Homework Challenge (2 Extra Points)

Given a simple linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

, we know how to interpret it: conditional on x_2, each 1 unit increase in x_1 is associated with β_1 units increase in y.

But given a general nonlinear model:

$$y = f(x_1, x_2)$$

, where $f(.)$ can be anything – a polynomial model, a cubic spline, how do we interpret the model? What is, for example, the "effect"[1] of x_1 on y?

A common misconception is that nonlinear models, in particular, "*black box*" models like boosted decision trees and neural networks, are only good for prediction, but cannot be interpreted. The good news is that we can indeed interpret them. Here is how:

Given $y = f(x_1, x_2)$, to gauge the "effect" of x_1 on y (at $x_1 = c$), we can either calculate the partial derivative

$$\left.\frac{\partial f(x_1, x_2 = a)}{\partial x_1}\right|_{x_1 = c} = \frac{f(x_1 = c + \epsilon, x_2 = a) - f(x_1 = c, x_2 = a)}{\epsilon} \tag{1}$$

, or the total derivative by integrating out x_2:

$$\begin{aligned} \left.\frac{df(x_1, x_2)}{dx_1}\right|_{x_1 = c} &= \int \left.\frac{\partial f(x_1, x_2)}{\partial x_1}\right|_{x_1 = c} p(x_2)\, dx_2 \\ &= \int \frac{f(c + \epsilon, x_2) - f(c, x_2)}{\epsilon} p(x_2)\, dx_2 \end{aligned} \tag{2}$$

[1] Note: here by "effect", we simply mean how much change in y is associated with a given change in x. We do *not* mean causal effect!

Calculate (1) or (2) at different c, i.e. for a range of different values of x_1, gives us the partial dependence plot of y on x_1 (i.e., the "effect" of x_1 on y).

To calculate the partial derivate, we need to hold x_2 constant. A common choice is to fix them at their median values (or other quantiles depending on your interest).

To calculate the total derivative, we need to integrate out x_2. How to do this numerically? The theory of **monte carlo integration** tells us that if we can draw many x_2 from the distribution $p(x_2)$, then

$$\left.\frac{df(x_1, x_2)}{dx_1}\right|_{x_1=c} \approx \sum_{i=1}^{M} \left.\frac{\partial f(x_1, x_{2i})}{\partial x_1}\right|_{x_1=c} = \sum_{i=1}^{M} \frac{f(c+\epsilon, x_{2i}) - f(c, x_{2i})}{\epsilon}$$

, where $\{x_{2i}\}_{i=1}^{M}$ are the x_2 points drawn from $p(x_2)$. Therefore, if we already have a large data set, then we can simply calculate $\left.\frac{\partial f(x_1, x_{2i})}{\partial x_1}\right|_{x_1=c}$ on all data points $i = 1, \ldots, N$, and average their results. If our data set is small, we can generate more x_{2i} by using resampling techniques such as the bootstrap.

Challenge

Task 1

Let $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and let $y = f(x_1, x_2) + e$. Simulate data from this model and calculate (1) and (2) for different values of x_1. Plot the resulting partial dependence relationship.

Task 2

Find (or simulate) any data set. Fit a decision tree and calculate (1) and (2) for different values of an input variable. Plot the resulting partial dependence relationship.