# Assignment #3: Predictive Modeling in Binary Classification

**Data:** The data for this assignment is the 'Spambase' data set. The data must be downloaded from the UC Irvine Machine Learning Repository. The data will need to be read into R using the function read.csv().

https://archive.ics.uci.edu/ml/datasets/Spambase

A description of the data and the statistical problem can be found on the author's website.

http://lorrie.cranor.org/pubs/spam/spam.html

**Assignment Instructions:**

This assignment will develop a predictive modeling, or learning, framework for a classification model. However, before we perform any statistical analysis we must always perform a data quality check and an exploratory data analysis.

(1) Data Quality Check:

The first step in any data analysis project is a data quality check. Some people refer to this data quality check as 'Exploratory Data Analysis' (EDA), however true EDA is much more than a data quality check. Instead, a data quality check is exactly what it sounds like – a quick summary of the values of your data. This summary could be tabular or graphical, but in general we want to know the value ranges, the shape of the distributions, and the number of missing values for each variable in our data set. The purpose of a data quality check is for the user to get to know the data. In practice one will seldom be given a data dictionary and the corresponding data summaries. You, or a team member, will always have to perform this step as a means of 'inventory' to see what you have.

- Does the data quality check indicate that there are any data anomalies or features in the data that might cause issues in a statistical analysis?
- A data quality check should begin with a description of the data. A table with the variable name, data type, and brief description is an effective way to describe a data set.
- A data quality check should provide an overview of missing values and potential outliers. How do we detect outliers?
- Use tables when needed. Do not simply cut and paste R output.

(2) Exploratory Data Analysis:

After we have performed a data quality check and determined that we have the correct data, 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. 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.

The format and structure of your EDA is determined by your statistical problem, which in turn is determined by your data. When you are performing, or designing, an exploratory data analysis, you will need to answer the following questions to make sure that you are performing an effective EDA. You want to use EDA to frame your statistical problem.

- What type of statistical problem do you have? Is it a regression problem or a classification problem?
- What types of EDA are appropriate for this statistical problem? The correct EDA for a regression problem is significantly different from the correct EDA for a classification problem.
- What interesting relationships do you find? How can these relationships be used to build a statistical model?
- Fit a tree model to the data and use the results for exploratory insights.

(3) The Model Build

In this course we are always working within the ***learning paradigm***. That means that we will always be interested in the predictive accuracy of our models, and hence we will always be using some form of cross-validation to evaluate our models. In this problem we will use a 70/30 training-testing split of the data. You will split your data into the two separate data sets, and then use the training data sets for all of your model development and the testing data set to evaluate each of your models out-of-sample.

Now begin to fit your model suite. Fit the following models: (1) a logistic regression model using variable selection, (2) a tree model, (3) a Support Vector Machine, and (4) Random Forest. Each of your models should have its own subsection in your report. Evaluate each of the goodness-of-fit for each of these models if applicable. Use tables when needed. Do not simply paste R output into your report.

(4) Model Comparison

Construct tables to compare model performance both in-sample and out-of-sample, and discuss your results. Which model performed best?

**Presentation:**

The presentation of your results is open to you. Remember that the objective of your analysis is to effectively **communicate the important information**. At a minimum your presentation should include a section for the data quality check (where you discuss the overall approach that you used and any interesting results), a section for the exploratory data analysis (where you present and discuss your interesting findings, and an appendix of relevant R code related to the results or graphics that you have presented.

**Assignment Document:**

Students should present their results in the form of a report. Reports should be well written. Results should be embedded into the report in the sections with the corresponding discussion. All figures and tables should be centered and labelled. Samples of relevant R code related to the output discussed in the report should be included in an appendix. This should not be all of your R code, only some of the relevant R code. For example if you made a very specific statistical graphic that you discuss in the report, then the R code for that statistical graphic should be included in your appendix. The R code used to fit any statistical models should always be included in the appendix.

The report document should be submitted in pdf format.