Sharmila KC
Professor Panagopoulos
Bus 5040
10/4/2023

Project 5

a. Constructing the confusion matrix and calculating the error rate

The error rate is calculated as the proportion of total predictions that are incorrect. In this case, the error rate is:

	Classification Confusion Matrix	
	Predicted Class	
Actual Class	**1 (Fraudulent)**	**0 (Non-Fraudulent)**
1(Fraudulent)	30	32
0(Non-Fraudulent)	58	920

Error Rate = (FP + FN) / Total Records
Error Rate = (58 + 32) / (88 + 952)
Error Rate = 90 / 1040
Error Rate = 0.0865 (or 8.65%)

So, the error rate for this model is approximately 8.65%.

b. Calculating the error rate if everything is classified as non-fraudulent

	Classification Confusion Matrix	
	Predicted Class	
Actual Class	**1 (Fraudulent)**	**0 (Non-Fraudulent)**
1(Fraudulent)	0	88
0(Non-Fraudulent)	0	952

If we classify everything as non-fraudulent, then the error rate will be the proportion of fraudulent transactions that are incorrectly classified as non-fraudulent. This is equal to the number of fraudulent transactions divided by the total number of transactions:

Error Rate = (FN) / Total Records
Error Rate = (0+88) / (0+88 + 0+952)
Error Rate =88 / 1040
Error Rate = 0.0846 (or 8.46%)

So, if we classify everything as non-fraudulent, the error rate would be approximately 8.46%.

c. Is the data mining routine useful?

Yes, the data mining routine is useful because it has a lower error rate than classifying everything as non-fraudulent. The data mining routine correctly classifies 90.35% of transactions, while classifying everything as non-fraudulent would only correctly classify 91.54% of transactions.

5.4.a:

The leftmost bar on the decile chart shows that the top 10% of records that are most likely to be fraudulent, according to the model, contain 6.5 times as many fraudulent records as a random sample of 10% of all records. The second bar from the left shows that the second decile of most likely fraudulent records contains 2.7 times as many fraudulent records as a random sample of 10% of all records.

5.4.b:

A bank could use the decile chart to identify credit card transactions that are likely to be fraudulent. The bank could then flag these transactions for review or block them altogether.

An insurance company could use the decile chart to identify insurance claims that are likely to be fraudulent. The insurance company could then investigate these claims more closely or deny them altogether.

An e-commerce company could use the decile chart to identify online orders that are likely to be fraudulent. The e-commerce company could then cancel these orders or require additional verification from the customer.

5.6.a

If the 1000 new leads are like those in the pilot, the company can expect the same mean profit per sale of $2128, or $2,128,000 for the 1000 leads. However, the cost of the sales effort is $2500 per lead, for a total cost of $2.5 million. Therefore, the company would lose money by selling to all 1000 leads without using predictive modeling to target its sales efforts.

5.6.b

If the company wants to double the sales effort cost, it needs to target customers who are likely to spend at least $5000. The decile chart shows that the top decile has an average revenue of $5036, which is more than twice the sales effort cost. Therefore, the company should focus its sales efforts on the top decile of customers.

5.6.c

If the company lowers the cutoff to $2500, it can tolerate a lift as low as 1.17. This means that the company can expect to make a profit by selling to customers in the top 5 deciles.