L10_Explainability

Зачем: Why Should I Trust You?

Why Is Model Interpretability so Important?
Machine learning is great in prediction accuracy, process efficiency, and research productivity. But computers usually do not explain their predictions. This becomes a barrier to the adoption of machine learning models. If the users do not trust a model or a prediction, they will not use or deploy it. Therefore the issue is how to help users to trust a model.

There are several great solutions including SHAP, LIME, and ELI5

Their approach is to gain the trust of users for individual predictions and then to trust the model as a whole

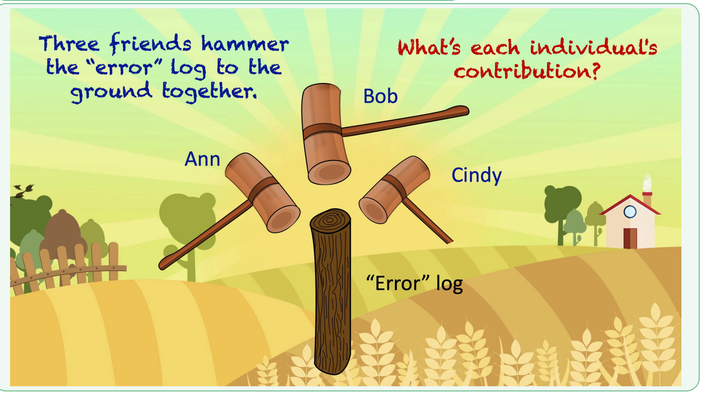
(a) Husky classified as wolf (b) Explanation

Trusting a prediction: a user will trust an individual prediction to act upon. No user wants to accept a model prediction on blind faith, especially if the consequences can be catastrophic.

Trusting a model: the user gains enough trust that the model will behave in reasonable ways when deployed. Although in the modeling stage accuracy metrics (such as AUC — Area under the curve) are used on multiple validation datasets to mimic the real-world data, there often exist significant differences in the real-world data. Besides using the accuracy metrics, we need to test the individual prediction explanations.

How Is LIME Different from SHAP?
In “Explain Your Model with the SHAP Values” I describe extensively how the SHAP (SHapley Additive exPlanations) is distinctly built on the Shapley value. The Shapley value is the average of the marginal contributions across all permutations. The Shapley values consider all possible permutations, thus SHAP is a united approach that provides global and local consistency and interpretability. However, its cost is time — it has to compute all permutations in order to give the results. In contrast, LIME (Local Interpretable Model-agnostic Explanations) builds sparse linear models around an individual prediction in its local vicinity. This is documented in Lundberg and Lee (2016) that LIME is actually a subset of SHAP but lacks the same properties.

The Advantage of LIME over SHAP — SPEED
Readers may ask: “If SHAP is already a united solution, why should we consider LIME?” Remember, the two methods emerge very differently. The advantage of LIME is speed. LIME perturbs data around an individual prediction to build a model, while SHAP has to compute all permutations globally to get local accuracy. Further, the SHAP Python module does not yet have specifically optimized algorithms for all types of algorithms (such as KNNs), as I have documented in “Explain Any Models with the SHAP Values — Use the KernelExplainer” that test models in KNN, SVM, Random Forest, GBM, or the H2O module.

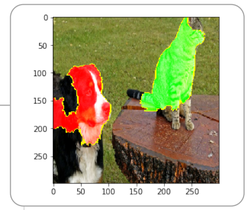


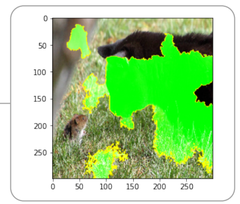
Как

Способы оценки влияния фич

Local Interpretable Model-Agnostic Explanations (LIME)

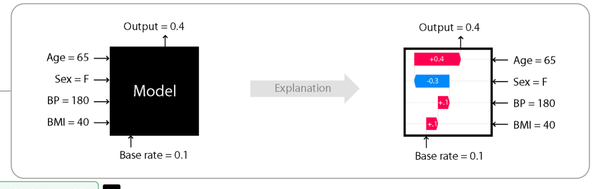
Using Lime with Pytorch





Faces and GradBoost

SHAP (SHapley Additive exPlanations)



Microsoft’s InterpretML

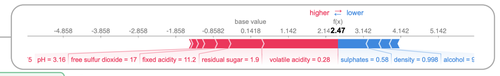
ELI5 - в основном для NLP

InterpretML

Примеры кода

H2)

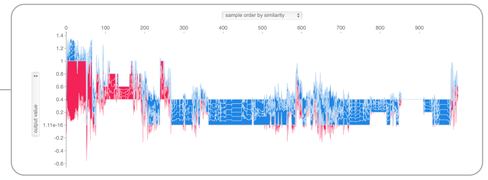
на базе вин



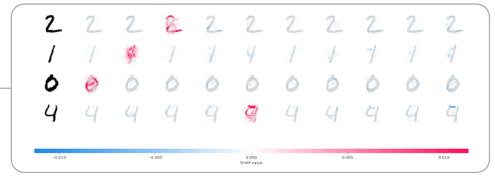
на базе ирисов

NLP

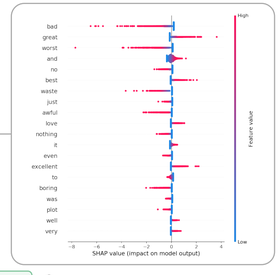
Adult Data Set



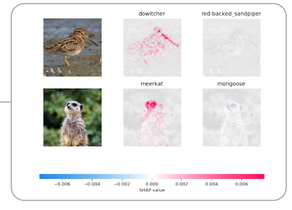
MNIST



Sentiment Analysis



ImageNet



ДЗ

Bonuses

Bonus: Visualizing the Loss Landscape of Neural Nets