MSIT423: Computer Homework 5
Due: June 1, 2:00 pm
Professor Malthouse

Work through the exercises on paper and submit the quiz on Canvas.

1. Please submit your final model for the bike data.

2. The data set crowd.csv is from a crowd-sourcing website that enables uses to submit and discuss ideas to improve the product. Your task is to build a machine classifier to screen through the ideas and identify the good ones, which will be evaluated by product-design teams. There are 7046 rows in the data set, and each row is an idea that was submitted in the past. The variable `y` is the dependent variable and equals 1 if the idea was implemented, and 0 if not. You have three types (the 3C's) of predictor variables:

   (a) *Contributor*: information about the contributor including the number of ideas accepted in the past (`pastaccept`) and the number of comments written by the contributor (`commentsC`).

   (b) *Content*: information about the idea itself. I did text mining on the text of the idea itself, and variables `X1-X11` summarize what is in the text (if you are interested, they are latent semantic indices). You also know the `age` of the idea (how long it has been since it was submitted) and the `month` it was submitted. I have also created a variable `diversity`, which is how different the idea is from previous ideas (large values means very different from other ideas while small means that it is a small modification of the product).

   (c) *Crowd*: You have the number of people who visited the side and voted for implementing the idea (`votes`) and the number of comments written about the idea (`comments`).

   Type the following code into R:

```
> setwd("put your working directory here")
> crowd = read.csv("crowd.csv")
> set.seed(12345)
> train = runif(nrow(crowd))<.5    # pick train/test split
> table(train)
train
FALSE  TRUE
 3540  3506
```

Note that the content and contributor variables are available at the time the idea was submitted, but the company must wait, sometimes weeks or months, for the crowd variables. There are three main questions I want you to answer:

- How well (in terms of AUC) can you predict `y` using only the content and contributor variables?
- How well can you predict `y` using all three sets of variables?
- Which variables are the strongest predictors and how do the relate to `y` (in terms of shape, e.g., diminishing returns, increasing returns, inverted-U, etc.)?

I am especially interesting the difference between linear (ridge/lasso) models and the fully automated ones (random forests/trees). Use only the training data (`train==TRUE`) for determining your model, and report the AUC values for the test set. Fill out this table with the **test-set** AUC values. You should probably use classification rather than regression trees (type `factor(y)` as the dependent variable).

Model	Contributor+content	All variables
Ridge		
Lasso		
GAM		
Tree		
RF		
GBM		