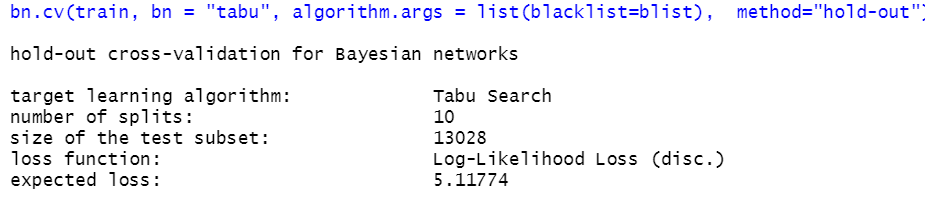
3-)

Five different methods (Hill Climbing, Tabu Search, Grow Shrink, Incremental Association and Max-Min Hill Climbing) which are scored-based, scored-based, constraint-based, constraint-based and hybrid algorithms respectively, are used to learn structure of BN with a given train data.

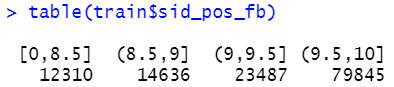
After whitelist and blacklist created, all algorithms make the same DAG and score. Then, cross-validation is made with two different methods (k-fold and hold-out). Tabu search algorithm expects minimum loss on both methods against other methods. Boosttrapping is also give the same DAG with tabu search. Thus, tabu search algorithm is used for the rest of the study.

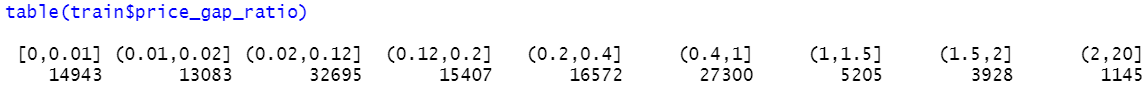


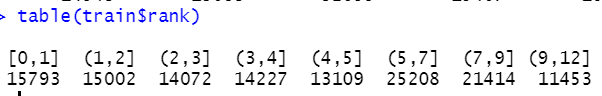
|  |  |  |  |
| --- | --- | --- | --- |
| BN | Algorithm | Method | Expected Loss |
| hc | Score-Based | k-fold | 5,135 |
| Tabu | Score-Based | k-fold | 5,117 |
| gs | Constrain-Based | k-fold | 5,150 |
| iamb | Constrain-Based | k-fold | 5,150 |
| mmhc | Hybrid | k-fold | 5,135 |
| hc | Score-Based | hold-out | 5,131 |
| Tabu | Score-Based | hold-out | 5,116 |
| gs | Constrain-Based | hold-out | 5,150 |
| iamb | Constrain-Based | hold-out | 5,145 |
| mmhc | Hybrid | hold-out | 5,130 |

"is\_Amazon", "price\_gap\_ratio", "sid\_pos\_fb", "rank" and "bbox" parameters are used to learn DAG. "price\_gap\_ratio", and "sid\_pos\_fb" parameters does not discrete and it is known that continues nodes cannot be parent of discrete nodes. Therefore, our model tends to be misinterpreted. To prevent this problem these two parameters transformed to discrete by making intervals so that it can be treated as factor.

While converting continues parameter to discrete intervals, two factor is considered. The first one is expected importance of value. For instance, customer’s decision may not change if the “sid\_pos\_rate” is 9.9 or 9.8. Therefore, we start with 0.5 interval but end with 0:7 interval since it does not make importance once it goes below 7. The second factor is almost equally and large enough sized interval. This part is important. If there is no data on train corresponding to test observation, algorithm cannot predict that test observation. To solve that interval length should be increased. However, to increase the effect of parameter interval length should be decreased. Here, it is tried to find optimum interval values which maximize the prediction accuracy.







|  |  |  |
| --- | --- | --- |
| price\_gap\_ratio | rank | sid\_pos\_fb |
| 0.00 - 0.01 | 0 | 0.0 - 8.0 |
| 0.01 - 0.02 | 1 | 8.5 - 9.0 |
| 0.02 - 0.12 | 2 | 9.0 - 9.5 |
| 0.12 - 0.20 | 3 | 9.5 - 10.0 |
| 0.20 - 0.40 | 4 |  |
| 0.40 - 1.00 | 5 |  |
| 1.00 - 1.50 | 6 - 7 |  |
| 1.50 - 2.00 | 7 - 8 |  |
| 2.00 - 20.00 | 9 -12 |  |

First DAG does not seem good model since we know some information between parameters. Therefore, whitelist and blacklist are used to force wanted arc and prevent unwanted arcs between parameters.

We are making whitelist and blacklist with the highest possible and logical information between nodes since learnt DAG is heuristic depends on whitelist and blacklist approval order. We start with whitelist since it is superior of blacklist. We expect all of chosen parameters should affect bbox. Therefore, we make whitelist for these and then continue to for the rest. The final whitelist and blacklist can be shown in below table.

|  |  |  |  |
| --- | --- | --- | --- |
| Whitelist | | Blacklist | |
| from | to | from | to |
| price\_gap\_ratio | bbox | is\_Amazon | price\_gap\_ratio |
| is\_Amazon | bbox | price\_gap\_ratio | is\_Amazon |
| rank | bbox | price\_gap\_ratio | sid\_pos\_fb |
| sid\_pos\_fb | bbox | sid\_pos\_fb | price\_gap\_ratio |
| is\_Amazon | sid\_pos\_fb | rank | price\_gap\_ratio |
|  |  | rank | sid\_pos\_fb |
|  |  | sid\_pos\_fb | rank |
|  |  | rank | is\_Amazon |

4 -) a-)

There are 2\*2\*9\*9\*4 parameters of model that is learnt by using training date. Since there are many parameters, only few of them are shown in the below section.

is\_Amazon

bbox 0 1

failure 0.9959866221

success 0.0040133779

, , price\_gap\_ratio = (0.4,1], rank = (2,3], sid\_pos\_fb = (9,9.5]

is\_Amazon

bbox 0 1

failure 0.9969040248

success 0.0030959752

, , price\_gap\_ratio = (1,1.5], rank = (2,3], sid\_pos\_fb = (9,9.5]

Parameters of node price\_gap\_ratio (multinomial distribution)

Conditional probability table:

, , price\_gap\_ratio = [0,0.01]

is\_Amazon

rank 0 1

[0,1] 0.365258462 0.935327635

(1,2] 0.188664393 0.006837607

(2,3] 0.121315490 0.023361823

(3,4] 0.116767253 0.001709402

(4,5] 0.059302020 0.015099715

(5,7] 0.057990029 0.017663818

(7,9] 0.060788944 0.000000000

(9,12] 0.029913409 0.000000000

, , price\_gap\_ratio = (0.01,0.02]

is\_Amazon

rank 0 1

[0,1] 0.077611940 0.005665722

(1,2] 0.101256874 0.056657224

(2,3] 0.107541241 0.314447592

(3,4] 0.097486253 0.235127479

(4,5] 0.103534957 0.198300283

(5,7] 0.207227023 0.184135977

(7,9] 0.207069914 0.005665722

(9,12] 0.098271799 0.000000000

Parameters of node sid\_pos\_fb (multinomial distribution)

Conditional probability table:

is\_Amazon

sid\_pos\_fb 0 1

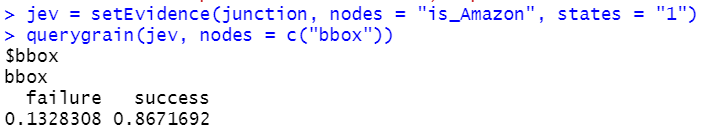
[0,8.5] 0.0989359 0.0000000

(8.5,9] 0.1176300 0.0000000

(9,9.5] 0.1887658 0.0000000

(9.5,10] 0.5946682 1.0000000

b-) The success probability of winning bbox of amazon as a saler P( bbox=success | sid=amazon ) is calculated as 86.7% with below function:



c-) Firstly, prediction model with the chosen dag and parameters ("is\_Amazon", "price\_gap\_ratio", "sid\_pos\_fb" and "rank") is created. Then, prediction made with training data to see the accuracy level of model and it is calculated as 98.3%. Once test result are uploaded, prediction made with test data on the same prediction model similarly. The accuracy of test model predictions calculated as 93.85%. Confusion matrix of both train and test data prediction results are shown in the below:

d-)

