Graded Homework #3: Part 2

Submit Assignment

Due Wednesday by 11:59pm **Points** 40 **Submitting** a file upload
File Types html and pdf **Available** Mar 27 at 8am - Apr 27 at 11:59pm about 1 month

Homework 3 - Part 2 is from Week 11 and 12 and weighs **4% of your grade**. It will require you to submit one file (HTML or PDF) which will be peer corrected by three of your peers. The TAs will also go through your submission and eventually assign the final marks.

Submit **ONE HTML or PDF file** with your code and answers to these 8 questions neatly labeled, clear and concise. Use RMarkdown to knit the file. You may work on the homework for as long as you like within the given window. As long as you do not click submit, you can enter and exit the assignment as many times as necessary during the time period that it is available. **Again, please note, you should only click "submit" when you are completely finished with the assignment and ready to submit it for grading.**

Also, please remember that you are to complete this assignment on your own. Any help given or received constitutes cheating. If you have any general questions about the assignment, please post it to the Piazza board. **If your question involves specific references to the answer to a question or questions, please be sure to mark your post as private.**

Good luck!

Instructions for Q.1 to 4

Please use the Facebook Ad dataset **KAG.csv** **(https://www.dropbox.com/s/it23ybqsqbn7uis/KAG.csv?dl=0)** for the next set of questions. We advise solving 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 below:

```
data <- read.csv("KAG_data.csv", stringsAsFactors = FALSE)
```

Q.1 Which ad (provide ad_id as the answer) among the ads that have the least CPC led to the most impressions?

[4 points]

Q.2 What campaign (provide compaign_id as the answer) had spent least efficiently on brand awareness on an average(i.e. most Cost per mille or CPM: use total cost for the campaign / total impressions in thousands)?

[4 points]

Q.3 Assume each conversion ('Total_Conversion') is worth **$5**, each approved conversion ('Approved_Conversion') is worth $50. ROAS (return on advertising spent) is revenue as a percentage of the advertising spent . Calculate ROAS and round it to two decimals.

Make a boxplot of the ROAS grouped by gender for interest = 15, 21, 101 (or interest_id = 15, 21, 101) in one graph. Also try to use the function '+ scale_y_log10()' in ggplot to make the visualization look better (to do so, you just need to add '+ scale_y_log10()' after your ggplot function). The x-axis label should be 'Interest ID' while the y-axis label should be ROAS. [8 points]

Q.4 Summarize the median and mean of ROAS by genders when campign_id == 1178.

[4 points]

Instructions for Q.5 to 8

Load the following libraries:

library(readr)

library(tidyverse)

library(correlationfunnel)

library(DataExplorer)

library(WVPlots)

library(ggthemes) [OPTIONAL]

library(ROCR)

library(caret)

library(corrplot)

Load the **advertising1 (https://www.dropbox.com/s/inyf1fgc6dkmuxb/advertising1.csv)** dataset using readr.

Convert the Clicked.on.Ad's datatype to factor through as.factor()

Q.5

a) We aim to explore the dataset so that we can better choose a model to implement. Plot histograms for at least 2 of the continuous variables in the dataset. Note it is acceptable to plot more than 2. [1 point]

b) Again on the track of exploring the dataset, plot at least 2 bar charts reflecting the counts of different values for different variables. Note it is acceptable to plot more than 2. [1 point]

c) Plot boxplots for Age, Area.Income, Daily.Internet.Usage and Daily.Time.Spent.on.Site separated by the variable Clicked.on.Ad. To clarify, we want to create 4 plots, each of which has 2 boxplots: 1 for people who clicked on the ad, one for those who didn't. [2 points]

d) Based on our preliminary boxplots, would you expect an older person to be more likely to click on the ad than someone younger? [2 points]

Q.6

Part (a) [3 points]

1. Make a scatter plot for Area.Income against Age. Separate the datapoints by different shapes based on if the datapoint has clicked on the ad or not.

2. Based on this plot, would you expect a 31-year-old person with an Area income of $62,000 to click on the ad or not?

Part (b) [3 points]

1. Similar to part a), create a scatter plot for Daily.Time.Spent.on.Site against Age. Separate the datapoints by different shapes based on if the datapoint has clicked on the ad or not.

2. Based on this plot, would you expect a 50-year-old person who spends 60 minutes daily on the site to click on the ad or not?

Q.7

Part (a) [2 points]

1. Now that we have done some exploratory data analysis to get a better understanding of our raw data, we can begin to move towards designing a model to predict advert clicks.

2. Generate a correlation funnel (using the correlation funnel package) to see which of the variable in the dataset have the most correlation with having clicked the advert.

- NOTE: Here we are creating the correlation funnel in regards to HAVING clicked the advert, rather than not. This will lead to a minor distinction in your code between the 2 cases. However, it will not affect your results and subsequent variable selection.

Part (b) [2 points]

1. Based on the generated correlation funnel, choose the 4 most covarying variables (with having clicked the advert) and run a logistic regression model for Clicked.on.Ad using these 4 variables.

2. Output the summary of this model.

Q.8 [4 points]

- Now that we have created our logistic regression model using variables of significance, we must test the model.
- When testing such models, it is always recommended to split the data into a training (from which we build the model) and test (on which we test the model) set. This is done to avoid bias, as testing the model on the data from which it is originally built from is unrepresentative of how the model will perform on new data.
- That said, for the case of simplicity, test the model on the full original dataset.
  - Use type ="response" to ensure we get the predicted probabilities of clicking the advert
  - Append the predicted probabilities to a new column in the original dataset or simply to a new data frame. The choice is up to you, but ensure you know how to reference this column of probabilities.
- Using a threshold of 80% (0.8), create a new column in the original dataset that represents if the model predicts a click or not for that person. Note this means probabilities above 80% should be treated as a click prediction.
- Now using the caret package, create a confusion matrix for the model predictions and actual clicks. Note you do not need to graph or plot this confusion matrix.
- How many false-negative occurrences do you observe? Recall false negative means the instances where the model predicts the case to be false when in reality it is true. For this example, this refers to cases where the ad is clicked but the model predicts that it isn’t