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.

**Manual 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. Variable selection will be based upon the strength of the multiple r-squared value on the training set with an eye to variables with higher absolute values for coefficients and t-values and high significance levels. Residual plots and Q-Q plots will be created to check for data skewing and outlier values. Evaluation will include calculating the r-squared for the predicted and targeted values along with the RMSE figures for the same. In addition to residual and Q-Q plots, gain plots will be created to further visually display model performance.

**Random Forest Models**

One random forest model will be created utilizing only the 14 law categories and another using the 132 specific law variables. Their creation is primarily to determine the degree to which law variables alone can predict a state’s FSR.

**Gradient Boost Models**

Like the random forest approach above, the primary reason for gradient boost modelling is to assess the ability of law variables alone to predict FSR. However, the gradient boost approach allows the chance to view variables in terms of their importance to the model. After assessing variable importance, an effort will be made to manually create a more effective regression model. Model evaluation will be assessed as above.

The machine learning R data file is available here:

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