**FINAL EXAM QUESTIONS**

**Extra Credit** **- 4 questions**

**Instructions for Q1**

These questions are related to the **Boston housing dataset** used in some other questions of the final exam.

We are interested in predicting which houses have median value greater than $30,000 using the variables in the **Boston housing dataset**.

The dataset ***Boston*** will be loaded into the environment once the following command is run:

library(MASS)

Please run the following code in R to learn more about the variables in the dataset:

?Boston

Create a new binary variable ***Result*** with a value of 1 if the ***medv*** (median value of owner-occupied homes in $1000s) variable is greater than $30k and 0 otherwise.

Create a logistic regression to model this question using all the variables in the ***Boston*** dataset, but please do not forget to remove the ***medv***variable while building the model. Use the information from the model to answer the following two questions. Select the closest answer.

**Q1)**

**What is the value of area under the ROC curve (AUC) for the model created? Please select the closest answer.**

1. 0.998
2. 0.986
3. 0.678
4. 0.845

**Answer**: B

**Explanation:** Refer to code below

**CODE:**

library(MASS)

library(ROCR)

Boston$Result <- ifelse(Boston$medv > 30,1,0)

names(Boston)

Boston$medv <- NULL

# Apply logistic regression algorithm on Boston data set train

logis <- glm(Result ~ ., data = Boston, family = binomial)

# Create a column for predicted values

Boston$pred\_prob\_model <- logis$fitted.values

pred <- prediction(Boston$pred\_prob\_model,Boston$Result ) # create a prediction object in R

perf <- performance(pred , "tpr", "fpr") # tpr and fpr are true and false positive rates

# AUC

auc.perf <- performance(pred, measure = "auc")

# Answer to Question 4

auc.perf@y.values



**Instructions for Q2 and Q3**

Please use the Facebook Ad dataset ***KAG\_conversion\_data\_wrangled.csv*** for the next set of questions. You should solve these questions using R (preferably using *dplyr* library wherever applicable) after reviewing the code provided for Week 11 and other resources provided for learning *dplyr* in R Learning Guide. Load the dataset as:

data <- read.csv("KAG\_conversion\_data\_wrangled.csv",stringsAsFactors = FALSE)

**Organic Impressions are ads that have generated impressions without any money spent on them. (i.e. Spent = 0).**

**Q2)**

**Which campaign (campaign\_id) has produced least number of ads with organic impressions?**

**Answer**: 1178

**Code**: data %>% filter(Spent == 0 & Impressions > 0) %>% group\_by(campaign\_id) %>% summarise(n\_ads = length(ad\_id)) %>% arrange(n\_ads)

**Q3)**

**Among ads with organic impressions, which ad (ad\_id) has the highest number of impressions?**

**Answer**: 1121094

**Code**: data %>% filter(Spent == 0 & Impressions > 0) %>% filter(Impressions == max(Impressions)) %>% select(ad\_id)

**Q4)**

Twenty samples of size 5 are taken from a stable process. The average of the sample means is 42.5 and the average range of the samples is 1.5. What is the upper control limit (UCL) for the R-chart?

a. 1.5

b. 43.37

c. 3.17

d. 0.00

**Answer**: C

**Explanation**:

UCLr = D4 \*

UCLr = 2.114 \* 1.5

UCLr = 3.17