**DSCI 5240 Assignment 6**

The following decision tree was built based on the data. The target variable has two values “Not default” (0) or “Default” (1).

No. of observations:

Node 1: 300

Node 2: 270

Node 3: 30

Node 4: 240

Node 5: 30

Node 1: DELINQ

0: 80%

1: 20%

>=1.5

<1.5

Node 2: DEROG

0: 84%

1: 16%

>=0.5

<0.5

1. How many leaves does the tree have? Please explain to the management the prediction of Node 4. Which type of applicants have the highest chance to default? (10 points)
2. In decision trees, a leaf is an end node that does not split further. From the above decision tree, there are several nodes, and we need to identify those that do not further split into additional nodes.

Node 1 splits into Node 2 and Node 3.

Node 2 splits into further nodes, while Node 3 does not.

Node 4 does not split further, and neither does Node 5.

From this, we can deduce that the decision tree has 3 leaves (Node 3, Node 4, and Node 5)

1. Node 4 represents applicants with low delinquent credit lines (DELINQ < 1.5) and low major bad credit signs (DEROG < 0.5). In this node, 87% of applicants are predicted to "Not default" on their loans, while 13% are predicted to "Default" on their loans. This means that applicants reaching Node 4 have a very high probability of not defaulting on their loans.

For management, this could be interpreted that the risk associated with these applicants is relatively low.

1. Node 5 represents applicants with low delinquent credit lines (DELINQ < 1.5) and high major bad credit signs (DEROG >= 0.5). In this node, 80% of applicants are predicted to "Default" on their loans, while only 20% are predicted to "Not default" on their loans. This would suggest that applicants routed to Node 5 are at the highest risk of defaulting.
2. An applicant is applying for a new loan. He has 2 delinquent credit lines. Will you, as a bank loan officer, approve his application? Please explain. (10 points)

With 2 delinquent credit lines, the decision tree indicates that this applicant would progress to a node were having delinquent credit lines >= 1.5. The applicant would land on Node 3 where 41% of applicants are predicted to "Not default" on their loans and 59% of applicants are predicted to "Default" on their loans. As the applicant is directed towards a higher risk node, the bank loan officer has a higher probability of rejecting the application.

However, based solely on the high number of delinquent lines, the risk of default appears elevated. A cautious approach would suggest a review of additional information or possibly not approving the loan without further risk mitigation.

1. Calculate the information gain from Node 1 to Node 2 and Node 3 based on entropy. (10 points)

Step 1: Calculate the Entropy

**Node 1**

Total observations: 300

Number of "Not default" (0) observations: 240

Number of "Default" (1) observations: 60

Entropy (Node 1) = - (240/300) \* log2(240/300) - (60/300) \* log2(60/300)

= - (0.8) (-0.3219) – (0.2) (-2.322)

= 0.25752 + 0.4644

= 0.72192

**Node 2**

Total observations: 270

Number of "Not default" (0) observations: 227

Number of "Default" (1) observations: 43

Entropy (Node 2) = - (227/270) \* log2(227/270) - (43/270) \* log2(43/270)

= - (0.8407) (-0.25034) – (0.1592) (-2.651)

= 0.21046 + 0.42203

= 0.63249

**Node 3**

Total observations: 30

Number of "Not default" (0) observations: 12

Number of "Default" (1) observations: 18

Entropy (Node 3) = - (12/30) \* log2(12/30) - (18/30) \* log2(18/30)

= - (0.4) (-1.322) – (0.6) (-0.737)

= 0.5288 + 0.4422

= 0.971

Step 2: Calculate the Weighted Average Entropy of the Split

Entropy (Node 2) = 0.63249

Entropy (Node 3) = 0.971

**Node 2**

Weighted Entropy = (270/300) \* Entropy (Node2)

= 0.9 \* 0.63249

= 0.569241

**Node 3**

Weighted Entropy = (30/300) \* Entropy (Node3)

= 0.1 \* 0.971

= 0.0971

Step 3: Calculate the Information Gain

Entropy (Node 1) = 0.72192

Weighted Average Entropy (Node 2) = 0.569241

Weighted Average Entropy (Node 3) = 0.0971

Information Gain (Node1, DELINQ) = Entropy (Node 1) – (Weighted Average Entropy (Node 2) + Weighted Average Entropy (Node 3))

= 0.72192 – (0.569241 + 0.0971)

= 0.72192 – 0.666341

= 0.055579

The Information Gain of Node 1 from the split (Node 2 and Node 3) is 0.055579