ISOM3530 HW4
Due: 5pm 16 May

- You can collaborate with your classmates for the assignments. If you work in groups, please list the names of your group members in the report. Submit by one member only.
- You need to submit both the report and the source code.

The entry of Generation Z into the workforce has led to significant changes in employment dynamics, particularly with their high turnover rates. High turnover rate has a significant negative impact in any company. Companies are going to have to be more proactive about employee retention. The follow dataset contains the details of current and former employee:

"department" - the department the employee belongs to.
"promoted" - 1 if the employee was promoted in the previous 24 months, 0 otherwise.
"review" - the composite score the employee received in their last evaluation.
"projects" - how many projects the employee is involved in.
"salary" - for confidentiality reasons, salary comes in three tiers: low, medium, high.
"tenure" - how many years the employee has been at the company.
"satisfaction" - a measure of employee satisfaction from surveys.
"bonus" - 1 if the employee received a bonus in the previous 24 months, 0 otherwise.
"avg_hrs_month" - the average hours the employee worked in a month.
"left" – 1 if the employee ended up leaving, 0 otherwise. The response variable.

Preliminary study

1. Check the employee retention/turnover rate.
2. Draw a side-by-side bar chart to see the distribution of left/stay among departments.

Preprocessing

1. Creating dummy variables for all categorical variables.
2. Take the first 7000 observations to be the train set, and the remaining observations to be test set.

Modeling

1. Use the train set, select predictors by
   a. Forward (AIC) method, then build a logistic model (by MLE) with the selected predictors (model.1) [remark: no need to use dummy variables for Forward selection under R]
   b. LASSO (use 1sd rule), then build a logistic model (by MLE) with the selected predictors (model.2) [remark: LASSO needs dummy variables for categorical variables]
2. Use the test set to compare the performances of the two models by AUC.
3. Refit the better model from above by the whole dataset (Final model)

Business Insights

1. From the final model, propose strategy for employee retention (~50 words).