Delta Capita

ML Interpretability Benchmarking

Huhan Yang

Version v1.0
June 2019

DC MINT

Evaluating Explanations

To take full advantage of outstanding performance from machine learning models, people need to be able to understand why the model is making particular predictions. Therefore, the machine learning models need to be interpretable by humans and they need to be able to evaluate the quality of explanations given.

Evaluating the quality of explanations

Machine learning interpretability has become a hot topic recently and different categories of techniques have been developed over the last few years. Model-agnostic techniques such as LIME and Shapley Values have become popular in this field and many other methods have also been developed for specific model types. Generally, according to their scope, interpretability methods can be split into global or local as well as model-specific or model-agnostic. Each method has its own advantages and disadvantages, but how can we evaluate and benchmark the quality of explanations produced by each method?

Unlike metrics of model performance, such as accuracy, precision, recall etc. there is no clear measure for assessing the quality of interpretability methods. But researchers have attempted to formulate approaches, Doshi-Velez and Kim [1] have proposed three levels:

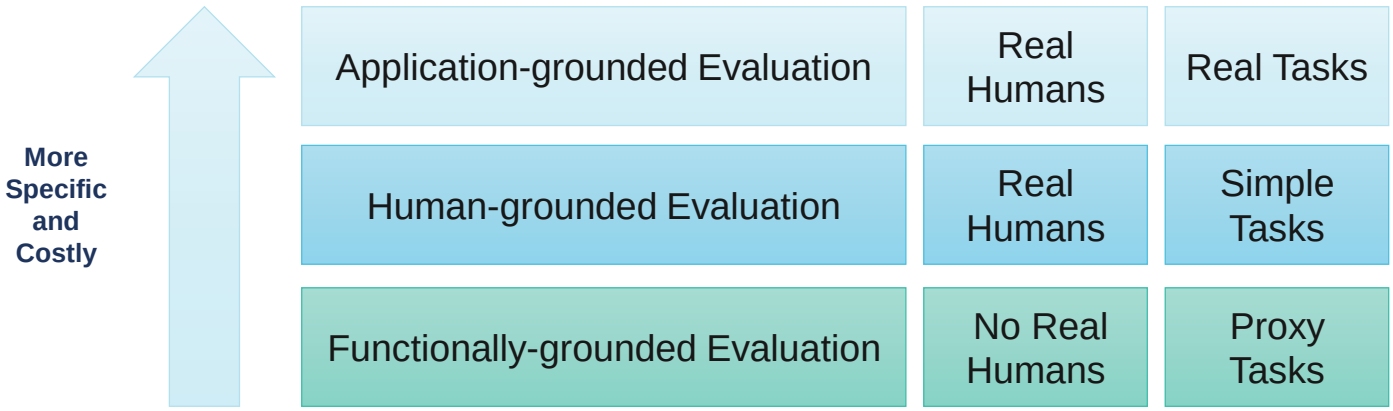
1. Application-grounded: done with expert users reviewing the model and explanations on real tasks;
2. Human-grounded: done with laypersons on real tasks, therefore, cheaper;
3. Functionally-grounded: uses formal definition of interpretability as proxy for explanation quality, no human tasks involved

M. R. Honegger[2] worked further on functionally-grounded evaluation methods and proposed an axiomatic framework which allows comparisons to be made between different interpretability methods to assess the quality of explanations. These three axioms are:

1. Identity: identical instances should have identical explanations
2. Separability: non-identical instances should not have identical explanations
3. Stability: similar instances should have similar explanations

1 Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017).

2 Honegger, M., 2018. Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions.



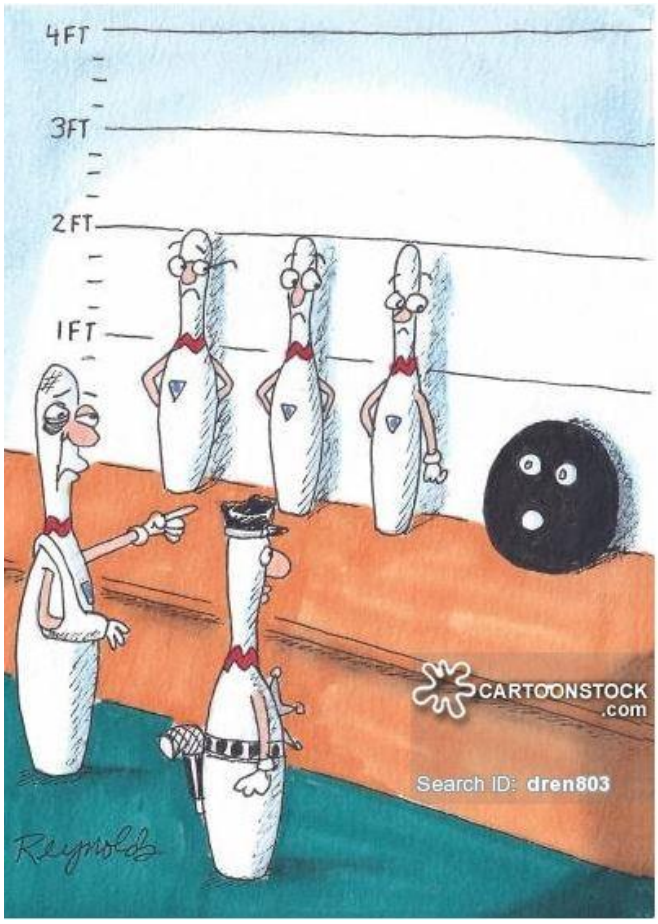
Evaluation Axioms

Axiom 1: Identity

Identity is satisfied if the explanation method is producing identical explanations on identical objects.

Imagine we are trying to explain why a mortgage is predicted to “default”. The first time we run the interpretability method, it tells us the “fico score” contributes the most to the predicted class. The second time we run the interpretability method on the same instance, it tells us the “loan amount” contributes the most to the predicted class. These results will be highly confusing and inconsistent for users, therefore, users will find it difficult to be able to tell which explanation is correct.

This will happen if the interpretability method involves some randomization processes when generating explanations, which is undesired.

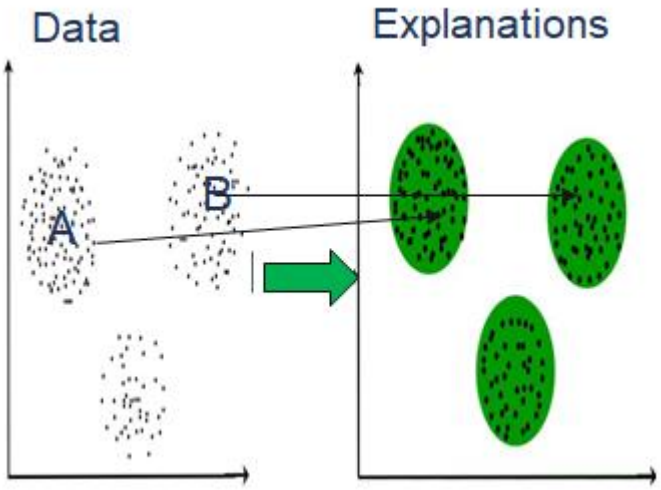


Axiom 2: Separability

Separability states that non-identical instances cannot have identical explanations.

Let us use the mortgage example again and imagine we have two instances; one has been classified as “default” and the other has been classified as “fully paid”. There must be some differences between these two instances, thus, the explanations of these two instances should also be different. Different explanations mean the feature importance for each instance is different. For example, in both cases, the feature “fico score” can be the feature that contributes most to the predicted classes, but in case A, “fico score” may have 51% contribution to its predicted class whereas “fico score” in case B has over 80% contribution.

Nevertheless, there is no reason why two distinct cases cannot be classified as the same class eventhough they have differences regarding their features, thus the explanations should still be non-identical. For instance, case A has a very bad “fico score”, therefore, its “fico score” is the main reason that A has been classified as “default”. In contrast, case C has a fairly good “fico score” but its “loan amount” is much larger, therefore, “loan amount” is the main reason that C has been classified as “default”.

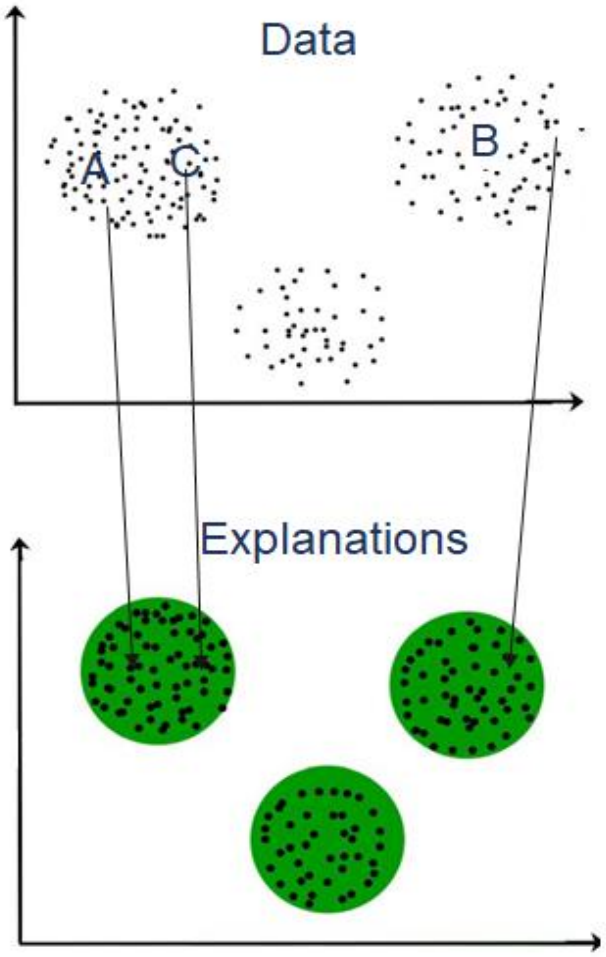


Evaluation Axioms

Axiom 3: Stability

The last axiom, stability, is to assure the explanations are stable, therefore, for slight different instances, their explanations should be similar.

In practice, it is still a bit fuzzy to define what similar explanations mean. To mitigate this, we can use clustering algorithms to evaluate this axiom. Returning to our mortgage example, imagine we have three instances, A, B and C. If we apply clustering on the data, and instances A and C have been assigned to the same group, whereas B has been assigned to another and the distance of these two groups is significantly large, then the corresponding explanations of A and C should be in the same group as well, and the explanation of instance B should be in a distinct group.

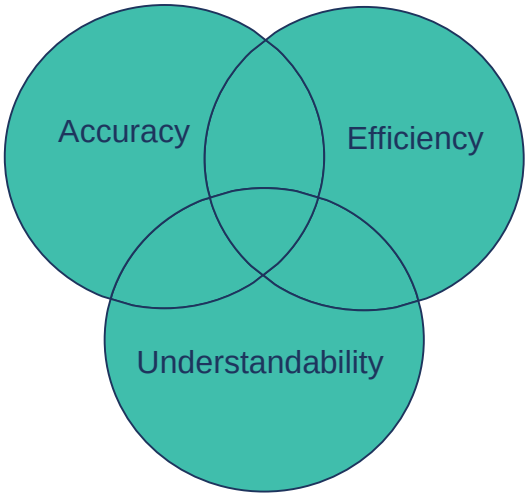


Goals of interpretability

Rüping[3] first proposed that interpretability is composed of three goals:

1. Accuracy: connection between the explanation and the prediction from the ML model;
2. Understandability: the explanation must be understandable by humans;
3. Efficiency: the explanation must be understandable in a finite amount of time

Rüping also affirms that these three goals are often connected and competing:



Conclusion

When using the three axioms to assess an interpretability method, these should be considered “soft” criteria and therefore, it is more meaningful to measure them in terms of the portion of the dataset in which these hold true. For example, for a certain interpretability model, one can say, 90% of cases satisfy axiom 1; 100% satisfy axiom 2.

The three axioms influence each of the three goals of interpretability. While, fulfilling these axioms does not necessarily guarantee that users are able to understand explanations these provide a very good measure of how suitable the method is and its ultimate ability to achieve the three interpretability goals.

3 Ruping, S. (2006). Learning interpretable models (PhD Thesis). Technical University of Dortmund.

DC MINT Evaluations/Benchmarking

DC MINT Benchmarking

Based on the three axioms we discussed in the previous sections, DC MINT has been evaluated/benchmarked on various neural networks with different types of data.

Datasets

The following datasets and models have been used for the evaluation of DC MINT:

- Financial Model – unbalanced (Numerical)
- Financial Model – balanced (Numerical)
- Mortgages (Numerical)
- Risk Model (Timeseries)
- Risk Model (Timeseries)
- Financial Model (Numerical)
- Financial Model (Image)
- Financial Model (Text)

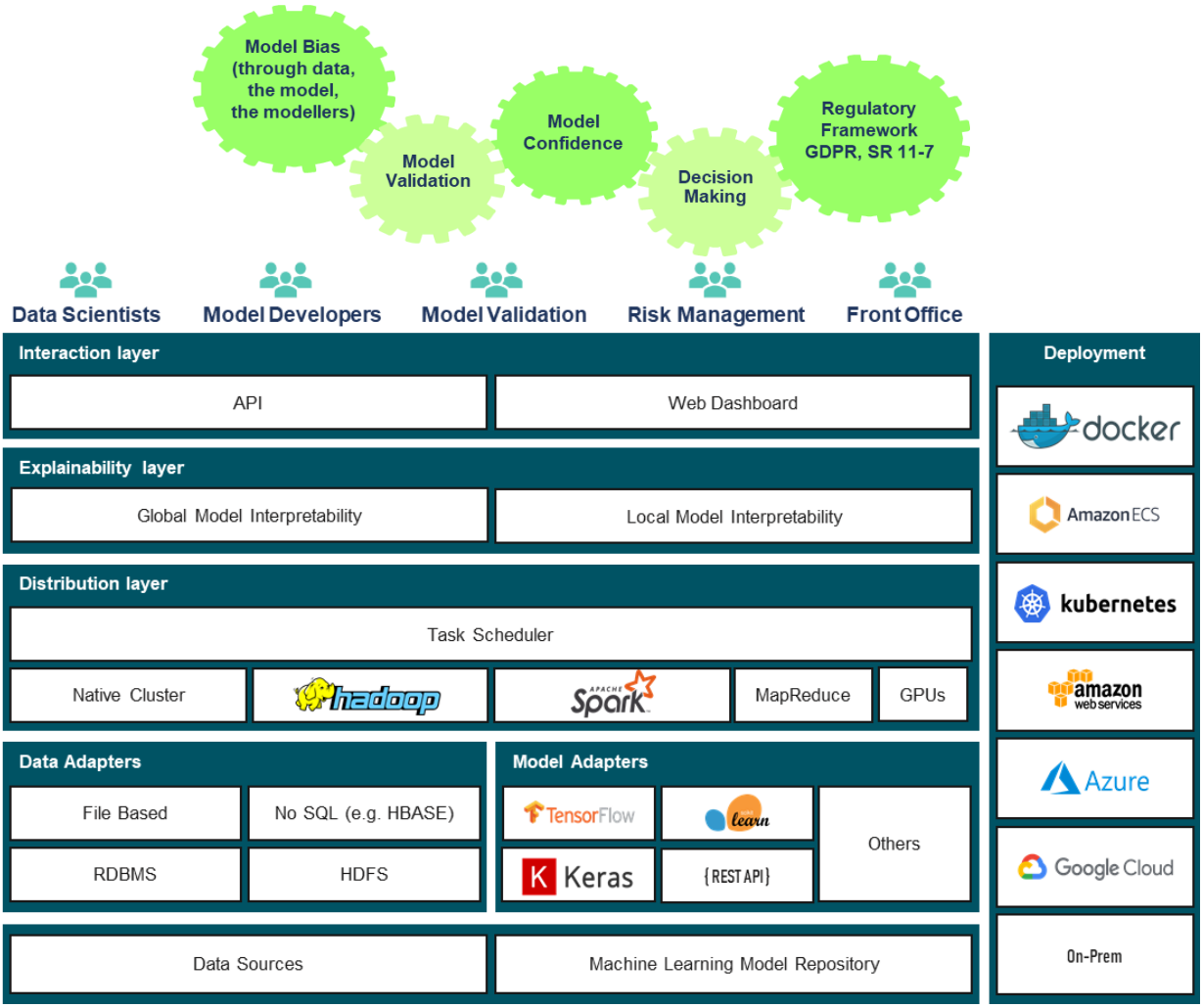
We have used both unbalanced and balanced data in order to test DC MINT's behaviour regarding different situations. The Synthetic Minority Over-sampling Technique (SMOTE) has been applied on unbalanced data.

Results

The results show that DC MINT produces consistent and descriptive explanations across different machine learning tasks, machine learning models and unbalanced/balanced data. Most of the three axioms have been fulfilled with 100% satisfaction. The stability axiom has been affected by the complexity of clustering on the data, hence this can cause slight mismatching between clusters of data points and explanations.

Data	Model	Task	Identity	Separability	Stability
Financial Model – unbalanced (Numerical)	MLP	Classification	100%	100%	99%
Financial Model – balanced (Numerical)	MLP	Classification	100%	100%	99%
Mortgages (Numerical)	MLP	Regression	100%	100%	100%
Risk Model (Timeseries)	LSTM	Classification	100%	100%	100%
Risk Model (Timeseries)	LSTM	Regression	100%	100%	100%
Financial Model (Numerical)	GBM	Regression	100%	100%	100%
Financial Model (Image)	MLP	Classification	100%	100%	100%
Financial Model (Text)	MLP	Classification	100%	100%	100%

DC MINT



Huhan Yang, Data Science Consultant

Data science consultant specializing in machine learning algorithms with experience in financial service, developing machine learning model-driven products and research.

Other contributors:
Sylvia Smit, Managing Partner
Ricardo Cruz, Senior Consultant
Khrystyna Andronova, Senior Consultant

Please contact Delta Capita at *capitalmarketsdelivery@deltacapita.com* if you are interested in any of the following areas:

- **DC MINT - Model Interpretability**
- **DC COMPRESSION - Model Compression and Model Optimisation**
- **DC DOCS - Automated Model Documentation**
- **DC VOICE - Voice/NLP on large data sets and Wealth applications**
- **DC RISK - Contagion Risk Models**

DC MINT