HR Analytics: Who Will Move to a New Job

Yongrong Chai

12/10/2021

Abstract

In this report, I did an HR analysis on job change of data scientists using data from the Kaggle(https://www.kaggle.com) to explore what and how affected factors make an employee leave his/her current job. I found that training hours, working experience, and city development index have a negative influence on candidates' willingness of finding a new job. Education level, company type, and enrolled university have different effects on the willingness of finding a new job. The report will be helpful on lower the company's employee turnover. There are 5 main parts in this report: Introduction, Method, Result, and Discussion.

Introduction

Companies that specialize 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. To address these questions, I used the factors related to my outcome variable, "target" with 0(Not looking for a job change) and 1(Looking for a job change), and built logistic multilevel models based on different groups with the mixed effect of education level, company type, and enrolled university, and fixed effect of numerical predictors training hours, working experience, and city development index. 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.

Method

Data Processing

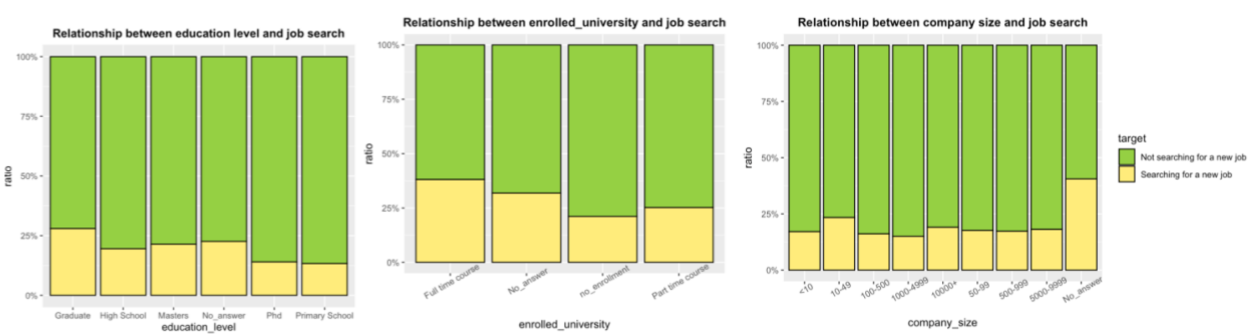
Variable Description:

Variables	Description
enrollee_id	Unique ID for candidate
city	City code
city_development_index	Development index of the city (scaled)

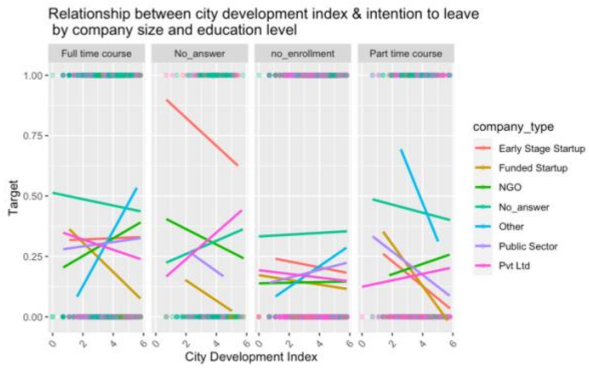
gender	Gender of candidate
relevent_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate total experience in years
company_size	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

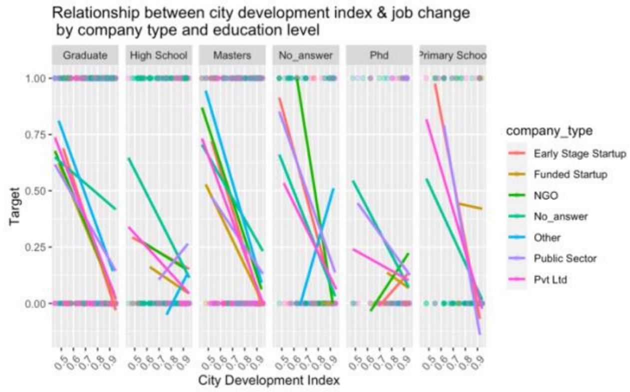
There are 19,158 individuals, and 14 variables with many missing values. Although the multilevel model I would use was not influenced by missing data, it is important to figure out why they are missing. In this case, candidates didn't provide their answers when they registered and enrolled, so the missing data were missing at random, deleting the instances with missing data does not lead to a biased inference. The HR analysis focuses on what factors may affect to willingness of finding a new job. From a company's 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 `training_hours`, `city_development_index`, and `experience` as numerical predictors, and `education_level`, `enrolled_university`, and `company_size`, `relevent_experience`, `major_discipline`, `last_new_job`, and `company_type` are categorical variables for the mixed effect.

Explanatory Data Analysis



Here are several findings from the above figures People with bachelor's and graduate degrees are more likely to change jobs than people at other education levels, with up to 25 percent of undergraduates considering changing jobs. People who do not currently enroll in college are more likely to stay than people who take full-time courses and part-time courses. Nearly 25 percent of employees in companies with 10 to 49 want to change jobs.





Model Fitting

The models I used are the multilevel logistic model. The response variable, target, is binary. I make the training_hours in log scale because this variable is skewed with a heavy tail. For each model, there is a fixed effect of numerical predictors training hours, working experience, and city development index, and one with a random effect of company_type, with a random effect of education_level, and one with a random effect of enrolled_university. Different education levels, company type, major discipline, and enrolled university, and company size have a very small effect on training hours and experience, so I exclude them from my models.

Model Checking

I used binned residual plot to check models, and most of fitted values seem to fall within the SE bands. (Please see Appendix Model Checking for details)

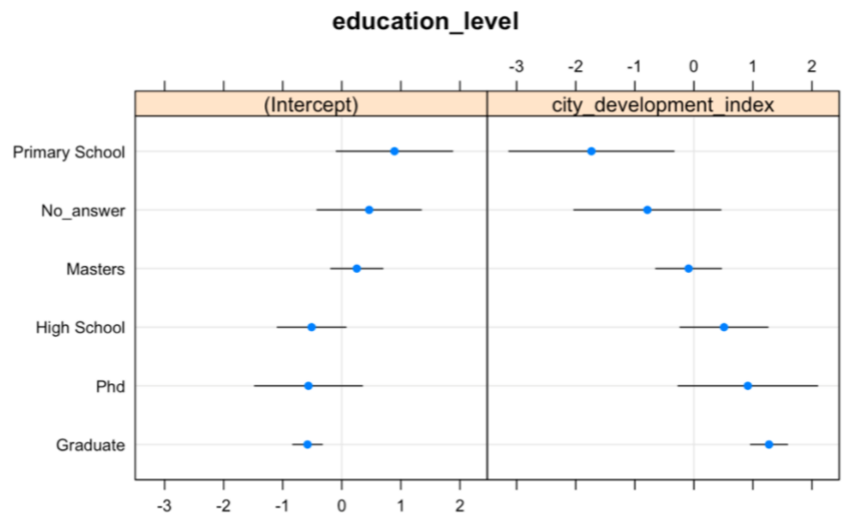
Result

People will be 0.5% to 0.75% less likely to change jobs when 1 unit of log of training hour increases. People will be 0.5% to 1% less likely to change jobs when 1 year of working experience is higher. People will be 134% to 168% less likely to change jobs when the city has 1 unit of city development index higher.

Model1:

$$\Pr(y_i = 1) = logit^{-1}(4.07 + (-0.03)\log(TrainingHours) + (-0.04)Experience + (-6.24)\,CityDevelopmentIndex) + Effect_{EducationLevel})$$

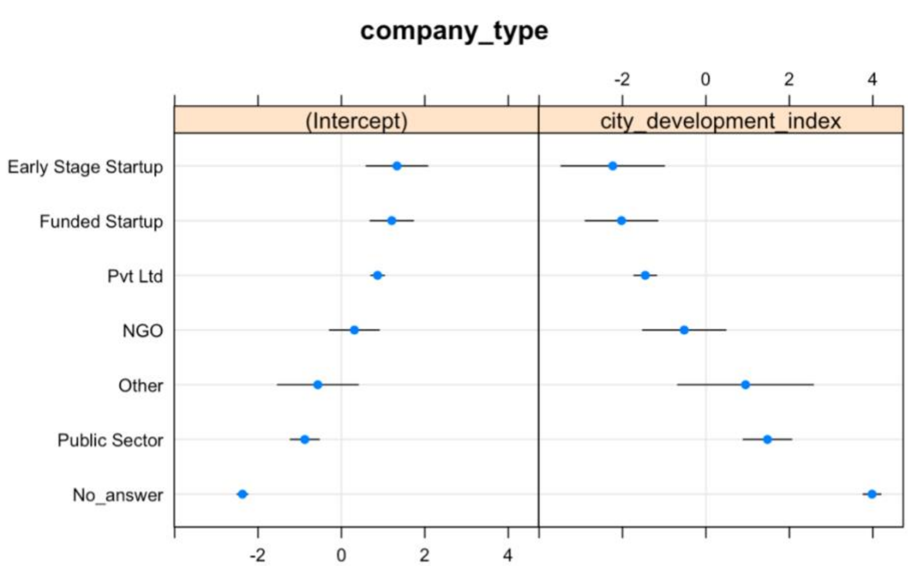
Education Level: For the education level of primary school, there is a strong negative correlation between the city development index and willingness of finding a new job. In the other words, if a person with an education level of primary school, he/she is more likely to stay in the current position if the city development index is high. If a person with an education level of high school, graduate school, and Ph.D., he/she is more likely to find a new job if the city development index is high. For people with master's degrees, their willingness of changing jobs is not affected by the city development index.



Model2:

$$Pr(y_i = 1) = logit^{-1}(4.42 + (-0.03)log(TrainingHours) + (-0.02)Experience + (-6.75)CityDevelopmentIndex) + Effect_{CompanyType})$$

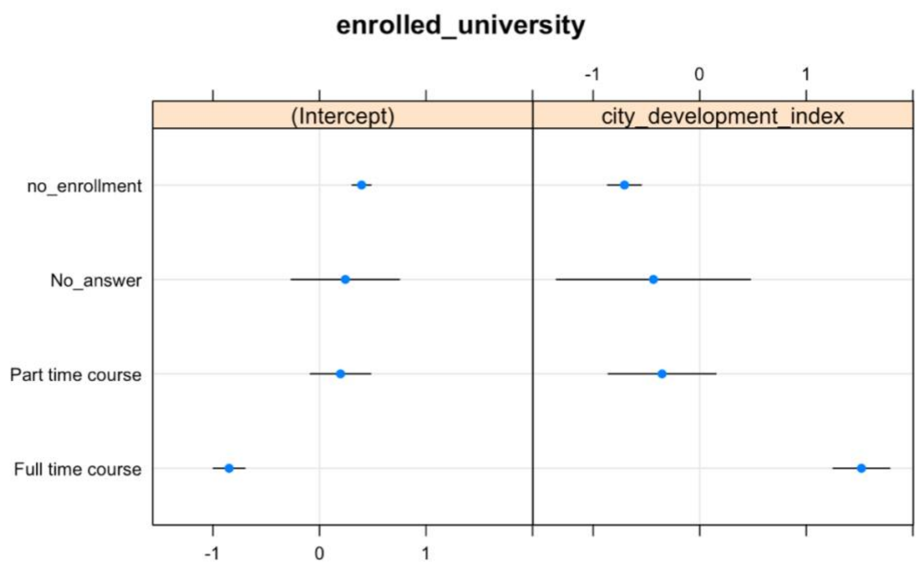
Company Type: For the company type of early-stage startup, funded startup, Pvt. Ltd., and NGO, there is a negative correlation between the city development index and willingness of finding a new job. In the other words, if people are in one of these companies, they are more likely to stay in the current position if the city development index is high. People in the public sector are more likely to stay in the current position if the city development index is low. This makes sense because people who work for nonprofits are motivated by their belief in their organization and its mission, so they don't make the city development index a criterion for choosing a job.



Model3:

$$Pr(y_i = 1) = logit^{-1}(3.57 + (-0.02)log(TrainingHours) + (-0.02)Experience + (-5.38)CityDevelopmentIndex) + Effect_{EnrolledUniversity})$$

Enrolled University: For candidates who currently take part-time courses and have no enrollment, there is a negative correlation between the city development index and willingness of finding a new job. In the other words, they are more likely to stay in the current position if the city development index is high. However, students who are full-time enrolled are more likely to change jobs if the city development index is high. It might be because full-time enrollment students may work as interns, so the current interns are not their final career choices.



Discussion

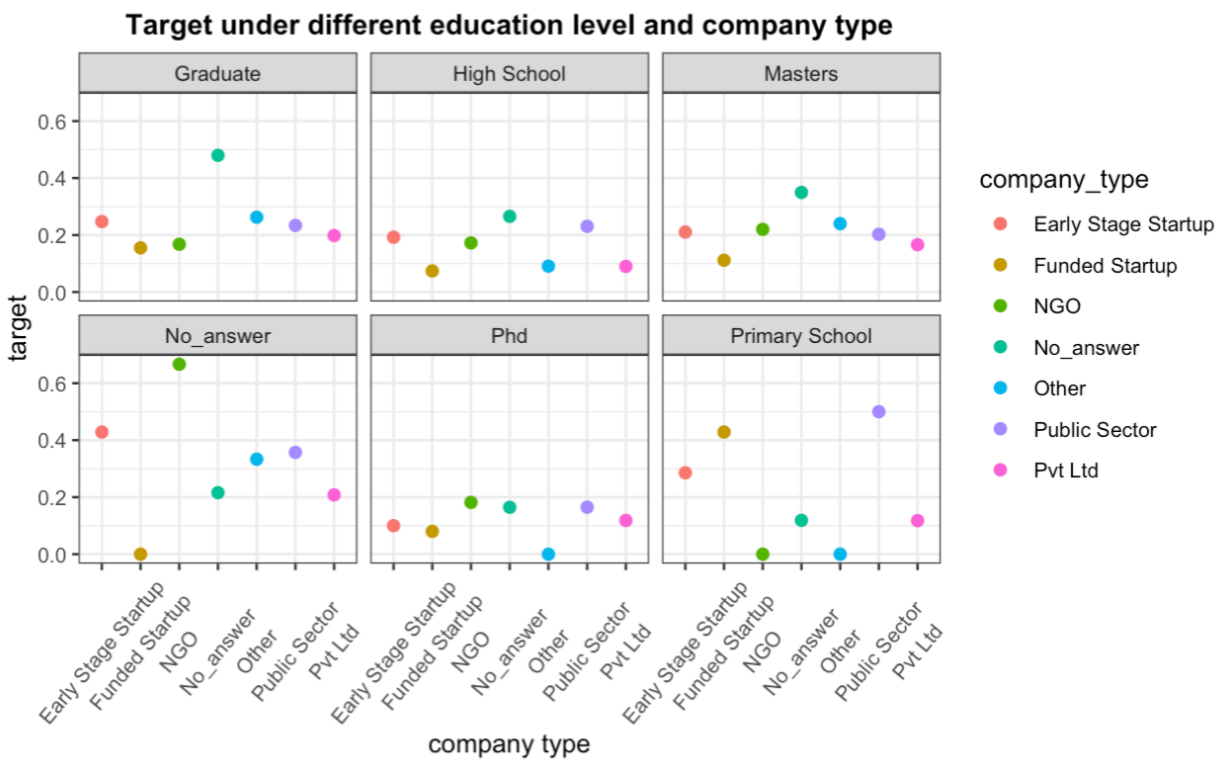
From the perspective of HR, the report is not very helpful for the recruitment of a company's HR department. A company's retention strategies should focus on company culture, career development, employee benefits. None of the predictors in this report should be used as a criterion for recruiting. It may be worth noting that candidates without higher education are less likely to consider changing jobs, and those with less working experience are more likely to do so.

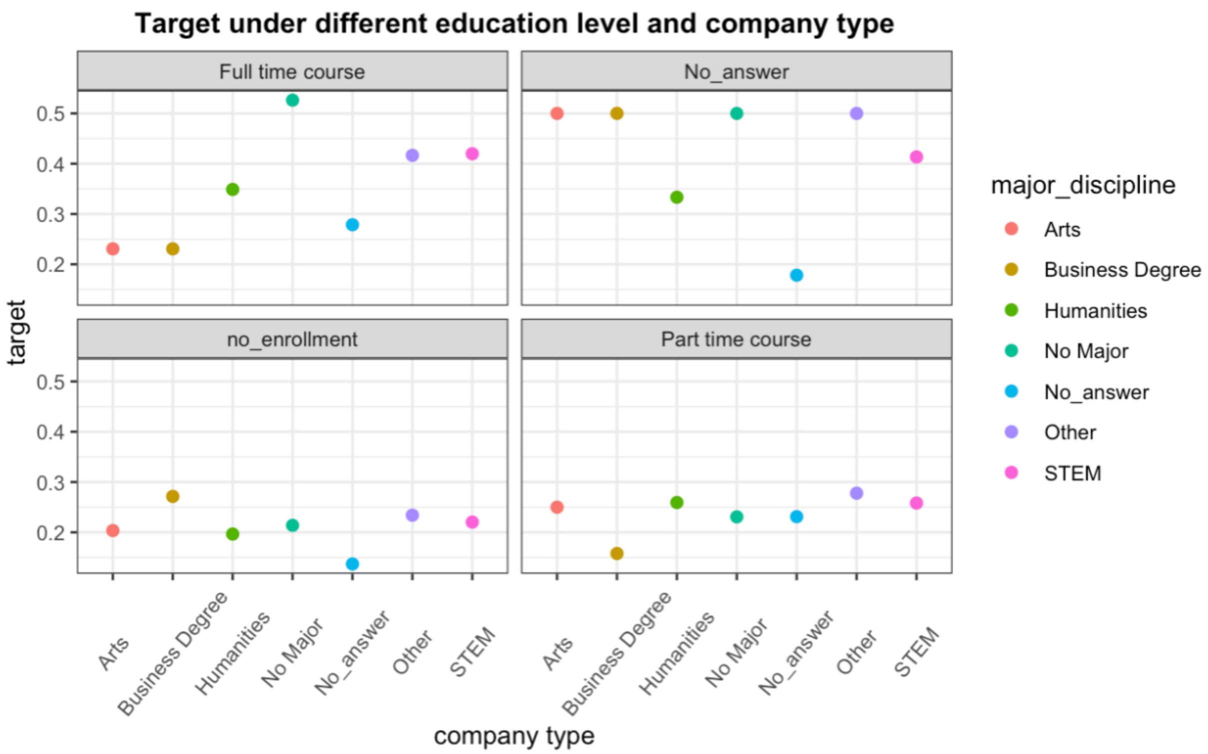
Appendix

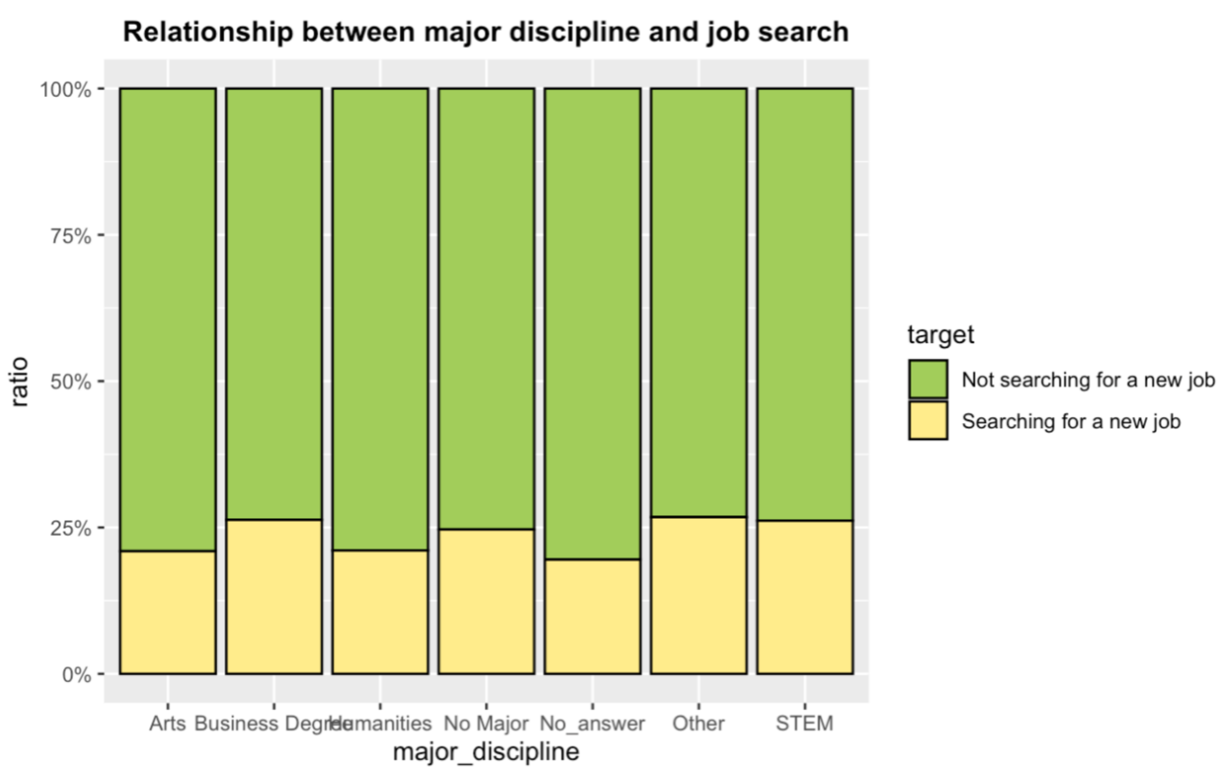
Missing Value

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

More EDA




Target under different education level and company type
Full time course
No_answer
no_enrollment
Part time course
target
0.5
0.4
0.3
0.2
Arts
Business Degree
Humanities
No Major
No_answer
Other
STEM
company type
major_discipline
Arts
Business Degree
Humanities
No Major
No_answer
Other
STEM


Relationship between major discipline and job search
ratio
100%
75%
50%
25%
0%
Arts
Business Degree
Humanities
No Major
No_answer
Other
STEM
major_discipline
target
Not searching for a new job
Searching for a new job

Full Result

Random effects of model1,2,3

```
## $education_level
##                 (Intercept) city_development_index
## Graduate         -0.5802383             1.27269832
## High School      -0.5111331             0.51084107
## Masters           0.2535280            -0.08910321
## No_answer         0.4629206            -0.78643803
## Phd              -0.5647452             0.91413720
## Primary School    0.8913085            -1.73352436
##
## with conditional variances for "education_level"
```

```
## $company_type
##                      (Intercept) city_development_index
## Early Stage Startup   1.3307984             -2.2335227
## Funded Startup        1.2060235             -2.0241088
## NGO                   0.3103263             -0.5208308
## No_answer            -2.3721380              3.9812371
## Other                -0.5658469              0.9496795
## Public Sector        -0.8786749              1.4747089
## Pvt Ltd               0.8666161             -1.4544702
##
## with conditional variances for "company_type"
```

```
## $enrolled_university
##                   (Intercept) city_development_index
## Full time course  -0.8483674              1.5179558
## No_answer          0.2427836             -0.4344047
## no_enrollment      0.3945340             -0.7059266
## Part time course   0.1973589             -0.3531277
##
## with conditional variances for "enrolled_university"
```

Fixed effects of model1,2,3

```
##            (Intercept)    log(training_hours)             experience
##             4.06807975            -0.02637786            -0.03585360
## city_development_index
##            -6.24003867
```

```
##            (Intercept)    log(training_hours)             experience
##             4.41791836            -0.02692614            -0.01831842
## city_development_index
##            -6.75463982
```

```
##            (Intercept)    log(training_hours)             experience
##             3.56668963            -0.02397697            -0.01906258
```

```
## city_development_index
##            -5.37825075
```

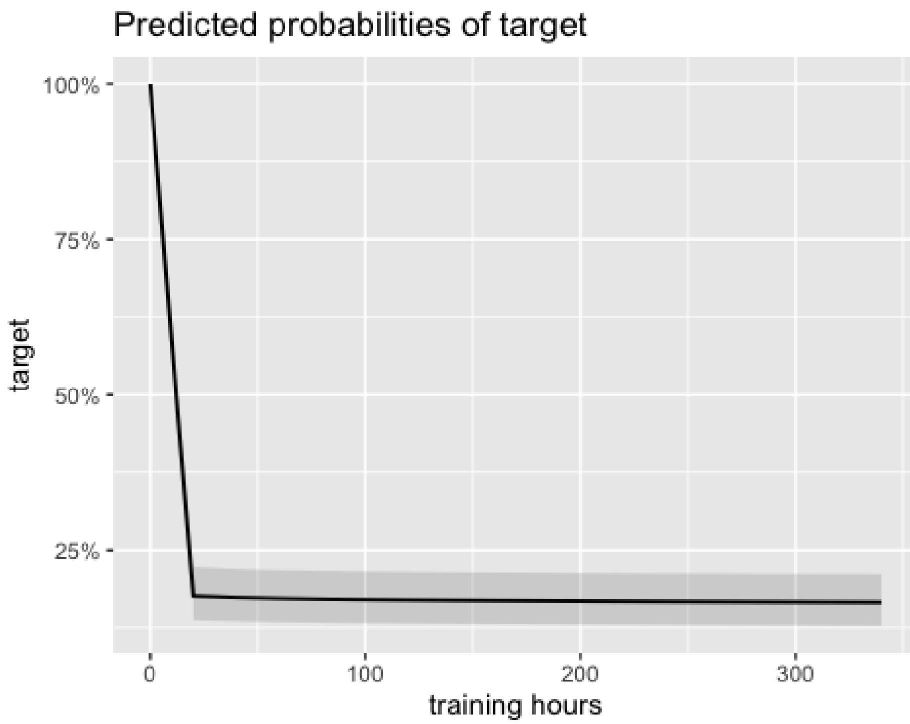
Coefficients of model

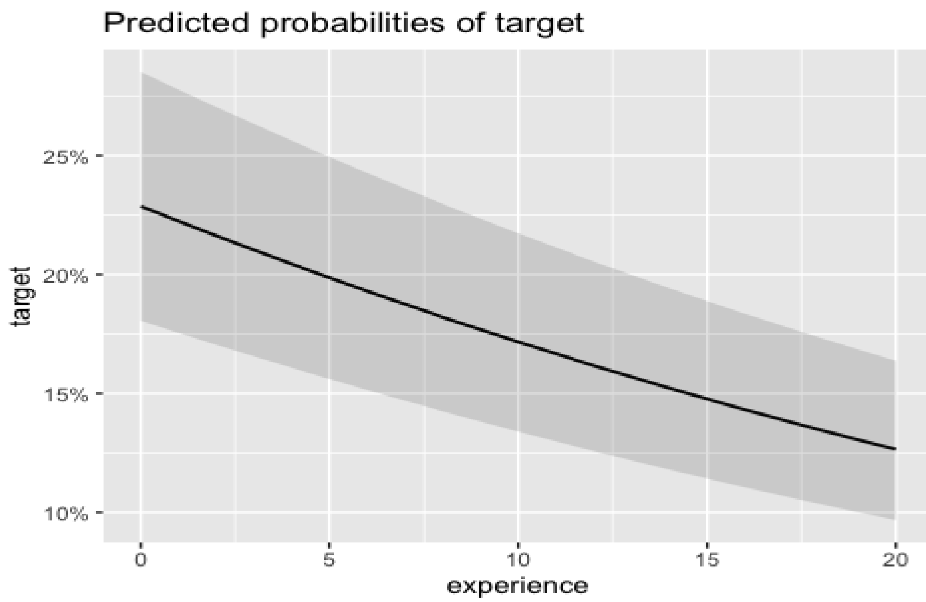
```
## $education_level
##                 (Intercept) log(training_hours) experience
## Graduate           3.487841         -0.02637786 -0.0358536
## High School        3.556947         -0.02637786 -0.0358536
## Masters            4.321608         -0.02637786 -0.0358536
## No_answer          4.531000         -0.02637786 -0.0358536
## Phd                3.503335         -0.02637786 -0.0358536
## Primary School     4.959388         -0.02637786 -0.0358536
##                 city_development_index
## Graduate                     -4.967340
## High School                  -5.729198
## Masters                      -6.329142
## No_answer                    -7.026477
## Phd                          -5.325901
## Primary School               -7.973563
##
## attr(,"class")
## [1] "coef.mer"

## $company_type
##                      (Intercept) log(training_hours)  experience
## Early Stage Startup     5.748717         -0.02692614 -0.01831842
## Funded Startup          5.623942         -0.02692614 -0.01831842
## NGO                     4.728245         -0.02692614 -0.01831842
## No_answer               2.045780         -0.02692614 -0.01831842
## Other                   3.852071         -0.02692614 -0.01831842
## Public Sector           3.539243         -0.02692614 -0.01831842
## Pvt Ltd                 5.284534         -0.02692614 -0.01831842
##                      city_development_index
## Early Stage Startup               -8.988163
## Funded Startup                    -8.778749
## NGO                               -7.275471
## No_answer                         -2.773403
## Other                             -5.804960
## Public Sector                     -5.279931
## Pvt Ltd                           -8.209110
##
## attr(,"class")
## [1] "coef.mer"

## $enrolled_university
##                    (Intercept) log(training_hours)  experience
## Full time course      2.718322         -0.02397697 -0.01906258
## No_answer             3.809473         -0.02397697 -0.01906258
## no_enrollment         3.961224         -0.02397697 -0.01906258
## Part time course      3.764049         -0.02397697 -0.01906258
```

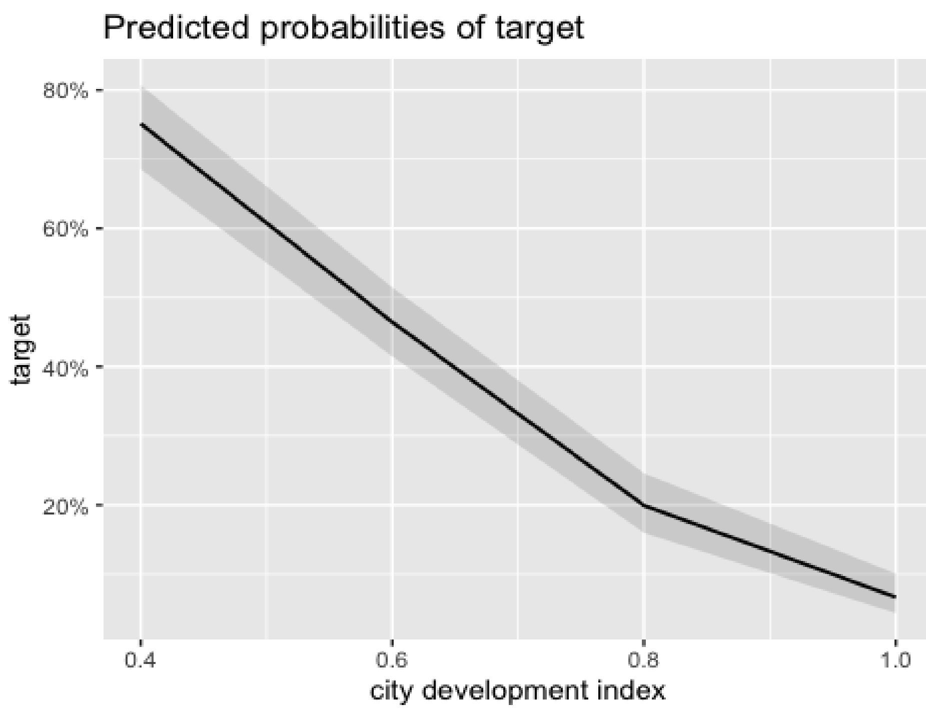
```
##                        city_development_index
## Full time course                   -3.860295
## No_answer                          -5.812655
## no_enrollment                      -6.084177
## Part time course                   -5.731378
##
## attr(,"class")
## [1] "coef.mer"
```

Effect models

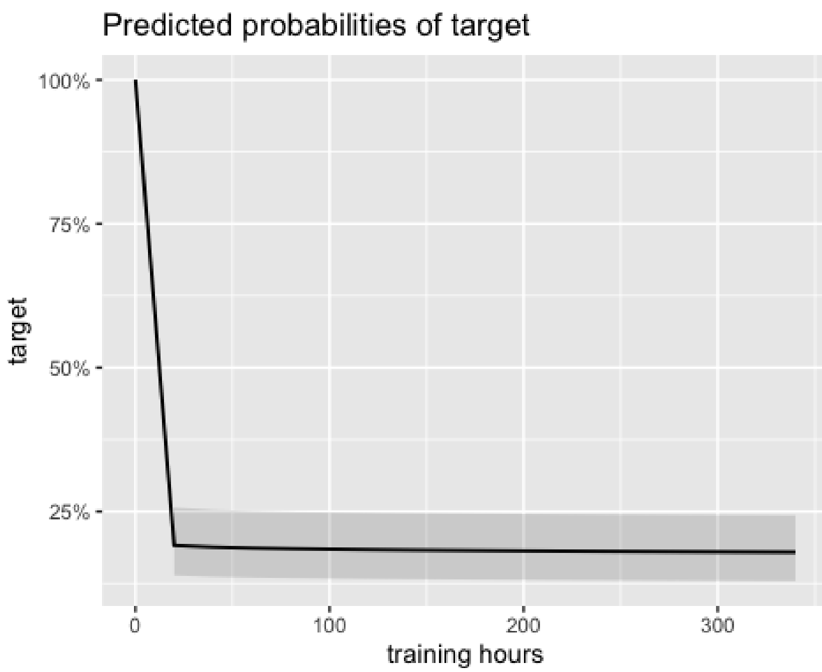
```r
plot_model(model1, type='eff')
```



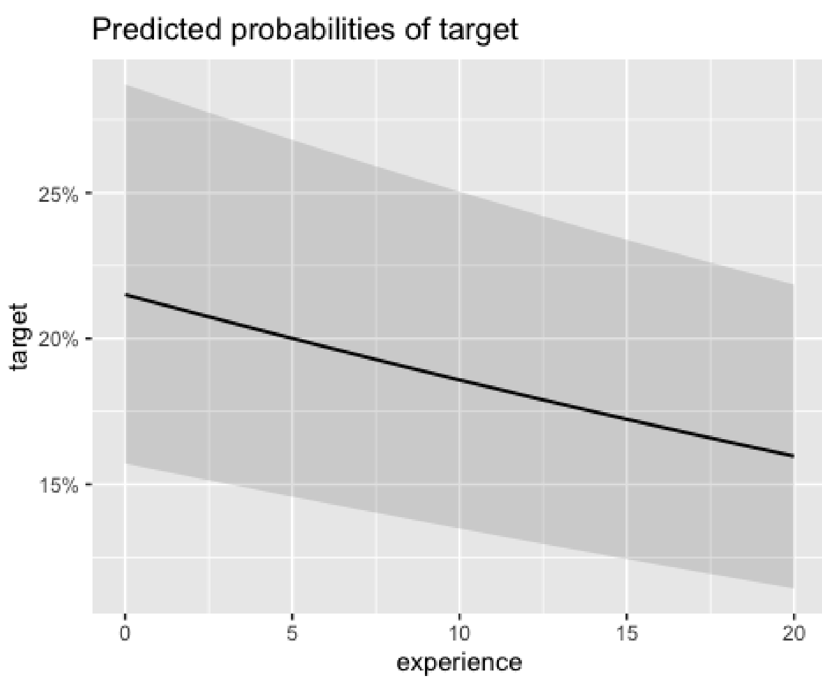




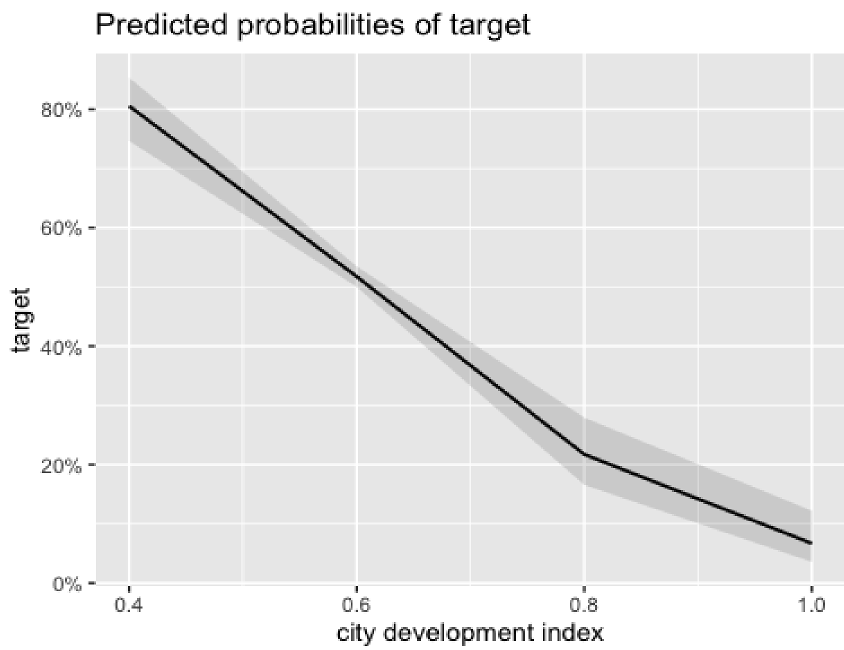

```
plot_model(model2, type='eff')
```



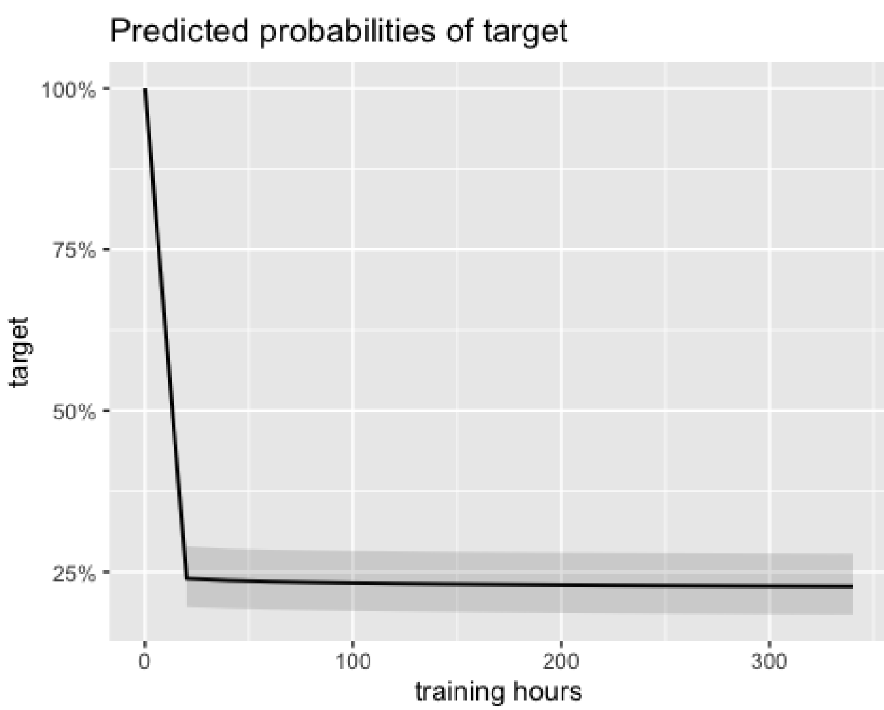
Predicted probabilities of target
100%
75%
50%
25%
target
0
100
200
300
training hours

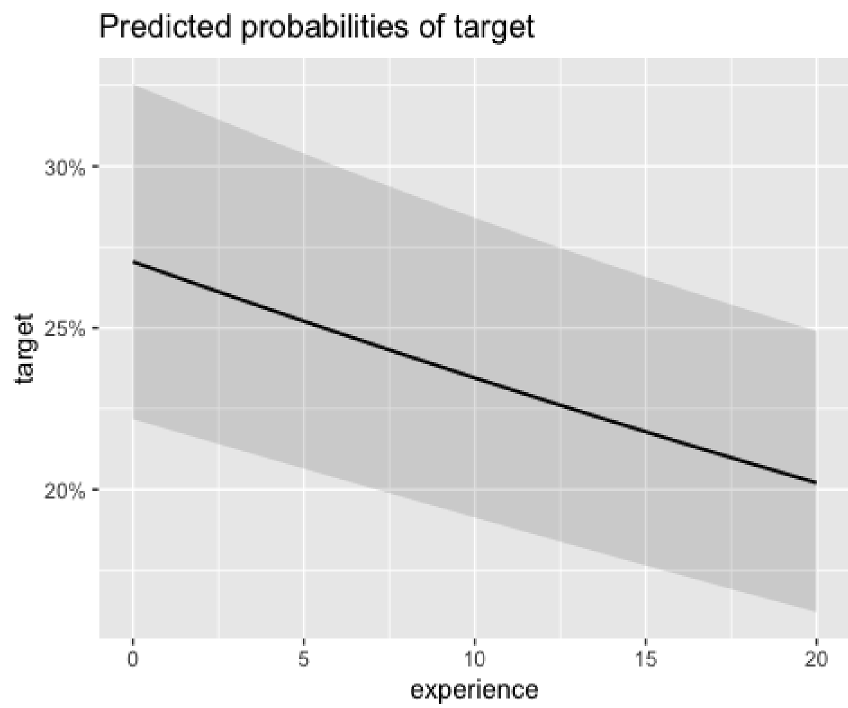


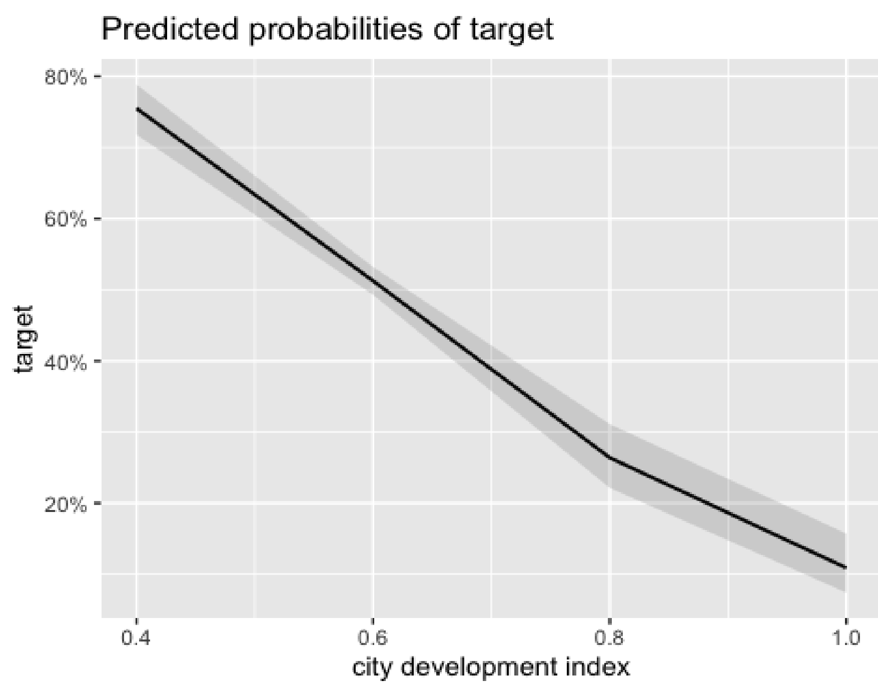
Predicted probabilities of target
25%
20%
15%
target
0
5
10
15
20
experience




```r
plot_model(model3, type='eff')
```

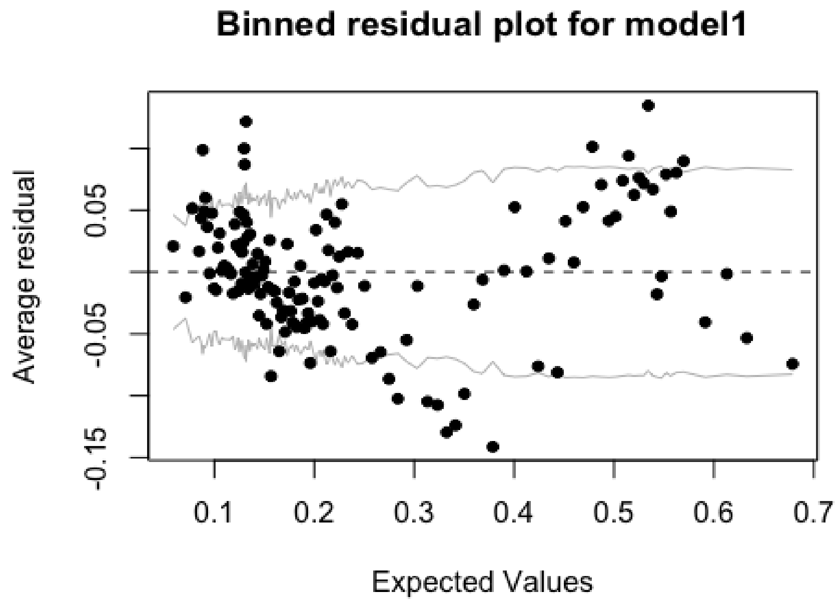




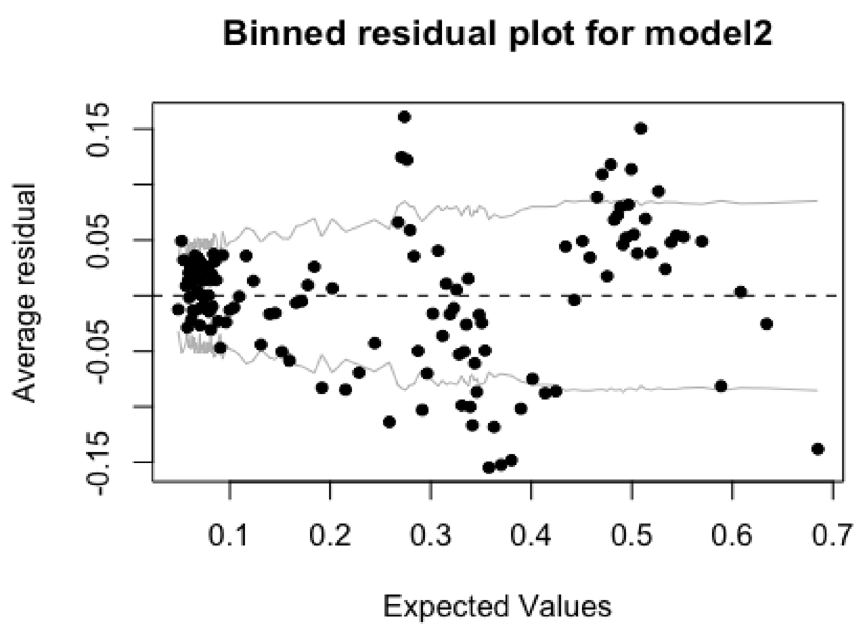


Model Checking

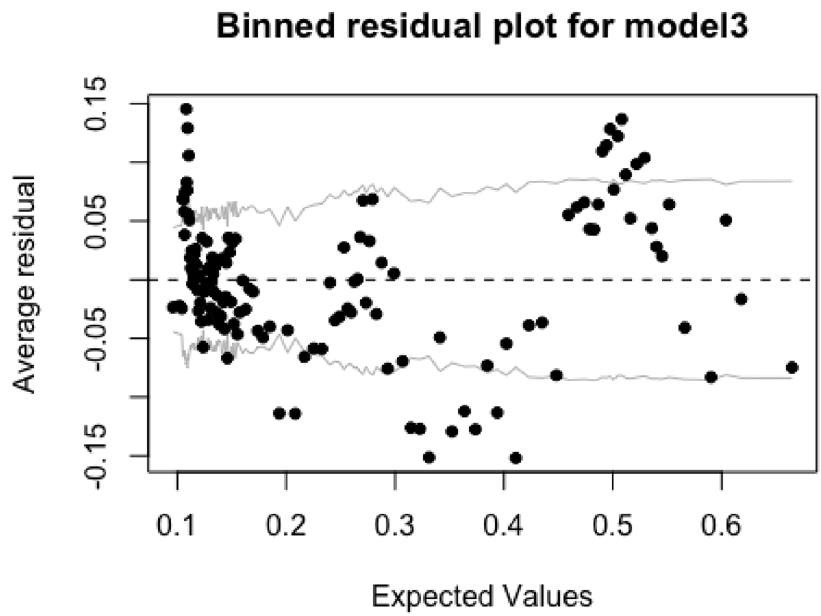
```
binnedplot(fitted(model1),resid(model1,type="response"),main="Binned residual
plot for model1")
```




```
binnedplot(fitted(model2),resid(model2,type="response"),main="Binned residual plot for model2")
```




```
binnedplot(fitted(model3),resid(model3,type="response"),main="Binned residual plot for model3")
```



Binned residual plot for model3
Average residual
0.15
0.05
-0.05
-0.15
0.1
0.2
0.3
0.4
0.5
0.6
Expected Values