David Troupe Big Data Algorithms 562
Analysis of National Corrections Data 1991 - 2014

I will be using the DS1: Term Records file found on this site please note you will need to create an account to download the file. I have uploaded a copy to google drive here and a description of the dataset can be found here.

Motivation

The United States has the largest prison population in the world. We incarcerate our population at an alarming rate.. A lot of interesting facts and visuals can hopefully be draw from this dataset with information on almost eleven million correctional cases. Lastly, it's important that we learn about criminals and the justice system so that we can make improvements to better our society. Currently about 67.8% [1] of criminals end up back in prison within 3 years of release. Thus we know that our justice system is failing to rehabilitate millions of prisoners.

Machine Learning Problem

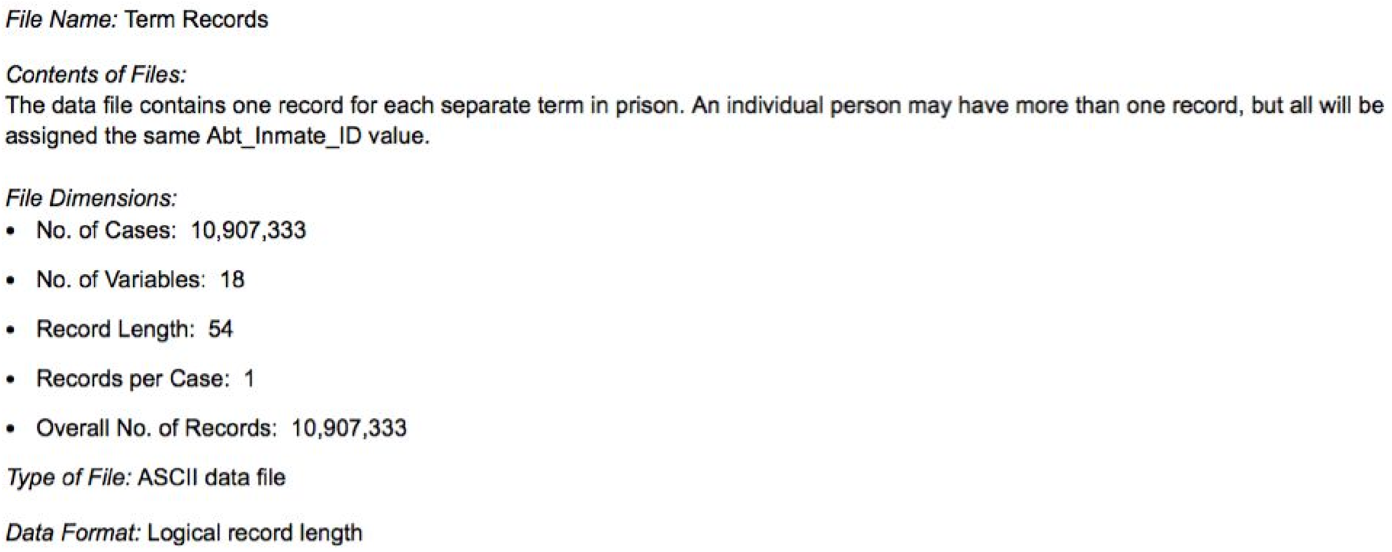
Can we predict the likelihood that a criminal will commit another crime? There are many applications of machine learning with regards to granting parole, sentencing, and even pre-crime detection. My motivation here is to design an algorithm that can predict the likelihood that an inmate will end up back in prison. Additionally, this specific application raises questions about the morals and machine learning.

Scikit-Learn

Various modules inside Scikit-Learn were used to analyze the dataset. Algorithms used included K nearest neighbors, logistic regression, and random forests. Additionally, some modules were used to preprocess ad standardize the data.

Dataset

700MB .tsv file



File Name: Term Records

Contents of Files:
The data file contains one record for each separate term in prison. An individual person may have more than one record, but all will be assigned the same Abt_Inmate_ID value.

File Dimensions:

- No. of Cases: 10,907,333
- No. of Variables: 18
- Record Length: 54
- Records per Case: 1
- Overall No. of Records: 10,907,333

Type of File: ASCII data file

Data Format: Logical record length

Results

My original hypothesis held true in that it was possible to predict which prisoners would be most likely to become repeat offenders. I first used logistic regression because there are only two possible outcomes repeat and non-repeat offenders. Unfortunately, this only resulted in a correct prediction about ~64% of the time. Thus my problem evolved and I tried K nearest neighbors (KNN) next. That algorithm resulted in a correct prediction about 71.7% of the time.

At this point the problem evolved into dimensional analysis. I analyzed the value of each feature using a random forest and KNN. I was about to remove 10 features from the dataset and improve the accuracy of my KNN algorithm to 74.7%. This was about a 1% reduction in the accuracy of the logistic regression, but the algorithm ran much faster.

Next I learned that software called COMPAS [3] was already being used in some states, and it is allegedly having a huge impact, reducing recidivism rates by up to 15% [2]. Northpoint uses separate algorithms to analyze male and female prisoners so I decided to try the same thing.

I split the data into two separate sets one that has just men and one that just has women. For the men there was seemingly no difference in the results. In fact the accuracy of KNN and logistic regression was almost exactly the same. The random forest ranked features in the same order.. This could be due to the fact that the original dataset was ~90% male to begin. For the female dataset the results were much more interesting, KNN was approximately three percent more accurate, and relied on two additional features sentence length, and release type. Lastly, logistic regression saw a major improvement when ran on the female dataset accuracy improved from 64% to 69.5%. Based on these results it seems evident that separate algorithms for men and women improve results, specifically when predicting female recidivism. The final version of the KNN on the female data set had 76.39% accuracy which is the best out of any set.

Sources

[1] Recidivism. (2014, June 17). Retrieved December 11, 2016, from https://www.nij.gov/topics/corrections/recidivism/pages/welcome.aspx

[2] Prison breakthrough. (2014). Retrieved December 11, 2016, from http://www.economist.com/news/united-states/21601009-big-data-can-help-states-decide-whom-release-prison-prison-breakthrough

[3] Northpointe Software Suite. (n.d.). Retrieved December 11, 2016, from http://www.northpointeinc.com/products/northpointe-software-suite