DATA SCIENCE MEI/1

University of Beira Interior
Department of Informatics

Hugo Pedro Proença,
hugomcp@di.ubi.pt, **2021/22**

Data Science: Interpretability/Explainability?

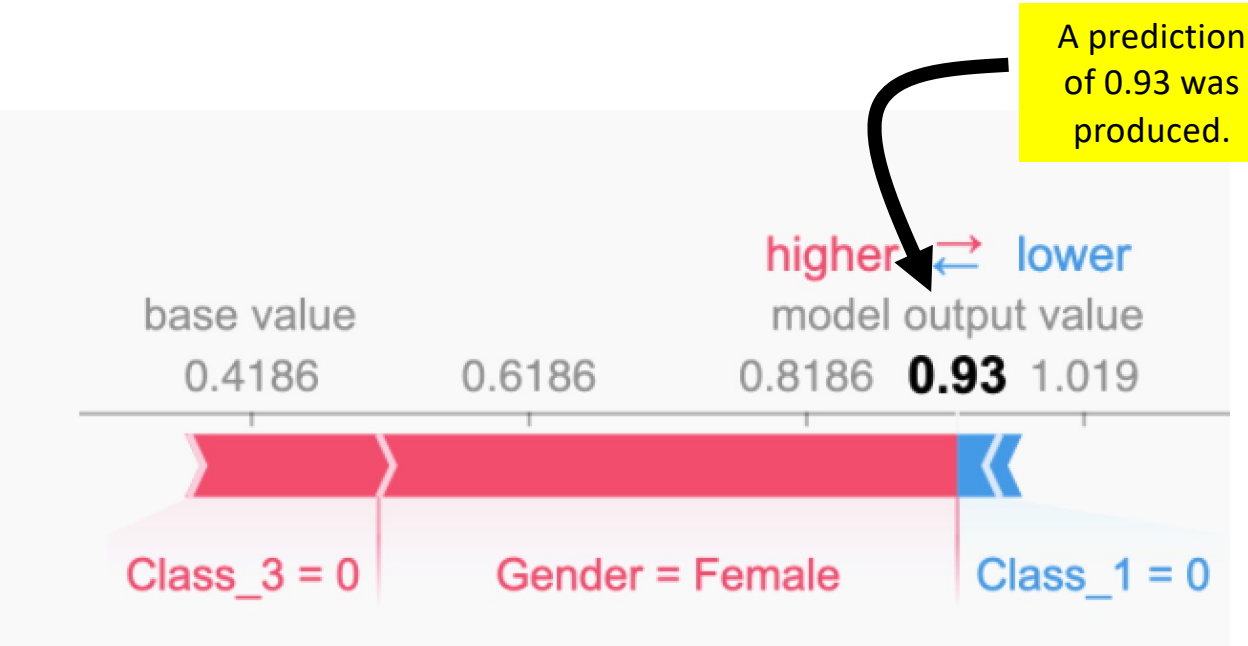
- In the context of Data Science, **explainability** and **interpretability** are often used interchangeably.
- **Interpretability** is about the extent to which a cause and effect can be observed within a system.
  - The extent to which we are able to *predict* what is going to happen, given a change in input or algorithmic parameters.
- **Explainability**, meanwhile, is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
- In summary, **interpretability** is about being able to discern the mechanics without necessarily knowing why. **Explainability** is being able to quite literally explain what is happening.
- Supposing that we are doing a science experiment at school.
  - The experiment might be **interpretable** insofar as you can see what you're doing,
  - However, it is only **explainable** once we dig into the chemistry behind.

Data Science Interpretability/Explainability?

- Automated models can be astonishingly good at making predictions, but they often can't yield explanations for their forecasts in terms that humans can easily understand.
  - The features from which they draw conclusions can be so **numerous**, and their calculations so **complex**, that it is often **impossible to perceive** exactly *why* a model produces a specific answer (explainability).
- Interpretability is also crucial for several reasons. If we don't understand how a model works, it is difficult to **transfer knowledge** into a different domain, for example.
- It is also important for guarding against embedded bias or **debugging an algorithm**. It helps to measure the effects of trade-offs in a model.
  - More broadly, as automated systems play an increasingly important role in society, understanding precisely how they produce their answers will become more critical.

Data Science Interpretability/Explainability?

- Another difficulty is that most of the well-known techniques that interpret/explain the output of a machine-learning system were developed for **tabular data**, i.e., data where the **number of features** is relatively **short**, and have an **obvious semantical meaning**.
  - For example, consider the following SHAP plot, interpreting the model that predicts the survival chances of Titanic passengers:



The SHAP plot illustrates the most important attributes the model:
- If the passenger **was not in third class**, survival chances **increase** substantially;
- If the passenger **was female**: survival chances **increase** even more;
- If the passenger was not in first class: survival chances **fall slightly**.

Explainability/Interpretability Taxonomy

- The existing techniques for Explainability/Interpretability are divided in terms of:
  - **Depth**
    - This is related to the length to which we explain a given model, i.e., whether the technique constraints the model's complexity to make it more transparent (***intrinsic* explainability**) or allows complexity and focuses on explaining exclusively the system outputs (***post hoc* explainability**).
  - **Scope**
    - Indicates the range that a technique possesses, i.e., if it explains individual(instance) predictions (***local***) or the whole model's behaviour (***global***).
  - **Applicability**
    - Divides the techniques based on their model affinity, i.e., whether they are only compatible with a specific family of models (***model-specific***) or any kind of model (***model-agnostic***).

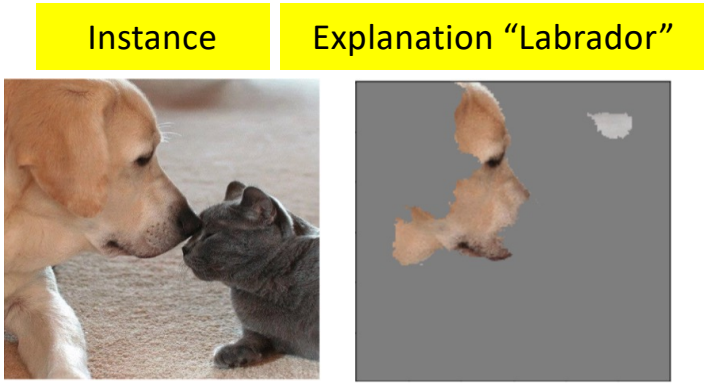
C. Molnar. Interpretable machine learning. A guide for making black box models explainable, 2019. https://christophm.github.io/interpretable-ml-book/

Explainability/Interpretability: Typical Techniques

- The three most common techniques include **LIME**, **Shapley values** (SHAP) and **Saliency Maps**.
  - **LIME** (Local interpretable model-agnostic explanations) use a surrogate linear model, trained on perturbed data (e.g., corrupted clusters of adjacent pixels), to locally approximate the behaviour of a complex black-box model.
  - **Shapley values** are based in game theory. Shapley values are assigned to the features based on how important they are to a given prediction.
  - **Saliency maps** use the derivative of a highly complex function (essentially, a CNN) with respect to a given input image, to determine which pixels need to be changed the least, while also changing the output class the most.
- Finally, other methods were proposed exclusively for visualisation purposes (e.g., **PDP** and **ALE**), which produce plots that correlate the independent

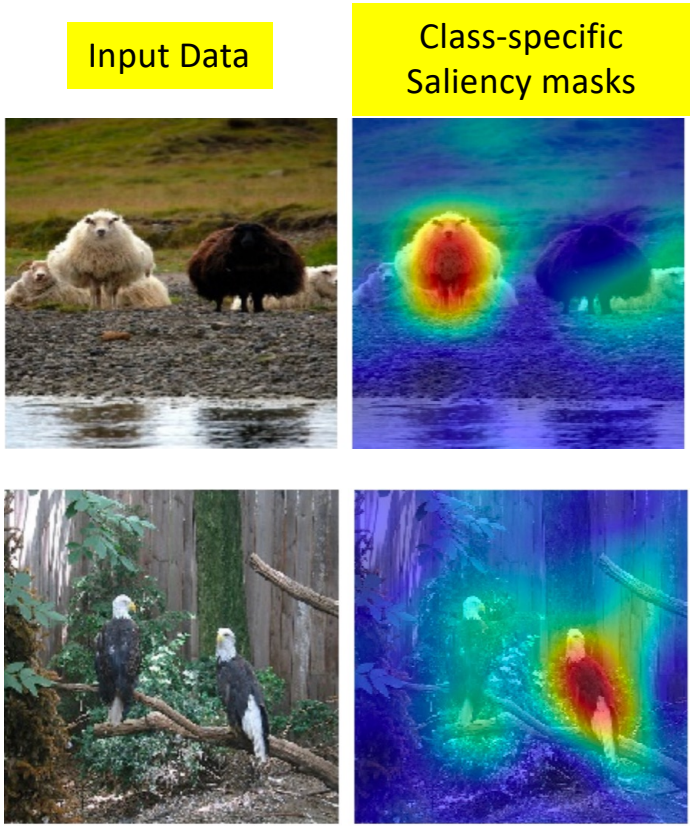
LIME

- Local interpretable model-agnostic explanations (LIME) (Ribeiro, M.T., Singh, S. and Guestrin, C.) is a method for fitting local, interpretable models that can **explain single predictions**.
- LIME explanations surrogate models, i.e., models (like a linear model or decision tree) that are learned on the predictions of the **original black box** model.
- **Algorithm**:
  - Choose one instance of interest for which we want to have an explanation of its black box prediction.
  - Perturb the data and get the black box predictions for these new points (e.g., super pixels).
  - Fit a novel weighted, interpretable model on the dataset with the variations (e.g., linear regression).
  - The explanation is given by the super-pixels with the largest weights.



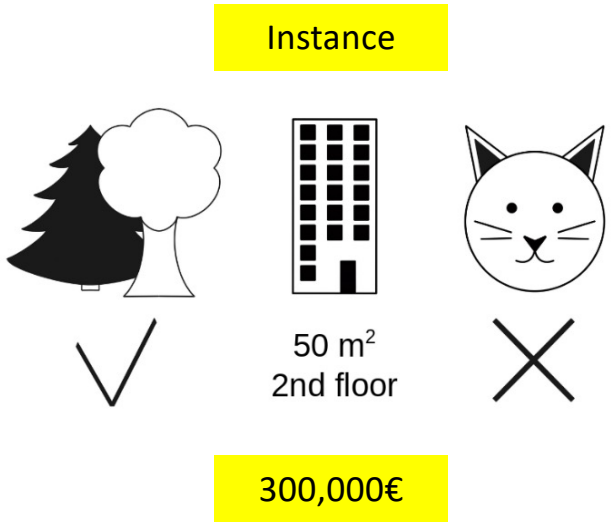
Saliency Maps

- In essence, saliency is what "**stands out**" in an image, enabling to focus on the most important regions.
- There are three main families of visual saliency methods:
  - **Static saliency:** This class of algorithms relies on image features and statistics to localize the most interesting regions of an image.
  - **Motion saliency:** Algorithms in this class typically rely on video or frame-by-frame inputs. The motion saliency algorithms process the frames, keeping track of objects that move.
  - **Objectness:** These algorithms that compute "objectness" generate a set of "proposals", or more simply bounding boxes of where it thinks an object/foreground may lie in an image.



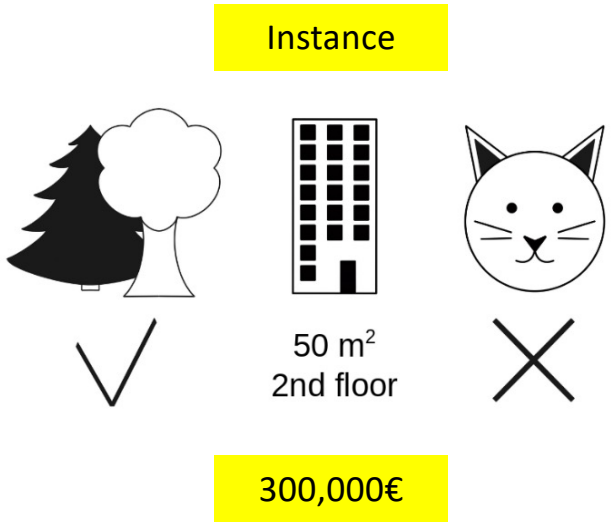
Shapley Values

- Are particularly suited for tabular data.
- Upon a **particular instance** and **feature** that we want to “explain”, the idea is to compare the responses provided by the system with all possible coalitions with/without the feature value.
  - Values outside the coalition are draw randomly.
- By **accumulating the differences** between the system responses when the feature **has/has not a particular value**, we obtain information about how important is that feature value.



Shapley Values: Example

- Suppose **we want to evaluate the contribution of the ”cat-banned” feature** value when it is added to a coalition of ”park-nearby” and ”area-50”.
- We simulate that only ”park-nearby”, ”cat-banned” and ”area-50” are in a coalition by randomly drawing another apartment from the data and using its value for the floor feature.
- The value floor-2nd was replaced by the randomly drawn floor-1st. Then we predict the price of the apartment with this combination (€310,000).
- In a second step, we remove ”cat-banned” from the coalition by replacing it with a random value of the cat allowed/banned feature from the randomly drawn apartment. In the example it was cat-allowed, but it could have been cat-banned again. We predict the apartment price for the coalition of ”park-nearby” and ”area-50” (€320,000).
- The contribution of cat-banned was €310,000 - €320,000 = -€10,000.



Shapley Values: Coallitions

- In the previous exemple, we cannot be certain that the observed difference (-10.000€ was due to the “cat banned” feature, or to any other of the values used in that particular case (multi-dimensional analysis).
- This can be measured by obtaining all the possible coallitions, with 0, 1, ...N features
- The final Shapley value yields from the average of the model responses in all coallitions.