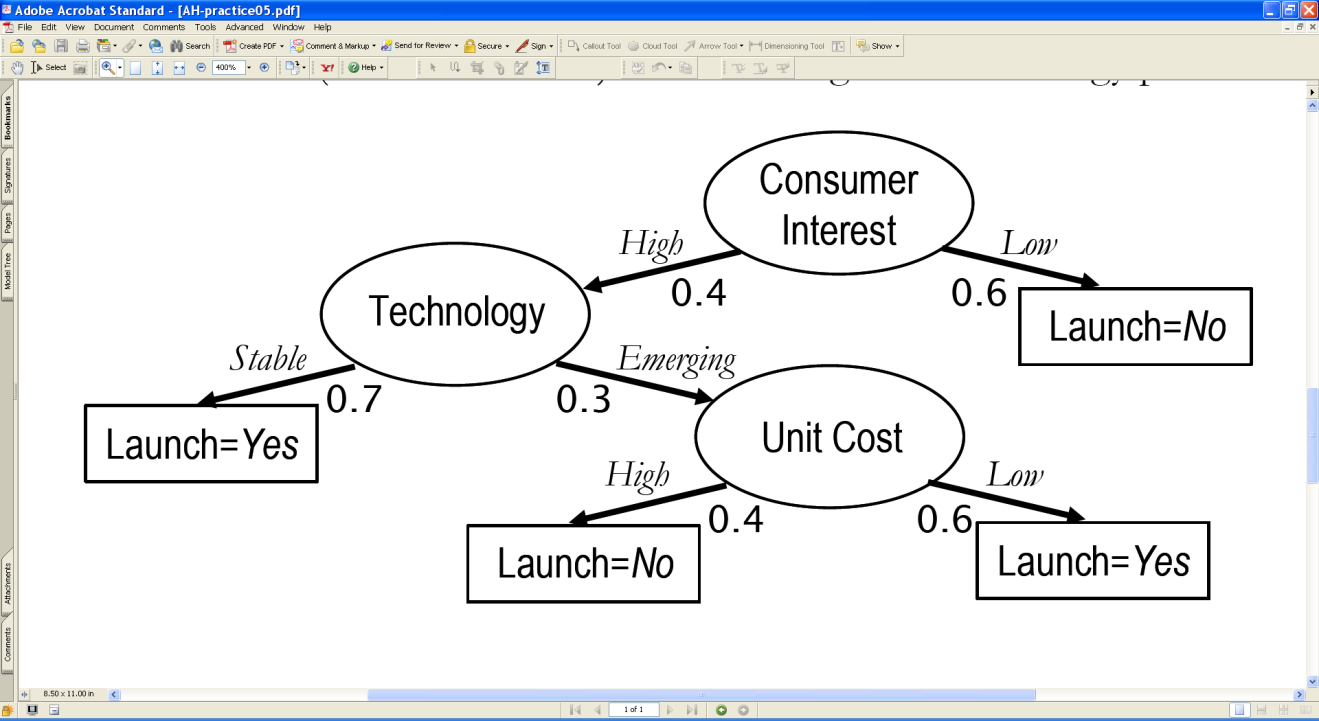
**BANA 273 Machine Learning for Analytics**

**Assignment 3**

This assignment must be completed individually. Submit Word file to online drop box on Canvas. Write your name in the Word file.

**Q.1**. Consider a decision tree (as shown below) for launching new technology products:

****

The branching probabilities are provided. Given this decision tree, find the probability Pr[Launch = Yes | evidence] in the following cases.

1. Evidence: Consumer Interest = *High*, and Technology = *Emerging*
2. Evidence: Consumer Interest = *High*

**A.1**

a) The probability of Launch = Yes is 0.6

b) The probability of Launch = Yes is 0.88

**Q. 2.** A winery maintains a dataset containing information about customers who subscribe to its tasting events and special offers for wine cases. The winery occasionally mails tasting samples of new wines in an effort to increase sales. The chief marketing officer is aiming to send samples of a newly produced wine to customers who are NOT likely to place an order for the new wine (when probability of purchase is less than 0.5). Based on a survey conducted amongst its customers and their willingness to buy a case of the new wine before tasting it, the firm collected three attributes – whether customers *prefer dry*, or have *preference for red wine* and their *ages*. The classification problem is to predict whether customers will buy or not buy wine, with certain confidence/probability. The following classification tree was induced:

Prefer

Dry?

Yes

No

Preference for

Red?

Age

No

Yes

<50

>=50

Buy

Buy

(with probability

Not buy

0.9)

(with probability

with probability

Not Buy

0.6

with probability

0.93

0.95

Answer the following questions:

1. After running a data mining software, suppose the above tree is generated. If you have to choose a single attribute to predict which customer is likely to place an order or not, which attribute would you use (choose one and briefly justify your choice)?

A. Preference for red wine

B. Age (whether the customer is older or younger than 50 years old)

C. Preference for dry wine

D. Impossible to determine given the information provided.

1. The Winery manager, Barbara, wants to consult with you if she should send a sample to a new customer named George. She tells you that George prefers dry wine and strongly prefers red wine. What’s your recommendation (on whether to send a sample to George)? Justify. (Note: The chief marketing officer is aiming to send samples of a newly produced wine to customers who are NOT likely to place an order for the new wine (when probability of purchase is less than 0.5).)
2. Assume that the cost of producing and shipping a wine sample to a customer is $8 and the revenue from each order is $100. Assume the company ascertained that the probability of purchase would become 17% for a customer who receives a sample of a new wine. What is the expected incremental revenue and the incremental cost of shipping a sample to a customer preferring dry and red wines? Would you suggest shipping a sample to customers preferring dry and red wines? Explain your answer.

**A. 2.**

1) The single attribute to predict which customer is likely to place an order or not is D. Impossible to determine given the information provided.

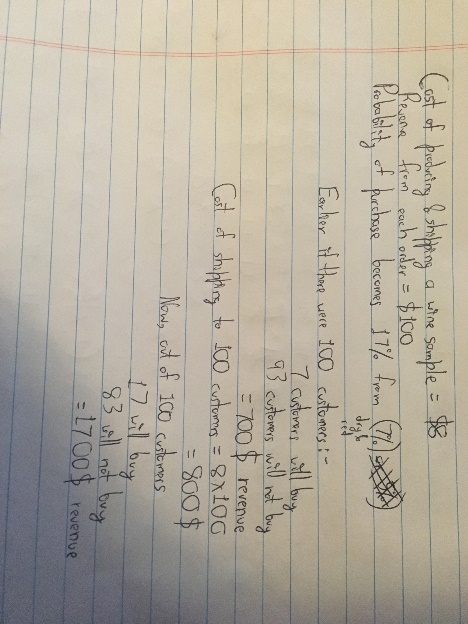
This is because the decision tree can work only in the format given. All the attributes are required for a decision to be taken. If dry wine is preferred, then the preference for red wine would be a good attribute to take a decision. Or if dry wine is not preferred, then age is a good attribute to take a decision. However, it is not possible to predict which customer is likely to place an order if only one attribute is to be taken into consideration.

2) My recommendation to Barbara is to send a sample to George. This is because looking at the decision tree, we can see that there is a 0.93 probability that George is not going to buy the new wine. As the chief marketing officer is aiming to send samples of a newly produced wine to customers who are not likely to place an order for the new wine, Barbara should send a sample to George.

3) The expected incremental revenue of shipping a sample to a customer preferring dry and red wines is 1700/100 - 7 = 10 dollars.

The incremental cost of shipping a sample to a customer preferring dry and red wines is 8 dollars.

Yes I would suggest shipping a sample to customers preferring dry and red wines. This is because shipping a sample to customers preferring dry and red wines will cause the profit to be 10-8 = 2 dollars per customer.



**Q. 3.** An NBA specialist is trying to predict each team’s likelihood of winning the championship (so this is the dependent variable). She decides to use two predictors (i.e. independent variables) – 1) whether a team won more than 55 games during the past season and 2) whether the team as a whole is healthy. The dataset below contains information about the top eight teams and whether the specialist thinks they have a chance to win.

|  |  |  |  |
| --- | --- | --- | --- |
| Team | Win less than 55 games? | Team Healthy? | Contender to win the championship? |
| Phoenix Suns | No | Fair | No |
| Detroit Pistons | No | Excellent | No |
| San Antonio Spurs | No | Fair | Yes |
| Miami Heat | No | Fair | Yes |
| Denver Nuggets | Yes | Fair | Yes |
| Seattle Supersonics | Yes | Excellent | No |
| Houston Rockets | Yes | Excellent | Yes |
| Dallas Mavericks | No | Excellent | No |

Use the examples in the above database to determine which attribute you should split on **first**, in order to build a decision tree to predict whether a team is a contender to win. Explain each step and show **all** relevant computations. To simplify your computation, you are given that the information gain if splitting on “team healthy” is 0.189.

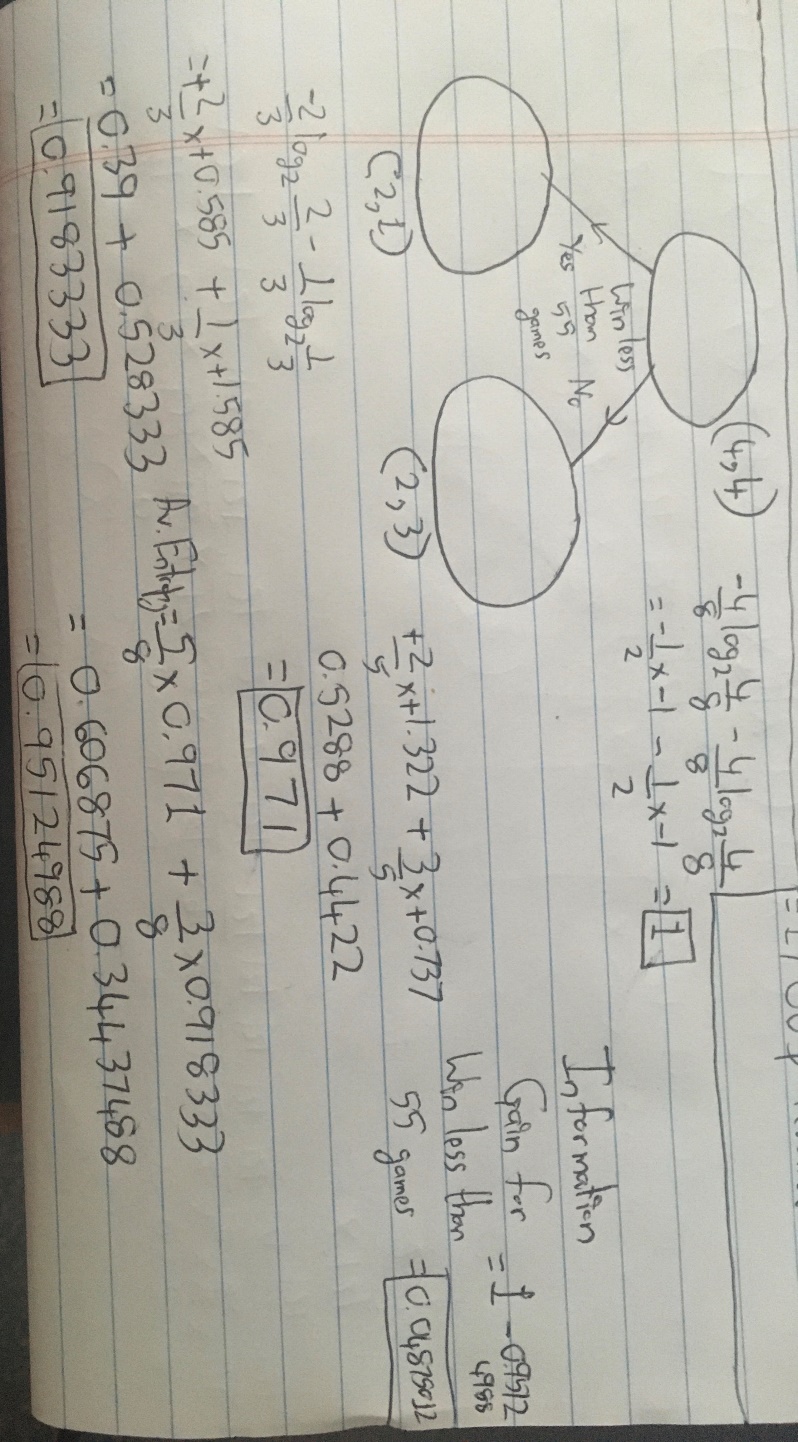
You may need the following logarithm values for answering the question:

 = -0.585,  = -1.585,  = -1,  = -0.737,  = -1.322

= -2,  = -0.415

**A. 3.**

The information gain for winning less than 55 games is given below.

****

The information gain for winning less than 55 games is less than the given information gain for team healthy.

Therefore the first attribute to be split on is the team healthy attribute.

**Q. 4.** A bank selected 12 inputs to predict whether each loan applicant defaults. The output variable (BAD) indicates whether an applicant defaulted on the home equity line of credit. The following table describes the variables used.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| **BAD** | Binary | 1=applicant defaulted on loan or seriously delinquent  0=applicant paid loan |
| **CLAGE** | Numeric | Age of oldest credit line in months |
| **CLNO** | Numeric | Number of credit lines |
| **DEBTINC** | Numeric | Debt-to-income ratio |
| **DELINQ** | Numeric | Number of delinquent credit lines |
| **DEROG** | Numeric | Number of major derogatory reports |
| **JOB** | Nominal | Occupational categories |
| **LOAN** | Numeric | Amount of the loan request |
| **MORTDUE** | Numeric | Amount due on existing mortgage |
| **NINQ** | Numeric | Number of recent credit inquiries |
| **REASON** | Binary | DebtCon=debt consolidation  HomeImp=home improvement |
| **VALUE** | Numeric | Value of current property |
| **YOJ** | Numeric | Years at present job |

**Question 4.a:** Download the file HMEQ.arff from Canvas.Load HMEQ.arff into WEKA. What’s the class distribution for BAD (i.e. percentage of 1s vs. percentage of 0s)? Build a classification model to predict whether a customer will default using decision tree model J48 (classifiers -> trees -> J48). Please use the default parameter setting for J48, and choose Percentage split (66%) for Test options. Please report the classification accuracy rate of the model built, as well as the confusion matrix. Use confusion matrix to comment on how good the model is.

**4.b** Now, go back to the Preprocess window where you have the HMEQ.arff file open. Under Filter, choose filters-> supervised->instance->Resample. In the parameter window for Resample, change biasToUniformClass to 1 while leaving other parameters unchanged (i.e. invertSelection=False, noReplacement=False, sampleSizePercent=100). Click on Apply and now report the class distribution of BAD. Use the same model set up in 4.a to build the model.Please report the classification accuracy rate of the model built, as well as the confusion matrix. Please compare with the prior class distribution and comment on how good the model is, and also use confusion matrix to comment on how good the model is, especially compared to the model in 4.a.

**A. 4**

**Answer 4.a:**

The class distribution of BAD is 80.05% 0s and 19.95% 1s.

The classification accuracy rate is 87.0681%

=== Confusion Matrix ===

a b <-- classified as

1572 59 | a = 0

203 192 | b = 1

It can be seen that the accuracy of the decision tree built is not very good. 203 of the instances that belong to b have been classified as belonging to a and only 192 of the instances that belong to b have been classified as belonging to b. That shows that if the instance belongs to b there is an approximately 50% chance of it being correctly classified as b.

**Answer 4.b:**

The class distribution of BAD is 50% 0s and 50% 1s.

The classification accuracy rate is 86.3277%

=== Confusion Matrix ===

a b <-- classified as

874 137 | a = 0

140 875 | b = 1

In this case, if out of the 1015 instances which belong to the b class, 875 instances were classified correctly as belonging to b and 140 were incorrectly classified as belonging to a. This model is much better at classifying the b instances than the first model even though the overall accuracy of the model has actually become slightly worse. As the b examples are the ones which are important to classify, we can assume that this model is actually better for our purposes.