COMPAS Data Modeling

Lucius Bynum, Preethi Seshadri

19 November 2017

Compas Scores Two Years

```
compas_scores <- read.csv('compas-scores-two-years.csv')
summary(compas_scores)
```

```
##        id                     name              first
##  Min.   :    1   anthony smith    :   3   michael    : 149
##  1st Qu.: 2735   angel santiago   :   2   christopher: 109
##  Median : 5510   anthony gonzalez :   2   james      :  84
##  Mean   : 5501   anthony louis    :   2   anthony    :  83
##  3rd Qu.: 8246   brandon whitfield:   2   robert     :  76
##  Max.   :11001   carlos vasquez   :   2   john       :  74
##                  (Other)          :7201   (Other)    :6639
##        last      compas_screening_date     sex              dob
##  williams:  83   2013-02-20:  32       Female:1395   1987-02-04:   5
##  johnson :  76   2013-03-20:  32       Male  :5819   1987-12-21:   5
##  brown   :  68   2013-02-07:  31                     1989-04-27:   5
##  smith   :  65   2013-04-20:  30                     1989-08-31:   5
##  jones   :  57   2013-01-03:  29                     1990-02-22:   5
##  davis   :  46   2013-04-25:  28                     1990-05-02:   5
##  (Other) :6819   (Other)   :7032                     (Other)   :7184
##       age                 age_cat                    race
##  Min.   :18.0   25 - 45        :4109   African-American:3696
##  1st Qu.:25.0   Greater than 45:1576   Asian           :  32
##  Median :31.0   Less than 25   :1529   Caucasian       :2454
##  Mean   :34.8                          Hispanic        : 637
##  3rd Qu.:42.0                          Native American :  18
##  Max.   :96.0                          Other           : 377
##
##  juv_fel_count     decile_score   juv_misd_count   juv_other_count
##  Min.   : 0.000   Min.   : 1.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 2.00   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 4.00   Median : 0.000   Median : 0.000
##  Mean   : 0.067   Mean   : 4.51   Mean   : 0.091   Mean   : 0.109
##  3rd Qu.: 0.000   3rd Qu.: 7.00   3rd Qu.: 0.000   3rd Qu.: 0.000
##  Max.   :20.000   Max.   :10.00   Max.   :13.000   Max.   :17.000
##
##   priors_count   days_b_screening_arrest              c_jail_in
##  Min.   : 0.00   Min.   :-414.0                           : 307
##  1st Qu.: 0.00   1st Qu.:  -1.0          2013-01-01 01:31:55:   1
##  Median : 2.00   Median :  -1.0          2013-01-01 03:16:15:   1
##  Mean   : 3.47   Mean   :   3.3          2013-01-01 03:28:03:   1
##  3rd Qu.: 5.00   3rd Qu.:   0.0          2013-01-01 04:17:22:   1
##  Max.   :38.00   Max.   :1057.0          2013-01-01 04:29:04:   1
##                  NA's   :307             (Other)            :6902
##                c_jail_out         c_case_number     c_offense_date
```

```
##                    : 307               :  22            :1159
##  2013-09-12 10:31:00:   3   00004068CF10A:   1   2013-01-14:  26
##  2013-09-14 05:58:00:   3   00022077MM10A:   1   2013-02-22:  26
##  2013-09-28 02:10:00:   3   01004839CF10A:   1   2013-03-01:  24
##  2013-02-06 10:01:51:   2   01006487CF10D:   1   2013-01-11:  23
##  2013-06-13 10:32:00:   2   01007205MM10A:   1   2013-02-16:  23
##  (Other)            :6894   (Other)      :7187   (Other)   :5933
##     c_arrest_date  c_days_from_compas c_charge_degree
##            :6077   Min.   :   0       F:4666
##  2013-02-06:   9   1st Qu.:   1       M:2548
##  2013-03-22:   8   Median :   1
##  2013-05-15:   8   Mean   :  58
##  2013-01-10:   7   3rd Qu.:   2
##  2013-01-11:   7   Max.   :9485
##  (Other)   :1098   NA's   :22
##                        c_charge_desc     is_recid           r_case_number
##  Battery                      :1156   Min.   :0.000                :3743
##  arrest case no charge        :1137   1st Qu.:0.000   13000349MM10A:   1
##  Possession of Cocaine        : 474   Median :0.000   13000445MM20A:   1
##  Grand Theft in the 3rd Degree: 425   Mean   :0.481   13000677MM20A:   1
##  Driving While License Revoked: 200   3rd Qu.:1.000   13000758MM30A:   1
##  Driving Under The Influence  : 135   Max.   :1.000   13000785MM30A:   1
##  (Other)                      :3687                   (Other)      :3466
##  r_charge_degree r_days_from_arrest    r_offense_date
##         :3743    Min.   : -1                  :3743
##  (M1)   :1201    1st Qu.:  0        2014-12-08:  12
##  (M2)   :1107    Median :  0        2015-01-28:  11
##  (F3)   : 892    Mean   : 20        2014-09-15:  10
##  (F2)   : 168    3rd Qu.:  1        2014-10-17:  10
##  (F1)   :  51    Max.   :993        2015-02-10:  10
##  (Other):  52    NA's   :4898       (Other)   :3418
##                             r_charge_desc       r_jail_in
##                                    :3801             :4898
##  Driving License Suspended        : 258   2014-05-27:   9
##  Possess Cannabis/20 Grams Or Less: 253   2013-11-22:   8
##  Resist/Obstruct W/O Violence     : 201   2014-06-05:   8
##  Battery                          : 192   2014-07-10:   8
##  Operating W/O Valid License      : 172   2014-10-17:   8
##  (Other)                          :2337   (Other)   :2275
##       r_jail_out   violent_recid  is_violent_recid       vr_case_number
##            :4898   Mode:logical   Min.   :0.000                :6395
##  2014-02-18:   9   NA's:7214      1st Qu.:0.000    13001383CF10A:   1
##  2014-12-09:   9                  Median :0.000    13001876CF10A:   1
##  2015-05-15:   9                  Mean   :0.114    13002119CF10A:   1
##  2013-11-13:   8                  3rd Qu.:0.000    13002546CF10A:   1
##  2014-07-11:   8                  Max.   :1.000    13003421CF10A:   1
##  (Other)   :2273                                   (Other)      : 814
##  vr_charge_degree   vr_offense_date                         vr_charge_desc
##         :6395              :6395                                :6395
##  (M1)   : 344     2015-08-15:   6   Battery                          : 329
##  (F3)   : 228     2013-11-14:   4   Battery on Law Enforc Officer :  38
##  (F2)   : 162     2014-02-18:   4   Felony Battery (Dom Strang)   :  38
##  (F1)   :  38     2014-10-29:   4   Aggravated Assault W/Dead Weap:  37
##  (M2)   :  19     2014-12-26:   4   Aggrav Battery w/Deadly Weapon:  34
```

```
##  (Other):  28     (Other)   : 797   (Other)                       : 343
##           type_of_assessment decile_score.1   score_text
##  Risk of Recidivism:7214     Min.   : 1.00   High  :1403
##                              1st Qu.: 2.00   Low   :3897
##                              Median : 4.00   Medium:1914
##                              Mean   : 4.51
##                              3rd Qu.: 7.00
##                              Max.   :10.00
##
##     screening_date       v_type_of_assessment v_decile_score
##  2013-02-20:  32   Risk of Violence:7214      Min.   : 1.00
##  2013-03-20:  32                             1st Qu.: 1.00
##  2013-02-07:  31                             Median : 3.00
##  2013-04-20:  30                             Mean   : 3.69
##  2013-01-03:  29                             3rd Qu.: 5.00
##  2013-04-25:  28                             Max.   :10.00
##  (Other)   :7032
##  v_score_text    v_screening_date      in_custody       out_custody
##  High  : 714   2013-02-20:  32              : 236             : 236
##  Low   :4761   2013-03-20:  32    2013-02-22:  20   2020-01-01:  61
##  Medium:1739   2013-02-07:  31    2013-12-12:  20   2013-05-14:  25
##                2013-04-20:  30    2014-01-04:  20   2014-02-04:  24
##                2013-01-03:  29    2014-01-22:  20   2013-11-26:  23
##                2013-04-25:  28    2013-01-27:  19   2013-02-15:  21
##                (Other)   :7032    (Other)   :6879   (Other)   :6824
##  priors_count.1      start            end           event
##  Min.   : 0.00   Min.   :  0.0   Min.   :   0   Min.   :0.000
##  1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.: 148   1st Qu.:0.000
##  Median : 2.00   Median :  0.0   Median : 530   Median :0.000
##  Mean   : 3.47   Mean   : 11.5   Mean   : 553   Mean   :0.383
##  3rd Qu.: 5.00   3rd Qu.:  1.0   3rd Qu.: 914   3rd Qu.:1.000
##  Max.   :38.00   Max.   :937.0   Max.   :1186   Max.   :1.000
##
##  two_year_recid
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.451
##  3rd Qu.:1.000
##  Max.   :1.000
##
```

```
nrow(compas_scores)
```

```
## [1] 7214
```

```
length(unique(compas_scores$c_charge_desc))
```

```
## [1] 438
```

Now we will perform the same cleaning/filtering they performed in order to maintain the same data.

```
compas_scores <- compas_scores %>%
  filter(days_b_screening_arrest <= 30) %>%
  filter(days_b_screening_arrest >= -30) %>%
  filter(is_recid != -1) %>%
```

```
  filter(c_charge_degree != "O") %>%
  filter(score_text != 'N/A')
```

We want to convert the jail in and out to time to how long they spent in jail. We will convert those to datetime format and take the difference in days.

```
compas_scores$c_jail_out <- strptime(compas_scores$c_jail_out, format = '%Y-%m-%d %H:%M:%S')
compas_scores$c_jail_in <- strptime(compas_scores$c_jail_in, format = '%Y-%m-%d %H:%M:%S')
compas_scores$time_spent <- difftime(compas_scores$c_jail_out, compas_scores$c_jail_in, units='days')
compas_scores$time_spent <- as.numeric(compas_scores$time_spent)
```

We will only use a subset of features. We will use sex, age, race, juvenile felony count, juvenile misdemeanor count, juvenile other count, priors count, the charge degree, and how long they were in jail for (for the crime directly linked to the Compas score). The response will be whether or not they recidivated. The goal with out model is to predict whether or not individuals will recidivate based off of the selected features.

```
#View(compas_scores)
names(compas_scores)
```

```
##  [1] "id"                     "name"
##  [3] "first"                  "last"
##  [5] "compas_screening_date"  "sex"
##  [7] "dob"                    "age"
##  [9] "age_cat"                "race"
## [11] "juv_fel_count"          "decile_score"
## [13] "juv_misd_count"         "juv_other_count"
## [15] "priors_count"           "days_b_screening_arrest"
## [17] "c_jail_in"              "c_jail_out"
## [19] "c_case_number"          "c_offense_date"
## [21] "c_arrest_date"          "c_days_from_compas"
## [23] "c_charge_degree"        "c_charge_desc"
## [25] "is_recid"               "r_case_number"
## [27] "r_charge_degree"        "r_days_from_arrest"
## [29] "r_offense_date"         "r_charge_desc"
## [31] "r_jail_in"              "r_jail_out"
## [33] "violent_recid"          "is_violent_recid"
## [35] "vr_case_number"         "vr_charge_degree"
## [37] "vr_offense_date"        "vr_charge_desc"
## [39] "type_of_assessment"     "decile_score.1"
## [41] "score_text"             "screening_date"
## [43] "v_type_of_assessment"   "v_decile_score"
## [45] "v_score_text"           "v_screening_date"
## [47] "in_custody"             "out_custody"
## [49] "priors_count.1"         "start"
## [51] "end"                    "event"
## [53] "two_year_recid"         "time_spent"
```

```
df_recid <- compas_scores[, c(6, 8, 10, 11, 13, 14, 15, 23, 25, 54)]
```

Split up our train and test data. Roughly a 75-25 percent split.

```
train <- sample(nrow(df_recid), 4500)
df.train <- df_recid[train, ]
df.test <- df_recid[-train, ]
```

Fit a logistic regression model.

```r
model <- glm(is_recid ~ ., data = df.train, family = binomial(link='logit'))
summary(model)
```

```
##
## Call:
## glm(formula = is_recid ~ ., family = binomial(link = "logit"),
##     data = df.train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.763  -1.015  -0.555   1.082   2.509
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.66032    0.12886    5.12  3.0e-07 ***
## sexMale             0.36470    0.08383    4.35  1.4e-05 ***
## age                -0.04314    0.00325  -13.29  < 2e-16 ***
## raceAsian          -0.04772    0.45979   -0.10  0.91734
## raceCaucasian      -0.09465    0.07436   -1.27  0.20304
## raceHispanic       -0.31996    0.12539   -2.55  0.01072 *
## raceNative American -0.05312    0.72482   -0.07  0.94158
## raceOther          -0.30208    0.14501   -2.08  0.03724 *
## juv_fel_count       0.04147    0.11271    0.37  0.71292
## juv_misd_count     -0.06295    0.08825   -0.71  0.47566
## juv_other_count     0.31903    0.09077    3.51  0.00044 ***
## priors_count        0.16550    0.01012   16.35  < 2e-16 ***
## c_charge_degreeM   -0.17017    0.06934   -2.45  0.01412 *
## time_spent          0.00238    0.00081    2.94  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6229.3  on 4499  degrees of freedom
## Residual deviance: 5452.5  on 4486  degrees of freedom
## AIC: 5481
##
## Number of Fisher Scoring iterations: 4
```

Look at the training performance for the model created above:

```r
preds <- predict(model, newdata = df.train, type = "response")
# use caret and compute a confusion matrix
confusionMatrix(data = as.numeric(preds>0.5), reference = df.train$is_recid)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1762  849
##          1  589 1300
##
##                Accuracy : 0.68
##                  95% CI : (0.667, 0.694)
##     No Information Rate : 0.522
```

```
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.356
##  Mcnemar's Test P-Value : 8.49e-12
##
##             Sensitivity : 0.749
##             Specificity : 0.605
##          Pos Pred Value : 0.675
##          Neg Pred Value : 0.688
##              Prevalence : 0.522
##          Detection Rate : 0.392
##    Detection Prevalence : 0.580
##       Balanced Accuracy : 0.677
##
##        'Positive' Class : 0
##
```

Look at the testing performance:

```
preds <- predict(model, newdata = df.test, type = "response")
# use caret and compute a confusion matrix
confusionMatrix(data = as.numeric(preds>0.5), reference = df.test$is_recid)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 635 326
##          1 196 515
##
##                Accuracy : 0.688
##                  95% CI : (0.665, 0.71)
##     No Information Rate : 0.503
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.376
##  Mcnemar's Test P-Value : 1.64e-08
##
##             Sensitivity : 0.764
##             Specificity : 0.612
##          Pos Pred Value : 0.661
##          Neg Pred Value : 0.724
##              Prevalence : 0.497
##          Detection Rate : 0.380
##    Detection Prevalence : 0.575
##       Balanced Accuracy : 0.688
##
##        'Positive' Class : 0
##
```

The training and testing performance are fairly comparable for the various metrics. The error for both could be significantly improved.