19. What was kind of evaluation you were doing in the production environment?

What you should consider more often in a production scenario is **revenue** for your model, and A/B test is a must Besides , you can check if the distribution of your prediction is consistent with that of ground truth concerning accuracy and stability for your model.

20. What was no. of requests (hits) your model was receiving on daily basis?

40K requests on daily basis. We have checked the log in the past and found that we usually receive 2.5k to 3.5k requests per hour for the model. Job is scheduled to process the bulk of requests from a file. The file size is usually about 370 MB to 440MB. Job is scheduled for the fixed time at night.

21. How you have implemented logging in the project for any failure cases?

We have implemented logging in our system. It captures each request and its status whether it was successful or failed. Even the exception of the system is captured properly. None of the detail can be missed.

We have the policy to keep a log of the previous 6 months and the remaining log we archive into backup tables.

We only keep a 13-month log including an archival log.

22. How you have integrated a notification (or Alarm) system for your project?

We have implemented our custom notification program. It sends notification emails to our team.

We have an application where an incident gets created automatically if any job failed or any exception occurs in our system. We have an SLA of 99.9%. Our application has been operational it’s repose time has to be less than 1.2 seconds. Else we get a notification then we investigate.

23. How you have implemented model monitoring?

Few of model monitoring we have implemented:

**Model Input Monitoring:** the set of expected values for an input feature, we can check that

a) the input values fall within an allowed set (for categorical inputs) or range (for numerical inputs) and

b) that the frequencies of each respective value within the set align with what we have seen in the past

Depending on our model configuration, we will allow certain input features to be null or not. This is something we can monitor. If features we expect generally to *not* be the null start to change, that could indicate a data skew or change in consumer behavior, both of which would be cause for further investigation.

Model Response time: The number of requests that can be processed data validation and cleaning was also included.

Model Versioning: Usually we retrain our model every weekend concerning new data we receive.

24. How you have derived the final KPI (Key Performance Indicator) for your client?

25. How many dashboards were there in your project?

We have 2 dashboards one is concerning application and one is for our model performance.

We only use it to work model performance-related dashboard.

Successful train rate. In our system, we do out-of-core learning because of the huge dataset. Our team is working to transform training steps in the production environment. We have started adopting mops currently we have not fully transformed our application with MLOps practices within the next 6 months our team will do it. We have started the planning.

26. On which platform do you have productions your model?

We have deployed our model at the AKS cluster. Sometimes we need to scale our API capabilities so we are using AKS.

27. What kind of API you have exposed to receive data for the model?.

We have a user interface to upload a file. Where user uploads their file in the backend file will be uploaded to s3 bucket. And we did the configuration in such a way that it can pick the file from s3 buck and it will start the validation file. If the file will be validated successfully. We transform data for prediction and again we upload the file at the s3 bucket. We have a UI where users can see their prediction file generate for the uploaded file. Users will get a notification that the file is available you can download the file.

28. What was the size of your final production environment (system configuration)?

We have 64 GB RAM.

We have GPU and CPU in our system. We have 5 TB of Hard disk.

29. What and all Databases you have used in the project?

The application uses a Microsoft SQL Server. But logging related to our machine learning operation we used MongoDB.

30. What kind of optimization you have done in your project, to what depth & explain the example.

31. Can you please talk about complete team structure and team size?

Our team was divided into MLOps, developers, data engineers, data scientists. 24 members were there in my team.

32. What was the duration of your complete project?

My project was all about 9 months.

33. What was your day-to-day responsibility in the last 2 months?

We usually try various random experiments and trying to analyze changes in data and estimate whether we should incorporate another mechanism to transform our data to get a more generalized prediction. Our meeting is scheduled with the client on weekly basis. If we want to highlight anything.

Previous it was daily as we were working proactively client decided to switch on weekly basis.

34. What kind of change request you have been receiving after your productions project

35. What kind of testing you have done in development, UAT, pre-pod, and prod?

Stress testing. Integration testing. Of course model evaluation is the core part of testing.

We tested API. We also simulated malicious requests to check robustness.

36. Have you used some of the predefined AI-OPS pipelines if yes explain.

We implement AIOps not completed but you can 60%. We have done the data version by ourself. We develop logic for that. We have deployed our application AKS.

We did Data Version COntro, CI and CD.

CML was not entirely implementing. We have to manually trigger training if it’s required or Schedule the training at a certain time.

37. Who has implemented AI-OPS in your project?

AIOPS team has done. 2 person was in AIOPS team who designed the whole workflow. We have used github and dvc and mlflow and circle ci actions.

38. What was the OPS stack you have been using?

GITHUB, GIT, DVC MLFLOW, CIRCLE CI , DOCKER HUB, elastic container service

39. What do you understand by CI-CD & have you implemented those in your project? If yes, what was the tech

stack you used for the CI-CD pipeline?

CI: Frequent changes have to be integrated with the whole application on daily basis. We should have automated test cases to check against every checkin of code. a notification has to be sent if test cases pass or fail.

CD: Our application can be deployed at any anytime so if we want to release a new version we should not discuss what date would be best because with help of a CD we can do it on daily basis but you should be careful if it’s required to meet demand

GITHUB, GIT, CIRCLECI, DOCKERHUB, and HEROKU.

40. What was the biggest challenge you faced in the project and how you have resolved it?

Data was huge we can not train our model on whole data at once we impleted incremental learning.

Where we have created mini-batches of the dataset and then trained our model.

41. Give me one scenario where you worked as a team player?

42. What was your overall learning from the current project?

I have seen end-to-end real work machine learning application. I have seen how team work is important to success for project delivery timeline. One of the most important like project planning and defining milestones to achieve within specific days and ensuring overtime if everything is going as per the plan if not what is the reason. What is another alternative to resolve the problem for time being. Evaluating risk and develop a strategy to encounter the worst-case scenarios.

Technically: I got familiar with cloud infrastructure and MLOps practices. Different types of orchestration toll-like airflow

43. How do you keep yourself updated on new technology?

I devote few hours to learn new technology on daily basis. I don’t push myself for long hour continuous learning. I usually decide what tech I am interested in then I use to start exploring those.

44. Have you designed an architecture for this project? If yes, define a strategy wrt to your current project.

No, I haven’t designed architecture for the project.

How many images you have taken to train your DL model?

80K

What is the size of the model that you have in your production system?

It was around some KB. But I am not sure what was the exact size of my model. It may be aound 150 KB to 250 KB

Have you tried optimizing this Vision or DL model?

65. Where you have hosted your Computer Vison model?

What was your frame per second?

What is the data filtration strategy you have defined for the CV project in production?

Have you used any edge device in this project, if yes, why?

What was the name of the camera & camera quality?

What was the outcome you were generating from these devices?

Have you processed the data in the local system or the cloud? Give reason.

72. How many number devices do you have productions (camera, edge devices, etc.)

Let’s suppose I am trying to build a solution to count the no. of the vehicle or to detect their no. plate or track their

speed. Then what is the dependency of distance, position & angle of the camera on your final model? What will

happen to your model? if we change position angle.

What was your data collection strategy in the CV project, have you received data from the client or you have

created the data? And how you have implemented it?