Community and Crime

Investigating the strongest and weakest attributes correlated with crime

David Gray
November 22, 2019
https://github.com/dsgray/Portfolio/tree/master/Crime

This project is based on the Communities and Crime Data Set. This dataset, available at the UCI Machine Learning Repository, was produced from combining the following:

- Socio-economic data from the 1990 US Census
- Law enforcement data from the 1990 US LEMAS survey
- Crime data from the 1995 FBI UCR

The goals of this project are to construct models that can predict the per capita violent crime rate for each community, and to determine the strongest and weakest attributes among the 122 attributes thought to have a plausible connection to crime that are selected from a wider set of attributes.

Cleaning Data
In examining the data, it became evident that there are a number of columns that contain multiple '?'s. Some of the entries were missing data due to the fact that police departments with less than one hundred officers generally did not participate in the LEMAS survey noted above, and, therefore, many of the communities do not have LEMAS data. All columns that had multiple entries of data missing were eliminated from the dataset because removing every row with instances of these question marks would substantially impact the amount of training and testing data. There was one '?' remaining after removing the columns just mentioned, and this community was dropped (the row was removed). If a prediction of crime for this community was needed, one of a variety of methods could be employed to impute a value to the missing datapoint.

Creating Models
Random forest and linear regression models were constructed to predict the per capita violent crime rate. The RMSE on the training dataset was 0.126, and the RMSE on the test dataset was 0.137. The score of random forest models is presented in the table below.

Tuning Parameters
I used GridSearchCV to find parameters that maximize the score of the random forest model. There was actually a marginal improvement in the score by experimenting with values of parameters other than found by optimization.

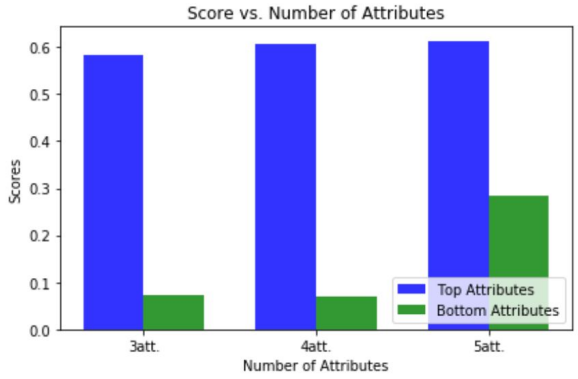
Finding Most and Least Important Attributes for Prediction

The strongest and weakest features were calculated with recursive feature elimination (RFE). RFE recursively removes features in building a model before analyzing with the remaining attributes. Also, "RFE attempts to eliminate dependencies and collinearity that may exist in the model," which may occur in this dataset. This approach results in a combination of the strongest or weakest attributes contributing to the prediction of the target (1A and 1B).

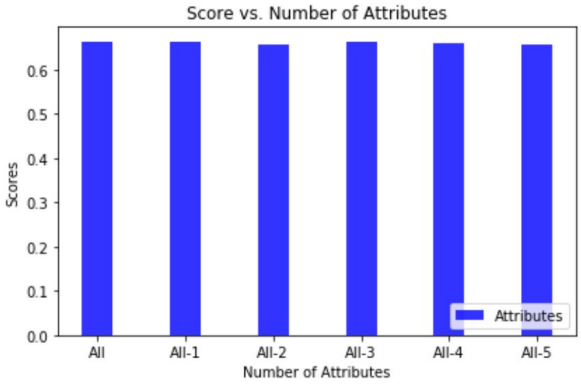
Eliminating Attributes

A small number of attributes were eliminated sequentially to see the effect on the score of the random forest model. Generally, there was a decrease in score with a decreasing number of attributes. This is what we expect (2A and 2B).

1A



2A



1B

	Score with top factors	Score with bottom factors
Three factors	0.582	0.0715
Four factors	0.605	0.0695
Five factors	0.613	0.283

Top five: PctFam2Par, PctIlleg, PctKids2Par, PctPersDenseHous, racePctWhite

Bottom five: perCapInc, PctRecImmig8, PctRecImmig10,pctUrban, MedNumBR

2B

Attributes	Score
All (allData)	0.665
All - 1 (allExceptPop)	0.662
All - 2 (allExceptPopHousehold)	0.666
All - 3 (allNumbUrban)	0.658
All - 4 (allHouseholdSize)	0.662
All - 5 (allMedIncome)	0.660
All - 6 (allpctWWage)	0.657

Recommended Next Steps

The following are recommended next steps:

- Compare the effectiveness of the random forest and linear regression models.
- Optimize parameters for the random forest model further with GridSearch.