SAS® GLOBAL FORUM 2018

USERS PROGRAM

How to Prevent the Best and Most-experienced Employee Turnover with SAS EM™

April 8 – 11 | Denver, CO

#SASGF

How to Prevent the Best and Most-experienced Employee Turnover with SAS EM™

Liyuan Liu, Lauren Staple, Jennifer Lewis Priestley

Analytics and Data Science, The Graduate College, Kennesaw State University

ABSTRACT

Attrition is a common issue every company faces. Many companies have investment in their employees and as such, are interested in their employee satisfaction and why some employees leave a company. In fact, companies often incorporate surveys as part of their annual review process.

A dataset was simulated by user ludoben (on the Kaggle website) from variables that any normal human resources department would have on their employees.

Our task was to predict which employees may leave from the ten available features. We found that the most at-risk employees for leaving were found in the most extreme regions of each feature.

DATA

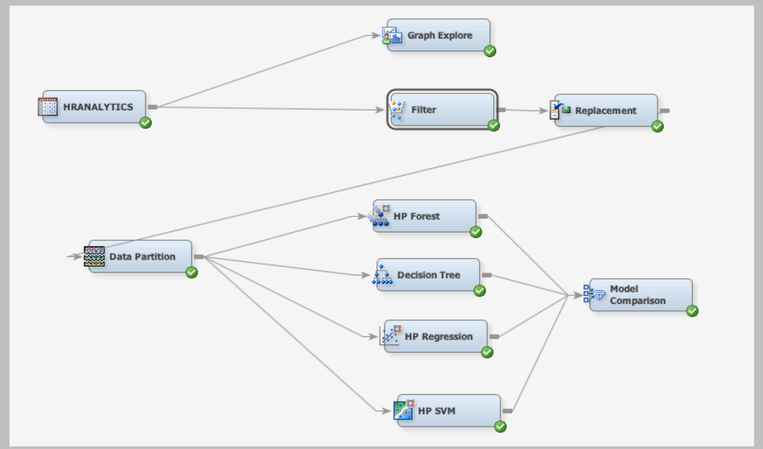
The dataset contained 14999 observations and 10 variables: satisfaction Level, last evaluation, number of project, average monthly hours, time spend company, work accident, promotion last five years, sales(department), salary, left. Left is the target variable, with "1" indicating the employee left.

METHODS OVERVIEW

- The goal was to run several different model approaches, compare the results, and select the model that gave the most accurate prediction on training data.
- First, we checked for missing values, which there were none. Had there been missing data, we'd have to impute them as a lot of the prediction models employed require no missing data.
- Then we performed data exploration, which is necessary to understand the features.
    - Often, patterns are obvious when examining distributions or scatterplots.
    - Collinearity should always be examined, modeling correlated variables will often give non-meaningful results.
- The data was split into two parts, 75 % for training, 25% for validation.
- We examined the following models:
    - A single Decision Tree,
    - A Random Forest (of decision trees),
    - A Logistic Regression Model,
    - And a Support Vector Machine learning (SVM).

WORKFLOW

Graph 1: Workflow. The top part of the workflow is the data exploration, and the bottom half is where the training set and test set are partitioned and the different models are run and compared.



VARIABLES CORRELATION

There does not appear to be any problematic correlation in the continuous variables for any level of salary (low, medium, high). Shown below is the correlation for the "low" level of salary. We took some help from base SAS Studio to create this matrix.

Table 1: Correlation of numeric variables for Salary="Low".

Correlation of Variables for Salary='Low'

The CORR Procedure

Pearson Correlation Coefficients, N = 7316

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years
satisfaction_level	1.00000	0.11107	-0.14036	-0.02785	-0.09710	0.05293	0.00977
last_evaluation	0.11107	1.00000	0.41141	0.38439	0.19020	-0.00700	-0.02072
number_project	-0.14036	0.41141	1.00000	0.46572	0.23341	-0.00455	-0.02684
average_montly_hours	-0.02785	0.38439	0.46572	1.00000	0.17481	-0.01317	-0.01885
time_spend_company	-0.09710	0.19020	0.23341	0.17481	1.00000	-0.01253	0.02448
Work_accident	0.05293	-0.00700	-0.00455	-0.01317	-0.01253	1.00000	0.03568
promotion_last_5years	0.00977	-0.02072	-0.02684	-0.01885	0.02448	0.03568	1.00000

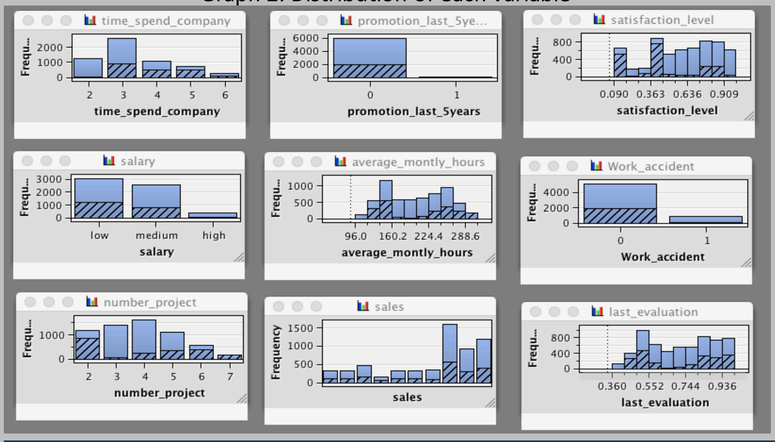
How to Prevent the Best and Most-experienced Employee Turnover with SAS EM™

Liyuan Liu, Lauren Staple, Jennifer Lewis Priestley

Analytics and Data Science, The Graduate College, Kennesaw State University

DATA EXPLORATION

Graph 2: Distribution of each variable



Key:

Employee Leaves: shaded with diagonal lines.

Employee Stays: non-shaded, no lines

CONCLUSION

- The data exploration node gives the most insight in the areas to focus for improving employee retention:
  - Employees generally left when they were underworked and overworked.
  - Employees generally left when they had too few or too many projects.
  - Employees have higher leaving risk after working in the company for 4-5 years.
  - Employees with either really high or low evaluations should be taken into consideration for high turnover rate.
- The top performing model for prediction on “future employees” (aka the validation data) was the random forest model.
- Top 3 important variables: satisfaction level, number project, time spend company.
- An application of this model might be to create risk scores for employees for company use.

The four models were evaluated based on their accuracy and their ROC curves. The model with the highest accuracy on the validation data was the Random Forest model, which also is shown to have the ideal characteristic of not sacrificing false negatives for accuracy.

The tradeoff between false positives and false negatives is situational. In our case, companies may be more concerned with false negatives (predicting an employee will not leave when in fact they will). A false positive is perhaps a less burdensome situation: a company may identify an individual as having a high risk of leaving when they will not, and as a result they simply put more investment or time into that individual.

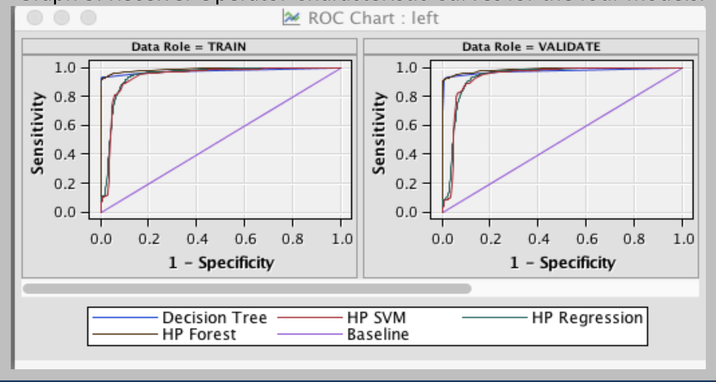
Table 2: Accuracy, False Positive Rate and False Negative Rate for the four models.

Models	Accuracy	False Positive	False Negative
Random Forest	98.8%	0.4%	8.9%
Decision Tree	97.7%	0.6%	8.2%
Logistic Regression	94.2%	7%	19%
SVM	93.5%	8.1%	19%

Table 3: Important Features

Variable Name	Number of Splitting Rules	Train: Gini Reduction
time_spend_company	535	0.060563
REP_satisfaction_level	516	0.082363
number_project	454	0.078273
REP_average_montly_hours	392	0.048593
REP_last_evaluation	332	0.033366
salary	151	0.001751
Work_accident	106	0.001612
sales	66	0.000380
promotion_last_5years	8	0.000082

Graph 3: Receiver Operator Characteristic Curves for the four models.



REFERENCE

Dataset titled “Human Resources Analytics” was created by user “ludoben” to the Kaggle website. Data was used in accordance with license CC BY-SA 4.0. Retrieved 10/8/2017 from url https://www.kaggle.com/ludobenistant/hr-analytics and used with no modifications.

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF