1 Equation 1: Permutation Variable Importance

Equation 1 is used to quantify the importance of a specific variable in a machine learning model. Specifically, it calculates the *permutation-based variable importance* for variable k by measuring how much the model's loss function increases when the values of that variable are randomly permuted while keeping other variables intact.

This method is useful because it provides insight into how much the accuracy (or performance) of the model depends on a particular feature. By observing how the model's prediction degrades when a variable is permuted, we can determine how relevant or critical that variable is to the model's overall predictions.

1.1 Main Usage, Indirect Benefits, and Intuition

1.1.1 Main Usage

- **Model Interpretation**: Permutation variable importance is widely used in machine learning to interpret black-box models (e.g., random forests, neural networks). It helps determine which variables the model relies on most to make its predictions.

- **Feature Effect Analysis**: It helps assess the contribution of each variable to the overall predictive performance of a model, which is crucial for making decisions based on the model's results, especially in high-stakes applications like healthcare, finance, and engineering.

1.1.2 Indirect Benefits

- **Feature Selection**: It aids in feature selection, allowing model developers to remove irrelevant or less important features, which can improve model generalization and efficiency.

- **Model Debugging**: When debugging models, permutation importance helps understand if the model is overfitting to noise or irrelevant features.

1.1.3 Intuition Behind the Calculation

The idea behind *permutation variable importance* is simple:

- Permuting (shuffling) a feature breaks the relationship between that feature and the target variable.

- If the model relies heavily on that feature to make accurate predictions, then shuffling the feature values should significantly degrade the model's performance.

- In contrast, if permuting a feature has little effect on the model's performance, that feature is considered less important.

1.2 Interpreting the Potential Outputs of Eq. 1

The results of the permutation variable importance can provide valuable insights into how a specific feature influences the model's performance. The potential outcomes and their interpretation are as follows:

- **High Increase in Loss Function**: If permuting the values of a feature leads to a significant increase in the loss function (or degradation in performance), it suggests that the model relies heavily on that feature to make accurate predictions. This feature is considered highly important.
- **Minimal Change in Loss Function**: If permuting the values of a feature results in little or no change in the model's performance, this indicates that the feature is not essential for the model's predictions. It may be a candidate for removal during feature selection to simplify the model.
- **Negative Impact**: In rare cases, permuting a feature could improve the model's performance, indicating that the feature introduces noise or irrelevant information into the model. This could signal a problem such as overfitting or the inclusion of redundant variables, warranting further investigation.

1.3 Numerical Example for Equation 1

Equation 1 is:

$$VI_k = \frac{1}{n}\sum_{i=1}^{n}[L(y_i, F(\tilde{x}_{i.k}, x_{i\sim k})) - L(y_i, F(x_i))]$$

Where:

- VI_k is the permutation variable importance for feature X_k,
- $F(\tilde{x}_{i.k}, x_{i\sim k})$ represents the prediction after randomly permuting X_k,
- $F(x_i)$ is the original prediction,
- L is the loss function, and
- n is the number of observations.

We will compute the permutation variable importance for X_1 using the following simple dataset:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

To calculate the permutation importance for X_1, we randomly permute the values of X_1 while keeping X_2 and X_3 unchanged. The permuted values for X_1 are:

Observation (i)	X_1 (Permuted)	X_2	X_3	New Prediction $F(\tilde{X}_1, X_2, X_3)$
1	8	3	4	45
2	2	6	7	20
3	5	9	10	40

2 Equation 2: Partial Dependence Functions

Equation 2 represents the Partial Dependence (PD) function, which is used to quantify the relationship between a specific subset of variables S and the model's prediction, while averaging out the effects of all other variables not in S. This approach is widely used in machine learning to understand how the selected variables influence the model's output, without being affected by interactions with other variables.

The partial dependence function is particularly helpful when trying to visualize the marginal effect of a variable or a subset of variables S on the model's prediction, while ignoring the influence of other variables.

2.1 Main Usage, Indirect Benefits, and Intuition

2.1.1 Main Usage

- **Model Interpretation**: The primary use of the partial dependence function is to interpret complex machine learning models (e.g., random forests, gradient-boosting machines, neural networks). It helps to visualize the marginal effect of one or more input features on the predicted output.

- **Feature Effect Analysis**: By isolating and visualizing the effect of a specific variable or a set of variables, practitioners can better understand how individual features influence the model's prediction and whether the relationships are linear, non-linear, or exhibit interactions.

2.1.2 Indirect Benefits

- **Interaction Effects Detection**: Although partial dependence plots ignore the interactions with other variables, they can give insight into whether such interactions may exist. For instance, a non-linear or unexpected shape in the plot could suggest that interactions between variables are present.

2.1.3 Intuition Behind the Calculation

The intuition behind the partial dependence function is straightforward: It represents the average effect of a variable or set of variables S on the model's output, while accounting for the distribution of the other variables.

- **Fix the Values of Variables in S**: Choose specific values for the variables of interest, x_S.
- **Vary All Other Variables**: For each observation i, keep the values of the other variables not in S (denoted $x_{i,\sim S}$) as they are.
- **Average the Model's Prediction**: Calculate the model's prediction for each observation, then average the predictions over all the observations. This averaging helps neutralize the effects of other variables, allowing you to isolate the influence of S.

In essence, the partial dependence function gives you a sense of how the model responds to changes in a particular variable or subset of variables, while averaging out the influence of all other variables.

2.2 Numerical Example for Equation 2

The partial dependence (PD) function $F_S(x_S)$ gives a prediction by the model for a prescribed set of values x_S, where S is a subset of variables, and x_S represents the values of the variables in this subset. The goal of the partial dependence function is to isolate the influence of the variables in S while averaging out the effects of all other variables that are not in S.

We will compute the partial dependence for X_1 using a simple dataset with three variables: X_1, X_2, and X_3. The dataset is as follows:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

We will fix $x_S = X_1 = 5$ (to see the effect of fixing $X_1 = 5$) and calculate the partial dependence function by averaging the model's predictions across the different values of X_2 and X_3 (i.e., $x_{i,\sim S}$).

Step-by-Step Calculation:

We calculate $F(x_S = 5, x_{i,\sim S})$ for each observation i:

- For observation 1, $x_S = 5$ and $x_{1,\sim S} = (X_2 = 3, X_3 = 4)$. The prediction is $F(5, 3, 4)$. Let's assume the model gives a prediction of 25.
- For observation 2, $x_S = 5$ and $x_{2,\sim S} = (X_2 = 6, X_3 = 7)$. The prediction is $F(5, 6, 7)$. Assume the model gives a prediction of 35.

- For observation 3, $x_S = 5$ and $x_{3,\sim S} = (X_2 = 9, X_3 = 10)$. The prediction is $F(5, 9, 10)$. Assume the model gives a prediction of 45.

Finally, we compute the partial dependence function $F_S(x_S = 5)$ by averaging these predictions:

$$F_S(5) = \frac{1}{3}(25 + 35 + 45) = \frac{105}{3} = 35$$

In this example, by fixing $X_1 = 5$ and averaging the model's predictions over different values of X_2 and X_3, we estimate that the partial dependence of $X_1 = 5$ on the model's predictions is 35. This approach allows us to isolate the effect of X_1 on the outcome, independent of X_2 and X_3.

2.3 Interpreting the Potential Outputs of Eq. 2

The results of the partial dependence function can provide key insights into the role of specific variables in the model's predictions. Here are the potential outcomes and their interpretations:

- **Linear Relationship**: If the partial dependence plot shows a linear relationship between a variable and the prediction, it suggests that the feature has a straightforward, proportional effect on the model's output.

- **Non-Linear Relationship**: A non-linear pattern in the plot indicates that the effect of the variable on the model's prediction varies at different values, suggesting that the feature has a more complex influence on the model's predictions.

- **Flat Relationship**: If the partial dependence plot is flat, it suggests that the variable has little to no impact on the model's predictions. This may indicate that the variable is irrelevant for the specific prediction task.

- **Potential Interaction Indications**: Unexpected patterns in the partial dependence plot, such as sudden jumps or drops, may indicate hidden interactions with other variables not captured by the plot, suggesting that further investigation into interactions is necessary.

3 Equation 3: H^2 Statistics

The H^2 statistic quantifies the proportion of variance in model predictions that is due to interactions between variables. Specifically, in the case of two-way interactions between variables k_1 and k_2, the H^2 statistic measures the proportion of variance in their two-way partial dependence function $F_{k1,k2}(x_{k1}, x_{k2})$ that is not attributed to the one-way partial dependence functions $F_{k1}(x_{k1})$ and $F_{k2}(x_{k2})$.

3.1 Main Usage, Indirect Benefits, and Intuition

3.1.1 Main Usage

- **Interaction Effects Quantifying**: The primary use of the H^2 statistic is to quantify the *interaction effects* between two variables in a machine learning model. It helps in understanding how much of the variance in model predictions is explained by the combined interaction of the two variables, rather than their individual effects.

3.1.2 Indirect Benefits

- **Model Interpretation**: By measuring interaction effects, the H^2 statistic helps interpret complex dependencies in machine learning models, making it easier to explain black-box models. This understanding of interactions enhances feature engineering and contributes to more accurate, interpretable models.

3.1.3 Intuition Behind the Calculation

The intuition behind the H^2 statistic lies in comparing the full two-way partial dependence function of two variables with the sum of their individual one-way partial dependence functions.

- **Two-way Partial Dependence Function**: This function, $F_{k1,k2}(x_{k1}, x_{k2})$, represents the model's prediction considering both k_1 and k_2 simultaneously.
- **One-way Partial Dependence Functions**: These functions, $F_{k1}(x_{k1})$ and $F_{k2}(x_{k2})$, represent the model's prediction considering each variable independently.
- **Interaction Term**: The interaction effect is detected by calculating how much the full two-way partial dependence deviates from the sum of the individual one-way effects. The H^2 statistic captures this deviation as a proportion of the total variance in the model's predictions.

The larger the H^2 value, the more of the variance in the model's predictions is explained by the interaction between k_1 and k_2, beyond their individual effects.

3.2 Interpreting the Potential Outputs of Eq. 3

The outputs of the H^2 statistic provide insights into how interactions between variables affect the model's predictions. The potential outcomes and their interpretations are as follows:

- **High H^2 Value**: A high H^2 value (close to 1) indicates a strong interaction effect between the two variables, meaning a significant portion of the variance in the model's predictions is driven by their combined effect, beyond their individual contributions.

- **Low H^2 Value**: A low H^2 value (close to 0) suggests that the interaction between the two variables has little to no influence on the model's predictions, and their individual effects are sufficient to explain the variance.
- **Intermediate H^2 Value**: An intermediate H^2 value indicates that while some of the variance is due to the interaction between the two variables, their individual effects still play a considerable role in explaining the model's predictions.

3.3 Formula

The H^2 statistic is given by:

$$H^2_{k1,k2} = \frac{\sum_{i=1}^{n} \left[F_{k1,k2}(x_{i,k1}, x_{i,k2}) - F_{k1}(x_{i,k1}) - F_{k2}(x_{i,k2}) \right]^2}{\sum_{i=1}^{n} F_{k1,k2}(x_{i,k1}, x_{i,k2})^2}$$

Where:

- $H^2_{k1,k2}$ quantifies the interaction effect between variables k_1 and k_2.
- The numerator captures the squared difference between the full two-way partial dependence function and the sum of the one-way partial dependence functions, representing the interaction term.
- The denominator is the total variance in the two-way partial dependence function, normalizing the interaction term as a proportion of the total variance.

3.4 Numerical Example for H^2 Statistic

3.4.1 H^2 Statistic for Two-Way Interactions (Equation 3)

Equation 3 for two-way interactions is:

$$H^2_{k_1,k_2} = \frac{\sum_{i=1}^{n} \left[F_{k_1,k_2}(x_{i,k_1}, x_{i,k_2}) - F_{k_1}(x_{i,k_1}) - F_{k_2}(x_{i,k_2}) \right]^2}{\sum_{i=1}^{n} F_{k_1,k_2}(x_{i,k_1}, x_{i,k_2})^2}$$

Where:

- $H^2_{k_1,k_2}$ is the interaction effect between variables X_{k_1} and X_{k_2},
- $F_{k_1,k_2}(x_{i,k_1}, x_{i,k_2})$ is the model prediction based on both variables X_{k_1} and X_{k_2},
- $F_{k_1}(x_{i,k_1})$ and $F_{k_2}(x_{i,k_2})$ are the marginal predictions for X_{k_1} and X_{k_2},
- n is the number of observations.

To illustrate the H^2 statistic, let's use the following dataset:

Observation (i)	X_1	X_2	Model Prediction $F(X_1, X_2)$
1	2	3	20
2	5	6	35
3	8	9	50

We will also compute the marginal predictions for X_1 and X_2 separately, assuming:

Observation (i)	Marginal Prediction $F(X_1)$	Marginal Prediction $F(X_2)$
1	19	18
2	34	33
3	49	48

Step-by-Step Calculation

We will now calculate the numerator of Equation 3 for each observation by computing the squared difference between the interaction and the sum of marginal effects:

- For observation 1: $(20 - 19 - 18)^2 = (-17)^2 = 289$
- For observation 2: $(35 - 34 - 33)^2 = (-32)^2 = 1024$
- For observation 3: $(50 - 49 - 48)^2 = (-47)^2 = 2209$

Now, calculate the denominator by summing the squares of the interaction predictions:

$$\sum_{i=1}^{3} F(X_1, X_2)^2 = 20^2 + 35^2 + 50^2 = 400 + 1225 + 2500 = 4125$$

Finally, the $H^2_{X_1,X_2}$ statistic is:

$$H^2_{X_1,X_2} = \frac{289 + 1024 + 2209}{4125} = \frac{3522}{4125} \approx 0.854$$

This shows how interaction effects between X_1 and X_2 contribute to the model predictions. The higher H^2 value indicates that the interaction between X_1 and X_2 has a significant effect on the model's prediction.

4 Plot 1 (Eq. 5)

Equation 5 introduces a modified version of the variable importance score from Equation 1. This version computes variable importance using partial dependence (PD) functions to replace the prediction $F(x_i)$ with the partial dependence function $F_{\sim k}(x_{\sim k})$, which excludes the feature x_k from the model. The primary goal of this modified variable importance score is to measure how much the model's loss function changes when a feature is replaced with its partial dependence estimate, thus allowing for computational efficiency in evaluating variable importance. This version of the variable importance score gives similar values to permutation-based importance scores but uses partial dependence functions to compute the contribution of each feature more efficiently.

4.1 Main Usage, Indirect Benefits, and Intuition

4.1.1 Main Usage

- **Total Variable Importance**:This equation calculates the total contribution of a feature to the model's predictive power. It does this by evaluating how much worse the model performs when that feature is replaced by the PD function, which provides a smoothed estimate of the feature's impact. The result reflects the overall importance of the variable, considering both its individual effect and interactions with other variables. The use of the PD function allows you to account for all possible influences the variable might have, without requiring computationally expensive permutations or retraining of the model. Instead of isolating specific linear, nonlinear, or interaction effects, this equation gives a single comprehensive measure of how important the variable is to the model's overall accuracy.

4.1.2 Indirect Benefits

- **Computational Efficiency**: By using partial dependence functions, this approach offers *computational savings* when evaluating variable importance, especially in complex models where calculating permutation importance for every feature might be computationally expensive.

- **Model Interpretation**: Like traditional variable importance scores, this method aids in interpreting black-box models by quantifying how each feature impacts model predictions. Additionally, by decomposing variable importance, this method provides more detailed insights into the nature of a variable's contribution (linear vs. nonlinear vs. interaction).

4.1.3 Intuition Behind the Calculation

The intuition behind this modified variable importance score is based on calculating the change in the model's loss function when a feature is replaced by its partial dependence estimate rather than permuted or excluded from the model entirely. Here's how it works:

- **Partial Dependence Function**: For each feature x_k, the partial dependence function $F_{\sim k}(x_{\sim k})$ is used. This function provides the model's average prediction when only considering the values of all other features (i.e., $x_{\sim k}$) and excluding x_k.

- **Replace the Prediction**: Instead of calculating the model's loss using the original prediction $F(x_i)$, the prediction is replaced by the average predictions from the partial dependence function for all values of x_k.
- **Compare Losses**: The modified variable importance score is computed as the difference between the original loss (computed with the model's prediction) and the new loss (computed after replacing the feature with its partial dependence estimate). This difference gives an indication of how important the feature is in the model's prediction process.

Mathematically:

$$\hat{VI}_k = \frac{1}{n}\sum_{i=1}^{n}[L(y_i, F_{\sim k}(x_{i,\sim k})) - L(y_i, F(x_i))]$$

Where:

- $\hat{VI}_k$ is the variable importance score for feature x_k,
- $F_{\sim k}(x_{\sim k})$ is the partial dependence function for all variables except x_k,
- L is the loss function,
- n is the number of observations.

This approach is useful when computational efficiency is needed, as the partial dependence function reduces the need for multiple permutations or recalculations for each feature. Additionally, it highlights both linear and nonlinear effects of the variable in the prediction process.

4.2 Interpreting the Potential Outputs of Eq. 5

The output of Equation 5 provides valuable insights into how the importance of a variable affects the model's overall performance. The potential outcomes and their interpretation are as follows:

- **Negative Importance Score**: A negative variable importance score indicates that replacing the variable x_k with its partial dependence estimate improves the model's performance. This suggests that the variable x_k might be introducing noise or unnecessary complexity into the model, reducing its accuracy. In such cases, the variable might be a candidate for removal or further investigation.
- **Positive Importance Score**: A positive score means that replacing the variable x_k with its partial dependence estimate degrades the model's performance, indicating that the feature is important for the model's predictions. The higher the positive value, the more critical the feature is for achieving accurate predictions.

- **Close to Zero Score**: If the score is close to zero, this suggests that the variable has little to no effect on the model's performance. The model can predict equally well with or without the feature. In such cases, the variable could be considered for exclusion to simplify the model without sacrificing predictive power.
- **High Absolute Score (positive or negative)**: A higher absolute value of the variable importance score (whether positive or negative) indicates that the feature plays a significant role in the model's performance. The direction (positive or negative) reveals whether the variable improves or worsens the model when replaced with its partial dependence estimate.

4.3 Numerical Example for Equation 5

Equation 5 presents a modified version of the variable importance score based on Partial Dependence (PD) functions. It computes the average change in model loss when replacing each model prediction with the partial dependence function for all variables except x_k. Here's the formula:

$$\hat{VI}_k = \frac{1}{n}\sum_{i=1}^{n}[L(y_i, F_{\sim k}(x_{i,\sim k})) - L(y_i, F(x_i))]$$

This can be simplified for computational efficiency by averaging the PD function over grid values $x_k^{(gr)}$, resulting in:

$$\hat{VI}_k = \frac{1}{n}\sum_{i=1}^{n}\left[L(y_i, \frac{1}{g_k}\sum_{j=1}^{g_k} F(x_{jk}^{(gr)}, x_{i,\sim k})) - L(y_i, F(x_i))\right]$$

Where:

- $\hat{VI}_k$ is the variable importance score for feature x_k,
- $L(y_i, F(x_i))$ is the loss function applied to the model's prediction,
- $F_{\sim k}(x_{i,\sim k})$ is the partial dependence function for all variables except x_k,
- g_k is the number of grid points used to approximate the PD function for x_k,
- $F(x_{jk}^{(gr)}, x_{i,\sim k})$ is the PD function evaluated at grid points for x_k,
- n is the number of observations.

We will use the same dataset and assume a simple loss function, such as Mean Squared Error (MSE), to illustrate the steps. The dataset is as follows:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Assumptions:

- True labels $y_1 = 22$, $y_2 = 36$, and $y_3 = 48$,
- We'll use the Mean Squared Error (MSE) as the loss function $L(y_i, F(x_i)) = (y_i - F(x_i))^2$,
- We assume there are 2 grid points for X_1, i.e., $g_1 = 2$, and similarly for X_2.

Loss Using Full Model Predictions First, we calculate the loss using the full model predictions:

- For observation 1: $L(22, 20) = (22 - 20)^2 = 4$,
- For observation 2: $L(36, 35) = (36 - 35)^2 = 1$,
- For observation 3: $L(48, 50) = (48 - 50)^2 = 4$.

Partial Dependence Function for X_1 We now calculate the partial dependence function for X_1. Assume that the model prediction is averaged over 2 grid values for X_1:

- For observation 1, assume the grid predictions $F(x_{11}^{(gr)}, X_2, X_3)$ are 21 and 23.
- For observation 2, assume the grid predictions are 36 and 37.
- For observation 3, assume the grid predictions are 49 and 51.

Now, compute the average predictions:

- For observation 1: $\frac{1}{2}(21 + 23) = 22$,
- For observation 2: $\frac{1}{2}(36 + 37) = 36.5$,
- For observation 3: $\frac{1}{2}(49 + 51) = 50$.

Loss Using PD Function for X_1 Next, we compute the loss using these partial dependence predictions:

- For observation 1: $L(22, 22) = (22 - 22)^2 = 0$,
- For observation 2: $L(36, 36.5) = (36 - 36.5)^2 = 0.25$,
- For observation 3: $L(48, 50) = (48 - 50)^2 = 4$.

Final Calculation of $\hat{VI}_1$ We now compute $\hat{VI}_1$ as the average change in loss:

$$\hat{VI}_1 = \frac{1}{3}\left[(0-4) + (0.25-1) + (4-4)\right]$$

$$\hat{VI}_1 = \frac{1}{3}\left[-4 + (-0.75) + 0\right] = \frac{1}{3} \times (-4.75) = -1.58$$

Thus, the modified variable importance score $\hat{VI}_1$ is approximately -1.58, which indicates that using the partial dependence function for X_1 improves the model's predictions by reducing the overall loss.

5 Plot 1 (Eq. 9-11)

Equations 9, 10, and 11 decompose variable importance into three components: **linear**, **nonlinear**, and **interaction** effects. These equations provide insights into how a feature affects the model predictions in different ways.

5.1 Main Usage

- **Feature Effect Decomposition**: These equations allow for a more nuanced analysis of feature importance by separating the effects into linear, nonlinear, and interaction components.

5.2 Indirect Benefits

- **Model Interpretation**: Decomposing variable importance helps explain how each feature influences the outcome, enabling better transparency and understanding of model predictions.
- **Feature Engineering**: Understanding linear and nonlinear effects helps improve feature engineering.
- **Interaction Effect Detection**: Equation 11 is particularly useful for identifying significant interaction effects between variables.

5.3 Intuition Behind the Calculation

- **Linear Importance**: Measures how well the variable's relationship with the outcome can be approximated by a straight line.
- **Nonlinear Importance**: Captures complex, curved relationships between the variable and the outcome.
- **Interaction Importance**: Measures how much of the variable's effect comes from its interaction with other variables.

5.4 Interpreting the Potential Outputs of Eq. 9-11

The output of Equations 9, 10, and 11 provides deeper insights into how the linear, nonlinear, and interaction effects of a variable impact the model's overall predictions. The potential outcomes and their interpretations are as follows:

- **High Positive Linear Importance** $\hat{VI}_k^{(lin)}$: A high positive value indicates that the variable's effect can largely be captured by a linear relationship with the outcome. This suggests the feature contributes to the model in a straightforward, proportional manner.
- **High Nonlinear Importance** $\hat{VI}_k^{(non)}$: A high nonlinear importance value indicates that the variable has a complex, nonlinear relationship with the outcome, which cannot be captured by a simple linear model. This can highlight variables with significant curvature or thresholds in their effect.
- **High Interaction Importance** $\hat{VI}_k^{(int)}$: A high interaction importance score suggests that the variable interacts significantly with other variables. This means the variable's effect on the outcome depends on the values of other features, and its influence cannot be fully captured by its individual linear or nonlinear effects.
- **Negative Importance Scores**: If any of the importance scores (linear, nonlinear, or interaction) are negative, this indicates that the variable's inclusion is detrimental to the model's accuracy for that component. Negative interaction importance, for example, suggests that the interactions between variables introduce noise or reduce model performance.
- **Close to Zero Scores**: If any of the scores are close to zero, the variable does not contribute significantly to the model in that component (linear, nonlinear, or interaction). This suggests the feature might be redundant for that aspect of the model.

5.5 Numerical Example for Equations 9, 10, and 11

Equations 9, 10, and 11 decompose the total variable importance score into linear, nonlinear, and interaction components.

$$\hat{VI}_k^{(lin)} = \frac{1}{n}\sum_{i=1}^{n}\left[L(y_i, F_{\sim k}^{(lin)}(x_i)) - L(y_i, F(x_i))\right]$$

$$\hat{VI}_k^{(non)} = \frac{1}{n}\sum_{i=1}^{n}\left[L(y_i, F_{\sim k}^{(non)}(x_i)) - L(y_i, F(x_i))\right]$$

$$\hat{VI}_k^{(int)} = \hat{VI}_k - \hat{VI}_k^{(lin)} - \hat{VI}_k^{(non)}$$

We will use the same dataset and assume a simple Mean Squared Error (MSE) loss function:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Assumptions:

- True labels $y_1 = 22$, $y_2 = 36$, and $y_3 = 48$,
- We'll use the Mean Squared Error (MSE) as the loss function $L(y_i, F(x_i)) = (y_i - F(x_i))^2$.

5.5.1 Step-by-Step Calculation

Loss Using Full Model Predictions: First, we calculate the loss using the full model predictions:

- For observation 1: $L(22, 20) = (22 - 20)^2 = 4$,
- For observation 2: $L(36, 35) = (36 - 35)^2 = 1$,
- For observation 3: $L(48, 50) = (48 - 50)^2 = 4$.

Linear Predictions $F_{\sim k}^{(lin)}(x_i)$: Assume the linear model for X_1 produces the following linear predictions:

- For observation 1: $F_{\sim 1}^{(lin)}(2) = 19$,
- For observation 2: $F_{\sim 1}^{(lin)}(5) = 34$,
- For observation 3: $F_{\sim 1}^{(lin)}(8) = 49$.

Now, compute the linear loss:

- For observation 1: $L(22, 19) = (22 - 19)^2 = 9$,
- For observation 2: $L(36, 34) = (36 - 34)^2 = 4$,
- For observation 3: $L(48, 49) = (48 - 49)^2 = 1$.

Nonlinear Predictions $F_{\sim k}^{(non)}(x_i)$: Assume the nonlinear model for X_1 produces the following nonlinear predictions:

- For observation 1: $F_{\sim 1}^{(non)}(2) = 20.5$,
- For observation 2: $F_{\sim 1}^{(non)}(5) = 35.5$,
- For observation 3: $F_{\sim 1}^{(non)}(8) = 50.5$.

Now, compute the nonlinear loss:

- For observation 1: $L(22, 20.5) = (22 - 20.5)^2 = 2.25$,
- For observation 2: $L(36, 35.5) = (36 - 35.5)^2 = 0.25$,
- For observation 3: $L(48, 50.5) = (48 - 50.5)^2 = 6.25$.

5.5.2 Compute Linear, Nonlinear, and Interaction Importance

Linear Importance $\hat{VI}_1^{(lin)}$:

$$\hat{VI}_1^{(lin)} = \frac{1}{3}\left[(9-4)+(4-1)+(1-4)\right] = \frac{1}{3} \times 5 = 1.67$$

Nonlinear Importance $\hat{VI}_1^{(non)}$:

$$\hat{VI}_1^{(non)} = \frac{1}{3}\left[(2.25-4)+(0.25-1)+(6.25-4)\right]$$

$$\hat{VI}_1^{(non)} = \frac{1}{3} \times (-1.75 - 0.75 + 2.25) = \frac{1}{3} \times (-0.25) = -0.08$$

Interaction Importance $\hat{VI}_1^{(int)}$: Using the total variable importance score $\hat{VI}_1 \approx -1.58$ from the previous example:

$$\hat{VI}_1^{(int)} = \hat{VI}_1 - \hat{VI}_1^{(lin)} - \hat{VI}_1^{(non)}$$

$$\hat{VI}_1^{(int)} = -1.58 - 1.67 - (-0.08) = -1.58 - 1.67 + 0.08 = -3.17$$

5.5.3 Final Results:

- $\hat{VI}_1^{(lin)} = 1.67$
- $\hat{VI}_1^{(non)} = -0.08$
- $\hat{VI}_1^{(int)} = -3.17$

Thus, the linear component of X_1 contributes positively to the model, the nonlinear component contributes slightly negatively, and the interaction term has a significant negative contribution.

6 Equations 12 and 13 (Plot 2): Inverse Propensity Weighting and Weighted Partial Dependence Functions

Equations 12 and 13 deal with adjusting the influence of data points within a given neighborhood set to ensure accurate Partial Dependence (PD) estimates. This is especially useful when data points are unevenly distributed across different regions of the feature space.

6.1 Main Usage, Applications, and Intuition

6.1.1 Equation 12: Weight Vector for Neighborhood Sets ($w^{(k:j)}$)

$$w^{(k:j)} = \min\left\{1, \frac{1}{p_{i,j,k}} Q\left(\frac{1}{p_{i,j,k}}, q_{\text{clip}}\right)\right\}$$

This equation defines the **inverse propensity weights**, which correct biases in the estimated Partial Dependence (PD) functions due to varying probabilities that a record falls into a specific neighborhood set. The weights adjust the influence of each observation based on how likely it is to belong to a neighborhood.

Main Usage:

- Bias correction for causal inference or treatment effect estimation: By using inverse propensity weights, the method helps ensure that the analysis accounts for selection bias, leading to more accurate estimation of causal effects or treatment outcomes.

Indirect Benefit:

- Data Balancing in ML Models: The weights help balance the data distribution across different neighborhood sets, improving model performance by ensuring that no neighborhood is over- or under-represented in the estimation process. This can enhance the robustness and fairness of the models.

Intuition: The weights ensure that under-represented points in a neighborhood are given stronger influence, and over-represented points are scaled down, ensuring the model's estimates reflect the true relationships in the data.

6.1.2 Equation 13: Weighted Partial Dependence Function with Inverse Propensity Weighting

$$F_{k_1}^{(k_2:j)}(x_{k_1}^{(gr)}) = \frac{1}{\sum_{i \in S(k_2,j)} w_i^{(k_2:j)}} \sum_{i \in S(k_2,j)} F(x_{k_1}^{(gr)}, x_{i,\sim k_1}) \times w_i^{(k_2:j)}$$

This equation estimates the **partial dependence function** for a variable x_{k_1}, weighted by the inverse propensity weights from Equation 12. It takes into account interactions with an interacting variable x_{k_2}, ensuring that the PD estimate reflects the true relationships between variables.

Main Usage

Accurate Partial Dependence (PD) Estimation: The equation provides more accurate PD function estimates by correcting for biases introduced by varying probabilities that a record falls into a specific neighborhood. This is particularly useful in situations where relationships between variables vary across different neighborhood sets $S(j, k_2)$.

Indirect Benefits

- **Model Interpretation:** By averaging across neighborhoods and adjusting weights appropriately, the method enhances the interpretability of the PD function. This makes it easier to understand how a specific variable influences the outcome in different contexts.

- **Interaction Detection:** The equation produces separate PD function estimates across different neighborhoods, highlighting potential interactions between the focal variable and the interacting variable x_{k_2}. This helps uncover effects that might be missed in standard PD analyses.

Thus, this equation ensures that the estimation of the PD function accounts for both the focal variable and the context provided by the interacting variable, leading to more nuanced insights and more reliable model interpretations.

6.2 Interpreting the Potential Outputs of Eq. 12 and 13

The results of Equations 12 and 13 provide insights into how to adjust the influence of different data points in a given neighborhood to ensure accurate Partial Dependence (PD) estimates. The potential outcomes and their interpretations are as follows:

- **Balanced Weights Across Neighborhoods**: If the weights are balanced across different neighborhoods, it suggests that the data points are evenly distributed, and the model does not need to heavily adjust the influence of any specific group. This leads to smoother PD estimates that are not skewed by over- or under-representation in certain regions of the feature space.

- **Heavily Weighted Data Points**: If certain data points receive significantly higher weights, it indicates that these observations are underrepresented in their neighborhood set and must be given more influence to ensure a balanced estimate. This can help correct biases due to sparse data in some areas.

- **Low Weight Adjustments**: If the weight adjustments are minimal, it suggests that the data is already well-represented across neighborhoods, and the PD estimates are already reflective of the true relationships between the variables without requiring significant corrections.

- **Improved Interaction Detection**: With Equation 13, separate PD estimates across different neighborhoods highlight interaction effects between the focal variable and the interacting variable. If significant differences are observed across neighborhoods, it suggests that interactions between the variables are driving the model's predictions in different ways, providing deeper insights into the model's behavior.

6.3 Numerical Example for Equations 12 and 13

Equations 12 and 13 involve calculating inverse propensity weights and using them to compute a partial dependence (PD) function for a focal variable x_{k1} with respect to an interacting variable x_{k2}.

6.3.1 Numerical Example

We use the following dataset:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Assume the logistic regression probabilities $p_{i,k}$ and the clipped weights are:

Observation (i)	Probability $p_{i,k}$	Clipped Weight $Q\left(\frac{1}{p_{i,k}}, q_{clip}\right)$
1	0.75	1.2
2	0.6	1.0
3	0.9	1.5

The final weights are:

Observation (i)	Weight $w_i^{(k2,j)}$
1	1
2	1
3	1

Assume model predictions $F(x_{k1}^{(gr)}, x_{i,\sim k1})$ at grid points $x_{k1}^{(gr)}$ are:

Grid Point $x_{k1}^{(gr)}$	Model Prediction $F(x_{k1}^{(gr)}, x_{i,\sim k1})$
$x_{11}^{(gr)} = 2$	21
$x_{12}^{(gr)} = 5$	36
$x_{13}^{(gr)} = 8$	49

PD Function Calculations: For $x_{11}^{(gr)} = 2$:

$$F_{k1}^{(k2,j)}(x_{11}^{(gr)}) = \frac{1}{1+1+1} \times (21 \times 1 + 21 \times 1 + 21 \times 1) = 21$$

For $x_{12}^{(gr)} = 5$:

$$F_{k1}^{(k2,j)}(x_{12}^{(gr)}) = \frac{1}{1+1+1} \times (36 \times 1 + 36 \times 1 + 36 \times 1) = 36$$

For $x_{13}^{(gr)} = 8$:

$$F_{k1}^{(k2,j)}(x_{13}^{(gr)}) = \frac{1}{1+1+1} \times (49 \times 1 + 49 \times 1 + 49 \times 1) = 49$$

6.3.2 Final Results:

- $F_{k1}^{(k2,j)}(x_{11}^{(gr)}) = 21$
- $F_{k1}^{(k2,j)}(x_{12}^{(gr)}) = 36$
- $F_{k1}^{(k2,j)}(x_{13}^{(gr)}) = 49$

7 Equations 17 and 18 (Plot 4)

Equations 17 and 18 define moderated Partial Dependence Functions (PDPs) for pairs of interacting variables. These equations split the feature space into different regions (e.g., high and low values of interacting variables) and estimate the PDP for each region separately. This allows for the visualization of how two variables interact in their influence on the model's predictions, with one variable set at high or low values and the other similarly moderated.

7.1 Main Usage, Applications, and Intuition

7.1.1 Equation 17: Moderated PDP for High Values of x_{k_2}

$$F_{k1}^{(k2,+)}(x_{k1}^{(gr)}) = \frac{1}{\sum_{i \in S(k2,+)} w_i^{(k2,+)}} \sum_{i \in S(k2,+)} F(x_{k1}^{(gr)}, x_{i,\sim k1}) \times w_i^{(k2,+)}$$

This equation computes the moderated partial dependence function for a focal variable x_{k_1} when the interacting variable x_{k_2} is at high levels.

7.1.2 Equation 18: Moderated PDP for Low Values of x_{k_2}

$$F_{k1}^{(k2,-)}(x_{k1}^{(gr)}) = \frac{1}{\sum_{i \in S(k2,-)} w_i^{(k2,-)}} \sum_{i \in S(k2,-)} F(x_{k1}^{(gr)}, x_{i,\sim k1}) \times w_i^{(k2,-)}$$

Similarly, this equation estimates the moderated partial dependence function for the same focal variable x_{k_1}, but when the interacting variable x_{k_2} is at low levels.

7.2 Main Usage

Pairwise Interaction Analysis: Equations 17 and 18 are primarily used for analyzing pairwise interactions between variables in machine learning models. By breaking down the interaction into high and low regions for an interacting variable, these equations help capture how two variables together influence the model's predictions. This is particularly useful in complex models where non-linear relationships between variables exist, allowing for a more detailed understanding of feature interactions.

7.3 Indirect Benefits

- **Model Interpretation:** By visualizing the joint effects of two variables, these equations offer a clearer view of how each variable contributes to the predictions, making the decision-making process of the model more interpretable and transparent.
- **Non-Linear Feature Interaction Exploration:** These equations are valuable for exploring non-linear interactions in models like random forests or gradient boosting machines, where feature interactions are not straightforward. This deeper exploration allows the detection of complex relationships that may be missed with traditional analysis.

7.3.1 Intuition Behind the Calculation

The intuition behind Equations 17 and 18 lies in moderating the relationship between a focal variable x_{k_1} and the model's prediction based on high or low values of an interacting variable x_{k_2}. By doing so, you can understand whether the interaction effect is strong or weak in different regions of the feature space.

- **Equation 17:** This equation calculates the partial dependence of x_{k_1} on the prediction when x_{k_2} is in a high region.
- **Equation 18:** This equation calculates the partial dependence of x_{k_1} on the prediction when x_{k_2} is in a low region.

7.4 Interpreting the Potential Outputs of Eq. 17 and 18

The results of Equations 17 and 18 provide insight into how two variables interact in their influence on the model's predictions when one variable is moderated at high or low levels. The potential outcomes and their interpretations are as follows:

- **Consistent Results Across High and Low Values**: If the moderated PDPs produce similar results for high and low levels of the interacting variable, it indicates that the interaction effect is weak, and the focal variable's impact on the model predictions is relatively consistent across different regions of the feature space.
- **Different Results Between High and Low Values**: If the moderated PDPs differ significantly between high and low values of the interacting variable, it suggests that the interaction effect is strong. The focal variable's influence on the predictions varies depending on whether the interacting variable is high or low, highlighting a complex relationship between the two features.
- **Flat PDP Curves**: If the moderated PDPs are flat across both high and low levels of the interacting variable, it indicates that the focal variable has

little to no effect on the model predictions, regardless of the interacting variable's values.

- **Non-Linear Patterns**: If the moderated PDPs exhibit non-linear patterns, it suggests that the interaction between the two variables is complex and non-linear, meaning their combined influence on the predictions cannot be captured by a simple linear model.

7.5 Numerical Example for Equations 17 and 18

We will use the following dataset to demonstrate how Equations 17 and 18 are used:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Assume the following probabilities and weights for high and low levels of the interacting variable X_2:

Observation (i)	Probability $p_{i,k}$	Weight $w_i^{(k2,+)}$
1	0.8	1.2
2	0.7	1.1
3	0.9	1.5

We will now calculate the moderated PDPs for X_1, assuming grid values of X_1 and the corresponding model predictions:

Grid Point $x_{k1}^{(gr)}$	Model Prediction $F(x_{k1}^{(gr)}, x_{i,\sim k1})$
$x_{11}^{(gr)} = 2$	21
$x_{12}^{(gr)} = 5$	36
$x_{13}^{(gr)} = 8$	49

7.5.1 Step-by-Step Calculation

For high levels of X_2 (Equation 17):

$$F_{k1}^{(k2,+)}(x_{11}^{(gr)}) = \frac{1}{1.2 + 1.1 + 1.5} \times (21 \times 1.2 + 21 \times 1.1 + 21 \times 1.5) = 21$$

$$F_{k1}^{(k2,+)}(x_{12}^{(gr)}) = \frac{1}{1.2 + 1.1 + 1.5} \times (36 \times 1.2 + 36 \times 1.1 + 36 \times 1.5) = 36$$

$$F_{k1}^{(k2,+)}(x_{13}^{(gr)}) = \frac{1}{1.2 + 1.1 + 1.5} \times (49 \times 1.2 + 49 \times 1.1 + 49 \times 1.5) = 49$$

For low levels of X_2 (Equation 18), assume the weights are:

Observation (i)	Weight $w_i^{(k2,-)}$
1	1.1
2	0.9
3	1.3

We perform similar calculations for low levels of X_2:

$$F_{k1}^{(k2,-)}(x_{11}^{(gr)}) = \frac{1}{1.1 + 0.9 + 1.3} \times (21 \times 1.1 + 21 \times 0.9 + 21 \times 1.3) = 21$$

$$F_{k1}^{(k2,-)}(x_{12}^{(gr)}) = \frac{1}{1.1 + 0.9 + 1.3} \times (36 \times 1.1 + 36 \times 0.9 + 36 \times 1.3) = 36$$

$$F_{k1}^{(k2,-)}(x_{13}^{(gr)}) = \frac{1}{1.1 + 0.9 + 1.3} \times (49 \times 1.1 + 49 \times 0.9 + 49 \times 1.3) = 49$$

7.5.2 Final Results:

- $F_{k1}^{(k2,+)}(x_{11}^{(gr)}) = 21$
- $F_{k1}^{(k2,+)}(x_{12}^{(gr)}) = 36$
- $F_{k1}^{(k2,+)}(x_{13}^{(gr)}) = 49$
- $F_{k1}^{(k2,-)}(x_{11}^{(gr)}) = 21$
- $F_{k1}^{(k2,-)}(x_{12}^{(gr)}) = 36$
- $F_{k1}^{(k2,-)}(x_{13}^{(gr)}) = 49$

Thus, the moderated partial dependence functions for X_1 with respect to high and low levels of X_2 both give the same results in this example.

8 Main Usage, Application, and Intuition Behind Equations 19 and 20 (Plot 5)

Equations 19 and 20 are designed to estimate the interaction effects between three variables x_{k1}, x_{k2}, x_{k3} in machine learning models, capturing their combined influence on model predictions.

8.1 Main Usage

Three-Way Interaction Importance: The primary purpose of these equations is to quantify the **importance of three-way interactions** in a machine learning model. Equation 19 estimates the partial dependence of a focal variable x_{k1} while considering its interaction with two other variables x_{k2} and x_{k3}. Equation 20 then computes the three-way H^2 interaction score, which provides a measure of the variance in predictions that can be attributed to the joint effect of these three variables. Together, these equations capture the significance of higher-order interactions, helping to identify how variable combinations influence the model's predictions.

8.2 Indirect Benefits

- **Model Interpretation:** By examining higher-order interactions, these equations help uncover the multi-dimensional relationships between features, leading to improved model interpretation, especially in complex models where interactions are non-linear and difficult to detect with simpler methods.

8.3 Intuition Behind Equations 19 and 20

- Equation 19 computes the average predictions for a focal variable x_{k1}, while weighting these predictions based on the interaction effects of x_{k2} and x_{k3}. This helps quantify how the interaction between these variables impacts the prediction of x_{k1}.

- Equation 20 extends the concept of the two-way H^2 statistic to capture the three-way interaction effects between three variables. It adjusts the three-way H^2 score by considering the interaction effects between pairs of the three variables.

8.4 Interpreting the Potential Outputs of Eq. 19 and 20

The results of Equations 19 and 20 provide insights into how three variables interact to influence model predictions. The potential outcomes and their interpretations are as follows:

- **High Three-Way PDP Values**: If the three-way Partial Dependence Function (PDP) values are high, it indicates that the interaction between the three variables x_{k1}, x_{k2}, x_{k3} plays a significant role in determining the model's predictions. This means that the combination of these variables has a strong influence on the outcome.

- **Low Three-Way PDP Values**: If the three-way PDP values are low, it suggests that the joint effect of the three variables on the model's predictions is weak. The focal variable x_{k1} may not be strongly affected by its interactions with x_{k2} and x_{k3}, meaning that their combined effect on the predictions is minimal.

- **High Three-Way H^2 Score**: A high three-way H^2 score indicates that a large proportion of the variance in the model's predictions can be explained by the interaction between the three variables. This suggests that the interactions between x_{k1}, x_{k2}, and x_{k3} are critical for accurate model predictions.

- **Low Three-Way H^2 Score**: A low three-way H^2 score means that only a small fraction of the variance in predictions is due to the three-way interaction between the variables. This implies that the individual contributions or two-way interactions between the variables are more important than their combined three-way interaction.
- **Intermediate H^2 Score**: An intermediate H^2 score indicates that while the three-way interaction is significant, it is not the primary driver of model predictions. Other factors, such as individual or pairwise interactions, may still play a significant role in explaining the variance in the predictions.

8.4.1 Numerical Example

We use the following dataset:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Assume the following weights for X_2 and X_3:

Observation (i)	Weight $w_i^{(k2,j2)}$	Weight $w_i^{(k3,j3)}$
1	1.2	0.9
2	1.0	1.3
3	0.8	1.1

Let's assume model predictions $F(x_{k1}^{(gr)}, x_{i,\sim k1})$ are:

Grid Point $x_{k1}^{(gr)}$	Model Prediction $F(x_{k1}^{(gr)}, x_{i,\sim k1})$
$x_{11}^{(gr)} = 2$	21
$x_{12}^{(gr)} = 5$	36
$x_{13}^{(gr)} = 8$	49

Three-Way PDP Function: For $x_{11}^{(gr)} = 2$:

$$F_{k1}^{(k2,j2),(k3,j3)}(x_{11}^{(gr)}) = \frac{1}{3.26} \times (21 \times 1.2 \times 0.9 + 21 \times 1.0 \times 1.3 + 21 \times 0.8 \times 1.1) = 21$$

For $x_{12}^{(gr)} = 5$:

$$F_{k1}^{(k2,j2),(k3,j3)}(x_{12}^{(gr)}) = \frac{1}{3.26} \times (36 \times 1.2 \times 0.9 + 36 \times 1.0 \times 1.3 + 36 \times 0.8 \times 1.1) = 36$$

For $x_{13}^{(gr)} = 8$:

$$F_{k1}^{(k2,j2),(k3,j3)}(x_{13}^{(gr)}) = \frac{1}{3.26} \times (49 \times 1.2 \times 0.9 + 49 \times 1.0 \times 1.3 + 49 \times 0.8 \times 1.1) = 49$$

Pair	H^2 Score
${H^2_{k1,k2}}^{(PD)}$	0.7
${H^2_{k1,k3}}^{(PD)}$	0.6
${H^2_{k2,k3}}^{(PD)}$	0.8

Table 1: Pairwise H^2 Scores

Three-Way H^2 Score: Assume the following pairwise H^2 scores:

The sum of the pairwise H^2 scores is:

$$ {H^2_{k1,k2}}^{(PD)} + {H^2_{k1,k3}}^{(PD)} + {H^2_{k2,k3}}^{(PD)} = 2.1 $$

The product of the pairwise H^2 scores is:

$$ {H^2_{k1,k2}}^{(PD)} \times {H^2_{k1,k3}}^{(PD)} \times {H^2_{k2,k3}}^{(PD)} = 0.336 $$

The cube root of the sum is:

$$ \left({H^2_{k1,k2}}^{(PD)} + {H^2_{k1,k3}}^{(PD)} + {H^2_{k2,k3}}^{(PD)} \right)^{1/3} = 2.1^{1/3} \approx 1.277 $$

Finally, the three-way H^2 score is:

$$ H^2_{k1,k2,k3} = 2.1 - \frac{0.336}{1.277} \approx 1.837 $$

9 Three-Way Interaction PDPs (Equation 21)

Equation 21 estimates the three-way Partial Dependence (PD) function, which captures interactions between three variables (x_{k1}, x_{k2}, x_{k3}) in a machine learning model. These interactions provide insights into how combinations of three variables jointly influence the model's predictions.

$$ F_{k1}^{(k2,**),(k3,**)}(x_{k1}^{(gr)}) = \frac{1}{Z} \sum_{i \in S(k_2,*) \cap S(k_3,*)} F(x_{k1}^{(gr)}, x_{i,\sim k1}) \times \left(w_i^{(k2,*)} w_i^{(k3,*)} \right) $$

Where:

- $F_{k1}^{(k2,**),(k3,**)}(x_{k1}^{(gr)})$ is the three-way PD function for variable x_{k1}, given interacting variables x_{k2} and x_{k3} at high or low values.
- Z is the sum of the right-hand-side weights.
- $S(k_2, *)$ and $S(k_3, *)$ represent neighborhoods of variables k_2 and k_3 at their high (+) or low (-) levels.
- $w_i^{(k2,*)}, w_i^{(k3,*)}$ are the weights from the inverse propensity score for each record.

9.1 Main Usage

- Three-Way Interaction Analysis: This equation helps detect how interactions between three variables affect model predictions.

9.2 Indirect Benefits

- Model Interpretation: By estimating and visualizing the effect of three-way interactions, it becomes easier to interpret complex machine learning models.

9.3 Intuition Behind the Calculation

The intuition is that we are trying to understand how combinations of three variables x_{k1}, x_{k2}, x_{k3} influence the model's output. Equation 21 computes an "adjusted" PD function by using weighted sums over neighborhoods of high and low values for x_{k2} and x_{k3}.

9.4 Interpreting the Potential Outputs of Equation 21

The results of Equation 21 provide insights into how three variables jointly affect the model's predictions at high and low levels of interacting variables. The potential outcomes and their interpretations are as follows:

- **High PDP Values at High Levels of Interactions**: If the three-way PDP values are high for high levels of the interacting variables x_{k2} and x_{k3}, it suggests that the interaction between the three variables strongly influences the model's predictions when these interacting variables are at high levels.

- **Low PDP Values at Low Levels of Interactions**: If the three-way PDP values are significantly lower for low levels of the interacting variables, it indicates that the joint influence of these variables on the model's predictions diminishes when they are at low levels.

- **Non-Linear Relationships**: If the results for high and low values of the interacting variables differ substantially, it may indicate non-linear interactions between the three variables. This can provide deeper insights into how the combination of variables affects the model's outcome.

- **Balanced PDP Values Across High and Low Levels**: If the three-way PDP values remain consistent across high and low levels of x_{k2} and x_{k3}, it suggests that the focal variable x_{k1} has a stable influence on the model's predictions, regardless of the interacting variables.

9.5 Numerical Example

Equation 21 calculates the four moderated Partial Dependence Plots (PDPs) for the focal variable x_{k1} at high and low levels of x_{k2} and x_{k3}.

$$F_{k1}^{(k2,*),(k3,*)}(x_{k1}^{(gr)}) = \frac{1}{Z} \sum_{i \in S(k_2,*) \cap S(k_3,*)} F(x_{k1}^{(gr)}, x_{i,\sim k1}) \times (w_i^{(k2,*)} w_i^{(k3,*)})$$

We use the following dataset:

Observation (i)	X_1	X_2	X_3	Model Prediction $F(X_1, X_2, X_3)$
1	2	3	4	20
2	5	6	7	35
3	8	9	10	50

Weights for high and low levels of X_2 and X_3 are as follows:

Observation (i)	$w_i^{(k2,+)}$	$w_i^{(k2,-)}$	$w_i^{(k3,+)}$	$w_i^{(k3,-)}$
1	1.2	0.8	1.3	0.9
2	1.5	1.0	1.4	1.2
3	1.1	1.3	1.0	1.5

Model predictions at grid points of X_1:

Grid Point $x_{k1}^{(gr)}$	Observation 1	Observation 2	Observation 3
$x_{11}^{(gr)} = 2$	22	30	45
$x_{12}^{(gr)} = 5$	36	48	58
$x_{13}^{(gr)} = 8$	52	64	74

Final Results for $F_{k1}^{(k2,+),(k3,+)}$:

- For $x_{11}^{(gr)} = 2$: $F_{k1}^{(k2,+),(k3,+)}(x_{11}^{(gr)}) \approx 30.84$
- For $x_{12}^{(gr)} = 5$: $F_{k1}^{(k2,+),(k3,+)}(x_{12}^{(gr)}) \approx 46.37$
- For $x_{13}^{(gr)} = 8$: $F_{k1}^{(k2,+),(k3,+)}(x_{13}^{(gr)}) \approx 62.35$

Equation Applications Summary Table

Equations	**Model Interpretation**	**Feature Selection**	**Interaction Detection**	**Decision Making**	**Feature Engineering**
Eq. 1	✓	✓	-	-	✓
Eq. 2	✓	-	✓	✓	✓
Eq. 3	✓	-	✓	-	✓
Eq. 5	✓	✓	✓	✓	✓
Eq. 9	✓	-	-	-	✓
Eq. 10	✓	-	-	-	✓
Eq. 11	✓	-	✓	-	✓
Eq. 12	-	-	✓	-	-
Eq. 13	✓	-	✓	-	-
Eq. 17	✓	-	✓	-	-
Eq. 18	✓	-	✓	-	-
Eq. 19	✓	-	✓	-	-
Eq. 20	✓	-	✓	-	-
Eq. 21	✓	-	✓	✓	-

Table 2: Applications of the equations across different machine learning tasks.