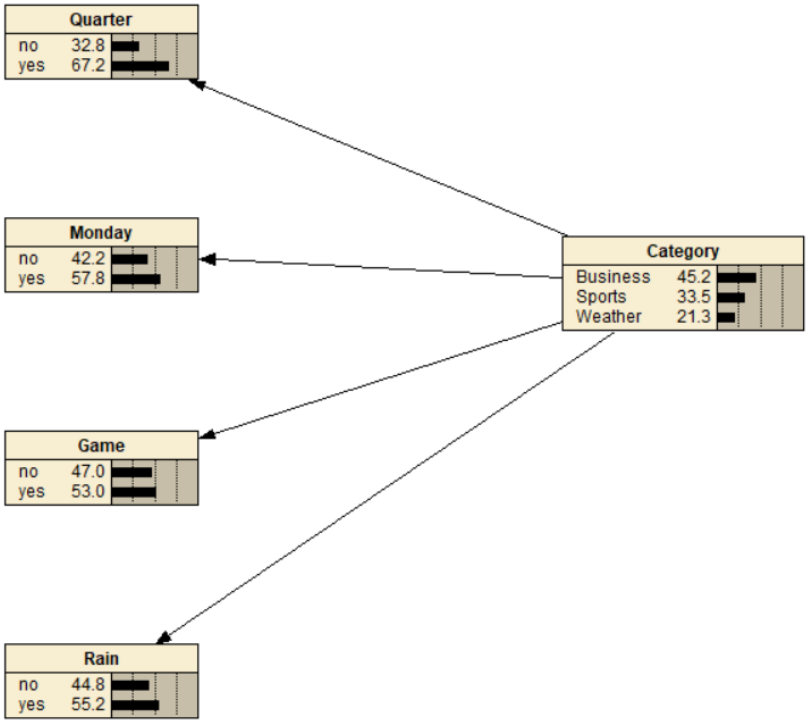
Problem 3 – Netica – Naïve Bayes

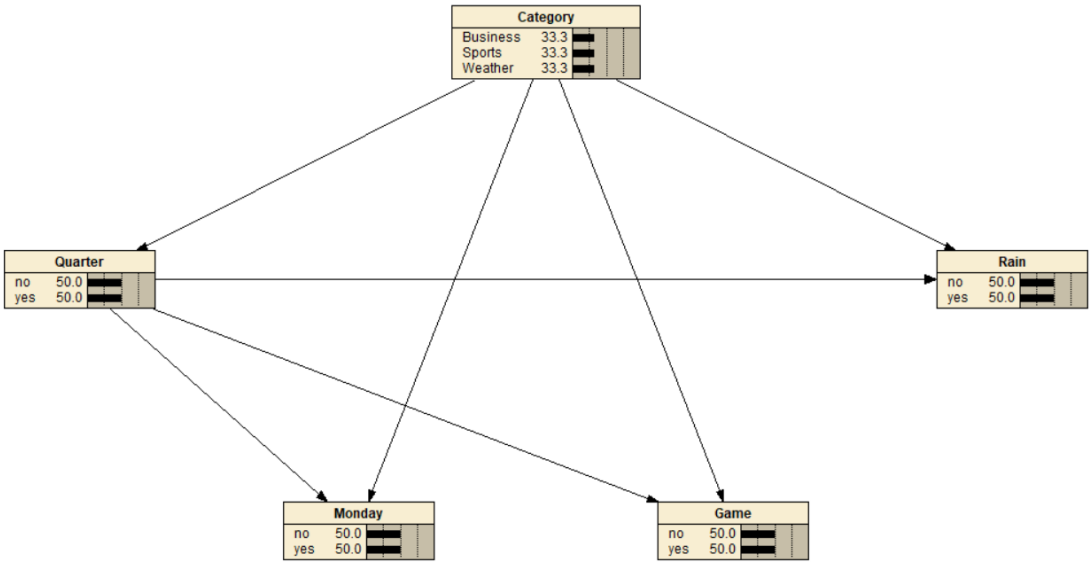
Results – Test set

Naïve Bayes Network – Newspaper Categorization articles based on words: Quarter, Monday, Game, Rain



```
Read 199 cases, and used 199 of them to test net.

For Category:
-------------

Confusion:
  .......Predicted......
  Busine  Sports  Weathe    Actual
  ------  ------  ------    ------
      69      19       0    Business
      15      62       0    Sports
      27       7       0    Weather

Error rate = 34.17%

Scoring Rule Results:
  Logarithmic loss = 0.8065
  Quadratic loss   = 0.4723
  Spherical payoff = 0.7223

Calibration:
  Business   0-25:     13    | 25-40:     40    | 40-60:     49.2 | 60-75:     78.9 |
  Sports     0-10:     6.9   | 10-25:     16.7  | 25-50:     35.9 | 50-70:     69.2 | 70-100:    90    |
  Weather    0-15:     5.56  | 15-25:     10.1  | 25-40:     39.5 | 40-100:    40    |
  Total      0-10:     6.74  | 10-15:     9.21  | 15-20:     5.56 | 20-25:     14.9 | 25-40:     37.1 | 40-50:     43.5 | 50-60:

Times Surprised (percentage):
               .................Predicted Probability...................
  State        < 1%             < 10%             > 90%             > 99%
  -----        ----             -----             -----             -----
  Business     0.00 (0/0)       18.18 (2/11)      0.00 (0/0)        0.00 (0/0)
  Sports       0.00 (0/0)        6.90 (4/58)      0.00 (0/0)        0.00 (0/0)
  Weather      0.00 (0/0)        0.00 (0/20)      0.00 (0/0)        0.00 (0/0)
  Total        0.00 (0/0)        6.74 (6/89)      0.00 (0/0)        0.00 (0/0)

Sensitivity of Test:
  Business   0     100    | 40    70.5   | 60    34.1   | 75    0      | 100   0      |
  Sports     0     100    | 20    89.6   | 50    70.1   | 70    23.4   | 80    11.7   | 85    0      | 100   0      |
  Weather    0     100    | 25    67.6   | 40    17.6   | 100   0      |

------------------------------------------------------------------------------
```

Tree Augmented Bayes Model - Newspaper Categorization articles based on words: Quarter, Monday, Game, Rain



```
Learning TAN structure to classify 'Category' based on nodes: Quarter, Monday, Game, Rain

Used 800 cases to learn the TAN structure of 5 nodes.
Case file to learn CPTs from:  D:\Documents\UMD_Docs\Grad-school\ENPM808Y\Assignments\HW-8\TextClassTrain.xlsx
Enter degree (normal is 1):  1

Correspondance between database columns and nodes of the BN during Caseset generation:
Database    =    Bayes net
--------         ---------
Category         Category

Used 800 cases to modify the CPT tables of 1 nodes.

**2334** No nodes selected (or in node-set named 'target') to run the tests for.
Case file for testing:  D:\Documents\UMD_Docs\Grad-school\ENPM808Y\Assignments\HW-8\TextClassTest.xlsx

Correspondance between database columns and nodes of the BN during Caseset generation:
Database    =    Bayes net
--------         ---------
Quarter          Quarter
Monday           Monday
Game             Game
Rain             Rain
Category         Category


**2448** Can't test Bayes net 'Untitled_1', because it hasn't been recently compiled.
Compiled to 3 cliques, with total table size (including sepsets) of 48.

**2760** Some node(s) (e.g. Quarter) don't have conditional probability tables (CPTs) (they will be taken as having uniform probabilities).
Case file for testing:  D:\Documents\UMD_Docs\Grad-school\ENPM808Y\Assignments\HW-8\TextClassTest.xlsx

Correspondance between database columns and nodes of the BN during Caseset generation:
Database    =    Bayes net
--------         ---------
Category         Category
Quarter          Quarter
Monday           Monday
Game             Game
Rain             Rain

Read 199 cases, and used 199 of them to test net.


For Category:
-------------

Confusion:
  .......Predicted......
  Busine  Sports  Weathe    Actual
  ------  ------  ------    ------
      88       0       0    Business
      77       0       0    Sports
      34       0       0    Weather

Error rate = 55.78%

Scoring Rule Results:
  Logarithmic loss = 1.039
  Quadratic loss   = 0.6301
  Spherical payoff = 0.6082

Calibration:
  Business   0-50:     44.2 |
  Sports     0-40:     38.7 |
  Weather    0-25:     17.1 |
  Total      0-25:     17.1 | 25-40:    38.7 | 40-50:    44.2 |

Times Surprised (percentage):
              .................Predicted Probability.....................
  State       < 1%             < 10%             > 90%             > 99%
  -----       ----             -----             -----             -----
  Business    0.00 (0/0)       0.00 (0/0)        0.00 (0/0)        0.00 (0/0)
  Sports      0.00 (0/0)       0.00 (0/0)        0.00 (0/0)        0.00 (0/0)
  Weather     0.00 (0/0)       0.00 (0/0)        0.00 (0/0)        0.00 (0/0)
  Total       0.00 (0/0)       0.00 (0/0)        0.00 (0/0)        0.00 (0/0)

Sensitivity of Test:
  Business   0     100    | 50    0      | 100   0      |
  Sports     0     100    | 40    0      | 100   0      |
  Weather    0     100    | 25    0      | 100   0      |

--------------------------------------------------------------------------------
```

C) comparison between Naïve Bayes and TAN model – error rate

The error rate in Naïve Bayes was about 34.17%, while TAN yielded 55.78%. The Naïve Bayes model had a simplistic connection between the effect (dependent) nodes and the target node, i.e. only a single link between cause and effect. For TAN, however, the link network was a bit complex where the effects depended on each other besides depending on the target node (classifier). This, as a result, without detailed specification can lead the classifier to incorrectly label the article, which is what's evident on the test data.