An Approach to Attrition and Performance Prediction

Victor Tapia

5/6/2019

Introduction

Working is the activity that most people will do most of their lives. With a working day of 8-9 hours on average and between 45 to 50 productive years, an average people will spend around 13 years working and 26 years in bed (Campbell 2017).

Given the amount of time every person will spend working, it's important to feel engaged. Schaufeli et al (2002, cited by Mäkikangas et al. 2013)

An important factor to consider is the organizational commitment that can be defined as a linkage, bond, or attachment of an individual to an organization (Klein, Molloy & Cooper, 2009, cited by Gomes Maia and Bittencourt Bastos 2015). They proposed a three-component model: affective, normative and continuance.

The affective component, highlights the emotional linkage between person and an organization. The normative component means that the link is based on a felling of obligation while the continuance component is due there is no other choice (Bastos et al, 2014, cited by Gomes Maia and Bittencourt Bastos 2015).

But the flip side of the coin, on average, workers will change job every 4.2 years, workers in management, professional and other related occupations every 5.0 years while workers in service has the lowest tenure with 2.9 years (Belli 2018).

About work change, employee turnover and attrition are two different types of employee churn and both of them are commonly used as synonym (Pawlewicz 2018).

Even though both of them decrease the number of employees on staff, attrition is tipically voluntary or natural, like retirement or resignation, while turnover can be either voluntary resign or involuntary termination or discharge (Pawlewicz 2018).

The dataset SAMPLE DATA: HR Employee Attrition and Performance (Stacker IV 2015) will be used. This is a fictional dataset created by IBM for practice.

For this project the following tasks will be performed:

1. String variables will be converted into factors and boolean variables will be converted in numeric 1 or 0.
2. The variables with zero variability will be removed as they doesn't add any value.
3. Some plots will be reviewed in order to understand how the sample is built.
4. Correlation analysis will be performed in order to drop the higher correlationed variables.
5. The following methods will be used in order to get the best ensemble: glm, lda, naive_bayes, svmLinear, knn, rf, ranger, Rborist, gbm, xgbTree, svmRadial, svmRadialCost and svmRadialSigma.

Methodology

This dataset consist of 35 variables and 1470 observations randomly generated. After the preparations, some fields were renamed and factors were created. Some fields were also removed due to zero variability:

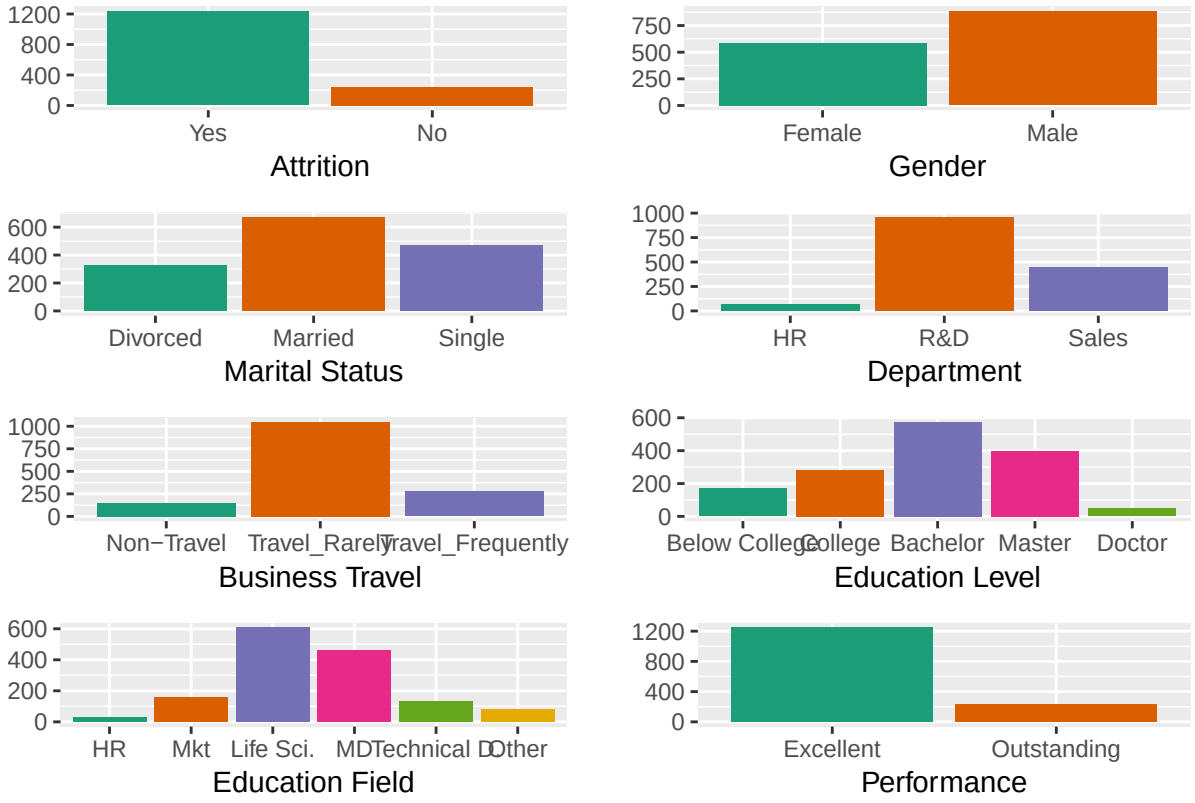
```
##               Unique Value
## employeeCount            1
## over18                   1
## standardHours           80
```

Before data partitioning, and cleansing the final dataset structure is as follows:

```
## 'data.frame':    1470 obs. of  32 variables:
##  $ age                : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ attrition          : Factor w/ 2 levels "Yes","No": 2 1 2 1 1 1 1 1 1 1 ...
##  $ travel             : Factor w/ 3 levels "Non-Travel","Travel_Rarely",..: 2 3 2 3 2 3 2 2 3 2 ...
##  $ employeeNumber     : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ dailyRate          : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ department         : Factor w/ 3 levels "HR","R&D","Sales": 3 2 2 2 2 2 2 2 2 2 ...
##  $ distHome           : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ education          : Factor w/ 5 levels "Below College",..: 2 1 2 4 1 2 3 1 3 3 ...
##  $ educationField     : Factor w/ 6 levels "HR","Mkt","Life Sci.",..: 3 3 6 3 4 3 4 3 3 4 ...
##  $ envSatisfaction    : Factor w/ 4 levels "Low","Medium",..: 2 3 4 4 1 4 3 4 4 3 ...
##  $ gender             : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ hourlyRate         : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ jobInvolvement     : Factor w/ 4 levels "Low","Medium",..: 3 2 2 3 3 3 4 3 2 3 ...
##  $ jobLevel           : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ jobRole            : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ jobSatisfaction    : Factor w/ 4 levels "Low","Medium",..: 4 2 3 3 2 4 1 3 3 3 ...
##  $ maritalStatus      : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ monthlyIncome      : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ monthlyRate        : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ numCompaniesWorked: int  8 1 6 1 9 0 4 1 0 6 ...
##  $ overTime           : num  1 0 1 1 0 0 1 0 0 0 ...
##  $ percentSalaryHike : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ performance        : Factor w/ 4 levels "Low","Good","Excellent",..: 3 4 3 3 3 3 4 4 4 3 ...
##  $ relSatisfaction    : Factor w/ 4 levels "Low","Medium",..: 1 4 2 3 4 3 1 2 2 2 ...
##  $ stocOptionLevel    : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ totalWkgYears      : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ trainingTimesLY    : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ workLifeBalance    : Factor w/ 4 levels "Bad","Good","Better",..: 1 3 3 3 3 2 2 3 3 2 ...
##  $ yearsAtCompany     : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ yearsInRole        : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ yearsSinceProm     : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ yearsWithManager   : int  5 7 0 0 2 6 0 0 8 7 ...
```

The first step is to start digging into the information, giving a quick look, we can see that the attrition rate is NA%. In the following plots we can see how the sample is distributed.

Fig 1. Distribution for most important categorical variables

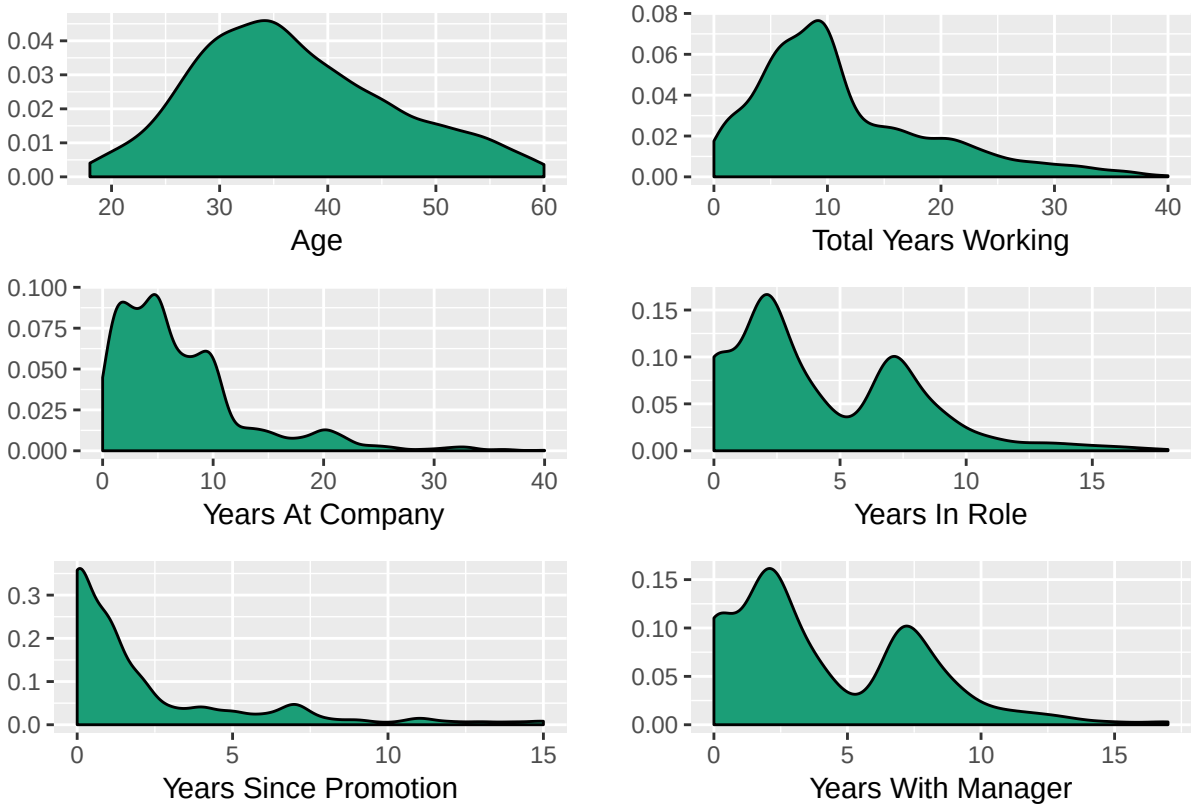


From the above plots we can notice the following:

1. Attrition: This plot confirms the attrition rate calculated before. This is the variable to be predicted.
2. Gender: The sample contains NA% of women and NA% of men.
3. Marital Status: The mode for this variable is Married, it is almost the double of the less frequent marital status which is Divorced.
4. Department: The sample contains only data from three departments: Research & Development (R&D), Sales and Human Resources (HR). The more frequent department is R&D (65.37%)
5. Business Travel: Most of the employees travel rarely (70.95%)
6. Education Level: The more frequent education level is Bachelor (38.91%) followed by Master (27.07%).
7. Education Field: Most of the employees studied a Life Science (41.22%) followed by Medical (31.56%)
8. Performance: This sample contains only employees with Excellent (0%) and Outstanding performance (0%).

The dataset also includes some time-related variables like age, total years working and others. Below are the density plots for this variables.

Fig 2. Time−related density plots



In this case, age seems to be the only normal distributed variable with an average age of NA and a standard deviation of NA. The older person is $-\infty$ years old, while the younger is ∞.

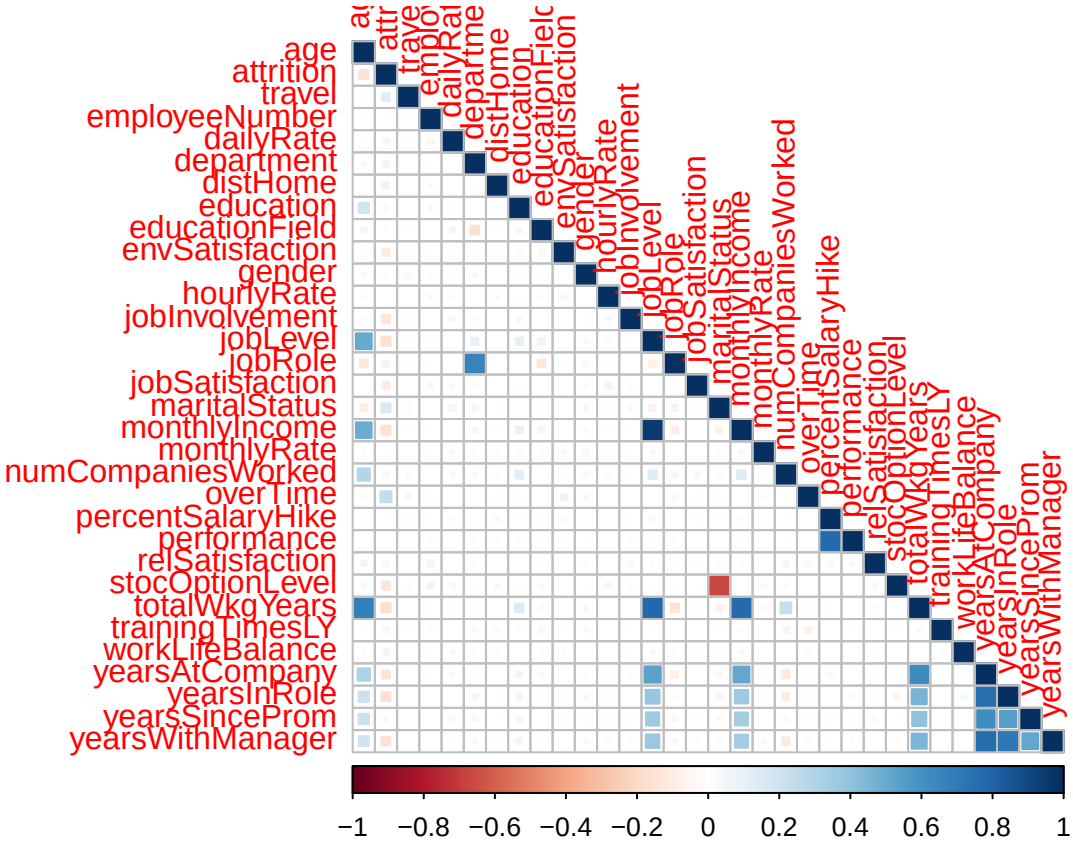
Another interesting finding is that Total Years Working and Years At Company seems to be similarly distributed. While Years in Role and Years with Manager shows a almost identical distribution.

There are other variables worth taking a look.

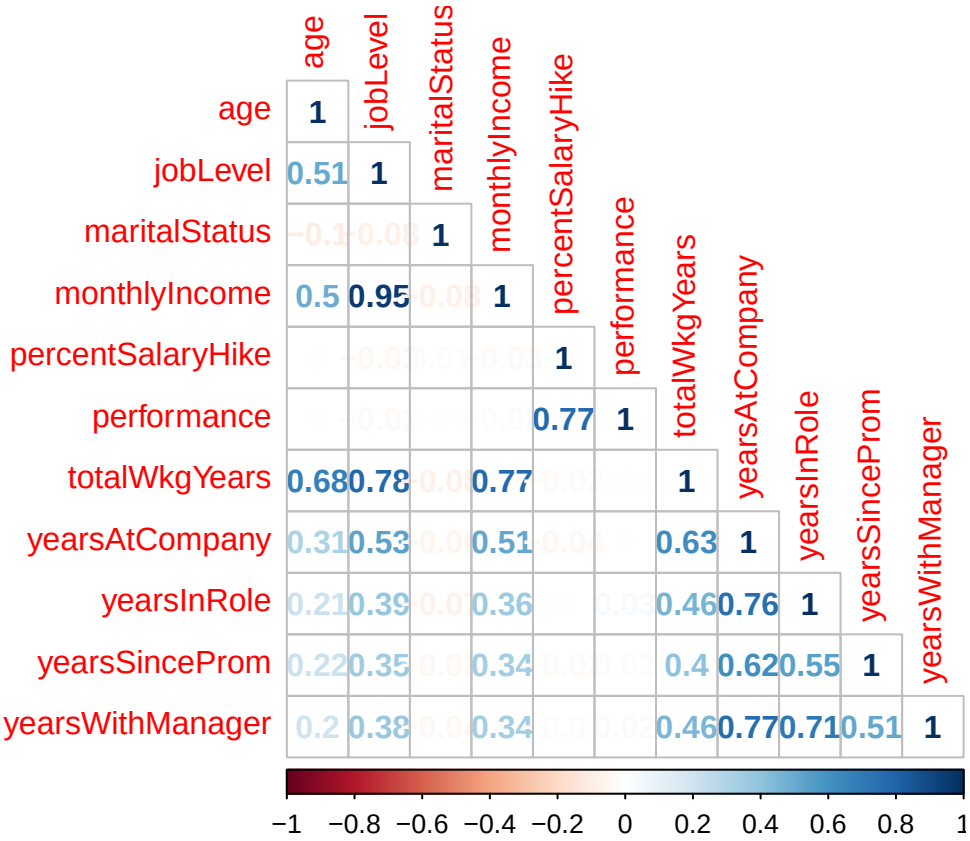
Fig 3. Density plots

Distance from Home and Percent Salary Hike seems to be a very high density at very low values. Daily Rate and Hourly rate seems to have a similar distribution and also the number of companies worked with the monthly income shows similarity.

The next step is to lookup for correlated variables, the following plot shows this correlations:



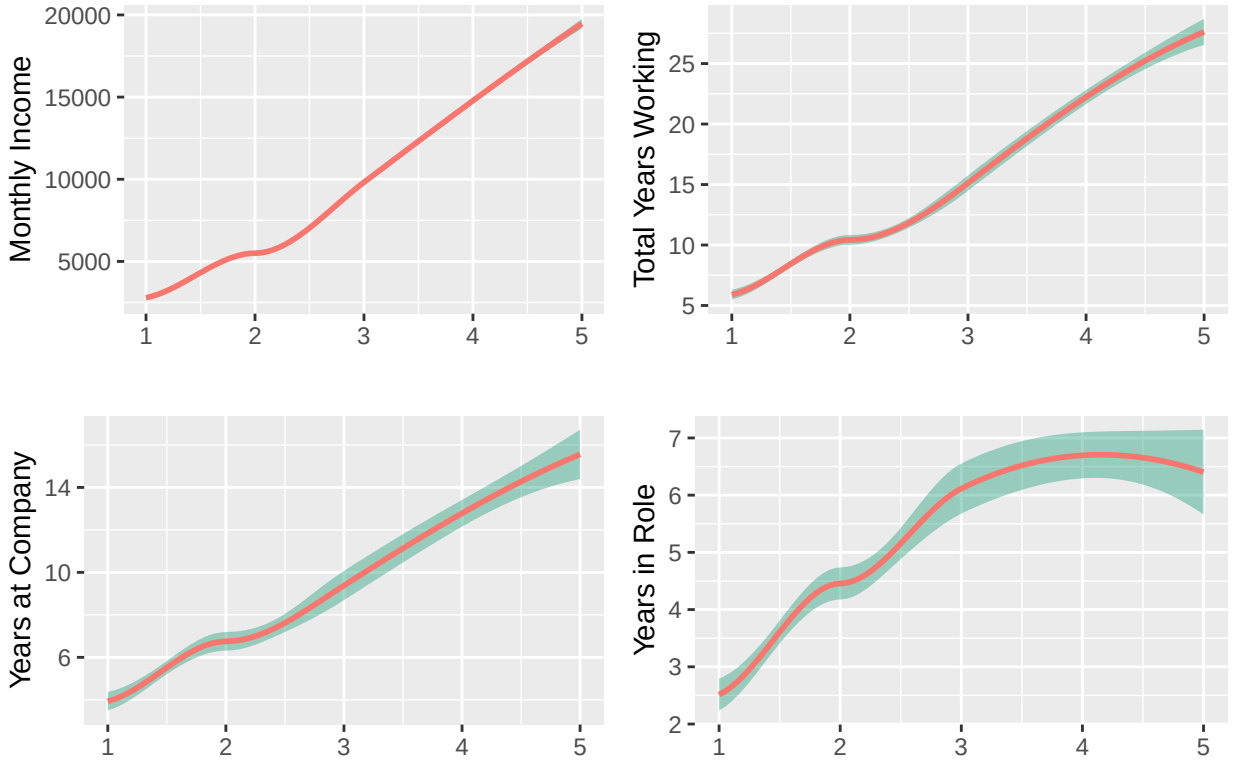
At first, the plot seems to be complicated to understand due the quantity of variables, but after removing those variables with low correlation, the new plot looks like this:



The variables Performance and Percent Salary Hike seems to be correlated. Given that assumption, the variable Performance will be removed. Age will be also removed as it shows correlation with Monthly Income.

The variable that seems to have more correlations with another variables is Job Level, specially Monthly Income, which is logical as the higher the position, the higher the income. These plots show those correlations:

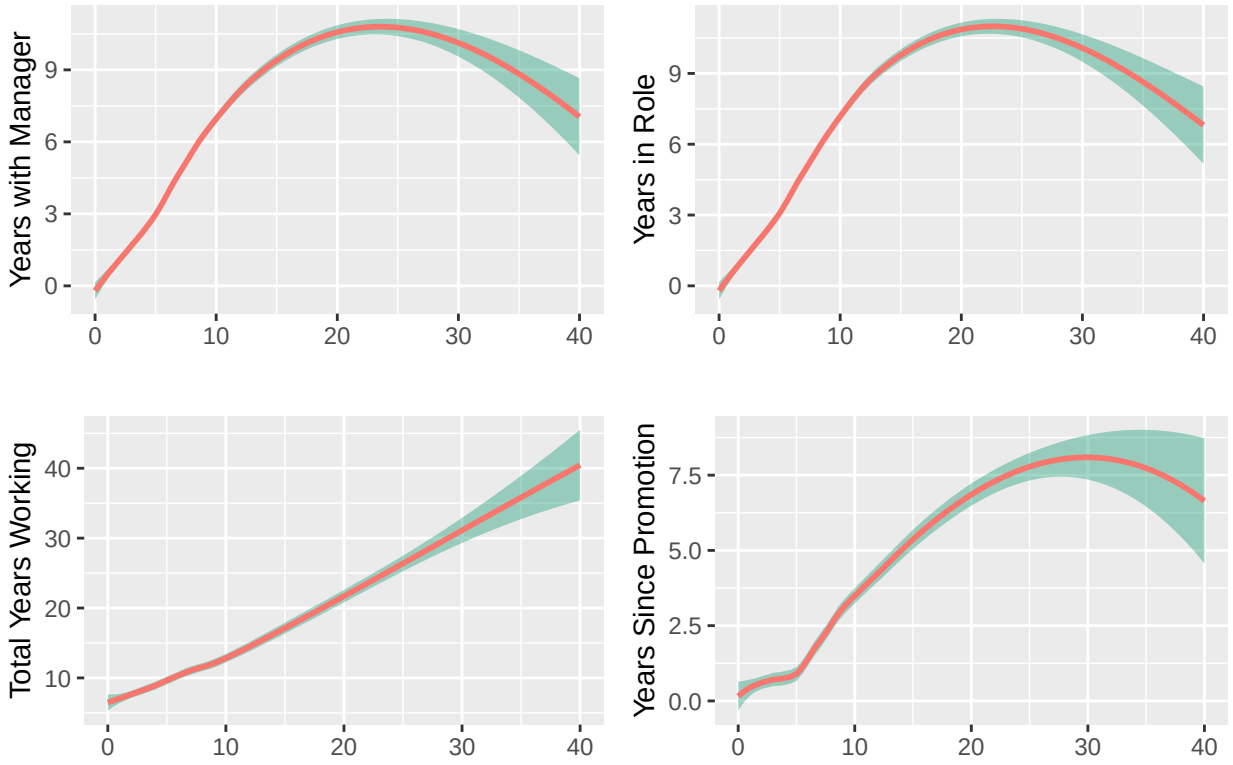
Fig 6. Strongest variables correlation and Job Level



Monthly Income seems to be an almost perfect correlation while the another variables are less correlated. Based on this plots, the variable Job Level will be removed from dataset. Years at Company and Total Years Working will be analized later.

The next variable to be reviewed is Years at Company, which seems to have correlation with another variables, here are the plots:

Fig 7. Strongest variables correlation and Years at Company



The variables Years at Company, Years with Manager, Years in Role and Total Years Working will be also removed.

The structure of the final dataset that will be partitioned is as follows:

```
## 'data.frame':    1470 obs. of  24 variables:
##  $ attrition         : Factor w/ 2 levels "Yes","No": 2 1 2 1 1 1 1 1 1 1 ...
##  $ travel            : Factor w/ 3 levels "Non-Travel","Travel_Rarely",..: 2 3 2 3 2 3 2 2 3 2 ...
##  $ dailyRate         : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ department        : Factor w/ 3 levels "HR","R&D","Sales": 3 2 2 2 2 2 2 2 2 2 ...
##  $ distHome          : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ education         : Factor w/ 5 levels "Below College",..: 2 1 2 4 1 2 3 1 3 3 ...
##  $ educationField    : Factor w/ 6 levels "HR","Mkt","Life Sci.",..: 3 3 6 3 4 3 4 3 3 4 ...
##  $ envSatisfaction   : Factor w/ 4 levels "Low","Medium",..: 2 3 4 4 1 4 3 4 4 3 ...
##  $ gender            : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ hourlyRate        : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ jobInvolvement    : Factor w/ 4 levels "Low","Medium",..: 3 2 2 3 3 3 4 3 2 3 ...
##  $ jobRole           : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...
##  $ jobSatisfaction   : Factor w/ 4 levels "Low","Medium",..: 4 2 3 3 2 4 1 3 3 3 ...
##  $ maritalStatus     : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ monthlyIncome     : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ monthlyRate       : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ numCompaniesWorked: int  8 1 6 1 9 0 4 1 0 6 ...
##  $ overTime          : num  1 0 1 1 0 0 1 0 0 0 ...
##  $ percentSalaryHike : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ relSatisfaction   : Factor w/ 4 levels "Low","Medium",..: 1 4 2 3 4 3 1 2 2 2 ...
##  $ stocOptionLevel   : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ trainingTimesLY   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ workLifeBalance   : Factor w/ 4 levels "Bad","Good","Better",..: 1 3 3 3 3 2 2 3 3 2 ...
```

```
##  $ yearsSinceProm    : int  0 1 0 3 2 3 0 0 1 7 ...
```

The training set will be partitioned to 80% of the sample (1175) and test set to 20% (295). For the algorithm creation, the ensemble will be created using the following methods: glm, lda, naive_bayes, svmLinear, knn, rf, ranger, Rborist, gbm, xgbTree, svmRadial, svmRadialCost, monmlp, kknn, mlp, wsrf, pcaNNet and svmRadialSigma. The method adaboost was also tested, but discarded due after several attempts, the computer crashed.

Results

After training, these are the acurracy values for each model:

	sort.acc..decreasing... TRUE.
glm	0.8711864
svmLinear	0.8711864
lda	0.8644068
xgbTree	0.8610169
gbm	0.8576271
rf	0.8508475
wsrf	0.8508475
monmlp	0.8474576
ranger	0.8440678
kknn	0.8406780
svmRadial	0.8406780
svmRadialCost	0.8406780
svmRadialSigma	0.8406780
naive_bayes	0.8372881
Rborist	0.8372881
pcaNNet	0.8372881
mlp	0.8372881
knn	0.8338983
rpart	0.8305085

The average for each model in the test set and the mean accuracy across all models is 84.71%. After building the ensemble, the accuracy of the model is 84.41%.

After comparing this accuracy to each model, only 8 methods will be used. These models are the following:

x
glm
lda
svmLinear
rf
monmlp
wsrf
gbm
xgbTree

With this models, the new estimated accuracy is the following: 82.1%.

Given this results, only the above mentioned methods will be used for the ensemble. After applying this ensemble to the data, the final accuracy is 86.44%.

Conclusion

An important part of managing work teams is to foster a working environment in which the people can develop their skills and have a strong sense of commitment. Attrition from the company's perspective can represent a high cost, but moreover, it can lead to a lowering in workplace morale, deterating of product or service quality, reduction in investment return, among others unwanted consequences.

Using Data Sciences techniques can be useful in prediction of employee attrition and change the perspective from a reactive to a proactive one.

I think that this kind of tools will become more frequent over the years to come and other applications will be implemented. Applications in problem solving like customer attrition, employee performance, burnout prevention and another will be developed.

For this specific dataset, some algorithms seems to be performed better than the ensemble, however, this is due partition process and random sampling. It would be interesting to evaluate the algorithm using real data.

References

Belli, Gina. 2018. "Here's How Many Years You'll Spend at Work in Your LifeTime." *Pay Scale*. https://www.payscale.com/career-news/2018/10/heres-how-many-years-youll-spend-work-in-your-lifetime.

Campbell, Leigh. 2017. "We've Broken down Your Entire Life into Years Spent Doing Tasks." *Huffington Post Australia*. https://www.huffingtonpost.com.au/2017/10/18/weve-broken-down-your-entire-life-into-years-spent-doing-tasks_a_23248153/.

Gomes Maia, Leticia, and Antonio Virgilio Bittencourt Bastos. 2015. "Organizational Commitment, Psychological Contract Fulfillment and Job Performance: A Longitudinal Quanti-Qualitative Study." *BAR - Brazilian Administration Review* 12 (3). Associação Nacional de Pós-Graduação e Pesquisa em Administração: 250–67.

Mäkikangas, Anne, Wilmar Schaufeli, Asko Tolvanen, and Taru Feldt. 2013. "Engaged Managers Are Not Workaholics: Evidence from a Longitudinal Person-Centered Analysis." *Journal of Work and Organizational Psychology* 29 (3). Colegio Oficial de Psicólogos de Madrid: 135–43.

Pawlewicz, Paul. 2018. "What Is the Difference Between Employee Turnover and Employee Attrition?" https://business.dailypay.com/blog/employee-turnover-vs-attrition.

Stacker IV, McKinley. 2015. "SAMPLE DATA: HR Employee Attrition and Performance." https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/.