1 Use of OLS to recover $F(\mathbf{x})$ in PD-AVID

Variable importance metrics have a demonstrated tendency to inflate in the presence of correlated variables (Strobl et al. 2008). Following past research on adjusting variable importance scores to account for this tendency (Strobl et al. 2008, Molnar et al. 2023), we use OLS regression to recover the prediction function $F(\boldsymbol{x}_i)$ from variable k's linear, nonlinear, and interaction components, rather than simply summing them.

To demonstrate our motivation for doing so, we examine the squared loss function $L(\mathbf{y}, F(X)) =_i (y_i - F(\boldsymbol{x}_i))^2$, commonly used in regression models and equivalent to the Brier Score loss function in binary classification models. For data points $i \in \{1, ..., n\}$, the mean squared error (or Brier Score, in the case of binary classification) is given by:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - F(\boldsymbol{x}_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{F}_k(x_{i,k}) - \tilde{F}_k(x_{i,k}) - F_{\sim k}(\boldsymbol{x}_{i\sim k}) - \Delta_{i,k})^2 \tag{1}$$

given our decomposition of the model prediction function in Equation **??**. Multiplying the right side yields:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i - F(\boldsymbol{x}_i))^2 = \frac{1}{n}\sum_{i=1}^{n}[&y_i^2 + \bar{F}_k^{\;2}(x_{i,k}) + \tilde{F}_k^{\;2}(x_{i,k}) + F_{\sim k}^2(\boldsymbol{x}_{i\sim k}) + \Delta_{i,k}^2 \\ &- 2(y_i)(\bar{F}_k(x_{i,k}) - \tilde{F}_k(x_{i,k}) - F_{\sim k}(\boldsymbol{x}_{i\sim k}) - \Delta_{i,k}) \\ &- 2(\bar{F}_k(x_{i,k}))(\tilde{F}_k(x_{i,k}) - F_{\sim k}(\boldsymbol{x}_{i\sim k}) - \Delta_{i,k}) \\ &- 2(\tilde{F}_k(x_{i,k}))(F_{\sim k}(\boldsymbol{x}_{i\sim k}) - \Delta_{i,k}) - 2(F_{\sim k}(\boldsymbol{x}_{i\sim k}))(\Delta_{i,k})]\end{aligned} \tag{2}$$

For the sake of this proof, we add the term $\frac{1}{n}\sum_{i=1}^{n} 3y_i^2 - \frac{1}{n}\sum_{i=1}^{n} 3y_i^2$ (which equals zero) to the right-hand side:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i - F(\boldsymbol{x}_i))^2 &= -\frac{1}{n}\sum_{i=1}^{n}3y_i^2 + \frac{1}{n}\sum_{i=1}^{n}[(y_i^2 - 2(y_i)\bar{F}_k(x_{i,k}) + \bar{F}_k^{\;2}(x_{i,k}))+ \\ &(y_i^2 - 2(y_i)\tilde{F}_k(x_{i,k}) + \tilde{F}_k^{\;2}(x_{i,k})) + (y_i^2 - 2(y_i)(F_{\sim k}(\boldsymbol{x}_{i\sim k})) + F_{\sim k}^2(\boldsymbol{x}_{i\sim k}))+ \\ &(y_i^2 - 2(y_i)\Delta_{i,k} + \Delta_{i,k}^2) - 2Z] \\ &= -\frac{1}{n}\sum_{i=1}^{n}3y_i^2 + \frac{1}{n}\sum_{i=1}^{n}[(y_i - \bar{F}_k(x_{i,k}))^2 + (y_i - \tilde{F}_k(x_{i,k}))^2+ \\ &(y_i - F_{\sim k}(\boldsymbol{x}_{i\sim k}))^2 + (y_i - \Delta_{i,k})^2 - 2Z]\end{aligned} \tag{3}$$

where Z is the sum of all pairwise cross-product terms between the prediction components $\bar{F}_k(x_{i,k})$, $\tilde{F}_k(x_{i,k})$, $F_{\sim k}(\boldsymbol{x}_{i\sim k})$, and $\Delta_{i,k}$, from the last two lines of Equation 2. Under this decomposition, it is clear that for a given variable k, the model's MSE is a function of the error due to variable k's linear component, its nonlinear component, its interaction with other variables, as well as the roles of all other variables *and* the covariance (cross-products) between these terms, captured in the cross-product term Z.

If we were to simply omit (i.e., replace with zero) one of variable k's prediction components $\bar{F}_k(x_{i,k})$, $\tilde{F}_k(x_{i,k})$, or $\Delta_{i,k}$, the MSE will increase as long as that prediction component of is positively correlated with y_i (i.e., $(\sum y_i - 0)^2 > \sum(y_i - F_k)^2$). But it will *also* cause the MSE to increase if the prediction component is positively correlated with the other prediction components, since its omission will shrink the cross-product term Z. This is, conceptually, why marginal variable importance metrics overestimate the importance of correlated variables; omitting correlated variables removes their contribution toward $F(\boldsymbol{x}_i)$ *and* shared variance with the other components of each observation's model prediction, all of which contribute to the model's squared error.

Estimating the prediction functions $F_k^{(\sim lin)}$ and $F_k^{(\sim non)}$ using OLS still omits information about variables' linear and nonlinear contributions (respectively), but does so in a way that the resulting predictions ($F_k^{(\sim lin)}(\boldsymbol{x}_i)$, $F_k^{(\sim non)}(\boldsymbol{x}_i)$) and the omitted portion (the residuals $F(\boldsymbol{x}_i) - F_k^{(\sim non)}(\boldsymbol{x}_i)$ and $F(\boldsymbol{x}_i) - F_k^{(\sim lin)}(\boldsymbol{x}_i)$) are uncorrelated. If we denote a given prediction component as $F_k^{(\sim \cdot)}(\boldsymbol{x}_i)$ and the residuals as $e_k^{\sim \cdot}(\boldsymbol{x}_i)$, we can show that substituting $F_k^{(\sim \cdot)}(\boldsymbol{x}_i)$ for $F(\boldsymbol{x}_i)$ does *not* omit information about the covariance between prediction components:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - F(\boldsymbol{x}_i))^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - F_k^{(\sim \cdot)}(\boldsymbol{x}_i) - e_k^{\sim \cdot}(\boldsymbol{x}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 + F_k^{(\sim \cdot)}(\boldsymbol{x}_i)^2 + e_k^{\sim \cdot}(\boldsymbol{x}_i)^2 - 2(y_i)F_k^{(\sim \cdot)}(\boldsymbol{x}_i) - 2(y_i)e_k^{\sim \cdot}(\boldsymbol{x}_i) - 2(F_k^{(\sim \cdot)}(\boldsymbol{x}_i))(e_k^{\sim \cdot}(\boldsymbol{x}_i)) \\
&= -\frac{1}{n}\sum_{i=1}^{n} y_i^2 + \frac{1}{n}\sum_{i=1}^{n}[(y_i - F_k^{(\sim \cdot)}(\boldsymbol{x}_i))^2 + (y_i - e_k^{(\sim \cdot)}(\boldsymbol{x}_i))^2] \qquad (4)
\end{aligned}$$

Because of the use of OLS to estimate $F_k^{(\sim \cdot)}(\boldsymbol{x}_i)$, the final cross-product term in the third line equals zero. Thus, when $e_k^{\sim \cdot}(\boldsymbol{x}_i)$ is set to zero, all information about that prediction component (*lin* or *non*) is removed, and the squared error increases *only* by the degree to which $\sum(y_i - 0)^2 > \sum(y_i - e_k^{\sim \cdot}(\boldsymbol{x}_i))^2$, i.e., the degree to which the omitted component correlates with outcomes. Note that, in cases where the feature $\mathbf{x}_k$ can be perfectly predicted from other features (e.g., one-hot encoded dummy categories), $F_k^{(\sim lin)}$ and $F_k^{(\sim non)}$ are computed by summing the prediction components rather than through OLS, since the latter would perfectly recover the prediction function and result in artificially deflated linear and nonlinear importance scores.

References

Molnar C, König G, Bischl B, Casalicchio G (2023) Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery* 1–39.

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC bioinformatics* 9:1–11.