Elijah McLaughlin

December 1, 2023

STA 3064

Assignment 4

* 1. A screenshot of a number of text

     Description automatically generated  
     Spider = -2.8489 + 7.0418x
  2. Since 0.0052 < 0.05, reject null hypothesis. The GrainSize variable is providing valuable information in predicting the probability of Spider occurrence.
  3. A graph with a line

     Description automatically generated  
     Plot appears to be linear. Assumptions of linearity are met.
  4. Since 0.4068 > 0.05, fail to reject null hypothesis. The Hosmer-Lemeshow test did not find evidence of a lack of fit in the model.
  5. What I found in (c) supports (d) because the assumptions are met, and the Hosmer-Lemeshow test did not find lack-of-fit
  6. A screenshot of a graph

     Description automatically generated  
     P-value for AB = 0.3331, greater than 0.05, not statistically significant  
     A screenshot of a graph

     Description automatically generated  
     P-value for H2B = 0.0003, less than 0.05, is statistically significant  
     A table with numbers and text

     Description automatically generated  
     P-value for RBI = <0.0001, less than 0.05, is statistically significant  
     A screenshot of a graph

     Description automatically generated  
     P-value for BB = 0.0952, less than 0.05, is statistically significant
  7. RBI is the most effective, as both itself and its intercept's P-value are < 0.0001.
  8. AB is the least effective, as both itself and its intercept are not statistically significant.
  9. Odds-ratio = 1.706, for each unit increase in RBI, the odds of winning the baseball game increase by 1.706.
  10. 95% confidence limit = [1.420, 2.049], with 95% certainty for each unit increase in RBI, the odds of winning the baseball game will fall between 1.420 and 2.049.
  11. 2-unit odds-ratio = 2.910. For each two unit increase in RBI, the odds of winning the baseball game increase by 2.910.
  12. A table with text on it

      Description automatically generated
  13. 0.706 probability for a correct prediction, 0.545 probability for an incorrect prediction. I think the cutoff being 0.5 is alright. If we want a higher probability of a correct prediction, then must lower the cutoff. However, if we want a higher probability of avoiding false positives, then must raise the cutoff.
  14. Sensitivity starts at 100% correct guess rate when the cutoff is 0, however falls to 0% when the cutoff is 1. Conversely, specificity starts at 0% correct guess rate when the cutoff is 0, however rises to 100% when the cutoff is 1. I would use the 0.5 cutoff for total accuracy, as it is the most balanced of the bunch.
  15. A graph of a curve

      Description automatically generated  
      Since c is 0.759 the model seems to perform reasonably well.

1. M
   1. A table with numbers and a few black text

      Description automatically generated with medium confidence
   2. AB, H2B, RBI are statistically significant to the model in (a), with a P-value less than 0.05.
   3. The selection model suggests AB, H2B, RBI as the predictors.
   4. Since 0.1488 > 0.05, there is no need to for a non-linear model according to the Hosmer-Lemeshow test.
   5. A graph of a number of cases

      Description automatically generated with medium confidence  
      Row 46 Delta Deviance = 6.6671, Row 58 Delta Deviance = 4.5400. These could be potential outliers, as delta deviance suggests that removing these points would have significant impact on the model.
   6. A graph of a model

      Description automatically generatedA graph of a model

      Description automatically generated  
      *c* = 0.845, so the model works well according to ROC.