Proiect Stiinta datelor in afaceri

Impactul diferitilor factori asupra tehnologiilor bazate pe containere -Dragomir Anca, Dragomir Gabriel, Dumittrescu Razvan, Iana Delia-Cristina

Motivatia alegerii

Stack Overflow annual Developer Survey este cea mai mare aplicatie de sondaje din lume ce include oameni care codeaza.

În fiecare an, ei creaza un sondaj care include mai multe topice, de la tehnologiile preferate ale dezvoltatorilor, pana la preferințele lor profesionale. Anul 2022 marchează al unsprezeceleaan în care au fost publicat rezultatele anuale ale sondajului iar numarul persoanelor care participa creste de la an la an.

In prezent, sunt disponibile multe opțiuni de filtrare utile, cum ar fi țara și sexul populatiei esantionului.

Exista 3 domenii principale de pe urma carora utilizatorii pot beneficia:

1. Tehnologie - Pentru a identifica instrumentele utilizate în mod obișnuit (de exemplu, limbaje și platforme de programare) și potențialul acestora. În plus, vom descoperi care sunt cele mai indragite si dorite platforme ale momentului potrivit esantionului.
2. Salariu - Pentru a afla distribuția salarială a diferitelor tipuri de dezvoltatori și modul în care acestea pot fi influențate de diplome și modul în care aceasta diferă de la o țară la alta.
3. analiza jobului - Pentru a identifica factorii care influenteaza alegerea locurilui de muncă și cum acesta diferă între bărbat și femeie și variază de la de la o țară la alta.

   Problema pe care vrem sa o rezolvam

   Sa se stabileasca, pentru un esantion de angajati din America de Nord, daca folosirea tehnologiilor bazate pe containere (docker / kubernetes) este influentata de:

- categoria de generatie a respondentilor (BabyBoomers, Boomers, GenX, GenZ, Millenials);

- dimensiunea organizatiei in cadrul careia lucreaza (small, medium, large);

- posesia de cunostinte din domeniul cloud (YES / NO).

Datele modelului de regresie

Pentru date am ales site-ul Stack Overflow Insights - Developer Hiring, Marketing, and User Research, si anume fisierul csv survey_results_public cu datele din 2022.

Ce solutii au fost incercate de alti oameni

Problema pe care noi am incercat sa o rezolvam nu a fost aprofundata de cercetatori inca, insa Ang Wei Xuan Dion, David Chow Jing Shan, Peh Anqi au facut o analiza asupra rezultatelor obtinute din urma sondajului si o regresie care sa arate impactul experientei profesionale asupra salariului si pozitiei in cazul developerilor. Sursa de date: IS428-Group3_DevBuzz_Research_Paper.pdf (smu.edu.sg)

```
knitr::opts_chunk$set(echo = TRUE)

library(ggplot2)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.3.0      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
v purrr   1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

```
#Initilizare setul de date.
data_2022<-read.csv("C:\\Users\\Admin\\Desktop\\survey_results_public.csv", header=TRUE, s

dim(data_2022)
```

```
[1] 83439    48
```

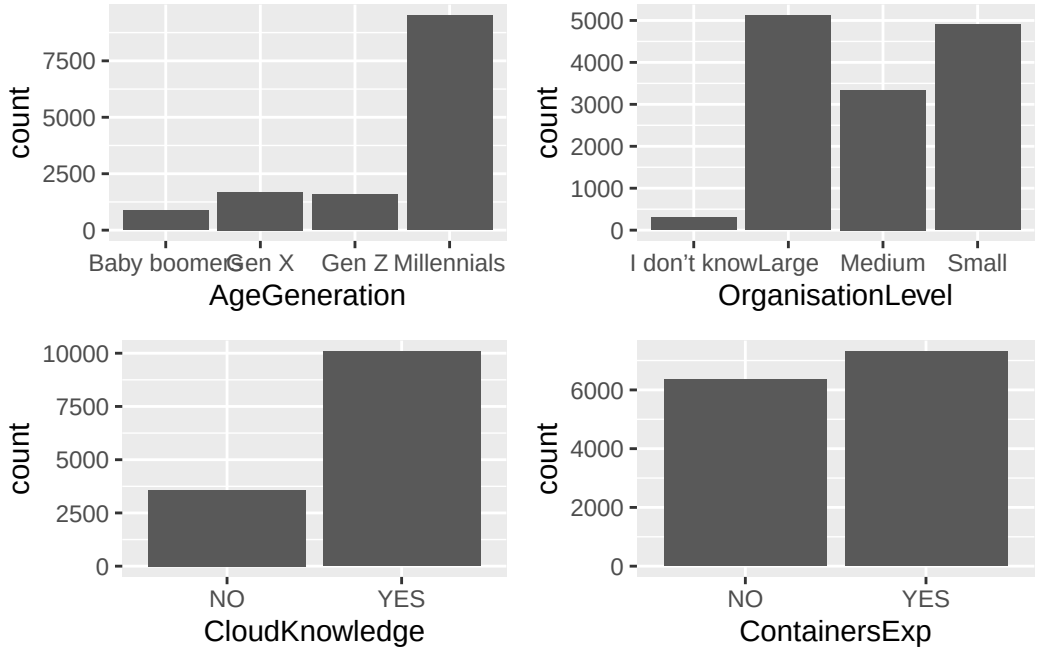
Filtram respondentii angajati din America de Nord.

```
library(dplyr)

employed_statuses = c("Employed full-time",
                      "Employed part-time",
                      "Independent contractor, freelancer, or self-employed")

selected_countries = c("United States of America", "Canada")


data_2022 <- data_2022 %>%
  filter(Employment %in% employed_statuses
         & Country %in% selected_countries)

dim(data_2022)
```

```
[1] 14731    48
```

```
#Calculam generatia din care face parte fiecare respondent intr-o variabila factor.

library(dplyr)

data_2022 <- data_2022 %>%
  filter(Age!= "Prefer not to say") %>%
  mutate("AgeGeneration" = as.factor(case_when(Age %in% c("55-64 years old", "65 years or
                                               Age %in% c("45-54 years old") ~ "Gen X",
                                               Age %in% c("25-34 years old", "35-44 years
                                               Age %in% c("Under 18 years old", "18-24 yea
                                               TRUE ~ Age)))

#Calculam nivelul organizatiei din care face parte fiecare respondent intr-o variabila fac

data_2022 <- data_2022 %>%
  filter(!is.na(OrgSize) & OrgSize != "I donâ€ t know") %>%
  mutate("OrganisationLevel" = as.factor(case_when(OrgSize %in% c("Just me - I am a freela
                                                   OrgSize %in% c("100 to 499 employees",
                                                   OrgSize %in% c("1,000 to 4,999 employee
                                                   TRUE ~ OrgSize)))

#Calculam nr. de respondenti care poseda / nu poseda cunostinte din domeniul cloud intr-o
```

```r
data_2022 <- data_2022 %>%
  mutate("CloudKnowledge" = as.factor(ifelse(!is.na(PlatformHaveWorkedWith), "YES", "NO"))

#Calculam nr. de respondenti care au / nu au experienta cu tehnologiile bazate pe containe

data_2022 <- data_2022 %>%
  mutate("ContainersExp" = as.factor(ifelse(grepl("Docker", ToolsTechHaveWorkedWith, fixed
                                            | grepl("Kubernetes", ToolsTechHaveWorkedWith,

#Pastram doar coloanele necesare formularii ecuatiei de regresie logistica.

data_2022 <- data_2022 %>%
  select(AgeGeneration, OrgSize, OrganisationLevel, CloudKnowledge, ContainersExp)

#Realizam sanity checks prin vizualizarea graficelor de distributie pentru variabilele exp

library(cowplot)

plot_AgeGeneration <-ggplot(data = data_2022, aes(x = AgeGeneration)) +
  geom_bar()

plot_OrganisationLevel <-ggplot(data = data_2022, aes(x = OrganisationLevel)) +
  geom_bar()

plot_CloudKnowledge <-ggplot(data = data_2022, aes(x = CloudKnowledge)) +
  geom_bar()

plot_ContainersExpl <-ggplot(data = data_2022, aes(x = ContainersExp)) +
  geom_bar()


plot_grid(plot_AgeGeneration, plot_OrganisationLevel, plot_CloudKnowledge, plot_Containers
```




```
summary(data_2022 %>%
          select(c(AgeGeneration, OrganisationLevel, CloudKnowledge, ContainersExp)
          ))
```

```
     AgeGeneration     OrganisationLevel CloudKnowledge ContainersExp
 Baby boomers: 866   I don't know: 299    NO : 3575      NO :6350
 Gen X       :1688   Large       :5122    YES:10088      YES:7313
 Gen Z       :1594   Medium      :3339
 Millennials :9515   Small       :4903
```

Nivelurile sunt destul de echilibrate in cadrul distributiilor, exceptie facand urmatoarele:

- generatia Milleanials care reprezinta aproximativ 65% din totalul respondentilor;

- numarul redus al respondentilor ce poseda cunostinte in domeniul cloud.

Distributia raspunsurilor pentru variabila explicata este foarte echilibrata, constituind un puternic avantaj in calitatea modelului.

```
summarise(data_2022)
```

```
data frame with 0 columns and 1 row
```

```r
logistical_regression <- glm(ContainersExp ~ AgeGeneration + OrganisationLevel + CloudKnow
                             data = data_2022,
                             family="binomial"
)

summary(logistical_regression)
```

```
Call:
glm(formula = ContainersExp ~ AgeGeneration + OrganisationLevel +
    CloudKnowledge, family = "binomial", data = data_2022)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4972  -1.2428   0.8883   0.9706   1.9059

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.87775    0.14722 -12.755  < 2e-16 ***
AgeGenerationGen X        0.33012    0.09018   3.661 0.000251 ***
AgeGenerationGen Z        0.53379    0.09113   5.858  4.7e-09 ***
AgeGenerationMillennials  0.68578    0.07734   8.867  < 2e-16 ***
OrganisationLevelLarge    0.40688    0.12692   3.206 0.001347 **
OrganisationLevelMedium   0.45721    0.12881   3.550 0.000386 ***
OrganisationLevelSmall    0.23912    0.12705   1.882 0.059830 .
CloudKnowledgeYES         1.46102    0.04307  33.919  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18873  on 13662  degrees of freedom
Residual deviance: 17407  on 13655  degrees of freedom
AIC: 17423

Number of Fisher Scoring iterations: 4
```

Se poate observa ca toate valorile variabilelor explicative sunt foarte semnificative din punct de vedere statistic, mai putin nivelul ~ Medium ~ ale companiilor la care sunt angajati respondentii. Se extrag urmatoarele concluzii din output-ul regresiei modelate:

- Se poate observa o puternica corespondenta inversa intre varsta (generatia) respondentilor si mediul de lucru cu tehnologii bazate pe containere (cu cat sunt mai tineri, cu atat sunt mai predispusi sa lucreze cu containere):

a) Gen X este cu 30.63% mai probabila decat generatia Baby Boomers de a avea experienta cu tehnologiile bazate pe containere;

b) Gen Z este cu 52.03% mai probabila decat generatia Baby Boomers de a avea experienta cu tehnologiile bazate pe containere;

c) Milleanials este cu 67.21% mai probabila decat generatia Baby Boomers de a avea experienta cu tehnologiile bazate pe containere.

- Se poate observa o corespondenta intre nivelul organizatiei si mediul de lucru cu tehnologii bazate pe containere:

a) Respondentii angajati in companii de nivel mediu sunt cu 5.009% mai probabili decat angajatii companiilor large sa lucreze cu tehnologii bazate pe containere (nesemnificativ stat.);

b) Respondentii angajati in companii de nivel mic sunt cu 16.82% mai putin probabili decat angajatii companiilor large sa lucreze cu tehnologii bazate pe containere;

- Se poate observa ca posesia de cunostinte in domeniul cloud mareste semnificativ sansa ca angajatii sa lucreze cu tehnologii bazate pe containere:

a) Respondentii care poseda cunostinte in domeniul cloud sunt de 1.46 ori mai probabili decat cei care nu poseda cunostinte in domeniul cloud sa lucreze cu tehnologii bazate pe containere.

Calculam acuratetea modelului sugerat.

```
fitted.results <- predict(logistical_regression, newdata = subset(data_2022, select =
                                                                  c(OrganisationLevel, A
fitted.results <- ifelse(fitted.results >= 0.5, "YES", "NO")
misClasificError <- mean(fitted.results != data_2022$ContainersExp)
print(paste('Accuracy', 1 - misClasificError))
```

```
[1] "Accuracy 0.652931274244309"
```

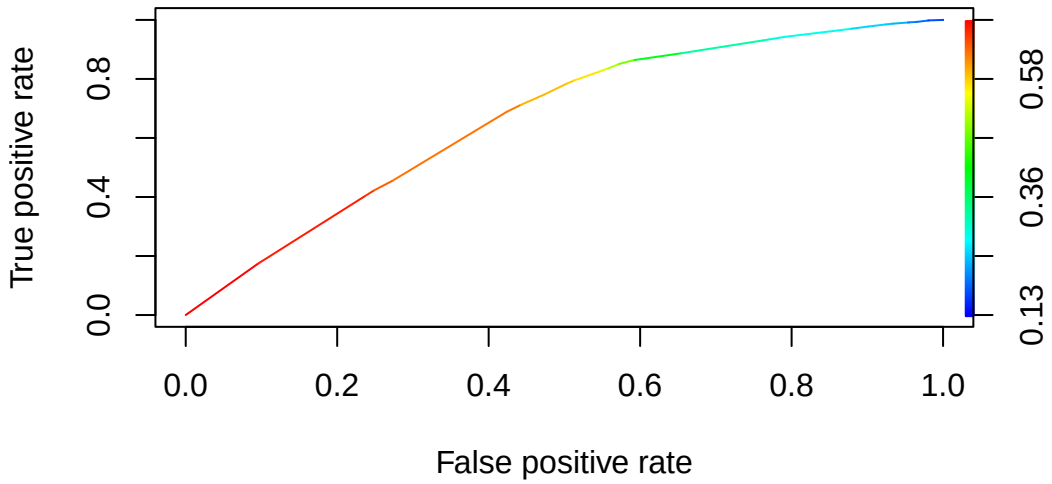
Astfel, reiese o acuratete de 65.54%, o acuratete de nivel satisfacator pentru un model valid.

```
install.packages("ROCR", repos = "http://cran.us.r-project.org")
```

```
Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
```

```
package 'ROCR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\Admin\AppData\Local\Temp\RtmpwLOOGG\downloaded_packages
```

```
library(ROCR)
yhat_regression <- predict(logistical_regression, type="response")
prediction_regression <- prediction( yhat_regression, data_2022$ContainersExp)

performance_regression <- performance(prediction_regression, "tpr", "fpr")
plot(performance_regression, colorize = TRUE)
```




```
install.packages("ROCR", repos = "http://cran.us.r-project.org")
```

```
Warning: package 'ROCR' is in use and will not be installed
```

```
library(ROCR)
 auc_regression <- performance(prediction_regression, "auc")
print(paste('AOC', auc_regression@y.values))
```

```
[1] "AOC 0.665889899445599"
```

Aria de sub curba ROC ocupa ~ 66.44% din suprafata totala, fapt ce indica ca am ales un clasificator mai bun decat unul aleator.

O valoare a pragului probabilitatii de 0.5 ar conduce la o rata a clasificarii pozitive corecte de aproximativ 0.9 si la o rata a clasificarii pozitive false de aproximativ 0.55, fapt ce justifica alegerea acestui punct optim de probabilitate de pe curba ROC.

```
install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
```

```
package 'caret' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
    C:\Users\Admin\AppData\Local\Temp\RtmpwLOOGG\downloaded_packages
```

```
library(caret)
```

```
Loading required package: lattice
```

```
Attaching package: 'caret'
```

```
The following object is masked from 'package:purrr':

    lift
```

```
confusionMatrix(factor(fitted.results), factor(data_2022$ContainersExp))
```

```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  2806 1198
       YES 3544 6115
```

```
               Accuracy : 0.6529
                 95% CI : (0.6449, 0.6609)
    No Information Rate : 0.5352
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.285

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4419
            Specificity : 0.8362
         Pos Pred Value : 0.7008
         Neg Pred Value : 0.6331
             Prevalence : 0.4648
         Detection Rate : 0.2054
   Detection Prevalence : 0.2931
      Balanced Accuracy : 0.6390

       'Positive' Class : NO
```

Senzitivitatea de 44.19% indica o acuratete modesta in predictia corecta a numarului de respondenti care au experienta cu tehnologiile bazate pe containere.

Specificitatea de 85.18% indica o acuratete mult mai ridicata in ceea ce priveste predictia corecta a numarului de respondenti care nu au experienta cu tehnologiile bazate pe containere.

Din matricea de confuzie se pot extrage urmatoarele date:

- true positive - 2628;

- false positive - 1065;

- true negative - 6119;

- false negative - 3552.

In concluzie, modelul are performante relativ bune, insa nu optime. Acest fapt se datoreaza altor variabile explicative neincluse in model care pot influenta daca un angajat din America de Nord are sau nu are experienta cu tehnologiile bazate pe containere (docker / kubernetes). Variabilele explicative care ar fi putut face parte din model sunt experienta utilizatorilor (anii de cand codeaza/ varsta la care au scris primul cod), dar si tara unde locuiesc.