テーブルデータを分析対象とするデータ分析コンペティションにおける LLMs の性能調査

高野 海斗 [1]
[1] 大阪公立大学
takaito0423@gmail.com

概要

本研究では，テーブルデータを分析対象とするデータ分析コンペティションにおいて，多くのコンペティションで入賞者が使用している GBDT 系のモデルと比較して，LLMs がどのような条件化において有効であるかを評価する．現状の LLMs は連続値のデータを扱う点において GBDT 系のモデルに大幅に劣るため，カテゴリカル変数のみに特徴量を絞った上で，学習データの件数により，性能がどのように推移するかを手法ごとに確認した．分析の結果，学習データの件数が少ない場合においては，事前学習が行われている LLMs の方が良好な結果を示し，データ件数が一定数を超えると GBDT 系のモデルの方が良好な結果を示すことが確認できた．

1 はじめに

近年，大規模言語モデル (Large Language Models; LLMs) の性能は日々向上しており，LLMs はテキストデータを分析対象とするデータ分析コンペティション (以降，コンペティションと呼ぶ) の上位入賞に欠かすことのできない要素になっている [1, 2]．また，テーブルデータを分析対象とするコンペティションにおいては，多くのコンペティションで勾配ブースティング決定木 (Gradient Boosting Decision Trees;GBDT) 系のモデルが上位入賞の解法に使用されてきた．テーブルデータを分析対象とするコンペティションにおいて，未だに GBDT 系のモデルが有力な手法ではあるが，LLMs を用いた手法も使われ始めており，上位入賞の全員が LLMs を使用していたコンペティションも存在する．例えば，2025 年に開催された atmaCup#20[3] では，テーブルデータの情報をプロンプトに埋め込み，BERT や LLMs をファインチューニングすることが，上位入賞に不可欠でした．コンペティションだけでなく，推薦やエージェントシミュレーションの分野など，様々な分野においても LLMs の活用は進んでおり，今後もその活用領域はより拡大していくことが予想される [4]．

このような背景を踏まえ，本研究では，テーブルデータに対してどのような条件化において LLMs が有効であるかを評価する．可能であれば atmaCup#20 のデータを用いて実証分析を行いたいが，atmaCup#20 の参加には NDA 契約を結ぶ必要があり，コンペティション終了後にデータの削除が必須であったため[1)]，本研究では，atmaCup#8[5] のデータを元に検証を行う．具体的には，カテゴリカル変数のみに特徴量を絞ったデータを用いて，学習データの件数により，性能がどのように推移するかを手法ごとに確認する．

2 実証分析

2.1 データセット

本研究では，atmaCup#8 で提供されている「train.csv」を使用する．このデータは，アカウントを持っていれば誰でも手元にダウンロードして分析が可能である．タスクは，ゲームに基づく与えられた情報を使って，ゲームの売上 (世界全体での販売量) を予測するものとなっている．したがって，目的変数は世界全体での販売量を使用する．また，特徴量はいくつか与えられているが，本研究では以下の 5 つのカテゴリカル変数を使用する．

- Platform: 動作するプラットフォーム名 (DS など)
- Publisher: 発売元
- Developer: 開発会社

1) NDA 契約が必要なコンペティションであったが，その内容や上位解法に関しては，YouTube 上に公式から動画がアップロードがされている．https://www.youtube.com/watch?v=FE_dB0fUAuQ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

- Rating: ゲームのレーティング (E10+など)
- Genre: ゲームのジャンル (Action など)

現状の LLMs は連続値のデータを扱う点において GBDT 系のモデルに大幅に劣るため，カテゴリカル変数のみに特徴量を絞っている．一方で，ゲーム名のようなユニークな値が大半を占めるテキストデータのカラム (High Cardinality) は，GBDT 系のモデルが不利になるため使用しない．

2.2 使用する手法

本研究では，LLMs として「gemma-2-2b-it-bnb-4bit[2)]」を使用する．また，GBDT 系のモデルに XGBoost を使用する．

LLMs で使用するプロンプトを以下に示す．アンダーライン部分は埋め込みを意味する．

プロンプト

[Platform] {Platform}
[Publisher] {Publisher}
[Developer] {Developer}
[Rating] {Rating}
[Genre] {Genre}
How well is it selling worldwide?

また，事前学習が有効であることを確認するために，以下に示す匿名化を行ったプロンプトによる実験も実施する．

匿名化プロンプト

[column0] {DummyPlatformID}
[column1] {DummyPublisherID}
[column2] {DummyDeveloperID}
[column3] {DummyRatingID}
[column4] {DummyGenreID}
How well is it selling worldwide?

DummyID は，頻度の多いものから順に 1 から始まる自然数に変換した変数を使用する．なお，欠損の場合の DummyID は 0 とする．

各手法のハイパーパラメータは以下の通りである．

2) https://huggingface.co/unsloth/gemma-2-2b-it-bnb-4bit

LLMs のハイパーパラメータ

LLMs の学習時のハイパーパラメータは以下の通りである．

- epochs: 10
- lr_scheduler: "linear"
- warmup_ratio: 0.1
- learning_rate: 2e-4
- ：batch_size 32
- optimizer: "adamw_8bit"
- seed: 42

LoRA ファインチューニングのハイパーパラメータは以下の通りである．

- r: 256
- lora_alpha: 256
- lora_dropout: 0.05
- target_modules: ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]

他のハイパーパラメータはデフォルト値を使用．

XGBoost のハイパーパラメータ

XGBoost の学習に設定したハイパーパラメータは以下の通りである．

- num_boost_round: 50000
- early_stopping_round: 250
- objective: "reg:squarederror"
- learning_rate: 0.05
- seed: 42

他のハイパーパラメータはデフォルト値を使用．

2.3 評価指標

評価指標はコンペティションで使用されていた RMSLE(Root Mean Squared Logarithmic Error) を使用する．式は以下の通りである．

$$\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(y_i+1\right)-\log\left(\hat{y}_i+1\right)\right)^2} \quad (1)$$

なお，直接評価指標を最適化しやすくするために，目的変数を事前に対数変換して，モデルの学習を行う．

2.4 実験設定

「train.csv」のデータを学習データと評価データにランダムに分けて，各手法のスコアを確認する．な

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

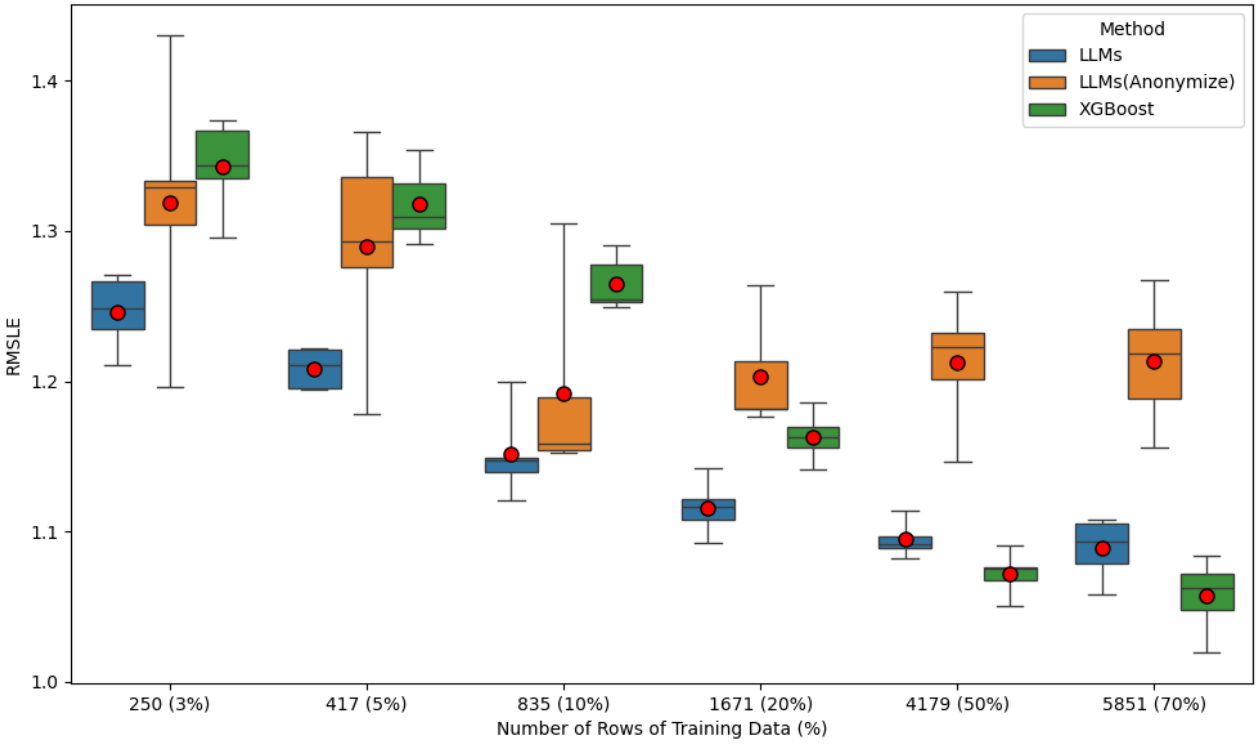

図 1　実験結果 (赤点は平均値)

お，学習データと評価データの分割は，学習データが 3%，5%，10%，20%，50%，70%の 6 つのパターンで分割を行う．また，推定誤差も考慮し，各パターンで乱数を変えて 5 回データ分割およびモデルの学習を行う．

2.5　結果

分析結果を図 1 に示す．

2.6　考察

図 1 の結果から atmaCup#8 のデータにおいて，LLMs は学習データの件数が 1671 件 (学習データの 20%) 以下の場合において，XGBoost よりも優れた結果を示している．このような結果が得られたのは，評価データにのみ存在するカテゴリカル変数に対して，事前学習がうまく機能していることに起因していると考えられる．もし，モデルサイズやニューラルネットワークなどの違いによるものであれば，匿名化したプロンプトを用いた LLMs は通常プロンプトを用いた LLMs と同等の結果が得られるはずだが，前者は後者に対して大きく性能が劣化していることが図 1 の結果から確認できる．

一方で，学習データの件数が増えてくると，アルゴリズムの優れている XGBoost の性能が改善していることが図 1 から確認できる．想定外だったのは，匿名化したプロンプトの LLMs の性能がデータ件数が増えても改善しなかったことである．この結果は，事前学習の効果およびプロンプトへの情報の組み入れ方が重要であることが示唆される結果である．

3　おわりに

本研究では，カテゴリカル変数のみに特徴量を絞った上で，学習データの件数により，性能がどのように推移するかを手法ごとに確認した．実証分析の結果，学習データの件数が少ない場合においては，事前学習が行われている LLMs の方が良好な結果を示し，データ件数が一定数を超えると GBDT 系のモデルの方が良好な結果を示すことが確認できた．

本研究では，一つのデータセットに対して一つの LLMs しか試すことができていないため，様々なデータセットに対して，いくつかの LLMs を用いて実証分析を行い，より詳細な傾向を確認したいと考えている．また，テキストデータのカラムがあれば，LLMs は GBDT 系のモデルよりも優れている一方で，連続値の変数をうまく扱うことができないため，今後はその部分の解決策を検討したい．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Wsdm cup - multilingual chatbot arena, 2026. https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings.

[2] Map - charting student math misunderstandings, 2026. https://www.kaggle.com/competitions/wsdm-cup-multilingual-chatbot-arena.

[3] atmacup20, 2026. https://www.guruguru.science/competitions/27.

[4] Kaito Takano, Masanori Hirano, and Kei Nakagawa. Modeling hawkish-dovish latent beliefs in multi-agent debate-based llms for monetary policy decision classification. In **The 26th International Conference on Principles and Practice of Multi-Agent Systems**, 2025.

[5] atmacup8, 2026. https://www.guruguru.science/competitions/13.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).