

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 12th, 2025

Contents

| | | |
|-------------------|--|----------|
| 1 | Introduction | 2 |
| 2 | Related works | 3 |
| 3 | Problem settings | 4 |
| 3.1 | Notation | 4 |
| 3.2 | Definition 1. Distributional shifts | 4 |
| 3.3 | Definition 2. Convex hull of dataset | 4 |
| 4 | Benchmark | 4 |
| 4.1 | Dataset | 4 |
| 4.2 | Train/Test splitting strategy | 4 |
| 4.3 | Models | 4 |
| 5 | Experimental setup | 4 |
| 5.1 | Evaluation metrics | 4 |
| 5.2 | Hyperparameters selection methods | 4 |
| 6 | Results | 4 |
| 7 | Analysis | 4 |
| 8 | Conclusion and Discussion | 4 |
| Appendices | | 8 |
| A | Datasets | 8 |
| B | Benchmark results | 8 |

Abstract

The **Wild-Tab** benchmark [1] evaluates models’ robustness under covariate, subpopulation, and hybrid shifts of large-sized real-world datasets. This paper also conducted an empirical study by implementing baseline DL models, e.g. MLP-PLR, FT-Transformer, and 10 other OOD generalization techniques, e.g. IRM, ERM, and GroupDRO, that are specifically designed for tabular regression tasks, and gave certain insightful analysis, i.e. ERM outperforms other models, and the impact of model architecture and hyperparameter tuning methods on models’ performance. In contrast, [2] offers a benchmark to compare between the tree-based models and DL models on OOD data in tabular regression tasks. More specifically, [2] leverages 45 tabular datasets, mostly from OpenML [3], and cleans and prepares properly. Later, they benchmarked 3 tree-based models, i.e. Random Forest, XGBoost, and Gradient Boosting Trees (or HistGradient Boosting Trees in case of categorical inputs) and 4 distinct DL models, i.e. classical MLP, ResNet, FT-Transformer, and SAINT. Also, they clearly interpreted why tree-based models outperform DL models in tabular regression on OOD data regardless of models’ complexity. Despite limited available benchmarks for tabular regression on OOD data, these prior works inspired us to conduct a more comprehensive empirical study to compare the performance of various ML/DL models on OOD data that are created by applying diverse splitting strategies to real-world datasets.

1 Introduction

Artificial Intelligence (AI) [4] 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 [5] 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 [6], 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 ([7], [8]), e.g. healthcare, finance, energy, and agriculture.

Since 1940s [9], 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 ([10], [11], [12], [13]) overlook the distributional shifts between the training data and deployment data, alternatively, they assume that the training and testing sets are independent and identically distributed (i.i.d.). These inherent distributional shifts are due to various factors, e.g. [11] sample biases, environment changing, data generation errors, and [14] learning spurious correlations. Many studies have been contributed to showing [14] distributional shifts in real data are roadblock to the ML/DL applications to other fields, and many [15] benchmarks and [16] AI practitioners failed to analyse the models’ performance on Out-of-Distribution (OOD) data. In Computer Vision, [17] showed that recent image classifiers fail to OOD generalization and [18] introduced a benchmark for performance evaluation among object detectors, namely ObjectNet, which showed performance drop of top models, from 2012 to 2018, in comparison to other standard benchmarks, e.g. ImageNet. Regarding Natural Language Processing, [19] proposed a benchmark, called CheckList, that leads to failure in semantic analysis due to tiny varia-

tions in testing texts compared to those in the training set.

Although ([1], [20]) tabular data is the primary data structure to retain data for training models in numerous areas, e.g. finance ([21], [22]), healthcare ([23], [24]) and manufacturing sectors [25], the number of standardized tabular regression benchmarks are very limited ([2], [1]). 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. The key contributions of our work are:

- Introduction to a comprehensive benchamrk consisting of 29 real-world low dimensional (up to 20 features) tabular datasets specifically for regression tasks. These datasets are subsequently partitioned into several train/test pairs aiming for evaluating the models’ OOD generalization. To our knowledge, this marks the first benchamrk that encompasses certain approaches to split datasets for OOD extrapolation assessment.
- Conduct an experimental study to quantify and rank the robustness of a broad spectrum of standard ML/DL models on unseen distributions.
- Give insightful analysis to explore what characteristics of the proposed ML/DL methods impact their performance when generalising to unseen data regions.

2 Related works

The aim of this paper is to investigate the performance of widely used ML models on Out-of-Distribution (OOD) tabular data. Thus, plenty of papers were introduced to formally define what it means for models to extrapolate on unseen data. Many papers ([26], [27], [28]) have supported that extrapolation to OOD data refers to models predicting data points lying beyond the **convex hull of training samples**. Nevertheless, a more prevalent definition of OOD data highlights the **distributional shifts between the training and testing sets**, and is commonly used to compare models’ generalization on OOD data ([12], [10], [15], [2], [14], [1]). To our knowledge, there are none of benchmarks for tabular regression tasks on OOD data that is built upon the first definition. Meanwhile, there are contributions, though still very limited, to benchmarking models on OOD data, based on distributional shifts, in tabular regression tasks, e.g. **Wild-Tab** [1], **Shifts Dataset** [10], and an unnamed benchmark provided by [2].

The **Wild-Tab** benchmark [1] evaluates models’ robustness under covariate, subpopulation, and hybrid shifts of large-sized real-world datasets. This paper also conducted an empirical study by implementing baseline DL models, e.g. MLP-PLR, FT-Transformer, and 10 other OOD generalization techniques, e.g. IRM, ERM, and GroupDRO, that are specifically designed for tabular regression tasks, and gave certain insightful analysis, i.e. ERM outperforms other models, and the impact of model architecture and hyperparameter tuning methods on models’ performance. In contrast, [2] offers a benchmark to compare between the tree-based models and DL models on OOD data in tabular regression tasks. More specifically, [2] leverages 45 tabular datasets, mostly from OpenML [3], and cleans and prepares properly. Later, they benchmarked 3 tree-based models, i.e. Random Forest, XGBoost, and Gradient Boosting Trees (or HistGradient Boosting Trees in case of categorical inputs) and 4 distinct DL models, i.e. classical MLP, ResNet,

FT-Transformer, and SAINT. Also, they clearly interpreted why tree-based models outperform DL models in tabular regression on OOD data regardless of models' complexity. Despite limited available benchmarks for tabular regression on OOD data, these prior works inspired us to conduct a more comprehensive empirical study to compare the performance of various ML/DL models on OOD data that are created by applying diverse splitting strategies to real-world datasets.

3 Problem settings

3.1 Notation

3.2 Definition 1. Distributional shifts

3.3 Definition 2. Convex hull of dataset

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

6 Results

7 Analysis

8 Conclusion and Discussion

References

- [1] S. Kolesnikov, “Wild-tab: A benchmark for out-of-distribution generalization in tabular regression,” 2023.
- [2] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on tabular data?,” 2022.
- [3] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, “Openml: networked science in machine learning,” *ACM SIGKDD Explorations Newsletter*, vol. 15, p. 49–60, June 2014.
- [4] 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.
- [5] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” *Electronic Markets*, vol. 31, p. 685–695, Apr. 2021.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [7] 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.
- [8] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [9] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Hoboken, NJ: Pearson, 4 ed., 2021.
- [10] A. Malinin, N. Band, G. Chesnokov, Y. Gal, M. Gales, A. Noskov, A. Ploskonosov, L. Ostroumova Prokhorenkova, I. Prosvirkov, 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.
- [11] L. Tamang, M. R. Bouadjenek, R. Dazeley, and S. Aryal, “Handling out-of-distribution data: A survey,” 2025.
- [12] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press, 2009.
- [13] 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.
- [14] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” 2020.
- [15] I. Deeva, N. Amerkhanova, and A. Kropacheva, “Evaluating robustness of tabular models under meta-features based shifts,” *arXiv preprint*, 2023. AI Institute, ITMO University.

- [16] A. Malinin, A. Athanasopoulos, M. Barakovic, M. B. Cuadra, M. J. F. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.-J. Lu, N. Molchanova, A. Nikitakis, V. Raina, F. L. Rosa, E. Sivena, V. Tsarsitalidis, E. Tsompopoulou, and E. Volf, “Shifts 2.0: Extending the dataset of real distributional shifts,” 2022.
- [17] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” 2020.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning in finance,” *arXiv preprint arXiv:1602.06561*, 2018.
- [22] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, “Deep learning for financial applications : A survey,” 2020.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] R. Balestriero, J. Pesenti, and Y. LeCun, “Learning in high dimension always amounts to extrapolation,” 2021.
- [27] S. Zhang, Y. Luo, Q. Wang, H. Chi, W. Li, B. Han, and J. Li, “Are all unseen data out-of-distribution?,” *CoRR*, vol. abs/2312.16243, 2023.

- [28] H. Hwang and J. Shin, “Uncertainty measurement of deep learning system based on the convex hull of training sets,” 2024.
- [29] A. Malinin, A. Athanasopoulos, M. Barakovic, M. Bach Cuadra, M. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.-J. Lu, N. Molchanova, A. Nikitakis, V. Raina, F. La Rosa, E. Sivena, V. Tsarsitalidis, E. Tsompopoulou, and E. Volf, “Shifts marine cargo vessel power consumption prediction dataset,” Sept. 2022.

Appendices

A Datasets

B Benchmark results