Deakin University

Faculty of Science, Engineering and Built Environment

School of IT

Performance of State-of-the-Art Machine Learning Methods on Out-of-Distribution Data in Tabular Regression: A Survey

Supervisor: Prof. Gleb Beliakov and A/Prof. Simon James

Author's Name: Bao Minh Tran

Email: s224236373@deakin.edu.au

November 12^{th}, 2025

Abstract

1 Introduction

Artificial Intelligence (AI) [1] is a broad computer science field that studies about softwares that can mimic the human's capability of learning and problem solving. Machine Learning (ML) is a sub-field of AI, which [2] applies statistical methods to explore the underlying correlation among the features and the target. Nonetheless, ML faces difficulties when dealing with high data volume and dimensions. A sub-field of ML has been studied to overcome these challenges, i.e. Deep Learning [3], which has proven its superiority over conventional ML approaches, and yielded prospective results in Natural Language Processing (NLP), Computer Vision (CV), Large Language Models (LLMs), and also being applied to other industries ([4], [5]), e.g. healthcare, finance, energy, and agriculture.

Since 1940s [6], several ML/DL models have been proposed, and various benchmarks have been introduced to compare the performance among them. A critical fault is that these benchmarks often ([7], [8], [9], [10]) overlook the distributional shifts between the training data and deployment data, alternatively, they assume that the traing and testing sets are independent and identically distributed (i.i.d). These inherent distributional shifts are due to various factors, e.g. [8] sample biases, environment changing, data generation errors, and [11] learning spurious correlations. Many studies have been contributed to showing [11] distributional shifts in real data are roadblock to the ML/DL applications to other fields. In Computer Vision, [12] showed that recent image classifiers fail to OOD generalization and [13] introduced a benchmark for performance evaluation among object detectors, namely ObjectNet, which showed performance drop of top models, from 2012 to 2018, in comparision to other standard benchmarks, e.g. ImageNet. Regarding Natural Language Processing, [14] proposed a benchmark, called CheckList, that leads to failure in semantic analysis due to tiny variations in the texts.

Although ([15], [16]) tabular data is the primary structure to retain data for training models in numerous areas, e.g. finance ([17], [18]), healthcare ([19], [20]) and manufacturing sectors [21], the number of standardized tabular regression benchmarks are very limited ([22], [15]). Therefore, this paper serves as a benchmark consisting of various train/test splitting strategies on real datasets to evaluate model's capability towards OOD generalization on tabular regression tasks.

2 Related works

Comment (delete after finished): this is a benchmark report, only cite the papers contributing to benchmarks.

3 Problem Statement

3.1 Problem Formulation

3.2 Notation

3.3 Definition 1. Distributional shifts

3.4 Definition 2. Convex hull of dataset - extrapolation & interpolation

4 Benchmark

4.1 Dataset

4.2 Train/Test splitting strategy

4.3 Models

5 Experimental setup

5.1 Evaluation metrics

5.2 Hyperparameters selection methods

For each traing data, the model selection method is re-run to find the most optimal hyperparameter setting w.r.t the dataset.

6 Results

7 Conclusion and Discussion

References

[1] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–6, 2018.

[2] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, p. 685–695, Apr. 2021.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] N. Rane, S. K. Mallick, O. Kaya, and J. Rane, "Applications of deep learning in healthcare, finance, agriculture, retail, energy, manufacturing, and transportation: A review," in *Applied Machine Learning and Deep Learning: Architectures and Techniques*, ch. 7, pp. 132–152, Deep Science Publishing, October 2024.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[6] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Hoboken, NJ: Pearson, 4 ed., 2021.

[7] A. Malinin, N. Band, G. Chesnokov, Y. Gal, M. Gales, A. Noskov, A. Ploskonosov, L. Ostroumova Prokhorenkova, I. Provilkov, V. Raina, V. Raina, M. Shmatova, P. Tigas, and B. Yangel, "Shifts: A dataset of real distributional shift across multiple large-scale tasks," 07 2021.

[8] L. Tamang, M. R. Bouadjenek, R. Dazeley, and S. Aryal, "Handling out-of-distribution data: A survey," 2025.

[9] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press, 2009.

[10] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. Lanas Phillips, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "Wilds: A benchmark of in-the-wild distribution shifts," *arXiv preprint arXiv:2012.07421*, 2020.

[11] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," 2020.

[12] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," 2020.

[13] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[14] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter,

and J. Tetreault, eds.), (Online), pp. 4902–4912, Association for Computational Linguistics, July 2020.

[15] S. Kolesnikov, "Wild-tab: A benchmark for out-of-distribution generalization in tabular regression," 2023.

[16] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, p. 7499–7519, June 2024.

[17] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning in finance," *arXiv preprint arXiv:1602.06561*, 2018.

[18] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications : A survey," 2020.

[19] A. Nayyar, L. Gadhavi, and N. Zaman, "Chapter 2 - machine learning in healthcare: review, opportunities and challenges," in *Machine Learning and the Internet of Medical Things in Healthcare* (K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar, eds.), pp. 23–45, Academic Press, 2021.

[20] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. V. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, 2018.

[21] J. Leukel, J. González, and M. Riekert, "Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review," *Journal of Manufacturing Systems*, vol. 61, pp. 87–96, 2021.

[22] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022.