Deep Learning on Tabular Data

Parul Chauhan • *pchauhan2022@fau.edu*

INTRODUCTION

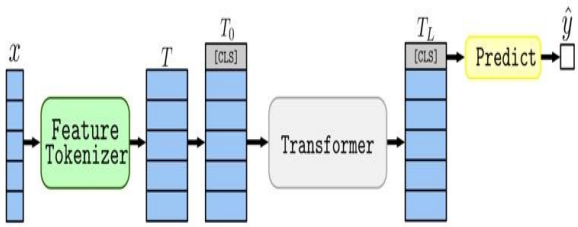
Deep learning on tabular data has become a popular topic in recent years due to the success of deep learning models in various applications and the abundance of tabular data in many domains. In this paper, we will provide a brief overview of deep learning on tabular data and discuss its potential benefits and challenges.

CURRENT SCENARIO

Deep learning is a subset of machine learning that uses deep neural networks to learn complex relationships from data. These networks are composed of multiple layers of interconnected nodes, which enable them to learn hierarchical representations of the data. This allows deep learning models to capture the complex non-linear relationships that often exist in data, which makes them well-suited for many applications.

Deep learning models may be tricky to analyze and comprehend, which makes it difficult to justify their predictions or modify the model. Some of the distinctive aspects of tabular data, such as missing values, outliers, or strongly correlated attributes, may be too complex for deep learning models to manage.

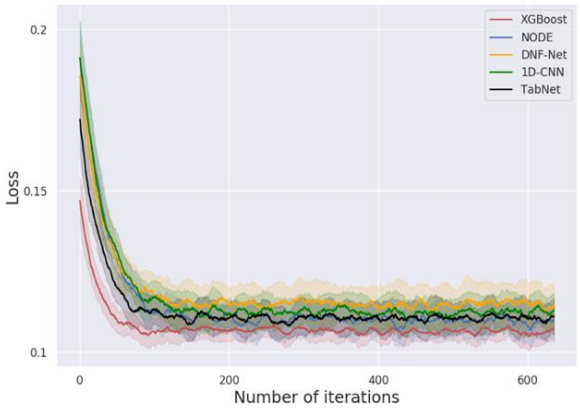
One of the potential benefits of using deep learning on tabular data is improved performance on predictive tasks. Deep learning models have been shown to outperform traditional machine learning methods on many tasks, and this is especially true when dealing with complex, non-linear relationships in the data. For example, in a study comparing deep learning and traditional machine learning methods on a financial dataset, the deep learning model was able to achieve a higher accuracy on the prediction task[2]. This is because deep learning models are able to capture the complex relationships between the different features in the data, which allows them to make more accurate predictions. When compared to deep learning models, conventional machine learning methods, such ridge regression, require the introduction of feature reduction strategies since they operate less effectively[4].



Source: Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021, November 10). *Revisiting deep learning models for tabular data*. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/2106.11959

Deep learning models are able to automatically learn which features are important and which can be ignored, which makes them well-suited for datasets with a large number of features[3]. There are variables like tuning parameter penalty and the L1 and L2 ratio that govern selection processes and the computational resources required to perform it for feature selection in regression models like elastic net and lasso[4]. Because of this, it is frequently difficult for the model to be effective when huge datasets are involved due to the needed processing power. This is the reason why deep learning models are particularly useful in fields such as healthcare, where datasets often contain a large number of features such as patient medical history, test results, and other information. By using deep learning models, it is possible to automatically select the most important features and use them to make predictions, which can improve the performance of the model and make it more efficient. In general, the decision between using an elastic net model or a deep learning model depends on the particular issue you are attempting to resolve as well as the features of your data. It is crucial to carefully assess which model is ideal for your use case because both types of models have benefits and drawbacks[2][4].

CHALLENGES



Source: Shwartz-Ziv, R., & Armon, A. (2021, November 23). *Tabular data: Deep Learning is not all you need*. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/2106.03253

Deep learning on tabular data is not without its difficulties, though. The requirement for a significant volume of labeled data is one of the key obstacles. When dealing with tabular data, deep learning models can be quite difficult to train since they need a lot of data. In many cases, the available datasets

may not be large enough to train a deep learning model, or they may not be well-labeled, which can make it difficult to train the model effectively. While deep learning has greatly advanced text and picture datasets, it is unclear if it is preferable for tabular data. We provide thorough benchmarks of well-known and cutting-edge deep learning techniques as well as tree-based models like Random Forests and XGBoost, spanning a wide range of datasets and hyperparameter combinations[5]. Additionally, deep learning models can be difficult to interpret, which can make it challenging to understand how they make predictions and identify potential errors. This can be a significant limitation, as it can make it difficult to understand the decisions made by the model and how to improve it.

Despite these challenges, deep learning on tabular data has the potential to provide significant benefits in a variety of applications. For example, in the healthcare field, deep learning models can be used to predict patient outcomes and identify potential risk factors. This can help doctors to make more informed decisions and provide better care to their patients. In the finance industry, deep learning models can be used to predict stock prices and identify potential investment opportunities. This can help investors to make more informed decisions and potentially increase their returns.

Conclusion

In conclusion, deep learning on tabular data has the potential to improve performance on predictive tasks and handle large numbers of features. However, it also presents challenges such as the need for large amounts of labeled data and difficulties in interpretation. Despite these challenges, deep learning on tabular data has the potential to provide significant benefits in a variety of ways. Overall, even though deep learning may be an effective tool for handling tabular data, it is crucial to carefully analyze these issues and decide if deep learning is the appropriate strategy for your particular use case.

References

[1] IEEE. (n.d.). *Submitted to the IEEE, June 2022 1 deep neural networks and ... - arxiv*. Deep Neural Networks and Tabular Data: A Survey. Retrieved December 12, 2022, from https://arxiv.org/pdf/2110.01889.pdf

[2] *Zou - Royal Statistical Society - Wiley Online Library*. (n.d.). Retrieved December 12, 2022, from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x

[3] Abutbul, A., Elidan, G., Katzir, L., & El-Yaniv, R. (2020, June 11). *DNF-net: A neural architecture for tabular data*. arXiv.org. Retrieved December 11, 2022, from https://arxiv.org/abs/2006.06465

[4] *Big Data Deep Learning: Challenges and perspectives | IEEE journals ...* (n.d.). Retrieved December 12, 2022, from https://ieeexplore.ieee.org/abstract/document/6817512

[5] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022, July 18). *Why do tree-based models still outperform deep learning on tabular data?* [2207.08815] Why do tree-based models still outperform deep learning on tabular data? Retrieved December 11, 2022, from http://export.arxiv.org/abs/2207.08815