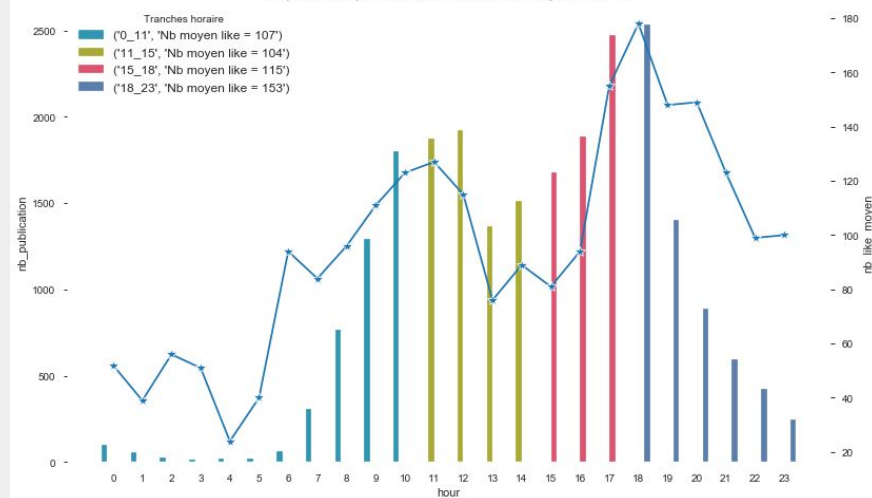
Pourcentage des publications et nombre moyen de like par format



Répartition par jour et nombre moyen de likes



Répartition par heure et nombre moyen de like



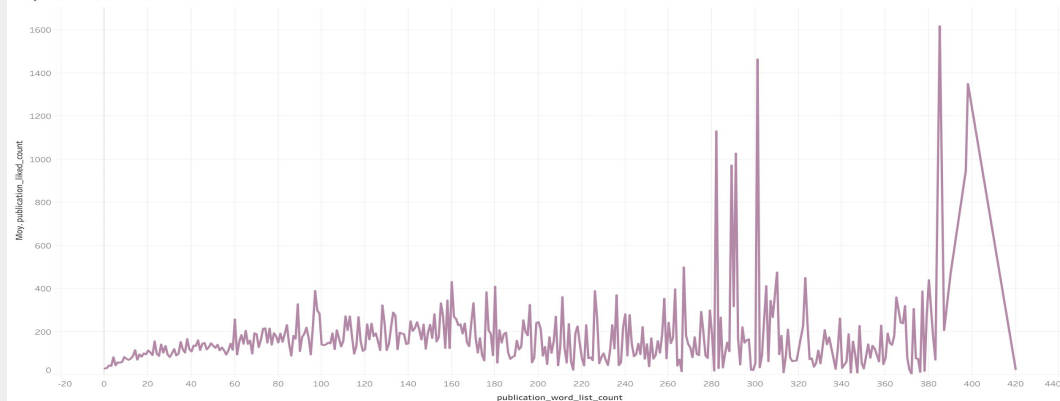
Datavisualisation

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

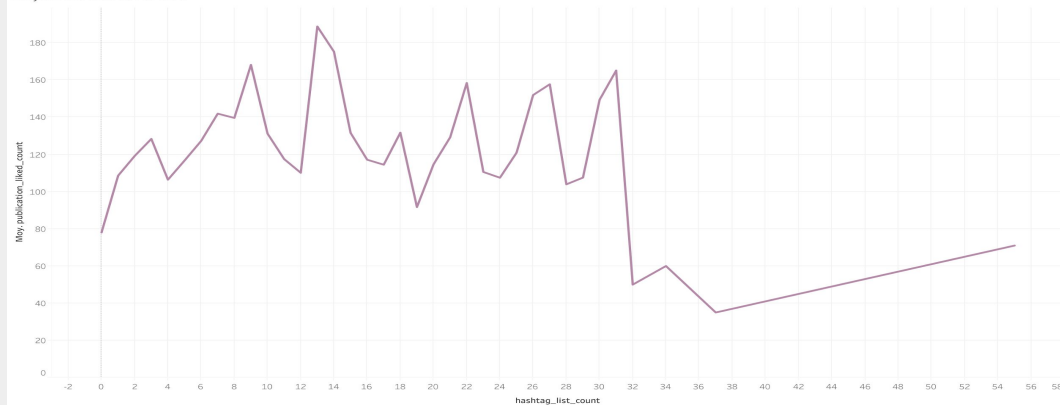
Nb publications hashtag_texte

	Sans_hashtag	Avec_hashtag
Avec_texte	3 761 82,7	19 145 130,4
Sans_texte	367 26,5	

moyenne like vs nb de mots



moyenne like vs nb de #



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

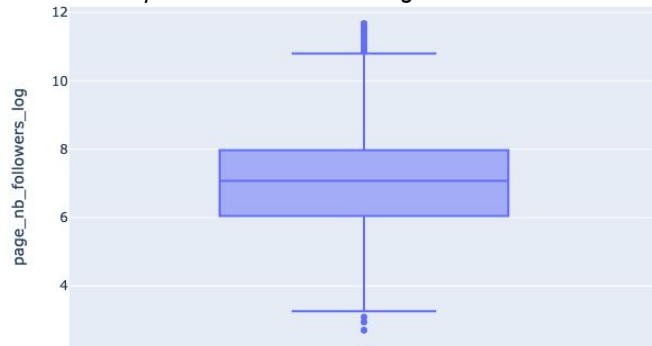
Machine Learning

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.

Box plot suite au calcul du logarithme des valeurs:



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

```
xgb_reg = XGBRegressor(objective = 'reg:squarederror', learning_rate=0.15,  
                       max_depth= 7,min_child_weight= 1,n_estimators = 150)  
  
# fitting the model  
xgb_reg.fit(X_train, y_train)  
  
# predict the response  
y_test_predict_xgb_reg = xgb_reg.predict(X_test)
```




```
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
```

Feature engineering

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.

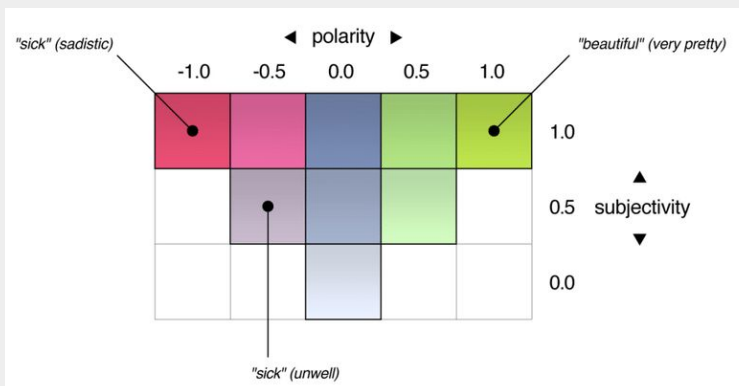


```
R2_score_train 87.0  
R2_score_test 80.0  
  
rmse_score_train 0.4318626797438654  
rmse_score_test 0.5526246942304651  
  
real_rmse_score_train 159.01949660828788  
real_rmse_score_test 162.37744761101908
```

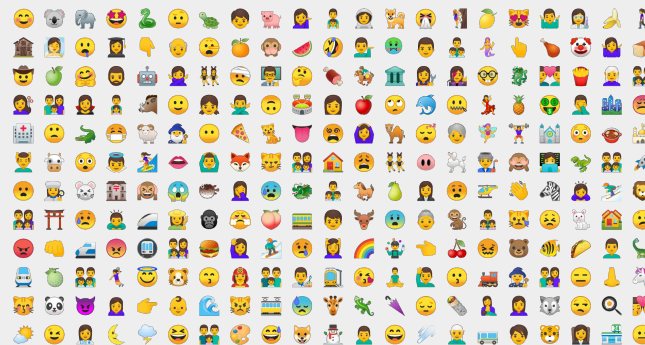
Résultat : Avec ce changement de variable, le R2 score gagne 2 point.

Feature engineering

Polarité / Subjectivité
bibliothèque TextBlob



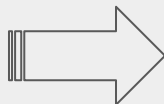
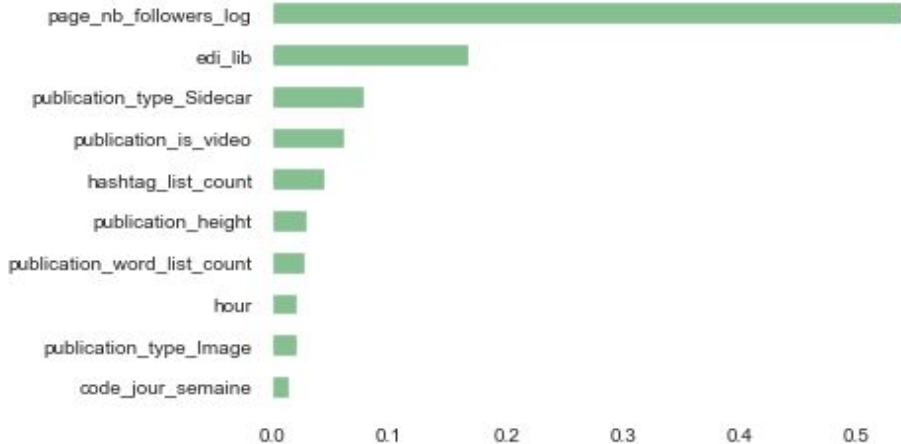
Nombre d'emoji
bibliothèque Demoji



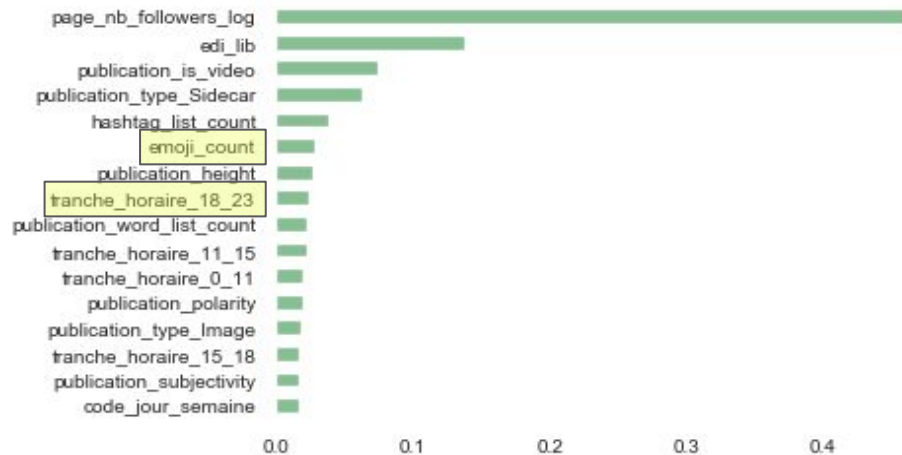
Résultat : Un R2 score légèrement meilleure qu'avec l'apprentissage initial.

Features importance

Features importance dataset d'origine



Features importance après feature engineering





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](#) 🖱



Sandrine Henry

...

MERCI !

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

