Anomaly Detection Summary Report:

Alon Firestein
314984402

I solved this problem by looking and the data using data exploration techniques taught in class and trying to visualize the data and using the Isolation Forest algorithm that we were taught, I would feed it all the necessary features of the dataset and get a relative score using the built-in decision function, and then we will predict if the data point was an outlier or not.

3a.) What is the approach you tried? Why them?

I went with the "all in" approach. There aren't many features in this dataset, which one of them being completely useless (record_ID). So, to let the algorithm work on this dataset I dropped that feature given that it does not help our model at all. The rest of the features are critical in finding the outliers in our data, and their values are all numeric and important. So, all of them went into the model.

I chose to use Isolation Forest for this approach because continuously “isolates” anomalies by creating decision trees over random attributes. Which over time can try and recognize the outliers in a field of normal data points. The random partitioning of the created decision trees produces noticeable shorter paths for anomalies, and that's why when a forest of random trees collectively produces shorter path lengths for some particular points, then they are highly likely to be anomalies.

Additional explanations about isolation forest and why I chose to go with that algorithm is located above the code with the "Isolation Forest" headline.

3b.) How do you know the algorithm is good?

To properly test my algorithm and if its results were good, I used sklearns metrics method to give it a score. Given that I was given a second dataset that does indeed contain the true labels, I imported that dataset and used the true labels to check my models scores at the end (accuracy, precision, f-score, recall).

Although the accuracy here doesn't have a real factor on our model because the dataset is so unbalanced that even if we told the model to predict "not an outlier" for every row, then we would have an accuracy score of over 99%. Therefore, it is important to better rate our recall, precision, and f-1 scores.

Therefore, given our F1 score, and our precision score has a good outcome given that size of the dataset (over 250,000 data rows) and that the sample size of anomalies in the data was small in comparison (a bit over 1000 anomalies). I would say that the outcome of my model is a success. I can see the model did not overfit the data and the scores were proper.