Can Explainable Artificial Intelligence Support Software Modellers in Model Comprehension?

Context and purpose of the experiment

Thank you for your willingness to participate in our experiment, which aims to evaluate the application of explainable methods for machine learning in the context of software modelling tasks. We have selected a software model for which a machine learning algorithm has returned a prediction. You will be asked to answer some questions based on the information provided and on the visual explanations generated by three explainable methods.

All the necessary information is contained in this document, and you do not need to run any code. In total, the survey is expected to take less than 30 minutes to complete. Please contact the researchers if you encounter any problems in understanding the task at hand or the meaning of some questions.

We ask you to use a stopwatch to keep track of the time spent on some questions. We deeply appreciate your participation.

Consent of participation

All information you provide will be treated confidentially and will only be used for research purposes. Some personal information about you may be collected if you choose to participate in the survey. However, this information will not be made public and responses to questions will be reported in aggregated form. We will not disclose your personal information or responses to third parties.

Participant information

Name:

Affiliation:

Position (student, professor, professional, etc.):

Experience in software modelling: Low / Medium / High

Experience in machine learning: Low / Medium / High

Experience in explainable artificial intelligence: Low / Medium / High

Problem: Detection of dummy models with machine learning

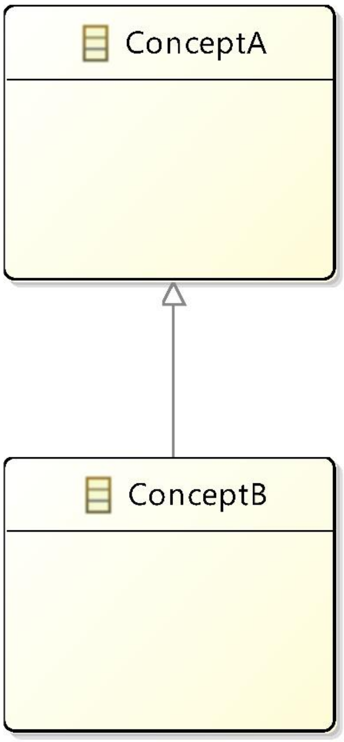
Introduction

We have trained a classifier using a machine learning algorithm to predict if a given Ecore model represents a dummy model (example, testing purposes) or a realistic model. The classifier uses the following features of model elements to make its prediction:

- Number of elements of type “reference”
- Number of elements of type “class”
- Number of elements of type “attribute”
- Number of elements of type “package”
- Number of elements of type “enum”
- Number of elements of type “datatype”

Example

The following Ecore model is correctly predicted as “dummy”:



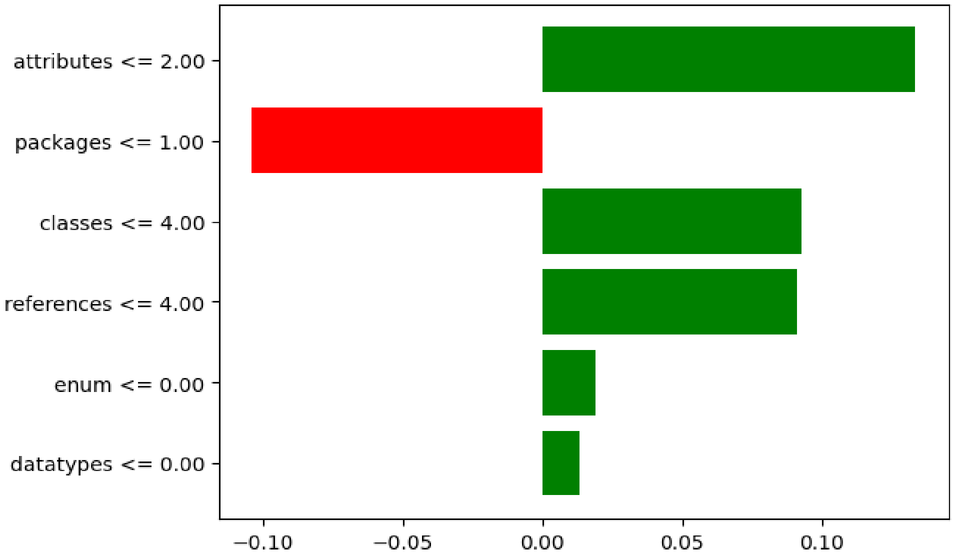
The feature values extracted from the model are:

Reference	Class	Attribute	Package	Enum	Datatype
0	2	0	1	0	0

Three explainable methods (LIME, SHAP and BreakDown) are executed to understand the relevant features that influence the prediction made by the algorithm. Each method generates a visual explanation that can be interpreted as follows:

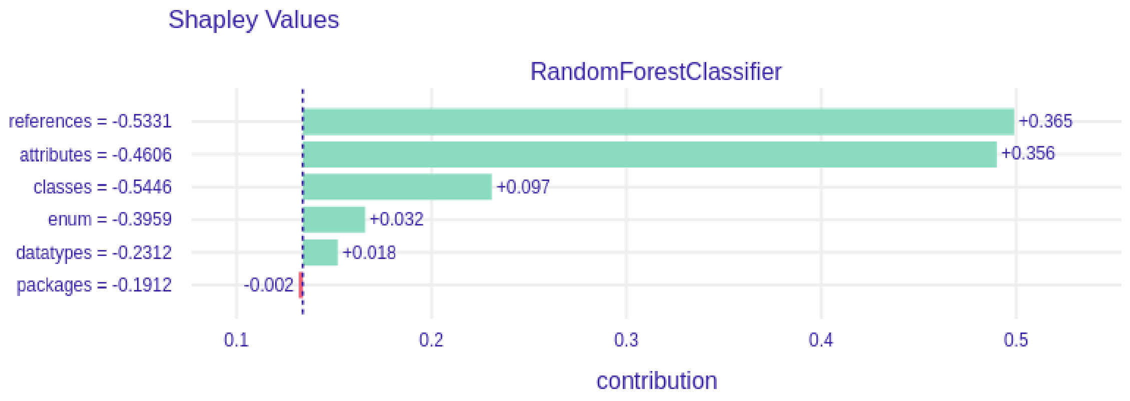
- The colour assigned to the feature indicates if the feature value positively (green) or negatively (red) contributes to the prediction of the model as "dummy".
- The importance assigned to the feature (how long the feature bar is) indicates how relevant the feature value is for predicting that the model is "dummy".

Method 1: LIME



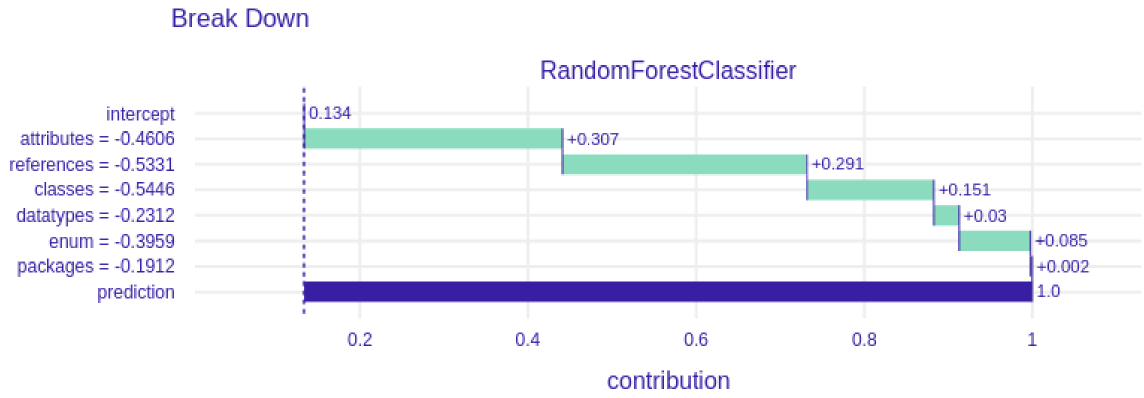
This explanation can be interpreted as follows: *The number of attributes (<=2), classes (<=4), references (<=4), enum types (<=0) and datatypes (<=0) support the prediction that the software model is "dummy". Among these features, the most relevant ones for the decision were the number of attributes, followed by the number or classes and references. In contrast, the number of packages (<=1) suggests that the model should not be predicted as "dummy".*

Method 2: SHAP



The values shown with the feature are the normalized values of the original feature values shown in the table of page 2. This explanation can be interpreted as follows: *The number of references, attributes, classes, enum types and datatypes support the prediction that the software model is "dummy". Among these features, the most relevant ones for the decision were the number of references, followed by the number or attributes and classes. In contrast, the number of packages suggests that the model should not be predicted as "dummy".*

Method 3: BreakDown

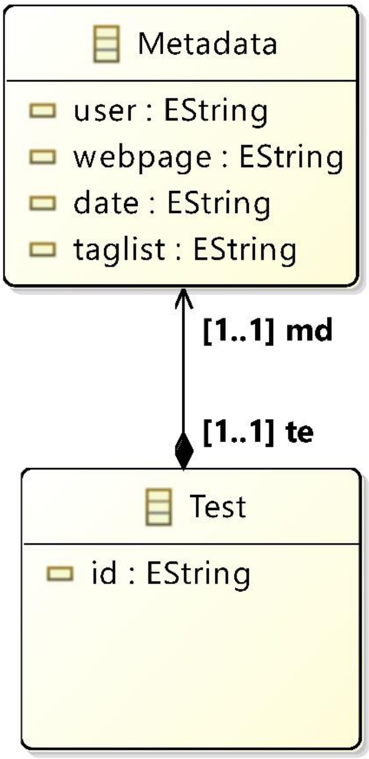


This explanation can be interpreted as follows: *The model is predicted as "dummy" with a 100% probability (contribution is equal to 1). This probability can be attributed to the feature values in the following order (from more to less important): number of attributes, number of references, number of classes, number of datatypes, number of enum types, number of packages.*

[From this point on, start the stopwatch and mark the time in the questions where you see the text "Timespan".]

Task

You will see a different model for which the classifier has made a prediction, together with the visual explanations generated by the three methods (LIME, SHAP, BreakDown). Based on your judgement and the previous example, please answer the questions.



The feature values extracted from the model are:

Reference	Class	Attribute	Package	Enum	Datatype
2	2	5	1	0	0

Question 1: The classifier has predicted that this model is "Not dummy". Do you agree? Mark with "X" inside the brackets.

[] Yes

[] No

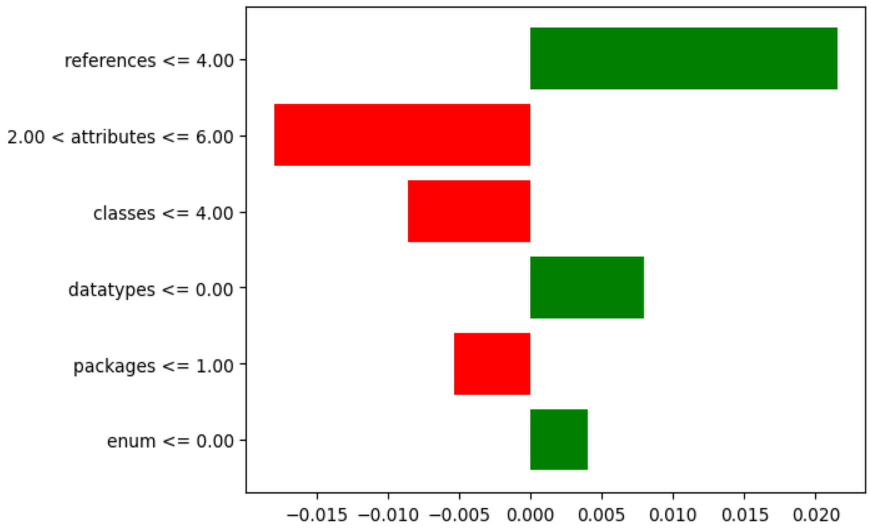
If you disagree, please specify why:

Answer:

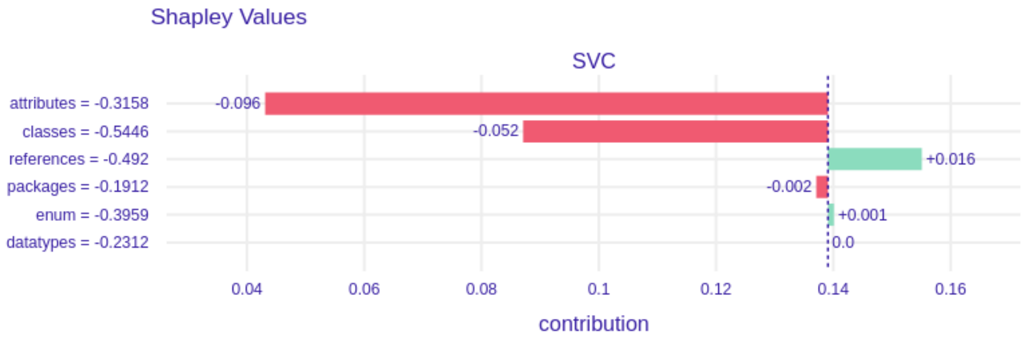
Timestamp: ____________________

Look at the three explanations generated by LIME, SHAP and BreakDown:

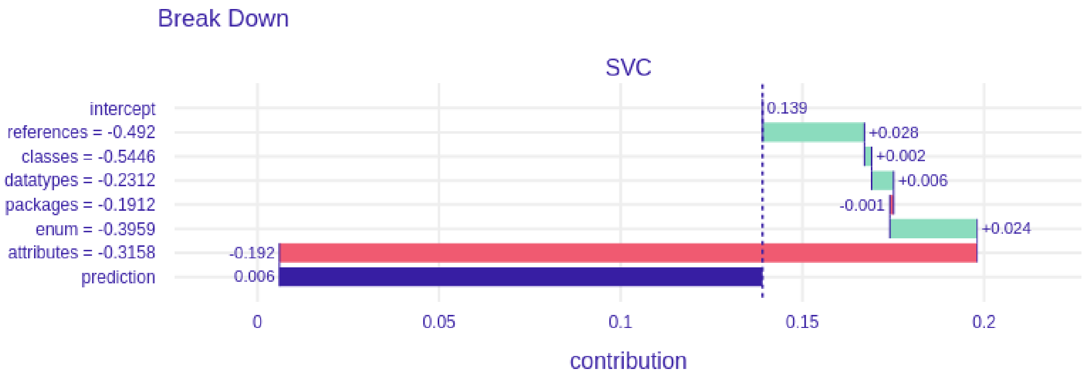
Method 1: LIME



Method 2: SHAP



Method 3: BreakDown



Timestamp: ________________

Please answer the following questions:

Question 2: The method whose explanation I understand more easily is:

[] LIME

[] SHAP

[] BreakDown

Question 3: The method whose explanation I agree with the most is:

[] LIME

[] SHAP

[] BreakDown

Timestamp: ____________________

Question 4: Compare the explanations provided by LIME and SHAP. To what extent do you think these methods agree in their explanations?

[] High disagreement

[] Low disagreement

[] Neutral

[] Some agreement

[] High agreement

Timestamp: ____________________

Question 5: Compare the explanations provided by LIME and BreakDown. To what extent do you think these methods agree in their explanations?

[] High disagreement

[] Low disagreement

[] Neutral

[] Some agreement

[] High agreement

Timestamp: ____________________

Question 6: Compare the explanations provided by SHAP and BreakDown. To what extent do you think these methods agree in their explanations?

[] High disagreement

[] Low disagreement

[] Neutral

[] Some agreement

[] High agreement

Timestamp: ____________________

Question 7: Based on your answers to the three questions above, which aspects of the explanation do you consider most relevant for assessing agreement between two explainable methods?

[] The colour assigned to the features, i.e., feature with positive or negative sign.

[] The feature importance value assigned to the feature.

[] The order of the feature according to the assigned importance value.

Question 8: In general, which aspects of the explanation do you consider most useful to understand the prediction?

[] The colour assigned to the features, i.e., feature with positive or negative sign.

[] The feature importance value assigned to the feature.

[] The order of the feature according to the assigned importance value.

Question 9: Rate from 1 (high disagreement) to 5 (high agreement) your level of agreement regarding the information provided by each explainable method (LIME, SHAP or BreakDown)

	1	**2**	**3**	**4**	**5**
LIME: The information is useful to understand the reasons behind the prediction					
LIME: The explanation is intuitive, i.e., easy to interpret					
LIME: The explanation is complete, i.e., I do not need additional information to understand the prediction					
SHAP: The information is useful to understand the reasons behind the prediction					

SHAP: The explanation is intuitive, i.e., easy to interpret					
SHAP: The explanation is complete, i.e., I do not need additional information to understand the prediction					
BreakDown: The information is useful to understand the reasons behind the prediction					
BreakDown: The explanation is intuitive, i.e., easy to interpret					
BreakDown: The explanation is complete, i.e., I do not need additional information to understand the prediction					
In general, I find this type of explanation useful to understand machine learning predictions in the context of software modelling					
In general, I would have been able to interpret the visual explanations without the text accompanying the explanations of the example					

Question 10: If your answer to the last two items in the table was 1 or 2, please specify what additional aspects you would include to improve the explanations in the context of software modelling.

Free comments [Optional]. Please include any additional comments here. Thank you very much.

Timestamp: ____________________