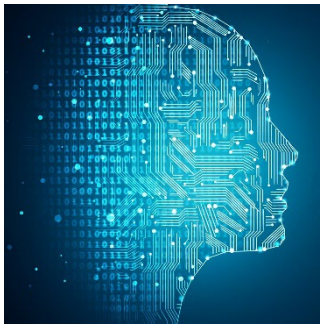
Machine Learning

Variable Importance

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655

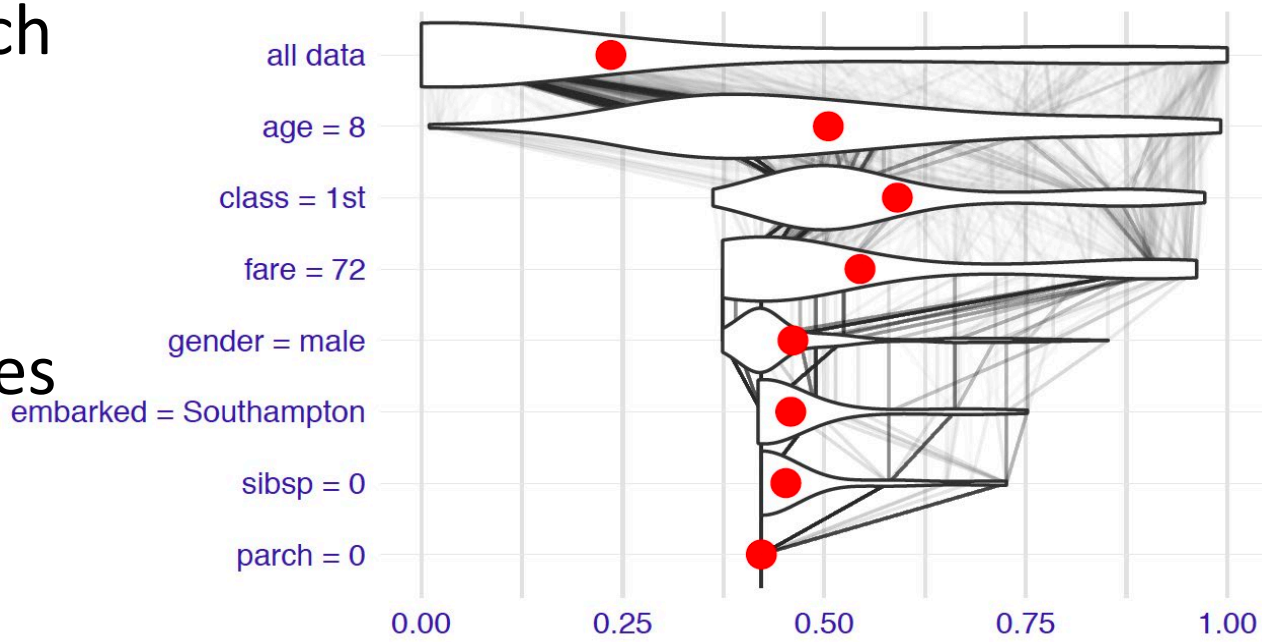


Variable Importance

- Multiple ways for computing variable importance
- We previously discussed two for RF:
  1. Mean decrease in impurity score
  2. Mean decrease in accuracy (permutation importance)
- Permutation importance is model agnostic
- We'll look at a few other ways variable importance can be analyzed

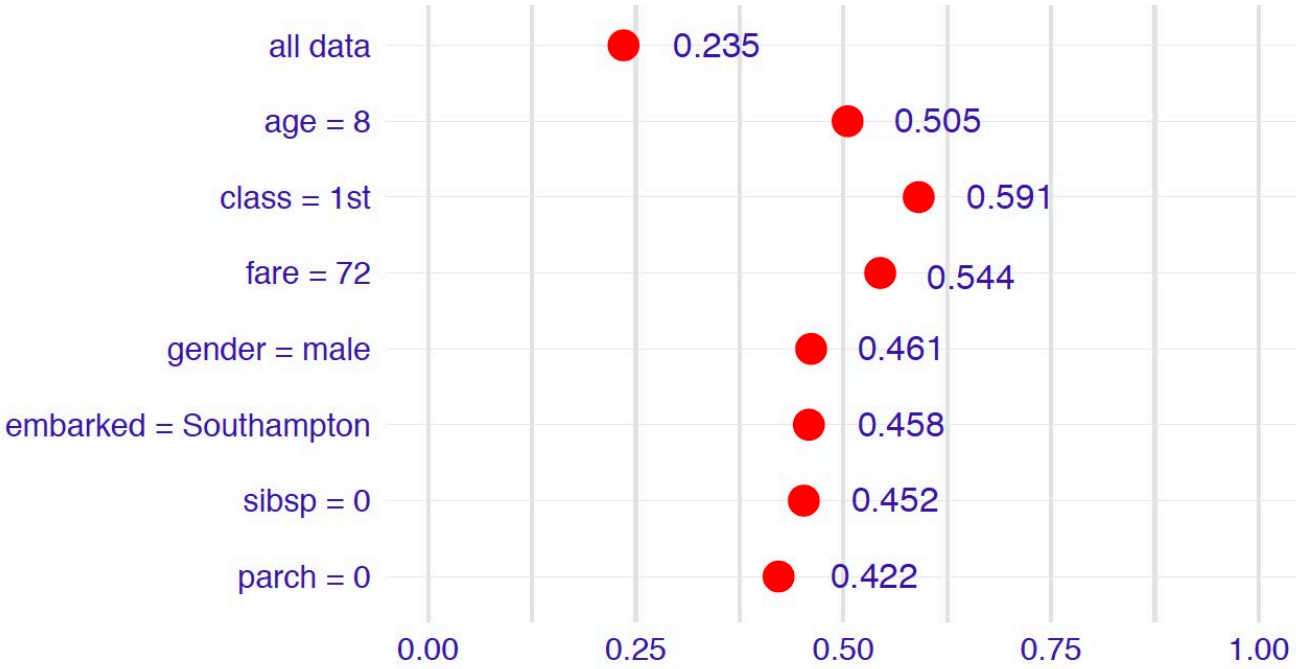
Break-down plots

- A measure of local variable importance
- Evaluates the contribution of individual variables to a particular sample prediction
- Violin plot for Johnny D in the Titanic dataset (predict whether someone was a survivor)
- Violin plot shows distribution of each variable
- Red dots are the mean
- Lines show changes in prediction



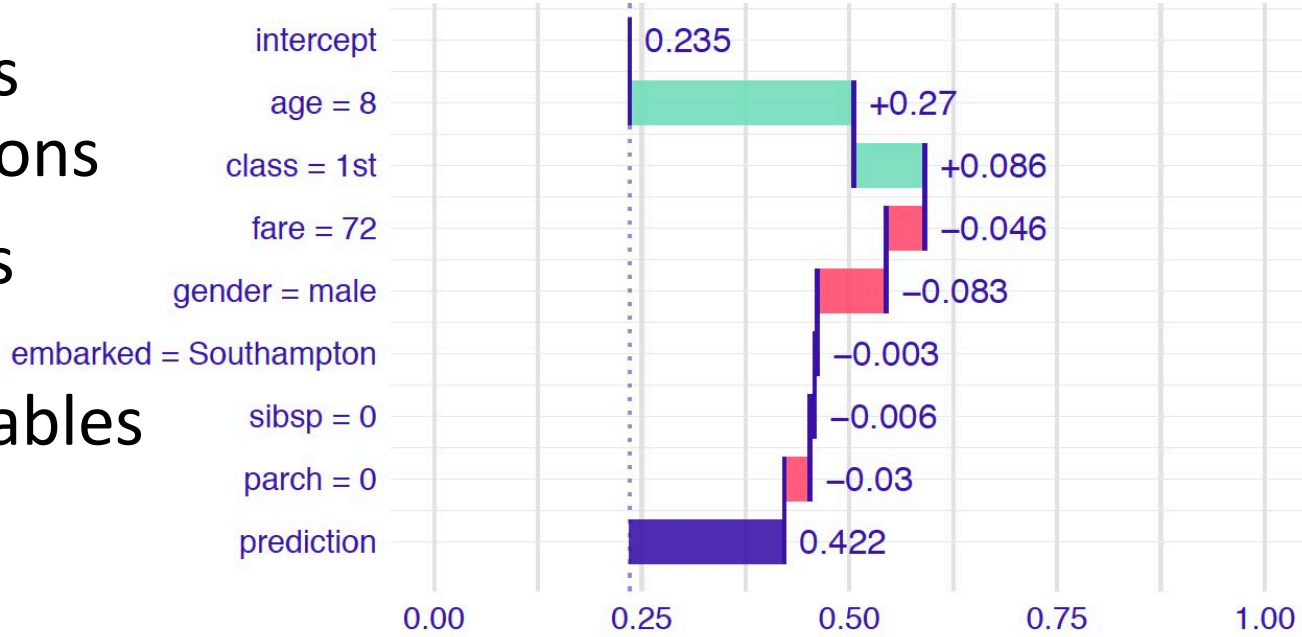
Break-down plots

- A measure of local variable importance
- Evaluates the contribution of individual variables to a particular sample prediction
- Simplified plot showing just the mean prediction



Break-down plots

- A measure of local variable importance
- Evaluates the contribution of individual variables to a particular sample prediction
- A breakdown plot for Johnny D
- Positive and negative changes in mean predictions
- I.e. contributions attributed to explanatory variables

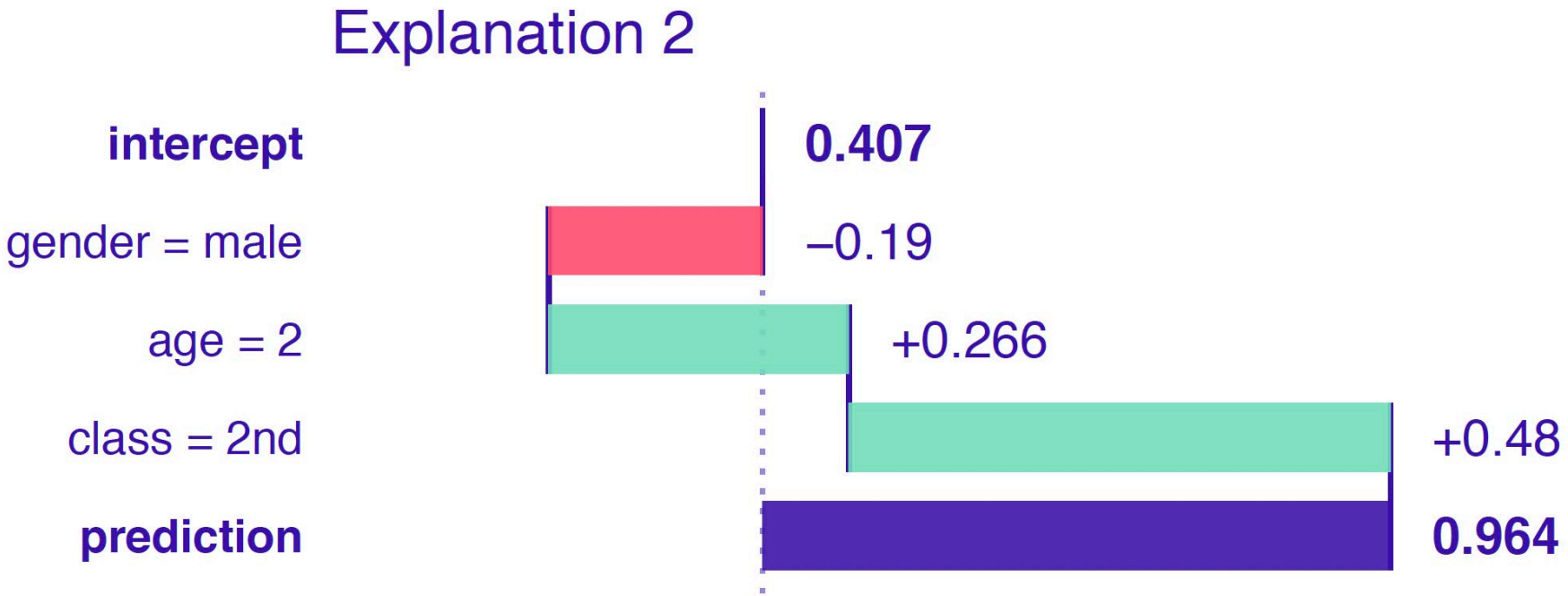


Order matters

- The order of variables matters
- **Example**: the Titanic data set with only 3 variables
- Focus on a 2-year old boy in 2nd class
- **Explanation 1**

Order matters

- The order of variables matters
- **Example**: the Titanic data set with only 3 variables
- Focus on a 2-year old boy in 2nd class
- **Explanation 2**



Break-down for linear models

- Assume a classical linear model
- Expected prediction conditioned on an instance x is

$$\mathbb{E}_Y[Y|\boldsymbol{x}] = f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$$

- Measure variable importance by evaluating how much the expected value of Y changes after conditioning on x_j:

$$v(j,\boldsymbol{x}) = \mathbb{E}_Y[Y|\boldsymbol{x}] - \mathbb{E}_{X_j}\left[\mathbb{E}\left[Y|\boldsymbol{x}_{j|=X_j}\right]\right]$$
$$= f(\boldsymbol{x}) - \mathbb{E}_{X_j}\left[f\left(\boldsymbol{x}_{j|=X_j}\right)\right]$$

  - $\boldsymbol{x}_{j|=X_j}$ indicates that the jth coordinate is a RV X_j

- This becomes

$$v(j,\boldsymbol{x}) = \beta_j\left(x_j - \mathbb{E}_{X_j}[X_j]\right)$$

Break-down for linear models

$$v(j, \boldsymbol{x}) = \beta_j \left(x_j - \mathbb{E}_{X_j}[X_j] \right)$$

- Can rewrite prediction function as:

$$f(\boldsymbol{x}) = v_0 + \sum_{j=1}^{d} v(j, \boldsymbol{x})$$

  - v_0 gives the mean prediction
- The variable contributions sum up the difference between the model's prediction for $\boldsymbol{x}$ and the mean prediction
- In practice, all the expected values (and the βs) are estimated from data
  - E.g. using the sample mean

Break-down for the general case

- Want "local accuracy"
  - Prediction equals sum of the importances

$$f(\boldsymbol{x}) = v_0 + \sum_{j=1}^{d} v(j, \boldsymbol{x})$$

- A natural definition:

$$v(j, \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{X}}\left[f(\boldsymbol{X}) | X_1 = x_1, \ldots, X_j = x_j\right] - \mathbb{E}_{\boldsymbol{X}}\left[f(\boldsymbol{X}) | X_1 = x_1, \ldots, X_{j-1} = x_{j-1}\right]$$

  - I.e. the difference between conditional expected values
  - Note this depends on the ordering of the variables
  - Let's consider more general cases

Break-down for the general case

- Let J be a subset of $K \leq d$ indices from $\{1,2,\dots,d\}$
  - I.e. $J = \{j_1, j_2, \dots, j_K\}$ with $j_k \in \{1,2,\dots,d\}$
- Let L be another subset of $M \leq d - K$ indices from $\{1,2,\dots,d\}$ distinct from J
  - $L = \{l_1, \dots, l_M\}$ with $l_m \in \{1,\dots,d\}$ and $J \cap L = \emptyset$
- Define:

$$\Delta_{L|J}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{X}}\left[f(\boldsymbol{X})|X_{l_1} = x_{l_1}, \dots X_{l_M} = x_{l_M}, X_{j_1} = x_{j_1}, \dots, X_{j_K} = x_{j_K}\right] - \mathbb{E}_{\boldsymbol{X}}\left[f(\boldsymbol{X})|X_{j_1} = x_{j_1}, \dots, X_{j_K} = x_{j_K}\right]$$

  - I.e. we consider the difference that the variables L make when already considering the variables in J
  - $\Delta_{l|J}$ considers the effect of the lth variable after considering J

Break-down for the general case

- Order can still matter, unless we consider all possible combos
- One heuristic is to choose the order based on largest contributions (in abs. value)

Break-down example

- Consider another Titanic passenger
- Variables are ordered in decreasing order of individual contributions

variable j	$E_{\underline{X}}\{f(\underline{X})\|X^j = x_*^j\}$	$\|\Delta^{j\|\emptyset}(\underline{x}_*)\|$
age = 8	0.5051210	0.2698115
class = 1st	0.4204449	0.1851354
fare = 72	0.3785383	0.1432288
gender = male	0.1102873	0.1250222
embarked = Southampton	0.2246035	0.0107060
sibsp = 0	0.2429597	0.0076502
parch = 0	0.2322655	0.0030440

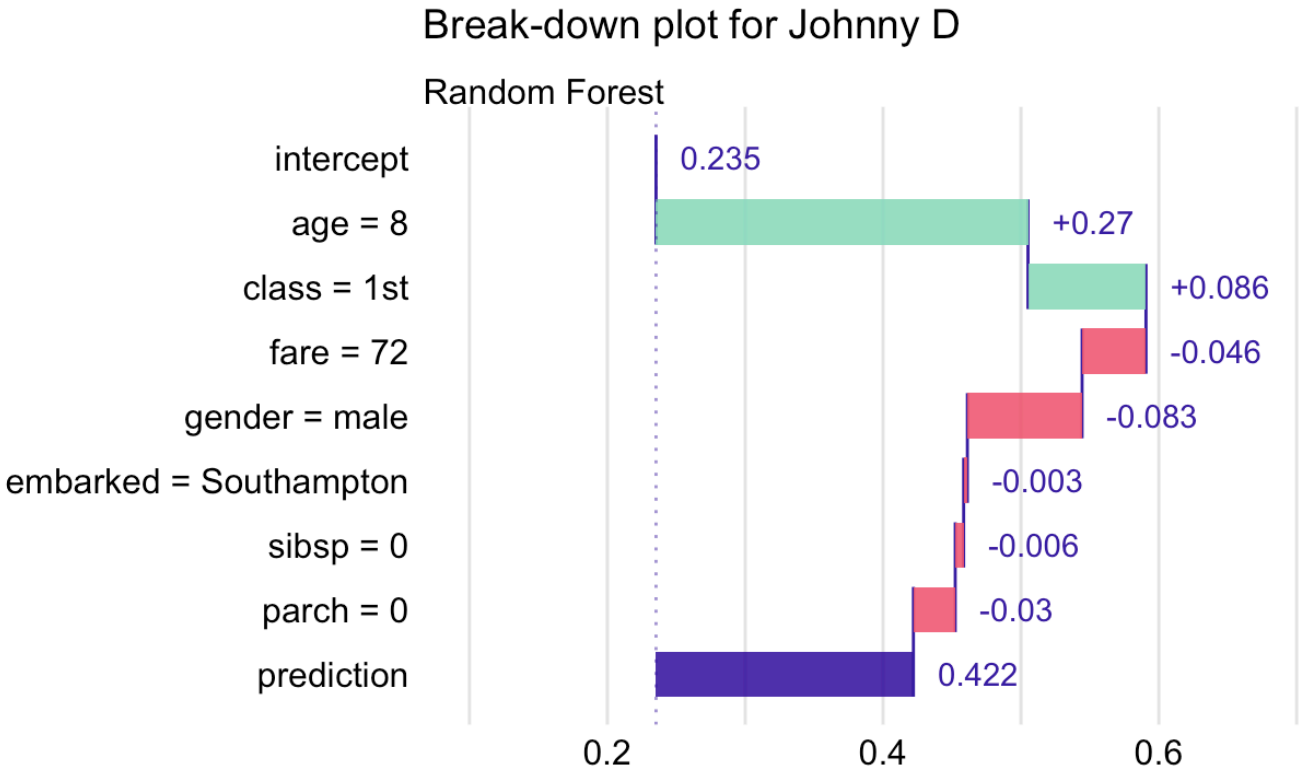
Break-down example

- Consider another Titanic passenger
- Variables are ordered in decreasing order of individual contributions

variable j	$E_{\underline{X}}\left\{f(\underline{X})\|\underline{X}^J = \underline{x}_*^J\right\}$	$\Delta^{j\|J}(\underline{x}_*)$
intercept (v_0)	0.2353095	0.2353095
age = 8	0.5051210	0.2698115
class = 1st	0.5906969	0.0855759
fare = 72	0.5443561	-0.0463407
gender = male	0.4611518	-0.0832043
embarked = Southampton	0.4584422	-0.0027096
sibsp = 0	0.4523398	-0.0061024
parch = 0	0.4220000	-0.0303398
prediction	0.4220000	0.4220000

Break-down example

- Consider another Titanic passenger
- Variables are ordered in decreasing order of individual contributions

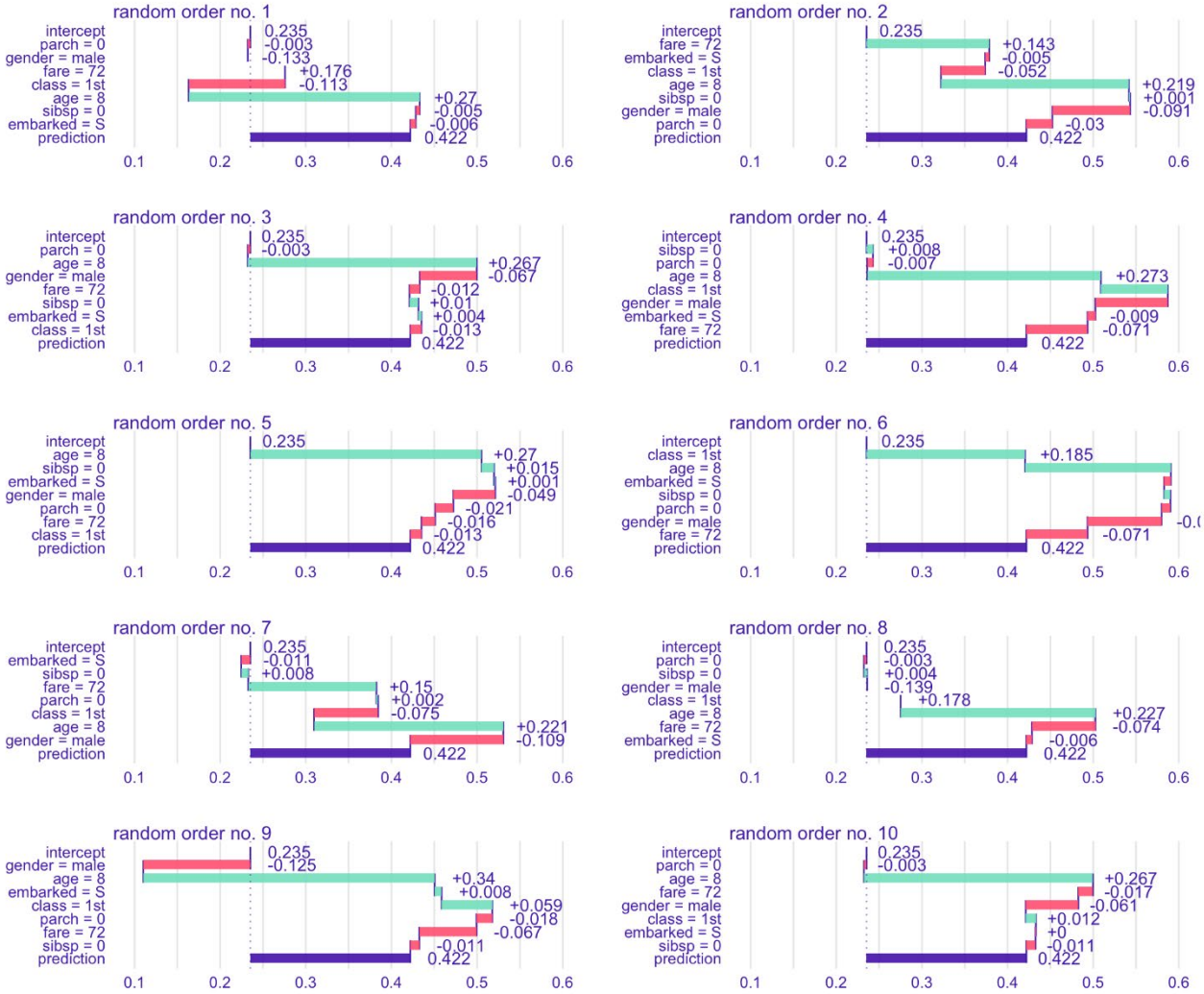


Break-down summary

- Advantages
  - Break-down is model agnostic
  - Provides generally easy to interpret plots
  - Good computational complexity
- Disadvantages
  - Misleading if there are interactions (ordering matters)
  - May be complex with many variables

Ordering affects break-down plots

- Fare and class change a lot



SHAP values

- **Idea**: average across all possible orderings
- This is what Shapley Additive exPlanations (SHAP) do
- Connected to game theory
  - A group of players cooperates and gains more
  - Players don't contribute the same amount
  - How do you distribute the excess gained?
- In prediction, variables are the players
  - Payoff is the model's prediction
  - How do you distribute the model's prediction across variables?
- Average across all possible orderings

SHAP values

- J a permutation of $\{1, \dots, d\}$
- $\pi(J, j) =$ the set of variables position in J before the jth variable
  - E.g. if the jth variable is first, $\pi(J, j) = \emptyset$
- SHAP value for the prediction at $\boldsymbol{x}$ is:

$$\phi(\boldsymbol{x}, j) = \frac{1}{d!} \sum_{J} \Delta_{j|\pi(J,j)}(\boldsymbol{x})$$

  - Sum is taken over all $d!$ Permutations
  - $\Delta_{j|\pi(J,j)}$ considers the effect of the jth variable after considering the set $\pi(J, j)$
  - I.e. $\phi(\boldsymbol{x}, j)$ averages variable importance across all orderings

SHAP values

- Note that $\Delta_{j|\pi(J,j)}(\boldsymbol{x})$ is constant for all permutations J that have the same subset $\pi(J,j)$
- Alternate form:

$$\phi(\boldsymbol{x},j)=\frac{1}{d}\sum_{s=0}^{d-1}\sum_{\substack{S\subseteq\{1,\ldots,d\}\setminus\{j\}\\|S|=s}}\left[\binom{d-1}{s}^{-1}\Delta_{j|S}(\boldsymbol{x})\right]$$

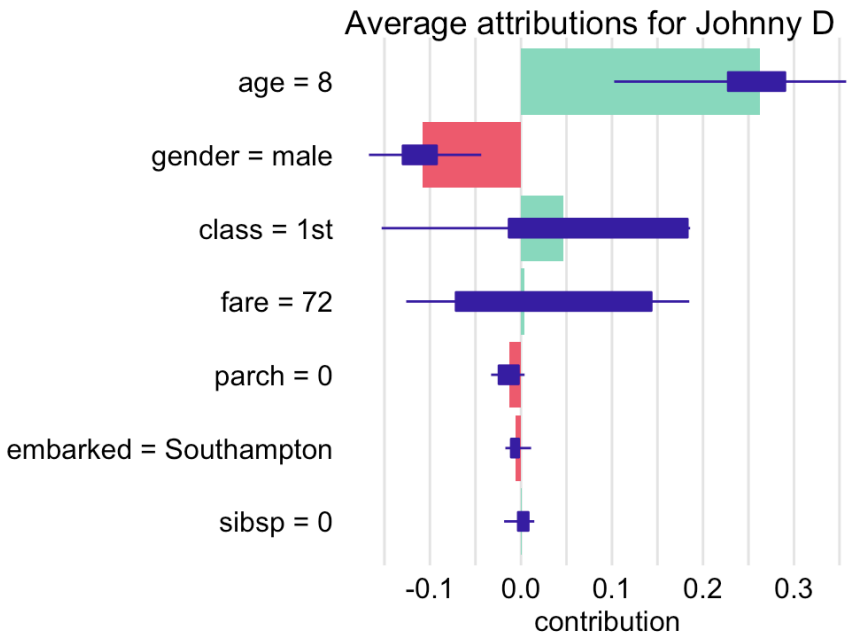
- Number of subsets from 0 to $d-1$ is 2^d-1
  - Smaller than $d!$ but still large
  - Can consider a sample of permutations for faster computation (this is what current implementations do)
  - Tree-based methods have effective implementations

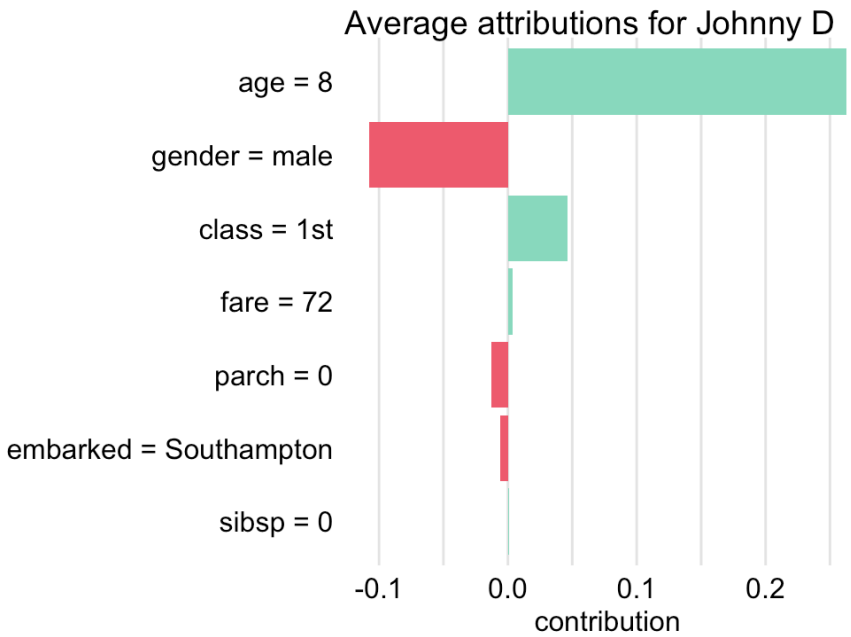
SHAP value properties

- **Symmetry**: if two variables are interchangeable, then their SHAP values are equal
- **Dummy feature**: if a variable does not contribute to any prediction, then its SHAP value is zero
- **Additivity**: if a model f is the sum of g and h, then the SHAP values for f are the sum of the SHAP values for g and h
- **Local accuracy**: sum of SHAP values equals the model prediction

SHAP value example

- Consider the Titanic passenger from before
- SHAP values based on 25 random orderings





SHAP value summary

- Advantages
  - Solid game theory backing
  - Unifies different approaches to additive variable importances
  - Efficient implementations
- Disadvantages
  - Can be misleading if the prediction model is not additive

Partial-dependence plots

- Variable importance measures give an overall measure
- Both break-down plots and SHAP values show the relevance for individual samples
- Sometimes, we want to know how important a variable is based on its value
  - E.g. for predicting someone's income, a person's degree has less of an effect if it is a PhD compared to an MBA
- Partial dependence (PD) plots look at how the expected value of model predictions change as a function of the variable

Partial-dependence plot uses

Also useful for comparing different models

- Can show agreement between models
  - If the PD plots for a complex and simple model agree, that suggests the complex model is not overfitting
- Disagreement may suggest ways to improve a model
  - If a simpler model has different PD plot than a complex model, then a variable transformation may help the simple model
- Evaluation of model performance at boundaries
  - Models typically behave differently at the boundaries
    - E.g. RFs tend to shrink predictions to the average, while the SVM has larger variance
    - Can see this using PD plots

Partial-dependence plots

- Consider a prediction model f and variable X_j at z
- PD profile:

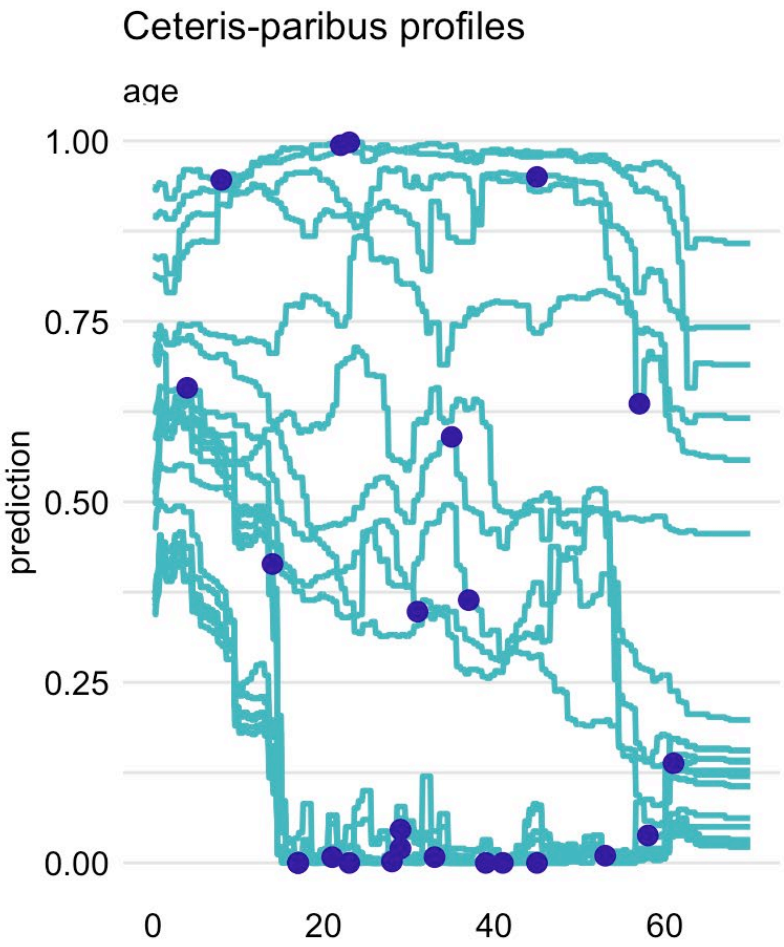
$$g_{PD}^{j}(z) = \mathbb{E}_{\boldsymbol{X}^{-j}}\left[f\left(X_{j|=z}\right)\right]$$

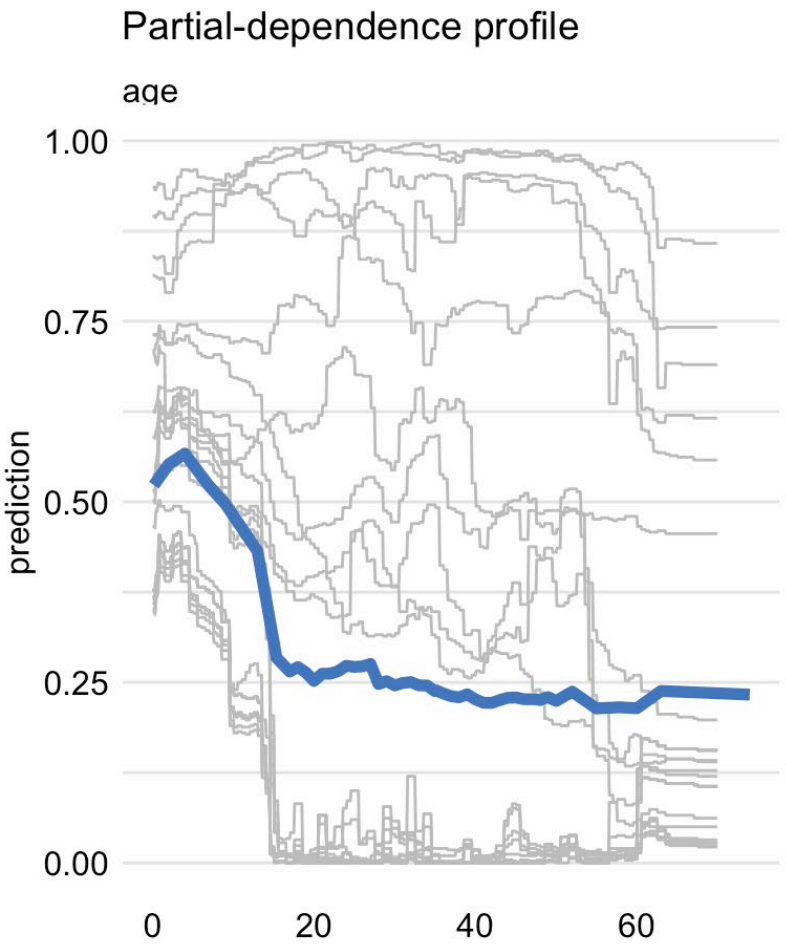
  - I.e. the expected value of the model predictions when X_j is fixed at z
  - Expectation is taken wrt to all other variables
- Typically estimate from data

$$\hat{g}_{PD}^{j}(z) = \frac{1}{n}\sum_{i=1}^{n} f\left(\boldsymbol{x}_{j|=z}(i)\right)$$

PD plot example

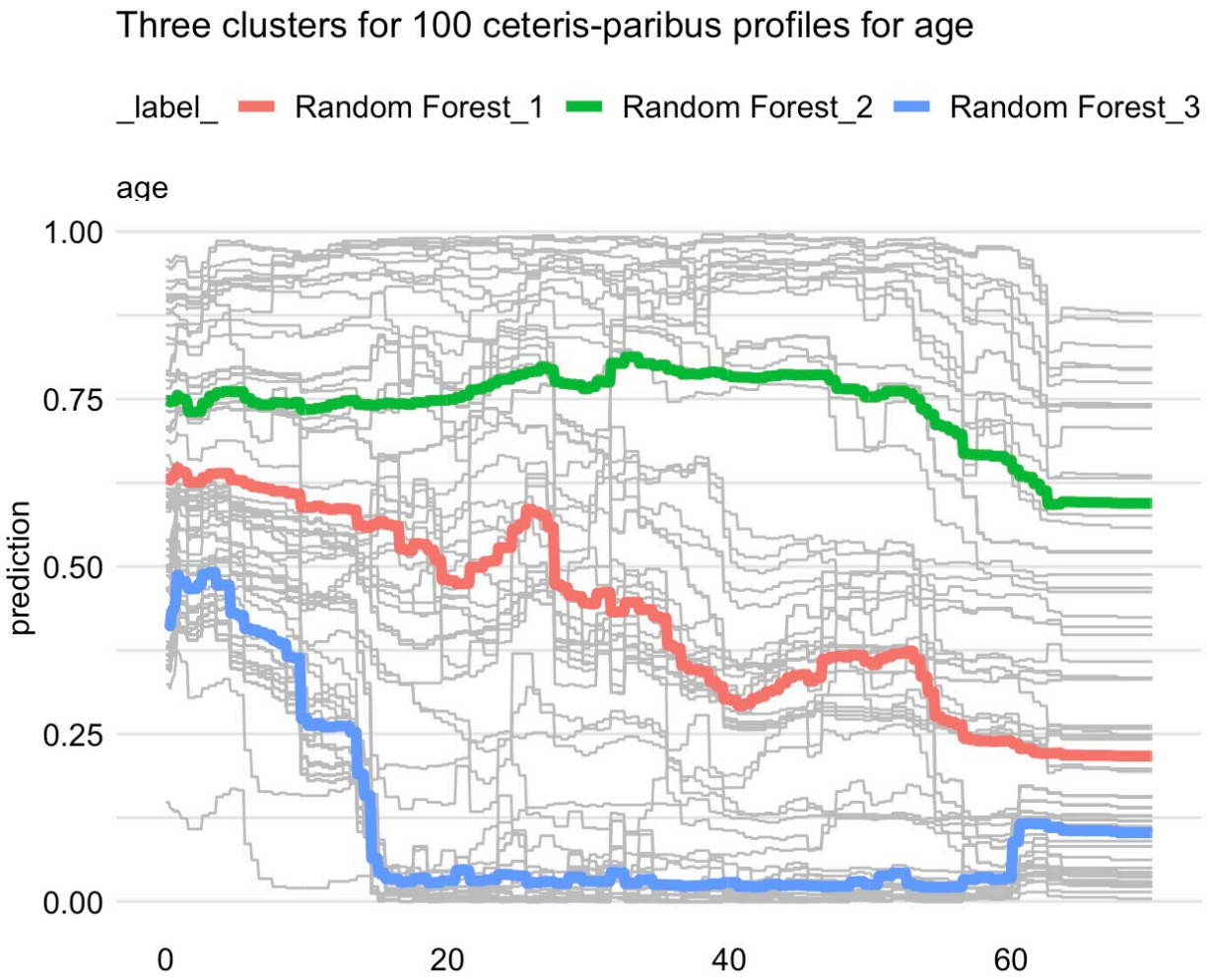
- Titanic dataset





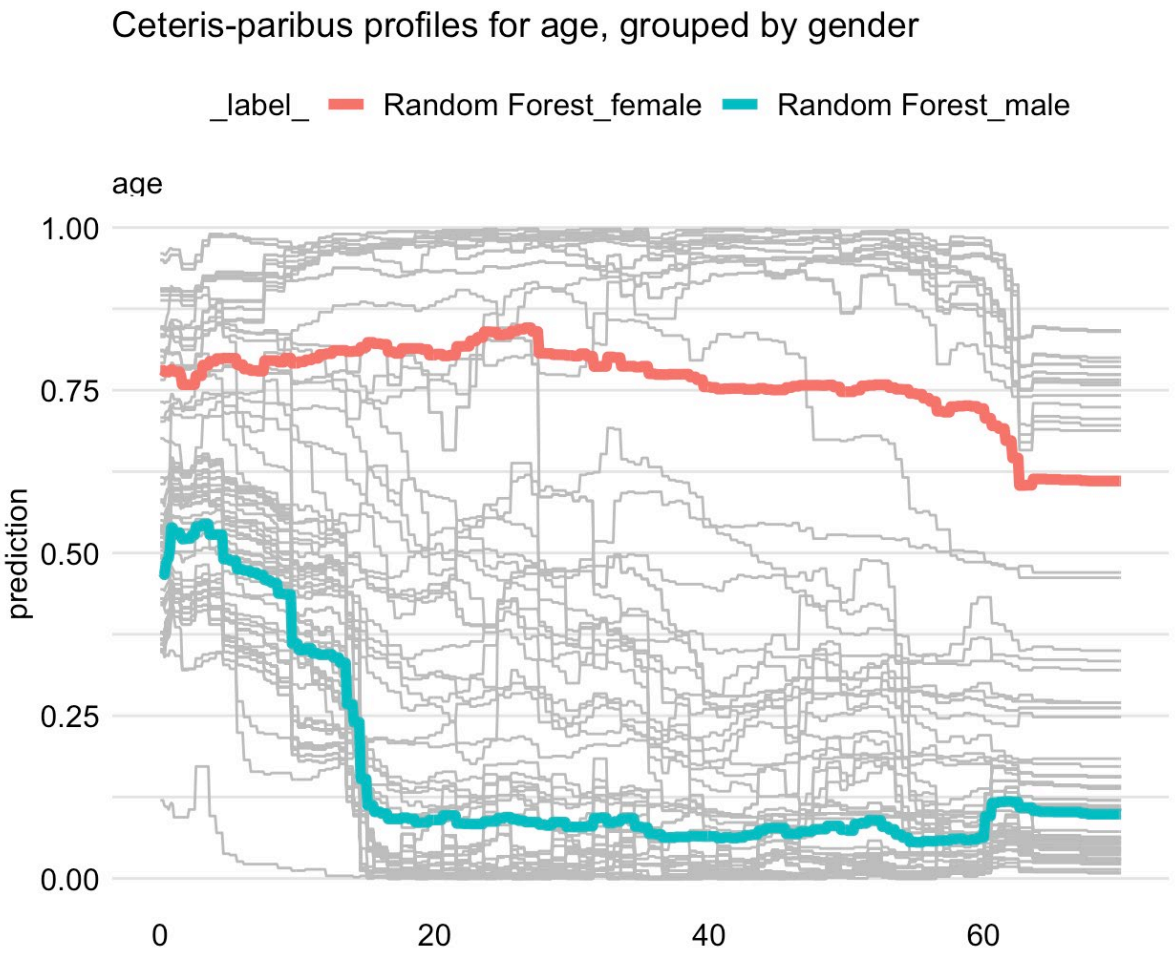
Clustered PD plots

- Can cluster instead of average



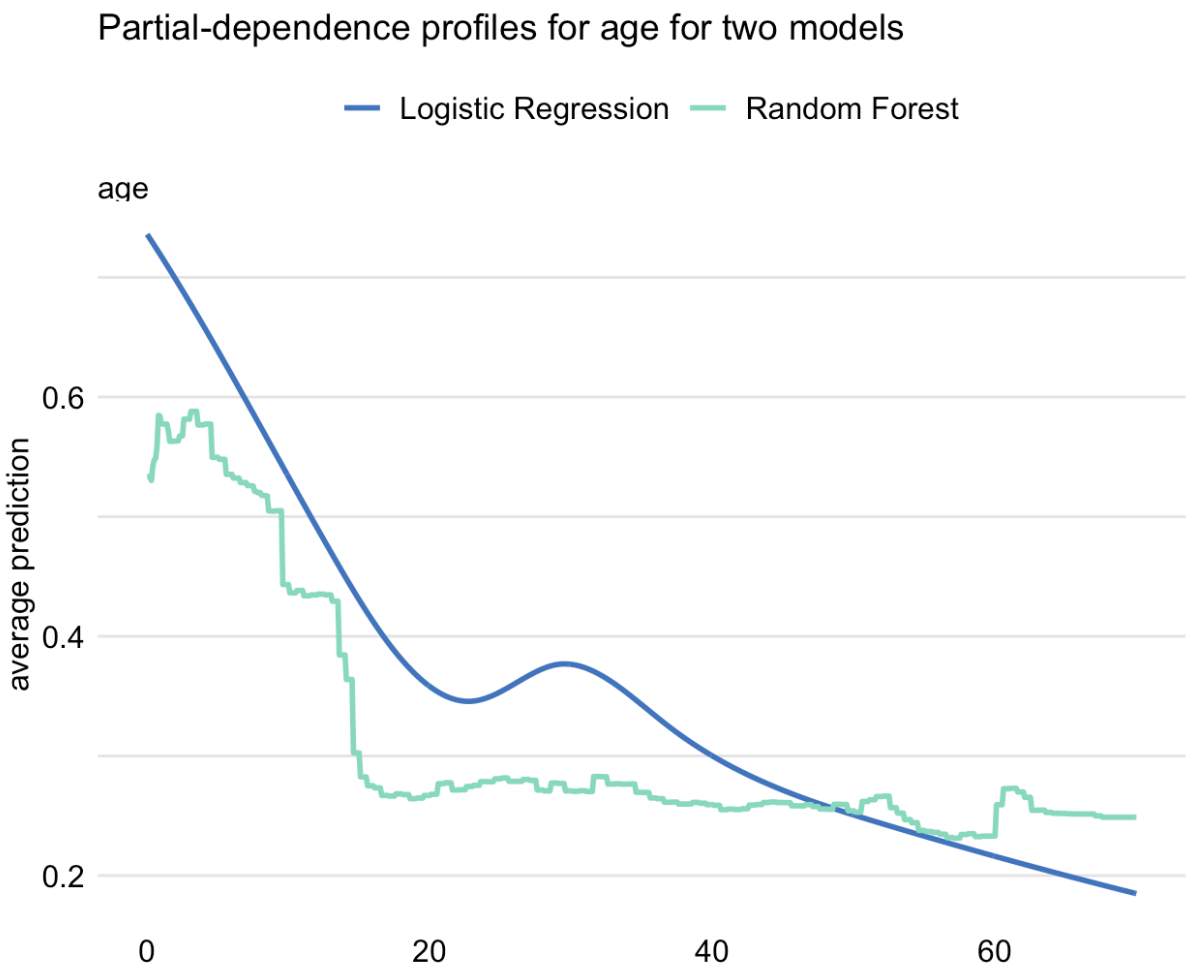
Grouped PD plots

- Can try to identify variable interactions



Contrastive PD plots

- Can compare PD plots for different models



Summary

- There are many ways for measuring variable importance and interactions
- Often a good idea to look at multiple measures
- There has been a lot of work done in explanatory model analysis

Further reading

- This book has a lot on explanatory model analysis: https://ema.drwhy.ai/