APSTA-GE 2047 - Fall 2018
Instructor: Ravi Shroff

HOMEWORK ASSIGNMENT 3

Instructions:

- Please submit this assignment on NYU classes by **11:00am on Thursday, 10/25/2018** (the beginning of Week 8 lecture). It is worth 100 pts.
- Work in groups of 2-3, turn in one assignment per group, and indicate group members. *Work together on all parts of the assignment.*
- A submission should include either:
    - a pdf document of all written work and any R scripts you wrote, or
    - one or more RMarkdown files.
- All scripts should be clearly written, commented, and self-contained so that I can easily run them to reproduce your analysis.
- You will be graded on completeness, writing and visualization quality, and effort/creativity.
- You have two weeks for this assignment, so please start early, and talk to me if you need help!

1. Cleaning stop-and-frisk [40 pts]

Download stop-and-frisk csv data for years 2008-2016 here.[1] Write an R script called import.R that imports, combines, cleans, and standardizes all years, and writes out one (big) csv file called sqf_08_16.csv. I have supplied you with a file called import08_09.R (and library.R) that do this for just 2008-2009 (this is what I used to create the data used in lecture 6); this problem asks you to create a similar script that covers a larger range of years. The dataset you create in this problem will be used in the next problem.

Here are specific criteria that your script should satisfy, along with some tips:

- Once you have the raw data files on your computer, you should import and combine all the 9 years of data using one 'for' loop (i.e., don't import each year individually before combining). You may need to deal with differences in columns and column names between the years.
- The output csv should have the same column names and types (e.g., logical, character, integer) as the file sqf.csv in the 'Lab 6 folder' in 'Resources' on NYU classes (i.e., what we used in lecture 6). In general, you should aim to follow the structure of the code in import_08_09.R.
- *Except* for the following fields, no variable should have more than 10% of its values missing in any year: dob, beat, officer.verbal, officer.shield, arrested.reason.
  Note: assuming your combined tibble of data is called sqf.data, the following command computes the proportion of missing values in all columns by year:

[1] The 2017 data is sufficiently different that we are not considering it for this assignment.

```
proportion.nas <- sqf.data %>% group_by(year) %>% summarize_all(funs(nas =
sum(is.na(.))/n()))
```

- 2014 is a particularly messy year. For all the use of force variables, 10 primary and 10 additional stop circumstances, frisk and search reasons, and found weapon variables, assume that ‘1’ is TRUE and NA is FALSE (in addition to the default assumption that ‘Y’ is TRUE and ‘N’ is FALSE).
- Do not subset the data to just stops where suspected.crime is ‘cpw’; keep all types of stops.

2. Making predictions with the stop-and-frisk data [60 pts]

Import the file sqf_08_16.csv that you created in question 1; you will use this dataset for all parts of this problem. Save all your code in an R script called predict.R, which should read in sqf_08_16.csv, and generate all statistics and plots used to answer the following questions. You should also submit a pdf file with your plots and answers to the written questions.

A) Restrict to stops where the suspected.crime is ‘cpw’, then train a logistic regression model on all of 2008, predicting whether or not a weapon is found. Use the following features as predictors, standardizing real-valued attributes:

- precinct;
- whether the stop occurred in transit, housing, or on the street;
- the ten additional stop circumstances (additional.*);
- the ten primary stop circumstances (stopped.bc.*);
- suspect age, build, sex, height , and weight;
- whether the stop occurred inside, and whether the stop was the result of a radio call;
- length of observation period;
- day, month, and time of day.

I. What are the ten largest and ten smallest coefficients? Pick one of these coefficients, and give a precise statement as to how that coefficient can be interpreted. **[5 pts]**

II. Suppose my (imaginary) friend, a 30 year old, six-foot tall, 165 lb man of medium build was stopped in the West 4th subway station on 10/4/2018 at 8pm (no weapon was found). Upon reviewing the UF-250 form filled out for his stop, you notice that he was suspected of criminal possession of a weapon, and was stopped because he had a suspicious bulge in his coat, and he was near a part of the station known for having a high incidence of weapon offenses. He was observed for 10 minutes before the stop was made, and the stop was not the result of a radio call. If your model were used to predict the ex-ante probability that my friend were carrying a weapon, what would this probability be? What if my friend were a woman, everything else being equal?[2] **[5 pts]**

[2] This suggests a statistical strategy for assessing discrimination. For example, if model-estimated ex-ante probabilities of weapon recovery were generally higher for women than for men, it might suggest

III. Compute the AUC of this model on all data from 2009, using the ROCR package (as in lecture 6) **[5 pts]**

IV. The AUC can be interpreted as the probability that a randomly chosen true instance will be ranked higher than a randomly chosen false instance. Check that this interpretation holds by sampling (with replacement) 10,000 random pairs of true (weapon is found) and false (weapon is not found) examples from 2009, and computing the proportion of pairs where your model predicts that the true example is more likely to find a weapon than the false example. Confirm that your answer is approximately equal to the answer computed in part III) **[10 pts]**

B) Using the same model from part A, make a plot where the x-axis shows each year from 2009-2016, and the y-axis shows the model AUC (computed using the ROCR package) when that year is used as the test set. Explain in a few sentences what you observe, why you think this might be happening, and what one might do about it. **[15 pts]**

C) For this question, you will generate a performance and calibration plot (like the ones created in lecture 6) for a classifier of your choice by following the steps below. *You must repeat this once for each team member* (e.g., if there are two people on your team, you must choose two classifiers and generate a performance and calibration plot for each). Write at least one paragraph (per classifier) explaining what you did and what you found. **[20 pts]**

- Choose a target variable that is not found.weapon or found.gun, e.g., whether a suspect is arrested, frisked, searched, whether a summons is issued, whether contraband is found, or whether force is used (or some specific type of force).
- If it makes sense, restrict to a subset of the data. For example, if your outcome measure is whether contraband is found, you may want to restrict to just stops where the suspected crime involves criminal sale/possession of “marihuana” and criminal sale/possession of a controlled substance.
- Choose a train/test split (you could do this temporally, e.g., training on 2008-2010 and testing on 2011, or randomly, e.g., training on 50% of 2008-2011 and testing on the remaining 50%).
- Choose a set of predictors (feel free to generate your own features, e.g., interaction terms, but make sure you only use pre-stop features) and a classification method (e.g., logistic regression), and fit your model.
- Generate a performance plot by sweeping over all possible thresholds, where for a given threshold, the x value represents the proportion of stops with estimated probability above the threshold, and the y-value represents the corresponding recall. For example, if half the stops have an estimated probability above the threshold of 0.3, and that subset of stops contains 3/4 of all positive cases, then (0.5, 0.75) would be a point on the plot.
- Generate a calibration plot. To generate the plot, first round the model predictions to the nearest percentage point. For each resulting bin of rounded predictions, plot the average

that officers had a higher ‘bar’ for stopping women compared to men. However, this interpretation is complicated by several statistical and substantive issues.

model prediction on the x-axis, and the empirical frequency of positive outcomes on the y-axis (points closer to the diagonal correspond to better calibration; make sure to plot the 45 degree line as well). Also, to see the distribution of model predictions, size each by the total number of events in that bin.