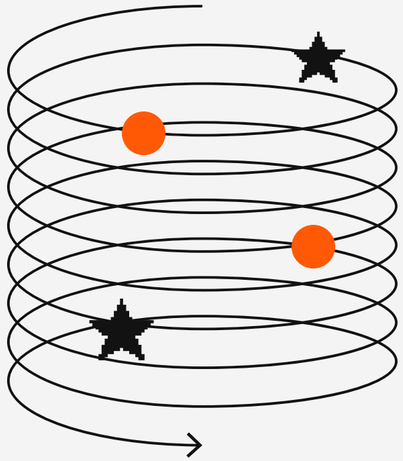
DEMYSTIFYING COMPLEX MODELS WITH SHAPLEY VALUES



@NEERAJP99

EXPLAINABLE AI

#XAI, #SHAP, #LIME &more

Neeraj Pandey

Pycon Taiwan 2023

[TUTORIAL OUTLINE]

Tip to Attendees: Please keep your questions for the end of each section.

TOPICS OF DISCUSSION
Explainable AI.
Significance and Need for Explainable AI.
Model Complexity and Interpretability Trade-off
Explainable AI methods, types
Hands on session - PDP, InterpretML, LIME, SHAP, explainerDashbaord

WHAT IS EXPLAINABLE AI?

Explainable AI focuses on making the behavior and predictions of machine learning systems transparent and understandable to humans, bridging the gap between model complexity and human interpretability."

HEALTHCARE DECISION SUPPORT

Example: Predictive models that determine patient risk for conditions like diabetes, heart failure, or cancer.

Why Explainable AI is Needed: Doctors need to understand why a model makes a certain prediction to trust it and to effectively communicate the risk to patients.

Benefit: Increased trust and actionable insights can lead to early interventions and better patient outcomes.

AUTONOMOUS VEHICLES

Example: Self-driving cars making decisions in real-time about speed, lane-changing, and collision avoidance.

Why Explainable AI is Needed: Engineers and regulators need to understand how the car makes decisions to ensure they comply with safety standards and can be audited.

Benefit: Regulatory approval is more straightforward, and in the event of an accident, it's easier to diagnose what went wrong.

FINANCIAL FRAUD DETECTION

Example: Machine learning models flagging unusual transactions that might indicate fraudulent activity.

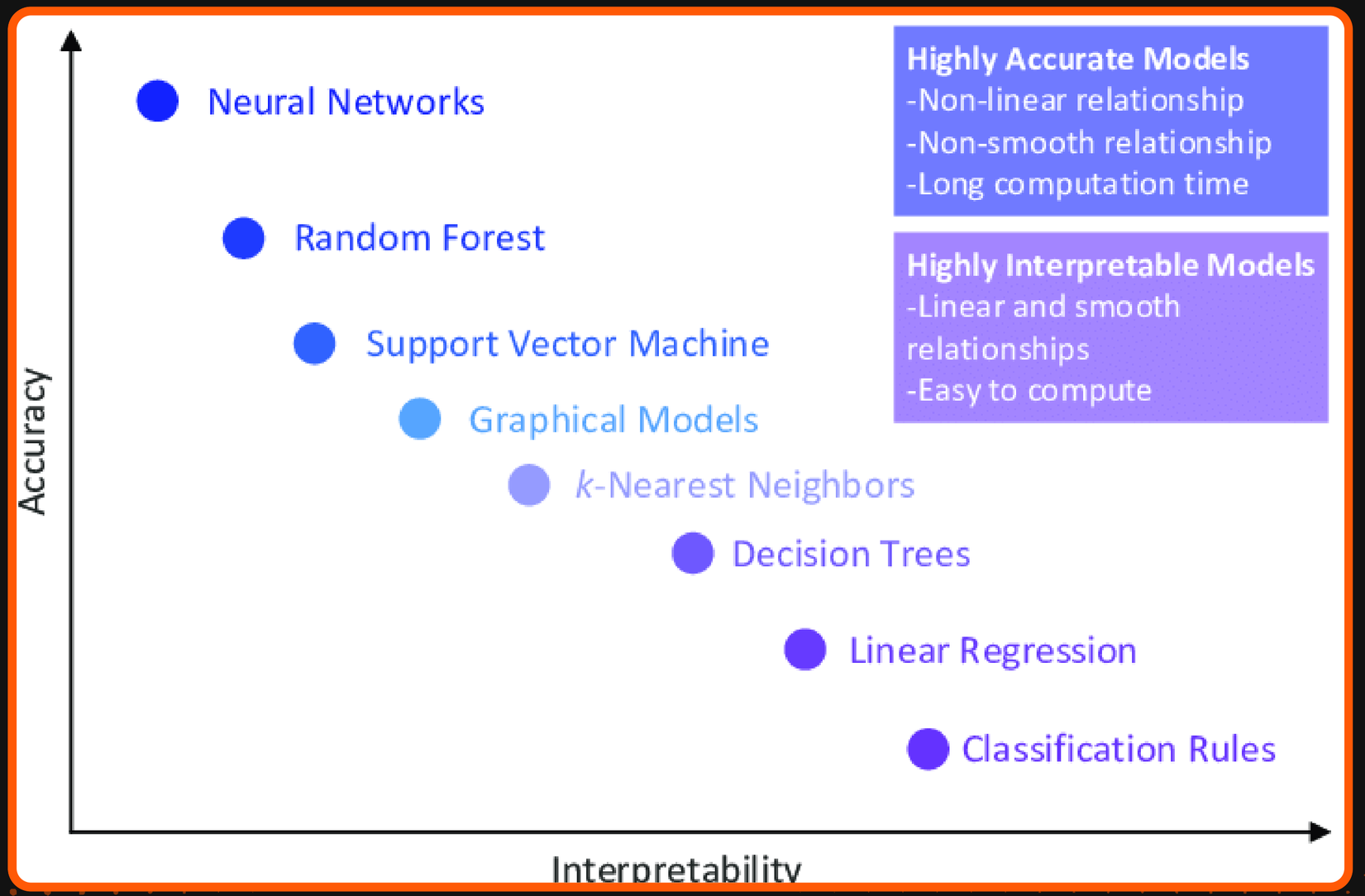
Why Explainable AI is Needed: Financial institutions must be able to explain to both customers and regulators why a transaction was flagged to maintain trust and compliance

Benefit: Improved customer relations and easier compliance with financial regulations.

THE BLACK BOX DILEMMA

CONSEQUENCES OF COMPLEXITY	THE SIMPLICITY SIDE OF THE COIN	NAVIGATING THE TRADE-OFF	POST-HOC INTERPRETABILITY
Deep Learning, Ensemble methods, etc., are highly accurate but hard to interpret.	Linear Regression, Decision Trees are easier to understand but might not capture all nuances in the data.	Some approaches try to balance complexity and interpretability, like ensemble models that combine simple and complex components.	Techniques like LIME and SHAP can be applied to complex models to generate interpretable insights after the model has made a prediction.

The term "Black Box" refers to systems where inputs and outputs are observable, but the internal workings are not transparent or easily understandable


Neural Networks
Random Forest
Support Vector Machine
Graphical Models
k-Nearest Neighbors
Decision Trees
Linear Regression
Classification Rules
Accuracy
Interpretability
Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute

HEALTHCARE - PREDICTING PATIENT OUTCOMES

Scenerio: A healthcare AI model classifies a patient as high-risk for a specific condition but doesn't clarify why.

Black Box Dilemma: Medical practitioners are hesitant to trust the AI's recommendations without understanding its reasoning, and there's potential legal liability.

Consequence: The patient may receive inadequate treatment, or worse, an incorrect treatment, based on an unexplained AI decision.

LOAN APPROVAL

Scenerio: A machine learning model denies a loan to an individual.

Black Box Dilemma:: The individual is entitled to know why the loan was denied, but the model cannot easily provide this explanation.

Consequence: Potential for mistrust, unfairness, and legal complications.

THE IMPERATIVES FOR EXPLAINABLE AI

As AI and machine learning become increasingly prevalent, the importance of Explainable AI has grown to address several key issues: fairness, privacy, reliability, causality, trust, and legal compliance.

Impartiality

Data Confidentiality

Model Stability

Cause-and-Effect

User Confidence

Lawful Accountability

IMPARTIALITY

Spotting Prejudice: How AI might inadvertently adopt biases from data or societal norms.

Universal Access: The need for unbiased predictive models.

eg: Consider a health risk prediction model that inaccurately assesses risk levels for a particular racial group.

EXPLAINABLE AI HELPS IN DETECTING AND MITIGATING BIASES, THEREBY FACILITATING MORE EQUITABLE DECISIONS.

FOR INSTANCE, AN AI TOOL USED IN COLLEGE ADMISSIONS SHOULD EVALUATE CANDIDATES ON ACADEMIC AND EXTRACURRICULAR MERIT, NOT ON DEMOGRAPHICS.

DATA CONFIDENTIALITY

Secure Information: The necessity to protect certain data elements.

Anonymization Methods: How AI can secure sensitive details.

eg: Imagine a healthcare AI that can diagnose diseases while keeping your personal data anonymous.

THE TRANSPARENCY AFFORDED BY EXPLAINABLE AI CAN INDICATE WHICH DATA POINTS THE MODEL USES, ENSURING THAT PRIVATE OR SENSITIVE INFORMATION IS NOT EXPLOITED.

THINK OF A CREDIT SCORING MODEL; YOU WOULDN'T WANT IT TO EXPOSE YOUR FINANCIAL HISTORY.

MODEL STABILITY

Tolerance for Error: Minor input modifications shouldn't result in major output fluctuations.

Uniform behaviour: The value of a dependable model.

eg: Think about Google Search; the algorithm should be consistent even if you misspell a word.

THROUGH EXPLAINABLE AI, WE CAN EXAMINE A MODEL'S RESILIENCE, SHEDDING LIGHT ON HOW VARYING INPUTS INFLUENCE OUTCOMES.

PICTURE A SELF-DRIVING CAR; A SLIGHT SENSOR ERROR SHOULDN'T RESULT IN A DRASTIC CHANGE IN DRIVING BEHAVIOR.

CAUSE-AND-EFFECT

Beyond Correlation: Why understanding causality is important

Informed Predictions: The added value of causal relationships.

eg: In marketing, it's crucial to know if an advertising campaign actually led to increased sales, rather than just correlated with them.

EXPLAINABLE AI CAN HELP DIFFERENTIATE BETWEEN MERE CORRELATIONS AND ACTUAL CAUSATIVE LINKS, BOLSTERING THE MODEL'S RELIABILITY AND PREDICTIVE ACCURACY.

FOR INSTANCE, PEOPLE MIGHT BUY MORE ICE CREAM WHEN CRIME RATES GO UP, BUT BUYING ICE CREAM DOESN'T CAUSE CRIME.

USER CONFIDENCE

AI-Human Synergy: The role of trust for successful interaction.

Openness as Trust Builder: How being transparent can foster trust.

eg: Consider a personalized recommendation engine that can tell you why it thinks you might like a certain product.

TRUST IS MORE EASILY ESTABLISHED WHEN AI SYSTEMS CAN CLEARLY ARTICULATE THEIR REASONING, ESPECIALLY IN CRITICAL SECTORS LIKE HEALTHCARE.

IMAGINE A MEDICAL DIAGNOSIS AI. TRUST WOULD SIGNIFICANTLY IMPROVE IF IT COULD EXPLAIN WHY IT SUGGESTS A PARTICULAR TREATMENT.

LAWFUL ACCOUNTABILITY

Right to Explanation: GDPR and other legal requirements.

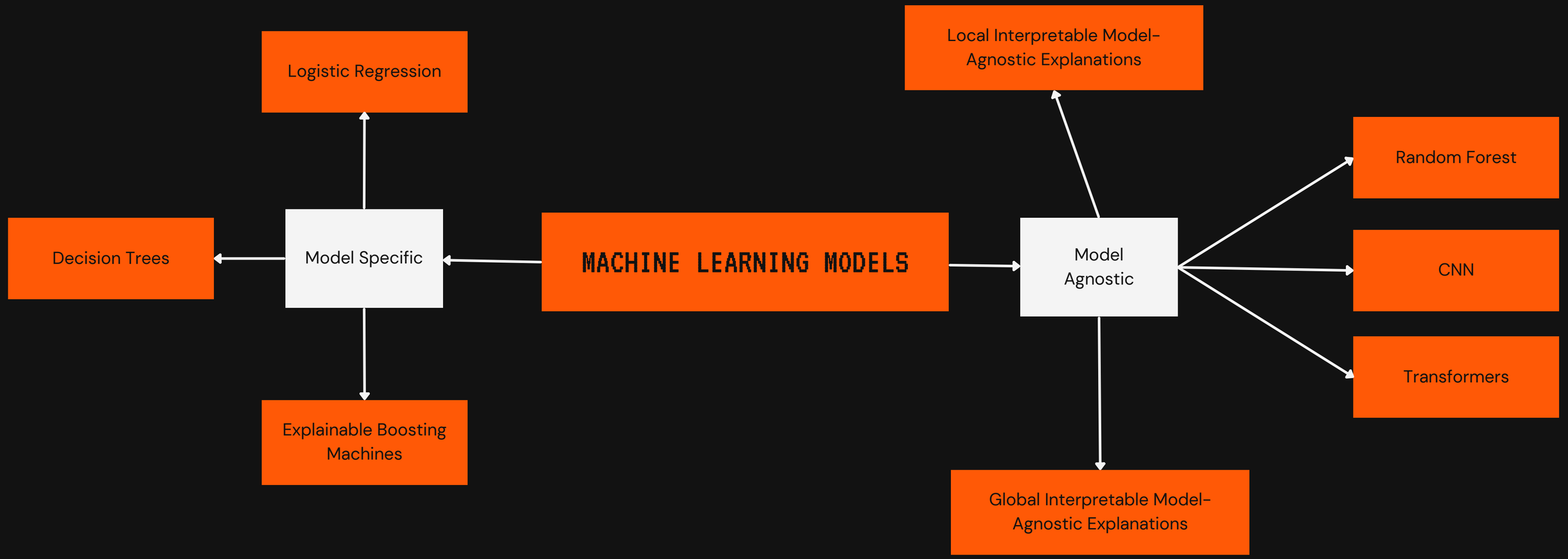
Accountability: Who is responsible for AI decisions?

eg: If an AI-driven drone were to make an incorrect delivery, who's accountable? Explainable AI could provide an answer.

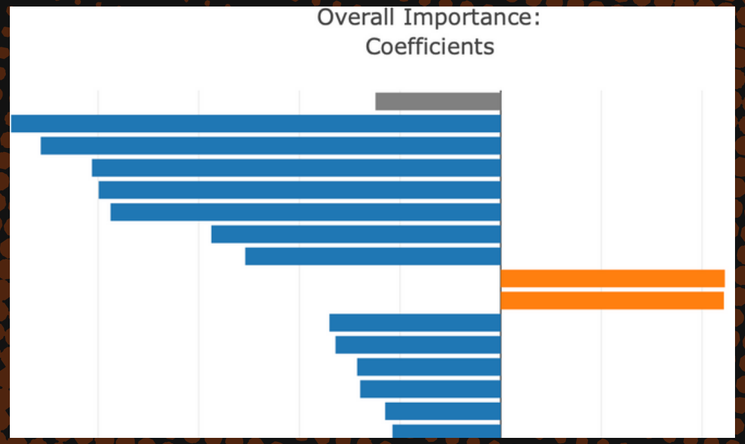
LEGAL FRAMEWORKS LIKE GDPR EMPHASIZE THE "RIGHT TO EXPLANATION," REQUIRING THAT AI SYSTEMS CAN PROVIDE UNDERSTANDABLE REASONING FOR THEIR DECISIONS.

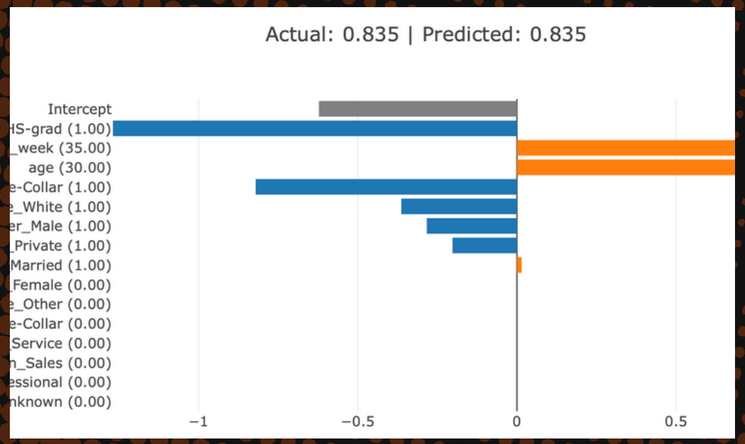
EXPLAINABLE AI ENSURES THAT SYSTEMS CAN BE HELD ACCOUNTABLE, THUS MEETING LEGAL REQUIREMENTS.

TYPES OF EXPLAINABLE AI METHODS



MODEL SCOPES





Global Explanations: These provide an overview of how a model makes decisions based on all the data it has been trained on. These are useful for understanding the overall behavior of the model.

Local Explanations: These are focused on individual predictions. They explain why the model made a specific prediction for a given data point.

GLOBAL EXPLANABILITY

Feature Importance: Techniques like permutation feature importance give a global view of which features are most important for a model's predictions.

Partial Dependence Plots (PDPs): These also offer a global view of the model by showing how individual features affect predictions across different data points.

Linear or Logistic Regression with Regularization: These models are inherently interpretable and can offer a global perspective on feature importance through their coefficients.

LOCAL EXPLANABILITY

LIME (Local Interpretable Model-agnostic Explanations): As the name suggests, LIME is geared for local explanations.

Counterfactual Explanations: These offer local insights by showing how a specific input could be minimally changed to alter its prediction.

SHAP (Shapley Additive Explanations) Values for Individual Predictions: Though SHAP can also be used for global explanations, it can break down a specific prediction to show the impact of each feature.

MODEL SPECIFIC

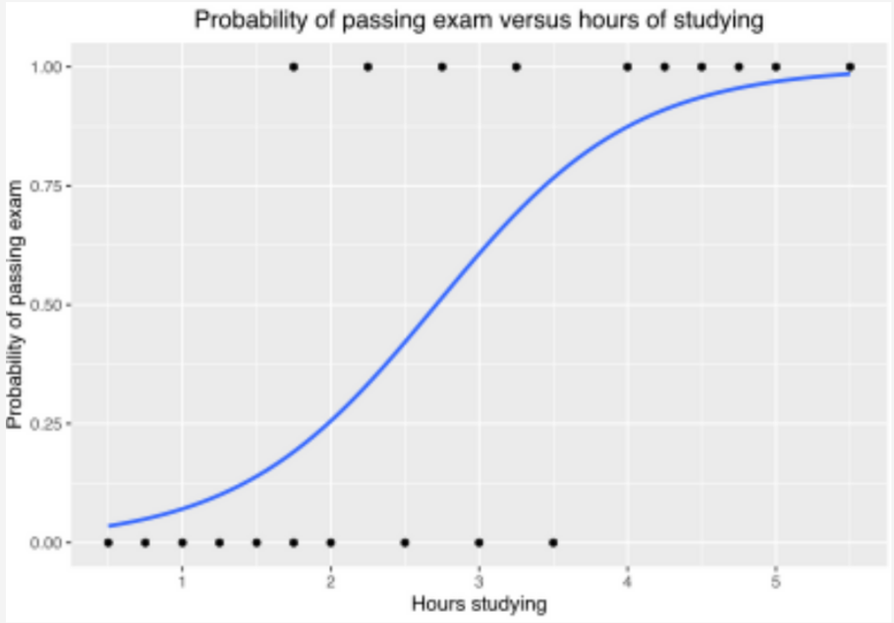
Example models that commonly benefit from model-specific approaches include Logistic Regression, Decision Trees, and Explainable Boosting Machines (EBMs).

Model-specific approaches are tailored to work with particular types of machine learning models.

Unlike generic interpretability methods, which aim to provide explanations for a wide range of models, model-specific approaches delve into the internal mechanics of specific models to generate more accurate and insightful explanations.

The benefit of using a model-specific approach is the ability to leverage the unique characteristics and mechanisms of a given model, thereby often yielding more precise and trustworthy interpretations.

Logistic Regression



Inherently Interpretable

Parameters in logistic regression are directly tied to feature importance.

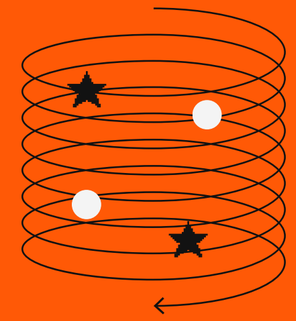
Probabilistic Output

Predicts the probability of an event, making it interpretable.

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + \ldots + b_n x_n$$

Coefficients as Importance Indicators

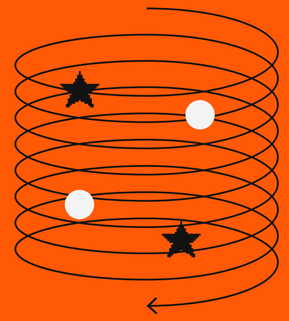
Positive and negative coefficients show the effect of features on the outcome.



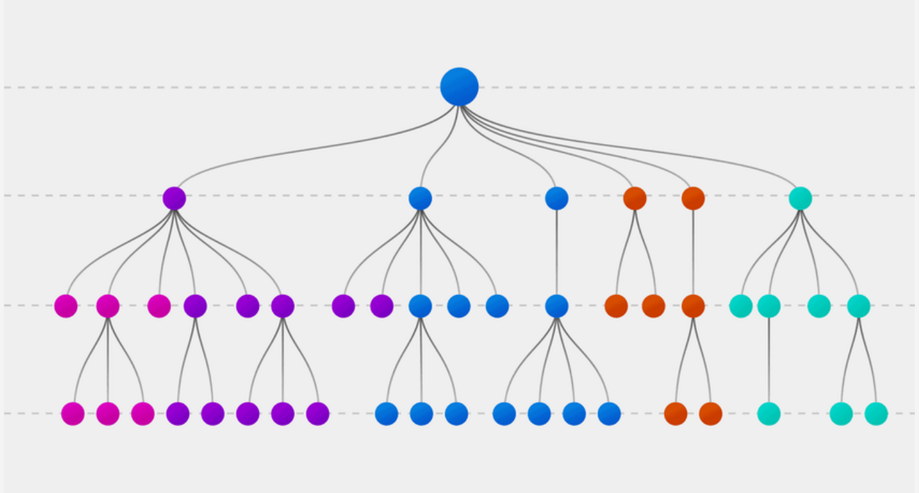
Logistic Regression

Predicting Customer Churn Scenario

A telecom company wants to predict which customers are likely to leave for a competitor.

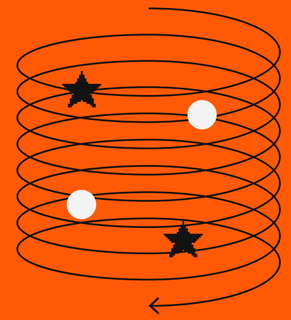


Decision Trees



Transparent Rules

Each path in the tree is essentially a 'rule' that leads to a decision.



Hierarchical Decision Making

Breaks down complex decisions into simpler decisions.

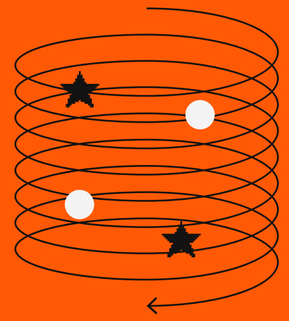
Easy to Visualise

Can be easily plotted and visualized for non-experts.

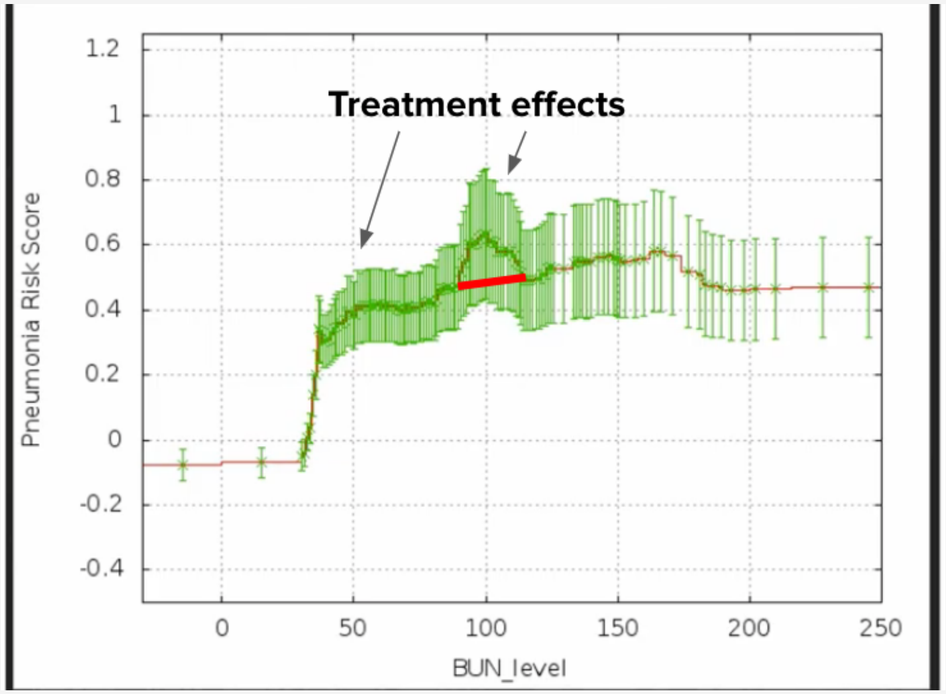
Decision Trees

Credit Risk Analysis

A bank wants to decide if they should offer a loan to individuals.



Explainable Boosting Machines



Generalized Additive Models (GAMs)

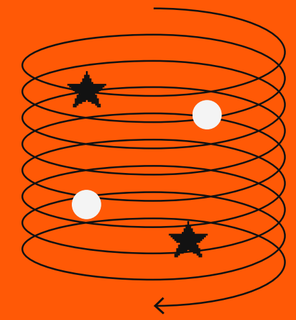
EBMs extend GAMs with modern ML capabilities.

Flexible and Interpretable

Allow non-linear feature interactions but maintains interpretability.

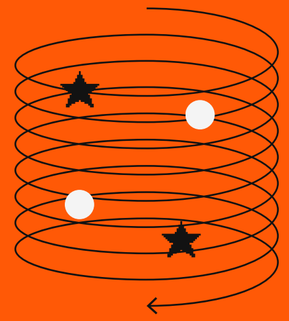
Visual Aids

Provides plots that help you see feature importance and interactions.



Explainable Boosting Machines

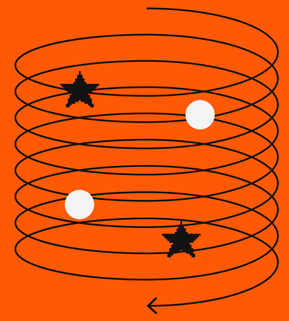
1. **Initialization:** Start with basic predictions, usually based on class prevalence.
2. **Feature Scanning:** Examine each feature to find the best simple model, like a one-level decision tree.
3. **Boosting Step:** Add the simple model to an ensemble, scaling it down with a learning rate.
4. **Residual Update:** Update prediction errors to guide the next iteration.
5. **Iterate:** Repeat steps 2-4, aiming to reduce prediction errors.
6. **Stop:** Cease iterations upon meeting stopping criteria or after a set number of rounds.



Explainable Boosting Machines

Healthcare Risk Prediction

Predicting the likelihood of a patient getting a particular disease.



MODEL AGNOSTIC

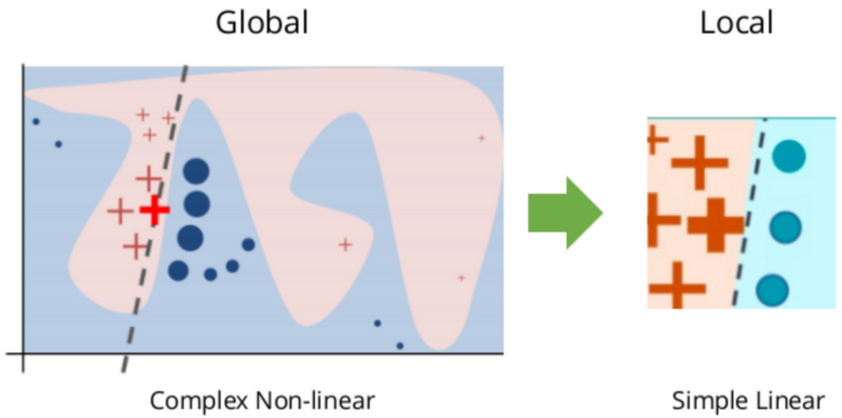
Popular model-agnostic methods include LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapley Additive Explanations), and Partial Dependence Plots (PDPs)

Model-agnostic approaches are designed to interpret any machine learning model without requiring access to their internal workings or data representations.

Unlike model-specific approaches, which are fine-tuned for a particular type of model, model-agnostic methods provide a generalized way to interpret predictions, making them highly versatile.

These approaches operate independently of the model's complexity or type, allowing for wide applicability. They can interpret anything from simple linear models to complex neural networks.

Local Interpretable Model-agnostic Explanations



LIME

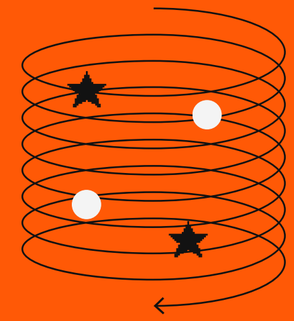
Based on locally approximating the decision boundary of the complex model. i.e - It approximates a black-box model with a simpler, interpretable model for individual predictions. For instance, in an image classification task, LIME can highlight regions in an image most responsible for a classification.

1. Sample Generation

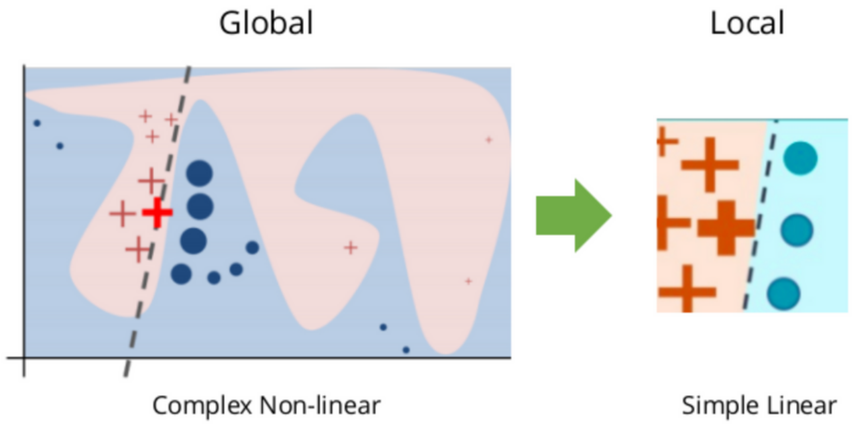
Create variations of the original input to see how predictions change.

2. Prediction on Perturbed Samples

Run these variations through the model to get new predictions



Local Interpretable Model-agnostic Explanations cont..



3. Weight Assignment

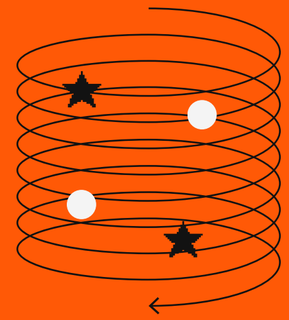
Assign weights to variations based on closeness to original input.

4. Model Training

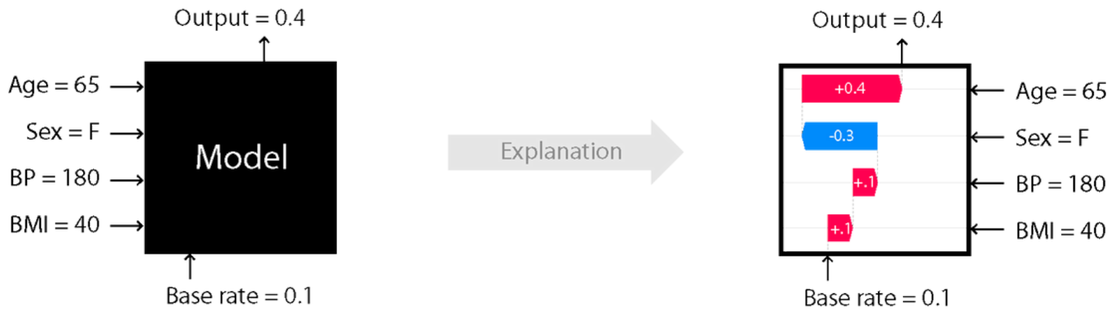
Use weights and predictions to train an easier-to-understand model.

5. Interpertation

The coefficients of the interpretable model are then used to explain the prediction of the complex model for the original input data.



SHapley Additive exPlanations

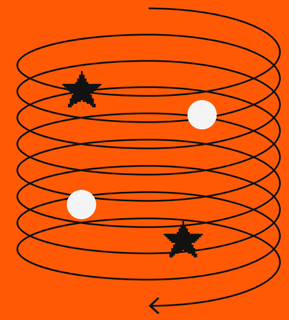


Cooperative Game Theory

In cooperative games, a set of players collaborate and obtain a certain gain (or cost). The question then is: how should the gain be distributed among the players?

Shapley Value

SHAP value is a solution from game theory which gives a unique distribution of gains to players. In the context of machine learning, each feature is considered a "player" in the game that collaborates to produce a prediction. The Shapley value allocates to each feature (player) an importance value for the prediction.



SHapley Additive exPlanations

TWO PLAYER GAME

A = 10, B = 20

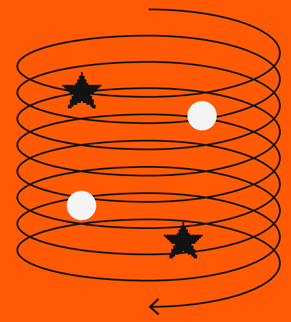
A, B (TOGHETHER) = 35

Prediction Game

The prediction for an instance minus the average prediction for all data is the "payout". Features "collaborate" to achieve this payout.

1. Compute Shapley Value

For a given feature, its marginal contribution is the additional payout achieved when adding that feature to a subset of features.



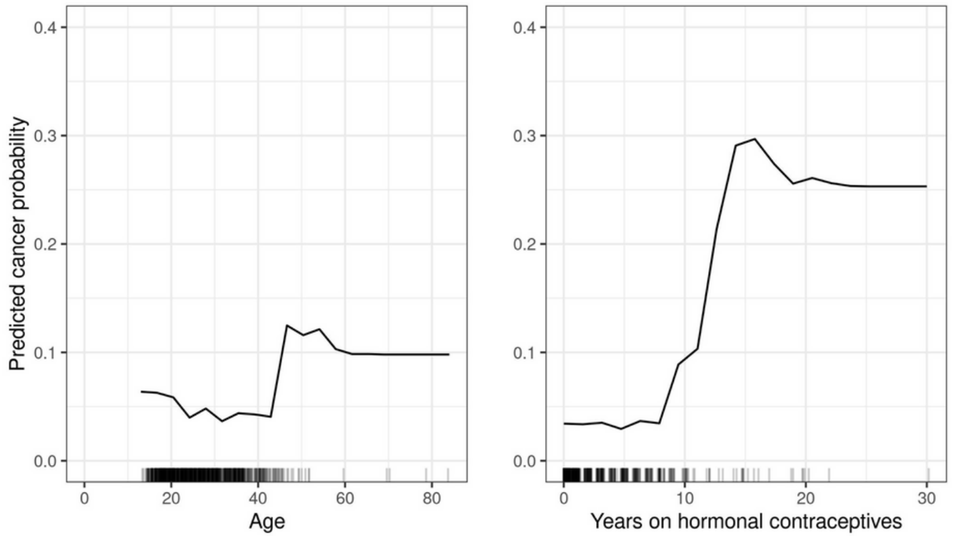
2. Local Explanations

Shapley values explain individual predictions.

3. Global Explanations

Aggregate Shapley values for a global understanding.

Partial Dependence Plots



PDP

Graphical representation that shows how a feature affects predictions.

1. Data Sampling

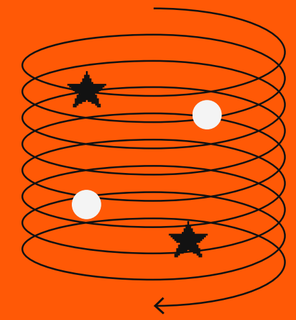
Choose a grid of values for the target feature.

2. Prediction

Make predictions by holding the target feature constant.

3. Plotting

Create plots for visual explanations



THANK YOU FOR JOINING.

PYCON TAIWAN 2023

@neerajp99