传统数据挖掘算法与AI大模型的协同范式

刘帅 | 爱心人寿信息技术部门负责人

场景：
训练保险反欺诈预测模型

- 样本数据：100万案例
- 特征指标：100个
- 标签：欺诈Yes, 非欺诈No

问题引入

梯度提升决策树(XGBoost)

Vs

AI大模型

数据集特点

结构化数据

（表格数据集 Tabular Data ）

非结构化数据

如文本、图像、声音等

同质数据集

数据类型多种，如数值、时间、类型等

异质数据集

数据类型相同

模型数值试验-数据集

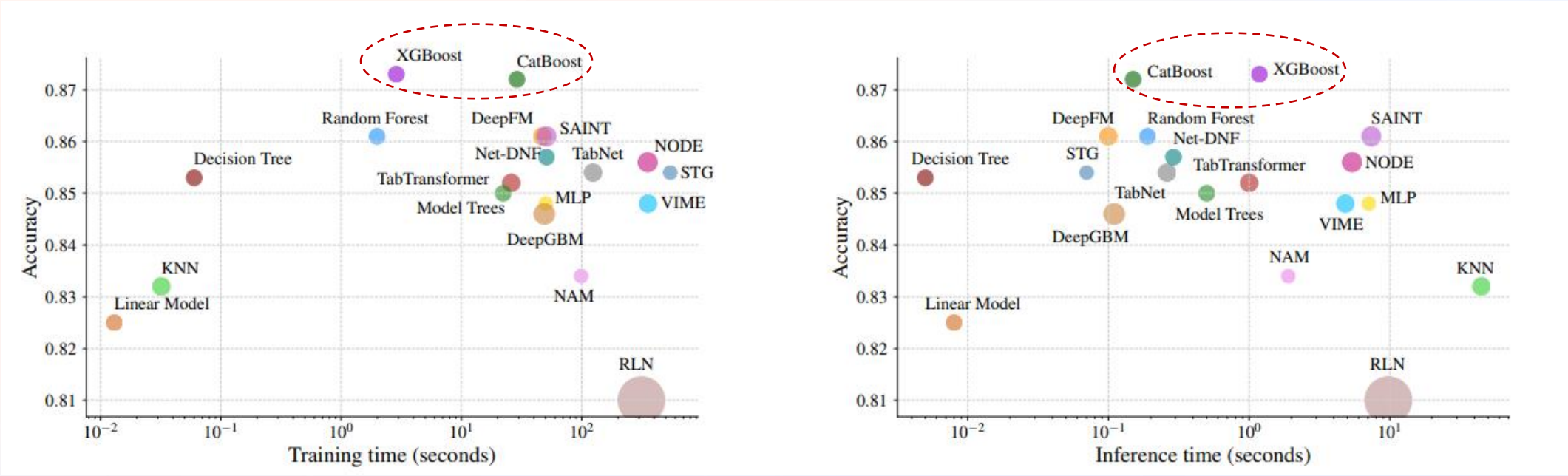
- **HELOC**（FICO提供的信用风险评估数据，真实用户金融记录）
- **Adult Income**（UCI人口普查数据，真实个人收入与人口统计信息）
- **HIGGS**（高能物理数据，真实粒子对撞实验记录）
- **Covertype**（森林植被类型数据，真实地理与环境测量）
- **California Housing**（加州房价数据，真实房产市场记录）

	HELOC	Adult Income	HIGGS	Covertype	California Housing
Samples	9.871	32.561	11 M.	581.012	20.640
Num. features	21	6	27	52	8
Cat. features	2	8	1	2	0
Task	Binary	Binary	Binary	Multi-Class	Regression
Classes	2	2	2	7	-

相关数据和观点引自：《Deep Neural Networks and Tabular Data: A Survey》， Vadim Borisov 等

模型数值试验-结果

- 除超大规模的HIGGS数据集外，梯度提升决策树集成模型（如XGBoost、CatBoost）在中小型数据集、回归任务及多分类任务中均优于所有深度学习方法。



原因分析

- 训练数据质量差：不完整
- 缺乏像图像/文本数据那样的显式空间（如像素邻接）或语义（如词序）关联，导致深度模型难以捕捉复杂特征交互
- 预处理依赖：严重依赖于预处理策略（尤其是类别特征处理），这可能引入信息损失或人为偏差
- 表格数据中的单个特征（尤其是类别或二值特征）可能对模型预测结果产生决定性影响，而深度神经网络（DNN）在处理这种特性时存在天然劣势

核心结论

- **传统方法仍主导中小型表格数据**：梯度提升树（GBDT）凭借高效特征选择与鲁棒性，仍是实际应用首选。
- **深度模型的曙光**：SAINT在超大规模数据中的突破，暗示未来可能通过架构优化与计算效率提升打破性能瓶颈。
- **领域依赖性**：深度模型性能高度依赖数据特性（如特征类型、规模），需“量体裁衣”选择模型。

现阶段的应用策略：传统数据挖掘算法+AI大模型

- **AI大模型**：解锁非结构化数据的价值，捕捉传统方法难以发现的模式
- **传统数据挖掘算法**：高效处理结构化数据，提供可解释、稳定的预测结果

数据预处理与特征工程

- ✓ 非结构化转为结构化数据
- ✓ 缺失数据补录
- ✓ 文本特征提取
- ✓ 自动化特征生成

+

模型训练环节：传统数据挖掘算法

前沿研究方向-AMFormer框架

AMFormer通过**并行加法和乘法注意力**分别建模线性和非线性特征交互，并通过**硬注意力**和**提示优化**提升效率与鲁棒性。

- **显式算术交互**：通过并行加法和乘法注意力机制，直接建模线性和非线性特征交互（如 x1+x2x1+x2 和 x1×x2x1×x2）
- **效率优化**：引入硬注意力（Top-k稀疏化）和提示向量（Prompt-based Optimization），降低计算复杂度
- **实验验证**：在合成数据和真实数据中，AMFormer在细粒度建模、数据效率和泛化能力上优于传统Transformer和树模型

Dataset	XGBoost	AMF-A	AMF-F
EP ↑	87.32 (8)	89.71 (2)	**89.83** (1)
HC ↑	74.59 (7)	75.57 (3)	**75.67** (1)
CO ↑	96.72 (3)	**97.36** (1)	97.26 (2)
MI ↓	0.5642 (3)	0.5606 (2)	**0.5557** (1)

说明：观点和数据来着于《Arithmetic Feature Interaction Is Necessary for Deep Tabular Learning 》Yi Cheng, Renjun Hu 等

谢谢大家的聆听！