**Report**

# Business Data Mining and Visualization - Trainee Choice

# 

Presented by

Ekko, Ha Chi Chong BB903133

Natalie, Lei Un Ieng BC000522

Tiffany, Wong Heng Cheng BC000827

Mark, Che Hou In BB904782

Leo, Luo Jun Lin BB906647

# 1.Executive summary

# Our company aims to know whether the trainee will stay to work or leave the company after the training. Predicting this intention will help reduce costs and time, and to monitor the turnover rate.Therefore we use R language to establish different models including KNN, decision tree and logistic regression model to show the result by going through the process of data preparation, data visualization, modeling and evaluation. Each model has different accuracy and predicting, We need to find the highest accuracy and prediction to support our consideration. Using the graph can clearly understand the dataset visually. Lastly, we will summarize some findings in our project.

# 2.Data descriptions

In this course project, we are using a dataset which is Trainee choice. There are in total 14 variables which are enrollee\_id, city, city\_development\_index, gender, relevent\_experience, enrolled\_university, education\_level, major\_discipline, experience, company\_size, company\_type, last\_new\_job, training\_hours and target with 19158 observations.

The biggest problem of this dataset is a lot of data is missing which is 20733, one-tenth of the total. To solve the missing value, there are several ways that can be used, delete the rows which have missing value, replace with the mean for numerical data, replace the missing categorical values with the occurs most times value (mode) or create a special categorical named “missing” for missing categorical values.

In this case, we decided to use the method which is to create a special categorical value named “missing\_value” . Because most of the data of the dataset is categorical. And to use replacing the missing categorical values with the occurs most times value (mode) for logistic regression part and kNN part.

We deleted the columns that are unnecessary for later analysis, since both “enrollee\_id” and “city” are only the identity of the enrollee, there is nothing valuable information can be extracted, else, we will take a lot more time on the running process.

This is the pre-process of the data that we have done in the early step.

# 3.Goals

business goals

The project’s business goal is to monitor the turnover rate, after analyzing this dataset, it can give a prediction on which type of trainees, with what conditions are more likely to stay or leave the company by considering different kinds of variables. With this information, the business can motivate and provide incentives to potential trainees who are more willing to stay, to better monitor the turnover rate of the business without the need of close supervision and observation of the trainees.

Data mining goals

While in data mining aspect, in order to find out the most accurate result, we use three methods which are namely kNN, decision tree and logistic regression, and each go through the process of data preparation, data visualization, modeling and evaluation, to find out which category is comparatively more influential and which are less influential on the result and only takes into account the more influential ones to make interpretation and analysis, aiming to maintain a result with highest accuracy and to figure out the most accurate approach among the three.

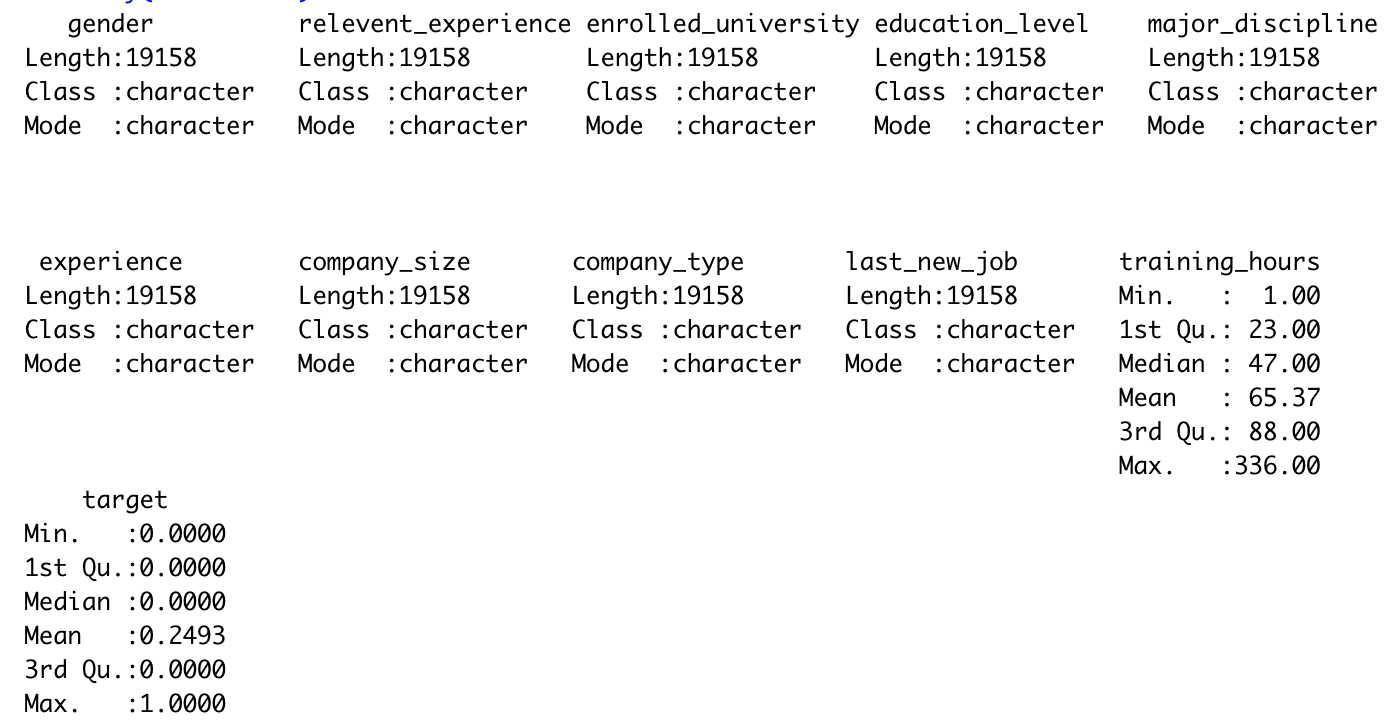
# 4.KNN

### 4.1 Data preparation

Step 1— Deleting unnecessary columns :

Before diving into the data visualization part, we need to work on the data preparation first. The first thing we did is to delete the columns that are not necessary for the analysis, so we deleted the first and second columns which are enrollee\_id and city. It is not useful for the analysis. In the column of city, it only shows city\_1, city\_2, etc and it is meaningless. Including excessive unnecessary data might make the analysis more time-consuming.

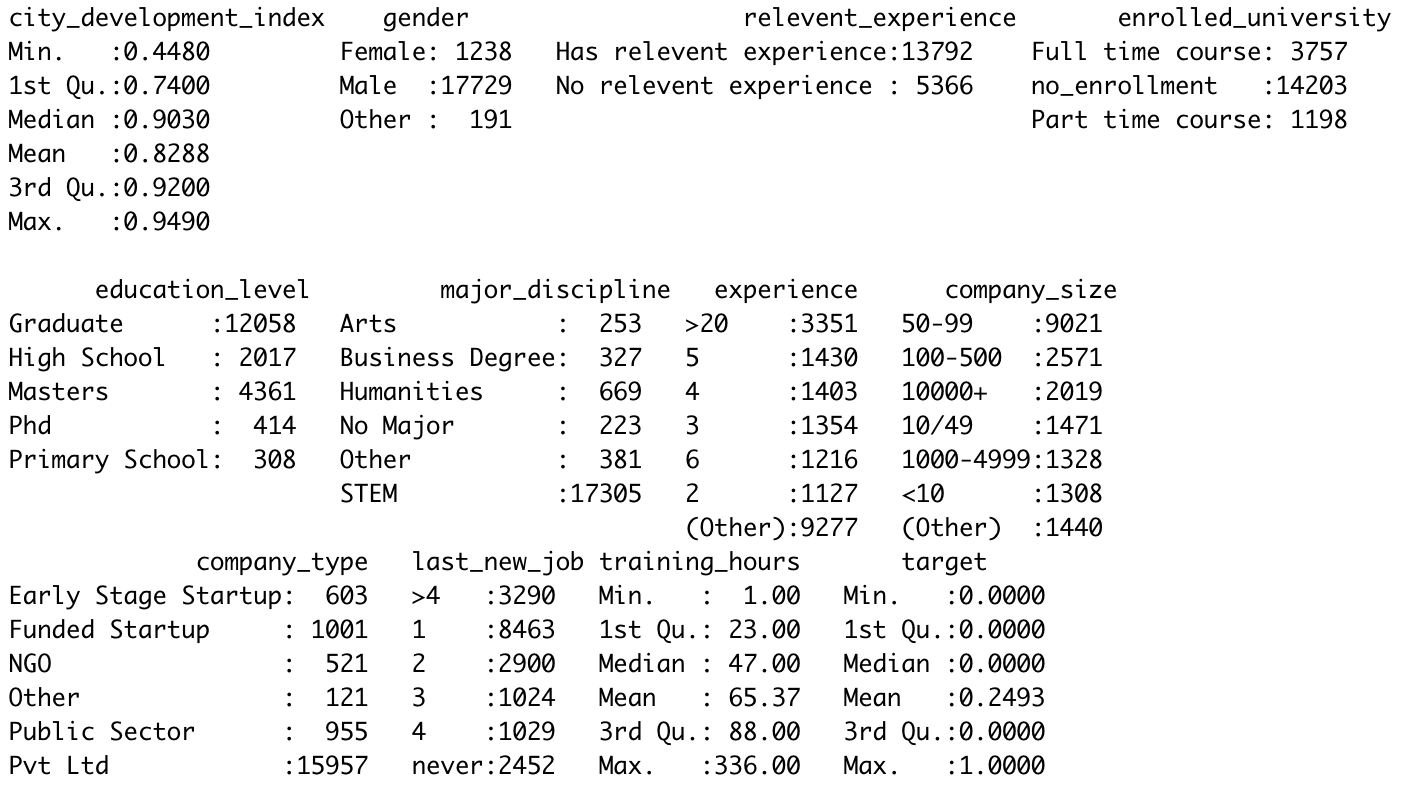
The following picture shows the summary of dataset after canceling the columns:



Step 2— Replacing missing value with mode :

As mentioned in the data description part, the largest problem of this dataset is that it has comparatively more missing value, since kNN requires only numeric value instead of categorical value, in this case, we replaced all the missing values into mode value which is the value that appears at most in all columns instead of creating a special category named “missing” for missing categorical values. After replacing, we can organize the data in a more clear way, and get the statistics from the dataset.

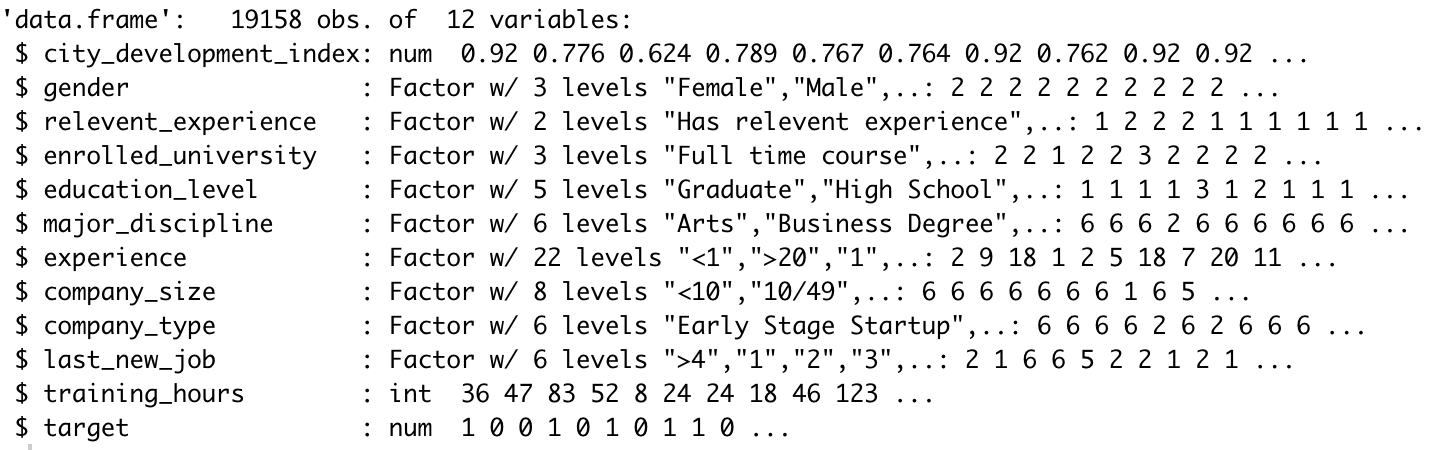
The following picture shows the summary of the dataset after modification:



Step 3— Converting categorical value to numerical value :

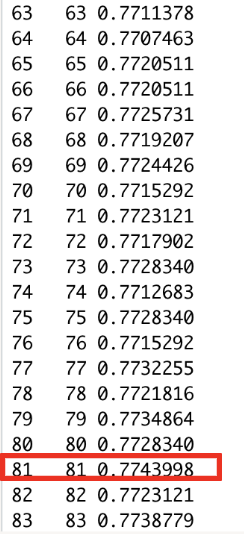
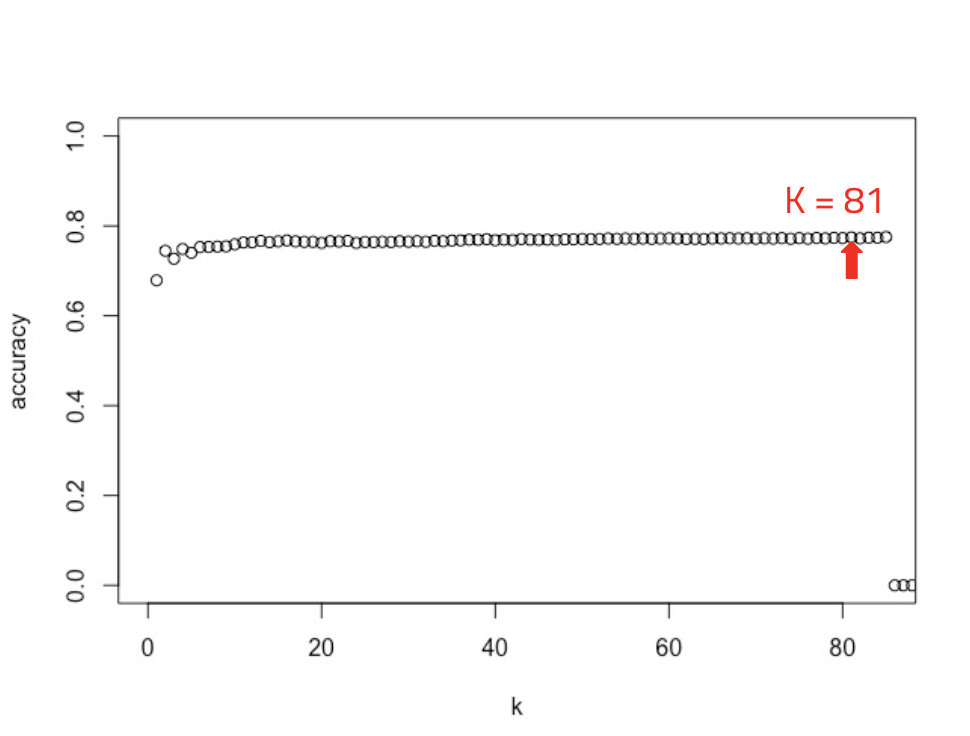
After the converting part, in order to maintain the result of kNN, the last step is to change all the categorical values into numerical values as kNN requires only numerical values to analyze. We remain numerical value like city\_development\_index, training\_hours and change all the categorical value like gender, relevent\_experience, enrolled\_university, education\_level, major\_discipline, experience, company\_size, company\_type and last\_new\_job into numerical value. After the conversion of datatype, we can see from the picture that the structure of the variables changes to factor, which simply means integer.

The following picture shows the structure of each variable:



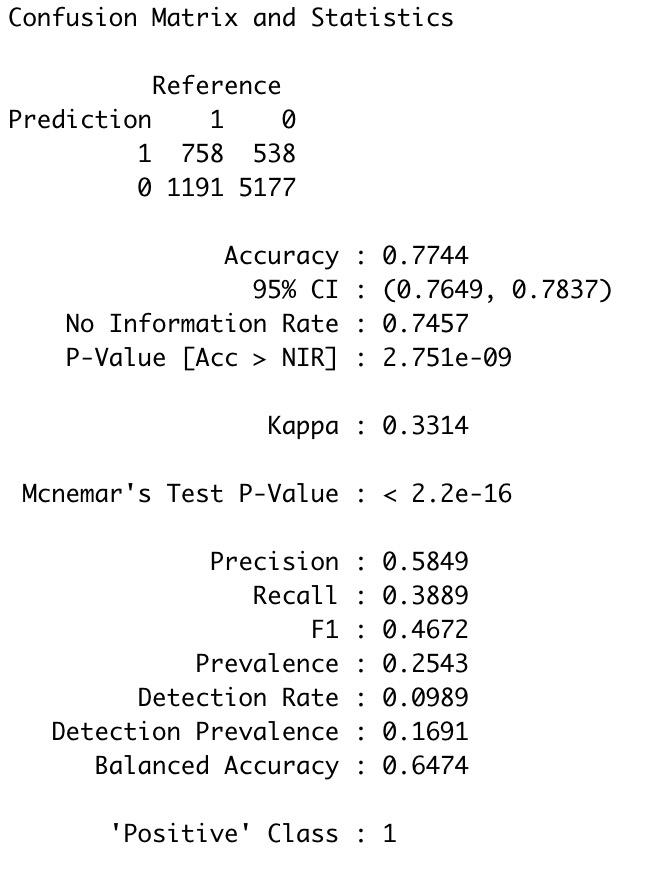
### 4.2 Visualization

With all the preparation of data, we can use knn to analyze the data. The following graph and table shows the different accuracy among the different numbers of k. We found out that the best k is in 81. To show a clear result, we demonstrate the value of k in the table on the right hand side, while the chart on the left hand side shows the k with the highest accuracy which is not obvious.



### 4.3 Modeling and evaluation

Below is the result when k is 81. We can see that the accuracy is 77.44%. And the precision and recall are 58.49% and 38.89% respectively. The accuracy is quite high but the recall is relatively low.



### 4.4 Limitations

Since KNN is a distance-based algorithm, although kNN achieves high accuracy, it is not suitable for large dataset like Trainee Choice with so many missing values as it affects the performance. It is time consuming and the system may not be affordable for such a large memory. On the other hand, there may be outliers in the dataset which we may not be aware of, we need to do a handful of preparation of data before proceeding to the knn algorithm. During our work, we have waited for 5 hours for a single code, which also affects the running of the whole project.

# 5. Logistic Regression

### 5.1 Data preparation

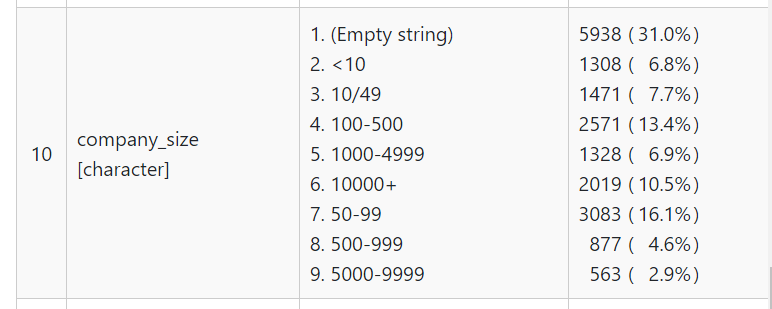
**Missing values**

Trying to preprocess the missing values in our dataset, since most of the missing values are categorical values, there appears to be two options that I can take:

1. Replace empty fields with a new category named “missing values”,
2. Replace empty fields with the category of the majority cases.

Because I’m concerned that it might be hard to make sense of the category “missing values” if I adopt the first option, so I went for the second option.

except for the one column, “company\_size”, where it appeared that the majority case is actually the empty fields. So I could choose no other option but to replace the empty fields with the “unknown” category.

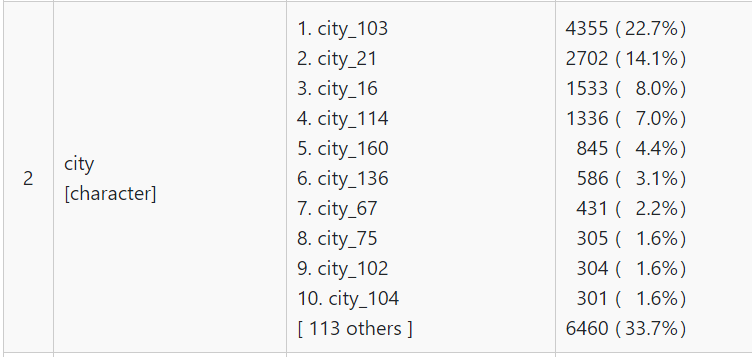


**Removed Column**

To get rid of irrelevant data, I removed two columns of this dataset: Enrollee\_id and city.

Enrollee\_id is a common column to be removed from every dataset. It’s just an identification field of each data record which has little to do with the outcome of the prediction.

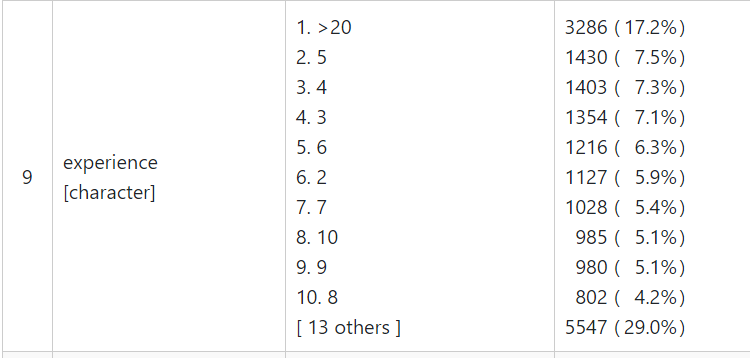
The reasons why I removed the city column are the following three:



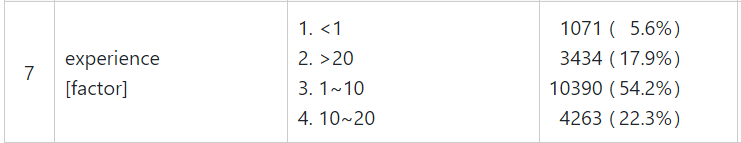
1. It has over 120 categories (branches), it may result in overfitting of our model.
2. Such a large pool of possible categories would require a lot of computation power.
3. There already exists another column called “city\_development\_index”. “City” column may have high correlation with the “city\_development\_index” column and result in redundancy.

**Data conversion**

Most of the data in the dataset already has reasonable datatype, so most remain unchanged. However there is one column namely “experience” (depicted below) that



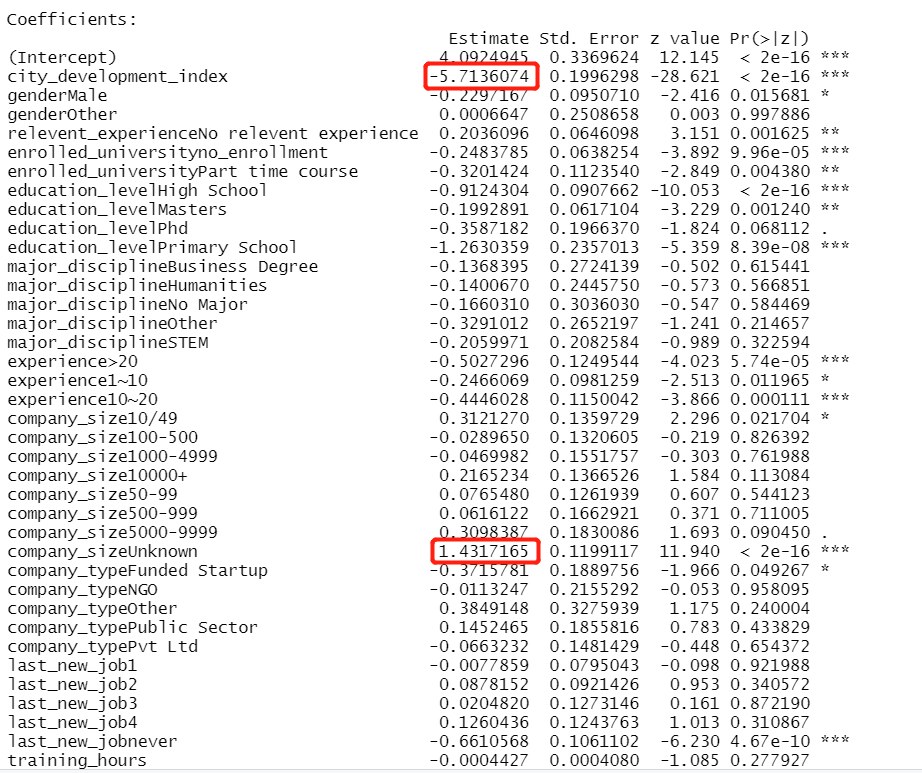
might require modifications. We can see that the distribution of the data is pretty violent - the proportion of category “>20” is significantly high, and the proportion is more or less evenly distributed among the rest. So I think it might improve model performance to compress this column data into 4 subcategories. Here is what I did with this column.



Whether or not it will actually improve the performance remains to be seen, and will be experimented later.

### 5.2 Modeling

With the above assumption, we have generated the below regression model:



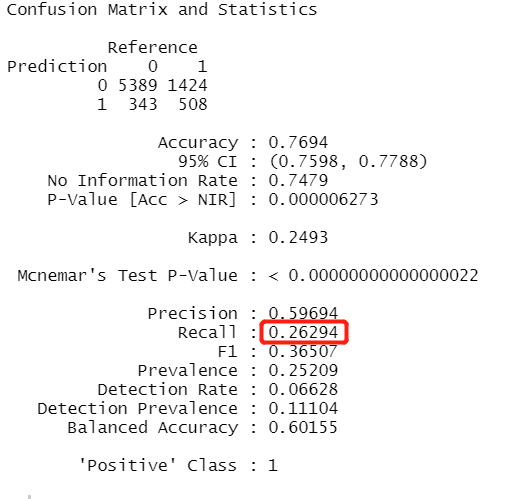
There is several interesting facts to be discovered:

1. city\_development\_index is significantly negatively correlated with the target column.

(one may want to start to look for more prospective trainees in the less developed areas)

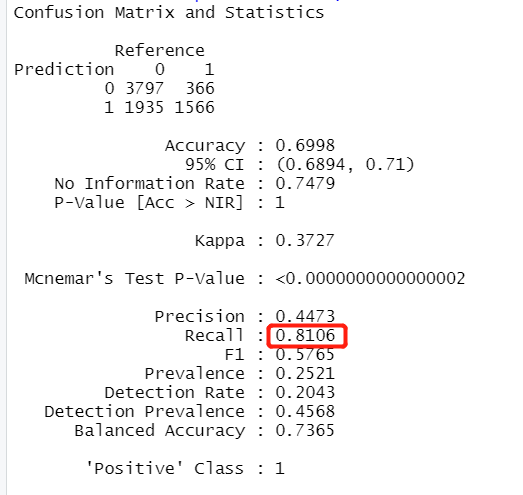
1. The new category “unknown” company actually has a pretty good correlation with the target column. It is encouraged that a deeper understanding of the dataset or further data collections is in place to make sense of it.

### 5.3 Evaluation



From the confusion matrix, we can see that although it has achieved an acceptable accuracy of 0.7694, the recall is dangerously low with only 0.263, meaning out of 100 prospective trainees, the model would only include roughly 26 of them.

But if one finds the recall unacceptable, we can adjust the cutoff to raise the recall at the expense of precision. Let’s experiment with a cutoff adjusted to 0.2.

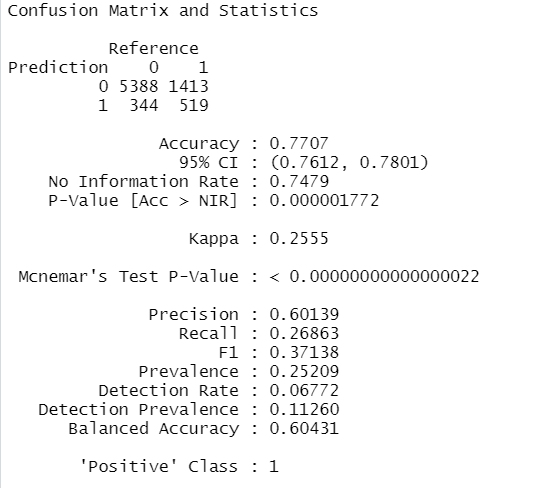


We see a significant boost in recall (from 0.263 to 0.810) with a minor drop in precision (from 0.597 to 0.447). It’s a very good deal for trading.

However, whether the company actually wants to do the trade or not really depends on the specific needs and goals of the company. More recall and less precision essentially means a larger pool of trainees with a smaller portion of them staying at the company in the end, the estimated cost associated with it is obviously higher than the one with less recall. So the company is only recommended to do this trade if they’re confident with their budget.

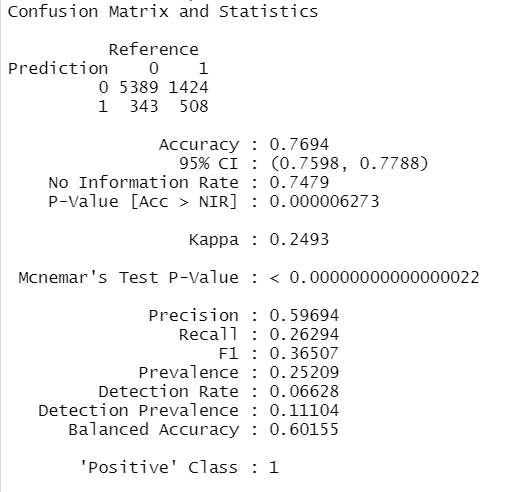
All in all, the performance of this model is pretty poor. It has mistakenly categorized most of the “1” into “0”.

Remember that I earlier compressed the experience column from over 120 categories into only 4 categories. Did that decision influence our accuracy? Can that be the reason why our model performed so poorly? So I tried building the model again using the original “experience” data categorization.



It turns out that the performance is not that different. So I finally decided to use the previous categorization of the “experience” column with only 4 categories, to avoid potential overfitting.

Final result:



### 5.4 Limitations

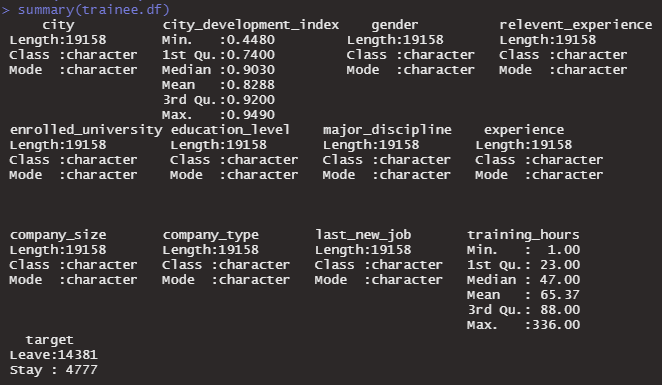
Although the model has good accuracy, it has a very low recall. And one of the largely correlated predictors is the category “unknown company size”. It’s a bit tricky for companies to use for both reasons.

# 6. Decision tree

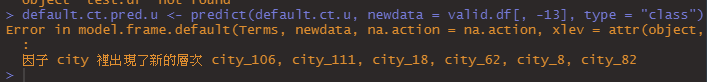
### 6.1 Data preparation

In the preparation part, I change the 0 and 1 to be “Leave” and “Stay” which are inside the target data and delete the first column which is enrollee\_id

I spend some time trying to find the methods to solve the problem of the missing value. The method that I use is to add a special categorical named “missing\_value” into the data. The reason that I choose this method is most of the data is categorical data in the dataset.



I also try to solve the problem of the data imbalance, but it always shows an error.

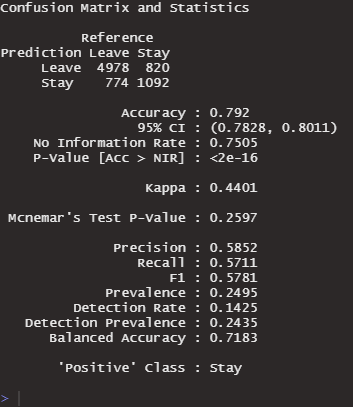


I think it is due to the data of city having too many categorical, so it is not easy to divide the data randomly into valid data and train data which group of the data can have all categorical at the same time. Due to the time limit, I can not find the seed number to divide the data successfully.

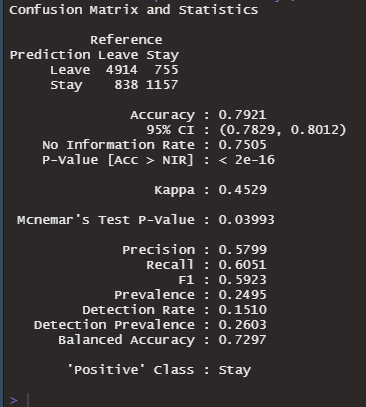
### 6.2 Modeling and evaluation

In the modeling part,first I divide the data into two groups. Choosing 60% data randomly to be training data and 40% to be validation data.

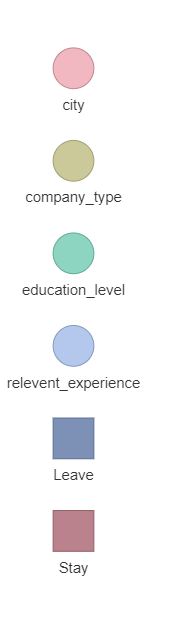
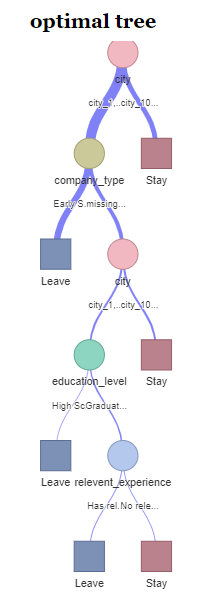
Second, I train the default tree. Using the default tree to predict the validation data and get the performance of it.

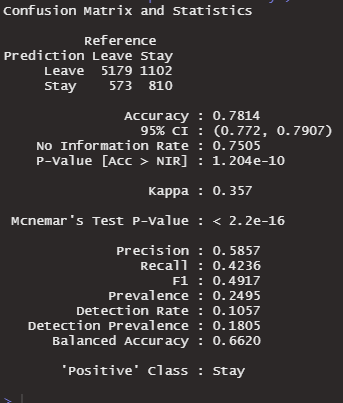


Third, I try to improve the tree model by finding the best CP value and using the new tree model to predict the validation data to get the performance.

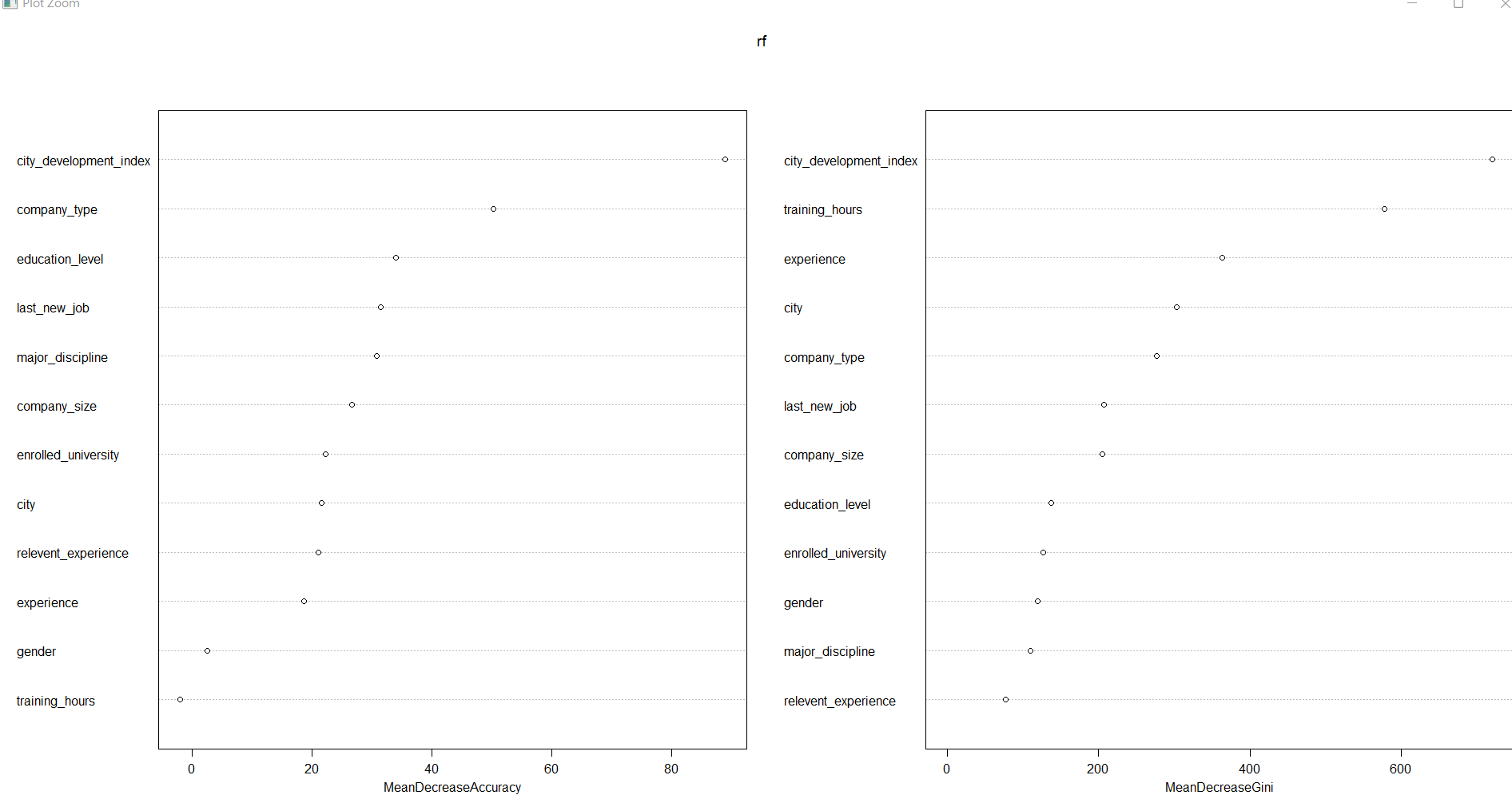


Finally, prune using the cp that minimizes error and get the decision tree diagram.

I also create a random forest model. Using a random forest model to create a diagram that can show variable importance.



As you can see, the most important variable is city\_development\_index. I also find the performance of the random forest model by predicting the validation data.

Random Forest

To increase the performance, I tried to change the missing value of company\_size to the majority and it did increase the performance.

To make a short conclusion in this part, the second tree model which uses the best CP value has the highest accuracy and the highest recall in those tree models.

### 6.3 Limitations

There are some limitations in those tree models.

First, as I had said before, the data imbalance problem is still there. I can not find a method to solve this problem before the deadline. I think it can be improved.

Second, precision and recall are not very high. I think the recall and precision also can be improved by solving the data imbalance problem.

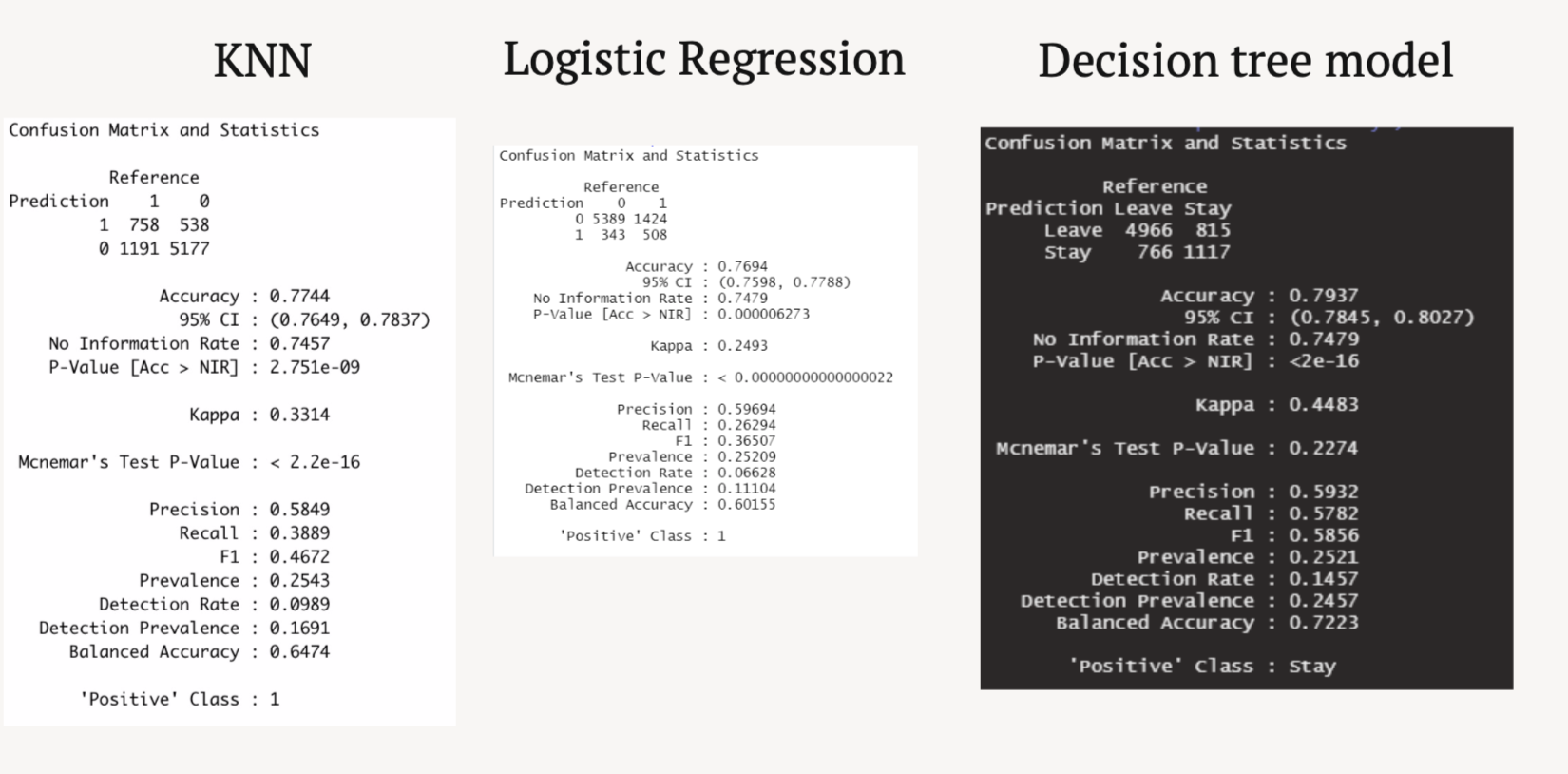
I think these two problems are the biggest limitations of these models.

# 7.Overall Limitations

The method used to process missing values in the data in the decision tree model is creating a new category named "missing value". This would result in a rather meaningless part in our result, as "missing value" doesn't mean any specific information. Even if it turns out that the branch "missing value" has a high correlation with "target", it doesn't provide any specific discovery or insight. The responsibility is now on us humans to decide ways to make sense of it.

# 8.Discussion and recommendation

To sum up, we use three methods of classification in our project .The precision of these three methods is roughly the same, but the recall of the Decision tree model is significantly higher than the other two, while achieving the highest accuracy of 0.7937. Using a decision tree model is better than other models .



In our dataset, there are total 14 columns include enrollee\_id, city, city\_development\_index, gender, relevent\_experience, enrolled\_university, education\_level, major\_discipline, experience, company\_size, company\_type, last\_new\_job, Training\_hours. Each feature will affect our precision , so we decide to delete enrollee\_id because enrollee\_id has not provided useful information.

A high turnover rate will cause the company to waste money and time. Although we understand the fact that every trainee trained must invest resources. If trainees choose to leave the company after training, it will bring high losses to the company.

To prevent potential trainees leaving the company, the most effective approach is to improve their motivation at work in the company. From the above figures, the most important variable is city\_development\_index , I think the reason why trainees want to stay in the higher urban development index is that their salary is higher, which can provide them with better development opportunities, better quality of life and higher status in society. In this case, we suggested the company provide more promotion opportunities to the potential employees, providing them with authority to manage other trainees, to increase their motivation and sense of belonging to stay in the company.

Although the accuracy of the three algorithms are quite high already, is it still not totally accurate. There is still uncertainty, there may be some trainees that will leave the company even if the analysis says he or she will stay as there are many more variables to affect the trainees’ decision on whether to stay or not. So we suggest the company not to over rely on the analytics based on the three algorithms, to prevent wasting of resources on unpotential or leaving trainees, the company can sign a yearly contract with all the trainees, for the trainees who leave the company within 1 year time, they might be required to pay for the training fee, in order to cover part of the loss by diversifying some of their financial burdens on these trainees.