Your Name: Arjun Bedi

MSDS 603 MLOps Assignment 1 – Part 2 (2.5%)
Requirements Gathering

In this assignment, you will gather the requirements for building a specific AI/ML-powered product. You will need to identify the business and technical requirements, assess potential risks, propose mitigation strategies, and outline the high-level components needed for successful implementation of the product. You **will not** need to actually build the product.

Learning Objectives

- Apply MLOps principles to a real-world product scenario
- Practice requirements gathering and analysis for ML systems
- Identify potential risks in ML systems and develop mitigation strategies
- Understand the core components required in an ML product pipeline

Scenario

An EdTech company is developing a personalized learning platform for K-12 students. The platform will use machine learning and AI to analyze student performance data from standardized tests and ongoing assessments within the platform to create customized learning pathways for each student. The system should adapt in real-time to student progress, identifying knowledge gaps, recommending appropriate learning activities, and adjusting difficulty levels to maximize learning outcomes while maintaining student engagement. The platform must eventually work across various subjects, but for now we will focus only on ***reading comprehension***. Ideally, it should accommodate different learning styles, comply with educational privacy regulations (like FERPA), and provide actionable insights to teachers and parents through intuitive dashboards.

Requirements

This assignment is done in **two parts**. Part One was already completed in class, and your answers to Part One should be available to you in Gradescope. Complete Part Two below at home and turn in to Canvas. If you did not attend class for Part One, you must accept a zero grade for this assignment since Part Two depends on your answers to Part One.

Part Two

In this part, use ***any resources you want*** (e.g. team members, internet, AI) to help you answer the below questions. Type your answers directly in this word doc.

Question 1: Define an additional two goals for this project.

- Maintaining student retention - Not only do we want to accelerate the rate of learning for these students, we want to make sure they remember the information for more than a day or week. If the information is flushing out quickly, they aren't really learning, but just memorizing. Retention is a big metric to help us understand how much this product is helping.

- Increasing student motivation for learning - Students who are forced to learn nor have the motivation or curiosity to go ahead will not advance at the same level as of those who are. Another goal for this product is to encourage this mindset, and to make learning fun for these students who do not look at it this way currently.

Question 2: For each additional goal from Question 1; define a metric to measure success of that goal.

- Maintaining student retention - Average exam scores difference after one month

- Increasing student motivation for learning - Session frequency and goal/lesson completion rate

Question 3: Briefly describe data governance considerations for the data sources you previously identified in Part One. Be sure to include data privacy and data quality requirements.

- Standardized Tests **-** FERPA regulations would need to be considered, which is used to protect students information being shared without parents consent. For data quality, we would need to have a record of which version tests are being used, to understand difficulty and assess scores better. Also, we should remove all practice tests out, as students tend to not try as much.

- Product recommended questions, answers, student responses - The recommended questions and answers should ideally be scraped from pre-existing learning materials, from textbooks, exams, and potentially other education tools/businesses. For privacy reasons, it should be obtained and used legally. For students answers, they should be anonymous when saved, to save personal information incase of a data breach, and to limit the people who can access the data when not anonymous (updating ML algorithms). Data quality is big here, as not only do we need to make sure the database is filled with good/appropriate questions, but also correct responses, and not obvious wrong answer choices. Also, we should save the students responses with important metadata for good practices.

- Privacy regulations data **-** Not all students/parents will agree to the same terms as the others will. Thus, it is important to save those preferences to always make sure we are not using

their data unethically. Additionally, it is important to keep this information private. We should also save the information regarding privacy regulations to make sure we are always complying with them.

Question 4: Identify an additional two risks associated with this product and the potential impact of each risk.

- Outdated model - When training, it should be imperative that this model is updated frequently. This is because then, some students may be recommended questions/lessons that they have already done or learned. We could see this lead to students being less engaged, or not trusting the product as much once seeing irrelevant content.

- Low interpretability - Parents and users should be able to understand why all learning decisions are made for that particular user. Without explaining why, or with low interpretable reasoning, impacts could include a reduction in engagement, or even accounts, along with not trusting the product. We could also see a blind trust, and going through with lesson plans that might not be the best for that student at that time.

Question 5: For each additional risk identified in Question 4; propose a strategy to mitigate the risk.

- Outdated model mitigation - We could use AirFlow and ML Flow in parallel to help mitigate this risk. We should be constantly monitoring the data as well, to make sure there is no data drift and our training data is met to the standards we set. We can then use AirFlow to automatically train our model at a frequency we choose, and ML Flow to track performance and make sure our model is only getting better.

- Low interpretability mitigation - To mitigate low interpretability, we could use more interpretable models, along with an agent hand in hand to help have more explainable predictions. Also, we could add a feature for users to ask for further clarification, where the agent would be able to dive deeper in the model for better explanations, research the internet for more insight, or use its own knowledge to give better feedback.

Question 6: Describe, in words, any additional major architectural components needed for this product that you did not already include in Part 1 and how those components interact with each other and with components that you described in Part 1.

- Data Preprocessing Service - I mentioned in Part 1 about having a data warehouse to store the data. However, we can not input this raw information in our model. We need a service that will handle all the preprocessing steps necessary before sending this data into our model, which will work hand in hand with the data warehouse, and ML flow to track data versioning.

- Model Development in Real Time - In Part 1, I wrote down both Airflow to automate our model development and ML flow to track metrics, however, I did not mention a model development service. We would need a service to develop our recommendation system/model,

which will send metadata regarding itself to ML Flow. We can use airflow to make these models update in real time.

Question 7: What other resources did you use to help answer these questions this time?

- I used LLMs to help generate new ideas and researched a few articles explaining some regulations, but only used the one regarding FERPA.

Question 8: Reflect on how you answered each question in Part One when you were working solo and compare it to Part Two. For each question 1-6, write down one thing you learned by answering the question again with assistance and resources. For example: "I learned about the existence of metric X, and that the metric I wrote down in Part One is actually not that useful for this problem."

- Reflection 1 **-** I learned that there can be multiple layers to the same umbrella goal, which we can split up to maximize performance.
- Reflection 2 - I learned that better metrics can lead to simpler enhancement of the product.
- Reflection 3 **-** I learned that although data governance is both privacy and quality, both are not as equally important depending on the context.
- Reflection 4 - I learned that risks should be assessed form the ground up. Before assessing risk for something like user engagement, consider how the model or data collection steps could be improved before hand, which will help mitigate those higher level risks.
- Reflection 5 **-** I learned the importance of mitigation strategies, and how considering these in the early stages will prevent many iterations in the later stages.
- Reflection 6 - I learned to not disregard more 'obvious' parts of the architecture. In part 1, I knew that there would be data preprocessing and model development needed, but did not include it in the major architecture. Without it, the rest of the diagram/architecture does not make sense.

Turning it in

Please type your name at the top of the first page, save as ***pdf***, and submit to Canvas.