Interpretable Machine Learning Questionnaire

This is an internal survey within VQD Data Science, part of the Master Thesis Research of Daniël de Bondt into different Machine Learning explanation methods.

It should take approximately 10-15 minutes to complete.

Thank you in advance!

***** Required

* This form will record your name, please fill your name.

Introduction

Within the Machine Learning community an increasing need is present for models to be interpretable, transparent and provide explanations. This interpretability can be achieved through two ways, either using explainer methods on complex models or develop and train inherently interpretable models. My research considers a comparison between these explanation methods, both in terms of performance and most importantly interpretability.

Four different methods are compared:
- Two recently developed inherently interpretable models: SLIM and EBM
- One explanation method to apply to complex Black Box models: SHAP
- One benchmark method that is also inherently interpretable: Logistic Regression

The ML task that is considered is binary classification, where from a set of input features x the output y needs to be predicted as either a zero or a one (success or failure). Examples include medical diagnosis, customer churn or credit card fraud prediction.

One specific data set is used in the survey, the UCI adult income data set. This is the same one used in a survey by Kaur et al. (2020). It has almost 50.000 instances of 14 features (like age, education, marital status etc.) that are used to predict wheter or not a person earns over 50K USD, where about 24% of observations falls into this category. The features are preprocessed into one-hot encoded binary variables (0 or 1) to incorporate categorical variables into linear models.
https://archive.ics.uci.edu/ml/datasets/adult (https://archive.ics.uci.edu/ml/datasets/adult)

First a few general questions will be asked.
Next, all four models will be visited with their own set of questions regarding the specific explanations.
Lastly, some of the experimental results will be shown followed by a few concluding questions.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan,J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–14.

1

Please give your consent that your response to this survey will be analysed and included in the research, where all results will be published anonymously. *

◯ Accept

2

How would you rate your background in Machine Learning?

- ◯ Some contact or general interest
- ◯ Some experience and great interest
- ◯ Considerable experience and daily work use

3

Have you ever seen or used the UCI adult income data set?

- ◯ Yes
- ◯ No

The Logistic Regression

$$\operatorname{logit} p(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x,$$

The Logistic Regression, or Logit model, is a very popular classification model with a linear formulation. It is part of the group of Generalized Linear Models and is widely used in practice.

In the formula above, p(x) represents the predicted probability (between 0 and 1) of a 1 prediction. After summing the product of all estimated beta coefficients with a data point, this probability is reached by applying the sigmoid function which is the inverse of the logit function (the middle term in the above formula). This probability can be turned into a prediction by checking if the probability exceeds a certain threshold, often 0,5.
The betas are estimated using Maximum Likelihood estimation on a training data set.

4

Please rank your familiarity with the Logistic Regression model

- ◯ Completely new
- ◯ Heard of it
- ◯ Familiar/Used it

5

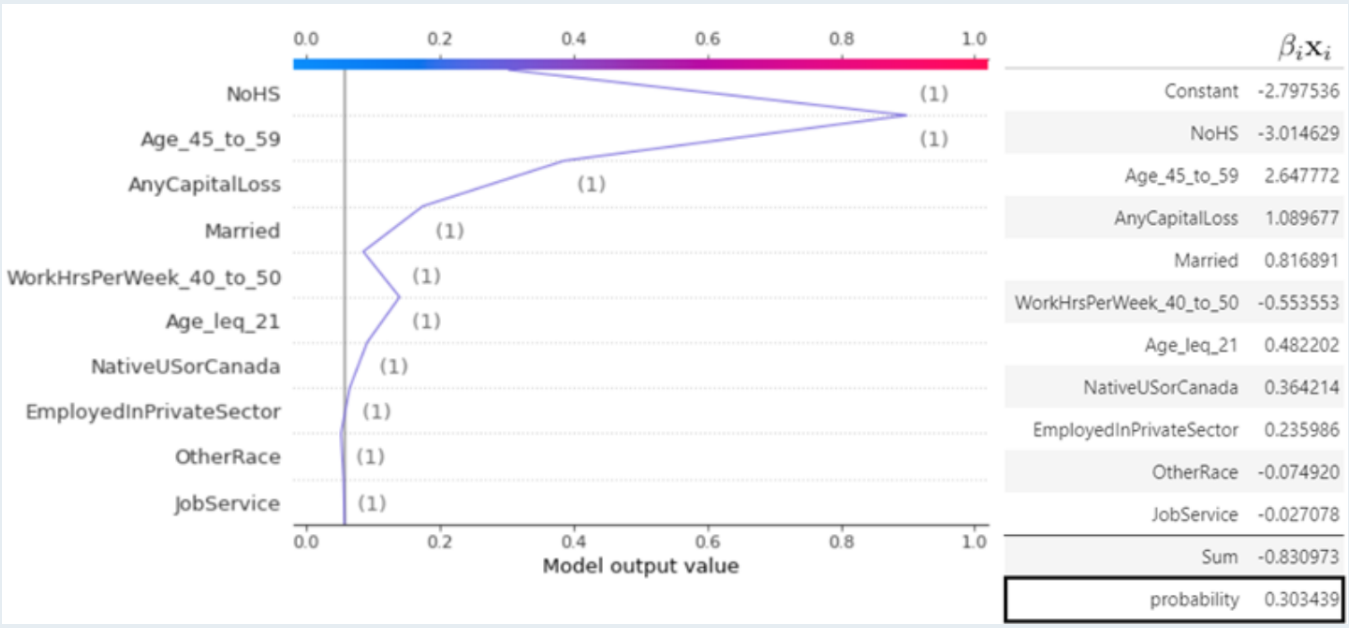
	beta		beta
Constant	-2.797536	WorkHrsPerWeek_40_to_50	-0.553553
NoHS	-3.014629	Age_leq_21	0.482202
Age_45_to_59	2.647772	NativeUSorCanada	0.364214
Age_30_to_44	2.215735	Graduate	-0.326301
Age_geq_60	2.114820	JobSkilledSpecialty	0.304789
HSDiploma	-1.867840	EmployedInPublicSector	0.276499
AnyCapitalGains	1.688607	EmployedInPrivateSector	0.235986
NeverMarried	-1.453965	Female	-0.221590
Age_22_to_29	1.267277	Black	-0.185929
DivorcedOrSeparated	-1.224919	JobAdministrative	0.140626
WorkHrsPerWeek_lt_40	-1.095446	NativeImmigrant	-0.075086
AnyCapitalLoss	1.089677	OtherRace	-0.074920
ProfVocOrAS	-1.083513	JobService	-0.027078
Widowed	-0.962988	SelfEmployedNotInc	-0.007092
JobAgriculture	-0.880738	JobArmedForces	0.000000
Married	0.816891	WorkHrsPerWeek_geq_50	0.000000
JobManagerial	0.787514	Male	0.000000
SelfEmployedInc	0.707983	White	0.000000
Bachelors	-0.701262		

Here, the estimated beta coefficients are displayed for the adult data set. Does the model make use of gender to explain/predict income?

- ◯ No, because the Male coefficient equals zero
- ◯ Yes, the coefficient for Female is lower than for Male
- ◯ Don't/Can't know / Need more information

6

Below an explained prediction is displayed for a single observation i from the adult income data. Would the model prediction (consider a threshold of 0.5) have changed if this person had not had any capital loss?



- ◯ Yes
- ◯ No
- ◯ Can't know

7

How confident are you in your understanding of this explanation method

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

8

How confident are you explaining this prediction to your stakeholders using this method?

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

9

How confident are you that the underlying model could be of value in practice?

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

The Supersparse Linear Integer Model

SLIM (Ustun et al., 2013) is a method to create scoring systems that are used for classification. Its formulation is similar to Logit, but coefficients are limited to integer values and the model is optimized using Mixed-Integer Programming. An example of such a scoring system is displayed below. It is highly intelligible and can be easily integrated with human decision making. It is commonly applied for medical diagnoses.

Ustun, B., Traca, S., and Rudin, C. (2013). Supersparse linear integer models for interpretable classification. arXiv preprint arXiv:1306.6677.

PREDICT MUSHROOM IS POISONOUS IF SCORE > 3

1.	$spore\_print\_color = green$	4 points		.....
2.	$stalk\_surface\_above\_ring = grooves$	2 points	+	.....
3.	$population = clustered$	2 points	+	.....
4.	$gill\_size = broad$	-2 points	+	.....
5.	$odor \in \{none, almond, anise\}$	-4 points	+	.....
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	.....

10

Please rank your familiarity with the SLIM model

- Completely new
- Heard of it
- Familiar/Used it

11

Below is the trained scoring system in practice for predicting a certain observation i from the adult data set. Would the model prediction be different if the person in question had had a high school diploma?

PREDICT 0 IF SCORE < -1		Contributions	X_i
Married	2 points	+2	1
Any capital gains	2 points	+0	0
Age 22 to 29	-2 points	+0	0
Highschool diploma	-2 points	+0	0
Work hours per week lower than 40	-2 points	+0	0
No Highschool diploma	-4 points	−4	1
ADD POINTS FROM ROWS 1 to 3	SCORE $= -2$		

- Yes
- No
- Can't be sure

12

How confident are you in your understanding of this explanation

1 2 3 4 5 6 7

13

How confident are you explaining this prediction to your stakeholders using this method?

1 2 3 4 5 6 7

14

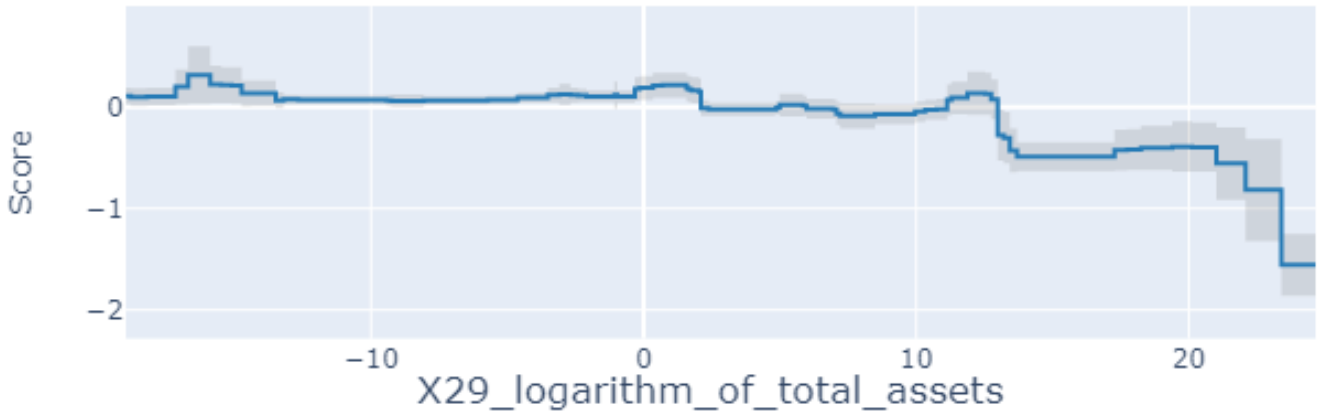
How confident are you that the underlying model could be of value in practice?

1 2 3 4 5 6 7

The Explainable Boosting Machine

The Explainable Boosting Machine is defined by Nori et al. (2019) as a Generalized Additive Model (GAM) that is trained using boosted trees. These trees result in a set of additive partial dependency plots for each feature (example below). These plots act like a look up table where for a specific observation, you add the scores (y-axis) from all of the specific feature values (x-axis) and then apply the sigmoid function (like the logit) to reach your estimated probability. This allows for a high model complexity while retaining transparency.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability.arXiv preprint arXiv:1909.09223.



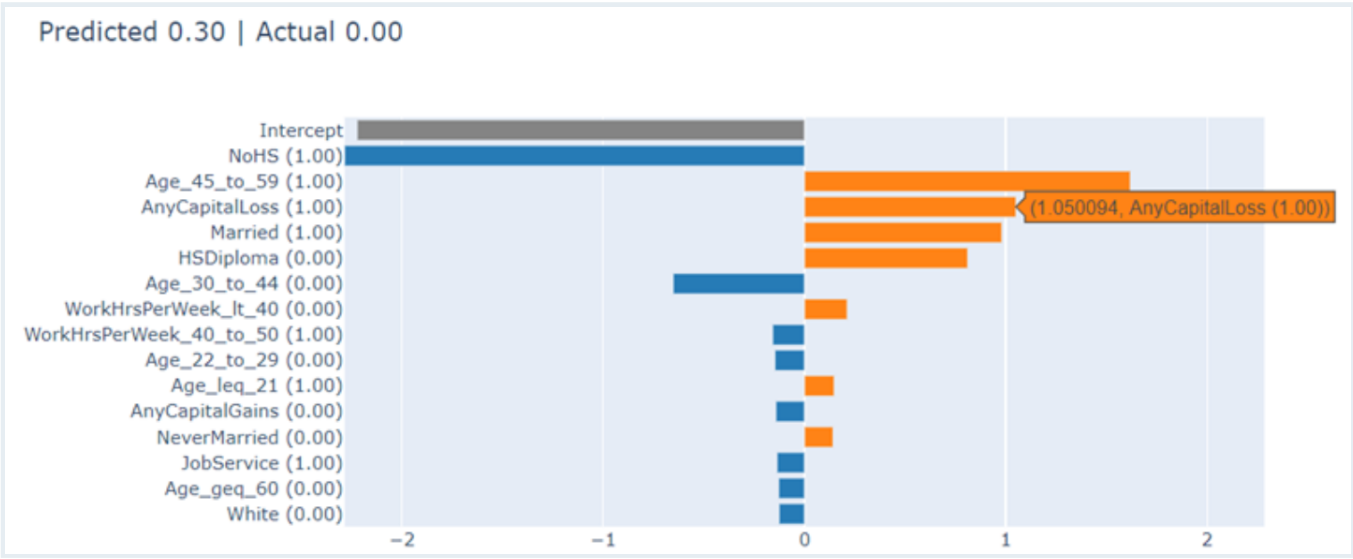
15

Please rank your familiarity with the EBM model

- ◯ Completely new
- ◯ Heard of it
- ◯ Familiar/Used it

16

Below you see an explanation for a prediction on the adult data set made by the EBM model.
Did the EBM make a correct prediction for this observation, considering a threshold of 0.5?



- ◯ Yes
- ◯ No
- ◯ Can't be distinguished

17

How confident are you in your understanding of this visualization

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

18

How confident are you explaining this prediction to your stakeholders using this method?

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

19

How confident are you that the underlying model could be of value in practice?

1 2 3 4 5 6 7

◯ ◯ ◯ ◯ ◯ ◯ ◯

SHAP (Shapley values)

SHapley Additive exPlanations (SHAP) is a framework introduced by Lundberg and Lee (2017) to provide model agnostic explanations. It uses coalition theory from Game Theory to compute Shapley values that approximate individual feature contributions on a local prediction level. It can be applied to any model, but several model specific techniques considerably improve computation efforts.

Here, SHAP is applied to an XGBoost classification model to provide interpretability to XGBoosts high performance.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems, pages 4765–4774.

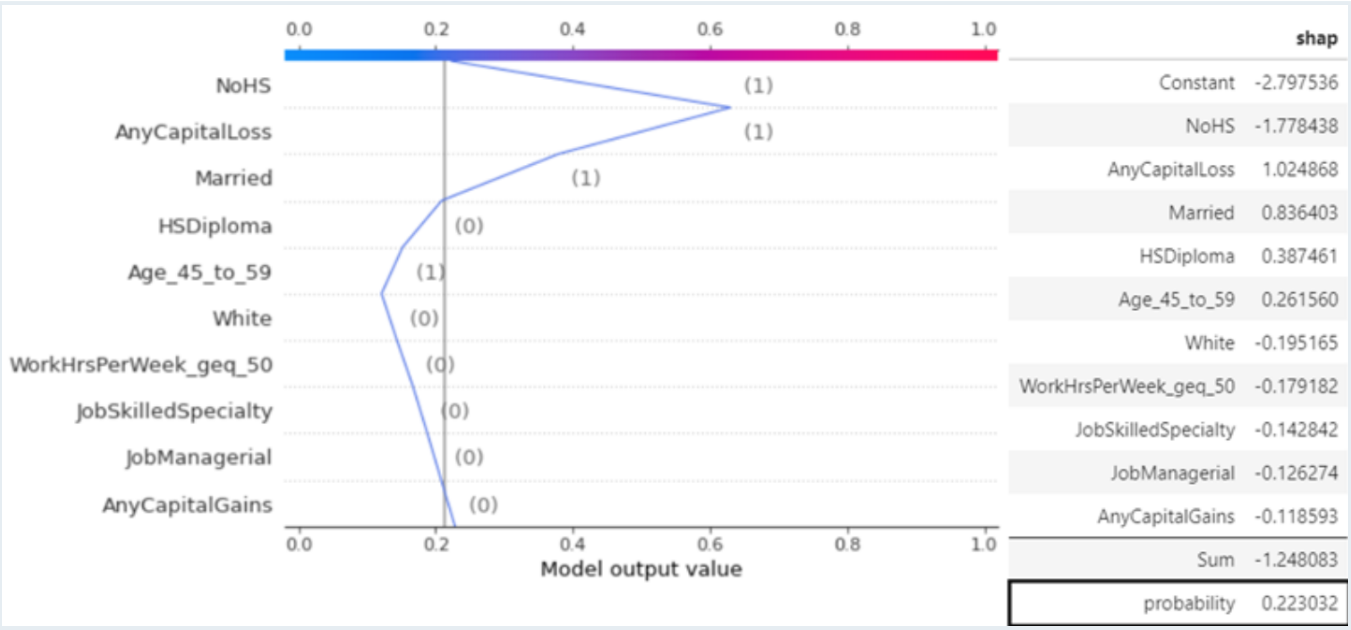
20

Please rank your familiarity with SHAP (Shapley values)

- ◯ Completely new
- ◯ Heard of it
- ◯ Familiar/Used it

21

Below an explained prediction is displayed for a single observation i from the adult income data.
Which feature is the strongest contributor in favor of a high income prediction despite a model prediction of only about 0.2



- ◯ Education level: no highschool diploma
- ◯ The existence of capital losses
- ◯ Marital status: married

22

How confident are you in your understanding of this visualization

1 2 3 4 5 6 7

23

How confident are you explaining this prediction to your stakeholders using this method?

1 2 3 4 5 6 7

24

How confident are you that the underlying model could be of value in practice?

1 2 3 4 5 6 7

General Preference

In the following table the classification performance of all models is displayed, along with how many of the 36 features are actually used. The AUC is defined as the area under the ROC curve which is a common measure that combines both the true positive rate (how many ones did we correctly classify out of all ones) and the false positive rate (how many zeros did we accidently flag as a one out of all zeros).

Table 2: The AUC performance statistic for different models on the Adult Income data set with best performing in bold. Additionally the amount of removed features (zero coefficient/contribution) is displayed.

model	EBM	Logit	SHAP (XGBoost)	SLIM
train	0.892	0.892	**0.905**	0.853
test	0.889	0.889	**0.894**	0.852
number of removed features	0	4 (11%)	2 (6%)	30 (83%)

25

Considering the previously shown explanations and the above performance, what would be your preferred model and explanation method of choice

- ◯ Logit
- ◯ SLIM
- ◯ EBM
- ◯ SHAP

26

Reasoning (Optional)

Final remarks

Thank you very much for completing this Survey. I look forward to presenting my Thesis (including survey results) during an upcoming knowledge share session.

Daniël de Bondt

27

Do you have any final comment or remark?

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms