Data Analysis

IBM Employees Dataset

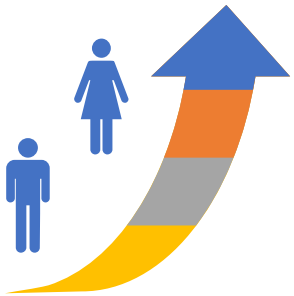
Warat Chokulket

Introduction

Introduction

An employee is the power of the company like an engine in a car. High potential employees mean high company capability. In order to obtain high potential employees, company need to invest a lot of time and money. If an employee you have invested so much time and money leaves, this would mean that you would have to spend even more time and money to hire somebody else.

Problem



The problem that we are looking to solve is …

01 What is the factor that lead to employee attrition?

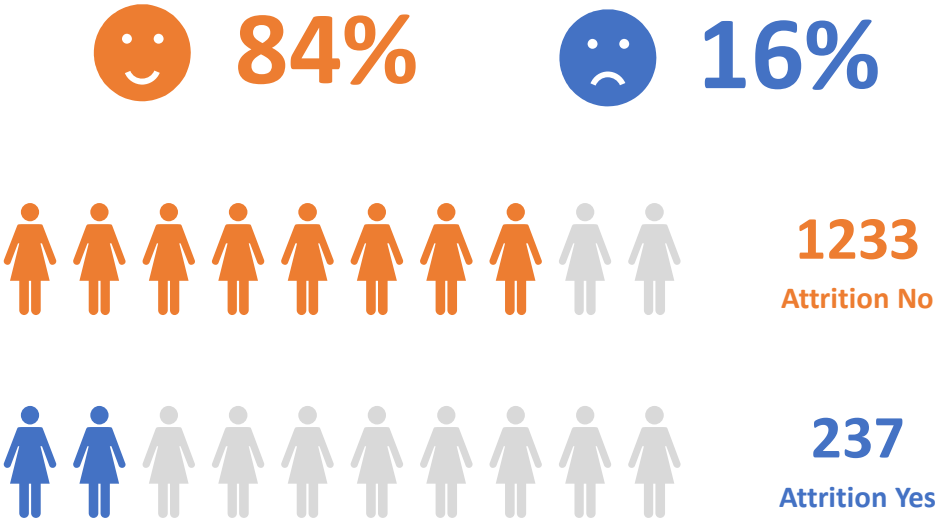
02 How can we keep employees happy and satisfied to company?

Our target is … **Attrition**

Exploratory Data Analysis

Exploratory Data Analysis

Dataset contained 1470 samples and 35 columns

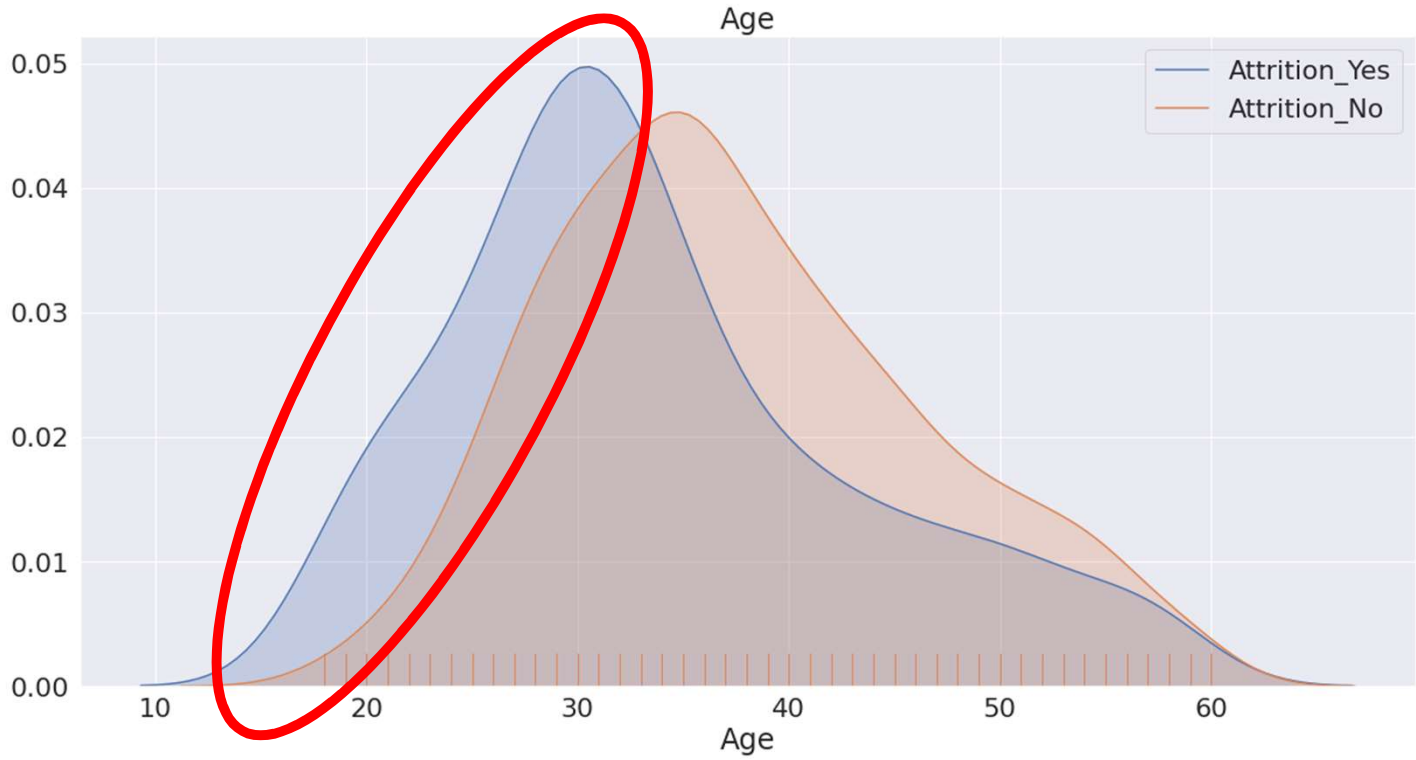


To find insight, we use KDE plot to show distribution of each features and display some interesting features.

KDE Plot

Age

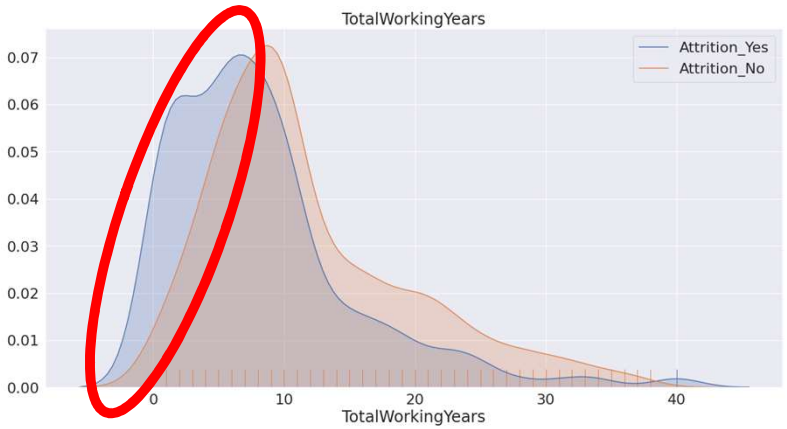
Age distribution is nearly the same for our target but the left blue area shows that younger people are more likely to quit.

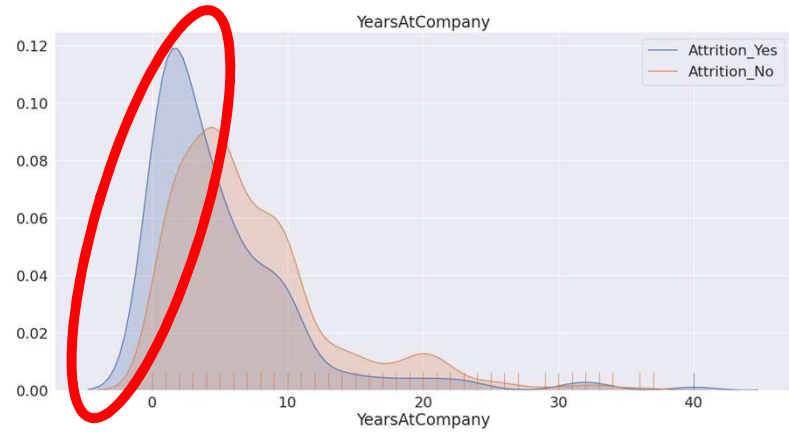


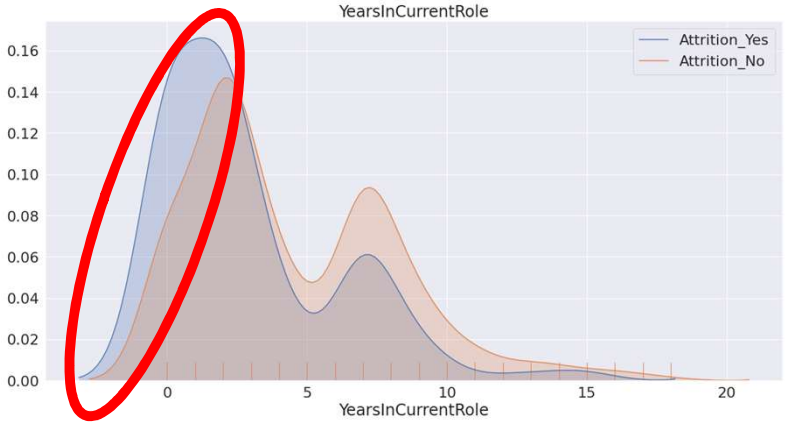
KDE Plot

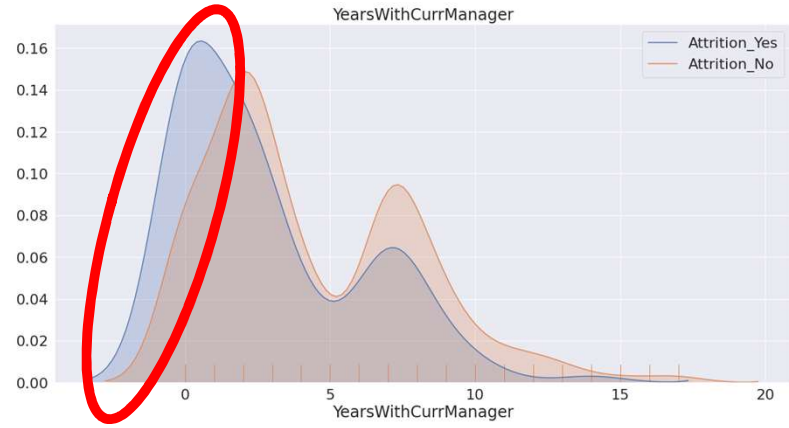
This 4 features are related to one another and related to Age feature too.

- **Total working years**
- **Years at company**
- **Years in current role**
- **Years with current manager**



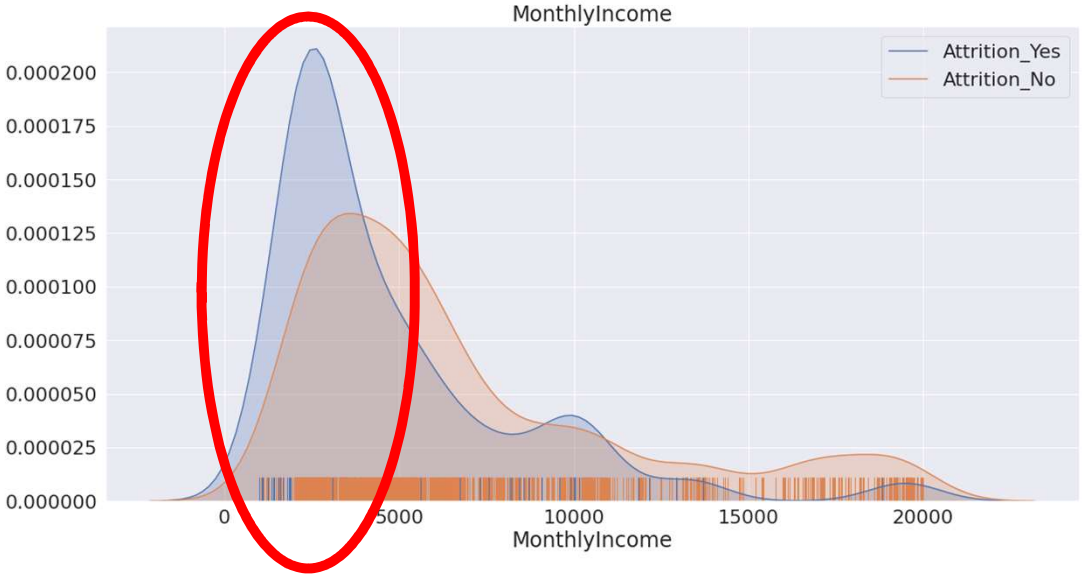






KDE Plot

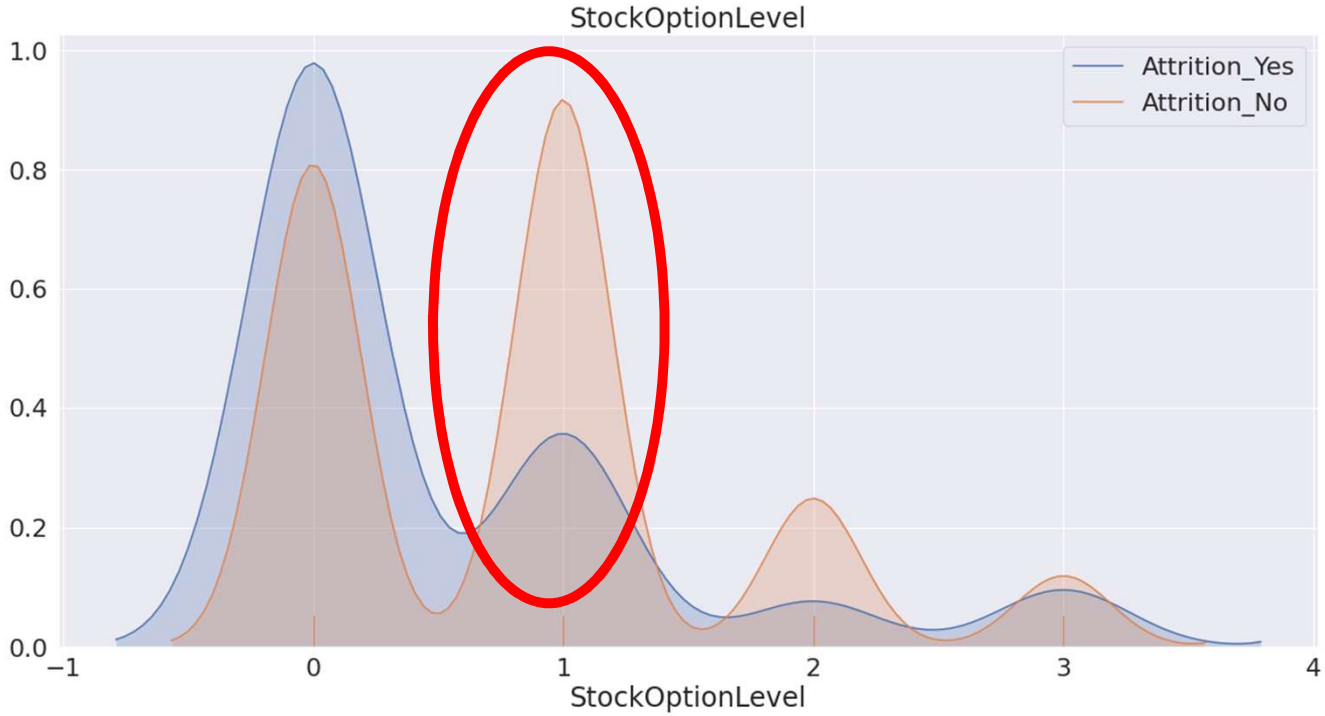
Job level and monthly income features seem interesting too.
With low job level and low monthly income, people are more likely to quit.



KDE Plot

Stock option level is the most interesting feature.

From blue area, we can presume that people are more likely to quit with zero stock option level. Moreover, stock option level 1 in orange area supports our hypothesis



Exploratory Data Analysis

84% 16%

1233
Attrition No

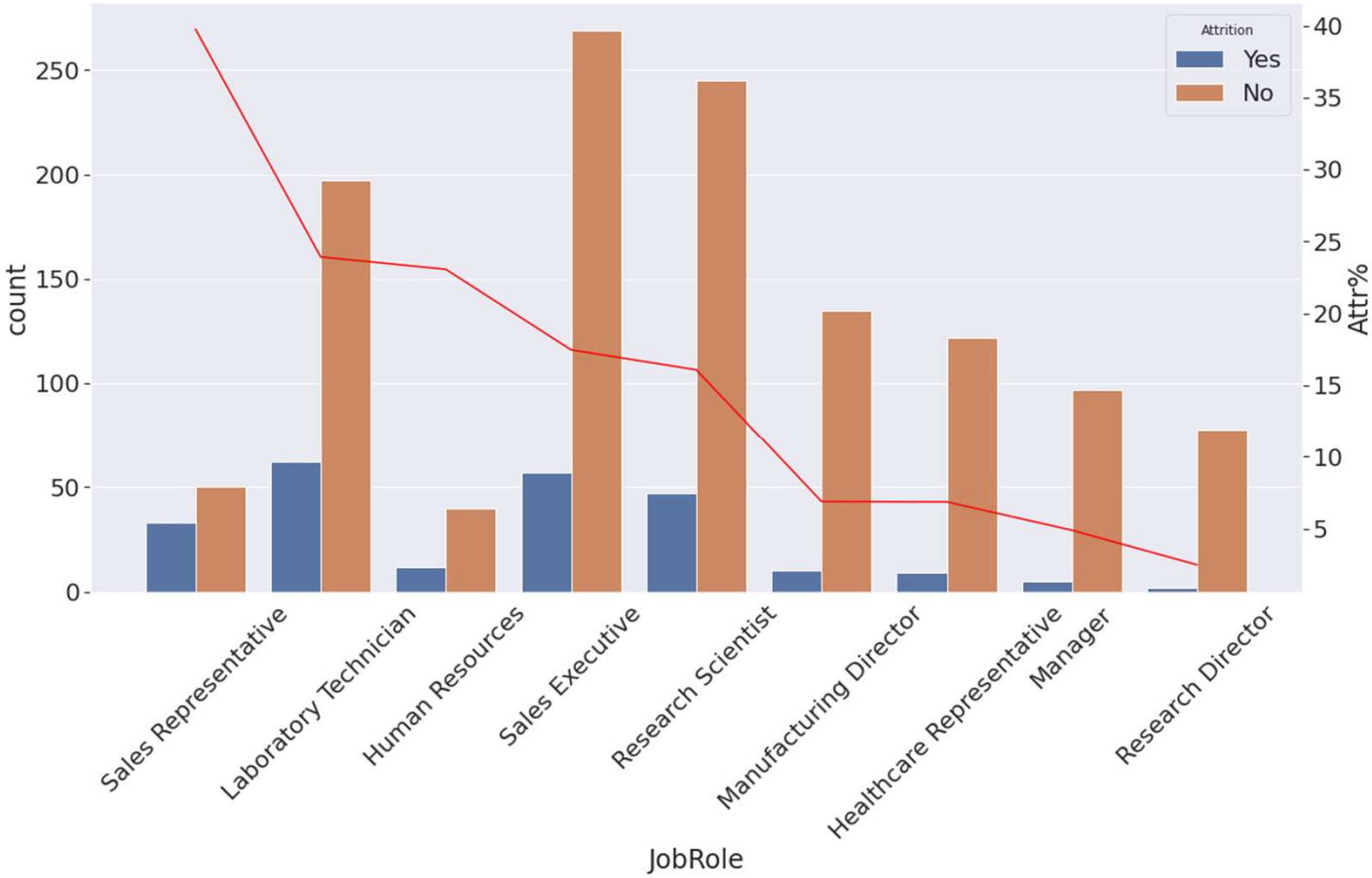
237
Attrition Yes

For categorical features, we use bar plot with percentage of attrition_yes / total_attrition to find insight.

Bar Plot

Job role

Sales representative is the highest attrition percentage at 40% that higher than Laboratory technician by 15%

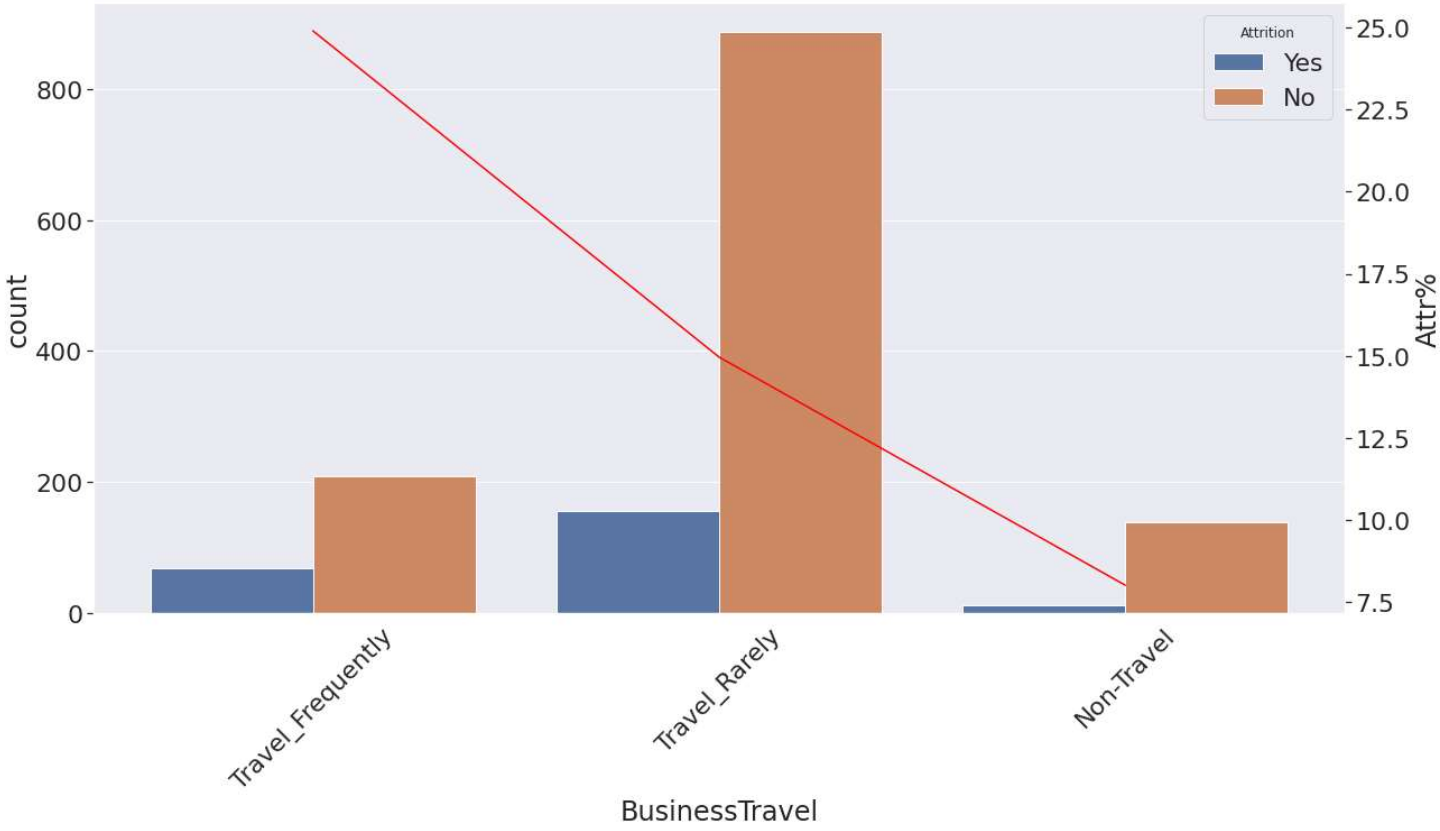


Bar Plot

Business travel

People who travel frequently are higher attrition percentage around 10%.

They may relate to sales representative role.

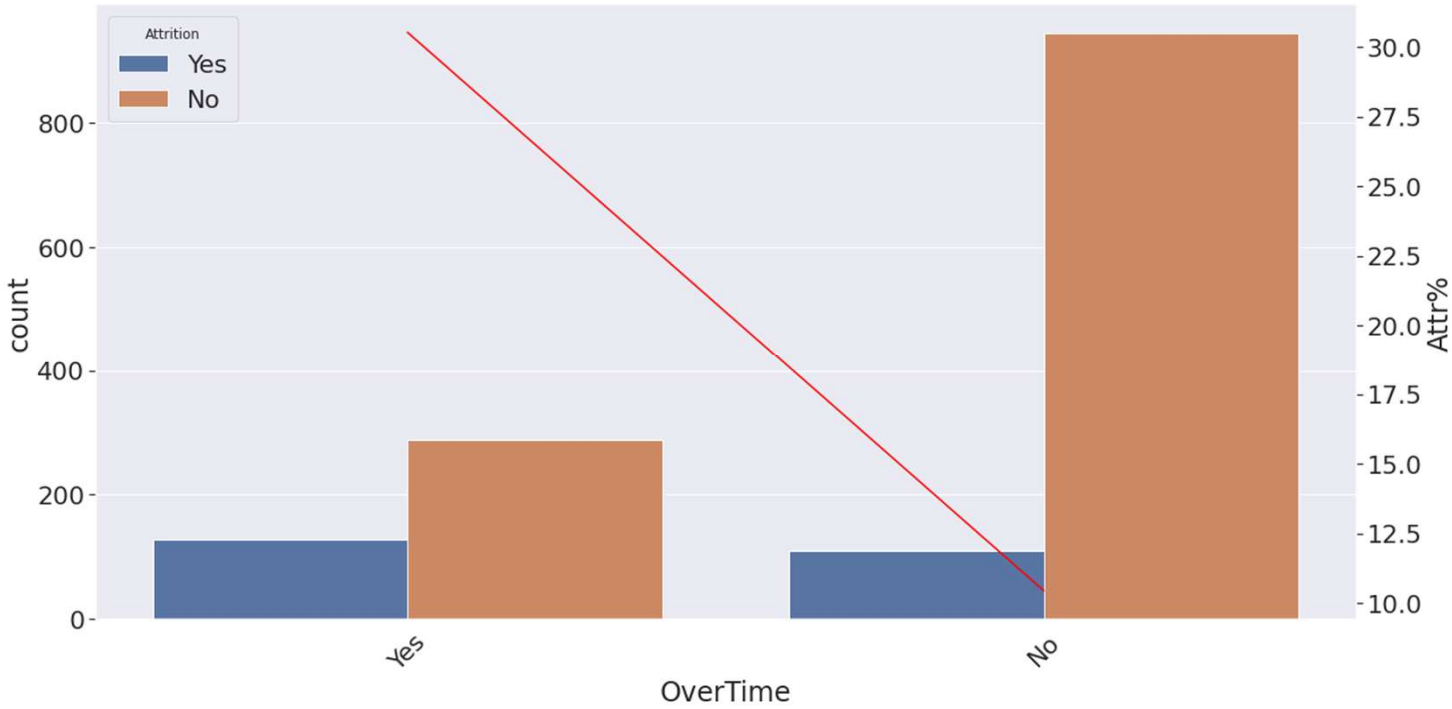


Bar Plot

Overtime

Attrition percentage for overtime is higher than no-overtime by 20%.

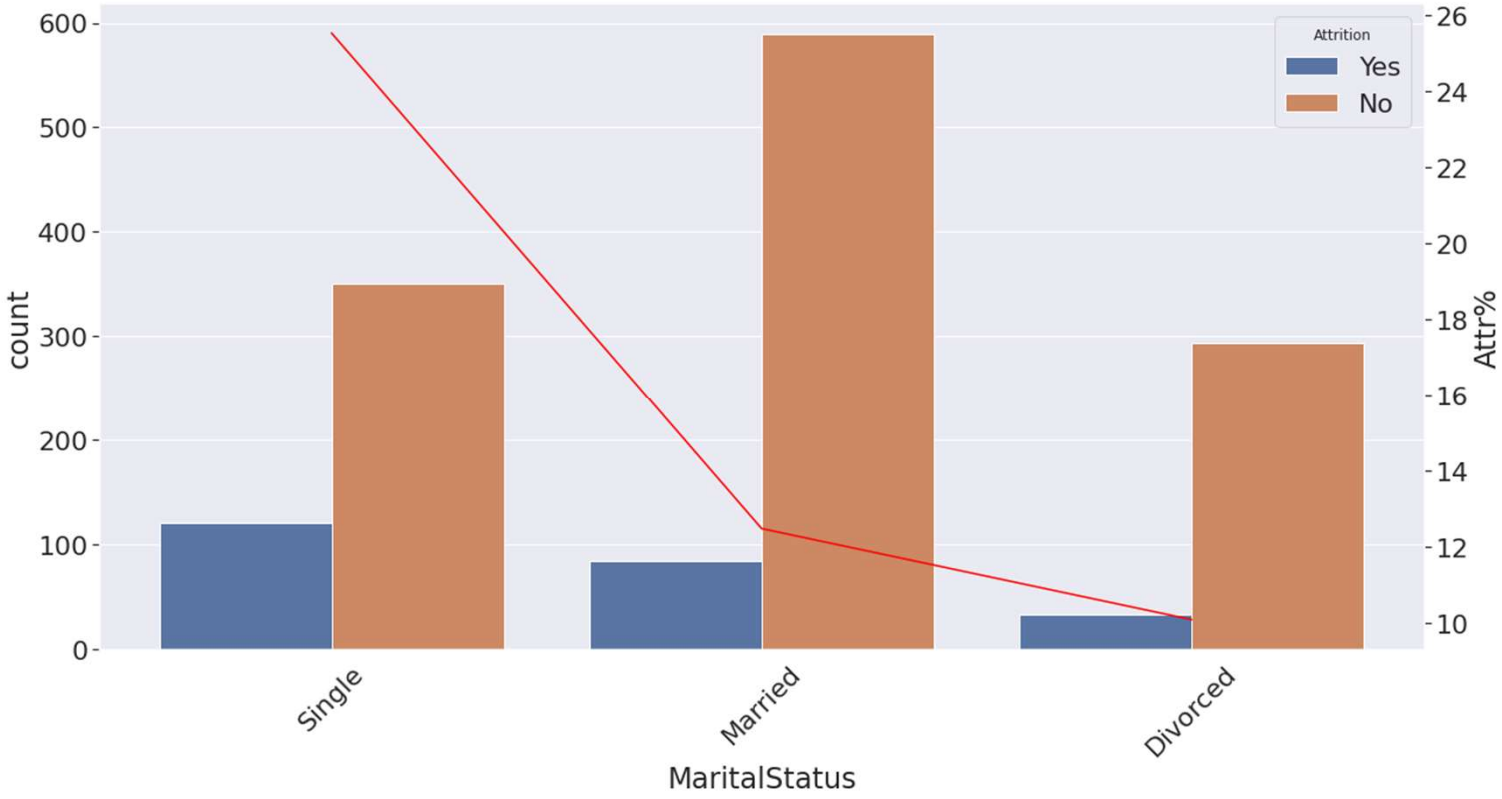
Overtime may affect to other features such as work life balance or job satisfaction.



Bar Plot

Marital status

Single status is higher attrition percentage than married status by 14%.



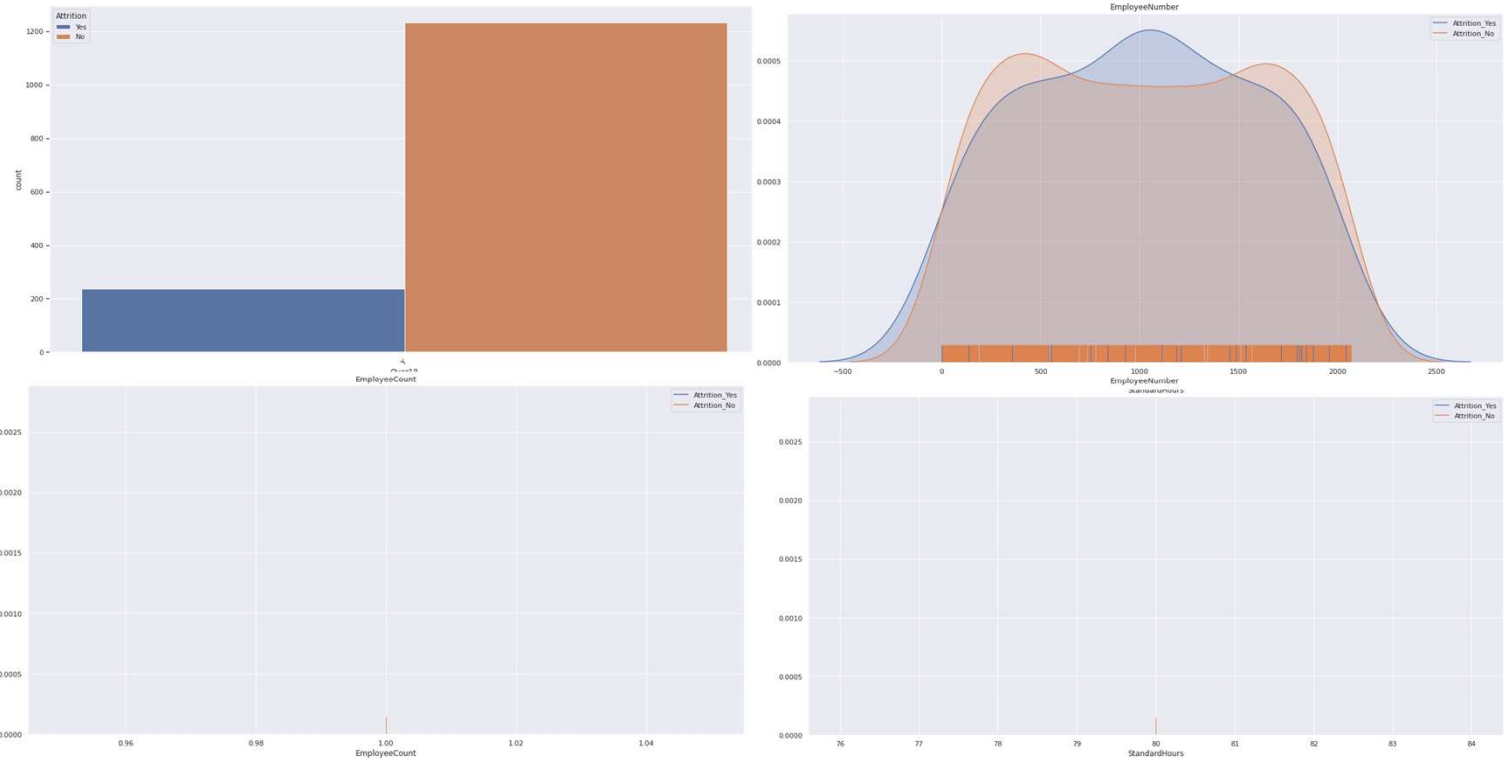
Feature Engineering & Data Preprocessing

Feature Engineering

Drop some unnecessary features

From the KDE plot, we will see some features that all values are the same or not have meaning.

- **Standard Hours**
- **Employee Count**
- **Over18**
- **Employee Number**

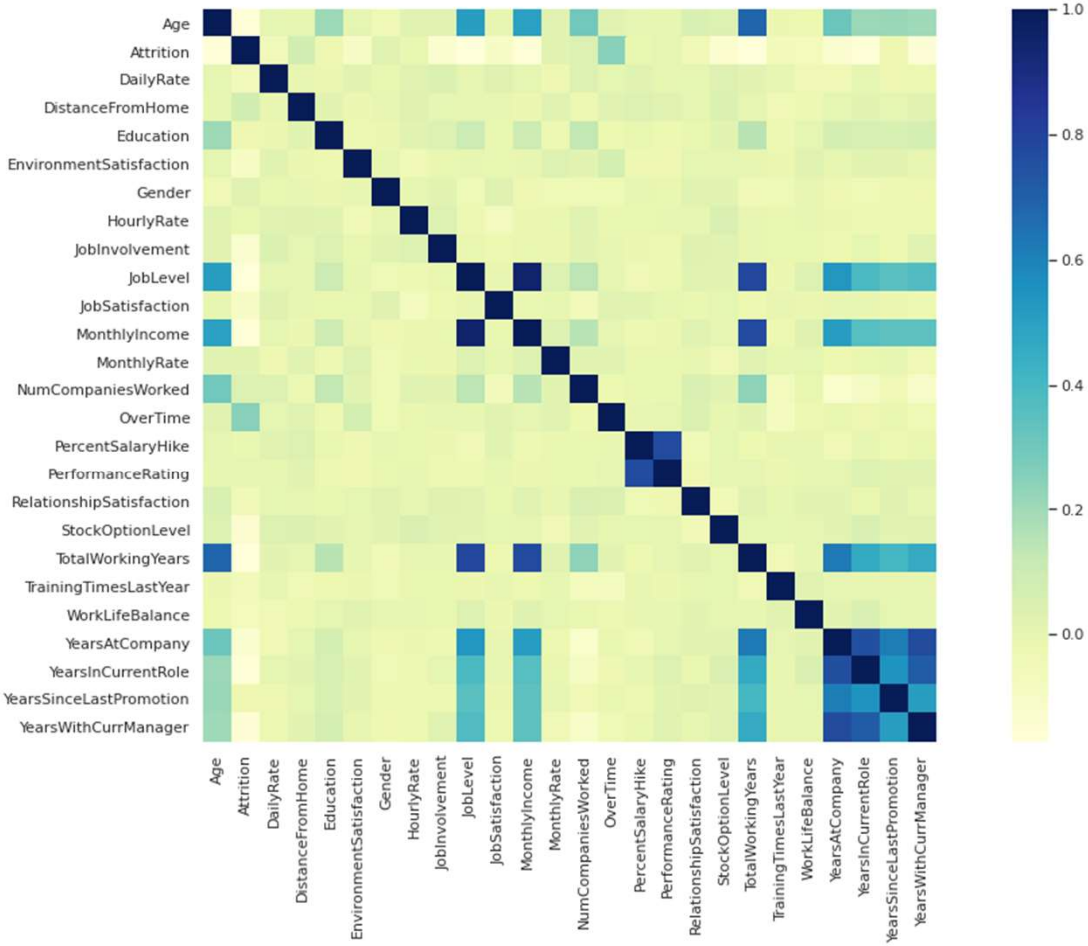


Feature Engineering

Drop colinear features

In order to accurate our machine learning model, we need to drop features that highly related to one another.

- Job Level
- Monthly Income
- Performance Rating
- Total Working Years
- Years At Company
- Years In Current Role
- Years Since Last Promotion
- Years With Current Manager



Data Preprocessing

Encode categorical features

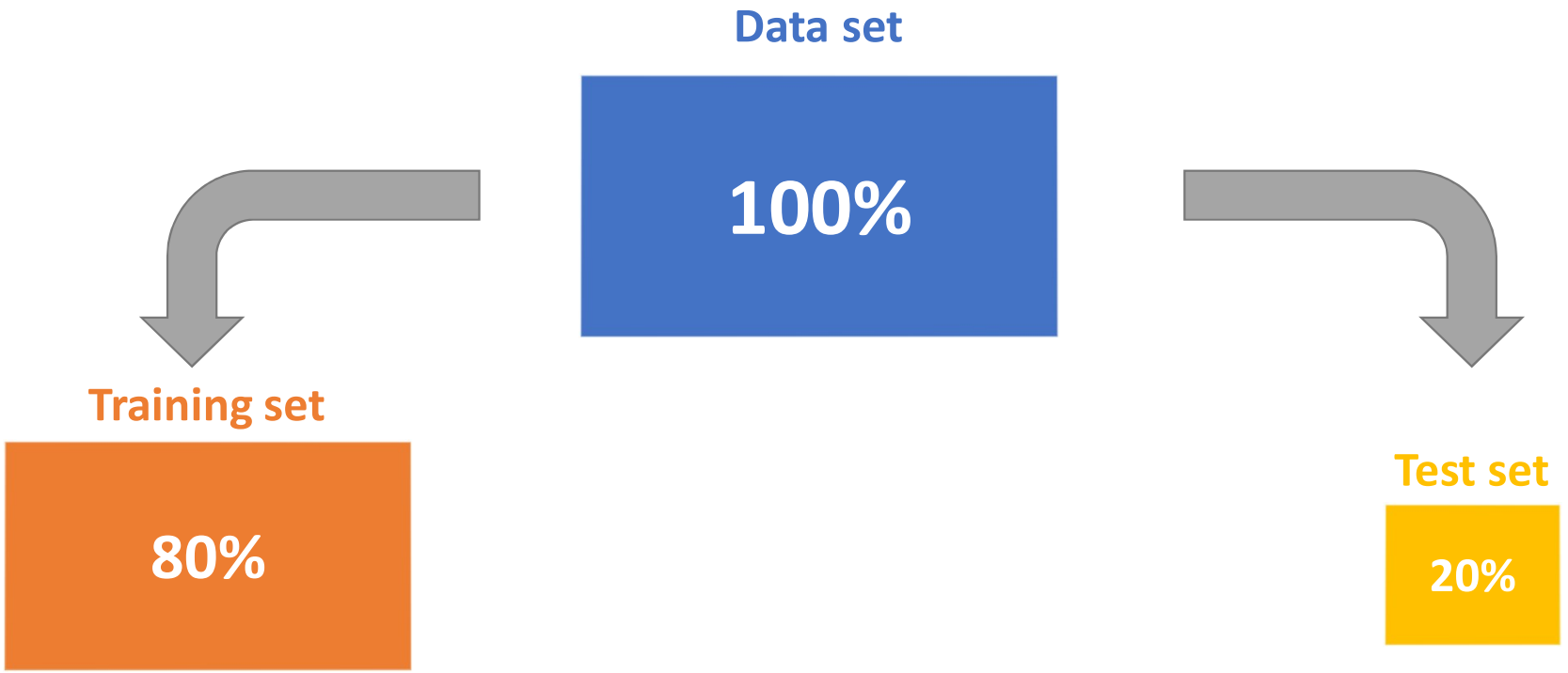
In order to train machine learning model, we need to use numerical data. So, we numerically encode all categorical features.

	BusinessTravel_Non-Travel	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely	Department_Human Resources	Department_Resea & Developm
0	0	0	1	0	
1	0	1	0	0	
2	0	0	1	0	
3	0	1	0	0	
4	0	0	1	0	

Machine Learning

Splitting Data

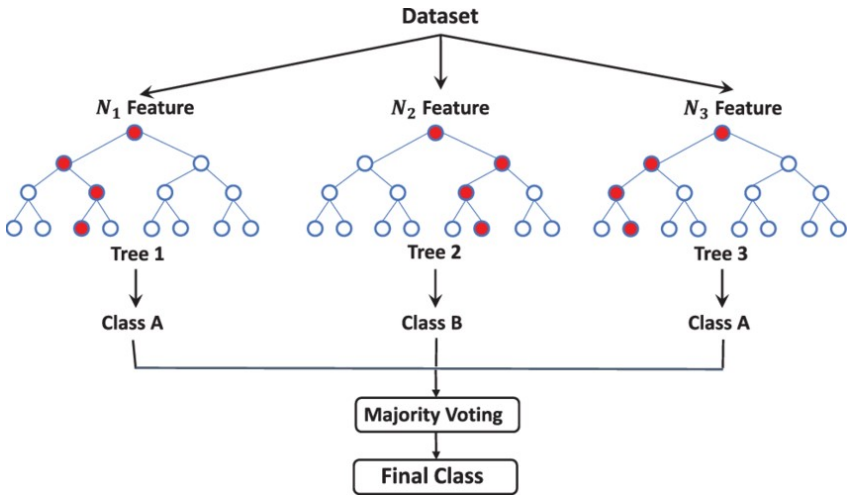
Split our dataset into training set and test set



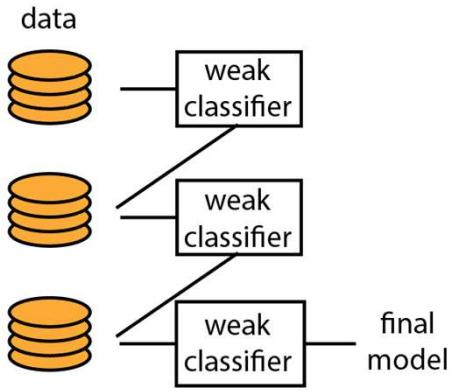
Model

In order to find the factor that lead to employee attrition, we can perform machine learning model and see "What features are importance for model?" and "It is the same as our hypothesis from data analysis or not? ".

Random Forest Model

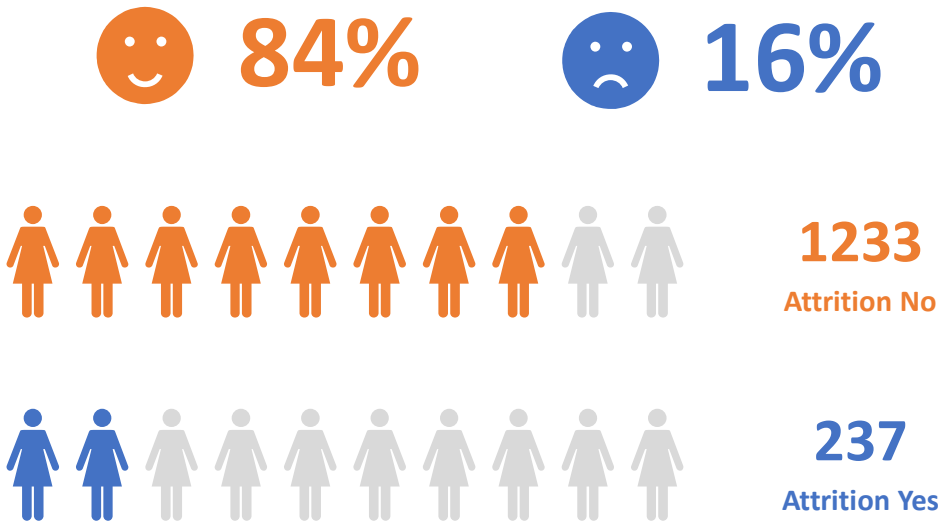


Gradient Boosting Model



Resampling training data

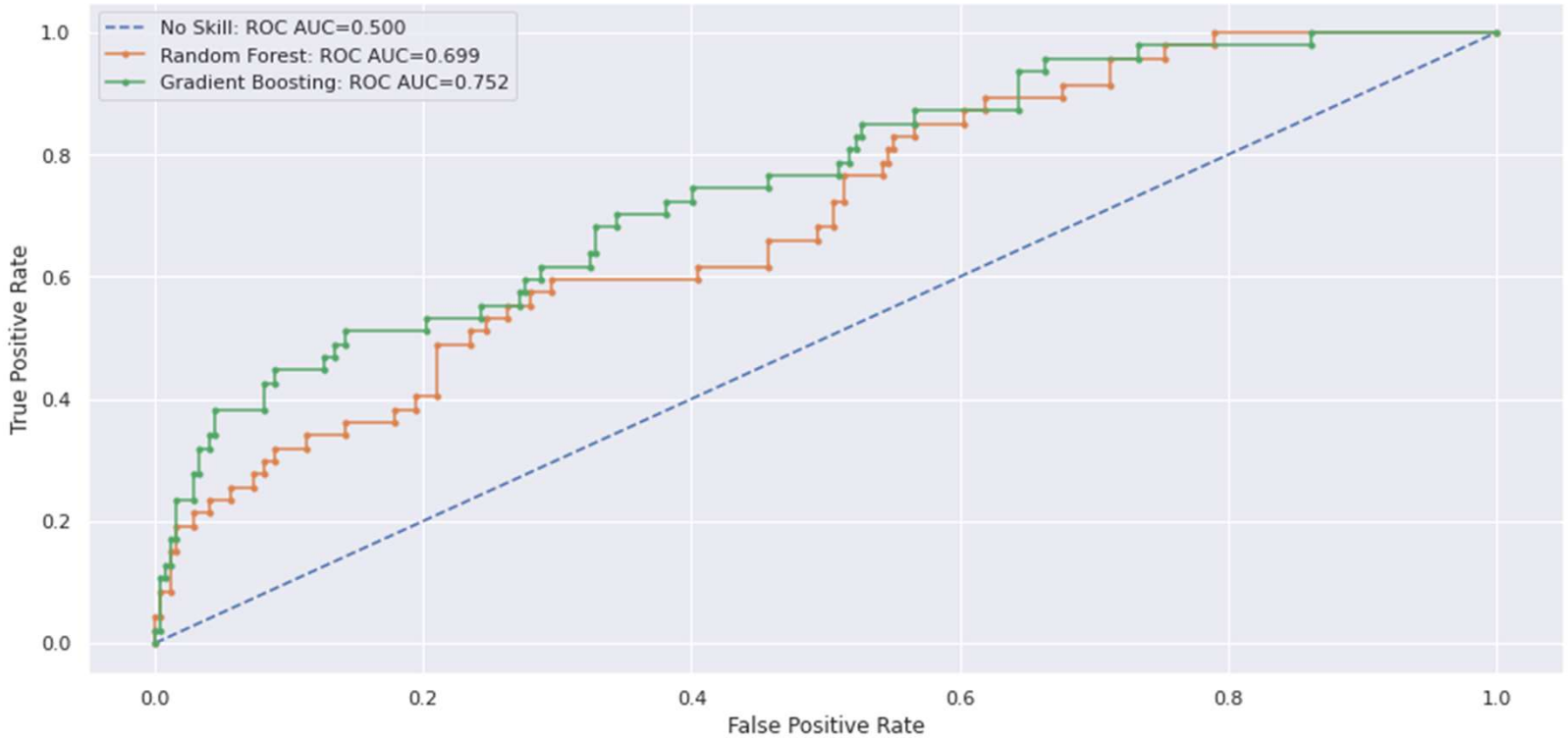
In order to reduce imbalance target problem, we use over-sampling and under-sampling process. Over-sampling increases the weight of the minority class by replicating the minority class examples and Under-sampling decreases the weight of the majority class by cutting the majority class examples



Result

ROC Curve

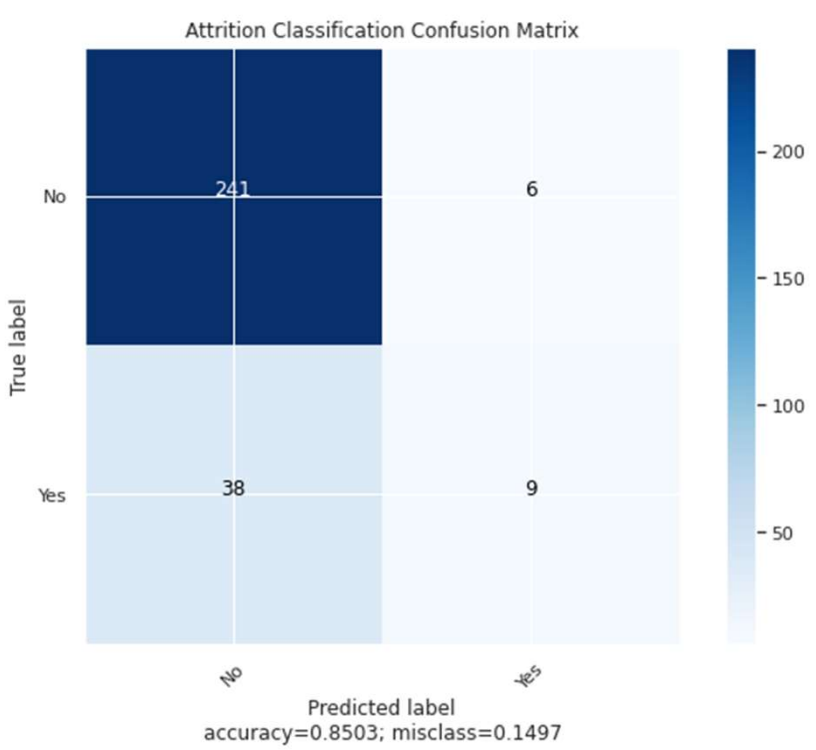
ROC scores from Gradient Boosting model is higher than Random Forest model at 0.752



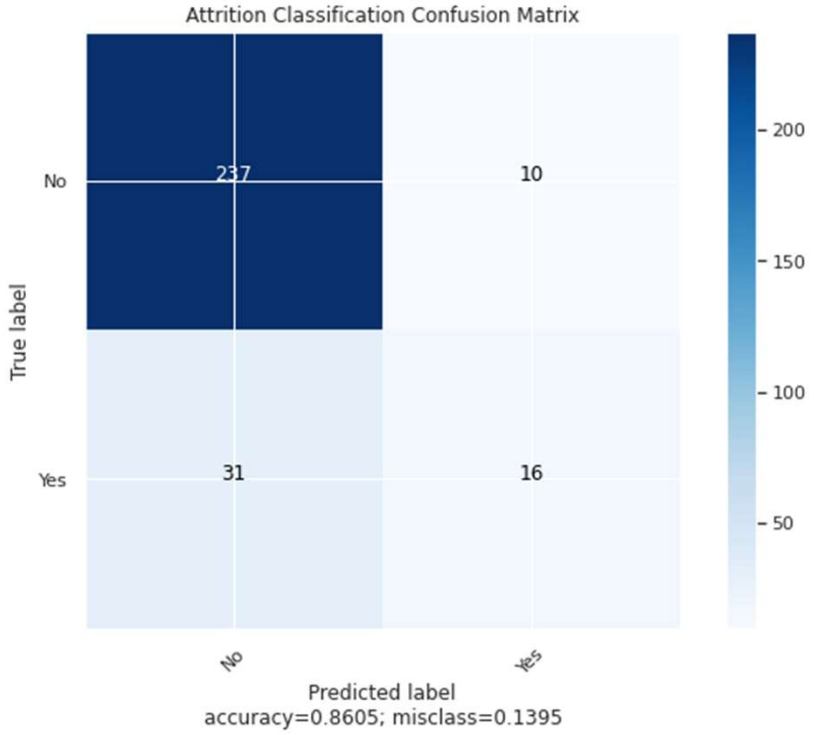
Confusion Matrix

Confusion Matrix **Create from test set data that we split before training**

Random Forest Model

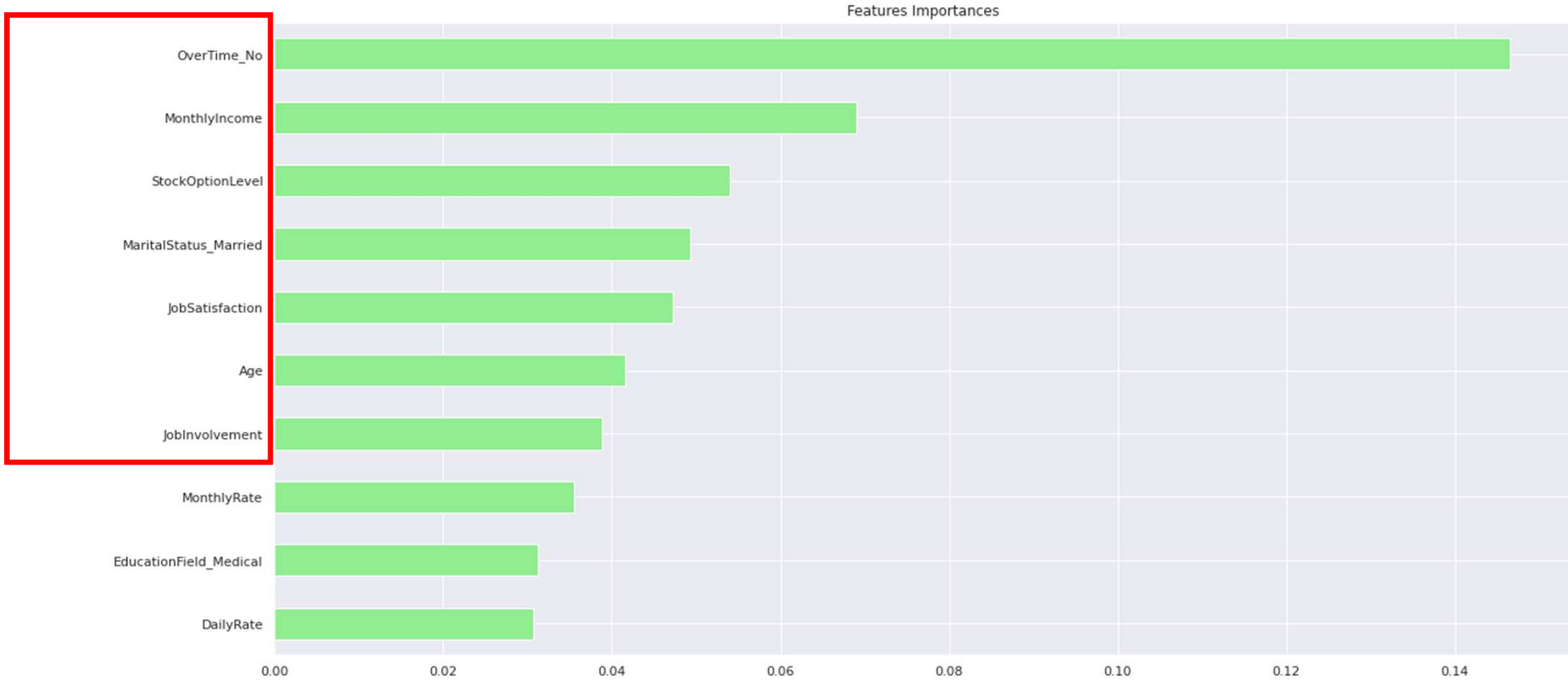


Gradient Boosting Model



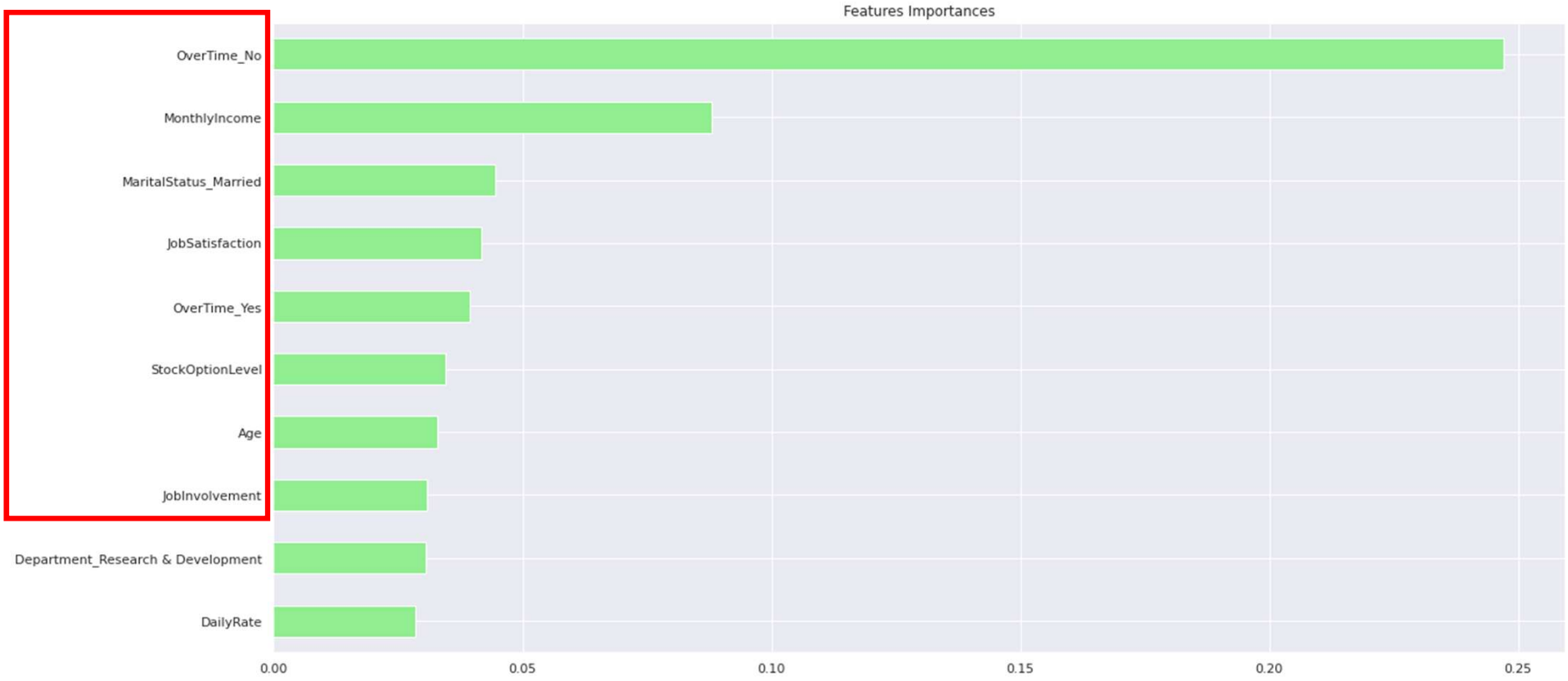
Importance Features

Random Forest Model



Importance Features

Gradient Boosting Model



Conclusion

Conclusion

Our Problem..

01 What is the factor that lead to employee attrition?

02 How can we keep employees happy and satisfied to company?

EDA

From EDA, we have 7 importance features

- Age and other that related to age
- Job level
- Monthly Income
- Stock Option Level
- Sales Representative job role
- Overtime
- Marital status

Conclusion

Machin Learning Model

From our two models, we have importance features that are the same from different model

- **Overtime = No**
- **Monthly Income**
- **Stock Option Level**
- **Job Satisfaction**
- **Married**
- **Age**
- **Job Involvement**

Conclusion

01 What is the factor that lead to employee attrition?

Expected

- Overtime
- Monthly Income ~~(Job level)~~
- Stock Option Level
- Marital status
- Age

Surprised

- Job Satisfaction
- Job Involvement

- ~~Sales Representative job role~~

02 How can we keep employees happy and satisfied to company?

The most importance factor is **Overtime**. To improve **Monthly Income** and **Stock Option**, we need to invest more money. But to reduce Overtime, we may just need to improve working process.

Job Satisfaction and **Job Involvement** are interesting factor. This two factors seem not related to other factor. So, we need to dig deeper to know “What affect to this two factors”.

THANK YOU