May Mon Thant - MMDT091

Mentor - Ma Nuwai Thet

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

What I understand is that it is detecting how much fraud is included in each transaction which is mostly used in the banking system. Not only about money but also the illegal transfer of data or resources can be regarded as fraud. This is a significant problem for business and governmental organizations.

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

Metric	40 % Testing	30 % Testing
Confusion Matrix	[8538 3 28 109]	[12793 3 39 180]
Precision	1.00	1.00
Recall	1.00	1.00
Precision	0.97	0.98
Recall	0.80	0.78
Accuracy	1.00	1.00
AUC	0.9186	0.9988

The above table shows the difference between 40 % testing and 30 % testing.

There are some slight improvements in both datasets.

When we check the confusion matrix, we can see more true positives for Class 1 in 30 % testing.

In Precision and Recall both models can identify fraud as it improves 0.97 and 0.98.

In recall there is a decrease in recall. Also in AUC in 30 % testing, we can see the improvement.

From the results, we can interpret that the model can get performance with more training data.

So, a larger training set with 70 % can have better model performance in precision and AUC.

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"?

I have tested with different KNN neighbors 1, 3,5, and 9.

Metric	K = 1	K =3	K =5	K =9
Confusion Matrix	[8519 22 25 112]	[8538 3 28 109]	[8538 3 32 105]	[8538 3 32 105]
Precision (Class 0)	1.00	1.00	1.00	1.00
Recall (Class 0)	1.00	1.00	1.00	1.00
Precision (Class	0.84	0.97	0.97	0.99

1)				
Recall (Class 1)	0.82	0.80	0.77	0.76
Accuracy	0.99	1.00	1.00	1.00
AUC	0.9075	0.9153	0.9252	0.9987

If we check K value = 1, the model has more false negatives for Class 1 which we can interpret that it has more missing fraud when compared to others.

As K increases, when it reaches to K = 9, there can be more accurate predictions. So, what if we reduced the number of neighbors, it is good for training but not so good for testing. Besides, as K increases precision and AUC also improves. There is not that much difference between neighbors 3 and 5. The best result might include 20 % of missing data in detection.