

Desafio Meantrix

Gustavo Konrad

1/23/2020

Análise exploratória

Começamos carregando pacotes que iremos utilizar e importando os dados para análise.

```
library(readr)
library(caret)
library(e1071)
library(ggplot2)
library(corrplot)
HR_Employee <- read_csv("../data/HR-Employee.csv")
summary(HR_Employee)
```

```
##      Age      Attrition      BusinessTravel      DailyRate
## Min.   :18.00  Length:1470      Length:1470      Min.    : 102.0
## 1st Qu.:30.00  Class :character  Class :character  1st Qu.: 465.0
## Median :36.00  Mode  :character  Mode  :character  Median : 802.0
## Mean   :36.92                                     Mean   : 802.5
## 3rd Qu.:43.00                                     3rd Qu.:1157.0
## Max.   :60.00                                     Max.   :1499.0
## Department      DistanceFromHome      Education      EducationField
## Length:1470      Min.    : 1.000      Min.    :1.000      Length:1470
## Class :character  1st Qu.: 2.000      1st Qu.:2.000      Class :character
## Mode  :character  Median : 7.000      Median :3.000      Mode  :character
##                                     Mean   : 9.193      Mean   :2.913
##                                     3rd Qu.:14.000     3rd Qu.:4.000
##                                     Max.    :29.000     Max.    :5.000
## EmployeeCount EmployeeNumber      EnvironmentSatisfaction      Gender
## Min.    :1      Min.    : 1.0      Min.    :1.000      Length:1470
## 1st Qu.:1      1st Qu.: 491.2      1st Qu.:2.000      Class :character
## Median :1      Median :1020.5      Median :3.000      Mode  :character
## Mean    :1      Mean    :1024.9      Mean    :2.722
## 3rd Qu.:1      3rd Qu.:1555.8      3rd Qu.:4.000
## Max.    :1      Max.    :2068.0      Max.    :4.000
## HourlyRate      JobInvolvement      JobLevel      JobRole
## Min.    : 30.00      Min.    :1.00      Min.    :1.000      Length:1470
## 1st Qu.: 48.00      1st Qu.:2.00      1st Qu.:1.000      Class :character
## Median : 66.00      Median :3.00      Median :2.000      Mode  :character
## Mean    : 65.89      Mean    :2.73      Mean    :2.064
## 3rd Qu.: 83.75      3rd Qu.:3.00      3rd Qu.:3.000
## Max.    :100.00      Max.    :4.00      Max.    :5.000
## JobSatisfaction      MaritalStatus      MonthlyIncome      MonthlyRate
## Min.    :1.000      Length:1470      Min.    : 1009      Min.    : 2094
## 1st Qu.:2.000      Class :character  1st Qu.: 2911      1st Qu.: 8047
```

```
## Median :3.000   Mode  :character   Median : 4919   Median :14236
## Mean   :2.729               Mean   : 6503   Mean   :14313
## 3rd Qu.:4.000               3rd Qu.: 8379   3rd Qu.:20462
## Max.   :4.000               Max.    :19999   Max.    :26999
## NumCompaniesWorked   Over18           OverTime           PercentSalaryHike
## Min.    :0.000       Length:1470       Length:1470       Min.    :11.00
## 1st Qu.:1.000       Class :character   Class :character   1st Qu.:12.00
## Median :2.000       Mode  :character   Mode  :character   Median :14.00
## Mean   :2.693               Mean   :15.21
## 3rd Qu.:4.000               3rd Qu.:18.00
## Max.   :9.000               Max.    :25.00
## PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## Min.    :3.000   Min.    :1.000           Min.    :80   Min.    :0.0000
## 1st Qu.:3.000   1st Qu.:2.000           1st Qu.:80   1st Qu.:0.0000
## Median :3.000   Median :3.000           Median :80   Median :1.0000
## Mean   :3.154   Mean   :2.712           Mean   :80   Mean   :0.7939
## 3rd Qu.:3.000   3rd Qu.:4.000           3rd Qu.:80   3rd Qu.:1.0000
## Max.   :4.000   Max.    :4.000           Max.    :80   Max.    :3.0000
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## Min.    : 0.00   Min.    :0.000           Min.    :1.000   Min.    : 0.000
## 1st Qu.: 6.00   1st Qu.:2.000           1st Qu.:2.000   1st Qu.: 3.000
## Median :10.00   Median :3.000           Median :3.000   Median : 5.000
## Mean   :11.28   Mean   :2.799           Mean   :2.761   Mean   : 7.008
## 3rd Qu.:15.00   3rd Qu.:3.000           3rd Qu.:3.000   3rd Qu.: 9.000
## Max.   :40.00   Max.    :6.000           Max.    :4.000   Max.    :40.000
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## Min.    : 0.000   Min.    : 0.000           Min.    : 0.000
## 1st Qu.: 2.000   1st Qu.: 0.000           1st Qu.: 2.000
## Median : 3.000   Median : 1.000           Median : 3.000
## Mean   : 4.229   Mean   : 2.188           Mean   : 4.123
## 3rd Qu.: 7.000   3rd Qu.: 3.000           3rd Qu.: 7.000
## Max.   :18.000   Max.    :15.000           Max.    :17.000
```

Variâncias próximas de zero

Com as variáveis categóricas codificadas, podemos identificar correlações entre variáveis independentes. Antes disso, no entanto, vamos utilizar a função `nearZeroVar` do pacote `caret` para identificar se temos variáveis com apenas um valor único.

```
zeroVars <- nearZeroVar(HR_Employee)
summary(HR_Employee[zeroVars])
```

```
## EmployeeCount   Over18           StandardHours
## Min.    :1       Length:1470       Min.    :80
## 1st Qu.:1       Class :character   1st Qu.:80
## Median :1       Mode  :character   Median :80
## Mean   :1               Mean   :80
## 3rd Qu.:1               3rd Qu.:80
## Max.   :1               Max.    :80
```

```
HR_Employee <- HR_Employee[-zeroVars]
```

A função `nearZeroVar` identifica, de maneira geral, variáveis com variância próxima de zero. Tais variáveis, que adicionam pouca ou nenhuma informação ao modelo, podem ser descartadas para melhorar a performance no treinamento. Podemos ver acima que as variáveis identificadas possuem variância nula, e portanto poderiam proveitosamente ser descartadas. Caso houvessem variáveis com variância baixa, mas não nula, seu descarte

deve ser considerado com cuidado.

Correlações

Podemos analisar correlações entre variáveis independentes e a variável dependente Attrition, mas antes precisamos gerar dummy variables.

```
dmy <- dummyVars("~.", HR_Employee, fullRank=T)
enc_HR <- data.frame(predict(dmy, HR_Employee))
correlations <- cor(enc_HR)
attrition_corrs <- correlations["AttritionYes",][-2] # removendo autocorrelação
attrition_corrs
```

```
##                Age  BusinessTravelTravel_Frequently
##                -0.1592050069                0.1151427655
##      BusinessTravelTravel_Rarely                DailyRate
##                -0.0495378384                -0.0566519919
## DepartmentResearch...Development                DepartmentSales
##                -0.0852929276                0.0808552021
##                DistanceFromHome                Education
##                0.0779235830                -0.0313728196
##      EducationFieldLife.Sciences                EducationFieldMarketing
##                -0.0327031477                0.0557806657
##                EducationFieldMedical                EducationFieldOther
##                -0.0469987159                -0.0178975168
##      EducationFieldTechnical.Degree                EmployeeNumber
##                0.0693545948                -0.0105772428
##                EnvironmentSatisfaction                GenderMale
##                -0.1033689783                0.0294532532
##                HourlyRate                JobInvolvement
##                -0.0068455496                -0.1300159568
##                JobLevel                JobRoleHuman.Resources
##                -0.1691047509                0.0362150821
##      JobRoleLaboratory.Technician                JobRoleManager
##                0.0982904855                -0.0833163842
##      JobRoleManufacturing.Director                JobRoleResearch.Director
##                -0.0829939241                -0.0888698417
##      JobRoleResearch.Scientist                JobRoleSales.Executive
##                -0.0003595713                0.0197743685
##      JobRoleSales.Representative                JobSatisfaction
##                0.1572342701                -0.1034811261
##                MaritalStatusMarried                MaritalStatusSingle
##                -0.0909836512                0.1754185536
##                MonthlyIncome                MonthlyRate
##                -0.1598395824                0.0151702125
##                NumCompaniesWorked                OverTimeYes
##                0.0434937391                0.2461179942
##                PercentSalaryHike                PerformanceRating
##                -0.0134782021                0.0028887517
##      RelationshipSatisfaction                StockOptionLevel
##                -0.0458722789                -0.1371449189
##                TotalWorkingYears                TrainingTimesLastYear
##                -0.1710632461                -0.0594777986
##                WorkLifeBalance                YearsAtCompany
##                -0.0639390472                -0.1343922140
```

```
##           YearsInCurrentRole           YearsSinceLastPromotion
##           -0.1605450043           -0.0330187751
##           YearsWithCurrManager
##           -0.1561993159
```

```
max(attrition_corrs)
```

```
## [1] 0.246118
```

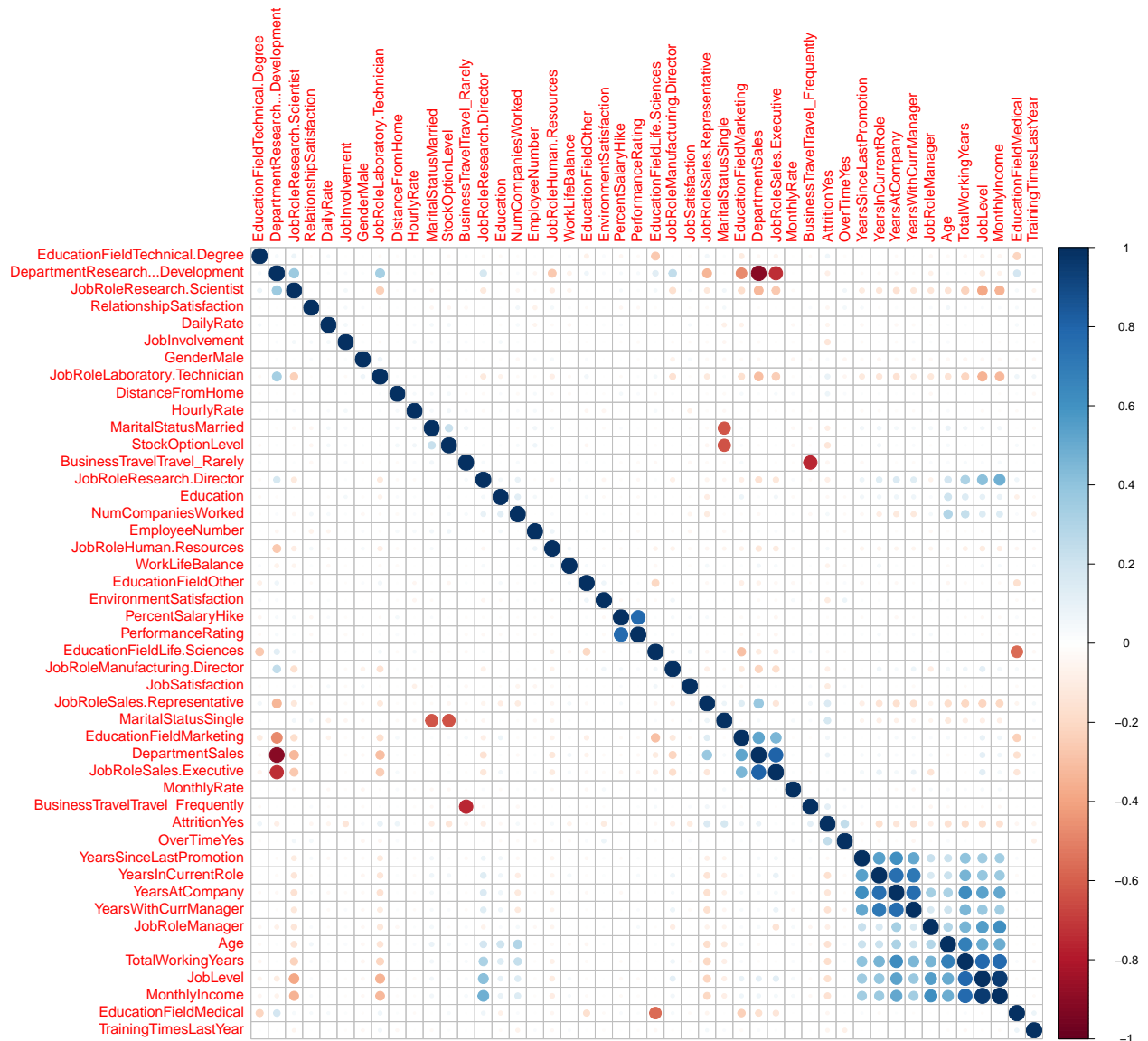
```
min(attrition_corrs)
```

```
## [1] -0.1710632
```

Muitas variáveis parecem contribuir em alguma medida para a variação em AttritionYes, com a variável OverTimeYes tendo a maior correlação positiva - indicando que sobrecarregamento do funcionário pode ser um dos principais fatores para a saída do mesmo - e a variável TotalWorkingYears tendo a maior correlação negativa - indicando que funcionários que estão à mais tempo no mercado de trabalho podem ter menor disposição à sair do seu emprego atual.

Podemos também plotar clusters de variáveis correlacionadas utilizando o pacote corrplot. Exportamos a imagem para um arquivo png para melhor visualização.

```
corrMatrix <- corrplot(correlations, order="hclust", tl.cex=1)
```



```
dev.print(file="corrplot.png", device=png, width=1920, height=1080)
```

```
## pdf
## 2
```

Identificamos correlações esperadas entre variáveis que indicam o tempo corrido desde algum evento (anos desde que entrou na companhia, anos desde a última promoção, etc). Correlações entre idade, tempo no mercado de trabalho e salário mensal também não são inesperadas. Caso tais correlações venham a ser problemáticas (como podem ser no caso de modelos lineares), ou caso queiramos otimizar o modelo, podemos aplicar Principal Component Analysis para gerar features independentes entre si.

Assimetria

Para verificar se temos features com distribuições assimétricas, utilizamos a função skewness do pacote e1071.

```
skewValues <- apply(enc_HR, 2, skewness)
skewValues
```

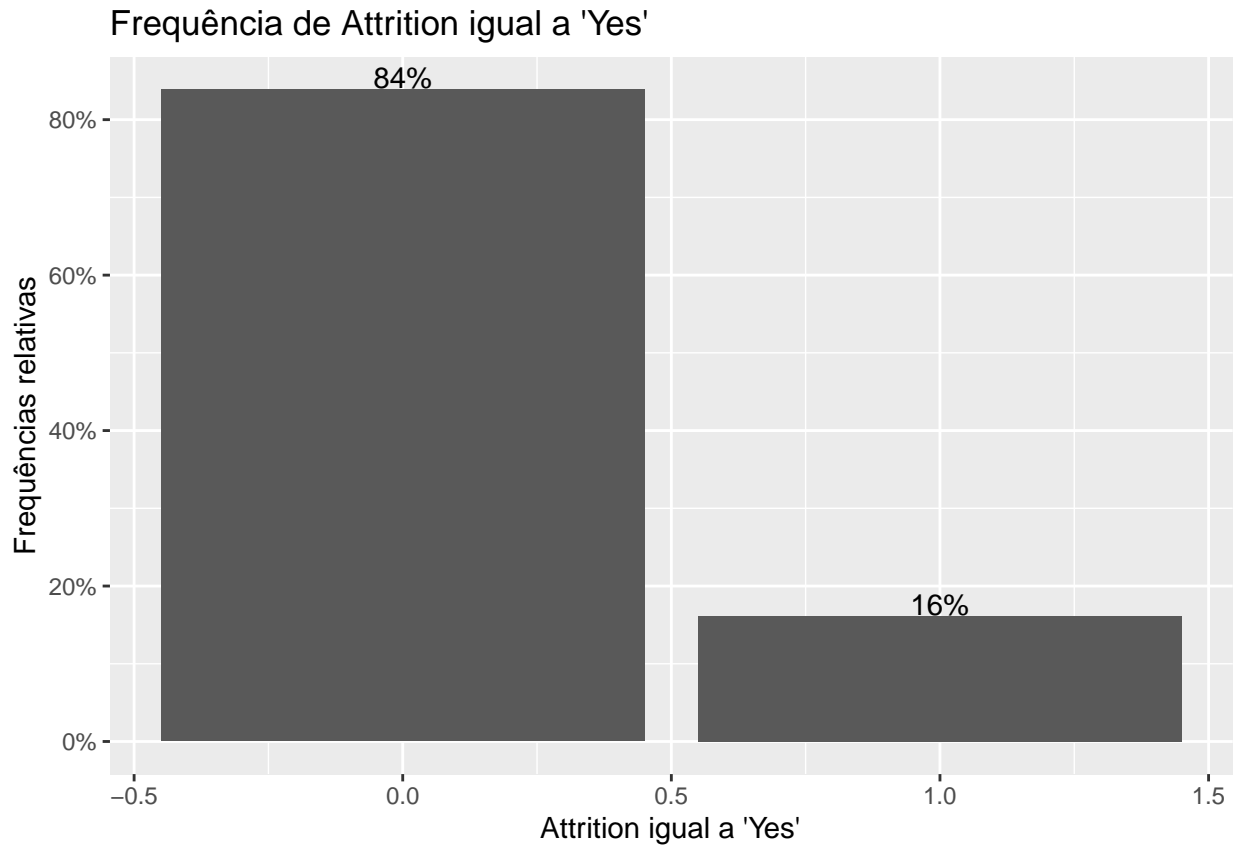
```
##                               Age                               AttritionYes
```

```
##          0.412443243          1.840603819
## BusinessTravelTravel_Frequently      BusinessTravelTravel_Rarely
##          1.591813043          -0.922106985
##          DailyRate      DepartmentResearch...Development
##          -0.003511391          -0.645616381
##          DepartmentSales          DistanceFromHome
##          0.854411573          0.956163540
##          Education          EducationFieldLife.Sciences
##          -0.289090164          0.356191223
##          EducationFieldMarketing          EducationFieldMedical
##          2.520630939          0.792497676
##          EducationFieldOther      EducationFieldTechnical.Degree
##          3.867214035          2.866744428
##          EmployeeNumber          EnvironmentSatisfaction
##          0.016540210          -0.320998308
##          GenderMale          HourlyRate
##          -0.407831781          -0.032245042
##          JobInvolvement          JobLevel
##          -0.497402643          1.023309576
##          JobRoleHuman.Resources      JobRoleLaboratory.Technician
##          5.025364918          1.698132940
##          JobRoleManager      JobRoleManufacturing.Director
##          3.385690560          2.689346485
##          JobRoleResearch.Director      JobRoleResearch.Scientist
##          3.924421023          1.509128923
##          JobRoleSales.Executive      JobRoleSales.Representative
##          1.338098805          3.839343825
##          JobSatisfaction          MaritalStatusMarried
##          -0.328999464          0.169138192
##          MaritalStatusSingle          MonthlyIncome
##          0.772295727          1.367022404
##          MonthlyRate          NumCompaniesWorked
##          0.018539911          1.024377223
##          OverTimeYes          PercentSalaryHike
##          0.962521412          0.819452964
##          PerformanceRating      RelationshipSatisfaction
##          1.917962271          -0.302209830
##          StockOptionLevel          TotalWorkingYears
##          0.967003703          1.114892944
##          TrainingTimesLastYear          WorkLifeBalance
##          0.551995858          -0.551353300
##          YearsAtCompany          YearsInCurrentRole
##          1.760930007          0.915491836
##          YearsSinceLastPromotion      YearsWithCurrManager
##          1.980242248          0.831750843
```

AttritionYes, que codifica Attrition, nossa variável target, é uma das variáveis que apresenta assimetria. Plotamos sua frequência relativa para inspeção visual.

```
ggplot(enc_HR, aes(x = AttritionYes)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels = scales::percent) +
  geom_text(aes(label= scales::percent(..prop..), y=..prop..), stat="count", vjust = -.075) +
  xlab("Attrition igual a 'Yes'") +
  ylab("Frequências relativas") +
```

```
ggtitle("Frequência de Attrition igual a 'Yes'")
```



```
dev.print(file="attrition.png", device=png, width=800)
```

```
## pdf  
## 2
```

Fica evidente que Attrition é igual a “Yes” (ou AttritionYes == 1) em apenas 16% da população. Um algoritmo que estimasse Attrition = “No” para todo e qualquer caso teria, portanto, uma exatidão próxima de 84% neste dataset. Dada tal assimetria, podemos inferir que uma medida de performance baseada não na exatidão, mas na precisão e recall (identificação correta de negativos e positivos vs. falsos negativos e falsos positivos), como a estatística F1 (média harmônica entre precisão e recall), pode representar melhor a qualidade do modelo.

Além disso, na presença de variáveis com distribuições altamente assimétricas, poderíamos, caso fosse conveniente, aplicar a transformação Box-Cox no pré-processamento dos dados para corrigir tal assimetria.