HR Analytics

1. Problem Description

Your client is a large MNC and they have 9 broad verticals across the organisation. One of the problem your client is facing is around identifying the right people for promotion (only for manager position and below) and prepare them in time. Currently the process, they are following is:

1. They first identify a set of employees based on recommendations/ past performance
2. Selected employees go through the separate training and evaluation program for each vertical. These programs are based on the required skill of each vertical
3. At the end of the program, based on various factors such as training performance, KPI completion (only employees with KPIs completed greater than 60% are considered) etc., employee gets promotion

For above mentioned process, the final promotions are only announced after the evaluation and this leads to delay in transition to their new roles. Hence, company needs your help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle.

They have provided multiple attributes around Employee's past and current performance along with demographics. Now, The task is to predict whether a potential promotee at checkpoint in the test set will be promoted or not after the evaluation process.

## Dataset Description

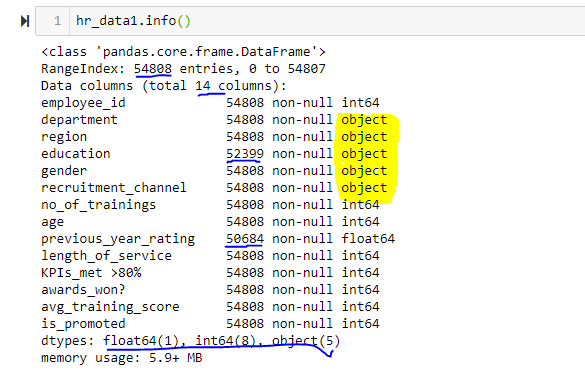
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Definition** | **Variable Type 1** | **Variable Type 2** | **Dummy Variable Need(Delete This)** |
| employee\_id | Unique ID for employee | Discrete | Independent |  |
| department | Department of employee | Nominal | Independent |  |
| region | Region of employment (unordered) | Nominal | Independent |  |
| education | Education Level | Ordinal | Independent |  |
| gender | Gender of Employee | Nominal | Independent |  |
| recruitment\_channel | Channel of recruitment for employee | Nominal | Independent |  |
| no\_of\_trainings | no of other trainings completed in previous year on soft skills, technical skills etc. | Discrete(Or Call it Numeric) | Independent |  |
| age | Age of Employee | Discrete | Independent |  |
| previous\_year\_rating | Employee Rating for the previous year | Discrete | Independent | Yes(Good,Better,Best) |
| length\_of\_service | Length of service in years | Discrete | Independent | Yes(<10,<15,>15) |
| KPIs\_met >80% | if Percent of KPIs(Key performance Indicators) >80% then 1 else 0 | Ordinal(Binary) | Independent |  |
| awards\_won? | if awards won during previous year then 1 else 0 | Ordinal | Independent |  |
| avg\_training\_score | Average score in current training evaluations | Ordinal | Independent | Yes(Good,Better,Best) |
| is\_promoted | (Target) Recommended for promotion | Ordinal(Binary) | Target Variable (Dependent) |  |

## **Data Wranglings Performed**

The data I had was in CSV format. I loaded the data through read\_csv function and performed below steps to get the feel of the data and to handle missing values.

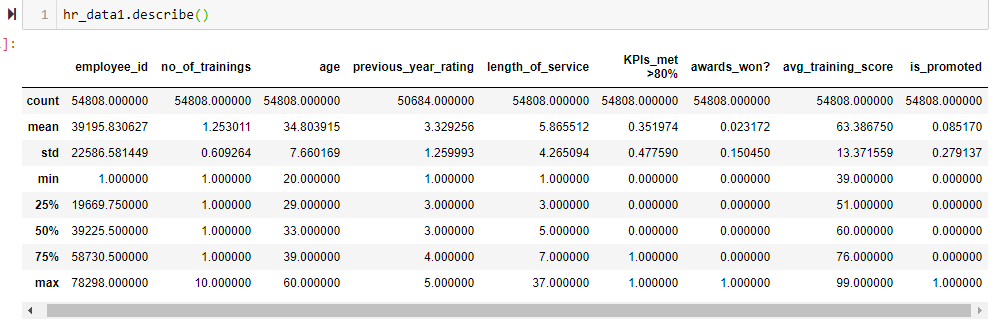
Step 1 : As soon as I loaded the data I ran info function and had below observations

1. Dataset had 14 columns and 54808 rows.
2. I noticed 2 columns (‘Education’ and ‘previous\_year\_rating’) with missing values.
3. 5 columns (highlighted in yellow below) were of type Object and may need conversion to ‘Categorical’ data type
4. Dataset had in today 3 data types ( Float, Int and Object)

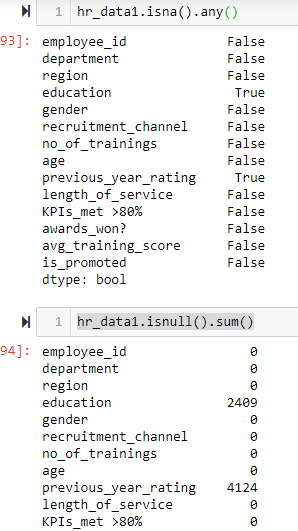


Step 2: Next I ran a describe function on the data set to get the feel on spread of the data. Below were my observations:

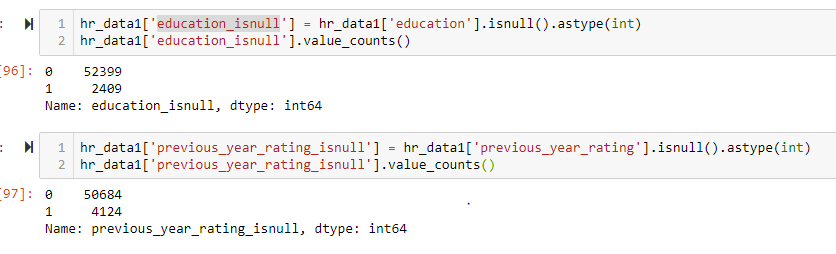
1. Employee\_ID max value was 78298 but count of employee\_id was 54808. So either dataset was not complete for many rows were deleted or employee\_id was not assigned in sequence.
2. I had an imbalanced data set as for my dependent variable( (is\_promoted) even 3rd quartile has value 0. (Also right skewed)
3. 50% or more of people had not got KPI > 80.(Right Skewed Data)
4. No\_of\_training data showed 75% of employees had not done more than 1 training and data was right skewed
5. Similarly I also ran .describe(include='all') function to find unique values for my categorical data.
6. Ang\_trainig\_score seemed almost uniform spread with mean value close to median..
7. Awards\_won showed 75% of employees had not won any awards and data was right skewed
8. Length\_of\_service shows uniform spread of data with Mean very close to median.
9. Age showed a similar case as length\_of\_service(Uniform spread).



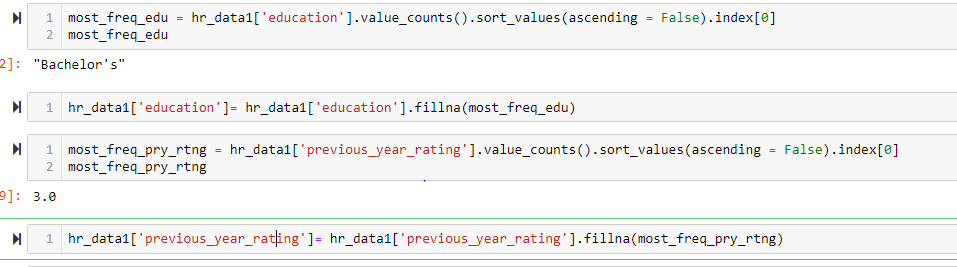
Step 3: Then I used isna and sum functions to confirm on how many columns had NULL values and count of NULL values in each column



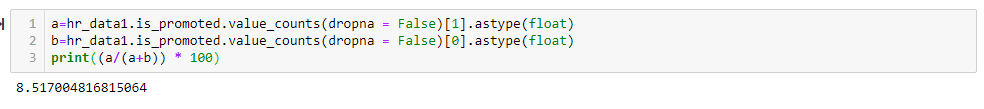
Step 4: Now before filling the NULL values for columns identified having NULLs, I created a new column with details on which row had a NULL value and which not for future reference.



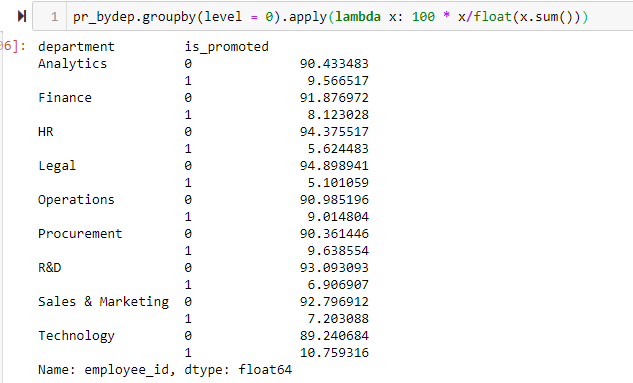
Step 5: Now for missing columns Education and previous\_year\_rating Imputed with the most frequently occurring data in the respective columns. Filled Education with ‘Bachelor’s’ and previous\_year\_rating with 3.0.

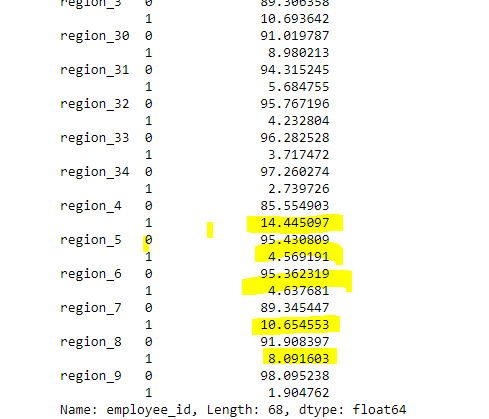


Step 6: Then to confirm the imbalanced dataset I checked the % ratio of positive values for my dependent variable is\_promoted to negative values. A 9% positive values confirmed a imbalanced dataset.



Step 7: Now to do some exploratory data analysis. I checked the % of people promoted for each department and region to check if they impact my dataset or not. Each department seemed to have uniform % or promotions but regions data did highlight an non uniformity on promotion %.



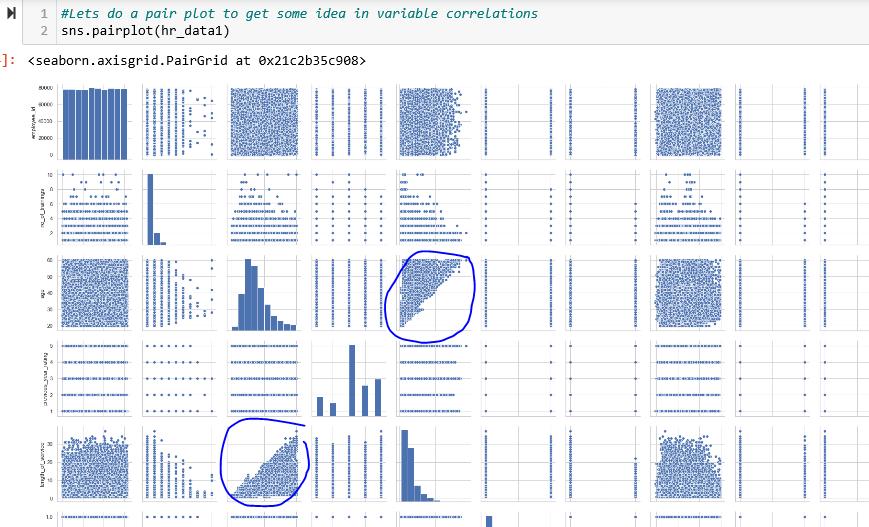


**Statistical Data Analysis**

After exploring visual descriptions of data using plots, as a next step I took a deeper dive into data to find out what is going behind them and if I can arrive on some conclusions with (un)certainty!

Considering I have a huge dataset(54K records) and armed with a few great pieces of advice from my mentor I choose the frequentist approach. Below I describe the steps I took:

Step 1: I generated a pairplot to get some idea in variable correlations. Below is the screenshot of the same

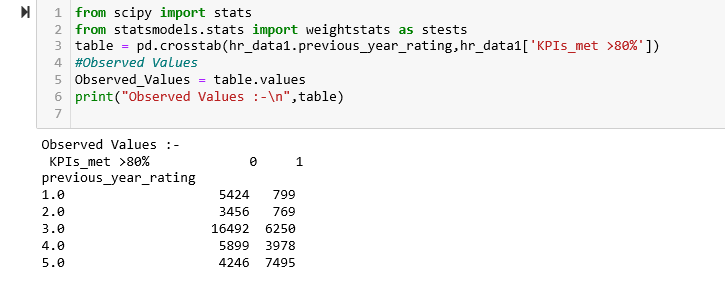


.

Step 2: Visualizing the data and looking at some other deep data analysis, I conceptualized two NULL hypotheses and performed the frequentist approach to prove/disprove the same. I will detail out one of them in the below steps.

Step3: One of correlation I noticed through visualization was between Previous year ratinings and how it influences if someone met there KPI > 80%.





Step 3: Visualizing and analyzing above I conceptualized the below NULL Hypothesis for further research:

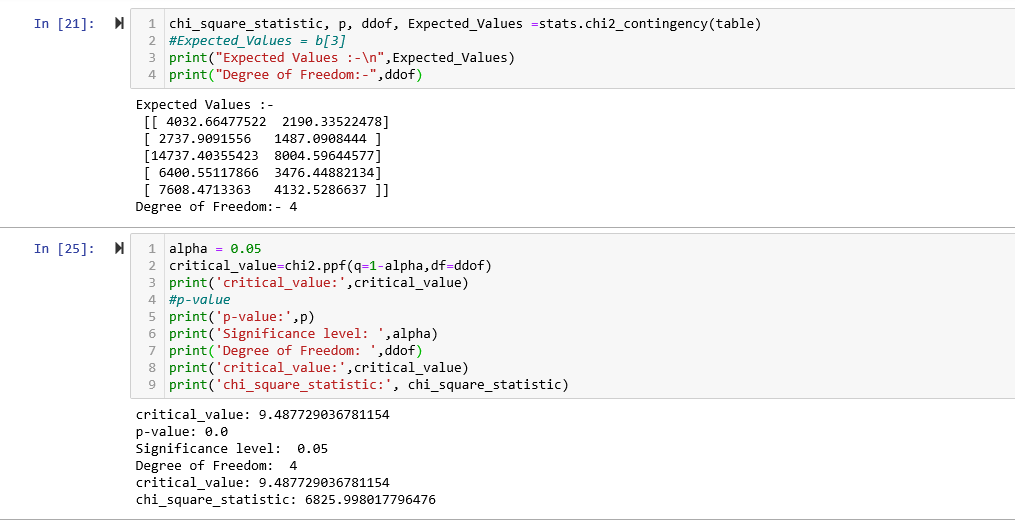
**My NULL Hypothesis:** Previous Year rating has no influence on if someone meets KPI > 80 % or in other words there is no statistical significance between both the variables.

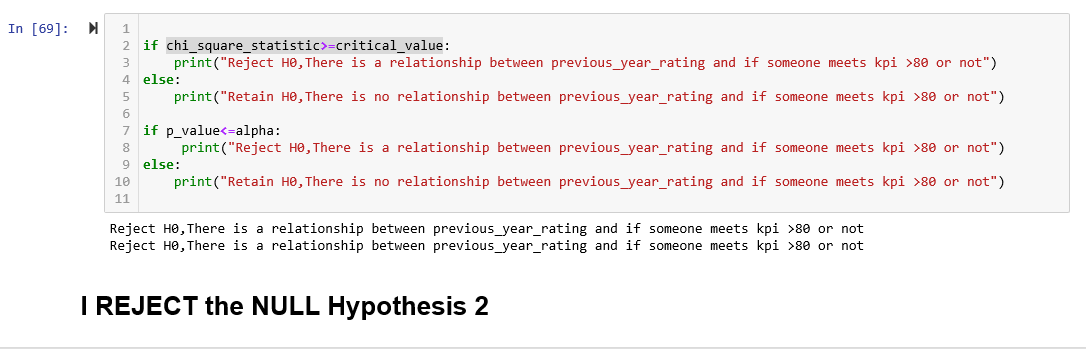
Since both the variables in consideration were **categorical variables - I used to Chi Squared Test**

Step 4: I went ahead with Significance level (Alpha) of 0.05 and would accept the NULL hypothesis with below conditions being true:

1. If my P value is greater than or equal to Alpha
2. If my chi\_square\_statistic is less than equal to critical\_value

Step 5: Below are screenshots of my results:





**Conclusions:** As we can refer above. Both my P Value(0.0) was less than my Alpha of 0.05 and Chi Squared Static (6826) was greater than my critical value (9.48). This led to me to conclude that the NULL hypothesis has to be rejected.

Or in other words the data that we visualized showing dependency between Previous Year Ration and KPI Met >80 **was not just any random observation and did had some statistical significance to the same.**

With the above observation I moved on to my next step - Machine Learning.