Further Explanation for Summary Plot And Force Plot

I have used the python SHAP library to do the feature importance test (https://shap.readthedocs.io/en/latest/) this week, I myself am slightly new to the plots the library is generating, so here comes my interpretation of what I see in the graphs for our dataset.

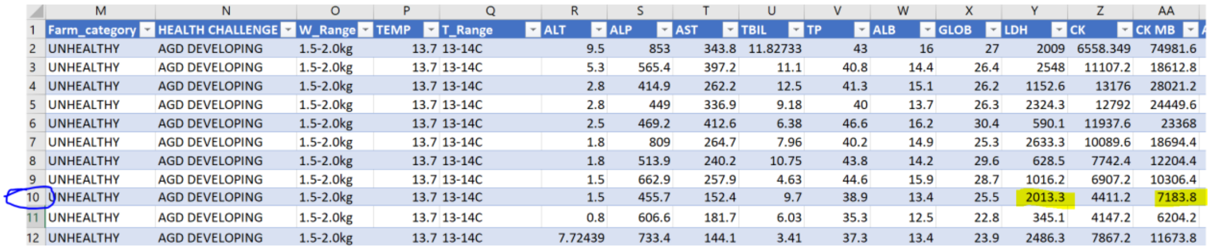
For any binary classification task, even if don't know anything there is a 50/50 prediction probability of a sample being either on class 0 or 1. However, in our dataset we have healthy and unhealthy ratio of 375:281, which is about 57:43, so in case we train our model to detect unhealthy fishes, the baseline probability that the fish will be unhealthy in this dataset is 0.43.

This value is reflected on the force plots we are creating for individual fishes. I have selected unhealthy as my positive case/class for these plots, so model can go up to 1 when it is 100% confident of its decision that the fish is unhealthy, when it is not it can go down to 0.

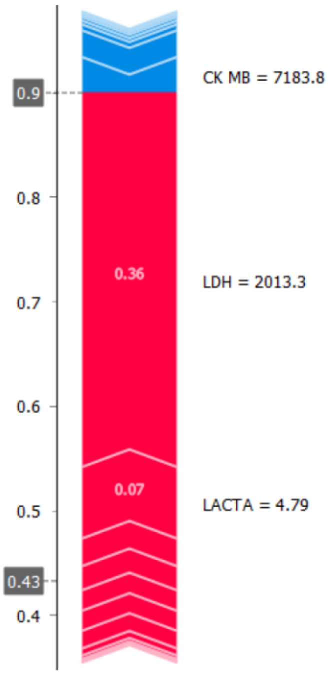
If the value is falling from the baseline value it would mean fish is on the healthy side of the spectrum.

The force plot starts with a baseline prediction value of a certain class: Red arrows represent feature effects (SHAP values) that drives the prediction value higher while blue arrows are those effects that drive the prediction value lower.

Consider the fish on row 10 of the attached file, LDH=2013.3 is on higher the range of LDH compared to all the LDH values in this subgroup.



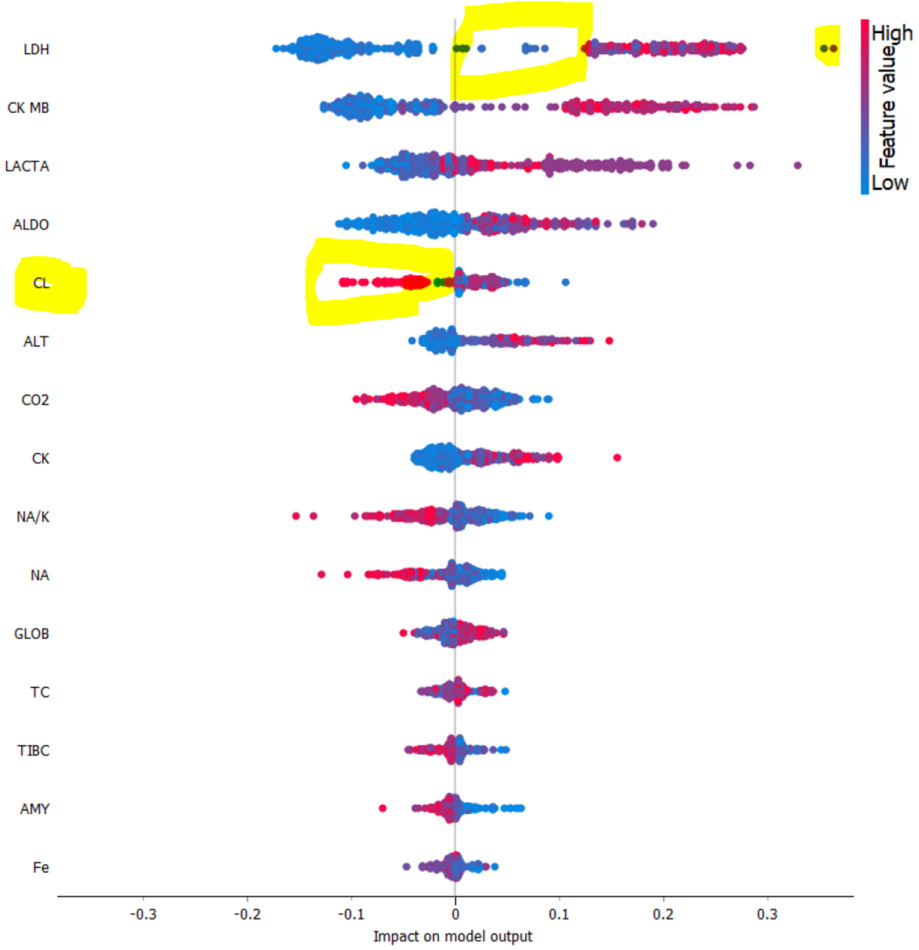
	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Farm_category	HEALTH CHALLENGE	W_Range	TEMP	T_Range	ALT	ALP	AST	TBIL	TP	ALB	GLOB	LDH	CK	CK MB
2	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	9.5	853	343.8	11.82733	43	16	27	2009	6558.349	74981.6
3	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	5.3	565.4	397.2	11.1	40.8	14.4	26.4	2548	11107.2	18612.8
4	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	2.8	414.9	262.2	12.5	41.3	15.1	26.2	1152.6	13176	28021.2
5	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	2.8	449	336.9	9.18	40	13.7	26.3	2324.3	12792	24449.6
6	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	2.5	469.2	412.6	6.38	46.6	16.2	30.4	590.1	11937.6	23368
7	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	1.8	809	264.7	7.96	40.2	14.9	25.3	2633.3	10089.6	18694.4
8	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	1.8	513.9	240.2	10.75	43.8	14.2	29.6	628.5	7742.4	12204.4
9	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	1.5	662.9	257.9	4.63	44.6	15.9	28.7	1016.2	6907.2	10306.4
10	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	1.5	455.7	152.4	9.7	38.9	13.4	25.5	2013.3	4411.2	7183.8
11	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	0.8	606.6	181.7	6.03	35.3	12.5	22.8	345.1	4147.2	6204.2
12	UNHEALTHY	AGD DEVELOPING	1.5-2.0kg	13.7	13-14C	7.72439	733.4	144.1	3.41	37.3	13.4	23.9	2486.3	7867.2	11673.8



But when we observe the CK-MB value, it is slightly on the lower value range compared to its unhealthy neighbours, this is quite nicely reflected in the force plots, that if High LDH value has increased the risk of fish being unhealthy by 0.36, the lower value of CK-MB has reduced that by 0.1 (probably that’s falling more within the range the model experienced from its training set that were labelled as healthy).

So the force plot is comparing each of the biomarkers with the models definition/baseline of healthy or unhealthy for the biomarker to compare it’s impact on the positive side or on the negative side based on this comparison.

The summary plot is holding all the samples in the dataset, consider, the fish on row 10 is a sample marked in yellow on the summary pot (impact=0.36 for its LDH value), so x-axis is holding either positive or negative impact on the models baseline value (shown as 0 in summary plots to maintain symmetry).



Our positive class again is the unhealthy class so an increase towards the positive means, the biomarker is reinforcing the models prediction of the fish being unhealthy, any increase on the negative direction indicates it is not agreeing with models prior/baseline values for unhealthy.

The Second thing to notice, the colour blue here means lower value for a specific feature in this plot, and red means high value of the feature. So in the plot there are samples with low value of LDH on the right side as well, unfortunately, it is not a linear scale, it is plotted sample by sample (or case by case), based on the impact value it had on model's decision of the sample being unhealthy.

To make the statement around colour for the feature value clear, let's look at chlorine, CL, the higher values are mostly on the left with lower values and some higher values being on the right. So, this can be an interesting way to look at interaction of various biomarker values how the mostly impacted model's decision making.

The third thing, the features are sorted in descending order in terms of their prediction capability. Suppose the values of Fe were a complete mix-up of high and low around the baseline, hence not a good predictor at all.

Another interesting explanation what the library does on the backend can be here from the developer of the library: https://www.youtube.com/watch?v=-taOhqkiuIo