Explainable AI (XAI)

What? What for? Why? And How?

Creator: Mohammad Amin Dadgar

Table Of Contents

Explaining XAI

Section 1

Terms & Definitions

- White/Glass box Methods (Transparent)[1,2]
  - Linear/Logistic Regression
  - Decision Trees
  - K-Nearest Neighbor
  - Rule Based learning
  - Bayesian models
- Black-box Methods [1]
  - Neural Networks
  - All transparent methods with huge number of features and correlations

[1] "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez, journal: Information Fusion, 2020, pp:82-115

[2] "From Machine Learning to Explainable AI", Andrew holzinger, journal: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)

Terms & Definitions

- Interpretations [1]
  - The reasons and rationales
- Interpretability [1,2]
  - The degree of human understandability
- Explanations [2]
  - Answer the broader question “Why?”
  - For each decisions of the model
- Explainability [2]
  - Answer globally to the question “Why?”
  - A Reason for the whole decision making process

[1] “Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond”, Li, Xuhong and Xiong, Haoyi and Li, Xingjian and Wu, Xuanyu and Zhang, Xiao and Liu, Ji and Bian, Jiang and Dou, Dejing, Computer and information sciences, 2021

[2] “Explainable Fuzzy Systems: Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems”, Alonso, Jose and Castiello, Ciro and Magdalena, Luis and Mencar, Corrado, 2021, pp:222-223

“Wh” questions

- What is XAI?
  - Ability to give comprehensible reasons for decision making process
  - Explaining model’s functionality in different tasks
- Why?
  - Understanding the rational process of decision making
  - In critical problems: Health care, banking, judgment

[1] “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez, journal: Information Fusion, 2020, pp:82-115

"Wh" questions (continue)

- What for?[1]
  - Trustworthiness
  - Casuality
  - Transferability
  - Confidence
  - Fairness
  - Informativeness
  - Privacy Awareness [1,2]

[1] "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez, journal: Information Fusion, 2020, pp:82-115
[2] AI Explainable AI: the basics - Royal Society, November 2019

How can it help?

- Transparent models
  - Simulatibility
  - Decomposability
  - Algorithmic Transparency
- Post-hoc methods
  - Text Explanations
  - Visual Explanations
  - Local/Global Explanations
  - Explanations by simplifications
  - Feature importance/relevance scores

[1] “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez, journal: Information Fusion, 2020, pp:82-115

Problems & Reactions

Section 2

Use cases

- Algorithms behind the Apple Credit Card are accused of being gender-biased [1]
  - Apple co-founder Steve Wozniak says Apple Card discriminated against his wife
- Amazon's system for curriculum-vitae screening was found to be biased against women [2]
  - Men were preferable than women to be hired
- Machine Bias for black and white people in predicting future criminals and judgements to be applied for criminals [3]
  - Huge Risk difference between a white person and black person

[1] Duffy, Clare. 2019. "Apple co-founder Steve Wozniak says Apple Card discriminated against his wife." *CNN Business*.
[2] Dastin, Jeffrey. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*
[3] There's software used across the country to predict future criminals. And it's biased against blacks. by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica 2016

Regulatory Acts

- **GDPR:** Article 22 empowers individuals with the right to demand an explanation of decision-making process of automated systems
- **California Consumer Privacy Act 2019:** Requires companies to align the process of collection, storage, and sharing of personal data with the new requirements of january 1, 2020
- **Washington Bill 1655:** Introduces measures for the use of automated decision systems to protect consumers, improve transparency and create more market predictability

[1] Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more, Book by Aditya Bhattacharya, 2022, pp:10-11

Regulatory Acts

- **Algorithmic Accountability 2019:** Mandates organizations to provide assessments of the risks of having automated decision systems to privacy, security, inaccurate, unfair, biased, or discriminatory outcomes impacting consumers.
- **Illinois House Bill 3415:** Establishes guidelines for not including information related to applicant race or zip code for predictive data analytics for the purpose of hiring financial services
- **Massachusetts Bill H.2701:** Establishes guidelines on automated decision-making, transparency, fairness and individual rights

[1] Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more, Book by Aditya Bhattacharya, 2022, pp:10-11

The Explanations Methods - Part I

Section 3

The Explanation Methods

- Transparent Models
  - Provides Explanations without the use of any other method
- Examples
  - Linear/Logistic Regression
  - Decision Trees
  - K-Nearest Neighbor
  - Rule Based learning
  - Bayesian models

[1] "Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond", Li, Xuhong and Xiong, Haoyi and Li, Xingjian and Wu, Xuanyu and Zhang, Xiao and Liu, Ji and Bian, Jiang and Dou, Dejing, Computer and information sciences, 2021

The Explanation Methods (continue)

- Post-hoc methods
  - Model agnostic [1]
    - LIME
    - SHAP
    - Break-down
  - Differentiable Model Specific [1]
    - DeepLift
    - Layer-wise Propagation (LRP)
    - Integrated Gradients (IG)
  - Model Specific
    - GNNExplainer [1]
    - RandomForestExplainer [2]

[1] "Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond", Li, Xuhong and Xiong, Haoyi and Li, Xingjian and Wu, Xuanyu and Zhang, Xiao and Liu, Ji and Bian, Jiang and Dou, Dejing, Computer and information sciences, 2021

[2] "Structure mining and knowledge extraction from random forest with applications to The Cancer Genome Atlas project", Master's thesis in the field of applied mathematics, University of Warsaw, 2017

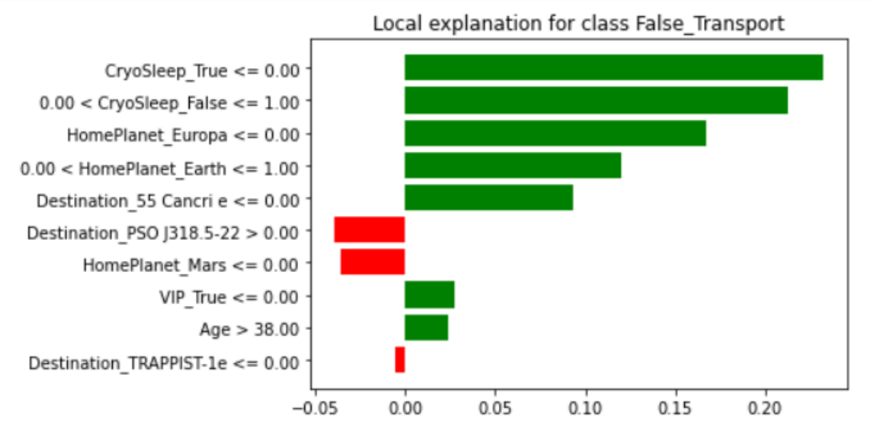
The Explanation Methods (continue)

- Local Explanations
  - Gives explanations for samples in dataset
  - E.g. Why a specific person cannot gets loan?
  - Useful for the end users
  - Methods like: LIME, SHAP, DeepLift
- Global Explanations
  - Gives explanations for the whole model's rationale
  - E.g. How the loan applications are accepted or not accepted?
  - Useful for developers to find if the model is not working correctly
  - Methods like: SP-LIME, XGNN

[1] AI Explainable AI: the basics - Royal Society, November 2019

LIME

- Local Interpretable Model-agnostic Explanations [1]
  - The category of surrogate models
  - Model-agnostic



[1] "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos, Computer and information sciences, 2016

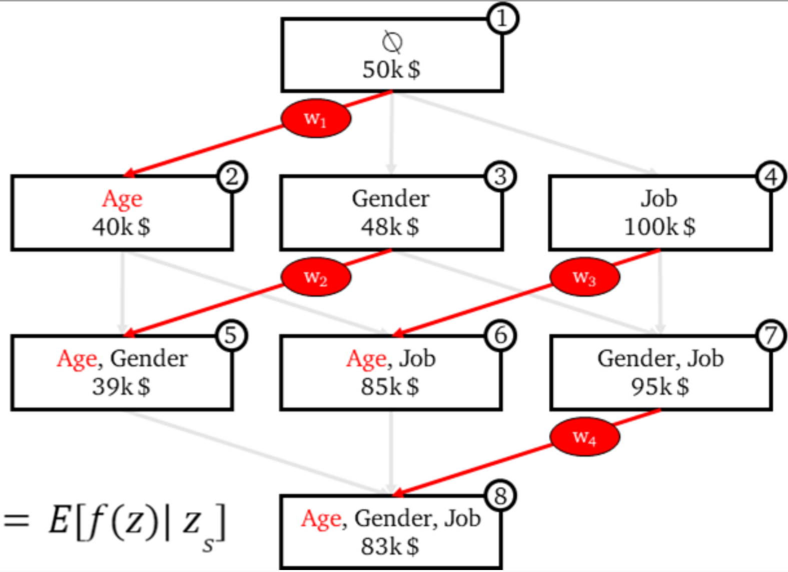
Shapley Values (SHAP)

- Additive feature method
- Finds the Expected value of the different formation of adding each feature
- The expected value is the feature importance

$w_1+w_2+w_3+w_4=1$

$w_1=w_2+w_3=w_4 \Rightarrow w_1=w_4=⅓, w_2=w_3=⅙$

Shapley value for *Age* = $w_1*(-10)+w_2*(-9)+w_3*(-15)+w_4*(-12) \simeq -11.3$

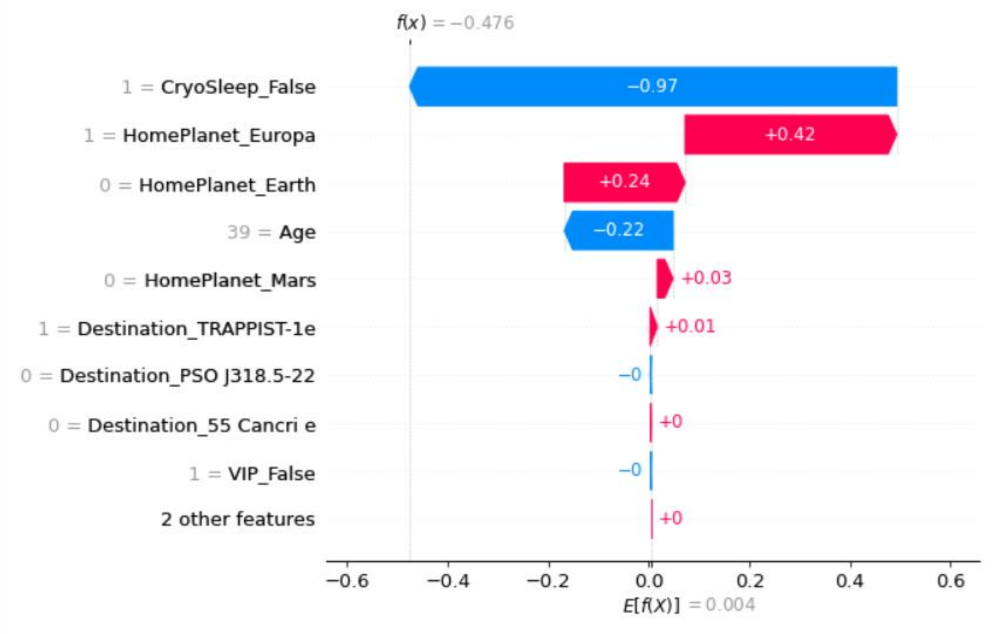


$$f_x(z_s) = E[f(z) | z_s]$$

[1] "A Unified Approach to Interpreting Model Predictions", I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach, Curran Associates, Inc., 2017

[2] Image Reference: "SHAP Values Explained Exactly How You Wished Someone Explained to You",Samuele Mazzanti, Towards Data Science, 2020

Shapley Values (SHAP)



DeepLift

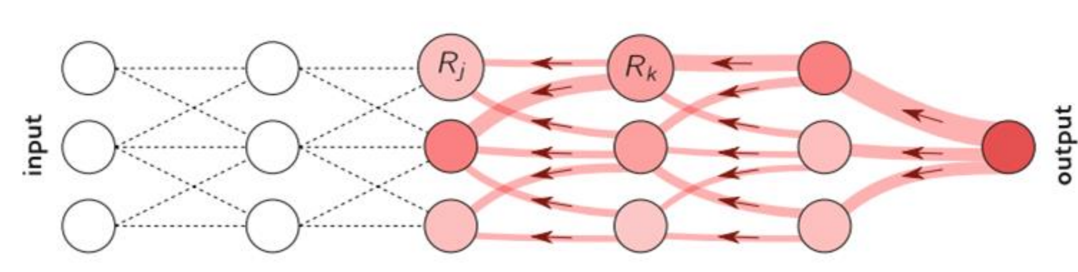
- Find feature importance by propagating once through network
- Uses
  - Difference from reference
    - Reference of intermediate neurons are calculated by propagating through network
- Propagate Twice through network
  - Find each neuron reference value
  - Feature importance found by contribution-score
  - contribution-score is found by multipliers

[1] "Learning Important Features Through Propagating Activation Differences", Shrikumar, Avanti and Greenside, Peyton and Kundaje, Anshul, Computer and information sciences, 2017

Layer-wise Relevance Propagation (LRP)

- Feature Relevance method
- Backpropagate through network
  - To find relevances of each neuron
- Advantages:
  - Is possible to find the model's relevance score for other decisions
  - Light weight

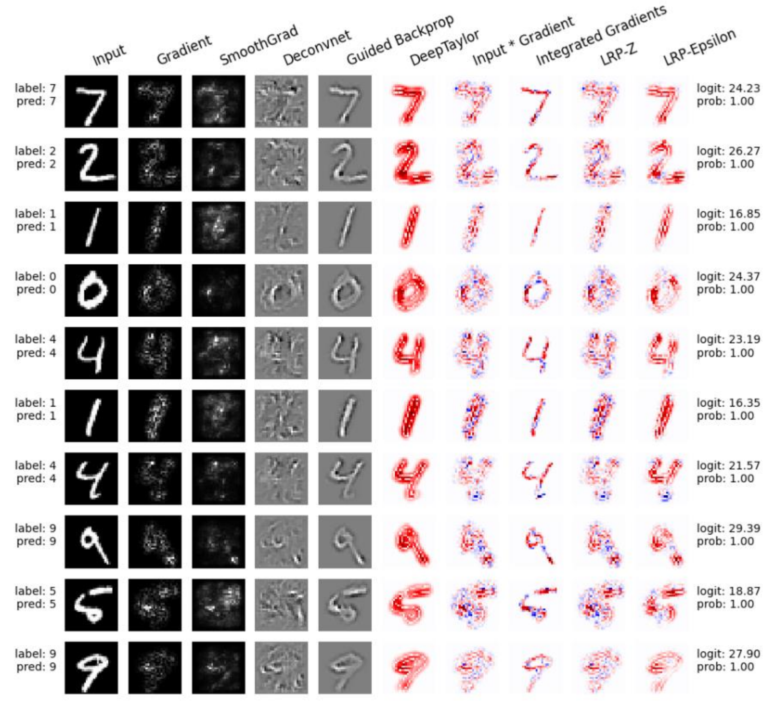
$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$



[1] "Layer-Wise Relevance Propagation: An Overview", Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, KR., Springer International Publishing, 2019, pp:193-209

Examples

- Src: https://github.com/albermax/innvestigate/blob/master/examples/mnist_compare_methods.ipynb



The Explanations Methods - Part II

Section 4

Partial Dependence Plots (PDP)

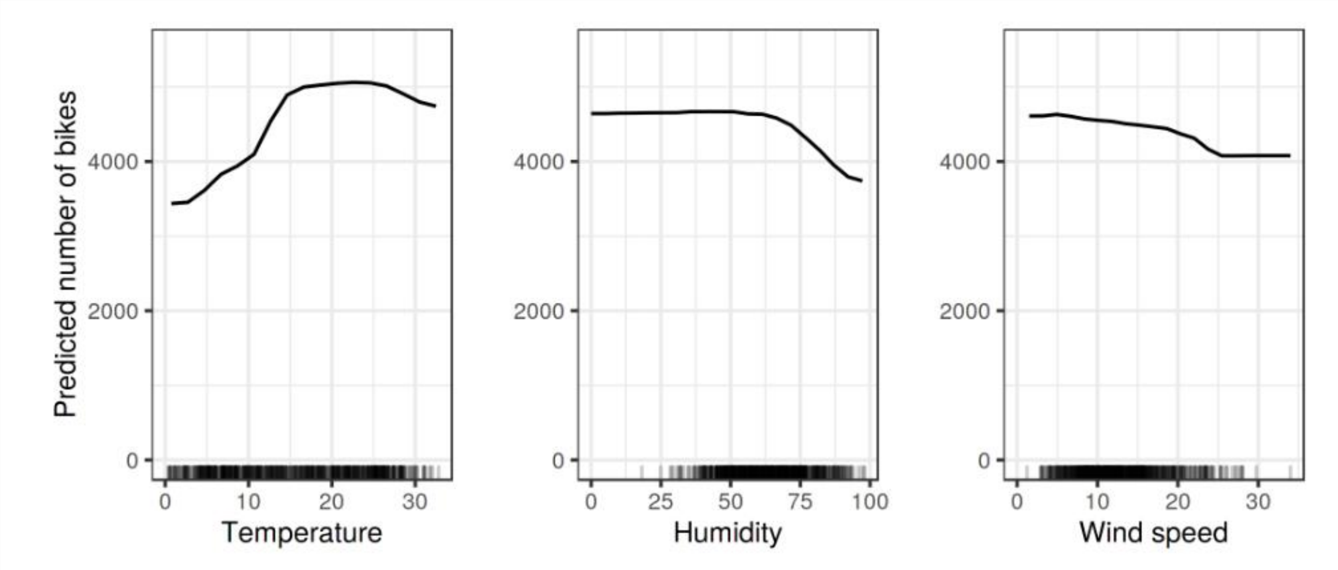
- Global Interpretation method
- Shows a feature effect on output
  - By marginalizing a set features over all other ones shows their dependency on output
  - Normally one or two features are shown (marginalized over all other features)

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)d\mathbb{P}(X_C)$$

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})$$

[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022

Partial Dependence Plots (PDP) - continue



[1] “Interpretable machine learning: a guide for making black box models explainable”, Christoph Molnar, leanpub, 2022

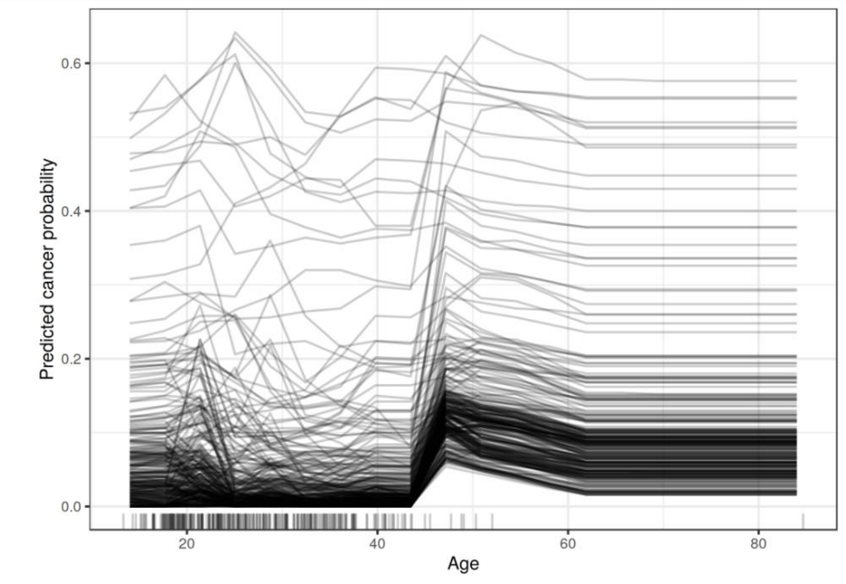
Discussions: PDP

- Easy to understand for people
- PDP feature importance is also proposed [2]
  - Measures the fluctuation of PDP, and represent it as a number
- Untrustworthy for correlated variables problem
  - Can create unreal values
    - A 2 meter height person with 40 Kg weight
  - Use the alternatives
    - ICE plots
    - ALE plots

[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022
[2] "A Simple and Effective Model-Based Variable Importance Measure", Greenwell, Brandon M. and Boehmke, Bradley C. and McCarthy, Andrew J., arXiv, 2018

Individual Condition Expectation (ICE) plots

- Local Interpretation method
- PDP for several individual data
- For each instance one feature is changing and all others are fixed



[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022

ARTIFICIAL Cassandra INTELLIGENCE

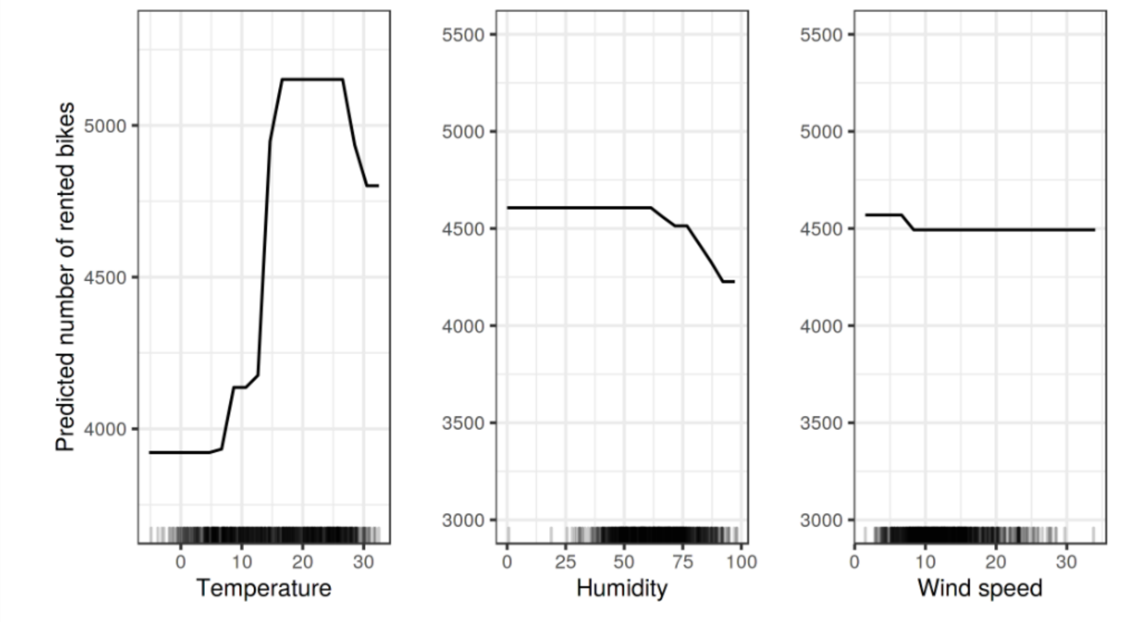
Accumulated Local Effects (ALE) plots

- Process to achieve ALE plots
  - *Step 1:* Divide up the a feature space into intervals
  - *Step 2:* The predicted output of the features is differed from interval bounds (is called effect)
  - *Step3:* Effects of the interval are summed and divided by the count of features in that interval
  - *Step 4:* go through step two for another interval till all the intervals end
  - *Step 5:* Accumulate all the values and divide them by the count of intervals

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[\hat{f}\left(z_{k,j}, x_{\setminus j}^{(i)}\right) - \hat{f}\left(z_{k-1,j}, x_{\setminus j}^{(i)}\right)\right]$$

[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022

Accumulated Local Effects (ALE) plots - continue



[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022

Discussions: ALE

- Unbiased Plots
  - Works for correlated features
- Faster to compute than PDP
  - PDP, calculates for all values of a feature
    - For a feature height, from 50 cm to 220 cm
  - ALE, calculates the difference from intervals for each feature
    - For a feature height, average of prediction differences in intervals 50 - 70, …
- The diagram could be shaky if the intervals are small
  - Or if the intervals are large, then the influences might become hidden

[1] "Interpretable machine learning: a guide for making black box models explainable", Christoph Molnar, leanpub, 2022

Discussions: PDP & ALE

- Partial Dependence plots (PDP)
  - "Let me show you what the model predicts on average when each data instance has the value v for that feature. I ignore whether the value v makes sense for all data instances."
- ALE plots:
  - "Let me show you how the model predictions change in a small "window" of the feature around v for data instances in that window."

[1] “Interpretable machine learning: a guide for making black box models explainable”, Christoph Molnar, leanpub, 2022

The Explanations Methods - Part III

Section 5

Explanations By Examples

- Factual Examples
  - By Extrapolation
    - The LIME family methods
    - live [1]
  - The SHAP family methods
  - Breakdown [1]
- Counterfactual Examples
  - DiCE [2]

[1] “Explanations of Model Predictions with live and breakDown Packages”, Mateusz Staniak and Przemys Biecek, The R Journal, 2019
[2] “Explaining machine learning classifiers through diverse counterfactual explanations”, Ramaravind K. Mothilal and Amit Sharma and Chenhao Tan, ACM, 2020

DiCE method

```
In [69]: # Visualize counterfactual explanation
         dice_exp.visualize_as_dataframe()
```

Query instance (original outcome : 0)

	Age	HomePlanet_Earth	HomePlanet_Europa	HomePlanet_Mars	CryoSleep_False	CryoSleep_True	VIP_False	VIP_True	Destination_55 Cancri e	Destination_PSO J318.5-22	De
0	44.0	1	0	0	1	0	1	0	0	1	

Diverse Counterfactual set (new outcome: 1.0)

	Age	HomePlanet_Earth	HomePlanet_Europa	HomePlanet_Mars	CryoSleep_False	CryoSleep_True	VIP_False	VIP_True	Destination_55 Cancri e	Destination_PSO J318.5-22	De
0	5.0	1.0	0.0	0.0	0	0.0	1.0	0.0	0.0	1.0	
1	44.0	0	0.0	0.0	0	0.0	1.0	0.0	0.0	1.0	
2	44.0	1.0	1	0.0	0	0.0	1.0	0.0	0.0	1.0	
3	44.0	1.0	1	0.0	1.0	1	1.0	0.0	0.0	1.0	

[1] "Explaining machine learning classifiers through diverse counterfactual explanations", Ramaravind K. Mothilal and Amit Sharma and Chenhao Tan, ACM, 2020

[2] "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR", Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, Computer and information sciences, 2017, pp: 17

Codes and Libraries

- Our code for LIME, SHAP, and DiCE
  - https://github.com/CASS-AI/XAI-model-agnostic-methods-attempt
- Libraries to use
  - LIME
  - SHAP
  - DiCE
  - Quantus
  - iNNvestigate
  - Aix360
- To get updates and have an access to a full list of libraries
  - https://github.com/stars/amindadgar/lists/xai-tools

References

- “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez, journal: Information Fusion, 2020, pp:82-115
- “From Machine Learning to Explainable AI”, Andrew holzinger, journal: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)
- “Explainable Fuzzy Systems: Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems”, Alonso, Jose and Castiello, Ciro and Magdalena, Luis and Mencar, Corrado, 2021
- AI Explainable AI: the basics - Royal Society, November 2019
- Duffy, Clare. 2019. “Apple co-founder Steve Wozniak says Apple Card discriminated against his wife.” *CNN Business*.
- Dastin, Jeffrey. 2018. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women.” *Reuters*
- There’s software used across the country to predict future criminals. And it’s biased against blacks. by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica 2016
- Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more, Book by Aditya Bhattacharya, 2022, pp:10-11
- “Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond”, Li, Xuhong and Xiong, Haoyi and Li, Xingjian and Wu, Xuanyu and Zhang, Xiao and Liu, Ji and Bian, Jiang and Dou, Dejing, Computer and information sciences, 2021
- “Structure mining and knowledge extraction from random forest with applications to The Cancer Genome Atlas project“, Master's thesis in the field of applied mathematics, University of Warsaw, 2017
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos, Computer and information sciences, 2016

References

- General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, Molnar, Christoph and König, Gunnar and Herbinger, Computer and information sciences, 2020
- “xxAI - Beyond Explainable AI”, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, Wojciech Samek International Workshop, Held in Conjunction with ICML 2020
- “A Unified Approach to Interpreting Model Predictions”, I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach, Curran Associates, Inc., 2017
- “SHAP Values Explained Exactly How You Wished Someone Explained to You”,Samuele Mazzanti, Towards Data Science, 2020
- “Gradients of Counterfactuals”, Sundararajan, Mukund and Taly, Ankur and Yan, Qiqi, Computer and information sciences, 2016
- “Learning Important Features Through Propagating Activation Differences”, Shrikumar, Avanti and Greenside, Peyton and Kundaje, Anshul, Computer and information sciences, 2017
- “Explanations of Model Predictions with live and breakDown Packages”, Mateusz Staniak and Przemys Biecek, The R Journal, 2019
- “Explaining machine learning classifiers through diverse counterfactual explanations”, Ramaravind K. Mothilal and Amit Sharma and Chenhao Tan, ACM, 2020
- “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”, Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, Computer and information sciences, 2017, pp:17
- “Interpretable machine learning: a guide for making black box models explainable”, Christoph Molnar, leanpub, 2022
- “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”, Kim, Been and Wattenberg, Martin and Gilmer, Justin and Cai, Carrie and Wexler, arXiv, 2017

Thanks for joining us

- If you have any questions feel free to ask on
  - Github Account Discussion
  - My Email: dadgaramin96@gmail.com
  - Telegram account: @mramin22

src: unsplash.com