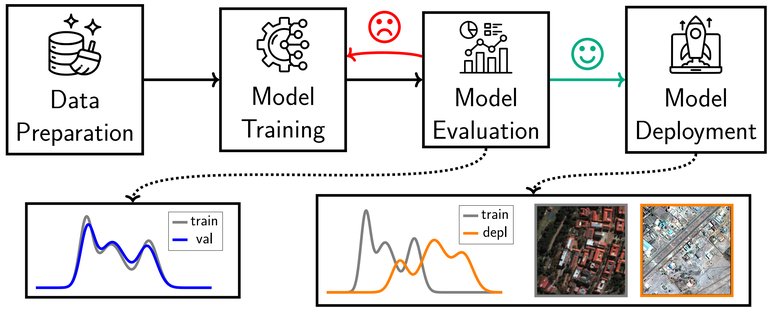
Suitability Filter: A Statistical Framework for Classifier Evaluation in Real-World Deployment Settings

Angéline Pouget[1] Mohammad Yaghini[2] Stephan Rabanser[2] Nicolas Papernot[2]

[1]ETH Zürich [2]University of Toronto & Vector Institute

Distribution Shift in Deployment Impacts Performance



Suitability Filters Detect Performance Deterioration

A classifier $M : \mathcal{X} \to \mathcal{Y}$ is **suitable** for use on *unlabelled* $D_u \sim \mathcal{D}_{\text{target}}$ iff the estimated accuracy of M on D_u deviates at most by m from the accuracy on $D_{\text{test}} \sim \mathcal{D}_{\text{source}}$:

$$\underbrace{\frac{1}{|D_u|}\sum_{x \in D_u} \mathbb{I}\{M(x) = \mathcal{O}(x)\}}_{\text{estimated accuracy on user data}} \geq \underbrace{\frac{1}{|D_{\text{test}}|}\sum_{(x,y) \in D_{\text{test}}} \mathbb{I}\{M(x) = y\}}_{\text{accuracy on test data}} - m.$$

Question

How can we proxy accuracy on the user data without access to labels?

A **suitability filter** $f_s : \mathcal{X} \to \{\text{SUITABLE, INCONCLUSIVE}\}$ outputs SUITABLE iff M is suitable for use on D_u with high probability and INCONCLUSIVE otherwise.

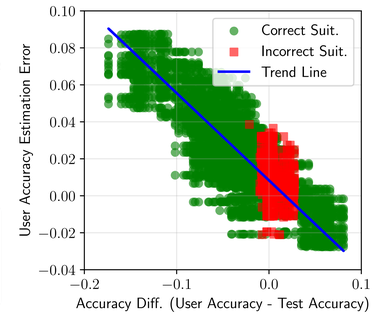
Margin Tuning Guarantees Suitability in Practice

δ-Calibration: $\mathbb{E}[p_c(x)] - \mathbb{P}[M(x) = y] = \delta$.

If C is δ-calibrated with δ_{source} and δ_{target}, define $m' = m + \delta_{\text{source}} - \delta_{\text{target}}$ and conduct a non-inferiority test with $H_0 : \mu_{\text{target}} < \mu_{\text{source}} - m'$. Then the FPR of this test is bounded by α.

Insight

Margin tuning ensures that C continues to provide reliable estimates, even in the presence of distribution shifts.

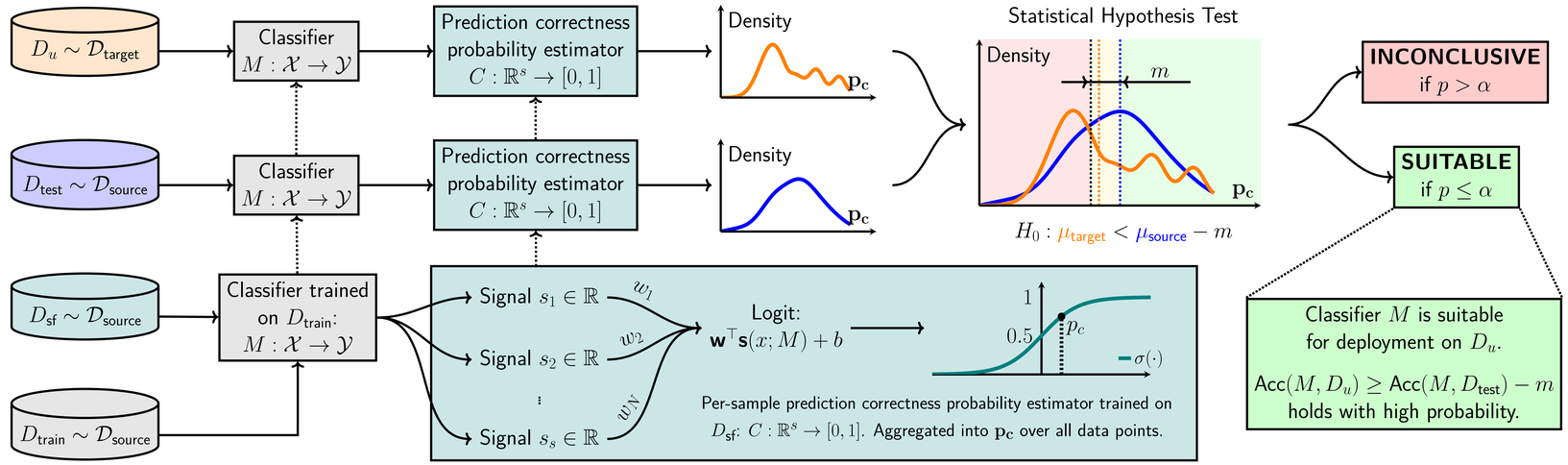


Main Contribution

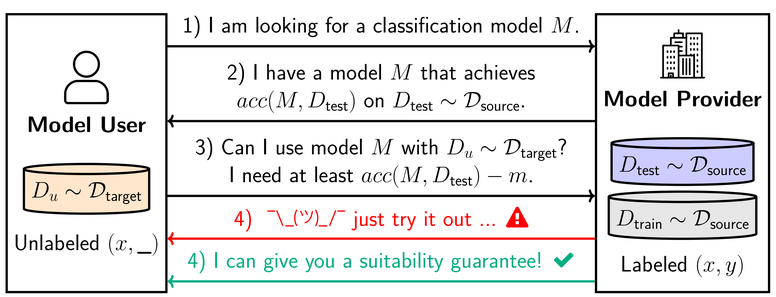
Detecting **performance drops on new, unlabeled user data** is a key challenge for deployed ML models. Our **suitability filter** compares model outputs from new and old data via **statistical testing** to determine if the model's accuracy has **degraded beyond a user-defined performance margin**. Our testing procedure even **guarantees bounded FPR** under some distribution shifts.

Learn more:

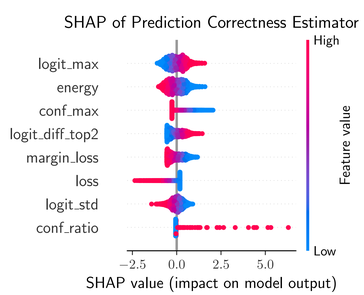
Paper & Code



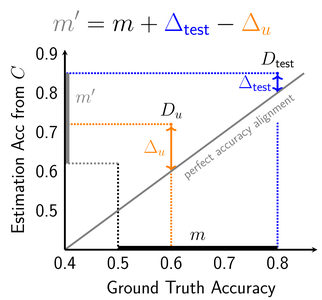
How Can a Model User Pick a Good Model?



Suitability Signals

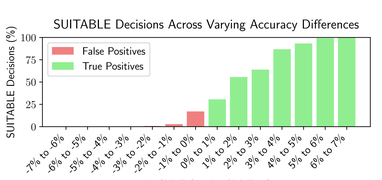


Margin Adjustment



Suitability Filters Work Across Domains

Dataset	Acc	FPR	ROC
FMoW ID	$81.8_{\pm 3.1}\%$	$0.027_{\pm 0.033}$	$0.969_{\pm 0.023}$
FMoW OOD	$91.9_{\pm 2.5}\%$	$0.018_{\pm 0.017}$	$0.965_{\pm 0.016}$
RxRx1 ID	$100.0_{\pm 0.0}\%$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$
RxRx1 OOD	$97.5_{\pm 7.2}\%$	$0.031_{\pm 0.088}$	$0.997_{\pm 0.006}$
CivilComm ID	$93.3_{\pm 5.3}\%$	$0.002_{\pm 0.007}$	$0.997_{\pm 0.008}$



Observation

On FMoW, we detect performance deterioration of $> 3\%$ with 100% accuracy.