



# SAS<sup>®</sup> GLOBAL FORUM 2018

---

## USERS PROGRAM

April 8 – 11 | Denver, CO  
Colorado Convention Center

#SASGF



# Interpreting Black-Box Machine Learning Models Using Partial Dependence and ICE Plots

#SASGF

Ray Wright, SAS Institute

Ray is Principal Machine Learning Developer at SAS.

# Interpreting Black-Box Machine Learning Models Using Partial Dependence and ICE Plots

# Topics

1. What a “black box” model is
2. Methods for understanding a model’s predictions
  - Visual
  - Model Agnostic
  - Post Hoc
3. Limitations of these methods and ways to scale up for big data

# Black Box Models

- Provide highly accurate predictions
- Are capable of discovering complicated interactions and nonlinearities.
- Are opaque due to their large numbers of parameters and many layers

*Examples: neural network, gradient boosting, and random forest.*

# Partial Dependence Plot

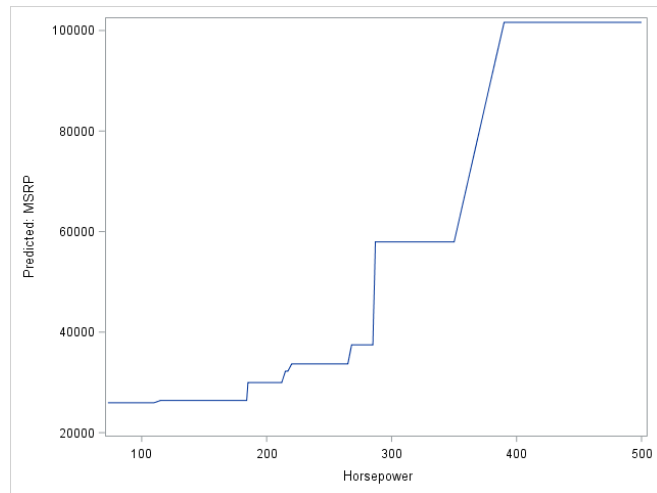
## What is it?

- A graph that depicts the functional relationship between a small number of model inputs and a model's predictions.
- Show how the model's predictions *partially depend* on values of the input variables of interest.

# Partial Dependence Plot

## One-Way Plots

- Plot the average model prediction for each value of a single model input.
- Valid if the plot variable does not interact strongly with other model inputs.

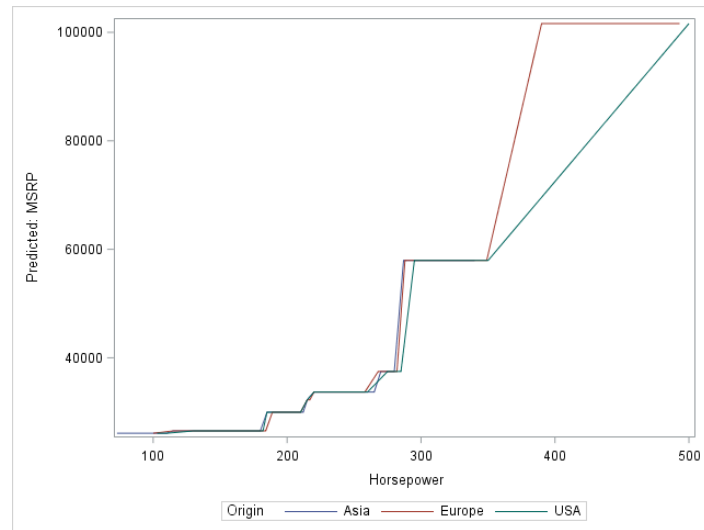


Overall relationship between horsepower and predicted MSRP for automobile models.

# Partial Dependence Plot

## Two-Way Plots

- In actual practice, interactions are common.
- You can use higher-order PD plots to check for specific interactions.



Horsepower by Origin by  
Predicted MSRP

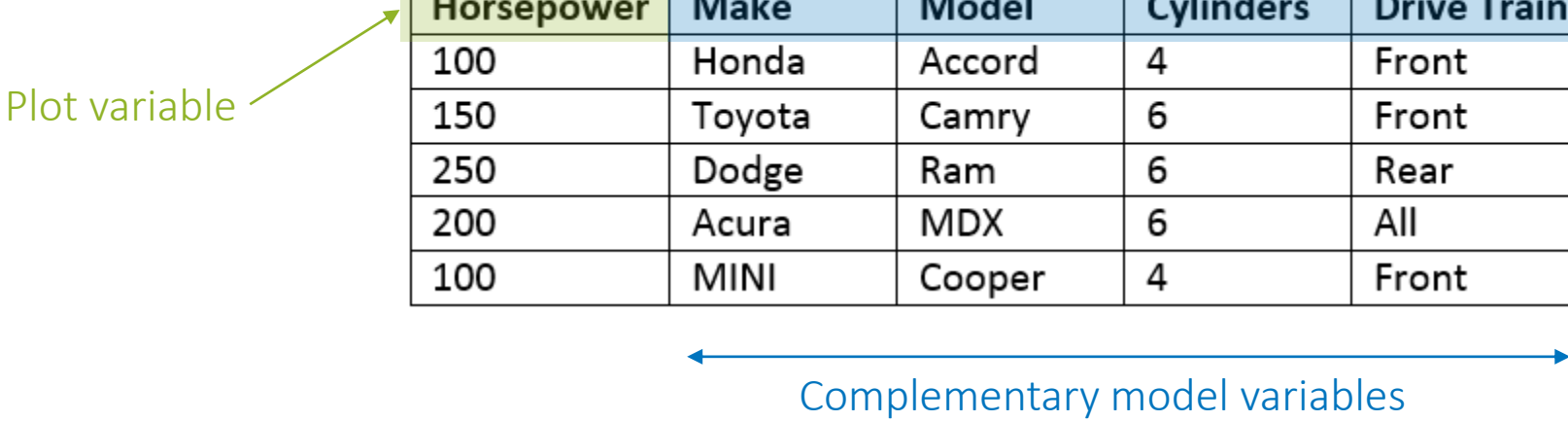


# Computing the Partial Dependence Function

## Step 1

Divide model inputs into two sets

Training Data (Hypothetical)



Horsepower	Make	Model	Cylinders	Drive Train
100	Honda	Accord	4	Front
150	Toyota	Camry	6	Front
250	Dodge	Ram	6	Rear
200	Acura	MDX	6	All
100	MINI	Cooper	4	Front

Plot variable

Complementary model variables

# Computing the Partial Dependence Function

## Step 2

Find the unique values of the plot variable

Horsepower
100
150
...
...
...

← First two  
values of  
horsepower

# Computing the Partial Dependence Function

## Step 3

Create one replicate of the training set for each value of the plot variable

Replicate #1	Horsepower	Make	Model	Cylinders	Drive Train
	100	Honda	Accord	4	Front
	100	Toyota	Camry	6	Front
	100	Dodge	Ram	6	Rear
	100	Acura	MDX	6	All
Replicate #2	100	MINI	Cooper	4	Front
	150	Honda	Accord	4	Front
	150	Toyota	Camry	6	Front
	150	Dodge	Ram	6	Rear
	150	Acura	MDX	6	All
	150	MINI	Cooper	4	Front



Cartesian  
Product

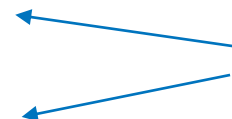
# Computing the Partial Dependence Function

## Step 4

Score the replicates

Horsepower	Make	Model	Cylinders	Drive Train	Predicted MSRP
100	Honda	Accord	4	Front	\$13000
100	Toyota	Camry	6	Front	\$15000
100	Dodge	Ram	6	Rear	\$15000
100	Acura	MDX	6	All	\$18000
100	MINI	Cooper	4	Front	\$16000
150	Honda	Accord	4	Front	\$15000
150	Toyota	Camry	6	Front	\$17000
150	Dodge	Ram	6	Rear	\$17000
150	Acura	MDX	6	All	\$20000
150	MINI	Cooper	4	Front	\$18000

Model  
scores



# Computing the Partial Dependence Function

## Step 5

Compute the average predicted value within each replicate

Horsepower	Average Predicted MSRP
100	\$15400
150	\$17400
...	
...	
...	



First two values  
of PD function

# Plots for Big Data

- Problem

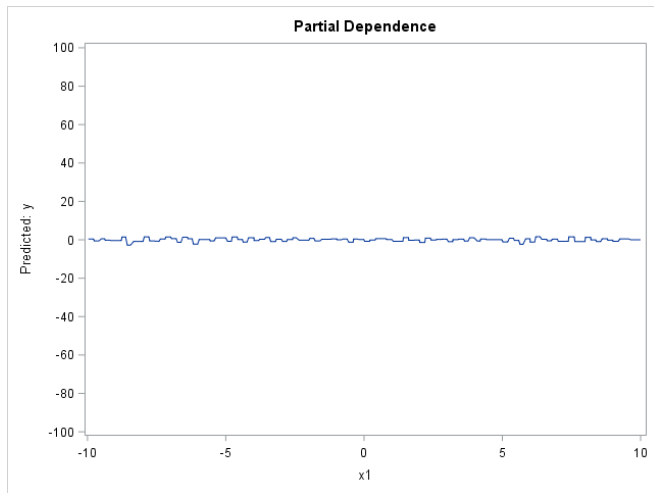
- As the number of unique values and observations increase, the number of replicated observations can grow out of hand.

- Solutions

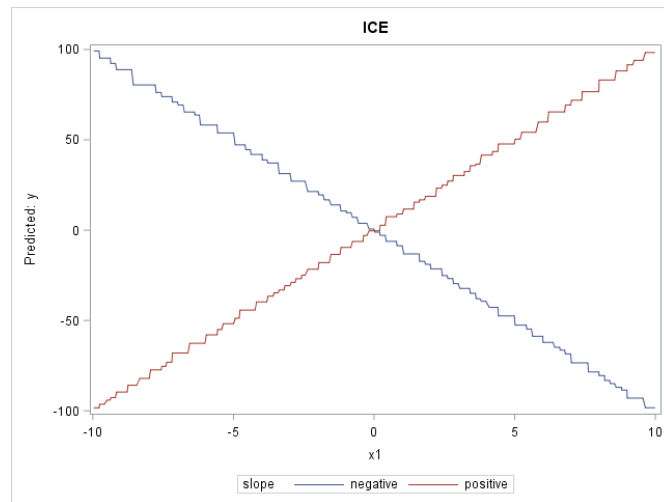
- Bin unique values of high-cardinality inputs such as income.
- Sample or cluster observations.
- Process the replicates one (or a few) at a time, keeping only the average predicted value for each replicate.

# Individual Conditional Expectation

## Toy Example



Partial Dependence



ICE for two individuals

# Individual Conditional Expectation

- Whereas PD plots provide a coarse view of a model's workings, ICE plots enable you to drill down to the level of individual observations.
- ICE plots disaggregate the PD function to reveal interactions and interesting subgroups.



# Computing the ICE Function

Unique Values of Plot Variable

Horsepower
100
150
200
250
300
...

Complementary Variables for One Observation

Make	Model	Cylinders	Drive Train
Honda	Accord	4	Front



Replicates

Horsepower	Make	Model	Cylinders	Drive Train
100	Honda	Accord	4	Front
150	Honda	Accord	4	Front
200	Honda	Accord	4	Front
250	Honda	Accord	4	Front
300	Honda	Accord	4	Front
...	Honda	Accord	4	Front



Scored Replicates

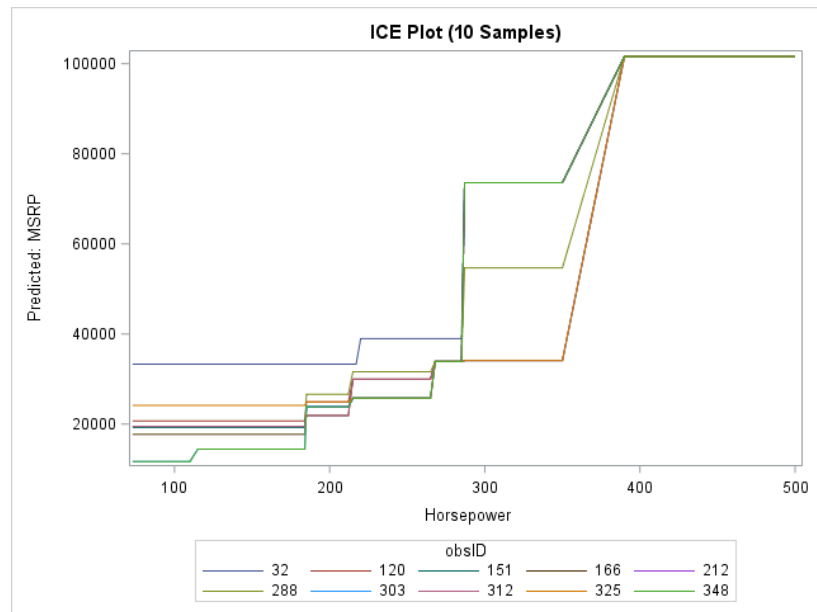
Horsepower	Make	Model	Cylinders	Drive Train	Predicted MSRP
100	Honda	Accord	4	Front	\$12,000
150	Honda	Accord	4	Front	\$15,000
200	Honda	Accord	4	Front	\$20,000
250	Honda	Accord	4	Front	\$26,000
300	Honda	Accord	4	Front	\$35,000
...	Honda	Accord	4	Front	....

Individual Car Model  
Varied over Horsepower

# Individual Conditional Expectation

## Managing Visual Overload

- Traditional ICE plots display one curve for each individual in the training set.
- You can manage the number of curves by sampling individuals or clustering the curves.



Sampled ICE curves

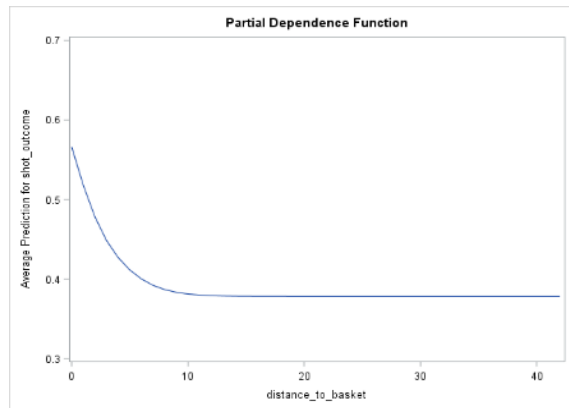
# Example: Predicting NBA Shot Success

## Model Variables

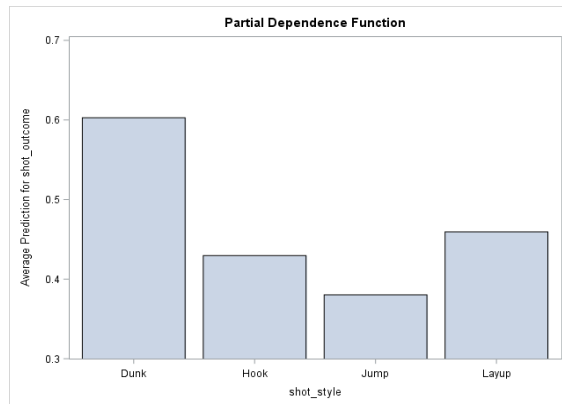
Variable	Role	Measurement Level	Values
Shot outcome	Target	Binary	0=made,1=missed
Distance to basket	Input	Interval	In feet
Player experience	Input	Interval	In years
Player height	Input	Interval	In inches
Player weight	Input	Interval	In pounds
Player position	Input	Nominal	Center, guard, or forward
Shot style	Input	Nominal	Jump, layup, hook, or dunk
Shot location	Input	Nominal	Right, left, center, left center, or right center
Shot area	Input	Nominal	Mid-range, restricted area (RA), in the paint (non-RA), above the break 3, right corner 3, left corner 3

# Predicting NBA Shot Success

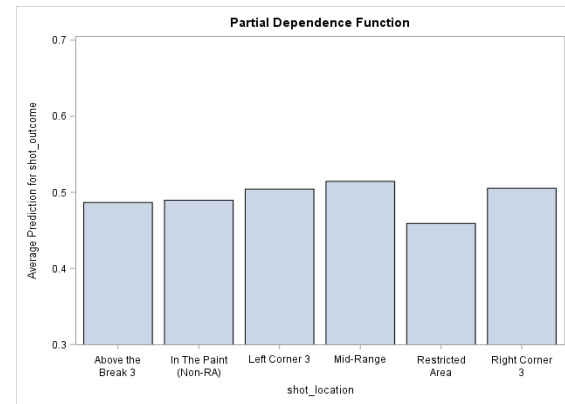
## PD Plots for Top Three Model Inputs



Distance to basket



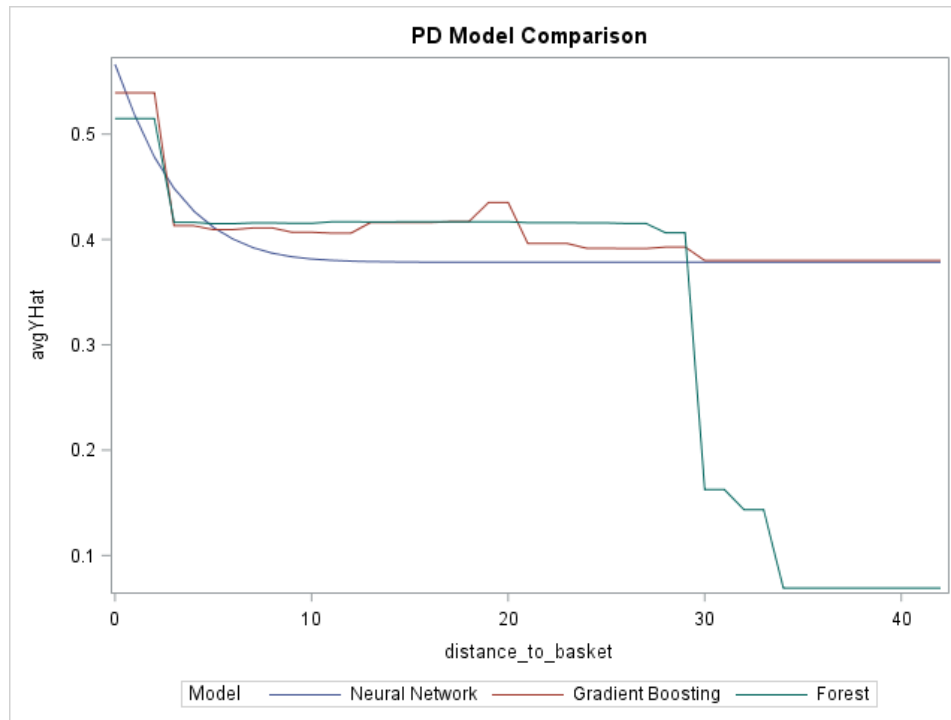
Shot style



Shot location

# Predicting NBA Shot Success

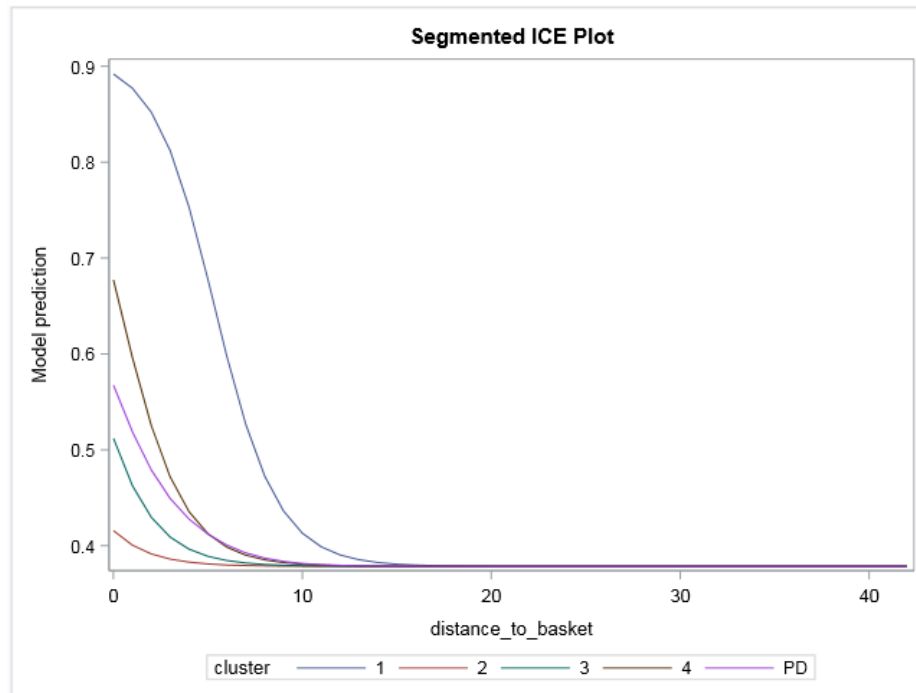
## Model Comparison



Partial Dependence Functions for Three Candidates

# Predicting NBA Shot Success

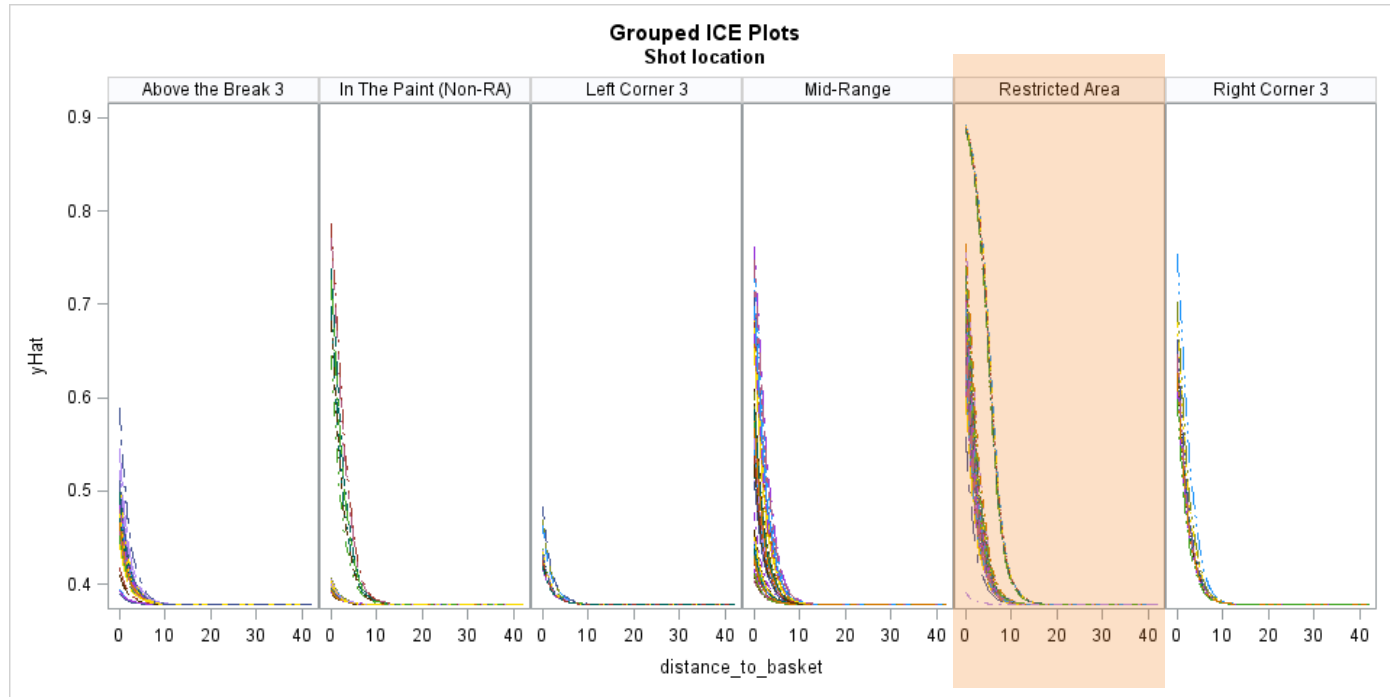
## Segmented ICE Plot



Centroids of clustered shot curves

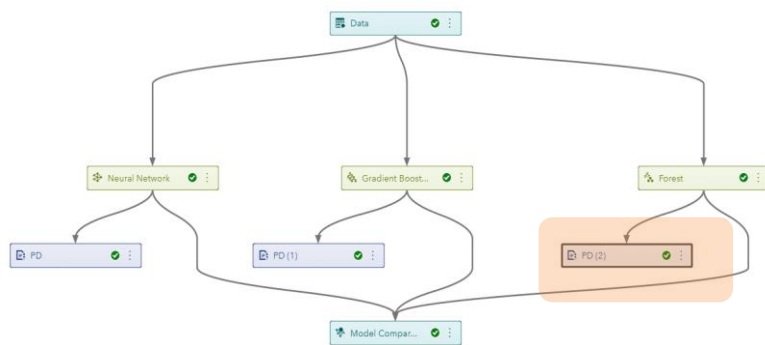
# Predicting NBA Shot Success

## Grouped ICE Plot

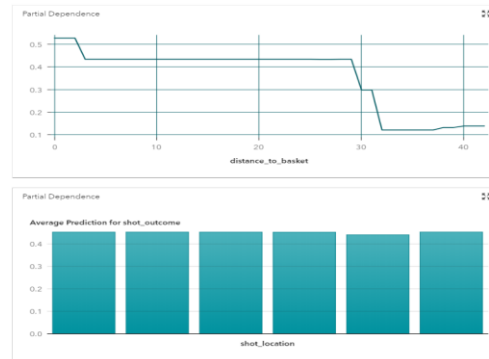


ICE curves grouped by shot location

# SAS Model Studio



Machine Learning Pipeline



Partial Dependence Plots



# Recap

- PD and ICE plots are visual, model-agnostic techniques that can help you interpret black box models.
- ICE plots let you drill down further to discover individual differences, interesting subgroups, and interactions among model variables.
- You may need to make adjustments for efficient computation.

# Closing Thoughts

- Both PD and ICE are *post hoc* methods and therefore approximations of the truth.
- To understand individual decisions, consider techniques like Locally Interpretable Model Agnostic Explanations (LIME).

# Want to Learn More?

- The SGF paper has code examples.
- Stop by the Data Mining and Machine Learning demo booth to chat.

# Your Feedback Counts!

Don't forget to complete the session survey  
in your conference mobile app.

1. Go to the Agenda icon in the conference app.
2. Find this session title and select it.
3. On the Sessions page, scroll down to Surveys and select the name of the survey.
4. Complete the survey and click Finish.

#SASGF

# SAS<sup>®</sup> GLOBAL FORUM 2018

April 8 - 11 | Denver, CO  
Colorado Convention Center