Analysis of Tabular Data Using LLM: A comparative Study

Supervisor:Dr. Florian Ertz

Supervisor: Abrar Ahmed

Presenter : Saurav Adhikari

Date: 23 Oct 2024

Universität Trier

Table of Contents

Introduction

- **Context**:
    - Data Driven World and complexity of modern datasets
    - Traditional Methods has served well.
    - Importance in health care
    - Why LLM?

Problem Statement

Limitations of Traditional Approaches:

- **Feature Extraction**: Traditional models struggle with automated feature extraction, especially in complex and unstructured datasets.
- **Pattern Recognition**: These models might overlook non-linear and hidden relationships.

Potential of LLMs:

- **Contextual Understanding**: LLMs can model interactions in the data more naturally and contextually.
- **Efficiency**: By leveraging pre-trained models, LLMs can require less manual tuning and yield more robust results.

Core Question: Will LLMs outperform traditional machine learning (ML) methods in analyzing tabular healthcare data, specifically in heart disease and breast cancer predictions?

Related Works

LLMs in Tabular Data Analysis:

- *TabLLM (Che et al., 2021)*: Pioneers the integration of LLMs in tabular data analysis by converting structured data into natural language representations.
- *UniPredict (Kumar et al., 2021)*: Applies LLMs to molecular property prediction, showing their versatility beyond language tasks.

Traditional Machine Learning:

- **Core Models**: Logistic Regression, Random Forests, Decision Trees, and SVMs (Bishop, 2006; Hastie et al., 2009). These are well-established but can be limited in handling feature complexity and large, noisy datasets.

LLMs in Healthcare:

- *Alsentzer et al. (2019)* and *Liu et al. (2020)*: Demonstrate the effectiveness of pre-trained LLMs like ClinicalBERT for tasks like medical record interpretation and hospital readmission predictions.

Research Questions

Comparative Performance:

- How do LLMs perform compared to traditional models (like Logistic Regression, SVMs, Random Forest) in key metrics such as accuracy, precision, recall, and F1-score?
- Will LLMs offer better scalability and computational efficiency for complex tabular datasets?

Enhancing Data Analysis with LLMs:

- How can LLMs enhance tabular data analysis through improved feature extraction, transfer learning, or augmenting traditional machine learning pipelines?
- Will LLMs lead to higher interpretability and generalization across different tabular datasets?

Dataset

Heart Disease Dataset (UCI Repository):

- **Features**: Age, gender, chest pain type, blood pressure, cholesterol, etc.
- **Objective**: Predict heart disease occurrence. Early diagnosis through predictive analytics could enable timely intervention and potentially save lives.
- **Significance**: Improving prediction accuracy can lead to proactive healthcare, reducing long-term costs and improving patient quality of life.

Breast Cancer Dataset (UCI Repository):

- **Features**: Tumor size, texture, margin, and other digitized image characteristics.
- **Objective**: Predict whether a tumor is benign or malignant. Early detection is crucial for treatment success and patient survival.
- **Significance**: Enhancing accuracy in classifying tumors could lead to faster and more accurate medical decisions.

Methodology

- **Model Selection**:
  - LLMs: BERT, Mistral
  - Traditional Models: Logistic Regression, Random Forest, SVM, Decision Trees.

- **Preprocessing**:
  - Handling missing values, encoding categorical data, and normalizing numerical features.
  - Resampling techniques (oversampling/undersampling) for imbalanced datasets.

- **Evaluation Metrics**: Accuracy, Precision, Recall, F1-Score, AUC-ROC.

Experimental setup

- **Data Preprocessing**:
    - One-hot encoding of categorical data (e.g., chest pain types).
    - Normalization of numerical features like age, blood pressure.
- **Cross-Validation**: Implement k-fold cross-validation to ensure model robustness by training and testing on different data splits. This ensures generalizability across subsets.
- **LLM Integration**:
    - PreTraining LLMs on specific tabular datasets.
    - Exploring In context Learning to see the effectiveness in case of a small sample size
    - Exploring how LLM embeddings can enhance traditional machine learning models through hybrid approaches.

Expected Results

Performance:

- Expect LLMs to achieve superior performance in terms of accuracy, precision, recall, and F1-score, especially for more complex datasets like breast cancer.
- LLMs should provide better scalability, especially as datasets grow in complexity and size.

Challenges:

- High computational costs for LLMs.
- Need for significant preprocessing to transform tabular data into formats compatible with LLMs (e.g., embeddings).
- Interpretability might be harder with LLMs compared to traditional, more transparent models like Decision Trees.

Challenges and limitations

Computational Complexity:

- Training and fine-tuning LLMs requires significant computational resources, which might limit their usability in real-time or low-resource environments.

Preprocessing Requirements:

- LLMs need careful data transformation, which may not always be straightforward, especially when dealing with tabular data formats.

Model Interpretability:

- Traditional models like Decision Trees offer higher interpretability, which is critical in healthcare decision-making. LLMs, while powerful, are often seen as "black boxes."

Solutions:

- Use optimized LLM versions (e.g., distilled models) for faster computation.
- Combine LLMs with traditional models to balance interpretability and predictive power.

Next Steps

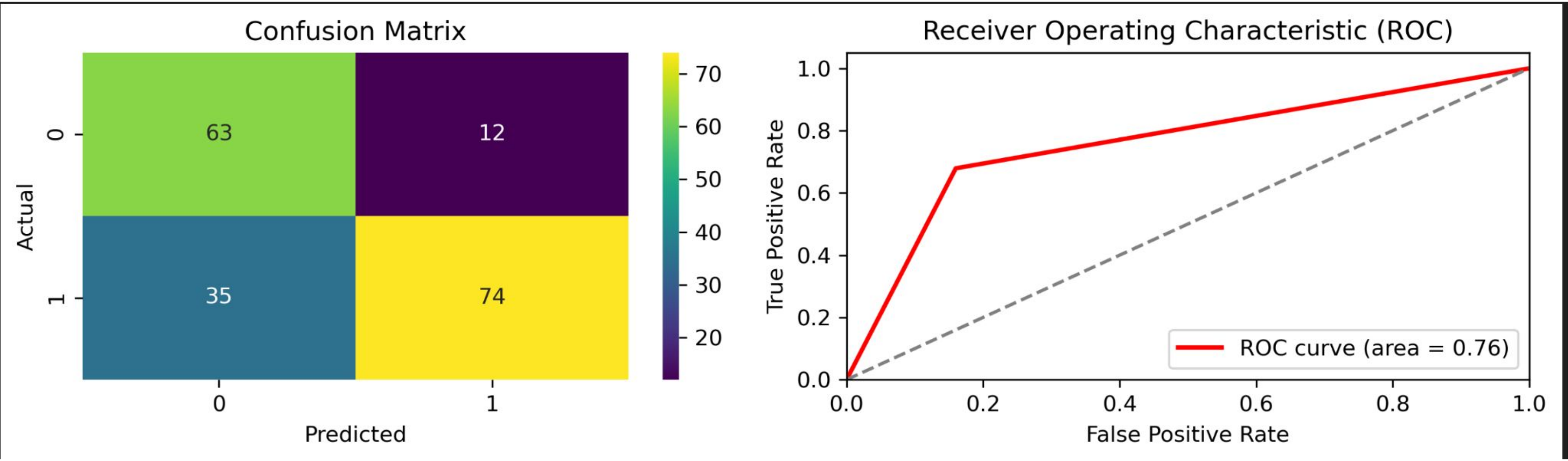
Complete Experimentation: Finish evaluating LLMs and traditional models on the heart disease and breast cancer datasets.

Analyze Results: Dive into the performance metrics and interpret the results to identify where LLMs excel or struggle compared to traditional models.

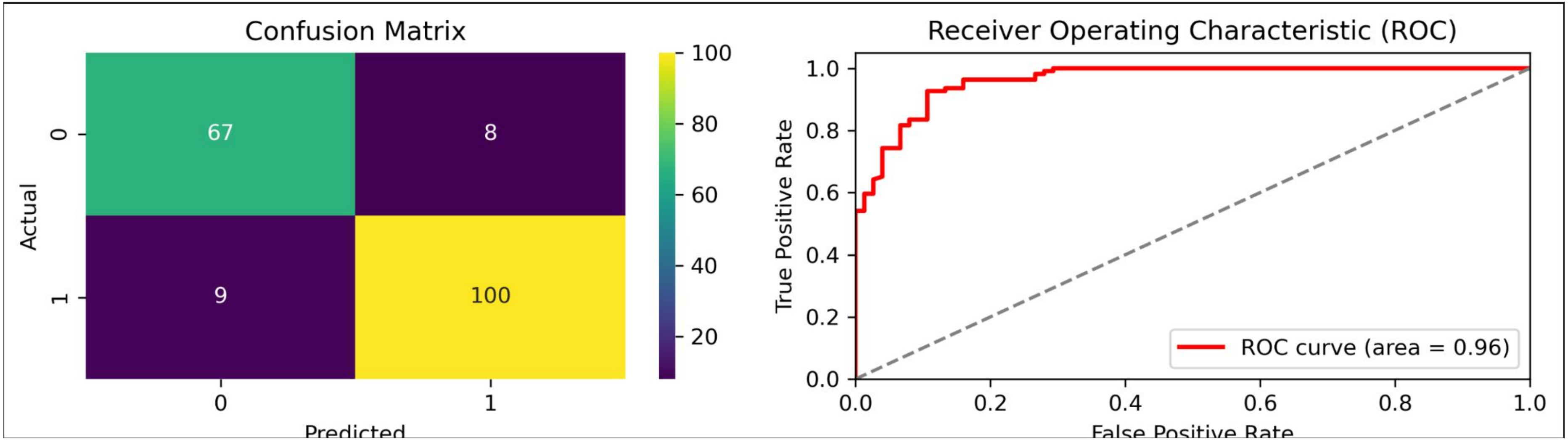
Simulation: Perform a simulation study to see the effectiveness of incontext learning with different sample sizes

Results

Predictions from bert



Predictions from traditional model



References

1. **Che, Z., et al. (2021)**. *TabLLM: Integrating Large Language Models with Tabular Data Analysis.* arXiv preprint arXiv:2110.00837.
2. **Kumar, A., et al. (2021)**. *UniPredict: Unified Molecular Property Prediction with Transformers.* arXiv preprint arXiv:2110.05078.
3. **Hastie, T., Tibshirani, R., & Friedman, J. (2009)**. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
4. **Bishop, C. M. (2006)**. *Pattern Recognition and Machine Learning.* Springer.
5. **Alsentzer, E., et al. (2019)**. *Publicly Available Clinical BERT Embeddings.* arXiv preprint arXiv:1904.03323.
6. **Liu, P. J., et al. (2020)**. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.* arXiv preprint arXiv:1904.05342.
7. **Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995)**. *Breast Cancer Wisconsin (Diagnostic).* UCI Machine Learning Repository. DOI: 10.24432/C5DW2B.
8. **Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988)**. *Heart Disease Dataset.* UCI Machine Learning Repository. DOI: 10.24432/C52P4X.