HR Analytics: Job Change of Data Scientists

- Predict who will move to a new job

Group Name: Headhunters

Group Member: Zhe Wang
Weijian Shu
Muqiao Cui
Yuqing Gao

Contents

Intro

1. **What is the problem?**
   - As the HR department in big data companies, it's important to know which of current employees are planning to leave the company, the reason behind it and then take corresponding actions. How can we do it?
2. **What is the project goal?**
   - The goal of this task is building model(s) that uses the current credentials,demographics,experience to predict whether a current employee is looking for a new job or not.
3. **Is it a classification or regression problem?**
   - It's classification problem
4. **What benefit will it bring?**
   - It helps to prevent the loss of precious human resources, feedback of the company current working environment and gives guidance in the future recruiting.

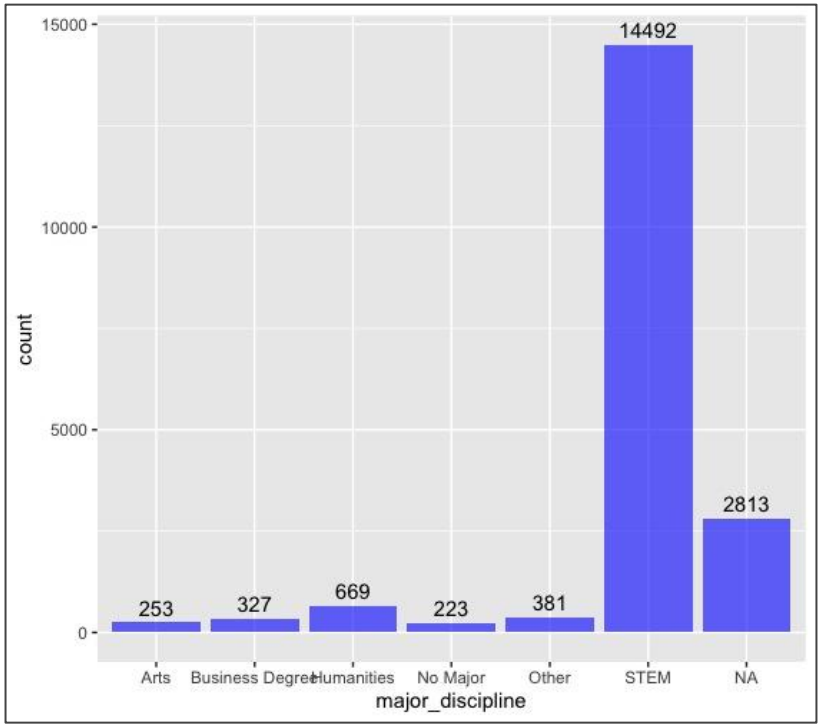
The Data

```
'data.frame':   19158 obs. of  14 variables:
 $ enrollee_id           : int  8949 29725 11561 33241 666 21651 28806 402 27107 699 ...
 $ city                  : chr  "city_103" "city_40" "city_21" "city_115" ...
 $ city_development_index: num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
 $ gender                : chr  "Male" "Male" NA NA ...
 $ relevent_experience   : chr  "Has relevent experience" "No relevent experience" "No relevent experience" "No relevent experience" ...
 $ enrolled_university   : chr  "no_enrollment" "no_enrollment" "Full time course" NA ...
 $ education_level       : chr  "Graduate" "Graduate" "Graduate" "Graduate" ...
 $ major_discipline      : chr  "STEM" "STEM" "STEM" "Business Degree" ...
 $ experience            : chr  ">20" "15" "5" "<1" ...
 $ company_size          : chr  NA "50-99" NA NA ...
 $ company_type          : chr  NA "Pvt Ltd" NA "Pvt Ltd" ...
 $ last_new_job          : chr  "1" ">4" "never" "never" ...
 $ training_hours        : int  36 47 83 52 8 24 24 18 46 123 ...
 $ target                : num  1 0 0 1 0 1 0 1 1 0 ...
```

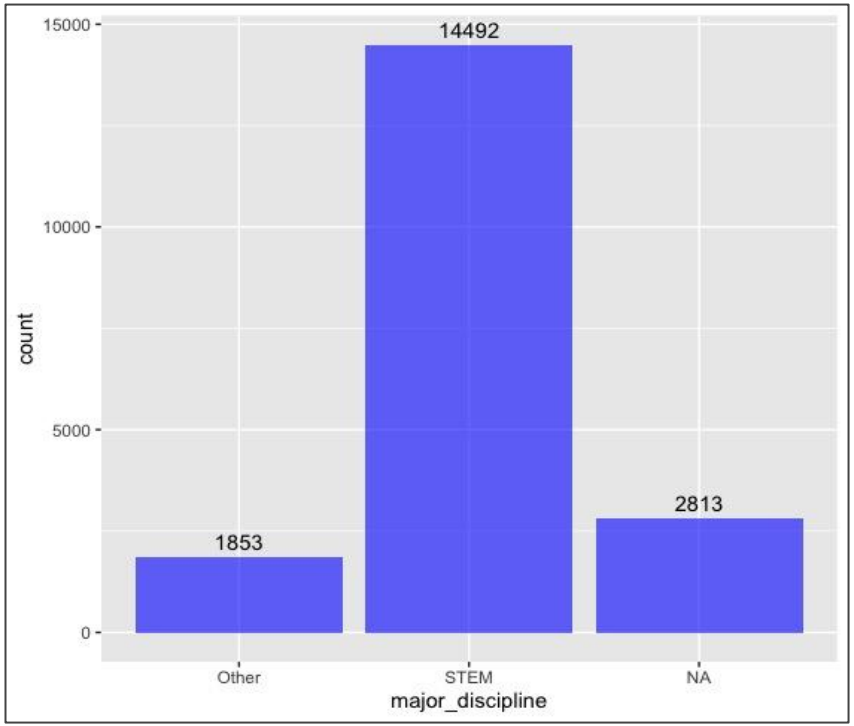
Data Preparation

Aggregating categorical variables - Major

- All other majors



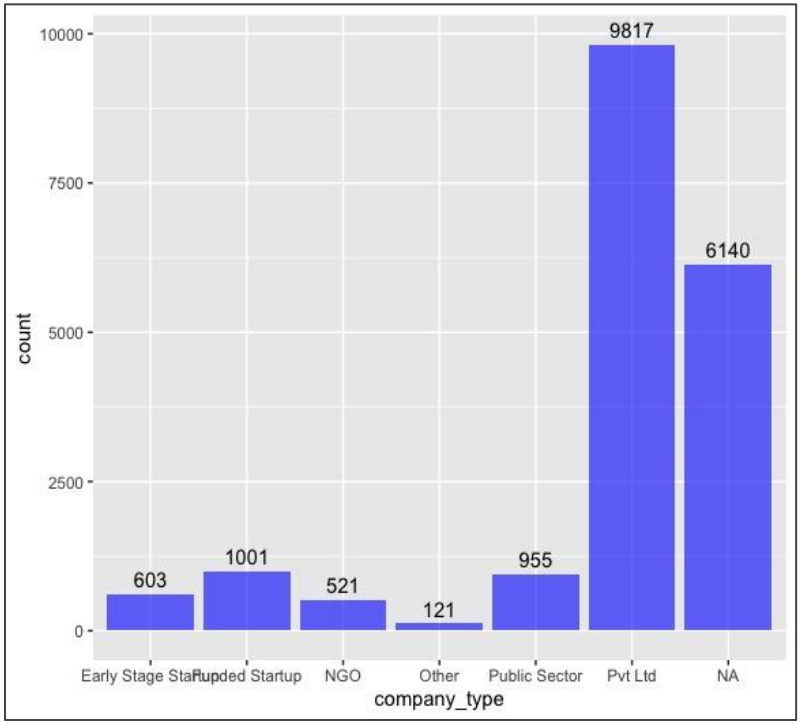
- Other



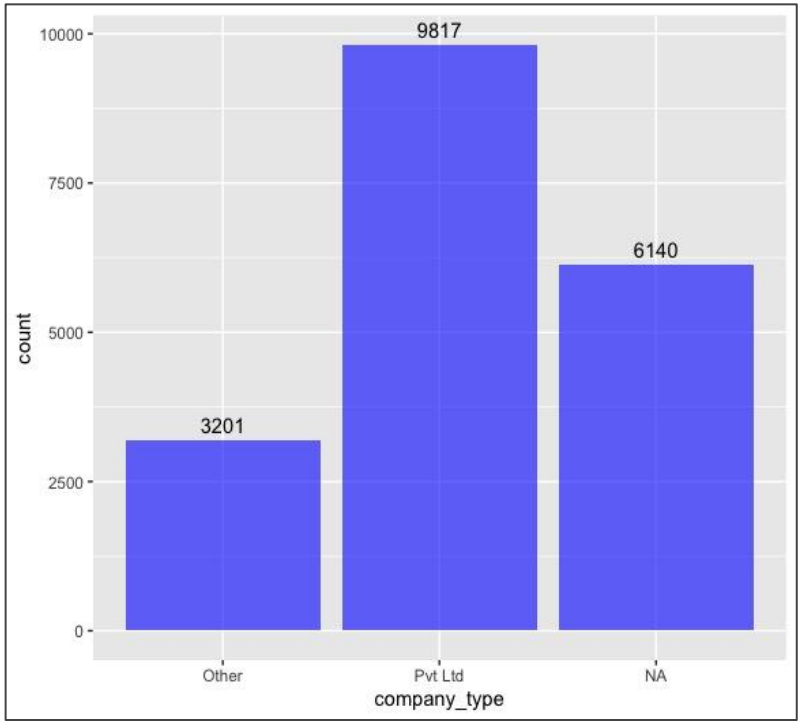
Data Preparation

Aggregating categorical variables - Company Type

- All other company type



- Other

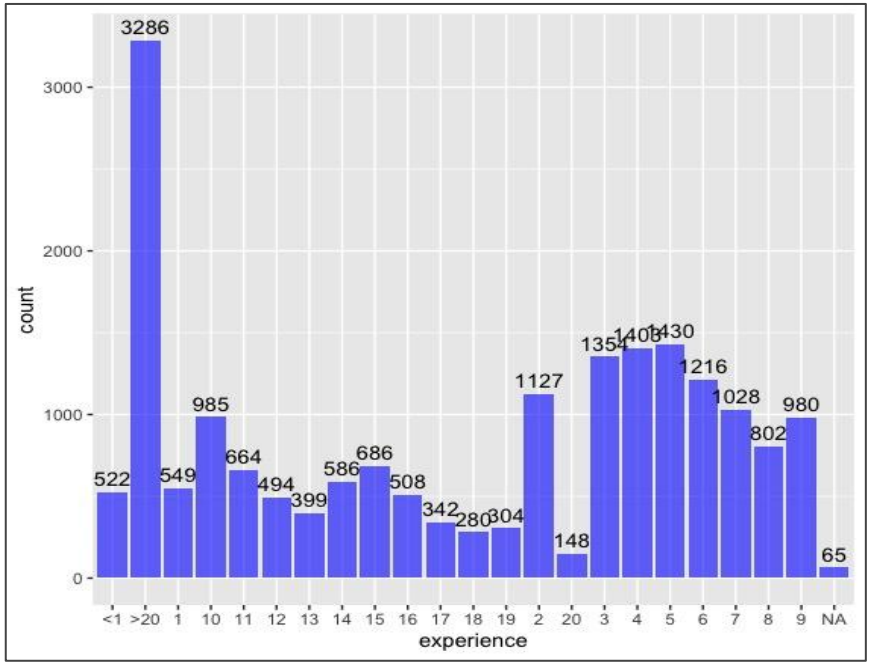


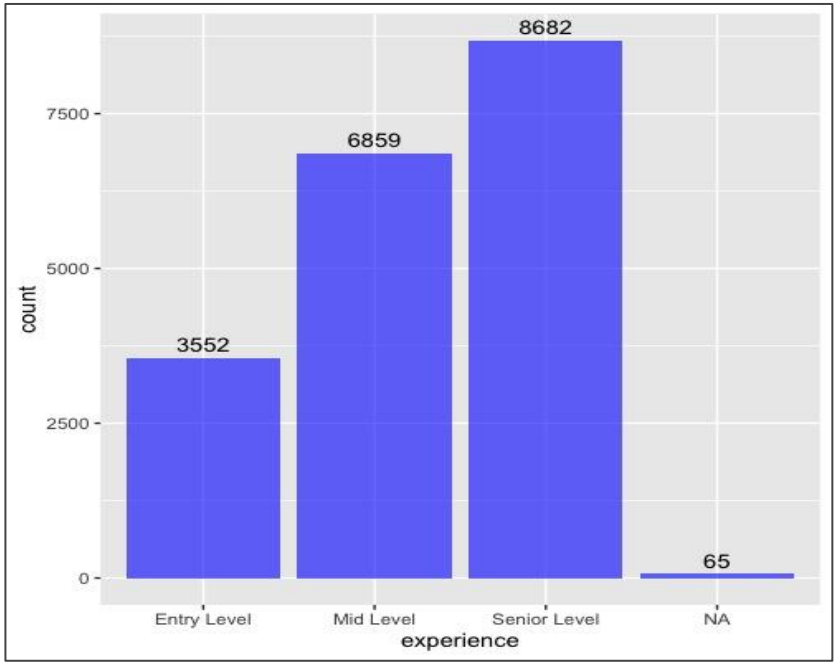
Data Preparation

Aggregating categorical variables - Experience

- Experience <= 3
- 4< Experience < 10
- Experience >= 10

- Entry Level
- Mid Level
- Senior Level

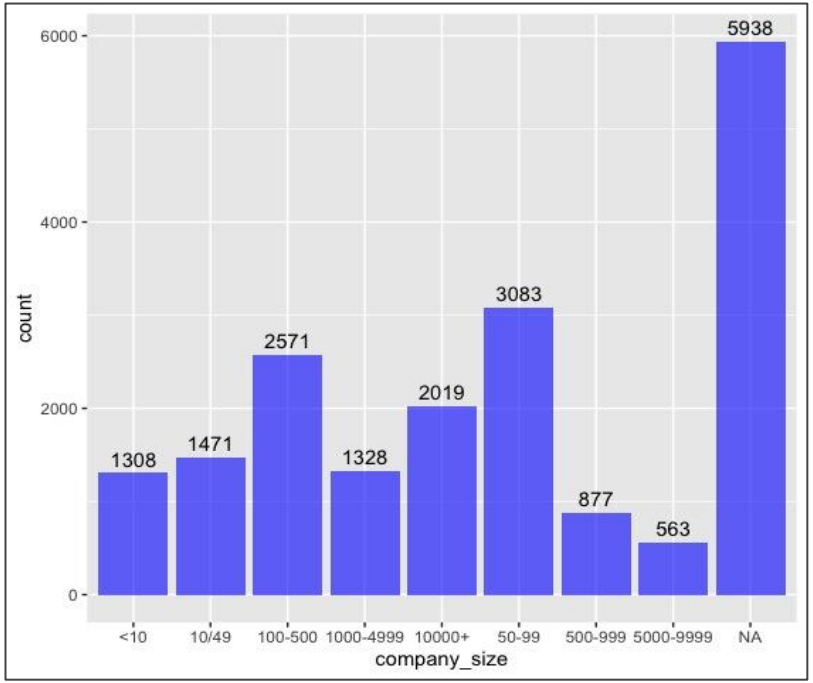




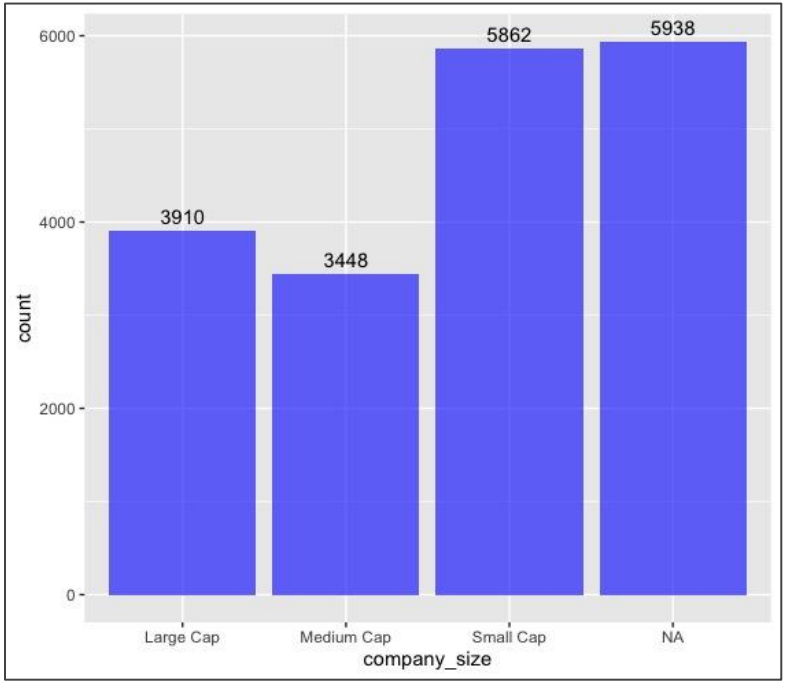
Data Preparation

Aggregating categorical variables - Company Size

- Company Size <= 100
- 100< Company Size < 1000
- Company Size >= 1000



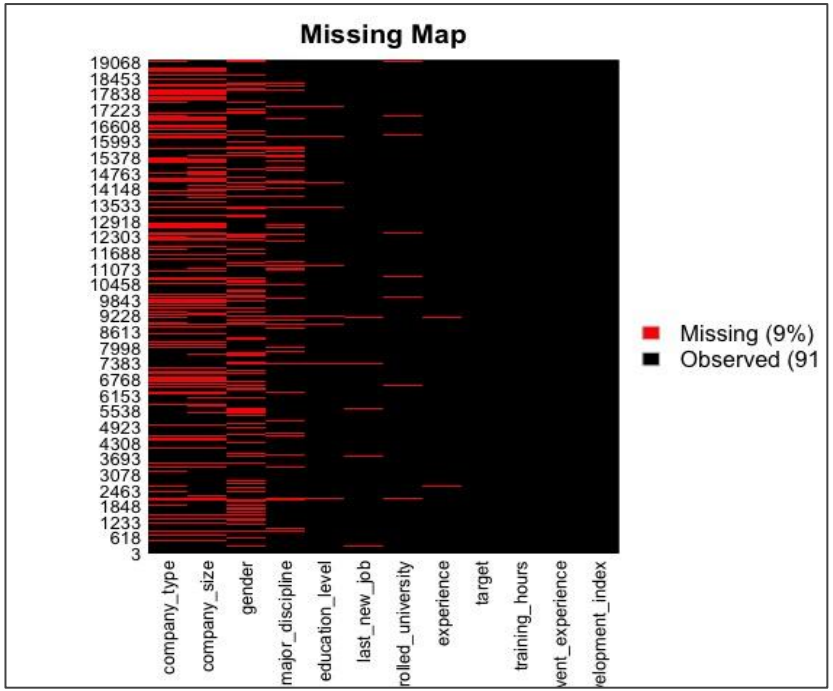
- Small Cap
- Medium Cap
- Large Cap

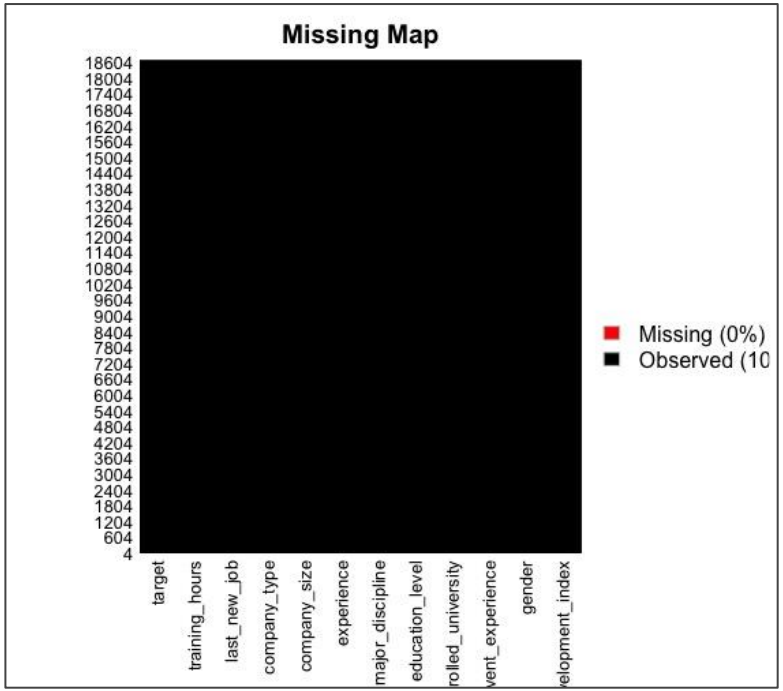


Data Preparation

Handle Missing Data

- Since most of our variables are categorical and there are lots of missing data, our approach is fill in the NAs with ‘Unknown’ or ‘Other’. If the missing value amount is small for some specific variables, we just simply drop them.

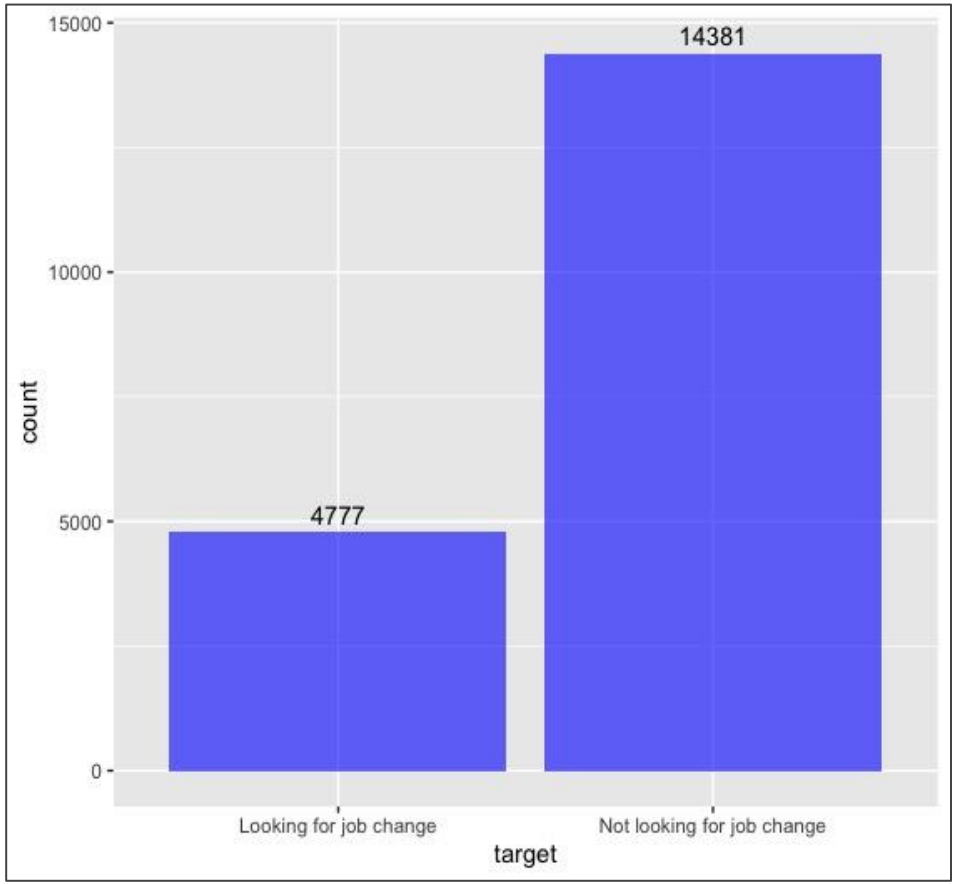




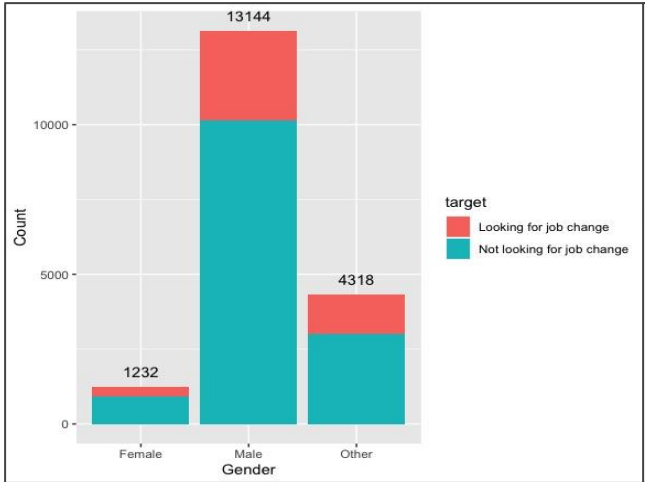
EDA

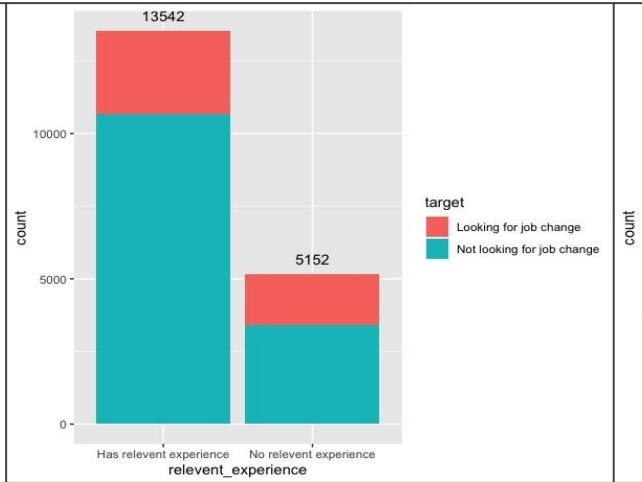
Target variable is not balanced:

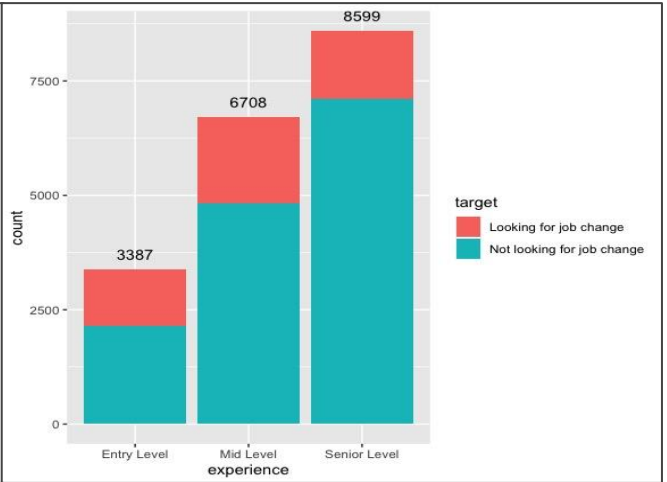
- **25% are looking for job change**
- **75% are not looking for job change**
- **Set a baseline for our model**

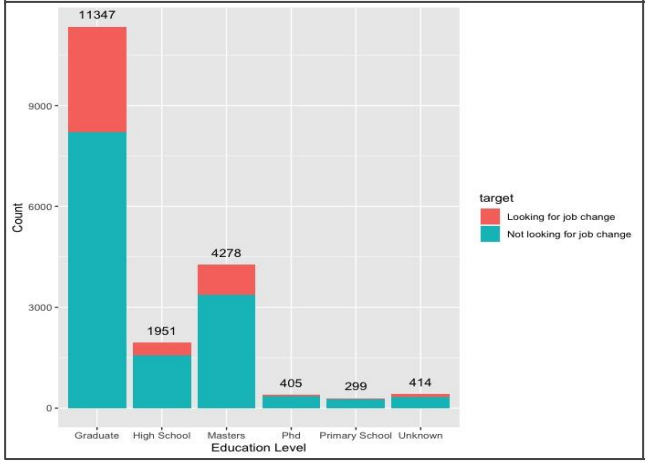


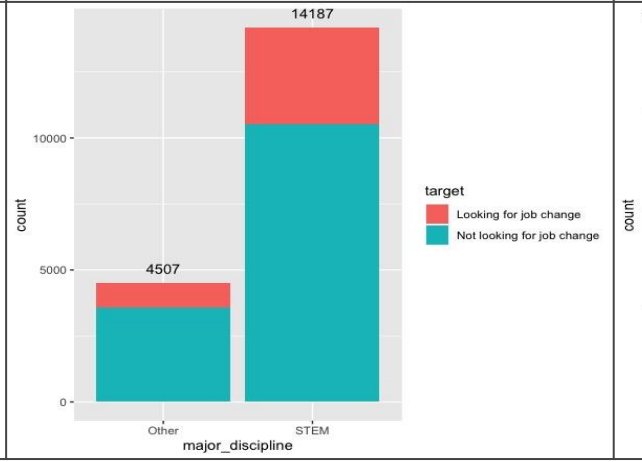
EDA - Count Plot

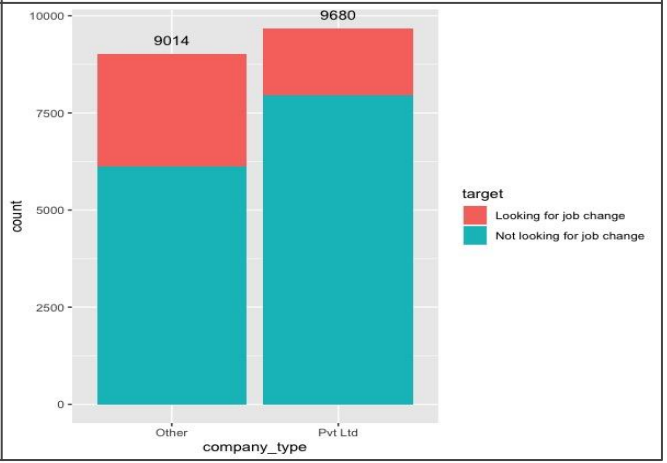




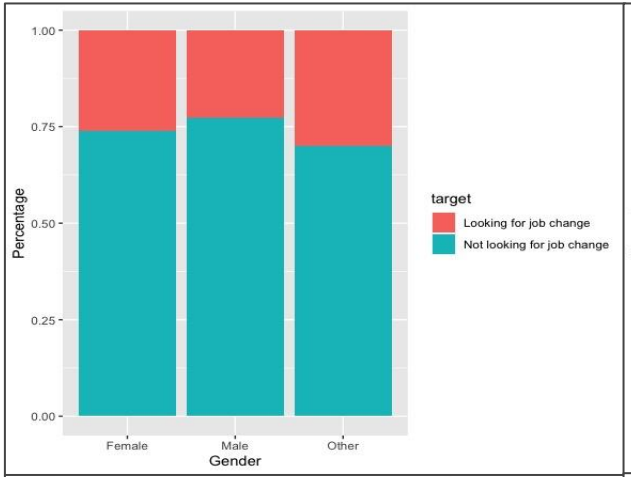


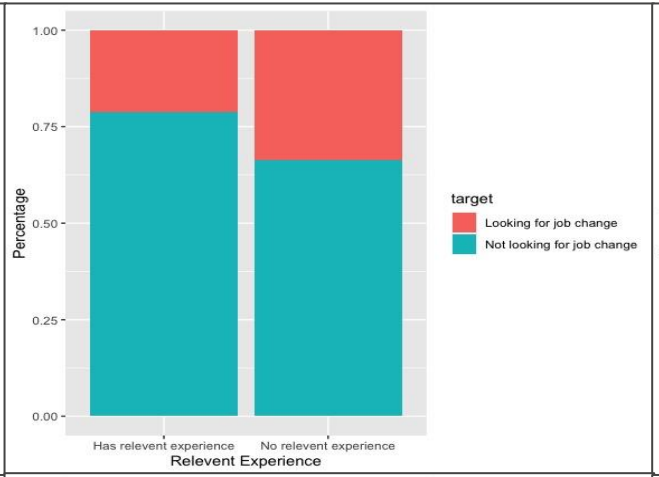


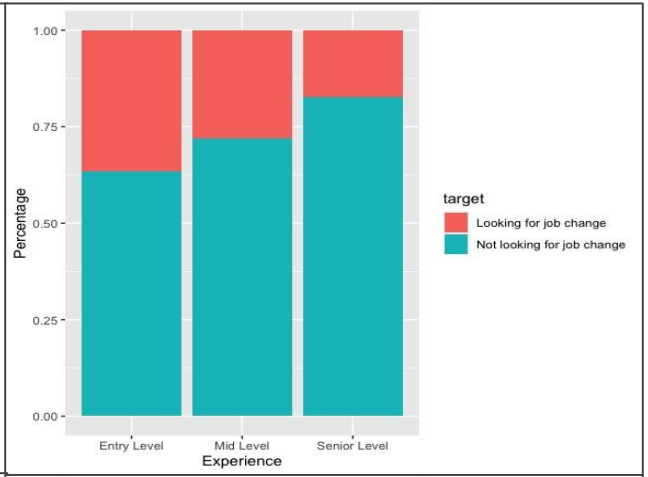


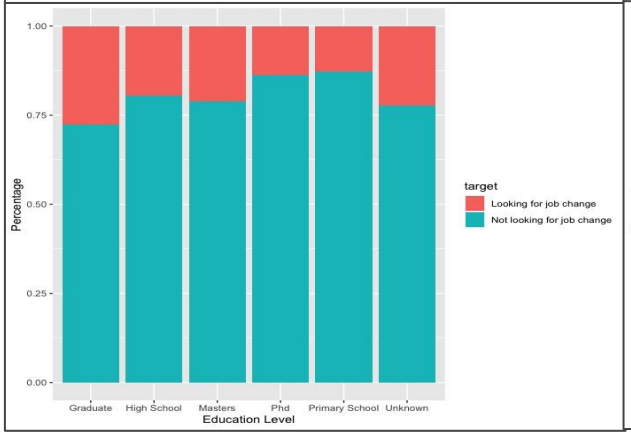


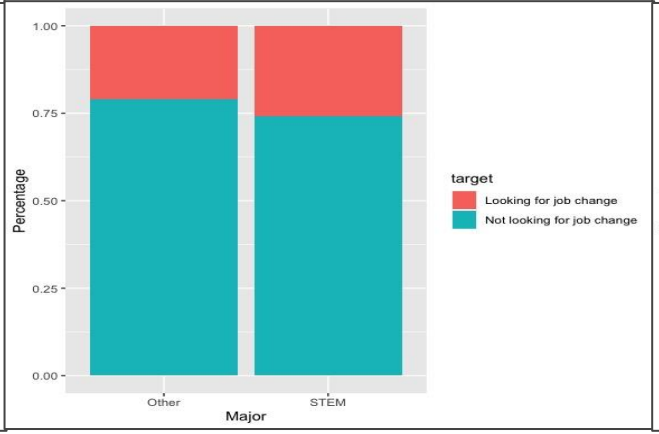
EDA - Percentage distribution

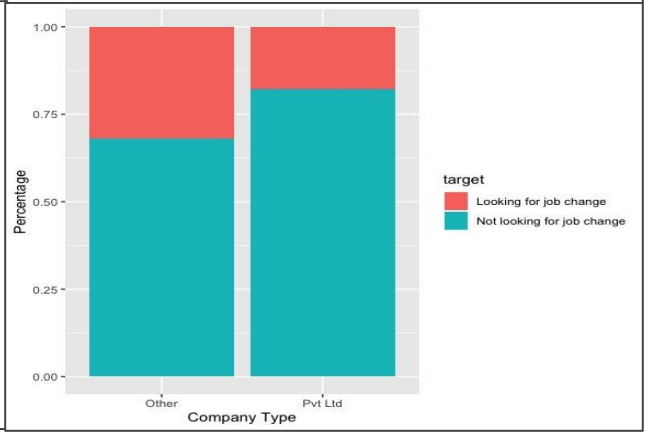










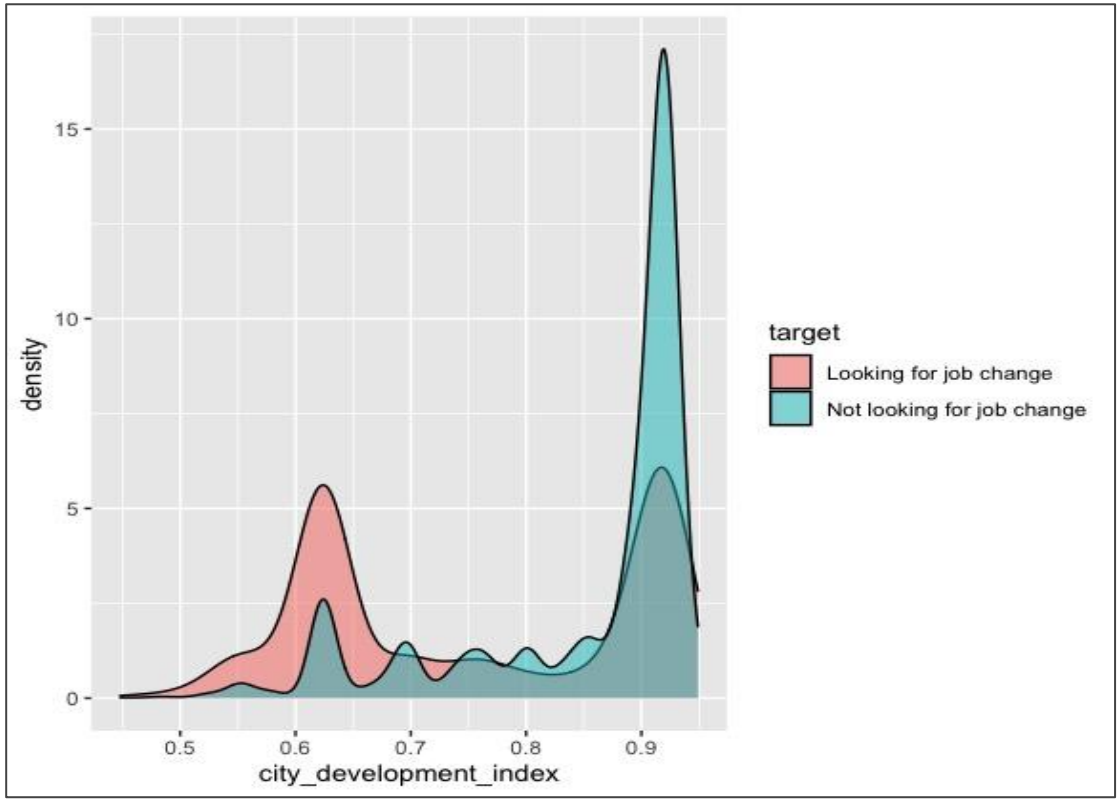


EDA

Distribution of job changing by city development index

Conclusion:

- The CDI plays a big role in the desire to change job
- Above 50% of people who work in a city with a low CDI are looking for a new job.
- On the other hand, in cities with a high CDI, the situation is the opposite, more than half of the people are not interested in finding a new job.

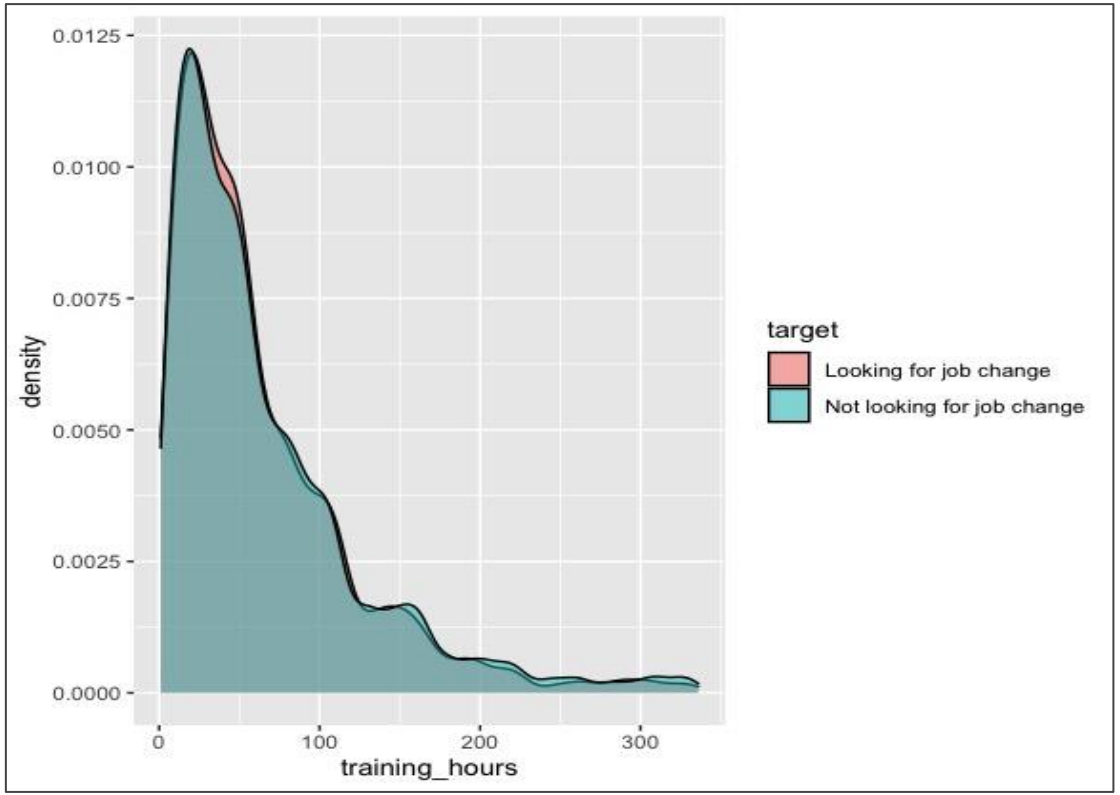


EDA

Distribution of job changing by training hours

Conclusion:

- Training hours means how many training hours the employee did before he/she got hired.
- There are not much difference in terms of training hours



EDA

EDA Conclusion:

- Female are slightly more likely than male to look for a new job.
- People with no relevant experience are more inclined to look for a new job.
- With the increase of working experience, people are less likely to look for new job
- People with graduate education are more likely than others to look for a new job.
- People who have a major discipline STEM (Science, Technology, Engineering and Mathematics) are more likely than others to look for a new job.
- People who work in other type of companies are more likely than people who work in the Private limited company to look for a new job.

Model Building

Random Forest:

- **Build the model on the training set**

  OOB estimate of error rate: 20.46%, thus an accuracy of 79.54%

- **Choose the best Tree size (parameter tuning)**

  We chose 50 trees

- **Variable importance plots**

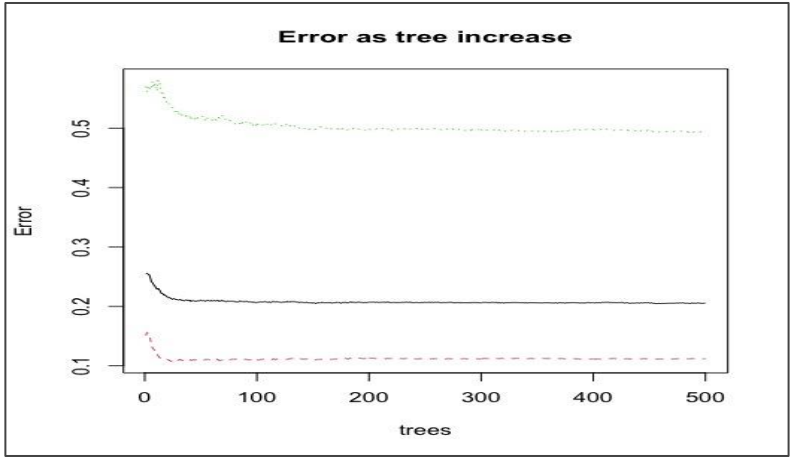
  City_development_index is much more important than others

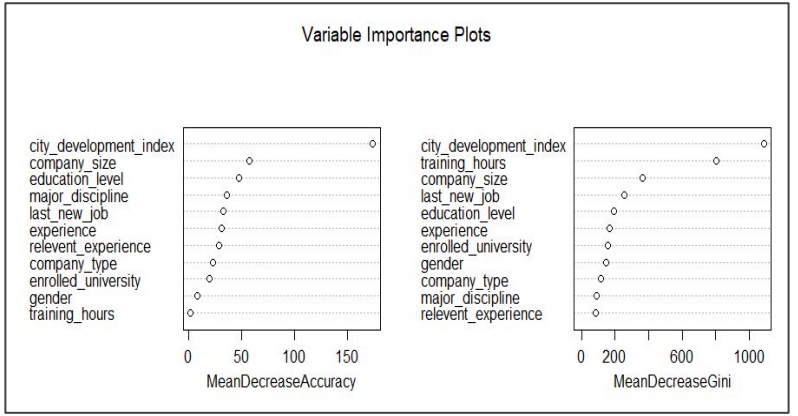
Final results on test set:

Acc: 0.79

Sens : 0.50

Pres : 0.57





Model Building

KNN

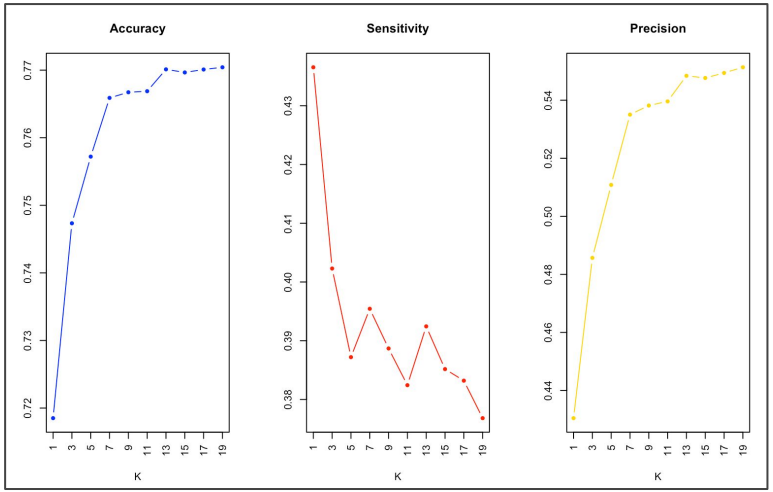
- Normalization
- Convert the categorical variables into dummy variables
- Model building (split data 80%/20%)
- Find optimal K using K-fold cross validation
- plot the results for each KNN model.
- find best k=13
- Show performance

Final results on test set:

Acc: 0.77

Sens: 0.39

Pres: 0.55



	K	accuracy	sensitivity	precision
1	1	0.7185036	0.4365468	0.4304651
2	3	0.7473391	0.4022930	0.4856848
3	5	0.7572048	0.3872061	0.5108213
4	7	0.7658927	0.3954379	0.5350346
5	9	0.7667262	0.3886874	0.5381643
6	11	0.7668840	0.3824370	0.5396045
7	13	0.7701182	0.3924447	0.5484088
8	15	0.7696600	0.3851723	0.5476651
9	17	0.7700957	0.3832030	0.5494015
10	19	0.7704234	0.3768221	0.5513429

Model Building

Logistic Regression

We built **THREE** models for logistic regression:

- Full model (all predictors)
- Reduced model (key predictors)
- One predictor model (citi development index)

Model Comparison based on **AIC** score:

- Full model: 13910.00
- Reduced model: 14063.95
- One model: 15153.37

Final results on test set:

Acc: 0.76

Sens: 0.28

Pres: 0.57

Results for the full model:

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             3.479733   0.205780  16.910  < 2e-16 ***
city_development_index                 -5.790661   0.175002 -33.089  < 2e-16 ***
genderMale                             -0.207171   0.084972  -2.438 0.014764 *
genderOther                            -0.125405   0.091395  -1.372 0.170028
relevent_experienceNo relevent experience 0.222657   0.057131   3.897 9.73e-05 ***
enrolled_universityno_enrollment       -0.215785   0.057607  -3.746 0.000180 ***
enrolled_universityPart time course    -0.245105   0.095919  -2.555 0.010609 *
enrolled_universityUnknown             -0.318530   0.151935  -2.096 0.036039 *
education_levelHigh School             -0.937532   0.104882  -8.939  < 2e-16 ***
education_levelMasters                 -0.243526   0.053632  -4.541 5.61e-06 ***
education_levelPhd                     -0.450773   0.176370  -2.556 0.010593 *
education_levelPrimary School          -1.460889   0.221649  -6.591 4.37e-11 ***
education_levelUnknown                 -0.831905   0.167175  -4.976 6.48e-07 ***
major_disciplineSTEM                    0.083825   0.074905   1.119 0.263107
experienceMid Level                    -0.143895   0.059166  -2.432 0.015014 *
experienceSenior Level                 -0.377073   0.069094  -5.457 4.83e-08 ***
company_sizeMedium Cap                 -0.248839   0.075380  -3.301 0.000963 ***
company_sizeSmall Cap                  -0.185466   0.066405  -2.793 0.005223 **
company_sizeUnknown                     1.121195   0.076046  14.744  < 2e-16 ***
company_typePvt Ltd                    -0.213135   0.055965  -3.808 0.000140 ***
last_new_job1                           0.596504   0.073148   8.155 3.50e-16 ***
last_new_job2                           0.694300   0.087044   7.976 1.51e-15 ***
last_new_job3                           0.731107   0.114373   6.392 1.63e-10 ***
last_new_job4                           0.738305   0.118504   6.230 4.66e-10 ***
last_new_job5+                          0.652998   0.092462   7.062 1.64e-12 ***
training_hours                         -0.001007   0.000359  -2.806 0.005019 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16732  on 14954  degrees of freedom
Residual deviance: 13858  on 14929  degrees of freedom
AIC: 13910
```

Model Comparison

	Random Forest:	Logistic Regression:	KNN:
Accuracy:	0.79	0.76	0.77
Sensitivity:	0.50	0.28	0.39
Precision:	0.57	0.57	0.55

Based on the problem statement: ‘Predict employees who are job seeking’, our model of choice will be Random Forest. This model has the highest accuracy, sensitivity and precision.

Conclusion

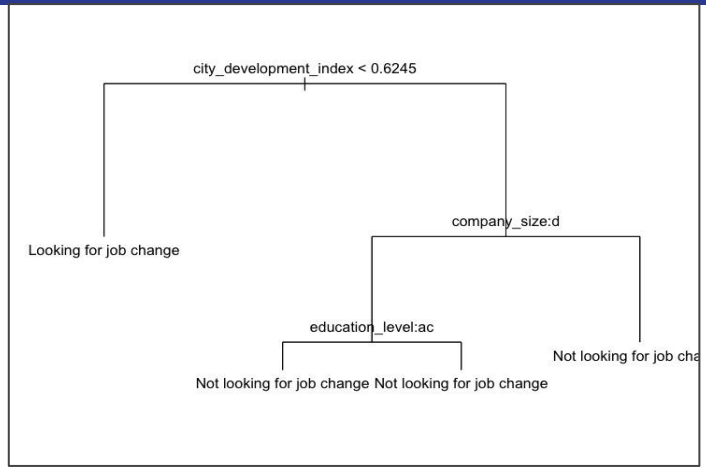
- In terms of accuracy, our random forest model does perform better than our baseline(0.79>0.75)
- Even though our final model outperforms than our other models, the sensitivity score is still relatively low (around 50%).
- Apparently, the most important feature in the task is city development index. Which is not quite good, because it predominates over other features.
- There are some other key variables that are not included in the dataset. For example, employee satisfaction score, salary, employee benefits, reputation of the company, etc. If we include these variables in our dataset, the results might be better.
- Further solution we come up with:
    - Feature insertion & feature engineering based on the most important features.
    - Try some other machine learning algorithms like neural networks, SMOTE Logistic Regression, etc.

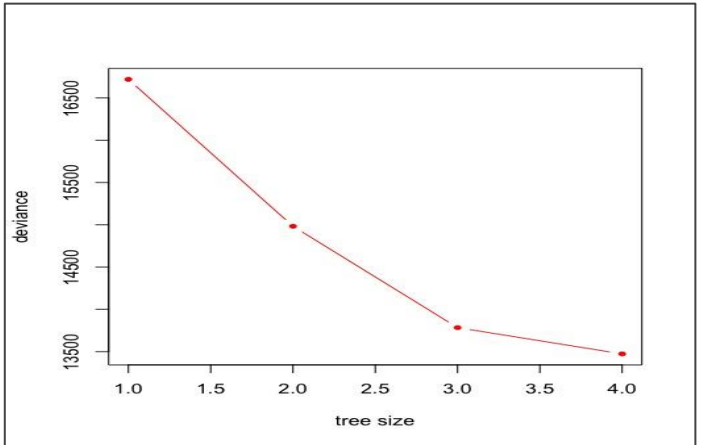
Thanks for watching!

Model Building

Decision Tree:

- We use 10-fold cross validation
- Plot the error by tree size
- Prune the tree and plot it again
- Using your pruned tree, make predictions
- Confusion matrix
- Calculate accuracy；sensitivity；precision
- Show results





Final results on test set:

Acc: 0.79

Sens：0.43

Pres：0.59