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

***Homework 5 (60 Points)***

**Homework Description:**

The attached file “HW5 Data.csv” is used in this homework assignment. The data file contains the status of people defaults on the credit card payment along with some potential explanatory variables.

You need to perform a predictive analytics modeling by using a logistic regression to forecast each individual’s probability of default. 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 the following logit model structure:**

**Q1.** Import the data and make the following variable as categorical: SEX, EDUCATION, MARRIAGE. 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). Apply the summary() function to check the training and validation dataset.

**Q2.** Based on the training dataset, fit the above logit model. Report the coefficients of the model and report the AIC of the model.

**Q3.** Based on this logit model, discuss whether the person is more likely to default on the credit card payment under the following different situations, assuming all other factors remain unchanged.

1. The person increases the payment amount
2. The person increases the bill amount
3. The person increases the limit balance
4. The person changes the marriage status from 3 to 1.

**Q4.** Based on this logit model, calculate the McFadden Pseudo R-Squared.

**Q5**. Based on this logit model, make a prediction of default in terms of probability for the following individual:

* LIMIT\_BAL = 50000
* SEX in the category of “1”
* EDUCATION in the category of “5”
* MARRIAGE in the category of “3”
* BILL.AMT = 2500
* PAY.AMT = 1000

**Q6.** Based on the training data, consider adding the variable AGE in the above logit model. Use the AIC index to determine whether the addition of the AGE variable improves the performance of the logit model. Provide explicit conclusion and numeric evidence.