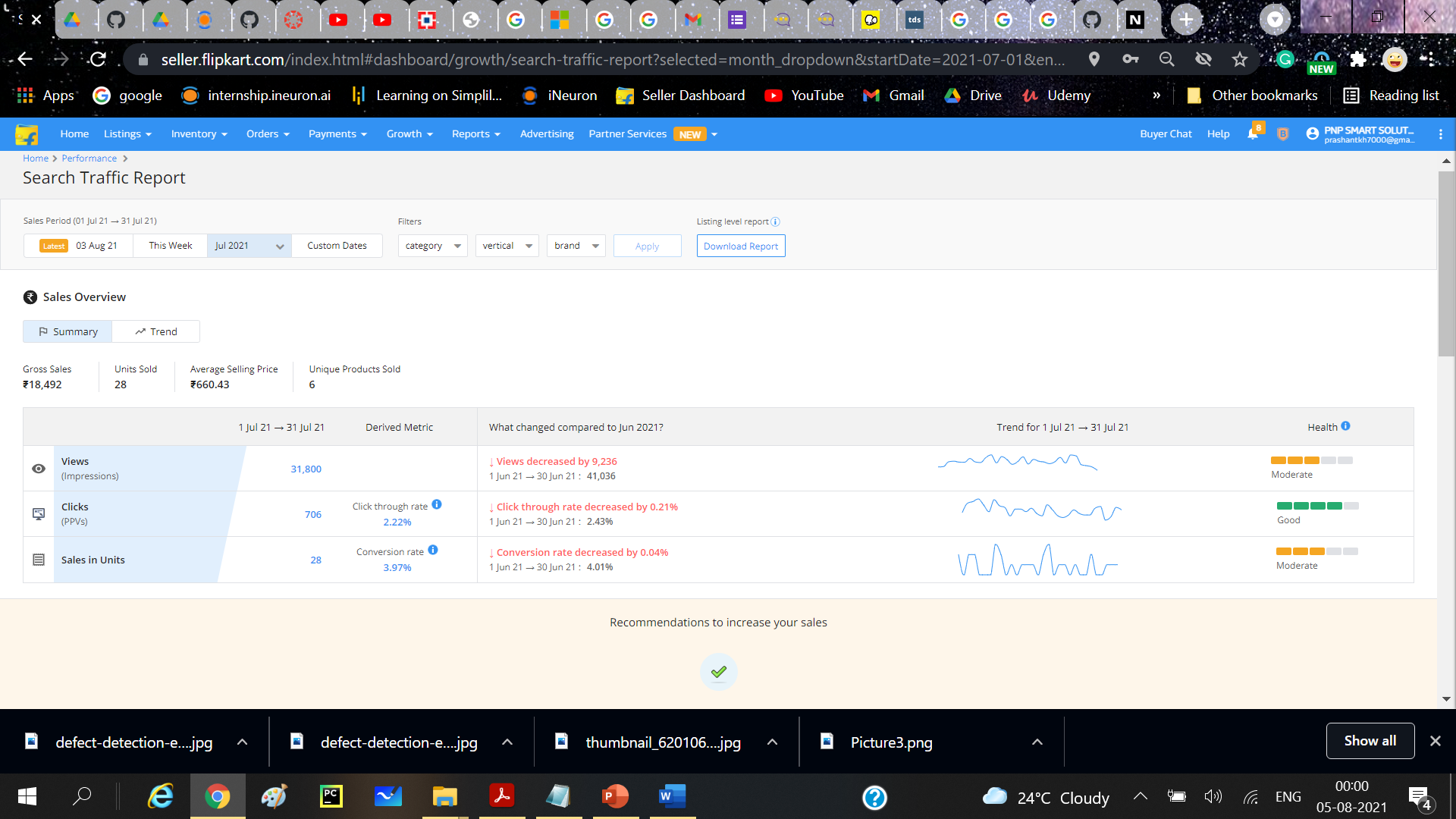
**Project Name: E-Commerce Search Traffic Prediction**

**Q1. Tell me about your current project.**

The project is called E-Commerce Search Traffic Prediction

This project is useful specially for E-commerce Vendors. Vendors need to manage the inventory as per their sales. Vendors must need to know what are the expected sales in next month. So they can plan accordingly and keep safety stock. As Vendors get fined, if they could not deliver product on time. This model is very much necessary for them. So vendors can see the prediction based on previous few year data. Also , Vendor may get information about, How are they growing on each month and what should be the expected plan.



**Q2. What was the size of the data?**

Total 3 year data is collected having 1095 rows

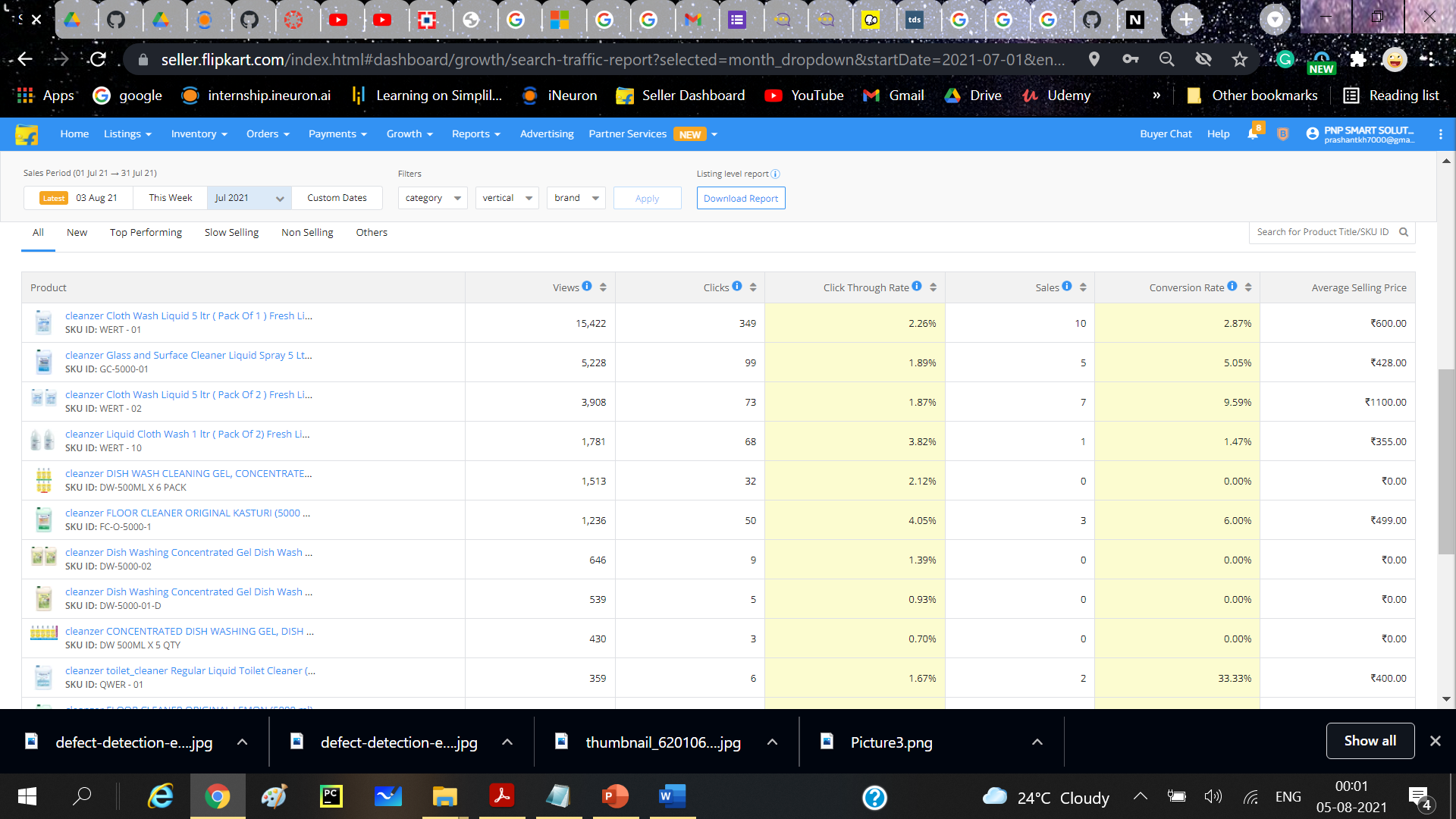
**Q3. What was the data type?**

**Answer:**

The data used for training this model consisted of columns and rows taken from e-commerce website in csv format

Columns names are

* Products
* Views
* Clicks
* Click Through Rate
* Sales
* Conversion Rate
* Average Selling Rate



**Q4. What was the team size and distribution?**

**Answer:**

The team consisted of:

* 1 Product Manager,
* 1 Solution Architect,
* 1 Lead,
* 2 Dev-Ops engineers,
* 2 QA engineers,
* 2 UI developers, and
* 3 Data Scientists.

**Q5. What Hadoop distribution were you using?**

The Hadoop distribution from Cloudera was used as it provides many of the much-needed capabilities out of the box like multi-function analytics, shared data experience with optimum security and governance, hybrid capabilities for support to clouds, on-premise servers as well as multi-clouds.

**Q6.What is the version of distribution?**

**Answer:**

CDH – 5.8.0

**Q7.What was the size of the cluster?**

**Answer:**

The cluster(production setup) consisted of 15 servers with

* Intel i7 processors
* 56 GB of RAM
* 500 GB of Secondary storage each
* Mounted NAS locations

**Q8. How many nodes were there in all the Dev, UAT, and Prod environments?**

**Answer:**

The necessary coding was done on one development server. But as a standalone machine won’t give enough speed to train the model in a short time, once we saw that the model’s loss is decreasing for a few numbers of epochs in the standalone machine, the same code was deployed to a cloud-based GPU machine for training. Once the model was trained there, we used the saved model file for prediction/classification. The same model file was deployed to the cloud UAT and Production environments.

In total, we had:

* 5 nodes in the dev environment,
* 5 nodes in UAT, and
* 15 nodes in production.

**Q9. How were you creating and maintaining the logs?**

The logs are maintained using MongoDB. The logging starts with the start of the application. The start time of the application gets logged. After that, there are loggings for entry and exits to the individual methods. There are loggings for the error scenarios and exception block as well.



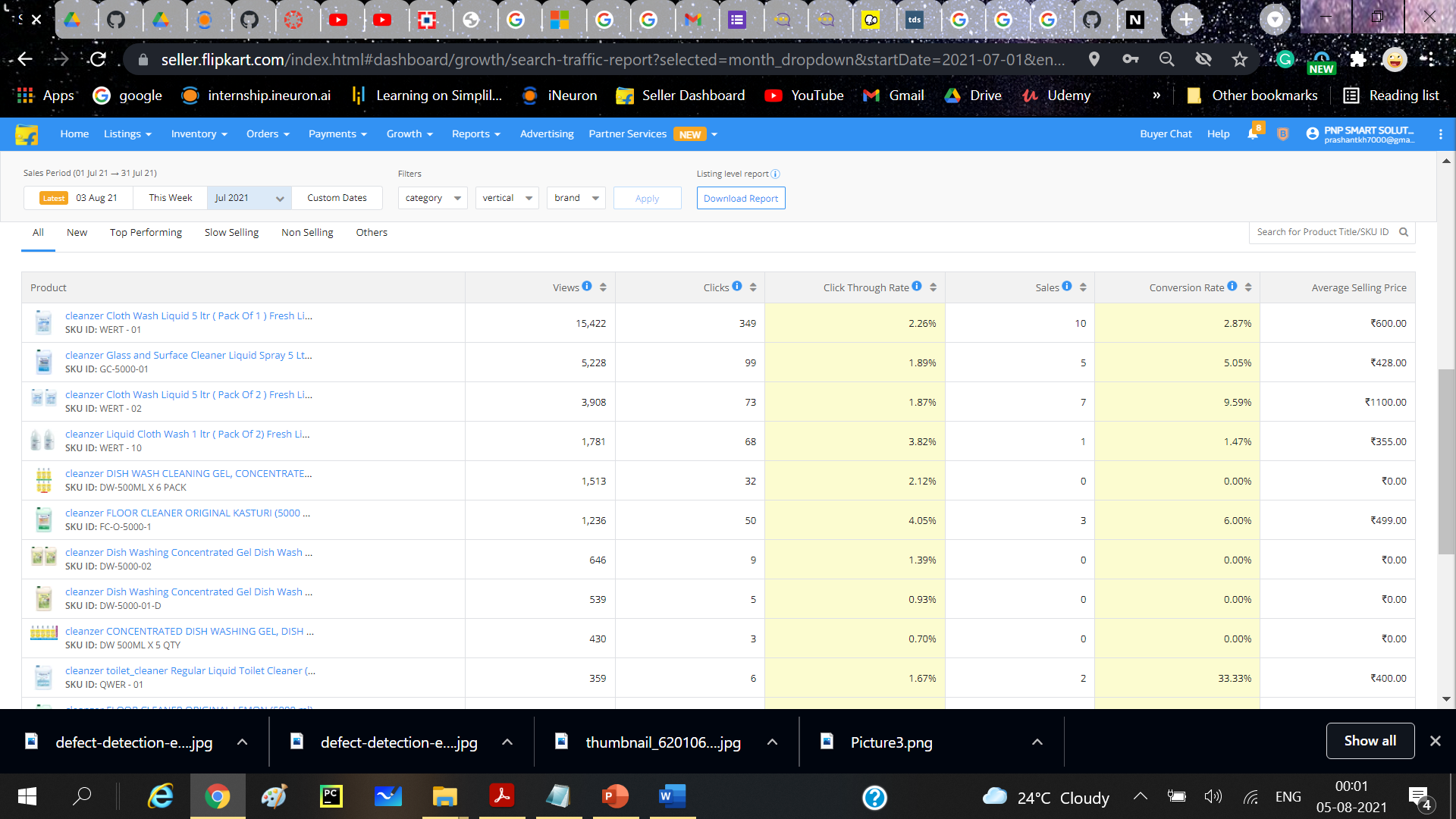
**Q10. What techniques were you using for data pre-processing for various data science use cases and visualization?**

**Answer:**

There are multiple steps that we do for data preprocessing, like data cleaning, data integration, data scaling, etc. Some of them are listed as follows:

While preparing data for a model, data should be verified using multiple tables or files to ensure data integrity.

Identifying and removing unnecessary attributes.



For example,

Here, the Conversion rate column does not contribute to the customer sales for purchasing the products. So, it can be dropped from the dataset.

Identifying, filling or dropping the rows/columns containing missing values based on the requirement.

**Q11. How were you maintaining the failure cases?**

**Answer:**

Let’s say that our model was not able to make a correct prediction. In that case, that data gets stored in the database. There will be a report triggered to the support team at the end of the day with all the failed scenarios where they can inspect the cause of failure. Once we have a sufficient number of data, we can retraining the model for better model performance.

**Q12.What kind of automation have you done for data processing?**

**Answer:**

We had a full-fledged ETL pipeline in place for data extraction. Employers already have images of their employees. That data can be easily used after doing pre-processing for training the image identification model.

**Q13.Have you used any scheduler?**

**Answer:**

Yes, a scheduler was used for retraining the model after a fixed time (20 days).

**Q14.How are you monitoring your job?**

**Answer:**

There are logging set-ups done. We regularly monitor the logs to see for any error scenarios. For fatal errors, we had email notifications in place. Whenever a specific error code, which has been classified as a fatal error occurs, email gets triggered to the concerned parties.

**Q15. What were your roles and responsibilities in the project?**

**Answer:**

My responsibilities consisted of gathering the dataset from e-commerce site, cleaning and pre-processing the data set, training the model on the prepared dataset, deploying the trained model to the cloud, monitoring the deployed model for any issues, providing QA support before deployment and then providing the warranty support post-deployment.

**Q16. What was your day to day task?**

My day to day tasks involved completing the tasks assigned to me, attending the scrum meetings, participating in design discussions and requirement gathering, doing the requirement analysis, data validation, Unit test for the models, providing UAT support, etc.

**Q17.In which area you have contributed the most?**

**Answer:**

I contributed the most to data pre-processing and model training areas. Also, we did a lot of brainstorming for finding and selecting the best algorithms for our use cases. After that, we identified and finalized the best practices for implementation, scalable deployment of the model, and best practices for seamless deployments as well.

**Q18.In which technology you are most comfortable?**

**Answer:**

I have worked with machine learning and deep learning.

I have keen interests towards NLP and I am learning it.

**Q19.How you rate yourself in big data technology?**

**Answer:**

I have introductory knowledge about Big Data. I have worked with AWS so far. But as I am focusing to become data scientist. I will get sufficient knowledge about Big data whichever is required.

**Q20. In how many projects you have already worked?**

**Answer:**

Till now, I have worked with three projects. I have worked in two small and one large scale projects, based on object detection, object classification, object identification, machine learning regression, and classification problems.

**Q21. How were you doing deployment?**

The mechanism of deployment depends on the client's requirement. For example, some clients want their models to be deployed in the cloud, and the real-time calls they take place from one cloud application to another. On the other hand, some clients want an on-premise deployment, and then they do API calls to the model. Generally, we prepare a model file first and then try to expose it through an API for predictions/classifications. The mechanism in which he API gets called depends on the client requirement.

**Q22. What kind of challenges have you faced during the project?**

The biggest challenge that we face is in terms of obtaining a good dataset, cleaning it to be fit for feeding it to a model, and then Pre-processing the prepared datasets. Pre-processing is a rigorous task and it burns a lot of hours. Then comes the task of finding the correct algorithm to be used for that business case. Then that model is optimized. If we are exposing the model as an API, then we need to work on the SLA for the API as well, so that it responds in optimum time.

**Q23.What will be your expectations?**

**Answer:**

It’s said that the best learning is what we learn on the job with experience. I expect to work on new projects which require a broad set of skills so that I can hone my existing skills and learn new things simultaneously.

**Q24. What is your future objective?**

**Answer:**

I always want to be a tech person who work with automation and AI industry. So I will be contributing towards this technologies by doing extensive research in this field.

**Q25.Why are you leaving your current organization?**

Currently I am handling a chemical manufacturing business. I have good knowledge of manufacturing industries. AI is future technology and I have already used some of the techniques in my master degree research work. I have used Regression analysis, Anova test, Concept of central tendency and many statistical concepts and tools. So I will be better fit in this industry.

**Q26. How did you do Data validation?**

**Answer:**

Data validation involves comparing the information entered against that available in reliable third-party databases. For example, when a customer enters his street address, this may be compared to the address saved in the voter records. If the address does not match, the entry may be flagged.

Customer data validation is verifying information like emails, mailing addresses, and phone numberwhich is crucial information that gives you a deeper understanding of who your customers are and keeps you in touch with them.

**Q27. How did you do Data enrichment?**

**Answer:**

With real-world data at their fingertips, the vendors got a view of their audience and discovered more than what meets the eye. With deep audience insights on behaviours, affinities, interests and intent, the vendors identify and activate their potential buyers, which, in turn, resulted in better understanding of their consumers and their data.

**Q28. How would you rate yourself in machine learning?**

**Answer:**

Well, honestly, my 10 and your 10 will be a lot different as we have different kinds of experiences. On my scale of 1 to 10, I’ll rate myself as an 7.2.

**Q29. How would you rate your self in distributed computation?**

**Answer:**

I’d rate myself a 5 out of 10.

**Q30.What are the areas of machine learning algorithms that you already have explored?**

I have explored various machine learning algorithms like Linear Regression, Logistic Regression, L1 and L2 Regression, Polynomial Regression, Multi Linear Regression,Decision Trees, Random Forests, , PCA, XG Boost, CAT Boost, ADA Boost, Gradient Boosting, K-Means,K-Means ++,LDA, KNN, SVM, SVR,Naïve Bayes, Agglomerative clustering, DBScan, Hierarchical clustering are some of them.