UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Estimate CO2 emissions from cars

Kaggle competition

Processus

Nettoyage

Nom des colonnes
Valeur des catégories

Filtre

Valeurs statiques
Valeurs manquantes

K-Fold

5 folds
Tirage train/valid

Imputation

Catégorielle:
par groupe/mode

Numérique:
par groupe/moyenne

Modélisation

Régression utilisée
Paramètres

Prédiction

Résultats
Assemblage des folds

1. Nettoyage

Opérations:

- Passage en minuscules
- Suppression des valeurs entre parenthèses
- Suppression des caractères blancs
- Remplacement des espaces par tiret bas

Electric range (km) devient electric_range

Application:

- Sur le nom des colonnes
- Sur les valeurs catégorielles

2. Filtre

Données manquantes

variable --- str	missing --- f64
mms	100.0
ernedc	100.0
de	100.0
vf	100.0
enedc	83.84
electric_range	82.96
z	77.98
erwltp	46.48
it	37.78
fuel_consumption	23.51
ec	13.51
mt	11.1
vfn	8.61
mp	6.41
at2	2.34
at1	2.19
date_of_registration	1.7
cn	1.52
ve	0.42
ep	0.25
tan	0.16
va	0.16
ct	0.16
w	0.16
t	0.02

Données statiques

variable --- str	unique --- u32
mms	1
r	1
ernedc	1
de	1
vf	1
status	1

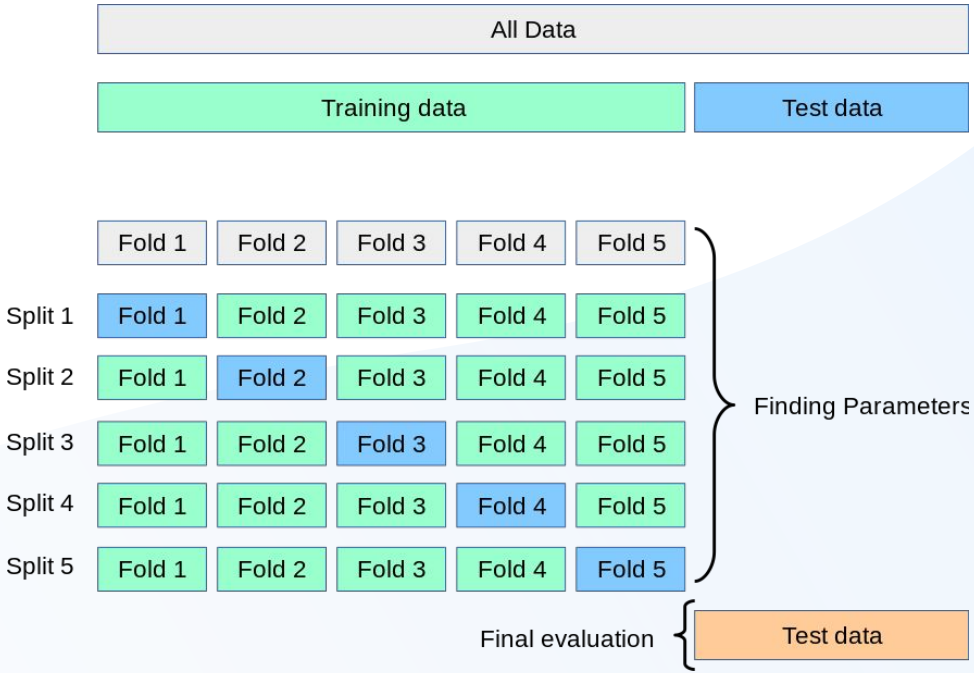
Colonnes conservées

variable --- str	missing --- f64
enedc	83.84
electric_range	82.96
z	77.98
erwltp	46.48
it	37.78
fuel_consumption	23.51
ec	13.51
mt	11.1
vfn	8.61
mp	6.41
at2	2.34
at1	2.19
date_of_registration	1.7
cn	1.52
ve	0.42
ep	0.25
tan	0.16
va	0.16
ct	0.16
w	0.16
t	0.02

3. K-Fold

Split train/test: 99.5 / 0.05 stratifié par type de fuel

~380k lignes de test



5 folds: découpage 80/20 pour train/valid

3.1. Imputation

Par groupe

variable	missing
str	f64
enedc	83.84
electric_range	82.96
z	77.98
erwltp	46.48
it	37.78
fuel_consumption	23.51
ec	13.51
mt	11.1
vfn	8.61
mp	6.41
at2	2.34
at1	2.19
date_of_registration	1.7
cn	1.52
ve	0.42
ep	0.25
tan	0.16
va	0.16
ct	0.16
w	0.16
t	0.02

- Groupe (cn / va / ve)
- Catégories imputées
- Valeurs numériques imputées
- Valeurs manquantes remplacées par 0

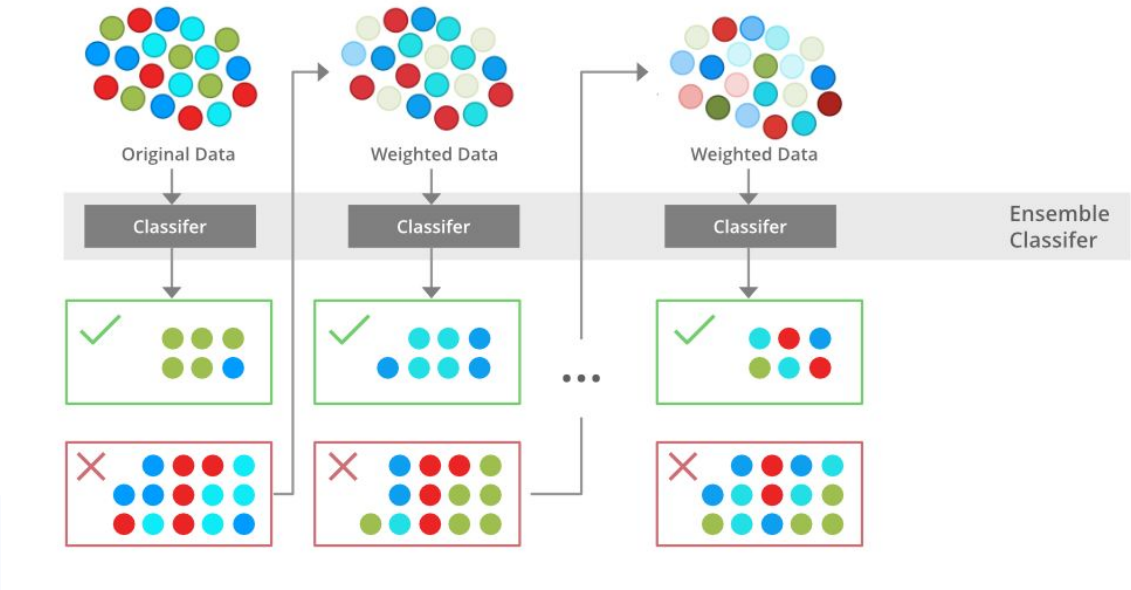
Méthode

Par mode pour les catégories

Par moyenne pour les numériques

Par zéro pour les électriques/non-électriques

3.2. Modélisation



Hyper-paramètres optimisés à l'aide d'Optuna:

max_depth min_child_weight subsample colsample_by(tree/level/node) reg_alpha reg_lambda

n_estimators gamma learning_rate

3.3. Prédiction

Chaque fold a généré un modèle

Amélioration par calcul sur test de out-of-fold

Avec la meilleure optimisation:
calcul de la moyenne des prédictions pour le set de soumission