**SHAP in Simple Terms**

SHAP values come from **game theory (Shapley values)**.

* Imagine a "game" where the **players** are features (Income, Credit Score, Debt Ratio, etc.).
* The **payout** is the model’s prediction (e.g., loan approval probability).
* SHAP measures **each feature’s fair share of the payout** by looking at how predictions change when you include/exclude that feature in different combinations.

**🔹 2. Model Specificity**

The SHAP value depends on the **model you built**:

* A decision tree, random forest, or gradient boosting model will split features differently, so the contribution of *Income* will vary.
* A neural network might scale Income differently (normalised input), so its influence will be calculated on that transformed scale.
* That’s why you cannot reuse SHAP values across models — they’re **model-specific**.

**🔹 3. Dataset Dependence**

The **baseline prediction** (starting point) comes from the **dataset average**.

* Example: In your dataset, the average loan approval probability might be **0.55** (55%).
* This is the "reference point" if we knew nothing about a specific applicant.

When we add *Income* into the calculation for Applicant A:

* Their predicted probability might shift from **0.55 → 0.80**.
* SHAP assigns the difference (+0.25) to the Income feature.

**🔹 4. Scaling / Normalisation**

Many models don’t use raw values directly (e.g., “Income = $120,000”).

* Features are often **scaled** (e.g., min-max scaling between 0 and 1, or z-score standardisation).
* SHAP works on the **scaled inputs** as the model sees them, not the raw business numbers.
* However, for documentation, most teams present SHAP contributions back in a **human-readable format** (e.g., “Income level contributed +0.25”).

**🔹 5. Putting it Together (Example)**

Let’s say:

* **Baseline prediction** (average applicant): 0.55 (55% chance of approval).
* Applicant’s profile fed into model gives **final prediction**: 0.80 (80%).
* SHAP explains the difference (0.80 – 0.55 = +0.25) across features.

For this applicant:

* Income (high) → +0.25
* Stable Employment → +0.10
* High Debt Ratio → -0.05
* Credit Score (excellent) → -0.05

So:  
0.55 (baseline) + 0.25 + 0.10 - 0.05 - 0.05 = **0.80 (final prediction)**

**✅ Why these matters in Governance**

* SHAP values are **not arbitrary scores**; they are mathematically derived contributions.
* They depend on:
  1. The **model type** (tree, neural net, logistic regression, etc.).
  2. The **dataset baseline** (average behaviour).
  3. Any **data transformations/scaling** applied during training.

That’s why SHAP documentation in regulated industries always includes:

* **Model name/version**
* **Dataset used**
* **Baseline probability**
* **Feature contributions (SHAP values)**

**SHAP Documentation Report – Example**

**Model Name:** Credit Risk Model V2  
**Business Purpose:** Loan approval decision support  
**Date:** 2025-09-09  
**Model Owner:** Data Science Lead – Retail Banking

**1. Overview of Prediction**

* **Prediction Result:** Loan Rejected
* **Customer ID:** 87453 (internal reference)
* **Prediction Probability:** 0.32 (below approval threshold of 0.6)

**2. Feature Contributions (SHAP Values)**

| **Feature** | **SHAP Value** | **Impact on Prediction** |
| --- | --- | --- |
| Credit Score | -0.35 | Strong negative |
| High Debt Ratio | -0.40 | Strong negative |
| Employment History | +0.20 | Positive |
| Income Level | +0.25 | Positive |

**3. Plain Language Explanation**

The loan application was rejected because the applicant’s **high debt ratio (-0.40)** and **low credit score (-0.35)** outweighed positive factors such as **stable employment (+0.20)** and **high income (+0.25)**.

**4. Compliance Notes**

* Aligned with MAS FEAT (Fairness, Ethics, Accountability, Transparency).
* Stored in Explainability Register.
* Reviewed by Compliance (✔).

**5. Next Review**

Quarterly model explainability check – due Dec 2025.