Evaluating Predictive Uncertainty and Robustness to Distributional Shift Using Real World Data

Kumud Lakara *
Dept. of Computer Science and Engineering
Manipal Institute of Technology, MAHE
Manipal, India
lakara.kumud@gmail.com

Akshat Bhandari *
Dept. of Computer Science and Engineering
Manipal Institute of Technology, MAHE
Manipal, India
akshatbhandari15@gmail.com

Pratinav Seth *
Dept. of Computer Applications
Manipal Institute of Technology, MAHE
Manipal, India
seth.pratinav@gmail.com

Ujjwal Verma
Dept. of Electronics and Comm. Engineering
Manipal Institute of Technology, MAHE
Manipal, India
ujjwal.verma@manipal.edu

Abstract

Most machine learning models operate under the assumption that the training, testing and deployment data is independent and identically distributed (i.i.d.). This assumption does not generally hold true in a natural setting. Usually, the deployment data is subject to various types of distributional shifts. The magnitude of a model's performance is proportional to this shift in the distribution of the dataset. Thus it becomes necessary to evaluate a model's uncertainty and robustness to distributional shifts to get a realistic estimate of its expected performance on real-world data. Present methods to evaluate uncertainty and model's robustness are lacking and often fail to paint the full picture. Moreover, most analysis so far has primarily focused on classification tasks. In this paper, we propose more insightful metrics for general regression tasks using the Shifts Weather Prediction Dataset. We also present an evaluation of the baseline methods using these metrics.

1 Introduction

Recent times have seen the growing deployment of machine learning models and deep neural networks in many mission-critical tasks such as medical image diagnosis [6], banking systems [2] and autonomous vehicles [15]. A prevalent assumption in machine learning is that the data the model sees after the development phase is independent and identically distributed (i.i.d). This implies that if a model performs well during the development phase it will be able to interpolate that performance to deployment. This however is seldom the case. Most data available to models in the real world is often out of the distribution of the data on which the models were trained and tested. When dealing with critical applications, the model's inability to generalize to the out of distribution deployment data can have disastrous consequences. Therefore, such tasks require not only for models to make accurate predictions but also a precise quantification of predictive uncertainty [1]. Thus, it is imperative to properly assess a model's robustness to distribution shift and its estimation of predictive uncertainty.

While the problem is apparent, yet there are limited resources and datasets that allow for proper evaluation of uncertainty estimates and robustness to distributional shift emulating real-world data. Most of the available datasets such as Imagenet - C [3], A [5], R [4], O [5] and WILDS [7]

*Equal contribution.

Bayesian Deep Learning workshop, NeurIPS 2021.

focus primarily on image classification tasks. The recently introduced Shifts Dataset [13] provides a favourable data setting. It is composed of three parts each corresponding to a different data modality: tabular weather prediction data, machine translation data and self-driving car data. All the data modalities are affected by distributional shift and pose challenges with respect to evaluating uncertainty predictions and robustness against distributional shift.

In this work, we present novel evaluation metrics for predictions on the tabular data related to weather prediction in the Shifts Dataset [13]. We introduce metrics for effectively evaluating both uncertainty predictions *and* robustness to distributional shift for regression tasks. We believe these metrics will be a much needed addition to the scarce pool of metrics available presently for regression tasks. In addition to this, we also validate the proposed metrics by evaluating the performance of baselines for predicting the temperature at a particular latitude/longitude and time, given all available measurements and climate model predictions.[2]

2 Evaluation Metrics

We believe a model with lower degradation in performance on shifted data is more robust to distribution shift. The quality of predictive uncertainty is governed by the model's ability to discriminate between i.i.d. and out of distribution data. We present metrics for the joint evaluation of predictive uncertainty and robustness to distributional shift. We validate our proposed metrics using the baseline Gradient Boosted Decision Trees (GBDT) models as used in [13].

To build baseline gradient boosted decision trees for the temperature prediction regression task, the open-source CatBoost gradient boosting library [16] is used. Similar to Malinin et. al [13], we consider an ensemble-based approach to uncertainty estimation for GBDT models [17]. We use a pre-trained ensemble of ten models on the train data from the canonical partition of the Weather Prediction dataset with different random seeds. The models are optimized with the loss function `RMSEWithUncertainty` [17] which predicts mean and variance of the normal distribution by optimizing the negative log-likelihood. Hyperparameter tuning is performed on the in-domain development data. Each model is constructed with a depth of 8 and is then trained for 20,000 iterations with a learning rate of 0.3.

2.1 RMSE/RMV Ratio

The root mean squared error (RMSE) and the root mean variance (RMV) can be viewed as indicators of the model accuracy and uncertainty respectively.

$$RMSE = \sqrt{\frac{1}{|B|}\sum_{b\epsilon B}(y_b - \hat{y_b})^2} \qquad (1) \qquad\qquad RMV = \sqrt{\frac{1}{|B|}\sum_{b\epsilon B}\sigma_b^2} \qquad (2)$$

where σ_b^2 is the variance, $|B|$ is the bin size and $y_b - \hat{y_b}$ is the error between the model's prediction and the ground truth value for example b ϵ B. The ratio of the two should be as close to one as possible. We reason this inference by considering the following paradigm:

If the error (RMSE) for a prediction is high then the model's uncertainty (RMV) about that prediction should also be high.

When plotting the retention curve [12][9] for the RMSE/RMV ratio, an ideal scenario would be a horizontal line close to one. This would indicate that RMSE is roughly equal to RMV. Figure 1(a) shows the RMSE/RMV retention curves (R3-curves) for single and ensemble models. The ensemble model clearly displays a more linear relationship between RMSE and RMV compared to the single model. The RMSE/RMV ratio is also much closer to one for the ensemble model compared to the single model.

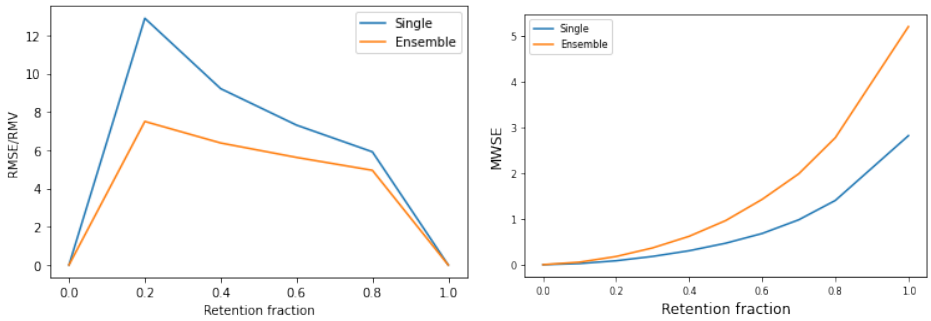
2.2 Logarithmic Expected Normalized Calibration Error (LENCE)

Levi et al. [10] propose a definition for calibration for regression by replacing mis-classification probability with mean squared error. This is called Expected Normalized Calibration Error (ENCE).

[2] `https://github.com/kumudlakara/Shifts_evaluation_metrics_regression`

Table 1: ENCE, C_v and LENCE for baseline GBDT ensemble and single models

Model	ENCE	C_v	LENCE
Ensemble	1.084	0.387	1.2968
Single	3.551	0.550	1.6804



(a)

Figure 1: (a) RMSE/RMV retention curve (b) MWSE retention curve

This is an improvement over the existing definition of calibrated regression uncertainty [8]. To calculate ENCE we arrange the data in increasing order of variance. Then we create bins of size S such that S divides the total number of examples into N bins. Each resulting bin represents an interval in the standard deviation axis. These intervals are non-overlapping and their boundary values are in increasing order [10]. This measure is now analogous to the Expected Calibration Error (ECE)[14] used in the case of classification problems. For finding the error in calibration of the model, we use the following formulation for ENCE:

$$ENCE = \frac{1}{N}\sum_{b=1}^{N}\frac{|RMV(b) - RMSE(b)|}{RMV(b)} \tag{3}$$

where N is the total number of bins. A shortcoming of ENCE is that if the model were to predict a homogeneous uncertainty for each example and that happened to match the model's empirical uncertainty for the entire population then ENCE would be zero. However, it would not be a correct evaluation of the models uncertainty or generalization ability. Levi et. al [10] propose using the Variation Coefficient of Standard Deviation, C_v as a secondary diagnostic tool with ENCE.

$$C_v = \frac{\sqrt{\frac{\sum_{t=1}^{T}(\sigma_t - \mu_\sigma)^2}{T-1}}}{\mu_\sigma} \tag{4}$$

where $\mu_\sigma = \frac{1}{T}\sum_{t=1}^{T}\sigma_t$ and T is the number of training examples. Evaluating two separate metrics and drawing relative inferences is complex and hence we propose the Logarithmic ENCE (LENCE). We incorporate C_v as a measure of dispersion of the uncertainty predictions by the model and ENCE as a measure of the calibration error into a single evaluation metric. We define LENCE as:

$$LENCE = log|ENCE + \frac{1}{C_v}| \tag{5}$$

Ideally, the uncertainty values will be well dispersed and C_v would be high. Then LENCE will approximately be $log|ENCE|$. In the absolute worst case when $ENCE = 0$ and $C_v \to 0$; $LENCE \to \infty$. ENCE, C_v and LENCE values for single and ensemble models are presented in Table 1. A lower LENCE value for the ensemble model indicates its superiority.

2.3 LL-Fisher Uncertainty (LL-FU)

We introduce the LL-Fisher metric as a joint measure of uncertainty and robustness to distributional shift. The LL-Fisher metric measures the localisation of a probability distribution function [18]. This

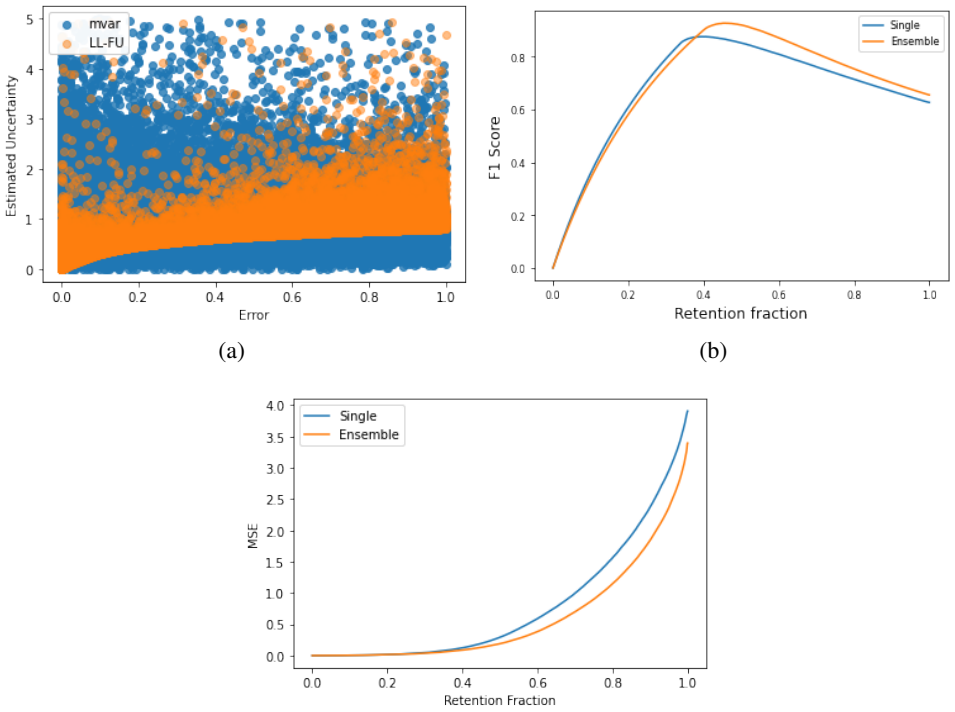


Figure 2: (a) LL-FU vs error scatter plot (b) LL-FU F1 retention curve (c) MSE - LLFU retention curve

Table 2: Ensemble model F1-AUC, F1@95 and R-AUC scores for uncertainty measures on the dev set of the Shifts Weather Prediction Dataset

	Uncertainty Metric				
	mvar	tvar	varm	epkl	LL-FU
F1-AUC	0.4951	0.5220	0.5012	0.5051	**0.7009**
F1@95	0.6581	0.6583	0.6579	0.6551	**0.6778**
R-AUC	1.4186	1.2708	1.2868	1.2266	**0.5838**

metric can be used to evaluate how well a prediction represents the distribution of data that the model used to make the prediction itself. Inspired by log-loss used for calculating Fisher Information [11] for a prediction x with mean μ and variance σ^2 following a continuous distribution.

We propose the LL-Fisher Uncertainty as follows:

$$LL\text{-}FU = max(0, \frac{1}{2}log(2\pi\sigma^2)) + \frac{(x-\mu)^2}{2\sigma^2} \tag{6}$$

We plot the F1-retention curve Figure 2(b), the MSE-retention curve Figure 2(c) and calculate the summary statistics: F1-AUC and R-AUC for the evaluation of our metric. We use F1@95 score to jointly evaluate uncertainty and robustness. A good uncertainty measure should achieve low R-AUC, high F1-AUC and high F1@95 scores [13]. We calculated these metrics for various uncertainty measures of prediction using a GBDT model as well as an ensemble of these models.

Results using ensemble and single models for various uncertainty measures are depicted in Table 2 and Table 3 showing superiority of our uncertainty measure. We compare both scores for ensemble and single model for LLFU. The scores for ensemble are better than the single model indicating the superiority of the ensembling approach. These results are presented in Table 4 .

Table 3: Single model F1-AUC, F1@95 and R-AUC scores for uncertainty measures on the dev set of the Shifts Weather Prediction Dataset.

	Uncertainty Metric				
	mvar	tvar	varm	epkl	LL-FU
F1-AUC	0.4425	0.4425	0.3851	0.4425	**0.6873**
F1@95	0.6276	0.6276	0.6115	0.6276	**0.6478**
R-AUC	1.8868	1.8868	1.9526	1.8868	**0.7740**

We observe that for out of distribution data, as the error of prediction increases, the lower bound of uncertainty (LL-FU) also increases. This validates the idea that with increasing shift in data distribution, not only does the likelihood of an error increase but so does the uncertainty corresponding to the prediction. We believe that LL-FU is unique in capturing this property when compared to other uncertainty metrics [13] as reflected in Figure 2(a).

Table 4: LL-FU for baseline GBDT ensemble and single models

Model	F1-AUC	F1@95	R-AUC
Ensemble	**0.7009**	**0.6778**	**0.5838**
Single	0.6873	0.6478	0.7740

2.4 Mean Weighted Squared Error (MWSE)

Intuitively, a model should be increasingly uncertain about it's prediction as the magnitude of distributional shift increases and hence will make more erroneous decisions. The error-retention curves [12] [9] do not directly evaluate the uncertainty estimates of the model and only take into account the correlation between error and uncertainty estimates. To more resiliently evaluate robustness and also take into account the magnitude of estimated uncertainty, we introduce MWSE. We propose MWSE as a secondary metric only to gain more insight into the models' behaviour by analysing their relative divergence.

$$MWSE = \frac{1}{N}\sum_{i=1}^{i=N}(\hat{y_i} - y)^2 * (UNC)_i \tag{7}$$

where $\hat{y_i}$ and y is the predicted and true value respectively and *UNC* is model's uncertainty estimate for each prediction.

Ideally datapoints with low MSE should have low uncertainty and those with high MSE should have high uncertainty which would mean their product would be very low or very high respectively. In Figure 1(c) the ensemble diverges more than the single model. As the retention fraction increases, the ensemble shows a greater divergence from the single model. From this observation we infer that the ensemble is comparatively more responsive to distributional shift than the single model. MWSE assesses the sensitivity of a model to distribution shift by using error and uncertainty in conjunction with one another. The MWSE-retention curve when studied in tandem with the MSE-retention curve gives more insight into the quality of estimated uncertainty.

3 Conclusion and Future Work

This work presents three primary metrics (RMSE/RMV ratio, LENCE and LL-FU) to quantify and evaluate predictive uncertainty under dataset shift. We used the Shifts Weather Prediction Dataset[13] for validation of the proposed metrics. Through these metrics we aimed to jointly evaluate predictive uncertainty and robustness of a model and its performance under dataset shift. While we focused only on regression tasks, we believe there is potential for these metrics to be extrapolated to classification as well as machine translation tasks. Another avenue of future research can be to incorporate the memory and computational efficiency of different methods into the evaluation metrics. We hope our

work helps the community and inspires further research on evaluation metrics for shifted multi-modal data.

Acknowledgments and Disclosure of Funding

We would like to thank *Mars Rover Manipal* for providing the necessary resources for our research.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] T. Boobier. *AI and the Future of Banking*. John Wiley & Sons, 2020.

[3] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[4] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[5] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[6] J. Ker, L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

[7] P. W. Koh, S. Sagawa, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[8] V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.

[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[10] D. Levi, L. Gispan, N. Giladi, and E. Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *arXiv preprint arXiv:1905.11659*, 2019.

[11] A. Ly, M. Marsman, J. Verhagen, R. P. Grasman, and E.-J. Wagenmakers. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

[12] A. Malinin. *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge, 2019.

[13] A. Malinin, N. Band, G. Chesnokov, Y. Gal, M. J. Gales, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

[14] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[15] A. M. Nascimento, L. F. Vismari, P. S. Cugnasca, J. Camargo, J. de Almeida, R. Inam, E. Fersman, A. Hata, and M. Marquezini. Concerns on the differences between ai and system safety mindsets impacting autonomous vehicles safety. In *International Conference on Computer Safety, Reliability, and Security*, pages 481–486. Springer, 2018.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*, 2017.

[17] A. Ustimenko, L. Prokhorenkova, and A. Malinin. Uncertainty in gradient boosting via ensembles. *ArXiv*, abs/2006.10562, 2021.

[18] C. Villani. A review of mathematical topics in collisional kinetic theory. *Handbook of mathematical fluid dynamics*, 1(71-305):3–8, 2002.