SHAP Interaction Values: Capturing Feature Interactions

φ_{ij} measures joint effect between features i and j

Interaction Matrix: Credit Scoring Example

	Age	Income	Debt	Credit
Age	0.15	0.08	0.02	0.04
Income	0.08	0.22	0.05	0.03
Debt	0.02	0.05	-0.18	0.02
Credit	0.04	0.03	0.02	0.12

Main Effect Strong (>0.06) Moderate (0.04-0.06) Weak (<0.04)

$$\varphi_{total} = \varphi_i + \Sigma_j \varphi_{ij}$$

Total effect = Main effect + Interactions

Age × Income Interaction

Strong interaction (0.08) indicates that age and income jointly affect credit approval more than their individual effects suggest.

Scenario:

Young + High Income → Higher approval than expected
Old + Low Income → Lower approval than expected

Matrix Properties

- Symmetric: $\varphi_{ij} = \varphi_{ji}$
- Diagonal: main effects (φ_{ii})
- Off-diagonal: interactions
- TreeSHAP: exact computation

Interpretation Guide

- Positive: synergistic effect
- Negative: antagonistic effect

- Large magnitude: strong coupling
- Near zero: independent features