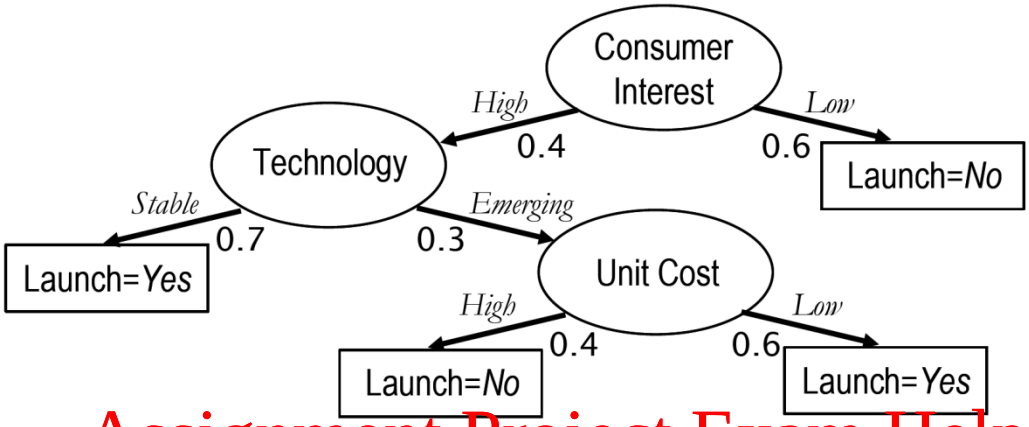
Assignment 3 Solution

This assignment must be completed individually. Submit Word file to 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 each of the following cases.

(a) Consumer Interest = *High*, Technology = *Emerging*

Pr [launch= Yes | Consumer Interest = High & Technology = emerging] =

$$\frac{Pr(Launch = Yes \ \& \ CI = high \ \& \ tech = emerging)}{Pr(\ launch = yes \ \& \ CI = high \ \& \ tech = emerging) + Pr(launch = no \ \& \ CI = high \ \& \ tech = emerging\ )}$$

$$= \frac{0.4*0.3*0.6}{0.4*0.3*0.6 + 0.4*0.3*0.4} = \mathbf{0.6}$$

(b) Consumer Interest = *High*, Unit cost = *Low*

Pr [launch= Yes | Consumer Interest = High & Unit Cost = low] =

$$\frac{Pr(Launch = Yes \ \& \ CI = high \ \& \ unit\ cost = low)}{Pr(\ launch = yes \ \& \ CI = high \ \& \ unit\ cost = low) + Pr(launch = no \ \& \ CI = high \ \& \ unit\ cost = low\ )}$$

$$= \frac{0.4*0.3*0.6}{0.4*0.3*0.6 + 0.4*0.3*0} = \mathbf{1}$$

(c) Technology = *Emerging*

Pr [launch= Yes | Technology = emerging] =

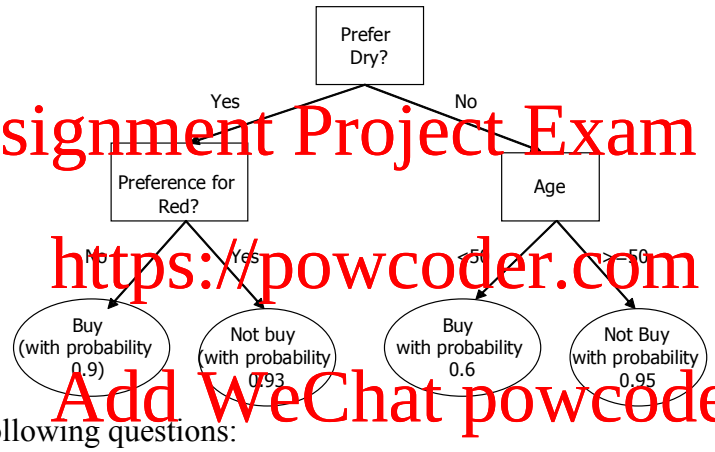
$$\frac{Pr(Launch = Yes \ \ tech = emerging)}{Pr(\ launch = yes \ \& \ tech = emerging) + Pr(launch = no \ \& \ \ tech = emerging\ )}$$

For $Pr(launch = no \ \& \ tech = emerging\ )$:

- Pr(launch = no & CI=high & tech = emerging) = 0.4*0.3*0.4

- Pr(launch = no & CI = low & tech = emerging) = 0.6
- $Pr(launch = no \,\&\, tech = emerging)$ = 0.4*0.3*0.4 + 0.6
- Pr [launch= Yes | Technology = emerging] = $\frac{0.4*0.3*0.6}{0.4*0.3*0.6 + 0.4*0.3*0.4 + 0.6}$ = **0.1**

Q2. 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 tree 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: (3 points)



Answer the following questions:

(a) After running the 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 (check 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.

C: The first split is the one that has the most information gain.

(b) 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).)

The likelihood of George not buying is 93%. Hence, the likelihood of George buying is 7%, which is much lower than the 0.5 threshold. I would recommend Barbara to send the sample to George.

(c) 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 will become 17% for a customer who receives a sample of a new wine. Would you suggest shipping a sample to customers preferring dry and red wines? Explain your answer.

Expected incremental revenue from a customer who receives sample = $100*(0.17-0.07)=$10. This exceeds the cost of sending the sample ($8). Net gain of $2.

Sending samples is only profitable if this net gain of $2 exceeds the cost of producing and shipping actual order. Assuming cost of producing and shipping the actual order does not exceed $2 (expected incremental revenue – cost of sample) **I would suggest shipping a sample to customers preferring dry and red wines.**

Q3. 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. (2 points)

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:

$\log_2 \frac{2}{3}$ = -0.585, $\log_2 \frac{1}{3}$ = -1.585, $\log_2 \frac{1}{2}$ = -1, $\log_2 \frac{3}{5}$ = -0.737, $\log_2 \frac{2}{5}$ = -1.322

$\log_2 \frac{1}{4}$ = -2, $\log_2 \frac{3}{4}$ = -0.415

- Information Gain with Win less than 55 games

Impurity (win less than 55 games)=-((2/3)* $\log_2 \frac{2}{3}$)-((1/3)* $\log_2 \frac{1}{3}$)=0.918

Impurity (win more than 55 games)=-((2/5) $\log_2 \frac{2}{5}$)-((-3/5) $\log_2 \frac{3}{5}$)=0.971

Average Entropy of children=((3/8)*0.918)+((5/8)*0.971)=0.951

Impurity (Entire population)=-((1/2) $\log_2 \frac{1}{2}$)-((1/2) $\log_2 \frac{1}{2}$)=1

Information gain=1-0.951=0.049

- Information Gain with Team Healthy

Impurity (Excellent)=-((1/4) $\log_2 \frac{1}{4}$)-((3/4) $\log_2 \frac{3}{4}$)=0.811

Impurity (Fair)=-((1/4)* $\log_2 \frac{1}{4}$)-((3/4) $\log_2 \frac{3}{4}$)=0.811

Average Entropy of children=((1/2)*0.811)+((1/2)*0.811)=0.811

Impurity (Entire population)=1

Information gain=1-0.811=0.189

When we compare these two information gain, Team Healthy : 0.189 > Win 55 : 0.049

Because the information gain of team healthy attribute is greater than the information gain of winning 55 games attribute, I should split on team healthy attribute first.

Q. 4. A bank selected 12 inputs to predict whether each 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.

Class distribution of BAD:

1s: 20% (1189/5960)

0s: 80% (4771/5960)

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.964	0.514	0.886	0.964	0.923	0.541	0.815	0.932	0
	0.486	0.036	0.765	0.486	0.594	0.541	0.815	0.604	1
Weighted Avg.	0.871	0.421	0.862	0.871	0.859	0.541	0.815	0.868	

=== Confusion Matrix ===

a	b	<-- classified as
1572	59 |	a = 0
203	192 |	b = 1

Accuracy is 87% (1572+192)/(1572+59+203+192). Although 87% seems high for a model, this 87% accuracy cannot be used in isolation. Type 1 error rate is 10% (203/(1572+59+203+192)). Given that loss incurred by the firm when a loan is not repaid by a customer greatly exceeds profit per customer, 10% type 1 error may be unacceptable to many lending institutions. Additionally, type 2 error is 2.9% (59/2026). Meaning, the firm may incorrectly reject 2.9% loan applications if it followed this model. Expected loss due to type 1 errors versus expected profit from each customer can be used to determine whether this model is good enough, but the 87% accuracy rate cannot be used in standalone basis. As a result of 10% type 1 error and insufficient class distribution for BAD=1 this model does not seem good.

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.

Class distribution of BAD:

1s: 50% (2980/5960)

0s: 50% (2980/5960)

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.864	0.138	0.862	0.864	0.863	0.727	0.919	0.921	0
	0.862	0.136	0.865	0.862	0.863	0.727	0.919	0.886	1
Weighted Avg.	0.863	0.137	0.863	0.863	0.863	0.727	0.919	0.903	

=== Confusion Matrix ===

a	b	<-- classified as
874	137	a = 0
140	875	b = 1

Accuracy:
In 4.a we had more data about loans that were repaid but in this model the data is equally split (approximately 50% of 0s and 1s for BAD). Accuracy of the model is similar. However, type 1 error rate reduced to 6.9% (140/(874+137+140+875)). This is a superior model than the one in 4.a not only because class distribution is even but also because type 1 error, which would cause severe losses for a lending institution, is reduced.