Project Proposal

Submitted By

Member 1

Member 2

Member 3

Member 4

1. About the dataset

The dataset we will be working on for the project for ANLY 506 course is "Communities and Crime Unnormalized Data Set" (http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized#). This data set combines data from 1995 FBI UCR and 90 census, law enforcement data. This data set has 2215 observations and 147 variables.

As per the data description on the UCI Machine learning website, out of the 147 variables 125 are predictive, 4 are non-predictive and 18 are potential goal.

The 18 potential goals or the one that can be predicted include 8 "Index Crime" as per FBI i.e. murders, rapes, robberies, assaults, burglaries, larcenies, auto theft and arsons. These 8 variables were grouped into two categories called Violent crimes and Non-violent crimes as below

Violent	Non-Violent
Murder	Burglaries
Rape	Larcenies
Robbery	Auto thefts
Assault	Arsons

.

2. Scope of work for the project

During the project we will use various graphical and statistical techniques learned during the course. Below is a tentative workflow that we will perform during the course of this project:

Pre-processing of the data

To maximize our insight into the data we will do Pre-processing on the dataset. Below is the list of activities we will perform as part of the pre-processing processing:

- Check and correct for the variable types
- Check and correct for missing values
- Check for outliers and anomalies
- Test for underlying assumptions
- Check for correlations and build parsimonious models if required

Compare area violent and non-violent crime with national averages

We will compare which community or area have violent and non-violent crime more than the national averages.

Check for effect of imbalances on the crime rate

We will check for effects of imbalances like income inequality, racial match of police, percent unemployed etc.

Check for Police constraints on crime rate

We will check for impact of Police workforce constraint like budget, number of personals etc., on violent and non-violent crime in an area. Apart from pre-processing we will run box plot and scatter plot with clusters, run multiple linear regression, etc.,

We will check for various box plots and the distribution of data to draw the conclusions. We will also use p-values and Adjusted R^2 from the regression model to draw the conclusion.