ISE 5103 Intelligent Data Analytics
Homework #6

Instructor: Charles Nicholson

See course website for due date

Learning objective: Perform predictive modeling using regression techniques.
Submission notes:

- Teams of 1, 2, or 3
- Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.)
- Only highly *relevant* computer output should be provided unless noted otherwise.
- You may use "R Markdown" to *help* with your submission. However, please edit the final submission to clearly and concisely respond to the questions. **The goal is to limit your complete homework submission to about 10 to 12 pages. (3 pt penalty *per page* once exceeded 14 pages.)**
- You will submit your complete R script (or RMarkdown file) – the code itself is part of your solution – make sure to *provide comments* on what your code is doing.
- Do not zip your files for submission. Submit exactly two files. Name the files "LastName(s)-HW6" with the appropriate file extension (that is, .pdf, .R, or .Rmd)

In one part of this assignment you will submit predictions to a private Kaggle competition hosted by OU. In order to join the competition, you need to create a Kaggle account. Only one account per student is allowed. Then, follow the following link to join the competition and access the assignment data and important informationd: `https://www.kaggle.com/t/14220c3388bb41058d911f9a40c586e7`

Once you join the competition, form a team following the name guidelines stated in the rules. Even if you want to work alone, you need to form a team. Download the data, read the problem description, read the rules and the rest of the information on the Kaggle site. Enjoy the competition!

1 Predicting violent crime rate

Given real-world data relating to various communities and their socio-demographics, law enforcement details, and crime statistics, your goal is to predict the community-level per capita violent crimes. The target variable is continuous and you may use any techniques at your disposal to produce a highly predictive model.

(a) (20 points) Explore the provided data, conduct any desired data preparation. Summarize your approach.

(b) (60 points) Using cross-validation of your choice, perform PLS, ridge regression, lasso, elastic net, and SVM-regression to tune the associated hyper-parameters and estimate the generalizable error for a model to predict the community-level per capita violent crimes. Provide a chart summarizing the hyper-parameter search for each approach.

Summarize your best tuned results in a table such as the following:

Model description	R package	tuned hyper-parameters	CV RMSE	CV R^2
PLS	PLS	number of components: 14	123.4	0.49
SVM - radial basis	kernlab	cost: 2.7	99.4	0.72
lasso	glmnet	lambda: 0.05	107.4	0.67
...	...	...	...	...

(c) (20 points) Using any regression technique or combination of techniques you prefer, predict the predict the community-level per capita violent crimes. You will submit a CSV file with two columns (Id and CrimeRate) based on your predictions, e.g.,

```
Id, CrimeRate
1, 0.1690
2, 0.8772
3, 0.1752
etc...
```

You will work on this problem in teams, but you are competing against other teams in the class for the best score on the private leaderboard. Top performing teams will receive bonus points. The lowest performing teams will receive a penalty. See the class competition site for details.

Please note the following:

- In order to join the competition, you need to create a Kaggle account. Only one account per student is allowed.
- Once you join Kaggle and the competition, to create a team have one person click on “Team”. Then, request a merge by searching for one of the other team member’s user name and “Request Merge”. Repeat this until all team members are merged to one group.
- Create a team name as stated in the instructions. Even if you want to work alone, you need to form a team so that we can recognize if you are online or on-campus student.
- Grades will, in part, be based on the quality of your predictions as compared to the other teams in the class.
- It is your responsibility to read the rules and the rest of the information on the site.