MLADS

MACHINE LEARNING, AI,
AND DATA SCIENCE CONFERENCE

June 2–4, 2020
June 9–11, 2020

Microsoft

Microsoft

How to interpret model prediction using interpretability tools on Azure Machine Learning?

Fatemeh Zamanian (Fatemeh.Zamanian@microsoft.com)
John Ehrlinger (John.Ehrlinger@microsoft.com)
Cheng Zhan (Zhan.Cheng@microsoft.com)

Session goals

- In this level 100 session, you will learn:
  - Why ML Interpretability is important,
  - An overview of SHAP, one of the most robust methods for ML Interpretability,
  - How to apply Azure Machine Learning Interpret tools for:
    - Regression use case – Tabular data,
    - Generic image classification use case,
  - Limitations of Azure Machine Learning Interpret tools.

Why ML Interpretability is important? What vs. Why?

- Modern ML techniques have enabled us to develop highly accurate models [the "What" question]
  - These generate very complex models which are often difficult to understand.
- For many high-stakes applications, understanding of these "black box" models are essential to gain the trust of stakeholders [the "Why" question]
  - This determines the success of the project, without which it is almost impossible to deploy these models into production.
- ML Interpretability addresses the trade-off between the "What" and the "Why"
  - Use complex model with high confidence.

What do we expect from a ML Interpretability tool?

From interpretably module, we will learn:

- Global point of view: From all the features in the model, which are most important.
- Local point of view: For any single prediction from a model, the marginal contribution of each feature in the data on that instance of prediction.

Why SHAP?

Here, we will learn about SHAP (**SH**apley **A**dditive **E**xplanation):

- Based on Shapley, from coalitional game theory,
- Mathematically very sound and robust,
- Provides both local and global explanation,
- Model agnostic,
- Can be applied to any type of data,
- Integrated within Azure Machine Learning Interpretability toolkit.

Intuition

For this data, this is “fit” from ML:

Black Box ML: $y = f(x_1, x_2)$

y	x_1	x_2
1	-1	1
14	1	4
1	2	-1
11	4	1
4	-1	2
-1	1	-1

Intuition

For this data, this is "fit" from ML:

Black Box ML: $y = f(x_1, x_2)$

How to break down the target between features?

y	x_1	x_2
1	-1	1
14	1	4
1	2	-1
11	4	1
4	-1	2
-1	1	-1

Intuition

For this data, this is "fit" from ML:

Black Box ML: $y = f(x_1, x_2)$

How to break down the target between features?

$$y = f(x_1, x_2) = 2x_1 + 3x_2$$

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

For this data, this is “fit” from ML:

Black Box ML: $y = f(x_1, x_2)$

How to break down the target between features?

$$y = f(x_1, x_2) = 2x_1 + 3x_2$$

- For instance i of the data (row i) the effect of $x_{1,i}$ on y_i is $2x_{1,i}$
- For instance i of the data (row i) the effect of $x_{2,i}$ on y_i is $3x_{2,i}$

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

- We know the "effect" of each feature for each instance of data.
- Which feature is "more" important?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

- We know the "effect" of each feature for each instance of data.
- Which feature is "more" important?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

- We know the "effect" of each feature for each instance of data.
- Which feature is "more" important?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

- We know the "effect" of each feature for each instance of data.
- Which feature is "more" important?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Local view: For some instances, x_1 has bigger impact and for others x_2 has bigger impact.

Intuition

- We know the “effect” of each feature for each instance of data.
- Which feature is “more” important?

<u>Local view:</u> For some instances, x_1 has bigger impact and for others x_2 has bigger impact.

How about **<u>global view</u>**?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

Intuition

- We know the "effect" of each feature for each instance of data.
- Which feature is "more" important?

Local view: For some instances, x_1 has bigger impact and for others x_2 has bigger impact.

How about **global view**?

y	x_1	x_2	I_{x_1}	I_{x_2}
1	-1	1	-2	3
14	1	4	2	12
1	2	-1	4	-3
11	4	1	8	3
4	-1	2	-2	6
-1	1	-1	2	-3

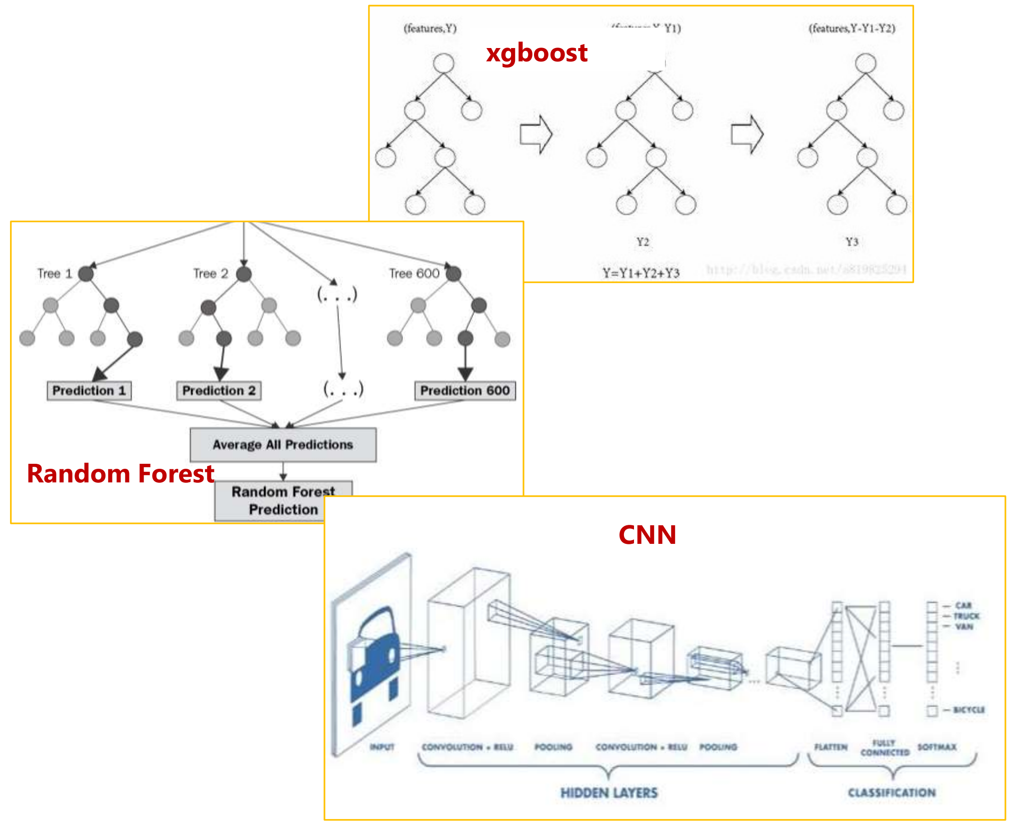
$$\mathrm{x}_{1,\mathrm{total}} = \frac{\sum \left|I_{x_{1,i}}\right|}{N} = \frac{20}{6} = 3.34$$

$$\mathrm{x}_{2,\mathrm{total}} = \frac{\sum \left|I_{x_{2,i}}\right|}{N} = \frac{30}{6} = 5$$

Globally, in this space, $-1 \leq x_1 \leq 4$ & $-1 \leq x_2 \leq 4$, x_2, comparing to x_1, is more important.

How about more complex models?

- How to answer these questions for a more complex ML model?
  - For any single prediction from a model, what is the **"fair"** contribution of each feature?
  - Which features in the model are more "important"?
- SHAP provides answers to both questions and "sheds light" on "black box" models.



Shapley Values

- The goal of SHAP is to explain the prediction of an instance, by computing the marginal contribution of each feature to the prediction.

- The SHAP explanation method computes Shapley values from coalitional game theory
  - The **feature** values of a data instance act as **players** in a coalition.
  - Shapley values tell us how to fairly distribute the "**payout**" (= the **prediction**) among all the features.

Target Space dim $N \times 1$

Feature Space dim $N \times M$

game	score	A	B	…	C
1	y_1	x_{11}	x_{12}	.	x_{1M}
2	y_2	x_{21}	x_{22}	.	x_{2M}
3	y_3	x_{31}	x_{32}	.	x_{3M}
4	y_4	x_{41}	x_{42}	.	x_{4M}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	y_N	x_{N1}	x_{N2}	.	x_{NM}

Coalitional Game Theory

- Suppose we have a game with 3 players, A, B, C and they have scored Y for a given game,
- Supposed we have trained a model, F, to predict target, Y,
- Let's see how to estimate "marginal" contribution of player (i.e. feature) A to score (i.e. target) Y for game i, based on coalition game theory, **this is the Shapley value of A for instance i**

game	score	A	B	C
i	Y	$\boldsymbol{A}$	B	C

Coalitional Game Theory

- Suppose we have a game with 3 players, A, B, C and they have scored Y for a given game,
- Supposed we have trained a model, F, to predict target, Y,
- Let's see how to estimate "marginal" contribution of player (i.e. feature) A to score (i.e. target) Y for game i, based on coalition game theory, **this is the Shapley value of A for instance i**

Coalition Exclude A	Target Estimation Exclude A
$\emptyset$	$\mathrm{F}(\emptyset)$
B	$\mathrm{F}(B)$
C	$\mathrm{F}(C)$
B, C	$\mathrm{F}(B, C)$

game	score	A	B	C
i	Y	$\boldsymbol{A}$	B	C

Coalitional Game Theory

- Suppose we have a game with 3 players, A, B, C and they have scored Y for a given game,
- Supposed we have trained a model, F, to predict target, Y,
- Let's see how to estimate "marginal" contribution of player (i.e. feature) A to score (i.e. target) Y for game i, based on coalition game theory, **this is the Shapley value of A for instance i**

game	score	A	B	C
i	Y	$\boldsymbol{A}$	B	C

Coalition Exclude A	Target Estimation Exclude A	Coalition Include A	Target Estimation Include A
$\emptyset$	$\mathrm{F}(\emptyset)$	A	$\mathrm{F}(A)$
B	$\mathrm{F}(B)$	A, B	$\mathrm{F}(A, B)$
C	$\mathrm{F}(C)$	A, C	$\mathrm{F}(A, C)$
B, C	$\mathrm{F}(B, C)$	A, B, C	$\mathrm{F}(A, B, C)$

Coalitional Game Theory

- Suppose we have a game with 3 players, A, B, C and they have scored Y for a given game,
- Supposed we have trained a model, F, to predict target, Y,
- Let's see how to estimate "marginal" contribution of player (i.e. feature) A to score (i.e. target) Y for game i, based on coalition game theory, **this is the Shapley value of A for instance i**

game	score	A	B	C
i	Y	$\boldsymbol{A}$	B	C

Coalition Exclude A	Target Estimation Exclude A	Coalition Include A	Target Estimation Include A	Difference in Target Estimation
$\emptyset$	$\mathrm{F}(\emptyset)$	A	$\mathrm{F}(A)$	$\mathrm{F}(A) - \mathrm{F}(\emptyset)$
B	$\mathrm{F}(B)$	A, B	$\mathrm{F}(A, B)$	$\mathrm{F}(A, B) - \mathrm{F}(B)$
C	$\mathrm{F}(C)$	A, C	$\mathrm{F}(A, C)$	$\mathrm{F}(A, C) - \mathrm{F}(C)$
B, C	$\mathrm{F}(B, C)$	A, B, C	$\mathrm{F}(A, B, C)$	$\mathrm{F}(A, B, C) - \mathrm{F}(B, C)$

Coalitional Game Theory

game	score	A	B	C
i	Y	A	B	C

Coalition Exclude A	Target Estimation Exclude A	Coalition Include A	Target Estimation Include A	Difference in Target Estimation	weight
$\emptyset$	$\mathrm{F}(\emptyset)$	A	$\mathrm{F}(A)$	$\mathrm{F}(A) - \mathrm{F}(\emptyset)$	1/3
B	$\mathrm{F}(B)$	A, B	$\mathrm{F}(A, B)$	$\mathrm{F}(A, B) - \mathrm{F}(B)$	1/6
C	$\mathrm{F}(C)$	A, C	$\mathrm{F}(A, C)$	$\mathrm{F}(A, C) - \mathrm{F}(C)$	1/6
B, C	$\mathrm{F}(B, C)$	A, B, C	$\mathrm{F}(A, B, C)$	$\mathrm{F}(A, B, C) - \mathrm{F}(B, C)$	1/3

Shapley value of A for instance i is the weighted average of marginal contributions for all the coalitions:

$$\frac{|S|!\,(N-|S|-1)!}{N!}$$

$$\frac{1}{3}\big(\mathrm{F}(A) - \mathrm{F}(\emptyset)\big) + \frac{1}{6}\big(\mathrm{F}(A, B) - \mathrm{F}(B)\big) + \frac{1}{6}\big(\mathrm{F}(A, C) - \mathrm{F}(C)\big) + \frac{1}{3}\big(\mathrm{F}(A, B, C) - \mathrm{F}(B, C)\big)$$

Robustness

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

The Shapley value is the only method that meets the criteria for a “fair” payout for each player (feature):

- **Efficiency:** sum of the Shapley values for all features in each instance, equals to the total coalition value, i.e. difference between prediction, and expected value.
- **Symmetry:** all features have a fair chance to join the game, i.e. if two instances of a feature contribute equally to all possible collations, the Shapley values are the same.
- **Dummy:** if a feature has no contribution to any coalition, then the Shapley value of that feature is zero.
- **Additivity:** the combined Shapley values of any pair of games, is the sum of the Shapley values of each game, $\varphi(v + v') = \varphi(v) + \varphi(v')$

What about SHAP?

- The disadvantage of using Shapley values is that it is computationally very expensive,
  - For M features and N observation, to calculate the entire Shapley values "matrix", we have $\mathrm{N} \times 2^M$ possible coalitions.
  - Plus, the absence of a feature in each coalition must be "simulated" by drawing random instances.
- What SHAP brings to the table, is to address these issues.

Target Space
dim $N \times 1$

Feature Space
dim $N \times M$

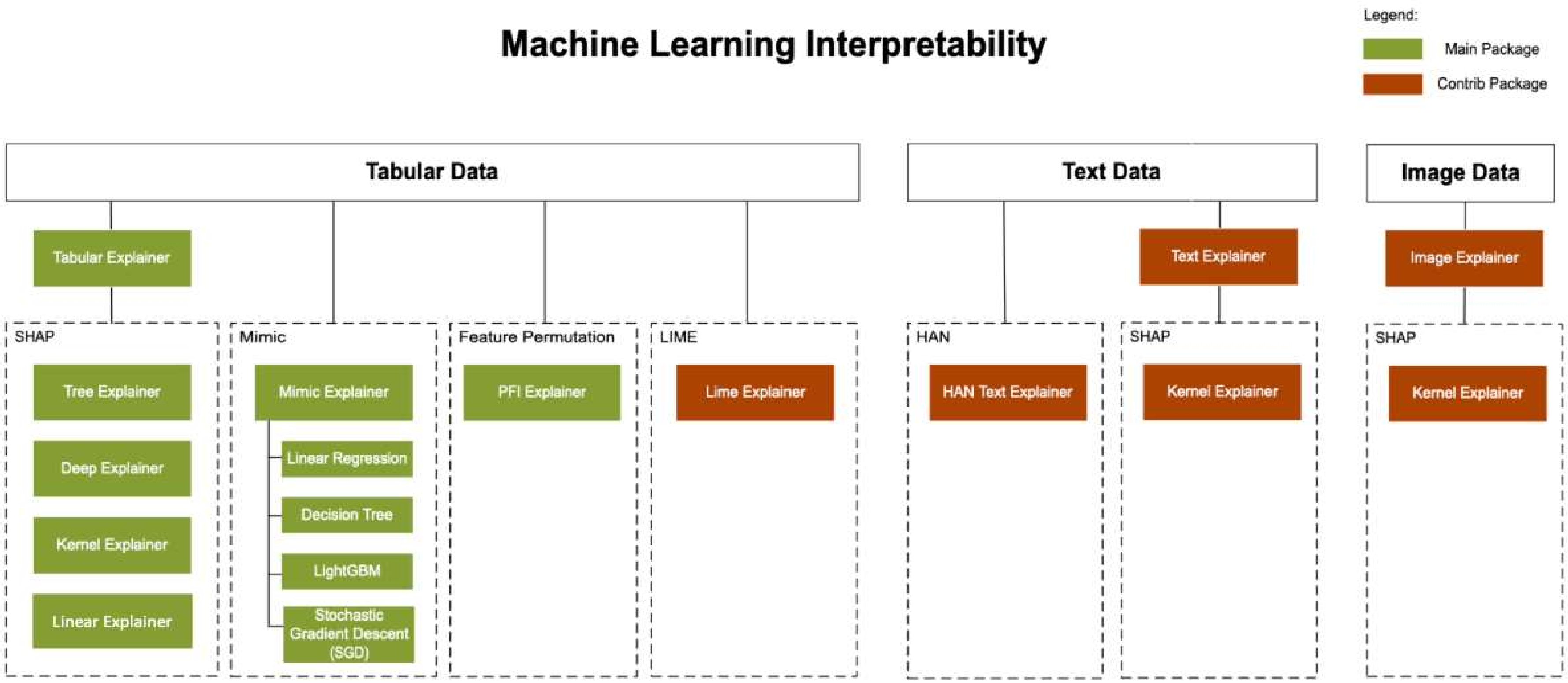
game	score	A	B	...	C
1	y_1	x_{11}	x_{12}	.	x_{1M}
2	y_2	x_{21}	x_{22}	.	x_{2M}
3	y_3	x_{31}	x_{32}	.	x_{3M}
4	y_4	x_{41}	x_{42}	.	x_{4M}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	y_N	x_{N1}	x_{N2}	.	x_{NM}

How SHAP helps with computational overhead?

SHAP optimizes the number of coalitions based on different scenarios:

1. LinearExplainer:
   - For linear models.
2. TreeExplainer:
   - Tree-based models,
   - This relaxes the dependency to background data,
   - Reduces the complexity space from $O(TL2^M)$ to $O(TLD^2)$.
3. KernelExplainer:
   - Model agnostic,
   - Reduces the complexity space from 2^M to $2M + 2048$, based on weights in Shapley formula,
   - Missing features are simulated from background data.
4. DeepExplainer:
   - This is a high-speed approximation algorithm for SHAP values in deep learning models.

Azure Machine Learning-Interpret



Overview of different methods in AML-interpret

SHAP
- Both local and global interpretability
- Can be applied to any type of data
- Both model agnostic and model specific tools
- Provides marginal contribution of features to prediction of target
- Incorporates the interactions between features

Mimic
- Both local and global interpretability
- Model agnostic, based on global surrogate models
- Approximate explanation
- Applicable to tabular data
- Provides marginal contribution of features to prediction of target
- Incorporates the interactions between features

Feature permutation
- Provides global explanation
- Model agnostic
- Univariate model explanation
- Applicable to tabular data
- Provides explanation based on the effect of the feature on model performance

LIME
- Provide local interpretability
- Model agnostic
- Applicable to tabular and image data

Demo

Notebooks

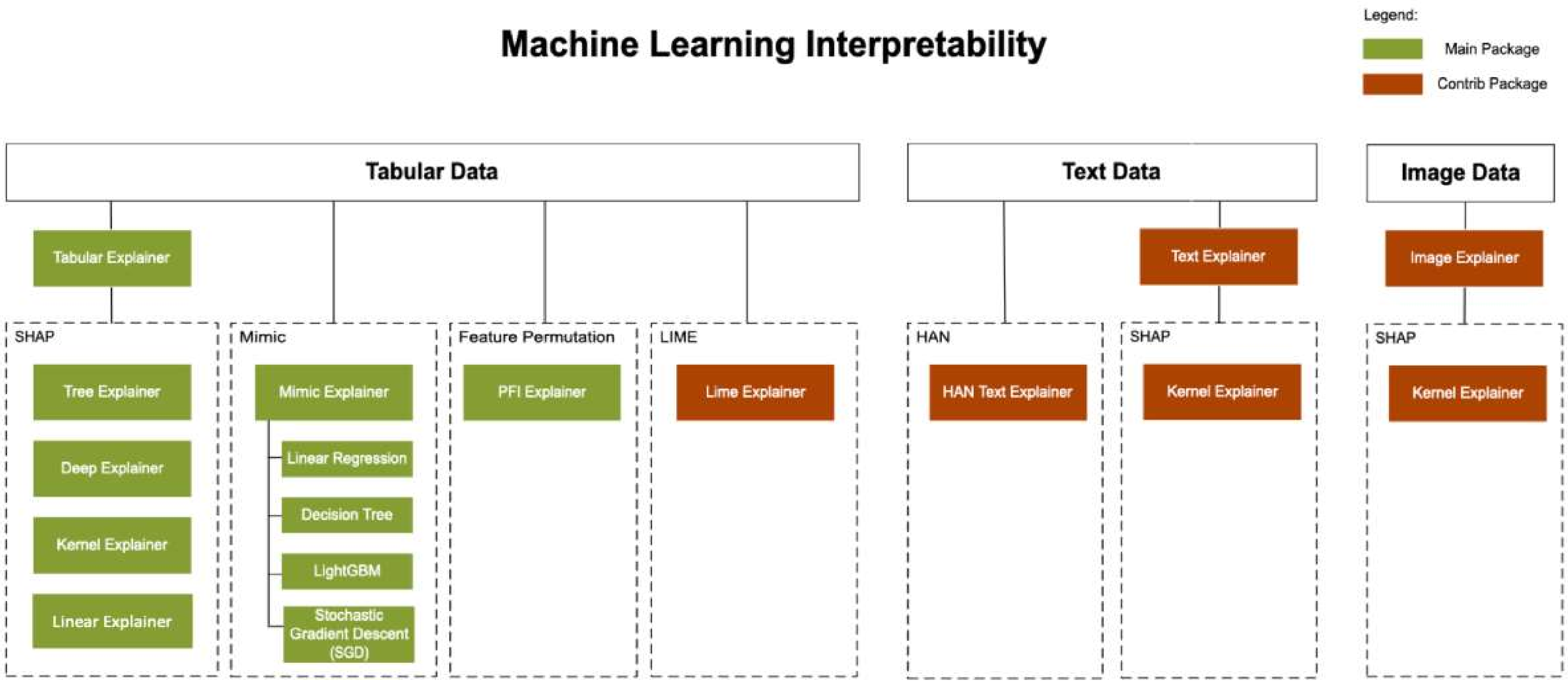
1. Regression use case using Azure ML Interpret
2. Image classification using Azure ML Interpret
3. ~~Image classification using OSS SHAP~~

Session Goals

- Why interpretability is important.
- SHAP, one of the most robust tools for ML interpretability:
  - is applicable to any type of model,
  - is applicable to any type of data,
  - provides both local and global interpretability.
- How to apply Azure Machine Learning Interpret tools for:
  - A regression use case,
  - A generic image classification use cases:
    - Explain on super-pixels from azureml toolkit,
- Limitations of Azure Machine Learning Interpret tools and work-arounds
- Notebooks are available for future reference
  https://github.com/microsoft/AML_Interpret_usecases

Reference 1
Reference 2

Azure Machine Learning-Interpret



Q&A

Thank you for attending the MLADS Conference and helping to build a strong community

To find recordings, presentations, and other resources from the event,
go to: http://aka.ms/spring2020mlads

More information about the Machine Learning Community: http://aka.ms/wwc-ml
More information about the Artificial Intelligence Community: http://aka.ms/wwc-ai

Microsoft