HR Analytics: Job Change of Data Scientists

Rodrigo Moraes Kunrath

7/25/2021

Introduction

Overview

The goal of this project is to predict whether an employee desires to leave his current job. It has been developed with R and is part of the Data Science's Capstone project from HarvardX.

The dataset was downloaded from a Kaggle proposed task and it is available in *https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists*. This project used only the *aug_train.csv* original file that was stored as *data/dataset.csv*. The whole project can be obtained through the GitHub repositoty in *https://github.com/rmkunrath/hr_analytics*.

The context that the dataset was created is that it comes from a poll generated by a company which is active in Big Data and Data Science and wants to hire data scientists among people who successfully pass some courses which were conducted by the company. The company wants to know which of these candidates are really looking to work for them after training. Information related to demographics, education, experience are from candidates signup and enrollment. In this project all enrollees were considered employees.

The general idea is to understand the factors that lead a person to leave their current job using models that use the current credentials, demographics, experience data. The final objective is to create a machine learning algorithim that predicts the employee desire to leave.

To achieve this goal, the data was imported, explored, wrangled, divided in training and testing datasets. Machine learning models were proposed with a subset of the training dataset and evaluated with another subset of the training. Finally a model was selected and applied to the testing dataset.

The dataset

The dataset has 19158 entries and is composed of the following 14 features:

- enrollee_id : Unique ID for candidate
- city: City code
- city_ development _index : Developement 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
- lastnewjob: 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

The collumn target is the value we are trying to understand and predict.

Bellow a summary of the dataset is shown:

```
summary(dt)
```

```
##   enrollee_id          city      city_development_index    gender
##  Min.   :    1   city_103:4355   Min.   :0.4480         Female: 1238
##  1st Qu.: 8554   city_21 :2702   1st Qu.:0.7400         Male  :13221
##  Median :16982   city_16 :1533   Median :0.9030         Other :  191
##  Mean   :16875   city_114:1336   Mean   :0.8288         NA's  : 4508
##  3rd Qu.:25170   city_160: 845   3rd Qu.:0.9200
##  Max.   :33380   city_136: 586   Max.   :0.9490
##                  (Other) :7801
##               relevent_experience       enrolled_university
##  Has relevent experience:13792    Full time course: 3757
##  No relevent experience : 5366    no_enrollment   :13817
##                                   Part time course: 1198
##                                   NA's            :  386
##
##
##
##        education_level         major_discipline   experience
##  Graduate      :11598   Arts           :  253   >20    : 3286
##  High School   : 2017   Business Degree:  327   5      : 1430
##  Masters       : 4361   Humanities     :  669   4      : 1403
##  Phd           :  414   No Major       :  223   3      : 1354
##  Primary School:  308   Other          :  381   6      : 1216
##  NA's          :  460   STEM           :14492   (Other):10404
##                         NA's           : 2813   NA's   :   65
##     company_size               company_type  last_new_job training_hours
##  50-99    :3083   Early Stage Startup: 603   >4   :3290   Min.   :  1.00
##  100-500  :2571   Funded Startup     :1001   1    :8040   1st Qu.: 23.00
##  10000+   :2019   NGO                : 521   2    :2900   Median : 47.00
##  10/49    :1471   Other              : 121   3    :1024   Mean   : 65.37
##  1000-4999:1328   Public Sector      : 955   4    :1029   3rd Qu.: 88.00
##  (Other)  :2748   Pvt Ltd            :9817   never:2452   Max.   :336.00
##  NA's     :5938   NA's               :6140   NA's : 423
##      target
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2493
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

Methods

In this section the methods applied to gain insight as well as the modeling developed are going to be described.

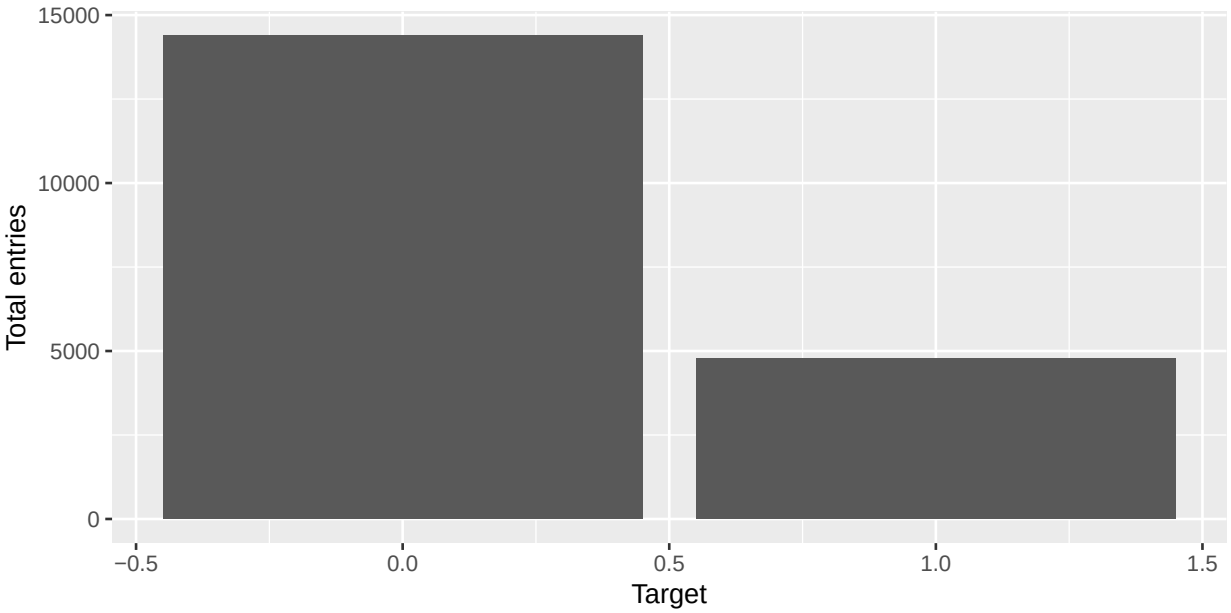
Data Exploration

The dataset is composed of 19158 entries and has 14 features. Its sumary has been shown in the *Dataset* section. The target is the feature this project wishes to predict and gain insight, it has a mean rate of about

25% and, in other words, it means that one fourth of the employees are looking for a new job.

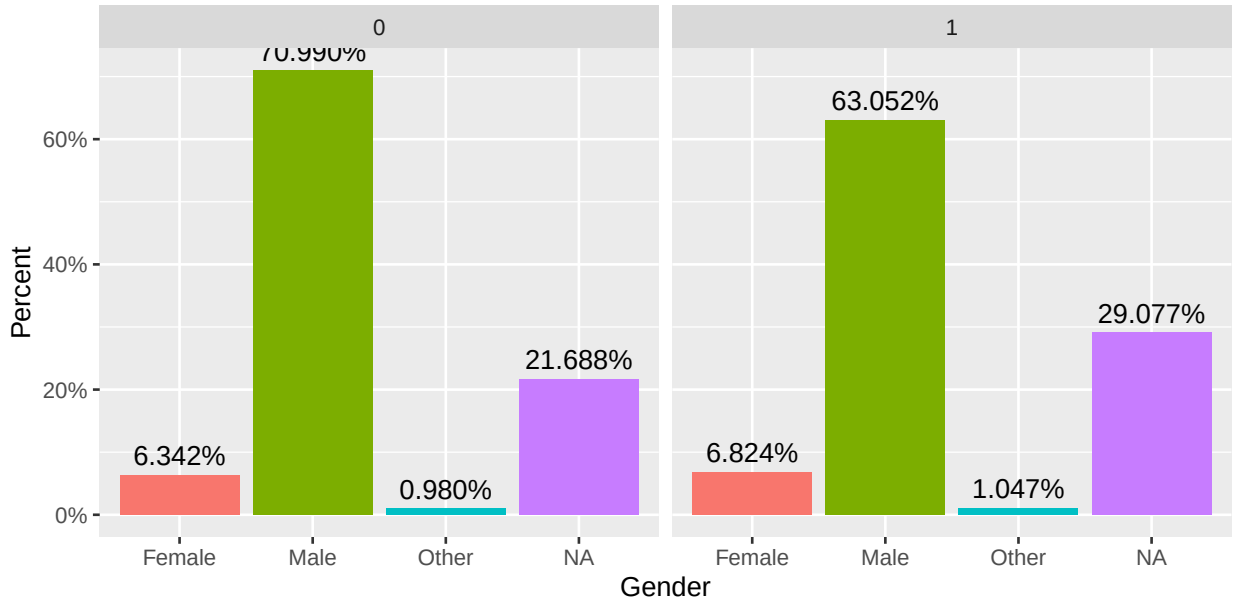
All information considered NA (not available) in this report are empty values in the dataset. It means that the employee hasn't filled the feature when asked.

Target mean rate
0.2493475

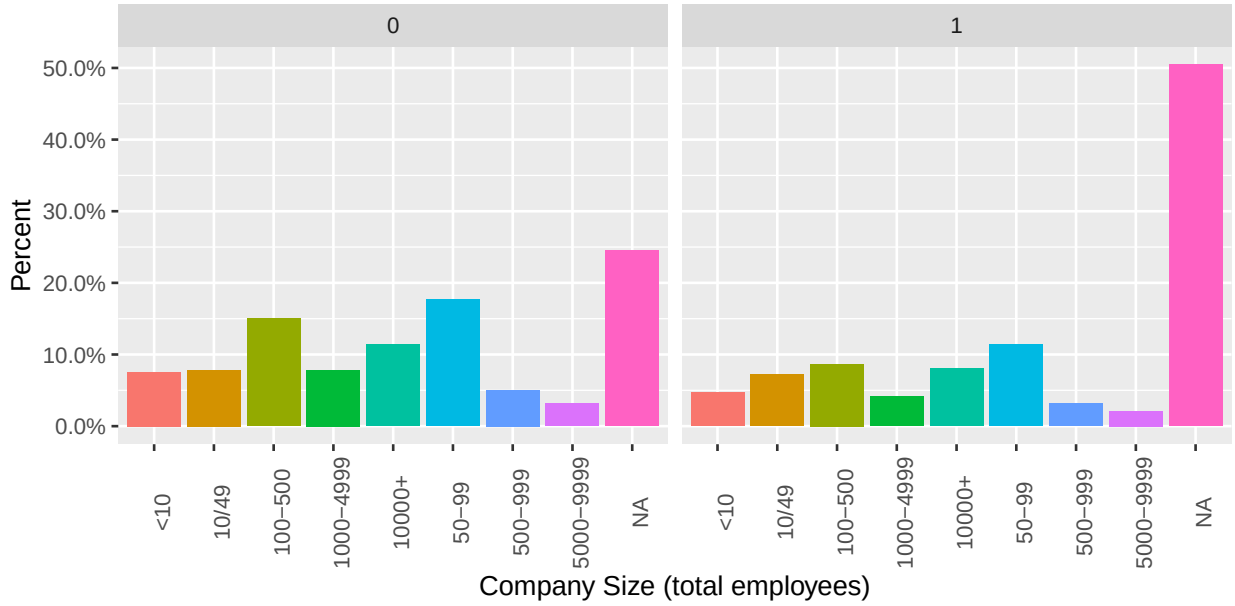


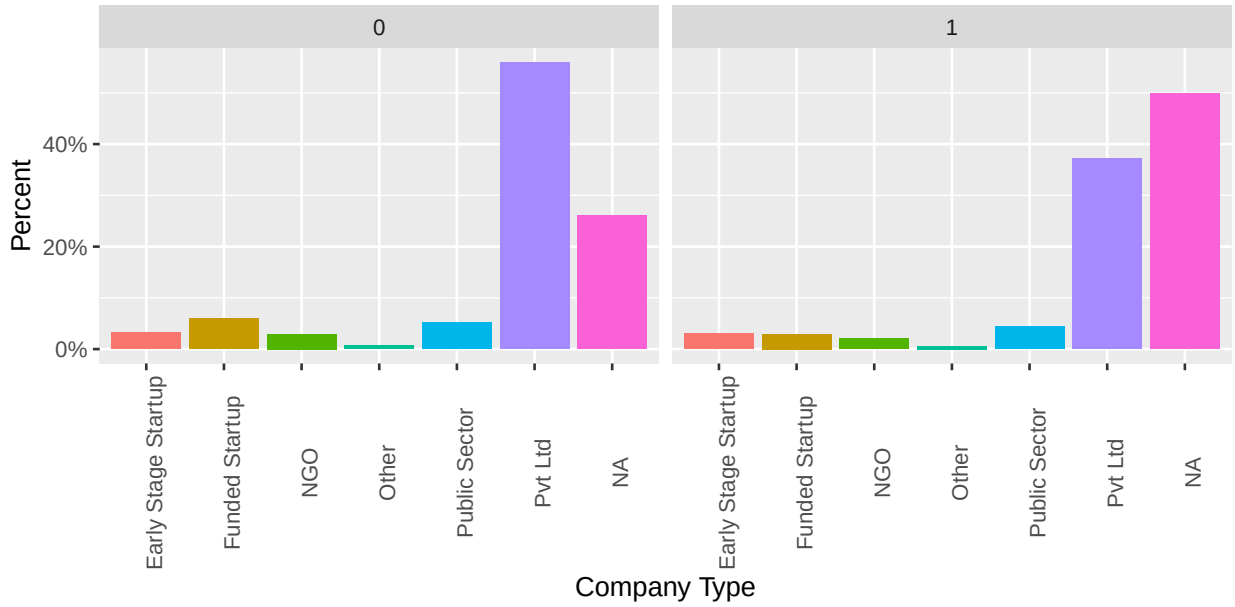
To better understand how the other features relate to the target, the desire that an employee has to look for a new job, a series of plots comparing the percentage distribution of this very features is going to be presented faceted by the target. Simply put, the distribution in the left contain data from employees that are not looking for a new job while the distribution in the right contains the employees that are looking for new positions.

When comparing the gender, there is a higher inclination of NAs with target 1. It is also noticiable that male employees consist of the majority of the dataset. Can men be concealing their gender during the polls when they desire to look for a new job?

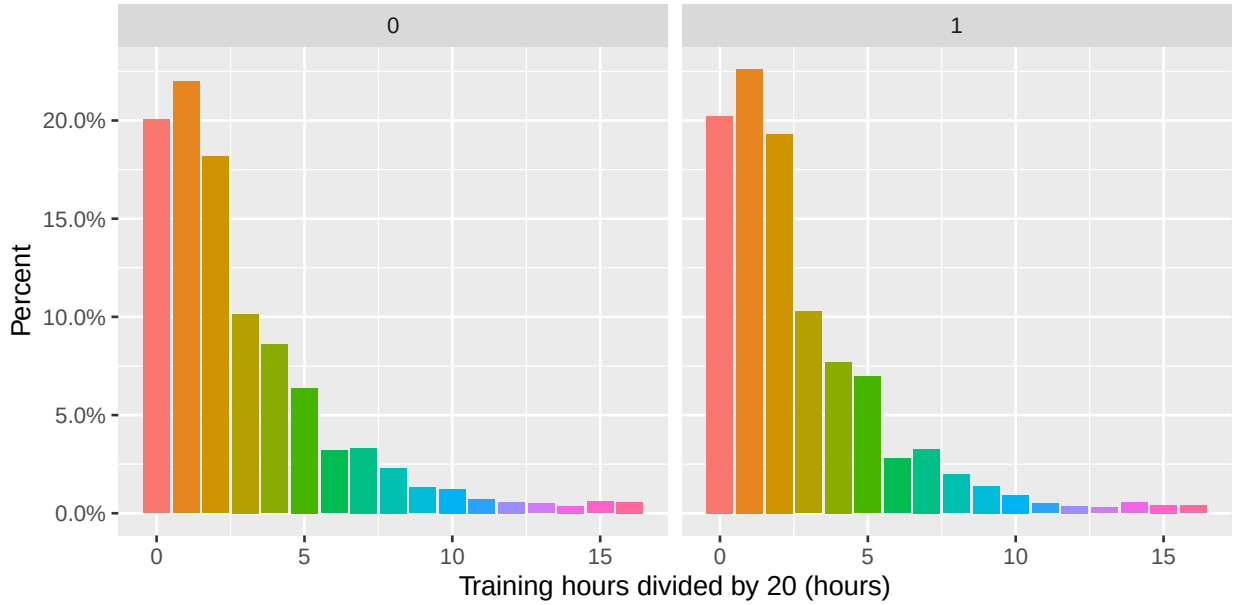


A similar behavior of higher NA values to target 1 is observed when comparing the company size and company type. The company size is the

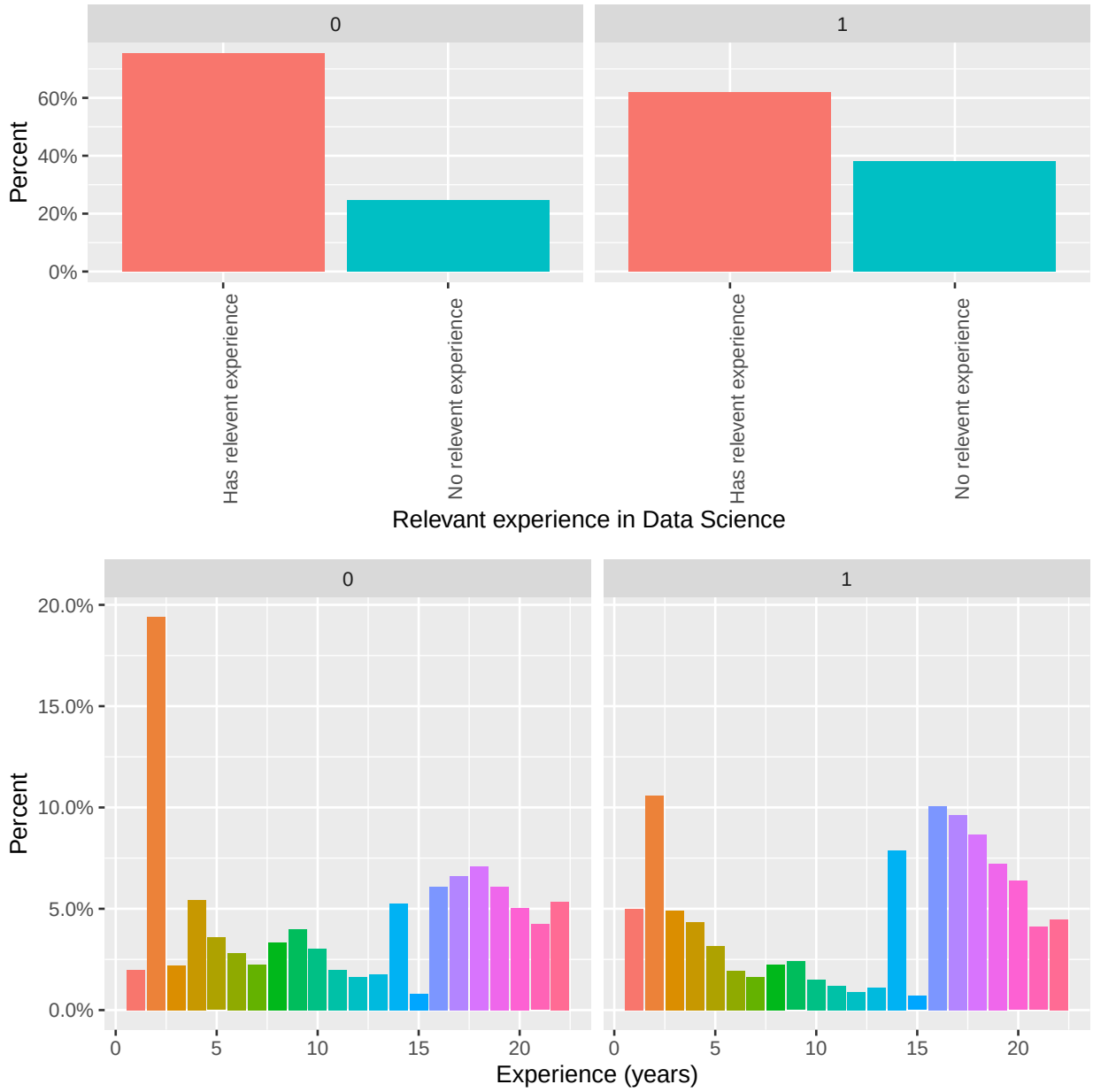




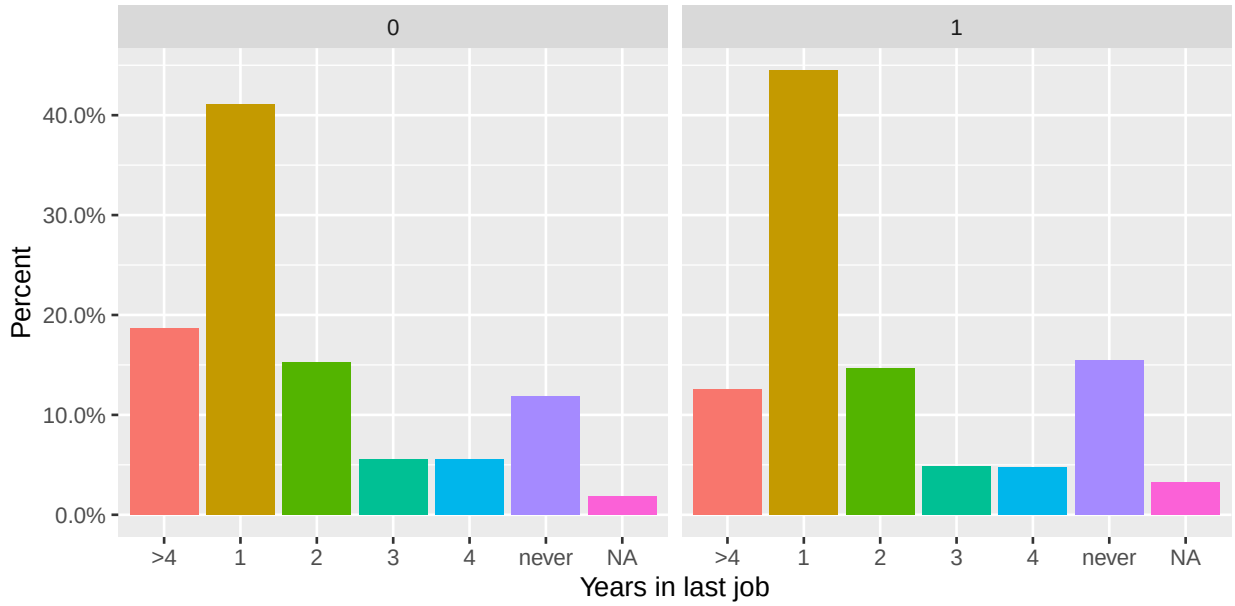
Other features, like the training hours, do not seem to interfere with the target at all as there is no substantial difference in distribution shapes.



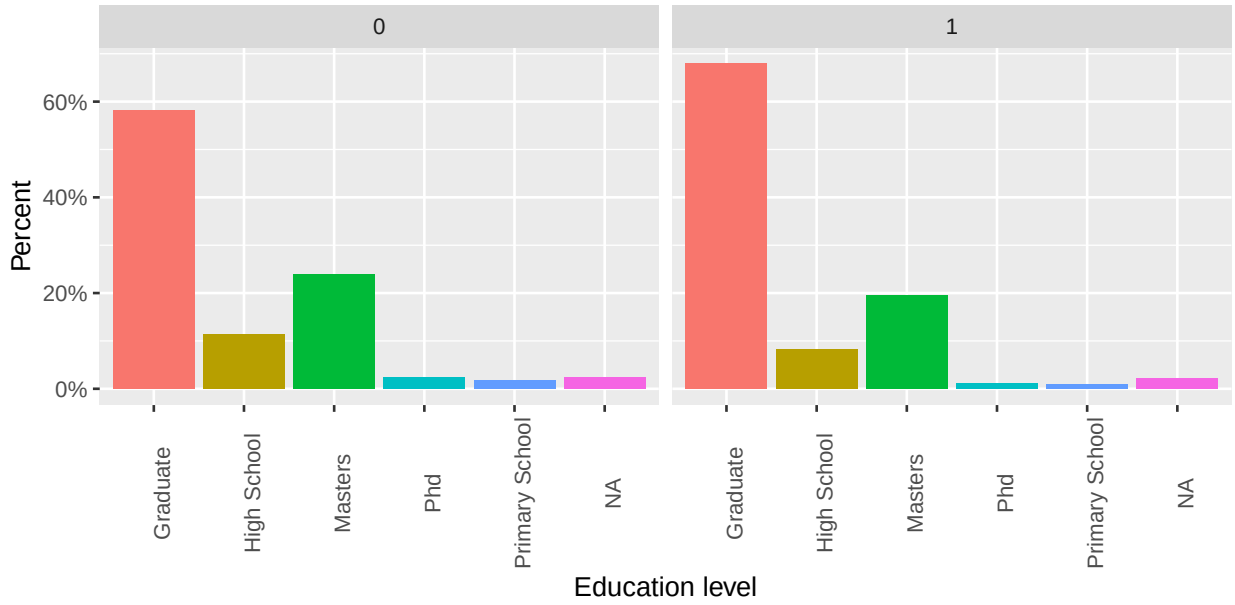
Another insight that data exploration give us is that, while employees with no relevant experience wish to work with Data Science, the ones with more than 14 years of experience are looking for new horizons. The persona that stars to build is someone with no relevant experience in Data Science but with more than 14 years in the market.

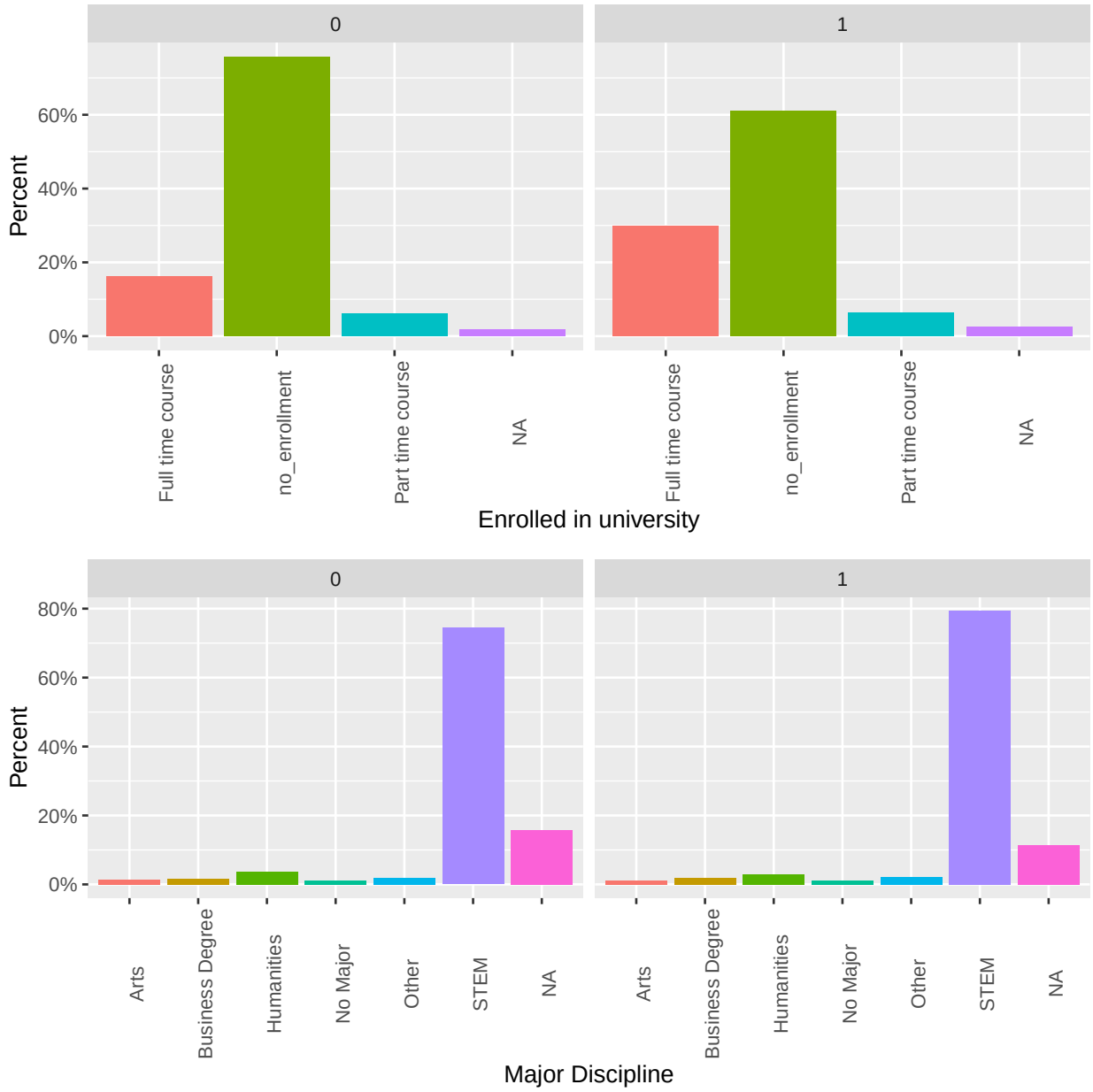


There is no great change in the last_new_job feature regarding the target.

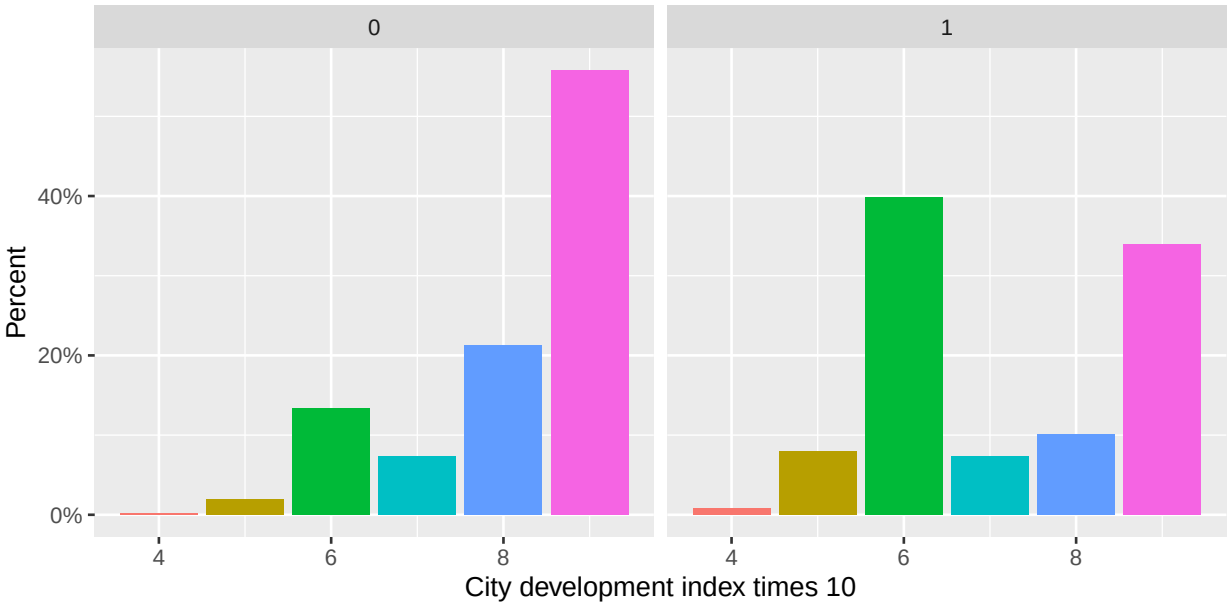


Regarding education, the fact that an employee is a graduate and enrolled in a full time course seem to alter the tendency to look for a new job. It can also be said that the fact that an employee is not enrolled in any course will reduce its tendency to look for a new job. The major discipline doesn't seem to affect much the inclination to target 1.





Finally, when analysing the demographic aspect it is clear that cities with lower development index (DI) have a higher rate of employees looking for new jobs. The target rate is way bigger than the 25% rate of the whole dataset.



City	Mean target	Entries count	HDI
city_11	0.5951417	247	0.550
city_21	0.5910437	2702	0.624
city_145	0.5873016	63	0.555
city_101	0.5733333	75	0.558
city_128	0.5652174	92	0.527
city_74	0.5192308	104	0.579
city_115	0.3888889	54	0.789
city_19	0.3781513	119	0.682
city_90	0.3147208	197	0.698
city_142	0.3018868	53	0.727

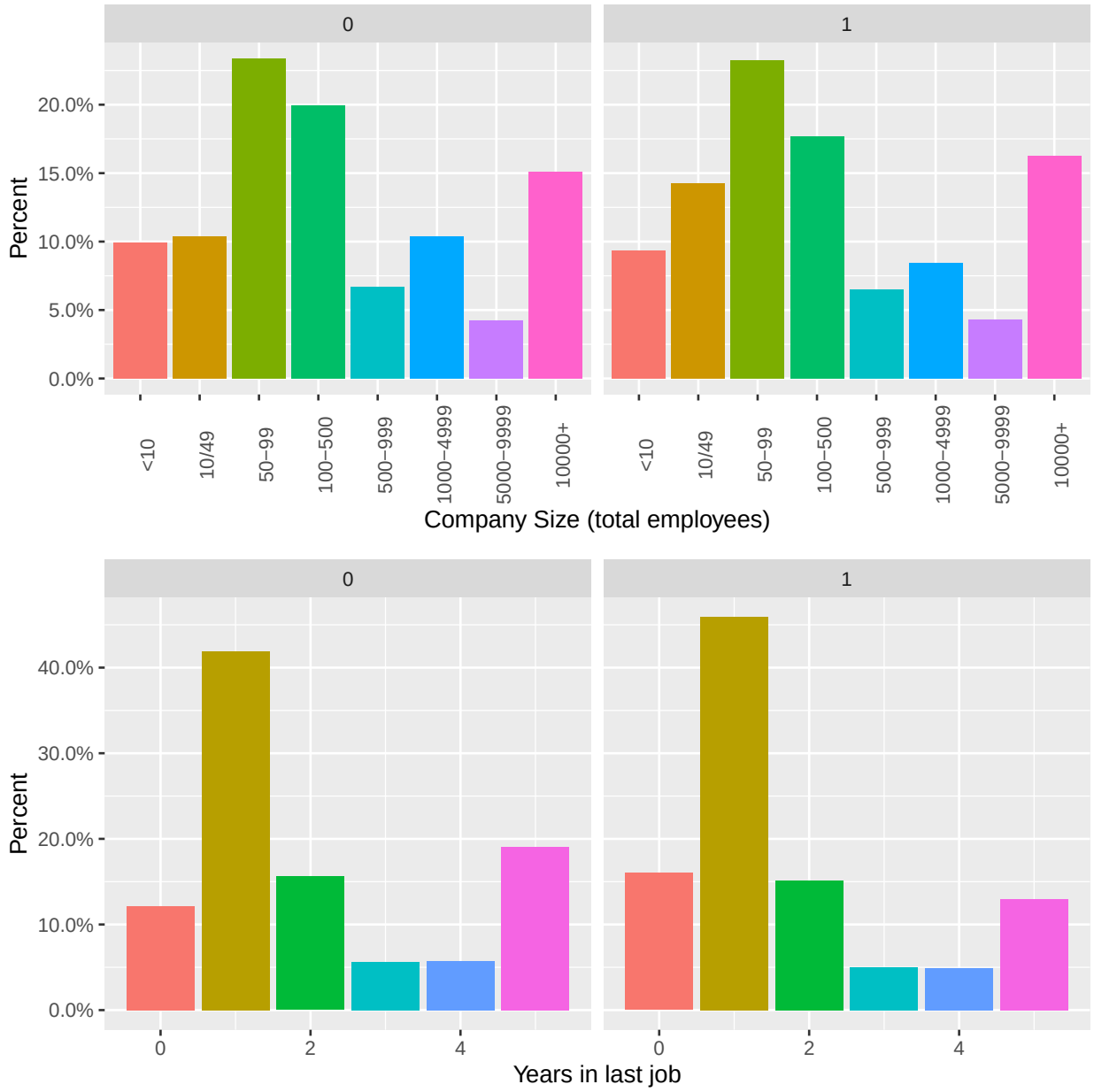
Data Wrangling

With the insight gained from data exploration, it is clear that the NA values cannot be disregarded for catagorical data from this dataset. With this in mind, categorical NAs were changed to a category called "ND" (not disclosed).

The levels of the categorical data were also factored and ordered using the insight gained. The features experience and last_new_job were wrangled and converted to numeric.

Also, a new column was added with the city target mean. Numerical NAs were omitted.

Using the new wrangling applied, new plots were created. These plots confirm that, without considering NAs, there is no significant distribution change in target with the last new job and companie size.



Data Splitting

To model the machine learning algorithms the data has been split twice. Firstly 10% of the data was randomly reserved for the final evaluation. The remaining 90% was again split, this time reserving 10% of it to testing development and 90% of it to training.

Modeling

In order to model the machine learning algorithm a table to evaluate and compare the techniques was created. It is composed of the method, accuracy, sensitivity, specificity and F1 score.

All predicted values use a 0.5 cutoff. When the prediction is higher than 0.5 the target is changed to 1, otherwise it is changed to 0.

Predicting the target rate

As a starting point, the metrics of just guessing the target with the dataset mean rate was used. Bellow the results are shown.

Method	Accuracy	Sensitivity	Specificity	F1
Guessing target rate	0.619	0.741	0.272	0.742

Linear Regression

The first linear regression considers the company_size, company_type, education_level, relevent_experience, experience, enrolled_university and city_development_index features.

A second linear regression was made. This time including also the city mean target rate. Below are the results.

Method	Accuracy	Sensitivity	Specificity	F1
Guessing target rate	0.619	0.741	0.272	0.742
Linear regression	0.774	0.955	0.258	0.862
Complete Linear regression	0.778	0.927	0.354	0.861

Although the first linear regression increased substantially the F1 score and accuracy, it had a slighter smaller specificity. Considering the city target rate seems to present a more robust approach with its higher specificity, once positive calls tend to actually have a higher chance of being positive.

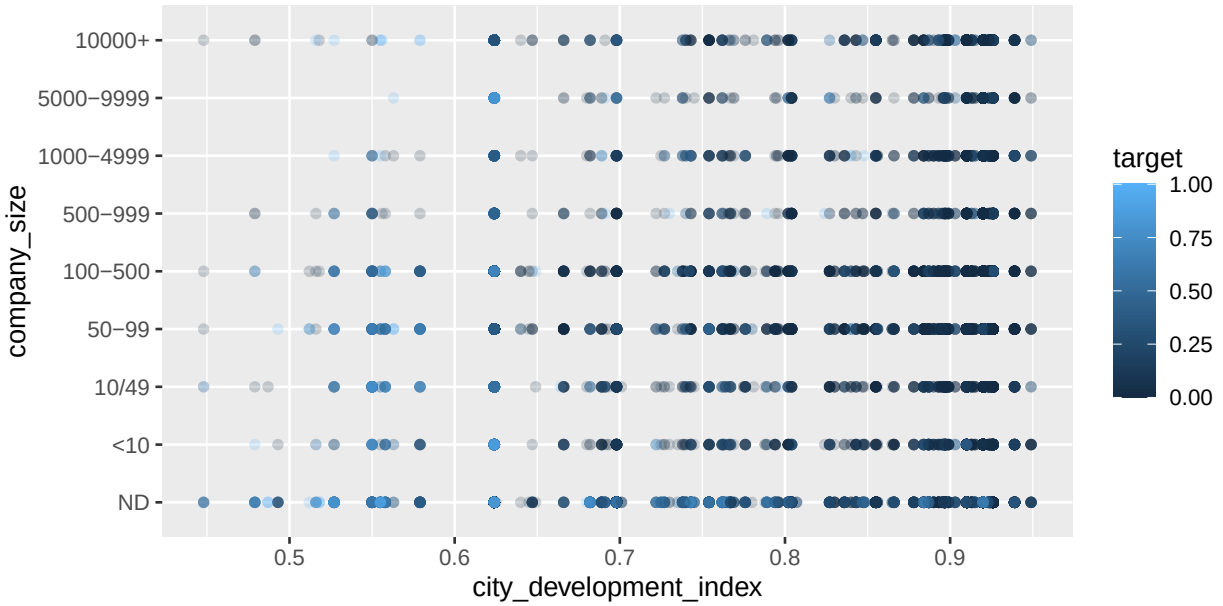
GAM Smoothing and Logistic Regression

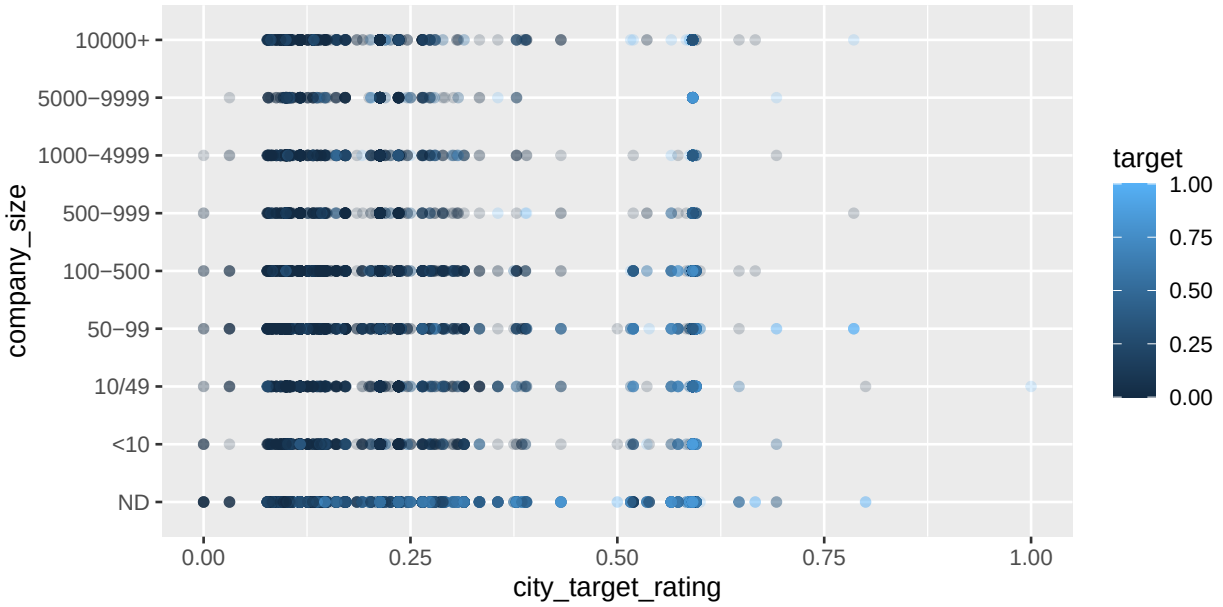
Both GAM Smoothing and Logistic regression do not improve the specificity and F1 Score in a way that the computational effort is worth.

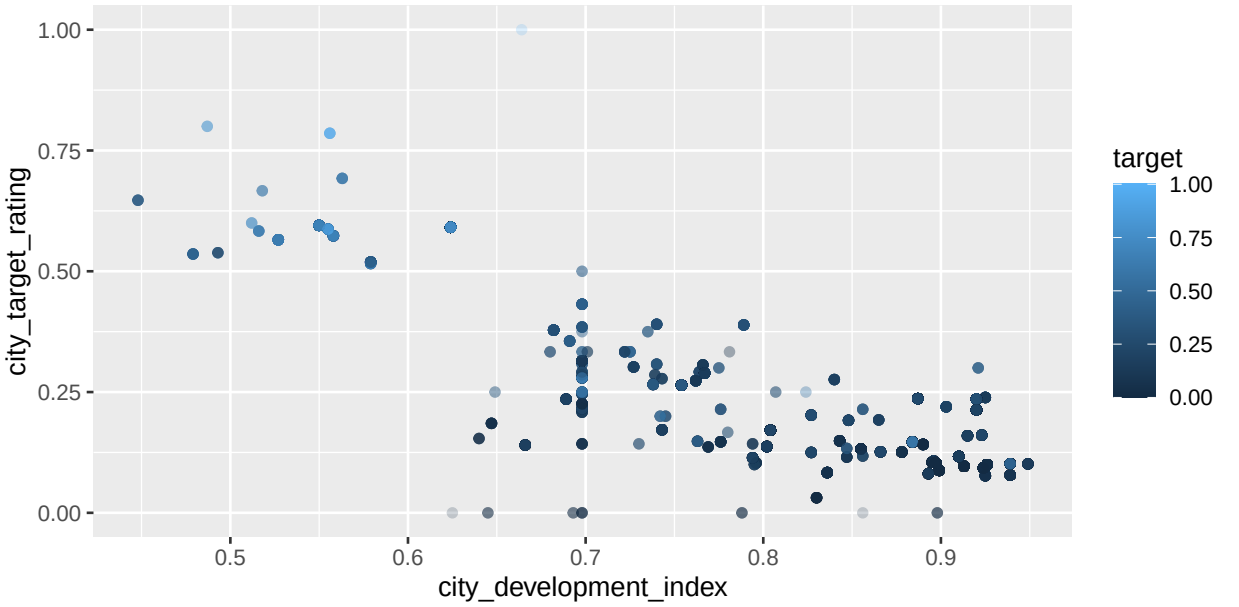
Method	Accuracy	Sensitivity	Specificity	F1
Guessing target rate	0.619	0.741	0.272	0.742
Linear regression	0.774	0.955	0.258	0.862
Complete Linear regression	0.778	0.927	0.354	0.861
GAM Smooth	0.779	0.933	0.342	0.862
Logistic Regression	0.778	0.927	0.354	0.861

Random Forest

The final approach is through a random forest algorithm with the *RBorist* package. In the graphics below it is shown that the features city_target_rating, city_development_index and company_size can be good predictors for the target, mainly when considering the NDs values.


company_size
10000+
5000−9999
1000−4999
500−999
100−500
50−99
10/49
<10
ND
0.5
0.6
0.7
0.8
0.9
city_development_index
target
1.00
0.75
0.50
0.25
0.00


company_size
10000+
5000−9999
1000−4999
500−999
100−500
50−99
10/49
<10
ND
0.00
0.25
0.50
0.75
1.00
city_target_rating
target
1.00
0.75
0.50
0.25
0.00



Four models have been created and their result can be seen at next table. There has been a great improvement in specificity. Curiously, when considering the three features at the same time the model performed worse in terms of specificity.

```
## note: only 1 unique complexity parameters in default grid. Truncating the grid to 1 .
```

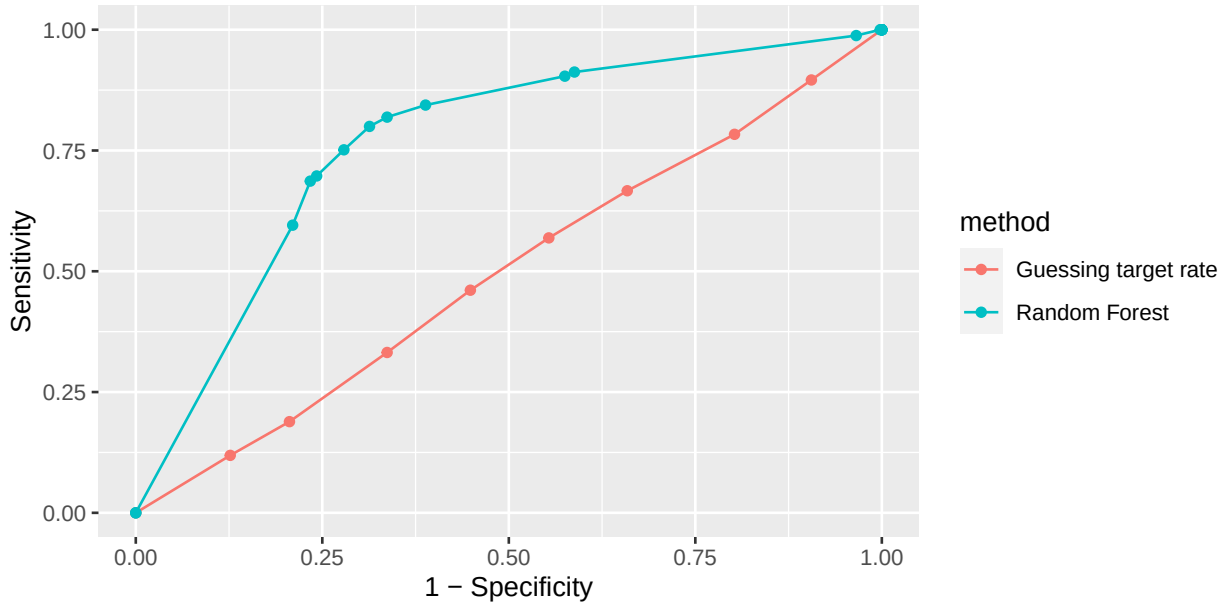
Method	Accuracy	Sensitivity	Specificity	F1
Guessing target rate	0.619	0.741	0.272	0.742
Linear regression	0.774	0.955	0.258	0.862
Complete Linear regression	0.778	0.927	0.354	0.861
GAM Smooth	0.779	0.933	0.342	0.862
Logistic Regression	0.778	0.927	0.354	0.861
Random Forest: Company Size X City DI	0.788	0.905	0.454	0.863
Random Forest: City Target Rating x Company Size	0.789	0.912	0.441	0.865
Random Forest: City Target Rating x City DI	0.787	0.906	0.447	0.863
Random Forest: City Target Rating x City DI x Company Size	0.775	0.959	0.251	0.863

Results

The Random Forest: Company Size X City DI model was chosen. It had the best specificity while also keeping a solid F1 score and accuracy. Below the final result is presented.

Method	Accuracy	Sensitivity	Specificity	F1
Random Forest: City Target Rating x Company Size	0.784	0.904	0.425	0.863

Below the ROC curve is presented. It was gererated with different cutoffs.



With the insight the ROC curve gives, it is reasonable to propose another cutoff in order to have a better specificity and move towards a point with greater projected area. A 0.4 cutoff is proposed and its result is compared to the previous modelling. The new modelling presents a much higher specificity whithout afecting much the F1 score. Both models can be used according to the desired sensitivity and specificity.

Method	Accuracy	Sensitivity	Specificity	F1
Random Forest: City Target Rating x Company Size	0.784	0.904	0.425	0.863
Random Forest with a 0.4 cutoff	0.786	0.832	0.648	0.854

Conclusion

Using Data Science a better approach was built to predict, with the curent dataset, when an employee is looking for a new job. The machine learning modelling created outperformed greatly the random guess of 25%.

Using the data presented, the persona built when thinking about one enrolle that has all the trends of someone looking for a new job in data science would be a person that

- Lives in a city with low DI
- Is full time enrolled in a university and in a graduation course
- Has no relevant experience in Data Science
- Has more than 14 years of experience
- Doesn't state the company size or type

A great limitation of the method chosen is that other features than the Company Size and City DI were not used. In data exploration it has been seen that whithout the NAs/NDs values the company size shouldn't be taken into accout. Another is that all enrollees were considered employees. In the end, the project is considering the fact that people that are looking for a new job are more concealing when given their information in polls.

A future work would be determining the primary components (PCA) that motivate one enrollee to look for a new job.