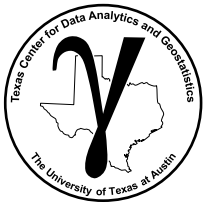


PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- Curse of Dimensionality
- Feature Selection
- **Shapley Values for Feature Importance and Model Explainability**
- Feature Selection Hands-on

Michael Pyrcz, The University of Texas at Austin

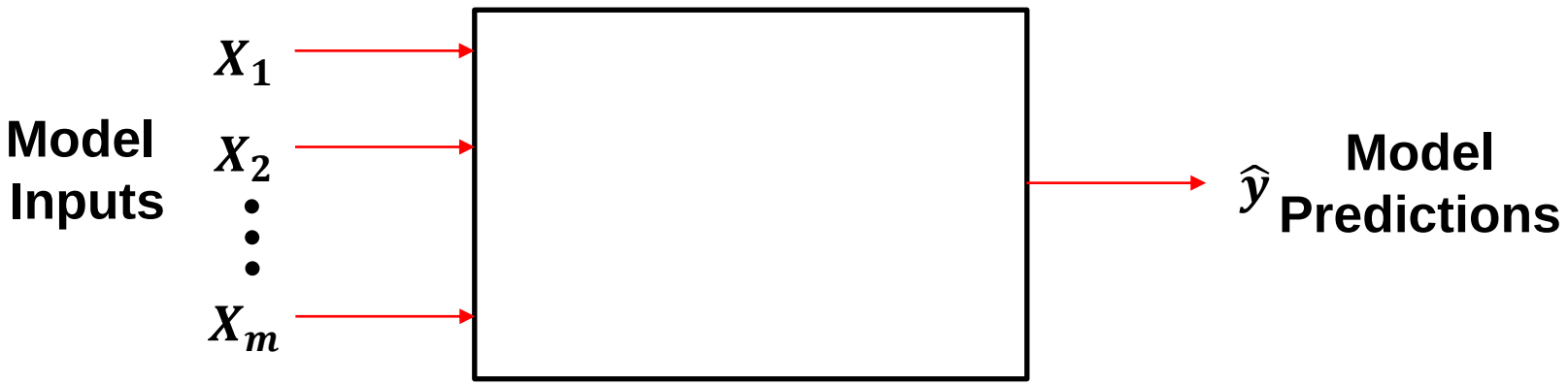
Shapley Values

Complicated models are often required but have low interpretability.

- There are 2 choices to improve model interpretability

1. reduce the complexity of the models, but also reduce model accuracy

2. develop improved, agnostic (for any model) model diagnostics



Shapley is a general method for feature importance, based on the model as a black box.

Shapley Values

Shapley Values (1953)

- The additive influence of each feature on the specific, **local**, model estimate.

$$f(x_1, \ldots, x_m) = \sum_{i=1}^{m} \phi_i$$

- where ϕ_i is a Shapley value for the impact of the feature i for the specific estimate, $\hat{y} = f(x_1, \ldots, x_m)$.

- Recall, lower case for the predictor features, $x_1, \ldots, x_m$, indicates a specific prediction case.

Shapley Values

Shapley Values' Properties

- There are a variety of Shapley Value properties that are useful, e.g. linearity, null player, and symmetry

- Efficiency – sum of the Shapley values over all features is the departure from the naïve model without the features.

$$\sum_{i=1}^{m} \phi_i = f(x_1, \ldots, x_m) - E[y]$$

Let's first explain Shapley value for games and then return to feature importance.

Shapley Values

Game Theory Approach

- Based on the Shapley value for allocating resources between 'players' based on a summarization of marginal contributions. Dividing up payment between players.

Marginal Contribution = Resource with i^{th} Player – Resource without i^{th} Player

Very Simple Shapley Example, 2 people join and work together

$$f(person_1) = \$50{,}000, f(person_2) = \$75{,}000, f(person_1, person_2) = \$125{,}000$$

Allocation player 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

$$\frac{1}{2}\$50k + \frac{1}{2}(\$125 - \$75k) = \$50k$$

Allocation player 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}\$75k + \frac{1}{2}(\$125 - \$50k) = \$75k$$

Shapley Values

Game Theory Approach

- Based on the Shapley value for allocating resources between 'players' based on a summarization of marginal contributions. Dividing up payment between players.

Marginal Contribution = Resource with i^{th} Player – Resource without i^{th} Player

Very Simple Shapley Example, 2 people join and work together with synergy!

$$f(person_1) = \$30{,}000, f(person_2) = \$70{,}000, f(person_1, person_2) = \$120{,}000$$

Allocation player 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

$$\frac{1}{2}\$30k + \frac{1}{2}(\$120 - \$70k) = \$40k$$

Allocation player 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}\$70k + \frac{1}{2}(\$120 - \$30k) = \$80k$$

Shapley Values

Game Theory Approach

- Based on the Shapley value for allocating resources between 'players' based on a summarization of marginal contributions. Dividing up payment between players.

Marginal Contribution = Resource with i^{th} Player – Resource without i^{th} Player

Very Simple Shapley Example, 2 people join and work together with dysergy!

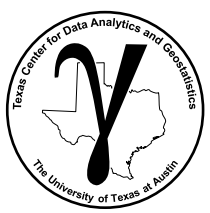
$$f(person_1) = \$30{,}000, f(person_2) = \$70{,}000, f(person_1, person_2) = \$80{,}000$$

Allocation player 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

$$\frac{1}{2}\$30k + \frac{1}{2}(\$80 - \$70k) = \$20k$$

Allocation player 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}\$70k + \frac{1}{2}(\$80 - \$30k) = \$60k$$



Shapley Values

We work out the contribution of each player through summarization over marginal contributions.

Now change:

- player → feature, x_i, and
- earnings → model prediction, $f(x)$

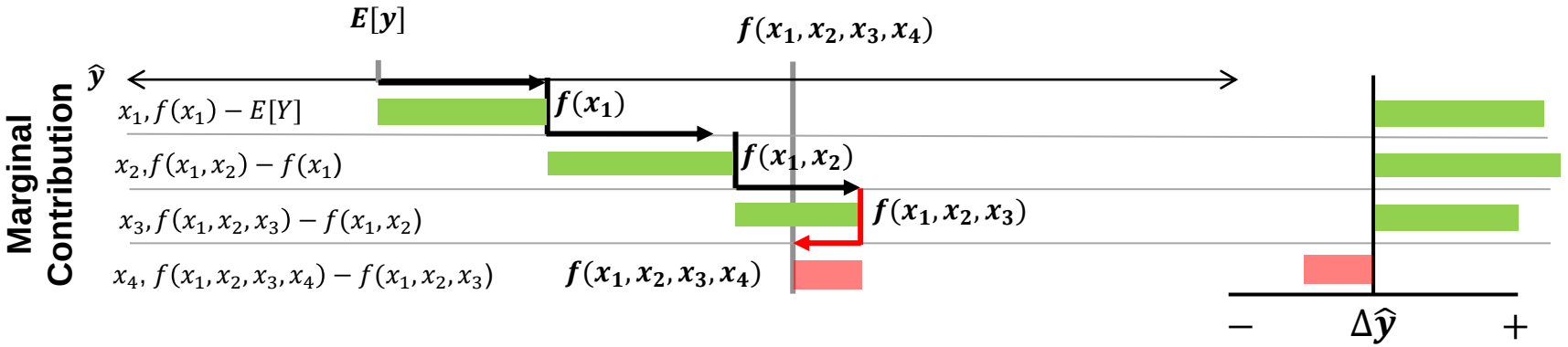
and we now have a measure of local feature importance.

Shapley Values

Feature Contribution via Local Feature Importance

- Local Feature Importance – representing a specific prediction case $(x_1, \ldots, x_m)$.
- Marginal contribution, additive components for each feature:

$$\boldsymbol{E[y] \rightarrow f(x_1) \rightarrow f(x_1, x_2) \rightarrow f(x_1, x_2, x_3) \rightarrow f(x_1, x_2, x_3, x_4)}$$



- Recall $E[y]$ is the expectation of all response training values, i.e. no information from the predictor features.

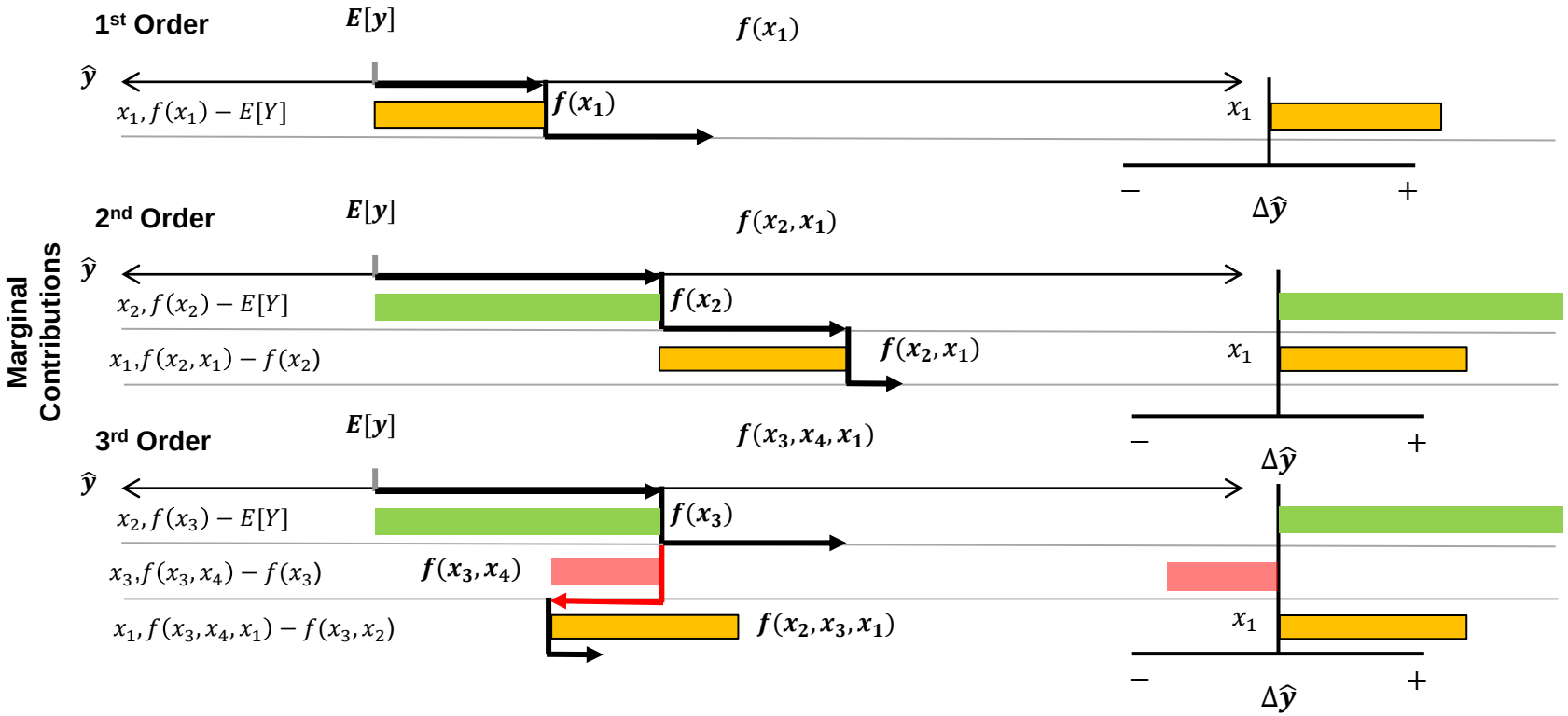
Remaining Issues:

1. What if we change the order?
2. We may not want to build multiple models (called the Naïve Case), we may want to assess feature important for a specific model.
3. We may want a global importance for all possible predictions, not a specific case, $x_1, \ldots, x_m$.

Shapley Values

Impact of Order, Number of Features:

- We could evaluate marginal contribution for any possible order, number of previous features added:



The number of previous features included, order, impacts the marginal contribution.

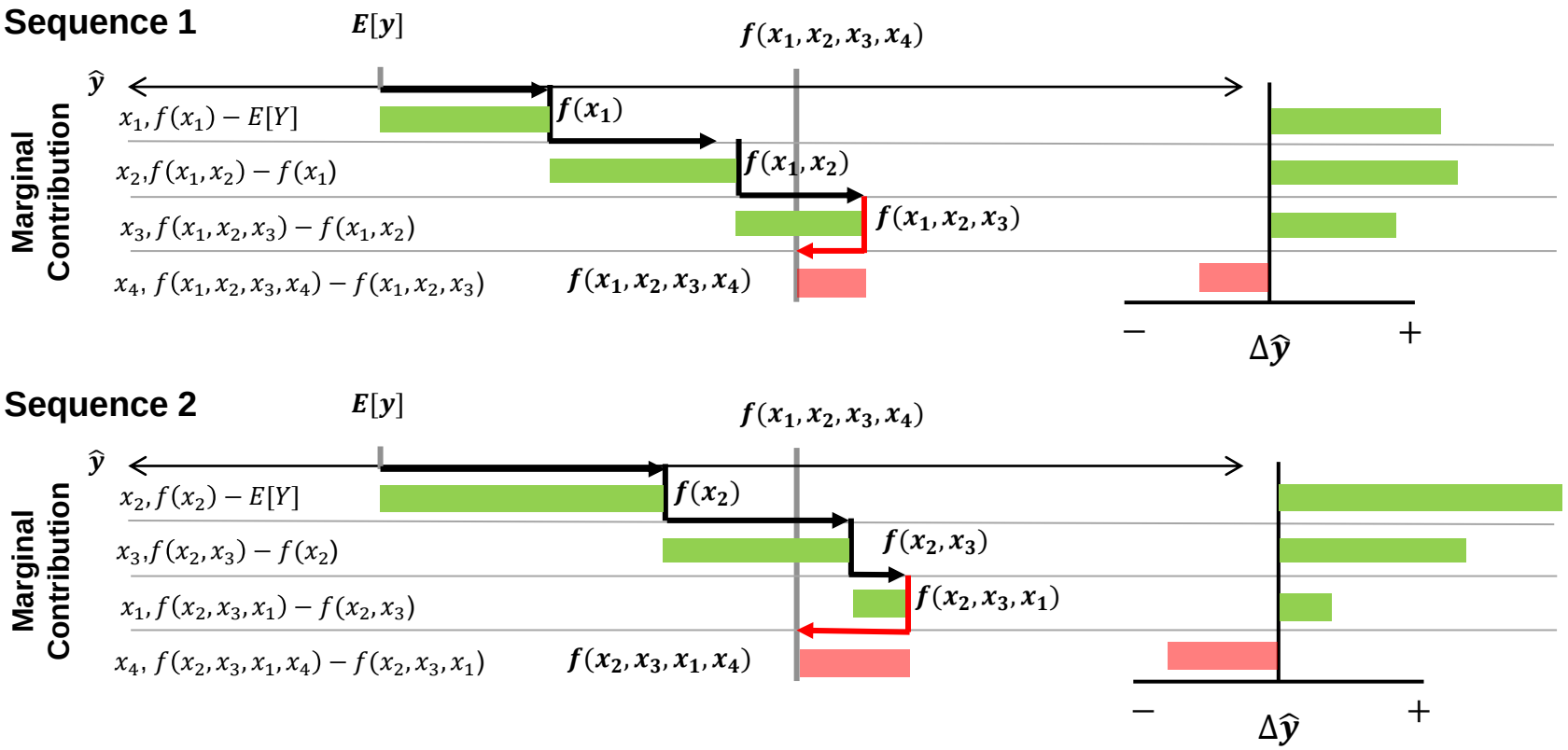
- We will average the marginal contribution over the combinatorial of orders.

Michael Pyrcz, The University of Texas at Austin

Shapley Values

Impact of Feature Sequence:

- We could evaluate marginal contribution for any possible sequence:



Due to interactions between predictor features, the feature sequence matters!

- We will average the marginal contribution over the combinatorial of sequences.

Shapley Values

We need to take a single model, $f(x_1, x_2, x_3, x_4)$, and make an estimate for all possible combinations features available!

- Note: the "naïve approach" is to train the full combinatorial of models. We don't want to do that if our goal is feature importance to diagnose our specific model, f. We want to **support model explainability**.

The variety of approaches are similar to imputation methods:

$$f(x_1, x_2, x_3) = f(x_1, x_2, x_3, x_4 = E[X_4])$$

$$f(x_1, x_2, x_3) = f(x_1, x_2, x_3, x_4 = P50_{x_4})$$

There is an unique method with tree-based models:

- Remove x_4 by averaging response prediction over all branches with x_4.

Shapley Values

Shapley Value Equation

Averaging over all possible subsets, orders of marginal contribution

Shapley value, ϕ_i, for the local importance of the i feature:

Marginal Contribution

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\ (|F| - |S| - 1)}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

where:

Weighted average, equal weight for all combinations

Our model prediction with i

Our model prediction without i

$|S|$ size of the subset before we add the i[th] feature

|F| number of features

$\left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$ the marginal contribution

$\frac{|S|!\ (|F| - |S| - 1)}{|F|!}$ is the weight for combinations for this occurrence

$S \subseteq F \setminus \{i\}$ is all possible subsets without i feature, so we can add i

$S \cup \{i\}$ is subset S with i added and S is a subset without i

Shapley Values

Shapley Value Equation

Let's explain the weighting applied to each case.

$$\frac{|S|!\ (|F| - |S| - 1)!}{|F|!}$$

Weight by the number same / reorder S cases divided by the total number of possible combinations.

$X_1, X_2 + X_3 = X_2, X_1 + X_3$

Example $F = X_1, X_2, X_3, X_4,\ \ S = X_1, X_2,\ \ i = X_3$

2 sequences for this case, we use 2 times weight.

$|S|!\,(|F| - |S| - 1)! : 2\ -\ X_1, X_2, X_3, X_4$ and X_2, X_1, X_3, X_4

$2!\,1! = 2$

$|F|! : 24\ -\ X_1, X_2, X_3, X_4,\ X_1, X_3, X_2, X_4,\ \ldots,\ X_4, X_3, X_2, X_1$

$4! = 24$

Michael Pyrcz, The University of Texas at Austin

Shapley Values

Shapley Value Equation

$$\sum_{S \subseteq F \setminus \{i\}}$$

We sum over all possible subsets without i, example subsets:

$$i = X_3, |F| = 3$$

Let's assume values, $x_1 = 10\%$, $x_2 = 150\ mD$, $x_3 = 13\%$, $x_4 = 0.54$

$f(X_3 = x_3) - E[y]$ ⟶ $f(X_1 = \overline{x_1}, X_2 = \overline{x_2}, X_3 = x_3) - f(X_1 = \overline{x_1}, X_2 = \overline{x_2}, X_3 = \overline{x_3})$

$f(X_1 = x_1, X_3 = x_3) - f(X_1 = x_1)$ ⟶ $f(X_1 = x_1, X_2 = \overline{x_2}, X_3 = x_3) - f(X_1 = x_1, X_2 = \overline{x_2}, X_3 = \overline{x_3})$

$f(X_1 = x_1, X_2 = x_2, X_3 = x_3) - f(X_1 = x_1, X_2 = x_2)$ ⟶ $f(X_1 = x_1, X_2 = x_2, X_3 = x_3) - f(X_1 = x_1, X_2 = x_2, X_3 = \overline{x_3})$

with feature X_3 without feature X_3

Shapley Values

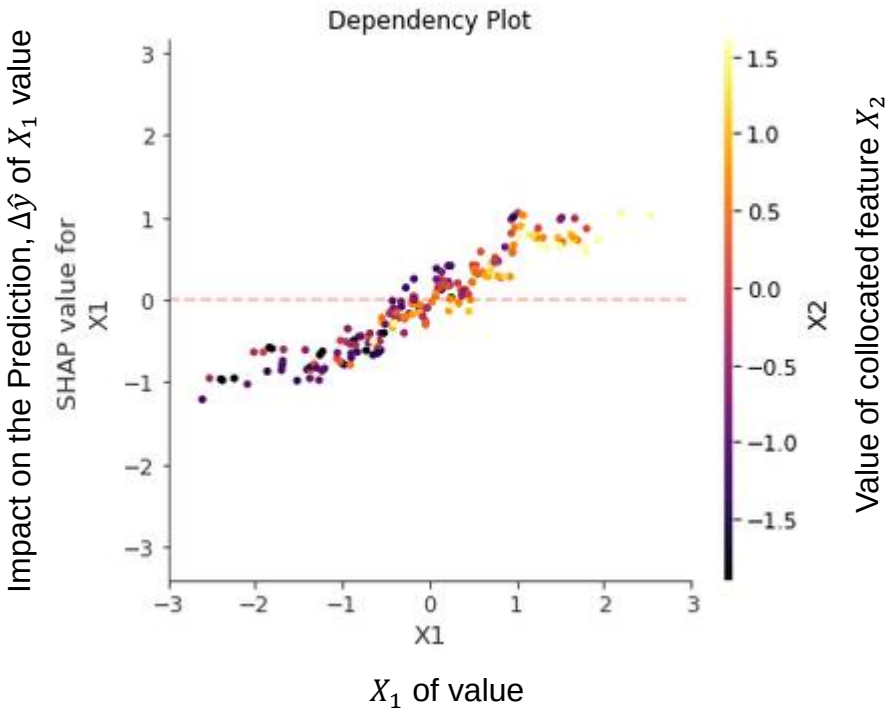
Local Feature Importance Summary by Dependency Plots

Scatter plot the local feature importance over all feature values, X_i, one plot per predictor feature.

Y-Axis:
Impact on the Prediction, $\hat{y}$ of X_i value
In units of change in the prediction Δy $\Delta\hat{y}$
Summarized over all sequence and order marginal contributions

X-Axis:
The value of the feature.

Points:
Each point is a specific background sample data, considered prediction cases.



Example dependency plot for a random forest regression model with 2 predictor features and 1 response feature.

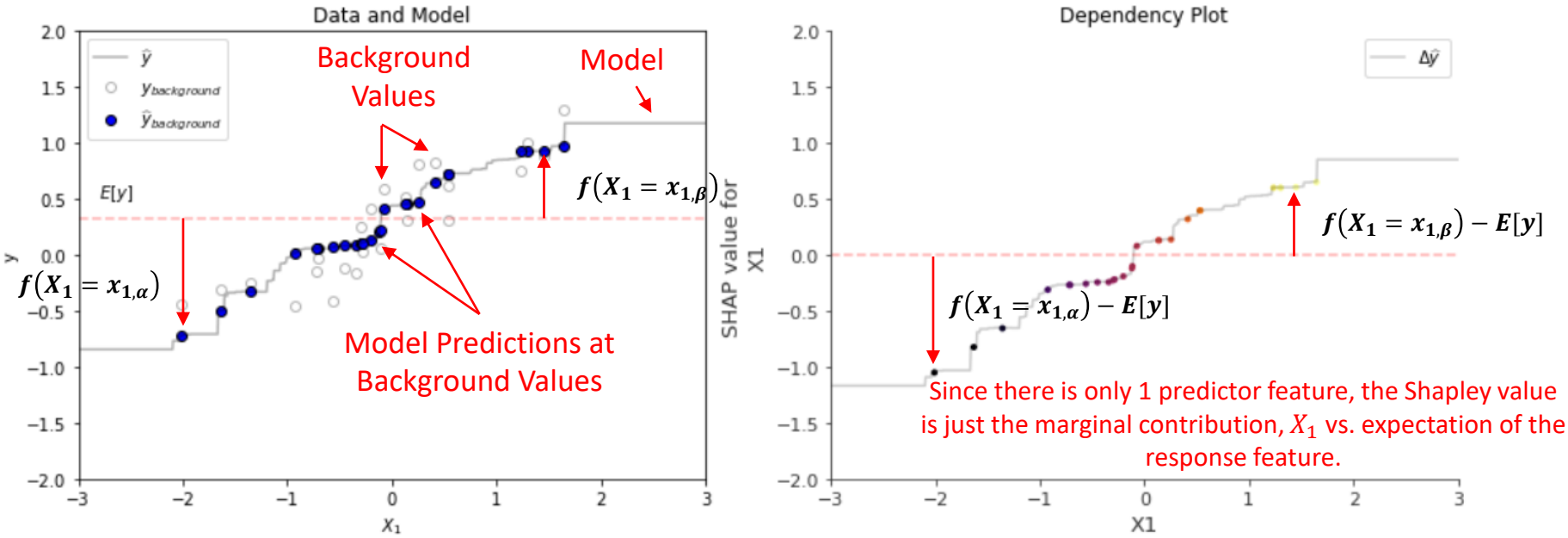
Michael Pyrcz, The University of Texas at Austin

Shapley Values

Local Feature Importance Summary by Dependency Plots

What is this plot? Let's look at 1 predictor and 1 response feature.

- With 1 predictor feature, the Shapley value is the model predictions centered on the expectation of the response feature.



Example dependency plot for a random forest regression model with 1 predictor feature and 1 response feature.

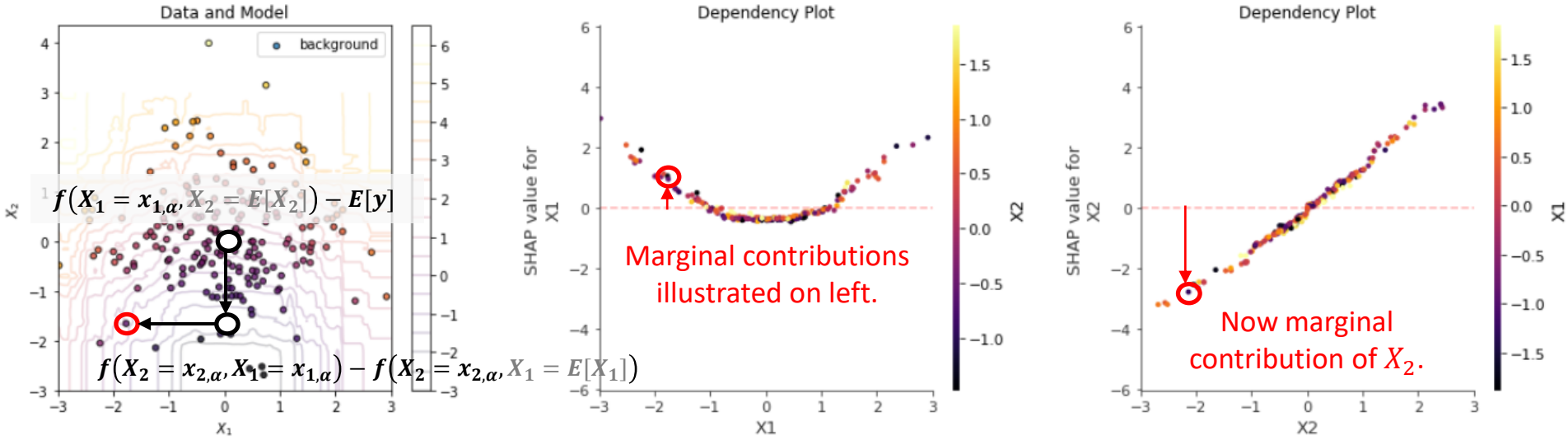
Michael Pyrcz, The University of Texas at Austin

Shapley Values

Local Feature Importance Summary by Dependency Plots

What is this plot? Now let's look at 2 predictor and 1 response feature.

- With 2 predictor feature, the Shapley value is the average of 2 marginal contributions, let's pick a background value and demonstrate



Example dependency plot for a random forest regression model with 2 predictor features and 1 response feature and the 2 marginal contributions applied in the Shapley value calculation illustrated. For illustration, we suggest that $f(X_1 = E[X_1], X_2 = E[X_2]) = E[Y]$, which of course may not be the case.

- The Shapley values' dependency plot visualizes the relationships for each feature.

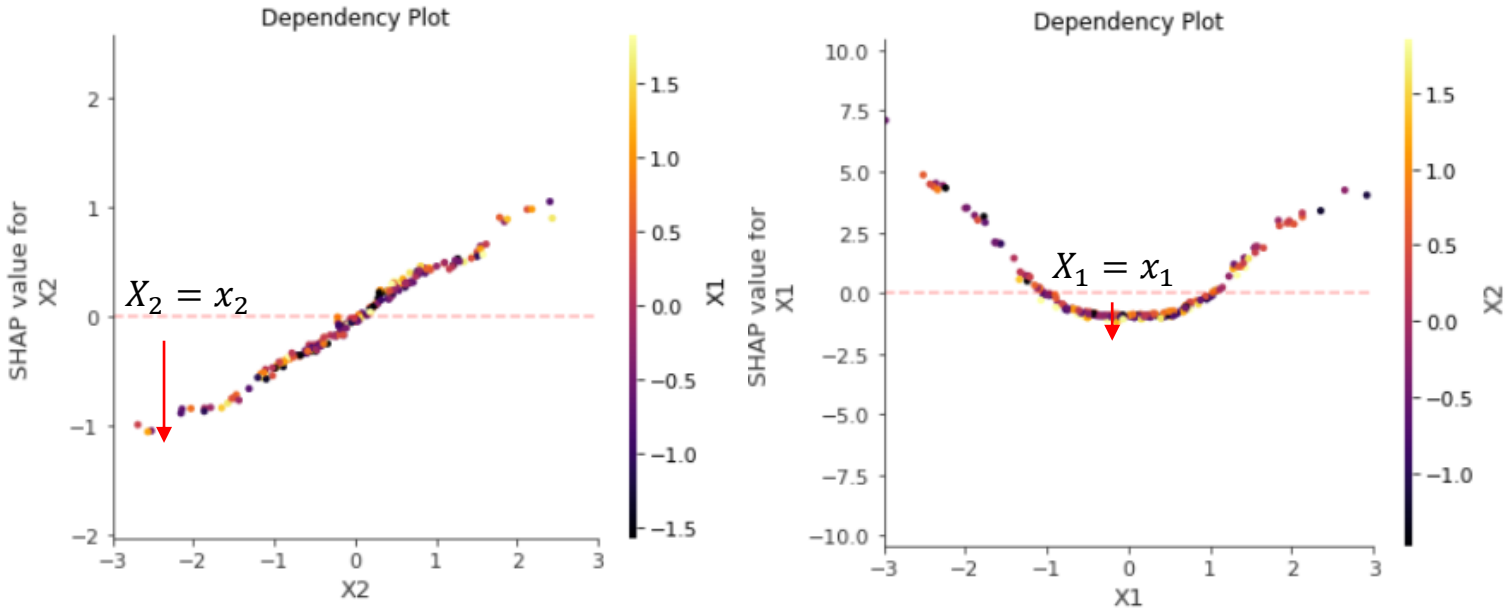
Michael Pyrcz, The University of Texas at Austin

Shapley Values

Local Feature Importance Summary by Dependency Plots

What is this plot? Now let's look at 2 predictor and 1 response feature.

- With 2 predictor feature, the Shapley value is the average of 2 marginal contributions, let's pick a background value and demonstrate



Example dependency plot for features with linear and non-linear relationships.

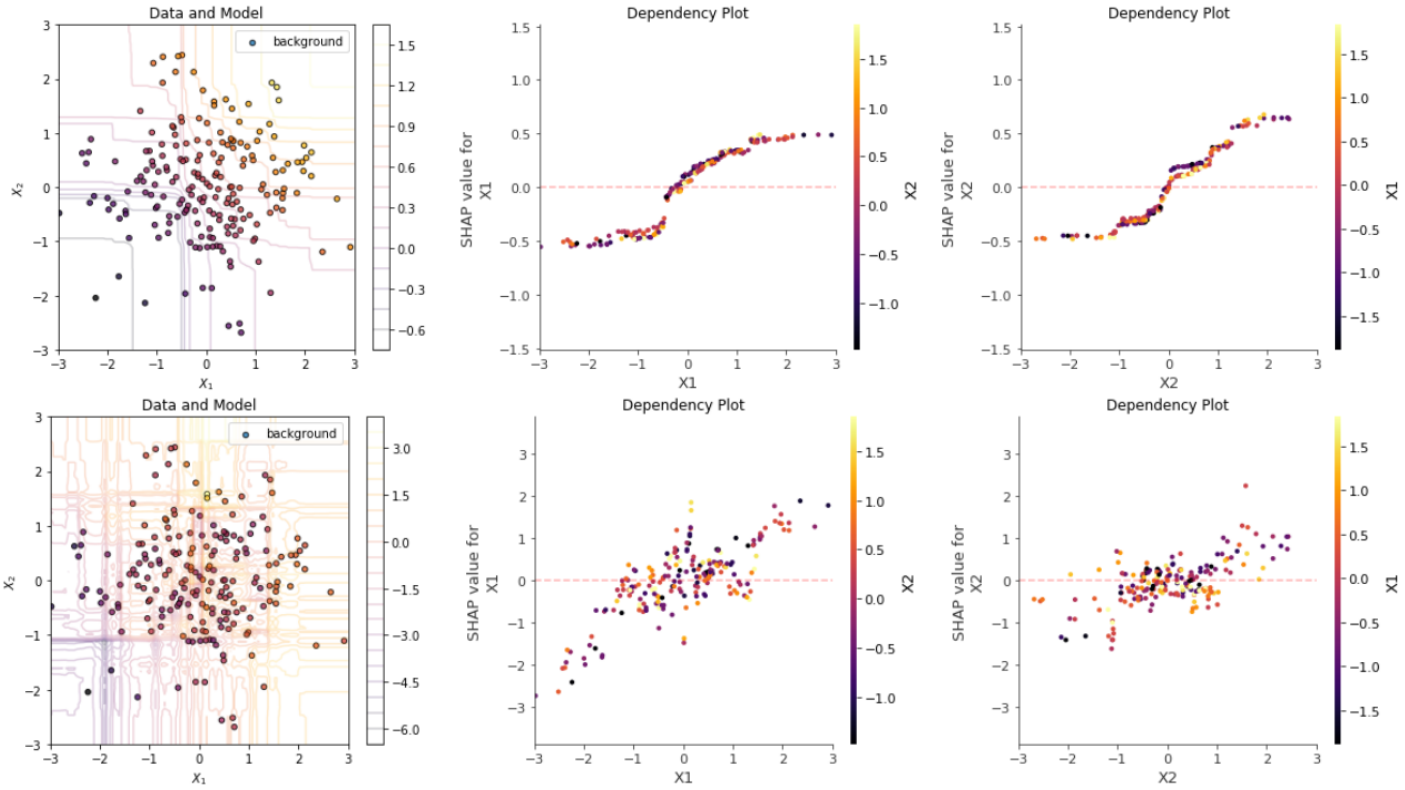
- The Shapley values' dependency plot visualizes the relationships for each feature.

Shapley Values

Local Feature Importance Summary by Dependency Plots

Feature Relationship with Noisy Data and Overfit Model

- Addition of random noise to data with complicated model.



Example dependency plot for features with reasonable and overfit model.

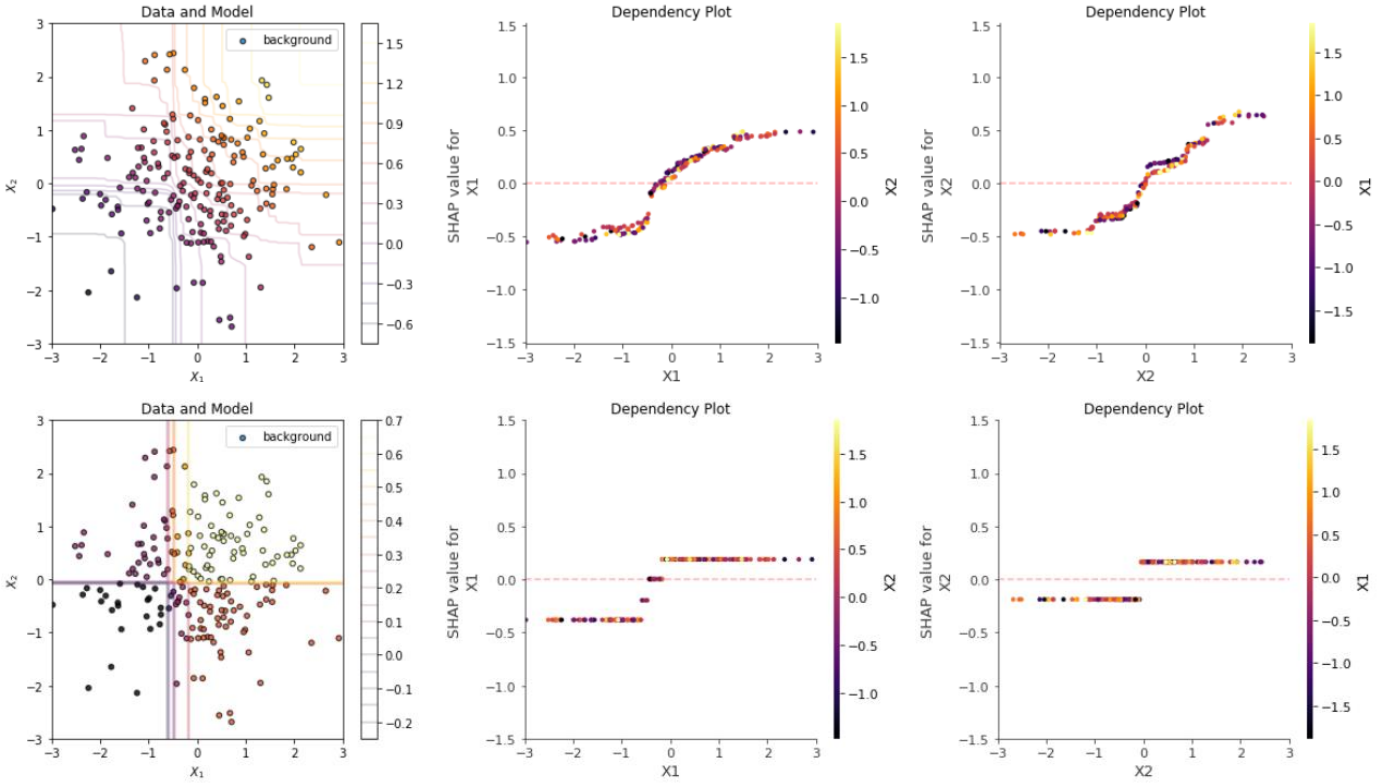
Michael Pyrcz, The University of Texas at Austin

Shapley Values

Local Feature Importance Summary by Dependency Plots

Feature Relationship with Underfit Model

- Underfit model obfuscates the feature relationships.



Example dependency plot for features with reasonable and overfit model.

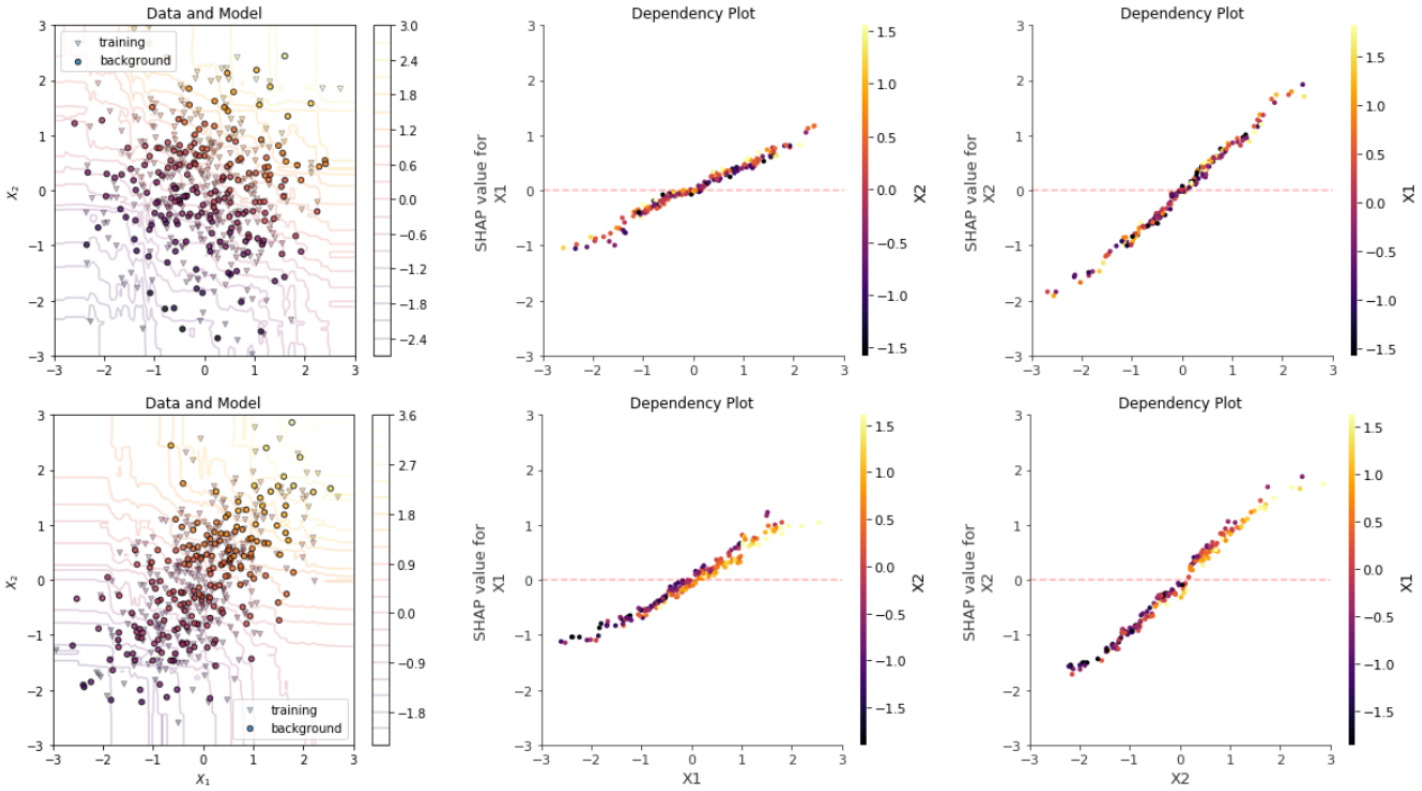
Michael Pyrcz, The University of Texas at Austin

Shapley Values

Local Feature Importance Summary by Dependency Plots

Feature Relationship with Feature Redundancy

- Imposes structure over the secondary feature on the dependency plot



Example dependency plot for predictor features with no correlation and correlation.

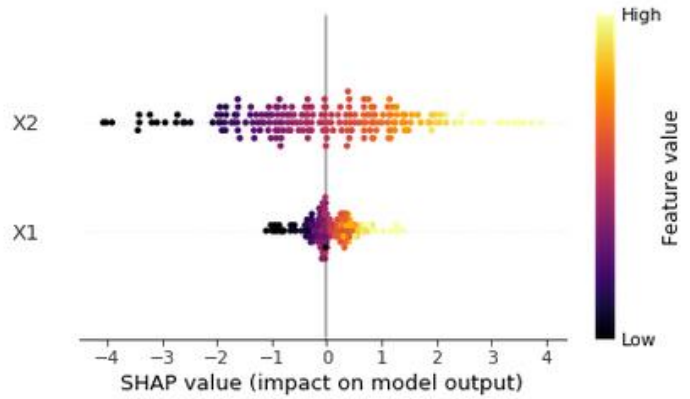
Michael Pyrcz, The University of Texas at Austin

Shapley Values

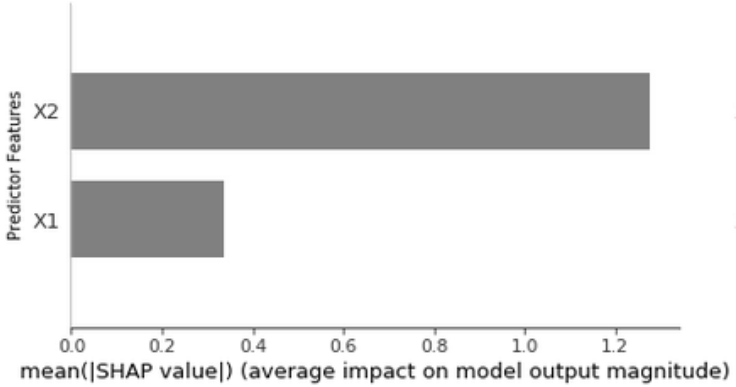
Global Feature Importance

Plot the local feature importance over all estimates, $\hat{y}$.

- Check for consistent SHAP and feature values.
- Summarize with the average of the absolute SHAP value over all background values.



Plot of Shapley Values by feature, jittered for observability and colored by secondary feature (e.g. X_2 for X_1).



Bar chart of mean absolute Shapley value.

Michael Pyrcz, The University of Texas at Austin

Shapley Value Demonstrations in Python

Demonstration of the wide array approach for feature selection with a documented workflow.

Data Analytics

Interactive Shapley Value Demonstration in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Shapley Values

Here's a demonstration of Shapley values for subsurface modeling in Python. This is part of my Subsuface Machine Learning Course at the Cockrell School of Engineering at the University of Texas at Austin.

Motivation

Complicated models are often required to model our natural settings, but have low interpretability. We have 2 choices to improve model interpretability:

1. reduce the complexity of the models, but also reduce model accuracy
2. develop improved, agnostic (for any model) model diagnostics

Shapley values are the later, they treat the model as a black box and use a data subset, known as background, with predictor features $X_1, \ldots, X_m$ and model predictions $\hat{y}$ to discover structures between each predictor feature and the model response.

$$f(x_1, \ldots, x_m) = \sum_{i=1}^{m} \phi_i$$

File Interactive_Shapley_Values.ipynb at https://git.io/Jt1os.

Subsurface Data Analytics

Feature Selection for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Subsurface Machine Learning: Feature Ranking for Subsurface Data Analytics

Here's a demonstration of feature ranking for subsurface modeling in Python. This is part of my Subsuface Machine Learning Course at the Cockrell School of Engineering at the University of Texas at Austin.

Variable Ranking

There are often many predictor features, input variables, available for us to work with for subsurface prediction. There are good reasons to be selective, throwing in every possible feature is not a good idea! In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Here's why:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?

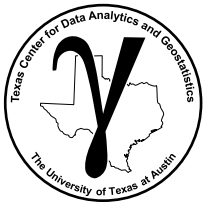
Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature. There are a wide variety of possible methods to accomplish this. My recommendation is a **'wide-array'** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking.

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries
3. Model-based

File SubsurfaceDataAnalytics_Feature_Ranking.ipynb at https://git.io/fjm4p.

Michael Pyrcz, The University of Texas at Austin



PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- Curse of Dimensionality
- Feature Selection
- **Shapley Values for Feature Importance and Model Explainability**
- Feature Selection Hands-on

Michael Pyrcz, The University of Texas at Austin