## 3. Results & Discussion

### 3.1 Results

```
[[1906237     130]
 [   1252    1167]]
0.9992759796016945
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1906367
           1       0.90      0.48      0.63      2419

    accuracy                           1.00   1908786
   macro avg       0.95      0.74      0.81   1908786
weighted avg       1.00      1.00      1.00   1908786
```

The confusion matrix has given us information on the True-Positive (TP), False-Positive (FP), True-Negative (TN) and False-Negative (FN) values which are used to determine the values for precision, recall and f1-score in the classification report. As we can see, our model has a very high precision for both 0 and 1 classes, but class 1 recall value is quite low as the number of samples inside the datasets for this class is not enough in comparison to class 0 which would then affect the f1-score. We could justify this by looking at the support values for class 0 and 1. Nevertheless, since our model has achieved a precision of 1.00 and 0.90 for the classes with an accuracy approximating to 1.00, it can be said that we are successful in training the model by implementing Logistic Regression.

## 3.2 Discussions

For this model, mainly we used Logistic Regression Algorithm and RandomisedSearchCV method for the training of our model. We also used Robust Scaler method, train-test-split method, and Stratified K Fold method in the model before the training occurs. The combination of these techniques helps us to enhance the logistic regression model in many ways. Robust Scaler Method, for an example helped to preprocess our dataset by scaling the data making it robust towards outliers. The Train-Test-Split Method in the other hand helped in training the logistic regression model and the testing is used to evaluate its performance. By splitting the data, we assessed how well our model generalizes to unseen data thus increasing its effectiveness in detecting fraudulent transactions. The stratified K-Fold method performs a cross-validation that helps to estimate the performance of the model making it more reliable. Our dataset works well with stratified sampling as our data is either to identify whether it is fraudulent or non-fraudulent. Lastly, the RandomisedSearchCV method helps us with its automated hyperparameter tuning of the Logistic Regression Algorithm. Therefore, all these methods help each other by ensuring our dataset is properly scaled, allowing for a more reliable model evaluation through cross-validation, thus, improving its performance.

## 3.3 Opinions on Result

Despite the fact that our model obtained accuracy close to 100%, we believe that this accuracy is due to the sample prediction being too shallow because the number of fraudulent cases in the dataset was less than 1% of the overall sample. There are also outliers that needed to be scaled, making our data to not be used as it is. Maybe if there are more fraudulent cases out of multiple types in our data set, the dataset can then be used as it is without having to scale to an extreme end, although

it might struggle to achieve almost 100% accuracy then but at least our model will show it flaws and allow us to improve it to the smallest of details.

### 3.4 Opinions on improvements

We have discussed and analysed potential enhancements to our model. Our perspective is that the datasets, algorithm selection, and dataset cleaning are the three main areas that need to be addressed in order to improve the accuracy and precision of our model.

In order to train fraud detection models, datasets are essential. Acquiring high-quality data sets that cover a sizable number of various fraudulent incidents is crucial to enhancing accuracy and precision. By expanding the dataset, the model is more able to identify and understand various fraudulent patterns, abnormalities, and complex linkages. To ensure a more complete portrayal of fraud incidents, data sets should contain a higher proportion of numerous fraudulent cases. As a result, the model will be better equipped to predict fraud trends with greater accuracy and precision.

Although logistic regression is a method that is widely used to solve classification issues, it might not always be the ideal choice for detecting fraud. Based in the literature review, potential benefits in this situation can be seen in the Random Forest algorithm, an ensemble learning technique that integrates different decision trees. Random Forest can effectively handle complex data and capture non-linear relationships between variables, thus enhancing the model ability to detect fraud. Exploring the implementation of the Random Forest algorithm to leverage its ability to handle complex data and capture non-linear relationships. This algorithm has shown promising results in various domains and has the potential to improve the accuracy and precision of fraud detection models.