Building a Robot Judge

16. Explanation and Interpretation

Elliott Ash

https://xkcd.com/

Characteristics of explanation methods

- Intrinsic or post hoc?
  - Intrinsic: interpretable due to simple structure, such as short decision trees or sparse linear models.
  - Post hoc: applied after model training (e.g. permutation feature importance)
- Model-specific or model-agnostic?
  - model-specific: tools limited to particular model classes (e.g. looking at parameters in linear regressions)
  - model-agnostic: applies to any model (e.g. analyzing input and out pairs)
- local or global?
  - does it explain an individual prediction, or try to explain an entire model?

How explanations are reported

- Feature summary statistic:
  - e.g. coefficient estimate, feature importance, or pairwise feature interaction strengths.
- Feature summary visualization:
  - e.g. coefficient plots, feature importance plots
  - partial dependence plots show a feature and the average predicted outcome.
- Model internals (e.g. learned weights):
  - coefficients in linear models
  - decision tree structure (features/thresholds for the splits) of decision trees.
  - visualization of filters in convolutional neural networks.
- Data points:
  - counterfactual explanations: find a similar data point by changing features which changes prediction
  - identification of prototypes of predicted classes.
  - works well for images and texts
- Intrinsically interpretable model:
  - approximate black box model (globally or locally) with a linear model or decision tree.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.
- **Social**: explanations should be targeted to the relevant audience.
- **Outlier-focused**: if one of the input features is abnormal, that should be the focus of the explanation.

Interpretable Models

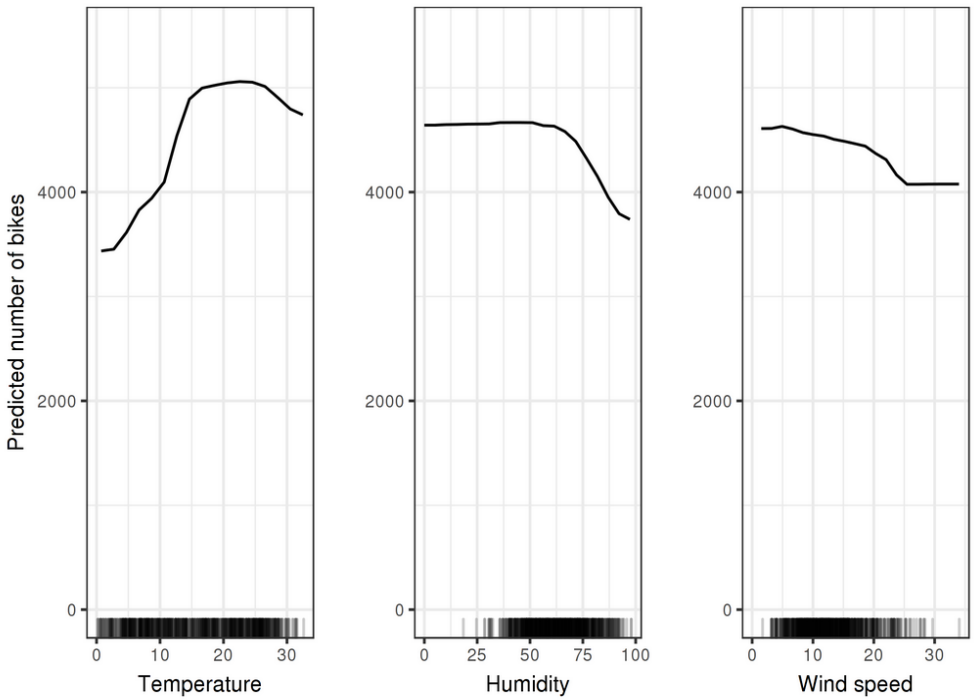
Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

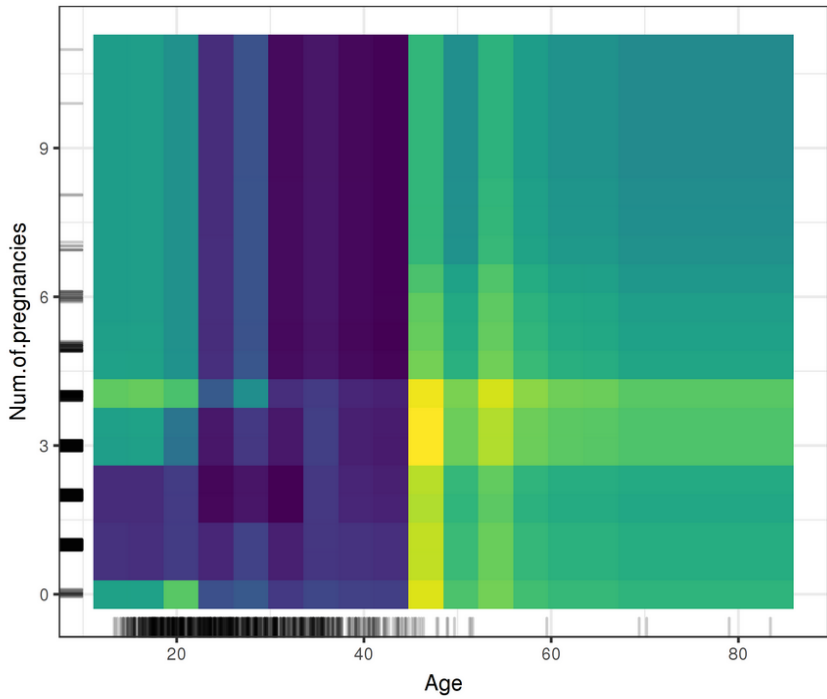
- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.
- **Monotonicity**: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature.
- **No interactions:** allowing interactions between features improves predictive performance but hurts interpretability.

Counterfactual Explanations

- Find the "closest" data point to the current one that changes the prediction.
- In text, could compute similarity (e.g. cosine similarity of document embeddings) between all pairs, and find closest document with a different prediction.

Partial Dependence Plots





Centered Individual Conditional Expectation

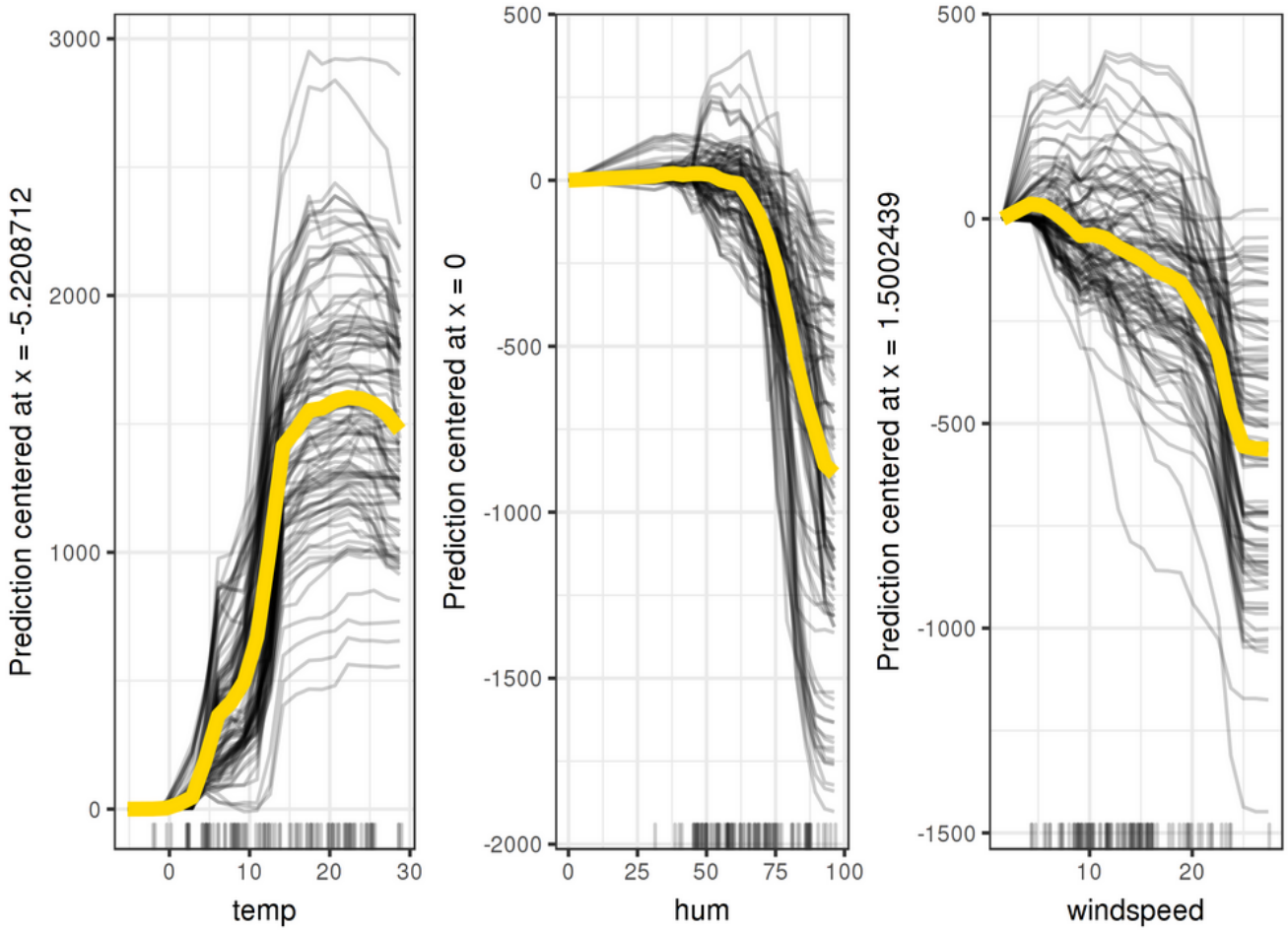


FIGURE 5.9: Centered ICE plots of predicted number of bikes by weather condition. The lines show the difference in prediction compared to the prediction with the respective feature value at its observed minimum.

Feature Importance

- What features are most important for prediction?

Permutation feature importance algorithm (Fisher, Rudin, and Dominici 2018):

- Estimate any model, compute mean squared error.
- For each feature j:
  - generate new dataset where feature j is permuted (scrambled)
  - generate predictions and estimate new error.
  - feature importance of j is (proportional or absolute) increase in error (ratio or difference).

Feature Importance Plot

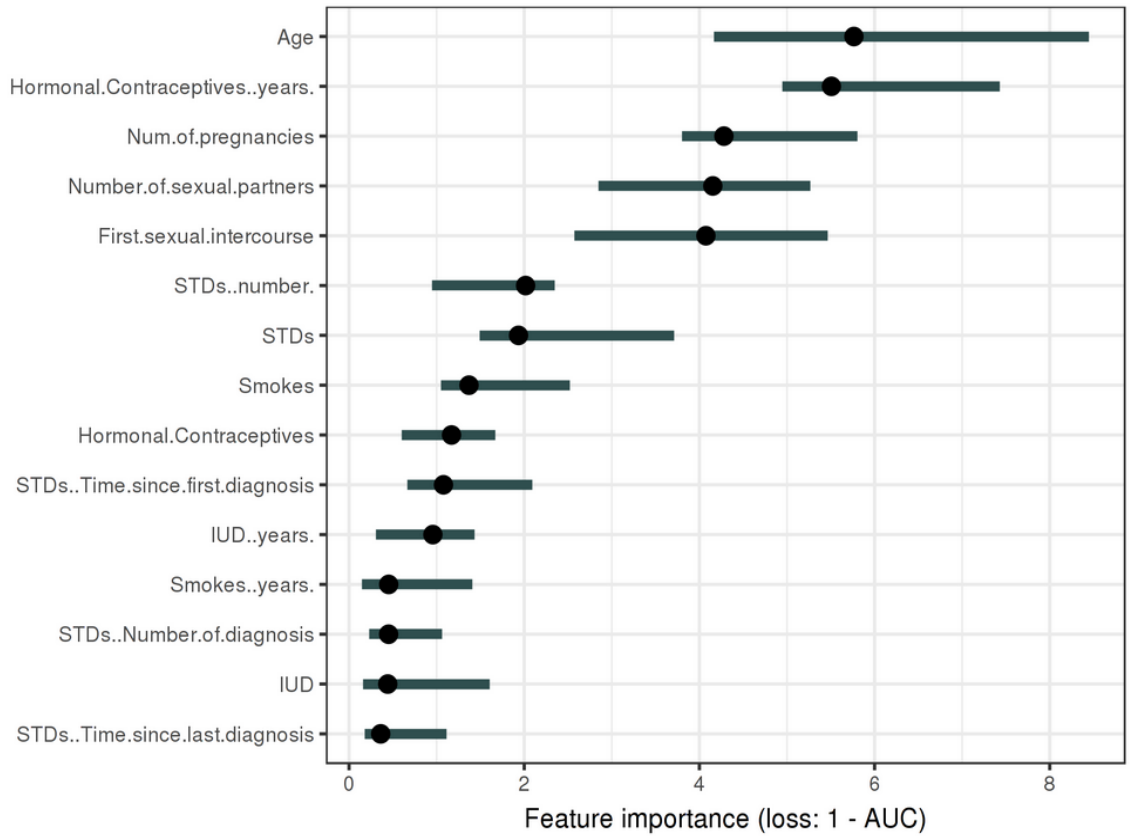


FIGURE 5.29: The importance of each of the features for predicting cervical cancer with a random forest. The most important feature was Age. Permuting Age resulted in an increase in 1-AUC by a factor of 5.76

Global Surrogate Model

- Approximate a black box model with an interpretable model.

1. Get predictions $\hat{y}$ of the black box model from the data X.
2. Train an interpretable model (lasso, decision tree, etc) on X with $\hat{y}$ as the label.
   - This is the surrogate model!
3. Validate that the surrogate model replicates the predictions of the black box model
   - e.g., compute R^2 between black box $\hat{y}$ and surrogate $\hat{\hat{y}}$

Local Surrogate (LIME)

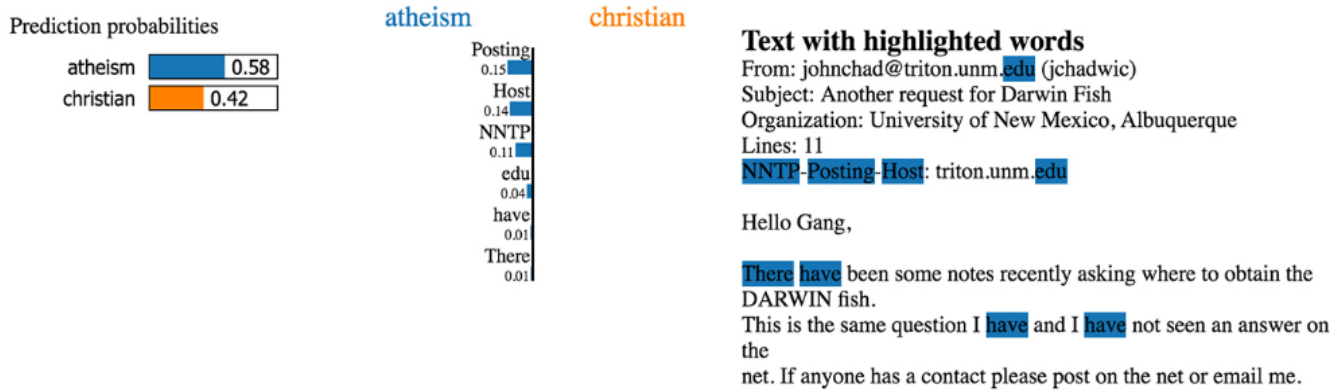
- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain
2. Perturb dataset (locally) and get black box predictions for the new points.
3. Train an interpretable surrogate model on the perturbed dataset (weighted by proximity to initial data point).
   - This is the "local" surrogate model.
   - use lasso with high L1 penalty to get a sparse explanation.

LIME for Text

1. Generate new texts by randomly *removing* words from the original document.
2. Form predictions $\hat{y}$ from black box model for these perturbed documents.
3. Train lasso on dataset of binary features for each word, equaling one if word appears, to predict $\hat{y}$.
   - weight by proximity to initial data point (one minus the proportion of words dropped)

```
exp = explainer.explain_instance(test_example,
classifier.predict_proba, num_features=6)
```



Other explanation methods

- This is an active area of research/development.
- Shapley values (Lundberg and Lee 2017) assign importance to features by relative contribution to prediction (based on solution concept in game theory)
  - (sometimes) better than LIME because adds value for interactions
  - but current implementations are slow.
- Model Understanding through Subspace Explanations (MUSE; Lakkaraju, Kamar, Caruana, and Leskovec 2019)
  - shows how to provide conditional global explanations
  - how the model works in different subsets of data
  - no code online yet