a) What is fraud detection, and why is it important?

Fraud detection is the process of identifying and preventing deceptive actions intended to result in unlawful financial gain or data misuse. It typically involves analyzing data (often in real-time) to recognize patterns that are consistent with fraudulent behavior. Fraud detection is important as it can cause massive financial loss, loss of customer trust, weak operational efficiency. Supervised learning such as logistics regression, random forest and XGBoost, etc can be used to label fraud data. But, class imbalance (fraud cases are rare), false positives(blocking real users should be avoided) and evolving pattern(fraud makers will change strategies) are major challenges.

b) If you change the training and testing split to 70% training and 30% testing, how does the model's performance change?

At first, the data is split into 60% training and 40% testing and the accuracies of different k values are calculated. Then, the accuracies are also calculated using data split of 70 % training data and 30% testing data.. According to comparison of accuracy score (figure) of two data split, 70% split's accuracy is considerably smaller than that of 60% split. And this result is consistent with that of confusion matrix of two different data split(Blue: 60% , Red: 70%)

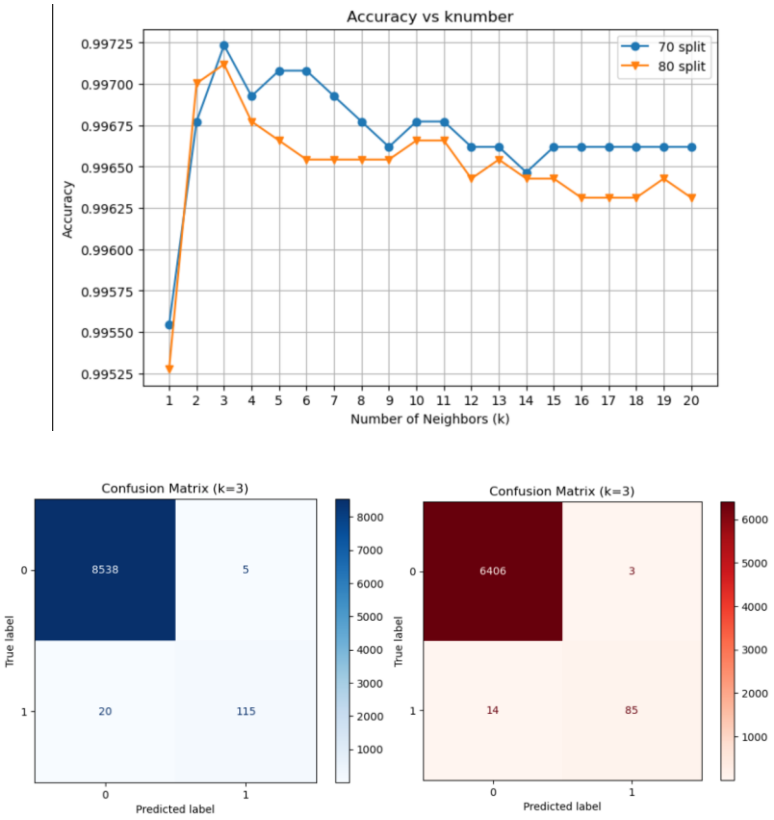

Figure 1 Comparison of 70% split and 60% split

c) Keeping the test size fixed at 40%, try changing the number of neighbors (in KNN). How does the model's performance vary with different K values? Which value gives the best result, and how do you define what makes it the "best"?

Accuracy score, ROC-AUC score, F1 score, precision and recall were calculated for model ranging from 1 to 20(figure 2). From F1 score and accuracy score, it can be concluded that the model's performance is the best when k = 3.