ANLY 506 Course Project

The course project uses the data set: Communities and Crime Un-normalized Data Set (see below)



This data set has 147 attributes and 2215 instances or observations. This data set is relatively recent, March 2, 2011. The data set will require pre-processing before beginning data mining. For example, missing data is designated by a “?”.

**Pre-Processing Information:**

The following information will help you in data pre-processing





You will use this data set to complete your course project which comprises an analysis of the data using a statistical and a graphical analysis. Your team should create a proposal to describe exactly what you will do. You are required, at a minimum, to assess whether or not an area will have a greater than or less than/equal to the national average rate of violent and non-violet crimes. Violent crimes include those which involve force or the threat of force.

Although you can use as many appropriate variables as you want, you should take into account at least: population; ethnicity; age; income; and, education. Note that you may have to pre-process the data set to obtain these variables or some set of equivalent normalized variables.

One thing that you need to specifically look for is how imbalances affect crime rates. That is, it is well known (look for articles and information online) that income inequality increases crime rates. It has also been demonstrated that other types of imbalances result in increased crime, e.g. when the demographics of an area are not reflected in groups such as law enforcement. This has been proposed as having had a major impact on the situation in Ferguson, MO.

**Proposal** (15% of the total project score) due week 9:

Create around 10 page proposal with sections including:

* What is the problem you are trying to solve or question you are trying to answer?
* What work do you plan to do in the project?
* Which algorithms/techniques/models do you plan to use/develop? Be as specific as you can?
* How will you evaluate what you’ve done?
* What do you expect to submit/accomplish by the end of the project?

**Status Report** (25% of the total project score) due week 12:

Create around 10 page status report with sections including:

* What the problem is that you are trying to solve or question you are trying to answer.
* All relevant background information including any relevant literature you have/will use.
* The overall process you will follow for the entire project.
* A description of any relevant, interesting exploratory data analyses.
* A description of the methods/techniques/tools/algorithms you have/will use to complete the project.
* A description of the challenges you have had working on the project so far.
* A discussion of the parts of the project that have been completed.
* A discussion of the parts of the project that remain to be completed.
* A discussion of how you will finish the final project report and presentation.

**Presentation:**

* Clear research question:   
  Data preparation including pre-processing, skew correctness, standardization, normalization, handling missing values
* Choice of suitable methodology including checking the requirements of the statistical method
* exploratory data analysis including but not limited to the use of either base, lattice or ggplot2 for plotting
* Results: depending on the chosen method, important results should be reported such as R-adj sq, ROC, accuracy, goodness of fitness
* Conclusion

**Project Final Report** and Presentation (60% of the total project score) due last day:

Create a final report with at least the following sections:

* Introduction, motivation and general description of the situation, problem or challenge.
  + Following the proposal and status report, what is the situation, problem or challenge you are addressing?
* Related work.
  + Provide a thorough background for the project; e.g. you can use information from other related work you have found online – don’t forget to properly cite others work.
* Data
  + Give a complete description of the data you use during the project, including any you reject.
    - Include your Code Book as an Appendix to your final report.
* Technical Approach
  + Give a detailed description of the process for your entire project including the analyses you completed.
  + Give a detailed description of the analytics you have used including any algorithms, methods, tools or techniques. You do not have to describe well known approaches themselves, e.g. linear regression. You do have to describe how you applied the approach you used.
* Test and evaluation
  + Describe how you tested your approach to ensure that it is valid.
  + Discuss the validity of your approach.
  + Describe how you evaluated your results and/or conclusions including any specific metrics, output data, completed analyses, etc.
  + Discuss how well your approach worked to address the situation or challenge, solve the problem or answer the research question.
  + Evaluate and report whether or not someone unfamiliar with your work could accurately replicate it.
* Written work and Presentation Style
  + Written work will be graded using the rubric provided.
  + Presentation style will be graded on comprehensiveness and inclusiveness, as well as using the rubric provided.

Create a final presentation. Your presentation should be for about 30 minutes. You should also allow about 20 minutes for a question and answer session.

Your reports should be in the standard format for graduate written reports, e.g. 12-point font, 1 inch margins, double spaced, with all citations and bibliography. Note that figures and tables do not count in the page count. The bibliography also does not count in the page count.