Introduction to Data Science

Instructor: Daniel D. Gutierrez

CLASS PROJECT

Open Data Hacking

ASSIGNMENT

This class project is designed for you to become productive with your new data science skills by working on something practical and useful – exploring the growing availability of open data and identifying a worthwhile project. You actually can help people and important causes by completing this project since much of this open data remains unexplored. You'll select from data sets over a wide spectrum of topic areas: transportation, public safety, health and social services, energy and environment, financial and economic, business development and much more. You should choose a project for which you have a passion – something you've always wondered about, an area where you'd like to make a difference.

The idea is to gain experience in trying out the principles of data science using R as discussed in this class. Please carry out the tasks outlined below for my review and assessment. Once complete, this project can serve as the first bullet-point in your new data science resume.

WARNING: Don't select a project that is too big or complex! You only have a few weeks to complete it. This project could be an initial, small part of a much larger project that you can continue to work on after this class.

TIP: After this class has ended, you should try to showcase your project by getting your results published somewhere, and let others know about what you've done using social media (Twitter, and LinkedIn). You may also try putting your code on the web as an interactive application using the Shiny tool from RStudio. You might also consider volunteering for DataKind (www.datakind.org) to use data for social good.

DATA SET

For the class project you'll need to use one of the many available open data repositories to find an appropriate data set for your chosen problem domain. Here are a few suggestions: data.sfgov.org, data.lacity.org, data.ny.gov and www.data.gov. Feel free to find a data repository and data set from your home city, state, or country. Be sure to choose a data set that supports the theme of your project, i.e. one that will help support your underlying hypothesis. The data set does not have to be large, just representative of your problem.

GOALS

The goal of the project is to engage many of the common tasks of the *Data Science Process* covered in class as shown in the figure below. Be sure to add subject domain value using your chosen data set by making predictions (supervised machine learning: regression or classification) and discoveries (unsupervised machine learning: clustering). You have free reign to explore the data and devise your own directions for the project. Be creative!

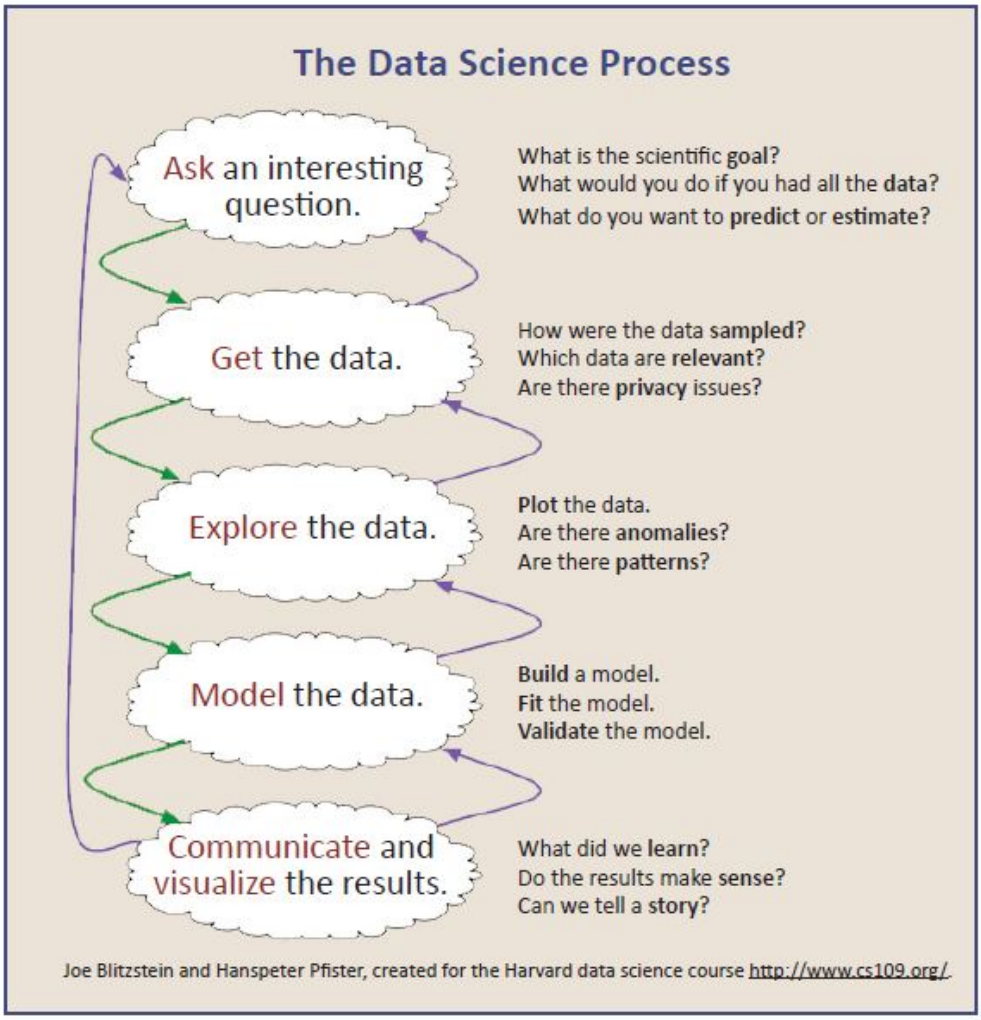
DATA SCIENCE TASKS TO CONSIDER

1. Data access: download the data set into your R environment.
2. Use the data repository to get a definition of all the variables in the data set. Perform feature engineering to select variables that support your hypothesis.
3. Perform any data transformations you feel are necessary to achieve the desired goals.
4. Use various EDA and simple statistical analysis techniques to gain a deep understanding for the data.
5. Use R's plotting features to produce both exploratory and expository data visualizations.
6. Select one or more of R's statistical learning algorithms to make predictions, and/or discoveries.

7. In the case of predictions, use the trained algorithm on new data and make a case for the algorithm's accuracy.
8. Prepare a report using techniques of "data storytelling" to present the results to a management-level audience – state the goals of the project, the data sets used, EDA results, data visualization, overview of how you used machine learning algorithms, and final conclusions.
9. Create your own GitHub repository and publish the results of the project: project description, data set used, well-commented R script(s), and the final report mentioned above.

PROJECT SUBMISSION

Submitting the project for grading is simple! Just email me the URL for your GitHub repo that includes the R code and final report. This site will be the way you can jump-start your data science career by showing others your work, getting comments and critiques, and letting future employers review your good work. Over time you can enhance your project, and/or add new projects. This site will be your technology soap box moving forward. Cultivate your new GitHub site and best of luck!


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/