Proiect sda

Duțu Cristina, Gavrilă Simona, Marinescu Sorina

1.Descrierea problemei

Compania care constituie subiectul de cercetare în cadrul proiectului se numește IBM- International Business Machines (Corporation). Veterană în domeniul tehnologiei, aceasta și-a pus amprenta în istoria marilor corporații prin invenții fără de care cele mai uzuale activități omenești nu ar mai fi posibile. În prezent problema principală cu care se confruntă este reprezentată de creșterea epuizării angajaților din cauza volumului mare de muncă.

Cu ajutorul acestui set de date, care reflectă performanța angajaților IBM, ne propunem să realizăm o previziune a numărului de angajați care vor deveni epuizați. Pentru aceasta, vom aplica regresia logistică binomială în funcție de venitul lunar și vechime; și KNN asociat problemei de clasificare.

2.Importanța problemei

De ce e importantă problema pentru firmă?

Productivitatea angajaților se reflectă în mod direct în performanța companiei. Un angajat suprasolicitat nu își îndeplinește sarcinile în mod corespunzător ceea ce va afecta rezultatele companiei într-un mod negativ. Astfel, IBM nu se va ridica la așteptările clienților, crescând riscul de a-și pierde clientela.

De ce e importantă problema pentru noi?

Din perspectiva viitorilor angajați există reținerea de a intra într-un mediu toxic de supraexploatare a personalului.

În concluzie, problema identificată afectează performanța firmei atât în plan intern cât și extern.

3.Ce măsuri s-au luat pentru combaterea problemei?

Măsuri la nivel global

În cadrul tuturor companiilor există riscul de suprasolicitare a angajaților, motiv pentru care au trebuit să adopte diverse măsuri precum: comunicarea constantă între angajați și conducere, suport, beneficii salariale, oferirea de tool-uri pentru eficientizarea sarcinilor de muncă.

Măsuri luate de IBM

Pentru a reduce riscul epuizării angajaților IBM a luat următoarele măsuri: training, program flexibil de lucru, posibilitatea de a lucra online, beneficii egale pentru toți angajații indiferent de rasă/ etnie.

Studii relevante cu privire la analizarea gradului de epuizare al angajaților IBM

În studiul "IBM Employee Attrition Analysis", realizat de Shenghuan Yang de la Jiangxi University of Finance and Economics și Md Tariqul Islam de la Syracuse University, s-au căutat principalele motive pentru care angajații aleg să demisioneze din IBM. Metodele utilizate de aceștia s-au concentrat pe exploatarea arborilor de decizie, găsind venitul lunar, vârsta și numărul de companii la care au lucrat ca având un impact semnificativ asupra angajaților. În continuare au clasificat persoanele în două grupuri prin utilizarea k-means clustering. În cele din urmă au efectuat o analiză cantitativă și anume o regresie logistică binară din care a rezultat că uzura persoanelor care au călătorit frecvent a fost de 2,4 ori mai mare decât cea a persoanelor care călătoreau rar. De asemenea, au constatat faptul că angajații care lucrează în domeniul resurselor umane au o tendință mai mare de a pleca.

4. Soluție + rezultate

Respectivul set de date a fost preluat de pe data.world și conține 14 atribute și 1470 de entități/indivizi. Vârsta, frecvența călătoriilor în interes de serviciu și sexul sunt atribute categoriale formate din 2 sau 3 clase, iar restul de coloane conține date cantitative. Scopul cercetării în această zonă este acela de a determina ce factori contribuie predominant la apariția uzurii utilizând tehnici de învățare precum clasificarea și regresia pentru a crea modele de predicție din datele colectate. Am prelucrat setul de date, adăugând o nouă coloană denumită EmployeeNr astfel încât fiecare angajat să aibă un id unic.

```
##Pachete
#install.packages("tidyverse")
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.1
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
#install.packages("dplyr")
library(dplyr)
#install.packages("caTools")
library(caTools)
#install.packages("class")
library(class)
#install.packages("rpart")
library(rpart)
#install.packages("rpart.plot")
library(rpart.plot)
#install.packages("e1071")
library(e1071)
library(class)
```

```r
## Incarcarea datelor
dateSDA<-read.csv("date_proiect_sda.csv", header=TRUE, sep=",")
View(dateSDA)
```

```r
## Prelucrarea setului de date
mat <- matrix (c(1:1470),nrow=1470, ncol=1, byrow=FALSE)
#View(mat)
colnames(mat) <- c("EmployeeNr")
date <- data.frame (mat, dateSDA)
#View(date)
date_proiect <- as_tibble(date)
str(date_proiect)
```

```
tibble [1,470 x 14] (S3: tbl_df/tbl/data.frame)
 $ EmployeeNr             : int [1:1470] 1 2 3 4 5 6 7 8 9 10 ...
 $ Age                    : int [1:1470] 41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition              : chr [1:1470] "Yes" "No" "Yes" "No" ...
 $ BusinessTravel         : chr [1:1470] "Travel_Rarely" "Travel_Frequently" "Travel_Rarely"
```

```
$ DistanceFromHome       : int [1:1470] 1 8 2 3 2 2 3 24 23 27 ...
$ Gender                 : chr [1:1470] "Female" "Male" "Male" "Female" ...
$ MonthlyIncome          : int [1:1470] 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 .
$ NumCompaniesWorked     : int [1:1470] 8 1 6 1 9 0 4 1 0 6 ...
$ TotalWorkingYears      : int [1:1470] 8 10 7 8 6 8 12 1 10 17 ...
$ TrainingTimesLastYear  : int [1:1470] 0 3 3 3 3 2 3 2 2 3 ...
$ YearsAtCompany         : int [1:1470] 6 10 0 8 2 7 1 1 9 7 ...
$ YearsInCurrentRole     : int [1:1470] 4 7 0 7 2 7 0 0 7 7 ...
$ YearsSinceLastPromotion: int [1:1470] 0 1 0 3 2 3 0 0 1 7 ...
$ YearsWithCurrManager   : int [1:1470] 5 7 0 0 2 6 0 0 8 7 ...
```

4.1.Regresia logistică binomială

Pentru aplicarea regresiei logistice de tip binomial, am utilizat, ca variabila dependenta, coloana Attrition, iar ca variabile independente, MonthlyIncome (venitul solicitantului) si TotalWorkingYears (numar ani vechime).

```
table(date_proiect$Attrition)
```

```
  No  Yes
1233  237
```

În cadrul setului de date, 1233 de persoane nu sunt epuizate, iar 237 de persoane resimt oboseala din cauza volumului mare de muncă.

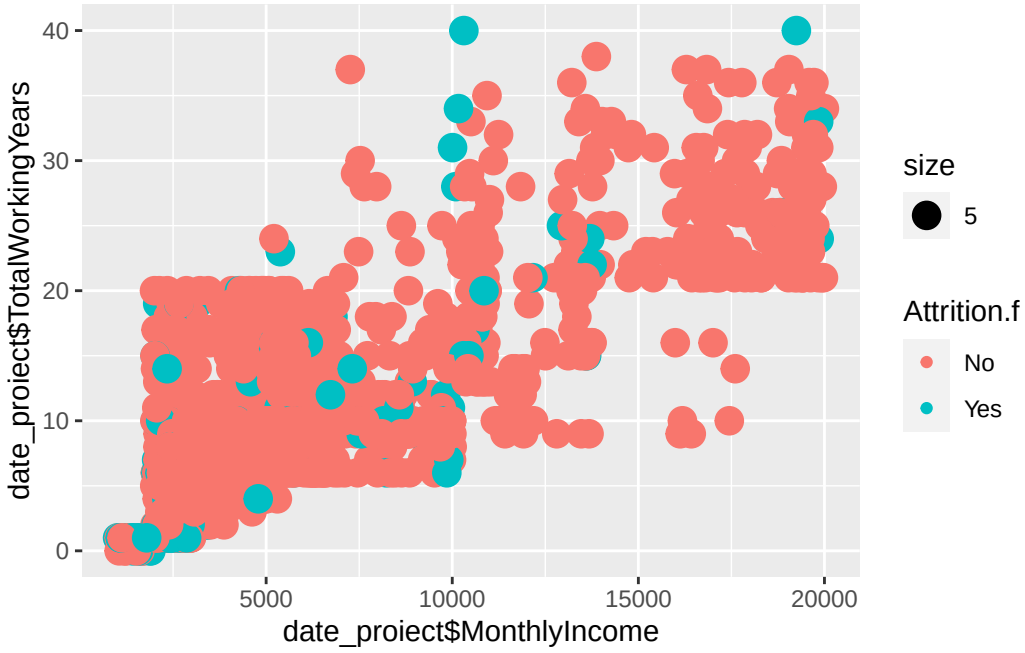
```
# Crearea variabilei factor
date_proiect$Attrition.f<-factor(date_proiect$Attrition)
View(date_proiect)

# Crearea setului de antrenare si testare
set.seed(88)
View(date_proiect)
split=sample.split(date_proiect$Attrition.f, SplitRatio=0.75)
setantrenare<-subset(date_proiect, split==TRUE)
settestare<-subset(date_proiect, split==FALSE)
```

```
plot<-ggplot(data=date_proiect, aes(x=date_proiect$MonthlyIncome, y=date_proiect$TotalWork
col=Attrition.f))
plot<-plot+geom_point(aes(size=5))
```

```
plot
```

```
Warning: Use of `date_proiect$MonthlyIncome` is discouraged.
i Use `MonthlyIncome` instead.
```

```
Warning: Use of `date_proiect$TotalWorkingYears` is discouraged.
i Use `TotalWorkingYears` instead.
```



În figura de mai sus, se poate observa și vizual distribuția persoanelor epuizate și în formă, în funcție de venit și vechimea în muncă.

```
## Model de regresie binomiala
model<-glm(Attrition.f~MonthlyIncome+TotalWorkingYears, data=setantrenare, family=binomial
summary(model)
```

```
Call:
glm(formula = Attrition.f ~ MonthlyIncome + TotalWorkingYears,
    family = binomial, data = setantrenare)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8259  -0.6594  -0.5633  -0.3299   2.5840

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.297e-01  1.581e-01  -5.249 1.53e-07 ***
MonthlyIncome     -6.704e-05  3.416e-05  -1.962   0.0497 *
TotalWorkingYears -4.453e-02  1.937e-02  -2.299   0.0215 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 974.94  on 1102  degrees of freedom
Residual deviance: 937.82  on 1100  degrees of freedom
AIC: 943.82

Number of Fisher Scoring iterations: 5
```

```
exp(coef(model))
```

```
      (Intercept)     MonthlyIncome TotalWorkingYears
        0.4361678         0.9999330         0.9564453
```

```
contrasts(date_proiect$Attrition.f)
```

```
    Yes
No    0
Yes   1
```

Prin intermediul acestui output pot fi evidențiate următoarele idei:

1.Forma ecuației de regresie:

Ln(p/1−p) = -0.82972823 – 0.00006704 * MonthlyIncome - 0.04453172 * TotalWorkingYears

Unde: p = probabilitatea ca o persoană să nu fie epuizata (Attrition = No), conform venitului și a vechimii;

1-p = probabilitatea ca o persoană să fie epuizată (Attrition = Yes), conform venitului și a vechimii.

Această ecuație indică faptul că ambele variabile afectează în mod negativ raportul Ln(p/1−p). Astfel, o creștere cu o unitate a venitului generează scăderea raportului șanselor ca o persoana sa fie epuizata cu 0.00006704 unități. Asemanator, o crestere cu o unitate a vechimii produce o scădere a șanselor ca persoana să fie epuizată cu 0.04453172 unități. Din punctul nostru de vedere, rezultatele obținute nu corespund așteptărilor noastre.

2.Devianța reziduală în valoare de 937.82 este mai mică decât devianța nulă egală cu 974.94, ceea ce înseamnă că modelul este unul relativ bun.

3.În ceea ce privește coeficienții funcției exponențiale, putem afirma faptul că: – Probabilitatea (Șansele) ca o persoană să nu fie epuizată crește de 0.9999330 ori dacă venitul persoanei crește cu 1 unitate. Raportul șanselor = p/1−p = 0.9999330
– Probabilitatea (Șansele) de a nu achita împrumutul crește de 0.9564453 ori dacă rata dobânzii crește cu 1 unitate. Raportul șanselor = □/ 1−p = 0.9564453

```
prob<-predict(model, settestare, type="response")
pred<-rep("Yes", dim(setantrenare)[1])
pred[prob>.5]="No"
# Matricea de confuzie pentru setul de antrenare
table(pred, setantrenare$Attrition.f)
```

```
pred   No Yes
  Yes 925 178
```

```
pred1<-rep("Yes", dim(settestare)[1])
pred1[prob>.5]="No"
# Matricea de confuzie pentru setul de testare
table(pred1, settestare$Attrition.f)
```

```
pred1  No Yes
  Yes 308  59
```

4.2. KNN pentru clasificare

Pentru această problemă vom ține cont de ultimele coloane din setul nostru de date și anume”MonthlyIncome”,“NumCompa
Year”,“YearsAtCompany”,YearsInCurrentRole”,“YearsSinceLastPromotion”,“YearsWithCurrM an-
ager”, și de asemenea vom considera coloana “Attrition” drept etichetă.

```
date1 <- date_proiect[,-c(1,2,3,4,5,6)]
View(date1)
date1<-data.frame(date1,date$Attrition)
View(date1)
```

Extragem două eșantioane, corespunzătoare setului de antrenare în proporție de 80% din date și setului de testare în proporție de 20%. Dimensiunea setului de antrenare este de 1186 observații, iar a setului de testare de 284.

```
set.seed(123)
ind<-sample(2, nrow(date1), replace=TRUE, prob=c(0.80,0.20))
antrenare<-date1[ind==1,1:8]
dim(antrenare)[1]
```

```
[1] 1186
```

```
testare<-date1[ind==2,1:8]
dim(testare)[1]
```

```
[1] 284
```

```
etichetaantrenare<- date1[ind==1,9]
head(etichetaantrenare, 10)
```

```
 [1] Yes No  Yes No  No  No  No  No  No  No
Levels: No Yes
```

```
length(etichetaantrenare)
```

```
[1] 1186
```

```
etichetatestare<- date1[ind==2,9]
head(etichetatestare, 10)
```

```
 [1] No No No No No No No No No No
Levels: No Yes
```

```
length(etichetatestare)
```

```
[1] 284
```

Fixăm cei mai apropiați patru vecini ai oricărui obiect (k=4) și realizăm predicția. Matricea de confuzie ne comunică faptul că pe diagonala principală se află numărul de observații corect previzionate, iar în rest sunt erori. Mai exact 226 de angajați au fost corect catalogați ca fiind încă în putere, iar 9 angajați au fost corect catalogați ca fiind epuizați. 17,25% din observațiile setului de testare sunt incorect clasificate.

```
predictie<-knn(train=antrenare, test=testare, cl=etichetaantrenare, k=4)
head(predictie, 10)
```

```
 [1] No No No No No No No No No No
Levels: No Yes
```

```
tab<-table(valori_reale=etichetatestare, valori_previzionate=predictie)
head(tab, 10)
```

```
            valori_previzionate
valori_reale  No Yes
         No  226  12
         Yes  37   9
```

```
mean(etichetatestare!=predictie)
```

```
[1] 0.1725352
```

Rata de acuratețe a setului de testare este de 17,25%.

5. Concluzii

În urma aplicării regresiei logistice binomiale, modelul creat are toți cei trei coefiecienți semnificativi, asta însemnând că o creștere cu o unitate a venitului generează scăderea raportului șanselor ca o persoana sa fie epuizata cu 0.00006704 unități. Asemanator, o crestere cu o unitate a vechimii produce o scădere a șanselor ca persoana să fie epuizată cu 0.04453172 unități. De asemenea, previziunea realizată pe baza modelului ne indică, prin intermediul matricii de confuzie, faptul că, pe setul de antrenare,

au fost previzionate corect 178 de persoane ca fiind epuizate, iar pe setul de testare, 59 de persoane. Deci, modelul nu este unul foarte relevant. Pe de altă parte, cu ajutorul analizei KNN pentru problema de clasificare, setând, în prealabil, 4 cei mai apropiați vecini, am obținut o previzionare corecta pentru 235 de persoane dintre care 226 nu sunt epuizate și 9 sunt epuizate, conform matricii de confuzie. Din cauza faptului că rata de acurateţe a setului de testare este mică, nu ne putem încrede pe deplin în analiză. Pe viitor, pentru previziuni mai bune se pot incerca arbori de decizie, regresia logistica multinomiala si o analiza SVM de clasificare.