Print

Quiz Feedback: Measuring Model Performance

Your Score: 100%

Congratulations! Your score of 100% indicates that you've mastered the topics in this lesson. If you'd like, you can review the feedback.

When you're finished, exit the lesson.

1. After you complete model assessment, which of the following models is the most appropriate choice?

- *a.* the simplest model that has the highest training fit measures
- *b.* the simplest model that has the highest validation fit measures
- *c.* the most complex model with the highest training fit measures
- *d.* the most complex model with the highest validation fit measures

Your answer: b
Correct answer: b

A reasonable rule is to select the simplest model associated with the highest validation fit statistic.

2. Which of the following statements is true regarding the preparation of the validation data?

- *a.* The validation data needs to be prepared for scoring the same way that the training data was prepared for model building.
- *b.* Missing values in the validation data set need to be replaced with the medians from the validation data set.
- *c.* The results of the analysis on the training data need to be recalculated on the validation data set.
- *d.* Validation data needs to be treated as if it were truly new data where the target variable values are known.

Your answer: a
Correct answer: a

Answer *b* is incorrect because the missing values in the validation data set need to be replaced with the medians from the training data set.

Answer *c* is incorrect because the results of the analysis on the training data set need to be applied to the validation data, not recalculated.

Answer *d* is incorrect because the validation data needs to be treated as if it were truly new data where the target variable values are unknown.

3. Which of the following statements is true regarding model performance measures?

- *a.* The sensitivity and specificity are affected by oversampling.
- *b.* The traditional ROC curve plots the sensitivity on the Y axis and the specificity on the X axis.
- *c.* If the lift value is 4 at a depth of 40%, then at that given depth, there are four times more responders targeted by the model than by random chance.
- *d.* The positive predictive value is computed as the number of true positives divided by the total actual positives.

Your answer: c
Correct answer: c

Answer *a* is incorrect because sensitivity and specificity are not affected by oversampling. They do not depend on the proportion of each class in the sample.

Answer *b* is incorrect because the traditional ROC curve plots the sensitivity on the Y axis and 1−specificity on the X axis.

Answer *d* is incorrect because the positive predictive value is computed as the number of true positives divided by the total predicted positives.

---

4. Which of the following statements is true regarding the gains chart?

- *a.* The horizontal axis is the depth, which refers to the proportion of cases exceeding different cutoffs of the predicted probabilities.
- *b.* The vertical axis is the sensitivity, which is the true positives divided by the total actual positives.
- *c.* The gains chart is not affected by oversampling.
- *d.* A model with good predictive power has an increasing response rate as the depth increases.

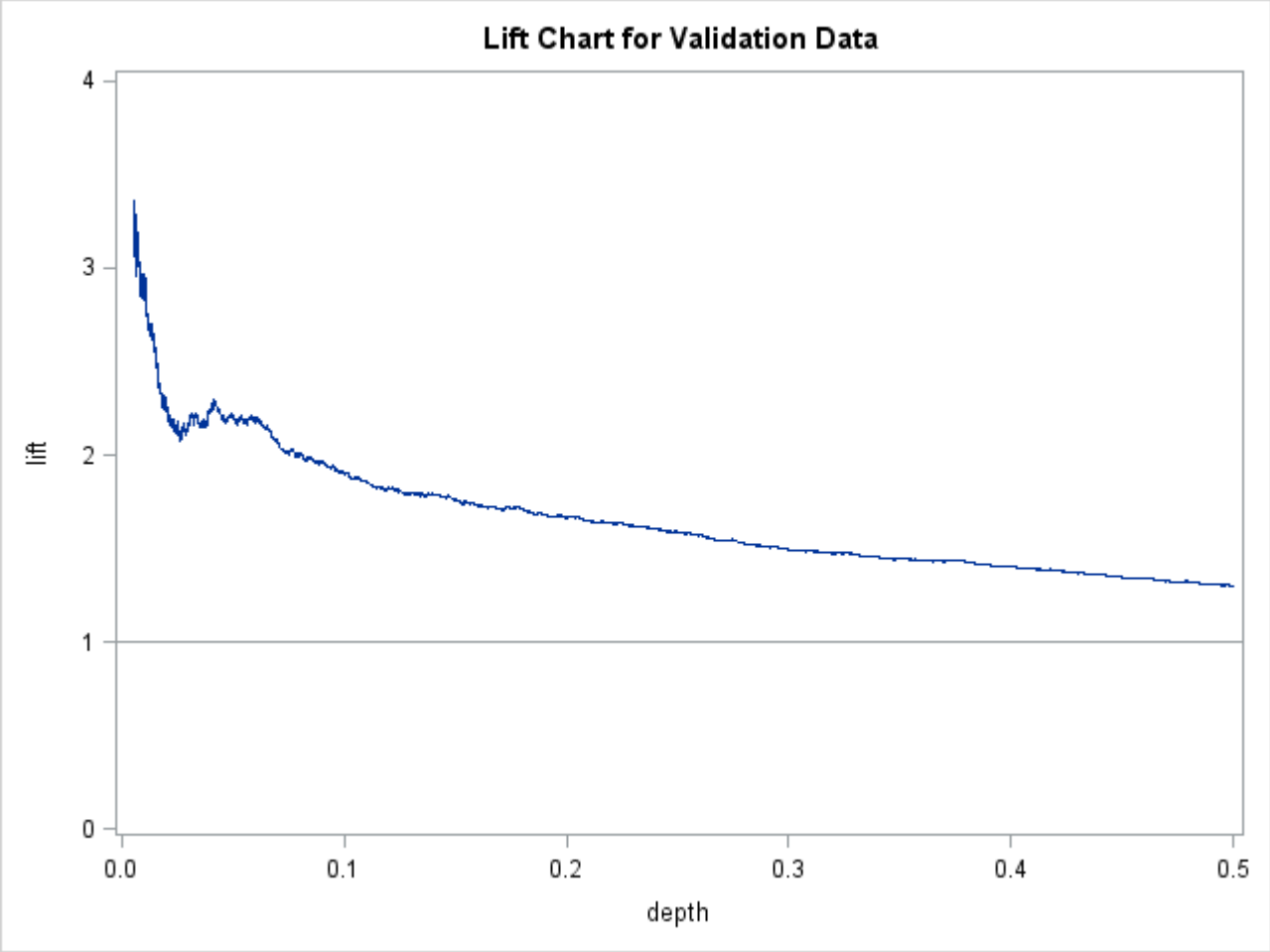
Your answer: a
Correct answer: a

Answer *b* is incorrect because the vertical axis for a gains chart is the positive predictive value.

Answer *c* is incorrect because the gains chart is affected by oversampling. The positive predictive value might be badly overestimated due to oversampling.

Answer *d* is incorrect because a model with good predictive power has an increasing response rate as the depth decreases.

5. The following lift chart is created for the validation data set:



What is the approximate lift at a depth of 10%?

- *a.* 1.3
- *b.* 1.9
- *c.* 2.5
- *d.* 2.7

Your answer: b
Correct answer: b

At a horizontal axis value of 0.10, the curve in the lift chart corresponds to a Y axis value of 1.9.

---

6. Which statement is true regarding decision rules in data mining?

- *a.* Higher cutoffs increase sensitivity and decrease specificity.
- *b.* If no profit matrix is given, always use 0.5 as the cutoff to achieve maximum profit.
- *c.* The plug-in Bayes' rule can achieve the maximum profit even if the estimate of the posterior probability is poorly estimated.
- *d.* A cutoff of the proportion of events in the population tends to maximize the mean of sensitivity and specificity.

Your answer: d
Correct answer: d

Answer *a* is incorrect because higher cutoffs decrease sensitivity and increase specificity.

Answer *b* is incorrect because the use of the central cutoff, π_1, is recommended when no profit matrix is given. The central cutoff tends to maximize the mean of sensitivity and specificity.

Answer *c* is incorrect because the plug-in Bayes rule might not achieve the maximum profit if the estimate of the posterior probability is poorly estimated.

---

7. If the profit margin of true positives is 24 times higher than the loss margins of false positives, then according to Bayes' rule, what is the cutoff that maximizes the total expected profit? (Assume zero profit and loss for true negatives and false negatives.)

   - *a.* 1/25
   - *b.* 1/24
   - *c.* 1/48
   - *d.* 2/25

Your answer: a
Correct answer: a

The formula in this problem is 1/(1+(24/1)) or 1/25.

---

8. Which statement is true regarding the ROC curve?

   - *a.* The area under the ROC curve is reported as the *c* statistic in PROC LOGISTIC.
   - *b.* Oversampling affects the area under the ROC curve.
   - *c.* The area under the ROC curve is equivalent to the *D* statistic in the Kolmogorov-Smirnov test.
   - *d.* A perfectly random model would have a *c* statistic of 0.

Your answer: a
Correct answer: a

Answer *b* is incorrect because the area under the ROC curve is determined by the sensitivity and 1-specificity, and these statistics are not affected by oversampling.

Answer *c* is incorrect because the area under the ROC curve is equivalent to the Wilcoxon-Mann-Whitney test, not the K-S test.

Answer *d* is incorrect because a perfectly random model would have a *c* statistic of 0.50.

---

9. The results of the Kolmogorov-Smirnov test are shown below. What value represents the maximum vertical difference between the two cumulative distributions?

Partial PROC NPAR1WAY Output

Kolmogorov-Smirnov Test for Variable P_1 Classified by Variable TARGET_B			
TARGET_B	**N**	**EDF at Maximum**	**Deviation from Mean at Maximum**
0	7264	0.632985	3.708858
1	2421	0.458901	-6.424373
Total	9685	0.589468	
Maximum Deviation Occurred at Observation 5357			
Value of P_1 at Maximum = 0.051465			

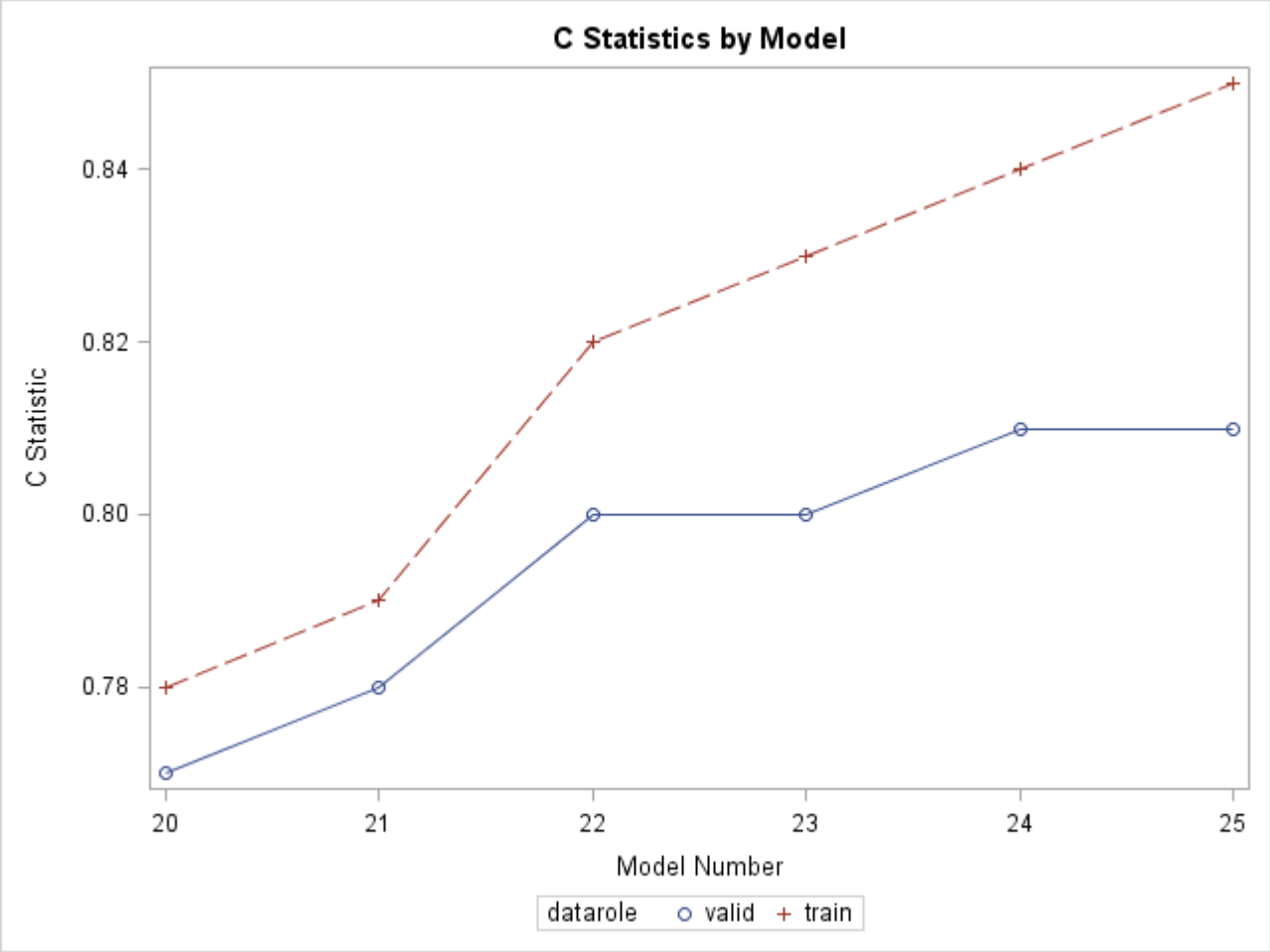
Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.075378	**D**	0.174083
KSa	7.418099	**Pr > KSa**	<.0001

- ○ *a.* 0.051465
- ○ *b.* 0.458901
- ○ *c.* 0.174083
- ○ *d.* 0.075378

Your answer: c
Correct answer: c

The maximum vertical difference between two cumulative distributions is represented by the *D* statistic, which is 0.174083 in this example.

10. The graph below shows the performance of several models on the training and validation data sets.



Which model had the highest shrinkage?

- ○ *a.* Model 20
- ○ *b.* Model 22
- ○ *c.* Model 24
- ○ *d.* Model 25

Your answer: d
Correct answer: d

Model 25 had the largest separation between the two curves on the graph.

Close