Analytics Edge

Mid Exam (Introduction to analytics, Linear Regression and Logistic regression)

**Note: This is an open book exam. Answer all the questions Duration:4hrs**

**Answer following questions from adult.csv.**

1. How many rows of data (observations) are in this dataset?
2. How many variables are in this dataset?
3. Using the "max" function, what is the maximum value of the variable "capital-gain"?
4. What is the minimum value of the variable " capital-loss "?
5. How many observations have value “Female” in the sex variable
6. How many observations have a “relationship” value of “Not-in-family”?

**Answer following questions from AirQualityUCI.csv.**

1. In what format are the entries in the variable Date?
2. What is the month and year of the median date in our dataset?
3. In which month the average of AH is low?
4. On which weekday the average of AH is high?
5. Find in each interval of (0-8, 8-18,18-0) hours the average of AH.
6. In general, does it look like AH increases or decreases from 2004 - 2005?
7. For what proportion of days in 2004 was the AH was more than 1

**Answer following questions from adult.csv.**

1. Which values are the top 3 education values for income >50K?
2. Which of the occupations is most popular in having income >50K?

**Build a regression model to estimate AH for data set AirQualityUCI.csv.**

1. Compute the model R2(the "Multiple R-squared" value)?
2. Which variables are significant in the model? We will consider a variable significant only if the p-value is below 0.05.
3. Compute the correlations between all the variables in the training set. Which of the following independent variables is CO highly correlated with (absolute correlation greater than 0.7)?
4. Which of the following independent variables is AH highly correlated with?
5. Given that the correlations are so high, let us focus on the CO variable and build a model with only variables which have correlation is between -0.2 to 0.2 with CO. Compute the coefficient of CO in this reduced model.
6. Compute the reduced model R2
7. Compute the R2 value of the model produced by the step function.

Compute the testing set R2 using the model produced from the step function, calculate AH predictions for the testing data set, using the predict function.

**Attribute Information:**

Input variables:  
# bank client data:  
1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)  
Output variable (desired target):

15 - y - has the client subscribed a term deposit? (binary: 'yes','no')

**Answer the following questions from the dataset “bank-full.csv”**

1. Read the dataset and split into test and training sets and before splitting set the seed to 1000 and 70% should go into training set.
2. Build a logistic regression model(model1) for predicting “y” with the help of the variables “age”, “balance”, ”campaign” and “duration”. Build another regression model (model2) with above mentioned attributes excluding “campaign”. Specify the AIC value in both the models and mention which is the best model among both.
3. Computer the values of Sensitivity, Specificity for above model (with campaign).
4. Make predictions on the test set and Compute the AUC of the “model1”
5. What proportions of the customers are “Married” and have “technician” job.

**Answer the following questions from Breast\_Cancer Dataset**

1. How many rows are there in the dataset “Breast\_Cancer.csv”? Set the seed to 1000 and split the dataframe into training and test datasets such that 65% of the records come into training set
2. Build a logistic regression model to predict “Class” using all variables except “Sample\_Code\_Number”. Mention the value of AIC that you get for this model.
3. What are the significant variables in the above created model. What is the value of AIC?
4. Make predictions on the test set and calculate the sensitivity and specificity for a probability of (i)>0.4,(ii)>0.5,(iii)>0.6
5. Compute the value of Accuracy for the above problem. Also mention the maximum predicted probability of Class
6. What is the test set AUC of the above model?
7. What is the proportion of cases belong to Class “2” and Marginal\_Adhesion value of minimum 5?