CSCE 585: Machine Learning Systems: Homework #2

Shima Oruji and Zach Thomas

October 22, 2022

Contents

1 Introduction

We interviewed with three senior ML engineers from amazon, walmart labs, and F5 networks. The amazon engineer works in the Last Mile Delivery system where they optimize the time and cost for amazon deliveries. The walmart engineer works on the ranking team where they seek to determine what to show and in which order when someone searches for a product in www.walmart.com. The F5 networks engineer works on the fraud detection team, where they are aiming at developing a fraud detection system in transactions for famous credit card companies such as Chase.

In the following sections, we list the questions and brief responses from these engineers. For each question, the responses are outlined in the order of amazon, walmart labs, and F5 networks, respectively.

2 Nature of ML task

- What is the ML task you are trying to solve?
  - I am working on a microservice whose goal is to determine how long it would take to deliver a grocery order for our customers. We show this time on our website when a customer makes a grocery order. The input of the model is the order details and the output of that is the rough delivery time.
  - I am working on an ML-based model where we have a certain number of products for a search (the result of the recommendation system) and we want to rank them and determine in which order we want to show them to the customer. The input of the model is a set of products and the output of that is the sorted products based on their probability of interest to our customer.
  - I am working on an ML-based fraud detection system where we want to automatically determine the potential suspicious credit card transactions. The input of the model is the transaction details and the output of that is whether the transaction is fraud or legitimate.
- Is it a classification or regression task?
  - This is a regression ML model.
  - This is a regression ML model.
  - This is a classification model.
- Are the class representations balanced?
  - NA
  - NA
  - No, the class representation is highly imbalanced here as most of the credit card transactions are legitimate.

3 Modeling and experimentation ideas

- How do you come up with experiment ideas?
  - It is based on a team discussion. amazon is a data driven company and if you want to change or propose a procedure, you must support it with some data analysis.
  - A/B testing is an important task usually involves a lot of details and analysis. In short we try to come up with fair experiments and evaluate the models' performances based on their ultimate goal.
  - This is really tricky in our application. We have a dedicated science team who are involved in designing the experimentation procedure for our models.
- What models do you use?
  - We use one of the variants of neural networks.
  - We use multiple models most of which are based on deep learning.
  - We use multiple anomaly detection models.
- How do you know if an experiment idea is good?
  - Online post-deployment evaluations usually verify whether an experiment idea is good or not.
  - If the experiment result is consistent with what happens in practice, it implies that the experiment is good.
  - Based on the real-world results of the models, we can judge the performance of experiment ideas.
- What fraction of your experiment ideas are good?
  - This depends on the team and application. In my case, we are changing the experiment ideas incrementally and in most of the cases (~99%), they are good.
  - A low percentage of experiment ideas are good.
  - Since we have a dedicated team for this, a lot of ideas need to be discussed and evaluated thoroughly before real-world deployment. So the overall number is pretty low.

4 Transition from development to production

- What processes do you follow for promoting a model from the development phase to production?
  - I cannot talk about the details due to NDA. It usually involves pilot deployment, A/B testing, and prod deployment.
  - It involves preprod evaluation, A/B testing, and prod deployment.
  - We have multiple steps for model evaluation before its deployment to production.
- How many pull requests do you make or review?
  - It depends, on average 10 per week.
  - 5-10 per week.
  - 5-10 per week.

- What do you look for in code reviews?
  - Simplicity, correctness, documentation, and how it got tested.
  - I usually compare it with how I would write it and give feedback based on that.
  - I check the structure, cleanness, and documentation.
- What automated tests run at this time?
  - We have unit tests before pushing to pipeline. In the pipeline, we have automated tests before moving to next stage in the pipeline.
  - We have unit tests, integration tests, and few other automated tests.
  - We have unit tests, and performance tests before the model fully goes into production.

5 Validation datasets

- How did you come up with the dataset to evaluate the model on?
  - We have access to a lot of historical data.
  - We have a large volume of historical data.
  - There are some historical data which we can access here.
- Do the validation datasets ever change?
  - Yes, they change very frequently.
  - Yes.
  - Yes. Criminals are always getting better at their job and we need to adapt our models with their advancements.
- Does every engineer working on this ML task use the same validation datasets?
  - Yes.
  - Yes.
  - Yes.

6 Monitoring

- Do you track the performance of your model?
  - Yes, we have online performance evaluation microservices.
  - Yes, we do.
  - Yes, always.
- If so, when and how do you refresh the metrics?
  - It is available in every second. We have some automated alarms if it gets below a certain threshold.
  - It is available in every minute.
  - We have hourly monitoring system.
- What information do you log?
  - Model inputs and outputs along with the performance metrics.
  - Model inputs and outputs.
  - Model inputs and output.

- Do you record provenance?
  - Yes
  - Yes
  - Yes
- How do you learn of an ML-related bug?
  - It is based on the tickets we receive. They could be manually or automatically created.
  - From our ticketing system.
  - From the tickets that are cut to our team.

7 Response

- What historical records (e.g., training code, training set) do you inspect in the debugging process?
  - Training code, training data, offline evaluation data, and ML model.
  - Training and evaluation code and data.
  - Training data and code.
- What organizational processes do you have for responding to ML-related bugs?
  - It is different based on the severity of the issue. For high sev tickets, we need to resolve them as soon as possible (< 24 hours), otherwise they get escalated to higher level engineers. For low sev tickets, they can remain pending for longer periods of time.
  - First the oncall engineer gets paged for the tickets and he can reassign the ticket to other team members if necessary. If the tickets remains pending for longer period of time, the higher level people get involved in the process.
  - If the OPS team is not able to resolve the issue, our oncall takes a look at the issue and involves other engineers if necessary.
- Do you make tickets (e.g., Jira) for these bugs?
  - Yes, we have an internal tool for that.
  - Yes, through our internal tool.
  - Yes.
- How do you react to these bugs?
  - As I mentioned, depending on the severity of the issue, we react differently. We usually first try to eliminate the customer impact as soon as possible. Then, we try dive deep and eliminate the root cause of the issue.
  - We go into the details and determine why we had this issue in the system. We try to fix the issue as soon as possible. Then, in the weekly team meeting we review the important issues and discuss the action items.
  - We try to address the bugs very quickly. The main bugs and action items are discussed during the weekly team meetings.
- When do you decide to retrain the model?

– It is either performance based or time based.
– We usually retrain model on regular basis.
– This is based on the age of the model and its online performance.