Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning

Sungwon Han, Jinsung Yoon, Sercan O Arik, Tomas Pfister

- ICML 2024

Importance of few-shot tabular learning

Datasets present labeling challenges.

- Tasks concerning rare diseases with few patients
- Tasks requiring specialized domain knowledge and expert input
- Tasks that are sensitive or private, making it hard to source annotators

...

In datasets with limited labels, conventional tabular models are prone to overfitting.
⇒ **Learn spurious correlations that do not reflect the actual patterns.**

Dealing with limited ground-truth labels

Leveraging prior knowledge about the problem **to provide an appropriate inductive bias** during model training

- Simultaneously train on various real-world benchmark tabular datasets [TransTab-NeurIPS'22]
- Utilize unlabeled dataset with self-supervised objective [SCARF-ICLR'21, STUNT-ICLR'23]
- Generate synthetic dataset with diverse distributions for pretraining [TabPFN-ICLR'23]
- **Utilize Large Language Model (LLM) for inference** [LIFT-NeurIPS'22, TabLLM-AISTATS'23, MediTab-Arxiv'23]

Three limitations of existing LLM-based approaches

1. At least one LLM query per sample is required for inference, making it **computationally expensive**.

2. Fine-tuning the LLM is often required, **limiting its application to full parameter accessible models**.

3. Most approaches are **not suitable with lengthy prompts** from high-dimensional tabular data.

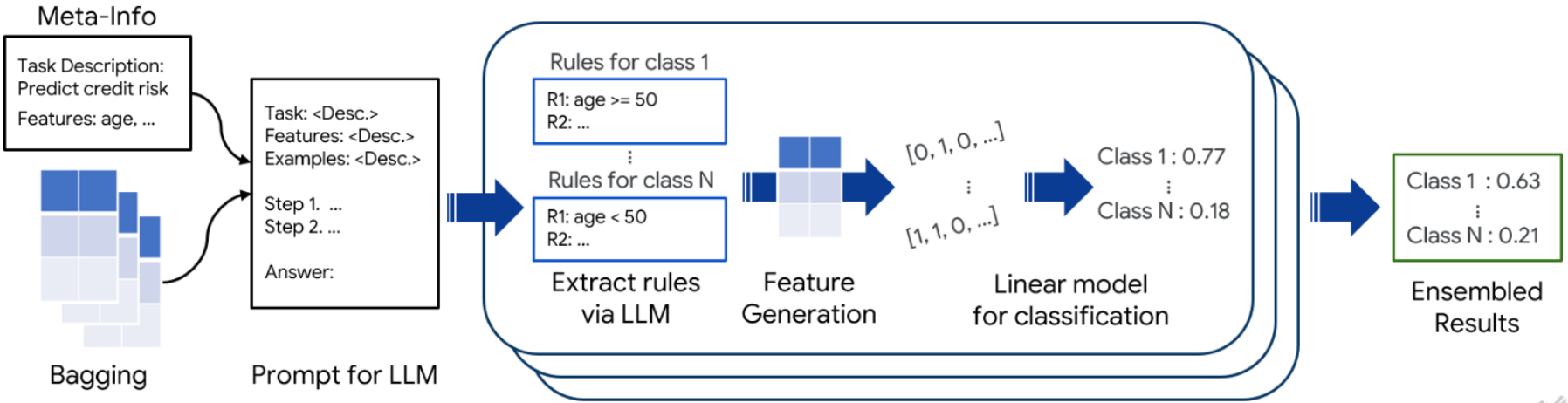
Why do these limitations occur?

⇒ Existing approaches utilize LLM to make an inference **sample by sample.**

Main Idea

Understand the “criteria” by which the LLM makes predictions.
Extract the underlying reasons, rather than running inference per each sample!

⇒ **Extract rules per each answer class!**



Highlighted Results

1. Evaluated over 11 tabular datasets, FeatLLM significantly outperforms baselines (10% on average) in few-shot settings.

Data	Shot	LogReg	XGBoost	SCARF	TabPFN	STUNT	In-context	TABLET	TabLLM	Ours
Average	4	65.47	50.00	58.22	62.93	62.36	68.44	68.69	70.26	**77.86**
	8	72.03	60.52	62.18	69.53	67.47	70.41	70.53	72.76	**79.31**
	16	76.33	69.72	71.69	74.37	69.72	72.72	73.02	76.22	**80.70**

Highlighted Results

2. FeatLLM even shows a relatively low inference time, comparable to that of conventional tabular methods (e.g., XGBoost).

Model	Training (in seconds)	Inference (in milliseconds)
LogReg	0.721	0.001
XGBoost	28.512	0.006
RandomForest	1.343	0.001
SCARF	426.859	0.002
TabPFN	0.440	1.149
STUNT	642.796	0.006
In-context†	N/A	463.000
TABLET†	0.813	523.254
TabLLM	251.242	335.127
FeatLLM†	860.094	0.006

† These models employ API queries, where the runtime is subject to the API's status at the time of use.

Highlighted Results

3. FeatLLM can handle high-dimensional tabular data (over 100 features) via feature bagging and ensembling.

Communities	Shots		
	4	8	16
LogReg	67.45±13.26	73.73±5.45	72.55±4.83
XGBoost	53.94±4.19	66.65±4.50	68.01±1.97
RandomForest	66.09±10.52	71.16±4.61	71.66±4.81
SCARF	66.18±9.13	72.69±3.79	73.09±2.84
STUNT	66.87±14.10	76.36±4.55	**77.29±2.56**
FeatLLM	**75.39±5.05**	**76.59±1.25**	76.25±0.64

Myocardial	Shots		
	4	8	16
LogReg	51.25±3.85	55.34±1.11	60.00±5.16
XGBoost	50.00±0.00	55.63±2.92	56.55±12.22
RandomForest	51.91±4.49	52.77±5.83	54.16±4.53
SCARF	47.70±4.10	49.37±3.41	54.31±1.42
STUNT	52.77±2.01	55.40±4.41	**61.22±3.45**
FeatLLM	**52.87±3.44**	**56.22±1.64**	55.32±9.15

Google KAIST

ICML International Conference On Machine Learning

Code:

Paper: