Université Paris-Dauphine
Machine Learning pour la finance
Examen — 28 juin 2023

Aucun document, calculatrice ou autres objets électroniques ne sont autorisés.

Prédiction de l'inflation

Vous intégrez en stage une unité dans une banque d'investissement et héritez du sujet de prédiction de l'inflation. Puisque vous n'êtes que stagiaire, le dataset que l'on vous fournit est anonymisé et standardisé. Cependant, en regardant les informations contenues dans le dataset vous comprenez que vous disposez d'une colonne renseignant le jour et une deuxième renseignant un pays européen. Vous identifiez également la colonne Y comme la variable à prédire.

Votre sujet est d'être capable de prédire l'inflation en France uniquement. Nous noterons `df` le dataframe Python correspondant à ce jeu de données, X la matrice correspondante aux informations présentes dans `df` et `y` la variable cible. On vous indique également que la variable à prédire a été normalisé : ces valeurs sont dans l'intervalle $[0, 1]$.

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 valeurs manquantes ni de valeurs aberrantes. Vous commencez donc la partie d'exploration des données, en prenant soin de vérifier ces informations tout de même.

2. La base de données est relativement pauvre : il n'y a que quinze champs exploitables. Quels indicateurs pouvez-vous créer qui pourraient être pertinents ?

3. Le dataset contient l'inflation pour chaque pays européen, mais votre sujet ne porte que sur la France. Quelle est la commande Python que vous allez utiliser pour ne conserver que les lignes relatives à l'inflation française, et pourquoi ?

   A. `df = df.loc[df["Country"] == "FR", ]`
   B. `df = df[df["Country"] == "FR", ]`
   C. Aucune des précédentes (précisez)

4. Quelle est la commande Python que vous allez utiliser pour séparer votre base de données en une base de données d'entraînement et une base de données 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

   Puisqu'il vous a été demandé de concevoir un algorithme performant, vous décidez de privilégier les méthodes de boosting et d'ensemble. On vous conseille cependant de comparer vos résultats à une régression linéaire.

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

6. Après ce travail réalisé, vous obtenez les résultats suivants :

   Commentez.

Algorithme	RMSE	Nombre d'arbres	Learning rate	Profondeur des arbres
Random Forest	0.8	50	-	7
XGBoost	0.6	75	0.2	5
Régression Linéaire	0.4	-	-	-

7. Un précédent stagiaire avait travaillé sur la même problématique et avait atteint une RMSE de 0.5 avec l'algorithme XGBoost avec 150 arbres de profondeurs 5. Si l'on mesure la complexité d'un algorithme à base d'arbre en comptant 1 pour chaque coupure réalisée, et en supposant que chaque arbre se développe complètement, quel est l'écart de complexité entre votre modèle et celui proposé ?

8. Un expert du domaine remarque dans votre compétition de modèle que lorsque les valeurs de l'indice d'inflation sont dans un intervalle de valeurs habituelles, alors XGBoost et Random Forest sont beaucoup plus performants que la régression linéaire. L'inverse se produit dès que l'indice d'inflation sort de l'intervalle de valeurs habituelles, et c'est encore plus marqué quand la valeur du dataset de test dépasse le maximum de la valeur du dataset de train. Est-ce un comportement prévisible ?

9. Vous parlez de cette problématique à un data scientist, et le lendemain il vous écrit sur un papier l'équation suivante :

$$\omega(x,y) = \max\{x,y\}\, e^{-|x-y|} + \min\{x,y\}\left(1 - e^{-|x-y|}\right)$$

   Et vous dit rapidement que ça peut vous permettre de combiner la prédiction d'XGBoost et la prédiction d'une régression linéaire. Expliquez.

   Suite à votre étude, vous êtes appelés pour la refaire sur l'ensemble des pays européens cette fois. On vous demande cependant de ne pas faire plus de 5 modèles différents. Vous décidez donc d'identifier les pays qui se *ressemblent* sur le plan de l'indice d'inflation.

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

11. Que pensez-vous de l'algorithme DBSCAN pour cette problématique ? Décrivez succinctement son fonctionnement.

12. Comment allez-vous mesurer la performance de votre regroupement ?

Génération de texte par un Large Langage Model

Avec l'engouement planétaire généré par les avancées récentes sur les Large Langage Model (LLM), plus particulièrement appliqué au ChatBot, votre entreprise cherche à se placer dans la course. On vous charge d'exploiter un algorithme pré-entraîné pour générer du texte.
Le data scientist référent sur le sujet vous a écrit un mail rapide avant de partir en vacance où il explique qu'un LLM génère à partir d'un contexte (question et phrase en cours de construction) un vecteur de taille n où chaque valeur du vecteur correspond à la probabilité de continuer la phrase avec le mot correspondant à la position.

Vous n'avez vu en cours que la possibilité de faire de la classification binaire. Supposons que l'on ait un problème avec 3 classes à prédire et que l'on ait 3 algorithmes, chacun dédié à prédire une classe. Pour une observation, vous obtenez donc trois probabilités y_1, y_2 et y_3. Sauf que ces trois probabilités ne forment pas un vecteur de probabilité : $y_1 + y_2 + y_3 \neq 1$ en général. Nous calculons donc :

$$\forall j \leqslant 3,\ \hat{y}_j = \frac{\exp\left(\frac{y_j}{\tau}\right)}{\sum_{i=1}^{3} \exp\left(\frac{y_i}{\tau}\right)}$$

Avec le paramètre τ un réel strictement positif, en général valant 1. On a donc clairement que $(\hat{y}_i)_{i\leqslant 3}$ est un vecteur de probabilité

1. Comment varie ce vecteur en fonction de τ ? Qu'est ce que cela veut dire en terme de *confiance* dans la prédiction ?

2. Généralisez l'exemple que l'on vient de donner avec un nombre $n > 2$ de classes.

   Si l'on effectue un tirage aléatoire suivant la distribution ainsi définie, on obtient fréquemment un texte incohérent. Pour limiter cela, une méthode proposée est d'effectuer un tirage aléatoire sur les k mots les plus probables.

3. Après avoir décrit comment définir une distribution à partir de ces k (fixé) mots les plus probables, expliquer en quoi cette solution permet effectivement d'avoir un texte plus cohérent.

4. En 2019 est proposée une nouvelle manière de résoudre le problème : sélectionner les k (variable) mots qui correspondent à 95% (par exemple) de la distribution. Expliquer en quoi cette solution permet également d'avoir un texte plus cohérent, et donner les avantages et inconvénients par rapport à la solution précédente.