Name:

ISE 365/465 – Applied Data Mining Mid-term Sample Exam Questions

1. **True/False: Circle True or False (2 points each)**
   a. **True** **False**: Decision Trees can create a non-linear function

   b. **True** **False**: Sensitivity tells the rate of true positive recognition

   c. **True** **False**: Lower propensity records are graphed further to the left on the x-axis of a gains chart

   d. **True** **False**: C5.0 Decision Trees allow missing values as an actual value in the decision tree

   e. **True** **False**: Homoscedasticity is an assumption for a valid linear regression.

   f. **True** **False**: Specificity tells the rate of true positive recognition

   g. **True** **False**: Sensitivity is more important than specificity in evaluating a model.

2. **Fill in the blank: Enter an answer for each problem (2 points each)**

   a. ___Confidence_____ is a measure of how strongly the model believes in its prediction

   b. To scale a variable between 0 and 1, use ___min/max normalization___.

3. **Multiple Choice: Circle the best choice (2 points each)**
   - **a.** The entropy calculation is part of which decision tree splitting technique?
     - **i.** CHAID ii. C5.0 iii. C&R Tree (CART) iv. QUEST v. Decision List
   - **b.** Which technique can be used to reduce the number of predictor variables in a data mining problem?
     - **i.** Normalization ii. Standardization iii. Subtraction iv. PCA v. none
   - **c.** Which of the following is a valid database schema?
     - **i.** Raindrop ii. Square iii. Snowflake iv. Heart-shaped v. Criss-cross
   - **d.** Circle all methods that can be used with a nominal target variable:
     - **i.** CHAID ii.C&R Tree iii. Logistic Regression iv. QUEST v. C5.0
   - **e.** Which measure does Enterprise Miner's Linear Node give for evaluating linear regression?
     - i. – 2 Log Likelihood ii. Adjusted R-square iii. Gini iv. Gain v. Lift

4. For the data in the table, do the following: (15 total points):

Car Type	Number of True Cases	Number of False Cases
Coupe	10	20
Sedan	40	30

a. Calculate the initial overall Gini Index for the data. Assume all data is represented in the table (no missing values in Car Type). Show your work. (5points).

$Gini(D) = 1-(50/100)^2 - (50/100)^2 = 0.5$

b. Calculate the Gini Index for the variable Car Type. Show your work (5 points).

$Gini_{Car\ Type}(D) = (30/100) * (1 - (10/30)^2 - (20/30)^2) + 70/100 * (1 - (40/70)^2 - (30/70)^2)$

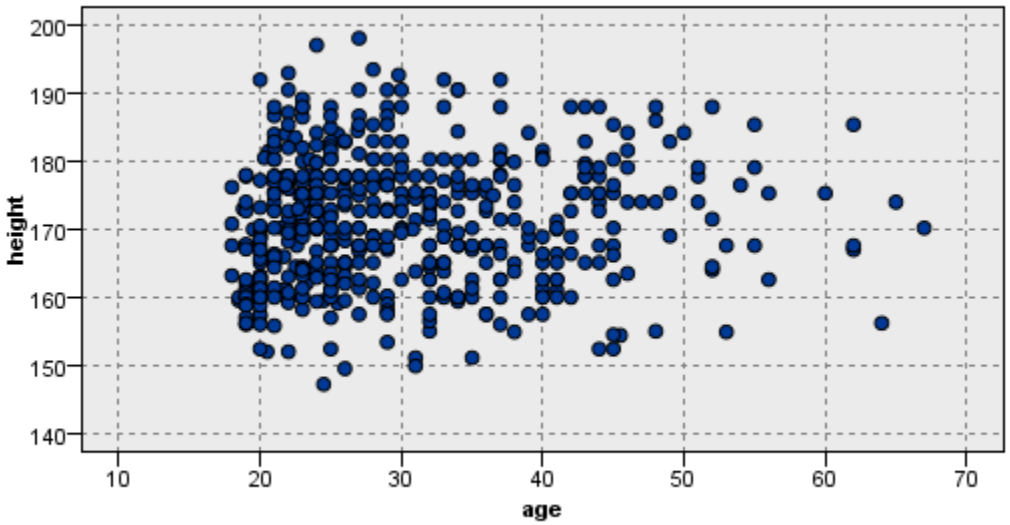
$= 0.476$

c. Assuming the best Reduction in Impurity for another variable is 0.2, will the variable Car Type enter the decision tree first. Show your work. (5 points)?

From calculations above:

0.5 – 0.476 < 0.2, so this means Car Type will not enter first as it provides less reduction of impurity compared to this other variable.

5. (10 points total) You are trying to fit a linear regression to predict the height of a sample population. The graph below shows data from this population. Answer the following questions:



a. From this graph, does it appear that age is a good variable to include in a linear regression model to predict height? Explain your answer (5 points)

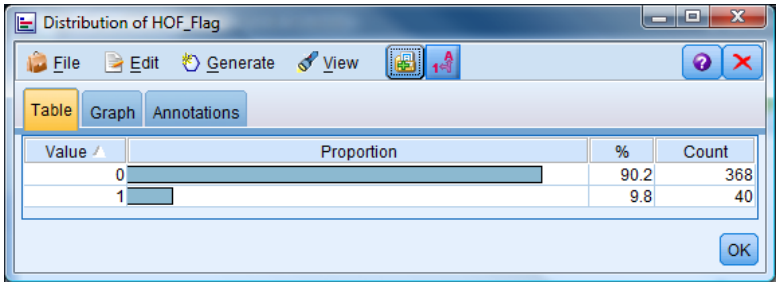
   No because there does not appear to be a strong linear relationship between age and height.

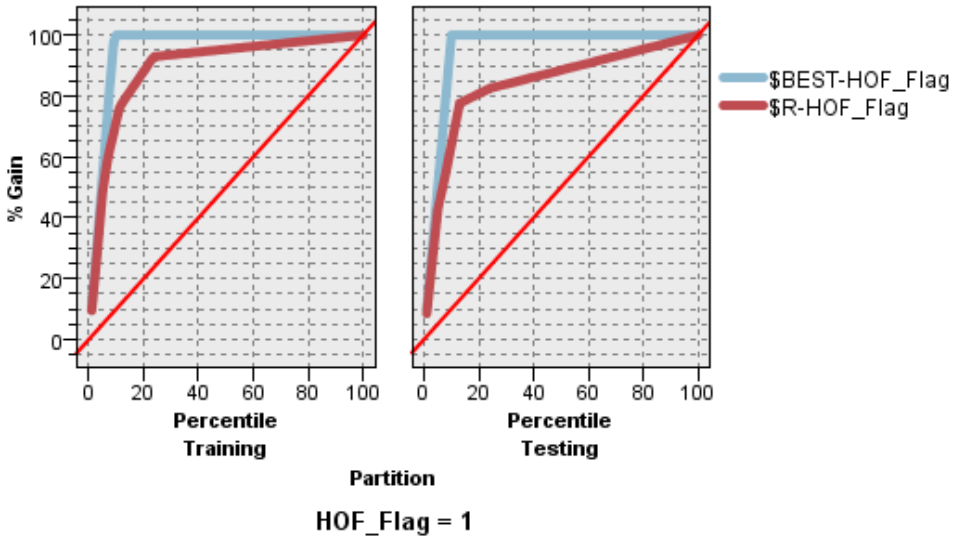
b. If the age of your population ranged from 2 to 16, do you think your answer to part a. would change? Explain why or why not. (5 points).

   Yes. Because the age and height should be linearly correlated in this age range since height increases with age at younger age.

6. (20 points total) The output of the distribution graph below shows the distribution of hall of fame players for the testing set. Use this graph and the gains chart testing set data to answer the questions below:

Name:





HOF_Flag = 1

a. What is the gain at a 20 % sample of the population for our model predicting hall of fame players (5 points)?

81%

b. What is the lift at a 40% sample of the population for predicting hall of fame players (5 points)? 86/40 = 2.15

c. How many players in the hall of fame were identified by our model with a 20% sample in the testing set? Round to the nearest whole number. Show your work (10 points).

0.81*40 = 32.4 = 32

7. (10 points) You are a statistician working for a hospital. You have conducted a study trying to predict who is at risk of dying of heart disease in the next three months. Your hospital has a procedure that can be performed to save the people you identify as at-risk, but they must come to the hospital for a day to get the procedure. You have constructed a decision tree model with the flag target corresponding to at-risk/not at-risk. Given the parameters described above, which statistic will you weight more heavily in your model evaluations - Sensitivity or Specificity? Clearly explain and defend your choice.

   Sensitivity – because in this case the cost of not identifying an at-risk person far outweighs the cost of falsely identifying a person not at-risk as being at risk. Therefore, we want to weight sensitivity more as high sensitivity ensures we identify as many at-risk patients as possible.

8. (5 points) Explain one advantage and one disadvantage of using Principal Components Analysis (PCA).

   Advantage: Can remove correlations in original predictors by reducing the number of variables to new variables created by PCA from the original variables which are not correlated with each other.

   Disadvantages: PCA is hard to explain. PCA only works on numeric data.

9. If the z-score of age for a record is 2, calculate the min-max normalized value normalized to [0, 1] for the same record's age using the following data

for the age variable across all records. Interpret what the min-max value you calculate means in words (10 points).

Variable Name	Mean	Min	Max	Variance
Age	26	5	55	49

Derive age from the z-score calculation:

2 = (age-26)/7 -> age = 40

Use the derived age to calculate min-max normalization:

(40– 5) / (55-5) = 35/50 = 0.7

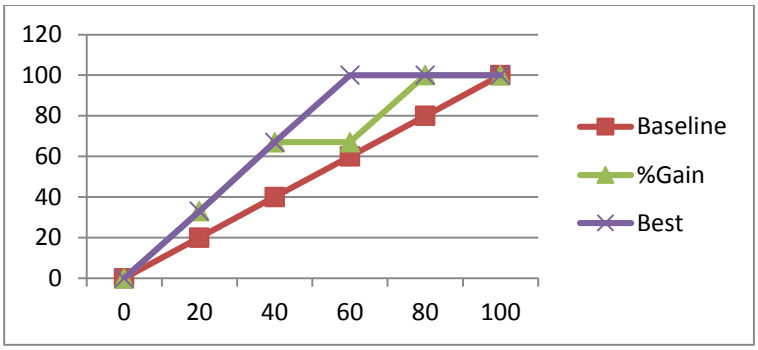
10. (5 points) Explain one advantage and one drawback of a decision tree with more tree depth than a decision tree with less tree depth.

Advantage: More complex functions can be approximated (Note: this does not necessarily mean a better accuracy depending on the problem)
Disadvantage: More difficult to explain, overfitting possible.

11. (10 points) Given the following table, answer the following questions:

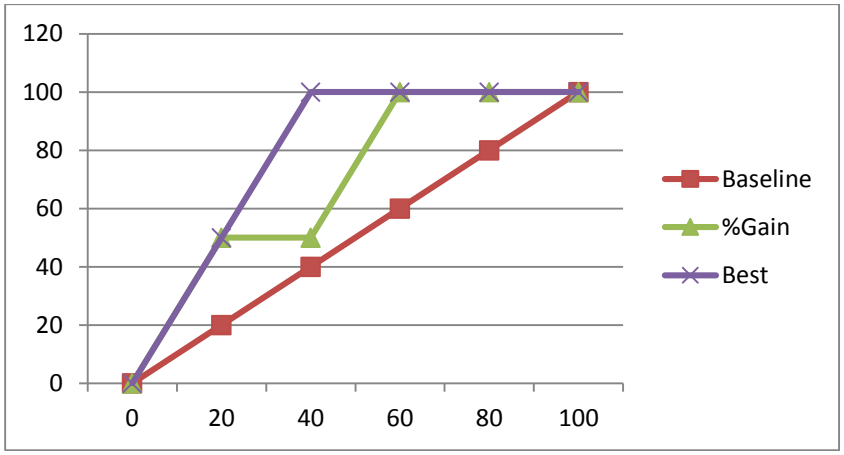
Record	Model Prediction	Actual Target	Prediction Confidence
1	True	False	0.1
2	False	True	0.2
3	True	True	0.3
4	False	False	0.4
5	True	True	0.5

a. (5 points) Draw a gains chart for predicting the "True" target value showing the best line and the line for the model results in the above table. Label the x-axis with percentile and the y-axis with % Gain. Connect the points for both the model represented by the table above and the best line (You will have two lines on your chart).



Point Order on the axis is 5-3-1-2-4. This is determined by ordering true predictions by descending confidence and then ordering false predictions in ascending order of confidence. The graph %Gain represents the actuals.

b. (5 points) Draw the gains chart as in part a, but for predicting "False" target values.



Point Order on the axis is 4-2-1-3-5. This is determined by ordering false predictions by descending confidence and then ordering true predictions in ascending order of confidence. The graph %Gain represents the actuals.

12. (5 points) You have been asked to perform a data mining task to predict which patients in an upcoming drug trial are likely to die during the trial. You have been given data with the following fields from a previous similar study to analyze: Patient ID, Gender, Age, Patient Diseases, a flag telling if the patient is alive today, and a flag telling if the patient died during the trial.
    a. (3 points) Which field(s) can you immediately eliminate as possible predictor variables? Explain your answer to get full credit.

    Patient ID – because this is patient specific

Patient alive today flag- this is a leak for predicting death in trial since the patient is dead and is probably more likely to have died in the original trial.

Patient died during trial – this is our target

b. (2 points) Which variable would you use as your target variable?

Patient died during trial - target

13. (10 points total) True/False: Circle True or False (2 points each)

a. **True** **False**: Stepwise variable selection in linear regression may not provide the optimal set of variables to use in the model

b. **True** **False**: The number of records resulting from an outer join is often larger than an inner join performed with the same data and key

c. **True** **False**: To remove a variable entirely from an Enterprise Miner project, set its role to "Rejected"

d. **True** **False**: You should use the Correlate Task in Enterprise Guide to find the correlation between hair color and gender

e. **True** **False**: Linear regression will always result in a better model for predicting continuous targets than decision trees

14. (10 points total) Fill in the blank: Enter an answer (2 points each)

a. Propensity is a measure of how strongly the model believes in a "positive" value of the target variable.

b. To generate statistics for each combination of gender and country in a data set you are analyzing, use the Summarize Statistics Task in Enterprise Guide.

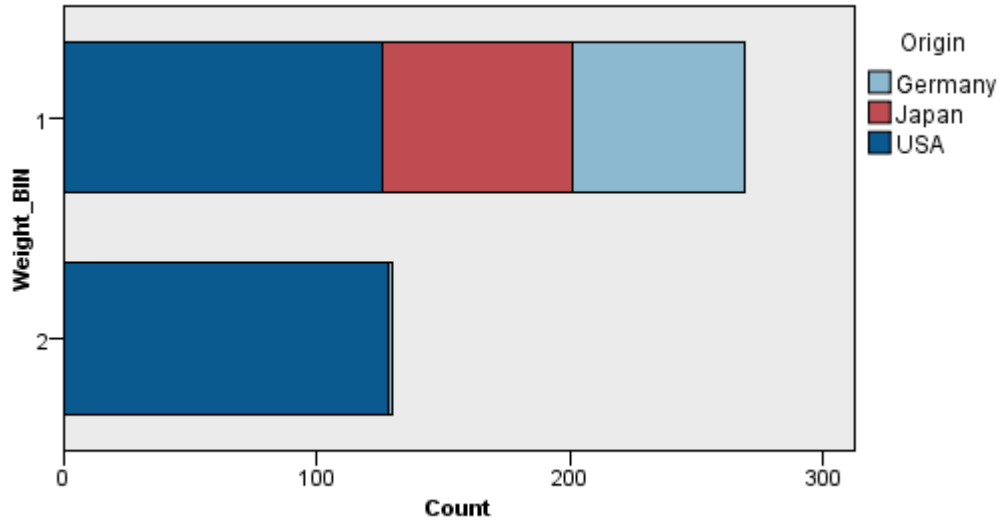
c. Curse of dimensionality limits your ability to fit a flexible model to noisy data when your data has a large number of input variables.

d. A 100 row data set with a 70%/30% training/validation partition having 10 rows with a missing input variable value will use 70 rows to build a CHAID Decision Tree model (assume proportional

assignment of the missing value rows to the training and validation set).

e. Redundancy is the term that indicates you used variables in a model which have a lot of similar information (which probably means you can reduce the complexity of your model).

15. (5 points) From this graph, does it appear visually that the binned weight variable should be removed from consideration from modeling which country a car is made in? Explain your answer

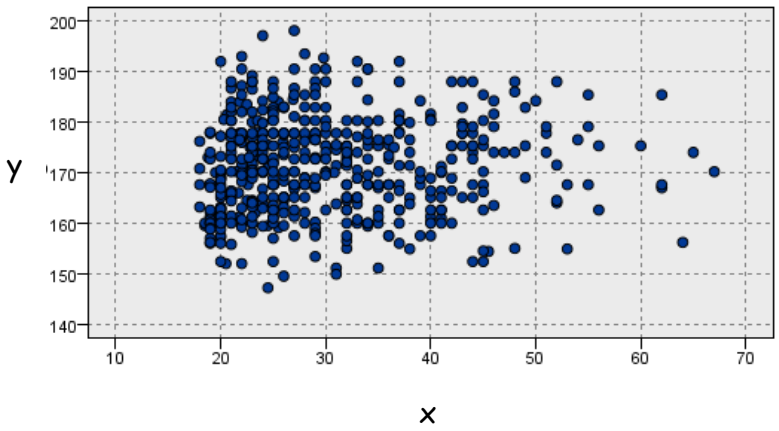


No, we cannot remove because this variable has value for predicting country since weight bin 2 has almost no Japanese and German cars whereas weight bin 1 has many more Japanese and German cars. Therefore, weight bin looks to have predictive value for country.

16. (10 points total) Explain the purpose of the Testing Set (3rd partition) in Enterprise Miner and how it is the same and different from the purpose of the Validation Set (2nd partition):

Both the validation and testing sets are hold-out sets to evaluate model results on data not used to build the model. However, the validation set is used iteratively as models as built to allow model comparison and changes to improve the model. The testing set is held out until the end of modeling to allow a final model evaluation of model contenders with data that hasn't been seen by the models or modeler at all during the modeling process.

17. (10 points) In which case below would you be more optimistic about your model if the x-axis variable is an input variable in a linear regression? **Explain in detail to get full credit.**
    a. The y-axis is the target variable with x as the only predictor in a linear regression
    b. The y-axis is the residual (error) of predictions from a multivariate linear regression with x as one of the predictors

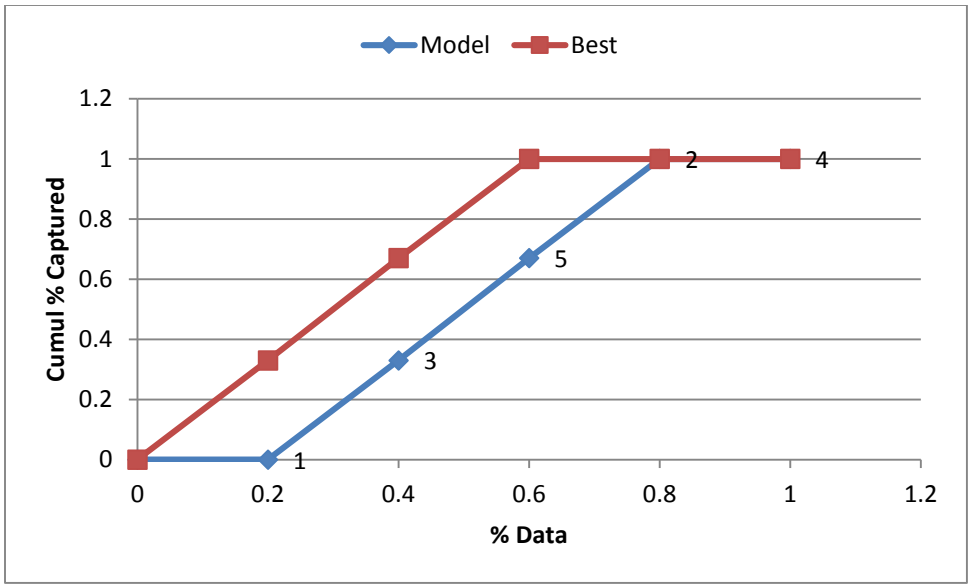


I would be more optimistic with case b because in case a, there is no linear relationship of y with x. However, the homoscedasicity of the residuals across the range of x values is acceptable. While we don't know how good the linear regression is in case b, we don't have any information as negative as case a and we do know the homoscedasicity shown on the graph is acceptable. Also, since there are mu;itple variables there is a chance some of these will have linear relationships with y.

18. (15 points total) Given the following model results assuming confidence is scaled between 0.5 and 1, answer the following question (use the definitions from the example given in class of confidence and propensity):

Record	Model Prediction	Actual Target	Model Confidence	Model Propensity
1	True	False		0.7
2	False	True	0.55	0.45
3	True	True		0.6
4	False	False		0.4
5	True	True	0.5	0.5

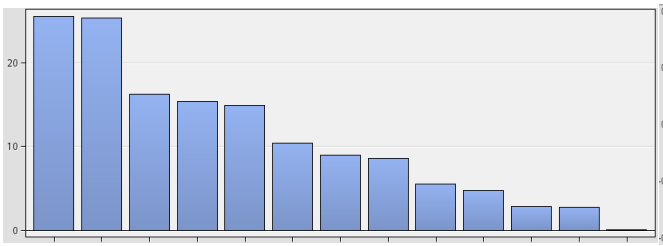
a. (10 points) Draw a cumulative % captured chart for predicting the "True" target value showing the best line and the line for the model results in the above table. Label the x-axis with percentile and the y-axis with Cumulative % Captured. Connect the points for both the model represented by the table above and the best line (You will have two lines on your chart).



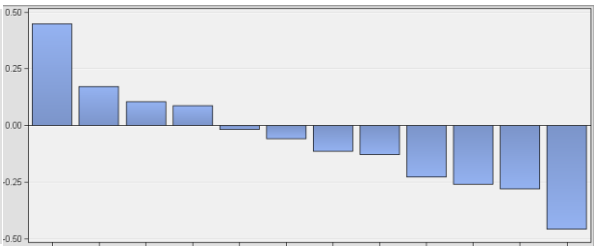
b. (5 points) What is the best lift achieved by this model (only choose from the points you graphed on the x-axis)? What % of the data did you need to inspect to achieve this lift? 1/0.8 = 1.25 lift at 80% of the data

19. (15 points total) You ran a Stat Explore Node for a supervised learning problem with a continuous target. Graph 1 and Graph 2 were results of the Stat Explore Node. Var_A and Var_B are input variables. Answer the following questions:

Graph 1



Graph 2



a. (2.5 points) What is the y-axis in Graph 1?

Worth

b. (2.5 points) What is the y-axis in Graph 2?

Correlation

c. (5 points) What do Graph 1 and/or Graph 2 tell you that Var_A and Var_B have in common?

Both A and B are important variables for predicting the target.

d. (5 points) What do Graph 1 and/or Graph 2 tell you that is different about Var_A and Var_B?

A is positively correlated with the target and B is negatively correlated with the target.

20. (10 points) You have generated a lift chart for a training and validation set of data using a model you built. Explain what you would look for in the shape of the lift charts to identify the possibility that your model overfit the data.

I would look to see if the lift curve in the training set is higher and to the right by a wide margin compared to the validation set. This means the performance on the training data set is much better than on the validation set which means overfitting is likely.

21. (15 points) For the following input tables, show what the final output table would contain for a stream that read each table separately, performed a left join in the Query Builder using the Car Origin field as the key, included all fields in the join output and used the Summarize Statistics Task with the Car Origin field as the Classification variable to calculate the mean, max, min, and Number of Observations for the MPG field as the Analysis Variable.

Car Origin (Left Table)	Make	Model Year
Japan	Toyota	1994
USA	Ford	1998

Car Origin (Right Table)	MPG	Horsepower	Cylinders

USA	5	220	8
Germany	10	155	6
United States	15	135	4
Germany	10	160	6

Summary Statistics

Results

The MEANS Procedure

Analysis Variable : MPG				
Car Origin	Mean	Minimum	Maximum	N
Japan	.	.	.	0
USA	5.0000000	5.0000000	5.0000000	1

Generated by the SAS System ('Local', X64_8PRO) on March 16, 2014 at 8:44:00 PM