Pourquoi forêt aléatoire (package ranger) ?

Les forêts aléatoires permettent de gérer un grand jeu de données. Les forêts aléatoires ne prennent pas en compte l’ordre chronologique mais le modèle peut trouver des lag patterns.

Avec les données temporelles ont va créer trois colonnes de décalage :

* H-1 : fluctuations à court terme, ex si il y a eu de forte pluie, les précipitations ont diminuer la quantité de particule fine dangereuse
* J-1 : cycles journaliers, ex changement d’humidité, de pollution dues à l’activité humaine (trafic)
* S-1 : cyles à long terme, ex tempete de sable, événement météorologique exceptionnels peuvent avoir un imact sur plusieur semaine.

On va chercher à prédire IQA (une variable catégorielle à 6 modalités calculé avec les 6 facteurs de pollutions et classés selon sa valeur dans :

* Bon
* Modéré
* Non sain pour sensible
* Non sain
* Très non sain
* Dangereux

On regarde si nos données sont équilibrées :

Une image contenant texte, Police, reçu, capture d’écran

Description générée automatiquement

Il y a un déséquilibre avec « bon » et « dangereux » qui sont sous représentées et « très non sain » qui est sur représenté. Le modèle risque de prédire la classe majoritaire au détriment des classes sous représentées.

Il faut donc ajouter un poids aux classes lors du calcul du modèle.

Données déséquilibrées :

Une image contenant texte, reçu, Police, capture d’écran

Description générée automatiquement

Données équilibrées avec inversement des fréquences, « bon » est encore très mal prédit et « dangereux » c’est pas encore ca.

Une image contenant texte, reçu, Police, capture d’écran

Description générée automatiquement

Poids à revoir car pour données déséquilibrée OOB error = 0.333555 et pour les autres OOB error = 0.3336492 => courant, cela peut vouloir dire qu’il faut revoir les poids choisi, les rentrer à la main, changer de mode de définition des poids…

Limite forêt aléatoire avec données temporelles :

Ne capturent pas les dépendances temporelles à long terme aussi efficacement que LSTM par exemple, spécialisé dans les séries temporelles.

On peut regarder d’autres lag par exemple h-2, j-2 et s-2.

Matrice de confusion

Une image contenant texte, reçu, Police, capture d’écran

Description générée automatiquement

66.6 % des données sont bien prédites

Kappa 🡪 accord modéré entre les valeurs observées et les valeurs prédites

Test de McNemar : p-value < 2.2e-16, suggérant que le modèle fait des erreurs systématiques dans certaines catégories (probablement dû au déséquilibre de classe)

* Sensibilité : combien de valeurs réelles sont bien classées
* Spécificité : combien de valeurs réelles ne sont pas mal classées
* PPV : combien de prédictions positives sont bonnes
* NPV : parmi toutes les prédictions négatives du modèle, quelle proportion est réellement correcte

Pour « bon » NPV = 0.97 donc 97% des valeurs que le modèle n’as pas prédit comme « bon » appartenait bien à une autre classe.

