ML Interpretation

What Types of Insights Are Possible

* What features in the data did the model think are most important?
* For any single prediction from a model, how did each feature in the data affect that particular prediction?
* How does each feature affect the model's predictions in a big-picture sense (what is its typical effect when considered over a large number of possible predictions)?

Why Are These Insights Valuable

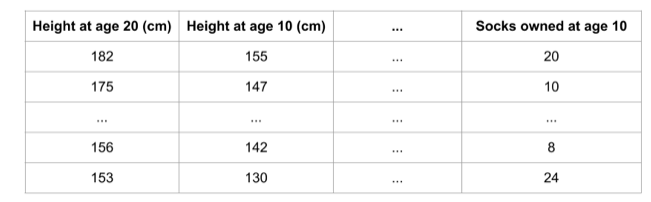
These insights have many uses, including

* Debugging
* Informing feature engineering
* Directing future data collection
* Informing human decision-making
* Building Trust

One of the most basic questions we might ask of a model is *What features have the biggest impact on predictions?* This concept is called *feature importance*

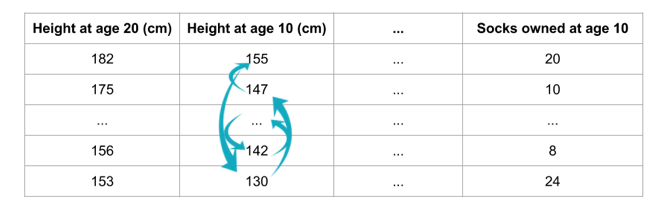
**Permutation importance**

Consider data with the following format:



We want to predict a person's height when they become 20 years old, using data that is available at age 10.

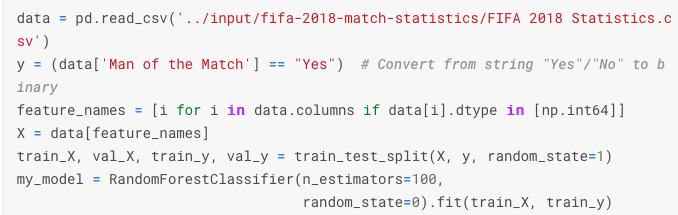
**Permutation importance is calculated after a model has been fitted.** So we won't change the model or change what predictions we'd get for a given value of height, sock-count, etc.

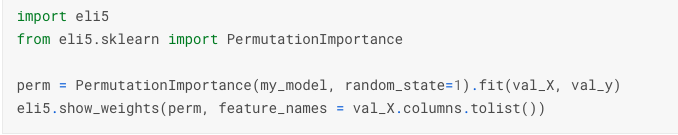


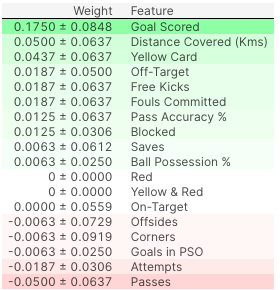
Randomly re-ordering a single column should cause less accurate predictions,

Code Example

Our example will use a model that predicts whether a soccer/football team will have the "Man of the Game" winner based on the team's statistics.







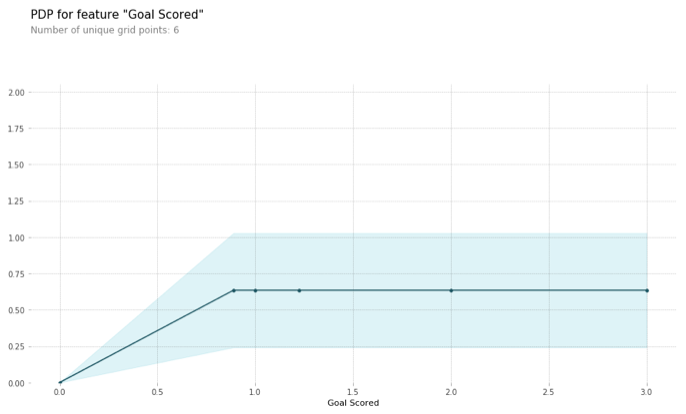
# Partial Dependence Plots

While feature importance shows **what** variables most affect predictions, partial dependence plots show **how** a feature affects predictions.

This is useful to answer questions like:

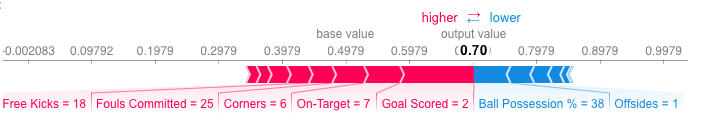
* Controlling for all other house features, what impact do longitude and latitude have on home prices? To restate this, how would similarly sized houses be priced in different areas?
* Are predicted health differences between two groups due to differences in their diets, or due to some other factor?

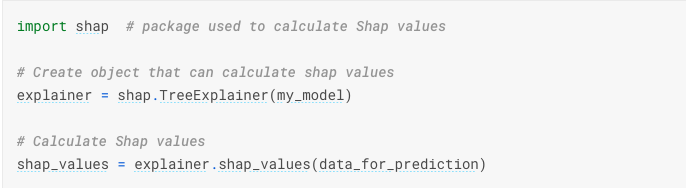
Like permutation importance, **partial dependence plots are calculated after a model has been fit.**The model is fit on real data that has not been artificially manipulated in any way.

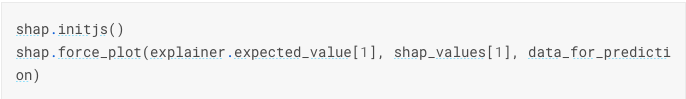


**SHAP Values (an acronym from SHapley Additive exPlanations)** break down a prediction to show the impact of each feature. Where could you use this?

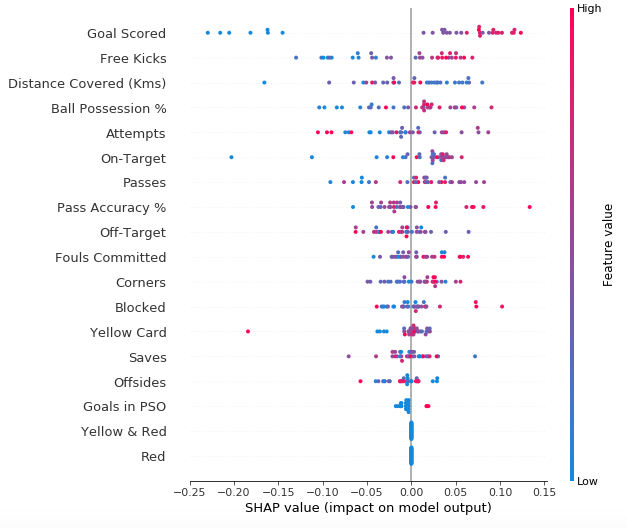
* A model says a bank shouldn't loan someone money, and the bank is legally required to explain the basis for each loan rejection
* A healthcare provider wants to identify what factors are driving each patient's risk of some disease so they can directly address those risk factors with targeted health interventions







**Advance use of Shap Values**



Some things you should be able to easily pick out:

* The model ignored the Red and Yellow & Red features.
* Usually Yellow Card doesn't affect the prediction, but there is an extreme case where a high value caused a much lower prediction.
* High values of Goal scored caused higher predictions, and low values caused low predictions

# SHAP Dependence Contribution Plots

# 

# 

The spread suggests that other features must interact with Ball Possession %.

# 

# Resource: [*https://www.kaggle.com/learn/machine-learning-explainability*](https://www.kaggle.com/learn/machine-learning-explainability)

By Dan Becker