**Case Problem**

**Predictive Analytics**

**Module 2 – Lesson 2**

**Professor: Lee**

#1. A data mining routine has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly and 58 incorrectly) and 952 as non-fraudulent (920 correctly and 32 incorrectly). Construct the confusion matrix and calculate the error rate. Assign 1 as fraudulent.

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **1** | **0** |
| **1** | 30 | 32 |
| **0** | 58 | 920 |

**Error Rate here is = (False Negative + False Positive)/ Total =**

**=58+32/1040=90/1040=.086 = 8.6 %**

#2. Suppose that we created a model to evaluate if our customers at the Ice Cream Shoppe should be offered a credit card. Assume we have past credit card data where 0=no default and 1=defaulted on credit card. The confusion matrix is shown.

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **0** | **1** |
| **0** | 410 | 12 |
| **1** | 12 | 72 |

1. Calculate the Accuracy

Accuracy= (True Positive + True negative) / Total= (72 + 410) / 506=1.04 = .952=95.2%

1. Calculate the % Error rate

Error Rate here is = (False Negative + False Positive)/Total= 24/506=.047=4.7%

1. Calculate the Sensitivity

Sensitivity = True Postive/( True Postive+ False positive)= 72/(72+12)=72/84=.857=85.7%

1. Calculate the false positive rate

False positive rate = False Positives/( True Negative + False positive)= 12 / (410+12)=.028=2.8%

1. Another analyst comments that you could improve the accuracy of the model by classifying everything as no default (0). Is this true? What would the error rate be if you classify everything as no default?

Original confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **0** | **1** |
| **0** | 410 | 12 |
| **1** | 12 | 72 |

Total No default(0)=410+12

Total default(1)=72+12 =84

Now we classify everyone as no default then it means our new matrix would be

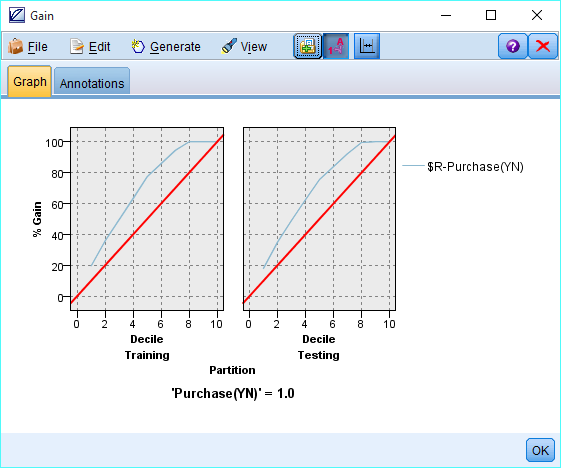
|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **0** | **1** |
| **0** | 422 | 0 |
| **1** | 84 | 0 |

**Now the error rate is** (False Negative + False Positive)/Total = 84 + 0 / 506 = .166 =16.6 %. We can see that error rate has actually increased . In our old matrix , the error rate was 4.7 % but in the new matrix it is increased to 16.6 %. Also, the accuracy with the new matrix is : (True Positive + True negative) / Total = 0+422 / 506 = 83.3 % which is less than the original value of 95.2 %.

Therefore, it is **not true** that model accuracy would incrase if we classify everything as no default(0).

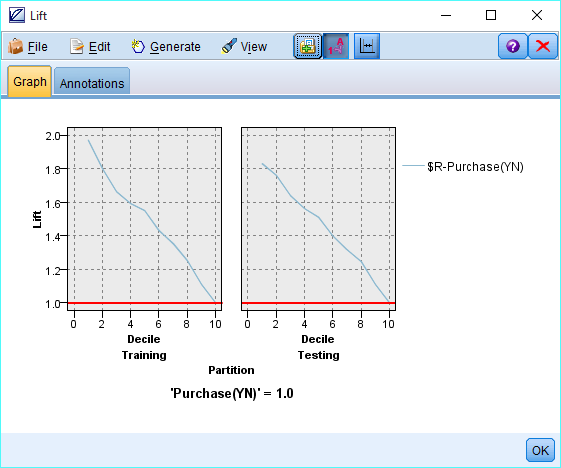
#3. Continue your case problem from last class (M2L1 case problem). Save the stream under a new name for today’s assignment.

* Use 3 ‘Evaluation’ Nodes under ‘Graphs’ to create 3 charts from your model predicting Purchase (Y/N):
  + **Gains – use deciles**
  + **Lift – use deciles**
  + **ROC**
* Insert screenshots of the 3 graphs and interpret.



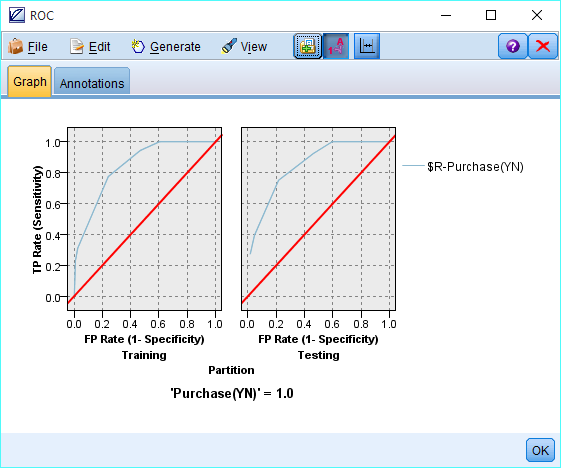
**Gain:**

The figures shows that through the 20 % depth , the model selects nearly 40 % of the 1s i.e purchases, nearly twice as many as a random draw, shown by the red line.



**Lift:**

This shows if we focus on the top 20% (Decile 2) of our observations, we will have a lift of 1.8 better than random choosing. The lift chart shows similar information to the gains chart, just expressed differently



**ROC:**

ROC shows a tradeoff between false positive(specificity) on x axis and True Positive(Sensitivity) on y axis. In the graph we can see that at Sensitivity(True Positive Rate) ~0.7 , specificity( the false alarm rate) is ~0.2 . If the business can tolerate higher false alarm rate then Sensitivity can have value of 1 at specificity of 0.6. End decision is based on the business objective.

**Submit your completed Word file and your stream file on Canvas. In the comments, put the names of the folks who worked on this assignment.**