NYC DATA SCIENCE ACADEMY

Spam Detection Case Study

Data Science Bootcamp

Spam Detection

The Task

- Congratulations! You have gotten to the group technical/coding interview round for a new job at Hewlett-Packard. You and a team of fellow Data Scientists must efficiently work together to solve the task at hand in order to impress the current Data Science team.

- The company is currently pushing a focus on email spam detection. The head Data Scientist has given you a dataset and almost no direction. He simply said:

"Tell me something interesting."

- As an up-and-coming Data Scientist, you instantly decode this request into answering the following two main research questions:
  - What are some interesting aspects about the nature of email? (Description)
  - How can we determine whether or not an email is spam? (Prediction)

The Teams

- To cover your bases, you choose to split into six subteams:
  1. Team Generalized Linear Models
  2. Team Principal Component Analysis
  3. Team Ridge & Lasso Regression
  4. Team Cluster Analysis
  5. Team Trees
  6. Team Support Vector Machines

- Teams will be randomly assigned:
  - Should a team member be **late**, you must fill them in on the details.
  - Should a team member be **absent**, you will make do -- man down.

- Teams may not use support from TAs, instructors, or staff. You are on your own.

  As always, you will find a way...

The Submission

- You and your subteam must work together to quickly uncover as many insights as possible and prepare to present to the Data Science team at **11:00am sharp**:
  - Each subteam will have a **strict** 10 minutes to present their findings and prove they know their stuff!
    - Every team member must present at least one insight.
  - Be warned -- the current data science team will be ready to fire questions!
    - Every team member must answer at least one question.

- You are expected to deliver a presentation in .pdf format; upload this file to the bootcamp Slack channel by the deadline. **Late submissions will not be accepted.**

- Laptops must be closed during presentations; no individual or team may have an unfair advantage for the battery. **No exceptions.**

The Guidelines

- In order to pass the interview, each team must address both of the main research questions:
  - For supervised learners, you will focus more on prediction; however, it is expected that you find some insights regarding the relationships among variables. You may freely use the `type` variable when modeling.
  - For unsupervised learners, you will focus more on description; however, it is expected that you find some insights regarding how your descriptions can help identify categories such as spam and not-spam (or other categories). You may only use the `type` variable in post-hoc analyses.

- While your subteam should focus on its own machine learning topic, do not forget about EDA! Some basic numerical and graphical analyses can always help tell a story.

The Data

- In order to complete this task, you need to obtain your data:
  - Load the `kernlab` library.
  - Use the `data(spam)` command to bring the data into your workspace.
  - Use the `help(spam)` command to understand the variables in the dataset.

The rest is up to you...

GOOD LUCK!!!