LOGIN / REGISTER (HTTPS://ID.ANALYTICSVIDHYA.COM/AUTH/LOGIN/?NEXT=HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2021/01/EXPLAIN-HOW-YOUR-MODEL-WORKS-USING-EXPLAINABLE-AI/)

Analytics Vidhya (https://www.analyticsvidhya.com/blog/)

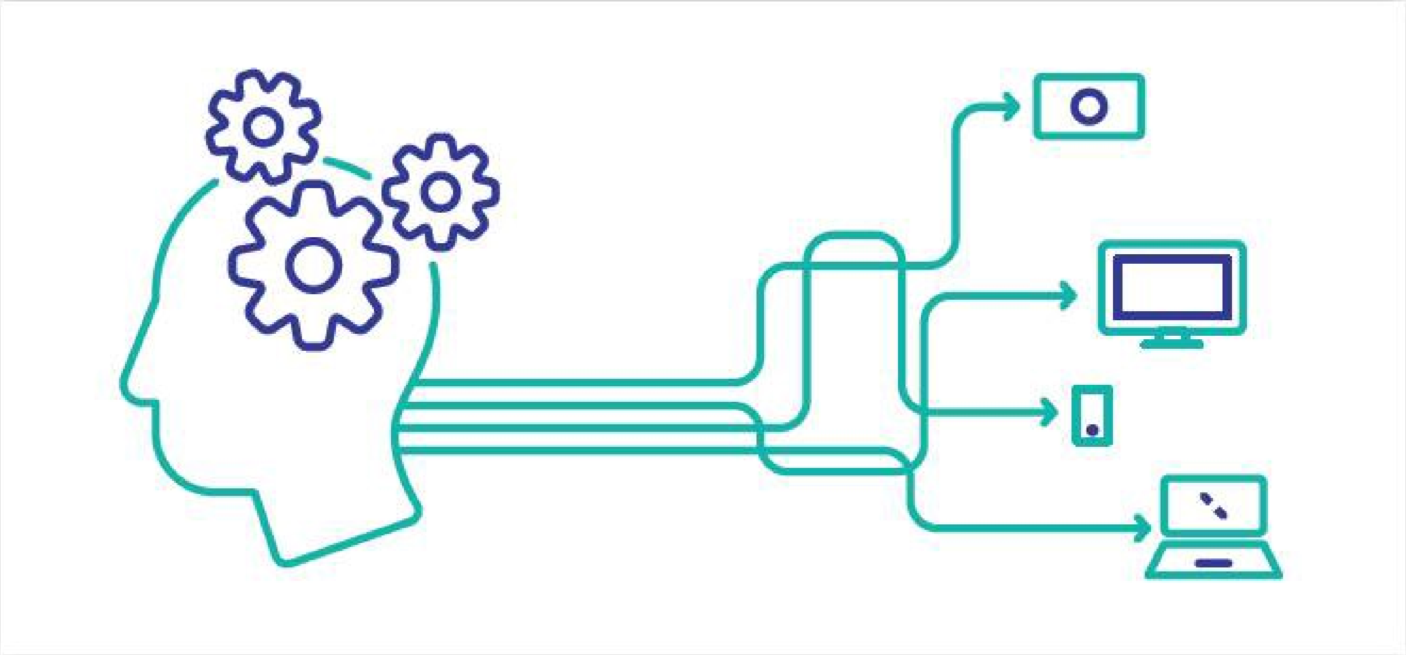
INTERMEDIATE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/INTERMEDIATE/)

MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

Explain How Your Model Works Using Explainable AI

GUEST BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/GUEST-BLOG/), JANUARY 7, 2021 LOGIN TO BOOKMARK THIS A…

Article Video Book



Can you explain how your model works?

Meet XAI!

Artificial intelligence techniques are used to solve real-world problems. We get the data, perform some operations to make it clean & ready for the following processes.

We basically **pick things from this world** and **take them into the world of machines**, represent it with numbers, and then feed it to a bunch of models. Try to improve them and eventually *"the winner model"* gets the test data. A vital question comes to the minds :

> *" How do we take this result back to real world ? "*

Explainable AI (with a cooler name: XAI)

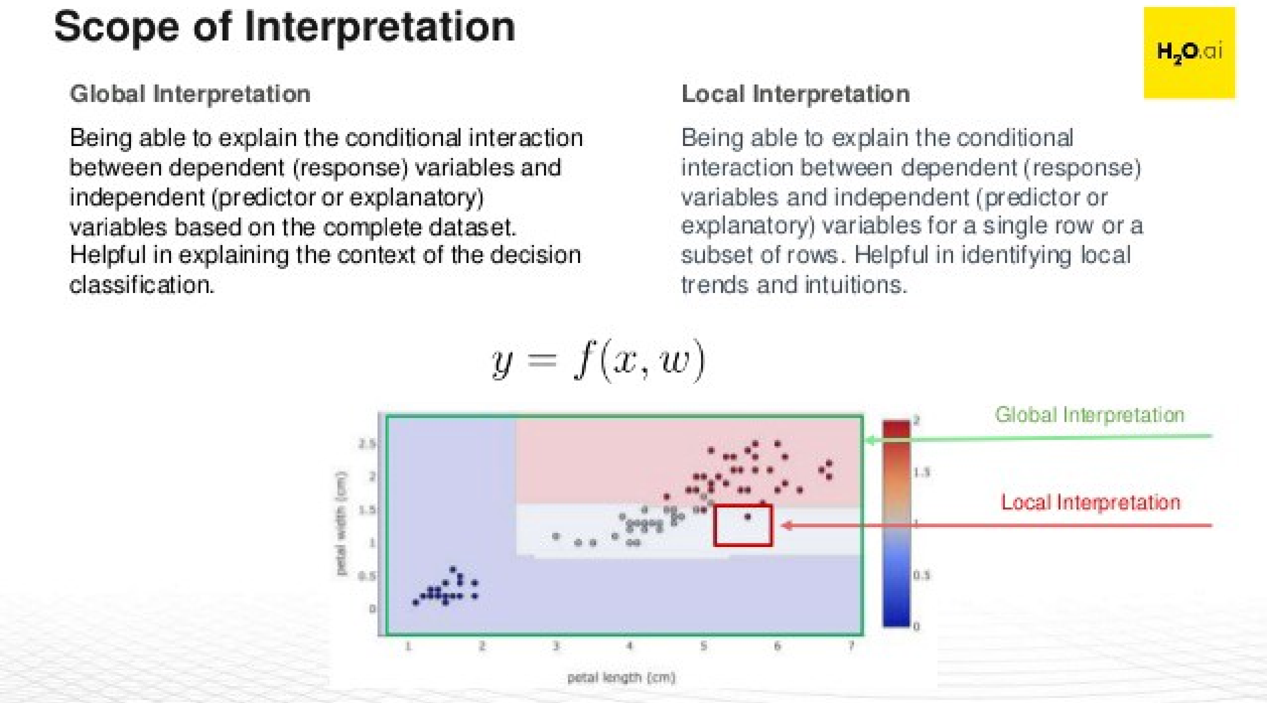
A formal definition: According to Wikipedia, Explainable AI refers to methods and techniques in the application of artificial intelligence technology such that the results of the solution can be understood by humans. [1]

In the early phases of AI adoption, it was okay to not understand what the model predicts in a certain way, as long as it gives the correct outputs. Explaining how they work was not the first priority. Now, the focus is turning to build ***human interpretable models***.

> *Three important aspects of model interpretation are:*
> *1. Transparency*
> *2. The ability to question*
> *3. The ease of understanding .[2]*

Model interpretability can be examined in two levels:

- **Global Interpretation:** Examines the model from a broader perspective. For example, let's say we are working on a house price dataset and we implemented a neural network. The global interpretation might say "Your model uses # of squared feet as an important feature to derive predictions"
- **Local Interpretation:** As the name suggests, this approach is focused on a certain observation/data point. Let's continue moving forward with our example. Prediction for a really small house turned out large. Local interpretation looks at the other features and it might say "Your model predicted this way because the location of the house is very close to the city center."



Scope of Interpretation

Global Interpretation

Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables based on the complete dataset. Helpful in explaining the context of the decision classification.

Local Interpretation

Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables for a single row or a subset of rows. Helpful in identifying local trends and intuitions.

$$y = f(x, w)$$

Source: [Sri Ambati, Get Hands on MLI (https://www.slideshare.net/0xdata/get-handson-with-explainable-ai-at-machine-learning-interpretabilitymli-gym)](https://www.slideshare.net/0xdata/get-handson-with-explainable-ai-at-machine-learning-interpretabilitymli-gym)

The Trade-off Between Accuracy and Interpretability

In the industry, you will often hear that **business stakeholders tend to prefer models that are more interpretable** like **linear models (linear\logistic regression)** and **trees** which are intuitive, easy to validate, and explain to a non-expert in data science. [2]

In contrast, when we look at the complex structure of real-life data, in the model building & selection phase, the interest is mostly shifted towards more *advanced models*. That way, we are more likely to obtain improved predictions.

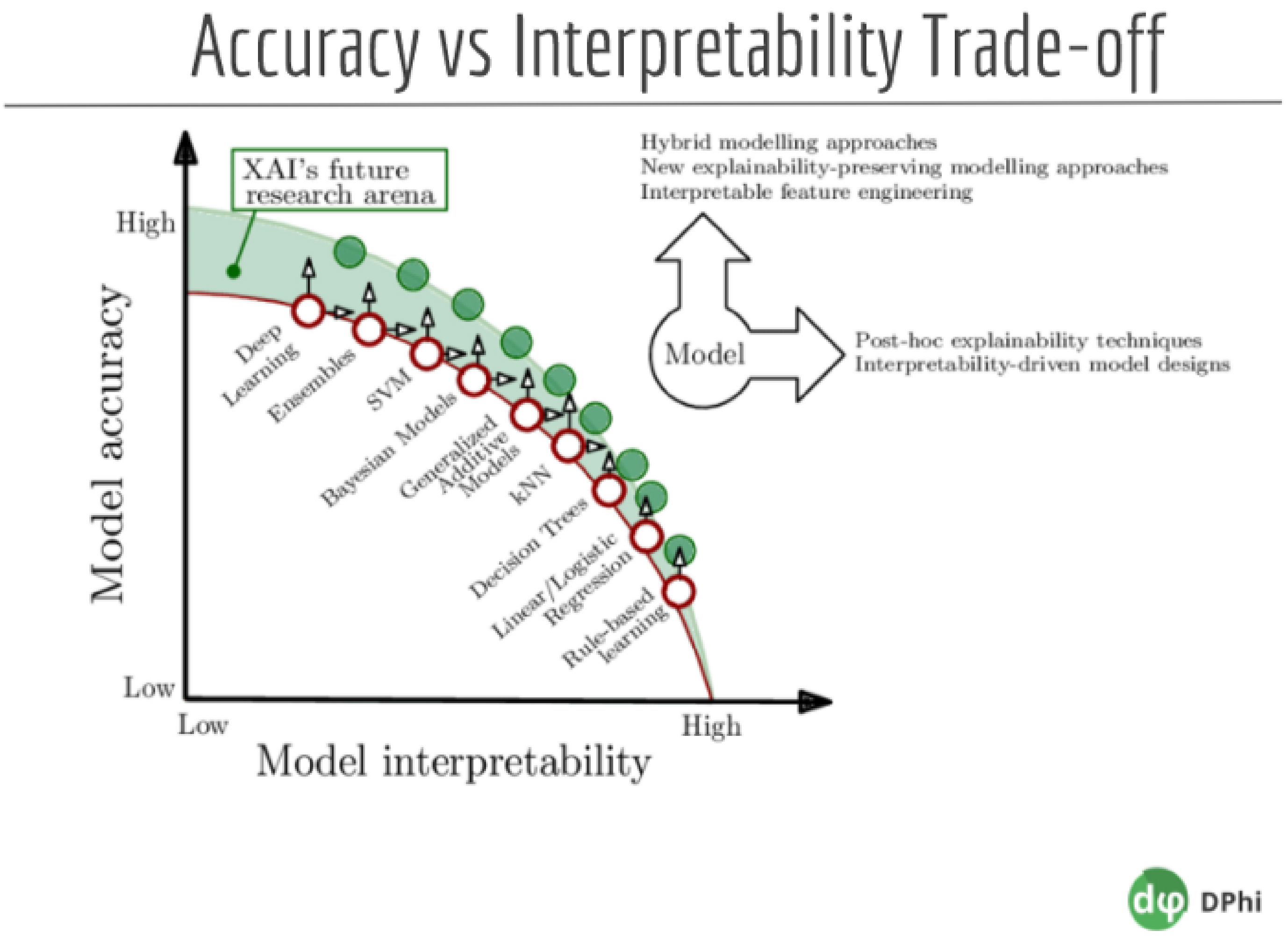
Models like these (ensembles, neural networks, etc.) are called **black-box** models. As the model gets more advanced, it becomes harder to explain how it works. Inputs magically go into a box and voila! We get amazing results.

But, HOW?

When we suggest this model to stakeholders, will they completely trust it and immediately start using it? **NO**. *They will ask questions* and we should be ready to answer them.

> *Why should I trust your model?*
> *Why did the model take a certain decision?*
> *What drives model predictions?*

We should consider both improving our model accuracy and not get lost in the explanation. There should be a balance between both.


Accuracy vs Interpretability Trade-off
XAI's future research arena
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs
DPhi

Source: [DPhi Advanced ML Bootcamp — Explainable AI (https://dphi.tech/lms/learn/ml-bootcamp-advanced/687)](https://dphi.tech/lms/learn/ml-bootcamp-advanced/687) [2]

Here, I would like to share a sentence from Dipanjan Sarkar's [medium post (https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476)](https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476) about explainable AI:

> *Any machine learning model at its heart has a response function which tries to map and explain relationships and patterns between the independent (input) variables and the dependent (target or response) variable(s). [3]*

So, models take inputs and process them to get outputs. *What if our data is biased?* It will also make our **model biased** and therefore **untrustworthy**. It is important to understand & be able to explain to our models so that we can also trust their predictions and maybe even detect issues and fix them before presenting them to others.

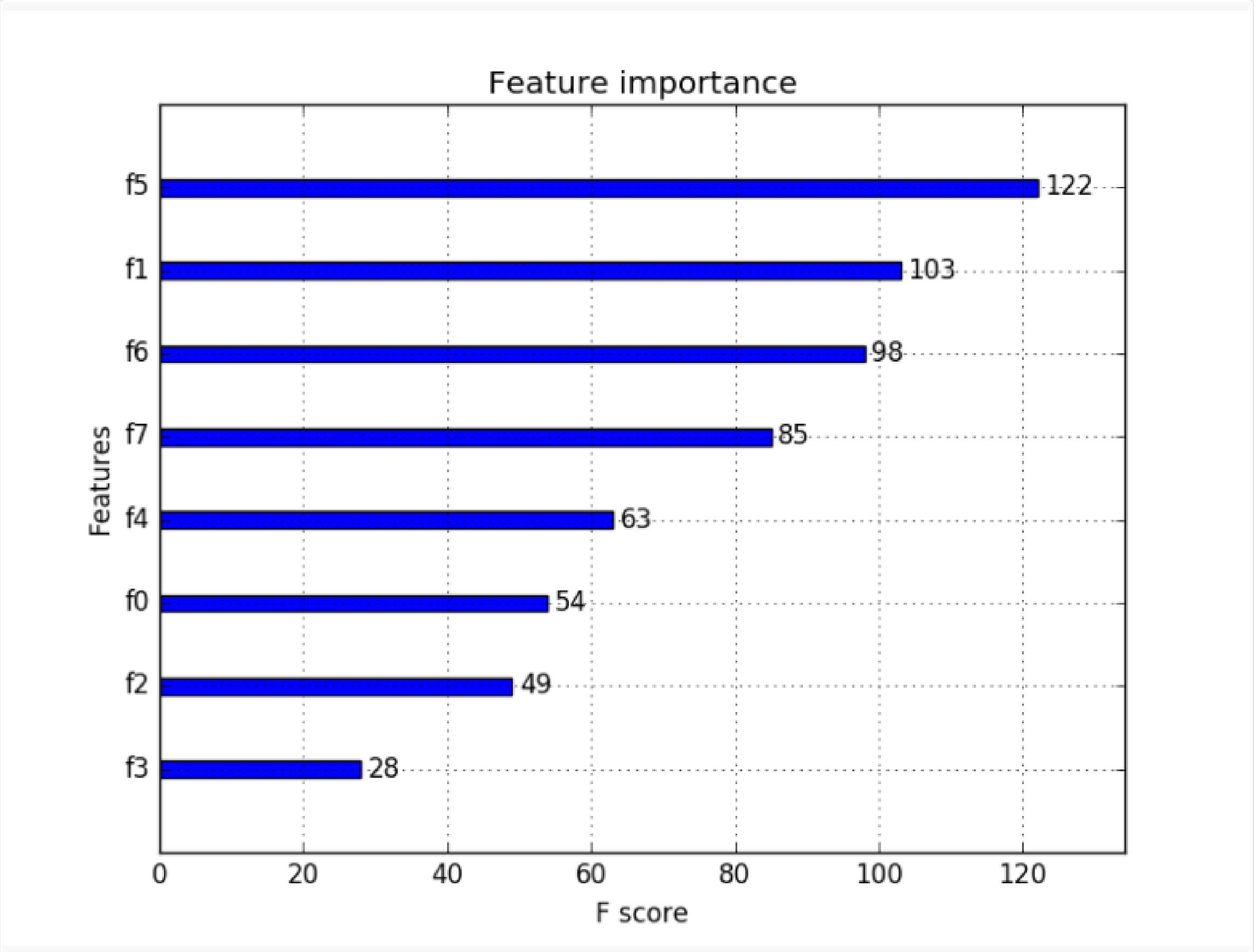
To improve the interpretability of our models, there are various techniques some of which we already know and implement. Traditional techniques are exploratory data analysis, visualizations, and model evaluation metrics. With the help of them, we can get an idea of the model's strategy. However, they have some limitations. To learn more about traditional ways and their limitations, check out this (https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739) amazing article by Dipanjan Sarkar.[4]

Other model interpretation techniques and libraries have been developed to overcome limitations. Some of these are :

- **LIME** (Local Interpretable Model-Agnostic Explanations)
- **SHAP** (Shapley Additive Explanations)
- **ELI5** (Explain Like I'm 5)
- **SKATER**

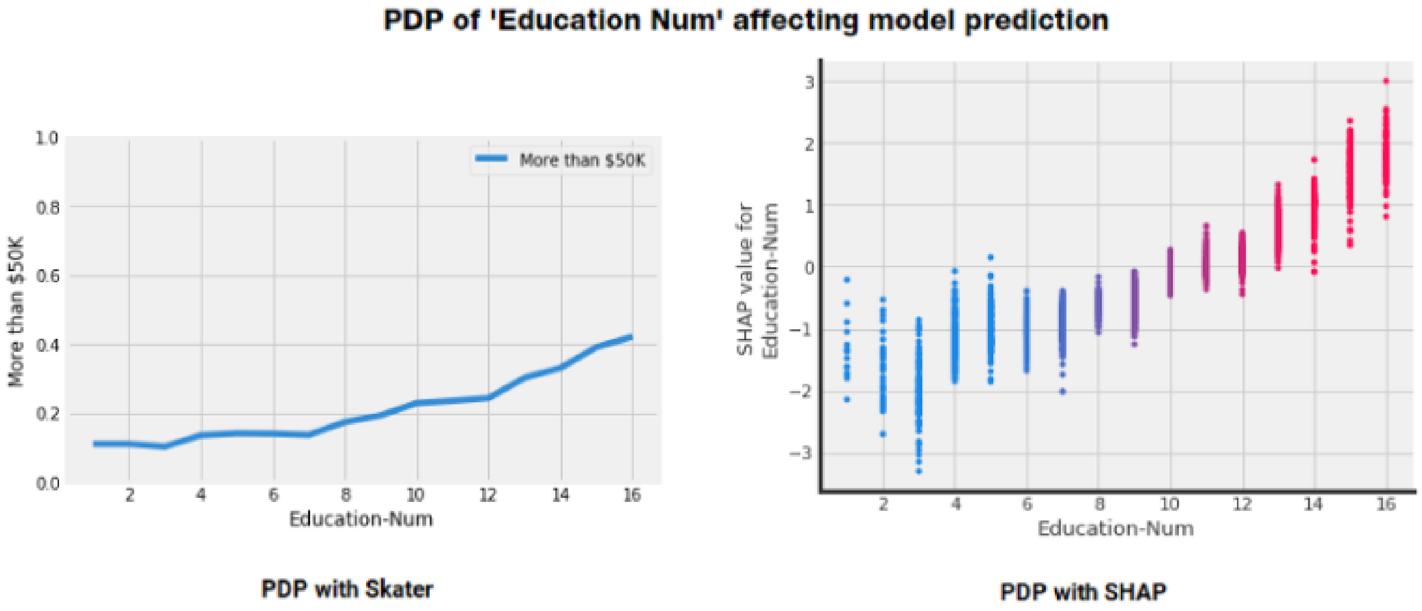
These libraries use feature importance, partial dependence plots, individual conditional expectation plots to explain less complex models such as linear regression, logistic regression, decision trees, etc.

Feature importance shows how a feature is important for the model. In other words, when we delete the feature from the model, how our error changes? If the error increases a lot, this means that a feature is important for our model to predict the target variable.



Source: Machine Learning Mastery, XGBoost Feature Importance Bar Chart (https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/)

Partial dependence plots visualize the effect of the change for a certain feature when everything else is held constant (with a cooler phrase: ceteris paribus). With the help of these, we can see a possible limit value, where this value is exceeded, it directs the model predictions the other way. When we are visualizing partial dependence plots, we are examining the model globally.



Source: Dipanjan (DJ) Sarkar, Model Interpretation Strategies (https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739)

Individual conditional expectation plots show the effect of changes for a certain feature, just like partial dependency plots. But this time, the point of view is local. We are interested to see the effect of changes for a certain feature **for all instances in our data**. A partial dependence plot is the average of the lines of an ICE plot. [5]

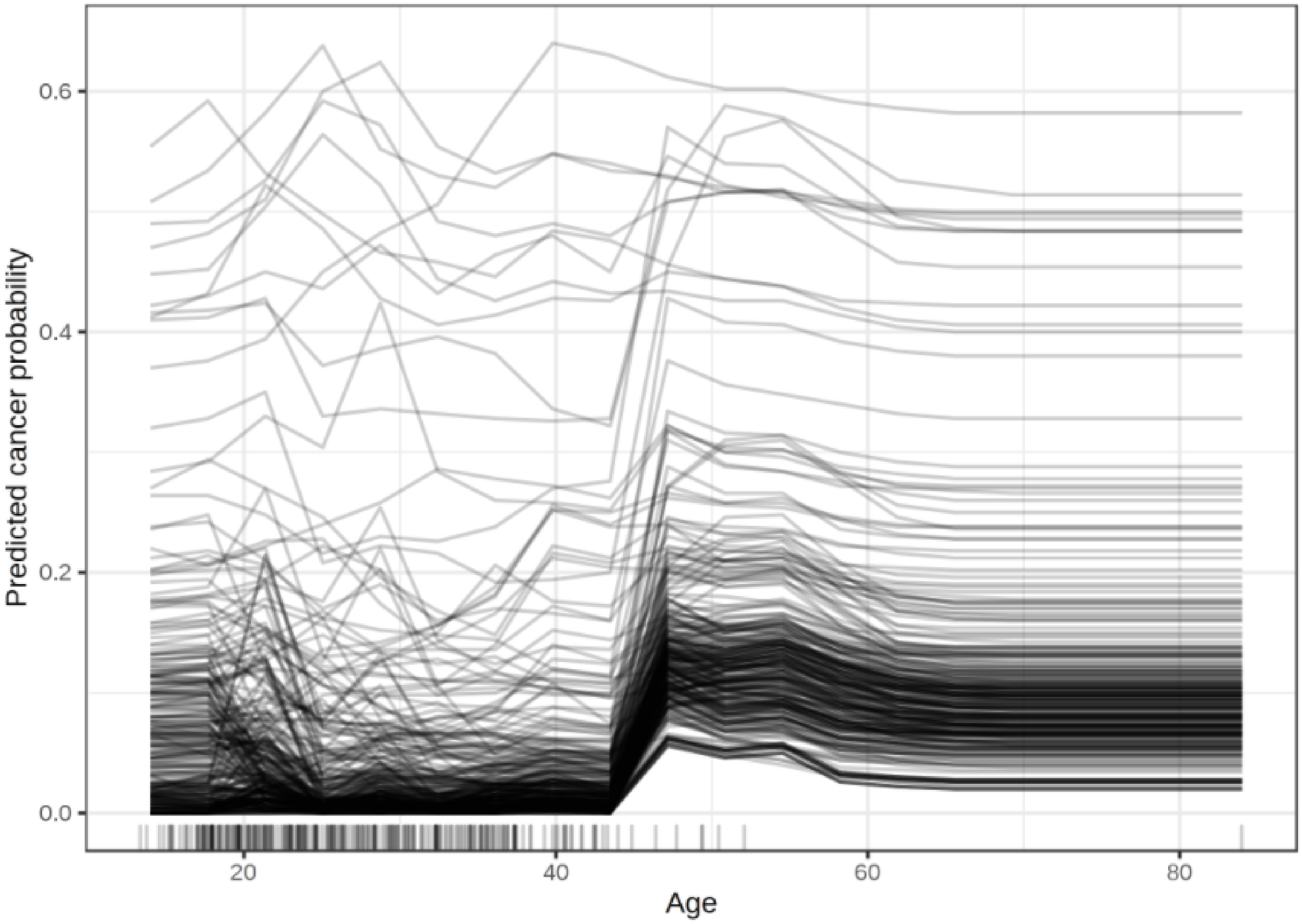


FIGURE 5.6: ICE plot of cervical cancer probability by age. Each line represents one woman. For most women there is an increase in predicted cancer probability with increasing age. For some women with a predicted cancer probability above 0.4, the prediction does not change much at higher age.

Source: Christoph Molnar, Interpretable Machine Learning- A Guide for Making Black Box Models Explainable (https://christophm.github.io/interpretable-ml-book/)

When it comes to explaining more advanced models, model-agnostic (does not depend on the model) techniques are used.

Global surrogate models take the original inputs and your black-box machine learning predictions. When this new dataset is used to train and test the appropriate global surrogate model (more interpretable models such as linear model, decision tree, etc.), it basically tries to *mimic your black-box model's predictions*. By interpreting and visualizing this *"easier"* model, we get a better understanding of how our actual model predicts in a certain way.

Other interpretability tools are **LIME, SHAP, ELI5**, and **SKATER** libraries. We will talk about them in the next post, over a *guided implementation*. Until then, I am sharing some amazing resources I used to form this post along with some extra links. Stay tuned for the next post, see you there!

Happy learning!

REFERENCES

[1] Wikipedia, Explainable AI, https://en.wikipedia.org/wiki/Explainable_artificial_intelligence (https://en.wikipedia.org/wiki/Explainable_artificial_intelligence)

[2] DPhi Tech, Explainable AI Course, https://dphi.tech/lms/learn/explainable-ai/563 (https://dphi.tech/lms/learn/explainable-ai/563)

[3] Dipanjan (DJ) Sarkar, The Importance of Human Interpretable Machine Learning, https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476 (https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476)

[4] Dipanjan (DJ) Sarkar, Model Interpretation Strategies, https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739 (https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739)

[5] Christoph Molnar, Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, 2019, https://christophm.github.io/interpretable-ml-book/ (https://christophm.github.io/interpretable-ml-book/)

Additional Resources

1. DPhi Tech, Importance of Human Interpretable models & Explainable A.I, https://www.youtube.com/watch?v=U92OB_gX7P8&feature=emb_logo (https://www.youtube.com/watch?v=U92OB_gX7P8&feature=emb_logo)
2. Sci-kit Learn Documentation, Partial Dependence Plots, https://scikit-learn.org/stable/modules/partial_dependence.html (https://scikit-learn.org/stable/modules/partial_dependence.html)
3. Yellowbrick Documentation, Feature Importances, https://www.scikit-yb.org/en/latest/api/model_selection/importances.html (https://www.scikit-yb.org/en/latest/api/model_selection/importances.html)

About the Author

Author

Semanur Kapusızoğlu

Recent graduate Industrial engineer aiming for a career in Data Science. Fascinated by how much one can do with data and determined to make an impact using this. Current research area: NLP – Deep Learning

LinkedIn: https://www.linkedin.com/in/semanurkapusizoglu/ (https://www.linkedin.com/in/semanurkapusizoglu/)
GitHub: https://github.com/semanurkps (https://github.com/semanurkps)

The media shown in this article are not owned by Analytics Vidhya and is used at the Author's discretion.

Related Articles

(https://www.analyticsvidhya.com/blog/2018/09/best-ted-talks-artificial-intelligence-must-watch/)
10 Mind-Blowing TED Talks on Artificial Intelligence Every Data Scientist & Business Leader Must Watch
(https://www.analyticsvidhya.com/blog/2018/09/best-ted-talks-artificial-intelligence-must-watch/)

(https://www.analyticsvidhya.com/blog/2020/12/top-ai-startups/)
Top 14 Artificial Intelligence Startups to watch out for in 2021!
(https://www.analyticsvidhya.com/blog/2020/12/top-ai-startups/)

(https://www.analyticsvidhya.com/blo
15-free-data-science-courses-to-kick-start-your-data-science-journey/)
Top 15 Free Data Science Courses to Kick Start your Data Science Journey!
(https://www.analyticsvidhya.com/blog/2020/12/top-15-free-data-science-courses-to-kick-start-your-data-science-journey/)

TAGS : EXPLAINABLE AI (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/EXPLAINABLE-AI/)

NEXT ARTICLE

Coca-Cola Bottle Image Recognition (with Python code)

(https://www.analyticsvidhya.com/blog/2021/01/coca-cola-bottle-image-recognition/)

•••

PREVIOUS ARTICLE

Kaggle Grandmaster Series – Exclusive Interview with 2x Kaggle Grandmaster Prashant Banerjee

(https://www.analyticsvidhya.com/blog/2021/01/kaggle-grandmaster-series-exclusive-interview-with-2x-kaggle-grandmaster-prashant-banerjee/)

(https://www.analyticsvidhya.com/blog/author/guest-blog/)

Guest Blog (Https://Www.Analyticsvidhya.Com/Blog/Author/Guest-Blog/)

LEAVE A REPLY

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

□ Notify me of new posts by email.

SUBMIT COMMENT

POPULAR POSTS

Commonly used Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

Introductory guide on Linear Programming for (aspiring) data scientists (https://www.analyticsvidhya.com/blog/2017/02/lintroductory-guide-on-linear-programming-explained-in-simple-english/)

40 Questions to test a data scientist on Machine Learning [Solution: SkillPower – Machine Learning, DataFest 2017] (https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/)

6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R (https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)

40 Questions to test a Data Scientist on Clustering Techniques (Skill test Solution) (https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/)

45 Questions to test a data scientist on basics of Deep Learning (along with solution) (https://www.analyticsvidhya.com/blog/2017/01/must-know-questions-deep-learning/)

30 Questions to test a data scientist on Linear Regression [Solution: Skilltest – Linear Regression] (https://www.analyticsvidhya.com/blog/2017/07/30-questions-to-test-a-data-scientist-on-linear-regression/)

Understanding Support Vector Machine(SVM) algorithm from examples (along with code) (https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)

CAREER RESOURCES

16 Key Questions You Should Answer Before Transitioning into Data Science (https://www.analyticsvidhya.com/16-key-questions-data-

science-career-transition/?&utm_source=Blog&utm_medium=CareerResourceWidget)

NOVEMBER 23, 2020

Here's What You Need to Know to Become a Data Scientist! (https://www.analyticsvidhya.com/blog/2021/01/heres-what-you-need-to-know-to-become-a-data-scientist/?&utm_source=Blog&utm_medium=CareerResourceWidget)

JANUARY 22, 2021

These 7 Signs Show you have Data Scientist Potential! (https://www.analyticsvidhya.com/blog/2020/12/these-7-signs-show-you-have-data-scientist-potential/?&utm_source=Blog&utm_medium=CareerResourceWidget)

DECEMBER 3, 2020

How To Have a Career in Data Science (Business Analytics)? (https://www.analyticsvidhya.com/blog/2020/11/how-to-have-a-career-in-data-science-business-analytics/?&utm_source=Blog&utm_medium=CareerResourceWidget)

NOVEMBER 26, 2020

Should I become a data scientist (or a business analyst)? (https://www.analyticsvidhya.com/blog/2020/11/become-data-scientist-business-analyst/?&utm_source=Blog&utm_medium=CareerResourceWidget)

NOVEMBER 24, 2020

RECENT POSTS

STANDARDIZED VS UNSTANDARDIZED REGRESSION COEFFICIENT (https://www.analyticsvidhya.com/blog/2021/03/standardized-vs-unstandardized-regression-coefficient/)

MARCH 21, 2021

Common terminologies used in Machine Learning and Artificial Intelligence (https://www.analyticsvidhya.com/blog/2021/03/common-terminologies-used-in-machine-learning-and-artificial-intelligence/)

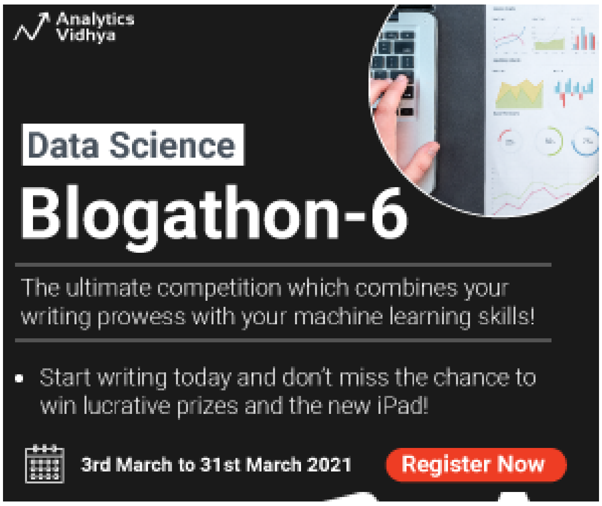
MARCH 20, 2021

Popular Python Data Structures: Comparison & Operations (https://www.analyticsvidhya.com/blog/2021/03/popular-python-data-structures-comparison-operations/)

MARCH 20, 2021

Introducing Machine Learning for Spatial Data Analysis (https://www.analyticsvidhya.com/blog/2021/03/introducing-machine-learning-for-spatial-data-analysis/)

MARCH 20, 2021



(https://www.analyticsvidhya.com/)

Download App

(https://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

(https://apps.apple.com/us/app/analytics-vidhya/id1470025572)

Analytics Vidhya

About Us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-vidhya/)
(https://www.facebook.com/AnalyticsVidhya/)
(https://twitter.com/analyticsvidhya)
(https://www.youtube.com/channel/UCH6gDteHtH4hg3o2343iObA)

Privacy Policy Terms of Use Refund Policy