PSA Challenge : Maintenance prédictive

PROBLEMATIQUE

L'objectif principal de ce défi est prédiction de panne au niveau de machines en fonction des données fournies par les capteurs de celles-ci. L'idée est de créer un algorithme facilement compréhensible et exploitable, sans besoin d'intervention d'un data scientist (bien que de nombreux outils effectuant ce travail existent déjà), dans le but de réduire les stocks de pièces de rechange en effectuant des prédictions précises. Cet algorithme est tenu de respecter les critères de performance (algorithme fonctionnel et entièrement automatisé), de qualité (la précision des prédictions doit être suffisamment élevée pour mieux anticiper la commande des pièces) et compréhensibilité (l'algorithme doit être compréhensible et exploitable sans besoin de travail supplémentaire pour le mettre à jour).

LE CŒUR DU TRAVAIL

I – DEBUT DE LA REALISATION ET PREMIER CONSTAT

1) *Un jeu de données peu exploitable*

Afin de pouvoir créer un modèle prédictif suffisamment performant, il nous faut un jeu de données précis et de taille assez conséquente (plus nous avons de données plus fiables et précises sont nos prédictions). Le souci : le jeu de données fournit par PSA était très peu complet et imprécis. Le manque d'informations concernant les données fournies, le jeu de données étant assez pauvre, la présence de nombreux champs vides (NaN ou Not a Number), sont des éléments qui compliquent l'eploitation des données et rendent nos modèles très imprécis et peu fiables.

2) *Une alternative au problème de données*

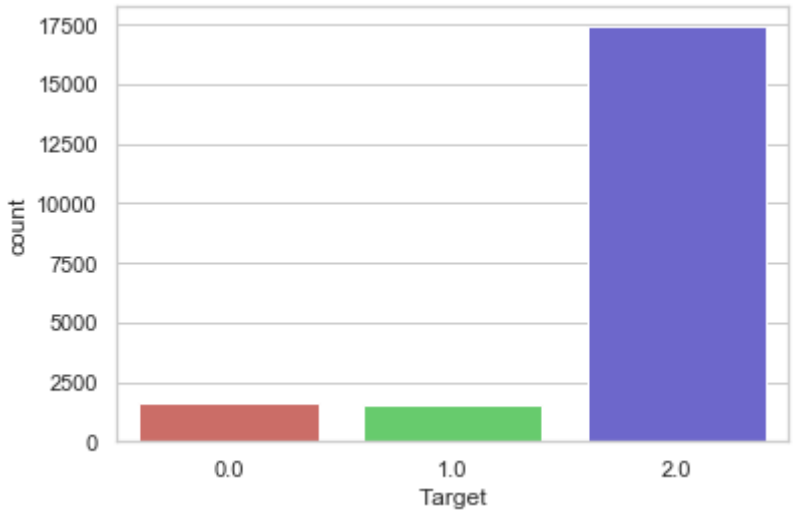
Sur le site de Kaggle, des jeux de données d'entraînement et de test ainsi qu'un algorithme de régression logistique dans le cadre de la maintenance prédictive étaient disponibles. Nous nous sommes donc inspirés de l'algorithme pour concevoir le notre en nous servant des fichiers d'entraînement et test dont ils se sont servis et qu'ils ont mis à disposition.

A partir de ça, nous avons établi un algorithme de classification exploitant la régression logistique mais également du Random Forest et du K-Nearest Neighbours, le but étant de faire une petite étude

comparative des 3 classificateurs et observer leur précision et leur comportement avec le jeu de données provenant de Kaggle.

II – PRESENTATION DES RESULTATS

Afin d'établir un modèle efficace, il était essentiel d'effectuer une séparation en classes et une égalisation de ces classes. Nous avons séparé nos données en 3 classes et, comme vous pouvez le voir sur la figure ci-dessous, le souci est que l'une des classes est prépondérante par rapport aux 2 autres.



Cela fausserait nos prédictions car le modèle choisirait par défaut la classe la plus prépondérante, étant donné qu'il aurait plus de chance d'être correcte. Il faut donc égaliser les données dans les classes afin d'avoir une probabilité équilibrée d'obtenir l'une des 3 classes (soit 1/3 pour chaque classe).

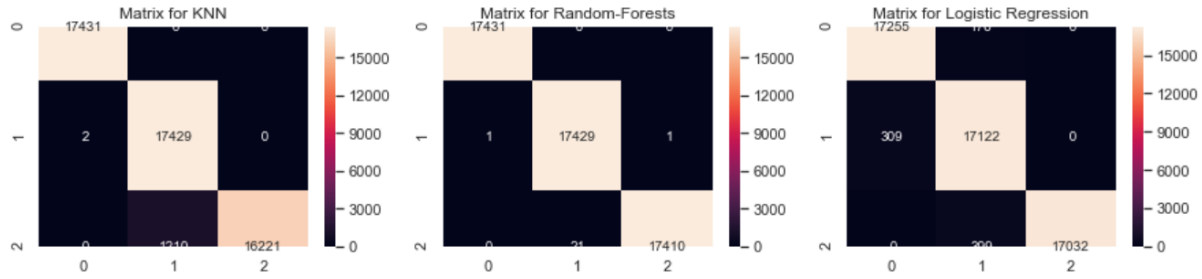
Ceci étant fait, nous obtenons la matrice de prédiction suivante (chaque ligne correspond à un cycle écoulé d'une machine, chaque colonne représente une classe : colonne 1 classe 0, colonne 2 classe 1, colonne 3 classe 2) :

```
[4.26957527e-56 2.47553188e-03 9.99975245e+01]
[8.75199814e-56 3.92538867e-03 9.99960746e+01]
[1.14742535e-55 6.05399973e-03 9.99939460e+01]
...
[8.10168755e+01 1.89830480e+01 7.65165228e-05]
[8.16956782e+01 1.83042728e+01 4.90154839e-05]
[8.13376877e+01 1.86622821e+01 3.01849020e-05]]
```

En appliquant la régression logistique, nous obtenons une précision de l'ordre de 90% et obtenons la matrice de confusion suivante :

```
[[7166  879    0]
 [   0  871    0]
 [   0  428 3652]]
```

Nous avons par la suite mené une étude comparative avec les algorithmes de classification Random Forest et KNN afin d'effectuer une étude comparative des classificateurs et déterminer lequel est le plus optimal pour de la régression logistique. Les résultats ont attesté que le Random Forest se classe devant les autres avec une précision de 98%, ce qui en fait le meilleur des 3 estimateurs pour de la maintenance prédictive. La matrice de confusion de chacun des 3 modèles est la suivante :



A partir de là, il nous est donc possible d'établir un modèle fiable et précis, certes pas entièrement, mais suffisamment pour réduire considérablement les stocks de pièces de rechange.