Churn analysis on human resources data, from Medium.com

Claudine Chen
May 30 2017

Improve your employee retention

- Keep your competitive edge with the best employees
- Use your existing human resources and survey data
- Make informed decisions on how to provide a better experience for your employees, and optimize income
- Machine learning and statistics offer tools to do exactly that, especially on a large scale.

The example data from Medium.com

- offered by Medium.com on Kaggle (https://www.kaggle.com/ludobenistant/hr-analytics)
- To understand employee churn - why are they leaving when they do
- contains 14999 simulated employees

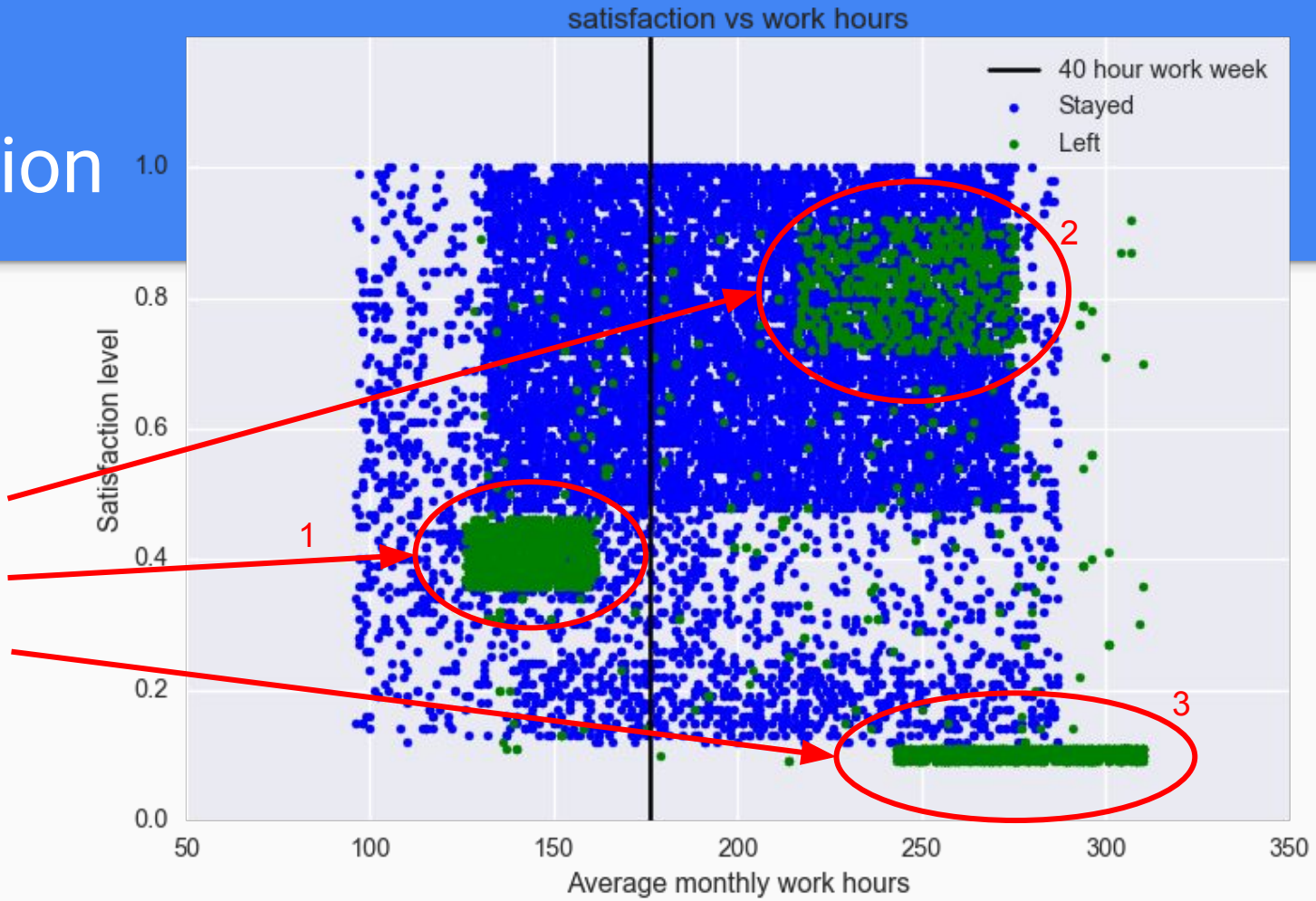
Features:

- employee satisfaction level
- their last evaluation
- time spent at the company
- whether they have had a work accident
- their department

- *Workload:*
  - number of projects
  - their average monthly hours
- *Rewards from employer:*
  - salary level
  - whether they have had a promotion in the last 5 years

Data Exploration

3 clusters of leavers when we look at satisfaction vs. work hours.



Demographics of our leavers

	Satisfaction level	Last evaluation	Monthly work hours	Number of projects
Cluster 1	0.3-0.5	0.4-0.5	125-165	2
Cluster 2	0.7-1.0	0.8-1.0	210-280	4-5
Cluster 3	0.1	0.7-1.0	240-310	6-7

- Are cluster 1 leavers underutilized?
- Cluster 2 leavers are happy high performers. What do they need to stay?
- Are cluster 3 leavers overworked?

Random Forest for the win

We tried various classifying machine learning models - non-linear does well, and **Random forest does best.**

Even “stupid” model that assumes everyone stays will have 76% accuracy.

Model	accuracy score	AUC-ROC	AUC-PR
baseline (assumes all stay)	0.76		
logistic regression (linear)	0.77	0.80	0.45
logistic regression (polynomial deg=2)	0.85	0.91	0.69
logistic regression (polynomial deg=3)	0.83	0.91	0.73
logistic regression (interaction terms)	0.91	0.93	0.76
random forest	**0.99**	**0.99**	**0.99**
SVM (linear kernel)	0.77	0.79	0.50
SVM (rbf kernel)	0.96	0.97	0.91
SVM (3rd deg polynomial kernel)	0.96	0.97	0.92

Random forest says...

The most important features are:

- satisfaction level
- last evaluation
- number of projects <- something the company can change
- monthly hours <- something the company can change
- time spent at the company

Promotion is not on this list, because only 2% were promoted in the last 5 years, but the churn rate on average is 24%, but of those who have been promoted, it is 6%. This is also something the company can change.

Recommendations to the employer (part 1)

- The majority of leavers leave within 5 years and have not had a promotion in those 5 years. Promote more people within 5 years.
- Cluster 2 people are happy high performers. Possible actions to prevent them from leaving are to provide perks that make the job more attractive, like higher salary, or anything else the company can offer at low to minimal cost such as more flexible working hours, more holiday, more stock options or equity
- Cluster 3 leavers have too many projects and have to work overtime. The employer could lessen the load of these unsatisfied high performers. This might reduce from the churn rate of people with 6-7 projects (64%) to the churn rate of people with 4-5 projects (14%)
- Cluster 1 leavers might want full time jobs, or more engaging work.

Recommendations to the employer (part 2)

- Know the cost to the company of losing highly productive people and retraining new employees - this cost could be invested in keeping these highly productive people.
- Improve the resolution of the data being collected to include satisfaction level of work hours, work topic, sense of responsibility, and advancement.