Witold Drzewakowski
XAI - hw 5, mimuw, 16.11.2023

Task 1

$f(x_1, x_2) = (x_1 + x_2)^2.$

Partial Dependence

$$g_1^{PD}(z) = \mathbb{E}[f(z, X_2)] = \mathbb{E}[(z + X_2)^2] = z^2 + 2z\mathbb{E}[X_2] + \mathbb{E}[X_2^2] = z^2 + \frac{1}{3},$$

because $\mathbb{E}[X_2] = 0$ and $\mathbb{E}[X_2^2] = \int_{-1}^{1} \frac{1}{2}x^2\,dx = \frac{1}{6}x^3|_{-1}^{1} = \frac{1}{3}$.

Marginal Effects

$$g_1^{ME}(z) = \mathbb{E}[f(z, X_2)|X_1 = z] = z^2 + 2z\mathbb{E}[X_2|X_1 = z] + \mathbb{E}[X_2^2|X_1 = z] = 4z^2.$$

Accumulated Local Effects

$$g_1^{AL}(z) = \int_{-1}^{z} \mathbb{E}[\frac{\partial f}{\partial v}(v, X_2)|X_1 = v]\,dv = \int_{-1}^{z} \mathbb{E}[2v + 2X_2|X_1 = v]\,dv =$$
$$\int_{-1}^{z} 4v\,dv = 2v^2|_{v=-1}^{v=z} = 2z^2 - 2.$$

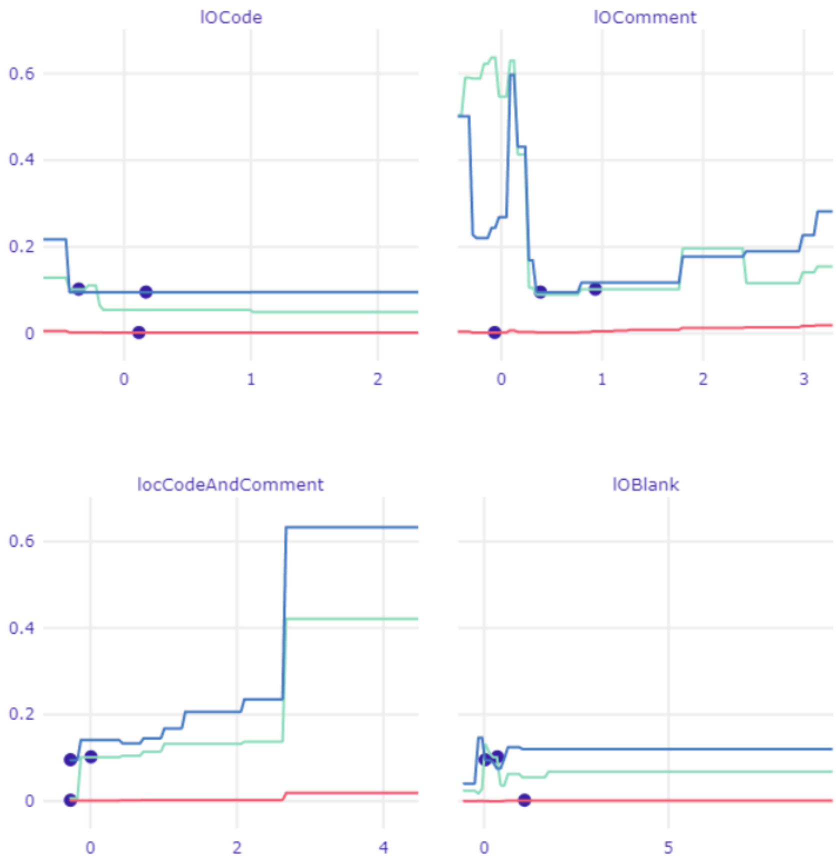
Task 2

In this task I explain the prediction of XGBoost (and a MLPClassifier in the last part) of dataset `pc1` from OpenML, using Ceteris Paribus and Partial Dependence profiles as implemented in dalex. In both figures I will present, in order to improve readability, I will display explanations for only 4 variables from the dataset. The full figures are available in `` `imgs_hw5` `` subdirectory.

In the first figure, we can see Ceteris Paribus plots for 4 selected observations. We can see that in thoseparticular examples, quite often the dynamics of those profiles are similar, but there are significant differences, e.g. for the variable IOComment, especially for its smaller values.

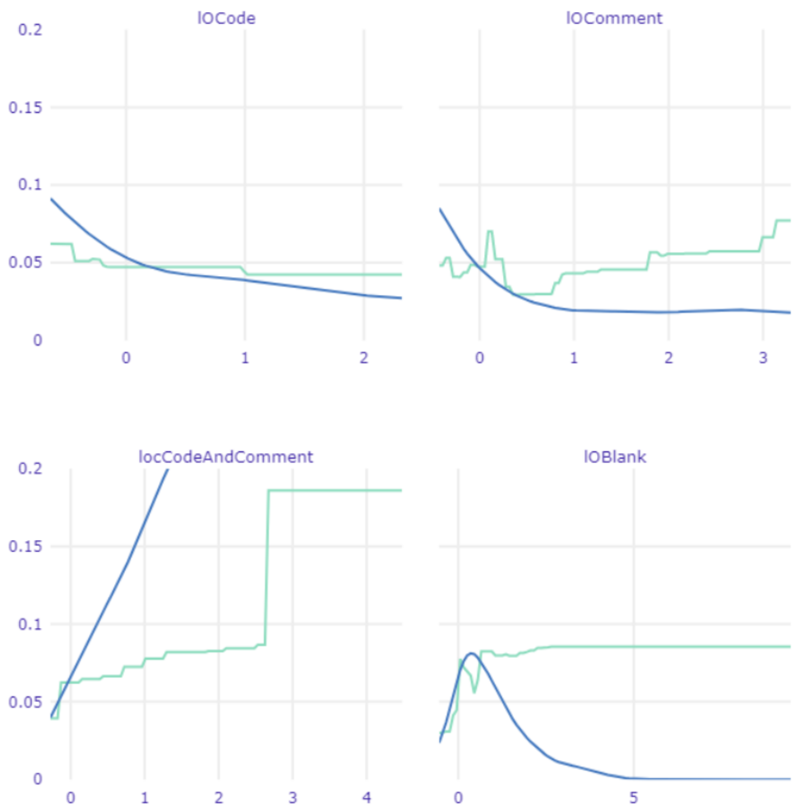
The second figure demonstrates the Partial Dependence profiles for XGBoost, MLPClassifier models trained on the same dataset. Even if we see a certain level of similarity, clearly the profiles for MLPClassifier are much smoother, as anticipated.

Fig 1:



lOCode
lOComment
locCodeAndComment
lOBlank
0.6
0.4
0.2
0
0
1
2
0
1
2
3
0.6
0.4
0.2
0
0
2
4
0
5

Fig 2:



lOCode
lOComment
locCodeAndComment
lOBlank
0.2
0.15
0.1
0.05
0
0
1
2
0
1
2
3
0.2
0.15
0.1
0.05
0
0
1
2
3
4
0
5