Name:

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

1. **(10 points total) True/False: Circle True or False (2 points each)**
   a. **True** False: Decision Trees can only be used with a categorical target
   b. **True** False: An inner join results in the most possible records after the join
   c. True **False**: Adding input variables to a linear regression may lead to overfitting
   d. **True** False: Proc Corr in SAS should be used to evaluate the relationship between nominal variables
   e. **True** False: Redundancy is desirable in input variables in a model

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

   a. Propensity can be used to order the records in a cumulative lift chart without knowing the prediction of the model.

   b. This type of graph shows the median and spread of the data - Box plot.

   c. Principal Component Analysis uses eignenvalues to remove correlation in input variables.

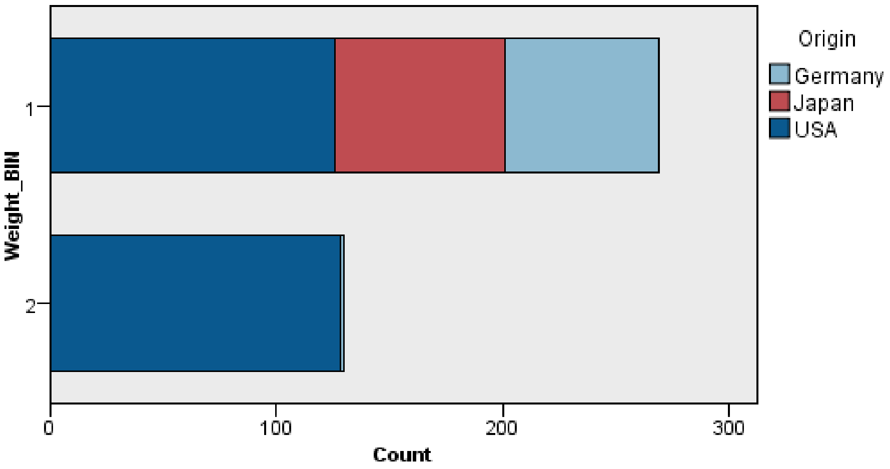
   d. Binning is used to segment an interval variable into buckets.

   e. Root Mean Square Error evaluates error in the same unit as the target variable.

3. (5 points) Explain the advantage of using gain ratio over information gain.

   Gain ratio divides information gain by Split Info to mitigate the bias information gain has for favoring variables with many values / splits.

4. (10 points) Does the weight_BIN input variable below look like a promising variable for predicting origin country of USA? Answer yes or no and **explain your answer to receive full credit.**



Yes, because the second weight bin has almost 100% USA, so a value of 2 for weight bin will almost always indicate an origin country of USA. Also, weight bin 2 has high enough count to make it relevant.

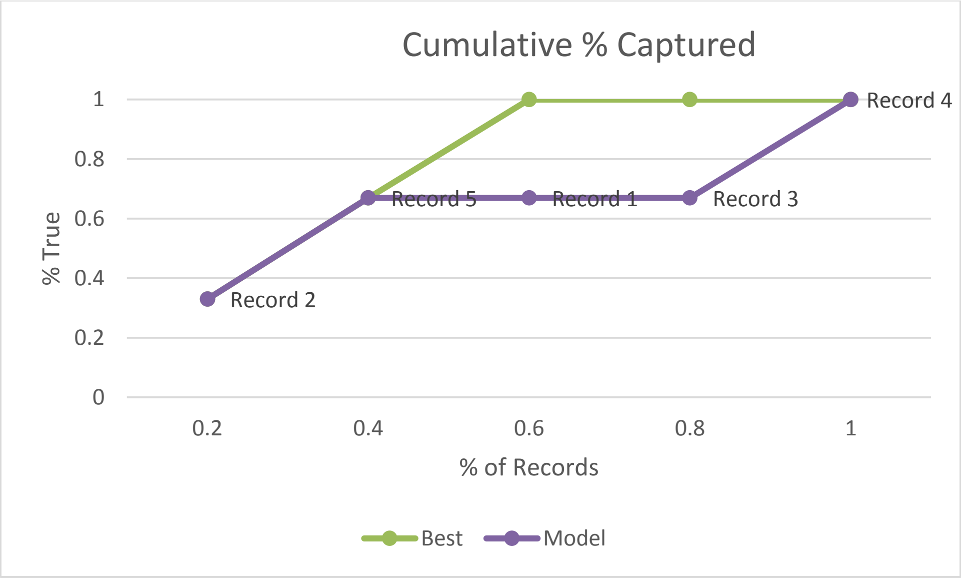
5. (5 points total) You have only 100 records to build and evaluate your model. Should you create 2 or 3 data partitions to build and test your model? **Explain your answer to get full credit.**

Since the amount of records is small, there is not enough data for a third data partition. Only 2 should be used.

6. (20 points total) Assuming confidence is scaled between 0.5 and 1 and propensity is scaled between 0 and 1, fill in the blank values in the table below and answer the following question (use the definitions of confidence and propensity from the example given in class):
(10 points for table)

Record	Model Prediction	Actual Target Value	Model Confidence	Model Propensity
1	False	False	0.6	**0.4**
2	True	True	**0.7**	0.7
3	False	False	0.7	**0.3**
4	**False**	True	0.8	0.2
5	True	True	0.5	**0.5**

a. (10 points) After filling in the blanks, draw the Cumulative Percent Captured Chart showing the best-possible and model lines on one graph.



7. (10 points) Give **2 characteristics** of a numeric input variable that would favor using ordinal measurement level instead of interval when building a decision tree. Hint: One is due to the variable itself and the other is due to the relationship of the variable with the target.

1. The variable has relatively few values so that it can be treated as a categorical variable.

2. You want the tree to group values for the variable which are not contiguous (next to each other) in the same node. Interval measurement level cannot do this as it must group by continuous ranges.

8. (10 points) You work at XYZ Transportation where your vehicles deliver packages to your customers throughout the US. You are building a model to predict vehicle maintenance costs. Should you use total annual maintenance cost as the target variable in your model? If not, propose a better metric and explain why it is better. If so, explain why.

You should not use total annual maintenance as this does not take into account that one vehicle may be used more than others. A better metric would create a ratio of maintenance costs to some measure of use such as miles driven, packages delivered, or a similar item. I would

choose Total Annual Maintenance Cost / Miles Driven to rate each trailer for maintenance relative to the workload of the trailer.

9. (20 points) For the following input tables, show what the final output table would contain for an Enterprise Guide stream that did the following in order:
   1) Read each table separately; 2) Performed a Summarize Statistics Task on the right table with the Car Origin field as the Classification Variable to calculate the mean and Number of Observations for Horsepower as the Analysis Variable; 3) Performed a right join in the Query Builder using the Car Origin field as the key and included all fields in the join output.

Left Table:

Car Origin	Make	Model Year
Japan	Toyota	1994
USA	Ford	1998

Right Table:

Car Origin	MPG	Horsepower
USA	5	220
Germany	12	150
US	15	135
Germany	10	160

Car Origin	**Mean Horsepower**	**Num Obs**	**Make**	**Model Year**
USA	**220**	**1**	**Ford**	**1998**
US	**135**	**1**	**NULL**	**NULL**
Germany	**155**	**2**	**NULL**	**NULL**