

Estimation du volume consommé à partir de capteurs de température

Sacha Alidadi Heran

22 août 2023

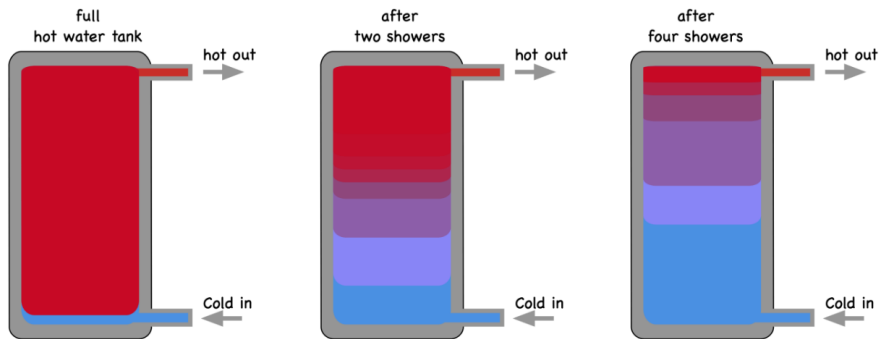
Sommaire

- 1 Contexte
- 2 EDA
- 3 Entraînement et analyse des résultats
- 4 Conclusion

Contexte du projet

- Les ingénieurs de BDR Thermea France, en particulier ceux travaillant sur le chauffe-eau Elensio, cherchent à comprendre et optimiser le volume d'eau consommé par leurs produits.
- Les débitmètres sont très chers et ont une empreinte carbone très élevé
- Le but est de créer un modèle d'apprentissage automatique qui prédit le volume consommé à partir des températures dans le ballon d'eau chaude

Pourquoi utiliser des capteurs de températures ?



Objectifs



- Prédire de manière précise la consommation d'eau en se basant sur les températures du ballon d'eau chaude "Elensio"

Contraintes

- Le matériel informatique embarqué dans le chauffe-eau Elensio impose des limites en termes de puissance de calcul
- Malgré leurs puissances, les modèles d'apprentissage profond seront à écarter.

Présentation du dataset

- Le dataset contient des données d'un ballon d'eau chaude d'un client de BDR Thermea. Elles ont été enregistrées chaque seconde, du 28 février 2023 au 13 avril 2023.
- On y retrouve entre autres des variables comme la température à différents niveaux du ballon, la consommation d'eau ...

Capteurs de températures

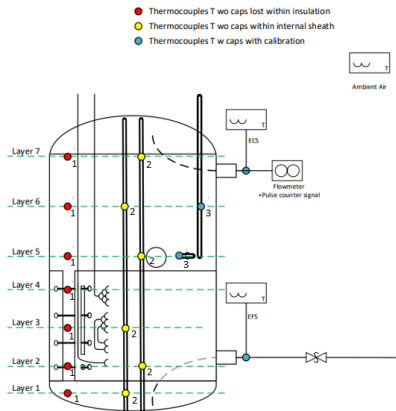


Figure – Répartition des différents capteurs de température dans le ballon

Approche utilisée

- Prédire le volume sur une tranche horaire
- Tranche horaire : période de temps non glissante (e.g 16 :20 ~ 16 :40)

Présentation des variables

- Notre dataset contient 29 colonnes initialement
- Pour optimiser le temps de calcul de notre modèle, nous avons créé 22 nouvelles features à partir des données existantes
- Entre autres : différence entre le début et la fin ou le max et le min de la tranche horaire ...

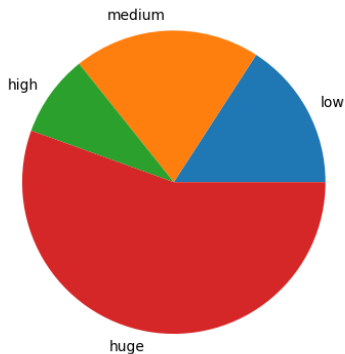
Identification des valeurs manquantes

- Aucune valeurs manquantes dans nos variables
- On retrouve des NaN lors de la création des features

Statistiques de nos sous-tirages

- 85% de nos tranches horaires n'ont pas de sous tirages
- La plupart de nos sous-tirages sont des sous-tirages faibles.
- Certaines journées ne contiennent que des sous-tirages faibles ou énormes.

Statistiques de nos sous-tirages

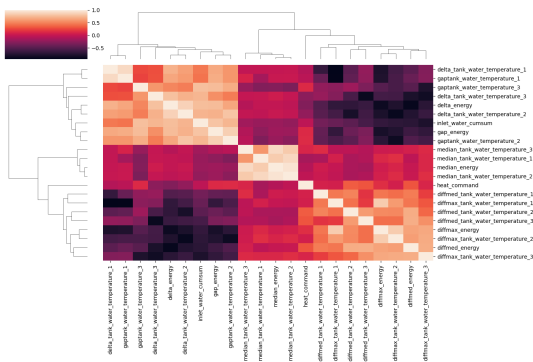


- En cumulé, plus de la moitié du volume vient des énormes sous-tirages

Sélection des meilleures features

- Pour des raisons d'efficacité, nous avons décidé de retirer les features les moins corrélés à notre variable target lors de l'entraînement
- Dressons une matrice de corrélation pour nos différentes features

Sélection des meilleures features



- Clustermap de la matrice de corrélation (regroupe les features les plus corrélées entre elles)
- Les features les plus corrélées sont les delta et les gap.

Modèles utilisées

- Régression linéaire
- Random Forest
- AdaBoost
- GradientBoosting

Métriques utilisées

- R^2 score
- Erreur Moyenne Quadratique
- Erreur Moyenne Absolue
- Erreur Max
- Erreur relative
- Erreur cumulée

Métriques utilisées

Nous avons regroupé ces métriques selon 7 catégories

- global : Toutes catégories confondues
- usage : Toutes catégories sauf "none"
- none : Pas de sous-tirages
- low : Sous-tirage faible
- medium : Sous-tirage moyen
- high : Sous-tirage gros
- huge : Sous-tirage énorme

Problèmes rencontrés

- Répartition inégale des catégories à cause de la définition arbitraire
- Bien répartir les catégories entre les données d'entraînement et de test

Meilleurs modèles

- Modèle (1) : GradientBoostingRegressor(n estimators=50)
- Avantages : R^2 score proche de 1, somme des erreurs presque nulle et erreur relative à peu près égale entre les catégories

Meilleurs modèles

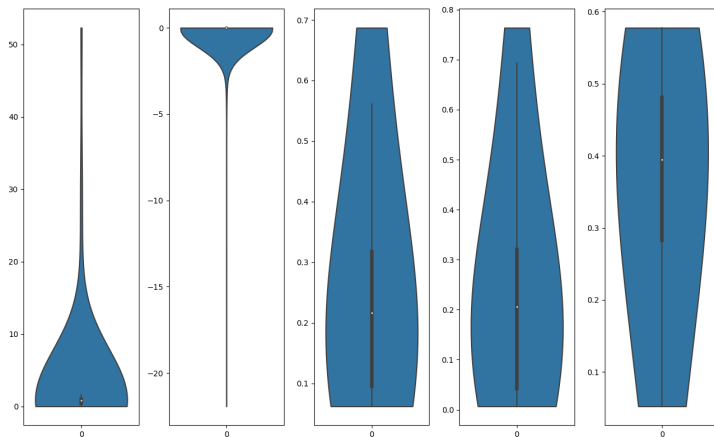


Figure – Diagramme en violon du modèle 1

Meilleurs modèles

- Modèle (2) : GradientBoostingRegressorAdaBoostRegressor(estimator=RandomForestRegressor(n_estimators=200))
- Avantages : Erreur relative très faible et à peu près identique entre les catégories
- Inconvénients : Somme des erreurs très élevée.

Meilleurs modèles

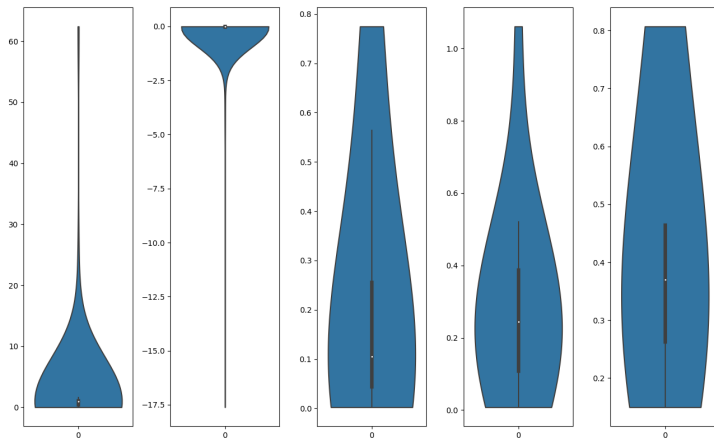


Figure – Diagramme en violon du modèle 2

Meilleurs modèles

- Modèle (3) :
`AdaBoostRegressor(estimator=RandomForestRegressor(n_estimators=50),n_estimators=100)`
- Avantages : R2 score élevé, MAE faible
- Inconvénients : Erreur relative à 75% inégale entre les catégories

Meilleurs modèles

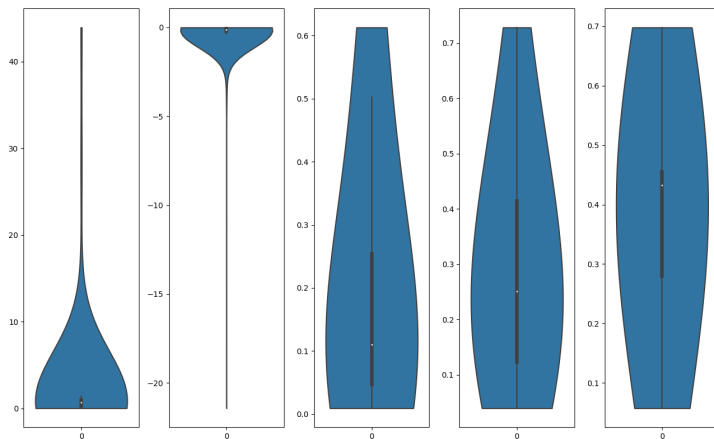


Figure – Diagramme en violon du modèle 3

Le meilleur modèle

- Des 3 modèles, le modèle (1) est le meilleur.
- Malgré quelques défauts de prédiction sur la catégorie "low"

Analyse des erreurs

```

gaptank_water_temperature_3    0.826338
gap_energy                     10.166163
diffmax_energy                 -0.04437
heat_command                   0.0
actual                         3.8
predict                        3.485864
Name: 2023-03-02 11:20:00, dtype: Float64
gaptank_water_temperature_3    0.710451
gap_energy                     3.055346
diffmax_energy                 -0.012588
heat_command                   1200.0
actual                         3.2
predict                        0.124686
Name: 2023-03-01 09:40:00, dtype: Float64

```

Figure – En haut, prédiction du modèle 1 avec une erreur de 8%. En bas, prédiction du modèle 2 avec une erreur de 2500%

Analyse des erreurs

```

gaptank_water_temperature_3    0.669077
gap_energy                     8.138509
diffmax_energy                 -0.046161
heat_command                   0.0
actual                         2.65
predict                        2.390605
Name: 2023-03-10 16:20:00, dtype: Float64
gaptank_water_temperature_3    0.714435
gap_energy                     3.297329
diffmax_energy                 -0.00923
heat_command                   558.0
actual                         2.25
predict                        0.006337
Name: 2023-03-02 19:20:00, dtype: Float64

```

Figure – En haut, prédiction du modèle 1 avec une erreur de 11%. En bas, prédiction du modèle 2 avec une erreur de 99%

Pistes d'amélioration

- Récolter plus de données du client
- Récolter des données d'autres clients
- Redéfinir nos catégories

Bilan

- Compétences acquises : Statistiques, apprentissage automatique, data science
- La plupart du stage s'est passé en remote, je suis juste allé à l'entreprise une fois pour récupérer l'ordinateur de travail et une autre fois pour le rendre.