When the Cost of Misclassification is Higher for the Customer than for the Business

Gaurav Sood

Abstract

Algorithmic decision systems create asymmetric costs where institutions face small, diffuse losses from false negatives while individuals face large, concentrated personal costs—a problem amplified when similar algorithms create correlated errors across entire markets. This leads to systematic under-investment in accuracy since decision-makers optimize for visible costs (bad hires, defaults) while externalizing invisible ones (missed talent, foregone profits). Mechanisms allowing affected parties to pay for enhanced evaluation can address this market failure, creating mutual benefits where individuals gain better decisions and institutions discover value they would have missed.

Traditional decision-making systems optimize for the decision maker's costs while largely ignoring the asymmetric losses that those subject to decisions face from misclassification. For instance, when an automated lending system falsely rejects a creditworthy applicant, the bank forgoes potential profits, a cost that's often diffuse and unmeasured. [1] However, the rejected applicant faces much larger, concentrated costs: missing out on a home purchase, business expansion, or emergency assistance.

This asymmetry becomes particularly damaging when multiple firms use similar systems, making correlated errors. When all banks rely on comparable credit scoring algorithms or employers filter using similar keywords and credentials, someone deemed "unsuitable" by one system faces systematic rejection everywhere. Individual classification mistakes become market-wide exclusion, amplifying personal costs while leaving institutional losses invisible.

[1] By the same token, employers who focus exclusively on avoiding false positives (bad hires) often overlook the substantial hidden costs of false negatives. These costs include prolonged recruiting cycles as they search for "perfect" candidates, project delays from unfilled positions, lost productivity from understaffing, and the opportunity cost of missed innovation from overlooked talent.

More generally, current systems systematically under-invest in accuracy because they optimize for visible costs (bad hires, defaults) while externalizing invisible ones (missed talent, foregone profits). This misalignment creates opportunities for mutually beneficial arrangements where affected parties pay for enhanced accuracy.

Consider voluntary comprehensive testing for job candidates. This directly addresses the market failure: candidates gain a mechanism to prove their worth despite algorithmic filtering, while employers discover talent they would have missed. Making such assessments portable across employers breaks the correlation of errors while improving overall accuracy.

The broader principle is that when algorithmic decisions create asymmetric costs and errors are correlated across firms, mechanisms allowing affected parties to pay for enhanced evaluation become not just beneficial but essential for market efficiency.

Formalization For the Lending Case

Let,

- $P_{\text{auto}}(x) = \Pr(\text{approve} \mid x)$: approval probability via automated system.
- $P_{\text{manual}}(x) = \Pr(\text{approve} \mid x, \text{review})$: approval probability after manual review.
- $\Pr(y = 1 \mid x)$: true repayment probability given features x.
- π_G: net gain when a good loan is approved.
- π_B: (negative) net gain when a bad loan (default) is approved.
- c_s: per-loan servicing cost.
- c_b: additional cost per manual review.
- L: applicant's loss if wrongly denied ($y = 1$, decision=deny).

1. Bank's Expected Profit

For mode $m \in \{\text{auto}, \text{manual}\}$, the bank's profit is

$$\Pi_m(x) = P_m(x)\big[\Pr(y=1 \mid x)\,\pi_G + (1 - \Pr(y=1 \mid x))\,\pi_B\big] - c_s - \mathbb{I}[m = \text{manual}]\,c_b.$$

2. Applicant Utility and Willingness to Pay

The applicant's expected utility under mode m (only false negatives matter) is

$$U_m(x) = -L\,\Pr(y=1 \mid x)\big[1 - P_m(x)\big].$$

Thus, the maximum fee the applicant will pay is

$$p_{\max}(x) = U_{\text{manual}}(x) - U_{\text{auto}}(x) = L\,\Pr(y=1 \mid x)\big[P_{\text{manual}}(x) - P_{\text{auto}}(x)\big].$$

The applicant often wouldn't know this, but would predict this. It could lead to some adverse selection, and we would need to monitor the calibration of the prediction function.

3. Pareto-Improving Manual Review

A manual review priced at p is Pareto-improving if both:

$$\Pi_{\text{manual}}(x) + p \geq \Pi_{\text{auto}}(x), \quad p \leq p_{\max}(x).$$

Combining gives the succinct criterion:

$$c_b \leq p \leq p_{\max}(x) \quad \Longleftrightarrow \quad p_{\max}(x) \geq c_b.$$

Interpretation

- The **bank** recovers its review cost c_b and may earn surplus.
- The **applicant** avoids a larger loss L by paying fee p.
- Both sides strictly improve their outcomes (Pareto improvement).

Extensions could include paying for a reduction in the uncertainty of the prediction, assuming that we output calibrated prediction intervals.