TDS3301 DATA MINING ASSIGNMENT

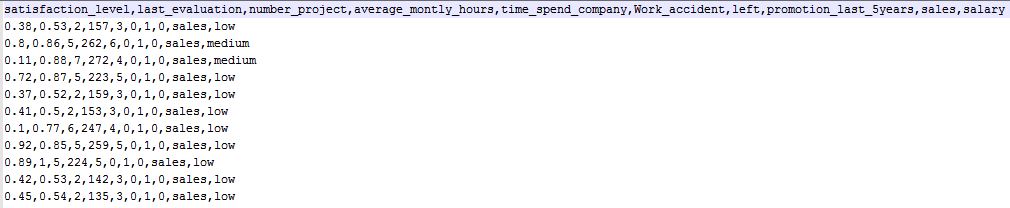
Dataset “Human Resources Analytics”.

A)

The uploader of the data asked the following question:

Why are our best and most experienced employees leaving prematurely?

The aim with the dataset is to predict which employees will leave next. When we analyse the data set maybe we can find a reason and generate rules.



The data set is comma separated.

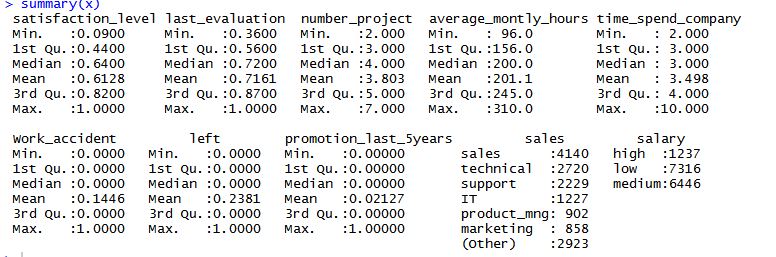
*read.csv(file = "HR\_comma\_sep.csv", header = TRUE, sep = ",")*

In the dataset, we have 10 columns and 14999 rows. One row describes one employee. With the command “summary”, you can easily analyse the data. In addition, we can see that there are 11428 people still employed and 3571 have left.

*Summary(data)*

|  |  |  |
| --- | --- | --- |
| Column | Description | Value |
| satisfaction\_  level | This column describes how stratified the company with an employee is. Data in % | Min: 0.1  Avg: 0.6  Max: 1.0 |
| last\_evaluation | When was the last evaluation of the satisfaction level? Data in % |  |
| number\_project | The number of Projects an employee is/was involved | Min: 2  Avg: 3.83  Max: 7 |
| average\_montly\_hours | The average number of a Worker per month  in hours | Min:96  AVG: 201.1  Max: 310 |
| time\_spend\_company | Here we can see the total time an employee spent in the company. Data is in Years | Min:2  Avg: 3.498  Max: 10 |
| Work\_accident | Amount of working accidents. | Min: 0  Avg: 0.1446  Max: 1 |
| left | Here we see the employee who are working or worked for the company | True (1) or false (0) |
| promotion\_last\_5years | The column shows us the employees who got a Promotion in the last 5 Years | True (1) or false (0) |
| sales | Here we can see the different departments an employee is working. There are 7 departments in the data set | Sales, technical, support, IT, Marketing, other, product mngt. |
| salary | In this column, we can see the how much an employee earns. There are 3 different categories | low  medium  high |

This table describe the characteristics of each features of our ABT. We can see different statistical measures of central tendency and variation. For example, we can see that our attrition rate is equal to 24%, the satisfaction level is around 62% and the performance average is around 71%. We see that on average people work on 3 to 4 projects a year and about 200 hours per months.



Correlations between the attributes: 

B)

When we analyse the data set we can find correlations between the attributes. Maybe there is a correlation between the working hours and the salary. Is the company paying to less to their employees? In this case, we can make a rule why the people are leaving the company and we can predict who will leave next. When we early know which of our top employee will leave next, we can make measures to keep them in the company. After grouping them per department, we could email the different managers to tell them which valuable employees might leave soon.

C)

In the Human Resource Analytics dataset, we can use the Classification as data mining technique. I would chose the column “left”. In the column “left”, we have two classes:

* Employed (0)
* Left (1)

With a tree, we can make a training set out of the Data with the employees who have left. After that, we can use the tree for the employees who are still in the company and make a prediction of the next leavers.

D)

One Problem with the data quality is the GIGO principle.

Garbage in – Garbage out

When you have a bad data quality and a good model, you will have garbage results. One of the classic problems of data analysis are outliers. Outliers differ significantly from the other data or are outside the usual range of values. It is often useful to examine metric attributes more closely. Whether such outliers should be scuttled or adapted for data mining or still in the original state, depends on the concrete context.

One of the main problems in data pre-processing are missing, inaccurate, contradictory, and contradictory values.

There is now a question: how you deal with such data?

Is the missing data ignorable or does it affect the results if the data mining. In many cases you can replace the data by special features like the negative numbers that cannot occur (for example negative numbers for stock items).

Another problem that can occur is the reduction of the data dimensionality. The reason for this is a limited computing capacity. In most cases are for one row many attributes. This leads us to high dimensionality. A high dimensionality always needs a lot of computing power and increases the runtime of the program. This is unacceptable for many applications.

The reduction of dimensions can be done in two ways. You can delete an attribute or summarize attributes who belong together.

A further problem in the data reduction can be that some attributes are in the wrong format so they are unsuitable for the mining process. For example, the data is in millimetres and you need it in kilometres. Similarly, a scaling of the data is necessary.

The aim of the data pre-processing is to improve the data quality.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | Outliers | Missing Value | Wrong Values | Reduction of Dimensions | Format Change |
| satisfaction\_  level | No | No | No | No | No |
| last\_evaluation | No | No | No | No | No |
| number\_project | No | No | No | No | No |
| average\_montly\_hours | No | No | No | No | Maybe |
| time\_spend\_company | No | No | No | No | No |
| Work\_accident | No | No | No | No | No |
| left | No | No | No | No | No |
| promotion\_last\_5years | No | No | No | Maybe | No |
| sales | - | No | No | No | No |
| salary | - | No | No | No | No |

In this dataset, the data quality is very good. There are no outliers or NA values. In addition, there are no wrong values. To decrease the run time of the program we can maybe delete the promotion attribute because it has a just a little correlation to the other attributes. However, not sure if that is useful. We can change the format from the average\_monthly\_hours to average\_daily\_hours. At least we could make the attribute salary scalable for example low =1 medium= 2 high = 3.

E)

The problem with this data set is that the pre-processing is done, so we already have a good data quality.

Sources

https://www.kaggle.com/ludobenistant/hr-analytics

https://de.wikipedia.org/wiki/Klassifizierung

Lecture Notes