Report of MA678 Midterm Project

Yuchen Liu

2021/12/8

Abstract

Nowadays there are many online courses website. They provide the professional training and bootcamp for people to learn. Here is a company provide the Big Data and Data Science training course for people who want to be data scientist in the future. Then, the company decided to take a survey of the candidates to see what kind of candidates want to change the job and work for this company after the training. Here are few questions about that: Does the gender, education level or the company size have relationship with changing the job? In order to solve this problem, I use models to analysis this dataset and get the results. From the analysis, we will know that PhD education level have positive impact with the target. This report have 5 main parts: Abstract, Introduction, Method, Result and Discussion.

Introduction

The company take the survey of candidates' city, gender, experience, university, education level, major discipline, company size, current company type, difference in years between previous job and current job and training hours. Finally, the company use 0 for not looking for job change and 1 for looking for a job change to count the survey. In this project, I will use logistic regression and multilevel regression to see the relationship between variables and final result. First of all, I need to clean up the data, to remove all the missing value, and make the data easier to make the model

Method

Data Cleaning and Processing

The main data set is published on Kaggle: HR analytics: Job change of data scientists.

At the first, for some missing value in this table, I need to clean them. I remove all of the missing value by "na.omit" code. The following figure 1 is the first page of the table look like.

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_unive
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment
2	666	city_162	0.767	Male	Has relevent experience	no_enrollment
3	402	city_46	0.762	Male	Has relevent experience	no_enrollment
4	27107	city_103	0.920	Male	Has relevent experience	no_enrollment
5	23853	city_103	0.920	Male	Has relevent experience	no_enrollment
6	25619	city_61	0.913	Male	Has relevent experience	no_enrollment
7	6588	city_114	0.926	Male	Has relevent experience	no_enrollment
8	31972	city_114	0.843	Male	Has relevent experience	no_enrollment
9	19061	city_114	0.926	Male	Has relevent experience	no_enrollment
10	7041	city_40	0.776	Male	Has relevent experience	no_enrollment
11	10408	city_21	0.624	Male	Has relevent experience	no_enrollment
12	14928	city_103	0.920	Male	Has relevent experience	no_enrollment

Figure 1

Then, I got the cleaned data with 8955 observations and 14 variables.

Exploratory Data Analysis

In order to make easy to read the data, I make some plot to see the relationship of each variables. And I also make some scatter plot to see the correlation between each variables. The relationship between variables and targets are important.

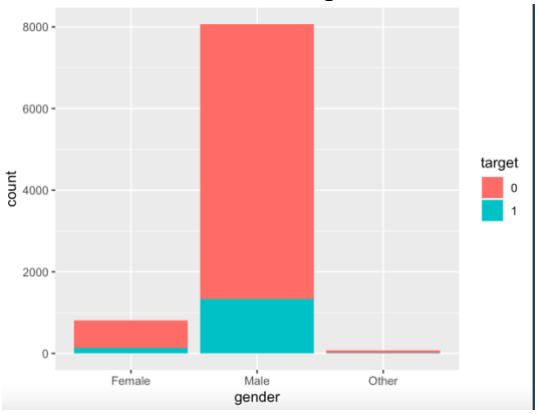


Figure 2

For this figure 2, I take gender verses target, which is the result of their survey. We can see from this plot, the male is the most in our survey. Most of people, whether male or female, the survey results are 0, which means most of people not looking for the job change.

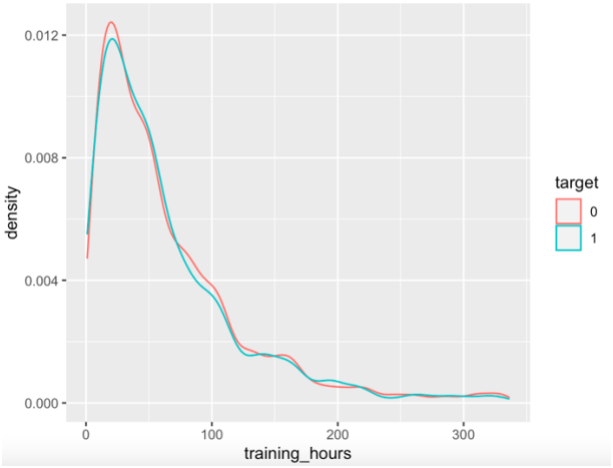


Figure 3

From figure 3, we can know that the training hours distribution are similar whether they are change or not change the work. The people who not want to change the work are a little bit more than people who wants in the lower training hours. From the model, we will know that the p-value of training hours is larger than 0.05, which means training hours doesn't affect the outcome, which is the target.

Results

Here I make a logistic regression model with most of useful variables at the first.

```
fit<-stan_glm(target~gender+city_development_index+training_hours+education_level
              +major_discipline+experience,family=binomial(link = 'logit'),
              data=train1)
```

Figure 4

From figure 4, I fit a logistic regression model for most of the useful variables. From the figure5, we can see that gender male, training hours, masters education level, other major discipline and 1year experience, these variables' confidence level are passed by 0 in the plot. That means they are not significant difference from 0. These variables are not affect the outcome.

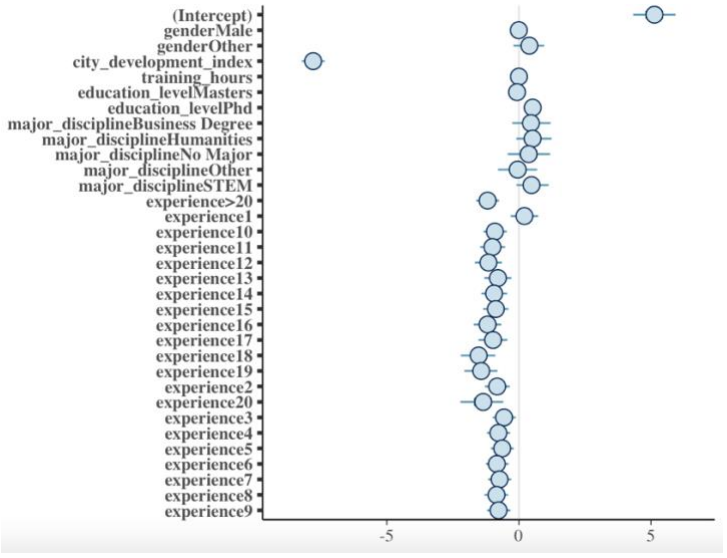


Figure 5

Then I make a multilevel regression model.

```
fit2 <- glmer(target~gender+city_development_index+training_hours+education_level
              +major_discipline+experience+(1|city),
              family=binomial(link = 'logit'),data=train1)
```

Figure 6

From the figure 6 and the full results, we can see that I made a multilevel regression model between target and most of variables. From the results, we will know that there

are some variables' p-value are less than 0.05, for example, the city development index, PhD education level, more than 20 years experiences, etc.. That means these variables have impacts on the target. Then, we can see that the PhD education level's estimate is larger than 0, which means this variable has positive impact on the outcome.

Discussion

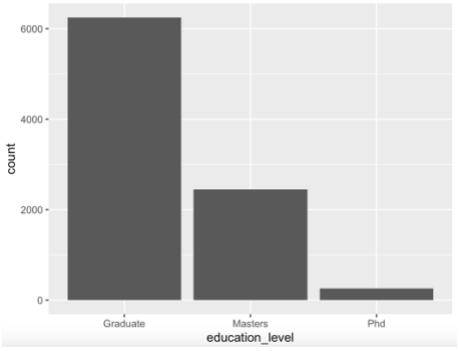
From the model, we can see that most of variables have some of impacts to our outcome. It is reasonable that the city, experiences and other variables can affect the person's willingness of changing the job. However, there are also some crucial variables that the data set doesn't include, for example, the people's satisfaction to this training and the reasons why they want to participate this training. If the data set include these data, the analysis will be more accurate.

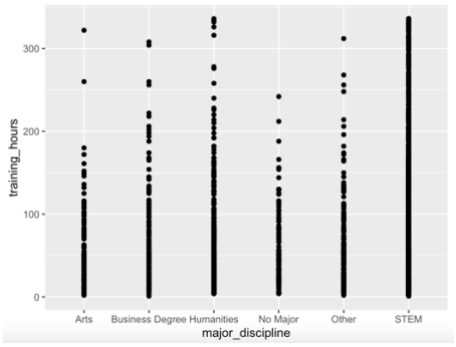
Citation

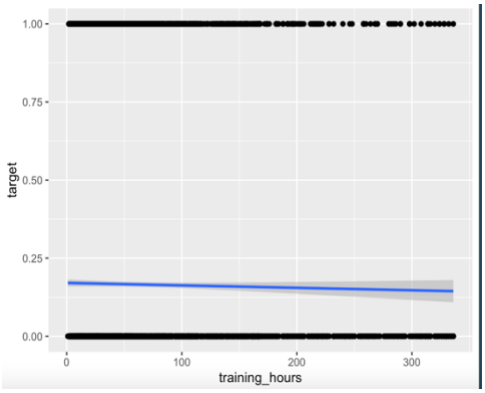
Kaggle, HR analytics
https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists/version/1?select=aug_train.csv

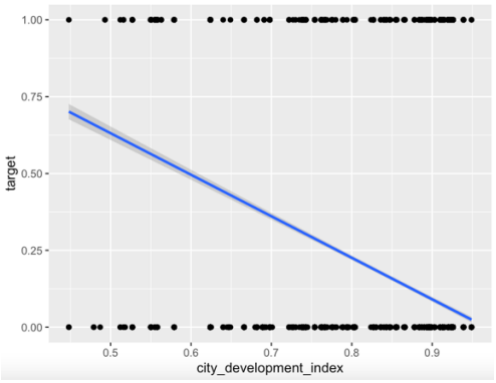
Appendix

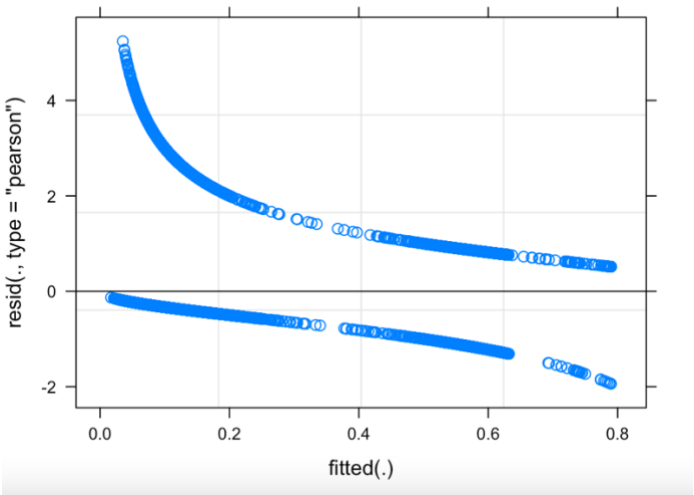
More EDA











Full Results

Results for linear regression model:
Estimates:

	mean	sd	10%	50%	90%
(Intercept)	5.1	0.5	4.5	5.1	5.8

genderMale	0.0	0.1	-0.1	0.0	0.2
genderOther	0.4	0.4	-0.1	0.4	0.8
city_development_index	-7.8	0.3	-8.1	-7.8	-7.4
training_hours	0.0	0.0	0.0	0.0	0.0
education_levelMasters	-0.1	0.1	-0.2	-0.1	0.0
education_levelPhd	0.5	0.2	0.3	0.5	0.8
major_disciplineBusiness Degree	0.5	0.4	-0.1	0.5	1.0
major_disciplineHumanities	0.5	0.4	0.0	0.5	1.1
major_disciplineNo Major	0.4	0.5	-0.2	0.4	1.0
major_disciplineOther	0.0	0.5	-0.6	0.0	0.6
major_disciplineSTEM	0.5	0.4	0.0	0.5	1.0
experience>20	-1.2	0.3	-1.5	-1.2	-0.8
experience1	0.2	0.3	-0.2	0.2	0.6
experience10	-0.9	0.3	-1.3	-0.9	-0.6
experience11	-1.0	0.3	-1.4	-1.0	-0.6
experience12	-1.2	0.3	-1.5	-1.2	-0.8
experience13	-0.8	0.3	-1.2	-0.8	-0.4
experience14	-0.9	0.3	-1.3	-0.9	-0.6
experience15	-0.9	0.3	-1.2	-0.9	-0.5
experience16	-1.2	0.3	-1.6	-1.2	-0.8
experience17	-1.0	0.3	-1.4	-1.0	-0.6
experience18	-1.5	0.4	-2.1	-1.5	-1.0
experience19	-1.4	0.4	-1.9	-1.4	-0.9
experience2	-0.8	0.3	-1.2	-0.8	-0.5
experience20	-1.4	0.5	-2.0	-1.4	-0.8
experience3	-0.6	0.3	-0.9	-0.6	-0.2
experience4	-0.8	0.3	-1.1	-0.8	-0.4
experience5	-0.6	0.3	-1.0	-0.6	-0.3
experience6	-0.8	0.3	-1.2	-0.8	-0.5
experience7	-0.7	0.3	-1.1	-0.7	-0.4
experience8	-0.9	0.3	-1.2	-0.8	-0.5
experience9	-0.8	0.3	-1.1	-0.8	-0.4

Fit Diagnostics:

	mean	sd	10%	50%	90%
mean_PPD	0.2	0.0	0.2	0.2	0.2

Results for multilevel regression model:

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.9757762	0.6906387	4.309	1.64e-05	***

genderMale	0.0194424	0.1153558	0.169	0.866156
genderOther	0.5311643	0.3487157	1.523	0.127708
city_development_index	-5.9005857	0.6766376	-8.720	< 2e-16 ***
training_hours	-0.0009488	0.0005540	-1.713	0.086746 .
education_levelMasters	-0.0500401	0.0787412	-0.636	0.525102
education_levelPhd	0.5534948	0.1984202	2.790	0.005279 **
major_disciplineBusiness Degree	0.4796448	0.4372351	1.097	0.272644
major_disciplineHumanities	0.5453370	0.3963851	1.376	0.168891
major_disciplineNo Major	0.5469976	0.4726681	1.157	0.247168
major_disciplineOther	-0.0431636	0.4511120	-0.096	0.923773
major_disciplineSTEM	0.5229727	0.3620533	1.444	0.148609
experience>20	-1.0314993	0.2657846	-3.881	0.000104 ***
experience1	0.2687189	0.3292296	0.816	0.414383
experience10	-0.7290948	0.2795064	-2.609	0.009094 **
experience11	-0.8077171	0.2978544	-2.712	0.006692 **
experience12	-0.9218195	0.3131476	-2.944	0.003243 **
experience13	-0.5783717	0.3167571	-1.826	0.067863 .
experience14	-0.6846838	0.3045684	-2.248	0.024573 *
experience15	-0.7207175	0.2970904	-2.426	0.015270 *
experience16	-0.9902543	0.3222356	-3.073	0.002119 **
experience17	-0.8108110	0.3357745	-2.415	0.015746 *
experience18	-1.2195592	0.3960258	-3.079	0.002074 **
experience19	-1.2473479	0.3836888	-3.251	0.001150 **
experience2	-0.7695989	0.2899624	-2.654	0.007951 **
experience20	-1.2042578	0.4826832	-2.495	0.012598 *
experience3	-0.5160176	0.2751791	-1.875	0.060764 .
experience4	-0.6900975	0.2754030	-2.506	0.012218 *
experience5	-0.5180797	0.2705439	-1.915	0.055498 .
experience6	-0.7096333	0.2741842	-2.588	0.009649 **
experience7	-0.6135086	0.2761808	-2.221	0.026324 *
experience8	-0.7701650	0.2859377	-2.693	0.007071 **
experience9	-0.6019134	0.2781397	-2.164	0.030459 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1