Project description: Classification

1. Find a data set for **binary classification** with $n > \max(500, 5p)$ and $p > 40$. Make sure to pick a dataset that has not been announced by any other student on blackboard. Post a paragraph about the data: 1) What is n, p? 2) What is the imbalance n_+/n_-? 3) What is the data about?

2. For $n_{train} = 0.9n$, repeat the following 50 times:

   (a) Randomly split the dataset into two mutually exclusive datasets $D_{validation}$ and D_{train} with size $n_{validation} = n - n_{train}$.

   (b) Use D_{learn} to fit random forrest, logistic elastic-net ($\alpha = 0.5$), logistic lasso, and logistic ridge. Use 10-fold CV to tune λ for logistic elastic-net ($\alpha = 0.5$), logistic lasso, and logistic ridge.

   (c) For each estimated model calculate the training, and test AUC.

3. Create a presentation with 5-7 slides. Your objective is to be clear and concise. Hence I recommend the following:

   (a) A brief description of the nature of the data, motivation and why it's important to analyze this data. Here we should understand why the data is important, how it was collected, what is the imbalance, what is n, p, the imbalance ratio, and what are the features. (1 slide)

   (b) Boxplots of the 50 AUCs (train and test) for $n_{train} = 0.9n$. Specifically show two plots, one for test AUCs and train AUCs, respectively. Make sure everything is clearly visible and legible. (1 slide)

   (c) For one on the 50 samples, create 10-fold CV curves for lasso, ridge and elastic-net. Record the time it takes to cross-validate ridge/lasso/elastic-net logistic regression. (1 slide)

   (d) Use all the data to do the following:

   - Also record the time it takes to fit a single ridge/lasso/elastic-net logistic regression (including the time needed to perform cross-validation parameter tuning), and random forrest. Create a table 4×2 table, the 4 rows corresponding to the 4 methods, and the two columns for test AUCs and time. Specifically, the first column should show the median of test AUCs among the the 50 samples, and the second column the time it takes to fit the model on all the data (as described in the sentences above). Is there a trade-off between the time it takes to train a model and it's predictive performance? (1 slide).
   - Present bar-plots of the standardized estimated coefficients (lasso, ridge, elastic-net), the importance of the parameters. Note that the graphs must be such that we can clearly see the differences. The idea is to see if these three statistical learning methods agree more or less on which features are important or not. Specifically, use the elastic-net estimated coefficients to

create an order based on largest to smallest coefficients. Use this order to present the estimated coefficients of lasso and ridge. Also use the same order to present the variable importance of random forrest. So in the end, we want to see a figure, with 4×1 subplots, each corresponding to the bar plots of the estimated coefficients (and variable importance). Say a few words about the most important predictors that are positively and negatively associated with the response. Is there any insight here to share? (1 slide)

(e) Concluding remarks (1 slide).