Model Selection and Prediction

Kanishk Deshwal

2026-02-08

```
comp <- read.csv("../data/indian_companies_transformed.csv", skipNul = TRUE)
comp$INDUSTRY = factor(comp$INDUSTRY)
comp$INDIA.HQ = factor(comp$INDIA.HQ)
```

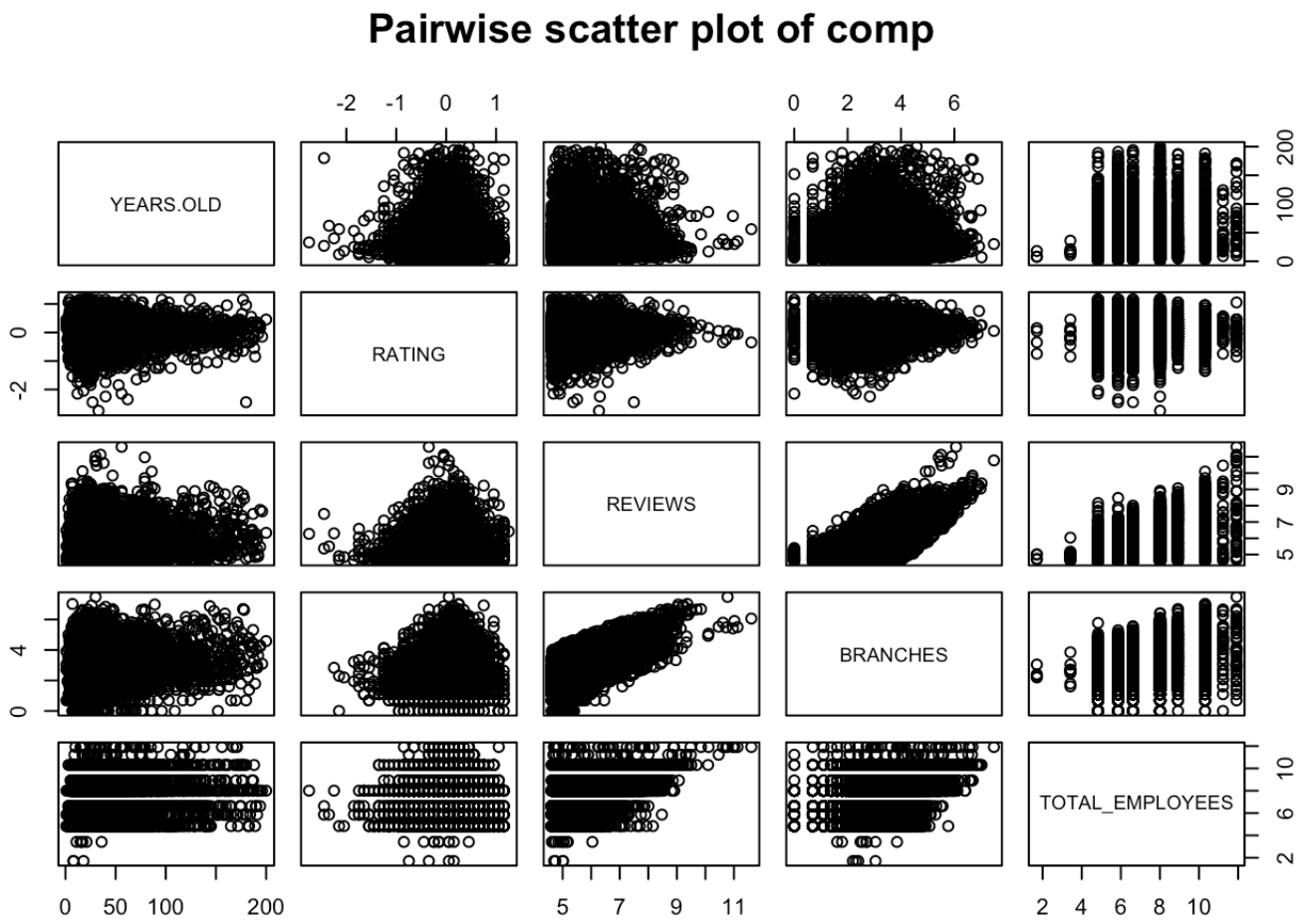
```
lmod = lm(PACKAGE ~ YEARS.OLD + INDUSTRY + INDIA.HQ + TOTAL_EMPLOYEES + BRANCHES + RA
TING + I(RATING^2) + REVIEWS, comp)
summary(lmod)
```

```
##
## Call:
## lm(formula = PACKAGE ~ YEARS.OLD + INDUSTRY + INDIA.HQ + TOTAL_EMPLOYEES +
##     BRANCHES + RATING + I(RATING^2) + REVIEWS, data = comp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8594 -0.1859  0.0265  0.2296  2.5620
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       -9.800e+00  4.353e-02 -225.135  < 2e-16 ***
## YEARS.OLD                         -5.603e-05  1.648e-04   -0.340 0.733963
## INDUSTRYEducation & Training      -3.014e-01  3.737e-02   -8.065 8.42e-16 ***
## INDUSTRYEngineering & Construction -8.144e-02  3.190e-02   -2.553 0.010686 *
## INDUSTRYFinancial Services         7.073e-02  3.639e-02    1.944 0.051981 .
## INDUSTRYHealthcare                -1.325e-01  3.601e-02   -3.680 0.000235 ***
## INDUSTRYIndustrial Machinery       3.113e-02  3.303e-02    0.942 0.345986
## INDUSTRYInternet                  -7.064e-02  3.677e-02   -1.921 0.054745 .
## INDUSTRYIT Services & Consulting   1.221e-01  2.700e-02    4.522 6.22e-06 ***
## INDUSTRYOther                     -5.426e-02  2.436e-02   -2.227 0.025951 *
## INDUSTRYPharma                    -5.822e-02  3.396e-02   -1.714 0.086498 .
## INDUSTRYReal Estate                2.298e-02  4.014e-02    0.573 0.566961
## INDUSTRYSoftware Product           1.883e-01  3.586e-02    5.251 1.56e-07 ***
## INDIA.HQChennai                   -9.712e-02  2.164e-02   -4.488 7.30e-06 ***
## INDIA.HQMumbai                    -1.572e-01  1.817e-02   -8.652  < 2e-16 ***
## INDIA.HQNew Delhi                 -2.459e-01  2.147e-02  -11.450  < 2e-16 ***
## INDIA.HQOther                     -2.009e-01  1.486e-02  -13.522  < 2e-16 ***
## INDIA.HQPune                      -4.040e-02  2.191e-02   -1.844 0.065242 .
## TOTAL_EMPLOYEES                   -2.299e-03  4.103e-03   -0.560 0.575339
## BRANCHES                          -5.467e-02  6.559e-03   -8.335  < 2e-16 ***
## RATING                            -5.029e-01  1.332e-02  -37.750  < 2e-16 ***
## I(RATING^2)                       -6.265e-01  1.902e-02  -32.946  < 2e-16 ***
## REVIEWS                            1.013e+00  8.351e-03  121.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4342 on 7753 degrees of freedom
##   (1647 observations deleted due to missingness)
## Multiple R-squared:  0.8305, Adjusted R-squared:   0.83
## F-statistic:  1727 on 22 and 7753 DF,  p-value: < 2.2e-16
```

Checking for Inflated Variance - Stabilizing β

Now, checking for multicoliearity using correlation matrix and calculating Condition number.

```
X = model.matrix(lmod)
cat_vars = c("YEARS.OLD","RATING","REVIEWS","BRANCHES","TOTAL_EMPLOYEES")
pairs(comp[cat_vars], labels = cat_vars, main="Pairwise scatter plot of comp")
```




```
cat("Correlation Matrix:\n")
```

```
## Correlation Matrix:
```

```
cor(X[, cat_vars])
```

```
##                  YEARS.OLD     RATING   REVIEWS   BRANCHES TOTAL_EMPLOYEES
## YEARS.OLD       1.00000000 0.01010944 0.1223101 0.09844222      0.19902371
## RATING          0.01010944 1.00000000 0.0697027 0.16506321      0.05653199
## REVIEWS         0.12231015 0.06970270 1.0000000 0.65146168      0.59629725
## BRANCHES        0.09844222 0.16506321 0.6514617 1.00000000      0.37548257
## TOTAL_EMPLOYEES 0.19902371 0.05653199 0.5962973 0.37548257      1.00000000
```

```
n = dim(X)[1]
p = dim(X)[2]

XtX = t(X)%*%X
lambda = eigen(XtX)$values
cat("\u03BB =", lambda, "\n")
```

```
## λ = 18364042 279703.6 8175.228 3551.718 2317.653 1966.791 1262.494 989.2817 730.05
63 661.1901 609.0224 500.9859 376.1625 327.0503 296.4258 262.7326 244.4292 242.0478 2
33.5391 215.2805 182.3725 118.0524 25.97879
```

```
K = sqrt(lambda[1]/lambda[p])
cat("Condition number is given by K = ", K)
```

```
## Condition number is given by K =  840.7651
```

High Condition Number states that this is an extreme case of Multi-collinearity which signals that the variance of estimators may be inflated. To further confirm this Variance Inflation Factor must be calculated.

```
library(faraway)
v = vif(lmod)
print(v)
```

```
##                          YEARS.OLD       INDUSTRYEducation & Training
##                           1.043861                           1.618579
## INDUSTRYEngineering & Construction         INDUSTRYFinancial Services
##                           2.047465                           1.673319
##                INDUSTRYHealthcare       INDUSTRYIndustrial Machinery
##                           1.664771                           1.854936
##                  INDUSTRYInternet   INDUSTRYIT Services & Consulting
##                           1.654398                           3.456023
##                     INDUSTRYOther                     INDUSTRYPharma
##                           6.117498                           1.821112
##               INDUSTRYReal Estate           INDUSTRYSoftware Product
##                           1.486156                           1.707909
##                  INDIA.HQChennai                     INDIA.HQMumbai
##                           1.429819                           1.805474
##                INDIA.HQNew Delhi                      INDIA.HQOther
##                           1.539695                           2.245544
##                     INDIA.HQPune                    TOTAL_EMPLOYEES
##                           1.429330                           1.596965
##                         BRANCHES                             RATING
##                           2.225271                           1.150603
##                       I(RATING^2)                            REVIEWS
##                           1.125056                           2.507384
```

```
cat("Average VIF for predictors: ", sum(v)/p, "\n")
```

```
## Average VIF for predictors:  1.878312
```

```
cat("Maximum value of VIF: ", max(v))
```

```
## Maximum value of VIF:  6.117498
```

Average and Maximum values of VIF depicts that the variance of each estimate is not inflated (atleast to a value that posses a concern). The huge condition number is due to design-matrix structure (dummy coding), not harmful collinearity in the regression coefficients. Furthermore, the pairwise correlations among predictors are all below 0.6, which is additional evidence that the independent variables do not exhibit strong linear relationships.

Model Selection

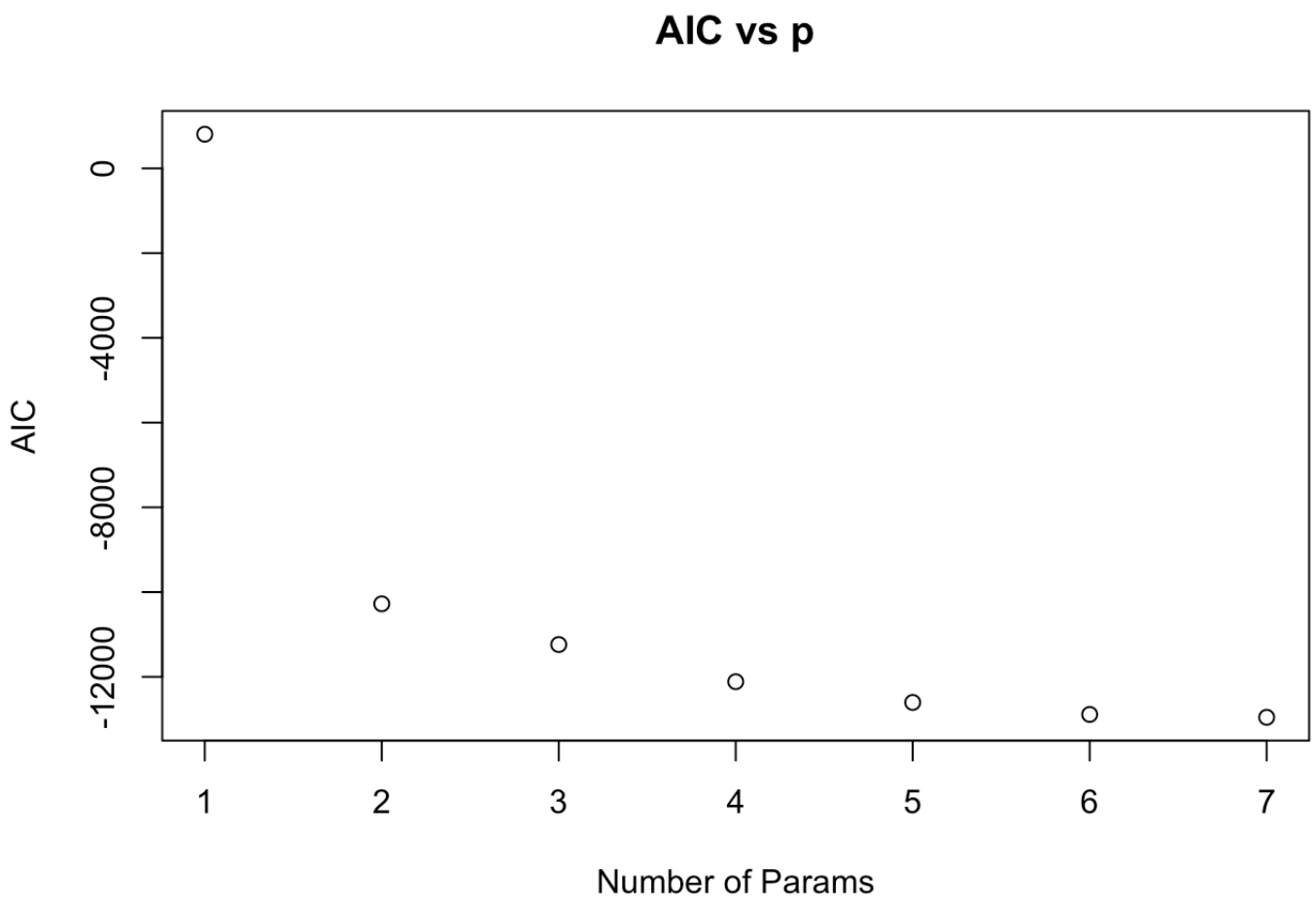
As the number of parameters p<<n, a balanced criterion is required to balance predictive power and complexity of the model. Thus, AIC is ideal criterion with Hybrid(Both) search.

```
null = lm(PACKAGE ~ 1, na.omit(comp))
out = step(null, scope = list(lower=~1, upper=formula(lmod)), direction = "both", tra
ce = FALSE)
cat("Order of parameters: ",out$anova[,1])
```

```
## Order of parameters:   + REVIEWS + RATING + I(RATING^2) + INDUSTRY + INDIA.HQ + BR
ANCHES
```

Here, Years.OLD and TOTAL_EMPLOYEES were dropped from the model as a result.

```
plot(out$anova[,6], xlab="Number of Params", ylab="AIC", main="AIC vs p")
```



Looking at the AIC vs Index plot, we see that AIC remains almost constant after 4. Replacing AIC with more strict criterion, BIC.

```
out = step(null, scope = list(lower=~1, upper=formula(lmod)), direction = "both", k=l
og(n), trace = FALSE)
cat("Order of parameters:", out$anova[,1], "\n")
```

```
## Order of parameters:  + REVIEWS + RATING + I(RATING^2) + INDUSTRY + INDIA.HQ + BRA
NCHES
```

```
cat("Adjusted R-squared: ", summary(out)$adj.r.squared)
```

```
## Adjusted R-squared:  0.8300548
```

```
summary(out)
```

```
##
## Call:
## lm(formula = PACKAGE ~ REVIEWS + RATING + I(RATING^2) + INDUSTRY +
##     INDIA.HQ + BRANCHES, data = na.omit(comp))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8601 -0.1850  0.0271  0.2302  2.5666
##
## Coefficients:
##                                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                      -9.804236   0.043075 -227.606  < 2e-16 ***
## REVIEWS                           1.010574   0.007357  137.370  < 2e-16 ***
## RATING                           -0.503053   0.013316  -37.778  < 2e-16 ***
## I(RATING^2)                      -0.626413   0.019013  -32.946  < 2e-16 ***
## INDUSTRYEducation & Training     -0.301205   0.037365   -8.061 8.68e-16 ***
## INDUSTRYEngineering & Construction -0.081468   0.031891   -2.555 0.010651 *
## INDUSTRYFinancial Services        0.070302   0.036379    1.933 0.053333 .
## INDUSTRYHealthcare               -0.132840   0.036008   -3.689 0.000227 ***
## INDUSTRYIndustrial Machinery      0.030920   0.033024    0.936 0.349153
## INDUSTRYInternet                 -0.070788   0.036758   -1.926 0.054169 .
## INDUSTRYIT Services & Consulting  0.121721   0.026993    4.509 6.60e-06 ***
## INDUSTRYOther                    -0.054445   0.024354   -2.236 0.025409 *
## INDUSTRYPharma                   -0.058538   0.033956   -1.724 0.084761 .
## INDUSTRYReal Estate               0.022775   0.040124    0.568 0.570305
## INDUSTRYSoftware Product          0.187972   0.035853    5.243 1.62e-07 ***
## INDIA.HQChennai                  -0.097095   0.021628   -4.489 7.25e-06 ***
## INDIA.HQMumbai                   -0.156967   0.018165   -8.641  < 2e-16 ***
## INDIA.HQNew Delhi                -0.245539   0.021464  -11.440  < 2e-16 ***
## INDIA.HQOther                    -0.200743   0.014852  -13.516  < 2e-16 ***
## INDIA.HQPune                     -0.040454   0.021906   -1.847 0.064828 .
## BRANCHES                         -0.054699   0.006555   -8.344  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4341 on 7755 degrees of freedom
## Multiple R-squared:  0.8305, Adjusted R-squared:  0.8301
## F-statistic:  1900 on 20 and 7755 DF,  p-value: < 2.2e-16
```

Prediction

Lets take examples of Indian companies with:

Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 4

Headquarter = Bangalore / Bengaluru
Industry = Software Product

```
## Average monthly package predicted:  0.74 L
```

Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 4
Headquarter = Bangalore / Bengaluru
Industry = Education & Training

```
## Average monthly package predicted:  0.45 L
```