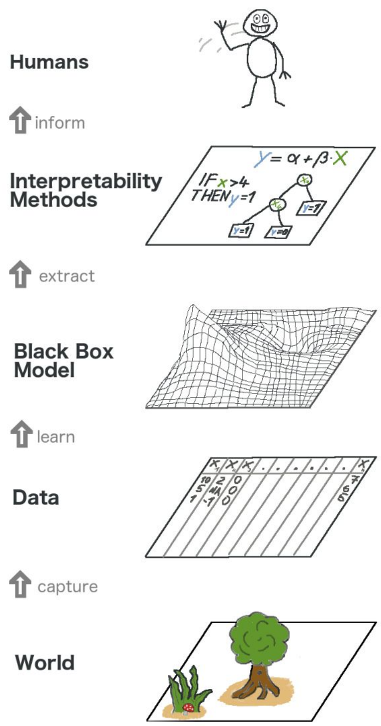
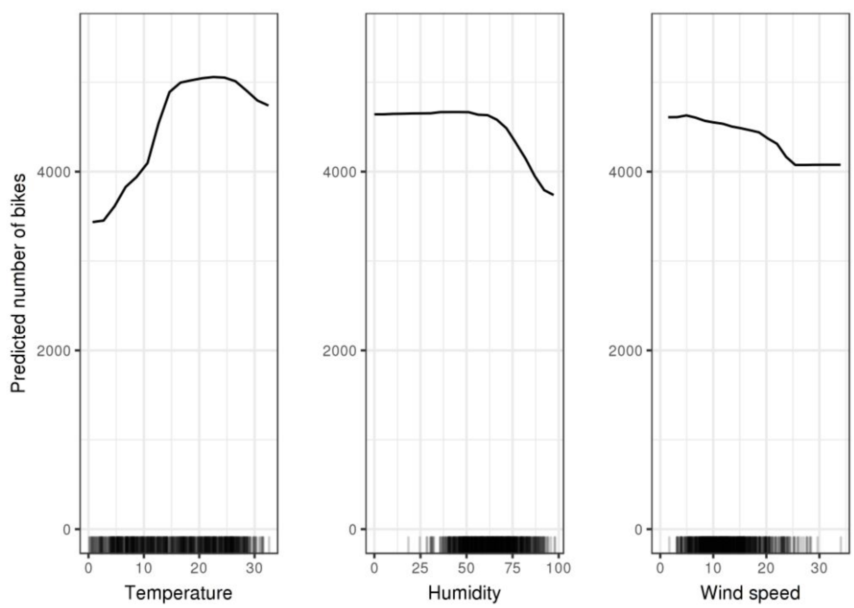
Notes taken from the Interpretable Machine Learning book by Christoph Molnar

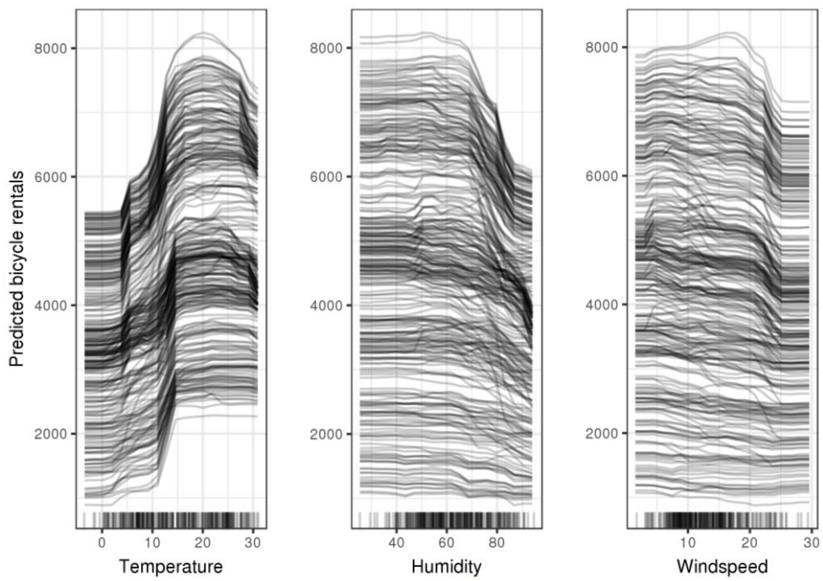
- Explainability - continually asking why (and how)?
  - Ex. Why stopped? 70% chance of child crossing the road. How did you calculate that? I took into account X,Y and Z, and combined them in this way. Why did you take these 3 features and not some other combination?
  - Explanations are contrastive - Why this and not that?
    - For a house price prediction, the house owner might be interested in why the predicted price was high compared to the lower price they had expected. If my loan application is rejected, I do not care to hear all the factors that generally speak for or against a rejection. I am interested in the factors in my application that would need to change to get the loan. I want to know the contrast between my application and the would-be-accepted version of my application.
  - The best explanation is the one that highlights the greatest difference between the object of interest and the reference object.
  - Explanations are selected - small list of causes (not all of them)
  - Explanations are social - know your audience
  - Explanations focus on the abnormal
    - If input features for a prediction was abnormal in any sense (eg rare category), and feature influenced the prediction, it should be included in an explanation
  - Explanations are truthful - should predict the event as truthfully as possible (called fidelity)
  - Explanations are general and probable - in contrast with them being abnormal above
- Properties of individual explanations
  - Accuracy - how well does explanation predict unseen data?
  - Fidelity - how well does explanation approximate prediction of black box model?
  - Consistency - how much does an explanation differ between models that have been trained on the same task and that produce similar predictions?
  - Stability - how similar are explanations for similar instances? (always desirable)
  - Comprehensibility
  - Certainty (confidence) that model has in individual predictions
  - Degree of importance - how well does explanation reflect importance of features or parts of the explanation
  - Novelty - is data instance an outlier? Then high novelty (and likely low certainty)
  - Representativeness - how many instances does an explanation cover?
- Model-agnostic methods



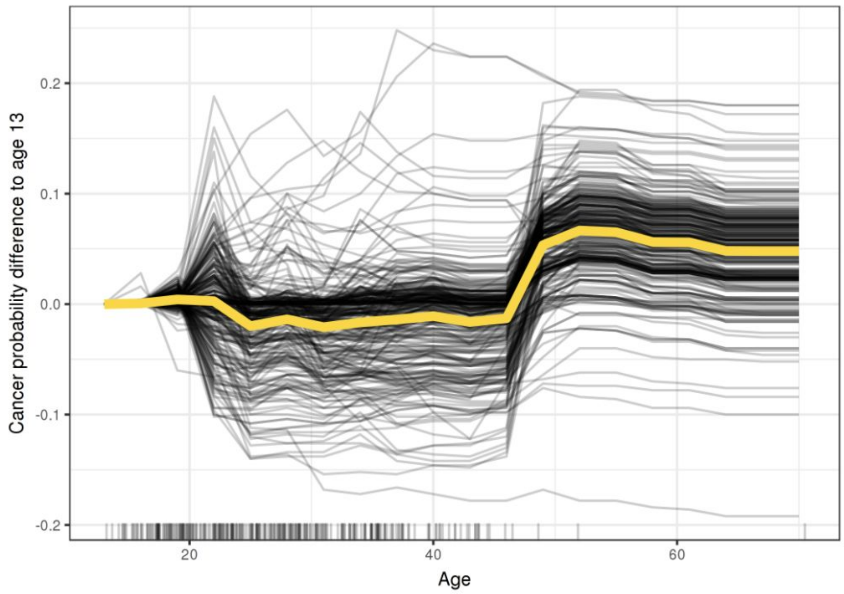
- Partial Dependence Plots



  - Advantages
    - Intuitive
    - Easy to implement
    - Has a causal interpretation (within the model, not necessarily the real world!)
  - Disadvantages

            - Low max number of features to represent
            - Need to show feature distribution on bottom
            - Assumption of independence (features could be correlated)
                - ALE plots help with this, work with conditional instead of marginal distribution
            - Heterogeneous effects might be hidden
                - Individual conditional expectation curves instead of aggregated line
    - Alternatives
        - ALE, ICE
- Individual Conditional Expectation (ICE)
    - Visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance. PDP is average of ICE plot
    - 

    - centered-ICE

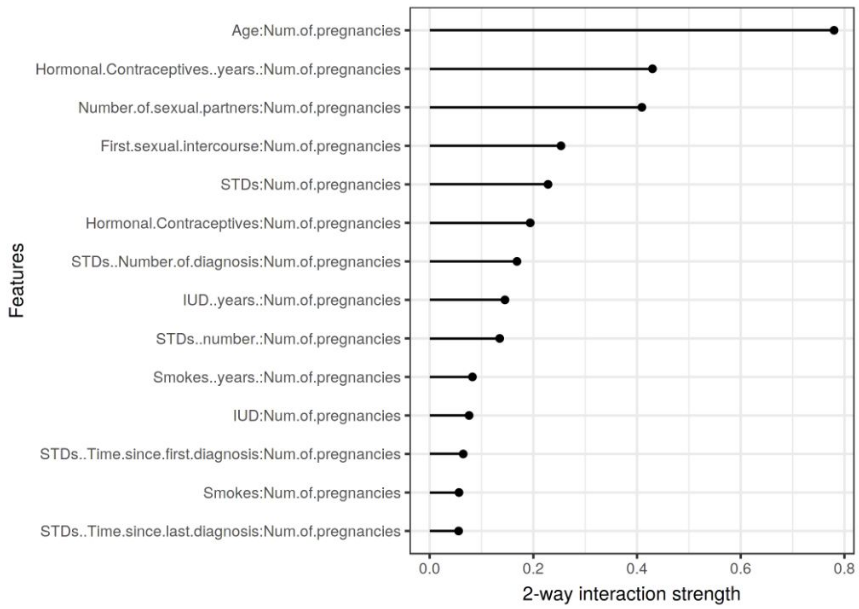
- 

- Advantages
  - Even more intuitive than PDP
  - Uncover heterogeneous relationships
- Disadvantages
  - Can only display 1 feature at a time
  - Feature correlation isn't dealt with meaningfully
  - Plot can become crowded/average not easy to see (easy to fix)

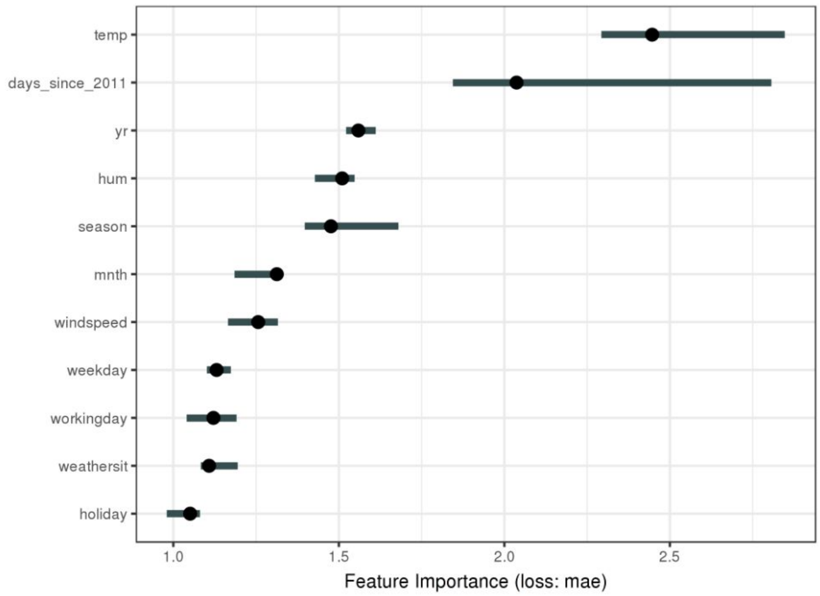
- Accumulated Local Effects (ALE)
  - Faster and unbiased alternative to PDPs
  - If features are correlated, the PDP cannot be trusted

$$\hat{\tilde{f}}_{j,ALE}(x)=\sum_{k=1}^{k_j(x)}\frac{1}{n_j(k)}\sum_{i:x_j^{(i)}\in N_j(k)}\left[f(z_{k,j},x^{(i)}_{\setminus j})-f(z_{k-1,j},x^{(i)}_{\setminus j})\right]$$

  - The value of the ALE can be interpreted as the main effect of the feature at a certain value compared to the average prediction of the data. For example, an ALE estimate of -2 at x_j=3 means that when the j-th feature has value 3, then the prediction is lower by 2 compared to the average prediction.
  - Advantages
    - Unbiased (work even when correlated)
    - Faster to compute than PDPs
    - Clear interpretation: conditional on a given value, the relative effect of changing the feature on the prediction can be read
    - In most situations, prefer ALE plots over PDPs
  - Disadvantages

        - Interpretation remains difficult when features are strongly correlated
        - ALE plots are not accompanied by ICE curves
- Feature Interaction
    - Want to know the share of variance that is explained by the interaction?
    - H-statistic!
    - 

    - Advantages
        - Has underlying theory
        - Always between 0 and 1, comparable across features and even models
        - Detects all kinds of interactions (even higher-order than 2)
    - Disadvantages
        - Computationally expensive
        - If sampling data, estimates have a certain variance and results can be unstable
        - Difficult to say when H-statistic is large enough to consider an interaction “strong”
        - If features are correlated, then integrate over feature combinations that are very unlikely in reality (same problem as with PDP)
- Feature Importance
    - Increase in the prediction error of the model after we permuted the feature’s values (randomize the values in that column)

- 

- Advantages
  - Nice interpretation
  - Highly compressed, global insight
  - FI is comparable across different problems (if use error ratio)
  - Automatically takes into account all interactions
  - No retraining
- Disadvantages
  - Unclear: use training or test data
  - Linked to error of model
  - Need access to the true outcome (need labeled data)
  - May be unstable
  - Correlated features are a problem, again (biased by unrealistic data points)

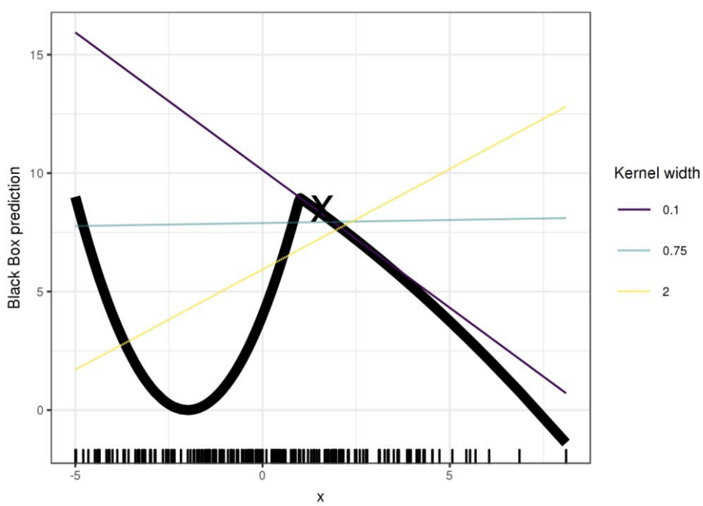
- Global Surrogate
  - Interpretable model that is trained to approximate the predictions of a black box model
  - Advantages
    - Flexible
    - Intuitive
    - Can easily measure how well surrogates are in approximating
  - Disadvantages
    - Draw conclusions about model, not data!
    - Surrogate model comes with advantages and disadvantages of that model

- Local Surrogate (LIME)

  The recipe for training local surrogate models:

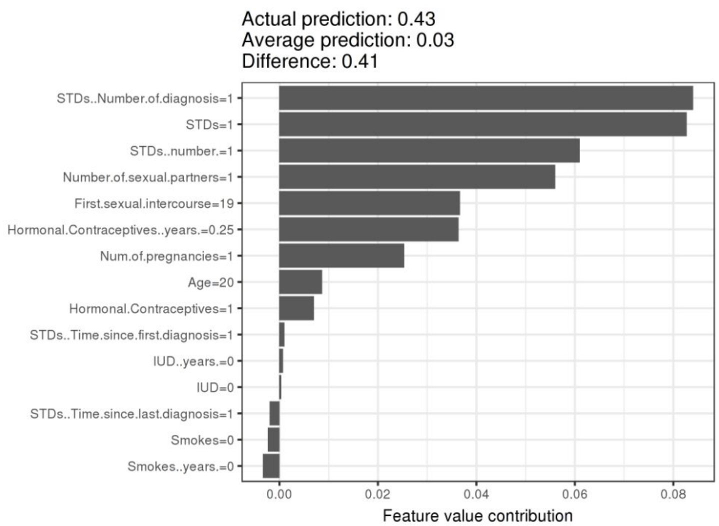
    - Select your instance of interest for which you want to have an explanation of its black box prediction.
    - Perturb your dataset and get the black box predictions for these new points.
    - Weight the new samples according to their proximity to the instance of interest.
    - Train a weighted, interpretable model on the dataset with the variations.
    - Explain the prediction by interpreting the local model.

  - Kernel width (neighborhood size) can make a large difference in interpretability



  -
  - Advantages
    - Model-agnostic
    - Explanations are short (selective) and possibly contrastive
    - Fidelity measure gives idea of reliability of interpretable model
    - Can use other features than original model
  - Disadvantages
    - Unclear neighborhood size
    - Better sampling
    - Complexity of explanation model is pre-defined

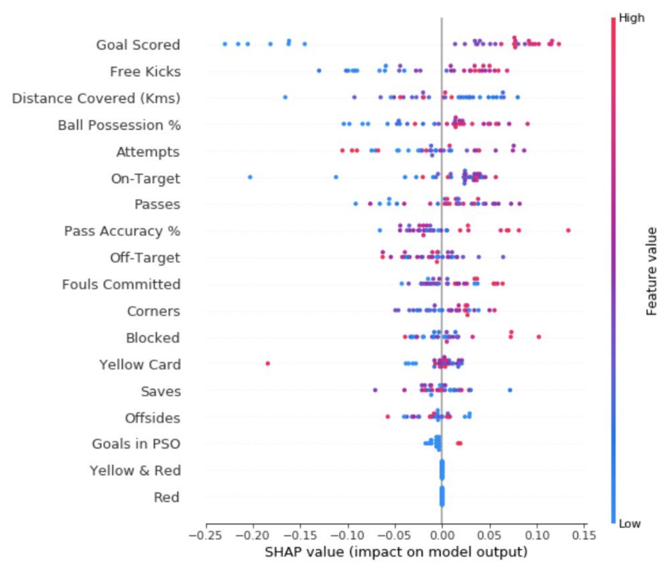
- Instability of explanations

- Shapley Values
  - The Shapley value is the average marginal contribution of a feature value across all possible coalitions
  - The Shapley value is NOT the difference in prediction when we would remove the feature from the model.
  - An intuitive way to understand the Shapley value is the following illustration: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.
  - 

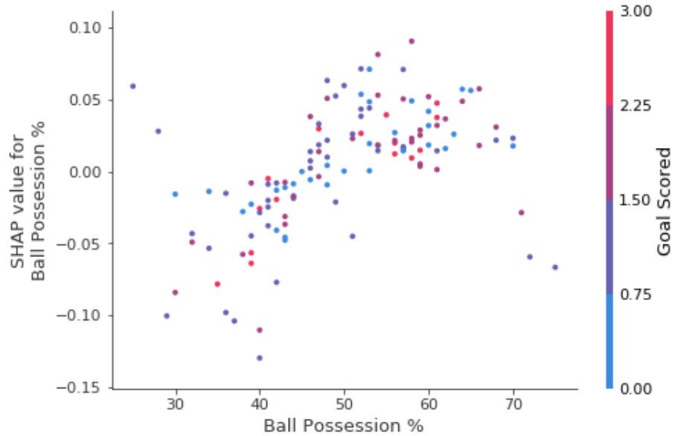
  - Advantages
    - Difference between prediction and average prediction is fairly distributed among feature values of the instance
    - Allows contrastive explanations
    - Solid theory
  - Disadvantages
    - Lots of computing time
    - Can be misinterpreted
    - Not parsimonious
    - No prediction model (like LIME)
    - Need access to data

- Inclusion of unrealistic data instances

- Kaggle
  - Feature importances (how much a feature affects)
    - Permutation - use **eli5**
  - Partial dependence plots (how a feature affects)
    - PDPBox
  - SHAP values
    - SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.
    - Shap library



    -


SHAP value for
Ball Possession %
0.10
0.05
0.00
−0.05
−0.10
−0.15
30
40
50
60
70
Ball Possession %
3.00
2.25
1.50
0.75
0.00
Goal Scored