Capstone Project – Machine Learning Plan

March 21, 2018

Jim Scotland

**The Machine Learning Problem**

This research seeks to identify factors that might addressed to reduce firearm suicide rates. Consequently, *the machine learning problem is to discover the variables that most effectively predict a state’s firearm suicide rate*. This question lends itself very naturally to a supervised multivariate regression approach.

**Critical Features of the Model**

The state level firearm suicide rate is the targeted dependent value. Initial statistical analysis suggests candidates for significant independent variables. These include the state level firearm ownership rates, the number of state level gun control laws – both overall total and counts for specific law categories, census geographic regions and state population density. Additionally, a single national level overall suicide rate for each year will be included to account for the steady rise in suicide rates, particularly since 2008.

**Gun Ownership Rate Concerns**

State level ownership rates are elusive, and the model will depend on rates for 2013 only. This means the model will not account for shifts in ownership rates over the 18-year analysis period. The reliability of the state level gun ownership data presents a second concern. Additional research uncovered another set of 30-year average proxy ownership rates that varies significantly from the modeled data. An effort is underway to access the annual proxy ownership rates underlying these averages from Boston University School of Public Health researchers. In the absence of that data, the model may be run utilizing both sets of gun ownership rates with the differences clearly noted.

**Time Sensitivity Issues**

An additional concern is the reliance on geographic region as an independent variable. This variable is also constant over time, meaning that two of the three primary predictive variables are time invariant. This is a significant drawback of the proposed model as the number of gun laws is the only critical independent variable that varies over time. This fact contributed to the inclusion of the national overall suicide rate to account for changes over time.

**Modelling Technique and Evaluation**

The predictions will be based upon a multivariate regression model using 70% of the data to train the model and the remaining 30% to test it. Success will be based upon the strength of the r-squared values for both the training set and the test set, with an eye to higher absolute t-values and high significance levels. Attention will be paid to both the “.hat “and “.cooksd” measures of influence and leverage to identify data points that may be skewing results. The machine learning R data file as it stands currently is available here:

<https://github.com/datahoundz/Springboard_Data_Science/blob/master/04_mach_learn.R>