SHAP (SHapley Additive exPlanations)

(approximating complex machine learning models)

Kao-Tai Tsai, PhD.
kaotai.tsai@gmail.com

Presented at the
Biopharmaceutical Applied Statistics Symposium (BASS)

November 6, 2024

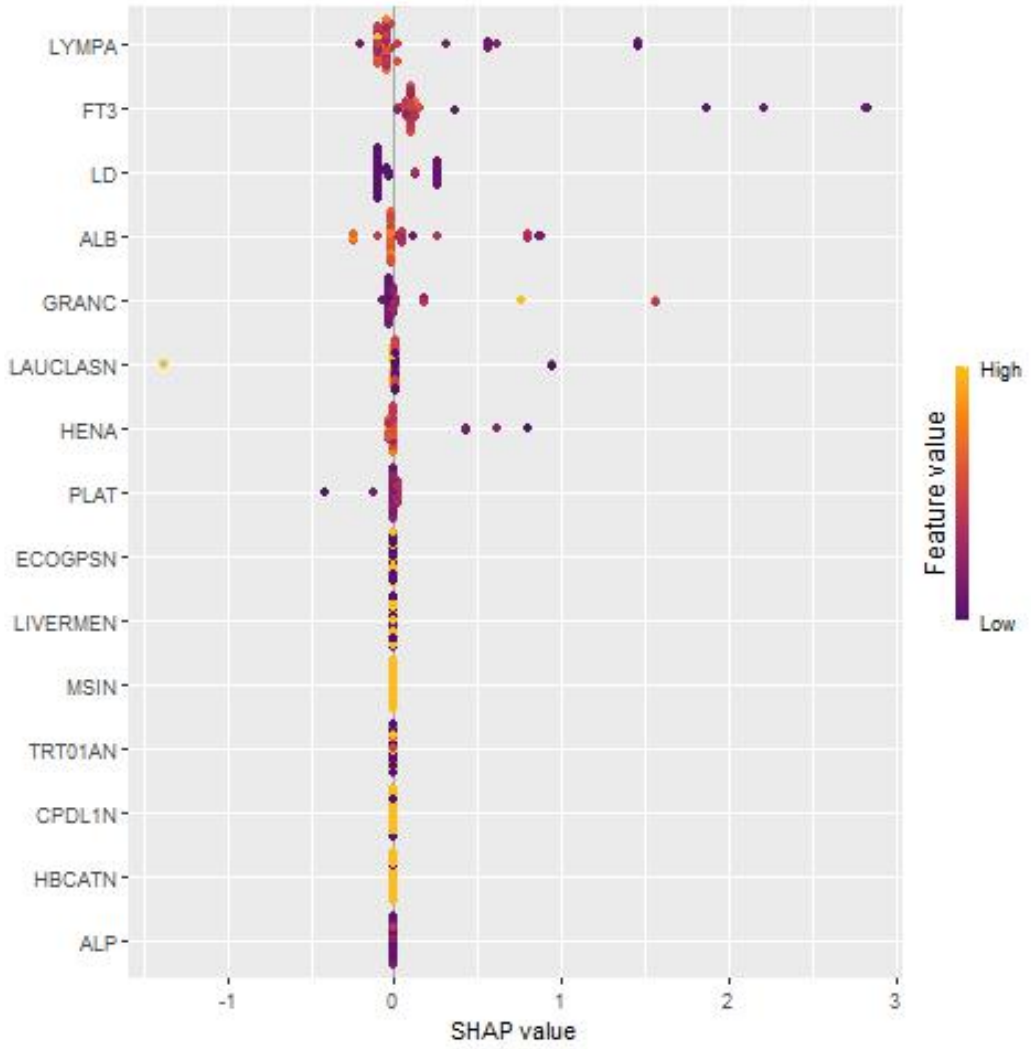
Outline of Presentation

A common scenario in study meetings - 1

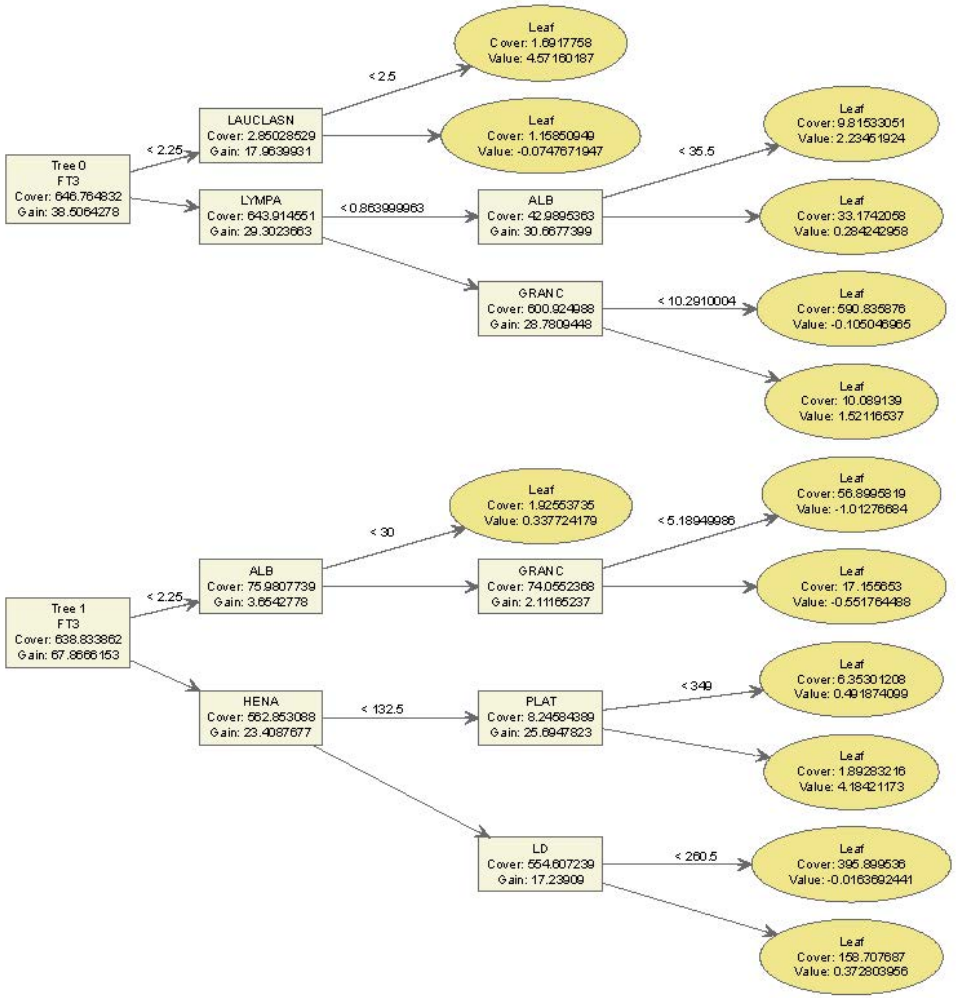
A real scenario in company study meetings:

- Statisticians/Data-analysts are presented a set of data and are asked to identify the important factors or even to build a predict model for the clinical outcomes.
- Many times statisticians will run a GLM model or Cox model and claim to have identified the important factors which can predict the clinical outcomes.
- Sometimes, they will run a machine learning program and present a colorful graph to stun the audience! (see next page)
- Many times, they even don't (or don't know) properly explain how the graph was constructed and the meaning.

A common scenario in study meetings - 2



A common scenario in study meetings - 3



Why SHAP - 1

- The growing availability of big data and computer programs for data analysis have increased the prevalence of using complex methods and models.
- These models very often are not transparent, therefore, it becomes important to consider the trade-off between validity, accuracy, and interpret-ability of a model's output.
- Chen, T and Guestrin, C. [1] had introduced the XGBoost sysytem in 2016 " XGBoost: A Scalable Tree Boosting System," and had been widely used ever since.
- XGBoost is a tree ensemble system and generate multiple trees for data analysis and inferences, however, it is not all that transparent and easily been understood.

Why SHAP - 2

- In their recent paper at the ACM SIGKDD conference, Ribeiro, et al. [2] had asked "Why should I trust you?" Similar questions had been also raised by many researchers.
- The ability to correctly interpret a prediction model's output is extremely important. It can establish appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled.

Why SHAP - 3

How to explain the results?

- The best explanation of a model is the model itself; it perfectly represents itself and can be better explained. For example: parametric models or simple tree structure models.
- For complex models, such as ensemble methods (e.g., randonForest, xgboost, etc.) or deep networks (neural networks, etc.), one cannot use the original model as its own best explanation because it is not easy to understand.
- Instead, simpler explanation models (*if feasible*) can be used, which serves as an interpretable approximation of the original model.

Procedures of approximation - 1

- Proposals of explanation models are available in the literature, e.g., Ribeiro, et al. [2], Lundberg & Lee [3], etc. The general rationale is briefly described below.
- Let f be the original prediction model and g be the corresponding explanation model.
- Local explanation focuses on methods designed to explain a prediction $f(x)$ based on a selected input x. (*Note: global explanation can be much more complicated.*)
- Explanation models often use simplified inputs x_0 that is mapped to the original inputs through a mapping function $x = h_x(x_0)$.
- Local methods try to ensure $g(z_0) \approx f(h_x(z_0))$ if $z_0 \approx x_0$.

Procedures of approximation - 2

- An explanation model is called **Additive Feature Attribution Method** if it is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in R$ (Ref: Lundberg & Lee [3]).

- Methods with explanation models matching this criteria attribute an effect ϕ_i to each feature, and summing the effects of all feature attributions approximates the output $f(x)$ of the original model.

Classic Shapley Value Estimation - 1

- Shapley regression values are feature importance for linear models in the presence of multi-collinearity.
- Note on Shapley regression:
  - if data can be explained using linear models, one most likely does not need methods such as xgboost, randomForest, or deep learning, etc.
  - if data can't be well-explained by linear models, using Shapley regression is unlikely to completely explain the data.
  - most of real data involves variables having various degrees of inter-correlation, with complex co-linearity, and unknown interactions, the true model is difficult to discover.

Classic Shapley Value Estimation - 2

- The estimation method requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all features.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature.
- To compute this effect, a model $f_{S\cup\{i\}}$ is trained with that feature present, and another model f_S is trained without that feature.
- Predictions from the two models are compared on the current input $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ - $f_S(x_S)$ where x_S represents the values of the input features in the set S.

Classic Shapley Value Estimation - 3

- Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S \subseteq F \setminus \{i\}$.
- The Shapley values are then computed and used as feature attributions. They are a weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (2)$$

Classic Shapley Value Estimation - 4

- For Shapley regression values, h_x maps 1 or 0 to the original input space, where 1 indicates the input is included in the model, and 0 indicates exclusion from the model.

- Shapley sampling values are meant to explain any model by: (i) applying sampling approximations to Equation (2), and (ii) approximating the effect of removing a variable from the model by integrating over samples from the training dataset.
- This eliminates the need to retrain the model and allows fewer than $2^{|F|}$ differences to be computed. Since the explanation model from Shapley sampling values is the same as that for Shapley regression values, it is also an additive feature attribution method.

Classic Shapley Value Estimation - 5

Shapley value estimation methods, has the following properties:

Property 1 (Local accuracy)

$$f(x) = g(x_0) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'. \tag{3}$$

Namely, the explanation model $g(x_0)$ matches the original model $f(x)$ when $x = h_x(x_0)$, where $\phi_0 = f(h_x(0))$ represents the model output with no simplified inputs.

In other words, when approximating the original model f for a specific input x, local accuracy requires the explanation model to at least match the output of f for the simplified input x_0 (which corresponds to the original input x).

Classic Shapley Value Estimation - 6

Property 2 (Missingness)

$$x_i' \Rightarrow \phi_i = 0 \tag{4}$$

Missingness constrains features where $x_0' = 0$ to have no attributed impact.

If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact.

Classic Shapley Value Estimation - 7

Property 3 (Consistency) Let $f_x(z') = f(h_x(z')), z'\backslash i$ denote $z'_i = 0$. For any two models f and f_0, if

$$f'_x(z'_i) - f'_x(z'_i\backslash i) > f_x(z'_i) - f_x(z'_i\backslash i)$$

for all inputs $z' \in \{0,1\}^M$, then

$$\phi_i(f', x) > \phi_i(f, x).$$

Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

Classic Shapley Value Estimation - 8

Lundberg & Lee [3] proved the following theorm:

Theorem: **only one possible explanation model** g **follows definition 1** and satisfies the Local accuracy, Missingness, and Consistency properties:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \setminus i)] \tag{5}$$

where $|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

Classic Shapley Value Estimation - 9

Then Lundberg & Lee [3] proposed

- SHAP values as a unified measure of feature importance.
- These are the Shapley values of a conditional expectation function of the original model; thus, they are the solution to Equation (5), where

$$f_x(z') = f(h_x(z')) = E[f(z)|z_S],$$

and S is the set of non-zero indexes in z'.

Stomach (gastric) cancer

Medical images of gastric cancer

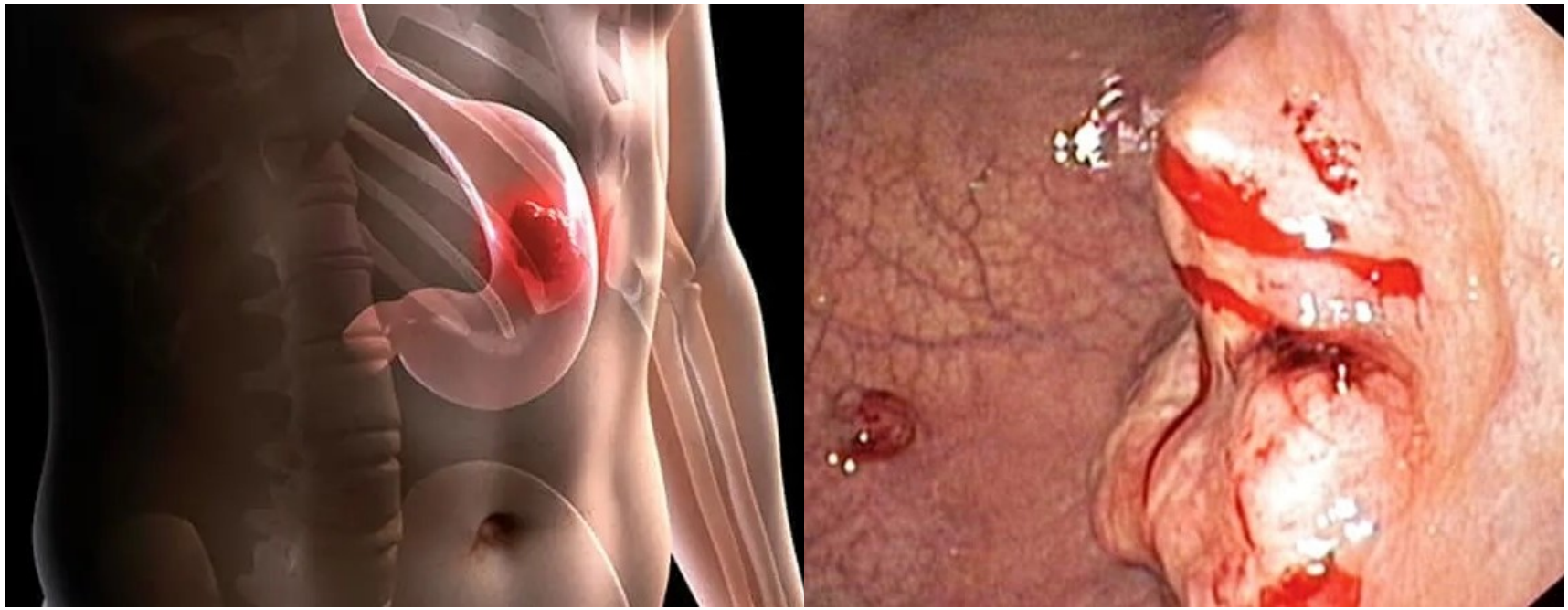


Figure 1: Stomach (gastric) cancer

Description of clinical study - 1

A Clinical Protocol: A randomized, multicenter, open-label, phase 3 study of EXP+chemo vs chemo in subjects with previously untreated advanced or metastatic Stomach Cancer (SC)

Objectives:

- Primary Objectives - EXP+chemo vs chemo: to compare OS and PFS via a 2-arm trial

Results of data analysis - 1

The xgboost computations

```
** Variables selected from databases - (adsl, efficacy, lab)

** Baseline values of input variables to run XGB
(only sample 100 cases as illustration)

 $ X        :'data.frame':        100 obs. of  24 variables:
  ..$ ECOGPSN : num [1:100] 0 0 0 1 1 0 1 1 0 1 ...
  ..$ LIVERMEN: num [1:100] 0 0 0 0 0 0 0 1 1 1 ...
  ..$ MSIN    : num [1:100] 2 2 2 2 2 2 2 2 2 2 ...
  ..$ TRT01AN : num [1:100] 3 2 2 1 1 3 3 2 3 2 ...
  ..$ CPDL1N  : num [1:100] 2 1 2 2 2 2 2 2 2 2 ...
  ..$ HBCATN  : num [1:100] 2 2 2 2 2 2 2 1 2 2 ...
  ..$ LAUCLASN: num [1:100] 2 1 2 2 1 2 2 1 2 1 ...
  ..$ ALB     : num [1:100] 35 42 38 47 37 42 38 39 46 47 ...
  ..$ ALP     : num [1:100] 155 100 192 130 121 58 89 139 199 130 ...
  ..$ AST     : num [1:100] 15 14 17 17 34 20 26 20 21 50 ...
  ..$ CA      : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
  ..$ CREAT   : num [1:100] 56 94 50 93 81 55 122 106 59 66 ...
```

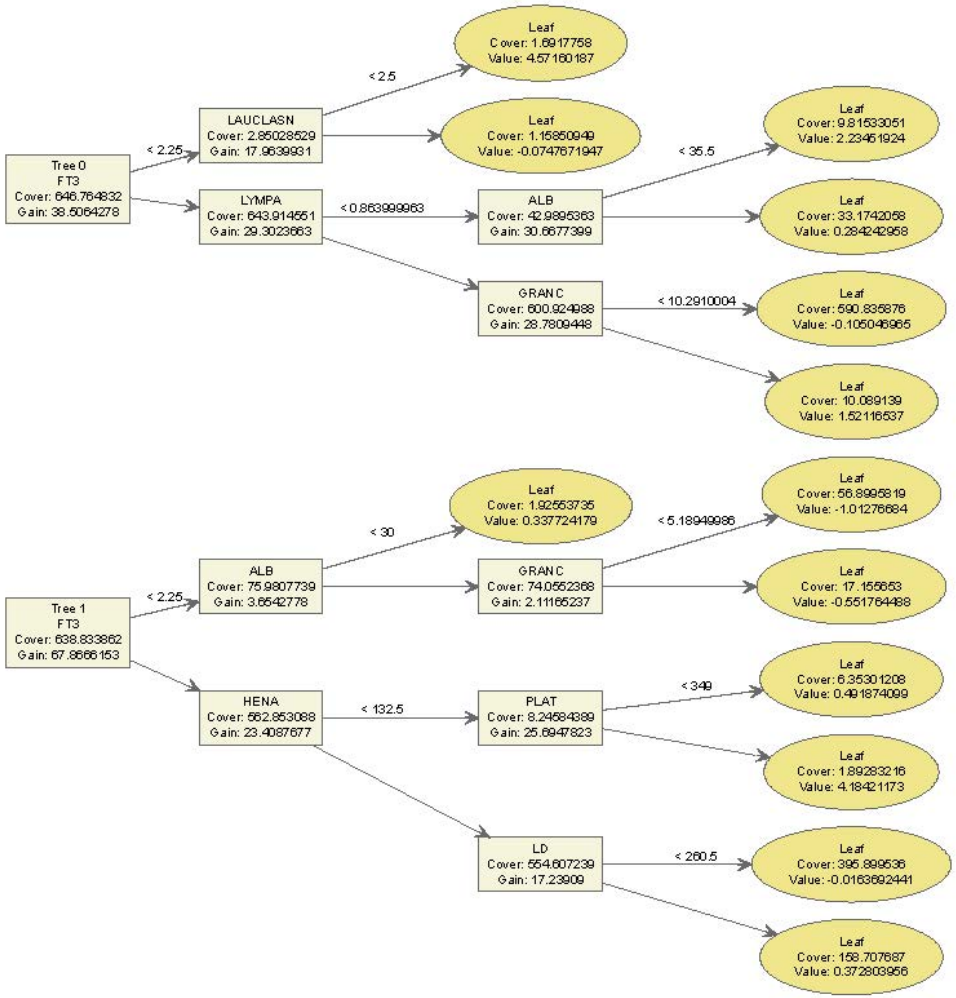
Results of data analysis - 2

```
..$ FT3      : num [1:100] 3.7 3.6 3.6 3 4.4 3.8 2.9 5.6 4.3 5.1 ...
..$ FT4      : num [1:100] 12.9 16.4 14.1 16 16.4 12.9 16.5 16.5 18.1 19.2 ...
..$ GRANC    : num [1:100] 3.73 6.35 2.04 3.89 5.07 ...
..$ HB       : num [1:100] 125 112 123 154 135 127 147 75 138 148 ...
..$ HECA     : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
..$ HENA     : num [1:100] 143 138 141 142 135 141 132 137 140 140 ...
..$ HOCA     : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
..$ HONA     : num [1:100] 143 138 141 142 135 141 132 137 140 140 ...
..$ LD       : num [1:100] 149 127 185 279 99 150 182 196 154 455 ...
..$ LYMPA    : num [1:100] 1.921 1.14 0.787 1.7 3.56 ...
..$ PLAT     : num [1:100] 320 275 188 232 346 370 477 634 174 172 ...
..$ TBILI    : num [1:100] 5.1 6 13.7 18.5 6.5 7 6.5 5.6 17.1 10.5 ...
```

Results of data analysis - 3

```
call:
xgb.train(params = param, data = dtrain, nrounds = 2, watchlist = watchlist)
params (as set within xgb.train):
max_depth = "3", eta = "1", verbose = "0", nthread = "2",
objective = "survival:cox", eval_metric = "cox-nloglik",
validate_parameters = "TRUE"
xgb.attributes: niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 24
niter: 2
nfeatures : 24
evaluation_log:
 iter train_cox_nloglik eval_cox_nloglik
    1          5.980229         5.980229
    2          5.973864         5.97386
```

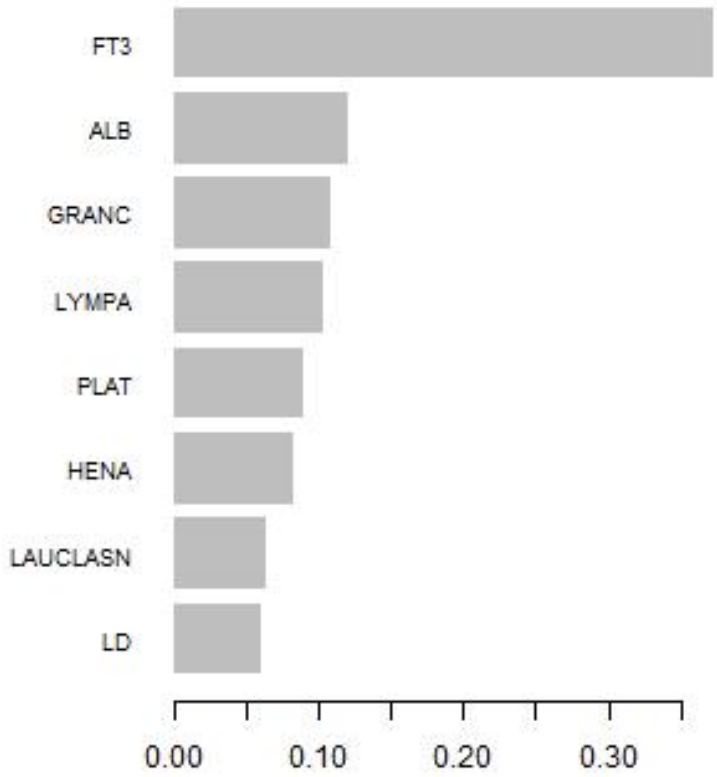
Results of data analysis - 4



Results of data analysis - 5

	Feature	Gain	Cover	Frequency
1:	FT3	0.37298139	0.3337465994	0.18181818
2:	ALB	0.12034509	0.0308851722	0.18181818
3:	GRANC	0.10832033	0.1752275935	0.18181818
4:	LYMPA	0.10274442	0.1671628190	0.09090909
5:	PLAT	0.09009496	0.0021406544	0.09090909
6:	HENA	0.08207939	0.1461189359	0.09090909
7:	LAUCLASN	0.06298809	0.0007399456	0.09090909
8:	LD	0.06044632	0.1439782800	0.09090909

Results of data analysis - 6



SHAP analysis - 1

The SHAP computations following xgb outputs

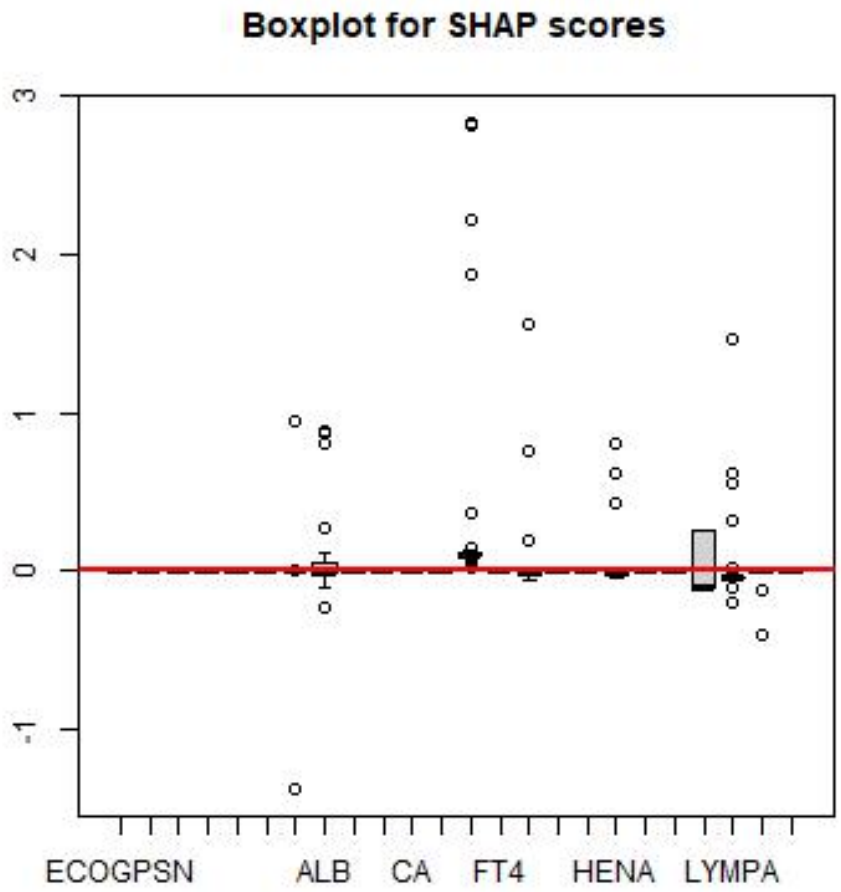
```
** SHAP importance scores of input variables for each case

'data.frame':   100 obs. of  24 variables:
 $ ECOGPSN : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LIVERMEN: num  0 0 0 0 0 0 0 0 0 0 ...
 $ MSIN    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TRT01AN : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CPDL1N  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ HBCATN  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LAUCLASN: num  0.00416 0.00416 0.00416 0.00416 0.00416 ...
 $ ALB     : num  0.0482 -0.0165 -0.2391 -0.0167 -0.0167 ...
 $ ALP     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AST     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CA      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CREAT   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ FT3     : num  0.1015 0.0742 0.1026 0.1245 0.1014 ...
 $ FT4     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ GRANC   : num  -0.03258 -0.00563 -0.01895 -0.03258 -0.03258 ...
```

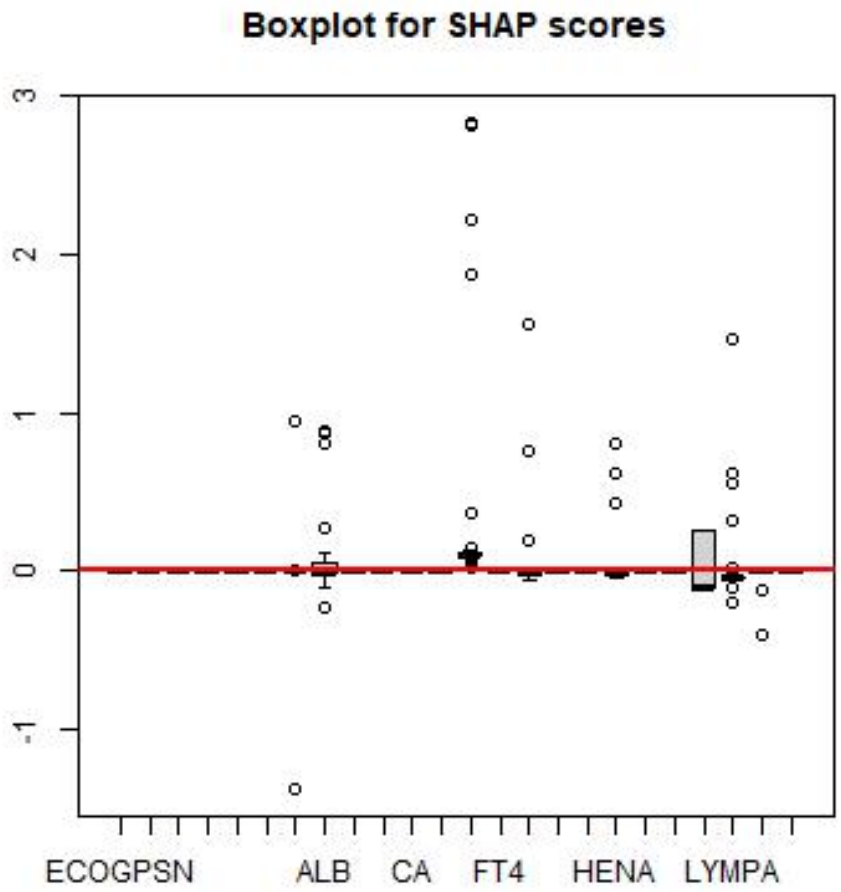
SHAP analysis - 2

```
$ HB      : num  0 0 0 0 0 0 0 0 0 0 ...
$ HECA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ HENA    : num  -0.01197 -0.01197 -0.01197 -0.00923 -0.01197 ...
$ HOCA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ HONA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ LD      : num  -0.104 -0.104 -0.104 0.259 -0.104 ...
$ LYMPA   : num  -0.1049 -0.0398 0.557 -0.0398 -0.0398 ...
$ PLAT    : num  -0.00572 -0.00572 -0.00572 -0.00572 -0.00572 ...
$ TBILI   : num  0 0 0 0 0 0 0 0 0 0 ...
```

SHAP scores via xgb model



SHAP scores via xgb model



Distributions of SHAP importance of FT3 and LYMPH

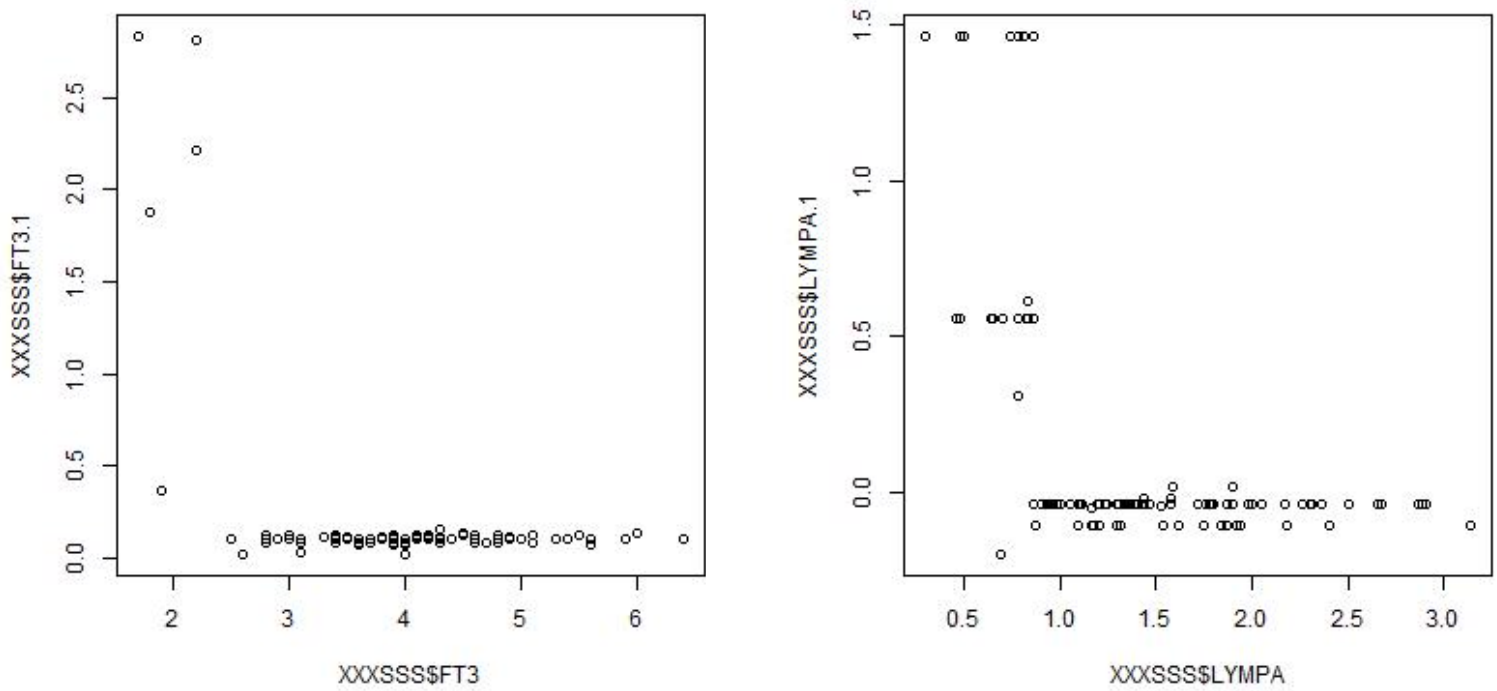
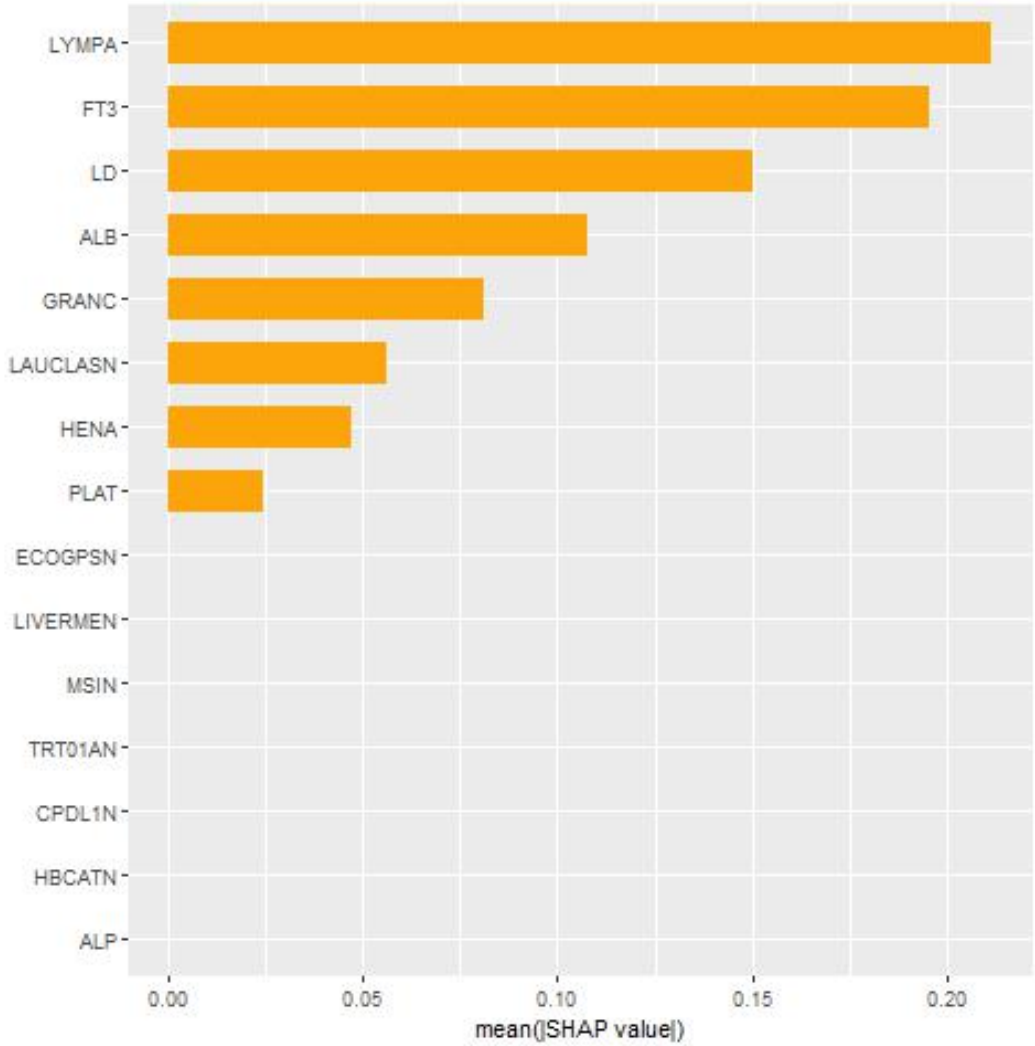
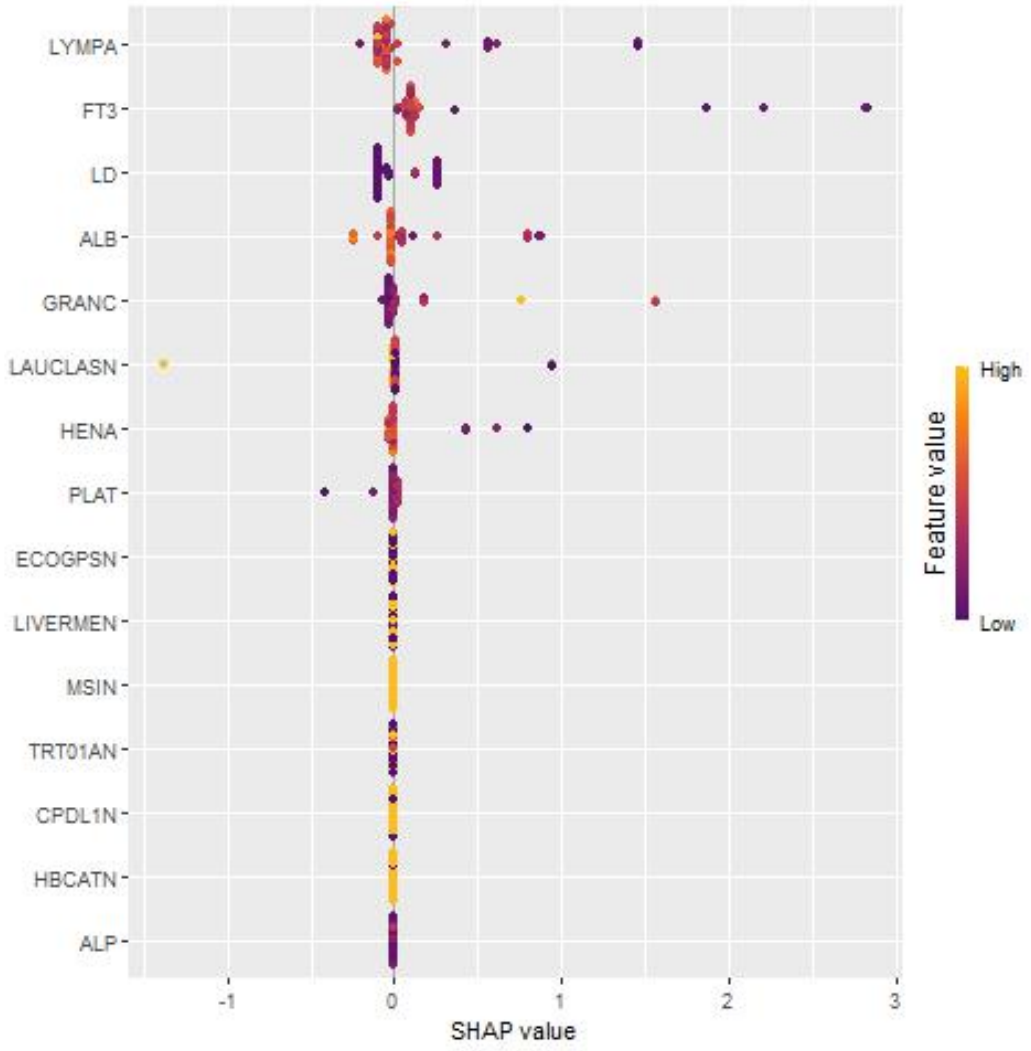


Figure 2: Distributions of SHAP importance of FT3 and LYMPH

SHAP (ordered by means) via xgb model



SHAP scores via xgb model



Summary - 1

- Broad availability of software programs and ease of use can make anyone becoming a data analyst.
- However, lack of the background knowledge of the subject matters in science, statistics, and programming algorithms, misleading results can easily be generated without been realized by the analysts and stakeholders.
- With the emphasis of evidence-based decision-makings, the correct and well-validated analytical results serve as the foundation of any further developments, especially in the recent enthusiasm of artificial intelligence.
- The technologies in machine learning provide important and convenient tools, in addition to the classical statistical data analysis methodologies.

Summary - 2

- Many of the ML methods require complicated computing and are not transparent in computations and background theory. Therefore, properly explaining the results they created becomes critical for it to be trusted by users .
- The SHAP tool is based on the outputs of xgboost, which is an recursive partitioning method based on ensemble of partitioning. Given its popularity among practitioners, it is essential to have proper understanding of its inner-workings.
- Th material discussed here provides the theory of SHAP so that it can be better understood and also raises proper caution when it is utilized.


Chen, T and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System, 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.


Ribeiro, M., Singh, S., and Guestrin, C.. (2016) Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016, pp. 1135-1144.


Lundberg, S. & Lee, S (2017) A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.