Income prediction

Q1. Tell me about your current project.

Ans :
The project is called income prediction. The purpose of this project is to use data transformation and machine learning to create a model that will predict a salary when given years of experience, job type. The purpose of this project is to use data transformation and machine learning to create a model that will predict a salary when given years of experience, job type.

Q2. What was the size of the data?
Ans : We used upto 10 to 90 mb data.

Q3. What was the data type?
Ans:
The data used for training this model consisted of employe salary deatials numercal and categoricals are converted. which have a float 32 representation

Q4. What was the team size and distribution?
Ans:
The team consisted of :

- 1 Product managers
- 1 Solution Architect,
- 1 Lead,
- 1 Devops Enigneer,
- 1 QA Engineer
- 2 UI Developers and
- Data Scientist

Q5.What is the version of distribution?
Ans:
IP- 4.2.1

Q6.What was the size of the cluster?
Ans:

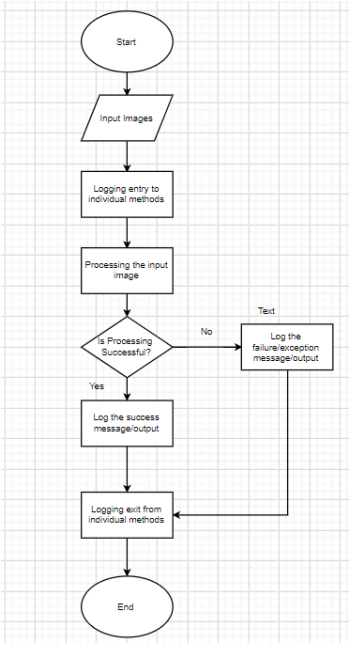
The cluster(production setup) consisted of 1servers with

- Intel i7 processors
- 8 GB of RAM
- 100 GB of Secondary storage each

Q7. How many nodes were there in all the Dev, UAT, and Prod environments?
Ans:

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.

Q8. How were you creating and maintaining the logs?
Ans: 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.



Q9. What techniques were you using for data pre-processing for various data science use cases and visualization?
Ans:
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.
- Identifying, filling or droping the rows/columns containing missing values based on the requirement.
- Identifying and removing outliers
- Converting the categorical data into numerical data.
  For example, gender data (Male or Female) is a categorical one. It can be converted to numeric values, as shown below:
  df['Gender']=df['Gender'].map({'F':0, 'M':1})
- Replacing or combining two or more attributes to generate a new attribute which serves the same purpose
- Trying out dimensionality reduction techniques like PCA(Principal Component Analysis), which tries to represent the same information but in a space with reduced dimensions
-

Q10.How were you maintaining the failure cases?
Ans:
Let’s say that our model was not able to make a correct prediction for an income. In that case, that salary details gets stored in the database. There will be a report triggered to the support team at the end of day with all the failed scenarios where they can inspect the cause of failure. Once we number of cases, we can label and include those salary details while retraining the model for better model performance.

Q11.What kind of automation have you done for data processing?
Ans:
We had a full-fledged ETL pipeline in place for data extraction. Income (salary) details already have of their employees. That data can be easily used after doing pre-processing for training the image identification model.

Q12.Have you used any scheduler?
Ans:

Yes, a scheduler was used for retraining the model after a fixed time.

Q13)How are you monitoring your job?
Ans:

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.

Q14. What were your roles and responsibilities in the project?
Ans:

My responsibilities consisted of gathering the dataset, labeling the images for the model training, 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.

Q15.What was your day to day task?
Ans:

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

Q16.In which area you have contributed the most?
Ans:

I contributed the most to image labeling 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.

Q17.In which technology you are most comfortable?
Ans:

I gave worked often in the Frontend and Backedn application. But my focus and extensive contribution have been as a data scientist.

Q18.How you rate yourself in big data technology?
Ans:

Actually i dont know about big data technology.

Q19. In how many projects you have already worked?
Ans:

It’s difficult to give a number. But I have worked in various small and large scale projects, e.g., object detection, object classification, object identification, NLP projects, chatbot building, machine learning regression, and classification problems.

Q20. How were you doing deployment?
Ans:
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.

Q21.What kind of challenges have you faced during the project?
Ans:
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 labeling the prepared datasets. Labeling 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

Q22.What will be your expectations?
Ans:
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.

Q23. What is your future objective?
Ans:
The field of data science is continuously changing. Almost daily, there is a research paper that changes the way we approach an AI problem. So, it really makes it exciting to work on things that are new to the entire world. My objective is to learn new things as fast as possible and try and implement that knowledge to the work that we do for better code, robust application and in turn, a better user/customer experience.

Q24. Why are you leaving your current organization?
Ans:
I was working on similar kinds of projects for some time now. But the market is rapidly changing, and the skill set required to be relevant in the market is changing as well. The reason for searching a new job is to work on several kinds of projects and improve my skill set.

Q25. How did you do Data validation?
Ans:
Data validation is done by taking the employes salary details gathered. By applying EDA using catogerical and numerical data.

Q26. How would you rate yourself in machine learning?
Ans:
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 8.2.

Q27. How would you rate your self in distributed computation?
Ans:
I'd rate myself a 8.5 out of 10.

Q28. What are the areas of machine learning algorithms that you already have explored?
Ans:
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, Extra Trees Classifier, PCA, TSnE, UMAP, XG Boost,CAT Boost, ADA Boost, Gradient Boosting, Light Boost, K-Means,K-Means ++,LDA,

QDA, KNN, SVM, SVR,Naïve Bayes, Agglomerative clustering, DBScan, Hierarchical clustering, TFIDF, Word to Vec, Bag of words, Doc to Vec, Kernel Density Estimation are some of them.

Q29. In which part of machine learning have you already worked on?
Ans:

> I have worked on both supervised and unsupervised machine learning approaches and building different models using the as per the user requirement.

Q30. What is your area of specialization in machine learning?
Ans:
I have worked on various algorithms. So, It’s difficult to point out one strong area. Let’s have a discussion on any specific requirement that you have, and then we can take it further from there.