Round-1

Arthur Sena

22-06-2015

```
library(dplyr, quietly = T, warn.conflicts = F)
library("C50")
library("gmodels")
```

Predição de Evasão dos feras de 2014.1

O objetivo desse problema é tentar criar um modelos que consiga predizer com uma boa eficiência se um aluno do primeiro período irá ou não evadir o curso baseado no histórico do mesmo. As nossas variáveis e suas descrições se encontram logo abaixo:

MATRICULA: identificador do aluno

PERIODO: identificador do periodo letivo da universidade (ano.semestre)

COD_CURSO: identificador do curso

CURSO: nome do curso. Cada curso tem seu COD_CURSO

CODIGO: identificador da disciplina que o aluno cursou no periodo

DISCIPLINA: nome da disciplina referente que o aluno cursou no periodo. Cada disciplina tem seu CODIGO

CREDITOS: numero de créditos referente a disciplina

DEPARTAMENTO: departamento que ofertou a disciplina

MEDIA: media do aluno na disciplina (0 a 10). Alunos reprovados por falta numa disciplina recebem 0 e alunos que trancaram a disciplina recebem NA.

STATUS: Aprovado, Reprovado Por Falta, Trancado ou Reprovado. Se refere ao estado final do aluno na disciplina

PERIODO_INGRESSO: periodo letivo da universidade em que o aluno ingressou no curso.

PERIODO_RELATIVO: numero de periodos que o aluno está matriculado na universidade. "1" refere-se ao aluno em seu primeiro periodo, "5" refere-se ao aluno no quinto periodo.

COD_EVASAO: identificador de evasao do aluno. "0" significa que o aluno continuou ativo na universidade no periodo seguinte e "1" significa que o aluno desistiu do curso nesse periodo e não voltou a se matricular no seguinte.

Uma pequena amostra dos dados pode ser visualizada logo abaixo:

```
  treino<- read.csv("training_without_accents.txt")
  summary(treino)
```

```
##        ID          MATRICULA           COD_CURSO
##  Min.   :    1   Min.   :  2636462   Min.   :12204100
##  1st Qu.: 4739   1st Qu.:249727234   1st Qu.:14123100
##  Median : 9478   Median :508993893   Median :14123100
##  Mean   : 9581   Mean   :501522998   Mean   :13677419
```

```
##  3rd Qu.:14216   3rd Qu.:745324313   3rd Qu.:14123100
##  Max.   :19536   Max.   :999280527   Max.   :14123100
##
##                  CURSO          PERIODO         CODIGO
##  ENFERMAGEM - D     : 4402   Min.   :2002   Min.   :1105013
##  ENGENHARIA ELETRICA:14552   1st Qu.:2009   1st Qu.:1109103
##                              Median :2011   Median :1201136
##                              Mean   :2010   Mean   :1246218
##                              3rd Qu.:2012   3rd Qu.:1404139
##                              Max.   :2013   Max.   :1503072
##
##                                    DISCIPLINA       CREDITOS
##  INTRODUCAO A PROGRAMACAO              : 1436   4      :11210
##  INTRODUCAO A ENGENHARIA ELETRICA      : 1392   3      : 2978
##  CIENCIAS DO AMBIENTE                  : 1377   2      : 2954
##  EXPRESSAO GRAFICA                     : 1355   1      :  739
##  ALGEBRA VETORIAL E GEOMETRIA ANALITICA: 1341   5      :  441
##  CALCULO DIFERENCIAL E INTEGRAL I      : 1334   8      :  323
##  (Other)                               :10719   (Other):  309
##                                   DEPARTAMENTO      MEDIA
##  UNID. ACAD. DE CIENCIAS DA SAUDE (UACS):4402   0      : 2158
##  UNID. ACAD. DE ENGENHARIA ELETRICA     :3492   7      : 1236
##  UNID. ACAD. DE MATEMATICA              :3374   5      :  558
##  UNID. ACAD. DE FISICA                  :2102   7.3    :  516
##  UNID. ACAD. DE SISTEMAS E COMPUTACAO   :1695   8      :  510
##  UNID. ACAD. DE ENGENHARIA CIVIL        :1420   (Other):13447
##  (Other)                                :2469   NA's   :  529
##                 SITUACAO     PERIODO_INGRESSO PERIODO_RELATIVO
##  0                   :    2   2010.1 :1882     Min.   :   1.000
##  Aprovado            :13438   2010.2 :1789     1st Qu.:   1.000
##  Reprovado           : 2906   2011.2 :1705     Median :   1.000
##  Reprovado por Falta: 2047   2011.1 :1597     Mean   :   2.353
##  Trancado            :  561   2009.1 :1426     3rd Qu.:   5.000
##                               2009.2 :1385     Max.   :2010.200
##                               (Other):9170
##    COD_EVASAO            X
##  Min.   :0.00000   Min.   :1
##  1st Qu.:0.00000   1st Qu.:1
##  Median :0.00000   Median :1
##  Mean   :0.09291   Mean   :1
##  3rd Qu.:0.00000   3rd Qu.:1
##  Max.   :1.00000   Max.   :1
##                    NA's   :18952
```

```
head(treino)
```

```
##   ID MATRICULA COD_CURSO              CURSO PERIODO  CODIGO
## 1  1 733623117  14123100 ENGENHARIA ELETRICA  2004.1 1404138
## 2  2 733623117  14123100 ENGENHARIA ELETRICA  2002.1 1503038
## 3  3 733623117  14123100 ENGENHARIA ELETRICA  2004.1 1404143
## 4  4 733623117  14123100 ENGENHARIA ELETRICA  2002.1 1109035
## 5  5 733623117  14123100 ENGENHARIA ELETRICA  2002.1 1109103
## 6  6 733623117  14123100 ENGENHARIA ELETRICA  2002.1 1108080
##                              DISCIPLINA CREDITOS
```

```
## 1                 CIRCUITOS ELETRICOS II        4
## 2                   CIENCIAS DO AMBIENTE        3
## 3                         ONDAS E LINHAS        4
## 4 ALGEBRA VETORIAL E GEOMETRIA ANALITICA        4
## 5       CALCULO DIFERENCIAL E INTEGRAL I        4
## 6                               FISICA I        4
##                        DEPARTAMENTO MEDIA SITUACAO PERIODO_INGRESSO
## 1 UNID. ACAD. DE ENGENHARIA ELETRICA     7 Aprovado           2002.1
## 2    UNID. ACAD. DE ENGENHARIA CIVIL   8.3 Aprovado           2002.1
## 3 UNID. ACAD. DE ENGENHARIA ELETRICA   7.6 Aprovado           2002.1
## 4          UNID. ACAD. DE MATEMATICA   9.5 Aprovado           2002.1
## 5          UNID. ACAD. DE MATEMATICA   9.8 Aprovado           2002.1
## 6              UNID. ACAD. DE FISICA   8.2 Aprovado           2002.1
##   PERIODO_RELATIVO COD_EVASAO  X
## 1                5          0 NA
## 2                1          0 NA
## 3                5          0 NA
## 4                1          0 NA
## 5                1          0 NA
## 6                1          0 NA
```

Observando os dados, notamos que os mesmos se encontram organizados por disciplinas, onde cada linha tem a média do aluno em uma determinada matéria. Ainda sim, podemos criar um modelo a partir desses dados e verificar o seu desempenho. As variáveis que eu achei que possam ter uma maior influência na evasão foram:

COD_CURSO: Acredito que dependendo do curso o usuário apresenta ou não uma maior chance de evasão, pois penso que cursos da área de exatas são mais propícios a um maior índice de evasão.

DISCIPLINA: Acredito que tal variável apresenta uma grande chance de influência na classificação, pois acho que ela engloba um conjunto de fatores que são determinantes para a permanência, ou não, de alguns alunos no curso. Por exemplo: Dificuldade da disciplina/ Método de Ensino do Professor/ Horário/ Conteúdo. Por tais fatores acredito que uma ou mais disciplinas podem fazer o aluno evadir o curso.

DEPARTAMENTO: No meu ponto de vista departamentos de cursos de exatas apresentam um maior problema de evasão de alunos.

MEDIA: Acredito que as notas do aluno podem desistimulá-lo a tanto desistir quanto prosseguir no curso.

SITUACAO: Por fim, acho que alunos que apresentem uma situação de aprovação maior que reprovação estão mais dispostos a continuarem no curso.

Agora, nós usaremos a biblioteca C50 para construir a árvore de classificação.

```
treino<- read.csv("training_without_accents.txt")
teste <- read.csv("test_without_accents.csv")

evadiu <- filter(treino, COD_EVASAO == 1, PERIODO_RELATIVO  == 1)
nao_evadiu <- filter(treino, COD_EVASAO == 0, PERIODO_RELATIVO  == 1)
treino2 <- rbind(evadiu[1:1000,],nao_evadiu[1:1200,])

treino2[,"COD_EVASAO"]<- as.factor(treino2$COD_EVASAO)
treino2[,"X"] <- NULL
evasao_model <- C5.0(treino2[,c(3,7,9,10,11)], treino2$COD_EVASAO)
treino2[,"MEDIA"] <- as.numeric(treino2$MEDIA)
result <- predict(evasao_model,newdata = treino)
CrossTable(treino$COD_EVASAO, result, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  18954
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |     10327 |      6866 |     17193 |
##              |     0.545 |     0.362 |           |
## -------------|-----------|-----------|-----------|
##            1 |       480 |      1281 |      1761 |
##              |     0.025 |     0.068 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10807 |      8147 |     18954 |
## -------------|-----------|-----------|-----------|
##
##
```

```
    recall <- 1281/1761
    precision <- 1281/8147
    Fmeasure <-  2 * precision * recall / (precision + recall)
    print(Fmeasure)
```

```
## [1] 0.2585789
```

Notamos que o F-measure apresenta um valor baixo, ou seja, esse muito provavelmente não é um bom modelo. O que realmente foi comprovado quando eu usei tal modelo para prever a evasão dos alunos e acabei obtendo uma pontuação de apenas 64% no Kaggle.

Melhorando Nosso Modelo A fim de melhorar nosso modelo, nós devemos criar um novo dataset que expresse melhor nossos dados e que contenha novas variáveis. Uma boa estratégia é criar uma dataset, onde cada observação seria um aluno, seu respectivo CURSO e seu CRE naquele período. Pois, acreditamos que o CRE que o individuo atinga no fim do primeiro período juntamente com o CURSO do aluno tenham uma forte influência na decisão de prosseguir ou não no curso. Com isso em mente só precisamos utilizar os nossos conhecimentos em R.

```
  alunos <- levels(as.factor(treino$MATRICULA))
  new_data <- data.frame()

    for (aluno in alunos){
         temp <- filter(treino, aluno == as.character(MATRICULA), na.omit = T)
         media <- sum(as.numeric(levels(temp$MEDIA))[temp$MEDIA],na.rm = T)/nrow(temp)
         cod_evasao <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
          CURSO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
         SITUACAO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
```

```
        new_data <- rbind(new_data,data.frame(aluno,media,cod_evasao,CURSO),warn.conflicts = F )
    }

  alunos_Teste <- levels(as.factor(treino2$MATRICULA))
  new_data2 <- data.frame()
  for (aluno in alunos_Teste){
       temp <- filter(treino2, aluno == as.character(MATRICULA), na.omit = T)
       media <- sum(as.numeric(levels(temp$MEDIA))[temp$MEDIA],na.rm = T)/nrow(temp)
       COD_EVASAO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
        CURSO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
       SITUACAO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
       new_data2 <- rbind(new_data2,data.frame(aluno,media,COD_EVASAO,CURSO))
  }

new_data[,"COD_EVASAO"] <- as.factor(new_data$COD_EVASAO)
evasao_model <- C5.0(new_data[,c(1,2,4)], new_data$COD_EVASAO, na.omit = T)
result <- predict(evasao_model,newdata = new_data2)
CrossTable(new_data2$COD_EVASAO, result, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('a
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2200
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1200 |         0 |      1200 |
##              |     0.545 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            1 |         0 |      1000 |      1000 |
##              |     0.000 |     0.455 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1200 |      1000 |      2200 |
## -------------|-----------|-----------|-----------|
##
##
```

```
recall <- 1000/1000
precision <- 1000/1000
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

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

Notamos que diferentemente do modelo anterior, esse apresenta um F-measure alto. O que realmente foi confirmado quando eu submeti a minha predição utilizando esse modelo, pois consegui obter uma pontuação de 90%. Ainda tentando melhorar esse modelo eu resolvi adicionar a variável situação e curso para observar o que aconteceria.

```
  alunos <- levels(as.factor(treino$MATRICULA))
  new_data <- data.frame()

    for (aluno in alunos){
         temp <- filter(treino, aluno == as.character(MATRICULA), na.omit = T)
         media <- sum(as.numeric(levels(temp$MEDIA))[temp$MEDIA],na.rm = T)/nrow(temp)
         cod_evasao <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
         CURSO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
         SITUACAO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
         new_data <- rbind(new_data,data.frame(aluno,media,cod_evasao,CURSO,SITUACAO),warn.conflicts = F
    }

  alunos_Teste <- levels(as.factor(treino2$MATRICULA))
  new_data2 <- data.frame()
  for (aluno in alunos_Teste){
       temp <- filter(treino2, aluno == as.character(MATRICULA), na.omit = T)
       media <- sum(as.numeric(levels(temp$MEDIA))[temp$MEDIA],na.rm = T)/nrow(temp)
       COD_EVASAO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
       CURSO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
       SITUACAO <- treino2 %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
       new_data2 <- rbind(new_data2,data.frame(aluno,media,COD_EVASAO,CURSO,SITUACAO))
  }

new_data[,"COD_EVASAO"] <- as.factor(new_data$COD_EVASAO)
evasao_model <- C5.0(new_data[,c(1,2,4,5)], new_data$COD_EVASAO, na.omit = T)
result <- predict(evasao_model,newdata = new_data2)
CrossTable(new_data2$COD_EVASAO, result, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('a
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2200
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1200 |         0 |      1200 |
##              |     0.545 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            1 |         0 |      1000 |      1000 |
##              |     0.000 |     0.455 |           |
## -------------|-----------|-----------|-----------|
```

```
## Column Total |      1200 |      1000 |      2200 |
## -------------|-----------|-----------|-----------|
##
##
```

```
recall <- 1000/1000
precision <- 1000/1000
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

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

Apesar do F-measure continuar alto, eu obtive um menor pontuação de 89% o que é um pouco menor que o modelo anterior.