A Overview of Semi-Automated Measurement Cycles

We present a high-level description of our proposed measurement pipeline for monitoring weighted rates. The details and formulas are provided in the main text.

Goal: Given a catalog domain and a measure of customer attention for each product (product importance), we want to estimate, at predetermined time intervals, the fraction of customer attention that goes to products with a certain characteristic. This is our *weighted rate* of interest.

Prerequisite: We need a pre-trained classifier for predicting product defects using the products' features. For classifiers that require thresholding a score to obtain the predictions, we provide options for dealing with the threshold.

Strategy: We propose *measurement cycles*, spanning over a predetermined number of months, that start with a *baseline* (or *time 0*) estimation based on audited data only, with *follow-up* (or *time t*) estimations in the cycle produced automatically.

- **Estimation at baseline (time 0)**: 1) Collect a random sample of catalog products and audit them for the defect of interest. 2) Using the audit results only, estimate the weighted rate, along with its variance and confidence interval. Use this as the official estimation at baseline. 3) Evaluate the classifier on the audited random sample, and estimate the true and false positive rates. If the classifier depends on a threshold, estimate these true and false positive rates for a grid of threshold values.
- **Estimation at follow-up (times t>0 in the cycle)**: 1) Evaluate the classifier on a large random sample from the catalog domain at time t to predicted their defect status (if feasible, evaluate on the entire catalog domain). If the classifier depends on a threshold, do this for the grid of threshold values used at baseline. 2) Compute a 'raw' estimate of the weighted rate using the classifier predictions. If the classifier depends on a threshold, obtain a 'raw' estimate per threshold in the grid used at baseline. 3) Compute the 'adjusted' estimator of the weighted rate, along with its variance and confidence interval. If the classifier depends on a threshold, follow one of the recommended threshold-handling approaches to obtain a final estimator of the weighted rate, along with its variance and confidence interval. The output of this step is the official estimation at follow-up time t.

B Analytic Variance Derivation

To derive the variance of $\hat{R}_t = (\hat{R}_t^{raw} - \hat{p}_{1|0})/(\hat{p}_{1|1} - \hat{p}_{1|0})$, we assume that the variance coming from $\hat{R}_t^{raw}$ is negligible, given that the evaluation sample at time t will be orders of magnitude larger than the audited sample at time 0. We then focus on obtaining the variance of $\hat{R}_t$ seen as a function of $\hat{p}_{1|1}, \hat{p}_{1|0}$. Assuming a simple random sample of size n_0 at baseline, we can estimate $p_{ab} = P(Y = a, \hat{Y} = b)$ as $\hat{p}_{ab} = n_{0ab}/n_0$, where n_{0ab} is number of elements with