Homework

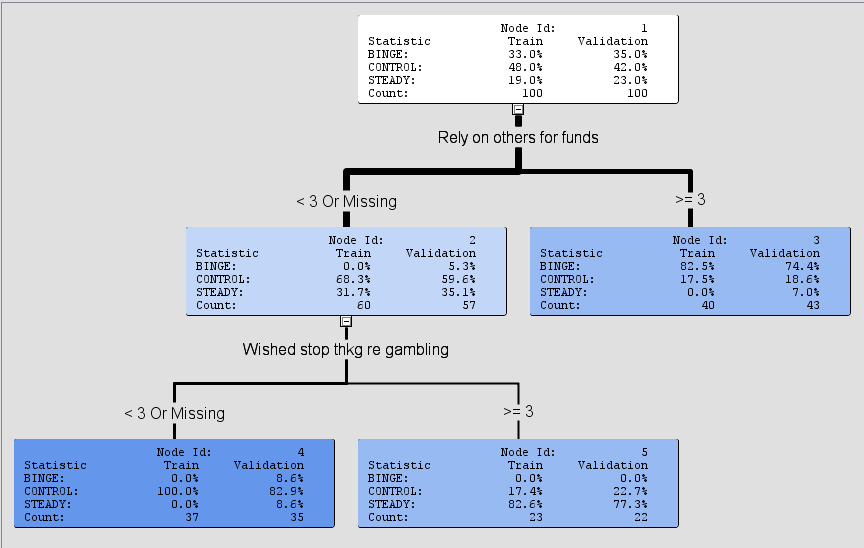
Suppose you work for an analytics consulting agency and have been hired by the Gamblers Anonymous group to evaluate their Gamblers Anonymous questionnaire in terms of its accuracy in diagnosing whether a person should be classified as a binge, control, or steady gambler. As you know, we have talked about one approach to doing this but have used the Diagnostic and Statistical Manual questionnaire to do so. For this one, let’s leave it open ended like this and see how creative people are in analyzing the data and reporting. Andrea and I are not “looking for” any particular details but I think we’ll be able to recognize a good analysis & report when we see it. No ground rules or designation of a particular technique on this one. Do what you think is best.

<http://support.sas.com/publishing/pubcat/chaps/55129.pdf>

Decision Trees

Combined DSM & GA

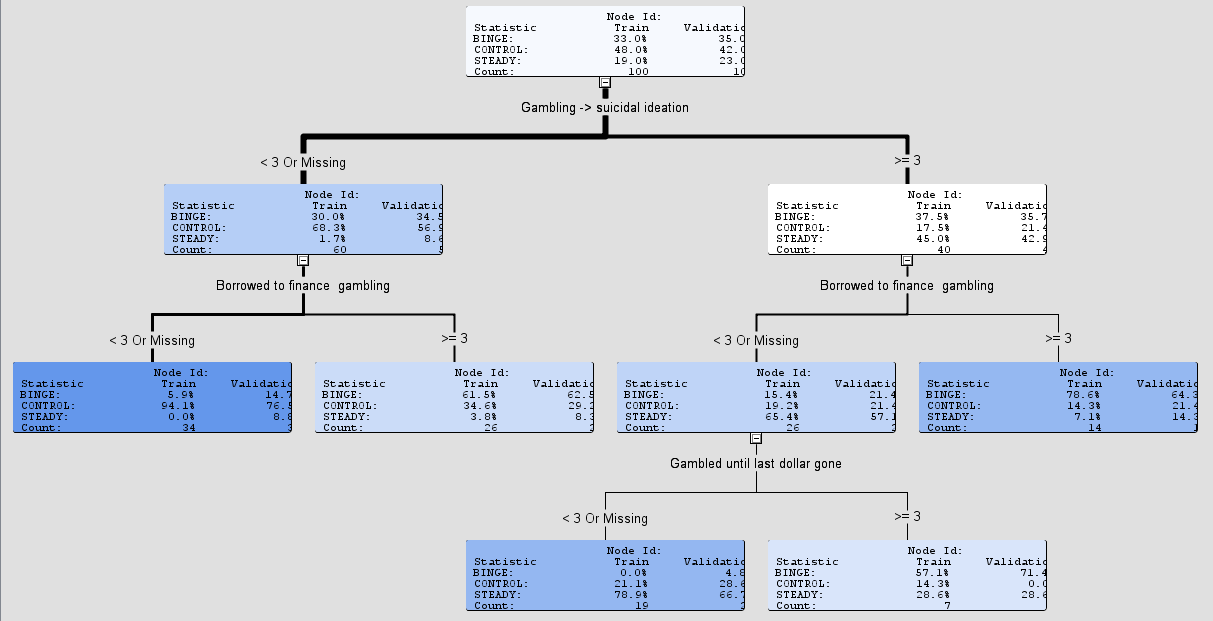
* Both questions came from DSM and were fairly accurate in predicting the 3 types.
* type \_MISC\_ Misclassification Rate 0.11 0.22 NaN



GA

3 questions

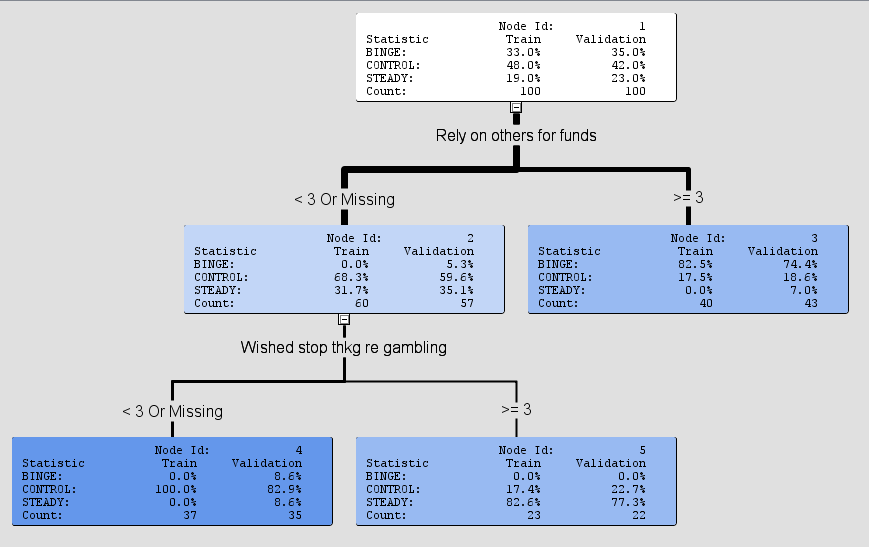
Type \_MISC\_Misclassification Rate 0.22 0.31 NaN



DSM

Same as combined

type \_MISC\_ Misclassification Rate 0.11 0.22 NaN



Discriminant Analysis

Need to rip through w/ GA

DSM

Full Model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Error Rate | Train | Validation | Train - Stepwise | Validation - Stepwise | Difference - Training | Diff - Validation |
| DSM | .04 | .1935 | .11 | .2607 |  |  |
| GA |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Training Data Set

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.GAMBLERS

Resubstitution Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into type** | | | | |
| --- | --- | --- | --- | --- |
| **From type** | **Binge** | **Control** | **Steady** | **Total** |
| **Binge** | |  | | --- | | 32 | | 96.97 | | |  | | --- | | 1 | | 3.03 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 33 | | 100.00 | |
| **Control** | |  | | --- | | 1 | | 2.08 | | |  | | --- | | 45 | | 93.75 | | |  | | --- | | 2 | | 4.17 | | |  | | --- | | 48 | | 100.00 | |
| **Steady** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 19 | | 100.00 | | |  | | --- | | 19 | | 100.00 | |
| **Total** | |  | | --- | | 33 | | 33.00 | | |  | | --- | | 46 | | 46.00 | | |  | | --- | | 21 | | 21.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.33 | |  | | |  | | --- | | 0.48 | |  | | |  | | --- | | 0.19 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for type** | | | | |
| --- | --- | --- | --- | --- |
|  | **Binge** | **Control** | **Steady** | **Total** |
| **Rate** | 0.0303 | 0.0625 | 0.0000 | 0.0400 |
| **Priors** | 0.3300 | 0.4800 | 0.1900 |  |

- Validation Data Set:

Classification Summary for Test Data: WORK.GAMBLE\_VALID

Classification Summary using Quadratic Discriminant Function

| **Observation Profile for Test Data** | |
| --- | --- |
| **Number of Observations Read** | 100 |
| **Number of Observations Used** | 100 |

| **Number of Observations and Percent Classified into type** | | | | |
| --- | --- | --- | --- | --- |
| **From type** | **Binge** | **Control** | **Steady** | **Total** |
| **Binge** | |  | | --- | | 26 | | 74.29 | | |  | | --- | | 9 | | 25.71 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 35 | | 100.00 | |
| **Control** | |  | | --- | | 2 | | 4.76 | | |  | | --- | | 39 | | 92.86 | | |  | | --- | | 1 | | 2.38 | | |  | | --- | | 42 | | 100.00 | |
| **Steady** | |  | | --- | | 1 | | 4.35 | | |  | | --- | | 8 | | 34.78 | | |  | | --- | | 14 | | 60.87 | | |  | | --- | | 23 | | 100.00 | |
| **Total** | |  | | --- | | 29 | | 29.00 | | |  | | --- | | 56 | | 56.00 | | |  | | --- | | 15 | | 15.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.33 | |  | | |  | | --- | | 0.48 | |  | | |  | | --- | | 0.19 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for type** | | | | |
| --- | --- | --- | --- | --- |
|  | **Binge** | **Control** | **Steady** | **Total** |
| **Rate** | 0.2571 | 0.0714 | 0.3913 | 0.1935 |
| **Priors** | 0.3300 | 0.4800 | 0.1900 |  |

Stepwise

Training Dataset

|  |
| --- |
| Gamblers: stepwise selection |
| validation of stepwise discriminant functions |

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.GAMBLERS

Resubstitution Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into type** | | | | |
| --- | --- | --- | --- | --- |
| **From type** | **Binge** | **Control** | **Steady** | **Total** |
| **Binge** | |  | | --- | | 32 | | 96.97 | | |  | | --- | | 1 | | 3.03 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 33 | | 100.00 | |
| **Control** | |  | | --- | | 7 | | 14.58 | | |  | | --- | | 39 | | 81.25 | | |  | | --- | | 2 | | 4.17 | | |  | | --- | | 48 | | 100.00 | |
| **Steady** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 5.26 | | |  | | --- | | 18 | | 94.74 | | |  | | --- | | 19 | | 100.00 | |
| **Total** | |  | | --- | | 39 | | 39.00 | | |  | | --- | | 41 | | 41.00 | | |  | | --- | | 20 | | 20.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.33 | |  | | |  | | --- | | 0.48 | |  | | |  | | --- | | 0.19 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for type** | | | | |
| --- | --- | --- | --- | --- |
|  | **Binge** | **Control** | **Steady** | **Total** |
| **Rate** | 0.0303 | 0.1875 | 0.0526 | 0.1100 |
| **Priors** | 0.3300 | 0.4800 | 0.1900 |  |

|  |
| --- |
| Gamblers: stepwise selection |
| validation of stepwise discriminant functions |

The DISCRIM Procedure

Classification Summary for Test Data: WORK.GAMBLE\_VALID

Classification Summary using Quadratic Discriminant Function

| **Observation Profile for Test Data** | |
| --- | --- |
| **Number of Observations Read** | 100 |
| **Number of Observations Used** | 100 |

| **Number of Observations and Percent Classified into type** | | | | |
| --- | --- | --- | --- | --- |
| **From type** | **Binge** | **Control** | **Steady** | **Total** |
| **Binge** | |  | | --- | | 31 | | 88.57 | | |  | | --- | | 4 | | 11.43 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 35 | | 100.00 | |
| **Control** | |  | | --- | | 7 | | 16.67 | | |  | | --- | | 33 | | 78.57 | | |  | | --- | | 2 | | 4.76 | | |  | | --- | | 42 | | 100.00 | |
| **Steady** | |  | | --- | | 3 | | 13.04 | | |  | | --- | | 5 | | 21.74 | | |  | | --- | | 15 | | 65.22 | | |  | | --- | | 23 | | 100.00 | |
| **Total** | |  | | --- | | 41 | | 41.00 | | |  | | --- | | 42 | | 42.00 | | |  | | --- | | 17 | | 17.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.33 | |  | | |  | | --- | | 0.48 | |  | | |  | | --- | | 0.19 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for type** | | | | |
| --- | --- | --- | --- | --- |
|  | **Binge** | **Control** | **Steady** | **Total** |
| **Rate** | 0.1143 | 0.2143 | 0.3478 | 0.2067 |
| **Priors** | 0.3300 | 0.4800 | 0.1900 |  |

| **Stepwise Selection Summary** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Number In** | **Entered** | **Removed** | **Label** | **Partial R-Square** | **F Value** | **Pr > F** | **Wilks' Lambda** | **Pr < Lambda** | **Average Squared Canonical Correlation** | **Pr > ASCC** |
| **1** | **1** | **dsm4** |  | Rely on others for funds | 0.5883 | 69.30 | <.0001 | 0.41169856 | <.0001 | 0.29415072 | <.0001 |
| **2** | **2** | **dsm1** |  | Wished stop thkg re gambling | 0.3443 | 25.21 | <.0001 | 0.26993945 | <.0001 | 0.46288291 | <.0001 |
| **3** | **3** | **dsm8** |  | Win back money next day | 0.0720 | 3.68 | 0.0288 | 0.25050786 | <.0001 | 0.48665727 | <.0001 |
| **4** | **4** | **dsm2** |  | Wished stop thkg re get money | 0.0405 | 1.98 | 0.1436 | 0.24037460 | <.0001 | 0.49606842 | <.0001 |

DSM

4,1,8,2

* RERUN ON GA and on Validation for Both

Logistic Regression

GA

6,9,10,20 via Stepwise and Forward

DSM

4,1 via Stepwise and Forward

Need to run on validation data set – see method from Log Reg Final Project

Need to add interactions

Future Recommendations

What are the costs of getting a prediction wrong? (i.e. Doctor’s making incorrect assessment on Cancer patients)

Questions

Principal Component Analysis

* interpretation (for negative numbers)

Discriminant Analysis

* How to score new data?
* How do you use the Linear Discriminant Function for Type table?
  + NOTE: Need to run for both GA and GSM (priors prop)