**OPIM 5894**

**Predicting Employee Attrition**

Semester: Fall 2017

Project due: 12/10/2017

Instructor: Jing Peng

Sub: 5894 - Data Science Using Python

Team Members:

Swaraj Machiraju

Sthiti Pragnya Mishra

Ravindra Kollapu

Constantine Macris

Shuge Li

Yunong Liu

|  |
| --- |
| **[**TEAM: PySparks**]** |
| HR Analytics |

Contents

[Objectives 3](#_Toc500714664)

[I. Data pre-processing 4](#_Toc500714665)

[II. Data visualization and pattern discovery 5](#_Toc500714666)

[III. Statistical Overview 8](#_Toc500714667)

[IV. Feature Selection 10](#_Toc500714668)

[V. Predictive Modeling 10](#_Toc500714669)

[1. Running Models 10](#_Toc500714670)

[2. Confusion Matrix and Classification Report 11](#_Toc500714671)

[3. ROC Curve of Models 11](#_Toc500714672)

[VI. Findings 13](#_Toc500714673)

SUMMARY

This project was conducted on the dataset generously provided by Kaggle on HR Analytics-Attrition - Why people leaving the organization?

The Goal is to analyze the attrition level for a company i.e. why the best employees leave the company prematurely. The factors contributing or leading an employee to leave this organization, will be recorded and summarized in the paper. The analysis can serve any company to deal with the losses incurred due to attrition.

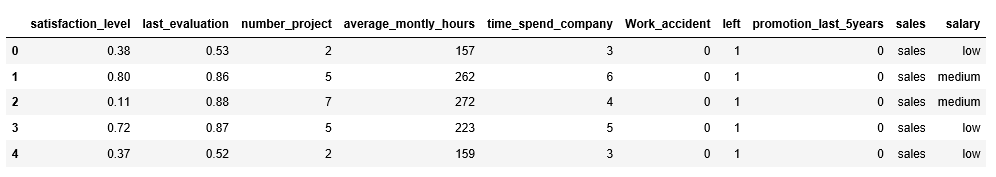
# Objectives

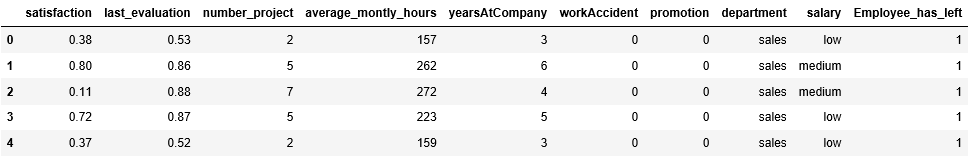
1. Data pre-processing
2. Data visualization and pattern discovery
3. Statistical Overview
4. Feature Selection
5. Predictive Modeling
6. Model Implementation
7. Recommendations

# Data pre-processing

The dataset used was a simulated set from Kaggle and included 14999 records with 9 fields listed below:

* **Satisfaction\_Level** 🡪 Continuous 0-1. Column renamed to **satisfaction**.
* **Last\_Evaluation** 🡪 Continuous 0-1
* **Number Project** 🡪 Continuous
* **Average Monthly Hours** 🡪 Continuous
* **Time\_spend\_company**🡪 Continuous. Column renamed to **yearsAtCompany.**
* **Work\_Accident** 🡪 Binary. Column renamed to **workAccident**
* **Left** 🡪 Binary. Column renamed to **Employee\_has\_left**.
* **Promotion\_last\_5years** 🡪 Binary. Column renamed to **promotion**.
* **Department** 🡪 Categorical
* **Salary** 🡪 Categorical H-M-L

Columns before renaming: 

Columns after renaming: 

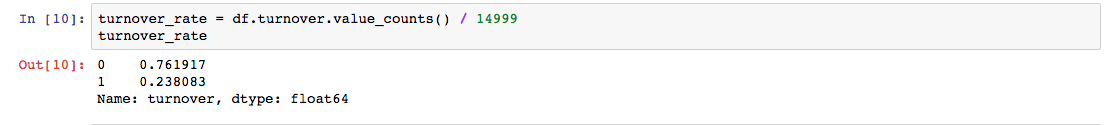
The target/response variable, also called as dependent variable, is “Left” or “turnover” and 9 independent variables to predict the response variable.

First, we checked for any null or missing values in the dataset, and we found there were no null or missing values.

Then we checked for outliers and there also we didn’t find any. So, moving ahead with our clean data.

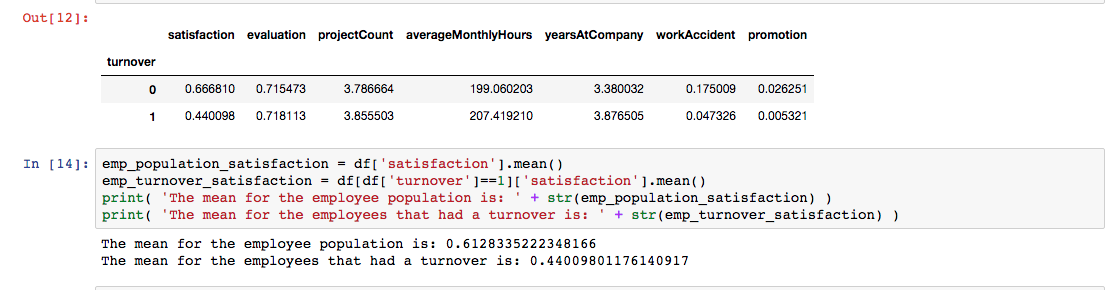
# Data visualization and pattern discovery

Upon investigating the target variable, we find that dataset is not ideally balanced. The dataset has only 23% turnover rows.



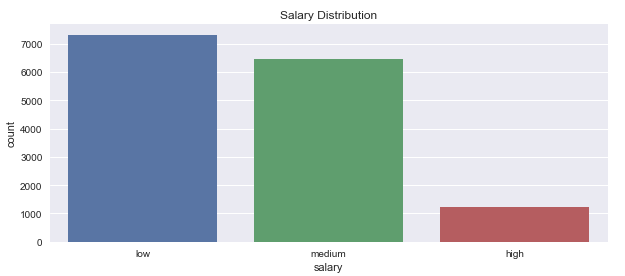
Proportion of target variable values

Now calculating the proportion of 0s and 1s in each column, the data is as below:



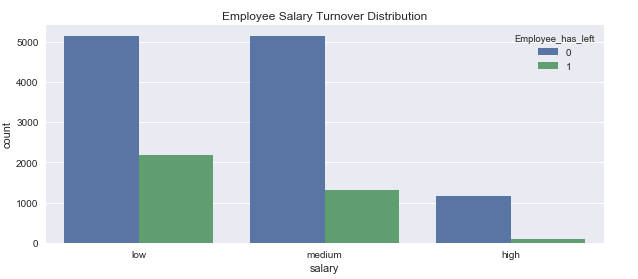
Proportion of target variable values for each column

Distribution of salary across the company:



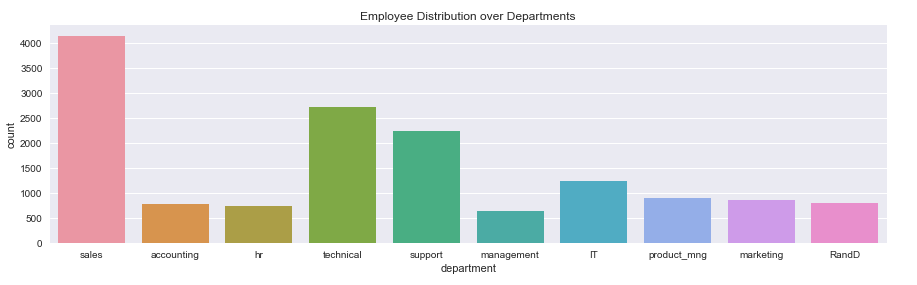
From above figure, it can be inferred that very few employees get high salary.

Exploring the attrition rate with respect to salary:

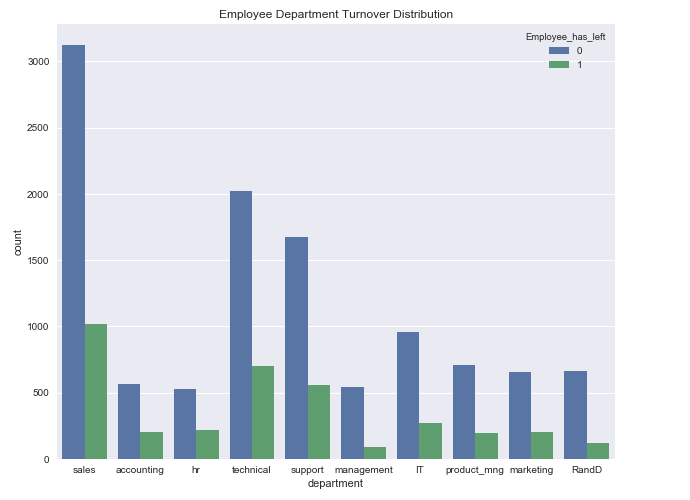


From the above figure, it can be inferred that employee getting low salary tend to leave the company more.

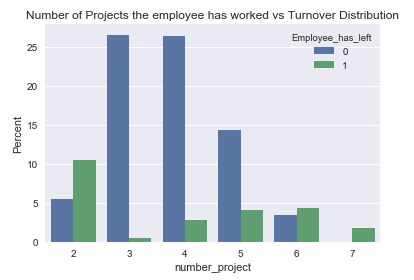
Exploring number of departments in the company:



Exploring the attrition rate with respect to department:

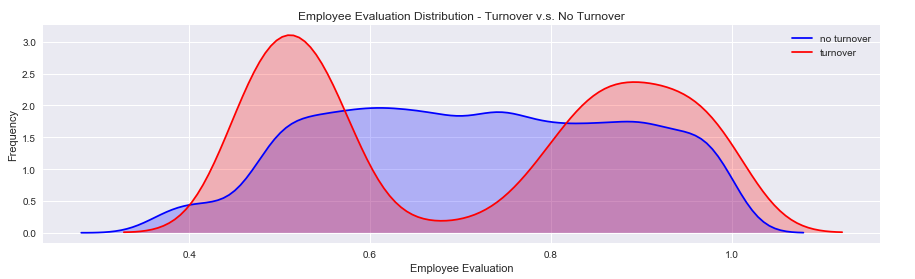


Exploring the turnover rate wrt. Number of project the employees have worked:



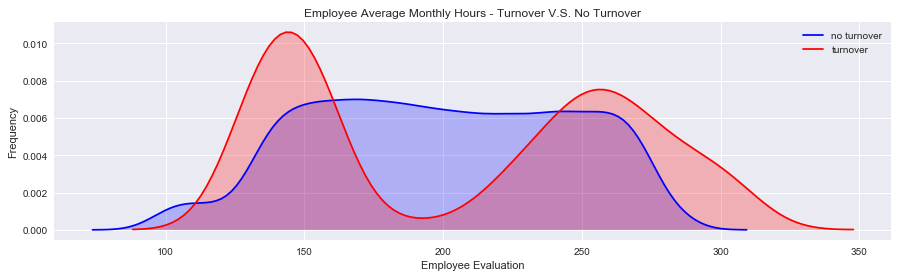
From above figure, it can be said that employees working on 3-5 projects, generally don’t leave the company.

Exploring the turnover rate wrt. Evaluation ratings of the employees:



From above figure, it can be inferred that employees getting very low or very high evaluation tend to leave the company.

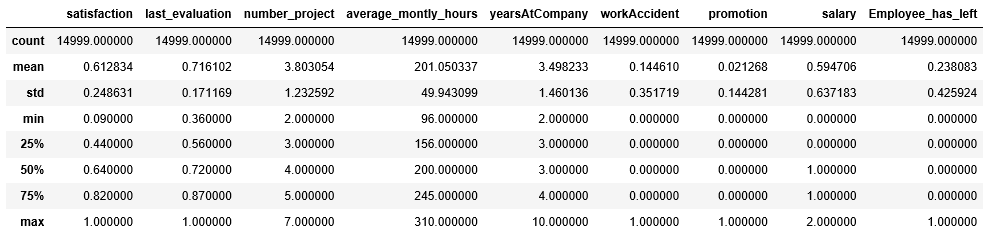
Exploring turnover rate wrt. Working hours of employees:



From above figure, it can be inferred that employees working for very less or too many hours, the attrition increases.

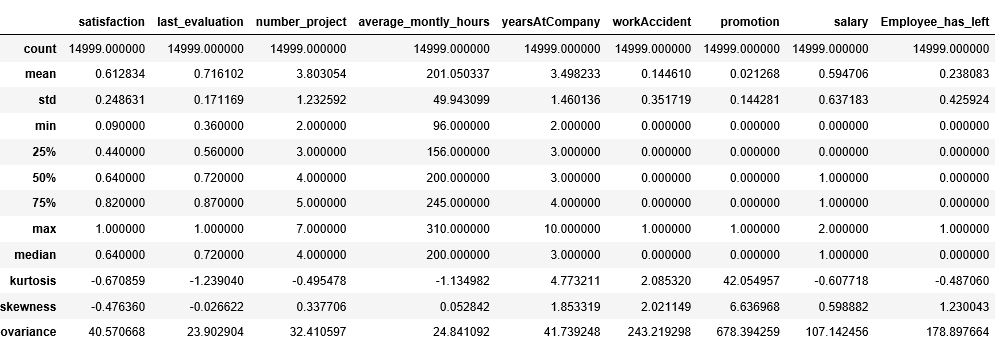
# Statistical Overview

Drawing a statistic of the data distribution:

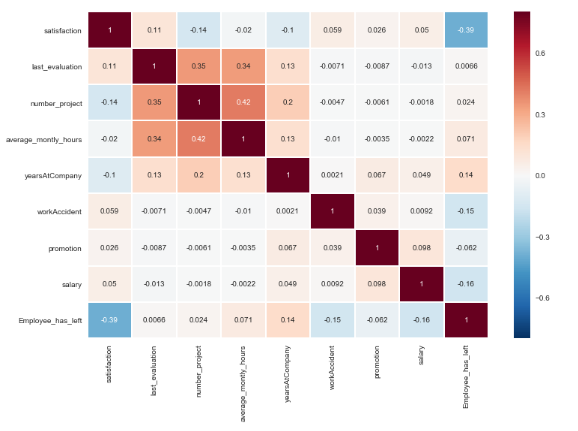


From the above, it can be inferred that the mean and median are almost equivalent, hence, can be said the data is normally distributed.

Drawing the skewness & kurtosis of the data:



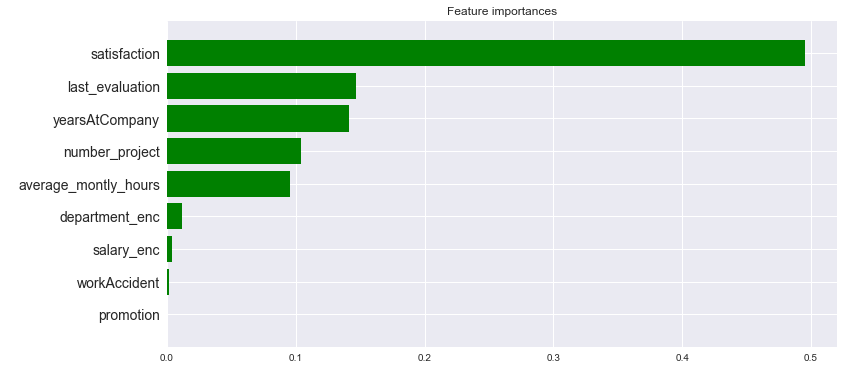
Drawing the correlation matrix of the independent variables:



From the above figure, it can be inferred that there doesn’t exist a high correlation between the independent variables.

# Feature Selection

Using decision tree when we categorize and assess the contributing columns, we find the following columns to bring in major changes. The columns are shown in decreasing order in terms of contribution for prediction:



From the above figure, it is evident that satisfaction of an employee plays a key role if the employee will leave the company or not. It is followed by last\_evaluation, yearsAtCompany, number\_project, average\_monthly\_hours. This means, if we focus on these attributes, we can predict precisely whether an employee will leave the company or not.

# Predictive Modeling

As stated in the Summary, we aim to predict the Employee Attrition of a company. We believe that in predicting these factors we might be able to help the Company to anticipate which employees will have a high risk of leaving the company. We may not completely predict the attrition rate with 100% accuracy, however, we would choose the best model which can help save company’s huge finance by predicting with greater accuracy.

# Running Models

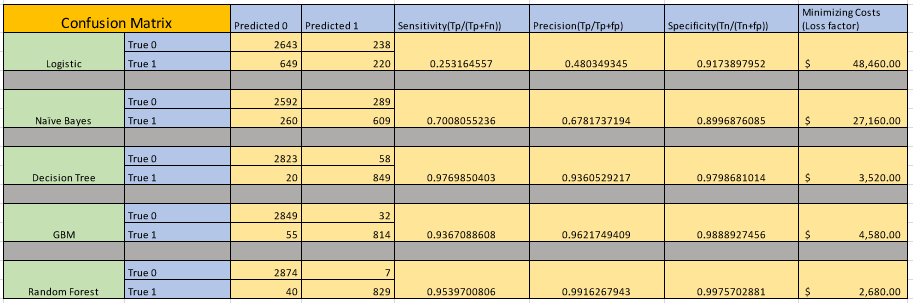
We introduced Prediction Modeling by running the following:

1. Logistic Regression Model
2. Naïve Bayes Model
3. Decision Tree Model
4. GBM Model
5. Random Forest Model

Two different data subsets were used: 1. Training, 2. Testing. For our conclusion and decision, we used the results from the Testing Dataset only.

# Confusion Matrix and Classification Report

After running the models, we build the confusion matrix and the classification report. The same is as below:

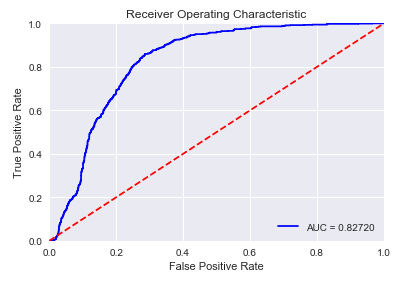


From the matrix report above, we can conclude that Random Forest Model predicts the number of employees who would be leaving the company more accurately with less percentage of misclassification. We have calculated the cost matrix for the same, which can help us decide which model to adopt. Thus, with the help of cost matrix, we are selecting the Random Forest model.

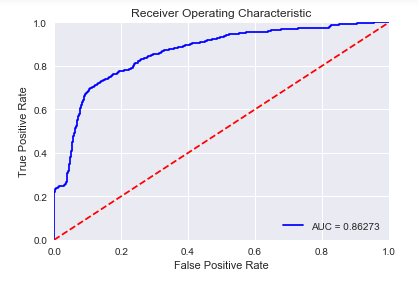
# ROC Curve of Models

The ROC curve of different models is as below:

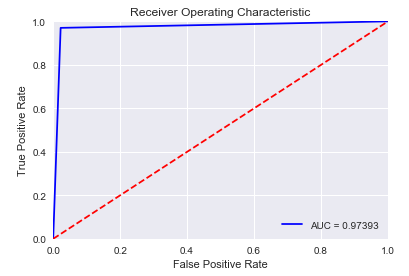
1. Logistic Model



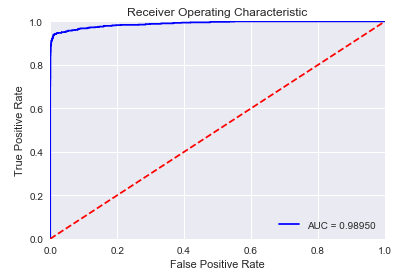
1. Naïve Bayes Classifier Model



1. Decision Tree Model



1. GBM Model



1. Random Forest Classifier Model



# Findings

As one could expect one of the most important ways to combat employee attrition is to keep your employees happy. Using the defined models high risk employees can be identified and provided incentives to keep them This is especially true for the group of high performing, highly satisfied employees that leave. These employees can then be focused on because they are high risk.

The top 5 important metrics are satisfaction, evaluation, years at company, number of projects and average monthly hours need to be considered with the top value being satisfaction. As discussed above it is critical to understand where this value is coming from since it is of high important. The data scientist needs to be confident as to how authenticate satisfaction is. A supervisor could assign a satisfaction rating; however this value may have errors due to the supervisor not fully understanding her employees. It would be suggested that this metric be broken out to what contributes to employee satisfaction with some rational to how the metric is developed.

With a good understanding of employee attrition and a model that can aid in prediction the management can now apply adequate motivation to employees to help keep high value employees. As this model is tested it can be refined and assist management in making better decisions and providing probabilistic interpretation to management in order to support deterministic decision making as it pertains to employee retention.

**References:**

<https://www.kaggle.com/randylaosat/predicting-employee-kernelover>

<https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

<https://en.wikipedia.org/wiki/Gradient_boosting>

<https://en.wikipedia.org/wiki/Random_forest>