Solution 1:

(a) Which of the following statement(s) apply to feature effect methods?

(i) The value of the PDP at a point x_j, corresponds to the point-wise average of the values of the ICE curves at this point. ⇒ **Correct**

(ii) The PDP of a feature provides information about possible interaction effects of the feature. ⇒ **Wrong**

(iii) ICE curves of a feature for multiple data points provide information about possible interaction effects of the feature with others. ⇒ **Correct**

(iv) If we center the ICE/PDPs for categorical features, the expected changes always refer to a selected reference category. ⇒ **Correct**

(v) ALE plots are based on conditional distributions, PDPs on marginal distributions. ⇒ **Correct**

(vi) If features are uncorrelated, ALE plots are equal to PDPs. ⇒ **Wrong**

(vii) ALE plots are faster to compute than PDPs if they are based on the same grid. ⇒ **Correct**

(b) You fitted a model that should predict the value of a property depending on the number of rooms and square meters. You want to compute feature effects using the following methods: PDP, M-plots and (uncentered) ALE plots. Which of the following strategies reflect which method?
The feature effect for a 30 m^2 corresponds to...

a) ... what the model predicts on average for flats that also have around 30 m^2, for example, 28 m^2 to 32 m^2. ⇒ **M-plot**

b) ... how the model's predictions change on average when flats with 28 m^2 to 32 m^2 have 32 m^2 vs. 28 m^2. ⇒ **uncentered ALE**

c) ... what the model predicts on average if all properties in the dataset have 30 m^2. ⇒ **PDP**

Solution 2:

(a) Both PDP and ALE plots show a strong linear effect of x_1, where higher values of x_1 lead to higher values of predicted value. The PDP and ALE plot of x_2 show a strong decreasing effect of x_2 on the prediction. The PDP of x_2 shows a steep jump for large values of x_2, while the ALE plot shows a strong linear effect over the whole value range of x_2. Interpretation at $x_1 = 0.5$:

- PDP: the model predicts on average a value of around 2.3 for y if for all data instances $x_1 = 0.5$.
- ALE: the model predicts on average an increase of around 2.5 of y for data instances with $x_1 = 0.5$ compared to the average prediction.

(b) PDPs assume that features are uncorrelated. We know from the GAM output above - as well as the scatter plot - that x_1 and x_2 are highly correlated. Since PDPs extrapolate over predictions of artificial points that are out of distribution, the interpretations might be misleading - especially in areas with low data density (high values of x_2) and if the model contains interactions. ALE on the other hand, does not predict in regions that are far away from the training data and therefore do not suffer from the extrapolation issue of PDPs.