1. A large number of insurance records are to be examined to develop a model for predicting fraudulent claims. Of the claims in the historical database, 1% were judged to be fraudulent. A sample is taken to develop a model, and oversampling is used to provide a balanced sample in light of the very low response rate. When applied to this sample (*N*=800), the model ends up correctly classifying 310 frauds, and 270 non-frauds. It missed 90 frauds, and classified 130 records incorrectly as frauds when they were not.
   a. Compute a confusion matrix for the sample as it stands.

	Predicted			Total
Actual		Fraud	Non-Fraud	
	Fraud	310	90	400
	Non-Fraud	130	270	400
Total		440	360	800

[4 points]

b. Find the misclassification rate for this

sample. (90+130)/800 = 27.5%

[2 points]

c. Find the adjusted misclassification rate (adjusting for the oversampling).

Let's suppose that we had 40,000 original records. A 1% fraudulent error rate would mean that 1% x 40,000 = 400 records would be fraudulent, and 40,000 – 400 = 39,600 would be non-fraudulent. We can assume the first row will not change and the second row will have the same ratio of fraud to non-fraud, and compute the second row from the totals that are left as follows:

	Predicted			Total
Actual		Fraud	Non-Fraud	
	Fraud	310	90	400
	Non-Fraud	12,870	26,730	39,600
Total		13,180	26,820	40,000

The adjusted mis-classification rate would be (12,870+90)/40,000 = 32.4%

[2 points]

d. What percentage of new records would be you expect to be classified as fraudulent?

We would have 13,180 records classified as fraudulent out of 40,000, which yields 13,180/40,000 = 33.0%.

[2 points]

2. Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment? Why?

If we are judging the models based upon their accuracy in the validation data then B is the better model than A. Often overfitting or choosing too many parameters will result in models that fit well in the training sample but do poorly in validation samples. However, if we consider other factors than A may be favored. For example, if A has fewer parameters or A makes more intuitive sense (or confirms agrees with our theory) than A may be better. [10 points]

3. For the following question use the following raw data:

Object	Attribute 1 (x)	Attribute 2 (y)
Customer A	1	1
Customer B	2	1
Customer C	4	3
Customer D	5	4

a. Cluster the following four objects (with (x, y) representing locations) into two clusters. Initial cluster centers are: Customer A (1, 1) and Customer B (2, 1). Use Euclidean distance. Use k-means algorithm to find the two cluster centers after the first iteration. (Hint: compute the distance matrix between all the observations.)

Iteration 1

	(1, 1)	(2, 1)	
Point	**Dist Mean 1 (Cust A)**	**Dist Mean 2 (Cust B)**	**Cluster**
(1, 1)			
(2, 1)			
(4, 3)			
(5, 4)			

First we list all objects / points in the first column of the table above. The initial cluster centers – means, are (1, 1) and (2, 1) - chosen randomly. Next, we will calculate the distance from the first point (1, 1) to each of the two means, by using the distance function - Euclidean distance.

The distance from the first point (1, 1) to the first mean – (1, 1) is = 0, because the point is equal to the mean.

point	mean2
x1, *y1*	*x2*, *y2*
(1, 1)	(2, 1)

The formula for Euclidean distance between two points i and j is:

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j1}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

where x_{i1} is the value of attribute 1 for i and x_{j1} is the value of attribute 1 for j, and so on, as many attributes we have … shown up to p - x_{ip} in the formula.

In our case, we only have 2 attributes. So, the Euclidean distance between our points point1 and mean2, which have attributes x and y would be calculated as follows:

$$\begin{aligned} d(\text{point1, mean2}) &= \sqrt{|x_{p1} – x_{p1}|^2 + |y_{p1} - y_{p2}|^2} \\ &= \sqrt{|1 – 2|^2 + |1 – 1|^2} \\ &= \sqrt{|1|^2 + |0|^2} \\ &= \sqrt{1 + 0} \\ &= \sqrt{1} \\ &= 1 \end{aligned}$$

So, we fill in these values in the table:

Iteration 1

	(1, 1)	(2, 1)	
Point	**Dist Mean 1**	**Dist Mean 2**	**Cluster**
(1, 1)	0	1	1
(2, 1)			
(4, 3)			
(5, 4)			

Which cluster should the point (1, 1) be placed in? The one, where the point has the shortest distance to the mean – that is mean 1 (cluster 1), since the distance is 0.

Cluster 1	Cluster 2
(1, 1)	

So, we go to the next point; and, analogically, we fill in the rest of the table.

Iteration 1

	(1, 1)	(2, 1)	
Point	**Dist Mean**	**Dist Mean 2**	**Cluster**

	1		
(1, 1)	0	1	1
(2, 1)	1	0	2
(4, 3)	3.60	2.83	2
(5, 4)	5	4.24	2

Cluster 1	Cluster 2
(1, 1)	(2, 1)
	(4, 3)
	(5, 4)

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

For Cluster 1, we only have one point A(1, 1), which was the old mean, so the cluster center remains the same.

For Cluster 2, we have ((2+4+5)/3, (1+3+4)/3) = (3.66, 2.66)

[6 points]

b. How should one decide upon the appropriate number of clusters?

There are many options that are available:

a) Choose the number of clusters based upon a predetermined criterion, like find the best solution with 3 clusters
b) Compute multiple solutions and find the one with the classification scheme that helps describe the data well
c) Compute multiple solutions and use the one with the best ratio of within to between errors

[4 points]

4. An insurance company has examined a random sample of 190 automobile accident claims for fraud. A logistic regression model is fitted to this data with the dependent variable being coded as one for a case that was fraudulent, and as zero otherwise. The five independent (predictor) variables included in the model are:

   i) CityCode: =1 if the claimant lived in a large city, =0 otherwise;
   ii) SexCode:=1 for males, =0 for females;
   iii) Age in years;
   iv) FaultCode:=1 if the fault in the accident was that of the policy holder, =0 otherwise;
   v) Deductible Amount (in dollars).

The model estimated for the logarithm of the odds of fraud is:

$$\log(\frac{p}{1-p}) = 53.119 - 0.081\times\text{CityCode} + 0.367\times \text{SexCode}$$

$$+ 0.060\times\text{Age} - 1.738\times \text{FaultCode} - 0.142\times\text{Deductible Amount}$$

a. Describe in words what the odds ratio means for the base case claimant (CityCode=0, SexCode=09, Age=0, FaultCode=0, Deductible Amount=0)? What are her odds for fraud?

The odds ratio is the ratio of the odds that an event occurs versus that it does not occur. In this case we are looking at the chance of fraud versus no fraud. The log odds for fraud in this case are: 53.199 - 0.081 x 0 + 0.367 x 0 + 0.060 x 0 – 1.738 x 1 – 0.142 x 0 = 53.199. Therefore, the odds = exp{ log(odds) } = exp{ 53.199 } = 1.3 x 10^23 (a very large number). [3 points]

b. What are the odds for fraud in an accident where the policyholder was at fault compared to one where the fault was not that of the policyholder, assuming all other variables take their base case values?

The log(odds) = 53.119 – 1.738 = 51.381 = 2.06 x 10^22 (a very large number) [2 points]

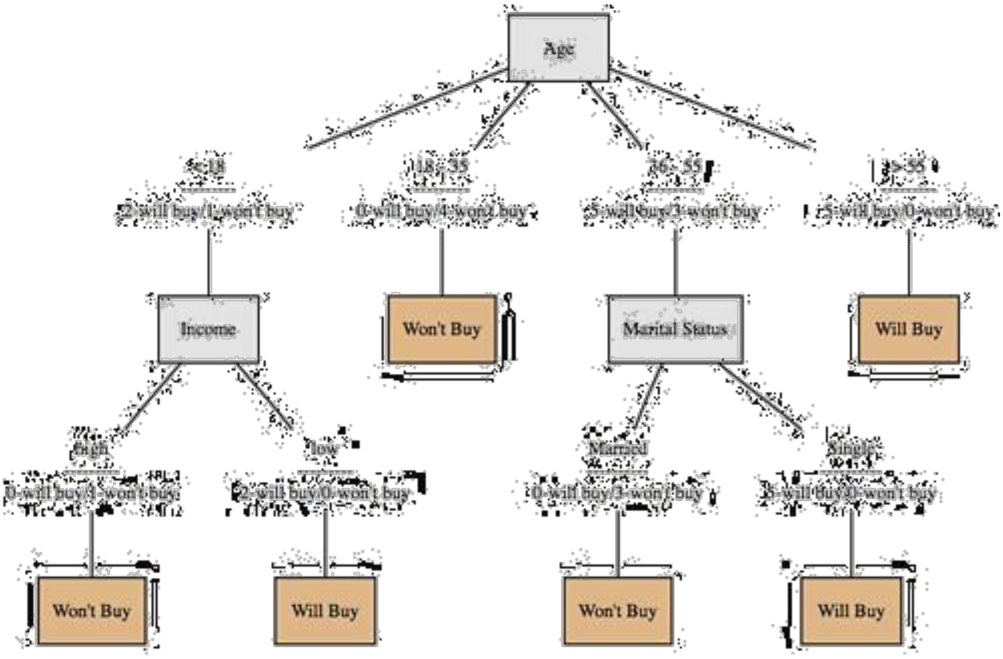
c. Does the probability for fraud increase or decrease with

age? The probability of fraud increases with age. [2 points]

d. What is the probability of fraud in a claim by a male policyholder aged 30 years, who lives in a major city, has a deductible of $400 and who was not at fault in the accident?

The LOR=log(odds) = 53.119 – 0.081 x 1 + 0.367 x 1 + 0.060 x 30 – 1.738 x 0 – 0.142 x 400 = -1.595, so the probability = exp(LOR)/(1+exp(LOR)) = 0.169 or there is a 16.9% chance of fraud in this accident. [3 points]

5. Suppose we have estimated the following classification and regression tree or decision tree. This tree has been pruned using a validation sample that is not provided. The tree is used to predict whether a consumer will purchase a product or not using age, income, marital status, and geographic location. (The number of consumer’s who will buy or won’t buy is stated underneath the variable level. For example, in the first age split for “<18”, we have 2 who will buy and 1 who will not buy.)



a. If a consumer is 38 years old, married, and has a high income, do you think this consumer is likely to buy or not? (Circle the path on the tree that you used to make this inference.)

Likely won't buy [2 points]

b. Why does geographic location not appear within this tree?

It is not predictive. It was likely dropped during pruning as unhelpful or resulting in overfitting. [4 points]

c. What segments (or clusters) would you suggest the company advertise to if they are interested in reaching those customers that are most likely to buy?

There are three potential segments: (a) Older individuals – those over 55, (b) middle age but single adults, (c) young and low income. [4 points]

- The End -