Term 1, 2022/2023

DSA211 Statistical Learning with R

Group Project – Part 2 (60%)

To be handed in through eLearn by the due

1. Based on **Bank2022P.csv** data set, your team is supposed to construct a predictive model to predict the account balance *Balance* with the given independent variables *Income, Limit, Rating, Cards, Age, Education, Gender, Married,* and *Ethnicity*.

   In your written report (it must be in pdf-format and within 8 pages, excluding codes and computer outputs in Appendix), you are required to provide the detailed documentation of how you search for your best predictive model and justify each step you take in data analysis. You are also expected to provide evidences (R-codes and computer outputs) with clear explanations that your recommended model is the best predictive model among all the models considered. Set the random seed to (1234) at the beginning of your codes. State your final recommended model clearly.

   Evaluation criterion of Question 1:

   Documentation: 10 marks

   Methodology used: 10 marks

   R-codes, computer outputs interpretation and final model recommendation: 20 marks

2. On 6th November at about 12:15am, you will be given a new data set Bank2022testP.csv as a test data set. Use your recommended model to calculate the test mean squared error with this data set. The report must be in pdf-format, including the R-codes, computer outputs, and the test mean squared error.

   The best two teams with the smallest test mean squared error among all teams will be awarded additional bonus points

   Evaluation criterion of Question 2:

   Final model performance based on the submitted test errors: 20 marks

-END-