Adversarial Perturbation Techniques and Their Impact on Alignment and Fairness in Tabular Foundation Models

Assignee Research

June 11, 2026

Abstract

Recent deep learning models for tabular data currently compete with the traditional ML models based on decision trees (GBDT). Unlike GBDT, deep models can additionally benefit from pretraining, which is a workhorse of DL for vision and NLP. For tabular problems, several pretraining methods were proposed, but it is not entirely clear if pretraining provides consistent noticeable improvements and what method should be used, since the methods are often not compared to each other or comparison is limited to the simplest MLP architectures. In this work, we aim to identify the best practices to pr

1 Introduction

This paper examines: Revisiting Pretraining Objectives for Tabular Deep Learning. Research question: How do different adversarial perturbation techniques during pretraining (e.g., FGSM, PGD) influence the alignment and fairness of tabular foundation models, as evaluated by metrics like demographic parity on tabular fairness benchmarks like TabFair?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.9/10.

3 Results

12 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 7.9/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Recent deep learning models for tabular data currently compete with traditional machine learning models based on decisio	✓	0.32
Unlike GBDT, deep models can benefit from pretraining.	✓	0.22
Pretraining is a workhorse of deep learning for vision and NLP.	✓	0.17
Several pretraining methods have been proposed for tabular problems.	✓	0.17
Existing pretraining methods for tabular data are often not compared to each other.	×	0.10
Existing comparisons of tabular pretraining methods are often limited to the simplest MLP architectures.	✓	0.18
Using object target labels during the pretraining stage is beneficial for downstream performance.	✓	0.30
Properly performed pretraining significantly increases the performance of tabular deep learning models.	✓	0.31
Properly performed pretraining often leads to tabular deep learning models achieving superiority over GBDTs.	✓	0.19

References

- http://arxiv.org/abs/2307.02055v1
- http://arxiv.org/abs/2207.03208v2
- http://arxiv.org/abs/2512.03307v1