Artificial Intelligence Seminar

UNIVERSITAT ROVIRA i VIRGILI

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAT DE BARCELONA

SHAP & Shapley Values

Explainable AI

Presented by : Hajar Lachheb & Clea Han

Table of contents

01

AI & Importance of Interpretability

02

Overview of SHAP and Shapley Values

03

SHAP Values and Feature Importance

04

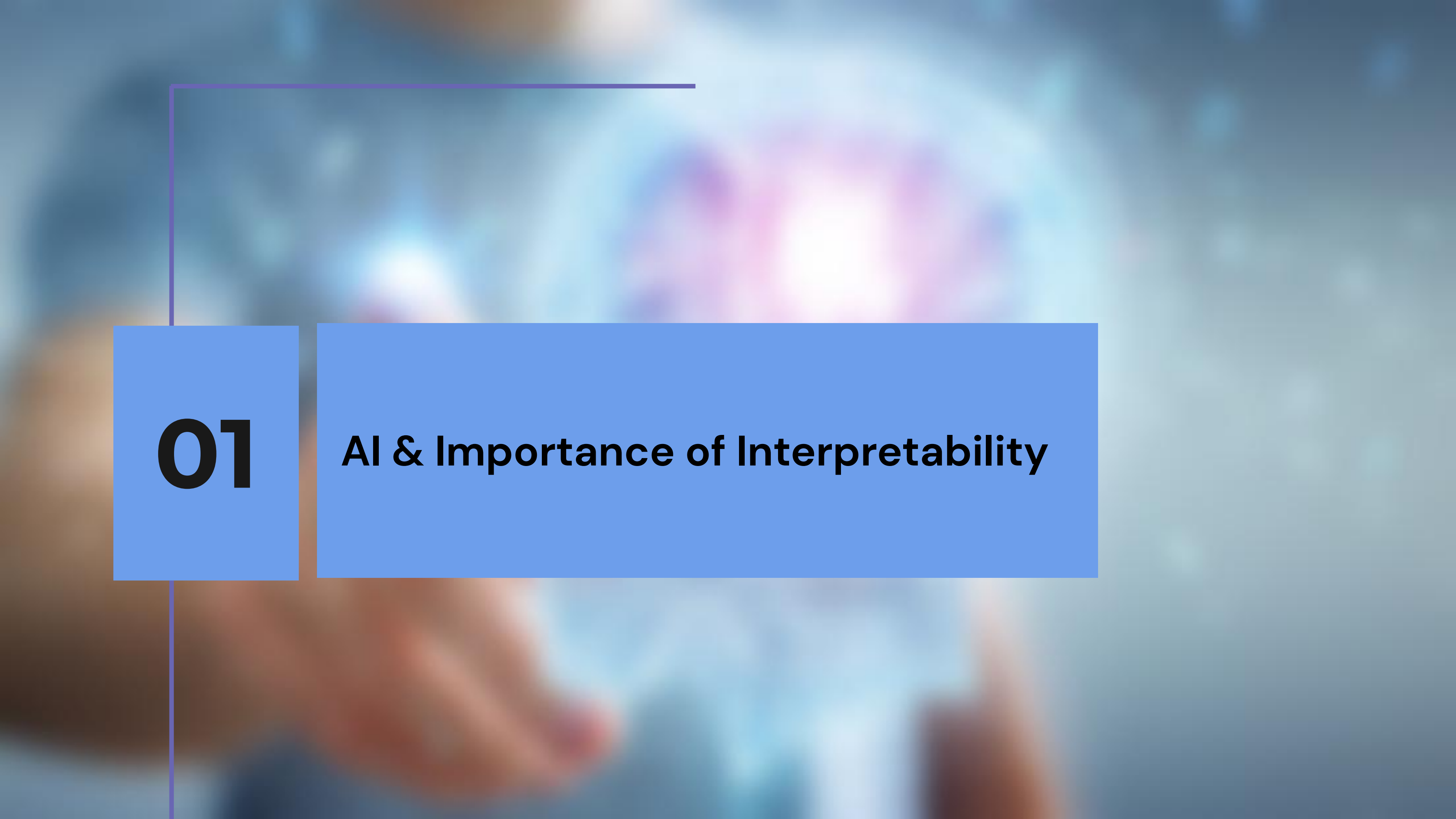
SHAP Values and Model Explanation

05

Challenges & Limitations

06

Conclusion & Future Directions



01

AI & Importance of Interpretability

AI & Importance of Interpretability

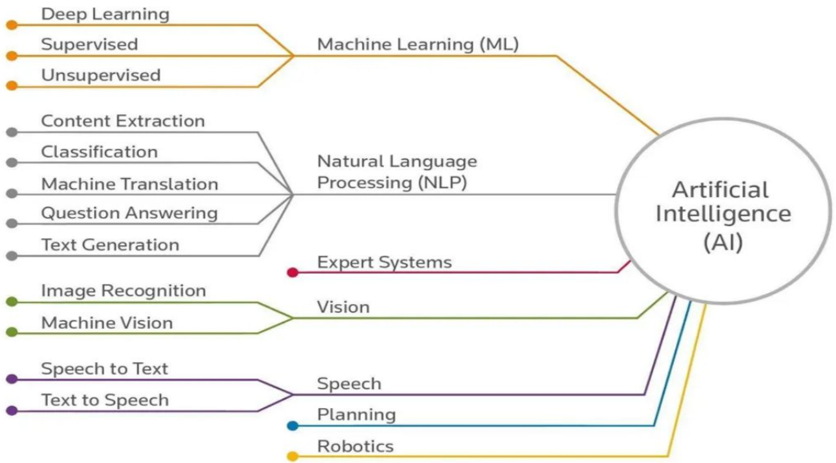
What is Artificial Intelligence ? The power of a machine to copy intelligent human behavior, right ?

AI te refers to the ability of machines to perform tasks that would typically require human intelligence, such as perception, reasoning, learning, problem-solving, and decision-making.

In recent years, there has been a rapid expansion in the use of AI and machine learning in various industries.

While these technologies offer many benefits, they can also pose significant challenges in terms of interpretability.

<u>*Especially when AI is making decisions that have significant consequences for individuals or society as a whole.</u>



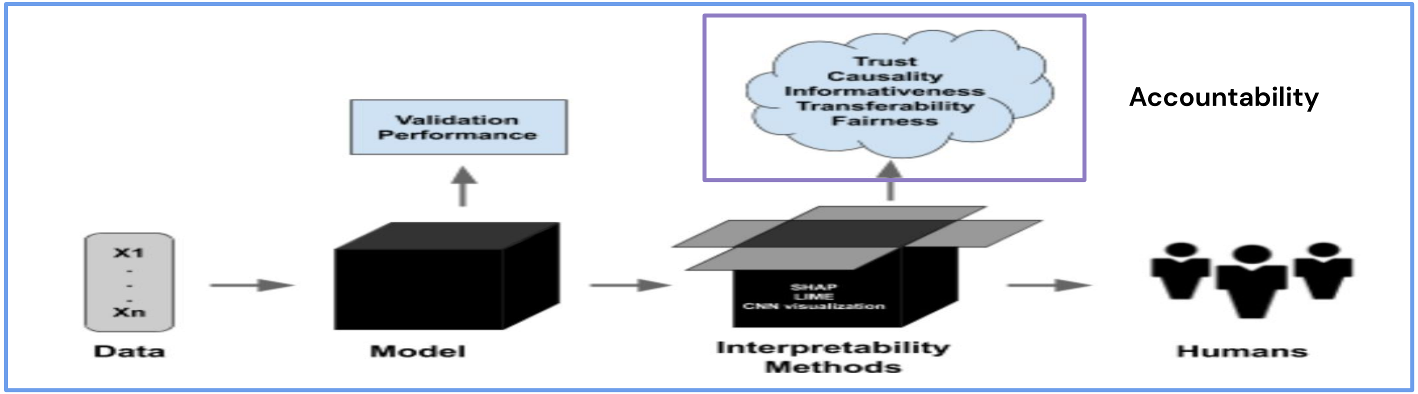
Notion of Interpretability

We say that something is interpretable if it is capable of being understood. With that in mind, we say a model is interpretable if it is capable of being understood by humans on its own.

But what about interpretability in Machine Learning ?

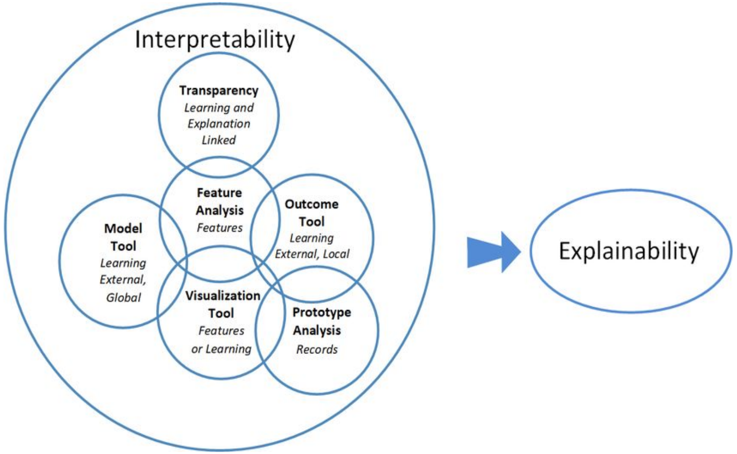
Interpretability allows us to understand what a model is learning, the other information it has to offer, and the reasons behind its judgments in light of the real-world issue we're attempting to address.

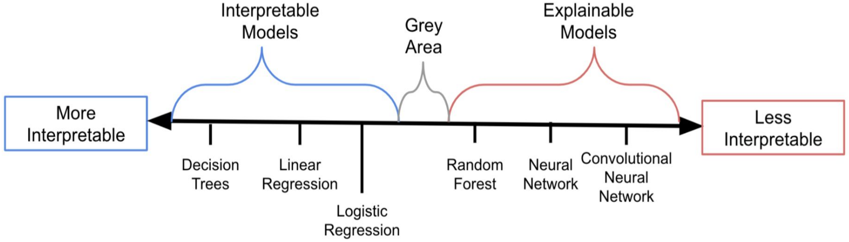
*When model metrics are insufficient, interpretability is required.



Explainable AI

Explainability and Interpretability



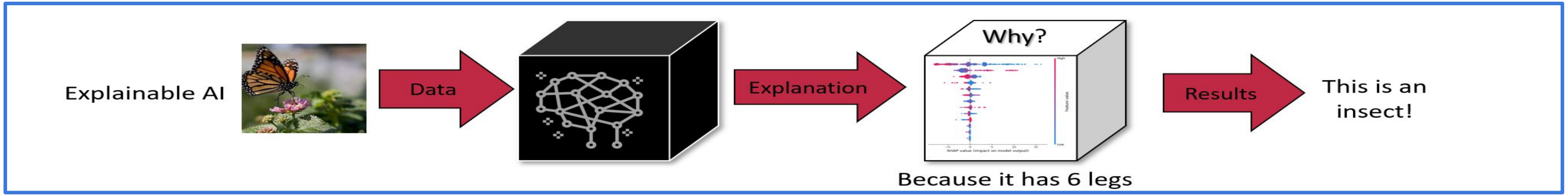


Interpretability focuses on **understanding** how a machine learning **model or system works**, explainability focuses on providing **clear and understandable explanations** for **its decisions or predictions.**

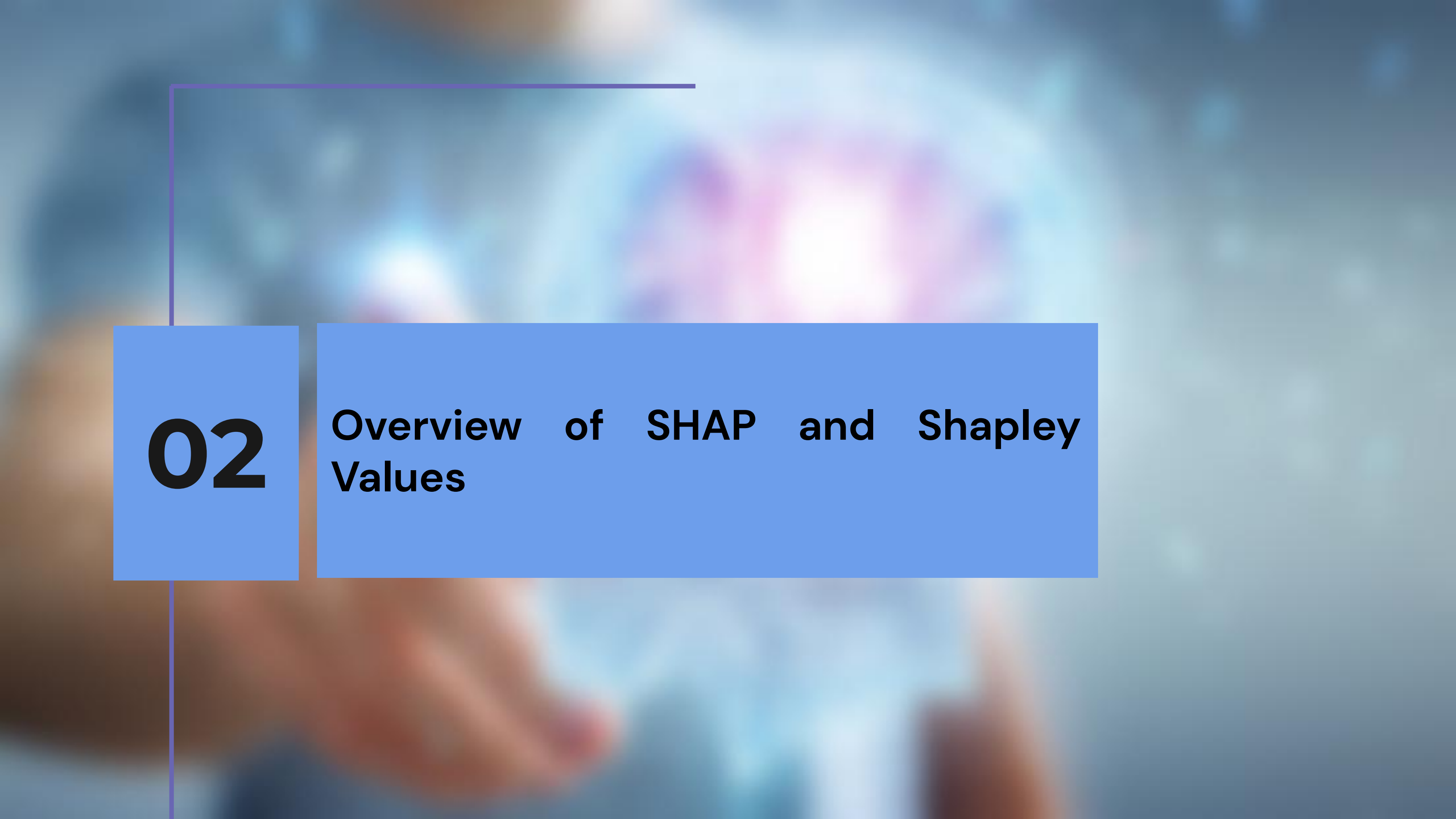
Both concepts are important for building **trust** and **transparency** in AI systems and ensuring that they are making decisions that are **fair, unbiased**, and **aligned** with ethical principles.

Why is it important ?

As scientists in AI, how would we feel if the scenario was always like this ?



- It is desired in use cases involving **accountability.**
- It is critical for situations involving **fairness** and **transparency** where there are scenarios with sensitive information or data associated with it.
- Enhanced **trust** between humans and machines.
- Higher **visibility** into model decision-making process and **transparency**.



02

Overview of SHAP and Shapley Values

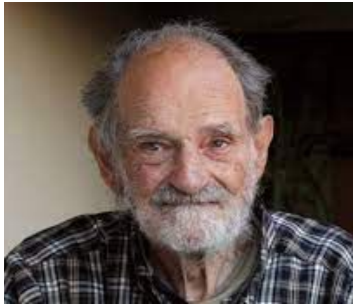
Brief History

Scott Lundberg, Microsoft Research AI team

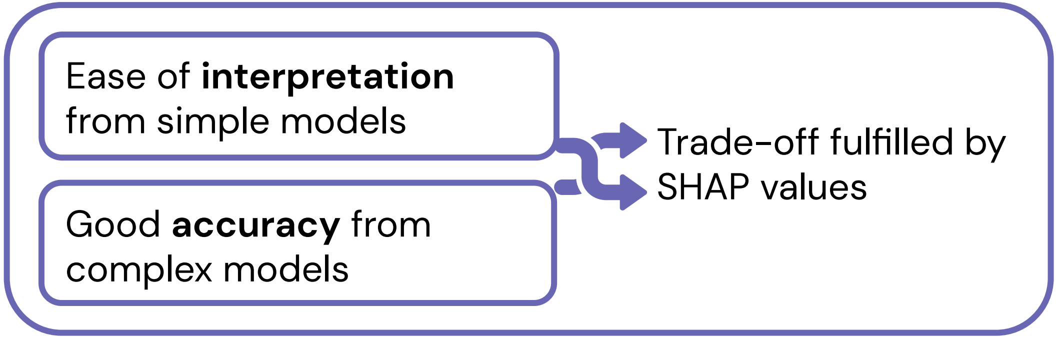
SHAP → **SH**apley **A**dditive ex**P**lanations

➢ **Explain Machine Learning models**

Cooperative game theory – How to make a fair payout in an interactive decision making process



Lloyd Shapley, 2012 Nobel Prize for Economics



Ease of **interpretation** from simple models

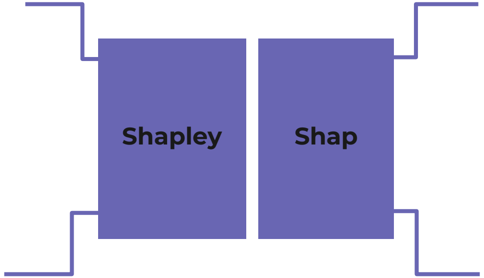
Good **accuracy** from complex models

Trade-off fulfilled by SHAP values

From Shapley values to SHAP values

Knowing the contribution of each player for fair rewarding in a game

One cooperative game with players interacting with each other



Quantifying the contributions of each feature relative to one prediction of a ML model

A model with different features that has effect on each other

SHAP values compute the **marginal contribution of a feature by averaging all its contributions in every kind of situation**

Interactions between players / features

Order of consideration of players / features

Computation of SHAP values

➢ **NP-hard** problem → Approximation algorithms

Model-agnostic approximation :

- Kernel SHAP (Linear LIME + Shapley values)

Model-specific approximation :

- Linear SHAP : for linear models
- Low-order SHAP : neglect interactions between features
- Max SHAP : sampling-based approximation
- Deep SHAP : for Deep Neural Networks
- Tree SHAP : for Tree-based models

Advantages of SHAP Values

Comparable : SHAP values are in the unit of the features

Local and Global explanations

➔ Unlike LIME (Local Interpretable Model-agnostic Explanations)

Global interpretability

➔ Unlike Anchors : do not capture the complexity of the model like SHAP values

Generate counterfactual explanations

⚠ SHAP values depend on the model itself, so sensible to underfitting or overfitting could lead to poor quality interpretations

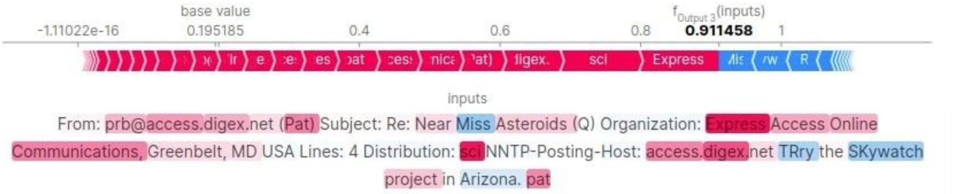
Applications of SHAP values

For model’s decision-making process

- Feature importance
- Feature engineering
- Model explanations :
  - **Debugging**
  - **Monitoring**
  - **Identifying biases**
- Building trust

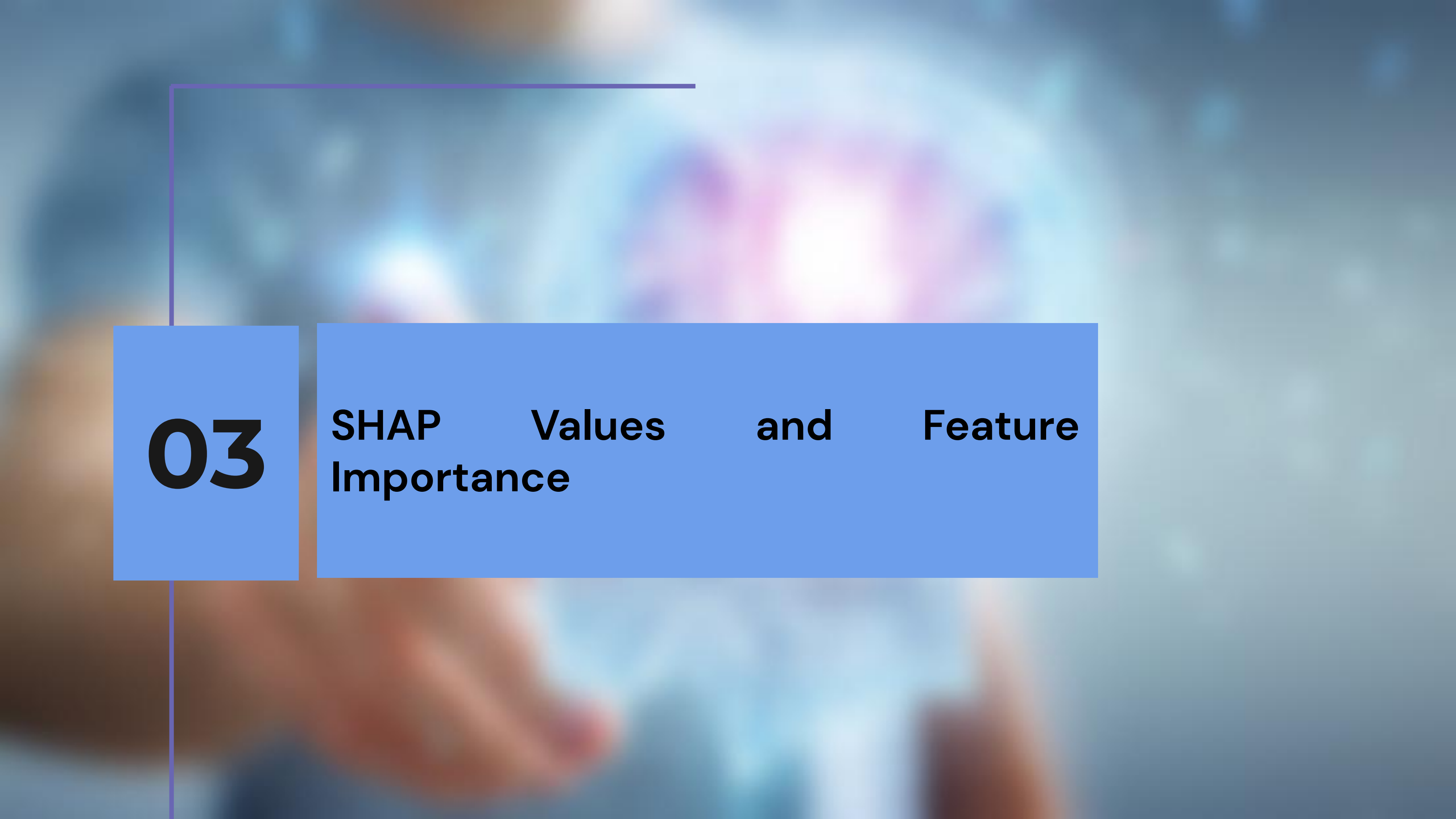
- In Finance : deciding whether or not to grand a loan for a bank client

- In Natural Language Processing



From: prb@access.digex.net (Pat) Subject: Re: Near Miss Asteroids (Q) Organization: Express Access Online Communications, Greenbelt, MD USA Lines: 4 Distribution: sci NNTP-Posting-Host: access.digex.net TRry the SKywatch project in Arizona. pat

- Explain the behavior of a metric, like the loss, to tackle the lack of accuracy



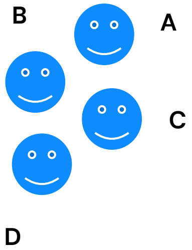
03

SHAP Values and Feature Importance

Shapley values and features contribution

To understand better where did the SHAP works when it comes to Machine Learning and when it comes to features importance, we can take into consideration this exemple.

Let’s imagine there is a group of students working together to win a coding competition.



We want to distribute the credits to each member of the team taking into account how much he contributed in the competition. We can take into consideration the amount of bugs fixed.

- Bugs fixed by just A: 20
- Bugs fixed by just B: 15
- Bugs fixed by just C: 10
- Bugs fixed by just D: 12
- Bugs fixed by A and B together: 30
- Bugs fixed by A and C together: 25
- Bugs fixed by A and D together: 28
- Bugs fixed by B and C together: 22
- Bugs fixed by B and D together: 24
- Bugs fixed by C and D together: 18
- Bugs fixed by the entire team: 35

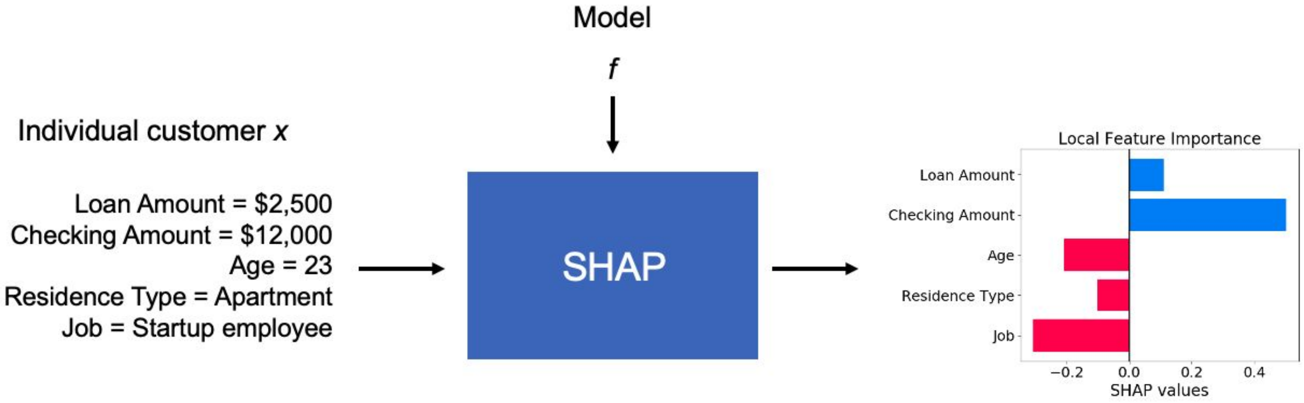
*To calculate the Shapley value for each team member, we consider their contribution when they are part of different subsets. The Shapley value represents the average marginal contribution of a player across all possible team combinations. For example, to calculate our Shapley value, we would consider the individual contribution of A (20 bugs fixed) and the contribution of A when paired with Alice (30 bugs fixed) and Bob (25 bugs fixed). The average of these contributions represents the Shapley value of A.

SHAP then will help us understand the contribution of each feature to a specific prediction in a machine learning model, taking into account feature interactions and the influence of different feature combinations.

Shap Values and Features

Is there any kind of link between SHAP Values and what we call feature importance ?

As mentioned earlier, SHAP values are a unified measure of feature importance that originated from cooperative game theory and have been adapted for machine learning interpretability.



Model

f

Individual customer *x*

Loan Amount = $2,500
Checking Amount = $12,000
Age = 23
Residence Type = Apartment
Job = Startup employee

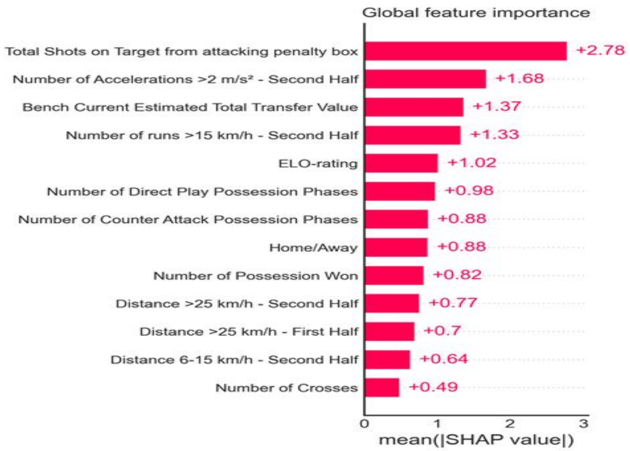
SHAP

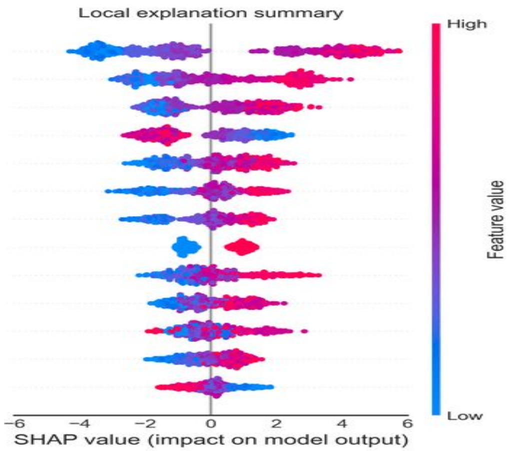
SHAP assigns each feature a value that represents whether it pushes the prediction f(x) higher or lower. For the customer, the loan amount and the checking amount push the probability of repayment higher, while his age, residence type and job push it lower.

Feature Importance

Feature importance refers to quantifying the relevance or significance of each feature in a machine learning model independently.

How does it work exactly ?



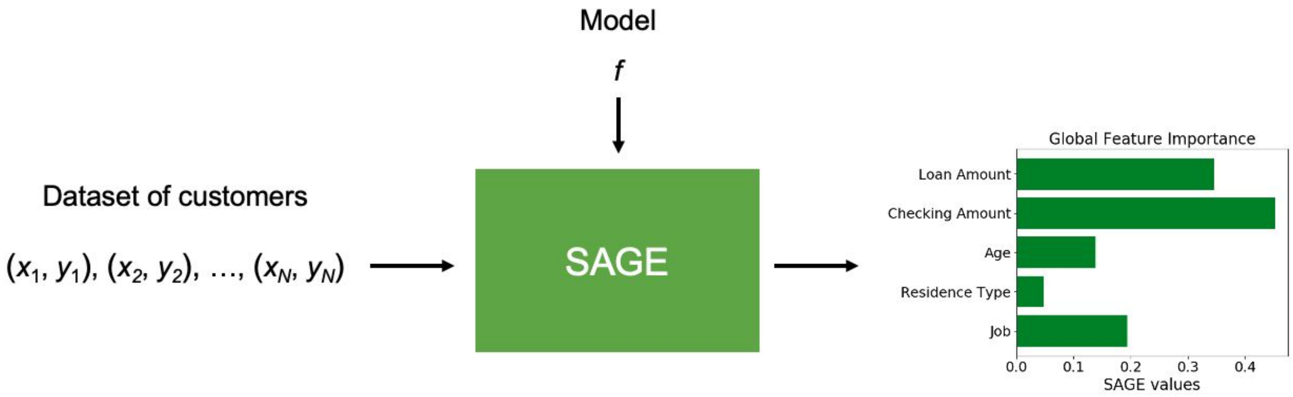


By analyzing SHAP values, we gain insights into how individual features contribute to predictions, accounting for both main effects and interactions. This understanding helps us identify which features are driving specific predictions and identify the most influential features and prioritize them for further analysis or feature engineering.

Local vs Global interpretability

As mentioned, analyzing SHAP values, we gain insights into how individual features contribute to predictions. We used an individual X as an exemple.

What about global interpretability ?



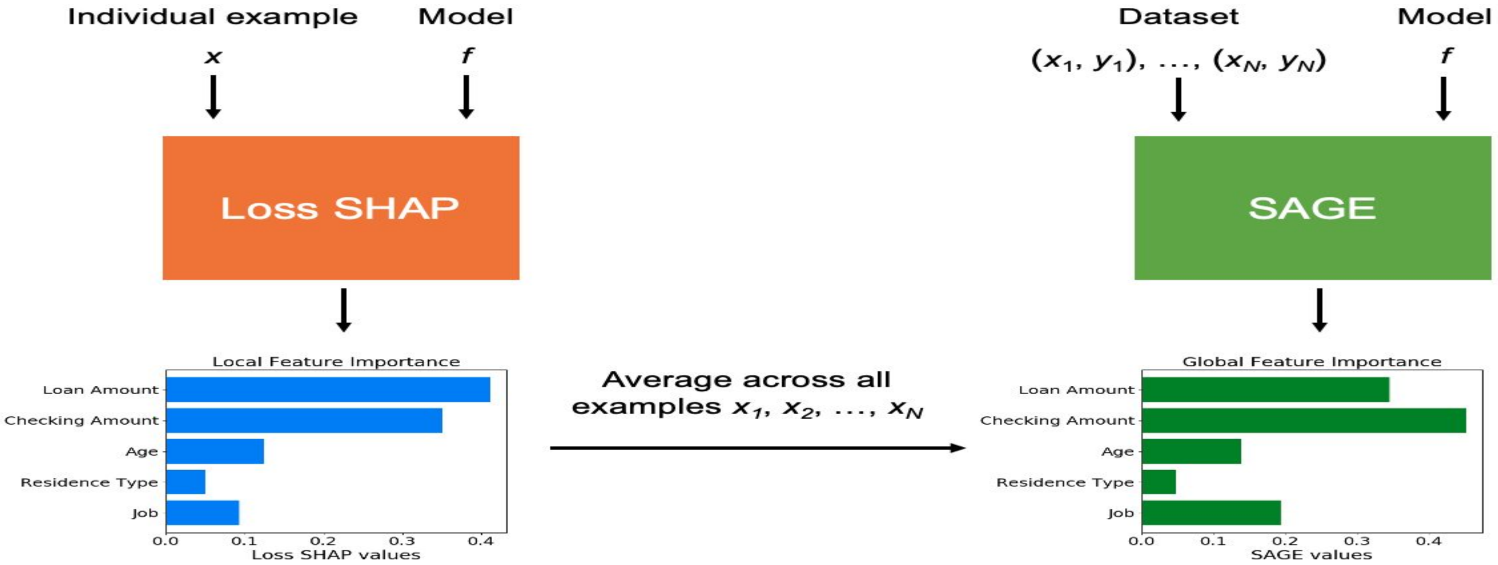
SAGE(Shapley Additive Global importancE) assigns each feature a value that represent how much the feature contributes to f's performance. The most important features have the highest values.

- ➔ **SHAP answers the question how much does each feature contribute to this individual prediction?**
- ➔ **SAGE answers the question how much does the model depend on each feature overall?**

SHAP and SAGE, what relation ?

SHAP and SAGE both use Shapley values, but since they answer fundamentally different questions about a model (local versus global interpretability).

Can we obtain global interpretability from local one ?



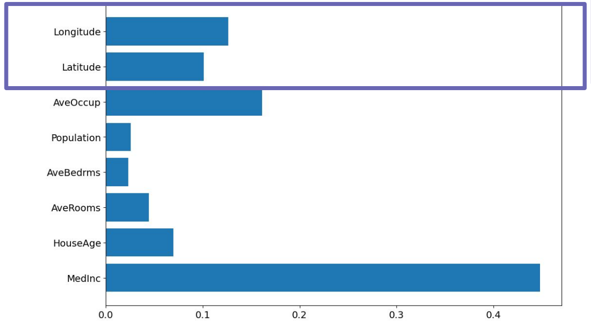
A global explanation can be obtained via many local explanations.

Implementation example

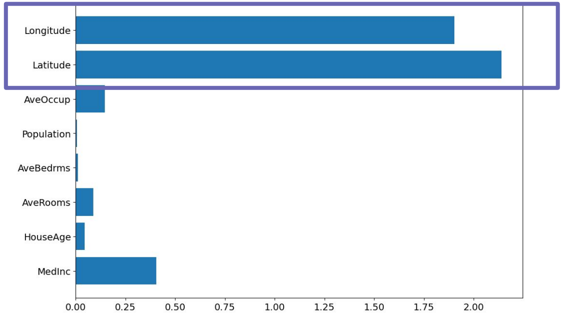
For the implementation example, we are using XGBoost and our California Housing dataset. XGBoost is used in our case since it's an implementation of gradient-boosting decision trees and it's compatible with SHAP Values library.

The results of the three methods used.

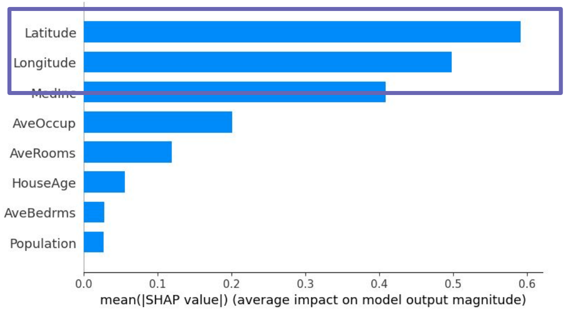
XGBoost Built in function



Permutation-based importance



SHAP Method

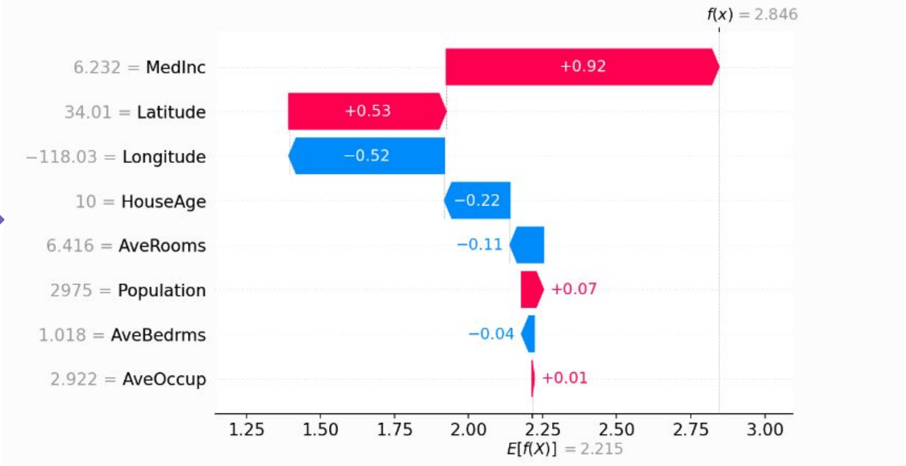


The results are indeed different from a method to an another. And this is due to the differences in their underlying principles and calculations.

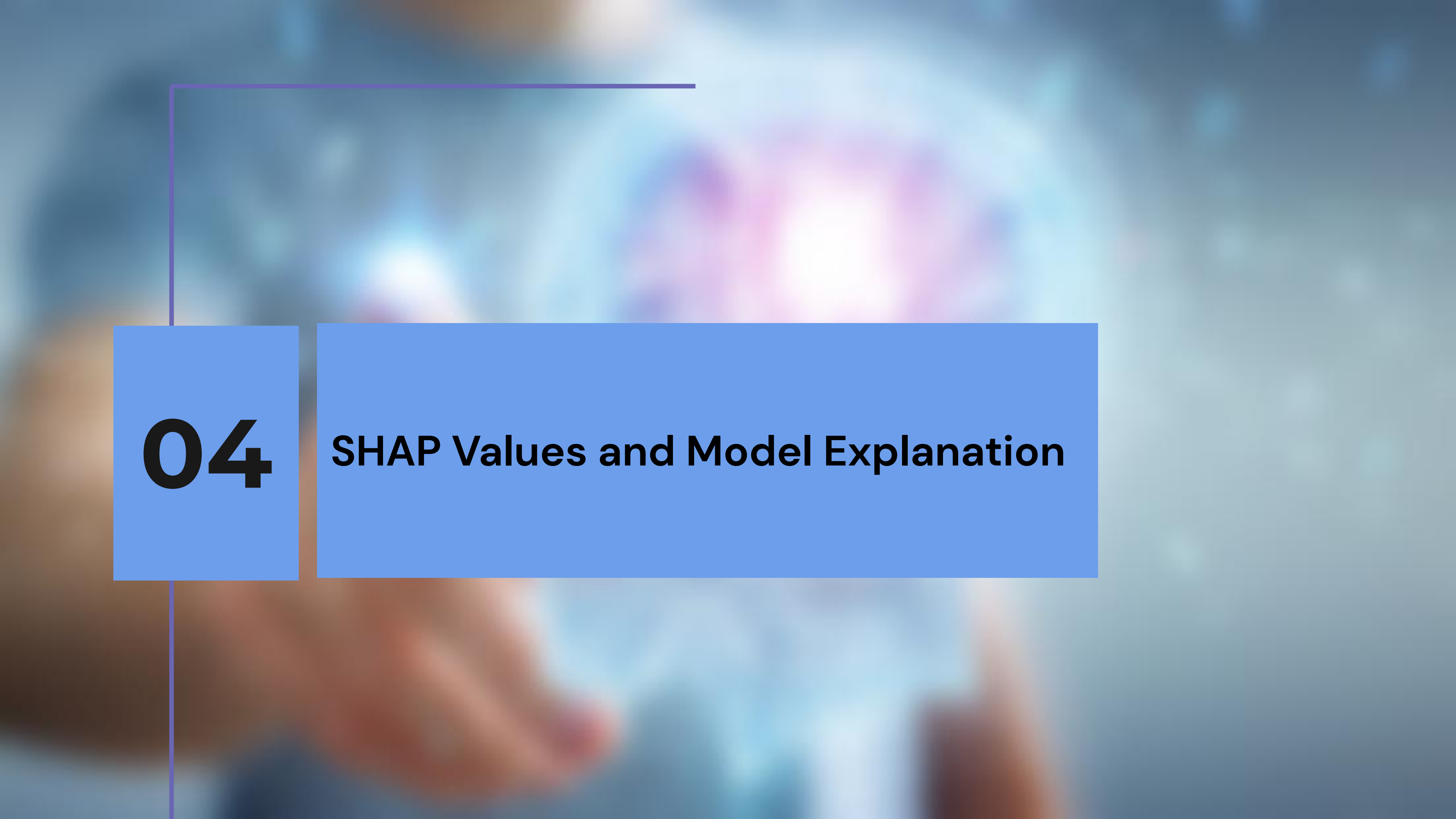
Additive nature of Shapley values

To observe better the waterfall of features, we can plot the following figure. In fact, the **sum of the SHAP values** (which represent the contributions of each feature) of **all the input features** will always add up to the **difference between the expected output** of the model when **no features are considered (baseline)** and the **current output** of the model for **the specific prediction we're trying to explain.**

We start at the baseline expectation (such as the **average home price**) and then adds features one by one, showing how each feature contributes to the final prediction made by the model.



Since the shapley value of the feature HouseAge is negative, we can conclude that having an old house tends to decrease the predicted price compared to the baseline expectation.

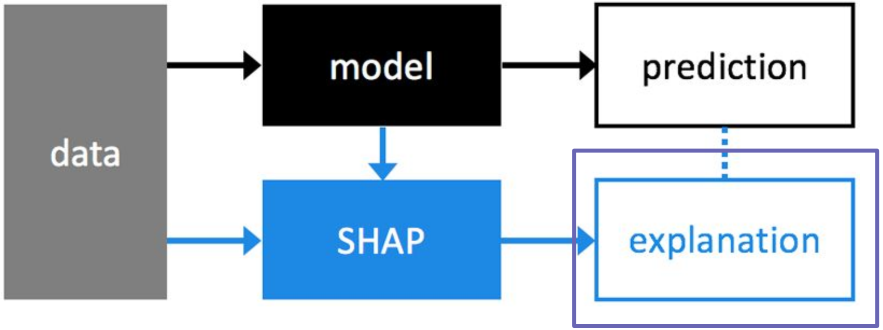


04

SHAP Values and Model Explanation

Model Explanation

Model explanation refers to the process of understanding and providing insights into how a machine learning model arrives at its predictions or decisions. It aims to shed light on the inner workings of the model and make its decision-making process transparent and interpretable to humans.



Model explanation techniques can provide insights into various aspects of the model, including: Feature Importance, Decision Rules , Data Influence and finally Uncertainty Estimation.

Adult income classifier with XGBoost

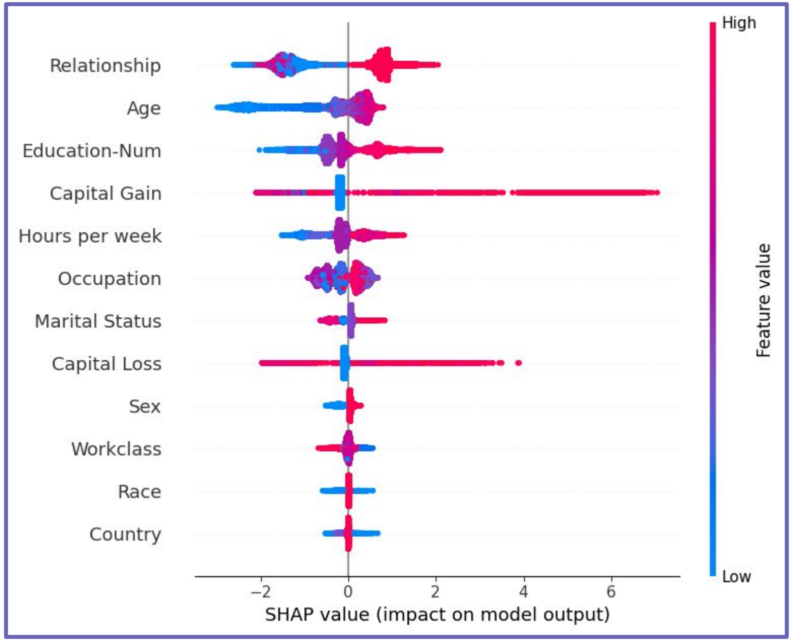
We could understand how SHAP Values will be helping us do the model explanation using the example below.

- **XGBoost :** a tree-based machine learning algorithm for regression and classification tasks, using gradient boosting
- **Dataset : Standard UCI Adult** income dataset uses 12 features : ***'Age', 'Workclass', 'Education-Num', 'Marital Status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Capital Gain', 'Capital Loss', 'Hours per week', 'Country'*.**
- **Goal :** Classify people which income exceeds **$50K / year** or not based on the dataset

XGBoost can provide feature importance score for each feature in the classification task, but in this context, they are contradicting each other so they are not consistent unlike SHAP values

SHAP values for XGBoost Classifier

SHAP summary plot



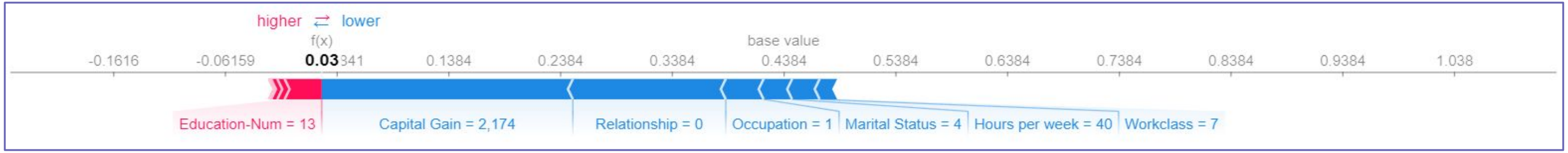
Visualizing model impact of each feature

- ***Capital gain* affects few predictions but intensely**
- ***Relationship feature* affects the overall model predictions but with a maximum of 2.**

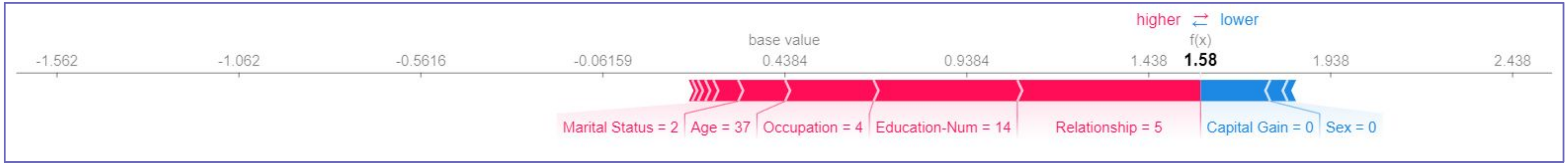
SHAP values for loss understanding

Identification of the features that contributed to the loss value

For a right classification :

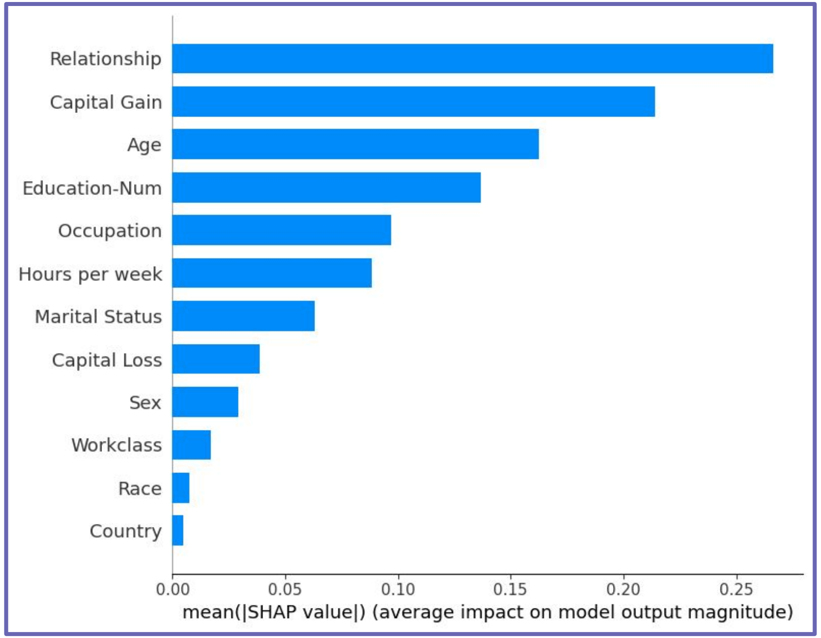


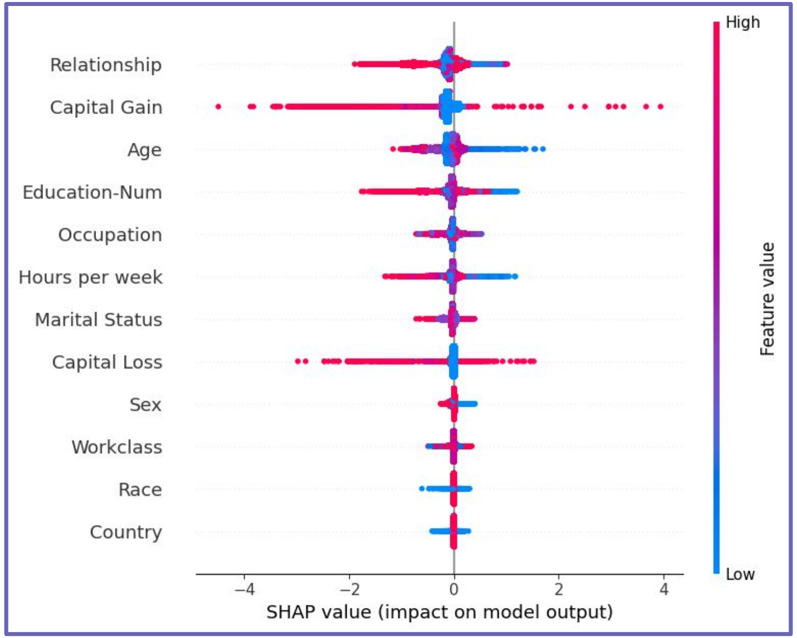
For a wrong classification :



SHAP values for loss understanding

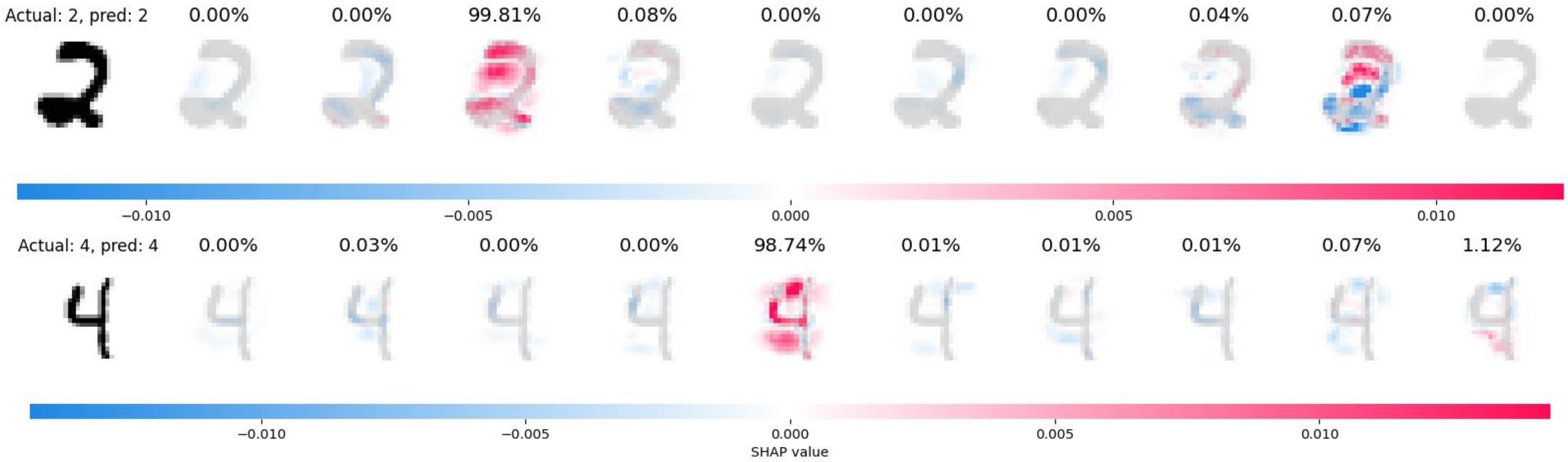
Summary plots for understanding the loss of the model





MNIST Image Classification

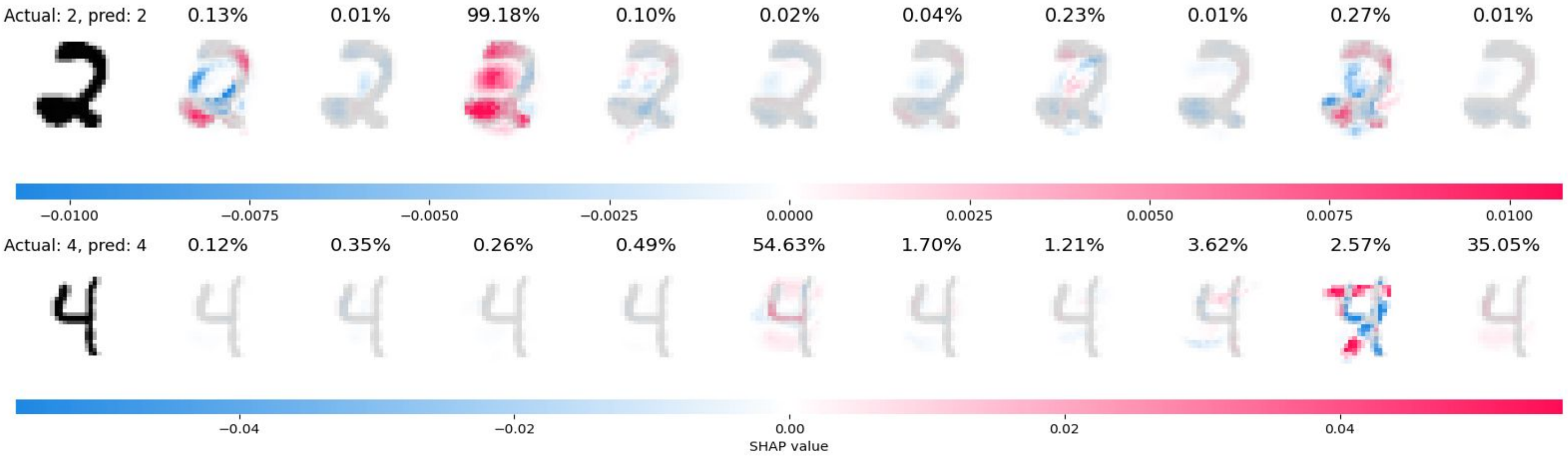
To understand better how SHAP works with Images, and how we can explain the predictions. We will use the MNIST Dataset. The following results are given using CNN after 8 epochs of training and an accuracy > 90%.



- Red pixels increases the probability of a class being predicted
- Blue pixels decrease the probability of a class being predicted

MNIST Image Classification

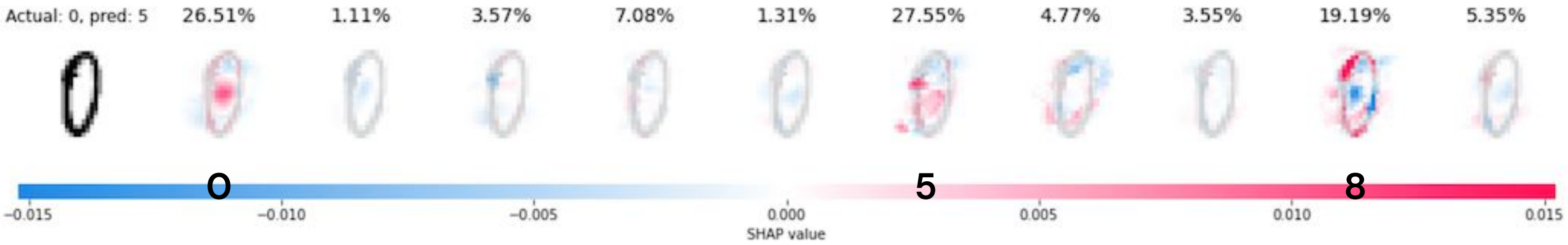
In this case, we are testing the results of SHAP when the predictions are not reliable. To test it, we are going to set the epochs of training to 1. (We can also do an attack)



Let’s put in mind that it's not only what is present that is important to decide what digit an image is, but also what is absent.

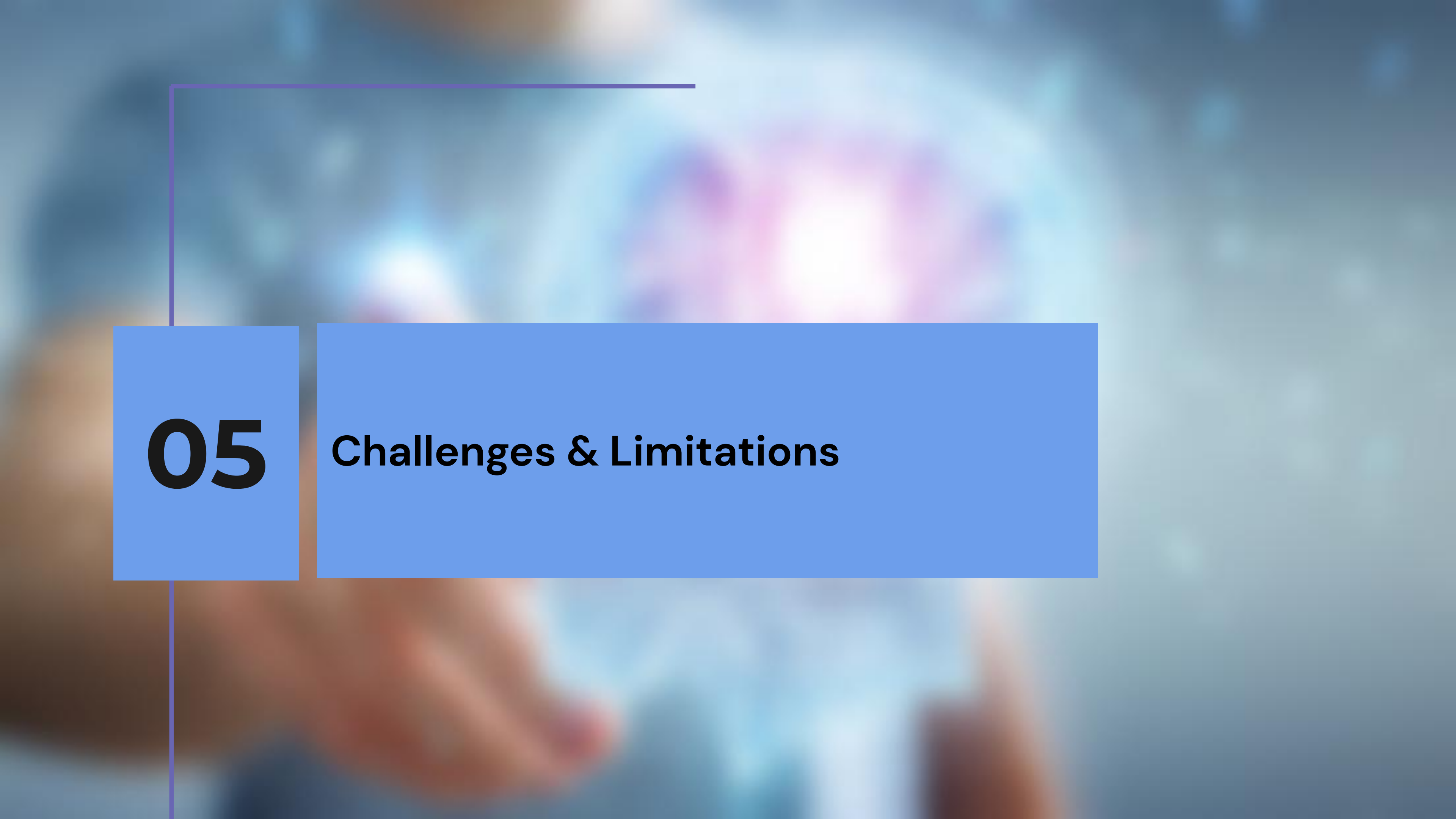
A case that shows SHAP Limitation

In this case, when the model is trained with one epoch, we have results that are wrong and there SHAP is explaining them as well.



Half of the probabilities are wrong. Yet 26% is predicted as 0 since the middle part of the 0 is empty. For 8, the top and down details are taken into account. But what about the 5 ? Few pixels in red were enough to predict that it’s 5 and not 0 with a probability higher than 0’s probability.

Black Box Models , as the exemple showcase, explain what the model is predicting, but do not attempt to explain if the predictions are correct.



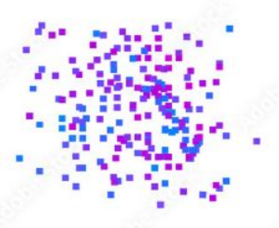
05

Challenges & Limitations

Challenges of the SHAP values

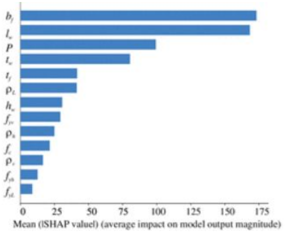
Computational Complexity & Scalability

When can we say that the calculation are complex ?



Computing SHAP values can be computationally expensive, especially for large datasets or complex models. The calculation requires examining all possible feature combinations, which can be time-consuming for high-dimensional data or models with a large number of features.

When can we say that we have a scalability issue ?



SHAP values can become less scalable as the number of features increases. Computing and visualizing SHAP values for models with a large number of features can become more cumbersome and less interpretable.

Challenges of the SHAP values

Ambiguity in Feature Interactions

SHAP values consider the interaction effects between features, but disentangling and interpreting these interactions can be complex. Identifying the specific nature and direction of interactions among features may not always be straightforward and may require additional analysis or domain knowledge.

Interpretability of SHAP Values

While SHAP values provide a measure of feature importance, interpreting the values themselves might be challenging. Understanding the magnitude and direction of SHAP values in relation to the prediction can require expertise and careful analysis, making it less intuitive for non-technical stakeholders.

Challenges of the SHAP values

User interface

For non technical people, it is very hard for them to understand the explanations and the graphs provided. Even if the code is easy, but it is indeed very complex for people to use SHAP with no previous coding experience.

Reliability and model dependency

SHAP values are model-specific and provide explanations for a particular model architecture and training data. They may not be directly transferable to different models or datasets.

When comparing SHAP values between different models, caution should be exercised as they may not be directly comparable due to variations in model behavior and underlying data.

06 Conclusion & Future Directions

Summary of SHAP values

- Based on cooperative game theory
- Method for interpreting and explaining the input features in a model and their contribution to one model output
- Calculate the average marginal contribution of each feature across all possible subsets of features → Quantify the importance of features in the model
- Provides local and global interpretations of the model’s behavior
- Advantageous relatively to other techniques with their ability to capture feature interactions and their possible applications to a wide range of machine learning models and problematics

Future directions

- Making informed decisions, especially for field where human lives are at stake

Perspectives for SHAP values

- More scalability for SHAP values to handle large datasets
- Combining SHAP values with other model explanation techniques
- Developing new applications for SHAP values

Overall, the future of SHAP values in AI looks promising, and further research and development could expand their use and potential applications in model explanation and interpretability in many more fields.