Employee Attrition Prediction

Vinisha   
Post Graduation Program in Data Science  
Praxis Business SchoolBangalore, Karnataka- 560102  
vinisha.maurya@gmail.com

*Problem Statement* — The company wants to use HR analytics to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees.

Keywords—Neural Network, Random Forest, Stacked Model

# Introduction

Employee attrition in any company is very serious problem. To maintain stability in the company and for smooth workflow, it is mandatory for the HR team to look out for the reasons affecting employee attrition.

Here is the detailed study on the variables affecting employee attrition and which model can help company to predict attrition better.

Steps involved should be:

a) Data Import (Target variable is "Attrition" column)  
b) Split the data in Dev & Hold Out sample (70:30)  
c) Perform Exploratory Data Analysis  
d) Identify columns which are of no use. drop those columns  
e) Write Hypothesis and validate the Hypothesis  
f) Build Neural Network Model (Development sample)  
g) Validate NN model on Hold Out. If need be improvise  
h) Build Random Forest Model  
i) Validate RF Model  
j) Compare NN with RF  
k) Combine NN and RF into Ensemble Model  
l) Check whether Ensemble Model Performance outperforms the individual RF & NN model

# Data Discription

Dataset consists of:

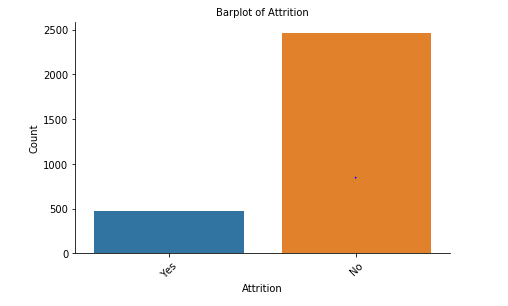
1. 35 variables and 2940 observations.
2. 9 Categorical variables and 26 Numerical variables.
3. Target Variable: ‘Attrition’.
4. No missing values.

# Exploratory data Analysis

## Univariate Analysis

Univariate analysis is done to check the distribution.

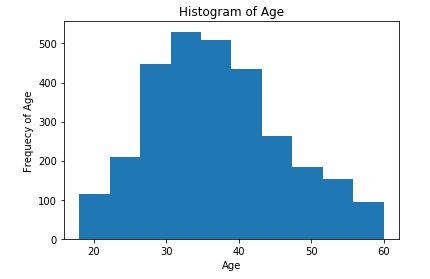
## Target Variable (Attrition):

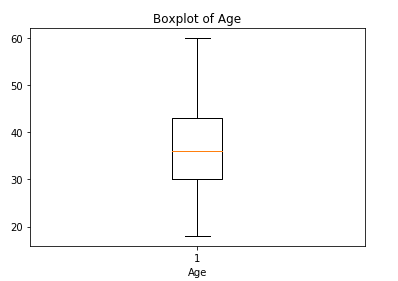


* The Yes : No ratio is 1:5. So, Data balancing is not required.
* Distribution is Yes – 84% and No – 16%.

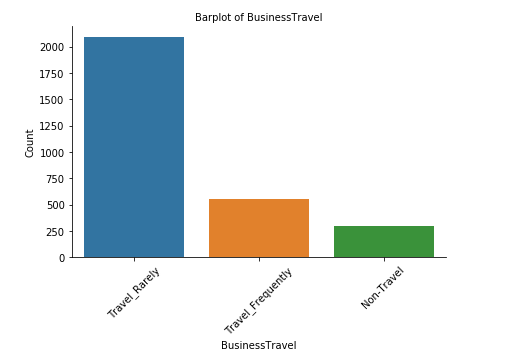
1. Age:

Age is normally distributed with no outliers.

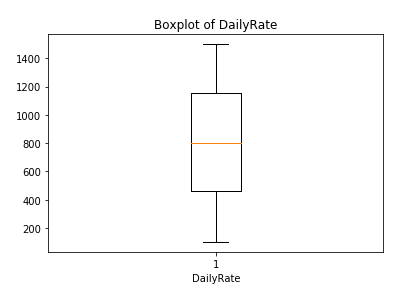




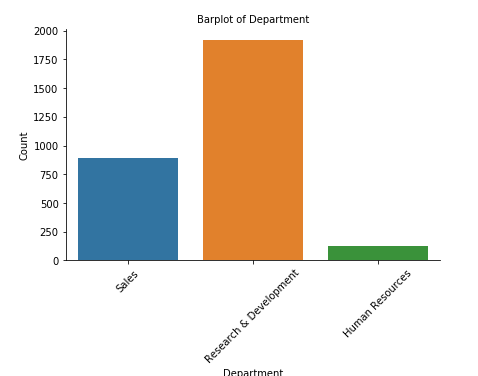
1. Business Travel:



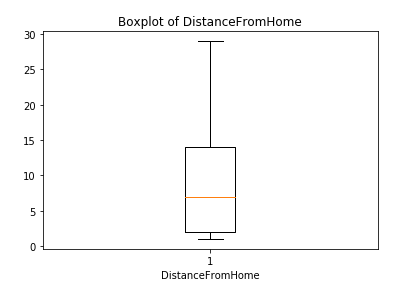
1. Daily Rate:



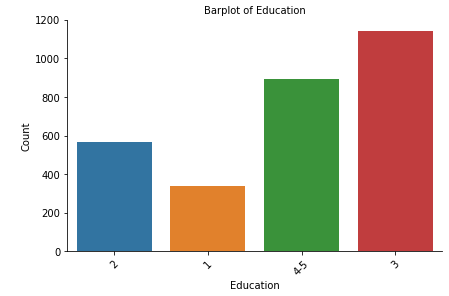
1. Department :



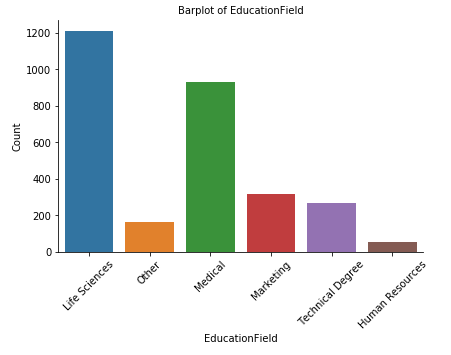
1. Distance from home:



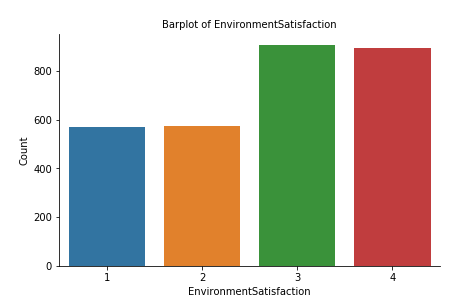
1. Education:



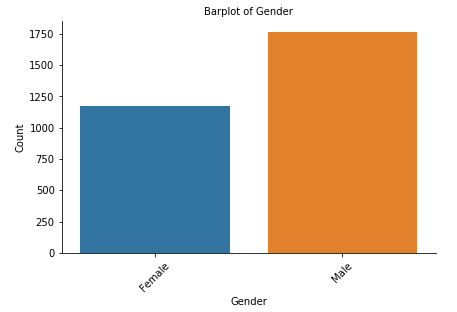
1. Education Field:



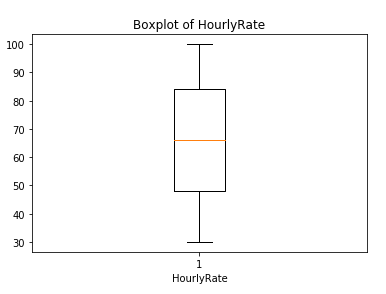
1. Environment Satisfaction:



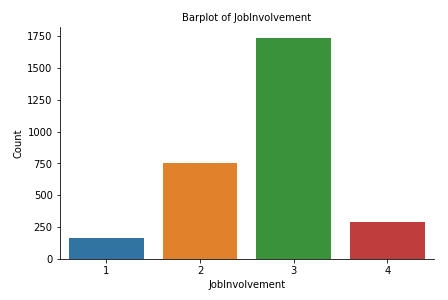
1. Gender:



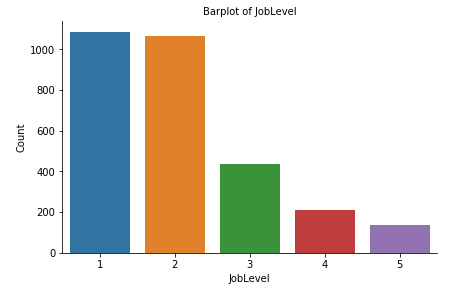
1. Hourly Rate:



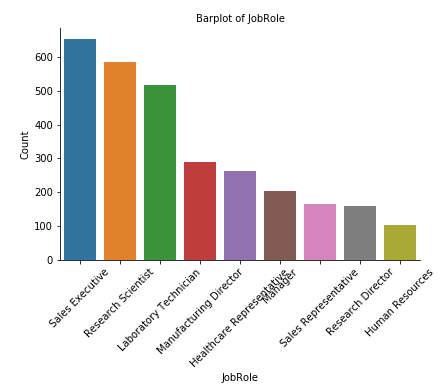
1. Job Involvement:



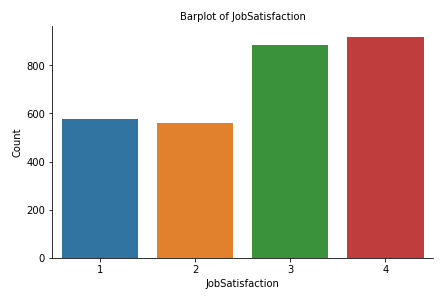
1. Job Level:



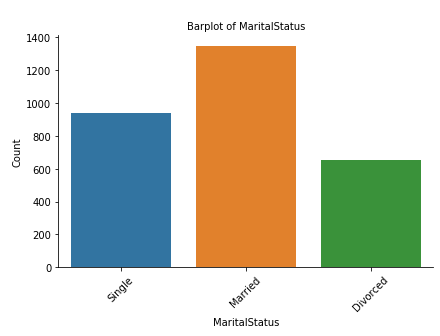
1. Job Role:



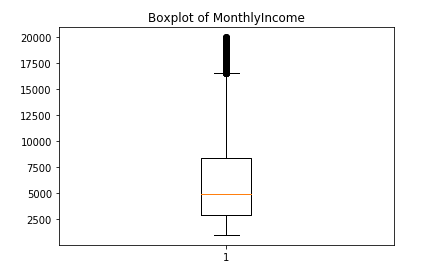
1. Job Satisfaction:



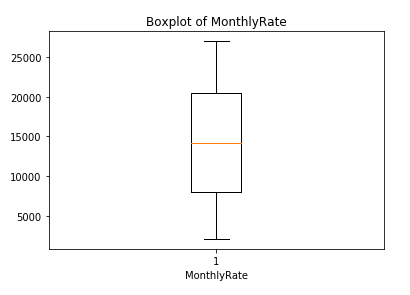
1. Marital Status:



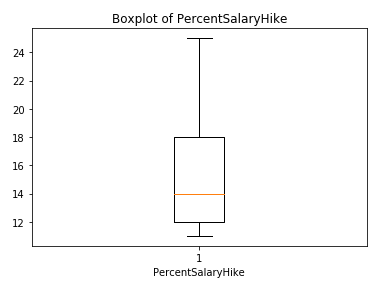
1. Monthly Income:



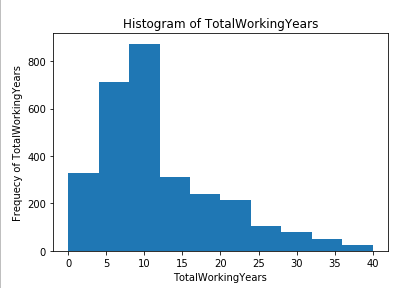
1. Monthly Rate:

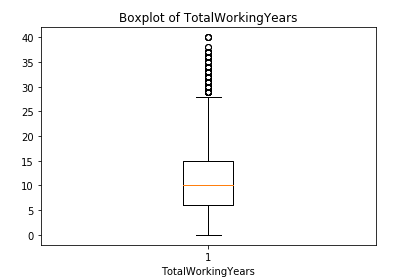


1. Salary Hike:



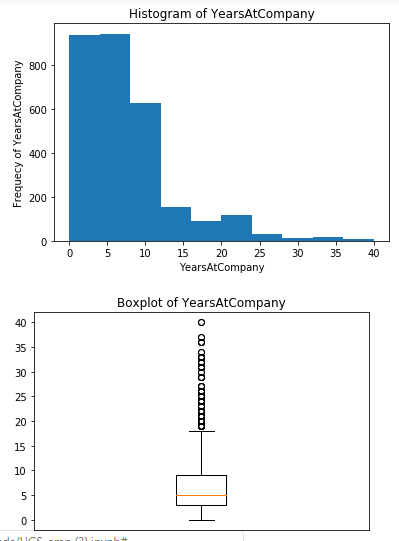
1. Total working years:



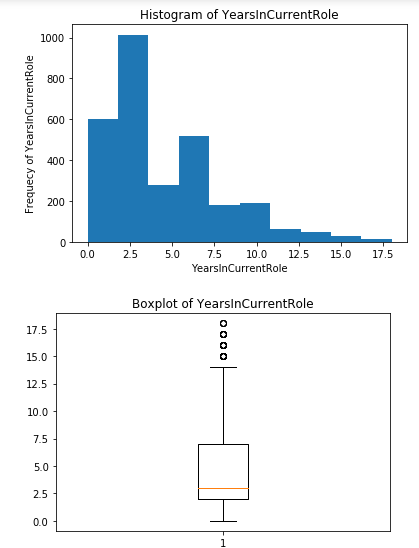


1. Years at company:

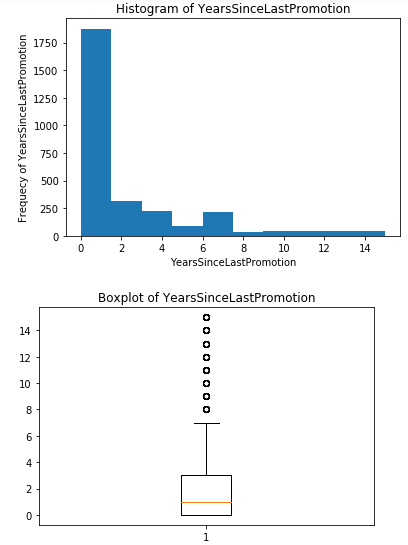
Years at company is right skewed with



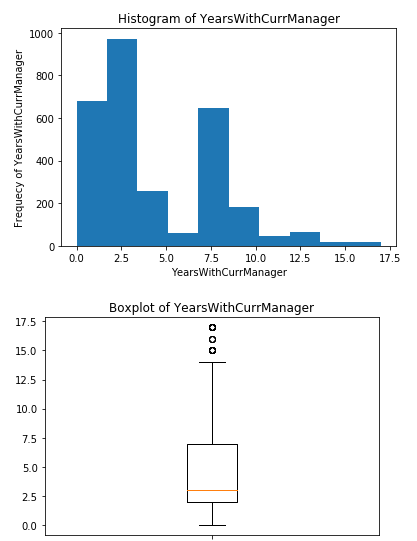
1. Year in Current role:



1. Years since last promotion:



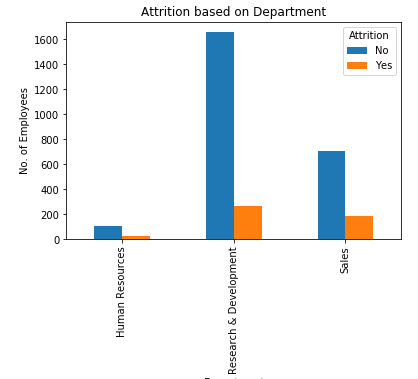
1. Years with current manager:



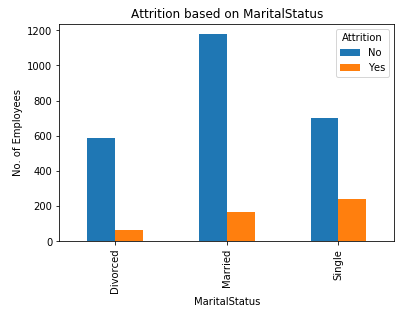
## Bivariate Analysis:

In bivariate analysis, the distribution of attrition with respect to other variables can be seen.

1. Attrition on different Departments:



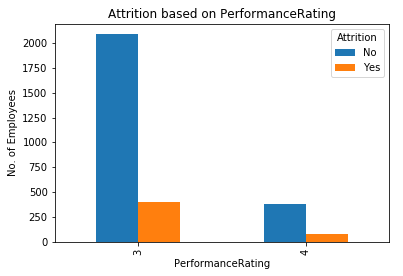
1. Attrition in different Marital status:



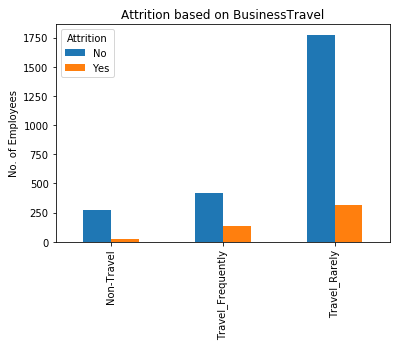
1. Attrition in different Satisfaction level of Jobs:



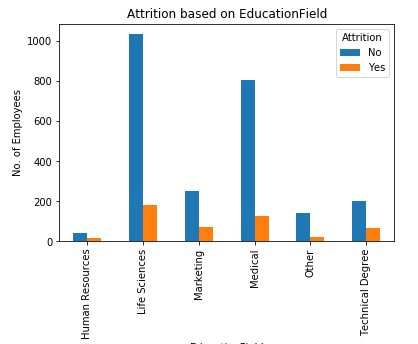
1. Attrition in different performance Rating:



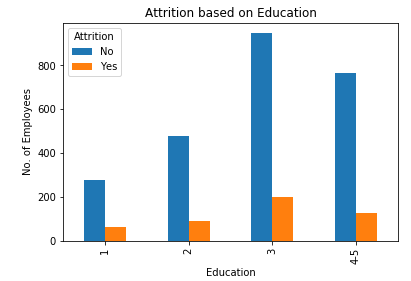
1. Attrition in type of Business Travel:



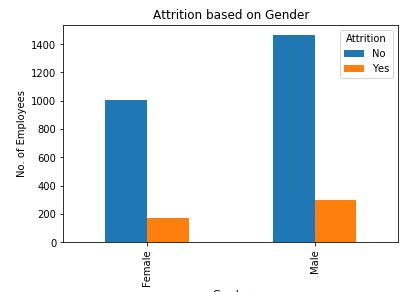
1. Attrition on different Educational Field:



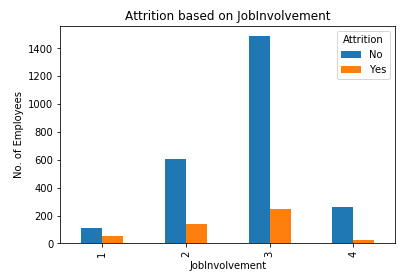
1. Attrition in different level of Education:



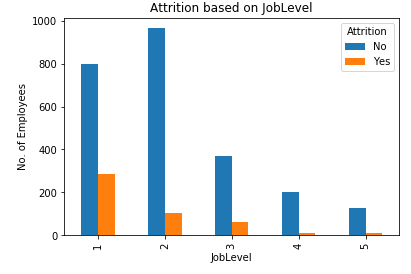
1. Attrition in Gender:

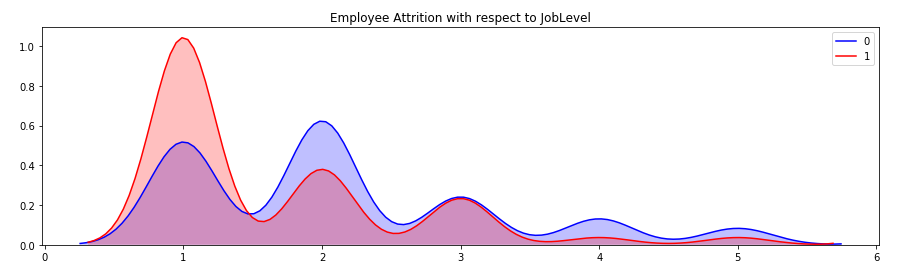


1. Attrition in different Job Involvement:

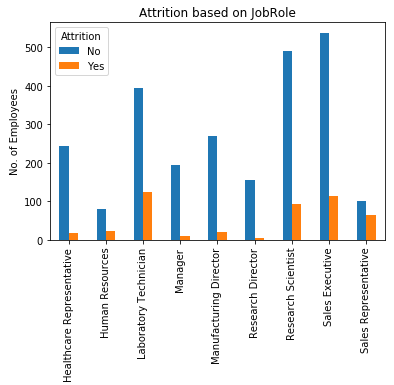


1. Attrition in different Job level:

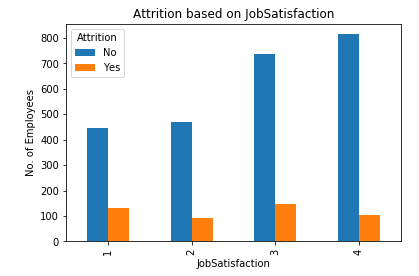




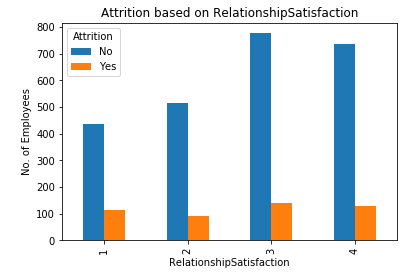
1. Attrition in Job Role:



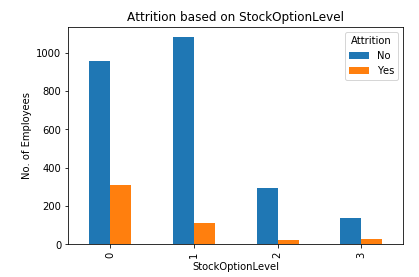
1. Attrition in different rating for job satisfaction:



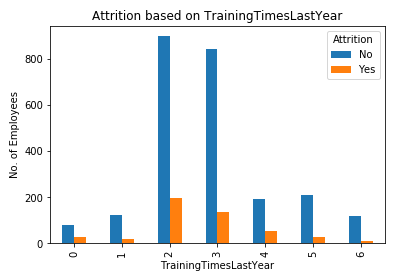
1. Attrition in different rating for Relationship Satisfaction:



1. Attrition in Stock Option Level:



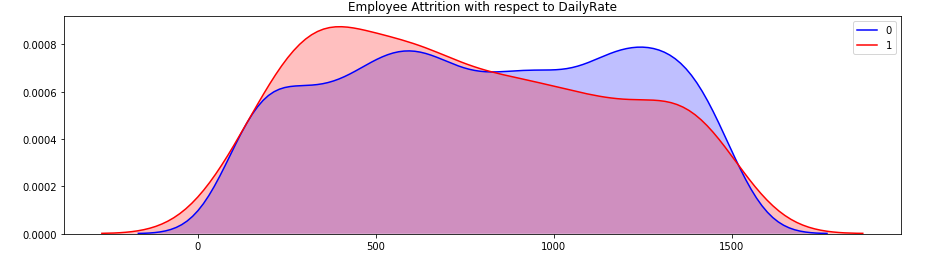
1. Attrition in Training times last year:



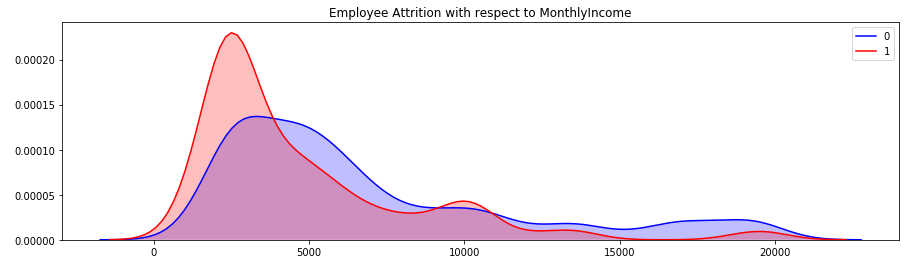
*KDE plot output:*

Kernel density estimation (**KDE**) is a non-parametric way to estimate the probability density function (PDF) of a random variable. This function uses Gaussian kernels and includes automatic bandwidth determination.

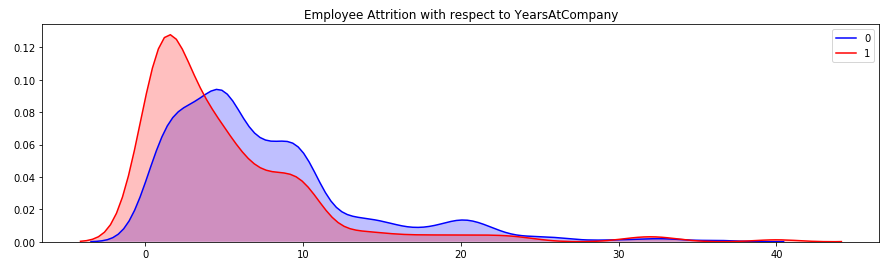
1. Attrition with respect to Daily Rate:



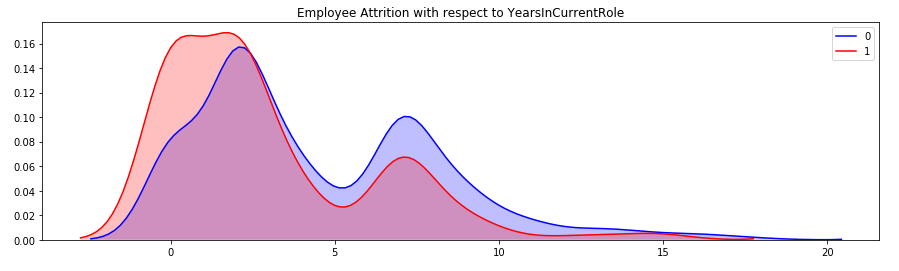
1. Attrition with respect to Monthly Income:



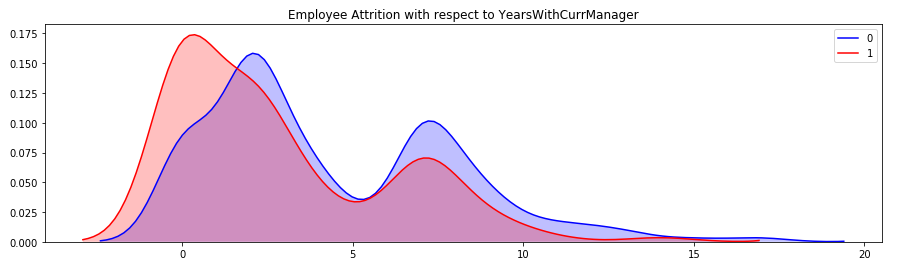
1. Attrition with repect to Years in Company:



1. Attrition with respect to Years in Current Role:



1. Attrition with respect to Years with Current Manager:



## Identifying Columns to Drop:

* Employee Count
* EmployeeNumber
* Over18
* StandardHours

# Hypothesis

1. JobSatisfaction, EnvironmentSatisfaction, Age, DailyRate, DistanceFromHome, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager contribute to Attrition in the company. (Performing T-test for the same to check the difference in the mean of the population who churned is different from overall mean.
2. Overtime is the important factor contributing towards attrition.
3. Ensembling different models should improve accuracy. (Checking this after building all the models)

## Performed one sample T-test on some factors that can cause Attrition:

A one-sample t-test checks whether a sample mean differs from the population mean.

**Hypothesis Testing:** Is done to check if there significant difference in the **means of variables** between employees who had a churned and employees who had not.

* **Null Hypothesis:** (H0: pTS = pES) The null hypothesis would be that there is **no** difference in satisfaction level between employees who did churned and those who did not.
* **Alternate Hypothesis:** (HA: pTS != pES) The alternative hypothesis would be that there **is** a difference in variable between employees who did turnover and those who did not.
  1. **Attributes contributed to attrition**

After performing T-test, it can easily be stated that JobSatisfaction, EnvironmentSatisfaction, Age, DailyRate, DistanceFromHome, MonthlyIncome, NumCompaniesWorked, RelationshipSatisfaction, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsWithCurrManager contribute to Attrition in the company.

All of the above attributes have p-value less than 0.05 and t-statistic beyond the 95% confidence interval. Hence, there is significance difference in the mean of above mentioned attributes for those whose attrition has been marked ‘Yes’. So, they may contribute to attrition in the company.

* 1. **Attributes do not contributed to attrition**

MonthlyRate, PercentSalaryHike, PerformanceRating, YearsSinceLastPromotion, these are some attributes that does not seem to contribute to attrition.

There t-statistic lie within the 95% confidence interval, so we accept the Ho, that means there is no significance difference in there mean.

One reason for above situation could be, there is not much data to generate insight for the population about attrition.

# **Model Building**

## Neural Network:

Steps involved:

1. Dummy Variable Creation for categorical data.
2. Feature Scaling
3. Parameter Tuning to get the best parameter:

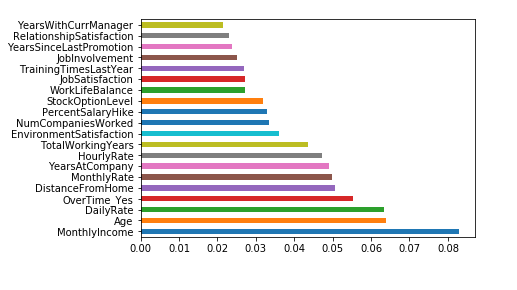
Building the NN model from the best parameters obtained above which is epoch-200, batch\_size-5, and optimizer-adam

1. Neural Network model building
2. K-fold NN model building, so that model can learn better and better accuracy is obtained.
3. Storing the prediction in a new variable [‘M1’] for both train and test data.

## Random Forest:

Steps involved:

1. Random forest classifier is trained on trained data.
2. Prediction is made on test data and stored in a new variable [‘M2’]
3. Accuracy is obtained.
4. K-fold Random forest is build. Prediction for the trained data is stored in training dataframe as [‘M2’].
5. Key Features obtained from Random Forest:



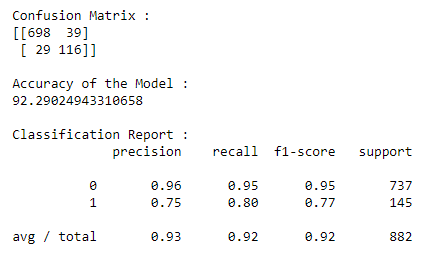
## Stacked model is builed to combine the result of NN and Random Forest and Ensembled model is used (XGBoost):

Steps involved:

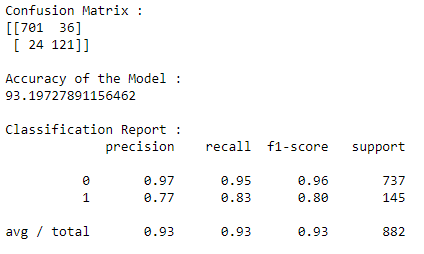
1. A new DataFrame X1 is created using Train[‘M1’, ‘M2’] created from above two models.
2. X1 now severs as the input variable to the ensemble model XGBoost.
3. XGboost model is trained on X1 data with Y\_train data created earlier.
4. Prediction is made on dataset test [‘M1’, ‘M2’].
5. Accuracy is calculated.

# **Model Comparision**

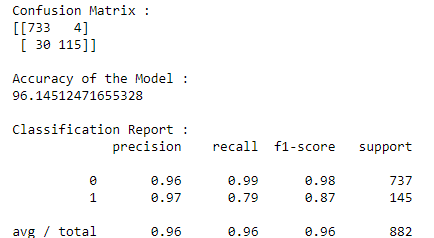
1. Accuracy of one Neural Network model:



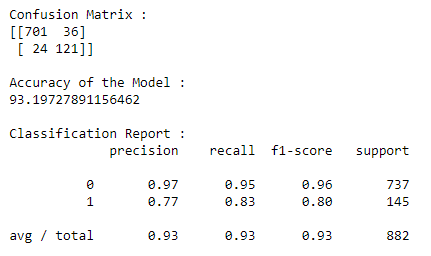
1. After improvisation of NN Model using k-fold the accuracy increases:



1. Random Forest accuracy:



1. Stacked Model accuracy:



The Stacked model didn’t performed well because, stacked model works well when we have strong base learner like NN. So, stacked classifier here, is preferring

NN model over Random Forest model. Hence the result are same as NN model.

From all above model, it is clearly visible that Random forest has out-performed NN and Stacked ensemble model.

So, our third hypothesis that Ensemble model will perform better is rejected, as it performed poor that Random Forest classifier.

# **Insights and conclusions**

1. From NN model and Random Forest model, Random Forest is performing with accuracy 96%.

2. Ensemble of NN and Random Forest is not performing good rather it is just imitating NN model. From all three models Random Forest is performing best.

3. Hypothesis:

1. Overtime is contributing to Attrition is validated from Feature importance model of Random Forest.

2. Ensemble model fails to provide a good model, this is validated from accuracy comparison of all three models.

4. Insights:

a. Monthly Income, Age, Overtime, Daily Rate, Distance from home are some of the key important features affecting Attrition.

b. Company should focus on these parameters to avoid high attrition rate in the company.

c. With this dataset Percentage salary hike,PerformanceRating, YearsSinceLastPromotion is not helping predicting attrition. May be larger dataset would help in this situation.