Regression_Report

Bowen Liu

8/1/2019

Data Preprocessing and Model Selecting

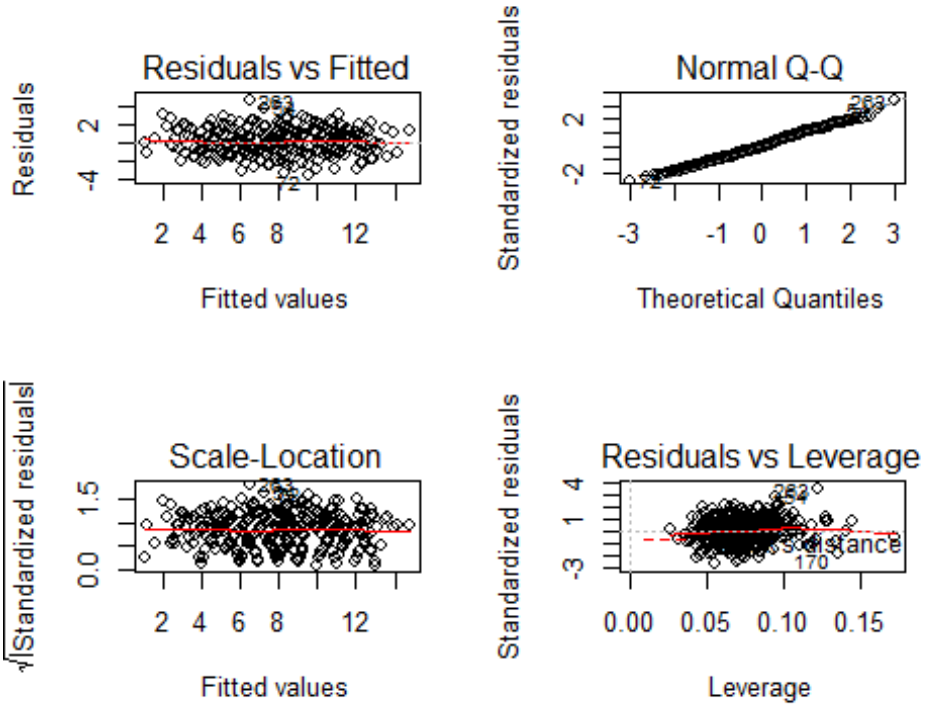
After viewing data, we don't see the need of any further cleaning. The train set of the data has 380 observations, 31 dependent variables, and 1 independent variable (Wins), as you can see in the dim function output below:

```
## [1] 380  32
```

So, we first run a linear model to visualize the data. Surprisingly, the assumtions of linearity, normalilty, and constant variance assumptions are all met very well, as you can see in the summary and plot below:

```
##
## Call:
## lm(formula = Wins ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5705 -1.0162 -0.0789  0.9942  4.4682
##
## Coefficients: (4 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12.0732281  3.6929277   3.269 0.001185 **
## ID                          -0.0005767  0.0006772  -0.852 0.395051
## Yards                       -0.0010514  0.0005722  -1.837 0.067002 .
## OffensivePlays              -0.0382241  0.0090049  -4.245 2.80e-05 ***
## TurnOversLost               -0.1205821  0.0197228  -6.114 2.58e-09 ***
## FumblesLost                 -0.0111965  0.0317473  -0.353 0.724543
## FirstDowns                   0.0299882  0.0087319   3.434 0.000665 ***
## PassesCompleted              0.0034943  0.0056849   0.615 0.539179
## PassesAttempted              0.0156055  0.0090996   1.715 0.087232 .
## YardsGainedPassing           0.0020855  0.0005484   3.803 0.000169 ***
## InterceptionsThrown                 NA         NA      NA       NA
## RushingAttempts              0.0337269  0.0086099   3.917 0.000108 ***
## YardsGainedRushing                  NA         NA      NA       NA
## PenaltiesCommitedByTeam      0.0132972  0.0114519   1.161 0.246372
## PenaltiesInYards            -0.0030049  0.0013524  -2.222 0.026924 *
## FirstDownsByPenalty         -0.0116435  0.0168043  -0.693 0.488837
## NumberOfDrives               0.0202147  0.0198995   1.016 0.310404
## OppYards                    -0.0001635  0.0005907  -0.277 0.782174
## OppOffensivePlays            0.0343151  0.0123382   2.781 0.005707 **
## OppTurnOversLost             0.1080579  0.0180442   5.989 5.23e-09 ***
```

```
## OppFumblesLost              -0.0187464  0.0292141  -0.642 0.521491
## OppFirstDowns               -0.0092511  0.0088419  -1.046 0.296150
## OppPassesCompleted          -0.0216525  0.0061177  -3.539 0.000455 ***
## OppPassesAttempted          -0.0080041  0.0128167  -0.625 0.532702
## OppYardsGainedPassing       -0.0011560  0.0006351  -1.820 0.069586 .
## OppInterceptionsThrown              NA         NA      NA       NA
## OppRushingAttempts          -0.0406338  0.0120225  -3.380 0.000807 ***
## OppYardsGainedRushing               NA         NA      NA       NA
## OppPenaltiesCommitedByTeam   0.0062637  0.0111429   0.562 0.574388
## OppPenaltiesInYards          0.0012886  0.0013327   0.967 0.334243
## OppFirstDownsByPenalty       0.0030725  0.0156386   0.196 0.844357
## OppNumberOfDrives           -0.0092546  0.0193598  -0.478 0.632924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 352 degrees of freedom
## Multiple R-squared:  0.8111, Adjusted R-squared:  0.7966
## F-statistic: 55.96 on 27 and 352 DF,  p-value: < 2.2e-16
```



Thus, we choose to use linear model, as it has perfect interpretability and good performance on this sepecfic data. Also, as this is a regression problem, linear regression would be great to deal with the overfitting problem. We will imporve our linear model, m1, in further steps.

Model Tuning

In the second step, we decide to use stepwise subset selection methods. In perticular, we use backward AIC for the selection. As you can see below, the result somehow matches with the significance of the fullmodel above:

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Wins ~ ID + Yards + OffensivePlays + TurnOversLost + FumblesLost +
##     FirstDowns + PassesCompleted + PassesAttempted + YardsGainedPassing +
##     InterceptionsThrown + RushingAttempts + YardsGainedRushing +
##     PenaltiesCommitedByTeam + PenaltiesInYards + FirstDownsByPenalty +
##     NumberOfDrives + OppYards + OppOffensivePlays + OppTurnOversLost +
##     OppFumblesLost + OppFirstDowns + OppPassesCompleted +
OppPassesAttempted +
##     OppYardsGainedPassing + OppInterceptionsThrown + OppRushingAttempts +
##     OppYardsGainedRushing + OppPenaltiesCommitedByTeam +
OppPenaltiesInYards +
##     OppFirstDownsByPenalty + OppNumberOfDrives
##
## Final Model:
## Wins ~ Yards + OffensivePlays + TurnOversLost + FirstDowns +
##     PassesAttempted + YardsGainedPassing + RushingAttempts +
##     PenaltiesInYards + OppOffensivePlays + OppTurnOversLost +
##     OppFirstDowns + OppPassesCompleted + OppYardsGainedPassing +
##     OppRushingAttempts + OppPenaltiesInYards
##
##
##                             Step Df   Deviance Resid. Df Resid. Dev
## 1                                                   352   704.3310
## 2       - OppYardsGainedRushing  0 0.00000000       352   704.3310
## 3      - OppInterceptionsThrown  0 0.00000000       352   704.3310
## 4          - YardsGainedRushing  0 0.00000000       352   704.3310
## 5         - InterceptionsThrown  0 0.00000000       352   704.3310
## 6      - OppFirstDownsByPenalty  1 0.07723622       353   704.4082
## 7                    - OppYards  1 0.22392698       354   704.6321
## 8                - FumblesLost  1 0.23520209       355   704.8673
## 9             - PassesCompleted  1 0.58806004       356   705.4554
## 10 - OppPenaltiesCommitedByTeam  1 0.68090276       357   706.1363
## 11          - OppNumberOfDrives  1 0.70556094       358   706.8418
## 12         - OppPassesAttempted  1 0.64402670       359   707.4859
## 13             - OppFumblesLost  1 0.48305471       360   707.9689
## 14        - FirstDownsByPenalty  1 1.12324563       361   709.0922
## 15                           - ID  1 1.16946953       362   710.2616
## 16             - NumberOfDrives  1 1.22561157       363   711.4872
## 17    - PenaltiesCommitedByTeam  1 3.11863835       364   714.6059
##         AIC
## 1  290.4893
```

```
## 2  290.4893
## 3  290.4893
## 4  290.4893
## 5  290.4893
## 6  288.5310
## 7  286.6517
## 8  284.7786
## 9  283.0955
## 10 281.4621
## 11 279.8416
## 12 278.1876
## 13 276.4470
## 14 275.0494
## 15 273.6756
## 16 272.3308
## 17 271.9928
```

As suggested by backwards AIC subset suggestions, there are 16 predictors in the final model. We than our model fit_first out of this. We can see the summary of the new model below:

```
##
## Call:
## lm(formula = Wins ~ Yards + OffensivePlays + TurnOversLost +
##     FirstDowns + PassesAttempted + YardsGainedPassing + RushingAttempts +
##     PenaltiesInYards + OppOffensivePlays + OppTurnOversLost +
##     OppFirstDowns + OppPassesCompleted + OppYardsGainedPassing +
##     OppRushingAttempts + OppPenaltiesInYards, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5040 -0.9974 -0.0890  1.0300  4.2078
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           14.1601264  3.3226715   4.262 2.59e-05 ***
## Yards                 -0.0008577  0.0004961  -1.729 0.084680 .
## OffensivePlays        -0.0359504  0.0083208  -4.321 2.01e-05 ***
## TurnOversLost         -0.1185666  0.0127397  -9.307  < 2e-16 ***
## FirstDowns             0.0250820  0.0066058   3.797 0.000172 ***
## PassesAttempted        0.0160009  0.0084222   1.900 0.058245 .
## YardsGainedPassing     0.0021336  0.0005017   4.253 2.69e-05 ***
## RushingAttempts        0.0316735  0.0081328   3.895 0.000117 ***
## PenaltiesInYards      -0.0015029  0.0005681  -2.645 0.008517 **
## OppOffensivePlays      0.0284877  0.0033311   8.552 3.40e-16 ***
## OppTurnOversLost       0.1051196  0.0125509   8.375 1.21e-15 ***
## OppFirstDowns         -0.0100716  0.0064157  -1.570 0.117324
## OppPassesCompleted    -0.0236346  0.0054353  -4.348 1.78e-05 ***
## OppYardsGainedPassing -0.0013281  0.0003671  -3.618 0.000339 ***
## OppRushingAttempts    -0.0364343  0.0044729  -8.146 6.10e-15 ***
```

```
## OppPenaltiesInYards    0.0015959  0.0006606   2.416 0.016193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.401 on 364 degrees of freedom
## Multiple R-squared:  0.8083, Adjusted R-squared:  0.8004
## F-statistic: 102.3 on 15 and 364 DF,  p-value: < 2.2e-16
```

As we can see in the summary of the new model above, some of the variables are not significant or of the highest level of significance. It would be helpful for us to generate new variables which are more representitive of the attributes in all these old variables, but have less redundancies. Especially for those variables selected by AIC but has less significance, it would be helpful to do manipulate them in a way - add some options of their combination of them interacting with other variables. So, we generate the combinations of all the variables, at pay close attentions to the conbinations related to Yards, PassesAttempted, OppFirstDowns, and OppPenaltiesInYards as they are not that significant. We found that four new iteractions related to these variables are worth consideraing:

- OffensivePlays and FirstDowns
- Yards and YardsGainedPassing
- PassesAttempted and RushingAttempts
- OppYardsGainedPassing and OppPassesCompleted

So, we add them as options and make a new model m2 for AIC to select again.

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Wins ~ ID + Yards + OffensivePlays + TurnOversLost + FumblesLost +
##     FirstDowns + PassesCompleted + PassesAttempted + YardsGainedPassing +
##     InterceptionsThrown + RushingAttempts + YardsGainedRushing +
##     PenaltiesCommitedByTeam + PenaltiesInYards + FirstDownsByPenalty +
##     NumberOfDrives + OppYards + OppOffensivePlays + OppTurnOversLost +
##     OppFumblesLost + OppFirstDowns + OppPassesCompleted +
OppPassesAttempted +
##     OppYardsGainedPassing + OppInterceptionsThrown + OppRushingAttempts +
##     OppYardsGainedRushing + OppPenaltiesCommitedByTeam +
OppPenaltiesInYards +
##     OppFirstDownsByPenalty + OppNumberOfDrives + r1 + r2 + r3 +
##     r4
##
## Final Model:
## Wins ~ Yards + OffensivePlays + TurnOversLost + FirstDowns +
##     PassesAttempted + YardsGainedPassing + RushingAttempts +
##     PenaltiesInYards + OppOffensivePlays + OppTurnOversLost +
##     OppFirstDowns + OppYardsGainedPassing + OppRushingAttempts +
##     OppPenaltiesInYards + r2 + r4
##
```

```
##
##                            Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                                                  348   686.8489 288.9384
## 2       - OppYardsGainedRushing  0 0.0000000       348   686.8489 288.9384
## 3      - OppInterceptionsThrown  0 0.0000000       348   686.8489 288.9384
## 4          - YardsGainedRushing  0 0.0000000       348   686.8489 288.9384
## 5         - InterceptionsThrown  0 0.0000000       348   686.8489 288.9384
## 6      - OppFirstDownsByPenalty  1 0.1037633       349   686.9526 286.9958
## 7                 - FumblesLost  1 0.2125611       350   687.1652 285.1133
## 8                    - OppYards  1 0.3156109       351   687.4808 283.2878
## 9             - PassesCompleted  1 0.3879452       352   687.8687 281.5022
## 10         - OppPassesAttempted  1 0.7211842       353   688.5899 279.9004
## 11 - OppPenaltiesCommitedByTeam  1 0.6859668       354   689.2759 278.2787
## 12        - FirstDownsByPenalty  1 0.6890489       355   689.9649 276.6584
## 13             - OppFumblesLost  1 0.8781758       356   690.8431 275.1418
## 14                           - r3  1 0.9221346       357   691.7653 273.6487
## 15          - OppNumberOfDrives  1 1.1295378       358   692.8948 272.2686
## 16             - NumberOfDrives  1 0.7928670       359   693.6877 270.7032
## 17         - OppPassesCompleted  1 1.3861540       360   695.0738 269.4618
## 18                           - ID  1 2.1606246       361   697.2344 268.6412
## 19                           - r1  1 2.1996827       362   699.4341 267.8381
## 20    - PenaltiesCommitedByTeam  1 2.8877292       363   702.3219 267.4038
```

As as the result shows, r2 and r4 are selected, along with many other variables we talked about before. So, we can make our final model of "fit_sec" for the prediction. We can see from the summary of fit_sec belwo that the R^2 increases compared to "fit_first":

```
##
## Call:
## lm(formula = Wins ~ Yards + OffensivePlays + TurnOversLost +
##     FirstDowns + PassesAttempted + YardsGainedPassing + RushingAttempts +
##     PenaltiesInYards + OppOffensivePlays + OppTurnOversLost +
##     OppFirstDowns + OppYardsGainedPassing + OppRushingAttempts +
##     OppPenaltiesInYards + r2 + r4, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5000 -0.9618 -0.0547  0.9290  4.2316
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.9972336  4.6162255  -0.216 0.829087
## Yards                 -0.0022443  0.0008221  -2.730 0.006645 **
## OffensivePlays        -0.0348483  0.0082508  -4.224 3.04e-05 ***
## TurnOversLost         -0.1209796  0.0126641  -9.553  < 2e-16 ***
## FirstDowns             0.0261703  0.0065743   3.981 8.30e-05 ***
## PassesAttempted        0.0149590  0.0083583   1.790 0.074333 .
## YardsGainedPassing     0.0041697  0.0010818   3.854 0.000137 ***
## RushingAttempts        0.0294726  0.0081046   3.637 0.000316 ***
## PenaltiesInYards      -0.0014965  0.0005659  -2.645 0.008534 **
```

```
## OppOffensivePlays      0.0295410  0.0033186   8.902  < 2e-16 ***
## OppTurnOversLost       0.1063831  0.0124318   8.557 3.30e-16 ***
## OppFirstDowns         -0.0104969  0.0063695  -1.648 0.100221
## OppYardsGainedPassing -0.0036631  0.0005158  -7.102 6.52e-12 ***
## OppRushingAttempts    -0.0376963  0.0044423  -8.486 5.52e-16 ***
## OppPenaltiesInYards    0.0015736  0.0006564   2.397 0.017022 *
## r2                     4.7118450  2.1942435   2.147 0.032426 *
## r4                     0.7766694  0.1656468   4.689 3.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.391 on 363 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.8033
## F-statistic: 97.73 on 16 and 363 DF,  p-value: < 2.2e-16
```

Modeling Fitting and Model Interpretation

```
ratio1<-test$OffensivePlays/test$FirstDowns
test$r1<-ratio1

ratio2<-test$Yards/test$YardsGainedPassing
test$r2<-ratio2

ratio3<-test$PassesAttempted/test$RushingAttempts
test$r3<-ratio3

ratio4<-test$OppYardsGainedPassing/test$OppPassesCompleted
test$r4<-ratio4

prefit_sec <- predict(fit_sec, test)
test_ids <- 381:544
output <- data.frame(ID = test_ids, Wins= prefit_sec)
#write.csv(output, 'upload22.csv', quote = FALSE, row.names = FALSE)
```

Using the "fit_sec" model, we predict on the test data and upload the file "upload22.csv" to Kaggle. This multiple linear regression model, using variables including Yards, OffensivePlays, TurnOversLost, FirstDowns, PassesAttempted, YardsGainedPassing, RushingAttempts, PenaltiesInYards, OppOffensivePlays, OppTurnOversLost, OppFirstDowns, OppYardsGainedPassing, OppRushingAttempts, OppPenaltiesInYards, interaction between Yards/YardsGainedPassing, and interaction between OppYardsGainedPassing/OppPassesCompleted. This model has a 1.37436 RMSE score.

Discussion of the Model Performance

In deed, this model has a good performance, as the R^2 is above 80%. However, the problem of multicolinearity may exist as some of the variables are still not that significant in the final model. It seems that some feature engineering could be more specific if we have more time. Also, it would be helpful if we can have better understanding of what the sport itself, as the domain knowledge is very specific and the orginal variable selection would be easier.