Preliminary design considerations for an AI autograder for high stakes decisions

Soumya Banerjee [1]

1 University of Cambridge, Cambridge, United Kingdom

Corresponding author sb2333@cam.ac.uk

Abstract

Introduction

Artificial intelligence (AI) is now ubiquitous. AI is being increasingly used in high-stakes scenarios such as healthcare [1]. Here, we describe the application of an AI autograder. This can potentially be used in deciding visa applications and eligibility for universities.

The key idea we want to explore in this work is the idea that there will always be tradeoffs in decisions and there are implicit costs to each automated decision. In certain cases, it may be feasible to triage the decisions: refer an automated decision to a human rater.

1 Methods

1.1 Decide What Error You Cannot Tolerate

In a high-stakes examination context, a **false pass** (i.e., the automarker predicts "fail" below the threshold, but the human rater would have passed the essay) or a **false fail** (i.e., the automarker predicts "pass" above the threshold, but the human rater would have failed the essay) may incur very different costs.

Typically, you would want near-perfect *precision* on **automated fails**: if the automarker confidently says "fail" it should be correct. Similarly, if you are automating "passes" you may prefer near-perfect *recall* on true fails. You must decide which type of error (false pass vs. false fail) is more critical and set your system's thresholds and tolerances accordingly.

Formally, you are optimising an objective function of the form:

$$\min_T \left[C_{\text{err}} \times \text{Error}_{\text{auto-only}}(T) + C_{\text{human}} \times \text{Workload}(T)\right]$$

where:

- $\text{Error}_{\text{auto-only}}(T)$ is the error rate (i.e., misclassifications) on all essays with confidence $\geq T$.
- $\text{Workload}(T)$ is the fraction of essays with confidence $< T$ (i.e., those routed to human raters).
- C_{err} and C_{human} are user-defined costs. For example, if a human review costs \$100 and a false pass costs \$1000, you would set the ratio $C_{\text{err}} : C_{\text{human}} = 10 : 1$.

2 Results

3 Sensitivity and specificity

We start by calculating the metrics sensitivity and specificity for the autograder.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The sensitivity (True Positive Rate) $= TP/(TP + FN) = 0.73$

The specificity (True Negative Rate) $= TN/(TN + FP) = 0.96$

3.1 Analysis of optimization model

Most of the time, the AI automarker makes high-confidence predictions (Figure 1). This suggests the idea that some of the low-confidence predictions could be triaged to a human rater. However there is a cost for this.

Optimising the following metric

$$\min_T \left[C_{\text{err}} \times \text{Error}_{\text{auto-only}}(T) + C_{\text{human}} \times \text{Workload}(T)\right]$$

yields an optimal value of T. This is shown in Figure 2. This yields an optimal $T = 0.99$. Ideally this process would need to be performed on a validation set.

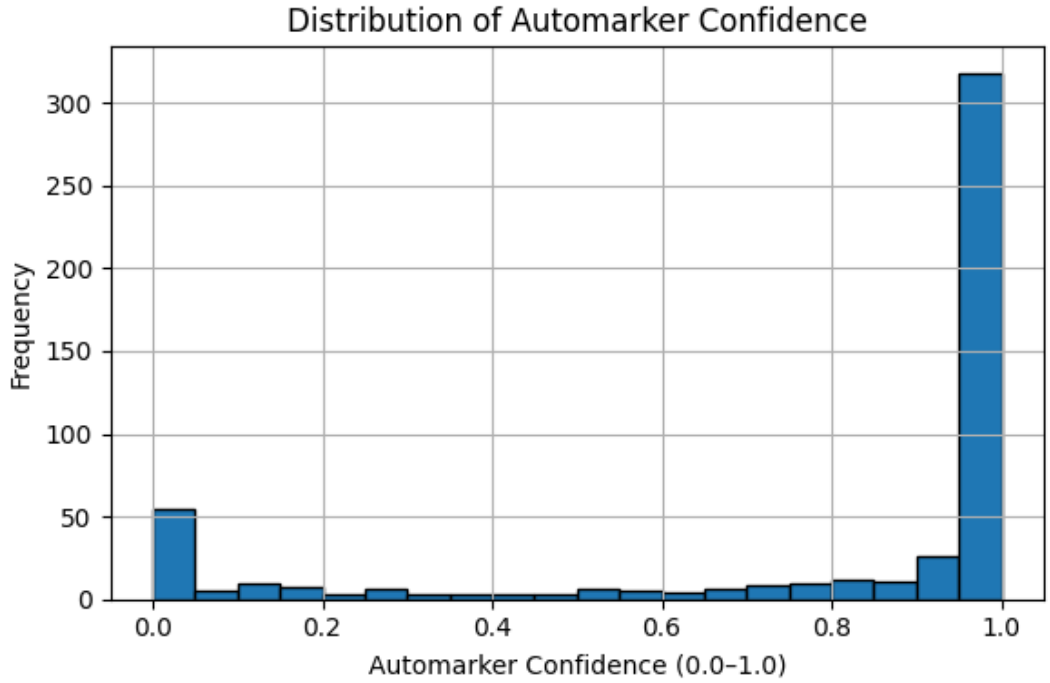

Figure 1. Histogram of confidence for automarker.

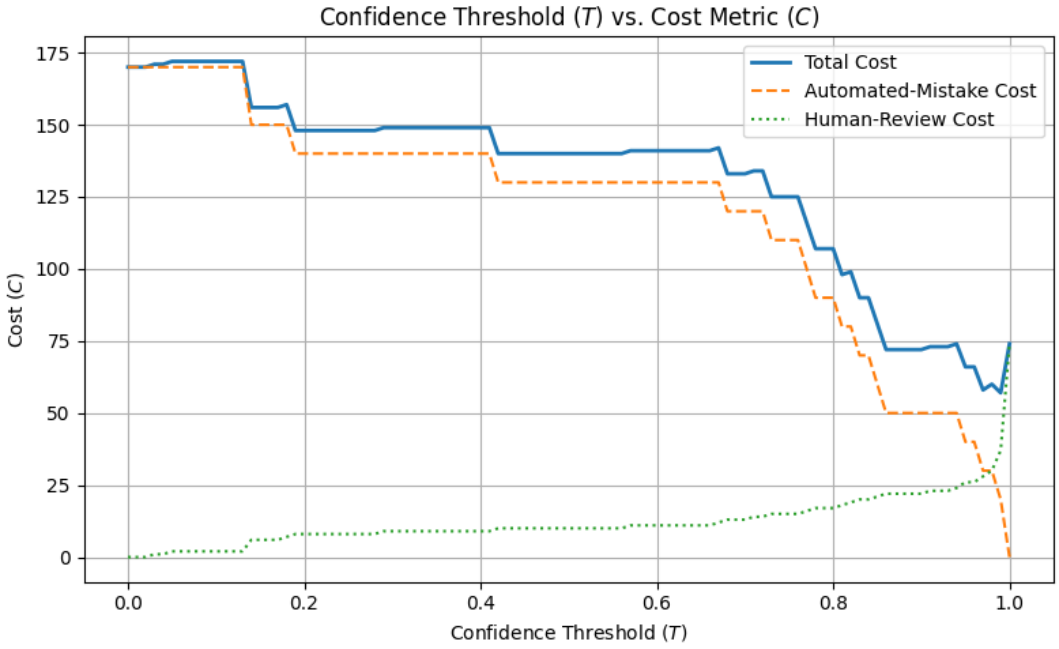

Figure 2. Plotting the values of the threshold T and the total cost $C = C_{\text{err}} \times \text{Error}_{\text{auto-only}}(T) + C_{\text{human}} \times \text{Workload}(T)$.

3.2 Alternative probabilistic approach

An alternative approach is a probabilistic approach. In this we assume that for each exam, with probability proportional to

$$C_{err} \cdot (1 - \text{autograder}_{\text{confidence}}) \tag{1}$$

we make a decision to refer to a human rater. With probability proportional to

$$C_{\text{human}} \cdot \text{autograder}_{\text{confidence}} \tag{2}$$

we use the autograder grade.

This would be a probabilistic version of the algorithm.

4 Recommendations

Based on these results, the autograder may be deployed in a limited trial.

1. High-stakes decisions made by the autograder (candidate failed or visa denied), should also be referred to a human rater.
2. The users should be informed that their examinations will be graded using an AI autograder. Transparent disclosure to users is essential for responsible use of AI.
3. Before final deployment of the AI tool, the feedback of users (on the AI autograder) should also be taken into account.
4. Some efforts should be made to make the tool explainable. If a candidate is being failed, the candidate deserves the right to know why they were failed.

Limitations and Discussion

Our work has a number of limitations and can be extended in many ways.

1. Here we have applied the approach to only the decision boundaries of pass and fail. Our approach can be extended to include multiple decision boundaries. For example, it can include costs for errors on different grade levels (A1 vs. A2, B1 vs. B2, etc.)

   The metric to optimise will be of the form

   $$\min_{T} \Sigma_{grade} \left[C_{\text{err}}^{\text{grade}} \times \text{Error}_{\text{auto-only}}(T) + C_{\text{human}}^{\text{grade}} \times \text{Workload}(T) \right]$$

   where the summation is over all grade levels (A1, B1, etc.) and there are different costs associated with each grade. For example, the cost associated with making a mistake on pass vs. fail would be greater than the cost associated with making an error on C1 vs. C2.

2. The probabilistic approach outlined here can be extended to yield an optimization framework.

3. We emphasise that there may also be a need to use explainable models [1] for high-stakes decisions.

   Simpler and interpretable models such as logistic regression and decision trees maybe used for this purpose [2] [3].

References

1. Banerjee S, Lio P, Jones PB, Cardinal RN (2021) A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. npj Schizophrenia 7: 1-13.

2. Gareth J, Daniela W, Trevor H, Robert T (2017) Introduction to Statistical Learning with Applications in R. Springer. URL `http://www-bcf.usc.edu/~gareth/ISL/`.

3. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1: 206-215.