EmployeePromotion

2023-02-18

Poner aquí documentación

Este proyecto está realizado por el grupo 3 formado por José Arturo Espaillat, Johnsiel Castaños, José Delgado, Salama Mohamed-fadel Sidna. El dataset elegido es llamado HR Analytics: Employer Promotion Data (https://www.kaggle.com/datasets/arashnic/hr-ana?select=test.csv) sacado de Kaggle, el cual cuenta con X columnas y X filas para analizar si un empleado será o no promocionado. Hablar más de esto, mirar Rúbica

Para organizarnos, hemos decidido crear un repositorio en GitHub (https://github.com/Josdelser/DataScienceEmployeePromotion/tree/develop) para trabajar simultáneamente. También hemos creado una bolsa de preguntas y las hemos asignado según las capacidades y aspiraciones de cada persona. En el caso de que algún miembro quiera obtener una calificación más alta, deberá realizar un mayor número de preguntas individuales. Todos los miembros realizarán una pregunta para la parte grupal y se ha intentado que estas preguntas estén relacionadas con las preguntas individuales, para poder afrontar mejor el reto. A continuación, se detalla la bolsa de preguntas y la asignación de cada miembro.

Jose Delgado:

-Grupal

```
¿Como predecir si será promovido un empleado?
```

-Individual:

```
¿Qué variables afectan fuertemente a la promoción?

¿Nos puede ayudar la visualización a sacar conclusiones temprana?

¿Existe alguna correlacion con variables que tipicamente la tienen como la edad, el tiempo
en la empresa, la eduación...?
```

Cada miembro realizará la documentación de ambas partes asignadas. Además en el documento se indacará donde empieza y acaba la parte de cada miembro para que así sea mas sencilla su evaluación.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Aquí empieza la parte de Jose Delgado

Información general del Dataset

```
data<- read.csv("train.csv")
data <- rename(data, awards_won = awards_won.)
head(data)
```

```
##   employee_id        department    region        education gender
## 1       65438 Sales & Marketing  region_7 Master's & above      f
## 2       65141        Operations region_22       Bachelor's      m
## 3        7513 Sales & Marketing region_19       Bachelor's      m
## 4        2542 Sales & Marketing region_23       Bachelor's      m
## 5       48945        Technology region_26       Bachelor's      m
## 6       58896         Analytics  region_2       Bachelor's      m
##   recruitment_channel no_of_trainings age previous_year_rating
## 1           sourcing               1  35                    5
## 2              other               1  30                    5
## 3           sourcing               1  34                    3
## 4              other               2  39                    1
## 5              other               1  45                    3
## 6           sourcing               2  31                    3
##   length_of_service awards_won avg_training_score is_promoted
## 1                 8          0                 49           0
## 2                 4          0                 60           0
## 3                 7          0                 50           0
## 4                10          0                 50           0
## 5                 2          0                 73           0
## 6                 7          0                 85           0
```

```
colnames(data)
```

```
##  [1] "employee_id"          "department"           "region"
##  [4] "education"            "gender"               "recruitment_channel"
##  [7] "no_of_trainings"      "age"                  "previous_year_rating"
## [10] "length_of_service"    "awards_won"           "avg_training_score"
## [13] "is_promoted"
```

```
attach(data)
```

```
str(data)
```

```
## 'data.frame':    54808 obs. of  13 variables:
##  $ employee_id         : int  65438 65141 7513 2542 48945 58896 20379 16290 73202 28911 ...
##  $ department          : chr  "Sales & Marketing" "Operations" "Sales & Marketing" "Sales & Marketing
##  $ region              : chr  "region_7" "region_22" "region_19" "region_23" ...
##  $ education           : chr  "Master's & above" "Bachelor's" "Bachelor's" "Bachelor's" ...
##  $ gender              : chr  "f" "m" "m" "m" ...
##  $ recruitment_channel : chr  "sourcing" "other" "sourcing" "other" ...
##  $ no_of_trainings     : int  1 1 1 2 1 2 1 1 1 1 ...
##  $ age                 : int  35 30 34 39 45 31 31 33 28 32 ...
##  $ previous_year_rating: num  5 5 3 1 3 3 3 3 4 5 ...
```

```
##  $ length_of_service   : int  8 4 7 10 2 7 5 6 5 5 ...
##  $ awards_won          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ avg_training_score  : int  49 60 50 50 73 85 59 63 83 54 ...
##  $ is_promoted         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
dim(data)
```

```
## [1] 54808    13
```

```
object.size(data)/1024
```

```
## 4073.3 bytes
```

```
print(paste("Número de valores faltantes totales:", sum(is.na(data))))
```

```
## [1] "Número de valores faltantes totales: 4124"
```

```
data_train <- na.omit(data)
data_train <- unique(data)
```

Una vez cargado el dataset y tratado minimamente vamos a pasar a hacer un analisis exploratorio para ver por donde van los tiros antes de hacer la prediccion y así ver variables que puedan afectar directamente.

Primero vamos a ver las proporciones de las variables. En la caso de las numericas vamos a sacar informacion estadistica como la mediana, max, min... En el caso de las categoricas veremos el numeros de observaciones según las categorias de la variable

```
summary(data)
```

```
##   employee_id     department          region           education
##  Min.   :    1   Length:54808       Length:54808       Length:54808
##  1st Qu.:19670   Class :character   Class :character   Class :character
##  Median :39226   Mode  :character   Mode  :character   Mode  :character
##  Mean   :39196
##  3rd Qu.:58731
##  Max.   :78298
##
##     gender          recruitment_channel no_of_trainings       age
##  Length:54808       Length:54808        Min.   : 1.000   Min.   :20.0
##  Class :character   Class :character    1st Qu.: 1.000   1st Qu.:29.0
##  Mode  :character   Mode  :character    Median : 1.000   Median :33.0
##                                         Mean   : 1.253   Mean   :34.8
##                                         3rd Qu.: 1.000   3rd Qu.:39.0
##                                         Max.   :10.000   Max.   :60.0
##
##  previous_year_rating length_of_service   awards_won      avg_training_score
##  Min.   :1.000        Min.   : 1.000    Min.   :0.00000   Min.   :39.00
##  1st Qu.:3.000        1st Qu.: 3.000    1st Qu.:0.00000   1st Qu.:51.00
##  Median :3.000        Median : 5.000    Median :0.00000   Median :60.00
##  Mean   :3.329        Mean   : 5.866    Mean   :0.02317   Mean   :63.39
##  3rd Qu.:4.000        3rd Qu.: 7.000    3rd Qu.:0.00000   3rd Qu.:76.00
```

```
##  Max.   :5.000        Max.   :37.000    Max.   :1.00000   Max.   :99.00
##  NA's   :4124
##   is_promoted
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08517
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

```
table(data$department)
```

```
##
##         Analytics           Finance                HR             Legal
##              5352              2536              2418              1039
##        Operations       Procurement               R&D Sales & Marketing
##             11348              7138               999             16840
##        Technology
##              7138
```

```
table(data$region)
```

```
##
##  region_1 region_10 region_11 region_12 region_13 region_14 region_15 region_16
##       610       648      1315       500      2648       827      2808      1465
## region_17 region_18 region_19  region_2 region_20 region_21 region_22 region_23
##       796        31       874     12343       850       411      6428      1175
## region_24 region_25 region_26 region_27 region_28 region_29  region_3 region_30
##       508       819      2260      1659      1318       994       346       657
## region_31 region_32 region_33 region_34  region_4  region_5  region_6  region_7
##      1935       945       269       292      1703       766       690      4843
##  region_8  region_9
##       655       420
```

```
table(data$education)
```

```
##
##                         Bachelor's  Below Secondary Master's & above
##             2409             36669               805             14925
```

```
table(data$gender)
```

```
##
##     f     m
## 16312 38496
```

```
table(data$recruitment_channel)
```

```
##
##    other referred sourcing
##    30446     1142    23220
```

```
table(data$awards_won)
```

```
##
##     0     1
## 53538  1270
```

Aqui poner conclusiones según esto