Python for data analysis

Facebook Comment Volume Dataset Data Set

Agathe Delas

Léopold Duverger

https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset

Origine du dataset

Notre dataset est https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset

Le dataset a été collecté par Kamaljot Singh en 2016

Son papier de recherche sur ce sujet: https://ijssst.info/Vol-16/No-5/paper16.pdf?fbclid=IwAR0UemwaW4tR7ROxxkVtPwYHpz5EIRCEYPI3OKZOSUGxW2kjZx52P2BsbxE

Contexte de l’étude

Facebook est fondé en 2004 par Mark Zuckerberg. Au 30 septembre dernier, Facebook comptabilisait 2,74 milliards d’utilisateurs actifs par mois et pas moins d’1,82 milliard d’utilisateurs actifs journaliers.

Face à ce phénomène récent impactant un grand nombre d’usagers, Il est intéressant d'étudier le comportement dynamique des utilisateurs vis-à-vis de ces services.

Les commentaires sur les sites permettent de gagner en visibilité, ils ont donc un impact sur l’économie. De plus, de nombreuses études concernent l’addiction aux réseaux sociaux et notamment une “addiction au like”. Cette étude peut donc permettre à différents acteurs de comprendre les comportements en ligne.

Objectif

Ce papier scientifique est un travail préliminaire pour étudier et modéliser l'activité des utilisateurs.

L’auteur de ce dataset a ciblé le service de réseautage social le plus actif, "Facebook", et surtout les "pages Facebook", pour les analyser.

Il s'agit ici de prédire le nombre de commentaires qu'un post devrait recevoir dans les prochaines heures

https://ijssst.info/Vol-16/No-5/paper16.pdf?fbclid=IwAR0UemwaW4tR7ROxxkVtPwYHpz5EIRCEYPI3OKZOSUGxW2kjZx52P2BsbxE

Le dataset

Le dataset contient 54 variables dont:

- 4 variables sur la page: décrivant l'intérêt de l’utilisateur pour cette page
- 25 variables dérivées: ce sont des calculs de minimum, maximum, moyenne, médiane et écart type des variables essentielles
- 5 variables essentielles: décrivent le pattern des commentaires sur un poste ou sur la page (C1 à C5)
- 5 autres variables: décrivant le post (sa taille, son nombre de partage…)
- 14 Variables du jour de la semaine: décrivant le jour où la publication est faite et celle de la publication à prévoir

Et la variable cible qui correspond donc au nombre de commentaires pour un post donné

Variables en détail: La page

Page_popularity	Définit la popularité ou le soutien à la source du document
Page_checkins	Décrit le nombre de personnes qui ont visité cet endroit jusqu'à présent. Cette caractéristique est uniquement associée aux lieux
Page_talking_about	Définit l'intérêt quotidien des individus envers la source du document/poste. Les personnes qui reviennent effectivement à la page, après avoir aimé la page. Cela inclut les activités telles que les commentaires, les goûts pour un poste, les partages, etc. des visiteurs de la page.
Page_category	Définit la catégorie de la source du document, par exemple : lieu, institution, marque, etc.†

Variables en détail: Les variables essentielles

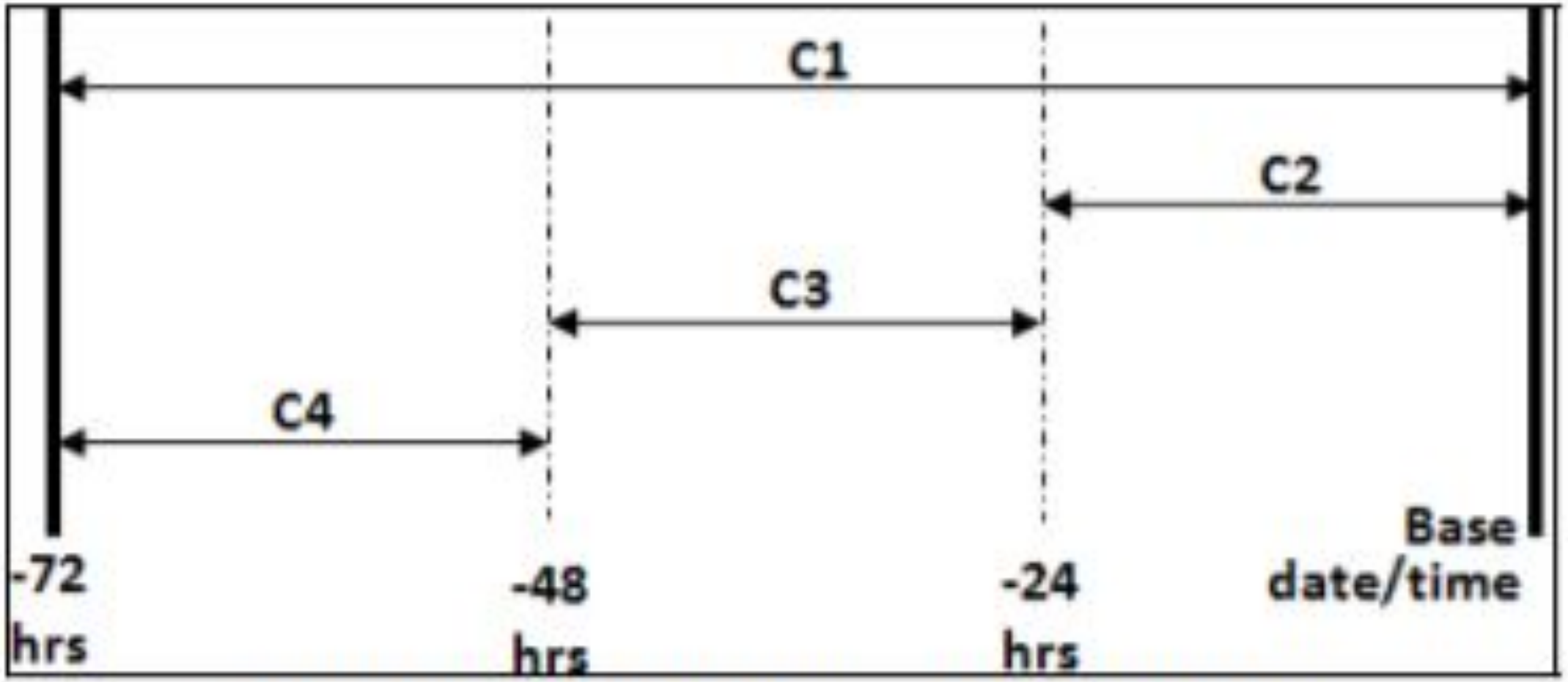


Figure 2. Demonstrating the essential feature details.

Variables crées

7 week_days in 1 column: Correspond au jour de publication, cette information était stockée dans les 7 variables de 39 à 46. Ce chiffre va donc de 1 à 7

1 étant lundi

Datetime_day in 1 column: Correspond au jour et l'heure du relevé du nombre de commentaires, cette information était stockée dans les 7 variables de 46 à 53.

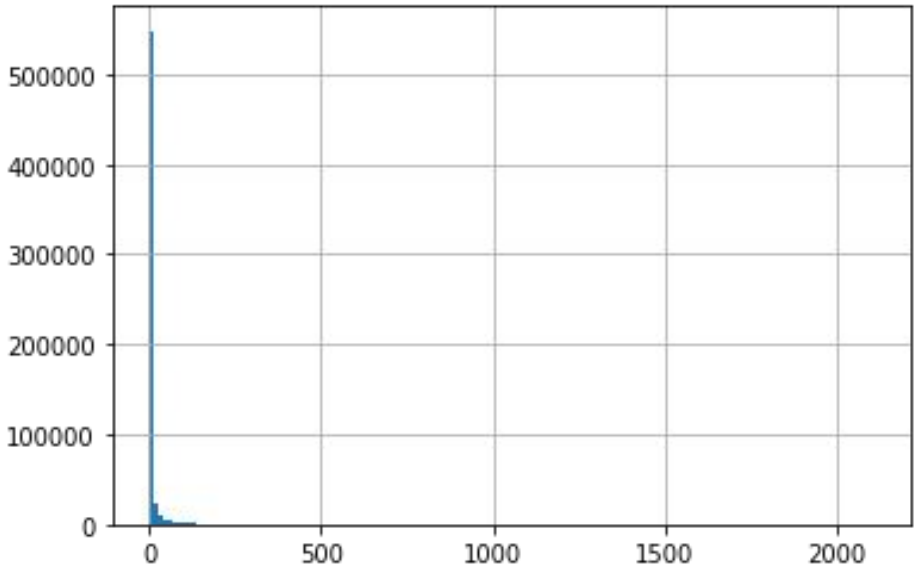
Résultats

En terme de précision et de viabilité d'industrialisation, nos résultats sont loin d'être excellents, nos meilleurs modèles font toujours beaucoup d'erreurs quand on sait que le nombre de commentaires du 3ème quartile est de 3, et que 65% des lignes avaient 0 commentaires. Mais nous avons vu une nette amélioration quand nous avons ignoré les lignes avec un nombre de commentaires > 100 (1.7 % des données). Nous sommes passés d'une erreur moyenne absolue de 20 à 7.

Pour nous cela témoigne de l'impact de la distribution très peu homogène des données. Ce qui rendait d'ailleurs certaines visualisations comme la distribution de l'erreur impossible à cause des écarts.

Mais le problème initial était de savoir s'il est possible d'estimer la quantité de commentaire à partir des caractéristiques d'une publication Facebook.

Et pour nous, nos résultats montrent que c'est un objectif qui peut être atteignable avec surement plus de feature engineering, et peut-être du fine-tuning de models avec une cross-validation peut-être, que nous n'avons pas pu réaliser par manque de temps et ressources de calculs.

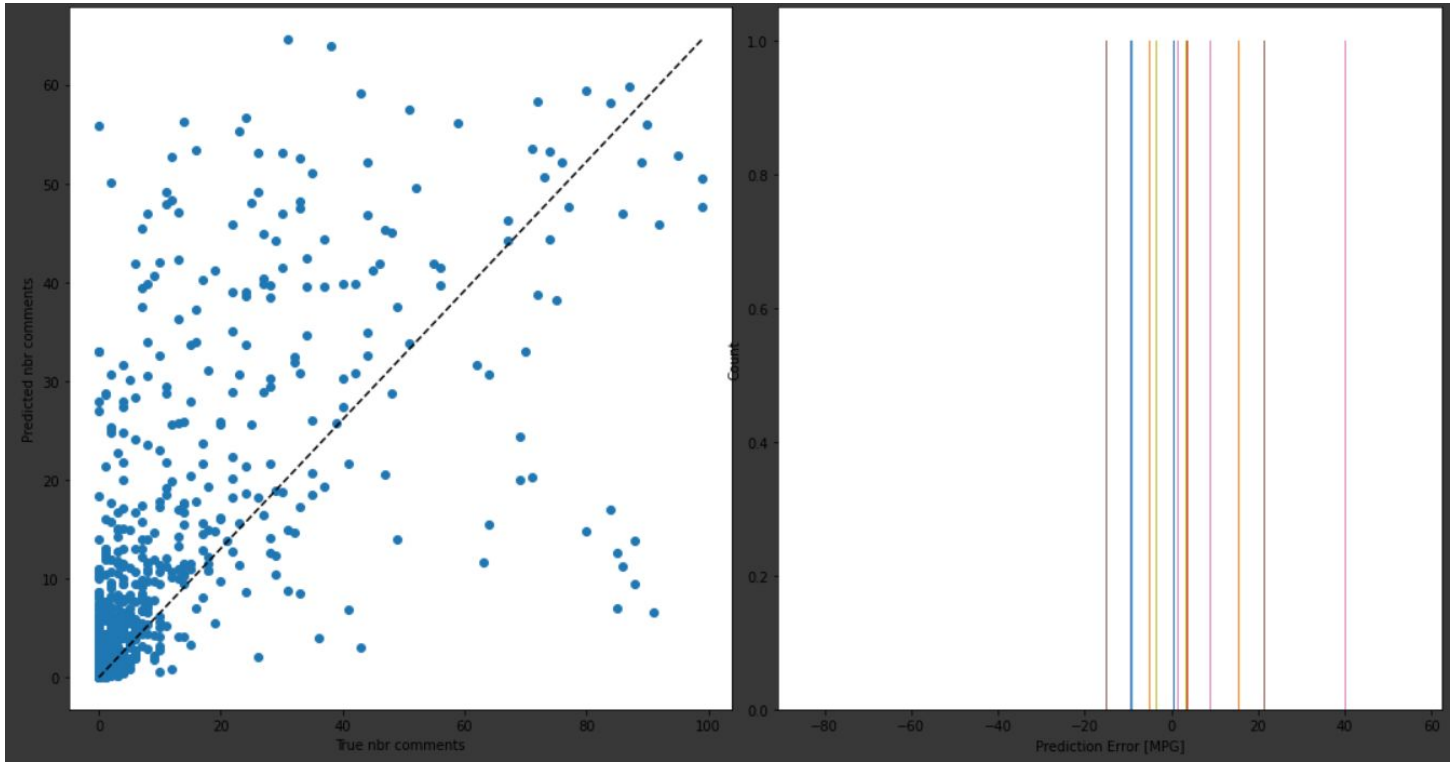


Distribution du nbr de commentaires

	Model	Mean_absolute_error	Root_mean_square_error	R_2_score
0	Linear regression	8.476264	18.844371	-0.115369
1	Gradient Boosting	6.767322	13.668256	0.413212
2	Polynomial Regression	8.522872	19.246600	-0.163492
3	Random Forest	6.554255	13.140796	0.457626
4	MLP	7.963365	21.472808	-0.448215
5	DNN	24.458314	103.346191	0.025842

Evaluation de différents modèles entrainés avec 12 variables et sans lignes avec un nbr de commentaires > 100

Exemple de résultats de performance pour notre Random Forest, et des difficultés de visualization



Reflexions / ouverture

Au vu de la distribution des données dans le dataset, nous nous sommes posés la question de comment contourner le problème de la surreprésentation des posts avec 0 commentaires.

Une solution potentiel que nous n'avons pas pu tester serait de faire 2 modèles au lieu d'un seul :

- un premier modèle de classification dont le rôle serait de prédire si un post *va recevoir des commentaires* ou *s'il y aura 0 commentaires*.
- puis, un autre modèle de régression, qui prédira le nombre de commentaire uniquement sur les posts étant identifiés comme "pouvant avoir des commentaires".

C'est une solution que nous avons vu fonctionner dans différents concours Kaggle !

Nous avons vu que l'auteur était plutôt parti sur la voie des réseaux de neurones, mais dans tous les cas nous avons vu le potentiel de ce jeu de données.
Et peut-être un jour verrons-nous une fonctionnalité d'estimation automatique du nombre de likes ou de commentaires au moment de l'écriture d'un post sur FB ?