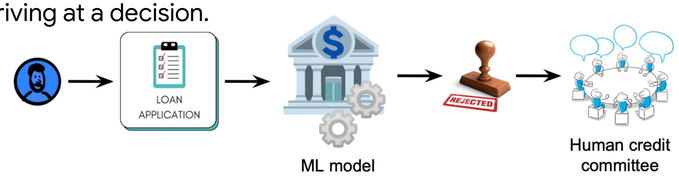
Selective Regression Under Fairness Criteria

Abhin Shah, Yuheng Bu, Joshua Ka-Wing Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, Gregory W. Wornell

MIT-IBM Watson AI Lab

Selective Prediction

- A trustworthy machine learning (ML) system should reliably communicate the uncertainty in its predictions.
- Consider a loan approval ML system designed to predict loan terms (e.g., loan approval, interest rate).
- If the model's uncertainty is high for an applicant, the prediction can be rejected to avoid potentially costly errors.
- The decision-maker can intervene and take remedial actions before arriving at a decision.



Uncertainty Measure

- If we have an uncertainty/confidence measure for each prediction, we can decide to abstain from decision making if our confidence is below a certain threshold.
- With a good confidence measure, increasing the threshold results in a better performance.
- The tradeoff is that we have predictions for a fewer samples (i.e., low coverage).

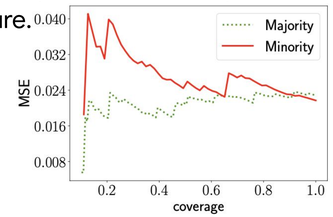
Selective Classification

- Classifiers can have good average performance but may perform poorly on certain subgroups [Jones et al., 2020].
- To mitigate such disparities, recent works [Lee et al., 2021 etc] proposed methods for performing fair selective classification.
- For a classification task, an uncertainty measure can be learned using the softmax output (of an existing classifier).
- However, there is no direct method to extract an uncertainty measure from an existing regressor designed only to predict the conditional mean!

We demonstrate and investigate the performance disparities across subgroups for selective regression as well as develop novel methods to mitigate such disparities.

Disparities between subgroups in selective regression:

- We consider predicting annual medical expenses charged to patients from age, BMI, number of children, etc as in the Insurance dataset.
- Following [Zaoui et al., 2020], we use conditional expectation as our predictor & conditional variance as our uncertainty measure.
- While decreasing the coverage improves the performance for the majority subgroup (i.e., females), the performance for the minority subgroup (i.e., males) degrades.



Monotonic Selective Risk (MSR): proposed novel fairness criteria

- MSR requires the risk of each subgroup to monotonically decrease with a decrease in coverage.
- We construct our predictor f and our uncertainty measure g using a feature representation Φ.
- The subgroup MSE for $d \in \mathcal{D}$, as a function of f and g, for a fixed coverage (parameterized by τ) is
  - $\mathrm{MSE}(f, g, \tau, d) = E[(Y - f(\Phi(X)))^2 \mid g(\Phi(X)) \le \tau, D = d]$.
- We say that f and g satisfy MSR if for any $\tau < \tau'$
  - $\mathrm{MSE}(f, g, \tau, d) \le \mathrm{MSE}(f, g, \tau', d) \ \forall\, d \in \mathcal{D}$.

Theorem: sufficiency ⇒ MSR

- A feature representation $\Phi(X)$ satisfies sufficiency if
  - $Y \perp D \mid \Phi(X)$.
- If $\Phi(X)$ is sufficient, then conditional mean as the predictor and conditional variance as the uncertainty measure ensures MSR.

Theorem: calibration for mean & variance ⇒ MSR

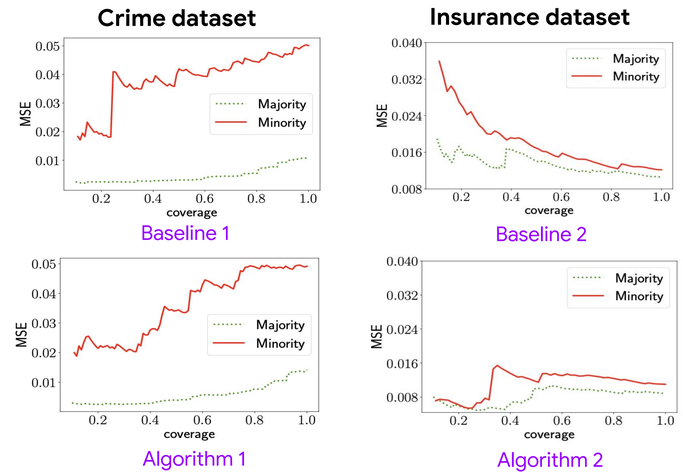
- A feature representation $\Phi(X)$ is calibrated for mean and variance if
  - $E[Y \mid \Phi(X), d] = E[Y \mid \Phi(X)] \ \forall\, d \in \mathcal{D}$
  - $\mathrm{Var}[Y \mid \Phi(X), d] = \mathrm{Var}[Y \mid \Phi(X)] \ \forall\, d \in \mathcal{D}$
- If $\Phi(X)$ is calibrated for mean and variance, then conditional mean as the predictor and conditional variance as the uncertainty measure ensures MSR.

Algorithm 1: Imposing sufficiency

- We use Heteroskedastic neural network (NN) which requires training a single NN by assuming $Y|X$ is Gaussian.
- We relax $Y \perp D \mid \Phi(X)$ by using conditional mutual information $I(Y; D \mid \Phi(X))$ which can be upper bounded (Lee et al,. 2021) by:
  - $I(Y; D \mid \Phi(X)) \le E_{\Phi(X),Y,D}[\log P(Y \mid \Phi(X), D)] - E_D[E_{\Phi(X),Y}[\log P(Y \mid \Phi(X), D)]]$.
- We train subgroup-specific Gaussian models to learn $P(Y \mid \Phi(X), D)$.

Algorithm 2: Imposing calibration for mean & variance

- We let $\Phi = (\Phi_1, \Phi_2)$ *and* use residual-based NN by letting (a) the predictor depend only on Φ_1 and (b) the uncertainty measure depend only on Φ_2.
- We relax $E[Y \mid \Phi(X), D] = E[Y \mid \Phi(X)]$ by the contrastive loss:
  - $E_D[E_{\Phi(X),Y}[(Y - E[Y \mid \Phi_1(X), D])^2]] - E_{\Phi(X),Y,D}[(Y - E[Y \mid \Phi_1(X), D])^2]$.
- We train subgroup-specific mean prediction models to learn $E[Y|\Phi_1(X), D]$.
- We impose calibration for variance similarly.



Quantitative comparison

Dataset	Algorithm	Area under curve (AUC)	AUC (Majority)	AUC (Minority)
Insurance	Baseline 1	0.0371	0.0342	0.0442
	Algorithm 1	0.0195	0.0207	0.0167
	Baseline 2	0.0142	0.0129	0.0175
	Algorithm 2	**0.0099**	**0.0087**	**0.0120**
Crime	Baseline 1	**0.0075**	**0.0040**	0.0345
	Algorithm 1	0.0079	0.0045	**0.0296**
	Baseline 2	0.0101	0.0060	0.0442
	Algorithm 2	0.0117	0.0082	0.0375