Report

XihaoCao

2021/12/9

Abstract

In order to find the factors that lead a candidate to look for a new job, I firstly conduct the Exploratory Data Analysis by a series of graphs. I notice that the data set is imbalanced in some variables, ex: gender, and I will not include them in the Model. I also notice that the influence of the education level a candidate have seems to vary among the cities, so, I will consider it as a random effect and use the city to construct groups. Since there are over 120 cities, in order to build the model more efficiently, I will only include the top 15 cities with where most candidates are from. Then I build the multilevel model and I find that working in the public sector, being current university enrolled and higher city development index all lead to a higher possibility for a candidate to look for a new job, while I have uncertainties for the other variables.

Introduction

A Big Data company has set up a series of data scientist training courses, and it wants to hire candidates who have successfully passed those courses. However, not all people want to quit their current jobs after the training, and the company wants to know which of these candidates are looking for a job change. Thus I will try to understand the factors, ex: their education level, previous experience, training hours, that lead candidates to quit their jobs. If we can build a efficient model to predict the probability that a candidate looks for a job change, we can help the company to reduce the cost and time planning the courses and categorization of candidates, and allow the employee to make better decisions.

Data Resource

I get this data set from the Kaggle database, and it contains 12 columns and 19158 observations. The columns are enrolled_id, city, city_development_index, gender, relevant_experience, enrolled_university, education_level, major_discipline, experience, company_size, company_type, last_new_job, training_hours, target. And the outcome is the target column which is a binary value indicating whether each candidate is looking for a job change.

Missing data and Data Organizations

The missing values in the data set are assigned a blank value rather than NA, in order to visualize and process easier, I replace all the missing value by a 'NA' string. And since all columns in the data set are stored as in the character type, I factorize all the columns that are not continuous to prepare for the model establishment.

EDA

Before I build the model, I conduct the exploratory data analysis to have a better understanding of the data set. In Figure 1, I explore the amount of missing values of each variable. As we can see, company size and company type have around 6000 missing values, which is almost one third of the total data set. Although these two variables have many variables, since we have a large data set that contains about 20000 pieces of observations, thus I will still include them in the model.

In Figure 2: I plot the number of candidates in each gender group, as we can see that almost 70% candidate are male, and we also have over 4500 missing values. Since the data set is imbalanced within the gender variable and the gender is commonly prohibited to be a admission factor, I will not include gender as a predictor in the model.

In Figure 3, I explore the size of each education level, and we can notice that 61% of the candidate hold a graduate degree.

In Figure 4, I explore how many candidates are there in each major_disciplines, and we can see that over 80% candidates come from the STEM majors and there are about 15% missing values. Thus the major_discipline column is also imbalanced, and I will not include it into the model.

In Figure 5, I check the size of each university enrollment groups, as we can see that over 70% candidate are not enrolled in a university.

In figure 6, I check the size of each experience group, and we can notice that the experience of candidates has a relative large range from 0 to 20.

In figure 7, as we can see that distributions of training hours of candidates both take the offer and not share almost the same shape. Thus we can conclude that the training hours has little impact on candidate' final decisions, and I will not include the training hours in my model.

In the last figure, I check the when were the last job of each candidate, as we can see the distribution is not imbalanced, and most candidates have quited last job within a year.

Figure1: number of missing value in each column

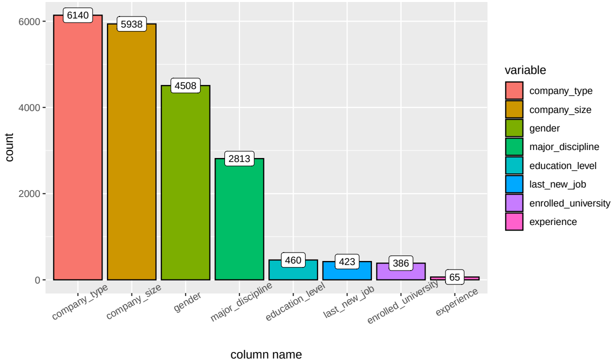


Figure2: proportion of each gender to the whole

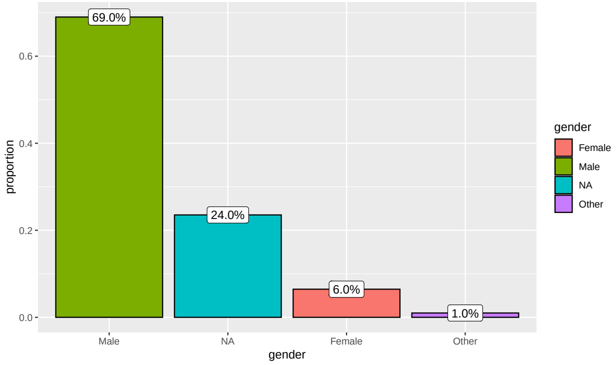


Figure3: proportion of each education level to the whole

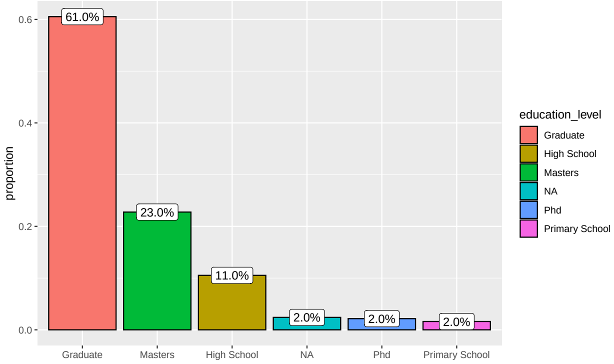


Figure4: proportion of each major discipline to the whole

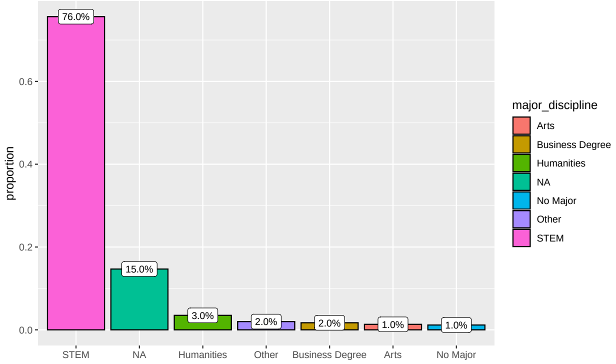


Figure5: proportion of each enrolled university condition to the whole

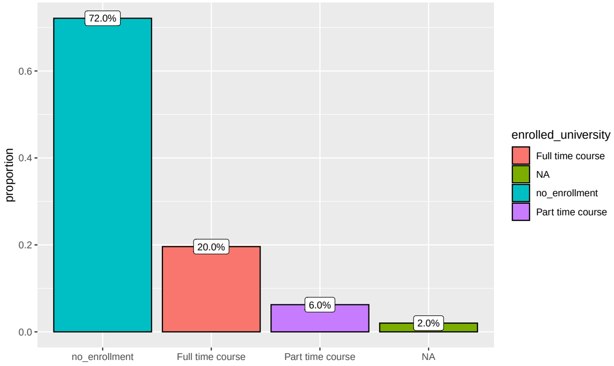


Figure6: proportion of each experience level to the whole

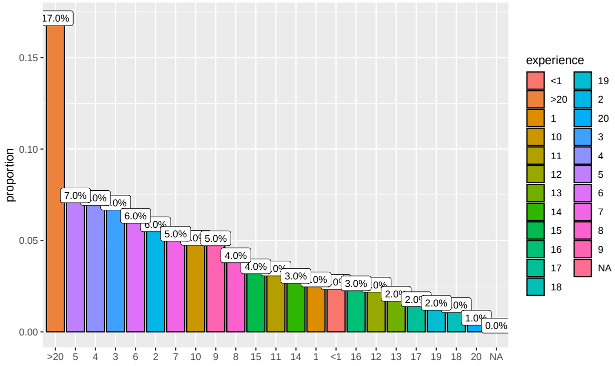


Figure7: Relationship between training hours and candidate's final decision

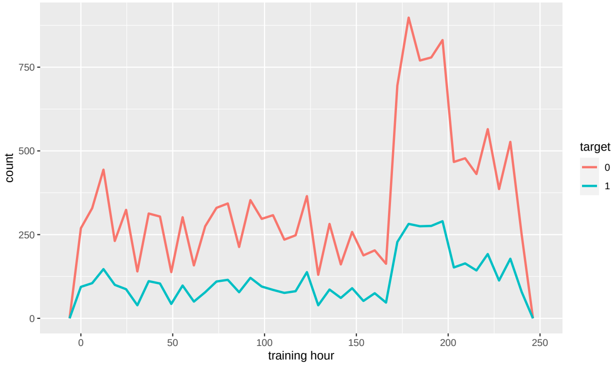
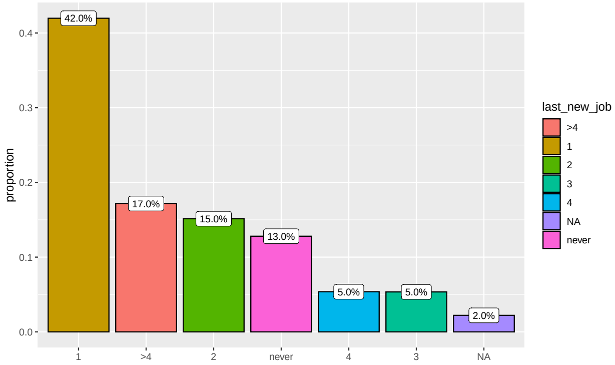


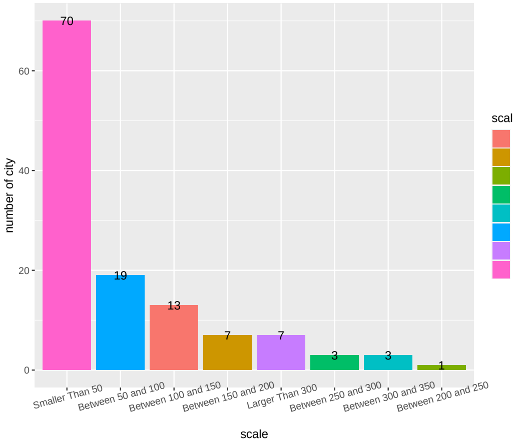
Figure8: proportion of each last_new_job level to the whole



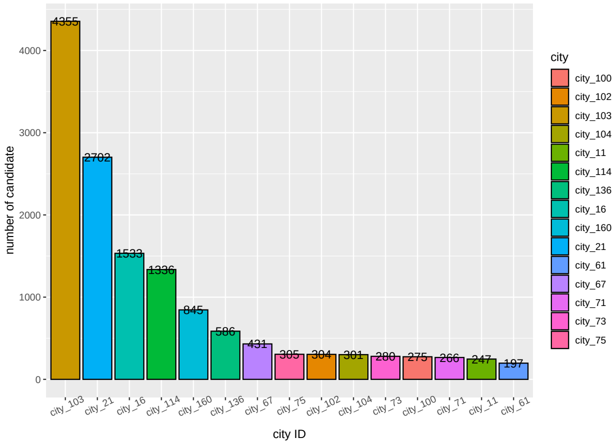
Use City to construct Groups

From the two figures, we notice that there are 70 cities who have a sample size smaller than 50, but at least one fourth of the total candidates are from the city 103 and the top 10 cities contain over 13500 candidate, which are over 70% of the total observations. So, it is not wise to build group for each city and I will only keep the top 15 cities to build 15 groups. The selected cities with ID are listed below.

number of city in each scale



cities where most candidates are from



```
## [1] "The selected city is city_103" "The selected city is city_21"
## [3] "The selected city is city_16"  "The selected city is city_114"
## [5] "The selected city is city_160" "The selected city is city_136"
## [7] "The selected city is city_67"  "The selected city is city_75"
##  [9] "The selected city is city_102" "The selected city is city_104"
## [11] "The selected city is city_73"  "The selected city is city_100"
## [13] "The selected city is city_71"  "The selected city is city_11"
## [15] "The selected city is city_61"
```

Model establishment

Since We use the top 15 cities most candidates live in to build groups, we have total 15 groups. Then I construct the initial Multilevel model including every variable into the model. Then we check the inital model coefficients by summary, I notice that only city_development_index, enrolled_university, education_level, last-new_job, and company type are significant at the 0.05 level. Thus I will only keep these variables and build my final multilevel model. And, according to our observations in the EDA section, we set the education level as a random effect and leave all others as fixed effects.

Our final model has the following structure:

```r
fit0 <- glmer(target ~ city_development_index + enrolled_university + last_new_job +
                company_type + (1 + education_level|city),
              family = binomial(link = 'logit'), data = model_train)
```

Notice that enrolled-university, last_new_job, and company_type are all factor variables, so we have coefficients for all the factor categories except for their baselines. And we can see that city_development_index, full time university enrollment, public sector conpany tyep and last_new_job in 3, 4, never are all significant at the 0.05 level.

	Estimate	Std. Error	Z value	Pr(>
(Intercept)	5.65792	0.90972	6.219	4.99e-10 ***
city_development_index	-9.06148	1.01555	-8.923	2e-16 ***
enrolled_universityno_enrollment	-0.37976	0.14049	-2.703	0.00687 **
enrolled_universityPart time course	-0.31751	0.24624	-1.289	0.19725
last_new_job1	0.09870	0.13085	0.754	0.45066
last_new_job2	0.15041	0.15213	0.989	0.32281
last_new_job3	0.41354	0.19582	2.112	0.03470 *
last_new_job4	0.43491	0.19367	2.246	0.02473 *
last_new_jobnever	0.50621	0.22246	2.275	0.02288 *
company_typeFunded Startup	-0.06352	0.26955	-0.236	0.81371
company_typeNGO	0.01462	0.31691	0.046	0.96321
company_typeOther	0.75507	0.49387	1.529	0.12629
company_typePublic Sector	0.81859	0.27545	2.972	0.00296 **
company_typePvt Ltd	0.16060	0.21966	0.731	0.46468

And for the random effect:

	Variance	Std.Dev.	Corr	Pr(>
(Intercept)	0.101853	0.31914		
education_levelMasters	0.003229	0.05683	0.39	0.97
education_levelPhd	0.846315	0.91995	0.15	0.97

Result

By the model coefficients tables above, we can have the following interpretations of the variables in the model.

- The lower the development index of the city a candidate live in, the less likely for this candidate to look for a new job.
- A candidate who is currently enrolled in a university is less likely to look for a new job compared with another candidate who has the same conditions, but not enrolled in a university.
- For Candidates who have changed jobs within 2 and 4 years or never are more likely to look for a new job compared with candidates who have changed jobs further than 4 years from now.
- Candidates who are currently working in a public sector are more likely to look for a new job than candidates from the Early-Stage companies.

Thus, we can conclude that working in the public sector, current university enrollment, having 2-4 years difference between previous job and current job, and higher city development index all lead to a higher possibility to look for a new job. Meanwhile, the PHD education level as a random effect has relative large variance across the 15 cities, which means the ability of the PhD education degree to influence a candidate's decision in each city is noticeably different. But, the variance of the Master degree is relatively small, which indicates that the Master degree has relatively the same ability to influence a candidate in each city.

Discussion

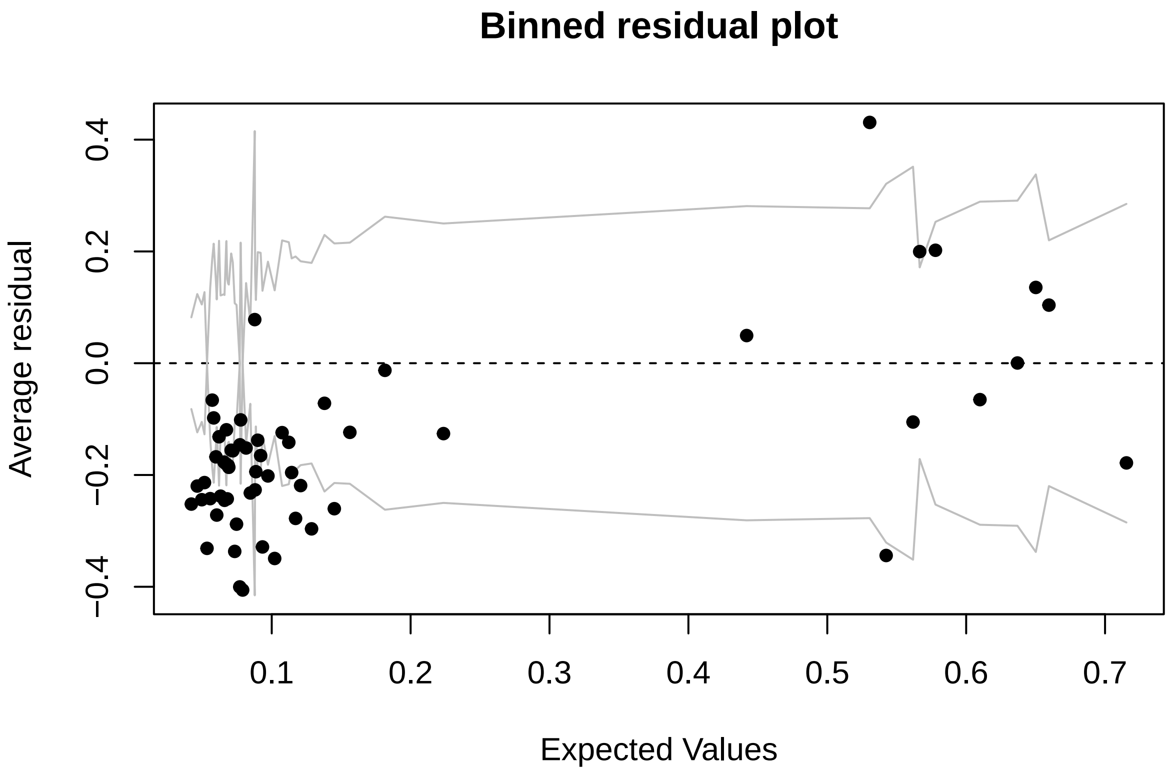
The model is relatively efficient, and the coefficient estimates are consistent with my assumption. By the accuracy test with a 0.5 probability threshold, the model has about 85% accuracy. The model suggests that the employee should focus on the university enrollment status, PhD degree, the city development index of the city living in, previous working history of a candidate to make decisions. However, I did not include gender, major discipline, experience, current working company size, and training hours in the model which may be considered important to some employees and some specific job positions.

Some concerns of this model is that the model residuals are bigger than expectations, I will spend more time improve the performance of the model later. Since the outcome is binary, we need to manually set a threshold between 0 and 1 to conduct the accuracy test, which means for predictions higher than this threshold, a candidate is considered to look for a new job. And the value of this threshold greatly determines the accuracy, I set the threshold to be 0.5, which is a common choice.

Reference

1. Kaggle: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists
2. Example report from Yu Zhe: https://learn.bu.edu/bbcswebdav/pid-9831929-dt-content-rid-61802940_1/xid-61802940_1
3. sjplot reference page: http://www.strengejacke.de/sjPlot/reference/plot_model.html

Appendix



Model Validation:

The following are the accuracy check with p=0.5 as the threshold.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Quit Stay
##       Quit  491  322
##       Stay  344 3593
##
##                Accuracy : 0.8598
##                  95% CI : (0.8496, 0.8695)
##     No Information Rate : 0.8242
##     P-Value [Acc > NIR] : 1.968e-11
##
##                   Kappa : 0.5111
##
##  Mcnemar's Test P-Value : 0.4158
##
##             Sensitivity : 0.5880
##             Specificity : 0.9178
##          Pos Pred Value : 0.6039
##          Neg Pred Value : 0.9126
##              Prevalence : 0.1758
##          Detection Rate : 0.1034
##    Detection Prevalence : 0.1712
##       Balanced Accuracy : 0.7529
##
##        'Positive' Class : Quit
##
```