Board Report

Leslie O'Bray

April 22, 2018

```
## Loading required package: grid

## corrplot 0.84 loaded

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Loading required package: lattice

## Loading required package: ggplot2

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-13

##
## Attaching package: 'glmnet'

## The following object is masked from 'package:pROC':
##
##     auc

##
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##
##     combn

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## The following object is masked from 'package:car':
##
##     logit
```

```
## Loading required package: survival
##
## Attaching package: 'survival'
## The following object is masked from 'package:caret':
##
##     cluster
##
## Attaching package: 'biostatUZH'
## The following object is masked from 'package:psych':
##
##     logit
## The following object is masked from 'package:car':
##
##     logit
##
## Attaching package: 'tidyr'
## The following object is masked from 'package:Matrix':
##
##     expand
##
## Attaching package: 'plotly'
## The following object is masked from 'package:ggplot2':
##
##     last_plot
## The following object is masked from 'package:stats':
##
##     filter
## The following object is masked from 'package:graphics':
##
##     layout
## The 'lsmeans' package is being deprecated.
## Users are encouraged to switch to 'emmeans'.
## See help('transition') for more information, including how
## to convert 'lsmeans' objects and scripts to work with 'emmeans'.
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
##
## Attaching package: 'plyr'
## The following object is masked from 'package:lubridate':
##
##     here
## The following objects are masked from 'package:plotly':
```

```
##
##     arrange, mutate, rename, summarise
```

Prepare the data and fit the full model:

```
##
##       0   1
##   1   1   0
##   2  38   1
##   3 133   7
##   4 427 214
##   5 150 476
##   6  13 163
```

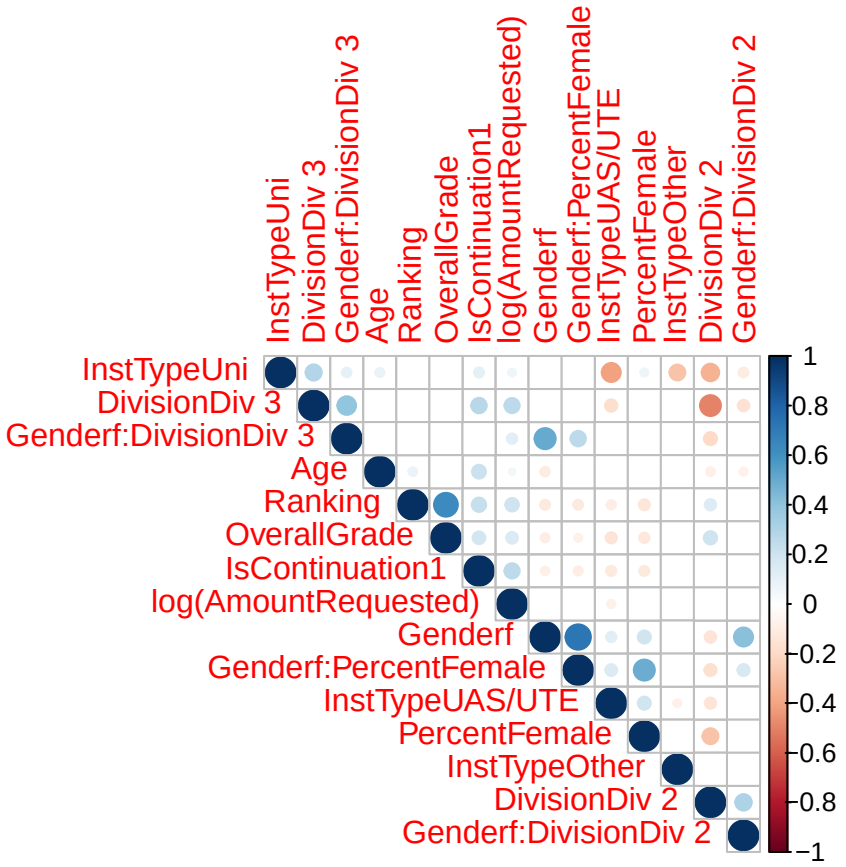
In fitting the regression, the summary shos that OverallGrade, Ranking, InstType and Age are all significant predictors.

```
##
## Call:
## glm(formula = board_data$IsApproved ~ Gender + Division + Age +
##     IsContinuation + InstType + log(AmountRequested) + Ranking +
##     OverallGrade + Gender:Division + PercentFemale + PercentFemale:Gender,
##     family = "binomial", data = board_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7280  -0.3684   0.0345   0.5559   2.7529
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -11.79012    2.13659  -5.518 3.43e-08 ***
## Genderf                 0.06480    0.41916   0.155   0.8771
## DivisionDiv 2          -0.20881    0.27616  -0.756   0.4496
## DivisionDiv 3           0.03066    0.26193   0.117   0.9068
## Age                    -0.02608    0.01048  -2.489   0.0128 *
## IsContinuation1         0.41781    0.21140   1.976   0.0481 *
## InstTypeOther           0.21352    0.38571   0.554   0.5799
## InstTypeUAS/UTE        -0.16225    0.36677  -0.442   0.6582
## InstTypeUni            -0.02601    0.22957  -0.113   0.9098
## log(AmountRequested)   -0.13366    0.15633  -0.855   0.3926
## Ranking                 3.14450    0.17590  17.877  < 2e-16 ***
## OverallGrade            0.65011    0.13110   4.959 7.09e-07 ***
## PercentFemale          -0.02877    0.48524  -0.059   0.9527
## Genderf:DivisionDiv 2  -0.31866    0.53096  -0.600   0.5484
## Genderf:DivisionDiv 3   0.18337    0.45242   0.405   0.6853
## Genderf:PercentFemale   0.22026    0.85577   0.257   0.7969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2243.9  on 1622  degrees of freedom
## Residual deviance:  996.4  on 1607  degrees of freedom
## AIC: 1028.4
##
## Number of Fisher Scoring iterations: 6
```

In checking for correlation among coefficients, we only don't see any alarming values (VIF > 5). However, gender is nearly 5 (4.99), so we would like to address this.

```
##                          GVIF Df GVIF^(1/(2*Df))
## Gender                4.989550  1        2.233730
## Division              2.954430  2        1.311047
## Age                   1.101439  1        1.049495
## IsContinuation        1.187186  1        1.089581
## InstType              1.581437  3        1.079382
## log(AmountRequested) 1.111318  1        1.054191
## Ranking               1.136836  1        1.066225
## OverallGrade          1.132667  1        1.064268
## PercentFemale         1.755519  1        1.324960
## Gender:Division       3.538740  2        1.371552
## Gender:PercentFemale 3.348380  1        1.829858
```

```
##                      Genderf DivisionDiv 2 DivisionDiv 3          Age
## Genderf         0.000000e+00  1.261550e-09  4.052810e-01 5.099710e-05
## DivisionDiv 2   1.261550e-09  0.000000e+00 8.307402e-100 4.136025e-04
## DivisionDiv 3   4.052810e-01 8.307402e-100  0.000000e+00 1.732005e-01
## Age             5.099710e-05  4.136025e-04  1.732005e-01 0.000000e+00
## IsContinuation1 6.370510e-04  2.545390e-01  1.507415e-28 9.864278e-20
## InstTypeOther   3.375759e-01  1.149066e-01  3.250971e-01 8.535204e-01
##                  IsContinuation1
## Genderf            6.370510e-04
## DivisionDiv 2      2.545390e-01
## DivisionDiv 3      1.507415e-28
## Age                9.864278e-20
## IsContinuation1    0.000000e+00
## InstTypeOther      6.855905e-01
```



Check VariableImportance. We see that Ranking is orders of magnitude more impactful than any of the other variables.

First, calculate Pseudo R^2, create a function:

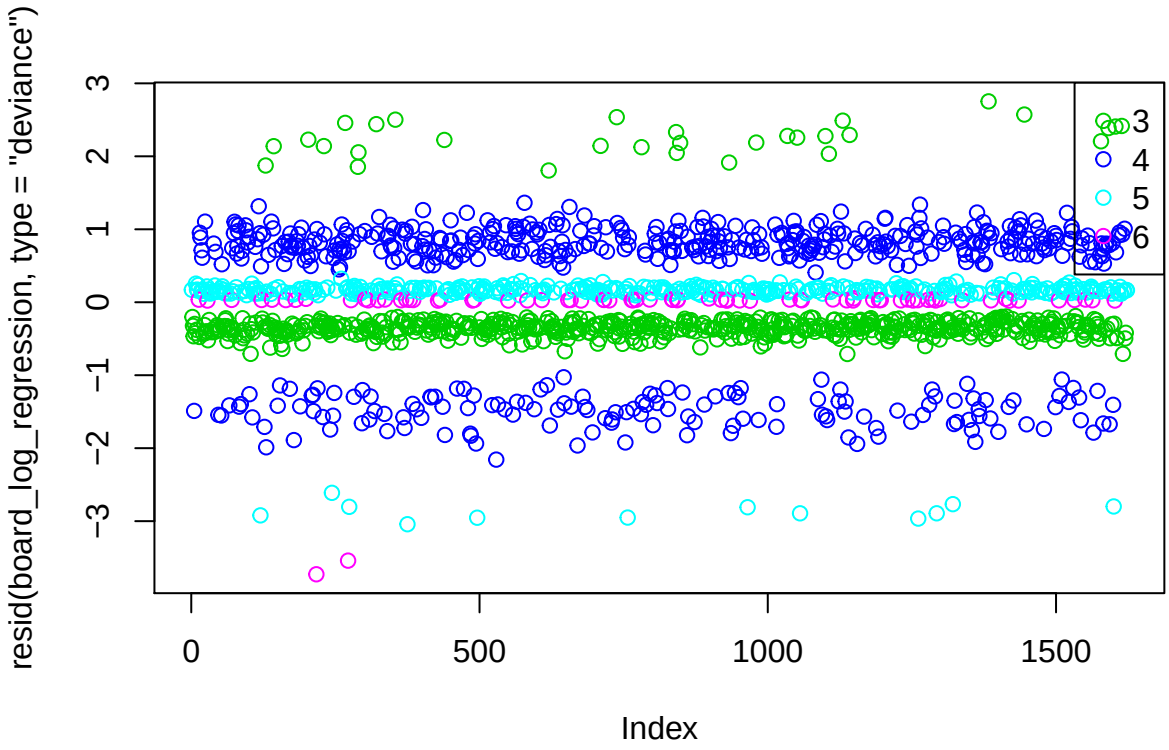
Alternative Method for seeing variable importance: For each predictor, permute the values. See difference in fit.

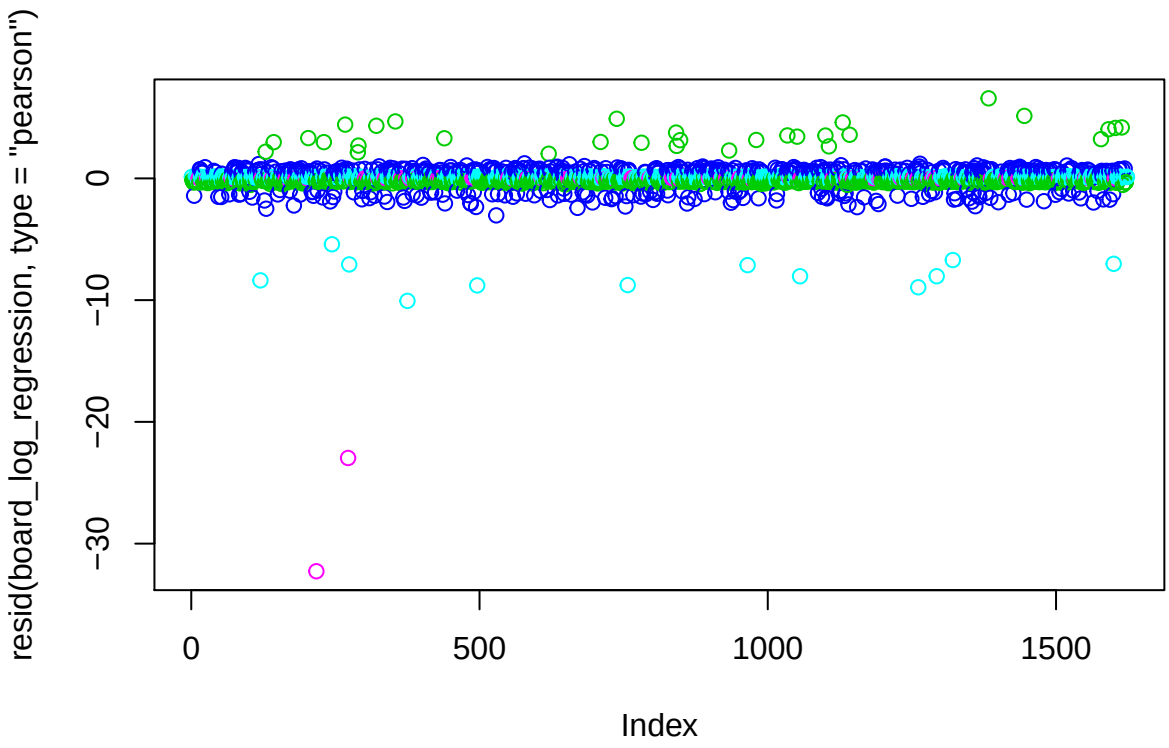
```
##           feature    importance
## 1          Gender  0.0011362186
## 5        InstType  0.0008263100
## 9   PercentFemale  0.0007938464
## 2        Division  0.0002428747
## 6 AmountRequested  0.0001523271
## 4  IsContinuation -0.0010754425
## 3             Age -0.0019168106
## 8    OverallGrade -0.0095010489
## 7         Ranking -0.2938943903
```

In this step, for each explanatory variable, I randomly permuted the values for that variable, and refitted the logistic regression model. I then compared the pseudo R^2 metric with the initial pseudo R^2 metric computed in the original model. In the output matrix, you can see how much the mean pseudo R^2 changed when that variable was permuted. In this case, we see that permuting Ranking had the biggest impact on the pseudo R^2 - decreasing it by 0.16. The next biggest impact was overall grade, which decreased the pseudo R^2 0.003. From this, we can conclude that Ranking is the most important explanatory variable in predicting IsApproved.

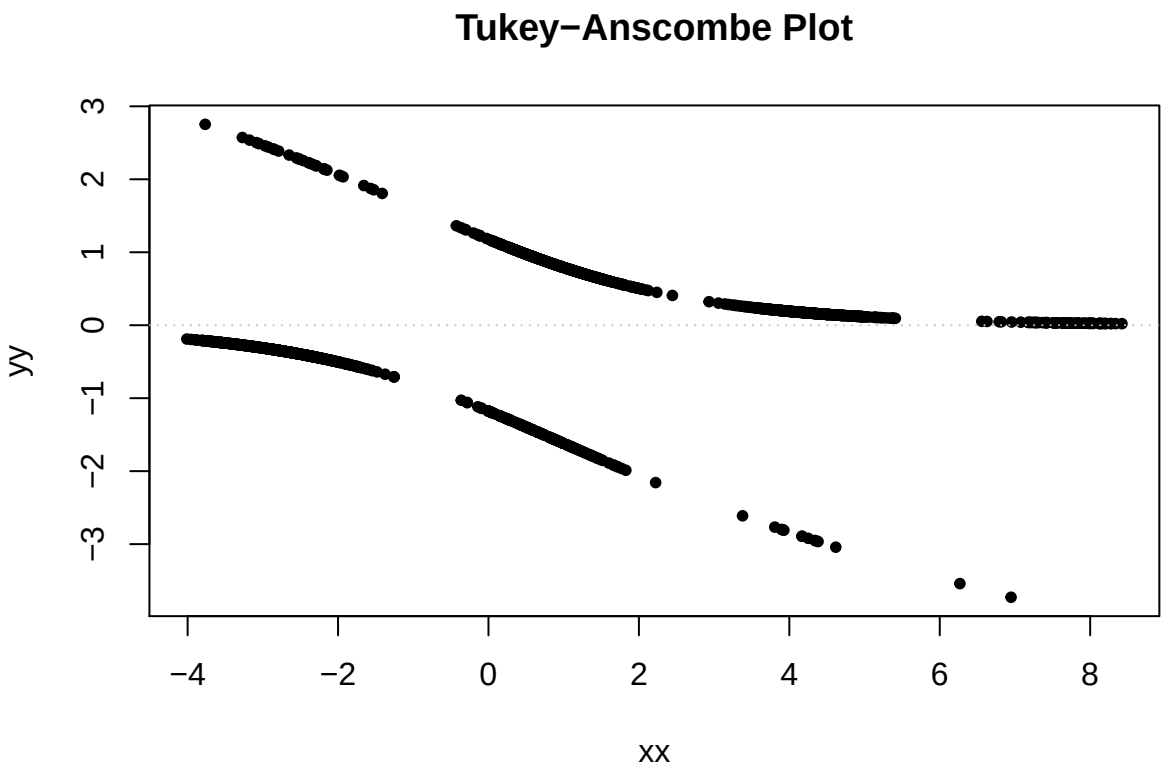
Check diagnostics: residuals. We see that we have smaller residuals for ranking 6 and 5, which intuitively makes sense that they get funded with consistency, and similarly we 1-3 dont get funded,

and thus are also classified correctly. Our resiudals with level 4 have the largest average deviance, and we have a few large outliers with 3 & 5, which likely means a 3 got funded, or a 5 did not get funded, despite our classification of the opposite. But this is a violation of our assumption of independently and identically distributed residuals, as there is clear structure in the residuals.





Checking the residuals, they seem to be expectation 0, with a few outliers. However, in looking at the deviance, there appears to be some structure in the data, explained by Ranking.



Do a bit of variable selection in order to optimize the AIC criterion:

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Gender + Division + Age + IsContinuation +
##     InstType + log(AmountRequested) + Ranking + OverallGrade +
##     Gender:Division + PercentFemale + PercentFemale:Gender
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     996.40 1028.4
## Age                    1  1002.65 1032.7   6.25   0.01242 *
## IsContinuation         1  1000.34 1030.3   3.94   0.04702 *
## InstType               3   997.12 1023.1   0.73   0.86718
## log(AmountRequested)   1   997.13 1027.1   0.73   0.39210
## Ranking                1  1628.14 1658.1 631.74 < 2.2e-16 ***
## OverallGrade           1  1021.83 1051.8  25.43 4.578e-07 ***
## Gender:Division        2   997.31 1025.3   0.92   0.63233
## Gender:PercentFemale   1   996.46 1026.5   0.07   0.79696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Gender + Division + Age + IsContinuation +
##     log(AmountRequested) + Ranking + OverallGrade + PercentFemale +
##     Gender:Division + Gender:PercentFemale
```

```
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     997.12 1023.1
## Age                    1  1003.39 1027.4   6.26   0.01232 *
## IsContinuation         1  1001.03 1025.0   3.90   0.04815 *
## log(AmountRequested)   1   997.84 1021.8   0.72   0.39704
## Ranking                1  1631.26 1655.3 634.13 < 2.2e-16 ***
## OverallGrade           1  1023.38 1047.4  26.26 2.986e-07 ***
## Gender:Division        2   998.04 1020.0   0.91   0.63377
## Gender:PercentFemale   1   997.17 1021.2   0.04   0.83263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Gender + Division + Age + IsContinuation +
##     log(AmountRequested) + Ranking + OverallGrade + PercentFemale +
##     Gender:PercentFemale
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     998.04 1020.0
## Division               2  1000.57 1018.6   2.53   0.28177
## Age                    1  1004.21 1024.2   6.17   0.01296 *
## IsContinuation         1  1001.79 1021.8   3.75   0.05274 .
## log(AmountRequested)   1   998.73 1018.7   0.70   0.40430
## Ranking                1  1631.43 1651.4 633.40 < 2.2e-16 ***
## OverallGrade           1  1024.36 1044.4  26.33 2.882e-07 ***
## Gender:PercentFemale   1   998.12 1018.1   0.09   0.77016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Gender + Division + Age + IsContinuation +
##     log(AmountRequested) + Ranking + OverallGrade + PercentFemale
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     998.12 1018.1
## Gender                 1   998.42 1016.4   0.30   0.58391
## Division               2  1000.60 1016.6   2.48   0.29001
## Age                    1  1004.29 1022.3   6.17   0.01298 *
## IsContinuation         1  1001.87 1019.9   3.75   0.05275 .
## log(AmountRequested)   1   998.80 1016.8   0.68   0.41050
## Ranking                1  1631.44 1649.4 633.32 < 2.2e-16 ***
## OverallGrade           1  1024.59 1042.6  26.47 2.674e-07 ***
## PercentFemale          1   998.12 1016.1   0.00   0.95801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Gender + Division + Age + IsContinuation +
##     log(AmountRequested) + Ranking + OverallGrade
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     998.12 1016.1
```

```
## Gender                1   998.44 1014.4   0.31   0.57582
## Division              2  1000.69 1014.7   2.57   0.27646
## Age                   1  1004.30 1020.3   6.17   0.01297 *
## IsContinuation        1  1001.88 1017.9   3.75   0.05274 .
## log(AmountRequested)  1   998.80 1014.8   0.68   0.41123
## Ranking               1  1633.57 1649.6 635.45 < 2.2e-16 ***
## OverallGrade          1  1024.59 1040.6  26.47 2.676e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Division + Age + IsContinuation + log(AmountRequested) +
##     Ranking + OverallGrade
##                      Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                    998.44 1014.4
## Division              2  1001.32 1013.3   2.88   0.23700
## Age                   1  1004.88 1018.9   6.45   0.01112 *
## IsContinuation        1  1002.16 1016.2   3.72   0.05367 .
## log(AmountRequested)  1   999.06 1013.1   0.63   0.42809
## Ranking               1  1633.87 1647.9 635.43 < 2.2e-16 ***
## OverallGrade          1  1024.77 1038.8  26.34 2.868e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Division + Age + IsContinuation + Ranking +
##     OverallGrade
##                Df Deviance    AIC    LRT  Pr(>Chi)
## <none>              999.06 1013.1
## Division        2  1001.70 1011.7   2.64   0.26714
## Age             1  1005.46 1017.5   6.39   0.01145 *
## IsContinuation  1  1002.46 1014.5   3.40   0.06525 .
## Ranking         1  1635.30 1647.3 636.23 < 2.2e-16 ***
## OverallGrade    1  1025.34 1037.3  26.27 2.964e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## board_data$IsApproved ~ Age + IsContinuation + Ranking + OverallGrade
##                Df Deviance    AIC
## <none>              1001.7 1011.7
## Age             1   1007.9 1015.9
## IsContinuation  1   1006.0 1014.0
## Ranking         1   1637.3 1645.3
## OverallGrade    1   1026.2 1034.2
```

```
## [1] "PseudoR^2 in Smaller Model"
```

```
## [1] 0.7140105
```

```
## [1] "PseudoR^2 in Full Model"
```

```
## [1] 0.7160376
```

The final model purely uses Age, IsContinuation, Ranking, & OverallGrade as predictors. Since the pseudo R^2 in the small model is nearly identical to the pseudo R^2 in the full model, we prefer the smaller model.

Let's check variable importance in the smaller model:

```
##          feature   importance
## 2 IsContinuation -0.001288568
## 1            Age -0.002085896
## 4   OverallGrade -0.009003748
## 3        Ranking -0.296981593
```