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

true status $Y = a$ and classified as $\hat{Y} = b$. Denoting $\hat{\mathbf{p}}_0 = (\hat{p}_{11}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{00})^T$, the multivariate Central Limit Theorem tells us that as $n_0 \to \infty$, $\sqrt{n}_0(\hat{\mathbf{p}}_0 - \mathbf{p}_0) \to Normal_4(\mathbf{0}_4, \text{diag}[\mathbf{p}_0] - \mathbf{p}_0\mathbf{p}_0^T)$, [2]. Given this, we can use the Multivariate Delta Method, as presented for instance in Chapter 14 of [2], to first obtain the asymptotic distribution of $(\hat{p}_{1|1},\ \hat{p}_{1|0})^T = (\hat{p}_{11}/(\hat{p}_{11} + \hat{p}_{10}), \hat{p}_{01}/(\hat{p}_{01} + \hat{p}_{00}))^T$, and then obtain the asymptotic distribution of $\hat{R}_t = (\hat{R}_t^{raw} - \hat{p}_{1|0})/(\hat{p}_{1|1} - \hat{p}_{1|0})$, which is found to be normal, centered at R_t, and with asymptotic variance given by

$$\text{var}_\infty(\hat{R}_t) = \frac{1}{n_0(p_{1|1} - p_{1|0})^2}\left[\frac{R_t^2\ p_{1|1}(1-p_{1|1})}{r_0} + \frac{(1-R_t)^2\ p_{1|0}(1-p_{1|0})}{1-r_0}\right],$$

under the extrapolation assumption. Our analytic estimator of the variance of $\hat{R}_t$ is obtained from replacing $\hat{p}_{1|1}, \hat{p}_{1|0}, \hat{R}_t$ and $\hat{r}_0 = (n_{010} + n_{011})/n_0$ in the formula of $\text{var}_\infty(\hat{R}_t)$.

C Details of Simulation Design

We generate synthetic catalogs under a variety of configurations. The synthetic catalogs are given by $\{(W_{it}, Y_{it}, \hat{Y}_{it}, X_{it})\}_{i=1}^{N_t}$, $N_t = 10^6$, one catalog for a baseline time $t = 0$ and one for a follow-up time $t > 0$. The catalog generation described below is done such that we can control characteristics that are representative of a variety of our use cases.

C.1 Baseline Catalog Generation

The data for each product i is generated independently of each other.
Characteristic of interest Y_{i0}:

- Given a value of the baseline unweighted rate $r_0 = P_0(Y = 1)$, we generate the indicator of the characteristic as $Y_{i0} \sim Bernoulli(r_0)$, $i = 1, \ldots, N_0$. We consider $r_0 = 0.1, 0.2, 0.3$.

Product weights W_{i0}:

- We generate the product weights among products without the characteristic as positive counts $W_{i0} \mid Y_{i0} = 0 \sim 1 + NegativeBinomial(s_0 = 0.0496, q_0 = 0.0008)$, such that the mean is $\mu_{W|0} = 1 + s_0q_0/(1-q_0)$ and its variance is $\sigma^2_{W|0} = s_0q_0/(1-q_0)^2$. The values s_0, q_0 were obtained from fitting a negative binomial distribution via the method of moments to real data on the numbers of visits to product pages on Amazon.com during a time period for a category of products. The particular configuration of s_0, q_0 makes it so that the distribution of $W_{i0} \mid Y_{i0} = 0$ is highly skewed, with most of the products in the synthetic catalog having a small integer as W_{i0}, but with a small fraction of products having extremely large W_{i0}.
- Similarly, we generate the weights among products with the characteristic as $W_{i0} \mid Y_{i0} = 1 \sim 1 + NegBinom(s_1, q_1)$, where the mean is $\mu_{W|1} = 1 + s_1q_1/(1-q_1)$. We take $s_1 = s_0$, and obtain q_1 from setting $\mu_{W|1}$ so that the weighted rate R_0 is a specific fraction d of the unweighted rate r_0. Specifically, here the expected weighted rate at baseline is $R_0 = E_0(WY)/E_0(W) = r_0\mu_{W|1}/[r_0\mu_{W|1} + (1-r_0)\mu_{W|0}]$. Given

that for many of our use cases we expect the characteristic of interest $Y = 1$ to be more prevalent among products with lower weights, we expect $R_0 < r_0$. In particular, we fix $R_0 = d\ r_0$ for $d = 1/4, 1/2, 3/4$. For a given value of R_0, we can find $\mu_{W|1}$ from the formula above, and thereby the appropriate q_1 to generate the weights $W_{i0} \mid Y_{i0} = 1$.

Product attributes X_{i0} used to predict Y_{i0}:

- We generate the attributes for products with and without the characteristic as $X_{i0} \mid Y_{i0} = a \sim Normal(\mu_{X|a}, \sigma^2_{X|a})$, for $a = 1, 0$. We choose the parameters $\mu_{X|0}, \sigma^2_{X|0}, \mu_{X|1}, \sigma^2_{X|1}$ so that we obtain different levels of classification difficulty, defined as follows. The optimal (Bayes or oracle) classifier according to the 0-1 loss is obtained as $h^*(x) = I[g^*(x) > 0.5]$ where $g^*(x) = P_0(Y = 1 \mid x)$, and $P_0(Y = 1 \mid x) = r_0 f_0(x \mid Y = 1)/[r_0 f_0(x \mid Y = 1) + (1 - r_0) f_0(x \mid Y = 0)]$, where $f_0(x \mid Y = 1)$ and $f_0(x \mid Y = 0)$ are the densities of the normal distributions above [19]. We can then characterize the classification difficulty of the problem by the true and false positive rates of the optimal classifier $h^*(x)$: $TPR = P_0[h^*(x) = 1 \mid Y = 1] = \int h^*(x) f_0(x \mid Y = 1) dx$ and $FPR = P_0[h^*(x) = 1 \mid Y = 0] = \int h^*(x) f_0(x \mid Y = 0) dx$. We consider three scenarios of classification difficulty by fixing $TPR = 0.5, 0.7, 0.9$ and fixing $FPR = 0.05$. The parameters $\mu_{X|0}, \sigma^2_{X|0}, \mu_{X|1}, \sigma^2_{X|1}$ were selected using a grid search so that they lead to the values of TPR and FPR above.

Predictions $\hat{Y}_{i0}$:

- Given that in our use cases we draw our predictions from large pre-trained machine learning models that are not retrained for the purposes of metric measurement, our estimation treats the scoring function $g(x)$ as fixed, and therefore in this simulation study we are interested in understanding the impact of the quality of the scoring function $g(x)$. For this reason, in this simulation study we simply take the model score $g(x)$ to be the oracle $g^*(x)$ described above, as we already designed $g^*(x)$ to represent different quality scenarios. We can then generate the scores $g(X_{i0})$ for each product in the synthetic catalog. Given a threshold c, these scores can be used to obtain predictions $\hat{Y}_{i0} = I[g(X_{i0}) > c]$.

C.2 Follow-Up Catalog Generation

To generate the catalog at time $t > 0$, we fix different values of the percent change of the weighted rate from time 0 to $t > 0$. The change is $\Delta = 100(R_t - R_0)/R_0$, where we take $\Delta = -50\%, -25\%, +25\%, +50\%$. The different combinations of Δ and R_0 considered here lead to a wide range of scenarios for the weighted rate R_t going from 1.25% to 33.75%. Given a value of R_t, and assuming that the conditional distributions of weights $W_{it} \mid Y_{it} = 0$ and $W_{it} \mid Y_{it} = 1$ remain the same for time $t > 0$ as for baseline time, we have that $R_t = r_t \mu_{W|1}/[r_t \mu_{W|1} + (1 - r_t)\mu_{W|0}]$, from which we can obtain the value of the unweighted rate at time t, r_t. Given a value of r_t, we generate the number of products in the catalog at time t that will have the characteristic as a draw from a $Binomial(N_t, r_t)$. If that number is different from the number for the baseline catalog, we change the characteristic for the appropriate number of products and re-generate the weights W_{it} and the features X_{it}. Products for which their status Y_{it} does not change from Y_{i0} keep their values $W_{it} = W_{i0}$ and $X_{it} = X_{i0}$.

C.3 Sampling and Estimation

Given a pair of synthetic catalogs for baseline and for time $t > 0$, we repeat 1000 times the following estimation process:

- Sample with replacement n_0 products from the baseline catalog, and record their ground truth values Y_{i0}, along with $W_{i0}, g(X_{i0})$. Here, n_0 represents the size of the sample that is audited in practice. We consider the performance of our proposed estimation approaches for $n_0 = 500, 1000, 2000$.
- For a predetermined grid of threshold values, $c_1, \ldots, c_u$, here taken as $0.01, 0.02, \ldots, 0.99$, compute the estimates of the true and false positive rates $p_{1|1}$ and $p_{1|0}$.
- Compute R_t^{raw} using the entire catalog at follow-up time $t > 0$ for the different threshold values, $c_1, \ldots, c_u$. This is a reasonable set-up, given that R_t^{raw} only depends on the classifier predictions $\hat{Y}_{it} = I[g(X_{it}) > c]$, which do not involve auditing resources. In practice, if computing $\hat{Y}_{it} = I[g(X_{it}) > c]$ on every product in the target catalog is computationally prohibitive, we can still estimate R_t^{raw} using a very large sample, so that its variability is negligible compared to the variability coming from the audited sample.
- Compute the corresponding $\hat{R}_t$ for each threshold $c_1, \ldots, c_u$.
- Obtain the final estimates of the weighted rate using each of the approaches outlined before.