ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS.

PROJET 3 - MARTIN VIELVOYE
OPEN CLASSROOM - INGÉNIEUR MACHINE LEARNING

PROBLÉMATIQUE

- Presentation problématique, l’interpreter, et les pistes de recherches envisagés

PRÉDIRE LES ÉMISSIONS DE CO2 ET LA CONSOMMATION TOTALE D'ÉNERGIE DE BÂTIMENTS POUR LESQUELS ELLES N'ONT PAS ENCORE ÉTÉ MESURÉES.

Axe principale :

- Prédire pour des bâtiments dont on ne possède que des caractéristiques et des données ‘*méta*’.

Axe secondaire :

- Prédire pour des bâtiments dont on possède des caractéristiques et des données « meta » ainsi que les performances de l’année précédentes.

Data set initial

- Relevé et des descriptions d'environ 3300 bâtiments de Seattle en 2015 et 2016.
- 47 features

PRÉPARATION

- Présentation du cleaning, du feature engineering et de l’exploration

Exploration

- Histogramme de la distribution de valeurs outputs de l'énergie et de l'émission.
- Etude des valeurs qualitatives et quantitatives.

Nettoyage

- Conservation des colonnes pertinentes pour les modèles.

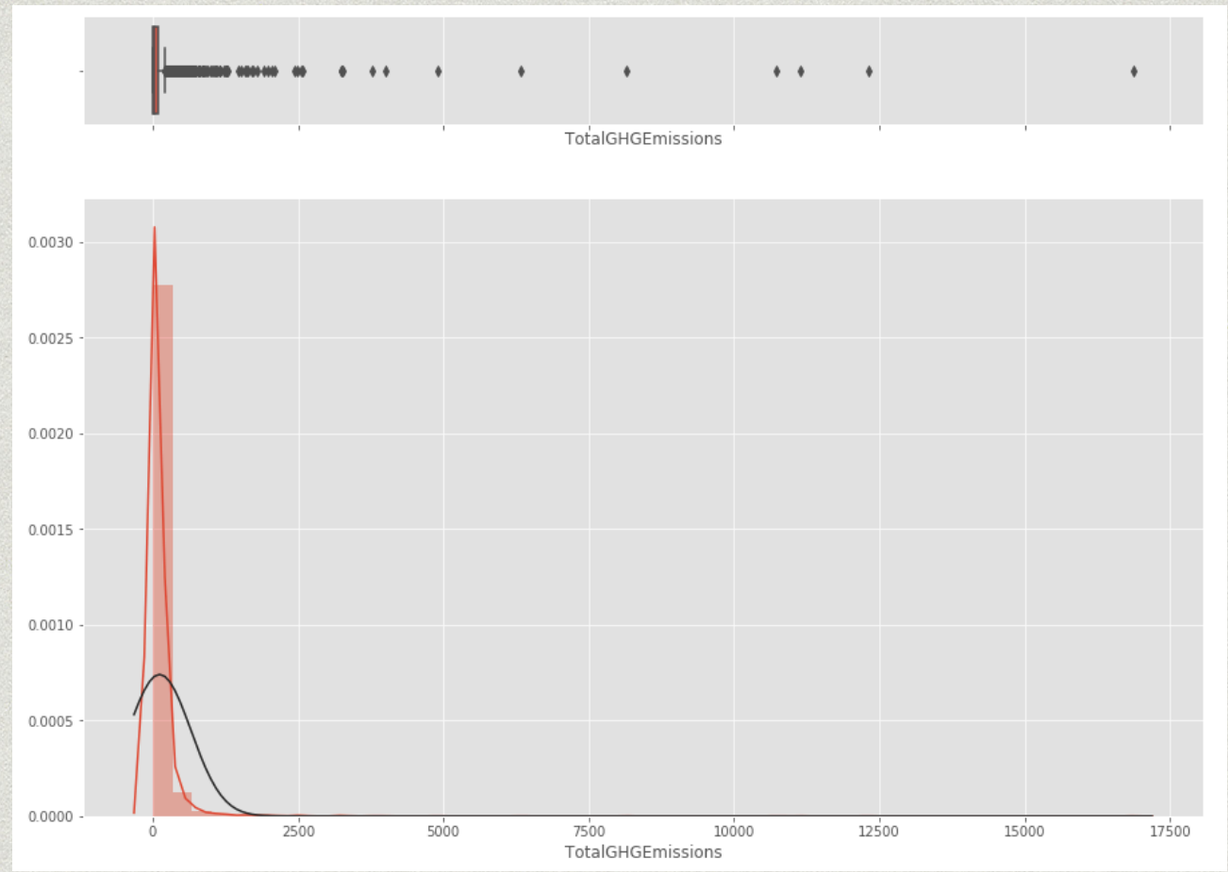
- Surpression des entrées où *SiteEnergy* ou *TotalGHG* sont des valeurs nulles.

- Analyse et nettoyage des dupliquas, des entrées nulles et des valeurs aberrantes.

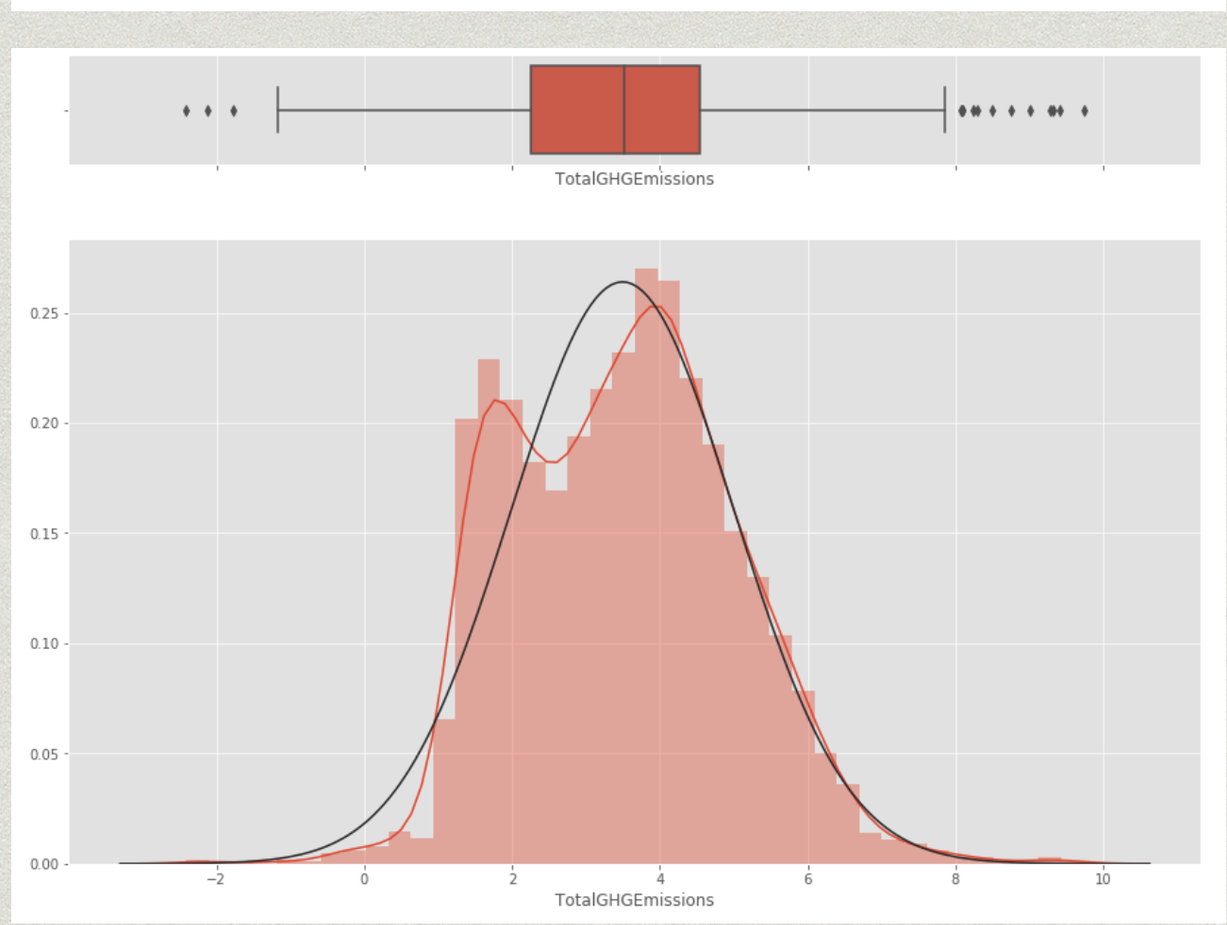
Feature engineering

- Regroupement des valeurs de la variable « année de construction » par groupes de 10.
- Transformation logarithmique des données outputs.
- OneHotEncoding pour les features qualitatives.
- Normalisation pour les features quantitatives.
- Préparation des jeux de données pour les différentes modèles.

Distribution initiale pour l'émission de gas



Distribution des mêmes données après une transformation logarithmique



MODÉLISATIONS

- Presentation des différentes pistes de modélisations effectué

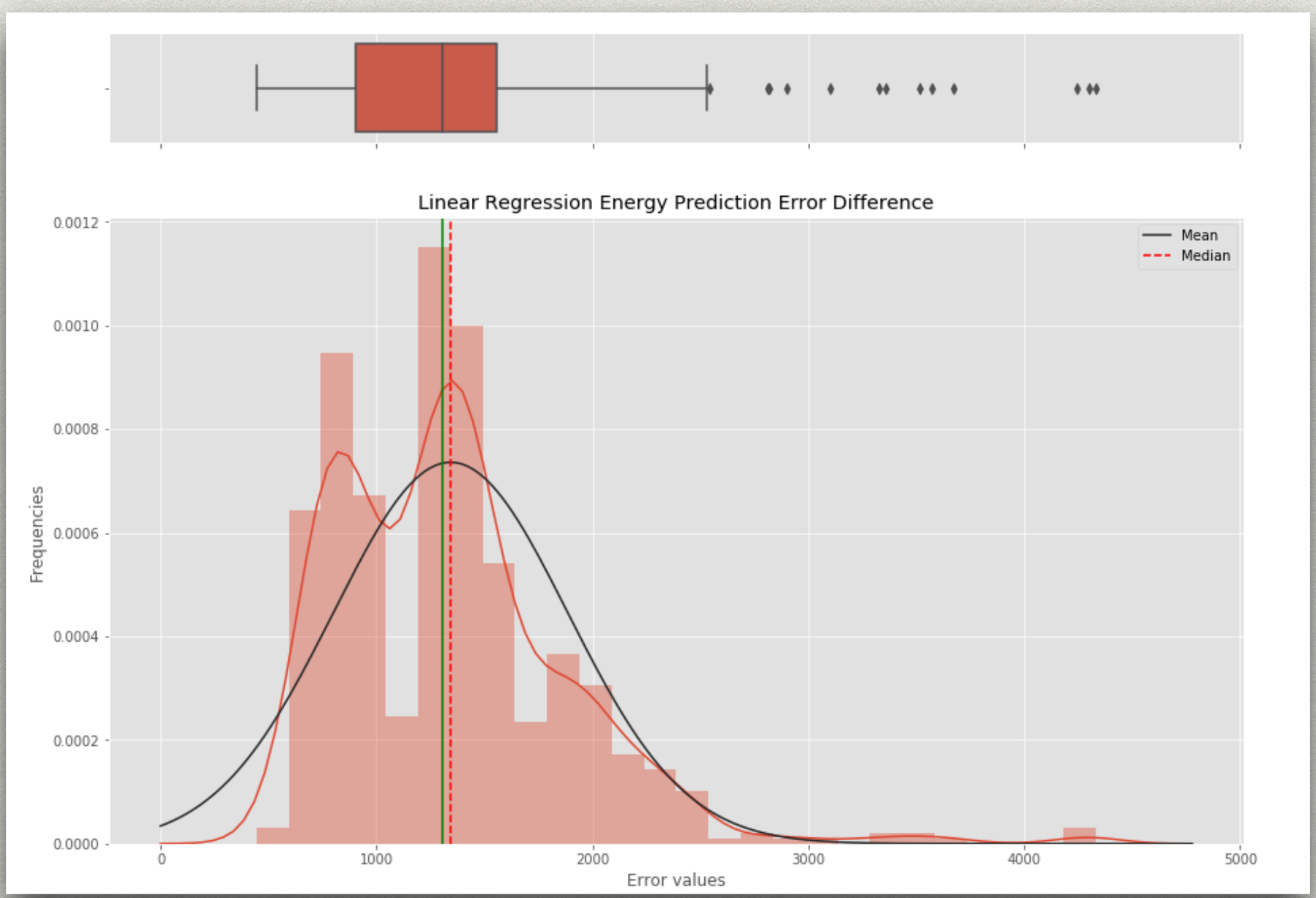
Pour chaque modèles…

- Metric : MSE and MAPE (*mean average percentage error*)
- Séparations des entrainements pour la prédictions de l’énergie ou des émissions de CO2.
- Analyse de la distribution des erreurs de prédictions par rapport aux valeurs réelles.

Régression

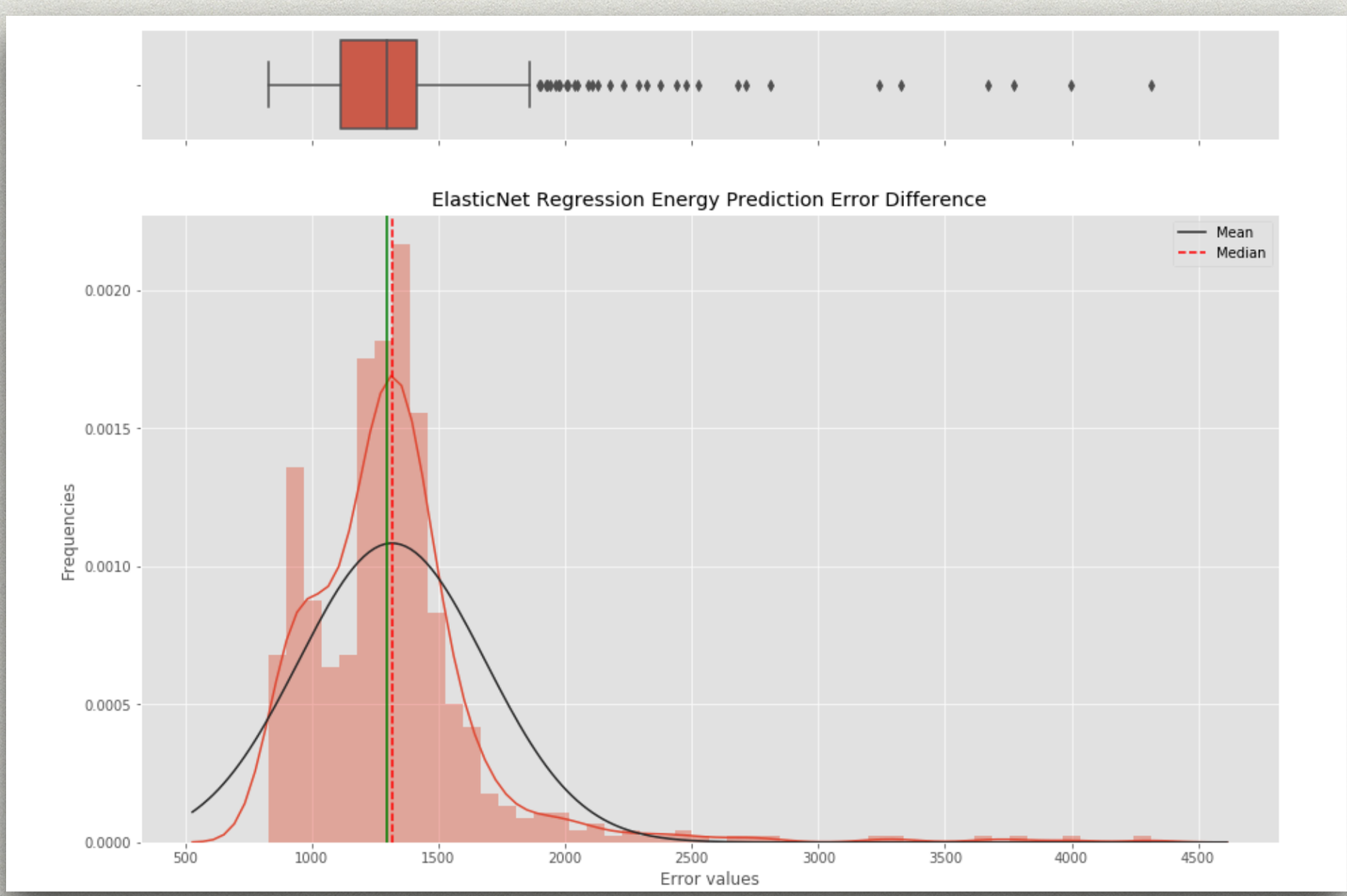
- Une régression *linéaire*
- Une régression « *ElasticNet CV* »
- Testé sur la prédictions d'émissions de CO2

Linear Regression



- MSE :1447.72
- Mean : 1342.36, Median : 1307.62

Elastic CV Regression



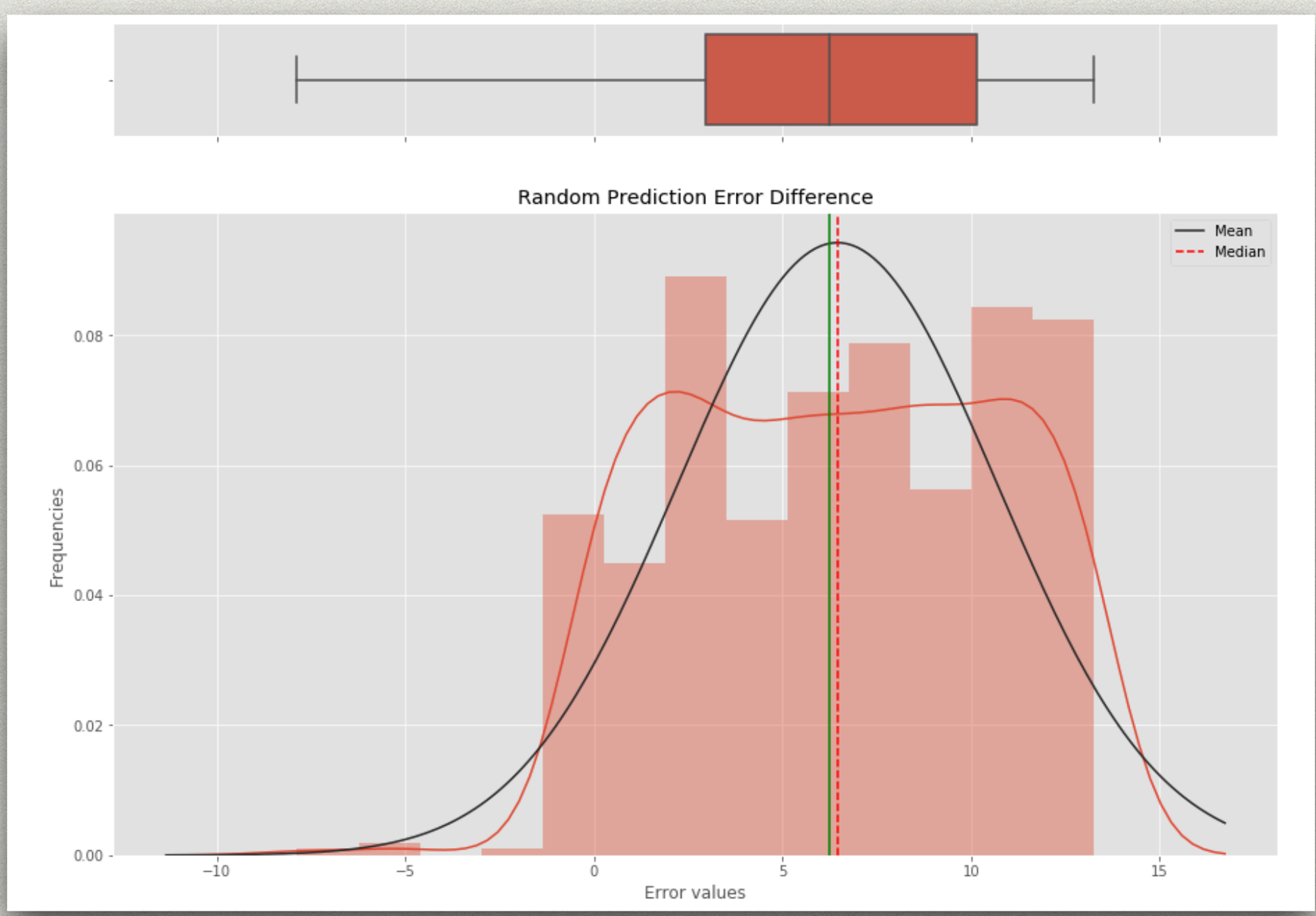
- MSE :1366.89
- Mean : 1316.35, Median : 1295.60

K-Nearest Neighbors

- Un modèle KNN « *aléatoire* »
- Un modèle KNN « *mean-dummy* »
- Un modèle KNN régulier

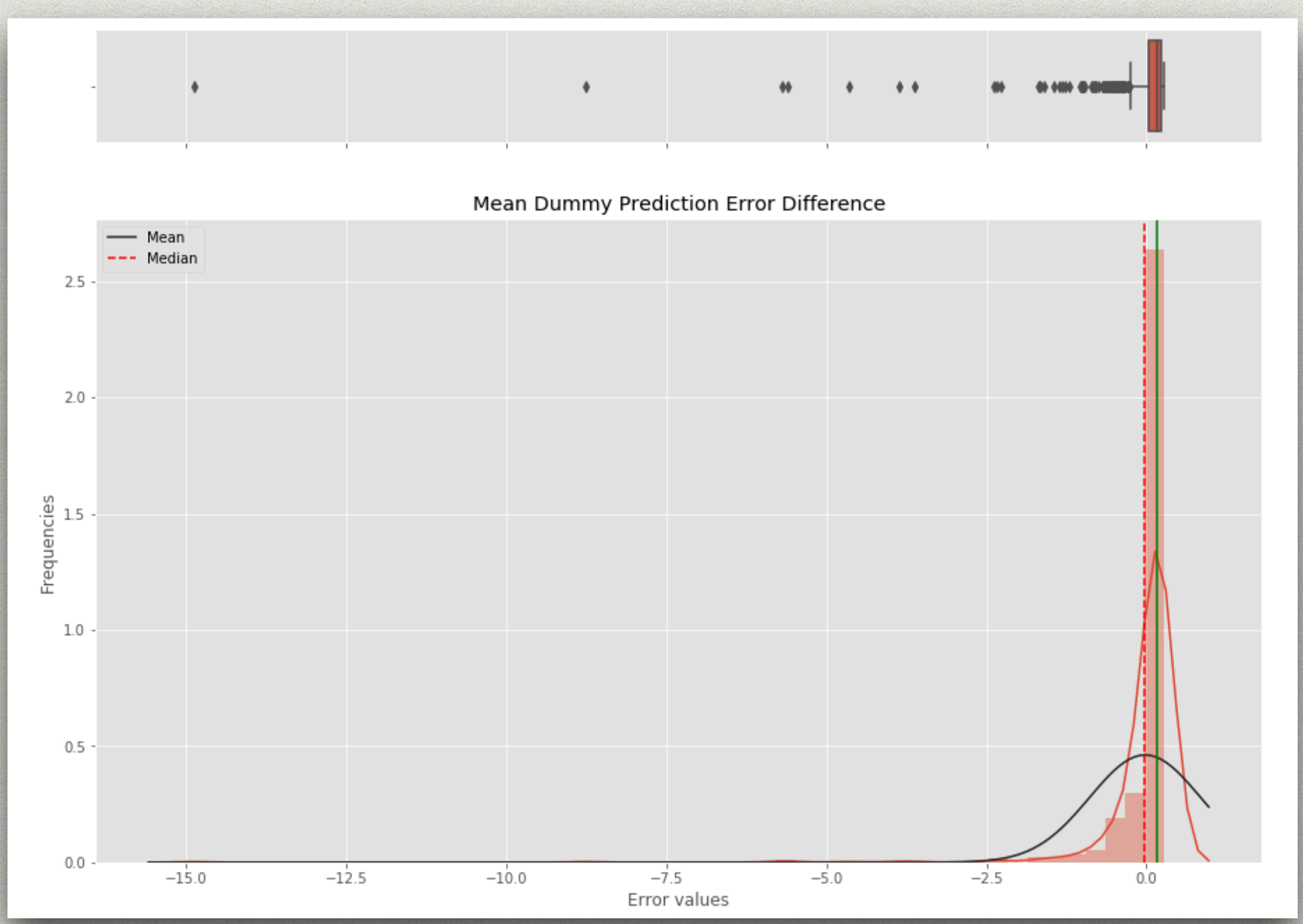
- Testé sur la prédictions d'émissions de CO2

Random k-Nearest Neighbors



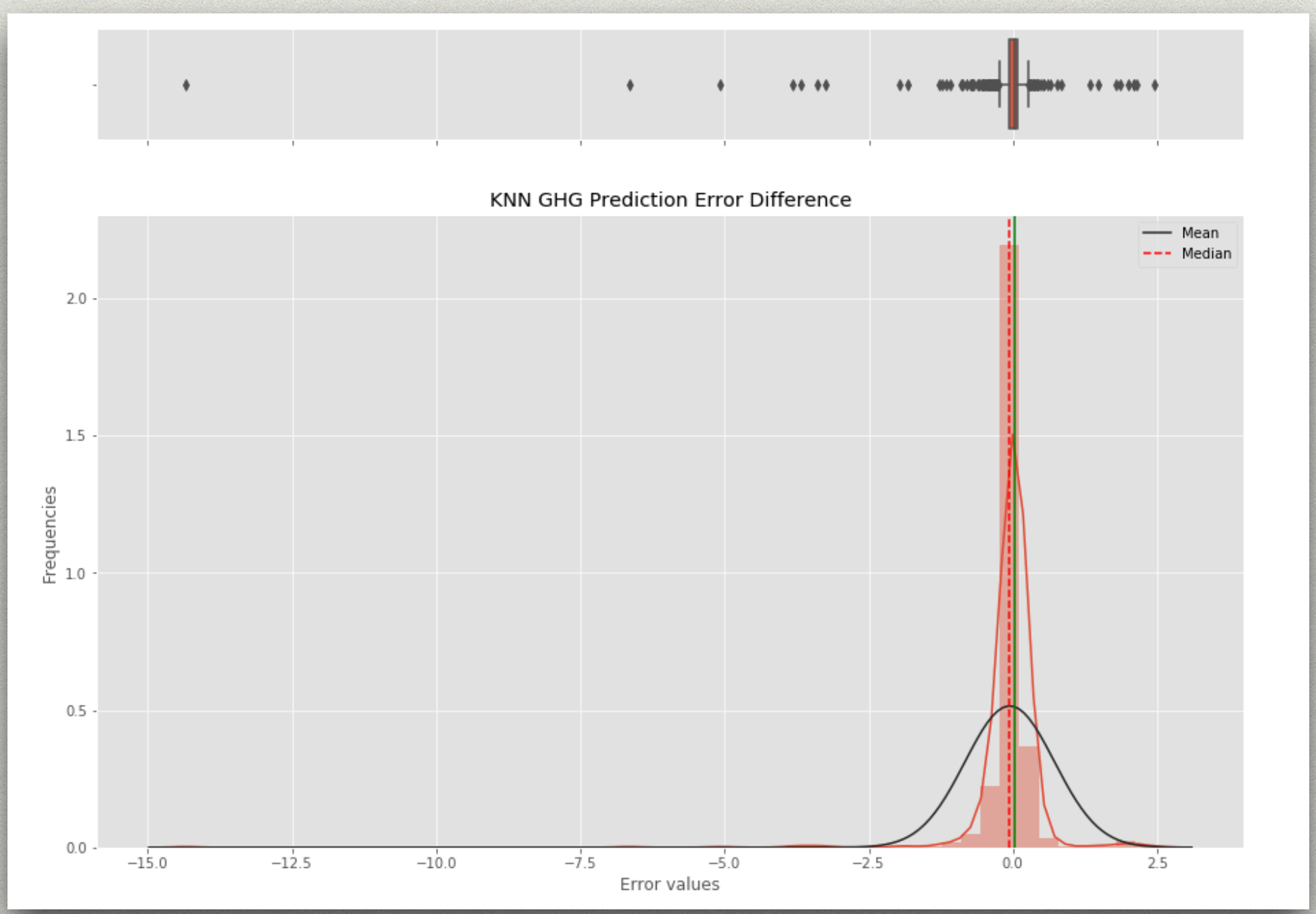
- MSE :7.720
- Mean : 6.457, Median : 6.252

Mean dummy k-Nearest Neighbors



- MSE :0.864
- Mean : -0.021, Median : 0.173

k-Nearest Neighbors

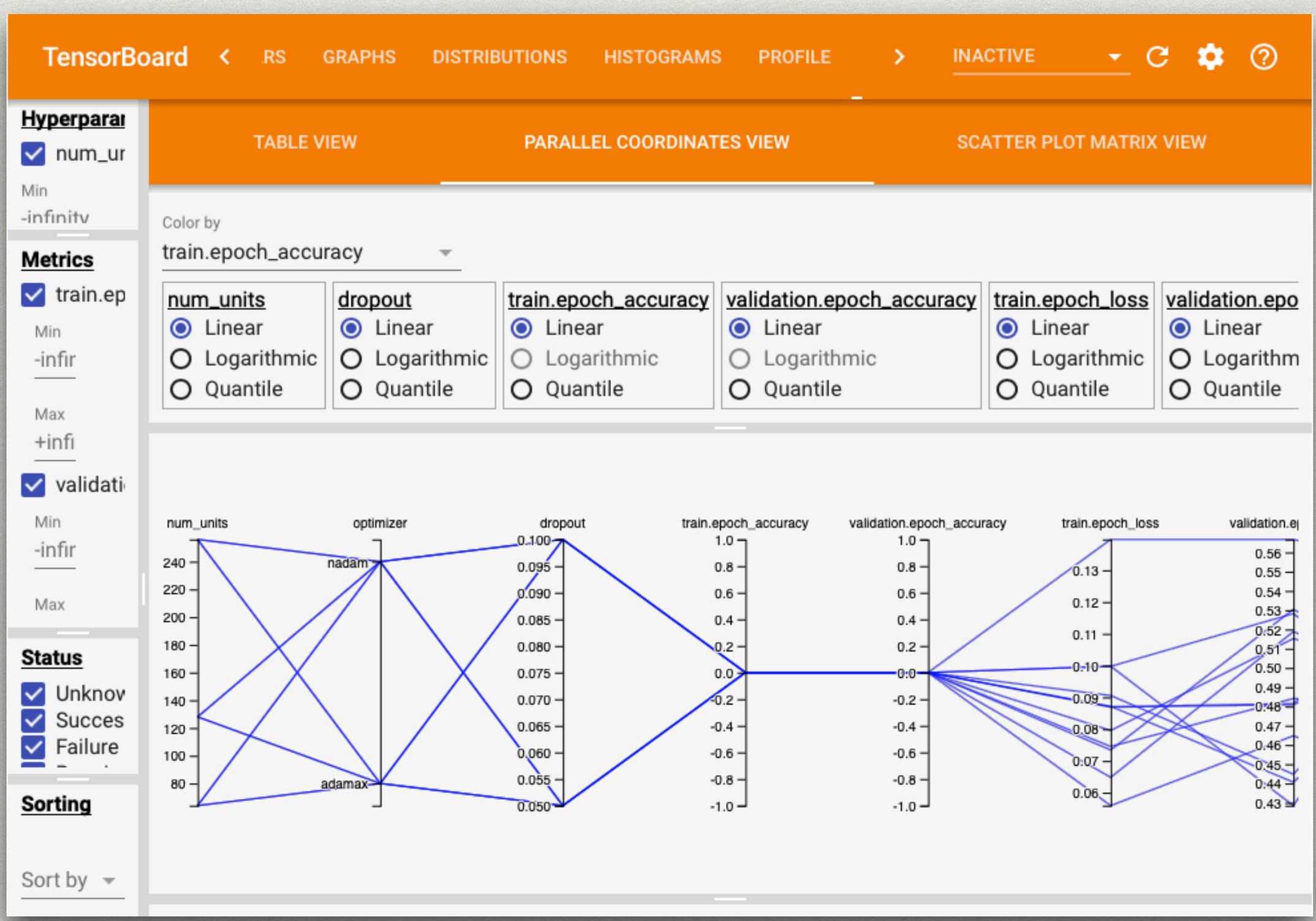


- MSE :0.776
- Mean : -0.056, Median : 0.023

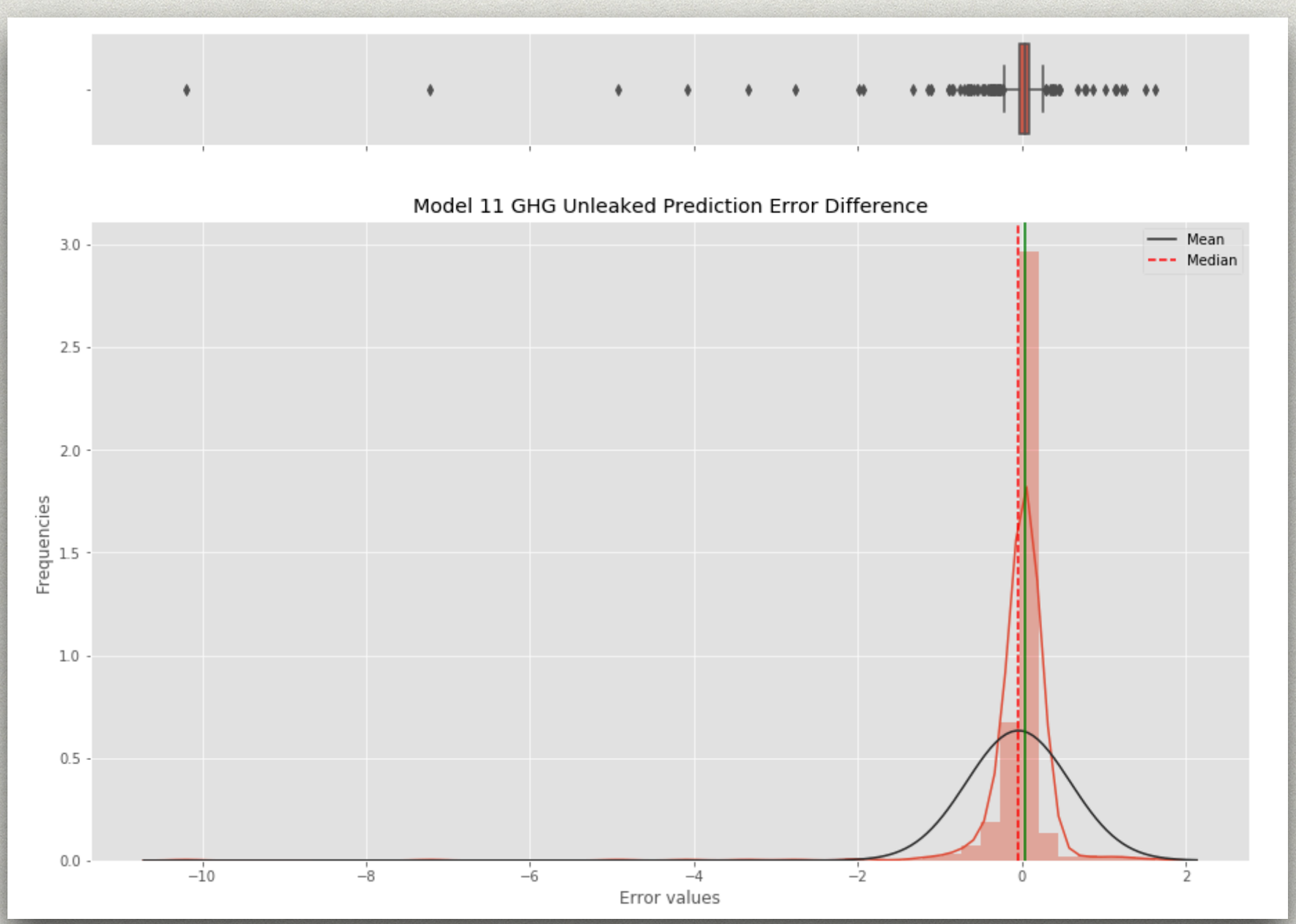
Deep Neural Networks

- Réseau neuronal de 5 couches pour 1 output
- Initialisation uniforme aléatoire des poids
- Activation « *relu* »
- Learning rate dépressif selon les epochs d'entrainement
- 12 combinaison d'hyper-paramètres on été testé
    - 3 valeurs d'unité initiale
    - 2 valeurs de dropout
    - 2 types d'optimizer
- Testé sur la prédictions d'émissions de CO2

TensorBoard



	description	test_mse
11	Model 11	0.633034
2	Model 2	0.646763
0	Model 0	0.656558
9	Model 9	0.657735
3	Model 3	0.663251
1	Model 1	0.671016
6	Model 6	0.673319
7	Model 7	0.678755
5	Model 5	0.684215
10	Model 10	0.692360
4	Model 4	0.693200
8	Model 8	0.709363



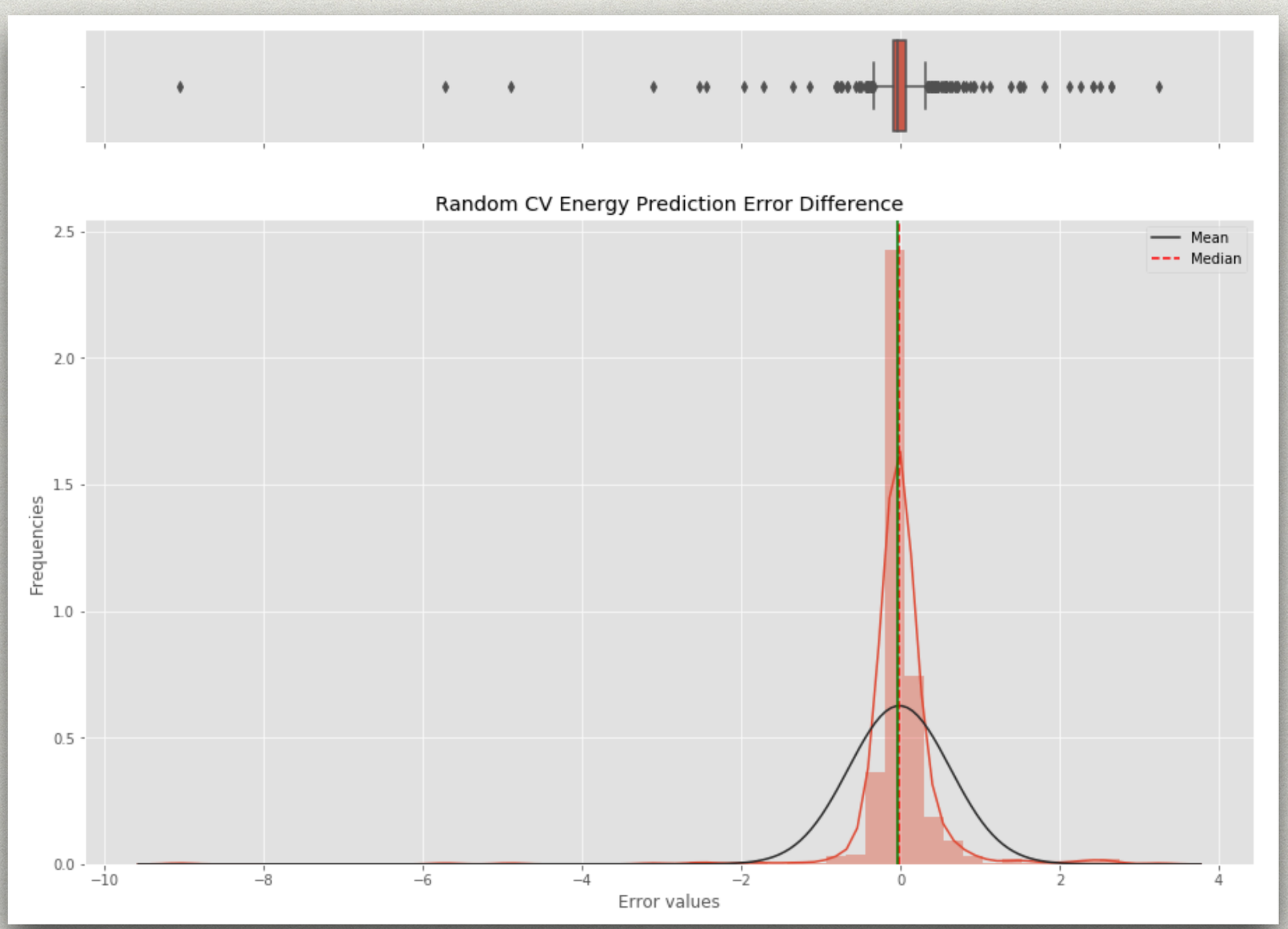
- MSE : 0.633
- Mean : -0.043, Median : 0.040

MODÉL FINAL

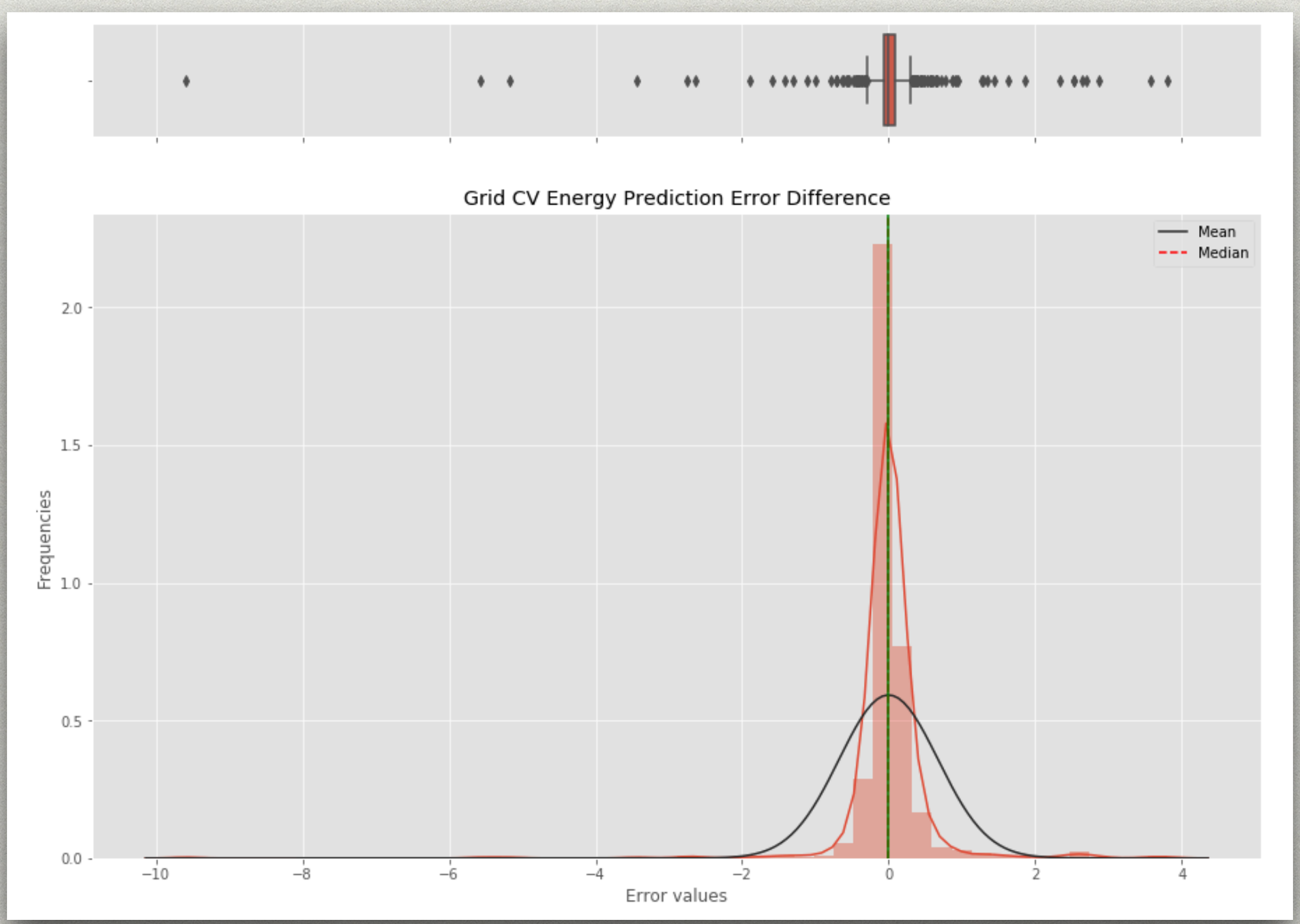
- Présentation du modele final sélectionné + ameliorations apportées

Random Deep Forest

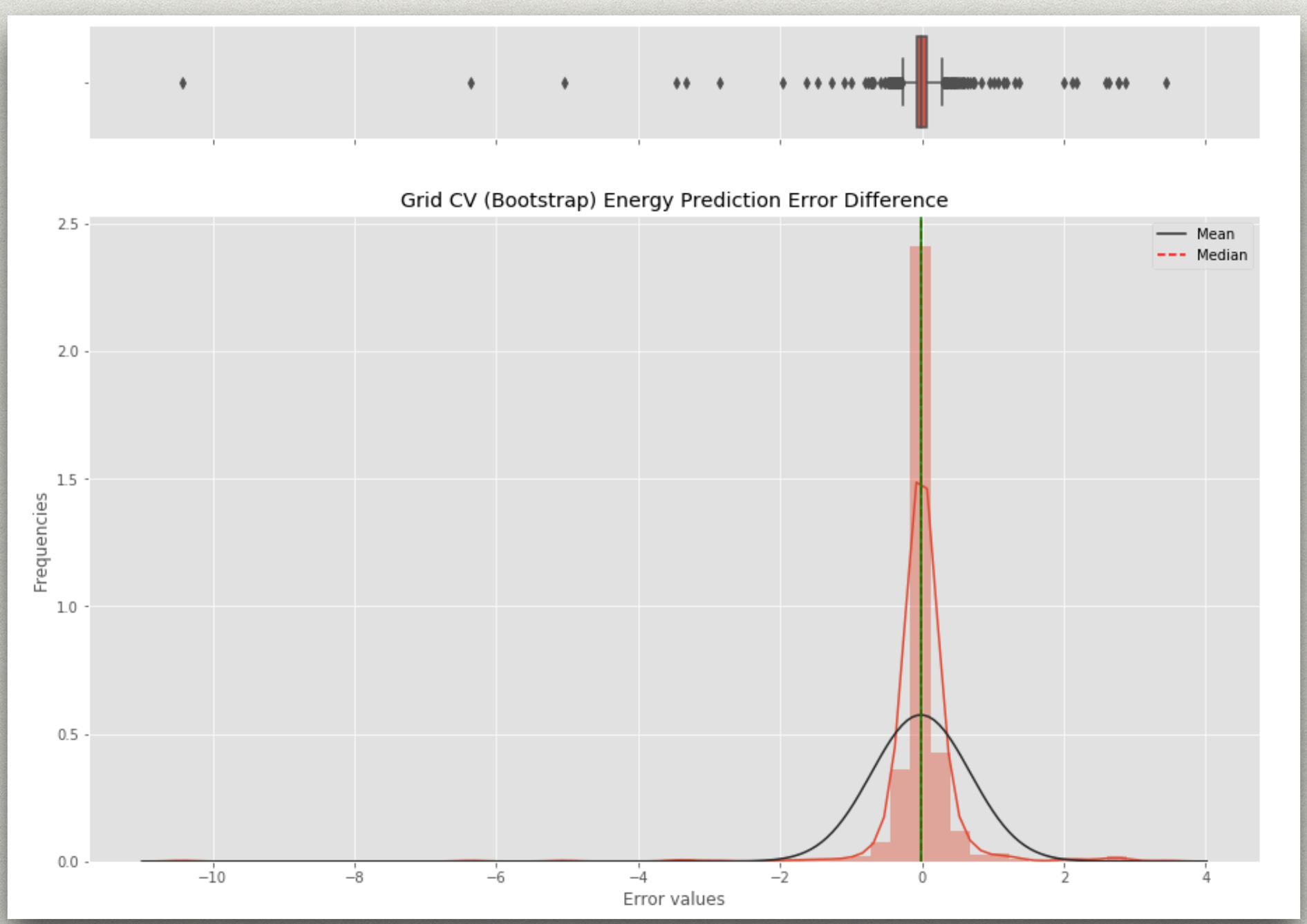
- Deux type de regression: Random CV et Grid CV.
- Utilisation du KFold pour enlever du biais dans l'entrainement.
- Combinaison de différents hyper-paramètres sur le bootstrap, le minimum de branches et de feuilles.
- Entrainement sur 100 iterations.



- MSE : 0.448
- Mean : -0.014, Median : -0.033

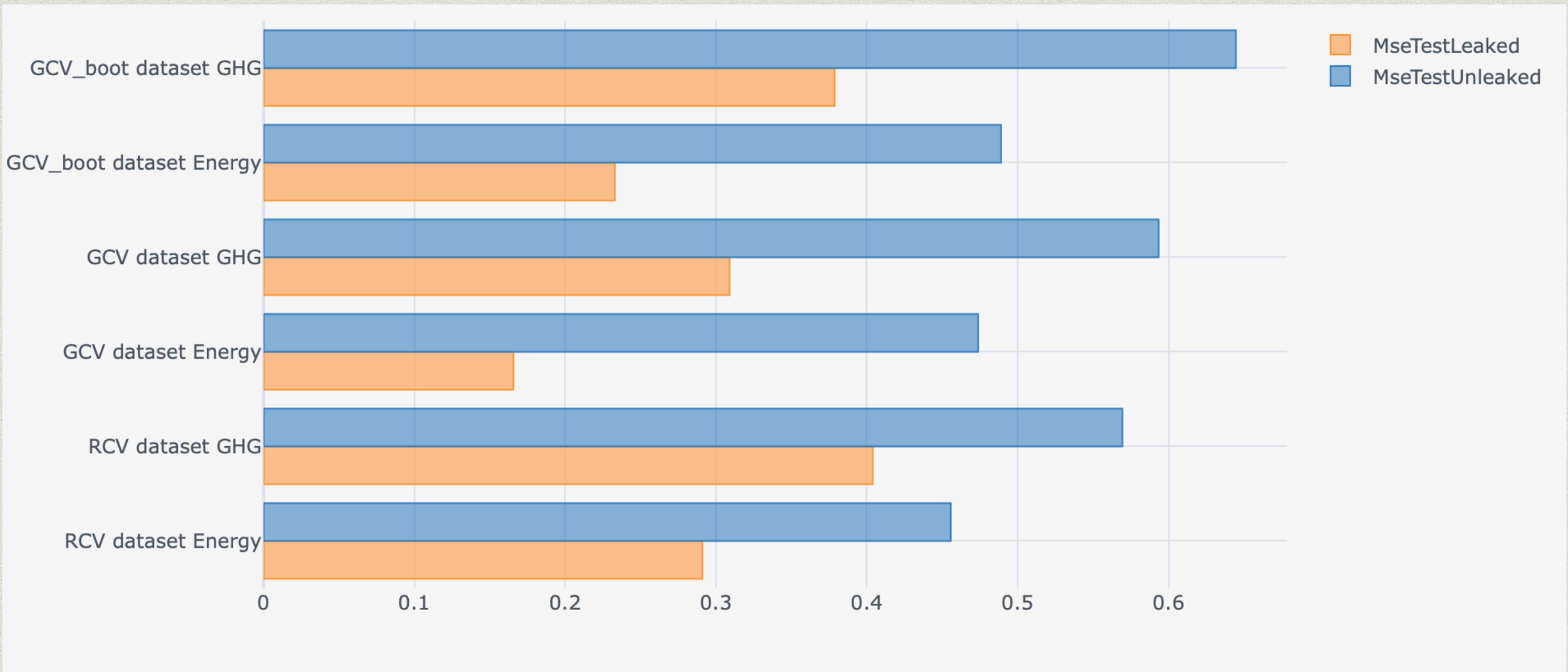


- MSE : 0.486
- Mean : -0.002, Median : -0.009



- MSE : 0.503
- Mean : -0.016, Median : -0.033

Étude des deux axes principaux de prédictions



Améliorations

- Comparaison des différentes combinaison d’hyper-paramètres
- Utilisation de FairML pour une évaluation de l’importance de chaque features pour le model.
- Manipulation à la main de quelques hyper-paramètres tel que le nombres d’itérations.