XGBoost - Utilisation avancée

Master 280 - Théo Lopès-Quintas

1 Introduction

Le KOSPI est l'indice des stocks coréen et le KOSPI200 regroupe les 200 stock avec le plus de valeurs. Le KIPSI200 fait environs 90% de la valeur totale du marché coréen de stock exchange. Le VKOSPI correspond à la volatilité implicite du KOSPI.

On s'intéresse ici à la prédiction de la valeur du VKOSPI à l'aide de plusieurs informations. Nous nous plaçons dans le cadre d'une prédiction journalière de la valeur. Il nous est demandé de faire le moins de surestimation de la valeur. Autrement dit, on préfère prédire 50 quand la vraie valeur est 51 que prédire 52.
Ainsi, la mesure pour évaluer la performance du modèle sera :

$$\Phi(y, \hat{y}) = (y - \hat{y})^2 \mathbb{1}_{\hat{y} \leqslant y} + 10 \times (y - \hat{y})^2 \mathbb{1}_{\hat{y} > y}$$

Le compte rendu du TP devra contenir un notebook qui répond aux questions posées dans la partie pratique, mais pas la partie théorique. Elle n'est présente que pour se questionner sur le cours qui a été donné, assurer de sa bonne compréhension et l'approfondir.

2 Partie théorique

Nous reprenons les notations du cours.

1. Cas particulier de la MSE
   (a) Expliciter dans ce cas la valeur de de O_{value}.
   (b) Montrer que le score de similarité peut s'écrire comme :

$$\text{Similarity Score} = \frac{\left(\sum_{i=1}^{n} y_i - \hat{y}_i^{(m-1)}\right)^2}{n + \lambda}$$

   (c) Commenter sur la création de feuille avec peu d'observations quand λ augmente.
2. On souhaite utiliser une fonction de perte différente de la MSE.
   (a) Pourquoi ne peut-on pas utiliser la MAE directement ?
   (b) Doit-on utiliser une fonction de perte symétrique ?

3 Partie pratique

Après avoir fait un travail d'exploration des données et un travail de création de features sans utiliser de données extérieure, répondre au problème présenté dans l'introduction.
On demande à ce que soit présent dans le compte-rendu :

1. La construction d'une baseline sans algorithme de Machine Learning

2. Présentation et implémentation d'une mise en place de recherche d'hyperparamètre
3. Comparaison de performance pour le cas d'usage donné entre la MSE et une fonction de perte bien choisie pour XGBoost
4. Implémentation d'une compétition de modèle et discussion sur les résultats