MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

OVERVIEW

Explore, analyze and model a data set containing information on crime for various neighborhoods of major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

OBJECTIVE

Objective is to build binary logistic regression model. Through this model, it is set to find out whether the neighborhood will be at risk for high crime levels.

DATA SET

There are two data sets given. They are:

05/10/2017 12:37 AM 2,542 crime-evaluation-data.csv
05/10/2017 12:37 AM 29,715 crime-training-data.csv

- zn: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- indus: proportion of non-retail business acres per suburb (predictor variable)
- chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
- nox: nitrogen oxides concentration (parts per 10 million) (predictor variable)
- rm: average number of rooms per dwelling (predictor variable)
- age: proportion of owner-occupied units built prior to 1940 (predictor variable)
- dis: weighted mean of distances to five Boston employment centers (predictor variable)
- rad: index of accessibility to radial highways (predictor variable)
- tax: full-value property-tax rate per $10,000 (predictor variable)
- ptratio: pupil-teacher ratio by town (predictor variable)
- black: $1000(B_k - 0.63)_2$ where B_k is the proportion of blacks by town (predictor variable)
- lstat: lower status of the population (percent) (predictor variable)
- medv: median value of owner-occupied homes in $1000s (predictor variable)

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

SAMPLE DATA LOADED IN EXCEL

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	369.3	3.7	50	1
0	19.58	1	0.871	5.403	100	1.3216	5	403	14.7	396.9	26.82	13.4	1
0	18.1	0	0.74	6.485	100	1.9784	24	666	20.2	386.73	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	374.71	5.19	23.7	0
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	394.12	4.82	37.9	0
0	8.56	0	0.52	6.781	71.3	2.8561	5	384	20.9	395.58	7.67	26.5	0
0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59	5	1
0	18.1	0	0.693	4.519	100	1.6582	24	666	20.2	88.27	36.98	7	1

DATA EXPLORATION

Various exploration of data set has been done through R Markdown and the program and the output has been attached

a) Summary of Statistics
b) Correlations of the data
c) Number of Rows and Columns
d) Structure of the data set
e) List all the variables of my data set
f) Statistical description of the data using additional packages install.packages("pastecs")

Sample from the output is given below:

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 - HOMEWORK 3 WEEK8

```
## 'data.frame':    466 obs. of  14 variables:
##  $ zn     : num  0 0 0 30 0 0 0 0 0 80 ...
##  $ indus  : num  19.58 19.58 18.1 4.93 2.46 ...
##  $ chas   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.605 0.871 0.74 0.428 0.488 0.52 0.693 0.693 0.515 0.392 ...
##  $ rm     : num  7.93 5.4 6.49 6.39 7.16 ...
##  $ age    : num  96.2 100 100 7.8 92.2 71.3 100 100 38.1 19.1 ...
##  $ dis    : num  2.05 1.32 1.98 7.04 2.7 ...
##  $ rad    : int  5 5 24 6 3 5 24 24 5 1 ...
##  $ tax    : int  403 403 666 300 193 384 666 666 224 315 ...
##  $ ptratio: num  14.7 14.7 20.2 16.6 17.8 20.9 20.2 20.2 20.2 16.4 ...
##  $ black  : num  369 397 387 375 394 ...
##  $ lstat  : num  3.7 26.82 18.85 5.19 4.82 ...
##  $ medv   : num  50 13.4 15.4 23.7 37.9 26.5 5 7 22.2 20.9 ...
##  $ target : int  1 1 1 0 0 0 1 1 0 0 ...
```

```
names(trgData)
```

```
##  [1] "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
##  [8] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"    "target"
```

```
head(trgData)
```

```
##   zn indus chas   nox    rm   age    dis rad tax ptratio  black lstat medv
## 1  0 19.58    0 0.605 7.929  96.2 2.0459   5 403    14.7 369.30  3.70 50.0
## 2  0 19.58    1 0.871 5.403 100.0 1.3216   5 403    14.7 396.90 26.82 13.4
## 3  0 18.10    0 0.740 6.485 100.0 1.9784  24 666    20.2 386.73 18.85 15.4
```

```
stat.desc(trgData)
```

```
##                      zn        indus         chas          nox
## nbr.val      466.000000  466.0000000 466.00000000 4.660000e+02
## nbr.null     339.000000    0.0000000 433.00000000 0.000000e+00
## nbr.na         0.000000    0.0000000   0.00000000 0.000000e+00
## min            0.000000    0.4600000   0.00000000 3.890000e-01
## max          100.000000   27.7400000   1.00000000 8.710000e-01
## range        100.000000   27.2800000   1.00000000 4.820000e-01
## sum         5395.000000 5174.9400000  33.00000000 2.583087e+02
## median         0.000000    9.6900000   0.00000000 5.380000e-01
```

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 - HOMEWORK 3 WEEK8

```
## [1] "Number of rows of Trainng Data Set->>>>>: 466"
```

```
print(paste0("Number of columns of Trainng Data Set->>>>>: ", ncol(trgData)))
```

```
## [1] "Number of columns of Trainng Data Set->>>>>: 14"
```

```
cor(trgData)
```

```
##                 zn       indus        chas         nox          rm
## zn       1.00000000 -0.53826643 -0.04016203 -0.51704518  0.31981410
## indus   -0.53826643  1.00000000  0.06118317  0.75963008 -0.39271181
## chas    -0.04016203  0.06118317  1.00000000  0.09745577  0.09050979
## nox     -0.51704518  0.75963008  0.09745577  1.00000000 -0.29548972
## rm       0.31981410 -0.39271181  0.09050979 -0.29548972  1.00000000
## age     -0.57258054  0.63958182  0.07888366  0.73512782 -0.23281251
## dis      0.66012434 -0.70361886 -0.09657711 -0.76888404  0.19901584
## rad     -0.31548119  0.60062839 -0.01590037  0.59582984 -0.20844570
## tax     -0.31928408  0.73222922 -0.04676476  0.65387804 -0.29693430
## ptratio -0.39103573  0.39468980 -0.12866058  0.17626871 -0.36034706
## black    0.17941504 -0.35813561  0.04444450 -0.38015487  0.13266756
## lstat   -0.43299252  0.60711023 -0.05142322  0.59624264 -0.63202445
## medv     0.37671713 -0.49617432  0.16156528 -0.43012267  0.70533679
## target  -0.43168176  0.60485074  0.08004187  0.72610622 -0.15255334
##                age         dis         rad         tax    ptratio
## zn      -0.57258054  0.66012434 -0.31548119 -0.31928408 -0.3910357
## indus    0.63958182 -0.70361886  0.60062839  0.73222922  0.3946898
## chas     0.07888366 -0.09657711 -0.01590037 -0.04676476 -0.1286606
```

DATA PREPARATION

1. In order to get the analysis, We have installed funModelling package for this. (viz. install.packages("funModeling")). This gives detailed report of about missing any data. – Please refer the output of RMARKDOWN file.

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

```
## funModeling v.1.6.2 :)
## Documentation at livebook.datascienceheroes.com

##    variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1        zn     339   72.75    0    0     0     0 numeric     26
## 2     indus       0    0.00    0    0     0     0 numeric     73
## 3      chas     433   92.92    0    0     0     0 integer      2
## 4       nox       0    0.00    0    0     0     0 numeric     79
## 5        rm       0    0.00    0    0     0     0 numeric    419
## 6       age       0    0.00    0    0     0     0 numeric    333
## 7       dis       0    0.00    0    0     0     0 numeric    380
## 8       rad       0    0.00    0    0     0     0 integer      9
## 9       tax       0    0.00    0    0     0     0 integer     63
## 10  ptratio       0    0.00    0    0     0     0 numeric     46
## 11    black       0    0.00    0    0     0     0 numeric    331
## 12    lstat       0    0.00    0    0     0     0 numeric    424
## 13     medv       0    0.00    0    0     0     0 numeric    218
## 14   target     237   50.86    0    0     0     0 integer      2

##    variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1        zn      33    82.5    0    0     0     0 integer      6
## 2     indus       0     0.0    0    0     0     0 numeric     22
## 3      chas      38    95.0    0    0     0     0 integer      2
## 4       nox       0     0.0    0    0     0     0 numeric     28
## 5        rm       0     0.0    0    0     0     0 numeric     40
## 6       age       0     0.0    0    0     0     0 numeric     39
## 7       dis       0     0.0    0    0     0     0 numeric     40
## 8       rad       0     0.0    0    0     0     0 integer      9
## 9       tax       0     0.0    0    0     0     0 integer     21
## 10  ptratio       0     0.0    0    0     0     0 numeric     17
## 11    black       0     0.0    0    0     0     0 numeric     32
## 12    lstat       0     0.0    0    0     0     0 numeric     40
## 13     medv       0     0.0    0    0     0     0 numeric     37
```

2. Checking for NULL or Infinite numbers. Please refer the output in RMarkdown file. Sample is given below

```
##      zn   indus    chas     nox      rm     age     dis     rad     tax
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## ptratio   black   lstat    medv  target
##   FALSE   FALSE   FALSE   FALSE   FALSE

##      zn   indus    chas     nox      rm     age     dis     rad     tax
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## ptratio   black   lstat    medv
##   FALSE   FALSE   FALSE   FALSE
```

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

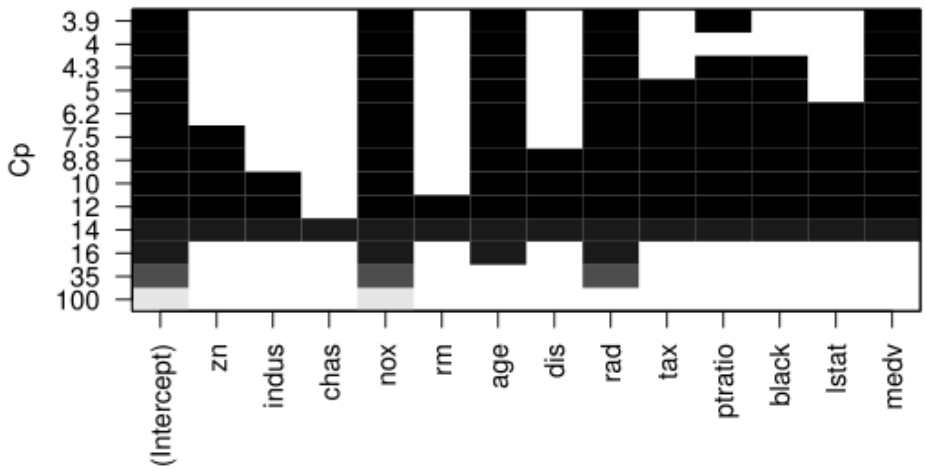
3. Put the Data into buckets

DATA VISUALIZATION

Sample data visualization has been given in RMARKDOWN output.

BUILD MODEL & SELECT MODEL

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8




```
##  (Intercept)          nox          age          rad      ptratio
## -1.412836094  1.956694224  0.003531713  0.017106647  0.012716341
##         medv
##  0.008021190

##
## Call:
## glm(formula = target ~ nox + age + rad + ptratio + medv, family = binomial,
##     data = trgData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.96654  -0.29783  -0.03987   0.00769   2.80829
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.936540   3.683449  -6.770 1.29e-11 ***
## nox          25.334778   4.084106   6.203 5.53e-10 ***
## age           0.019403   0.009308   2.085  0.03711 *
## rad           0.512600   0.114818   4.464 8.03e-06 ***
## ptratio       0.274193   0.098737   2.777  0.00549 **
## medv          0.085445   0.027979   3.054  0.00226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Please see the output of the markdown file.

CONCLUSION

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

Rmarkdown file

Output pdf file

Data file

This word document converted to PDF file

Are all available in https://github.com/muthukumars/DATA-621/tree/master/Week8-Homework3

THANK YOU