Project Description

The project is to build a training model to predict property crime in a city given the following information: name, population, # violent crimes (broken out by murders, rapes, robberies, assualts).

This document describes

* the iterations taken during the training process.
* the chosen model
* the model performance against 6 different testing sets

Training

|  |  |  |
| --- | --- | --- |
| Training Iteration 1  Training Data: NY, GA, IL, CA, UT, KY, CO  Population Cutoff 400k  Cities processing: 1975  Cities Removed due to pop above 400000:  City Population  216 New York 8396126.000  1351 Los Angeles 3878725.000  684 Chicago3, 4 2720554.000  1470 San Diego 1349306.000  1477 San Jose 992143.000  1473 San Francisco 833863.000  1774 Louisville Metro 671120.000  1906 Denver 648981.000  1265 Fresno 508876.000  1464 Sacramento 478182.000  1347 Long Beach 469665.000  366 Atlanta 451020.000  1896 Colorado Springs 436108.000  1399 Oakland 403887.000 | Property\_Crime ~ Population+Violent\_Crime  Parameter P-Value  Intercept -3.355 0.760  Population 0.017 0.000  Violent\_Crime 3.180 0.000  R-Squared: 0.898170155862724 |  |
| Training Iteration 2  Training Set Same as above  Separate Violent Crime into each individual component | Property\_Crime ~ Population+Murder+Rape2+Robbery+Assault  Parameter P-Value  Intercept 7.270 0.472  Population 0.014 0.000  Murder -11.231 0.055  Rape2 20.588 0.000  Robbery 6.541 0.000  Assault 0.650 0.000  R-Squared: 0.9155839697176992 |  |
| Training Iteration 3  Training Set Same as above  Remove Murder  Sum Roberry and Assualt | Property\_Crime ~ Population+Rape2+Robbery\_Assault\_Sum  Parameter P-Value  Intercept -9.773 0.349  Population 0.015 0.000  Rape2 18.322 0.000  Robbery\_Assault\_Sum 2.630 0.000  R-Squared: 0.9085913599218914 |  |
| Training Iteration 4  Training Set Same as above  Square Population | Property\_Crime ~ Population\_Square+Rape2+Robbery\_Assault\_Sum  Parameter P-Value  Intercept 179.441 0.000  Population\_Square 0.000 0.000  Rape2 24.452 0.000  Robbery\_Assault\_Sum 2.987 0.000  R-Squared: 0.8797021003954226 |  |
| Training Iteration 5  Training Set Same as above  Added new binary variables to account for variability due to town type:  small = pop >15k  large = pop >75k | Property\_Crime ~ Population\_Square+Rape2+Robbery\_Assault\_Sum+small+large  Parameter P-Value  Intercept 422.303 0.000  Population\_Square 0.000 0.000  Rape2 19.148 0.000  Robbery\_Assault\_Sum 2.680 0.000  small -349.988 0.000  large 534.517 0.000  R-Squared: 0.9033201467740077 |  |
| Training Iteration 5  Same model. Changed Training Group to include population up to 1 million instead of 400k  Cities processing: 1985  Cities Removed due to pop above 1000000:  City Population  216 New York 8396126.000  1351 Los Angeles 3878725.000  684 Chicago3, 4 2720554.000  1470 San Diego 1349306.000 | Property\_Crime ~ Population\_Square+Rape2+Robbery\_Assault\_Sum+small+large  Parameter P-Value  Intercept 454.003 0.000  Population\_Square 0.000 0.000  Rape2 10.433 0.000  Robbery\_Assault\_Sum 3.512 0.000  small -378.559 0.000  large 942.844 0.000  R-Squared: 0.9321407398138103 | Population\_Square Rape2 Robbery\_Assault\_Sum small \  Population\_Square 1.000 0.667 0.715 -0.154  Rape2 0.667 1.000 0.700 -0.311  Robbery\_Assault\_Sum 0.715 0.700 1.000 -0.272  small -0.154 -0.311 -0.272 1.000  large 0.346 0.513 0.495 -0.399  large  Population\_Square 0.346  Rape2 0.513  Robbery\_Assault\_Sum 0.495  small -0.399  large 1.000 |

Testing

|  |  |  |
| --- | --- | --- |
| Test Group 1 - New York  R Squared: 0.9037200124926436  Cities processing: 347  Cities Removed due to pop above 1000000:  City Population  216 New York 8396126.000 | Test Group 2 - Georgia  R Squared: 0.9213296871216533  Cities processing: 253  Cities Removed due to pop above 1000000:  None | Test Group 3 – Illinois  R Squared: 0.7963358715274659  Cities processing: 505  Cities Removed due to pop above 1000000:  City Population  77 Chicago3, 4 2720554.000 |
| Test Group 4 – Alabama  R Squared: 0.9315341317565912  Cities processing: 196  Cities Removed due to pop above 1000000:  None | Test Group 5 – Washington  R Squared: 0.8238075138192604  Cities processing: 182  Cities Removed due to pop above 1000000:  None | Test Group 6 – Ohio  R Squared: 0.9400928218203343  Cities processing: 343  Cities Removed due to pop above 1000000:  None |