Prompt written by: Cody York

LLM: OpenAI ChatGPT(May 2023)

Version: GPT-4

Date: 05/18/2023

LLM Prompt: Write a one-page paper on how companies use data science and machine learning in production to perform Data Engineering. Data Engineering is referring to managing large changing datasets. Include citations for 3 sources used in the paper.

Rewritten Paper

Data Engineering in the Age of Data Science and Machine Learning:

An Insight into Modern Companies

Data Engineering has emerged as a critical domain in the modern technological landscape. The advent of big data, data science, and machine learning has propelled a paradigm shift in the handling and management of complex and dynamic datasets in production (Kitchin, Rob. (2014)). To harness this massive amount of data we need optimized machine learning models to get better insights into what these large datasets can tell us.

Companies are increasingly using Data Science and Machine Learning (ML) to optimize data engineering processes and to derive actionable insights from the voluminous data that they generate daily. Data Science can include tasks such as: extracting data, processing data, and visualizing data. These tasks can help companies recognize patterns and trends in the data more efficiently. On the other hand, ML automates the predictive analysis of data, freeing up data engineers to focus on more complex tasks, such as architecting scalable data pipelines that move data to its destination and perform necessary cleaning and data preprocessing. Machine learning is attracting a great amount of interest because it offers new tools to reveal generalizable patterns successfully from very complex and unregulated big data. Machine learning achieves such a breakthrough since it turns the deductive problem of finding a rule to an inductive one by letting the data inform us the best rule characterizing data (Wang, H., Yao, Y., & Salhi, S. (2020)).

To manage large, changing datasets in production, companies are integrating ML algorithms with traditional data engineering techniques. The outcome is a dynamic data processing pipeline that can adapt to the evolving data landscape. For example, in the article “Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database” they discuss using ML algorithms to automatically assign Lab codes to the data. This was successful and was summarized as “this scalable, automated algorithm may improve data quality and interoperability, while substantially reducing the manual effort currently required to accurately map laboratory data (Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME (2018)).”

Companies are also utilizing ML for data imputation, the process of filling missing values in a dataset. This is a common step in the pre-processing of data in a data pipeline. Traditional methods, which relied on statistical techniques, were often insufficient when faced with complex datasets with high dimensionality or many different variables. ML algorithms, in contrast, can handle multiple variables simultaneously and more accurately predict missing values, thereby improving the overall data quality.

In the realm of data storage and retrieval, ML is enabling companies to structure and classify unstructured data, which comprises a significant portion of Big Data likely found at a company. Techniques such as Natural Language Processing (NLP) and image recognition algorithms are used to extract meaningful features from unstructured data like images. This data can then be organized and stored in a structured format, such as a database, making it accessible for future analysis.

In conclusion, Data Science and Machine Learning are transforming the face of Data Engineering in production environments. They allow companies to manage large, changing datasets more efficiently, accurately, and are able to scale the data properly, leading to improved business intelligence and decision-making insights from the data.

References:

1. Kitchin, Rob. (2014). Big Data, New Epistemologies and Paradigm Shift. Big Data & Society. 1. 1-12. 10.1177/2053951714528481.

2. Wang, H., Yao, Y., & Salhi, S. (2020). Tension in big data using machine learning: Analysis and applications. Technological Forecasting and Social Change, 158, 1.1.

3. Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. J Am Med Inform Assoc. 2018 Oct 1;25(10):1292-1300. doi: 10.1093/jamia/ocy110. PMID: 30137378; PMCID: PMC7646911.