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 (4 points)

    b. Find the misclassification rate for this sample (2 points)

    c. Find the adjusted misclassification rate (adjusting for the oversampling) (2 points)

    d. What percentage of new records would be you expect to be classified as fraudulent? (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? (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.) (6 points)

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

    $$\log(p/1-p) = 53.119 - 0.081\times\text{CityCode} + 0.367\times \text{SexCode } 1 + 0.060\times\text{Age} - 1.738\times \text{FaultCode} - 0.142\times\text{Deductible Amount}$$

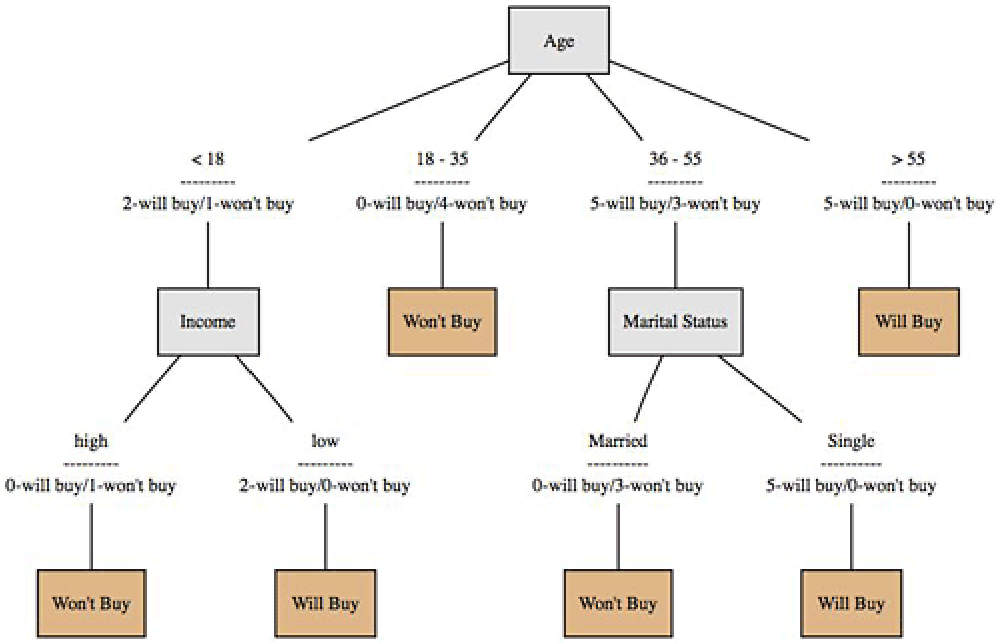
a. Describe in words what the odds ratio means for the base case claimant (CityCode=0, SexCode=0, Age=0, FaultCode=0, Deductible Amount=0)? What are her odds for fraud? (3 points)

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? (2 points)

c. Does the probability for fraud increase or decrease 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? (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.) (2 points)

b. Why does geographic location not appear within this tree? (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? (4 points)