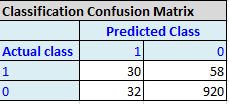
Assignment #9 Assessment of Pred. Models by Joshua Troup

Q1: A data mining routine has been applied to a transaction dataset and has classified: 88 records as fraudulent (30 correctly so) and 952 as nonfraudulent (920 correctly so). • Construct the classification (confusion) matrix and calculate the error (misclassification) rate. Also, calculate the accuracy. (5 points)

****

**Accuracy= A+D/A+B+C+D**

**Accuracy= 30+920/30+58+32+920 = .9134 or 91.34% Accuracy**

**Error Rate = B+C/A+B+C+D**

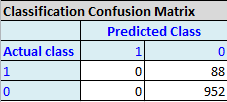
**Error Rate= 58+32/30+58+32+920 = .0865 or 8.65% Error Rate**

Q2: Consider the Figure below showing the decile wise lift chart for the transaction data model in Q1, applied to new data. (5 points)

• Comment /interpret the meaning of the first and second bars from the left.

**The first bars decile 1 (10%) ranks the most probable 1s or in this case frauds. It will catch 6.5 times as many frauds compared to a random selection of the top 10% records. The second decile or the second 10%, it will catch 2.7 times as many frauds than if it was a random selection of 10% of records.**

Q3: Another analyst comments that you could improve the accuracy of the model by classifying everything as nonfraudulent for the transactional data in Q1. (5 points) • If so, what is the misclassification (error) rate? What is the accuracy?

****

**Misclassification (error) rate = 88/1040 = .0846 = 8.46%**

**Accuracy = 952/1040 = .9153 = 91.53%**

Q4: A firm that sells software services has been piloting a new product and has records of 500 customers who have either bought the services or decided not to. • The target value is the estimated profit from each sale. • The global mean is $2128. • However, the cost of the sales effort is not cheap. The company figures it comes to $2500 for each of the 500 customers (whether they buy or not). • The firm developed a predictive model in hopes of being able to identify the top spenders in the future. The lift and decile charts are shown in Figure below. Q4: If the company begins working with a new set of 1000 leads to sell the same services, like the 500 in the pilot study, without any use of predictive modeling to target sales efforts, what is the estimated profit? Will it be a profitable move? (5 points)

**The lift chart shows at 200 cases the predicted cumulative is 400,000. The visual line seems to be at peak at this point therefore if 1000 cases were graphed, I see the cumulative remaining at 400,000 or lower. Based on average cumulative sales (non predictive) the graph shows the cumulative rising at a steady rate with no plateau. 200 cases is 400,000 cumulative. We can assume at 1000 cases, the cumulative would be 2,000,000.**

**Current = 500 customers \* cost $2500 = 1,250,000**

**500 customers \* mean $2128 = 1,064,000**

**1000 customers \* cost $2500 = 2,500,000**

**1000 customers \* mean $2128 = 2,128,000**

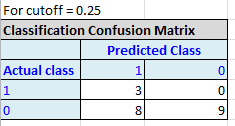
**2,500,000-2,128,000 = 372,000**

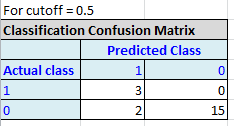
**Does not seem very profitable**

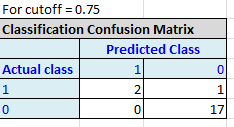
Q5: If the firm wants the average profit on each sale to at least double the sales effort cost, and applies an appropriate cutoff with this predictive model to a new set of 1000 leads, how far down the new list of 1000 should it proceed (how many deciles)? (5 points)

**Decile 1 & 2 would be proficient obtaining double the sales effort cost ($5000). Decile 3 is showing a 1.75 times the average profit across all customer which is not satisfactory to the question.**

Q6: Table 5.5 shows a small set of predictive model validation results for a classification model, with both actual values and propensities. a. Calculate the Confusion Matrix for cutoff values of 0.25, 0.5, and 0.75. At what cutoff, the overall error rate is lowest? (5 points)

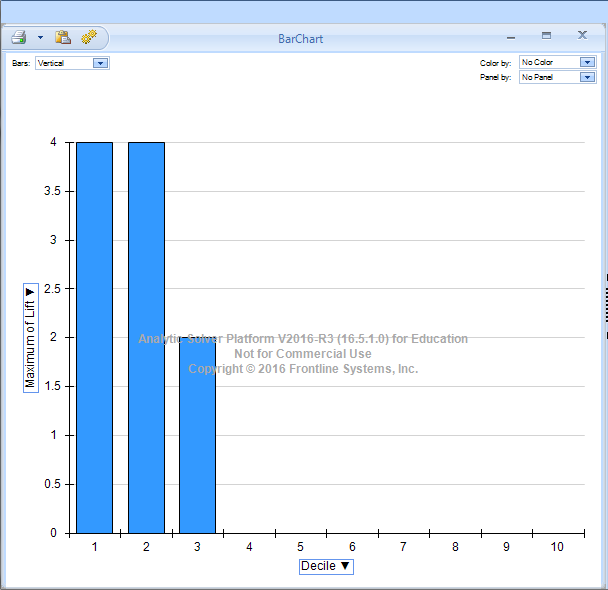
0+8/20 = Error Rate = .4

0+2/20 = Error Rate = .1

1+0/20 = Error Rate = .05

**.75 has the lowest overall error rate of .05**

b. Using a cutoff of 0.5, create a decile lift chart. (5 points)

****