Developing a Large Tabular Model for Predicting Credit Approvals & Analyzing Customer Behavior in Financial Institutions

Columbia University Data Science Institute

Atharva Kulkarni, Nhat Minh Dang, Nicolás Corgono, Ryu Sonoda, Varun Agarwal, Yu Zheng Lim

Industry Mentors: Raveena Samtani, Urvi Palvankar | **Faculty Mentor:** Sining Chen

TD Bank

Introduction

Traditional machine learning models (TradML) often struggle to generalize to new datasets or tasks without extensive retraining and data preprocessing. Through this project, we explore the LLaMA 2 Generative Tabular Learning (GTL) models (LLaMA 2 pretrained with generic tabular data) to tackle these challenges, harnessing their tabular data generalization capabilities, with no retraining/fine-tuning. With capabilities like zero-shot generalization & in-context learning, we aim to predict credit approval for TD Bank by leveraging a publicly available dataset.

Methodology

Dataset, preprocessing: Kaggle. Features - removed credit score (data leakage), impute missing data, OHE. Target - 4 approval classes (P1-3 "approved" as Class 0, P4 "not approved" as Class 1)

Baseline: TradML models (Logistic Regression, Decision Tree, Random Forest, XGBoost). XGBoost, if given full training data & features, had the best F1 score (0.758). But to compare fairly with GTL models (4096 token limit) that can only see limited training data & features, we impose the same limited-data/feature treatment to traditional ML models.

GTL models: Tested with 2 prompt templates. T-table (uses feature, label description and column header for data) and T-anony (providing only data). Tuned hyperparameters below.

1 T-Table Template:

You are an expert in the financial sector and banking industry with expertise in analyzing customer credit data to make actual prediction about loan approvals. Based on the credit information of individuals, please predict the Approval_Flag. I will supply multiple instances with features and the corresponding label for your reference. Please refer to the table below for detailed descriptions of the features and label:

Initial Prompt

- - - - - feature description - - - - -
time_since_recent_enq: Duration since the customer made a recent credit enquiry
enq_L12m: Total number of enquiries in the last 12 months
...

Feature Description

- - - - - label description - - - - -
Approved_Flag: The flag which signifies if loan is approved for the customer or not.

Label Description

- - - - - data - - - - -
|3 | 6 | 4 |... | 1 |
|1 | 5 | 1|.....|<MASK>|
...

In-context examples

Please use the supplied data to predict the <MASK> Approved_Flag.
Answer:

Query

Hyperparameters

Number of features
(5, 10, 20, 30, 40)
In-context examples
(0, 8, 16, 32, 64)
Class 1 proportion
(0.1, 0.3, 0.5)

2 T-anony Template:

~~Feature Description~~ *~~Label Description~~*

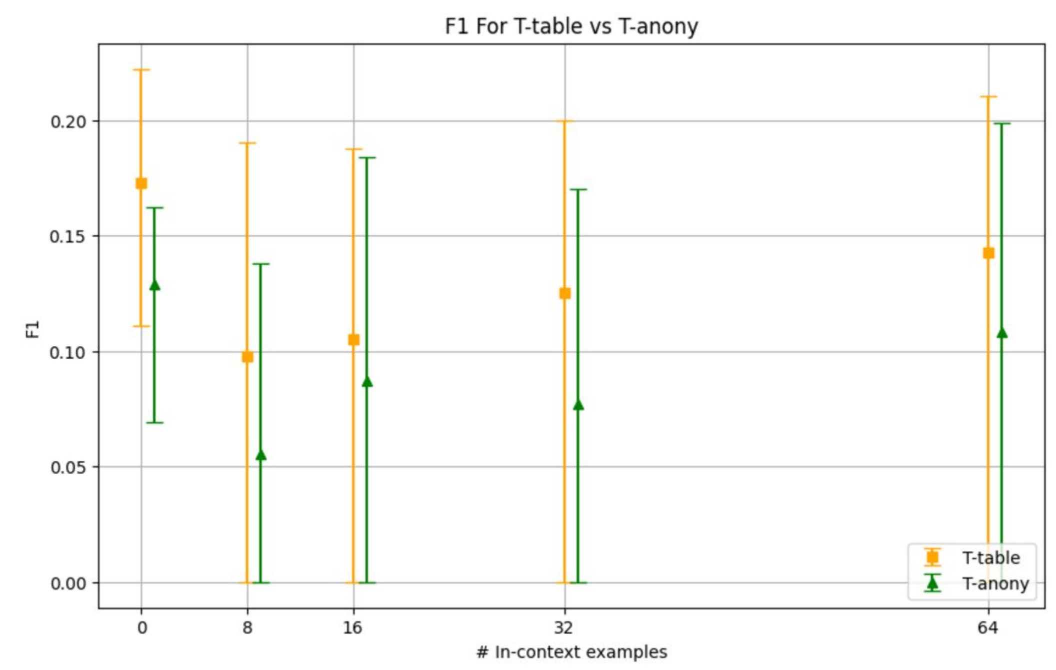
LLaMa2 GTL Models: 7B GTL (Original) | 7B 8 bit Quantized | 13B 8 bit Quantized

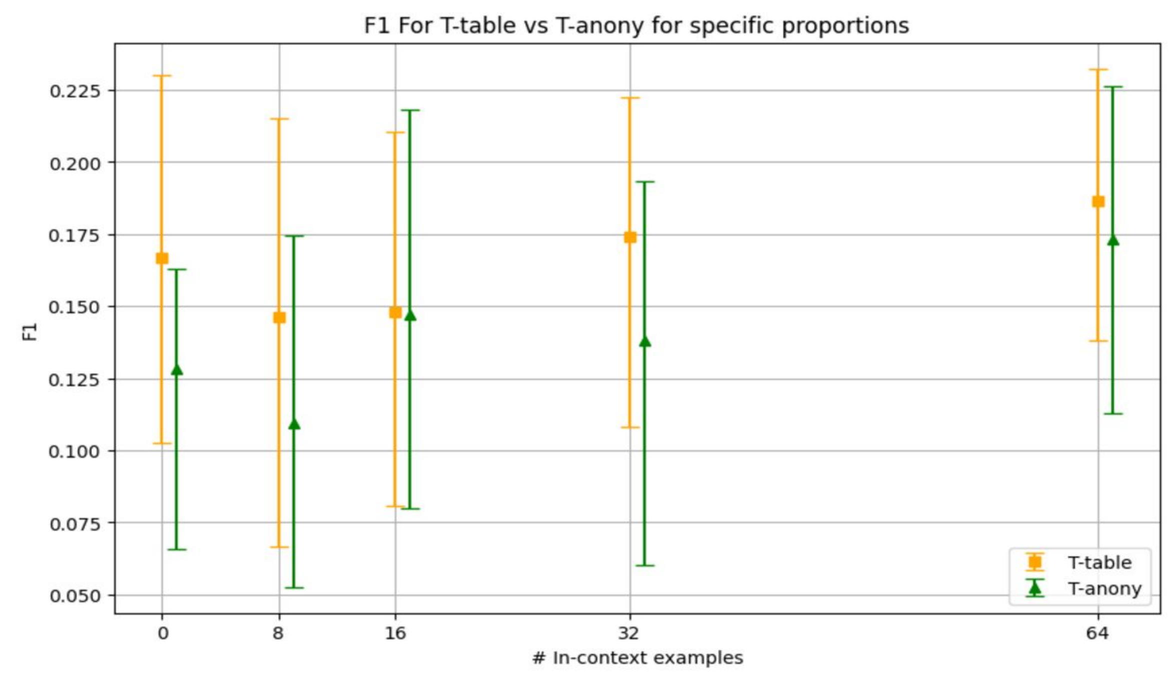
Analysis of Experiments & Key Observations

Main evaluation metric: F1 score due to class imbalance; and we are predicting classes (0 or 1) instead of probabilities (hence not AUROC). We use Kruskal-Wallis and Dunn's test to check for statistical significance, confirming F1 differences due to hyperparameters are significant.

Key Observations

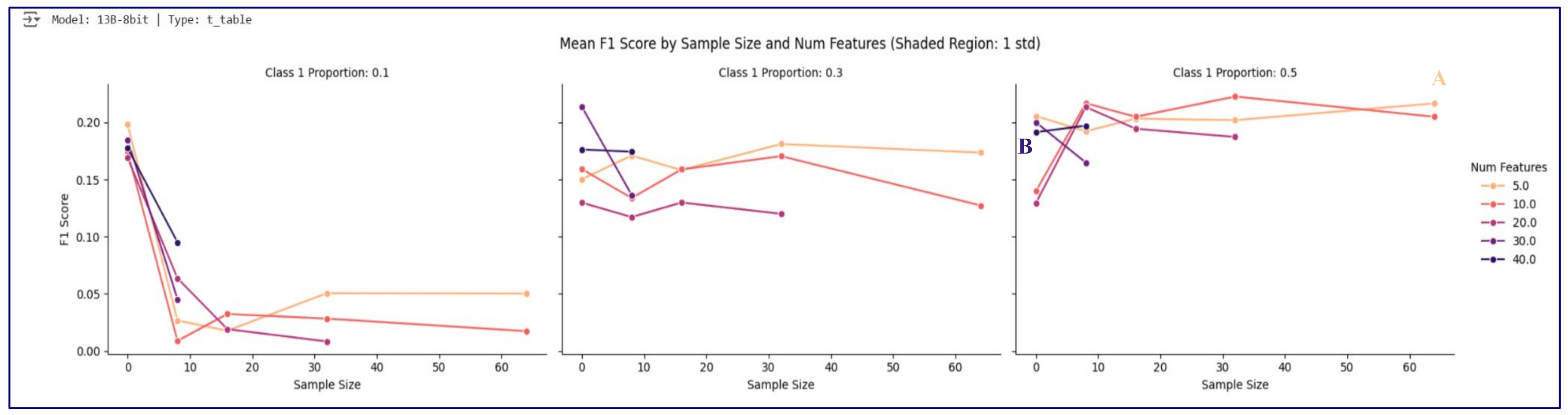
1. **T-Table beats T-anony**





2. **13B-GTL-8bit beats 7B-GTL-8bit (Table 1)**
   7B-8bit & 7B-Unquant show no significant difference in F1 score. Use 7B-8bit to save resources

3. **Higher the Class 1 proportion (more class-balanced in-context examples), better the F1 score**



4. **Given 4096 token limit & balanced in-context examples, ↑Sample Size ↓Num Features (A) beats ↓Sample Size ↑Num Features (B)**

5. **Best GTL model (13B-GTL-8bit) loses to Logistic Regression but beats other traditional ML models**

Table 1:

F1 Score Means:

Prompt Type	Model \ Sample Size	0.0	8.0	16.0	32.0	64.0	Row Mean
null	DecisionTree	NaN	0.043	0.079	0.119	0.134	0.094
	LogisticRegression	NaN	0.347	0.215	0.155	0.114	0.208
	RandomForest	NaN	0.036	0.053	0.076	0.093	0.065
	XGBoost	NaN	0.065	0.100	0.150	0.162	0.120
t_anony	13B-GTL-8bit	0.119	0.093	0.111	0.097	0.113	0.107
	7B-GTL-8bit	0.130	0.073	0.084	0.097	0.101	0.097
	7B-GTL-unquant	0.112	0.069	0.102	0.078	0.111	0.094
t_table	13B-GTL-8bit	0.173	0.131	0.124	0.130	0.132	0.138
	7B-GTL-8bit	0.146	0.084	0.099	0.114	0.143	0.117
	7B-GTL-unquant	0.202	0.096	0.095	0.108	0.124	0.125

Conclusion: With 0 retraining on a new task, limited data, & with/out feature descriptions, GTL beats some TradML models. Although Logreg seems best, GTL may beat Logreg with: ↑in-context samples & a larger model.

Limitations and Future Work

Limited compute prevents running ≥13B Llama2-GTL. 4096-token limit restricts experiments.
Future work: Train larger models with higher token limits. Predict other target variable(s).

Reference: Xumeng Wen, Han Zhang, Shun Zheng, Wei Xu, Jiang Bian. 2023. From Supervised to Generative: A Novel Paradigm for Tabular Deep Learning with Large Language Models. In KDD.