Soutenance Défi IA 2022

Vendredi 7 janvier 2022

5GMM

Groupe 8 : Ikrame Amine, Jérôme Deveaux, Marie Le Chevère, Folke Skrunes et Aleksander Stangeland

Encadrant : David Bertoin

Pré-traitement des données

Construction de l'ensemble des données

- Charger les .csv : *X_station*, *baseline_forecast*, *baseline_observation*, *stations_coordinates*.
- Pivoter les lignes par rapport à l'heure du relevé du capteur.
- Permet au modèle d'avoir accès aux observations par station,par jour et par heure.

Feature engineering

- Moyenne et écart-type des autres variables de météorologie par jour
- Précipitations cumulées des voisins
- Moyenne des précipitations cumulées des voisins
- Précipitations cumulées pour les jours précédents
- Précipitations moyennes sur n-jours

Données manquantes

- Taux de données manquantes énorme
- Pour Xgboost : il le traite lui-même
- Remplacement par la valeur précédente de la donnée manquante par station pour notre réseau de neurones convolutionnels

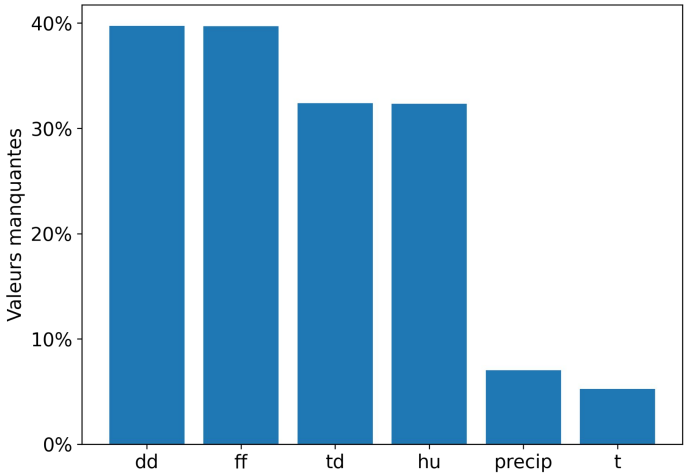


Figure 1 : Diagramme en barres des valeurs manquantes (en %) des variables de *X_station_train*.

Choix du modèle

XGBoost (eXtreme Gradient Boosting)

Avantages:

- Bonnes performances
- Bien adapté à des jeux de données structurés
- Tolérant aux données manquantes

Plusieurs paramètres à régler.

Loss function

Objectif: minimiser le Mean Absolute Percentage Error (MAPE)

MSE: minimise l'erreur *additive*

MSLE: minimise l'erreur *multiplicative*

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2$$

Optimisation bayésienne

- Permet d'optimiser les hyperparamètres du modèle
- Exploration / exploitation
- On optimise 8 paramètres du modèle XGBoost:

n_estimators, *max_depth*, *learning_rate*, *colsample_bylevel*

gamma, *min_child_weight*, *max_delta_step*, *colsample_bynode*

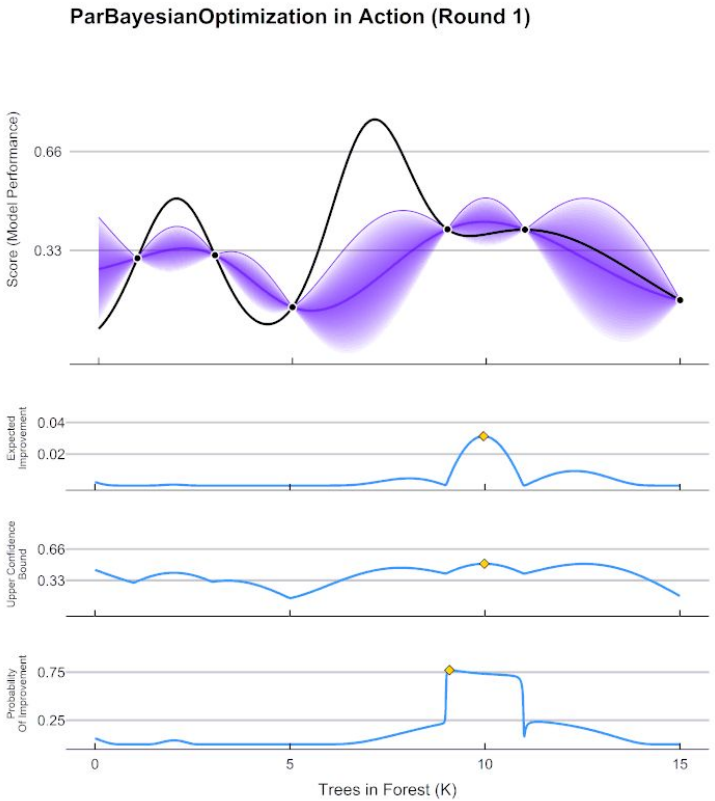


Figure 2 : Optimisation bayésienne

Source : AnotherSamWilson. (2019, 12 décembre). WIKIPEDIA. *Bayesian optimization*. https://commons.wikimedia.org/w/index.php?curid=84842869

Optimisation bayésienne

- Permet d'optimiser les hyperparamètres du modèle
- Exploration / exploitation
- On optimise 8 paramètres du modèle XGBoost:

n_estimators, *max_depth*, *learning_rate*, *colsample_bylevel*

gamma, *min_child_weight*, *max_delta_step*, *colsample_bynode*

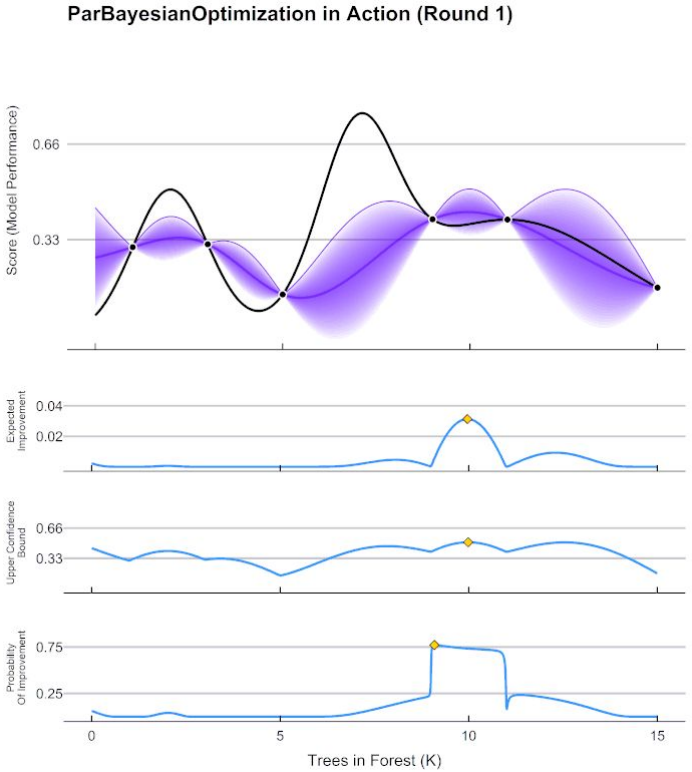


Figure 2 : Optimisation bayésienne

Source : AnotherSamWilson. (2019, 12 décembre). WIKIPEDIA. *Bayesian optimization*. https://commons.wikimedia.org/w/index.php?curid=84842869

Optimisation bayésienne

- Permet d'optimiser les hyperparamètres du modèle
- Exploration / exploitation
- On optimise 8 paramètres du modèle XGBoost:

n_estimators, *max_depth*, *learning_rate*, *colsample_bylevel*

gamma, *min_child_weight*, *max_delta_step*, *colsample_bynode*

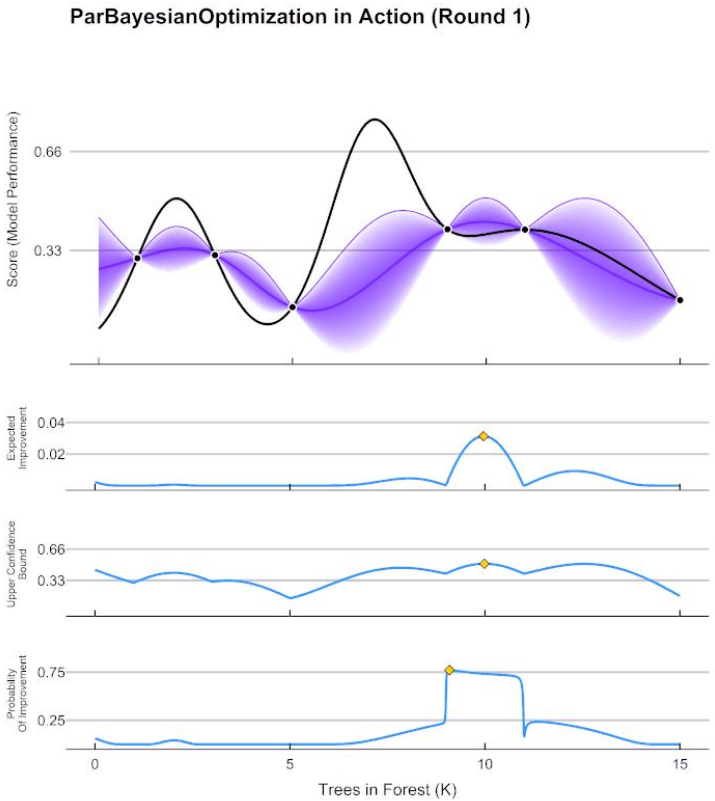
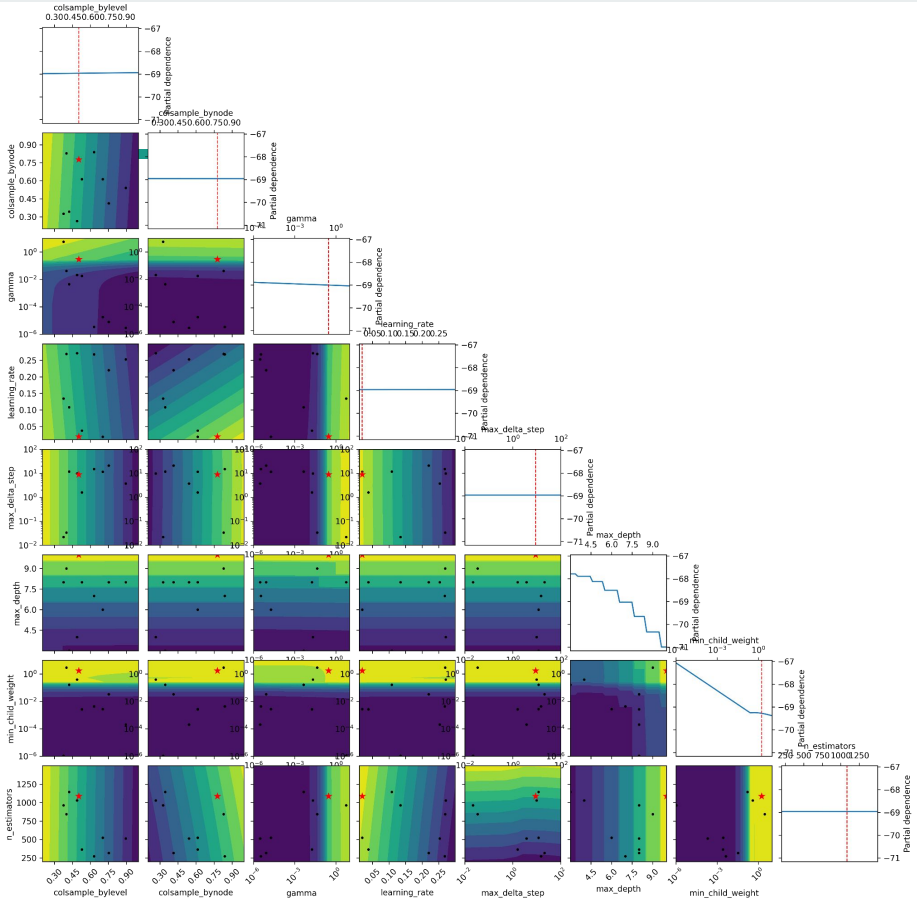


Figure 2 : Optimisation bayésienne

Source : AnotherSamWilson. (2019, 12 décembre). WIKIPEDIA. *Bayesian optimization*. https://commons.wikimedia.org/w/index.php?curid=84842869

Optimisation bayésienne



MAPE: ~27.5 sur le jeu de données *train*

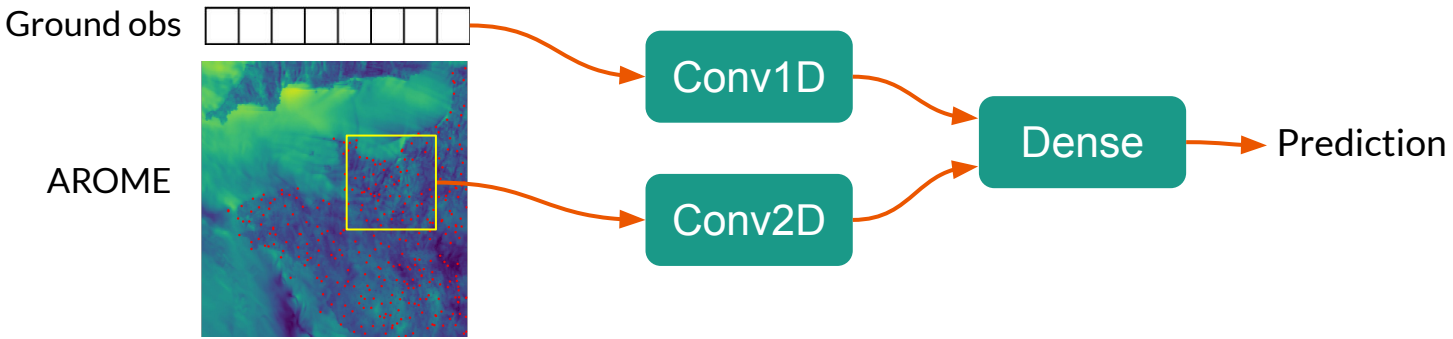
Figure 3 : Paramètres optimaux par optimisation bayésienne

Réseau de neurones convolutionnels (CNN)

Motivation: possibilité d'intégrer les données spatiales (AROME, ARPEGE)

1ère version: convolution 1D sur les 24 observations de chaque variable

2ème version: convolution 2D sur la prédiction du modèle AROME



Réseau de neurones convolutionnels (CNN)

Inconvénients:

- Nécessitent plus de prétraitement (imputation des valeurs manquantes, normalisation)
- Résultats inférieurs à ceux de XGboost
- On n'a pas eu le temps d'implémenter la version 2 avec les données AROME

Résultats

30.54 MAPE

Avec le modèle XGBoost

Améliorations:

- Trouver les paramètres optimaux
- Limiter le surapprentissage
- Autres méthodes d'imputation
- Exploitation des données AROME

Conclusion

- Le modèle XGBoost a eu les meilleures performances.
- Le CNN a eu des résultats inférieurs mais permet d'intégrer les données AROME.
- Avec quelques modifications, nous pensons pouvoir améliorer notre prédiction.

Merci de votre attention

C'est le moment des questions