INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Année universitaire 2019-2020

Rapport CheckYourSmile

Tutorial Project

LE Viet Minh Thong, PHAN Dinh Triem

Mai 2020

Institut National des Sciences Appliquées de Toulouse
Département Génie Mathématique et Modélisation

Contents

```
      prénom          Semestre   Filière     snapshot.1       snapshot.2    Snapshot.2...4m CYS.S3
ALEXANDRE:  3   S3 2017-18: 38   EEA:181   Min.   : 1.000   Min.   : 1.75          :127     non:120
ALEXIS   :  3   S3 2018-19:143             1st Qu.: 4.500   1st Qu.: 7.00   -      : 15     oui: 61
HUGO     :  3                              Median : 6.750   Median : 8.75   11     :  3
LUCAS    :  3                              Mean   : 6.442   Mean   : 8.82   11.75  :  3
NICOLAS  :  3                              3rd Qu.: 8.000   3rd Qu.:10.50   14.5   :  3
VINCENT  :  3                              Max.   :13.000   Max.   :15.75   6.75   :  3
(Other)  :163                                                               (Other): 27
CYS.S4    TP.S3    TP.S4     CMI        Groupe.S3     Groupe.S4     Prof.TP
   :137   FR:132     :143   non:150   Siuban :49           :154         :168
non: 21   GB: 49   FR: 15   oui: 31   Akane  :34   Nadia   : 15   Didier:  6
oui: 23            GB: 23             Alba   :29   Virginia: 12   Pierre:  7
                                      Nadia  :24
                                      Steven :23
                                      Yolanda:15
                                      (Other): 7
```

Figure 1: Summary of data semester 3

1 Introduction

CheckYourSmile (CYS) is a web platform project for learning speciality vocabulary in foreign languages (eg IT English: networks / databases), led by Dr Nadia Yassine-Diab, where users can learn through a set of "serious" games. The objective is to provide a complement to face-to-face language courses in higher education courses, where few hours can be devoted to teaching speciality vocabulary, despite its importance for the professional integration of students. Note that one of the innovations of CYS is to offer a collaborative system to propose and validate lexical entries: thus, everyone participates in the construction of knowledge (cf. crowdsourcing).

The first CYS prototype was released in 2014 (currently online at www.checkyoursmile.fr). IDEX (Initiative of Excellence) funding from the University of Toulouse in 2016 made it possible to hire several developers, trainees and post-docs to develop the site; a new version was released in January 2017, including new games and new features. The platform is free and licensed under the Creative Commons license. The previous prototypes have already served us to demonstrate the concept and to propose a stable and functional version of the site which now includes 6 games that work on the 4 skills of learning a language (French as a foreign language and English for French). 'instant, Spanish). For more information, see the project's Facebook page, the Twitter channel, the YouTube account and the Linked in page.

Our objective is to obtain indicators on the plus-value of Check Your Smile in a university context and on the combinations of variables which make it possible to obtain the best results in order to improve the effects of the determined tool.

2 Materials

The subject aims to study a database acquired during 3 academic years (2016-7, 2017-8 and 2018-9). It contains the evaluation results of students of different UPS courses as well as details on their courses and the particularities of the received language teaching (English or French TP, CMI engineering courses, use from CYS or not ...)

In fact, in semester 3, we carried out an assessment of 181 students including:
- 38 in 2017-2018 and 143 in 2018-2019
- All 181 in the sector EEA
- 120 used the CYS's tool while 41 did not use it
- 132 had practical works in French while 49 used English
- 160 are in CMI while 31 are not

We will analyse semester 4 and insert a table which indicates missing data

3 Methods

We applied statistical tests on data of both semesters (3 and 4 respectively)

Firstly, descriptive statistics were carried out to determine the most influential factors among considered variables. "Summary" in R provided a range of descriptive statistics at once. Moreover, charts like "boxplot" illustrated which

variables should be more important than others. Furthermore,"interaction.plot" showed how variables interacted mutually.

Secondly, linear regression led to a linear formula to study how multiple variables affected the progressions of students simultaneously including their mutual interactions. We have used AIC/BIC as criterion to choose the best fitting model to the data. Model ANOVA pointed out the effect of 3 qualitative variables on the progression of students as a term of difference between 2 Snapshots and as a term of ratio. On another hand, model ANCOVA pointed out the effect of 3 qualitative variables and Snapshot1 on Snapshot2. By comparing the R^2 values, we maintained the model whose R^2 value is higher.

Thirdly, non-linear regression (decision tree) with the tree graph demonstrated how multiple variables affected the progressions of students.

According to cross-validated predictions, we kept the model which possessed the minimal complexity parameter. As the same method we applied on linear regression, we studied 2 cases:

* The effect of 3 qualitative variables on the progression of students as a term of difference between 2 Snapshots.
* The effect of 3 qualitative variables and Snapshot1 on Snapshot2

We calculated the cross-validation errors of those models to reach the proper models.

Developments were carried out in R.

4 Results

4.1 Descriptive statistics

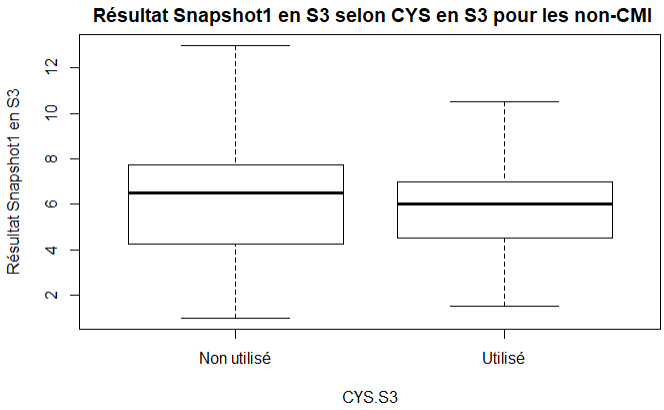

Figure 2: Snapshot1 of non-CMIs

According to the boxplots (Figure 2,3), we found the fact that:

- Generally, a non-CMI student had a higher Snapshot 1 score when he/she used the CYS tool.
- However, among non CMI students, those who used CYS had higher Snapshot 2 scores than those who did not use it.

We have observed a positive effect of the CYS tool applied to non CMI students

According to the boxplots(Figure 4,5), we found the fact that:

- Generally, a CMI student had a lower Snapshot 1 score when he/she did not use the CYS tool.
- Among non CMI students, those who used CYS had lower Snapshot 2 scores than those who did not use it.

We will study for this case the result evolution

Thus, CMI students progressed better when they did not use CYS.

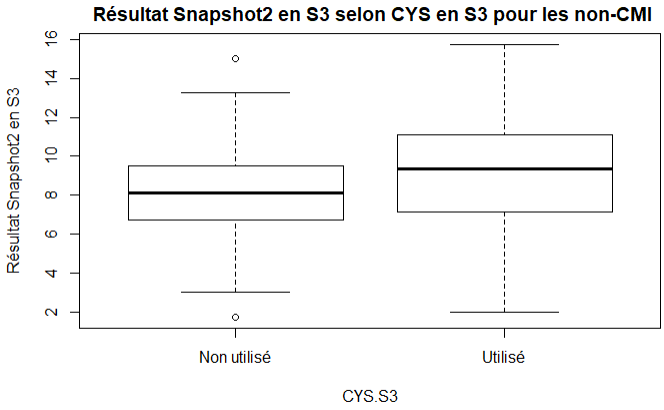

Figure 3: Snapshot2 of non-CMIs

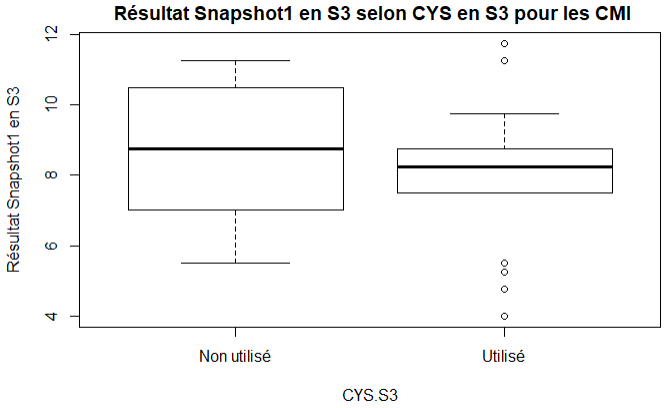

Figure 4: Snapshot1 of CMIs

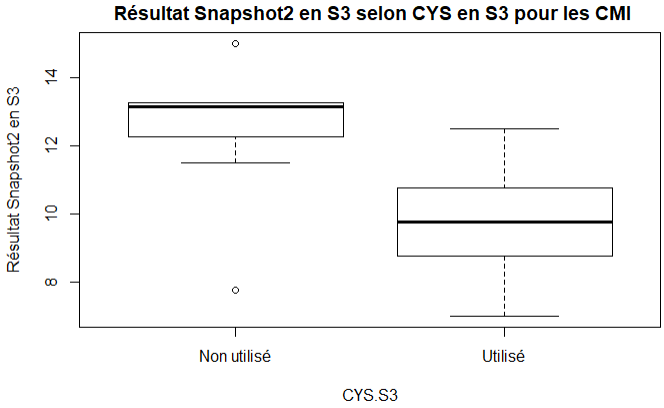

Figure 5: Snapshot2 of CMIs

We found that:

- The interaction between CYS and CMI was important: A student using CYS progressed less if he was in CMI. On the other hand, a student not using CYS progressed more if he was in CMI.

- The interaction between TP and CMI did not exist because there was no CMI student using French for TPs.

- The interaction between CYS and CMI existed: A student using CYS progressed a little more if he practiced TPs in English. However, a student not using CYS progressed much more if he practiced TPs in English.

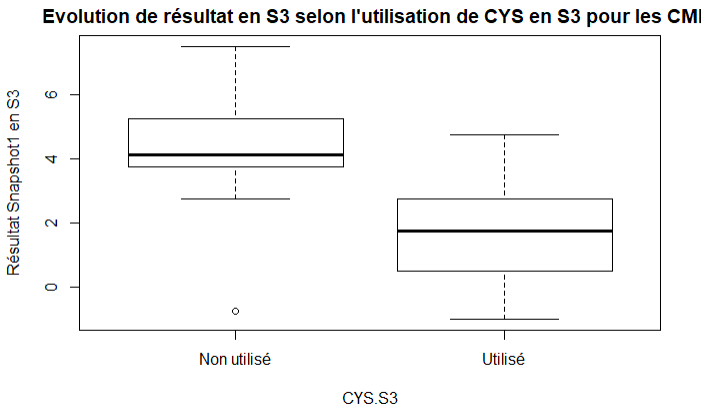

Figure 6: Evolution of CMIs

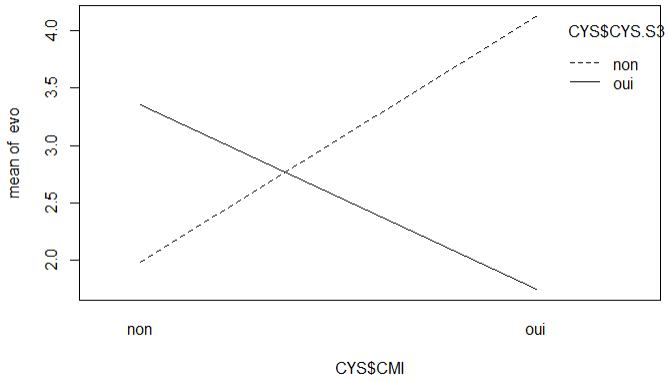

Figure 7: Interaction between CMI and CYS

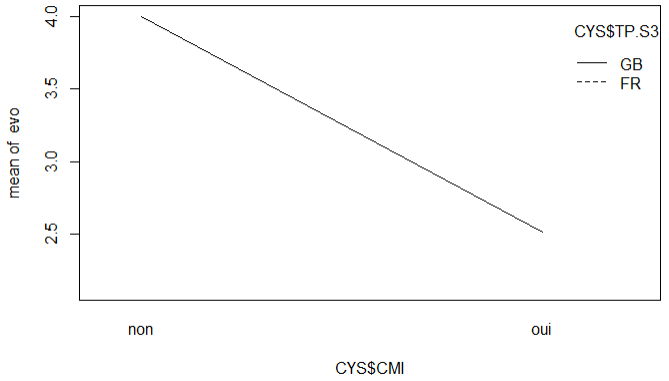

Figure 8: Interaction between CMI and TP

4.2 Linear regression

We started to conduct ANOVA models to study the impact of CMI, CYS and TP language factors on the evolution of results between Snapshot 1 and Snapshot 2 in semester 3

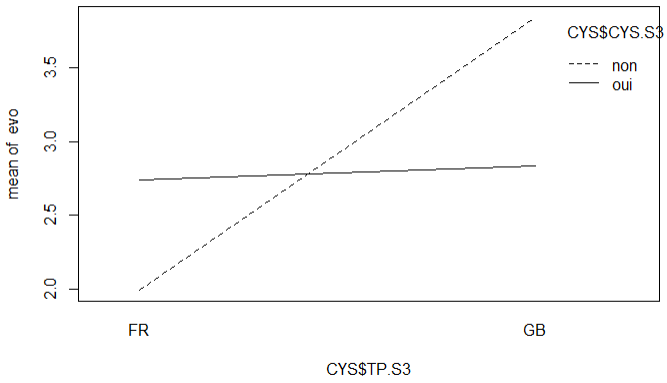

Figure 9: Interaction between CYS and TP

```
Call:
lm(formula = dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,
    data = CYS)

Residuals:
   Min     1Q Median     3Q    Max
-4.875 -1.600  0.000  1.264  5.400

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.9864     0.2068   9.606  < 2e-16 ***
CYS$CYS.S3oui               1.3636     0.4005   3.405 0.000818 ***
CYS$CMIoui                  2.1386     0.7164   2.985 0.003233 **
CYS$CYS.S3oui:CYS$CMIoui   -3.7386     0.9245  -4.044 7.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.169 on 177 degrees of freedom
Multiple R-squared:  0.1009,    Adjusted R-squared:  0.08569
F-statistic: 6.623 on 3 and 177 DF,  p-value: 0.000289
```

Figure 10: ANOVA model in case difference

```
Call:
lm(formula = ratio_snap ~ CYS$CMI, data = CYS)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8666 -0.4505 -0.1435  0.1981  3.4334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.56657    0.05769  27.153   <2e-16 ***
CYS$CMIoui  -0.20431    0.13941  -1.466    0.145
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7066 on 179 degrees of freedom
Multiple R-squared:  0.01186,   Adjusted R-squared:  0.006336
F-statistic: 2.148 on 1 and 179 DF,  p-value: 0.1445
```

Figure 11: ANOVA model in case ratio

4.2.1 ANOVA

By applying the BIC criterion on the complete model, we found the "modBIC2" model, where we did not find

the impact of the TP language:

- When we considered the difference between 2 snapshots we observed the modBIC1 model where this difference depended on the use of CYS, the fact of being in CMI and an interaction term between these two.

- When we considered the ratio between 2 snapshots we observed the modBIC2 model where this ratio depended only on being in CMI.

However we had the R-adjusted values too small (less than 0.1) so we decided to go further towards the ANCOVA model

4.2.2 ANCOVA

```
Call:
lm(formula = CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 +
    CYS$CMI + CYS$CYS.S3:CYS$CMI, data = CYS)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8965 -1.2725 -0.0205  1.1728  5.8232

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.62384    0.44064  10.493  < 2e-16 ***
CYS$snapshot.1             0.56912    0.06511   8.740 1.85e-15 ***
CYS$TP.S3GB                1.90361    0.59802   3.183  0.00172 **
CYS$CYS.S3oui              0.43506    0.43744   0.995  0.32132
CYS$CMIoui                 1.28154    0.86629   1.479  0.14084
CYS$CYS.S3oui:CYS$CMIoui -3.02652    0.85983  -3.520  0.00055 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.932 on 175 degrees of freedom
Multiple R-squared:  0.4764,    Adjusted R-squared:  0.4615
F-statistic: 31.85 on 5 and 175 DF,  p-value: < 2.2e-16
```

Figure 12: ANCOVA model

According to the BIC test, we found the best modBIC2 model:

$$snapshot.2 \sim snapshot.1 + TP.S3 + CYS.S3 + CMI + CYS.S3 : CMI$$

In fact, we estimated the result of Snapshot2 by the model:

$$(modbest) : Snapshot2_{ijkl} = \mu + \alpha Snapshot1_{ijkl} + \beta_i + \gamma_j + \theta_k + \delta_{jk} + \varepsilon_{ijkl},\ \forall i = 1,\ 2,\ \forall j = 1, 2, \forall k = 1, 2$$

where:

i, j, k are modality indices of the qualitative variables TP.S3, CYS.S3 and CMI, respectively (1 for the answer *No* and 2 for the answer *Yes*, in the case of TP 1 for *FR* and 2 for *GB*)

The index *ijkl* is to indicate the l-th individual having modalities *i, j, k* for TP.S3, CYS.S3 and CMI, respectively. $varepsilon_{ijkl}$ is errors in the estimation of the individual with the index *ijkl*.

From where:

$$\mu = 4.62384$$
$$\alpha = 0.56912$$
$$\beta_1 = \gamma_1 = \theta_1 = \delta_{11} = \delta_{12} = \delta_{21} = 0$$
$$\beta_2 = 1.90361$$
$$\gamma_2 = 0.43506$$
$$\delta_{22} = -3.02652$$

We decided to model the score of Snapshot2 according to Snapshot1, the use of the CheckYourSmile tool, the language of TP and the fact that the student is in CMI or not.

So, under the ANCOVA model, we found that the three qualitative factors had an impact on the Snapshot2 result. However, Snaphot1 had a big effect on the result of Snapshot2. There was also an interaction term between the variable CYS.S3 and the variable CMI and the latter had a significant negative effect on Snapshot2.

ie, a CMI student using the Check Your Smile tool tended to downgrade by about 2.6 points (-3.02652 + 0.43506) and a non-CMI student using the Check Your Smile tool tended to progress around 0 , 4 points (0.43506)

We saw that the ANCOVA model gave us a much better R-fit than the ANOVA model. (0.4615 ¿¿ 0.006 and 0.4615 ¿¿ 0.09)

For the purpose of obtaining a higher value of R we will pass under nonlinear models

4.3 Decision tree

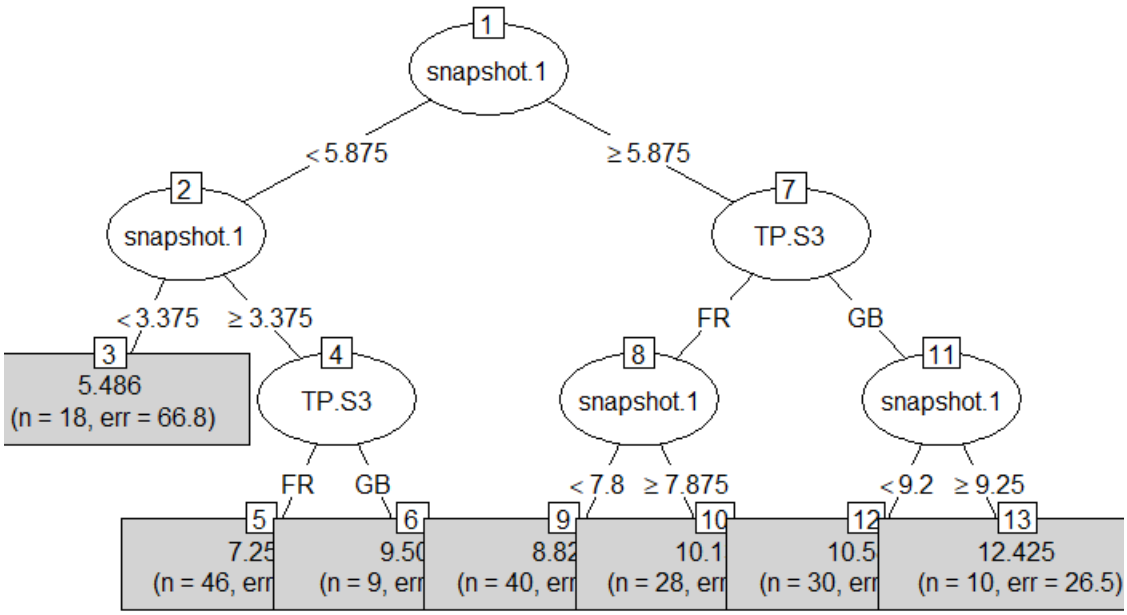

Figure 13: Regression tree

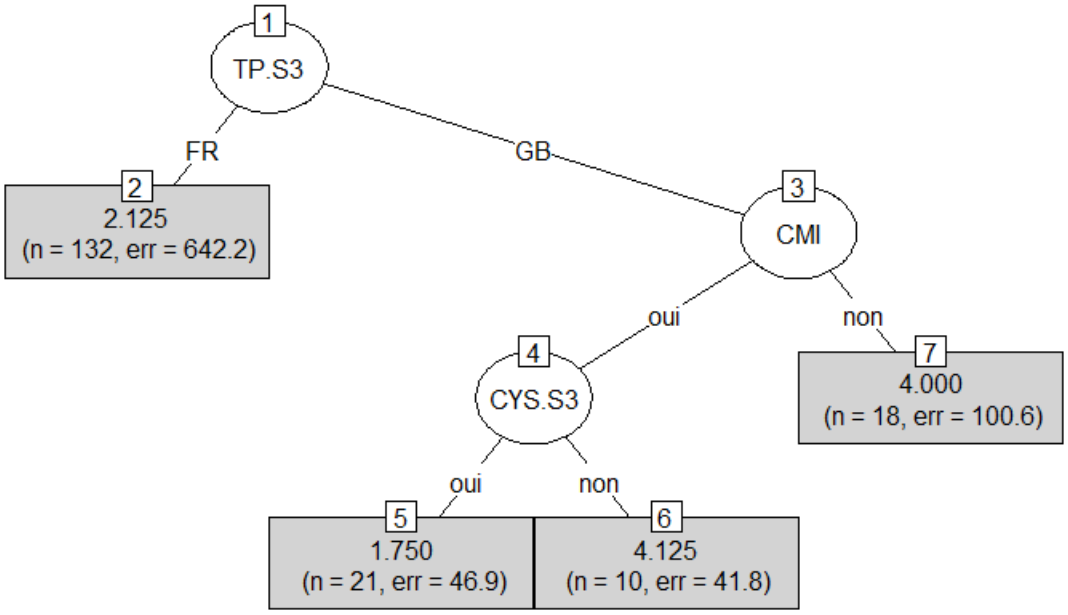

Figure 14: Regression tree of difference

We tried several times to study the behavior of cross-validation errors of these two types of the tree and we did not see the impact of CYS in the first case.

When we took the second:

- Among students practicing TPs in English, a student using the CYS tool progressed less than another not using CYS. (see sheets 5,6,7). In addition, among students practicing TPs in English and using the CYS tool, a CMI student progressed less than another non-CMI.

- The binary regression tree allowed us to conclude that the effect of the CYS tool in the progression of students was not remarkable.

- The cross validation error of the ANCOVA model was smaller than that of the decision tree. However, we found the same phenomenon for CMIs on the effect of CYS on student progression.

In term of cross-validation error:

* The first tree gave us 122.9607

* The second tree gave us 123.3507

By the way, ANCOVA gave us 100.7896 as cross-validation error.

5 Conclusion

Both linear(ANCOVA) and non-linear(decision tree) models led us to the fact that CYS only had a positive impact on those who were not in CMI while it had a negative impact on those who were in CMI

However, the non-linear model showed us how the variable TP impacted on the result of students using CYS while linear model could not.

References