Formal Fairness Conditions for Financial ML Models

Definition 1 (Demographic Parity)**.** A classifier satisfies demographic parity if

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1).$$

Definition 2 (Equal Opportunity)**.** A classifier satisfies equal opportunity if

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1).$$

Definition 3 (Average Odds)**.** A classifier satisfies average odds if both

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1),$$

and

$$P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1).$$

Definition 4 (Counterfactual Fairness)**.** A predictor is counterfactually fair if for any individual,

$$\hat{Y}_{A \leftarrow a}(U) = \hat{Y}_{A \leftarrow a'}(U),$$

where U represents latent background variables, and $A \leftarrow a$ denotes intervention setting A to a.