Round-2

Arthur Sena

06/24/2015

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

Predição de Evasão de alunos do quinto período

Nesse segundo round da competição, nós utilizaremos os mesmos dados, porém tetaremos criar um bom modelo para prever a evasão de alunos que estão no quinto período do curso. Os nossos dados podem ser, novamente, vistos abaixo:

```
    treino<- read.csv("treino_2Round.csv")
    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   Min.   :0.000
##  INTRODUCAO A ENGENHARIA ELETRICA      : 1392   1st Qu.:3.000
##  CIENCIAS DO AMBIENTE                  : 1377   Median :4.000
##  EXPRESSAO GRAFICA                     : 1355   Mean   :3.536
##  ALGEBRA VETORIAL E GEOMETRIA ANALITICA: 1341   3rd Qu.:4.000
##  CALCULO DIFERENCIAL E INTEGRAL I      : 1334   Max.   :8.000
##  (Other)                               :10719
##                                  DEPARTAMENTO      MEDIA
##  UNID. ACAD. DE CIENCIAS DA SAUDE (UACS):4402   Min.   : 0.000
##  UNID. ACAD. DE ENGENHARIA ELETRICA     :3494   1st Qu.: 4.000
##  UNID. ACAD. DE MATEMATICA              :3374   Median : 7.000
##  UNID. ACAD. DE FISICA                  :2102   Mean   : 5.904
##  UNID. ACAD. DE SISTEMAS E COMPUTACAO   :1695   3rd Qu.: 8.200
##  UNID. ACAD. DE ENGENHARIA CIVIL        :1420   Max.   :10.000
##  (Other)                                :2467   NA's   :529
```

```
##                 SITUACAO     PERIODO_INGRESSO PERIODO_RELATIVO
##  Aprovado           :13438   Min.   :2002     Min.   :1.000
##  Reprovado          : 2906   1st Qu.:2008     1st Qu.:1.000
##  Reprovado por Falta: 2049   Median :2010     Median :1.000
##  Trancado           :  561   Mean   :2009     Mean   :2.141
##                              3rd Qu.:2011     3rd Qu.:5.000
##                              Max.   :2013     Max.   :5.000
##
##    COD_EVASAO
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.09291
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

```
    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   7.0 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
## 1                5          0
## 2                1          0
## 3                5          0
## 4                1          0
## 5                1          0
## 6                1          0
```

Primeiro Modelo

No meu primeiro modelo eu considerei as variáveis: Curso e Média do aluno. Eu construir um novo dataset a partir do original, onde cada linha representa um aluno com sua média do período e seu curso. A partir disso,

utilizei a biblioteca C50 a fim de criar um modelo para explicar a variável evasão do aluno. O código para isso se encontrar logo abaixo:

```
    treino<- read.csv("treino_2Round.csv")
    teste <- read.csv("teste_2round.csv")

    evadiu <- filter(treino, COD_EVASAO == 1, PERIODO_RELATIVO  == 5)
    nao_evadiu <- filter(treino, COD_EVASAO == 0, PERIODO_RELATIVO  == 5)
    treino2 <- rbind(evadiu[1:1000,],nao_evadiu[1:1200,])

 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(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))
    }

  alunos_Teste <- levels(as.factor(treino2$MATRICULA))
  new_data2 <- data.frame()
  for (aluno in alunos_Teste){
       temp <- filter(treino2, aluno == as.character(MATRICULA), na.omit = T)
       media <- sum(temp$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('
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1519
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1200 |         0 |      1200 |
```

```
##              |     0.790 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            1 |       316 |         3 |       319 |
##              |     0.208 |     0.002 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1516 |         3 |      1519 |
## -------------|-----------|-----------|-----------|
##
##
```

```
recall <- 3/319
precision <-3/3
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

```
## [1] 0.01863354
```

Notamos que o F-measure apresenta um valor baixo, ainda sim eu consegui uma pontuação de 94% no kaggle quando submeti tal modelo.

Segundo Modelo

No meu segundo modelo, eu resolvi usar a variável situação e observar como meu modelo se sairia.

```
  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(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,CURSO,SITUACAO,cod_evasao))
    }

  alunos_Teste <- levels(as.factor(treino2$MATRICULA))
  new_data2 <- data.frame()
  for (aluno in alunos_Teste){
       temp <- filter(treino2, aluno == as.character(MATRICULA), na.omit = T)
       media <- sum(temp$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,CURSO,SITUACAO,COD_EVASAO))
  }

new_data[,"COD_EVASAO"] <- as.factor(new_data$COD_EVASAO)
new_data[,"aluno"] <- as.integer(new_data$aluno)
new_data2[,"aluno"] <- as.integer(new_data2$aluno)
evasao_model <- C5.0(new_data[,c(1,2,3,4)], new_data$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('
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1519
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       991 |       209 |      1200 |
##              |     0.652 |     0.138 |           |
## -------------|-----------|-----------|-----------|
##            1 |       207 |       112 |       319 |
##              |     0.136 |     0.074 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1198 |       321 |      1519 |
## -------------|-----------|-----------|-----------|
##
##
```

```
recall <- 319/319
precision <-319/159
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

```
## [1] 1.334728
```

Esse modelo obteve um F-measure melhor que o anterior, porém não consegui melhorar minha pontuação no kaggle, pois mantive um score de 94%.

Outros Modelos.

Eu construi mais 7 diferentes modelos tentando melhorar minha pontuação, mas infelizmente não consegui. Alguns desses modelos conseguiam prever muito bem os dados de treino, onde obtive o F-measure de valor 1. Ainda sim, quando eu fazia minha submissão via Kaggle minha pontuação continuava a mesma. O que me leva a acreditar que minhas árvores de decisão nao conseguiam generalizar suas previsões muito bem (Overfitting). Abaixo se encontra algumas árvores que eu tinha criado e seus respectivos F-measure.

Adicionando Variável Total de Créditos cursado pelo aluno

```
  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(temp$MEDIA)/nrow(temp)
         cod_evasao <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
          CURSO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
         SITUACAO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
         CREDITOS <- sum(temp$CREDITOS)/nrow(temp)
         new_data <- rbind(new_data,data.frame(aluno,media,CURSO,SITUACAO,CREDITOS,cod_evasao))
    }

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

new_data[,"COD_EVASAO"] <- as.factor(new_data$COD_EVASAO)
evasao_model <- C5.0(new_data[,c(1,2,3,4,5)], new_data$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('
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1519
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1200 |         0 |      1200 |
##              |     0.790 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            1 |       160 |       159 |       319 |
##              |     0.105 |     0.105 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1360 |       159 |      1519 |
## -------------|-----------|-----------|-----------|
```

```
##
##
```

```
recall <- 159/319
precision <- 159/159
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

```
## [1] 0.665272
```

Adicionando Variável Período que o aluno se encontra

```
  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(temp$MEDIA)/nrow(temp)
         cod_evasao <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(COD_EVASAO)
          CURSO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(CURSO)
         SITUACAO <- treino %>% filter( aluno ==  as.character(MATRICULA)) %>% select(SITUACAO)
         CREDITOS <- sum(temp$CREDITOS)/nrow(temp)
         PERIODO <- temp$PERIODO_RELATIVO
         new_data <- rbind(new_data,data.frame(aluno,media,CURSO,SITUACAO,CREDITOS, PERIODO,cod_evasao))
    }

  alunos_Teste <- levels(as.factor(treino2$MATRICULA))
  new_data2 <- data.frame()
  for (aluno in alunos_Teste){
       temp <- filter(treino2, aluno == as.character(MATRICULA), na.omit = T)
       media <- sum(temp$MEDIA)/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)
       CREDITOS <- sum(temp$CREDITOS)/nrow(temp)
        PERIODO <- temp$PERIODO_RELATIVO
       new_data2 <- rbind(new_data2,data.frame(aluno,media,CURSO,SITUACAO,CREDITOS, PERIODO ,COD_EVASAO
  }

new_data[,"COD_EVASAO"] <- as.factor(new_data$COD_EVASAO)
evasao_model <- C5.0(new_data[,c(1,2,3,4,5,6)], new_data$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('
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
```

```
##
## Total Observations in Table:  1519
##
##
##              | predicted
##       actual |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1200 |         0 |      1200 |
##              |     0.790 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            1 |         0 |       319 |       319 |
##              |     0.000 |     0.210 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1200 |       319 |      1519 |
## -------------|-----------|-----------|-----------|
##
##
```

```
recall <- 319/319
precision <-319/159
Fmeasure <-  2 * precision * recall / (precision + recall)
print(Fmeasure)
```

```
## [1] 1.334728
```