Determine which employees are prone to leave next

TakenMind Internship July 2019

Santiago Frias Moreno, Tech Developer
Tarragona, Catalonia(ES)

Goal: determine which employees are prone to leave next

Target:

1. CREATE A MS-PRESENTATION (MAXIMUM 12 SLIDES) EXPLAINING THE REASONS EMPLOYEE ARE PRONE TO LEAVE AFTER ANALYSING THIS DATA SET.

2. EXPLAIN WHAT TYPE OF EMPLOYEE ARE PRONE TO LEAVE THE COMPANY.

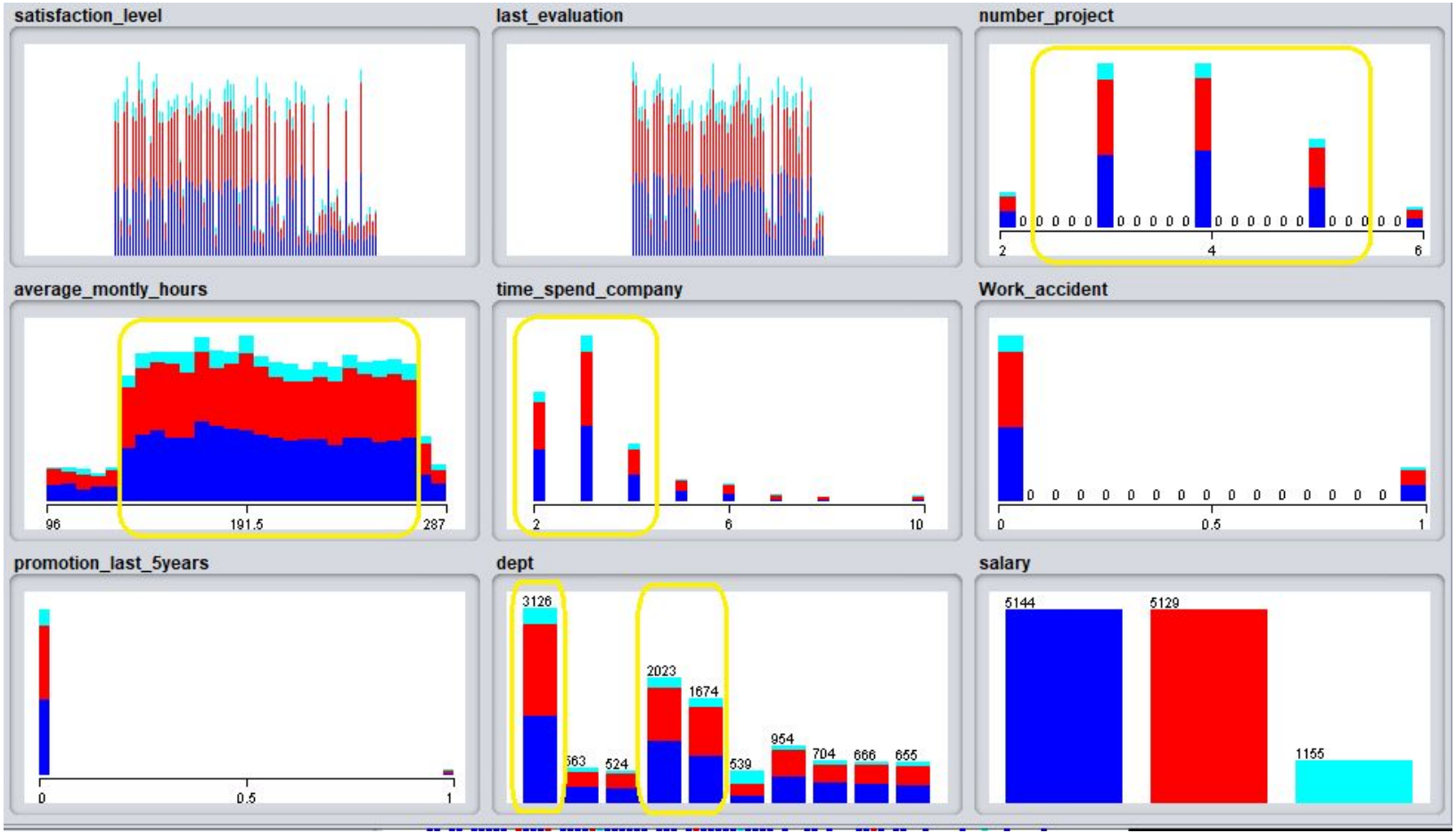
3. PREDICT THE FUTURE EMPLOYEE WHO WOULD TEND TO LEAVE THE COMPANY.

Dataset: same on two parts merged (left and not left employees)

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

Exploratory for more significants who have left to select best prediction methode

Histograms



First Insights: profile of all employees

- With number of projects between three to five-

- With delimited average of montly hours

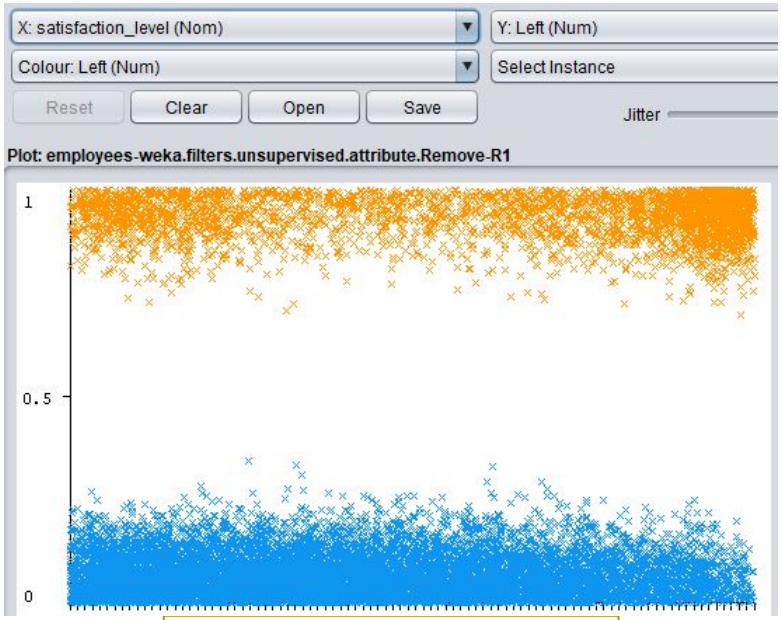
- Two group of departments where there is greater impact (more than 1.5 standard deviation):

    > First group: sales.

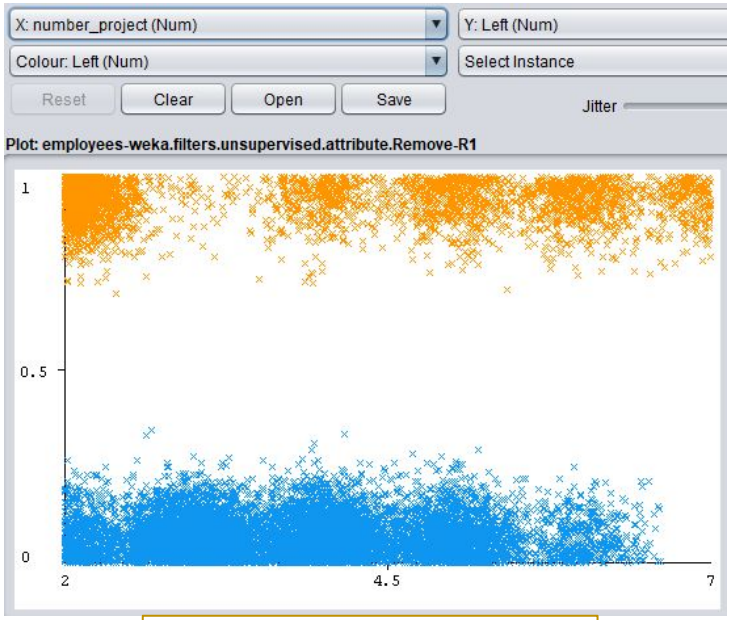
    > Second Group: technical and support.

- preferably low and medium salaries.

Compare left group to non left group (I)



Satisfaction level

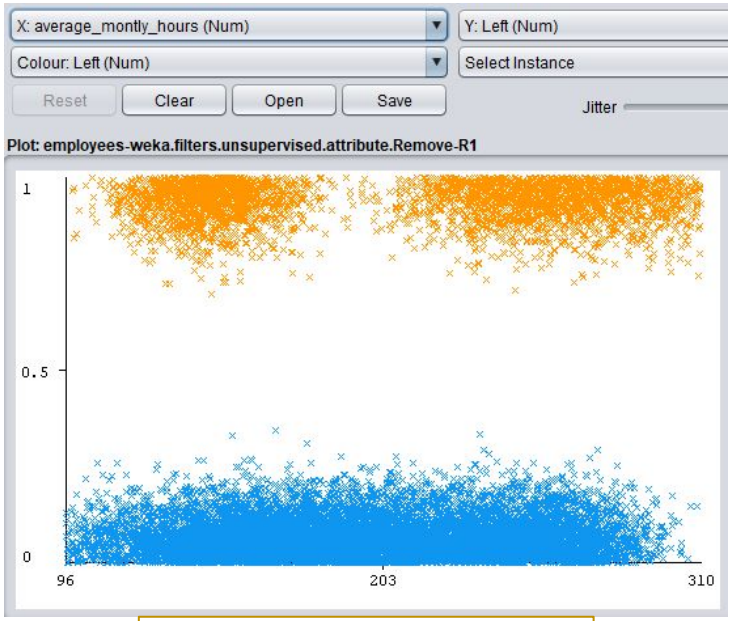


Number of projects

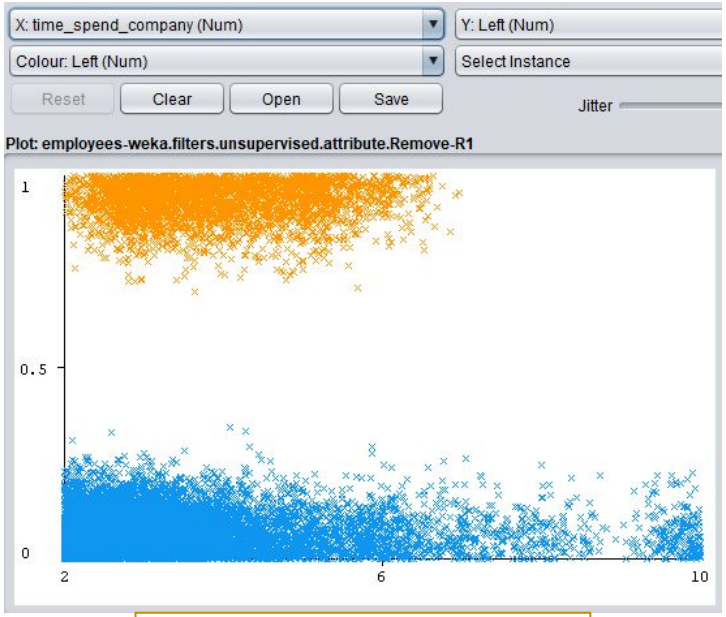
Satisfaction level higher are slightly more prone to left out (Breach of expectations?)

For two projects, left out is lower.

Compare left group to non left group (II)



Avg.month hours

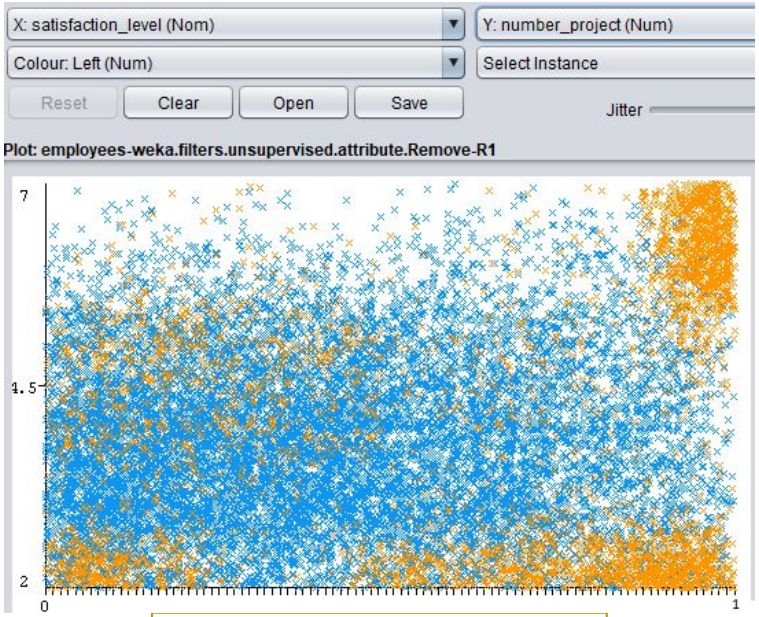


time spent company

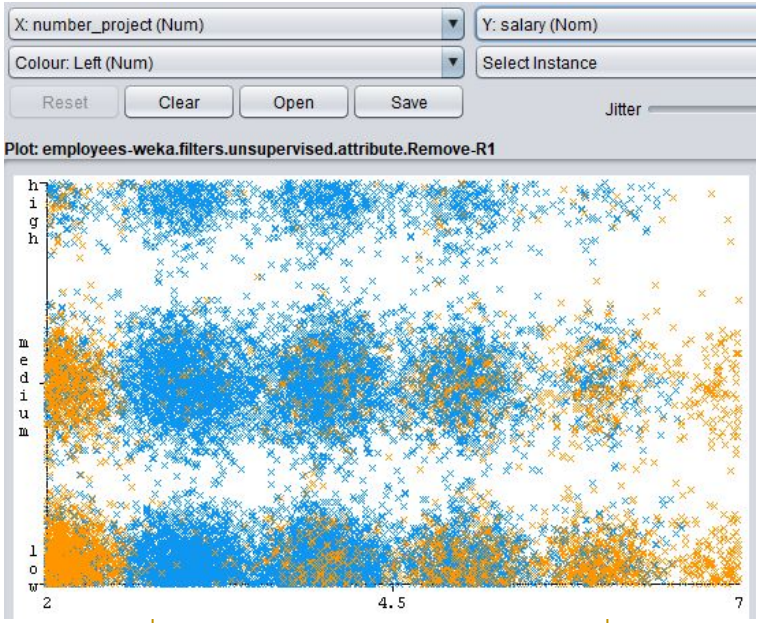
Avg. month hours between 190h to 220h. has less left out.

time spent company slightly higher for time spent company 4y. to 5y.

Cross relationships (I)



Satisfact/n.projects



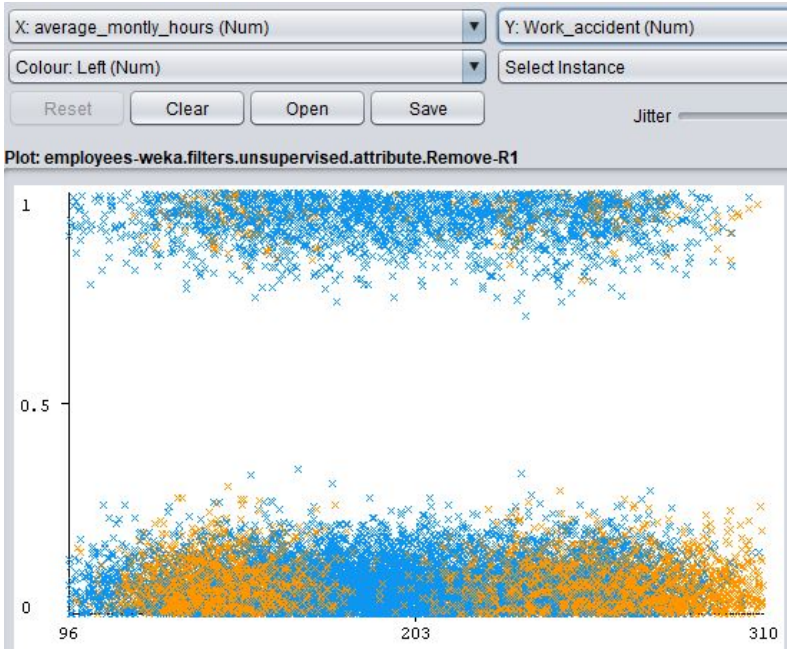
n.project/salary

Distribution of left out is different for Satisfaction/number projects representation.

Left out concentrated on medium/low salary on n.project/salary representation.

Cross Relationships (II)

avg.month.h/time spent company



avg.month.h/work accident

Avg. Month hours vs. years spent company has two focused zones: 100-126h&3y, 126-160h&4-5y

Work accidents occurs on this focused zones.

Roadmap to automate learning model and infering attrition problem

- Load two sheets of datasheet with pandas, concatenate, and profile attributes.
- Filter values and rows, discarding stathistical outliers.
- Determine good points for bin intervals of model (unsupervised learning groups, EM-Means)
- Generate a simple model (e.g. supervised classification model) to infere upshot.
- Validate model and know best parameters to define model inference as reliable.

Why:

- ❏ Structure data can change over time.
- ❏ If data changes along time, we can train newly model, cutting or adding new dataframe series.