Python for Data Analysis

KDD Cup 1999 Data

Elodie TOROSSIAN
Khang TRUONG
Camille VERNERIE

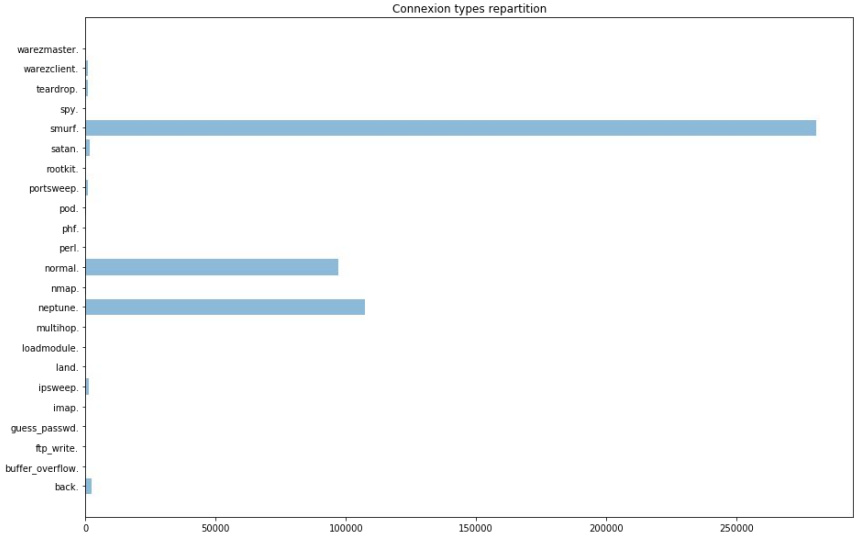
Le contexte de la base de données

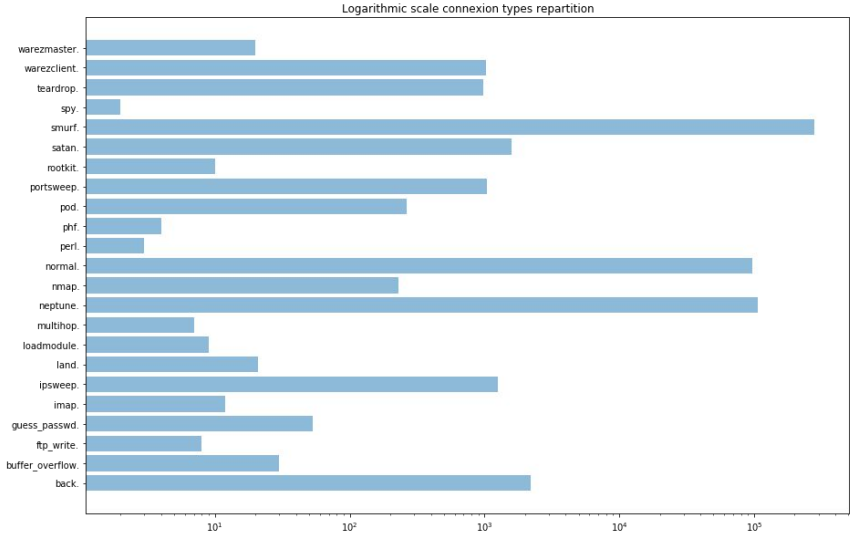
En 1998 le DARPA a enregistré 9 semaines de connexions TCP ayant eu lieu dans une simulation de réseau LAN typique de l'US Air Force, en les parsemant avec des connexions malveillantes.

Le but de ces données était de pouvoir créer un modèle capable de détecter les intrusions sur le réseau.

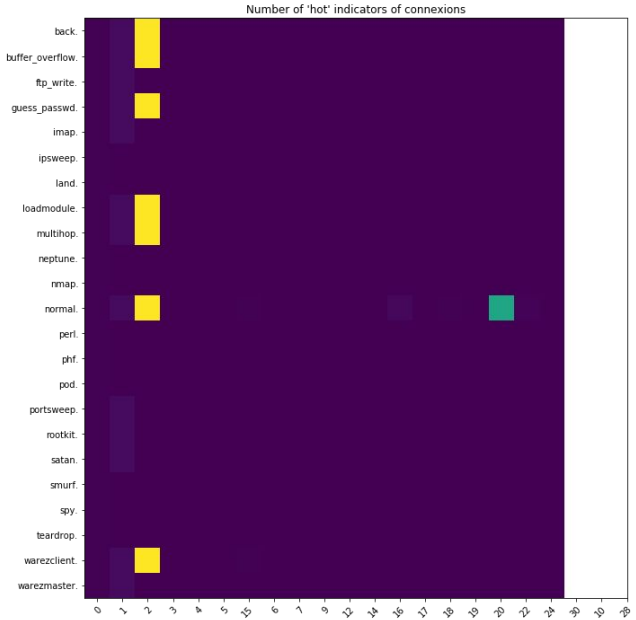
Notre dataset est extrait de cette base de données, et a été utilisé lors de la KDD Cup de 1999.

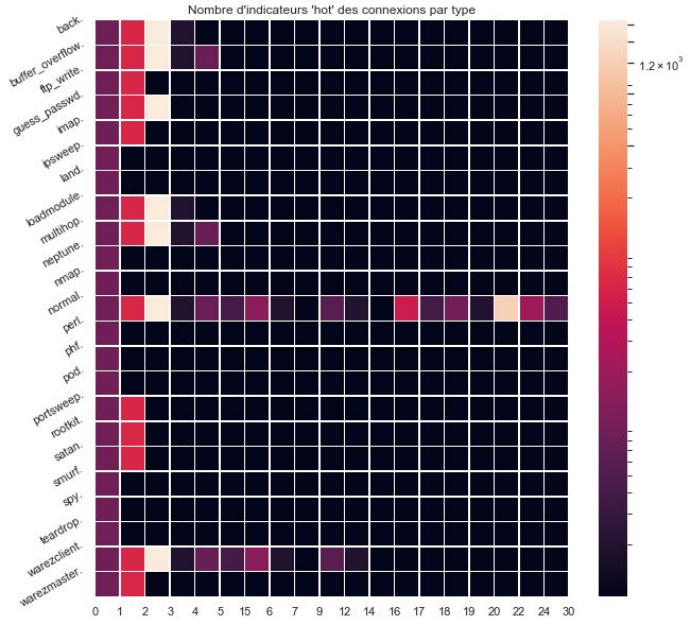
Le contexte de la base de données





Le contexte de la base de données





Les différentes features disponibles

Ce dataset est composé de 42 colonnes qui regroupent différentes informations sur la connexion comme par exemple :

- la durée de la connexion
- le flag de la connexion
- le nombre de paquets urgents
- le nombre de mauvais indicateurs
- le pourcentage d'erreur de différentes choses
- ...

Les étapes de la récupération de données

Nous nous sommes servis du dataset avec peu de données car le dataset complet prenait trop de temps à charger

Nous avons chargé les données avec la fonction read_csv de pandas

Le feature engineering

Le feature engineering n’a pas été nécessaire pour ce dataset car la majorité des données sont numériques et :

- Les données ne comportent pas de valeur NULL
- Les données ne comportent pas de NA

Les modèles

Nous avons essayé plusieurs modèles pour faire la classification des données, comme :

- L’arbre de décision
- Le random forest
- La Gaussian Naive Bayes

Les scores de nos modèles étaient supérieurs à 95% ce qui nous a paru curieux. Nous n’avons pas réussi à trouver la cause de ces scores trop élevés.

Le test de différents hyperparamètres

Pour tester différents hyperparamètres nous avons utilisé une grid search.

Grâce à cela nous avons fait changer différents paramètres (comme le nombre max de features, le nombre d'estimateurs..) du random forest

Le gain de performance des différents modèles

Nos modèles étaient globalement excellents (ce trop bon score nous parait suspect)

Le meilleur de tous les modèle était le random forest

Il est possible qu'il y ait un problème dans notre modèle car les résultats sont trop bon. Il y a peut être dans données une colonne qui est colinéaire à la valeur recherchée.