**BSAN440 (Fall 2019) Assignment 5**

**Shaobo Li**

**Due on 11/11/2019**

**Instruction**

1. There are two parts for this homework. Part I requires individual work, and part II is group work.
2. Part I requires **paper submission by 3pm** on Monday (11/11/2019).
3. Part II requires electronic submission on Bb as before. One submission per group.
4. Even for group work, you are strongly recommended to solve each question by yourself before or after your group meetings.

**Please copy the sentence below to the first page of your submission document, and sign.**

By signing below, I certify that all results are original and produced by members in our group.

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_

**Part I** (50 pts). Print out the file “HW5\_Part1.pdf”, and carefully read the instruction.

**Part II** (50 pts). Predicting movie success. **(DO NOT paste R code and raw R output. You are expected to write a managerial report. All metrics/numbers should be reported in clearly labeled tables. All questions should be answered with complete and clear statement.)**

IMDb score is a key indicator of movie success. In this problem, you again need to work on the IMDb dataset you have explored before. However, your job is to build linear regression to identify important predictors to predict IMDb score. The entire analysis procedure should roughly follow following steps.

1. Starting with EDA, use scatter plot and bar chart (you need to categorize imdb\_score) to visually explore at least six predictors that may potentially affect imdb\_score. Then quantitatively (calculate correlation coefficient for scatter plot and Chi-square test for bar chart) justify your conclusion.
2. Clean your data. You may impute missing values using median or mean; take logarithm transformation on heavy tailed variables.
3. Random split your data to training (4000 rows) and testing (the rest rows) sample. Build linear regression using all potential predictors you explored in (1) on the training sample. Report following:
   1. Coefficient estimates. Are there any insignificant variables?
   2. Model assessment metrics: adjusted R^2, MSE, AIC, and BIC.
   3. Model diagnostic plots: are there any assumption violation? Are there potential outliers?
4. Calculate mean squared prediction error (MSPE) on the testing sample.
5. Apply automatic variable selection techniques, such as forward, backward and stepwise. Which variables are selected in the new model? For this new model. report all metrics as in (3) and (4). Which model do you prefer?