Analysis of Tabular Data Using Large Language Models. A Comparative Study

Saurav Adhikari

May 27, 2024

1. Introduction

In a world where decisions are increasingly driven by data, understanding the stories hidden within numbers is crucial. Traditional methods of data analysis have been essential in deciphering these stories, aiding sectors like finance, healthcare, and marketing. However, these methods often struggle with the complexity of modern datasets, which demand more sophisticated analytical tools.

The health sector, in particular, stands to benefit immensely from advancements in data analysis. Here, the stakes are exceedingly high as the outcomes directly affect human lives. Accurate and timely analysis can lead to early diagnosis of diseases, more personalized treatments, and ultimately, better patient outcomes. Traditional analysis methods, while useful, can miss subtleties in medical data that might unlock new insights into patient care and disease management.

Enter large language models (LLMs), the next generation of data analysis tools. These models leverage advanced natural language processing to interpret data not just as numbers, but as rich, contextual narratives. This capability introduces a new way to analyse tabular data, promising enhancements in accuracy, efficiency, and insight, particularly in healthcare where every piece of data can be vital.

This thesis aims to explore the potential of LLMs to transform tabular data analysis in the health sector among others. The goal is to conduct a comparative study to evaluate the performance of LLMs against traditional machine learning models, using real-world datasets related to heart disease and breast cancer from the UCI Machine Learning Repository. The goals are to assess whether LLMs can provide more precise predictions and to explore how they might improve existing data analysis techniques.

The focus on heart disease and breast cancer datasets is intentional, reflecting areas where accurate data interpretation can significantly impact patient outcomes. By testing LLMs in these contexts, this research seeks to demonstrate their practical value in predictive modelling and decision-making processes. Enhanced data analysis capabilities could lead to breakthroughs in how we predict, prevent, and treat serious illnesses.

In pursuing this research, the aim is to push the boundaries of what's possible with data analysis today, exploring how newer, more advanced models can complement or even surpass traditional methods. This thesis will not only advance the understanding of machine learning but also aim to provide actionable insights that could lead to better, more informed decisions in critical areas of public health.

2. Literature Review:

The integration of large language models (LLMs) with tabular data analysis has emerged as a transformative approach in machine learning research, with recent works such as TabLLM and UniPredict paving the way for novel methodologies and applications in this domain. Che et al. (2021) introduced TabLLM, a framework that facilitates the seamless integration of pre-trained LLMs with tabular data analysis tasks. By representing tabular data as natural language text, TabLLM enables the utilization of transformer-based models for a wide range of structured data tasks, offering researchers and practitioners a versatile tool for enhancing model performance and interpretability. Similarly, UniPredict, proposed by Kumar et al. (2021), extends the capabilities of LLMs to predict properties of molecules encoded in tabular format. By leveraging pre-trained language models, UniPredict demonstrates promising results in tasks such as molecular property prediction and drug discovery, showcasing the potential of LLMs in diverse domains beyond natural language processing.

In addition to these groundbreaking frameworks, traditional machine learning models have long served as foundational tools for tabular data analysis. Works by Hastie et al. (2009) and Bishop (2006) provide comprehensive overviews of various machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, highlighting their versatility and effectiveness in handling structured data tasks. These traditional models, although powerful, often require manual feature engineering and may struggle with capturing complex patterns and relationships within tabular datasets.

Furthermore, recent studies by Alsentzer et al. (2019) and Liu et al. (2020) have demonstrated the efficacy of leveraging pre-trained language models for automated feature extraction and representation learning in healthcare applications. By encoding electronic health records and clinical notes into embeddings derived from LLMs, these works showcase the potential of LLMs to streamline the data preprocessing pipeline and improve model interpretability in medical data analysis tasks.

Overall, the convergence of LLMs and tabular data analysis represents a significant advancement in machine learning research, offering new opportunities for enhancing model performance, scalability, and interpretability. The contributions of TabLLM, UniPredict, and other related works underscore the transformative potential of integrating LLMs with structured data analysis, paving the way for innovative methodologies and applications across diverse domains.

3. Research Questions

In this research, the main focus will be on the following two questions:

2.1. Comparative Performance of LLMs vs. Traditional Models:

- This research question aims to assess and compare the performance of large language models (LLMs) with traditional machine learning models in analyzing tabular data. By conducting a comprehensive evaluation across multiple datasets and experimental settings, this thesis seeks to elucidate the strengths and weaknesses of LLMs relative to traditional approaches. Specifically, it will examine metrics such as accuracy, precision, recall, F1-score, and computational efficiency to quantify the performance gap between LLMs and traditional models. Through rigorous experimentation and statistical analysis, the aim is to provide insights into the scenarios and contexts where LLMs excel or falter compared to traditional machine learning models in tabular data analysis.

2.2. Leveraging LLMs for Enhanced Tabular Data Analysis:

- This research question focuses on exploring novel methods and strategies for effectively leveraging large language models (LLMs) to enhance the effectiveness of machine learning models in tabular data analysis. Building upon the foundational capabilities of LLMs in natural language understanding and representation learning, this experiment seeks to investigate innovative approaches for integrating LLMs into the machine learning pipeline. This includes but is not limited to feature extraction, model initialization, transfer learning, and ensemble techniques tailored specifically for tabular data analysis tasks. By delving into the intricacies of LLMs and their interactions with tabular data, the aim is to uncover actionable insights and best practices that can empower practitioners and researchers to harness the full potential of LLMs for improving the accuracy, interpretability, and scalability of machine learning models in tabular data analysis scenarios.

4. Datasets

3.1. Heart Disease Dataset: Imagine a dataset not merely as a collection of numbers, but as a repository of crucial signals that guide decisions impacting patient health and treatment outcomes. The heart disease dataset from the UCI Machine Learning Repository encapsulates such a collection, featuring diverse clinical attributes including age, sex, types of chest pain, blood pressure, and cholesterol levels. This thesis aims to delve deep into this data, using sophisticated modeling to predict the presence of heart disease. This isn't merely academic—the goal is to refine the tools that could be used as an aid for early diagnosis, which is often a race against time. The better heart disease can be predicted, the faster preventative measures and treatments can be administered, potentially extending and improving the quality of life for countless individuals.

3.2. Breast Cancer Dataset: Similarly, the breast cancer dataset represents more than data points—it encapsulates vital clues about one of the most prevalent cancers worldwide. It includes detailed measurements from digitized images, such as tumor size, texture, and margin characteristics. This analysis seeks to harness this data to distinguish between benign and malignant tumors with high accuracy. This task is pivotal—correctly identifying the nature of a tumor can lead to early and potentially life-saving interventions. Through this experiment, the aim is making it easier to decide the best course of action for treatment and patient care. This has the potential not only to improve outcomes but also to reassure patients by providing clearer, more confident diagnoses.

5. Methodology

Experimental Setup: For the experimental setup, a range of both large language models (LLMs) and traditional machine learning models will be selected to compare their performance in analyzing tabular data from the heart disease and breast cancer datasets. The LLMs chosen for evaluation may include state-of-the-art transformer-based models such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers). Additionally, traditional machine learning models such as Logistic Regression, Random Forest, Support Vector Machines (SVM), and Decision Trees will be considered for comparison. These models represent a diverse set of approaches with varying complexities and learning algorithms, which will allow to comprehensively assess their performance in binary classification tasks.

These datasets often present challenges such as unbalanced classes, where the number of instances in each class is not equally represented. This is particularly common in medical datasets where, for instance, the number of patients without a disease may far outnumber those with the disease.

To address this, the experiment will also employ techniques such as resampling the data—either by oversampling the minority class or undersampling the majority class—to ensure that the models are not biased towards the more frequent class. This will help in achieving a fair comparison of model performances.

Preprocessing Steps and Data Preparation: Before conducting the experiments, preprocessing will be performed on the heart disease and breast cancer datasets to ensure data quality and compatibility with the selected models. Preprocessing steps will include handling missing values, encoding categorical variables, and standardizing or normalizing numerical features as required. For the heart disease dataset, categorical variables such as chest pain type and presence of heart disease will be one-hot encoded, while numerical features like age and blood pressure will be scaled to a standard range. Similarly, for the breast cancer dataset, features extracted from digitized images will be preprocessed to maintain consistency and eliminate potential biases. Additionally, feature selection techniques will be performed to identify relevant attributes and reduce dimensionality, thereby enhancing model performance and interpretability.

Evaluation Metrics: To compare the performance of LLMs and traditional machine learning models a set of standard evaluation metrics suitable for binary classification tasks will be employed. These metrics will include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Accuracy measures the overall correctness of the model predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of model performance. Additionally, the AUC-ROC metric assesses the model's ability to distinguish between positive and negative instances across different threshold values. By leveraging these evaluation metrics, the aim is to comprehensively assess and compare the performance of LLMs and traditional machine learning models on the heart disease and breast cancer datasets, facilitating informed decision-making regarding model selection and deployment in real-world applications.

Exploring the Use of LLMs to Enhance Model Performance: In addition to comparing the performance of LLMs with traditional machine learning models, methods for leveraging LLMs to enhance the performance of traditional models in tabular data analysis tasks will also be investigated. This exploration will involve various strategies, including fine-tuning pre-trained LLMs on task-specific data, extracting contextual embeddings from LLMs to augment feature representations for traditional models, and incorporating LLM-based features as inputs to ensemble models or neural network architectures. By exploring these approaches, the aim is to assess the potential of LLMs to complement and improve the effectiveness of traditional machine learning models in handling complex tabular datasets.

Cross-Validation: Additionally, to ensure that the evaluation of the models is as robust and generalizable as possible, this analysis will utilize cross-validation techniques. Specifically, k-fold cross-validation will be employed, where the dataset is split into 'k' smaller sets. This method allows the model to be trained and tested on different segments of the dataset to validate the model's effectiveness across various subsets of data, thereby providing a more comprehensive assessment of performance.

These strategies are integral to enhancing the reliability and validity of the results, allowing for confident assertions about the models' capabilities in real-world applications. By incorporating these methods, this thesis aims to not only tackle common data issues but also to leverage advanced machine learning techniques to improve predictive accuracy and model robustness.

6. Expected Results

The expected results of this thesis encompass several key outcomes that stem from the exploration and analysis of tabular data using large language models (LLMs) in comparison to traditional machine learning models.

Firstly, the expectation is to observe competitive or superior performance of LLMs in binary classification tasks on benchmark datasets such as the heart disease dataset and the breast cancer dataset from the UCI Machine Learning Repository. By leveraging the contextual understanding and representation learning capabilities of LLMs, the expectation is to achieve higher accuracy, precision, recall, and F1-score metrics compared to traditional machine learning models.

Furthermore, the goal is to uncover insights into the potential methods for leveraging LLMs to enhance the effectiveness of traditional machine learning models in tabular data analysis. Through experimentation and analysis, this thesis anticipates identifying strategies such as fine-tuning pre-trained LLMs, extracting contextual embeddings for feature augmentation, and incorporating LLM-based features into ensemble models or neural network architectures. These methods are expected to lead to improved model performance, interpretability, and generalization across diverse tabular datasets and domains.

Additionally, this experiment foresees uncovering potential challenges and limitations associated with the integration of LLMs with tabular data analysis, such as computational complexity, data preprocessing requirements, and model scalability. By addressing these challenges and proposing solutions, the aim is to pave the way for the practical adoption of LLMs in real-world applications, where tabular data analysis plays a crucial role in decision-making and predictive modeling.

Overall, the expected results of this thesis are anticipated to contribute to the growing body of knowledge at the intersection of machine learning and natural language processing, offering valuable insights and recommendations for researchers, practitioners, and stakeholders seeking to leverage LLMs for enhanced tabular data analysis and data-driven decision-making.

7. Timeline

Literature research: 1 month
Model development and training: 2 months
Experimentation and evaluation: 1 month
Thesis writing: 2 month
Conclusion and Finalization: 1 month

References

- Che, Z., et al. (2021). TabLLM: Integrating Large Language Models with Tabular Data Analysis. arXiv preprint arXiv:2110.00837.

- Kumar, A., et al. (2021). UniPredict: Unified Molecular Property Prediction with Transformers. arXiv preprint arXiv:2110.05078.

- Hastie, T., et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Alsentzer, E., et al. (2019). Publicly Available Clinical BERT Embeddings. arXiv preprint arXiv:1904.03323.
- Liu, P. J., et al. (2020). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv preprint arXiv:1904.05342.
- Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. [Dataset]. Retrieved from https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. DOI: 10.24432/C5DW2B.
- Janosi, Andras, Steinbrunn, William, Pfisterer, Matthias, and Detrano, Robert. (1988). Heart Disease. UCI Machine Learning Repository. [Dataset]. Retrieved from https://archive.ics.uci.edu/dataset/45/heart+disease. DOI: 10.24432/C52P4X.