Université Paris-Dauphine
Machine Learning pour la finance
Examen — XX juillet 2023

Aucun documents, calculatrice ou autre objets électroniques ne sont autorisés.

Prédiction du prix de l'électricité

Le marché des futures sur l'électricité est un marché dynamique en Europe. Vous travaillez pour un cabinet de conseil qui a pour mission de proposer un algorithme de prédiction des prix des futures. Vous travaillez spécifiquement sur la France et l'Allemagne.
Vous disposez d'un dataset avec le prix à prédire, contenant les informations suivantes :

- Date du jour, température, pluie et vent global par pays, ainsi qu'un identifiant du pays
- Production d'énergie : gaz, charbon, hydrolique, nucléaire, photovoltaïque et éolienne par pays
- Utilisation électrique : électricité totale consommé, électricité consommé après l'utilisation des énergies renouvelables, électricité importée depuis l'Europe, électricité exporté vers l'Europe, électricité échangée entre l'Allemagne et la France et inversement

Nous noterons `df` le dataframe Python correspondant à ce jeu de donnée, X la matrice correspondante aux information présentent dans `df` et `y` la variable cible. Pour assurer la sécurité des informations, les prix ont été normalisé entre 0 et 100, de moyenne 50.

1. Dans quel cadre sommes-nous : apprentissage supervisé ou non-supervisé ? Et quel sous-type ? On décide de travailler avec la métrique RMSE : rappeler sa définition.

   Votre data engineer vous affirme que `df` ne contient pas de valeur manquante ni de valeur aberrante. Vous commencez donc la partie d'exploration des données, en prenant soin de vérifier ces informations tout de même.

2. Quelle est la commande Python que vous allez utiliser pour séparer votre base de donnée en une base de donnée d'entraînement et une base de donnée de test, et pourquoi ?

   A. `X_train, X_test, y_train, y_test = train_test_split(X, y)`
   B. `X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y)`
   C. Aucune des précédentes

3. La base de données est relativement pauvre. Quel features pouvez-vous créer qui pourraient être pertinente ?

   Puisqu'il vous a été demandé de concevoir un algorithme performant, vous décidez de privilégier les méthodes de boosting et d'ensemble.

4. Pour commencer, vous décidez de ne travailler qu'avec les informations traitant d'information énergétique ou financière. Vous écartez pour le moment les données climatique et du jour. Sans réglage et sans sélection de variable, vous calculez les performances des trois algorithmes :

Algorithme	RMSE
Random Forest	25
XGBoost	20
AdaBoost	23

   Commentez.

5. Pour avoir de meilleure performance, vous décidez de trouver les meilleurs paramètres pour chaque algorithme, comment vous-y prenez-vous ?

6. Voici les performances avec les meilleurs paramètres :

Algorithme	RMSE	Nombre d'arbres	Learning rate	Profondeur des arbres
Random Forest	23	50	-	7
XGBoost	14	75	0.1	5
AdaBoost	15	250	0.1	2

   On suppose que chaque arbre se développe sans atteindre la limite d'observation minimum par feuille. Calculer le nombre de conditions qui sont présente dans chaque modèle. Quel algorithme décidez-vous de sélectionner ?

7. Lors de votre restitution, on vous demande d'expliciter la différence entre une méthode de Boosting et une méthode d'ensemble. Expliquer la différence, et également celle de fonctionnement entre XGBoost et AdaBoost.

   Votre algorithme semble performant et est testé pendant un an. A la fin de cette période, l'entreprise vous présente les performances mesurée. Il semblerait que votre algorithme ne s'adapte pas très bien au saisonnalité.

8. Un data scientist de votre équipe vous propose de travailler la donnée du jour de la manière suivante : les jours sont numérotés de 0 à 365 puis on utilise un sinus et un cosinus de période $\frac{365}{4} = 91.25$. Il a écrit les formules suivantes pour calculer les deux features :

$$\begin{cases} \mathsf{sin\_day} & = \sin\left(\frac{2\pi x}{\frac{365}{4}}\right) \\ \mathsf{cos\_day} & = \cos\left(\frac{2\pi x}{\frac{365}{4}}\right) \end{cases}$$

   Est-ce que son approche permet de répondre à la problématique soulevée ? Si non, peut-on l'adapter ?

   Suite à votre étude, vous êtes appelés pour réutiliser la base de donnée mais cette fois pour de la visualisation. On se pose la question s'il est capable de visualiser en deux dimensions la base de données. Vous savez que c'est possible, et vous considérer deux algorithmes : Analyse par composantes principales et UMAP.

9. Dans quel cadre sommes-nous : apprentissage supervisé ou non-supervisé ? Qu'est-ce que cela change par rapport au travail précédent ?

10. Est-t-on certain que l'on aura une bonne projection en deux dimensions pour les deux algorithmes ?

Support Vector Machine pour la régression

Nous travaillons avec un data scientist qui nous affirme que nous pouvons utiliser les SVM pour traiter des problèmes de régression. Pourtant, nous n'avons jamais vu cela en cours. Sans plus d'explications, il écrit sur un papier le problème suivant :

$$
\begin{aligned}
(w,b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\min} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\xi_i^* \\
\text{tel que} \quad & \forall i \leqslant n,\ \langle w, x_i\rangle - y_i \leqslant \varepsilon + \xi_i \\
& \forall i \leqslant n,\ y_i - \langle w, x_i\rangle \leqslant \varepsilon + \xi_i^* \\
& \forall i \leqslant n,\ 0 \leqslant \xi_i \text{ et } 0 \leqslant \xi_i^*
\end{aligned}
$$

1. Rappeler le fonctionnement des SVM dans le cadre d'une classification. On pourra s'aider d'un schéma.
2. Comment est adapté le SVM dans le cadre d'une régression si le problème formulé est correct ?
3. Quel est l'impact de l'hyper paramètre C ? Et celui de ε ?