Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning

Simon A. Lee[1], Kyoka Ono[2]

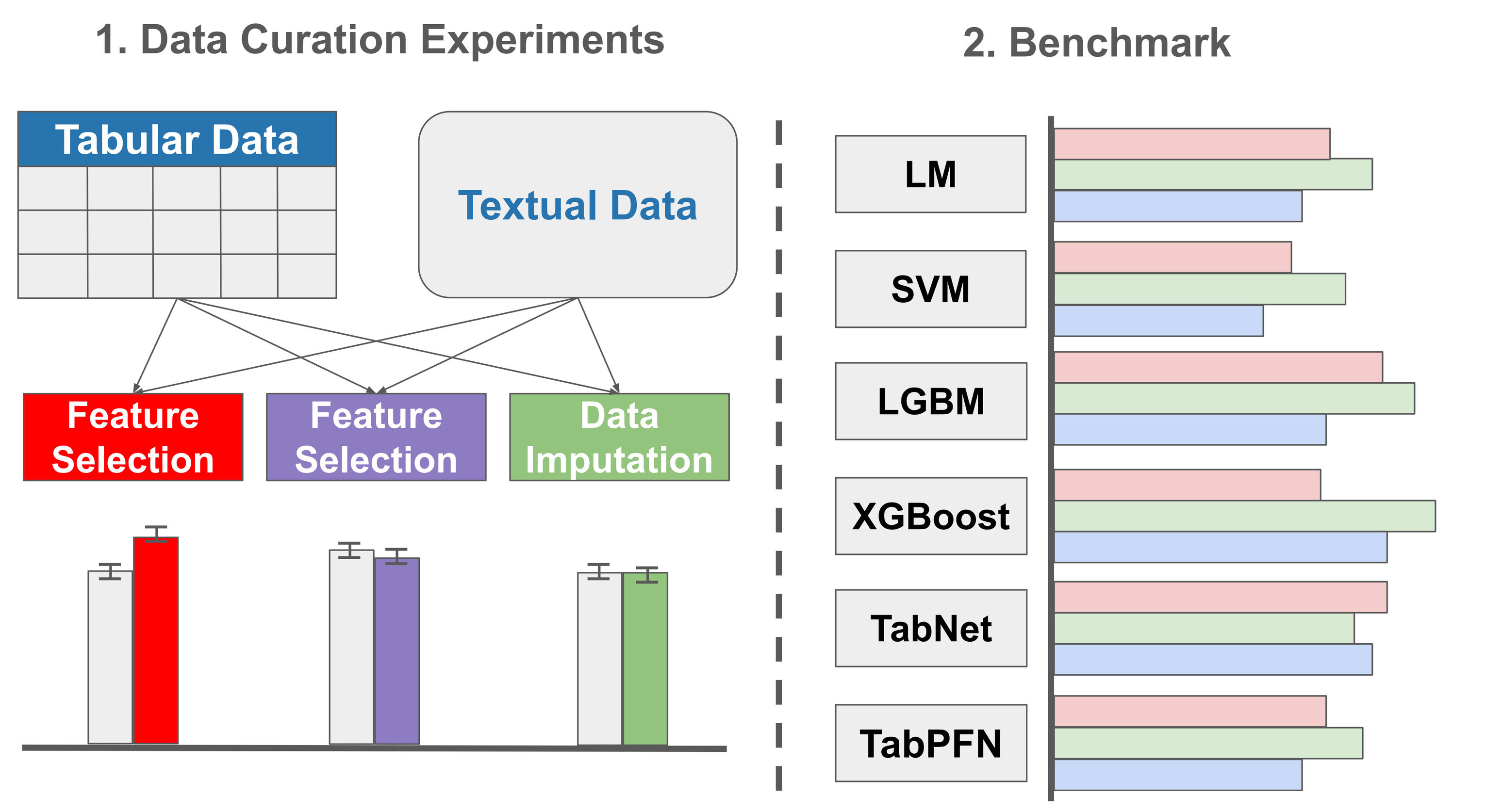
1. Department of Computational Medicine, University of California, Los Angeles
2. Department of Statistics & Data Science, University of California, Los Angeles

1. Motivation

Background: The most common machine learning (ML) tasks use tabular datasets, organized in table format. Recent advancements in language models (LMs) prompt a need to understand how these models and methods align with traditional ML paradigms.

Problem Formulation: This research aims to address two questions:

1. Does Text serialization require similar data curation techniques as tabular data?
2. How do pre-trained language models with supervised fine tuning (SFT) compare to traditional ML models and deep learning tabular models?



2. Text Serialization

Text Serialization: Converting Data From Tabular to Text Derived from TabLLM [1]. Analogous to a game of Madlibs.

Titanic Dataset		
Name	**Age**	**Sex**
…	26	M
…	54	F

Text Serialization

$$\Phi : X \rightarrow T$$

Passenger John Doe is a 26 year old male...

Passenger [Name] is a [Age] year old [Sex]...

A methodology for generating text from tabular data

```
template = f"Passenger {df[name]} is a {df[age]} year old
{df[sex]}"
```

3. Datasets

Dataset	Sample Size (n)	# of Features (m)	Binary	B/E
Iris	150	4	✗	B
Diabetes	784	8	✓	B
Titanic	891	11	✓	B
Wine	178	13	✗	B
HELOC	10,459	23	✓	E
Fraud	284,807	30	✓	E
Crime	878,049	8	✗	E
Cancer	801	20,533	✗	E

- **Baseline Datasets (B):** The baseline datasets, primarily sourced from the UCI Machine Learning Repository [2], are used for pre-processing experiments and benchmark studies.
- **Experimental Datasets (E):** The experimental datasets consist of tabular data with unique characteristics for benchmark studies, including distribution shifts, bias, high dimensionality, and class imbalance.

4. Pretrained-Model Selection 🤗

Model	Loss	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC	Runtime (s)	Samples/s
Bert	0.4903	0.7821	0.7536	0.7027	0.7273	0.8483	0.8262	5.0933	35.144
DistilBert	**0.4535**	0.8045	0.7097	**0.8919**	0.7904	**0.8743**	**0.8426**	2.6072	68.656
RoBERTa	0.5547	0.7989	0.7317	0.8108	0.7692	0.8206	0.7448	4.7434	37.737
Electra	0.4583	**0.8268**	0.7529	0.8649	**0.8050**	0.8515	0.7665	5.1101	35.029
XLNet	0.5574	0.7821	0.7536	0.7027	0.7273	0.8529	0.8222	17.336	10.325
Albert	0.4802	0.7989	0.7262	0.8243	0.7722	0.8387	0.7637	5.8252	30.729
DeBERTa	0.5057	0.7933	0.7342	0.7838	0.7582	0.8059	0.7006	3.2567	54.964
GPT2	0.6947	0.6592	0.8824	0.2027	0.3297	0.8408	0.7877	**2.0704**	**86.456**
Longformer	0.5092	0.7989	0.7436	0.7838	0.7632	0.8138	0.6742	3.7726	47.447
GTE-large	0.5226	0.7933	0.7761	0.7027	0.7376	0.8704	0.7947	6.4885	27.587
GTE Base	0.5336	0.7821	**0.9070**	0.5270	0.6667	0.8725	0.8139	2.1677	82.575

- **Benchmark:** We select a model through benchmarks involving various pre-trained encoder language models, including those from the Massive Text Embedding Benchmark (MTEB) to represent modern methods. Due to compute limitations, we chose embedding models with fewer than 1 billion parameters.

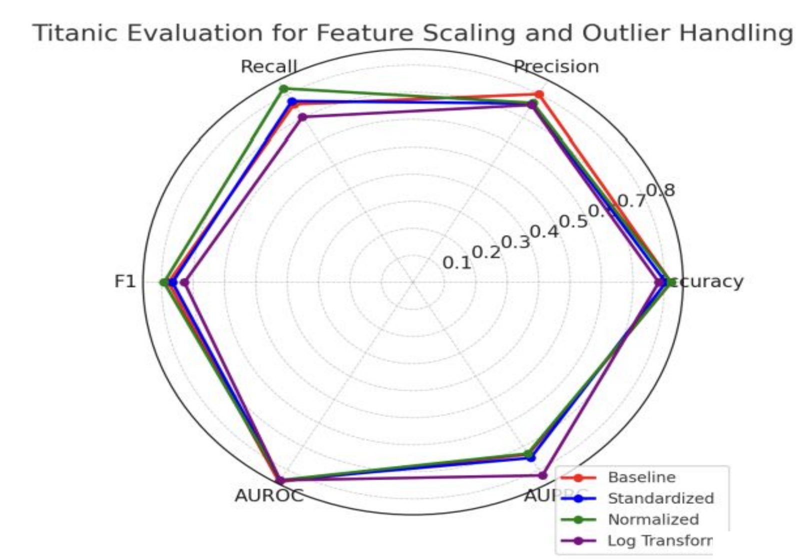
5. Data Curation Experiments

Feature Selection

- In two datasets, it has demonstrated that feature selection plays a critical role in optimizing performance.
- **This observation is analogous to established practices tabular machine learning.**

Table 2. Benchmark study with and without feature selection

Dataset	Without Feature Selection		With Feature Selection		Improved?
Metrics	AUROC	F1	AUROC	F1	
Iris	**1.000**	**1.000**	**1.000**	**1.000**	—
Wine	0.952	0.944	**0.976**	**0.972**	✓
Diabetes	0.654	0.621	**0.659**	**0.659**	✓
Titanic ♡	**0.786**	**0.871**	0.777	0.852	✗

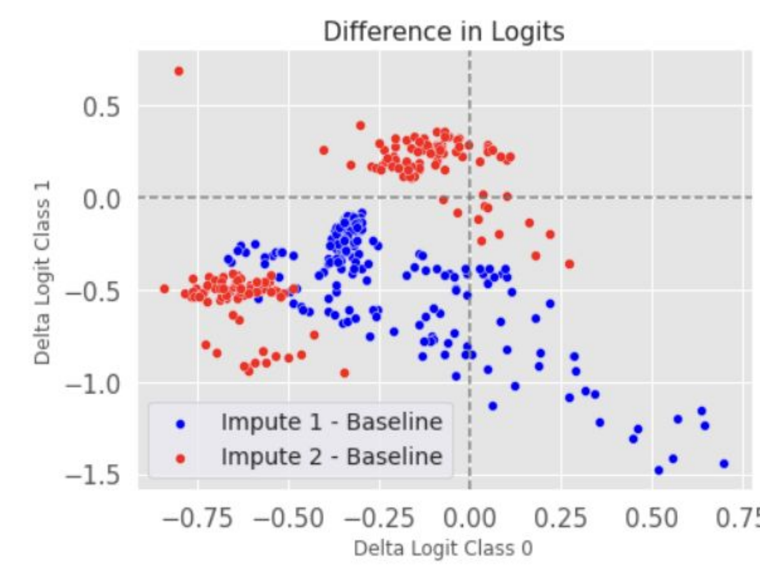


Feature Scaling

- Scaling numerical data may not always be effective, particularly when the dataset contains outliers or non-normal distribution.
- **This observation is different from established tabular machine learning practices.**

Data Imputation

- The method employed for data imputation significantly influences the probability of the final outcome.
- **This data curation technique should be further explored.**



6. Benchmark Results

Metrics: Accuracy, F1 Scores, Area Under the Receiver Operating Characteristic, Matthews Correlation Coefficient (MCC).
(Macro-averaging for non-binary classification cases)

State of the Art Evaluation - Baseline Datasets

Dataset	Method	Accuracy	F1	AUROC	MCC	Current State of the Art	TabLM SOTA?
Iris	SVM (RBF)	1.0000	1.0000	1.0000	1.1870	1.0000 (Acc)(Ojha & Nicosia, 2020)	✗
	LGBM	1.0000	1.0000	1.0000	1.1870		
	XGBoost	1.0000	1.0000	1.0000	1.1870		
	TabNet	1.0000	1.0000	1.0000	1.1870		
	TabPFN	1.0000	1.0000	—	1.1870		
	TabLM	1.0000	1.0000	1.0000	1.1870		
Wine	SVM (RBF)	0.8333	0.8107	0.9414	1.2004	0.9800 (Acc) (Di et al., 2020)	✗
	LGBM	**1.0000**	**1.0000**	**1.0000**	1.2089		
	XGBoost	0.9722	0.9663	1.0000	1.2133		
	TabNet	0.8333	0.8497	0.9503	0.7306		
	TabPFN	0.9800	0.9785	—	0.9704		
	TabLM	0.9722	0.9761	1.0000	**1.2147**		
Diabetes	SVM (RBF)	**0.7662**	0.7411	0.8044	0.4833	0.7879 (Acc) (Sarkar, 2022)	✗
	LGBM	0.7532	0.7334	0.8129	0.4671		
	XGBoost	0.7597	0.7301	0.8235	0.4640		
	TabNet	0.7273	0.6250	**0.8525**	0.4329		
	TabPFN	**0.7662**	**0.7433**	0.8211	**0.4870**		
	TabLM	0.6423	0.6594	0.6593	0.3962		
Titanic	SVM (RBF)	0.7765	0.7687	0.8654	0.5376	0.7985 (Acc) (Sarkar, 2022)	✓
	LGBM	0.7877	0.7747	**0.8995**	0.5572		
	XGBoost	0.7989	**0.7889**	0.8958	0.5812		
	TabNet	**0.8212**	0.7612	0.8938	**0.6192**		
	TabPFN	0.8101	0.7344	0.4747	0.5923		
	TabLM	**0.8212**	0.7777	0.8521	0.6001		

State of the Art Evaluation - Experimental Datasets

Dataset	Method	Accuracy	F1	AUROC	MCC	Current State of the Art	TabLm SOTA
HELOC	SVM (RBF)	0.7223	0.7207	0.7903	0.4426	N/A	✗
	LGBM	**0.7280**	**0.7267**	0.7958	**0.4541**		
	XGBoost	0.7170	0.7157	0.7746	0.4321		
	TabNet	0.7275	0.7070	**0.7966**	0.4532		
	TabPFN	0.7500*	0.7253*	0.4519*	0.5014*		
	TabLM	0.7157	0.7025	0.7939	0.4331		
Fraud	SVM (RBF)	0.9983	0.4996	0.4790	0.0000	0.9530 (AUROC) (Xu et al., 2023)	✗
	LGBM	0.9994	0.9075	0.9083	0.8167		
	XGBoost	**0.9996**	**0.9293**	**0.9811**	**0.8635**		
	TabNet	0.9994	0.8218	0.9640	0.8215		
	TabPFN*	—	—	—	—		
	TabLM	0.9988	0.9211	0.9155	0.8545		
Crime	SVM (RBF)	0.2006	0.0088	0.4849	0.2310	N/A	✓
	LGBM	0.2636	**0.0764**	0.6291	0.2395		
	XGBoost	0.2606	0.0756	0.6467	0.2389		
	TabNet	0.3087	0.0502	**0.7193**	0.2097		
	TabPFN*	—	—	—	—		
	TabLM	**0.3212**	0.0671	0.6789	**0.2437**		
Cancer	SVM (RBF)	1.0000	1.0000	1.0000	1.1428	N/A	✗
	LGBM	1.0000	1.0000	1.0000	1.1428		
	XGBoost	1.0000	1.0000	1.0000	1.1428		
	TabNet	0.9814	0.9735	0.9994	0.9749		
	TabPFN*	—	—	—	—		
	TabLM	0.9833	0.9826	0.9864	0.9792		

Verdict: Our benchmarking study reveals that pre-trained models with supervised fine-tuning (SFT) currently do not outperform traditional machine learning models and some deep learning tabular models, indicating future research directions for language models in solving tabular tasks.

7. Conclusions

- We identified that text serialization **differs** from tabular machine learning for data curation.
- We also identified that pre-trained models with supervised fine-tuning **do not represent a state of the art methodology** for tabular ML.

8. Citations

[1] Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., & Sontag, D. (2023). *TabLLM: Few-shot classification of tabular data with large language models.* arXiv.
[2] Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository
[3] Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). *MTEB: Massive Text Embedding Benchmark.* arXiv.

External Links

Poster Number

91