Report date 25/11/2024

Paper report:

**Overview:**

The paper has been considered since it's been cited in “Data Science Through the Looking Glass“ and consists in a more interview based analysis of the ML/Data Science world.

The content of the paper ended up not being very relevant to our goal, but there are some interesting points to it like the importance given to data management and the overall highlighting of the importance of data.

The paper describes a study done at Microsoft that analyzed what ML workers consider important also with regard to their experience.

The paper also introduces a reference workflow for ML applications consisting of 9 steps, starting from the Model Requirements ending up to the Model Monitoring.

The whole first part of the paper regards ML workflows and shows how ML development is different from traditional Software Engineering development due to the need of more experimentation and more complex and unpredictable feedback loops at essentially every step of the workflow.

The study itself is based on a questionnaire developed starting from personal interviews with 14 software engineers coming from different sectors, with most of them being in leadership roles and all working with ML in their teams.

The questionnaire has then been answered by 551 SWE that have been split in 3 buckets based on their experience (expressed in number of years working with AI).

The most interesting outcome for us is that among all the challenges the interviewees had to define a priority, the most important one, throughout all the 3 groups was “Data Availability, Collection, Cleaning and Management”, highlighting the importance of data management, data preparation and data sourcing in the AI field. Another positive feedback is that the most experienced group considered this topic to be most important 60% of the times more than the other 2 groups. Meaning that this is a challenge that is even more important for experienced practitioners.

In particular they mention the need to use data and model versioning techniques to keep track of that used data and cope with further data evolution. Overall, even if never explicitly mentioned, they highlight the importance of ML-Ops.

Some correlations between the responses and the group that gave them have been highlighted (like: people with less AI experience consider Education and Training a more important challenge).

The paper then discuss a way to model the experience level through another survey that provided them with an Activity Maturity Index which gave a similar representation of the experience when compared with the simple year of experience.

The paper end by highlighting the main discoveries:

* Data discovery and management: ML is all about the data that powers learning models, highlighting the importance of data and everything concerning it (management, preparation, cleaning, labeling)
* Customization and reuse: it's quite hard to reuse components in ML scenarios due to how entangled different components of the same pipelines are. The fact that models are build and developed on the data for that specific data that's being used makes it even harder to use the same model/pipeline for different projects that are based on different data. This is an important difference from the more classic software development environment.
* ML Modularity: linked with the previous challenge.

**Limitations:**

* The test has been conducted within Microsoft so the findings may be specific to Microsoft teams and workflows, making the findings potentially less generalizable
* The questionnaire is not available on the paper and it's been developed based on interviews with only 14 SWE
* The reported data and content coming from the interview doesn't seem very relevant in terms of size and completeness (few questions, few options…). The raw data could be more useful that only the one they've reported

**Repo:**

None

**Authors:**

Microsoft

Saleema Amershi et ali

**Year:**

2019