Brown Bag: Interpretable Machine Learning Model Summaries

Data Science Team

???Date???

Andrew Wheeler, PhD

andrew.wheeler@hms.com

Overview of the Problem

- Machine learning models are very difficult to understand how the inputs produce the prediction

- Can be difficult to explain to others how the model works

- Some actions need not only prediction, but *why* a case is predicted high risk.

4 Types of Interpretable Summaries

- What variables are *important* for prediction
- When you change *x*, how does *y* change?
- What variables interact with one another?
- Why is a particular prediction high/low?

What variables are important for prediction

- Either leave variable out, or permutate feature and redo predictions
- Can either be absolute or relative
- Can use whatever “accuracy” metric you want.

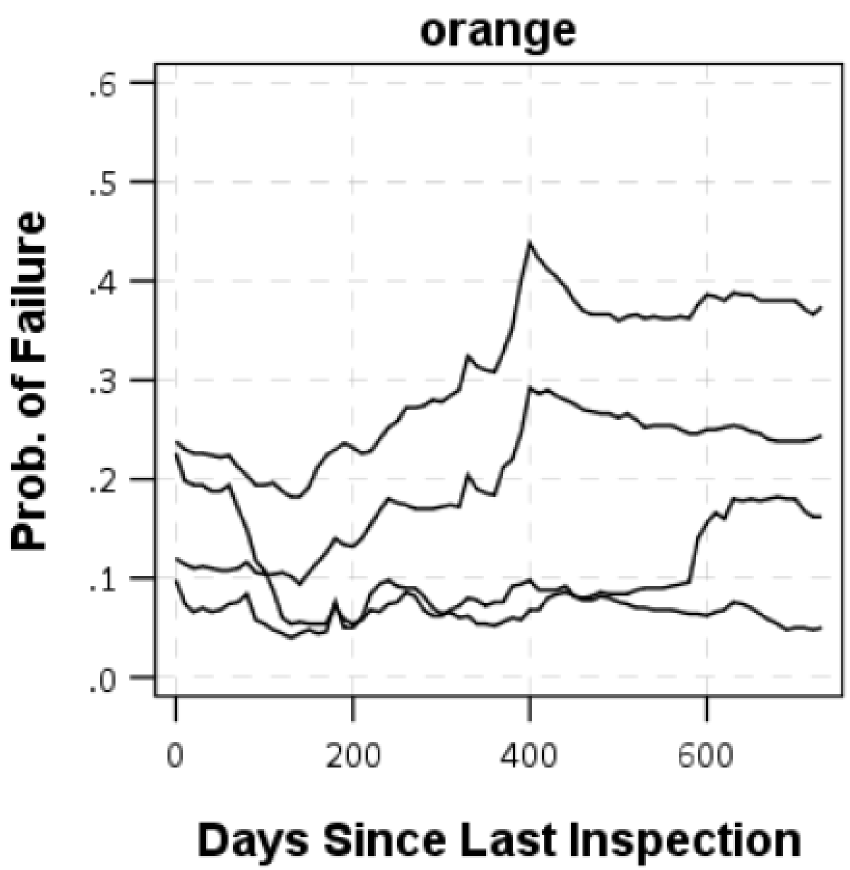
What variables are important for prediction

- Why do we care?
  - Can be used as a general way to evaluate model / EDA / sanity check

- To be aware of
  - Very volatile in my experience (slight change in model produces very different rankings) [more features correlated, bigger problem]
  - Documentation is very poor for different tools

When we change *x*, what happens to *y*?

- Simplest approach, calculate $\mathbb{E}[Y \mid X = x, Z]$ and put in a graph, varying only *x.*
- Example, predicting prob. of failed food inspection
  - Vary days since last inspection (x axis)
  - Different lines are for # of prior offenses
  - Orange is the rater
  - Hold constant several other factors



When we change *x*, what happens to *y*?

- Requires we pick arbitrary inputs to hold constant, may not be reasonable

- Other alternatives combat this by averaging those lines over all other observed samples -> partial dependence plots

When we change *x*, what happens to *y*?

- Why do we care?
  - We may actually want to change x to produce a particular outcome
    - E.g. missing data on age produces a high probability of soft-denial. Suggests spend more time getting that age data to begin with

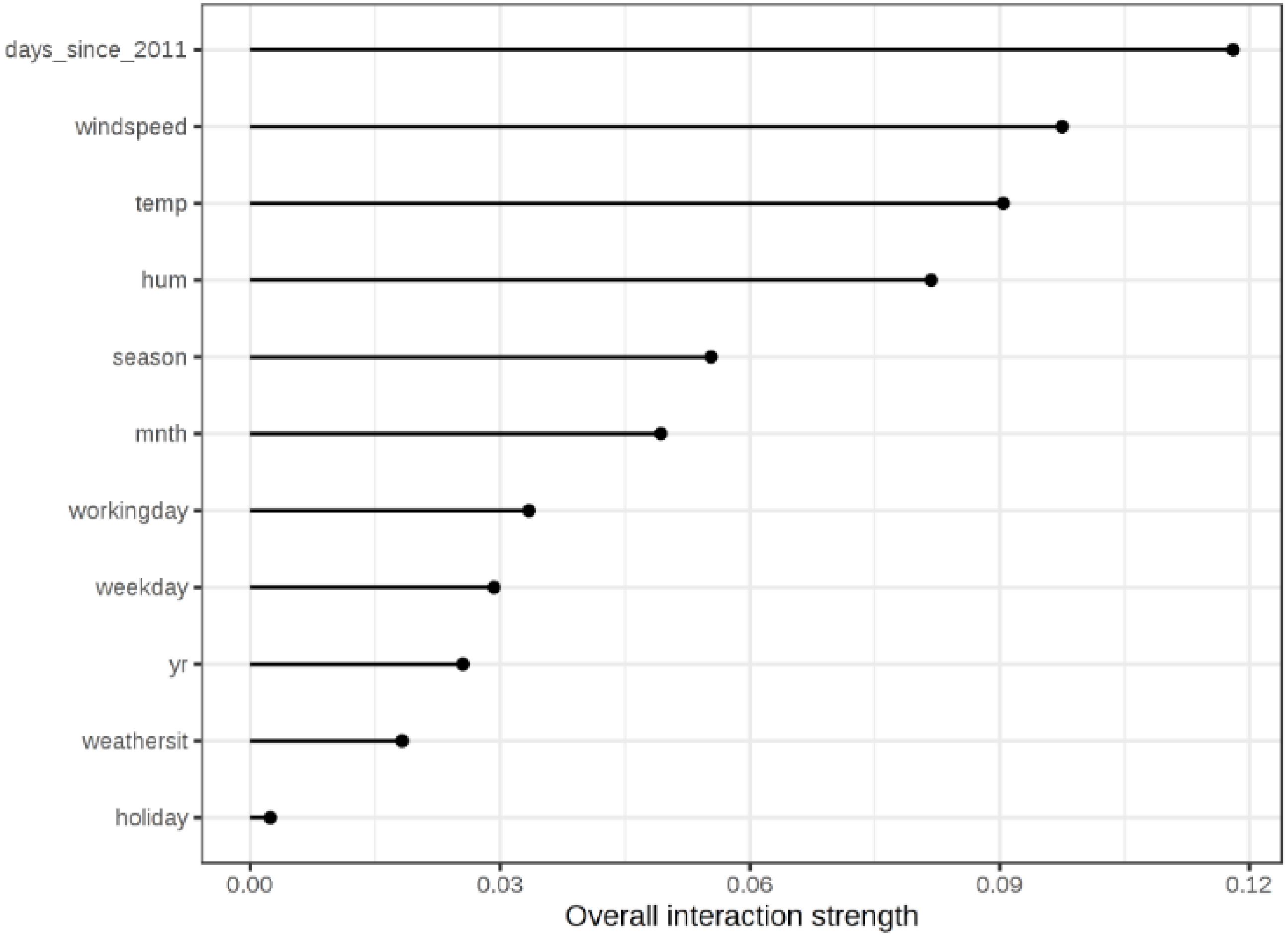
- To be aware of
  - Is not guaranteed to be a causal relationship, may be spurious with another factor

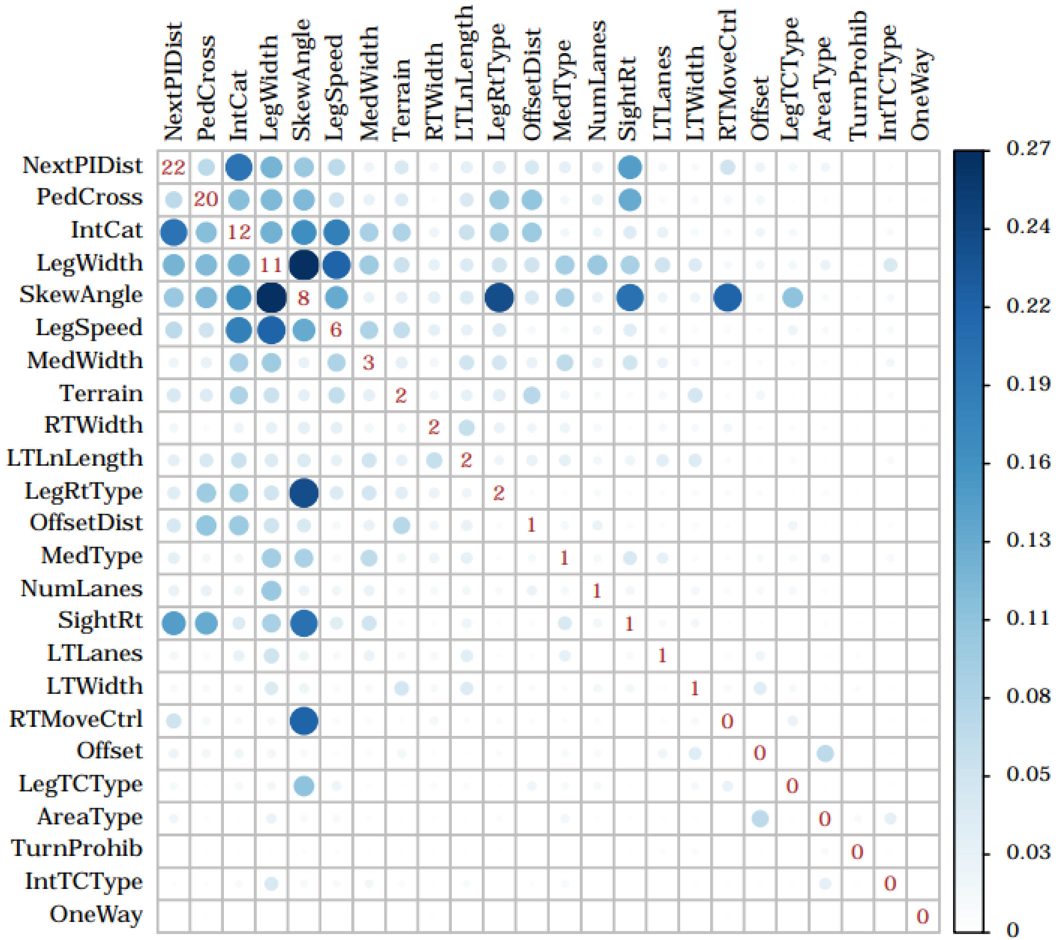
What variables interact with one another?

- Friedman's H statistic
  - Partitions variance from partial dependence between two variables, or between one variable and every other variable
  - Sort of like $\frac{\mathbb{V}(A+B) - \mathbb{V}(A) - \mathbb{V}(B)}{\mathbb{V}(A+B)}$, how variance of prediction changes when changing just A, just B, or both A & B at the same time
  - On a scale of 0-1, so a value of 0.2 would mean 20% of variance is due to interaction

- Why do we care?
  - High H values signify other reduced form summary metrics may not be accurate

What variables interact with one another?





Why did we get this particular prediction?

- Local interpretation for a specific case.

- Reduced form summary
  - LIME – simulate data, estimate a regularized regression and pick top N variables
  - Shapely values – simulate data, and see how much X changes on average when other variables change
  - Decision Tree – simulate data, and estimate 1 decision tree

Why did we get this particular prediction?

- Why do we care?
  - Human in the loop needs to act on that information.
    - E.g. predicted high probability of soft-denial, why? Missing fields for x and large claim

- To be aware of
  - Model can be highly non-linear & have interactions, so reduced form is inaccurate
  - Correlated features can often swap out for one another

Applying in the Future

- Data Robot
  - Provides feature importance
  - Can do ourselves the ‘when you change x’ partial dependence type summaries
  - Very difficult to do ‘why particular predictions’ & interaction statistics, code ourselves and API intensive

- Why particular predictions can be data intensive, so can’t just run them overnight for *all* cases

- May want to stick with an interpretable model to begin with if these interpretable summaries are very important (and black box is not much of an improvement)
  - E.g. linear regression, association-rules, naïve Bayes, k-nearest-neighbors

Links to Resources

- Jupyter notebook giving example use, ??????
- Molnar’s online book, https://christophm.github.io/interpretable-ml-book/
- Shorter article with python references, https://towardsdatascience.com/an-overview-of-model-explainability-in-modern-machine-learning-fc0f22c8c29a

Questions?

Brown Bag: Interpretable Machine Learning Model Summaries

Data Science Team

???Date???

Andrew Wheeler, PhD

andrew.wheeler@hms.com