Professional Elective-IV (Group F)

Explainable Artificial Intelligence (**AI374TFB**)

Explanation

Very Purpose

- “Explanation” has to do with reasoning and making the reasoning explicit.
- Characteristics of Good Explanation??

- Clarity, Relevance, Accuracy, Completeness, Causality, Simplicity, Understandability, Coherence, Transparency, Engagement

Explainable
Very Purpose

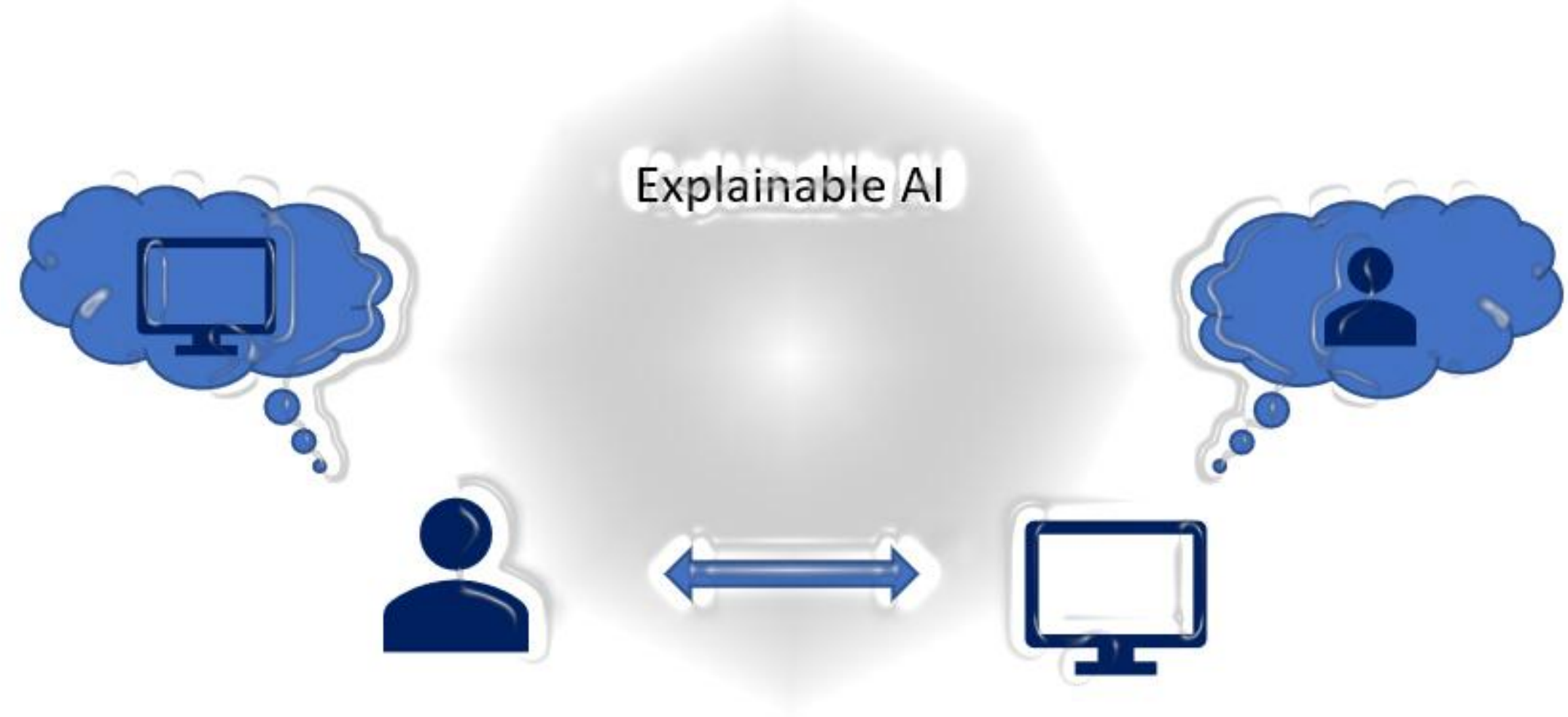
- Explainable could mean *interpretable*

- Interpretability may even be more important than explainability

AI

- AI is conditional probability for the most part.
- Achieving dependable output(s) is set by the threshold(s) you choose for your application.
- Some thresholds can be machine-determined, **which will be more common in the future.**
- **Healthcare/Banking/AVs/E-Commerce/Manufacturing/Etc.**

Explainable AI



Explainable AI

Identify the Hidden Rule and Explain

Input String	Output
“aioua”	A
“smart”	B
“boat”	A
“flight”	B
“strong”	C
“echo”	A

If a string has more vowels than consonants **and** length < 6 → Output A,
Else if it ends with ‘t’ → Output B,
Else → Output C.

Explainable AI

Identify the Hidden Rule and Explain

Income (₹L)	Credit Score	Area	Output
7	680	Zone-A	Approve
6	720	Zone-C	Deny
8	650	Zone-B	Approve
5.5	690	Zone-C	Deny
9	720	Zone-A	Approve
4.5	750	Zone-B	Deny

If income > 5L **and** credit score > 650 → Approve,
but if area = Zone-C, Deny regardless of score

Explainable AI

Identify the Hidden Rule and Explain

Temperature (°C)	Humidity (%)	Wind Speed (km/h)	Day	Output
35	90	40	Monday	Storm Risk
30	80	20	Sunday	No Risk
40	70	30	Wednesday	Storm Risk
25	85	50	Friday	No Risk
32	88	25	Saturday	Storm Risk
36	60	20	Sunday	No Risk

If (temperature × humidity / 100) + (wind_speed / 10) > 60 **and** day ≠ Sunday → Output **Storm Risk**, else **No Risk**.

Explainable AI
Key Objectives

- Makes AI models more _____ and _____ to humans
- Makes AI models more **interpretable** and **understandable** to humans
- **Interpretable AI**: Traceability and Justification
- **Explainability Tools**: LIME(Local Interpratable Model-Agnostic Explanations), DHAP(Shapley Additive exPlanations), What-If Tool, Model Cards, etc.

Explainable AI
Key Objectives

- Defines ____, _____, _____ and _____ results

- Defines **model correctness, fairness, transparency** and **decision-making** results

- Amazon Hiring Algorithm Bias (favored male candidates)
- Apple's Credit Card (Higher credit limits to men)
- Facial Recognition (Higher error rates for Darker-skinned people)
- PredPol (Biased to certain neighborhoods and ethnic groups)

Explainable AI
Key Objectives

- Helps organizations to gain the _____ of stakeholders
- Helps organizations to gain the **confidence** of stakeholders
- Google Photos Tagging Error (2015): Incorrectly labeled Black people as Gorillas
- LLMs: “The nurse said ___ went to get the patient’s chart.”

  “The engineer said ___ fixed the machine.”

Overfitting to linguistic patterns instead of verified data.

Explainable AI
Key Objectives

- Aids in organizations' adoption of a _____ development of AI

- Aids in organizations' adoption of a **responsible** development of AI

Responsible AI
Examples

- Microsoft Azure Cognitive Services (like Face API) introduced strict data-handling, bias reduction, and limited access to facial recognition for sensitive use cases.
- BMW created a *Responsible AI Policy:* AI systems for driver assistance are tested under ethical review boards, ensuring transparency and human oversight.
- Mastercard's AI models for fraud prevention undergo bias auditing and explainability validation.
- IBM introduced transparent AI tools to assist clinicians with treatment options, always showing reasoning paths and data sources.
- Google's *What-If Tool* and *Model Card Toolkit* enable developers to visualize bias and explain decisions.

Explainable AI
Key Objectives

- Vital to satisfy ____ requirements

- Vital to satisfy **legal** requirements

  - Copyright infringement – Using lyrics to generate responses
  - Use of pirated books for LLM training
  - Using shadow libraries for the LLM training
  - Scraping the original works of Artists for Image generation
  - Fakecase citations in law assistants

Who needs Explainability?

Practitioners

Data scientists and engineers who are familiar with machine learning

Observers

Business stakeholders and regulators who have some familiarity with machine learning but are primarily concerned with the function and overall performance of the ML system

End users

Domain experts and affected users who may have little-to-no knowledge of ML and are also recipients of the ML system's output

Challenges in Explainability

- How do I know this explanation is accurate?
- What does it mean that one pixel is highlighted more than another?
- What happens if we remove these pixels?
- Why did it highlight these pixels instead of what we think is important over here?
- Could you do this for the entire dataset?

Challenges in Explainability

- Demonstrating the semantic correctness of explanation techniques above and beyond the theoretical soundness of the underlying mathematics
- Combining different explanation techniques in an easy and safe way that enhances understanding rather than generating more confusion
- Building tools that allow consumers to easily explore, probe, and build richer explanations
- Generating explanations that are computationally efficient
- Building a strong framework for determining the robustness of explanation techniques

An Overview of Explainability

The Model Works really well, but how?

Practitioners' focus was to ensure that the ML is working as expected, rather than providing an explanation

What are Explanations?

Figuring out why can help in building a better comprehension of what influences a model, and how the influence occurs

Pure explanations are often unsatisfactory ??

A **pure explanation** is an explanation that **states the reason for an outcome — without comparing it to alternatives** or showing **what could have been different**.

Answers only: Why did this happen? But not "Why did this happen **instead of something else**?"

What are Explanations?

Pure explanation: “You got a B because you scored 80.”

“You needed 90 to get an A; improving your essay clarity could raise your score.”

What are Explanations?

Counterfactual explanations for the original situation

A **pure explanation** only tells the **factors that influenced the model's decision**, while a **contrastive or counterfactual explanation** tells *how to change the outcome*.

A **pure explanation** gives the *reason* for an event,
While a **contrastive** or **counterfactual explanation** gives the *context* or *alternative*.

What are Explanations?

Causal Explanations are Logical but not always true

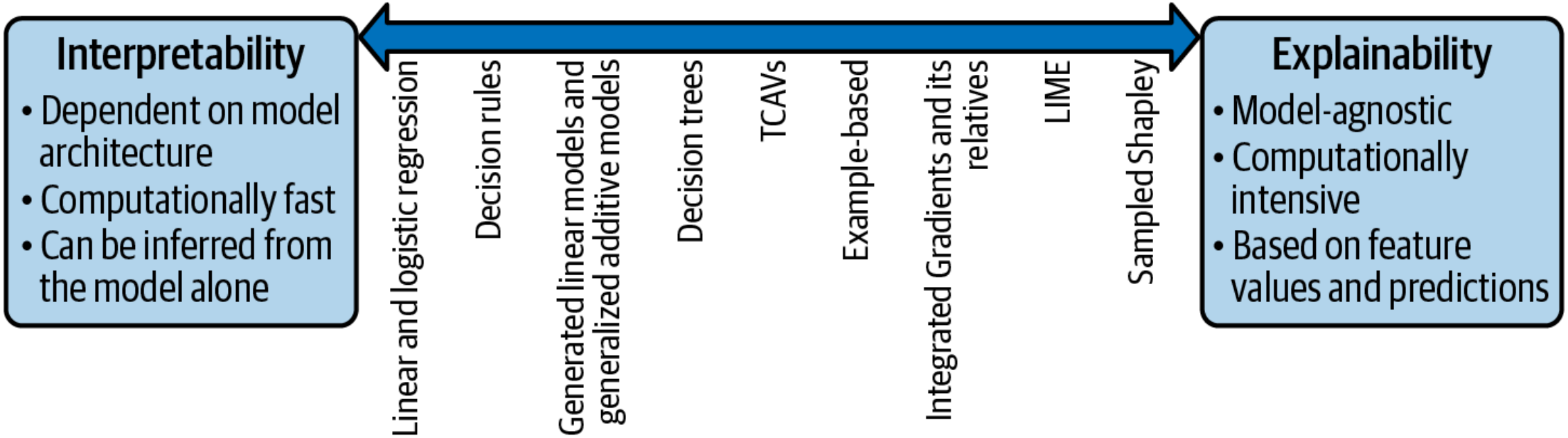
The rooster’s crow causes the sun to rise

I passed the exam because I used my lucky pen.

The model predicts success because the candidate is male.

The model predicts an employee is high performing because he/she is from a top university

Interpretability and Explainability



Interpretability and Explainability

A downside of XAI is reliance on predictions, which are more computationally intensive than explainability

Explanations can be much slower than a single prediction

It is a limitation in real-time systems

Solutions: Approximate XAI methods with limited samples

Precompute explanations for common input clusters

TCAV (Testing with Concept Activation Vectors)

- **Attendance** (% of classes attended)
- **Homework completion** (average % completed)
- **Participation in group activities** (Low/Medium/High)
- **Past internal assessment scores**
- **End-of-semester exam scores** (target variable)

Your task is to use **TCAV** to analyze how the following **concepts** influence predicted grades:

1. **Regular Attendance** – students with >90% attendance
2. **Homework Completion** – students completing >80% assignments
3. **Active Participation** – students frequently engaging in class or group work

Explainability Consumers

- Limitation: It is not of type Conversational
- Scope for Smart Explainability Techniques – adaptive type
- Practitioners – Use explainability to improve the model performance
- Observers-
  - stakeholders->nontechnical explanation->business focused questions
  - Regulators->Adherence to criteria and standards
- End Users
  - Domain Experts->for decision support->quality control
  - Affected users-> little or no understanding of how the model works

Types of Explanations

- Explainability can be at any stage of the ML pipeline
- Data models -> Types of explanations
- Explanations
  - Single Prediction type
  - Set of Predictions -> Overall model behavior

Premodeling Explainability

- Depends on understanding the data before the model development
- Depends on the feature attribution
- Exploratory Data Analysis (EDA) -> Statistics and Visualization Tools
- Used to get statistical properties of the dataset
- Multivariate Analysis: Statistical relationships analysis between features and the target

Comment on the Advantages and Limitations of this type of Explainability

Intrinsic Versus Post Hoc Explainability

- Intrinsic: Part of model's prediction->Model itself explains it's decisions
- Post Hoc: After the model has been trained
- Intrinsic: Change in the model->inherently interpretable models

$$\text{Pred}(y) = 0.1 \cdot \text{feature}_A + 0.7 \cdot \text{feature}_B + 0.4 \cdot \text{feature}_C$$

Comment on the Advantages and Limitations of each of this type of Explainability

Identify which type of explainability is suitable for the following cases and why?
Loan Approval, Medical diagnostic Support, Detecting cracks in ships, Student grade prediction, Predictive maintenance of equipment's

Local, Cohort and Global Explanations

- Local: Context and predictions for a **single** prediction -> Very detailed -> -> **Example?**
- Global: Overall behavior of the model across **all** predictions -> Broad -> **Example?**
- Cohort: **Subset** of predictions -> Intermediate -> **Example?**
- It’s not safe to believe the Local explanations
- One cohort can be compared with another cohort

Attributions, Conterfactual and Example-Based Explanations

- Attributions-based: Highlights relevant properties of the system
- Example-based: gives an analogous scenario
- Counterfactual: “It does not rain when it is sunny”
  - Clouds: Proponent of Counterfactual explanations
  - Sunny: Opponent of Counterfactual explanations
- Finding the causes behind proponents and opponents depends on the data modality

Classroom Exercise

Case Study**: Fair Credit – Understanding Bias in Loan Approval Models**

Use **Microsoft InterpretML** to:

- Build a transparent machine learning model for loan approval.
- Explore global and local explanations.
- Compare model behavior across cohorts (age groups, gender, income level).
- Identify potential bias and interpret the model's fairness.

Dataset: UCI Credit Approval Dataset

Themes throughout Explainability

Feature Attributions

Deals with the influence of a feature on the prediction

Feature attribution can be absolute -> Predicting the house price based on location

Taking absolute value of a feature -> makes feature is always important -> neglecting the negative influence of the feature -> **Example??**

Feature attributions can also be relative, representing a percentage of influence compared to other features used by the model -> depends on normalization -> change in one attribute would effect all other percentages -> **Example??**

Shapley Values

Root is in Game Theory (Lloyd Shapley, 1953)-> Cooperative Game Theory -> To compute the contribution of each player on the overall outcome of the game

In 2017, Scott Lundberg and Su-in Lee adapted Shapley values to the needs of machine learning with the idea of a SHAP value as computed by kernelSHAP

Used to attribute the influence of each feature on overall prediction->the Shapley value is the average expected marginal contribution of one player after all possible combinations have been considered

The Shapley value gives a fair share of the total prediction to each feature, considering all possible combinations of features.

It’s calculated by averaging the feature’s marginal contribution across all possible feature subsets.
Computational cost grows exponentially as variables and observations are added

Maths behind Shapley Values

Assume teamwork is needed to finish a project. The team, *T*, has *p* members.

Knowing that the contribution of team members during the work was not the same, how can we distribute the total value achieved through this teamwork, *v=v(T)* , among team members?

Shapley value, $\varphi_m(v)$, is the fair share or payout to be given to each team member m. The $\varphi_m(v)$ is defined as

$$\varphi_m(v) = \frac{1}{p}\sum_{S}\frac{[v(S \cup \{m\}) - v(S)]}{\binom{p-1}{k(S)}}, \qquad m = 1,2,3,\ldots p$$

Maths behind Shapley Values

For a given member, *m*, the summation is over all the subsets *S*, of the team, *T*={1,2,3,…,p}, that one can construct after excluding m.

In the formula, *k(S)* is the size of *S*, *v(S)* is the value achieved by subteam *S*, and *v*(*S*∪{*m*}) is the realized value after *m* joins *S*.

$$\varphi_m(v) = \frac{1}{p}\sum_{S} \frac{[v(S \cup \{m\}) - v(S)]}{\binom{p-1}{k(S)}}, \qquad m = 1,2,3,\ldots p$$

Maths behind Shapley Values

1. Efficiency: The total of individual contributions is equal to the team's realized value:

$$\sum_{m=1}^{m=p} \varphi_m(v) = v(T)$$

2. Symmetry: Two team members with the same added value have the same share:

if $v(S \cup \{m\}) = v(S \cup \{n\})$, then $\varphi_m(v) = \varphi_n(v)$

3. Linearity: If the team participates in several projects (say two projects), each yielding $v(T), u(T)$, then adding the share of each member in the different projects is the same as finding his/her share using the total gain $v(T)+u(T)$.

The shares are additive:

$\varphi_m(v + u) = \varphi_m(v) + \varphi_m(u)$

And, $\varphi_m(av) = a\,\varphi_m(v)$

Example

Model predicts House Price (in ₹ thousands) using:
- Feature 1: Income (I)
- Feature 2: Credit Score (C)

Model outputs: None: {Ø}:100, {I}: 130, {C}: 120, {I,C}:160

Compute Shapley Value for Income (I)

Subsets not containing I: Ø, {C}
Marginal contributions: f({I}) – f(Ø) = 30, f({I, C}) – f({C}) = 40
Weights = ½ each
ϕ(I) = (½ × 30) + (½ × 40) = 35
Income adds ₹35k to the prediction.

Compute Shapley Value for Credit Score (C)

Subsets not containing C: Ø, {I}
Marginal contributions: f({C}) – f(Ø) = 20, f({I, C}) – f({I}) = 30
Weights = ½ each
ϕ(C) = (½ × 20) + (½ × 30) = 25
Credit Score adds ₹25k to the prediction.

Example

Verify Additivity

Baseline + Shapley values = Full Prediction
100 + 35 + 25 = 160
Model explanation is consistent.

Relative Influence

Total contribution above baseline = 60 (35 + 25)
Income: 35 (58.33%)
Credit Score: 25 (41.67%)
Income contributes more strongly to the price prediction.

Interpretation

Baseline house price = ₹100k
Income raises price by ₹35k
Credit Score raises price by ₹25k
Final predicted price = ₹160k

Shapley values show fair, additive influence distribution.

Gradient-based Techniques

Deep learning models are differentiable

Built using mathematical functions that have well-defined derivatives

Learning from gradients: The computed gradients tell the model how to adjust each weight to reduce the error

Derivative measures the rate of change of a function at a point

Measuring the gradient of a model function with respect to its inputs gives valuable information as to how the model predictions may change if the inputs change as well.

Gradient-based methods are particularly common for image-based models -> filters are **optimized using gradients**

Gradients can be computed in parallel across millions of pixels using GPUs, making it feasible to train large image models.

Saliency Maps and Feature Attributions

Saliency maps, broadly, refer to any technique that aims to determine particular regions or pixels of an image that are somehow more important than others.

The MIT/Tuebingen Saliency Benchmark dataset is a benchmark dataset for eye movement tasks

Saliency methods serve a similar purpose; that is, they emphasize the regions or pixels of an image that indicate where a trained model focuses its attention to arrive at a given prediction

Doctors and regulators need to understand *why* an AI model predicted a certain diagnosis.

Highlighting **regions of X-rays, MRIs, or CT scans** that led to a disease prediction

Security and Surveillance: AI systems must explain what they're detecting to ensure fairness and avoid bias.

Think of examples: Agriculture, Manufacturing, Document Processing, …

Surrogate Models

The *surrogate*, to give more explanatory power directly from observing the architecture of the model

A surrogate model is a simplified or approximate model that is used to mimic or substitute a more complex or expensive model.

A surrogate provides a faster or interpretable approximation for tasks such as optimization, sensitivity analysis, or model explanation.

Design optimization-> Fit a surrogate (e.g., Gaussian Process, Polynomial Regression, Neural Net) to approximate the true function.

BMW uses surrogate-assisted optimization for vehicle crash simulations and engine performance tuning.

Siemens uses surrogate models to optimize turbine efficiency.

In drug discovery, they approximate outputs of molecular dynamics simulations to predict binding affinity.