IBM Performance Analysis

WDyZnLiP4:郑慧琳 5141619006; 邓迪 5141209241; 夏欣羽 515010910004;
张婧文 516120910184; 李玥沁 516120910175

I. Abstract

Our project aims to find out the features that related to the performance between the employees and the company. We first do some data description to have a general view of all the features. We draw the Pearson Map to find the inner-related features. Then, we draw some histogram to observe the distribution of satisfaction. We assume that job satisfaction, environment satisfaction and relationship satisfaction are the three most relevant factors to job performance and we explore which factors have influences on the satisfaction mentioned above.

We also use machine learning method to do the visualization, mainly use decision tree. We prepare the decision tree picture to demonstrate the outcome. We finally compare the square root errors and root mean square errors of KNN, Linear Regression, Decision Tree Regressor and Random Forest Regressor involving all factors, finding that Random Forest Regressor is the best way.

II. Data description

2.1 Data

We use the IBM Employee Attrition dataset in our project. In this dataset, there are many specific statistics of the employees such as business travel, daily rate, department, distance from home, education, education field, work life balance, age, years at company, monthly income, overtime, etc. We import those data into our program and then use the build-in function to calculate the average value, minimum value and the median of each set of value.

At the very beginning, as for the reason that we need all the values in the form of integer, we transfer the type of the data into integer before further steps. After this, we can start to explore how those features affect the performance between the company and the employees.

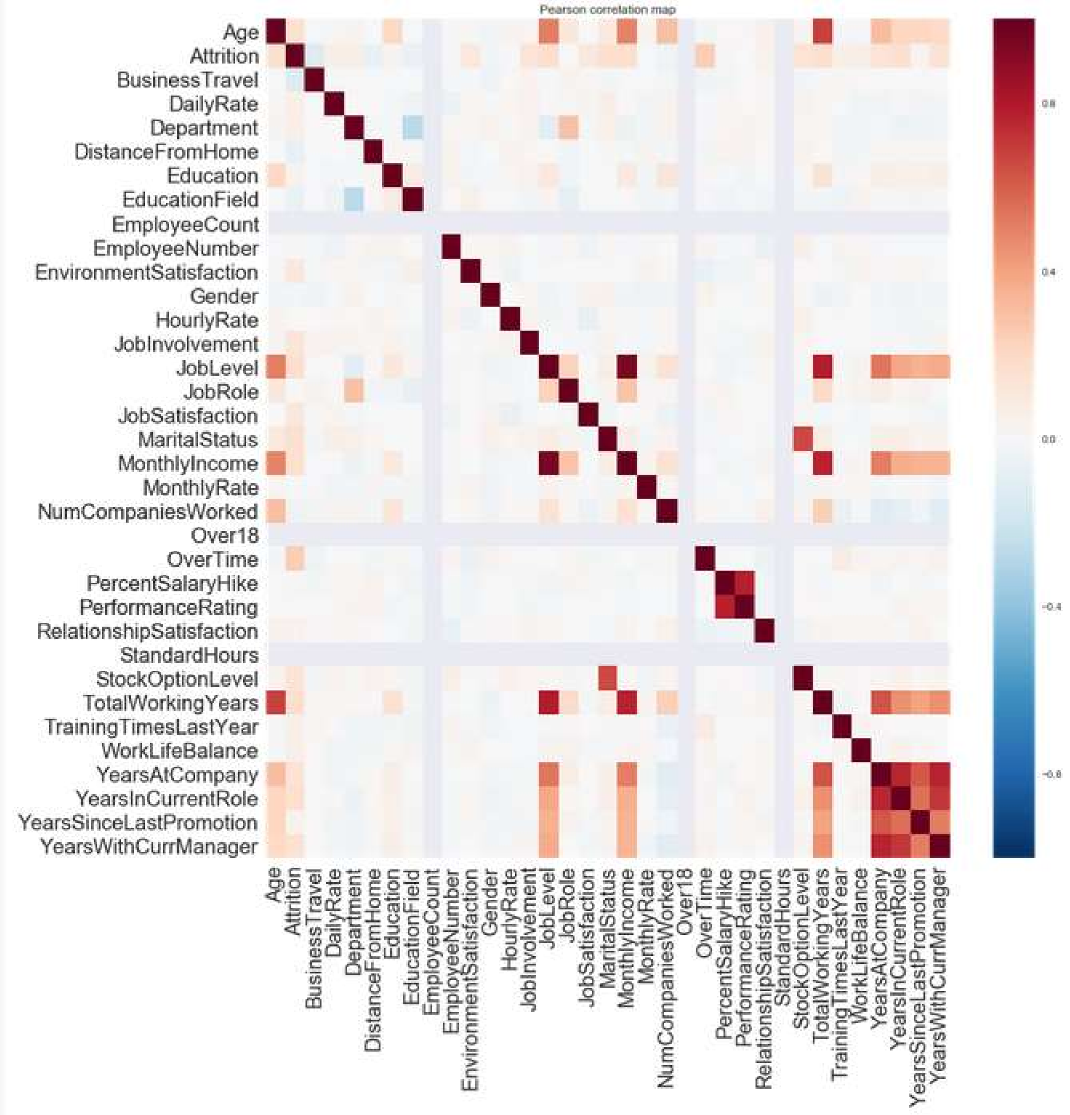
2.2 Pearson Map

To start up, we first want to see how those features related to each other. We draw the Pearson Map to see how the features are distributed with one another, and we can conclude which of the features are related to the other most closely.

As is shown below, we find out four sets of most inner-related features, which are:

- Job level and monthly income;

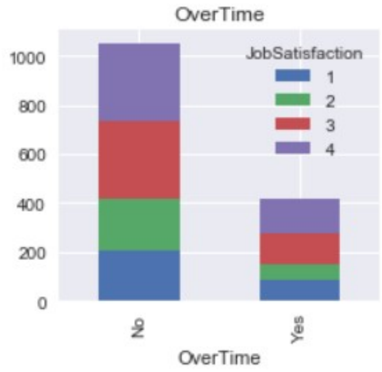
- Age and total working years;
- Job level and total working years;
- Monthly income and total working years;

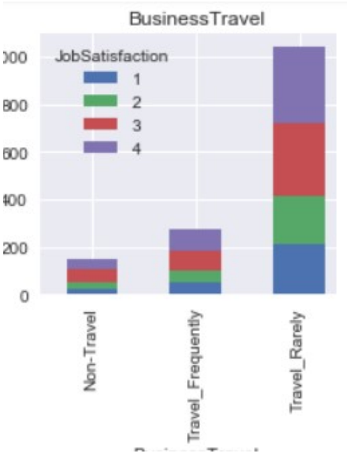


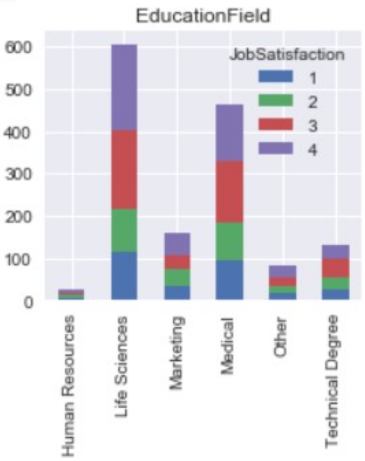
2.3 Plot histogram

We firstly plot histogram to observe the distribution of the satisfaction.

Because there are continuous variables and discrete variables, we program to draw histogram for continuous variables and draw bar chart for the discrete variables. The charts below show the distribution of the job satisfaction. We can see that some factors are more related with job satisfaction, such as business travel and over time, while other factors are not, like education field. The percentage of each satisfaction level for several dimensions of the factors is different, the more the difference is, the more related it is.







Several factors are related to relationship satisfaction, such as marital status and education field, while job role has less influence on it. For environment satisfaction, over time and education field is more related while gender is less related.

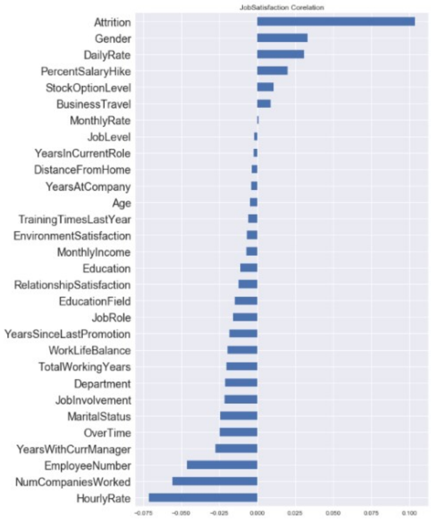
The problem of this part is that the diagram for the continuous variables could not show completely, so we could improve it in the future.

2.4 Satisfaction correlation

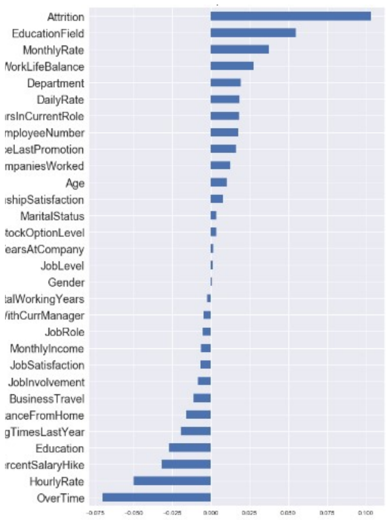
At first, we find the column that has only one value for all rows and delete them (employee count, over 18, standard hours, data type).

We then draw three correlation diagrams to find out the relationship between the satisfaction and other factors.

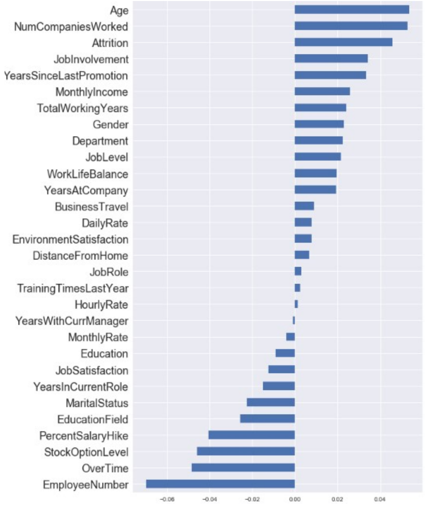
Job satisfaction



Relationship satisfaction



Environment satisfaction



We get the factors that have negative, positive or non influence on each satisfaction.

For job satisfaction, gender, daily rate, percenta salary hike have positive influence on it, while hourly rate, num companies worked, employee number influence it negatively.

As for relationship satisfaction, it is shown that education field, monthly rate and work like balance have positive effect on it, and over time, hourly rate and percentage salary hike have negative one.

When it comes to environment satisafction, there are more positive factors than negative factors. Age,number companies worked,job involvement are positive factors while employee number, over time and stock option level are negative one.

In addition, several factors have relatively no effect on satisfaction, such as monthly rate for job satisfaction, job level for relationship, and years with current manager.

Above all, we could find the significant factors that influence the sastisfaction based on the plot histogram and correlation diagram.

III. Machine learning

3.1 Preparation of the data

Due to the reason that we need the standardized version of data to get ready for visualization, we turn the data into the integer version. Take the factor of business travel as an example. For key None-Business Travel, we give it the value 1. And value 2 is given to key Travel-Rarely, while value 3 is given to key Travel-Frequently. We do the similar measure to attrition and other factors, including department, educational field, gender, job role, age, time and marital status.

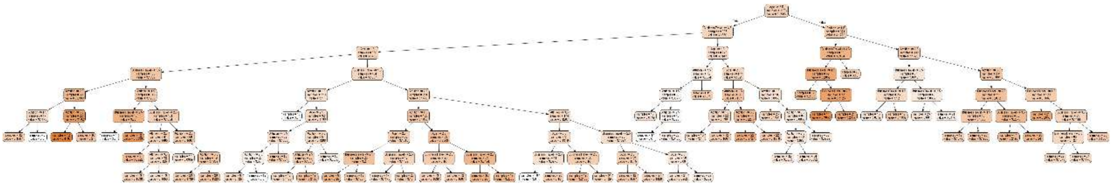
3.2 Machine learning method

3.2.1 Models for the influence coefficient of satisfaction on performance

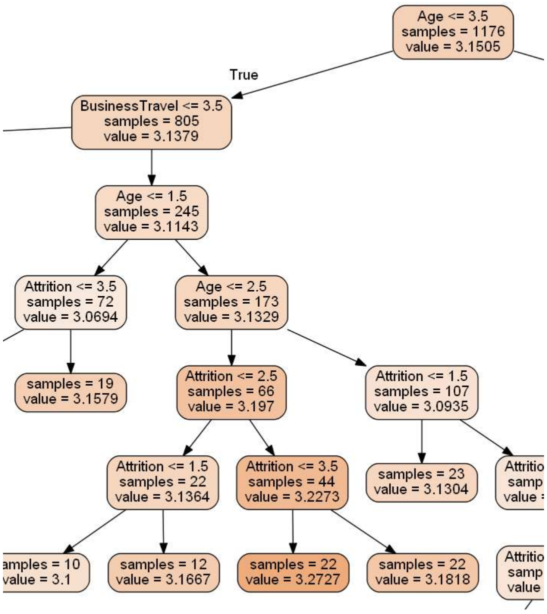
We use three satisfaction dimensions, including job satisfaction, environment satisfaction and relationship satisfaction, which have been totally discussed before to predict employees’ performance. We use machine learning to automatically learn the patterns and rules. Machine learning algorithms that we choose are K Nearest Neighbors Regressor(KNN), Linear Regression, Decision Tree Regressor and Random Forest Regressor. Test size is 0.2, which means that 50 groups are train data, while 10 groups are test data to be predicted. Then, we us both mean square errors and root mean square errors of each machine learning model to evaluate the accuracy and find the most accurate model among the four algorithms above. It turns out that the best way is the random forest, while in our project, due to our model’s defects，we use the Linear Regression.

3.2.2 The decision tree model

Considering the advantages of the visualization of decision tree, we especially use decision tree to do machine learning and predict the job performance. In this case, we use job satisfaction, environment satisfaction and relationship satisfaction, which are our assumptions before. Then, we use graphiz to draw the picture of decision tree.

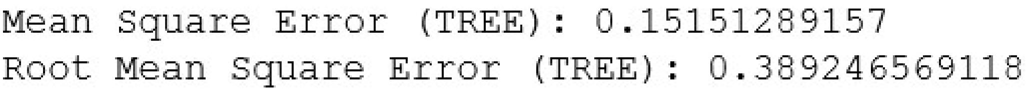


The picture is too large, so part of it is showed below to make it clear.



In the picture of the visualization of decision tree, “sample” is the number of data, while “value” represents the value of the corresponding sample. Take the part of Business Travel as an example. The ‘805’ means that there are 805 samples that meets the requirement of Business Travel <=3.5, and 3.1379 means the value of these samples.

As one of the most human-explainable models, the visualization of decision tree can be of great benefit. The effect of every factor on the result can be clearly represented, making it easy for us to understand. Moreover, we can clearly see the minor different influence of different factors.

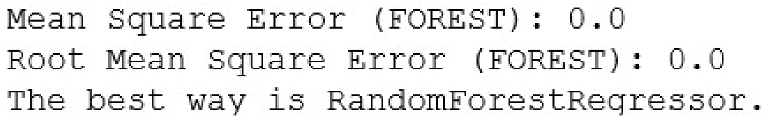


```
Mean Square Error (TREE): 0.15151289157
Root Mean Square Error (TREE): 0.389246569118
```

3.2.3 Machine learning models (involve all factors)

To test whether our model, which uses three special satisfaction dimensions, involving job satisfaction, environment satisfaction and relationship satisfaction can efficiently predict the job performance, we do the control group, using all the factor to train the model and do the prediction. To keep other elements the same as previous, we use machine learning algorithms, including K Nearest Neighbors Regressor(KNN), Linear Regression, Decision Tree Regressor and Random Forest Regressor, and keep the test size 0.2.

After comparing both the mean square root errors and root mean square errors, we can find that the best way is random forest regressor if all factors involves, and that using all factors is slightly better than using three satisfaction factors when it comes to accuracy.



```
Mean Square Error (FOREST): 0.0
Root Mean Square Error (FOREST): 0.0
The best way is RandomForestRegressor.
```

IV. Summary of the project and future study

Although it is of great complexity, there are certain circumstances where we can use machine learning algorisms to predict job performance. We also creatively use the method of visualization to make the results clearer.

Our group makes the assumption that job satisfaction, environment satisfaction and relationship satisfaction are the three most relevant factors to job performance and makes deep research on these three factors. After using different machine learning methods, especially decision tree, we learn the accuracy of the models. Despite the fact that using all factors can produce the most accurate results, models based on our assumptions have good results. Additionally, when it comes to huge numbers of statistics, our method can use less time and energy to achieve good results.

Our model has strong practical significance in how to improve performance, by promoting the 3 satisfactions which can be extended to many other corporations.
Our future study may focus on finding more and better factors and the model of machine learning to predict job performance more accurately. And there are also some defects in our model, to be dealt with. For example, in the histogram part, because the continuous variable will be covered, better visualization need to be developed. And more factors influencing the job satisfaction, environment satisfaction and relationship satisfaction also need to be taken into consideration.

In the end, thanks for Mr. Baoyang’s great contribution to our project.