10/12/2013

Orange Team 6:

*Team Lead:* Phillip Domschke

*Other Contributors:* Jake Frost, Wesley Ledebuhr, Steve Neola, and Marc Zimmerman

**GAMBLERS ANONYMOUS QUESTIONNAIRE ANALYSIS**

# TABLE OF CONTENTS

TABLE OF CONTENTS 1

EXECUTIVE SUMMARY 2

Future Considerations 2

ANALYSIS 3

Decision Trees 3

Logistic Regression 3

Discriminant Analysis 4

Principal Components & Factor Analysis 4

CONCLUSION 5

Appendix A: Questionnaires 6

Appendix B: Decision Trees 7

# 

# EXECUTIVE SUMMARY

We were asked by the Steering Committee of Gamblers Anonymous (GA) to analyze their questionnaire in terms of its predictive capabilities of classifying participants into one of three types of gamblers: Control, Steady, or Binge. To solve this problem, we applied multiple analytical techniques to the Gamblers Anonymous questionnaire and corresponding data as well as to the Diagnostic and Statistical Manual (DSM) questionnaire. We also combined the data from both questionnaires to see how they compared to each other and if better results could be obtained with the aggregated data.

We found that the number of questions on both questionnaires could be significantly reduced while retaining the majority of the predictive capabilities of the questionnaire. The table below identifies which questions should be retained based on the results of each analytical method. (See Appendix A for a list of the questions.)

|  |  |  |  |
| --- | --- | --- | --- |
| Analytical Method | DSM | GA | Combined |
| Decision Trees | 1, 4 | 9, 10, 20 | Same as DSM |
| Logistic Regression | 1, 4 | 6, 9, 10, 20 | DSM 1,4, GA 20 |
| Discriminant Analysis | 1, 2, 4, 8 | 6, 9, 10, 20 | DSM 1,4,8 GA 10,20 |

Interestingly, there is some overlap in the nature of the questions between the final GA and DSM questions. For example, DSM 4 - *Rely on others for funds* is very similar to GA 10 - *Borrowed to finance gambling.*  Likewise, DSM 8 - *Win back money next day* is very similar to GA 9 - *Gambled until last dollar gone*. This qualitative comparison adds validity to the quantitative findings.

To assess the predictive power of the questions, we used Principal Component Analysis and Factor Analysis to check the clusters of results and found that there are, in fact, logical groupings that map to the three types of gamblers.

In the body of the report, we will walk through the methodology and results for each analytical technique we applied for both questionnaires. We will then compare the techniques and discuss why each method was chosen.

### Future Considerations

Prior to reducing the number of survey questions, it will be important to understand the costs of getting a prediction wrong for the Gamblers Anonymous organization as well as for the clients. For example, how do cost and duration of the treatment plans differ based on the classification of a gambler? Would there be any repercussions (i.e. distrust, lawsuits) to the organization or any of its employees if a client was misclassified? Lastly, there must be a shared understanding of what an acceptable level of misclassification prior to any questions being removed.

Once the costs of a misclassification are understood, the costs of respondents not completing the survey due to the current length need to be investigated. For example, how many potential clients are not completing the survey or are racing through it due to the number of questions?

With this combined cost information, we could find the optimal balance between keeping a minimal number of questions while maintaining a high degree of predictive power.

# ANALYSIS

## Decision Trees

The decision trees were able to reduce the questionnaire to a small number of questions as shown in the table below. Without doing any pruning or modifications to the classification settings, the trees produced very actionable results.

|  |  |  |  |
| --- | --- | --- | --- |
| Analytical Method | DSM | GA | Combined |
| Decision Trees | 1, 4 | 9, 10, 20 | Same as DSM |

Decision Trees proved to be a very effective method for reducing the number of questionnaires without losing a lot of predictive power. As shown in the table below, the Combined and DSM performed the best. The reason they have the same results is because the Combined decision tree only used chose questions from the DSM questionnaire to split on.

|  |  |  |
| --- | --- | --- |
| Decision Tree | Misclassification - Training | Misclassification - Validation |
| Combined | 0.11 | 0.22 |
| DSM | 0.11 | 0.22 |
| GA | 0.22 | 0.31 |

See Appendix B: Decision Trees for the actual decision trees.

## Logistic Regression

We applied Forward and Stepwise variable selection techniques to determine which questions were significant predictors of the gambler classifications. Compared to Decision Trees, Logistic Regression only added GA question 6 when running GA by itself and added GA question 20 when running the combined data set. The results are shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Analytical Method | DSM | GA | Combined |
| Logistic Regression | 1, 4 | 6, 9, 10, 20 | DSM 1,4, GA 20 |

We also applied the Backward selection technique and found multiple two-way interactions between the questions. However, since all of the other tests across all the other methods did not include these, we decided to leave them out of the results.

Lastly, we were hoping that we could treat the classifications of gamblers as ordinal variables to improve the interpretability of the results. After running the SCORE Test for Proportional Odds, we found that we needed to treat the classifications as nominal inputs. This limitation would only allow us to compare the odds that a client is more likely to be in a category versus the control group given they are not in the third group (as opposed to the odds of being in that category compared to all the groups).

## Discriminant Analysis

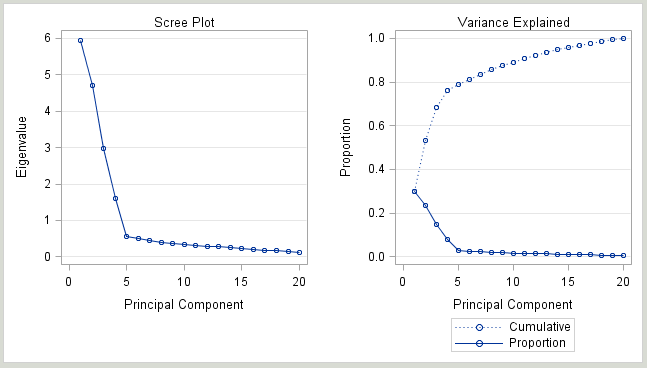
Disciminant Analysis produced a slightly larger number of questions. In comparison to Decision Trees and Logistic Regression, Discriminant Analysis added question 2 and 8 from the DSM and question 6 from GA. It also included a wider set of questions in the combined set. The results are shown in the table below.

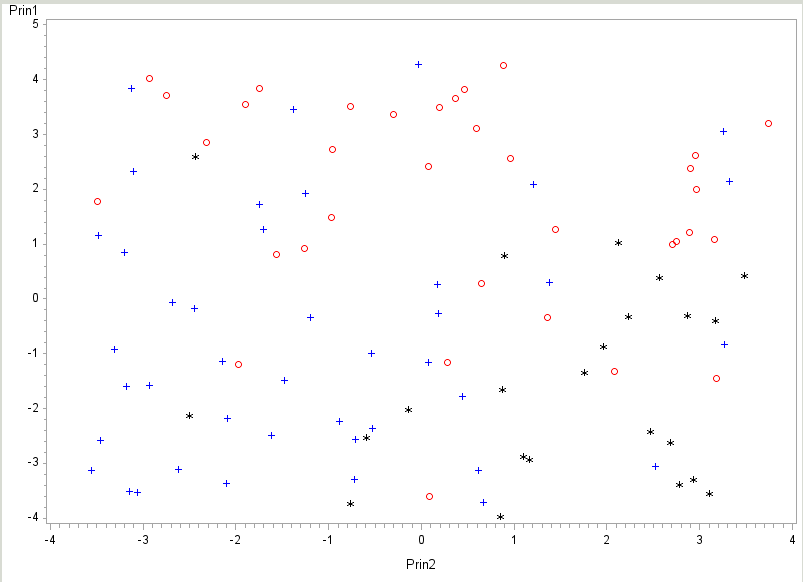
|  |  |  |  |
| --- | --- | --- | --- |
| Analytical Method | DSM | GA | Combined |
| Discriminant Analysis | 1, 2, 4, 8 | 6, 9, 10, 20 | DSM 1,4,8 GA 10,20 |

When using the Stepwise variable selection method, the model performed fairly well as the error rate only increased by .13 between the training and the validation GA data sets. These missclassification rates are also very similar to the Decision Tree Results – a further validation of the results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Questionnaire | Train – All Variables | Validation - All Variables | Train - Stepwise | Validation – Stepwise | Difference |
| DSM | 0.04 | 0.19 | 0.11 | 0.26 | 0.16 |
| GA | 0.17 | 0.29 | 0.16 | 0.29 | 0.13 |

## Principal Components & Factor Analysis

The first three principal components in the training data set and the first four in the validation data set explained over 70% of the variance in the data (as shown below in the validation data set Scree and Variance Explained Plots).

In the principal component graph below, there are three clear clusters. The eigenvectors confirm these findings as there are multiple questions that not only predicted a lot of the variation but are logically themed together based on the three types of gamblers.

We also ran factor analysis on the significant questions from the GA data set and had similar results.

# CONCLUSION

To determine which technique is best for predicting which questions classify the client into one of the gambling categories, we compared the misclassification rates of the models in aggregate and by each type of category. In aggregate, both Decision Trees and Discriminant Analysis showed very similar misclassification rates on the validation data sets. Logistic Regression validated the question selection as well in terms of picking which questions were significant predictors. Lastly, Principal Component Analysis was helpful to validate that the three categories of gamblers do, in fact, show up in the data in very distinct ways. If you could only pick one type of analysis, we recommend discriminant analysis due to the accuracy of predicting each category (see table below).

|  |  |  |  |
| --- | --- | --- | --- |
| Error Rates – GA | Control | Steady | Binge |
| Decision Trees | 23.5% | 33.3% | 35.7% |
| Logistic Regression | 34.2% | 23.8% | 43.5% |
| Discriminant Analysis | 6.6% | 13.7% | 8.2% |
| Factor Analysis | 34.88 | 21.43 | 25.71 |

# Appendix A: Questionnaires

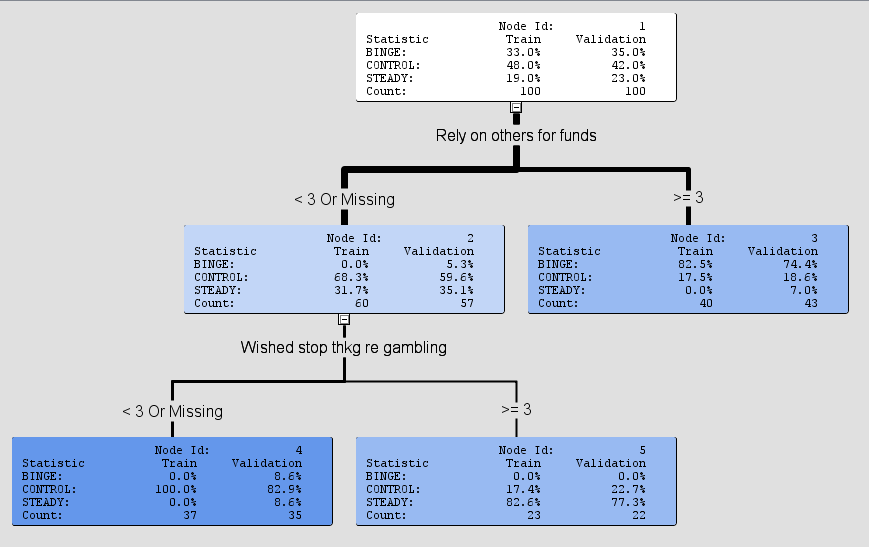
The table below identifies the questions that are included in each organization’s questionnaire.

|  |  |
| --- | --- |
| ID | Questions |
| dsm1 | Wished stop thkg re gambling |
| dsm2 | Wished stop thkg re get money |
| dsm3 | Felt need to bet more and more |
| dsm4 | Rely on others for funds |
| dsm5 | Gamble to escape |
| dsm6 | Lie about how much I gamble |
| dsm7 | Relaxing difficult if not gambling |
| dsm8 | Win back money next day |
| dsm9 | Felt I should cut back on gambling |
| dsm10 | Illegal acts to pay for gambling |
| dsm11 | Danger of losing relationship |
| dsm12 | Danger of losing job |
| ga1 | Lost time from work from gambling |
| ga2 | Gambling made home life unhappy |
| ga3 | Gambling affected reputation |
| ga4 | Felt remorse after gambling |
| ga5 | Gamble to get money for debts |
| ga6 | Caused decreased ambition/efficiency |
| ga7 | Felt must return win back losses |
| ga8 | After win want to return win more |
| ga9 | Gambled until last dollar gone |
| ga10 | Borrowed to finance gambling |
| ga11 | Sold things to finance gambling |
| ga12 | Kept gambling money for gambling |
| ga13 | Gambling->Careless of self/family |
| ga14 | Gambled longer than planned |
| ga15 | Gambled to escape worry/trouble |
| ga16 | Illegal act to finance gambling |
| ga17 | Gambling caused difficulty sleeping |
| ga18 | Arguments, frustration -> gambling |
| ga19 | Good fortune-> gambling |
| ga20 | Gambling -> suicidal ideation |

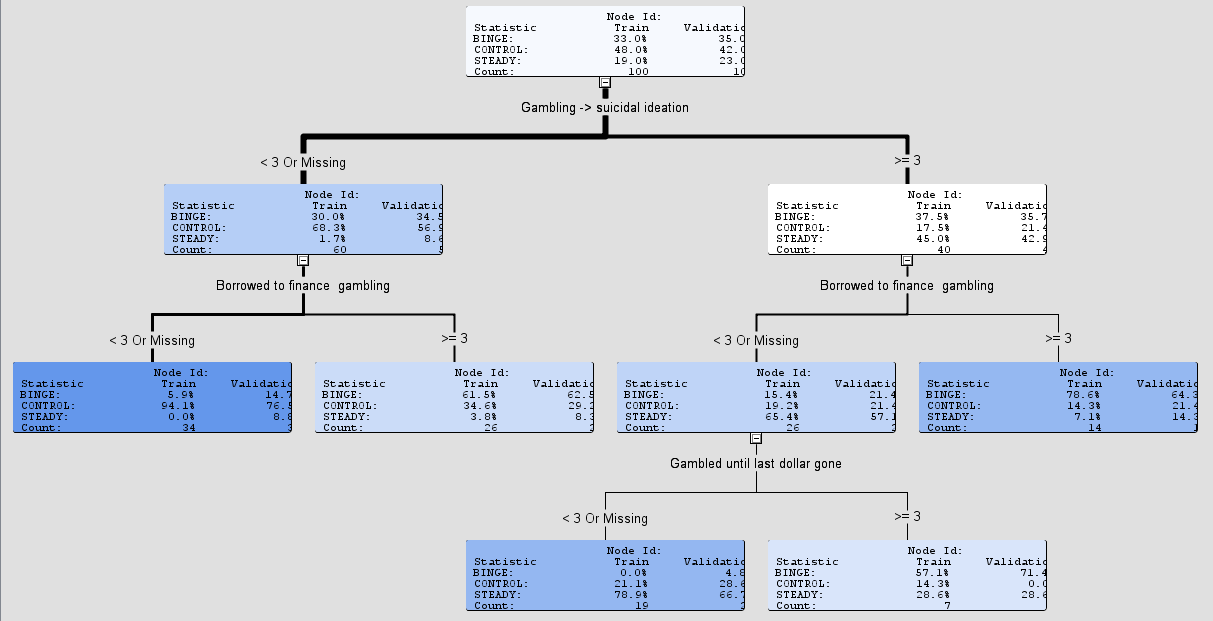
# Appendix B: Decision Trees

The decision tree outputs are shown below.

**DSM**



**GA**



**Combined**

