Report of MA678 Midterm Project

Yongrong Chai

11/30/2021

Abstract

The dataset I found was from Kaggle, and it was about HR Analytics: Job Change of Data Scientists. The dataset was collected from companies with information about candidates of data scientists candidates. In this project, I will answer two questions: 1. What affected factors make an employee leave his/her current job? 2. What's the probability that a candidate will work for the company? To address there questions, I used the factors related with my outcome variable, "target" with 0(Not looking for a job change) and 1(Looking for a job change), and built a logistic multilevel model. The model showed that the variables (————————————————–). There are 5 main parts in this report: Introduction, Method, Result and Discussion.

Introduction

Companies that specializes in Big Data and Data Science are looking to hire data scientists, and they also provide some courses for their candidates. Although candidates have signed up for the training, some of them still want to leave for another new job. The company needs to know which of these candidates genuinely wants to work for the company after training or looking for a new job because it helps to cut costs and time while also improving the quality of training or course preparation and categorization. Candidates' demographics, education, and experience are all in the hands of those who sign up and enroll.

Method

Data Cleaning and Processing

Variable Description:

Variables	**Description**
enrollee_id	Unique ID for candidate
city	City code
city_development_index	Development index of the city (scaled)
gender	Gender of candidate
relevent_experience	Relevant experience of candidate

enrolled_university	Type of University course enrolled if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate total experience in years
company_size	No of employees in current employer's company
company_type	Type of current employer
last_new_job	Difference in years between previous job and current job
training_hours	Training hours completed
target	0–Not looking for job change, 1–Looking for a job change

Let's take a look at the data's head and tail:

```
# Viewing the datasets using the "kableExtra" package
HR[c(1:4,19155: 19158), ]%>%
  kbl(caption = '<b>Job Change of Data Scientists</b>') %>%
  kable_classic(full_width = F, html_font = "times") %>%
  kable_styling(bootstrap_options = c("striped",
                                      "hover")) %>%
  scroll_box(width = "100%")
```

Job Change of Data Scientists

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	edu
1	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Gra
2	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Gra
3	11561	city_21	0.624		No relevent experience	Full time course	Gra
4	33241	city_115	0.789		No relevent experience		Gra
19155	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Gra

19156	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Gra
19157	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	Hig
19158	23834	city_67	0.855		No relevent experience	no_enrollment	Pri

Missing data:

```
HR[HR == ""] = NA
colSums(is.na(HR))
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##                 gender    relevent_experience    enrolled_university
##                   4508                      0                    386
##         education_level       major_discipline             experience
##                    460                   2813                     65
##            company_size           company_type           last_new_job
##                   5938                   6140                    423
##          training_hours                 target
##                      0                      0
```

```
#colSums(is.na(HR)) / nrow(HR)
```

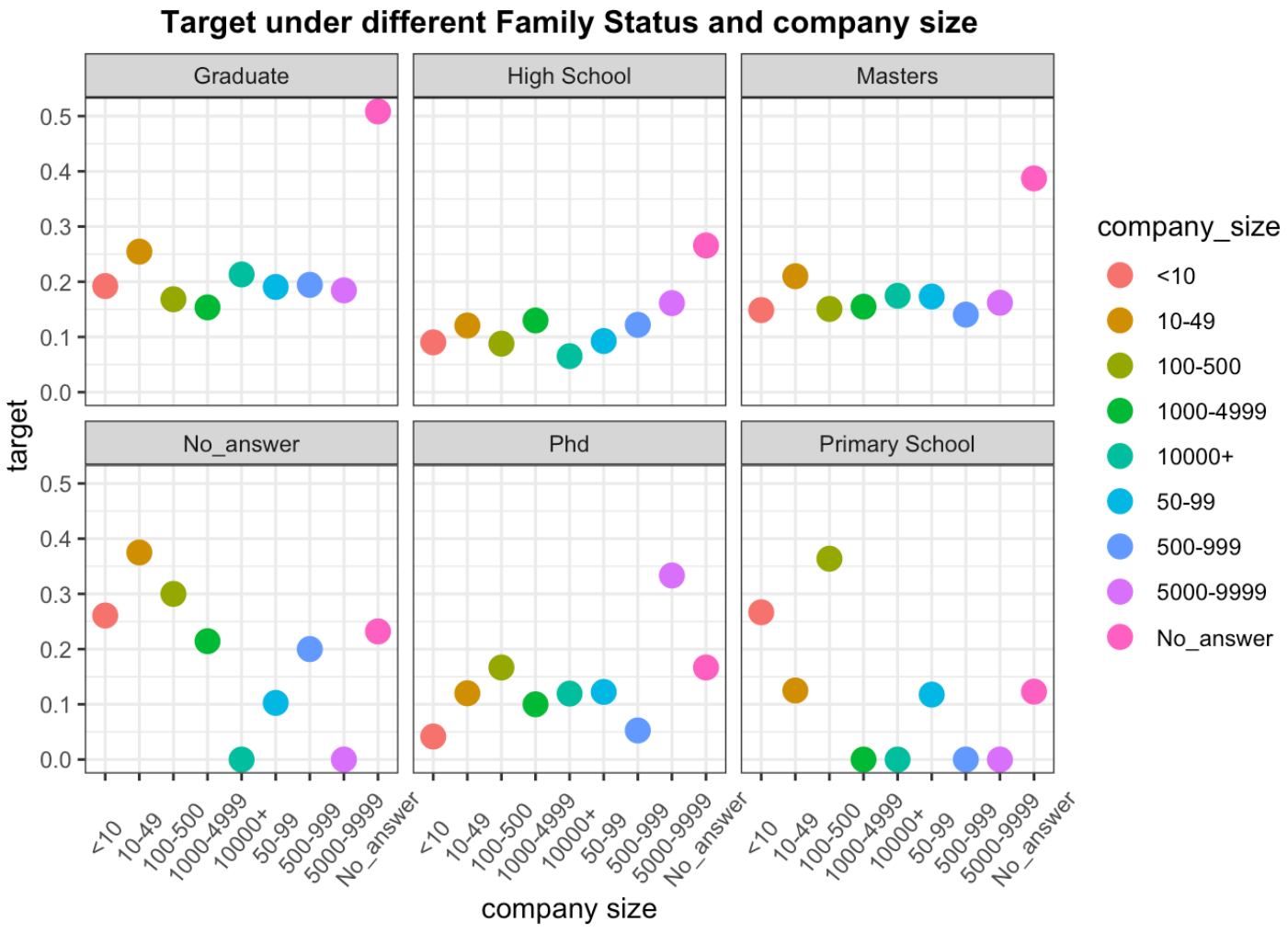
Although the multilevel model I would use was not influenced by missing data. However, it is important to appreciate why they are missing when faced with missing data. In this case, candidates didn't provide their answers when they registered and enrolled, so the missing data was missing at random, deleting the instances with missing data does not lead to biased inference.

Data Selection: From a corporate perspective, gender discrimination is something to avoid when hiring and making decisions, so I don't include `gender` as one of my predictors. I chose numerical variables: `training_hours` , `city_development_index` , and `experience` ; categorical variables: `education_level` and `company_size` .

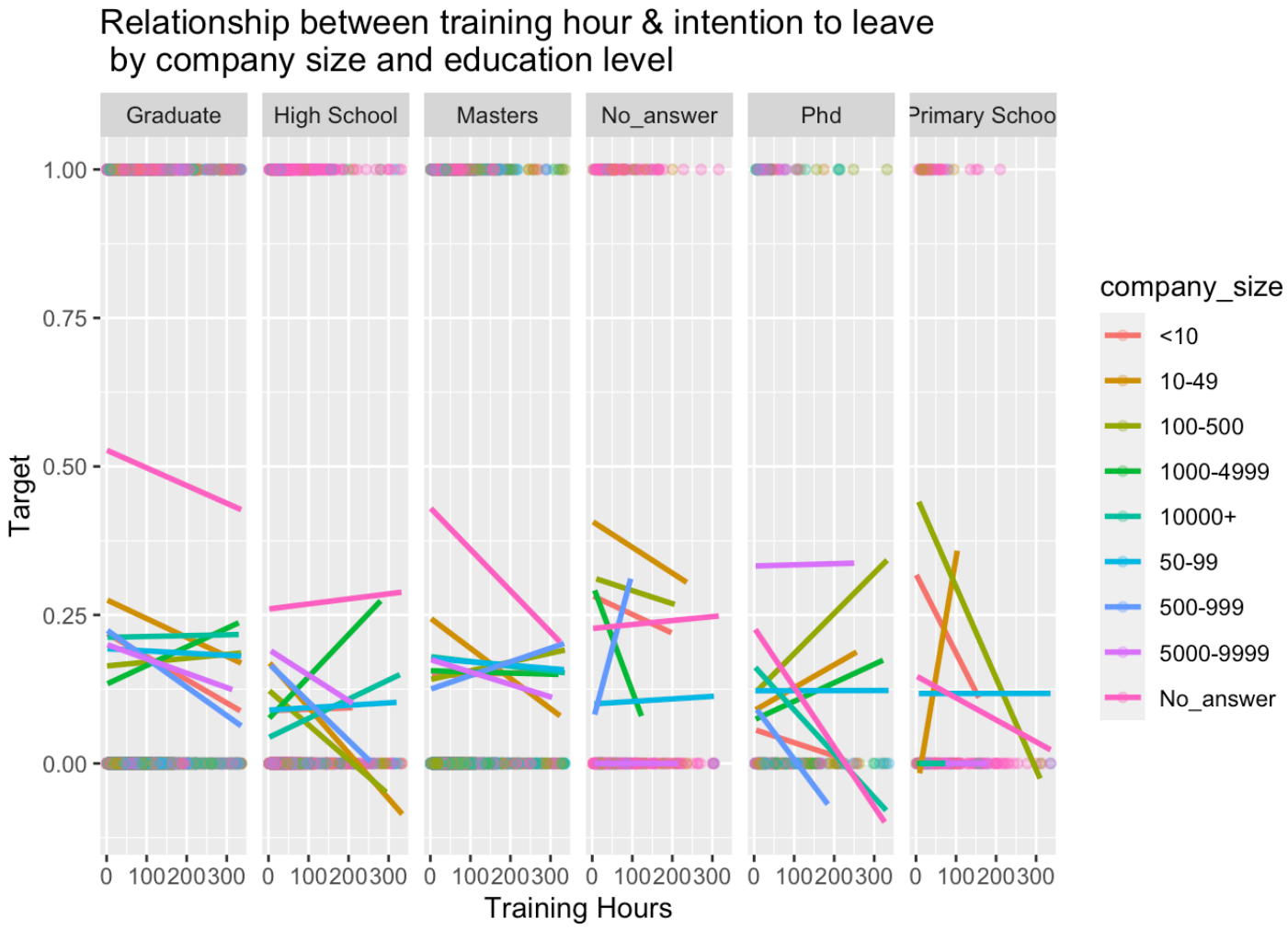
```
HR <- read.csv("aug_train.csv", header = T)
HR$experience<-gsub(c(">","<"),"",as.character(HR$experience))
HR$experience<-gsub("<1","1",as.character(HR$experience))
HR$company_size<-gsub("/","-",as.character(HR$company_size))
HR$experience <- as.numeric(HR$experience)
#str(HR)

HR$education_level[HR$education_level == ""] <- "No_answer"
HR$company_size[HR$company_size == ""] <- "No_answer"
```

Explanatory Data Analysis

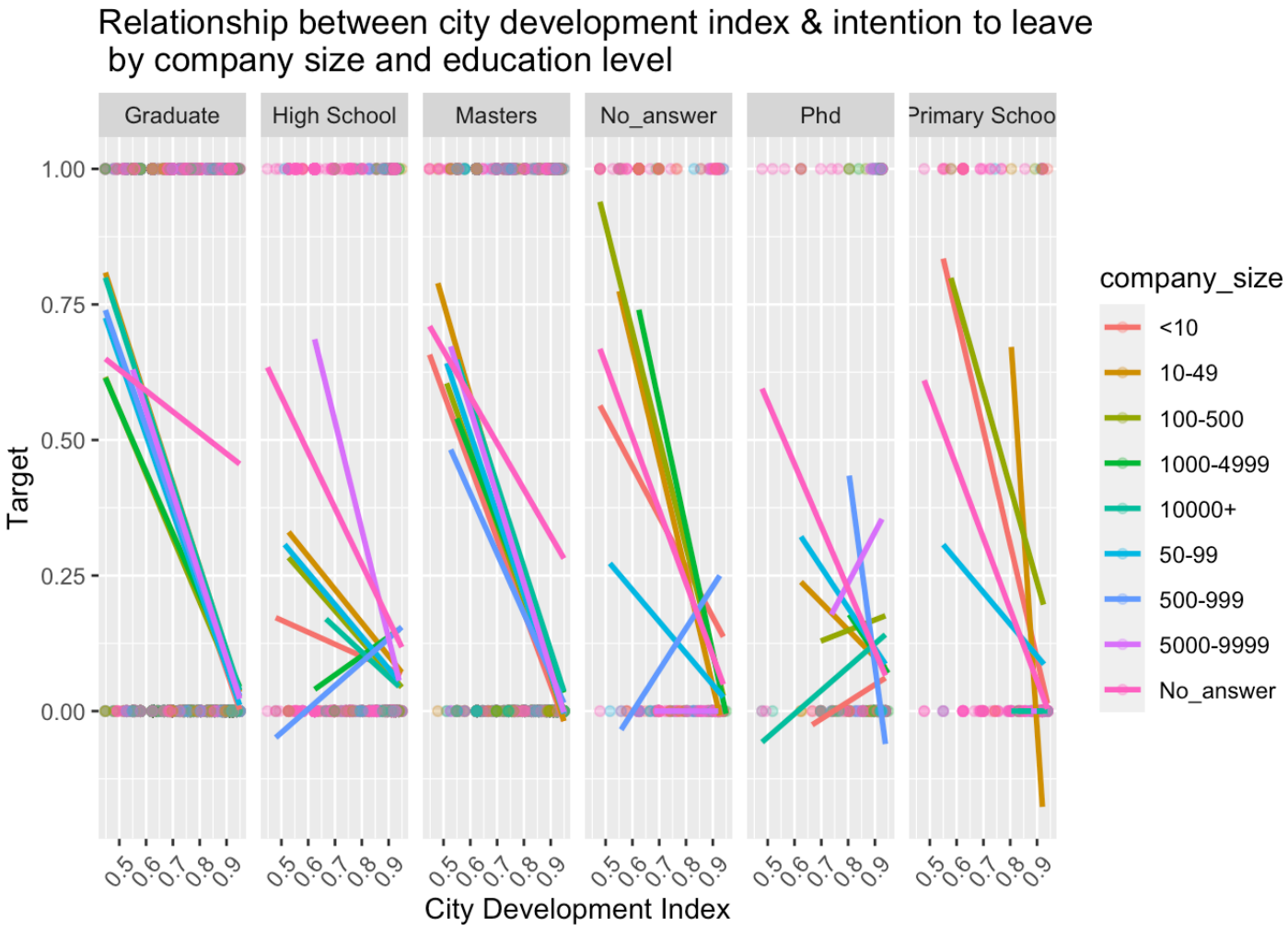
```
source(file = "Final_EDA.R")
p_2v
```




```
line1
```


Relationship between training hour & intention to leave
by company size and education level
Graduate
High School
Masters
No_answer
Phd
Primary Schoo
Target
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Training Hours
company_size
<10
10-49
100-500
1000-4999
10000+
50-99
500-999
5000-9999
No_answer

```
line2
```



Model Fitting

The model I used is the multilevel logistic model with varying slope and varying intercept. Categorical predictors I used are `company_size` and `education_level` because they have higher ICC which measures reliability of ratings or measurements for clusters. Here is the model: glmer(formula = target ~ city_development_index + training_hours + experience + (1|education_level) + (1|company_size), family = binomial(link = “logit”), data = HR_data) summary(model2)

```
HR$training_hours = as.numeric(HR$training_hours)
model1 <- glmer(formula = target ~ city_development_index + training_hours + experien
ce + (1|company_size), family = binomial(link = "logit"), data = HR)

model2 <- glmer(formula = target ~ city_development_index + training_hours + experien
ce + (1|education_level), family = binomial(link = "logit"), data = HR)

model3 <- glmer(formula = target ~ city_development_index + training_hours + experien
ce + (1|education_level) + (1|company_size), family = binomial(link = "logit"), data
= HR)
summary(model2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: target ~ city_development_index + training_hours + experience +
##     (1 | education_level)
##    Data: HR
##
##      AIC      BIC   logLik deviance df.resid
##  19120.5  19159.8  -9555.3  19110.5    19088
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8304 -0.5077 -0.4027 -0.2709  3.9481
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  education_level (Intercept) 0.07941  0.2818
## Number of obs: 19093, groups:  education_level, 6
##
## Fixed effects:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.3715884  0.1652870  20.398  < 2e-16 ***
## city_development_index -5.3883308  0.1425646 -37.796  < 2e-16 ***
## training_hours         -0.0009421  0.0003046  -3.093  0.00198 **
## experience             -0.0355314  0.0032118 -11.063  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cty_d_ trnng_
## cty_dvlpmn_ -0.639
## trainng_hrs -0.127  0.014
## experience   0.061 -0.318 -0.002
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

```
anova(model1, model2, model3)
```

```
## Data: HR
## Models:
## model1: target ~ city_development_index + training_hours + experience + (1 | compa
ny_size)
## model2: target ~ city_development_index + training_hours + experience + (1 | educa
tion_level)
## model3: target ~ city_development_index + training_hours + experience + (1 | educa
tion_level) + (1 | company_size)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1    5 18334 18374 -9162.1    18324
## model2    5 19120 19160 -9555.3    19110    0.0  0
## model3    6 17965 18012 -8976.3    17953 1157.9  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#summ(model3)
```

Result

Model Coefficients

fixed effects:

	Est. S	.E. z	val.	p
(Intercept)	2.91	0.32	9.18	0.00
city_development_index	-5.60	0.15	-37.45	0.00
training_hours	-0.00	0.00	-2.68	0.01
experience	-0.03	0.00	-9.24	0.00

random effects:

Group P	arameter St	d. Dev.
company_size	(Intercept)	0.46
education_level	(Intercept)	0.61