

HR Analytics Report

UPx academy certification exam

Manoj Kalamkar | Data Science Super Specialization | June 10th, 2017

# Problem Statement

This is a dataset of employees from some company with following fields:

 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

 Department

 Salary

 Whether the employee has left

\*This dataset is simulated

Objective

Find out why the best and most experienced employees leaving prematurely? Predict which valuable employees will leave next.

Guidelines

Explore and prepare the data

Create training and testing data for the model

Train the model

Test the model

Show or visualize the output

# Exploratory Data Analysis

## Understanding Data

> dim(train\_hr)

[1] 14999 10

> str(train\_hr)

'data.frame': 14999 obs. of 10 variables:

$ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...

$ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...

$ number\_project : int 2 5 7 5 2 2 6 5 5 2 ...

$ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...

$ time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...

$ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...

$ left : int 1 1 1 1 1 1 1 1 1 1 ...

$ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...

$ sales : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...

$ salary : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...

## Univariate Analysis





Observations:-

Many employees joined very recently 2 to 3 years are more

Number of projects employees worked are higher for 3 & 4 projects

About 23% employees left company with given data



Observations:-

Very few employees got promotion in last 5 years. Only around 2%

Seem to be higher work accident





Observations:-

There are more Sales, Technical employees in company than others

Low and Medium salaries are higher than High

Overall employees had less higher than Low/Medium salaries as expected

## Bivariate Analysis

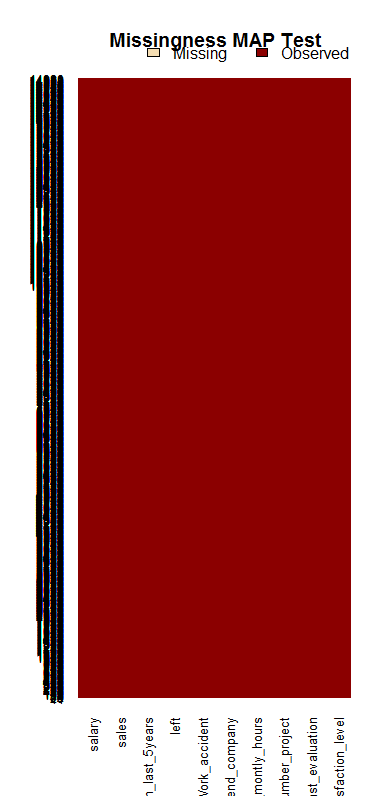
Observations:-

Employees who left seem to have wide range of last evaluation

Satisfaction level seemed to be low for employees who left. This seems to be strong predictor

Employees who left have more time spend in company

# Feature Engineering



Observations:-

Seems to be no missing values

# Build Predictive Models

> summary(results)

Call:

summary.resamples(object = results)

Models: m\_lda, m\_glm, m\_cart, m\_knn, m\_svm, m\_rf, m\_gbm

Number of resamples: 10

Accuracy

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

m\_lda 0.7615 0.7752 0.7795 0.7802 0.7868 0.8010 0

m\_glm 0.7715 0.7855 0.7942 0.7913 0.7987 0.8033 0

m\_cart 0.9000 0.9044 0.9138 0.9203 0.9363 0.9483 0

m\_knn 0.9308 0.9344 0.9391 0.9411 0.9455 0.9558 0

m\_svm 0.9400 0.9419 0.9492 0.9493 0.9556 0.9608 0

m\_rf 0.9842 0.9892 0.9908 0.9906 0.9931 0.9967 0

m\_gbm 0.9667 0.9725 0.9746 0.9746 0.9771 0.9825 0

Kappa

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

m\_lda 0.2263 0.2619 0.2911 0.2899 0.3102 0.3709 0

m\_glm 0.2567 0.3125 0.3354 0.3321 0.3603 0.3724 0

m\_cart 0.6912 0.7104 0.7399 0.7659 0.8301 0.8603 0

m\_knn 0.8119 0.8180 0.8363 0.8389 0.8511 0.8799 0

m\_svm 0.8319 0.8413 0.8603 0.8604 0.8773 0.8907 0

m\_rf 0.9556 0.9698 0.9745 0.9738 0.9810 0.9908 0

m\_gbm 0.9062 0.9229 0.9291 0.9289 0.9361 0.9514 0

Observations:-

Random Forest seems to be standing out as best model given data

# Predict Models with validation set

> confusionMatrix(predict\_rf,val\_hr\_left$left)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 2289 20

1 3 687

Accuracy : 0.9923

95% CI : (0.9885, 0.9951)

No Information Rate : 0.7643

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9785

Mcnemar's Test P-Value : 0.0008492

Sensitivity : 0.9987

Specificity : 0.9717

Pos Pred Value : 0.9913

Neg Pred Value : 0.9957

Prevalence : 0.7643

Detection Rate : 0.7633

Detection Prevalence : 0.7699

Balanced Accuracy : 0.9852

'Positive' Class : 0

Observations:-

Random Forest is giving best accuracy and selected as best model for this set of data

Satisfaction Level turned out to be the best predictor followed by time spend in company, # of projects worked, Avg monthly hours spent and last evaluation

# AUC Curve



Observations:-

Another proof of Random Forest being perfect model for this data