Machine learning and Analysis on IBM's Employee Attrition and performance

By

Sushant Ghorpade

EM 624-Informatics for Engineering Management

Instructor: Dr.Carlo Lipizzi

Introduction:

With recent pandemic, we have seen its effect on businesses and services. May people have lost their jobs since companies had to layoff employees due to current reduction in demand. Also there has been serious deductions in salaries. But comparing to normal situation, were people tend to leave their current job due to various factors such as lack of growth, being overworked, lack in recognition etc. From organizations point of view this is import since hiring new employees consumes both money and time, hence they look to retain talent. Using the dataset, we are going to analyze factors that may lead to attrition within organization.

Business Understanding

In the realm of human resources, attrition is defined as both the voluntary and involuntary reduction of a company's workforce through deaths, employee retirements, transfers, resignations, and terminations. While some attrition is to be expected in normal business operations, a high level of reduction can lead to problems and a lack of manpower.

Some of the ways human resources professionals do their part to keep top-performing employees happy and attrition rates low is design and implement company compensation programs, motivation systems and a company culture. Besides retaining top performing employees, business owners try to keep their attrition rates as low as possible to keep from having to spend money on advertising for, hiring, training and completing paperwork for new employees.

The key to success in any organization is attracting and retaining top talent. As an HR analyst one of the key tasks is to determine which factors keep employees at the company and which prompt others to leave. Given in the data is a set of data points on the employees who are either currently working within the company or have resigned. The objective is to identify and improve these factors to prevent loss of good people.

Data Understanding

For the project, we are using IMB attrition data. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset is the link to dataset I took from Kaggle. There are 1470 rows and 35 columns of which 9 are categorical data while 26 are integer data type. The dataset using was artificially created by IBM itself for to understand and learn from situation. To get a better look at our variables, they have been listed below. The column names are self-explanatory.

1. AGE Numerical Value

2. ATTRITION Employee leaving the company (0=no, 1=yes)

3. BUSINESS TRAVEL (1=No Travel, 2=Travel Frequently, 3=Travel Rarely)

4. DAILY RATE Numerical Value - Salary Level

5. DEPARTMENT (1=HR, 2=R&D, 3=Sales)

6. DISTANCE FROM HOME Numerical Value - THE DISTANCE FROM WORK TO HOME

7. EDUCATION Numerical Value. (1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor')

8. EDUCATION FIELD (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TECHNICAL)

9. EMPLOYEE COUNT Numerical Value

10. EMPLOYEE NUMBER Numerical Value - EMPLOYEE ID

11. ENVIRONMENT SATISFACTION Numerical Value - SATISFACTION WITH THE ENVIRONMENT (1 'Low' 2 'Medium' 3 'High' 4 'Very High')

12. GENDER (1=FEMALE, 2=MALE)

13. HOURLY RATE Numerical Value - HOURLY SALARY

14. JOB INVOLVEMENT Numerical Value - JOB INVOLVEMENT (1 'Low' 2 'Medium' 3 'High' 4 'Very High')

15. JOB LEVEL Numerical Value - LEVEL OF JOB

16. JOB ROLE (1=HR REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= RESEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIVE, 9= SALES REPRESENTATIVE)

17. JOB SATISFACTION Numerical Value - SATISFACTION WITH THE JOB (1 'Low' 2 'Medium' 3 'High' 4 'Very High')

18. MARITAL STATUS (1=DIVORCED, 2=MARRIED, 3=SINGLE)

19. MONTHLY INCOME Numerical Value - MONTHLY SALARY

20. MONTHLY RATE Numerical Value - MONTHLY RATE 21. NUMCOMPANIES WORKED Numerical Value - NO. OF COMPANIES WORKED AT

22. OVER 18 (1=YES, 2=NO)

23. OVERTIME (1=NO, 2=YES) 24. PERCENT SALARY HIKE Numerical Value - PERCENTAGE INCREASE IN SALARY

25. PERFORMANCE RATING Numerical Value - PERFORMANCE RATING

26. RELATIONS SATISFACTION Numerical Value - RELATIONS SATISFACTION

27. STANDARD HOURS Numerical Value - STANDARD HOURS

28. STOCK OPTIONS LEVEL Numerical Value - STOCK OPTIONS (Higher the number, the more stock option an employee has)

29. TOTAL WORKING YEARS Numerical Value - TOTAL YEARS WORKED

30. TRAINING TIMES LAST YEAR Numerical Value - HOURS SPENT TRAINING

31. WORK LIFE BALANCE Numerical Value - TIME SPENT BETWEEN WORK AND OUTSIDE

32. YEARS AT COMPANY Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPANY

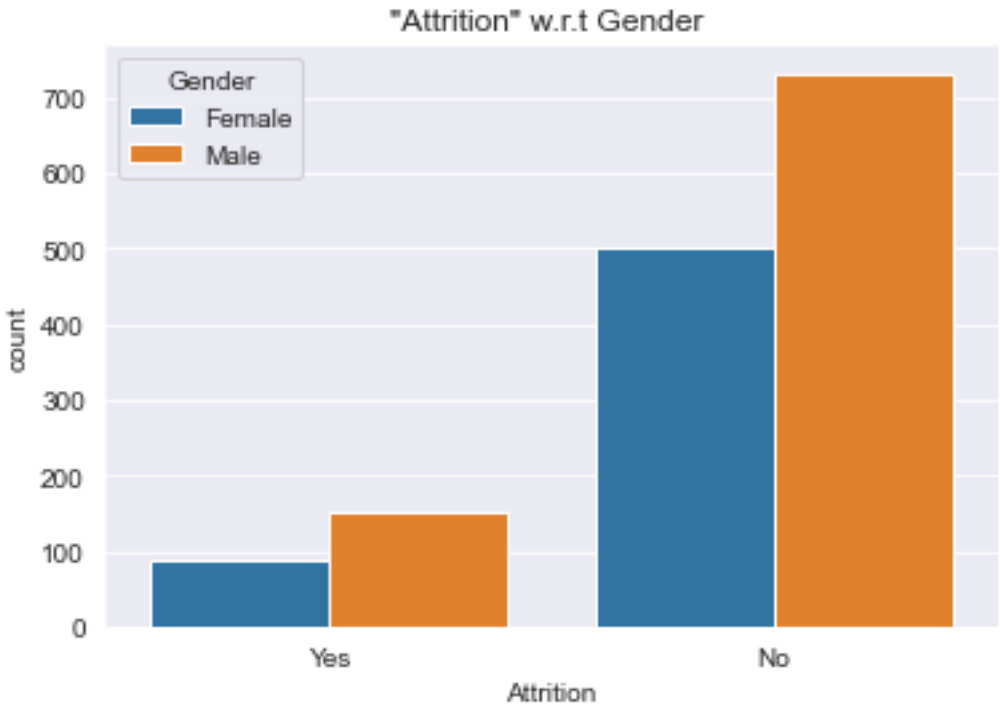
33. YEARS IN CURRENT ROLE Numerical Value -YEARS IN CURRENT ROLE

34. YEARS SINCE LAST PROMOTION Numerical Value - LAST PROMOTION

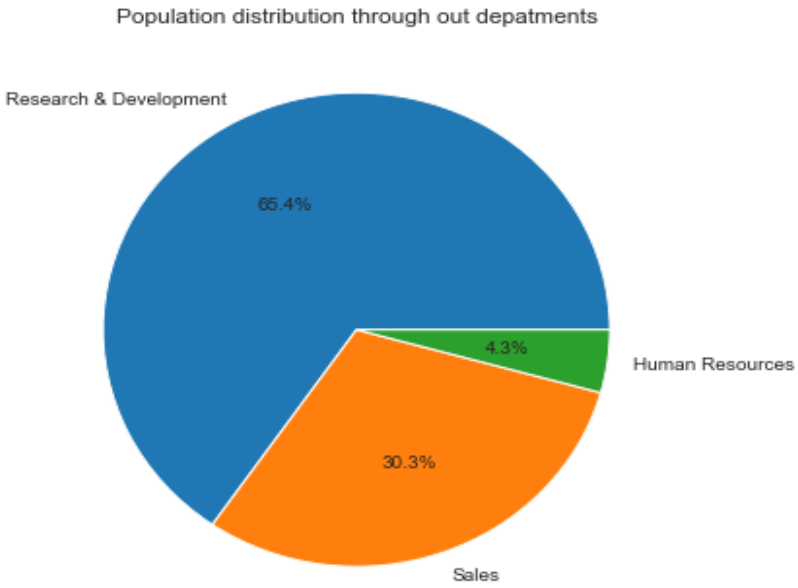
35. YEARS WITH CURRENT MANAGER Numerical Value - YEARS SPENT WITH CURRENT MANAGER

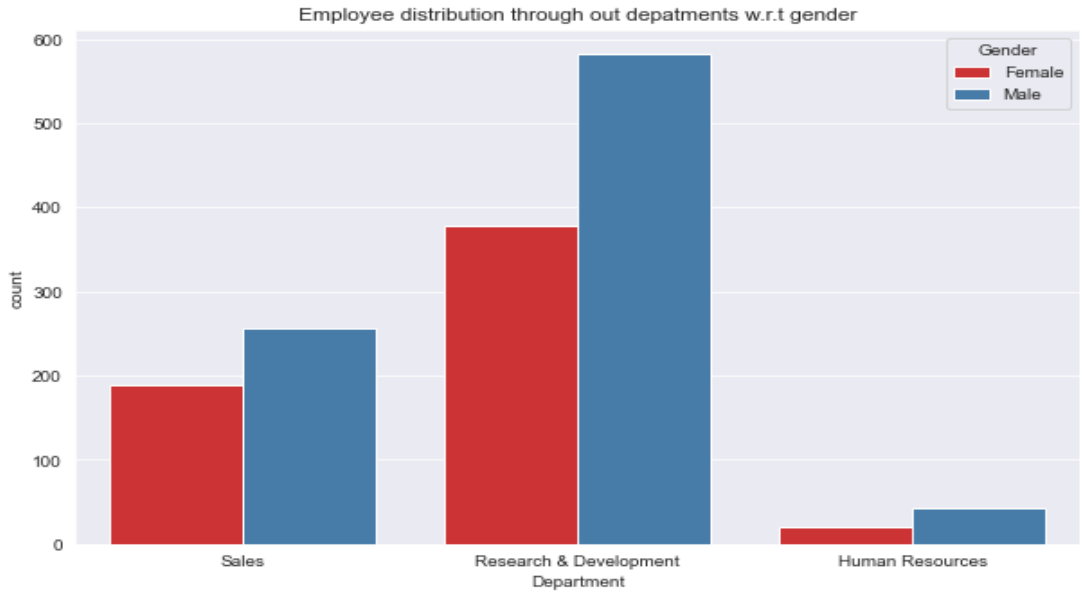
Below we have done exploratory data analysis using data visualization on few of our reverent variables

1. ***Attrition distribution with respect to gender:***
    a. Looking at the graph below we can say there are more of males as compared to females within organization
    b. Although, there is not much difference between both when we compare attrition percentage between both genders

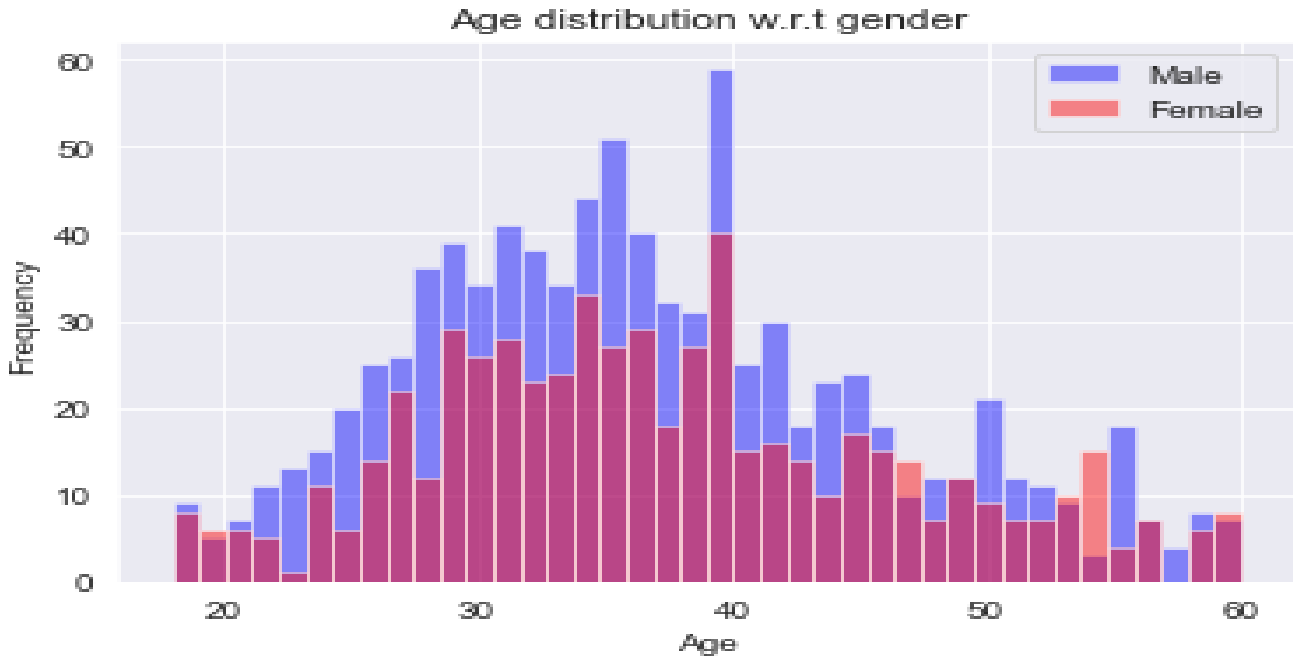


2. ***Employee distribution throughout the departments and gender***
   a. By looking at pie chart we can say research and development has 65.4% of people working, 30.3% of sales and Human resources as 4.3%
   b. Research and development have heights employees while sales human resources have lowest





3. ***Histogram on age with respect to gender:***
   c. By looking at the histogram we can say there isn't much difference between both genders when it comes to age distribution
   d. The mean for males and females is approximately 36 and 37 respectively. Though the minimum and maximum are 18 and 60 for both
   e. The standard deviation between age of males is 9.2 while for females it's 9.

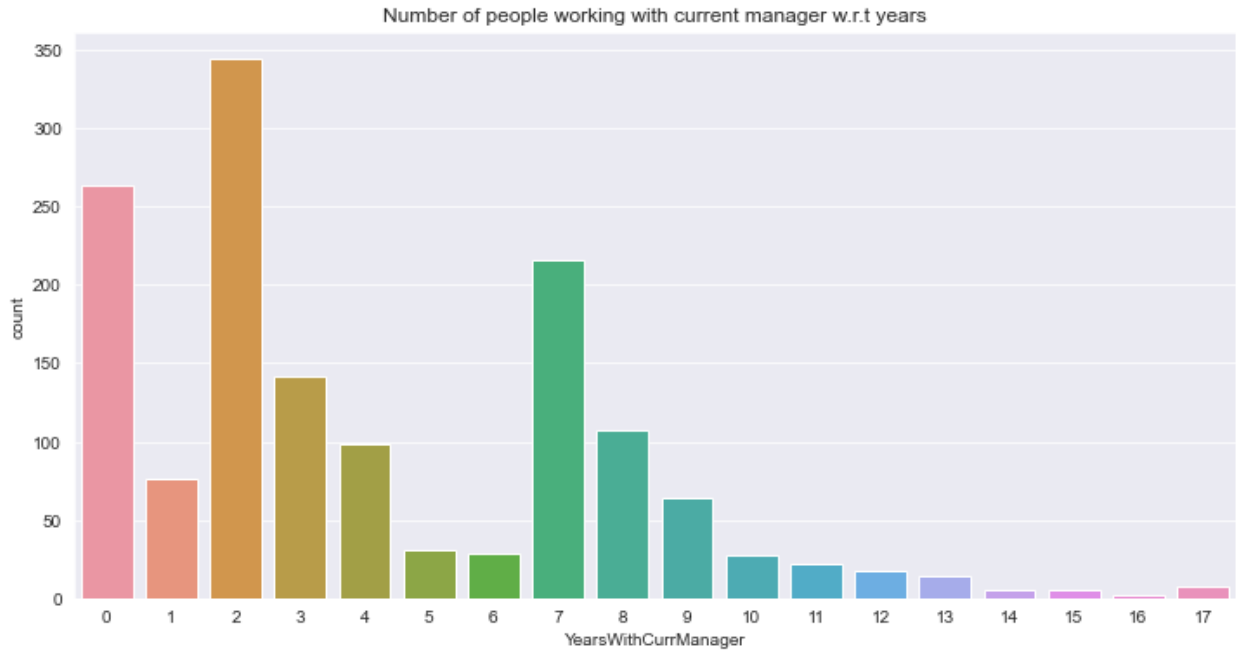


4. ***Relationship between employee income and age:***
   a. We plot monthly income of employees verse employee age
   b. It is seen that there is a positive linear relationship between both variables for both genders in all departments

c. One thing to observe it that, females are paid more that males in human resource department as compared to other two departments

5. *Manager and work duration*

a. The graph below depicts on how many years employees have worked under the same manager
b. We can say not many people work more than 2 year under same manager, a lot of them change department or job in less than a year
c. Although median service years under same manager would be approximately 4 years
d. While people have still served up to 17 years maximum



Data Preparation

Before we start, we need to import packages used for data preparation, visualization and machine learning. Below are packages we shall be using for it.

1. Pandas
2. Matplotlib
3. Seaborn
4. Scikit learn

With the help of pandas library, we shall import our dataset to environment. Once imported we check for its first five rows using ".head()" function, just to have glace of our variables. Use ".isnull().sum()" to check total missing values in each column although we do not have any. Next, use the ".info()" to get information on data type of each variable. We also use " .describe()" to get basic statistics on each integer type data like min, max, medians and quartiles.

Once this is done, we need to see correlation between variables. Checking for each is difficult so we create a correlation plot using heatmap from seaborn. We can use the color scale to determine correlation between variables. On analyzing we can see there is not much correlation between the variables. Correlation is only created for the variables which are integer in nature. To get better results we need to convert categorical into integer type. Categorical columns having two unique values can be replaced by 0 and 1 while for other we are going to create dummy variables A dummy variable (aka, an indicator variable) is a numeric variable that represents categorical data, the 9 columns in our dataset Technically, dummy variables are dichotomous, quantitative variables.

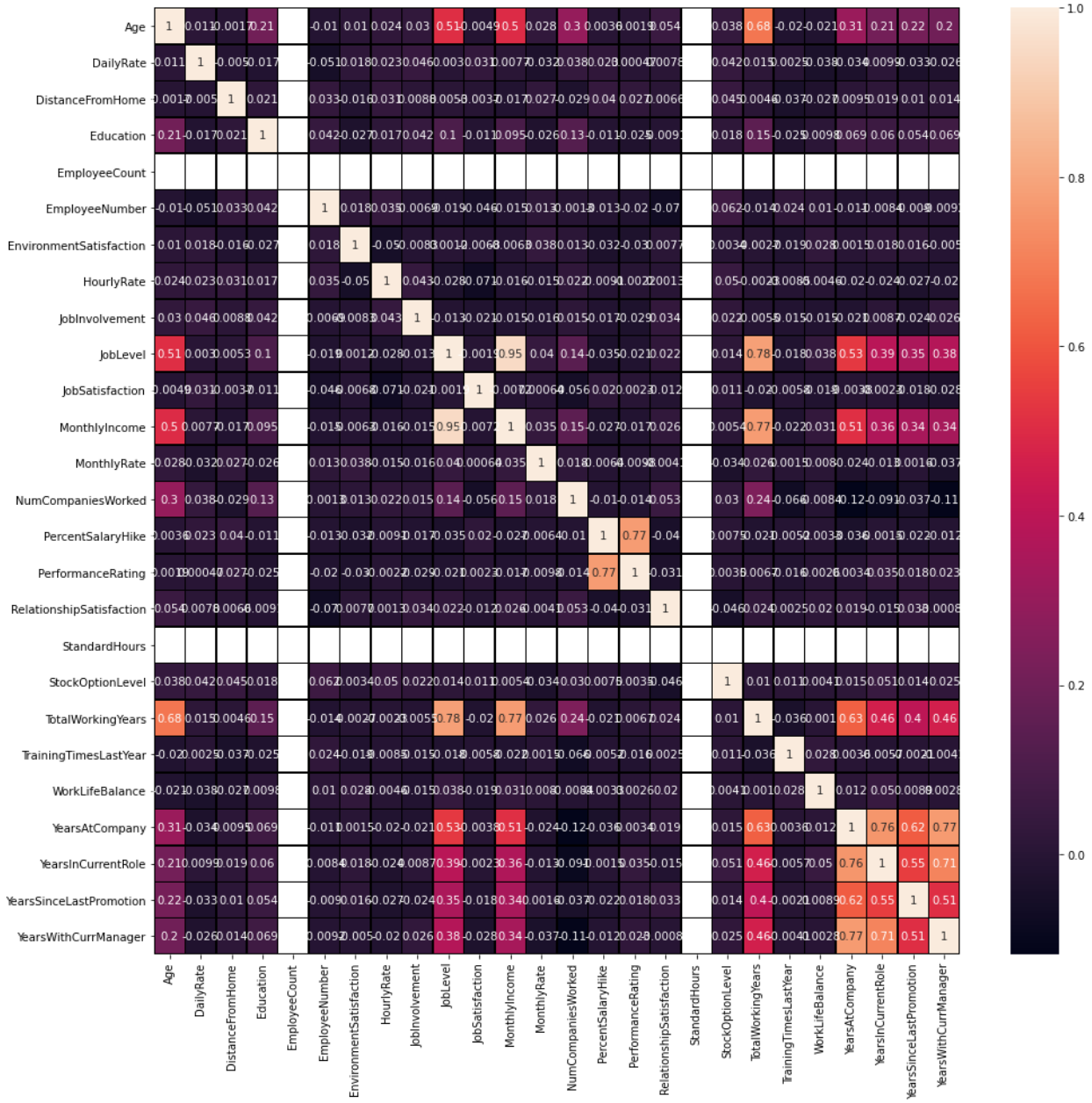


Figure A: Correlation before dummy variables

To create dummy variables, we segregate them from integer type variables. Since we will we working on classification problem with attrition as our target we'll drop that variables and store it in dataframe. We'll use "pandas.get_dummies()" to create dataframe with dummy variables. We'll also store numeric variables in dataframe. And check number of observations for both so we can concatenate them together.

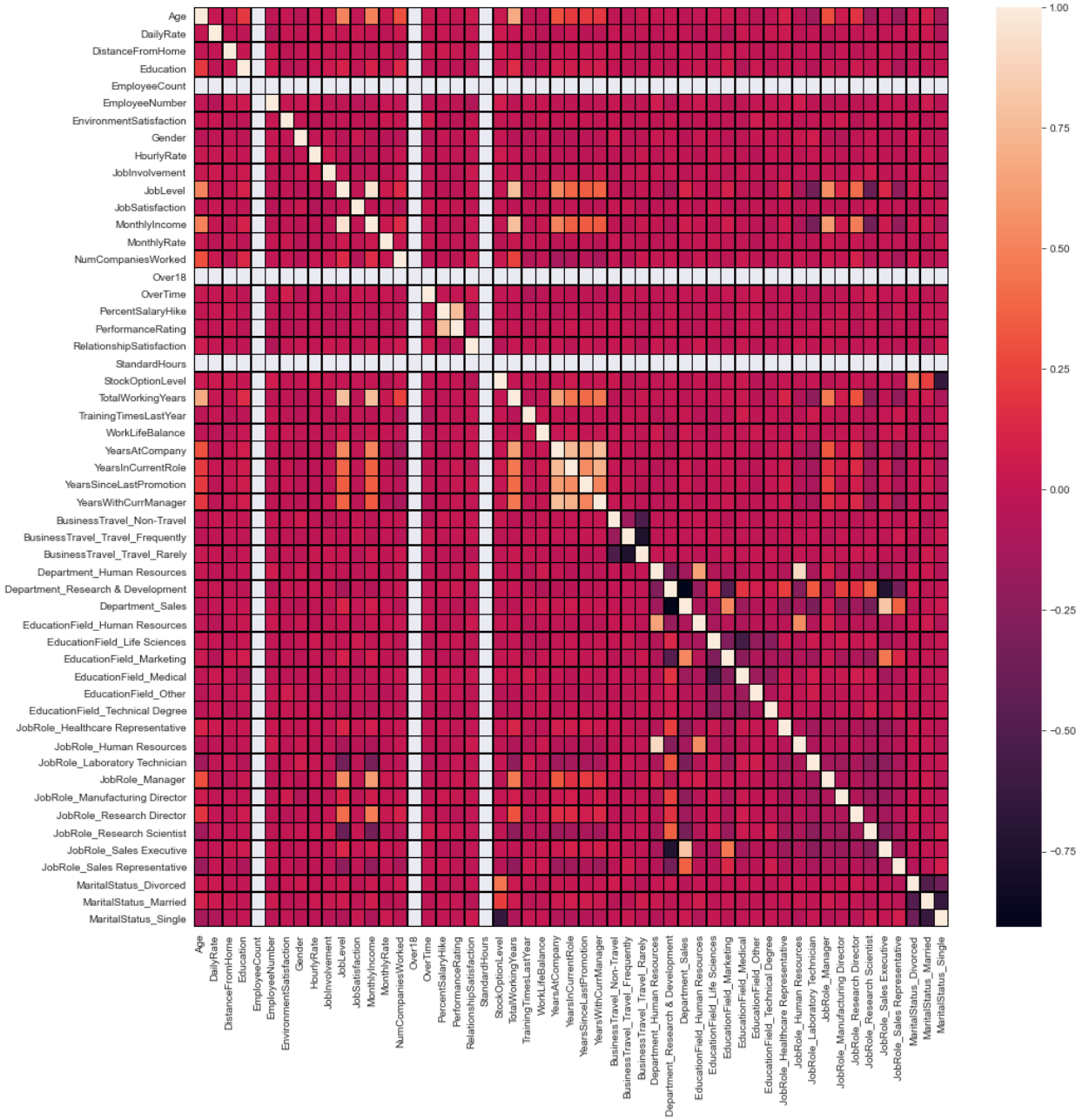


Figure B: Correlation after dummy variables

Modeling

For modeling, we keep “attrition” as our target variable and rest as in dependent variables in X. for machine learning we need to split our data in training and testing. We will have 70 percent of our data for training on which we will train our model and 30 percent for testing our model.

All this can be easily performed and evaluated using scikit learns different inbuilt machine learning modules. For our machine learning classification, we are going to use the below algorithms to classify “Attrition”.

1. **Logistic Regression:**
   In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Each object being assigned a probability between 0 and 1, with a sum of one. Logistic regression uses the sigmoid function which is mathematical function having S-shape sigmoid curve.

2. **Decision tree:**
   Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

3. **Naive Bayes:**
   In statistics, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. There are three types of Naive Bayes models.

   a. ***Gaussian:*** It is used in classification and it assumes that features follow a normal distribution.
   b. ***Multinomial:*** It is used for discrete counts. For example, let's say we have a text classification problem. Here we can consider Bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number x is observed over the n trials".
   c. ***Bernoulli:*** This model is useful if feature vectors are binary i.e., 0 and 1.

   Hence, for our model we are using Bernoulli Naïve Bayes classifier

4. **Random Forest:**
   Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

Evaluation

Below are results for our classification models. The tables consist of models used and their accuracies i.e. how well the algorithm has classified the labels. By looking at the table we can random forest has height accuracy compared to others especially for our problem. Hence, it is recommended that we choose it over others.

Sr.no.	Model	Accuracy (percentage)
1.	Logistic Regression	84 %
2.	Decision Tree	78%
3.	Naive Bayes	84%
4.	Random Forest	86%

We use confusion/error matrix to understand results. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. Below is confusion matrix from our Random Forest model.

	Attrition (Yes)	Attrition (No)
Attrition (Yes)	364	7
Attrition (No)	59	11

- Accuracy is calculated based on ratio of **true positive** and **true negative** by sum of all parameters in confusion matrix
- A **true positive** is an outcome where the model correctly predicts the positive class. Similarly, a **true negative** is an outcome where the model correctly predicts the negative class. A **false positive** is an outcome where the model incorrectly predicts the **positive class**.

Conclusion:

There are many factors for a person to changer his/her job. We have discussed some parameters in our exploratory data analysis although there are many more parameter as we expand and have a more in-depth look at our variables. We saw this when we saw our correlation plot to make better sense after we created dummy variable. This method of onehot encoding helped to improve our model for better results.

Companies can us such data for people insights to find pattens which create loop-holes within organization and use it to improve and retain talent. Hence, I feel working attrition is important.