

## 7.2 Accumulated Local Effects (ALE) Plot

Accumulated local effects<sup>4</sup> describe how features influence the prediction of a machine learning model on average. ALE plots are a faster and unbiased alternative to partial dependence plots (PDPs).

I recommend reading the chapter on partial dependence plots first, as they are easier to understand and both methods share the same goal: Both describe how a feature affects the prediction on average. In the following section, I want to convince you that partial dependence plots have a serious problem when the features are correlated.

### 7.2.1 Motivation and Intuition

If features of a machine learning model are correlated, the partial dependence plot cannot be trusted. The computation of a partial dependence plot for a feature that is strongly correlated with other features involves averaging predictions of artificial data instances that are unlikely in reality. This can greatly bias the estimated feature effect. Imagine calculating partial dependence plots for a machine learning model that predicts the value of a house depending on the number of rooms and the size of the living area. We are interested in the effect of the living area on the predicted value. As a reminder, the recipe for partial dependence plots is: 1) Select feature. 2) Define grid. 3) Per grid value: a) Replace feature with grid value and b) average predictions. 4) Draw curve. For the calculation of the first grid value of the PDP – say  $30 \text{ m}^2$  – we replace the living area for all instances by  $30 \text{ m}^2$ , even for houses with 10 rooms. Sounds to me like a very unusual house. The partial dependence plot includes these unrealistic houses in the feature effect estimation and pretends that everything is fine. The following figure illustrates two correlated features and how it comes that the partial dependence plot method averages predictions of unlikely instances.

What can we do to get a feature effect estimate that respects the correlation of the features? We could average over the conditional distribution of the feature, meaning at a grid value of  $x_1$ , we average the predictions of instances with a similar  $x_1$  value. The solution for calculating feature effects using the conditional distribution is called Marginal Plots, or M-Plots (confusing name, since they are based on the conditional, not the marginal distribution). Wait, did I not promise you to talk about ALE plots? M-Plots are not the solution we are looking for. Why do M-Plots not solve our problem? If we average the predictions of all houses of about  $30 \text{ m}^2$ , we estimate the combined effect of living area and of number of rooms, because of their correlation. Suppose that the living area has no effect on the predicted value of a house, only the number of rooms has.

accumulate 积累 .

ALE是一种快速的、无歧视的 PDP代替方案

PDP在 features 之间有关联的情况下会有 bias .

unlikely 不可能发生的；  
难以置信的。

怎么 define grid ?

averages prediction ?

Monte Carlo Method.

→ Fig 7.5

<sup>4</sup>Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086.

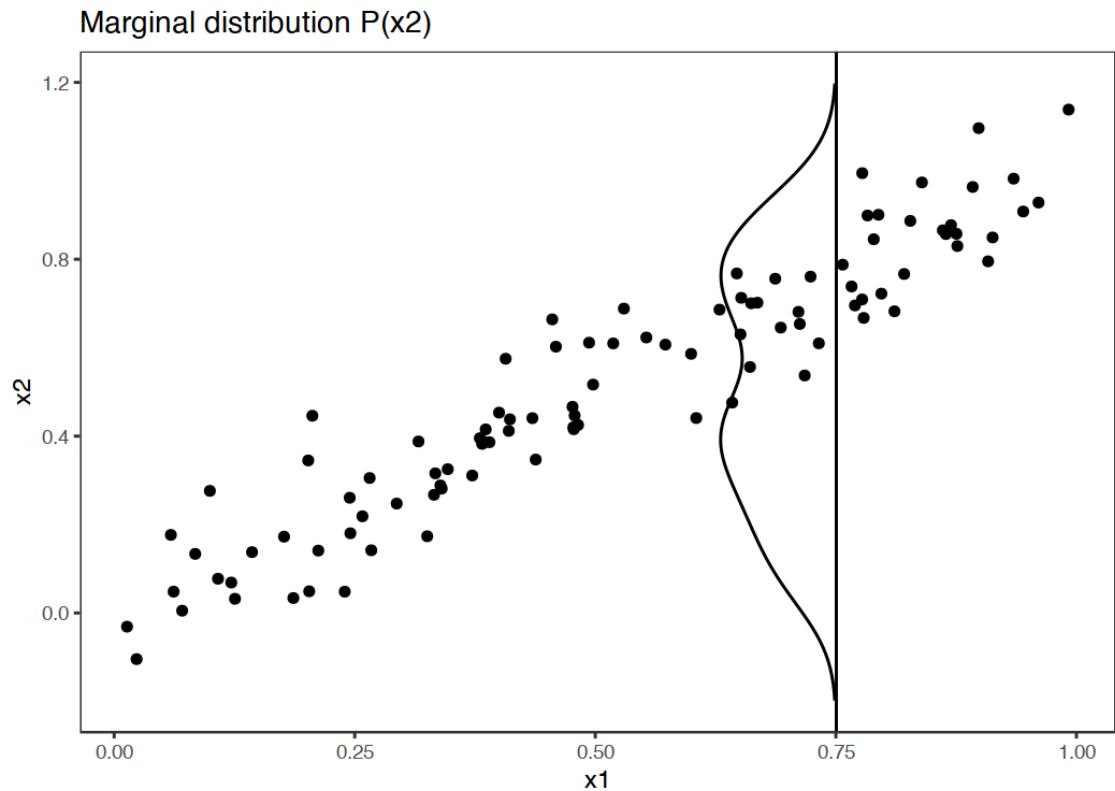


Figure 7.5: Strongly correlated features  $x_1$  and  $x_2$ . To calculate the feature effect of  $x_1$  at 0.75, the PDP replaces  $x_1$  of all instances with 0.75, falsely assuming that the distribution of  $x_2$  at  $x_1 = 0.75$  is the same as the marginal distribution of  $x_2$  (vertical line). This results in unlikely combinations of  $x_1$  and  $x_2$  (e.g.  $x_2=0.2$  at  $x_1=0.75$ ), which the PDP uses for the calculation of the average effect.

The M-Plot would still show that the size of the living area increases the predicted value, since the number of rooms increases with the living area. The following plot shows for two correlated features how M-Plots work.

M-Plots avoid averaging predictions of unlikely data instances, but they mix the effect of a feature with the effects of all correlated features. ALE plots solve this problem by calculating – also based on the conditional distribution of the features – differences in predictions instead of averages. For the effect of living area at  $30 \text{ m}^2$ , the ALE method uses all houses with about  $30 \text{ m}^2$ , gets the model predictions pretending these houses were  $31 \text{ m}^2$  minus the prediction pretending they were  $29 \text{ m}^2$ . This gives us the pure effect of the living area and is not mixing the effect with the effects of correlated features. The use of differences blocks the effect of other features. The following graphic provides intuition how ALE plots are calculated.

为了计算  $x_1$  在 0.75 上的效应，  
PDP 用 0.75 替换所有样本中的  $x_1$ ，  
它错误地假设了  $x_2$  在  $x_1 = 0.75$  处  
的分布与  $x_2$  的边际分布相同（垂直  
线），这导致了不太可能的  $x_1$  与  $x_2$   
组合 ( $x_1 = 0.75, x_2 = 0.2$ )，PDP 仍  
将其用于计算平均效应。

Fig 7.6

ALE 通过计算 预测值之间的  
差异 而不是其平均值来解决  
features 之间的交互影响  
要得到  $30 \text{ m}^2$  的影响，ALE 用  
假装所有的  $30 \text{ m}^2$  的该 feature  
数值是  $31 \text{ m}^2$ ，得出的预测  
减去数值是  $29 \text{ m}^2$  的预测值

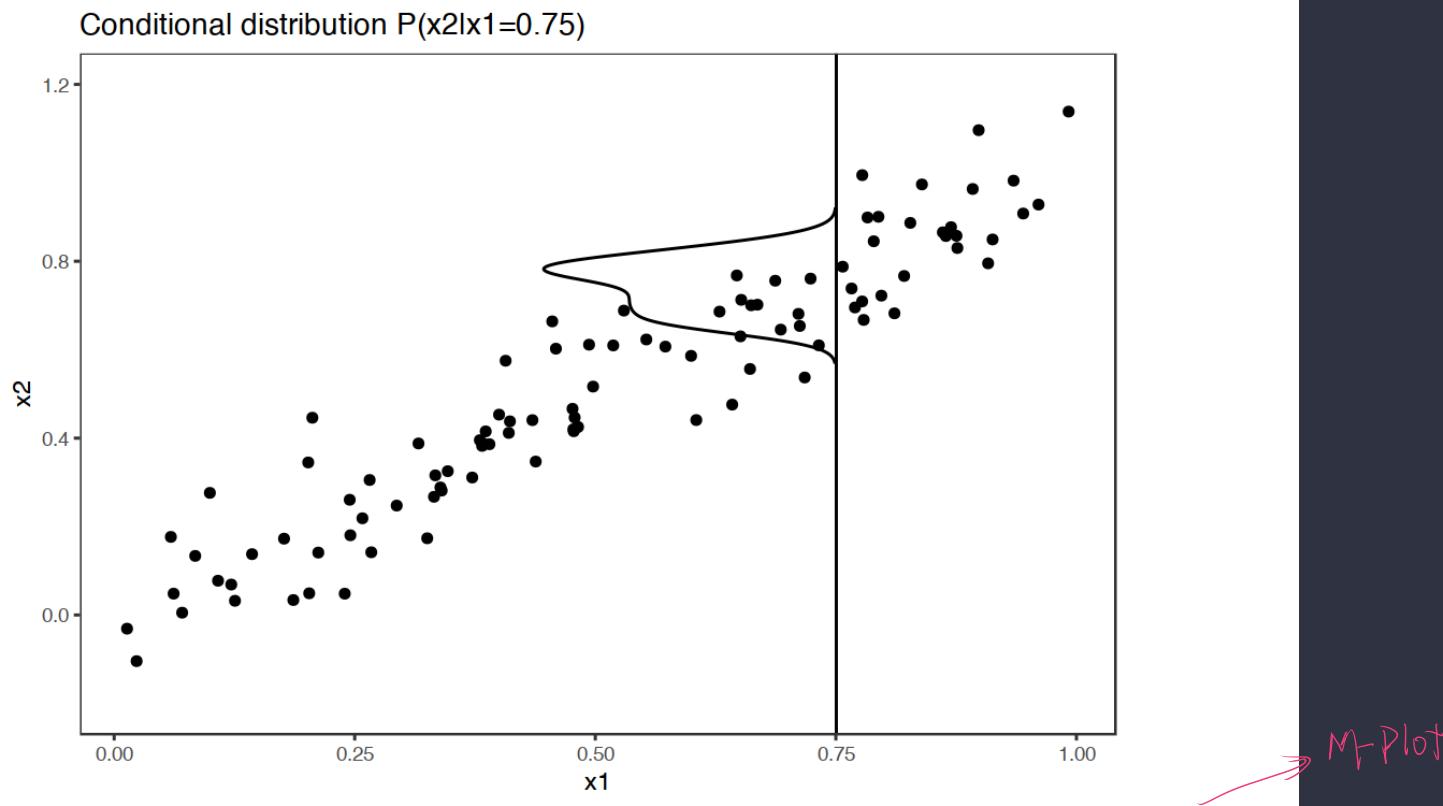


Figure 7.6: Strongly correlated features  $x_1$  and  $x_2$ . M-Plots average over the conditional distribution. Here the conditional distribution of  $x_2$  at  $x_1 = 0.75$ . Averaging the local predictions leads to mixing the effects of both features.

To summarize how each type of plot (PDP, M, ALE) calculates the effect of a feature at a certain grid value  $v$ :

**Partial Dependence Plots:** “Let me show you what the model predicts on average when each data instance has the value  $v$  for that feature. I ignore whether the value  $v$  makes sense for all data instances.”

**M-Plots:** “Let me show you what the model predicts on average for data instances that have values close to  $v$  for that feature. The effect could be due to that feature, but also due to correlated features.”

**ALE plots:** “Let me show you how the model predictions change in a small “window” of the feature around  $v$  for data instances in that window.”

在  $v$  的窗口范围内以窗口  
展示预测的变化

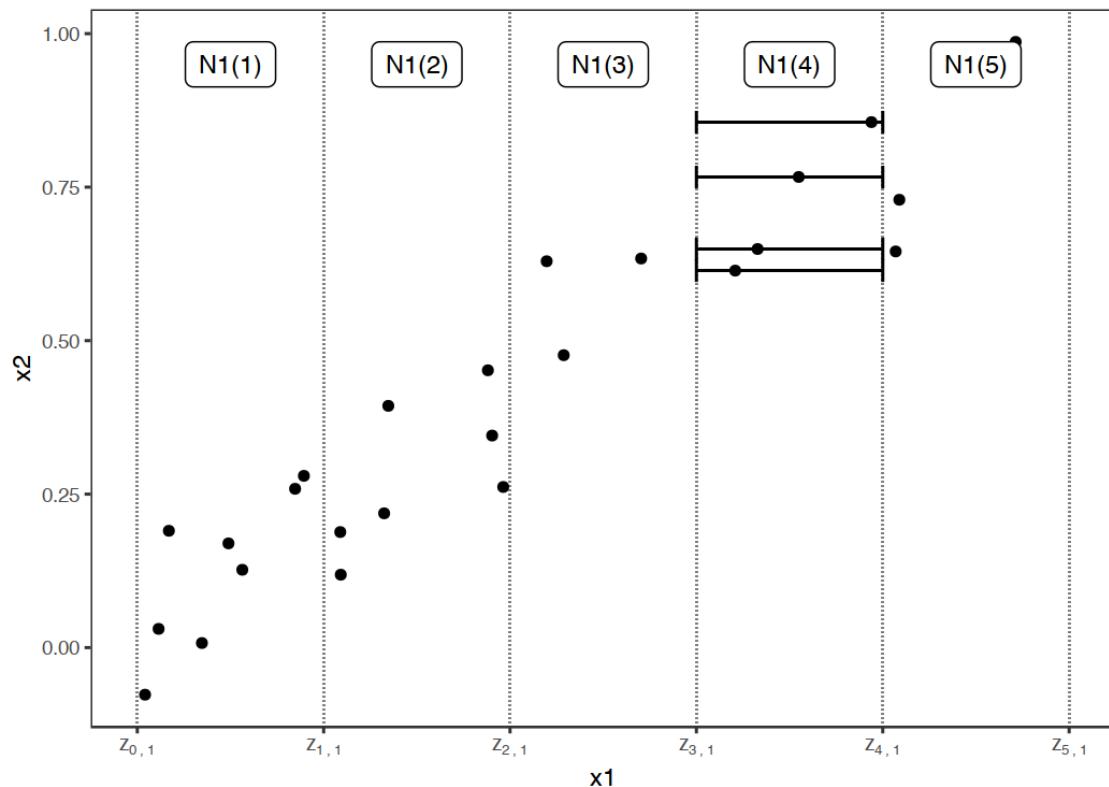


Figure 7.7: Calculation of ALE for feature  $x_1$ , which is correlated with  $x_2$ . First, we divide the feature into intervals (vertical lines). For the data instances (points) in an interval, we calculate the difference in the prediction when we replace the feature with the upper and lower limit of the interval (horizontal lines). These differences are later accumulated and centered, resulting in the ALE curve.

## 7.2.2 Theory

How do PD, M and ALE plots differ mathematically? Common to all three methods is that they reduce the complex prediction function  $f$  to a function that depends on only one (or two) features. All three methods reduce the function by averaging the effects of the other features, but they differ in whether averages of predictions or of **differences in predictions** are calculated and whether averaging is done over the marginal or conditional distribution.

Partial dependence plots average the predictions over the marginal distribution.

首先我们将 feature 的 value 分区，在分区内的数据就用两端的值算出预测差。这些差值随后积累并中心化，形成 ALE 曲线。

PDP、M-Plot、ALE 的区别：

- ① 在预测值上取平均还是在预测差上取
- ② 边缘分布还是条件分布

PDP 是边缘分布的预测上取平均

$$\begin{aligned}\hat{f}_{S,PDP}(x) &= E_{X_C} [\hat{f}(x_S, X_C)] \\ &= \int_{X_C} \hat{f}(x_S, X_C) d\mathbb{P}(X_C)\end{aligned}$$

This is the value of the prediction function  $f$ , at feature value(s)  $x_S$ , averaged over all features in  $X_C$  (here treated as random variables). Averaging means calculating the marginal expectation  $E$  over the features in set  $C$ , which is the integral over the predictions weighted by the probability distribution. Sounds fancy, but to calculate the expected value over the marginal distribution, we simply take all our data instances, force them to have a certain grid value for the features in set  $S$ , and average the predictions for this manipulated dataset. This procedure ensures that we average over the marginal distribution of the features.

M-plots average the predictions over the conditional distribution.

$$\begin{aligned}\hat{f}_{S,M}(x_S) &= E_{X_C|X_S} [\hat{f}(X_S, X_C)|X_S = x_S] \\ &= \int_{X_C} \hat{f}(x_S, X_C) d\mathbb{P}(X_C|X_S = x_S)\end{aligned}$$

The only thing that changes compared to PDPs is that we average the predictions conditional on each grid value of the feature of interest, instead of assuming the marginal distribution at each grid value. In practice, this means that we have to define a neighborhood, for example for the calculation of the effect of  $30 \text{ m}^2$  on the predicted house value, we could average the predictions of all houses between  $28$  and  $32 \text{ m}^2$ .

ALE plots average the changes in the predictions and accumulate them over the grid (more on the calculation later).

$$\begin{aligned}\hat{f}_{S,ALE}(x_S) &= \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S} [\hat{f}^S(X_s, X_c)|X_S = z_S] dz_S - \text{constant} \\ &= \int_{z_{0,S}}^{x_S} (\int_{x_C} \hat{f}^S(z_S, X_c) d\mathbb{P}(X_C|X_S = z_S)) dz_S - \text{constant}\end{aligned}$$

The formula reveals three differences to M-Plots. First, we average the changes of predictions, not the predictions itself. The change is defined as the partial derivative (but later, for the actual computation, replaced by the differences in the predictions over an interval).

平均意味着计算非被测feature集的边缘期待

M-Plot 在条件分布上的预测值取平均。

ALE 在预测变化之中求平均，并将它们累积在网格上。

ALE 和 M-Plot 的区别：

① 在预测变化上取平均而不是预测本身。

理论上，预测变化被定义为部分导数 partial derivative，但在实际计算中可以用预测在区间上的差异来代替。

$$\hat{f}^S(x_s, x_c) = \frac{\partial \hat{f}(x_S, x_C)}{\partial x_S}$$

The second difference is the additional **integral** over z. We accumulate the local partial derivatives over the range of features in set S, which gives us the effect of the feature on the prediction. For the actual computation, the z's are replaced by a grid of intervals over which we compute the changes in the prediction. Instead of directly averaging the predictions, the ALE method calculates the prediction differences conditional on features S and **integrates the derivative over features S to estimate the effect**. Well, that sounds stupid. Derivation and integration usually cancel each other out, like first subtracting, then adding the same number. Why does it make sense here? The derivative (or interval difference) isolates the effect of the feature of interest and blocks the effect of correlated features.

The third difference of ALE plots to M-plots is that we subtract a constant from the results. This step centers the ALE plot so that the average effect over the data is zero.

**One problem remains:** Not all models come with a gradient, for example random forests have no gradient. But as you will see, the actual computation works without gradients and uses intervals. Let us delve a little deeper into the estimation of ALE plots.

### 7.2.3 Estimation

First I will describe how ALE plots are estimated for a single numerical feature, later for two numerical features and for a single categorical feature. To estimate local effects, we **divide the feature into many intervals and compute the differences in the predictions**. This procedure approximates the derivatives and also works for models without derivatives.

First we estimate the uncentered effect:

$$\hat{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} [\hat{f}(z_{k,j}, x_j^{(i)}) - \hat{f}(z_{k-1,j}, x_j^{(i)})]$$

Let us break this formula down, starting from the right side. The name **Accumulated Local Effects** nicely reflects all the individual components of this formula. At its core, the ALE method calculates the differences in predictions, **whereby we replace the feature of interest with grid values z**. The difference in prediction is the **Effect** the feature has for an individual instance in a certain interval. The sum on the right adds up the effects of all instances within an interval which appears in the formula as neighborhood  $N_j(k)$ . We divide this sum by the number of instances in this interval to obtain the average difference of the predictions for this interval. This average in the interval is covered by the term

integral 积分.

integrate 求...积分.

区别 2：

局部偏导数在S集的范围上累  
积. 实际计算上, 它被网格区  
间所代替.

ALE有条件地计算 feature S  
上的预测差异, 而对 S 的  
导数求积分来估算 effect.

区别 3：

ALE 去除结果中的一个常数.

这让 ALE 的 plot 中心化.

问题: 不能梯度的 model 怎么办?  
可以用区间代替.

将 feature 值分在

$x_j^{(i)}$ : value of feature j of  
the single instance i

Step:

1. 用网格值去替换被测  
feature 的值.

2. 高区预测值减去低区预  
测值得到的差即是该区间  
在该区间的 effect

3. 区间内所有实测的 effect  
求和, 再除以实例数量.

这个区间平均 effect 代表了 Local

**Local** in the name ALE. The left sum symbol means that we accumulate the average effects across all intervals. The (uncentered) ALE of a feature value that lies, for example, in the third interval is the sum of the effects of the first, second and third intervals. The word **Accumulated** in ALE reflects this.

This effect is centered so that the mean effect is zero.

$$\hat{f}_{j,ALE}(x) = \tilde{f}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^n \tilde{f}_{j,ALE}(x_j^{(i)})$$

减去期望值  
使其中心化

The value of the ALE can be interpreted as the main effect of the feature at a certain value compared to the average prediction of the data. For example, an ALE estimate of -2 at  $x_j = 3$  means that when the j-th feature has value 3, then the prediction is lower by 2 compared to the average prediction.

The **quantiles** of the distribution of the feature are used as the grid that defines the intervals. Using the quantiles ensures that there is the same number of data instances in each of the intervals. Quantiles have the disadvantage that the intervals can have very different lengths. This can lead to some weird ALE plots if the feature of interest is very skewed, for example many low values and only a few very high values.

### ALE plots for the interaction of two features

ALE plots can also show the interaction effect of two features. The calculation principles are the same as for a single feature, but we work with rectangular cells instead of intervals, because we have to accumulate the effects in two dimensions. In addition to adjusting for the overall mean effect, we also adjust for the main effects of both features. This means that ALE for two features estimate the second-order effect, which does not include the main effects of the features. In other words, ALE for two features only shows the additional interaction effect of the two features. I spare you the formulas for 2D ALE plots because they are long and unpleasant to read. If you are interested in the calculation, I refer you to the paper, formulas (13) – (16). I will rely on visualizations to develop intuition about the second-order ALE calculation.

In the previous figure, many cells are empty due to the correlation. In the ALE plot this can be visualized with a grayed out or darkened box. Alternatively, you can replace the missing ALE estimate of an empty cell with the ALE estimate of the nearest non-empty cell.

Since the ALE estimates for two features only show the second-order effect of the features, the interpretation requires special attention. The second-order effect is the additional interaction effect of the features after we have accounted for the main effects of the features. Suppose two features do not interact, but each has a linear effect on the predicted outcome. In the 1D ALE plot for each feature, we would see a straight line as the estimated ALE curve. But when we plot the 2D ALE estimates, they should be close to zero, because

4. 最左边的求和代表对所有区间的平均 effect 求和。

比如：第三个区间的 ALE 是第一、二、三个区间的 effects 总和，这个特点体现了

Accumulate  
这个 effect 是中心化的所以中位 effect 是 0。

quantile 分位数；分位点

feature 分布的 分位数 作为间隔

How to pick quantile?

→ 用于展示两个特征交互的 ALE

两个 features 的 ALE 只展示它们之间的额外交互 effect

Fig 7.8

如果两个 features 没有交互，那么它们的 2D ALE plot 应接近于 0。

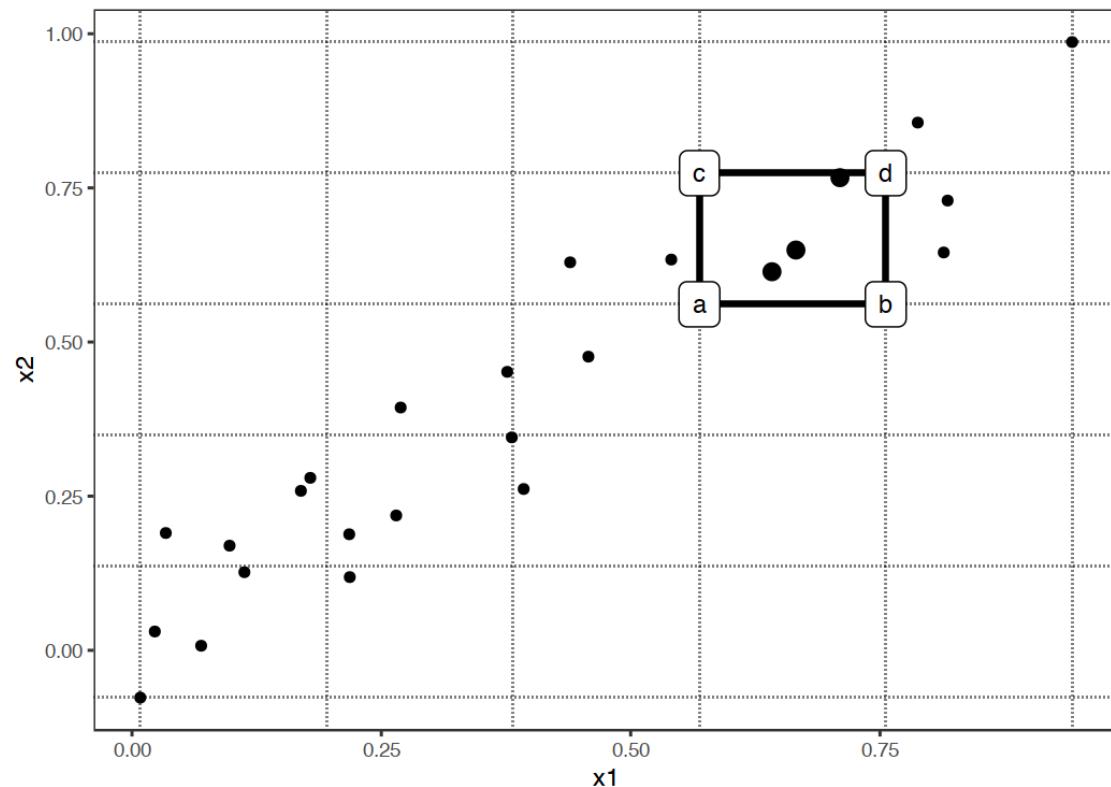


Figure 7.8: Calculation of 2D-ALE. We place a grid over the two features. In each grid cell we calculate the 2nd-order differences for all instance within. We first replace values of  $x_1$  and  $x_2$  with the values from the cell corners. If  $a$ ,  $b$ ,  $c$  and  $d$  represent the "corner"-predictions of a manipulated instance (as labeled in the graphic), then the 2nd-order difference is  $(d - c) - (b - a)$ . The mean 2nd-order difference in each cell is accumulated over the grid and centered.

the second-order effect is only the additional effect of the interaction. ALE plots and PD plots differ in this regard: PDPs always show the total effect, ALE plots show the first- or second-order effect. These are design decisions that do not depend on the underlying math. You can subtract the lower-order effects in a partial dependence plot to get the pure main or second-order effects or, you can get an estimate of the total ALE plots by refraining from subtracting the lower-order effects.

The accumulated local effects could also be calculated for arbitrarily higher orders (interactions of three or more features), but as argued in the [PDP chapter](#), only up to two features makes sense, because higher interactions cannot be visualized or even interpreted meaningfully.

## ALE for categorical features

The accumulated local effects method needs – by definition – the feature values to have an order, because the method accumulates effects in a certain direction. Categorical features do not have any natural order. To compute an ALE plot for a categorical feature we have to somehow create or find an order. The order of the categories influences the calculation and interpretation of the accumulated local effects.

One solution is to order the categories according to their similarity based on the other features. The distance between two categories is the sum over the distances of each feature. The feature-wise distance compares either the cumulative distribution in both categories, also called Kolmogorov-Smirnov distance (for numerical features) or the relative frequency tables (for categorical features). Once we have the distances between all categories, we use multi-dimensional scaling to reduce the distance matrix to a one-dimensional distance measure. This gives us a similarity-based order of the categories.

To make this a little bit clearer, here is one example: Let us assume we have the two categorical features “season” and “weather” and a numerical feature “temperature”. For the first categorical feature (season) we want to calculate the ALEs. The feature has the categories “spring”, “summer”, “fall”, “winter”. We start to calculate the distance between categories “spring” and “summer”. The distance is the sum of distances over the features temperature and weather. For the temperature, we take all instances with season “spring”, calculate the empirical cumulative distribution function and do the same for instances with season “summer” and measure their distance with the Kolmogorov-Smirnov statistic. For the weather feature we calculate for all “spring” instances the probabilities for each weather type, do the same for the “summer” instances and sum up the absolute distances in the probability distribution. If “spring” and “summer” have very different temperatures and weather, the total category-distance is large. We repeat the procedure with the other seasonal pairs and reduce the resulting distance matrix to a single dimension by multi-dimensional scaling.

### 7.2.4 Examples

Let us see ALE plots in action. I have constructed a scenario in which partial dependence plots fail. The scenario consists of a prediction model and two strongly correlated features. The prediction model is mostly a linear regression model, but does something weird at a combination of the two features for which we have never observed instances.

Is this a realistic, relevant scenario at all? When you train a model, the learning algorithm minimizes the loss for the existing training data instances. Weird stuff can happen outside the distribution of training data, because the model is not penalized for doing weird stuff in these areas. Leaving the data distribution is called extrapolation, which can also be used to fool machine learning models, described in the chapter on adversarial examples.

对于非数值的features，  
需要找到或者创造顺序。

一种方法是基于其它类别  
的相似性来排序。

春天和夏天的距离是温度  
和天气距离的和？

WTF?

当PDP图出错的时候

→ Fig 7.9

远离数据分布也叫外推法，  
这能被用于欺骗数据model

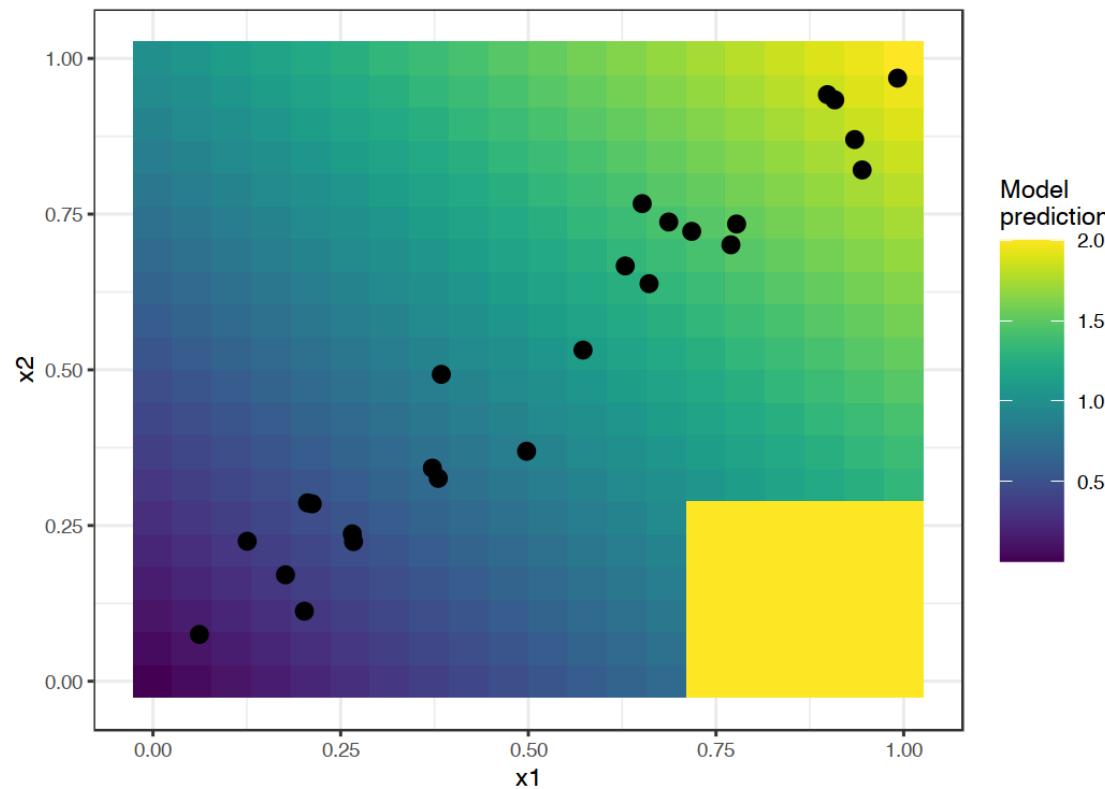


Figure 7.9: Two features and the predicted outcome. The model predicts the sum of the two features (shaded background), with the exception that if  $x_1$  is greater than 0.7 and  $x_2$  less than 0.3, the model always predicts 2. This area is far from the distribution of data (point cloud) and does not affect the performance of the model and also should not affect its interpretation.

See in our little example how the partial dependence plots behave compared to ALE plots.

→ Fig. 7.10

But is it not interesting to see that our model behaves oddly at  $x_1 > 0.7$  and  $x_2 < 0.3$ ? Well, yes and no. Since these are data instances that might be physically impossible or at least extremely unlikely, it is usually irrelevant to look into these instances. But if you suspect that your test distribution might be slightly different and some instances are actually in that range, then it would be interesting to include this area in the calculation of feature effects. But it has to be a conscious decision to include areas where we have not observed data yet and it should not be a side-effect of the method of choice like PDP. If you suspect that the model will later be used with differently distributed data, I recommend to use ALE plots and simulate the distribution of data you are expecting.

当  $x_1 > 0.7$ ,  $x_2 < 0.3$  时,  
model 总是预测 2.  
这个值远大于数据分布，  
但它没有什么影响。

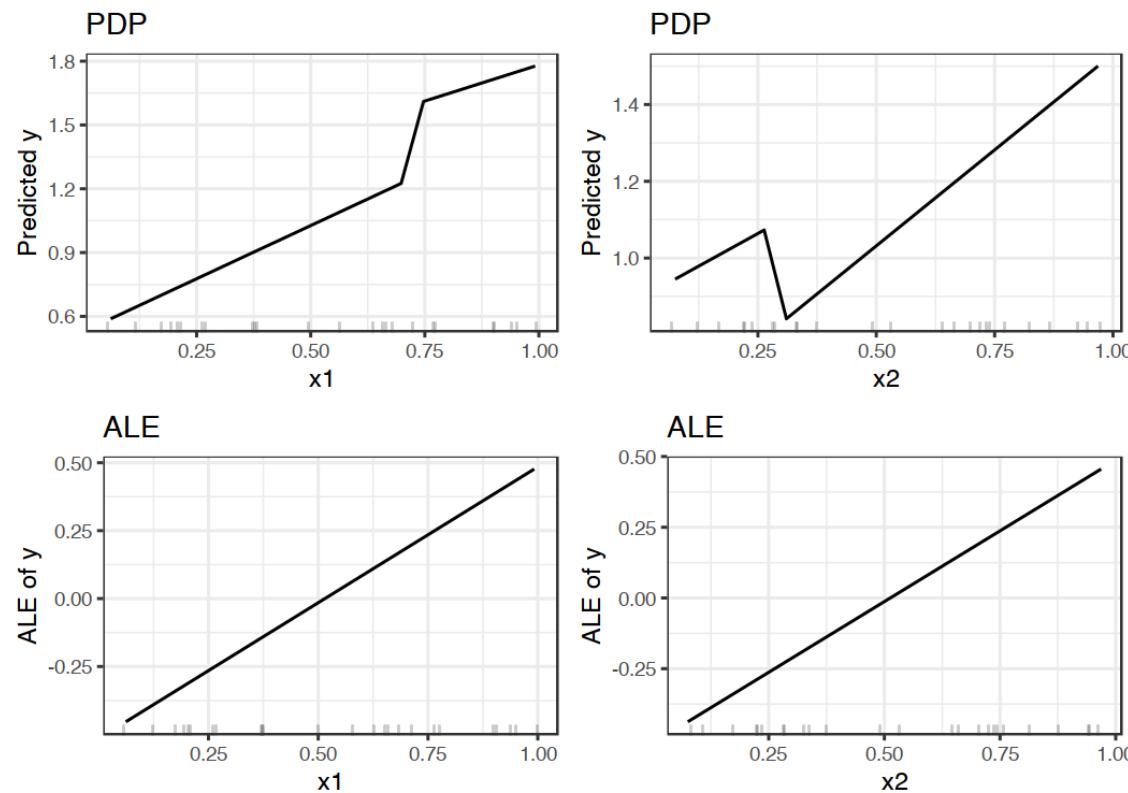


Figure 7.10: Comparison of the feature effects computed with PDP (upper row) and ALE (lower row). The PDP estimates are influenced by the odd behavior of the model outside the data distribution (steep jumps in the plots). The ALE plots correctly identify that the machine learning model has a linear relationship between features and prediction, ignoring areas without data.

ALE修正了PDP的错误

Turning to a real dataset, let us predict the **number of rented bikes** based on weather and day and check if the ALE plots really work as well as promised. We train a regression tree to predict the number of rented bicycles on a given day and use ALE plots to analyze how temperature, relative humidity and wind speed influence the predictions. Let us look at what the ALE plots say:

Let us look at the correlation between temperature, humidity and wind speed and all other features. Since the data also contains categorical features, we cannot only use the **Pearson correlation coefficient**, which only works if both features are numerical. Instead, I train a linear model to predict, for example, temperature based on one of the other features as input. Then I measure how much variance the other feature in the linear model explains and take the square root. If the other feature was numerical, then the result is equal to the absolute value of the standard Pearson correlation coefficient. But this model-

→ Fig 7.11

Pearson correlation coefficient  
What?

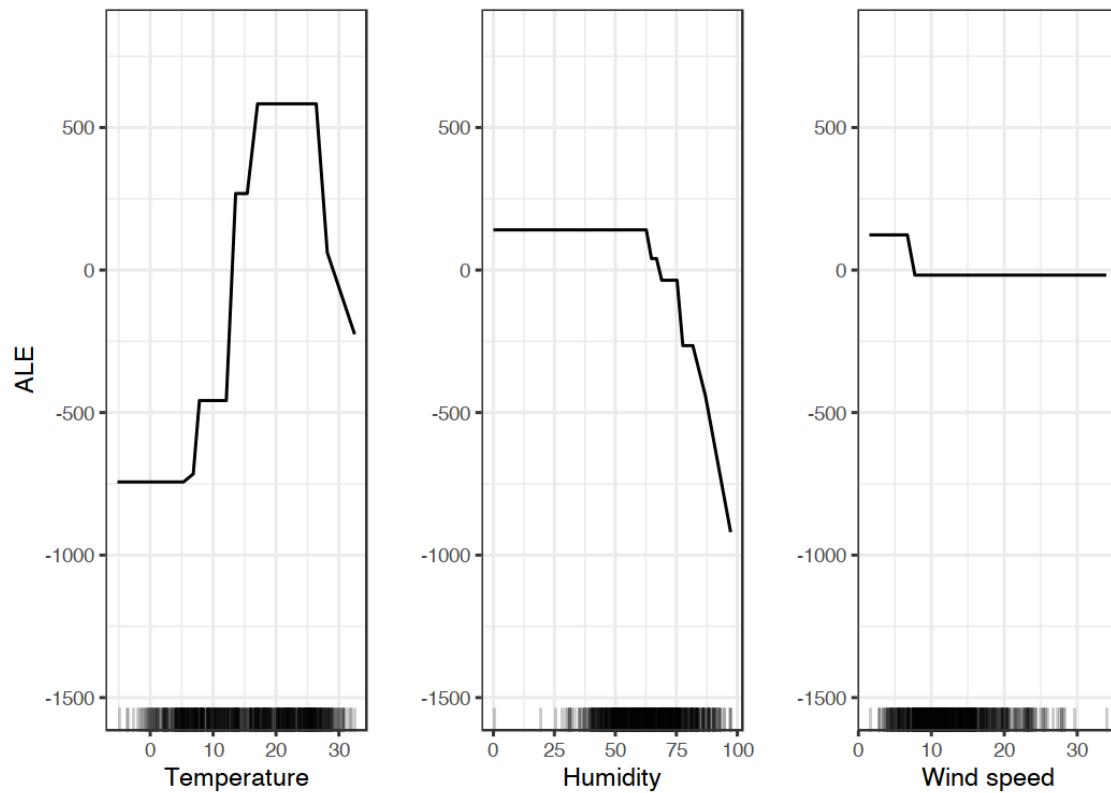


Figure 7.11: ALE plots for the bike prediction model by temperature, humidity and wind speed. The temperature has a strong effect on the prediction. The average prediction rises with increasing temperature, but falls again above 25 degrees Celsius. Humidity has a negative effect: When above 60%, the higher the relative humidity, the lower the prediction. The wind speed does not affect the predictions much.

based approach of “variance-explained” (also called ANOVA, which stands for ANalysis Of VAriance) works even if the other feature is categorical. The “variance-explained” measure lies always between 0 (no association) and 1 (temperature can be perfectly predicted from the other feature). We calculate the explained variance of temperature, humidity and wind speed with all the other features. The higher the explained variance (correlation), the more (potential) problems with PD plots. The following figure visualizes how strongly the weather features are correlated with other features.

This correlation analysis reveals that we may encounter problems with partial dependence plots, especially for the temperature feature. Well, see for yourself:

Next, let us see ALE plots in action for a categorical feature. The month is a categorical

ALE 圖

model-based 基于模型的  
Variance-explained 方差解释  
analysis of variance 方差分析

方差解释总是在0到1之间，  
值越高, feature 的 PDP 越有  
问题

→ Fig 7.12

→ Fig 7.13



Figure 7.12: The strength of the correlation between temperature, humidity and wind speed with all features, measured as the amount of variance explained, when we train a linear model with e.g. temperature to predict and season as feature. For temperature we observe – not surprisingly – a high correlation with season and month. Humidity correlates with weather situation.

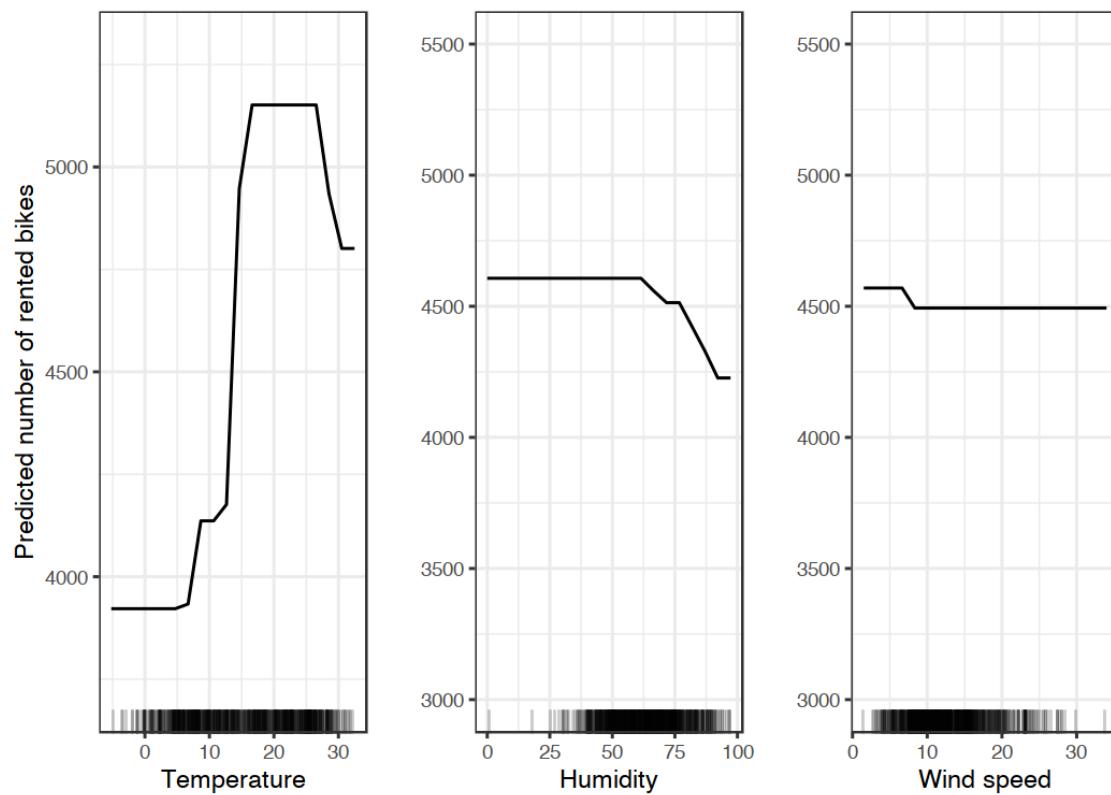
feature for which we want to analyze the effect on the predicted number of bikes. Arguably, the months already have a certain order (January to December), but let us try to see what happens if we first reorder the categories by similarity and then compute the effects. The months are ordered by the similarity of days of each month based on the other features, such as temperature or whether it is a holiday.

Since many of the features are related to weather, the order of the months strongly reflects how similar the weather is between the months. All colder months are on the left side (February to April) and the warmer months on the right side (October to August). Keep in mind that non-weather features have also been included in the similarity calculation, for example relative frequency of holidays has the same weight as the temperature for calculating the similarity between the months.

月份本身是有序的 Categorical feature, 但当根据其他 features 的相似度对月份进行重排序的话：

→ Fig 7.14.

所有较冷的月份在左边，暖的在右边。



PDP 的问题

Figure 7.13: PDPs for temperature, humidity and wind speed. Compared to the ALE plots, the PDPs show a smaller decrease in predicted number of bikes for high temperature or high humidity. The PDP uses all data instances to calculate the effect of high temperatures, even if they are, for example, instances with the season "winter". The ALE plots are more reliable.

Next, we consider the second-order effect of humidity and temperature on the predicted number of bikes. Remember that the second-order effect is the additional interaction effect of the two features and does not include the main effects. This means that, for example, you will not see the main effect that high humidity leads to a lower number of predicted bikes on average in the second-order ALE plot.

Keep in mind that both main effects of humidity and temperature say that the predicted number of bikes decreases in very hot and humid weather. In hot and humid weather, the combined effect of temperature and humidity is therefore not the sum of the main effects, but larger than the sum. To emphasize the difference between the pure second-order effect (the 2D ALE plot you just saw) and the total effect, let us look at the partial dependence plot. The PDP shows the total effect, which combines the mean prediction, the two main effects and the second-order effect (the interaction).

→ Fig 7.15

→ Fig 7.16

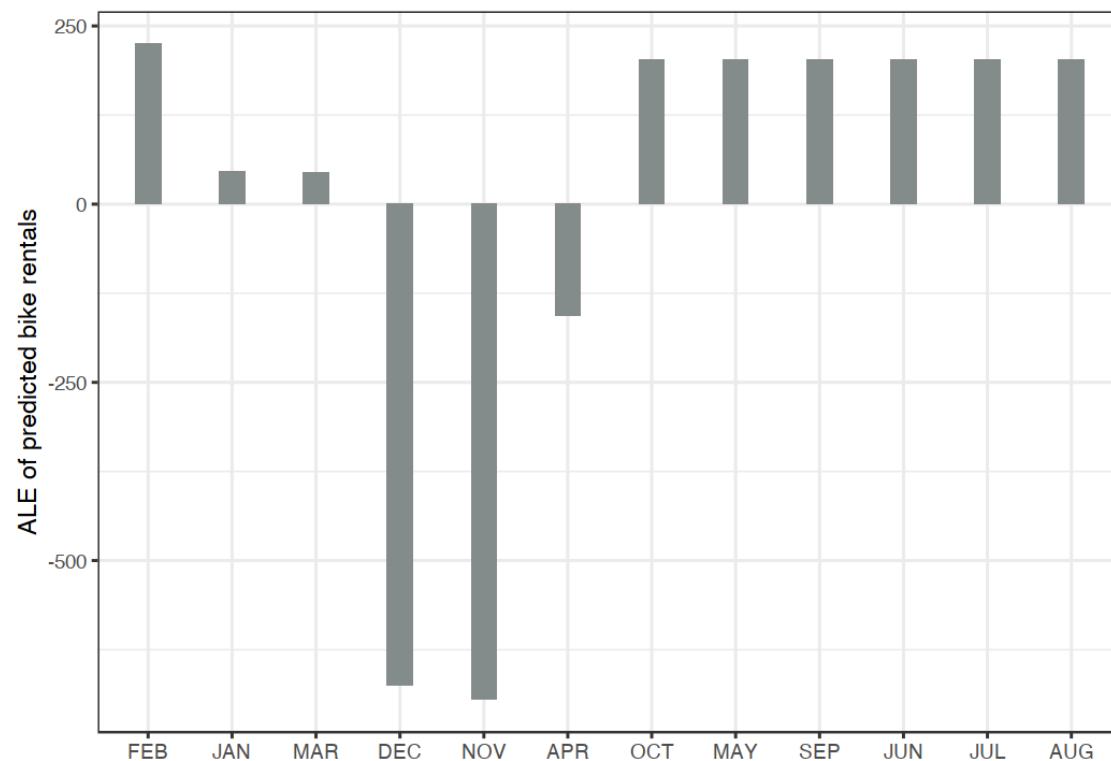


Figure 7.14: ALE plot for the categorical feature month. The months are ordered by their similarity to each other, based on the distributions of the other features by month. We observe that January, March and April, but especially December and November, have a lower effect on the predicted number of rented bikes compared to the other months.

If you are only interested in the interaction, you should look at the second-order effects, because the total effect mixes the main effects into the plot. But if you want to know the combined effect of the features, you should look at the total effect (which the PDP shows). For example, if you want to know the expected number of bikes at 30 degrees Celsius and 80 percent humidity, you can read it directly from the 2D PDP. If you want to read the same from the ALE plots, you need to look at three plots: The ALE plot for temperature, for humidity and for temperature + humidity and you also need to know the overall mean prediction. In a scenario where two features have no interaction, the total effect plot of the two features could be misleading because it probably shows a complex landscape, suggesting some interaction, but it is simply the product of the two main effects. The second-order effect would immediately show that there is no interaction.

Enough bicycles for now, let's turn to a classification task. We train a random forest

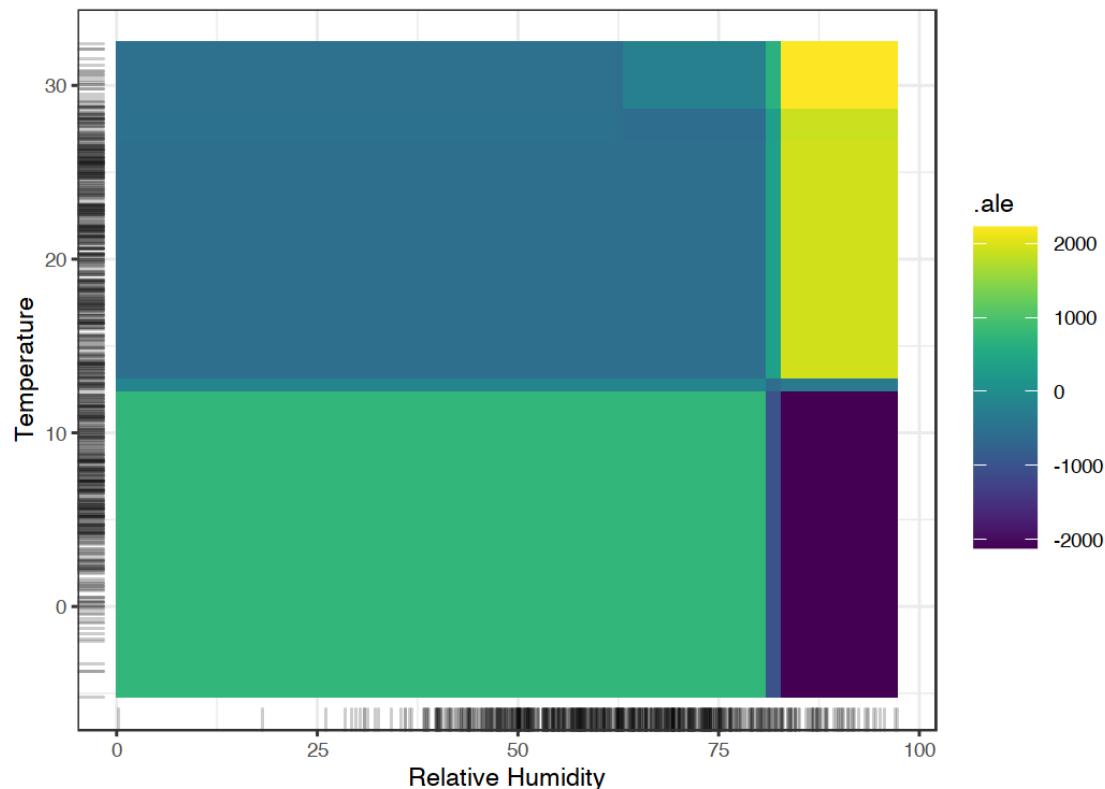


Figure 7.15: ALE plot for the 2nd-order effect of humidity and temperature on the predicted number of rented bikes. Lighter shade indicates an above average and darker shade a below average prediction when the main effects are already taken into account. The plot reveals an interaction between temperature and humidity: Hot and humid weather increases the prediction. In cold and humid weather an additional negative effect on the number of predicted bikes is shown.

to predict the probability of **cervical cancer** based on risk factors. We visualize the accumulated local effects for two of the features:

Next, we look at the interaction between number of pregnancies and age.

### 7.2.5 Advantages

ALE plots are **unbiased**, which means they still work when features are correlated. Partial dependence plots fail in this scenario because they marginalize over unlikely or even physically impossible combinations of feature values.

优点：

① unbiased.

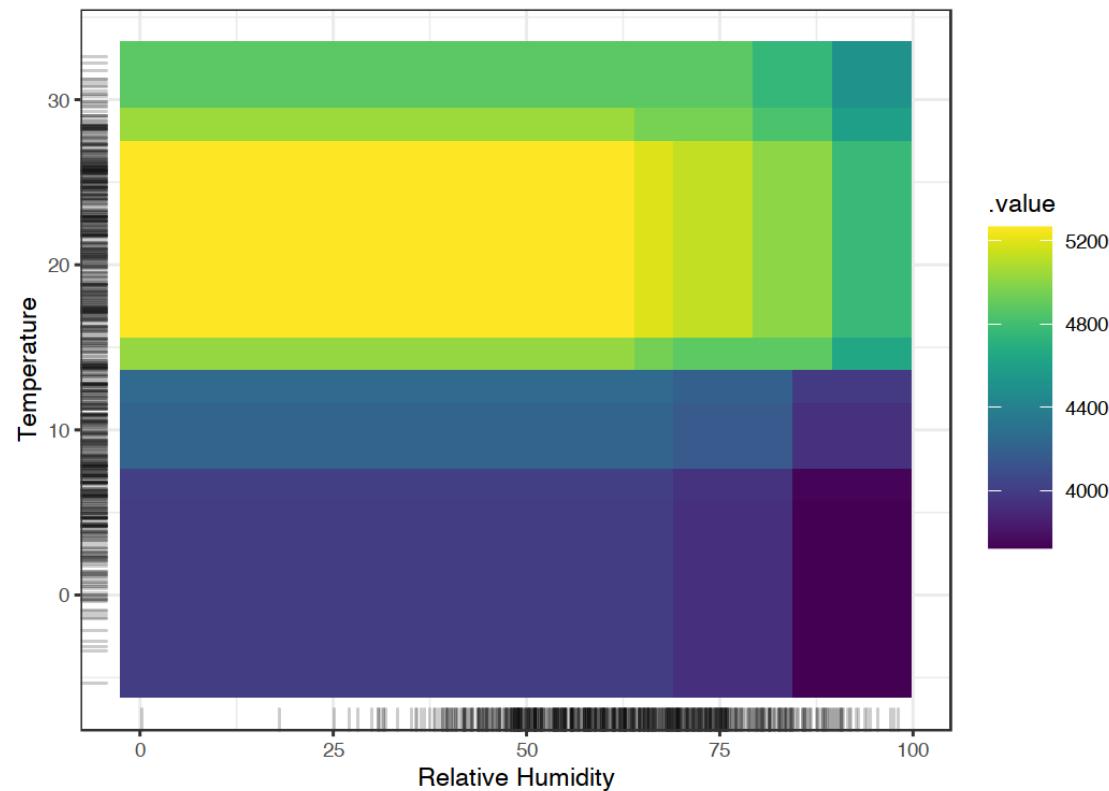


Figure 7.16: PDP of the total effect of temperature and humidity on the predicted number of bikes. The plot combines the main effect of each of the features and their interaction effect, as opposed to the 2D-ALE plot which only shows the interaction.

**ALE plots are faster to compute** than PDPs and scale with  $O(n)$ , since the largest possible number of intervals is the number of instances with one interval per instance. The PDP requires  $n$  times the number of grid points estimations. For 20 grid points, PDPs require 20 times more predictions than the worst case ALE plot where as many intervals as instances are used.

The **interpretation of ALE plots is clear**: Conditional on a given value, the relative effect of changing the feature on the prediction can be read from the ALE plot. **ALE plots are centered at zero**. This makes their interpretation nice, because the value at each point of the ALE curve is the difference to the mean prediction. **The 2D ALE plot only shows the interaction**: If two features do not interact, the plot shows nothing.

The entire **prediction function can be decomposed** into a sum of lower-dimensional ALE functions, as explained in the chapter on **functional decomposition**.

② 计算快,

③ ALE的解释很清楚.

ALE以0为中心,ALE的  
每个点都是平均预测之差.  
ALE的2D仅显示交互作用.

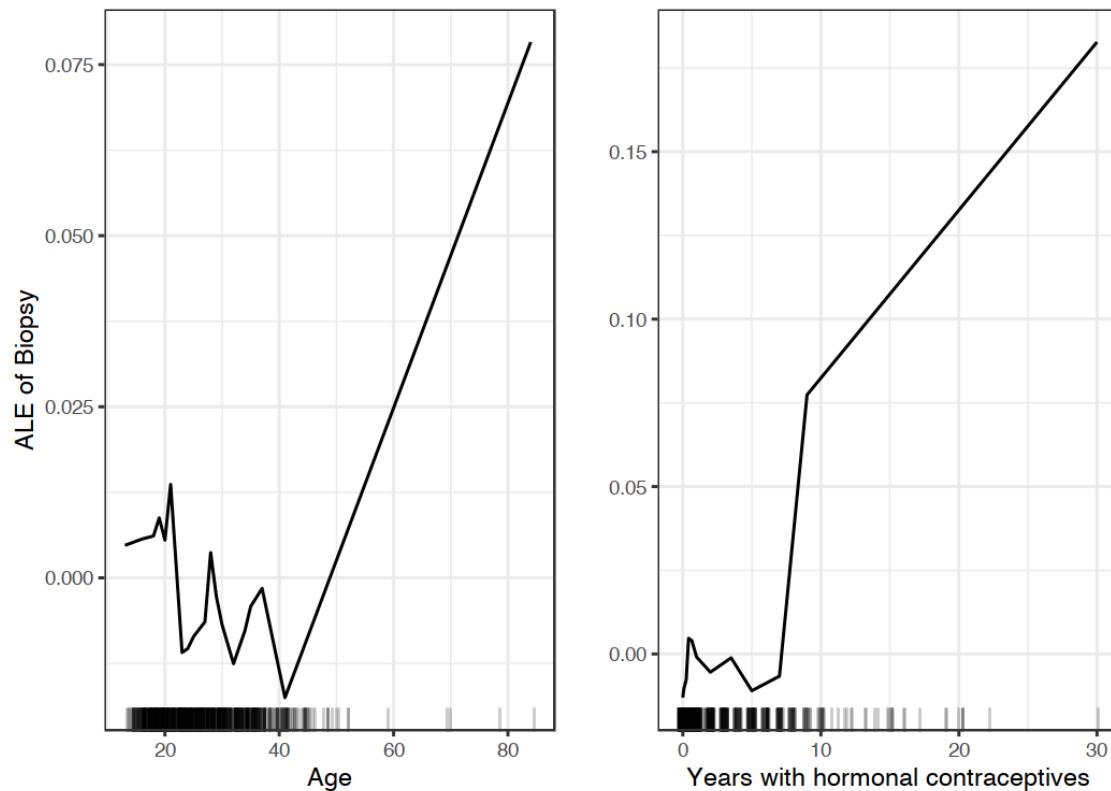


Figure 7.17: ALE plots for the effect of age and years with hormonal contraceptives on the predicted probability of cervical cancer. For the age feature, the ALE plot shows that the predicted cancer probability is low on average up to age 40 and increases after that. The number of years with hormonal contraceptives is associated with a higher predicted cancer risk after 8 years.

All in all, in most situations I would **prefer ALE plots over PDPs**, because features are usually correlated to some extent.

### 7.2.6 Disadvantages

An interpretation of the effect across intervals is not permissible if the features are strongly correlated. Consider the case where your features are highly correlated, and you are looking at the left end of a 1D-ALE plot. The ALE curve might invite the following misinterpretation: “The ALE curve shows how the prediction changes, on average, when we gradually change the value of the respective feature for a data instance, and keeping the instances other feature values fixed.” The effects are computed per interval (locally) and therefore the interpretation of the effect can only be local. For convenience, the

缺点是：

- ① 对于跨区间之间的 effect 解释是禁止的，当 features 之间高度相关联、而你在观察 plot 的左边时，ALE 曲线可能会造成这个误解：ALE 代表了一个渐渐改变、求平均的过程。
- 但总要记住：ALE 是局部计算的。

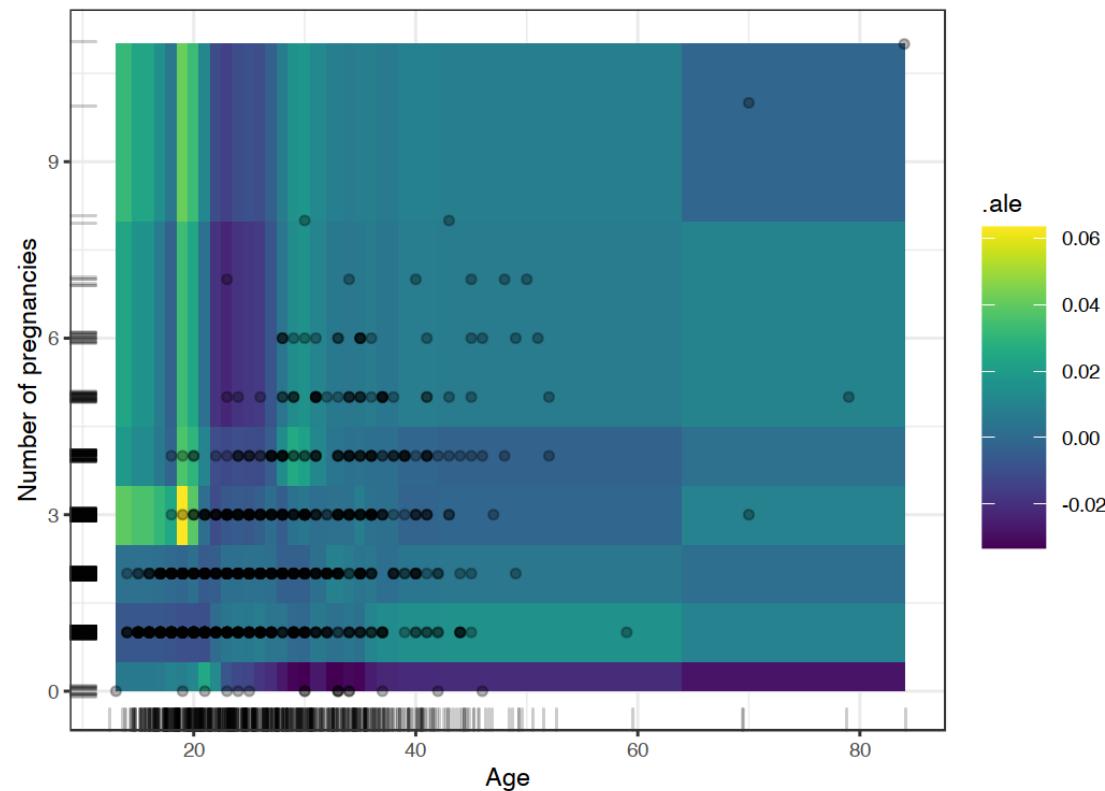


Figure 7.18: ALE plot of the 2nd-order effect of number of pregnancies and age. The interpretation of the plot is a bit inconclusive, showing what seems like overfitting. For example, the plot shows an odd model behavior at age of 18-20 and more than 3 pregnancies (up to 5 percentage point increase in cancer probability). There are not many women in the data with this constellation of age and number of pregnancies (actual data are displayed as points), so the model is not severely penalized during the training for making mistakes for those women.

interval-wise effects are accumulated to show a smooth curve, but keep in mind that each interval is created with different data instances.

ALE effects **may differ from the coefficients specified in a linear regression model** when features interact and are correlated. Grömping (2020)<sup>5</sup> showed that in a linear model with two correlated features and an additional interaction term ( $\hat{f}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ ), the first-order ALE plots do not show a straight line. Instead,

ALE 可能与线性回归中的  
系数不同。

<sup>5</sup>Grömping, Ulrike. "Model-Agnostic Effects Plots for Interpreting Machine Learning Models." Reports in Mathematics, Physics and Chemistry: Department II, Beuth University of Applied Sciences Berlin. Report 1/2020 (2020)

they are slightly curved because they incorporate parts of the multiplicative interaction of the features. To understand what is happening here, I recommend reading the chapter on **function decomposition**. In short, ALE defines first-order (or 1D) effects differently than the linear formula describes them. This is not necessarily wrong, because when features are correlated, the attribution of interactions is not as clear. But it is certainly unintuitive that ALE and linear coefficient do not match.

**ALE plots can become a bit shaky** (many small ups and downs) with a high number of intervals. In this case, reducing the number of intervals makes the estimates more stable, but also smoothes out and hides some of the true complexity of the prediction model. There is **no perfect solution for setting the number of intervals**. If the number is too small, the ALE plots might not be very accurate. If the number is too high, the curve can become shaky.

Unlike PDPs, **ALE plots are not accompanied by ICE curves**. For PDPs, ICE curves are great because they can reveal heterogeneity in the feature effect, which means that the effect of a feature looks different for subsets of the data. For ALE plots you can only check per interval whether the effect is different between the instances, but each interval has different instances so it is not the same as ICE curves.

**Second-order ALE estimates have a varying stability across the feature space, which is not visualized in any way.** The reason for this is that each estimation of a local effect in a cell uses a different number of data instances. As a result, all estimates have a different accuracy (but they are still the best possible estimates). The problem exists in a less severe version for main effect ALE plots. The number of instances is the same in all intervals, thanks to the use of quantiles as grid, but in some areas there will be many short intervals and the ALE curve will consist of many more estimates. But for long intervals, which can make up a big part of the entire curve, there are comparatively fewer instances. This happened in the cervical cancer prediction ALE plot for high age for example.

**Second-order effect plots can be a bit annoying to interpret**, as you always have to keep the main effects in mind. It is tempting to read the heat maps as the total effect of the two features, but it is only the additional effect of the interaction. The pure second-order effect is interesting for discovering and exploring interactions, but for interpreting what the effect looks like, I think it makes more sense to integrate the main effects into the plot.

The **implementation of ALE plots is much more complex** and less intuitive compared to partial dependence plots.

Even though ALE plots are not biased in case of correlated features, **interpretation remains difficult when features are strongly correlated**. Because if they have a very strong correlation, it only makes sense to analyze the effect of changing both features

ALE图可能会不稳定  
起起伏伏

没有完美的方法来设置区间大小

ALE不附带ICE曲线

二阶 ALE 有着不同的稳定性，这个点是不可忽视的。

二阶 ALE 的解释要谨慎

ALE 比 PDP 难实现

ALE虽不 bias, 但一阶 ALE  
还是不要用子解释单 feature  
的 effect - 那样没有意义

together and not in isolation. This disadvantage is not specific to ALE plots, but a general problem of strongly correlated features.

If the features are uncorrelated and computation time is not a problem, PDPs are slightly preferable because they are easier to understand and can be plotted along with ICE curves.

The list of disadvantages has become quite long, but do not be fooled by the number of words I use: As a rule of thumb: Use ALE instead of PDP.

经验之谈：  
能ALE就别PDP

### 7.2.7 Implementation and Alternatives

Did I mention that [partial dependence plots](#) and [individual conditional expectation curves](#) are an alternative? =)

ALE plots are implemented in R in the `ALEPlot` R package<sup>6</sup> by the inventor himself and once in the `iml` package<sup>7</sup>. ALE also has at least two Python implementations with the `ALEPython` package<sup>8</sup> and in `Alibi`<sup>9</sup>.

<sup>6</sup><https://cran.r-project.org/web/packages/ALEPlot/index.html>

<sup>7</sup><https://cran.r-project.org/web/packages/iml/index.html>

<sup>8</sup><https://github.com/blent-ai/ALEPython>

<sup>9</sup><https://docs.seldon.io/projects/alibi/en/stable/index.html>