Machine Learning - Classifier

Groupe 13
Dupont Loïc - Alexander Jérôme - Dascotte Louis

Université de Mons

UMONS Université de Mons
Faculté des Sciences

12 mai 2023

Quel procédure choisir ?

- Comme énoncé dans le rapport, nous avions le choix entre 2 procédures. L'une consistait à enlever toutes les lignes contenant les *NaN* et l'autre était d'en conserver et éliminer uniquement celle où la valeur *NaN* se trouvait à la colonne Y.
- Les 2 approches ont été testées et comparées avec la log_loss.

Quel procédure choisir ?

- Comme énoncé dans le rapport, nous avions le choix entre 2 procédures. L'une consistait à enlever toutes les lignes contenant les *NaN* et l'autre était d'en conserver et éliminer uniquement celle où la valeur *NaN* se trouvait à la colonne Y.
- Les 2 approches ont été testées et comparées avec la log_loss.
- Celle qui éliminait toutes les lignes a donc été retenue.

Quel modèle choisir ?

- Comme pour la slide précédente, nous nous sommes basés au préalable sur l'exploration des données pour sélectionner le modèle le plus adapté à ce que nous possédions.
- Nous avons rapidemment éliminer *kNN* et *LR* de par la forme des données et également *GaussienNB* car les données ne suivaient pas une loi normale.

Quel modèle choisir ?

- Comme pour la slide précédente, nous nous sommes basés au préalable sur l'exploration des données pour sélectionner le modèle le plus adapté à ce que nous possédions.
- Nous avons rapidemment éliminer *kNN* et *LR* de par la forme des données et également *GaussienNB* car les données ne suivaient pas une loi normale.
- Nous nous sommes donc penchés sur le *random forest* ainsi que le *gradient boosting* pour plus d'efficacité. Les arbres nous serviront pour prendre de meilleures décisions que ceux cités précédemment.

Pour le Random forest

- Ce modèle fonctionne sur un dataset sur lequel a été effectué un bootstrap.

Pour le Random forest

- Ce modèle fonctionne sur un dataset sur lequel a été effectué un bootstrap.
- Il crée des arbres de décisions aléatoires, en utilisant une sélection aléatoire des prédicteurs pour les séparations et en choisissant le meilleur d'entre eux. Chaque arbre utilise un dataset qui a reçu un bootstrap différent des autres arbres.

Pour le Random forest

- Ce modèle fonctionne sur un dataset sur lequel a été effectué un bootstrap.
- Il crée des arbres de décisions aléatoires, en utilisant une sélection aléatoire des prédicteurs pour les séparations et en choisissant le meilleur d'entre eux. Chaque arbre utilise un dataset qui a reçu un bootstrap différent des autres arbres.
- Pour les tests, on passe par chaque arbre et on regarde les résultats.

Pour le Gradient Boosting

- Le Gradient Boosting utilise lui aussi des arbres de décisions, mais différemment du Random Forest.

Pour le Gradient Boosting

- Le Gradient Boosting utilise lui aussi des arbres de décisions, mais différemment du Random Forest.
- On commence avec un seul arbre de décision qu'on fit sur les résiduels. Avec les résultats, on construit un nouvel arbre.

Pour le Gradient Boosting

- Le Gradient Boosting utilise lui aussi des arbres de décisions, mais différemment du Random Forest.
- On commence avec un seul arbre de décision qu'on fit sur les résiduels. Avec les résultats, on construit un nouvel arbre.
- À chaque étape, un nouvel arbre sera construit en utilisant les arbres précédents. Le modèle va s'améliorer lentement au fur et à mesure.

ExtraTreesClassifier et HistGradientBoosting

- Ce sont des modèles basés sur ceux vu précédemment.

ExtraTreesClassifier et HistGradientBoosting

- Ce sont des modèles basés sur ceux vu précédemment.
- Pour l'ExtraTreesClassifier, au lieu d'utiliser des datasets qui ont reçu un bootstrap, on utilise le dataset de base en entier. Et pour choisir le prédicteur pour les séparations, il ne prend pas le meilleur mais le choisit aléatoirement.

ExtraTreesClassifier et HistGradientBoosting

- Ce sont des modèles basés sur ceux vu précédemment.
- Pour l'ExtraTreesClassifier, au lieu d'utiliser des datasets qui ont reçu un bootstrap, on utilise le dataset de base en entier. Et pour choisir le prédicteur pour les séparations, il ne prend pas le meilleur mais le choisit aléatoirement.
- Pour l'HistGradientBoosting, nos données de départ vont être rassemblées ensemble (binning) pour créer des histogrammes. Ceux-ci vont ensuite être utilisés pour créer les arbres.

PCA

- Maintenant qu'on a notre modèle, on souhaite améliorer nos performances et pour cela, on utilisera une analyse en composantes principales (ACP/ PCA en anglais).

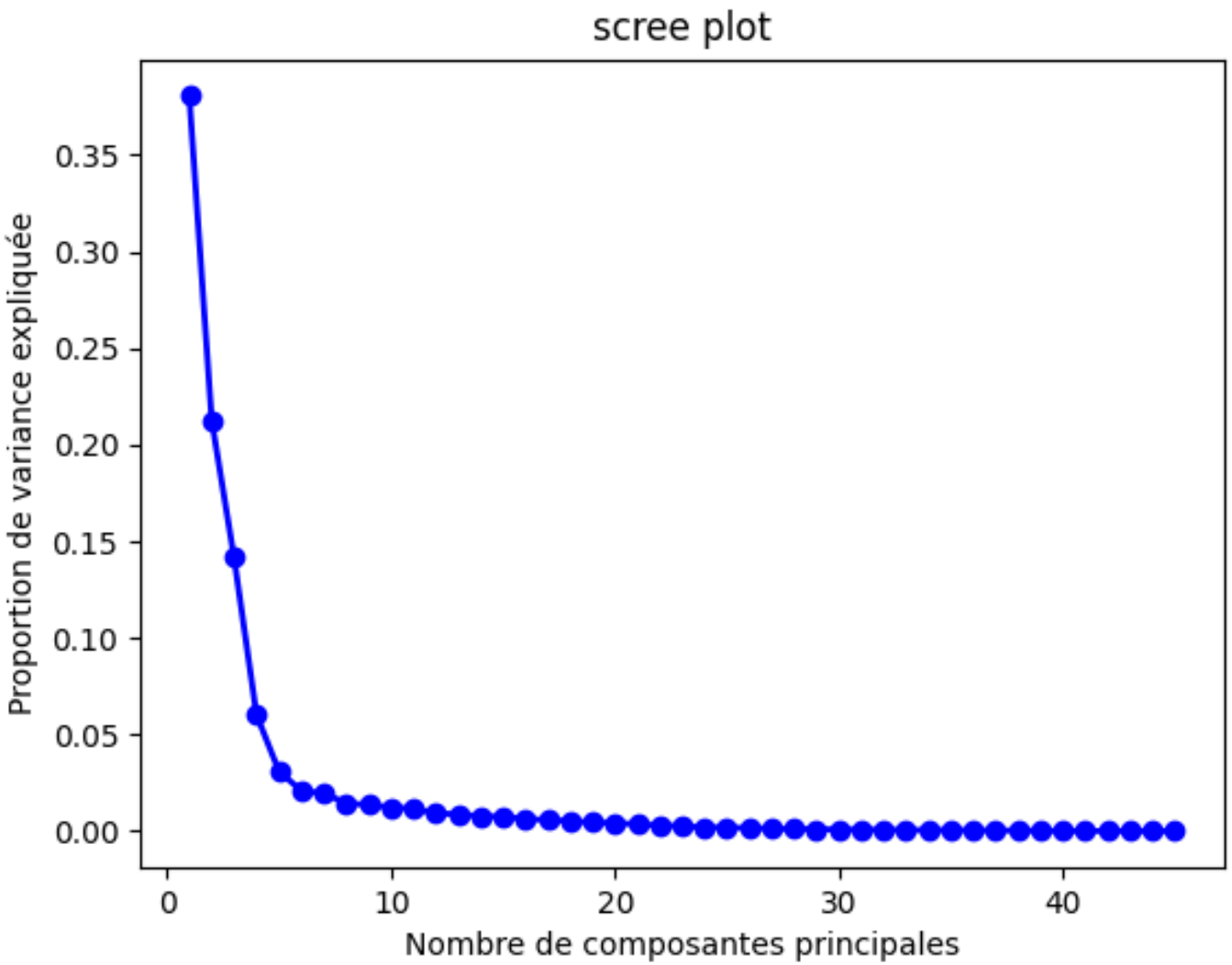
PCA

- Maintenant qu'on a notre modèle, on souhaite améliorer nos performances et pour cela, on utilisera une analyse en composantes principales (ACP/ PCA en anglais).
- L'ACP réduit la dimensionnalité de nos données tout en maximisant l'information.

PCA

- Maintenant qu'on a notre modèle, on souhaite améliorer nos performances et pour cela, on utilisera une analyse en composantes principales (ACP/ PCA en anglais).
- L'ACP réduit la dimensionnalité de nos données tout en maximisant l'information.
- Il va prendre des variables corrélées et les transformer en un plus petit nombre de variables non-corrélées. Ça permettra de réduire la complexité du modèle tout en gardant le maximum d'information.

Scree Plot


scree plot
Proportion de variance expliquée
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
Nombre de composantes principales

Log_loss

Modèle	Log_loss
Random Forest	0.2898170168090797
ExtraTree	0.2751090721395433
Gradient Boost	0.4617706737506961
HistGradientBoost	0.32933181558343994