POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Large Language Models in Data Preparation: Opportunities and Challenges,

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica,

Author: Anna Barberio

Advisor: Prof. Cinzia Cappiello

Co-advisor: Dott. Ing Camilla Sancricca

Academic year: 2022-2023

---

1. Introduction

In a world in which data-driven decision-making has become very widespread, many tools have been designed to assist users in performing data analysis tasks to achieve the ultimate goal of achieving good analysis results and then apply them to the decision-making process. These tools assist users in discovering solutions to improve the quality of datasets by executing the data preparation pipeline. Indeed, the complexity of designing data preparation processes stems from the challenge of managing diverse data sources, formats, and structures. Expertise is essential for effectively navigating and transforming the data to ensure the quality and reliability of the prepared datasets. This applications present users with an improved dataset but often lack explanations, leaving users unaware and uninformed about the changes made and, consequently, potentially fostering greater trust in the system.

To address this gap, several studies, particularly those conducted by the Georgia Institute of Technology, underscore how incorporating explanations can furnish users with guidelines, thereby reducing the opacity of automated machine learning processes. For this reasons, our thesis work has the primary objective of presenting users with valid explanations to enhance their understanding. The second pivotal aspect of the work, linked to the initial point, revolves around the format of explanations. According to additional studies conducted by Miller comparing diverse forms of explanation, natural language explanations have proven to be the most effective. Consequently, within the methodolgy we propose, the format of explanations will be entrusted to a Natural Language Processing (NLP) tool such as ChatGPT, capable of presenting explanations in a form most comprehensible to users. In addition to studying and analyzing a methodology aimed at integrating explanations into a common data preparation tool, my contribution also involved an examination of ChatGPT and its potential utility in the context of providing explanations for a data preparation pipeline.

2. Background Concepts

Before delving into the actual methodology, it is essential to outline the fundamental background knowledge elements one must possess to best comprehend the thesis work.

Within the tool, two primary guidelines will be

proposed for the user to follow based on the type of objective they aim to achieve. It is crucial to elucidate the distinction between 'description' and 'explanation.'
A 'description' is a comprehensive representation providing clear and factual information about an object, process, or concept. It objectively presents facts without delving into the reasoning or rationale behind the subject matter. Descriptions play a crucial role in offering an essential overview or context on a specific topic. On the other hand, 'explanation,' as defined by the European Commission's High-Level Expert Group on AI (AI HLEG), involves the AI system's ability to clarify its decision-making process. Explanations are typically more detailed, attempting to illuminate connections within a concept or process.[2]
Both descriptions and explanations within the tool are intricately tied to the text generation facilitated by ChatGPT.
ChatGPT, an advanced Natural Language Processing (NLP) system developed by OpenAI, is designed to generate conversations resembling human interactions. It comprehends contextual nuances and generates suitable responses, leveraging the underlying GPT-3 model trained on diverse conversational contexts. The tool excels in textual transformation, enhancing content user-friendliness. This is why it has been chosen to generate the textual part of descriptions and explanations of the thesis tool.[1]
The understanding of LIME will also prove valuable. LIME, which stands for "Local Interpretable Model-Agnostic Explanations," serves as a potent instrument crafted to illuminate the intricate mechanisms within complex machine learning models. In our present context, LIME plays a pivotal role in deciphering the mystery of which factors have exerted the most significant, whether positive or negative, impact on recommending the optimal data preparation technique.

3. Methodology

After introducing the foundational knowledge concepts necessary for a better understanding of the work, we can now proceed to the presentation of the actual methodology. Given a common data preparation tool that assisted users in data exploration and in creating a data preparation pipeline for a loaded dataset, the main objective of the thesis was to introduce explanations that would serve as guidelines throughout all phases of analysis.

3.1. NLP tool

Before delving into the detailed analysis of the methodology, it is essential to thoroughly describe the innovative elements that my thesis work has brought to the common data preparation tool. Firstly, we need to define the manner in which explanations are presented to the user. Through the use of large language models, we have integrated a new component that leverages ChatGPT to articulate information within the tool in a language understandable even to non-expert users. Within the tool's code, OpenAI's GPT-3.5 APIs are invoked, taking varied information inputs based on the analysis phase and generating text that transforms the input into comprehensible language.
Let's categorize the different types of support the user can receive.

- Description: Involves text elaboration to better describe the data, especially focusing on graph descriptions and various features of the uploaded dataset in the initial analysis phase. The input passed, in this case, to obtain the text description as output, is the result of the code that, depending on the section of the tool, calculates the fundamental features extracted during the data profiling and data quality assessment phase.
- Static explanations: not dependent on the use of the large language model, explaining the ranking of quality dimensions.
- Dynamic explanation: Inextricably linked to the utilization of LIME is the objective of providing a clearer understanding of which features in the uploaded dataset exerted the most significant influence on recommending a specific method. This integral aspect played a pivotal role in not only simplifying but also rendering transparent the responses generated by the NLP tool in the project. Consequently, it enabled users to comprehensively grasp the rationale behind dataset characteristics and the proposed choices. The input given in this case is the explanation extracted from LIME

through Python code. This code takes a trained classifier and a specific example from the dataset as input. It generates perturbed versions of the example, evaluates them using the classifier, and records predictions. LIME then uses this local dataset to train an interpretable model, approximating the classifier's behavior near the chosen instance. The interpretable model identifies and explains which features of the dataset significantly influenced the classifier's choice of the best method.

3.2. Architecture

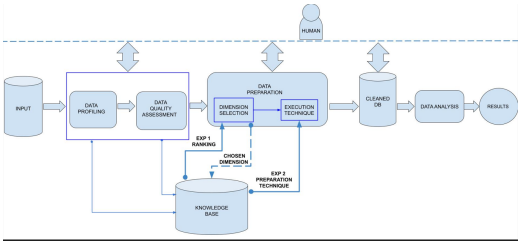

Figure 1: Data Analysis Pipeline

Let's now proceed with the step-by-step description of the methodology of the initial tool, which our thesis work has integrated at various points in the pipeline with innovations such as explanations. These additions have provided clear guidelines for improved reliability and user understanding. As illustrated in Figure 1, the process begins with the selection of a dataset. Once the dataset is loaded, the initial phases of data profiling and data quality assessment start. In these phases, information from the dataset loading stage and the data profiling techniques are utilized to present the user data characteristics in both tabular and graphical formats.
During this stage of the pipeline, descriptions, the first output of our NLP tool, play a crucial role in helping users gain a more thorough understanding of graphs, relationships between dataset features, and the overall presentation of tables. Once this step is completed, users become aware of both the positive and negative characteristics of the dataset that require improvement in the subsequent phase.
The next phase is data preparation, where the last two types of support mentioned earlier come into play. It's important to note that in this phase, the knowledge base becomes increasingly vital as it aids in formulating a strategy for dimension ranking and the selection of the most appropriate data preparation techniques to enhance dataset quality.
The initial step in this phase involves generating a ranking of quality dimensions ordered by urgency for improvement. This ranking is the outcome of the data quality assessment and the information in the knowledge base regarding the impact of applying machine learning.
In this context, the first explanation is displayed to the user to help them understand how the combined dimensions ranking was generated. As mentioned earlier, this form of support embraces a static approach. This text serves to instill confidence in the final ranking, encouraging the user to place trust in the prioritized order for enhancing these dimensions. Consequently, users can freely choose to follow the recommended ranking or create their own pipeline.
In this phase, the user is presented with the last form of support: the second explanation. This explanation utilizes LIME to generate explanations. To better define the second explanation, it's useful to know that the proposed best method is the result of a trained classifier.The classifier was trained using the K-Nearest Neighbors (KNN) algorithm, which classifies a data point based on the majority class among its nearest neighbors in the training set. The output of this classifier determines the optimal imputation method for the loaded dataset. The final step, in fact, takes information derived from the application of LIME to the classifier and other ML-based techniques, providing textual explanations through the large language model.

4. Implementation

After providing a detailed overview of the thesis methodology, the subsequent focus will shift to the actual implementation. Before delving into this description, it is advisable to briefly list the technologies used for development, namely Python Flask, HTTP Methods, Ajax POST requests, OpenAI, and LIME. Now, let's move on to the presentation of the tool's implementation.
The workflow begins with the user uploading the data source, which can be selected from the computer's folders or dragged and dropped into the designated area. In this new view, a tabular representation of the dataset is displayed along-

side sections related to various choices the user needs to make. The user is prompted to select the features they are most interested in, choose the machine learning algorithm they intend to use in the analysis phase, and indicate whether they would like assistance in designing the data preparation pipeline. The first novelty contribution introduced by thesis work is incorporated in this page, in particular to the initial description of the tool. By clicking the 'More Detailed about the dataset' button, the response generated by the NLP tool describing the dataset in all its columns is shown. When the user clearly understands the data they will be working with after reading the initial description, they will be able to comprehend what will be presented on the subsequent page. The elements described in this section are divided into different graphs and tables, each describing a different aspect of the dataset (overview and alerts, missing values, correlation, and variables). In Figure 2, an example of a generated description to better explain missing values is shown.

Figure 2: More Details

In the 'variables' section, on the other hand, the description of statistics and distributions is included. They are presented in the form of tables and interactive graphs, accompanied by descriptions generated by the NLP tool, aiding in a better understanding of the results of interactions with the graphs. After the user has all the descriptions, they can proceed to the data preparation phase. In this subsequent step, the user is guided on enhancing three quality dimensions: Accuracy, Completeness, and Uniqueness. The user is presented with a suggested improvement ranking. Once again, the reasoning behind this seemingly mysterious classification is unveiled by clicking a button that reveals a textual explanation; this belongs to the second type of support. Figure 3 provides an illustration of what will be explained to the user.

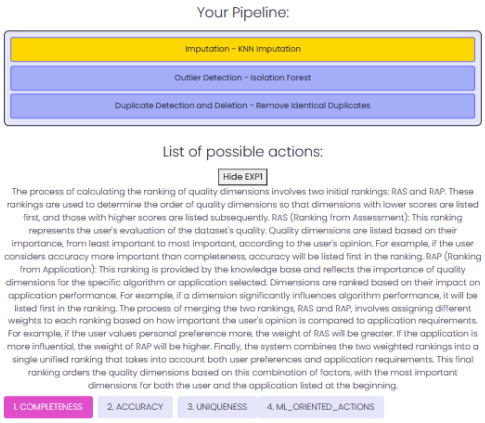

Figure 3: EXP1

In this concluding section, following the user's earlier guidance in prioritizing one of the three dimensions, the attention turns to how to enhance them. Utilizing the already mentioned trained classifier, the user receives guidance on the optimal technique to improve the specified quality dimension. The user has the option to accept or reject this advice. To reinforce the recommendation, a second explanation, triggered by clicking the 'exp2' button and generated by ChatGPT, is provided. An exemplified summary of 'exp2' is illustrated in Figure 4. .

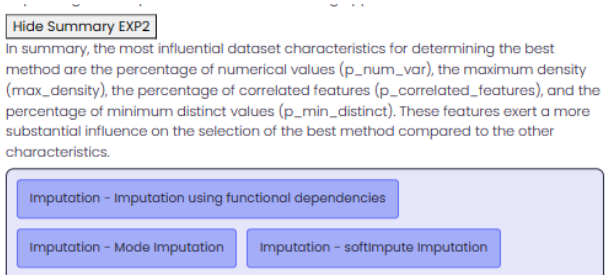

Figure 4: EXP2

5. ChatGPT

In this paragraph, there is a brief description of the additional contribution made to my thesis. We conducted research aimed at understanding the actual qualities in the field of data preparation pipeline of ChatGPT. This study explores the use of ChatGPT, a generative language model, in various phases to assess its effectiveness in aiding data analysis. The initial experimentation involved posing questions to an online tool associated with ChatGPT-3, revealing

challenges such as format conversion and limited response completeness. The comparison with a locally integrated tool and subsequent adjustments addressed some issues, enhancing accuracy and contextualization in responses. While ChatGPT proves valuable for generating textual explanations, there are limitations, including generic suggestions and challenges in visualization and comprehensive dataset analysis. Despite some drawbacks, ChatGPT can be considered useful for enhancing textual aspects of results obtained through familiar formulas and functions. However, the tool has limitations, including a lack of access to external information sources or internet browsing, impacting its ability to provide accurate or up-to-date information on various topics. It may also face challenges with intricate or unconventional queries and occasionally produce bias. These limitations, alongside strengths, were identified in our research, offering insights into OpenAI's role in user explanations.

6. Conclusions

The proposed solution is a user-friendly platform designed to simplify data analysis processing for individuals with varying expertise levels, particularly those seeking to optimize their data for machine learning analyses. The innovation lies in integrating explanations generated by a large language model into the data preparation pipeline. Textual explanations empower users to make informed decisions during data exploration, highlighting dataset features and weaknesses. The tool, complemented by generative AI like ChatGPT, simplifies complex results into understandable text. The future evolution of the project can involve implementing a dual workflow system tailored for both expert and non-expert users. This enhancement aims to provide personalized experiences, accommodating diverse user needs and expertise levels. The strategic refinement reflects a proactive approach towards adaptability and user-centred in subsequent developments.

References

[1] Jianyang Deng and Yijia Lin. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83, 2022.

[2] Tim Miller. Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 333–342. ACM, 2023.