***BUS 347.01 Introduction to Business Analytics***

***Homework 6 (60 Points)***

**Homework Description:**

The attached file “HW6 Data.csv” is used in this homework assignment. “HW6 Data.csv” contains exactly the same information as the datafile “HW5 Data.csv” in your previous homework.

You need to evaluate the performance of the classification model through the confusion matrix, accuracy, sensitivity & specificity, roc curve and lift chart. Answer each question in your analytics report. Your submission should be an analytics report in the word format knitted from RMarkdown code.

In your analytics report, you need clearly label the following items:

* Question Number
* R Code
* R Output
* Conclusion, if applicable

You need to submit the knitted analytics report on Blackboard. Please carefully check your work before the submission, as you can only submit your work once. Late submissions will not be accepted.

**Homework Questions (Each Question has 10 point)**

**You are considering choosing one of the following models as described below:**

* **Model 1:**
* **Model 2:**

Following previous homework instructions, you need to make the following variable as categorical: SEX, EDUCATION, MARRIAGE. Then, use a simple random sampling method to select 20,000 observations as training data and the remaining observations as validation data (set your random seed to be 12345).

**Q1.** Based on the training dataset, fit Model 1. Produce a confusion matrix for this logit model for both training and validation dataset. Use the cutoff probability 0.3 in producing the confusion matrix. Report the in-sample and out-of-sample confusion matrix.

**Q2.** Based on the Model 1 and confusion matrix in Q1, calculate the in-sample and out-of-sample accuracy, sensitivity and specificity.

**Q3.** Based on the Model 1 in Q1, plot a roc curve and find the value of in-sample and out-of-sample AUC.

**Q4**. Based on the Model 1 in Q1, calculate the lift for the first decile (top 10% scores) for in-sample and out-of-sample data.

**Q5.** Repeat the calculation of accuracy, sensitivity and specificity for Model 2 (use the same cutoff probability 0.3). Discuss which model produces higher in-sample accuracy and in-sample sensitivity; which model produces higher out-of-sample accuracy and out-of-sample sensitivity.

**Q6.** Repeat the analysis through roc curve and lift chart for Model 2. Discuss which model produces higher in-sample performance based on i) AUC ii) First Lift Index; which model produces higher out-of-sample performance based on the same criteria.