Assignment #4: Using Statistics to Identify Spam

Data: The data for this case study is not located on the book website due to poor website maintenance. The data will be provided by the instructor in the course shell.

Note that these files will produce some minor discrepancies from the case study in the book.

Assignment Instructions:

This assignment will follow the case study in Chapter 3 of DSR. However, students should note that the case study walks you through the steps of the analysis, however, your statistical report is NOT a report or diary of your actions or the case study. Your report is a presentation of your ANALYSIS and RESULTS. Your report should focus on communicating your modeling problem, your data, your exploratory data analysis, your models, and your modeling results.

Your report should follow the following format.

(1) The Modeling Problem:

State the modeling problem. Provide enough detail that an intelligent reader could read your statement of the modeling problem, and then understand what you are modeling and why your modeling approach is appropriate.

(2) The Data:

Describe your data. If your data is simple and already cleaned, then consider a data dictionary. If your data starts out very raw, then consider a general description of the raw data followed by a description of how the data was cleaned and manipulated (or processed) and then followed by a description, sample, or data dictionary of the final modeling data.

(3) Exploratory Data Analysis:

After we have processed the data to a modeling format, we can then begin to analyze our data and glean information from it. The primary purpose of EDA is to look for **interesting relationships** in the data, typically relationships between the response variable and the predictor variables. While we are performing our EDA, we will uncover many uninteresting relationships in our data. As a matter of good practice, we typically store these uninteresting relationships in documentation for our own personal use, but we do not report the uninteresting relationships. Reporting uninteresting relationships distracts us and our audience from the more important details. We want to report the interesting results so that an intelligent reader (a modeler) can get a feel for the data.

(4) Model Comparison:

The case study will have you fit the following model suite.

Model #1: Naïve Bayes

Model #2: Decision Tree

Model #3: Random Forest

Model #4: Logistic Regression using the variable selection algorithm of your choice

Model #5: Support Vector Machine

Compare model performance both in-sample and out-of-sample and discuss your results. Which model performed best? Report the Type I error, Type II error, and Area Under the Curve (AUC) for each model. Create a summary table that has each model identified, Type I error, Type II error, and AUC. Discuss the results. Which model performed the best?

The report for this assignment should be between 8-15 pages long.

Assignment Document:

Students should present their results in the form of a report. Reports should be well written and well organized. Results should be embedded into the report in the sections with the corresponding discussion. All figures and tables should be centered and labelled.

The report document should be submitted in pdf format. File should be named Assignment4_LastName_FirstName.pdf.