Human Resources Analytics

Dilsa Sajan
Department of Computer Application
Amal Jyothi College of Engineeing
Kanjirappally
dilsasajan@mca.ajce.in

Jetty Benjamin
Department of Computer Application Amal Jyothi College of Engineeing
Kanjirappally
jettybenjamin@amaljyothi.ac.in

***Abstract*— In the past the machines were more concentrated in the centres but over time the focus is now on the work and activities of the workers as to why they are leaving the company and the means to prevent it. The Human Resources Analytics dataset, is used to explain the first steps in the data analysis path. In this first part is presented how to get familiarize itself with the data set by performing the descriptive analysis. Techniques such as exploratory data analysis (EDA) allow us to present the data in a more meaningful way, applying general statistical methods and exploratory graphics, that allow a simpler interpretation before engage a machine learning algorithm.**

Keywords— Human Resources Analytics, jupyter, satisfaction level, last evaluation, number of projects.

I. HUMAN RESOURCE ANALYTICS

It is the simulated dataset form Kaggle and it focus to understand why the employees is leaving the company. By the exploration of this it is possible to extract good insights and problems that the human resource department deals. In many industries retain their best employee it's a question of long-term strategy, and it impact the grow financial risk, for the employees leave to work at the competitor.

II. Exploratory Data Analysis(EDA)

All variables in the dataset to:

- Catch mistakes
- Generate hypotheses
- See patterns in the data
- Extract important variables
- Detect outliers and anomalies
- Gain deep familiarity with the dataset
- Refine selection of features that will be used to build the machine learning models.
- Do not skip the EDA process, because can generate inaccurate models or accurate models on the wrong data.
- This dataset contains 14999 objects and 10 attributes described below:

satisfaction_level	satisfaction level
last_evaluation	Last evaluation
number_project	Number of Projects
average_monthly_hours	Average monthly hours
time_spend_company	Time spent at the comp any
Work_accident	Whether they have had a work accident
left	Whether the employee h as left
promotion_last_5years	Whether had a promotio n in the last 5 years
sales	Departments (c olumn sales)
salary	Salary

III. Prepossessing the dataset

Before starting the process, its important to answer if it's clear what kind of problem we are dealing with, because in many cases isn't so simple to identy it. A good understanding of the problem will help to choose the right data mining and machine learning techniques to make the right predictions. Thus, the first step, is preprocessing the data to look for missing, incomplete or noise values.

NumPy: Is a fundamental package to use linear algebra and random number capabilities.

Pandas: Is a package to work with relational data as tables.

DOI: 10.5281/zenodo.6364787
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

```
In [2]: import numpy as np
        import pandas as pd
```

(Fig:1)

i. Load the data

To load the dataset we use a Pandas method called **read_csv** that read CSV(comma-separated) files and covert into DataFrame.

```
In [4]: data = pd.read_csv(r'E:\HR_comma_sep.csv')
```

(Fig:2)

Other useful method are used to store the summary of the dataset, like number of observations, columns, variable type and the total memory usage. The dataset have 14999 observations, 10 columns and with no null values. The data types of the variables are divided in 2 float, 6 integer and 2 object.

```
In [5]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   Work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
```

(Fig:3)

The first 5 lines of the dataset. The **head** method list first *N* rows from the Data Frame and the method **tail**, returns the last *N* rows.

```
data.head(5)
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

```
data.tail(5)
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	sala
14994	0.40	0.57	2	151	3	0	1	0	support	l
14995	0.37	0.48	2	160	3	0	1	0	support	l
14996	0.37	0.53	2	143	3	0	1	0	support	l
14997	0.11	0.96	6	280	4	0	1	0	support	l
14998	0.37	0.52	2	158	3	0	1	0	support	l

(Fig:4)

sample is a easy way to get a few data quickly.

```
data.sample(10)
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales
4222	0.98	0.87	4	152	2	0	0	0	sales
5980	0.65	0.70	3	243	3	0	0	0	sales
12776	0.10	0.97	6	264	5	0	1	0	sales
8272	0.60	0.76	3	270	2	1	0	0	marketing
4161	0.90	0.74	3	260	4	0	0	0	technical
10827	0.65	0.98	3	252	2	0	0	0	product_mng
8304	0.55	0.97	4	267	4	0	0	0	management
2041	0.98	0.91	4	240	3	0	0	0	technical
10776	0.89	0.99	4	205	2	0	0	1	sales
906	0.10	0.78	7	286	4	0	1	0	product_mng

(Fig:5)

ii. Variables transformations

To plot some statistical graphics and for better understanding, we make some transformations in the variables:

sales: Rename to department

salary: Convert the type of the variable from categorical to numerical.

```
data.rename(columns={'sales': 'department'}, inplace = True)

data['salary'] = data['salary'].map({'low':1, 'medium':2, 'high':3})
```

IV. Descriptive Analysis

The descriptive Analysis is used to simplify and summarize the mainly characteristics of the dataset. In other words, show what kind of information the dataset has. The Pandas method **describe** generates a descriptive statistic that summarize the central tendency, dispersion and shape of the dataset. By using this method in Human Resource dataset important insights is possible to see:

That approximately 24% os the employees left the company.

The satisfaction level is around 62% and performance is around 72%.

Employees work in average on 4 projects with 200 hours worked per month.

We have to find:

How many employees works in each department?

Depending on how many employees work in each department, you can learn more about the type of company segment.

How many employees per salary range?

The employees salary is divided in Low (1), Medium (2) and High (3), distributed as follows:

How many employees per salary range and department?

DOI: 10.5281/zenodo.6364787
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

```
print(data['department'].value_counts())
```

```
sales          4140
technical      2720
support        2229
IT             1227
product_mng     902
marketing       858
RandD           787
accounting      767
hr              739
management      630
Name: department, dtype: int64
```

```
print(data['salary'].value_counts())
```

```
1    7316
2    6446
3    1237
Name: salary, dtype: int64
```

(Fig:6)

```
table = data.pivot_table(values="satisfaction_level", index="department", columns="salary",aggfunc=np.count_nonzero)
table
```

department	1	2	3
IT	609	535	83
RandD	364	372	51
accounting	358	335	74
hr	335	359	45
management	180	225	225
marketing	402	376	80
product_mng	451	383	68
sales	2099	1772	269
support	1146	942	141
technical	1372	1147	201

(Fig:7)

V. How plot graphics?

In descriptive analysis is very useful to use graphics to represent the data. For that, is necessary to import the libraries:

Matplotlib: is a plotting lybrary, usefull to plot statistical graphics. See: www.matplotlib.org

Seaborn: is a library based on matplotlib that can draw attrative statistical graphics. See: seaborn.pydata.org/index.html

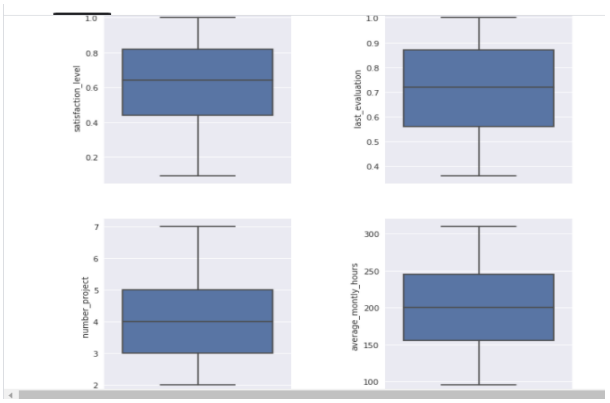
The boxplots below, give the information about the data distributions:

Satisfaction level and Last evaluation has a skewed left (negative) ditribuitions.

Number of projects has a skewed right(positive)ditribution.

Average monthly hours has a simetric ditribution.

Analyse de distribution of the variables is important due the fact that many statistical tests assume normal distribution.



In the boxplots below it is possibel to see that only time_spend_company has outliers.

Let's explain what kind of information is possible to conclude:

The employees with more time in the company have 10 years, so is possible to say that is a relatively young company.

Most of the employees have between 3 or 4 years in the company.

VI. Hypothesis

Now let's extract some more information and testing some hypothesis

We have to find the list of employees left form company.

- **First Hypothesis**

The first hypothesis is that salary is the reason why the employees left the company. Let's see if is this correct.

- **Second Hypothesis**

It is a dangerous job?

The second hypothesis is: employees leave the company because work is not safe.

- Performance Analysis

There are 2 distincts groups of employees. A group with poor performance and other with high performance employees. It's natural that employees that don't work well leave the company, but the main problem is that the high performance employees is leaving too and it's necessary to understand why.

It is possible to see that 98% of employees with few projects that left also have poor performance.

And 95% of the employees with 5 or more projects that left the company had the highest performance.

- Satisfaction Level

It is possible to see 3 interesting peaks in the satisfaction levels of the employees that left the company.

We have a peak of employees who are totally disappointed.

DOI: 10.5281/zenodo.6364787
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Another peak at 0.4, representing another group with the satisfaction level below the average.

And another amount in the range 0.7 and 0.9, with employees that left, although the high satisfaction.

Conclusion

It is a relatively young company, on average, employees have 3 or 4 years in the company and the oldest employees are working 10 years. The biggest difference in the salary from who stayed and those who left, was found in the management department, in the others departments although the salaries of who stayed be higher in average, it is not a big difference. The employees with 4 years in the company have the lowest average satisfaction level of all the company with (0.47).The satisfaction drops when the employees are working in 5 or more projects. A number of 3 or 4 projects seems to be ideal independent of the time spend in the company. The employees with 5 or more projects that left also worked at least 20% more hours than the average of the company. The satisfaction level of the employees that left is grouped in totally disappointed, below the average satisfaction and satisfied.

REFERENCES

[1] :https://www.sciencedirect.com/science/article/abs/pii/S1053482220300681

[2] :https://www.sciencedirect.com/science/article/pii/S2214785320401774

[3] :https://www.investopedia.com/terms/d/descriptive-analytics.asp

[4] :https://www.sciencedirect.com/science/article/abs/pii/S0301420719305860

[5] :https://www.sciencedirect.com/science/article/abs/pii/S0301420719305860

DOI: 10.5281/zenodo.6364787
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam