Dear Interviewer,

To proceed with my hiring process, here is my process to solve the tasks and challenges.
I am pleased to submit a report on the recent data science project I have been working on, which involves a classification tasks for SnappBOX. The project aimed to analyze a dataset related to delivery services and develop a predictive model to classify certain outcomes.

The project included preparing a suite of data manipulation and visualization tools.
The exploratory data analysis (EDA) phase followed, where I reviewed the dataset to understand it. During the EDA, I took a deep dive into the quality and structure of the data. I examined the relationships between different variables, the distribution and range of the numerical values, and the prevalence of any categorical data. This process was instrumental in uncovering insights that would later inform the feature engineering and model training phases.
For example, handling data for missing values in each column.

Part 1 Problem Definitions:

In Summary part 1 is about a task that takes information about an order and predicts whether it will be hyper-ack or not (binary classification).

After all the processes of EDA, I started the feature engineering and model training phases.
In the feature engineering section, I created a column named **geo_cluster**. Apply K-Means to cluster provided geo-locations in the dataset, and then use this as a feature.
After that, I preprocessed the data and dropped unnecessary columns, then split the data into test and train for the training and evaluation phases.

As you mentioned I train data with 3 Algorithms Logistic Regression, XGBoost, SVM.
In the evaluation phase, I plot some results to visualize them for a better understanding of the classification and the prediction of my models.
As you can see in my results and visualizations **XGBoost** has the best results because, with the same data, it gains the highest recall, precision, and accuracy compared to other models.

Part 2 Problem Definitions:

In Summary, part 1 is about the Fraud Detection Algorithm task.
You provide a fraud samples dataset that helps me to find the best features in fraud detection.
In the fraud samples dataset, a column is reasons that indicate the most important fraud reasons for me. I do feature engineering and combine the reasons to create the is_fruad column for the main part2 dataset and the add fraud sample to it, then I put is_fraud True for them and other False. After all, I created a new dataset named part2-dataset-with-is_fraud.csv.
After creating a new dataset I do some EDA and data visualization on the new dataset like plotting the Distribution of Fraudulent vs Non-Fraudulent which gives me insight into the imbalance data. To figure out imbalance data I used oversampling to create more balance data for my training.
In order to do fraud detection I use both **Anomaly detection and classification.**
I used anomaly detection because of imbalance data, since we talked about it in my last interview we can do this fraud detection by an anomaly. As you know fraud usually is an anomaly in our data because It happens rarely and we can solve it by anomaly detection.

For anomaly detection, I used IsolationForest Algorithm and for the classification solution, I used GradientBoostingClassifier which can provide a good classification model.
After training the model I evaluated the model and plotted the result in a confusion matrix that gave us a good insight into the results.

Best regards,

Shahriar