**Module Assignment**

**Module 8**

**QMB-6304 Analytical Methods for Business**



Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load the file “Module 8 Assignment Data Set.xlsx” into R. This file contains information on 20,807 employees of the City of Chicago who received overtime pay during 2016. This will be your master data set. Variables included are:
   1. department.name: A character variable identifying the department in which the employee works. This data set includes only the five largest (by headcount) in the City of Chicago government in 2016.
   2. employee.name: The name of the employee.
   3. title: the job title of the employee
   4. january through december: the amount of overtime pay the employee received in each of the 12 months in 2016.
   5. total: the total amount of overtime pay the employee received in 2016.
   6. nummos: the number of months in 2016 in which the employee received overtime pay.
   7. over5000: a binary variable indicating whether the employee earned $5000 or more in overtime pay during 2016.
2. Using the numerical portion of your U number as a random number seed, take a random sample of 3000 cases from the master data set using the method presented in class. This will be your primary data set for analysis.

**Analysis (Using the Reduced Data Set)**

1. Parameterize a logistic regression model with over5000 as the dependent and department.name and nummos independent variables.
2. Using the *summary()* command report the results of the model from Step 1.
3. State whether you believe the Residual Deviance of your model is markedly different from the Null Deviance.
4. Given your model from Part 1 and ignoring p values, which variable or variable/level will have the greatest influence in increasing the modeled probability that an employee earned $5000 or more in 2016?
5. Given your model from Part 1 and ignoring p values, which variable will have the greatest influence in decreasing the modeled probability that an employee earned $5000 or more in 2016?
6. Using the *expand.grid()* command develop a prediction file with all independent variables in the Step 1 model. For independent variables in this case use the *unique()* qualifier. R will by default calculate predicted probabilities to many decimal places. For convenience in reporting round your stored predictions to only 4 decimal places. Show the predicted probabilities for ONLY the first ten cases appearing in your prediction file.
7. Based on your predictions generated in Step 6, find the maximum and minimum predicted probabilities generated. State the values of the independent variables for these max and min cases.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is “Assignment 5” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.