Multinom regression and random forest • Funding decision with budget cuts

Tommaso Portaluri

06/06/2018

Contents

Introduction

The Swiss National Science Foundation (SNF) is a research funding agency which disseminates yearly, on behalf of the Swiss Government, billions of CHF to the best researchers in Switzerland. This report contains a statistical analysis performed on three datasets provided by SNF, cointaing information on the applications for funding received in 2016, the corresponding evaluations and the scores given by both internal and external evaluators.
In addition to the two research questions foreseen at the beginning of the project, this report tries to answer an additional question, which looks at the cut in the budget of grant requests, even if approved. In particular, the research question is as follows:

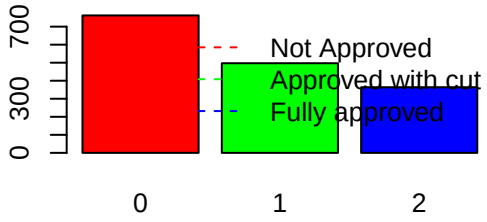
Which factors determine whether a request is approved and fully funded, approved with a cut in the budget or not approved?

Exploratory analysis

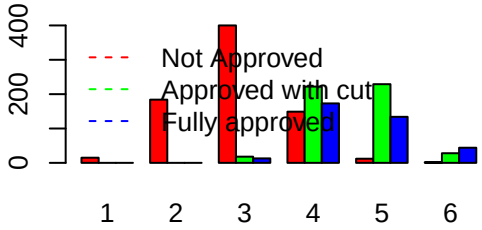
Before looking into the model, we will have a look at some visualizations, to better design the model. To start with, it is necessary to create an additional level in the variable IsApproved: not approved (0), approved with a cut (1), approved without cut (2). IsApproved was thus modified to become a three-level factor. To this end, we first checked that there are no cases in which the AmountGranted is higher than the amount requested; we will consider these cases as "fully granted" (2). We then create an additional class for applications approved with cut, to obtain a three-level IsAprroved variable. We also will consider the log of the AmountRequested, which is skewed to the left.

Now the dataset is ready for some visualizations.

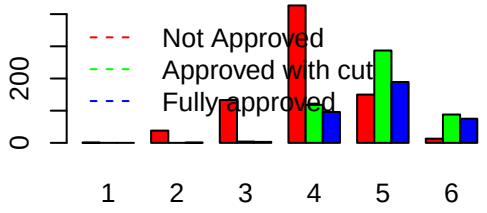
Plot 1 − Overview of funded project with



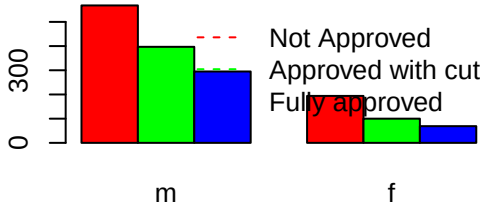
Plot 2 − Internal Ranking vs. Funding deci



t 3 − External Overall Grade vs. Funding c



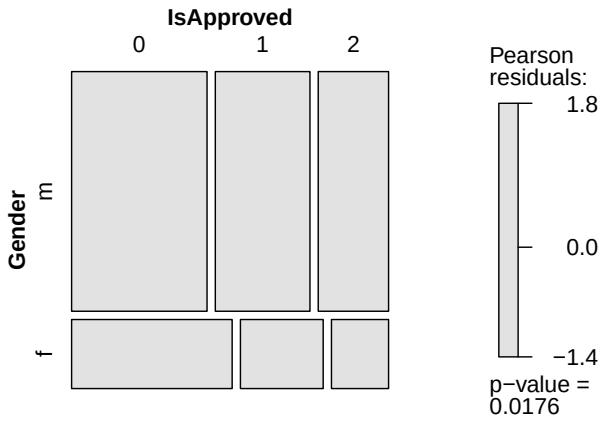
Plot 4 − Gender vs. Funding decision



As we can see in plot 1, most approved projects are approved with a cut (almost 58% of the total of approved projects). We are thus interested in analysing which demographic data of the applicant may determine the cut and at which step of the process this is more likely to happen (if with the external reviewers or with the internal referees). At a first glance, by having a look at the distribution of of cuts per both internal (plot 2) and external ranking (plot 3), both external referees and internal reviewers seem to affect the decision, with the latter displaying a stronger link (projects evaluated as "outstading" by internal revieers are very unlikely

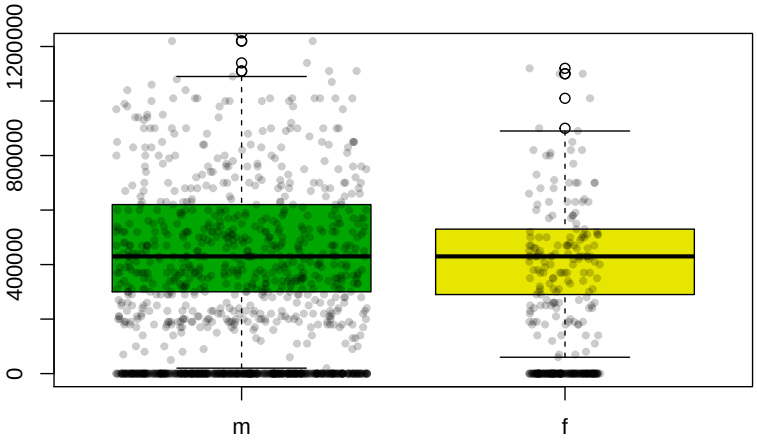
not to be funded and receive less cuts than those receiving the same grade by referees). Gender, instead, does not seem to be so meaningful (plot 4). We can invest this further with mosaic plots.

Plot 5 − Gender vs. Funding decision



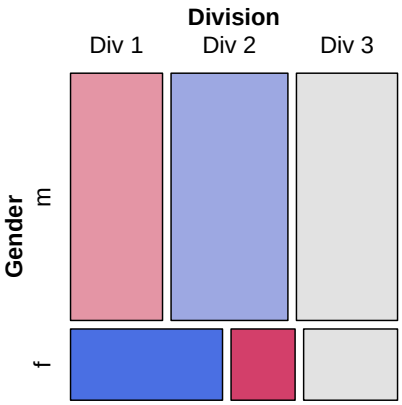
The mosaic plot (plot 5) suggests that gender may have an influence.

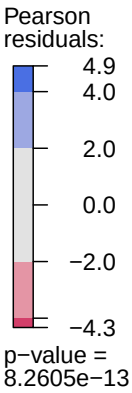
Plot 6 − Amount granted vs. Gender



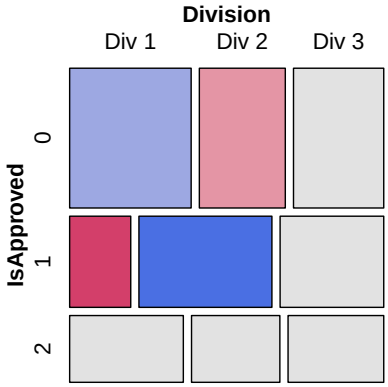
In plot 6, we can see that the variation in the amount granted is larger for men, which is somehow expected considering that male applications are much more.

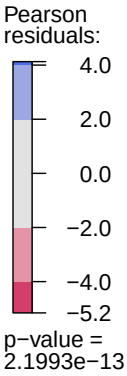
Plot 7 − Gender vs. Division





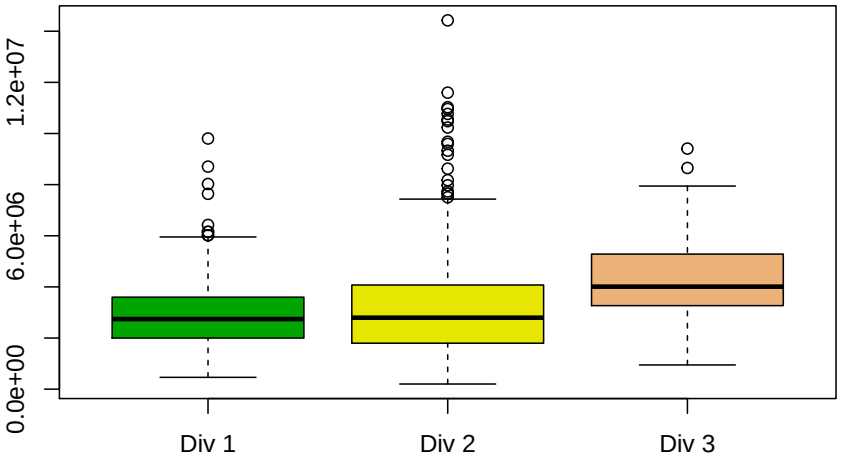
Plot 8 − Funding decision vs. Division





An interesting further interaction to investigate could be the intaraction between gender and division. Plot 7 indicates that the number of female applicants varies signficantly across divisions. Plot 8 shows clearly that, when looking within divisions, we find different "propension to cutting" the budget - e.g., Div2 is that with most cuts (in percentage).

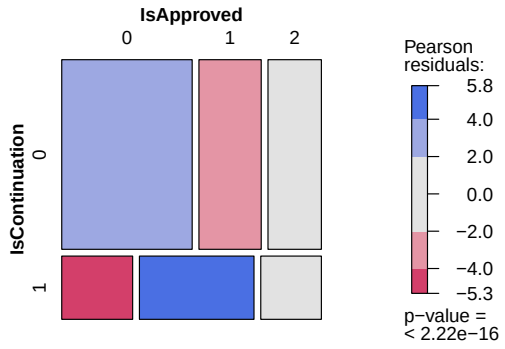
Plot 9 − Percentage of amount granted of the amount requested



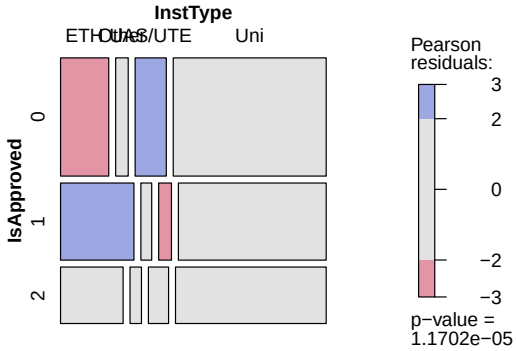
Differences across divisions do not just pertain to the percentage of projects which get cut but also to the extent of the cut. We can see in plot 9 that Division2 is also the division making the most important cuts in terms of budget (up to 30%).

We then look at other variables that might have an effect, such as IsContinuation and Institute Type.

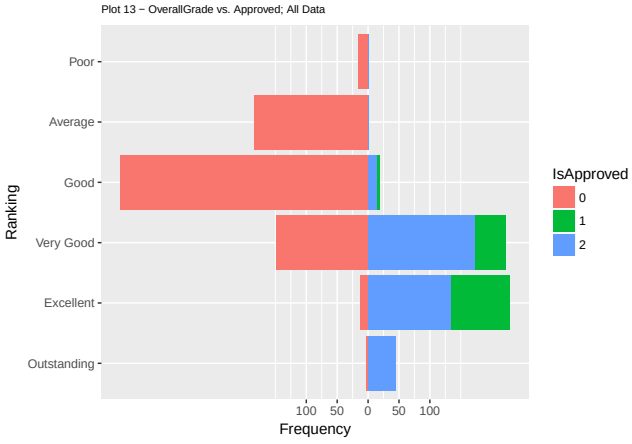
Plot 10 − IsContinuation vs. Funding decision



Plot 11 − Funding decision vs. Institute Type



Finally, some mirror plots can help us getting some useful insights. For instance, on the relationship between the funding decision and the internal and external evaluation (already considered in plot 2 and 3). In plot 12 and 13, we see again that applications rated as Outstanding by internal referees are more likely to get fully funded, with respect to those receiving the same grade by external referees. This suggests an higher influence of the internal process, which is in line with founding of previous analysis.



The model

We want to investigate the relative importance of both applicants' demographic data and of the assessment on the final decision - either (i) approving by fully granting the amount requested, (ii) approving by cutting it or (iii) not approving. Since we have a three-level response variable, we will fit a multinomial regression. To this end, we will use the multinom() function from the nnet package, which fits a multinomial log-linear model via neural networks. Multinomial regression works as follows: suppose we have a k-level response variable, the multinom() function will then compute k-1 equations whose coefficients should be interpreted with respect to a reference level set in advance. In this case, we choose as reference level the case of fully funded projects (IsApproved = 2), which means that positive coefficients will have to be interpreted as increasing the probability of being cut or even not approved (and conversely for negative coefficients). We should also set a reference level for factor predictors - we set both OverallGrade and Ranking raference to grade 6 = Outstanding. Note: to use the multinom() function, the assumpation of Independence of Irrelevant Alternatives (IIA), according to which adding or deleting alternative outcome categories should note affact the odds of the remaining outcomes, must hold. For the purposes of this report, we will assume IIA to hold.

In the model we will include all the variables that, according to the exploratory analysis, pontentially impact the funding decision (building also on the logistic regression already performed) and will perform variable selection afterwards. To allow a more meaningful interpretation, also predictors such as Ranking and

OverallGrade are considered here as non-ordered factors.

Variable selection was performed by using trough the function stepAIC (MASS package), which performs stepwise model selection by AIC in both directions.

```
##   (Intercept) DivisionDiv 2 DivisionDiv 3 OverallGrade1 OverallGrade2
## 0   -19.00675     0.7132202    -0.2758736     0.5743852     0.8508308
## 1   -25.07695     0.8043337    -0.1790771    -0.8180840   -11.2678001
##   OverallGrade3 OverallGrade4 OverallGrade5  Ranking1  Ranking2 Ranking3
## 0     1.9124740     1.4582044     0.8772315 26.208824 31.218126 6.531055
## 1     0.3709525     0.2088114     0.2949862  1.999492  8.379337 1.917222
##   Ranking4 Ranking5          Age AmountRequested
## 0 2.949732 0.737908 0.0259995450        1.041362
## 1 1.578151 1.500235 0.0004638001        1.796119
```

The summary gives the log odds for each of the two other levels, in each variable. With these values, it is possible to compute the probabilities of interest, by referring to the log of the ratio between the probability of the level chosen and the probability of the level of reference.

Apparently, gender is not significant, as it is not included in the final model.

Concerning the the interpretation of the coefficients, we have to look at some theory behind the multinomial models. As mentioned, the multinom() function computes k-1 equations, with respect to a reference level. The coefficients can be interpreted as generating a variation in the log odds Zk, which are equal to the log of the ratio between the probability of level K and the probability of the level of reference. In our case, for coefficients referring to "granted with cut" we would have:

ln(P(IsApproved = 1)/P(IsApproved = 2)) = Intercept + Sum(Coeff • Predictors) = Z1

Z1 is then then log odd for the Approved-with-cut response.

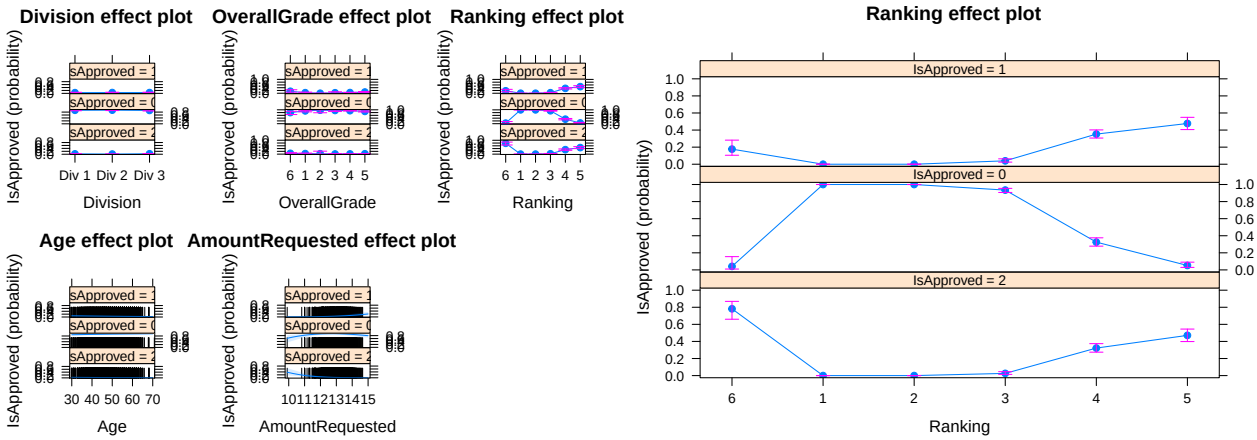
If we now look at the coefficients in the summary of the final model (see Appendix, p. 10), we can thus interpret them ad follows:

- an increase by one in the variable Age will determine an increase in the odds of not getting funded with respect to getting to fully funded (P(Y=0)/P(y=2)) by the factor exp(0.022). Which is to say: P(Y = 0 | age +1)/P(Y=2|age + 1) = exp(0.026) * P(Y=0 | no change in age)/P(Y=2 | no change in age)

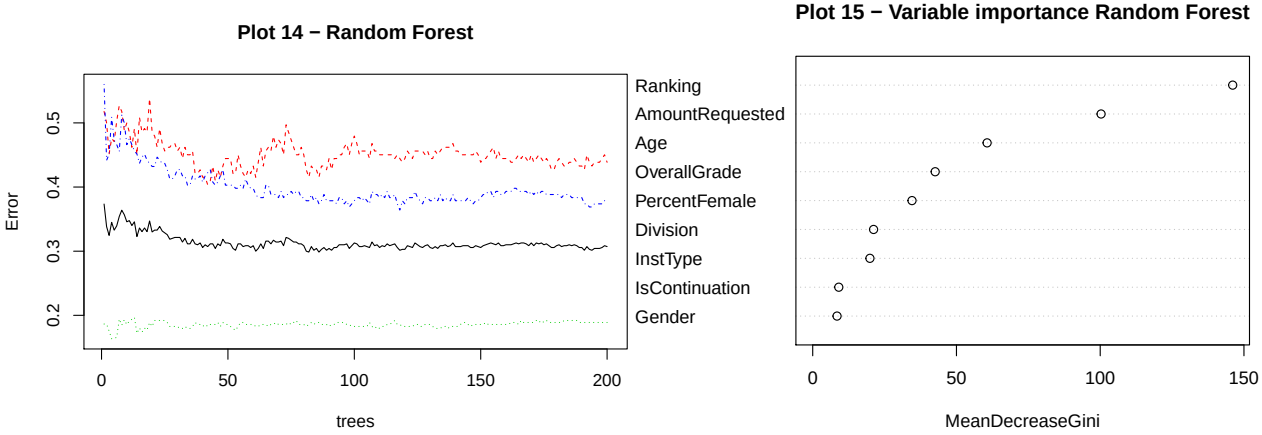
The interpretation of factor variables must be done with respect to the reference level. For instance:

- when moving from Ranking 6 (our reference level) to Ranking 5, there will be an increase in the odds of not getting funded with respect to getting to fully funded (P(Y=0)/P(y=2)) by the factor exp(0.94), and by the factor exp(0.32) in the odds of getting funded with a cut with respect to getting to fully funded (P(Y=1)/P(y=2));
- similarly, when moving from OverallGrade 6 to OverallGrade 3 there will be an increase in the odds of not getting funded with respect to getting to fully funded (P(Y=0)/P(y=2)) by the factor exp(2), and by the factor exp(0.43) in the odds of getting funded with a cut with respect to getting to fully funded (P(Y=1)/P(y=2));
- when moving from Div1 to Div2, there will be an increase in the odds of not getting funded with respect to getting to fully funded (P(Y=0)/P(y=2)) by the factor exp(0.62), and by the factor exp(0.66) in the odds of getting funded with a cut with respect to getting to fully funded (P(Y=1)/P(y=2)).

To get some insights into the relative importance of the variable we tried an analysis through the effects package. For the final model selected via stepAIC(), we get the following effect plots.



However, apart from Ranking effect plot, all the others do not seem to be very meaningful ways for interpretation. Hence, to better assess the variable importance, we decide then to run a random forest. To this end we use randomForest() function from the randomForest package, which is a classification algorithm that can be used for assessing proximities among data points in unsupervised mode. The randomForest function requires to define the number of trees to grow (high enough to ensure repeated predictions) and the number of variable to be randomly sampled at each split (usually set to the square root of the number of columns). It also required to create a train subset. After an exploratory analysis, we set at 200 the number of trees and then plot the importance of the variable with function varImpPlot (package caret), which gives as output a dotchart of variable importance as measured by Random Forest.



As we can see from plot 15, Ranking is confirmed to be the most important variable followed by AmonutRequested and Age. Gender appears to be the least important.

Conclusion

The multinomial regression suggests, in accordance with the other models performed, that gender has no effect on the final decision (it seems acutally the least important). The percentage of female reviewers is for instance more important than the gender of the applicant. Moreover, the grades of the internal referees seem to be those having a major effect in determining the final funding decision. AmountRequested and Age also seem to play a relevant role.

Appendix

Multinomial model

```
## Call:
## multinom(formula = IsApproved ~ Division + OverallGrade + Ranking +
##     Age + AmountRequested, data = regression_data)
##
## Coefficients:
##   (Intercept) DivisionDiv 2 DivisionDiv 3 OverallGrade1 OverallGrade2
## 0   -19.00675     0.7132202    -0.2758736     0.5743852     0.8508308
## 1   -25.07695     0.8043337    -0.1790771    -0.8180840   -11.2678001
##   OverallGrade3 OverallGrade4 OverallGrade5  Ranking1  Ranking2 Ranking3
## 0     1.9124740     1.4582044     0.8772315 26.208824 31.218126 6.531055
## 1     0.3709525     0.2088114     0.2949862  1.999492  8.379337 1.917222
##   Ranking4 Ranking5          Age AmountRequested
## 0 2.949732 0.737908 0.0259995450        1.041362
## 1 1.578151 1.500235 0.0004638001        1.796119
##
## Std. Errors:
##   (Intercept) DivisionDiv 2 DivisionDiv 3 OverallGrade1 OverallGrade2
## 0    2.695539     0.2396115     0.2416194  9.934539e-10  1.182053e+00
## 1    2.263618     0.1962667     0.2052653  9.930548e-10  1.319313e-05
##   OverallGrade3 OverallGrade4 OverallGrade5     Ranking1     Ranking2
## 0     0.7728456     0.4180058      0.407902 8.517010e-09 1.761607e-07
## 1     0.8228539     0.2587743      0.217179 2.467146e-09 1.758926e-07
##    Ranking3  Ranking4  Ranking5         Age AmountRequested
## 0 0.8257159 0.7637914 0.8010810 0.011684214       0.1868762
## 1 0.5136408 0.3321987 0.3159849 0.009564067       0.1644341
##
## Residual Deviance: 1928.177
## AIC: 1988.177
```

Random forest model

```
##
## Call:
##  randomForest(formula = IsApproved ~ Gender + Division + OverallGrade +      Ranking + Gender:Divisi
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 30.71%
## Confusion matrix:
##    2   0   1 class.error
## 2 96  24  51   0.4385965
## 0 27 262  34   0.1888545
## 1 60  19 127   0.3834951
```